Four Reflections
The Road to a Complete Unification of Physics

© 2019 Copyright Ty K Coburn

Until and unless other notice is given, copies of material found herein may only be made by students, instructors, and others solely for their personal use. To include any original material, or original presentation of known material, contained herein in a publication, written prior approval from Ty K Coburn, or his heirs, is required.

All Rights Reserved. No part of this book may be reproduced, stored in a retrieval system, transmitted or translated into machine language, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the author, his heirs if he is no longer alive, or the publisher, except for brief quotations embodied in research articles and reviews. Any permissible public use must cite this book as a source.
BOOK III: The Third Reflection

MODERN PHYSICS

First Addition
Table of Contents

Chapter 1 .............................................................................................................................................. 20
The General Approach to Physics and the Concepts of ‘Motion’ and ‘Rest’ .............................................. 20
1.0 Introduction .................................................................................................................................. 20
1.1 The Foundations of Physics ........................................................................................................ 20
  1.1.1 Extension ............................................................................................................................... 20
  1.1.2 Time ......................................................................................................................................... 20
  1.1.3 Mass ......................................................................................................................................... 21
1.2 Dimensions ................................................................................................................................... 21
  1.2.1 Systems of Units ..................................................................................................................... 21
  1.2.2 Dimensional Analysis .............................................................................................................. 22
1.3 Constants of Nature ..................................................................................................................... 23
  1.3.1 Dimensionless Constants ........................................................................................................ 24
1.4 Motion and Rest ............................................................................................................................ 24
  1.4.1 Constant Motion in a Straight Line ......................................................................................... 25
  1.4.2 Instantaneous Velocity ............................................................................................................ 25
  1.4.3 Acceleration ............................................................................................................................. 26
    1.4.3.1 Constant Acceleration ...................................................................................................... 26
1.5 Angular Velocity and Angular Acceleration ................................................................................... 27
  1.5.1 Instantaneous Angular Velocity ............................................................................................. 28
  1.5.2 Instantaneous Angular Acceleration ...................................................................................... 28
  1.5.3 Constant Angular Motion ...................................................................................................... 28
1.6 The Relationship between Angular Speed and Linear Speed ......................................................... 29
1.7 Harmonic Motion ............................................................................................................................ 30
  1.7.1 Modelling Harmonic Motion using Complex Functions ...................................................... 31
  1.7.2 Simple Harmonic Motion and Phase ..................................................................................... 32
    1.7.2.1 Simple Harmonic Motion along Different Axes ................................................................. 32
  1.7.3 Wave Motion .......................................................................................................................... 33
    1.7.3.1 The Wave Equation ......................................................................................................... 34
    1.7.3.2 Wave Dispersion and Group Velocity .............................................................................. 34
    1.7.3.3 Standing Waves ............................................................................................................... 36
1.8 Concluding Remarks ...................................................................................................................... 36
Chapter 2 ............................................................................................................................................... 38
5.8 Electromagnetic Induction

5.6 Electromotive Force and Circuits

5.5 Moving Charges

5.4 Electric Potential

5.3 The Electric Field and Gauss’ Law

5.2 Electric Current

5.1 Electrostatics

5.0 Introduction

Electricity and Magnetism
7.1.1 The Three Fundamental Laws of Newtonian Mechanics ............................................ 121
7.1.2 The Tacit Assumptions of Newtonian Physics .......................................................... 121
  7.1.2.1 Inertial Reference Frames ..................................................................................... 121
  7.1.2.2 The Time Parameter ............................................................................................. 121
  7.1.2.3 The Relationship between Force and Mass .......................................................... 122
7.1.3 The Laws of Motion for a Constant Force ................................................................. 123
  7.1.3.1 Kinds of Acceleration ............................................................................................ 123
7.1.4 Other Inertial Reference Frames .................................................................................. 124
7.1.5 Work, Power and Kinetic Energy .................................................................................. 125
  7.1.5.1 Conservation of Momentum and Energy ............................................................... 125
7.1.6 Newtonian Mechanics – A Summary ......................................................................... 126
7.2 Central Forces ............................................................................................................... 126
  7.2.1 The Laws of Central Forces Written in Polar Coordinates ....................................... 127
  7.2.2 Laws of Planetary Motion ........................................................................................ 128
    7.2.2.1 Kepler’s Three Laws of Planetary Motion .......................................................... 129
7.3 The Simple Harmonic Oscillator .................................................................................... 130
  7.3.1 Hooke’s Law ............................................................................................................ 131
  7.3.2 Simple Harmonic Motion as a Conservative System ............................................... 131
7.4 Systems of Bodies ......................................................................................................... 132
  7.4.1 Newton’s Second Law for a System of Bodies .......................................................... 132
  7.4.2 Newton’s First Law for a System of Bodies ............................................................... 134
  7.4.3 Angular Momentum and Torque of a System of Bodies .......................................... 134
    7.4.4 The Energy of a System of Bodies .......................................................................... 135
      7.4.4.1 The Kinetic Energy of a System of Bodies ......................................................... 136
      7.4.4.2 Potential Energy and the Conservation of Energy of a System of Bodies ....... 136
7.5 Mechanics in Non-Inertial Reference Frames ............................................................... 137
  7.5.1 Newton’s Second Law in Non-Inertial Reference Frames ........................................ 138
    7.5.1.1 Newton’s Second Law in a Rotating Frame ....................................................... 139
7.6 Maxwell’s Equations ..................................................................................................... 141
  7.6.1 Gauss’ Law – the First of Maxwell’s Equations ....................................................... 142
    7.6.1.1 Calculating Charge Using Gauss’ Law .............................................................. 143
  7.6.2 Gauss’ Law for the Magnetic Field – the Second of Maxwell’s Equations ............... 143
  7.6.3 Faraday’s Law of Induction – the Third of Maxwell’s Equations ............................. 144
  7.6.4 The Ampere-Maxwell Law – the Fourth of Maxwell’s Equations ......................... 146
7.6.4.1 The Magnetic Field Associated with a Changing Electric Field .................................. 147
7.6.5 A Summary of Maxwell’s Equations ............................................................................ 148
7.7 Electromagnetic Waves .................................................................................................. 149
  7.7.1 The Poynting Vector ................................................................................................. 151
7.8 Generalizations of Classical Mechanics .......................................................................... 152
  7.8.1 Poisson’s Equation .................................................................................................. 153
7.9 Concluding Remarks ...................................................................................................... 155
Chapter 8 .............................................................................................................................. 157
The Principles of Analytical Mechanics ............................................................................... 157
8.0 Introduction ..................................................................................................................... 157
8.1 Curvilinear Coordinates ................................................................................................. 158
  8.1.1 Contravariant and Covariant Components of a Vector ............................................. 158
  8.1.2 The Generalized Metric ......................................................................................... 158
    8.1.2.1 The Derivation of the Contravariant Components of a Vector ......................... 158
    8.1.2.2 The Derivation of the Covariant Components of a Vector ............................ 160
    8.1.2.3 The Derivation of the General Metric ........................................................... 162
8.2 Configuration Space ...................................................................................................... 163
  8.2.1 Degrees of Freedom ............................................................................................... 164
8.3 The Fundamental Principle of Classical Mechanics ....................................................... 164
  8.3.1 Generalized Velocity .............................................................................................. 164
  8.3.2 Generalized Force and Kinetic Energy .................................................................... 165
  8.3.3 The Principle of Least Action ................................................................................ 166
8.4 The Hamiltonian Formalism of Classical Mechanics ..................................................... 167
8.5 Poisson Brackets .......................................................................................................... 170
  8.5.1 The Rules for Poisson Brackets ............................................................................ 171
  8.5.2 Canonically Conjugate Variables .......................................................................... 171
    8.5.2.1 The Idea of a Symmetry .................................................................................. 172
8.6 Concluding Remarks ..................................................................................................... 173
Chapter 9 .............................................................................................................................. 176
The Theory of Special Relativity .......................................................................................... 176
9.0 Introduction ..................................................................................................................... 176
9.1 Wave and Particle Physics ............................................................................................ 176
  9.1.1 The Motion of a Particle ......................................................................................... 176
  9.1.2 The Motion of a Wave ............................................................................................ 177
9.2 The Michelson and Morley Experiment ................................................................. 177
  9.2.1 Calculating Fringe Shift .................................................................................. 179
9.3 The Consequences of the Michelson and Morley Experiment .......................... 181
  9.3.1 Violation of Galilean Relativity ..................................................................... 181
  9.3.2 The Lorentz Theory ...................................................................................... 182
9.4 Einstein’s Principle of Relativity ............................................................. 182
  9.4.1 The Lorentz Transformations ...................................................................... 183
    9.4.1.1 Time Dilation .......................................................................................... 184
    9.4.1.2 Length Contraction .............................................................................. 185
  9.4.2 The Twin Paradox ....................................................................................... 185
  9.4.3 All Observers will Measure the Speed of Light to be $c$ ............................. 185
  9.4.4 The Oddities of Special Relativity .............................................................. 186
9.5 Space-Time ........................................................................................................ 187
  9.5.1 World Points and World Lines ............................................................... 187
    9.5.2 The Equivalence of Space-Time with the Special Theory of Relativity .... 188
      9.5.2.1 The Synchronization of Clocks ............................................................. 188
      9.5.2.2 Deriving the Lorentz Transformations from Space-Time ............... 189
  9.5.3 Invariance in Space-Time ............................................................................ 190
9.6 Proper Time ....................................................................................................... 192
  9.6.1 Proper Time in Three Dimensions ............................................................... 192
    9.6.1.1 Light-Like Distances ............................................................................. 193
    9.6.1.1 Time-Like Distances ............................................................................. 193
    9.6.1.2 Space-Like Distances ........................................................................... 194
9.7 4-Vectors ............................................................................................................ 194
  9.7.1 The 4-Velocity ............................................................................................. 195
    9.7.1.1 The Relationship between 4-Velocity and 3-Velocity ...................... 195
9.8 Applying Hamilton’s Principle to Space-Time ............................................. 196
  9.8.1 Conservation of the 4-Momentum ............................................................... 197
  9.8.2 Energy in Space-Time ................................................................................. 198
9.9 Classical Field Theory and Space-Time .................................................... 199
  9.9.1 The Action for a Field ............................................................................... 200
  9.9.2 The Euler-Lagrange Equations for a Field ............................................... 200
  9.9.3 The Principle of Least Action as Applied to Special Relativity .............. 201
    9.9.3.1 Particle/Field Interactions in Special Relativity ............................... 202
10.5 Mathematical Preliminaries

10.4 Flat vs. Curved Spaces and Coordinate

10.3 The Concept of Curvature

10.2 Gravity and the Curvature of Space

10.1 Flat Ear...

9.10 A Notational Excursion

9.11 The Klein-Gordon Equation

9.12 Electro-Dynamics in Special Relativity

9.12.1 The Principles of Relativistic Physics

9.12.2 The Vector Potential and the Lorentz Force Law

9.12.2.1 The Proper Force

9.12.3 Gauge Invariance

9.13 A Relook at Maxwell’s Equations in Terms of Special Relativity

9.13.1 The Vector Potential and Maxwell’s Equations

9.13.2 The Bianchi Identity

9.13.3 The Last Two of Maxwell’s Equations

9.14 Concluding Remarks

Chapter 10

The General Theory of Relativity

10.0 Introduction

10.1 Flat Earth Gravity

10.1.1 The Principle of Equivalence

10.1.2 Distance Dependent and Tidal Forces

10.1.3 The Field Formulation of Newton’s Law

10.1.3.1 Poisson’s Equation

10.2 Gravity and the Curvature of Space

10.2.1 The Path of a Light Ray in a Uniform Gravitational Field

10.2.1.1 Uniform Acceleration and Space-Time

10.3 The Concept of Curvature

10.3.1 Extrinsic vs. Intrinsic Geometry

10.3.2 Distance as an Intrinsic Concept

10.4 Flat vs. Curved Spaces and Coordinate Transformations

10.4.1 The Consequences of Describing Gravity Intrinsically

10.5 Mathematical Preliminaries

10.5.1 The Gradient of a Function

10.5.2 Tensors Revisited

10.5.2.1 Scalars

10.5.2.2 Vectors

10.5.2.3 Higher Ranked Tensors
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.9.2 The Time Evolution of Quantum Systems</td>
<td>273</td>
</tr>
<tr>
<td>11.9.2.1 The Dirac Delta Function</td>
<td>274</td>
</tr>
<tr>
<td>11.9.3 The Generalization of Eigenvalue Equations in Quantum Mechanics</td>
<td>275</td>
</tr>
<tr>
<td>11.9.3.1 The Momentum Operator</td>
<td>276</td>
</tr>
<tr>
<td>11.9.4 The Nature of Operators in Quantum Mechanics</td>
<td>279</td>
</tr>
<tr>
<td>11.10 The Two Slot Interference Experiment (Revisited)</td>
<td>280</td>
</tr>
<tr>
<td>11.10.1 Position and Momentum as Fourier Transforms</td>
<td>281</td>
</tr>
<tr>
<td>11.11 Light Polarization and Quantum Operators</td>
<td>282</td>
</tr>
<tr>
<td>11.11.1 Vertically and Horizontally Polarized Light</td>
<td>283</td>
</tr>
<tr>
<td>11.11.2 The Quantization of Polarization</td>
<td>284</td>
</tr>
<tr>
<td>11.11.3 Circularly and Elliptically Polarized Light</td>
<td>286</td>
</tr>
<tr>
<td>11.11.4 Time Evolution and Schrödinger’s Equation</td>
<td>287</td>
</tr>
<tr>
<td>11.11.4.1 Schrödinger’s Equation</td>
<td>288</td>
</tr>
<tr>
<td>11.11.4.2 The Expected Value of an Operator over Time</td>
<td>291</td>
</tr>
<tr>
<td>11.11.4.3 The Role of Time in Quantum Mechanics</td>
<td>291</td>
</tr>
<tr>
<td>11.11.4.4 Time as a Parameter</td>
<td>292</td>
</tr>
<tr>
<td>11.12 The Relationship between Quantum and Classical Physics</td>
<td>293</td>
</tr>
<tr>
<td>11.13 Symmetry</td>
<td>295</td>
</tr>
<tr>
<td>11.13.1 Symmetry and Time Evolution</td>
<td>295</td>
</tr>
<tr>
<td>11.13.2 Continuous Symmetries</td>
<td>296</td>
</tr>
<tr>
<td>11.13.3 Rotational Symmetry</td>
<td>296</td>
</tr>
<tr>
<td>11.13.3.1 Degenerate States</td>
<td>297</td>
</tr>
<tr>
<td>11.14 Angular Momentum in Quantum Mechanics</td>
<td>298</td>
</tr>
<tr>
<td>11.14.1 Angular Momentum Commutation Operators</td>
<td>300</td>
</tr>
<tr>
<td>11.14.2 Angular Momentum Raising and Lowering Operators</td>
<td>300</td>
</tr>
<tr>
<td>11.14.3 Quantizing Angular Momentum</td>
<td>301</td>
</tr>
<tr>
<td>11.15 A Particle Moving in a Coulomb Potential</td>
<td>302</td>
</tr>
<tr>
<td>11.15.1 Angular Momentum and Energy in a Coulomb Potential</td>
<td>303</td>
</tr>
<tr>
<td>11.16 The Quantum Harmonic Oscillator</td>
<td>304</td>
</tr>
<tr>
<td>11.16.1 Quantizing the Energy</td>
<td>305</td>
</tr>
<tr>
<td>11.16.2 Raising and Lowering Operators for Energy</td>
<td>307</td>
</tr>
<tr>
<td>11.16.2.1 Normalizing Raising and Lowering Energy Operators</td>
<td>308</td>
</tr>
<tr>
<td>11.17 Classical vs. Quantum Logic via Electron Spin</td>
<td>309</td>
</tr>
<tr>
<td>11.17.1 The Qubit and Electron Spin</td>
<td>311</td>
</tr>
</tbody>
</table>
11.17.1.1 Electron Spin is Quantized ................................................................. 312
11.17.2 Information in Quantum Physics is Probabilistic ............................... 312
11.17.3 Measuring Electron Spin .................................................................... 312
   11.17.3.1 Spin Operators ............................................................................... 313
   11.17.3.2 The Pauli Exclusion Principle ......................................................... 313
11.17.4 Measuring Spin in an Arbitrary Direction ............................................. 315
11.18 Quantum Entanglement ....................................................................... 319
11.19 Bell’s Theorem .................................................................................. 320
   11.19.1 Bell’s Inequality Applied to Electron Spin ........................................ 322
   11.19.2 Conducting Electron Spin Experiments ............................................. 322
       11.19.2.1 Results of Electron Spin Experiments ......................................... 324
11.20 Concluding Remarks ........................................................................ 325
Chapter 12 .................................................................................................. 329
Quantum Field Theory ............................................................................. 329
12.0 Introduction ....................................................................................... 329
12.1 Why Quantum Field Theory (QFT)? ................................................... 330
   12.1.1 Quantizing the Field ........................................................................ 330
12.2 The Development of a Relativistic Field Equation .................................. 331
   12.2.1 The Klein-Gordon Equation ............................................................... 332
   12.2.2 The Continuity Equations and Locality ............................................. 333
       12.2.2.1 Relativistic Continuity Equations ................................................ 334
12.3 The Quantum Field ............................................................................. 336
   12.3.1 Periodic Functions ........................................................................... 336
   12.3.2 The Quantum Field as an Operator ................................................ 337
       12.3.2.1 The State Vector in QFT ............................................................... 338
       12.3.2.2 Creating Field Operators .............................................................. 338
       12.3.2.3 Rising and Lowering Energy Operators ......................................... 339
12.4 QFT Overview .................................................................................... 340
   12.4.1 Creation and Annihilation Operators for Bosons ............................... 341
   12.4.2 The Frock Space ............................................................................... 342
   12.4.3 The Conservation of the Inner Product of States over Time ................. 343
   12.4.4 The Number Operator in QFT ........................................................... 343
   12.4.5 The Energy Operator in QFT ............................................................. 344
   12.4.6 Creation and Annihilation Operators of Position and Momentum .......... 344
12.4.7 Momentum Conservation in QFT ................................................................. 345
12.4.8 Coupling Constants .............................................................................. 347
12.4.9 The Ordering of Operators .................................................................. 347
12.4.10 Creation and Annihilation Operators for Fermions ......................... 349
12.5 The Dirac Equation .................................................................................. 350
  12.5.1 Solving the One Direction Problem .................................................. 352
  12.5.2 Solving the Mass Problem ................................................................. 352
  12.5.3 Extending Dirac’s Ideas to Three Dimensions .................................. 353
    12.5.3.1 Pauli Matrices and Spin ............................................................... 353
    12.5.3.2 Spin and Relativity ...................................................................... 354
12.6 Dirac’s Ideas Applied to QFT .................................................................. 354
  12.6.1 Identifying Particles by their Spin and Angular Momentum ............. 355
  12.6.2 Orbital Angular Momentum ............................................................... 355
  12.6.3 Spin for Fermions .............................................................................. 357
  12.6.4 Spin for Bosons ................................................................................ 358
  12.6.5 Spin and Energy ................................................................................ 358
12.7 The Role of the Lagrangian in QFT .......................................................... 359
  12.7.1 The Equations of Motion for a Field .................................................. 359
12.8 The Path Integral Approach ................................................................... 361
  12.8.1 Phasors .............................................................................................. 361
  12.8.2 Calculating the Amplitude with Path Integrals ................................ 361
    12.8.2.1 Sum over Histories .................................................................... 363
    12.8.2.2 The Meaning of the Lagrangian in the Path Integral Approach ...... 364
12.9 Dirac’s Equations and a Lagrangian ....................................................... 364
  12.9.1 Dirac Fields ....................................................................................... 365
12.10 Concluding Remarks ............................................................................. 366
Chapter 13 ....................................................................................................... 368
  The Standard Model of Particle Physics ...................................................... 368
13.0 Introduction .............................................................................................. 368
13.1 The Standard Model Particle Classification Scheme ......................... 370
  13.1.1 Types of Elementary Particles ........................................................... 372
  13.1.2 Quarks ............................................................................................... 372
  13.1.3 Leptons .............................................................................................. 373
13.2 The Sectors of the Standard Model ........................................................ 374
13.3 Quantum Electrodynamics ($QED$) ................................................................................. 375
  13.3.1 $U(1)$ Symmetry ........................................................................................................ 376
13.4 Elementary Quantum Chromodynamics ..................................................................... 376
  13.4.1 The Yukawa Model of the Strong Interaction .......................................................... 377
    13.4.1.1 The Conservation of Baryon Number ................................................................. 378
  13.4.2 The Connection between Isospin and $SU(3)$ Symmetry ........................................ 378
    13.4.2.1 The Isospin Quantum Number ........................................................................... 379
    13.4.2.2 The Introduction of the Quark and the Eight-Fold Way .................................... 380
    13.4.2.3 The Relationship between Spin and Isospin ...................................................... 380
  13.4.3 Giving the Quark a Mass ......................................................................................... 382
  13.4.4 Gluons .................................................................................................................... 382
    13.4.4.1 Color States of Gluons ...................................................................................... 383
    13.4.4.2 The Interactions of Gluons ............................................................................... 383
    13.4.4.3 Gluon Mass ........................................................................................................ 384
13.5 Rotation Operators ..................................................................................................... 384
  13.5.1 The Rotation Group .................................................................................................. 385
  13.5.2 The Invariance of the Inner Product of Rotations .................................................... 386
  13.5.3 Rotations Applied to the Spin of a Particle .............................................................. 386
  13.5.4 The Group $SU(2)$ ................................................................................................ 386
  13.5.5 The Pauli Matrices .................................................................................................. 387
  13.5.6 The Group $SU(3)$ ................................................................................................ 388
    13.5.6.1 $SU(3)$ Representations .................................................................................... 388
13.6 The Postulates of Quantum Chromodynamics .............................................................. 389
  13.6.1 Gauge Theories ....................................................................................................... 390
  13.6.2 Gluon Fields ............................................................................................................ 391
  13.6.3 The $QCD$ Lagrangian ............................................................................................ 392
13.7 The Weak Interaction ................................................................................................ 393
  13.7.1 Parity Violation in Weak Interactions ..................................................................... 393
  13.7.2 The Cabibbo Model ................................................................................................ 395
  13.7.3 The Rise of the Electroweak Theory ..................................................................... 396
  13.7.4 The Intermediate Vector Boson Theory ................................................................ 396
13.8 A Summary of the Standard Model of Particle Physics ............................................. 399
  13.8.1 Permissible Interactions ......................................................................................... 400
  13.8.2 Coupling Constants ............................................................................................... 402
13.8.3 The Anatomy of Feynman Diagrams ......................................................... 402
13.8.4 Quantum Tunelling .................................................................................. 404
13.8.5 Symmetry Breaking .................................................................................. 405
  13.8.5.1 Types of Symmetry Breaking .............................................................. 405
  13.8.5.2 Symmetry Breaking within a Lagrangian ............................................. 406
  13.8.5.3 The Goldstone Boson and Mass ......................................................... 408
13.8.6 Gauge Invariance ..................................................................................... 410
13.8.7 The Vector Potential ............................................................................... 410
13.8.8 The Higgs Mechanism ............................................................................ 411
  13.8.8.1 The Dispersion Relation ................................................................... 412
  13.8.8.2 The Higgs Field ................................................................................ 412
  13.8.8.3 Fermion Mass .................................................................................... 414
  13.8.8.4 Chirality and Helicity ....................................................................... 418
  13.8.8.5 The Dynamics of the Higgs Boson .................................................... 419
13.9 The Relative Strength of the Interactions ................................................. 420
13.10 Concluding Remarks ................................................................................ 421
Chapter 14 ......................................................................................................... 427
String Theory ...................................................................................................... 427
14.0 Introduction ................................................................................................ 427
14.1 Regge Trajectories and Fundamental Particles ......................................... 427
14.2 The Concept of a ‘String’ .......................................................................... 428
14.3 Open Strings ............................................................................................... 429
  14.3.1 What is a Particle? .............................................................................. 430
  14.3.2 The Mathematics of an Open String ................................................... 431
  14.3.2.1 The Dirichlet and Neumann Boundary Conditions ............................ 431
  14.3.2.2 Fourier Analysis of an Open String .................................................. 432
  14.3.2.3 The Energy of an Open String .......................................................... 432
  14.3.2.4 The Force on the Ends of an Open String ........................................ 433
  14.3.2.5 The Energy of an Open String - Boosted and Non-Boosted ............. 433
  14.3.3 Quantizing the Open String .................................................................. 434
  14.3.3.1 Creation and Annihilation Operators for an Open String ............... 435
  14.3.3.2 The Spin of a Massless Particle ....................................................... 437
  14.3.3.3 The Tachyon .................................................................................... 438
14.4 Closed Strings ............................................................................................ 439
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.4.1</td>
<td>String Coupling Constant</td>
<td>440</td>
</tr>
<tr>
<td>14.4.2</td>
<td>The Mathematics of Closed Strings</td>
<td>440</td>
</tr>
<tr>
<td>14.4.2.1</td>
<td>The Lagrangian for a Closed String</td>
<td>441</td>
</tr>
<tr>
<td>14.4.2.2</td>
<td>The Energy in a Closed String</td>
<td>441</td>
</tr>
<tr>
<td>14.4.2.3</td>
<td>Level Matching</td>
<td>442</td>
</tr>
<tr>
<td>14.5</td>
<td>String Tension</td>
<td>443</td>
</tr>
<tr>
<td>14.5.1</td>
<td>The Units Used in String Theory</td>
<td>444</td>
</tr>
<tr>
<td>14.6</td>
<td>Why Does String Theory have 25 Spatial Dimensions?</td>
<td>445</td>
</tr>
<tr>
<td>14.7</td>
<td>The Veneziano Amplitude</td>
<td>446</td>
</tr>
<tr>
<td>14.7.1</td>
<td>Mandelstam Variables</td>
<td>447</td>
</tr>
<tr>
<td>14.7.2</td>
<td>$S$ and $T$ Channel Amplitudes</td>
<td>449</td>
</tr>
<tr>
<td>14.7.3</td>
<td>The Physics of the Veneziano Amplitude</td>
<td>450</td>
</tr>
<tr>
<td>14.7.3.1</td>
<td>The Breaking and Joining of Strings</td>
<td>450</td>
</tr>
<tr>
<td>14.8</td>
<td>Conformal Mappings</td>
<td>452</td>
</tr>
<tr>
<td>14.8.1</td>
<td>The Theory of Conformal Mappings</td>
<td>455</td>
</tr>
<tr>
<td>14.8.1.1</td>
<td>The Role of $ln z$ in String Theory</td>
<td>457</td>
</tr>
<tr>
<td>14.8.1.2</td>
<td>Conformal Mappings in String Theory</td>
<td>458</td>
</tr>
<tr>
<td>14.9</td>
<td>String Theory, Gravity and Multiple Compactified Dimensions</td>
<td>460</td>
</tr>
<tr>
<td>14.9.1</td>
<td>String Theory in Flat Geometries</td>
<td>460</td>
</tr>
<tr>
<td>14.9.2</td>
<td>String Theory in Curved Geometries</td>
<td>461</td>
</tr>
<tr>
<td>14.9.3</td>
<td>Conditions for a Sensible String Theory</td>
<td>462</td>
</tr>
<tr>
<td>14.10</td>
<td>Compactification</td>
<td>463</td>
</tr>
<tr>
<td>14.10.1</td>
<td>The Torus as a Compactified Space</td>
<td>463</td>
</tr>
<tr>
<td>14.10.2</td>
<td>Momentum in the Compactified Dimension</td>
<td>464</td>
</tr>
<tr>
<td>14.10.2.1</td>
<td>$T$-Dual Strings</td>
<td>465</td>
</tr>
<tr>
<td>10.11</td>
<td>$T$-Duality and $D$-Branes</td>
<td>465</td>
</tr>
<tr>
<td>10.11.1</td>
<td>Winding Number</td>
<td>465</td>
</tr>
<tr>
<td>14.11.2</td>
<td>Radius of Compactification</td>
<td>465</td>
</tr>
<tr>
<td>14.11.3</td>
<td>The Significance of Wound and Unwound Strings</td>
<td>466</td>
</tr>
<tr>
<td>14.11.4</td>
<td>$T$-Dual Theories for Open Strings and $D_n$-Branes</td>
<td>467</td>
</tr>
<tr>
<td>14.11.4.1</td>
<td>$D_n$-Branes</td>
<td>468</td>
</tr>
<tr>
<td>14.12</td>
<td>Concluding Remarks</td>
<td>469</td>
</tr>
</tbody>
</table>
Chapter 1

The General Approach to Physics and the Concepts of ‘Motion’ and ‘Rest’

“Consequently, he who wishes to attain to human perfection must therefore first study Logic, next the various branches of Mathematics in their proper order, then Physics, and lastly Metaphysics.”

- Maimonides

1.0 Introduction

Physics is written in the language of mathematics. Mathematicians construct formal systems that are generally considered inventions of the human mind. A mathematician would consider the statement \( a + b = c \) as expressing an unambiguous relationship between the elements \( a, b, c \). A physicist, however, employs numbers associated with a dimension, something that can be measured. The equation \( a \,[\text{cm}] + b \,[\text{sec}] = c \) is not physically meaningful, since numbers associated with different dimensions cannot be added together. The physicist’s job is further compounded, since the abstract mathematical inventions used in describing the physical world must, in some sense, mirror measurements taken independently of the mathematics. Mathematicians rarely concern themselves with elements outside the formal system of interest. But physicists describe phenomena external to a given mathematical system and generally regard mathematics as a means to an end. Ideally, the physicist would prefer a mathematical system that can be put into one-to-one correspondence with the measurements associated with a physical system of interest. This makes describing the system using mathematics much easier.

1.1 The Foundations of Physics

Physics is simple. Virtually the entire subject relies on only three notions: ‘space’, ‘time’ and ‘matter’. These notions, however, are subject to degrees of belief and perspective. It is generally agreed that the three notions are understood through dimension, in other words, by how they are measured.

1.1.1 Extension

The first of these notions is ‘extension’ (space), which is, metaphysically, the simplest to understand. Its dimension is ‘length’ denoted by \([L]\). Length has a standard measure: the ‘meter’. The meter is the length light travels in a vacuum during the time interval \(1/299,792,458 \,[\text{sec}]\), where 299,792,458 is the measured speed of light [9].

1.1.2 Time

The second notion ‘time’, denoted \([T]\), is measured in both a linear and periodic manner. Time orders the events in space; this event occurs after, before or simultaneously with that event etc. There is a sense that time is unidirectional (linear).
Yet, in general, it is measured periodically, like the time it takes the moon to complete one orbit around the Earth or for the Earth to complete one orbit of the Sun etc. In non-relativistic physics, space and time are separate notions. While two observers might disagree on the time of an event, no observer would claim that a unique event in space happened at two different times as measured by a single clock. In the non-relativistic physics, space is conceived as fixed and it is time that changes. Bodies that change positions in space move in time. Since time has no spatial direction, space is said to be ‘parameterized’ by time.

However, the theory of relativity gives time a direction. The notions of space and time become unified into a four dimensional ‘space-time’. This exacerbates the problem of answering the question: “What is time?”. Time has a standard measure: the ‘second’ [sec]. One second is the time occupied by 9,192,631,770 vibrations of the light emitted by a cesium atom. Two synchronized cesium clocks would run for about 6000 years before their readings would differ by 1 [sec] [9].

1.1.3 Mass

The final notion ‘mass’ is the hardest to understand metaphysically. Mass is usually measured in terms of a quantity of matter, denoted ‘[M]’. The notion of ‘mass’ requires a definition of matter, which is problematic. Measuring mass raises no great difficulty in Newtonian mechanics, since, in that system, anything which carries weight is affected by gravity, and hence, is considered ‘matter’. But, in modern physics, particularly in the theory of relativity, light is affected by gravity, but has zero mass. If something has mass, it has energy. But the converse is not true. Light carries energy, but has no mass. Moreover, how particles, under modern concepts, gain mass remains a topic of discussion. In fact, in quantum physics, mass is not described as a quantity of matter, but as a certain vibration that arises in a field. Such a description can be modelled mathematically, but for a truly theoretical explanation for mass, modern physics gives an unsatisfactory answer.

1.2 Dimensions

Any direct observation of a physical nature ordinarily results in a number expressing the magnitude of the thing measured. Virtually all concepts in physics are expressed in terms of the fundamental notions of ‘length’, ‘time’ and ‘mass’ [L, T, M]. For example, velocity is defined as length divided by time i.e. [LT\(^{-1}\)]. The magnitude of a measurement depends on the scale employed. Length, for example, can be expressed in units of ‘centimeters’, ‘miles’ or ‘light years’ etc. When the magnitude depends on the units chosen, it is said to have a ‘dimension’ [224].

1.2.1 Systems of Units

Dimensions are ratios, and therefore, any unit of measure can be converted into any other unit of measure within the same dimension. For example,

\[ 1 \text{ [inch]} = 2.54 \text{ [centimeters]} \]
Both are measures of length. Dividing through by 2.54 gives 1 [centimeter] = 1/2.54 [inch]. The system of units employed largely depends on the scale of the physics involved. At small scales, commonly, the centimeter-gram-second [cgs] system of units is chosen. In other cases, the meter-kilogram-second [mks] system of units is employed. But, whatever system of units is chosen, measurements must be consistently made in those units to avoid errors. Throughout this study, the ‘[cgs]’ system of units will be employed whenever convenient.

1.2.2 Dimensional Analysis

The fact that equations must balance dimensionally provides a powerful tool. If a certain conceptual relationship exists between the variables and the constants of a problem, dimensional equality demands that the variables combine so that the exponents on each side of an equation are identical. In many problems, the functional form of the equation can be completely determined, except for numerical factors that have no dimensions. For example, the characteristics of a simple pendulum are its length (l), the mass (m) of the bob and the period (t) of the swing. The period ‘t’ depends on the acceleration ‘g’ due to gravity; this, for practical purposes, can be considered constant:

\[ g = G \frac{M_E}{R_E^2} [LT^{-2}] \]

where \( M_E \) and \( R_E \) are the mass and the radius of the Earth respectfully. Hence,

\[ t \propto g^\alpha m^\beta l^\gamma, \]

where \( \alpha, \beta, \gamma \) are the numerical exponents to be determined. Since \( g [LT^{-2}] \) has no mass dimension, it contributes nothing to \( m \), contributes linearly to \( l \) and quadratically to \( T \). Therefore,

\[ t = [M^\beta L^\gamma + m^\alpha T^{-2\alpha}] \]

Since \( t \) has no mass dimension, \( \beta = 0 \); and no length dimension, hence, \( \gamma + \alpha = 0 \). Therefore, \( -2\alpha = 1 \). These linear equations in three unknowns have the unique solution: \( \beta = 0, \alpha = -1/2, \gamma = 1/2 \). Hence,

\[ t \propto (l/g)^{1/2} \rightarrow t = 2\pi \sqrt{l/g} \]

The numerical factor ‘2π’ is provided as a reference. It cannot be determined by the procedure just illustrated, but only by experimentation. The technique demonstrated above is called ‘dimensional analysis’. It finds immediate application in many problems in physics [224].
1.3 Constants of Nature

Physical constants enter naturally into the equations of physics. For example, Newton's law of gravitational attraction:

\[ F = -G \frac{m_1 m_2}{r^2} \rightarrow -\frac{r^2}{m_1 m_2} F = G, \]

where \( G \) is a constant, \( r \) represents a distance, and the \( m_i \)'s are masses. Some physical constants are ratios of proportionality that correlate the relationship between various physical concepts. The constant '\( G \)' is one of these. Many such constants exist. An extensive list can be found in [225]. Other constants are associated directly with a dimension, the velocity of light, for example, designated '\( c \ [L/T] \)' is a constant of Nature.

Constants sometimes appear singly or in combinations. The equations are intended to quantitatively represent the laws of Nature; the values of the constants are found by experiment. In classical physics each successive identical experiment is supposed to give identical results (within experimental error). Hence, these constants are generally referred to as 'constants of Nature' [224]. In this study, the primary interest will be in three fundamental constants of Nature:

- **gravitational (\( G \)) = 6.67 \times 10^{-8} \ [gm^{-1} \times cm^3 \times sec^{-2}] [M^{-1}L^3T^{-2}]**
- **velocity of light (\( c \)) = 2.99793 \times 10^{10} \ [cm \times sec] \ [LT^{-1}]**
- **Planck's constant (\( h \)) = 6.62517 \times 10^{-27} \ [gm \times cm^2 \times sec^{-1}] [ML^2T^{-1}]**

Constants occur in conceptual combinations such as 'velocity', 'energy' etc. along with the variables in equations. An equation is a statement of the functional relationship between the concepts. The two sides of a meaningful physics equation must agree dimensionally as well as numerically [224]. Take, for instance,

\[ F = -G \frac{m_1 m_2}{r^2} \]

The dimensions of force (\( F \)) are [\( ML^2T^{-2} \)]. The dimensions of \( m_1 m_2/r^2 \) are [\( M^2L^{-2} \)]. Note that [\( MLT^{-2} \) \( \neq \) [\( M^2L^{-2} \)]. In order that the equation provides a meaningful physical statement, \( G \) must be assigned dimensions '[\( M^{-1}L^3T^{-2} \)]' i.e.

\[ F \ [MLT^{-2}] = G \ [M^{-1}L^3T^{-2}] \frac{m_1 m_2}{r^2} [M^2L^{-2}] = G \frac{m_1 m_2}{r^2} [M^{-1}L^3T^{-2}] [M^2L^{-2}] = G \frac{m_1 m_2}{r^2} [MLT^{-2}] \]
1.3.1 Dimensionless Constants

In physics there are dimensionless combinations of physical constants, for instance, between ‘m, ε, c’, where m is the mass of an electron, ε, electric charge and c, the speed of light. Suppose

\[ A = m^\alpha \varepsilon^\beta c^\gamma \rightarrow [M^{\alpha+\beta/2+\beta\gamma}T^{-\beta-\gamma} = 1], \]

where A is a constant. However, in this case, the linear equations involving α, β, γ are inconsistent. There is no solution. But, if Plank’s constant ‘ℏ’ is included, then

\[ A = m^\alpha \varepsilon^\beta c^\gamma \hbar^\delta \rightarrow [M^{\alpha+\beta/2+\beta\gamma+2\delta}T^{-\beta-\gamma-\delta} = 1] \]

The three equations in four knowns can be solved in terms of δ. The result is α = 0, β = −2δ, γ = δ. The constant becomes (ℏc/ε²)δ. For the simplest form, let δ = 1. This constant appears in the theory of atomic spectra related to the fine structure of spectral lines and is usually written ‘ℏc/2πε²’, which includes the dimensionless numerical factor ‘1/2π’. The numerical value of this constant is ≈ 137.29 [224].

1.4 Motion and Rest

Motion is a change in spatial position with respect to time. It is typically described in terms of displacement - a velocity or an acceleration of a point measured in terms of the change in the position of the point relative to the stationary points in the frame of reference.

Geometrically, motion is described by a space curve ‘s(t)’, which is generally a function of time. As time ticks away, s(t) draws out a space curve (see fig. 1.4-1) against a fixed background of points or ‘frame of reference’. Motion, described in this manner, is normally associated with the motion of a particle – one of the two major modes of representing physical phenomena; the other mode is the ‘wave’. Classically, the motion of a body is simple, since classical physics assumes that the entire mass of a body can be concentrated at a point, called its ‘center of gravity’ or ‘center of mass’. Hence, the motion of a body is synonymous with the motion of a point particle in space.

![Figure 1.4-1](image-url)
1.4.1 Constant Motion in a Straight Line

The average speed ‘\(\bar{v}\)’ of a point drawn out by a space curve is defined as

\[
\bar{v} = \frac{s(t)}{t} \rightarrow s(t) = \bar{v}t
\]

If \(v\) is a constant, then \(v = \bar{v}\), so that \(s(t) = vt\), where \(v\) is the slope of the line. Note that the equation ‘\(s(t) = vt\)’ is a valid physical statement, since

\[
s(t)[L] = v[L^{-1}T]t[T] = vt[L^{-1}T] = vt[L],
\]

where the dimensions on both sides of the equation are equal.

Constant rectilinear motion is described geometrically by a straight line. If \(t = 0\), then \(s(0) = 0\) represents the ‘present’. If \(\forall t(v = 0 \rightarrow s(t) = 0)\), time continues, but no distance is covered. The point is ‘at rest’, which represents no spatial motion of the point relative to the frame of reference. If \(v = \infty\), then \(t \rightarrow 0\), no time ticks by. If a signal travels at an infinite speed toward two observers, who reside at different spatial locations, the observers will claim that the signal reached them simultaneously. This, of course, is an assumption and will be found fallacious. But, in non-relativistic physics, this is the assumption made.

If \(v < 0\), then \(v\) is given a direction as well as a magnitude; in this case, the variable ‘\(v\)’ becomes a vector quantity and \(s(t)| = t|v|\) becomes a vector equation. For example, if \(s_1(t) = vt\) and \(s_2(t) = -vt\), then the two points ‘\(s_1, s_2\)’ move in opposite directions, but at the same speed. If \(s = s_1 + s_2\), then \(s = vt - vt = 0\). If the two points are regarded as one system, the system is said to be in ‘equilibrium’. In modern physics the word ‘equilibrium’ is often replaced by the word ‘symmetry’.

1.4.2 Instantaneous Velocity

If \(v\) is not a constant, the equation ‘\(s(t) = v(t)t\)’ is not a straight line, but the average velocity ‘\(\bar{v}\)’ can be defined as

\[
\bar{v} = \frac{s_2 - s_1}{t_2 - t_1} = \frac{\Delta s}{\Delta t}
\]

However, computing the average ‘\(\bar{v}\)’ has a disadvantage. It does not provide a velocity at each instant of time, but only the mean velocity over an interval of time. To remedy this, the calculus is employed:

\[
|v| = \lim_{\Delta t \to 0} \frac{\Delta |s|}{\Delta t} = \frac{d|s|}{dt}
\]
The expression $|v|^¢$ is called the ‘instantaneous velocity’. The space curve describing this motion must be sufficiently smooth so that the derivative $d|s>/dt$ exists at each point along the curve.

Example: $|s(t)| = |7.8 + 9.2t - 2.1t^3| \rightarrow |v| = d|s>/dt = |9.2 - 6.3t^2|$. Hence, $t = 3.5$ [sec] $\rightarrow |v| = |9.2 - 6.3(3.5)^2| = -|68| [cm/sec]$. The speed then becomes $||v|| = 68$ [cm/sec] at $t = 3.5$ [sec].

1.4.3 Acceleration

If the velocity is not constant in a time interval, then the particle ‘accelerates’ in that interval. This can take one of two forms: 1) the particle moves in a straight line, but its speed changes during the time interval, 2) the magnitude of the velocity remains constant, but the direction of the particle changes. In any case, the average acceleration $\overline{a}$ in a time interval is

$$\overline{a} = \frac{v_2 - v_1}{t_2 - t_1} = \frac{\Delta v}{\Delta t} \rightarrow |a| = \lim_{\Delta t \rightarrow 0} \frac{\Delta |v|}{\Delta t} = \frac{d|v|}{dt},$$

assuming the derivatives exists at all points in the interval. The variable ‘$|a|$’ is called the ‘instantaneous acceleration’.

Example: $|s(t)| = |7.8 + 9.2t - 2.1t^3| \rightarrow |v| = d|s>/dt = |9.2 - 6.3t^2| \rightarrow |a| = d^2|s>/dt^2 = d|v>/dt = -|12.6t|$. Hence, $t = 3.5$ [sec] $\rightarrow |a| = -|12.6(3.5)| = -|44.1| [cm/sec^2]$.

If the velocity is constant, then $|v| = d|s>/dt = |c|, \ c = a \ constant \rightarrow d|v>/dt = |0|$. As expected, a particle moving with a constant velocity does not accelerate.

1.4.3.1 Constant Acceleration

Let

$$d|v>/dt = -|a| \rightarrow |v| = -\int |a| \ dt = |v_0| - t|a| \rightarrow |s| = \int (|v_0| - t|a|) \ dt$$

$$= t|v_0| - \frac{1}{2}t^2|a| + |s_0|, \ \ a = \ a \ constant$$

These are the equations of motion of a particle moving with a ‘constant acceleration’. Such equations are very useful, since, near the Earth’s surface, the acceleration due to gravity can, for practical purposes, be considered constant:

$$|a| = -|g| = -G\frac{M}{R^2} |R_E| [LT^{-2}],$$

where $|R_E|$ is a unit vector in the direction of the radius between the center of the Earth and an object on the surface and $g$ is a constant. The minus sign indicates that the

---

1 In most treatments of the subject the bracket ‘|⟩’ is omitted.
acceleration is directed toward the center of Earth rather than away from it. The equation is not entirely correct, since the Earth is not a perfect sphere, but slightly flattened at the poles. In this case, $\langle g \rangle$ is slightly greater at the poles than at the equator. But the difference is so slight, it can, for most problems, be ignored. The equations of motion in a constant gravitational field describe the flight of baseballs, artillery shells and many other phenomena with a relatively high degree of accuracy so long as other forces ancillary to gravity are ignored. When $\langle a \rangle = -\langle g \rangle$, this is often referred to as the "flat Earth" model of the gravitational field.

Example: Find the horizontal distance ($s_x$) a shell travels if fired from a gun with muzzle velocity 'v' at an angle 'θ'. If $t$ is the time it takes to hit the ground, then $s_x = v_x t \rightarrow t = s_x / v_x$, where $v_x$ is the velocity in the horizontal direction and $s_x$ is the horizontal distance the shell travels. When the shell strikes the ground, the vertical distance is zero. Hence,

$$0 = v_y t - \frac{1}{2} t^2 g \rightarrow t = \frac{2v_y}{g}$$

So,

$$\frac{2v_y}{g} = \frac{s_x}{v_x} \rightarrow s_x = \frac{2v_y v_x}{g} = \frac{2(v \cos \theta)(v \sin \theta)}{g} = \frac{v^2 \sin 2\theta}{g},$$

since $2 \sin \theta \cos \theta = \sin 2\theta$.

**1.5 Angular Velocity and Angular Acceleration**

Any point a constant distance 'r' from a fixed axis draws out a circle as it moves around the axis with constant speed, described mathematically by the ratio

$$\frac{\theta}{2\pi} = \frac{s}{C}$$

where $\theta$ is the angle that subtends 's', the distance traveled along the circle; the angle '2π' subtends 'C', the circumference of the circle. Since $C = 2\pi r$,

$$\frac{\theta}{2\pi} = \frac{s}{2\pi r} \rightarrow s = \theta r$$

Since $s = vt \rightarrow v = s/t$, then

$$v = \frac{s}{t} = \theta r$$

Hence, $v$ is proportional to the arc subtended by the angle 'θ'. It is customary to replace $\theta / t$ by 'ω' and simply write 'v = ωr', where ω is called the 'angular speed'. This can be a bit confusing, since speed, generally, has dimensions $[LT^{-1}]$, but ω is measured in
The tangential velocity of points on a revolving disc depends on their distance from the axis of rotation, but the angular velocity of all points on the disc is identical.

### 1.5.1 Instantaneous Angular Velocity

The average angular speed \( \bar{\omega} \) is defined as

\[
\bar{\omega} = \frac{\theta_2 - \theta_1}{t_2 - t_1} = \frac{\Delta\theta}{\Delta t} \rightarrow \omega = \lim_{\Delta t \to 0} \frac{\Delta\theta}{\Delta t} = \frac{d\theta}{dt},
\]

so long as the limit exists, where \( \omega \) is called ‘instantaneous angular velocity’.

### 1.5.2 Instantaneous Angular Acceleration

Similarly, the average angular acceleration \( \bar{\alpha} \) is defined as

\[
\bar{\alpha} = \frac{\omega_2 - \omega_1}{t_2 - t_1} = \frac{\Delta\omega}{\Delta t} \rightarrow \alpha = \lim_{\Delta t \to 0} \frac{\Delta\omega}{\Delta t} = \frac{d\omega}{dt},
\]

so long as the limit exists, where \( \alpha \) is called ‘instantaneous angular acceleration’ and has dimensions \([rad/T^2]\). Angular motion is useful in describing the motion of rotating bodies, like a merry-go-round, a rotating wheel or a spinning ice skater.

The variable \( \omega \) is actually a vector. Normally, but not always, vectors point in the direction of motion, but angular motion is an exception. Rotating points can move in only two directions around a fixed axis: ‘clockwise’ or ‘counterclockwise’. If the rotation is counterclockwise, \(|\omega\rangle \) points in the positive vertical direction along the axis of rotation. If the rotation is clockwise, \(|\omega\rangle \) points in the negative direction. Evidently, a fixed axis can always be oriented in the vertical direction.

### 1.5.3 Constant Angular Motion

Recall that purely translational motion of a particle \((v = \text{constant})\) is rectilinear, the acceleration being zero. But if \(\omega = \text{constant}\), the motion is equivalent to constant accelerated motion, since the direction of the motion is changing at every moment.
To see this, suppose a point is moving along a circle of radius ‘r’ from a point ‘p’ to a point ‘q’ (see fig. 1.5.3-1). The tangential velocity ‘v’ at ‘p’ can be divided into horizontal (\(v_{px}\)) and vertical (\(v_{py}\)) components. Since |\(v\)| is a constant, |\(|v|\)| at p and q is the same, but |\(v\)| points in a different direction at q than it does at p. Now, \(v_{px} = v \cos \theta\), \(v_{py} = v \sin \theta\), \(v_{qx} = v \cos \theta\), \(v_{qy} = -v \sin \theta\). The distance \(pq\) along the circle is

\[
\overline{pq} = v(t_2 - t_1) = v \Delta t \rightarrow \Delta t = \frac{\overline{pq}}{v} = \frac{2\theta r}{v}
\]

Therefore, the average acceleration ‘\(\overline{a}_x\)’ in the horizontal direction is

\[
\overline{a}_x = \frac{v_{qx} - v_{px}}{\Delta t} = \frac{v \cos \theta - v \cos \theta}{\Delta t} = 0
\]

But the average acceleration ‘\(\overline{a}_y\)’ in the vertical direction is

\[
\overline{a}_y = \frac{v_{qy} - v_{py}}{\Delta t} = \frac{-v \sin \theta - v \sin \theta}{\Delta t} = -\frac{2v \sin \theta}{\Delta t} = -\left(\frac{v^2}{\theta}ight)\left(\frac{\sin \theta}{\theta}\right)
\]

If p and q are brought closer and closer together, and ultimately, coincide, \(\theta \rightarrow 0\). But, \(\lim_{\theta \rightarrow 0} \sin \theta/\theta = 1\). This can be verified by L’Hospital’s Rule. Since \(\sin 0/0 = 0/0\), then \(d(\sin \theta)/d\theta = \cos \theta\) and \(d\theta/d\theta = 1\), so \(\lim_{\theta \rightarrow 0} (\cos \theta)/1 = (\cos 0)/1 = 1/1 = 1\). Hence,

\[
|\overline{a}_y| = -\frac{v^2}{r}
\]

Since v and r are constant, |\(\overline{a}_y\)| is constant. Seen as a vector, |\(\overline{a}\)| always points in the direction of the axis of rotation, where ‘\(\overline{a}\)’ is called the ‘instantaneous centripetal acceleration’.

1.6 The Relationship between Angular Speed and Linear Speed

The linear speeds of points on a rotating disc depend on their distance from the axis of rotation, but those points have the same angular speed. What is the relationship between linear and angular speed? Let \(s = \theta r\), \(r = a\ constant\). Differentiating both sides of this equation with respect to time leaves

\[
\frac{ds}{dt} = r \frac{d\theta}{dt}
\]

Since \(ds/dt = v\) and \(d\theta/dt = \omega\), then \(v = \omega r\). Since \(dv/dt = r \ d\omega/dt\), then the tangential acceleration ‘\(a_t\)’ is

\[
a_t = \frac{d\omega}{dt} r = \alpha r, \quad \alpha = \frac{d\omega}{dt}
\]

and the radial acceleration ‘\(a_r\)’ is
\[ a_r = \left( \frac{v^2}{r} \right) = \left( \frac{v^2}{r} \right) \left( \frac{r}{r^2} \right) = \frac{v^2}{r^2} r = \omega^2 r \]

Therefore, the acceleration of a point on a rotating disc has two components - the radial component \(a_r\), which is always present so long as the disc is rotating. But if \( \omega = \text{a constant} \), \( a_t = \frac{dv}{dt} = r \frac{d\omega}{dt} = 0 \) and there is no tangential acceleration.

### 1.7 Harmonic Motion

The simplest periodic process, the vibration, is described mathematically by the sine or cosine function:

\[ x(t) = A \cos 2\pi \nu t, \quad x(t) = A \sin 2\pi \nu t \]

The function \( x(t) \) varies between \(+A\) and \(-A\) because the sine and cosine functions vary between \(+1\) and \(-1\). Physically, a point oscillates back and forth along an axis between \(+A\) and \(-A\) at regular intervals around an equilibrium point \( x(t) = 0 \) (see fig. 1.7-1). If \( \forall t (t \in R \rightarrow x(t) = 0) \), the point is in ‘equilibrium’ and does not oscillate.

![Figure 1.7-1](image)

The sine and cosine functions are periodic in time. A periodic function \( f(t) \) is defined \( f(t) = f(t + T) \), where \( T \) is called the ‘period’.

That the sine function satisfies the periodic condition can be seen from

\[ x(t + T) = A \sin (2\pi \nu (t + T)) = A \sin \left( \frac{2\pi}{T} (t + T) \right) = A \sin \left( \frac{2\pi t}{T} + 2\pi \right) = A \sin \frac{2\pi t}{T} = x(t), \]

where \( \nu = \frac{1}{T} \) \([sec^{-1}]\) and is called the ‘frequency’ of the vibration. If \( T >> 1 \), then the oscillations are very slow. If \( T << 1 \), then the oscillations are rapid. Often the equations are simplified by letting \( \omega = \frac{2\pi}{T} = 2\pi \nu \), hence,

\[ x(t) = A \sin \omega t, \]

where ‘\( \omega [rad/sec]^2 \)’ is called the ‘angular frequency’ of the vibration.

---

\(^2\) Unfortunately, the same symbol ‘\( \omega \)’ is used for angular speed in circular motion. For uniform circular motion the angular speed is equal to the angular frequency but for non-uniform
1.7.1 Modelling Harmonic Motion using Complex Functions

Simple harmonic motion can be modelled using complex functions, avoiding the cumbersome addition theorems associated with trigonometric functions. By Euler’s formula:

\[ e^{i\omega t} = \cos \omega t + i \sin \omega t \rightarrow Ae^{i\omega t} = A \cos \omega t + iA \sin \omega t = z \]

The function \( z = Ae^{i\omega t} \) describes a point moving in a complex circle of radius \( A \) with angular speed \( \omega \). The projections of this motion onto the real or imaginary axis of the complex plane constitute simple harmonic motion i.e. \( Re(e^{i\omega t}) = A \cos \omega t \) and \( Im(e^{i\omega t}) = A \sin \omega t \). It is, therefore, possible to take either the real or the imaginary part of \( z \) as the actual physical statement [226].

The function \( z = Ae^{i\omega t} \) also provides a simple construction for taking derivatives:

\[
\frac{d}{dt} z = i\omega e^{i\omega t} = i\omega z \rightarrow \frac{d^2}{dt^2} z = -\omega^2 e^{i\omega t} = -\omega^2 z
\]

Simple harmonic motion satisfies the differential equation

\[
\frac{d^2 z}{dt^2} = -\omega^2 z
\]

To see this, let \( z = \cos \omega t \), then

\[
\frac{dz}{dt} = -\omega \sin \omega t \rightarrow \frac{d^2 z}{dt^2} = -\omega^2 \cos \omega t = -\omega^2 z
\]

If \( w = \sin \omega t \),

\[
\frac{dw}{dt} = \omega \cos \omega t \rightarrow \frac{d^2 w}{dt^2} = -\omega^2 \sin \omega t = -\omega^2 w
\]

Hence, \( w \) is a solution. Moreover, \( z + w \) is also a solution, since

\[
\frac{d^2 z}{dt^2} + \frac{d^2 w}{dt^2} = -\omega^2 z - \omega^2 w = -\omega^2 (z + w)
\]

The most general solution is \( z = A \cos \omega t + B \sin \omega t \), where \( A \) and \( B \) are constants. For a more general discussion of periodic functions, see Book II: Chapter 16.

motion the angular speed is not constant. The angular frequency for simple harmonic motion is a constant by definition.
1.7.2 Simple Harmonic Motion and Phase

Suppose there are two simple harmonic oscillators \( x_1(t), x_2(t) \) described by the equations

\[
x_1(t) = A_1 \cos \omega t, \quad x_2(t) = A_2 \cos(\omega t + \delta)
\]

The two oscillators are 'out of phase' by an amount '\( \delta \)', the max and min values occurring at different times. Since the two oscillations lie along the same axis, combining them into one equation will result in simple harmonic motion. To see this,

\[
x_1(t) + x_2(t) = A_1 \cos \omega t + A_2 \cos(\omega t + \delta) = A_1 \cos \omega t + A_2 \cos \omega t \cos \delta - A_2 \sin \delta \sin \omega t
\]

The above equation follows from a trigonometric identity regarding the cosine sum of two angles. Algebraic manipulation (see [226] for details) leaves

\[
x(t) = x_1(t) + x_2(t) = \sqrt{A_1^2 + A_2^2 + 2A_1^2 A_2^2 \cos \delta \cos(\omega t + \phi)} = A \cos(\omega t + \phi),
\]

where \( A = \sqrt{A_1^2 + A_2^2 + 2A_1^2 A_2^2 \cos \delta} \).

The function \( x(t) \) is a simple harmonic oscillator with phase \( \phi \).

1.7.2.1 Simple Harmonic Motion along Different Axes

Suppose

\[
x(t) = a \cos \omega t, \quad y(t) = b \cos(\omega t - \delta)
\]

These two harmonic oscillators have the same frequency, but oscillate along different perpendicular axes, out of phase by an amount '\( \delta \)'. What geometrical figures are represented by different values of \( \delta \)? If \( \delta = 0 \), then

\[
\frac{y}{x} = \frac{b}{a},
\]

which is the diagonal of a rectangle with sides \('a'\) and \('b'\). The motion is called the 'linear form' of vibration. If \( 0 < \delta < \pi/2 \), then the elimination of \( t \) between \( x(t) \) and \( y(t) \) gives

\[
y = b \frac{x}{a} \cos \delta + b \sqrt{1 - \frac{x^2}{a^2} \sin \delta} \to y - \frac{bx}{a} \cos \delta = b \sqrt{1 - \frac{x^2}{a^2} \sin \delta} \to \left(y - \frac{bx}{a} \cos \delta \right)^2
\]

\[
= b^2 \left(1 - \frac{x^2}{a^2}\right) \sin^2 \delta
\]

The curve must be an ellipse. Its principle axes are determined by the methods of analytic geometry. For instance, if \( \delta = \pi/2 \), then

\[
x(t) = a \cos \omega t, \quad y(t) = b \sin \omega t
\]
Therefore,

\[ x^2(t) = a^2 \cos^2 \omega t, \quad y^2(t) = b^2 \sin^2 \omega t \rightarrow \frac{x^2}{a^2} + \frac{y^2}{b^2} = \cos^2 \omega t + \sin^2 \omega t = 1 \]

But

\[ \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \]

is the equation of an ellipse, where the principle axes are the coordinate axes. If \( \delta = \pi \), then

\[ x(t) = a \cos \omega t, \quad y(t) = -b \cos \omega t \]

and

\[ \frac{y}{x} = -\frac{b}{a} \]

This is the original linear vibration reflected in the \( y \)-axis. Setting \( \delta = 3\pi/2 \) results in the same ellipse as \( \delta = \pi/2 \), but the curve is traversed in the opposite direction. In the case where the frequency difference \( \varepsilon = \omega_2 - \omega_1 \) is small so that

\[ x(t) = a \cos \omega_1 t, \quad y(t) = b \cos[\omega_1 t + (\omega_2 - \omega_1) t] = b \cos(\omega_1 t + \varepsilon t), \]

then the phase difference becomes a linear function of time and the point traverses in smooth succession all the elliptical curves corresponding to the various values of \( \delta \) [226].

### 1.7.3 Wave Motion

Simple harmonic motion can be described by the two parameter function

\[ y(t, x) = Ae^{i\omega t} \]

Since \( y \) is a function only of \( t \), \( x \) does not change and is arbitrary. But remembering that \( x = vt \rightarrow t = x/v \rightarrow t - x/v = 0 \), consider the function, where at time \( 't = 0' \), the same condition exists as at time \( 't - x/v' \), then

\[ y(t, x) = Ae^{i\omega(t-x/v)} \]

and \( y \) becomes a function of both \( t \) and \( x \). If the same state of vibration exists at \( x \neq 0 \) as at \( x = 0 \), but with a difference of phase corresponding to the finite velocity of the propagation \( 'v' \) of the phase, then the motion is called a 'wave'. Waves not only oscillate in time, but move through space.

The instantaneous value \( 't = 0' \) recurs if

\[ x = \frac{2\pi v}{\omega} = \frac{v}{\nu}, \]

33
since, in this case, \( \omega x / v = 2\pi \), is the period of the wave, \( y(t, x) \) remains unchanged. The distance, which separates the points of equal phase, is called the ‘wave-length’, designated ‘\( \lambda \)’. Letting \( x = \lambda \) in the equation above, then \( \lambda v = v \). Hence, the ‘wave length \times frequency’ is equal to the velocity of the propagation of the phase, where the direction of propagation is in the \( x \)-direction. Such a wave is called a ‘plane wave’.

1.7.3.1 The Wave Equation

If \( y(t, x) = Ae^{i\omega(t-x/v)} \) is differentiated twice with respect to \( x \) and \( t \), then

\[
\frac{\partial^2 y}{dt^2} = -\omega^2 Ae^{i\omega(t-x/v)}, \quad \frac{\partial^2 y}{dx^2} = -\frac{\omega^2}{v^2} Ae^{i\omega(t-x/v)}
\]

It follows immediately that

\[
\frac{\partial^2 y}{dx^2} = \frac{1}{v^2} \frac{\partial^2 y}{dt^2}
\]

The equation above is called the ‘wave equation’. It plays a very important role in both quantum and classical physics. The arbitrary function ‘\( f(t-x/v) \)’ satisfies the wave equation, but so does the function ‘\( g(t+x/v) \)’. The latter corresponds to a wave moving in the negative \( x \)-direction. The general solution is

\[
y = f(t-x/v) + g(t+x/v)
\]

Furthermore, if

\[
u(t, x) = Ae^{i\omega(t-(x \cos \alpha + y \cos \beta + z \cos \gamma)/v)},\]

where, evidently, \( \alpha, \beta, \gamma \) are direction cosines, then differentiating \( u \) twice with respect to each coordinate and twice with respect to \( t \) gives

\[
\frac{\partial^2 u}{dx^2} + \frac{\partial^2 u}{dy^2} + \frac{\partial^2 u}{dz^2} = \frac{1}{v^2} \frac{\partial^2 u}{dt^2}
\]

This is the three dimensional wave equation [226].

1.7.3.2 Wave Dispersion and Group Velocity

If the partial vibrations composing a complicated vibrational phenomenon are propagated with different velocities, the superposition of all the vibrations creates a non-periodic condition of which little can be said. But if \( \omega \) changes only slightly, the wave is said to ‘disperse’. Let

\[
\omega_1 = \omega + \Delta \omega, \quad v_1 = v + \frac{dv}{d\omega} \Delta \omega
\]
There are two waves:

\[ y = \cos \omega_1 \left( t - \frac{x}{v_1} \right), \quad z = \cos \omega \left( t - \frac{x}{v} \right) \]

\[ u(x, t) = y + z = \cos \omega_1 \left( t - \frac{x}{v_1} \right) + \cos \omega \left( t - \frac{x}{v} \right) \]

\[ = 2 \cos \left\{ \frac{\omega_1}{2} \left( t - \frac{x}{v_1} \right) + \frac{\omega}{2} \left( t - \frac{x}{v} \right) \right\} \cos \left\{ \frac{\omega_1}{2} \left( t - \frac{x}{v_1} \right) - \frac{\omega}{2} \left( t - \frac{x}{v} \right) \right\} \]

Since \( \Delta \omega \) is small, setting \( \omega_1 = \omega \) and \( v_1 = v \) will not result in an appreciable error in the first sum enclosed in ‘\{\}', so

\[ u(x, t) = 2 \cos \omega \left( t - \frac{x}{v} \right) \cos \left\{ \frac{\Delta \omega}{2} t - \frac{1}{2} \left( \frac{\omega_1}{v_1} - \frac{\omega}{v} \right) x \right\} \]

Neglecting higher powers gives

\[ \frac{\omega_1}{v_1} - \frac{\omega}{v} = \Delta \omega = \frac{1}{v} \left( \frac{\omega}{v^2} \frac{dv}{d\omega} \right) \Delta \omega \]

Hence,

\[ u(x, t) = 2 \cos \frac{\Delta \omega}{2} \left( t - \frac{x}{V} \right) \cos \omega \left( t - \frac{x}{v} \right), \]

where

\[ \frac{1}{v} - \frac{\omega}{v^2} \frac{dv}{d\omega} = \frac{d}{dv} \left( \frac{\omega}{v} \right) = \frac{1}{v} \]

A simple wave with constant amplitude ‘\( A \)' is written

\[ u(x, t) = A \cos \omega \left( t - \frac{x}{v} \right) \]

However, in this case, the amplitude of wave is subject to a wave-like vibration given by

\[ A = 2 \cos \frac{\Delta \omega}{2} \left( t - \frac{x}{V} \right) \]

While the phase of the wave is propagated with the velocity ‘\( v' \)', the velocity of the dispersion of the wave is propagated with velocity ‘\( V' \' and is called the ‘group velocity’. Note that since \( v/v = \lambda \),

\[ V = \frac{dv}{d \left( \frac{1}{\lambda} \right)} = -\lambda^2 \frac{dv}{d\lambda} \]
Since \( v = \nu \lambda \),

\[
\lambda \frac{d\nu}{d\lambda} + \nu = \frac{d\nu}{d\lambda}
\]

by the chain rule. Hence,

\[
V = -\lambda^2 \frac{d\nu}{d\lambda} = -\lambda \nu \frac{d\nu}{d\lambda} = -\lambda \left( -\nu + \frac{d\nu}{d\lambda} \right) = \lambda \nu - \lambda \frac{d\nu}{d\lambda} = v - \lambda \frac{d\nu}{d\lambda}
\]

Thus, the group velocity is a function of the phase velocity and the wave length [226].

1.7.3.3 Standing Waves

Finally, let

\[
u_1(x, t) = Ae^{i\omega(t-x/v)}, \quad \nu_2(x, t) = Ae^{i\omega(t+x/v)+i\delta}
\]

The wave \( \nu_2 \) has the same amplitude and frequency as \( \nu_1 \), but travels in the opposite direction. Now

\[
u_1 + \nu_2 = Ae^{i\omega(t-x/v)} + Ae^{i\omega(t+x/v)+i\delta} = Ae^{i(\omega t+\delta/2)} \left\{ e^{-i(\omega x/v+\delta/2)} + e^{i(\omega x/v+\delta/2)} \right\}
\]

Note that

\[
Ae^{i(\omega t/v+\delta/2)}
\]

represents a vibration having the same phase all through space. The factor \( \cos(\omega x/v + \delta/2) \) shows that the amplitude is a periodic function of \( x \). The period is \( v/v = \lambda \). The chief characteristic of this wave is that the finite velocity of propagation of the phase is entirely absent. Therefore, the wave is called a ‘standing wave’. If the amplitudes are unequal, there exists, in addition to the standing wave, a wave traveling in the direction of the stronger partial wave, whose amplitude is the difference of the amplitudes of the component waves [226].

1.8 Concluding Remarks

Basic physics emulates from three fundamental notions: ‘length’, ‘time’ and ‘matter’. Virtually all concepts in physics are described, in one form or another, in terms of these notions. Notions in physics are associated with measurements, expressed in a certain system of units. The magnitude of a measurement depends on the scale employed. Whenever a measurement depends on the units chosen, it is said to have a ‘dimension’. Concepts in physics are communicated by equations, where the expressions on both sides of the equation must be equal numerically and dimensionally.

The fundamental notions are used to define the two most important concepts in physics: the ‘particle’ and the ‘wave’. Both particles and waves are described in terms of motion and rest. Since a body has a center of mass, its motions are synonymous with particle motions, with a definite position in space, moving smoothly from position to position in a
given amount of time. The wave, on the other hand, has no center of mass, being
spread out over space, and therefore, occupies many spatial positions at once, does not
move smoothly through space, but jiggles and suffers dispersion, and hence, its form
can deteriorate of over time.

Much of the current research in physics focuses on finding a single theory that
accommodates both the particle and the wave concepts. The theory of relativity and
quantum mechanics make outstandingly accurate predictions in their domains of
applicability, but are logically incompatible. In quantum mechanics, energy ‘E’ is
described by the formula ‘E = hν’, where ν is a frequency and is normally associated
with wave behavior. In classical physics, the energy ‘E’ is described by the formula ‘E =
\frac{mv^2}{2} + V’ or \( E = mc^2 \) if \( v = 0 \), where the amount of mass determines, in part, the
amount of energy. But mass, described in this way, is primarily associated with a
particle. The mass of a particle does not emerge naturally at the quantum level. How
are these two fundamental concepts of energy reconciled? In quantum mechanics, ‘h’
is the constant that establishes a relationship between the energy and frequency of a
wave. In classical physics, c is the constant that establishes the relationship between
energy and mass. But \( h \) is on the order of \( \approx 10^{-27} \) and \( c^2 \) is on the order of \( \approx 10^{20} \).
Why is \( c \) so much greater than \( h \)? This dilemma of scale is referred to as the
‘hierarchical problem’.
Chapter 2
The Physics of a Single Rigid Body

“The law of conservation of energy tells us we can't get something for nothing, but we refuse to believe it.”

- Isaac Asimov

2.0 Introduction

Intuitively, mass is any manner of bulky substance conceived as a single entity with spatial boundaries. In classical physics, lumps of matter have an internal point called the 'center of gravity', where the entire mass of the lump can be conceived as located. This allows describing the motion of a mass, usually referred to as a 'body', by regarding its mass as concentrated at a single point. By employing the concepts of 'motion' and 'rest' developed in Chapter 1, nonrelativistic classical physics describes the motion of a body as a space curve drawn out by its center of gravity as it runs over a set of stationary points.

However, the idea of 'mass', just described, will not suffice for quantum physics. A discussion of the quantum mechanical concept of 'mass' will be deferred until a later chapter, where quantum mechanics is discussed in detail.

Virtually all of physics revolves around the concept of 'conservation'. Emmy Noether showed that, mathematically, any continuous symmetry is associated with a conservation law. Most investigations in physics involve identifying the symmetries of the problem. Any quantities associated with a symmetry are conserved.

This chapter discusses the physics associated with the motion of a single particle (body), including changes in the motion and in what sense the motion is conserved.

2.1 Inertia

The simplest motion of a body is called 'momentum' 'p' defined

\[ p = mv \ [gm \times cm/sec], \]

or alternatively,

\[ p = m \frac{ds}{dt}, \quad \frac{ds}{dt} = v = \text{a constant} \]

where \( m \) is the mass of the body and \( ds/dt = v \) is the velocity of the center of gravity of a body as measured against a stationary set of points. If a body with mass '\( m \)' is at rest or moving with a constant velocity, it is in an 'inertial state'. In this case, the mass is referred to as 'inertial' mass. If no external forces act on the body, it will continue in uniform motion in a straight line or remain at rest indefinitely. This rule is known as 'Newton’s First Law' of motion. Momentum can be written in vector form i.e. \( |p\rangle = m|v\rangle \).
2.1.1 Force (Newton’s Second Law of Motion)

If the velocity of the center of mass of a body changes over time, it is said to be influenced by a ‘force’ \( F \). The average force \( \bar{F} \), in an interval of time, is defined

\[
\bar{F} = m \frac{p_2 - p_1}{t_2 - t_1} = m \frac{\Delta p}{\Delta t}
\]

Note that a force is described by the change in the inertia of the body’s center of gravity per time interval.

Let

\[
m \frac{\Delta p}{\Delta t} \to \lim_{\Delta t \to 0} \frac{\Delta p}{\Delta t} = \frac{dp}{dt} \to F = m \frac{d^2s}{dt^2} = ma \ [gm \times cm/sec^2],
\]

where \( d^2s/dt^2 = a \) is the ‘acceleration’ of the body and \( \bar{F} \) is called the ‘instantaneous force’. Force can also be written in vector form:

\[
|F| = m|a|
\]

If \( a = a \text{ constant} \), a body will either move in a straight line with constantly accelerating motion or move in a circle with constant angular velocity. Mentioned previously, the acceleration ‘\( g \)’ near the Earth’s surface, for practical purposes, can be considered constant, where \( w = mg \) is called the ‘weight’ of a body.

2.1.2 Equilibrium (Newton’s Third Law of Motion)

Moreover, someone standing on the Earth’s surface, whose weight is \( mg \), is at rest, which implies that the Earth is pushing back with a force ‘\(-F_E\)’ of equal magnitude, but opposite direction i.e. \( mg = -F_E \to mg + F_E = 0 \). In general, a body, if the vector sum of all the external forces that act on the body is zero i.e. \( \sum_{i=1}^{n} |F_i| = |0| \), then the body is in ‘equilibrium’. If \( \sum_{i=1}^{n} |F_i| \neq |0| \), the body will suffer an acceleration in the direction of the vector sum of the forces. The body is then said to be affected by an ‘unbalanced’ force.

2.1.3 Inertial Mass

The equation ‘\( F = ma \)’ implies that if \( m \ll 1 \), the force required to accelerate the mass to ‘\( a \)’ is small. If \( m \gg 1 \), the force is large, suggesting that inertia is synonymous with the ability of a body to resist a change in its motion.

2.2 Work, Power and Energy

Suppose a force moves the center of mass of a body a distance ‘\( r \)’ in time ‘\( t \)’, then let

\[
\frac{W_2 - W_1}{t_2 - t_1} = \frac{\Delta W}{\Delta t} = F \left( \frac{r_2 - r_1}{t_2 - t_1} \right) = F \frac{\Delta r}{\Delta t} \to \lim_{\Delta t \to 0} \frac{\Delta W}{\Delta t} = \lim_{\Delta t \to 0} F \frac{\Delta r}{\Delta t} \to \frac{dW}{dt} = F \frac{dr}{dt},
\]

\[W_2 = F \cdot r_2, \quad W_1 = F \cdot r_1\]
The force $|F|$ is a vector, but $W$, called the ‘work’ done on a body, is a scalar i.e. $dW = \langle F|dr \rangle$, since only the component of $F$ in the direction $|dr\rangle$ contributes to the motion of the body [227]. Another way of looking at this is

$$dW = F \cos \theta \, dr = \|F\|\|dr\| \cos \theta = \langle F|dr \rangle$$

Hence, if a force moves a particle from position $|P_1\rangle$ to a position $|P_2\rangle$ along a space curve, then

$$W = \int_{P_1}^{P_2} \langle F|dr \rangle = \int_{|r_1\rangle}^{|r_2\rangle} \langle F|dr \rangle \left[ gm \times cm^2/sec^2 \right],$$

where $|r_2\rangle, |r_1\rangle$ are the position vectors of $P_2, P_1$ respectfully and ‘$W$’ is called the ‘total work’ done by the force.

2.2.1 Power

Moreover,

$$P = \frac{dW}{dt} = \langle F| \frac{dr}{dt} \rangle = \langle F|v \rangle \left[ gm \times cm^2/sec^3 \right],$$

where $v$ is the velocity of the body, $dW/dt$ is the time rate of doing work and $P$ is called the ‘instantaneous power’ or the ‘power’ applied to the body [227].

2.2.2 Kinetic Energy

If along any space curve ‘$C$’,

$$\frac{dW}{dt} = \langle F| \frac{dr}{dt} \rangle \rightarrow W = \int_{t_1}^{t_2} \langle F|v \rangle \, dt = m \int_{t_1}^{t_2} \langle \frac{dv}{dt}|v \rangle \, dt = m \int_{v_1}^{v_2} \langle v|dv \rangle$$

$$= \frac{1}{2}mv^2\bigg|_{v_1}^{v_2} = \frac{1}{2}mv^2_2 - \frac{1}{2}mv^2_1, \quad T_2 = \frac{1}{2}mv^2_2, \quad T_1 = \frac{1}{2}mv^2_1$$

The quantity ‘$mv^2/2$’ is called the ‘kinetic energy’ of the body often denoted ‘$T$’. Kinetic energy is proportional to the velocity of a particle. It takes energy to move a body from one point to another. Hence, the total work is equivalent to the change in the kinetic energy i.e. $W = T_2 - T_1$.

Example: Let a particle with mass ‘$m$’ move along a space curve such that

$$|r(t)\rangle = a \cos \omega t |i\rangle + b \cos \omega t |j\rangle, \quad a > b, \quad a, b, \omega > 0$$

Recall that $|i\rangle$ and $|j\rangle$ are unit vectors along the $x$- and $y$-axis respectfully. Note that $x = a \cos \omega t$ and $y = b \sin \omega t$ are the parametric equations of the ellipse:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \cos^2 \omega t + \sin^2 \omega t = 1$$
So,

\[ |F\rangle = m|a\rangle = m\frac{d^2r}{dt^2} = m\frac{d^2}{dt^2}[a \cos \omega t|i] + b \sin \omega t|j]\rangle = -m\omega^2[a \cos \omega t|i] + b \sin \omega t|j]\rangle

Moreover,

\[ |v\rangle = \frac{d|r\rangle}{dt} = -\omega a \sin \omega t|i + \omega b \cos \omega t|j\rangle \rightarrow T = \frac{1}{2}m(\omega^2a^2 \sin^2 \omega t + \omega^2b^2 \cos^2 \omega t)

At \( t = 0 \), \( T_1 = 1/2 \omega^2b^2 \) and at \( t = \pi/2 \), \( T_2 = 1/2 \omega^2a^2 \). Hence,

\[ W = T_2 - T_1 = 1/2 \omega^2a^2 - 1/2 \omega^2b^2 = \frac{m\omega^2(a^2 - b^2)}{2}

2.3 Impulse, Torque and Angular Momentum

Another way of defining force is

\[ l_2 - l_1 = m\left(\frac{v_2}{t_2} - \frac{v_1}{t_1}\right) \rightarrow \frac{\Delta l}{\Delta t} = \frac{\Delta v}{\Delta t} \rightarrow \lim_{\Delta t \to 0} \frac{\Delta l}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} \rightarrow \frac{d|l\rangle}{dt} = m|a\rangle = |F\rangle,

l_2 = mv_2, \quad l_1 = mv_1

Let

\[ I = \int_{t_1}^{t_2} F dt \quad [gm \times cm/sec] \text{ or } \int F dt + p_0, \quad p_0 = a \text{ constant}

The quantity 'I' is called the 'impulse' of a force. It is defined as the amount of time a force is acting on a body.

Notice that

\[ |I\rangle = \int_{t_1}^{t_2} |F\rangle dt = m\int_{t_1}^{t_2} \frac{d|v\rangle}{dt} dt = m\int_{t_1}^{t_2} \frac{d|v_2\rangle}{dt} = m|v_2\rangle - m|v_1\rangle = |p_2\rangle - |p_1\rangle

The impulse of a particle is equal to the change in the momentum of the particle.

2.3.1 Torque

Let

\[ |\Lambda\rangle = (r \times |F\rangle

The magnitude of \( |\Lambda\rangle \) measures to what extent a force causes a body to rotate. The quantity \( |\Lambda\rangle \) is referred to as 'torque'.
2.3.2 Angular Momentum

Now
\[ |\Lambda\rangle = \langle r | \times |F\rangle = \langle r | \times \frac{d(mv)}{dt} \rangle \rightarrow |\Lambda\rangle \, dt = \langle r | \times |mv\rangle = \langle r | \times |p\rangle = |\Omega\rangle \]

The quantity '|\Omega\rangle' is called the 'angular momentum' of a particle about a fixed axis. Note that
\[ \frac{d|\Omega\rangle}{dt} = \frac{d}{dt} \langle r | \times |mv\rangle = \left( \frac{d|v\rangle}{dt} \times |mv\rangle \right) + \langle r | \times \frac{mdv}{dt} \rangle = |0\rangle + \langle r | \times |F\rangle = \langle r | \times |F\rangle = |\Lambda\rangle \text{ [gm cm}^2/\text{sec}^2] \]

The last step follows because any vector crossed with itself produces the zero vector. Hence, the torque is the time rate of change of the angular momentum of a body. Note that |\Lambda\rangle has units of energy.

2.4 Conservation

Recall from vector analysis that if \( V(x, y, z) = V(r) \) is a scalar function differentiable at every point in a region, then
\[ \nabla V = \left( \frac{\partial}{\partial x} |i\rangle + \frac{\partial}{\partial y} |j\rangle + \frac{\partial}{\partial z} |k\rangle \right) V = \frac{\partial V}{\partial x} |i\rangle + \frac{\partial V}{\partial y} |j\rangle + \frac{\partial V}{\partial z} |k\rangle \]

is called the 'gradient' of \( V \), where \( \nabla V \) is a 'vector field'. Moreover,
\[ \int_{P_1}^{P_2} \langle \nabla V | dr \rangle = \int_{P_1}^{P_2} \left( \frac{\partial V}{\partial x} |i\rangle + \frac{\partial V}{\partial y} |j\rangle + \frac{\partial V}{\partial z} |k\rangle \right) \left( dx|i\rangle + dy|j\rangle + dz|k\rangle \right) \\
= \int_{P_1}^{P_2} \frac{\partial V}{\partial x} \, dx + \frac{\partial V}{\partial y} \, dy + \frac{\partial V}{\partial z} \, dz = \int_{P_1}^{P_2} dV = V(P_2) - V(P_1), \]

where
\[ dV = \frac{\partial V}{\partial x} \, dx + \frac{\partial V}{\partial y} \, dy + \frac{\partial V}{\partial z} \, dz \]

is a total differential and \( \int_{P_1}^{P_2} dV \) is the total change in the vector field as a body moves from a point 'P_1' to 'P_2'. Since the integral depends only on \( P_1 \) and \( P_2 \), the end points, it is independent of the path taken by the particle to get from \( P_1 \) to \( P_2 \).

2.4.1 Conservative Forces

If \(|F\rangle = -|\nabla V\rangle\) in a region, \(|F\rangle\) is called a 'conservative' force field and \( V \) is called its 'scalar potential' or just 'potential'. A force field is conservative if and only if \( \langle \nabla | \times |F\rangle = \)
But this follows immediately if $|F| = -|\nabla V|$, since $(\nabla \times |F|) = 0$ by the rules of differentiation involving $\nabla$. Recall that $\langle \nabla \times |F| \rangle$ is just the curl of $F$:

$$\langle \nabla \times |F| \rangle = \begin{vmatrix} \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & \frac{\partial}{\partial x} \end{vmatrix} |i\rangle + \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & \frac{\partial}{\partial x} \end{vmatrix} |j\rangle + \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \end{vmatrix} |k\rangle$$

$$= \left( \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) |i\rangle + \left( \frac{\partial F_2}{\partial z} - \frac{\partial F_3}{\partial x} \right) |j\rangle + \left( \frac{\partial F_1}{\partial x} - \frac{\partial F_2}{\partial y} \right) |k\rangle$$

There are two requirements if $|F|$ is to represent a conservative force field. First, the force field must be a function of position only. If, in addition to position, a force is a function of time, velocity, friction etc. or equivalently, if $(\nabla \times |F|) \neq 0$, then $|F|$ is called a ‘non-conservative’ force. Secondly, the kinetic energy required to move a particle from a point ‘$P_1$’ to a point ‘$P_2$’ in a conservative force field is independent of the path taken along the space curve that joins the points.

Example: Let $F = -m\omega^2 |r\rangle = -m\omega^2 (x|i\rangle + y|j\rangle)$, then

$$\langle \nabla \times |F| \rangle = \begin{vmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ -m\omega^2 x & -m\omega^2 y & 0 \end{vmatrix}$$

$$= \left( \frac{\partial}{\partial z} - \frac{\partial}{\partial y} \right) |i\rangle + \left( \frac{\partial}{\partial x} - \frac{\partial}{\partial x} \right) |j\rangle + \left( \frac{\partial}{\partial z} - \frac{\partial}{\partial y} \right) |k\rangle = 0$$

Hence, $|F|$ is a conservative force field.

### 2.4.1.1 A Stable Force Field

If, at a point, a particle is in equilibrium, then $|F| = -|\nabla V| = 0$. Moreover, in a conservative force field, if a particle is displaced from a point of equilibrium, but the particle tends to return to that point, the particle is called ‘stable’. A necessary and sufficient condition that a particle be stable is that the potential at that point be a minimum [227].
Example: Let $V = kx^2/2$, $k > 0$. Hence,

$$\frac{dV}{dx} = kx \rightarrow x = 0$$

is a minimum, since $d^2V/dx^2 = k > 0$. Therefore, $x = 0$ is the only possible point of stability.

### 2.4.2 The Law of Energy Conservation

If $|F⟩ = -|∇V⟩$, then $|F⟩ + |∇V⟩ = |0⟩$. Hence, $\int⟨F|dr⟩ + \int⟨∇V|dr⟩ = a constant$. To see this:

$$\int⟨F|dr⟩ = \int(F_1|i⟩ + F_2|j⟩ + F_3|k⟩|dx|i⟩ + dy|j⟩ + dz|k⟩) = \int F_1 dx + F_2 dy + F_3 dz$$

$$= \int F_1 dx + \int F_2 dy + \int F_3 dz = T$$

and

$$\int⟨∇V|dr⟩ = \int\left(\frac{∂V}{∂x}|i⟩ + \frac{∂V}{∂y}|j⟩ + \frac{∂V}{∂z}|k⟩|dx|i⟩ + dy|j⟩ + dz|k⟩\right) = \int \frac{∂V}{∂x} dx + \frac{∂V}{∂y} dy + \frac{∂V}{∂z} dz$$

$$= \int dV = V$$

Therefore,

$$\int⟨F|dr⟩ + \int⟨∇V|dr⟩ = T + V = E, \quad E = a constant,$$

where $E$ is the ‘total energy’ of the system. Hence, in a conservative force system, the total energy is a constant. This is called the law of the ‘conservation of energy’; the quantity ‘$T$’ is called the ‘kinetic’ energy and $V$ is called the ‘potential’ energy of the system. All systems that conserve energy must have a potential.

### 2.4.3 Conservation of Momentum

If $|F⟩ = |0⟩$, then

$$\frac{d}{dt}(m|v⟩) = |0⟩ \rightarrow m|v⟩ = a constant$$

If the net external force acting on a particle is zero, its momentum remains invariant. This is called the ‘conservation of momentum’. As an example, if two elastic bodies collide, their momentum before the impact must equal the momentum after the impact, i.e.
\[ m_1 |v_1⟩ + m_2 |v_2⟩ = m_1 |u_1⟩ + m_2 |u_2⟩, \]

where the \( v_i’ \)'s are the velocities before the impact and the \( u_i’ \)'s are the velocities after impact. Moreover, the idea of ‘elasticity’ comes from the ability of a body to return to its original shape after a collision. The coefficient of restitution ‘\( e \)’ is defined

\[ e = \frac{v_2 - v_1}{u_2 - u_1} \]

If \( e = 1 \), the collision is completely ‘elastic’. If \( e = 0 \), the collision is completely ‘inelastic’. Most collision are somewhere in between. The collision between two billiard balls is highly elastic whereas the collision between two automobiles or two cotton balls is highly inelastic.

**2.4.4 Conservation of Angular Momentum**

If \( \Lambda = |0⟩ \), the net external torque acting on a particle is zero, then

\[ \frac{d}{dt}(m⟨r| × |v⟩) = |0⟩ → m⟨r| × |v⟩ = a \text{ constant} \]

Therefore, the angular momentum of the particle remains invariant and this is called the ‘conservation of angular momentum’.

**2.5 Rotating Bodies**

The rotational inertia of a body depends not only on its mass, but how the mass is distributed. A baton with most of its mass located around its center of gravity will rotate much faster and it will take less effort to initiate the motion than if the baton’s mass is mostly located at the ends. ‘Rotational inertia’ or ‘moment of inertia’ \( I \) is defined as

\[ I = \sum_i m_i r_i^2, \]

where the \( m_i’ \)'s are all the little masses that make up the body and the \( r_i’ \)'s are the distances of all those little masses from the axis of rotation. If the body’s mass is roughly a continuous substance, then

\[ \Delta I = \sum r^2 (m_i - m_{i-1}) = \sum r^2 \Delta m \rightarrow I = \lim_{\Delta m \to 0} \sum r^2 \Delta m = \int r^2 \, dm \]

For instance, for a uniform rod of length \( L \), the mass element \( dm \) a distance ‘\( x \)’ from the axis of rotation, where the mass per unit length is \( m/L \), is

\[ dm = \left(\frac{m}{L}\right) dx \]
Hence,
\[ \int_{-L/2}^{L/2} x^2 \, dm = \int_{-L/2}^{L/2} x^2 \left( \frac{m}{L} \right) \, dx = \frac{1}{12} mL^2 \]

2.5.1 The Kinetic Energy of a Rotating Body

Intuitively, the kinetic energy “\( T \)” of a rotating body should be
\[ T = \frac{1}{2} \sum_i m_i v_i^2 \]

But since \( v \) is the tangential velocity, the kinetic energy at each point along the body is different. To get around this, \( v \) is replaced by \( \omega r_i \), where \( \omega \) is the angular velocity, which is the same for each mass element:
\[ T = \frac{1}{2} \omega^2 \sum_i m_i r_i^2 \rightarrow T = \frac{1}{2} I \omega^2 \]

2.5.2 The Torque of a Rotating Body

The torque ‘\( \Lambda \)’ on a body is defined as
\[ |\Lambda| = \langle r \times F \rangle = rm|a_t| \]

Recall that the tangential acceleration ‘\( a_t \)’ in radian measure is \( ar \). Therefore,
\[ \Lambda = m(ar)r = mr^2\alpha = l\alpha, \quad I = mr^2 \]

2.5.3 The Work Done by a Rotating Body

If a body of mass ‘\( m \)’, fastened to the end of a rod (of negligible mass), rotates through an angle ‘\( \theta \)’, then
\[ dW = \langle F|ds \rangle = F_t r d\theta \rightarrow \frac{dW}{d\theta} = \Lambda \rightarrow W = \int_{\theta_1}^{\theta_2} \Lambda \, d\theta \]

The above equations hold for all rigid bodies rotating about a fixed axis. Moreover,
\[ W = \int_{\theta_1}^{\theta_2} \Lambda \, d\theta = \int_{\theta_1}^{\theta_2} l\alpha \, d\theta = \int_{\theta_1}^{\theta_2} l \left( \frac{d\omega}{dt} \right) \, d\theta = \int_{\omega_1}^{\omega_2} l \left( \frac{d\theta}{dt} \right) \, d\omega = \int_{\omega_1}^{\omega_2} l \omega \, d\omega \]
2.5.4 The Power of a Rotating Body

Finally, the power:

\[ P = \frac{dW}{dt} = \Lambda \frac{d\theta}{dt} = \Lambda \omega \]

2.6 Concluding Remarks

The two great concepts in physics are the ‘particle’ and the ‘wave’. Both are described in terms of motion and rest. This chapter emphasized the motion of a rigid body. Conveniently, all the mass of the body is concentrated at its center of gravity. This allows describing the physics of a rigid body as though it was a point particle. The simple mathematical idea of a ‘space curve’ describes the motion of a body.

A body will remain at rest or move with a constant velocity, unless a force accelerates it. If a particle is in motion, various concepts associated with that motion can be defined i.e. momentum, energy, torque etc. A particle can move more or less in one direction, change direction or rotate around a fixed axis or some combination of all three.

Most problems in physics involve determining what factors in the problem remain invariant, while other aspects of the problem change. This normally involves identifying the symmetries of the problem, since symmetries are associated with ‘conservation’. If a particle has a predisposition for motion, it is said to have a ‘potential’. A conservative force is equal to the change in the particle’s potential for motion. If a particle’s potential for motion changes, but the force does not change, then the force is said to be ‘non-conservative’. A conservative force system is one in which the total energy of the system remains invariant i.e. energy is conserved.
Chapter 3

Fluids

“The ocean, whose tides respond ... to the pull of the moon.... is unstable and threatening as the earth is not; it spawns new life daily, yet swallows up lives; it is changeable like the moon, unregulated, yet indestructible and eternal.”

- Adrienne Rich

3.0 Introduction

The discussion so far has revolved around the physics of lumps of matter, where the lump is fairly stable, not readily changing shape or configuration. The center of gravity of the lump can, theoretically, be located and the dynamics of the lump can be described as though all the mass is concentrated there. The motion of the lump can be modelled by a space curve. But, unless confined to a small volume of space, substances, whose parts spread out, moving in unpredictable directions and acquire unpredictable shapes, make identifying where the center of mass is at any point in time difficult. Therefore, other methods are employed to describe the physics of ‘fluidic’ or ‘gaseous’ substances having variable configurations.

3.1 Density and Pressure (Fluids at Rest)

Physically describing a fluid requires replacing the ideas of ‘mass’ and ‘force’ with the concepts of ‘density’ and ‘pressure’. The mass of a fluid confined to a small volume of space can be measured by weighing it. The average density ‘\( \bar{\rho} \)’ of a specified material is defined as

\[
\bar{\rho} = \frac{\rho}{V} = \frac{m}{V} = \frac{\Delta m}{\Delta V} \Rightarrow \rho = \lim_{\Delta V \to 0} \frac{\Delta m}{\Delta V} = \frac{dm}{dV} \text{ [gm/cm}^3\text{]} \Rightarrow \rho g = g \frac{dm}{dV},
\]

where \( m \) is the mass, \( V \) the volume\(^3\) and \( g dm/dV \text{ [gm}^2/\text{sec}^2 \times \text{cm}^2\text{]} \) is called the ‘density weight’ of the fluid (\( g \) is the acceleration due to the gravitational force of the Earth). In the flat Earth approximation, \( g \) is a constant.

3.1.1 Specific Gravity

Different materials have different densities. The specific gravity ‘\( S_g \)’ of any substance is defined as the ratio of the density of a substance to the density (density of the same unit volume) of a reference substance i.e.

\[
S_g = \frac{\rho_m}{\rho_w},
\]

\(^3\) This assumes that the fluid can be represented as a continuous substance.
where, \( \rho_m \) represents the density of the material of interest and \( \rho_w \) is the density of water, which is the reference substance. The quantity '\( S_w \)' denotes the specific gravity of water, chosen so that \( S_w = 1 \). The density and specific gravity are numerically equal when the density is measured in units \([gm/cm^3]\), since the density of water is \(1\ [gm/cm^3]\). For example, the specific gravity of aluminum is 2.70, since its density is \(2.70\ [gm/cm^3]\) [228].

### 3.1.2 Pressure

Suppose there is a fluid enclosed in a rigid volume open at the top. If a piston of area \('\Delta A'\) is inserted snugly through the top onto the fluid and if a force \('\Delta F'\) is exerted perpendicularly onto the fluid, then average pressure \('\bar{p}'\) on the fluid is defined as

\[
\bar{p} = \frac{\Delta F}{\Delta A} \rightarrow p = \lim_{\Delta A \to 0} \frac{\Delta F}{\Delta A} = \frac{dF}{dA} \ [gm/(cm \times sec^2)]
\]

Pressure is a scalar quantity, since measuring it, at any point and in any direction in the fluid, gives the same result.

Suppose a hypothetical cylindrical canister with area \('A'\) is filled with water and dropped into a tub filled with water. At some point below the surface, it will reach equilibrium, where the upward force on the canister will equal the downward force \(F_u = F_d\). Suppose the top of the canister is a distance \('y_1'\) and the bottom of the canister is a distance \('y_2'\) from the surface, then

\[
F_u = F_d \rightarrow p_A = p_1A + \rho Ag(y_1 - y_2) \rightarrow p_2 = p_1 + \rho g(y_1 - y_2).
\]

where \(p_2A\) is the upward force created by buoyancy, \(p_1A\), the downward force due to atmospheric pressure and \(\rho Ag(y_1 - y_2)\) is the downward gravitational force on the canister (\(\rho\) is the density of water). To compute the pressure \('p'\) at a distance \('h'\) below the surface of the water, set \(y_1 = 0\), \(p_1 = p_0\), \(y_2 = -h\), \(p_2 = p\). Therefore,

\[
p = p_0 + \rho gh,
\]

where \(p_0\) is the atmospheric pressure at the surface\(^4\) and \(\rho gh\) is the pressure due to gravity. Note that the pressure depends only on the depth and not on any horizontal dimension. Thus, the pressure on a dam at its base depends only on the depth of the water and is independent of the amount of water behind the dam. Lake Mead, for example, extends many miles behind Hoover Dam [9].

#### 3.1.2.1 Pascal’s Principle

In the equation above, \('p_0'\) can be thought of as the external pressure acting on the contained fluid. By adding more external pressure \('p_{ext}'\) to \(p_0\), then the equation above can be written

\[p = p_0 + p_{ext} + \rho gh\]

\(^4\)The equation \('p = p_0 + \rho gh'\) assumes the density \(\rho\) is a constant in every sector of the fluid. Such a fluid is called ‘incompressible’. If \(\rho\) is a function of, say, \(h\), the fluid is called ‘compressible’. 
\[ p = p_{\text{ext}} + \rho gh, \]

since \( \rho gh \), the pressure from gravity, does not change.

Let \( p_1 \) be the input pressure and \( p_2 \) be the output pressure of the system shown in fig. 3.1.2.1-1. If \( p_1 - p_2 = 0 \), then

\[ \frac{F_1}{A_1} = \frac{F_2}{A_2} \rightarrow F_1A_2 = F_2A_1 \rightarrow V = A_1d_1 = A_2d_2, \]

where \( V \) is the volume of fluid moved. Note that if \( A_2 \gg 1 \), then \( F_1 \) is small, but if \( A_2 \ll 1 \) \( F_1 \) is large. From the equations above, \( W = F_1d_1 = F_2d_2 \), where \( W \) is the work done by the forces. The work done on the input piston by the applied force is equal to the work done by the output piston in lifting the load placed on it.

![Diagram of Pascal's principle](image)

**Figure 3.1.2.1-1**

The principle demonstrated here is called ‘Pascal’s principle’ named for the French mathematician, physicist, inventor, writer and Christian philosopher Blaise Pascal (19 June 1623 – 19 August 1662). It shows how a given force, exerted over a given distance, can be transformed into a larger force over a smaller distance. Stated formally:

* A change in the pressure applied to an enclosed fluid is transmitted undiminished to every portion of the fluid and to the walls of the containing vessel [9].

Pascal’s principle governs hydraulic levers. An automobile mechanic cannot lift an automobile, but welcomes the availability of a hydraulic jack, even though the jack handle has to be pumped farther than the automobile rises [9]. But it is not necessary to become an auto mechanic to appreciate Pascal’s principle, as anyone who has squeezed a tube of toothpaste will attest.

**3.1.2.2 Archimedes’ Principle**

If an aluminum boat is put into sea water, it will float. The reason it floats is because the upward force due to buoyancy just matches the downward force on the boat due to
gravity. The weight of the boat \( W_b \) is \( g \rho_b V_b \), where \( \rho_b \) is the density of aluminum and \( V_b \) is the volume of the boat. The weight \( W_w \) of the displaced water is \( W_w = g \rho_w V_w \), where \( \rho_w \) is the density of sea water and \( V_w \) is the volume of the displaced water. Since the system is in equilibrium,

\[
W_b = W_w \rightarrow \rho_w V_w g = \rho_b V_b g \rightarrow \rho_w V_w = \rho_b V_b
\]

The equations above illustrate ‘Archimedes’ principle’, named for the Ancient Greek mathematician, physicist, engineer, inventor, and astronomer Archimedes of Syracuse (287 BC – 212 BC). It states that a body wholly or partially immersed in a fluid will be buoyed by a force equal to the weight of the fluid it displaces [9].

### 3.2 Fluids in Motion

A fluid in motion is difficult to describe physically because there are so many variables involved. The most common approach involves making a series of simplifying assumptions and then calling the fluid that complies with the assumptions an 'ideal fluid'. Although there are no truly ideal fluids in Nature, the dynamics of real fluids can be approximated by the physics of an ideal fluid without egregious error. An ideal fluid, 1) has a velocity that is constant both in magnitude and direction, 2) is incompressible and non-viscous i.e. there is no friction slowing the fluid flow, and 3) the fluid does not rotate, and therefore, no accounting need be taken of torque.

#### 3.2.1 The Physics of Fluid Flow

Fluid flow ‘\( Q \)’ is defined

\[
Q = Av \ [cm^3/sec],
\]

where \( A \) is the cross sectional area of say, a tube, through which the fluid is flowing, and \( v \) is the velocity of the fluid. The volume ‘\( \Delta V \)’ of flow defined in a short time interval is

\[
\Delta V = Q \Delta t = Av \Delta t \ [cm^3].
\]

Since the fluid is incompressible, it can neither be created nor destroyed. Hence, the same volume of fluid exists at any point along the tube i.e.

\[
\lim_{\Delta t \to 0} \frac{\Delta V}{\Delta t} = \frac{dV}{dt} = Av = a \text{ constant}
\]

The flow does not change at any point along the tube i.e.

\[
\int dV = V = \int A_1 v_1 dt = \int A_2 v_2 dt \rightarrow A_1 v_1 t = A_2 v_2 t \rightarrow A_1 v_1 = A_2 v_2
\]

The flow passing a point in one section of the tube must equal the flow passing a point in any other section of the tube. Note that if \( A_1 v_1 = A_2 v_2 \), then decreasing \( A_1 \) increases \( v_1 \) and increasing \( A_1 \) decreases \( v_1 \). Anyone who has operated a simple garden hose will have experienced this phenomenon.
3.2.2 Bernoulli’s Equations

If the fluid is a continuous stream, the sum of the pressure energy, kinetic energy and potential energy at any point in a stream is equal to the sum of those energies at any other point along the stream:

\[ p_1 \frac{m}{\rho} + \frac{1}{2} \rho v_1^2 + mg h_1 = p_2 \frac{m}{\rho} + \frac{1}{2} \rho v_2^2 + mg h_2, \]

where \( m \), evidently, is the mass of the entire fluid, \( \rho \) is the density of the fluid, \( p_1, v_1, h_1 \), \( p_2, v_2, h_2 \) are the pressure, velocity and height respectfully at any two distinct points along the stream. Since energy is conserved in this system,

\[ p \frac{m}{\rho} + \frac{1}{2} \rho v + mg h = \text{a constant} \]

Since measuring the mass of an entire fluid is problematic, the equation above can be multiplied through by \( \rho m \) leaving

\[ p_1 + \frac{1}{2} \rho v_1^2 + g \rho h_1 = p_2 + \frac{1}{2} \rho v_2^2 + g \rho h_2, \]

which eliminates \( m \) from the equation and gives the law in terms of the pressure along the stream [228]. The equations above are called ‘Bernoulli’s equations’, named for the Swiss mathematician and physicist Daniel Bernoulli (8 February 1700 – 17 March 1782), who was one of the many prominent mathematicians in the Bernoulli family.

If the speed of a fluid increases as it travels along, the pressure of the fluid must decrease, and conversely. A proof of Bernoulli’s equation can be found in [9].

Figure 3.2.1-1

Figure 3.2.1-1 shows the flow of the air streamlines around an airplane wing. The air speed below the wing is less than above the wing. Hence, by Bernoulli’s principle, the pressure below the wing must be greater than the pressure above, providing lift.

3.3 Concluding Remarks

Substances that spread out, where the smaller parts freely move in unpredictable directions and acquire unpredictable shapes are usually referred to as ‘fluidic’ or ‘gaseous’. Because of the nature of these substances, identifying a point where all the
mass of the substance can be located is normally impractical. Hence, the ideas of 'mass' and 'force', associated with rigid bodies, are replaced with the concepts of 'density' and 'pressure' that generally do not require measuring mass and force directly. The density of a substance depends on the material it is made of and pressure is defined in terms of density, acceleration and distance.

There are conservation laws associated with fluidic substances. The first is 'Pascal’s principle': the work done by applying a force to a system is equal to the work done coming out of system i.e. $W = F_1 d_1 = F_2 d_2$.

The second of these conservations laws is ‘Archimedes’ principle’: a body completely or partially submerged in a fluid is acted upon by an upward or buoyant force, the magnitude of which equals the weight of the fluid displaced by the body i.e. $\rho_f V_f = \rho_b V_b$.

If the fluid is a continuous stream, the sum of the pressure energy, kinetic energy and potential energy at any point in a steam is equal to the sum of those energies at any other point along the stream. This conservation law is called ‘Bernoulli’s principle’. It is what makes airplanes fly.
Chapter 4
Heat

“With thermodynamics, one can calculate almost everything crudely; with kinetic theory, one can calculate fewer things, but more accurately; and with statistical mechanics, one can calculate almost nothing exactly.”

— Eugene Wigner

4.0 Introduction

The discussion so far has revolved around how external pressures and forces change the motion or composition of a body or a fluid or some similar substance. However, in describing these systems, a tacit assumption is that the systems are free from changes in temperature. The properties of substances change when their environments are altered. Ice, for example, has a rigid crystalline structure. But, when heated, ice melts into water, a fluid. Higher temperatures turn water into steam, a gas. This suggests that matter is just a form of energy, which can be transformed but neither created nor destroyed.

Thermodynamics is the physics of the internal changes of a substance, rather than how a substance responds when acted upon by external forces or pressures. The description of thermodynamics replaces terms like ‘force’, ‘kinetic energy’ and ‘acceleration’ with ‘temperature’, ‘heat’, ‘internal energy’ and ‘entropy’.

Scientists worry about temperature changes in the Earth’s surface because our existence seems to depend on the most delicate circumstances. Our fragile metabolism speeds up and slows down with changes in temperature. The Earth’s overall temperature of just a little bit lower would cause unlivable freezing conditions; just a bit higher and the molecules in our bodies would break apart. Our very existence hovers between fire and ice in the tiniest of ecological niches [9].

4.1 Temperature

Temperature is not a substance, but a measure of the internal energy of a system, usually referred to as ‘heat’. Since the human body, under normal conditions, remains at a relatively constant temperature, it can gauge internal temperature changes quite readily. A change of only a couple of degrees within the human body is easily noticeable and a temperature change of just six or eight degrees is life threatening. Because of the consistency of body temperature, human beings can easily sense changes in the temperature of external objects. However, our judgement of the changes in the temperature of external things is not as reliable as our sense of internal changes in our own body temperature. On a cold day, an iron pipe seems much colder than a wooden stick, but both are actually the same temperature. The difference in sensation comes because iron is a better conductor of heat than is wood, which makes iron seem colder.
4.1.1 Measuring the Amount of Heat

Determining the temperature of something, however, is surprisingly complicated. The first step requires building a ‘thermoscope’ – a device that measures changes in its own temperature. If a thermoscope has an internal temperature ‘\(x\)’ and if it is placed in an environment with thermal temperature ‘\(y\)’, after a time both the thermoscope and the environment will reach a temperature ‘\(z\)’. At that point, the thermoscope and the environment are in ‘thermal equilibrium’. The fact that thermal equilibrium is always obtained is a law of Nature, called ‘the second law of thermodynamics’. Without the second law, it would be virtually impossible to measure temperature.

Once the thermoscope is stable, measuring no changes in its own temperature, it is brought into contact with another body. If the temperatures of the two bodies are unequal, the gauge on the thermoscope will begin to vary, but, after a time, will return to thermal equilibrium. The temperature of the two bodies will be the same. If the temperature of a third body is measured and found to be the same as the two bodies in thermal equilibrium, then all three bodies have the same temperature. This can be stated formally:

4.1.1.1 The Zero\(^{\text{th}}\) Law of Thermodynamics

\[
\text{If bodies } A \text{ and } B \text{ are each in thermal equilibrium with a third body } T, \text{ then they are in thermal equilibrium with each other.} \]

This is sometimes referred to as the ‘zero\(^{\text{th}}\) law of thermodynamics’.

4.1.2 The Thermometer

The next step is to turn the thermoscope into a thermometer – a device for measuring an amount of heat. In physics, temperature is measured on the ‘Kelvin’ scale (unit symbol ‘\(K\)’), a unit of measure based upon an absolute scale, where the lowest temperature, called ‘absolute zero’, is the temperature at which all thermal motion ceases in the classical description of thermodynamics. Theoretically, there is no highest temperature. This is possibly not the case, but there is no known way of verifying whether a highest temperature exists or not. The lowest temperature ‘absolute zero’ is a hypothetical temperature. It would take an infinite amount of energy to lower the temperature of any substance to absolute zero. In essence, absolute zero is an unattainable temperature for any material substance.

4.1.2.1. The Triple Point of Water

The ‘Kelvin’, named for the British mathematical physicist and engineer William Thomson, (Lord) Kelvin (26 June 1824 – 17 December 1907), is the fraction ‘\(1/273.16\)’ of the thermodynamic temperature of the triple point of water ‘\(T_3\)’. By fiat, \(T_3 = 273.16 \text{ }[K]\). But what is the ‘triple point’ of water? Liquid water, solid ice and water vapor can co-exist in thermal equilibrium at only one set of values of pressure and temperature. The triple point of water can be achieved in the laboratory using a ‘triple
point cell’. By international agreement the triple point of water is assigned the value ‘273.16 [K]’. This is the standard measure of calibration for all thermometers.

4.1.2.2 The Standard Thermometer

For physical reasons, the standard thermometer, against which all other thermometers are calibrated, is based on the pressure exerted by a gas confined to a fixed volume (see fig. 4.1.2.2-1).

![Diagram of standard thermometer](image)

Figure 4.1.2.2-1

At a certain temperature ‘\(T\)’, the gas will exert pressure on the substance confined to a volume i.e. \(T [K] = C p\), where \(p\) is the pressure exerted by the gas and \(C\) is an undetermined constant. The pressure is determined by

\[
p = p_0 + \rho gh,
\]

where \(p_0\) is the atmospheric pressure, \(\rho\) is the density of mercury and \(h\) is the level difference. When the temperature rises, the pressure increases, and the mercury goes down.

If the gas is immersed in a triple point cell, then \(T_3 = C p_3\). Hence,

\[
\frac{T_3}{p_3} = C = \frac{T}{p} \rightarrow T = T_3 \left(\frac{p}{p_3}\right) = 273.16 [K] \left(\frac{p}{p_3}\right)
\]

The equation above is preliminary because the amount and type of gas has not been specified. Different gases will give different readings. However, if smaller and smaller amounts of gas are utilized, the readings converge regardless of what type of gas is used. Therefore,

\[
T = 273.16 K \left(\lim_{m \to 0} \left(\frac{p}{p_3}\right)\right),
\]

where \(T\) is called the ‘ideal gas temperature’, which gives the same temperature independently of the type of gas used.
4.1.2.3 Temperature Scales

There are other temperatures scales in common use. The ‘Celsius’ scale uses the same degree size as the Kelvin scale, but the zero of the temperature is shifted to a more convenient value:

\[ T_C[^\circ C] = T - 273.15^\circ \]

The measure is then in ‘degrees Celsius’ [\(^\circ C\)].

The ‘Fahrenheit’ scale, used in the United States, employs a smaller degree than the Celsius scale and its zero point is set to a different temperature:

\[ T_F[^\circ F] = \frac{9}{5} T_C[^\circ C] + 32[^\circ C] \]

The measure is then in ‘degrees Fahrenheit’ [\(^\circ F\)].

4.2 Thermal Expansion

When heated, most substances expand. If, after the temperature of a metal rod of length \(L_1\) is raised by an amount \(T_2 - T_1 = \Delta T\), the rod expands its length by \(L_2 - L_1 = \Delta L\), then

\[ L_2 = L_1 + \alpha L_1 \Delta T = L_1(1 + \alpha \Delta T) \]

The factor \(\alpha\), called the ‘coefficient of linear expansion’, has units of \([1/\text{K}]\) and depends on the material being heated and on the temperature range of interest. If the material is three dimensional, then the expansion occurs in all three dimensions.

If a liquid is heated and the temperature changes by \(\Delta T\), the increase in the volume \(\Delta V\) of the liquid is

\[ \Delta V = \beta V \Delta T \rightarrow \frac{dV}{dT} = \beta V, \]

where \(\beta [1/\text{K}]\) is called the ‘coefficient of volume expansion’. Experiments have shown that \(\beta \approx 3\alpha\) for most materials [228]. If the lid on a glass jar is stuck, applying heat will expand the lid more than the glass, which has a smaller coefficient of linear expansion, making the lid easier to remove.

4.3 Heat

If an object is heated and then placed in a room of lower temperature, it will cool until reaching thermal equilibrium with its environment. Heat is defined as the energy that flows between a system and its environment by virtue of the temperature difference between them [9].
Since heat is energy, its dimensions are \([ML^2T^{-2}]\). Specifically, 1 [\text{cal}] is the energy required to raise the temperature of 1 [\text{gm}] of water from 16.5 to 17.5 \([\text{C}^\circ]\); 1 [\text{cal}] \approx 4.184 [\text{joules}]. Recall that 1 [\text{joule}] \approx 10^7 [\text{gm \times cm^2/sec^2}].

4.3.1. Heat Capacity

The rate ‘\(C\)’ at which heat energy ‘\(Q\)’ is transferred to a substance is called ‘heat capacity’ and is proportional to the change in temperature:

\[
Q = C(T_f - T_i),
\]

where \(C\) [\text{cal/K}] is the heat capacity and \(T_f, T_i\) are the final and initial temperatures respectfully; the quantity ‘\(Q\)’ is in units ‘[\text{cal}]’ (energy).

4.3.1.1 Specific Heat Capacity

The heat capacity depends on the type of material and is proportional to the mass of the material. Hence, for any given material,

\[
Q = cm(T_f - T_i),
\]

where \(c\) [\text{cal/(gm \times C^\circ)}] is a constant called the ‘specific heat capacity’ and \(m\) is the mass of the material of interest. Of note, the concept of ‘specific heat capacity’ assumes the substance enjoys relatively stable environmental conditions, that the environmental temperature and pressure conditions remain stable, usually at room temperature. Some substances vaporize in the process. At this point, the temperature does not change with added heat, but the substance transforms from a liquid to a vapor, for example. If, however, the environmental temperature is made hot enough, virtually all substances absorb energy in the form of heat in relatively the same way.

4.3.2 Heat Transfers (Conduction, Convection and Radiation)

The mechanisms by which heat is transferred are essentially three: ‘conduction’, ‘convection’ and ‘radiation’.

4.3.2.1 Conduction

Heating a metal rod at one end will cause the atoms in the metal to vibrate. The vibrations passed along from atom to atom will cause the overall temperature of the rod to gradually increase. The process is called ‘conduction’. If the rod has length ‘\(L\)’ and cross-sectional area ‘\(A\)’ whose ends are maintained at temperatures ‘\(T_H\)’ and ‘\(T_C\)’, then

\[
\frac{Q}{t} = kA \frac{T_H - T_C}{L},
\]

where \(Q\) is the heat, \(Q/t\) is called the ‘heat flow’ and \(k\) [\text{cal/(K \times sec \times cm)}] is called ‘thermal conductivity constant’, which depends on the type of material. Materials with large \(k\) values are called ‘good conductors’ of heat. Those with small \(k\) values are ‘poor conductors’ and are usually referred to as ‘insulators’. Insulators are normally described
in terms of their ability to resist heat conduction or by ‘R-value’. The higher the R-value of a substance, the better insulator it is.

4.3.2.2 Convection

Convection is heat transfer by mass motion of a fluid when the heated fluid moves away from the source of heat, carrying energy with it. Convection above a hot surface occurs because hot air expands, becomes less dense, and rises. Hot water is less dense than cold water and rises, causing convection currents which transport energy. Convection can also lead to circulation in a liquid, as in the heating of a pot of water over a flame. Heated water expands and becomes more buoyant. Cooler, denser water near the surface descends forming patterns of circulation. A flame of a candle rising is heat energy transported by convection [9].

4.3.2.3 Radiation

Energy carried by electromagnetic waves is called ‘radiation’. Unlike conduction and convection, the heat carried in electromagnetic waves does not require a medium, such as steel or air through which to travel, as anyone who has suffered sunburn will attest. All objects emit radiation because of their temperature and also absorb some of the radiation that falls on them. For instance, the average temperature of the Earth levels off at about 300 [K], since, at that temperature, the Earth absorbs about as much radiation as it emits [9].

4.4 The First Law of Thermodynamics

The energy associated with heat is described as the amount of heat transfer between a system and its environment. At times it is difficult to define precisely what constitutes the environment and what constitutes the system. But if this difficulty is ignored, then if heat flows from the environment into the system, by fiat, the energy transfer is ‘positive’. If heat flows from the system into the environment, the energy transfer is ‘negative’.

Recall that an element of work ‘\(dW\)’ is defined as the amount of force necessary to move a body a distance ‘\(ds\)’ and is given by

\[
dW = \langle f|ds \rangle \rightarrow \langle f|v \rangle \ dt = \frac{dW}{dt} = \langle f|v \rangle
\]

There is another way of looking at this. Suppose an ideal fluid is flowing along a tube, then \(p = f/A \rightarrow f = pA\). If the fluid is flowing in the ‘\(x\)’ direction, then

\[
dW = \langle f|dx \rangle = pAdx = pdV \rightarrow \frac{dW}{dt} = pAV = \frac{\partial^2 W}{\partial S \partial t} = pv \rightarrow \frac{dW}{dt} = \int pv \ dS,
\]

where \(p\) is the pressure, \(dV\) an element of the volume, \(v\), the velocity of the fluid and \(dS\) is an element of the cross-sectional area of the tube. More generally, \(dW/dt\) is the rate of energy being gained or lost by any type of material through its surface area.
If a volume of material is hotter than its environment, it will lose heat to the environment until the material and the environment reach thermal equilibrium. According to the definition of specific heat, the amount of heat transfer \( \frac{dE}{dT} \) with temperature change \( \frac{dT}{dt} \) is

\[
dE = -c_\rho dV dT = \frac{dE}{dV} = -c_\rho dT \rightarrow dE = -\int c_\rho dTdV \rightarrow \frac{dE}{dt} = -\int c_\rho \frac{\partial T}{\partial t} dV,
\]

where \( \rho \) is the density of the material, \( dV \) is an element of volume and the negative sign indicates a net energy loss. Hence, the rate of work done on the environment must be equal to the rate of energy loss from the material i.e.

\[
\frac{dW}{dt} = \frac{dE}{dt} \rightarrow \int p\nu dS = -\int c_\rho \frac{\partial T}{\partial t} dV
\]

If any heat is generated within the volume, a term \( \frac{dQ}{dt} = \varepsilon dV = dQ/dt = d\varepsilon/dt dV \) must be added to the right-hand side of the equation above, where \( d\varepsilon/dt \) is the rate of heat generation per unit volume and \( Q \) is the amount of energy added. Therefore,

\[
\frac{dW}{dt} = \frac{dE}{dt} \rightarrow \int p\nu dS = \int -c_\rho \frac{\partial T}{\partial t} dV + \frac{dQ}{dt}
\]

Substances like a gas, for instance, possess internal energy due to the kinetic and potential energy of its molecules or atoms. This rate of internal energy generation is represented by the term \( dQ/dt \) in the equation above. Integrating with respect to \( t \),

\[
\iiint (p\nu dS) dt = -\iiint (c_\rho \frac{\partial T}{\partial t} dV) dt + Q \rightarrow W = -\Delta U + Q \rightarrow \Delta U = Q - W,
\]

where \( \Delta U = \iiint (c_\rho \frac{\partial T}{\partial t} dV) dt \) represents the work done by the system on the environment. If the volume of material is in thermal equilibrium with its environment, then \( \Delta U = 0 \rightarrow d(\Delta U)/dt = 0 \). So,

\[
\frac{dE}{dt} = \frac{dW}{dt} \rightarrow \frac{dQ}{dt} - \frac{dW}{dt} = 0 \rightarrow \int dQ - \int dW = Q - W = a \text{ constant}
\]

This is the 'first law of thermodynamics'. It is a restatement of the law of the conservation of energy phrased in thermodynamic terms.

### 4.5 The Kinetic Theory of Gases

So far, the discussion of heat has been examined in terms of volume, pressure and temperature. But substantial materials are made of atoms, and yet, there has been no discussion of the relationship between heat and the atoms that make up common materials. The internal kinetic energy of a gas, for example, is somehow related to the
jiggling of the molecules that make up the material and has something to do with its internal temperature. The kinetic theory of gases addresses this omission.

### 4.5.1 The Number of Molecules in a Given Volume of Material

How many atoms or molecules are in a volume of material? The early 19th Century Italian scientist Amedeo Avogadro (9 August 1776 – 9 July 1856), in 1811, proposed that the volume of a gas (at a given pressure and temperature) is proportional to the number of atoms or molecules regardless of the nature of the gas [230]. The number \( N_A = 6.02 \times 10^{23} \text{ [mol}^{-1}] \) is called ‘Avogadro constant’, given in inverse moles. A mole is the amount of any substance that contains as many elementary entities as there are atoms in a 12 [gm] sample of carbon-12. The term ‘elementary entities’ usually refers to atoms or molecules, but not necessarily. There are moles of water as well as moles of tennis balls. In other words, there are \( 6.02 \times 10^{23} \) tennis balls in one mole. One mole of tennis balls would equal to the volume of seven Earth moons [9].

The number of moles \( n \) contained in a sample of material is

\[
\begin{align*}
n & = \frac{\text{number of molecules in the sample}}{N_A} = \frac{\text{mass of sample}}{\text{molecular weight}} \\
& = \frac{\text{mass of sample}}{(\text{mass of one molecule})N_A}
\end{align*}
\]

### 4.5.2 The Ideal Gas Law

If an amount of several different types of gases are separately confined to a given volume and kept at a given temperature, there measured pressures will be nearly equal, but not quite. However, if the density of the gases is reduced, the small differences in measured pressures will disappear. At low enough densities, all gases obey the law

\[
pV = nRT,
\]

where \( p \) is the pressure, \( n \) is the number of moles, \( R = 8.31 \text{ [J/(mol} \times K)\text{]} \) is a constant, \( T \) is the temperature in Kelvin and \( V \) is the volume. The equation above represents the ‘ideal gas law’. There are no ideal gases in Nature. But, the physics of an ideal gas approximates the physics of real gases. Note that \( pV \) has the dimensions of energy.

### 4.5.3 The Work Done by an Ideal Gas at a Constant Temperature

Suppose a sample of \( n \) moles of a gas, confined to a piston-cylinder arrangement, are allowed to expand from a volume \( \mathcal{V}_i \) to a volume \( \mathcal{V}_f \). The work \( W \) done by the gas is

\[
W = \int_{\mathcal{V}_i}^{\mathcal{V}_f} p \, dV
\]
If the gas is ideal, \( p = nRT/V \). So,

\[
W = nRT \int_{V_i}^{V_f} \frac{dV}{V} \rightarrow W = nRT \ln \left( \frac{V_f}{V_i} \right),
\]

since \( n, R \) are constants and \( T \) is held constant. If \( V_f > V_i \), then \( W > 0 \). If \( V_f < V_i \), then \( W < 0 \).

4.5.4 The Internal Energy of an Ideal Gas

Suppose \( n \) moles of an ideal gas are confined to a box of volume ‘\( V \)’, where the temperature inside the box is held constant. The molecules will bounce around inside the box in random directions. At any time, a given molecule will have velocity ‘\(|v\rangle\)’, which can be given by its components ‘\(|v_x\rangle, |v_y\rangle, |v_z\rangle\)’. To simplify the situation, the collisions between the molecules will be ignored and only collisions with the walls of the box considered. The velocity in the \( x \)-direction after slamming into a wall will have its direction reversed, while the \( y, z \)-directions will remain unchanged. Hence, the change in momentum will be

\[-m|v_x\rangle - m|v_x\rangle = -2m|v_x\rangle\]

The molecule will hit the wall several times. The time ‘\( \Delta t \)’ between collisions is equal to the travel time to and from the opposite face i.e. \( \Delta t = 2L/|v_x\rangle \), where \( L \) is the distance between the faces of the box. Therefore, the total rate ‘\( \Delta p/\Delta t \)’ at which momentum ‘\( p \)’ is delivered to a face of the box by a single molecule is

\[
\frac{\Delta p}{\Delta t} = \frac{2m(v_x|v_x\rangle}{2L} = \frac{mv_x^2}{L}
\]

i.e. \(|f\rangle = d|p\rangle/dt\). The total force is the sum of all the contributions of the other molecules that strike the face allowing for the possibility that all of them have different speeds. The pressure ‘\( P \)’ is given by

\[
P = \frac{mv_{x1}^2/L + mv_{x2}^2/L + mv_{x3}^2/L + \cdots}{L^2} = \frac{m}{L^3} (v_{x1}^2 + v_{x2}^2 + v_{x3}^2 + \cdots)
\]

There are \( nN_A \) number of terms in the sum enclosed in the parentheses of the equation above. Hence,

\[
P = \frac{nmN_A}{L^3} \bar{v}_x^2 \rightarrow P = \frac{nM}{V} \bar{v}_x^2
\]

where \( \bar{v}_x^2 \) is the square of the average velocity in the \( x \)-direction, \( M \) is the molecular weight of the gas and \( L^3 \) is the volume of the box. Since the average velocities in any direction are equal, the squares of the components are equal i.e. \( v_x^2 = 1/3 v^2 \), since \( v^2 = v_x^2 + v_y^2 + v_z^2 \). Hence,

\[
P = \frac{nM}{3V} \bar{v}^2
\]
In the derivation above, the collisions between molecules were ignored. But a molecule that moves around in the box will have changes in speed from time to time as it collides with other molecules. Its transitional kinetic energy ‘KE’ is $KE = \frac{1}{2} mv^2$. Over time the average kinetic energy is

$$\overline{KE} = \frac{1}{2}mv^2 = \frac{1}{2}m \left( \frac{3RT}{M} \right) = \frac{3RT}{2N_A} = \frac{3}{2}kT,$$

where $k = R/N_A = 1.38 \times 10^{-23} [J/K]$ is called ‘Boltzmann’s constant’, named for the Austrian physicist and philosopher Ludwig Eduard Boltzmann (February 20, 1844 – September 5, 1906). The equation above stated in words:

\textit{At a given temperature, all gas molecules – no matter what their mass – have the same average translational kinetic energy i.e. $3/2kT$. Measuring the temperature of a gas is the same as measuring the average translational kinetic energy of its molecules [9].}

\subsection{4.5.5 Monatomic Gases}

A monatomic gas is made of a single type of atom. The internal energy ‘$U$’ of the gas is the sum of the transitional kinetic energy of its molecules. A single molecule has average kinetic energy

$$\overline{KE} = \frac{3}{2}kT$$

A sample of $n$ moles contains $nN_A$ molecules. The internal energy ‘$U$’ of the gas is then

$$U = (nN_A)\overline{KE} = \frac{3}{2}nRT$$

Therefore, the internal energy is a function of only the temperature. Suppose an amount of heat ‘$Q$’ is added to the gas. If the volume is held constant, there is a change in temperature ‘$T + \Delta T$’ and a change in pressure ‘$P + \Delta P$’. Define the molecular heat capacity ‘$C_V$’ at a constant volume as $Q = nC_V\Delta T$. Recall from the first law of thermodynamics that $\Delta U + W = Q$. Hence,

$$\Delta U + P\Delta V = nC_V\Delta T$$

Since the volume is assumed constant, $\Delta V = 0$ and

$$\Delta U = nC_V\Delta T \rightarrow C_V = \frac{1}{n} \frac{\Delta U}{\Delta T}$$

Since $\Delta U/\Delta T = 3/2nR$, then

$$C_V = \frac{3}{2}R = 12.5 [J/(mol \times K)]$$

The result agrees well with experiment, so long as the atoms in the gas are monatomic.
### 4.5.6 Diatomic and Polyatomic Gases

For diatomic and polyatomic gases, the theoretical predictions for $C_V$ are significantly different. For diatomic and polyatomic gases, $C_V = (f/2)R$, where $f$ is the number of degrees of transitional and rotational freedom in the molecules of the gas.

### 4.5.7 Adiabatic Systems

If, instead of the volume, the pressure is held constant, then $Q = nC_P\Delta T$. From the first law of thermodynamics,

$$\Delta U + P\Delta V = Q = nC_P\Delta T$$

From the ideal gas law, $PV = nRT$ and so

$$P\Delta V = nR\Delta T,$$

for constant pressure.

Hence,

$$nC_P = \frac{\Delta U}{\Delta T} + nR$$

Since $U$ depends only on the temperature, $\Delta U / \Delta T$ can be replaced by $nC_V$, therefore

$$nC_P = nC_V + nR \rightarrow C_P = C_V + R \rightarrow C_P - C_V = R$$

If $Q = 0$, there is no internal temperature change i.e. $\Delta T = 0$. This is called an ‘adiabatic system’. By removing weight from a piston-cylinder system the gas expands adiabatically. As weight is removed the volume of the container increases, and hence, both the pressure and temperature drop, since, if not the case, the first law of thermodynamics would be violated. By the first law of thermodynamics

$$Q = \Delta U + P\Delta V \rightarrow \Delta U = -P\Delta V, \quad Q = 0$$

Since the gas is ideal, $\Delta U = nC_V\Delta T$. Therefore,

$$\Delta U = -P\Delta V \rightarrow nC_V\Delta T = -P\Delta V$$

From the ideal gas law, $PV = nRT$. Hence,

$$n\Delta T = -\left(\frac{P}{C_V}\right)\Delta V$$

Note that the ideal gas law is a function of temperature only. Taking the derivative with respect to the temperature, by the chain rule

$$P \frac{dV}{dT} + V \frac{dP}{dT} = nR \rightarrow PdV + VdP = nRdT.$$
Cheating a little by replacing the 'd's' with 'Δ's' leaves

\[ P \Delta V + V \Delta P = nRT \]

And replacing \( R \) by \( C_p - C_v \), then

\[ n \Delta T = - \left( \frac{P}{C_v} \right) \Delta V = \frac{P \Delta V + V \Delta P}{C_p - C_v} \to \frac{\Delta P}{P} + \frac{C_p}{C_v} \left( \frac{\Delta V}{V} \right) = 0 \]

Next,

\[ \lim_{\Delta P, \Delta V \to 0} \left[ \frac{\Delta P}{P} + \frac{C_p}{C_v} \left( \frac{\Delta V}{V} \right) \right] \to \frac{dP}{P} + \gamma \left( \frac{dV}{V} \right) = 0, \quad \frac{C_p}{C_v} = \gamma \]

Integrating

\[ \ln P + \gamma \ln V = \text{a constant} \to PV^\gamma = \text{a constant} \]

The above equation gives the pressure/volume relationship for adiabatic processes. Recalling that, if the gas is ideal, then

\[ PV = nRT \to PV^\gamma = \frac{nRT}{V}V^\gamma \to TV^{\gamma-1} = \text{a constant} \]

### 4.6 The Second Law of Thermodynamics

Why doesn’t a spoon spontaneously get hot at one end and cool at the other? Why don’t the molecules in a room all move to one corner and stay there? These events do not require energy. The hot end of the spoon could get its energy from the cool end and the molecules do not change their kinetic energies, but just their positions [9]. If a room is cold, a heater can take the cooler air, do work and make the air warmer. The heater is an example of a ‘simple engine’. An air conditioner can take hotter air, do work and make cooler air, the concept behind a ‘refrigerator’. Hence,

\[ |Q_c| + |W| = |Q_H| \to |W| = |Q_H| - |Q_c|, \]

where \( W \) is the work done by the engine or refrigerator and \( Q_c \) and \( Q_H \) are the energies at the lower and higher temperatures respectfully.

#### 4.6.1 The Efficiency of an Engine

Multiplying the equation above by \( 1/|Q_H| \) gives

\[ e = \frac{|W|}{|Q_H|} = \frac{|Q_H| - |Q_c|}{|Q_H|} \]

where \( e \) is called the ‘efficiency’ of an engine. And

\[ |W| = |Q_H| - |Q_c| \to \frac{1}{|Q_H| - |Q_c|} = \frac{1}{|W|} \to \frac{|Q_c|}{|Q_H| - |Q_c|} = \frac{|Q_c|}{|W|} = K \]
In this case, $K$ is called the ‘efficiency’ of a refrigerator. Note that in the cases above, the engine and refrigerator are perfect, no heat is lost through outside forces like friction etc. Hence, the energy of the system can be given in terms of temperature. To see this, by the first law of thermodynamics, if $\Delta U = 0$, then

$$|Q_H| = |W_H| = nRT_H \ln \left( \frac{V_b}{V_a} \right),$$

by the ideal gas law, where $V_a$ and $V_b$ are the volumes before and after the work respectfully. Similarly,

$$|Q_C| = |W_C| = nRT_C \ln \left( \frac{V_c}{V_d} \right).$$

Hence,

$$\frac{|Q_H|}{|Q_C|} = \frac{T_H \ln \left( \frac{V_b}{V_a} \right)}{T_C \ln \left( \frac{V_c}{V_d} \right)}$$

By the law of adiabatic processes,

$$T_H V_b^{\gamma - 1} = T_C V_c^{\gamma - 1}, \quad T_H V_a^{\gamma - 1} = T_C V_d^{\gamma - 1} \rightarrow V_b/V_a = V_c/V_d$$

Hence,

$$\frac{|Q_H|}{|Q_C|} = \frac{T_H}{T_C} \rightarrow e = \frac{T_H - T_C}{T_H} \rightarrow K = \frac{T_C}{T_H - T_C},$$

where $e$ and $K$ are the efficiencies of a perfect engine and a perfect refrigerator respectfully. From above,

$$\frac{|Q_H|}{|Q_C|} = \frac{T_H}{T_C} \rightarrow |Q_H|T_C = |Q_C|T_H$$

If $T_C = 0$, $T_H \neq 0$, this represents a perfect engine with perfect efficiency, since $e = 1$. If $T_H \neq 0$ then $|Q_C| = 0$. And if $|Q_C| = 0$, then $|Q_H| = 0$. The amount of work, then, is $|W| = |Q_H| - |Q_C| = 0$. A perfect engine with perfect efficiency does no work or, in order for the engine to do work, an infinite amount of heat would be needed. This is why perpetual motion machines cannot exist.

4.6.1 The Law of Entropy

The change in the amount of energy ‘$dQ$’ per given temperature ‘$T$’ is called the entropy ‘$\Delta S$’ of a system i.e.

$$S_f - S_i = \int_l^f dS = \int_l^f \frac{dQ}{T} = \frac{1}{T} \int_l^f dQ,$$
where $i, f$ represent the initial state and final state respectfully.

Entropy is the reason an ice cube left at room temperature melts, why humans get older, never younger and why clean rooms become messy in the future. Certain things happen in one direction and not the other. The thermodynamic arrow of time (entropy) is the measurement of disorder within a system. Denoted as ‘$\Delta S$’, the change of entropy suggests that time itself is asymmetric with respect to the order of an isolated system i.e. a system will become more disorderly as time passes. To see this,

$$\Delta S_{sys} = \Delta S_{eng/ref} + \Delta S_{envir}$$

The above equation shows that the change in entropy of the system is equal to the change in the entropy of the engine/refrigerator plus the change in the entropy of the environment. By the ideal gas law,

$$Q = nRT \ln \left( \frac{V_2}{V_1} \right)$$

Since the heat absorbed by the engine is equal to the amount lost by the environment,

$$\Delta S_{sys} = \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} = \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} \rightarrow \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} - \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} = 0$$

If the engine is replaced by a refrigerator, the process is reversed and

$$\frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} = - \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} \rightarrow \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} + \frac{nRT \ln \left( \frac{V_2}{V_1} \right)}{T} > 0$$

Hence, $\Delta S_{sys} \geq 0$. This leads to the following:

In any thermodynamic process that proceeds from one equilibrium state to another, the entropy of the machine (engine/refrigerator) plus the entropy of its environment remains unchanged or increases [9].

The statement above is called the ‘second law of thermodynamics’.

4.6 Concluding Remarks

Heat is considered a form of energy and is measured in joules or ergs or some other units of energy. Measuring the amount of heat in a system requires a thermometer, which measures temperature. The thermometer is a surprisingly complicated device, at least, theoretically. Before a thermometer can be created, it is necessary to build a thermoscope – a device that measures changes in its own temperature. The amount of heat contained within an object is measured relative to its environment. Once a thermoscope reaches a stable condition, where its temperature does not change, it is in ‘thermal’ equilibrium with its environment. All objects in thermal equilibrium with each other have the same temperature. By international agreement, the triple point of water, the standard measure of calibration for all thermometers, is assigned the value
‘273.16 [K]’. All thermometers should be constructed so that the factors not related to temperature are as insignificant as possible. For instance, the volume and type of gas can impact the measurement, among other things.

Heat is defined as the energy that flows between a system and its environment by virtue of the temperature difference between the two. When heat flows between a system and its environment, the system does ‘work’ on the environment. The amount of work is proportional to the difference in temperature between the system and its environment. By fiat, if heat flows from the system to the environment, the work done takes on a negative value. If the flow is in the opposite direction, the amount of work takes on a positive value. Some substances absorb and lose heat at different rates. The rate at which a substance gains and loses heat is called its ‘heat capacity’.

The source of heat for a system comes in two forms: 1) the heat gained or lost from some external source, 2) the internal heat gained from the kinetic energy associated with the movement of the molecules the system is made of. Hence, the total work ‘W’ done by a system is

\[
W = \Delta U + Q,
\]

where \(\Delta U\) is the work associated with heat gained from or lost to the environment and \(Q\) is the internal energy of the system. The equation above represents the ‘first law of thermodynamics’. It is a statement of the law of the conservation of energy stated in terms of heat.

The characteristics of heat are normally examined in terms of the volume, pressure and temperature of a mass, without any reference to the material the substance is made of. Thermodynamics, discussed in terms of the atoms that make up common materials, is called the ‘kinetic theory of gases’. All gases at low enough densities will obey

\[
pV = nRT,
\]

called the ‘ideal gas law’. According to this law, the pressure \(p\) in any volume \(V\) of gas is proportional to the temperature and the number of gas molecules. The work ‘W’ done by an ideal gas with a change in volume is given by

\[
W = nRT \ln \left( \frac{V_f}{V_i} \right),
\]

where \(V_f\) is the final and \(V_i\) is the initial volume of the gas. The average kinetic energy \(\overline{KE}\) due to the movement of the molecules of an ideal gas is given by

\[
\overline{KE} = \frac{3}{2} kT,
\]

which is called ‘Boltzmann’s law’. Hence, the average kinetic energy due to the movement of the molecules of an ideal gas is proportional to its temperature.
The reason a spoon does not spontaneously get hot at one end and cool at the other and why the molecules in a room don’t all move to one corner and stay there is called the ‘second law of thermodynamics’. The second law is the reason that perfect engines and perfect refrigerators do no work, and hence, explain why there are no perpetual motion machines.

The change in the amount of energy per a given temperature is called the ‘entropy’ of a system. Entropy is a measure of the amount of disorder in a system. The second law demands that, in any system, entropy never decreases, but always stays the same or increases.
Chapter 5  
Electricity and Magnetism

“God runs electromagnetics on Monday, Wednesday, and Friday by the wave theory, and the devil runs it by quantum theory on Tuesday, Thursday, and Saturday.”

— Sir Lawrence Bragg

5.0 Introduction

 Barely a thought is given to the degree that modern societies depend on electronic devices. Yet, the understanding of electric phenomena was slow to materialize. In 600 B.C., the Greek philosopher Thales of Miletus (636-546 B.C.) observed that, when rubbed with silk, sparks fly from amber and that the silk attracted small pieces of straw. Moreover, Thales noticed that loadstone had similar attractive properties. But anything resembling a scientific study did not appear until William Gilbert (1540-1603) of England published ‘De Magnete’, which involved systematic experiments with electricity and magnetism. He hypothesized that the Earth’s movement around Sun was completely electromagnetic. His hypothesis was wrong, but he invented the electroscope, an instrument useful for observing electrostatic effects.

 In 1748, Benjamin Franklin established the law of the conservation of electric charge and discovered two kinds of charges – ‘positive’ and ‘negative’. He described electricity as a fluid like substance.

Charles Augustin de Coulomb (1736-1806) performed measurements of electric and magnetic forces with a delicate torsion balance, which led him to discover ‘Coulomb's law’ i.e. $F = \frac{kq_1q_2}{r^2}$, where the force between two static electric charges $q_1, q_2$ is proportional to the square of the distance ‘$r$’ between them.

The idea of ‘static charges’ didn’t square with Franklin’s “electric fluid” concept. Franklin’s idea that electricity was a fluidic substance was somewhat validated when Alessandro Volta (1745-1827) of Italy produced electric currents through a voltaic cell (electric battery) by connecting the cells in a series. And later, Hans Christian Oersted (1777-1851), a Danish Professor of Physics, showed a link between electricity and magnetism demonstrating, quite by accident, that a wire carrying a current caused a nearby compass needle to deflect. Michael Faraday, in 1831, demonstrated the reverse, that a moving magnetic field produced an electric current. In 1820, the French Physicist Andre Marie Ampere (1775-1836), through the invention of the solenoidal coil, proposed that atoms are magnetized by tiny electric currents circulating in them. And Georg Simon Ohm (1787-1854) of Germany discovered the relationship between voltage, current and resistance in a conductor.

James Clerk Maxwell, a professor at Cambridge University, England unified the previously discovered laws of electricity by an elegant demonstration of the interdependence of electricity and magnetism and showed that light is an
electromagnetic wave. His theory suffered slow acceptance. But, when Heinrich Hertz (1857-1894), in 1888, demonstrated the existence of electromagnetic waves and Guglielmo Marconi (1874-1931), in 1901, sent radio signals across the Atlantic Ocean, Maxwell’s theory won acceptance.

In 1905, Einstein showed that the manifestation of the electric or magnetic fields depended upon the relative motion of the fields. He also showed that many metals emit electrons when light shines upon them. But if light was a wave, as Maxwell had shown, how did it interact with an electron (a particle)?

In 1928, Dirac laid the foundations for quantum electrodynamics (QED) by discovering a wave equation that described the motion and spin of electrons that incorporated both quantum mechanics and the theory of special relativity; QED was fully developed and refined during the 1940’s by Richard P. Feynman, Julian S. Schwinger and Tomonaga Shin’ichirō, who proffered the idea that charged particles (electrons and positrons) interact by emitting and absorbing photons, the particles that transmit electromagnetic forces. These photons are undetectable because their existence violates the conservation of energy and momentum. The photon exchange is merely the “force” of the interaction, where particles change their speed and direction of travel by absorbing the energy of the photon. Photons are also emitted in a free state, observed as light or other forms of electromagnetic radiation.

5.1 Electrostatics

Electric charge is a mysterious property of matter. The physics of electric charge can be described, but its origins are unknown. At one time, electricity was considered a continuous fluid, but proposals arose that hypothesized that atoms were made from discrete particles that carried an electrical charge. Atoms were made of just as many positive as negative charges and are roughly electrically neutral.

Electric charge is quantized, found only in integer multiples of \( e = 1.6 \times 10^{-19} \text{ [C]} \), where \( C \) represents ‘Coulombs’, the fundamental unit of electrical charge. When an atom becomes electrically imbalanced, its net charge ‘\( q \)’ will be an integer multiple of \( e \) i.e. \( q = ne \), \( n = \pm 1, \pm 2, \pm 3, \ldots \).

5.1.1 Coulomb’s Law

The electrostatic force ‘\( F_e \)’ between two charged particles is represented by ‘Coulomb’s law’ expressed

\[
F_e = k \frac{q_1 q_2}{r^2},
\]

where \( k \) is a constant depending on the chosen units of measurement and possibly on the medium through which the force prorogates, \( q_1, q_2 \) are two charged particles and \( r \) is the distance between the charges.

The form of the equation for \( F_e \) is the same as the Newtonian law of gravity, suggesting there may be an intimate relationship between gravity and electricity. But the two forces
differ in three important aspects: 1) the electrostatic force is about $10^{38}$ times stronger than the gravitational force, 2) the gravitational force is exclusively attractive, while the electrostatic force can be either attractive or repulsive, depending on whether the charges have like or opposite charge, 3) while Newton’s law of gravity fails at the quantum level, Coulomb’s law continues to give correct answers. As yet, no relationship between electromagnetism and gravity has been found.

However, the two forces agree in three aspects: 1) both obey the law of superposition i.e. the force, on say the $q_i^{th}$ charge, is $F_i = \sum_{j=1}^{n} F_{ij}$, 2) a uniform spherical shell, where the charges exist entirely on the surface of the shell, can be treated as though all the charge exists at the center of the shell, 3) a uniform spherical shell of charges exerts no force on a charged particle inside the shell. These same characteristics hold for the gravitational force.

### 5.1.2 Conservation of Electric Charge

Particles can annihilate each other forming other particles or an atom can decay into different atom. In either case, electric charge is conserved. For instance,

$$U^{238} \rightarrow Th^{234} + He^4$$

represents a uranium ‘$U^{238}$’ nucleus decaying into thorium-234 ($Th^{234}$) and a alpha particle ‘$He^4$’. The parent nucleus ‘$U^{238}$’ has 92 protons, $He^4$ has 2 protons and $Th^{234}$ has 90 protons. Since $92 = 90 + 2$, electric charge is conserved.

Moreover, an electron and a positron can annihilate forming two oppositely directed neutrally charged gamma rays i.e. $e^- + e^+ \rightarrow \gamma + \gamma$. Let the charge on an electron be ‘−1’ and the charge on a positron ‘+1’. Since $−1 + 1 = 0$, charge is conserved. Energy is conserved if a resting electron decays into two oppositely directed gamma rays i.e. $e^- \rightarrow \gamma + \gamma$. But this decay never happens in Nature, since it would violate conservation of electric charge [9].

### 5.2 Electric Current

An imbalance in electric charge can be measured. Nature always acts to eliminate this imbalance, which is most often conceived as a ‘current’ that arises when electrons move through a copper wire or some other substance. A current persists so long as the imbalance remains. When the charge imbalance vanishes, the current ceases. Substances that allow the easy flow of electrons are called ‘conductors’, while other substances that substantially resist the flow are called ‘insulators’. A semiconductor falls somewhere in between, conducting electricity under some conditions, but not others. Its conductance varies depending on the current or voltage applied or on the intensity of irradiation by infrared, visible, ultraviolet light or $X$-rays. Superconductors are materials that do not resist current flow. If an electric current is set up in a superconductor, it will persist indefinitely, without relying on a battery or other source of power.
Current ‘i’ is defined as amount of charge ‘q’ passing through a cross section of material in one second i.e.

\[ \frac{dq}{dt} = i, \]

where the current is usually measured in amperes ([A]).

5.3 The Electric Field and Gauss’ Law

While Coulomb’s law has never been violated so far, like Newton’s law of gravitation, its “action-at-a-distance” formulation is philosophically untenable because it postulates that a force acts instantaneously over a distance. In reality, there is a time lag of at least \( \frac{L}{c} \) before the force can act, where \( c \) is the speed of light and \( L \) is the distance between the charges. Gauss formulated a ‘field’ law entirely equivalent to Coulomb’s law, but without the philosophical objections associated with the “action-at-a-distance” formulation.

The concept of a ‘field’ is one where every point in space has an assigned value. If the value at a point does not depend on direction, like temperature or pressure, the field is called a ‘scalar field’. If the value at a point is a vector quantity, the field is called a ‘vector field’. For instance, a body near the Earth’s surface will suffer an acceleration due to the Earth’s gravitational field i.e.

\[ |g\rangle = \frac{1}{m} |F_G\rangle \]

The gravitational field is a vector field, where a body with mass ‘\( m \)’ is accelerated in the direction of the force.

By the same token, if a body, carrying a positive electric charge ‘+q’, is placed near a charged rod an electrostatic force ‘|\( F_E \rangle \)’ will act on it. Hence, ‘+q’ is in a vector field called the ‘electric field’ (\( |E\rangle \)) i.e.

\[ |E\rangle = \frac{1}{q} |F_E\rangle, \]

where the charge ‘q’ will suffer an acceleration in the direction of the force.
5.3.1 Lines of Force

Electric fields can be visualized by employing lines of force.

Figure 5.3.1-1 shows the field lines of two charges of equal magnitude, one positive, one negative. The tangent to the field line on any surface gives the direction of $|E\rangle$ at any point. The field lines are drawn so that the number of lines per unit area in the plane at right angles to the line is proportional to the magnitude of $|E\rangle$. If the field lines are close together, $|E\rangle$ is large. When farther apart, $|E\rangle$ is smaller.

5.3.2 Electric Flux

Figure 5.3.2-1 shows a thin loop of wire immersed in a vector field. The rate ‘$\Phi$’ at which the vectors pass through the surface with area ‘$A$’ is called the ‘flux’, given by the equation

$$\Phi = Av \cos \theta,$$

where $v$ is the magnitude of the vectors. The electric flux is always defined on a closed surface, essentially some kind of volume. If the surface area of the volume is divided up into small squares or circles of area ‘$\Delta A$’, small enough so that each element of area can be considered a plane, then the area can be represented by a vector ‘$|\Delta A\rangle$’ whose magnitude is equal to the area ‘$\Delta A$’ and whose direction is normal to the surface (see figure 5.3.2-1). If each element of area has a constant electric field vector ‘$|E\rangle$’, there will be an angle ‘$\theta$’ between $|\Delta A\rangle$ and $|E\rangle$. The flux ‘$\Phi$’ in each element of area is given by $\Phi = \langle E | \Delta A \rangle$. The flux over the entire surface is the total of the fluxes of each individual surface element i.e.

$$\Phi = \sum \langle E | \Delta A \rangle \rightarrow \Phi = \lim_{\Delta A \to 0} \sum \langle E | \Delta A \rangle = \oint \langle E | dA \rangle,$$
where ‘∮’ indicates that the integral is along a closed surface.

\[ \varepsilon_0 \Phi = \varepsilon_0 \oint (E \, dA) = q, \]

where \( \varepsilon_0 \) is a constant that depends on the medium in which the field is immersed and \( q \) is the ‘net charge’.

\textit{Gauss’s law: the total flux over a closed surface is proportional to the net charge over the surface.}

By fiat, flux lines exiting a surface have charge ‘+ne’, where \( n \) is the number of lines exiting; those entering have charge ‘−me’, where \( m \) is the number of flux lines entering. Hence, the surface ‘\( S_1 \)’ of figure 5.3.1-1 has a net flux of \( +q/\varepsilon_0 \), since all the flux lines are exiting. The surface ‘\( S_2 \)’ has a net flux of \( -q/\varepsilon_0 \), since all the flux lines are entering the surface. And the surface ‘\( S_3 \)’ has a net flux of zero, since just as many flux lines are entering as exiting. Finally, the surface ‘\( S_4 \)’ has a net flux of \( (q - q)/\varepsilon_0 = 0 \), since there are just as many flux lines entering as exiting.

Gauss’ law requires that if an excess charge is put on an isolated conductor, the charge will move entirely to the surface of the conductor. None of the excess charge will be found within the body of the conductor \([9]\). If this were not so, there would be a field in the interior of the conductor, where just as many flux lines would exit as enter, and hence, the net charge would be zero.
5.3.3.1 The Equivalence of Gauss’ Law and Coulomb’s Law

Gauss’ ‘field law’ is entirely equivalent to Coulomb’s law. To see this,

Figure 5.3.3-1

Figure 5.3.3-1 shows a thin ring of radius ‘a’ with a constant charge density ‘\( \lambda = Q/2\pi a \)’, where \( Q \) is the total charge on the ring. Coulomb’s law cannot be applied directly, since the ring is not a point charge. To overcome this, the ring is broken up into charge elements that are small enough that Coulomb’s law can be applied i.e. \( dQ = \lambda \, dl \). Hence,

\[
dE = k \frac{\lambda dl}{x^2 + a^2},
\]

\[
\cos \theta = \frac{x}{\sqrt{x^2 + a^2}} \rightarrow \beta = \frac{\lambda (2\pi a)}{x^2 + a^2} \rightarrow \beta = k \frac{\lambda}{(x^2 + a^2)^{3/2}} \rightarrow E = k \frac{\lambda}{(x^2 + a^2)^{3/2}} \int dl
\]

If \( x \gg a \), then without appreciable error, set \( a = 0 \). Thus,

\[
E \approx k \frac{Q}{x^2} \rightarrow F = Q' E \approx k \frac{QQ'}{r^2},
\]

which is Coulomb’s law if \( x \) is replaced by \( r \).

Moreover, Gauss’s law can be written

\[
\epsilon_0 \iiint_S \left< E \right| dS = q,
\]

where \( S \) indicates a surface area. By the divergence theorem,

\[
\iiint_V \left< \nabla \right| E \right> dV = \iint_S \left< E \right| dS = \frac{q}{\epsilon_0}.
\]
for any volume ‘𝑉’ containing net charge ‘𝑞’. By the relation between charge and charge density, this equation is equivalent to

\[
\iiint_{𝑉} \langle \nabla |𝐸⟩ \, d𝑉 = \iiint_{𝑉} \frac{ρ}{\varepsilon₀} \, d𝑉 \rightarrow \frac{d}{d𝑉} \left( \iiint_{𝑉} \langle \nabla |𝐸⟩ \, d𝑉 \right) = \frac{d}{d𝑉} \left( \iiint_{𝑉} \frac{ρ}{\varepsilon₀} \, d𝑉 \right) \rightarrow \langle \nabla |𝐸⟩ = \frac{ρ}{\varepsilon₀},
\]

for any volume ‘𝑉’, where ρ is the ‘charge density’. The equation on the right-hand side above is the differential form of Gauss’ law, which is equivalent to its integral form.

5.4 Electric Potential

Gravity is a conservative force, and therefore, an object in a gravitational field has potential energy (\( P.E. \)). If a body with mass ‘\( m \)’ is lifted to a height ‘\( h \)’ above the surface of the Earth, \( P.E. = mgh \), where \( g \) is the acceleration due to gravity near the Earth’s surface. At some arbitrary point, the potential energy must be zero. This point is defined as \( h = 0 \), a point on the surface of the Earth. Like gravity, the electric force is conservative, and thus, a charge in the electric field has potential energy. In the gravitational case, the change in potential energy ‘\( \Delta U \)’ is defined as

\[
\Delta U = U_f - U_i = -W_{if},
\]

the negative of the work required to move a given mass from an initial position ‘\( i \)’ to a final position ‘\( f \)’. In the electric case, the mass is replaced by a charge ‘\( q \)’, which is placed in an electric field. The zero electric potential energy relative to any point ‘\( P \)’ is defined to be at an infinite distance from \( P \). Hence, \( U_i = 0 \) in the equation above and

\[
U = -W_\infty
\]

The potential energy ‘\( U \)’ of a charge ‘\( q_0 \)’ at any point is equal to the negative of the work ‘\( W_\infty \)’ done on the charge by an electric field as the charge moves from infinity to an arbitrary point ‘\( P \)’. If the equation above is divided by \( q_0 \), then

\[
\frac{U}{q_0} = -\frac{W_\infty}{q_0} = V \text{ [volts]}, \quad 1 \text{ [volt]} = 1 \text{ [} \frac{J}{C} \text{]},
\]

where \( V \) is called the ‘electric potential’. Energy associated with electromagnetic phenomena is normally designated in the convenient units of ‘electron volts’ i.e. one electron volt ([1 eV]) is the work required to move a single element of charge ‘\( e \)’ through a potential of 1 [volt] i.e.

\[
1 \text{ [eV]} = e\Delta V = (1.6 \times 10^{-19} \text{ [C]})(\text{ [} \frac{J}{C} \text{]}) = 1.6 \times 10^{-19} \text{ [J]}
\]

The potential difference ‘\( \Delta V \)’ is defined as \( V_f - V_i = -W_{if}/q_0 \), where \( i, f \) are the initial and final positions of \( q_0 \) within the field respectively. The work done on the charge ‘\( q_0 \)’ by the electric field ‘\( E \)’ is

\[
W_{if} = \int_i^f \langle F |ds \rangle = \int_i^f \langle q_0E |ds \rangle = q_0\int_i^f \langle E |ds \rangle \rightarrow -W_{if} = V_f - V_i = -\int_i^f \langle E |ds \rangle,
\]
5.4.1 Equipotential Surfaces

If \( q_0 \) takes a path perpendicular to \( |E| \), then \( \langle E|ds \rangle = 0 \) at all points along the path. Such a surface is called an ‘equipotential surface’. In this case, \( V_f = V_i \), so no work is performed in moving a charge along an equipotential surface.

5.4.2 Calculating the Electric Potential

Figure 5.4.2-1 shows a charge ‘+q’ that sets up an electric field ‘E’ which points radially outward in all directions. What is the potential at the point ‘P’? First,

\[
\langle E|ds \rangle = E \cos \pi (-dr) = E \, dr \rightarrow V = -\int_{l}^{f} \langle E|ds \rangle = -\int_{l}^{f} E \, dr,
\]

since \( V_i = 0 \) because \( q_0 \) comes in from infinity.

![Figure 5.4.2-1](image)

The magnitude of \( E \) at the location of \( q_0 \) is

\[
E = k \frac{q}{r^2} \rightarrow V = -kq \int_{\infty}^{R} \frac{1}{r^2} \, dr = -kq \left( -\frac{1}{r} \right)\bigg|_{\infty}^{R} = k \frac{q}{R},
\]

which is the potential at \( P \).

If there is a group of point charges, then the potential can be found by the superposition principle i.e.

\[
V = \sum_{n} V_n = k \frac{q_n}{r_n}
\]

For example, the potential for an electric dipole (one positive charge, one equal in magnitude negative charge) is

\[
\sum_{n} V_n = k \left( \frac{q}{r_+} + \frac{-q}{r_-} \right) = kq \frac{r_- - r_+}{r_- r_+}, \quad n = 2
\]

If a charge ‘\( q_0 \)’ moves through a displacement ‘\( |ds| \)’ from one equipotential surface to a neighboring one, the work ‘\( W_{if} \)’ done on the charge is
The work done by the electric field is

\[ W_{if} = -q_0 dV \]

Hence,

\[ -q_0 dV = q_0 E \cos \theta \, ds \rightarrow E \cos \theta = -\frac{dV}{ds}, \]

which gives the electric field in terms of the potential difference. Note that \( E \cos \theta \) is the component of \(|E|\) in the direction of \( |s| \). Hence,

\[ E_s = -\frac{\partial V}{\partial s} = -(\frac{\partial V}{\partial x} + \frac{\partial V}{\partial y} + \frac{\partial V}{\partial z}) \]

Finally, the potential energy between two charges \( 'q_1' \) and \( 'q_2' \) separated by a distance \( 'r' \) is

\[ U = W = q_2 V = k \frac{q_1 q_2}{r}, \quad V = k \frac{q_1}{r} \]

If the charges have the same sign, an external force would have to do positive work to push the charges together against their mutual repulsion. Hence, their potential energy would be positive. If the charges were opposite in sign, the force would do negative work to restrain their mutual attraction. The potential energy would then be negative \([9]\).

### 5.5 Moving Charges

The physics of moving charges is much like the physics of a fluid. Most materials that conduct electricity carry no electrical charge. In order for a conducting material to carry a current, there must be a non-zero potential deference in the conductor, provided by a dipole i.e. a battery, where the voltage of the battery is higher at one end than the other. When a potential difference is created, an electric current arises, conceived as electrons moving through the conductor, where slightly more electrons move in one direction than in the opposite direction. The current \( 'i' \) in a conductor is defined

\[ \frac{dq}{dt} = i \quad [1 \text{ampere} = 1 \text{Coulomb/sec}], \]

the amount of charge that passes by a cross sectional area of the conductor in time \( 't' \). Since charge is conserved, if a current \( 'i_0' \) is split into two branches \( 'i_1' \) and \( 'i_2' \), then \( i_0 = i_1 + i_2 \). Unless there is some outside influence within an isolated system, if the potential difference is constant, the current remains constant. Note that

\[ q = \int dq = \int_0^t i \, dt, \]
where \( i \), the current, may or may not be a function of time \([9]\).

### 5.5.1 Current Density

In cases where the current is not constant, the flow of charge is defined at a point along the conductor. A positive charge carrier, at a given point, will flow in the direction of the electric field ‘\( |E| \)’ at that point. The flow at this point is described by a vector ‘\( |J| \)’, called the ‘current density’, which points in the direction of \( |E| \). Hence, \( J = i / A \), where \( A \) is the cross sectional area of the conductor. For any surface, \( |J| \) is related to \( i \) by the equation

\[
i = \oint |J| dA
\]

The equation above describes current as the flux of the current density through a surface \([9]\).

### 5.5.2 Drift Speed

If the potential difference in a conductor is zero, on average, just as many electrons will move in one direction as in the other, the average velocity of the electrons is effectively zero. But if the potential difference in a conductor is not zero, there will be a non-zero average velocity ‘\( \bar{v}_d \)’, called the ‘drift speed’. The average number of charge carriers ‘\( n \)’, with drift speed ‘\( \bar{v}_d \)’, in a given length ‘\( L \)’ of the conductor, is \( nAL \), where \( A \) is the cross sectional area of the conductor and \( n \) is the average number of carriers per unit volume with drift speed ‘\( \bar{v}_d \)’. The charge ‘\( \Delta q \)’ is related to \( nAL \) by

\[
\Delta q = (nAL)e,
\]

where \( e \) is the charge on an electron; the quantity ‘\( \Delta q \)’ is the amount of charge that passes out of the volume in time ‘\( \Delta t = L / \bar{v}_d \)’. Hence,

\[
i = \frac{\Delta q}{\Delta t} = \frac{nALe}{L / \bar{v}_d} = nAe \bar{v}_d \rightarrow \bar{v}_d = \frac{i}{nAe} = \frac{J}{ne} \rightarrow |J| = ne|\bar{v}_d|
\]

Therefore, the current density ‘\( J \)’ is directly proportional to the drift speed.

### 5.5.3 Resistance and Resistivity

Conceptually, some atoms that make up a certain conductor allow easy electron flow compared to other conductors. If the same potential difference exists in two different conductors, the current in each conductor will be different. The concept that enters here is called ‘resistance’ (\( R \)), defined as \( R = V / i \) [1 ohm = 1 \( \Omega \) = 1 volt/ampere]. If the resistance varies from point to point along a conductor, then \( |J| = \rho |E| \) [\( \Omega \times m \)], where \( \rho \) is called the ‘resistivity for isotropic materials’, where the electrical properties of the material do not vary over the conductor. Since \( E = V / L \) and \( J = i / A \), \( \rho \) is defined as

\[
\rho = \frac{E}{J} = \frac{V/L}{i/A} \rightarrow \rho = \frac{R}{L} \rightarrow R = \frac{L}{A}
\]
5.5.3.1 Ohm’s Law

If \( R \) is a constant, the equation ‘\( iR = V \)’ is called ‘Ohm’s law’. If a conductor obeys Ohm’s law, the resistance between any two points is independent of the magnitude and polarity of the potential difference applied between those points on the conductor. Ohm’s law can also be expressed \( |E⟩ = \rho |J⟩ \), where \( \rho \) must be constant. In other words, a conductor obeys Ohm’s law if its resistivity is independent of the magnitude and direction of the applied electric field \([9]\). All homogeneous materials obey Ohm’s law for some range of values of the electric field, but departures result if the field is too strong.

To see this, if an electron with mass ‘\( m \)’ is in an electric field, it will accelerate by an amount

\[
a = \frac{F}{m} = \frac{eE}{m}
\]

There are generally so many free electrons in a conductor that collisions between them can be considered random. On average, those collisions will occur every \( \tau \) seconds. A typical electron will change its velocity by an amount ‘\( a\tau \)’, which is equivalent to the drift speed of the electron, since, if there were no potential difference, \( a\tau = 0 \), since \( a = 0 \). Therefore,

\[
\bar{v}_d = a\tau = \frac{eE\tau}{m}
\]

Since \( \bar{v}_d = J/ne \), then

\[
|E⟩ = \left( \frac{m}{e^2n\tau} \right) |J⟩ \rightarrow |E⟩ = \rho |J⟩, \quad \rho = \frac{m}{e^2n\tau}
\]

This equation implies that for a material to obey Ohm’s law, \( m/e^2n\tau \) must be a constant.

5.5.3.2 The Change in Resistivity with Temperature

Ohm’s law is violated if the temperature varies. The assumption is that metals that obey Ohm’s law remain at roughly room temperature. A good approximation of a change in resistivity with temperature is given by

\[
\rho - \rho_0 = \frac{\rho_0}{\rho} \alpha (T - T_0),
\]

where \( \rho_0 \) and \( T_0 \) are the resistivity at room temperature and room temperature respectfully and \( \alpha \approx 1[Ω \times m/K] \).
If an element of charge moves through a light bulb (see figure 5.5.3.2-1), its potential energy will be reduced by $dqV_L$, since, by the law of conservation of energy, some amount of energy is needed to light the light bulb. The energy element $dU$ transferred to the light bulb is given by

$$dU = dqV_L = i dt V_L \rightarrow \frac{dU}{dt} = i V_L$$

The quantity $dU/dt$ is called the ‘rate of energy transfer’. If the materials involved in the circuit obey Ohm’s law, then

$$\frac{dU}{dt} = i^2 R,$$

since $V_L = iR$.

### 5.6 Electromotive Force and Circuits

Figure 5.6-1 shows a circuit with one resistor. Any charge element $dq$, which passes through the battery, must move from a lower to a higher potential. Hence, the battery must do work on $dq$ to allow it to move:

$$E = \frac{dW}{dq}.$$
where $E [j/C = V]$ is called the ‘electromotive force’ (efm) and $E \, dq = dW$ is the element of work required to move an element of charge around the circuit. It is assumed that the battery recharges itself and maintains a constant potential difference.

Since $dU/dt = i^2 R$, then $dW = dU$ and

$$dW = E \, dq = Ei \, dt \rightarrow Ei = i^2 R \rightarrow E = \frac{i}{R}$$

The equation on the far right above follows from the law of the conservation of energy, requiring that the work done by the battery equal the thermal energy appearing in the resistor.

![Figure 5.6-2](image)

Example: Examine the circuit in figure 5.6-2. Calculate the potential at position ‘2’ around the circuit back to position ‘2’: $V_2 - 3i + 9V = V_2 \rightarrow i = 9V/3 = 3V$. In general, $V_2 - iR + E = V_2 \rightarrow i = E/R$.

5.6.1 The Loop Rule

This example illustrates the ‘loop rule’:

*The algebraic sum of the changes in the potential in a complete traversal of any closed circuit must be zero.*

5.6.1.1 The Resistor and EMF Rule

The example also illustrates two additional universal rules, 1) the resistor rule: if the traversal is done in the direction of the current, the change in potential is $-iR$; if the traversal is in the opposite direction, the change in potential is $+iR$, 2) the *emf* rule: if the traversal is in the direction of the current, the change in potential is $+E$; in the opposite direction, the change in potential is $-E$ [9].

5.6.1.2 Resistors in Series and in Parallel

In the example above, the resistance ‘r’ within the battery was ignored. If included, then $V_2 - iR - ir + E = V_2 \rightarrow i = E/(R + r)$, which illustrates an additional general rule of circuits:

*If resistors in a circuit are connected in a series, then the total resistance $R_{tot}$ is the sum of each individual resistance i.e. $R_{tot} = \sum R_n$.\*
Resistors connected in parallel produce more than one current. For example, let \(i_1 = V/R_1\), \(i_2 = V/R_2\), \(i_3 = V/R_3\), then by conservation of current

\[
i = i_1 + i_2 + i_3 = V \left( \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right)
\]

In general,

\[
\frac{1}{R_\parallel} = \sum_n \frac{1}{R_n}
\]

These general rules make analyzing complicated circuits with many efm’s, resistors and other devices in the circuits easier.

In figure 5.6-2, what is the potential difference between positions ‘2’ and ‘3’? Use of the circuit rules gives

\[
V_2 - iR = V_3 \rightarrow V_2 - V_3 = 3i \rightarrow V_2 - V_3 = 3(\mathcal{E}/(R + r))
\]

In general, \(V_a - V_b = \mathcal{E}(R/(R + r))\) i.e. to find the potential difference between any two points in the circuit, start at one point, traverse the circuit to the other point, following any path, then add the changes in potential.

### 5.6.1.3 Circuits with Varying Current

So far the circuits analyzed assumed that the current in the circuit remains constant. Adding a capacitor to a circuit stores electrical energy so that the current does not remain constant. A capacitor is a device for storing charge, made up of two parallel plates with a space between them. The plates have an equal and opposite charge on them, creating a potential difference between the plates. A capacitor can be made of conductors of any geometry. The charge ‘\(q\)’ and potential difference ‘\(V\)’ on a capacitor are related by the equation

\[
q = CV,
\]

where \(C\) is called the ‘capacitance’ of the capacitor.

The current on either side of a capacitor is different.
Figure 5.6.1.3-1 shows a circuit with an emf source (battery), resistor and a capacitor hooked in series. Using the loop rule, $\mathcal{E} - iR - q/C = 0$, where $q/C$ is the potential difference between the capacitor plates. This equation is not solvable, since there is only one equation and two variables 'i' and 'q'. But since $i = dq/dt$, $\mathcal{E} - iR - q/C = 0 \rightarrow \mathcal{E} - \frac{dq}{dt} R - q/C = 0 \rightarrow R \frac{dq}{dt} + \frac{q}{C} = \mathcal{E}$

The equation on the far right is a differential equation with solution

$$q = C \mathcal{E} (1 - e^{-t/RC}),$$

If the initial conditions are $t = 0 \rightarrow q = 0$, then

$$i = \frac{dq}{dt} = \frac{\mathcal{E}}{R} e^{-t/RC},$$

5.7 The Magnetic Field

Is the magnetic field a mirror image of the electric field? Several contemporary theories predict the existence of magnetic monopoles – magnetic charges that are a counterpart to the electric charge. But, so far, no magnetic monopoles have been observed in Nature. The idea of ‘magnetic monopoles’ survives, in part, because lodestones carry magnetic properties without the apparent existence of an electric current. But, with no evidence for the existence of magnetic monopoles, the best available explanation is that the magnetic effects of lodestones are due to the spinning and circulating of electrons within the material. Hence, contemporary theory maintains that the magnetic force arises from moving electric charges. A moving electric charge not only produces a force in the direction of the electric field, but also a magnetic force i.e. $|F_{em}| = |F_e| + |F_m|$.

5.7.1 Lorentz’s Law

The law that governs the electromagnetic force ‘$|F_{em}|$’ is called ‘Lorentz’s law’ given by

$$|F_{em}| = q(|E| + \langle v | \times |B\rangle), \quad |F_e| = q|E|, \quad |F_m| = q\langle v | \times |B\rangle$$

where $q$ is the charge, $|E|$ the electric field, $|v|$ is the velocity of the moving charge and $|B|$ represents the ‘magnetic field’. Note that if $v = 0$, the magnetic force vanishes, confirming that a magnetic force arises only when charges are moving. If the charges are stationary, the equation above reduces to the one for the electric field ‘$|F_e|$’.

Moreover, the magnetic force always acts perpendicularly to the electric force, illustrated by the rules of vector algebra, since $|F_m| = q\langle v | \times |B\rangle$ and $|v|$ points in the direction of $|E|$, $\langle v | \times |B\rangle$ is a vector perpendicular to $|E|$ i.e. $F_e \perp F_m$. Since $|F_m| = q\langle v | \times |B\rangle$, its magnitude is $F_m = qvB \sin \theta$. Note that any external electric charges moving parallel to the magnetic field i.e. $\theta = 0$, would suffer no magnetic force. Charges moving a right angles suffer a maximum force ‘$F_m = qvB$', since $\sin(\pi/2) = 1$. Finally, the direction of the deflection of a charge moving in a magnetic field depends on
whether the charge ‘\(q\)’ is positive or negative. Positive charges deflect in one direction, negative charges in the opposite direction.

In the discussion of magnetic fields, a moving charge that creates the magnetic field is always distinct from a separate charge that moves through the field. The unit associated with \(|B|\) is the ‘\([\text{tesla}] \rightarrow 1 \ [T] = 1 \ [N/(A \times m)]\)’.

5.7.2 The Biot-Savart Law

Recall that calculating the magnitude of an electric field involves taking a charge element ‘\(dq\)’ on a surface, which corresponds to an electric field element, and integrating over the entire surface i.e.

\[
|dE| = k \frac{q}{r^3} |r| \int dE = k \frac{1}{r^2} \int dq \rightarrow E = k \frac{q}{r^2}
\]

Figure 5.7.2-1 shows a current ‘\(I\)’ moving through an arbitrary conductor. The moving current sets up a magnetic field around the conductor. By analogy to the electric case the magnitude of \(|dB|\) at \(P\) is

\[
|dB| = k \frac{I(dL) \times |r|}{r^3}, \quad k = \frac{\mu_0}{4\pi},
\]

where \(\mu_0\) is a constant that depends on the medium the field is in and \(dL\) is an element of distance along the conductor. This is called the ‘Biot-Savart law’, named for the French physicist, astronomer and mathematician Jean-Baptiste Biot (21 April 1774 – 3 February 1862) and physicist Félix Savart ((30 June 1791 – 16 March 1841).

The magnitude of the field is

\[
dB = k \frac{I \sin \theta}{r^2} dL
\]
For instance, figure 5.7.2-2 shows a current moving in an infinitely long straight wire. Hence

\[ B = \int dB = kl \int_{-\infty}^{\infty} \frac{\sin \theta}{r^2} \, dl \]

The variables \( \theta, l, r \) are related by

\[ \sin \theta = \sin(\pi - \theta) = \frac{R}{\sqrt{l^2 + R^2}}, \quad r = \sqrt{l^2 + R^2} \]

Therefore,

\[ B = kI \int_{-\infty}^{\infty} \frac{R}{(l^2 + R^2)^{3/2}} \, dl = kI \frac{l}{R \sqrt{l^2 + R^2}} \bigg|_{l=\infty}^{l=-\infty} = 2k \frac{I}{R} \]

suggesting that the value of the magnetic field at any point depends on the geometry of the conductor that carries the current.

Suppose two straight parallel conductors carry currents \( I_1 \) and \( I_2 \) respectfully (see figure 5.7.2-3).
Wire ‘1’ produces a magnetic field ‘$|B_1\rangle$’, where the magnitude at wire ‘2’ is

$$B_1 = 2k \frac{I}{d}$$

Therefore, a length ‘$L$’ of wire ‘2’ will experience a magnetic force ‘$F_2$’ with magnitude

$$F_2 = 2k \frac{I_1I_2L}{d}$$

The magnetic force ‘$F_1$’ that wire ‘2’ exerts on wire ‘1’ is

$$F_1 = -2k \frac{I_1I_2L}{d},$$

but in the opposite direction. So, $|F_1\rangle = -|F_2\rangle$. There are two things of note: 1) the two wires form an action – reaction pair, 2) the external magnetic field that each wire finds itself in is the intrinsic field of the other wire. Hence, parallel currents attract each other, while anti-parallel currents repel.

The force acting between currents in parallel wires forms a basis for the definition of the ‘ampere’:

*The ‘ampere’ is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce on each of the conductors a force equal to $2 \times 10^{-7} \text{[N/m]}$ [9].*

5.7.2.1 Ampere’s Law

The Biot-Savart law, like Coulomb’s law, is formulated as “action at a distance”. And like Gauss’ law, which replaces Coulomb’s law with a field concept, Ampere’s law replaces the Biot-Savart law with a field formulation. In terms of total current, Ampere’s law is defined as the line integral of the magnetic field (in teslas, [T]) around a closed curve ‘$C$’ proportional to the total current passing through a surface ‘$S$’ (enclosed by $C$):

$$\oint(A\times|B|)ds = \mu_0i = \mu_0 \int_S (|J|dS)$$

where $|B\rangle$ is the magnetic field, $ds$ the line element, $|J\rangle$ the charge density, $dS$ the surface element and $i$ is the current. Moreover, by the Stokes’ theorem,

$$\int_S (\nabla \times |B\rangle) dS = \oint (A\times|B\rangle)ds = \mu_0 \int_S (|J|dS) \to \frac{d}{dS}\left(\int_S (\nabla \times |B\rangle) dS\right) = \mu_0 \frac{d}{dS}\left(\int_S (|J|dS)\right) = \nabla \times |B\rangle = \mu_0|J|$$

The final equation on the right-hand side above is the differential form of Ampere’s law, which is equivalent to the integral form.
Figure 5.7.2.1-1

Figure 5.7.2.1-1 shows an ‘Amperian loop’ with two infinitely straight wires carrying anti-parallel currents ‘$i_1, i_2$’ directed into and out of the plane of the page and encircled by the loop. There is a third current ‘$i_3$’ parallel to $i_1$, but outside the loop. The loop represents an arbitrary magnetic field line that sits entirely in the plane of the page. Dividing the loop into differential line segments, the line integral around the loop is

$$\oint (B | ds) = \oint B \cos \theta \, ds$$

The value of the right-side of the equation above is proportional to the algebraic sum of the currents inside the loop i.e.

$$\oint (B | ds) = \mu_0 (i_1 - i_2),$$

where, for a counterclockwise traversal of the loop, currents pointing out of the loop are positive and currents pointing inwardly are negative.

Ampere’s law depends heavily on the geometry of the circuitry and, in particular, the symmetry of the geometry. If there is only one current inside an Amperian loop, then the magnetic field lines are circles, which have continuous symmetry. Hence,

$$\oint (B | ds) = \oint B \cos \theta \, ds = B \oint ds = 2B \pi r = \mu_0 i \rightarrow B = \frac{\mu_0 i}{2 \pi r},$$

which gives the same answer as the Biot-Savart law.
Figure 5.7.2-2

Figure 5.7.2-1-2 shows a magnetic field set up in a long tightly wound helical coil called a 'solenoid'. The solenoid field is the vector sum of the fields set up in the individual turns in the wire. The solenoid behaves magnetically like a long straight wire, where the lines of $|\mathbf{B}\rangle$ associated with each turn in the wire are almost concentric circles. For points like $P$, the upper part of the solenoid field directs to the left, which cancels the field set up by the lower part of the field, which points to the right. The longer the solenoid, the more the magnitude of $|\mathbf{B}\rangle$ outside the solenoid approaches zero. If the solenoid is grasped in the right hand, the fingers curling in the direction of the wire, the thumb will point in the direction of the axial magnetic field. So, the axial magnetic field is fairly strong, where the external field is essentially zero.

If Ampere’s law is applied to the loop ‘$a, b, c, d$’ in fig. 5.7.2.1-2 (b), then

$$\oint \langle B|ds \rangle = \int_a^b \langle B|ds \rangle + \int_b^c \langle B|ds \rangle + \int_c^d \langle B|ds \rangle + \int_d^a \langle B|ds \rangle = Bh + 0 + 0 + 0 = Bh$$

The first integral is equal to $Bh$. The second and fourth integrals are zero, because the field lines are perpendicular to the path, and hence, $\langle B|ds \rangle = 0$. The third integral is zero, because $|B\rangle = |0\rangle$ outside the solenoid. The net current ‘i’ enclosed by the rectangular loop is not the same as the current ‘$i_0$’ in the windings, because the windings pass more than once through the loop. If $n$ is the number of turns in the solenoid, then

$$i = nhi_0$$

Hence,

$$Bh = \mu_0 nhi_0 \rightarrow B = \mu_0 ni_0$$

The magnetic moment of a magnet is synonymous with the torque experienced in an external magnetic field. A loop of electric current, a bar magnet, an electron (revolving around a nucleus) all have magnetic moments (see figure 5.7.2.1-3).
The torque $|\tau\rangle$ is given by

$$|\tau\rangle = \langle m \times |B\rangle \rightarrow \tau = mB \sin \phi,$$

where $|m\rangle$ is called the 'magnetic dipole moment'. Suppose there is a circular ring of radius $'R'$ carrying a current $'i'$. What magnetic field does the current set up in the surrounding space?

If there is a point $'P'$ that lies along the direction of the magnetic dipole moment a distance $'z'$ from the center of the ring, then the element of $|B\rangle$ at $P$ can be broken up into two components $|dB\rangle_\parallel$ and $|dB\rangle_\perp$; the element $'|dB\rangle_\parallel'$ lies along the magnetic dipole moment axis at $P$ and $'|dB\rangle_\perp'$ is perpendicular to it. The element $'|dB\rangle_\parallel'$ is the only component that contributes to $|B\rangle$, since $|dB\rangle_\perp$ is at right angles to the direction of the field i.e. $\int |dB\rangle_\perp = |0\rangle$. Therefore, $|B\rangle = \int |dB\rangle_\parallel$ (see fig. 5.7.2.1-4).

From the Biot-Savart law

$$dB = \frac{\mu_0}{4\pi} \frac{lds \sin(\pi/2)}{r^2} \rightarrow dB_\parallel = dB \cos \alpha \rightarrow dB_\parallel = \frac{\mu_0}{4\pi} \frac{lds \cos \alpha}{r^2},$$

where $\alpha$ is the angle between a point on the ring and a line drawn to $P$ of length $'r'$.
\[
\cos \alpha = \frac{R}{r} = \frac{R}{\sqrt{z^2 + R^2}}
\]

Therefore,
\[
dB = \frac{\mu_0}{4\pi} \frac{iR}{(z^2 + R^2)^{3/2}} ds \rightarrow B = \int dB = \frac{\mu_0}{4\pi} \frac{iR}{(z^2 + R^2)^{3/2}} \int ds = \frac{\mu_0}{2} \frac{iR^2}{(z^2 + R^2)^{3/2}}
\]

since \(\int ds = 2\pi R\), the circumference of the ring. Note that the Boit-Savant law was employed here because there is not enough symmetry in the problem to use Ampere’s law.

5.8 Electromagnetic Induction

Moving charges in a conductor generate a current, which, in turn, creates a magnetic field in the vicinity of the conductor. Does a moving magnetic field create a current?

![Diagram of magnets and conducting ring](image)

Figure 5.8-1

Figure 5.8-1 (b) shows magnets in the vicinity of a conducting ring. The magnets, surrounded by a magnetic field, produce a current in the ring if the magnet is moving (in the direction indicated in the figure) or if the magnet is stationary and the ring is moving toward the magnet. Note that if the north pole of the magnet moves though the ring in the direction indicated, the current flows in one direction, but if the south pole moves though the ring in the same direction, the current flows in the opposite direction. Moreover, if the magnet moves opposite the direction indicated in the figure, the direction of the current reverses. The potential difference through the conductor is proportional to the relative velocity of the magnetic and the conductor.

5.8.1 Faraday’s Law of Induction

Consider a surface bounded by a closed loop. The number of magnetic lines that pass through the surface is given by the magnetic flux i.e.

\[
\Phi_B = \oint (B \cdot dA) \ [1 \text{ tesla } \times \text{ meters}^2 = 1 \text{ weber}],
\]
where $|dA|$ is an element of surface area. This leads to Faraday’s law of induction:

$$\mathcal{E} = -\frac{d\Phi_B}{dt} \text{ [volts]}$$

The induced emf in a circuit is equal (except for sign) to the rate of change of the magnetic flux through the circuit.

If the change in magnetic flux is through a coil consisting of $n$ turns, an induced emf is produced in each turn and Faraday’s law then becomes

$$\mathcal{E} = -n \frac{d\Phi_B}{dt}$$

### 5.8.2 Lenz’s Law

The direction of the magnetic field created by a current moving through a conductor is called the ‘intrinsic magnetic field’. If this field is in the same direction as the external magnetic field, it would violate the law of the conservation of energy. Hence, ‘Lenz’s law’ states:

*An emf generated by a change in magnetic flux according to Faraday’s Law, the polarity of the induced emf produces a current whose magnetic field opposes the change which produces it.*

If the direction of the magnetic field that creates a current and the magnetic field of the current in a conductor are in the same direction, the two magnetic fields would add together producing a current of twice the magnitude, which, in turn, would create a higher magnitude magnetic field, causing more current. The process would continue indefinitely, leading to a violation of the law of conservation of energy. The negative sign in Faraday’s law of electromagnetic induction indicates that the induced emf ($\mathcal{E}$) and the change in magnetic flux ($d\Phi_B$) oppose one and other, having opposite signs:

$$\mathcal{E} = -n \frac{d\Phi_B}{dt} \to \mathcal{E} + n \frac{d\Phi_B}{dt} = 0 \to \mathcal{E} \int dt + n \int d\Phi_B = \mathcal{E}t + n\Phi_B = a \text{ constant},$$

which is consistent with the conservation of energy.

### 5.8.3 Induction and Energy

Mechanical work ‘$W = \langle F|ds\rangle$’ implies that the rate work ‘$dW/dt$’ done is

$$\frac{dW}{dt} = \langle F|\frac{ds}{dt}\rangle = \langle F|v\rangle$$

In figure 5.8-1 (a), since the current carrying loop is moving through an external magnetic field, the forces ‘$|F_1\rangle$, $|F_2\rangle$, $|F_3\rangle$’ will act on three of the conductors that make up the loop. Their magnitudes and directions follow from

$$|F\rangle = i\langle L| \times |B\rangle$$
The forces \(|F_2\rangle, |F_3\rangle\) cancel, being equal and opposite. Hence,

\[
|F\rangle = |F_1\rangle = i|L| \times |B\rangle \rightarrow iLB \sin \left(\frac{\pi}{2}\right) = iLB
\]

The induced current \('i'\) in the loop is \(i = \mathcal{E}/R\). By Faraday's law, \(\mathcal{E} = d\Phi_B/dt\). The absolute value of the flux \('\Phi_B'\) enclosed by the loop in the fig. 5.8-1 (a) is \(\Phi_B = BLx\), where \(Lx\) is that area of the loop in which the magnetic field is not zero. Hence,

\[
\mathcal{E} = \frac{d}{dt} (BLx) = BL \frac{dx}{dt} = BLv \rightarrow i = \frac{\mathcal{E}}{R} = \frac{BLv}{R} \rightarrow |F\rangle = iLB = \frac{B^2L^2v}{R} \rightarrow \frac{dW}{dt} = \langle F|v\rangle = \frac{(BLv)^2}{R}
\]

The last equation on the right-side is the work done by pulling the loop through the magnetic field. The thermal energy built up in the loop as a consequence of the work done by pulling the loop through the magnetic field is \(dW/dt = i^2R\). Hence,

\[
\frac{dW}{dt} = i^2R = \frac{(BLv)^2}{R} \rightarrow \frac{dW}{dt} - \frac{(BLv)^2}{R} = 0,
\]

This is expected, since, if not the case, it would violate the law of the conservation of energy.

**5.8.4 Induced Electric Fields**

If an external magnetic field is changing, it will produce a current, which, in turn, produces and electric field \(|E\rangle\), that, in turn, will exert a force \('q_0|E\rangle'\) on an external charge \('q_0'\). Hence, a changing magnetic field produces an electric field. Interestingly, the field appears even if there is no conducting material.

The work done by the electric field in moving a charge \('q_0'\) around a circle with radius \('r'\) is \(W = q_0\mathcal{E} = q_0E2\pi r\). In the more general case, Faraday's law of induction can be written

\[
\mathcal{E} = \oint (E|ds) = -\frac{d\Phi_B}{dt}
\]

The magnetic flux is given by

\[
\Phi_B = \oint (B|ds) \rightarrow \frac{d\Phi_B}{dt} = \oint \left(-\frac{\partial |B\rangle}{\partial t}\right|ds
\]

Moreover, by Stokes' theorem

\[
\int_S \langle \nabla \times |E\rangle|ds = \oint (E|ds)
\]
Hence,
\[
\int_s \langle \nabla \times |E| \rangle |dS| = \oint \langle - \frac{\partial |B|}{\partial t} \rangle |dS| = \frac{d}{ds} \int_s \langle \nabla \times |E| \rangle |dS| = \frac{d}{dt} \oint \langle - \frac{\partial |B|}{\partial t} \rangle |dS| = \langle \nabla \times |E| \rangle
\]
\[
= - \frac{\partial |B|}{\partial t}
\]

The far right-hand equation above is the differential form of Faraday’s law, which is equivalent to the integral form.

Recall that the potential difference between two points is
\[
V_f - V_i = \frac{W_{if}}{q_0} = - \int_i^f \langle E \rangle |ds|
\]

However, the electric fields involved were produced by static charges; if \( i = f \), then
\[
V_f - V_i = - \int_i^f \langle E \rangle |ds| = 0
\]

However, when a changing magnetic flux is present, \(- \frac{d\Phi_B}{dt} \neq 0\). This implies that electric potential has no meaning for electric fields associated with induction \([9]\).

### 5.8.5 Inductance

If a capacitor is placed in a circuit and if a charge ‘\( q \)’ is placed on it, a potential difference ‘\( V \)’ appears across the capacitor, its magnitude given by \( V = \frac{q}{C} \), where \( C \) is the capacitance in \([\text{farads}]\). The current on either side of the capacitor will be different. An inductor is an arrangement, where, if a current ‘\( i \)’ is set up in its windings, then a magnetic flux ‘\( \Phi \)’ appears in each winding. The ‘inductance’ (\( L \)) is defined by
\[
L = \frac{n\Phi}{i} \quad [1 \text{ henry} = 1 (T \times m^2)/A],
\]

where \( n \) is the number of turns in the inductor.

If two inductors are near each other, a current ‘\( i \)’ in one of the inductors will produce a magnetic flux ‘\( \Phi \)’ through the second inductor. If the flux is altered by changing the current, an induced \( \text{emf} \) will appear in the second coil according to Faraday’s law. Hence, since \( n\Phi = LI \) and \( \mathcal{E} = - \frac{nd\Phi}{dt} \), then
\[
\mathcal{E} = -L \frac{di}{dt}
\]

In any inductor a self-induced \( \text{emf} \) appears whenever the current changes. The minus sign in the equation above reflects Lenz’s law – that the induced \( \text{emf} \) opposes the change that brings it about.
If an *emf* ‘ℰ’ is introduced into a single loop circuit containing a resistor ‘R’ and a capacitor ‘C’, the charge built up in the capacitor is not immediate, but is a function of time:

\[ q(t) = CE(1 - e^{-t/\tau_c}), \quad \tau_c = RC = a \text{ constant} \]

If the *efm* is suddenly removed, the decrease in charge is

\[ q(t) = q_0 e^{-t/\tau_c} \]

An analogous rise and fall of the current occurs by introducing a current into a single-loop circuit with a resistor and an inductor ‘L’ (see figure 5.8.5-1).

![Figure 5.8.5-1](image)

The inductor causes an *emf* ‘ℰₗ’ to build up which opposes the *emf* ‘V’ in the battery. Hence, the resistance in the resistor will correspond to the constant one due to V and a variable one due to the self-inductance ‘ℰₗ = −L di/dt’. Hence, \( R = V/i - \mathcal{E}_l/i \rightarrow R < V/i \). Applying the loop rule to the circuit in figure 5.8.5-1 (assumed closed at \( t = 0 \), \( i(0) = 0 \)) gives

\[-iR - L \frac{di}{dt} + \mathcal{E} = 0 \rightarrow iR + L \frac{di}{dt} = \mathcal{E}\]

The equation above is a differential equation with solution

\[ i(t) = \frac{\mathcal{E}}{R} \left(1 - e^{-t/\tau_L}\right), \quad \tau_L = L/R = a \text{ constant}, \]

where \( L/R \) has dimensions [sec]:

\[
\left[\frac{\text{1 henry}}{\text{ohm}}\right]\left[\frac{\text{1 volt \times sec}}{\text{henry \times ampere}}\right]\left[\frac{\text{1 ohm \times ampere}}{\text{volt}}\right] = 1 \text{ [sec]}
\]

Over time, \( \mathcal{E}_L \rightarrow 0 \) at which point \( i = \mathcal{E}/R \). If, at that point, the battery is removed from the circuit, then

\[ iR + L \frac{di}{dt} = 0 \]
If \( i(0) = i_0 = \mathcal{E}/R \), then

\[
i(t) = \frac{\mathcal{E}}{R} e^{-t/\tau_L}
\]

is the decay in the current over time, which will eventually drop to zero asymptotically.

Since

\[
\mathcal{E} = iR + L \frac{di}{dt} \rightarrow i\mathcal{E} = i^2 R + Li \frac{di}{dt},
\]

if a charge ‘\( dq \)’ passes through the battery in time ‘\( dt \)’, the battery does work at the rate of \( \mathcal{E} dq/dt = i\mathcal{E} \). Hence, \( i\mathcal{E} \) is the rate at which the battery delivers energy to the circuit. The term ‘\( i^2 R \)’ is the rate at which thermal energy is built up in the resistor. Finally, ‘\( Li \; di/dt = dU_B/dt \)’ is the rate at which energy is stored in the magnetic field. Hence,

\[
dU_B = Li \; di \rightarrow \int dU_B = \int Li \; di \rightarrow U_B = \frac{1}{2} Li^2,
\]

which represents the total energy stored by an inductor ‘\( L \)’ carrying a current ‘\( i \)’.

### 5.8.6 Gauss’ Law for Magnetism

An ordinary wall magnet cut in half makes two wall magnets. Two magnets each cut in half result in four magnets, then eight magnets and so on. Expressed differently, isolated magnetic poles do not exist in Nature, at least, none that have been discovered so far. The physical law that articulates this is called ‘Gauss’ law for magnetism’, given by the equation

\[
\Phi_B = \oint \langle B | dA \rangle = 0
\]

In other words, the magnetic flux through any Gaussian surface is zero. Contrast this to Gauss’ law for electricity i.e.

\[
\Phi_E = \varepsilon_0 \oint \langle E | dA \rangle = q,
\]

the electric flux through any Gaussian surface is \( q \neq 0 \), which predicts the existence of an isolated charge. Moreover, Gauss’s law for magnetism can be written

\[
\Phi_B = \oiint_S \langle B | dS \rangle = 0
\]

By the divergence theorem, this equation is equivalent to

\[
\iiint_V \langle \nabla | B \rangle \; dV = \oiint_S \langle B | dS \rangle = 0,
\]
for any volume \( V \). Hence,
\[
\iint_V \langle \nabla \mathbf{B} \rangle \, dV = 0 \rightarrow \frac{d}{dV} \left( \iint_V \langle \nabla \mathbf{B} \rangle \, dV \right) = \frac{d}{dV}(0) \rightarrow \langle \nabla \mathbf{B} \rangle = 0,
\]
for any volume \( V \). The final equation on the right-hand side above is the differential form of Gauss’ law for magnetism, which is equivalent to the integral form.

### 5.9 Electromagnetic Oscillations

Are electromagnetic and mechanical phenomena governed by the same physics? An indication of this can be found in the oscillations of \( LC \) circuits. In a simple closed \( LC \) circuit, if the capacitor carries a charge \( q \) and the current \( i \) in the inductor is zero, the energy \( U_E \), stored in the capacitor is \( U_E = q^2 / 2C \) and the energy \( U_B \) in the inductor is \( U_B = 1/2 Li^2 = 0 \), since \( i = 0 \).

Recall from Chapter 1 that harmonic motion satisfies the differential equation
\[
\frac{d^2z}{dt^2} = -\omega^2z, \quad \omega = \text{a constant}
\]
In a simple conservative mechanical system
\[
U = U_T + U_P,
\]
where \( U \) is the total energy, \( U_T \), the kinetic energy and \( U_P \) is the potential energy. Since the system is assumed conservative, \( dU / dt = 0 \). For a simple vibrating system
\[
U = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \rightarrow \frac{dU}{dt} = \frac{d}{dt} \left( \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \right) = mv \frac{dv}{dt} + kx \frac{dx}{dt} = 0,
\]
\[
k = \text{a constant}
\]
But \( dx / dt = v \) and \( dv / dt = d^2x / dt^2 \). Hence,
\[
mv \frac{dv}{dt} + kx \frac{dx}{dt} = 0 \rightarrow m \frac{d^2x}{dt^2} = -kx,
\]
which is the equation of simple harmonic motion for a mechanical system. The solution ‘\( x \)’ to this differential equation is \( x(t) = A \cos(\omega t + \phi) \).

An \( LC \) circuit is also a conservative system. Hence,
\[
U = U_B + U_E = \frac{1}{2}Li^2 + \frac{1}{2}q^2 / C \rightarrow \frac{dU}{dt} = 0
\]
Therefore,
\[
\frac{dU}{dt} = \frac{d}{dt} \left( \frac{1}{2}Li^2 + \frac{1}{2}q^2 / C \right) = Li \frac{di}{dt} + \frac{q}{C} \frac{dq}{dt} = 0
\]
But \( \frac{dq}{dt} = i \) and \( \frac{di}{dt} = \frac{d^2q}{dt^2} \). Hence,
\[
L \frac{d^2q}{dt^2} = -\frac{1}{C} q
\]
The solution ‘\( q \)’ to this differential equation is \( q(t) = Q \cos(\omega t + \phi) \). Note that the \( LC \) oscillator has the same solution as the mechanical oscillator, the only difference is the labeling, in which \( x \) is replaced by \( q \) and \( k \) is replaced by \( 1/C \). However, \( \omega \) in the \( LC \) circuit does not equal ‘\( \omega \)’ in the mechanical system i.e. \( \omega = \sqrt{k/m} \) in the mechanical system, while \( \omega = 1/\sqrt{LC} \) in the \( LC \) system. To see this,
\[
\frac{dq}{dt} = -\omega Q \sin(\omega t + \phi), \quad \frac{d^2q}{dt^2} = -\omega^2 Q \cos(\omega t + \phi)
\]
Hence,
\[
L \frac{d^2q}{dt^2} + \frac{1}{C} q = -L\omega^2 Q \cos(\omega t + \phi) + \frac{1}{C} Q \cos(\omega t + \phi) = 0 \rightarrow \omega = \frac{1}{\sqrt{LC}}
\]
The analysis above assumed that there was no resistance in the circuit. If electromagnetic energy is lost to a thermal heat buildup in the resistor, then the circuit is no longer conservative and \( dU/dt = -i^2R \). Hence,
\[
Li \frac{di}{dt} + \frac{q dq}{C} = -i^2R \rightarrow L \frac{di}{dt} + R \frac{dq}{dt} + \frac{q}{C} = 0
\]
The equation above reduces to the conservative case if \( R = 0 \).

5.10 Concluding Remarks

Isolated electric charges exist in Nature. Electric charge is quantized, coming in basic units ‘\( e = 1.6 \times 10^{-19} \) [C]’. The law governing the force ‘\( F_e \)’ between two charges is called ‘Coulomb’s law’ express by
\[
F_e = k \frac{q_1 q_2}{r^2}
\]
Electric charge is conserved in all elementary particle annihilations and decays. Gauss’s law replaces Coulomb’s law with an equivalent field formulation in which the charge creates an electric field ‘\( |E| \)’, where the total flux over a closed surface is proportional to the net charge over the surface:
\[
\epsilon_0 \Phi = \epsilon_0 \oint \langle E \rangle dA = q
\]
The electric force is conservative. A charge in an electric field has potential energy. If a charge is an infinite distance from an arbitrary point ‘\( P \)’, the potential energy at \( P \) is zero. The potential energy ‘\( U \)’ of a charge ‘\( q_0 \)’ at any point is equal to the negative of the work
‘$W_\infty$’ done on the charge by an electric field as the charge moves from infinity to an arbitrary point ‘$P$’ i.e.

$$U = -W_\infty$$

A moving charge in a conductor is called a ‘current’. There must be a potential difference in a conductor for a current to exist. Most conductors obey Ohm’s law ($iR = V$), which gives the relationship between the current, the resistance in the conductor and the potential difference at all points in the conductor. If there is a current in a circuit, there must be a force that creates the current. This force is called the ‘electromotive force’ ($emf$). Ohm’s law can then be written $iR = \mathcal{E}$, where $\mathcal{E}$ denotes the seat of the power that creates the potential difference in a circuit, and hence, the current.

A moving electric charge not only produces a force in the direction of the electric field, but produces a magnetic force i.e. $|F_{em}| = |F_e| + |F_m|$. The law that governs this electromagnetic force is called ‘Lorentz’s law’ given by

$$|F_{em}| = q(|E| + \langle v \rangle \times |B|)$$

The law governing the magnetic field ‘$|B|$’ is called the ‘Biot-Savart law’:

$$|dB| = k\frac{I(|dL| \times |r|)}{r^3}, \quad k = \frac{\mu_0}{4\pi}$$

Ampere’s law replaces the Biot-Savart “action at a distance” formulation with the field formulation defined as

$$\oint \langle B |dS \rangle = \mu_0 i = \mu_0 \int_{S} \langle j |dS \rangle$$

Moving charges in a conductor create a magnetic field and a moving magnetic field creates a current, and hence, an $emf$. This is called ‘electromagnetic induction’ and is governed by Faraday’s law of induction:

$$\mathcal{E} = -\frac{d\Phi_B}{dt}$$

Lenz’s law states that the $emf$ generated by a change in magnetic flux according to Faraday’s Law produces a current whose magnetic field opposes the change which produces it.

There appears to be no isolated magnetic poles in Nature, expressed by ‘Gauss’s law for magnetism’:

$$\Phi_B = \oint \langle B |dA \rangle = 0$$

Are electromagnetic phenomena and the mechanics of bodies governed by the same physics? The physics of $LC$ circuits is governed by the same equations as mechanical
oscillations. But no definite link between, say gravity (mechanical oscillations) and electromagnetism has been found.
Chapter 6
Light and Optics

“All the fifty years of conscious brooding have brought me no closer to answer the question, “What are light quanta?” Of course today every rascal thinks he knows the answer, but he is deluding himself.”

— Albert Einstein

6.0 Introduction

Is light a particle or a wave? The corpuscular (particle) theory of light, advanced by the French philosopher, priest, scientist, astronomer and mathematician Pierre Gassendi (22 January 1592 – 24 October 1655) and the English political philosopher Thomas Hobbes of Malmesbury (5 April 1588 – 4 December 1679), argues that light is made of small discrete particles that possess energy and travel in straight lines with a finite velocity. Isaac Newton, having passed a beam of white light through a prism at such an angle that the light splits into a spectrum found that the light recomposed into white light after passing through a second prism. He concluded that the “geometric nature of reflection and refraction” could only be explained if light was corpuscular, because waves tended not to travel in straight lines.

Newton’s theory took precedence over Huygens’ wave front theory for more than 100 years partly because of Newton’s great prestige. But in 1801, the English physician and physicist Thomas Young (June 13, 1773 — May 10, 1829) allowed light to pass through two closely set pinholes onto a screen and found that the light beams spread apart and interfered producing bands of alternating bright light and darkness. With his demonstration of ‘interference’, Young established the wave nature of light. In 1817, he proposed that light waves were transverse (vibrating at right angles to the direction of travel), rather than longitudinal (vibrating in the direction of travel), as had long been assumed, and this explained the polarization of light.

Young’s work was disparaged by most English scientists, since any opposition to a theory of Newton’s was inconceivable. It was only through the work of the French engineer and physicist Augustin-Jean Fresnel (1788 – 1827) and the French mathematician, physicist, astronomer and freemason Dominique François Jean Arago (1786 – 1853), who showed that the corpuscular theory failed to adequately explain diffraction, interference and polarization, did the wave theory of light finally achieved acceptance in Europe.

To some extent, Newton’s corpuscular theory re-emerged in the twentieth century. The current theory of light has aspects of both the corpuscular and wave concepts, where light, as presently explained, is both a particle and a wave, depending on how is it measured or how it interacts with its environment.
6.1 Geometric Optics

Reflection is the abrupt change in the direction of propagation of light that strikes the boundary between two different media. The reflected part of the light stays in the same medium it started in. Refraction is the change in direction of propagation of light when it passes from one medium into another. Light waves are refracted when crossing the boundary from one transparent medium into another because the speed of light is different in different media.

6.1.1 The Law of Reflection

A light ray changes direction when it strikes the boundary between two different media. At least some part of the incoming ray stays in the same medium. That part is said to be ‘reflected’. If the incoming light ray makes an angle ‘\( \theta_1 \)’ with the normal to the boundary, the reflected ray makes an angle ‘\( \theta'_1 \)’ with this normal and lies in the same plane as the incident ray such that (see figure 6.1.1.1-1)

\[ \theta_1 = \theta'_1 \]

Smooth reflection occurs at plane boundaries. Reflection at rough, irregular boundaries is ‘diffuse’.

6.1.1.1 Reflectivity

The ‘reflectivity’ or ‘reflectance’ of a surface material is the fraction of energy in the incoming ray that is reflected. The reflectivity of a mirror is close to 1, virtually the entire ray is reflected.

![Figure 6.1.1.1-1](image)

6.1.2 Refraction

Refraction is the change in direction of a light ray when it passes from one medium into another. The change in direction is due to a variation in light speed. The amount of light refracted when crossing the boundary from one transparent medium into another is a function of the media through which the light passes. Different media affect the speed of light differently. If a light ray encounters the plane surface of a piece of glass after traveling initially through air (see figure 6.1.1.1-1), the speed of light in glass is less than...
the speed of light in a vacuum or air. The speed of light in a given substance is given by

\[ v = c/n \rightarrow n = c/v, \]

where \( n \) is the ‘index of refraction’ of the substance, \( c \) is the speed of light in a vacuum (air) and \( v \) is the speed of light in the medium (glass).

**6.1.2.1 Snell’s Law**

Light rays will bend when passing from air into glass. The amount of bending depends on the angle of incidence and on the indices of refraction of glass and air, which determine the change in speed. From experiments,

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2, \]

where \( n_1, n_2 \) are the indices of refraction of the given media (see figure 6.1.1.1-1).

When light passes from one transparent medium to another, the rays are bent toward the surface normal if the speed of light is smaller in the second medium than the first. Rays are bent away from the normal if the speed of light in the second medium is greater than the first. In the “glass” example, one part of the wave is reflected and the other part is refracted as it passes from air into glass. Since the index of refraction of glass is greater than the index of refraction of air, the rays are bent towards the normal. If the ray passes from glass into air, the rays would be bent away from the normal.

**6.1.3 Total Internal Reflection**

Total internal reflection occurs when a ray strikes a medium boundary at an angle larger than a particular ‘critical angle’ with respect to the normal to the surface. If the refractive index is lower on the other side of the boundary and the incident angle is greater than the critical angle, the wave cannot pass through and is entirely reflected (see fig. 6.1.3-1).

![Figure 6.1.3-1](image)

From the equation above

\[ \theta_c = \sin^{-1}(\frac{n_2}{n_1}), \]

where \( \theta_c \) is determined by setting \( \theta_2 = \pi/2 \); the quantity ‘\( \theta_c \)’ is the ‘critical angle’, where at angles larger than \( \theta_c \), there is no refracted ray, but total reflection (see fig. 6.1.3-1).
The sine of an angle cannot exceed unity. Hence, \( n_2 < n_1 \) must hold, otherwise total internal reflection cannot occur, since the incident light would be in a medium of lower index of refraction. The reflected ray does not lose intensity (the rate at which light energy is delivered to a unit of surface). By example, the reflection from a mirror has an intensity loss of about 4% [9].

6.1.4 Light Polarization

A light wave vibrating in more than one plane is called ‘un-polarized’. Light emitted by the sun, by a lamp, or by a candle flame is un-polarized light. Un-polarized light waves are created by electric charges that vibrate in a variety of directions and can be conceptualized as a wave having half its vibrations in the horizontal plane and half in the vertical plane. Polarized light vibrates in a single plane. Un-polarized light can be transformed into polarized light by passing it through a ‘polarizer’.

![Figure 6.1.4-1](image_url)

6.1.4.1 Brewster’s Law

Figure 6.1.4-1 shows an un-polarized beam falling on a glass surface. The electric field vector for each wave train can be resolved into its perpendicular, represented by the dots, and its parallel components, represented by the arrows. Un-polarized light has both components equal in amplitude. For most dielectric materials, in this case, glass, there is a particular angle of incidence, called the ‘Brewster angle’ \((\theta_B)\), at which the reflection of the parallel component is zero. By experiment, the angle of refraction \(\theta_r\) is perpendicular to the angle of reflection i.e. \(\theta_B + \theta_r = \pi/2\). By the law of refraction

\[
n_1 \sin \theta_B = n_2 \sin \theta_r = n_2 \sin(\pi/2 - \theta_B) = n_2 \cos \theta_B \rightarrow \theta_B = \tan^{-1} \left( \frac{n_2}{n_1} \right)\]

The equation on the far right is called ‘Brewster’s law’, named for the Scottish physicist, mathematician, astronomer and inventor Sir David Brewster (11 December 1781 – 10 February 1868).
6.1.5 The Optics of a Plane Mirror

The light rays striking a plane mirror, for practical purposes, are totally reflected. Since the angle of reflection equals the angle of incidence, a plane mirror makes an image of the objects in front of it; these images will appear to be behind the plane in which the mirror lies. The location of the image is found by extending the reflected rays in a straight line in the direction behind the plane of the mirror (see figure 6.1.5-1). The image will appear where the extended rays intersect. The image formed is ‘virtual’, since the light rays do not actually come from the image. The image will be upright and of the same shape and size as the object it is reflecting, but will appear rotated 180° from its object i.e. if a person is reflected in a plane mirror, their right hand will appear to be the left hand of their image.

![Figure 6.1.5-1](image)

6.1.6 The Optics of a Spherical Mirror

Plane mirror images appear upright and of the same shape and size as their objects. Moreover, the apparent distance of the image from the mirror will be the same as the distance of the object from the mirror i.e. $d_o = -d_i$ (see figure 6.1.6-1 a)), the minus sign signifying that the image appears on the other side of the mirror relative to its object (virtual image).
6.1.6.1 Spherically Concave Mirrors

If the mirror is bent towards the object into a semi-circular shape, the mirror is 'spherically concave'. The distance $\overline{CV} = R$ is the 'radius of curvature' and the point 'C' is called the 'center of the mirror'. The distance $\overline{FV} = f$ is called the 'focal distance' and the point 'F' is called the 'focal point' (see figure 6.1.6-1 b)).

If an object (black) is placed at the center of the mirror, the image will appear at the center. The image is 'real' because it appears on the same side of the mirror as its object, and therefore, the light rays are actually producing the image. Moreover, the image (green) will appear inverted with the same size as its object (see figure 6.1.6-1 b)). To find the image, two rays are sent from the top of the object, one parallel to $\overline{CV}$, which passes through the focal point after reflection. The second ray passes through
the focal point and then parallel to $\overline{CV}$ upon reflection. The intersection point of these two reflected rays gives the location of the image.

If the object is moved away from the center of the mirror, the image formed is real, inverted, appears shrunken, but closer (see figure 6.1.6-1 c)). If the object is moved between the center of the mirror and the focal point, the image formed is real, inverted, appears magnified, but farther away (see figure 6.1.6-1 d)). If the object is placed at the focal point, the reflected rays run parallel and do not intersect. The image is formed at infinity and will disappear (see figure 6.1.6-1 e)). If the object is placed between the focal point and the mirror, the image formed is virtual. Since the rays reflected from the mirror do not intersect, their extensions intersect behind the mirror, forming a virtual image. The image is magnified and upright (see figure 6.1.6-1 f)).

### 6.1.6.2 Spherically Convex Mirrors

If the mirror is bent away from the object into a semi-circular shape, the mirror is 'spherically convex' (see figure 6.1.6.2-1). A ray travelling from the top of the object parallel to $\overline{CV}$ is reflected, its virtual extension passes through the virtual focal point 'F'. A ray aimed at the virtual center ‘C’ is reflected back along its own path. A ray aimed at the virtual focal point is reflected back along a path parallel to $\overline{CV}$. Its virtual image will also be parallel to $\overline{CV}$. The Image will appear upright and de-magnified where the three dashed lines intersect. For convex mirrors, no matter where the object is placed in front of the mirror, a virtual, upright, de-magnified image is formed "behind the mirror" between F and V. If the image is at F, its size will reduce to zero and the object is considered an infinite distance from the mirror. As the object is moved closer to the mirror, the image size increases, until, at V, the image and the object are almost the same size.

![Figure 6.1.6.2-1](image.png)

### 6.1.7 The Relationship between Image Distance and Object Distance

Whether a spherical mirror is concave or convex, there is a relationship between the distance '$d_o$' of the object from the mirror, the distance '$d_i$' of the image from the mirror and the focal distance i.e.
\[ \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}, \quad f = \frac{R}{2}, \]

where \( R \) is the radius of curvature. A proof of this relationship can be found in [9]. Note that if \( R \to \infty \), then \( 1/f \to 0 \) and \( d_o = -d_i \), which is the distance equation of a plane mirror.

### 6.1.7.1 Magnification

The lateral magnification ‘\( m \)’ of a mirror is given by \( m = -d_i/d_o \). For a plane mirror, \( d_i = -d_o \to m = +1 \). If \( m > 0 \), the image is upright. If \( m < 0 \), the image is inverted. The equations above hold good not only for mirrors, but spherical refracting surfaces and lenses, subject to the following rules:

1. \( d_o > 0 \) if the object is real (on the real-side of the mirror)
2. \( d_i > 0 \) if the image is real
3. \( R > 0 \) if \( C \) is real
4. \( f > 0 \) if \( F \) is real
5. \( m > 0 \) if the image is upright

For magnification, note that triangle \( OFe \) is similar to the triangle \( Ifb \) (see figure 6.1.7.1-1). Hence, by the law of similar triangles

\[ \frac{Ib}{Oa} = \frac{IF}{aF} = \frac{IV}{aV} \]

But, \( IV = d_i \) and \( aV = d_o \), then letting

\[ m = Ib/Oa \to m = d_i/d_o \]

Since the image is inverted, \( m < 0 \), so \( m = -d_i/d_o \), which is the equation for magnification [9].

### 6.1.8 Spherical Refracting Surfaces

If light rays emanate from an object ‘\( O \)’ onto a convex spherical refracting surface, the rays, instead of being completely reflected, are refracted. The surface separates into two media with index of refraction ‘\( n_1 \)’ and ‘\( n_2 \)’ respectively. After refraction, the rays join
to form a real image at ‘I’ (see figure 6.1.8-1). Note that, since the surface is refracting, the rays emulate from the virtual side (V-side) and transmit to the real side (R-side), unlike mirrors, where rays emanate from the real side.

Note that, if the surface is concave, the transmitted rays diverge and their virtual extensions form a virtual image at ‘I’.

\[ \theta_1 = \alpha + \beta, \quad \theta_2 + \gamma = \beta \]
\[ \alpha \approx \tan \alpha = \frac{AV}{o}, \quad \beta = \frac{AV}{r}, \quad \gamma = \frac{AV}{l}, \]
if \( \alpha \) is small.

Consider the convex spherical surface shown in figure 6.1.8-2. By applying the exterior angle theorem to the triangles ‘OAC’ and ‘IAC’, then

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2 \]
\[ n_1 \theta_1 = n_2 \theta_2 \]

Figure 6.1.8-2
Remembering that \( n_1 \sin \theta_1 = n_2 \sin \theta_2 \) (Snell’s law), then
\[
n_1 \sin \theta_1 = n_2 \sin \theta_2 \rightarrow n_1 (\alpha + \beta) = n_2 (\beta - \gamma) \rightarrow n_1 \alpha + n_2 \gamma = (n_2 - n_1) \beta,
\]
since, if \( \theta \) is small, \( \sin \theta \approx \theta \). Hence,
\[
n_1 \frac{AV}{o} + n_2 \frac{AV}{i} = \frac{AV}{r} (n_2 - n_1) = \frac{n_1}{o} + \frac{n_2}{i} = \frac{(n_2 - n_1)}{r}
\]
Letting \( o = d_o \), \( i = d_i \), \( r = R \), then
\[
\frac{n_1}{d_o} + \frac{n_2}{d_i} = \frac{n_2 - n_1}{R}
\]
where \( d_o \) is the distance of the object from \( V \), \( d_i \) is the distance of the image from \( V \), \( R \) is the radius of curvature and \( n_2, n_1 \) are the indices of refraction for each media. The equation above is general and holds whether the refracting surface is concave or convex and even when \( n_2 < n_1 \). The rules for signs for a single refracting surface are the same as for spherical mirrors [9].

6.1.9 Thin Lenses

In most cases, there is more than one refracting surface which is true of thin lenses. If the thickness of the lens is thin compared to \( d_o \), then
\[
\frac{1}{d_0} + \frac{1}{d_i} = \frac{1}{f}, \quad f = (n - 1) \left( \frac{1}{r_1} - \frac{1}{r_2} \right)
\]
where \( u = d_o \), \( v = d_i \), \( r_1 \) and \( r_2 \) are the radii of curvature on each side of the lens and \( n = n_{lens}/n_{medium} \) (see figure 6.1.9-1). The rules for signs are the same as for mirrors and refracting surfaces. A proof of the equation above can be found in [9].

![Figure 6.1.9-1](image)

A convergent lens has two focal points ‘\( F_1, F_2 \)’. It takes three rays to define the position and size of the image. Ray ‘1’ runs parallel to the axis of the lens on the virtual side and passes through \( F_2 \). Ray ‘2’ passes through \( F_1 \) and emerges parallel to the axis of the lens. Ray ‘3’ is directed from the virtual side towards the center of the lens and emerges un-deflected. The converging lens produces a real inverted image.
Figure 6.1.9-2 shows a divergent lens, which produces a virtual upright image.

![Figure 6.1.9-2](image)

**6.2 The Wave Theory of Light**

The first to advance a wave theory of light was the prominent Dutch mathematician and scientist Christiaan Huygens (14 April 1629 – 8 July 1695) in 1678. To its great merit, it accounted for the laws of reflection and refraction in terms of the wave picture, which gave a physical meaning to the index of refraction [9]. It also explained light phenomena that the corpuscular theory could not.

**6.2.1 Huygens Principle**

Huygens principle describes where a wave front will be at any future time ‘t’ if the location of the wave front is known at a previous time ‘t₀’. Huygens principle states:

*All points on a wave front serve as point sources of spherical secondary wavelets. After a time ‘t – t₀’, the new position of the wave front will be the surface of tangency to these secondary wavelets* (see figure 6.2.1-1).

![Figure 6.2.1-1](image)
In time \( t - t_0 \) the radius of the spherical wavelets is \( c(t - t_0) \), where \( c \) is the speed of light. Since all the wavelets travel at a constant speed, a new wave front parallel to the old wave front appears so long as the waves pass through the same medium.

### 6.2.2 The Wave Theory and Snell’s Law

Figure 6.2.2-2 shows a light wave front passing from air into glass. The distance between wave fronts will be shorter in the glass than in air, since the waves travel a smaller distance per period \( T \). If \( v \) is the frequency of the wave and \( T = 1/\nu \) is the period, i.e. the time interval between successive crests passing a fixed point in space, then \( \lambda_1 = v_1T \) and \( \lambda_2 = v_2T \), where \( v_1, v_2 \) are the velocities of light and \( \lambda_1, \lambda_2 \) are the wave lengths in air and in glass respectfully. Since \( v = c/n \), then

\[
\lambda_1 = \frac{c}{n_1} T, \quad \lambda_2 = \frac{c}{n_2} T
\]

Hence,

\[
\frac{\lambda_1}{\lambda_2} = \frac{\frac{c}{n_1}T}{\frac{c}{n_2}T} = \frac{n_2}{n_1}
\]

![Figure 6.2.2-2](image)

The light rays will bend as the wave front passes from air into glass because the wave fronts do not travel as far in one cycle in the glass compared to air. The amount of bending depends on the angle of incidence and on the indices of refraction of glass and air, which determine the change in light speed. From Figure 6.2.2-2, \( \lambda_1/\lambda_2 = \sin \theta_1/\sin \theta_2 \). But, \( \lambda_1/\lambda_2 = n_2/n_1 \). Therefore,

\[
\frac{n_2}{n_1} = \frac{\sin \theta_1}{\sin \theta_2} \rightarrow n_1 \sin \theta_1 = n_2 \sin \theta_2,
\]

which is Snell’s law of refraction.
6.2.3 Thomas Young’s Experiment

If parallel wave fronts fall on a barrier with an opening of similar dimensions to the wavelength, the wave front will flare out as it passes through the opening. This is called ‘diffraction’ (see figure 6.2.3-1). The smaller the opening, the more the light diffracts. Diffraction is difficult to explain using the corpuscular theory of light. This is one reason the wave theory replaced the corpuscular theory.

![Figure 6.2.3-1](image)

6.2.3.1 Interference

Young’s experiment is shown in figure 6.2.3.1-1, where a wave front falls on a barrier ‘A’ and passes through a small slit ‘S₀’ . As it does, it diffracts and the wave front continues eventually falling upon a second barrier ‘B’, where it passes through two slits ‘S₁’ and ‘S₂’. The wave front then breaks into two wave fronts, which pass through one another. Upon the screen ‘C’, where the two waves have reinforced, appear light bands labeled ‘Max’. Dark bands appear between the ‘Maxes’, where the crests and troughs of the two waves cancel. This light phenomenon is called ‘interference’, one of the prime characteristics of light. Optical interference is critical to understanding many natural phenomena, ranging from color shifting in butterfly wings to intensity patterns formed by small apertures, which cannot be explained by geometrical optics.

![Figure 6.2.3.1-1](image)
6.2.3.2 Coherent Waves

Consider the setup shown in fig. 6.2.3.2-1. The electric fields at any time ‘t’, at a point ‘P’ in space radiated by the two slits ‘S₁’ and ‘S₂’ will be given by

\[ E_1 = E_0 \cos \left( 2\pi \frac{r_1}{\lambda} - \omega t \right), \quad E_2 = E_0 \cos \left( 2\pi \frac{r_2}{\lambda} - \omega t \right), \]

where \( r_1 = 2\pi r_1 / \lambda \) and \( r_2 = 2\pi r_2 / \lambda \). By the trigonometric identity

\[ \cos A + \cos B = 2 \cos \frac{A + B}{2} \cos \frac{A - B}{2} \rightarrow E = 2E_0 \cos \left( \frac{\phi_1 - \phi_2}{2} \right) \cos \left( \frac{\phi_1 + \phi_2}{2} - \omega t \right), \]

by letting \( A = \phi_1 - \omega t \) and \( B = \phi_2 - \omega t \). This is an oscillating electric field with magnitude

\[ E_m = 2E_0 \cos \left( \frac{\phi_1 - \phi_2}{2} \right) \rightarrow E = E_m \cos \left( \frac{\phi_1 + \phi_2}{2} - \omega t \right) \]

Assuming the amplitude of both waves is \( E_0 \) and the initial phase ‘\( \phi \)’ is zero. The net electric field ‘\( E \)’ at point ‘\( P \)’ is

\[ E = E_0 \cos \left( 2\pi \frac{r_1}{\lambda} - \omega t \right) + E_0 \cos \left( 2\pi \frac{r_2}{\lambda} - \omega t \right) \rightarrow E = E_0 \cos(\phi_1 - \omega t) + E_0 \cos(\phi_2 - \omega t), \]

Consider an observation point making an angle ‘\( \theta \)’ with an axis passing through the middle of the two sources ‘\( S_1 \)’ and ‘\( S_2 \)’ and assume \( r_1, r_2 \gg d \). In other words, the position of \( P \) is very large in comparison to the distance ‘\( d \)’ between the two light sources (see figure 6.2.3.1-1) i.e. \( y \gg d \).

Consistent with the above assumption, \( r_1 \parallel r_2 \). The path difference \( r_2 - r_1 = S_1 b \) can be given by

\[ r_2 - r_1 = d \sin \theta \]
Since
\[
\frac{\phi_1 - \phi_2}{2} = \frac{\pi}{\lambda} (r_2 - r_1) \rightarrow \frac{\phi_1 - \phi_2}{2} = \frac{\pi}{\lambda} d \sin \theta,
\]
the electric field magnitude is
\[
E_m = 2E_0 \cos \left( \frac{\pi}{\lambda} d \sin \theta \right)
\]
The quantity \(E_m\) will be a maximum whenever
\[
\frac{\pi}{\lambda} d \sin \theta = m\pi \rightarrow d \sin \theta = m\lambda, \quad m = 0, 1, 2, ...
\]
Moreover, \(E_m\) will be a minimum whenever
\[
\frac{\pi}{\lambda} d \sin \theta = \left( m + \frac{1}{2} \right) \pi \rightarrow d \sin \theta = \left( m + \frac{1}{2} \right) \lambda, \quad m = 0, 1, 2, ...
\]
In the Young experiment there is a tacit assumption that both the sources are of the same frequency and phase. Hence, the electric fields at any time \(t\), at a point \(P\) in space radiated by the two sources will be given by
\[
E_1 = E_{10} \cos \left( 2\pi \frac{r_1}{\lambda} - \omega t + \varphi \right), \quad E_2 = E_{20} \cos \left( 2\pi \frac{r_2}{\lambda} - \omega t + \varphi \right)
\]
The phase difference of these two fields at any point in space does not vary in time i.e.
\[
\left( 2\pi \frac{r_1}{\lambda} - \omega t \right) - \left( 2\pi \frac{r_2}{\lambda} - \omega t \right) = 2\pi \frac{r_1}{\lambda} - 2\pi \frac{r_2}{\lambda}
\]
Any two light waves whose phase difference does not vary with time are called 'coherent'. Interference produced by coherent waves is stable in time. If light wave generators synchronously oscillate with the same frequency, the waves are coherent. Laser light sources are examples of coherent light waves. All the atoms/molecules in a laser source emit light at the same frequency. If the atoms/molecules in two sources of light emit different frequencies and phases, conducting Young’s interference experiment would not produce distinct light and dark bands, produced by coherent sources. Interference would occur at different times so that the human eye would be unable to distinguish an interference pattern. Light produced by an electric bulb is an example of incoherent light as all the atoms in its filament emit light in a random fashion, independent of each other [231].

6.3 Concluding Remarks

Is light a particle or a wave? Opinions on this question have changed over time. At first, the corpuscular theory of light dominated, with Newton being its strongest advocate. About one hundred years after Newton’s death, Thomas Young’s experiments showed that light exhibited wave qualities – interference, diffraction and polarization that the corpuscular theory had difficulty explaining. Shortly thereafter,
Clerk Maxwell united the theories of optics, light and electromagnetism into a single theory, which revealed that light was an electromagnetic wave. Once Marconi sent light waves across the Atlantic, Maxwell's wave theory was virtually universally accepted and dominated until the twentieth century. However, the corpuscular theory partly reemerged when it was discovered that light, when falling on a filament of gold foil, ejected electrons from the foil, which could only be explained if light had corpuscular properties.

This led to a dilemma. How could light be both a particle and wave? The predicament was partially resolved by the quantum theory of light, where light is explained, not as a thing itself, but in terms of how it interacts with the environment. According to the quantum theory of light, those experiments that measure the wave properties of light will view light as a wave. Those experiments that measure the corpuscular properties of light will view light as a particle.

Optics is the branch of physics that studies how light interacts with matter and how certain instruments impact the behavior of light. Optics usually involves visible, ultraviolet and infrared light. When light is conceived as an electromagnetic wave, other forms of electromagnetic radiation such as X-rays, microwaves, and radio waves exhibit similar properties [258].

Geometrical optics describes the propagation of light in terms of "rays", which travel in straight lines. The paths of light rays are governed by the laws of reflection and refraction [259]. These laws have been used in the design of optical components and instruments and can be summarized as follows:

When a ray of light hits the boundary between two transparent materials, it is divided into a reflected and a refracted ray:

1. **The law of reflection**: the reflected ray lies in the plane of incidence and the angle of reflection equals the angle of incidence.

2. **The law of refraction**: the refracted ray lies in the plane of incidence, and the sine of the angle of refraction divided by the sine of the angle of incidence is a constant.

Reflections can be divided into two types: ‘specular reflection’ and ‘diffuse reflection’. Specular reflection describes the gloss of surfaces such as mirrors, which reflect light in a simple, predictable way, allowing for production of reflected images that can be associated with a real or virtual location in space.

Diffuse reflection describes light reflected from non-glossy materials, such as paper or rock. The reflections from these surfaces can only be described statistically, with the exact distribution of the reflected light depending on the microscopic structure of the material.

Specular reflected rays are determined by the angle the incident ray makes with the surface normal, a line perpendicular to the surface at the point where the ray hits. The incident and reflected rays and the normal lie in a single plane and the angle between
the reflected ray and the surface normal is the same as that between the incident ray and the normal [260]. This effect is known as the ‘Law of Reflection’.

For flat mirrors, the law of reflection implies that images of objects are upright and the same distance behind the mirror as the objects are in front of the mirror. The image size is the same as the object size. The law also implies that mirror images are parity inverted, a left-right inversion [260].

Mirrors with curved surfaces can be modelled by tracing the rays and using the law of reflection at each point on the surface. For mirrors with parabolic surfaces, parallel rays incident on the mirror produce reflected rays that converge at a common focus. Other curved surfaces may also focus light, but with aberrations due to the diverging shape of the surface causing the focus to be smeared out. In particular, spherical mirrors exhibit spherical aberration. Curved mirrors can form images with magnification greater or less than one. Magnification can be negative, indicating that the image is inverted. An upright image formed by reflection in a mirror is always ‘virtual’, while an inverted image is real and can be projected onto a screen [260].

Refraction occurs when light travels through an area of space that has a changing index of refraction, allowing for the production of lenses and the focusing of light. The simplest case of refraction is governed by Snell's Law:

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2, \]

where \( \theta_1 \) and \( \theta_2 \) are the angles between the normal (to the interface) and the incident and refracted waves respectively [260]. The index of refraction ‘\( n \)’ of a medium is related to the speed ‘\( v \)’ of light in that medium:

\[ n = \frac{c}{v} \]

where \( c \) is the speed of light in vacuum.

For light rays travelling from a material with a high index of refraction to a material with a low index of refraction, Snell's law predicts that there is no transmission; all the light is reflected. This is called ‘total internal reflection’ and is the concept behind fibre optic technology. As light travels down an optical fibre, it undergoes total internal reflection allowing no light to be lost over the length of the cable [260].

A device which produces converging or diverging light rays due to refraction is known as a ‘lens’. Lenses are characterized by their focal length: a converging lens has positive focal length, while a diverging lens has negative focal length. A smaller focal length indicates that the lens has a stronger converging or diverging effect.

Ray tracing can be used to show how images are formed by a lens. For a thin lens in air, the location of the image is given by the equation

\[ \frac{1}{d_0} + \frac{1}{d_i} = \frac{1}{f}, \quad f = (n - 1) \left( \frac{1}{r_1} - \frac{1}{r_2} \right) \]
where $d_0$ is the distance from the object to the lens and $d_i$ is the distance from the lens to the image, $r_1$ and $r_2$ are the radius of curvature on each side of the lens and $n = n_{\text{lens}} / n_{\text{medium}}$.

Incoming parallel rays are focused by a converging lens onto a spot one focal length from the lens, on the far side of the lens. This is called the ‘rear focal point’ of the lens. Rays from an object at a finite distance are focused further from the lens than the focal distance; the closer the object is to the lens, the further the image is from the lens.

Diverging lenses diverge the incoming parallel rays after going through the lens so that the original rays appear one focal length in front of the lens. This is the lens’ ‘front focal point’. Rays from an object at a finite distance are associated with a virtual image that is closer to the lens than the focal point and on the same side of the lens as the object. The closer the object is to the lens, the closer the virtual image is to the lens. As with mirrors, upright images produced by a single lens are virtual, while inverted images are real [260].

Physical optics is synonymous with the ‘wave theory of light’, which predicts interference and diffraction, not explained well by geometric optics. The wave model can make predictions about how an optical system will behave without requiring an explanation of what is "waving" in what medium. Until the middle of the 19th century, most physicists believed in an "ethereal" medium through which the light disturbance propagated. The existence of electromagnetic waves was predicted in 1865 by Maxwell. These waves propagate at the speed of light, require no medium and have varying electric and magnetic fields orthogonal to one another and also to the direction of propagation of the waves. In other words, light is polarized. Light waves are treated as electromagnetic waves except when quantum mechanical effects must be considered.

Polarization is a general property of waves that describes the orientation of their oscillations. For transverse waves, such as electromagnetic waves, polarization is the orientation of the oscillations in the plane perpendicular to the wave’s direction of travel. The oscillations may be oriented in a single direction (linear polarization) or may rotate as the wave travels (circular or elliptical polarization). Circularly polarized waves can rotate rightward or leftward in the direction of travel and which of the two rotations is present in a wave is called ‘chirality’ [260].
Chapter 7
Non-Relativistic Classical Physics

“As in Mathematics, so in Natural Philosophy, the Investigation of difficult Things by the Method of Analysis ought ever to precede the Method of Composition. This Analysis consists in making Experiments and Observations, and in drawing general Conclusions from them by Induction, and admitting of no Objections against the Conclusions, but such as are taken from Experiments, or other certain Truths. For Hypotheses are not to be regarded in experimental Philosophy.”

- Sir Isaac Newton

7.0 Introduction

So far, the discussion has revolved around physical concepts. As yet, there has been little discussion of how those concepts arise from the underlying assumptions of a theory. The next few chapters will remedy this omission.

There are several methods for investigating the laws of Nature. Which method is chosen largely depends on convenience and the problem to be solved. Nevertheless, the general trend in physics has been to gravitate toward methods that promote the unification of the subject. In other words, the objective is to explain the maximum number of phenomena with as few assumptions as possible, keeping in mind that the validity of those assumptions relies on the test of observation and experiment. It will be instructive to examine how the methods have evolved over time towards unification.

7.1 Newtonian Mechanics

The first comprehensive system of mechanics and dynamics, the ‘Philosophiæ Naturalis Principia Mathematica’ (Latin for "Mathematical Principles of Natural Philosophy"), written by I. Newton, first appeared 5 July 1687. The book describes the laws of motion, forming the foundation of classical mechanics, the law of universal gravitation and solidifies Kepler's laws of planetary motion. The Principia is "justly regarded as one of the most important works in the history of science".

The French mathematical physicist Alexis Clairaut assessed it in 1747: "The famous book of mathematical Principles of Natural Philosophy marked the epoch of a great revolution in physics ... spread the light of mathematics on a science, which up to then, remained in the darkness of conjectures and hypotheses." While acceptance of Newton's theories was not immediate, by the end of a century after its publication in 1687, "no one could deny that .... a science had emerged that, at least in certain respects, so far exceeded anything that had ever gone before that it stood alone as the ultimate exemplar of science generally." [232].
7.1.1 The Three Fundamental Laws of Newtonian Mechanics

Newtonian mechanics proceeds on just three postulates elevated to Natural laws:

1. A body remains in a state of rest or in uniform motion in a straight line unless a force acts on it.

2. If a body ‘m’ is acted on by an impressed force ‘F’, its acceleration ‘a’ (change of motion) is in the direction of the force i.e. \(|F\) = \(m|a\).

3. If a body exerts a force ‘|F_{12}\rangle’ on a second body, the second body exerts the same force ‘|F_{21}\rangle’ on the first body, but in the opposite direction i.e.

\(|F_{12}\rangle = -|F_{21}\rangle \rightarrow |F_{12}\rangle + |F_{21}\rangle = |0\rangle\)

7.1.2 The Tacit Assumptions of Newtonian Physics

However, Newton used notions not explicitly stated in the laws above. The first of these is the idea of a ‘stationary space’, where a point ‘P’ is specified by a position vector ‘|r\rangle’, giving its distance and direction from a fixed origin ‘0’. If \(|r\rangle\) is given by components \(\langle x, y, z \rangle\), called ‘coordinates’, in the directions of three chosen perpendicular axes, then

\(|r\rangle = x|i\rangle + y|j\rangle + z|k\rangle,\)

where \(x, y, z \in \mathbb{R}\) and \(|i\rangle, |j\rangle, |k\rangle\) are unit vectors in the direction of each axis respectfully. Every point in a stationary space can be located by specifying its coordinates \(\langle x, y, z \rangle\) [214].

7.1.2.1 Inertial Reference Frames

A reference frame is a choice of spatial origin, good for locating positions and for specifying a temporal origin to measure times. The frame allows observers to make theoretically measurements relative to it. Not all reference frames are equivalent. Newton’s laws hold only in inertial frames, ones that are stationary or moving with a constant velocity in accordance with Newton’s first law. If a second frame is accelerating or rotating relative to an inertial frame, it is ‘non-inertial’ frame [214].

7.1.2.2 The Time Parameter

Moreover, Newton’s mechanics treats time as a universal parameter ‘t’. All observers equipped with accurate clocks, properly synchronized, will agree on the time that an event occurs in space. Hence, the position vector ‘|r\rangle’ is generally given as a function of time, where, as time passes, the position vector draws out a space curve that describes the path of movement i.e.

\(|r(t)\rangle = x(t)|i\rangle + y(t)|j\rangle + z(t)|k\rangle,\)

where the spatial coordinates \(\langle x, y, z \rangle\) are functions of time.
7.1.2.1 Velocity and Acceleration Defined in terms of Time

The concept of a ‘velocity’ can be defined by differentiating $|r(t)\rangle$ with respect to time:

$$\frac{d|r\rangle}{dt} = \frac{dx}{dt}|i\rangle + \frac{dy}{dt}|j\rangle + \frac{dz}{dt}|k\rangle = |\dot{r}\rangle = |v\rangle$$

where $|v\rangle$ is the velocity of an object at time ‘$t$’. Mathematically, this is not quite correct. The correct approach requires use of the product rule. For instance,

$$\frac{d|r\rangle}{dt} = \frac{dx}{dt}|i\rangle + x\frac{d|i\rangle}{dt} + \frac{dy}{dt}|j\rangle + y\frac{d|j\rangle}{dt} + \frac{dz}{dt}|k\rangle + z\frac{d|k\rangle}{dt}$$

But, for example, since $|i\rangle$ is a constant, $d|i\rangle/dt = |0\rangle$. Therefore, all the second terms in each direction vanish so that

$$\frac{d|r\rangle}{dt} = \frac{dx}{dt}|i\rangle + \frac{dy}{dt}|j\rangle + \frac{dz}{dt}|k\rangle$$

If this equation is differentiated again with respected to time, then

$$\frac{d^2|r\rangle}{dt^2} = \frac{d^2x}{dt^2}|i\rangle + \frac{d^2y}{dt^2}|j\rangle + \frac{d^2z}{dt^2}|k\rangle = |\ddot{r}\rangle = |\dot{a}\rangle,$$

where $|\dot{a}\rangle$ is the ‘acceleration’ of the object at time ‘$t$’. Analytically, there is the tacit assumption that the space curve drawn by $|r\rangle$ is sufficiently smooth so that second order derivatives exist at all points along the curve. This is a theoretical assumption and is usually satisfied in most physical situations.

7.1.2.3 The Relationship between Force and Mass

Finally, there must be a way of describing the resistance to a directional or a speed change of a body. The notion here is that all bodies possess ‘mass’ ($m$) that resist a change in speed or direction. If $m$ is large, the resistance to change is large, if small, the resistance is small. Moreover, Newtonian mechanics treats all objects as though their mass is concentrated at a single point, its ‘center of mass’. Hence, Newton’s second law can be written

$$|F\rangle = m\frac{d^2|r\rangle}{dt^2} = m|\ddot{r}\rangle = m|\dot{a}\rangle$$

The body moves in the direction of the force, because the force is conceived as acting at its center of mass point.

If a force ‘$|F\rangle$’ does not change the speed or the direction of a body of mass ‘$m$’, then

$$|F\rangle = m|\dot{a}\rangle = |0\rangle,$$
and, by integration

\[|p⟩ = \int |F⟩ \, dt − |c⟩ = m \int \frac{d^2|r⟩}{dt^2} \, dt − |c⟩ = m|v⟩ − |c⟩ = m|\dot{r}⟩ − |c⟩ = 0 \rightarrow |p⟩ = m|v⟩\]

\[= m|\dot{r}⟩ = |c⟩, \quad c = a \text{ constant},\]

which is Newton’s first law – a body remains at rest or in uniform rectilinear motion unless a force is impressed upon it; the quantity ‘|p⟩’ is the ‘momentum’ of a body.

### 7.1.3 The Laws of Motion for a Constant Force

Newton’s second law ‘\(F = d^2|r⟩/dt^2\)’ is a differential equation, where an unknown function ‘\(r(t)\)’ solves the equation and involves the derivatives of \(r(t)\). For the motion along the \(x\)-axis, Newton’s second law can be written

\[|\ddot{x}⟩ = |F_x⟩/m,\]

where \(|F_x⟩\) is that part of a constant force directed along the \(x\)-axis. By integration,

\[\int |\ddot{x}⟩ \, dt = |\dot{x}⟩ = \frac{|F_x⟩}{m} t + v_x,\]

where \(v_x\) is a constant of integration. Moreover,

\[\int |\dot{x}⟩ \, dt = |x(t)⟩ = \frac{1}{2} \frac{|F_x⟩}{m} t^2 + v_x t + x_x,\]

where \(x_x\) is also a constant of integration. Equations for the \(y\)- and \(z\)-directions are derived similarly:

\[|y(t)⟩ = \frac{1}{2} \frac{|F_y⟩}{m} t^2 + v_y t + y_y\]

\[|z(t)⟩ = \frac{1}{2} \frac{|F_z⟩}{m} t^2 + v_z t + z_z\]

The solutions to these differential equations require specifying two constants of integration, 1) the ‘initial instantaneous velocity’ of the body, 2) the ‘initial position’ of the body. There are three equations of motion, one for each direction. Hence, a total of six constants of integration must be specified, two for each equation.

#### 7.1.3.1 Kinds of Acceleration

If \(|F⟩ = m|a⟩ \neq 0\), then either, 1) the body continues in a straight line, but accelerates along that line, 2) the velocity remains constant, but the body changes direction or 3) both the can happen at the same time. To handle this, acceleration is divided into two components, 1) ‘\(|a_T⟩\)’ is the ‘tangential acceleration’, which signifies the straight line changes in velocity; the other ‘\(|a_n⟩\)’, signifies the ‘normal acceleration’, acceleration due to a change in direction. To see this, note that
\[ |v\rangle = |r\rangle = \frac{d|r\rangle}{dt} = \frac{d|r\rangle}{ds} \frac{ds}{dt} = v|T\rangle, \]

In accordance with the theory of space curves, \( |d|r\rangle/ds | = 1 \), so \(|T\rangle\) is a unit vector in the direction tangent to the space curve. Now

\[ |a\rangle = \frac{d|v\rangle}{dt} |T\rangle + \frac{d|T\rangle}{ds} \frac{ds}{dt} \frac{d|v\rangle}{dt} |T\rangle + \frac{v^2}{\rho} |n\rangle = |a_T\rangle + |a_n\rangle \]

The vector \(|n\rangle\) is a unit vector pointing in the direction of the concave side of the space curve. Readers unfamiliar with how the second term in the above equation is derived should consult Chapter 1, Sec. 1.2.3; the quantity \( \rho \) is the radius of curvature of the space curve. Hence,

\[ ||a_T|| = \frac{d^2r}{dt^2}, \quad ||a_n|| = \frac{v^2}{\rho} \]

If the motion is in a straight line, only the tangential component is present. If \(|a_T\rangle = |0\rangle\), then the tangential velocity is a constant and there is only the normal component of the acceleration. This constitutes uniform circular motion (see Chapter 1, Sec. 1.2.3).

### 7.1.4 Other Inertial Reference Frames

While Cartesian coordinates are generally the simplest, there are problems more easily formulated by using various non-Cartesian coordinate systems, even though these coordinate systems are generally more complicated mathematically. For example, polar coordinates \( r, \phi \) can be written in terms of rectangular coordinates \( x, y \):

\[
\begin{align*}
x &= r \cos \phi & r &= \sqrt{x^2 + y^2} \\
y &= r \sin \phi & \phi &= \tan^{-1}(y/x)
\end{align*}
\]

Let \(|e_r\rangle\) be a unit vector pointing in the direction of increasing \( r \) with \( \phi \) fixed; let \(|e_\phi\rangle\) be the unit vector pointing in the direction of increasing \( \phi \) with \( r \) fixed. The vectors \(|i\rangle\) and \(|j\rangle\) are the same at all points in the plane, whereas the new vectors \(|e_r\rangle\) and \(|e_\phi\rangle\) change their directions as the position vector \(|r\rangle\) changes. Note that \(|e_r\rangle \perp |e_\phi\rangle\). To see this,

\[ |e_r\rangle = \cos \phi |i\rangle + \sin \phi |j\rangle \rightarrow \langle e_r| e_r\rangle = \cos^2 \phi + \sin^2 \phi = 1 \]
\[ |e_\phi\rangle = -\sin \phi |i\rangle + \cos \phi |j\rangle \rightarrow \langle e_\phi| e_\phi\rangle = \sin^2 \phi + \cos^2 \phi = 1 \]

Hence,

\[ \langle e_r| e_\phi\rangle = -\cos \phi \sin \phi + \cos \phi \sin \phi = 0 \rightarrow \langle e_r\rangle \perp |e_\phi\rangle \]

Since the derivative of a unit vector is perpendicular to the vector itself, then \( \langle de_r/dt | \parallel |e_\phi\rangle \) and \( \langle de_\phi/dt | \parallel |e_r\rangle \). Hence,
\[ \frac{d|e_r\rangle}{dt} = \frac{d\phi}{dt}|e_\phi\rangle, \quad \frac{d|e_\phi\rangle}{dt} = -\frac{d\phi}{dt}|e_r\rangle \]

To see the first of these,
\[ \frac{d|e_r\rangle}{dt} = \frac{d|e_r\rangle}{dr} \frac{dr}{dt} + \frac{d|e_r\rangle}{d\phi} \frac{d\phi}{dt} = \frac{dr}{dt}|0\rangle + \left(-\sin \phi |i\rangle + \cos \phi |j\rangle\right) \frac{d\phi}{dt} = \frac{d\phi}{dt}|e_\phi\rangle \]

Since, \(|r\rangle = r|e_r\rangle|0\rangle = r|e_r\rangle|e_\phi\rangle\),
\[ |\dot{r}\rangle = \dot{r}|e_r\rangle + r\dot{e}_r = \dot{r}|e_r\rangle + r\dot{\phi}|e_\phi\rangle, \quad \dot{\phi} = \frac{d\phi}{dt}, \]
by the product rule. Differentiating again with respect to \(t\) and combining terms leaves
\[ |\ddot{r}\rangle = \ddot{r}|e_r\rangle + \dot{r}\ddot{e}_r + r\dddot{\phi}|e_\phi\rangle + \dot{\phi}\dot{\phi}|e_\phi\rangle + \dddot{\phi}r|e_\phi\rangle = (\ddot{r} - r\dddot{\phi}^2)|e_r\rangle + (r\dddot{\phi} + 2\dddot{\phi})|e_\phi\rangle \]
Hence [226],
\[ \left\{ \begin{array}{l} a_r = (\ddot{r} - r\dddot{\phi}^2) \\
 a_\phi = (r\dddot{\phi} + 2\dddot{\phi}) \end{array} \right\} \rightarrow |F\rangle = m|a\rangle = \left\{ \begin{array}{l} |F_r\rangle = m|a_r\rangle = m(\ddot{r} - r\dddot{\phi}^2)|e_r\rangle \\
 |F_\phi\rangle = m|a_\phi\rangle = m(r\dddot{\phi} + 2\dddot{\phi})|e_\phi\rangle \right\} \]
which is Newton’s second law in polar coordinates.

### 7.1.5 Work, Power and Kinetic Energy

Recall that the work (Chapter 2, Sec. 2.2) done in moving a body an infinitesimal distance ‘\(dW\)’ is
\[ dW = \langle F|dr\rangle \rightarrow W = \int_{r_1}^{r_2} \langle F|dr\rangle, \quad d|r\rangle = dx|i\rangle + dy|j\rangle + dz|k\rangle \]
Moreover,
\[ P \ (\text{power}) = \frac{dW}{dt} = \langle F|\frac{dr}{dt}\rangle = \langle F|v\rangle \rightarrow W = \int_{t_1}^{t_2} \langle F|dr\rangle dt = \int_{t_1}^{t_2} \langle F|v\rangle dt = m \int_{t_1}^{t_2} \frac{dv}{dt} \langle v\rangle dt \]
\[ = m \int_{v_1}^{v_2} \langle v|dv\rangle = \frac{1}{2}mv^2 \bigg|_{v_1}^{v_2} = \frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2, \]
where the quantity ‘\(T = 1/2 \, mv^2\)’ is called the ‘kinetic energy’ of a body.

#### 7.1.5.1 Conservation of Momentum and Energy

If \(|F\rangle\) is a conserved force (Chapter 2, Sec. 2.4), then \(|F\rangle = -\nabla V \rightarrow |F\rangle + \nabla V = |0\rangle\). Hence,
\[ \int \langle F|dr\rangle + \int \langle \nabla V|dr\rangle = T + V = E, \quad E = a \text{ constant} \]
where $E$ is the ‘total energy’ of the system and $V$ is the ‘potential energy’. All conservative force fields have a ‘potential’. In a conservative force system, the total energy is a constant, which is called the law of the ‘conservation of energy’.

If $|F⟩ = |0⟩$, then

$$|F⟩ = m|a⟩ = \frac{d}{dt}(m|v⟩) = |0⟩ \rightarrow m|v⟩ = a \text{ constant}$$

If the net external force acting on a body is zero, its momentum remains invariant. This is called the ‘conservation of momentum’ and is a restatement of Newton’s first law.

Recall that the torque ‘$|Λ⟩$’ on a body is defined as

$$|Λ⟩ = ⟨r| × |F⟩ \rightarrow ⟨r| × |F⟩ = m\frac{d}{dt}(r| × |v⟩ \rightarrow ⟨r| × |p⟩ = |Ω⟩),$$

where $|Ω⟩$ is called the ‘angular momentum’ of a body.

If $|Λ⟩ = |0⟩$, then

$$\frac{d}{dt}(m(r| × |v⟩) = |0⟩ \rightarrow ⟨r| × |p⟩ = a \text{ constant}$$

The angular momentum of the body remains invariant and this is called the ‘conservation of angular momentum’.

**7.1.6 Newtonian Mechanics – A Summary**

This sums up Newtonian Mechanics in a nutshell. Using Newton’s three simple laws of motion, a host of physical and engineering problems can be solved. Even problems which involve non-conservative forces can be solved by adding additional constraints to the conservative equations. The number of problems to which Newtonian mechanics applies is vast and only a couple of pertinent examples can be discussed here. For additional details, the reader is referred to a textbook on theoretical mechanics, such as [214, 226, 227].

**7.2 Central Forces**

If a force acts on a body such that 1) the force is directed toward or away from a fixed point ‘$O$’, 2) its magnitude depends only on the radius vector ‘$|r⟩$’, then such a force is called a ‘central force’. Central forces have three important properties:

1. *The space curve describing the motion of the body must be a plane curve*
2. *The angular momentum of the body is conserved*
3. *The radius vector ‘$|r⟩$’ sweeps out equal areas in equal times*
To see 1, since the magnitude of the force depends only on the radius vector, all central forces can be written in the form \(|F| = \hat{f}(r)|e_r\rangle\), where \(|e_r\rangle\) is a unit vector in the direction of \(r\). Hence, \(\langle r| \times |F| = \hat{f}(r)(r| \times |e_r\rangle = |0\rangle\), since \(\langle r| \parallel |e_r\rangle\). Since,

\[
|F| = m \frac{d|v|}{dt} \rightarrow m \langle r| \times \frac{d|v|}{dt} \rangle = \frac{d}{dt} \langle r| \times |v\rangle = |0\rangle,
\]

upon integration,

\[
\langle r| \times |v\rangle = |h\rangle, \quad h = a \text{ constant}
\]

Moreover,

\[
\langle r|\langle r| \times |v\rangle\rangle = \langle \langle r| \times |r\rangle |v\rangle = \langle r|h\rangle = 0
\]

Therefore, \(\langle r| \perp |h\rangle\), showing that the motion is in a plane.

To see 2, note that \(\langle r| \times |v\rangle = |h\rangle \rightarrow m\langle r| \times |v\rangle = m|h\rangle\). Since \(m\langle r| \times |v\rangle\) is the angular momentum of the body and \(|h\rangle\) is a constant, the angular momentum is conserved in all central force fields [227].

### 7.2.1 The Laws of Central Forces Written in Polar Coordinates

Recall that Newton’s second law in polar coordinates is given by

\[
\left\{\begin{array}{l}
|F_r\rangle = m|a_r\rangle = m(\ddot{r} - r\dot{\phi}^2)|e_r\rangle \\
|F_\phi\rangle = m|a_\phi\rangle = m(r\ddot{\phi} + 2\dot{r}\dot{\phi})|e_\phi\rangle
\end{array}\right.
\]

The equations of motion for a central force field can then be written

\[
|F_r\rangle + |F_\phi\rangle = m(\ddot{r} - r\dot{\phi}^2)|e_r\rangle + m(r\ddot{\phi} + 2\dot{r}\dot{\phi})|e_\phi\rangle = f(r)|e_r\rangle,
\]

which implies that

\[
m(\ddot{r} - r\dot{\phi}^2) = f(r), \quad m(r\ddot{\phi} + 2\dot{r}\dot{\phi}) = 0
\]

Since

\[
m(r\ddot{\phi} + 2\dot{r}\dot{\phi}) = \frac{m}{r} (r^2\ddot{\phi} + 2\dot{r}\dot{\phi}) = \frac{m}{r} \frac{d}{dt} (r^2\ddot{\phi}) = 0 \rightarrow r^2\ddot{\phi} = h, \quad h = a \text{ constant}
\]

Suppose that in time \(\Delta t\), the body moves from point \(P_1\) to point \(P_2\). If \(\Delta t\) is small, the curve forms one side of a triangle whose area \(\Delta A\) is \(1/2 |\langle r| \times |\Delta r\rangle||\). Now

\[
\frac{\Delta A}{\Delta t} = 1/2 |\langle r| \times |v\rangle| \rightarrow \lim_{\Delta t \to 0} \left(\frac{\Delta A}{\Delta t}\right) = \dot{A} = 1/2 |\langle r| \times \left|\frac{dr}{dt}\right|\|
\]

But, \(|\langle r| \times |v\rangle| = h\), a constant. Hence, \(\dot{A} = 1/2 |h|\). This proves 3 above - the radius vector \(|r|\) sweeps out equal areas in equal times.
Since a central force is a function only of position, it is a conservative force, and hence, has a potential i.e. $V = -\int f(r) \, dr$. Since

$$m(\ddot{r} - r\dot{\phi}^2) = f(r) \rightarrow m(\ddot{r} - \dot{r} r\dot{\phi}^2) = f(r)\dot{r}$$

$$m(r^2 \dot{\phi} \ddot{\phi} + 2\dot{\phi}^2 \dot{r}) = 0$$

Adding the two equations above leaves

$$m(\ddot{r} + r^2 \dot{\phi} \ddot{\phi} + \ddot{r} r \dot{\phi}^2) = f(r)\dot{r} = \frac{1}{2} m \frac{d}{dt} \left( \dot{r}^2 + r^2 \dot{\phi}^2 \right) = \frac{d}{dt} \int f(r) \, dr$$

Integrating gives

$$\frac{1}{2} m(\dot{r}^2 + r^2 \dot{\phi}^2) - \int f(r) \, dr = T + V = E, \quad E = a \text{ constant},$$

which is the law of the conservation of energy equation for central forces.

### 7.2.2 Laws of Planetary Motion

By the definition of a conic section, given any point ‘$P$’ on the conic, $r/d = \epsilon$, $\epsilon = a \text{ constant}$. For the point ‘$Q$’, $p/D = \epsilon$ (see fig. 7.2.2-1). Now $D = d + r \cos \theta$.

Hence,

$$p = \epsilon D = \epsilon (d + r \cos \theta) \rightarrow D = \frac{r}{\epsilon} + r \cos \theta = \frac{r}{\epsilon} \left( 1 + \epsilon \cos \theta \right) \rightarrow r = \frac{p}{1 + \epsilon \cos \theta}$$

The equation on the far right is a circle if $\epsilon = 0$, an ellipse if $0 < \epsilon < 1$, a parabola if $\epsilon = 1$ and a hyperbola if $\epsilon > 1$ [227].

Recall that

$$r^2 \ddot{\phi} = h \rightarrow \dot{\phi} = h/r^2, \quad h = a \text{ constant}$$
Letting $u = 1/r \rightarrow \phi = hu^2$ and since $m(\ddot{r} - r\dot{\phi}^2) = f(r)$, then

$$m(\ddot{r} - h^2/r^3) = f(r),$$

since $r\dot{\phi}^2 = h^2/r^3$.

Note that

$$\dot{r} = \frac{dr}{dt} = \frac{dr}{d\phi} \frac{d\phi}{dt} = \frac{dr}{d\phi} \frac{h}{r^2} = -h \frac{du}{d\phi},$$

since $r = 1/u$, then

$$\frac{dr}{d\phi} = -\frac{1}{u^2} \frac{du}{d\phi} = -r^2 \frac{du}{d\phi}$$

And

$$\ddot{r} = \frac{d\dot{r}}{dt} = \frac{d}{dt} \left( -h \frac{du}{d\phi} \right) = \frac{d}{d\phi} \left( -h \frac{du}{d\phi} \right) \frac{d\phi}{dt} = -h^2 u^2 \frac{d^2u}{d\phi^2}$$

Recalling that

$$m\left(\ddot{r} - \frac{h^2}{r^3}\right) = f(r) \rightarrow m\left(-h^2 u^2 \frac{d^2u}{d\phi^2} - h^2 u^3\right) = f(1/u),$$

then

$$\frac{d^2u}{d\phi^2} + u = -\frac{f(1/u)}{mh^2 u^2},$$

which is a differential equation in the variable ‘$u$’.

### 7.2.2.1 Kepler’s Three Laws of Planetary Motion

Using the voluminous data collected by the Danish nobleman Tyge Ottesen Brahe (14 December 1546 – 24 October 1601), the German mathematician, astronomer and astrologer Johannes Kepler (December 27, 1571 – November 15, 1630) formulated his three laws of planetary motion:

1. **Every planet moves in an elliptical orbit with the Sun at one focus**

2. **The radius vector drawn from the Sun to any planet sweeps out equal areas in equal times**

3. **The squares of the periods of revolution of the planets are proportional to the cubes of the semi-major axes of their orbits**
Since
\[ r = \frac{p}{1 + \epsilon \cos \phi} \rightarrow u = \frac{1}{r} = \frac{1}{p} + \frac{\epsilon}{p} \cos \phi, \]
where \( \epsilon < 1 \), this is a statement of Kepler’s first law. Every planet moves in an elliptical orbit with the Sun at one focus.

Hence,
\[ \frac{d^2u}{d\phi^2} + u = -\frac{f(1/u)}{mh^2u^2} \rightarrow -mh^2u^2 \left( \frac{d^2u}{d\phi^2} + u \right) = f(1/u) = -\frac{mh^2u^2}{p} \]

The last equation on the right is obtained by substituting the value for \( u \) i.e. \( u = 1/p + \epsilon/p \cos \theta \) into the equation above and working through the algebra. Upon replacing \( u \) by \( 1/r \),
\[ f(r) = -\frac{mh^2}{pr^2} \]

Letting \( GM = h^2/p \), then
\[ f(r) = -\frac{GmM}{r^2} \]

The equation above is Newton’s universal law of gravitation, which states that any two bodies attract each other with a force proportional to the square of the distance between them. Newton concluded that if \( r = R \), the radius of the Earth, then objects close to the Earth’s surface accelerate by a constant ‘\( g \)’ i.e. \( F = mg \), where \( m \) is the mass of the object and \( M \) is the mass of the Earth. Therefore,
\[ f(r) = -\frac{GmM}{R^2} = mg \rightarrow g = -\frac{GM}{R^2} \]

### 7.3 The Simple Harmonic Oscillator

In figure 7.3-1 a body of mass ‘\( m \)’ is connected to a spring and its center of gravity is pulled to the position ‘\( x_m \)’ and released. The body will perform harmonic motion. If friction and the mass of the spring are insubstantial, the body will exhibit ‘simple harmonic motion’, which complies with Hooke’s law, named for the English natural philosopher, architect and polymath Robert Hooke (18 July 1635 – 3 March 1703).
7.3.1 Hooke’s Law

Hooke’s law states that the force ‘$F_x$’ needed to extend or compress a spring by a distance ‘$x$’ is proportional to that distance i.e. $|F_x| = -kx|i|$, where $k > 0$ is a constant factor, which depends on the stiffness of the spring. The point ‘$x_0$’ is called the ‘point of equilibrium’ about which the body will oscillate symmetrically.

Hence,

$$|F_x| = -kx|i| = m \frac{d^2x|i|}{dt^2} = -kx|i| \to \ddot{x} + \frac{k}{m} x = 0$$

The equation on the right above is a differential equation with general solution

$$x(t) = A \cos(\omega t) + B \sin(\omega t), \quad \omega = \sqrt{k/m},$$

where $A$ and $B$ are constants determined by the initial conditions of the problem. For instance, if $x = A$, $dx/dt = 0$ at $t = 0$ are the initial conditions, then $B = 0$, and $x(t) = A \cos \omega t$ becomes the particular solution (see Chapter 1, Sec. 1.3).

7.3.2 Simple Harmonic Motion as a Conservative System

Note that

$$\langle \nabla \times |F_x| \rangle = \begin{vmatrix}
  |i| & |j| & |k| \\
  \partial & \partial & \partial \\
  \partial x & \partial y & \partial z \\
  -\frac{k}{m} x|i| & 0 & 0 \\
\end{vmatrix} = |0|,$$

which shows that a simple harmonic oscillator is a conservative system, and hence, has a potential ‘$V$’. To see this, note that

$$m\ddot{x} + kx = 0 \to m \frac{dv}{dt} = -kx \to m \frac{dv}{dx} \frac{dx}{dt} = -kx \to m v \frac{dv}{dx} + kx = 0$$
Upon integration by $x$

$$m \int v \frac{dv}{dx} dx + k \int x \, dx = \frac{1}{2} m v^2 + \frac{1}{2} k x^2 = T + V = E, \quad E = a \text{ constant},$$

where $V = k x^2/2$.

If friction ‘$F_D$’ is included, say $F_D = -\beta \frac{dx}{dt} |i|$, $\beta = a \text{ constant} > 0$, the mass will not slide smoothly across the surface. The system exhausts energy through heat loss and the friction will act like a damping force, causing the oscillator to slow down and eventually stop oscillating. In this case, the system is no longer conservative. Nevertheless, the non-conservative motion can be modeled by the equation

$$\ddot{x} + 2\gamma \dot{x} + \omega^2 x = 0,$$

where $2\gamma = \beta / m$ and $\omega^2 = k / m$. The equation above is a differential equation with general solution

$$x(t) = e^{-\gamma t} (A e^{\alpha t} + B e^{-\alpha t}), \quad \alpha = \sqrt{\gamma^2 - \omega^2},$$

where $A$ and $B$ are constants determined by the initial conditions.

Simple harmonic motion can be modeled in more than one dimension. Let

$$|F| = -k_1 x|i| - k_2 y|j| - k_3 z|k|$$

Since $|F|$ is linear, the general solutions become

$$x(t) = A_1 \cos \left( \sqrt{\frac{k_1}{m}} t \right) + B_1 \sin \left( \sqrt{\frac{k_1}{m}} t \right), \quad y(t) = A_2 \cos \left( \sqrt{\frac{k_2}{m}} t \right) + B_2 \sin \left( \sqrt{\frac{k_2}{m}} t \right),$$

$$z(t) = A_3 \cos \left( \sqrt{\frac{k_3}{m}} t \right) + B_3 \sin \left( \sqrt{\frac{k_3}{m}} t \right),$$

where the $A_i$’s and the $B_i$’s are constants.

7.4 Systems of Bodies

Thus far, the discussion has revolved around applying Newton’s laws to a point mass - a body treated as though all its mass is concentrated at a point, where there is no need to consider the forces between bodies.

7.4.1 Newton’s Second Law for a System of Bodies

A system of $N$ bodies will be subject to two kinds of forces 1) ‘external forces’ i.e. $f_1 + f_2 + \cdots + f_N = F_k$, interpreted as the resultant external force on the $k^{th}$ body and 2) ‘internal forces’, the forces ‘$F_{ij}$’ that represent the force that the $j^{th}$ body exerts on the $i^{th}$ body. If these two kinds of forces are added up for all $N$ bodies, then
\[ \sum_{k} m_k \frac{d^2 |r_k\rangle}{dt^2} = \sum_{k} |F_k\rangle + \sum_{k} \sum_{i} |F_{ik}\rangle, \quad i, k = 1, \ldots, N, \]

where \(|r_k\rangle\) is the radius vector for the \(k^{th}\) body. According to Newton’s third law, \(|F_{ik}\rangle = -|F_{ki}\rangle \rightarrow |F_{ik}\rangle + |F_{ki}\rangle = |0\rangle\). But, in the double sum on the right side of the equation above, there are just as many \(F_{ik}\)'s as \(F_{ki}\)'s. Moreover, \(\forall i \in N \rightarrow |F_{ii}\rangle = |0\rangle\), since a body does not apply an internal force on itself. Hence, the double sum on the right side of the equation above vanishes i.e.

\[ \sum_{k} \sum_{i} |F_{ik}\rangle = |0\rangle \rightarrow \sum_{k} m_k \frac{d^2 |r_k\rangle}{dt^2} = \sum_{k} |F_k\rangle \]

If \(|r_1\rangle + |r_2\rangle + \cdots + |r_N\rangle\) are the positions of the \(N\) bodies, let

\[ |\bar{r}\rangle = \frac{m_1 |r_1\rangle + m_2 |r_2\rangle + \cdots + m_N |r_N\rangle}{m_1 + m_2 + \cdots + m_N} = \frac{\sum_{i=1}^{N} m_i |r_i\rangle}{\sum_{i=1}^{N} m_i} = \frac{\sum_{i=1}^{N} m_i |r_i\rangle}{M}, \]

where \(M = \sum_{i=1}^{N} m_i\) is called the ‘total mass’ of the system and is a constant; the vector quantity ‘\(|\bar{r}\rangle\)’ is the position of the ‘center of mass’ of the system. Therefore,

\[ \sum_{k} m_k \frac{d^2 |r_k\rangle}{dt^2} = \sum_{k} |F_k\rangle \rightarrow M \frac{d^2 |\bar{r}\rangle}{dt^2} = \sum_{k} |F_k\rangle, \]

which leads to the following simple result:

*The center of mass of a system of bodies moves as though all the mass of the system resides at the center of mass and with the resultant of the applied external forces acting at that point.*

Hence, Newton’s second law for a system of particles is identical to that of a single body. All the mass is located at the center of mass of the system and that the external forces act there – a very fortunate result.

As an aside, there are times when a body can be considered smooth, where the methods of the calculus can be employed to find the center and total mass of a system. In this case, it is convenient to define a density function ‘\(\sigma\)’ such that

\[ \sigma = \lim_{\tau \to 0} \frac{\Delta M}{\Delta \tau} = \frac{dM}{d\tau} \]

where \(d\tau\) specifies an element of volume. Hence,

\[ |\bar{r}\rangle = \frac{\int_{\mathbb{R}} \sigma |r\rangle \, dr}{\int_{\mathbb{R}} \sigma \, d\tau}, \quad M = \int_{\mathbb{R}} \sigma \, d\tau, \]
where the integral is taken over the entire region ‘$\mathcal{R}$’. In practice, it is a simple matter to replace the discrete case with the continuous case by merely substituting integrals for summations.

### 7.4.2 Newton’s First Law for a System of Bodies

If $\sum_k |F_k\rangle = |0\rangle$, in other words, if no external forces act on the system, then

$$
M \frac{d^2 |\vec{r}\rangle}{dt^2} = |0\rangle \rightarrow M \int \frac{d^2 |\vec{r}\rangle}{dt^2} dt = M \frac{d |\vec{r}\rangle}{dt} = M |\vec{v}\rangle = |\vec{p}\rangle = a \text{ constant},
$$

where $|\vec{v}\rangle = |\vec{\dot{r}}\rangle$ is the ‘velocity’ and $|\vec{p}\rangle$ is the ‘momentum’ of the center of mass. This is Newton’s first law for a system of bodies, which is identical to his first law for a single body.

### 7.4.3 Angular Momentum and Torque of a System of Bodies

The quantity

$$
|\Omega\rangle = \sum_{\nu=1}^{N} m_\nu \langle r_\nu | \times |\dot{r}_\nu\rangle
$$

is called the ‘total angular momentum’ of the system about an origin ‘$O$’, and

$$
|\Lambda\rangle = \sum_{\nu=1}^{N} \langle r_\nu | \times |F_\nu\rangle
$$

is called the ‘torque’ on the system of bodies. By Newton’s second law, the force on the $\nu^{th}$ body is

$$
\frac{d |p_\nu\rangle}{dt} = F_\nu + \sum_i |F_{i\nu}\rangle, \quad i = 1, \ldots, N,
$$

Multiplying both sides of the equation above by $\langle r_\nu | \times$, then

$$
\langle r_\nu | \times \frac{d |p_\nu\rangle}{dt} = \langle r_\nu | \times |F_\nu\rangle + \sum_i \langle r_\nu | \times |F_{i\nu}\rangle = \frac{d}{dt} (m_\nu \langle r_\nu | \times |v_\nu\rangle)
$$

Summing on $\nu$ to pick up all the bodies gives

$$
\sum_\nu \langle r_\nu | \times |F_\nu\rangle + \sum_\nu \sum_i \langle r_\nu | \times |F_{i\nu}\rangle = \frac{d}{dt} \left( \sum_\nu m_\nu \langle r_\nu | \times |v_\nu\rangle \right), \quad \nu, i = 1, \ldots, N
$$

The double sum has terms like $\langle r_\nu | \times |F_{i\nu}\rangle + \langle r_i | \times |F_{\nu i}\rangle$. Applying Newton’s third law

$$
|F_{i\nu}\rangle = -|F_{\nu i}\rangle \rightarrow \langle r_\nu | \times |F_{i\nu}\rangle - \langle r_i | \times |F_{\nu i}\rangle = \langle r_\nu - r_i | \times |F_{\nu i}\rangle
$$
If the internal forces are central forces, which is most often the case, then the internal forces are in the same direction as their radius vectors and the double sum vanishes i.e. \( \langle r_v - r_i \rangle \times \langle F_{vi} \rangle = 0 \) because \( \langle r_v - r_i \parallel \langle F_{vi} \rangle \). Hence,

\[
\sum_v \langle r_v | F_v \rangle = \frac{d}{dt} \left( \sum_v m_v \langle r_v | v_v \rangle \right) \rightarrow |\Lambda \rangle = \frac{d|\Omega \rangle}{dt},
\]

which is the same result as with a single body. This result leads to the following:

*The total external torque on a system of particles is equal to the time rate of change of the angular momentum of the system, provided the internal forces between particles are central forces.*

Hence, if \( |\Lambda \rangle = |0\rangle \), \( |\Omega \rangle = a \text{ constant} \) and this is the law of the ‘conservation of angular momentum’ for a system of bodies.

7.4.4 The Energy of a System of Bodies

Recall that the equation of motion for a system of bodies is

\[
\frac{d}{dt} |p_v \rangle = F_v \text{ (External force)} + \sum_i |F_{iv} \rangle \text{ (Internal forces)}, \quad i = 1, ..., N,
\]

Taking the inner product of both sides with \( |\dot{r}_v \rangle \), then

\[
\frac{d}{dt} \langle \dot{r}_v | m_v \dot{r}_v \rangle = \langle \dot{r}_v | F_v \rangle + \langle \dot{r}_v | \sum_i |F_{iv} \rangle \rightarrow \frac{d}{dt} \langle \dot{r}_v | m_v \dot{r}_v \rangle = \frac{1}{2} \frac{d}{dt} (m_v \dot{r}_v^2) = \langle \dot{r}_v | F_v \rangle + \langle \dot{r}_v | \sum_i |F_{iv} \rangle
\]

Summing over \( v \) to include all the bodies gives

\[
\frac{1}{2} \frac{d}{dt} \left( \sum_v m_v \dot{r}_v^2 \right) = \sum_v \langle \dot{r}_v | F_v \rangle + \sum_v \sum_i \langle \dot{r}_v | F_{iv} \rangle
\]

Integrating both sides with respect to \( t \) gives

\[
W = \frac{1}{2} \sum_v \int_{t_1}^{t_2} \frac{d}{dt} (m_v \dot{r}_v^2) \ dt = \sum_v \int_{t_1}^{t_2} \langle \dot{r}_v | F_v \rangle \ dt + \sum_v \sum_i \int_{t_1}^{t_2} \langle \dot{r}_v | F_{iv} \rangle \ dt
\]

Since \( \dot{r}_v \ dt = dr_v \), then

\[
\frac{1}{2} \sum_v m_v \dot{r}_v^2 \bigg|_{t_1}^{t_2} = \sum_v \int_{r_1(t_1)}^{r_2(t_2)} \langle F_v | dr_v \rangle \quad \text{(External forces)} + \sum_v \sum_i \int_{r_1(t_1)}^{r_2(t_2)} \langle F_{iv} | dr_v \rangle = T_{t_2} - T_{t_1},
\]
which is the difference in the kinetic energy of the system of particles at \( t_1 \) and \( t_2 \) or the total work done on the system. Note that the double sum, indicating the work done by the internal forces, cannot be reduced to zero either by using Newton’s third law or by assuming central forces [227].

7.4.4.1 The Kinetic Energy of a System of Bodies

The kinetic energy of a system of bodies can be divided into two parts. Let \( 0' \) be located in a second coordinate system, where \( 0' \) is at the center of gravity of the system i.e. \( |r_v⟩ = |\vec{r}⟩ + |r'_v⟩ \). Then

\[
\frac{1}{2} \sum_v m_v |\dot{r}_v⟩^2 = \frac{1}{2} |\dot{\vec{r}}⟩^2 \sum_v m_v + |\dot{r}_v⟩^2 \sum_v m_v + \frac{1}{2} \sum_v m_v |\dot{r}'_v⟩^2
\]

By hypothesis, \( \sum_v m_v |r_v⟩ / M = |0⟩ \), since it is the radius vector of the center of mass and this, by hypothesis, is zero in the primed coordinates. Hence, the middle term on the right side of the equation above vanishes. So,

\[
\frac{1}{2} \sum_v m_v |\dot{r}_v⟩^2 = \frac{1}{2} |\dot{\vec{r}}⟩^2 \sum_v m_v + \frac{1}{2} \sum_v m_v |\dot{r}'_v⟩^2
\]

The following interpretation can be given to the result above [226]:

*The total kinetic energy is equal to the translational kinetic energy of the entire mass, considered concentrated at the center of gravity, plus the energy of motion of the parts of the system relative to the center of gravity.*

7.4.4.2 Potential Energy and the Conservation of Energy of a System of Bodies

If the internal forces of a system of bodies are conservative forces, let \( V_{iv}(r_{iv}) \) be the system potential, where

\[
r_{iv} = \sqrt{(x_i - x_v)^2 + (y_i - y_v)^2 + (z_i - z_v)^2}
\]

is the distance between the bodies ‘\( i \)’ and ‘\( v \)’. Then the force acting on body ‘\( v \)’ is

\[
|f_{iv}⟩ = - \frac{\partial V_{iv}}{\partial x_v} |i⟩ - \frac{\partial V_{iv}}{\partial y_v} |j⟩ - \frac{\partial V_{iv}}{\partial z_v} |k⟩ = -\nabla_v V_{iv}
\]

The force acting on body ‘\( i \)’ is

\[
|f_{iv}⟩ = - \frac{\partial V_{iv}}{\partial x_i} |i⟩ - \frac{\partial V_{iv}}{\partial y_i} |j⟩ - \frac{\partial V_{iv}}{\partial z_i} |k⟩ = -\nabla_i V_{iv} = -|f_{vi}⟩,
\]

by Newton’s third law. The work done by these forces, producing the displacements ‘\( dr_v \)’ and ‘\( dr_i \)’ respectfully, is
\begin{align*}
\langle f_v | d r_v \rangle + \langle f_i | d r_i \rangle &= - \left( \frac{\partial V_{iv}}{\partial x_v} d r_v + \frac{\partial V_{iv}}{\partial y_v} d r_v + \frac{\partial V_{iv}}{\partial z_v} d r_v + \frac{\partial V_{iv}}{\partial x_i} d r_i + \frac{\partial V_{iv}}{\partial y_i} d r_i + \frac{\partial V_{iv}}{\partial z_i} d r_i \right) \\
&= - d V_{vi}
\end{align*}

The total work done by all the bodies is

\[ \sum_v \sum_i \langle f_v | d r_v \rangle = - \frac{1}{2} \sum_v \sum_i d V_{vi}, \quad v \neq i \]

the factor ‘1/2’ entering on the right because, otherwise, the terms in the summation would be counted twice. Hence,

\[ \sum_v \sum_i \int_{r_1}^{r_2} \langle f_v | d r_v \rangle = - \frac{1}{2} \sum_v \sum_i \int_{r_1}^{r_2} d V_{iv} = V_2^{int} - V_1^{int}, \]

where \( V_1^{int} \) and \( V_2^{int} \) signify the internal potentials.

If both the external and internal forces for a system are conservative, let \( V_1 = V_1^{ext} + V_1^{int} \) and \( V_2 = V_2^{ext} + V_2^{int} \), then

\[ T_2 - T_1 = V_1 - V_2 = V_1^{ext} - V_2^{ext} + V_1^{int} - V_2^{int} \rightarrow T_2 + V_2 = T_1 + V_1 \]

or \( T + V = \text{a constant} \), which is the principle of the conservation of energy for a system of bodies.

**7.5 Mechanics in Non-Inertial Reference Frames**

Newton’s laws are valid only in inertial reference frames — frames that are neither accelerating nor rotating. To see this, suppose ‘0′ signifies the origin of a coordinate system ‘\( S_0 \)’ at rest and ‘\( 0′ \)’ signifies the origin of a coordinate system \( S_{0′} \) in motion relative to \( S_0 \). Let \( |r⟩ \) be the position vector as measured from 0 and \( |r′⟩ \) be the position vector as measured from ‘\( 0′ \). Then Newton’s second law, given in the coordinate systems ‘\( S_0′ \)’ and ‘\( S_{0′}′ \)’ respectfully is

\[ |F⟩ = m \frac{d^2 |r⟩}{dt^2}, \quad |F′⟩ = m \frac{d^2 |r′⟩}{dt^2} \]

Hence,

\[ |F⟩ - |F′⟩ = m \frac{d^2 |r⟩}{dt^2} - m \frac{d^2 |r′⟩}{dt^2} = m \frac{d^2}{dt^2} (|r⟩ - |r′⟩) \]

Therefore,

\[ m \frac{d^2}{dt^2} (|r⟩ - |r′⟩) = |0⟩ \rightarrow m \frac{d}{dt} (|r⟩ - |r′⟩) = \text{a constant} \]

Newton’s second law holds if and only if the two reference frames are at rest or moving with a constant velocity relative to one another.
In some situations, it is desirable to consider a non-inertial frame. For example, to an excellent approximation, the Earth rotating on its axis can be considered an inertial frame, but, technically, a reference frame fixed to the Earth rotates. While most problems can ignore the fact that someone standing on the Earth is in a non-inertial frame, there are situations, for instance, firing of a long-range rocket, where the earth’s rotation has important consequences and other situations, where the problem is easier to formulate in a non-inertial frame [214].

7.5.1 Newton’s Second Law in Non-Inertial Reference Frames

Consider an inertial frame ‘$S_0$’ and a second frame ‘$S$’ that is accelerating relative to $S_0$ with acceleration ‘$|A\rangle$’, not necessarily constant, but without a change in direction. In the reference frame ‘$S_0$’, Newton’s second law reads $|F\rangle = m|i_0\rangle$. The velocity of a body measured in $S_0$ is $|i_0\rangle$. The velocity of a body measured in $S$ is $|\dot{r}\rangle + |V\rangle$, where $|V\rangle$ is the velocity of $S$ relative to $S_0$. By the law of Galilean relativity, $i_0 = \dot{r} + |V\rangle$ i.e.

$$
(body \ velocity \ relative \ to \ S_0) = (body \ velocity \ relative \ to \ S) + (S's \ velocity \ relative \ to \ S_0)
$$

So,

$$
\frac{d}{dt} |i_0\rangle = \frac{d}{dt} |\dot{r}\rangle + \frac{d}{dt} |V\rangle = |i_0\rangle = |\dot{r}\rangle + |A\rangle \rightarrow m|\dot{r}\rangle = |F\rangle - m|A\rangle
$$

Newton’s second law in the non-inertial frame ‘$S$’ is valid provided the extra force term ‘$-m|A\rangle$’ is added to the force in the inertial frame [214].

For example, consider a simple pendulum (mass ‘$m$’ and length ‘$L$’) mounted inside a railroad car that is accelerating to the right with constant acceleration ‘$|A\rangle$’. There are two forces on the bob of the pendulum, the tension ‘$|T\rangle$’ in the string and the weight ‘$m|g\rangle$’; thus, the net force in the inertial frame is $|F\rangle = |T\rangle + m|g\rangle$. In the non-inertial frame ‘$S$’

$$
F_S = m|\ddot{r}\rangle = |T\rangle + m|g\rangle - m|A\rangle = |T\rangle + m(|g\rangle - |A\rangle) = |T\rangle + m|g_S\rangle,
$$

where $|g_S\rangle = |g\rangle - |A\rangle$ [214].

Since

$$
m|\ddot{r}\rangle = |F\rangle - m|A\rangle \rightarrow |F\rangle = m|\ddot{r}\rangle + m|A\rangle,
$$
in an inertial frame, anchored to the ground, only two forces act on the bob ($|T\rangle + m|g\rangle$). If the pendulum is to remain at rest in the railroad car, then (as seen from the ground) it must accelerate by $|A\rangle$. Therefore, the net force $|T\rangle + m|g\rangle = m|A\rangle$. Suppose that $|A\rangle = |g\rangle$, then the inertial force exactly balances the gravitational force ‘$-mg$’ i.e. $|F\rangle = |0\rangle$. An observer inside a box accelerating toward Earth with a constant acceleration can perform no experiment that would detect that acceleration. However, if the box is placed where no gravitational field exists, but instead, moves upward with acceleration ‘$+g|k\rangle$’, then the observer inside the box would claim to be in the Earth’s gravitational
field. In order for this statement to be true, gravitational mass must equal inertial mass. This fact is at the heart of the theory of general relativity.

### 7.5.1.1 Newton’s Second Law in a Rotating Frame

Now suppose a moving coordinate system ‘M’ is rotating around a point ‘0’ relative to a fixed system ‘F’. Given a vector ‘\( |A\rangle = A_1|i\rangle + A_2|j\rangle + A_3|k\rangle \)’, its time rate of change relative to \( M \) is

\[
\frac{d|A\rangle}{dt} = \frac{dA_1}{dt}|i\rangle + \frac{dA_2}{dt}|j\rangle + \frac{dA_3}{dt}|k\rangle
\]

What is the time rate of change of the vector ‘\( |A\rangle \)’ relative to \( F \)? To an observer in \( F \), the unit vectors ‘\(|i\rangle, |j\rangle, |k\rangle\)’ change with time i.e.

\[
\frac{d|A\rangle}{dt} = \frac{d|A\rangle}{dt} + A_1 \frac{d|i\rangle}{dt} + A_2 \frac{d|j\rangle}{dt} + A_3 \frac{d|k\rangle}{dt}
\]

Since \( \langle di/dt | \perp i \rangle \), then \( d|i\rangle/dt \) must lie in \( |j\rangle, |k\rangle \)-plane i.e.

\[
d|i\rangle/dt = \alpha_1|j\rangle + \alpha_2|k\rangle
\]

Similarly,

\[
d|j\rangle/dt = \alpha_3|k\rangle + \alpha_4|i\rangle, \quad d|k\rangle/dt = \alpha_5|i\rangle + \alpha_6|j\rangle
\]

Since \( \langle i|j \rangle = 0 \), by differentiation

\[
\langle i | \frac{dj}{dt} \rangle + \langle di/dt | j \rangle = 0 \quad \Rightarrow \langle i | \frac{dj}{dt} \rangle = \alpha_4, \quad \langle di/dt | j \rangle = \alpha_1 \rightarrow \alpha_4 = -\alpha_1
\]

Similarly,

\[
\langle j | \frac{dk}{dt} \rangle + \langle dj/dt | k \rangle = 0 \quad \Rightarrow \alpha_6 = -\alpha_3, \quad \langle i | \frac{dk}{dt} \rangle + \langle di/dt | k \rangle = 0 \quad \Rightarrow \alpha_5 = -\alpha_2
\]

Then

\[
d|i\rangle/dt = \alpha_1|j\rangle + \alpha_2|k\rangle, \quad d|j\rangle/dt = \alpha_3|k\rangle - \alpha_1|i\rangle, \quad d|k\rangle/dt = -\alpha_2|i\rangle - \alpha_3|j\rangle
\]

Hence,

\[
A_1d|i\rangle/dt = A_1\alpha_1|j\rangle + A_1\alpha_2|k\rangle, \quad A_2d|j\rangle/dt = A_2\alpha_3|k\rangle - A_2\alpha_1|i\rangle, \quad A_3d|k\rangle/dt = -A_3\alpha_2|i\rangle - A_3\alpha_3|j\rangle
\]

Therefore,

\[
A_1 \frac{d|i\rangle}{dt} + A_2 \frac{d|j\rangle}{dt} + A_3 \frac{d|k\rangle}{dt} = (-\alpha_1A_2 - \alpha_2A_3)|i\rangle + (\alpha_1A_1 - \alpha_3A_3)|j\rangle + (\alpha_2A_1 + \alpha_3A_2)|k\rangle
\]
which can be written
\[
A_1 \frac{d|i\rangle}{dt} + A_2 \frac{d|j\rangle}{dt} + A_3 \frac{d|k\rangle}{dt} = \begin{vmatrix} |i\rangle & |j\rangle & |k\rangle \\ \alpha_3 & -\alpha_2 & \alpha_1 \\ A_1 & A_2 & A_3 \end{vmatrix} = \langle \omega | \times |A\rangle
\]

where \( \omega_1 = \alpha_3, \ \omega_2 = -\alpha_2, \ \omega_3 = \alpha_1 \). Hence,
\[
\frac{d|A\rangle}{dt} = \frac{d|A\rangle}{dt} + \langle \omega | \times |A\rangle,
\]

where \( \omega = \omega_1|i\rangle + \omega_2|j\rangle + \omega_3|k\rangle \).

The equation above describes the rate of change of the vector \(|A\rangle\) with respect to the fixed frame of reference.

Differentiating, once again with respect to \(t\),
\[
\frac{d^2|A\rangle}{dt^2} = \frac{d^2|A\rangle}{dt^2} + \langle d\omega | \times |A\rangle + \langle 2\omega | \times |\frac{dA}{dt}\rangle + \langle \omega | \times \langle \omega | \times |A\rangle\rangle,
\]

where the terms on the right reference the rotating coordinate system [227]. If \(|A\rangle\) is replaced by \(|r\rangle\) in the equation above, then
\[
\frac{m \ d^2|r\rangle}{dt^2} = \frac{m \ d^2|r\rangle}{dt^2} + m \langle d\omega | \times |r\rangle + m \langle 2\omega | \times |\frac{dr}{dt}\rangle + m \langle \omega | \times \langle \omega | \times |r\rangle\rangle
\]
gives Newton’s second law in a rotating system relative to the fixed system of coordinates.

### 7.5.1.1.1 Coriolis and Centrifugal Force

The last two terms on the right side of this equation
\[
m \langle 2\omega | \times |\frac{dr}{dt}\rangle, \quad m \langle \omega | \times \langle \omega | \times |r\rangle\rangle
\]

are called the ‘Coriolis force’, named for the French mathematician, mechanical engineer and scientist Gustave de Coriolis (21 May 1792 – 19 September 1843), and ‘centrifugal force’ respectfully.

To put things in perspective, the angular velocity ‘\(\omega\)’ of the Earth, which rotates on its axis about once every 24 hours is
\[
\omega = \frac{2\pi}{24 \times 3600} \approx 7.3 \times 10^{-5} \ [\text{rads/sec}],
\]
an angular velocity so small it is often ignored. Nevertheless, the earth’s rotation does have measurable effects on the motion of projectiles, pendulums, and other systems [214].
The Coriolis force on an object is proportional to the object's velocity relative to the rotating frame. The force is zero for any object at rest in the rotating frame, and it is negligible for objects moving slowly.

In free-fall acceleration, $|g|$ is the initial acceleration, relative to the Earth, of an object that is released from rest near the Earth's surface. The equation of motion (relative to the Earth) is $m\ddot{r} = -F_{\text{gravity}} + F_{\text{Cor}} + F_{\text{CF}}$. The Coriolis force is zero if the body is released from rest. When the object speeds up, the Coriolis force eventually becomes important and the acceleration changes. Hence, ignoring the Coriolis force,

$$m\ddot{r} = -F_{\text{gravity}} + F_{\text{CF}} = -mg_0 + m\omega^2 R \sin \theta,$$

where $\theta$ is the angle between the axis of rotation of the Earth and the place on the Earth were the event is taking place. Note that, at the poles, $\theta = 0$ and $m\ddot{r} = -F_{\text{gravity}}$. At the equator $\theta = \pi/2 \rightarrow \sin \theta = 1$ and

$$\omega^2 R = (7.3 \times 10^{-5})^2 \times (6.4 \times 10^6 \text{ [meters]}) \approx .034 \text{ [meters/sec}^2]\]$$

Since $g_0$ is about $9.8 \text{ m/sec}^2$, because of the centrifugal force, the value of $g$ at the equator is about 0.3% less than at the poles\(^5\) [214].

Viewed from above the North Pole, since the earth rotates to the east, the angular velocity tends to deflect moving bodies to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. Because of the Coriolis effect, due to the earth's rotation, long-range gunners must aim to the left of their target in the Northern and to the right of their target in the Southern Hemisphere. Moreover, the air is deflected to the right and begins to circulate counterclockwise in the Northern Hemisphere — clockwise in the Southern Hemisphere explaining why a cyclone, hurricane, or typhoon rotates counterclockwise in the Northern Hemisphere, but clockwise in the Southern Hemisphere [214].

### 7.6 Maxwell’s Equations

Newtonian mechanics was the first monumental work in modern physics. Its three fundamental laws explained the force of gravity, how bodies moved and acted under the influence of forces and unified a vast number of formerly disparate physical phenomena. Later, Newton published the ‘Opticks’, which explained the nature of light. But about one hundred years after its publication, his corpuscular theory of light was challenged by experiments showing that light was a wave. Moreover, in Newton's time, there was little evidence that light, electricity and magnetism were one and the same phenomenon. A theoretical explanation for the nature of electricity and magnetism was complicated by the fact that the electric force was not only attractive, but repulsive. Electric forces were much stronger than gravitational forces, but unlike gravity, only showed their effects under special circumstances. Experiments with electricity varied widely and gave contradictory results. Electromagnetic phenomena sometimes

---

\(^5\) Assuming that the earth is perfectly spherically symmetric; this, although a very good approximation, is not exactly correct.
behaved like a fluid, but other times like a corpuscle. It was not obvious how the laws of electricity should be formulated. Gravity required only a single equation to describe it. Electromagnetism requires four equations. A truly unified explanation for electricity and magnetism had to wait for Michael Faraday and James Clerk Maxwell to complete their investigations, some two hundred years after Newton published his monumental work.

The fundamental force ‘\( F_E \)’ between two electric charges, discovered by Coulomb i.e. 
\[
F_E = k q_1 q_2 / r^2, \quad k = 1/4\pi\varepsilon
\]
was an “action-at-a-distance” formulation. Later, Gauss introduced an equivalent field concept, describing the electric field in terms of just one charge i.e.
\[
E = k \frac{q}{r^2}, \quad k = \frac{1}{4\pi\varepsilon}
\]
Instead of a charge ‘\(+q\)’ initiating a force on another charge ‘\(Q\)’, \(q\) creates a field that emanates equally in all directions (see figure 7.6-1) and the charge ‘\(Q\)’ in the vicinity of \(q\) suffers the effects of the field set up by \(q\). The quantity ‘\(E\)’ gives the strength of the field a distance ‘\(r\)’ from the charge ‘\(+q\)’.

7.6.1 Gauss’ Law – the First of Maxwell’s Equations

Gauss’ law is predicated on the idea of ‘flux’ – the amount of the electric field that passes through a surface. Mathematically, a small flux ‘\(d\phi\)’ is defined as ‘\(d\phi = (E|dA)\)’,
where \(dA\) is a small patch of surface area. If the surface is spherical, then
\[
\phi = \int d\phi = \oint (E|dA) = E4\pi r^2,
\]
where \(\phi\) is the total flux and \(4\pi r^2\) is the surface area of a sphere. But
\[
E = \frac{1}{4\pi\varepsilon} \frac{q}{r^2} \quad \phi = E4\pi r^2 = \frac{1}{4\pi\varepsilon} \frac{q}{r^2} 4\pi r^2 = \frac{q}{\varepsilon}
\]
If there is more than one charge within the sphere, then
\[
\phi = \sum_i q_i \varepsilon
\]
the total flux through a surface is equal to the sum of the charges within the surface. This is ‘Gauss’ law’ for electric charges. It is the first of Maxwell’s equations, usually written

\[ \phi = \oint (E \mid dA) = \sum q_i / \varepsilon \]

### 7.6.1.1 Calculating Charge Using Gauss’ Law

![Figure 7.6.1.1-1](image)

If a charge ‘+Q’ is evenly distributed on the surface of the sphere (see figure 7.6.1.1-1), what is the value of the electric field at the point ‘P’? If a sphere is drawn around P such that the charge ‘+Q’ is inside the sphere, then applying Gauss’s law results in

\[ \phi = E 4 \pi r^2 = Q / \varepsilon \rightarrow E = Q / (4 \pi r^2 \varepsilon), \]

where P is a distance ‘r’ from the center of the sphere. At any point located external to a charged sphere, the value of the electric field at that point is the same as if all the charge is located at the center of the sphere.

Consider the point ‘P’, where the charge ‘+Q’ is moved to the external sphere (see figure 7.6.1.1-1). The point ‘P’ is now located wholly within the charged sphere. The amount of charge on the internal sphere is zero, since all the charge is on the external sphere. According to Gauss’s law

\[ \phi = E \int dA = E 4 \pi r^2 = Q / \varepsilon = 0 \rightarrow E = 0, \]

since \(4 \pi r^2 \neq 0\). Since P is internal to the sphere, a charge placed there suffers no electric force.

### 7.6.2 Gauss’ Law for the Magnetic Field – the Second of Maxwell’s Equations

What is Gauss’ law for a magnetic field? By hypothesis,

\[ \phi = \oint (B \mid dA) = \sum m_i / k, \]
where $B$ is the value of the magnetic field and $\Sigma_i m_i$ is the sum of the magnetic charges inside, say a sphere, and $k$ is a constant. However, magnets are dipoles, there are just as many positive as negative charges i.e. there are just as many $+m_i$'s as $-m_i$'s. Hence, $\Sigma_i m_i = 0$ and

$$\phi = \oint \langle B | d\mathbf{A} \rangle = 0$$

This is Maxwell's second equation, which implies that there are no magnetic monopoles in Nature. Note that simply because the flux is zero does not imply that the field does not exist, but only that the amount of flux out of the surface equals the amount of flux into the surface, and hence, the flux's cancel.

### 7.6.3 Faraday’s Law of Induction – the Third of Maxwell’s Equations

By experiment, if a magnet is moved through a coil of wire, a current is created in the wire (see figure 7.6.3-1). And if there is a current in the wire, there must be a potential difference or $emf$ ($\mathcal{E}$) that accounts for the current. In order for a current to arise, the magnet must move relative to the coil. If both the wire and magnet are stationary, no current emerges. Since the motion of the magnetic lines of force creates the current,

$$\mathcal{E} \propto \frac{d|B|}{dt}$$

In other words, the $emf$ is proportional to the time rate of change of the magnetic field.

![Figure 7.6.3-1](image)

Also, by experiment,

$$\mathcal{E} \propto A,$$

where $A$ is the surface area enclosed by the current loop, which implies that

$$\mathcal{E} \propto \frac{d|B|}{dt} A$$
Recall that the magnetic flux is defined as
\[ d\phi_B = \langle B|dA \rangle \to \frac{d\phi_B}{dt} = \langle dB|dA \rangle \to d\mathcal{E} \propto \langle dB|dA \rangle = \frac{d}{dt} \langle B|dA \rangle \]

If the plates of a large capacitor are far enough apart so that the charge around the edges of the plates can be ignored, consider just one of the plates (see figure 7.6.3-2). If the plate is flat, the electric field \( |E\rangle \) is at right angles to the plate. If the charge density of the plate is \( \sigma \), then, according to Gauss’ law, the flux through a surface of, in this case a cylinder, is
\[ d\phi = \langle E|dA \rangle \to \phi = E2\pi r^2 = \sum \frac{q}{\varepsilon} \]

The quantity \( \phi = E2\pi r^2 \) is \( E \) times the area of a circle at the bottom and top of the cylinder i.e. \( A = 2\pi r^2 \), since the area of a circle is \( \pi r^2 \). No flux passes through the sides of the cylinder, since the sides are parallel to the field lines. So, \( \phi = \sum q/\varepsilon = (\sigma/\varepsilon)\pi r^2 \), the charge density \( \sigma \) times the area of the circle that sits on the plate i.e. \( \phi = (\sigma/\varepsilon)\pi r^2 \). Hence,
\[ \phi = E2\pi r^2 = \frac{\sigma}{\varepsilon} \pi r^2 \to E = \frac{\sigma}{2\varepsilon} \]

If the second plate is added, with an opposite charge density to the first plate, then the field on the outsides of the two plates cancel and \( |E\rangle = |0\rangle \). However, between the plates the field lines point in the same direction and have a value of \( 2E \). Hence, the field value inside the plates is
\[ E = \frac{\sigma}{\varepsilon} \]

Figure 7.6.3-2

The potential energy is the work \( 'W' \) done to move a charge \( 'q' \) a distance \( 'd' \), defined as \( W = \langle F|d \rangle \). The magnitude of the force between two electric charges is
\[ F = \frac{qQ}{4\pi\varepsilon r^2} \]
The work done to move a charge a distance ‘r’ is
\[
W = \frac{qQ}{4\pi\varepsilon r^2} = \frac{qQ}{4\pi\varepsilon r}
\]
Note that the potential energy at infinity is zero, since \( \lim_{r \to \infty} |F| = 0 \). But the voltage ‘V’ (potential difference) is defined as potential energy per unit charge. Hence,
\[
V = \frac{Q}{4\pi\varepsilon r} \rightarrow V = Er \rightarrow E = \frac{V}{r}
\]
If \( E \) is constant, \( V \) must vary as the distance from the charge changes. Moreover, even if the charge density is not along a flat plate, a curved wire, for instance, the potential difference can be written as a differential i.e.
\[
dV = d\mathcal{E} = \langle E|ds \rangle \rightarrow \mathcal{E} = \oint \langle E|ds \rangle
\]
where \( ds \) is an infinitesimal distance along, say, a conductor. Recall that
\[
\mathcal{E} = \frac{d\phi}{dt} \propto \oint \left( \frac{dB}{dt} \left| dA \right| \right) \rightarrow \frac{d}{dt} \oint \langle B|dA \rangle = -\oint \langle E|ds \rangle,
\]
This is Faraday’s law of induction and is the third of Maxwell’s equations. The minus sign on the right-hand side of the equation above comes from Lenz’s law, which says that an induced current must resist the magnetic field which creates it.

**7.6.4 The Ampere-Maxwell Law – the Fourth of Maxwell’s Equations**

If a current ‘I’ is sent through a long straight wire, the current will produce a magnetic field ‘\( |B| \)’ around the wire (see figure 7.6.4-1).

![Figure 7.6.4-1](image)

The magnitude of ‘\( B \)’ a distance ‘\( r \)’ from the wire is proportional to ‘\( I/r \)’ i.e.
\[
B \propto \frac{I}{r} \rightarrow B = k \frac{I}{r}, \quad k = \frac{\mu_0}{2\pi}
\]
Note that

\[ \langle B|ds \rangle \rightarrow B \oint ds \cos \theta = B \oint ds = B2\pi r, \]

since \( \theta = \pi/2 \rightarrow \cos \theta = 1 \). The field is in the form of a circle, and therefore, the distance around the flux line is \( 2\pi r \). Hence,

\[ \oint (B|ds) = B2\pi r = \frac{\mu_0 I}{2\pi r}2\pi r = \mu_0 I \]

The equation above is called ‘Ampere’s law’.

**7.6.4.1 The Magnetic Field Associated with a Changing Electric Field**

Is there a magnetic field associated with a changing electric field? If so, then the right-hand side of the equation above will have to be modified to reflect the “symmetry” between the electric and magnetic fields ‘\(|E\rangle\)’ and ‘\(|B\rangle\)’.

To see how magnetic fields are associated with a time-varying electric field, consider the process of charging a capacitor. During the charging process, the electric field strength increases with time as more charge is accumulated on the plates. The conduction current that carries the charges also produces a magnetic field. In order to apply Ampere’s law to this field, choose the curve ‘\(C_1\)’ as the Amperian loop, shown in fig. 7.6.4.1-1.

If the surface bounded by the path is the flat surface ‘\(S_1\)’, then the enclosed current is \(I\). On the other hand, if \(S_2\) is the surface bounded by the curve, then \(I = 0\), because there is no current through \(S_2\). An ambiguity exists in choosing the appropriate surface bounded by the curve ‘\(C_1\)’.

![Diagram](image_url)

**Figure 7.6.4.1-1**

Ampere’s law requires that the current pass through the surface. But in the situation illustrated by fig. 7.6.4.1-1, the surface sits between two capacitor plates where no current flows (assuming the capacitor is still charging), and hence, there should be no magnetic field around the wire. But, of course, there is. To handle this situation, note that
\[ \mu_0 I_d = \mu_0 \frac{dq}{dt} = \mu_0 \frac{\varepsilon \cos \theta A dq}{\varepsilon \cos \theta A dt} = \mu_0 \varepsilon \cos \theta A \frac{d}{dt} \left( \frac{q}{\varepsilon \cos \theta A} \right) \]

Recall that

\[ \phi = \oint (E|dA) = E \cos \theta \oint dA = E \cos \theta A = \frac{q}{\varepsilon} \rightarrow E = \frac{q}{\varepsilon \cos \theta A} \]

Therefore,

\[ \mu_0 I_d = \mu_0 \varepsilon \cos \theta A \frac{dE}{dt} \rightarrow \mu_0 I_d = \mu_0 \varepsilon \frac{d}{dt} \left( E \cos \theta A \right) = \mu_0 \varepsilon \frac{(E|dA)}{dt} = \mu_0 \varepsilon \frac{d\phi}{dt}, \]

since \( (E|dA) = d\phi/dt \) – the change in the flux with time. Hence,

\[ \oint (B|ds) = \mu_0 I + \mu_0 \varepsilon \frac{d\phi}{dt} = \mu_0 (I + I_d), \]

where the second term on the right-hand side of the equation above is called the ‘displacement current’ and takes care of the situations where Ampere’s law, in its original form, is not adequate. This is Maxwell’s fourth equation.

**7.6.5 A Summary of Maxwell’s Equations**

The four equations that form the foundation of electromagnetic phenomena, listed in table 7.6.5-1, are called ‘Maxwell’s equations’. The equations have equivalent differential and integral forms. Virtually all classical electromagnetic phenomena can be derived from these equations including interference, diffraction and polarization. Consult [226] for details.
<table>
<thead>
<tr>
<th>Law</th>
<th>Differential Form</th>
<th>Integral Form</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gauss' Law</strong></td>
<td>$\langle \nabla</td>
<td>E\rangle = \sigma/\varepsilon$</td>
<td>$\oint (E</td>
</tr>
<tr>
<td><strong>Gauss' Law (Magnetism)</strong></td>
<td>$\langle \nabla</td>
<td>B\rangle = 0$</td>
<td>$\oint (B</td>
</tr>
<tr>
<td><strong>Faraday's Law</strong></td>
<td>$\langle \nabla \times</td>
<td>E\rangle = -\frac{\partial</td>
<td>B\rangle}{\partial t}$</td>
</tr>
<tr>
<td><strong>Ampere-Maxwell Law</strong></td>
<td>$\langle \nabla \times</td>
<td>B\rangle = \mu_0i + \mu_0\varepsilon \frac{\partial</td>
<td>E\rangle}{\partial t}$</td>
</tr>
</tbody>
</table>

Table 7.6.5-1

### 7.7 Electromagnetic Waves

Experiments had shown, fairly conclusively, that light was a wave, which contradicted Newton’s corpuscular theory. But Maxwell went further by showing that light consisted of polarized electromagnetic waves that traveled through empty space at the speed of light.

![Figure 7.7-1](image)

To see this, assume that the electric and magnetic fields, ‘$|E\rangle$’ and ‘$|B\rangle$’, are constrained to the $x,y$- and $x,z$-planes respectfully and are both functions of $x$ and $t$ i.e.

$$|E(x,t)) = E(x,t)|j\rangle, \quad |B(x,t)) = B(x,t)|k\rangle,$$
which results in a linearly polarized plane wave travelling in the $+x$-direction (see fig. 7.7-1). If the wave is travelling in free space, $i, q = 0$ and Maxwell’s equations can be written

$$\langle \nabla |E\rangle = 0, \quad \langle \nabla |B\rangle = 0, \quad \langle \nabla \times |E\rangle = -\frac{\partial |B\rangle}{\partial t}, \quad \langle \nabla \times |B\rangle = \mu_0 \varepsilon \frac{\partial |E\rangle}{\partial t}$$

Recall that

$$\langle \nabla \times |E(x, t)\rangle \frac{\partial}{\partial x} \begin{pmatrix} |i\rangle \\ |j\rangle \\ |k\rangle \end{pmatrix} = \frac{\partial E(x, t)}{\partial x} |k\rangle$$

But, by Faraday’s law

$$|\langle \nabla \times |E\rangle| = -\frac{\partial B}{\partial t} = \frac{\partial E}{\partial x}$$

So

$$\langle \nabla \times |B\rangle \frac{\partial}{\partial x} \begin{pmatrix} |i\rangle \\ |j\rangle \\ |k\rangle \end{pmatrix} = -\frac{\partial B(x, t)}{\partial x} |j\rangle \rightarrow \frac{\partial B}{\partial x} = -\mu_0 \varepsilon \frac{\partial E}{\partial t},$$

which indicates that the spatial variation of the magnetic field produces a time-varying electric field and visa-versa.

Therefore,

$$\left(\frac{\partial^2 E}{\partial x^2} \frac{\partial}{\partial x} \frac{\partial E}{\partial x} = -\frac{\partial B}{\partial x} \frac{\partial B}{\partial t} = -\frac{\partial \partial B}{\partial t} \partial x = -\frac{\partial E}{\partial t} \left(\frac{\partial \partial E}{\partial t} = \mu_0 \varepsilon \frac{\partial^2 E}{\partial t^2}\right)\right)^6$$

Similarly,

$$\frac{\partial^2 B}{\partial x^2} = \mu_0 \varepsilon \frac{\partial^2 B}{\partial t^2}$$

Recall (Chapter 1, Sec. 1.3.3.1) that the form of the wave equation is

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2}$$

---

6 The third equality assumes that $B(x, t)$ is analytic, so that the order of the partial derivatives is immaterial.
If \( y \) is replaced by \( E \) or by \( B \) and \( 1/v^2 \) is replaced by \( \mu_0 \varepsilon \), this showed that light was an electromagnetic wave, travelling in free space at a velocity \( c = 1/\sqrt{\mu_0 \varepsilon} \approx 3 \times 10^{10} \text{[cm/sec]} \) [233].

A solution to the wave equation is

\[
E(x, t) = E_0 \cos(kx - \omega t), \quad B(x, t) = B_0 \cos(kx - \omega t),
\]

where \( k = 2\pi/\lambda \), \( \lambda \) is the ‘wave length’ and \( \omega = 2\pi v \) is the ‘frequency’ of the wave. Hence,

\[
\frac{\partial E(x, t)}{\partial x} = -kE_0 \sin(kx - \omega t), \quad \frac{\partial B(x, t)}{\partial t} = \omega B_0 \sin(kx - \omega t)
\]

Since

\[
-\frac{\partial B}{\partial t} = \frac{\partial E}{\partial x} \rightarrow -kE_0 \sin(kx - \omega t) = -\omega B_0 \sin(kx - \omega t) \rightarrow E_0 = \frac{\omega}{k} B_0
\]

But \( \omega/k = 2\pi v/2\pi/\lambda = v \lambda = v = c \). Electromagnetic waves travel at a velocity ‘\( c \)’, the speed of light in free space. Hence, \( E_0/B_0 = \omega/k = c \). At every instant, the ratio of the electric field to the magnetic field in an electromagnetic wave equals the speed of light [233]. This implies that \( E_0 = cB_0 \).

### 7.7.1 The Poynting Vector

Energy can also be transported by electromagnetic waves. Consider a plane electromagnetic wave passing through a small volume element of area ‘\( A \)’ and thickness ‘\( dx \)’. The total energy ‘\( dU \)’ stored in the electromagnetic field in the volume element is

\[
dU = uA dx = (u_E + u_B) A dx = \frac{1}{2} \left( \varepsilon E^2 + \frac{B^2}{\mu_0} \right) A dx,
\]

where \( u_E = \varepsilon E^2/2 \) and \( u_B = B^2/2\mu_0 \) are the energy densities associated with the electric and magnetic fields respectively. Because the electromagnetic wave propagates with speed ‘\( c \)’, the amount of time it takes for the wave to move through the volume element is \( dt = dx/c \). Thus, the rate of change of energy per unit area, denoted by the symbol ‘\( S \)’, is

\[
S = \frac{dU}{Adt} = \frac{c}{2} \left( \varepsilon E^2 + \frac{B^2}{\mu_0} \right)
\]

The magnitude of the field satisfies \( E = cB, \) \( c = 1/\sqrt{\mu_0 \varepsilon} \). Therefore, the equation above can be written

\[
S = \frac{c}{2} \left( \varepsilon E^2 + \frac{B^2}{\mu_0} \right) = \frac{cB^2}{\mu_0} = \frac{EB}{\mu_0}
\]
This energy flow can be turned into a vector by assigning it a direction, the direction of propagation. The rate of energy flow per unit area is called the ‘Poynting vector’ \(|\mathbf{S}|\) named after the British physicist John Henry Poynting (9 September 1852 – 30 March 1914), and is defined by the vector product

\[ |\mathbf{S}| = \frac{1}{\mu_0} \langle \mathbf{E} \times \mathbf{B} \rangle \]

For a plane electromagnetic wave

\[ S = \frac{EB}{\mu_0} = \frac{E^2}{c\mu_0} = \frac{cB^2}{\mu_0} \]

To see this, recall that

\[ |E(x,t)| = E(x,t)|j|, \quad |B(x,t)| = B(x,t)|k| \Rightarrow E(x,t) = E_0 \cos(kx - \omega t)|j|, \quad B(x,t) = B_0 \cos(kx - \omega t)|k| \]

Hence,

\[ |\mathbf{S}| = \frac{1}{\mu_0} \langle \mathbf{E} \times \mathbf{B} \rangle = \langle E_0 \cos(kx - \omega t)|j| \times B_0 \cos(kx - \omega t)|k| \rangle = \frac{E_0 B_0}{\mu_0} \cos^2 (kx - \omega t)|i| \]

As expected, \(|\mathbf{S}|\) points in the direction of wave propagation.

The time average of ‘\(S\)’ over one or more cycles is called the wave intensity ‘\(I\)’, which gives the intensity of a light wave and is proportional to the square of the amplitude of the electric or magnetic fields:

\[ I = \bar{S} = \frac{E_0^2}{2c\mu_0} = \frac{cB_0^2}{2\mu_0} \]

(The 1/2 comes from averaging the cosine squared terms.) Intensity is a measureable quantity. It is the loudness in a sound wave or the brightness of light and serves as an important conceptual analog in the study of quantum physics in the following way: “Intensity is to the electric field as probability is to the wave function.” [233].

### 7.8 Generalizations of Classical Mechanics

There is a natural inclination to think, since the gravitational force between two masses and the electric force between two charges have the same mathematical form, that somehow, the two forces reflect a higher, still unknown, unifying concept. But if so, there remain several unanswered questions:
1. Why are gravitational forces only attractive while electric forces attract and repel?

2. What accounts for the difference in magnitudes of the two forces?

3. What is the gravitational analog for magnetism?

4. Since light is propagated by electromagnetic waves, does gravity propagate by gravitational waves?

Satisfactory answers to these questions and the successful discovery of a unifying principle have yet to materialize. On the other hand, at the time, less was known about the difficulties in pursuing a higher unified theory.

### 7.8.1 Poisson’s Equation

The first step in pursuing a unified theory of electromagnetism and gravity is to put both theories on the same philosophical foundation. Newtonian gravity, an "action-at-a-distance" concept, was reformulated to make it consistent with Maxwell’s field concept.

The amount of ‘flux’ ($\phi$) through a closed surface is

$$\phi = \oint (g|dA)$$

where $|g\rangle$ is the field. Gravitational flux depends on the inertial mass being equivalent to gravitational mass i.e.

$$m|g\rangle = -\frac{GmM}{r^2}|e_r\rangle \rightarrow |g\rangle = -\frac{GM}{r^2}|e_r\rangle,$$

where $|g\rangle$ is now called the ‘gravitational field’. Hence,

$$\phi = -\frac{GM}{r^2} \oint (e_r|dA)$$

is interpreted as the ‘gravitational flux’ through a surface.

If the surface is a spherical surface, then

$$\phi = -\frac{GM}{r^2} \oint (e_r|dA) = -\frac{GM}{r^2} \oint dA = -\frac{GM}{r^2} 4\pi r^2 = -4\pi GM$$

The equation above can be derived for a generic closed surface (not necessary a sphere) and for a generic distribution or system of bodies. In the latter case, $M$ is replaced by $\sum_i M_i$, and hence,

$$\phi = -4\pi G \sum_i M_i$$
Using the theorem of divergence,

\[ \phi = \oint (g \mid dA) = \int_\tau (\nabla \mid g) \, d\tau = -4\pi G \int_\tau \rho \, d\tau \rightarrow (\nabla \mid g) = -4\pi G \rho \]
i.e.

\[ \phi = -\iiint \left( \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} \right) d\tau = -\iiint (\nabla \mid \nabla V) \, d\tau = -4\pi GM = -4\pi G \iiint \rho \, d\tau, \]

where \( d\tau \) is an element of volume and \( \rho \) is the density of matter. Hence,

\[ \nabla^2 V = 4\pi G \rho \]

In regions free of matter \( \rho = 0 \rightarrow \nabla^2 V = 0 \). The equation above is called ‘Poisson’s equation’. For electric charges, Poisson’s equation becomes

\[ \nabla^2 V = -4\pi \rho_e, \]

where \( \rho_e \) is the density of electric charges per unit volume.

In the equation ‘\( \phi = -4\pi GM \)’, ‘\( 4\pi G \)’ is interpreted as the number of lines of force emanating from each unit of mass i.e. the number of lines penetrating the closed surface of a volume containing mass \( M \) must be \( 4\pi GM [224] \).

Recalling that gravity is a central force, and therefore, conservative, the quantity ‘\( V \)’ is the ‘potential’ and the force can be calculated from

\[ |F\rangle = -\nabla V \]
i.e.

\[ dV = -\frac{G}{R} \, dm = -\frac{G}{R} \rho \, d\tau = -\frac{G}{R} \rho \, dx \, dy \, dz \rightarrow dF = -\frac{\partial}{\partial R} (dV) = -\frac{G \rho}{R^2} \, dx \, dy \, dz, \]

\[ dm = \rho \, dx \, dy \, dz \]

\[ d\tau = r^2 \sin \theta \, dr \, d\theta \, d\phi \]

![Figure 7.8.1-1](image-url)
The potential \( V \) at a point \( P \) a distance \( R \) from a homogeneous spherical shell of matter of radius \( r \) and thickness \( dr \) is (see fig. 7.8.1-1)

\[
V = -G \rho \iiint \frac{dx dy dz}{R} = -G \rho \int_r^{r+dr} \int_0^{2\pi} \int_0^{\pi} \frac{r^2 \sin \theta \, dr d\theta d\phi}{R},
\]

where the volume element has been expressed in spherical coordinates. Recalling the law of cosines:

\[
R^2 = R_0^2 + r^2 - 2R_0 r \cos \theta \to \frac{d}{d\theta} (R^2 = R_0^2 + r^2 - 2R_0 r \cos \theta) \to \frac{R dR}{R_0} = r \sin \theta \, d\theta
\]

\[
V_E = -G \rho \int_r^{r+dr} \int_{R_0-r}^{R_0+r} \int_0^{2\pi} \frac{r}{R_0} \, dr \, dR \, d\phi = -\frac{4\pi G \rho}{R_0} r^2 dr,
\]

\[
V_I = -G \rho \int_r^{r+dr} \int_{r-R_0}^{r+R_0} \int_0^{2\pi} \frac{r}{R_0} \, dr \, dR \, d\phi = -4\pi G \rho r \, dr,
\]

where \( V_E \) and \( V_I \) represent the potentials when a point \( P \) is external or internal to the sphere respectfully. Since \( V_I \) is independent of \( R_0 \), its derivative with respect to \( R_0 \) is zero. Hence, the potential at any point within a hollow spherical shell is zero. If the mass of the shell is \( dM_0 = 4\pi \rho r^2 \, dr \), then the potential at any external point becomes

\[
V_E = -\frac{4\pi G}{R_0} \int_0^R \rho r^2 \, dr = -\frac{GM}{R_0}
\]

The value \( M \) becomes the total mass of the sphere.

### 7.9 Concluding Remarks

In practice, physical objects, ranging from those larger than atoms and molecules to objects in the macroscopic and astronomical realm, can be well-described by classical mechanics. Beginning at the atomic scale and smaller, the laws of classical physics break down. Electromagnetic fields and forces can be described well by classical electrodynamics at length scales and field strengths large enough that quantum mechanical effects are negligible. Unlike quantum physics, classical physics is characterized by the principle of ‘complete determinism’, where, if the initial conditions are known, the future state of any physical system can be predicted with certainty. Classical mechanics includes the theory of relativity.

Non-relativistic classical physics refers to theories that do not use quantization (do not include quantum mechanics) or the theory of relativity. In short, non-relativistic classical theories are those that obey Galilean relativity. Among the topics included: non-relativistic classical mechanics, Newton's laws of motion, the laws of planetary motion,
classical Lagrangian and Hamiltonian formalisms, classical electrodynamics (Maxwell’s Equations) and classical thermodynamics.

The predictions of non-relativistic physics are significantly different from those of general and special relativity, particularly concerning the passage of time, the geometry of space, the motion of bodies in free fall and the propagation of light. Mathematically, the equations of non-relativistic classical physics do not include Planck’s constant and no accounting is made of the disposition of the observer viewing a physical event. According to the correspondence principle and Ehrenfest’s theorem, as a system becomes larger or more massive, quantum mechanical predictions merge with non-relativistic classical predictions, although there are some exceptions. When dealing with everyday objects, quantum mechanics can generally be ignored. However, one of the most vigorous on-going fields of research in physics is classical-quantum correspondence, concerned with the discovery of how the laws of quantum physics give rise to classical physics found at the limit of the large scales at the classical level.
Chapter 8
The Principles of Analytical Mechanics

“The object of physics is the unfolding of the laws of the intelligible world; the object of pure mathematics is the unfolding of the laws of human intelligence.”

- James Joseph Sylvester

8.0 Introduction

Analytical mechanics is a collection of alternative formulations of classical mechanics. Its development involved many scientists and mathematicians during the 18th century and onward after Newtonian mechanics was completed. While Newtonian mechanics primarily uses vector quantities, particularly when describing accelerations, momenta and forces of the constituents of the system, analytical mechanics employs scalar functions to represent a system—functions related to the kinetic and potential energy of a system [267].

8.1 Curvilinear Coordinates

Recall that if \( \langle x, y, z \rangle \) is the position of a point ‘\( P \)’ expressed in rectangular coordinates, then

\[
\langle x, y, z \rangle \equiv \langle r \rangle = x|e_1\rangle + y|e_2\rangle + z|e_3\rangle,
\]

where \( |e_i\rangle \)’s are unit orthogonal vectors. If

\[
x = x(q_1, q_2, q_3), \quad y = y(q_1, q_2, q_3), \quad z = z(q_1, q_2, q_3),
\]

then the variables ‘\( x, y, z \)’ are functions of ‘\( q_1, q_2, q_3 \)’ and if the ‘\( q_1, q_2, q_3 \)’ are also functions of \( x, y, z \) such that

\[
q_1 = q_1(x, y, z), \quad q_2 = q_2(x, y, z), \quad q_3 = q_3(x, y, z),
\]

where each of the functions above are assumed single valued having derivatives at every point, making the correspondence between the two functions unique, then \( \langle x, y, z \rangle \) and \( \langle q_1, q_2, q_3 \rangle \) specify the same point ‘\( P \)’ in two different coordinate systems. The coordinates ‘\( \langle q_1, q_2, q_3 \rangle \)’ are called the ‘curvilinear coordinates’ of \( P \); the equations above correspond to the transformation equations that allow a conversion from one set of coordinates to another [205].
8.1.1 Contravariant and Covariant Components of a Vector

If

\[ |A| = C_1 \frac{\partial |r|}{\partial q_1} + C_2 \frac{\partial |r|}{\partial q_2} + C_3 \frac{\partial |r|}{\partial q_3}, \quad |A| = c_1 \nabla |q_1| + c_2 \nabla |q_2| + c_3 \nabla |q_3|, \]

\[ \nabla = \frac{\partial}{\partial x} |e_1| + \frac{\partial}{\partial y} |e_2| + \frac{\partial}{\partial z} |e_3| \]

then \( C_1, C_2, C_3 \) are called the ‘contravariant components’ and \( c_1, c_2, c_3 \) the ‘covariant components’ of ‘\(|A|\)’. Consider

\[ A^2 = \sum_{j=1}^{3} \sum_{i=1}^{3} C_i C_j \frac{\partial |r|}{\partial q_i} \nabla |q_j| \]

If the system of coordinates is orthogonal, \( \frac{\partial |r|\nabla |q_j|}{\partial q_i} = \delta_{ij} \), since \( \nabla |q_j| = \partial q_j/\partial |r| \). Moreover, \( C_i = c_i, \ i = 1,2,3 \) (see Book II: Chapter 12, Sec. 12.8.1). Hence,

\[ \langle A|A\rangle = A^2 = \sum_{j=1}^{3} \sum_{i=1}^{3} C_i C_j \frac{\partial |r|}{\partial q_i} \nabla |q_j| = C_1^2 + C_2^2 + C_3^2 = c_1^2 + c_2^2 + c_3^2, \quad i = 1,2,3 \]

8.1.2 The Generalized Metric

If the restriction to orthogonality is lifted, the contravariant and covariant components of a vector are not necessarily equal. Let the transformation equations of a point \((x,y,z)\) into two different curvilinear system ‘\((q_1, q_2, q_3)\)’ and ‘\((\bar{q}_1, \bar{q}_2, \bar{q}_3)\)’ be

\[ x = x(q_1, q_2, q_3), \quad y = y(q_1, q_2, q_3), \quad z = z(q_1, q_2, q_3) \]
\[ x = x(\bar{q}_1, \bar{q}_2, \bar{q}_3), \quad y = y(\bar{q}_1, \bar{q}_2, \bar{q}_3), \quad z = z(\bar{q}_1, \bar{q}_2, \bar{q}_3), \]

then there is a transformation directly from \((q_1, q_2, q_3)\) to \((\bar{q}_1, \bar{q}_2, \bar{q}_3)\) defined by

\[ q_1 = q_1(\bar{q}_1, \bar{q}_2, \bar{q}_3), \quad q_2 = q_2(\bar{q}_1, \bar{q}_2, \bar{q}_3), \quad q_3 = q_3(\bar{q}_1, \bar{q}_2, \bar{q}_3) \]

8.1.2.1 The Derivation of the Contravariant Components of a Vector

From the three transformation equations above

\[ dq_1 = \frac{\partial q_1}{\partial \bar{q}_1} d\bar{q}_1 + \frac{\partial q_1}{\partial \bar{q}_2} d\bar{q}_2 + \frac{\partial q_1}{\partial \bar{q}_3} d\bar{q}_3 \]
\[ dq_2 = \frac{\partial q_2}{\partial \bar{q}_1} d\bar{q}_1 + \frac{\partial q_2}{\partial \bar{q}_2} d\bar{q}_2 + \frac{\partial q_2}{\partial \bar{q}_3} d\bar{q}_3 \]
\[ dq_3 = \frac{\partial q_3}{\partial \bar{q}_1} d\bar{q}_1 + \frac{\partial q_3}{\partial \bar{q}_2} d\bar{q}_2 + \frac{\partial q_3}{\partial \bar{q}_3} d\bar{q}_3 \]
Recall that
\[
d |r\rangle = \frac{\partial |r\rangle}{\partial q_1} d q_1 + \frac{\partial |r\rangle}{\partial q_2} d q_2 + \frac{\partial |r\rangle}{\partial q_3} d q_3 = \alpha_1 d q_1 + \alpha_2 d q_2 + \alpha_3 d q_3, \quad \alpha_i = \frac{\partial |r\rangle}{\partial q_i},
\]
\[
d |r\rangle = \frac{\partial |r\rangle}{\partial q_1} d \bar{q}_1 + \frac{\partial |r\rangle}{\partial q_2} d \bar{q}_2 + \frac{\partial |r\rangle}{\partial q_3} d \bar{q}_3 = \bar{\alpha}_1 d \bar{q}_1 + \bar{\alpha}_2 d \bar{q}_2 + \bar{\alpha}_3 d \bar{q}_3, \quad \bar{\alpha}_i = \frac{\partial |r\rangle}{\partial \bar{q}_i}.
\]
Hence,
\[
\alpha_1 d q_1 + \alpha_2 d q_2 + \alpha_3 d q_3 = \bar{\alpha}_1 d \bar{q}_1 + \bar{\alpha}_2 d \bar{q}_2 + \bar{\alpha}_3 d \bar{q}_3
\]
\[
\rightarrow \alpha_1 \left( \frac{\partial q_1}{\partial q_1} d \bar{q}_1 + \frac{\partial q_1}{\partial q_2} d \bar{q}_2 + \frac{\partial q_1}{\partial q_3} d \bar{q}_3 \right) + \alpha_2 \left( \frac{\partial q_2}{\partial q_1} d \bar{q}_1 + \frac{\partial q_2}{\partial q_2} d \bar{q}_2 + \frac{\partial q_2}{\partial q_3} d \bar{q}_3 \right)
\]
\[
+ \alpha_3 \left( \frac{\partial q_3}{\partial q_1} d \bar{q}_1 + \frac{\partial q_3}{\partial q_2} d \bar{q}_2 + \frac{\partial q_3}{\partial q_3} d \bar{q}_3 \right) = \bar{\alpha}_1 d \bar{q}_1 + \bar{\alpha}_2 d \bar{q}_2 + \bar{\alpha}_3 d \bar{q}_3
\]
And this implies that
\[
\alpha_1 \frac{\partial q_1}{\partial q_1} d \bar{q}_1 + \alpha_2 \frac{\partial q_2}{\partial q_1} d \bar{q}_1 + \alpha_3 \frac{\partial q_3}{\partial q_1} d \bar{q}_1 = \left( \alpha_1 \frac{\partial q_1}{\partial q_1} + \alpha_2 \frac{\partial q_2}{\partial q_1} + \alpha_3 \frac{\partial q_3}{\partial q_1} \right) d \bar{q}_1 = \bar{\alpha}_1 d \bar{q}_1 \rightarrow \bar{\alpha}_1
\]
Similarly,
\[
\bar{\alpha}_2 = \alpha_1 \frac{\partial q_1}{\partial \bar{q}_1} + \alpha_2 \frac{\partial q_2}{\partial \bar{q}_1} + \alpha_3 \frac{\partial q_3}{\partial \bar{q}_1}, \quad \bar{\alpha}_3 = \alpha_1 \frac{\partial q_1}{\partial \bar{q}_3} + \alpha_2 \frac{\partial q_2}{\partial \bar{q}_3} + \alpha_3 \frac{\partial q_3}{\partial \bar{q}_3}
\]
Let
\[
|A\rangle = C_1 |\alpha_1\rangle + C_2 |\alpha_2\rangle + C_3 |\alpha_3\rangle, \quad |A\rangle = \tilde{C}_1 |\bar{\alpha}_1\rangle + \tilde{C}_2 |\bar{\alpha}_2\rangle + \tilde{C}_3 |\bar{\alpha}_3\rangle,
\]
where $|A\rangle = (C_1, C_2, C_3)$ and $|A\rangle = (\tilde{C}_1, \tilde{C}_2, \tilde{C}_3)$ so that $|A\rangle$ is specified in two different curvilinear coordinate systems. Since
\[
C_1 |\alpha_1\rangle + C_2 |\alpha_2\rangle + C_3 |\alpha_3\rangle
\]
\[
= \tilde{C}_1 \left( \frac{\partial q_1}{\partial \bar{q}_1} |\alpha_1\rangle + \frac{\partial q_2}{\partial \bar{q}_1} |\alpha_2\rangle + \frac{\partial q_3}{\partial \bar{q}_1} |\alpha_3\rangle \right) + \tilde{C}_2 \left( \frac{\partial q_1}{\partial \bar{q}_1} |\alpha_1\rangle + \frac{\partial q_2}{\partial \bar{q}_1} |\alpha_2\rangle + \frac{\partial q_3}{\partial \bar{q}_1} |\alpha_3\rangle \right)
\]
\[
+ \tilde{C}_3 \left( \frac{\partial q_1}{\partial \bar{q}_3} |\alpha_1\rangle + \frac{\partial q_2}{\partial \bar{q}_3} |\alpha_2\rangle + \frac{\partial q_3}{\partial \bar{q}_3} |\alpha_3\rangle \right)
\]
Hence,

\[ C_1|_\alpha_1 = \tilde{C}_1 \frac{\partial q_1}{\partial \bar{q}_1}|_\alpha_1 + \tilde{C}_2 \frac{\partial q_1}{\partial \bar{q}_1}|_\alpha_1 + \tilde{C}_3 \frac{\partial q_1}{\partial \bar{q}_3}|_\alpha_1 = \left( \tilde{C}_1 \frac{\partial q_1}{\partial \bar{q}_1} + \tilde{C}_2 \frac{\partial q_1}{\partial \bar{q}_1} + \tilde{C}_3 \frac{\partial q_1}{\partial \bar{q}_3} \right) |_\alpha_1 \rightarrow C_1 \]

\[ = \tilde{C}_1 \frac{\partial q_1}{\partial \bar{q}_1} + \tilde{C}_2 \frac{\partial q_1}{\partial \bar{q}_1} + \tilde{C}_3 \frac{\partial q_1}{\partial \bar{q}_3} \]

Similarly,

\[ C_2 = \tilde{C}_1 \frac{\partial q_2}{\partial \bar{q}_1} + \tilde{C}_2 \frac{\partial q_2}{\partial \bar{q}_1} + \tilde{C}_3 \frac{\partial q_2}{\partial \bar{q}_3}, \quad C_3 = \tilde{C}_1 \frac{\partial q_3}{\partial \bar{q}_1} + \tilde{C}_2 \frac{\partial q_3}{\partial \bar{q}_1} + \tilde{C}_3 \frac{\partial q_3}{\partial \bar{q}_3} \]

Hence, the components \( \langle C_1, C_2, C_3 \rangle \) can be written in terms of the components \( \langle \tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \rangle \). The relationships above can be written in shorthand:

\[ C_\rho = \sum_{i=1}^{3} \tilde{C}_i \frac{\partial q_\rho}{\partial \bar{q}_i}, \quad \rho = 1,2,3 \]

Note that

\[ C_\rho = \tilde{C}_i \frac{\partial q_\rho}{\partial \bar{q}_i} \rightarrow C_\rho \frac{\partial q_\rho}{\partial \bar{q}_i} = \tilde{C}_i \frac{\partial q_\rho}{\partial \bar{q}_i} \frac{\partial \bar{q}_i}{\partial q_\rho} = \tilde{C}_i \rightarrow \tilde{C}_\rho = \sum_{i=1}^{3} \tilde{C}_i \frac{\partial q_\rho}{\partial \bar{q}_i}, \quad \rho = 1,2,3 \]

These results lead to the following definition:

*If the quantities \( \langle C_1, C_2, C_3 \rangle \) specified in a coordinate system \( \langle q_1, q_2, q_3 \rangle \) are related to three other quantities \( \langle \tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \rangle \) specified in a coordinate system \( \langle \bar{q}_1, \bar{q}_2, \bar{q}_3 \rangle \) by the transformation equations*

\[ C_\rho = \sum_{i=1}^{3} \tilde{C}_i \frac{\partial q_\rho}{\partial \bar{q}_i}, \quad \tilde{C}_\rho = \sum_{i=1}^{3} C_i \frac{\partial \bar{q}_\rho}{\partial q_i}, \quad \rho = 1,2,3, \]

Then \( C_\rho \) and \( \tilde{C}_\rho \) are called the ‘components of a contravariant vector’.

### 8.1.2.2 The Derivation of the Covariant Components of a Vector

Conversely, suppose

\[ |A| = c_1 \nabla q_1 + c_2 \nabla q_2 + c_3 \nabla q_3 \]

\[ |A| = \tilde{c}_1 \nabla \bar{q}_1 + \tilde{c}_2 \nabla \bar{q}_2 + \tilde{c}_3 \nabla \bar{q}_3 \rightarrow c_1 \nabla q_1 + c_2 \nabla q_2 + c_3 \nabla q_3 \]

\[ = \tilde{c}_1 \nabla \bar{q}_1 + \tilde{c}_2 \nabla \bar{q}_2 + \tilde{c}_3 \nabla \bar{q}_3 \]
Recalling that \( \bar{q}_\rho = \bar{q}_\rho(q_1, q_2, q_3) \) and employing the chain rule

\[
\frac{\partial \bar{q}_\rho}{\partial x} = \frac{\partial \bar{q}_\rho}{\partial q_1} \frac{\partial q_1}{\partial x} + \frac{\partial \bar{q}_\rho}{\partial q_2} \frac{\partial q_2}{\partial x} + \frac{\partial \bar{q}_\rho}{\partial q_3} \frac{\partial q_3}{\partial x}, \quad \frac{\partial \bar{q}_\rho}{\partial y} = \frac{\partial \bar{q}_\rho}{\partial q_1} \frac{\partial q_1}{\partial y} + \frac{\partial \bar{q}_\rho}{\partial q_2} \frac{\partial q_2}{\partial y} + \frac{\partial \bar{q}_\rho}{\partial q_3} \frac{\partial q_3}{\partial y}, \\
\frac{\partial \bar{q}_\rho}{\partial z} = \frac{\partial \bar{q}_\rho}{\partial q_1} \frac{\partial q_1}{\partial z} + \frac{\partial \bar{q}_\rho}{\partial q_2} \frac{\partial q_2}{\partial z} + \frac{\partial \bar{q}_\rho}{\partial q_3} \frac{\partial q_3}{\partial z}
\]

Now

\[
c_1 \nabla q_1 + c_2 \nabla q_2 + c_3 \nabla q_3 = \left( c_1 \frac{\partial q_1}{\partial x} + c_2 \frac{\partial q_2}{\partial y} + c_3 \frac{\partial q_3}{\partial z} \right) |e_1\rangle + \left( c_1 \frac{\partial q_1}{\partial y} + c_2 \frac{\partial q_2}{\partial x} + c_3 \frac{\partial q_3}{\partial z} \right) |e_2\rangle + \left( c_1 \frac{\partial q_1}{\partial z} + c_2 \frac{\partial q_2}{\partial x} + c_3 \frac{\partial q_3}{\partial x} \right) |e_3\rangle
\]

Conversely,

\[
\tilde{c}_1 \nabla \tilde{q}_1 + \tilde{c}_2 \nabla \tilde{q}_2 + \tilde{c}_3 \nabla \tilde{q}_3 = \left( \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial x} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial y} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial z} \right) |e_1\rangle + \left( \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial y} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial x} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial y} \right) |e_2\rangle + \left( \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial z} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial x} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial x} \right) |e_3\rangle
\]

Then

\[
c_1 \frac{\partial q_1}{\partial x} + c_2 \frac{\partial q_2}{\partial y} + c_3 \frac{\partial q_3}{\partial z} = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial x} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial y} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial z}
\]

\[
c_1 \frac{\partial q_1}{\partial y} + c_2 \frac{\partial q_2}{\partial x} + c_3 \frac{\partial q_3}{\partial z} = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial y} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial x} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial y}
\]

\[
c_1 \frac{\partial q_1}{\partial z} + c_2 \frac{\partial q_2}{\partial x} + c_3 \frac{\partial q_3}{\partial x} = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial z} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial x} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial x}
\]

If the first equation above is multiplied by \( \frac{\partial x}{\partial q_1} \)

\[
c_1 \frac{\partial q_1}{\partial x} \frac{\partial x}{\partial q_1} + c_2 \frac{\partial q_2}{\partial x} \frac{\partial x}{\partial q_1} + c_3 \frac{\partial q_3}{\partial x} \frac{\partial x}{\partial q_1} = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial x} \frac{\partial x}{\partial q_1} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial x} \frac{\partial x}{\partial q_1} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial x} \frac{\partial x}{\partial q_1} \Rightarrow c_1
\]

\[= \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial q_1} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial q_1} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial q_1}\]

Similarly,

\[c_2 = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial q_2} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial q_2} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial q_2}, \quad c_3 = \tilde{c}_1 \frac{\partial \tilde{q}_1}{\partial q_3} + \tilde{c}_2 \frac{\partial \tilde{q}_2}{\partial q_3} + \tilde{c}_3 \frac{\partial \tilde{q}_3}{\partial q_3}\]
The relationships above can be written in shorthand:

\[ c_\rho = \sum_{i=1}^{3} \tilde{c}_i \frac{\partial \tilde{q}_i}{\partial q_\rho}, \quad \rho = 1,2,3 \]

Note that

\[ c_\rho = \tilde{c}_i \frac{\partial \tilde{q}_i}{\partial q_\rho} \rightarrow c_\rho \frac{\partial q_\rho}{\partial \tilde{q}_i} = \tilde{c}_i \frac{\partial \tilde{q}_i}{\partial q_\rho} = \tilde{c}_\rho = \sum_{i=1}^{3} c_i \frac{\partial q_i}{\partial \tilde{q}_\rho}, \quad \rho = 1,2,3 \]

These results lead to the following definition:

*If the quantities \( \langle c_1, c_2, c_3 \rangle \) specified in a coordinate system \( 'q_1, q_2, q_3' \) are related to three other quantities \( \langle \tilde{c}_1, \tilde{c}_2, \tilde{c}_3 \rangle \) specified in a coordinate system \( '\tilde{q}_1, \tilde{q}_2, \tilde{q}_3' \) by the transformation equations

\[ c_\rho = \sum_{i=1}^{3} \tilde{c}_i \frac{\partial \tilde{q}_i}{\partial q_\rho}, \quad \tilde{c}_\rho = \sum_{i=1}^{3} c_i \frac{\partial q_i}{\partial \tilde{q}_\rho}, \quad \rho = 1,2,3, \]

Then \( c_\rho \) and \( \tilde{c}_\rho \) are called the 'components of a covariant vector'.

### 8.1.2.3 The Derivation of the General Metric

If \( q^i = q^j(x^j) \) and \( x^i = x^i(q^j) \), \( j = 1,2,3 \), where the superscripts are now labels rather than exponents, let

\[ |r\rangle = x^1|i_1\rangle + x^2|i_2\rangle + x^3|i_3\rangle, \quad \text{where} \ |i_j\rangle = 1, \quad j = 1,2,3 \]

Therefore,

\[ d|r\rangle = \left( \frac{\partial x^1}{\partial q^1} |i_1\rangle + \frac{\partial x^2}{\partial q^1} |i_2\rangle + \frac{\partial x^3}{\partial q^1} |i_3\rangle \right) dq^1 + \left( \frac{\partial x^1}{\partial q^2} |i_1\rangle + \frac{\partial x^2}{\partial q^2} |i_2\rangle + \frac{\partial x^3}{\partial q^2} |i_3\rangle \right) dq^2 + \left( \frac{\partial x^1}{\partial q^3} |i_1\rangle + \frac{\partial x^2}{\partial q^3} |i_2\rangle + \frac{\partial x^3}{\partial q^3} |i_3\rangle \right) dq^3 \]

Simplifying, let

\[ |e_1\rangle = \left( \frac{\partial x^1}{\partial q^1} |i_1\rangle + \frac{\partial x^2}{\partial q^1} |i_2\rangle + \frac{\partial x^3}{\partial q^1} |i_3\rangle \right), \quad |e_2\rangle = \left( \frac{\partial x^1}{\partial q^2} |i_1\rangle + \frac{\partial x^2}{\partial q^2} |i_2\rangle + \frac{\partial x^3}{\partial q^2} |i_3\rangle \right), \quad |e_3\rangle = \left( \frac{\partial x^1}{\partial q^3} |i_1\rangle + \frac{\partial x^2}{\partial q^3} |i_2\rangle + \frac{\partial x^3}{\partial q^3} |i_3\rangle \right) \]

\[ |e_1\rangle = \frac{\partial |r\rangle}{\partial q^1}, \quad |e_2\rangle = \frac{\partial |r\rangle}{\partial q^2}, \quad |e_3\rangle = \frac{\partial |r\rangle}{\partial q^3} \]
So,
\[ d|r\rangle = dq^1|e_1\rangle + dq^2|e_2\rangle + dq^3|e_3\rangle \]

The vector \(d|r\rangle\) is said to be a vector with contravariant components \(dq^i, i = 1,2,3\). Let
\[
(ds)^2 = \langle dr|dr\rangle
= dq^1dq^1(e_1|e_1) + dq^1dq^2(e_1|e_2) + dq^1dq^3(e_1|e_3) + dq^2dq^1(e_2|e_1)
+ dq^2dq^2(e_2|e_2) + dq^2dq^3(e_2|e_3) + dq^3dq^1(e_3|e_1) + dq^3dq^2(e_3|e_2)
+ dq^3dq^3(e_3|e_3),
\]

where \(\langle e_i|e_j\rangle\) does not necessarily equal zero if \(i \neq j\). To simplify matters even further, let \(g_{ij} = \langle e_i|e_j\rangle\), then
\[
(ds)^2 = \langle dr|dr\rangle = \sum_{i=1}^{3} \sum_{j=1}^{3} g_{ij} dq^i dq^j
\]

The expression \(ds\) called the ‘generalized metric’ or ‘line element’ because it measures distance along a curve in any given set of coordinates.

8.2 Configuration Space

Given the line element
\[
(ds)^2 = \langle dr|dr\rangle = \sum_{i=1}^{3} \sum_{j=1}^{3} g_{ij} dq^i dq^j,
\]

\(g_{ij}\) represents a more complicated expression i.e.
\[
g_{ij} = \frac{\partial x}{\partial dq^i} \frac{\partial x}{\partial dq^j} + \frac{\partial y}{\partial dq^i} \frac{\partial y}{\partial dq^j} + \frac{\partial z}{\partial dq^i} \frac{\partial z}{\partial dq^j} = g_{ji}
\]

The coordinates \(q_i, i = 1,2,3\) represent the position of a single particle in three dimensional space. If there are \(n\) particles, then there are \(3n\) such coordinates. For example, suppose there are two particles ‘a’ and ‘b’, one constrained to move along the \(x\)-axis, the other along the \(y\)-axis. Instead of representing their positions as two points in one-dimensional space, the same information can be represented symbolically by one point ‘c’ in two-dimensional space (see figure 8.2-1). The path that \(c\) takes is not the path of the particle; but if the position of \(c\) is defined with respect to time, the motions of \(a\) and \(b\) are also defined [224].
The same idea can be extended to $n$ particles. Instead of following the histories of $n$ points in three dimensions, an assembly of particles can be described by one point in $3n$ dimensions. Such a $3n$-dimensional space is called a ‘configuration space’. The $n$-body problem is then reduced to defining a single trajectory [224].

The configuration of a mechanical system is composed of a number or vector consisting of values of a set of independent coordinates that completely describe the system, where none of the numbers in a set of coordinates ‘$\{q_i(t)\}$’ is dependent on the others. The coordinates in the configuration can be independently changed and each set of values describes something physically different. Given a set of values of the coordinates at some particular time, the configuration describes what the system looks like at that time. A configuration is a mathematical snapshot of the system at a given time. Problems in classical mechanics can be solved by knowing how the configuration of the system evolves with time because the configuration provides a mathematical motion picture that describes how the system moves.

8.2.1 Degrees of Freedom

The number of independent components ‘$q_i$’ of a configuration is called the ‘degrees of freedom’, the number of independent ways in which the system can move. In the two particle example above, the configuration space has two degrees of freedom.

8.3 The Fundamental Principle of Classical Mechanics

The totality of classical mechanics is predicated on the following:

*The motion of a mechanical system is completely determined if the configuration of the $q_i$’s and the velocities ‘$\dot{q}_i = dq_i/dt$’ are given at any time.*

8.3.1 Generalized Velocity

Consider the $i^{th}$ particle. Suppose the system of particles conforms to

$$x_i = x_i(q_1, q_2, ..., q_{3n}), \quad y_i = y_i(q_1, q_2, ..., q_{3n}), \quad z_i = z_i(q_1, q_2, ..., q_{3n}),$$
then

\[
\frac{dx_i}{dt} = \dot{x}_i = \frac{\partial x_i}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial x_i}{\partial q_2} \frac{dq_2}{dt} + \cdots + \frac{\partial x_i}{\partial q_{3n}} \frac{dq_{3n}}{dt}
\]

So,

\[
\dot{x}_i = \sum_{i=1}^{3n} \frac{\partial x_i}{\partial q_i} \dot{q}_i, \quad \dot{y}_i = \sum_{i=1}^{3n} \frac{\partial y_i}{\partial q_i} \dot{q}_i, \quad \dot{z}_i = \sum_{i=1}^{3n} \frac{\partial z_i}{\partial q_i} \dot{q}_i
\]

The \( \dot{q}_i \)'s are called the 'generalized velocities'. The products \( (\partial x_i/\partial q_i)\dot{q}_i \) etc. must have dimensions \([\text{L} \cdot \text{T}^{-1}]\).

### 8.3.2 Generalized Force and Kinetic Energy

Note that

\[
\frac{\partial \dot{x}_i}{\partial \dot{q}_i} = \frac{\frac{\partial x_i}{dt}}{\partial \dot{q}_i} = \frac{\partial x_i}{\partial q_i} \frac{dt}{\partial \dot{q}_i} = \frac{\partial x_i}{\partial q_i},
\]

which allows the formation of a 'generalized force'. If \( V \) is the 'potential', then

\[
F_j = -\frac{\partial V}{\partial q_j} = -\sum_{i=1}^{n} \left( \frac{\partial V}{\partial x_i} \frac{\partial x_i}{\partial q_j} + \frac{\partial V}{\partial y_i} \frac{\partial y_i}{\partial q_j} + \frac{\partial V}{\partial z_i} \frac{\partial z_i}{\partial q_j} \right), \quad V = V(q_j)
\]

The total kinetic energy \( T \) of the configuration is

\[
T = \frac{1}{2} \sum_{i=1}^{n} m_i (\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2),
\]

where \( m_i \) is the 'mass' of the \( i \)th particle. Form the derivatives

\[
\frac{\partial T}{\partial \dot{x}_i} = m_i \dot{x}_i, \quad \frac{\partial T}{\partial \dot{y}_i} = m_i \dot{y}_i, \quad \frac{\partial T}{\partial \dot{z}_i} = m_i \dot{z}_i, \quad \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}_i} \right) = m_i \ddot{x}_i, \quad \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{y}_l} \right) = m_i \ddot{y}_l,
\]

let \( F_{xi} = m_i \ddot{x}_i, \ F_{yi} = m_i \ddot{y}_i, \ F_{zi} = m_i \ddot{z}_i \), then

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}_i} \right) = F_{xi} = -\frac{\partial V}{\partial x_i} + \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}_i} \right) + \frac{\partial V}{\partial x_i} = 0
\]
Similarly,

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial y} \right) = F_{yi} = -\frac{\partial V}{\partial y_i} \rightarrow \frac{d}{dt} \left( \frac{\partial T}{\partial y} \right) + \frac{\partial V}{\partial y_i} = 0,
\]

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial z} \right) = F_{zi} = -\frac{\partial V}{\partial z_i} \rightarrow \frac{d}{dt} \left( \frac{\partial T}{\partial z} \right) + \frac{\partial V}{\partial z_i} = 0
\]

Therefore, there are three equations for each particle.

8.3.3 The Principle of Least Action

Recalling that

\[
F_j = -\frac{\partial V}{\partial q_j} = -\sum_{i=1}^{n} \left( \frac{\partial V}{\partial x_i} \frac{\partial x_i}{\partial q_i} + \frac{\partial V}{\partial y_i} \frac{\partial y_i}{\partial q_i} + \frac{\partial V}{\partial z_i} \frac{\partial z_i}{\partial q_i} \right),
\]

then, for example,

\[
\frac{d}{dt} \frac{\partial T}{\partial x} = -\frac{\partial V}{\partial x_i}
\]

By substitution

\[
F_j = \sum_{i=1}^{n} \left( \frac{\partial x_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial x} + \frac{\partial y_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial y} + \frac{\partial z_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial z} \right) = -\frac{\partial V}{\partial q_j}
\]

\[
\rightarrow \sum_{i=1}^{n} \left( \frac{\partial x_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial x} + \frac{\partial y_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial y} + \frac{\partial z_i}{\partial q_j} \frac{d}{dt} \frac{\partial T}{\partial z} \right) + \frac{\partial V}{\partial q_j} = 0
\]

Again, recalling that

\[
\dot{x}_i = \sum_{i=1}^{3n} \frac{\partial x_i}{\partial q_i} \dot{q}_i, \quad \dot{y}_i = \sum_{i=1}^{3n} \frac{\partial y_i}{\partial q_i} \dot{q}_i, \quad \dot{z}_i = \sum_{i=1}^{3n} \frac{\partial z_i}{\partial q_i} \dot{q}_i
\]

then, if the equations above are differentiated with respect to \( q_k \), then

\[
\frac{\partial \dot{x}_i}{\partial q_k} = \sum_{i=1}^{3n} \frac{\partial^2 x_i}{\partial q_i \partial q_k} \dot{q}_i = \sum_{i=1}^{3n} \frac{\partial}{\partial q_i} \left( \frac{\partial x_i}{\partial q_k} \right) \dot{q}_i = \frac{d}{dt} \left( \frac{\partial x_i}{\partial q_k} \right).
\]

Similarly,

\[
\frac{\partial \dot{y}_i}{\partial q_k} = \sum_{i=1}^{3n} \frac{\partial^2 y_i}{\partial q_i \partial q_k} \dot{q}_i = \sum_{i=1}^{3n} \frac{\partial}{\partial q_i} \left( \frac{\partial y_i}{\partial q_k} \right) \dot{q}_i = \frac{d}{dt} \left( \frac{\partial y_i}{\partial q_k} \right).
\]
\[ \frac{\partial z_i}{\partial q_k} = \sum_{i=1}^{3n} \frac{\partial^2 z_i}{\partial q_i \partial q_k} \frac{\dot{q}_i}{\partial q_k} = \sum_{i=1}^{3n} \frac{\partial}{\partial q_i} \left( \frac{\partial z_i}{\partial q_k} \right) \dot{q}_i = \frac{d}{dt} \left( \frac{\partial z_i}{\partial q_k} \right) \]

By the product rule for differentiation

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial x_i} \right) = \frac{\partial x_i}{\partial q_j} \frac{d T}{\partial \dot{x}_i} \frac{dt}{\partial \dot{x}_i} \frac{dt}{\partial q_j} + \frac{\partial T}{\partial x_i} \frac{d x_i}{\partial q_j} \frac{dt}{\partial q_j} \]

Hence,

\[ \sum_{i=1}^{n} \left( \frac{\partial x_i}{\partial q_j} \frac{d T}{\partial \dot{x}_i} \frac{dt}{\partial \dot{x}_i} + \frac{\partial T}{\partial x_i} \frac{d x_i}{\partial q_j} \frac{dt}{\partial q_j} \right) + \frac{d V}{\partial q_j} = 0 \]

becomes

\[ \sum_{i=1}^{n} \left[ \frac{d}{dt} \left( \frac{\partial T}{\partial x_i} \right) \frac{d T}{\partial \dot{x}_i} \frac{dt}{\partial \dot{x}_i} + \frac{\partial T}{\partial x_i} \frac{d x_i}{\partial q_j} \frac{dt}{\partial q_j} \right] + \frac{d V}{\partial q_j} = 0 \]

Now

\[ T = \frac{1}{2} \sum_{i=1}^{n} m_i (\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2) \rightarrow \frac{\partial T}{\partial \dot{q}_j} = \sum_{i=1}^{n} \left( \frac{\partial T}{\partial \dot{x}_i} \frac{d \dot{x}_i}{\partial \dot{q}_j} + \frac{\partial T}{\partial \dot{y}_i} \frac{d \dot{y}_i}{\partial \dot{q}_j} + \frac{\partial T}{\partial \dot{z}_i} \frac{d \dot{z}_i}{\partial \dot{q}_j} \right) \frac{\partial T}{\partial \dot{q}_j} \\
= \sum_{i=1}^{n} \left( \frac{\partial T}{\partial \dot{x}_i} \frac{d \dot{x}_i}{\partial \dot{q}_j} + \frac{\partial T}{\partial \dot{y}_i} \frac{d \dot{y}_i}{\partial \dot{q}_j} + \frac{\partial T}{\partial \dot{z}_i} \frac{d \dot{z}_i}{\partial \dot{q}_j} \right) \frac{\partial T}{\partial \dot{q}_j} \]

So then

\[ \sum_{i=1}^{n} \left[ \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}_i} \right) \frac{d \dot{x}_i}{\partial \dot{q}_j} + \frac{\partial T}{\partial \dot{x}_i} \frac{d \dot{x}_i}{\partial \dot{q}_j} \frac{dt}{\partial \dot{q}_j} \right] + \frac{d V}{\partial q_j} = 0 \]

can be written

\[ \frac{d}{dt} \left( \frac{\partial T}{\partial q_j} \right) - \frac{\partial T}{\partial q_j} + \frac{d V}{\partial q_j} = 0 \rightarrow \frac{d}{dt} \left( \frac{\partial T}{\partial q_j} \right) - \frac{\partial (T - V)}{\partial q_j} = 0 \]

Let \( L = T - V \) and regard \( V \) as function of the coordinates and not of the velocities. Note that \( \partial L/\partial q_i = -V \), since \( T \) is a function of the velocities only. And \( \partial L/\partial \dot{q}_i = T \), since \( V \) is a function of position only. Therefore, the equation above becomes

\[ \frac{d}{dt} \left( \frac{\partial L}{\partial q_j} \right) - \frac{\partial L}{\partial q_j} = 0 \]

There are three sets of the equations for each particle. Since the equation above was deduced from
\[
\frac{d}{dt} \frac{\partial T}{\partial \dot{x}} = m_i \ddot{x}_i, \quad \frac{d}{dt} \frac{\partial T}{\partial \dot{y}} = m_i \ddot{y}_i, \quad \frac{d}{dt} \frac{\partial T}{\partial \dot{z}} = m_i \ddot{z}_i,
\]

it is entirely equivalent to Newton's second law of motion.

Recalling that
\[
\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0
\]

is the Euler-Lagrange equation of least action (see Book II, Chapter 17, Sec. 17.1.1) [224], then

*The actual path a particle follows between two points in the time interval \('t_1 < t < t_2'\)* is given by the action

\[
S = \int_{t_1}^{t_2} L \, dt,
\]

which is stationary along the actual path.

Therefore, every particle or body in classical mechanics follows the path of least action, that is, bodies comply with \(|F| = m|a|\), Newton’s second law.

### 8.4 The Hamiltonian Formalism of Classical Mechanics

There is another approach to analytical Mechanics called the ‘Hamiltonian formalism’. Let

\[
H = \sum_{i=1}^{3n} p_i \dot{q}_i - L(q_1, q_2, ..., q_{3n}; \dot{q}_1, \dot{q}_2, ..., \dot{q}_{3n}) \rightarrow dH
\]

\[
= \sum_{i=1}^{3n} \left( p_i \dot{q}_i + \dot{q}_i dp_i - \frac{\partial L}{\partial q_i} dq_i - \frac{\partial L}{\partial \dot{q}_i} d\dot{q}_i \right)
\]

Note that

\[
p_i = \frac{\partial L}{\partial \dot{q}_i}, \quad \dot{p}_i = \frac{\partial L}{\partial q_i},
\]

since

\[
\frac{\partial L}{\partial \dot{q}_i} = \frac{\partial (T - V)}{\partial \dot{q}_i} = \frac{\partial T}{\partial \dot{q}_i} = \frac{\partial}{\partial \dot{q}_i} \left( \frac{1}{2} m_i \dot{q}_i^2 \right) = m_i \dot{q}_i = p_i
\]

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = 0 \rightarrow \frac{d}{dt} \left( \frac{\partial L}{\partial q_i} \right) = \frac{\partial L}{\partial q_i} \rightarrow \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_i} \right) = \dot{p}_i = \frac{\partial L}{\partial q_i}, \quad \frac{\partial V}{\partial \dot{q}_i} = 0
\]
Therefore,
\[
\frac{dH}{dt} = \sum_{i=1}^{3n} \left( p_i \dot{q}_i + \dot{q}_i d_{q_i} - \frac{\partial L}{\partial q_i} d_{q_i} - \frac{\partial L}{\partial \dot{q}_i} d_{\dot{q}_i} \right) = \sum_{i=1}^{3n} (p_i \dot{q}_i + \dot{q}_i d_{q_i} - \dot{p}_i d_{q_i} - p_i d_{\dot{q}_i})
\]

By direct differentiation
\[
\frac{dH}{dt} = \sum_{i=1}^{3n} \left( \frac{\partial H}{\partial p_i} d_{p_i} + \frac{\partial H}{\partial q_i} d_{q_i} \right)
\]

Now
\[
\sum_{i=1}^{3n} (\dot{q}_i d_{p_i} - \dot{p}_i d_{q_i}) = \sum_{i=1}^{3n} \left( \frac{\partial H}{\partial p_i} d_{p_i} + \frac{\partial H}{\partial q_i} d_{q_i} \right) \rightarrow \frac{\partial H}{\partial p_i} = \dot{q}_i, \quad \frac{\partial H}{\partial q_i} = -\dot{p}_i
\]

Hence,
\[
\frac{dH}{dt} = \frac{\partial}{\partial t} \left( \sum_{i=1}^{3n} (\dot{q}_i d_{p_i} - \dot{p}_i d_{q_i}) \right) = \sum_{i=1}^{3n} (\dot{q}_i \dot{p}_i - \dot{p}_i \dot{q}_i) = 0 \rightarrow H = a \text{ constant}
\]

Hence, \( H = a \text{ constant} \) in a conservative system [224].

In rectangular coordinates,
\[
T = \frac{1}{2} \sum_{i=1}^{n} m_i (\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2)
\]

But
\[
\dot{x}_i = \sum_{i=1}^{3n} \frac{\partial x_i}{\partial q_i} \dot{q}_i, \quad \dot{y}_i = \sum_{i=1}^{3n} \frac{\partial y_i}{\partial q_i} \dot{q}_i, \quad \dot{z}_i = \sum_{i=1}^{3n} \frac{\partial z_i}{\partial q_i} \dot{q}_i
\]

Substituting the values for \( \dot{x}_i, \dot{y}_i, \dot{z}_i \) into \( T \) leaves
\[
T = \frac{1}{2} \sum_{j=1}^{3n} \sum_{k=1}^{3n} A_{jk} \dot{q}_j \dot{q}_k, \quad A_{jk} = \frac{1}{2} \sum_{i=1}^{n} m_i \left( \frac{\partial x_i}{\partial q_j} \frac{\partial x_i}{\partial q_k} + \frac{\partial y_i}{\partial q_j} \frac{\partial y_i}{\partial q_k} + \frac{\partial z_i}{\partial q_j} \frac{\partial z_i}{\partial q_k} \right)
\]
The terms in the summation for which \( j = k \) are to be counted only once. Hence,

\[
\frac{\partial T}{\partial \dot{q}_j} = 2 \sum_k A_{jk} \dot{q}_k.
\]

The factor ‘2’ appearing because each term of the summation is counted twice when \( j \neq k \). If \( j = k \), the resulting single term is \( \dot{q}_i^2 \), which, when differentiated, again introduces a factor of \( 2 \). Hence,

\[
\sum_l p_l \dot{q}_l = \sum_l \frac{\partial L}{\partial \dot{q}_i} \dot{q}_l = \sum_l \frac{\partial T}{\partial \dot{q}_i} \dot{q}_l = 2 \sum_k \sum_j A_{jk} \dot{q}_i \dot{q}_k = 2T \rightarrow H = 2T - L = 2T - (T - V)
\]

\[
= T + V = a \text{ constant,}
\]

the total energy of the system \([224]\). If the system is not conservative, \( H \) no longer represents the total energy, but is merely a function of the coordinates as expressed in

\[
H = \sum_{i=1}^{3n} p_i \dot{q}_i - L(q_1, q_2, ..., q_{3n}; \dot{q}_1, \dot{q}_2, ..., \dot{q}_{3n})
\]

### 8.5 Poisson Brackets

Suppose that \( u = u(p_i, q_i) \) and \( v = v(p_i, q_i) \) are continuous and differential functions of the coordinates and momenta. Let

\[
[u, v] = \sum_{i=1}^{3n} \frac{\partial u}{\partial p_i} \frac{\partial v}{\partial q_i} - \frac{\partial v}{\partial p_i} \frac{\partial u}{\partial q_i}
\]

The expression above is called the ‘Poisson bracket’, named for a French mathematician, geometer, and physicist Siméon Denis Poisson (21 June 1781 – 25 April 1840).

Note that

\[
\dot{u} = \sum_{i=1}^{3n} \frac{\partial u}{\partial q_i} \dot{q}_i + \frac{\partial u}{\partial p_i} \dot{p}_i = \sum_{i=1}^{3n} \frac{\partial u}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial u}{\partial p_i} \frac{\partial H}{\partial q_i} = [H, u], \quad \frac{\partial H}{\partial p_i} = \dot{q}_i, \quad \frac{\partial H}{\partial q_i} = -\dot{p}_i
\]

Let \( u = q_i \), then \( \dot{q}_i = [H, q_i] \) and if \( u = p_i \), then \( \dot{p}_i = [H, p_i] \). The following relationships should be noted:

\[
[q_i, q_k] = [p_i, p_k] = 0, \quad i \neq k, \quad [p_i, q_k] = \delta_{ik}
\]

Hence, \([p_i, q_i] = 1 \) and \([q_i, p_i] = -1 \). To see this, note that

\[
[p_i, q_k] = \sum_{i=1}^{3n} \frac{\partial p_j}{\partial q_i} \frac{\partial q_k}{\partial q_i} - \frac{\partial q_k}{\partial q_i} \frac{\partial p_j}{\partial q_i}
\]
The first factor in the first term of the sum is 1. Since the coordinates are independent, \( \frac{\partial q_k}{\partial q_i} = 0 \), unless \( k = i \) in which case \( \frac{\partial q_k}{\partial q_k} = 1 \), which happens in only one term of the summation. Hence, for all the terms of the sum

\[
\frac{\partial p_i}{\partial q_k} \frac{\partial q_k}{\partial p_i} = 0
\]

except for the \( k^{th} \) term, where

\[
\frac{\partial p_i}{\partial q_k} \frac{\partial q_k}{\partial q_k} = 1
\]

The factor ‘\( \partial q_k / \partial p_i \)’ is zero for all the terms of the sum, since the derivative of a coordinate with respect to a velocity is zero. Using the same line of reasoning, the factor ‘\( \partial p_i / \partial q_i \)’ is zero for all the terms in the sum. Therefore, \([q_i, p_k] = \delta_{ik}\). A similar line of reasoning will show that \([q_k, p_i] = -\delta_{ik}\). The Poisson brackets have limited use in classical mechanics, but have found great utility in quantum mechanics.

### 8.5.1 The Rules for Poisson Brackets

The following rules for Poisson brackets listed in table 8.5.1 hold universally.

<table>
<thead>
<tr>
<th>Rules for Poisson Brackets, where ((f, g, h)) are analytic functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Constant Function Rule</td>
</tr>
<tr>
<td>Anti-commutativity</td>
</tr>
<tr>
<td>Distributive Property</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Product Rule</td>
</tr>
<tr>
<td>Jacobi Identity</td>
</tr>
</tbody>
</table>

| Table 8.5-1 |

### 8.5.2 Canonically Conjugate Variables

Variables ‘\( u_i \)’ and ‘\( v_i \)’ that comply with the conditions ‘\([u_i, u_k] = 0 = [v_i, v_k], \quad i \neq k\), \([v_i, u_k] = \delta_{ik}\)’ are called ‘canonically conjugate’.

Consider two sets of functions ‘\( P_k \)’ and ‘\( Q_k \)’ that are related to the dynamical variables ‘\( p_i \)’ and ‘\( q_i \)’ so that

\[
\begin{align*}
P_k &= P_k(p_1, p_2, ..., p_{3n}, q_1, q_2, ..., q_{3n}), & \quad Q_k &= Q_k(p_1, p_2, ..., p_{3n}, q_1, q_2, ..., q_{3n}), \\
p_i &= p_i(P_1, P_2, ..., P_{3n}, Q_1, Q_2, ..., Q_{3n}), & \quad q_i &= q_i(P_1, P_2, ..., P_{3n}, Q_1, Q_2, ..., Q_{3n})
\end{align*}
\]
From the time derivatives:
\[
\frac{dP_k}{dt} = \dot{P}_k = \sum_i \left( \frac{\partial P_k}{\partial q_i} \dot{q}_i + \frac{\partial P_k}{\partial p_i} \dot{p}_i \right) = \sum_i \left( \frac{\partial P_k}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial P_k}{\partial p_i} \frac{\partial H}{\partial q_i} \right) \rightarrow \dot{P}_k = [H, P_k],
\]
\[\dot{Q}_k = [H, Q_k]\]

By direct differentiation,
\[
\frac{\partial H}{\partial p_i} = \sum_i \left( \frac{\partial H}{\partial p_i} \frac{\partial p_i}{\partial p_i} + \frac{\partial H}{\partial q_i} \frac{\partial q_i}{\partial p_i} \right), \quad \frac{\partial H}{\partial q_i} = \sum_i \left( \frac{\partial H}{\partial p_i} \frac{\partial p_i}{\partial q_i} + \frac{\partial H}{\partial q_i} \frac{\partial q_i}{\partial q_i} \right)
\]

Recalling that
\[
\dot{P}_k = \sum_i \left( \frac{\partial P_k}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial P_k}{\partial p_i} \frac{\partial H}{\partial q_i} \right),
\]
then
\[
\dot{P}_k = \sum_i \left( \frac{\partial P_k}{\partial q_i} \left( \sum_j \frac{\partial H}{\partial p_j} \frac{\partial p_j}{\partial p_j} - \frac{\partial H}{\partial q_j} \frac{\partial q_j}{\partial p_j} \right) - \frac{\partial P_k}{\partial p_i} \left( \sum_j \frac{\partial H}{\partial q_j} \frac{\partial q_j}{\partial p_j} - \frac{\partial H}{\partial q_i} \frac{\partial q_i}{\partial q_i} \right) \right) \rightarrow \dot{P}_k
\]
\[
= \sum_i \frac{\partial H}{\partial p_i} \sum_j \left( \frac{\partial P_k}{\partial p_j} \frac{\partial P_k}{\partial q_j} - \frac{\partial P_k}{\partial p_j} \frac{\partial P_k}{\partial q_j} \right) + \sum_i \frac{\partial H}{\partial q_i} \sum_j \left( \frac{\partial Q_k}{\partial p_j} \frac{\partial Q_k}{\partial q_j} - \frac{\partial Q_k}{\partial p_j} \frac{\partial Q_k}{\partial q_j} \right) \rightarrow \dot{P}_k
\]
\[
= [H, P_k] = \sum_i \left( \frac{\partial H}{\partial p_i} [P_i, P_k] - \frac{\partial H}{\partial q_i} [P_k, Q_i] \right), \quad [P_i, P_k] = 0
\]

If the \(p_i\)'s are canonically conjugate to the \(q_i\)'s, then the \(P_i\)'s are canonically conjugate to the \(Q_i\)'s. Hence,
\[
[P_i, Q_k] = \delta_{ik}, \quad \dot{P}_k = [H, P_k] = -\frac{\partial H}{\partial Q_k}, \quad \dot{Q}_k = [H, Q_k] = \frac{\partial H}{\partial P_k}
\]

### 8.5.2.1 The Idea of a Symmetry

However, the equations \(\dot{P}_k = [H, P_k], \dot{Q}_k = [H, Q_k]\) hold generally. Hence, if \([H, P_k] = 0\), then either \(P_k = 0\) or \(P_k = \text{a constant}\). A dynamic variable is a constant during the motion of the system if and only if it commutes with the Hamiltonian \([224]\).

For instance, consider \([H, L_z]\), where \(L_z\) is the \(x\)-component of angular momentum for a particle of mass 'm'. By the usual definition of the curl, \(L_z = yp_z - zp_y\) and

\[
H = T + V = \frac{1}{2m} \left( p_x^2 + p_y^2 + p_z^2 \right) + V = \frac{\partial V}{\partial y} = \frac{\partial H}{\partial y}, \quad \frac{\partial H}{\partial z} = \frac{\partial V}{\partial z},
\]

since \(T\) is a function of the velocities and not of the coordinates. If the number of particles is 1, then
\[ [H, L_x] = y \frac{\partial V}{\partial z} - z \frac{\partial V}{\partial y} \]

Hence, \( H \) and \( L_x \) do not commute unless \([H, L_x] = 0\). The left-side of the equation above is the \( x \)-component of \( \langle r \times F \rangle \) i.e. the torque around the \( x \)-axis. Therefore, \( H \) and \( L_x \) commute only when the torque vanishes and the angular momentum is constant, which is consistent with the definition of the conservation of angular momentum. Moreover, if a quantity is conserved during the motion of the system, then it is associated with a symmetry of the system, by Noether’s theorem. This suggests that all the dynamical variables of a system that commute with the Hamiltonian represent symmetries of the system, and hence, are conserved quantities.

The Poisson bracket is primarily used in quantum mechanics to investigate the relationship between dynamical variables [224]. It represents a transition from classical to quantum mechanics.

**8.6 Concluding Remarks**

Analytical mechanics takes advantage of a system’s constraints to solve problems. The constraints limit the degrees of freedom of a system, reducing the number of coordinates needed to solve for the motion of a particle or systems of particles. The formalism is well suited to arbitrary choices of coordinates, known as ‘generalized coordinates’. The kinetic and potential energies are expressed in these generalized coordinates to formulate the equations of motion. In most cases, analytical mechanics solves problems more efficiency than vectorial methods. It is not suitable for solving problems where the forces involved are non-conservative, in which case, a reversion to Newtonian mechanics or some other reasonable facsimile is necessary.

The two dominant branches of analytical mechanics are 1) ‘Lagrangian mechanics’ (using generalized coordinates and corresponding generalized velocities in configuration space) and 2) ‘Hamiltonian mechanics’ (using coordinates and corresponding momenta in phase space). Both formulations are equivalent, containing the same information for describing the dynamics of a system. Analytical mechanics does not introduce new physics and is not more general than Newtonian mechanics. Rather, it is a collection of equivalent formalisms which have broad application. In fact, the same principles and formalisms can be used in relativistic mechanics, and, with some modification, quantum mechanics and quantum field theory. The methods of analytical mechanics apply to discrete particles, each with a finite number of degrees of freedom, and can be modified to describe continuous fields or fluids, which have infinite degrees of freedom.

Generally, Newtonian mechanics uses all three Cartesian coordinates or other 3\( D \) coordinate systems to describe the motion of a body. Most physical systems have constraints that prevent a body from taking certain directions and pathways. Specifying a full set of Cartesian coordinates is often unnecessary, as the constraints determine the relations among the coordinates. Such relations can be modeled by equations corresponding to the constraints. In the Lagrangian and Hamiltonian formalisms, the constraints are incorporated into the system’s geometry, reducing the number of
coordinates to the minimum needed to model the motion of a system. These are known as 'generalized coordinates', denoted \( q_i (i = 1, 2, 3...) \) [5].

Generalized coordinates incorporate constraints on the system. There is one generalized coordinate \( q_i \) for each degree of freedom i.e. each way the system can change its configuration, such as curvilinear lengths or angles of rotation. Generalized coordinates are not the same as curvilinear coordinates. The number of curvilinear coordinates equals the dimension of the position space (usually 3D space), while the number of generalized coordinates does not necessarily equal this dimension; constraints can reduce the number of degrees of freedom (hence the number of generalized coordinates required to define the configuration of the system), following the general rule [262]:

\[
\text{[dimension of position space (usually 3)] × [number of constituents of the system ("particles")]} - \text{[number of constraints]} = \text{[number of degrees of freedom]} = \text{[number of generalized coordinates]}
\]

All dynamical variables can be derived from a position \('r'\), momentum \('p'\), and time \('t'\) and written as a function of these i.e. \( A = A(q, p, t) \). If \( A(q, p, t) \) and \( B(q, p, t) \) are two scalar valued dynamical variables, the Poisson bracket is defined by the generalized coordinates and momenta:

\[
[p_i, q_k] = \sum_{i=1}^{3n} \frac{\partial p_i}{\partial q_i} \frac{\partial q_k}{\partial q_i} - \frac{\partial q_k}{\partial p_i} \frac{\partial p_i}{\partial q_i} = \delta_{ik}
\]

Calculating the total derivative of one of these, say \('A'\), and substituting Hamilton's equations into the result leads to the time evolution of \( A \):

\[
\frac{dA}{dt} = [A, H] + \frac{\partial A}{\partial t}
\]

This equation is closely related to the equation of motion in the Heisenberg picture of quantum mechanics, where classical dynamical variables become quantum operators and the Poisson bracket is replaced by the 'commutator' of operators via Dirac's canonical quantization.

All equations of motion for particles and fields, in any formalism, can be derived from the widely applicable result called the 'principle of least action'. One result is Noether's theorem, a statement which connects conservation laws to their associated symmetries.

A general way to find the equations of motion from the principle of least action is:

*The actual path a particle follows between two points in the time interval \('t_1 < t < t_2'\) is given by the action\)

\[
S = \int_{t_1}^{t_2} L \, dt,
\]

*which is stationary along the actual path,\)**
where the departure ‘$t_1$’ and arrival ‘$t_2$’ times are fixed. The term ‘path’ or ‘trajectory’ refers to the time evolution of the system through configuration space. In other words, $q(t)$ traces out a path. The path for which the action is stationary is the path taken by the system.

From this principle, all equations of motion in classical mechanics can be derived. This approach can be extended to fields and underlies the path integral formulation of quantum mechanics [60] and is used for calculating geodesic motion in general relativity [263].
Chapter 9
The Theory of Special Relativity

“When you are courting a nice girl an hour seems like a second. When you sit on a red-hot cinder a second seems like an hour. That’s relativity.”

- Albert Einstein

9.0 Introduction

This discussion of special relativity emphasizes its logical foundations and follows closely the lectures of Dr. Bob Eagle and Prof. Leonard Susskind of Stanford University. Readers, interested in a deeper analysis of the physics of special and general relativity are encouraged to view their excellent online presentations.

9.1 Wave and Particle Physics

The conditions that existed just before its publication made the discovery of the theory of special relativity almost inevitable. At the time, there were two mutually exclusive ways of describing physical phenomena: ‘particle-like’ and ‘wave-like’. Particles could travel through empty space. Waves were conceived as disturbances in a medium, such as water or air, where the particles in the medium vibrated. Unless the medium itself was moving, the particles in the medium maintained their relative positions, while the wave propagated through the medium, like a wave that propagates through a stretched sting or rope (see fig. 9-1).

![Figure 9-1](image)

9.1.1 The Motion of a Particle

The physics of waves and particles were different, even under similar experimental conditions, which justified the belief that waves and particles were, in some sense, independent notions. Consider the following thought experiment involving two observers ‘P₁’ and ‘P₂’. Suppose ‘P₁’ is strapped to the front of a train moving with a velocity ‘v₁’ and, at a certain time ‘t₁’, fires an arrow that travels with velocity ‘vₐ’, in the same direction as the train, at a target a distance ‘D’ away, the arrow reaching the
target at time ‘$t_2’$. If $P_2$, at rest relative to the train, fires an arrow at time ‘$t_1’ = t_1’ at the
same target a distance ‘$D’ away with velocity ‘$v_a’$, $P_2$’s arrow will reach the target at time
‘$t_2’$. Since $P_1$’s arrow travels at a velocity ‘$v_a + v_t’$, while $P_2$’s arrow travels with velocity
‘$v_a’’, then

$$t_2' - t_1' = \frac{D}{v_a} > t_2 - t_1 = \frac{D}{v_t + v_a},$$

as measured by $P_2$, who is at rest relative to the train. In this scenario, the train and the
arrow act like free particles, their velocities add.

9.1.2 The Motion of a Wave

If the same experiment is conducted with a sound wave, the results are different. The
velocity of a wave depends only on the medium through which it travels. The observer
‘$P_2’$, at rest relative to the train, would measure the sound wave velocity to be $v_w$, whether or not it was emitted from the moving train or from a source at rest relative to
the train. Unlike particle velocities, wave velocities do not add, unless the actual
medium through which the wave travels is moving relative to the observer. In this case,

$$t_2' - t_1' = t_2 - t_1 = \frac{D}{v_w},$$

as measured by $P_2$, who is at rest relative to the moving train.

The observer ‘$P_1’$, stationary relative to but riding on the train, would measure the speed
of the sound wave to be $v_w - v_t$, the velocity of the wave minus the velocity of the train.
If observer ‘$P_1’’ is considered at rest, then the relative speed of the medium is $-v_t$.
Wave motion (not additive) was thought to be entirely different than free particle motion
(additive), providing more evidence that waves and particles were distinct phenomena.

The belief, at the time, was that waves required a physical substance through which to
travel. Experiments had shown that sound waves, for example, could not travel through
empty space. And Maxwell had shown that light consisted of electromagnet waves,
which traveled through space at a constant speed ‘$c’ ($c ≈ 3 \times 10^{10} \text{[cm/sec]}$). That light
was a wave suggested that empty space was not really empty, but contained a physical
substance - a ‘luminous aether’ through which light traveled. The aether was supposed
stationary, and if so, the Earth’s motion through the aether would create an aether flow,
which could be detected.

9.2 The Michelson and Morley Experiment

Suppose a light ray travels at speed ‘$c’ from ‘A’ to ‘B’ and back again, parallel to the
aether flow. If the Earth travels through the aether at a speed ‘$v’ in the direction shown
in fig. 9.2-1, then the light ray moving against the aether flow has velocity $c - v$.

7This assumes that the medium through which the wave travels remains stationary relative to
the motion of the train.
Conversely, when the light ray moves in the same direction of the flow, its velocity is \( c + v \). If \( AB = AC = a \), the time it takes to go from \( A \) to \( B \) is \( a/(c - v) \) and the time it takes to go from \( B \) to \( A \) is \( a/(c + v) \). Hence, the total time \( t_{AB} \) is

\[
t_{AB} = \frac{a}{c - v} + \frac{a}{c + v} = \frac{2ac}{c^2 - v^2} = \frac{2a}{c} \left( \frac{1}{1 - \frac{v^2}{c^2}} \right),
\]

Figure 9.2-1

When the light ray is moving perpendicular to the flow, its velocity can be computed using the Pythagorean Theorem, as shown in fig. 9-2.2.

As the ray moves from \( A \) to \( C \), its velocity is \( \sqrt{c^2 - v^2} \). On the return trip, its velocity is also \( \sqrt{c^2 - v^2} \). So the time it takes to go from \( A \) to \( C \) is \( a/\sqrt{c^2 - v^2} \) and the time it takes to go from \( C \) to \( A \) is \( a/\sqrt{c^2 - v^2} \). Hence, the total time \( t_{AC} \) is

\[
t_{AC} = \frac{a}{\sqrt{c^2 - v^2}} + \frac{a}{\sqrt{c^2 - v^2}} = \frac{2a}{\sqrt{c^2 - v^2}} = \frac{2a}{c} \left( \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right).
\]
Note that $t_{AC} \neq t_{AB}$. In fact,

$$\frac{t_{AB}}{t_{AC}} = \frac{\frac{2ac}{c^2 - v^2}}{\frac{2a\sqrt{c^2 - v^2}}{\sqrt{c^2 - v^2}}} = \frac{2ac\sqrt{c^2 - v^2}}{2a(c^2 - v^2)} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad \Rightarrow \quad t_{AB} = \frac{t_{AC}}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Hence, if $0 < v < c$, then $1 - \frac{v^2}{c^2} < 1$. Therefore, $t_{AB} > t_{AC}$. Since the times are not equal, the motion of the Earth through the aether should be detectable by experiment.

Just such an experiment was conducted by American physicists Albert Michelson (1852-1931) and Edward Morley (1838-1923) in 1887. Suppose a light beam emanating from a source is split into two beams, where one of the beams is directed to a mirror along the path perpendicular to the aether flow, while the other beam covers a path parallel to the aether flow. The apparatus is constructed so that both beams return to the same viewing screen (detector) (see fig. 9.2-3 A).

Since light is a wave, if the path lengths of the two beams are equal, the beams will arrive at the screen in phase, interfere constructively yielding a bright field of view. But, if the aether flow is present, the beams have different transit times, no longer arriving in phase and will interfere destructively. If either of the beam paths varied in length, the interference fringes would move across the detector screen as reinforcement and cancellation of the waves succeed one another at each point (see fig. 9.2-3 B). If the apparatus remains stationary, nothing can be said about the time difference between the two paths. But if the apparatus is rotated by $90^\circ$, the two paths change their orientations relative to the aether flow. The beam formerly requiring time $t_{AB}$ to make the trip, now requires time $t_{AC}$ and vice versa. Since these times are different, the fringes will move across the screen during the rotation [45].

### 9.2.1 Calculating Fringe Shift

With the knowledge of wave refraction covered in Chapter 6, Sec 6.1, the shift in the fringes can be calculated. The time difference between the two paths owing to the aether flow is
\[ \Delta t = t_{AB} - t_{AC} = 2 \frac{L}{c} \left( \frac{1}{1 - \frac{v^2}{c^2}} \right) - 2 \frac{L}{c} \left( \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right), \]

where \( v \) is the speed of the Earth (\( v = 3 \times 10^4 \text{ [m/sec]} \)) through the aether and \( c \) is the speed of light (\( c \approx 3 \times 10^8 \text{ [m/sec]} \)). Since \( \frac{v^2}{c^2} \approx 10^{-8} \ll 1 \), by the binomial theorem

\[
(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)x^2}{2!} \pm \frac{n(n-1)(n-2)x^3}{3!} + \ldots
\]

If \( x = \frac{v^2}{c^2} \approx 10^{-8} \), then terms of order \( x^2 \) and higher in the binomial expansion add virtually nothing to the sum and can be discarded. Therefore, let

\[
(1 \pm x)^n \approx 1 \pm nx
\]

Since

\[
t_{AB} = \frac{2L}{c} (1 - \frac{v^2}{c^2})^{-1}, \quad t_{AC} = \frac{2L}{c} (1 - \frac{v^2}{c^2})^{-1/2},
\]

then the series expansion for \( t_{AB} \), where \( n = -1 \) and \( x = \frac{v^2}{c^2} \) is

\[
t_{AB} = \frac{2L}{c} (1 - \frac{v^2}{c^2})^{-1} \to \frac{2L}{c} \left( 1 - (-1) \frac{v^2}{c^2} \right) = \frac{2L}{c} \left( 1 + \frac{v^2}{c^2} \right)
\]

and the series expansion for \( t_{AC} \), where \( n = -1/2 \) and \( x = \frac{v^2}{c^2} \) is

\[
t_{AC} = \frac{2L}{c} (1 - \frac{v^2}{c^2})^{-1/2} \to \frac{2L}{c} \left( 1 - \left( -\frac{1}{2} \right) \frac{v^2}{c^2} \right) = \frac{2L}{c} \left( 1 + \frac{1}{2} \frac{v^2}{c^2} \right)
\]

Hence,

\[
\Delta t = t_{AB} - t_{AC} = \frac{2L}{c} \left( 1 + \frac{v^2}{c^2} \right) - \frac{2L}{c} \left( 1 + \frac{1}{2} \frac{v^2}{c^2} \right) = \frac{2L}{c} \left( \frac{v^2}{c^2} - \frac{1}{2} \frac{v^2}{c^2} \right) = \frac{L}{c} \left( \frac{v^2}{c^2} \right)
\]

The path difference 'd' corresponding to the time difference '\( \Delta t \)' is

\[
d = c\Delta t,
\]

since the waves travel at the speed of light 'c'. If \( d \) corresponds to the shifting of \( n \) fringes, then

\[
d = n\lambda,
\]

where \( \lambda \) is the wavelength of the light used in the experiment. Hence,

\[
c\Delta t = n\lambda \to n = \frac{c\Delta t}{\lambda} = \frac{Lv^2}{\lambda c^2}
\]
In the actual experiment, \( L = 10 \, [m] \), \( v = 3 \times 10^4 \, [m/sec] \), \( \lambda = 5 \times 10^{-7} \, [m] \), \( c = 3 \times 10^8 \, [m/sec] \). So,

\[
n = \frac{10 \, [m] \times (3 \times 10^4 \, [m/sec])^2}{5 \times 10^{-7} \, [m] \times (3 \times 10^8 \, [m/sec])^2} = 0.2
\]

Since both paths experience the fringe shift, the actual shift is \( 2n = 0.4 \). If the aether theory was correct, the rotation of the apparatus should show 0.4 of a fringe shift. While small, such a shift is readily detectable.

The experiment revealed no fringe shift. Later, more delicate experiments produced the same result - no fringe shift.

### 9.3 The Consequences of the Michelson and Morley Experiment

If the aether did exist, the results of the Michelson/Morley experiment suggested that Galilean relativity was violated. Since Galilean relativity was fundamental to Newtonian mechanics, and if incorrect, then the entire basis, upon which all the physics known at the time depended, would need reevaluation.

#### 9.3.1 Violation of Galilean Relativity

To see this, recall that in a reference frame \( 'S' \), if an event occurs at a location \( '(x, y, z)' \) at time \( 't' \) and if a reference frame \( 'S'' \) is moving relative to \( S \) with a constant velocity \( 'v' \), then the event viewed from \( S' \) happens at a different location \( '(x', y', z')' \). If the velocity of \( S' \) is restricted to the \(+x\)-direction and if the time in both systems is measured from the instant that the origins of \( S \) and \( S' \) coincide, then distance measurements made in the \( x \)-direction in \( S \) will exceed those in \( S' \) by the amount \( 'vt' \). The relationships between the coordinates in \( S \) and \( S' \) are given by

\[
x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t
\]

The velocities are computed by taking the derivatives with respect to time i.e.

\[
\frac{dx'}{dt'} = \frac{dx}{dt} - v, \quad \frac{dy'}{dt'} = \frac{dy}{dt}, \quad \frac{dz'}{dt'} = \frac{dz}{dt'}, \quad dt' = dt
\]

Galilean relativity predicts that someone at rest relative to the aether would measure the speed of light to be \( c \). Hence, \( dx/dt = c \). If \( dx'/dt' = c' \), someone moving with the aether flow would measure the speed of light as \( c' = c - v \), where \( v \) is velocity of the aether flow. But the failure of the Michelson/Morley to detect the aether suggested that \( c' = c \). Therefore, there was a direct contradiction between the principle of Galilean relativity and the results of Michelson/Morley experiment. And the great physics challenge of the day was to explain the discrepancy.

Either the aether did not exist or there was something Michelson and Morley overlooked. Neither of these scenarios portended a happy outcome. If the aether did exist, then the principle of ‘Galilean relativity’ was violated, which meant that the entire basis, upon which virtually all of physics relied, would require complete reformulation.
The belief at the time was that particles could move through empty space, but waves could not. And if a light wave could travel through empty space, just like a particle, then the distinction between wave-like and particle-like phenomena became less obvious. Of course, this is eventually what happened. The aether theory was ultimately abandoned. But the consequences of giving up the aether theory would not be fully appreciated until the development of quantum physics.

During this period many physicists, reluctant to concede that the principle of ‘Galilean relativity’ was wrong, suspected that Michelson and Morley overlooked something. That a light wave could travel through empty space without some mechanism of transportation, violating all previous experimental results involving wave phenomena, seemed, to any reasonable mind, improbable. Moreover, Galilean relativity had withstood the scrutiny of virtually all experimental tests, its violation seemed equally improbable.

9.3.2 The Lorentz Theory

An advocate of this view was the physicist H. A. Lorentz, whose solution to the problem was quite ingenious, but turned out to be deeply unsatisfactory [234]. He argued that the length of the arms of the apparatus used in the Michelson/Morley experiment shortened by a factor of \( L\sqrt{1 - v^2/c^2} \) when the arm pointed in the direction of the aether flow - that the molecules in the arm were shortened in that direction. The contraction was just enough to account for the absence of the fringe shift and explained why the two beams arrived at the detector simultaneously. The slower light beam didn’t travel as far. Lorentz also showed that his ideas were consistent with Maxwell’s theory. His solution preserved both the aether theory and Galilean relativity. The aether could change the dimensions of physical objects, although by what mechanism remained unexplained.

Upon this last point the critics pounced, particularly the French mathematician, theoretical physicist, engineer and a philosopher of science Jules Henri Poincaré (29 April 1854 – 17 July 1912), who argued that assuming the aether existed was superfluous. Those who supported the contraction theory provided no basis for believing that the contractions actually happened. An adequate response to this critique never materialized. But Poincaré was a friendly critic, who aided Lorentz in formulating the transformation equations that now bear his name. But the aether theory was destined for replacement by another explanation.

9.4 Einstein’s Principle of Relativity

The theory of special relativity, as formulated by Einstein, applies only to inertial systems - reference frames that move with a constant velocity relative to one another. If two reference frames ‘S’ and ‘S’’ move relative to one another along a direction parallel to the \( x \)-axis with relative velocity \( v \), then the principle of relativity states:

\[
\text{It is impossible to determine whether } S \text{ is at rest and } S' \text{ is moving in the } +x- \text{direction or } S' \text{ is at rest and } S \text{ moving in the } -x- \text{direction.}
\]
Both observers have an equal right to claim the “at rest” disposition while asserting the other observer is the one moving.

9.4.1 The Lorentz Transformations

Suppose there is a light source a distance ‘$x$’ from $S$ and $S'$. At time $t = t' = 0$, a flash of light emanates from the source. As the light wave travels through space, $S'$ travels a distance ‘$vt$’. The light wave reaches $S'$ after traveling a distance ‘$x'$’ and reaches $S$ after traveling a distance ‘$x$’ (see fig. 9.4.1-1), where ‘$t$’ is the time measured in the $S$ frame and ‘$t'$’ is the time measured in the $S'$ frame. From the figure,

$$x = x' + vt \rightarrow x' = x - vt$$

![Figure 9.4.1-1](image)

Now $x' = c't'$, where $c'$ specifies the speed of light as measured in the $S'$ frame. Hence,

$$x = x' + vt \rightarrow x = c't' + vt$$

Moreover, $x = ct$, where $c$ specifies the speed of light as measured in the $S$ frame. Therefore,

$$x = c't' + vt \rightarrow ct = c't' + vt \rightarrow (c - v)t = c't'$$

If $t = t'$, which is consistent with Galilean relativity, then $c - v = c'$. But the results of the Michelson/Morley experiment requires that $c = c'$. This suggests that $t \neq t'$. Note that the observer ‘$S''$’ calculates $x' = c't' \rightarrow c' = x'/t'$. The observer ‘$S'$’ calculates $x = ct \rightarrow c = x/t$. Since $c = c'$, one of the observers is measuring distance incorrectly, but which one?

The observer ‘$S''$’ asserts that $x = \gamma(x' + vt')$ and the observer ‘$S'$’ asserts that $x' = \gamma(x - vt)$, where $\gamma$ is the error that the observer ‘$S''$’ asserts that the observer ‘$S'$’ is making in the measurement of $x$; The factor ‘$\gamma$’ is also the error that the observer ‘$S'$’ asserts that the observer ‘$S''$’ is making in measuring $x'$. Note that

$$xx' = \gamma^2(x' + vt')(x - vt) = \gamma^2(xx' - x'vt + xvt' - v^2t't)$$

Remembering that $t' = x'/c$ and $t = x/c$, the equation above becomes

$$xx' = \gamma^2 \left( xx' - x'v \frac{x}{c} + xv \frac{x'}{c} - v^2 \frac{x'x}{c^2} \right)$$
Dividing through by $xx'$ gives

$$1 = \gamma^2 \left(1 - \frac{v}{c} + \frac{v^2}{c^2}\right) = \gamma^2 \left(1 - \frac{v^2}{c^2}\right) \Rightarrow \gamma^2 = \frac{1}{1 - \frac{v^2}{c^2}} \Rightarrow \gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Hence,

$$x = \gamma(x' + vt) = \frac{(x' + vt)}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad x' = \gamma(x - vt) = \frac{(x - vt)}{\sqrt{1 - \frac{v^2}{c^2}}}$$

To calculate the times,

$$x' = \gamma(x - vt) = \gamma \left(x - \frac{v}{c}x\right) \Rightarrow x' = \gamma \left(x - \frac{v}{c^2}x\right) \Rightarrow t' = \gamma \left(t - \frac{v}{c^2}x\right) = \frac{t - \frac{v}{c^2}x}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$x = \gamma(x' + vt') = \gamma \left(x' + \frac{v}{c}x'\right) \Rightarrow x = \gamma \left(x' + \frac{v}{c^2}x'\right) \Rightarrow t = \gamma \left(t' + \frac{v}{c^2}x'\right) = \frac{t' + \frac{v}{c^2}x'}{\sqrt{1 - \frac{v^2}{c^2}}}$$

The equations above for $x, x', t, t'$ are called the ‘Lorentz transformations’, which give the distances and times in terms of observers moving at a constant velocity relative to one another.

**9.4.1.1 Time Dilation**

To see this pictorially,

![Diagram](image_url)

suppose at $t = t' = 0$, a light ray is sent from $A$ to $B$ in the time it takes $S'$ to move from $A$ to $C'$ at a constant velocity ‘$v$’. By the time the light ray reaches $C$, $S'$ has moved a distance $\overline{BC'} = vt'$ as measured by $S$. In the meantime, $S$ remains stationary relative to $S'$. The distance the light ray travels as measured by $S'$ is $\overline{CC'} = \overline{AB'}$. That distance, as
measured by a clock stationary with respect to \( S' \), is \( ct' \). On the other hand, to an observer in \( S \), the light ray will travel a distance \( \overline{AC} \). That distance, measured on a clock stationary in \( S \), is \( ct \). Moreover, from the standpoint of \( S \), \( S' \) has moved distance \( \overline{BC} = vt \) as measured on a clock in \( S \) (see fig. 9.4.1.2).

Applying the Pythagorean theorem,

\[
(ct)^2 = (ct')^2 + (vt)^2 \rightarrow c^2(t')^2 = c^2t^2 - v^2t^2 \rightarrow (t')^2 = \left(1 - \frac{v^2}{c^2}\right)t^2 \rightarrow t' = \sqrt{1 - \frac{v^2}{c^2}}\]

The time \( t' \) measured by a clock in \( S \) will run slow by a factor of \( \sqrt{1 - v^2/c^2} \). This is called ‘time dilation’.

### 9.4.1.2 Length Contraction

Moreover, since \( x = ct \), if \( t = t'/\sqrt{1 - v^2/c^2} \), then

\[
x = ct = c \frac{t'}{\sqrt{1 - \frac{v^2}{c^2}}} \rightarrow x \sqrt{1 - \frac{v^2}{c^2}} = ct' = x'
\]

Hence, a distance \( x' \) measured by a ruler stationary in \( S' \) will appear shorter by an amount \( \sqrt{1 - v^2/c^2} \). This is called ‘length contraction’.

### 9.4.2 The Twin Paradox

The principle of relativity maintains that the observers in both reference frames can lay claim to the “at rest” status relative to the other frame. Observers in both \( S \) and \( S' \) can claim that the clock in the other frame is running slow relative to their clock. This seems quite contradictory. How can both clocks “be running slow” at the same time? This has led to something called the ‘twin paradox’. Twins moving at a constant velocity relative to one another will believe that the other twin is younger, since each twin measures the other twin’s clock as running slow relative to their own clock. But there is really no paradox. The ‘twin thought experiment’ actually requires three observers. The third observer must choose which twin is at rest and which is moving. Once a choice is made, the paradox disappears.

### 9.4.3 All Observers will Measure the Speed of Light to be \( c \)

If a particle is moving relative to \( S \) at a speed \( v' \), the distance \( x' \) that the particle travels in time \( t' \) is \( x = vt \rightarrow v = x/t \), measured by an observer in \( S \). If \( S' \) is moving relative to \( S \) with velocity \( u' \), then the distance \( x'' \) that the particle travels in time \( t'' \) is \( x' = wt' \rightarrow \)
\[ w = \frac{x'}{t'}, \text{ where } w \text{ is the velocity of the particle as measured by an observer at rest in } S'. \] Hence,

\[ w = \frac{x'}{t'} = \gamma \frac{x - ut}{\gamma \left( t - \frac{u}{c^2} x \right)} = \frac{\gamma (x - ut)}{\gamma t - \frac{u}{c^2} x} = \frac{x - u}{1 - \frac{u^2}{c^2}} = \frac{v - u}{1 - \frac{uv}{c^2}} \]

\[ v = \frac{x}{t} = \gamma \left( \frac{x' + ut'}{\gamma \left( t' + \frac{u}{c^2} x' \right)} \right) = \frac{x' + u}{1 + \frac{u x'}{c^2 t'}} = \frac{w + u}{1 + \frac{uw}{c^2}} \]

If \( w = c \), then

\[ v = \frac{c + u}{1 + \frac{uc}{c^2}} = \frac{c + u}{1 + \frac{u}{c}} = \frac{c(c + u)}{c + u} = c \]

If a particle could move at the speed of light (not possible for a massive particle), then the speed of the particle will be \( c \) as measured in both \( S \) and \( S' \), regardless of the velocity of \( S' \) relative to \( S \) so long as \( u \leq c \), where \( u \) is the relative velocity of \( S' \) with respect to \( S \).

### 9.4.4 The Oddities of Special Relativity

Is it possible that an event in the Universe that happened 4 years ago could occur after an event that happened 2 years ago? Suppose there are two stars, one 8-light-years from an observer, the other 2-light-years away. A light signal sent 2-light-years ago at a place 2-light-years away will reach the observer now. But, a light signal send 4-light-years ago at a place 8-light-years away will reach the same observer in 4 years. It is possible for a sequence of events to happen in reverse order. This occurs because the speed of light is finite.

If one event causes another, is it possible that some observer would see the sequence of events in reverse order? If a light signal is sent 8-light-years from an observer, the observer will detect the light signal 8-light-years from now. But, if the light from 8-light-years away causes another event to happen, an amount of time \( t' \) must pass before the second event can begin. The soonest possible time the observer will witness the latter event is \( 8 + t \ (t \geq 0) \) light-years from now; if one event causes another, it is not possible that the sequence of events can be seen by any observer as happening in reverse order.

If two cars collide at some place and time in the Universe, is it possible that some observer in the Universe will see the two cars missing each other, since the two events do not necessarily occur simultaneously? At the place and time of the event, the difference in location and time is zero i.e. \( \Delta x = 0, \Delta t = 0 \). Since

\[ t' = \gamma \left( t - \frac{v}{c^2} x \right) = \gamma \left( 0 - \frac{v}{c^2} 0 \right) = 0, \]

186
all observers will see the two cars collide.

9.5 Space-Time

Einstein’s account of special relativity emphasized the constancy of the velocity of light and the “principle of relativity”. In 1908 his insights received an eloquent, but purely mathematical accounting courtesy of the German-Jewish mathematician Hermann Minkowski (22 June 1864 – 12 January 1909) [235]. Minkowski’s contributions served as a foundation for the general theory of relativity and showed that the theory of special relativity resulted in a union between space and time, known as ‘space-time’. It was no longer possible to think of space and time independently.

9.5.1 World Points and World Lines

Minkowski’s idea was to let the ordered set ‘(t,x,y,z)’ designate a point in space-time called a ‘world point’. The collection of all world points is called ‘the world’. All events happening in the world happen at a certain location ‘(x,y,z)’ and at a certain time ‘t’ designated ‘(t,x,y,z)’. A curve connecting any two space-time points ‘(t,x,y,z)’ and ‘(t₁,x₁,y₁,z₁)’ is called a ‘world line’. Mathematically, ‘(t,x,y,z)’ is not dimensionally consistent, as (x,y,z) has dimensions ‘[L]’, and t has dimensions ‘[T]’. To remedy the inconsistency, the ordered set becomes (ct,x,y,z), where c is the speed of light. The set becomes dimensionally consistent. To simplify matters, the discussion will be restricted to one spatial dimension and one time dimension i.e. (x,ct).

To Illustrate this geometrically, the ‘x’ and ‘t’ axes are drawn perpendicularly to one another (see fig. 9.5.1-1). The x, t-coordinates represents the system ‘S’, the stationary system. Since the speed of light is so enormous, the line ‘x = ct’ is virtually horizontal. To make the picture easier to draw, set c = 1. The equation x = t represents the path of a light ray in space-time, the dashed line in the figure. The frame ‘S’ is represented by the line x = vt or as x’ = 0, since an observer in the S’ frame is ‘at rest’ relative to S’. Hence x – vt = 0 = x’. Assume that t = t’, where t is the time measured in S and t’ is the time measured in S’. Along the dashed line

Figure 9.5.1-1
\[ x = ct \rightarrow x' = ct - vt = (c - v)t' \]

But, this is wrong, since all observers should measure the speed of light to be \( c \).

9.5.2 The Equivalence of Space-Time with the Special Theory of Relativity

Suppose there are three observers \( P_1, P_2, P_3 \) at rest relative to each other in \( S' \), spaced 1 unit apart (see fig 9.5.2-1). What is the time \( t' \) at point \( a' \)? The equation of the line running through \( Oa' \) is \( x = t \). Hence,

\[ t = vt + 1 \rightarrow t = \frac{1}{1 - v} \]

Figure 9.5.2-1

Since \( x = t \) at \( a \), then \( x = 1/(1 - v) \).

9.5.2.1 The Synchronization of Clocks

How is it determined if the clocks worn by observers \( P_1, P_3 \) are synchronous? Suppose \( P_1 \) and \( P_3 \) agree to send a light signal to \( P_2 \) when their clocks read a certain time. When \( P_1 \) sends out the light signal, it will reach \( P_2 \) in time \( t' \) at point \( a' \). In order for \( P_3 \) to send a light signal that will reach point \( a' \) at time \( t' \), in other words, will reach \( P_2 \) simultaneously with the light signal sent by \( P_1 \), \( P_3 \)'s light ray must make an angle of 45° with the \( x \)-axis (see fig 9.5.2-1).

What is the time \( t' \) at point \( b' \)? Note that the equation \( x = t \) has slope \( '1' \). So, a line running through \( Oa' \) has a slope \( '1' \). Therefore, a line passing through \( ab' \) has slope \( '−1' \). The equation of this line is \( x = −t + K \rightarrow x + t = K \), where \( K \) is a constant. If \( t = 0 \), then \( x = K \). So, \( K \) is just the point where the line meets the \( x \)-axis. But,

\[ x + t = \frac{1}{1 - v} + \frac{1}{1 - v} = \frac{2}{1 - v} \]

Hence, \( K = 2/(1 - v) \). Since \( x - vt = 2 \) and \( x + t = 2/(1 - v) \), solving these two equations simultaneously leaves
\[
t(1 + v) = \frac{2}{1-v} - \frac{2(1-v)}{1-v} = \frac{2v}{1-v} \rightarrow t = \frac{2v}{(1-v)(1+v)} = \frac{2v}{1-v^2}
\]

What is the value of \( x \) at point \( 'b' \)? Remembering that \( x + t = 2/(1-v) \), then

\[
x = \frac{2}{1-v} - t = \frac{2}{(1-v)} - \frac{2v}{1-v^2} = \frac{2}{1-v^2}
\]

To determine the equation of the line passing through \( \overline{O\,b} \), recall that, in general, the equation of an arbitrary line is \( t = mx + K \), where \( m \) is the slope and \( K \) is the \( t \)-intercept.

The slope of the line running through \( b \) is

\[
\frac{t}{x} = m = \frac{2v}{2} = v
\]

Since \( t = 0 \) when \( x = 0, K = 0 \), therefore, the equation of the line is \( t = vx \). This line represents the times at which the “at rest” observers in \( S \) will agree that their clocks are synchronous. Note that the line \( 'x = vt' \) has an inverse slope relative to the line \( 't = vx' \) i.e. \( t = x/v \rightarrow x = vt \). There is a symmetry between the lines \( 't = vx' \) and \( 'x = vt' \); the two lines are mirror images of one another with the line \( 't = x' \) representing the axis of symmetry.

### 9.5.2.2 Deriving the Lorentz Transformations from Space-Time

What are these equations in terms of \( x' \) and \( t' \)? Remembering that \( x' = x - vt \) and \( t' = t - vx \), then if \( x = t \), then \( x' = t' \), since the speed of light must be the same for all observers. Hence,

\[
x' = (x - vt)f(v), \quad t' = (t - vx)g(v),
\]

where \( f(v) \) and \( g(v) \) are functions of \( v \) that would make \( x' = t' \). But for \( x' \) to be equal to \( t' \), \( f(v) = g(v) \). Solving \( x' = (x - vt)f(v) \) and \( t' = (t - vx)f(v) \) simultaneously for \( x \) and \( t \) respectfully leaves

\[
x = (x' + vt')f(v), \quad t = (t' + vx')f(v)
\]

The only way the equations \( 'x', t', t, x' \) can be compatible with one another is if, by plugging the equation for \( x' \) into the equation for \( x \), the result would be \( x = x \). The same can be said of \( t \) and \( t' \). Hence,

\[
x = ((x - vt)f(v) + v(t - vx)f(v))f(v) = (x - vt)f^2(v) + v(t - vx)f^2(v)
\]

\[
= xf^2(v) - vtf^2(v) + vtf^2(v) - v^2xf^2(v) = xf^2(v) - v^2xf^2(v)
\]

\[
= x(1 - v^2)f^2(v) \rightarrow f^2(v) = \frac{1}{1 - v^2} \rightarrow f(v) = \pm \frac{1}{\sqrt{1 - v^2}}
\]

189
Note that \( f(v) \neq -1/\sqrt{1-v^2} \), since, if this were the case, \( x = -x \) and the two equations for \( x' \) and \( x \) would be incompatible. Therefore,

\[
x' = \frac{x - vt}{\sqrt{1 - v^2}}, \quad x = \frac{x' + vt'}{\sqrt{1 - v^2}}
\]
\[
t' = \frac{t - vx}{\sqrt{1 - v^2}}, \quad t = \frac{t' + vx'}{\sqrt{1 - v^2}}
\]

To simplify the geometric explanation of special relativity \( c \), the speed of light, was set equal to 1. To restore the above equations to the form that includes \( c \),

\[
x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}}, \quad x = \frac{x' + vt'}{\sqrt{1 - v^2/c^2}}
\]
\[
t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}, \quad t = \frac{t' + vx'/c^2}{\sqrt{1 - v^2/c^2}}
\]

These equations are the ‘Lorentz transformations’, which show that Minkowski’s ‘space-time’ is entirely equivalent to Einstein’s theory of special relativity.

9.5.3 Invariance in Space-Time

Suppose there are two coordinate systems \( 'X, Y' \) and \( 'X', Y'' \), where the \( X', Y'' \)-system is rotated relative to the \( X, Y \)-system by an angle \( '\theta' \) around a common origin (see fig. 9.5.3-1). If the coordinates of the point \( 'P' \) in the \( X, Y \)-system are designed \( '(x, y)' \) and in the \( X', Y'' \)-system designated \( '(x', y')' \), then, \( x \neq x' \) and \( y \neq y' \), but the distance \( 'OP' \) is the same in both systems. Using the Pythagorean theorem,

\[
x^2 + y^2 = (x')^2 + (y')^2 = (OP)^2
\]
The relationship between \( (x, y) \) and \( (x', y') \) is given by the following transformation equations:

\[
\begin{align*}
    x &= x' \cos \theta - y' \sin \theta \\
    y &= x' \sin \theta + y' \cos \theta
\end{align*}
\]

In Euclidean space the distance \( \overline{OP} \) is invariant, independent of the orientation of the coordinate system. Is there an analogous invariance associated with the Lorentz transformations? In the equations above, let \( y = t \) and \( y' = t' \). Is \( x^2 + t^2 = (x')^2 + (t')^2 \) invariant in a Lorentz space? Let

\[
\begin{align*}
    x^2 + t^2 &\equiv (x')^2 + (t')^2 \equiv \left( \frac{x - vt}{\sqrt{1 - v^2}} \right)^2 + \left( \frac{t - vx}{\sqrt{1 - v^2}} \right)^2 \\
    &\equiv \frac{x^2 + v^2 t^2 - 2xvt}{1 - v^2} + \frac{t^2 + v^2 x^2 - 2xvt}{1 - v^2}
\end{align*}
\]

A glance at the equation above reveals that \( x^2 + t^2 \neq (x')^2 + (t')^2 \), and hence, is not invariant in space-time. However, consider \( t^2 - x^2 \), then

\[
(t')^2 - (x')^2 = \left( \frac{t - vx}{\sqrt{1 - v^2}} \right)^2 - \left( \frac{x - vt}{\sqrt{1 - v^2}} \right)^2 = \frac{t^2 + v^2 x^2 - 2xvt}{1 - v^2} - \frac{x^2 + v^2 t^2 - 2xvt}{1 - v^2} = t^2 - x^2
\]

Therefore, \( t^2 - x^2 \) is invariant in space-time. This transformation is sometimes referred to as 'Lorentz invariance'. The formal derivation of this invariance was discovered by Minkowski. Recall the Lorentz transformations:

\[
\begin{align*}
    x &= \gamma (x' + vt'), \\
    t &= \gamma (t' + vx'), \\
    \gamma &= \frac{1}{\sqrt{1 - v^2}}
\end{align*}
\]

Letting \( t = it, \ t' = it' \), \( i = \sqrt{-1} \), then

\[
\begin{align*}
    x &= \gamma x' + i\gamma vt', \\
    it &= i\gamma t' + \gamma vx' \rightarrow t = \gamma t' - i\gamma vx'
\end{align*}
\]

Substituting \( \cos \theta \) for \( \gamma \) and \( \sin \theta \) for \( -i\gamma v \), then

\[
\begin{align*}
    x &= x' \cos \theta - t' \sin \theta, \\
    t &= x' \sin \theta + t' \cos \theta
\end{align*}
\]

The equations above are the transformation equations for a Euclidean space, where one Euclidean coordinate system is rotated relative to another (see fig. 9.5.3-1). However, because of the imaginary angle \( \theta = \sin^{-1}(-i\gamma v) \), Minkowski’s space-time is not Euclidean.

Note that

\[
x^2 = (x')^2 \cos^2 \theta - 2x' t' \cos \theta \sin \theta + (t')^2 \sin^2 \theta
\]
\[ t^2 = (x')^2 \sin^2 \theta + 2x't' \sin \theta \cos \theta + (t')^2 \cos^2 \theta \]

Adding the two equations above leaves
\[ x^2 + t^2 = (x')^2(\cos^2 \theta + \sin^2 \theta) + (t')^2(\cos^2 \theta + \sin^2 \theta) = (x')^2 + (t')^2 \]

### 9.6 Proper Time

Consider a clock moving in the \( S' \) frame along the line \( 'x = vt' \) or \( 'x' = 0' \) (see fig. 9.5.2-1). Since \( t^2 - x^2 = (t')^2 - (x')^2 \) and \( x' = 0 \) along this line, then \( t^2 - x^2 = (t')^2 \); the time \( 't' \) is called the 'proper time' and is customarily denoted by the Greek letter \( \tau ' \). The proper time is significant because it is invariant, all observers in space-time will agree on the proper time.

More generally, all observers within a system will agree on invariant values. Invariant quantities become universal laws of physics within the constraints of the physical system under consideration. In the present case, all systems must be inertial systems.

#### 9.6.1 Proper Time in Three Dimensions

So far motion has been restricted to one dimension, along the \( x \)-axis. However, the proper time can be extended to higher dimensions. In general,
\[ \tau^2 = t^2 - x^2 - y^2 - z^2 \]

is invariant. Note that the equation for \( \tau^2 \) seems dimensionally inconsistent, the term \( \tau^2 \) has dimension \( '[T]^2' \), whereas \( x^2 \) has dimension \( '[L]^2' \). To make the proper time dimensionally consistent, reinsert \( c \), the velocity of light, into the equation. This gives
\[ (ct)^2 = (ct)^2 - x^2 - y^2 - z^2, \]

Dividing through by \( c^2 \) leaves
\[ \tau^2 = t^2 - \left(\frac{x}{c}\right)^2 - \left(\frac{y}{c}\right)^2 - \left(\frac{z}{c}\right)^2 \rightarrow t^2 = t^2 - x^2 - y^2 - z^2, \quad c = 1 \]

What the proper time represents physically is the distance between two space-time points. If \( t^2 - x^2 = 0 \), then \( x = \pm t \) or \( x, t = 0 \). Note that in a Euclidean space, the only way \( x^2 + y^2 = 0 \) is if \( x, y = 0 \). And if \( x = x' \) and \( y = y' \), then \( (x, y) \) and \( (x', y') \) are the same point. But in space-time, \( t^2 - x^2 = 0 \) does not necessarily imply that \( x, t = 0 \). In fact, \( t \) could be equal to \( +x \) i.e. \( t = x \), which represents the path of a light ray. Hence, if a clock could move at the speed of light, its proper time would be zero i.e. \( \tau = 0 \).

Geometrically, the invariant space-time interval or proper time can be represented by two hyperbolas with their asymptotes denoting the path of a light ray (see fig. 9.6.1-1).
9.6.1.1 Light-Like Distances

If an object, like a photon, travels along the path '$x = t$' or '$x = -t$', then $t^2 - x^2 = 0$. The distance between two points along this path is called a 'light-like' distance. Only an object travelling at the speed of light could reach the point 'G' from 'N' or from 'R' to 'N' (see fig. 9.6.1-1). Geometrically, all light rays travel along the surfaces of cones. The 'future' cone represents a light ray moving from 'now' into the 'future'.

9.6.1.1 Time-Like Distances

If $t^2 - x^2 > 0$, the distance between two points is called a 'time-like' distance. All points lying within the cone are time-like. Any path that joins two time-like points represents a possible future of an observer moving along that path.

Moreover, if a particle is moving along a time-like trajectory, it is always possible to choose a reference frame in which the spatial location of the particle does not change i.e. since $t^2 - x^2 = (t')^2 - (x')^2$, choose the reference frame where $x' = 0$. Hence, $t^2 - x^2 = \tau^2$ and this is the trajectory corresponding to the proper time.
9.6.1.2 Space-Like Distances

If \( t^2 - x^2 < 0 \), the distance between two points is called a 'space-like' distance. A light ray cannot reach points that are space-like separated. Note that along the line \( t = vx \) or \( t' = 0 \) that \( t^2 - x^2 = -(x')^2 \) or \( (x')^2 = x^2 - t^2 \). Hence, \( x' \) is an invariant quantity called the 'proper distance'. If points are space-like separated, it is always possible to choose a reference frame in which time does not change i.e. since \( t^2 - x^2 = (t')^2 - (x')^2 \), choose the reference frame where \( t' = 0 \). Hence, \( x^2 - t^2 = (x')^2 \), where \( x' \) is the proper distance and is invariant.

9.7 4-Vectors

A vector, in the "physics" sense, is a directed line segment, represented by an ordered set of numbers. The ordered set \( \langle x, y, z \rangle \equiv \langle x^1, x^2, x^3 \rangle = \langle x^i \rangle, \ i = 1, ..., 3 \) is called a '3-vector', where the superscripts are labels, not exponents. If time is included i.e.

\[
\langle t, x, y, z \rangle \equiv \langle x^0, x^1, x^2, x^3 \rangle = \langle x^\mu \rangle, \quad t = x^0, \quad \mu = 0, ..., 3,
\]

the vector becomes four dimensional. Not all four-dimensional vectors are '4-vectors'. To be a '4-vector', a four dimensional vector must also be Lorentz invariant.

The square of the proper time \( \tau^2 \) can be represented as

\[
\tau^2 = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2, \quad t = x^0
\]

![Figure 9.7-1](image)

If a particle is moving along a world line (see fig. 9.7-1), and if the world line is divided into tiny segments \( \langle \Delta x^0, \Delta x^1, \Delta x^2, \Delta x^3 \rangle = \langle \Delta x^\mu \rangle, \ \mu = 0, ..., 3 \), \( \Delta x^\mu \) represents the space-time distance between two adjacent world points. Note that

\[
\Delta \tau^2 = (\Delta x^0)^2 - (\Delta x^1)^2 - (\Delta x^2)^2 - (\Delta x^3)^2 \rightarrow \Delta \tau = \sqrt{(\Delta x^0)^2 - (\Delta x^1)^2 - (\Delta x^2)^2 - (\Delta x^3)^2}
\]

194
9.7.1 The 4-Velocity

The expression \( U^\mu = \Delta x^\mu / \Delta \tau , \mu = 0, ..., 3 \) is called the ‘4-velocity’. In terms of the proper time, the 4-velocity is an invariant quantity. The 4-velocity is the analog to the velocity in 3-dimensional space i.e. \( (dx/dt, dy/dt, dz/dt) = dx^i/dt = v^i, i = 1, ..., 3 \).

A particle always moves along a time-like trajectory i.e. a particle moves in accordance with the proper time. If a particle moved along a space-like trajectory, it would travel faster than light, which, according to the theory of special relativity, is impossible.

9.7.1.1 The Relationship between 4-Velocity and 3-Velocity

What is the relationship between \( U^\mu \) and \( v^i \)? If \( \lim_{\Delta \tau \to 0} (\Delta \tau) = d\tau \), then

\[
\begin{align*}
\Delta \tau^2 &= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 \to d\tau = \sqrt{(dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2} \\
\end{align*}
\]

If \( U^\mu = dx^\mu / d\tau \), \( \mu = 0, ..., 3 \), then

\[
\begin{align*}
v^i &= \frac{dx^i}{d\tau} \frac{d\tau}{dt} = U^i \frac{dt}{d\tau}, \quad i = 1, ..., 3,
\end{align*}
\]

Where \( U^0 = dx^0 / d\tau = dt / d\tau \). But

\[
\begin{align*}
d\tau^2 &= (dx^0)^2 - \sum_{i=1}^{3} dx^i dx^i \to d\tau = \sqrt{(dt)^2 - \sum_{i=1}^{3} dx^i dx^i} = dt \sqrt{1 - \sum_{i=1}^{3} \frac{dx^i dx^i}{dt^2}} \\
&= dt \sqrt{1 - \dot{x}^i \dot{x}_i} = dt \sqrt{1 - v^2} \to \frac{d\tau}{dt} = \sqrt{1 - v^2}
\end{align*}
\]

Therefore,

\[
\begin{align*}
v^i &= \frac{dx^i}{d\tau} \frac{d\tau}{dt} = U^i \sqrt{1 - v^2} \to U^i = \frac{v^i}{\sqrt{1 - v^2}}, \quad i = 1, ..., 3
\end{align*}
\]

Now,

\[
\begin{align*}
\sum_{i=1}^{3} U^i U^i &= \frac{\sum_{i=1}^{3} v^i v^i}{\sqrt{1 - v^2} \sqrt{1 - v^2}} = (U^1)^2 + (U^2)^2 + (U^3)^2 = \frac{v^2}{1 - v^2}, \quad (U^0)^2 = \frac{1}{1 - v^2},
\end{align*}
\]

So,

\[
(U^0)^2 - U^i U^i = \frac{1}{1 - v^2} - \frac{v^2}{1 - v^2} = \frac{1 - v^2}{1 - v^2} = 1
\]

The equation above shows that in space-time, space and time are not independent, but in some sense represent a whole.
9.8 Applying Hamilton’s Principle to Space-Time

A complete physical system requires developing descriptions for momentum and energy. Preferably, the description should be invariant i.e. the same in all inertial coordinate systems. Given these requirements, how are momentum and energy described in space-time?

If a particle is traveling along a time-like world line ‘W’ (see fig. 9.8-1), about the only invariant along the world line is the proper time ‘τ’. If possible, momentum and energy should be described in terms of τ, since all observers will agree on the value of τ in any inertial reference frame. If W is divided into a number of small pieces (Δτ’s), then \( \sum \Delta \tau = \tau \), the total proper time along the world line. As it turns out, \( \sum \Delta \tau \) is an instance of the least-action principle associated with classical mechanics. The action along any trajectory is stationary (usually a minimum), so that if the action is multiplied by any constant, it remains stationary. Replacing \( \sum \Delta \tau \) by \( -m \sum \Delta \tau \), \( m = \text{a constant} \), does not affect the stationarity of the action.

Since τ is invariant, all observers will agree on this trajectory. If the world line is continuously smooth, \( \lim_{\Delta \tau \to 0} \Delta \tau = d\tau \). Hence, \( -m \sum \Delta \tau \) can be replaced by \( -m \int d\tau \) and

\[
-m \int d\tau = -m \int \sqrt{dt^2 - dx^i dx_i} = -m \int \sqrt{1 - \dot{x}^i \dot{x}_i} \, dt, \quad dx^i dx_i = dx^2 + dy^2 + dz^2
\]

The integrand \( \sqrt{1 - \dot{x}^i \dot{x}_i} \) is a function of velocity only. The integral is of the form

\[
A = \int \mathcal{L} \, dt, \quad \mathcal{L} = -m \sqrt{1 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2}
\]

where \( A \) is the ‘action’ and \( \mathcal{L} \) a ‘Lagrangian density function’. Since \( \mathcal{L} \) is function of velocity and not of position, \( V \), the potential, is zero, which indicates that there are no forces present. The particle simply moves along a trajectory determined by the
geometry of the world line. If the velocity of the particle is \( v \ll c \), then if \( \mathcal{L} = -m\sqrt{1 - v^2} \), and, by the binomial theorem

\[
(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)x^2}{2!} \pm \frac{n(n-1)(n-2)x^3}{3!} + \ldots \rightarrow (1 - v^2)^{1/2} \approx 1 - \frac{1}{2}v^2,
\]
to a first approximation.

Hence,

\[
-m\sqrt{1 - v^2} \approx -m \left(1 - \frac{1}{2}v^2\right) = -m + \frac{1}{2}mv^2 = \mathcal{L}
\]

Recall the definition of momentum \( p_i = \partial\mathcal{L}/\partial\dot{x}_i \). It follows then that

\[
\frac{\partial\mathcal{L}}{\partial\dot{x}} = p_x = \frac{\partial \left(-m\sqrt{1 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2}\right)}{\partial \dot{x}} = \frac{mx}{\sqrt{1 - v^2}}
\]

But \( U^x = U^1 = v^1/\sqrt{1 - v^2} \). Hence, \( p_x = mU^1 \). Likewise,

\[
\frac{\partial\mathcal{L}}{\partial\dot{y}} = p_y = \frac{\partial \left(-m\sqrt{1 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2}\right)}{\partial \dot{y}} = \frac{my}{\sqrt{1 - v^2}}
\]

\[
\frac{\partial\mathcal{L}}{\partial\dot{z}} = p_z = \frac{\partial \left(-m\sqrt{1 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2}\right)}{\partial \dot{z}} = \frac{m\dot{z}}{\sqrt{1 - v^2}}
\]

So,

\[
p_x = mU^1, \quad p_y = mU^2, \quad p_z = mU^3 \rightarrow p_i = mU^i, \quad i = 1, \ldots, 3
\]

### 9.8.1 Conservation of the 4-Momentum

If spatial momentum is conserved, then \( p_i = p_j \rightarrow p_i - p_j = 0 \), where \( i \) signifies the initial momentum and \( j \) the final momentum of a particle. However, spatial momentum is a 3-vector and, in space-time, is not independent of the time dimension. Spatial momentum is not, in general, conserved in space-time. There must be a forth component \( p_0 \), which makes the 4-momentum conserved. In physics, normally, energy is conserved. But does \( p_0 = mU^0 \) represent the energy? Recall that the Hamiltonian \( H = T + V \) is the total energy. Moreover,

\[
H = \sum_i \dot{x}_i p_i - \mathcal{L}
\]
Now $p_i = m\dot{x}_i/\sqrt{1 - v^2}$ and $L = -m\sqrt{1 - v^2}$. Hence,

$$H = \sum_i \frac{m\dot{x}_i}{\sqrt{1 - v^2}} + m\sqrt{1 - v^2} = \sum_i \frac{m\dot{x}_i^2}{\sqrt{1 - v^2}} + m \frac{1 - v^2}{\sqrt{1 - v^2}} = m\frac{v^2}{\sqrt{1 - v^2}} + m \frac{1 - v^2}{\sqrt{1 - v^2}}$$

$$= \frac{m}{\sqrt{1 - v^2}}$$

But, $U_0 = 1/\sqrt{1 - v^2} \rightarrow p_0 = m/\sqrt{1 - v^2}$. The conservation of momentum in three dimensional space becomes the conservation of the 4-momentum $\langle p_0, p_1, p_2, p_3 \rangle$ in space-time.

### 9.8.2 Energy in Space-Time

By the binomial theorem,

$$\frac{m}{\sqrt{1 - v^2}} \approx m + \frac{mv^2}{2}$$

The above equation is not dimensionally consistent, since the first term on the right-hand side has dimensions $'[M]'$ and the second term has dimensions $'[ML^2T^{-2}]'$. To make the equation dimensionally consistent, replace $v$ by $v/c$, then

$$\frac{m}{\sqrt{1 - (\frac{v}{c})^2}} \approx m + \frac{mv^2}{2c^2},$$

which has dimensions $'[M]'$ on both sides of the equation. Multiplying through by $c^2$ leaves

$$\frac{mc^2}{\sqrt{1 - (\frac{v}{c})^2}} \approx mc^2 + \frac{mv^2}{2},$$

which has dimensions $'[ML^2T^{-2}]' \text{ (energy)}$. Let $E = mc^2/\sqrt{1 - (v/c)^2}$, then

$$E = mc^2 + \frac{mv^2}{2} \rightarrow E = mc^2, \quad \text{if} \ v = 0$$

In other words, a particle of mass $'m'$ has rest energy $'mc^2'$. If $c = 1$, then

$$(U_0)^2 - U_i U_i = 1 = (U_0)^2 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2 \rightarrow (mU_0)^2 - m^2 U_i U_i$$

$$= m^2(U_0)^2 - m^2 x^2 - m^2 y^2 - m^2 z^2 = m^2 \rightarrow m^2 = (mU_0)^2 - p_x^2 - p_y^2 - p_z^2$$
Now \( U^0 = 1/\sqrt{1 - v^2} \rightarrow mU^0 = m/\sqrt{1 - v^2} \). Hence,

\[
m^2 = \left(\frac{m}{\sqrt{1 - v^2}}\right)^2 - p_x^2 - p_y^2 - p_z^2
\]

To make the equation above dimensionally consistent, write

\[
(mc^2)^2 = \left(\frac{mc^2}{\sqrt{1 - v^2/c^2}}\right)^2 - c^2p_x^2 - c^2p_y^2 - c^2p_z^2 \rightarrow m^2c^4 = E^2 - p^2c^2 \rightarrow E^2
\]

\[
= p^2c^2 + m^2c^4 \rightarrow E = \sqrt{p^2c^2 + m^2c^4}
\]

The concept of ‘energy’ \( (E) \) in space-time is very different from concept of ‘energy’ in non-relativistic mechanics, where mass is conserved. Particles in non-relativistic mechanics all possess mass and if an experiment is performed which transforms one substance into another, the weight of the substance before the transformation is equal to the weight of the substance after the transformation.

But relativistically, if \( m = 0 \),

\[
E = \sqrt{p^2c^2 + 0^2c^4} = \sqrt{p^2c^2} = c|p|
\]

which suggests the idea of a ‘massless particle’, for example, the ‘photon’. Massless particles travel at the speed of light, but can have different energies. Their energy does not depend on velocity, since all massless particles travel at the speed of light in empty space. Moreover, particles with mass can transform into particles without mass. These types of transformations do happen in Nature. What is conserved in relativistic mechanics is the 4-momentum \( 'mU^\mu = (p_0, p_1, p_2, p_3)' \), not the mass, giving a slight glimmer into the world of quantum mechanics. Einstein was one of the first to recognize the ramifications of the particle nature of light, which led to the development of quantum mechanics – a development he would later despise.

**9.9 Classical Field Theory and Space-Time**

A function ‘\( \varphi(t, x) \)’ that gives the value for \( \varphi \) at all points in space-time is called a ‘field’. Fields can be defined in any space, not just space-time.
If all the values of \( \varphi(t, x) \) are simply numbers, then \( \varphi(t, x) \) is called a ‘scalar field’. If \( |\varphi(t, x)\rangle \) is a vector, then \( |\varphi(t, x)\rangle \) is called a ‘vector field’. Fields are not restricted to a single spatial dimension, but can have any number of spatial dimensions i.e. \( \varphi(t, x^i), \) \( i = 1, \ldots, n \). It is rare to find a field with more than one time dimension. If \( \varphi \) specifies the trajectory of a particle, then \( \varphi \) is purely a function of time. The trajectory of a particle in classical physics is given by the action principle i.e.

\[
A = \int_a^b L \, dt, \quad L = \frac{1}{2} m \left( \frac{d\varphi}{dt} \right)^2 - V(\varphi) \ ,
\]

where the end points \( 'a, b' \) are fixed and \( L = \text{kinetic energy} - \text{potential energy} \). The equations of motion are

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \varphi} \right) - \frac{\partial L}{\partial \varphi} = 0 \rightarrow m\ddot{\varphi} = -\frac{\partial V}{\partial \varphi},
\]

the Euler-Lagrange equations, which are equivalent to Newton’s second law of motion.

Both classical and relativistic mechanics are governed by the principle of least action. Describing the action of a field, regardless of the number of spatial dimensions involved, requires minimizing the action. The action is minimized if it complies with the Euler-Lagrange equations.

**9.9.1 The Action for a Field**

If \( \varphi = \varphi(t, x^i) \), one time dimension and three spatial dimensions, then \( 'dt' \) in the action ‘\( A' \) must be replaced by \( dt dx dy dz \) i.e.

\[
A = \int L \, dt dx dy dz = \int L \, dx^\mu,
\]

where \( dx^\mu = dt dx dy dz \). The Lagrangian density function ‘\( L' \) becomes

\[
L = L\left( \varphi, \frac{\partial \varphi}{\partial t}, \frac{\partial \varphi}{\partial x^i}, \frac{\partial \varphi}{\partial y}, \frac{\partial \varphi}{\partial z} \right) = L\left( \varphi, \frac{\partial \varphi}{\partial x^\mu} \right), \quad \mu = 0, \ldots, 3
\]

Sometimes \( \partial \varphi/\partial t \) is written \( '\partial_t' \), and similarly, \( \partial \varphi/\partial x \) is written \( '\partial_x' \) etc.

**9.9.2 The Euler-Lagrange Equations for a Field**

The equations of motion, upon replacing \( d/dt \) with \( \partial/\partial x^\mu \) become

\[
\sum_\mu \frac{\partial}{\partial x^\mu}\left( \frac{\partial L}{\partial \left( \frac{\partial \varphi}{\partial x^\mu} \right)} \right) - \frac{\partial L}{\partial \varphi} = 0, \quad \mu = 0, \ldots, 3,
\]
which represents the equations of motion in 4-dimensional space-time. A Lagrangian ‘\( \mathcal{L} \)’ in 4-dimensional space-time might look like

\[
\mathcal{L} = \frac{1}{2} m \left[ \left( \frac{\partial \varphi}{\partial t} \right)^2 - \left( \frac{\partial \varphi}{\partial x} \right)^2 - \left( \frac{\partial \varphi}{\partial y} \right)^2 - \left( \frac{\partial \varphi}{\partial z} \right)^2 \right] - V(\varphi)
\]

Note that

\[
\frac{\partial \mathcal{L}}{\partial x^0} = \frac{\partial \mathcal{L}}{\partial t} = m \dot{\varphi} \quad \Rightarrow \quad \frac{\partial \mathcal{L}}{\partial t} \left( \frac{\partial \mathcal{L}}{\partial t} \right) = m \ddot{\varphi}, \quad x^0 = t
\]

Likewise,

\[
\frac{\partial \mathcal{L}}{\partial x^i} = -m \frac{\partial \varphi}{\partial x^i} \rightarrow \frac{\partial \mathcal{L}}{\partial x^i} \left( \frac{\partial \mathcal{L}}{\partial x^i} \right) = -m \frac{\partial^2 \varphi}{\partial^2 (x^i)}, \quad x^1 = x, \quad x^2 = y, \quad x^3 = z
\]

Hence,

\[
\sum_{\mu} \frac{\partial}{\partial x^\mu} \left( \frac{\partial \mathcal{L}}{\partial \varphi} \right) - \frac{\partial \mathcal{L}}{\partial \varphi} = m \left[ \frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} - \frac{\partial^2 \varphi}{\partial y^2} - \frac{\partial^2 \varphi}{\partial z^2} \right] + \frac{\partial V}{\partial \varphi} = 0,
\]

where \( 1/c^2 \) has been reinserted for dimensional consistency. In the case where \( \partial V/\partial \varphi = 0 \) and if the \( y, z \)-coordinates are ignored, then the expression in the square brackets becomes

\[
\frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} = 0,
\]

which is the equation of a wave, moving in the direction of the \( x \)-axis.

However, \( \partial V/\partial \varphi \) does not always equal zero. Suppose \( V(\varphi) = (\mu^2/2)\varphi^2 \), then \( \partial V/\partial \varphi = \mu^2 \varphi \). The equation above becomes

\[
\frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} + \mu^2 \varphi = 0
\]

Accept for the middle term ‘\( - \frac{\partial^2 \varphi}{\partial x^2} \)’, this is the equation of a harmonic oscillator, where \( \mu^2 \) represents the spring constant.

**9.9.3 The Principle of Least Action as Applied to Special Relativity**

Applying the principle of least action to special relativity requires that the action be invariant under Lorentz transformations, in other words, specified in terms of the proper time. Given the scalar field ‘\( \varphi(t, x^i) \)’, then \( \varphi(t, x^i) = \varphi(t', (x')^i) \), the value \( \varphi \) at a point in space-time should be the same, regardless of the frame of reference used. Scalar fields are invariant in space-time.
Under the rules for Lorentz transformations:

\[ t' = \frac{t - vx}{\sqrt{1 - v^2}}, \quad x' = \frac{x - vt}{\sqrt{1 - v^2}}, \quad y' = y, \quad z' = z, \quad c = 1 \]

Given any 4-vector \( 'A^\mu' \), then

\[
(A')^0 = \frac{A^0 - vA^1}{\sqrt{1 - v^2}}, \quad (A')^1 = \frac{A^1 - vA^0}{\sqrt{1 - v^2}}, \quad (A')^2 = A^2, \quad (A')^3 = A^3
\]

If \( 'A^\mu = (A^0, iA^1, iA^2, iA^3) \), \( i = \sqrt{-1} \) is a 4-vector, then

\[
\langle A^\mu | A^\mu \rangle = (A^0)^2 - (A^1)^2 - (A^2)^2 - (A^3)^2
\]

is a scalar, and hence, invariant. Moreover, \( \partial \varphi / \partial x^\mu \) is a 4-vector. And therefore, \( \langle \partial \varphi / \partial x^\mu \mid \partial \varphi / \partial x^\mu \rangle \) is a scalar. The upshot is that in order to make the principle of least action the same in every reference frame, the Lagrangian \( 'L' \) should be a scalar.

### 9.9.3.1 Particle/Field Interactions in Special Relativity

Both classical mechanics and the theory of special relativity treat the idea of a ‘field’ and the idea of a ‘particle’ as separate notions. But in some situations, electromagnetism, for example, moving electrons create fields and fields affect the trajectory of electrons. When particles and fields interact, there is an action/reaction relationship between the field and particle. At this point, the happy marriage between pure mathematics and physics begins to diverge. Up to now, the concept of ‘space-time’ has been pretty much synonymous the geometry of world points and world lines and the ‘field’, which describes wave-like phenomena.

Describing particle/field interactions requires developing what could be called an ‘artificial invention’, inventions not rooted in foundational mathematics. Of course, the idea that the ‘particle’ and the ‘field’ are separate notions breaks down in the quantum realm, but the first signs of trouble begin by attempting to describe particle/field interactions in relativity.

Suppose a field \( \varphi(t, x) \) is given. The Lagrangian for the particle is given in terms of the proper time \( \tau \) and since \( \tau^2 = t^2 - x^2 \),

\[
L_{\text{particle}} = -m \frac{d\tau}{dt} = -m \frac{1}{dt} \sqrt{dt^2 - dx^2} = -m \sqrt{1 - \dot{x}^2}, \quad \dot{x} = \frac{dx}{dt}
\]

The action for a particle is

\[
A_{\text{particle}} = \int L_{\text{particle}} dt = \int -m \sqrt{1 - \dot{x}^2} dt
\]

What about the particle’s interaction with the field? A very simple interaction \( 'A_{\text{interact}}' \) might be described
\[ A_{\text{interact}} = \int -[m + g\varphi(t,x)]\sqrt{1 - \dot{x}^2} \, dt \rightarrow A_{\text{interact}} = \int -[mc^2 + g\varphi(t,x)]\sqrt{1 - \frac{\dot{x}^2}{c^2}} \, dt, \]

where \( c \) has been reinserted, giving the action in terms of the energy and \( g \) is a constant. To further simplify the problem, \( \sqrt{1 - \dot{x}^2/c^2} \) will be replaced by its binomial approximation \( '1 - \dot{x}^2/2c^2' \), where the particle is assumed to move slowly compared to the speed of light. Hence,

\[ A_{\text{interact}} = \int -[mc^2 + g\varphi(t,x)] \left(1 - \frac{\dot{x}^2}{2c^2}\right) \, dt \]

The first term \(-mc^2\) in the equation above is a constant, and hence, its derivative is zero. Therefore, it does not appear in the equations of motion and can be discarded. The last term is too small to count, since \( \dot{x} \ll c \). This leaves

\[ A_{\text{interact}} = \int \left(\frac{m\ddot{x}^2}{2} - g\varphi(t,x)\right) \, dt \rightarrow \mathcal{L}_{\text{interact}} = \frac{m\dot{x}^2}{2} \text{ kinetic energy} - \frac{g\varphi(t,x)}{2} \text{ potential energy} \]

Since the interest is in minimizing the action of both the particle and the field, there must be a Lagrangian \( \mathcal{L}_F \) for the field, say

\[ \mathcal{L}_F = \frac{1}{2} \left(\frac{\partial \varphi}{\partial t}\right)^2 - \frac{1}{2} \frac{\partial \varphi}{\partial x^i} \frac{\partial \varphi}{\partial x^i} \rightarrow A_F = \int \left[ \frac{1}{2} \left(\frac{\partial \varphi}{\partial t}\right)^2 - \frac{1}{2} \frac{\partial \varphi}{\partial x^i} \frac{\partial \varphi}{\partial x^i} \right] \, d^4x \]

How is the particle distinguished from the field? If the particle is at rest, let its spatial position in the field be \( (x,y,z) = (0,0,0) \). Hence, \( \dot{x} = 0 \), which implies that

\[ \mathcal{L}_{\text{interact}} = -g\varphi(t,0) \]

This is where things go off the rails a bit. Note that \( \mathcal{L}_{\text{interact}} \) is given completely in terms of the field. The particle is represented by introducing the ‘Dirac delta function’. The name is a misnomer. The Dirac delta function \( '\delta(t)' \) is not a function in the normal “mathematical” sense. Its definition is

\[ \delta(t) = \begin{cases} 0, & t \neq 0 \\ \infty, & t = 0 \end{cases} \]

Given any function \( 'f(x)' \), the defining property of \( \delta(t) \) is

\[ \int f(x)\delta(x) \, dx = f(0) \]
In the case of $\mathcal{L}_{\text{interact}} = -g \varphi(t, x^i)$,
\[
\int \mathcal{L} \, dx^i = \int -g \varphi(t, x^i) \delta(x^i) \, dx^i = \int -g \varphi(t, x, y, z) \delta(x) \delta(y) \delta(z) \, dx \, dy \, dz = -g \varphi(t, 0), \quad i = 1, \ldots, 3
\]
The action $\mathcal{A}_{P/F}$ for both the particle and the field is then
\[
\mathcal{A}_{P/F} = \int \left[ \frac{1}{2} \left( \frac{\partial \varphi}{\partial t} \right)^2 - \frac{1}{2} \frac{\partial \varphi}{\partial x^i} \frac{\partial \varphi}{\partial x^i} - g \varphi(t, x^i) \delta^3(x) \right] \, d^4x,
\]
where the second term in the square brackets represents the effect of the particle on the field and the Dirac delta function simply picks out where the particle is in the field. The Lagrangian for this particular problem is the quantity in the square brackets.

What are the equations of motion? Recall that
\[
\sum \frac{\partial}{\partial x^\mu} \left( \frac{\partial \mathcal{L}}{\partial \left( \frac{\partial \varphi}{\partial x^\mu} \right)} \right) - \frac{\partial \mathcal{L}}{\partial \varphi} = 0 \rightarrow \frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} - \frac{\partial^2 \varphi}{\partial y^2} - \frac{\partial^2 \varphi}{\partial z^2} + g \delta^3(x) = 0
\]
The motion is of a three dimensional wave, except at the point $(x, y, z) = (0, 0, 0)'$, which designates the position of a particle. Since the particle is not moving, suppose the wave (field) is a standing wave, not varying with time. The equation above then becomes
\[
\nabla^2 \varphi = \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} = g \delta^3(x)
\]
This is the ‘Poisson’s equation’. Recall that Poisson’s equation for gravity is
\[
\nabla^2 \varphi = -4\pi G \rho
\]
If the particle was located not at $(x, y, z) = (0, 0, 0)'$ but at $a' = (a, b, c)$, then $\delta^3(x)$ becomes $\delta^3(x - a')$. If the motion of the particle is time dependent, then $\delta^3(x)$ becomes $\delta^3(x - a'(t))$. In this case,
\[
\frac{\partial^2 \varphi}{\partial t^2} - \nabla^2 \varphi = -g \delta^3(x - a'(t))
\]
Since the particle affects the field, there is no possibility of time-dependent particle motion in a time-independent field. The field will change with time, because of the interaction of the particle with the field.
9.10 A Notational Excursion

A general four dimensional vector, written in tensor notation, is given by $A^\mu$, $\mu = 0, \ldots, 3$.

Let

\[
\eta_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \rightarrow \eta_{\mu\nu} A^\mu = A_\nu = \begin{bmatrix} A^0 \\ -A^1 \\ -A^2 \\ -A^3 \end{bmatrix}, \quad A^\mu = [A^0 \quad A^1 \quad A^2 \quad A^3]
\]

The vector ‘$A^\mu$’ is called a ‘contravariant vector’ and $A_\nu$ is called a ‘covariant vector’. In special relativity, the only difference between a contravariant and a covariant vector is a sign difference in the spatial components of the vector. Note that

\[
\eta_{\mu\nu} A^\mu A^\nu = \langle A_\nu | A^\nu \rangle = (A^0)^2 - (A^1)^2 - (A^2)^2 - (A^3)^2,
\]

which is a scalar, and hence, an invariant in space-time. Note that

\[
\langle A^\mu | B_\mu \rangle = A^0 B_0 + A^1 B_1 + A^2 B_2 + A^3 B_3,
\]

which is also a scalar.

The value of a function ‘$\varphi(x^\mu)$’ at two neighboring points is given by the total differential:

\[
\frac{\partial \varphi}{\partial x^\mu} dx^\mu
\]

Since $\varphi$ is a scalar, then $\partial \varphi / \partial x^\mu$ is a 4-vector i.e.

\[
\frac{\partial \varphi}{\partial x^\mu} \equiv \begin{bmatrix} \frac{\partial \varphi}{\partial t} \\ -\frac{\partial \varphi}{\partial x^1} \\ -\frac{\partial \varphi}{\partial x^2} \\ -\frac{\partial \varphi}{\partial x^3} \end{bmatrix}
\]

It is a covariant 4-vector. Note that

\[
\frac{\partial \varphi}{\partial x^\mu} dx^\mu
\]

is a scalar, since the $\mu$’s are summed over. Hence, $dx^\mu$ must be a 4-vector. In fact, it is a contravariant 4-vector. Now $\partial \varphi / \partial x^\mu$ is normally written $\partial_\mu \varphi$ to indicate that it is a covariant 4-vector and $\partial^\mu \varphi$ indicates a contravariant 4-vector i.e.
\[
\frac{\partial}{\partial t} \varphi = \begin{bmatrix}
\frac{\partial \varphi}{\partial t} \\
\frac{\partial \varphi}{\partial \varphi} \\
\frac{\partial \varphi}{\partial x^1} \\
\frac{\partial \varphi}{\partial x^2} \\
\frac{\partial \varphi}{\partial x^3}
\end{bmatrix}, \quad \varphi = \begin{bmatrix}
\frac{\partial \varphi}{\partial t} \\
\frac{\partial \varphi}{\partial x^1} \\
\frac{\partial \varphi}{\partial x^2} \\
\frac{\partial \varphi}{\partial x^3}
\end{bmatrix}
\]

So,

\[
\varphi \frac{\partial}{\partial \varphi} \varphi = (\frac{\partial \varphi}{\partial t})^2 - (\frac{\partial \varphi}{\partial x^1})^2 - (\frac{\partial \varphi}{\partial x^2})^2 - (\frac{\partial \varphi}{\partial x^3})^2
\]

The equation above is a valid Lagrangian density function and is a scalar, and hence, invariant in every inertial reference frame.

The Lorentz transformation can be expressed in tensor notation i.e.

\[
(A')^\mu = L_\nu^\mu A^\nu, \quad L_\nu^\mu = \begin{bmatrix}
1 \\
\frac{-v}{\sqrt{1-v^2}} & 0 & 0 \\
\frac{-v}{\sqrt{1-v^2}} & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

where \((A')^\mu\) and \(A^\nu\) are 4-vectors. The transformation shown above is the transformation of a contravariant 4-vector. The transformation for a covariant 4-vector is given by

\[
A'_\mu = M_\mu^\nu A^\nu, \quad M = \eta L \eta
\]

### 9.11 The Klein-Gordon Equation

Suppose

\[
\mathcal{L} = \frac{1}{2} \left[ (\frac{\partial \varphi}{\partial t})^2 - (\frac{\partial \varphi}{\partial x^1})^2 - (\frac{\partial \varphi}{\partial x^2})^2 - (\frac{\partial \varphi}{\partial x^3})^2 - \mu^2 \varphi^2 \right],
\]

which is a scalar. Recall that

\[
\sum_\mu \frac{\partial}{\partial x^\mu} \left( \frac{\partial \mathcal{L}}{\partial (\frac{\partial \varphi}{\partial x^\mu})} \right) - \frac{\partial \mathcal{L}}{\partial \varphi} = 0 \rightarrow \frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} - \frac{\partial^2 \varphi}{\partial y^2} - \frac{\partial^2 \varphi}{\partial z^2} + \mu^2 \varphi = 0
\]

The equation on the right above is called the ‘Klein-Gordon equation’, named for Oskar Klein and the German theoretical physicist Walter Gordon (August 13, 1893 –
December 24, 1939). It is often described as a relativistic version of the Schrödinger equation, although it has some undesirable properties.

Let

$$\varphi = e^{-i\omega t} e^{-i(k_1 x^1 + k_2 x^2 + k_3 x^3)} = e^{-i(k_\mu x^\mu)}, \quad k_\mu = a \text{ constant}, \quad k_0 = \omega$$

then

$$\frac{\partial^2 \varphi}{\partial t^2} = -\omega^2 \varphi, \quad \frac{\partial^2 \varphi}{\partial x_i} = -k_i \varphi, \quad i = 1, \ldots, 3$$

Hence, the equation of motion must be

$$\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} - \frac{\partial^2 \varphi}{\partial y^2} - \frac{\partial^2 \varphi}{\partial z^2} + \mu^2 \varphi = (-\omega^2 + k_1^2 + k_2^2 + k_3^2 + \mu^2) \varphi = 0$$

Since \(\varphi\) is arbitrary, the expression in the parentheses is zero. Solving for \(\omega\),

$$\omega = \pm \sqrt{k_1^2 + k_2^2 + k_3^2 + \mu^2}$$

This is the solution to the Klein-Gordon equation.

**9.12 Electro-Dynamics in Special Relativity**

A tensor \(T^{ab}\) is not, in general, equal to the tensor \(T^{ba}\). If \(T^{ab} = T^{ba}\), the tensor is called 'symmetric'. If \(T^{ab} = -T^{ba}\), the tensor is called 'anti-symmetric' (see Book II: Chapter 13). If \(T^{ab}\) is a \(4 \times 4\) anti-symmetric tensor, it can be represented by a square matrix. For \(T^{ab}\) to be anti-symmetric, its diagonal elements must be zero. For instance,

$$T^{ab} = \begin{bmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & -B_3 & B_2 \\ E_2 & B_3 & 0 & -B_1 \\ E_3 & -B_2 & B_1 & 0 \end{bmatrix}$$

is anti-symmetric.

**9.12.1 The Principles of Relativistic Physics**

Relativistic physics is based on the following collection of principles:

1. Least action
2. Locality
3. Lorentz Invariance
4. Gauge invariance
1. The concepts of ‘symmetry’ and ‘conservation’ would not be possible without
the principle of ‘least action’; the path a particle takes is stationary, and hence,
invariant

2. Locality: the principle that says whatever happens at some point in a system
affects things only in the immediate vicinity of the point, not things that are
sufficiently far away. Mathematically, locality is expressed by a Lagrangian that
depends only on the value of the field at a particular point and its first derivatives
i.e. \( A_P = \int L(x, \dot{x}) \, dt \) and \( A_F = \int L(\varphi, \varphi_\mu) \, d^4x \), where \( A_P \) is the ‘action of a
particle’ and \( A_F \) is the ‘action of a field’; note that \( dt \) and \( d^4x \) represent small
displacements, either in the trajectory of the particle or in the wave motion of the
field at the place where the displacement occurs

3. Lorentz invariance: the laws of physics should be the same for all inertial
observers, in other words, the Lagrangian should be a scalar

4. Gauge invariance will be discussed shortly

9.12.2 The Vector Potential and the Lorentz Force Law

Recall the Lorentz force law for electromagnetism:

\[
|f| = m|a| = -e(\langle E \rangle + \langle v \times B \rangle),
\]

where \( e \) is the electric charge, \( \langle E \rangle \), the electric field, \( \langle B \rangle \), the magnetic field and \( \langle v \rangle \) is the
velocity of the particle.

The action of the particle is

\[
A_P = \int L(x, \dot{x}) \, dt = \int -m \frac{dx}{dt} \, dt = \int -m \sqrt{1 - \dot{x}^i \dot{x}_i} \, dt, \quad i = 1, \ldots, 3
\]

The quantity ‘\( A_P \)’ is invariant because \( m \, dt \) is a scalar. The action for the field is given by

\[
A_F = \int L(\varphi, \varphi_\mu) \, d^4x = \int -e A_\mu dx^\mu = \int -e A_\mu \frac{dx^\mu}{dt} \, dt
\]

\[
= \int -e \left( A_0 + A_x \frac{dx}{dt} + A_y \frac{dy}{dt} + A_z \frac{dz}{dt} \right) \, dt = \int -e \left( A_0 + A_i \dot{x}_i \right) \, dt,
\]

where \( A_\mu \equiv A_\mu (x^\mu) \) is a function of a 4-vector and \( e A_\mu dx^\mu \) is a scalar, since \( e \) is a scalar
and \( A_\mu dx^\mu \) is the inner product of two 4-vectors, which is a scalar. Hence,

\[
A_P + A_F = \int -m \sqrt{1 - \dot{x}^i \dot{x}_i} \, dt + \int -e \left( A_0 + A_i \dot{x}_i \right) \, dt
\]

\[
= \int \left[ -m \sqrt{1 - \dot{x}^i \dot{x}_i} - e \left( A_0 + A_i \dot{x}_i \right) \right] \, dt = \int \mathcal{L} \, dt,
\]

\[
\mathcal{L} = -m \sqrt{1 - \dot{x}^i \dot{x}_i} - e \left( A_0 + A_i \dot{x}_i \right)
\]
Note that $\mathcal{L}$ is Lorentz invariant.

The next step is to calculate the equations of motion:

$$
\frac{\partial \mathcal{L}}{\partial \dot{x}^i} = m \frac{\dot{x}^i}{\sqrt{1 - \dot{x}^i \dot{x}^i}} - eA_i = mU^i - eA_i \rightarrow \frac{d}{dt} \left( \frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) = \frac{d}{dt} (mU^i) - e \left( \frac{\partial A_i}{\partial x^0} + \frac{\partial A_i}{\partial x^n} \dot{x}^n \right),
$$

$n = 1, \ldots, 3$

since $A_\mu(t, x^n)$.

And,

$$
\frac{\partial \mathcal{L}}{\partial x^i} = -e \left( \frac{\partial A_0}{\partial x^i} + \frac{\partial A_n}{\partial x^i} \dot{x}^n \right), \quad i = 0, 1, 2, 3
$$

Therefore,

$$
\frac{d}{dt} \left( \frac{\partial \mathcal{L}}{\partial \dot{x}^i} \right) - \frac{\partial \mathcal{L}}{\partial x^i} = \frac{d}{dt} (mU^i) - e \left( \frac{\partial A_i}{\partial x^0} + \frac{\partial A_i}{\partial x^n} \dot{x}^n \right) + e \left( \frac{\partial A_0}{\partial x^i} + \frac{\partial A_n}{\partial x^i} \dot{x}^n \right) = 0 \rightarrow \frac{d}{dt} (mU^i)
$$

Grouping the terms together, on the right-side of this equation, which do not have coefficients of $\dot{x}^n$ and doing the same for the terms which do have coefficients of $\dot{x}^n$ gives

$$
\frac{d}{dt} (mU^i) = e \left( \frac{\partial A_0}{\partial x^0} - \frac{\partial A_0}{\partial x^i} \right) + e \left( \frac{\partial A_n}{\partial x^n} - \frac{\partial A_n}{\partial x^i} \right) \dot{x}^n
$$

Now let

$$
E_i = \frac{\partial A_i}{\partial x^0} - \frac{\partial A_0}{\partial x^i}, \quad \left( \frac{\partial A_i}{\partial x^n} - \frac{\partial A_n}{\partial x^i} \right) \dot{x}^n = \langle v | \times | B \rangle_i,
$$

then

$$
\frac{d}{dt} (mU^i) = e (E_i + (v \times B)_i),
$$

which is in the form of the Lorentz force law.
The left-side of the equation above is not invariant, since it is not written in terms of $\tau$. To remedy this, let

$$\frac{d}{dt} (mU^i) = m \frac{d\tau}{d\tau} \frac{dU^i}{d\tau} = e \left( \frac{\partial A_i}{\partial x^0} - \frac{\partial A_0}{\partial x^i} \right) \frac{dx}{d\tau} + e \left( \frac{\partial A_n}{\partial x^i} - \frac{\partial A_i}{\partial x^n} \right) \frac{dx}{d\tau} \frac{d^2x^i}{d\tau^2} \cdot$$

Note that

$$m \frac{dU^i}{d\tau} = m \frac{d^2x^i}{d\tau^2},$$

which is the equation for a force called the 'proper force', since it is given in terms of $\tau$. Hence,

$$m \frac{d^2x^i}{d\tau^2} = e \left( \frac{\partial A_i}{\partial x^0} - \frac{\partial A_0}{\partial x^i} \right) U^0 - e \left( \frac{\partial A_n}{\partial x^i} - \frac{\partial A_i}{\partial x^n} \right) U^n \rightarrow m \frac{d^2x^\mu}{d\tau^2} = F_\nu^\mu U^\nu$$

Note that, in the equation on the right, the index ‘$i$’, which indicates a 3-vector, has been replaced with $\mu$, which indicates a 4-vector. This is allowed because if three components of a 4-vector are correct, the fourth component must be correct.

In the equation above,

$$\left( \frac{\partial A_i}{\partial x^0} - \frac{\partial A_0}{\partial x^i} \right), \quad \left( \frac{\partial A_n}{\partial x^i} - \frac{\partial A_i}{\partial x^n} \right)$$

have been replaced by the tensor ‘$F_\nu^\mu$’. Note that $F_\nu^\mu U^\nu$ is a 4-vector, where

$$F_\nu^\mu = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix}$$

If an index is lowered in $F_\nu^\mu$ to make $F_\nu^\mu$, the only difference in the matrix is a sign difference in the space related components, since, in special relativity, the only difference between a contravariant and covariant 4-vector is a sign change in the spatial components. The signs of the time components remain unchanged.
9.12.3 Gauge Invariance

Suppose there is a field ‘\(\varphi(t, x^i)\)’ in which the Lagrangian involving \(\varphi\) is invariant under some change. For example, if the equation of motion is
\[
\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^i \partial x^i} = 0,
\]
then if \(\varphi' = \varphi + C, \ C = a \text{ constant}\) is substituted for \(\varphi\), this substitution does not change the equations of motion.

However, suppose the equations of motion are altered to read
\[
\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^i \partial x^i} + \mu^2 \varphi = 0,
\]
then the substitution \(\varphi' = \varphi + C\) does not leave the equations of motion invariant, since \(\mu^2 \varphi \neq \mu^2 (\varphi + C)\).

Consider \(-e \int A_\mu dx^\mu\) and let \(A_\mu\) be replaced by \(A_\mu + \partial S/\partial x^\mu\), where \(S\) is an arbitrary scalar. This substitution changes the action i.e.
\[
A = -e \int_{P_1}^{P_2} A_\mu dx^\mu \rightarrow -e \int_{P_1}^{P_2} A_\mu dx^\mu + \partial S/\partial x^\mu dx^\mu = -e \int_{P_1}^{P_2} A_\mu dx^\mu - e \int_{P_1}^{P_2} \partial S/\partial x^\mu dx^\mu \neq A
\]
But note that
\[
\int_{P_1}^{P_2} \frac{\partial S}{\partial x^\mu} dx^\mu = S(P_2) - S(P_1),
\]
which only depends on the end points ‘\(S(P_2), S(P_1)\)’. Since the principle of least action requires that the end points be fixed, adding any 4-vector gradient ‘\(\partial S/\partial x^\mu\)’ to the action does not alter the equations of motion. To see this, recall that the motion in an electromagnetic field is given by
\[
m \frac{d^2 x^\mu}{d\tau^2} = F^\mu_v U^v
\]
Note that \(U^v\) is a 4-vector that depends on \(\tau\). So anything that does not change \(F^\mu_v\), does not change the equations of motion. If
\[
F^\mu_v = \frac{\partial A_\mu}{\partial x^v} - \frac{\partial A_v}{\partial x^\mu},
\]
make the substitution ‘\(A_\mu \rightarrow A_\mu + \partial S/\partial x^\mu\)’. Then
If the motion of the field is sufficiently smooth, then

$$\frac{\partial^2 S}{\partial x^\mu \partial x^\nu} - \frac{\partial^2 S}{\partial x^\nu \partial x^\mu} = 0,$$

since the order in which the partial derivatives are taken is immaterial. Any substitution '$$A_\mu \rightarrow A_\mu + \frac{\partial S}{\partial x^\mu}$$', where the gradient of a scalar is added to the Lagrangian, does not alter the equations of motion. Such a substitution is called 'gauge invariant', which is useful in simplifying many problems and in emphasizing certain aspects of the physics, but does not change the value of the electromagnetic field.

### 9.13 A Relook at Maxwell’s Equations in Terms of Special Relativity

If there is a magnetic field ‘$$|B_y\rangle$$’ confined to the $$x, y$$-plane and if a copper wire, represented by the dashed line (see fig. 9.13-1), is moving in the $$x$$-direction, then the magnetic field (at rest) will exert a force (emf) on the wire creating an electric current in the wire. According to the principle of relativity, an electric current will flow whether the wire is moving and the magnetic field is stationary or the wire is stationary and the magnetic field is moving.

**Figure 9.13-1.**

The electromagnetic tensor is given by

$$F_{\mu\nu} = \begin{bmatrix} t & x & y & z \\ 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{bmatrix}, \quad t = 0, \quad x = 1, \quad y = 2, \quad z = 3$$
The electric field components specified by \( F_0 \) and \( F_\nu \) are space-time components. The \( F_{ij} \)'s, \( i, j = 1, \ldots, 3 \) specify the components of the magnetic field and are purely spatial components. Recall that the Lorentz transformation can be written

\[
(x')^\mu = L^\mu_\nu x^\nu \rightarrow \begin{bmatrix} x'^0 \\ x'^1 \\ x'^2 \\ x'^3 \end{bmatrix} = L \begin{bmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{bmatrix}, \quad L = \begin{bmatrix} 1 \\ \frac{-v}{\sqrt{1-v^2}} \\ \frac{1}{\sqrt{1-v^2}} \\ 0 \end{bmatrix} \begin{bmatrix} \frac{-v}{\sqrt{1-v^2}} \\ \frac{1}{\sqrt{1-v^2}} \\ 0 \end{bmatrix}
\]

If the motion is along the \( x \)-axis, then

\[
t' = x'^0 = \frac{t - vx}{\sqrt{1-v^2}}, \quad x' = x'^1 = \frac{x - vt}{\sqrt{1-v^2}}, \quad y' = x'^2 = y, \quad z' = x'^3 = z
\]

It is always possible to rotate a coordinate system so that the motion is along a chosen axis. Since \( F_{\mu\nu} \) is a tensor, it transforms as

\[
F'^{\mu}_{\nu} = L^\mu_\sigma L^\nu_\tau F_{\sigma\tau}
\]

In fig. 9.13-1, if the magnetic field is stationary in the \( x, y \)-plane and the wire is moving in the \( x \)-direction, an electric force is created in the \( z \)-direction i.e.

\[
F'_{03} = E'_3 = L^0_1 L^3_3 F_{13}
\]

Note that \( L^3_3 = 1 \) and \( F_{13} = -B_y \). Hence,

\[
F'_{03} = E'_3 = L^0_1 (-B_y) = \frac{v}{\sqrt{1-v^2}} B_y \neq 0, \quad L^0_1 = \frac{-v}{\sqrt{1-v^2}}
\]

In the primed frame, the wire is stationary and the magnet is moving, but the wire still experiences an electric force in the \( z \)-direction, which is consistent with the principle of relativity.

9.13.1 The Vector Potential and Maxwell’s Equations

If \( A_i \), \( i = 1, \ldots, 3 \) is an arbitrary 3-vector, then \( \langle \nabla \times |A_i| \rangle \) is the ‘curl’ of \( A \). Moreover, by ordinary vector operations, \( \langle \nabla | \langle \nabla \times |A_i| \rangle \rangle = 0 \) and \( \langle \nabla \rangle \times \langle \nabla S \rangle = 0 \), where \( S \) is any scalar. These are simply identities – true of any 3-vector. The magnetic field \( |B| \) is defined as \( \langle B \rangle = \langle \nabla \times |A_i| \rangle \), where, in this case, \( A_i \) represents the vector potential. Hence, \( \langle \nabla |B| \rangle = \langle \nabla | \langle \nabla \times |A_i| \rangle \rangle = 0 \). The equation \( \langle \nabla |B| \rangle = 0 \) is one of Maxwell’s equations. Moreover, \( |E| = \partial A_i / \partial t - \langle \nabla |A_0 | \rangle \). Hence,

\[
\langle \nabla \times |E| \rangle = \frac{\partial}{\partial t} \langle \nabla \times |A_i| \rangle - \langle \nabla \rangle \times \langle \nabla |A_0 | \rangle,
\]
the second term on the right-side of the equation above is $|0\rangle$. Therefore,

$$
\langle \nabla \times |E\rangle = \frac{\partial}{\partial t} \langle \nabla \times |A_i\rangle = -\frac{\partial |B\rangle}{\partial t} \rightarrow \langle \nabla \times |E\rangle + \frac{\partial |B\rangle}{\partial t} = |0\rangle
$$

The equation on the right-side is another of Maxwell’s equations; a changing magnetic flux is associated with an electric field.

If an amount of charge is moving through a small volume of space $\Delta x\Delta y\Delta z$, then the amount of charge within that volume in any instant of time is called the ‘charge density’ ($\rho$), defined as $\rho = \Delta q/\Delta V$, where $\Delta q$ is the amount of charge and $\Delta V$ is the volume element (see fig. 9.13.1-1). The rhombus marked ‘$\rho$’ in the figure is really a volume as it is difficult to draw a four-dimensional figure. The surface of the rhombus is perpendicular to the $t$-axis.

![Figure 9.13.1-1](image)

The rhombus marked ‘$J$’ is also a volume, representing the amount of charge passing through an area $A_{yz}$ per unit time parallel to the $x$-axis. The surface area $A_{yz}$ is perpendicular to the $x$-axis; the amount of charge in this volume is called the ‘current density’ ($J_x$) in the $x$-direction i.e.

$$
J_x = \frac{\Delta q}{A_{yz}\Delta t} \rightarrow J_y = \frac{\Delta q}{A_{xz}\Delta t} \rightarrow J_z = \frac{\Delta q}{A_{xy}\Delta t}
$$

If the amount of charge in a unit volume changes over time, then $d\rho/dt = \dot{\rho}$. If the current density changes, then

$$
\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z},
$$

where, evidently, by convention, a ‘−’ sign indicates an increase in current density and a ‘+’ sign indicates a decrease in current density.
There are two more Maxwell equations in addition to the ones discussed above:

\[ \langle \nabla | E \rangle = \rho, \quad \langle \nabla \times | B \rangle - \frac{\partial | E \rangle}{\partial t} = J \]

Note that

\[ \langle \nabla | \frac{\partial E}{\partial t} \rangle = \dot{\rho}, \quad \langle \nabla \langle \nabla \times | B \rangle \rangle - \langle \nabla \frac{\partial E}{\partial t} \rangle = \frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} \]

Adding together the two equations above gives

\[ \langle \nabla \frac{\partial E}{\partial t} \rangle + \langle \nabla \langle \nabla \times | B \rangle \rangle - \langle \nabla \frac{\partial E}{\partial t} \rangle = \dot{\rho} + \frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} = 0, \]

since \( \langle \nabla \langle \nabla \times | B \rangle \rangle = 0 \), the divergence of the curl of a spatial vector is zero. Hence,

\[ \int \left( \dot{\rho} + \frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z} \right) d^4x = C, \quad C = \text{a constant} \]

Therefore, charge is locally conserved. Such a conservation law is a consequence of Maxwell’s equations.

9.13.2 The Bianchi Identity

Recall two of Maxwell equations: \( \langle \nabla | B \rangle = 0 \) and \( \langle \nabla \times | E \rangle + \partial | B \rangle / \partial t = |0 \). And consider the following equation:

\[ \partial_\sigma F_{\nu \tau} + \partial_\nu F_{\sigma \tau} + \partial_\tau F_{\sigma \nu} = 0, \quad F_{\mu \nu} = \begin{bmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{bmatrix} \]

Suppose that \( \sigma = x, \ \nu = y, \ \tau = z \), then

\[ \partial_x F_{yz} + \partial_y F_{zx} + \partial_z F_{xy} = -\partial_x B_x - \partial_y B_y - \partial_z B_z = -\langle \nabla | B \rangle = 0, \]

the first of Maxwell’s laws. Now suppose \( \sigma = x, \ \nu = y, \ \tau = t \), then

\[ \partial_x F_{yt} + \partial_y F_{tx} + \partial_t F_{xy} = -\partial_x E_y + \partial_y E_x - \partial_t B_z = -\left( \langle \nabla \times | E \rangle + \frac{\partial | B \rangle}{\partial t} \right) = |0, \]

the second of Maxwell’s equations. The upshot is that the equation

\[ \partial_\sigma F_{\nu \tau} + \partial_\nu F_{\sigma \tau} + \partial_\tau F_{\sigma \nu} = 0 \]

is equivalent to Maxwell’s first and second law. This identity is called the ‘Bianchi identity’.
9.13.3 The Last Two of Maxwell’s Equations

What about Maxwell’s third and fourth laws i.e. \( \langle \nabla | E \rangle = \rho \), \( \langle \nabla \times | B \rangle - \partial | E \rangle / \partial t = | J \rangle \)? First, let \( \rho = J^0 \), so that \( \partial \mu F^{\mu \nu} = J^\nu \) and note that \( J^\nu \) is a 4-vector. The third and fourth equations of Maxwell’s are derived from the principle of least action, and hence, from a Lagrangian.

Whatever Lagrangian is chosen, it must comply with the principle of locality:

\[
\int L(\varphi_n, \varphi_{n,\mu}) d^4x, \quad \varphi_{n,\mu} = \frac{\partial \varphi_n}{\partial x^\mu} = \partial_\mu \varphi_n, \quad n \in N
\]

where ‘\( \varphi_n \)’ signifies that there are \( n \) fields in the Lagrangian. The Lagrangian chosen can only depend on the \( \varphi_n \)’s and their first derivatives. In addition, \( L \) should be Lorentz invariant i.e. \( L = \text{a scalar} \). Since electromagnetic phenomena are gauge invariant, this suggests selecting a Lagrangian that is gauge invariant.

If \( F_{\mu \nu} \) is chosen for the Lagrangian, it would be gauge invariant, but \( F_{\mu \nu} \) is a tensor, not a scalar. One way of making \( F_{\mu \nu} \) into a scalar is to raise one of the indices and then contract i.e. \( F_{\mu \nu} \rightarrow F^\nu_\mu \rightarrow F^\mu_\mu \). But, \( F^\mu_\mu \) represents the sum of the diagonal elements, which are all zero, making \( F^\mu_\mu = 0 \) trivial. What about \( F_{\mu \nu} F^{\mu \nu} \)? It is gauge invariant and a scalar; but what does \( F_{\mu \nu} F^{\mu \nu} \) represent? Note that

\[
F_{0i} F^{0i} = -2E^2, \quad i = 1, \ldots, 3,
\]

since raising the time index changes the sign, hence, the minus sign. The ‘2’ enters, since, for every \( F_{0i} \), there is a \( F_{i0} \). Moreover,

\[
F_{ij} F^{ij} = 2B^2, \quad i, j = 1, \ldots, 3
\]

Hence,

\[
F_{\mu \nu} F^{\mu \nu} = -2E^2 + 2B^2 \rightarrow -\frac{1}{4} F_{\mu \nu} F^{\mu \nu} = \frac{1}{2} (E^2 - B^2),
\]

the ‘\(-1/4\)’ added for purely conventional reasons.

Recall that

\[
F_{\mu \nu} = \frac{\partial A_\mu}{\partial x^\nu} - \frac{\partial A_\nu}{\partial x^\mu} = A_{\mu,\nu} - A_{\nu,\mu},
\]

where \( A_\mu \) is the vector potential and consider the Lagrangian

\[
L = -\frac{1}{4} F_{\mu \nu} F^{\mu \nu} = -\frac{1}{4} (A_{\mu,\nu} - A_{\nu,\mu}) (A^\mu_\nu - A^\nu_\mu)
\]
The equations of motion are
\[ \frac{\partial}{\partial x^\nu} \left( \frac{\partial \mathcal{L}}{\partial A_{\mu,\nu}} \right) = \frac{\partial \mathcal{L}}{\partial A_\mu} \]

There is an equation of motion for each \( A_\mu \), four total. Suppose a particular term of \( \partial \mathcal{L}/\partial A_{\mu,\nu} \) is computed, say \( 'A_{x,y}' \). That term is
\[ \mathcal{L} = -\frac{1}{2} \left( \frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x} \right)^2 = -\frac{1}{2} \left[ \left( \frac{\partial A_x}{\partial y} \right)^2 - 2 \frac{\partial A_x}{\partial y} \frac{\partial A_y}{\partial x} + \left( \frac{\partial A_y}{\partial x} \right)^2 \right] \]

The \( '1/2' \) comes in because of repeated terms. There are other terms in the Lagrangian, but do not involve \( x \) and \( y \), and therefore, their derivatives are zero. Hence,
\[ \frac{\partial \mathcal{L}}{\partial A_{x,y}} = -\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} = -A_{x,y} + A_{y,x} = A_{y,x} - A_{x,y} = -F_{xy} \]

In general,
\[ \frac{\partial \mathcal{L}}{\partial A_{\mu,\nu}} = -F_{\mu\nu} \rightarrow \frac{\partial}{\partial x^\nu} \left( \frac{\partial \mathcal{L}}{\partial A_{\mu,\nu}} \right) = -F_{\mu\nu} \]

The equation on the right is the left-hand side of the equation of motion. The right-hand side is \( \partial \mathcal{L}/\partial A_\mu \), but \( A_\mu \) does not appear undifferentiated in \( \mathcal{L} \). Thus, \( \partial \mathcal{L}/\partial A_\mu = 0 \).
Hence, the equations of motion are
\[ \frac{\partial F_{\mu\nu}}{\partial x^\nu} = 0 \]

But this is not quite the whole story. Recall the continuity equations \( \partial_\mu J^\mu = 0 \) and consider \( 'A = -\int J^\mu A_\mu d^4x \rightarrow \mathcal{L} = J^\mu A_\mu' \). Note that \( J^\mu A_\mu \) is a scalar. It is gauge invariant. To see, suppose
\[ A = -\int J^\mu A_\mu d^4x \rightarrow A = -\int J^\mu A_\mu d^4x - \int J^\mu \frac{\partial S}{\partial x^\mu} d^4x, \]
the last term representing a gauge transformation on the action. Integration by parts gives
\[ \int J^\mu \frac{\partial S}{\partial x^\mu} d^4x = -\int \frac{\partial J^\mu}{\partial x^\mu} S d^4x = 0, \]
because \( \partial J^\mu/\partial x^\mu = \partial_\mu J^\mu = 0 \); therefore, \( \mathcal{L} = J^\mu A_\mu \) is a scalar and is gauge invariant.
Hence, the term \( J^\mu A_\mu \) can be added to the Lagrangian i.e.
\[ \mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - J^\mu A_\mu \]

Note that the equation above is true if and only if the current is conserved. How does the addition of \( J^\mu A_\mu \) change the equations of motion? Recall that

\[ \frac{\partial}{\partial x^\nu} \left( \frac{\partial \mathcal{L}}{\partial A_{\mu,\nu}} \right) = -\frac{\partial F_{\mu\nu}}{\partial x^\nu} = 0 \]

But,

\[ \frac{\partial \mathcal{L}}{\partial A_\mu} \left( -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - J^\mu A_\mu \right) = -J^\mu \]

Hence,

\[ \frac{\partial}{\partial x^\nu} \left( \frac{\partial \mathcal{L}}{\partial A_{\mu,\nu}} \right) = \frac{\partial \mathcal{L}}{\partial A_\mu} \rightarrow -\frac{\partial F_{\mu\nu}}{\partial x^\nu} = -J^\mu \rightarrow \partial_\nu F_{\mu\nu} = J^\mu \]

The equation on the right above is the third and fourth of Maxwell’s equations written in tensor notation.

**9.14 Concluding Remarks**

The theory of special relativity arose because of a need to explain the null result of the Michelson/Morley experiment. How could a light wave travel through empty space without a conducting medium? When no medium was detected, it threatened the trustworthiness of Galilean relativity, a principle upon which all physics known at the time depended. The principle of ‘Galilean relativity’ was saved by a radical shift in perspective. The concept of ‘absolute rest’ was given up, replaced by the “principle of relativity” and the constancy of the speed of light in empty space. Absolute time and space faded away. Each observer was equipped with their own clock and measuring rod and the measurements obtained from those measuring devices depended on the relative motion of one observer with respect to another. Moreover, the notions of ‘space’ and ‘time’ became part of a new concept called ‘space-time’, where space and time were, in some sense, inseparable. Any change in the motion of an observer not only affected spatial measurements, but time measurements as well.

Since space and time were no longer separate concepts, physical laws written solely in terms of them led to laws that were not universal, but depended on the state of motion of an observer. Creating universal physical laws became an exercise in identifying the invariants in Nature. And these, in general, were mathematical symmetries, described in terms of physical conservation laws, embodied in Hamilton’s ‘principle of least action’.

Finally, the conflict between Maxwell’s equations and Galilean relativity was overcome by showing that those equations could be derived from a vector potential in the form of a 4-vector, which complied with the principle of relativity. In other words, Maxwell’s
equations were a manifestation of the principle of relativity and the constancy of the speed of light.
Chapter 10

The General Theory of Relativity

“In relativity, movement is continuous, causally determinate and well defined; while in quantum mechanics it is discontinuous, not causally determinate and not well defined.”

- David Bohm

10.0 Introduction

Michael Green once lamented that no matter the effort, teaching your cat general relativity would end in failure. The odds are not much greater when the pupil is an uninitiated human learner, since comprehending general relativity requires the student to grasp a myriad of sophisticated mathematical concepts. My treatment of the subject follows closely the lectures of Prof. Leonard Susskind of Stanford University and Dr. Bob Eagle. The reader is encouraged to view their excellent online lectures.

The force of gravity dominates at the larger scales of the Universe, but plays virtually no role in smaller scale physics, except in extreme cases, the physics of a black hole, for example. The laws of gravity are embodied in the general theory of relativity.

10.1 Flat Earth Gravity

The flat Earth approximation to Newton’s law of gravitation considers the Earth infinitely flat, where Newton’s second law becomes $|F⟩ = m|g⟩$. The acceleration vector $|g⟩$ is a constant vector that points downward towards the Earth. The magnitude of the force does not depend on the height above the Earth. The flat Earth approximation is a reasonably good estimate of Newton’s law near the Earth’s surface.

10.1.1 The Principle of Equivalence

The force of gravity is uniquely proportional to the mass of a body. No other force in Nature is strictly proportional to mass. Moreover,

$$|F⟩ = m|g⟩ = m \frac{dz^2}{dt^2} \rightarrow |g⟩ = \frac{dz^2}{dt^2}$$

The acceleration of a body in a gravitational field is independent of its mass, in other words, gravitationally, all bodies accelerate at the same rate. This leads to the ‘principle of equivalence’:

No experiment can be performed which distinguishes between a body falling freely in a uniform gravitational field and one that is at rest in free space, where no gravitational field exists.
In other words, gravitational mass is equivalent to inertial mass\(^8\). Such reckoning led Einstein to the realization that, in a small enough region of space, gravity could be described by the flat Earth approximation i.e. the laws of gravity reduce to the theory of special relativity.

### 10.1.2 Distance Dependent and Tidal Forces

Newton’s gravitational force law ‘\(F = -G M m / r^2\)’, where \(M\) and \(m\) represent the masses of the bodies, \(G\), a constant and \(r\) is the distance between the center of mass of the bodies, is different from the flat Earth approximation in a couple of respects: 1) the force varies with distance ‘\(r\)’, 2) the force does not act in only a single direction. For example, the Moon exerts a gravitational force on the Earth, but the force exerted on the parts of the Earth farther from the center of the Moon will be slightly weaker than the force exerted on the parts of the Earth closer to the Moon; the magnitude of the force depends on distance. Moreover, the gravitational force exerted by the Moon on the top of the Earth differs in direction from the force exerted on the bottom of the Earth (see fig. 10.1.2-1). Such forces are called ‘tidal forces’. Tidal forces represent a deviation from the flat Earth approximation, and therefore, from the principle of equivalence.

![Figure 10.1.2-1](image)

### 10.1.3 The Field Formulation of Newton’s Law

Newton formulated his theory of gravitation as a force that acted instantaneously between two bodies over a distance, which is philosophically problematic. It was more acceptable to describe gravity in terms of accelerations that happened locally, at every point in space. Suppose there are a number of particles of various masses distributed randomly in space. The gravitational force ‘\(|F_i\)’ on the \(i^{th}\) particle due to the presence of all the other particles is

\[
|F_i\rangle = m_i |A_i\rangle = G m_i \sum_{j \neq i} \frac{M_j |R_{ij}\rangle}{R_{ij}^3} \rightarrow |A_i\rangle = G \sum_{j \neq i} \frac{M_j |R_{ij}\rangle}{R_{ij}^3}
\]

The vector ‘\(|A_i(x)\rangle\)’ gives the magnitude and direction of the acceleration at position ‘\(x\)’. There is an ‘\(|A(x)\rangle\)’ at every point in space.

If a vector field ‘\(|A(x)\rangle\)’ depends on position, then the divergence ‘\(\nabla |A\rangle\)’ of \(|A\rangle\) is given by

---

\(^8\)This assumes the gravitational force between the bodies is considered so weak it can be ignored.
\[
\langle \nabla | A \rangle = \frac{\partial A}{\partial x} + \frac{\partial A}{\partial y} + \frac{\partial A}{\partial z},
\]
which is a scalar. Recall that the electric field is predicated on the idea of ‘flux’ \((\phi)\) through a closed surface. By the divergence theorem of Gauss, if \(V\) is the volume bounded by a closed surface ‘\(S\)’ and if \(|A|\) is a vector function of position, then

\[
\iiint_V \langle \nabla | A \rangle \, dV = \int_S \langle A | n \rangle \, dS = \iiint_S \langle A | dS \rangle \rightarrow \phi = \iiint_S \langle A | dS \rangle,
\]
where \(|n\) is the positive unit normal to all points of \(S\) and \(|A|\) represents the field.

**10.1.3.1 Poisson’s Equation**

Gravitational flux depends on the principle of equivalence (inertial mass equivalent to gravitational mass) i.e.

\[
m|a| = -\frac{GmM}{r^2} |e_r| \rightarrow |a| = -\frac{GM}{r^2} |e_r|,
\]
where \(|a|\) is called the ‘gravitational field’ and \(|e_r|\) is a unit vector in the direction of \(r\). Hence,

\[
\phi = -\frac{GM}{r^2} \oint (e_r | dS)
\]

If the surface is a sphere, then

\[
\phi = -\frac{GM}{r^2} \oint (e_r | dS) = -\frac{GM}{r^2} \oint dS = -\frac{GM}{r^2} 4\pi r^2 = -4\pi GM
\]

For a system of bodies, \(M\) is replaced by \(\sum_i M_i\):

\[
\phi = -4\pi G \sum_i M_i
\]

Using the theorem of divergence,

\[
\phi = \oint \langle a | dS \rangle = \int_V \langle \nabla | a \rangle \, d\tau = -4\pi G \int_V \rho \, d\tau \rightarrow \langle \nabla | a \rangle = -4\pi G \rho
\]
i.e.

\[
\phi = -\iiint \left( \frac{\partial^2 a}{\partial x^2} + \frac{\partial^2 a}{\partial y^2} + \frac{\partial^2 a}{\partial z^2} \right) d\tau = -\iiint \langle \nabla \nabla a \rangle \, d\tau = -4\pi GM = -4\pi G \iiint \rho \, d\tau,
\]
where \(d\tau\) is an element of volume and \(\rho\) is the density of matter. Hence,

\[
\langle \nabla^2 | a \rangle = -4\pi G \rho
\]
The equation above, called ‘Poisson’s equation’, is entirely equivalent to Newton’s gravitational force law, but is defined in local terms, rather than by "action-at-a-distance". The force law can be interpreted as though all the mass ‘$M$’ is concentrated at a single point, where the vector field points radially inward toward that point. Any test particle external to $M$ will experience a force that points in the direction of the vector field.

10.2 Gravity and the Curvature of Space

Suppose a spaceship, far away from any other bodies, at time ‘$t = 0$’ instantaneously accelerates with magnitude ‘$g = a\ constant$’ directly upward. If, at $t = 0$, a light ray is shot horizontally across the ship and, since the speed of light is finite, the time for the light ray to travel from one side of the ship to the other will be finite (see fig. 10.2-1). To an observer on the ground, the light ray travels in a straight line. Since the ship is accelerating upward, to an observer in the ship, the light ray will move along a downwardly curved path. By the principle of equivalence, experiments conducted in a uniformly accelerating rocket ship do not differ from those conducted in a flat Earth gravitational field. Therefore, light rays should move along curved paths in a uniform gravitational field.

10.2.1 The Path of a Light Ray in a Uniform Gravitational Field

According to the special theory of relativity, light particles (photons) travel at the speed of light (in a vacuum) and have zero mass. Newton’s law ‘$F = -\frac{GmM}{r^2}$’ predicts, since photons have zero mass, $m = 0 \rightarrow F = 0$. Correspondingly, photons that experience no gravitational force should travel in a straight line. But the principle of equivalence maintains that light should move along a curved path in a gravitational field, which directly contradicts Newton’s law. This discrepancy gave birth to the theory of general relativity.
Ignoring special relativity for the moment and returning to Galilean relativity, suppose the point 'P' in fig. 10.2.1-1 is stationary in the 'x, t'-frame of reference. The spatial coordinates of P are labeled 'x'. Hence, \( x = \text{a constant} \), since 'P' is stationary in the x, t-frame. If the x', t'-frame is moving relative to the x, t-frame with a constant velocity \( v \), then \( x' = x - vt \); since x is a constant, \( dx'/dt = -v \). In the x', t'-frame, the point 'P' moves in the \(-x\)-direction with constant velocity \( v \). Note that \( d^2x'/dt^2 = 0 \). The point 'P' does not accelerate in the x', t'-frame.

Suppose the x', t'-frame is accelerating relative to the x, t-frame with a constant acceleration \(-g\), then \( x' = x - \frac{1}{2} gt^2 \). To see this, note that

\[
\frac{dx'}{dt} = -gt \rightarrow \frac{d^2x'}{dt^2} = -g, \quad dt = dt' \]

### 10.2.1.1 Uniform Acceleration and Space-Time

A uniformly accelerated point in space-time moves along a curved path. This suggests that uniform acceleration, which, by the equivalence principle, is equivalent to a uniform gravitational field, is related to the curvature of space-time. Tidal forces, however, represent an obstruction to formulating a flat Earth theory of gravity consistent with the equivalence principle. Real gravitational fields are functions of distance and direction, which cannot be described by a flat Earth model. But in small enough regions of space and in short enough intervals of time, tidal forces and distance can be ignored and gravity reduces to special relativity, where the principle of equivalence holds.

### 10.3 The Concept of Curvature

Gaining an understanding of the theory of general relativity requires grasping the concept of 'curvature', which is the primary notion upon which Einstein's theory depends. Curvature is a geometrical concept.

#### 10.3.1 Extrinsic vs. Intrinsic Geometry

Euclidean geometry is an 'extrinsic geometry'. All geometric concepts in Euclidean geometry require a fixed background. Whether specified or not, figures and other geometrical concepts in Euclidean geometry are described relative to a fixed coordinate
system. All geometrical figures are imbedded in the background. For instance, the equation of a sphere i.e.

\[(x - a)^2 + (y - b)^2 + (z - c)^2 = r^2,\]

where the point \((a, b, c)\) signifies the center of the sphere, is completely described in terms of the fixed points that make up the background. This is true of virtually all geometric figures in Euclidean geometry (lines, planes, rectangles etc.).

By contrast, geometric concepts, for example curvature on a surface, described without reference to a background space, are called 'intrinsic' geometrical concepts. If a geometry has no background space, it is called 'intrinsic'. There are no completely intrinsic geometries, but there are geometrical concepts, curvature, for example, that are intrinsic. The theory of general relativity is entirely based on intrinsic concepts.

As an aside, the fact that general relativity is geometrically intrinsic creates an obstacle when trying to unify it with quantum mechanics. So far, a way of describing matter fields, which represent particles in quantum mechanics, without first defining a fixed background, has eluded researchers. The key to understanding general relativity is to comprehend its intrinsic geometrical properties and appreciate the role that curvature plays in the theory.

### 10.3.2 Distance as an Intrinsic Concept

To get an inkling of the role that intrinsic geometry and curvature play in general relativity, consider a circle in 2-dimensional Euclidean space, centered at the origin. Since Euclidean geometry is extrinsic, the distance between two arbitrarily chosen points remains constant. In essence, the fixed background determines the metric of the space i.e. how distance is measured. For instance, in 2-dimensional Euclidean space, a distance 's' is given by

\[s = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2},\]

where \((x_1, y_1)\) and \((x_2, y_2)\) are two fixed points in the space. If \(x_1 = x_2\) and \(y_1 = y_2\), then \(s = 0\). There is no distance between a point and itself.

By contrast, if the geometric concept is intrinsic, distance, for example, is measured by first specifying two infinitely close neighboring points \((x, y)\) and \((x + dx, y + dy)\). Where is the point \((x, y)\) in this space? It is everywhere and nowhere. If the geometry is intrinsic, the metric defines not only how distance is measured, but the geometry of the space itself. If the metric is

\[ds = \sqrt{(y + dy - y)^2 + (x + dx - x)^2} = \sqrt{(dy)^2 + (dx)^2},\]

the space is called Euclidean 'flat'. There are many different metrics that define an intrinsic geometry. In general "the metric" is specified by \(ds^2 = g_{ij}x^i x^j\) (note that the \(i, j\)'s are summed over). If the metric is 2-dimensional, then
In a 2-dimentional Euclidean flat space, \( g_{11} = g_{22} = 1, \ g_{12} = g_{21} = 0 \).

Figure 10.3.2-1

Intrinsic metrics are not unique. For instance, a flat 2-dimentional plane can be described by any of the coordinates systems illustrated in fig. 10.3.2-1. If the coordinates are orthogonal (fig. 10.3.2-1 A), then \( g_{12} = g_{21} = 0 \). If the coordinates are not orthogonal (fig. 10.3.2-1 B), then \( g_{12} = g_{21} \neq 0 \). If the coordinate axes are not the same distance apart, then \( g_{11} \neq g_{22} \). If the coordinate axes are curved (fig. 10.3.2-1 C), the \( g_{ij}'s \) are not constant, but functions of position.

10.4 Flat vs. Curved Spaces and Coordinate Transformations

Given any 2-metric, if there is a coordinate transformation into

\[
ds^2 = g_{11}(x^1)^2 + g_{22}(x^2)^2, \quad g_{11}, g_{22} = \text{a constant},
\]

then the geometry is called ‘flat’. This is also true of higher dimensional geometries. If there is no coordinate transformation into \( ds^2 = g_{11}(x^1)^2 + g_{22}(x^2)^2 \), then the geometry of the space has ‘curvature’. Note that simply because a space is expressed in curvilinear coordinates does not imply it has curvature. For instance, a space can be described in polar coordinates \('(r, \theta)'\), which are curvilinear, but the geometry of the space can be flat. A cylinder does not look flat, but can be cut lengthwise and laid flat into a rectangle, without sacrificing any surface material from the cylinder. Hence, the geometry of a cylinder is flat. The ability to flatten the cylinder without removing any surface material is the analog of finding a coordinate transformation that gives a flat metric.

On the other hand, a flat metric cannot describe the geometry of a sphere. A sphere is not a flat surface. The rule is that if a coordinate transformation can be found such that the metric is Euclidean, then the geometry of the surface is intrinsically flat. The concepts of ‘flat' and 'curved' spaces play a fundamental role in the theory of gravity,
since flat spaces are associated with flat Earth gravity, while curved spaces are the analog to the tidal forces.

10.4.1 The Consequences of Describing Gravity Intrinsically

However, describing gravity in terms of intrinsic geometry comes with unanticipated and, in some sense, unwelcomed consequences. Since the metric $\text{d}s$ defines the geometry as well as distance, what is the significance of $\text{d}s = 0$ i.e. $\forall i \forall j (i, j \in N \rightarrow g_{ij} = 0)$? If the geometry is extrinsic, the fact that $s = 0$ simply indicates there is no distance between a point and itself. But since $\text{d}s$ defines the geometry of the space, if $\forall i \forall j (i, j \in N \rightarrow g_{ij} = 0)$, there is no geometry, and hence, no space. This becomes problematical because the curvature of space, roughly speaking, is inversely proportional to the metric. It follows, then, that

$$\forall i \forall j (i, j \in N \rightarrow g_{ij} = 0 \rightarrow 1/g_{ij} = \infty)$$

A black hole has infinite curvature. One of the reasons a black hole is called ‘black’ is because it does not have a geometry. It has no form. Moreover, the Universe supposedly began from a singular point in accordance with the big bang theory. But in order to answer the question "where is the point of creation?", the geometry of space-time must be extrinsic. An intrinsically described space-time does not have a specified location in space. If the Universe began $\forall i \forall j (i, j \in N \rightarrow g_{ij} = 0)$, then how did it evolve to $\exists i \exists j (i, j \in N \rightarrow g_{ij} \neq 0)$? In other words, how was space-time created virtually from nothing?

This leads to all manner of speculative ideas – multiple parallel universes, theories on how the big bang began and what the Universe will be like in the future. Will it go on forever or die in a big crunch? However, one thing is certain, when the metric gets too small the physics of the general theory of relativity breaks down, failing to predict the correct physical outcomes and sending physicists to the safer harbingers of quantum physics. Nevertheless, the concepts of ‘intrinsic geometry’ and ‘curvature’ are fundamental to general relativity and the remainder of this chapter will be devoted to understanding how these concepts evolve into a theory of gravity.

10.5 Mathematical Preliminaries

Given a scalar function of position $\varphi(x_i), \ i = 1, ..., n$, its total differential is

$$d\varphi = \frac{\partial \varphi}{\partial x_i} dx_i$$

(the $i$’s are repeated, and hence, summed over). The total differential represents the change in the field $\varphi$ given a slight change in position.
10.5.1 The Gradient of a Function

If \( \varphi \) represents the height above sea level, \( \frac{\partial \varphi}{\partial x_i} \) specifies the change in height given a change in position \( \text{d}x_i \). The factor \( \frac{\partial \varphi}{\partial x_i} \) is called the ‘gradient’ of \( \varphi \).

Suppose \( \varphi \) is represented in two different coordinate systems \( \varphi(x_i) \) and \( \varphi(y_i) \). If there is a one-to-one and onto function \( y_k \) such that \( y_k = y_k(x_1, x_2, \ldots, x_n), \ k = 1, \ldots, n \), continuous and differentiable to any degree necessary, then there is an allowable coordinate transformation between the \( x_i \)'s and \( y_i \)'s. To find the change in \( \varphi \) in terms of the \( x_i \)'s and \( y_i \)'s, divide the equation above by \( \partial y_k \) and replace the \( \text{d}'s \) with \( \partial' \)’s i.e.

\[
\frac{\partial \varphi}{\partial y_k} = \frac{\partial \varphi}{\partial x_i} \frac{\partial x_i}{\partial y_k}
\]

There are \( n \) of these equations, one for each spatial dimension. Mathematically, each equation is an instance of applying the chain rule associated with the differential calculus.

10.5.2 Tensors Revisited

Since a geometry can be portrayed in many different coordinate systems. Physical laws, written independently of the coordinate system that describes the geometry of the space, are said to be ‘Laws of Nature’. This suggests using tensor equations (see Book II: Chapter 13).

10.5.2.1 Scalars

A tensor of rank zero is a scalar. If \( x_i \) and \( y_i \) represent the same point in space, but are specified in a different coordinate system, then \( \varphi(x_i) = \varphi(y_i) \) at every point in space. Scalars do not transform. For instance, if the temperature at \( x_i \) is \( 10^\circ \text{C} \), then the temperature at \( y_i \) is \( 10^\circ \text{C} \), provided \( x_i \) and \( y_i \) represent the same point.

10.5.2.2 Vectors

A vector is a tensor of rank one. The simplest vector is the small displacement \( \text{d}x^i \). Expressing this vector in another set of coordinates uses the equation

\[
d\varphi = \frac{\partial \varphi}{\partial x_i} \text{d}x_i \rightarrow d\varphi = \frac{\partial \varphi}{\partial x^i} \text{d}x^i,
\]

where \( d\varphi \) has replaced \( \text{d}x^i \). In general,

\[
V^j = \frac{\partial y^j}{\partial x^i} V^i,
\]

where \( V^j \) has replaced \( d\varphi \) and \( V^i \) has replaced \( \text{d}x^i \). The vector \( V^i \) denotes the vector in one frame of reference and \( V^i \) denotes the same vector in a second frame of reference. The equation above is the transformation of an arbitrary contravariant vector from one coordinate system to another.
10.5.2.3 Higher Ranked Tensors

Higher rank tensors can be constructed simply by repeated applications of the same rule. If $V^j$ is a contravariant vector in one frame of reference and $W^k$ is a contravariant vector in the same frame of reference, then

$$V^j = \frac{\partial y^j}{\partial x^l} V^l, \quad W^k = \frac{\partial y^k}{\partial x^l} W^l \rightarrow V^j W^k = \frac{\partial y^j}{\partial x^l} \frac{\partial y^k}{\partial x^l} V^l W^l \rightarrow T^{jk} = \frac{\partial y^j}{\partial x^l} \frac{\partial y^k}{\partial x^l} T^{il}$$

where $T^{jk} = V^j W^k$ and $T^{il} = V^l W^l$ are tensors of ‘rank two’. A tensor of rank two can be thought of as two vectors acting at a point, not necessarily in the same direction. Sometimes this is called a ‘stress’. Higher ranking tensors than two can be created similarly.

A covariant vector $'A_j'$ transforms like

$$A_j = \frac{\partial x^i}{\partial y_j} A_i, \quad B_k = \frac{\partial x^i}{\partial y_k} B_l \rightarrow A_j B_k = \frac{\partial x^i}{\partial y_j} \frac{\partial x^l}{\partial y_k} A_i B_l \rightarrow T_{jk} = \frac{\partial x^i}{\partial y_j} \frac{\partial x^l}{\partial y_k} T_{il}, \quad T_{jk} = A_j B_k,$$

$$T_{il} = A_i B_l$$

Note that the transformation of a covariant vector comes from the equation

$$\frac{\partial \varphi}{\partial y_k} = \frac{\partial \varphi}{\partial x_i} \frac{\partial x_i}{\partial y_k},$$

since if $A_k$ is substituted for $\varphi/\partial y_k$ and $A_i$ is substituted for $\varphi/\partial x_i$, then the equation above becomes

$$A_k = \frac{\partial x^i}{\partial y_k} A_i,$$

which is the transformation of covariant vector.

10.5.2.4 The Advantages of Using Tensors

To see the advantage of using tensors, suppose a force is acting on block that lies on a frictionless plane (see fig. 10.5.2.4-1).
The work 'W' done in the x-direction is given by \( W = \int \langle F \rangle dx \), where \( \langle F \rangle dx = F \cos \theta \, dx \). If \( F \) acts directly downward, then
\[
\theta = \frac{\pi}{2} \rightarrow \cos \left( \frac{\pi}{2} \right) = 0 \rightarrow W = 0
\]

Hence, the block will not move. Moreover, if \( W = 0 \) in one frame of reference, it equals zero in all frames of reference. This fact plays an exceptionally important role in the use of tensors.

To sum up, covariant vectors represent gradients and contravariant vectors represent coordinate displacements. There can be mixed tensors i.e.
\[
T^i_j = \frac{\partial y^i}{\partial x^s} \frac{\partial x^s}{\partial y^j} T^r_s \rightarrow T^i_j = \frac{\partial y^i}{\partial x^s} \frac{\partial x^s}{\partial y^j} T^r_s = \delta^s_r T^s_r = T^r_r,
\]
since \( \frac{\partial x^r}{\partial x^s} = 1 \), if \( r = s \), \( \frac{\partial x^r}{\partial x^s} = 0 \), if \( r \neq s \). Note that \( T^i_j \) is a scalar. It is invariant, the same in all coordinate systems.

10.6 The Metric Tensor

Suppose there is a displacement 'ds' in an ordinary 2-dimensional Euclidean flat space. Intrinsically, \( ds^2 = (dx^1)^2 + (dx^2)^2 \). Written more compactly,
\[
ds^2 = \delta_{ij} dx^i dx^j, \quad \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases} \quad i, j = 1,2
\]

Note that \( ds^2 \) can be expanded to n-dimensional spaces simply by letting \( i, j = 1, ..., n \).

The quantity 'ds^2' can be represented in another coordinate system by the transformation
\[
dx^i = \frac{\partial x^i}{\partial y^r} dy^r, \\
\]
\[
dx^j = \frac{\partial x^j}{\partial y^s} dy^s \rightarrow dx^i dx^j = ds^2 = \delta_{ij} \frac{\partial x^i}{\partial y^s} \frac{\partial x^j}{\partial y^s} dy^r dy^s \rightarrow ds^2 = g_{rs} dy^r dy^s,
\]

where \( g_{rs} \) is called the 'metric tensor'. Note that if \( \delta_{ij} \) is replaced by \( g_{ij} \), then
\[
ds^2 = g_{ij} \frac{\partial x^i}{\partial y^r} \frac{\partial x^j}{\partial y^s} dy^r dy^s,
\]
which is just the form for the transformation of a covariant tensor. By the quotient rule for tensors, \( g_{ij} \) is a tensor. More importantly, if there exists a coordinate transformation such that
\[ \delta_{rs} = g_{ij} \frac{\partial x^i}{\partial y^r} \frac{\partial x^j}{\partial y^s} \]

then the geometry of the space is flat, otherwise the geometry of the space has ‘curvature’.

Example: Let \( ds^2 = \delta_{ij} dx^i dx^j = (dx^1)^2 + (dx^2)^2 \), \( i, j = 1, 2 \). Suppose \( ds^2 \) is written in polar coordinates \( (r, \theta) \), then \( x^1 = r \cos \theta \) and \( x^2 = r \sin \theta \). So,

\[
\begin{align*}
    dx^1 &= \cos \theta \, dr - r \sin \theta \, d\theta, \\
    dx^2 &= \sin \theta \, dr + r \cos \theta \, d\theta \\
    ds^2 &= (dx^1)^2 + (dx^2)^2 = dr^2 + r^2 d\theta^2
\end{align*}
\]

Hence,

\[
g_{rr} = 1, \quad g_{r\theta} = g_{\theta r} = 0, \quad g_{\theta\theta} = r^2
\]

Since the transformation sends flat coordinates into curvilinear coordinates and vice versa, the geometry of the space is flat.

10.6.1 The Conjugate of the Metric Tensor

The metric tensor \( g_{jk} \), having \( n^2 \) elements, can be represented by an \( n \times n \) matrix. If \( g = |g_{jk}| \), the determinate, and if \( g \neq 0 \), define

\[
g^j_k = \frac{\text{cofactor of } g_{jk}}{g} \]

Let the cofactor of \( g_{jk} \) be denoted \( 'G(j, k)' \). If \( n = 3 \), by the theory of determinates,

\[
G(2,1) = (-1)^{2+1} \begin{vmatrix} g_{12} & g_{13} \\ g_{32} & g_{33} \end{vmatrix}, \quad G(2,2) = (-1)^{2+2} \begin{vmatrix} g_{11} & g_{13} \\ g_{31} & g_{33} \end{vmatrix}, \quad G(2,3) = (-1)^{2+3} \begin{vmatrix} g_{11} & g_{12} \\ g_{31} & g_{32} \end{vmatrix} \]

Hence,

\[
g_{21}G(2,1) + g_{22}G(2,2) + g_{23}G(2,3) = g
\]

The same idea can be extended to \( n \) coordinates. In general, \( g_{jk}G(p,k) = 0 \) if \( j \neq p \). Since

\[
g^j_k = \frac{\text{cofactor of } g_{jk}}{g} = \frac{G(j, k)}{g}
\]

then

\[
g_{jk}g^j_k = g_{jk} \frac{\text{cofactor of } g_{jk}}{g} = g_{jk} \frac{G(j, k)}{g} = g \frac{g}{g} = 1
\]
In this case, the summation is over $k$ only. Note that $g_{jk}g^{pk} = 0$, if $p \neq j$. Hence, $g_{jk}g^{pk} = \delta^p_j$. Since $\delta^p_j$ is a tensor, $g^{jk}$ is a tensor by the quotient rule. Moreover, $g^{jk}$ is a symmetric tensor, since $g_{jk}$ is; the tensor ‘$g^{jk}$’ is called the ‘conjugate’ of $g_{jk}$.

Moreover, a contravariant component of a tensor can be changed into a covariant component or vice versa by taking the inner product of the given tensor with the metric tensor or its conjugate. To see this,

$$A_j = g_{jk} \to A^k g^{jq} A_j = g^{jq} g_{jk} A^k = \delta^q_k A^k = A^q$$

10.7 Space-Time Revisited

In ordinary Cartesian coordinates, more than one coordinate system can represent the same space. For instance, the point ‘$P$’ could be represented in an $X,Y$-system or an $X',Y'$-system. A transformation of Cartesian coordinates simply represents a rotation and/or a translation of the coordinates (see fig. 10.7-1).

10.7.1 Space-Time and the Metric Tensor

Regardless of which Cartesian coordinates are selected, the metric ‘$ds$’ is given by

$$ds^2 = \delta_{ij} dx^i dx^j$$

In other words, the metric in the $X,Y$-system is $ds^2 = dx^2 + dy^2$ and metric in the $X',Y'$-system is $ds^2 = (dx')^2 + (dy')^2$. The Pythagorean form of the metric is preserved.

10.7.1.1 The Metric Tensor and Proper Time

In space-time, the square of the proper time element ‘$d\tau$’ is given by

$$d\tau^2 = dt^2 - \frac{dx^i dx^i}{c^2} = dt^2 - dx^i dx^i, \quad c = 1$$

---

9 Strictly speaking, this is not true. The quotient rule only applies if $\delta^p_j$ is an arbitrary tensor, which it is not. But, use of an arbitrary tensor will lead to the same result [205].
The proper time is invariant, all observers will agree on the proper time. Suppose there is a set of space-time coordinates \( \{x^\mu\} \) and a set \( \{y^\mu\} \), which represent the same geometrical space. Assuming that for every point \( x^\mu \), there is a \( y^\mu \), then there will be a coordinate transformation from the set of coordinates \( \{x^\mu\} \) to the set of coordinates \( \{y^\mu\} \). Note that

\[
d\tau^2 = g_{\mu\nu}(x^\mu)dx^\mu dx^\nu
\]

has the same form as \( ds^2 \). To see this, the Minkowski form of the metric tensor is

\[
g_{\mu\nu} = \eta_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \rightarrow d\tau^2 = \eta_{\mu\nu}dx^\mu dx^\nu
\]

Since \( dx^\mu = \partial x^\mu / \partial y^r dy^r \), then

\[
d\tau^2 = \eta_{\mu\nu} \frac{\partial x^\mu}{\partial y^r} \frac{\partial x^\nu}{\partial y^s} dy^r dy^s = g_{rs}dy^r dy^s, \quad g_{rs} = \eta_{\mu\nu} \frac{x^\mu}{\partial y^r} \frac{x^\nu}{\partial y^s}
\]

which is identical to the form for the metric in ordinary spatial coordinates.

### 10.8 The Derivative of a Tensor

Ordinarily, expressing a change in a function from point to point is accomplished by taking its ‘derivative’. But, in general, the ordinary derivative of a tensor is not a tensor. So taking the ordinary derivative of a tensor gives an inaccurate description of how a tensor changes as its position in space changes. To see this, suppose that \( \varphi(x^\mu) \) is a scalar function of position, then \( d\varphi = \partial \varphi / \partial x^\mu dx^\mu \). Hence,

\[
\frac{\partial \varphi}{\partial x^\mu} \frac{\partial \varphi}{\partial y^\mu} = 0 \rightarrow \varphi = C, \quad C = a \text{ constant}
\]

Therefore, \( \varphi \) is constant at every point in space, regardless of which coordinate system it is specified in. Given a constant vector \( V_\mu \), its magnitude and direction remains constant with a change of position. If \( V_\mu \) is specified in Cartesian coordinates, its components are the projections of \( V_\mu \) onto the coordinate axes and the components are also constant. Hence, \( \partial V_\mu / \partial x^\mu = 0 \). But if \( V_\mu \) is specified in curvilinear coordinates \( y^\mu \) and if \( V_\mu \) is projected onto the coordinate axes, the components of \( V_\mu \) will not be constant, since the coordinates \( y^\mu \) vary from place to place. Hence,

\[
\partial V_\mu / \partial y^\mu \neq 0 \rightarrow \partial V_\mu / \partial x^\mu \neq \partial V_\mu / \partial y^\mu
\]

Consequently, the ordinary derivative of a vector in one coordinate system does not necessarily equal the ordinary derivative of a vector in another coordinate system.
10.8.1 The Covariant Derivative of a Tensor

If ‘$T_{mn}$’ is a tensor such that $T_{mn}(x) = \partial V_m / \partial x^n$, where $V_m$ is a vector, then

$$ T_{mn}(y) = \frac{\partial x^r}{\partial y^m} \frac{\partial x^s}{\partial y^n} T_{rs}(x) \to T_{mn}(y) = \frac{\partial x^r}{\partial y^m} \frac{\partial x^s}{\partial x^t} V_r(x) $$

Is the following equation true?

$$ \frac{\partial x^r}{\partial y^m} \frac{\partial x^s}{\partial x^t} V_r(x) \equiv \frac{\partial V_m(y)}{\partial y^n} $$

To see, note that

$$ \frac{\partial x^r}{\partial y^m} \frac{\partial x^s}{\partial x^t} V_r(x) = \frac{\partial x^r}{\partial y^m} \frac{\partial V_r(x)}{\partial y^t} = \frac{\partial}{\partial y^n} \left( \frac{\partial x^r}{\partial y^m} V_r(x) \right) $$

And that

$$ \frac{\partial}{\partial y^n} \left( \frac{\partial x^r}{\partial y^m} V_r(x) \right) = \frac{\partial x^r}{\partial y^m} \frac{\partial V_r(x)}{\partial y^n} + \frac{\partial^2 x^r}{\partial y^n \partial y^m} V_r(x) $$

Hence,

$$ \frac{\partial V_m(y)}{\partial y^n} \neq \frac{\partial}{\partial y^n} \left( \frac{\partial x^r}{\partial y^m} V_r(x) \right) $$

The derivative of a vector has the extra term

$$ \frac{\partial^2 x^r}{\partial y^n \partial y^m} V_r(x) $$

because the coordinates are curvilinear, as opposed to rectilinear.

Let

$$ \frac{\partial^2 x^r}{\partial y^n \partial y^m} V_r(x) \equiv \Gamma^r_{mn} V_r $$

If some mathematical object can be found such that

$$ \frac{\partial V_m(x)}{\partial y^n} + \Gamma^r_{mn} V_r(x) = T_{mn}(y) = \nabla_n V_m(x), $$

where $T_{mn}$ is a tensor, then the factor ‘$\Gamma^r_{mn}$’ would account for coordinates changes with respect to a change in position and the above expression would represent the ‘derivative of a vector’ ($\nabla_n$).
The covariant derivative \( \nabla_p (T_{mn}) \) of \( T_{mn} \) by definition is

\[
\nabla_p (T_{mn}) = \frac{\partial T_{mn}}{\partial y^p} - \Gamma^r_{pm} T_{rn} - \Gamma^r_{pn} T_{rm}
\]

The equation above is obtained by repeated applications of the rule for obtaining the covariant derivative of a vector (see Book II, Sec. 13.11).

### 10.8.1.1 The Covariant Derivative of the Metric Tensor

In flat space, \( g_{ij} = \delta_{ij} \). And since all the components \( \delta_{ij} \) are constant, \( \Gamma = 0 \), and hence, \( \nabla_r (\delta_{ij}) = 0 \). But if all the components of a tensor are zero in one frame of reference, they must be zero in all frames of reference. Is \( \nabla_r (g_{ij}) \) a tensor? Note that

\[
\nabla_p (g_{mn}) = \frac{\partial g_{mn}}{\partial y^p} - \Gamma^r_{pm} g_{rn} - \Gamma^r_{pn} g_{rm} = 0
\]

To see this, recall that

\[
\frac{\partial g_{mn}}{\partial y^p} = [mp, n] + [np, m] = \frac{1}{2} \left( \frac{\partial g_{mn}}{\partial y^p} + \frac{\partial g_{pn}}{\partial y^m} - \frac{\partial g_{mp}}{\partial y^n} \right) + \frac{1}{2} \left( \frac{\partial g_{mn}}{\partial y^p} + \frac{\partial g_{pm}}{\partial y^n} - \frac{\partial g_{np}}{\partial y^m} \right)
\]

But,

\[
[mp, n] = \Gamma^r_{pm} g_{nr}, \quad [np, m] = \Gamma^r_{pn} g_{rm}
\]

Hence,

\[
\frac{\partial g_{mn}}{\partial y^p} = \Gamma^r_{pm} g_{nr} + \Gamma^r_{pn} g_{rm} \to \frac{\partial g_{mn}}{\partial y^p} - \Gamma^r_{pm} g_{nr} - \Gamma^r_{pn} g_{rm} = 0 = \nabla_p (g_{mn}),
\]

where

\[
\Gamma^r_{pm} = \frac{1}{2} g^{nr} \left( \frac{\partial g_{mn}}{\partial y^p} + \frac{\partial g_{pn}}{\partial y^m} - \frac{\partial g_{mp}}{\partial y^n} \right)
\]

While \( \Gamma \) is not a tensor, \( \nabla_p (g_{mn}) \) is a tensor. To see this, suppose that \( A_p \) is a vector, then

\[
\frac{\partial A_j}{\partial x^k} - \Gamma^r_{jk} A_n = \frac{\partial x^p}{\partial x^j} \frac{\partial x^q}{\partial x^k} \left( \frac{\partial A_p}{\partial x^q} - \Gamma^r_{pq} A_s \right)
\]

Hence,

\[
\frac{\partial A_p}{\partial x^q} - \Gamma^r_{pq} A_s = \nabla_q A_p
\]
is a tensor - the covariant derivative of the covariant vector \(A_p\). If \(A_p\) is a contravariant vector, then

\[
\frac{\partial A^p}{\partial x^q} + \Gamma^p_{qs} A^s = \nabla_q A^p,
\]

where \(\nabla_q A^p\) is the 'covariant derivative of \(A^p\)' (see Book II: Chapter 13, Sec. 13.11.2).

**10.8.2 The Significance of the Covariant Derivative**

If the space is flat, it is always possible to find a set of coordinates in which the Christoffel symbols \(\Gamma^r_{mn}'s\) vanish. In this case, the covariant derivative simply reduces to the ordinary derivative. However, if a flat space is described in curvilinear coordinates, in general, the Christoffel symbols will not vanish, but a coordinate transformation into a Euclidean metric will make the Christoffel symbols vanish. On the other hand, in a space with curvature, finding a coordinate transformation where the Christoffel symbols vanish is not possible. This is the power of the covariant derivative. It provides a way of distinguishing between a ‘flat space’ and a ‘curved space’. Ultimately it provides a way of distinguishing between a flat Earth gravitational field and one with tidal forces.

**10.8.3 The Covariant Derivative of the Tangent Vector**

Suppose there is a curve \(C\) in flat space immersed in a vector field (see fig. 10.8.3-1). Choose a parameter that varies along the curve such as the arc length \('s\'. How the function \('\varphi\') varies along the curve is computed using the chain rule:

\[
\frac{d\varphi}{ds} = \frac{\partial \varphi}{\partial y^m} \frac{dy^m}{ds},
\]

**Figure 10.8.3-1**

Note that \(dy^m/ds\) is the unit tangent vector along the curve i.e. \(|dy^m/ds| = 1\).

Now replace \(\varphi\) with a vector \('V^n'\) i.e.

\[
\frac{dV^n}{ds} = \frac{\partial V^n}{\partial y^m} \frac{dy^m}{ds}
\]
Note that \(\partial V^n/\partial y^m\) is the ordinary derivative of a vector, which is not, in general, a tensor. To make \(\partial V^n/\partial y^m\) a tensor, it must be replaced by the covariant derivative. Let

\[
\nabla_s V^n(y) = \nabla_m V^n(y) \frac{dy^m}{ds} \rightarrow \frac{\partial V^n}{\partial y^m} \frac{dy^m}{ds} + \Gamma^r_{mn} \frac{dy^m}{ds} = \left(\frac{dV^n}{ds} + \Gamma^r_{mr} V^r\right) \frac{dy^m}{ds}
\]

The equation above represents the covariant derivative of a vector along a curve, which is a tensor. The tangent vector along a curve is found by simply substituting the tangent vector for \(V^n\) in the equation above i.e. \(V^n \rightarrow dy^n/ds\). Hence,

\[
\nabla_s V^n(y) = \frac{dV^n}{ds} + \Gamma^r_{mr} \frac{dy^r}{ds} \rightarrow \nabla_s V^n(y) = \frac{d^2y^n}{ds^2} + \Gamma^r_{mr} \frac{dy^r}{ds} \frac{dy^m}{ds}
\]

The equation on the right above is the covariant derivative of the tangent vector and is a tensor. Because the covariant derivative is a tensor, if it is zero in one reference frame, it is zero in all reference frames. Moreover, the covariant derivative of a straight line is zero everywhere along the line. Hence, if the points along a curve are given in terms of some set of curvilinear coordinates, and if the covariant derivative of the tangent vector is zero along the curve, then the curve is a straight line.

10.9 Curvature

Intuitively, curvature is the geometric characteristic which imposes an obstruction to finding a coordinate transformation into a flat metric. Determining whether a space is ‘curved’ or ‘flat’ requires the concept of the ‘parallel transport’ of a vector along a curve in a geometry.

10.9.1 The Parallel Transport of a Vector

If, in flat 2-space, a vector \(V^n\) is transported along a curve from point \(q\) to \(p\), then \(V^n\) is ‘parallel transported’ if its length remains constant and if the lines drawn perpendicular to the vector at every point are all parallel (see fig. 10.9.1-1).
Normally, in the intuitive description of parallel transport, the reference to perpendicular lines is omitted, since it is only true in 2-space. In 3-space, the definition is more complicated, but, for learning purposes, the reference to parallel lines is included.

Mathematically, the parallel transport of an arbitrary vector along a curve is defined as

\[ \nabla_s V^n(y) = \frac{dV^n}{ds} + \Gamma^m_{mn} V^r \frac{dy^m}{ds} = 0 \rightarrow dV^n = -\Gamma^m_{mn} V^r dy^m \]

In other words, if the covariant derivative of a vector is equal to zero all along the curve, then the vector is ‘parallel transported’ along the curve. If \( V^n \) is the tangent vector, then the curve is called a ‘geodesic’.

**10.9.1.1 Parallel Transport in Curved vs. Flat Spaces**

If an arbitrary vector is parallel transported along a closed curve from a point ‘\( P \)’, and upon returning to \( P \) is the same vector, then the space is flat. If a vector is parallel transported along a closed curve from a point ‘\( P \)’ and if, upon returning to \( P \), is not the same vector, then the space is geometrically curved. A geodesic is a curve whose curvature is entirely due to the geometry of the space. If the curve has curvature of its own, independent of the curvature of the space, it is not a geodesic.

**Figure 10.9.1.1-1**

Example: Consider a cone (fig. 10.9.1.1-1 B), where a cut is made along the line between the points ‘\( A \)’ and ‘\( B \)’. The cone can then be flattened (fig. 10.9.1.1-1 A). Let a vector at point ‘\( C \)’ be parallel transported along the curve, ending at point ‘\( C' \)’. Note that the points ‘\( C \)’ and ‘\( C' \)’ are the same point on the cone. If the cone is closed up so that the points ‘\( C \)’ and ‘\( C' \)’ coincide, the two vectors will not be the same. Both vectors will have the same magnitude, but point in different directions. Hence, a cone is a geometric space with curvature. The amount of curvature is proportional to the angle ‘\( \theta \)’ between the initial vector and the one that has been parallel transported along the curve. There is only one point on the cone with curvature, the tip ‘\( O \)’. If any other point on the surface of a cone is circumvented by parallel transportation, the vector will not change. That part of the surface of the cone is spatially flat.
10.10 The Curvature Tensor

In a curved space, a small angle deficit ‘$\delta \theta$’ is proportional to the area bounded by closed curve i.e. $\delta \theta = R \delta A$, where $R$ is proportional to the ‘curvature’ and $\delta A$ is the area bounded by the curve. Let $\theta + \phi = 2\pi$. If $\theta > 0$, then $\phi < 2\pi$ and the space has ‘positive’ curvature. If $\theta = 0$, then $\phi = 2\pi$ and the space is flat (zero curvature). If $\theta < 0$, then $\phi > 2\pi$ and the space has ‘negative’ curvature. Intuitively, positive, zero and negative curvature can be visualized as shown in fig. 10.10-1.

![Figure 10.10-1](image)

Given a point on a surface, where a closed curve circumvents that point, its boundary enclosing a very small area of the surface, and if an arbitrary vector is parallel transported along the curve, then $d\theta/dA = R$, where $d\theta$ is the change in the angle the vector makes with itself upon its return trip, $dA$ is the area enclosed by the curve and $R$ is proportional to the curvature around the point. If $d\theta/dA = R = 0$, there is no curvature, but, $d\theta/dA = R \neq 0$ indicates curvature around the point.

10.10.1 Measuring Curvature

To measure curvature, two pieces of information are required: 1) the orientation of the plane of the curve, 2) the change in angle the vector makes with itself as it circumvents the curve.

![Figure 10.10.1-1](image)

Let the coordinates ‘$x^\mu$’ and ‘$x^\nu$’ define the plane in which a point on a surface is being circumvented, then choose a small rectangle in the plane to represent the closed
surface (see fig. 10.10.1-1). Note that the area ‘dA ∝ dxμdxν’. The area will not exactly equal dxμdxν because the coordinate axes are not necessarily perpendicular.

If an arbitrary vector 'Vμ' is parallel transported counterclockwise around the rectangle beginning at O and returning to O and if there is curvature, there will be a change in the vector proportional to the area enclosed by the rectangle i.e.

\[ dV^\mu \propto dx^\sigma dx^\nu V^\tau \]

To change ‘∝’ to ‘=’ requires including a factor of the form ‘R^\mu_{\sigma\nu\tau}' i.e.

\[ dV^\mu = R^\mu_{\sigma\nu\tau}dx^\sigma dx^\nu V^\tau \]

Recalling the definition of parallel transport i.e.

\[ dV^m = -\Gamma^m_{\nu r}V^\nu dx^r, \]

dV^m will depend on the Γ^m_{\nu r}’s, which vary from point to point along the curve.

### 10.10.1.1 The Riemann-Christoffel Tensor

This suggests that the change in the vector depends on the derivatives of the Christoffel symbols. One choice of a tensor that depends on the derivatives of the Christoffel symbols is the ‘Riemann-Christoffel’ tensor. Its derivation can be found in Book II: Chapter 13, Sec. 13.12. Note that within R^\mu_{\sigma\nu\tau}, the two indices ‘\sigma, \nu’ are associated with the plane of circumvention and the indices ‘\mu, \tau’ are associated with how much the vector changes angle as it is parallel transported around the curve.

Recall that the commutator ‘[A, B]’ is defined as [A, B] = AB − BA. If A, B are simply numbers, then [A, B] = 0. But consider

\[ \left[ \frac{\partial}{\partial x}, f(x) \right] = \frac{\partial f}{\partial x} - f \frac{\partial}{\partial x} \frac{\partial f}{\partial x} V - f \frac{\partial V}{\partial x}, \]

The V comes in because ∂/∂x must act on something, in this case, V. Note that the terms of the expression on the right above are the differentiation of a product. Hence,

\[ \frac{\partial}{\partial x}(fV) - f \frac{\partial V}{\partial x} = \frac{\partial f}{\partial x} V + f \frac{\partial V}{\partial x} - f \frac{\partial V}{\partial x} V \rightarrow \left[ \frac{\partial}{\partial x}, f(x) \right] = \frac{\partial f}{\partial x} \]

\[ \begin{array}{c}
\text{Figure 10.10.1.1-1}
\end{array} \]
Suppose an arbitrary vector is parallel transported around the loop 'OBCDO' (see fig. 10.10.1-1). Note that O, O' are the same point, labeled differently, since the vector, when it returns to O' from O, may not be the same vector.

Consider

\[(V_C - V_D) - (V_B - V_O)\]
\[(V_C - V_B) - (V_D - V_{O'})\],

The quantity '(V_C - V_D)' represents the difference in the vector 'V' at positions 'C' and 'D'. Hence, (V_C - V_D) - (V_B - V_O) represents the difference in the vector 'V' along the \(dx^\nu\)-direction. Likewise, (V_C - V_B) - (V_D - V_{O'}) represents the difference of the vector 'V' along the \(dx^\mu\)-direction. Note that

\[(V_C - V_D) - (V_B - V_O) - [(V_C - V_B) - (V_D - V_{O'})] = V_O - V_{O'} = dV,\]

where \(dV\) is the change in the vector 'V' as it is parallel transported around the loop 'OBCDO'.

Now

\[V_C - V_D = \frac{\partial V}{\partial x^\mu} dx^\mu V \rightarrow V_C - V_D = \nabla_\mu dx^\mu V,\]

where the ordinary derivative \(\partial V/\partial x^\mu\) has been replaced by the covariant derivative \(\nabla_\mu\). Moreover,

\[(V_C - V_D) - (V_B - V_O) = \nabla_\nu dx^\nu \nabla_\mu dx^\mu V = \nabla_\nu \nabla_\mu dx^\mu dx^\nu V,\]

since \((V_C - V_D)\) and \((V_B - V_O)\) are separated by \(dx^\nu\). Similarly,

\[(V_C - V_B) - (V_D - V_{O'}) = \nabla_\mu \nabla_\nu dx^\mu dx^\nu V\]

Hence,

\[(V_C - V_D) - (V_B - V_O) - [(V_C - V_B) - (V_D - V_{O'})] = V_O - V_{O'}\]

\[= \nabla_\nu \nabla_\mu dx^\mu dx^\nu V - \nabla_\mu \nabla_\nu dx^\mu dx^\nu V = (\nabla_\nu \nabla_\mu - \nabla_\mu \nabla_\nu)dx^\mu dx^\nu V = dV\]

Note that \(\nabla_\nu \nabla_\mu - \nabla_\mu \nabla_\nu = [\nabla_\nu, \nabla_\mu]\) and if \([\nabla_\nu, \nabla_\mu] = 0\), \(dV = 0\) and the space is flat. If \([\nabla_\nu, \nabla_\mu] \neq 0\), the space is curved.

Recall that the covariant derivative takes the form

\[\nabla_\nu = \frac{\partial}{\partial x^\nu} + \Gamma_{\nu r}^m\]
Hence,
\[ [\nabla_v, \nabla_\mu] = \left( \frac{\partial}{\partial x^v} + \Gamma_v \right) \left( \frac{\partial}{\partial x^\mu} + \Gamma_\mu \right) - \left( \frac{\partial}{\partial x^\mu} + \Gamma_\mu \right) \left( \frac{\partial}{\partial x^v} + \Gamma_v \right), \]
where, for brevity, \( \Gamma_v \) represents the Christoffel symbol corresponding to the \( v \)-direction and so on. Therefore,
\[ [\nabla_v, \nabla_\mu] = \left( \frac{\partial}{\partial x^v} + \Gamma_v \right) \left( \frac{\partial}{\partial x^\mu} + \Gamma_\mu \right) - \left( \frac{\partial}{\partial x^\mu} + \Gamma_\mu \right) \left( \frac{\partial}{\partial x^v} + \Gamma_v \right) = - \left[ \frac{\partial}{\partial x^\mu}, \Gamma_v \right] + \left[ \frac{\partial}{\partial x^v}, \Gamma_\mu \right] + \left[ \Gamma_v, \Gamma_\mu \right] \]
Recall that
\[ \left[ \frac{\partial}{\partial x^\alpha}, f(x) \right] = \frac{\partial f}{\partial x}, \quad \left[ \frac{\partial}{\partial x^\mu}, \Gamma_v \right] = \frac{\partial \Gamma_v}{\partial x^\mu}, \quad \left[ \frac{\partial}{\partial x^v}, \Gamma_\mu \right] = \frac{\partial \Gamma_\mu}{\partial x^v} \]
Therefore, \([\nabla_v, \nabla_\mu] \) is made up of Christoffel symbols and their derivatives. Note that \( [\nabla_v, \nabla_\mu] = R^\alpha_{\mu\nu\beta} \) is a tensor, since the outer multiplication of two tensors is a tensor. Subtracting two tensors of the same rank results in a tensor i.e.
\[ dV^\alpha = \left[ \frac{\partial \Gamma^\alpha_{\mu\beta}}{\partial x^\nu} - \frac{\partial \Gamma^\alpha_{\nu\beta}}{\partial x^\mu} + \Gamma^\alpha_\nu \Gamma^\delta_\mu - \Gamma^\alpha_\mu \Gamma^\delta_\nu \right] d\mu d\nu V^\beta = R^\alpha_{\mu\nu\beta} d\mu d\nu V^\beta \]

**10.10.1.2 The Ricci Tensor**

In particular, \( R^\alpha_{\mu\nu\beta} \) can be contracted in \( \alpha, \nu \), i.e. \( R^\alpha_{\mu\alpha\beta} \), which gives a new tensor \( R^\mu_{\mu\beta} \), the 'Ricci tensor', named for the Italian mathematician Gregorio Ricci-Curbastro (12 January 1853 – 6 August 1925) i.e.
\[ dV = R^\mu_{\mu\beta} d\mu V^\beta \]
While \( R^\alpha_{\mu\nu\beta} \) is not a symmetric tensor, \( R_{\mu\beta} \) is symmetric. Furthermore, \( R_{\mu\beta} \) can be contracted i.e. \( g^{\mu\beta} R_{\mu\beta} = R \), where \( R \) is called the 'curvature scalar'. If \( R \neq 0 \), then the space is curved. Interestingly, if \( R = 0 \), this does not necessarily indicate that the space is flat. The condition \( 'R = 0' \) is a necessary, but not sufficient a condition for a flat space. If \( R = 0 \) and the dimension of the space is either two or three, then the space is flat. But this is not true in 4-dimensional space.

**10.11 The Energy-Momentum Tensor**

In a small volume of space, \( dQ/dV = \rho_Q \), where \( Q \) represents charge, \( V \) represents spatial volume and \( \rho_Q \) is the 'charge density'. The equivalent notion to 'charge density' in space-time is the amount of charge in a spatial volume 'dV' in an instant of time.
Moreover, the amount of charge flowing through a spatial area ‘A’ in time ‘t’ is the ‘current density’ i.e. \( dQ / (dA \cdot dt) = J^i \). The \( A_i \)’s have orientations, one for each spatial direction. Hence, there are three current densities \( J^1, J^2, J^3 \). Recall from special relativity that \( (J^0, J^1, J^2, J^3) = J^\mu \), \( J^0 = \rho \) is a 4-vector.

If charge is flowing out of a volume, a net current flows out of the volume. If charge is flowing onto a volume, a net current flows into the volume. Mathematically, this is expressed by the equation ‘\( \partial \rho / \partial t + \langle \nabla | J \rangle = 0 \)’, which simply states that charge and current density are locally conserved. Not that

\[
\frac{\partial \rho}{\partial t} + \frac{\partial J^0}{\partial x^0} + \frac{\partial J^1}{\partial x^1} + \frac{\partial J^2}{\partial x^2} + \frac{\partial J^3}{\partial x^3} = \frac{\partial J^\mu}{\partial x^\mu} = 0
\]

If \( \langle p^0, p^1, p^2, p^3 \rangle = p^\mu \), then ‘\( p^0 = E \)’ is the time component and \( p^i, i = 1,2,3 \)’ are the spatial components. If energy is flowing out of a volume, a net momentum flows out of the volume. If energy flows onto a volume, a net momentum flows into the volume. Hence, there must notions of ‘energy and momentum density’ and ‘flow of energy and momentum’. Energy density could be signified by \( T_{\text{energy,density}} = T^{00} \), \( T^{00} \) signifies energy and \( T^{00} \) signifies ‘energy-density’. Similarly, \( T_{\text{energy,flow}} = T^{01} \) is the flow of energy in the \( x^1 \)-direction. Moreover, \( T_{\text{momentum,density}} = T^{10}, T^{10} \) the momentum density in the \( x^1 \)-direction. Likewise, the flow of momentum in the \( x^1 \)-direction would be given by \( T^{11} \). The quantity \( T^{12} \) represents the flow from the \( x \)-direction toward the \( y \)-direction.

By similar arguments, all the other \( T^{\nu\mu} \)'s can be constructed. Hence,

\[
T^{\mu\nu} = \begin{bmatrix}
T^{00} & T^{01} & T^{02} & T^{03} \\
T^{10} & T^{11} & T^{12} & T^{13} \\
T^{20} & T^{21} & T^{22} & T^{23} \\
T^{30} & T^{31} & T^{32} & T^{33}
\end{bmatrix}
\]

where \( T^{\mu\nu} \) is called the ‘momentum-energy’ tensor.

10.11.1 Conservation of Momentum and Energy

There are four local conservation equations, one for each row of \( T^{\mu\nu} \) i.e.

\[
\frac{\partial T^{00}}{\partial t} + \frac{\partial T^{01}}{\partial x^1} + \frac{\partial T^{02}}{\partial x^2} + \frac{\partial T^{03}}{\partial x^3} = \frac{\partial T^{0\mu}}{\partial x^\mu} = 0
\]

Similarly,

\[
\frac{\partial T^{1\mu}}{\partial x^\mu} = 0, \quad \frac{\partial T^{2\mu}}{\partial x^\mu} = 0, \quad \frac{\partial T^{3\mu}}{\partial x^\mu} = 0 \rightarrow \frac{\partial T^{\nu\mu}}{\partial x^\mu} = 0,
\]

which is the local conservation of energy and momentum associated with special relativity. Note that \( \partial T^{\nu\mu} / \partial x^\mu \) is the ordinary derivative of \( T^{\nu\mu} \), and hence, \( \partial T^{\nu\mu} / \partial x^\mu \) is not, in general, a tensor. In order to jump from special relativity to general relativity, \( \partial T^{\nu\mu} / \partial x^\mu \) must be replaced by the covariant derivative.
10.12 Einstein’s Gravitational Field Equations

The Einstein field equations were basically determined by trial and error. The final equations should be of the form

\[ g_{\mu\nu} \rightarrow KT_{\mu\nu}, \quad K = \text{a constant} \]

In other words, the equations should be some functions of the \( g_{\mu\nu} \)'s, which define the curvature of space-time, mapped into the \( T_{\mu\nu} \)'s, which define how particles move. Preferably, the equations should be tensor equations, which are invariant under coordinate transformations.

One of the postulates of general relativity is that particles (bodies) move, in a gravitational field, along geodesics. The equations of a geodesic are given by

\[ \frac{d^2 x^\mu}{d\tau^2} + \Gamma^\mu_{\nu r} \frac{dx^\nu}{d\tau} \frac{dx^r}{d\tau} = 0, \]

where \( \tau \) signifies the proper time i.e. \( d\tau^2 = g_{ij}dx^i dx^j \). The proper time is invariant, all observers will agree on the proper time. So, a geodesic in space-time is just a world line. Specifically, the equations of a geodesic are the parallel transporting of the tangent vector \( dx^\mu/d\tau \) along a world line. And if the covariant derivatives of the tangent vector all along world line are equal to zero, the world line is a geodesic.

10.12.1 A First Approximation to Einstein’s Gravitational Field Equations

As a first approximation to finding the proper gravitational equations, suppose that all particles move very slowly compared to the speed of light and that the gravitational forces are weak (no black holes). Under these conditions, a few simplifying assumptions can be made. For one \( g_{\mu\nu} \approx \eta_{\mu\nu} \), the Minkowski metric, plus some small corrections. So, \( \tau \approx t \rightarrow d\tau/dt \approx 1 \). To further simplify, suppose only one spatial direction is chosen, say ‘\( x \)’, then \( d^2 x^i/d\tau^2 \rightarrow d^2 x/dt^2 \), \( i = 1,2,3 \), which is just the ordinary derivative in the \( x \)-direction. Note that \( dx/d\tau \ll 1 \), since, by assumption, everything is moving slowly. The only large term is \( dx^0/d\tau \approx 1 \). Hence, the equation for a geodesic becomes

\[ \frac{d^2 x}{dt^2} + \Gamma^x_{00} \left( \frac{dx^0}{dt} \right)^2 = \frac{d^2 x}{dt^2} + \Gamma^x_{00} = 0 \rightarrow \frac{d^2 x}{dt^2} = -\Gamma^x_{00}, \quad \Gamma^x_{00} = \frac{1}{2} g^{xx} \left( \frac{\partial g_{00}}{\partial x^0} + \frac{\partial g_{0x}}{\partial x^0} - \frac{\partial g_{0x}}{\partial x} \right) \]

Since \( g^{xx} \approx \eta^{xx} = -1 \) and, since everything is moving slowly, \( \partial g_{00}/\partial x^0 = \partial g_{0x}/\partial x^0 \approx 0 \). Therefore,

\[ \Gamma^x_{00} = \frac{1}{2} \frac{\partial g_{00}}{\partial x} \rightarrow \frac{d^2 x}{dt^2} = -\Gamma^x_{00} = -\frac{1}{2} \frac{\partial g_{00}}{\partial x}, \]

which is a statement of Newton’s force law i.e. \( F = ma \rightarrow a = F/m \). This would suggest that \( \Gamma^x_{00} \) is associated with the acceleration or force. In the field formulation of Newton’s
force law, \( F = - \frac{\partial \phi}{\partial x} \), where \( \phi \) is the gravitational potential energy i.e. \( \phi = mgx \rightarrow \frac{\partial \phi}{\partial x} = mg \). If \( \Gamma_{00} \equiv F \), then

\[
\frac{1}{2} \frac{\partial g_{00}}{\partial x} = - \frac{\partial \phi}{\partial x} \rightarrow \phi = - \frac{1}{2} g_{00} + K, \quad K = a \text{ constant}
\]

Recall that \( F = - \frac{\partial \phi}{\partial x} = -\langle \nabla | \phi \rangle \), which is simply the divergence of the gravitational potential.

Recall that Newton’s gravitational force law \( (F = - GMm/r^2) \) and suppose there is a body of mass ‘\( m = 1 \)’ a distance ‘\( r \)’ from a larger body with mass ‘\( M \)’ (see fig. 10.12.1-1).

Hence, \( F = - GM/r^2 \). The flux moving through a sphere of radius ‘\( r \)’ is given by \( \int \langle F|dA \rangle \), where \( dA \) is an element of the surface area. But,

\[
\int \langle F|dA \rangle = - \frac{GM}{r^2} 4\pi r^2 = -4\pi GM
\]

By the divergence theorem

\[
\iiint_V \langle \nabla |F \rangle \, dV = \iiint_S \langle F|dS \rangle = \int \langle F|dA \rangle
\]

Since \( dM/dV = \rho \rightarrow M = \int \rho \, dV \), then

\[
\iiint_V \langle \nabla |F \rangle \, dV = -4\pi G \iiint_V \rho \, dV
\]

Differentiating both sides of the equation above by \( dV \) gives

\[
\langle \nabla |F \rangle = \langle \nabla |\nabla |\phi \rangle \rangle = \langle \nabla^2 |\phi \rangle = -4\pi G \rho
\]

Since \( \phi = - \frac{1}{2} g_{00} + K \), then

\[
-\langle \nabla^2 |\frac{1}{2} g_{00} \rangle = -4\pi G \rho \rightarrow \langle \nabla^2 |g_{00} \rangle = -8\pi G \rho
\]

245
Letting $T_{00} = \rho$, the equation above becomes

$$\langle \nabla^2 | g_{00} \rangle = -8\pi G T_{00},$$

which is the required form relating the curvature of space-time to the energy and momentum of the system. The equation above is not a tensor equation, since it was derived in a frame of reference where everything moved slowly and gravitational forces are weak. If this law is to hold in general, a tensor equation is required.

### 10.12.2 Einstein’s Search for a Tensor Equation

Einstein’s needed to find a tensor ‘$G_{\mu\nu}$’ such that

$$G_{\mu\nu} = -8\pi G T_{\mu\nu}$$

It had to be of rank ‘2’ and contain the derivatives of $g_{\mu\nu}$’s, the metric tensor. The most obvious candidate was the Ricci tensor ‘$R_{\mu\nu}$’ and this is what Einstein tried first i.e.

$$R_{\mu\nu} = -8\pi G T_{\mu\nu}$$

He immediately ran into difficulties. Recall that $T_{\mu\nu}$, the energy-momentum tensor, is locally conserved. Therefore, it must follow that

$$\nabla_\mu T_{\mu\nu} = 0$$

But, in general, $\nabla_\mu R_{\mu\nu} \neq 0$. The equation ‘$R_{\mu\nu} = -8\pi G T_{\mu\nu}$’ simply will not work. What Einstein was looking for was something along the lines of what Maxwell achieved. Since charge conservation was a consequence of Maxwell’s equations, Einstein wanted energy and momentum conservation to be a consequence of the curvature of space-time. Interestingly, he did not achieve this.

However, working out the calculation for $\nabla_\mu R_{\mu\nu}$ gives

$$\nabla_\mu R_{\mu\nu} = \frac{1}{2} \nabla_\mu (g_{\mu\nu} R) = \frac{1}{2} \nabla_\mu g_{\mu\nu} R + \frac{1}{2} g_{\mu\nu} \nabla_\mu R = \frac{1}{2} g_{\mu\nu} \nabla_\mu R, \quad \nabla_\mu g_{\mu\nu} = 0$$

where $R$ is the ‘curvature scalar’ and the covariant derivative of a scalar is just the ordinary derivative. Clearly, $\nabla_\mu R_{\mu\nu} \neq 0$.

Hence,

$$\nabla_\mu \left[ R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R \right] = 0 \rightarrow \nabla_\mu \left[ R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R \right] = -8\pi G \nabla_\mu T_{\mu\nu} \rightarrow G_{\mu\nu} = -8\pi G T_{\mu\nu},$$

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R$$

$G_{\mu\nu} = -8\pi G T_{\mu\nu}$ are Einstein’s equations for the gravitational field.
10.12.3 The Non-Conservation of Einstein’s Field Equations

Note that

\[ T_{\mu\nu} = 0 \rightarrow R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 0 \]

And consider

\[ g^{\mu\nu} \left( R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R \right) = R - \frac{1}{2} g^{\mu\nu} g_{\mu\nu} R = R - \frac{1}{2} \delta^\mu_\mu R = 2R - R = 0 \rightarrow R = 0 \]

If \( R = 0 \), then \( R_{\mu\nu} = 0 \). In regions of space-time where there is no energy or momentum, the Ricci tensor is equal to zero. If the Ricci tensor is zero in 2-space or 3-space, then the space is flat. But, in 4-dimensional space-time, \( R_{\mu\nu} = 0 \) is not enough to guarantee that the space-time geometry is flat. There are cases where the Riemann tensor is not equal to zero, but the Ricci tensor is equal to zero. The geometric conditions, where \( R_{\mu\nu}^{\alpha\beta} \neq 0 \), but \( R_{\mu\nu} = 0 \) are the existence of gravitational waves.

Evidently, these waves propagate in empty space-time and carry momentum and energy. Hence, momentum and energy are not conserved as in electromagnetic theory. Theoretically, no explanation is given as to where gravitational waves come from. They are simply a consequence of Einstein’s equations.

The non-conservation of gravitational waves creates a problem when trying to unify general relativity with quantum physics, where momentum and energy are conserved, at least within the constraints of the uncertainty principle. General relativity is the only theory in physics where there exists an open-ended violation of the conservation of momentum and energy.

10.12.4 Gravitational Waves

Consider a simple scalar wave equation i.e.

\[ \frac{\partial^2 \varphi}{\partial t^2} = c^2 \left( \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} \right) \rightarrow \frac{\partial^2 \varphi}{\partial t^2} = \left( \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} \right), \quad \text{if } c = 1, \]

which represents a light wave propagating in some arbitrary direction in empty space. The equation above can be written in the following form:

\[ \frac{\partial^2 \varphi}{\partial t^2} = \left( \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} \right) \rightarrow \frac{\partial^2 \varphi}{\partial t^2} - \left( \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} \right) = \eta^{\mu\nu} \frac{\partial^2 \varphi}{\partial x^\mu \partial x^\nu} = 0, \]

\[ \eta^{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \]

However, the equation on the right above is not a tensor equation. But \( \partial \varphi / \partial x^\mu \) is a tensor, since \( \varphi \) is a scalar. In other words, the covariant derivative of \( \varphi \) is just the
ordinary derivative. Since $g^{\mu\nu}$ is a tensor, $g^{\mu\nu} \frac{\partial \phi}{\partial x^\mu}$ is a tensor by the quotient rule. In fact, it is a contravariant vector. But
\[\frac{\partial}{\partial x^\nu} \left( g^{\mu\nu} \frac{\partial \phi}{\partial x^\mu} \right)\]
is not a tensor. To make the expression above a tensor, the ordinary derivative \('\frac{\partial \phi}{\partial x^\nu}'\) is replaced by its covariant derivative i.e.
\[\frac{\partial}{\partial x^\mu} \left( g^{\mu\nu} \frac{\partial \phi}{\partial x^\nu} \right) + \Gamma^\mu_{\alpha\beta} g^{\alpha\beta} \frac{\partial \phi}{\partial x^\mu} = \nabla_\mu g^{\mu\nu} \frac{\partial \phi}{\partial x^\nu} = 0\]
The equation above is, in fact, a tensor equation. In particular, it is the wave equation in curved space-time.

The wave equation in flat space-time can be written
\[\eta^{\mu\nu} \frac{\partial^2 \phi}{\partial x^\mu \partial x^\nu} = 0,\]
but is not a tensor equation. A tensor equation must satisfy the continuity equation i.e.
\[\frac{\partial T_{\mu\nu}}{\partial x^\mu} = 0\]
An equation that satisfies the criterion above is
\[T_{\mu\nu} = \phi \frac{\partial}{\partial x^\mu} \frac{\partial}{\partial x^\nu} - \frac{1}{2} \eta_{\mu\nu} \frac{\partial \phi}{\partial x^\alpha} \frac{\partial \phi}{\partial x^\alpha}\]
By direct calculation
\[\nabla_\mu T_{\mu\nu} = 0\]
So,
\[T_{\mu\nu} = \phi \frac{\partial}{\partial x^\mu} \frac{\partial}{\partial x^\nu} - \frac{1}{2} \eta_{\mu\nu} \frac{\partial \phi}{\partial x^\alpha} \frac{\partial \phi}{\partial x^\alpha}\]
satisfies the continuity equations i.e. energy and momentum are conserved. Note that
\[T_{00} = \dot{\phi}^2 - \left( \frac{\partial \phi}{2 \partial t} - \frac{\partial \phi \partial \phi}{\partial x^i \partial x^i} \right) = \frac{1}{2} \dot{\phi}^2 + \frac{\partial \phi \partial \phi}{\partial x^i \partial x^i}\]
The first term in the parentheses is the kinetic energy and the second term is the potential energy. Hence, $T_{00}$ gives the correct value for the energy of a wave. The other components of $T_{\mu\nu}$ also give the correct momentum and energy of a wave. Therefore, $T_{\mu\nu}$ can be inserted into Einstein’s equation i.e.
\[G_{\mu\nu} = -8\pi G T_{\mu\nu}\]
and the left-hand side of this equation will be the geometry of space-time. Note that $T_{\mu\nu}$ is in the form of waves propagating in space-time.

**10.12.5 The Cosmological Constant**

There is one more term which could be added to Einstein’s field equations. This term arises because $\nabla_\mu g_{\mu\nu} = 0$. Hence,

$$G_{\mu\nu} = -8\pi G T_{\mu\nu} \rightarrow G_{\mu\nu} + \Lambda g_{\mu\nu} = -8\pi G T_{\mu\nu},$$

where $\Lambda$ is called the ‘cosmological constant’. Interestingly, $\Lambda$ represents a force which counteracts the gravitational force. It is a repulsive force proportional to distance ‘$r$’. The value of ‘$\Lambda$’ is tiny, and plays virtually no role in the gravitational equations where $r$ is small, but only where $r$ is extremely large, in fact, cosmologically large.

The cosmological constant is included in Einstein’s equations to account for the expansion of the Universe. When it was discovered that the Universe was expanding, the only thing that could account for this expansion was the cosmological constant.

Einstein’s equations can be written

$$G_{\mu\nu} = -8\pi G T_{\mu\nu} - \Lambda g_{\mu\nu}$$

Therefore, ‘$\Lambda g_{\mu\nu}$’ can be thought of as part of the energy-momentum tensor. Suppose $T_{\mu\nu} = 0$, then

$$G_{\mu\nu} = -\Lambda g_{\mu\nu}$$

In a universe devoid of matter and energy, there would still be an expansion of space-time. In effect, space-time would be created.

**10.13 Concluding Remarks**

Gravity is a unique force, the only one proportional the mass of an object. Foundationally, gravity is predicated on the ‘equivalence principle’, which states that there is no difference between falling freely in a flat Earth gravitational field and being at rest in free space, where no gravitational field exists. The amount of energy determines the geometry of space-time, so that space-time is modelled by an intrinsic geometry. If the space-time geometry is flat, gravity is described wholly in terms of the equivalence principle. Non-flat space-time has curvature, which represents an obstruction to describing gravity solely terms the equivalence principle. In essence, curvature in space-time represents such things as tidal forces.

In any sufficiently small area of space-time, the space-time geometry can be considered flat, and hence, modeled using special relativity. General relativity comes about by simply replacing the equations of special relativity with tensors equations and replacing ordinary derivatives with covariant derivatives, which are also tensors. The tensor equations give the laws of gravity independently of the coordinate system used to
describe the space-time geometry. The laws of gravity described in this way will be the same for all observers. This is the beauty of general relativity.
Chapter 11
Quantum Mechanics

“Bohr was inconsistent, unclear, willfully obscure and right. Einstein was consistent, clear, down-to-earth and wrong.”

- John S. Bell

11.0 Introduction

The term ‘quantum mechanics’ applies to a variety of quantum theories. ‘Non-relativistic’ quantum mechanics, introduced by Schrödinger and Heisenberg, includes the basic postulates of quantum mechanics, but does not incorporate relativistic ideas. Paul Dirac and others founded ‘relativistic quantum mechanics’, which incorporates Einstein’s special theory of relativity and led to the discovery of the antiparticle. But relativistic quantum mechanics predicts negative energy states for the electron, states not observed in Nature. Moreover, the antiparticle is regarded, not as a real particle, but the absence of a particle. Neither non-relativistic nor relativistic quantum mechanics easily accounts for multi-particle states. Non-relativistic quantum mechanics requires that the number of particles be conserved in particle interactions, but experiments show that the number of particles before and after a given interaction does not always stay the same. ‘Quantum field theory’, incorporated into the ‘standard model of particle physics’, solved the “multi-particle states” problem by associating each particle with a different field, where particles arise through the quantization of the fields. Quantum field theory treats the antiparticle as a particle, rather than the absence of a particle.

In this chapter, the treatment given quantum mechanics follows closely the lectures by Prof. Leonard Susskind of Stanford University, Dr. Bob Eagle and Dr. Paul Langacker. Readers are encouraged to view their excellent online presentations. The emphasis in this discussion will be on the logical foundations, rather than the calculational aspects of the subject and will highlight non-relativistic quantum mechanics, where particles move much slower than the speed of light.

11.1 Blackbody Radiation

Trouble for classical physics began when Max Planck initiated a study of how radiant energy was absorbed and emitted from the walls of a continuously heated hollow body. As the body was heated, the amount of energy absorbed by the body was proportional to the temperature. But as heating increased, a temperature would eventually be reached where the body emitted just as much radiant energy as it absorbed. At the time, the absorption and emission of heat energy by a blackbody was described by the Rayleigh-Jeans law:

\[ U(\nu, T) \propto T\nu^2, \]

where \( U \) is the energy, \( T \) the temperature, \( \nu \) the frequency of the radiation. While the Rayleigh-Jeans law was accurate for low frequencies, as the temperature increased, it
predicted that the amount of radiant energy emissions would increase without bound. This was contrary to what was observed experimentally. A rough description of the discrepancy is shown in fig. 11.1-1.

**Figure 11.1-1**

### 11.1.2 Wien’s Law

In 1896, the German physicist Wilhelm Wien (13 January 1864 – 30 August 1928) introduced Wien’s displacement law:

\[
U(v) \, dv \approx v^3 e^{-hν/kT} \, dv
\]

His equation accurately described short wavelength (high frequency) thermal emissions, but its predictions for low frequency emissions contradicted experimental results. The locus of his equation gave a curve similar to the one Planck ultimately settled on. His equation did not depend on the temperature, but only on the frequency of the radiation and included the expression ‘\(hν\).

### 11.1.3 Planck’s Solution

Planck’s analysis was entirely empirical. The revolutionary nature of his results did not occur to him. He accepted the wave nature of light and several years would pass before an entirely new theory for describing light phenomena emerged. To solve the problem of the emission and absorption of radiant energy, Planck assumed that the walls of the cavity only emitted or absorb energy in discrete multiples of the basic unit ‘\(hν\), where \(h\) is a constant (\(h \approx 6.6 \times 10^{-27} \ [\text{erg} \times \text{sec}]\)). This basic unit of energy is proportional to the frequency of the radiation i.e.

\[
E_n = nhν, \quad n \in N, \quad n = 1, \ldots, \infty
\]

Classically, there is no restriction on the energy that anything that oscillates can obtain. Experience with clocks and pendulums seemed to confirm this. Planck’s assumption was plausible only because \(h\) was an extremely small number. At low frequencies, such as those with clocks and pendulums, the tiny quantization of energy is
unobservable. However, at light frequencies of $10^{15}$ [cycles/sec], quantization can be detected with sensitive instruments.

Moreover, Planck assumed that the probability that a quantum of energy would either be absorbed or emitted was a function of the energy $E_n$. He settled upon

$$ P(n) \propto e^{-E_n/\kappa T} = e^{-h\nu/\kappa T} \rightarrow P(n) = \frac{e^{-h\nu/\kappa T}}{\sum_{n=0}^{\infty} e^{-h\nu/\kappa T}} = e^{-h\nu/\kappa T} \left(1 - e^{-h\nu/\kappa T}\right), $$

$$ \frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \rightarrow \frac{1}{\sum_{n=0}^{\infty} e^{-h\nu/\kappa T}} = \left(1 - e^{-h\nu/\kappa T}\right), \quad x = e^{-h\nu/\kappa T}, $$

where $P(n) = e^{-E_n/\kappa T}$ represents the probability that the cavity will emit or absorb of an amount of energy $nh\nu$. The average energy $\langle E \rangle$ is

$$ \langle E \rangle = \sum_{n=0}^{\infty} E_n P(n) = (1 - e^{-h\nu/\kappa T}) \sum_{n=0}^{\infty} e^{-h\nu/\kappa T} nh\nu = h\nu (1 - e^{-h\nu/\kappa T}) \sum_{n=0}^{\infty} ne^{-h\nu/\kappa T} $$

By the well-known summation formula

$$ \sum_{n=0}^{\infty} ne^{-na} = \frac{\frac{1}{n+1}}{(1-e^{-a})^2} $$

and letting $a = h\nu/\kappa T$, then

$$ \langle E \rangle = \frac{h\nu \left(e^{-h\nu/\kappa T}\right)}{1 - e^{-h\nu/\kappa T}} $$

Recall the Rayleigh-Jeans law:

$$ U(v, T) = K\kappa T v^2 = K \langle E \rangle v^2, $$

where $\langle E \rangle = \kappa T$. Planck substituted the value for $\langle E \rangle$ into the Rayleigh-Jeans law obtaining

$$ U(v, T) = K \langle E \rangle v^2 = K \frac{h\nu^3 \left(e^{-h\nu/\kappa T}\right)}{1 - e^{-h\nu/\kappa T}}, $$

where $K = 8\pi N / c^3$. The equation above is the one Planck untimely settled on. If $h\nu/\kappa T$ is small, Planck’s predictions agree with the Rayleigh-Jeans law. As $h\nu/\kappa T$ becomes larger, its predictions agree with Wien’s law [24].

### 11.2 The Photoelectric Effect

Shining light of a given frequency upon certain metals will knock electrons out of the metal. This is called the ‘photoelectric effect’. Classical electromagnetic theory predicts that, if a light beam with electromagnetic energy of $10^{-6}$ [watts/m$^2$] falls on a sodium
surface, it would take $10^7 \text{[sec]}$ or more than a year before a single electron would be emitted [45].

But experiments revealed no emission time lag anywhere near the predictions of classical theory. Moreover, unless the light reached or exceeded a certain threshold frequency $\nu_0$, no electrons would be emitted from the metal, regardless of the light intensity or the length of exposure to the light. This suggested that the energy of the emitted electrons depended on the frequency of the light waves, not the intensity of the light. In other words, faint blue light, having higher frequency, would emit electrons of higher energy than bright red light, which has lower frequency.

11.2.1 The Einstein/Millikan Solution

Einstein proposed that a beam of light was not a wave propagating through space, but rather, a collection of discrete wave packets (photons), each with energy $'hv'$. His hypothesis aligned with Planck's discovery linking energy ($E$) and frequency ($\nu$) through the quantization of energy. In 1905, Einstein published a paper explaining that the photoelectric was the result of light energy carried in discrete packets. In 1914, experiments conducted by Robert Millikan confirmed Einstein's law on the photoelectric effect. Einstein was awarded the Nobel Prize in 1921 for the “law of the photoelectric effect” and Millikan won the Nobel Prize in 1923 for the “elementary charge of electricity and on the photoelectric effect”.

The solution is modelled as though the electron is confined to a well. A certain amount of energy is required to kick the electron out of the well. The well represents the force that binds the electron to the nucleus of the atom. If a photon strikes an electron, it may or may not provide enough energy to kick the electron out of orbit. If the amount of energy necessary to push the electron out of orbit is specified by $\nu_0$, then the fundamental formula for Einstein's law is given by

$$T_{\text{max}} = h(\nu - \nu_0),$$

where $h
\nu_0$ is the threshold energy, below which no electron emissions occur. The quantity $\nu_0$ depends on the type of substance the light is falling upon. The quantity $T_{\text{max}}$ is the maximum amount of electron energy and $'hv'$ is the energy content of each quantum of light (see fig. 11.2.1-1).
11.2.2 The Compton Effect

Initially, Einstein’s proposal elicited incredulity from his contemporaries. But there was
one way of testing whether light quanta, except for the absence of a rest mass, behaved
like particles. If a photon collided with an electron, the amount of energy lost should
equal the amount of kinetic energy ‘\(T\)’ gained by the electron. In other words, since \(E = h\nu\),
if the light quanta has frequency ‘\(\nu\)’ before the collision, then it should have
frequency ‘\(\nu'\)’ after the collision (\(\nu' < \nu\)) i.e. the loss in photon energy should equal the
gain in electron energy:

\[
h\nu - h\nu' = T
\]

Recall from the theory of special relativity that \(E = \sqrt{m_0^2c^4 + p^2c^2}\) and, since a photon
has no rest mass, \(m_0 = 0 \rightarrow E = pc\). Since \(E = h\nu\), then

\[
p = \frac{E}{c} = \frac{h\nu}{c}
\]

Momentum must be conserved in each of two mutually perpendicular directions. In the
direction that the photon is moving

\[
\frac{h\nu}{c} + 0 = \frac{h\nu'}{c}\cos \phi + p\cos \theta,
\]

where \(h\nu/c\) and \(h\nu'/c\) are the initial and final momentum of the photon respectively and
‘0’ and ‘\(p\)’ are the initial and final momentum of the electron respectively. In the
direction perpendicular to this

\[
0 = \frac{h\nu'}{c}\sin \phi - p\sin \theta,
\]
where $\phi$ is the angle between the initial and scattered photon and $\theta$ is the angle between the initial photon and the scattered electron. Multiplying both equations above by $c$ gives

\[ pc \cos \theta = h\nu - h\nu' \cos \phi, \quad pc \sin \theta = h\nu' \sin \phi \]

Squaring and adding the two equations together leaves

\[ p^2 c^2 = (h\nu)^2 - 2(h\nu)(h'\nu') \cos \phi + (h'\nu')^2 \]

Since

\[ E = T + m_0c^2, \quad E = \sqrt{m_0^2c^4 + p^2c^2}, \]

then

\[ (T + m_0c^2)^2 = m_0^2c^4 + p^2c^2, \quad p^2c^2 = T^2 + 2m_0c^2T \]

Since $T = h\nu - h\nu'$, then

\[ p^2c^2 = (h\nu)^2 - 2(h\nu)(h'\nu') + (h'\nu')^2 + 2m_0c^2(h\nu - h'\nu') \]

Recalling that $p^2c^2 = (h\nu)^2 - 2(h\nu)(h'\nu') \cos \phi + (h'\nu')^2$, then

\[ m_0c^2(h\nu - h'\nu') = (h\nu)(h'\nu')(1 - \cos \phi) \]

Dividing the equation above by $(hc)^2$ leaves

\[ \frac{m_0c}{h} \left( \frac{\nu}{c} - \frac{\nu'}{c} \right) = \frac{\nu\nu'}{cc} (1 - \cos \phi) \rightarrow \frac{m_0c}{h} \left( \frac{1}{\lambda} - \frac{1}{\lambda'} \right) = \frac{1 - \cos \phi}{\lambda\lambda'} \rightarrow \lambda' - \lambda = \frac{h}{m_0c} (1 - \cos \phi), \]

where $\lambda = c/\nu$ is the wave length. The equation on the right above was derived by Arthur Compton and also experimentally confirmed by him. It was a spectacular verification that Einstein’s supposition that light consisted of discrete quanta should be taken seriously [45].

However, the fact that light consisted of discrete wave packets created a theoretical dilemma. The classical wave theory of light provided the sole explanation for a host of optical effects, like diffraction and interference etc. The energy of a wave was distributed continuously throughout the wave pattern. So successful was the wave theory, there had to be some level of truth to it. The quantum theory of light described the photoelectric effect, but, in this case, light was described as a particle. In classical physics, nothing behaved like a particle and a wave. And this was not a case similar to relativistic verses Newtonian mechanics, where the latter could be explained as an approximation to the former. There simply was no a way of deriving the quantum theory of light from the wave theory and vice versa.
11.2.3 The Theory of Atoms

The idea of ‘discrete quantities of energy’ eventually led to the conjecture that the fundamental elements of Nature consisted of ‘atoms’. Rutherford hypothesized a miniature solar system model in which the atom consisted of a central core of positive charges orbited by a collection of an equal number of negatively charged electrons, making the atom roughly electrically neutral. Rutherford’s conjectures eventually led to particle physics. Several properties of Rutherford’s model agreed with experiments, but two issues arose: 1) if the core consisted of a collection of repulsive positive charges, how did it remain intact? And 2) the electrons orbiting the core were accelerating. And accelerating electrons lose radiant energy, forcing them into smaller orbits. Eventually, the electrons would spiral into the core effectively destroying the atom. But atoms are stable and the spiraling effect does not occur. Modern theory maintains that the nuclear core is held together by the ‘strong nuclear force’, which overcomes the repulsive force associated with the positive charges in the core. The ‘spiraling into the nucleus problem’ was solved by Bohr, who hypothesized that electrons could only assume certain orbits.

11.3 Wave/Particle Duality

Despite solid experimental evidence, Einstein’s hypothesis that light consisted of discrete quanta was ceremoniously ignored for several years. Interestingly, by the time, in 1924, that Louis de Broglie discovered that, not only did light have particle properties, but all particles had wave properties, the intellectual climate had transformed so dramatically that de Broglie’s hypothesis was accepted almost immediately without strong experimental evidence to support it. The existence of de Broglie waves was not experimentally demonstrated until 1927. de Broglie hypothesized that, since a photon of light had momentum

\[ p = \frac{h \nu}{c} = \frac{h}{\lambda} \rightarrow \lambda = \frac{h}{p}, \quad \lambda \nu = c, \]

and, since \( p = mv \) for any particle, the wavelength of a particle would be

\[ \lambda = \frac{h}{mv}, \quad m = \frac{m_0}{\sqrt{1 - v^2/c^2}}, \quad m_0 = \text{the rest mass} \]

His hypothesis was confirmed by diffraction experiments involving fast moving X-rays.
11.3.1 The Quantization of Linear and Angular Momentum

de Broglie’s discovery gave credence to the Rutherford model of the atom by suggesting that the electron, instead of moving around the nucleus in roughly a circle, vibrated as a ‘standing wave’. A standing wave requires that the radius ‘$R$’ of the circumference of the orbit of the electron must be an integer multiple of the wavelength ‘$\lambda$’ i.e. $n\lambda = 2\pi R$, $n \in N$. The electron wave maintains its integrity by returning to the same point each time it orbits the nucleus (see fig. 11.3-1) i.e.

$$n\lambda = 2\pi R \rightarrow \frac{nh}{p} = 2\pi R \rightarrow pR = \frac{nh}{2\pi} = nh, \quad \lambda = \frac{h}{p}, \quad \hbar = \frac{h}{2\pi}$$

Note that $pR = L$ is the angular momentum of the electron, and hence, $L = n\hbar$. In other words, both momentum and angular momentum are quantized, restricted to only certain discrete values proportional to $\hbar$.

11.3.2 Rutherford’s Modified Model of the Atom

The radius of the orbit ‘$R$’ is also quantized. The electron can only assume certain discrete orbits. This prevents it from spiraling into the nucleus. If it loses energy, the electron must jump from one allowable orbit to another.

If the electron is both a particle and a wave, how is it described in terms of its wave motion? The general equation for a wave ‘$\Psi$’ is

$$\Psi(x, t) = Ae^{i(kx - \omega t)} = A(\cos(kx - \omega t) + i \sin(kx - \omega t))$$

where $k = 2\pi/\lambda$ and $\omega = 2\pi v$. Since $\lambda = h/p$, $p = h/k \rightarrow p = k$, $\hbar = 1$

Hence, $k$ is proportional to the momentum of the electron. Moreover, since $E = h\nu$, $\omega$ is proportional to the energy of the electron i.e.

$$\omega = E/\hbar \rightarrow E = \omega\hbar = \omega, \quad \hbar = 1$$
Rutherford had proposed a model of the atom as a miniature solar system. But his model predicted that circling electrons would lose energy causing them to spiral into the nucleus. Atoms would be unstable. Rutherford considered scrapping his miniature solar system model until Bohr showed that if the electrons circled the nucleus in orbits with an integer number of de Broglie wavelengths, the electrons would maintain orbits without losing energy. Bohr’s insights saved the Rutherford model, but meant de Broglie’s hypothesis that particles possessed wave properties had to be taken seriously.

11.4 Classical Rationality vs. Quantum Weirdness

That classical physics failed to explain a wide variety of atomic phenomena made finding an entirely new theory inevitable. From about 1925 onward, ‘quantum mechanics’ would evolve through several iterations, eventually producing the ‘standard model of particle physics’. The remainder of this chapter examines the evolution of the subject.

Classical physics makes sense. Its underpinnings are, more or less, based on propositional logic which, for all intent and purposes, is well understood. In a certain sense, all classical physics is deterministic, even if the systems being investigated are treated statistically. If enough information is known about a system, classical theory dictates that all the future states of the system can be predicted with certainty.

11.4.1 The Anatomy of a System

Consider the simple system of tossing a coin, where, intuitively, on any given toss, the probability that the coin shows a head or a tail is 1/2. The reasoning is sound if nothing interferes with the symmetry of the coin, which is the most significant factor in determining how often it lands heads or tails. The same thing can be said of a perfectly symmetrical die. The probability of tossing a 1, or any other number on the die, is 1/6.

Since there are no perfectly symmetrical coins or dies, researchers study systems dynamically. There are two aspects to this: 1) determine the configuration of the system, 2) formulate a rule describing how the system changes over time. For instance, if the system is a die, then its configuration is a six-sided cube. Suppose the die shows a certain number, then, over time, changes constantly from number to number as
shown in fig. 11.4-1 a). If the time periods are short enough, the die will be no more likely to show a ‘1’ than any other number. A reasonable conclusion over a sufficiently long period of time is the probability that the die shows a given number at a given time would be 1/6.

11.4.1.1 Conservative Systems

But suppose the system follows the rule shown in fig. 11.4-1 b). In this case, the evolution rule depends on which number the die initially shows. There is a probability that the system will be governed by the first and a probability the system will be governed by the second rule. But once it is known which rule governs the system, the die will follow that rule. This illustrates the notion of ‘conservation’. Once it is determined which rule governs the system, the rule does not change.

11.4.1.2 Non-Classical Systems

Consider the evolution rule shown in fig. 11.4-1 c). No matter what number the die shows initially, it migrates to ‘1’. Over time, the probability that the die would show a number other than ‘1’ is close to zero. Unlike the two previous examples, in this case, if the die shows a ‘1’, it is impossible to know what number the die showed previously.

Classical systems always comply with the rule structures illustrated in figures 11.4-1 a) or b), but never follow rule structures like those illustrated in fig. 11.4-1 c). In other words, if a system begins with a number of characteristic distinctions, it will maintain those distinctions over time. This is often called the ‘conservation of distinctions’, exemplified by Liouville’s theorem.

11.4.1.3 Liouville’s Theorem

There are two aspects to Liouville’s theorem: 1) given the initial status of a system at some initial time ‘\(D(t = 0)\)’, if the system has a number of distinct characteristics, then those characteristics will remain distinct at a later time ‘\(t\)’, designated ‘\(D(t)\)’; 2) if the volume of the system at \(D(t = 0)\) is \(V_0\) and if the volume of the system at \(D(t)\) is \(V_t\), then \(V_0 = V_t\). In other words, the volume of the system is conserved over time (see fig. 11.4.1.3-1).

![Figure 11.4.1.3-1](image)
If at $t = 0$ a system contains a small collection of distinct initial conditions on a set of trajectories, the trajectories will evolve in time according to Hamilton's equations of motion, and, at a later time '$t$', will be located in a new volume element as shown in fig. 11.4.1.3-1, exemplified by Liouville's theorem:

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{n} \left( \frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) = 0,$$

where $\rho$ is the volume element and $q_i$ and $p_i$ are the positions and momentums respectively of the distinctions in the system. Liouville's theorem implies that, given a set of distinct initial conditions, a one-to-one and onto mapping exists between the initial distinctions and the distinctions in the modified volume element.

All classical systems comply with Liouville's theorem. Assuming nothing interferes with the rule by which a system evolves, then, if time is reversed, the system will have exactly the same collection of distinct initial conditions as it started with.

### 11.4.2 The Two-Slot Interference Experiment

Consider the two-slot interference experiment illustrated in fig. 11.4.2-1.

![Two-Slot Interference Experiment](image)

Figure 11.4.2-1

A stream of electrons moves toward a wall that has two small openings, labeled ‘1’ and ‘2’. Some of the electrons will pass through the openings making an impact on the absorber. The electrons will pass through the openings at various angles making an impact at different locations along the absorber. If one of the openings, say ‘2’, is closed, then the distribution of strikes will look like the pattern shown by ‘$I_1$’ (see fig. 11.4.2-1).

The distribution of points ‘$I_1$’ can be thought of as a probability distribution, where the graph represents the probability that an electron will strike the absorber at a given location. According to the graph, the most likely location is directly across from the opening the electron passes through. Electron strikes some distance away from this
“center point” become less and less likely as the distance from the center point increases in either direction. If slot ‘1’ is closed, then the electron impact distribution is described by the pattern ‘$I_2$’ (see fig. 11.4.2-1 (b)).

Classical theory predicts that, if both slots are open, the probability distribution that emerges is the sum ‘$I_1 + I_2$’ of the two single slot distributions. But this prediction does not agree with what is observed experimentally. Instead, the distribution looks like the one shown in fig. 11.4.2-1 (c). Classical theory predicts that if both slots are open, there will be a non-zero probability that a given electron will strike the detector. But, in fact, if both slots are open, the probability that an electron strikes the detector is zero. Moreover, there are many locations along the absorber where the probability of an electron striking there is zero.

11.4.2.1 Nature’s Conspiracy

What is the correct solution? Logically speaking, the answer generally given is not entirely satisfactory. If only one slot is open, the electron behaves like a particle. The physics of particle mechanics can be applied. But, if both slots are open, the electron displays interference effects, a property of a wave. To some extent, wave mechanics must be part of the solution. So, is an electron a particle or a wave?

If the experiment is slowed down, letting the electron gun emit only one electron at a time, intuitively, the electron should pass through only one of the two slots. But, this does not change the results of the experiment. With both slots open, over time, the absorber shows an interference pattern characteristic of a wave.

Additionally, a light beam could be directed on to the wall to see which slot the electron passes through. However, in order to observe the electron, the light beam must contain enough energy, in other words, have sufficiently short frequency so that the electron can be detected. But when this level of light energy is obtained, the interference pattern on the absorber disappears. It seems Nature has conspired, making it impossible to know “a-priori” whether an electron is a wave or a particle. This knowledge can only be obtained upon taking a measurement.

11.4.2.2 What is a Measurement?

But what constitutes a measurement? Evidently, a measurement is something that interferes with the internal evolution of a system. For instance, consider the die system illustrated in fig. 11.4-1 a). Suppose the system starts with the die showing a ‘6’, then evolves in accordance with the given rule. Suppose a deity intervenes, declaring that on the $10^{th}$ go round, when the system shows a ‘6’, it will continue to show a ‘6’. The intervention changes the rule by which the system is evolving. If time is reversed and the system suffers no interference, the system will return to its initial condition, showing a ‘6’. But, if the evolution of the system is interfered with, upon time reversal, the system can return to its initial state only probabilistically.

Evidently, the act of measuring changes the rule by which a system evolves. In fact, in quantum physics, the act of measuring determines if the electron behaves like a particle
or a wave. Those experiments designed to measure the particle properties of the electron will see the electron as a particle, while experiments designed to measure the wave properties will see the electron as a wave. Prior to taking a measurement, there is no way of determining whether the electron is a wave or a particle. This explanation is very different from a classical one, where it is assumed that measurements taken on systems do not appreciably interfere with the internal evolution of the system.

### 11.4.3 The Heisenberg Uncertainty Principle

Classical physics assumes that light can be conceived as a wave, and therefore, divided into as small a segment as desired. As such, the light energy needed to observe a particle can be made as small as desired and will have no significant impact on the nature of the particle. But the fact that light comes in discrete wave packets with energy ‘\( \hbar \nu \)’ levies a fundamental limitation on how much the energy in the wave can be reduced. The exact location of an electron can be found by shining a beam of light on the electron. Light has energy, and hence, momentum i.e.

\[
E = cp = \hbar \nu \rightarrow p = \hbar \nu / c = \hbar / \lambda,
\]

where \( p \) is the momentum, \( c \), the speed of light and \( \lambda \) the wave length of the light. In order to obtain a sharpened photo, the wavelength of the light illuminating the object must be shorter than the object being photographed. To get the exact location of the electron, light of a very high frequency (very high momentum) is required. But a very high momentum light packet smashing into the electron will dramatically alter the momentum of the electron. At the very moment a position measurement is obtained, the momentum of the electron becomes significantly different after, than before the position measurement. Note that \( \Delta x \), the ‘error’ in the position measurement, is proportional to \( \lambda \). Recall that \( p = \hbar / \lambda \rightarrow p \propto \hbar / \Delta x \). Hence, if \( \Delta x = 0 \) (no error in the position measurement), then \( p = \infty \), momentum becomes indefinite.

This is the logic underlying the Heisenberg uncertainty principle, which says that multiple characteristics of a system cannot be measured to any desired degree of accuracy at the same time. The logic of the quantum, whatever it may be, seems entirely different than the logic that supports classical physics.

### 11.5 Quantum Mathematics

Classical physics is based on the idea of the ‘point set in phase space’\(^{10}\) (see fig. 11.5-1). All classical systems can be described by a set of points in phase space, where the points represent the momentum and position of the system at any moment in time. Since classical physics assumes that momentum and position can be measured simultaneously, all classical systems evolve in time as a set of points in phase space that comply with Hamilton’s principle of least action and Liouville’s theorem.

---

\(^{10}\) Phase space is a multidimensional space in which each axis corresponds to one of the coordinates required to specify the state of a physical system, all the coordinates being thus represented so that a point in the space corresponds to a state of the system.
11.5.1 Linear Algebra and Vector Spaces

Conversely, rather than a point set in phase space, the fundamental notion in a quantum mechanical system is the ‘vector’ in a ‘vector space’, designated ‘\((V, +, \cdot)\)’. Vectors represent the ‘states’ of a system. Vectors comply with the algebra of vector spaces:

\[
\begin{align*}
\text{i)} & \quad |A\rangle + |B\rangle + |C\rangle = |A\rangle + (|B\rangle + |C\rangle) \\
\text{ii)} & \quad |A\rangle + |0\rangle = |A\rangle \\
\text{iii)} & \quad |A\rangle - |A\rangle = |0\rangle \\
\text{iv)} & \quad |A\rangle + |B\rangle = |B\rangle + |A\rangle \\
\text{v)} & \quad k \cdot (|B\rangle + |A\rangle) = k \cdot |A\rangle + k \cdot |B\rangle \\
\text{vi)} & \quad (h + k) \cdot |A\rangle = h \cdot |A\rangle + k \cdot |A\rangle \\
\text{vii)} & \quad (h \cdot k) \cdot |A\rangle = h \cdot (k \cdot |A\rangle) \\
\text{viii)} & \quad 1 \cdot |A\rangle = |A\rangle \text{ and } 0 \cdot |A\rangle = |0\rangle,
\end{align*}
\]

where \(|A\rangle, |B\rangle, |C\rangle\) signify vectors and \(k, h, 1\) and \(0\) are ‘scalars’ (numbers) which belong to some number field ‘\(K\)’. The set ‘\(V\)’ specifies a collection of vectors, while ‘+’ stipulates vector addition and ‘\(\cdot\)’ indicates scalar multiplication. Readers unfamiliar with vector spaces should review Book II: Chapter 8.

11.5.1.1 Inner Product Spaces

The vectors associated with quantum mechanical systems are of a special kind: 1) the number field is complex i.e. \(K = \mathbb{C}\), 2) any two vectors have an inner product, which, in general, is a complex number. This extension of a vector space to inner products is sometimes called an ‘inner product’ or ‘Hilbert’ space, designated ‘\((\langle V \rangle, +, \cdot)\)’.

Example: Suppose \(V\) consists of the set of complex functions i.e. \(V = \{\Psi_j(x), \ x \in \mathbb{C}\}\). Note that \(\Psi_j(x) = Re(\Psi_j(x)) + i Im(\Psi_j(x))\), where \(Re(\Psi_j(x))\) specifies the ‘real part’ of \(\Psi_j\) and \(Im(\Psi_j)\), the ‘imaginary part’ and that \((V, +, \cdot)\) forms a vector space over the complex numbers. To see this, note that \(\Psi_j + \Psi_h = \Psi_k \in V\) and \(\alpha \Psi_j = \Psi_s \in V, \ \alpha \in \mathbb{C}\). Hence, the \(\Psi_j\)’s comply with the algebraic rules of vector addition and scalar multiplication, and therefore, constitute a vector space. Notably \(\{\Psi_j(x), \ x \in \mathbb{R}\}, +, \cdot\) does not constitute a vector space over the real numbers, since if \(\alpha \in \mathbb{R}, \alpha \Psi_j \notin \mathbb{R}\). The inner product ‘\(\langle \Psi | \Phi \rangle\)’ of two complex functions is defined as
\[
\langle \Psi | \Phi \rangle = \int \Psi^\dagger(x) \Phi(x) \, dx, \quad \Psi, \Phi \in V
\]

### 11.5.1.2 Row and Column Vectors

Given the complex number \( z = a + ib, \ a, b \in R \)', \( z^\dagger = a - ib \) is called the ‘complex conjugate’ of \( z \). More generally, suppose

\[
|a\rangle = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \quad a_i \in C \rightarrow \langle a | = \begin{bmatrix} a_1^\dagger & a_2^\dagger & a_3^\dagger & a_4^\dagger \end{bmatrix}
\]

In other words, \( \langle a | \) is the row vector whose entries are the complex conjugates of the entries of the column vector \( |a\rangle \). The inner product of \( |a\rangle \) with itself is given by

\[
\langle a|a\rangle = \sum_{i=1}^{n} a_i^\dagger a_i \in R,
\]

Hence, \( \langle a|a\rangle \) is a real number, since a complex number multiplied by its complex conjugate is a real number. In general, the inner product of a vector \( |a\rangle \) with a vector \( |b\rangle \) is a complex number i.e.

\[
\langle a|b\rangle = \sum_{i=1}^{n} a_i^\dagger b_i \in C
\]

Inner products of vectors have the following properties:

1. \( \langle a|\beta|b\rangle = \beta \langle a|b\rangle, \ \beta \in C \)
2. \( \langle a|(|b\rangle + |c\rangle) = \langle a|b\rangle + \langle a|c\rangle \)
3. \( \langle a|b\rangle = \langle b|a\rangle^\dagger \)
4. \( \langle a|a\rangle \geq 0 \in R \)

#### 11.5.1.2.1 The Dual Space

There is a one-to-one and onto mapping between \( |a\rangle \) and \( \langle a \rangle \) i.e. \( f(|a\rangle) = \langle a | \). Since \( f \) is one-to-one and onto, the set ‘\( \{|a_i\}\)’ is isomorphic to ‘\( \langle a_i \rangle \)’. And if \( (\{|a_i\}, +, \cdot) \) represents a vector space, then \( (\langle a_i \rangle, +, \cdot) \) represents the same vector space, often called the ‘dual’ of \( (\{|a_i\}, +, \cdot) \).

The rules stated above hold for matrices. If \( H \) is a matrix with complex entries, then \( H^\dagger \) is the matrix formed by transposing the row vectors of \( H \) into column vectors and replacing all the entries of \( H \) by their complex conjugates. Note that \( H \) is a square matrix.
11.5.2 Basis Vectors

What is the minimum number of vectors required to form a vector space? Consider the following:

\[
|a\rangle = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \rightarrow |a\rangle = a_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + a_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + a_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}
\]

It requires a minimum of four vectors to write every vector in this vector space. The vectors

\[
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}
\]

are called ‘basis vectors’, which are not unique. The vector ‘\(|a\rangle\)’ is said to be a ‘linear combination’ of the basis vectors.

11.5.2.1 The Independence of Basis Vectors

Basis vectors are independent. No basis vector can be written as a linear combination the other basis vectors. In general, if

\[
\forall i \left( |v_i\rangle \in V \land c_i \in C \land \sum_{i=1}^{n} c_i |v_i\rangle = |0\rangle \rightarrow c_i = 0 \right),
\]

then the vector space ‘(V, +, ∙)’ has dimension ‘n’. A vector space of dimension ‘n’ must have n basis vectors.

Example: The set ‘V’ of complex functions ‘Ψ_j(x)’ has an infinite number of basis vectors.

11.5.2.2 Orthogonal and Orthonormal Basis Vectors

Suppose \{ |b_i\rangle \} is a set of vectors. If \( \forall i (|b_i\rangle | = 1) \), then the set of vectors \‘\{ |b_i\rangle \}’ is called a ‘normalized’ set. If \( \forall i \forall j (\langle b_i | b_j \rangle = \delta_{ij}) \), the set of vectors \‘\{ |b_i\rangle \}’ is called an ‘orthonormal’ set. If a set of vectors is orthonormal, then the vectors form a basis of a vector space. Moreover, any vector belonging to the vector space can be written as a linear combination of orthonormal basis vectors i.e.

\[
|v\rangle \in V \rightarrow |v\rangle = \sum_{j} v_j |b_j\rangle, \quad v_j \in C
\]
Note that

\[ |v\rangle \in V \rightarrow \langle b_i | v \rangle = \sum_j v_j \langle b_i | b_j \rangle = v_i, \]

since \( \langle b_i | b_j \rangle = 0 \) if \( i \neq j \) and \( \langle b_i | b_j \rangle = 1 \) if \( i = j \). Hence,

\[ |v\rangle \in V \rightarrow |v\rangle = \sum_j v_j |b_j\rangle = \sum_j |b_j\rangle \langle b_j | v \rangle = \sum_j |b_j\rangle \langle b_j | b_j \rangle = 1 \]

11.6 Quantum States

In classical mechanics, states are associated with points in a set. A coin can be in two states \( |H\rangle \) or \( |T\rangle \), representing heads or tails. States in classical physics are either distinct or the same. The conception of a coin in classical physics is that it can show a head or a tail, but not at the same time. However, in quantum mechanics, statements like \( \alpha |H\rangle + \beta |T\rangle \), \( \alpha, \beta \in \mathbb{C} \) are considered meaningful states, states which are linear combinations of distinct states. This does not happen in classical physics.

Two quantum mechanical states \( |a\rangle \) and \( |b\rangle \) are 'distinct' if \( \langle a | b \rangle = 0 \). In other words, the two vectors, representing states, are orthogonal. If, in addition, \( \langle a | a \rangle = \langle b | b \rangle = 1 \), the states are orthonormal.

11.6.1 The Probabilistic Nature of Quantum States

What is the meaning of the state \( |v\rangle = \alpha |H\rangle + \beta |T\rangle \)? The coefficients \( \alpha, \beta \) are associated with the probability that the system is in state \( |H\rangle \) or in state \( |T\rangle \). To find the probability relationship between \( \alpha \) and \( \beta \) take the inner product \( \langle v | v \rangle \) i.e.

\[
\langle v | v \rangle = \left( \langle H | \alpha^\dagger + \langle T | \beta^\dagger \right) | \alpha | H \rangle + \beta | T \rangle \right) = \alpha^\dagger \alpha \langle H | H \rangle + \alpha^\dagger \beta \langle H | T \rangle + \beta^\dagger \alpha \langle T | H \rangle + \beta^\dagger \beta \langle T | T \rangle
\]

The quantity \( P_H = \alpha^\dagger \alpha \) is the probability that the system will be found in state \( |H\rangle \) and \( P_T = \beta^\dagger \beta \) is the probability that the system will be found in state \( |T\rangle \). Since a coin can be found in only two distinct states \( |H\rangle \) or \( |T\rangle \), then

\[
\alpha^\dagger \alpha + \beta^\dagger \beta = P_H + P_T = 1
\]

In any quantum system, if

\[ |v\rangle = \sum_j \alpha_j |b_j\rangle, \]
where $|b_j\rangle$'s are orthonormal basis vectors, then

$$\langle v|v \rangle = \sum_j \alpha_j^\dagger \alpha_j \langle b_j|b_j \rangle = \sum_j \alpha_j^\dagger \alpha_j = 1$$

### 11.7 Quantum Operators

Not all states within a quantum mechanical system are observable. For instance, a quantum system is never observed in a superposition of states, but always in one of the possible distinct states. If a coin is placed in a box and the box opened, the coin will show a head or a tail, but never found in a state of “both heads and tails”.

The ‘operator’ represents an observable in quantum mechanics. There is an operator for each observable. Operators are linear, having square matrix representations. Customarily, operators are represented by a letter with a hat on top, say ‘$\hat{O}$’ (read $O$-hat), or just by a capital letter, say ‘$O$’. Mathematically, operators act on vectors to produce other vectors i.e. $\hat{T}|u\rangle = |v\rangle$, where $\hat{T}$ is an operator and $|u\rangle, |v\rangle$ are vectors in a vector space. Linear operators have two important properties:

1. $\hat{T}\alpha|u\rangle = \alpha\hat{T}|u\rangle = \alpha |v\rangle$, $\alpha \in \mathbb{C}$
2. $\hat{T}(\alpha|u\rangle + \beta|v\rangle) = \alpha\hat{T}|u\rangle + \beta\hat{T}|v\rangle$, $\alpha, \beta \in \mathbb{C}$

In other words, linear operators comply with the algebraic rules of vector addition and scalar multiplication, and hence, form a vector space.

#### 11.7.1 Matrix Representations of Operators

Linear operators can be represented by a square matrix. Recall that if $\hat{T}|u\rangle = |v\rangle$, then $|v\rangle$ is simply another vector. Now consider

$$\langle w|\hat{T}|u \rangle = \langle w|v \rangle = T_{wu},$$

where $T_{wu}$ is the $(w, u)^{th}$ entry of the matrix ‘$\hat{T}$’. Suppose $\{|u_i\rangle\}$ is a set of $n$ basis vectors. In general,

$$\hat{T} = [\hat{T}_{ij}] = \langle u_i|\hat{T}|u_j \rangle = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n} \\ T_{21} & T_{22} & \cdots & T_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{bmatrix} = \begin{bmatrix} \langle u_1|\hat{T}|u_1 \rangle & \langle u_1|\hat{T}|u_2 \rangle & \cdots & \langle u_1|\hat{T}|u_n \rangle \\ \langle u_2|\hat{T}|u_1 \rangle & \langle u_2|\hat{T}|u_2 \rangle & \cdots & \langle u_2|\hat{T}|u_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_n|\hat{T}|u_1 \rangle & \langle u_n|\hat{T}|u_2 \rangle & \cdots & \langle u_n|\hat{T}|u_n \rangle \end{bmatrix}$$

Example: Let $\{|u_i\rangle\}$, $i = 1, \ldots, 3$ be a set of three basis vectors. Suppose an operator ‘$\hat{A}$’ acts on the basis vectors in the following way:

$$\hat{A}|u_1\rangle = 2|u_1\rangle, \quad \hat{A}|u_2\rangle = 3|u_1\rangle - i|u_3\rangle, \quad \hat{A}|u_3\rangle = -|u_2\rangle,$$
Hence, \( A_{ij} = \begin{bmatrix} \langle u_1 | \hat{A} | u_1 \rangle & \langle u_1 | \hat{A} | u_2 \rangle & \langle u_1 | \hat{A} | u_3 \rangle \\ \langle u_2 | \hat{A} | u_1 \rangle & \langle u_2 | \hat{A} | u_2 \rangle & \langle u_2 | \hat{A} | u_3 \rangle \\ \langle u_3 | \hat{A} | u_1 \rangle & \langle u_3 | \hat{A} | u_2 \rangle & \langle u_3 | \hat{A} | u_3 \rangle \end{bmatrix} = \begin{bmatrix} \langle u_1 | 2 | u_1 \rangle & \langle u_1 | 3 | u_1 \rangle - i | u_3 \rangle & \langle u_1 | - | u_2 \rangle \\ \langle u_2 | 2 | u_1 \rangle & \langle u_2 | 3 | u_1 \rangle - i | u_3 \rangle & \langle u_2 | - | u_2 \rangle \\ \langle u_3 | 2 | u_1 \rangle & \langle u_3 | 3 | u_1 \rangle - i | u_3 \rangle & \langle u_3 | - | u_2 \rangle \end{bmatrix} \)

= \begin{bmatrix} 2 \langle u_1 | u_1 \rangle & 3 \langle u_1 | u_1 \rangle - i \langle u_1 | u_3 \rangle & -\langle u_1 | u_2 \rangle \\ 2 \langle u_2 | u_1 \rangle & 3 \langle u_2 | u_1 \rangle - i \langle u_2 | u_3 \rangle & -\langle u_2 | u_2 \rangle \\ 2 \langle u_3 | u_1 \rangle & 3 \langle u_3 | u_1 \rangle - i \langle u_3 | u_3 \rangle & -\langle u_3 | u_2 \rangle \end{bmatrix} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 0 & -1 \\ 0 & -i & 0 \end{bmatrix}

Any vector \(|v\rangle\) in a vector space can be written as a linear combination of the basis vectors \(|u_i\rangle\):

\[ |v\rangle = \sum_i v_i |u_i\rangle \rightarrow \langle u_j | v\rangle = v_i \rightarrow |v\rangle = \sum_i |u_i\rangle \langle u_i | v\rangle = \left( \sum_i |u_i\rangle \langle u_i | \right) |v\rangle \rightarrow \sum_i |u_i\rangle \langle u_i | = 1, \]

Note that

\[ \langle u_i | \hat{T} | u_j \rangle \rightarrow \sum_j \langle u_i | \hat{T} | u_j \rangle \langle u_j | v\rangle = \sum_j T_{ij} v_j, \quad \langle u_i | \hat{T} | u_j \rangle = T_{ij}, \quad \langle u_j | v\rangle = v_j \]

Hence, \( \langle u_i | \hat{T} | u_j \rangle \) is simply a square matrix acting on the components of the column vector.

**11.7.2 Multiple Operators**

A vector can be operated on by more than one operator i.e.

\[ \hat{S} \hat{T} | u \rangle = \hat{S}[\hat{T} | u \rangle] \]

Hence,

\[ \langle u_i | \hat{S} \hat{T} | u_j \rangle \rightarrow \langle u_j | v \rangle = \sum_r \langle u_r | \hat{T} | u_j \rangle \rightarrow \langle u_i | \hat{S} \hat{T} | u_j \rangle = \sum_r \langle u_i | \hat{S} | u_r \rangle \langle u_r | \hat{T} | u_j \rangle \rightarrow \sum_r S_{ir} T_{rj} \]

\[ = (ST)_{ij} \]
11.7.3 Hermitian and Unitary Operators

A Hermitian operator $\hat{A}$ has the property $\hat{A} = \hat{A}^\dagger$. In other words, $\langle a|\hat{A}|b \rangle = \langle b|\hat{A}^\dagger|a \rangle$.

Example: Let

$$\hat{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & i \\ 0 & -i & 1 \end{bmatrix} \rightarrow \hat{A}^T = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & -i \\ 0 & i & 1 \end{bmatrix}$$

Since $\hat{A} = \hat{A}^\dagger$, $\hat{A}$ is Hermitian.

Recall that a linear operator $\hat{L}$ acts on a vector $|\psi\rangle$ producing another vector i.e. $\hat{L}|\psi\rangle = |\phi\rangle$. If the vector is a column vector, it is called a 'ket'. By the definition of a 'bra' vector (row vector), $\langle\psi|\hat{L}^\dagger = \langle w|$. If $\langle\psi|\hat{L}^\dagger \hat{L} |\psi\rangle = \langle w|w \rangle = \langle \psi|\psi \rangle$, then $\hat{L}^\dagger \hat{L} = I$, the identity matrix. Note that $\hat{L}^\dagger \hat{L} = I \rightarrow L^\dagger L = L^{-1}$. Such operators are called 'unitary', often designated $\hat{U}$.

Note that inner products are invariant under unitary operations.

11.8 Eigenvalues and Eigenvectors

In general, if an operator $\hat{A}$ acts on a vector $|\psi\rangle$ i.e. $\hat{A}|\psi\rangle = |\phi\rangle$, $|\phi\rangle$ will not be a multiple of $|\psi\rangle$. If it is, then

$$\hat{A}|\psi\rangle = |\phi\rangle = \lambda |\psi\rangle, \quad \lambda \in \mathbb{C},$$

where the scalar $\lambda$ is called the 'eigenvalue' of $\hat{A}$ and $|\psi\rangle$ is the 'eigenvector' of $\hat{A}$.

11.8.1 The Eigenvalues and Eigenvectors of Hermitian Operators

Hermitian operators always have eigenvalues and eigenvectors with the following properties:

1. All eigenvalues of Hermitian operators are real numbers

To see this, let $\hat{A}|\psi\rangle = \lambda |\psi\rangle$, $||\psi\rangle|| = 1$, then $\langle \psi|\hat{A}|\psi\rangle = \lambda \langle \psi|\psi\rangle = \lambda$. In general, $\langle a|b \rangle = \langle b|a \rangle^\dagger \rightarrow \langle v|\hat{A}^\dagger |v\rangle = \lambda^\dagger$. But $\hat{A} = \hat{A}^\dagger \rightarrow \lambda = \lambda^\dagger$, which can only happen if $\lambda$ is real.

2. If $\lambda_1, \lambda_2, \lambda_1 \neq \lambda_2$ are eigenvalues of a Hermitian operator $\hat{A}$, with eigenvectors $|v_1\rangle$ and $|v_2\rangle$ respectfully, then $|v_1\rangle, |v_2\rangle$ are orthogonal

To see this, let $\hat{A}|v_1\rangle = \lambda_1 |v_1\rangle$ and $\hat{A}|v_2\rangle = \lambda_2 |v_2\rangle$, $\lambda_1 \neq \lambda_2$. Now $\langle v_2|\hat{A}|v_1\rangle = \lambda_1 \langle v_2|v_1\rangle$. Similarly,

$$\langle v_1|\hat{A}|v_2\rangle = \lambda_2 \langle v_1|v_2\rangle, \quad \langle v_1|\hat{A}|v_2\rangle^\dagger = \lambda_2^\dagger \langle v_1|v_2\rangle^\dagger = \langle v_2|\hat{A}|v_1\rangle = \lambda_2 \langle v_2|v_1\rangle$$
Hence,
\[ \langle v_2 | H | v_1 \rangle - \langle v_2 | H | v_1 \rangle = (\lambda_1 - \lambda_2)\langle v_2 | v_1 \rangle = 0 \]

Since \( \lambda_1 \neq \lambda_2 \), then \( \langle v_2 | v_1 \rangle = 0 \) and \( |v_1\rangle \) and \( |v_2\rangle \) are orthogonal.

3. The number of distinct eigenvectors is equal to the dimension of the vector space

This follows immediately from 2. Since \( |v_i\rangle, |v_j\rangle \in V, i \neq j \) are orthogonal, this implies that all the eigenvectors are independent, and hence, basis vectors, which determine the dimension of the vector space.

Example: Let

\[ H = \begin{bmatrix} h_{11} & 0 \\ 0 & h_{22} \end{bmatrix}, \quad |v_1\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad |v_2\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} h_{11} & 0 \\ 0 & h_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = h_{11} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \]

Hence, \( h_{11} \) is an eigenvalue associated with the eigenvector \( |v_1\rangle \) and \( h_{22} \) is an eigenvalue associated with the eigenvector \( |v_2\rangle \).

11.8.2 The Expected Value of an Operator

Now consider the expression \( \langle \Psi | \hat{R} | \Psi \rangle \), where \( \hat{R} \) is any linear operator and \( |\Psi\rangle \) is a vector. If \( \hat{R}|u_i\rangle = \lambda_i |u_i\rangle \), where the \( \lambda_i \)'s are the eigenvalues and the \( |u_i\rangle \)'s are the orthonormal basis vectors or eigenvectors of \( \hat{R} \), then

\[ \langle \Psi | \hat{R} | \Psi \rangle = \sum_i \langle \Psi | \hat{R} | u_i \rangle \langle u_i | \Psi \rangle = \sum_i \lambda_i \langle \Psi | u_i \rangle \langle u_i | \Psi \rangle = \sum_i \lambda_i P_i = \langle \hat{R} \rangle, \]

since \( P_i = \langle \Psi | u_i \rangle \langle u_i | \Psi \rangle \). The quantity \( \langle \hat{R} \rangle \) is called the ‘average’ or the ‘expected value’ of \( \hat{R} \).
11.9 Postulates of Quantum Mechanics

The postulates of quantum mechanics spring from the seed of experimental physics. There are roughly five postulates:

**Postulate 1**: States of systems are described by vectors, one vector per state

**Postulate 2**: An observable is represented by a Hermitian operator

**Postulate 3**: The possible values of a measurement are the eigenvalues of Hermitian operators

**Postulate 4**: A definite value of a measurement on a system can be obtained only if the states of the system are represented by the eigenvectors of Hermitian operators

**Postulate 5**: If a system is in an arbitrary state \(|\psi\rangle\) and if the values for any observable \(\hat{H}\) consist of the set \(\{\lambda_i\}\) of eigenvalues with eigenvectors \(\{|\lambda_i\rangle\}\), then if \(|\psi\rangle\) is measured at any point in time, the probability \(P(\lambda_i)\) that the measurement will result in \(\lambda_i\) is \(P(\lambda_i) = \langle \lambda_i | \psi \rangle \langle \psi | \lambda_i \rangle\) (no summation on \(i\))

11.9.1 Measurements in Quantum Mechanics

There are a couple of measurements that are almost always of interest in quantum mechanics. The first is the ‘average value’ of a measurement. Secondly, suppose a system is prepared such that \(\hat{R} |\psi\rangle = \alpha |\psi\rangle\). In other words \(\hat{R}\) operates on \(|\psi\rangle\) giving eigenvalue \(\alpha\). Furthermore, suppose, in the same system, \(\hat{L} |\phi\rangle = \beta |\phi\rangle\). What is the probability that if \(\hat{L}\) is measured, given that the system is in state \(|\psi\rangle\), the result will be \(\beta\)? The answer to this question is always the same:

\[
P_{\psi\phi} = \langle \psi | \phi \rangle \langle \phi | \psi \rangle
\]

where \(P_{\psi\phi}\) is the probability that if the system is in state \(|\psi\rangle\) and \(\hat{L}\) is measured, the system will be found in state \(|\phi\rangle\).

In classical mechanics, states are the same or completely different. A flip of a coin results a head or a tail. A toss of a die gives a 1,2,3 \(\ldots\) etc. All classical states, being distinct, are said to have an ‘excluded middle’. Quantum states are different. Two states \(|a\rangle\) and \(|b\rangle\) are distinct if \(\langle a | b \rangle = 0\). States in quantum mechanics are generally normalized i.e. \(\langle a | a \rangle = 1\). If \(\langle a | b \rangle \neq 0 \lor \langle a | b \rangle \neq 1\), this reflects the degree of non-distinctness of the two quantum states.
11.9.1.1 Gauge Transformations in Quantum Mechanics

A complex number can be written in different forms i.e.

\[ z = a + ib = r \cos \theta + ir \sin \theta = re^{i\theta} \]

If \( r = 1 \), then \( z(\theta) = e^{i\theta} \). Note that \( z^\dagger(\theta) = e^{-i\theta} \). The points representing the function \( e^{i\theta} \) lie on a complex unit circle; the expression \( e^{i\theta} \) is called a 'phase'. Let \( |b\rangle = e^{i\theta}|a\rangle \). Note that \( |b\rangle \neq |a\rangle \). But \( \langle b| = \langle a|e^{-i\theta} \). Hence,

\[ \langle b|b\rangle = \langle a|e^{-i\theta}e^{i\theta}|a\rangle = e^{-i\theta}e^{i\theta}\langle a|a\rangle = \langle a|a\rangle = 1 \]

Moreover,

\[ \langle b|\vec{R}|b\rangle = \langle a|e^{-i\theta}\vec{R}e^{i\theta}|a\rangle = e^{-i\theta}e^{i\theta}\langle a|\vec{R}|a\rangle = \langle a|\vec{R}|a\rangle \]

In other words, the quantum mechanical properties of a state are indistinguishable from that same state multiplied by a phase. Multiplying a state by a phase is called a 'gauge transformation'. The quantum mechanical properties of a system are invariant under gauge transformations.

11.9.2 The Time Evolution of Quantum Systems

Quantum mechanical systems can evolve over time. The quantum analog to the classical 'Liouville's theorem' is that inner products are conserved i.e. if \(|a\rangle \xrightarrow{t} |a'\rangle \) and \(|b\rangle \xrightarrow{t} |b'\rangle \), then \( \langle a|b\rangle = \langle a'|b'\rangle \). Therefore, in spatial terms, angles between vectors remain fixed over time. It is sometimes said that, in quantum mechanical systems, information is conserved.

The operators that preserve inner products are called 'unitary' operators. To see this, let \( \hat{U}|a\rangle = |a'\rangle \) and \( \hat{U}|b\rangle = |b'\rangle \), then

\[ \langle a|b\rangle = \langle a|\hat{U}^\dagger \hat{U}|b\rangle = \langle a'|b'\rangle, \quad \hat{U}^\dagger \hat{U} = I \]

Example: Suppose there is a particle that moves along a horizontal line. In classical physics, the motion of the particle is described by a real function \( x(t) \), expressing where the particle is located at time 't'. In quantum mechanics, the state of a particle is described by a complex function \( '\Psi(x) = \langle x|\Psi' \). If \( \Psi, \Phi \in \mathcal{V} \), then

\[ \langle \Psi|\Phi \rangle = \int \Psi^\dagger \Phi \, dx \]

If an operator \( '\hat{X}' \) is defined as \( \hat{X}|\Psi \rangle = x|\Psi \rangle \), then it has the effect of simply multiplying the function \( '\Psi(x)' \) by its location on the line. According to the postulates of quantum mechanics, \( \hat{X} \) must be Hermitian. Note that
\[ \langle \Psi | \hat{X} | \Psi \rangle = \int \psi^\dagger x \psi \, dx = \int x \psi^\dagger \psi \, dx \]

Since \( x, \psi^\dagger \psi \in \mathbb{R} \), \( \hat{X} \) is Hermitian.

### 11.9.2.1 The Dirac Delta Function

By the definition of an eigenvalue,

\[ \hat{X} | \Psi \rangle = \lambda | \Psi \rangle \rightarrow x | \Psi \rangle = \lambda | \Psi \rangle \rightarrow (x - \lambda) | \Psi \rangle = | 0 \rangle \]

In order to satisfy the equation \( (x - \lambda) | \Psi \rangle = | 0 \rangle \), \( \Psi(x) \) must equal zero for all \( x \) except at the point \( x = \lambda \). The solution to this problem was invented by Dirac. He set \( \Psi(x) = \delta(x - x_0) \), where

\[ \delta(x - x_0) = \begin{cases} 1/\varepsilon, & \text{if } x = x_0, \\ 0, & \text{if } x \neq x_0, \end{cases} \quad \varepsilon \ll 1 \]

The expression \( \delta(t - t_0) \) is called the ‘Dirac delta function’ (see fig. 11.9.2.1-1).

![Figure 11.9.2.1-1](image)

The Dirac delta function not a function in the mathematical sense, more like an invention, sometimes referred to as a ‘generalized function’. Looking at fig. 11.9.2.1-1,

\[ \int_{t_0 - \varepsilon/2}^{t_0 + \varepsilon/2} \delta(t) \, dt = \int_{t_0 - \varepsilon/2}^{t_0 + \varepsilon/2} \frac{1}{\varepsilon} \, dt = \frac{t_0 + \varepsilon}{2} - \left( \frac{t_0 - \varepsilon}{2} \right) = \frac{\varepsilon}{\varepsilon} = 1 \]

The function \( \Psi(x) = \delta(x - \lambda) \) becomes the eigenvector of \( \hat{X} \) with eigenvalue \( \lambda \). There are an infinite number of eigenvectors, since there are an infinite number of points on a line. The eigenvectors \( | \delta(x - \lambda) \rangle \) are orthogonal (not normalized), since

\[ \langle \delta(x - \lambda) | \delta(x - \lambda') \rangle = \int \delta(x - \lambda)^\dagger \delta(x - \lambda') \, dx = 0, \quad \lambda \neq \lambda' \]
Note that
\[ \hat{X}|\Psi\rangle = \lambda|\delta(x - \lambda)\rangle \rightarrow \lambda \int \delta(x - \lambda) \, dx = \lambda, \quad |\lambda\rangle = |\delta(t - \lambda)\rangle \]

Hence,
\[ \langle \lambda|\Psi\rangle = \int \delta(x - \lambda)\Psi(x) \, dx = \int \delta(x - \lambda)\delta(x - \lambda) \, dx = \int \delta^2(x - \lambda) \, dx = 1 \]
since \( \int \delta(x - \lambda) \, dx = 0 \) except where \( x = \lambda \), then \( \int \delta(x - \lambda) \, dx = 1 \). Since \( x = \lambda \),
\[ \langle \Psi|x\rangle = \Psi^\dagger(x), \quad \langle x|\Psi\rangle = \Psi(x) \]

Moreover, the probability \('P_x'\) that the particle will be found at position \('x'\) is
\[ P_x = \langle \Psi|x\rangle(x|\Psi\rangle = \Psi^\dagger\Psi \]

### 11.9.3 The Generalization of Eigenvalue Equations in Quantum Mechanics

In general, if a quantum system is prepared in a state \('|x\rangle\)' and if \(|x\rangle\) is an eigenvector of an operator \('\hat{R}'\), then the system will definitely be found in state \('|x\rangle\)' i.e. \(\hat{R}|x\rangle = x|x\rangle\). If \(|x\rangle\) represents position, the system will be found at \(x\). The same can be said for other operators, for example, \(\hat{R}|p\rangle = p|p\rangle\). If the system is prepared with momentum \('p'\) and \(\hat{R}\) is the 'momentum' operator, the system will definitely be found with momentum \('p'\).

Note that
\[ \langle p|\hat{X}|x\rangle = x\langle p|x\rangle, \quad \langle x|\hat{R}|p\rangle = p\langle x|p\rangle \]

Both \(\langle p|\hat{X}|x\rangle\) and \(\langle x|\hat{R}|p\rangle\) are expected values. If \(x\) replaces \(p\) in the first expected value, then \(\langle x|\hat{R}|x\rangle = x\langle x|x\rangle = x\). The system would be found at \(x\) with probability \('1'\). Likewise,
\[ \langle x|\hat{R}|p\rangle \rightarrow \langle p|\hat{R}|p\rangle = p\langle p|p\rangle = p \]

The system would have momentum \('p'\) with probability \('1'\). Both \(\langle p|p\rangle\) and \(\langle x|x\rangle\) can be interpreted as probabilities. In the cases
\[ \langle p|\hat{X}|x\rangle = x\langle p|x\rangle, \quad \langle x|\hat{R}|p\rangle = p\langle x|p\rangle, \]

\(\langle p|x\rangle\) is interpreted as: if the system is definitely at position \('x'\), what is the probability it has momentum \('p'\)? And \(\langle x|p\rangle\) can be interpreted as: if the system definitely has momentum \('p'\), what is the probability that it is at position \('x'\). Hence, \(\langle p|x\rangle\langle x|p\rangle\) is the probability that the system has both momentum \('p'\) and is at position \('x'\).
11.9.3.1 The Momentum Operator

Position is not the only observable that can be measured. Different observables are described by different Hermitian operators. Consider

\[
\frac{\partial \Psi(x)}{\partial x}
\]

The operator \( \hat{H} = \partial / \partial x \) is linear, since

\[
\frac{\partial}{\partial x} (\alpha \Psi + \beta \Phi) = \alpha \frac{\partial \Psi}{\partial x} + \beta \frac{\partial \Phi}{\partial x}, \quad \alpha, \beta \in \mathbb{C},
\]

but is not Hermitian. It is actually an anti-Hermitian operator, since \( \langle a | \hat{H} | b \rangle = -\langle b | \hat{H}^\dagger | a \rangle \). But \( \hat{H} \) can be made Hermitian by multiplying it by \(-i\). To see this, note that

\[
\langle \Psi | -i\hat{H} | \Psi \rangle = \int \Psi^\dagger (-i \frac{\partial \Psi}{\partial x}) \, dx = -i \int \Psi^\dagger \frac{\partial \Psi}{\partial x} \, dx
\]

Integrating the expression on the right above by parts gives

\[
i \int \frac{\partial \Psi^\dagger}{\partial x} \Psi \, dx
\]

Taking complex conjugate leaves

\[
-i \int \Psi^\dagger \frac{\partial \Psi}{\partial x} \, dx
\]

Since \( \langle \Psi | -i\hat{H} | \Psi \rangle \) is its own complex conjugate, \(-i\hat{H}\) is Hermitian.

Let \( \hat{K} = -i \partial / \partial x \), then the eigenvalue equation becomes

\[
\hat{K}\Psi(x) = -i \frac{\partial \Psi(x)}{\partial x} = k\Psi(x),
\]

where \( k \) is the eigenvalue. The solution to this equation is \( \Psi(x) \propto e^{ikx} \). To see this, note that

\[
\frac{\partial \Psi}{\partial x} = ik\Psi \rightarrow -i \frac{\partial \Psi}{\partial x} = k\Psi
\]

Moreover \( \Psi(x) = \alpha e^{ikx}, \ \alpha \in \mathbb{C} \) is also a solution and

\[
\Psi^\dagger \Psi = \alpha^\dagger e^{-ikx} \alpha e^{ikx} = \alpha^\dagger \alpha e^{-ikx} e^{ikx} = \alpha^\dagger \alpha
\]
So \( \alpha^\dagger \alpha \) is a probability. Note that

\[
\Psi(x) = e^{ikx} = \alpha \cos(kx) + \alpha i \sin(kx)
\]

Unlike the Dirac delta function, which is localized at a point, \( \Psi(x) \) is a wave function that spreads out over the entire line.

The period of \( \sin(kx) \), \( \cos(kx) \) is \( 2\pi \). If \( x \) must move a distance ‘\( \lambda \)’ before the wave repeats itself, then \( k\lambda = 2\pi \Rightarrow \lambda = 2\pi/k \). Since \( p = h/\lambda \), where \( p \) is the momentum of a particle, then \( p = h/\lambda = \hbar k \), \( \hbar = h/2\pi \). Hence, the momentum of a particle is proportional to the eigenvalue ‘\( k \)’. Since

\[
\Psi^\dagger \Psi = \alpha^\dagger e^{-ikx} \alpha e^{ikx} = \alpha^\dagger \alpha = \alpha^\dagger \alpha,
\]

then \( \alpha^\dagger \alpha \) is the probability of finding the particle at position ‘\( x \)’, given that the momentum is \( k \). If \( \alpha = 1 \), then \( \Psi^\dagger \Psi = 1 \). In this case, the particle is definitely somewhere on the line, but no information is given as to where on the line. The position spreads out over the whole line. The upshot is that, if the momentum of the particle is known, the only information about the position of the particle is that it is somewhere on the line.

11.9.3.1 The Quantization of Momentum

The function ‘\( \Psi(x) = e^{ikx} \)’ is very different from the Dirac delta function ‘\( \delta(x - x_0) \)’; \( \Psi(x) = e^{ikx} \) lies on a circle in the complex plane, whereas \( \delta(x - x_0) \) is located at a point on the line. There is no obvious way of knowing \( \Psi(x) \) if the position of the particle is known and vice versa. The function ‘\( \Psi(x) \)’ rotates in a plane perpendicular to the line. The eigenvalues of \( \hat{K} \) are completely independent of the eigenvalues of \( \hat{X} \) (see fig. 11.9.3.1-1). In other words, no eigenvalues of momentum are eigenvalues of position.

![Figure 11.9.3.1.1-1](image)

Example: Suppose a particle is moving along a circular path (see fig. 11.9.3.1.1-2). The circle can be laid flat so that at one end \( x = 0 \) and, at the other end, \( x = 2\pi r \), the
circumference of the circle. The functions that define the states in this space are ‘periodic’ i.e. \( \Psi(x) = \Psi(x \pm 2\pi r) \). Note that

\[
\alpha \Psi(x) + \beta \Phi(x) = \alpha \Psi(x \pm 2\pi r) + \beta \Phi(x \pm 2\pi r)
\]

Figure 11.9.3.1.1-2

Functions in the form \( \Psi(x) = \Psi(x \pm 2\pi r) \) comprise a vector space. Recall the momentum operator:

\[
\hat{K} \Psi(x) = -i \frac{\partial \Psi(x)}{\partial x} = k \Psi(x) \rightarrow \hbar \hat{K} \Psi(x) = -i \hbar \frac{\partial \Psi(x)}{\partial x} = \hbar k \Psi(x) = p \Psi(x),
\]

where \( p = \hbar k \) is the momentum. A solution to this equation is \( \Psi(x) = e^{i p x / \hbar} \). Now

\[
\int_0^{2\pi r} \Psi^\dagger \Psi \, dx = \int_0^{2\pi r} e^{-i p x / \hbar} e^{i p x / \hbar} \, dx = \int_0^{2\pi r} \, dx = 2\pi r \rightarrow \int_0^{2\pi r} \frac{\Psi^\dagger \Psi}{\sqrt{2\pi r} \sqrt{2\pi r}} \, dx = 1,
\]

where \( |\Psi\rangle = 1 / \sqrt{2\pi r} e^{i p x / \hbar} \) is normalized. If \( |\Psi_p\rangle = e^{i p x / \hbar} \) and \( |\Psi_q\rangle = e^{i q x / \hbar} \), then

\[
\langle \Psi_p | \Psi_q \rangle = \int \Psi_p^\dagger \Psi_q = 0
\]

i.e. the eigenstates are orthogonal. In this case, since \( \Psi \) is periodic,

\[
\Psi(x) = \Psi(x \pm 2\pi r) \rightarrow e^{i p (x+2\pi r) / \hbar} = e^{i p x / \hbar} e^{i p 2\pi r / \hbar} = e^{i p x / \hbar} \rightarrow e^{i p 2\pi r / \hbar} = 1
\]

The relation \( e^{i p 2\pi r / \hbar} = 1 \) is the necessary and sufficient condition to ensure that \( \Psi \) is periodic. Recall that for any exponential, \( e^{i 2\pi n} = 1 \), \( n \in \mathbb{Z} \). Therefore, in order for \( \Psi \) to be periodic,

\[
p 2\pi r / \hbar = n 2\pi \rightarrow pr / \hbar = n,
\]

278
placing a restriction on the eigenvalues of momentum. The quantities \( p = n \hbar / r \) are the only allowable values for the momentum of a particle moving along a circle of radius \( r \). This is a surprising result. Even though the particle is moving continuously in a circle, the values of its momentum are discrete. In other words, the momenta are quantized.

Classical physics allows the momentum of a particle to take on continuous values. The relationship between the quantum result and the classical result is, if \( r \to \infty \), then the differences between the momenta \( \Delta p \) go to zero. In other words, as the circles become larger and larger, the quantum result approaches the classical result.

If a particle is moving in a circle, it has an angular momentum \( L = pr \). But

\[
p = n\hbar/r \to L = n\hbar
\]

Hence, in quantum mechanical systems, angular momentum is quantized. Note that the quantization of the angular momentum does not depend on the radius of the circle. In quantum mechanical systems, angular momentum is always quantized.

### 11.9.4 The Nature of Operators in Quantum Mechanics

Suppose \( \{ |v_i \rangle \} \in V \), where \( (V, +, \cdot) \) is a vector space. Let

\[
\hat{A}|v_i\rangle = \alpha_i|v_i\rangle, \quad \hat{B}|v_i\rangle = \beta_i|v_i\rangle,
\]

where the \( \alpha_i \)'s are the eigenvalues of the operator \( \hat{A} \) and the \( \beta_i \)'s are the eigenvalues of the operator \( \hat{B} \). Note that

\[
\hat{B}\hat{A}|v_i\rangle = \alpha_i\hat{B}|v_i\rangle = \alpha_i\beta_i|v_i\rangle, \quad \hat{A}\hat{B}|v_i\rangle = \beta_i\hat{A}|v_i\rangle = \beta_i\alpha_i|v_i\rangle
\]

Hence,

\[
\hat{A}\hat{B}|v_i\rangle - \hat{B}\hat{A}|v_i\rangle = \beta_i\alpha_i|v_i\rangle - \alpha_i\beta_i|v_i\rangle = (\beta_i\alpha_i - \alpha_i\beta_i)|v_i\rangle = 0|v_i\rangle = |0\rangle \to \hat{A}\hat{B} - \hat{B}\hat{A} = 0
\]

If two observables represented by the operators \( \hat{A}, \hat{B} \) have a common set of eigenvectors, then \( \hat{A}\hat{B} - \hat{B}\hat{A} = 0 \). The two observables can be measured simultaneously. Moreover, if

\[
\hat{A}|v_i\rangle = \alpha_i|v_i\rangle \to \hat{A}|v_i\rangle - \alpha_i|v_i\rangle = |0\rangle, \quad \hat{B}|u_i\rangle = \alpha_i|u_i\rangle \to \hat{B}|u_i\rangle - \alpha_i|u_i\rangle = |0\rangle \to \hat{A}|v_i\rangle - \alpha_i|v_i\rangle = \hat{B}|u_i\rangle - \alpha_i|u_i\rangle
\]

So,

\[
\hat{B}|u_i\rangle = \hat{A}|v_i\rangle + \alpha_i|u_i\rangle - \alpha_i|v_i\rangle \to \hat{A}\hat{B}|u_i\rangle = \hat{A}\alpha_i|v_i\rangle + \alpha_i\hat{A}|u_i\rangle - \alpha_i\hat{A}|v_i\rangle \\
\to \hat{A}\hat{B}|u_i\rangle = \alpha_i\hat{A}|v_i\rangle + \alpha_i\hat{A}|u_i\rangle - \alpha_i\hat{A}|v_i\rangle = \alpha_i\hat{A}|u_i\rangle \to \hat{A}\hat{B} = \alpha_i\hat{A}
\]

279
Similarly, $\hat{B}\hat{A}|v_i\rangle = \alpha_i \hat{B}|v_i\rangle \rightarrow \hat{B}\hat{A} = \alpha_i \hat{B}$. Hence,

$$\hat{A}\hat{B} - \hat{B}\hat{A} = \alpha_i \hat{A} - \alpha_i \hat{B} = \alpha_i (\hat{A} - \hat{B})$$

If $\hat{A}\hat{B} - \hat{B}\hat{A} = 0$ and the operators ‘$\hat{A}$’ and ‘$\hat{B}$’ have the same eigenvalues, then $\hat{A} = \hat{B}$.

Hence, if $\hat{A}\hat{B} - \hat{B}\hat{A} \neq 0$, the operators ‘$\hat{A}$’ and ‘$\hat{B}$’ do not measure the same observable or, more importantly, cannot have common states.

Example: If ‘$\hat{X}$’ is the position and ‘$\hat{K}$’ the momentum operator, then

$$\hbar \hat{X}\hat{K} - \hbar \hat{K}\hat{X} = \hbar \left[ \hat{X} \left( -i \frac{\partial}{\partial x} \right) - \left( -i \frac{\partial}{\partial x} \right) \hat{X} \right]$$

If $\hat{X}\hbar \hat{K} - \hbar \hat{K}\hat{X}$ operates on a complex function ‘$\Psi(x)$’, then

$$[\hat{X}\hbar \hat{K} - \hbar \hat{K}\hat{X}]\Psi(x) = \hbar \left[ x \left( -i \frac{\partial}{\partial x} \right) - \left( -i \frac{\partial}{\partial x} \right) x \right] \Psi(x) = x \left( -i \hbar \frac{\partial \Psi(x)}{\partial x} \right) + \left( i \hbar \frac{\partial (x\Psi(x))}{\partial x} \right)$$

$$= -i \hbar x \frac{\partial \Psi(x)}{\partial x} + i \hbar \Psi(x) + i \hbar x \frac{\partial \Psi(x)}{\partial x} = i \hbar \Psi(x) \rightarrow [\hat{X}\hbar \hat{K} - \hbar \hat{K}\hat{X}] = [\hat{X},\hat{K}] = i$$

Hence, there are no common eigenvectors between the position operator ‘$\hat{X}$’ and the momentum operator ‘$\hat{K}$’. It is not possible to measure position and momentum simultaneously.

### 11.10 The Two Slot Interference Experiment (Revisited)

Returning to the two slot interference experiment, suppose $\Psi = e^{ipy/\hbar}$. If only one slot is open, because photon strikes against the absorber will be closer together at points farther away from the point directly across from the open slot, the function ‘$\Psi = e^{ipy/\hbar}$’ is a poor approximation of how the strikes against the absorber are distributed. But if the interval ‘$\Delta y$’ is small enough, $\Psi = e^{i\Delta y/\hbar}$ will be a good approximation of the wave distribution (see fig. 11.10-1).

![Figure 11.10-1](image)

Suppose $\Psi_1 = e^{ip\Delta y/\hbar}$ describes the wave passing through the upper slot and $\Psi_2 = e^{iq\Delta y/\hbar}$ the wave passing through the lower slot (see fig. 11.10-1). If only the upper slot
is open, \( \Psi_1 \Psi_1^\dagger = e^{ip\Delta y/h} e^{-ip\Delta y/h} = 1 \) i.e. the probability that the electron is dislodged in the interval ‘\( \Delta y \)’ is 1. It does not specify where in the interval ‘\( \Delta y \)’ the electron is dislodged.

If both slots are open, then

\[
\Psi = \Psi_1 + \Psi_2 = e^{ip\Delta y/h} + e^{iq\Delta y/h}
\]

The probability that an electron is dislodged from the absorber in the interval ‘\( \Delta y \)’ is

\[
\psi \psi^\dagger = (e^{ip\Delta y/h} + e^{iq\Delta y/h})(e^{-ip\Delta y/h} + e^{-iq\Delta y/h}) = 1 + 1 + e^{i((p-q)/h)\Delta y} + e^{i((q-p)/h)\Delta y} = 2 + 2 \cos\left(\left(\frac{p-q}{\hbar}\right) \Delta y\right).
\]

where the last equation on the right follows from a trigonometric identity. The equation above is not quite correct, since \( \Psi \) has not been normalized, but this is not important in this case. There are points on the absorber where

\[
\cos\left(\left(\frac{p-q}{\hbar}\right) \Delta y\right) = -1 \rightarrow \psi \psi^\dagger = 0
\]

With both slots open, there are places along the absorber where the probability of an electron being dislodged is zero. This is a different result than if only one slot is open, where the electron can be dislodged from any position in the interval ‘\( \Delta y \)’. Moreover, the “two slots are open” case cannot be described by classical mechanics, which does not allow a superposition of states.

11.10.1 Position and Momentum as Fourier Transforms

Recall that the position operator ‘\( \hat{X} \)’ acting on a state vector ‘\( |\Psi(x)\rangle \)’ gives \( \hat{X} |\Psi(x)\rangle = x |\Psi(x)\rangle \) and that \( \langle x |\Psi(x)\rangle = \Psi(x) \). On the other hand, \( \langle x |k\rangle = e^{-ikx/\sqrt{2\pi}} \), \( k = p/\hbar \), where, for convenience, the radius ‘\( r \)’ of the circle has been set to 1. Moreover, let

\[
\bar{\Psi}(k) = \int \langle k|x\rangle \langle x|\Psi \rangle dx = \frac{1}{\sqrt{2\pi}} \int e^{-ikx} \Psi(x) dx
\]

Recall the definition of a Fourier transform:

\[
F(\alpha) = \int_{-\infty}^{\infty} f(u) e^{-i\alpha u} du \rightarrow f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\alpha) e^{i\alpha x} d\alpha
\]

The function ‘\( F(\alpha) \)’ is called the ‘Fourier transform’ of \( f(x) \), while

\[
f(x) \equiv \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\alpha) e^{i\alpha x} d\alpha
\]

is called the ‘inverse Fourier transform’. The function ‘\( F(\alpha) \)’ is often called the ‘\( \alpha \) – representation’ and ‘\( f(x) \)’ is called the ‘\( x \) – representation’ (see Book II: Chapter 16, Sec. 16.8.2). By letting \( \alpha = k \) and \( F = \bar{\Psi} \), \( f = \Psi \) and \( u = x \), then
\[ \Psi(k) = \int \langle k|x\rangle\langle x|\Psi \rangle \, dx = \frac{1}{\sqrt{2\pi}} \int e^{-ikx} \Psi(x) \, dx \]

The position and momentum eigenvectors are related by a Fourier transformation. To see this, note that

\[ \Psi(x) = \int \langle x|k\rangle\langle k|\Psi \rangle \, dk = \frac{1}{\sqrt{2\pi}} \int e^{ikx} \bar{\Psi}(k) \, dk, \quad x \rightarrow k, \quad k \rightarrow x \]

The following relationships hold for position and momentum operators:

1. \( \hat{X}|\Psi(x)\rangle = x|\Psi(x)\rangle \)
2. \( \hat{P}|\Psi(x)\rangle = -i \frac{\partial \Psi(x)}{\partial x} \)
3. \( \hat{P}|\bar{\Psi}(k)\rangle = k|\bar{\Psi}(k)\rangle \)
4. \( \hat{X}|\bar{\Psi}(k)\rangle = i \frac{\partial \bar{\Psi}(k)}{\partial k} \)

### 11.11 Light Polarization and Quantum Operators

Classical physics describes light as an electromagnetic plane wave. If the electric field \(|E\rangle\) oscillates in the vertical plane, then the magnetic field \(|B\rangle\) oscillates in the horizontal plane. Electric and the magnetic fields are orthogonal. The ‘polarization’ of light is defined as the plane in which the electric field is oscillating. If the electric field is oscillating vertically, then the light is ‘vertically polarized’.

Un-polarized light, where the electric field oscillates in many directions, can be polarized by a ‘polarizer’. Light that passes through a vertical polarizer will have its electric field ‘vertically polarized’. A horizontal polarizer will only permit light through polarized in the horizontal direction (see fig. 11.11-1).

![Figure 11.11-1](image)

If un-polarized light passes through a vertical polarizer and then a horizontal polarizer, no light will pass through both polarizers. However, quantum mechanics conceives light as a ‘photon’, a wave packet that acts much like a particle. Classical physics predicts that, if the photon encounters a polarizer, part of the photon would pass through and part would not. This prediction, however, does not agree with what happens experimentally, which shows that either the photon passes through the polarizer or it
does not. There is no such thing as a partial photon or partial photon energy. The photon is indivisible.

Quantum mechanically, the physics of photons is described probabilistically. A single photon that passes through a vertical polarizer is definitely polarized vertically. If the photon encounters a second vertical polarizer, the probability it will pass through is 1. If a vertical polarized photon encounters horizontal polarizer, the probability it will pass through is zero.

A polarizer is a device that measures the polarity of light, suggesting that, in quantum mechanics, the polarizer can be represented by an operator and the “state of the light” represented by the direction of polarization.

11.11.1 Vertically and Horizontally Polarized Light

For instance, let $|0^\circ\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ represent light polarized vertically and $|90^\circ\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ represent light polarized horizontally. The two states represent an orthonormal basis of a two dimensional vector space. Note that $\langle 0^\circ | 90^\circ \rangle = 0$ and $\langle 0^\circ | 0^\circ \rangle = \langle 90^\circ | 90^\circ \rangle = 1$. Define a polarization operator $\hat{P}_\theta$ such that if $\hat{P}_\theta$ acts on $|0^\circ\rangle$, it produces eigenvalue ‘1’ and if $\hat{P}_\theta$ acts on $|90^\circ\rangle$, it produces eigenvalue ‘−1’. In other words, the eigenvector equations become

$$\hat{P}_\theta |0^\circ\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{P}_\theta |90^\circ\rangle = -\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Note that $\hat{P}_\theta$ measures horizontal or vertical polarization. Since $\hat{P}_\theta$ is a linear operator, it has a square matrix representation:

$$\hat{P}_\theta = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \rightarrow \hat{P}_\theta |0^\circ\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{P}_\theta |90^\circ\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

And

$$\hat{P}_\theta |90^\circ\rangle = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} = -\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Suppose that the light is polarized in some other direction, say 45°, represented by $|45^\circ\rangle$. Since the state vectors $|0^\circ\rangle$ and $|90^\circ\rangle$ are basis vectors, $|45^\circ\rangle$ can be represented by a linear combination of the basis vectors i.e.

$$|45^\circ\rangle = \alpha|0^\circ\rangle + \beta|90^\circ\rangle, \quad \alpha, \beta \in \mathbb{C}$$

By direct calculation, $\alpha = \beta = 1/\sqrt{2}$. To see this,

$$\langle 0^\circ | \alpha^\dagger |45^\circ\rangle = \langle 0^\circ | \alpha^\dagger \alpha |0^\circ\rangle + \langle 0^\circ | \alpha^\dagger \beta |90^\circ\rangle = \alpha \alpha^\dagger \langle 0^\circ | 0^\circ \rangle + \alpha^\dagger \beta \langle 0^\circ | 90^\circ \rangle = \left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2}$$
Similarly, \( \langle 90^\circ | \beta^\dagger | 45^\circ \rangle = 1/2 \). Note that \( \alpha \alpha^\dagger + \beta \beta^\dagger = 1/2 + 1/2 = 1 \) and this implies that

\[
|45^\circ \rangle = \frac{1}{\sqrt{2}} |0^\circ \rangle + \frac{1}{\sqrt{2}} |90^\circ \rangle = \begin{bmatrix} 1 / \sqrt{2} \\ 1 / \sqrt{2} \end{bmatrix} \rightarrow | -45^\circ \rangle = \frac{1}{\sqrt{2}} |0^\circ \rangle - \frac{1}{\sqrt{2}} |90^\circ \rangle = \begin{bmatrix} 1 \\ -1 / \sqrt{2} \end{bmatrix}
\]

By direct calculation, \( \langle -45^\circ | 45^\circ \rangle = 0 \) and \( \langle -45^\circ | -45^\circ \rangle = \langle 45^\circ | 45^\circ \rangle = 1 \). The probability that \( 45^\circ \) polarized light passes through a horizontal polarizer is

\[
\langle 90^\circ | 45^\circ \rangle \langle 45^\circ | 90^\circ \rangle = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 / \sqrt{2} \\ 1 / \sqrt{2} \end{bmatrix}^2 = \left( \frac{1}{\sqrt{2}} \right)^2 = \frac{1}{2}
\]

An operator ‘\( \mathbb{P}_\bigcirc \)’ can be found such that if \( \mathbb{P}_\bigcirc \) acts on \( | 45^\circ \rangle \), then \( \mathbb{P}_\bigcirc \) produces eigenvalue ‘1’ and if \( \mathbb{P}_\bigcirc \) acts on \( | -45^\circ \rangle \), then \( \mathbb{P}_\bigcirc \) produces eigenvalue ‘−1’ i.e.

\[
\mathbb{P}_\bigcirc = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \rightarrow \mathbb{P}_\bigcirc | 45^\circ \rangle = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 / \sqrt{2} \\ 1 / \sqrt{2} \end{bmatrix} = \frac{1}{\sqrt{2}} = | 45^\circ \rangle,
\]

\[
\mathbb{P}_\bigcirc | -45^\circ \rangle = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 / \sqrt{2} \\ -1 / \sqrt{2} \end{bmatrix} = -\begin{bmatrix} 1 / \sqrt{2} \\ -1 / \sqrt{2} \end{bmatrix} = -| -45^\circ \rangle
\]

The observables represented by the two operators ‘\( \mathbb{P}_\bigcirc \)’ and ‘\( \mathbb{P}_\bigcirc \)’ are incompatible, having no common eigenstates, and hence, cannot be measured simultaneously.

### 11.11.2 The Quantization of Polarization

If the polarizer is set to some arbitrary direction relative to the horizontal polarizer and if the two states are labeled ‘\( | \theta \rangle \)’ and ‘\( | \theta + \pi / 2 \rangle \)’, where \( \theta \) is the polarization angle, then

\[
| \theta \rangle = \cos \theta | 0^\circ \rangle + \sin \theta | 90^\circ \rangle = \cos \theta \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \sin \theta \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix},
\]

\[
| \theta + \pi / 2 \rangle = \cos(\theta + \pi / 2) | 0^\circ \rangle + \sin(\theta + \pi / 2) | 90^\circ \rangle = -\sin \theta \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \cos \theta \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

\[= \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \]
Note that
\[ \langle \theta + \pi/2 | \theta \rangle = -\sin \theta \cos \theta + \sin \theta \cos \theta = 0, \]
\[ \langle \theta | \theta \rangle = \langle \theta + \pi/2 | \theta + \pi/2 \rangle = \cos^2 \theta + \sin^2 \theta = 1 \]

Hence, \(|\theta\rangle\) and \(|\theta + \pi/2\rangle\) represent orthonormal basis vectors in a two dimensional vector space, and therefore, can represent quantum mechanical states.

The probability that a photon polarized in the \(\theta\)-direction passes through a polarizer polarized in the vertical direction is
\[ \langle 0^\circ | \theta \rangle \langle \theta | 0^\circ \rangle = \left( \begin{array}{c} 1 \\ 0 \end{array} \right) \left[ \begin{array}{c} \cos \theta \\ \sin \theta \end{array} \right] = \cos^2 \theta \]

Similarly,
\[ \langle 90^\circ | \theta \rangle \langle \theta | 90^\circ \rangle = \left( \begin{array}{c} 0 \\ 1 \end{array} \right) \left[ \begin{array}{c} \cos \theta \\ \sin \theta \end{array} \right] = \sin^2 \theta \]

In general, the probability that a photon polarized in the \(\alpha\)-direction passes through a polarizer polarized in the \(\beta\)-direction, where \(\alpha, \beta\) are arbitrary angles with the horizontal, is \(\langle \beta | \alpha \rangle \langle \alpha | \beta \rangle = \cos^2 (\alpha - \beta)\). To see this, note that
\[ |\alpha\rangle = \left[ \begin{array}{c} \cos \alpha \\ \sin \alpha \end{array} \right] \rightarrow \langle \beta | \alpha \rangle = \left[ \begin{array}{c} \cos \beta \\ \sin \beta \end{array} \right] \left[ \begin{array}{c} \cos \alpha \\ \sin \alpha \end{array} \right] = \cos \beta \cos \alpha + \sin \beta \sin \alpha = \cos (\alpha - \beta) \]
\[ \rightarrow \langle \beta | \alpha \rangle \langle \alpha | \beta \rangle = \cos^2 (\alpha - \beta), \]

The probability that \(\alpha\)-polarized light passes through a \((\alpha + \pi/2)\)-polarizer is zero.

Suppose a vertical polarizer is placed between an \(\alpha\)-polarizer and a \((\alpha + \pi/2)\)-polarizer. The probability that the \(\alpha\)-polarized light passes through the vertical polarizer is \(\cos^2 \alpha\). But if the photon passes through the vertical polarizer, it is vertically polarized. Hence, the probability the photon passes through the \((\alpha + \pi/2)\)-polarizer is \(\sin^2 \alpha\). The probability the photon passes through both the vertical and \(\beta\)-polarizer is \(\cos^2 \alpha \sin^2 \alpha\). By contrast, there is zero probability the photon passes through the \((\alpha + \pi/2)\)-polarizer if the vertical polarizer is removed. Intuitively, placing a barrier between two polarizers should reduce the chances of a photon getting through. But, in this case, it increases the chance. In this sense, quantum mechanics seems counterintuitive.

An operator \(\hat{\mathbb{P}}_\theta\) such that \(\hat{\mathbb{P}}_\theta |\theta\rangle = |\theta\rangle\) and \(\hat{\mathbb{P}}_\theta |\theta + \pi/2\rangle = -|\theta + \pi/2\rangle\) is given by
\[
\hat{\mathbb{P}}_\theta = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} = \begin{bmatrix} \cos^2 \theta - \sin^2 \theta & 2 \sin \theta \cos \theta \\ 2 \sin \theta \cos \theta & \sin^2 \theta - \cos^2 \theta \end{bmatrix} \rightarrow \hat{\mathbb{P}}_\theta |\theta\rangle
= \begin{bmatrix} \cos \theta (\cos^2 \theta - \sin^2 \theta) + 2 \sin^2 \theta \cos \theta \\ 2 \sin \theta \cos^2 \theta + (\sin^2 \theta - \cos^2 \theta) \sin \theta \end{bmatrix} = \begin{bmatrix} \cos \theta (\cos^2 \theta + \sin^2 \theta) \\ (\sin^2 \theta + \cos^2 \theta) \sin \theta \end{bmatrix}
= \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}
\]
Similarly,

\[ \mathbb{P}_\theta |\theta + \pi/2\rangle = \begin{bmatrix} \cos^2 \theta - \sin^2 \theta & 2 \sin \theta \cos \theta \\ 2 \sin \theta \cos \theta & \sin^2 \theta - \cos^2 \theta \end{bmatrix} \begin{bmatrix} -\sin \theta \\ -\cos \theta \end{bmatrix} = \begin{bmatrix} \sin \theta \\ -\cos \theta \end{bmatrix} = -\begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \]

On the face of it, this is a very strange result, since the sine and cosine are continuous functions, but a measurement of the system will find only two discrete outcomes - either the photon passes through the polarizer or it does not.

11.11.3 Circularly and Elliptically Polarized Light

Linear polarized states can be represented by state vectors with real entries. State vectors with complex entries represent circularly polarized light. For example, let

\[ |\varphi\rangle = \begin{bmatrix} 1 \\ \sqrt{2}i \\ i \\ \sqrt{2} \end{bmatrix}, \quad |\psi\rangle = \begin{bmatrix} 1 \\ \sqrt{2}i \\ -i \\ \sqrt{2} \end{bmatrix} \]

Note that \( \langle \psi | \varphi \rangle = 0 \) and \( \langle \varphi | \varphi \rangle = \langle \psi | \psi \rangle = 1 \). Hence, \( |\varphi\rangle \) and \( |\psi\rangle \) form an orthonormal basis of a two dimensional vector space.

From the standpoint of the wave theory of light, circular polarization is described as light oscillating in two perpendicular planes out of phase by 90° (see fig. 11.11.3-1). If the light is moving in the \( z \)-direction, then \( E_y = \cos(z - ct) \) and \( E_x = \sin(z - ct) \), where \( E_y \) and \( E_x \) are the components of the electric field circling around the \( z \)-axis. Light can also be elliptically polarized, where, in one of the planes, the electric field is tilted (see fig. 11.11.3-1).
The probability ‘\( P \)’ that a circularly polarized photon passes through a linear polarizer is

\[
\langle \theta | \mathcal{U} \rangle = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix} \rightarrow P = \langle \theta | \mathcal{U} \rangle \langle \mathcal{U} | \theta \rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} \cos \theta & -i \sin \theta \\ i \sin \theta & \cos \theta \end{bmatrix}
\]

\[
= \frac{1}{2} (\cos^2 \theta + \sin^2 \theta) = \frac{1}{2}
\]

Note that the probability that a circularly polarized photon passes through a linear polarizer is independent of \( \theta \).

The operator ‘\( \mathcal{P}_\circlearrowleft \)’ for a circular polarizer such that \( \mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = | \mathcal{U} \rangle \) and \( \mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = -| \mathcal{U} \rangle \) must be constructed so that

\[
\mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = | \mathcal{U} \rangle = \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad \mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = -| \mathcal{U} \rangle = -\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}
\]

Hence,

\[
\mathcal{P}_\circlearrowleft = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}
\]

To see this,

\[
\mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad \mathcal{P}_\circlearrowleft | \mathcal{U} \rangle = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -i \end{bmatrix} = \begin{bmatrix} -1 \\ -i \end{bmatrix} = -\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}
\]

### 11.11.4 Time Evolution and Schrödinger’s Equation

In classical mechanics, states are points in phase space. Each state is described by a function \( f(p(t), q(t)) \), where \( p \) is the momentum and \( q \) the position of the system at any moment in time. There are only two relationships between the points in phase space that make sense:

\[
f(p, q) = f(p', q'), \quad f(p, q) \neq f(p', q')
\]

In other words, states in phase space are either the same or different. Moreover, if a state evolves in time

\[
f(p, q) \xrightarrow{t} f(r, s), \quad f(p, q) = f(p', q') \rightarrow f(p', q') \xrightarrow{t} f(r', s') \rightarrow f(r, s) = f(r', s')
\]

Classical systems demand that indistinguishable points remain indistinguishable and distinguishable points remain distinguishable over time. States, in classical physics, are conserved.
Systems in classical mechanics move in a smooth continuous fashion and are governed by the Hamiltonian function \( H(p, q) \) and the principle of least action. Recall that the Hamiltonian represents the energy of the system. Since the Hamiltonian is a continuous function, the derivatives of \( H \) describe how the function changes with position and momentum i.e. \( \partial H / \partial p = \dot{q} \) and \( \partial H / \partial q = -\dot{p} \), where \( \dot{q} = \partial q / \partial t \), \( \dot{p} = \partial p / \partial t \). For instance, if \( H(p, q) = p^2 / 2m + U(q) \), then \( \partial H / \partial p = \dot{q} \) and \( \partial H / \partial q = \partial U(q) / \partial t = -\dot{p} \).

If there is an arbitrary function \( f(p, q) \), then
\[
\frac{df}{dt} = \frac{\partial f}{\partial p} \frac{dp}{dt} + \frac{\partial f}{\partial q} \frac{dq}{dt} = \frac{\partial f}{\partial p} \dot{p} + \frac{\partial f}{\partial q} \dot{q} = -\frac{\partial f}{\partial q} \frac{\partial H}{\partial q} + \frac{\partial f}{\partial q} \frac{\partial H}{\partial p} = \{f, H\},
\]
where \( \{f, H\} \) is called the ‘Poisson bracket’. In some sense, the equation above encapsulates the entirety of classical mechanics; the quantity \( 'H' \) describes the flow of the system and if there is any interesting feature of the system \( 'f' \), then \( df / dt \) describes how this feature changes with the flow.

In quantum mechanics, unlike classical physics, it is not enough to write \( |a\rangle = |b\rangle \) and claim that the two states are the same. Recall that if \( |b\rangle = e^{i\theta} |a\rangle \), then \( \langle b|b\rangle = \langle a|a\rangle \), but \( |a\rangle \neq |b\rangle \). The distinction between states is determined by their inner product. Two states are different if \( \langle a|b\rangle = 0 \). If \( \langle a|b\rangle \neq 0 \) \( \forall \ 1 \), then \( \langle a|b\rangle \) is a measure of the similarity between the states.

There is a quantum mechanical analog to the classical conservation law, namely, that the inner product of states is conserved over time i.e. if \( |a\rangle \rightarrow \ |a'\rangle \) and \( |b\rangle \rightarrow |b'\rangle \), then \( \langle a|b\rangle = \langle a'|b'\rangle \). Let \( \mathcal{U}|a\rangle = |a'\rangle \) and \( \mathcal{U}|b\rangle = |b'\rangle \), where \( \mathcal{U} \) is an operator and a function of time. Hence, \( \langle b'|a'\rangle = \langle b|\mathcal{U}^\dagger \mathcal{U}|a\rangle = \langle b|a\rangle \) implies that \( \mathcal{U}^\dagger \mathcal{U} = I \) i.e. \( \mathcal{U} \) is a unitary operator.

**11.11.4.1 Schrödinger’s Equation**

If \( t \rightarrow t + \Delta t \), then \( |\Psi(t + \Delta t)\rangle = \mathcal{U}(\Delta t)|\Psi(t)\rangle \), where \( |\Psi(t)\rangle \) is an arbitrary state vector. If \( \Delta t = 0 \), then \( |\Psi(t + \Delta t)\rangle = |\Psi(t)\rangle \rightarrow |\mathcal{U}(0)\rangle = I \). Now suppose \( \mathcal{U}(\Delta t) \) is a function written in a power series i.e.
\[
\mathcal{U}(\varepsilon) = 1 + s_1 \varepsilon + s_2 \varepsilon^2 + \cdots
\]
If \( \varepsilon \ll 1 \), the terms in the series that contain \( \varepsilon^2 \) or higher can be ignored, being too small to count. Recall that \( e^{\varepsilon h} = 1 + \varepsilon h \), if terms of \( \varepsilon^2 \) or higher are ignored. Now
\[
(1 + \varepsilon h)^2 = 1 + 2\varepsilon h + \varepsilon^2 h^2 = 1 + 2\varepsilon h = e^{2\varepsilon h}
\]
to a first order. In general, \( (1 + \varepsilon h)^n = e^{n\varepsilon h} \). As \( n \rightarrow \infty \), \( n\varepsilon = t \) and this implies that \( (1 + \varepsilon h)^n \rightarrow e^{\varepsilon h} \), where \( t \) represents time. Hence, if \( s_1 = -i\mathcal{H}/h \), then
\[
\mathcal{U}(\varepsilon) = 1 + s_1 \varepsilon \rightarrow \mathcal{U}(\varepsilon) = 1 - \frac{i\mathcal{H}\varepsilon}{h} \rightarrow \mathcal{U}(\varepsilon) \approx e^{-i\mathcal{H}t/h}
\]

288
Since \( \hat{U} \) is unitary,
\[
\hat{U}^\dagger \hat{U} = 1 \rightarrow \hat{U}^\dagger(\varepsilon) \hat{U}(\varepsilon) = \left(1 + \frac{i \hat{H} \varepsilon}{\hbar}\right) \left(1 - \frac{i \hat{H} \varepsilon}{\hbar}\right) = 1 - \frac{i \varepsilon}{\hbar} + \frac{\dot{\varepsilon}}{\hbar^2},
\]
where \( \hat{H} \hat{H}^\dagger \varepsilon^2 / \hbar^2 \) has been discarded as too small to count. Note that \( \hat{H} \) is Hermitian. Hence, \( \hat{U} \), since it is related directly to \( \hat{H} \), must be a measurable quantity, where \( \hat{H} \) has real eigenvalues. The operator \( \hat{H} \) turns out to be associated with the energy of a quantum system, an analog to the Hamiltonian in classical physics. But for now, \( \hat{H} \) is just a Hermitian operator.

A state \( |\Psi(t + \varepsilon)\rangle \) changes with time if
\[
|\Psi(t + \varepsilon)\rangle = \hat{U}(\varepsilon)|\Psi(t)\rangle \rightarrow \frac{|\Psi(t + \varepsilon)\rangle - |\Psi(t)\rangle}{\varepsilon} = -\frac{i \hat{H}}{\hbar} |\Psi(t)\rangle
\]
\[
\frac{\partial |\Psi\rangle}{\partial t} = -\frac{i \hat{H}}{\hbar} |\Psi\rangle \rightarrow i \hbar \frac{\partial |\Psi\rangle}{\partial t} = \hat{H} |\Psi\rangle
\]
Note that
\[
\frac{\partial |\Psi\rangle}{\partial t} = -\frac{i \hat{H}}{\hbar} |\Psi\rangle \rightarrow i \hbar \frac{\partial |\Psi\rangle}{\partial t} = \hat{H} |\Psi\rangle
\]

The equation on the right-side above is the famed ‘Schrödinger’s equation’. It is the fundamental equation of quantum mechanics.

Example: Suppose a material is constructed in such a way that if vertically polarized light passes through the material, then the energy of the light is \( E_1 \) and if horizontally polarized, then the energy is \( E_2 \) i.e.
\[
\hat{H}|0\rangle = E_1|0\rangle, \quad \hat{H}|90\rangle = E_2|90\rangle
\]
If the state of a photon is
\[
|\Psi(t)\rangle = \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix},
\]
where \( \alpha(t) \) is associated with the probability that the photon is vertically polarized and \( \beta(t) \) associated with the probability that the photon is horizontally polarized, then
\[
\begin{bmatrix} \dot{\alpha}(t) \\ \dot{\beta}(t) \end{bmatrix} = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} = \begin{bmatrix} E_1 \alpha(t) \\ E_2 \beta(t) \end{bmatrix}
\]
Hence,
\[
\begin{bmatrix} \frac{d\alpha}{dt} = E_1 \alpha(t) \\ \frac{d\beta}{dt} = E_2 \beta(t) \end{bmatrix} \rightarrow \frac{d\alpha}{dt} = -\frac{i}{\hbar} E_1 \alpha(t), \quad \frac{d\beta}{dt} = -\frac{i}{\hbar} E_2 \beta(t)
\]
The solutions to the above equations are

\[ \alpha(t) = \alpha(0)e^{-iE_1t/\hbar}, \quad \beta(t) = \beta(0)e^{-iE_2t/\hbar}, \]

which implies that

\[ |\Psi(t)\rangle = \begin{bmatrix} \alpha e^{-iE_1t/\hbar} \\ \beta e^{-iE_2t/\hbar} \end{bmatrix}, \]

where \( \alpha(0) = \alpha \) and \( \beta(0) = \beta \). The probability \( P_{0^\circ} \) that the photon passes vertically through the material is simply

\[ P_{0^\circ} = \alpha e^{-iE_1t/\hbar}\alpha^\dagger e^{iE_1t/\hbar} = \alpha\alpha^\dagger \]

Likewise, the probability \( P_{90^\circ} \) that the photon passes through horizontally polarized is \( \beta\beta^\dagger \). If a photon is prepared vertically polarized, then passes through the material, the probability it will pass through a second vertically polarized polarizer is \( \alpha\alpha^\dagger \). Time has no effect on the probability that the photon passes through a second vertical polarizer. The same can be said of a horizontally polarized photon.

If the photon is run through a \( 45^\circ \) polarizer, then

\[ |45^\circ\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \rightarrow \quad \langle 45^\circ|\Psi\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \left[ \begin{bmatrix} \alpha e^{-iE_1t/\hbar} \\ \beta e^{-iE_2t/\hbar} \end{bmatrix} = \frac{\alpha}{\sqrt{2}} e^{-iE_1t/\hbar} + \frac{\beta}{\sqrt{2}} e^{-iE_2t/\hbar} \right] \rightarrow P_{45^\circ} \]

\[ = \left( \frac{\alpha}{\sqrt{2}} e^{-iE_1t/\hbar} + \frac{\beta}{\sqrt{2}} e^{-iE_2t/\hbar} \right) \left( \frac{\alpha}{\sqrt{2}} e^{-iE_1t/\hbar} + \frac{\beta}{\sqrt{2}} e^{-iE_2t/\hbar} \right)^\dagger \]

To simplify matters, let \( \alpha = \beta = 1/\sqrt{2} \), which corresponds to the photon entering a second polarizer polarized at \( 45^\circ \), then

\[ P_{45^\circ} = \langle 45^\circ|\Psi\rangle\langle\Psi|45^\circ\rangle = \left( \frac{1}{2} e^{-iE_1t/\hbar} + \frac{1}{2} e^{-iE_2t/\hbar} \right) \left( \frac{1}{2} e^{-iE_1t/\hbar} + \frac{1}{2} e^{-iE_2t/\hbar} \right)^\dagger \]

\[ = \frac{1}{4} \left( 2 + e^{i(E_2-E_1)t/\hbar} + e^{-i(E_2-E_1)t/\hbar} \right) = \frac{1}{4} [2 + 2 \cos(\Delta Et)], \]

where \( \Delta E = E_2 - E_2 \).

What is the probability that a photon polarized to \( 45^\circ \) passes through a material that gives different energies to the vertical and horizontal polarization states and then passes through a second polarizer also polarized \( 45^\circ \)? In this case, the probability that the photon passes through the second polarizer oscillates with time. At \( t = 0 \), \( P_{45^\circ} = 1 \), which is to be expected. But at \( t = \pi \), \( P_{45^\circ} = 0 \).
11.11.4.2 The Expected Value of an Operator over Time

Suppose there is an arbitrary operator \( \hat{R} \) that acts on a state vector \(|\Psi\rangle\). How does the average value of \( \hat{R} \) i.e. \( \langle \hat{R} \rangle = \langle \Psi | \hat{R} | \Psi \rangle \) change with time? Here \( \hat{R} \) is definite and does not change with time, but the state vector \(|\Psi\rangle\) does. To compute this, use Schrödinger’s equation:

\[
 ih \frac{\partial |\Psi\rangle}{\partial t} = \hat{H} |\Psi\rangle \rightarrow \frac{\partial |\Psi\rangle}{\partial t} = -\frac{i}{\hbar} \hat{H} |\Psi\rangle, \quad \frac{\partial \langle \Psi |}{\partial t} = \frac{i}{\hbar} \langle \Psi | \hat{H}^\dagger \]

Now

\[
 \frac{d\langle \hat{R} \rangle}{dt} = \frac{d\langle \Psi | \hat{R} | \Psi \rangle}{dt} = \langle \Psi | \hat{R} | \Psi \rangle + \langle \Psi | \hat{R} | \Psi \rangle = \frac{i}{\hbar} \langle \Psi | \hat{H}^\dagger \hat{R} | \Psi \rangle - \frac{i}{\hbar} \langle \Psi | \hat{R} \hat{H} | \Psi \rangle
\]

\[
 = -\frac{i}{\hbar} \langle \Psi | [\hat{R}, \hat{H}] | \Psi \rangle,
\]

where \([\hat{R}, \hat{H}] = \hat{R} \hat{H} - \hat{H} \hat{R} \), the commutator of \( \hat{R} \) with \( \hat{H} \). The last equation on the right is justified because \( \hat{H} \) is Hermitian. Hence,

\[
 \frac{d\langle \hat{R} \rangle}{dt} = -\frac{i}{\hbar} [\hat{R}, \hat{H}]
\]

When \([\hat{R}, \hat{H}] = 0 \), then \( \langle \hat{R} \rangle = a \text{ constant} \) and the observable represented by \( \hat{R} \) is conserved. If an arbitrary operator \( \hat{R} \) commutes with \( \hat{H} \), \( \langle \hat{R} \rangle \) does not change with time. Recall that in classical mechanics, to determine if a state changes with time, the Hamiltonian is inserted into the Poisson bracket. Something very similar happens in quantum mechanics, where the commutator \([\hat{R}, \hat{H}]\) determines if some measurable quantity changes over time.

11.11.4.3 The Role of Time in Quantum Mechanics

If a particle moves along a horizontal line with no forces acting on it, then \( \langle x | \Psi(t) \rangle = \Psi(x, t) \). By Schrödinger’s equation

\[
 \frac{\partial |\Psi\rangle}{\partial t} = -\frac{i}{\hbar} \hat{H} |\Psi\rangle
\]

In classical physics, \( E = H = p^2 / 2m \), since \( E = 1/2 mv^2 \) and \( p = mv \), suggesting that the ‘\( H \)’ in quantum mechanics might be equal to the energy. Hence,

\[
 \frac{\partial |\Psi\rangle}{\partial t} = -\frac{i}{\hbar} \hat{p}^2 |\Psi\rangle
\]

where, in this case, \( \hat{p} \) is an operator. But

\[
 \hat{p} = -i\hbar \frac{\partial}{\partial x} \rightarrow \hat{p}^2 |\Psi\rangle = \left(-i\hbar \frac{\partial}{\partial x}\right) \left(-i\hbar \frac{\partial}{\partial x}\right) |\Psi\rangle \rightarrow \frac{\partial |\Psi\rangle}{\partial t} = \frac{i}{2m} \frac{\partial^2 |\Psi\rangle}{\partial x^2}
\]
Let $\Psi = f(t)e^{ikx}$, where $k = p/h$. The function ‘$f(t)$’ is included to make $\Psi$ time dependent. Hence,

$$\frac{\partial |\Psi\rangle}{\partial t} = |\dot{\Psi}\rangle = f(ik)^2 e^{ikx} = -f \frac{i\hbar}{2m} k^2 e^{ikx} \rightarrow \dot{f} = -f \frac{i\hbar}{2m} k^2 \rightarrow f = e^{-\frac{i p^2}{2m t}}$$

$$\rightarrow |\Psi\rangle = e^{ikx - \frac{i p^2}{2m t}}$$

In general, if a particle moves with no forces acting on it, then its motion in time is described by the Schrödinger equation i.e.

$$i\hbar \frac{\partial |\Psi\rangle}{\partial t} = \hat{H} |\Psi\rangle = \frac{\hat{p}^2}{2m} |\Psi\rangle = \frac{1}{2m} \left(-i\hbar \frac{\partial}{\partial x} \right) \left(-i\hbar \frac{\partial}{\partial x} \right) |\Psi\rangle = \frac{\hbar^2}{2m} \frac{\partial^2 |\Psi\rangle}{\partial x^2},$$

where $\hat{p} = -i\hbar \partial / \partial x \rightarrow \hat{H} = \hat{p}^2 / 2m$.

In this case, $\Psi(x, t)$ is a function of position and time. The position, say ‘$x$’, can be an eigenvalue of $\Psi$, but the time ‘$t$’ cannot. This can be a bit confusing. There is a quantum mechanical operator ‘$\hat{X}$’ for position. It makes sense to ask, “What is the probability that a particle resides at a position ‘$x$’ at time ‘$t$’?” But there is no ‘time’ operator in quantum mechanics. It makes no sense to ask, “What is the probability that the time is ‘$t$’, given that the position of the particle is $x$?” Position and time play logically different roles within quantum mechanics. In fact, time plays a role in quantum mechanics similar to the role it plays in Newtonian mechanics. It is essentially a parameter.

### 11.11.4.4 Time as a Parameter

Schrödinger’s equation for a free particle has plane wave solutions i.e.

$$|\Psi(x, t)\rangle \propto e^{ipx/\hbar}e^{i\omega t}$$

The particle oscillates with position and time. Plugging this solution into the Schrödinger equation leaves

$$i\omega \hbar |\Psi\rangle = \frac{\hbar^2}{2m} \frac{p^2}{\hbar^2} |\Psi\rangle \rightarrow -\omega = \frac{p^2}{2m\hbar} \rightarrow \omega = -\frac{p^2}{2m\hbar} \rightarrow |\Psi\rangle = e^{ipx/\hbar} e^{-ip^2t/2m\hbar}$$

There is another solution, since $p$ can be plus or minus i.e.

$$|\Psi\rangle = e^{-ipx/\hbar} e^{-ip^2t/2m\hbar}$$

The quantities ‘$\pm p$’ represent particles moving in opposite directions. According to the postulates of quantum mechanics, the most general solution is a superposition of states i.e.

$$|\Psi\rangle = \alpha e^{ipx/\hbar} e^{-ip^2t/2m\hbar} + \beta e^{-ipx/\hbar} e^{-ip^2t/2m\hbar},$$
where \( \alpha \alpha^\dagger \) is the probability that the wave is moving in the positive direction and \( \beta \beta^\dagger \) is the probability that the wave is moving in the negative direction. The probability that the particle has energy \( E = p^2/2m \) is 1, no matter which direction the particle is moving.

As described above, \( |\Psi\rangle \) corresponds to a particle that has a specified energy. Suppose the particle has different energies at different times, then

\[
|\Psi\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\Psi(p)\rangle e^{ipx/h} e^{-ip^2t/2mh} dp,
\]

where \( |\Psi(p)\rangle \) has replaced \( \alpha \). It is unnecessary to include waves moving in both directions, since the integration is over the interval \( (-\infty, \infty) \), which indicates that the line the particle is moving along can be of infinite length. The equation above represents a system where the particle can have different momentums at different times. The probability density \( P_E \) of a particle with energy \( E \) is, since \( \pm p = 2mE \),

\[
P_E = \langle \Psi(2mE)|\Psi(2mE)\rangle + \langle \Psi(-2mE)|\Psi(-2mE)\rangle
\]

At \( t = 0 \),

\[
|\Psi\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\Psi(p)\rangle e^{ipx/h} dp
\]

Recall that

\[
|\Psi(p)\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\Psi(x)\rangle e^{-ipx/h} dx \to \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\Psi(x)\rangle e^{-ip^2t/2mh} dx dp = |\Psi(t)\rangle
\]

If the wave function \( |\Psi(x)\rangle \) happens to be known, it is possible to know it at a later time.

### 11.12 The Relationship between Quantum and Classical Physics

Recall the Poisson bracket \( \{f, H\} \), where \( H \) describes the flow of the system and if there is any interesting feature of the system \( f \), then \( df/dt \) describes how this feature changes with the flow:

\[
\frac{df}{dt} = \frac{\partial f}{\partial p} \dot{p} + \frac{\partial f}{\partial q} \dot{q} = -\frac{\partial f}{\partial p} \frac{\partial H}{\partial q} + \frac{\partial f}{\partial q} \frac{\partial H}{\partial p} = \{f, H\},
\]

What is the connection between commutators and Poisson brackets?

Note that

\[
\{f, p\} = \frac{\partial f}{\partial x} \frac{\partial p}{\partial x} - \frac{\partial f}{\partial x} \frac{\partial p}{\partial x} = \frac{\partial f}{\partial x}, \quad H = p
\]
since \((\partial f/\partial p)(\partial p/\partial x) = 0\) and \(\partial p/\partial p = 1\). Recall that
\[
\frac{\partial}{\partial x} (fV) - f \frac{\partial V}{\partial x} = \frac{\partial f}{\partial x} V + f \frac{\partial V}{\partial x} - f \frac{\partial V}{\partial x} = \frac{\partial f}{\partial x} V \rightarrow \left[ \frac{\partial}{\partial x}, f(x) \right] = \frac{\partial f}{\partial x} \rightarrow \left[ f, \frac{\partial}{\partial x} \right] = -\frac{\partial f}{\partial x},
\]
since commutators anti-commute.

Hence,
\[
[\hat{p}, f] = \left[ -i\hbar \frac{\partial}{\partial x}, f \right] = -i\hbar \frac{\partial f}{\partial x} \rightarrow [f, \hat{p}] = \left[ f, -i\hbar \frac{\partial}{\partial x} \right] = i\hbar \frac{\partial f}{\partial x} \rightarrow [\hat{x}, f(p)] = i\hbar \frac{\partial f}{\partial p}.
\]
The equation on the right above follows because of the Fourier relationships between \(x\) and \(p\). Moreover, if \(f(x) = x\), then
\[
[\hat{x}, \hat{p}] = \left[ x, -i\hbar \frac{\partial}{\partial x} \right] = i\hbar \frac{\partial x}{\partial x} = i\hbar
\]
And if \(f(p) = p\), then \([\hat{x}, \hat{p}] = i\hbar\). Note that \([x, p] = \partial x/\partial x = 1\). Hence, the relationship between commutators and Poisson brackets is \(i\hbar \{a, b\} = [\hat{a}, \hat{b}]\). Moreover, commutators, in general, have a much smaller value than Poisson brackets, since \(\hbar\) is \(\approx 10^{-27}\). In classical mechanics, \([x, p] = 0\), which implies that \(\lim_{\hbar \to 0} i\hbar \{a, b\} = [a, b]\).

Both commutators and Poisson brackets are anti-symmetric i.e.
\[
\{a, b\} = -\{b, a\}, \quad [\hat{a}, \hat{b}] = -[\hat{b}, \hat{a}].
\]

Now consider
\[
\{a, \{b, c\}\} = \frac{\partial a b}{\partial x} \frac{\partial c}{\partial p} - \frac{\partial a b}{\partial p} \frac{\partial c}{\partial x} = (a \partial_x b + b \partial_x a) \partial_p c - (a \partial_p b + b \partial_p a) \partial_x c = a \{b, c\} + b \{a, c\}
\]
The above equation and \([\hat{a} \hat{b}, \hat{c}] = \hat{a} [\hat{b}, \hat{c}] + [\hat{a}, \hat{c}] \hat{b}\) are almost identical, except that Poisson brackets consist of functions or numbers, which always commute, but commutators contain operators that do not always commute.

Suppose a particle is not free, but has a force applied to it i.e. moves in a potential. The system complies with Schrödinger’s equation
\[
i\hbar \frac{\partial |\Psi\rangle}{\partial t} = H|\Psi\rangle,
\]
except now, \(H = \hat{p}^2/2m + \hat{U}(x)\), where \(\hat{U}\) represent a potential operator. Hence,
\[
i\hbar \frac{\partial |\Psi\rangle}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} |\Psi\rangle + \hat{U}(x)|\Psi\rangle
\]
If \(\hat{U}(x)\) is a smooth continuous function, then how does the average ‘\(\langle |\Psi\rangle \rangle\)’ change with time? Recall that for any operator ‘\(\hat{R}\)’, its expected value is

294
\[
\frac{d\langle \hat{R} \rangle}{dt} = -\frac{i}{\hbar} [\hat{R}, \hat{H}]
\]

Suppose \( |\Psi\rangle = \hat{X} \)

\[
\frac{d}{dt} \langle \hat{X} \rangle = \frac{i}{\hbar} \hat{p}^2/2m + \hat{U}(x), \hat{X} \rangle = \frac{i}{\hbar} \hat{p}^2/2m, \hat{X} \rangle = \frac{i}{2m\hbar} \hat{p}^2, \hat{X} \rangle = \frac{i}{2m\hbar} (\hat{p}[\hat{p}, \hat{X}] + [\hat{p}, \hat{X}]\hat{p})
\]

\[
= \frac{i}{2m\hbar} (-2i\hbar \hat{p}) = \frac{\hat{p}}{m} \rightarrow \frac{d}{dt} \langle \hat{X} \rangle = \frac{\hat{p}}{m} = \langle \hat{v} \rangle,
\]

since \( [\hat{U}, \hat{X}] = 0 \). Moreover, if \( |\Psi\rangle = \hat{p} \)

\[
\frac{d}{dt} \langle \hat{p} \rangle = \frac{i}{\hbar} \hat{p}^2/2m + \hat{U}(x), \hat{p} \rangle = \frac{i}{\hbar} \hat{U}, \hat{p} \rangle = \frac{i}{\hbar} i\hbar \frac{\partial U}{\partial x} = -\frac{\partial U}{\partial x} = \langle F \rangle,
\]

since \( \hat{p}^2/2m, \hat{p} \rangle = 0 \) and \( F \) is the force. On average the time rate of change of the position operator agrees with Newtonian mechanics, where the change in position with time is equal to the velocity. And, on average, the time rate of change of the momentum operator is equal to the force, which also agrees with Newtonian mechanics.

11.13 Symmetry

If some property of a system remains unchanged after a number of transformations or operations, this indicates there is ‘a symmetry’ in the system. Symmetries in physics are generally synonymous with conservation laws, revealed in Noether’s theorem:

Every continuous symmetry is associated with a conservation law.

In quantum physics symmetries determine which particle interactions can take place and which cannot. Symmetries in physics appear when variables within a mathematical equation describing a system change, but the value of the equation remains unchanged. For instance, suppose \( \hat{U} \) is a unitary operator and consider \( \hat{U} |a\rangle \) and \( \hat{U} |b\rangle \). If \( \langle a | b \rangle = c \), then \( \langle a | \hat{U}^\dagger \hat{U} | b \rangle = \langle a | | b \rangle = \langle a | b \rangle = c \). A symmetry is indicated here, since the inner product of two state vectors remains unchanged if those state vectors are operated on by a unitary operator.

11.13.1 Symmetry and Time Evolution

Now suppose \( \hat{V} |\Psi\rangle = |\Psi'\rangle \), where \( V^\dagger V = I \). Further suppose that \( \hat{U} |\Psi_1\rangle = |\Psi_2\rangle \). Over time, the state \( |\Psi_1\rangle \) migrates to the state \( |\Psi_2\rangle \). Since \( V \) is a symmetry operation,

\[
\hat{V} |\Psi\rangle = |\Psi'\rangle \rightarrow \hat{U} \hat{V} |\Psi_1\rangle = |\Psi_2\rangle = \hat{V} \hat{U} |\Psi_1\rangle \rightarrow [\hat{U}, \hat{V}] = 0
\]

Hence, any symmetry operator commutes with the time evolution operator, which is also unitary, suggesting that finding a symmetry in a system is associated with operators that commute with the time evolution operator.
Recall that
\[
\hat{U}(\varepsilon) = 1 - \frac{i\hat{A}\varepsilon}{\hbar} \rightarrow \hat{V} \left( 1 - \frac{i\hat{A}\varepsilon}{\hbar} \right) = \left( 1 - \frac{i\hat{A}\varepsilon}{\hbar} \right) \hat{V} = \frac{i\hat{A}\varepsilon}{\hbar} \hat{V} \rightarrow \hat{V} \frac{i\hat{A}\varepsilon}{\hbar} - \frac{i\hat{A}\varepsilon}{\hbar} \hat{V} \\
= \frac{i\varepsilon}{\hbar} (\hat{V}\hat{A} - \hat{A}\hat{V}) = 0 \rightarrow [\hat{V}, \hat{A}] = 0
\]

Hence, \(\hat{V}\) commutes with the Hamiltonian. Therefore, a symmetry is an operation which commutes with the Hamiltonian. Note that since \([\hat{V}, \hat{A}] = 0\), whatever operator \(\hat{V}\) represents, it is conserved. This is a direct consequence of Noether's theorem.

### 11.13.2 Continuous Symmetries

There are two types of symmetries: 'discrete' and 'continuous'. Discrete symmetries have a finite number of symmetry operations. For instance, a square has nine symmetry operations, which leave the square looking the same before as after the operation is performed.

Any continuous symmetry '\(\hat{V}\)' can be represented by \(\hat{V} = 1 - i\varepsilon \hat{G}\), where \(\varepsilon \ll 1\). Recall that
\[
[\hat{V}, \hat{A}] = 0 \rightarrow [1 - i\varepsilon \hat{G}, \hat{A}] = [\hat{G}, \hat{A}] = 0
\]

The operator '\(\hat{G}\)' is called the 'generator' of a continuous symmetry. Finding symmetries within a system is important because the symmetries of a system are synonymous with the characteristics of a system that are conserved.

Example: Suppose the wave function '|\(\Psi(x)\)\)' is translated to a position '|\(\Psi(x - \varepsilon)\)\', where \(\varepsilon \ll 1\). This represents a wave moving to the right and implies that
\[
\hat{V}|\Psi(x)\rangle = |\Psi(x - \varepsilon)\rangle = |\Psi(x)\rangle - \varepsilon \partial |\Psi\rangle / \partial x \rightarrow \hat{V} = 1 - \varepsilon \partial / \partial x
\]

Recall the momentum operator '\(\hat{K} = -i\hbar \partial / \partial x\) \rightarrow \hat{K}/\partial x = \hat{R}/\hbar\)'. Hence,
\[
\hat{V} = 1 - \frac{i\varepsilon \hat{R}}{\hbar}
\]

Note that \(\hat{R}\) is the generator of translations. Whether or not the momentum along the \(x\) -axis is conserved depends on the Hamiltonian. The Hamiltonian for free particle is \(H = \hat{K}^2/2m\) and, since \(\hat{R}\) commutes with itself, the momentum of a free particle is conserved.

### 11.13.3 Rotational Symmetry

A rotation is a continuous symmetry. Large rotations are built from small ones. If a particle is moving freely along a circle, its wave function only depends on the angle the radius vector '\(r\)' makes with the horizontal and the magnitude of \(r\). If \(r\) is a constant, the wave function becomes a function only of the angle '\(\theta\)' i.e. \(\Psi(\theta)\). Note that \(\Psi\Psi^\dagger\) is the
probability that the particle is located at position \( r\theta \). Suppose \( \Psi(\theta) \) is rotated by a small amount i.e. \( \Psi(\theta) \rightarrow \Psi(\theta + \varepsilon) \), \( \varepsilon \ll 1 \). Note that \( d\Psi = \varepsilon \partial \Psi / \partial \theta \), since

\[
\frac{\Psi(\theta + \varepsilon) - \Psi(\theta)}{\varepsilon} \approx \frac{\partial \Psi}{\partial \theta} \rightarrow \Psi(\theta + \varepsilon) - \Psi(\theta) = d\Psi \approx \varepsilon \frac{\partial \Psi}{\partial \theta}
\]

Define

\[
\hat{L} = -\frac{i}{\hbar} \frac{\partial}{\partial \theta} \rightarrow d|\Psi\rangle = \frac{\varepsilon}{\hbar} \hat{L}|\Psi\rangle,
\]

since, if \( \Psi \) is to be quantum mechanical, the wave function must contain an \( \hbar \), where \( \Psi \) has been replaced by the ket \('|\Psi\rangle'\) to indicate it is a state vector. Note that the operator \( \hat{L} \) is the generator of rotations, and physically, represents ‘angular momentum’. If \( \hat{L} \) commutes with \( \hat{H} \), the Hamiltonian, then \( \hat{L} \) is conserved, and hence, a symmetry.

### 11.13.3.1 Degenerate States

Now suppose \( \hat{L} \) acts on \(|\Psi\rangle\) i.e.

\[
\hat{L}|\Psi(\theta)\rangle = m|\Psi(\theta)\rangle = -i \frac{\partial}{\partial \theta} |\Psi(\theta)\rangle = m|\Psi(\theta)\rangle \rightarrow |\Psi(\theta)\rangle = e^{im\theta}, \quad \hbar = 1
\]

Since the particle is moving in a circle,

\[
e^{im\theta} = e^{i2\pi n m} = 1
\]

Hence, \( m \) must be an integer and \( \hat{L} \), the angular momentum, can be defined

\[
L = m, \quad m \in \mathbb{Z}
\]

The energy of a particle moving along a circle depends on its angular momentum. The energy of a particle with angular momentum \( 'm' \) should equal the energy of a particle with angular momentum \( '-m' \), since the only difference between a particle with angular momentum \( 'm' \) and one with angular momentum \( '-m' \) is the direction in which the particle is traversing the circle. But if \( E(m) = E(-m) \), then different states of the particle have the same energy, but different eigenvalues of angular momentum. Upon measurement, if two or more states of a quantum mechanical system give the same value for the energy, the states are said to be ‘degenerate’. As it turns out, in addition to rotational symmetry, if the system has mirror reflection symmetry, then the system is degenerate.

![Figure 11.13.3-1](image-url)
Figure 11.13.3-1 shows an example of mirror reflection symmetry. If a particle is moving in a circle, its mirror reflection rotates in the opposite direction. Introduce a mirror reflection operator ‘\( \hat{M} \)’ such that \( \hat{M} |\Psi(\theta)\rangle = |\Psi(-\theta)\rangle \). Hence,

\[
\hat{M} L e^{im\theta} = \hat{M} m e^{im\theta} = me^{-im\theta}, \quad \Psi(\theta) = e^{im\theta}
\]

But,

\[
\hat{L} \hat{M} e^{im\theta} = \hat{L} e^{-im\theta} = -me^{-im\theta}
\]

Hence, \([\hat{M}, \hat{L}] = 2m\) i.e. \(\hat{L}\) and \(\hat{M}\) do not commute. If symmetries do not commute, the system has degenerate energy levels.

Recall that if \(\hat{A}\) is the generator of an operator and \([\hat{A}, \hat{H}] = 0\) i.e. if \(\hat{A}\) commutes with the Hamiltonian, then \(\hat{A}\) is a ‘symmetry’ of the system. If \([\hat{A}, \hat{H}] = 0\) and \([\hat{B}, \hat{H}] = 0\), it does not necessarily follow that \([\hat{A}, \hat{B}] = 0\). However, if \([\hat{A}, \hat{B}] = i\hat{C}\) and \(\hat{C}\) is not a linear combination of \(\hat{A}\) and \(\hat{B}\), then \(\hat{C}\) is a new symmetry, since \([\hat{C}, \hat{H}] = 0\). To see this, note that

\[
[\hat{C}, \hat{H}] = \hat{A}\hat{B}\hat{H} - \hat{H}\hat{A}\hat{B} = \hat{H}\hat{A}\hat{B} - \hat{H}\hat{A}\hat{B} = 0,
\]

since \(\hat{A}\) and \(\hat{B}\) commute with \(\hat{H}\). If this process is repeated, it generates all the symmetries of the system. The collection or set of symmetries of a system is called a ‘symmetry group’.

### 11.14 Angular Momentum in Quantum Mechanics

Classically, the angular momentum ‘\(L\)’ of a particle is defined as

\[
L = \langle r | \times |p\rangle,
\]

where \(r\) is the radius vector and \(p\) is the linear momentum of the particle measured tangent to the curve drawn out by the radius vector. If the particle is moving in a circle in the \(x, y\)-plane, then its angular momentum is given by ‘\(|L_z\rangle\)’, which is normal to the \(x, y\)-plane and parallel to the \(z\)-axis (see fig. 11.14-1). If the system has rotational symmetry, the system can be rotated, for example, so that particle is rotating in the \(x, z\)-plane and its angular momentum is given by ‘\(|L_y\rangle\)’. In either case, the energy of the particle remains unchanged.
If a particle is at a certain position in the $x, y$-plane, which is rotated slightly by a small amount $\varepsilon$, then the particle's position changes slightly relative to the $x, y$-plane. The change in the $x$-direction is $dx = -\varepsilon y$ and the change in the $y$-direction is $dy = \varepsilon x$ (see fig. 11.14-2).

Hence,

$$x = r \cos \varepsilon, \quad y = r \sin \varepsilon \rightarrow dx \approx -\varepsilon r \sin \varepsilon = -\varepsilon y, \quad dy = \varepsilon r \cos \varepsilon = \varepsilon x$$

By the definition of a differential

$$d\Psi = \frac{\partial\Psi}{\partial x} dx + \frac{\partial\Psi}{\partial y} dy = -i\varepsilon \frac{\partial\Psi}{\partial x} y + i\varepsilon \frac{\partial\Psi}{\partial y} x \rightarrow d\Psi = i\varepsilon \hat{L}_z \Psi, \quad \hbar = 1$$

Recall that

$$\hat{R}\Psi = -i\hbar \frac{\partial\Psi}{\partial x} \rightarrow i\hat{R}\Psi \rightarrow \frac{\partial}{\partial x} = \hat{p}_x, \quad \hbar = 1$$

Therefore,

$$L = \langle r \times p \rangle \rightarrow -i\varepsilon y \hat{p}_x + i\varepsilon x \hat{p}_y = i\varepsilon \hat{L}_z \rightarrow \hat{L}_z = x\hat{p}_y - y\hat{p}_x$$

Angular momentum in quantum mechanics is defined the same way it is in classical mechanics, except that $L$ is replaced by the operator $\hat{L}$. By similar arguments,
\[
\hat{L}_x = y\hat{p}_z - z\hat{p}_y, \quad \hat{L}_y = z\hat{p}_x - x\hat{p}_z
\]

If the system has rotational symmetry, then \([\hat{L}_i, \hat{H}] = 0, \ i = x, y, z\). In other words, the \(\hat{L}_i\)'s commute with the Hamiltonian and angular momentum would be conserved along the \(x, y, z\)-axes. Moreover, angular momentum would be conserved under rotational symmetry in any direction, since an arbitrary direction of \(\hat{L}\) can be written as a linear combination of the \(\hat{L}_i\)'s. And any linear combination of the \(\hat{L}_i\)'s would commute with the Hamiltonian, and therefore, be conserved.

11.14.1 Angular Momentum Commutation Operators

Does \([\hat{L}_i, \hat{L}_j] = 0\)? If \(i = j\), then \([\hat{L}_i, \hat{L}_j] = 0\). All operators commute with themselves. If \(i \neq j\), then

\[
[j, \hat{p}_j] = i\hbar = i, \quad h = 1, \quad j = x, y, z
\]

For example,

\[
[\hat{L}_x, \hat{L}_y] = [\hat{y}\hat{p}_z - \hat{z}\hat{p}_y, \hat{z}\hat{p}_x - \hat{x}\hat{p}_z] = [\hat{y}\hat{p}_z, \hat{z}\hat{p}_x] + [\hat{z}\hat{p}_y, \hat{x}\hat{p}_z] - [\hat{y}\hat{p}_z, \hat{x}\hat{p}_z] - [\hat{z}\hat{p}_y, \hat{z}\hat{p}_x]
\]

The last two commutators in the equation above vanish leaving

\[
[\hat{L}_x, \hat{L}_y] = [\hat{y}\hat{p}_z, \hat{z}\hat{p}_x] + [\hat{z}\hat{p}_y, \hat{x}\hat{p}_z]
\]

So,

\[
[\hat{y}\hat{p}_z, \hat{z}\hat{p}_x] = yp_zxp_x - zp_xyp_z = yp_x[p_z, \hat{z}] = -iy\hbar \hat{p}_x
\]

Similarly,

\[
[\hat{z}\hat{p}_y, \hat{x}\hat{p}_z] = ix\hbar \hat{p}_y
\]

Therefore,

\[
[\hat{L}_x, \hat{L}_y] = ix\hbar \hat{p}_y - iy\hbar \hat{p}_x = i\hat{L}_z, \quad h = 1
\]

Similarly, \([\hat{L}_y, \hat{L}_z] = i\hat{L}_x\) and \([\hat{L}_z, \hat{L}_x] = i\hat{L}_y\). Hence, the \(\hat{L}_i\)'s do not commute.

11.14.2 Angular Momentum Raising and Lowering Operators

Recall that \([\hat{L}_x, \hat{L}_z] = -i\hat{L}_y\) and \([i\hat{L}_y, \hat{L}_z] = -\hat{L}_x\) and define \(\hat{L}_+ = \hat{L}_x + i\hat{L}_y\) and \(\hat{L}_- = i\hat{L}_y - \hat{L}_x\) = \(\hat{L}_-\). Note that

\[
[\hat{L}_+, \hat{L}_z] = [\hat{L}_x + i\hat{L}_y, \hat{L}_z] = [\hat{L}_x, \hat{L}_z] + i[\hat{L}_y, \hat{L}_z] = -i\hat{L}_y - \hat{L}_x = -\hat{L}_+
\]

Similarly, \([\hat{L}_-, \hat{L}_z] = \hat{L}_-\).
Suppose \( \hat{L}_z | m \rangle = m | m \rangle \). In other words, suppose \( \hat{L}_z \) acting on a state \( | m \rangle \) gives eigenvalue \( 'm' \), then

\[
[\hat{L}_+, \hat{L}_z] | m \rangle = (\hat{L}_+ \hat{L}_z - \hat{L}_z \hat{L}_+) | m \rangle = -\hat{L}_+ | m \rangle \rightarrow m \hat{L}_+ | m \rangle - \hat{L}_z \hat{L}_+ | m \rangle = -\hat{L}_+ | m \rangle \\
\rightarrow m \hat{L}_+ | m \rangle + \hat{L}_+ | m \rangle = \hat{L}_z \hat{L}_+ | m \rangle \rightarrow (m + 1) \hat{L}_+ | m \rangle = \hat{L}_z \hat{L}_+ | m \rangle \rightarrow \hat{L}_z | \Psi \rangle \\
= (m + 1) | \Psi \rangle, \quad | \Psi \rangle = \hat{L}_+ | m \rangle
\]

Hence, \( m + 1 \) is an eigenvalue of \( \hat{L}_z \) with eigenvector \( '\hat{L}_+ | m \rangle' \). If \( \hat{L}_+ \) operates on the state \( '|m\rangle' \), it raises the angular momentum of the system by one unit. Conversely, if \( \hat{L}_- \) operates on the state \( '|m\rangle' \), it lowers the angular momentum by one unit. Moreover, let \( \hat{H} | m \rangle = E | m \rangle \). Note that

\[
\hat{H} \hat{L}_+ | m \rangle = \hat{L}_+ \hat{H} | m \rangle \rightarrow \hat{H} | m + 1 \rangle = \hat{L}_+ E | m \rangle = E | m + 1 \rangle
\]

Because \( \hat{L}_+ \) is a linear combination of \( \hat{L}_x \) and \( \hat{L}_y \), it commutes with the Hamiltonian. Raising the angular momentum by one unit does not change the energy. The upshot is that, in a system with rotational symmetry, if the symmetry operators do not commute, then the system has degenerate states.

### 11.14.3 Quantizing Angular Momentum

In a system having rotational symmetry, if \( \hat{L}_z \) has eigenvalue \( 'm' \), then \( -\hat{L}_z \), which is \( \hat{L}_z \) rotated by \( \pi \), has eigenvalue \( '-m' \). Note that \( \hat{L}_+ | - m \rangle = | - m + 1 \rangle \) and \( \hat{L}_+ | - m + 1 \rangle = | - m + 2 \rangle \) and so on. The only way this can happen is if \( 'm' \) is an integer or a half-integer. No other set of numbers will have rotational symmetry around \( m = 0 \).

In some systems, \( m \) can be infinite, but ordinarily, \( m \) is finite. If \( m \) is a maximum, this is represented by \( \hat{L}_+ | m \rangle = | 0 \rangle \). Hence, \( -m \) is the minimum value. Let \( l = m \) and \( -l = -m \). Therefore, there are \(-l, -l + 1, ... , 0, ... , l - 1, l = 2l + 1, l \in Z \) states of angular momentum of the system.

Now consider \( \hat{L}^2 \) i.e.

\[
\hat{L}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2 = \hat{L}_x^2 + (\hat{L}_x - i \hat{L}_y)(\hat{L}_x + i \hat{L}_y) - i[\hat{L}_x, \hat{L}_y] = \hat{L}_z^2 + \hat{L}_x + \hat{L}_y
\]

The term \( i[\hat{L}_x, \hat{L}_y] \) in the equation above is required because \( \hat{L}_x \) and \( \hat{L}_y \) do not commute. Hence, if \( \hat{L}_z | l \rangle = l | l \rangle \), then

\[
\hat{L}^2 | l \rangle = \hat{L}_z^2 | l \rangle + \hat{L}_x | l \rangle + \hat{L}_y | l \rangle = l^2 + l = l(l + 1),
\]

since \( | l \rangle \) is the maximum, \( \hat{L}_+ | l \rangle = | 0 \rangle \). Moreover, since \( \hat{L}_- | l \rangle = | l - 1 \rangle \), then

\[
\hat{L}^2 \hat{L}_- | l \rangle \rightarrow \hat{L}_- \hat{L}^2 | l \rangle = l(l + 1) \hat{L}_- | l \rangle = l(l + 1) | l - 1 \rangle
\]

The equation above is true if and only if \( [\hat{L}^2, \hat{L}_-] = 0 \), which can be confirmed by direct calculation. Hence, \( l(l + 1) \) is an eigenvalue for every state of angular momentum of a system. Note that all the states of angular momentum have the same energy under
rotational symmetry, since the angular momentum operator commutes with the Hamiltonian.

11.15 A Particle Moving in a Coulomb Potential

Classically, the Hamiltonian ‘$H$’ for a particle in a Coulomb potential is

$$H = \frac{p^2}{2m} + V(r),$$

where $V(r)$ represents the potential energy. In classical physics, angular momentum ‘$L = \langle r \times |p\rangle$’ is conserved i.e. the magnitude of the angular momentum and the plane of the orbit are conserved.

Figure 11.15-1

Figure 11.15-1 illustrates a particle moving in a coulomb potential. Note that

$$H = \frac{p^2_\theta + p^2_r}{2m} + V(r)$$

Hence, $|L| = |p_\theta||r| \rightarrow L^2 = p^2_\theta r^2 \rightarrow p^2_\theta = L^2/r^2$. Therefore,

$$H = \frac{p^2_r}{2m} + \frac{L^2}{2mr^2} + V(r)$$

In classical physics, angular momentum is conserved. In other words, $L$ is a constant, and hence, $H$ is a function of $r$ only.

Figure 11.15-2
Figure 11.15-2 shows a plot of the potential energy vs. the distance ‘r’ of a particle moving in a coulomb potential. If \( r \ll 1 \), then the centripetal force (repulsive) dominates, shown by the dotted blue line above \( E_\omega = 0 \). If \( r \gg 1 \), then the centripetal force (repulsive) dies out rapidly and \( V(r) \) (attractive force) dominates, shown by the dotted blue line below \( E_\omega = 0 \). The combined potential is illustrated by the red line.

Quantum mechanically, a central force system is described by a wave function ‘\( \Psi(r) \)’. Inserted into Schrödinger’s equation, \( \hat{H} |\Psi(r)\rangle = E |\Psi(r)\rangle \). The momentum operator is defined as \( \hat{P}_r = i \frac{\partial}{\partial r} \), \( \hbar = 1 \). So,

\[
\hat{H} = -\frac{1}{2m} \frac{\partial^2}{\partial r^2} + \frac{1}{2mr^2} + V(r) \quad \rightarrow \quad \hat{H} = -\frac{1}{2m} \frac{\partial^2}{\partial r^2} + \frac{l(l+1)}{2mr^2} + V(r), \quad \hat{L}^2 = l(l+1)
\]

Therefore,

\[
\hat{H} |\Psi(r)\rangle = E |\Psi(r)\rangle \rightarrow \left( -\frac{1}{2m} \frac{\partial^2 |\Psi(r)\rangle}{\partial r^2} + \frac{l(l+1)}{2mr^2} \right) |\Psi(r)\rangle + V(r) |\Psi(r)\rangle = E |\Psi(r)\rangle
\]

Figure 11.15-3 shows schematically the types of \( \Psi(r) \)'s that are solutions to the Schrödinger equation for a particle moving in a Coulomb potential. The \( \Psi(r) \)'s are ‘bound’, which means the orbits are closed. If the energy gets too high, the particle will fly away. The '\( \Psi(r) \)'s' are characterized by the number of nodes (number of wiggles); the more nodes, the more wiggles, the higher the momentum, the higher the energy (see fig. 11.15-3). In essence, the \( \Psi(r) \)'s become little harmonic oscillators.

**Figure 11.15-3**

**11.15.1 Angular Momentum and Energy in a Coulomb Potential**

A hydrogen atom has a single electron bound to the nucleus. If a number of factors are ignored, such as the spin of the electron or that nucleuses are finite, along with many other factors, and if consideration is confined to the energy and angular momentum of the electron, then fig. 11.15-4 demonstrates how the energy and angular momentum are related.
If $l = 0$, then $\Psi_n$ can be in only one possible state at each energy level, since the number of possible states is $2l + 1$; The function $'\Psi_0'$ has no nodes, representing the lowest energy level; the function $'\Psi_1'$ represents a wave function with one node, and has higher energy than $\Psi_0$; the function $'\Psi_2'$ represents a wave function with two nodes and has even higher energy. If $l = 1$, then $\Psi_0$ represents a wave function with no nodes, but since it has non-zero angular momentum, its energy is higher than $\Psi_0$ if $l = 0$. Note that if $l = 1$, $\Psi_0$ has three possible states at each energy level, since the number of states is $2l + 1$. Interestingly, $\Psi_0$ with $l = 1$ has the same energy as $\Psi_1$ with $l = 0$. This is a general pattern. For instance, $\Psi_0$ with $l = 2$, has the same energy as $\Psi_1$ with $l = 1$ and the same energy as $\Psi_2$ with $l = 0$ (see fig. 11.15-4). The levels of degeneracy of energy states can be read off the figure. If $l = 0$ for $\Psi_0$, then there is only one state, the 'ground state'. At the second energy level, there are four states. At the third, nine states and so on.

The analysis above is of an idealized hydrogen atom. However, the ground state with $l = 0$ actually has two possible states instead of one. The second energy level has eight states rather than four. There is something missing in the analysis by a factor of two. The "something missing" is 'electron spin', discussed subsequently.

### 11.16 The Quantum Harmonic Oscillator

A classical harmonic oscillator, a pendulum, for example, or a box sliding back and forth along a frictionless plane, shown in fig 11.16-1, oscillates around a point of equilibrium '0'. The system is ideal, since no external forces act on the system, and, once the system is set in motion, it will oscillate at a certain frequency indefinitely.
The Lagrangian '$\mathcal{L}$' for a classical harmonic oscillator is

$$\mathcal{L} = K - U = \frac{1}{2} m \dot{x}^2 - \frac{m \omega^2}{2} x^2 \rightarrow \mathcal{L} = \frac{1}{2} \dot{x}^2 - \frac{\omega^2}{2} x^2, \quad m = 1, \quad \omega = \sqrt{\frac{k}{m}}$$

where $K$ is the 'kinetic energy', $U$ is the 'potential energy', $m = 1$ and $\omega$ is the 'frequency' of the oscillator, which is proportional to the force required to displace the system from its equilibrium point. Note that the momentum '$p$' is $\dot{x}$. The equations of motion are given by Euler’s law:

$$\frac{d}{dt} \left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right) = \ddot{x} = \frac{\partial \mathcal{L}}{\partial x} = -\omega^2 x \rightarrow \ddot{x} + \omega^2 x = 0,$$

where the general solution is

$$x = \alpha \cos(\omega t) + \beta \sin(\omega t), \quad \alpha, \beta = \text{constants}$$

If $\Psi(x)$ is a wave function such that $\Psi^\dagger \Psi$ is the probability that the particle can be found at position 'x', then the Hermitian operator 'â$H$' acting on $|\Psi\rangle$ is likely related to the classical Hamiltonian i.e.

$$H = p\dot{x} - \mathcal{L} = \dot{x}^2 - \frac{1}{2} \dot{x}^2 + \frac{\omega^2}{2} x^2 = \frac{1}{2} \dot{x}^2 + \frac{\omega^2}{2} x^2 = \frac{1}{2} p^2 + \frac{\omega^2}{2} x^2 = \frac{1}{2} \left( p^2 + \omega^2 x^2 \right)$$

**11.16.1 Quantizing the Energy**

Making $H$ quantum mechanical requires replacing the numbers '$p$' and 'x' by the operators 'â$\hat{p}$' and 'â$\hat{x}$', so that $H \rightarrow \hat{H}$. Moreover, the frequency comes in discrete units of $\hbar$. Therefore, $\omega \rightarrow \hbar \omega$ so that
\[ \hat{H}|\Psi\rangle = \frac{1}{2} (\hat{p}^2 + (\hbar \omega)^2 \hat{X}^2)|\Psi\rangle = \frac{1}{2} \left( -i\hbar \frac{\partial}{\partial x} \right) \left( -i\hbar \frac{\partial}{\partial x} \right) |\Psi\rangle + \frac{\hbar^2 \omega^2}{2} x^2 |\Psi\rangle \]

\[ = -\frac{\hbar^2}{2} \frac{\partial^2}{\partial x^2} + \frac{\hbar^2 \omega^2}{2} x^2 |\Psi\rangle \]

\[ \hat{H} = -\frac{\hbar^2}{2} \frac{\partial^2}{\partial x^2} + \frac{\hbar^2 \omega^2}{2} x^2 \]

Since, \( \hat{X} |\Psi\rangle = x|\Psi\rangle \). Schrödinger’s equation becomes

\[ i\hbar \frac{\partial |\Psi\rangle}{\partial t} = \hat{H}|\Psi\rangle \rightarrow \frac{\partial |\Psi\rangle}{\partial t} = -\frac{\hbar \partial^2 |\Psi\rangle}{2 \partial x^2} + \frac{\hbar \omega^2}{2} x^2 |\Psi\rangle \rightarrow \frac{\hat{H}}{\hbar} = -\frac{\hbar}{2} \frac{\partial^2 |\Psi\rangle}{\partial x^2} + \frac{\hbar \omega^2}{2} x^2 |\Psi\rangle \]

Since \( \hat{H} \) represents the energy of the system,

\[ \frac{\hat{H}}{\hbar} |\Psi_E\rangle = E |\Psi_E\rangle \rightarrow \hat{H} |\Psi_E\rangle = \hbar E |\Psi_E\rangle \]

where \( \hbar E \) is the eigenvalue of \( \hat{H} \), suggesting the energy ‘\( E \)’ comes in discrete units of \( \hbar \). Hence,

\[ \hat{H} |\Psi\rangle = \hbar E |\Psi\rangle \rightarrow \left( -\frac{\hbar}{2} \frac{\partial^2}{\partial x^2} + \frac{\hbar \omega^2}{2} x^2 \right) |\Psi\rangle = E |\Psi\rangle \]

The equation above has two solutions i.e. \( |\Psi\rangle = e^{\alpha x^2} \), \( |\Psi\rangle = e^{-\alpha x^2} \), \( \alpha = a \) a constant.

The first solution is impossible. To see this, suppose \( \alpha = a + ib \), then

\[ |\Psi\rangle = e^{(a+ib)x^2} \quad |\Psi^\dagger\rangle = e^{(a-ib)x^2} \rightarrow \langle \Psi^\dagger |\Psi\rangle = \int |\Psi^\dagger\rangle |\Psi\rangle d\chi = \int e^{2ax^2} dx = \infty \]

In this case, the probability that the system has energy ‘\( E \)’ is infinite. But quantum mechanics requires that \( \Psi \) must be square integrable i.e. \( \langle \Psi^\dagger |\Psi\rangle < \infty \). Let

\[ |\Psi\rangle = e^{-\frac{\omega}{2} x^2} \quad \alpha = \frac{\omega}{2} \]

then

\[ \left( -\frac{\hbar}{2} \frac{\partial^2}{\partial x^2} + \frac{\hbar \omega^2}{2} x^2 \right) |\Psi\rangle = E |\Psi\rangle \rightarrow -\frac{\hbar}{2} \left( \frac{\partial^2}{\partial x^2} \right) |\Psi\rangle + \frac{\hbar \omega^2}{2} \frac{\partial^2}{\partial x^2} |\Psi\rangle = E |\Psi\rangle \]

\[ \rightarrow -\frac{\hbar}{2} \left( -\omega e^{-\frac{\omega}{2} x^2} + \omega^2 x^2 e^{-\frac{\omega}{2} x^2} \right) + \frac{\hbar \omega^2}{2} x^2 e^{-\frac{\omega}{2} x^2} = E e^{-\frac{\omega}{2} x^2} \rightarrow \frac{\hbar \omega}{2} e^{-\frac{\omega}{2} x^2} \]

\[ = E e^{-\frac{\omega}{2} x^2} \rightarrow E = \frac{\hbar \omega}{2} \]

The quantity ‘\( \hbar \omega/2 \)’ is an eigenvalue of \( \hat{H} \), and hence, a possible energy state of the system. It is also the lowest energy state i.e. the ‘ground state’. As predicted, the energy is an integer multiple of \( \hbar \). Interestingly, in classical systems, the ground state or ‘vacuum state’ is zero. There is no energy in a classical vacuum. But in quantum systems, the ground state can never be zero i.e. \( \omega \neq 0 \). This is a consequence of the Heisenberg uncertainty principle.
11.16.2 Raising and Lowering Operators for Energy

Recall that

\[ \hat{H} = \frac{1}{2} \left( \hat{p}^2 + (\omega \hbar)^2 \hat{X}^2 \right) = \frac{1}{2} \left( \hat{\hat{p}}^2 + (\omega \hbar)^2 \hat{X}^2 \right) = \frac{1}{2} \left( \hat{\hat{p}}^2 + (\omega \hbar)^2 \hat{X}^2 + i \omega \hbar [\hat{X}, \hat{p}] \right) = \frac{1}{2} \left( \hat{\hat{p}}^2 + (\omega \hbar)^2 \hat{X}^2 + i \omega \hbar \hbar \right) \]

since \( \hat{X} \hat{p} - \hat{p} \hat{X} = [\hat{X}, \hat{p}] = i \hbar \). Hence,

\[ \hat{H} = \frac{1}{2} \left( \hat{\hat{p}}^2 + (\omega \hbar)^2 \hat{X}^2 \right) = \frac{1}{2} \left( \hat{\hat{p}}^2 + (\omega \hbar)^2 \hat{X}^2 - \hbar^2 \omega \right) + \frac{\hbar^2 \omega}{2} \to \hat{H} = \frac{1}{2} \left( \hat{\hat{p}}^2 + \omega^2 \hat{X}^2 - \omega \right) + \frac{\omega}{2}, \]

\( \hbar = 1 \)

Now

\[ \hat{H} = \frac{1}{2} \left( \frac{(\hat{\hat{p}} + i \omega \hat{X}) (\hat{\hat{p}} - i \omega \hat{X})}{\hat{b}^+} \right) \to (\hat{\hat{p}} - i \omega \hat{X}) |\Psi\rangle = \left( \hat{\hat{p}} - i \omega \hat{X} \right) e^{-\frac{\omega x^2}{2}} = -i \left( \frac{\partial}{\partial x} + \omega x \right) e^{-\frac{\omega x^2}{2}} \]

Note that \( \hat{b}^- = \hat{\hat{p}} - i \omega \hat{X} \) acting on \(|\Psi\rangle = e^{-\omega x^2/2}\) annihilates it. This suggests that

\[ |\Psi_0\rangle = |g\rangle = e^{-\omega x^2/2} \to \hat{b}^- |g\rangle = |0\rangle, \]

where \(|g\rangle\) is the ‘ground’ or ‘lowest’ energy state.

Let \( \hat{b}^+ = \hat{\hat{p}} + i \omega \hat{X} \) and \( \hat{b}^- = \hat{\hat{p}} - i \omega \hat{X} \), then

\[ [\hat{b}^+, \hat{b}^-] = [\hat{\hat{p}} + i \omega \hat{X}, \hat{\hat{p}} - i \omega \hat{X}] = -2 \omega \]

To see this,

\[ [\hat{\hat{p}} + i \omega \hat{X}, \hat{\hat{p}} - i \omega \hat{X}] = [\hat{\hat{p}}, \hat{\hat{p}}] - \omega^2 [\hat{X}, \hat{X}] + i \omega [\hat{X}, \hat{p}] - i \omega [\hat{p}, \hat{X}] = i \omega i + i \omega i = -2 \omega \]

Let \( \hat{b}^+ / \sqrt{2 \omega} = \hat{a}^+ \) and \( \hat{b}^- / \sqrt{2 \omega} = \hat{a}^- \), then \([\hat{a}^+, \hat{a}^-] = -1\).

Hence,

\[ \hat{H} = \frac{1}{2} \left( \frac{(\hat{\hat{p}} + i \omega \hat{X}) (\hat{\hat{p}} - i \omega \hat{X})}{\hat{b}^+} \right) + \frac{\omega}{2} = \frac{\hat{b}^+ \hat{b}^-}{2} + \frac{\omega}{2} = \frac{1}{2} (\hat{a}^+ \sqrt{2 \omega} \hat{a}^- \sqrt{2 \omega}) + \frac{\omega}{2} \]

\[ = \frac{1}{2} (2 \omega \hat{a}^+ \hat{a}^-) + \frac{\omega}{2} \to \hat{H} = \omega \hat{a}^+ \hat{a}^- + \frac{\omega}{2} \]

307
The energy levels for the operator $\hat{H}$ are the eigenvalues of $\hat{a}^+, \hat{a}^-$ given that $[\hat{a}^+, \hat{a}^-] = -1$.

Since $[\hat{a}^+, \hat{a}^-] = -1$, let $\hat{a}^+ \hat{a}^- |n\rangle = n|n\rangle$. Note that

$$[\hat{a}^+, \hat{a}^-]|n\rangle = -|n\rangle, \quad [\hat{a}^-, \hat{a}^+]|n\rangle = |n\rangle \rightarrow \hat{a}^+(\hat{a}^- \hat{a}^- - \hat{a}^- \hat{a}^+)|n\rangle = \hat{a}^+|n\rangle \rightarrow (\hat{a}^+ \hat{a}^- \hat{a}^- - \hat{a}^+ \hat{a}^- \hat{a}^-)|n\rangle = \hat{a}^+|n\rangle$$

But, $\hat{a}^+ \hat{a}^- |n\rangle = n|n\rangle$, hence

$$(\hat{a}^+ \hat{a}^- \hat{a}^+ - \hat{a}^+ \hat{a}^- \hat{a}^-)|n\rangle = \hat{a}^+|n\rangle \rightarrow \hat{a}^+ \hat{a}^- \hat{a}^-|n\rangle - \hat{a}^+ \hat{a}^- \hat{a}^-|n\rangle = \hat{a}^+|n\rangle \rightarrow \hat{a}^+ \hat{a}^- \hat{a}^-|n\rangle = n\hat{a}^+|n\rangle + \hat{a}^+|n\rangle$$

Recalling that $\hat{H} = \omega \hat{a}^+ \hat{a}^- + \omega / 2$, let $\hat{H}' = \hat{H} - \omega / 2$. Hence, $\hat{H}' = \omega \hat{a}^+ \hat{a}^-$. 

$$\hat{a}^+ \hat{a}^- \hat{a}^+|n\rangle = (n + 1)\hat{a}^+|n\rangle \rightarrow \omega \hat{a}^+ \hat{a}^- \hat{a}^+|n\rangle = (n + 1)\omega \hat{a}^+|n\rangle \rightarrow \hat{H}' \hat{a}^+|n\rangle = (n + 1)\omega \hat{a}^+|n\rangle$$

$$\rightarrow \hat{H}' \hat{a}^+|n\rangle = \left[(n + 1)\omega + \frac{\omega}{2}\right] \hat{a}^+|n\rangle = \left(n + 1\right)\omega \hat{a}^+|n\rangle$$

The operator $\hat{a}^+$ raises the energy level by $\omega \hbar$, and conversely, $\hat{a}^-$ lowers the energy level by $\omega \hbar$. As such, $\hat{a}^+, \hat{a}^-$ are called ‘raising’ and ‘lowering’ operators respectfully.

Since $\hat{a}^+$ raises the energy level, the energy of a system, theoretically, can be raised indefinitely. But applying $\hat{a}^-$, since the harmonic oscillator operator $\hat{H}'$ contains only squares of momentum and position, the energy must be positive definite or, at least, not negative. Hence, $\hat{a}^-|g\rangle = |0\rangle$, $|g\rangle$ the ground state.

### 11.16.2.1 Normalizing Raising and Lowering Energy Operators

Of note, the operators $\hat{a}^+$ and $\hat{a}^-$ have not been normalized. To be mathematically consistent, $\hat{a}^+|n\rangle = \sqrt{n + 1}|n + 1\rangle$ and $\hat{a}^-|n\rangle = \sqrt{n}|n - 1\rangle$. To see this, intuitively,

$$\hat{a}^+|n\rangle \rightarrow |n + 1\rangle \rightarrow \hat{a}^+|g\rangle \rightarrow |g + 1\rangle, \quad \hat{a}^-|n\rangle \rightarrow |n - 1\rangle \rightarrow \hat{a}^-|g\rangle \rightarrow |g - 1\rangle$$

But this is wrong, since

$$\hat{a}^+ \hat{a}^-|n\rangle = \hat{a}^+|n - 1\rangle = |n\rangle, \quad \hat{a}^- \hat{a}^+|n\rangle = \hat{a}^-|n + 1\rangle = |n\rangle$$

implies that $\hat{a}^+ \hat{a}^- - \hat{a}^- \hat{a}^+ = [\hat{a}^+, \hat{a}^-] = 0$. But $[\hat{a}^+, \hat{a}^-] = -1$.

Now consider

$$\hat{a}^+|n\rangle = \sqrt{n + 1}|n + 1\rangle, \quad \hat{a}^-|n\rangle = \sqrt{n}|n - 1\rangle,$$

then

$$\hat{a}^+ \hat{a}^-|n\rangle = \sqrt{n} \hat{a}^+ |n - 1\rangle = \sqrt{n} \sqrt{n}|n\rangle = n|n\rangle,$$

$$\hat{a}^- \hat{a}^+|n\rangle = \sqrt{n + 1} \hat{a}^- |n + 1\rangle = \sqrt{n + 1} \sqrt{n + 1}|n\rangle = (n + 1)|n\rangle$$
Hence,

\[ \hat{a}^+ \hat{a}^- - \hat{a}^- \hat{a}^+ = [\hat{a}^+, \hat{a}^-] = n - (n + 1) = -1, \]

which is the correct commutation relationship.

Since \( |\Psi_0\rangle = |g\rangle = e^{-\omega x^2/2} \), then

\[
|\Psi_1\rangle = \hat{a}^+ |\Psi_0\rangle = \frac{1}{\sqrt{2\omega}} (\hat{p} + i\omega x) |\Psi_0\rangle = \frac{1}{\sqrt{2\omega}} (-i \frac{\partial}{\partial x} + i\omega x) e^{-\frac{1}{2} \omega x^2} = \frac{2i\omega x e^{-\frac{1}{2} \omega x^2}}{\sqrt{2\omega}}
\]

Note that \( |\Psi_0\rangle \) has no nodes, whereas \( |\Psi_1\rangle \) has one node. In fact, each time \( \hat{a}^+ \) acts on \( |\Psi_n\rangle \), it adds a node, until no more nodes can be added (see fig. 11.15-3). Therefore, \( |\Psi_n\rangle \) represents a wave function with more energy than \( |\Psi_{n-1}\rangle \).

Importantly, although \( \Psi(x, t) \) is a continuous function, the energy levels of quantum mechanical harmonic oscillators are discrete. Only certain values of energy are achievable.

**11.17 Classical vs. Quantum Logic via Electron Spin**

One of the best ways to explain the difference between classical and quantum logic is in terms of information. A piece of classical information is called a ‘bit’. All bits have two states: ‘up’ or ‘down’, ‘left’ or ‘right’, ‘heads’ or ‘tails’, ‘on’ or ‘off’ etc. The states of a bit can be denoted ‘|0⟩, |1⟩’. A bit can be in one state or the other, but not both at the same time. Lining up several bits can give a large amount of information. For instance, \( |001011\rangle \) has six bits; the first two are in state ‘|0⟩’ and the third is in state ‘|1⟩’ and so forth. Since a bit can be in two states, the number ‘\( N \)’ of configurations is \( 2^n \). Hence,

\[ N = 2^n \rightarrow \log_2 N = \log_2 2^n = n \]

The idea of a bit can be generalized. Suppose the fundamental piece of information is the ‘die-bit’. A die-bit has six states. The number of die-bit configurations would be \( N = 6^n \rightarrow \log_6 N = \log_6 6^n = n \). Any string of bits can represent a number or at least an approximation to a number.

Theoretical physics is an exercise in 1) determining the configuration of the system at any moment in time, 2) determining how the configuration of the system is updated over time. In most cases, the configuration of the system is simply an exercise in determining the configuration of bits. Any concept in physics, if its quantitative representation does not need to be too precise, can be thought of in terms of bits.
Fig. 11.17-1 shows a simple system. It has four states ‘A, B, C, D’. The arrows represent updates to the system over time. If the system starts at A, in the next instant, it goes to B, then C, then D. If it starts at B, it goes to C, then D, then B etc. In this system, information is lost. If the system is in the state ‘B’, it is not possible to determine the state of the system during the previous time period. It could have been in state ‘A’ or state ‘D’. No classical system is configured in the way illustrated by fig. 11.17-1. Classical systems do not lose information. It is always possible, if the system’s current configuration is known, to determine the configuration in the previous time period. This is also true of quantum systems.

Consider the simple system, shown in fig. 11.17-2, with time sequence ‘A, B, C, D, A’ etc. Suppose $H|A\rangle = |B\rangle$. If $H$ is to describe the time evolution of the system, then

$$H|A\rangle = |B\rangle = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = |B\rangle,$$

$$|A\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow H|B\rangle = |C\rangle = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = |C\rangle,$$

and so on. Each time the matrix ‘$H$’ operates on a state, it moves the system to the next state; the matrix ‘$H^{-1}$’ would update the system in the opposite direction. The operator ‘$H$’ shows how the time evolution of classical states can be described. In general,

$$H|A\rangle = |B\rangle, \quad H|B\rangle = |C\rangle \rightarrow H^2|A\rangle = |C\rangle$$

and so on. The status of a system after $n$ time periods is $H^n|A\rangle$.  

310
The system shown in fig. 11.17-2 is deterministic. The initial state of the system determines the next state and so on. A system can also be updated randomly. Suppose, on the flip of a fair coin, if the coin turns up heads, $H$ acts on the system, but does not act if the coin shows tails. The time sequence of the system is now unpredictable. With each flip of the coin, there is a 50% chance of landing heads and a 50% chance of landing tails. In this case, the time sequence of the system cannot be predicted with certainty. But classical physics assumes that with unlimited information about the coin and its environment, theoretically, it is possible to predict whether or not the coin would land heads or tails on any given flip. Probability in classical physics arises because of a lack of information. However, quantum physics considers probability fundamental. Even with an infinite degree of information about a quantum system, predicting when certain events occur is probabilistic.

### 11.17.1 The Qubit and Electron Spin

What is the difference between a ‘bit’ of classic information and a ‘qubit’ of quantum information? Electrons are little magnets with a characteristic magnetic moment, generally referred to as ‘spin’. Spin can be conceptualized as an arrow pointing in a certain direction.

Figure 11.17.1-1 shows an imaginary electron placed between the poles of an electromagnet. The arrow represents the spin direction of the electron. The ‘$N$’, at the end of the arrow, represents the north pole of the electron’s magnetic field. If the electromagnet is turned on, the spin direction of the electron will precess, eventually aligning with the south pole of the electromagnet. As it transitions, the electron gives off energy in the form of radiation.

According to classical physics, the amount of energy given off is proportional to the angle ‘$\theta$’, representing the ‘spin direction’. If $\theta = -\pi/2$, the electron gives off the maximum amount of energy. If $\theta = \pi/2$, the electron gives off no energy. Classical physics predicts that as $\theta = -\pi/2 \rightarrow \theta = \pi/2$, the amount of energy given off by the
electron will continuously decrease until reaching $\theta = \pi/2$, where it drops to zero. But if spin is actually measured, the results are different.

**11.17.1.1 Electron Spin is Quantized**

If an electron of a certain energy is placed between the poles of the electromagnet, regardless of its initial spin direction, it will emit one photon of energy or it will not. The spin of an electron has only two states — ‘up’ or ‘down’, nothing in between. The electron either emits one photon of energy or it does not. If it emits a photon, its spin is in the ‘down’ state. If not, its spin is in the ‘up’ state. Those are the only two states.

**11.17.2 Information in Quantum Physics is Probabilistic**

However, electron spin can be prepared in some other state than vertical (up or down). The only way of predicting what happens when an electron is placed in a magnetic field is by giving a probability that the electron will emit a photon. If the electron is prepared in the horizontal direction, then placed in a vertical magnetic field, the probability it will give off a photon is $1/2$. There is no way of predicting with certainty whether or not an electron will emit a photon. No amount of information about an electron will allow a certain prediction of photon emission, unless, of course, the electron is prepared in the ‘down’ state, then it will emit a photon with certainty. An electron can be prepared in any spin direction, but, when measured, will only be found in the ‘up’ or ‘down’ state. The only information about how the electron was initially prepared is given by a probability.

**11.17.3 Measuring Electron Spin**

Electron spin can be measured along any axis, but only two values along that axis are possible: ‘spin-up’ (+1) or ‘spin-down’ (−1). Quantum mechanically, spin is measured by first defining an operator that measures spin along an axis. The operator acts on an eigenvector in that direction and produces an eigenvalue of ‘1’ or ‘−1’ (spin-up or spin-down). For instance, the operators measuring spin along the $x, y, z$ axes are

$$
\hat{\sigma}_z = \begin{bmatrix}
1 & 0 \\
0 & -1
\end{bmatrix}, \quad \hat{\sigma}_x = \begin{bmatrix}
0 & 1 \\
1 & 0
\end{bmatrix}, \quad \hat{\sigma}_y = \begin{bmatrix}
0 & -i \\
i & 0
\end{bmatrix} = i \begin{bmatrix}
0 & -1 \\
1 & 0
\end{bmatrix}
$$

Their corresponding eigenvectors are respectfully:

$$
|\text{up}\rangle_z = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad |\text{down}\rangle_z = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad |\text{up}\rangle_x = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad |\text{down}\rangle_x = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix},
$$

$$
|\text{up}\rangle_y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad |\text{down}\rangle_y = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -i \end{bmatrix}
$$

For example,

$$
\hat{\sigma}_z |\text{up}\rangle_z = \begin{bmatrix}
1 & 0 \\
0 & -1
\end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{\sigma}_z |\text{down}\rangle_z = \begin{bmatrix}
1 & 0 \\
0 & -1
\end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = - \begin{bmatrix} 0 \\ 1 \end{bmatrix}
$$
The eigenvalues of $\hat{\sigma}_z$ are $\pm 1$ as expected. This is also true of $\hat{\sigma}_x$ and $\hat{\sigma}_y$, which can be easily checked. Moreover, $\hat{\sigma}_z^2 = \hat{\sigma}_x^2 = \hat{\sigma}_y^2 = I$. This confirms that the eigenvalues of $\hat{\sigma}_z$, $\hat{\sigma}_x$ and $\hat{\sigma}_y$ can only be 1 or $-1$ i.e. spin-up or spin-down.

11.17.3 Spin Operators

Spin is a kind of angular momentum. Recall that $[\hat{L}_z, \hat{L}_x] = i\hat{L}_y$, where $\hat{L}_i$'s represent the orbital angular momentum of a system. Now

$$[\hat{\sigma}_z, \hat{\sigma}_x] = \hat{\sigma}_x \hat{\sigma}_z - \hat{\sigma}_z \hat{\sigma}_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = 2i \hat{\sigma}_y$$

Let $\hat{S}_i = \hat{\sigma}_i/2$, $i = x, y, z$, then

$$\left[ \frac{\hat{\sigma}_z}{2}, \frac{\hat{\sigma}_x}{2} \right] = \left[ \hat{S}_z, \hat{S}_x \right] = i \hat{S}_y$$

Similarly, $[\hat{S}_x, \hat{S}_y] = i \hat{S}_z$ and $[\hat{S}_y, \hat{S}_z] = i \hat{S}_x$. Note that the spin commutation relations are identical in form to the ones for orbital angular momentum. Moreover,

$$\hat{S}_z = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \hat{S}_x = \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \hat{S}_y = \frac{1}{2} \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$$

11.17.3.2 The Pauli Exclusion Principle

If the orbital angular momentum of a hydrogen atom is designated 'l', then $|E_0\rangle$ is called the ‘ground’ state energy if $l = 0$ and is the only possible state for the electron. A helium atom has two electrons. If $l = 0$, one of the electrons could be in the state $|E_0\rangle$, but the other must be in state $|E_1\rangle$, since only one electron is allowed in the state $|E_0\rangle$ with $l = 0$. But experiments had shown that helium could have both electrons in the state $|E_0\rangle$ with $l = 0$. This suggested that the electron possessed some property in addition to orbital angular momentum. Of course, the other property is ‘spin’. Two electrons can occupy the same state if one of the electrons is spin-up and the other spin-down, a rule known as the ‘Pauli exclusion principle’. In the $|E_0\rangle$, $l = 0$’ state, there can be two electrons. In the $|E_1\rangle$ state, there can be eight electrons, since there are four possible states.

11.17.3.2.1 Fermions and Bosons

If many characteristics of an atom are ignored (interaction between the electrons orbiting the nucleus amongst other things), the Pauli exclusion principle became a rationale for the periodic table of the elements, where the number of electrons at each energy level in an atom is explained by the Pauli exclusion principle. While the periodic table is only approximate, the Pauli exclusion principle is fundamental in describing the two types of particles found in Nature: ‘fermions’ and ‘bosons’. All fermions are spin-1/2 particles that obey the Pauli exclusion principle. The term ‘spin-1/2’ is a general term, indicating a particle with 1/2 integer spin. All bosons have integer spin and do not obey the Pauli exclusion principle.
There can be multi-particle states. For instance, $|x_1, x_2\rangle$ represents a state where particle ‘1’ is at position ‘$x_1$’ and particle ‘2’ is at position ‘$x_2$’. Let ‘$\hat{S}$’ be an operator that switches the position of the two particles i.e. $\hat{S}|x_1, x_2\rangle = |x_2, x_1\rangle$. Note that $\hat{S}^2 = I$. In other words, $\hat{S}^2 |x_1, x_2\rangle = |x_1, x_2\rangle$. The operator ‘$\hat{S}$’ is unitary and its eigenvalues are $\pm 1$. In general, $(x_i | \Psi) = \Psi(x_i) = \Psi(x_1, x_2, ..., x_n)$, where $\Psi^\dagger$ is the probability of finding particle ‘1’ at position ‘$x_1$’ and particle ‘2’ at position ‘$x_2$’ etc.

If $\hat{S}|x_1, x_2\rangle = +1|x_2, x_1\rangle$, then the particle is a ‘boson’. If $\hat{S}|x_1, x_2\rangle = -1|x_2, x_1\rangle$, the particle is a ‘fermion’. Suppose there are two particles ‘$\Psi_0(x_1), \Psi_0(x_2)$’ in the ground state. If

$$
\Psi_0(x_1)\Psi_0(x_2) = \Psi_0(x_2)\Psi_0(x_1) \rightarrow \Psi_0(x_1)\Psi_0(x_2) - \Psi_0(x_2)\Psi_0(x_1) = 0,
$$

the two particles are bosons (here spin has been ignored). Bosons can be in the same state at the same time, but fermions cannot. The equation

$$
\Psi_0(x_1)\Psi_1(x_2) = \Psi_0(x_2)\Psi_1(x_1)
$$

is false for both bosons and fermions, since a particle cannot to be in two different states at the same time. In quantum mechanics, the statement

$$
\Psi_0(x_1)\Psi_1(x_2) + \Psi_0(x_2)\Psi_1(x_1),
$$

which says particle ‘1’ is in the ground state and particle ‘2’ is in state ‘1’ or particle ‘2’ is in the ground state and particle ‘1’ is in state ‘1’, is valid. Such a statement can only apply to a boson, since it is possible for two bosons to occupy the same state at the same time.

The configuration of states

$$
\Psi_0(x_1)\Psi_1(x_2) - \Psi_0(x_2)\Psi_1(x_1)
$$

is valid, but represents a fermion, since it is possible to have particle ‘1’ in the ground state and particle ‘2’ in state ‘1’, but not possible, at the same time, to have particle ‘2’ in the ground state and particle ‘1’ in state ‘1’. Fermions cannot occupy the same state at the same time.

In general, if

$$
\Psi(x_1, x_2, ... x_i, x_j, ... , x_n) = \Psi(x_1, x_2, ... x_i, x_i, ... , x_n),
$$

the particle is a ‘boson’. If

$$
\Psi(x_1, x_2, ... x_i, x_j, ... , x_n) = -\Psi(x_1, x_2, ... x_j, x_i, ... , x_n),
$$

the particle is a ‘fermion’. Keep in mind that the $\Psi$’s are the wave functions of a particle. If spin is included, the wave function becomes a function of position and spin. In general,

$$
\Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2, ... \hat{\sigma}_i x_i, \hat{\sigma}_j x_j, ... , \hat{\sigma}_n x_n) = \Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2, ... \hat{\sigma}_j x_j, \hat{\sigma}_i x_i, ... , \hat{\sigma}_n x_n)
$$
for a boson and
\[ \Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2, \ldots \hat{\sigma}_i x_i, \hat{\sigma}_j x_j, \ldots, \hat{\sigma}_n x_n) = -\Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2, \ldots \hat{\sigma}_j x_j, \hat{\sigma}_i x_i, \ldots, \hat{\sigma}_n x_n), \]
for a fermion.

Consider a wave function \( \Psi(x, \hat{\sigma}) \) of position and spin and let \( \hat{J}_z = \hat{L}_z + \hat{S}_z \). Since \( \hat{J}_z \) is angular momentum,
\[ \hat{J}_z |\Psi\rangle = -i \frac{\partial |\Psi\rangle}{\partial \theta}, \]
where \( \theta \) is an angle. Let \( \hat{J}_z |\Psi\rangle = m |\Psi\rangle \), where \( m \) is the eigenvalue of \( \hat{J}_z \). A solution to the equation above is \( |\Psi(\theta)\rangle = e^{im\theta} \). Since \( \hat{J}_z \) is an angular momentum operator, \( m \) must be integer or half-integer. If \( \theta = 2\pi \), intuitively, \( \hat{J}_z |\Psi\rangle = |\Psi\rangle \). There should no effect on the wave function. But, if \( m = 1/2 \), then
\[ e^{im2\pi} = e^{i\pi} = \cos \pi + i \sin \pi = -1 \]
In this case, \( \hat{J}_z |\Psi\rangle = -|\Psi\rangle \). Note that if \( m = 1/2 \), the particle is a fermion. And if \( m = 1 \), the particle is a boson.

Suppose there is a wave function \( \Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) \) where the angular momentum of one of the particles is rotated by \( 2\pi \) and then the particles exchanged, given that the two particles are fermions. If \( \hat{\sigma}_2 \) is rotated by \( 2\pi \), then
\[ \Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) \rightarrow \Psi(\hat{\sigma}_1 x_1, -\hat{\sigma}_2 x_2) = -\Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) \]
If the two particles are exchanged, then
\[ -\Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) \rightarrow -\Psi(-\hat{\sigma}_2 x_2, \hat{\sigma}_1 x_1) = -\Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) = \Psi(\hat{\sigma}_1 x_1, \hat{\sigma}_2 x_2) \]
Hence, a rotation of one of the spins by \( 2\pi \) and then an exchange leaves the state unchanged. The same can be said of bosons.

11.17.4 Measuring Spin in an Arbitrary Direction

Suppose an electron is prepared spin-up in the \( z \)-direction i.e. let \( |b\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \). What is the probability \( 'P_x(up)' \) if spin is measured in the \( x \)-direction, it is found ‘spin-up’? In other words, what is the probability the spin will be found in the state
\[ |a\rangle = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \end{bmatrix} \]

315
The answer, as always, is

\[ P_x(\text{up}) = \langle a|b\rangle\langle b|a \rangle = \left(\frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \left(\begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \end{bmatrix}\right) = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} \]

Since there are only two possibilities i.e. spin-up or spin-down, then \( P_x(\text{down}) = \frac{1}{2} \). In general, if the electron is prepared in a certain direction and then measured in a perpendicular direction, the probability that the spin is up or down is \( \frac{1}{2} \). Remember that

\[ [\hat{\sigma}_i, \hat{\sigma}_j] \neq 0, \quad i \neq j, \quad i, j = x, y, z \]

i.e. the \( x, y, z \)-spin operators do not commute. Obtaining exact spins measurements along perpendicular axes simultaneously is not possible. In other words, if an electron's spin is measured along the \( z \)-axis and found spin-up, it is impossible to know simultaneously whether it is spin-up or spin-down along the \( x \)-axis.

Suppose there is an arbitrary unit vector \( \langle n \rangle \) which has components

\[ |n\rangle = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix}, \quad ||n|| = 1 \]

The components of spin along an arbitrary \( n \)-axis are

\[ \sum_i \hat{\sigma}_i n_i, \quad i = x, y, z \]

The \( \hat{\sigma}_i \)'s are matrices and the \( n_i \)'s are just numbers, hence

\[ \sum_i \hat{\sigma}_i n_i = \begin{bmatrix} 0 & n_1 & \hat{\sigma}_1 n_1 \\ n_1 & 0 & \hat{\sigma}_2 n_1 \\ \hat{\sigma}_1 n_2 & \hat{\sigma}_2 n_2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -in_2 & \hat{\sigma}_3 n_2 \\ in_2 & 0 & \hat{\sigma}_1 n_2 \\ \hat{\sigma}_3 n_2 & \hat{\sigma}_1 n_2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \hat{\sigma}_3 n_3 \\ 0 & 0 & \hat{\sigma}_1 n_3 \\ \hat{\sigma}_3 n_3 & \hat{\sigma}_1 n_3 & 0 \end{bmatrix} = \hat{\sigma}_{\langle n \rangle} \]

Note that the matrix on the right above is Hermitian. Hence, it is a candidate for an operator in quantum mechanics and is, in fact, the operator that measures electron spin in the \( n \)-direction. Note that \( \hat{\sigma}_{\langle n \rangle}^2 = I \). To see this

\[ \left( \sum_i \hat{\sigma}_i n_i \right)^2 = (\hat{\sigma}_1 n_1 + \hat{\sigma}_2 n_2 + \hat{\sigma}_3 n_3)(\hat{\sigma}_1 n_1 + \hat{\sigma}_2 n_2 + \hat{\sigma}_3 n_3) \]

\[ = \hat{\sigma}_1^2 n_1^2 + \hat{\sigma}_2^2 n_2^2 + \hat{\sigma}_3^2 n_3^2 + n_1 n_2 (\hat{\sigma}_1 \hat{\sigma}_2 + \hat{\sigma}_2 \hat{\sigma}_1) + n_1 n_3 (\hat{\sigma}_1 \hat{\sigma}_3 + \hat{\sigma}_3 \hat{\sigma}_1) + n_2 n_3 (\hat{\sigma}_2 \hat{\sigma}_3 + \hat{\sigma}_3 \hat{\sigma}_2) \]

However, the \( \hat{\sigma}_i \)'s are operators, which anti-commute i.e. \( \hat{\sigma}_i \hat{\sigma}_j = -\hat{\sigma}_j \hat{\sigma}_i, \quad i \neq j \). Therefore, the last three terms in the expression on the right-side above vanish and this leaves
\[
\left( \sum_i \hat{\sigma}_i n_i \right)^2 = \hat{\sigma}_1^2 n_1^2 + \hat{\sigma}_2^2 n_2^2 + \hat{\sigma}_3^2 n_3^2
\]

But, \( \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2 = 1 \). Hence,
\[
\left( \sum_i \hat{\sigma}_i n_i \right)^2 = (n_1^2 + n_2^2 + n_3^2) I = I,
\]
since \( n_1^2 + n_2^2 + n_3^2 = 1 \).

There can be only two eigenvalues in the \( n \)-direction, ‘1’ and ‘−1’ i.e. ‘spin-up’ or ‘spin-down’. In general, if an electron is prepared spin-up along the \( n \)-direction, what is the probability its spin is up (or down) along an arbitrary \( m \)-direction? Let
\[
|\hat{\sigma}_i n_i = 1\rangle \rightarrow \hat{\sigma} \cdot n |\hat{\sigma}_i n_i = 1\rangle = |\hat{\sigma}_i n_i = 1\rangle,
\]
such that if \( \hat{\sigma} \cdot n \) operates on \( |\hat{\sigma}_i n_i = 1\rangle \), the eigenvalue is 1. Similarly, \( \hat{\sigma} \cdot m |\hat{\sigma}_i m_i = 1\rangle = |\hat{\sigma}_i m_i = 1\rangle \) is spin-up along the \( m \)-direction. In other words, if the electron is prepared spin-up along the \( n \)-direction, then measured along the \( m \)-direction, the probability it is spin-up along \( m \)-direction is
\[
|\langle \hat{\sigma}_i m_i = 1 |\hat{\sigma}_i n_i = 1\rangle|^2
\]
The first step is to find a vector \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \) such that
\[
\begin{bmatrix} n_3 \\ n_1 + in_2 \\ -n_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}
\]
There is no loss in generality if \( \alpha \) is set to 1 and, for convenience, let \( \beta = z \), then
\[
\begin{bmatrix} n_3 \\ n_1 + in_2 \\ -n_3 \end{bmatrix} \begin{bmatrix} 1 \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ z \end{bmatrix}
\]
Note that \( n_3 + n_-z = 1 \), where \( n_- = n_1 - in_2 \), then \( z = (1 - n_3)/n_- \). So
\[
\begin{bmatrix} n_3 \\ n_+ \\ n_- \end{bmatrix} \begin{bmatrix} 1 \\ n_- \end{bmatrix} = \begin{bmatrix} 1 \\ n_- \end{bmatrix}, \quad n_+ = n_1 + in_2
\]
Normalizing \( \begin{bmatrix} 1 \\ z \end{bmatrix} \) gives
\[
\begin{bmatrix} 1 \\ n_+ \end{bmatrix} \begin{bmatrix} 1 \\ n_- \end{bmatrix} = 1 + (1 - n_3)^2 = \frac{n_+ n_-}{n_+ n_-} + \frac{1 + n_3^2 - 2n_3}{1 - n_3^2} = \frac{1 - n_3^2}{1 - n_3^2} + \frac{1 + n_3^2 - 2n_3}{1 - n_3^2}
\]
\[
= \frac{2 - 2n_3}{(1 - n_3)(1 + n_3)} = \frac{2}{1 + n_3}, \quad n_+ n_- = n_1^2 + n_2^2 = 1 - n_3^2
\]
i.e.

\[
\begin{bmatrix}
1 \\
z
\end{bmatrix} = \begin{bmatrix}
\sqrt{\frac{1 + n_3}{2}} \\
\sqrt{\frac{1 + n_3}{2} \left(1 - \frac{n_3}{n_-}\right)}
\end{bmatrix}
\]

The expression on the right above is the normalized eigenvector of the operator

\[
\begin{bmatrix}
n_3 & n_- \\
n_+ & -n_3
\end{bmatrix}
\]

To see this,

\[
\left(\sqrt{\frac{1 + n_3}{2}}\right)^2 + \left(\sqrt{\frac{1 + n_3}{2} \left(1 - \frac{n_3}{n_-}\right)}\right)^2 = \frac{1 + n_3}{2} + \frac{1 + n_3}{2} \left(\frac{1 - n_3}{n_-}\right)^2
\]

\[
= \frac{1 + n_3}{2} + \frac{1 + n_3}{2} \frac{(1 - n_3)^2}{n_-} = \frac{1 + n_3}{2} + \frac{1 + n_3}{2} \frac{(1 - n_3)^2}{1 - n_3^2}
\]

\[
= \frac{(1 + n_3)(1 - n_3)}{2(1 - n_3^2)} + \frac{1 + n_3}{2} \left(\frac{(1 - n_3)^2}{1 - n_3^2}\right)
\]

\[
= \frac{(1 + n_3)(1 + n_3)(1 - n_3)}{2(1 + n_3)(1 - n_3)} + \frac{1 + n_3}{2} \left(\frac{(1 - n_3)(1 - n_3)}{(1 + n_3)(1 - n_3)}\right)
\]

\[
= \frac{1 + n_3}{2} + \frac{(1 - n_3)}{2} = 1
\]

So, if the spin of the system is prepared spin-up in the \(n\)-direction, the probability it will be found spin-up in the \(m\)-direction is

\[
|\langle \sigma_i m_i = 1 | \sigma_i n_i = 1 \rangle|^2 = \left| \begin{bmatrix}
\sqrt{\frac{1 + m_3}{2}} \\
\sqrt{\frac{1 + m_3}{2} \left(1 - \frac{m_3}{m_+}\right)}
\end{bmatrix} \right|^2
\]

\[
= \frac{1 + n_1 m_1 + n_2 m_2 + n_3 m_3}{2} = \frac{1 + \langle n | m \rangle}{2} = \frac{1 + \cos \theta}{2},
\]

where \(\theta\) is the angle between the \(m\)-direction and the \(n\)-direction.

Recall that a measurement interferes with the evolution of a system. When a measurement is taken, the system is left in a new state, the measurement being the eigenvalue of that new state. In other words, the system is left in the state corresponding to the measured eigenvalue and loses any memory of being in the
prepared state. For instance, if the spin along an axis is measured spin-up, then the system is left in a state 'spin-up' along that axis.

11.18 Quantum Entanglement

For any electron whose spin state is given by \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \), there is always a direction ‘\( n \)’ for which the spin is definitely ‘spin-up’ i.e. \( \hat{\sigma} \cdot n \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \). Now suppose there are two electrons, then the operators act on a spin-up or spin-down electron in the following manner:

\[
\begin{align*}
\hat{\sigma}_1 |u\rangle &= |d\rangle, & \hat{\sigma}_1 |d\rangle &= |u\rangle, & \hat{\sigma}_2 |u\rangle &= i|d\rangle, \\
\hat{\sigma}_2 |d\rangle &= -i|u\rangle, & \hat{\sigma}_3 |u\rangle &= |u\rangle, & \hat{\sigma}_3 |d\rangle &= -|d\rangle,
\end{align*}
\]

where \( |u\rangle \) signifies ‘spin-up’ and \( |d\rangle \) signifies ‘spin-down’.

For example,

\[
\begin{align*}
\hat{\sigma}_1 |u\rangle &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & \hat{\sigma}_1 |d\rangle &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.
\end{align*}
\]

Consider a state with two electrons. There are four possibilities: \( |u, u\rangle, |d, u\rangle, |u, d\rangle, |d, d\rangle \), where, for example, ‘\( |u, u\rangle \)’ signifies that the first electron is spin-up and the second electron is spin-up and ‘\( |d, u\rangle \)’ signifies that the first electron is spin-down and the second is spin-up. Note that the four states are orthonormal.

How do the sigma operators operate on states with two electrons? First, define the ‘tau’ operators ‘\( \hat{\tau}_i, \ i = 1,2,3 \)’ to be exactly like the \( \hat{\sigma}_i \)'s, except that the \( \hat{\tau}_i \)'s act only on the second electron and the \( \hat{\sigma}_i \)'s act only on the first electron. For example, \( \hat{\sigma}_1 |u, u\rangle = |d, u\rangle \) and \( \hat{\tau}_1 |u, u\rangle = |u, d\rangle \).

In quantum mechanics, a system can be in a superposition of states i.e.

\[
\alpha |u, u\rangle + \beta |d, u\rangle + \gamma |u, d\rangle + \delta |d, d\rangle, \quad \alpha, \beta, \gamma, \delta \in \mathbb{C}
\]

Suppose the first electron is prepared in the state ‘\( \alpha_1 |u\rangle + \beta_1 |d\rangle \)’ and the second electron prepared in the state ‘\( \alpha_2 |u\rangle + \beta_2 |d\rangle \)’, then

\[
(\alpha_1 |u\rangle + \beta_1 |d\rangle)(\alpha_2 |u\rangle + \beta_2 |d\rangle) = \alpha_1 \alpha_2 |u, u\rangle + \alpha_1 \beta_2 |u, d\rangle + \beta_1 \alpha_2 |d, u\rangle + \beta_1 \beta_2 |d, d\rangle
\]

The equation above is called a ‘product state’. In this case, the spin states of the two electrons are independent. In other words, there will always be a direction ‘\( n \)’ in which the first electron is definitely spin-up and a direction ‘\( m \)’ in which the second electron is definitely spin-up. Since the \( \hat{\tau}_i \)'s only act on the second and the \( \hat{\sigma}_i \)'s only on the first electron, the spin states are independent.

Consider the two electron state ‘\( 1/\sqrt{2} (|u, d\rangle + |d, u\rangle) \)’. The expected value ‘\( \langle \hat{\sigma}_1 \rangle \)’ given that the state is \( 1/\sqrt{2} (|u, d\rangle + |d, u\rangle) \) is
\[ \langle \hat{\sigma}_1 \rangle = \frac{1}{2} \langle (u, d) + (d, u) | \hat{\sigma}_1 | u, d \rangle = \langle (u, d) + (d, u) | (d, d) + |u, u\rangle = 0 \]

Similar calculations will find \( \langle \hat{\sigma}_2 \rangle, \langle \hat{\sigma}_3 \rangle = 0 \). If the spin states of two electrons are independent, then it is always possible to find a direction ‘n’ in which \( \langle \sigma \cdot n \rangle = 1 \). But, if the spin states of the two electrons are given by \( 1/\sqrt{2} (|u, d\rangle + |d, u\rangle) \), then there is no direction ‘n’ where \( \langle \sigma \cdot n \rangle = 1 \). In this case, the spin states of the two electrons are dependent. As such, the electrons are said to be ‘entangled’.

If the spin states of two electrons are independent, knowing the spin state of one of the electrons reveals no information about the spin state of the second electron. The only way to know the spin state of the second electron is to measure it. However, if the spin states are entangled, if the spin state of one electron is known, the other spin state can be known without measuring it.

Consider the states ‘\( 1/\sqrt{2} (|u, d\rangle \pm |d, u\rangle) \)’. The sums ‘\( \hat{\sigma}_i + \hat{\tau}_i, \ i = 1, 2, 3 \)’ acting on \( 1/\sqrt{2} (|u, d\rangle \pm |d, u\rangle) \) produce

\[ \frac{1}{\sqrt{2}} (\hat{\sigma}_3 + \hat{\tau}_3) \{ |u, d\rangle \pm |d, u\rangle \} = |u, d\rangle - |u, d\rangle \pm |d, u\rangle \pm |d, u\rangle = |0\rangle \]

\[ \frac{1}{\sqrt{2}} (\hat{\sigma}_1 + \hat{\tau}_1) \{ |u, d\rangle \pm |d, u\rangle \} = |d, d\rangle + |u, u\rangle \pm |u, u\rangle \pm |d, d\rangle = 2(|d, d\rangle + |u, u\rangle) \neq |0\rangle \]

In the second equation above, if the sign is ‘+’, then

\[ (\hat{\sigma}_1 + \hat{\tau}_1)(|u, d\rangle + |d, u\rangle) \neq |0\rangle \]

But, if the sign is ‘−’, then \( (\hat{\sigma}_1 + \hat{\tau}_1)(|u, d\rangle - |d, u\rangle) = |0\rangle \). To see this,

\[ (\hat{\sigma}_1 + \hat{\tau}_1)(|u, d\rangle - |d, u\rangle) = |u, u\rangle - |d, d\rangle + |d, d\rangle - |u, u\rangle = |0\rangle \]

By direct calculation \( (\hat{\sigma}_2 + \hat{\tau}_2)(|u, d\rangle - |d, u\rangle) = |0\rangle \) and \( (\hat{\sigma}_3 + \hat{\tau}_3)(|u, d\rangle - |d, u\rangle) = |0\rangle \). Hence, \( (\hat{\sigma}_i + \hat{\tau}_i)(|u, d\rangle - |d, u\rangle) = |0\rangle, \ i = x, y, z \). This implies that \( \langle (\hat{\sigma}_i + \hat{\tau}_i) \cdot n \rangle (|u, d\rangle - |d, u\rangle) = |0\rangle \), where \( n \) is any arbitrary direction. If a component ‘\( \hat{\sigma}_i \)’ is measured, then the component ‘\( \hat{\tau}_i \)’ can be known without measuring it.

11.19 Bell’s Theorem

Book I: Sec. 4.4.3 discussed Bell’s inequality. To review, suppose a set of objects possess three distinctly different characteristics, labeled ‘A, B’ and ‘C’ respectfully. Symbolically, Bell’s inequality can be written

\[ N(A, \sim B) + N(B, \sim C) \geq N(A, \sim C) \]

which states that the number of objects that possess characteristic ‘A’, but not characteristic ‘B’ (the “~” symbol indicates “not”) plus the number of objects that
possess characteristic ‘\(B\)’, but not characteristic ‘\(C\)’ is greater than or equal to the number of objects that possess characteristic ‘\(A\)’, but not characteristic ‘\(C\)’.

Bell’s inequality is simply an exercise in propositional logic. Examine the Venn diagram in fig. 11.19-1. In classical logic, an argument is valid if its conclusion entails its premises. In Bell’s inequality \(N(\neg A, \neg B), N(B, \neg C)\) are the premises and ‘\(N(A, \neg C)\)’ is the conclusion. In the Venn diagram below, there are three circles labeled ‘\(A, B, C\)’ divided into distinct areas marked ‘\(a_1, a_2, a_3 \ldots\)’. The individual circles represent each characteristic (\(A, B\) or \(C\)); the label ‘\(a_1\)’ represents those objects that possess characteristic ‘\(A\)’, but not characteristics ‘\(B\)’ or ‘\(C\)’. Area ‘\(a_2\)’ signifies those objects that possess characteristics ‘\(A\)’ and ‘\(B\)’, but not characteristic ‘\(C\)’. Similarly, area ‘\(a_4\)’ signifies objects that possess all three characteristics.

In terms of Bell’s inequality, as it relates to the Venn diagram,

\[
N(A, \neg B) \equiv \text{"the number of objects that fall within the areas } a_1 \text{ and } a_5\text{"}
\]

\[
N(B, \neg C) \equiv \text{"the number of objects that fall within the areas } a_2 \text{ and } a_3\text{"}
\]

\[
N(A, \neg C) \equiv \text{"the number of objects that fall within the areas } a_1 \text{ and } a_3\text{"}
\]

Remembering that Bell’s inequality is written

\[N(A, \neg B) + N(B, \neg C) \geq N(A, \neg C),\]

by substituting the Venn diagram values for their equivalent symbols in Bell’s inequality gives

\[N(a_1 + a_5) + N(a_2 + a_3) \geq N(a_1 + a_2)\]

The number of objects counted in areas ‘\(a_1 + a_5\)’ plus the number of objects counted in areas ‘\(a_2 + a_3\)’ is greater than or equal to the number of objects counted in areas ‘\(a_1 + a_2\)’. The inequality is logically true, since the right-hand side of the inequality (the conclusion) is entailed in the left-hand side (the premises).
11.19.1 Bell’s Inequality Applied to Electron Spin

Bell’s inequality is applied to the electron spin by counting the number of electron spin detections in various detector configurations and then examining the results. The characteristics, in this case, are the detection angles:

A. 0°  
B. 45°  
C. 90°

In spin correlation experiments, Bell’s inequality can be written

\[ N(0°, \sim 45°) + N(45°, \sim 90°) \geq N(0°, 90°) \]

The degree numbers in parentheses represent detector configurations. For example, \( N(0°, \sim 45°) \) indicates that the first detector is set to 0° and the second detector set to 45° and \( N(45°, \sim 90°) \) indicates that the first detector is set to 45° and the second detector is set to 90°. The values inserted into Bell’s inequality are the number of detections registered in the 0° detector, but not in the 45° detector when the detectors are in the 0° − 45° configuration. Added to that amount is the number of detections registered in the 45° detector, but not in the 90° detector when the detectors are configured 45° − 90°. According to Bell’s inequality, that sum should be greater than or equal to the number of detections registered in the 0° detector, but not in the 90° detector when the detectors are 0° − 90° configured.

11.19.2 Conducting Electron Spin Experiments

The experiment is actually performed by putting electron pairs into the singlet state \( |u, d⟩ \sim |d, u⟩ \). Each electron pair will be entangled. Characteristic ‘A’ means that the first electron is measured spin-up along the \( z \)-direction. Characteristic ‘B’ means that the first electron is measured spin-up along the 45°-direction and characteristic ‘C’ says that the first electron is measured spin-up along the \( x \)-direction. Therefore, ‘\( \sim B \)’ means the second electron is spin-up along the 45°-direction. And ‘\( \sim C \)’ means the second electron is spin-up along the \( x \)-direction. Hence, Bell’s inequality can be written

\[ N(0°, 45°) + N(45°, 90°) \geq N(0°, 90°), \]

where \( N(0°, 45°) \) indicates that the first electron is measured spin-up along the \( z \)-direction and the second electron is measured spin-up along the 45°-direction. Moreover, \( N(0°, 45°) \) and \( N(45°, 90°) \) will have identical values, since \( N(45°, 90°) \) is simply a rotation of 45° from the first pair configuration. Bell’s inequality can be simplified to read

\[ 2N(0°, 45°) \geq N(0°, 90°) \]
Counting the number of electron pairs in a certain configuration is the same as determining the probability that the electron pairs are in a given configuration. How are these probabilities calculated? Recall that

\[ |\Psi\rangle = \sum_n a_n |n\rangle, \]

where the \( |n\rangle \)'s are basis vectors. Hence, \( |\Psi\rangle \) is simply a linear combination of the basis vectors. Note that

\[ \langle m | \Psi \rangle = \sum_n a_n \langle m | n \rangle = a_n \to |\Psi\rangle = \sum_n |n\rangle \langle n | \Psi \rangle = \sum_n |n\rangle \langle n | = I \]

In general, the factor \( |a\rangle \langle b| \) is a linear operator. Hence, \( |a\rangle \langle b| \Psi \rangle = \lambda |a\rangle \). In other words, \( |a\rangle \langle b| = \lambda |a\rangle \) is an eigenvalue equation. Suppose

\[ \hat{K} |A\rangle = \lambda |A\rangle, \quad \hat{K} |B\rangle = \lambda |B\rangle \]

If

\[ |\Psi\rangle = \alpha |A\rangle + \beta |B\rangle \to \hat{K} |\Psi\rangle = \alpha \hat{K} |A\rangle + \beta \hat{K} |B\rangle = \alpha \lambda |A\rangle + \beta \lambda |B\rangle = \lambda (\alpha |A\rangle + \beta |B\rangle) = \lambda |\Psi\rangle \]

In other words, if \( \lambda \) is an eigenvalue of two different eigenvectors, it is also an eigenvalue of any linear combination of those eigenvectors.

On the other hand, suppose

\[ |\Phi\rangle = \alpha |A\rangle + \beta |B\rangle + \gamma |C\rangle, \quad \hat{K} |C\rangle = 0 |C\rangle = |0\rangle \]

Hence,

\[ \hat{K} |\Phi\rangle = \alpha \hat{K} |A\rangle + \beta \hat{K} |B\rangle + \gamma \hat{K} |C\rangle = \lambda (\alpha |A\rangle + \beta |B\rangle) = \lambda |\Psi\rangle \]

Note that \( |\Psi\rangle \) is in the subspace of \( |\Phi\rangle \). Now, \( \langle A | \Psi \rangle = \alpha \) and \( \langle B | \Psi \rangle = \beta \), hence, define

\[ \sum_n |n\rangle \langle n | \Psi \rangle = \hat{P}_{\lambda} = \lambda, \]

where \( \hat{P} \) is called the ‘projection operator’. Note that \( |n\rangle \langle n | = I \) if \( \hat{K} |\Phi\rangle \) has eigenvalue \('\lambda'\). Otherwise, \( \hat{P} \) eliminates all vectors that do not have eigenvalue \('\lambda'\). The projection operator \( \hat{P} |u\rangle = |u_p\rangle \)' takes any vector \( |u\rangle \) and projects that vector onto another vector, in this case, \( |u_p\rangle \). In other words, \( \hat{P} \) takes \( |u\rangle \) and eliminates all the components of \( |u\rangle \) that are not in the \( u_p \)-direction. For instance, \( \hat{P}_{\sigma_3 = 1} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \). Hence, \( \hat{P}_{\sigma_3 = 1} \) removes all the components of a vector which are not spin-up along the z-direction. Similarly, \( \hat{P}_{\sigma_3 = -1} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ \beta \end{bmatrix} \). So,

\[ \hat{P}_{\sigma_3 = 1} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \]
By direct calculation $\hat{P}_{\sigma_3=1} = (\sigma_3 + I)/2$, since

$$(\sigma_3 + I)/2 = \frac{1}{2} \left( \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Similarly, $\hat{P}_{\sigma_1=1} = (\sigma_1 + I)/2$. In general, the probability that a certain state vector $|\Psi\rangle$ has a certain property is designated $\langle \Psi | \hat{P} | \Psi \rangle = \langle \hat{P} \rangle$, the average value of $\hat{P}$. For example, let $|\Psi\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, then

$$\langle \psi | \hat{P}_{\sigma_3=1} | \psi \rangle = [\alpha^\dagger \quad \beta^\dagger] \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} [\alpha \quad \beta] = \alpha^\dagger \alpha,$$

which is the probability that $|\Psi\rangle$ is spin-up along the $z$-direction. This gives enough information to calculate the probabilities associated with Bell’s inequality.

11.19.2.1 Results of Electron Spin Experiments

Recall Bell’s inequality for electron spin:

$$2N(0^\circ, 45^\circ) \geq N(0^\circ, 90^\circ)$$

The characteristic ‘$A = 0^\circ$’ means spin-up along the $z$-direction. The projection operator associated with spin-up along the $z$-direction is $\hat{P}_{\sigma_3=1} = (\sigma_3 + 1)/2$ and the projection operator associated with spin-up along the $45^\circ$-direction is

$$\hat{P}_{45^\circ} = \left( \frac{\tau_3 + \tau_1}{2\sqrt{2}} + \frac{1}{2} \right)$$

Hence,

$$\langle (u, d) | -(d, u) | \sigma_3 + 1 \left( \frac{\tau_3 + \tau_1}{2\sqrt{2}} + \frac{1}{2} \right) | u, d \rangle - | d, u \rangle \rangle$$

Note that when $\sigma_3$ acts on $|d, u\rangle$, $|d, u\rangle$ simply vanishes, since $\sigma_3$ measures spin-up on the first electron and since the state of the first electron in $|d, u\rangle$ is spin-down, $\sigma_3$ throws it away. This leaves

$$\langle (u, d) | -(d, u) | \left( \frac{\tau_3 + \tau_1}{2\sqrt{2}} + \frac{1}{2} \right) | u, d \rangle \rangle$$

When $\tau_1$ acts on $|u, d\rangle$, it changes to $|u, u\rangle$. But in $\langle u, d | -(d, u)$ there is no $|u, u\rangle$. Hence, the inner product simply vanishes. Hence, $\tau_1$ contributes nothing to the inner product. This leaves

$$\langle (u, d) | -(d, u) | \left( \frac{\tau_3 + 1}{2\sqrt{2}} \right) | u, d \rangle \rangle$$
When $\tau_3$ acts on $|u, d\rangle$, it gives $-1$. So,

$$\langle (u, d) - (d, u) \mid \left(\frac{-1}{2\sqrt{2}} + \frac{1}{2}\right) \mid u, d\rangle = \frac{1}{2} \left(\frac{-1}{2\sqrt{2}} + \frac{1}{2}\right) = 0.075$$

The above calculation represents the probability that the first electron is spin-up along the $z$-direction and that the second electron is spin-up along the 45°-direction. The left-side of Bell’s inequality becomes $2P(0°, 45°) = 2(0.075) = 0.15$.

The right-side of Bell’s inequality becomes

$$\langle (u, d) - (d, u) \mid \left(\sigma_3 + 1\right)/2 \left(\frac{1}{2} + \frac{1}{2}\right) \mid u, d\rangle = \langle (u, d) - (d, u) \mid \left(\frac{1}{2} + \frac{1}{2}\right) \mid u, d\rangle$$

$$= \left(\frac{1}{2} \mid u, d\rangle = 0.25$$

Therefore, Bell’s inequality becomes

$$2P(0°, 45°) \geq P(0°, 90°) \rightarrow 0.15 \geq 0.25$$

The inequality above is false. Bell’s inequality is violated. But Bell’s theorem is simply an exercise in propositional logic – the logic that governs classical physics, showing beyond doubt that quantum entanglement cannot be explained in terms of classical logic. This is an impediment when trying to unite general relativity with quantum mechanics, since the theory of general relativity complies with traditional logic, but quantum mechanics does not.

11.20 Concluding Remarks

Quantum mechanics, including quantum field theory, describes the nature of atoms and subatomic particles i.e. physics at the smallest scales [264]. Classical physics assumes all physical transitions are smooth and can be measured to any desired degree of accuracy, but quantum mechanics maintains that physical transitions are discrete (quantized). Energy, momentum and other measurable quantities can only assume certain values.

Quantum mechanics arose because certain characteristics of atoms could not be explained by classical physics. In 1803, Thomas Young performed a double-slit experiment that promoted the general acceptance of the wave theory of light. When Maxwell finished his monumental in which he, in contrast to Newton’s corpuscle theory, argued that light consisted of the oscillations of electromagnetic waves, the question was all but decided.

But in 1900, Max Planck hypothesized that energy was radiated and absorbed in discrete "quanta" (energy packets). His results precisely matched the observed patterns of black-body radiation [265]. Planck insisted that his discovery was simply an aspect of the processes of absorption and emission of radiation and had nothing to do with the physical reality of the radiation itself [266]. In fact, he considered his quantum hypothesis a mathematical invention to get the right answer rather than a significant
discovery [267]. However, in 1905, Albert Einstein employed Planck’s quantum hypothesis to explain the photoelectric effect, in which the shining of light on certain materials ejected electrons from the material. Einstein found it necessary to describe light as streams of massless particles, later called ‘photons’. In the 1920’s, Louis de Broglie and Erwin Schrödinger found that an electron, a particle, under some circumstances, showed wave properties. In order to account for the energies of the stable states of atoms, the notion that electrons are little planet-like objects in orbit around the atomic nucleus was rebuked in favor of describing electrons as waves, fitting around the nucleus of an atom like sound waves fitting into an organ pipe.

The world’s categories had become muddled [268]. At about the same time, Ernest Rutherford hypothesized the nuclear model of the atom, for which Niels Bohr developed his theory of the atomic structure, later confirmed by the experiments of Henry Moseley. Eventually, the conclusion was that subatomic particles and electromagnetic waves are neither simply particle nor wave, but have certain properties of each. This originated the concept of ‘wave/particle duality’.

The new discoveries of the mid-1920’s confirmed that, if accepted, wave/particle duality could not be reconciled with classical physics and that an innovative, standard formulation for atomic physics was required. In the summer of 1925, Bohr and Heisenberg published results that better described the particle-like behavior in certain processes and measurements. In 1926, Erwin Schrödinger discovered a partial differential equation for the wave functions of particles like electrons. And when effectively restricted to a finite region, the equation allowed only certain modes, corresponding to discrete quantum states [269]. From Einstein’s simple postulation, the entire field of quantum physics emerged, leading to its wider acceptance at the Fifth Solvay Conference in 1927.

But the new quantum theory had troubling philosophical consequences. The electron waves were not waves of electronic matter, in the way that ocean waves are waves of water, but waves of probability. When a free electron collides with an atom, there is no way of predicting in what direction it will bounce off. The electron wave, after encountering the atom, spreads out in all directions, like an ocean wave after striking a reef, which does not mean that the electron itself spreads out, but instead, goes in some, but not in a precisely predictable direction [268].

Although probability was not an unfamiliar notion, it generally represented an imperfect knowledge of whatever was under study, not an indeterminism in the underlying physical laws. Newton’s theories of motion and gravitation had set the standard of deterministic laws. When a reasonably precise knowledge of the location and velocity of each body in the solar system is known at a given moment, Newton’s laws predict where those bodies will be for a long time into the future. Probability enters Newtonian physics only when knowledge of a system is imperfect, as, for example, when a pair of dice is thrown. But the new quantum mechanics contends that the moment-to-moment determinism of the laws of physics is lost [268].

Quantum mechanics is essential for understanding the physics of systems at atomic length scales and smaller. If the physical nature of an atom were solely described by
classical mechanics, electrons would not orbit the nucleus, since orbiting electrons emit radiation (due to circular motion), but would eventually collide with the nucleus as energy was lost. This framework was unable to explain the stability of atoms. Instead, electrons remain in an uncertain, indeterministic, smeared, probabilistic wave/particle orbit about the nucleus, defying the traditional assumptions of classical mechanics and electromagnetism [270].

In the formalism of quantum mechanics, the state of a system at a given time is described by a complex ‘wave function’, a ‘state vector’ in a complex vector space [271]. Such an abstract mathematical object allows calculating the probabilities of outcomes of concrete experiments. For example, it allows for computing the probability of finding an electron in a particular region around the nucleus at a particular time. Contrary to classical mechanics, simultaneous predictions of conjugate variables, such as position and momentum, to arbitrary precision, is not possible. For instance, electrons may be considered (to a certain probability) located somewhere within a given region of space, but with their exact positions unknown. Heisenberg's uncertainty principle quantifies the inability to precisely locate the particle given its momentum.

The probabilistic nature of quantum mechanics stems from the act of measurement. This is one of the most difficult aspects of quantum systems to understand. It was the central topic in the famous Bohr–Einstein debates, in which the two scientists attempted to clarify these fundamental principles by conducting thought experiments. In the decades after the formulation of quantum mechanics, the question of what constitutes a "measurement" has been extensively studied. Newer interpretations of quantum mechanics do away with the concept of "wave function collapse". The basic idea is that when a quantum system interacts with a measuring apparatus, their respective wave functions become entangled, so that the original quantum system ceases to exist as an independent entity. Although extensively studied, the 'measurement problem' in quantum mechanics has yet to be adequately resolved.

But the most troubling aspect of quantum mechanics is not that it is probabilistic, but that quantum states are described by a complex wave function, which, at any moment in time, gives two values (real and imaginary). This suggests that a quantum system can be in two states at once i.e. in a ‘superposition’ of states. Incredibly, systems are never observed in superposition. If an observation is made, the system will be found in only one of the possible states. If electron spin is measured, the electron will be ‘spin-up’ or ‘spin-down’ in any given direction, never both. The way wave functions change with time is described by Schrödinger’s equation, which does not involve probabilities. It is just as deterministic as Newton’s equations of motion and gravitation. Given the wave function at any moment, Schrödinger’s equation predicts precisely what the wave function will be at any future time. There is not even the possibility of chaos, the extreme sensitivity to initial conditions that is possible in Newtonian mechanics. So if the whole process of measurement is governed by the equations of quantum mechanics, and these equations are perfectly deterministic, how do probabilities enter into quantum mechanics [268]? The whole situation has led to heated debates about the nature of quantum mechanics. The question “what is quantum mechanics” has yet to find a satisfactory answer.
But one thing is certain. The logic that underpins classical physics is not the same as the logic that underpins quantum physics. Quantum Logic was developed in an attempt to construct a propositional structure that would allow describing events in quantum mechanics. Quantum Logic replaced the Boolean structure, which, although suitable for the discourse of classical physics, was inadequate for representing the atomic realm. The mathematical structure of classical systems is a ‘power set’, partially ordered by set inclusion, with a pair of operations that represent conjunction and disjunction. This algebra is consistent with both classical and relativistic physics, but inconsistent in a theory that prohibits, for example, giving simultaneous truth values to the following propositions: “The system possesses this velocity” and “The system is in this place.” The founders of Quantum Logic tried to replace the Boolean structure of classical logic by a weaker structure which relaxed the distributive properties of conjunction and disjunction. But, if there is any hope of uniting quantum mechanics with relativity, a logical system must be found that can accommodate both theories.
"The mathematical framework of quantum theory has passed countless successful tests and is now universally accepted as a consistent and accurate description of all atomic phenomena. The verbal interpretation, on the other hand – i.e., the metaphysics of quantum theory – is on far less solid ground. In fact, in more than forty years physicists have not been able to provide a clear metaphysical model.”

— Fritjof Capra, 
The Tao of Physics

12.0 Introduction

Not much angst surfaced when Maxwell’s equations did not apply to gravitation, since, at that time, no a link between electricity and gravity was expected. But eventually, it was discovered that light was at once a wave and a particle. How could its proper size and shape be imagined? To produce interference it must spread out, but to bounce off electrons, it must be minutely localized. A stalemate in the wave-photon battle troubled the soul of every physicist. How could light be two such contradictory things? Such an unresolved dualism gnawed at the ideals and traditions of science, yet the evidence could not be denied and much water flowed beneath the bridges before a way out of the quandary was found [237].

According to ‘wave-particle duality’, the properties of electrons and photons are fundamentally similar. Despite obvious differences in their mass and charge, under the right circumstances, both show wave-like diffraction and spread as well as particle-like localization. Yet, in classical physics, these objects are very different. Electrons and other particles are conceived as elementary constituents of Nature. By contrast, light arises as a ripple of the electromagnetic field. Is the particle fundamental, with the electromagnetic field arising from the jiggling of a collection of quantum photons? Or is the field fundamental, with the photon appearing only when there is a local disturbance in the field? If the latter view is correct, then there should be an “electron field”, whose ripples give rise to particles with mass and charge. But, if this is the case, why didn’t Faraday, Maxwell and other classical physicists introduce matter fields, analogous to the electromagnetic field [236]?

In quantum field theory, the second viewpoint is the most useful. The field is primary. Particles appear only after ‘quantization’. Photons arise from the quantization of the electromagnetic field. Massive, charged particles, such as electrons, arise from the quantization of matter fields. To describe the fundamental laws of Nature, it is necessary to introduce electron fields, quark fields, neutrino fields, gluon fields, $W$- and $Z$-boson fields and the Higgs field. In fact, there will be a field for each type of particle [236].
Introducing the ‘field’ in classical physics allows creating local laws of Nature. The old laws of Coulomb and Newton involved “action at a distance”, where the force felt by an electron (or planet) changes immediately if a distant proton (or star) moves. Such a description is not only philosophically untenable, but does not align with experiments. The field theories of Maxwell and Einstein require a time lag between a disturbance occurring in the field and when it is felt some distance away, giving a local interpretation to the spread of forces and other physical phenomena. The requirement for locality remains for quantum fields as well [236].

The unification of quantum mechanics with special relativity showed that the number of particles in any interaction is not necessarily conserved. Particles are not indestructible objects, but can be created and destroyed. The discovery of the anti-particle, first predicted by Dirac [236], doomed any attempt at writing down a relativistic version of the one-particle Schrödinger equation. Any attempt at a relativistic version of the one-particle Schrödinger equation meets with serious problems (negative probabilities, infinite towers of negative energy states or a breakdown in causality). When entering the relativistic regime, a new formalism which can handle an unspecified number of particles is needed. This formalism is ‘quantum field theory’ (QFT) [236].

12.1 Why Quantum Field Theory (QFT)?

Interestingly, two protons are exactly the same, whether one came from a cosmic ray associated with a supernova lying eight billion light-years away and the other came from a particle accelerator here on Earth. Why are there no errors in proton production? How can two objects, manufactured so far apart in space and time, be identical? One explanation is that there is a sea of “proton stuff” filling the Universe. The “stuff” is the proton field or, if looked at closely enough, the quark field. Identical quantum particles are truly indistinguishable.

Of the known particles in Nature, there are only two types: the ‘boson’ and the ‘fermion’. In QFT, these particles are represented mathematically by a state vector \(|p_1, p_2\rangle\), where \(p_1\) represents the first particle and \(p_2\) the second. Swapping two particles around leaves the state of a system completely unchanged — apart from a possible minus sign, which determines the kind of particle. If \(|p_1, p_2\rangle = |p_2, p_1\rangle\), the particles are ‘bosons’. If \(|p_1, p_2\rangle = -|p_2, p_1\rangle\), the particles are ‘fermions’. In ordinary quantum mechanics, so as to agree with experiments, this information is put in by hand, Bose statistics (no minus sign) for integer spin particles (bosons) and Fermi statistics (minus sign) for half-integer spin particles (fermions). In quantum field theory, the relationship between spin and particle statistics is not contrived, but a consequence of the framework of QFT [236].

12.1.1 Quantizing the Field

A field is essentially a wave, conceived as infinitely long, spreading out over an entire space. But an infinitely long wave is difficult to describe mathematically, so waves are often modeled as string vibrations between two fixed points, like a guitar or violin string. This kind of wave is called a ‘standing wave’. A standing wave travels in a given direction until it encounters one of the fixed ends, then it reverses direction. The direction reversal creates a problem, since, in this case, the momentum of the wave is
not conserved. Because most fields encountered in QFT require momentum conservation, the wave is imagined as travelling in a closed loop. Once the wave completes the loop, it simply continues in the same direction, which preserves momentum. The price paid for conserving momentum is quantization, where the wave length \( \lambda \) must be an integer multiple of the distance \( S \) around the closed loop i.e. \( S = m\lambda, \ m \in Z \). If this were not the case, the wave would lose its composition as it moved around the loop becoming difficult or impossible to describe. If wave length \( \lambda \) is quantized, momentum is quantized, since momentum \( p \) is defined in terms of \( \lambda \) i.e. \( p = \hbar/\lambda \rightarrow p = m\hbar/S \). Quantization, therefore, is an essential feature of quantum mechanical systems.

If the wave moves in a circle of radius \( r \), then

\[
p = m\hbar/S = \frac{m\hbar}{2\pi r}
\]

The angular momentum \( L \) is given by

\[
L = mvr = pr \rightarrow L = \frac{m\hbar}{2} = mh, \quad m \in Z, \quad \hbar = \frac{\hbar}{2\pi}
\]

In all quantum mechanical systems, angular momentum is quantized, regardless of the path of the wave.

Intuitively, QFT requires that the Universe consists of a variety of fields. Values for the field exist at every point of space. Particles arise from local vibrations in the fields. Some fields can excite vibrations in other fields. The formal mathematics of QFT describes the kinds of excitations and vibrations that can take place.

### 12.2 The Development of a Relativistic Field Equation

Mathematically, a field in QFT is a collection harmonic oscillators. The total energy \( E \) in a harmonic oscillator is

\[
E = K.E. + P.E. \rightarrow \hat{H} = \frac{\hat{p}^2}{2m_{k.E.}} + \frac{1}{2}k\hat{x}^2 = \frac{\hat{p}^2}{2m_{k.E.}} + \frac{m\omega^2}{2P.E.}\hat{x}^2, \quad \omega = \sqrt{\frac{k}{m}} \rightarrow \omega^2 = \frac{k}{m}, \quad E \rightarrow \hat{H},
\]

where \( E \) has been replaced by the operator \( \hat{H} \) and \( k \) is the spring constant. To associate the equation above with the energy in a wave, write

\[
\hat{H} = \frac{\hat{p}^2}{2m_{k.E.}} + \frac{1}{2}k\hat{x}^2 \rightarrow \hat{H} = \frac{\hat{p}^2}{2m_{k.E.}} + \frac{1}{2}\omega^2\hat{x}^2 = \frac{1}{2}(\hat{p} + i\omega\hat{x})(\hat{p} - i\omega\hat{x}), \quad m = 1, \quad \omega^2 = k,
\]

Earlier attempts to incorporate relativity into quantum mechanics began with Schrödinger’s equation:
\[
\frac{i\hbar}{\partial t} \Psi = \hat{H} \Psi, \quad \hat{H} = \frac{\hat{p}^2}{2m} + V = -\frac{\hbar^2}{2m} \frac{\partial}{\partial x_j} \partial_j + V, \quad j = x, y, z, \quad \hat{p} = -i\hbar \frac{\partial}{\partial x_j}
\]

The 4-momentum in special relativity is length invariant. Hence,

\[
p'^\mu p_\mu = mc^2 = g_{\mu\nu} p'^\mu p_\nu = [E/c \ p^x \ p^y \ p^z] \left[ \begin{array}{c} E/c \\ -p_x \\ -p_y \\ -p_z \end{array} \right] \rightarrow \frac{E}{c^2} = p^j p_j + mc^2 \rightarrow E^2
\]

Changing the variables to operators, \( E \rightarrow \hat{H}, \ p \rightarrow \hat{p} = -i\hbar \partial/\partial x_j \), then

\[
\hat{H} = \sqrt{(\hat{p}c)^2 + mc^4} = \sqrt{-\hbar^2 c^2 \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_j} + mc^4}
\]

Dirac attempted this approach, but could make no sense of \( \hat{H} \) as a square root. He ultimately abandoned the approach.

### 12.2.1 The Klein-Gordon Equation

However, a more fruitful tactic was found by using \( \hat{H}^2 \) rather than \( \hat{H} \):

\[
\hat{H}^2 = -\hbar^2 c^2 \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_j} + mc^4 \rightarrow \left( i\hbar \frac{\partial}{\partial t} \right) \left( i\hbar \frac{\partial}{\partial t} \right) |\Psi\rangle = -\hbar^2 \frac{\partial^2}{\partial t^2} |\Psi\rangle
\]

\[
= \left( -\hbar^2 c^2 \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_j} + mc^4 \right) |\Psi\rangle \rightarrow -\frac{1}{c^2} \frac{\partial^2}{\partial t^2} |\Psi\rangle = \left( -\frac{\partial}{\partial x_j} \frac{\partial}{\partial x_j} + \frac{mc^2}{\hbar^2} \right) |\Psi\rangle
\]

\[
\rightarrow \frac{\partial}{\partial x^0} \frac{\partial}{\partial x^0} |\Psi\rangle - \left( \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_j} - \frac{m}{\hbar^2} \right) |\Psi\rangle = |0\rangle \rightarrow \left( \frac{\partial}{\partial x^0} \frac{\partial}{\partial x^0} + \mu^2 \right) |\Psi\rangle = |0\rangle,
\]

\[c = 1, \quad \mu = \frac{m}{\hbar^2}, \quad \nu = 0, ..., 3\]

The equation on the right above is the ‘Klein-Gordon equation’. The solution is

\[
|\Psi(x, t)\rangle = \sum_{n=0}^{\infty} \frac{1}{\sqrt{2V E_n / \hbar}} \left( A_n e^{i(E_n t - \hat{p}_n |x\rangle)} + B_n^+ e^{iE_n t - (\hat{p}_n |x\rangle)} \right)
\]

To make \( |\Psi\rangle \) time independent, let \( E \rightarrow \omega, \ t = 0, \ n \rightarrow k \), then the solution becomes

\[
|\Psi(x)\rangle = \sum_{k=0}^{\infty} \frac{1}{\sqrt{2V \omega_k}} \left( A_k e^{ikx} + B_k^+ e^{-ikx} \right), \quad \hbar = 1
\]

since \( k_n = 2\pi/\lambda_n \) and each discrete wavelength fits an integer number of times inside the volume ‘V’. In other words, if time is ignored, \( |\Psi(x)\rangle \) becomes a relativistic time-independent solution. Let
\[ |\Psi_{k,A} \rangle = \frac{e^{ikx}}{\sqrt{V}} \]

then

\[ \langle \Psi_{k,A} | \Psi_{k,A} \rangle = \int \Psi_{k,A}^\dagger \Psi_{k,A} \, d^3 x = \int \frac{e^{-ikx} e^{ikx}}{\sqrt{V} \sqrt{V}} \, d^3 x = 1 \]

More generally, the states \(|\Psi_{k,A}\rangle\) are orthonormal, therefore

\[ \int \Psi_{k,A}^\dagger \Psi_{k',A} \, d^3 x = \int \frac{e^{-ikx} e^{ik'x}}{\sqrt{V} \sqrt{V}} \, d^3 x = \delta_{kk'} \]

12.2.2 The Continuity Equations and Locality

There is a relationship between Schrödinger’s equation and the continuity equation of classical physics. Continuity equations are a stronger, local form of conservation laws. For example, a weak version of the law of conservation of energy asserts that energy can neither be created nor destroyed, but does not rule out the possibility that a quantity of energy could disappear from one point while simultaneously appearing at another point. If energy is locally conserved, not only can energy not be created nor destroyed, it cannot “teleport” from one place to another. It can only move by a continuous flow. A continuity equation is the mathematical way of expressing local conservation laws.

Recall the continuity equation of classical physics:

\[ \frac{\partial \rho}{\partial t} + \langle \nabla | j \rangle = 0, \quad \nabla = \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z} \]

where \(\rho\) is the ‘charge/mass density’, \(j\) is the ‘current or flow density’ and \(\rho = 0\) outside a given volume ‘\(V\)’, but is nonzero and can change inside \(V\). Recall Schrödinger’s equation

\[ \frac{\partial \Psi}{\partial t} = \frac{1}{i\hbar} \left( -\frac{\hbar^2}{2m} \nabla^2 + V \right) \Psi \rightarrow \left\{ \Psi^\dagger \frac{\partial \Psi}{\partial t} = \frac{\Psi^\dagger}{i\hbar} \left( -\frac{\hbar^2}{2m} \nabla^2 + V \right) \Psi \right\} \]

The conjugate Schrödinger’s equation is

\[ \frac{\partial \Psi^\dagger}{\partial t} = \frac{-1}{i\hbar} \left( -\frac{\hbar^2}{2m} \nabla^2 + V^\dagger \right) \Psi^\dagger \rightarrow \left\{ \Psi \frac{\partial \Psi^\dagger}{\partial t} = \frac{-\Psi}{i\hbar} \left( -\frac{\hbar^2}{2m} \nabla^2 + V^\dagger \right) \Psi^\dagger \right\} \]
Adding the two equations on the right above gives

\[ \begin{align*}
\{ \psi^\dagger \frac{\partial \psi}{\partial t} &= \frac{\psi^\dagger}{i\hbar} \left( -\frac{\hbar^2}{2m}\nabla^2 + V \right) \psi \} + \{ \psi \frac{\partial \psi^\dagger}{\partial t} &= -\frac{\psi}{i\hbar} \left( -\frac{\hbar^2}{2m}\nabla^2 + V^\dagger \right) \psi^\dagger \} \\
&\rightarrow \psi^\dagger \frac{\partial \psi}{\partial t} + \psi \frac{\partial \psi^\dagger}{\partial t} \\
&= \frac{\psi^\dagger}{i\hbar} \left( -\frac{\hbar^2}{2m}\nabla^2 + V \right) + \left( -\frac{\psi}{i\hbar} \left( -\frac{\hbar^2}{2m}\nabla^2 + V + V^\dagger \right) \right) \\
&= -\frac{\hbar}{2im} (\psi^\dagger \nabla^2 \psi - \psi \nabla^2 \psi^\dagger) + \frac{\nabla^\dagger \psi^\dagger \psi - \psi^\dagger \nabla \psi}{i\hbar}
\end{align*} \]

The last term on the right above vanishes, since \( V = V^\dagger \) and \( V \) commutes with \( \psi \).

Hence,

\[ \frac{\partial (\psi^\dagger \psi)}{\partial t} = -\frac{\hbar}{2im} (\psi^\dagger \nabla^2 \psi - \psi \nabla^2 \psi^\dagger) = -\frac{\hbar}{2im} (\nabla (\nabla^\dagger \psi^\dagger) \psi - (\nabla \psi^\dagger) \psi) \]

Letting

\[ \rho = \psi^\dagger \psi \ , \quad J = \frac{\hbar}{2im} \left( (\nabla \psi^\dagger) \psi - (\nabla^\dagger \psi) \psi \right), \]

then

\[ \frac{\partial \rho}{\partial t} + \langle \nabla J \rangle = 0, \]

which is the continuity equation.

In quantum mechanics, \( \rho \) is called the ‘probability density’ and \( J \) is called the ‘probability current’ [272]. Note that \( \int \rho \ dV = 1 \). However, the quantum mechanical continuity equation differs from the classical, because, in the classical case, \( \rho \) and \( J \) represent physically measurable quantities. Whereas the quantum mechanically continuity equation is associated with a probability.

12.2.2.1 Relativistic Continuity Equations

A similar derivation can be accomplished in relativistic quantum mechanics, except that Schrödinger's equation is replaced by the Klein-Gordon equation:

\[ \frac{\partial^2 \psi}{\partial t^2} = (\nabla^2 - \mu^2) \psi, \quad \mu = \frac{mc^2}{\hbar^2} \rightarrow \left\{ \frac{\partial^2 \psi}{\partial t^2} = (\nabla^2 - \mu^2) \psi \right\} \psi^\dagger, \quad \left\{ \frac{\partial^2 \psi^\dagger}{\partial t^2} = (\nabla^2 - \mu^2) \psi^\dagger \right\} \psi. \]
This time subtracting the second equation from the first
\[
\left\{ \frac{\partial^2 \psi}{\partial t^2} = (\nabla^2 - \mu^2) \psi \right\} \psi^+ - \left\{ \frac{\partial^2 \psi}{\partial t^2} = (\nabla^2 - \mu^2) \psi^+ \right\} \psi = \frac{\partial^2 \psi}{\partial t^2} \psi^+ - \frac{\partial^2 \psi}{\partial t^2} \psi
\]
\[
= \{ (\nabla^2 - \mu^2) \psi \} \psi^+ - \{ (\nabla^2 - \mu^2) \psi \} \psi = \frac{\partial}{\partial t} \left( \frac{\partial \psi}{\partial t} \psi^+ - \frac{\partial \psi^+}{\partial t} \psi \right) =
\]
\[
= (\nabla) (\nabla \psi) \psi^+ - (\nabla \psi^+) \psi
\]

Letting
\[
\rho = \frac{\partial \psi}{\partial t} \psi^+ - \frac{\partial \psi^+}{\partial t} \psi, \quad J = \left( (\nabla \psi) \psi^+ - (\nabla \psi^+) \psi \right) \to \frac{\partial \rho}{\partial t} + (\nabla J) = 0
\]
The last equation on the right above is the continuity equation [272]. Recall the solution to the Klein-Gordon equation:
\[
\psi(x, t) = \sum_{k=0}^{\infty} \frac{1}{\sqrt{2V \omega_k}} (A_k e^{i(\omega_k t - kx)} + B_k^+ e^{-i(\omega_k t - kx)}), \quad \hbar = 1
\]
To simplify matters, let all the $B_k^+$'s equal zero, then
\[
\psi(x, t) = \sum_{k=0}^{\infty} \frac{A_k e^{i(\omega_k t - kx)}}{\sqrt{2V \omega_k}} \to \rho = \frac{\partial \psi}{\partial t} \psi^+ - \frac{\partial \psi^+}{\partial t} \psi
\]
\[
= i \left( \sum_{k=0}^{\infty} \frac{\omega_k A_k}{\sqrt{2V \omega_k}} e^{i(\omega_k t - kx)} \right) \left( \sum_{k'=0}^{\infty} \frac{A_{k'}^*}{\sqrt{2V \omega_{k'}}} e^{-i(\omega_{k'} t - k'x)} \right) + i \left( \sum_{k'=0}^{\infty} \frac{\omega_{k'} A_{k'}^*}{\sqrt{2V \omega_{k'}}} e^{-i(\omega_{k'} t - k'x)} \right) \left( \sum_{k=0}^{\infty} \frac{A_k}{\sqrt{2V \omega_k}} e^{i(\omega_k t - kx)} \right) \to \rho(x, 0)
\]
\[
= i \left( \sum_{k=0}^{\infty} \frac{\omega_k A_k}{\sqrt{2V \omega_k}} e^{-ikx} \right) \left( \sum_{k'=0}^{\infty} \frac{A_{k'}^*}{\sqrt{2V \omega_{k'}}} e^{ik'x} \right) + i \left( \sum_{k'=0}^{\infty} \frac{\omega_{k'} A_{k'}^*}{\sqrt{2V \omega_{k'}}} e^{-ik'x} \right) \left( \sum_{k=0}^{\infty} \frac{A_k}{\sqrt{2V \omega_k}} e^{ikx} \right),
\]
Since the states are orthogonal, all the terms where $k \neq k'$ vanish. Hence,
\[
\rho(x, 0) = 2i \int_{k=0}^{\infty} \frac{\omega_k A_k A_k^+}{2V \omega_k} = i \int_{k=0}^{\infty} \frac{A_k A_k^+}{V} \to \int \rho \, d^3x = i \int d^3x \left( \sum_{k=0}^{\infty} \frac{A_k A_k^+}{V} \right) = \sum_{k=0}^{\infty} A_k A_k^+
\]
Note that $\int \rho \, d^3x = 1$. Hence, $\sum_k A_k A_k^+ = 1$ and $A_k A_k^+$ becomes a probability.
However, the Klein-Gordon approach uses the square of the energy i.e.
\[
E^2 = (\hbar \omega)^2 \to E = \pm \hbar \omega,
\]
producing negative energies. Suppose that $\omega_k$ is replaced with $-\omega_k$, negative energies states. In this case, $\rho \to -\rho$ and

335
\[-\int \rho \, d^3x = -i \sum_k \sum_k A_k A_k^\dagger, = -i,\]

which implies not only negative energies, but negative probabilities. Negative probabilities are meaningless. This feature of the Klein-Gordon approach doomed it as a viable option for a theory, but the exercise eventually led to quantum field theory.

12.3 The Quantum Field

One of the advantages to the Klein-Gordon approach was that $A_k, A_k^\dagger$ and $B_k, B_k^\dagger$, the amplitudes in the solution, were eventually identified as operators, specifically, raising and lowering operators (see [272] for details).

The wave function ‘$\psi(x)$’ in quantum mechanics takes the form ‘$\psi(x) \propto e^{ikx}$’, where $k$ can be positive or negative. If the wave travels in a closed loop, where the distance around the loop is signified by ‘$S$’, then $e^{iks} = 1$, since

$$e^{2\pi i} = \cos 2\pi + i \sin 2\pi = 1,$$

the angle around any closed loop being $2\pi$. Therefore,

$$kS = 2\pi m, \quad m \in N \rightarrow k = \frac{2\pi m}{S},$$

where $S/m = \lambda$, the wave length. Hence, $k\lambda = 2\pi \rightarrow k = 2\pi/\lambda$. The momentum ‘$p$’ is

$$\frac{h}{\lambda} \rightarrow p = \frac{hk}{2\pi} = \frac{2\pi m}{S},$$

$$h = \frac{\hbar}{2\pi}$$

12.3.1 Periodic Functions

If a function ‘$f(x)$’ is, 1) single valued and continuous in an interval ‘$[a, b]$’, 2) $f(a) = f(b)$, then $f(x)$ is ‘periodic’ with period ‘$p$’ if the cycle repeats itself (see Figure 12.3-1). What Fourier found was that any wave ‘$f(x)$’, regardless of the complexity of its shape or configuration, could be represented by sine or cosine functions, so long as the two requirements mentioned above are satisfied.

$$f(x) = f(x + p)$$

Figure 12.3-1
A field \( \Psi(x) \) in classical physics is defined

\[
\Psi(x) = \sum_k \alpha_k e^{ikx}, \quad \Psi^\dagger(x) = \sum_k \alpha_k^* e^{-ikx} \rightarrow \Psi(x) = \sum_k \alpha_k (\cos kx + i \sin kx),
\]

\[
\Psi^\dagger(x) = \sum_k \alpha_k^* (\cos kx - i \sin kx), \quad \alpha_k = \frac{1}{\pi} \int_{-\infty}^{\infty} \Psi(x) \cos kx \, dx,
\]

\[
\alpha_k^* = \frac{1}{\pi} \int_{-\infty}^{\infty} \Psi^\dagger(x) \sin kx \, dx
\]

where \( \tilde{\Psi}(x) \) and \( \tilde{\Psi}^\dagger(x) \) are Fourier transforms of each other (see Book II: Sec. 16.8). The \( \alpha_k \)'s are the various amplitudes of the individual waves that make up the composite wave. The function \( \Psi(x) \) is said to be a 'wave' with amplitude \( \alpha_k \) and wave number \( k \), where \( k = 2\pi/\lambda \).

### 12.3.2 The Quantum Field as an Operator

Recall that in quantum mechanics \( p = \hbar k \rightarrow p = k, \ \hbar = 1 \). To create a quantum field \( \tilde{\Psi}(x) \), the \( \alpha_k \)'s are replaced by the creation and annihilation operators \( \tilde{\alpha}_k^+ \) and \( \tilde{\alpha}_k^- \) respectfully i.e.

\[
\tilde{\Psi}(x) = \sum_k \tilde{\alpha}_k^- e^{ikx}, \quad \tilde{\Psi}^\dagger(x) = \sum_k \tilde{\alpha}_k^+ e^{-ikx}
\]

For instance,

\[
\tilde{\Psi}^\dagger(x) |g\rangle = \sum_k e^{-ikx} \tilde{\alpha}_k^+ |g\rangle = e^{-ikx} |k\rangle
\]

In other words, \( \tilde{\Psi}^\dagger(x) \) acting on \( |g\rangle \), the ground state, creates a particle at position \( x \) with momentum \( k \). On the other hand,

\[
\tilde{\Psi}(x) |k\rangle = \sum_k e^{ikx} \tilde{\alpha}_k^- |k\rangle = e^{ikx} |g\rangle
\]

The operator \( \tilde{\Psi}(x) \) acting on \( |k\rangle \), annihilates a particle with momentum \( k \) and returns the vacuum state \( |g\rangle \).

In QFT, quantum states are particles and operators are fields that create and annihilate particles. In ordinary quantum mechanics, a state is represented by a state vector \( |\psi(x)\rangle \), which is a field, and, if operated on by \( \hat{X} \), represents a particle at position \( x \). If \( |\phi(x)\rangle \) is operated on by \( \hat{R} \), the momentum operator, it represents a particle with momentum \( k \) at position \( x \). In ordinary quantum mechanics states are fields and operators operate on fields to create particles, which is the opposite of what takes place in QFT.
12.3.2.1 The State Vector in QFT

In ordinary quantum mechanics, the number of particles is fixed in any system. In QFT, the \( \hat{\Psi} \)'s represent operators, which signify observables. The function \( \hat{\Psi} \) is a function of a single position and describes a system of any number of particles. The state, in QFT, is represented by a state vector, for example, \( | n_{k_1}, n_{k_2}, n_{k_3}, \ldots \rangle \), where \( n_{k_1} \) represents the number of particles with momentum \( 'k_1' \) and \( n_{k_2} \), the number of particles with momentum \( 'k_2' \) etc.

To understand this in more detail, recall that \( \psi(x) \propto e^{ikx} \), where the \( k \)'s represent the allowed values of momentum i.e. \( k = 2\pi m/S, \ m \in \mathbb{Z} \) and \( S \) is the distance around a closed loop. Note that

\[
\int_{-\frac{S}{2}}^{\frac{S}{2}} e^{ikx} dx = \begin{cases} \frac{S}{2} & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \implies \int_{-\frac{S}{2}}^{\frac{S}{2}} e^{ikx} dx = 2\pi \delta(k)
\]

To see this, recall that the Dirac delta function with a width of \( '2\pi/S' \) has height \( 'S' \). Hence, \( 2\pi S/S = 2\pi \). If \( k \neq 0 \), then

\[
\int_{-\frac{S}{2}}^{\frac{S}{2}} e^{ikx} dx = \frac{1}{ik} e^{ikx} \left|_{-\frac{S}{2}}^{\frac{S}{2}} \right. = \frac{1}{ik} \left[ \cos(kx) + i \sin(kx) \right] \left|_{-\frac{S}{2}}^{\frac{S}{2}} \right.
\]

\[
= \frac{1}{ik} \left[ \cos \left( \frac{kS}{2} \right) + i \sin \left( \frac{kS}{2} \right) - \cos \left( -\frac{kS}{2} \right) - i \sin \left( -\frac{kS}{2} \right) \right]
\]

\[
= \frac{1}{ik} \left[ \cos \left( \frac{kS}{2} \right) + i \sin \left( \frac{kS}{2} \right) - \cos \left( \frac{kS}{2} \right) - i \sin \left( \frac{kS}{2} \right) \right] = \frac{2}{k} \sin \left( \frac{kS}{2} \right) = \frac{2}{k} \sin(\pi m)
\]

Therefore,

\[
\lim_{S \to \infty} \int_{-\frac{S}{2}}^{\frac{S}{2}} e^{ikx} dx = \int_{-\infty}^{\infty} e^{ikx} dx = 2\pi \delta(k) \to \int_{-\infty}^{\infty} e^{ikx} dk = 2\pi \delta(x)
\]

The equation above implies that if the exact momentum of a particle is known, then its location is spread over the entire real line. Conversely, if the exact position of a particle is known, then its momentum is spread over the entire real line.

12.3.2.2 Creating Field Operators

The equation

\[
\hat{\Psi}(x) = \sum_k \hat{a}_k e^{ikx} \to \hat{\Phi}(x, t) = \sum_k \hat{a}_k e^{i(kx - \omega t)}
\]

where \( \hat{\Phi}(x, t) \) is now a traveling wave rather than a vibration. Recalling that

338
\[ E = h \omega \rightarrow E = \omega \rightarrow E = p^2/2m \rightarrow E = k^2/2m, \quad \hbar = 1, \quad p = k, \]

then
\[ \frac{d \hat{\Psi}}{dt} = \sum_k \hat{a}_k^-( -i \omega) e^{i(kx - \omega t)}, \]
\[ \frac{d^2 \hat{\Psi}}{dx^2} = \sum_k \hat{a}_k^-( i k)^2 e^{i(kx - \omega t)} \rightarrow -\frac{1}{i \omega} \frac{d \hat{\Psi}}{dt} = - \frac{1}{k^2} \frac{d^2 \hat{\Psi}}{dx^2} \rightarrow i \frac{d \hat{\Psi}}{\omega dt} = - \frac{1}{k^2} \frac{d^2 \hat{\Psi}}{ dx^2} \]

However, \( E = \omega \rightarrow k^2 = 2m \omega \). Hence,
\[ i \frac{d \hat{\Psi}}{\omega dt} = - \frac{1}{2m \omega} \frac{d^2 \hat{\Psi}}{dx^2} \rightarrow i \frac{d \hat{\Psi}}{dt} = - \frac{1}{2m} \frac{d^2 \hat{\Psi}}{dx^2} \]

Reinserting the \( \hbar \)'s and inserting a potential '\( V \)' gives
\[ i \hbar \frac{d \hat{\Psi}}{dt} = - \frac{\hbar^2}{2m} \frac{d^2 \hat{\Psi}}{dx^2} + V \]

The equation above is the time dependent Schrödinger's equation, except \( \hat{\Psi} \) is now an operator rather than a state vector.

12.3.2.3 Rising and Lowering Energy Operators

Recall the creation '\( \hat{a}^+ \)' and annihilation '\( \hat{a}^- \)' operators that create and destroy harmonic oscillators:
\[ \hat{a}^+ = \hat{p} - i \omega \hat{x} \sqrt{2 \omega}, \quad \hat{a}^- = \hat{p} + i \omega \hat{x} \sqrt{2 \omega}, \quad \hat{a}^+ \hat{a}^- = \hat{N}, \]

where \( \hat{N} \) is called the 'number operator' or an 'occupation number'.

The energy '\( \hat{H} \)' of a harmonic oscillator is given by \( \hat{H} = \hbar \omega (\hat{N} + 1/2) \), where the '1/2' is generally omitted, since it does not add appreciable content. The quantity '\( \hat{a}^+ \)', called the 'raising operator', acting on a state '\( |n\rangle \)', raises the energy of that state by one unit i.e.
\[ \hat{a}^+ |n\rangle = \sqrt{n + 1} |n + 1\rangle \]

The lowering operator '\( \hat{a}^- |n\rangle = \sqrt{n} |n - 1\rangle \)' lowers the energy by one unit. Of note,
\[ \hat{a}^+ \hat{a}^- |n\rangle = \sqrt{n} \hat{a}^+ |n - 1\rangle = \sqrt{n} \sqrt{n} |n\rangle = n |n\rangle \rightarrow \hat{a}^+ \hat{a}^- = n \rightarrow \hbar \omega \hat{a}^+ \hat{a}^- = n \hbar \omega = E \]
Hence, energy in all quantum mechanical systems is quantized.

Since

\[ \hat{a}^\dagger \hat{a}^\dagger |n\rangle = \sqrt{n + 1} \hat{a}^\dagger |n + 1\rangle = \sqrt{n + 1} \sqrt{n + 1} |n\rangle = (n + 1) |n\rangle \rightarrow \hat{a}^\dagger \hat{a}^\dagger = n + 1 \]
\[ \rightarrow \hat{a}^\dagger \hat{a}^\dagger - \hat{a}^\dagger \hat{a}^\dagger = [\hat{a}^\dagger, \hat{a}^\dagger] = n - (n + 1) = -1 \rightarrow [\hat{a}^\dagger, \hat{a}^\dagger] = 1, \]

the commutation relations for \( a^+, a^- \) are

\[ [\hat{a}^+, \hat{a}^+] = 0, \quad [\hat{a}^-, \hat{a}^-] = 0, \quad [\hat{a}^-, \hat{a}^+] = 1 \]

This is all the information required to describe quantum harmonic oscillators.

Suppose there are a number of harmonic oscillators, then

\[ \hat{a}^+_j = \frac{\hat{p}_j - i \omega_j \hat{x}_j}{\sqrt{2 \omega_j}}, \quad \hat{a}^-_j = \frac{\hat{p}_j + i \omega_j \hat{x}_j}{\sqrt{2 \omega_j}}, \]

where \( \hat{a}^+_j, \hat{a}^-_j \) are creation and annihilation operators for the \( j^{th} \) harmonic oscillator. Hence,

\[ \hat{H} = \hbar \sum_j \omega_j \hat{N}_j, \quad \hat{a}^+_j \hat{a}^-_j = \hat{N}_j \]

The commutation relations for \( \hat{a}^+_j, \hat{a}^-_j \) are

\[ [\hat{a}^+_j, \hat{a}^-_j] = 0, \quad [\hat{a}^-_j, \hat{a}^-_j] = 0, \quad [\hat{a}^-_j, \hat{a}^+_k] = \delta_{jk} \]

States for a number of harmonic oscillators are given by \( |n_1, n_2, n_3, \ldots \rangle \), where \( n_1 \) is the number of harmonic oscillators in state ‘1’ and \( n_2 \) is the number of harmonic oscillators in state ‘2’ etc. The number of harmonic oscillators can be infinite or finite. Here, it is assumed that the harmonic oscillators do not interact.

**12.4 QFT Overview**

Suppose the initial state of a system is specified ‘\( |\text{int}\rangle \)’ and the final state by ‘\( \langle \text{final}| \)’. Recall that ‘\( \langle \text{final}| \text{int}\rangle \)’ is associated with the probability that, if the system is prepared in the “initial state”, it will end up in the “final state”. If \( \hat{a}^+ |n\rangle = \sqrt{n + 1} |n + 1\rangle \) and \( \hat{a}^- |n\rangle = \sqrt{n} |n - 1\rangle \), what do \( \langle n|\hat{a}^+ \rangle \) and \( \langle n|\hat{a}^- \rangle \) represent?

Note that \( \langle n|\hat{a}^+ |m\rangle = \sqrt{m + 1} \langle n|m + 1\rangle \) and

\[ \langle n|m + 1\rangle = \begin{cases} 1, & \text{if } n = m + 1 \\ 0, & \text{otherwise} \end{cases} \]

if the states are normalized.
So,

\[ \langle n|\hat{a}^+ = \langle n - 1|\sqrt{n}, \quad \langle n|\hat{a}^- = \langle n + 1|\sqrt{n+1} \]

Example:

\[ \langle n|\hat{a}^+\hat{a}^-|n \rangle = \sqrt{n}\langle n - 1|\hat{a}^-|n \rangle = \sqrt{n}\sqrt{n}|n|n \rangle = n, \quad \langle n|n \rangle = 1 \]

### 12.4.1 Creation and Annihilation Operators for Bosons

The \( \Psi_i \)'s can be thought of as creation and annihilation operators, synonymous with creation and annihilation operators for harmonic oscillators. For instance, suppose a particle 'a' with momentum 'k' decays into two particles 'b' and 'c' at position 'x' with momentum 'l' and 'm' respectfully. In QFT, the process is represented thusly:

\[
\Psi_m^\dagger(x)\Psi_l^\dagger(x)\Psi_k(x)|0,0,0,...,k,...,0,0,0...\rangle = \left( \sum_m \hat{a}_m^+ e^{-imx} \right) \left( \sum_l \hat{a}_l^+ e^{-ilx} \right) \left( \sum_k \hat{a}_k^- e^{ikx} \right) |0,0,0,...,k,...,0,0,0...\rangle
\]

First, \( \Psi_k(x) \) acts on \(|0,0,0,...,k,...,0,0,0...\rangle\) and will cycle through all the allowable values of \( k \) and does nothing until it finds a particle with momentum 'k'. At that point, it annihilates the particle, leaving the vacuum state \(|g\rangle\) along with a factor \( e^{ikx} \). Next,

\[
\Psi_m^\dagger \Psi_l^\dagger (x)|g\rangle = e^{ikx} \left( \sum_m \hat{a}_m^+ e^{-imx} \right) \left( \sum_l \hat{a}_l^+ e^{-ilx} \right) |g\rangle = e^{ikx} e^{ilx} \sum_m \hat{a}_m^+ e^{-imx} |l\rangle
\]

The creation operator \( \Psi_l^\dagger(x) \) acts on \(|g\rangle\) creating a particle with momentum 'l' at position 'x', leaving the state \(|l\rangle\) along with a factor \( e^{ilx} \). i.e. it creates a particle at position 'x' with momentum 'l'. Finally, \( \Psi_m^\dagger \) acts on \(|l\rangle\) creating a particle with momentum 'm' at position 'x':

\[
\Psi_m^\dagger |l\rangle = e^{ikx} e^{-ilx} \left( \sum_m \hat{a}_m^+ e^{-imx} \right) |l\rangle = e^{ikx} e^{-ilx} e^{-imx} |l,m\rangle
\]

The coefficients of \(|l,m\rangle\) i.e. \( e^{ikx}, e^{-ilx}, e^{-imx} \), are related to the probability that this particular interaction takes place. The creation and annihilation operators just described can create and annihilate bosons.

Example: A particle with momentum \( 'k_1' \) slams into a completely rigid target that runs the entire length of the vertical axis. The particle is scattered by the target and ends up with momentum \( 'k_F' \) (see fig. 12.4.1-1). The initial state of the particle is given by

\[
|0,0,0,...,\frac{1}{k_1},...0,0,0...\rangle = |k_1\rangle,
\]
which represents a single particle with momentum \( 'k_f' \). The problem is formulated as

\[
\langle k_F | g \int \Phi^\dagger(0, t)\Phi(0, t) \, dt | k_i \rangle = \text{a number}
\]

The annihilation of the particle with momentum \( 'k_f' \) and the creation of the particle with momentum \( 'k_f' \) occurs at \( 'x = 0' \). At some arbitrary time \( 't' \), a particle with momentum \( 'k' \) is annihilated by \( \Phi \) and, at the same time, a particle with momentum \( 'k_f' \) is created by \( \Phi^\dagger \). The factor \( 'g' \) is called the 'coupling constant' and is related to the probability that this particular interaction occurs. Recall that the probability amplitude \( \langle y|x \rangle \), in ordinary quantum mechanics, is related to the probability that, if a system is prepared in state \( '|y\rangle' \), then undergoes a transformation, it will end up in state \( '|x\rangle' \). Hence,

\[
\langle k_F | g \int \Phi^\dagger(0, t)\Phi(0, t) \, dt | k_i \rangle = \langle k_F | g \int (\sum_k \hat{a}_k^+ e^{-i(kx-\omega_F t)})(\sum_k \hat{a}_k e^{i(kx-\omega e x t)}) \, dt | k_i \rangle \\
= \langle k_F | g \int (\sum_k \hat{a}_k^+ e^{i\omega F t})(\sum_k \hat{a}_k^+ e^{-i\omega t}) \, dt | k_i \rangle \\
= \langle k_F | g \int (\sum_k \hat{a}_k^+ e^{i\omega_F t} e^{-i\omega t}) \, dt | g \rangle = \langle g | g \sum_i e^{i\omega_F t} e^{-i\omega t} \, dt | g \rangle \\
= g \int e^{i\omega_F t} \, dt \langle g | g \rangle = 2\pi g \delta(\omega_F - \omega_i) = 2\pi g,
\]

since \( \delta(\omega_i - \omega_F) = 0 \) unless \( \omega_i = \omega_F \), in which case, \( \delta(\omega_i - \omega_F) = 1 \). This is an example of time translation invariance, which is synonymous with the conservation of energy i.e. \( \hbar \omega_i = \hbar \omega_F \). Moreover, the probability that this interaction takes place is proportional to \( (2\pi g)^2 \).

Neither \( \Phi(x), \Phi^\dagger(x) \) is Hermitian, since \( \hat{a}_i^-, \hat{a}_i^+ \) are not Hermitian. But it is a simple matter to create Hermitian operators from \( \hat{a}_i^-, \hat{a}_i^+ \) by adding or subtracting them.

### 12.4.2 The Fock Space

Both \( \Phi(x), \Phi^\dagger(x) \) act on the state \( '|n_1, n_2, n_3, ... \rangle' \), which is called a 'Fock space'. Recall that \( \sum_i |i\rangle\langle i|x \rangle = |x\rangle \), \( \langle i|x \rangle = \psi_i^\dagger(x) \). Therefore, \( \sum_i |i\rangle\langle i|x \rangle = \sum_i \psi_i^\dagger(x) |i\rangle = |x\rangle \). Now if \( |i\rangle = \hat{a}_i^\dagger |g\rangle \), then

\[
\sum_i \psi_i^\dagger(x) |i\rangle = \sum_i \psi_i^\dagger(x) \hat{a}_i^\dagger |g\rangle = |x\rangle, \quad \sum_i \psi_i^\dagger(x) \hat{a}_i^+ = \Phi^\dagger(x).
\]
where $|g\rangle$ represents the ground state. Hence, $\bar{\Psi}^\dagger(x)$ acting on the ground state creates a particle at position ‘x’. This is written ‘$\bar{\Psi}^\dagger(x)|g\rangle = |x\rangle$’. Likewise, $\bar{\Psi}(x)|g\rangle = |0\rangle$. The operator ‘$\bar{\Psi}(x)$’, acting on the ground state annihilates the state. By the same token, $\bar{\Psi}^\dagger(x)\bar{\Psi}^\dagger(y)|g\rangle = |y, x\rangle$; the operator ‘$\bar{\Psi}^\dagger(y)$’ first creates a particle at ‘y’ and then $\bar{\Psi}^\dagger(x)$ creates a particle at ‘x’. Note that

$$
\bar{\Psi}^\dagger(x)\bar{\Psi}^\dagger(y)|g\rangle = |y, x\rangle = \bar{\Psi}^\dagger(y)\bar{\Psi}^\dagger(x)|g\rangle = |x, y\rangle
$$

since $[\bar{\Psi}^\dagger, \bar{\Psi}^\dagger] = 0$. Interchanging $\bar{\Psi}^\dagger(y)$ with $\bar{\Psi}^\dagger(x)$ does not change the sign of the state vector. The two operators ‘$\bar{\Psi}$, $\bar{\Psi}^\dagger$’ are creation and annihilation operators for bosons, which can be in the same state at the same time. If two bosons are interchanged, the state remains unchanged. There is a distinct quantum field for each boson.

12.4.3 The Conservation of the Inner Product of States over Time

Note that $\bar{\Psi}\bar{\Psi}^\dagger = e^{i\alpha}\bar{\Psi}^\dagger e^{-i\alpha}\bar{\Psi}$, $\alpha \in \mathbb{C}$. If the fields ‘$\bar{\Psi}$, $\bar{\Psi}^\dagger$’ carry electric charge, the number of $\bar{\Psi}^\dagger$s must equal the number of $\bar{\Psi}$s. The inner product of states must be conserved over time. To see this, suppose there is a one particle state described by the following process:

$$
\langle F|\bar{\Psi}^\dagger\bar{\Psi}|I\rangle \to e^{i\alpha x}\langle F|\bar{\Psi}^\dagger\bar{\Psi}|g\rangle = e^{-i\beta x}e^{i\alpha x}\langle g|\bar{\Psi}^\dagger|g\rangle = e^{-i\gamma x}e^{-i\beta x}e^{i\alpha x}\langle g|g\rangle = 0
$$

Such a process cannot take place, since $\langle g|g\rangle = 0$. Charge is conserved in Nature. If there are an unequal number of annihilation and creation operators, the phases do not cancel and the inner products are not conserved over time, a violation of the postulates of quantum mechanics.

12.4.4 The Number Operator in QFT

Recall that if $\{|i\rangle\}$ is a basis of states, then $\sum_i |i\rangle\langle i| = I$. Note that

$$
\langle y|x\rangle = \sum_i \langle y|i\rangle \langle i|x\rangle = \sum_i \psi_i(x)\psi_i^\dagger(x) = \delta(x - y)
$$

This is because $\sum_i \psi_i(x)\psi_i^\dagger(x) = 0$ unless $x = y$, then it equals 1.

In ordinary quantum mechanics,$\int \psi_i^\dagger(x)\psi_i(x) \, dx = 1$. Suppose the $\psi_i$’s are replaced by the $\bar{\Psi}_i$’s, then

$$
\int \psi_i^\dagger(x)\psi_i(x) \, dx = \int \sum_{ij} \bar{\hat{a}}_j^\dagger\bar{\Psi}_j^\dagger(x)\bar{\hat{a}}_i\psi_i(x) \, dx = \sum_{ij} \bar{\hat{a}}_j^\dagger \bar{\hat{a}}_i \delta_{ij} = \sum_i \bar{\hat{a}}_i^\dagger \bar{\hat{a}}_i = \sum_i \bar{N}_i,
$$

since $\int \psi_i(x)\psi_i^\dagger(x) \, dx = \delta_{ij}$. The integral ‘$\int \psi^\dagger(x)\psi(x) \, dx$’ is equal to the number of particles; the expression ‘$\Psi^\dagger(x)\Psi(x)$’ is called the ‘particle density’. Note that $\Psi^\dagger(x)\Psi(x)$ is Hermitian, since multiplying anything by its complex conjugate is Hermitian.
12.4.5 The Energy Operator in QFT

The total energy ‘$E$’ of a number of particles is given by $E = \sum_i n_i \omega_i$, where $n_i$ is the number of particles in state $|i\rangle$ with energy $\omega_i$. If the particles do not interact in any way,

$$E = \sum_i n_i \omega_i \to \sum_i \omega_i \hat{N}_i = \sum_i \omega_i \hat{a}_i^+ \hat{a}_i^-, \tag{12.4.5.1}$$

where $E$ has been turned into an operator. Each state $|i\rangle$ is an energy eigenstate whose wave function is $\psi_i(x)$.

The time independent Schrödinger equation for a one particle system is $\hat{H} |\psi_i\rangle = \omega_i |\psi_i\rangle$, where the $\omega_i$’s are the energy eigenvalues of $\hat{H}$; the operator ‘$\hat{H}$’ represents the Hamiltonian i.e. $\hat{H} = \hat{p}^2 / 2m + V(x)$. Hence,

$$\left[ \frac{\hat{p}^2}{2m} + V(x) \right] |\psi_j\rangle = \left[ -\frac{\nabla^2}{2m} + V(x) \right] |\psi_j\rangle = \omega_j |\psi_j\rangle, \quad \hat{p} = -i \frac{\partial}{\partial x}, \quad \hbar = 1 \tag{12.4.5.2}$$

Note that

$$\langle \psi | \hat{H} | \psi \rangle = \int \psi^+(x) \left[ -\frac{\nabla^2}{2m} + V(x) \right] \psi(x) \, dx, \tag{12.4.5.3}$$

where $\langle \psi | \hat{H} | \psi \rangle$ signifies the ‘average value’ of $\hat{H}$. Upon replacing $\psi$ by $\Phi$ in the equation above,

$$\int \Phi^+(x) \left[ -\frac{\nabla^2}{2m} + V(x) \right] \Phi(x) \, dx = \int \sum_{ij} \hat{a}_i^+ \psi_i^+(x) \left[ -\frac{\nabla^2}{2m} + V(x) \right] \hat{a}_j^+ \psi_j(x) \, dx$$

$$= \int \sum_{ij} \hat{a}_i^+ \hat{a}_j^+ \psi_i^+(x) \psi_j(x) \omega_j \, dx = \sum_{ij} \hat{a}_i^+ \hat{a}_j^+ \delta_{ij} \omega_j = \hat{E}, \tag{12.4.5.4}$$

since $[-\nabla^2 / 2m + V(x)] = \omega_j$ and $\int \psi_i(x) \psi_i^+(x) \, dx = \delta_{ij}$.  

12.4.6 Creation and Annihilation Operators of Position and Momentum

Recall that $\psi(x) \psi^+(x) \propto P(x)$ and $\bar{\psi}(p) \to \bar{\psi}(p) \psi^+(p) \propto P(p)$, where $P(x)$ is the probability that a particle is at position ‘$x$’ and $P(p)$ is the probability that a particle has momentum ‘$p$’. Note that $\psi$ and $\bar{\psi}$ are Fourier conjugates of each other:

$$\bar{\psi}(p) = \frac{1}{\sqrt{2\pi}} \int \psi(x) e^{-ipx} \, dx \to \psi(x) = \frac{1}{\sqrt{2\pi}} \int \bar{\psi}(p) e^{ipx} \, dp$$
And recall that \( \bar{\Psi}(x) = \sum_j \hat{a}_j^\dagger \psi_j(x) \) and \( \bar{\Psi}^\dagger(x) = \sum_j \hat{a}_j \psi_j^*(x) \). This implies that for a free particle with no forces acting on it

\[
\bar{\Psi}^\dagger(x) = \frac{1}{\sqrt{2\pi}} \int \hat{a}_k^\dagger e^{-ikx} \, dk,
\]

where

\[
\hat{a}_k^\dagger = \frac{1}{\sqrt{2\pi}} \int \bar{\Psi}^\dagger(x) e^{ikx} \, dx,
\]

and \( \bar{\Psi}^\dagger(x) \) and \( \hat{a}_k^\dagger \) are Fourier conjugates of each other. Since \( [\hat{a}_i^\dagger, \hat{a}_j^\dagger] = \delta_{ij} \),

\[
[\bar{\Psi}(y), \bar{\Psi}^\dagger(x)] = \delta(x - y),
\]

which confirms that fields at different points in space do not interfere. This maintains ‘locality’.

Consider a situation where a particle with momentum ‘\( k_1 \)’ decays into two particles with momentum ‘\( k_1, k_2 \)’ respectfully. In this case, at some particular time ‘\( t = 0 \)’, a particle with momentum ‘\( k_1 \)’ is annihilated by \( \bar{\Psi} \) and, at the same time, two particles with momentum ‘\( k_1, k_2 \)’ are created by \( \bar{\Psi}^\dagger \). If \( g \) is the coupling constant, then

\[
\langle k_2, k_3 \mid g \int \bar{\Psi}^\dagger(x, 0) \bar{\Psi}^\dagger(x, 0) \bar{\Psi}(x, 0) \, dx \mid k_1 \rangle
= \langle k_2, k_3 \mid g \int (\sum_k a_k^\dagger e^{-ikx}) (\sum_k a_k^\dagger e^{-ikx}) \, dx \mid k_1 \rangle
= g \int e^{i(k_1-k_2-k_3)x} \, dx \langle g \mid g \rangle = g \int e^{i(k_1-k_2-k_3)x} \, dx = 2\pi g \delta(k_1 - k_2 - k_3)
= 2\pi g
\]

Note that \( \delta(k_1 - k_2 - k_3) = 0 \) unless \( k_1 = k_2 + k_3 \), in which case, \( \delta(k_1 - k_2 - k_3) = 1 \). This is an example of space translation invariance, which is synonymous with the conservation of momentum.

12.4.7 Momentum Conservation in QFT

Recall that the average ‘\( \langle \hat{H} \rangle \)’ is given by

\[
\langle \hat{H} \rangle = \int \bar{\Psi}^\dagger(x) \left[ -\frac{\nabla^2}{2m} + V(x) \right] \bar{\Psi}(x) \, dx = \int \left( \bar{\Psi}^\dagger(x) \left[ -\frac{\nabla^2}{2m} \bar{\Psi}(x) \right] + V(x)\bar{\Psi}^\dagger(x)\bar{\Psi}(x) \right) \, dx
\]

The second term in the integral indicates that there are \( \int \bar{\Psi}^\dagger(x)\bar{\Psi}(x) \, dx \) number of particles with potential energy ‘\( V(x) \)’, since \( \bar{\Psi}^\dagger(x)\bar{\Psi}(x) \) is the particle density. For the simplest case, let \( V(x) = mc^2 \), which is just a constant. What does \( \langle \hat{H} \rangle \) do when it acts on a state vector ‘\( \mid \phi \rangle \)’? To see, note that

345
\[ |\phi(t + \varepsilon) = (1 - i\varepsilon\langle \hat{H} \rangle)|\phi(t)\rangle = |\phi(t)\rangle - i\varepsilon\langle \hat{H} \rangle|\phi(t)\rangle \]

The state \(|\phi(t + \varepsilon)\rangle\) is the state of the system as time changes by a small amount \(\varepsilon\). If momentum is conserved in a system, then \(\langle \hat{H} \rangle\) acting on \(|\phi(t)\rangle\) will not change the momentum. If

\[
\Psi(x) = \frac{1}{\sqrt{2\pi}} \int \bar{\Psi}(p)e^{ipx} dp, \quad \Psi^+(x) = \frac{1}{\sqrt{2\pi}} \int \bar{\Psi}^+(q)e^{-iqx} dq
\]

then,

\[
mc^2 \bar{\Psi}^+(x)\Psi(x) dx = mc^2 \frac{1}{2\pi} \int \int \bar{\Psi}^+(q)\bar{\Psi}(p)e^{i(p-q)x} dx dp dq
\]

\[
= mc^2 \int \bar{\Psi}^+(q)\bar{\Psi}(p)\delta(q-p) dp dq = mc^2 \int \bar{\Psi}^+(p)\bar{\Psi}(p) dp,
\]

since \(\delta(q-p) = 0\), unless \(q = p\), then \(\delta(q-p) = 1\). When \(\langle \hat{H} \rangle\) acts on the potential it removes a particle of momentum \(p\) and then puts it back. Momentum is conserved. The same idea can be extended to a many particle system, mathematically written

\[
mce^2 \frac{1}{2\pi} \int \bar{\Psi}^+(q_i)\bar{\Psi}(p_i)e^{i(\Sigma_i p_i-q_i)x} dx dp_i dq_i, \quad i = 1,2,3, ...
\]

Recall that the first term in \(\langle \hat{H} \rangle\) is

\[
\langle \hat{H} \rangle(\text{first term}) = \int \bar{\Psi}^+(x) \left[ -\frac{\nabla^2}{2m} \Psi(x) \right] dx \rightarrow -\frac{\nabla^2}{2m} \Psi(x) = \frac{1}{\sqrt{2\pi}} \int \bar{\Psi}(p) \frac{p^2}{2m} e^{ipx} dp
\]

\[
\rightarrow \frac{1}{2\pi} \int \bar{\Psi}^+(p)e^{-ipx} \frac{\bar{\Psi}(p)}{2m} p^2 e^{ipx} dx dp = \frac{1}{2\pi} \int \bar{\Psi}^+(p)\bar{\Psi}(p) \frac{p^2}{2m} dp
\]

Momentum in the first term of \(\langle \hat{H} \rangle\) is conserved. If \(\langle \hat{H} \rangle\) acts on a state vector, momentum is conserved and, since the operator ‘\(\hat{\Psi}\)’ is a function of the spatial coordinates, momentum is translationally invariant.

Suppose there is a two particle bosonic system with particles labeled ‘\(A\)’ and ‘\(B\)’. A possible \(\langle \hat{H} \rangle\) for the system is

\[
\langle \hat{H} \rangle = \int \bar{\Psi}^+_A(x) \left[ -\frac{\nabla^2}{2m_A} \right] \Psi_A(x) + \int \bar{\Psi}^+_B(x) \left[ -\frac{\nabla^2}{2m_B} \right] \Psi_B(x) + \bar{\Psi}^+_A(x)\bar{\Psi}^+_B(x)\bar{\Psi}(A)\bar{\Psi}(B) \Psi_A(x) \Psi_B(x) dx,
\]

which is somewhat unrealistic, but illustrates how \(QFT\) works in practice.
In figure 12.4.7-1, \( \langle \hat{H} \rangle \) only acts if the two particles ‘\( A \)’ and ‘\( B \)’ are found at the same place ‘\( x \)’. The operator ‘\( \Psi_A^\dagger \)’ annihilates particle ‘\( A \)’ and ‘\( \Psi_B^\dagger \)’ annihilates particle ‘\( B \)’ at position ‘\( x \)’. If \( \langle \hat{H} \rangle \) finds the particles at two different positions, it does nothing. Moreover, \( \Psi_A^\dagger \) creates a particle ‘\( A \)’ at ‘\( x \)’ and \( \Psi_B^\dagger \) creates a particle ‘\( B \)’ at ‘\( x \)’. This is an example of scattering and is typical of the kinds of problems that quantum field theory handles well. While the momentum of each particle may not be conserved, the total momentum of the system is conserved. Moreover, for particles to interact, they must be at the same place, preserving locality.

### 12.4.8 Coupling Constants

There is another value that can go into the Hamiltonian, a constant ‘\( g \)’, called the ‘coupling constant’:

\[
\langle \hat{H} \rangle = \int \Psi_A^\dagger(x) \left(-\frac{\nabla^2}{2m_A}\right) \Psi_A(x) + \int \Psi_B^\dagger(x) \left(-\frac{\nabla^2}{2m_B}\right) \Psi_B(x) + g \Psi_A^\dagger(x) \Psi_B^\dagger(x) \Psi_B(x) \Psi_A(x) \, dx
\]

A large value of \( g \) indicates a high probability, while a small \( g \) is associated with a small probability that the interaction takes place.

### 12.4.9 The Ordering of Operators

Suppose a particle ‘\( a \)’ decays into two particles ‘\( b \)’ and ‘\( c \)’ (see figure 12.4.9-1 a)). The operators used in the Hamiltonian must be Hermitian. The factor ‘\( \Psi_c^\dagger(x) \Psi_b^\dagger(x) \Psi_a(x) \)’ is not Hermitian, since the \( \Psi \)'s are not Hermitian. In order to make the Hamiltonian Hermitian, the complex conjugate of each operator must be included, for example,

\[
g \int \Psi_c^\dagger(x) \Psi_b^\dagger(x) \Psi_a(x) + \Psi_a^\dagger(x) \Psi_b(x) \Psi_c(x) \, dx
\]
In the second term every creation operator has been replaced by an annihilation operator and vice versa. This implies that if one particle can become two particles, then two can become one (see figure 12.4.9-1 b)). Moreover, the fact that the operators must be Hermitian places a restriction on the type of particle interactions and decays that can occur in Nature.

Consider the expansion of $|\phi(t + \epsilon)\rangle$ to second order i.e.

$$|\phi(t + \epsilon)\rangle = |\phi(t)\rangle - i\epsilon\langle\hat{H}|\phi(t)\rangle - \frac{\epsilon^2}{2!}\langle\hat{H}\rangle^2|\phi(t)\rangle$$

The effect of $\langle\hat{H}\rangle$ is to add a term to the Hamiltonian in the form

$$g^2 \int \bar{\Psi}_c^\dagger(x)\bar{\Psi}_b^\dagger(x)\bar{\Psi}_a(x)\bar{\Psi}_c(x)\bar{\Psi}_b(x) dx$$

Two particles 'b' and 'c' are annihilated at 'x' and a particle 'a' is created, then annihilated and two particles 'b' and 'c' are created. This process represents the scattering of two particles. The particle 'a' simply plays an intermediary role. The scattered particles conserve momentum, not necessarily individually, but total system momentum.

Consider

$$\langle b, c | g \int \bar{\Psi}_c^\dagger(x)\bar{\Psi}_b^\dagger(x)\bar{\Psi}_a(x) dx | a \rangle = \langle b, c | g \int a_a^+ a_b^+ a_a^+ e^{iax} e^{-i(b+c)x} dx | a \rangle$$

$$= \langle g | g \int e^{iax} e^{-i(b+c)x} dx | g \rangle = g \int e^{iax} e^{-i(b+c)x} dx | g | g \rangle$$

$$= g \int e^{iax} e^{-i(b+c)x} dx = 2\pi g \delta[a - (b + c)]$$

Since the particles 'a, b, c' are in momentum eigenstates, the only time the equations above give a non-zero result is if $a - (b + c) = 0$. Thus, the equations above show that
Moreover, \( \langle \hat{H} \rangle \) can have an enormous number of particle scattering and decaying possibilities. The pictorial representations of these processes are called ‘Feynman diagrams’.

12.4.10 Creation and Annihilation Operators for Fermions

It is impossible to distinguish between multi-particle states if their only difference is an exchange of bosons i.e. \( |x, y\rangle = |y, x\rangle \). In other words, if \( \hat{\Psi}^\dagger \) creates a boson, then \( \hat{\Psi}^\dagger (x) \hat{\Psi}^\dagger (y) |g\rangle = \hat{\Psi}^\dagger (y) \hat{\Psi}^\dagger (x) |g\rangle \), where \( |g\rangle \) signifies the ground state. This is true because \( [\hat{\Psi}^\dagger (x), \hat{\Psi}^\dagger (y)] = 0 \), boson field operators commute.

The rules for fermions are different. Fermions cannot occupy the same state at the same time. The mathematical rule for fermions is anti-commutation i.e. \( |x, y\rangle = -|y, x\rangle \):

\[
\hat{\Psi}^\dagger (x) \hat{\Psi}^\dagger (y) |g\rangle = -\hat{\Psi}^\dagger (y) \hat{\Psi}^\dagger (x) |g\rangle = \hat{\Psi}^\dagger (x) \hat{\Psi}^\dagger (y) |g\rangle + \hat{\Psi}^\dagger (y) \hat{\Psi}^\dagger (x) |g\rangle = |0\rangle
\]

The form of the creation and annihilation operators for bosons and fermions are identical, but the creation and annihilation operators for fermions are signified by \( \hat{c}^+, \hat{c}^- \), respectfully:

\[
\hat{\Psi}(x) = \sum_k \hat{c}_k^- \psi_k(x) = \sum_k \hat{c}_k^- e^{ikx}, \quad \hat{\Psi}^\dagger (x) = \sum_k \hat{c}_k^+ \psi_k(x) = \sum_k \hat{c}_k^+ e^{-ikx},
\]

What separates fermions from bosons is the commutation relations for \( \hat{c}^+, \hat{c}^- \):

\[
\hat{c}^+ |g\rangle = |1\rangle, \quad \hat{c}^+ |1\rangle = |0\rangle, \quad \hat{c}^- |g\rangle = |0\rangle, \quad \hat{c}^- |1\rangle = |g\rangle
\]

Fermions cannot occupy the same state at the same time. There are only two fermion states ‘\( |1\rangle \)’ and ‘\( |g\rangle \)’ – a state with one particle or a state with no particles. If \( \hat{c}^+ \) acts on the ground state, it puts a particle there. If it acts on an occupied state, it annihilates the state. Conversely, if \( \hat{c}^- \) acts on the ground state, it annihilates it, since it cannot remove a particle if no particle is there. If it acts on an occupied state, it removes the particle. Of note:

\[
\hat{c}^+ \hat{c}^- |g\rangle = \hat{c}^+ |0\rangle = |0\rangle, \quad \hat{c}^- \hat{c}^+ |g\rangle = \hat{c}^- |1\rangle = |g\rangle
\]

\[
\hat{c}^+ \hat{c}^- |1\rangle = \hat{c}^+ |g\rangle = |1\rangle, \quad \hat{c}^- \hat{c}^+ |1\rangle = \hat{c}^- |0\rangle = |0\rangle
\]

Moreover,

\[
(\hat{c}^+ \hat{c}^- + \hat{c}^- \hat{c}^+) |s\rangle = |s\rangle \rightarrow (\hat{c}^+ \hat{c}^- + \hat{c}^- \hat{c}^+) = I,
\]

space translation symmetry is synonymous with the conservation of momentum, since \( b, c \) must have the same momentum as the initial particle ‘\( a \)’.
where \( |s\rangle \) is some arbitrary state. The quantity \((\hat{c}^+\hat{c}^- + \hat{c}^-\hat{c}^+)\) is called an ‘anti-commutator’, defined \(\{A, B\} = AB + BA\). The fact that fermion operators anti-commute is generally signified by \(\{\hat{\Psi}^\dagger(x), \hat{\Psi}^\dagger(y)\} = 0\), where

\[
\{\hat{\Psi}^\dagger(x), \hat{\Psi}^\dagger(y)\} = \hat{\Psi}^\dagger(x)\hat{\Psi}^\dagger(y) + \hat{\Psi}^\dagger(y)\hat{\Psi}^\dagger(x) = 0
\]

Note that if \(x = y\), then

\[
\hat{\Psi}^\dagger(x)\hat{\Psi}^\dagger(y)|g\rangle + \hat{\Psi}^\dagger(y)\hat{\Psi}^\dagger(x)|g\rangle = 2\hat{\Psi}^\dagger(x)\hat{\Psi}^\dagger(x)|g\rangle = |0\rangle,
\]

which guarantees the impossibility of creating two fermions that occupy the same state. Note that

\[
\{\hat{c}^+, \hat{c}^-\} = \{\hat{c}^-, \hat{c}^+\} = 1, \quad \{\hat{c}^+, \hat{c}^+\} = \{\hat{c}^-, \hat{c}^-\} = 0
\]

Mathematically, \(\hat{c}^+, \hat{c}^-\) are indistinguishable, although obviously representing different functions.

### 12.5 The Dirac Equation

Dirac first successfully applied quantum mechanical methods to relativistic ideas, although his efforts were limited to the special theory of relativity. In classical physics, the kinetic energy of a free particle \(E\) is given by \(E = \frac{p^2}{2m}\). In quantum theory,

\[
E \to \hat{H} = i \frac{\partial}{\partial t}, \quad \hat{p} \to \hat{\rho} = -i \frac{\partial}{\partial x}, \quad \hbar = 1
\]

In relativity theory, the energy \(E\) is given in terms of the momentum and mass of a particle i.e. \(E^2 = p^2 + m^2\), \(c = 1\). Replacing the variables with their associated operators gives

\[
-\frac{\partial^2}{\partial t^2} \phi(x, t) = -\frac{\partial^2}{\partial x^2} \phi(x, t) + m^2 \phi(x, t)
\]

This is the ‘Klein-Gordon’ equation. It complies with the laws of special relativity and is Lorentz invariant.

Dirac was dissatisfied with the ‘Klein-Gordon’ approach because it was second order in time, whereas Schrödinger’s approach was first order in time. As a first attempt, he considered

\[
i \frac{\partial}{\partial t} |\psi\rangle = E |\psi\rangle, \quad E = \sqrt{-\frac{\partial^2}{\partial x^2} + m^2},
\]
but abandoned the approach because no sense could be made of \( \sqrt{-\partial^2/\partial x^2 + m^2} \). He finally settled on the following: suppose the problem is simplified by assuming that the particle has no mass i.e. \( m = 0 \), and that it moves in one dimension along a line. The previous equation becomes

\[
i \frac{\partial}{\partial t} \ket{\psi} = -i \frac{\partial}{\partial x} \ket{\psi} \rightarrow \frac{\partial}{\partial t} \ket{\psi} + \frac{\partial}{\partial x} \ket{\psi} = \ket{0},
\]

\[
E = \sqrt{-\partial^2/\partial x^2} \rightarrow -E^2 = \frac{\partial^2}{\partial x^2} \rightarrow iE = \frac{\partial}{\partial x} \rightarrow E = -i \frac{\partial}{\partial x}
\]

The general solution to the equation above takes the form

\[
\ket{\psi} = \alpha (x - t), \quad \alpha = a \text{ constant}
\]

Three issues arose with this approach: 1) a particle could move to the left or to the right, in which case its momentum (and hence, its energy) would be positive if moving to the right and negative if moving to the left, 2) the equation allowed a particle to move in only one direction, 3) the particle’s mass must be zero.

If \( \ket{\psi} \) is considered a field instead of a state vector, then

\[
\psi_L = e^{i(kx+\omega t)}, \quad \psi_R = e^{i(kx-\omega t)},
\]

where \( \psi_L \) represents a wave moving to the left and \( \psi_R \) represents a wave moving to the right. Hence,

\[
\frac{d\psi_L}{dx} = ik\psi_L, \quad \frac{d\psi_L}{dt} = i\omega \psi_L \rightarrow \frac{d\psi_L}{dkx} = \frac{d\psi_L}{i\omega dt} \rightarrow \frac{d\psi_L}{dt} = \frac{\omega d\psi_L}{k dx} = \nu \frac{d\psi_L}{dx},
\]

since \( \omega/k = \nu \). Similarly,

\[
\frac{d\psi_R}{dx} = ik\psi_R, \quad \frac{d\psi_R}{dt} = -i\omega \psi_R \rightarrow \frac{d\psi_R}{dkx} = \frac{d\psi_R}{-i\omega dt} \rightarrow \frac{d\psi_R}{dt} = -\frac{\omega d\psi_R}{k dx} = -\nu \frac{d\psi_R}{dx}
\]

\[\text{Figure 12.5-1}\]
If the energy \( \omega \) is plotted against the momentum \( k \), then the velocity \( v \) is represented by the diagonal lines in fig. 12.5-1. Below the \( k \)-axis, the velocities dip into negative energies. In fact, the number of negative energy levels is infinite. If there was no lowest energy state, electrons, for example, would simply fill the endless number of negative energy states and there would be no positive energy electrons. This simply did not agree with what was observed, where positively energy electrons were abundant. Dirac solved the problem by recognizing that the equation he had developed represented a fermion. The only fermion known at the time was the ‘electron’. Electrons could not occupy the same state per the Pauli exclusion principle. Dirac hypothesized that Nature was endowed with a huge vacuum state, where electrons with negative energy filled the vacuum. This vacuum state was called the ‘Dirac sea’. Positive energy electrons could not enter the vacuum, since it was completely filled and no two electrons could occupy the same state. Occasionally, a negative energy electron is bumped out of the vacuum, creating a positive energy electron. This left a “hole” in the vacuum, which had a positive charge, since the absence of a negative energy electron is the same as a positively charged ‘something’. The positively charged ‘something’ was later identified as the ‘positron’ (anti-electron).

**12.5.1 Solving the One Direction Problem**

To solve the “one direction” problem, it was necessary to create two states, one for right and one for left moving particles:

\[
\frac{\partial}{\partial t} |\psi_L\rangle - \frac{\partial}{\partial x} |\psi_L\rangle = 0, \quad \frac{\partial}{\partial t} |\psi_R\rangle + \frac{\partial}{\partial x} |\psi_R\rangle = 0
\]

Putting the two ideas together and letting \( \hat{H} = \hat{a} \hat{p} \), then

\[
\hat{H} = \hat{a} \hat{p} \rightarrow i \frac{\partial}{\partial t} \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} = -i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial}{\partial x} \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} \rightarrow i \frac{\partial}{\partial t} |\psi\rangle = -i \hat{a} \frac{\partial}{\partial x} |\psi\rangle, \quad |\psi\rangle = \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix}
\]

where

\[
\hat{a} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}
\]

In the equation above, the particle is massless.

**12.5.2 Solving the Mass Problem**

To include particles with mass, the Hamiltonian must be extended to the form \( \hat{H} = \hat{a} \hat{p} + \hat{\beta} m \), where \( \hat{\beta} \) is an undetermined operator. Since

\[
E^2 = p^2 + m^2, \quad c = 1, \quad E = \hbar \omega = E = \omega, \quad \hbar = 1, \quad p^2 = k^2 \rightarrow \omega^2 = k^2 + m^2,
\]
then

\[ H^2 = (\hat{\alpha} k + \hat{\beta} m)(\hat{\alpha} k + \hat{\beta} m) = (\hat{\alpha} k)^2 + (\hat{\beta} m)^2 + (\hat{\alpha} \hat{\beta} + \hat{\beta} \hat{\alpha}) km \]

The equation above is satisfied if \( \hat{\alpha}^2 = \hat{\beta}^2 = I \) and \( \hat{\alpha} \hat{\beta} + \hat{\beta} \hat{\alpha} = 0 \). Dirac realized that the Pauli matrices satisfied the necessary conditions. Hence,

\[ \hat{\beta} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \rightarrow H|\psi\rangle = \hat{\alpha} \hat{p}|\psi\rangle + \hat{\beta} m|\psi\rangle \rightarrow i \frac{\partial}{\partial t} \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} \]

\[ = -i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial}{\partial x} \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} + m \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} \]

Furthermore,

\[ i \frac{\partial}{\partial t} |\psi_R\rangle + i \frac{\partial}{\partial x} |\psi_R\rangle = m |\psi_L\rangle, \quad i \frac{\partial}{\partial t} |\psi_L\rangle - i \frac{\partial}{\partial x} |\psi_L\rangle = m |\psi_R\rangle \]

Because of the mass term, the two equations are no longer independent. The approach got rid of the “left-right moving particle” problem and gave the particle mass.

### 12.5.3 Extending Dirac’s Ideas to Three Dimensions

Dirac’s ideas can be extended to three spatial dimensions. If the particle is massless, he hypothesized that

\[ H = \hat{\alpha} \cdot \hat{p} = \hat{\alpha}_x \hat{p}_x + \hat{\alpha}_y \hat{p}_y + \hat{\alpha}_z \hat{p}_z \]

The necessary and sufficient conditions for \( H^2 = \hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2 \) to be true are

\[ \hat{\alpha}_x^2 = \hat{\alpha}_y^2 = \hat{\alpha}_z^2 = I, \quad \hat{\alpha}_i \hat{\alpha}_j + \hat{\alpha}_j \hat{\alpha}_i = 0, \quad i \neq j, \quad i,j = x,y,z \]

Dirac realized that the operators that met the conditions above were the Pauli matrices:

\[ \hat{\sigma}_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \hat{\sigma}_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \hat{\sigma}_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \]

### 12.5.3.1 Pauli Matrices and Spin

The Pauli matrices are related to the spin of a particle. This meant that the energy of a massless particle was proportional to its spin along the direction of motion. But since there were two spin directions along any axis, particles could have either positive or negative energy depending on whether their spin was up or down along a given direction.
To include mass, the Hamiltonian must be extended to the form

$$\hat{H} = \hat{\alpha} \cdot \hat{p} + m\hat{\beta}$$

Again if $\hat{R}^2 = \hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2$, then the necessary and sufficient conditions for this are:

$$\hat{\alpha}_x^2 = \hat{\alpha}_y^2 = \hat{\alpha}_z^2 = \hat{p}^2 = 1, \quad \hat{\alpha}_i \hat{\alpha}_j + \hat{\alpha}_j \hat{\alpha}_i = 0, \quad \hat{\alpha}_i \hat{\beta} + \hat{\beta} \hat{\alpha}_i = 0, \quad i \neq j, \quad i, j = x, y, z$$

**12.5.3.2 Spin and Relativity**

Interestingly, there are no $2 \times 2$ or $3 \times 3$ matrices that satisfy all the conditions above. The smallest dimensional operators that satisfy these conditions are the $4 \times 4$ matrices

$$\alpha_i = \begin{bmatrix} \hat{\alpha}_i & 0 \\ 0 & -\hat{\alpha}_i \end{bmatrix}, \quad \beta = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = x, y, z$$

Recall that

$$i \frac{\partial}{\partial t} |\psi_R\rangle + i \frac{\partial}{\partial x} |\psi_R\rangle = m|\psi_L\rangle,$$

$$i \frac{\partial}{\partial t} |\psi_L\rangle - i \frac{\partial}{\partial x} |\psi_L\rangle = m|\psi_R\rangle \rightarrow i \frac{\partial}{\partial t} \langle\psi_L|\psi_R\rangle + i \frac{\partial}{\partial x} \langle\psi_L|\psi_R\rangle = m,$$

$$i \frac{\partial}{\partial t} \langle\psi_R|\psi_L\rangle - i \frac{\partial}{\partial x} \langle\psi_R|\psi_L\rangle = m, \quad \langle\psi_R|\psi_R\rangle = \langle\psi_L|\psi_L\rangle = 1,$$

Hence,

$$i \frac{\partial}{\partial t} \langle\psi_L|\psi_R\rangle + i \frac{\partial}{\partial x} \langle\psi_L|\psi_R\rangle = i \frac{\partial}{\partial t} \langle\psi_R|\psi_L\rangle - i \frac{\partial}{\partial x} \langle\psi_R|\psi_L\rangle$$

$$\rightarrow i \frac{\partial}{\partial t} \langle\psi_L|\psi_R\rangle + i \frac{\partial}{\partial x} \langle\psi_L|\psi_R\rangle - i \frac{\partial}{\partial t} \langle\psi_R|\psi_L\rangle + i \frac{\partial}{\partial x} \langle\psi_R|\psi_L\rangle = 0$$

Let $\langle\psi_L|\psi_R\rangle = \phi_1(x)$ and $\langle\psi_R|\psi_L\rangle = \phi_2(x)$, then

$$i \frac{\partial}{\partial t} \phi_1(x) + i \frac{\partial}{\partial x} \phi_1(x) - i \frac{\partial}{\partial t} \phi_2(x) + i \frac{\partial}{\partial x} \phi_2(x) = 0$$

If $\exists x(\langle\phi_1|\phi_2\rangle \neq 0)$ and $\phi_1(x), \phi_2(x)$ satisfy the equation above, then the particle has a mass at 'x'. Otherwise, it is a massless particle.

**12.6 Dirac’s Ideas Applied to QFT**

Field theory gives a slightly different interpretation to the description above. The annihilation operator 'Ψ' is given by
\[ \hat{\Psi} = \int \hat{c}^-(p)e^{ipx}dp, \]

where \( \hat{c}^- \) is an annihilation operator for a fermion. There are two parts to this:

\[ \hat{\Psi}_+ = \int_0^\infty \hat{c}^-(p)e^{ipx}dp, \quad \hat{\Psi}_- = \int_{-\infty}^0 \hat{c}^-(p)e^{ipx}dp, \]

where \( \hat{\Psi}_+ \) annihilates a positive energy and \( \hat{\Psi}_- \) annihilates a negative energy electron. Since the commutation relations for fermions do not make a distinction between creation and annihilation operators, then

\[ \hat{\Psi}_- = \int_{-\infty}^0 \hat{c}^-(p)e^{ipx}dp = \int_0^\infty \hat{b}^+(p)e^{-ipx}dp, \]

where \( \hat{b}^+ \) is a creation operator for a positron. In other words, the annihilation of an electron is the same as the creation of a positron. In QFT, the positron is considered a particle, not the absence of a particle. The idea of the 'Dirac Sea' is replaced with the idea that the 'positron is a particle'.

### 12.6.1 Identifying Particles by their Spin and Angular Momentum

Classically, the angular momentum \( 'L' \) of a particle is defined \( L = \langle r \times p \rangle \), where \( r \) is the radius vector and \( p \) is the linear momentum of the particle tangent to the curve at the point of the radius vector. If the particle is moving in a circle in the \( x, y \)-plane, its angular momentum is \( '|L_z|' \), a vector normal to the \( x, y \)-plane parallel to the \( z \)-axis. The Earth, rotating around the Sun, has classical angular momentum, sometimes called 'orbital momentum'. Spin, on the other hand, is conceptualized classically as rotation around an axis. The Earth, rotating around its axis, has 'classical spin'. Both the concepts of 'orbital angular momentum' and 'spin' are carried over into quantum mechanics, but their definitions are not identical. Mathematically, the two terms mean something slightly different in classical vs. quantum physics.

#### 12.6.2 Orbital Angular Momentum

Mathematically, what is the difference between orbital angular momentum and spin? If a vector does not change its position, but only its direction, this is paramount to 'spin'. A vector changing its position, but not its direction is paramount to 'orbital angular momentum' (see fig. 12.6.2-1).
Vectors that change both position and direction are represented by \( J = S + L \).

Classically,

\[
L_x = yp_z - zp_y, \quad L_y = zp_x - xp_z, \quad L_z = xp_y - yp_x
\]

Quantum mechanically, an electron is conceived as moving around the nucleus in a standing wave motion, where

\[
\lambda n = 2\pi r, \quad p = \frac{\hbar}{\lambda} \rightarrow pr = L = nh, \quad n \in N
\]

Hence, the orbital angular momentum \( L \) is quantized, coming in integer multiples of \( \hbar \). Moreover, since \( p \) is quantized,

\[
[\hat{X}, \hat{p}_x] = xp_x - p_xx = i, \quad \hbar = 1
\]

Recall that (see Chapter 11, Sec 11.14.1)

\[
[\hat{L}_x, \hat{L}_y] = i\hbar \hat{L}_z, \quad [\hat{L}_y, \hat{L}_z] = i\hbar \hat{L}_x, \quad [\hat{L}_z, \hat{L}_x] = i\hbar \hat{L}_y
\]
Quantum mechanically, orbital angular momentum cannot be measured simultaneously along the \(x, y, z\)-directions. Recall the definitions \(\hat{L}_+ = \hat{L}_x + i\hat{L}_y\) and \(\hat{L}_- = \hat{L}_x - i\hat{L}_y\) and remember that 

\[
[\hat{L}_+, \hat{L}_z] = -\hbar\hat{L}_+, \quad [\hat{L}_-, \hat{L}_z] = \hbar\hat{L}_-
\]

If the operator \(\hat{L}_z\) acts on a state \(|m\rangle\) i.e. \(\hat{L}_z|m\rangle = m|m\rangle\), it returns the value \(m\) of the angular momentum measured along the \(z\)-axis.

If \(\hat{L}_z\hat{L}_+|m\rangle\), then

\[
[\hat{L}_+, \hat{L}_z] = \hat{L}_+\hat{L}_z - \hat{L}_z\hat{L}_+ = -\hbar\hat{L}_+ \rightarrow \hat{L}_z\hat{L}_+ = \hat{L}_+\hat{L}_z + \hat{L}_+,
\]

\[
h = 1 \rightarrow \hat{L}_z\hat{L}_+|m\rangle = \hat{L}_+\hat{L}_z|m\rangle + \hat{L}_+|m\rangle = m\hat{L}_+|m\rangle + \hat{L}_+|m\rangle = (m + 1)\hat{L}_+|m\rangle
\]

Hence, \(\hat{L}_z\hat{L}_+\) acting on \(|m\rangle\) increases the angular momentum by one unit in the \(z\)-direction. Conversely, \(\hat{L}_-\) decreases the angular momentum by a unit measured along the \(z\)-axis. The operator \(\hat{L}_+\) operating on an eigenstate of angular momentum can continue to increase the angular momentum until the vector is aligned with the \(z\)-axis. If the value of the angular momentum along the \(z\)-axis is \(m_+\), then \(\hat{L}_+|m_+\rangle = |0\rangle\), since \(m_+\) is a maximum (see fig. 12.6.2-2). Similarly, \(\hat{L}_-|m_-\rangle = |0\rangle\), since \(m_-\) is a minimum.

By the Pythagorean theorem, \(\hat{L}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2\). Note that

\[
\hat{L}_-\hat{L}_+ = (\hat{L}_x - i\hat{L}_y)(\hat{L}_x + i\hat{L}_y) = \hat{L}_x^2 + \hat{L}_y^2 + (i\hat{L}_x\hat{L}_y - i\hat{L}_y\hat{L}_x) = \hat{L}_x^2 + \hat{L}_y^2 + i[\hat{L}_x, \hat{L}_y]
\]

\[
= \hat{L}_x^2 + \hat{L}_y^2 - \hat{L}_z \rightarrow \hat{L}^2 = \hat{L}_-\hat{L}_+ + \hat{L}_z + \hat{L}_z
\]

Consider \(\hat{L}^2|m_+\rangle\), then

\[
\hat{L}^2|m_+\rangle = \hat{L}_-\hat{L}_+|m_+\rangle + \hat{L}_z|m_+\rangle + \hat{L}_z^2|m_+\rangle = m_+|m_+\rangle + m_+^2|m_+\rangle = (m_+^2 + m_+)|m_+\rangle
\]

Comparing \(\hat{L}^2 = m_+^2 + m_+\) with \(L_z^2 = m_+^2\) illustrates the difference between classical and quantum results. Classically, if all the momentum is along the \(z\)-axis, then \(L_z^2 = m_+^2\). Quantum mechanically, if all the momentum is along the \(z\)-axis, then \(\hat{L}^2 = m_+^2 + m_+\), which includes the classical result \(m_+^2\) plus an extra term \(m_+\). There is an extra amount of angular momentum \(m_+\) distributed along the \(x\)- and \(y\)-axes, but quantum mechanics forbids measuring how much of \(m_+\) lies along the \(x\)- and how much along the \(y\)-axes, since the angular momentum operators along different axes do not commute.

**12.6.3 Spin for Fermions**

Whether a particle is a boson or a fermion is determined by spin. All fermions are spin-\(1/2\) particles, with spin coming in integer multiples of \(\hbar/2\). For example, the electron
has only two spin states: \( |\frac{h}{2}\rangle \) and \( |\frac{-h}{2}\rangle \). The spin operator \( \hat{S}_i \) has the same commutation rules as \( \hat{L}_i \), the angular momentum operator i.e.

\[
[\hat{S}_x, \hat{S}_y] = i\hat{S}_z, \quad [\hat{S}_y, \hat{S}_z] = i\hat{S}_x, \quad [\hat{S}_z, \hat{S}_x] = i\hat{S}_y, \quad \hat{S}_\pm = \hat{S}_x \pm i\hat{S}_y, \quad h = 1,
\]

where \( \hat{S}_+ \) and \( \hat{S}_- \) are the raising and lowering operators for spin and

\[
|u_z\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad |d_z\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad |u_x\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad |d_x\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad |u_y\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}, \quad |d_y\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}
\]

where \( |u_i\rangle \)'s are the eigenvectors along a given axis. For instance, \( \hat{S}_+ |u_z\rangle = |0\rangle \), \( \hat{S}_+ |d_z\rangle = |u_z\rangle \), \( \hat{S}_- |u_z\rangle = |d_z\rangle \) and \( \hat{S}_- |d_z\rangle = |0\rangle \).

### 12.6.4 Spin for Bosons

Bosons have integer spin. For example, a spin-1 particle has three possible eigenvalues: ‘\( m = -1,0,1 \)’. So, if

\[
\hat{S}_z = i \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \hat{S}_y = i \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad \hat{S}_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}
\]

and \( \hat{S}_z \) acts on the following states

\[
|m = 1\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \\ 0 \end{bmatrix}, \quad |m = -1\rangle = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix}, \quad |m = 0\rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
\]

the results would give the eigenvalues ‘\( 1,-1,0 \)’ respectfully.

### 12.6.5 Spin and Energy

The electron can be in four states, signified by ‘\( \omega |u\rangle, -\omega |u\rangle, \omega |d\rangle, -\omega |d\rangle \)’ respectfully. The electron could be in an ‘up-spin state’ and have positive energy ‘\( \omega \)’ or in an ‘up-spin state’ and have negative energy ‘\( -\omega \)’ etc. Assuming the electron has zero momentum \( (k = 0) \), then the Dirac equation can be written

\[
i \frac{\partial |\psi\rangle}{\partial t} = m\beta |\psi\rangle = m \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} |\psi\rangle, \quad |\psi\rangle = \begin{bmatrix} |\psi_R\rangle \\ |\psi_L\rangle \end{bmatrix} \rightarrow i \frac{\partial |\psi_R\rangle}{\partial t} = m |\psi_L\rangle, \quad i \frac{\partial |\psi_L\rangle}{\partial t} = m |\psi_R\rangle
\]
Adding and subtracting the equations on the right above gives

\[ i \frac{\partial}{\partial t} |\psi_R + \psi_L \rangle = m |\psi_R + \psi_L \rangle, \quad i \frac{\partial}{\partial t} |\psi_R - \psi_L \rangle = -m |\psi_R - \psi_L \rangle \]

But \( i \frac{\partial}{\partial t} |\psi_R + \psi_L \rangle \) and \( i \frac{\partial}{\partial t} |\psi_R - \psi_L \rangle \) are associated with the energy. To see this, note that if \( |\psi_R \rangle \propto e^{i k x} e^{-i \omega t} \) and if \( k = 0 \), then \( |\psi_R \rangle \propto e^{-i \omega t} \). Likewise, \( |\psi_L \rangle \propto -e^{-i \omega t} \).

Hence,

\[ i \frac{\partial}{\partial t} |\psi_R + \psi_L \rangle = i \frac{\partial}{\partial t} (e^{-i \omega t} - e^{-i \omega t}) = m(e^{-i \omega t} - e^{-i \omega t}) \rightarrow i(-i \omega e^{-i \omega t} + i \omega e^{-i \omega t}) \]

\[ = m(e^{-i \omega t} - e^{-i \omega t}) \rightarrow \omega = m \]

\[ i \frac{\partial}{\partial t} |\psi_R - \psi_L \rangle = i \frac{\partial}{\partial t} (e^{-i \omega t} + e^{-i \omega t}) = -m(e^{-i \omega t} + e^{-i \omega t}) \rightarrow i(-i \omega e^{-i \omega t} - i \omega e^{-i \omega t}) \]

\[ = -m(e^{-i \omega t} + e^{-i \omega t}) \rightarrow \omega = -m \]

The analysis above is intended to show, heuristically, that whether the electron is spin-up or spin-down, the energy of the electron is the same, except for a sign, either positive or negative. However, the mathematics is a bit dubious, since, in the first equation above, the quantity in parentheses is zero. Evidently, the idea is for the quantity \( e^{-i \omega t} - e^{-i \omega t} \) to cancel on both sides of the equation. The second equation, however, does not suffer the same problem.

12.7 The Role of the Lagrangian in QFT

In classical physics, the equations of motion are given in terms of a Lagrangian density function \( \mathcal{L} = \mathcal{L}(q(t), \dot{q}(t), t) \), where \( \mathcal{L} \) is a ‘functional’ of the single function \( q \) and its derivatives. Similarly, the equations of motion for fields are given in terms of a Lagrangian density function \( \mathcal{L} = \mathcal{L}(\phi, \partial_\mu \phi) \), where, in this case, \( \mathcal{L} \) is a ‘functional’ of the single field \( \phi \) and its derivatives with respect to \( t, x, y, z \). In either case, the equations of motion are governed by the action \( \mathcal{S} \) defined as

\[ \mathcal{S}(q) = \int_{t_1}^{t_2} \mathcal{L}(q(t), \dot{q}(t), t) \, dt, \quad \mathcal{S}(\phi) = \int_{\mu_1}^{\mu_2} \mathcal{L}(\phi, \partial_\mu \phi) \, d\mu, \]

where \( \mathcal{S}(q) \) is the action for a single particle and \( \mathcal{S}(\phi) \) is the action for a single field.

12.7.1 The Equations of Motion for a Field

The action for a field is minimized if \( \delta \mathcal{S} = 0 \). In other words, the action over the interval ‘\( \mu_2 < \mu < \mu_1 \)’ remains stationary if and only if

\[ \partial_\mu \left( \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} \right) = \frac{\partial \mathcal{L}}{\partial \phi}, \quad \mu = t, x, y, z \]
In the case of a classical particle, the action represents the path of a single particle, or, in the case of a field, describes the motion of the field.

For example, let

\[
\mathcal{L} = \frac{1}{2} \left( \frac{\partial \phi}{\partial t} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial y} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial z} \right)^2 - V(\phi), \quad V(\phi) = \frac{m^2}{2} \phi^2,
\]

then

\[
\partial_\mu \left( \frac{\partial \mathcal{L}}{\partial \phi} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0 \Rightarrow \frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} - \frac{\partial^2 \phi}{\partial z^2} = -m^2 \phi
\]

Now let \( \phi = e^{i(k_x x + k_y y + k_z z - \omega t)} \), then

\[
\frac{\partial^2 \phi}{\partial t^2} = -\omega^2 e^{i(k_x x + k_y y + k_z z - \omega t)}, \quad \frac{\partial^2 \phi}{\partial x^2} = k_x^2 e^{i(k_x x + k_y y + k_z z - \omega t)},
\]

\[
- \frac{\partial^2 \phi}{\partial y^2} = k_y^2 e^{i(k_x x + k_y y + k_z z - \omega t)}, \quad - \frac{\partial^2 \phi}{\partial z^2} = k_z^2 e^{i(k_x x + k_y y + k_z z - \omega t)}
\]

So,

\[
(-\omega^2 + k_x^2 + k_y^2 + k_z^2)e^{i(k_x x + k_y y + k_z z - \omega t)} = -m^2 e^{i(k_x x + k_y y + k_z z - \omega t)} \rightarrow -\omega^2 + k^2 = -m^2
\]

\[
\omega^2 = k^2 + m^2, \quad k^2 = k_x^2 + k_y^2 + k_z^2
\]

From relativity, \( E^2 = p^2 + m^2 \), \( k = p \), \( c = 1 \), where \( \mathcal{L} \), in this case, is Lorentz invariant, so \( \phi \) must be a scalar function. Note that \( \delta \mathcal{L} \) is linear in \( \phi \). The waves represented by \( \phi \) may interfere, but do not scatter, maintaining their integrity throughout the process. But suppose that \( \delta \mathcal{L} \) is non-linear. For instance, let \( V(\phi) = m^2 \phi^2/2 + g \phi^3 \), then \( \partial V / \partial \phi = m^2 \phi + 3 g \phi^2 \). In this case, \( \delta \mathcal{L} \) contains non-linear terms in \( \phi \) indicating that the field scatters.

Suppose \( \mathcal{L} \) is made up of two fields \'\( \phi, \rho \)\':

\[
\mathcal{L} = \frac{1}{2} (\partial_\mu \phi)^2 + \frac{1}{2} (\partial_\mu \rho)^2 - \frac{m^2}{2} \phi^2 - \frac{M^2}{2} \rho^2, \quad \mu = t, x, y, z,
\]

then

\[
\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} - \frac{\partial^2 \phi}{\partial z^2} = -m^2 \phi, \quad \frac{\partial^2 \rho}{\partial t^2} - \frac{\partial^2 \rho}{\partial x^2} - \frac{\partial^2 \rho}{\partial y^2} - \frac{\partial^2 \rho}{\partial z^2} = -M^2 \rho
\]

In this case, the fields \'\( \phi, \rho \)\' do not interact. But, suppose

\[
V(\phi) = m^2 \phi^2/2 + M^2 \rho^2/2 + \rho \phi^2,
\]
then
\[ \frac{\partial^2 \phi}{\partial t^2} - \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial y^2} - \frac{\partial^2 \phi}{\partial z^2} = -m^2 \phi - 2\rho \phi, \quad \frac{\partial^2 \rho}{\partial t^2} - \frac{\partial^2 \rho}{\partial x^2} - \frac{\partial^2 \rho}{\partial y^2} - \frac{\partial^2 \rho}{\partial z^2} = -M^2 \rho - \phi^2 \]

The two equations above are no longer independent, since the two fields interact with one another.

### 12.8 The Path Integral Approach

The path integral approach to QFT conceives a wave/particle, such as an electron, as travelling along all paths between two space-time events. In other words, the wave/particle travels along all world lines (an infinite number of them) between the two space-time events.

**12.8.1 Phasors**

The idea begins with the concept of a ‘phasor’. The equation of a phasor ‘ψ’ is \( \psi = e^{i\omega t} \). Pictorially, a phasor can be imagined as an ‘arrow’ that rotates counterclockwise as the wave moves from point ‘a’ to point ‘b’ (see fig. 12.8.1-1). It also can be thought of as rotating on a unit circle. As the wave makes a complete cycle, the phasor rotates by \( 2\pi \). The phasor points in the same direction at ‘a’ as it does at ‘b’.

Figure 12.8.1-1

If two identical waves are in phase, their phasors add i.e. \( \Uparrow \Uparrow = 2 \uparrow \), \( \Downarrow \Downarrow = 2 \downarrow \). If the two waves are \( \pi \text{ [rads]} \) out of phase, then \( \Uparrow \Downarrow = 0 \), the phasors cancel. Phasors have unit length. If two waves are out of phase, the difference between their phases is measured by the angle between their respective phasors. Note that the phasor is proportional to the amplitude of the wave. And since light has energy \( E = h\nu \), the frequency of the light is equal to the phasor. So, if two light waves travel from ‘a’ to ‘b’ and are out of phase by an angle \( \theta \neq 0 \), then \( \theta \propto U^2 \), where \( U \) is the amplitude of the wave.

**12.8.2 Calculating the Amplitude with Path Integrals**

In QFT, as in quantum mechanics, it is necessary to calculate an ‘amplitude’, since the square of the amplitude is proportional to the probability that a given process occurs. The amplitude is given in terms of the action:
$$U = \sum_{n} e^{\frac{-i}{\hbar} \tilde{\phi}^{(\mu_2)} L_n (\Phi, \partial \mu \Phi) d \mu}$$

where fields are replaced by creation and annihilation operators and the sum signifies all the possible ways of moving from the state \(\tilde{\phi}(\mu_1)\) to \(\tilde{\phi}(\mu_2)\). Since a particle/wave takes all possible paths,

$$e^{\frac{-i}{\hbar} \tilde{\phi}^{(\mu_2)} L_j (\Phi, \partial \mu \Phi) d \mu}$$

represents the \(j^{th}\) phasor. The sum taken over all the phasors results in a resultant phasor, proportional to \(U^2\), associated with the probability that the system begins at \(\tilde{\phi}(\mu_1)\) and ends at \(\tilde{\phi}(\mu_2)\). Note that the field \(\tilde{\phi}\) takes on different values for different paths.

A heuristic example is shown in fig. 12.8.2-1. A photon is emitted from ‘S’ and ends up at ‘P’. Classically, the photon would take the path ‘SGP’, the shortest path. But, according to Feynman, it takes all possible paths, 13 of which are shown in the figure. The phasors for each path are shown below the grid. Note that the phasors associated with the paths close to the classical path tend to reinforce, while those farther away tend to cancel. Paths close to the classical path have a higher probability of occurring than those farther away. This was Feynman’s interpretation of the two slot interference experiment in quantum mechanics. An electron could take any path to a point on a detector. Reinforced phasors give a high probability of the electron landing there.
Conversely, those points, where the phasor resultants were low, have a smaller probability of landing at a given point on the detector.

### 12.8.2.1 Sum over Histories

To simplify matters, divide up space-time into small cells (see fig. 12.8.2.1-1).

Let

\[
\mathcal{S}(\hat{\phi}) = \int_{\hat{\phi}(\mu_1)}^{\hat{\phi}(\mu_2)} \mathcal{L}(\hat{\phi}, \partial_\mu \hat{\phi}) d\mu \Rightarrow \sum_{ij} \mathcal{L}_{ij}
\]

where \(\mathcal{L}_{ij}\) signifies the value of the Lagrangian in the \(i,j\)-cell. If \(\hat{\phi}(\mu_{ij})\) gives the value of the field in the \(i,j\)-cell and if \(\hat{\phi}(\mu_1) = \hat{\phi}(\mu_{11})\) and \(\hat{\phi}(\mu_2) = \hat{\phi}(\mu_{33})\), the initial state of a particle being \(\hat{\phi}(\mu_{11})\), then what is the probability, after a time, that the particle ends up in a final state \(\hat{\phi}(\mu_{33})\)? Let \(\psi_1 = A_1 e^{i/\hbar \mathcal{S}_1}\), where \(e^{i/\hbar \mathcal{S}_1}\) is the phasor for the first path. Since a particle can travel many paths at once,

\[
e^{-i\mathcal{S}_1} + e^{-i\mathcal{S}_2} + e^{-i\mathcal{S}_3} + \cdots = A_b e^{i\omega t \text{sum}},
\]

where \(A_b\) is the amplitude of the sum. As the number of paths approaches infinity, \(A_b A_b^\dagger\) becomes proportional to the final state at event \('b'\). The amplitude, then, is the sum over all possible ways of moving from the initial field \(\hat{\phi}(\mu_{11})\) to the final field \(\hat{\phi}(\mu_{33})\). This approach is often referred to as the 'sum over histories'. For example,

\[
\hat{\phi}(\mu_{11}) \rightarrow \hat{\phi}(\mu_{12}) \rightarrow \hat{\phi}(\mu_{13}) \rightarrow \hat{\phi}(\mu_{23}) \rightarrow \hat{\phi}(\mu_{33})
\]

represents a possible history of a particle, but only one history. The amplitude is found by summing over every possible history. Depending on the Lagrangian, not all paths are possible. What paths are possible depends on the symmetry governing a particular Lagrangian. For example, if charge is conserved in any system, the Lagrangian must generate only paths that conserve charge.
12.8.2.2 The Meaning of the Lagrangian in the Path Integral Approach

The quadratic terms in the Lagrangian represent movement of a particle from one point to another. For instance,

\[ \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 \approx \frac{1}{2} \left( \phi(t, x) - \phi(t, x') \right)^2, \quad \frac{1}{2} \left( \frac{\partial \phi}{\partial t} \right)^2 \approx \frac{1}{2} \left( \phi(t, x) - \phi(t', x) \right)^2 \]

Terms like \( \phi^2(t, x) \) or \( \phi^2(t', x) \) do little but annihilate a particle then create one or create a particle and annihilate it at a place and time. But terms involving, for example, \( \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 \) move a particle from \( x \) to \( x' \) or vice versa, where \( x, x' \) are neighboring points. Recall that \( e^\varepsilon = 1 + \varepsilon \), \( \varepsilon \ll 1 \) or to higher powers

\[ e^\varepsilon = 1 + \varepsilon + \varepsilon^2 / 2! + \varepsilon^3 / 3! + \cdots \]

Hence, \( e^{-i \sum \phi(x) \dot{\phi}(x')} \) can be expanded in the following form:

\[ 1 - i \sum \phi(x) \dot{\phi}(x') - \sum \dot{\phi}(x) \dot{\phi}(x') \frac{\sum \dot{\phi}(x) \dot{\phi}(x'')}{2!} + \cdots \]

The second term in the expansion represents operations like annihilating a particle at \( x \) and creating one at \( x' \) or vice versa. These operations involve moving a particle to neighboring space-time points. The third term involves moving a particle two space-time points or, if the fields are different, two particles can be moved to a neighboring space-time point. If

\[ \mathcal{L} = \frac{1}{2} \left( \frac{\partial \phi}{\partial t} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial y} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial z} \right)^2 - m^2 \frac{\phi^2}{2}, \]

then the paths containing the \( 'm^2' \) term are weighed heavier. If the Lagrangian \( \mathcal{L} \) looks like this:

\[ \mathcal{L} = \frac{1}{2} \left( \frac{\partial \phi}{\partial t} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial y} \right)^2 - \frac{1}{2} \left( \frac{\partial \phi}{\partial z} \right)^2 - m^2 \frac{\phi^2}{2} + g \phi^3, \]

then the \( g \phi^3 \) term represents creation and annihilation in threes, which allows, for example, a particle to split into two particles and recombine into one etc. Very complicated system processes can be described by adding higher terms to the Lagrangian. In essence, the Lagrangian codifies particle interactions and movements in a condensed fashion.

12.9 Dirac’s Equations and a Lagrangian

The Dirac equation can be written in the form of a Lagrangian:

\[ \mathcal{L} = \frac{i}{2} \left( \Phi^\dagger \frac{\partial \Phi}{\partial t} + \Phi^\dagger \alpha \frac{\partial \Phi}{\partial x} \right) - m \Phi^\dagger \beta \Phi, \]

364
where the fields have been replaced by field operators, made up of creation and annihilation operators. Note that there is no derivative of $\bar{\Psi}^\dagger$ in $\mathcal{L}$. Hence, by Euler’s equation

$$\frac{\partial}{\partial x_\mu} \left( \frac{\partial \mathcal{L}}{\partial \partial x_\mu \bar{\Psi}^\dagger} \right) - \frac{\partial \mathcal{L}}{\partial \bar{\Psi}^\dagger} = 0 \rightarrow \frac{\partial}{\partial x_\mu} \left( \frac{\partial \mathcal{L}}{\partial \partial x_\mu \bar{\Psi}^\dagger} \right) = 0$$

since $\frac{\partial \mathcal{L}}{\partial \bar{\Psi}^\dagger} = 0$. Therefore,

$$\alpha_\mu \frac{\partial \bar{\Psi}}{\partial x_\mu} + m\beta \bar{\Psi} = 0 \rightarrow \left( \alpha_\mu \frac{\partial}{\partial x_\mu} + m\beta \right) \bar{\Psi} = 0,$$

which is the Dirac equation.

If terms are added to the Dirac Lagrangian that contain, say, $\phi^3$, and $\phi$ is of the form $\phi \propto (a^+ + a^-)$, then $\phi^3$ represents particles coming and going in triplets.

Two particles can come in, be annihilated, and one particle created going out. One particle can be annihilated and two created. Three particles can be created or three can be annihilated (see fig. 12.9-1). All this happens at a point ‘$x$’. This is the essence of Feynman diagrams in QFT.

**12.9.1 Dirac Fields**

The Dirac fields in QFT are defined thusly:

$$\bar{\Psi} = \bar{c}_e^- (+) + \bar{c}_e^- (-), \quad \bar{c}_e^- (-) = \bar{c}_e^+ (+), \quad \bar{\Psi}^\dagger = \bar{c}_e^+ (+) + \bar{c}_e^+ (-), \quad \bar{c}_e^+ (-) = \bar{c}_e^- (+),$$

where $\bar{c}_e^- (+)$ signifies annihilating a positively charged electron and $\bar{c}_e^- (-)$ signifies annihilating a negatively charged electron. Whereas $\bar{c}_e^- (+)$ creates a positively charged electron, $\bar{c}_e^+ (-)$ creates a negatively charged electron. But, $\bar{c}_e^- (-) = \bar{c}_e^+ (+)$ i.e. annihilating a negatively charged electron is the same as creating a positron and $\bar{c}_e^+ (+) = \bar{c}_e^- (-)$ signifies that creating a positron is the same as annihilating an electron. This is how the Dirac fields are interpreted in QFT.

For instance, basic interactions within quantum electrodynamics come in the form

$$e\bar{\Psi}^\dagger \bar{\Psi} \lambda_0 + \bar{\Psi}^\dagger \bar{\Psi} a \lambda_i.$$
where $\hat{\Lambda}_\mu$ signifies the photon field i.e. $\hat{\Lambda}_\mu$ contains creation and annihilation operators for photons.

![Figure 12.9.1-1](image)

The interaction shown in fig. 12.9.1-1 a) illustrates an electron coming in, being annihilated at $x$, then an electron created and a photon emitted. This is a valid interaction within quantum electrodynamics. Virtually any electrodynamic interaction is permitted so long as charge is conserved. The interaction shown in Fig. 12.9.1-1 b) shows an electron coming in, being annihilated at $x$, then a positron created and a photon emitted. Charge is not conserved within this interaction, so it cannot happen in Nature. The Lagrangian governing quantum electrodynamics is $U(1)'$ invariant, electric charge is conserved. Any Lagrangian that describes quantum electrodynamics must be $U(1)$ invariant. Mathematically, $U(1)$ represents a group of transformations.

Each particle is represented by a different operator; the operators $\hat{\Psi}_p, \hat{\Psi}_n, \hat{\pi}^+_+, \hat{\pi}^-_-, \hat{\pi}^+_0$, for example, represent the proton, neutron, positive $\pi$-meson, negative $\pi$-meson and neutral $\pi$-meson respectfully. The interaction $\hat{\Psi}_p^+ \hat{\Psi}_n \hat{\pi}^+_+$ is a valid electrodynamic interaction. A positive $\pi$-meson and a neutron are annihilated and a proton created. Charge is conserved. Whereas, $\hat{\Psi}_p \hat{\Psi}_p^+ \hat{\Psi}_n \hat{\pi}^+_+$ is not a valid interaction.

### 12.10 Concluding Remarks

By the turn of the twentieth century, physics had grappled with an abundance of new discoveries, the most dramatic being the realization that light and the electron were not entirely different in their makeup. At one time, light was thought to arise from the oscillations of the electromagnetic field and that the electron was a particle. Waves were spread out over an entire space and particles were minutely located, their masses concentrated at a point. But Einstein showed that, under certain circumstances, light behaved like a particle and de Broglie confirmed that, under other circumstances, particles exhibited wave properties.

The first attempts to explain wave/particle behavior were made by Schrödinger, Heisenberg and their contemporaries. The results were mixed. Particles no longer took definite paths. Concrete physical outcomes were replaced by probabilistic predictions, certainty with uncertainty. No physical reality could be given to the particle/waves. Whether a thing was a particle or a wave depended on how it interacted with its environment. And “the environment” was synonymous with “taking a measurement”. Moreover, the quantum mechanical explanation for particle/waves was restricted to non-
relativistic physics. Paul Dirac showed that “particle spin” was a manifestation of Einstein’s special theory of relativity, leading to the discovery of the antiparticle. But even Dirac’s brilliant insights left electrons in negative energy states, which did not agree with observations. Moreover, Dirac showed that particles interacted and decayed into a myriad of different particles, not necessarily conserving the number of particles.

To handle the additional demands, quantum field theory was developed. It took a radical new direction. Quantum operators that acted on fields were replaced by field operators that created and annihilated particles, which gave a new interpretation to the wave/particle problem. The ‘field’ became fundamental. Fields became the “stuff” from which all particles were created or destroyed. There were matter fields (proton, electron etc.), fields for the photon and even a field that gave particles mass, the ‘Higgs field’. The non-conservation of the number of particles was no longer an issue and particle interactions and decays could be explained by the appropriate use of creation and annihilation by field operators. The problem of the 'Dirac sea' was partially solved by recognizing that destroying a negative electron was identical to creating a positron and creating an electron was identical to destroying a positron. Hence, the problem of negative electron energies was resolved in a somewhat acceptable manner by considering the positron as a particle.

But this did not settle all outstanding issues. QFT is still plagued by occasional infinities that crop up in its mathematics from time to time. If the coupling constants in the infinite sums that occur in QFT are small enough, the sums will converge, but this does not always happen. Nevertheless, the development of QFT allowed the creation of the 'standard model of particle physics', which explains most of what is known about the current state of our Universe and is the subject of the next chapter.

The artistry in QFT comes from choosing an appropriate Lagrangian. The events that can happen in Nature are governed the symmetries associated with the Lagrangian. In other words, the symmetries dictate which processes can and cannot occur in Nature. The way physicists see it today, the Universe could be explained by a giant Lagrangian. There are more than 100 particles considered somewhat fundamental and presumably would be included in the Lagrangian along with a multitude of particle interactions. The single Lagrangian describing all of this would easily fill over two closely space pages, a colossal mess. The untidy nature of this approach begs the question whether it is useful to produce a field theory that describes the Universe or whether a new approach is needed. At this point, a suitable answer to this question remains wanting.
Chapter 13
The Standard Model of Particle Physics

“I often feel a discomfort, a kind of embarrassment, when I explain elementary-particle physics to laypeople. It all seems so arbitrary - the ridiculous collection of fundamental particles, the lack of pattern to their masses.”

— Leonard Susskind

13.0 Introduction

The standard model includes three fundamental concepts: ‘particles’, ‘fields’ and ‘forces’ (see fig. 13-1). For every field, there is a particle. In quantum field theory (QFT), fields are fundamental and particles arise as disturbances in the fields. These disturbances arise by ‘quantization’. The idea begins with the ‘vacuum state’. In classical physics, the vacuum is roughly synonymous with nothingness, no particles or fields, just nothing. In quantum physics, the ‘vacuum state’ is not very well understood, but generally conceived as a state that is not truly empty, but instead, contains fleeting electromagnetic waves and particles that pop into and out of existence in a state of equilibrium.

Particles ↔ Fields

\[ \text{Forces} \]

Figure 13-1

Physics is largely a study of things that remain constant in a sea of change. Constancy is synonymous with conservation laws, which, in turn, are associated with symmetries of one kind or another. In the standard model, the vacuum state is in unstable equilibrium. In other words, it is described by a symmetry, but because the symmetry is unstable, the symmetry is broken, creating disturbances that are called ‘particles’.

There are two types of fields: ‘matter’ and ‘force’ fields. Matter fields are associated with the particles directly observed, like electrons and protons. Force carrying particles are unobservable, but are conceived as particles exchanged between the matter particles. These exchanges fashion the interactions amongst the matter particles.

The standard model is divided into three sectors: the ‘electrodynamic’, the ‘weak’ and the ‘strong’ sectors. Each sector represents a different theory. The theories are similar, described in terms of the interactions between the matter particles mediated by the force carrying particles. Each sector is governed by a symmetry. The equations associated with a sector remain invariant under a given symmetry transformation. The symmetries that govern the sector determine which particle interactions can take place and which cannot. Such theories are generally referred to as ‘gauge theories’. The general theory of relativity is not, strictly speaking, a gauge theory. It does not describe
gravity in terms of a force carrying particle (the graviton) that interacts with a matter particle. The challenge, if physics is to be unified into a single theory, is to find some way of reconciling the gauge theories of the standard model with the theory of gravity.

Interestingly, force fields are not independent, but exist because of matter fields. Each force field must have a source. The source is a matter particle. The electrostatic force around an electron is a force field whose source is the electron. Classically, the energy contained in the force field is proportional to the electric charge i.e.

\[ \text{field energy} \approx (eE)^2 \]

The total energy over an entire region is simply

\[ \text{Total Energy} \approx \int (eE)^2 \, dV, \]

where \( V \) is a volume and \( E \) is the electric field. If there are two electrons, then the field energy is proportional to \((e_1|E_1| + e_2|E_2|)^2\) i.e.

\[ \text{Total Energy} \approx \int [(e_1|E_1|)^2 + (e_2|E_2|)^2 + 2e_1e_2\langle E_1|E_2 \rangle] \, dV \]

If the two electrons are far apart, then \( \langle E_1|E_2 \rangle \approx 0 \) and the third term in the integrand can be ignored, making the total energy equal to the sum of the energies associated with the two electrons. Of note, \((e_1|E_1|)^2 + (e_2|E_2|)^2\) is proportional to the masses of the electrons. The term \('2e_1e_2\langle E_1|E_2 \rangle'\) acts as the interaction term and is roughly equivalent to the inverse square law for the electrostatic force \(F_e = Ke_1e_2/r^2\). As the electrons get closer together, their fields interact creating a stronger and stronger force as the distance between them diminishes.

Given two protons \('P_1, P_2'\) and an electron circling around one of the protons, the electron would assume the lowest energy state around one of the protons. If the protons are far enough apart, the energy of the system is just that of a single hydrogen atom. There is a probability that the electron is in the lowest energy state around \( P_1 \) and a probability it is in the lowest energy state around \( P_2 \). If the protons are brought closer together, quantum mechanically, the protons morph into a superposition of states, where the electron jumps between the two protons, changing the wave function of the electron. The jumping of the electron reduces the repulsive electrostatic force between the protons. The closer the protons are to each other, the more the electron jumps, diminishing further the electrostatic force between the protons. The idea that a force carrying particle is exchanged between two matter particles establishes a relationship between forces and particles that does not exist in classical physics, but is fundamental in quantum physics. In the analysis above, the electron plays the role of the force carrier and the proton the role of the matter particle. In the standard model, the electron is actually a matter particle, but the idea serves to show how force carrying particles exchanged between matter particles can change the behavior of the matter particles.

Quantum mechanically, an electron can be thought of as a particle surrounded by a cloud of photons, which are emitted and absorbed by the electron. So long as the
electron emits just as many photons as it absorbs, within the constraints of the uncertainty principle, the electron remains in equilibrium. If two electrons are brought close together, every once and a while, a photon is exchanged between them and this creates a force between the two electrons. This force is the quantum mechanical substitute for the electrostatic inverse square law of classical physics.

13.1 The Standard Model Particle Classification Scheme

By the mid 1960’s about 30 varieties of relatively stable particles had been discovered. For convenience, table 2.2.5.2.3-1 has been reconstructed here. It shows a particle classification scheme developed prior to the discovery of the quark. In 1964, physicists Murray Gell-Mann and George Zweig hypothesized that hadrons were composed of combinations of more elementary particles that Gell-Mann named ‘quarks’. Like other elementary particles, each quark has an antiquark [70,77].

<table>
<thead>
<tr>
<th>QUARKS</th>
<th>CHARM</th>
<th>TOP</th>
<th>GLUON</th>
<th>HIGGS BOSON</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP</td>
<td>mass 2.3 MeV/c^2</td>
<td>charge 2/3</td>
<td>spin 1/2</td>
<td>126 GeV/c^2</td>
</tr>
<tr>
<td>DOWN</td>
<td>mass 4.8 MeV/c^2</td>
<td>charge -1/3</td>
<td>spin 1/2</td>
<td>0</td>
</tr>
<tr>
<td>STRANGE</td>
<td>mass 95 MeV/c^2</td>
<td>charge -1/3</td>
<td>spin 1/2</td>
<td>0</td>
</tr>
<tr>
<td>BOTTOM</td>
<td>mass 4.18 GeV/c^2</td>
<td>charge -1/3</td>
<td>spin 1/2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LEPTONS</th>
<th>ELECTRON</th>
<th>MUON</th>
<th>TAU</th>
<th>W BOSON</th>
</tr>
</thead>
<tbody>
<tr>
<td>mass 0.511 MeV/c^2</td>
<td>mass 105.7 MeV/c^2</td>
<td>mass 1.777 GeV/c^2</td>
<td>mass 80.4 GeV/c^2</td>
<td></td>
</tr>
<tr>
<td>charge -1</td>
<td>charge -1</td>
<td>charge -1</td>
<td>charge -1</td>
<td></td>
</tr>
<tr>
<td>spin 1/2</td>
<td>spin 1/2</td>
<td>spin 1/2</td>
<td>spin 1</td>
<td></td>
</tr>
</tbody>
</table>

Table 13.1-1

Once all the quarks were discovered, the number of fundamental particles was reduced to about 17 (see table 13.1-1).

---

<table>
<thead>
<tr>
<th>Class</th>
<th>Name</th>
<th>Particle</th>
<th>Antiparticle</th>
<th>Spin</th>
<th>Rest Mass $m_e$</th>
<th>Rest Mass $Mev$</th>
<th>Half Life, Sec</th>
<th>L</th>
<th>M</th>
<th>B</th>
<th>S</th>
<th>Y</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photon</td>
<td>Photon</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>Stable</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lepton</td>
<td>e-neutrino</td>
<td>$\nu_e$</td>
<td>$\bar{\nu}_e$</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>Stable</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\mu$-neutrino</td>
<td>$\nu_\mu$</td>
<td>$\bar{\nu}_\mu$</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>Stable</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Electron</td>
<td>e</td>
<td>e$^-$</td>
<td>e$^+$</td>
<td>1/2</td>
<td>1</td>
<td>.51</td>
<td>Stable</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\mu$-meson</td>
<td>$\mu^-$</td>
<td>$\mu^+$</td>
<td>1/2</td>
<td>207</td>
<td>106</td>
<td>1.5X10$^{-6}$</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Meson</td>
<td>$\pi^-$meson</td>
<td>$\pi^0$</td>
<td>$\bar{\pi}^0$</td>
<td>0</td>
<td>264</td>
<td>135</td>
<td>7X10$^{-17}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\pi^+$</td>
<td>$\pi^+$</td>
<td>$\pi^+$</td>
<td>0</td>
<td>273</td>
<td>140</td>
<td>1.8X10$^{-8}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$K^-$meson</td>
<td>$K^-$</td>
<td>$\bar{K}^-$</td>
<td>0</td>
<td>966</td>
<td>494</td>
<td>8X10$^{-9}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>+1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$K^+$</td>
<td>$K^+$</td>
<td>$\bar{K}^+$</td>
<td>0</td>
<td>974</td>
<td>498</td>
<td>7X10$^{-11}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>+1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$\eta$-meson</td>
<td>$\eta^0$</td>
<td>$\bar{\eta}^0$</td>
<td>0</td>
<td>1037</td>
<td>548</td>
<td>~10$^{-18}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Baryon</td>
<td>Nucleon</td>
<td>$p$</td>
<td>$\bar{p}$</td>
<td>1/2</td>
<td>1836</td>
<td>938</td>
<td>Stable</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>+1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$n$</td>
<td>$\bar{n}$</td>
<td>$\bar{n}$</td>
<td>1/2</td>
<td>1839</td>
<td>940</td>
<td>7X10$^{-2}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>+1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$\Lambda$-hyperon</td>
<td>$\Lambda^0$</td>
<td>$\bar{\Lambda}^0$</td>
<td>1/2</td>
<td>2182</td>
<td>1115</td>
<td>1.7X10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\Sigma$-hyperon</td>
<td>$\Sigma^0$</td>
<td>$\bar{\Sigma}^0$</td>
<td>1/2</td>
<td>2328</td>
<td>1192</td>
<td>6X10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\Sigma^+$</td>
<td>$\Sigma^+$</td>
<td>$\Sigma^+$</td>
<td>1/2</td>
<td>2332</td>
<td>1194</td>
<td>&lt;10$^{-12}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\Sigma^-$</td>
<td>$\Sigma^-$</td>
<td>$\Sigma^-$</td>
<td>1/2</td>
<td>2341</td>
<td>1197</td>
<td>1.2X10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\Xi$-hyperon</td>
<td>$\Xi^0$</td>
<td>$\Xi^0$</td>
<td>1/2</td>
<td>2571</td>
<td>1310</td>
<td>10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-2</td>
<td>-1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$\Xi^+$</td>
<td>$\Xi^+$</td>
<td>$\Xi^+$</td>
<td>1/2</td>
<td>2583</td>
<td>1320</td>
<td>.9X10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-2</td>
<td>-1</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>$\Omega$-hyperon</td>
<td>$\Omega^-$</td>
<td>$\bar{\Omega}^+$</td>
<td>3/2</td>
<td>3290</td>
<td>1676</td>
<td>~10$^{-10}$</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
</tr>
</tbody>
</table>

Elementary Particles Stable against Decay by the Strong Interaction

Table 2.2.5.2.3-1

371
Since $E = mc^2$, mass can be thought of as energy in units where $c = 1$. In the standard model, mass has units of ‘millions of electron volts’ ([MeV]), a unit of energy:

$$1 \text{ [eV]} \approx 1.6 \times 10^{-19} \text{[Joule]}$$

Each particle has an associated field, which is represented in a Lagrangian density function in accordance with a particular interaction. Recall that each field is made from creation and annihilation operators. For example, let $\Psi_e^\dagger$ and $\Psi_e$ be creation and annihilation operators for the electron and let $A^\dagger$ be a creation operator for the photon, then the term ‘$\Psi_e^\dagger A^\dagger \Psi_e$’ represents the annihilation of an electron, the creation of a photon and the creation of an electron. The term ‘$\Psi_e^\dagger A^\dagger \Psi_e$’ would appear in the Lagrangian and represented by a Feynman diagram:

13.1.1 Types of Elementary Particles

The two major elementary particle groups are the ‘fermions’ (matter particles) and the ‘bosons’ (force carrying particles). Fermions have half-integer spin, while bosons have integer spin; most bosons are spin-1 particles, including the photon. There are ‘$W^+$’, ‘$W^-$’ and ‘$Z^0$’ bosons, the force carriers of the weak interaction, also spin-1 particles. There are the ‘gluons’, the force carrying particles of the strong interaction. Each elementary particle is associated with an antiparticle with the same mass, but opposite quantum number.

13.1.2 Quarks

Note that the standard model particle classification scheme contains no baryons, including the proton and the neutron, and no mesons. This is because baryons and mesons are made from quarks. Quarks are never found singly, but come in triplets (baryons) or doublets (mesons). Quarks come in family pairs: ‘up, down’ ($u, d$), ‘charm, strange’ ($c, s$) and ‘bottom, top’ ($b, t$). It is a mystery why there are three generations of quarks and leptons. For instance, the charm- and top-quarks are exact replicates of the up-quark except for mass. All three are spin-1/2 particles having electric charge ‘$+2/3$’. The strange- and bottom-quarks are just heavier versions the down-quark, being spin-1/2 particles with electric charge ‘$-1/3$’. All quarks have baryon number ‘$1/3$’. Table 13.1.2-1 shows the ‘quark makeup’ of various baryons and mesons [70,84,85,86,87].

The proton is made of two up-quarks and a down-quark ($uud$). Note that the hyperon ‘$\Sigma^+$’ is made from two up-quarks and a strange-quark ($uus$), just a heavy proton. The same can be said of the ‘$\Sigma^0$’ particle, a heavy neutron. It is not clear why these heavier
replicates occur in Nature, since they decay rapidly into lighter protons and neutrons, while emitting other particles in various ways.

<table>
<thead>
<tr>
<th>Mesons</th>
<th>Baryons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Particle</strong></td>
</tr>
<tr>
<td>π-meson</td>
<td>π⁰</td>
</tr>
<tr>
<td>π-meson</td>
<td>π⁺</td>
</tr>
<tr>
<td>K-meson</td>
<td>K⁺</td>
</tr>
<tr>
<td>η-meson</td>
<td>η⁰</td>
</tr>
<tr>
<td></td>
<td>Hyperon</td>
</tr>
<tr>
<td></td>
<td>Hyperon</td>
</tr>
<tr>
<td></td>
<td>Hyperon</td>
</tr>
<tr>
<td></td>
<td>Hyperon</td>
</tr>
<tr>
<td></td>
<td>Hyperon</td>
</tr>
</tbody>
</table>

**Table 13.1.2-1**

All mesons are made from quark/antiquark pairs. The π⁺-meson (ud) has electric charge `+1`. Replacing the π⁺-meson by its antiquark/quark pair i.e. (u̅d) represents a π⁻-meson with electric charge `-1`. The π⁰-meson, represented by a quantum superposition of states i.e. 1/√2(|uu⟩ – |d̅d⟩), has no electric charge. The K⁺-meson (u̅s) has electric charge `+1` and is simply a heavy version of a π⁺-meson. The K⁻-meson (s̅u), has electric charge `-1`, a heavy π⁻-meson. As with the baryons, the K⁻-mesons rapidly decay into π⁻-mesons and other particles in various ways.

**13.1.3 Leptons**

The leptons or “small particles” are found singly in Nature, while quarks are not. Leptons are considered free particles. The muon and Tau particles are just heavy versions of the electron. The electron, muon and Tau all have associated neutrinos that carry no electric charge. The muon- and Tau-neutrinos are just heavier versions of the electron-neutrino. All leptons are spin-1/2 particles (fermions). Why there are three generations of leptons remains a mystery. All neutrinos in the standard model are considered massless particles. But this does not square with experiments, which show
that neutrinos must have a small, but non-zero mass, giving rise to the realization that
the standard model is in need of significant modification.

13.2 The Sectors of the Standard Model

The standard model is divided into three sectors – the ‘QCD sector’ (strong interactions),
the ‘QED’ sector (electrodynamics) and the ‘QFD sector’ (weak interactions):

1. **The Strong Interaction:** The force that holds ordinary matter together,
   confining quarks into hadrons, creating the proton and neutron and forming the
   nucleus of the atom by binding neutrons and protons together. Most of the
   mass/energy of a proton or neutron is due to the strong force field energy; the
   individual quarks provide only about 1% of the mass/energy of a proton. This
   force (created by the exchange of gluons between quarks) holds together
   protons, neutrons, and other hadronic particles and is called the ‘color force’.
   The strength of the color force is so high that hadrons, when struck by high-
   energy particles, give rise to new hadrons instead of emitting radiation, a
   property called ‘color confinement’. In the context of binding protons and
   neutrons in atomic nuclei, the strong interaction is called the ‘nuclear force’. The
   strong interaction obeys a different distance-dependent behavior between
   nucleons, than when it acts to bind quarks within the nucleus of an atom.

   Gluons interact with quarks and other gluons by way of ‘color charge’. Color
   charge is analogous to electromagnetic charge, but comes in three types (+/-
   red, +/- green, +/- blue) rather than one, which results in a different type of
   force, with unique interaction rules, detailed in the theory of quantum
   chromodynamics (QCD) i.e. quark-gluon interactions.

2. **The Weak Interaction:** Sometimes called ‘quantum flavor dynamics’ (QFD), is
   responsible for radioactive decay and plays the essential role in nuclear fission.
   In weak interactions, fermions exchange three distinct types of force carriers
   known as the ‘W^+’, ‘W^−’ and ‘Z’ bosons. Quarks come in six ‘flavors’ – ‘up’,
   ‘down’, ‘strange’, ‘charm’, ‘top’ and ‘bottom’. The weak interaction allows quarks
   to swap one flavor for another, mediated by the force carrying bosons. For
   example, during ‘beta minus’ decay, a down-quark within a neutron is changed
   into an up-quark, converting a neutron to a proton and emitting an electron and
   an electron-antineutrino. Uniquely, the weak interaction breaks parity symmetry,
   and similarly, breaks charge conjugation symmetry. This does not happen within
   the other interactions.

3. **Quantum Electrodynamics (QED):** Responsible for a relativistic quantum
   field theory of electrodynamics. It describes how light and matter interacts and is
   the first theory where full agreement between quantum mechanics and special
   relativity is achieved. Mathematically, QED describes electrically charged
   particles interacting by means of an exchange of photons and represents the
   quantum counterpart of classical electromagnetism, giving a complete account of
   matter and light interaction.
Each sector of the standard model is described by a Lagrangian density function, made of field operators that portray the interaction. The fields are related to the field amplitude, which, in turn, is related to the probability that a given interaction takes place. Each of the sector Lagrangians is constrained by a symmetry, which dictates the kinds of particle interactions that can take place. Each of the sector Lagrangians remain invariant under the symmetry transformations that constrain the particular sector.

13.3 Quantum Electrodynamics (QED)

Quantum electrodynamics for a free particle is governed by a Lagrangian 'L_{QED}', first discovered by Dirac. Recall the Dirac equation:

\[ i \partial_t |\psi\rangle = -i\alpha_j \partial_j |\psi\rangle + m\beta |\psi\rangle, \quad j = 1,2,3 \]

which can be written in the form of a Lagrangian:

\[ L_{QED} = \frac{i}{2} \left( \Phi^+ \frac{\partial \Phi}{\partial t} + \Phi^+ \alpha \frac{\partial \Phi}{\partial x} \right) - m \Phi^+ \beta \Phi \rightarrow L_{QED} = \Phi^+ \left( i \frac{\partial}{\partial \tau} - \beta m \right) \Phi, \quad \delta = \gamma^\mu \partial_\mu, \]

where the fields have been replaced by field operators, made up of creation and annihilation operators.

In field theory, a slightly different interpretation is given to the description above. The annihilation operator '\( \Phi \)' given by

\[ \Phi = \int c^- (p) e^{ipx} dp, \]

where \( c^- \) annihilates a fermion, is broken up into two parts:

\[ \Phi_+ = \int_0^\infty c^- (p) e^{ipx} dp, \quad \Phi_- = \int_{-\infty}^0 c^- (p) e^{ipx} dp, \]

where \( \Phi_+ \) annihilates a positive energy electron and \( \Phi_- \) annihilates a negative energy electron. But because commutation relations for fermions do not distinguish between creation and annihilation operators, then

\[ \Phi_- = \int_{-\infty}^0 c^- (p) e^{ipx} dp = \int_0^\infty b^+ (p) e^{-ipx} dp, \]

where \( b^+ \) is a creation operator for a positron. In other words, annihilating an electron is the same as creating a positron.
Consider

\[ \Psi(x) = \frac{1}{(2\pi)^3 2E_p} \int \sum_{s=1}^{2} u(s, \hat{\rho}) c^-(\hat{\rho}, s) e^{i\hat{\rho}x} + v(s, \hat{\rho}) b^+(\hat{\rho}, s) e^{-i\hat{\rho}x} \, d\hat{\rho}^3, \quad E_\rho = \sqrt{\hat{\rho}^2 + m^2}, \]

where \( u(s, \hat{\rho}), v(s, \hat{\rho}) \) are related to the spin-states of the particle.

Understanding the details of the equation above is not terribly important at this point, but the fact that there are only two spin-states indicates the particle is massless. In fact, there are no massless spin-1 particles with three spin-states. All massless particles have two spin-states. The reason for this will be explained later. But, of course, the electron has a nonzero mass. How the electron acquires mass will also be covered later.

Note that \( \Psi(x) \) has two terms. The first annihilates an electron and the second creates a positron. This reflects the fact that annihilating an electron is the same as creating a positron. Note that \( c^+ |g\rangle = |e^-\rangle \) and \( b^+ |g\rangle = |e^+\rangle \). In other words, \( c^+ \) acting on the vacuum creates an electron and \( b^+ \) acting on the vacuum creates a positron. Hence, \( \Psi(x) \) destroys an electron and creates a positron, \( \Psi^+(x) \) does the opposite.

### 13.3.1 U(1) Symmetry

Moreover, the Lagrangian \( \mathcal{L}_{QED} \) remains invariant under \( U(1) \) symmetry transformations. In other words, if

\[ \Psi(x) \rightarrow \Psi'(x) = e^{i\theta} \Psi(x), \quad \Psi^+ \rightarrow \Psi'^+ (x) = e^{-i\theta} \Psi^+(x), \]

then \( \mathcal{L}'_{QED} = \mathcal{L}_{QED} \), since

\[ \Psi'(x) \Psi'^+(x) = e^{i\theta} \Psi(x) e^{-i\theta} \Psi^+(x) = \Psi(x) \Psi^+(x) \]

Because \( \Psi(x) \Psi^+(x) \) is associated with the probability that an interaction takes place, the probabilities remain unchanged with this substitution. Since \( \theta = a \text{ constant} \), \( U(1) \) is a global symmetry. If \( \theta \) depends on the space-time coordinates, in other words, if \( \theta(x) \neq a \text{ constant} \), the physics becomes localized and the symmetry is lost. The symmetry can be regained by adding terms to the Lagrangian. The procedure for accomplishing this will be discussed later in the chapter. Local symmetries are called 'gauge symmetries'.

### 13.4 Elementary Quantum Chromodynamics

Quantum mechanically, angular momentum can be measured along a single direction only, since angular momentum operators in different directions do not commute i.e. \([\hat{L}_x, \hat{L}_y] = i\hbar \hat{L}_z\). The same can be said of spin operators i.e. \([\hat{S}_x, \hat{S}_y] = i\hbar \hat{S}_z\). In fact, the algebra of spin operators is identical to the algebra of angular momentum operators.
Spin operators and angular momentum operators have the same Lie algebra (consult Book II: Chapter 11, Sec 11.11).

Both angular momentum and spin are quantized in integer multiples of \( h \). A particle can have \( 2l + 1 \) spin states, where \( l \) represents the highest spin-state. A spin-0 particle has one spin-state, a spin-1/2 particle two states: \(|1/2\rangle, |−1/2\rangle\), and a spin-1 particle has three states: \(|1\rangle, |0\rangle, |−1\rangle\) and so on.

Two half-spin fermions can make a spin-1 boson \((1/2 + 1/2 = 1)\). Hence, there should be three states: \(|1\rangle, |0\rangle, |−1\rangle\). However, there are four possible states: \(|↑↑\rangle\) (spin-state-1), \(|↓↓\rangle\) (spin-state-\((-1)\)), \(|↑↓\rangle\) (spin-state-0) and \(|↓↑\rangle\) (spin-state-0). But, spin-1 particles have only three states. What is the significance of the extra spin-state-0?

States can be in superposition i.e. \( 1/\sqrt{2} (|↑↓\rangle + |↓↑\rangle) \) or \( 1/\sqrt{2} (|↑↓\rangle - |↓↑\rangle) \). In the first case, if \(|↑↓\rangle\) is replaced by \(|↓↑\rangle\) and vice versa, \( 1/\sqrt{2} (|↑↓\rangle + |↓↑\rangle) \) remains unchanged. It is symmetric with regard to the interchange of spin-state-0’s. So, \( 1/\sqrt{2} (|↑↓\rangle + |↓↑\rangle) \) represents a spin-1 particle in spin-state-0. On the other hand, \( 1/\sqrt{2} (|↑↓\rangle - |↓↑\rangle) \) is not symmetric with regard to the interchange of spin-state-0’s. Therefore, the state \( '1/\sqrt{2} (|↑↓\rangle - |↓↑\rangle)' \) represents a spin-0 particle, which only has one state.

13.4.1 The Yukawa Model of the Strong Interaction

Isospin is an approximate \( SU(2) \) symmetry associated with the strong interactions. It was first noticed by Yukawa in the 1960’s. He defined two fields:

\[
\Psi = \begin{bmatrix} \Psi_p \\ \Psi_n \end{bmatrix}, \quad \pi = \begin{bmatrix} \pi^1 \\ \pi^2 \\ \pi^3 \end{bmatrix}, \quad \pi_k = \pi_k^+, \quad k = 1, 2, 3, \quad L^i = \frac{\gamma^i}{2}, \quad L^i_k = -i\varepsilon_{ijk},
\]

where \( \Psi_p, \Psi_n \) are fields for the proton and neutron respectively, the \( \pi^j \)’s are fields for the \( \pi^\pm, \pi^0 \) pions, \( \gamma^j \) are the Pauli matrices and \( L^i \)’s are the commutation relations.

Eventually, a Lagrangian ‘\( \mathcal{L}_Y \)’ was discovered that was invariant under certain symmetry transformations:

\[
\mathcal{L}_Y = \bar{\Psi} [i\partial - ml + ig_\pi \gamma^5 \gamma \pi] \Psi + \frac{1}{2} \left[ \partial_\mu \pi - \mu^2 \langle \pi | \pi \rangle \right] - \lambda \langle \pi | \pi \rangle,
\]

\[
\frac{\gamma \pi}{\sqrt{2}} = \frac{3}{\sqrt{2}} \begin{bmatrix} \pi^0 \\ \pi^+ \\ \pi^- \end{bmatrix}, \quad \langle \pi | \pi \rangle = \frac{3}{2} \pi^i \pi^i = \frac{1}{2} \text{Trace}(\gamma \pi)^2,
\]

\[
\text{Trace} \left( \frac{\gamma^i \gamma^k}{2} \right) = \frac{1}{2} \delta^{ik}, \quad \gamma^5 = i\gamma^0 \gamma^1 \gamma^2 \gamma^3
\]

It is not important to understand the Lagrangian ‘\( \mathcal{L}_Y \)’, other than it describes strong interactions that could occur, where, for example, a neutron could be transformed into a proton by the emission of pion i.e.
Moreover, if
\[
\Psi \rightarrow e^{i\theta_2 \gamma} \Psi, \quad \bar{\Psi} = \Psi^\dagger \gamma^0, \quad \Psi^\dagger \rightarrow e^{-i\theta_2 \gamma} \Psi^\dagger, \quad \gamma \pi \rightarrow e^{i\theta_2 \gamma \pi} e^{-i\theta_2 \gamma},
\]
\[
\langle \pi | \pi \rangle \rightarrow e^{i\theta_2 \gamma \langle \pi | \pi \rangle} e^{-i\theta_2 \gamma},
\]
then \( \mathcal{L}_Y \) remained invariant. Note that \( \mathcal{L}_Y \) has self-interacting terms i.e. terms involving \( \langle \pi | \pi \rangle \), indicating that the pions have a mass \( (\mu^2 \langle \pi | \pi \rangle) \). The strength of the interaction is \( g_\pi \approx 14 \). Because \( g_\pi \) was so large, it could not be associated with a probability. This suggested that there was something else at work. The ‘something else’ eventually led to the quark model.

### 13.4.1.1 The Conservation of Baryon Number

But \( \mathcal{L}_Y \) had an additional symmetry ‘\( SU(2) \times U(1)_S \)’, called ‘conservation of baryon number’. So far as is known, baryon number is conserved in Nature. Under baryon number conservation
\[
\bar{\Psi} \rightarrow e^{i\beta \theta} \bar{\Psi}, \quad \bar{\Psi}^\dagger \rightarrow e^{-i\beta \theta} \bar{\Psi}^\dagger, \quad \pi^i \rightarrow \pi^i
\]
For any field \( \Phi' = e^{i\beta \theta} \Phi \), \( \beta \) is conserved i.e. baryon number is a global symmetry. Under the following types of interactions,
\[
p \rightarrow p + \pi^0, \quad p \rightarrow n + \pi^+, \quad p + \bar{p} \rightarrow \pi^0,
\]
not only is electric charge conserved, but baryon number is conserved.

### 13.4.2 The Connection between Isospin and \( SU(3) \) Symmetry

There were particles other than the proton, neutron and the pions. For example, there were the ‘\( K \)-mesons’ or ‘kaons’, which did not seem to fit the Yukawa model. For instance, \( K^+ \rightarrow \pi^+ + \pi^0 \), \( \pi^+ + \pi^- \rightarrow K^+ + K^- \). These particles were labelled ‘strange’ particles. Gell-Mann and Yuval Ne’eman (14 May 1925 – 26 April 2006) noticed that in order to describe these particles adequately, another quantum number was required. The new quantum number became known as ‘strangeness’ \( (s) \).

It was not clear why some heavy particles decayed into lighter ones with the emission of a gamma ray, while other, seemingly equally permissible, decays did not occur. For instance, the ‘\( \Sigma^0 \)’ decayed into a \( \Lambda^0 \) and a gamma ray, but the ‘\( \Sigma^+ \)’ did not decay into a proton and a gamma ray. Moreover, these heavier ‘strange particles’ were never found singly, but in combinations of two or more at a time. To deal with these situations, a
new quantum number called ‘strangeness’ \((s)\) was introduced. In all electromagnetic and strong interactions, \(s\) is conserved. This explained why certain baryon decays occurred while others did not. Only in weak interactions did \(s\) change, but only by \(+1\) or \(-1\) in any given decay. It was also convenient to introduce a quantity called ‘hypercharge’ \((Y)\) \((Y = S + B)\); the quantity ‘\(Y\)’ was conserved in all strong interactions [45].

<table>
<thead>
<tr>
<th>Strange</th>
<th>Baryons</th>
<th>Mesons</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s = 0)</td>
<td>(n)</td>
<td>(\pi^-)</td>
</tr>
<tr>
<td>(s = -1)</td>
<td>(\Sigma^-)</td>
<td>(\Sigma^0)</td>
</tr>
<tr>
<td>(s = 1)</td>
<td></td>
<td>(K^-)</td>
</tr>
<tr>
<td>(s = -2)</td>
<td>(\Xi^-)</td>
<td>(\Xi^0)</td>
</tr>
</tbody>
</table>

Table 13.4.2-1

13.4.2.1 The Isospin Quantum Number

Looking at Table 2.2.5.2.3-1, there are a number of particle families whose members have the same mass and interaction properties, but a different charge. These families can be thought of as a single entity, but in a different charge state. One way of categorizing these families is by assigning the family a number, call it ‘\(I\)’, where the number of charge states is given by the formula ‘\(2I + 1\)’. For the nucleon, \(I = 1/2\), so that \(2I + 1 = 2\). The two states are the ‘proton’ and the ‘neutron’. The \(\pi\)-meson family has \(I = 1\), so \(2I + 1 = 3\). The three charge states are the ‘\(\pi^+\)’, ‘\(\pi^0\)’, and ‘\(\pi^-\)’. The number ‘\(I\)’ was named ‘isotopic spin’, but has nothing to do with the spin-state of a particle. In analogy to quantum angular momentum, a new quantum number ‘\(I_3\)’ was designated, restricted to the values \(I, I - 1, \ldots, 0, \ldots, -(I - 1), -I\); the quantity ‘\(I_3\)’ takes on half-integer values if \(I\) is half-integer and integer values if \(I\) is an integer. For the nucleon, \(I = 1/2\), and hence, \(I_3 = 1/2\) or \(-1/2\); the former represents the proton and the latter the neutron. In the case of \(\pi\)-mesons, \(I = 1\), so \(I_3 = 1\) corresponds to \(\pi^+\), \(I_3 = 0\) to \(\pi^0\) and \(I_3 = -1\) to \(\pi^-\). All baryons have baryon number ‘\(1\)’ and all mesons have baryon number ‘\(0\)’.

The charge of a meson or baryon is related to its hypercharge ‘\(Y\)’ and \(I_3\) by

\[
q = e \left( I_3 + \frac{Y}{2} \right) = e \left( I_3 + \frac{S + B}{2} \right)
\]

In the case of the nucleons, the proton has \(I_3 = 1/2\), \(B = 1\) and \(S = 0\), so \(q = e\), the charge on a proton. For the neutron, \(I_3 = -1/2\), \(B = 1\) and \(S = 0\), so \(q = 0\). The other meson and baryon charge states can be calculated in a similar fashion. Charge and baryon number ‘\(B\)’ are conserved in all interactions; \(I_3\) is conserved whenever \(s\) is, namely, in strong and electromagnetic, but not in weak interactions [45].
13.4.2.2 The Introduction of the Quark and the Eight-Fold Way

What Gell-mann and Ne’eman noticed (see table 13.4.2-1) was that particles that had the same charge could be assigned a different ‘strangeness number’ that would put each fermion in a different state. Moreover, the strangeness classifications came in threes. Baryons had strange numbers ‘$s = 0, -1, -2$’ and the mesons had strange numbers ‘$s = 0, +1, -1$’.

Eventually, Gell-Mann and George Zweig introduced three new particles: the ‘up-’, ‘down-’ and ‘strange-quarks’ ($u, d, s$). The up- and down-quarks ($u, d$) came in $SU(2)$ doublets. Any Lagrangian containing fields that described these particles would remain invariant under $SU(2)$ rotations. The $s$-quark was a singlet. However, quarks do not come in singlets, but, at the time, this is the way the quark model was constructed. Eventually, a Lagrangian $\mathcal{L}_{\text{quark}}$ for non-interacting quarks was developed:

$$\mathcal{L}_{\text{quark}} = \sum_{a=u,d,s} \overline{q}_a (i\gamma^\mu - m_a) q_a , \quad q = \begin{bmatrix} u \\ d \\ s \end{bmatrix} \rightarrow e^{\frac{\beta \lambda}{2} q}$$

The Lagrangian above is invariant under $SU(2)$ and $SU(3)$ symmetry transformations if the differences in the $u$, $d$- and $s$-quark masses are ignored. Recall that hypercharge ‘$y$’ is given by $y = \beta + s$, where $\beta$ is the baryon number, $I_3 = \lambda^3/2$, where $\lambda^3$ represents one of the three generators of $SU(2)$; the quantity ‘$y = I_3 + \lambda^8/\sqrt{3}$’, where $\lambda^8$ represents one of the eight generators of $SU(3)$. The charge

$$q = \frac{\lambda^3}{2} + \frac{\lambda^8}{2\sqrt{3}} = \begin{bmatrix} \frac{2}{3} & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}, \quad y = \frac{\lambda^8}{\sqrt{3}} = \frac{1}{3} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \rightarrow \frac{\lambda^3}{2} = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $q_u = +2/3$, $q_d = q_s = -1/3$ are the charges on the up-, down- and strange-quark respectfully.

Gell-Mann named his system the ‘Eight-Fold-Way’. At first, it earned a cool reception from other particle physicists because, even though the three quarks were in a bound state, the strange-quark was a singlet. No one had ever heard of or seen a single particle with a fractional electric charge. For some time, the ‘quark’ was considered simply a mathematical convenience, rather than a real particle. But additional insights and ingeniously designed experiments would change the perception.

13.4.2.3 The Relationship between Spin and Isospin

In quantum mechanics, quarks have a characteristic analogous to spin, called ‘isospin’ or ‘isotopic spin’. Leptons, for example, have two spin states ($\uparrow, \downarrow$). Isospin is characterized by two states ($|u\rangle, |d\rangle$), representing the up- and the down-quark. There is a certain symmetry here. A neutron can be created from the proton by replacing an
up-quark with a down-quark and vice versa. The math associated with $|u\rangle, |d\rangle$ quarks is isomorphic to the math associated with the spin-states of leptons. The symmetry is only approximate. For one thing, the mass of an up-quark is not the same as the mass of a down-quark. Moreover, if the nucleus of an atom gets too large, then the electrostatic force between the protons in the nucleus breaks the symmetry. But if the small difference between the masses of an up- and down-quark are ignored, and if the nucleus of the atom is not too large, then an approximate symmetry exists and is called 'isospin'.

Consider a spin-3/2 particle with isospin-3/2. Since it is a spin-3/2 particle, it has four spin-states and four isospin-states:

$$
\begin{pmatrix}
|isospin\rangle & |spin\rangle \\
uuu & \uparrow\uparrow\uparrow \\
\bar{d}d & \downarrow\downarrow\downarrow \\
uud & \uparrow\uparrow\downarrow \\
\bar{d}u & \downarrow\uparrow\uparrow
\end{pmatrix} \rightarrow \Delta_{3/2}
$$

This particle, called the 'delta-3/2 particle', actually exists in Nature. It is a short lived particle that decays rapidly into other particles. For instance, the $uuu$-state decays into a proton ($uud$) and a $\pi^+$ ($\bar{d}u$) in the following way:

$$uuu \rightarrow uud + \bar{d}u = p + \pi^+$$

![Figure 13.2.1-1](image)

Figure 13.2.1-1

Note that, in this decay, charge is conserved. The state 'uuu' has charge $2/3 + 2/3 + 2/3 = 2$, a proton having charge '+1' and a $\pi^+$ charge '+1' for a total charge of '+2'. Care must be taken not to confuse the 'spin-state' of a particle with the 'spin' of a particle. One of the spin-states of the $\Delta_{3/2}$ is coded 'uud', the same quark makeup as a proton. But in this case, $uud$ is a $\Delta_{3/2}$ in spin-state-1/2. On the other hand, a proton is a spin-1/2 particle. The proton and the $\Delta_{3/2}$ in spin-state-1/2 also have different masses.

The $\Delta_{3/2}$ is a fermion and so are the individual quarks that comprise it. But if the $\Delta_{3/2}$ is in the state 'uuu', the three fermions are in the same quantum state, which is forbidden by the Pauli exclusion principle. This gave a clue that quarks must have another property in addition to spin and isospin, first recognized by the Japanese/American
physicist Yoichiro Nambu (18 January 1921 – 5 July 2015). He was awarded the Nobel Prize in Physics in 2008 for this discovery. The additional property eventually became known as ‘color’. The property itself has nothing to do with color, but is simply a means of tagging each quark with a different label. There are ‘red-quarks’, ‘blue-quarks’ and ‘green-quarks’ \((u_R, u_B, u_G)\). A red up-quark is in a different quantum state than a blue up-quark.

13.4.3 Giving the Quark a Mass

The Lagrangian \(\mathcal{L}_{\text{quark}}\) can be written in a form where the symmetries are broken:

\[
\mathcal{L}_{\text{quark}} = \bar{q}(i\partial - R)q - m_3 \bar{q}\lambda^3 q - m_8 \bar{q}\lambda^8 q,
\]

where, evidently,

\[
m_u = R - m_3 + \frac{m_8}{\sqrt{3}}, \quad m_d = R + m_3 + \frac{m_8}{\sqrt{3}}, \quad m_s = R + \frac{2m_8}{\sqrt{3}}
\]

The masses \(m_u, m_d, m_s\) are the masses of the up-, down- and strange-quark respectfully, \(R\) is a mass to be determined and the term \(m_3 \bar{q}\lambda^3 q\) represents the breaking of the \(SU(2)\) symmetry and \(m_8 \bar{q}\lambda^8 q\) represents the breaking or the \(SU(3)\) symmetry. This suggested that the symmetry is broken by giving a particle mass.

13.4.4 Gluons

Since the nucleus of an atom is made of protons and neutrons, the repulsive electrostatic force between the protons should cause the nucleus to fly apart. So what holds the nucleus together? The force carrying particles that hold the nucleus together are called ‘gluons’. The force between gluons and quarks is sometimes referred to as the ‘strong force’. It is a powerful, but short ranged force that acts over a distance of about \(10^{-15}\) [meters], which is roughly the diameter of an atomic nucleus. Once the nucleus reaches a reasonably high number protons and neutrons, the electrostatic force associated with the protons begins to overwhelm the short ranged gluon force and the nucleus will begin to decay.

The primary interaction in electrodynamics, where an electron is annihilated, an electron created and a photon emitted as the force carrying particle, can be described in terms of a Feynman diagram:
In this case, the photon is represented by an electron/positron pair \((e, \bar{e})\). Technically, the electron/positron annihilate, creating a photon. Similarly, a quark can emit a gluon. And a gluon can be thought of as made up of a quark and an anti-quark.

### 13.4.4.1 Color States of Gluons

Gluons come in color states: ‘red-state’ \((R)\), ‘blue-state’ \((B)\) and ‘green-state’ \((G)\). All gluons come in color/anti-color quark doublets:

\[
\begin{bmatrix}
u_R \bar{u}_R & u_R \bar{u}_G & u_R \bar{u}_B \\
u_G \bar{u}_R & u_G \bar{u}_G & u_G \bar{u}_B \\
u_B \bar{u}_R & u_B \bar{u}_G & u_B \bar{u}_B
\end{bmatrix}
\]

The matrix above suggests there are nine gluons, but there are only eight. The reason is that the Lagrangian for the QCD sector of the standard model must be invariant under \(SU(3)\) symmetry transformations. The number of group transformations under \(SU(N)\) symmetry is \(N^2 - 1\), where \(N\) is the number of elements in the set associated with the group. In this case, \(N = 3\) \((R, B, G)\). Hence, \(3^2 - 1 = 8\) symmetry elements.

### 13.4.4.2 The Interactions of Gluons

But unlike photons, which do not interact, gluons interact to produce other configurations of gluons.

![Diagram](image)

Figure 13.4.4.2-1

Figure 13.4.4.2-1 shows two gluons ‘\(u_R \bar{u}_B\), \(u_G \bar{u}_G\)’ coming together and producing \(u_R \bar{u}_G, u_G \bar{u}_B\). The fact that gluons create gluons is something new. Photons, for example, do not interact to make other photons. Not only do gluons create gluons, but gluons exert forces on each other as shown in the diagram below:
Figure 13.4.4.2-2 shows a $u_R\bar{u}_B$ and a $u_B\bar{u}_B$ exchanging a $u_G\bar{u}_B$ gluon. The ability to create gluons from other gluons and for gluons to exert forces on each other makes describing gluon interaction more complicated, but more interesting.

13.4.4.3 Gluon Mass

Quarks are never found singly in Nature (quark confinement), but come in either triplets or doublets. Quarks are bound together in a kind of “soup of gluons”, which makes measuring the masses of quarks particularly ambiguous. Two terms are used in referring to “quark mass”: 1) ‘current quark mass’, which refers to the mass of a quark and 2) ‘constituent quark mass’, which refers to the current quark mass plus the mass of the surrounding gluon field [239]. These masses typically have very different values. The amount of hadron mass comes primarily from the gluons that bind the constituent quarks together, rather than from the quarks themselves. Gluons possess energy, specifically, quantum chromodynamics binding energy, which greatly contributes to the overall mass of the hadron. For example, a proton has a mass of approximately 938 [MeV]. The mass of its three valence quarks only contributes about 9 [MeV]; much of the remainder of the proton’s mass can be attributed to the field energy associated with the gluons [240,241].

13.5 Rotation Operators

Given an axis $|\hat{n}\rangle$ and an angle $\theta$, rotations of a vector around $|\hat{n}\rangle$ constitute a group structure, in other words, a symmetry. A rotation matrix $\hat{R}$ acts as a linear operator performing a rotation in Euclidean space. For example, $\hat{R}$ might rotate points in the $x, y$-Cartesian plane counter-clockwise through an angle $\theta$ about the origin of the coordinate system, where each point is represented by a column vector $|v\rangle$ containing the coordinates of the point. A rotated vector is obtained by using matrix multiplication i.e. $|v\rangle \rightarrow \hat{R}|v\rangle = |v'\rangle$. For example, let

$$|v\rangle = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \hat{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \rightarrow \hat{R}|v\rangle = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \cos \theta - y \sin \theta \\ x \sin \theta + y \cos \theta \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} = |v'\rangle,$$
where $\hat{R}$ rotates the vector $|\nu\rangle$ around the origin. It is usually assumed that the rotations are counter-clockwise in a right-handed coordinate system. If the axes are rotated instead of the vectors, then the inverse of the matrix is used, which coincides precisely with its transpose:

$$\hat{R}^{-1} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

Rotation matrices are square orthogonal matrices with real entries and determinant ‘1’; a square matrix $\hat{R}$ is a rotation matrix if $\hat{R}^T = \hat{R}^{-1}$ and $\det \hat{R} = 1$. In some of the literature, the term ‘rotation’ includes ‘improper rotations’, orthogonal matrices with determinant ‘−1’. Improper rotations are reflections, which invert orientation.

### 13.5.1 The Rotation Group

The set of all orthogonal matrices of size ‘n’ with determinant ‘+1’ forms a group called the ‘special orthogonal group’ ($SO(n)$), a special case being the rotation group ‘$SO(3)$’. The set of all orthogonal matrices of size ‘n’ with determinant ‘+1’ or ‘−1’ forms the (general) orthogonal group ‘$O(n)$’. Rotations form a symmetry group because any rotation followed by another rotation is also a rotation. Rotations are closed under matrix multiplication and $\hat{R}_1(\hat{R}_2 \hat{R}_3) = (\hat{R}_1 \hat{R}_2) \hat{R}_3$, $\hat{R}_i = I$ and $\hat{R}_i^T = \hat{R}_i^{-1}$, all belonging to the rotation group. Hence, all the group properties are satisfied. Interestingly, the rotation group in two dimensions is an Abelian group i.e. $\hat{R}_i \hat{R}_j = \hat{R}_j \hat{R}_i$, but, in three dimension, it is non-Abelian.

Example: A basic rotation is defined around a given coordinate axis. The following three basic rotation matrices rotate vectors by an angle ‘$\theta$’ about the $x, y$ or $z$-axis, using the right-hand rule (the same matrices can also represent a clockwise rotation of the axes):

$$\hat{R}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}, \hat{R}_y(\theta) = \begin{bmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{bmatrix}, \hat{R}_z(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For column vectors, each of the above rotations appears counterclockwise when the axis about which the rotation occurs is toward the observer; the coordinate system is said to be ‘right-handed’ and the angle ‘$\theta$’ positive. The rotation $\hat{R}_z$, for instance, would rotate toward the $y$-axis a vector aligned with the $x$-axis:

$$\hat{R}_z\left(\frac{\pi}{2}\right) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos \frac{\pi}{2} & -\sin \frac{\pi}{2} & 0 \\ \sin \frac{\pi}{2} & \cos \frac{\pi}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Given a $3 \times 3$ rotation matrix ‘$\hat{R}$’, a vector $|\hat{n}\rangle$ parallel to the rotation axis must satisfy $\hat{R}|\hat{n}\rangle = |\hat{n}\rangle$, since the rotation of $|\hat{n}\rangle$ around the rotation axis must result in $|\hat{n}\rangle$. The equation above may be solved for $|\hat{n}\rangle$ i.e. $(\hat{R} - I)|\hat{n}\rangle = |0\rangle$. Hence, $|\hat{n}\rangle$ is an eigenvector of $\hat{R}$ corresponding to eigenvalue ‘1’. Every rotation matrix must have this eigenvalue,
the other two eigenvalues being complex conjugates of each other. For detailed information on rotations, consult Book II: Chapter 11, Sec 11.9.

13.5.2 The Invariance of the Inner Product of Rotations

Suppose there is a $3 \times 3$ rotation matrix $'\hat{R}_{ij}(\hat{n}, \theta)'$ such that $\hat{R}_{ij}(\hat{n}, \theta)|v_j \rangle = |v'_j \rangle$. In other words, $\hat{R}_{ij}$ rotates $|v_j \rangle$ to produce $|v'_j \rangle$. Since this is a simple rotation, the length of the vector remains invariant i.e. $\sum_{j} v_j v_j = \sum_{l} v'_l v'_l$. Hence,

$$\sum_{l} v'_l v'_l = \sum_{i,j,k} \hat{R}_{ij} \hat{R}_{ik} v_j v_k = v_j v_k$$

The only way the above relationship can be true is if $\hat{R}_{ij} \hat{R}_{ik} = \delta_{jk}$. Since $\hat{R}$ is a rotation matrix, $\hat{R}^\dagger \hat{R} = I$. Hence, $\hat{R}_{ij}$ is a unitary matrix.

13.5.3 Rotations Applied to the Spin of a Particle

Recall that a spin-1 particle can have three spin-states '$|1\rangle, |0\rangle, |−1\rangle$'. The amplitude associated with the three spin-states is given by a 3-vector. Because the spin-states are rotated by the $3 \times 3$ rotation matrices just discussed, spin-1 particles are often referred to as 'vector bosons'. Spin-0 particles are represented rotationally by $\hat{R} = I$. There is no rotation of a spin-0 particle.

13.5.4 The Group $SU(2)$

For a spin-1/2 particle, the two spin-states are: '$|u\rangle = [1\,0\,0], |d\rangle = [0\,0\,1]$'. If their amplitude is given by the matrix $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, then $\alpha \alpha^\dagger + \beta \beta^\dagger = 1$. Given a unitary matrix $'\hat{U}'$ such that

$$\hat{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha' \\ \beta' \end{bmatrix},$$

then $1 = \alpha \alpha^\dagger + \beta \beta^\dagger = \alpha'(\alpha')^\dagger + \beta'(\beta')^\dagger$. In other words, the probabilities for the states 'spin-up' and 'spin-down' must be preserved. The only way this can happen is if $\hat{U}^\dagger \hat{U} = I$. Since $\hat{U}$ is a $2 \times 2$ unitary matrix, in general, it has complex entries, and therefore, has 8 components, two for each entry (real and imaginary). The condition $'\hat{U}^\dagger \hat{U} = I'$ reduces the number of components to 4. The further restriction to $'\det \hat{U} = \det \hat{U}^\dagger = 1'$ reduces the number of components to 3. The group formed by these restrictions is called 'SU(2)'. To see this, the group $'U(2)'$ is the set of all unitary $2 \times 2$ matrices. A complex $n \times n$ matrix has $2n^2$ real parameters. The unitary condition removes $n^2$ of these. To see this, let

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a^\dagger & c^\dagger \\ b^\dagger & d^\dagger \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

386
Hence,
\[ aa^\dagger + bb^\dagger = 1, \quad ac^\dagger + bd^\dagger = 0, \quad ca^\dagger + db^\dagger = 0, \quad cc^\dagger + dd^\dagger = 1 \]

Moreover,
\[
\begin{bmatrix} a^\dagger & c^\dagger \\ b^\dagger & d^\dagger \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow aa^\dagger + cc^\dagger = 1, \quad a^\dagger b + c^\dagger d = 0, \quad b^\dagger a + d^\dagger c = 0, \quad b^\dagger b + d^\dagger d = 1
\]

From the equations above and the conditions for unitarity, then
\[
c = \frac{b^\dagger b}{c^\dagger}, \quad d = \frac{a^\dagger a}{d^\dagger} \rightarrow \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} a \\ b^\dagger b a^\dagger \\ c^\dagger \\ d^\dagger \end{bmatrix} = \begin{bmatrix} a \\ -b^\dagger a \\ a^\dagger a \\ b^\dagger \end{bmatrix} \rightarrow \begin{bmatrix} a \\ -b^\dagger \\ a^\dagger \\ b^\dagger \end{bmatrix}
\]

### 13.5.5 The Pauli Matrices

The group \( 'U(2)' \) has four generators and four parameters i.e. \( E(\theta_0, \theta_1, \theta_2, \theta_3) = e^{-i\theta_j F_j}, \ j = 0,1,2,3 \). The generators \( F_j = \hat{\theta}_j/2 \) are:

\[
F_0 = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad F_1 = \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad F_2 = \frac{1}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad F_3 = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}
\]

The group \( 'SU(2)' \) is the set of all unitary \( 2 \times 2 \) matrices with unit determinant. The unit determinant constraint removes one more parameter; the group \( 'SU(2)' \) then has three generators i.e. \( E(\theta_1, \theta_2, \theta_3) = e^{-i\theta_j F_j}, \ j = 1,2,3 \), and three parameters:

\[
F_1 = \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad F_2 = \frac{1}{2} \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad F_3 = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}
\]

The matrices above are the ‘Pauli matrices’. To see this, if the exponential function is expanded to first order, then \( \hat{U} = 1 + i \varepsilon M, \ \varepsilon \ll 1 \), where \( M \) is a matrix. Then

\[
\hat{U} \hat{U}^\dagger = (1 + i \varepsilon M)(1 - i \varepsilon M^\dagger) = 1 - i \varepsilon M^\dagger + i \varepsilon M = 1 \rightarrow \varepsilon (M - M^\dagger) = 0 \rightarrow M = M^\dagger,
\]

where the \( \varepsilon^2 \) term has been omitted as too small to count.

Moreover, let
\[
I + M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \quad m_{ij} \ll 1,
\]

where \( M \) represents a very small deviation from the unit matrix, then
\[
det(I + M) = det\left[ \begin{bmatrix} 1 + m_{11} & m_{12} \\ m_{21} & 1 + m_{22} \end{bmatrix} \right] = (1 + m_{11})(1 + m_{22}) - m_{12}m_{21} = 1
\]

\[
\rightarrow det\left[ \begin{bmatrix} 1 + m_{11} & m_{12} \\ m_{21} & 1 + m_{22} \end{bmatrix} \right] = (1 + m_{22} + m_{11} + m_{11}m_{22}) - m_{12}m_{21} = 1
\]

\[
\rightarrow m_{22} + m_{11} = 0,
\]
since the quadratic terms ‘$m_{11}m_{22}$’ and ‘$m_{12}m_{21}$’ are too small to count. Hence, $M$ must be traceless. The Pauli matrices meet this requirement. Therefore, the generators of $SU(2)$ are the set of three linearly independent, traceless $2 \times 2$ Hermitian matrices. In fact, there is a one-to-one correspondence between $SU(2)$ and $SO(3)$, the group of rotations in three dimensions, such that $SU(2)$ represents the set of all possible rotations of two dimensional complex vectors (spinors) in a real three dimensional space. However, this is not quite true. There are actually two $SU(2)$ rotations for every $SO(3)$ rotation. For $SU(2)$, the rotation ‘$\theta + 2\pi$’ is not the same as the rotation ‘$\theta$’. There is a sign difference. A range for $\theta$ of the group ‘$SU(2)$’ is $0 \leq \theta \leq 4\pi$. The set with range ‘$0 \leq \theta \leq 2\pi$’ corresponds to the complete set of $SO(3)$ rotations.

### 13.5.6 The Group $SU(3)$

The group ‘$U(3)$’ is the set of all $3 \times 3$ unitary matrices. A complex $n \times n$ matrix has $2n^2$ real parameters. The unitary condition removes $n^2$ of these. Hence, $U(3)$ has $18 - 9 = 9$ parameters. The group ‘$SU(3)$’ is the set of all unitary matrices with unit determinants. The additional “unit determinant” requirement reduces the number of elements by 1. Hence, $SU(3)$ has $18 - 9 - 1 = 8$ elements:

$$
\lambda_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \lambda_2 = \begin{bmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \lambda_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \lambda_4 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}
$$

$$
\lambda_5 = \begin{bmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{bmatrix}, \quad \lambda_6 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \lambda_7 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{bmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}
$$

The matrices above are traceless with determinant ‘1’. It is not a coincidence that the group ‘$SU(3)$’ has eight elements and that there are eight gluons in the standard model.

#### 13.5.6.1 $SU(3)$ Representations

Suppose that $\tilde{U}$ is a $3 \times 3$ unitary operator that acts on a lone quark i.e. $\tilde{U}|q\rangle = |q'\rangle$. Quarks can be in three states ‘$R, B, G$’. Hence,

$$
\tilde{U} \begin{bmatrix} \hat{q}_R \\ \hat{q}_B \\ \hat{q}_G \end{bmatrix} \equiv '3',
$$

where the elements of the column vector represent the amplitudes such that $\hat{q}_R \hat{q}_R^\dagger + \hat{q}_B \hat{q}_B^\dagger + \hat{q}_G \hat{q}_G^\dagger = 1$, where $\hat{q}_R \hat{q}_R^\dagger$ represents the probability that the quark is in the red-state and so on. The expression above is called a ‘3’ representation. In a similar fashion, $\tilde{U}^*|q^\dagger\rangle = |q^{'\dagger}\rangle$, where the $q^{'\dagger}$'s represent antiquarks and ‘$\tilde{U}^*$’ represents a $3 \times 3$ unitary operator such that the complex components have been replaced by their complex conjugates. The operator ‘$\tilde{U}^*$’ acts on the column vector '|$q^\dagger$\rangle' to produce |$q^{'\dagger}$\rangle. This is called a ‘$\bar{3}$’ representation.

388
13.5.6.1.1 The Direct Products of Representations

Consider the direct products of these representations. For example, $3 \otimes 3$ represents a 2-quark state. It has 9 elements. It happens that $3 \otimes 3 = \bar{3}$ or 6. In other words, $3 \otimes 3$ results in an antiquark or something called a ‘$6$’ $SU(3)$ representation. It turns out that $6$’s do not exist in Nature for reasons that will be explained subsequently. Moreover, $3 \otimes \bar{3} = 1$ or 8. This is a quark-antiquark state. The ‘1’ is represented by $q_R \bar{q}_R + q_B \bar{q}_B + q_G \bar{q}_G$ and is called a ‘singlet’. It does not exist in Nature. The ‘8’ is associated with the elements of the symmetry group ‘$SU(3)$’. Such quark-antiquark pairs are called ‘gluons’.

What about $3 \otimes 3 \otimes 3$? This represents a 3-quark state. So,

$$3 \otimes 3 \otimes 3 = \downarrow$$
$$3 \otimes 3 = 1 \text{ or } 8$$
$$6 \otimes 3 = 10 \text{ or } 8$$

In other words, $3 \otimes 3 \otimes 3$ results in a ‘1’ or an ‘8’ or a ‘10’ or an ‘8’. The only 3-quark state that exists in Nature is the singlet state $[q_R q_B q_G]$.

13.6 The Postulates of Quantum Chromodynamics

This exercise illustrates the postulates of quantum chromodynamics:

1. $q = 3$
2. $\bar{q} = \bar{3}$
3. gluons = 8
4. Free particles = 1

**Postulate 1**: Quarks come in three colors ($R, B, G$)
**Postulate 2**: Antiquarks come in three colors ($\bar{R}, \bar{B}, \bar{G}$)
**Postulate 3**: There are eight gluons, which transform in accordance with $SU(3)$
**Postulate 4**: Free particles are color singlets made from three quarks ($q_R, q_B, q_G$)

The particles made from three quark singlets ($3 \otimes 3 \otimes 3$) are the baryons. Mesons are made from a quark and an antiquark singlet ($3 \otimes \bar{3}$), which together make up the hadrons, including the proton and neutron. Gluons act like quark-antiquark pairs, but are not singlets and are subject to quark confinement. It is possible to create particles from the gluons. For example, $8 \otimes 8 = 1$ or $63$. Two gluons produce a singlet called the ‘glueball’. The fact that particles come in quark color singlets is consistent with charge conservation. Three quarks cannot have a fractional electric charge. Neither can quark-antiquark pairs.

The mathematics of the direct products of these representations is quite complicated. To take a simple example, suppose two spin-$1/2$ particles are combined. This is signified by ‘$2 \otimes 2$’, since there are two spin-states (up, down). The spin of the two particles can be represented by two column vectors:
If $\psi^+_u \Phi^+_d - \psi^+_d \Phi^+_u$ is acted on by an $SU(2)$ operator, it simply reproduces itself. It is, therefore, a singlet and represents a spin-0 particle. If an $SU(2)$ operator acts on the states

$$
\begin{bmatrix}
\psi_u^+ \\
\psi_d^+
\end{bmatrix}, \quad
\begin{bmatrix}
\Phi_u^+ \\
\Phi_d^+
\end{bmatrix}
\rightarrow
\begin{bmatrix}
\psi_u^+ \Phi_u^+ \\
\psi_d^+ \Phi_d^+ \\
\psi_u^+ \Phi_d^+ \\
\psi_d^+ \Phi_u^+
\end{bmatrix}
$$

they transform into each other, which has three spin-states, and hence, represents a spin-1 particle.

13.6.1 Gauge Theories

Gluons play the same role in $QCD$ as photons play in $QED$. Gluons have charge (color charge) while photons carry no charge. The source of the photon is the electromagnetic field and the conserved quantity is electric charge, in other words, the electric field has $U(1)$ symmetry. Color is the source of the gluon field, except that there are eight gluon fields instead of one. If two electric charges have opposite charge, the field energy around them slowly disappears as the distance between them diminishes. If they have the same charge, the energy increases as the distance between them decreases, creating a large magnetic field around the charged particles.

If the electric charge was much bigger, most particles in Nature would be electrically neutral. Charged particles would simply bind together, making it difficult to pull them apart. In the case of electromagnetism, the coupling constant is relatively small, $\approx 1/137$. Pulling two electric charges apart does not require much energy. Nambu hypothesized that if the coupling constant for color charge was large enough, then it simply required too much energy to pull the quarks apart. The color singlet $[u_R u_B u_G]$ was the analog of a neutrally charged particle and pulling these singlets apart simply required too much energy, giving a reason why quarks are never found in isolation.

All continuously conserved quantities are associated with a symmetry, the essence of Noether’s theorem. Quantum electrodynamics is predicated on the conservation of electric charge i.e. on the $U(1)$ symmetry group. Quantum chromodynamics requires the conservation of color charge, epitomized by the $SU(3)$ symmetry group. Angular momentum and spin of a particle are conserved quantities associated with the symmetry group ‘$SU(2)$’. Any theory based on a local symmetry group is called a ‘gauge theory’. The standard model is a collection of gauge theories, often stipulated ‘$U(1) \times SU(2) \times SU(3)$’.
13.6.2 Gluon Fields

In quantum chromodynamics, the analog of the photon is the gluon. But, unlike the photon, gluons carry color charge. Electromagnetic waves do not exert a force on each other and photons do not create other photons. However, gluons exert forces on each other and gluons can create other gluons. Gluon interaction is non-linear, making the description of gluon interactions more complicated.

The fields lines associated with electric charge radiate outward from the source uniformly. There is no sense in which a field line exerts a force on another field line. But the lines of flux associated with gluons attract forming tube-like structures (see Fig. 13.6.2-1):

![Figure 13.6.2-1](image)

Figure 13.6.2-1

An electromagnetic field weakens as it moves out from its source, the electric charge. But a gluon force field increases roughly linearly with distance. A quark-antiquark pair is a ‘singlet’ as is a three quark combination. If it was possible to have non-singlets in Nature, then the flux lines would not end on another quark or antiquark and would be infinitely long. Evidently, these non-singlets do not exist in Nature. However, the gluon flux-tubes can break in half, creating two quark-antiquark pairs, rather than a single quark.

There is a relationship between lines of flux and charges. All flux lines must end on a charge. Given an electric charge, the flux lines spread out to infinity. Flux lines vanishing instantaneously with the removal of the electric charge would violate the postulate that nothing can travel faster than light. Sending signals instantaneously over large spatial distances does not happen in Nature. Hence, electric charge must be conserved.

The fields associated with quantum chromodynamics are the gluon fields, designated ‘\(A_{ij}\)’, where the \(i^{th}\) index is associated with an antiquark and the \(j^{th}\) index is associated with a quark.

![Figure 13.6.2-2](image)

Figure 13.6.2-2
For example, an \( h \)-quark can transform into a \( k \)-quark by emitting a gluon represented by the operator ‘\( A_{kh} \)’, where \( \text{Tr} (A_{kh}) = 0 \) (see fig. 13.6.2-2). Since gluons come in quark-antiquark pairs, a gluon can emit other gluons (see fig. 13.6.2-3). Evidently, mathematically, \( A_{ij} = A_{ik}A_{kj} \), where the \( k' \)s are summed over.

\[
\begin{align*}
\text{Figure 13.6.2-3}
\end{align*}
\]

### 13.6.3 The QCD Lagrangian

The whole of QCD can be summed up in a single equation that has the following constituent parts: there are six quarks or ‘flavors’ (\( u, d, c, s, t, b \)); each quark carries a fractional electric charge: \( q_u = q_c = q_t = +2/3 \) and \( q_d = q_s = q_b = -1/3 \); let \( \psi_j, j = 1, \ldots, 6 \) represent the quark field for each quark, then \( \bar{\psi}_j \) represents the six antiquarks and each quark comes in three colors (red, blue, green). Hence,

\[
\mathcal{L}_{\text{QCD}} = \frac{1}{4 g^2} G_{\mu
u}^a G^{a\mu\nu} + \sum_j \psi_j (i\gamma^\mu D_\mu + m_j) \bar{\psi}_j
\]

where

\[
G_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + f_{bc}^a A_\mu^b A_\nu^c, \quad D_\mu = \partial_\mu + i t^a A_\mu^a
\]

The \( G_{\mu\nu}^a \)’s play the same role in QCD as \( F_{\mu\nu} \) does in QED, except there are eight of them instead of one. The \( m_j \)’s are the masses of the quarks, \( g \) is the coupling constant, the \( A \)’s are the gluon fields and the \( a, b, c \)’s are color indices. The numerical coefficients ‘\( f \)’ and ‘\( t \)’ guarantee \( SU(3) \) color symmetry. The factor ‘\( D_\mu \)’ is the covariant derivative.

The Lagrangian ‘\( \mathcal{L}_{\text{QCD}} \)’ has chiral symmetry, in other words, \( \mathcal{L}_{\text{QCD}} \) remains invariant whether left- or right-handed fields are used in the equation. The masses ‘\( m_j \)’s \( \neq 0 \)’ will break the symmetries of the theory. In the limit ‘\( m_u = m_d = 0 \)’, \( \mathcal{L}_{\text{QCD}} \) has a \( SU(2)_L \times SU(2)_R \times U(1) \) flavor symmetry, where the \( U(1) \) symmetry is associated with baryon number.
13.7 The Weak Interaction

The first attempt at describing the weak interaction, which, at the time, was thought responsible for beta decay, was introduced by Enrico Fermi in 1934. Fermi described his theory in terms of current flow, where, for example, a proton was created from a neutron, producing a current. The theory possessed a couple of undesirable features: 1) it was non-renormalizable, producing infinites that could not be absorbed into the theory; 2) there was a violation of the conservation of energy. Even with these flaws, the theory made reasonably good predictions. Mathematically,

\[ -\mathcal{L} = H \approx G_F J_{\mu}^\dagger J^\mu, \quad J_{\mu}^\dagger \neq J^\mu, \quad J_{\mu}^\dagger = \bar{p} \gamma_{\mu} n + \bar{v}_e \gamma_{\mu} e, \quad J^\mu = \bar{n} \gamma^\mu p + \bar{\epsilon} \gamma^\mu \nu_e \]

where \( p, n \) are hadrons, \( e, \nu \), leptons and \( G_F \approx 1.17 \times 10^{-5} \text{[GeV}^{-2}] \), called ‘Fermi’s constant’.

The partial Feynman diagrams associated with \( \mathcal{L} \) are shown in fig. 13.7.4-1,

\[ \begin{array}{c}
\text{Figure 13.7.4-1}
\end{array} \]

which show only the currents. The Fermi theory was modified several times, improved upon, with updates that included parity violation, decays of heavy leptons, strangeness, the introduction of the quarks, and extended to include \( CP \)-violation. But even with all the modifications, the basic idea that Fermi put forward survived.

13.7.1 Parity Violation in Weak Interactions

The first improvement involved parity violation. Experimentally, the electrons produced in weak interactions all showed left-handed chirality, never right-handed. The emitted electrons lacked a mirror image reflection. Mirror image reflections are usually referred to as ‘parity’. The Lagrangian of Fermi’s theory was updated to reflect this experimental fact:
\[ -\mathcal{L} = H \approx \frac{G_F}{\sqrt{2}} J^\mu_i J^\mu_j, \quad J^\mu = J^\mu_\mu + J^\mu_\rho, \]

\[ J^\mu = \bar{e} \gamma^\mu (1 - \gamma^5) v_e + \bar{\nu} \gamma^\mu (1 - \gamma^5) v_\nu + \{\tau\} = 2\bar{e}_L \gamma^\mu v_{eL} + 2\bar{\nu}_L \gamma^\mu v_{\nu L}, \]

\[ \Psi_L = \frac{(1 - \gamma^5)}{2} \psi \equiv \text{left-handed chirality}, \quad \Psi_R = \frac{(1 + \gamma^5)}{2} \psi \]

where \((1 - \gamma^5)/2\) turns any field into a left-handed field and \(\{\tau\}\) represents the '\(\tau\)' portion of the Lagrangian, which is not important for this part of the discussion. Note that

\[ H \approx \frac{G_F}{\sqrt{2}} J^\mu_i J^\mu_j = \frac{G_F}{\sqrt{2}} (J^\mu_\mu + J^\mu_\rho)(J^\mu_\mu + J^\mu_\rho) = H^{ll}_{\text{Lepton}} + H_{sl}^{\text{Non-Lepton}} + H_{hh}^{\text{Hadronic}} \]

In other words, the multiplication above produces three terms labeled '\(H_l, H_{sl}, H_{hh}\)' respectfully. Each of the three terms represents a different process. For example, \(H_l\) represents processes where leptons decay into other leptons i.e. \(\mu^- \to e^+ v_\mu + \bar{v}_e\); the term '\(H_{sl}\)' represents decays that are partially hadronic and partially leptonic i.e. \(n \to p + e + \bar{v}_e\) and \(H_{hh}\) represents hadrons decaying into other hadrons i.e. \(K^+ \to \pi^+ + \pi^0\). It can also represent the scattering of hadrons. There are hundreds and hundreds of these processes. The rules governing \(\mathcal{L}\) determine which processes take place and which are forbidden.

Experimentally, around 1956, two particles were discovered named '\(T, \theta\)' respectfully, which decayed thusly:

\[ \theta \to \pi^+ + \pi^0, \quad T \to \pi^+ + \pi^0 + \pi^0 \]

Both the \(\theta\) and the \(T\) had charge '+1', were spin-0, had the same mass and the same life time. However,

\[ \hat{P}|\pi^+, \pi^0\rangle = +1|\pi^+, \pi^0\rangle, \quad \hat{P}|\pi^+, \pi^0, \pi^0\rangle = -|\pi^+, \pi^0, \pi^0\rangle, \]

where \(\hat{P}\) signifies the parity operator. The process shown above assumes that the angular momentum of both particles is zero. The only reasonable explanation was to regard the \(\theta\) and the \(T\) as the same particle, called the '\(K^+\)', and conclude that parity symmetry is violated in weak interactions. Hence,

\[ J^\mu = \bar{e} \gamma^\mu (1 - \gamma^5) v_e = 2\bar{e}_L \gamma^\mu v_{eL} + \{\mu\} + \{\tau\} = -\bar{v}_e (1 + \gamma^5) \bar{e} = 2\bar{v}_{eR} \gamma^\mu \bar{e}_R, \]

where \(\{\mu\}, \{\tau\}\) signify the \(\{\mu\}\) and \(\{\tau\}\) pieces of the current, but will be ignored for now. Hence, \(e_L, \bar{e}_R\) and \(v_{eL}, \bar{v}_{eR}\) participate in weak interactions, but \(e_R, v_{eR}\) do not. Under the parity operator \(\hat{P}\):

\[ J^\mu_L(x) \to J^\mu_R(\bar{x}), \quad J^{\mu\nu}_L(x) \to J_{\nu}^{\mu}(\bar{x}) \]
Under charge conjugation ‘\( \hat{C} \)’, where \( \hat{C} \) changes the charge on a particle,

\[
J_{L}^{t\mu}(x) \rightarrow -J_{R}^{t\mu}(x), \quad J_{L}^{t\mu}(x) \rightarrow -J_{R}^{t\mu}(x), \quad x = (t, \vec{x}), \quad \vec{x} = (t, -\vec{x})
\]

Hence, parity followed by charge conjugation gives

\[
H \approx 4G_{F} \sqrt{2} \int \mu(x) \int \mu(\vec{x}) J_{L}^{t\mu}(x) J_{R}^{t\mu}(\vec{x})
\]

Note that the second expression and the last expression are essentially the same save for an \( \vec{x} \). Evidently, Nature is not symmetric under parity or charge conjugation, but, in the Fermi model, \( CP \) is a symmetry. It was discovered that ‘\( CP \)’ is not a symmetry of Nature. So, the Fermi model required additional improvements.

13.7.2 The Cabibbo Model

The next revision of the Fermi model was provided by the Italian physicist Nicola Cabibbo (10 April 1935 – 16 August 2010). Beta decay ‘\( n \rightarrow p + e + \bar{\nu}_{e} \)’ was modelled as follows:

\[
J_{\mu}^{t\mu} = \bar{p} \gamma_{\mu}(1 - \gamma^{5})n
\]

But there were many similar processes, \( \pi^{+} \rightarrow \mu^{+} + \nu_{\mu} \), for example, each of them representing another term making the equation unmanageably complicated. The complexity was improved upon by inventing quarks i.e.

\[
J_{\mu}^{t\mu} = \bar{u} \gamma_{\mu}(1 - \gamma^{5})d
\]

But beta decays represented only about 95% of the decays. About 5% were \( \mu \) decays. In beta decays the strangeness number was conserved i.e. \( \Delta s = 0 \). But there were decays, \( K^{+} \rightarrow \mu^{+} + \nu_{\mu} \), for example, where \( \Delta s \neq 0 \). Experimentally, decays such as \( K^{+} \rightarrow \mu^{+} + \nu_{\mu} \) and \( \Sigma^{-} \rightarrow \mu + e + \bar{\nu}_{e} \) occurred in Nature, but seemingly equally likely decays such as \( \Sigma^{+} \rightarrow n + \bar{e} + \nu_{\mu} \) and \( \Sigma^{-} \rightarrow \bar{p} + e + \bar{e} \) did not. What Cabibbo noticed was that in the decays that were permitted, it was always the case that \( \Delta s = \Delta q \neq 0 \), where \( q \) is the electric charge on the particle. Cabibbo modified the equation above to read

\[
J_{\mu}^{t\mu} = \bar{u} \gamma_{\mu}(1 - \gamma^{5})^{d'} \quad d' = d \cos \theta_{C} + s \sin \theta_{C}, \quad \sin \theta_{C} \approx 0.23,
\]

where \( \theta_{C} \) is called the ‘Cabibbo angle’. Hence,

\[
J_{\mu}^{t\mu} = \cos \theta_{C} \bar{u} \gamma_{\mu}(1 - \gamma^{5})d + \sin \theta_{C} \bar{u} \gamma_{\mu}(1 - \gamma^{5})s
\]

So the probability that a particle in a weak interaction changes its strangeness number is much smaller than if strangeness is conserved and only happens if the change in strangeness equals the change in electric charge. The Cabibbo model is \( CP \)-invariant.
13.7.3 The Rise of the Electroweak Theory

In spite of a myriad of improvements, the Fermi model broke down at high energies:

1. The theory was not unitary
2. It was non-renormalizable
3. It made incorrect predictions involving kaon physics

\[
\begin{align*}
\text{Figure 13.7.4.3-1}
\end{align*}
\]

The scattering \(\nu_e + e \rightarrow e + \nu_e\)', in the Fermi theory, is conceived as happening at one space-time point, where \(\sigma_s \rightarrow G_F^2 s / \pi, \ s = E_{cm}^2, \ E_{cm} \gg m_e\) and \(E_{cm}\) is the two particle center of mass and \(\sigma_s\) is the cross sectional area of scattering. It turned out that \(\sigma_s < 16 \pi / s\) i.e. it is bounded for any high energy process. If \(E_{em} / 2 \geq \sqrt{\pi / G_F} \approx 500 \text{ GeV}\), the theory lost unitarity, so that the probability amplitudes exceeded 1. Moreover, at the scattering point, the 4-momentum of the particles diverges i.e. \(\int d^4 k = \infty\). There was no way of choosing a sensible cutoff for \(\sigma_s\) that worked in every case. In other words, the theory was non-renormalizable. The fact that the theory made incorrect predictions for kaon physics was corrected by introducing the 'charm quark'. Although the Fermi theory made good approximations at low energy scales, it broke down at higher energies (fig. 13.7.4.3-1)

13.7.4 The Intermediate Vector Boson Theory

Recall that if \(\hbar = 1\), then \(E = \omega\) and the energy is equal to the frequency of the wave. The units for frequency are \([1/\sec]\). In QFT, distance ‘s’ and time ‘t’ are not distinct. If \(c = 1\), then \(s = t\). Hence, distance is just the inverse of the energy i.e. \(1 \text{ [eV}^{-1}\) \approx 10^{-7} \text{[m]}\). Table 13.7.4.4-1 shows some rough orders of magnitude for various measurements of a typical atom and a typical hadron. The ‘transit time’ is the time it takes light to travel across the diameter of an atom or a hadron.
In QED the coupling constant, which is roughly equal to the probability that an electron emits a photon, is \( \approx 1/137 \). In other words, if 137 electrons slammed into a cathode ray tube, then, on average, one photon would be emitted. In this sense, electromagnetism is called a ‘weak force’ or a ‘weak interaction’. All the times listed for the atom in table 13.7.4.4-1 are related to the fine structure constant ‘1/137’. Because the fine structure constant is small, the acceleration of the electron around the atom is, comparatively speaking, slower than the hadron times. The time for the electron in an excited state, for example, to move to the ground state and give off a photon is comparatively long, all due to the fact that the fine structure constant is small.

On the other hand, the orbital time, the time for a quark to swing around the interior of a proton, for example, is much smaller than the electron orbital time. The typical decay time for an excited proton to emit a pion is very short. This can only happen if the analog to the fine structure constant is much larger, \( \approx 1/4 \), which is the reason color charge is considered strong. The probability that a quark will emit a gluon is \( \approx 1 \).

In general, weak interactions are called ‘weak’ because their decay times are comparatively long in duration. For example, neutron decay \( n \rightarrow e + p + \bar{\nu} \) takes about 12 [minutes]. Proton decay is a hypothetical form of radioactive decay in which the proton decays into lighter subatomic particles, such as a neutral pion and a positron. But if the proton decays, it takes \( \approx 10^{34} \) [years]. According to the Standard Model, protons are stable because baryon number is conserved. Protons will not decay into other particles on their own, being the lightest (least energetic) baryon. There are experiments currently on-going that attempt to detect proton decay. To date, all attempts have failed.

One of the reasons the neutron is stable is that the mass of the proton is only slightly less than the mass of the neutron. The mass \( e + p \) is \( \approx 939.5 \) [MeV]. The mass of a neutron is about 940 [MeV]. There is not much energy left for this decay to occur. If the neutron was only slightly less massive, the decay could not occur at all, since it would violate the law of conservation of energy.

But the slowness of the decays in weak interactions is not entirely due to energy differences. For instance, the decays \( \pi^- \rightarrow e + \bar{\nu} \) and \( \pi^+ \rightarrow \bar{e} + \nu \) take about
10 [nanoseconds], which, comparatively speaking, is a fairly long time. Yet, there is enough energy available for these decays to occur in a much shorter time. There must be more to the explanation than the amount of available energy.

The difficulties associated with the Fermi model were overcome by the 'intermediate vector boson theory' introduced by the American theoretical physicist Julian Seymour Schwinger (February 12, 1918 – July 16, 1994). He won the Nobel Prize for this work in 1965. Schwinger’s idea was to introduce an intermediate particle (boson) that mediated the interaction between the leptonic and hadronic particles, making it look like a non-Abelian gauge theory similar to $QED$. The introduction of the boson created a time lag between the interactions and effectively got rid of the non-renormalization and unitarity issues associated with the Fermi theory. Figure 13.7.2-1 shows the Feynman diagram of a neutron decaying into a proton. The mediating particle is the $W^-$ boson. The strength of the interaction is given by

$$G_F \approx \frac{g^2}{q^2 + m_W^2}, \quad q^2 \ll m_W^2 \rightarrow G_F \approx \frac{g^2}{m_W^2},$$

where $g$ is the coupling constant and $m_W$ is the mass of the $W^-$ boson, which lives for a finite amount of time, eliminating the non-renormalizable issues with the Fermi theory. However, there are other processes like the one shown in the diagram below,

where an electron and a positron can decay into a $W^-$ and $W^+$ boson. The same issues arise as in the Fermi theory, because there is no mediating particle. There must be a third weak interaction mediating particle to handle the kind of interaction shown in fig. 13.7.4.4-2.
In 1961, Sheldon Glashow introduced a $SU(2) \times U(1)$ electroweak theory using the intermediate bosons $W^\pm, W^0, B \rightarrow W^0 B \rightarrow Z$. Glashow had to put the masses of the mediating particles in by hand, which violated the gauge invariance of the theory. But in 1967, Weinberg and Salam modified Glashow’s theory by incorporating the Higgs mechanism, which restored the gauge invariance and this theory made it into the standard model of particle physics. For their work, Glashow, Weinberg and Salam shared the 1975 Nobel Prize.

The electroweak theory has an enormously complicated Lagrangian:

$$\mathcal{L}_{SU(2)\times U(1)} = \mathcal{L}_{gauge} + \mathcal{L}_\phi + \mathcal{L}_f + \mathcal{L}_{int}$$

Without revealing the details, $\mathcal{L}_{SU(2)\times U(1)}$ is made up of four pieces: 1) ‘$\mathcal{L}_{gauge}$’ contains two terms, one for $SU(2)$ and the other for the $U(1)$ symmetry of the theory, 2) ‘$\mathcal{L}_\phi$’ is a scalar field associated with the Higgs phenomenon (discussed in more detail later), 3) ‘$\mathcal{L}_f$’ describes fermion fields and ‘$\mathcal{L}_{int}$’ covers the interactions between the Higgs field ‘$\phi$’ and the fermion fields.

### 13.8 A Summary of the Standard Model of Particle Physics

All quarks come in three colors ($R, B, G$) (see table 13.8–1). The color fields associated with the $SU(3)$ symmetry group, governing the strong interactions, are represented vertically in the table. A red $u$-quark can be rotated into a blue or green $u$-quark or a red $d$-quark can transform into a blue or green $d$-quark, but $u$-quarks cannot be transformed into $d$-quarks, for example, under a $SU(3)$ rotation. In other words, $SU(3)$ compliant field operators will not change the flavor of a quark.
The fields associated with the $SU(2)$ symmetry group, governing the weak interactions, are represented horizontally in the table. The operators associated with $SU(2)$ operate on doublets. In other words, those fields can change a red $u$-quark into a red $d$-quark, but $SU(2)$ field operators conserve color charge. The fields also cannot jump to other generations of quarks. A $u$-quark cannot be transformed into a $c$-quark and so on.

Note that in table 13.8-1, when moving horizontally, the difference in the charge of the two particles is 1. For instance,

$$ch(u - quark) - ch(d - quark) = 2/3 - (-1/3) = 1$$

$$ch(\nu_e) - ch(e) = 0 - (-1) = 1$$

Two spin-1/2 particles can make a spin-1 particle, a boson. Consider

$$e + \bar{\nu}_e \rightarrow W^- \leftarrow d + \bar{u}$$

Note that the electron plus antineutrino has charge ‘−1’ as does the pair ‘down-quark-anti-up-quark ($-1/3 + (-2/3) = -1$). Similarly,

$$\nu_e + \bar{e} \rightarrow W^+ \leftarrow u + \bar{d}$$

has charge ‘+1’.

13.8.1 Permissible Interactions

What kinds of interactions can be made from this? An electron-neutrino can be produced from an electron by emitting a $W^-$ or a down-quark can be turned into an up-quark by emitting a $W^-$, which plays the same role in weak interactions as the gluons in strong interactions and the photon in electrodynamic interactions (see fig. 13.8.1-1 a),
b)). The $W^+$ is the antiparticle of the $W^-$. For example, a $W^+$ can be absorbed by an electron to create an electron-neutrino. Hence, gluons, the $W^\pm$, $Z$ and the photon are all called 'gauge bosons'.

Interestingly, a $c$-quark can be created from a $s$-quark by emitting the same $W^-$. An $s$-quark could be changed into an $c$-quark by emitting a different gauge boson than the $W^-$, but the Lagrangian associated with weak interactions is constrained by $SU(2)$, which requires that the same gauge boson be involved in all the transformations of the generations of leptons and quarks. Moreover, this agrees with direct observations.

All the rules of Feynman diagrams hold. For example, a $d$-quark an $\bar{u}$-quark can annihilate each other to produce a $W^-$. This leads to a whole conglomeration of possible decays. For example,

a $\pi^-$ consists of a $d$-quark an $\bar{u}$-quark. They can annihilate to create a $W^-$, which, in turn, decays into an electron and an anti-electron-neutrino (see fig. 13.8.1-2). All possible weak interactions occur in Nature, except in cases where not enough energy is available. Since mass is associated with energy, generally speaking, heavier particles decay into lighter ones. For instance, the mass of a $\pi^-$ is $\approx 140 [MeV]$, where the mass
of $e + \bar{\nu}$ is $\approx 100 \ [MeV]$. Since energy is conserved, the difference must come in the form of kinetic energy.

Referring to table 13.8-1, the $SU(3)$ operators, for example, annihilate a red-quark and create a blue-quark by emitting a gluon (see fig. 13.8.1-3). The $SU(3)$ symmetry group requires that the same gluon be emitted if the color operator acts on, say, a $\tau$-quark. In other words, the same gluon is emitted when a red $\tau$-quark turns into a blue $\tau$-quark as when a red $u$-quark turns into a blue $u$-quark.

![Figure 13.8.1-3](image)

By the same token, the $SU(2)$ symmetry group requires that the same gauge boson is emitted if the flavor operator acts on, say, a $\tau$-quark instead of a $u$-quark. In other words, the same gauge boson is emitted when a red $\tau$-quark turns into a red $b$-quark as when a red $u$-quark turns into a red $d$-quark.

13.8.2 Coupling Constants

The coupling constants for the electrodynamic, strong and weak interactions are proportional to the probability that the interactions take place. The electrodynamic coupling constant '$\alpha \approx 1/137$' is proportional to the electric charge i.e. $\alpha \propto e^2$. The strong interaction coupling constant '$g_S \approx 1/4$' is much larger than the electrodynamic coupling constant, making the emission of a gluon by a quark more likely than the exchange of a photon between two electrons. The fact that weak interactions occur far less often than electrodynamic interactions suggests that the weak interaction 'coupling constant' would be extremely small, close to zero. But, in fact, this is not the case. So, why are weak interactions so unlikely?

13.8.3 The Anatomy of Feynman Diagrams

Figure 13.8.3-1 shows the Feynman diagram of a neutron decaying into a proton. In this case, one of the down-quarks is changed into an up-quark by emitting a $W^-$, which then decays into an electron and an electron-antineutrino.
There are two important factors associated with Feynman diagrams: ‘vertices’ and ‘propagators’. The vertices represent points where the interaction takes place and are associated with the coupling constant. In this case, the weak interaction coupling constant ‘$g_w$’ acts twice - once when the down-quark is changed into an up-quark and again when the $W^-$ decays into an electron and an electron-antineutrino. Since it acts twice, the probability that these interactions take place is proportional to $g_w^2$.

The propagator ‘$\Delta$’ is associated with the amount of time that the virtual boson exists before it decays i.e. $\Delta = \text{(function of distance)}$ (recall that in QFT time and distance are the same measure). In fact, $\Delta$ is a function of the space-time distance between the creation and annihilation of a propagator.

The lifetime of a propagator is highly dependent on its mass. If it is massless, its lifetime is longer than if it carries a non-zero mass. The more mass it has, the shorter its lifetime (see fig. 13.8.3-1).

To see this, recall that $d = 1/m$, where $d$ is the distance and $m$ is the mass i.e.

$$[m] = [E] = [1/\text{length}] = [p] \rightarrow [\text{length} = 1/m] \ (c = \hbar = 1)$$
The propagator ‘Δ’ is associated with the probability that a gauge boson lasts a certain amount of time before it decays. Experimentally, Δ is found by measuring the difference between the momentums, for example, of a neutron and a proton, when the neutron decays into the proton. This change in the momentums is just \[1/\text{length}\]. The momentum of Δ is related to its position by a Fourier transform:
\[
\tilde{\Delta}(k_{W^-}) = \frac{1}{k^2 + m_{W}^2}
\]

The momentum of the \(W^-\) is equal to 1 over the square of the difference between the momentum of the neutron and proton \(k^2\), plus the square of the mass of the \(W^-\); the quantity \(k\) is usually pretty small, hence, \(\Delta(k_{W^-}) \approx 1/m_{W}^2\). Since Δ is proportional to the amplitude i.e. \(amplitude \propto g_{W}^2/m_{W}^2\), a very small number. The reason weak interactions take a long time is because the \(W^-\) is a relatively massive particle. Note that since \([E] = [1/\text{length}] = [p] \rightarrow [\text{length} = 1/m]\) and \(d (\text{length}) = t\), the \(W^-\) can live for a comparatively short period of time.

Rates of decays are influenced by the following factors:

1. The value of the coupling constants
2. The masses of the intermediate particles
3. The masses of the resultant particles

**13.8.4 Quantum Tunneling**

When a neutron decays into a proton, because the \(W^-\) is relatively massive, there is not enough energy to create the \(W^-\). In fact, the \(W^-\) is about 80 times heavier than the neutron. If this process happens, it should violate the law of conservation of energy. However, according to quantum mechanics, the law of conservation of energy can be violated so long as \(\Delta E \Delta t \leq \hbar\). An amount of energy \(\Delta E\) can be borrowed, but not for longer than \(\hbar/\Delta E\). This process is called ‘quantum tunneling’. Classically, there is not enough energy for quantum tunneling to take place. However, the uncertainty principle allows borrowing energy for a time, not to exceed \(\hbar/\Delta E\), so long as the energy is paid back in time so that the law of conservation of energy holds for the observable part of the interaction. Like many processes in quantum mechanics, quantum tunneling is not directly observable. Any attempt at measuring a gauge boson in a ‘non-conservation energy state’ will destroy the experiment.

![Diagram of Quantum Tunneling](image-url)
Since the mass of a proton is lighter than the neutron mass, proton decay, if it occurs, requires extra energy, for example, the extra energy contained in the Sun. Since mass is basically equivalent to energy, a proton could change into a neutron only if there where extra energy available. There are theories that predict proton decay, but the source of the extra energy must be explained, since there is more energy after the interaction than before – a clear violation of conservation of energy (see fig. 13.8.4-1).

### 13.8.5 Symmetry Breaking

While symmetry and conservation laws are the foundation of modern physics, Nature is asymmetrical. Systems in Nature migrate to their lowest energy state unless acted on by an external force. A system configured like the one shown in fig. 13.8.5-1 a), reaches its lowest energy state when \( \phi = 0 \). At that point, the potential energy is zero and the system will remain there unless an external force acts on it. This is an example of ‘no symmetry breaking’.

![Figure 13.8.5-1](image)

#### 13.8.5.1 Types of Symmetry Breaking

There are actually two kinds of symmetry breaking: ‘spontaneous symmetry breaking’ and ‘explicit symmetry breaking’. Symmetry breaking is usually associated with the lowest energy state of the system. If the lowest energy level of the system moves away from its symmetry point \( \phi = 0 \), then the symmetry is broken. If a system has an internal bias, for example, if the system shown in fig. 13.8.5-1 a) was tilted to the left or to the right, this is paramount to ‘explicit symmetry breaking’. On the other hand, the system shown in fig. 13.8.5-1 b), where, at \( \phi = 0 \), \( V(\phi) = 0 \), the system is symmetric, but not in its lowest energy state. In this case, the system is in an unstable condition and is inclined to migrate to one of its lowest energy states, either \( \phi = R \) or \( \phi = -R \). In both cases, \( V(\phi) < 0 \). If the system has no preference between \( \phi = R \) or \( \phi = -R \), then the symmetry is ‘spontaneously broken’.

Quantum physics requires that the symmetries of Nature be broken at some point. In fact, particles gain their mass through the process of ‘spontaneous symmetry breaking’. Theoretically, all particles in the standard model are massless. Particles only gain mass through the Higgs mechanism, which is a kind of breaking of the symmetries that exist in Nature.
Symmetries can be ‘exact’ or ‘approximate’. Most symmetries in Nature are approximate i.e. the symmetries are broken in some fashion. Mathematically, if 
\[ [\hat{T}^i, \mathcal{L}] = 0, \quad \hat{T}^i |g\rangle = |0\rangle \) and \( e^{i\beta \hat{T}^i} |g\rangle = |g\rangle \), where \( \hat{T}^i \)'s are the generators of the symmetry group that constrain the Lagrangian ‘\( \mathcal{L} \)' in other words, if the generators of the symmetry group commute with the Lagrangian and \( e^{i\beta \hat{T}^i} \) acting on a state leaves the state invariant, then the symmetry is ‘exact’. If \( \mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1 \), where \( [\hat{T}^i, \mathcal{L}_0] = 0 \), but \( [\hat{T}^i, \mathcal{L}_1] \neq 0 \), then the symmetry is said to be ‘explicitly broken’. If \( \hat{T}^i |g\rangle \neq |0\rangle \) and \( e^{i\beta \hat{T}^i} |g\rangle \neq |g\rangle \), then the symmetry is ‘spontaneously broken’.

13.8.5.2 Symmetry Breaking within a Lagrangian

Suppose there is a real field ‘\( \phi \)' that can take on values from \(-\infty \) to \( \infty \) and a Lagrangian ‘\( \mathcal{L} \)' such that
\[
\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - V(\phi), \quad V(\phi) = a\phi + \frac{\mu^2}{2} \phi^2 + \sigma\phi^3 + \lambda \frac{\phi^4}{4}
\]
Note that \( V \) is a polynomial. The Lagrangian above is not symmetric in \( \phi \). In other words, \( \mathcal{L} \) is not constrained by a symmetry. But a symmetry can be forced on \( \mathcal{L} \) by insisting that \( \mathcal{L} = -\mathcal{L} \). In that case, if \( \phi \to -\phi \), \( \mathcal{L} \) would remain invariant. In order for this to happen, \( a = \sigma = 0 \) in \( V \). The modified Lagrangian reads
\[
\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \left( \frac{\mu^2}{2} \phi^2 + \lambda \frac{\phi^4}{4} \right)
\]
The Euler-Lagrange equation associated with \( \mathcal{L} \) is:
\[
\left( \frac{\partial^2}{\partial t^2} - \nabla \right) \phi = \frac{\partial V}{\partial \phi} = -(\mu^2 + \lambda \phi^3)\phi
\]
The classical solution to the equation above is
\[
\frac{\partial}{\partial t} \langle \phi \rangle = 0 = \frac{\partial V}{\partial \phi} \bigg|_{\langle \phi \rangle}
\]
If \( \mu^2 > 0, \lambda > 0 \), then the system is represented by fig. 13.8.5-1 a). In this case, the system is in equilibrium when \( \phi = 0 \) and in the lowest state of potential energy. There is no propensity to move away from the state of equilibrium (no symmetry breaking). If \( \mu^2 > 0, \lambda < 0 \), then the system is unstable, since its lowest energy state would be \(-\infty \). Such systems do not appear in Nature. Systems in Nature always have \( \lambda > 0 \). If \( \mu^2 < 0, \lambda > 0 \), then the system is represented by fig. 13.8.5-1 b), where the symmetry point ‘\( \phi = 0 \)' is unstable, but the lowest energy levels are finite i.e \( \langle \phi \rangle_{\text{min}} = \pm R = \pm \sqrt{-\mu^2/2} \). In this case, there are two ground states ‘\( \pm R \)' representing spontaneous symmetry breaking, where the system does not favor either state ‘\( \phi = +R \)' or ‘\( \phi = -R \)’. Without loss of generality, let \( \langle \phi \rangle_{\text{min}} = R = \sqrt{-\mu^2/2} \) and let \( \phi = R + \phi \)
\[ \mathcal{L} = \frac{1}{2} \partial_{\mu} \phi' \partial^{\mu} \phi' + V'(\phi'), \quad V'(\phi') = \frac{\mu^4}{4\lambda} - \frac{\mu^2}{2} (\phi')^2 + \lambda R (\phi')^3 + \frac{\lambda}{4} (\phi')^4 \]

Note that the substitution \( \phi = R + \phi' \) breaks the symmetry. Once the system is found in a state, say \( \phi = R' \), then it costs energy to put the system in the state \( \phi = -R' \). Mathematically, the kinetic energy terms in the Lagrangian would be large.

The example above discussed discrete symmetries, where a system tends toward the state \( +R \) or \( -R \), but with no preference. The standard model deals in continuous symmetries. For example, fig. 13.8.5-1 b) shows that the lowest energy states lie on a circle of radius \( R \). In this case, \( \phi \) is best represented by a complex field: \( \phi \neq \phi^\dagger \). The Lagrangian associated with this field becomes:

\[ \mathcal{L} = \frac{1}{2} \frac{\partial_{\mu} \phi^\dagger \partial^{\mu} \phi - V(\phi)}{\kappa.E} - V(\phi) \]

Under what conditions does \( \mathcal{L} \) have a symmetry? Suppose that \( \phi' = e^{i\theta} \phi \) and \( \phi'^\dagger = e^{-i\theta} \phi^\dagger \). Then

\[ \partial^{\mu} \phi'_{\mu} = e^{i\theta} \partial^{\mu} \phi_{\mu}, \quad \partial_{\mu} \phi'^\dagger_{\mu} = e^{-i\theta} \partial_{\mu} \phi^\dagger_{\mu} \]

If \( V \) is a function of \( \phi_{\mu} \phi^\dagger_{\mu} \) i.e. \( V(\phi_{\mu} \phi^\dagger_{\mu}) \), then \( \mathcal{L} \) remains invariant if \( \phi_{\mu} \) is replaced by \( \phi'_{\mu} \) and \( \phi^\dagger_{\mu} \) is replaced by \( \phi'^\dagger_{\mu} \). To see this, note that

\[ \partial^{\mu} \phi_{\mu} \partial^{\mu} \phi'^\dagger_{\mu} = e^{i\theta} e^{-i\theta} \partial^{\mu} \phi_{\mu} \partial^{\mu} \phi'^\dagger_{\mu} = \partial^{\mu} \phi_{\mu} \partial^{\mu} \phi'^\dagger_{\mu}, \]

\[ V(\phi_{\mu} \phi^\dagger_{\mu}) = V(\phi'^\dagger_{\mu} \phi'^\dagger_{\mu}) = V(\phi_{\mu} \phi^\dagger_{\mu}) \]

In this case, \( \mathcal{L} \) has a \( U(1) \) symmetry, the one associated with the electromagnetic field, which rotates the complex plane by an angle \( \theta \).

Now suppose that

\[ \phi = \frac{\phi_R + i\phi_I}{\sqrt{2}}, \quad \phi^\dagger = \frac{\phi_R - i\phi_I}{\sqrt{2}}, \quad \phi_R = \phi^\dagger_R, \quad \phi_I = \phi^\dagger_I \]
Then

\[ L = \frac{1}{2} \frac{\partial_{\mu} \phi^{\dagger} \partial_{\mu} \phi - V(\phi)}{K.E} \]

\[ = \frac{1}{2} \left[ (\partial \phi_R)^2 + (\partial \phi_I)^2 - \left[ (\partial \phi_{R_x})^2 + (\partial \phi_{I_x})^2 \right] - \left[ (\partial \phi_{R_y})^2 + (\partial \phi_{I_y})^2 \right] \right] \]

\[ - \left[ (\partial \phi_{R_z})^2 + (\partial \phi_{I_z})^2 \right] - V(\phi_R, \phi_I), \]

\[ V(\phi_R, \phi_I) = \frac{\mu^2}{2} (\phi_R^2 + \phi_I^2) + \frac{\lambda}{4} (\phi_R^2 + \phi_I^2)^2 \]

In the case above, \( L \) manifests an \( O(2) \) or rotational symmetry i.e.

\[ \begin{bmatrix} \phi_R' \\ \phi_I' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \phi_R \\ \phi_I \end{bmatrix} \]

Note that \( V \), the potential, can be chosen so that the symmetry is maintained or is broken. For instance, if \( \min[V(\phi_R, \phi_I^\dagger)] = 0 \), then the symmetry is not broken. In this case, \( R_R = R_i = 0, \mu^2 > 0 \).

But if \( \mu^2 < 0 \), suppose

\[ V = -R \phi_{\mu} \phi_{\mu}^\dagger + R(\phi_{\mu} \phi_{\mu}^\dagger)^2, \quad R^2 = -\mu^2/2 > 0, \]

then \( \min[V(\phi_{\mu} \phi_{\mu}^\dagger)] \neq 0 \) and the symmetry is broken, where the locus of \( V \) vs. \( \phi_{\mu} \) is shown in fig. 13.8.5-1 b). The potential has many minimum values that lie on a circle (see fig. 13.8.5-1 c)). The ground states of the system lie at any point along the circle. However, it costs energy to move from one point to another along the circle, primarily associated with the gradient terms in the Lagrangian.

**13.8.5.3 The Goldstone Boson and Mass**

The next step is to quantize around a point. Since all the points along the circle are rotationally invariant, quantization can take place around any of the points, say \( \phi = R \). If \( \phi_R = R + \phi_R' \) and \( \phi_I = \phi_I \), then

\[ V(\phi_R, \phi_I) = -\frac{\mu^4}{4\lambda} - \frac{\mu^2}{2} (\phi_R')^2 + \lambda R((\phi_R')^2 + (\phi_I')^2) + \frac{\lambda}{4} ((\phi_R')^2 + (\phi_I')^2)^2 \]

\[ \text{min}_{m_{\phi_R'}^2 = -2\mu^2 > 0} \]

Note that there is a mass associated with the field \( \phi_R' \), given by the term \( -\mu^2(\phi_R')^2/2 \), but no mass term associated with \( \phi_I' \) i.e. \( m_{\phi_I'}^2 = 0 \). Hence, a particle with mass is associated with oscillations around the minimum.
In quantum physics, when a particle moves such that its potential energy is at a minimum and the only cost in energy are changes in the gradient terms (changes in kinetic energy), then the particle is massless. In other words, if a wave function, say \( \phi(k) = R(k)e^{ikx} \), is such that \( \lim_{k \to 0} \phi(k) = 0 \), then \( \phi \) is a field associated with a massless particle. If the wavelength becomes longer and longer as \( k \to 0 \), where \( k \) signifies ‘momentum’, then \( \phi \) is associated with a massless particle. On the other hand, if \( \lim_{k \to 0} \phi(k) \neq 0 \), then \( \phi \) is associated with a particle with non-zero mass. For example, if \( V = m^2 \phi \phi^\dagger /2 \), then any movement away from \( \phi \phi^\dagger = 0 \) would cost energy, corresponding to a particle of non-zero mass.

To better understand this, suppose \( \mathcal{L} = \partial \phi^\dagger \partial \phi - V(\phi^\dagger \phi) \) and \( \phi = Re^{i\alpha} \), where both \( R \) and \( \alpha \) are fields, then

\[
\partial \phi = \partial Re^{i\alpha} + iR\partial \alpha e^{i\alpha} = (\partial R + iR\partial \alpha)e^{i\alpha}, \quad \partial \phi^\dagger = (\partial R - iR\partial \alpha)e^{-i\alpha}
\]

So,

\[
\partial \phi \partial \phi^\dagger = (\partial R + iR\partial \alpha)e^{i\alpha}(\partial R - iR\partial \alpha)e^{-i\alpha} = (\partial R + iR\partial \alpha)(\partial R - iR\partial \alpha) = (\partial R)^2 + R^2(\partial \alpha)^2
\]

and

\[
V(\phi^\dagger \phi) = V(Re^{-i\alpha}Re^{i\alpha}) = V(R^2)
\]

Hence,

\[
\mathcal{L} = (\partial R)^2 + R^2(\partial \alpha)^2 - V(R^2)
\]

Note that the potential \( 'V' \) is only a function of \( R^2 \). If \( R \) is a constant, then \( (\partial R)^2 = 0 \) and the only kinetic term is \( R^2(\partial \alpha)^2 \), which depends only on the angle ‘\( \alpha \)’. Since \( V \) does not change, it is always a minimum, corresponding to fig. 13.8.5-1 b), in which the particle continuously stays in the lowest state of potential energy. This particle is massless. If \( R \) varies, then the potential moves away from the minimum, corresponding to a particle with a non-zero mass. The upshot is that mass is associated with oscillations in the potential.

Note that the situation illustrated in fig. 13.8.5-1 b) corresponds to a rotational symmetry. Anytime there are massless particles associated with a rotational symmetry, the particles are called ‘Goldstone bosons’. The concept of a ‘Goldstone boson’ is crucial to understanding the ‘Higgs mechanism’, which gives all particles mass.

As an aside, in quantum mechanics, a constant can be added to the Lagrangian without altering the physics, since the probabilities only depend on the fields. But in any attempt at uniting quantum theory with general relativity, adding a constant is generally associated with the cosmological constant within Einstein’s field equations. Comparatively speaking, the constants added to the Lagrangians of each sector of the standard model are large relative to the value of the cosmological constant by some 120 orders of magnitude. The situation represents one of the biggest impediments to uniting
quantum field theory with general relativity. There seems no way of producing a small cosmological constant within quantum field theory.

13.8.6 Gauge Invariance

Recall that the Lagrangian of quantum electrodynamics remains invariant under a $U(1)$ symmetry transformation, given mathematically by

$$\phi' = e^{i\theta(x)} \phi, \quad \phi'^\dagger = e^{-i\theta(x)} \phi \to \phi' \phi'^\dagger = \phi \phi$$

This transformation is invariant so long as $\theta(x) = \text{a constant}$. But if $\theta(x)$ varies over space-time and

$$\mathcal{L} = \partial \phi^\dagger \partial \phi - V(\phi^\dagger \phi)$$

then,

$$\partial \phi' = e^{i\theta} \partial \phi + i\phi \frac{\partial \theta}{\partial x} e^{i\theta} = \left( \partial \phi + i\phi \frac{\partial \theta}{\partial x} \right) e^{i\theta},$$

$$\partial \phi'^\dagger = \left( \partial \phi'^\dagger - i\phi^\dagger \frac{\partial \theta}{\partial x} \right) e^{-i\theta} \to \partial \phi' \partial \phi'^\dagger$$

$$\to \left( \partial \phi + i\phi \frac{\partial \theta}{\partial x} \right) e^{i\theta} \left( \partial \phi'^\dagger - i\phi^\dagger \frac{\partial \theta}{\partial x} \right) e^{-i\theta}$$

$$= \partial \phi \partial \phi'^\dagger + i(\partial \phi'^\dagger \phi - \partial \phi \phi^\dagger) \frac{\partial \theta}{\partial x} + \phi \phi^\dagger \left( \frac{\partial \theta}{\partial x} \right)^2$$

Therefore,

$$\mathcal{L}' = \partial \phi \partial \phi'^\dagger + i(\partial \phi'^\dagger \phi - \partial \phi \phi^\dagger) \frac{\partial \theta}{\partial x} + \phi \phi^\dagger \left( \frac{\partial \theta}{\partial x} \right)^2 - V(\phi'^\dagger \phi')$$

If the transformation ‘$\phi \to \phi'$’ was a symmetry, then $\mathcal{L}'$ should equal $\mathcal{L}$. But this is clearly not the case. So, if $\theta$ is a varying parameter, the symmetry is lost.

13.8.7 The Vector Potential

If $\theta$ varies, the symmetry can be recovered by introducing new fields into the Lagrangian, in particular, the vector potential ‘$A_{\mu}$’. Recall that $A_0$ is associated with the electrostatic potential and $\partial A_0$ is associated with the electric field; $A_i, \ i = 1,2,3$ is associated with the magnetic force and $\text{curl} (A_i)$ is associated with the magnetic field.

Let $A'_\mu = A_\mu - \partial \theta_\mu$ and in the Lagrangian replace all partial derivatives with the following:

$$\partial \phi_\mu \to D_\mu \phi = \partial_\mu \phi + iA_\mu \phi, \quad \partial \phi_\mu'^\dagger \to D_\mu \phi'^\dagger = \partial_\mu \phi'^\dagger - iA_\mu \phi'^\dagger$$
The Lagrangian then becomes
\[ \mathcal{L} = D_\mu \phi^+ D_\mu \phi - V(\phi^+ \phi) \]

Note that
\[ D_\mu \phi' = (\partial \phi_\mu + i \phi_\mu \partial \theta) e^{i\vartheta} + i(A_\mu - \partial \theta) \phi e^{i\vartheta} = (\partial \phi_\mu + i A_\mu \phi) e^{i\vartheta} \]
\[ D_\mu \phi^+ = (\partial \mu \phi^+ - i \phi^+ \partial \theta) e^{-i\vartheta} - i(A_\mu - \partial \theta) \phi^+ e^{-i\vartheta} = (\partial \mu \phi^+ - i A_\mu \phi^+) e^{-i\vartheta} \]

Hence,
\[ \mathcal{L}' = (\partial_\mu \phi^+ - i A_\mu \phi^+) e^{-i\vartheta} (\partial \phi_\mu + i A_\mu \phi) e^{i\vartheta} - V(\phi^+ \phi) = D_\mu \phi^+ D_\mu \phi - V(\phi^+ \phi) = \mathcal{L}, \]

making the Lagrangian gauge invariant. Recall that the electromagnetic field tensor is given by
\[ F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \]
and that \( F_{\mu\nu} = - F_{\nu\mu} \), where
\[ F_{\mu\nu} = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} \]

The matrix \( 'F_{\mu\nu}' \) is traceless. The terms \( 'F_{0i}' \) are associated with the electric field and the terms \( 'F_{ij}' \) associated with the magnetic field i.e. \( F_{xy} = -B_z \). Hence, \( F_{xy} \) is the magnitude of the magnetic field in the z-direction. Classically, the Lagrangian \( '\mathcal{L}' \) for the electromagnetic field is \( \mathcal{L} = F_{\mu\nu} F^{\mu\nu} \). Is \( \mathcal{L} \) gauge invariant? To see, note that
\[ \mathcal{L}' = \mathcal{L} \]

Hence, \( '\mathcal{L} = F_{\mu\nu} F^{\mu\nu}' \) is gauge invariant. It follows immediately that
\[ \mathcal{L} = D_\mu \phi^+ D_\mu \phi - V(\phi^+ \phi) + F_{\mu\nu} F^{\mu\nu} \]
is also gauge invariant.

The Lagrangian described above is associated with a massless boson, namely, the photon. If a term \( m^2/2 A_\mu A^\mu \) was added to \( \mathcal{L} \), it would be like giving the photon a mass. However, \( \mathcal{L} \) would no longer be gauge invariant i.e. \( m^2 A_\mu A^\mu /2 \neq m^2 A_\mu A^\mu /2 \).

13.8.8 The Higgs Mechanism

The Higgs mechanism is an explanation of how particles gain mass. The notion of mass in the standard model is association with oscillations.
An electromagnetic wave propagates through empty space at a speed 'c'. In fact, this is true of any wave. A sound wave, for example, moves at the speed of sound, and while sound waves do not propagate in empty space, the relationship 'ω ∝ k' still holds i.e. \( \lim_{k \to 0} \omega = 0 \). In other words, there is no energy in a wave that does not propagate. If it happens that \( \lim_{k \to 0} \omega \neq 0 \), then the energy associated with the wave must be due to a non-zero mass (see fig. 13.8.8-1 a)), \( k \to 0 \to p = 0 \) and the only way \( E \neq 0 \) is if \( m \neq 0 \), \( m \) being the mass. Note that \( \lim_{k \to 0} \omega \neq 0 \to \lim_{k \to 0} \lambda = \infty \), then, if a wave with an infinite wavelength has a non-zero energy, the wave oscillates uniformly in all parts of space and this oscillation is associated with a mass (see fig. 13.8.8-1 a), b)).

13.8.8.1 The Dispersion Relation

Evidently, in a gauge theory, the fields that propagate must have a source. For example, the electromagnetic field cannot exist without electric charge. The same can be said of gluon fields, which end on quarks. In the case where \( \lim_{k \to 0} \lambda = \infty \), then the vibration of the wave cannot be associated with wave oscillations and if \( \lim_{k \to 0} \lambda = \infty \to \omega = 0 \), then the particle is massless. On the other hand, if \( \omega \neq 0 \), then the particle has a non-zero mass. The relationship '\( d\omega/dk \)' is called the 'dispersion relation'. If \( d\omega/dk = a \text{ constant} \), the particle is massless. If \( d\omega/dk \neq a \text{ constant} \), then the particle is has a non-zero mass.

13.8.8.2 The Higgs Field

The relationship between energy and mass is given by the Einstein equation '\( E = mc^2 \)' or '\( E = m, c = 1 \)'. In a system, where the symmetry of the system is spontaneously broken (see fig. 13.8.5-1 b)), the field(s) connected with that system are associated with two particles, one with mass and one massless. Interestingly, the two particles cannot exist at the same moment. Either the system is in a massless state or in a massive state, never both at the same time.

Figure 13.8.8-1
To gain a better understanding of this, in the standard model, fields are mathematically complex: $\phi = \phi_R + i \phi_I$, where $\phi_R$ and $\phi_I$ are the real and imaginary parts of $\phi$ respectfully. For instance, let

$$ L = \partial \phi^\dagger \partial \phi - V(\phi^\dagger \phi) = \partial \phi^\dagger \partial \phi - \frac{m^2}{2}(\phi_R^2 + \phi_I^2) $$

be the Lagrangian of a system. Note that $L$ is gauge invariant, since if $\phi$ is replaced with $\phi' = e^{i\theta} \phi$, the Lagrangian remains unchanged. If $m = 0$, then $L$ depends only on the kinetic terms and not on a potential, the particle is massless. Whether or not a system has mass depends on the nature of potential terms in the Lagrangian.

The Lagrangian ‘$L$’ above can be written in polar form:

$$ L = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi^\dagger \phi) = (\partial_\mu \rho)^2 + \rho^2(\partial^\mu \alpha)^2 - V(\rho), $$

where $\phi = \rho e^{i\alpha}$ and $\phi^\dagger = \rho e^{-i\alpha}$ (see figure 13.8.2-1), where $\rho$ and $\alpha$ are fields.

Let $V(\rho) = \rho^2$. Note that $V(\rho)$ is minimum when $\rho = R$ (see fig. 13.8.5-1 b)) and $V(\rho)$ moves away from a minimum when $\rho = R + H$, $H > 0$. Now if $\rho = R$, the potential is stuck at a minimum and the only contribution to $L$ is the term involving $\alpha$, i.e. $R^2(\partial^\mu \alpha)^2$. Since $R$ is constant, only the angle ‘$\alpha$’ varies. The only thing that changes the energy of the system are changes in the kinetic terms in $L$, which corresponds to a massless particle. In fact, this represents the ‘Goldstone boson’.

\[ \text{Figure 13.8.8.2-1} \]

Recall that the transformation ‘$\phi' = e^{i\theta(x)}$’ is gauge invariant so long as $\theta(x)$ = a constant. But if $\theta$ varies with $x$, then the symmetry is lost, but can be regained by making the substitution ‘$\partial_\mu \phi \rightarrow \partial_\mu \phi + i A_\mu \phi'$, where $A_\mu$ is the vector potential of the electromagnetic field. The Lagrangian then becomes

$$ L = D_\mu \phi^\dagger D^\mu \phi - V(\phi^\dagger \phi), $$

where $D_\mu \phi = \partial_\mu \phi + i A_\mu \phi'$. Moreover, the term $F_{\mu \nu} F^{\mu \nu}$ is also gauge invariant, since $F_{\mu \nu}$ is a linear function of the vector potential. Hence, the Lagrangian

$$ L = D_\mu \phi^\dagger D^\mu \phi - V(\phi^\dagger \phi) + F_{\mu \nu} F^{\mu \nu} $$

413
is gauge invariant. The Lagrangian above describes a two particle system, the field ‘$\phi$’ being a charged boson, a Goldstone boson. The other particle is associated with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Note that $F_{\mu\nu}$ is a gradient term. There is no potential associated with it. Hence, the field ‘$F_{\mu\nu}F^{\mu\nu}$’ indicates a massless particle, the photon.

In order to give the photon a mass, a potential term must be added to $\mathcal{L}$ proportional to the square of the vector potential. The Lagrangian would then look something like

$$\mathcal{L} = D_\mu \phi^\dagger D_\mu \phi - V(\phi^\dagger \phi) + F_{\mu\nu}F^{\mu\nu} + \frac{m^2}{2} A_\mu A^\mu$$

The price paid for introducing a mass is that $\mathcal{L}$ is no longer gauge invariant.

However, consider

$$\mathcal{L} = D_\mu \phi^\dagger D_\mu \phi - V(\phi^\dagger \phi) + F_{\mu\nu}F^{\mu\nu}$$

which is gauge invariant. Recall that $D_\mu \phi = \partial_\mu \phi + iA_\mu \phi$ and $D_\mu \phi^\dagger = \partial_\mu \phi^\dagger - iA_\mu \phi^\dagger$. Hence,

$$\mathcal{L} = D_\mu \phi^\dagger D_\mu \phi - V(\phi^\dagger \phi) + F_{\mu\nu}F^{\mu\nu} = (\partial^\mu \phi^\dagger - iA_\mu \phi^\dagger)(\partial_\mu \phi + iA_\mu \phi) - V(\phi^\dagger \phi) + F_{\mu\nu}F^{\mu\nu}$$

The term ‘$A_\mu A^\mu \phi^\dagger \phi = R^2 \phi^\dagger \phi$’ acts as though the photon has a mass. So, where spontaneous symmetry breaking exists, the photon acquires a mass. In this case, the Goldstone boson disappears, since the replacement ‘$\alpha \rightarrow \alpha + \theta$’ does not change the symmetry of the system, and therefore, adds no energy. Hence, the Goldstone boson is replaced by a photon with a non-zero mass. The process of creating a particle with mass through symmetry breaking is called the ‘Higgs mechanism’; the field ‘$\phi$’ is called the ‘Higgs field’ and while the analysis above was done entirely in the context of quantum electrodynamics, the process is quite general. In fact, the non-zero mass particles created are not photons with mass, but the $W^\pm, Z$ bosons of the weak interaction. Moreover, the Higgs mechanism gives all particles, at least all known particles, mass.

There are hypothetical particles that acquire mass without spontaneous symmetry breaking, but, so far, no particles of this kind have been found experimentally. Speculation is that these kinds of particles exist but are too massive to be detected by current particle accelerators. The conjecture is that these heavy particles are related to ‘dark matter’.

13.8.8.3 Fermion Mass

So far, the discussion has revolved around how bosons gain mass. How do fermions obtain mass? The Lagrangian for the photon is given in terms of $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ i.e.
\[ \mathcal{L} = F_{\mu\nu}F^{\mu\nu} \]

Since \( \mathcal{L} \) involves only gradient terms, the photon is massless. Now suppose a charged field \( '\phi = \rho e^{i\alpha} ' \) is added to \( \mathcal{L} \). The charge associated with \( \phi \) is not the charge on an electron, but the one associated with the \( W^\pm, Z \) bosons i.e. the charge associated with the weak interaction. Then the Lagrangian becomes

\[ \mathcal{L} = D_\mu \phi^\dagger D_\mu \phi - V(\phi^\dagger \phi) + F_{\mu\nu}F^{\mu\nu}, \]

where \( V \) is the potential associated with the Higgs field. Recall that \( \phi \) is a minimum when \( \rho = R \) (see fig. 13.8.5-1 b)), then

\[ D_\mu \phi = \partial_\mu \phi + iA_\mu \phi = i(\partial_\mu \alpha + A_\mu)Re^{i\alpha} \rightarrow D_\mu \phi^\dagger D_\mu \phi \]

\[ = -i(\partial_\mu \alpha + A_\mu)Re^{-i\alpha} i(\partial_\mu \alpha + A_\mu)Re^{i\alpha} = (\partial_\mu \alpha + A_\mu)^2 R^2 \]

Hence,

\[ \mathcal{L} = (\partial_\mu \alpha + A_\mu)^2 R^2 - V(R^2) + F_{\mu\nu}F^{\mu\nu} \]

Recall that \( A_\mu \rightarrow A_\mu + \partial \theta \) is a gauge transformation and does not impact the physics. Let \( A'_\mu = \partial_\mu \alpha + A_\mu \), then

\[ \mathcal{L} = (A'_\mu)^2 R^2 - V(R^2) + F_{\mu\nu}F^{\mu\nu} \]

Note that \( F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = \partial_\mu A'_\nu - \partial_\nu A'_\mu \). Therefore, the term \( '(A'_\mu)^2 R^2 ' \) in \( \mathcal{L} \) gives the photon a mass. Of course, this particle is not the photon, but the gauge bosons of the weak interaction. When a field shifts away from \( \phi = 0 \) to a minimum at \( R \), the symmetry is spontaneously broken, the condition that gives a particle mass. Interestingly, the field \( '\alpha ' \), associated with the Goldstone boson, that existed when \( R \) does not vary or varies only slightly, vanishes. So when the vector potential is introduced and the symmetry broken, the massless Goldstone boson is replaced by a particle with a mass.

If there is enough energy in the system, then \( \rho \) can vary significantly and \( \phi \) will oscillate around \( \rho = R \). This is also associated with a mass, the mass of the Higgs boson.

13.8.8.3.1 Spatial Parity and the Higgs

Prior to 1956, the neutrino and antineutrino were thought to be the same particle. Neutrinos carried no electric charge, possessed no mass and differed only in spin direction. At the time, it was believed the physical description of a system, whether processes were described by a right- or left-handed coordinate system, were identical. This was called the ‘law of Leibniz’, a universally accepted principle referred to as ‘spatial parity’, a symmetry of Nature. The neutrino and antineutrino seemed mirror images of one another, and hence, identical particles. The law of Leibniz holds for quantum electrodynamics and quantum chromodynamics. But T. D. Lee and C. N.
Yang noticed that many serious theoretical discrepancies disappeared if neutrinos and antineutrinos had different ‘handedness’, even though that would violate spatial parity.

Handedness is related to the spin of a particle. If an electron, for example, is moving and its spin points in the same direction, the electron is ‘left-handed’. If its spin is opposite its direction of motion, it is ‘right-handed’. Surprisingly, in the process of a neutron decaying into a proton, electron and an antineutrino, the electrons emitted were, in every case, left-handed. In fact, in all weak interactions, where a $W$ boson decays into an electron and a neutrino, the emitted electron is always left-handed. It is as if, relative to the $W$ boson (not the photon), the left-handed electron has charge, but the right-handed electron does not. Of course, all right-handed electrons have charge, but the process of weak decays proceeds as though the right-handed electron has no charge.

Consider a Dirac field ‘$\Psi$’ and the Dirac equation

$$i \left( \frac{\partial \Psi}{\partial t} + \alpha_j \frac{\partial \Psi}{\partial x^j} \right) = 0, \quad j = 1,2,3$$

Since the Dirac equation only involves gradient terms, the field ‘$\Psi$’ is associated with a massless particle. To give $\Psi$ a mass, the equation must be modified:

$$i \left( \frac{\partial \Psi}{\partial t} + \alpha_j \frac{\partial \Psi}{\partial x^j} \right) = m\beta \Psi, \quad j = 1,2,3$$

Recall that $\Psi$ is a four-vector. It has four components: ‘up-spin’, ‘down-spin’, ‘positive energy’ and ‘negative energy’. In QFT, the negative energy states are replaced by antiparticles, but the mathematics is the same. The spin direction can be the same as the direction of motion or in the opposite direction. These two cases can be designated ‘$\Psi_l$’ and ‘$\Psi_r$’. Hence, the Dirac equation can be written in a left-handed and right-handed form:

$$i \left( \frac{\partial \Psi_r}{\partial t} + \alpha_j \frac{\partial \Psi_r}{\partial x^j} \right) = m\beta \Psi_r = m\Psi_l, \quad i \left( \frac{\partial \Psi_l}{\partial t} - \alpha_j \frac{\partial \Psi_l}{\partial x^j} \right) = m\beta \Psi_l = m\Psi_r,$$

where ‘$\beta$’ is a matrix that reverses the direction of the field. In the equations above the left-handed field and right-handed field are dependent. When a field is associated with a mass, there is an interaction between the left- and right-handed fields.

Suppose that $\Psi$ represents the electron and, relative to the $W$ boson, $\Psi_r$ has no electric charge, but $\Psi_l$ carries charge. Therefore, in any transformation, $\Psi_r' \rightarrow \Psi_r$ and $\Psi_l' \rightarrow e^{i\theta} \Psi_l$. But if these substitutions are made in the massive Dirac equations, charge conservation is violated; a charge-less $\Psi_r$ could be turned into a charged $\Psi_l$.

Consider an additional charged boson field ‘$\phi$’. Let the gauge transformation properties of the three fields be $\Psi_r' \rightarrow \Psi_r$, $\Psi_l' \rightarrow e^{i\theta} \Psi_l$ and $\phi' = e^{i\theta} \phi$. Now suppose the massive Dirac equations are modified to read:
\[
i \left( \frac{\partial \Psi_r}{\partial t} + \alpha_j \frac{\partial \Psi_r}{\partial x^j} \right) = g \phi^\dagger \Psi_l, \quad \left( \frac{\partial \Psi_l}{\partial t} - \alpha_j \frac{\partial \Psi_l}{\partial x^j} \right) = g \phi \Psi_r,
\]

where \( m \) has been replaced by \( 'g \phi' \), \( g \) the 'coupling constant'. The equations above have \( U(1) \) symmetry. By introducing \( 'g \phi' \), charge conservation has been reestablished.

\[
\begin{align*}
\text{a) } & \quad \Psi_r & \phi \\
\text{b) } & \quad \Psi_l & \phi
\end{align*}
\]

Figure 13.8.8.3.1-1

From a Feynman diagram perspective, an uncharged field \('\Psi_r'\) can decay into a charged field \('\Psi_l'\) and a field \('\phi'\) of the opposite charge or a charged field \('\Psi_l'\) can decay into a uncharged field \('\Psi_r'\) and a field \('\phi'\) of the same charge (see fig. 13.8.8.3.1-1 a),b)).

Suppose that \( \phi = \rho e^{i\alpha} \), \( \phi \) a Higgs field. If the symmetry is spontaneously broken, then \( \phi = R \), the lowest energy state. Hence,

\[
\begin{align*}
i \left( \frac{\partial \Psi_r}{\partial t} + \alpha_j \frac{\partial \Psi_r}{\partial x^j} \right) &= gR \Psi_l, \\
i \left( \frac{\partial \Psi_l}{\partial t} - \alpha_j \frac{\partial \Psi_l}{\partial x^j} \right) &= gR \Psi_r, \quad j = 1,2,3
\end{align*}
\]

The coefficient \( 'gR' \) plays the same role as the mass in the Dirac equations. Note that \( \Psi_l \) and \( \Psi_r \) have the same mass. The upshot is that when a symmetry is spontaneously broken, the Higgs field gives the fermion a mass. Not only does the Higgs field give a fermion mass, but also the bosons including itself. The \('g'\) is related to the probability that the interaction takes place. There is a different \('g'\) for each fermion. The mass of the \( W \) boson is equal to \( R \).

It could be argued that the \( g' \)'s are simply found by experimentally knowing the mass of the associated fermion and there is really no theoretical basis for the values of the coupling constants. But the same coupling constants can produce a fermion and a Higgs boson:

\[
\begin{align*}
\text{Figure 13.8.8.3.1-2}
\end{align*}
\]
A left-handed fermion can decay into a right-handed fermion and a Higgs boson. In this case, the same coupling constant ‘$g$’ associated with the fermion mass is also associated with the fermion plus the Higgs mass i.e. $g R \Psi_l \rightarrow g (R + H) \Psi_l$ (see fig. 13.8.8.3.1-1). It might be possible to find a theoretical explanation for the $g$’s. Furthermore, all massive particles of the standard model acquire mass through the process of spontaneous symmetry breaking associated with the Higgs phenomenon. All the masses of the particles in the standard model are proportional to $R$.

13.8.8.4 Chirality and Helicity

Recall the Dirac equation

$$i \partial_t \Psi = -i \alpha_j \partial_j \Psi + m \beta \Psi \rightarrow (i \partial_t + i \alpha_j \partial_j - m \beta) \Psi = |0\rangle$$

Note that

$$(i \partial_t \Psi + i \alpha_j \partial_j \Psi - m \beta \Psi) \Psi^\dagger = i \left( \frac{\partial \Psi}{\partial t} \Psi^\dagger + \Psi^\dagger \alpha_j \partial_j \Psi \right) - m \Psi^\dagger \beta \Psi = |0\rangle \rightarrow L$$

$$= i \left( \frac{\partial \Psi}{\partial t} \Psi^\dagger + \Psi^\dagger \alpha_j \partial_j \Psi \right) - m \Psi^\dagger \beta \Psi$$

Let $\beta \Psi^\dagger = \overline{\Psi}$, then $\Psi^\dagger = \beta \overline{\Psi}$, since $\beta^2 = I$. Hence,

$$L = (i \partial_t \Psi + i \alpha_j \partial_j \Psi - m \beta \Psi) \Psi^\dagger = (i \partial_t \Psi + i \alpha_j \partial_j \Psi - m \beta \Psi) \beta \overline{\Psi}$$

$$= i (\beta \overline{\Psi} \partial_t \Psi + \beta \overline{\Psi} \alpha_j \partial_j \Psi) - \beta \overline{\Psi} m \beta \Psi = i (\gamma^0 \overline{\Psi} \partial_t \Psi + \gamma^j \overline{\Psi} \partial_j \Psi) - m \overline{\Psi} \Psi \rightarrow L$$

$$= \overline{\Psi} (i \gamma^0 \partial_t - m) \Psi, \quad \beta = \gamma^0, \quad \gamma^j = \beta \alpha_j$$

Let $\gamma^0 \gamma^1 \gamma^2 \gamma^3 = \gamma^5$, where $\gamma^5$ is associated with the handedness of a particle, referred to as ‘chirality’. The eigenvalues of $\gamma^5$ are $\pm 1$, referring to particles that are left- or right-handed. The spin of a particle can be used to define ‘handedness’ or ‘helicity’ of a particle. If the particle is massless, this is the same as chirality. A symmetry transformation between a left-handed and right-handed particle is called ‘parity’. Invariance under parity by a fermion is called ‘chiral symmetry’. An experiment on the weak decay of cobalt-60 nuclei carried out by Chien-Shiung Wu (May 31, 1912 – February 16, 1997) and collaborators in 1957 demonstrated that parity is not a symmetry of the Universe. It can be shown that

$$\psi \overline{\psi} = \psi^\dagger_l \psi_r + \psi^\dagger_r \psi_l$$

The term ‘$i \gamma^\mu \partial_\mu$’ in $L$ is a gradient term describing how a particle moves from one point to another, but the mass term involves $\Psi \overline{\Psi}$, which mixes up the right- and left-handed fields. It flips the chirality back and forth and this is synonymous with a particle having a mass. If $m = 0$, then the particle has a definite chirality i.e. it is a massless particle, where there is no oscillation in the particle’s chirality.
13.8.8.5 The Dynamics of the Higgs Boson

Note that if $\Psi'_r \to \Psi_r$ and $\Psi'_l = e^{i\theta} \Psi_l$, in other words, if $\Psi_l$ carries charge relative to the $W$ boson, but $\Psi_r$ does not, then

$$\Psi' \bar{\Psi}' = e^{-i\theta} \Psi'_l \Psi_r + \Psi'_r e^{i\theta} \Psi_l,$$

which violates charge conservation. To make $\Psi \bar{\Psi}$ a valid particle interaction, the Higgs ‘$\phi' = e^{i\theta} \phi$’ must be introduced such that

$$\Psi \bar{\Psi} = \Psi'_l \phi \Psi_r + \Psi'_r \phi \Psi_l \to \Psi' \bar{\Psi}' = e^{-i\theta} \Psi'_l e^{i\theta} \phi \Psi_r + \Psi'_r e^{i\theta} \Psi_l \phi e^{-i\theta} = \Psi \bar{\Psi}$$

Charge is then conserved. This shows the necessity of a Higgs boson, which gives a fermion a mass. The term ‘$\Psi'_l \phi \Psi_r$’ annihilates a right-handed particle and creates a left-handed particle and emits a Higgs boson:

But $H$ is associated with spontaneous symmetry breaking, since if there are oscillations around $R$, $\phi = R + H$. Hence,

$$\Psi \bar{\Psi} = (R + H)(\Psi_l^\dagger \Psi_r + \Psi_r^\dagger \Psi_l) \to g_y \Psi \bar{\Psi} = g_y R (\Psi'_l \Psi_r + \Psi'_r \Psi_l) + g_y H (\Psi'_l \Psi_r + \Psi'_r \Psi_l),$$

where $g_y$ is a coupling constant. The quantity ‘$R$’ is associated with the mass of the $W$ boson and $g_y$ is associated with the mass of the fermion. There is a different ‘$g_y$’ for every fermion. The term ‘$g_y R (\Psi'_l \Psi_r + \Psi'_r \Psi_l)$’ in a Lagrangian represents the annihilation of a right- or left-handed fermion and the creation of a fermion of opposite handedness and the emission of a $W$ boson; the term ‘$g_y H (\Psi'_l \Psi_r + \Psi'_r \Psi_l)$’ represents the annihilation of a right- or left-handed fermion and the creation of a fermion with opposite handedness and the emission of a Higgs boson.

The Higgs phenomenon is associated with spontaneous symmetry breaking (see fig. 13.8.5-1 b)). The shape of the curve in the figure very close to $\phi = 0$ is an upside down parabola whose equation is something like

$$V(\phi) = -\frac{\mu^2}{2} \phi^2$$

At $\phi = R$ the graph of the function begins to turn upward and a term must be added to reflect this i.e.
\[ V(\phi) = -\frac{\mu^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4, \]

where \( \lambda/4 \) is a number and \( \phi^4 \) is chosen so that \( V(\phi) \) remains symmetric. The potential \( V' \) is a minimum when

\[ \frac{dV}{d\phi} = -\mu^2 \phi + \lambda \phi^3 = 0 \rightarrow \lambda \phi^2 = \mu^2 \rightarrow \phi = \pm \frac{\mu}{\sqrt{\lambda}} = \pm R \]

It turns out that all the masses of massive particles in the standard model are proportional to \( R \). And all particles in the standard model gain their mass through spontaneous symmetry breaking. Why all the masses of the particles in the standard model are proportional to \( R \) remains a mystery.

### 13.9 The Relative Strength of the Interactions

The Coulomb law \( F \propto \frac{e^2}{r^2} \) governs the electric force between two charged particles in a vacuum, where the force is attractive if the charges are opposite and repulsive if the charges are the same. To write this in terms of potential energy, multiply by \( r \) i.e. \( E \propto \frac{e^2}{r} \). In a dielectric material, the charges are surrounded by opposite charges, creating a current which discharges the charge, making the material electrically neutral. Conceptually, the idea is called 'screening', where, for example, a negative charged cloud surrounds the positive charge, screening it from a distance negative charge. At this point,

\[ \lim_{r \to \infty} \frac{e^2}{r^2} = 0 \]

If \( r \ll 1 \), at some point, the two charges will be closer together than their respective surrounding charged clouds and the Coulomb law will be reestablished. The distance \( 'r_T' \) where this takes place depends on the material. The charge \( 'e^2' \) between two charged particles can be thought of as a function of \( r \) i.e. \( E \propto \frac{e^2}{r} \). In the case of the electromagnetic force, the charge gets weaker with increasing distance, but becomes stronger as the distance shortens.

The same kind of phenomenon occurs in QFT, where the ‘vacuum’ is conceived as equal amounts of positive and negative charges, making it electrically neutral. But if the vacuum is agitated by an electric field, then the charges are separated. The vacuum becomes a dielectric and electrons, for example, are screened slightly by a positive charged cloud. As the electrons get closer and closer to each other, the screening effect diminishes. What determines how close electrons can get to each other depends on the energy of the electrons. There is no set distance where the full Coulomb force is realized. The higher the energy of the dielectric pair, the closer they can get.

However, the strong force is quite different than the electromagnetic force. Because of the attractive force of the gluons, the strong force \( 'E_S' \) is weaker at short distances and stronger at longer distances, the opposite of the electromagnetic force i.e. \( E_S = \lambda r \). The
The force associated with weak interactions is somewhere in between (see fig. 13.9-1).

![Graph showing the force associated with strong, weak, and electromagnetic interactions.](image)

**Figure 13.9.7-1**

### 13.10 Concluding Remarks

The theoretical part of the standard model of particle physics includes three gauge theories, which describe the three fundamental forces of Nature – the ‘strong nuclear force’, the ‘weak nuclear force’ and the ‘electromagnetic force’. The standard model is not entirely theoretical, but is supplemented by parameter inputs retrieved from experiments and by additional mathematical modeling that are not part of the gauge theories. All three gauge theories are structurally the same. Particles interact through the exchange of gauge bosons, which decay into fermions. The gauge bosons are described mathematically by fields distinguished by the symmetry group that governs them - $SU(3)$ for the gluons, $SU(2)$ for the $W^-, W^+, Z$ bosons and $U(1)$ for the photon.

Classically, a field is described by a Lagrangian density function, its motion computed by the least action on the fields that make up the Lagrangian. In quantum mechanics, an entirely different question is asked. The fields in the Lagrangian are associated with a ‘probability amplitude’ that represents the probability that if a quantum system starts in a state $|x\rangle$, it will end up in a state $|y\rangle$.

The strong nuclear force is mediated by eight gauge gluons, whose sources are the quark fields. Quarks are fermions responsible for the hadronic particles. But a hadron could be made of three up-quarks ($uuu$), for example, and since hadrons are fermions, the quarks could be in the same quantum state, violating the Pauli exclusion principle. The three up-quarks are required to be in different quantum states. Hence, the theory introduces red-, blue- and green- and anti-red-, anti-blue- and anti-green-quarks, each representing a different quantum state. The Lagrangian of the strong nuclear force is governed by the $SU(3)$ symmetry group, which restricts the kinds of transformations that can take place in Nature. Red-quarks can be transformed into blue-quarks and blue- into green-quarks through the mediation of a gluon. In other words, the Lagrangian of the strong nuclear force remains invariant under $SU(3)$ symmetry transformations. This
is an exact symmetry, unlike the approximate flavor symmetry of the weak interaction [247].

The weak interaction is mediated by three gauge bosons ($W^-, W^+, Z$), whose source is the quark flavors. In the process of weak interactions, up-quarks are converted into down-quarks and vice versa, but the color of the quarks remains unchanged. The Lagrangian of the weak nuclear force is governed by the $SU(2)$ symmetry group, which restricts the kinds of transformations that can take place in Nature. In other words, the Lagrangian of the weak nuclear force remains invariant under $SU(2)$ symmetry transformations. This is an approximate symmetry, since the mass of, for example, the up-quark and the down-quark are not identical, but under the opuses of the symmetry, the difference in their masses is ignored.

The electromagnetic interaction is mediated by one gauge boson – the photon, whose source is the electromagnetic field. The Lagrangian associated with electromagnetic phenomenon is constrained by the $U(1)$ symmetry group, which restricts the kinds of transformations that can take place in Nature. The Lagrangian of the electromagnetic force remains invariant under $U(1)$ symmetry transformations. The conserved quantity is electric charge.

Physically, the three sectors ($QED, QCD, QFD$) are distinguished by the strength of the coupling constants associated with the forces within the sectors, which are proportional to the probability that the interaction takes place and the distance over which the force acts. Strong interactions happen with virtual certainty, but their range is $\approx 10^{-15} [m]$. Recall that in $QCD$, distance = time. Hence, the lifetimes of the gluons that mediate the force are extremely short. The electromagnetic force is much weaker than the strong force, $\approx 1/137$ that of the strong force. Its range is $\infty$, so the photon lives forever unless it decays into other particles.

The weak nuclear force is $\approx 10^{-6}$ weaker than the strong force. The probability that weak interactions occur in Nature is relatively low, not because the associated coupling constant is small, but because the mediating gauge bosons ($W^-, W^+, Z$) are relatively massive requiring a lot of energy to produce them. And, in accordance with the uncertainty principle, these gauge bosons are extremely short lived. The range of the force is even shorter than the strong force ($\approx 10^{-18} [m]$).

Finally, gravity is the weakest of all the forces. It is not part of the standard model of particle physics. The range of the gravitational force is $\infty$. There is a hypothetical gauge boson associated with the gravitational force – the 'graviton'. Whether or not the graviton actually exists in Nature is a matter of speculation. The strength of the gravitational force is $\approx 10^{-38}$ that of the strong force, which is also associated with the probability that two particles exchange a graviton (see table 13.10-1).
<table>
<thead>
<tr>
<th>Force</th>
<th>Strength of the Force</th>
<th>Range</th>
<th>Gauge Boson</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNF</td>
<td>≈ 1</td>
<td>≈ 10^{-15} [meters]</td>
<td>Gluon</td>
</tr>
<tr>
<td>EM</td>
<td>≈ 1/137</td>
<td>∞</td>
<td>Photon</td>
</tr>
<tr>
<td>WNF</td>
<td>≈ 10^{-6}</td>
<td>≈ 10^{-18} [meters]</td>
<td>W^-, W^+, Z</td>
</tr>
<tr>
<td>Gravity</td>
<td>≈ 10^{-38}</td>
<td>∞</td>
<td>(Graviton)</td>
</tr>
</tbody>
</table>

Table 13.10-1

There is an interesting fact associated with the coupling constants related to each of the three interactions within the standard model. If experiments are conducted at relatively long distances, then the coupling constants for each of the three interactions have very different values. As the distance gets smaller and smaller, the coupling constants of the three interactions almost, but not quite, merge, their values become almost equal. This forms the basis for grand unification theories.

A Grand Unified Theory (GUT) is one in which at high energy, the three gauge interactions of the standard model, the electromagnetic, weak, and strong interactions, merged into a single force. A GUT is characterized by one larger gauge symmetry ‘SU(5)’, where several force carriers have one coupling constant. There is the possibility of a grand unification epoch in the early Universe in which the fundamental forces were indistinct. The coming together of the coupling constants happens at a very high energy ≈ 10^{16} [GeV].

What about gravity? The non-relativistic gravitational force law is given by

\[ F = G \frac{m_1 m_2}{r^2} \rightarrow E = G \frac{m_1 m_2}{r} \]

Relativistically, if the masses are at rest, then \( E = mc^2 \rightarrow m = E/c^2 \). Hence,

\[ E = G \frac{E_1 E_2}{c^4 r} \]

If \( r \ll 1 \), the quantum theory says that \( p \), the momentum, is very uncertain i.e. \( p \) is very large. In fact, \( p \approx \hbar/r \rightarrow E \approx \hbar c/r \). Hence,

\[ E = G \frac{E_1 E_2}{c^4 r} \rightarrow E = G \frac{\hbar^2 c^2}{c^4 r^3} = G \frac{\hbar^2}{c^2 r^3} \]

Recall that the energy associated with the electromagnetic force is given by \( E \propto e^2/r \). If \( e \) is a dimensionless number, then the equation for the energy associated with the electromagnetic force must be modified to read ‘\( E = \hbar c/r \)’ to be dimensionally consistent, where \( e \approx 1 \). At what distance is the energy associated with the gravitational force equal to the energy associated with the electromagnetic force i.e. when is \( E_G \approx E_e \)? To find out, write

423
\[ E_G \approx E_e \rightarrow G \frac{\hbar^2 c^2}{c^4 r^3} \approx \frac{\hbar c}{r} \rightarrow r^2 \approx G \frac{\hbar}{c^3} \rightarrow r \approx \sqrt{G \frac{\hbar}{c^3}} \]

The number \( \sqrt{G\hbar/c^3} \) is called the ‘Planck length’. Its value is \( \approx 1.6 \times 10^{-33} \text{[m]} \). But a very small distance is associated with a very high energy:

\[ E = \frac{hc}{r} \approx \frac{10^{-34} \times 10^{11}}{10^{-33}} \approx 10^7 \text{Joules} \approx 10^{17} \text{[GeV]} \]

This becomes the motivation for unification. The current thinking is that the four forces of Nature unify at somewhere around the Planck length. The greatest puzzlement associated with this conjecture is not why the energy at unification is so high, but why is \( R \), the number associated with the symmetry breaking masses of all known particles, so small? The value of \( R \) is some fourteen orders of magnitude smaller than it should be at the scale where unification is thought to take place. This is called the ‘hierarchy problem’. There is no known theoretical calculation derived from \( QFT \) that produces a small \( R \).

There are several shortcomings in the standard model. The first is the issue of neutrino mass. When Pauli first hypothesized the existence of the neutrino, he thought that, since neutrinos have no electric charge, did not participate in the strong interaction, if they had a mass, it would never be detected. But within the standard model, decays involving neutrinos must take place if energy is to be conserved. However, the masses of all the neutrinos within the standard model are zero. According to the Higgs mechanism, particles in the vacuum acquire mass by colliding with the Higgs boson. Photons do not interact with the Higgs boson, and therefore, have no mass. All particles change handedness when they collide with the Higgs boson; left-handed particles become right-handed and vice versa. Experiments have shown that neutrinos are always left-handed. Moreover, the left-handed neutrino becomes a right-handed antineutrino through \( CP \)-conjugation i.e. \( \nu_e \leftrightarrow \bar{\nu}_e \rightarrow \nu_e \). There was no reason to introduce a right-handed neutrino into standard model. In fact, since right-handed neutrinos do not exist in the standard model, the theory predicts that neutrinos can never acquire mass.

However, it is believed that the Sun produces generous amounts of cosmic neutrinos. Around 1970, Ray Davis, a chemist, put 100,000 [gals] of cleaning fluid deep in a mine shaft, which over time, detected the neutron decay:

\[ \nu_e + n \rightarrow p + e \]

But, there were only about 1/3 the number of these decays as expected. Around 1990 decays involving neutrinos were observed:

\[ \pi^+ \rightarrow \mu^+ + \nu_e, \quad \mu^+ \rightarrow \bar{e} + \bar{\nu}_\mu + \nu_e \]
By 1998, $\nu_\mu$ atmospheric oscillations were confirmed. But, neutrino oscillations were paramount to saying, quantum mechanically, that the neutrino has a mass. The neutrino masses are extremely small, $< 1\, eV$. Moreover, according to Dirac’s theory, if a fermion gains mass, the chirality of the particle must be reversed. That meant there must be right-handed neutrinos.

Now that neutrinos appear to have mass, there are two issues: 1) how to overcome the contradiction between left-handedness and mass, 2) understand why the neutrino mass is so small compared with other particle masses — electrons being at least 500,000 times more massive than neutrinos.

There are two ways of extending the standard model thereby making neutrinos massive: 1) introduce new particles called ‘Dirac neutrinos’, 2) introduce a completely different type of particle called the ‘Majorana neutrino’. The Dirac neutrino theory is a simple idea with a serious flaw. According to this approach, the reason that right-handed neutrinos have escaped detection is that their interactions are at least 26 orders of magnitude weaker than ordinary neutrinos. Dirac neutrinos arise through the Higgs mechanism. However, the theory also suggests that neutrinos should have similar masses to the other particles in the standard model. To give neutrinos the smaller masses consistent with experiments, the strength of neutrino interactions with the Higgs boson must be at least 12 orders of magnitude weaker than that of the top quark. Few physicists accept such a tiny number as a fundamental constant of nature [249].

The second way of extending the standard model involves particles that are called ‘Majorana neutrinos’. This approach does not require the invocation of right-handed neutrinos with extremely weak interactions. The price paid is giving up the fundamental distinction between matter and antimatter, neutrinos and antineutrinos can be identical, because they have no electric charge. Massive neutrinos sit naturally within this framework. An observer travelling at the speed of light who overtakes a left-handed neutrino sees a right-handed neutrino. If neutrinos and antineutrinos are the same particle, the observer really sees a right-handed antineutrino. In this scheme, it is possible for right-handed neutrinos to have a mass of their own without relying on the Higgs boson. But all particles in the standard model gain mass through the Higgs mechanism. There are theories that postulate particles having masses that do not rely on the Higgs mechanism. But these particles are predicted so massive that current technologies are not able to detect them. Hence, these conjectures are considered highly speculative. Right-handed neutrinos must, therefore, be very heavy, as predicted by grand unified theories that aim at combining electromagnetism with the strong and weak interactions [249].

The standard model makes good predictions down to a distance of about $10^{-16} \, [cm]$ and is expected to hold up to about an energy of $10^{12} \, [eV] \approx 1 \, [TeV]$. Within these bounds, it is a fully consistent renormalizable field theory. Beyond these levels, some of its predictions are clearly wrong. The standard model does not provide an explanation for dark matter. There are many theories of dark phenomena involving speculative ideas like supersymmetry or the theory of axions, but current theory, including the theory of relativity, provides no universally accepted candidates for dark phenomena.
There is no explanation in the standard model for the predominance of matter over antimatter. The model is complicated and fine-tuned. It has 20 or so unexplained parameters that require input into the model before it can make predictions. The unexplained parameters include the masses of the particles. All the sectors of the standard model are gauge theories. But each gauge theory has unique characteristics that are absent in the others. For instance, why are the gauge bosons of the $QED$ and $QCD$ sectors massless, while the gauge bosons of the $QFD$ sector are massive? Why do gluons directly interact with each other while photons do not? Why are there three families of leptons and quarks and where do their masses come from? At lower energy levels the mass of the Higgs boson is $\approx 125 \, [GeV]$, but at higher energies the mass of the Higgs must be fine-tuned. If the standard model is to make good predictions up to energies at or near the Planck scale ($10^{18} \, [GeV]$), the corrections to the Higgs mass would be on the order of $\approx 10^{34}$. This is why most people believe that the standard model cannot extend to the Planck scale.
Chapter 14
String Theory

“I just think too many nice things have happened in string theory for it to be all wrong. Humans do not understand it very well, but I just don’t believe there is a big cosmic conspiracy that created this incredible thing that has nothing to do with the real world.”

- Edward Witten

“To the extent that we even understand string theory, it may imply a massive number of possible different universes with different laws of physics in each universe, and there may be no way of distinguishing between them or saying why the laws of physics are the way they are. And if I can predict anything, then I haven’t explained anything.”

- Lawrence M. Krauss

14.0 Introduction

The previous chapters dealt with physics validated extensively through observation and experimentation. In this context, it would be inappropriate to discuss speculative ideas, such as string theory. But string theory is so widely believed linked to the next advancements in physics, it seems fitting to discuss it, if only to clarify the myriad of misconceptions of what string theory currently represents. The two quotations above exemplify the wide range of perspectives that string theory engenders in the people who study it. I should point out that my treatment of the subject follows closely the lectures of Prof. Leonard Susskind of Stanford University.

String theory was originally developed as a potential theory of hadrons. At the time, the theory of quarks was in its infancy and the theory of gluons did not yet exist. Hadrons are the particles that make up the nucleus of an atom. In order that the nucleus did not fly apart, there had to be some kind of force counteracting the repulsive electrostatic force between the protons in the nucleus. Not much was known about the force, other than it behaved something like a rubber band, which, when stretched, eventually breaks. The force got stronger as the distance between protons grew, but its range was limited to roughly the diameter of an atom.

14.1 Regge Trajectories and Fundamental Particles

Hadrons come in families of particles, but when their angular momentum is graphed against the square of the mass of each hadron, the plot, called a ‘Chew-Frautschi plot’, is a family of parallel lines. The lines were called ‘Regge trajectories’, named for Italian theoretical physicist Tullio Eugenio Regge (July 11, 1931 – October 23, 2014) (see fig. 14.1-1). It was hard to imagine that the plots were coincidental. In all of the particle families, it took about the same amount of energy to increase the angular momentum by one unit. The spin of an electron, for example, cannot be increased in this way and the
electron was a fundamental particle. The message that finally came out was that hadrons were likely not elementary, but composite particles.

14.2 The Concept of a ‘String’

Experimentally, two $\pi$-mesons could annihilate creating $\rho$-mesons and then recombine into two $\pi$-mesons ($\pi$-meson scattering). But according to $QFT$, if $\pi$-meson scattering was possible, then so was the exchange of a $\rho$-meson between two $\pi$-mesons. At the time, that two of one kind of particle could exchange a particle of the same kind violated fundamental principles of quantum physics. So, physicists started drawing diagrams like the one shown in fig. 14.2-1.

If a horizontal line divides the picture, two quark/antiquark pairs produce two quark/antiquark pairs, much like, for example, $\pi$-meson scattering. If, on the other hand, a vertical line divides the picture, it might represent a quark/antiquark pair exchanging a particle with another quark/antiquark pair, which suggests a force carrying particle. But there must have been something holding the quarks and the antiquarks together. In this case, the ‘something’ was a tiny one dimensional object i.e. a little string (see fig. 14.2-1). These ideas led, more or less, to the concept of a ‘string’. Moreover, if strings held quarks and antiquarks together, then the quark/antiquark pairs could rotate around one another, which could not happen if there was only one particle.

String theory was eventually replaced by the theory of quantum chromodynamics ($QCD$). While string theory held promise, it did not quite describe the physics of hadrons.
correctly. The primary issue was, regardless of how it was formulated, every version of string theory contained a massless spin-2 particle. But there were no massless spin-2 particles in hadron physics. John Swartz and Joel Scherk, a French theoretical physicist, realized that string theory was better suited to a theory of quantum gravity and that the massless spin-2 particle was likely a ‘graviton-like’ object, the particle thought responsible for carrying the gravitational force.

14.3 Open Strings

In classical physics, the energy ‘\( E \)’ of a particle is given by the equation ‘\( E = \frac{p^2}{2m} + B \)’, where \( p \) is the momentum, \( m \), the mass and \( B \) is the internal energy, which arises because the particle exists. Note that \( B \) cannot depend on \( p \). If the system is relativistic, then \( B = mc^2 \), where \( c \) is the speed of light. If the system is made from a multitude of particles, then

\[
E = \sum_i \frac{p_i^2}{2m_i} + B_i, \quad i = 1, ..., n,
\]

where the total energy is the sum of the energies associated with all \( n \) particles. If the system is relativistic, then

\[
E = \sum_i \sqrt{p_i^2 c^2 + m_i^2 c^4}
\]

If \( p \) is small, the system is non-relativistic. As an approximation,

\[
E = \sum_i \sqrt{p_i^2 c^2 + m_i^2 c^4} = \sum_i m_i c^2 \sqrt{1 + \frac{p_i^2 c^2}{m_i^2 c^4}}
\]

Note that \( \frac{p_i^2 c^2}{m_i^2 c^4} \) is an extremely small number and recall that

\[
\sqrt{1 + \varepsilon} \approx 1 + \varepsilon/2 + \cdots,
\]

\[
\varepsilon \ll 1 \rightarrow m_i c^2 \sqrt{1 + \frac{p_i^2 c^2}{2m_i^2 c^4}} \approx m_i c^2 + m_i c^2 \frac{p_i^2 c^2}{2m_i^2 c^4} \approx m_i c^2 + \frac{p_i^2}{2m_i},
\]

\[
\varepsilon = \frac{p_i^2 c^2}{m_i^2 c^4}
\]

which is the energy of a non-relativistic system for the \( i^{th} \) particle.
If $p$ is large along a given axis, say the $z$-axis, then

$$E = \sum_i \sqrt{p_{iz}^2 + p_{iy}^2 + p_{ix}^2 + m_i^2}, \quad c = 1 \rightarrow E = \sum_i p_{iz} \sqrt{1 + \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{p_{iz}^2}},$$

$$\frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{p_{iz}^2} \ll 1$$

So,

$$p_{iz} \sqrt{1 + \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{p_{iz}^2}} \approx p_{iz} \left(1 + \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}^2}\right) = p_{iz} \left(\frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}}\right) \rightarrow E$$

$$\sum_i p_{iz} + \sum_i \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}}, \quad \epsilon = \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{p_{iz}^2}$$

Since energy is conserved, $\sum_i p_{iz}$ is conserved i.e. is a constant. Hence,

$$E - \sum_i p_{iz} \approx \sum_i \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}} \rightarrow \sum_i \left(\frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}}\right) = \sum_i \frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}} + \frac{m_i^2}{2p_{iz}} \rightarrow H$$

$$\sum_i \left(\frac{p_{iy}^2 + p_{ix}^2 + m_i^2}{2p_{iz}}\right)$$

is small, if $H$ is associated with the quantum mechanical definition of the Hamiltonian, then $H = \hat{h} \frac{\partial}{\partial t}$, which implies that changes in the system take place very slowly. This is an example of time dilation, where a clock moving rapidly in the $z$-direction, measured by someone at rest, would appear to run slow. Measuring the change in the position of a particle in a direction perpendicular to the $z$-direction would be non-relativistic. The only requirement is that $m_i^2/2p_{iz}$ be independent of $p_{iy}^2 + p_{ix}^2$. Note that the internal energy is proportional to $m^2$.

A relativistic string can be described non-relativistically as an idealized rubber band, cut in half and free to move, wiggle and stretch in a two dimensional plane perpendicular to its direction of motion.

**14.3.1 What is a Particle?**

What is the relationship between a string and particle? The modern view contends that there is no such thing as a point particle. Even an elementary particle, such as an electron, teems with activity, swarmed by a cloud of virtual photons. The photon can
decay into an electron and a positron. In string theory, particles are associated with string vibrations. The rationale for this is somewhat arbitrary and practical rather than theoretically supportable. But if the energy spectrum of the string is sufficiently distinct, then the string behaves like a particle.

It is virtually impossible to excite an electron to a heavier mass. A proton can be excited to a heavier mass, but the energy required to produce each mass level is relatively distinct. A cup of coffee, cooled to its lowest energy state, would require a vanishingly small amount of energy to raise its mass level. From a string theory standpoint, a cup of coffee is not a particle, even though it has a center of mass and a position in space. Its mass levels are not sufficiently distinct to be described by a string.

Philosophically, this approach to defining “a particle” is a bit unconvincing. At what point do the excitation levels become close enough together that the physics of a string can no longer be described as particle-like? It is a valid question, but the current definition will have to do, since the strings discussed here are considered particle-like.

14.3.2 The Mathematics of an Open String

A string can be thought of as a collection of point masses held together by tiny springs. Consider a function \( x(\sigma) \) that can be approximated by a series of \( N \) discrete points (see fig. 14.3.2.3-1). Let \( \sigma \) run from 0 to \( \pi \) and

\[
\Delta x_i = x_i - x_{i-1} \approx \frac{\partial x}{\partial \sigma} \Delta \sigma, \quad \Delta \sigma = \frac{\pi}{N}
\]

Note that

\[
\lim_{n \to \infty} \sum_{i=1}^{N} x_i \Delta \sigma = \frac{1}{\pi} \int_{0}^{\pi} x(\sigma) \, d\sigma,
\]

if the limit exists.

The function above can be Fourier analyzed so long as certain boundary conditions are met.

14.3.2.1 The Dirichlet and Neumann Boundary Conditions

Since the function has only one parameter, the boundaries are the endpoints. There are two possible boundary conditions: ‘Dirichlet’ and ‘Neumann’. Dirichlet boundary conditions require the endpoints to be fixed i.e. \( x(0) = x(\pi) = 0 \). Neumann boundary conditions require that \( \frac{\partial x(0)}{\partial \sigma} , \frac{\partial x(\pi)}{\partial \sigma} = 0 \). In other words, the derivative at the endpoints must be a relative maximum or minimum.
14.3.2.2 Fourier Analysis of an Open String

Both conditions, whether Dirichlet or Neumann, can be Fourier analyzed as follows: suppose

\[ x(\sigma) = \sum_{n=1}^{\infty} X_n \sin n\sigma \text{ (Dirichlet)}, \quad x(\sigma) = \sum_{n=0}^{\infty} X_n \cos n\sigma \text{ (Neumann)}, \]

then

\[ \int_{0}^{\pi} \cos n\sigma \cos m\sigma \, d\sigma = \begin{cases} \frac{\pi}{2}, & m, n \neq 0 \\ \pi, & m = n = 0 \end{cases} \]

14.3.2.3 The Energy of an Open String

Hooke’s law for a harmonic oscillator i.e. the energy ‘E’ associated with a simple harmonic oscillator is given by

\[ E = m \frac{\dot{x}^2}{2} + \frac{\kappa}{2} x^2 \rightarrow E = \frac{\dot{x}^2}{2} + \frac{\omega^2}{2} x^2, \quad \omega = \sqrt{\kappa}, \quad m = 1 \]

where \( \kappa \) is the spring constant.

Consider a series of point masses held together by little springs (see fig. 14.3.2.3-1). The total energy is given by

\[ E = \sum_{i} m_i \frac{\dot{x}_i^2}{2} + \frac{km_i(x_{i+1} - x_i)^2}{2}, \]

where \( k \) is the spring constant.

If an open string is approximated by a set of \( N \) masses and \( N - 1 \) little springs between the masses (see fig. 14.3.2.3-1), an approximation to the energy ‘\( E \)’ of an open string is
\[ E \approx \sum_{i=1}^{N} \mu \left( \frac{\dot{x}_i^2 + \dot{y}_i^2}{2} \right) + \frac{k}{2} \sum_{i=2}^{N} \Delta x_i^2 + \Delta y_i^2, \]

where \( k \) is the spring constant. The mass of the string remains fixed. If \( \mu = 1/N \), then the spring constant 'k' must become large as \( N \) gets large, since it takes more force to pull masses apart that are close together than if they are far apart. Let \( k = N/\pi^2, \pi^2 \) being somewhat arbitrary. In the limit, the energy 'E' of the string becomes

\[
\lim_{k \to \infty, N \to \infty} E = H = \frac{1}{2\pi} \int_0^{\pi} \dot{x}^2(\sigma) + \dot{y}^2(\sigma) + \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \sigma} \right)^2 \, d\sigma \rightarrow \mathcal{L}
\]

The lower limit '0' and upper limit '\pi' of integration are conveniently chosen to represent the energy of an open string. The string is parameterized by '\sigma' and '\tau', where '\tau' represents the proper time. The equation above gives the energy of a vibrating string. The relative motion of the parts of the string, vibrations and stretching, are considered internal motions of the string as opposed to the displacement of its center of mass. Note that \( H \) is a simple wave field in the form of a Lagrangian.

### 14.3.2.4 The Force on the Ends of an Open String

Because the string is open, there is a different force on the ends than on the internal parts of the string. For instance, if the string is approximated by \( N \) small masses held together by tiny springs, then the force on the \( N^{th} \) mass can be estimated by Newton's law. The force 'F' on the \( N^{th} \) particle is proportional to \( k\Delta x \) i.e.

\[
F \propto \frac{N}{\pi^2} \Delta x \approx \frac{N}{\pi^2} \Delta \sigma = \frac{N}{\pi^2} \frac{\partial x}{\partial \sigma} N = \frac{1}{\pi} \frac{\partial x}{\partial \sigma} \rightarrow F \propto \frac{\partial x}{\partial \sigma} = \mu \ddot{x} = \frac{N}{\pi} \ddot{x} 
\]

As \( N \to \infty \), the acceleration becomes infinite, which does not happen. To resolve this, the Neumann boundary conditions are imposed at the endpoints of the string i.e.

\[
\partial x/\partial \sigma, \partial y/\partial \sigma = 0.
\]

### 14.3.2.5 The Energy of an Open String - Boosted and Non-Boosted

Note that the internal energy of the string is proportional to the mass squared i.e.

\[
H = \frac{1}{2\pi} \int_0^{\pi} \left( \dot{x}^2(\sigma) + \dot{y}^2(\sigma) + \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \sigma} \right)^2 \right) \, d\sigma \propto m^2,
\]

since the equation above describes a string whose center of mass is at rest in the \( x,y \)-plane. If the string is not vibrating i.e. \( \dot{x}^2(\sigma) + \dot{y}^2(\sigma) = 0 \), then

\[
H = \frac{1}{2\pi} \int_0^{\pi} \left( \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \sigma} \right)^2 \right) \, d\sigma \propto m^2,
\]

433
It turns out that $\partial x / \partial \sigma \propto L$, where $L$ is the length of the string. This makes the energy of the string proportional to its length i.e. $(\partial x / \partial \sigma)^2 + (\partial y / \partial \sigma)^2 \propto L^2 \propto m^2 \rightarrow L \propto m$. If the string has no kinetic energy, then the energy of the string in a boosted frame of reference is proportional to $m^2$, but in the non-boosted frame, the energy of the string is proportional to $m$. In the non-boosted environment, the string does not follow Hooke’s law. This was one feature that made string theory a potential theory of hadrons, since, experimentally, the forces between hadrons were proportional to the linear distance between quark/antiquark pairs.

Experiments had suggested that hadrons were composite particles, with masses proportional to the length of something. That something was eventually hypothesized to be the string. There were two types of strings, both emulating from the same energy source i.e. $E = \sqrt{p^2 + m^2}$. A small $p$ and large $m$ represented a non-relativistic string in the rest frame, whose energy was proportional to its length. If $p$ was large in a single direction and $m$ small, this also represented a non-relativistic string, but in one less dimension, where, the energy was proportional to $m^2$.

### 14.3.3 Quantizing the Open String

If a wave equation can be written for an open string, then it can be quantized. To begin, let

$$x(\sigma, \tau) = \sum_{n=0}^{\infty} X_n \cos n\sigma, \quad y(\sigma, \tau) = \sum_{n=0}^{\infty} Y_n \cos n\sigma,$$

where $X_n$ and $Y_n$ are functions of $\tau$. Next, plug the expressions for $x(\sigma, \tau)$ and $y(\sigma, \tau)$ into the Lagrangian

$$\mathcal{L} = \frac{1}{2\pi} \int_0^\pi \left( \left( \frac{\partial x}{\partial \tau} \right)^2 + \left( \frac{\partial y}{\partial \tau} \right)^2 \right) - \left( \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \sigma} \right)^2 \right) \ d\sigma \rightarrow \frac{\partial x}{\partial \tau} = \dot{x}(\sigma) = \sum_{n=0}^{\infty} \dot{X}_n \cos n\sigma \rightarrow \left( \frac{\partial x}{\partial \tau} \right)^2 = \dot{X}_n^2 \cos n\sigma \rightarrow \sum_{n=0}^{\infty} \dot{X}_n^2 \cos n\sigma \rightarrow \dot{x}^2 = \dot{X}_n^2 \cos n\sigma \rightarrow \sum_{n=0}^{\infty} \dot{X}_n^2 \cos n\sigma$$

$$= \frac{1}{2\pi} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \dot{X}_n \dot{X}_m \int_0^\pi \cos n\sigma \cos m\sigma \ d\sigma = \frac{1}{2\pi} \sum_{n=0}^{\infty} \dot{X}_n^2 \left\{ \begin{array}{ll} 0, & m \neq n \\ \frac{\pi}{2}, & m = n \neq 0 \\ \pi, & m = n = 0 \end{array} \right.$$ 

Therefore,

$$\dot{x}^2 + \dot{y}^2 = \left( \frac{\partial x}{\partial \tau} \right)^2 + \left( \frac{\partial y}{\partial \tau} \right)^2 = \frac{\dot{X}_0^2}{2} + \frac{\dot{Y}_0^2}{2} + \frac{1}{4} \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2$$
\[ \frac{\partial x}{\partial \sigma} = -X_n \sum_{n=1}^{\infty} n \sin n\sigma \rightarrow (\frac{\partial x}{\partial \sigma})^2 = \frac{1}{2\pi} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} mnX_mX_n \int_0^\pi \sin n\sigma \sin m\sigma \ d\sigma \]

\[ = \frac{1}{2\pi} \sum_{n=1}^{\infty} n^2X_n^2 \left\{ \begin{array}{ll} 0, & \text{if } m \neq n \\ \pi, & \text{if } m = n \end{array} \right\} \]

\[ \rightarrow \left( \frac{\partial x}{\partial \sigma} \right)^2 = \frac{1}{4} \sum_{n=1}^{\infty} n^2X_n^2, \]

\[ \left( \frac{\partial y}{\partial \sigma} \right)^2 = \frac{1}{4} \sum_{n=1}^{\infty} n^2Y_n^2 \]

Evidently then

\[ L = \frac{\dot{X}_0^2}{2} + \frac{\dot{Y}_0^2}{2} + \frac{1}{4} \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2 - \frac{1}{4} \sum_{n=1}^{\infty} n^2(X_n^2 + Y_n^2) \]

The Lagrangian of an open string is an infinite number of harmonic oscillators, where \( \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2 \) is proportional to the kinetic energy, \( \sum_{n=1}^{\infty} n^2(X_n^2 + Y_n^2) \) is proportional to the restoring force for each oscillator and \( \dot{X}_0^2 + \dot{Y}_0^2 \) is proportional to the movement of the mass of the string. The frequency \( \omega_n \) of each oscillator is equal to \( n \). The terms

\[ \frac{1}{4} \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2 - \frac{1}{4} \sum_{n=1}^{\infty} n^2(X_n^2 + Y_n^2) \]

are associated with the internal energy of the string, which is proportional to the square of the mass of the string.

In terms of hadron physics, open strings have a quark and an antiquark at the opposite ends of the string. Two strings can come together, where the antiquark at one end can annihilate the quark at one end of the other string, forming a longer string. But also, one end of a string can join its other end, forming a closed string. There are string theories that contain only closed strings, but there are no string theories that contain only open strings. This was the first clue that string theory was not a good candidate for describing hadron physics, which could not be modelled by a closed string. It was later realized that closed strings corresponded to a graviton-like object.

### 14.3.3.1 Creation and Annihilation Operators for an Open String

Consider the Lagrangian \( 'L' \) of an open string:

\[ L = \frac{\dot{X}_0^2}{2} + \frac{\dot{Y}_0^2}{2} + \frac{1}{4} \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2 - \frac{1}{4} \sum_{n=1}^{\infty} n^2X_n^2 + n^2Y_n^2 \rightarrow H \]

\[ = \frac{\dot{X}_0^2}{2} + \frac{\dot{Y}_0^2}{2} + \frac{1}{4} \sum_{n=1}^{\infty} \dot{X}_n^2 + \dot{Y}_n^2 + \frac{1}{4} \sum_{n=1}^{\infty} n^2X_n^2 + n^2Y_n^2 \]
For simplicity, consider the $i^{th}$ term in one dimension and ignore the center of mass terms, then

$$H_i = \frac{\dot{X}_i^2}{4} + \frac{i^2 X_i^2}{4}$$

Quantum mechanically, $\partial H / \partial \dot{X} = \dot{p} = \dot{X}/2$. Hence,

$$\hat{H}_j = \hat{p}_j^2 + \frac{j^2 \hat{x}_j^2}{4} = \left(\frac{j \hat{x}_j}{2} + i \hat{p}_{x_j}\right) \left(\frac{j \hat{x}_j}{2} - i \hat{p}_{x_j}\right),$$

where $\hat{p}_{x_j}$ and $\hat{x}_j$ are now operators. The candidates for the creation `\(\hat{a}^+\)` and the annihilation `\(\hat{a}^-\)` operators are

$$\hat{a}^-_j = \left(\frac{j \hat{x}_j}{2} + i \hat{p}_{x_j}\right), \quad \hat{a}^+_j = \left(\frac{j \hat{x}_j}{2} - i \hat{p}_{x_j}\right)$$

Remembering that $[\hat{a}^-, \hat{a}^+] = 1$ and $[\hat{x}_j, \hat{p}_{x_j}] = i$, $\hbar = 1$, then

$$[\hat{a}^-, \hat{a}^+_j] = \left(\frac{j \hat{x}_j}{2} + i \hat{p}_{x_j}\right) \left(\frac{j \hat{x}_j}{2} - i \hat{p}_{x_j}\right) - \left(\frac{j \hat{x}_j}{2} - i \hat{p}_{x_j}\right) \left(\frac{j \hat{x}_j}{2} + i \hat{p}_{x_j}\right) = \frac{j}{2} (-i)(i) + \frac{j}{2} (-i)(i) = j$$

However,

$$\left[\frac{1}{\sqrt{j}} \left(\frac{j \hat{x}_j}{2} + i \hat{p}_{x_j}\right), \frac{1}{\sqrt{j}} \left(\frac{j \hat{x}_j}{2} - i \hat{p}_{x_j}\right)\right] = 1 \rightarrow \hat{a}^-_j + \hat{a}^+_j = \sqrt{j} \hat{x}_j \rightarrow \hat{x}_j = \frac{\hat{a}^-_j + \hat{a}^+_j}{\sqrt{j}}$$

If a system’s momentum is boosted in the $z$-direction, then there are two dimensions in which the string oscillates non-relativistically. So,

$$\hat{b}^-_j = \left(\frac{\sqrt{j} \hat{y}_j}{2} + i \hat{b}_{y_j}\right), \quad \hat{b}^+_j = \left(\frac{\sqrt{j} \hat{y}_j}{2} + \frac{i}{\sqrt{j}} \hat{b}_{y_j}\right) \rightarrow [\hat{b}^-_j, \hat{b}^+_j] = 1,$$

which represents creation and annihilation operators in the $y$-direction.

Recall that

$$\hat{x}(\sigma) = \sum_{j=0}^{\infty} \hat{x}_j \cos j \sigma, \quad \hat{y}(\sigma) = \sum_{j=0}^{\infty} \hat{y}_j \cos j \sigma \rightarrow \hat{x}(\sigma) = \sum_{j=0}^{\infty} \hat{a}^-_j + \hat{a}^+_j \cos j \sigma,$$

$$\hat{y}(\sigma) = \sum_{j=0}^{\infty} \frac{\hat{b}^-_j + \hat{b}^+_j}{\sqrt{j}} \cos j \sigma$$
Hence, the parameterization of an open string can be expressed in terms of creation and annihilation operators. There is a slight mathematical problem here. At \( j = 0 \), \( \hat{x}(\sigma) \) as well as \( \hat{y}(\sigma) \) are undefined. Evidently, this is overcome by parameterizing on \( \sigma \) i.e.

\[
\hat{x}(0) = \sum_{j=1}^{\infty} \frac{\hat{a}_{j} - \hat{a}_{j}^{+}}{\sqrt{j}} \cos 0 = \sum_{j=1}^{\infty} \frac{\hat{a}_{j} - \hat{a}_{j}^{+}}{\sqrt{j}} \rightarrow \frac{\partial \hat{x}}{\partial \sigma} = - \sum_{j=1}^{\infty} \frac{\hat{a}_{j} - \hat{a}_{j}^{+}}{\sqrt{j}} \sin j \sigma \rightarrow \frac{\partial \hat{x}(0)}{\partial \sigma} = - \sum_{j=1}^{\infty} \frac{\hat{a}_{j} - \hat{a}_{j}^{+}}{\sqrt{j}} \sin 0 = 0
\]

14.3.3.2 The Spin of a Massless Particle

Recall that particles have \( 2j + 1, \ j \in N \cup \{0\} \) spin states. A spin-0 particle has one state, a spin-1/2 particle has two states, a spin-1 particle three states and so on. This rule applies only to massive particles. Massless particles have only two spin states.

All massive particles have a ‘rest mass’. Hypothetically, a massive spin-1 particle can be brought to rest, its spin direction rotated perpendicular to the direction of motion, regain its velocity while maintaining the new spin direction. It, therefore, has three spin states \( (1,0,-1) \). But massless particles cannot be brought to rest. The spin direction of a massless particle cannot be rotated, and hence, only two spin directions are possible, spin along the direction of motion or in the opposite direction.

14.3.3.2.1 Spin Direction and Polarization

The spin of a massless particle is synonymous with the polarization of light. Linear polarization is a superposition of states between vertical and horizontal polarization. Linearly polarized light is defined by the plane in which the electric field is oscillating. If the electric field moves in the \( y, z \)-plane, then the magnetic field moves in the \( x, y \)-plane. If a photon is traversing along the \( z \)-axis, then there are two polarized states: \( |y\rangle \) (vertical), \( |x\rangle \) (horizontal).

Light can be circularly polarized, where the light beam rotates clockwise or counterclockwise. The two polarization states are

\[
|r\rangle = |x\rangle + i|y\rangle, \quad |l\rangle = |x\rangle - i|y\rangle
\]

In other words, the polarization of a photon occurs in a plane perpendicular to its direction of motion.

Consider the ground state \( |g\rangle \) of a string that is not vibrating. Applying the annihilation operators \( \hat{a}_{j}^{-}, \hat{b}_{j}^{-} \) to \( |g\rangle \) gives \( \hat{a}_{j}^{-}|g\rangle = |0\rangle \) and \( \hat{b}_{j}^{-}|g\rangle = |0\rangle \). In other words, the annihilation operators, applied to the ground state of the string, annihilate the string. Since the energy ‘\( E \)’ of the string is proportional to its mass squared i.e. \( E \propto m^2 \), suppose that the energy ‘\( E_{|g\rangle} \)’ of the ground state is given by \( m_0^2 \) i.e. \( E_{|g\rangle} = m_0^2 \). If \( \hat{a}_{j}^{+}|g\rangle = |g + 1\rangle \) or \( \hat{b}_{j}^{+}|g\rangle = |g + 1\rangle \), whatever the state \( |g + 1\rangle \) is, its energy ‘\( E_{|g + 1\rangle} \)’ is increased by one unit i.e. \( E_{|g + 1\rangle} = m_0^2 + 1 \). Moreover, in accordance with quantum
mechanics, any linear combination of \( \hat{a}_1^+ |g\rangle \) and \( \hat{b}_1^+ |g\rangle \) will give an energy level of \( m_0^2 + 1 \). Since the \( \hat{a}_j^+ \)'s are functions of \( x \) and the \( \hat{b}_j^+ \)'s are functions of \( y \), the \( \hat{a}_j^+ \)'s and \( \hat{b}_j^+ \)'s act like vectors in the \( x,y \)-plane in terms of rotations. In fact, the \( a_j^+ \)'s and \( b_j^+ \)'s act analogously to linearly polarized light. Moreover, \( (\hat{a}_j^+ \pm i \hat{b}_j^+) |g\rangle \) acts analogously to circularly polarized light. There are two states: \( (\hat{a}_j^+ + i \hat{b}_j^+) |g\rangle \) and \( (\hat{a}_j^+ - i \hat{b}_j^+) |g\rangle \). Since there are only two states, these objects are massless, identified as photon-like objects within string theory. But if energy is associated with spin direction, the spin can be in the \( z \)-direction or in the \( -z \)-direction, but no spin, and hence, no energy, in the \( x \)- or \( y \)-direction. So,

\[
E_{|g+1\rangle} = 0 \rightarrow m_0^2 = -1
\]

### 14.3.3.3 The Tachyon

This is disastrous. Masses are not negative. In particular, the square of a mass is certainly not negative. If \( m_0^2 = -1 \), then \( m_0 = \pm i \). In other words, \( m_0 \) is an imaginary mass. There is actually a string (particle) with an imaginary mass, called a ‘tachyon’ within string theory, a hypothetical particle that travels faster than light. According to relativistic physics, faster-than-light particles cannot exist. If such particles did exist, they could be used to send signals faster than light, and this, according to special relativity, would violate causality. To see why tachyons travel faster than light, recall the kinetic energy ‘\( K.E. \)’ of a non-relativistic particle:

\[
K.E. = \frac{p^2}{2m} \rightarrow \frac{\partial (K.E.)}{\partial p} = \frac{p}{m} = v,
\]

where \( v \) is the velocity of the particle. For a relativistic particle

\[
E = \sqrt{p^2 + m^2} \rightarrow \frac{\partial E}{\partial p} = \frac{p}{\sqrt{p^2 + m^2}} < 1,
\]

if \( m^2 > 0 \), the velocity of a massive particle cannot exceed the speed of light. But if \( m^2 < 0 \), then

\[
\frac{\partial E}{\partial p} = \frac{p}{\sqrt{p^2 - m^2}} > 1,
\]

the velocity of a particle with a negative mass must exceed the speed of light.

Moreover, consider a simple wave equation

\[
\frac{\partial^2 \phi}{\partial t^2} = \frac{\partial^2 \phi}{\partial x^2}
\]

Recall that \( \partial^2 \phi / \partial t^2 \) is associated with \(-\omega^2\), where \( \omega \) is the frequency of the wave and \( \partial^2 \phi / \partial x^2 \) is associated with \(-k^2\), where \( k \) is the momentum of the wave. Hence,
\[-\omega^2 = -k^2 \rightarrow \frac{\omega}{k} = 1,\]

which shows the wave travels at the speed of light. To produce a wave that travels at less than the speed of light, add a mass term:

\[
\frac{\partial^2 \phi}{\partial t^2} = \frac{\partial^2 \phi}{\partial x^2} - \frac{m^2}{2} \phi
\]

The energy ‘H’ associated with this wave is

\[
H \approx \frac{1}{2} \left( \frac{\partial \phi}{\partial t} \right)^2 + \frac{1}{2} \left( \frac{\partial \phi}{\partial x} \right)^2 + \frac{m^2}{2} \phi^2, \quad V(\phi) = \frac{m^2}{2} \phi^2,
\]

With changes in \( \phi \), the potential ‘V’ remains positive. The locus of \( V \) is a parabola with the lowest energy level at the base, the point around which a particle oscillates. This describes a perfectly legitimate physical system. But if \( m^2 < 0 \), then the locus of \( V \) is an upside-down parabola and the system would become unstable, its lowest energy level becoming \(-\infty\). The upshot is that systems containing tachyons become unstable, and hence, highly undesirable. The Universe is simply not unstable in the way tachyon theories suggest. There are string theories that contain tachyons and some that do not, but purging the tachyon from string theory requires supersymmetry. If the issues with the tachyon are ignored, string theory produces photon-like objects. At the time, however, string theory was a candidate for a theory of hadrons, but there were no massless hadrons. Attempts were made to expel the photon-like objects from string theory, but all efforts failed.

### 14.4 Closed Strings

So far the discussion has focused on freely moving open strings propagating in space-time. However, like particles, strings can interact. For instance, the end of one string can join with the end of another open string to form a new string. And, in fact, this type of interaction is synonymous with two particles annihilating and creating a third particle (see figure 14.4-1). There are diagrams in string theory analogous to the Feynman diagrams found in QFT.

![Figure 14.4-1](image-url)
14.4.1 String Coupling Constant

Like in QFT, string theory requires that if a process is possible, its reverse is possible. A string can break into two strings or two strings can join to make one string. Strings do not continually join or break apart. What determines when and if strings break apart or join is the ‘string coupling constant’. It is associated with the probability that two strings will join or one string will break creating two strings. If the coupling constant was large, strings would continually break and join, creating an unstable environment. So, one of the assumptions of string theory is that the string coupling constant is relatively small. A small coupling constant is associated with a small probability, making the joining or breaking of strings a relatively rare occurrence. The string, then, can be considered a somewhat stable object.

An open string can join with the end of another open string or join its ends. A string with joined ends is called a ‘closed string’. Any string theory containing interacting open strings will contain closed strings. There are exclusively closed string theories, but there are no exclusively open string theories. All interacting open string theories include closed strings. Since closed strings are associated with the graviton, string theory predicts gravity. Once it was realized that string theory contained graviton-like objects, it became clear that string theory went well beyond Newton’s or Einstein’s theories, which simply described gravity. String theory predicts it!

14.4.2 The Mathematics of Closed Strings

Like open strings, closed strings are described by a parameter ‘σ’, which runs from 0 to $2\pi$. The parameter ‘σ’ is intrinsic to the string, but not to space-time. Waves can move along the string. The direction of the waves is described in terms of σ (see figure 14.4.2-1).

![Figure 14.4.2-1](image)

A wave can move in the direction of increasing or decreasing σ. These waves are referred to as a ‘right-moving’ or ‘left-moving’ waves. The string can be parameterized in space-time by designating functions ‘$x(\sigma, \tau)$’ and ‘$y(\sigma, \tau)$’ or, if there are more than two, adding more dimensions. The constraints on $x(\sigma, \tau)$ and $y(\sigma, \tau)$ are $x(0, \tau) =$
The general form of a function ‘$x(\sigma, \tau)$’ or ‘$y(\sigma, \tau)$’ with period ‘$2\pi$’ is given by

$$x(\sigma, \tau) = \sum_j X_j e^{ij\sigma},$$

$$y(\sigma, \tau) = \sum_j Y_j e^{ij\sigma} \rightarrow x(\sigma, \tau)$$

**Right Moving Wave**

$$\sum_{j>0} X_j e^{ij\sigma} \quad \text{Center of Mass}$$

$$\sum_{j=0} X_0$$

**Left Moving Wave**

$$\sum_{j>0} X_{-j} e^{-ij\sigma}$$

$$\sum_{j=0} X_0$$

where the first term represents a right moving wave, the second, a left moving wave, and the third the center of mass of the string. The frequency of each mode of oscillation is given by $j$.

### 14.4.2.1 The Lagrangian for a Closed String

The next step is to write the Lagrangian

$$\mathcal{L} = \int_0^{2\pi} \left( \frac{\partial x}{\partial \tau} \right)^2 + \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \tau} \right)^2 - \left( \frac{\partial y}{\partial \sigma} \right)^2 \, d\sigma \rightarrow H$$

$$= \int_0^{2\pi} \left( \frac{\partial x}{\partial \tau} \right)^2 + \left( \frac{\partial x}{\partial \sigma} \right)^2 + \left( \frac{\partial y}{\partial \tau} \right)^2 + \left( \frac{\partial y}{\partial \sigma} \right)^2 \, d\sigma$$

$$= \int_0^{2\pi} \frac{1}{2} \left( \frac{\partial x}{\partial \tau} + \frac{\partial x}{\partial \sigma} \right)^2 + \frac{1}{2} \left( \frac{\partial y}{\partial \tau} - \frac{\partial x}{\partial \sigma} \right)^2 + \frac{1}{2} \left( \frac{\partial y}{\partial \tau} + \frac{\partial y}{\partial \sigma} \right)^2 + \frac{1}{2} \left( \frac{\partial y}{\partial \tau} - \frac{\partial y}{\partial \sigma} \right)^2 \, d\sigma$$

There are creation operators ‘$\hat{a}_j^+$’, ‘$\hat{a}_{-j}^+$’, ‘$\hat{a}_j^-$’ creates a wave of frequency ‘$j$’, $\sigma$ positive and in the $x$-direction and $\hat{a}_{-j}^+$ creates a wave of frequency ‘$-j$’, $\sigma$ negative and in the $x$-direction. Similarly, there are creation operators ‘$\hat{b}_j^+$’, ‘$\hat{b}_{-j}^+$’ in the $y$-direction along with four complimentary annihilation operators. This is essentially the whole structure of closed strings.

### 14.4.2.2 The Energy in a Closed String

Suppose the string is in the ground state ‘$|g\rangle$’. Let the energy ‘$E_{|g\rangle} = m_0^2$’. Hypothetically, there are four distinct states that give the least energy to the system above the ground state: ‘$\hat{a}_{i,j}^+ |g\rangle$, $\hat{b}_{i,j}^+ |g\rangle$’ or a linear combinations thereof. To what do these states correspond? There is no way of producing the state ‘$E = 0$’, using the given operators. It is possible to get $m_0^2 + 1$ or $m_0^2 - 1$, but no other state. Hence, whatever
particle this string represents, it must be a massless particle i.e. \( m = 0 \), since it only has two states.

Another requirement for establishing legitimate states for a closed string is called ‘level matching’, which requires that the energy in a left-moving wave must match the energy in a right-moving wave along the string. For example, \( \hat{a}^+_1 \langle g \rangle \) is not a legitimate state and neither is \( \hat{a}^+_1 \langle g \rangle \), since the energy in the right- and left-moving waves do not match. But, \( \hat{a}^+_1 \hat{a}^+_1 \langle g \rangle \) and \( \hat{b}^+_1 \hat{b}^+_1 \langle g \rangle \) are legitimate states, both states having two units of energy.

A circular polarized wave can have four states:

\[
\begin{align*}
(\hat{a}^+_1 + i\hat{b}^+_1)(\hat{a}^-_1 + i\hat{b}^-_1)\langle g \rangle & \rightarrow m = 2 \\
(\hat{a}^+_1 - i\hat{b}^+_1)(\hat{a}^-_1 - i\hat{b}^-_1)\langle g \rangle & \rightarrow m = -2 \\
(\hat{a}^+_1 + i\hat{b}^+_1)(\hat{a}^-_1 - i\hat{b}^-_1)\langle g \rangle & \rightarrow m = 0 \\
(\hat{a}^+_1 - i\hat{b}^+_1)(\hat{a}^-_1 + i\hat{b}^-_1)\langle g \rangle & \rightarrow m = 0,
\end{align*}
\]

where \( m \) is the angular momentum of the wave. The first two states are analogous to a spin-2 massless particle, a candidate for the graviton. The last two states represent massless spin-0 particles, one called the ‘dilaton’ and the other the ‘axion’. While both are legitimate states within string theory, neither the dilaton nor the axion have been discovered experimentally. There are, however, string theories that exclude the dilaton and the axion.

14.4.2.3 Level Matching

Why is there a requirement that the energy in all right-moving waves match the energy in all left-moving waves along a string? Consider a point on a closed string labelled ‘\( \sigma = 0 \)’. Since the string is closed, \( \sigma = 0 \) can represent any point on the string. If the string is conceived as a discrete set of points, then each point can be labelled ‘\( \langle x_i, y_i \rangle \)’. The wave function could be given by

\[
\Psi((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))
\]

If \( (x_i, y_i) \rightarrow (x_{i+1}, y_{i+1}) \), then

\[
\Psi((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \rightarrow \Psi((x_2, y_2), \ldots, (x_n, y_n), (x_1, y_1)) \rightarrow \lim_{n \to \infty} \Psi
\]

\[
= \Psi(x(\sigma, \tau), y(\sigma, \tau)) = \Psi(x(\sigma + \varepsilon, \tau), y(\sigma + \varepsilon, \tau)),
\]

\[
\varepsilon \ll 1 \rightarrow \Psi(x(\sigma, \tau), y(\sigma, \tau)) - \Psi(x(\sigma + \varepsilon, \tau), y(\sigma + \varepsilon, \tau)) = 0
\]

\[
\rightarrow \int \frac{\partial \Psi}{\partial x(\sigma)} \frac{\partial x(\sigma)}{\partial \sigma} + \frac{\partial \Psi}{\partial y(\sigma)} \frac{\partial y(\sigma)}{\partial \sigma} d\sigma = 0
\]

What does \( \partial \Psi / \partial x \) represent? Recall from quantum mechanics that \( \hat{p} = -i \partial / \partial q \), \( \hbar = 1 \rightarrow -i \partial \Psi / \partial x = \hat{p} \Psi \), then
\[
\int \hat{p}_x(\sigma) \frac{\partial x(\sigma)}{\partial \sigma}, \quad \hat{p}_y(\sigma) \frac{\partial y(\sigma)}{\partial \sigma} \, d\sigma = \int \dot{x} \frac{\partial x(\sigma)}{\partial \sigma}, \dot{y} \frac{\partial y(\sigma)}{\partial \sigma} \, d\sigma = \int \frac{\partial x}{\partial \tau} \frac{\partial x(\sigma)}{\partial \sigma} - \frac{\partial y}{\partial \tau} \frac{\partial y(\sigma)}{\partial \sigma} \, d\sigma
\]

If
\[
H' = \int_0^{2\pi} \frac{1}{2} \left( \frac{\partial x}{\partial \tau} + \frac{\partial x}{\partial \sigma} \right)^2 - \frac{1}{2} \left( \frac{\partial x}{\partial \tau} - \frac{\partial x}{\partial \sigma} \right)^2 + \frac{1}{2} \left( \frac{\partial y}{\partial \tau} + \frac{\partial y}{\partial \sigma} \right)^2 - \frac{1}{2} \left( \frac{\partial y}{\partial \tau} - \frac{\partial y}{\partial \sigma} \right)^2 \, d\sigma
\]

and noting that
\[
\frac{1}{2} \left( \frac{\partial x}{\partial \tau} + \frac{\partial x}{\partial \sigma} \right)^2 - \frac{1}{2} \left( \frac{\partial x}{\partial \tau} - \frac{\partial x}{\partial \sigma} \right)^2 = \frac{1}{2} \left[ \left( \frac{\partial x}{\partial \tau} \right)^2 + 2 \left( \frac{\partial x}{\partial \tau} \right) \left( \frac{\partial x}{\partial \sigma} \right) + \left( \frac{\partial x}{\partial \sigma} \right)^2 \right] - \frac{1}{2} \left[ \left( \frac{\partial x}{\partial \tau} \right)^2 - 2 \left( \frac{\partial x}{\partial \tau} \right) \left( \frac{\partial x}{\partial \sigma} \right) + \left( \frac{\partial x}{\partial \sigma} \right)^2 \right]
\]
\[
= 2 \left( \frac{\partial x}{\partial \tau} \right) \left( \frac{\partial x}{\partial \sigma} \right).
\]

then the necessary and sufficient condition that \( \sigma = 0 \) not be a preferred point is
\[
\int \frac{\partial x}{\partial \tau} \frac{\partial x(\sigma)}{\partial \sigma} \, d\sigma = 0, \quad \int \frac{\partial y}{\partial \tau} \frac{\partial y(\sigma)}{\partial \sigma} \, d\sigma = 0
\]
i.e. the right-moving energy and the left-moving energy of the wave on the string should match. The level matching requirement eliminates many possible string states that do not have corresponding particles in Nature, making string theory better suited for a theory of particle physics.

**14.5 String Tension**

If a string is stretched to a length ‘\( L \)’, then the tension in the string is proportional to \( L \). A stretched string acts somewhat like a spring and by Hooke’s law
\[
P.E. = \frac{k}{2} L^2
\]
i.e. the potential energy is proportional to the square of the length of the string. But in string theory, \( E \propto m^2 \), where \( m \) is the mass of the string. Hence,
\[
P.E. = m^2 = \frac{k}{2} L^2 \rightarrow m = \sqrt{\frac{k}{2} L}
\]
The mass of the string is proportional to its length. The factor ‘\( \sqrt{k/2} \)’ is called the ‘string tension’ and is often designated ‘\( T \)’ i.e. \( m = TL \). The dimensions associated with \( T \) are ‘[energy per unit length]’. Since \( E = mc^2 \), \( c = 1 \rightarrow E = m \), the dimensions in string theory are associated with the string tension ‘\( T \)’. If \( c = \hbar = 1 \), \( E \) has units [1/L], which
implies that $T$ has units $[1/L^2]$. The amount of energy required to excite the string to a higher energy state is proportional to $T$. The string tension in fundamental strings is extremely high, somewhere near the Plank scale. The oscillations of these fundamental strings are extremely rapid.

### 14.5.1 The Units Used in String Theory

String theorists prefer Planck scale dimensions. It makes developing the physics easier. There are three fundamental universal constants of Nature, 'c', the speed of light, 'h', Planck’s constant and 'G', the universal gravitational constant. String theorists work in units where $c = h = G = 1$. The reason for this is that strings operate on or near the Planck scale.

There is a way of writing the three fundamental dimensions: 'L' (length), 'M' (mass) and 'T' (time) in terms of the three fundamental constants ($c, h, G$). This is accomplished through dimensional analysis. For example,

\[
[L^2] = \left[ \frac{L^2}{T^2} \right], \quad [c] = \left[ \frac{L}{T} \right], \quad [h] = \left[ \frac{L^2 M}{T} \right], \\
[G] = \left[ \frac{L^3}{M T^2} \right] \rightarrow [L^2] = \left[ \left( \frac{L^3}{M T^2} \right)^p \left( \frac{L^2 M}{T} \right)^q \left( \frac{L}{T} \right)^r \right],
\]

where the square brackets indicate that units are being equated, not numerical values, and the exponents 'p, q, r' are the powers of the units.

Since there are no mass units in $L^2$, $p = q$:

\[
[L^2] = \left[ \left( \frac{L^3}{M T^2} \right)^p \left( \frac{L^2 M}{T} \right)^q \left( \frac{L}{T} \right)^r \right] = \left[ \left( \frac{L^3}{M T^2} \right)^p \left( \frac{L^2 M}{T} \right)^q \left( \frac{L}{T} \right)^r \right] \rightarrow r = -3p
\]

Setting $p = 1$,

\[
L^2 = G h c^{-3} \rightarrow L_p = \sqrt{G h c^{-3}} \approx 10^{-35} [meters]
\]

Similarly, the Planck time 'T_p' is the time it takes a light ray to travel the Planck length i.e.

\[
T_p \approx 10^{-43} [sec] \rightarrow M_p = \sqrt{\frac{h c}{G}} \approx 10^{-8} [Kilograms],
\]

where $M_p$ is the 'Planck mass'. This is the scale on which fundamental strings are thought to exist. To put things in perspective, the wavelengths of fundamental strings are around the Planck length. The frequency, the rate at which the strings vibrate, is about $10^{43} [cycles \ per \ sec]$ and the energy to excite an electron, if an electron can be thought of as a string, is about a Planck mass. Note that 1 [Planck mass] $\approx 10^{19} [GeV]$. 

---

444
14.6 Why Does String Theory have 25 Spatial Dimensions?

The energy of the ground state of an open string is proportional to $m^2$, the mass of the string squared. To produce a photon-like object, $m^2$ is set to $-1$, so that when a creation operator acts on the ground state, it produces a massless string. Where does the ‘$-1$’ come from?

Strings have ground state energies that are the sum total of all the harmonic oscillators that make up the string. The ground state energies are all positive and depend on the frequency of the oscillator i.e.

$$E = \frac{1}{2} \hbar \omega = \frac{1}{2} \omega, \quad \hbar = 1 \rightarrow E = \frac{n}{2}, \quad \omega = n,$$

where $n$ corresponds to the $n^{th}$ oscillator. To get the total energy $E_T$ of the string:

$$E_T = \frac{1}{2} \sum_{n=1}^{\infty} n$$

This is not a particularly encouraging result, since $\sum_{n=1}^{\infty} n = \infty$, not $-1$. Recall that a string with a very high momentum ‘$p$’ in a direction is given by

$$E = p + \frac{m^2}{2p} \rightarrow 2pE = 2p^2 + m^2$$

If $p$ grows infinitely large, the equation above acquires an infinite piece ‘$2p^2$’ and a finite piece ‘$m^2$’. The approach to overcoming the problem of infinite energies uses what some would regard as the “worse aspects of the process of renormalization” to eliminate the infinite piece leaving just the finite piece. To begin, the sum ‘$\sum_{n=1}^{\infty} n$’ is replaced by

$$e^{-\varepsilon} + 2e^{-2\varepsilon} + 3e^{-3\varepsilon} + \cdots = \sum_{n=1}^{\infty} ne^{-n\varepsilon}$$

This is justified by noting that if $\varepsilon \rightarrow 0$,

$$\sum_{n=1}^{\infty} ne^{-n\varepsilon} \rightarrow \sum_{n=1}^{\infty} n$$

But, note that $\sum_{n=1}^{\infty} ne^{-n\varepsilon}$ converges if $\varepsilon \neq 0$, while $\sum_{n=1}^{\infty} n$ does not. The process involves adding the series up, let $\varepsilon \rightarrow 0$ and see what happens. Note that

$$\sum_{n=1}^{\infty} ne^{-n\varepsilon} = -\frac{\partial}{\partial \varepsilon} \sum_{n=1}^{\infty} e^{-n\varepsilon} = -\frac{\partial}{\partial \varepsilon} e^{-\varepsilon}(1 + e^{-\varepsilon} + e^{-2\varepsilon} + \cdots) = -\frac{\partial}{\partial \varepsilon} \left( \frac{e^{-\varepsilon}}{1 - e^{-\varepsilon}} \right),$$

since $(1 + e^{-\varepsilon} + e^{-2\varepsilon} + \cdots)$ is a geometric series. Now
The first term in the final equation on the right goes to infinity as \( \varepsilon \to 0 \). As the renormalization argument goes, this term can be absorbed into the total energy of the system. Hence,

\[
\frac{e^{-\varepsilon}}{1 - e^{-\varepsilon}} = \frac{1 - \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3!} + \cdots}{1 - (1 - \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3!} + \cdots)} = \frac{1}{\varepsilon} \left( \frac{1 - \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3!} + \cdots}{1 - \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3!} + \cdots} \right)
\]

\[
= \frac{1}{\varepsilon} \left( 1 - \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3!} + \cdots \right) \left( 1 + \varepsilon + \frac{\varepsilon^2}{12} - \cdots \right) = \frac{1}{\varepsilon} \left( 1 - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{12} - \cdots \right)
\]

\[
\rightarrow - \frac{\partial}{\partial \varepsilon} \left( 1 - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{12} \right) = \frac{1}{\varepsilon^2} \left( 1 - \frac{1}{12} \right) = \frac{1}{2\varepsilon^2} - \frac{1}{24}
\]

But \( -1/24 \neq -1 \), which is the ground state of the string. The only available explanation is that the ground state energy of the string is \( -1/24 \), but that there are 24 dimensions in which the string vibrates. The energy of the ground state becomes \( -d/24 \), where \( d = 24 \). Adding in the dimension in which the string travels plus the time dimension gives a total of 26 dimensions. Hence, bosonic string theory has 26 dimensions, 25 spatial dimensions and one time dimension.

14.7 The Veneziano Amplitude

Gabriele Veneziano first formulated the rudiments of string theory in 1968 when he discovered that the Euler ‘Beta function’, interpreted as a scattering amplitude, explained the physical properties of strongly interacting particles. The ‘Veneziano amplitude’ is interpreted as the scattering amplitude of four open string tachyons. In retrospect, his work is considered the founding of string theory, although, at the time, there was no hint that it might lead to a theory of quantum gravity. The rise of quantum chromodynamics as an explanation of strong interactions temporarily diluted interest in string theory, but by the 1980’s, it had realized a revival. In 1991, Veneziano published a paper showing how an inflationary cosmological model can be obtained from string theory, opening the door to a description of pre-big bang scenarios.

About the only way of experimentally studying the properties of elementary particles is through scattering them and studying particle decays. The life of a particle physicist boils down to finding ways of computing the scattering amplitudes of a system of particles. When particles interact, something happens; the particles reappear, but not necessarily in the same numbers. The post interacting particles carry various characteristics: spin, energy, momentum and many others.
14.7.1 Mandelstam Variables

Suppose the only property of interest is momentum. Momentum is conserved (see fig. 14.7.1-1). Each particle has a four-momentum i.e.

\[
(E, p_x, p_y, p_z) = k_\mu \rightarrow E^2 = p_i^2 + m^2 \rightarrow p_i^2 = -m^2 \rightarrow k_\mu^2 = k^2
\]

\[
t = -m^2, \quad c = h = 1, \quad i = 1, ..., 3, \quad \mu = 0, ..., 3
\]

To simplify matters, suppose two particles with momentum ‘\(k_{\mu_1}, k_{\mu_2}\)’ interact to produce two new particles with momentum ‘\(q_{\mu_3}, q_{\mu_4}\)’. If momentum and energy are conserved, then

\[
k_{\mu_1} + k_{\mu_2} = q_{\mu_3} + q_{\mu_4}
\]

In other words, the four-momentums of the outgoing particles must be equal the four-momentums of the incoming particles. Now let \(q_{\mu_3} = -k_{\mu_3}\) and \(q_{\mu_4} = -k_{\mu_4}\), then

\[
k_{\mu_1} + k_{\mu_2} + k_{\mu_3} + k_{\mu_4} = 0
\]

![Diagram](https://via.placeholder.com/150)

Figure 14.7.1-1

The scattering amplitude ‘\(A(k_{\mu_1}, k_{\mu_2}, k_{\mu_3}, k_{\mu_4})\)’ is a function of the \(k_{\mu_i}\)’s, which are not independent. For one thing, \(k_{\mu_i}^2 = -m_i^2\). Moreover, it is always possible to formulate the problem where the center of mass is at rest. In this frame, the total spatial momentum is zero and the spatial momentum of each particle is equal, but opposite. Just the energy ‘\(E_{cm}\)’, the energy of the center of mass of the particles requires computation. Additionally, the system can be rotated so its center of mass is moving along a given axis, say the \(x\)-axis. Since momentum and energy are conserved, the only thing that can differ between the incoming and outgoing particles is the scattering angle ‘\(\theta\)’. Hence, the scattering amplitude ‘\(A\)’, instead of depending on 16 independent variables, depends only on \(E_{cm}\) and \(\theta\).

Consider

\[
(k_{\mu_1} + k_{\mu_2})^2 = (k_{\mu_{i1}} + k_{\mu_{i2}})^2 - (k_{\mu_{01}} + k_{\mu_{02}})^2, \quad i = 1, ..., 3
\]

The quantity ‘\((k_{\mu_{i1}} + k_{\mu_{i2}})^2 - (k_{\mu_{01}} + k_{\mu_{02}})^2\)’, is relativistically invariant. In the center of mass frame \(k_{\mu_{i1}} + k_{\mu_{i2}} = 0\), since the two quantities are equal but opposite.
Also, $k_{\mu_01} = k_{\mu_02}$ in the center of mass frame. Hence,

$$(k_{\mu_01} + k_{\mu_02})^2 = (2k_{\mu_01})^2 \rightarrow (2k_{\mu_01})^2 = -S \rightarrow E_{cm}^2 = S,$$

where $S$ is called the ‘center of mass energy’ squared. Note that

$$(k_{\mu_3} + k_{\mu_4})^2 = (k_{\mu_1} + k_{\mu_2})^2$$

What about $(k_{\mu_1} + k_{\mu_3})^2$? In the center of mass frame $k_{\mu_01} + k_{\mu_03} = 0$. So by direct calculation

$$(k_{\mu_1} + k_{\mu_3})^2 = 2(E_{cm}^2 - m^2)(1 - \cos \theta) = -T,$$

where $T$ is Lorentz invariant. Note that

$$(k_{\mu_1} + k_{\mu_3})^2 = (k_{\mu_2} + k_{\mu_4})^2 = -T$$

And finally,

$$(k_{\mu_1} + k_{\mu_4})^2 = -U$$

The quantities ‘$S, T, U$’ are called ‘Mandelstam variables’, which have some interesting relationships. For one,

$$S + T + U = m_1^2 + m_2^2 + m_3^2 + m_4^2$$

To see this,

$$-S + (-T) + (-U) = (k_{\mu_1} + k_{\mu_2})^2 + (k_{\mu_1} + k_{\mu_3})^2 + (k_{\mu_1} + k_{\mu_4})^2$$

$$= 3k_{\mu_1}^2 + k_{\mu_2}^2 + k_{\mu_3}^2 + k_{\mu_4}^2 + 2k_{\mu_1}k_{\mu_2} + 2k_{\mu_1}k_{\mu_3} + 2k_{\mu_1}k_{\mu_4}$$

$$= k_{\mu_1}^2 + k_{\mu_2}^2 + k_{\mu_3}^2 + k_{\mu_4}^2 + 2k_{\mu_1}(k_{\mu_2} + k_{\mu_3} + k_{\mu_4})$$

$$= k_{\mu_1}^2 + k_{\mu_2}^2 + k_{\mu_3}^2 + k_{\mu_4}^2 = m_1^2 + m_2^2 + m_3^2 + m_4^2$$

Figure 14.7.1-2
14.7.2 S and T Channel Amplitudes

The two diagrams in fig. 14.7.1-a, b) show that the amplitude \( A_a \) for the system in fig. 14.7.1-a) is

\[
A_a = g^2 \frac{1}{S - M^2},
\]

where \( g^2 \) is the coupling constant and \( 1/(S - M^2) \) is the propagator for this interaction. The amplitudes of type \( A_a \) are called 'S-channel' amplitudes. The amplitude \( A_b \) for the system shown in fig. 14.7.1-b) is

\[
A_b = g^2 \frac{1}{T - M^2}
\]

Note that the diagram in fig. 14.7.1-b) is the same as the diagram in fig. 14.7.1-a) except tilted on its side. The amplitudes of the type \( A_b \) are called 'T-channel' amplitudes. Moreover, if \( S \) is interchanged with \( T \), \( A_a + A_b \) remains invariant.

The diagrams of the type shown in fig. 14.7.1-2 are not particularly useful for describing meson scattering. There are so many possible configurations, the problem becomes intractable. But in the process of attempting to count all the possible combinations, a remarkable discovery was made, based on the gamma function:

\[
\Gamma(x) = \int_0^\infty t^{x-1}e^{-t} \, dt
\]

The integral is absolutely convergent for \( x \geq 1 \). But for \( x < 1 \), the integrand becomes infinitely large as \( t \to 0 \) through positive values. In fact, the integral is defined for all \( x > 0 \), since \( \lim_{t \to 0^+} \int_0^\infty t^{x-1}e^{-t} \, dt \) exists because \( \forall x(x > 0 \to t^{x-1}e^{-t} \leq t^{x-1}) \). Note that

\[
\Gamma(x + 1) = \int_0^\infty t^x e^{-t} \, dt = x\Gamma(x), \quad x > 0
\]

To see this, integrate by parts: \( u = t^x \) and \( dv = e^{-t} \, dt \), then

\[
\int u \, dv = uv - \int v \, du \to \int t^x e^{-t} \, dt = -t^x e^{-t} + \int xt^{x-1}e^{-t} \, dt \to \int t^x e^{-t} \, dt
\]

\[
= -t^x e^{-t} \bigg|_0^\infty + \int xt^{x-1}e^{-t} \, dt \to \int t^x e^{-t} \, dt = \int xt^{x-1}e^{-t} \, dt \to \Gamma(x + 1)
\]

\[
= x\Gamma(x), \quad x > 0
\]

If \( x \in \mathbb{N} \), then \( \Gamma(x + 1) = x! \). The gamma function provides a way of giving meaning to the factorial of any positive real number.

What Veneziano discovered was an amplitude \( A_V \) that describes the scattering of all \( S \) and \( T \)-channel amplitudes for all possible particle interactions:
\[ A_V = g^2 \frac{\Gamma(-S) \Gamma(-T)}{\Gamma(-S - T)} \]

The Veneziano amplitude was a kind of sum over Feynman diagrams that make up all possible particle interactions. This was something new. But what kind of physics gave rise to the Veneziano amplitude? The answer was the ‘string’.

14.7.3 The Physics of the Veneziano Amplitude

When an open string is propagating in space-time, it draws out a worldsheet as opposed to a worldline (see fig. 14.7.3-1).

![Diagram of worldsheet](image)

Figure 14.7.3-1

Rather than a single variable ‘\( \tau \)', a string has two parameters ‘\( \tau, \sigma \)', where \( 0 \leq \tau \leq \infty \) and \( 0 \leq \sigma \leq \pi \). The wave equation that describes the oscillations of the string is

\[ \frac{\partial^2 x_\mu}{\partial \tau^2} - \frac{\partial^2 x_\mu}{\partial x_i} = 0, \]

where \( x_\mu \) is the space-time point of a discrete particle that makes up the string. Each of the tiny masses that make up the string can be thought of as lying along infinitesimal worldlines. The wave function ‘\( \Psi \)' for the string can be parameterized by

\[ \Psi(x_{\mu_1}, x_{\mu_2}, ..., x_{\mu_n}) \]

14.7.3.1 The Breaking and Joining of Strings

Consider two open strings

\[ \Psi_1(x_{\mu_1}, x_{\mu_2}, ..., x_{\mu_n}), \quad \Psi_2(x_{\mu(n+1)}, x_{\mu(n+2)}, ..., x_{\mu2n}), \]

where \( \Psi_1 \) is the wave function for the first and \( \Psi_2 \) the wave function for the second string. There is a probability that the two strings join to form one string. The string persists for a time until it separates again becoming two strings.
The state of a quantum system with momentum \( k_1 \) is given by
\[
\Psi_1(x_{\mu 1}, x_{\mu 2}, \ldots, x_{\mu n}) = e^{\frac{i k_1}{\hbar} (x_{\mu 1} + x_{\mu 2} + \ldots + x_{\mu n})} \Psi_{01}(x_{\mu 1}, x_{\mu 2}, \ldots, x_{\mu n}),
\]
where \( \Psi_{01} \) is the ground state of the system. Similarly,
\[
\Psi_2(x_{\mu(n+1)}, x_{\mu(n+2)}, \ldots, x_{\mu 2n}) = e^{\frac{i k_2}{\hbar} (x_{\mu(n+1)} + x_{\mu(n+2)} + \ldots + x_{\mu 2n})} \Psi_{02}(x_{\mu(n+1)}, x_{\mu(n+2)}, \ldots, x_{\mu 2n}),
\]
where \( \Psi_{02} \) is the ground state of the second string. If the two strings coalesce, then \( x_{\mu n} = x_{\mu(n+1)} \) and
\[
\Psi_3(x_{\mu 1}, x_{\mu 2}, \ldots, x_{\mu n}, x_{\mu(n+2)}, \ldots, x_{\mu 2n}) = e^{\frac{i (k_1 + k_2)}{2\hbar} (x_{\mu 1} + x_{\mu 2} + \ldots + x_{\mu n} + x_{\mu(n+2)} + \ldots + x_{\mu 2n})} \Psi_{03}(x_{\mu 1}, x_{\mu 2}, \ldots, x_{\mu n}, x_{\mu(n+2)}, \ldots, x_{\mu 2n}),
\]
where \( \Psi_3 \) is the wave function for the coalesced string; the new state \( \Psi_3 \) evolves in time. Hence,
\[
i\hbar \frac{\partial \Psi_3}{\partial \tau} = \hat{H} \Psi_3 = e^{iH\tau} \Psi_3, \quad \hat{H} = e^{iH\tau}
\]
where \( \hat{H} \) is just a collection of harmonic oscillators.

The last step breaks the coalesced string into two strings with momentum \( k_3 \) and \( k_4 \) respectfully. The amplitude \( A_V \) is given by
\[
A_V = \int e^{-\tau} e^{\tau(S+1)} (1 - e^{-\tau})^{(-T-1)} d\tau,
\]
where \( S = E_{cm}^2 \) and \( -T = 2(E_{cm}^2 - m^2)(1 - \cos \theta) \). Now \( A_V \) does not look symmetric with respect to \( S \) and \( T \). Let \( e^{-\tau} = z \), then \( -e^{-\tau} d\tau = dz \). Hence,
\[
A_V = -\int z^{-S-1}(1 - z)^{(-T-1)} dz
\]
Note that \( \tau = 0 \rightarrow z = 1 \) and \( \tau = \infty \rightarrow z = 0 \). Hence, let\[
\beta(-S, -T) = -\int_1^0 z^{-S-1}(1 - z)^{(-T-1)} dz = \int_0^1 z^{-S-1}(1 - z)^{(-T-1)} dz,
\]
and note that \( A_V \) is now invariant upon interchanging \( S \) with \( T \).

It turns out that
\[
\beta(-S, -T) = \frac{\Gamma(-S)\Gamma(-T)}{\Gamma(-S-T)},
\]
which is the Veneziano amplitude.
The upshot is that two strings can coalesce then break apart into two or more strings (see 14.7.3.1-1). Hence, there is the potential of modelling particle interactions by employing string diagrams, where the mathematics is similar to the mathematics of quantum field theory.

14.8 Conformal Mappings

Recall that classical systems are governed by the ‘principle of least action’. If the system is non-relativistic, then

$$S = \frac{1}{2} \int \dot{x}^2 dt, \quad m = 1,$$

where $S$, the ‘action’, is minimized over all possible paths $x(t)$ such that

$$x(t) \rightarrow x(t) + \delta x(t), \quad \delta x(t_1) = \delta x(t_2) = 0$$

This is the path the particle takes, called the ‘stationary’ path. If the system is relativistic, then

$$S \propto \int \frac{\partial x^\mu}{\partial \tau} \frac{\partial x^\mu}{\partial \tau} d\tau,$$

where the space-time coordinates $x^\mu$ are functions of the proper time $\tau$. As with the nonrelativistic case, the path of a particle is found by minimizing $S$, subject to $\delta x^\mu(\tau_1) = \delta x^\mu(\tau_2) = 0$.

In quantum mechanics, the exact path of a particle cannot be predicted with certainty, so the quantity sought is the ‘amplitude’. An approximation for the amplitude $A(1,2)$ can be found by the path integral, computed as proportional to
\begin{equation*}
A(1,2) \propto \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} e^{\frac{1}{2} \int_{t_1}^{t_2} \left( \frac{dx_j}{dt} \right)^2 dt} dx_1 \ldots dx_{n-1}, \quad j = 0, \ldots, n,
\end{equation*}

where \( dx_j \) corresponds to the position at the \( j^{th} \) time step, if the time integral is approximated by a sum of \( n \) terms. The amplitude \( 'A(1,2)' \) is associated with the probability that if a particle is found at position \( '1' \) at time \( 't_1' \), it will be found at position \( '2' \) at time \( 't_2' \). Each path, weighted by its action, is integrated over. This is an enormously complicated calculation. In most circumstances, the integral diverges.

Normally, the integrand

\( e^{\frac{1}{2} \int S_j dt} \)

has a very large value for trajectories that deviate significantly from the minimum trajectory. The primary reason that the integrals diverge is because each path is given equal weight.

The problem is overcome by noting that experiments show that the trajectories close to the minimum are more likely to occur, and hence, should be weighted more heavily. The approach begins by letting \( \tau = \alpha s_j \rightarrow d\tau = \alpha ds_j \). Hence,

\begin{equation*}
S_j = \int_{s_1}^{s_2} \frac{1}{\alpha} \left( \frac{dx_j}{ds_j} \right)^2 ds_j
\end{equation*}

Now let \( \alpha = -i \), then

\begin{equation*}
e^{\frac{1}{2} \int S_j dt} = e^{\frac{i}{2} \int_{s_1}^{s_2} \left( \frac{dx_j}{ds_j} \right)^2 ds_j} = e^{\frac{-1}{2} \int_{s_1}^{s_2} \left( \frac{dx_j}{ds_j} \right)^2 ds_j}, \quad j = 0, \ldots, n
\end{equation*}

This substitution allows the exponential to be small when the action is large. In other words, trajectories far from the minimum do not contribute much to the amplitude. On the other hand, small, close to the minimum trajectories, contribute more to the amplitude, allowing the integrals over all the trajectories to converge.

In string theory, strings draw out worldsheets rather than worldlines. Any space-time point on a worldsheet is parameterized by two instead of one parameter. Hence, a space-time point on a worldsheet is given by \( x^\mu(\tau, \sigma) \). The action \( 'S_j' \) for the \( j^{th} \) worldsheets is

\begin{equation*}
S_j = \int_{\tau_1}^{\tau_2} \int_0^\pi \left( \frac{\partial x^\mu}{\partial \tau} \right)^2 - \left( \frac{\partial x^\mu}{\partial \sigma} \right)^2 d\sigma d\tau \rightarrow \text{Amplitude} \propto \int \ldots \int e^{iS_j} dA_j, \quad j = 0, \ldots, n,
\end{equation*}

where \( A_j \) is the surface area of the \( j^{th} \) worldsheets.

String theory amplitudes suffer the same problem as quantum mechanical amplitudes. Wild worldsheets trajectories cause the integrals to diverge. Fixing this problem follows exactly the same approach as in the quantum mechanical case. Let \( \tau \rightarrow -i\tau \), then
\[ iS_j = i \int_{\tau_1}^{\tau_2} \int_0^\pi \left( \frac{\partial x^\mu_j}{\partial \tau} \right)^2 - \left( \frac{\partial x^\mu_j}{\partial \sigma} \right)^2 \, d\sigma d\tau \rightarrow \text{Amplitude} \propto e^{i \int iS_j d\sigma d\tau} \]

\[ = \int \ldots \int e^{-i \int_{\tau_2}^{\tau_1} \sigma \left( \frac{\partial x^\mu_j}{\partial \tau} \right)^2 + \left( \frac{\partial x^\mu_j}{\partial \sigma} \right)^2} \, d\sigma d\tau \, dA_j, \quad j = 1, \ldots, n - 1 \]

Note that

\[ \left( \frac{\partial x^\mu_j}{\partial \tau} \right)^2 + \left( \frac{\partial x^\mu_j}{\partial \sigma} \right)^2 \geq 0, \quad \text{for all } j \]

Hence, trajectories far from the minimum will be large and the exponential will be small. Large trajectories will not contribute much to the amplitude, allowing the amplitude calculation to converge.

If a string moves along a worldsheet, then the equation above is a Lagrangian that describes the motion of the string (see fig. 14.8-1).

![Figure 14.8-1](image)

Under what coordinate transformations of \( \sigma, \tau \) does the Lagrangian remain invariant? If the action is varied and the stationary path computed, then

\[ \left( \frac{\partial x^\mu_j}{\partial \tau} \right)^2 + \left( \frac{\partial x^\mu_j}{\partial \sigma} \right)^2 \rightarrow \frac{\partial^2 x^\mu_j}{\partial \tau^2} + \frac{\partial^2 x^\mu_j}{\partial \sigma^2} = 0 \]

The equation on the right above is in the form of a Laplacian. Suppose a set of points along the \( \tau \)-axis are labelled ‘1,2,3’, then the first derivative can be estimated by

\[ \frac{\partial x^\mu_j}{\partial \tau} \approx \frac{x^\mu_j(3) - x^\mu_j(2)}{\varepsilon} - \frac{x^\mu_j(2) - x^\mu_j(1)}{\varepsilon} \approx \lim_{\varepsilon \to 0} \frac{x^\mu_j(3) - x^\mu_j(1) - 2x^\mu_j(2)}{\varepsilon} \]
The second derivative with respect to $\tau$ and $\sigma$ respectfully can be estimated by

$$\frac{\partial^2 x^\mu_j}{\partial \tau^2} + \frac{\partial^2 x^\mu_j}{\partial \sigma^2} \approx x^\mu_j(1) + x^\mu_j(3) - 2x^\mu_j(5) + x^\mu_j(2) + x^\mu_j(4) - 2x^\mu_j(5) = 0$$

$$\rightarrow x^\mu_j(1) + x^\mu_j(2) + x^\mu_j(3) + x^\mu_j(4) = 4x^\mu_j(5)$$

$$\frac{x^\mu_j(1) + x^\mu_j(2) + x^\mu_j(3) + x^\mu_j(4)}{4} = x^\mu_j(5),$$

see fig. 14.8-2. If the geometrical interpretation of the intersection of the neighboring $\sigma, \tau$-points is an infinitesimal square, then the value of the field $'x^\mu_j'$ at $x^\mu_j(5)$ is just the sum of the average values of the field $'x^\mu_j'$ at the corners of the square. Hence, the Laplacian equation is invariant under rotations of the $\sigma, \tau$-plane. Moreover, the Laplacian equation is invariant if any infinitesimal square is transformed into any other infinitesimal square.

\[\text{Figure 14.8-2}\]

Mappings that take infinitesimal squares into infinitesimal squares leaving the Laplacian invariant are called 'conformal' mappings. Intuitively, a small 8" × 12" map of the United States can be enlarged into an 8′ × 12′ map, but the relative positions of the points on the map remain unchanged. This is an example of a conformal mapping. Conformal mappings preserve angles in the transformation. If the angles are preserved, infinitesimal squares are mapped to infinitesimal squares, since if the angles changed in the transformation, then a square would be mapped to some other shape and the angles would be different, violating the definition of a conformal mapping.

**14.8.1 The Theory of Conformal Mappings**

The theory of electrostatics is described by an electric potential $'\phi'$, where $\phi$ is a field and $
abla|E\rangle$ is defined by the equation $'\nabla\phi = |E\rangle$, $\nabla = \partial / \partial x + \partial / \partial y$ in two dimensions. The charge density is

$$\langle \nabla|E\rangle = \rho = \langle \nabla|\nabla\phi\rangle, \quad \langle \nabla|\nabla\rangle = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \rightarrow \nabla^2 \phi = 0, \quad \rho = 0,$$
where $\rho$ is the ‘charge density’; The equation ‘$\nabla^2 \phi = 0$’ is invariant under a conformal mapping. To see this, let $z = x + iy$ be a point in the $z$-plane and $w = u + iv$ be a point in the $w$-plane, $x, y, u, v \in \mathbb{R}$. If $W$ maps points in the $z$-plane into points in the $w$-plane i.e. $W(z) = w = u + iv$, under what circumstances does $W$ have a derivative (assuming that $W$ is a one-to-one and onto function)?

For $W$ to have a derivative, $W$ must map $z + \Delta z$ into $w + \Delta w$. In other words,  

$$
\lim_{\Delta z \to 0} \frac{\Delta W}{\Delta z} = \frac{dW}{dz}
$$

But for $dW/dz$ to exist, $\Delta \to 0$ must be independent of direction. If $\Delta \to 0$ along the $x$-axis, then $dy = 0$ and 

$$
\frac{dW}{dz} = \frac{du + idv}{dx + idy} = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}
$$

If $\Delta \to 0$ along the $y$-axis, then $dx = 0$ and 

$$
\frac{dW}{dz} = \frac{du + idv}{dx + idy} = \frac{\partial v}{\partial y} - i \frac{\partial u}{\partial y}
$$

Setting the real and imaginary parts of the two equations above equal gives  

$$
\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}
$$

If $dW/dz$ exists in the $x$- and $y$-directions, then $dW/dz$ is independent of direction. Hence, the necessary and sufficient conditions for $W$ to have a derivative, at every point in its domain, are given by the ‘Cauchy-Riemann’ equations. For readers unfamiliar with these equations consult Book II: Chapter 14, Sec 14.5.1. Functions that satisfy the Cauchy-Riemann equations are called ‘analytic’.

Note that  

$$
\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y}, \quad \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial x \partial y} \to \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 v}{\partial x \partial y} - \frac{\partial^2 v}{\partial x \partial y} = 0
$$

Likewise,  

$$
\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0
$$

Suppose there is a small complex number ‘$\delta z$’, where $\delta z = \rho e^{i\theta}$, $\theta$ an angle and let $\Delta z = \rho' e^{i\theta'}$ be a second small complex number. Then  

$$
\frac{\delta z}{\Delta z} = \frac{\rho}{\rho'} e^{i(\theta - \theta')}
$$
If there is a function \( W \) such that \( z \to w \) and if the derivative is taken along the \( \delta z \)-direction, then \( \delta z \to \delta w \) and \( \delta z = \rho e^{i\theta} \), \( \delta w = \rho' e^{i\theta'} \). But if the derivative is taken along the \( \Delta z \)-direction, then \( \Delta z \to \Delta w \) and \( \Delta z = \rho e^{i\varphi} \) and \( \Delta w = \rho' e^{i\varphi'} \). If \( W \) is analytic, then it must follow that

\[
\frac{\delta w}{\delta z} = \frac{\Delta w}{\Delta z} \to \frac{\Delta w}{\delta w} = \frac{\theta - \theta'}{\varphi - \varphi'}
\]

Since \( \theta - \theta' = \varphi - \varphi' \), the angles between vectors in two-dimensional complex space are preserved. Hence, analytic functions are conformal mappings.

Example: Let \( W(z) = z^2 \), then

\[
(x + iy)(x + iy) = x^2 - y^2 + 2ixy \to u = x^2 - y^2, \quad v = 2xy \to \frac{\partial u}{\partial x} = 2x, \quad \frac{\partial v}{\partial y} = 2x,
\]

\[
\frac{\partial v}{\partial x} = 2y, \quad -\frac{\partial u}{\partial y} = 2y
\]

Since the Cauchy-Riemann equations are satisfied, \( W \) is analytic, and hence, conformal.

Example: Let \( W(z) = z^\dagger \), then \( u = x, \quad v = -y \). Now \( \partial u / \partial x = 1 \) and \( \partial v / \partial y = -1 \). Since \( \partial u / \partial x \neq \partial v / \partial y \), the Cauchy-Riemann equations are not satisfied, so \( W \) is not analytic i.e. not conformal.

14.8.1.1 The Role of \( \ln z \) in String Theory

Note that \( W(z) = e^z \) is conformal, which makes the function \( W(z) = \ln z \) conformal, since \( \ln z \) is the inverse function of \( e^z \) and if angles are preserved by a one-to-one and onto function \( W(z) \), then \( W^{-1}(z) \) must also preserve angles.

The function \( W(z) = \ln z \) plays an important role in string theory, because the complex half-plane can be mapped into a worldsheet (see fig. 14.8.1.1-1).

The radial lines, represented on the complex plane, are mapped by \( W \) into the horizontal lines at constant angle and the semicircles are mapped into the vertical lines.
The vertical direction is associated with the parameter ‘\(\sigma\)' which lies on the string and the horizontal direction is associated with \(\tau\). The string can then be described by a function \(x^\mu(\sigma, \tau)\), which moves along the worldsheet. If \(\theta\) varies from \(-\pi/2\) to \(\pi/2\), then the worldsheet models an open string. If \(\theta\) varies from \(-\pi/2\) to \(-\pi/2 + 2\pi\), then the worldsheet models a closed string. The two parameters ‘\(\sigma, \tau\)' describe where the string is on the worldsheet and \(x^\mu\) gives the space-time location of the string.

### 14.8.1.2 Conformal Mappings in String Theory

Recall that the action of the \(j^{th}\) worldsheet is

\[
S_j = \int_{\tau_1}^{\tau_2} \int_0^\pi \left( \frac{\partial x^\mu_j}{\partial \tau} \right)^2 - \left( \frac{\partial x^\mu_j}{\partial \sigma} \right)^2 \, d\sigma d\tau \rightarrow \text{Amplitude} \propto \int \ldots e^{iS_j d\sigma d\tau} \, dA_j,
\]

where \(A_j\) is the surface area of the \(j^{th}\) worldsheet. The upshot is that any appropriate bounded shape can be conformally mapped to any other bounded shape, a theorem on analytic functions. Recall that two open or closed strings can join to form a single string and then break apart to form two or more strings.

The string is modelled by \(x^\mu(\sigma, \tau)\), which remains invariant under conformal mappings. Hence, the process of two open or closed strings joining to form a single string, then breaking apart to form two or more strings, can be conformally mapped into a series of concentric circles (see fig. 14.8.1.2-1). The \(x^\mu(\sigma, \tau)\)'s are conceived as living on the concentric circle except for four exceptional points at infinity that must be treated specially. In other words, there is a one-to-one correspondence between the \(x^\mu(\sigma, \tau)\)'s that lie on the upper figure and those points lying on the concentric circles, except for the exceptional points, which, depending on the shape of the upper figure, are mapped to different points on the circles.
Where on the circular discs do the exceptional points go? If the four exceptional points have up-down as well as left-right symmetry (see fig. 14.8.1.2-2 a)), then there are two extreme cases, shown in fig. 14.8.1.2-2 b) and c). The exceptional points represent the amount of time the two strings stay coalesced before breaking apart. The situation shown in fig. 14.8.1.2-2 c) represents two strings that remain coalesced for a long time. The other extreme, shown in fig. 14.8.1.2-2 b), represents two strings that remain coalesced for a short time.

The only thing missing from the amplitude calculation is the momentum of the external particles that come in, basically from infinity, with a certain momentum. This omission is resolved by including the following in the amplitude calculation:

\[
\text{Amplitude} \propto \int e^{ik_{\mu}\gamma^\mu(z_{rh})} \int e^{iS_j\sigma \sigma \tau} dA_j dz_{rh}, \quad j = 0, \ldots, n, \quad \mu = 0 \ldots, 25
\]

The \( k_{u} \)'s are fixed, but to compute the amplitude, the integration must be taken over all the space-time locations that particles with momentum \( 'k_{u}' \) could enter and leave. Note that the momentum has 26 components as well as 26 space-time components. In essence, there are 26 electrostatic problems added up and then summed over all possible worldsheets to obtain the amplitude of two particles with momentum \( 'k_1' \) and \( 'k_2' \) coming in, coalescing, then breaking into two particles with momentum \( 'k_3' \) and \( 'k_4' \) going out. If this can be done for two particles, it can be done for any number of particles.

If the momentums of the two incoming strings are \( k_1 \) and \( k_2 \) respectfully and the momentums of the two outgoing strings are \( k_3 \) and \( k_4 \) respectfully, then fig. 14.8.1.2-2 c) corresponds to \( (k_1 + k_2)^2 = -S = E_{cm}^2 \), where \( S \) is a Mandelstam variable and fig. 14.8.1.2-2 b) corresponds to \( (k_1 + k_3)^2 = -T = (E_{cm}^2 - M^2)(1 - \cos \theta) \), where \( T \) is also a Mandelstam variable. Moreover, since \( T \) and \( S \) are invariant under exchange, the process in which strings come together, coalesce, then break apart is much like particle decay and the exchange of a particle between two or more particles i.e. force carriers. Hence, one basic structure can represent both the \( S \)- and \( T \)-channel processes. This
was a major discovery in string theory. In fact, this represents the entire essence of open sting theory.

14.9 String Theory, Gravity and Multiple Compactified Dimensions

In classical physics, if no other forces are present, a particle, confined to a surface, moves along a geodesic, the shortest path between two points on a curved surface. If the surface is Euclidean flat, the path is a straight line. The term 'geodesic' generalizes the notion of a 'straight line' to broader mathematical spaces, often referred to as 'differential manifolds'. Such spaces (in any number of dimensions) must be smooth enough that the requisite number of derivatives exists. This guarantees that the motion of the particle is totally measurable along the surface or hypersurface.

In quantum mechanics, it is not possible to predict the exact path of a particle; the least action principle does not apply. Instead, the path integral approach is used instead, where the amplitude is computed and is related to the probability of the particle starting at position 'a' and ending up at position 'b'. All possible paths must be integrated and summed over, a far more complicated problem than the least action approach. But, with some mathematical trickery, in most cases, the integrals exist, and a finite amplitude can be computed.

If the particle is confined to a sphere, then, classically, the equation of motion is \( E = \frac{1}{2} m v^2 = \frac{p^2}{2m} \). While the quantity 'v' is confined to the sphere, the particle's motion can be described by some form of Newton's or Einstein's equations and the principle of least action. Quantum mechanics requires that the angular momentum 'L' be quantized i.e. \( L = pr \propto i\hbar \). Hence,

\[
E = \frac{1}{2} m v^2 = \frac{p^2}{2m} = \frac{L^2}{2mr^2} = \frac{L^2}{2I}, \quad I = mr^2
\]

where \( r \) is the radius of the sphere and \( I \) is the 'moment of inertia'.

Whether finding the least action in classical mechanics, or the sum of the path integrals in quantum mechanics, in general, the limits exist, and the calculations can be executed. In string theory, the string, confined to a sphere, is spread out over a part of the surface, rather than located at one particular point.

14.9.1 String Theory in Flat Geometries

Recall that for an open string

\[
\hat{x}(\sigma, \tau) = \sum_{j=0}^{\infty} \hat{x}_j \cos j \sigma = \sum_{j=0}^{\infty} \frac{\hat{a}_j^- + \hat{a}_j^+}{\sqrt{j}} \cos j \sigma, \quad \hat{x}_j = \frac{\hat{a}_j^- + \hat{a}_j^+}{\sqrt{j}}
\]
then

\[ \hat{x}^2(\sigma, \tau) = \sum_{j,k=0}^{\infty} \frac{\hat{a}_j^- + \hat{a}_j^+ \hat{a}_k^- + \hat{a}_k^+}{\sqrt{j} \sqrt{k}} \cos j\sigma \cos k\sigma \]

The average value \( \langle g | \hat{x}^2 | g \rangle \) of \( \hat{x}^2 \) in the ground state \( |g\rangle \) of the string is

\[ \langle g | \sum_{j,k=0}^{\infty} \frac{\hat{a}_j^- + \hat{a}_j^+ \hat{a}_k^- + \hat{a}_k^+}{\sqrt{j} \sqrt{k}} \cos j\sigma \cos k\sigma | g \rangle \]

In quantum mechanics, the ground state or vacuum energy is not zero, as it is in classical physics. What is the ground state energy of an open string? The terms in the equation above that involve \( \hat{a}_j^- \hat{a}_k^- \) annihilate the ground state and contribute nothing to the ground state energy. Likewise, the terms involving \( \hat{a}_k^+ \hat{a}_j^+ \) that create energy states two levels above the ground state contribute nothing to the ground state energy. The only terms contributing to the ground state energy involve \( \hat{a}_j^- \hat{a}_k^+ \), but only if \( j = k \), since it is only possible to annihilate an oscillator that has been created. Hence,

\[ \langle g | x^2 | g \rangle = \frac{1}{2} \sum_{j=1}^{\infty} \frac{1}{j} \approx \ln j_{\text{max}} \]

since \( \langle g | \sum_{j=0}^{\infty} \cos^2 j\sigma | g \rangle \approx 1/2 \) and \( a_j^- a_j^+ = 1 \). Note that \( x^2 - x_{cm}^2 \) is the distance of each point on the string from its center of mass and is a measure of the length of the string. However, the string is infinitely long. One way of making \( x^2 - x_{cm}^2 \) finite is to cut off the sum above some very high value of \( j \), say \( j_{\text{max}} \). While \( \sum_{j=1}^{j_{\text{max}}} 1/j \) is finite, string theory requires taking a limit. It is a remarkable mathematical fact that in flat space a fortunate number of cancellations make \( x^2 - x_{cm}^2 \approx \ln j_{\text{max}} \). It is immaterial that the string is spread out, its center of mass moves as \( p^2 / 2m = L^2 / 2mr^2 \). The fact that the string is moving with an infinite number of oscillating modes is of no consequence.

**14.9.2 String Theory in Curved Geometries**

Suppose the string is confined to a sphere, which is a curved surface. If the string is not oscillating, it moves like a point particle and the rules of Einsteinian, Newtonian or quantum mechanics apply. In particular, for any point on the sphere, the distance from any axis of rotation is always the same. Its angular momentum around that axis is simply a function of the velocity of the particle. But, if the string has non-zero modes of oscillation, it spreads out over the surface of the sphere. The more modes of oscillation, the more the string spreads out. Although the center of mass of the string will move along a great circle (geodesic), some of the points along the string will be closer to the axis of rotation than other points. This affects the moment of inertia calculation, and hence, the energy of the string. In general, the string’s moment of inertia will be less than if it was not oscillating. There are a couple of ways of representing this: 1) the string covers more and more of the sphere as modes of oscillation are added, 2) the string is conceived as a point particle that moves on a smaller and smaller sphere as
more modes of oscillation are added. But there is no limit to the number of modes that can be added. In either case, as $j_{max} \to \infty$, $l \propto mr^2 \to 0$ and the string has infinite energy, and hence, cannot describe the true nature of the string. This happens solely as a result of the curvature of the space and is true in any number of dimensions. The upshot is that, in general, string theories do not make sense in curved geometries.

14.9.3 Conditions for a Sensible String Theory

Under what conditions does it make sense to develop a theory of strings? When more and more modes of oscillation are added to the string, the most desirable condition is, if the string is confined to a geometrical surface, the surface does not change, but tends to a limit. Recall that a geometry is described by a metric tensor $'g_{\mu\nu}(x_{\mu})'$. If $g_{\mu\nu}(x_{\mu})$ represents a certain geometry, the change in the geometry $\delta g_{\mu\nu}$ as more and more modes of oscillation are added to the string depends on the curvature of the geometry. If the geometry is flat, then $\delta g_{\mu\nu} = 0$ and there is no change. But if the space has curvature, in general, $\delta g_{\mu\nu} \neq 0$. Recall that

$$\delta g_{\mu\nu} \propto R_{\alpha\beta\gamma}^{\delta} \to \delta g_{\mu\nu} \propto R_{\mu\nu},$$

where $R_{\alpha\beta\gamma}^{\delta}$ is the curvature tensor, which has been contracted to $R_{\mu\nu}$. It turns out that $\delta g_{\mu\nu} \propto -R_{\mu\nu}$, which corresponds to the geometry getting smaller as more modes of oscillation are added to the string. So, $\delta g_{\mu\nu} = 0 \to R_{\mu\nu} = 0$. In other words, geometries, where string theory makes sense, are those where $R_{\mu\nu} = 0$. Such geometries are called ‘Ricci flat’. The term ‘Ricci flat’ does not mean that the geometry is Euclidean flat. The Ricci tensor $'R_{\mu\nu}'$ is part of Einstein’s gravitational equations i.e.

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R_{\alpha}^{\alpha} = T_{\mu\nu}$$

The condition that $R_{\mu\nu} = 0 \to T_{\mu\nu} = 0$ is the Einstein gravitational equation for the vacuum, which contains just gravity, no mass, no energy. Note that

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R_{\alpha}^{\alpha} = 0 \to R_{\mu}^{\nu} - \frac{1}{2} g_{\mu}^{\nu} R_{\alpha}^{\alpha} = 0 \to R_{\mu}^{\nu} - \frac{1}{2} g_{\mu}^{\nu} R_{\alpha}^{\alpha} = 0, \quad g_{\mu}^{\nu} = \delta_{\mu}^{\nu}$$

Hence,

$$R_{\alpha}^{\alpha} - \frac{1}{2} \delta_{\alpha}^{\alpha} R_{\alpha}^{\alpha} = 0 \to R_{\alpha}^{\alpha} = \frac{1}{2} \delta_{\alpha}^{\alpha} R_{\alpha}^{\alpha} = 2R_{\alpha}^{\alpha} \to R_{\alpha}^{\alpha} = 0 \to G_{\mu\nu} = R_{\mu\nu} = 0,$$

the Einstein equations for the gravitational field. The upshot is that the only geometries in which string theory makes sense are those in which the Einstein gravitational field equations are satisfied. Not only must the geometries be Ricci flat, but must have the requisite number of dimensions, 10 for superstring theory and 26 for bosonic string theory.

It is quite remarkable that the only geometries that support a sensible string theory are those where the Einstein gravitational field equations are satisfied. But, it is a two-
edged sword. Certainly, there are geometries in which $T_{\mu\nu} \neq 0$, where evidently, a sensible string theory cannot be developed.

14.10 Compactification

How is a theory that has 10 or 26 dimensions made sense of? Ideally, at some point, the extra dimensions should vanish. But, if that is not possible, the next best thing is to roll extra dimensions up into tiny spaces that don’t contribute much to the overall theory. The process of rolling up dimensions into tiny spaces is called ‘compactification’.

Figure 14.10-1

Figure 14.10-1 shows a line (one dimensional space) transformed into an infinite cylinder, which has one non-compactified dimension and one compactified dimension. If confined to the line, particles can only move in one dimension, but in the compactified case, a particle can move in two dimensions. But, when a particle moves in the compactified dimension, it can move in a circle, the simplest example of compactification. It is possible to compactify two of three dimensions or six of ten dimensions, as is done within superstring theory.

14.10.1 The Torus as a Compactified Space

The simplest surface for compactification is a torus (see Figure 14.10.1-1 a)). If the torus is cut as in Figure 14.10.1-1 a) and opened up, it becomes a cylinder (see Figure 14.10.1-1 b)). However, the points along the cut must be identified with one another. For instance, if a particle leaves the surface at point ‘a’ on the right, it reappears immediately at point ‘a’ on the left, since it is the same point. The same can be said for point ‘b’ (see Figure 14.10.1-1 b)). If, in turn, the cylinder is cut horizontally as in Figure 14.10.1-1 b) and opened, it becomes a rectangle (see Figure 14.10.1-1 c)). The rectangle is compactified in the sense that it does not have edges. If a point leaves the surface at, say point ‘b’ on the upper edge, it immediately reappears at point ‘b’ on the lower edge. The same can be said for point ‘a’ (see Figure 14.10.1-1 c)). In other words, points can move on the surface, but cannot leave the surface. The upshot is that a torus is topologically equivalent to a compactified rectangle, a flat surface. A sensible string theory can be developed on a torus. The torus in the figure is a two-dimensional surface, but mathematically, it can be extended to any number of dimensions, say six. In this case, there would be six compactified dimensions. Combining four non-compactified dimensions with six compactified dimensions makes a theory in ten-dimensions. Superstring theory has four non-compactified dimensions and six compactified dimensions.

463
Toroidal compactifications are Ricci flat spaces suitable for defining a sensible string theory, where the characteristics of the string are well defined. While the simplest, toroidal compactification does not lend itself to real world physics. There is another class of Ricci flat spaces, called ‘Calabi–Yau’ spaces, that yield applications in theoretical physics, particularly in superstring theory. The extra dimensions of space-time are described on a 6-dimensional Calabi–Yau manifold, which leads to the idea of mirror symmetry. The mathematics of a Calabi–Yau manifold goes well beyond the scope of this discussion, but is mentioned as a motivation for further study.

14.10.2 Momentum in the Compactified Dimension

Consider the simplest case of a particle moving in one non-compactified and one compactified dimension (see Figure 14.10-1). A particle moving along this surface can have momentum in the non-compactified and in the compactified dimension. The momentum in the compactified dimension is quantized. To see this, if the compactified dimension is in the form of a circle, the circumference is $2\pi r$. If the momentum in the compactified dimension is $p_c$, then

$$p_c r \propto L_c = n\hbar, \quad n \in \mathbb{N} \rightarrow p_c = \frac{in}{r}, \quad \hbar = 1$$

If the particle is massless, then its energy is equal to its momentum i.e. $E = |p|$, $c = 1$. If the particle has no momentum in the non-compactified dimension, it can have energy in the compactified dimension proportional to $n/r$. Recall that $E = m$ in units where $c = 1$. Interestingly, this describes a particle at rest in the non-compactified dimension, but has momentum in the compactified dimension, and hence, in principle, has a mass.

In addition, the string can be wrapped around the compactified dimension. The tension in the string is proportional to its length, which, in turn, is proportional to the energy of the string i.e. to the mass of the string. The tension of the string is a function of the circumference of the circle ($2\pi r$), and therefore, the mass of the string is proportional to $r$. The energy of the string, then, is proportional to $nr$, where $n$ is the number of times the string is wrapped around the compactified dimension, sometimes referred to as the ‘winding number’. For a wrapped string, $E \propto nr$. 

464
14.10.2.1 T-Dual Strings

The two cases, momentum in the compactified direction and wound strings represent two kinds of particles with an interesting symmetry. If \( r \ll 1 \), a very small amount of energy is required to wind a string around the compactified dimension, so the energy levels for the wound string are very close together, while the energy levels for the momentum string are far apart. As \( r \) grows, the two types of strings morph into one another. When \( r = 1 \), the energy levels for the two types of strings are identical. If \( r \gg 1 \), the two types of strings reverse roles. The high energy strings are the wound strings and the low energy strings are the momentum strings. This symmetry only happens if the strings are closed, suggesting that there is a smaller sized scale in string theory. Theories that have this kind of symmetry are called "T-dual". The "T" stands for torus.

10.11 T-Duality and D-Branes

Theories that are T-dual are compactified on an \( n \)-dimensional torus, which is Ricci flat in any number of dimensions. A 1-dimensional torus is topologically equivalent to a line; a 2-dimensional torus is topologically equivalent to a parallelogram with the opposite edges identified with one another and a 3-dimensional torus is topologically equivalent to a parallelepiped with the opposite faces identified with one another.

10.11.1 Winding Number

Consider a closed string moving in a space that has one non-compactified and one compactified dimension illustrated in fig. 14.10-1. The closed string can be oriented so that there is a distinction between traversing the sting in one direction and traversing the sting in the opposite direction. If the string is not wound around the compact direction, it is free to move either in the compactified or in the non-compactified dimension; it has momentum in both directions. The momentum along the compactified dimension is quantized. If a string is wound around the compactified dimension, it can be wound in a positive or negative orientation. Wound strings can joint and spit. In either case, orientation is conserved. A string oriented in the positive direction has winding number \( +1 \). A string oriented in the negative direction has winding number \( -1 \). All unwound closed strings have winding number \( 0 \).

14.11.2 Radius of Compactification

If the string is not wound, its momentum \( p_c \) in the compactified direction is \( p_c = E_c \propto n/r \), \( n \in \mathbb{N} \). If working in units where the mass of the string is 1, then \( n/r = v = dy/d\tau \), where \( y \) signifies the compactified dimension and \( v \) is the velocity of the center of mass of the string. If the string is wound, the string tension is proportional to \( E/L = w r \), where \( E \) is the 'energy in the string', \( L \), the 'length of the string' and \( w \in \mathbb{Z} \) is the 'winding number'. In both cases, \( r \) is the radius of the compactified dimension and is called the 'radius of compactification'. If \( r \ll 1 \), then the energy levels of the unwound strings are far apart, but those of the wound strings are very close together. If \( r \gg 1 \), the energy levels of the unwound strings are close together and those of the wound strings far apart. The energy levels of both the wound and unwound string are integer multiplies of
Interestingly, if the only thing that can be measured is the energy of the string, it is impossible to know whether the string is wound or unwound. In other words, interchanging the momentum number with the winding number does not change the physics of the problem. Remarkably, this is true of all calculations performed in higher dimensional string theories. Hence, there is a symmetry between the two types of string classifications: ‘wound’ and ‘unwound’.

If \( \sigma = y/r \), where \( \sigma \) is the ‘string parameter’ for a wound string, then \( \partial \sigma / \partial y = 1/r \).

Recall that \( n \approx r \int \partial y / \partial \tau \) and \( w \approx 1/r \int \partial y / \partial \sigma \). So, \( T \)-duality involves interchanging \( n \) with \( w \) with \( 1/r \) and \( \partial y / \partial \tau \) with \( \partial y / \partial \sigma \). In other words, \( T \)-duality is an exact symmetry of string theory.

### 14.11.3 The Significance of Wound and Unwound Strings

Wound and unwound closed strings have properties closely related to charged particles, the electron, for example. Two closed strings having their compactified momentum in the same direction repel and two closed strings having compactified momentum in opposite directions attract. These attractions and repulsions result from the gravitational characteristics in the higher dimensional string theories. To get a basic understanding, recall that Einstein’s theory of gravitation is given in terms of the metric tensor \( g_{\mu \nu} \), \( \mu, \nu = 0,1,2,3 \). Now consider a metric tensor \( g_{mn} \), \( m,n = 0,1,2,3,4 \), in five dimensions, one time and four spatial dimensions, where the fifth dimension is compactified; the metric \( g_{mn} \) can be written

\[
g_{mn} = g_{\mu \nu} + 2 \sum_{\mu} \frac{g_{\mu 5}}{A_{\mu}} + g_{55}
\]

The tensor \( g_{\mu \nu} \) is associated with the ordinary four dimensional space-time of general relativity; the quantity \( g_{\mu 5} \) is a 4-vector which is the analog to the electromagnetic potential and \( g_{55} \) is an undefined scalar.

Interestingly, string theory predicts two kinds of electromagnetic phenomena. One is associated with \( n \), the momentum of an unwound string, and the other with \( w \), the winding number. Suppose there is a closed string in the ground state \( |g \rangle \). The string can be excited from the ground state using a creation operator \( a_i^{+} \), where \( i \) labels directions in space and \( n \) labels the energy level. In other words, \( a_i^{+} \) starts the string vibrating in every direction. In this case, one of the spatial dimensions is compactified. So, \( n \) is related to the frequency and \( i \) is related to the direction in space. Moreover, closed strings are oriented; vibrations in the string can move in one direction around the string or in the opposite direction, labeled \( a_i^{+}(L) \) and \( a_i^{+}(R) \) respectfully. Level matching requires that the energy of the \( L \)-moving vibrations equal the energy of the \( R \)-moving vibrations around the string. Firstly, the string’s lowest state of energy is the tachyon with \( -2 \) units of energy, but represents something that should not be there. The next possibility is to excite the string with one unit of energy i.e. \( a_i^{+}(L)|g \rangle \) or \( a_i^{+}(R)|g \rangle \). Neither of these states is possible, since both violate level matching. The simplest states that can actually exist in string theory are of the form
\begin{align*}
\hat{\alpha}_1(L)\hat{\alpha}_1(R)\langle g
\end{align*}

For example, \(\hat{\alpha}_1(L)\hat{\alpha}_1(R)\langle g\) or \(\hat{\alpha}_1(R)\hat{\alpha}_1(L)\langle g\) are expressions structurally similar to a photon. A photon-like object has momentum in the non-compactified dimensions and circles around (has polarity) in the compactified dimension. Interestingly, there are two types of photon-like fields. Quantum mechanically, the two fields are represented in superposition i.e.

\begin{align*}
\hat{\alpha}_1(L)\hat{\alpha}_1(R)\langle g + \hat{\alpha}_1(R)\hat{\alpha}_1(L)\langle g
\end{align*}

or

\begin{align*}
\hat{\alpha}_1(L)\hat{\alpha}_1(R)\langle g - \hat{\alpha}_1(R)\hat{\alpha}_1(L)\langle g
\end{align*}

The first is associated with the graviton field and the second is associated with the winding number.

A summary of T-duality is shown in table 10.7-1.

<table>
<thead>
<tr>
<th>T-duality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Characteristic</strong></td>
</tr>
<tr>
<td>Quantification</td>
</tr>
<tr>
<td>Radius of Compactification</td>
</tr>
<tr>
<td>Parameterization</td>
</tr>
<tr>
<td>Compactified Metric</td>
</tr>
</tbody>
</table>

Table 10.7-1

In T-dual theories, the energy of an unwound string is proportional to \(n/r\). In wound string theory, the energy is proportional to \(wr\), where \(n \in \mathbb{N}\), \(w \in \mathbb{Z}\). Recall that the string itself is parameterized by \(\sigma\) and \(\tau\) i.e. \(x_\mu(\sigma, \tau)\). In the unwound theory, the rate of change in \(x_\mu\) is relative to \(\tau\) and in the wound theory the rate of change in \(x_\mu\) is relative to \(\sigma\). In the unwound theory, the object described is photon-like i.e. the graviton. In the wound theory, the object described is called a 'dilaton' and is associated with the winding number.

**14.11.4 T-Dual Theories for Open Strings and Dn-Branes**

The compactified theories discussed so far involved only closed strings. What about open strings moving in compactified dimensions? Open strings cannot be wound and have winding number '0'. Moreover, the boundary conditions placed on open strings require that \(\partial x_{\mu 2}/\partial \sigma = \partial x_{\mu 1}/\partial \sigma = 0\), where \(x_{\mu 1}\) and \(x_{\mu 2}\) correspond to the endpoints of the string and where \(\mu\) includes the compact directions. So, in what sense is an open string T-dual?
Suppose $x_{51}$ and $x_{52}$ are the endpoints of the string in the compact dimension, then the equation $\partial x_{51} / \partial \sigma = 0$ implies that the string cannot move in that dimension. And if it cannot move, then it must be attached to something. That something is called a $D$-brane. A $T$-dual open string theory must contain $D$-branes, objects open strings attach themselves to. If a string is constrained from moving in one of three dimensions, then it is free to move in two dimensions. If a string theory has 10-dimensions, one time and nine spatial dimensions, if the open string is constrained from moving in one of the dimensions, then the dimension of the $D$-brane is eight i.e. is a $D_8$-brane. $D$-branes can be created in any number of the compactified dimensions. For example, a theory can have nine spatial dimensions, eight of which are compactified. An open string in such a theory is constrained to move on a line (not necessarily straight). A line is called a $D_1$-brane. There is also a $D_0$-brane, which represents an open string that cannot move. Interestingly, $D_0$-branes represent a new particle in string theory.

A $D_1$-brane is essentially a line, structurally identical to a string. The difference between a $D_1$-brane and a string is that a $D_1$-brane is heavier than a string. This is not a very satisfactory answer, but the one that string theorists give.

A $D_2$-brane is essentially a surface. Open strings attached to the $D_2$-brane can move around on the surface. In fact, their endpoints can joint, lift off the surface, creating a new string attached to the $D_2$-brane. They can also break apart, each end attaching itself to the $D_2$-brane, forming two open strings from one string (see Fig. 14.11.4.1-1).

![Figure 14.11.4.1-1](image)

There can be more than one brane. Suppose there are three $D_2$-branes, call them ‘red’, ‘green’ and ‘blue’. An open string attached to the red $D_2$-brane could be called a ‘red-red’ string. One end of an open string could be attached to the red $D_2$-brane and the other end to the green $D_2$-brane. This string could be called a ‘red-green’ string. Note the similarities to the gluon interactions within QFT. In fact, string theorists are attempting to build string interactions that mirror gluon interactions, making rules for the splitting and joining of the strings that mirror the Yang-Mills rules in quantum field theory. Finally, quarks are represented by open strings in which only one end of string is attached to the $D_2$-brane. Since stings are oriented, a quark is oriented into the brane and an antiquark is oriented away from the brane. If there is only one $D_2$-brane, then the theory is similar to quantum electrodynamics, where strings with both ends attached are photon-like objects and those with just one end attached to the brane are electron-
like objects that carry charge. Strings coming into the brane have positive charge. Those going away from the $D_2$-brane carry negative charge.

Recall that a $D_1$-brane is just a heavy string. A $D_1$-brane can end on a $D_2$-brane. This can only occur in a space with three or more spatial dimensions. What do $D_1$-branes represent? Speculatively, one of the things they could represent is magnetic monopoles. Interestingly, string theory predicts the existence of magnetic monopoles, which have yet to be discovered. This is one of the more controversial aspects of string theory. All string theories carry extra ingredients, many with no apparent physical explanation and it is sometimes difficult to remove unwanted ingredients from the theory.

14.12 Concluding Remarks

So, what is string theory? In its present form, it is not a theory of the Universe. However, it possesses many features a theory of the Universe is expected to include. It has graviton-like objects and many features of the gauge theories associated with the standard model. It contains photon and electromagnetic-like objects associated with the theory of electrodynamics. Currently, physicists regard string theory primarily as a sophisticated learning environment, where its flexibility provides an ideal setting for trying out new ideas and for understanding how string concepts might apply to the physical world. But if string theory is to become a theory of the Universe, it must overcome at least three fairly significant hurdles.

Firstly, it is almost universally agreed that the 26-dimensional bosonic string theories are poor candidates for a theory of the Universe. At this point, the most promising string theories are the 10-dimensional superstring theories that rely on ‘supersymmetry’, required for mathematical consistency. Without supersymmetry, 10-dimensional superstring theories suffer the curse of infinities, which ordinarily dooms imaginative conjectures as potential theories.

Supersymmetry is the conjecture that every boson has a fermionic super-partner and vice versa. None of the super-partners are part of the standard model of particle physics. Moreover, supersymmetry does not exist in Nature. If it did, the super-partners would have been discovered, since their masses would be the same as their counterparts and the energies to produce them are well within current experimental capabilities. To make supersymmetry viable, somewhere along the way, the symmetry must have been broken. And the only way for symmetry breaking to make sense is if the super-partners are much heavier than their counterparts, how much heavier is a matter of conjecture. When the super-partners become too heavy, it is increasingly difficult to create a string theory consistent with the standard model. So, one challenge is to describe how the symmetry is broken, predict the masses of the super-partners and provide a theory consistent with standard model.

Secondly, string theory comes with a host of mysterious ingredients. The discovery that the only way of making sense of opens strings in compactified spaces was by introducing the ‘brane’ concept delivered an enormous amount of flexibility to string theory. The flexibility was both a blessing and a curse. It opened a rich environment,
where theorists could experiment with creative ideas. It became possible to produce an inordinate number of string theories, about $10^{500}$ or more. But the euphoria of possibilities levies an obligation to explain what all those theories represent. Moreover, all the string theories created so far contain mysterious objects, the tachyon and the dilaton, for example, that are either unwanted or not well understood. It is sometimes difficult to eliminate unwanted objects or to explain the role these objects play within the theory.

Philosophically, the research environment created by string theory has produced a plethora of speculative and sometimes controversial ideas that claim to explain what all the many string theories represent, not the least of which is the ‘anthropic principle’. The ‘string theory landscape’ refers to the huge number of possible string theories. The large number of theoretically allowed configurations promotes the idea that certain physical mysteries, particularly related to the fine-tuning of the constants of Nature, are explained, not by a physical mechanism, but by assuming that the many different string theories represent different universes, referred to as ‘multiverses’. The large number of possibilities arises from different choices of Calabi–Yau spaces and different values of generalized magnetic fluxes over different geometries [15].

The ‘anthropic principle’ is the idea that the observed Universe must be compatible with the conscious life that observes it. The principle was formulated because the laws of Nature seemed to assume values consistent with the conditions for life. The anthropic principle regards this as a necessity, since, if life were impossible, no living entity could observe it, and thus, such laws of Nature could not be known.

The anthropic principle comes in two versions:

1. **The strong anthropic principle**: contends that the Universe is compelled to produce conscious life; it is often used as an argument for the existence of a supreme deity.

2. **The weak anthropic principle**: portends a Universe with fined tuned constants, a result of a selection bias; only in a universe capable of eventually supporting life will there be living beings capable of observing and reflecting upon the fine-tuning. In other words, out of so many possible universes, there will be a statistical probability that at least some of the universes will be compatible with life. Proponents of the weak anthropic principle often argue that it is unremarkable that this Universe should have fundamental constants that fall within the narrow range compatible with life. This is often used as an argument against the existence of a supreme deity [246].

The ‘anthropic landscape’ refers to the collection of string theories that are suitable for supporting intelligent life, an application of the anthropic principle that selects a certain subset of the otherwise possible universes. This, somewhat radical interpretation, is given as a possible explanation for why there is such an enormous number of string theories, each having its own fundamental constants and parameters. And at least some of the possible string theories would have parameters consistent with our Universe. But, no string theory developed so far is compatible with the standard model.
The anthropic principle is often criticized for being unscientific and speculative. It remains to be seen if these criticisms can be overcome, where skeptics will find worthiness in string theory as a viable theory of Nature.

Perhaps the biggest obstacle for string theory resides in what it is best known for, predicting gravity. No string theory, at least no known string theory, exists without gravity. This fact is often sighted as to the viability of string theory as a candidate for a theory of the Universe. If the gauge theories associated with the standard model can be reproduced in string theory, gravity comes along for ride. Therefore, all four known fundamental forces could be explained using one concept, namely, the ‘string’. The general theory of relativity, the current theory gravity, is written in terms of the curvature tensor $R_{abcd}$. In terms of 4-dimensional space-time, $R_{abcd}$ has 256 equations. The prospects for solving the equations are not quite so daunting, since some of the equations are duplicates. But, in general, finding a solution is difficult. One thing is known, $R_{abcd} = 0$ has only one solution, the ‘zero’ solution. In other words, if $R_{abcd} = 0$, the geometry is Euclidean flat. Unless an external force causes a change in speed or direction, the object moves in a straight line or stands still.

Einstein’s field equations take the form $G_{\mu\nu} \propto T_{\mu\nu}$, where $G_{\mu\nu}$ is associated with the geometry of space-time and $T_{\mu\nu}$ is associated with the amount of matter/energy at the locality where the objects of interest reside; the tensor $G_{\mu\nu}$ is derived from a contraction of the curvature tensor $R_{abcd}$. Interestingly, unlike $R_{abcd} = 0$, the equations $G_{\mu\nu} = 0$ have nonzero solutions. The zero solution is a solution to $G_{\mu\nu} = 0$, but there are others. Geometries associated with $G_{\mu\nu}$ are not necessarily Euclidean flat. Mathematicians call $G_{\mu\nu} = 0$ ‘topologically’ or ‘Ricci’ flat. A way of thinking about this is to visualize a surface where cuts made along certain portions of the surface allow the flattening of the surface without removing any of the surface material – a ‘donut’ or ‘torus’ is an example. If material must be removed to make the surface flat, then the surface is ‘curved’ – a sphere is an example.

The upshot is that string theories are written on Calabi–Yau spaces. All Calabi–Yau spaces are Ricci flat. Writing a sensible string theory on a curved space-time, where $G_{\mu\nu} \neq 0$, seems out of reach. It is difficult to see how string theory, restricted to Calabi–Yau spaces, could be fully compatible with general relativity.