

The White Hole Existence Principle (Mathematics, Science and NS* Version)

WHEP Framework

Created by Prof. Nils Efverman

|

2025

Project by FÈUE

Contact: feue.com1@outlook.com

DISCLAIMER: THIS IS NOT ONLY FACTS, BUT ALSO A

THEORY

FÈUE copyright 2025

Thing that is'ent yet in the document:

10.7 Halogens

10.7.1 Fluorine (F)

10.7.2 Chlorine (Cl)

10.7.3 Bromine (Br)

10.7.4 Iodine (I)

10.7.5 Astatine (At)

10.7.6 Tennessine (Ts)

10.8 Noble Gases

10.8.1 Helium (He)

10.8.2 Neon (Ne)

10.8.3 Argon (Ar)

10.8.4 Krypton (Kr)

10.8.5 Xenon (Xe)

10.8.6 Radon (Rn)

10.8.7 Oganesson (Og)

10.9 Lanthanides

10.9.1 Lanthanum (La)

10.9.2 Cerium (Ce)

10.9.3 Praseodymium (Pr)

10.9.4 Neodymium (Nd)

Written and created by Nils Efverman. This project is powered by FÈUE inc. FÈUE

office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

10.9.5 Promethium (Pm)

Written and created by Nils Efverman. This project is powered by FÈUE inc. FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on: <https://tinyurl.com/3xceeda5>

10.9.6 Samarium (Sm)

10.9.7 Europium (Eu)

10.9.8 Gadolinium (Gd)

10.9.9 Terbium (Tb)

10.9.10 Dysprosium (Dy)

10.9.11 Holmium (Ho)

10.9.12 Erbium (Er)

10.9.13 Thulium (Tm)

10.9.14 Ytterbium (Yb)

10.9.15 Lutetium (Lu)

10.10 Actinides

10.10.1 Actinium (Ac)

10.10.2 Thorium (Th)

10.10.3 Protactinium (Pa)

10.10.4 Uranium (U)

10.10.5 Neptunium (Np)

10.10.6 Plutonium (Pu)

10.10.7 Americium (Am)

10.10.8 Curium (Cm)

10.10.9 Berkelium (Bk)

10.10.10 Californium (Cf)

10.10.11 Einsteinium (Es)

10.10.12 Fermium (Fm)

10.10.13 Mendelevium (Md)

10.10.14 Nobelium (No)

10.10.15 Lawrencium (Lr)

10.11 Superheavy Elements

10.11.1 Rutherfordium (Rf)

10.11.2 Dubnium (Db)

10.11.3 Seaborgium (Sg)

10.11.4 Bohrium (Bh)

- 10.11.5 Hassium (Hs)
- 10.11.6 Meitnerium (Mt)
- 10.11.7 Darmstadtium (Ds)
- 10.11.8 Roentgenium (Rg)
- 10.11.9 Copernicium (Cn)
- 10.11.10 Nihonium (Nh)
- 10.11.11 Flerovium (Fl)
- 10.11.12 Moscovium (Mc)
- 10.11.13 Livermorium (Lv)
- 10.11.14 Tennessine (Ts)
- 10.11.15 Oganesson (Og)

11. Molecules

11.1 Simple Diatomic Molecules

- 11.1.1 Hydrogen (H₂)
- 11.1.2 Oxygen (O₂)
- 11.1.3 Nitrogen (N₂)
- 11.1.4 Carbon monoxide (CO)
- 11.1.5 Hydrogen chloride (HCl)

11.2 Water and Related Compounds

- 11.2.1 Water (H₂O)
- 11.2.2 Hydrogen peroxide (H₂O₂)
- 11.2.3 Hydroxide ion (OH⁻)

11.3 Carbon Compounds

- 11.3.1 Methane (CH₄)
- 11.3.2 Ethane (C₂H₆)
- 11.3.3 Ethene (C₂H₄)
- 11.3.4 Ethyne (C₂H₂)
- 11.3.5 Glucose (C₆H₁₂O₆)

11.4 Salts

- 11.4.1 Sodium chloride (NaCl)
- 11.4.2 Potassium chloride (KCl)

11.4.3 Calcium carbonate (CaCO₃)

11.5 Acids and Bases

11.5.1 Hydrochloric acid (HCl)

11.5.2 Sulfuric acid (H₂SO₄)

11.5.3 Nitric acid (HNO₃)

11.5.4 Ammonia (NH₃)

12. Computer Sciences

12.1 Classical Computers

12.1.1 Hardware Components

12.1.1.1 CPU (Central Processing Unit)

12.1.1.2 GPU (Graphics Processing Unit)

12.1.1.3 RAM (Random Access Memory)

12.1.1.4 Storage Devices (HDD, SSD)

12.1.1.5 Motherboard

12.1.1.6 Input/Output Devices

12.1.2 Software Components

12.1.2.1 Operating Systems

12.1.2.2 Applications

12.1.2.3 Programming Languages

12.1.2.4 Drivers

12.1.2.5 Utilities

12.1.3 Computing Concepts

12.1.3.1 Binary System

12.1.3.2 Algorithms

12.1.3.3 Data Structures

12.1.3.4 Networking Basics

12.1.3.5 Security & Encryption

12.2 Quantum Computers

12.2.1 Quantum Hardware

12.2.1.1 Qubits

12.2.1.2 Quantum Gates

- 12.2.1.3 Quantum Circuits
- 12.2.1.4 Quantum Memory
- 12.2.1.5 Quantum Error Correction

12.2.2 Quantum Software

- 12.2.2.1 Quantum Algorithms
- 12.2.2.2 Quantum Programming Languages
- 12.2.2.3 Simulators
- 12.2.2.4 Hybrid Classical-Quantum Systems

12.2.3 Quantum Concepts

- 12.2.3.1 Superposition
- 12.2.3.2 Entanglement
- 12.2.3.3 Quantum Decoherence
- 12.2.3.4 Quantum Cryptography

13. BHR and WHR:

- 13.1 Black Hole Radiation (BHR)
- 13.2 White Hole Radiation (WHR)

14. Black and White Holes content (in atom level):

- 14.1 Black Hole
 - 14.1.1 The accretion disks
 - 14.1.2 The singularity
- 14.2 White Hole
 - 14.2.1 The singularity
 - 14.2.2 White hole radiation (WHR)

15. Education In NS*

15.1 Physics

15.1.1 Quantum Mechanics

15.1.1.1 Wave Functions

15.1.1.2 Schrödinger Equation

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

15.1.1.3 Heisenberg Uncertainty Principle

15.1.1.4 Quantum Entanglement

15.1.2 Classical Physics

15.1.2.1 Newtonian Mechanics

15.1.2.2 Electromagnetism

15.1.2.3 Thermodynamics

15.1.2.4 Fluid Dynamics

15.1.3 Universe and Space

15.1.3.1 Cosmology

15.1.3.2 Relativity (Special and General)

15.1.3.3 Astrophysics

15.1.3.4 Particle Physics

15.2 Chemistry

15.2.1 Matter

15.2.1.1 States of Matter

15.2.1.2 Properties of Matter

15.2.2 Atoms

15.2.2.1 Atomic Structure

15.2.2.2 Electron Configuration

15.2.3 Subatomic Particles

15.2.3.1 Protons, Neutrons, Electrons

15.2.3.2 Quarks and Leptons

15.2.4 Periodic Table

15.2.4.1 Groups and Periods

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

15.2.4.2 Element Properties

15.2.5 Physical Chemistry

15.2.5.1 Thermodynamics

15.2.5.2 Kinetics

15.2.5.3 Quantum Chemistry Basics

15.2.6 Electrochemistry

15.2.6.1 Redox Reactions

15.2.6.2 Electrolysis

15.2.7 Computational Chemistry

15.2.7.1 Molecular Modeling

15.2.7.2 Simulation Methods

15.2.8 Quantum Chemistry

15.2.8.1 Molecular Orbitals

15.2.8.2 Schrödinger Equation Applications

15.3 Biology

15.3.1 Bioinformatics

15.3.1.1 Sequence Analysis

15.3.1.2 Genomics and Proteomics

15.3.2 Quantum Biology

15.3.2.1 Photosynthesis Mechanisms

15.3.2.2 Enzyme Dynamics

15.3.3 Molecular Biology

15.3.3.1 DNA/RNA Structure and Function

15.3.3.2 Protein Synthesis

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

15.3.3.3 Cell Signaling

15.4 Earth Science

15.4.1 Meteorology

15.4.1.1 Weather Systems

15.4.1.2 Climate Dynamics

15.4.2 Oceanography

15.4.2.1 Ocean Currents

15.4.2.2 Marine Ecosystems

15.4.3 Astronomy

15.4.3.1 Planetary Science

15.4.3.2 Stellar Evolution

15.4.4 Environmental Science

15.4.4.1 Pollution Studies

15.4.4.2 Sustainability

15.4.5 Geology

15.4.5.1 Rock and Mineral Formation

15.4.5.2 Plate Tectonics

16. Education In Math

16.1 Pure Math

16.1.1 Algebra

16.1.2 Geometry

16.1.3 Topology

16.1.4 Number Theory

16.2 Applied Math

Written and created by Prof. Nils Efverman. This project is powered by FEUE inc.
FEUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

16.2.1 Differential Equations

16.2.2 Mathematical Modeling

16.2.3 Optimization

16.3 Mathematical Finance

16.3.1 Risk Analysis

16.3.2 Financial Modeling

16.4 Structures (Algebraic/Geometric)

16.4.1 Groups, Rings, Fields

16.4.2 Vector Spaces

16.5 Number Systems

16.5.1 Integers, Rationals, Reals, Complex Numbers

16.6 Computer Science

16.6.1 Algorithms and Data Structures

16.6.2 Computational Complexity

16.7 Foundations of Mathematics

16.7.1 Logic

16.7.2 Set Theory

16.7.3 Proof Techniques

16.8 Complex Analysis

16.8.1 Analytic Functions

16.8.2 Cauchy's Theorem

16.9 Cryptography

16.9.1 Symmetric and Asymmetric Encryption

16.9.2 Hash Functions

1. Introduction and information:

1.1 Introduction:

The nearest black hole is 1560 light years away.

This black hole is believed to have formed around 13.75 billion years ago. The name of this black hole is "Gaia BH1". Under Gaia BH1's lifetime, Gaia BH1 has an average intake of $3,15 \times 10^{23}$ kg of material. And if any material goes over the event horizon it never comes back. What we know is that when something has crossed the line from the outside it comes as if the object has stopped and never moves, but for the object comes it splits into pieces and a minute or so. When this piece comes to the singularity of Gaia BH1 (or other black hole) no one knows what happens, but that is what theory is for. So, we mostly know how a black hole can in the first place exist. The black hole creates when two supernovae collapse together, this happens when Big Bang exploded, and stars exploded and with the power off the force from Big Bang only then could two supernovae that exploded at the same time and collide. And now we know how a black hole creates we can look up what the black hole eats. A black hole can eat whatever is in the path of the black hole and this also includes the fourth-dimension time. When we know what it eats, we can check out what happens when we come singularly.

1.2 Information:

* NS = Natural Science

Natural Science has:

- Physics
- Chemistry
- Biology
- Earth Science

DISCLAIMER: THIS IS NOT ONLY FACTS, BUT ALSO A THEORY

For understanding WHEP as good as possible you need education in:

- Physics
- Chemistry
- Biology
- Earth Science

- Math

Education about physics: Chapter 15.1

Education about chemistry: Chapter 15.2

Education about biology: Chapter 15.3

Education about earth science: Chapter 15.4

Education about math: Chapter 16.

In physics you need to learn about: all quantum mechanics and physics, classical physics and about the universe and space. In chemistry you must learn about matter, atoms, subatomic particles, the periodic table, physical chemistry, electro chemistry, computational chemistry and quantum chemistry. In biology you must learn about: bioinformatics, quantum biology and molecular biology. In earth science you need to learn: geology, meteorology, oceanography, astronomy and environmental science. In the language of all science; math you need to learn about: pure math, applied math, mathematical finance, structures, number systems, computer science, foundations, complex analysis and cryptography.

2. Dimensions and Formulas:

So, we know first a singularity is an object that has broken all laws of physics.

2.1 Dimensions:

Singularity is an object that we do not know very much about, but this theory says that singularity is the only object in the universe that has five dimensions:

1. Length (x)
2. Width (y)
3. Height (z)
4. Time (t)
5. Wong (w)

This fifth dimension is a dimension that has broken laws of physics. This dimension is called Wong. Wong is like the deep of the singularity. Wong can you not see with your eyes but feel it directly. Our standard 3D spectrum can only see three of the totals and with tools the fourth. Wong has a curved depth that we can feel but not see. And in a long spectrum you can also feel the fourth dimension. The 5D spectrum is like a rainbow, we can only see the standard color and in the 5D spectrum we can only see standard dimensions. Wong is also impossible to see Wong because it is in the object and that is also why an object cannot only have a Wong dimension. Wong is also circle, but it has depth that breaks the law of physics and how we see our own 3D spectrum forever. Now we can understand better what happened when you meet the singularity in a black hole. When you come to the singularity that second you cannot feel, you cannot see, you do not remember anything. This is because you can only follow the laws of physics (else it comes be catastrophe of the balance in the universe), but an object can break the laws of physics and that is why this happens. When this ten second that feels you has died has ended then you come die and all material of you

be pieces the size of an atom. These 7×10^{27} atoms come travel true a tunnel and the diameter of an atom.

2.2 Gaia BH1 Formulas:

Gaia GH1 form can simplest be explained by this:

$$D_n = 176 \cdot (1.125)^n$$

$$C_n = \pi \cdot 176^2 \cdot (1.125)^{2n}$$

$$R_n = \pi^2 \cdot 176^3 \cdot t \cdot (1.125)^{2n+nt}$$

Or for making the fifth-dimension core for the Gaia BH1 singularity (if it was possible):
feue256.github.io/BH.html

3 Black Hole and White Hole Physics:

3.1 Gravitational Field Strength Equations:

And in the end, when the object has a diameter equal to that of an atom, then a white hole comes on the other side of the tunnel. A White Hole is created when the first material from the black hole comes. The creation of a white hole can be so long that it can be a black hole, and then in the middle of the tunnel it comes to create a super nova from the start, and it has a chance of creating infinity of black hole and destroying the world. But when the creation of a white hole successfully comes do the opposite of what a black hole does. A white hole also has singularity, but it has negative dimensions. Gravitational Field Strength Near the Singularity for a black hole is defined as:

$$g_{\text{black}}(r, w) = GM(r^2 + \gamma \sin^2(w))(1 - e^{-w^2})$$

Near the singularity of a black hole, space is curved not only in 3D but also warped by the Wong dimension. The $\gamma \cdot \sin^2(w)$ acts like a dimensional distortion buffer, preventing division by zero as $w \rightarrow 0$, while $(1 - e^{-w^2})$ models how the Wong dimension intensifies the curvature the deeper you go. Gravity here becomes "ultra-curved", but not infinite due to 5D constraints.

Gravitational Field Strength Near the Singularity for a white hole is defined as:

$$g_{\text{white}}(r, w) = GM(r^2 + \delta \cosh(w))(1 + \tanh(\beta w))$$

The gravitational field strength near a black hole singularity, extended with the Wong dimension, is defined as:

$$g(r, w) = G \cdot M \cdot r^2 \cdot e^w$$

Explanation:

A white hole emits matter and time, reversed from the black hole. The hyperbolic cosine $\cosh(w)$ exponential dimensional repulsion, increasing rapidly with Wong depth. The $\tanh(\beta w)$ term causes an energy rebound that stabilizes the repulsion. The negative sign flips the force direction: explosive instead of attractive.

3.2 Event Horizon and Singularities:

When you are in the tunnel behind the singularity with the Wong dimension your atoms then come feel that it has stop (if atoms could feel), but if you somehow could see it from outer space and if the tunnel was of glass then you would see the atoms in approximately 953.34 lightyears/secund. If you stand on the singularity, then if you also had a telescope and zoom in on Tellus then it would be the year 465 Anno Domini. Why is this happening? This is because the light has a speed of approximately 300 000 km/s (or 186 411 miles/s), so when we look at an object extremely far away in the universe then we see them as they were when the light left them. On the same set, if something could observe Tellus from Gaia BH1, it comes to see the past. The tunnel between the black hole and the white hole is very stable, but the first ten years after birth of the black hole is the tunnel extremely unstable. The start of the universe is what we all know is Big Bang, but it can be more truthful than what we know about our universe. It can be so that when the creation of a black hole and the unstable tunnel between the black hole and the white go into pieces and then a new black hole creates and go so infinity, then we have infinity of black holes. The first black hole is the biggest and it can eat another black holes. Now when it is only one black hole in the hole universe and from the spectrum of the hole universe, this black hole is extremely small like Big Bang. When so much material is in one place and cannot go to a white hole it becomes a white hole. The black hole comes in one Pico secunds ($1 \text{ ps} = 10^{-12} \text{ s}$) into a white hole. This explosion is exactly like Big Bang, but the first world with the black hole how could that black hole come to the first universe? The answer to that is:

The fourth dimension Time in nothing can create material. This material can create a force like a black hole and when it has had all that material it can create a black hole. That means all material in the whole universe is from time itself.

3.3 Types of Black Holes:

Our universe that we have at the latest 13,8 billion years has approximately 10^{21} black holes. That are three types of black holes what we know:

Stellar Black Holes
Intermediate-Mass Black Holes
Supermassive Black Holes

Stellar Black Holes creates when two supernovae collapse. This type is 3-20 times bigger than our sun.

Intermediate-Mass Black Holes are the middle size of black holes. This type of black hole is the rarest of the three normal black holes. This is 100-1000 times the size of our sun.

Supermassive Black Holes is the biggest type of black holes. This black hole is in nearly every galaxy in the universe. These black holes can be millions to billions of times bigger than our sun.

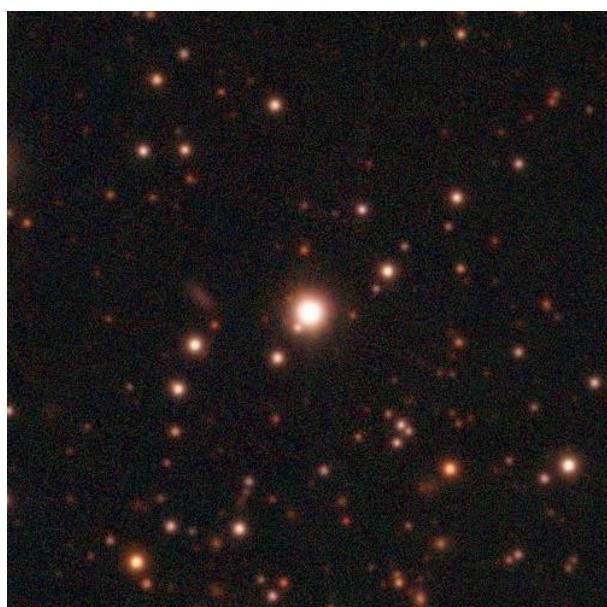
There are also two other types of black hole, but these ones are only theoretical:

Primordial Black Holes is a black hole that can only been created a brief time after Big Bang due to the frequent fluctuations in the early stage of the universe. The size of the Primordial Black Hole can variant all from extremely small to quite big.

Micro Black Holes are very very extremely small black holes. This Black Holes can be created in labs around the world. This would be especially useful för testing different theories.

3.4 White Hole Formation:

White Holes are terribly slow when they are spitting out everything from the black hole. This is because Supermassive Black Holes has a Micro White Hole. Every black hole has its opposite size than what the white hole has. The white hole is not a star, even if that is the opposite of what a black hole is. Every white hole is in another multiverse, and in this multiverse there are no black holes. When a one thousand multiverse it will become a mutivhope. Gaia BH1's real name is Gaia DR3 4373465352415301632. Image of Gaia BH1:



3.5 Gaia BH1 Case Study:

The tunnel between the black hole and white hole is made of the first material that the hole eats at the start of the black hole life. This process can take up years. Where does the black color come from in a black hole? A black hole is black because light cannot escape from within its event horizon. Since no light can be reflected or emitted back, the black hole appears completely dark to any observer. The accretion disk on a black hole can reach temperatures up to ten million Celsius (eighteen million Fahrenheit). When you look at a black hole from the side of the black hole you see the accretion disk bend on the top, but when you

look at the black hole above you see the accretion disk in a full circle. This is because the black holes gravity has a so enormous power and then it bends the light.

4. Cosmological Material Flow Analogy:

4.1 Waterfall Analogy for Black Holes:

The Supermassive Black Hole in the center of our galaxy Milky Way Sagittarius A* is the size

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.

FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

of 4,3 million times our sun. One of the biggest black holes we have ever seen is Ton 618. Ton 618 is sixty-six billion times our sun. Ton 618's diameter can fit ~40 of our solar systems. If you look at a black hole you will never see anything, go into the black hole. As an example:

Material in space is water. The black hole is a fall. Then would even water go into the black hole and create a waterfall. There the edge of the waterfall is an event horizon. After the water has fallen out from the edge it is faster than the speed of light down.

When you are going into the singularity, gravity increases. When you are far away from the event horizon you come to start the spaghettification. Spaghettification is when long come exceptionally long to you sardar into pieces. This is happening when the gravity with your head is lower than what it is with your feet. In the singularity all time ends. When something crosses the event horizon, there is no way back. Only radiation escapes from the black hole. This radiation consists of old particles from the tunnel between the black hole and the white hole. The radiation is extremely radioactive – up to one thousand times more dangerous than Polonium-210. This radiation is called BHR. We can never see BHR because the BHR come disappears into the accretion disk. When the white hole explodes like Big Bang all stuff come to stay in the white hole in approximately 5 seconds and then put it in another black hole and time starts again and time itself creates the start practice for a new universe starts it is new life. If a black hole were the size as an atom it would weigh like the biggest mountain on Tellus.

4.2 Black Hole Mergers and White Hole Events:

We cannot see white holes because the tunnel between black and white holes is going true another universe with only white holes in it. If something goes over the Event Horizon, it cannot come back (if it is not BHR), on same set cannot anything go into a white hole. When two black holes eat another black hole, it can go three separate ways:

1. The big black hole eats the small and it be a large black hole.
2. Two black holes of equal mass are orbiting each other. When a third black hole approaches, the system becomes unstable — eventually leading to the end of their lives through a massive merger called the closing of a black hole system.
3. A small black hole consumes a larger one. This causes a white hole to form for just one picosecond (10^{-12} seconds), before the system explodes.

5. Relativity and Metrics:

5.1 Schwarzschild Metric:

As solutions to the Einstein field equations, black holes are defined in general relativity. These metrics detail the shape of spacetime around different type of black holes. Specifically, the following are two instances of systems of these invariants:

The Schwarzschild-metrics for a non-rotating, neutral black hole can be defined by:

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$$

The Schwarzschild-metrics for a non-rotating, charged black hole can be defined by:

$$ds^2 = - \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right) dt^2 + \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$$

The Schwarzschild metric is a solution to Einstein's field equations in general relativity that describes the spacetime around a non-rotating, uncharged (neutral) black hole or a massive spherical object.

5.1 Reissner-Nordström Metric:

Reissner–Nordström solution (for charged black holes):

$$r_{\pm} = \frac{GM}{c^2} \pm \sqrt{\left(\frac{GM}{c^2}\right)^2 - \frac{GQ^2}{4\pi\varepsilon_0 c^4}}$$

5.2 Kerr Metric:

Kerr solution (for rotating black holes):

$$ds^2 = -\left(1 - \frac{2GMr}{\rho^2 c^2}\right) c^2 dt^2 - \frac{4GMar \sin^2 \theta}{\rho^2 c} dt d\phi + \frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \left(r^2 + a^2 + \frac{2GMa^2 r \sin^2 \theta}{\rho^2 c^2}\right) \sin^2 \theta d\phi^2$$

5.3 Hawking Radiation (BHR) and Temperature:

This equation below is Stephen Hawking temperature that a black hole gives out in BHR:

$$T_H = \frac{\hbar c^3}{8\pi GM k_B}$$

\hbar = reduced Planck constant

k_B = Boltzmann's constant

The White holes singularity also has the wong dimension in it. The BH in Gaia BH1 stands for Black Hole and the one in Gaia BH1 stands for the first black holes that is found in the Gaia project by ESA. The European Space Agency (ESA) serves as the European counterpart to NASA, focusing on space exploration, satellite development, and scientific research. ESA's head office is in Paris, France. ESA has twenty-two member countries. At the time of writing this is Sławosz Uznański-Wiśniewski from Poland on ISS (International Space Station). ISS is a space station 400 km (248.55 miles) above Tellus. ISS was starting construction in 1998. ISS has been mannded in nearly 25 years 24 hours a day in 7 day a week. ISS weighs 420 000 kg, and it has a speed of 28 000 km/h (17398.39/h). ISS size is approximately as big as a football field. Participating countries include NASA (USA), Roscosmos (Russia), ESA (Europe), JAXA (Japan), CSA (Canada), and many others.

6. Time, Material Creation, and Wong Dimension:

6.1 Time Creating Matter:

How can time create from nothing create something?

This is because the time near a white hole changes to the opposite of how time changes when the black hole changes it. The black hole can take the BHR before the BHR has exploded and save it into the event horizon, later the black hole can eat up the old event horizon and it can send in small packages of BHR to the white hole. Then the white hole takes this small packages with a force out from the singularity in the white holes center. Now the time can build a shield over the BHR, and it can build a stone. This process can take up to billions and billions of light years.

6.2 Wong Dimension Explanation:

What is the physics or mathematical nature of the Wong dimension, and how can it be experimentally or observationally verified to exist beyond theoretical speculation?

This dimension can only be found in the singularity of a black hole, which is why we cannot experiment with it. If we try, we come die by spaghettification. And the singularity of a black hole cannot exist without breaking the laws of physics, because whatever that is passes the event horizon that's not BHR (and it's new) can go into the singularity of the black hole. As example our sun cannot go into a space that is is this center of the black hole it is most be something behind. And it is with the wong and going into the tunnel to the white hole.

The wong dimension curves because the tunnel to the white hole is always not straight. The wong dimension is time like it always goes to the white hole if the tunnel is stable and if it is not stable it is spacelike. The wong and singularity are not objects it's something we can't understand. Wong is timelike but it's also curved. In the tunnel between a black and white hole all math and physics fail, and all that's not the transaction of atoms fails. When all fail, wong can't fail. When all else fails, Wong does not. The Wong dimension continues no matter what.

6.3 Tunnel Between Black and White Hole:

The tunnel between the black hole and the white hole is made of a unique material that naturally forms as element 119 to 139 in the periodic table, called Hongstone. If we ever discover Hongstone, it will officially become the 119 to 139th element, opening new doors to understanding black holes and white holes like never before. This discovery could allow us to build a stable tunnel connecting a micro black hole to a white hole. The tunnel must be made of Hongstone; otherwise, its instability would cause a supernova-like explosion, breaking Tellus (Earth) into tiny pieces. However, we face three major challenges. First, the Hongstone tunnel must have a diameter as small as a single atom. Second, we must create a micro black hole the size of an H₂O molecule but with the mass of Jupiter. Third, and most difficult, the white hole must be supermassive to perfectly balance with the micro black hole. Solving these problems would help us finally understand what the singularity at the center of a black hole truly is and how these cosmic phenomena work. For it to work it must be in quantum gravity. Quantum gravity isn't possible with the tool we have today. Quantum gravity can be possible with a quantum-computer-chamber on ISS with zero gravity and if it's in a black hole. This project is very dangers because we don't know how much the black hole comes eat and the force from the white hole, if something goes wrong the whole ISS can go into pieces, (but would that second after because they must be in a black hole). We must have a black hole for creating another black hole.

6.4 Hongstone:

Atomic Number: 139

Hypothetical Symbol (WHEP): Hsx

Isotopic Structure (WHEP): one proton, one neutron, 182,983 electrons, 5 HE-bosons

Hongstone is a hypothetical superheavy gaseous element, predicted to exist under extreme WHEP conditions and in the tunnel between the black hole and the white hole. Its minimal nucleus consists of a single proton and one neutron, while the 182,983 electrons form a highly diffuse, high-energy electron cloud. This configuration allows the atom to remain in a gaseous phase, despite its enormous electron count.

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.

FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

The 5 HE-bosons embedded in the nucleus mediate intense hyperelectric interactions, which govern the behavior of both the proton-neutron core and the surrounding electron cloud. In the gas phase, these interactions generate dynamic electron shell fluctuations and HE-quark excitations, producing a highly reactive yet non-condensed atomic state.

As a gas, Hongstone exhibits:

- Extreme hyperelectric conductivity: Electron clouds move freely while interacting with HE-bosons, creating transient fields.
- High quantum fluctuation rates: Valence electrons and HE-core couplings produce rapid spin-dependent excitations.
- Minimal nuclear influence: The single proton and neutron nucleus are dominated by HE-boson effects, enabling clear observation of WHE-quark dynamics.
- Unique gas-phase properties: Low interatomic cohesion allows free expansion, while HE-boson-mediated interactions maintain subtle quantum correlations between atoms.

WHEP predicts that gaseous Hongstone is the ultimate testbed for hyperelectric and nuclear quantum phenomena, where extreme electron populations and HE-boson interactions can be studied without solid-state or liquid-phase interference. Its gaseous state provides direct experimental access to WHE-quark excitations and HE-core energy shifts across an isolated atomic system.

6.5 Quantum physics:

Hongstone's strength is due to its structure using quantum bits for stability on the Wong scale. Unlike regular atomic bonds, Hongstone forms connections that are not held together by traditional forces like electromagnetism, but by entangled qubit states that only exist under Wong dimensional curvature. These qubit bonds allow the tunnel walls between the black hole and white hole to self-repair and resist spaghettification, even as the fabric of spacetime collapses around them. Without this quantum-layered strength, any normal matter would disintegrate instantly under the stress of Wong field fluctuations. That's why Hongstone isn't just a material, it's a quantum-stabilized state of existence. Hongstone has come to be the first element with computers and quantum-physics in an element. The Black Holes we can create with hongstone can only last a secund before the explode. Hongstone can also be used as the main material in CPU's for quantum computers. This CPU's can make and find patters in what ever you give then. This is only going to be possible with quantum computers, and they most have this special hongstone CPU's. This computers can solve math problems that we see as impossible, but this computers can solve this problem in seconds. This can also lead to a world of AI that's so smart that it can do things we believe is thing we think of because today we don't understand how to do this thing. If hongstone creates or find's then nobody is allowed to do anything with it because in specific condition can it open a portal to another dimension and that's would be extremely dangerous. We don't know the conditions needed for this to happen and before that it must be illegal to owe it because of this danger that can happen. What we know may be right, but we could still be wrong.

7. Pico Doom Virus (PDV) and Cosmic Implications:

7.1 PDV Behavior in Black Hole Singularities:

The Black Hole Radiation (BHR) is not only a source of extreme ionizing radiation but also a carrier of a theoretical pathogen referred to as Pico Doom Vires (PDV). PDV is a bio-quantum agent with an effective reproduction number (R_0) of infinity. Unlike conventional viruses, PDV does not require cellular machinery to replicate; instead, it propagates through fundamental interactions at the subatomic level, targeting nucleic acid structures (DNA/RNA) directly. Upon exposure—whether airborne, dermal, or via particulate carriers—all known biological systems experience immediate and irreversible molecular collapse. This makes PDV not only the most infectious agent theoretically proposed but also a universal terminator of life, functioning beyond the constraints of biochemistry as we know .

The presence of PDV within BHR emissions suggests that certain black holes may serve not only as gravitational singularities but also as distribution hubs for entropic viral fields, potentially reshaping our understanding of cosmic sterilization events. This raises serious implications for interstellar exploration, planetary quarantine protocols, and the theoretical upper bounds of biological survivability. The Pico Doom Virus (PDV) has a theoretical survival probability of 0% under standard conditions in the observable universe due to its extreme bio-instability outside of specialized quantum environments. However, within the singularity of a black hole, this survival probability approaches 100%.

This anomaly arises due to the black hole's ability to warp quantum physical laws. As spacetime and quantum fields are intensely curved near or within the event horizon, standard probabilities—such as decay or disintegration rates—no longer apply in their classical form. Since PDV is classified as a bio-quantum virus, it is intrinsically entangled with quantum mechanical behavior rather than classical biological constraints.

7.2 Implications for Interstellar Exploration:

As quantum mechanics is deformed near the singularity, so too is the statistical framework governing particle and viral decay. This deformation results in a unique condition in which PDV's quantum structure stabilizes rather than deteriorates, effectively reversing its normal survival odds. The PDV is the size of a picometer.

The black hole can grow because the BHR can build a magnetic field and when the PDV particle collides with the magnetic field it can create an intense gravitational field and makes stone that the black hole eats and then grows.

7.3 Impact on DNA/RNA Structure:

The interaction of PDV with DNA and RNA in biological systems is one of the most significant aspects of this theoretical virus. Its bio-quantum nature suggests that it could fundamentally alter the structure and function of genetic material in ways we have yet to fully comprehend.

1. Direct Molecular Collapse:

- PDV targets nucleic acids (DNA and RNA) directly at a quantum level, bypassing the need for a host cell. This would result in irreversible molecular collapse upon exposure, as the virus interacts with the subatomic structure of

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.

FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

the nucleic acids.

- The viral particles could induce quantum fluctuations in the chemical bonds that hold the nucleotides together, causing them to break apart, resulting in severe genetic instability.

2. Quantum Entanglement with DNA/RNA:

- Due to its quantum nature, PDV could be capable of entangling with DNA or RNA molecules, causing the virus to share the same quantum states as the nucleic acids. This interaction would likely destabilize the double helix structure of DNA, potentially leading to structural distortions and breaks in the strands.
- The molecular collapse could disrupt the genetic code, causing random mutations or deletions of genetic sequences that are essential for cellular function.

3. Altered Replication and Transcription:

- PDV-induced changes in the genetic material could block or distort DNA replication and RNA transcription. This disruption would prevent normal gene expression and protein synthesis, leading to cellular dysfunction.
- The virus may also alter the three-dimensional folding of RNA molecules, which is essential for their proper functioning in protein synthesis.

4. Epigenetic Changes:

- PDV might induce epigenetic changes in DNA that affect how genes are expressed without altering the underlying sequence. For example, it could modify DNA methylation patterns or histone modifications, leading to long-term changes in gene expression.
- Such epigenetic shifts could permanently affect an organism's biology, potentially rendering the infected species nonviable or causing their extinction.

5. Unpredictable Biological Effects:

- Since PDV operates beyond the typical constraints of biochemistry, the exact impact it would have on RNA and DNA is unpredictable. In some cases, it could lead to revolutionary changes in genetic makeup, potentially altering evolutionary processes or creating entirely new forms of life.
- This unpredictability introduces a new level of biological uncertainty when dealing with cosmic pathogens, where traditional biological models would not suffice to explain the virus's behavior.

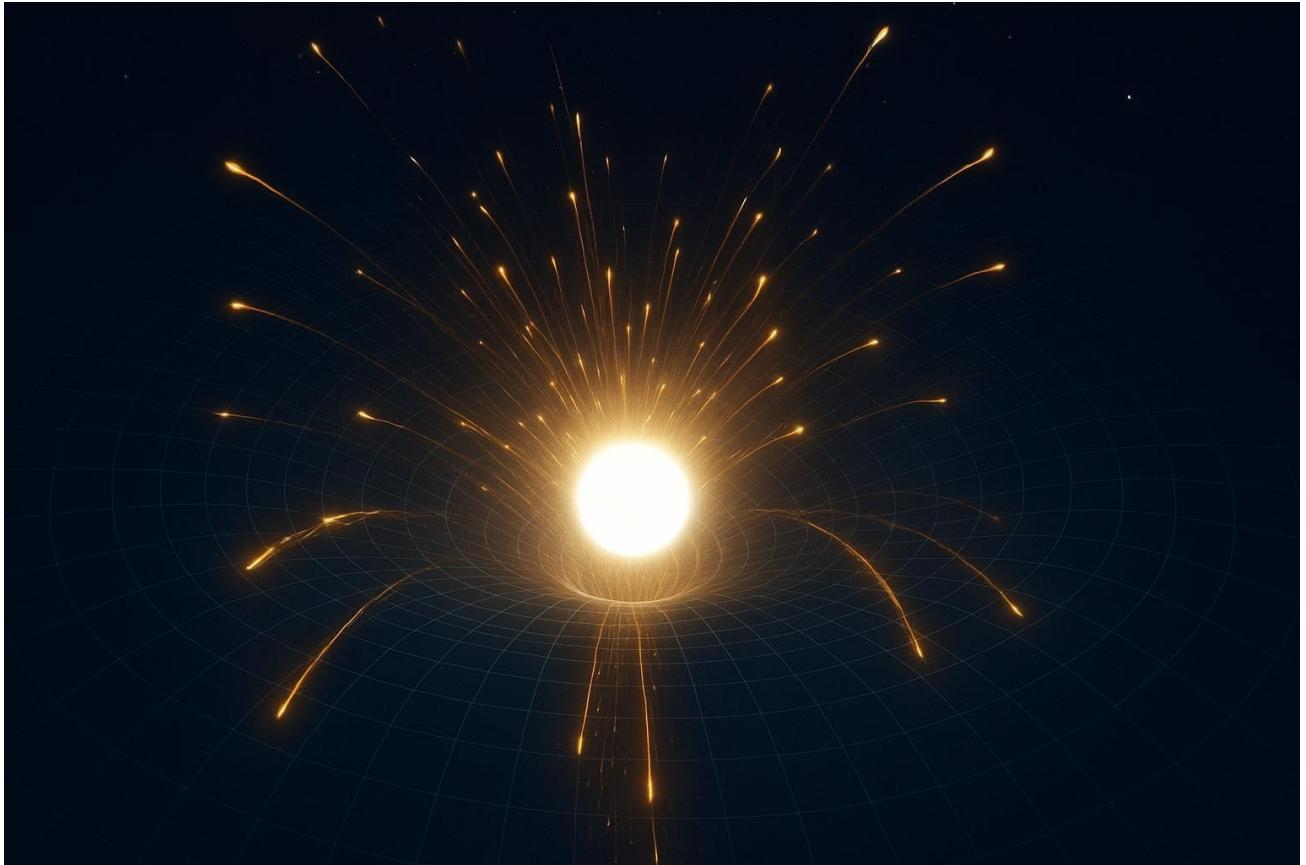
8. Online data:

8.1 Description of how light bends:

Data Source:

https://en.wikipedia.org/wiki/Black_hole#/media/File:Black_Hole_Shadow.gif

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
 FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>



The GIF is playing a short video that describes how light bends near a black hole. The yellow is the light, and the black is the singularity.

This tells us how light waves react to a static black hole and how the black holes gravity reacts to the light waves. This GIF is from the wiki article en.wikipedia.org/wiki/Black_hole under the label “Physical properties”

8.2

particles

Data Source:

Quantum
white holes compared

in the image ,
experience extreme
absorbed as in
ejected

trajectories.
such as

Description of
how
quantum
react to white holes:

particles behave differently near
to conventional matter
interactions. As
illustrated
particles approaching a white hole
repulsive energy flows: instead of being
black holes, they are
along highly
directional
Quantum effects
wavefunction dispersion,

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

tunneling, and near the event horizon of a white density distribution and probable maintain stable orbits and are forced quantum radiation patterns.

entanglement shifts may occur hole. The data visualization shows particle ejection paths, indicating that particles cannot outward, which may produce observable

8.3 Description of quantum particles / the standard model of particle physics:

Data Source :

<https://live-production.wcms.abc-cdn.net.au/2c70e6d5c36be4979a8f1be0e10d3810?src>

This image tells us two separate groups of particles in the Standard Model. The first group is the fermions, and the second group is the bosons. The fermions are the particles that make all matter in the universe. The bosons are the particles that are force carriers or exchange particles. In the fermions there are two subgroups in the fermions, the quarks, and the leptons. The bosons have the same principle; they also have two separate groups as the fermions. They have the gauge bosons and the scale boson. The thing that tells them apart is the spin of the particles. The fermions have a spin of $\frac{1}{2}$ and the bosons have a spin of one or for the scalar boson is zero.

9.The standard model of particle physics:

9.1Fermions:

9.1.1Quarks:

9.1.1.1 Up quark:

The up quark is one of the most popular quarks. It has noticeable mass at only approximately $2,3 \text{ MeV}/c^2$. This makes the up quark very stable. That's why the up quark is extremely popular in the universe. The up quark has a spin of $\frac{1}{2}$ and a charge of $\frac{2}{3}$. The symbol for the up quark is "u", the first letter of up.

9.1.1.2 Down quark:

The down quark is a quark with the symbol "d" because that is the first letter of the quark full name (down). It's the most popular quark on the same place as the up quark. This quark has the same spin as all quarks ($\frac{1}{2}$) and a charge of $-\frac{1}{3}$. The mass of the down quark is a little bite heavier than the up quark on approximately $4.8 \text{ MeV}/c^2$. It's still stable. The down, strange and the bottom quark always have the same charge, and the up, charm and top quarks also have the same charge.

9.1.1.3 Charm quark:

The charm quark is a heavier quark with the symbol "c", the first letter of charm. It has a spin of $1/2$ and a charge of $2/3$, just like the up quark. The mass of the charm quark is approximately $1.27 \text{ GeV}/c^2$, making it much heavier than the up and down quarks. The charm quark is less common in everyday matter but appears in high-energy particle collisions.

9.1.1.4 Strange quark:

The strange quark has the symbol "s", the first letter of strange. It has a spin of $1/2$ and a charge of $-1/3$, the same as the down and bottom quarks. The mass of the strange quark is around $95 \text{ MeV}/c^2$. Strange quarks are responsible for forming "strange" particles and appear in cosmic events and particle experiments.

9.1.1.5 Top quark:

The top quark is the heaviest quark and has the symbol "t", from top. Its spin is one-half, and its charge is two-thirds, like the up and charm quarks. The mass of the top quark is about $173 \text{ GeV}/c^2$. It is extremely unstable and decays almost immediately after formation, so it is only seen in high-energy accelerators.

9.1.1.6 Bottom quark:

The bottom quark has the symbol "b", from bottom. It has a spin of $1/2$ and a charge of $-1/3$, like the down and strange quarks. Its mass is approximately $4.18 \text{ GeV}/c^2$. The bottom quark is heavier than the strange and down quarks but lighter than the top quark. It forms particles called bottom-hadrons, such as B-mesons, and is important in studies of quantum effects like CP-symmetry. In particle physics CP means "Change Parity".

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.

FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

9.1.1.7 WHE Quark:

The WHE quark is a hypothetical particle introduced in the context of the White Hole Existence Principle (WHEP). It is theorized to be a key component of the exotic matter associated with white holes and their unique properties. The WHE quark may interact with the newly defined hyperelectromagnetic force (HEF), mediated by the HE-boson.

Properties:

- Charge: The WHE quark could possess a unique type of hyper electric charge that is linked to the hyperelectromagnetic force (HEF). This charge would allow it to interact with the HE-boson, the mediator of the hyperelectromagnetic force.
- Mass: The mass of the WHE quark may vary depending on its interaction with gravitational fields and exotic matter near white holes.
- Spin: Like other quarks, the WHE quark is predicted to have spin one-half.
- Interactions: The WHE quark is theorized to interact with other quarks via the strong nuclear force, mediated by gluons, but it may also participate in hyperelectromagnetic interactions through the HE-boson. Furthermore, it may interact with gravitons, especially in the environment of white holes.

Theoretical Role:

- White Hole Matter: The WHE quark is expected to be a fundamental particle that contributes to the exotic matter found inside white holes. It may help explain the emission of energy and particles from these objects.
- HEF Interaction: Due to its unique hyper electric charge, the WHE quark would play a critical role in mediating interactions within the hyper electromagnetic force (HEF), which is described by the HE-boson. This interaction could explain the formation and dynamics of exotic matter within and around white holes.

Mathematical Representation:

The WHE quark is included in the quantum field theory of WHEP. A possible Lagrangian term for the WHE quark's interactions is as follows:

$$\begin{matrix} L \\ W \\ H \\ E \\ = g \\ W \\ H \\ E \\ \psi \\ - \end{matrix}$$
$$\begin{matrix} W \\ H \\ E \\ \gamma \\ \mu \\ \psi \\ W \\ H \\ E \\ \cdot H \\ E \\ \mu \\ +g \\ W \\ H \\ E \\ -s \\ t \\ r \\ o \\ n \\ g \\ \psi \\ - \end{matrix}$$
$$\begin{matrix} W \\ H \\ E \\ \gamma \\ \mu \\ \psi \\ W \\ H \\ E \\ \cdot G \\ \mu \end{matrix}$$

In this equation:

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc.
FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on:
<https://tinyurl.com/3xceeda5>

- $\frac{g}{W}$ is the coupling constant for the interaction between the WHE quark and the H
 E hyperelectromagnetic field.
- $gWHE\text{-strong}$ is the coupling constant for the strong interaction between the WHE quark and gluons.
- $\frac{H}{E}$ represents the hyperelectromagnetic field.
 μ
- G
 μ is the gluon field, which mediates the strong interaction.

This Lagrangian expresses the coupling between the WHE quark and the HE-boson, as well as its interaction with gluons.

9.1.2 Leptons:

9.1.2.1 Electron:

The electron is a fundamental particle, classified as a lepton. It has a negative electric charge and is not composed of smaller particles like quarks. The electron has a spin of $\frac{1}{2}$ and a mass of approximately $0.511 \text{ MeV}/c^2$.

9.1.2.2 Muon:

The muon is like the electron but has a much greater mass. It is unstable and decays into electrons, neutrinos, and other particles. Its mass is approximately $105.7 \text{ MeV}/c^2$ and it has a spin of $\frac{1}{2}$.

9.1.2.3 Tau:

The tau is the heaviest of the charged leptons. It is unstable and decays into lighter particles such as electrons or muons and neutrinos. The tau's mass is $1.777 \text{ GeV}/c^2$ and it has a spin of $\frac{1}{2}$.

9.1.2.4 Electron Neutrino:

The electron neutrino is a neutral lepton associated with the electron. It has an exceedingly small mass, which is still under investigation, and does not carry any electric charge. Its spin is $\frac{1}{2}$.

9.1.2.5 Muon Neutrino:

The muon neutrino is a neutral lepton associated with the muon. It has an exceedingly small mass, and like other neutrinos, interacts weakly with matter. Its spin is $\frac{1}{2}$.

9.1.2.6 Tau Neutrino:

The tau neutrino is a neutral lepton associated with the tau particle. It has an exceedingly small mass and interacts very weakly with matter. Its spin is $\frac{1}{2}$.

9.2 Bosons:

Bosons are particles that carry the fundamental forces of nature. They have integer spin (0, 1, 2...) and play a critical role in mediating interactions between other particles.

9.2.1 Gauge Bosons:

Gauge bosons are the force-carrier particles that mediate the fundamental forces in the Standard Model of particle physics. These include the gluon, photon, W bosons, and the Z boson.

9.2.1.1Gluon:

The gluon is the force carrier for the strong nuclear force, which binds quarks together inside protons and neutrons. Gluons are massless and interact only with quarks and other gluons. This interaction is fundamental to the structure of atomic nuclei.

9.2.1.2Photon:

The photon is the force carrier for electromagnetic interactions. It is responsible for electromagnetic radiation (such as light) and has zero mass and no electric charge. Photons mediate the interaction between electrically charged particles.

9.2.1.3Z Boson:

The Z boson mediates the weak nuclear force and is electrically neutral. It is involved in processes like neutrino interactions and beta decay. The Z boson, along with the W boson, is crucial in explaining phenomena such as the decay of unstable particles.

9.2.2.4W Boson:

The W boson mediates the weak nuclear force, which is responsible for particle decay processes like beta decay. There are two types of W bosons:

- W^+ : The positively charged W boson.
- W^- : The negatively charged W boson.
The W boson plays a key role in interactions that change the type (or flavor) of particles.

9.2.2.5Gravitons:

The graviton is the one of two new thing in the boson group for the subatomic particle system. The graviton is a force carrier. This new boson has an unlikely spin of two that is specific to the graviton boson. This boson come have the force of gravity. This particle is not really used in the universe, but it has a minor impact on the hongstones atomic core.

9.2.2.6HE boson:

The HE-boson (Hyperelectromagnetic Boson) is a hypothetical particle introduced to mediate the hyperelectromagnetic force (HEF) within the framework of the White Hole Existence Principle (WHEP). The HE-boson is theorized to play a significant role in interactions involving exotic matter and WHE quarks, particularly in the environment of white holes.

Properties:

- Charge: The HE-boson is expected to be electrically neutral but is involved in the mediation of the hyperelectromagnetic force through its interaction with WHE quarks and other exotic particles.
- Mass: The HE-boson may possess a non-zero mass, which could vary depending on the gravitational conditions within a white hole. The mass is speculated to be relatively small, though it could fluctuate due to interactions with gravitational fields.
- Spin: The HE-boson is theorized to have spin one, classifying it as a gauge boson. This means it is responsible for mediating the hyperelectromagnetic force, much like the photon mediates the electromagnetic force or the gluon mediates the powerful force.
- Interaction: The HE-boson mediates the hyperelectromagnetic force (HEF), which governs the interactions of WHE quarks and other exotic matter. It is a critical component of the theoretical structure of white holes and the exotic matter contained within them.

Theoretical Role:

- Hyperelectromagnetic Force: The HE-boson is hypothesized to be the force carrier for the hyperelectromagnetic force (HEF). This force is an extension of the electromagnetic force, theorized to act on particles with hyper electric charge, such as the WHE quark. The HE-boson would be responsible for transferring energy between these particles.
- White Hole Dynamics: In the context of white holes, the HE-boson could explain the interactions between exotic matter and the energy released by these objects. The HE-boson is theorized to be central to the unique radiation and particle emission processes in white hole environments.
- Potential Gravitational Coupling: Although the HE-boson is primarily associated with the hyperelectromagnetic force, there may also be indirect interactions with gravitational fields, especially in the extreme conditions of a white hole.

Mathematical Representation:

The interaction of the HE-boson with WHE quarks and other exotic particles can be described through the following Lagrangian term:

$$\begin{matrix} L \\ H \\ E \\ = g \\ H \\ E \\ \psi \end{matrix}$$
$$\begin{matrix} W \\ H \\ E \\ \gamma \\ \mu \\ \psi \\ W \\ H \\ E \\ \cdot H \\ E \\ \mu \end{matrix}$$

Where:

$\frac{g}{E}$ is the coupling constant for the interaction between the WHE quark and the HE-boson.

- $\begin{matrix} \psi \\ \bar{W} \\ H \\ E \end{matrix}$ and $\begin{matrix} \psi \\ \bar{W} \\ H \\ E \end{matrix}$ are the WHE quark fields.
- $HE\mu HE\mu HE\mu$ represents the hyperelectromagnetic field, mediated by the HE-boson.

This term describes the interaction of the WHE quark with the hyperelectromagnetic force, mediated by the HE-boson.

9.2.1 Scalar Boson

A scalar boson is a type of elementary particle that has no directionality of spin, meaning it has a spin of zero. Unlike vector bosons, which mediate forces and have a directional spin, scalar bosons are theorized to interact differently, primarily in the context of mass generation for fundamental particles. Scalar bosons are part of many quantum field theories, and they often play a critical role in the mechanism that gives mass to other particles in the Standard Model.

9.2.1.1 Higgs Boson

The Higgs boson is a specific type of scalar boson that is central to the Higgs mechanism in the Standard Model of particle physics. It was proposed in the 1960s by physicist Peter Higgs and created by Prof. Nils Efverman. This project is powered by FÉUE inc.

FÉUE office website: feue256.github.io. Learn more about Gaia BH1 on:

<https://tinyurl.com/3xceeda5>

Higgs and others, as part of the mechanism that explains how fundamental particles acquire mass.

Properties of the Higgs Boson:

- Spin: The Higgs boson has a spin of zero, meaning it is a scalar boson.
- Mass: The Higgs boson has a relatively large mass compared to other fundamental particles, approximately $125 \text{ GeV}/c^2$.
- Charge: The Higgs boson is electrically neutral.

Role in the Standard Model:

The Higgs boson is the quantum excitation of the Higgs field, a scalar field that permeates all of space. The interaction of particles with the Higgs field gives them mass. This process is essential because, without the Higgs field, particles such as the W and Z bosons (responsible for the weak nuclear force) would be massless, and the weak force would not behave as we observe in nature.

The Higgs mechanism explains the mass of fundamental particles by interacting with them. The stronger a particle interacts with the Higgs field, the more massive it becomes. For example, the top quark, which has a large mass, interacts very strongly with the Higgs field, while the photon (the force carrier for electromagnetism) does not interact with the Higgs field at all, and remains massless.

Discovery:

The discovery of the Higgs boson was one of the most significant milestones in modern physics. It was detected in July 2012 by scientists at the Large Hadron Collider (LHC) at CERN, confirming the existence of the Higgs field and validating the Standard Model. The discovery was awarded the 2013 Nobel Prize in Physics to Peter Higgs and François Englert.

Implications:

The detection of the Higgs boson not only confirmed the existence of the Higgs field but also solidified our understanding of the mass generation mechanism in the universe. It has profound implications for further research in particle physics, particularly in areas such as:

- Quantum Field Theory (QFT): The Higgs boson provides crucial evidence for the existence of scalar fields in quantum field theory.
- New Physics: While the discovery of the Higgs boson completed the Standard Model, it has also raised questions about what lies beyond it, such as potential connections to dark matter, supersymmetry, and other exotic theories.

Conclusion:

The Higgs boson is a pivotal component of the Standard Model and is critical for understanding the mass of elementary particles. Its discovery has profound implications not only for particle physics but also for the fundamental understanding of the universe's structure and the forces at play within it.¹⁰

10. Atoms

Atoms are what everything is made of, on a bigger level them quarks. All atoms are sorted in the periodic table of elements. Picture of the periodic table of elements:

Periodic Table of the Elements																		
1 IA H Hydrogen 1.008 22.990	2 IIA Be Beryllium 9.0122	3 Na Sodium 22.990 22.990	4 Mg Magnesium 24.310	5 VIB Sc Scandium 44.956	6 VB Ti Titanium 47.887	7 VIB Cr Chromium 50.942	8 VIIA Mn Manganese 54.938	9 VIIA Fe Iron 55.845	10 VIIA Co Cobalt 58.933	11 VIIA Ni Nickel 58.673	12 VIIA Cu Copper 63.546	13 IIIA Zn Zinc 65.38	14 IIIA Ga Gallium 69.723	15 IIIA Ge Germanium 72.630	16 IIIA As Arsenic 74.922	17 IIIA Se Selenium 78.911	18 IIIA Br Bromine 79.904	19 VIIA Kr Krypton 83.780
20 VIIA Ca Calcium 40.078	21 IIIB Sc Scandium 44.956	22 IVB Ti Titanium 47.887	23 VIB V Vanadium 50.942	24 VIIA Cr Chromium 51.996	25 VIIA Mn Manganese 54.938	26 VIIA Fe Iron 55.845	27 VIIA Co Cobalt 58.933	28 VIIA Ni Nickel 58.673	29 VIIA Cu Copper 63.546	30 VIIA Zn Zinc 65.38	31 VIIA Ga Gallium 69.723	32 VIIA Ge Germanium 72.630	33 VIIA As Arsenic 74.922	34 VIIA Se Selenium 78.911	35 VIIA Br Bromine 79.904	36 VIIA Kr Krypton 83.780		
37 VIIA Rb Rubidium 80.448	38 VIIA Sr Strontium 80.448	39 VIIA Y Yttrium 88.904	40 IIIB Zr Zirconium 91.230	41 IIIB Nb Niobium 91.946	42 VIIA Mo Molybdenum 95.946	43 VIIA Tc Technetium 95.946	44 VIIA Ru Ruthenium 96.946	45 VIIA Rh Rhodium 96.946	46 VIIA Pd Palladium 96.946	47 VIIA Ag Silver 107.87	48 VIIA Cd Cadmium 112.42	49 VIIA In Indium 113.42	50 VIIA Sn Tin 118.70	51 VIIA Sb Antimony 121.76	52 VIIA Te Tellurium 121.46	53 VIIA I Iodine 126.90	54 VIIA Xe Xenon 131.30	
55 Cs Cesium 132.911	56 Ba Barium 137.33	57-71 Lanthanides Ce Lanthanum 140.91	72 VIIA Hf Hafnium 178.09	73 VIIA Ta Tantalum 183.84	74 VIIA W Tungsten 183.95	75 VIIA Re Rhenium 191.23	76 VIIA Os Osmium 191.23	77 VIIA Ir Iridium 191.23	78 VIIA Pt Platinum 191.00	79 VIIA Au Gold 196.97	80 VIIA Hg Mercury 201.59	81 VIIA Tl Thallium 204.38	82 VIIA Pb Lead 207.23	83 VIIA Bi Bismuth 209.98	84 VIIA Po Polonium (209)	85 VIIA At Astatine (210)	86 VIIA Rn Radium (222)	
87 Fr Francium (223)	88 Ra Radium (226)	89-103 Actinides Rf Rutherfordium (267)	104 VIIA Db Dubnium (268)	105 VIIA Sg Seaborgium (269)	106 VIIA Bh Bohrium (260)	107 VIIA Hs Hassium (261)	108 VIIA Mt Meitnerium (262)	109 VIIA Ds Darmstadtium (263)	110 VIIA Rg Roentgenium (264)	111 VIIA Cn Copernicium (265)	112 VIIA Nh Nihonium (266)	113 VIIA Fl Flerovium (267)	114 VIIA Mc Livermorium (268)	115 VIIA Lv Tennessine (269)	116 VIIA Ts Oganesson (269)	117 VIIA Og Livermorium (269)	118 VIIA Og Oganesson (269)	
57 La Lanthanum 139.90	58 Ce Cerium 140.01	59 Pr Praseodymium 140.91	60 Nd Neodymium 140.91	61 Pm Promethium 141.91	62 Sm Samarium 141.91	63 Eu Europium 141.91	64 Gd Gadolinium 141.91	65 Tb Terbium 141.91	66 Dy Dysprosium 142.91	67 Ho Holmium 142.91	68 Er Erbium 142.91	69 Tm Thulium 142.91	70 Yb Ytterbium 142.91	71 Lu Lutetium 142.91				
89 Ac Actinium (227)	90 Th Thorium (232)	91 Pa Protactinium (231)	92 U Uranium (238)	93 Np Neptunium (237)	94 Pu Plutonium (239)	95 Am Americium (243)	96 Cm Curium (247)	97 Bk Berkelium (247)	98 Cf Californium (251)	99 Es Einsteinium (257)	100 Fm Fermium (257)	101 Md Mendelevium (258)	102 No Nobelium (258)	103 Lr Lawrencium (264)				

10.1 Alkali Metals:

10.1.1 Hydrogen (H):

Hydrogen is the simplest and most fundamental element in the universe. It has atomic number 1 and in its most common isotope (protium) it consists of a single proton and one electron. The proton is composed of two up quarks (u) and one down quark (d) (uud), bound together by gluons through the strong nuclear force. Unlike many other nuclei, hydrogen-1 contains no neutron, which makes it the most elementary example of baryonic matter. Its electron, belonging to the lepton family, is not made of quarks but is an indivisible elementary particle.

Chemically, hydrogen is highly reactive and often forms compounds with electronegative elements such as oxygen, nitrogen, and halogens. In stars, hydrogen undergoes nuclear fusion, where four protons fuse under extreme conditions to form helium nuclei. This fusion process is the foundation of stellar energy production and the origin of heavier elements in the universe.

From the perspective of the White Hole Existence Principle (WHEP), hydrogen plays a key role as the most abundant element in the cosmos. White holes are predicted to generate hyperelectric fields mediated by the HE-boson, which may interact with hydrogen nuclei at the quantum level. The proton's quark structure could temporarily shift under such conditions, possibly involving hypothetical WHE-quarks, leading to exotic emissions of both electromagnetic radiation and novel particle streams. This makes hydrogen not only the

basis of ordinary cosmic matter but also a critical testing ground for how matter behaves near white holes and under extreme hyperelectric influence.

In summary, hydrogen is both the building block of stars and galaxies and a cornerstone in WHEP, bridging ordinary matter and exotic quantum phenomena in white hole environments.

10.1.2 Lithium (Li)

Lithium, with atomic number 3, is the lightest of the alkali metals and one of the most essential elements for both chemistry and modern technology. Its nucleus in the most stable isotope, Lithium-7, consists of three protons and four neutrons. Each proton is composed of two up-quarks and one down-quark (uud), while each neutron consists of two down-quarks and one up-quark (udd). In total, the lithium-7 nucleus contains seventeen up-quarks and sixteen down-quarks, bound together by gluons via the strong interaction. The three electrons surrounding the nucleus are elementary leptons and are not made of quarks.

Chemically, lithium is highly reactive, especially with water, where it produces hydrogen gas and lithium hydroxide. Due to its low density, lithium can even float on water. Its position as the first true alkali metal gives it properties like sodium and potassium, but it is less reactive compared to its heavier counterparts. Lithium compounds are widely used in rechargeable batteries, ceramics, and in psychiatric medicine as a mood stabilizer.

From the WHEP perspective, lithium has a unique role. Being both light and stable, it is considered one of the "transition" elements between hydrogen/helium (formed in the Big Bang) and heavier elements created in stars. In white hole environments, lithium nuclei are hypothesized to undergo transformations under the influence of hyperelectric fields mediated by the HE-boson. The delicate balance between protons and neutrons in lithium makes it a possible candidate for quark-level resonance states, where the inclusion of a WHE-quark could momentarily stabilize exotic isotopes. Such states might explain anomalous radiation signatures near white holes that cannot be accounted for by hydrogen alone.

In summary, lithium bridges the gap between the simplicity of hydrogen and the complexity of heavier elements. In WHEP, it may act as an experimental cornerstone to study how ordinary baryonic matter transitions into exotic matter under extreme quantum-gravitational conditions.

10.1.3 Sodium (Na)

Sodium, with atomic number 11, is one of the most well-known alkali metals. In its most common isotope, Sodium-23, the nucleus contains eleven protons and twelve neutrons. Each proton is built of two up-quarks and one down-quark (uud), while each neutron is composed of two down-quarks and one up-quark (udd). This means the sodium-23 nucleus contains a total of sixty-seven up-quarks and sixty-eight down-quarks, all bound by gluons in the strong nuclear force. Surrounding the nucleus are eleven electrons, arranged in the configuration [Ne]3s¹, with a single valence electron in the outermost shell. This lone electron is what makes sodium highly reactive.

Chemically, sodium reacts violently with water, producing hydrogen gas and sodium hydroxide, often with enough heat to ignite the hydrogen. Its salts, such as sodium chloride (NaCl), are essential to life and are the foundation of biological ionic balance. Sodium ions play a critical role in nerve impulse transmission and cellular processes, making sodium indispensable for biology. Despite its intense reactivity as a pure element, in compounds it is both stable and life supporting.

In the WHEP framework, sodium plays a significant role as a mid-level element between the exceptionally light alkali metals and the much heavier ones. The relatively balanced number of protons and neutrons in sodium-23 provides a testing ground for studying HE-boson field interactions. The hyperelectric influence of white holes is theorized to distort the proton-neutron ratio in certain isotopes, possibly leading to transient states where WHE-quarks integrate into nuclear matter. These exotic sodium isotopes could contribute to unexplained emission lines observed near white hole candidates.

From an astrophysical perspective, sodium is notable for its strong yellow doublet emission (the sodium D-lines), which is often used in spectroscopy to identify its presence in stars and nebulae. Under WHEP, these emissions may be further modified by white hole fields, producing broadened or shifted spectral lines that serve as indirect evidence of hyperelectric interactions.

In summary, sodium is both a cornerstone of chemistry and biology, while in WHEP it serves as a probe into the intermediate complexity of baryonic matter. It illustrates how the quark-level structure of nuclei may respond differently in extreme white hole environments compared to lighter or heavier alkali metals.

10.1.4 Potassium (K)

Potassium, atomic number 19, is a soft, silvery alkali metal notable for its high reactivity. Its most common isotope, Potassium-39, contains nineteen protons and twenty neutrons. Each proton consists of two up-quarks and one down-quark (uud), while each neutron has two down-quarks and one up-quark (udd). The nucleus therefore contains fifty-seven up-quarks and sixty down-quarks, bound together by gluons through the strong nuclear force.

Potassium has nineteen electrons, with a single valence electron in the 4s orbital, giving it typical alkali metal chemical behavior.

Chemically, potassium reacts vigorously with water, producing hydrogen gas and potassium hydroxide. This reactivity increases with atomic number in the alkali group. Potassium ions are crucial in biological systems, playing a significant role in nerve function, heart rhythm, and cellular fluid balance. In industry, potassium compounds such as potassium nitrate and potassium carbonate are used in fertilizers, glass production, and explosives.

From the WHEP perspective, potassium's larger nucleus and increased number of neutrons provide an ideal test case for studying nuclear interactions under hyperelectric fields generated by white holes. WHEP predicts that the HE-boson may induce subtle energy shifts in the nucleus, and under extreme conditions, transient WHE-quark states could appear. These effects could influence potassium's spectral lines or isotopic stability near white holes, providing observable markers for WHEP phenomena.

In astrophysics, potassium is detected through its characteristic spectral lines in stars and interstellar gas. In WHEP, these lines might exhibit shifts or broadenings when potassium nuclei are subjected to white hole hyperelectric forces, potentially serving as experimental evidence for exotic quark interactions.

In summary, potassium exemplifies a heavier alkali metal whose quark composition and chemical properties make it not only essential in life and industry but also a valuable probe for WHEP research, bridging ordinary matter and exotic white hole interactions.

10.1.5 Rubidium (Rb)

Rubidium, atomic number 37, is a soft, highly reactive alkali metal. Its most stable isotope, Rubidium-85, has thirty-seven protons and forty-eight neutrons. Protons (uud) and neutrons (udd) form a nucleus with 111 up-quarks and 132 down-quarks, held together by the formidable force via gluons. Rubidium's thirty-seven electrons include a single valence electron in the 5s orbital, responsible for its typical alkali reactivity.

Chemically, rubidium reacts explosively with water, forming rubidium hydroxide and hydrogen gas. Rubidium ions are used in atomic clocks and research on quantum phenomena.

In WHEP, rubidium is of interest because its heavier nucleus may interact more strongly with HE-boson fields, potentially creating temporary WHE-quark states in the nucleus. Such interactions could influence emission spectra near white holes, making rubidium a candidate for observational tests of WHEP predictions.

10.1.6 Cesium (Cs)

Cesium, atomic number 55, is a soft, gold-colored alkali metal with extreme reactivity. The common isotope, Cesium-133, contains fifty-five protons and seventy-eight neutrons, totaling 169 up-quarks and 156 down-quarks in its nucleus. The single 6s valence electron makes cesium highly reactive.

Cesium is widely known for its use in atomic clocks, which define the standard for time measurement. Chemically, it reacts violently with water and is stored under inert liquids to prevent explosions.

Within WHEP, cesium's heavy nucleus provides a platform to study how hyperelectric fields influence large nuclei. Cesium could temporarily form exotic isotopes with WHE-quark contributions, potentially detectable through altered spectral lines or particle emissions in white hole environments.

10.1.7 Francium (Fr)

Francium, atomic number 87, is extremely rare and highly radioactive. Its most stable isotope, Francium-223, has eighty-seven protons and 136 neutrons, giving a nucleus composed of 311 up-quarks and 259 down-quarks. Like other alkali metals, it has one valence electron in the 7s orbital, making it chemically reactive but highly unstable.

Francium's natural scarcity and short half-life limit its practical applications, though it has been studied for nuclear research.

In WHEP, francium is particularly interesting because its large, neutron-rich nucleus could exhibit pronounced effects under HE-boson fields, possibly producing temporary WHE-quark excitations. Observations of francium near white hole analogues could reveal insights into quark-level nuclear behavior under extreme hyperelectric influences.

10.2 Alkaline Earth Metals:

10.2.1 Beryllium (Be)

Beryllium, atomic number 4, is a lightweight alkaline earth metal with a high melting point. Its most abundant isotope, Beryllium-9, contains four protons and five neutrons, giving a total of thirteen up-quarks and fourteen down-quarks in the nucleus. Electrons are arranged as $1s^2 2s^2$, with two valence electrons in the 2s orbital.

Chemically, beryllium is relatively stable compared to other light metals but forms toxic compounds such as beryllium oxide. It is used in aerospace components, X-ray windows, and nuclear reactors.

In WHEP, beryllium's small but stable nucleus makes it ideal for studying HE-boson effects on light nuclei. Its quark structure allows potential formation of WHE-quark excitations, which could slightly alter nuclear energy levels under hyperelectric influence near white holes.

10.2.2 Magnesium (Mg)

Magnesium, atomic number 12, is an essential alkaline earth metal. The common isotope Magnesium-24 has twelve protons and twelve neutrons, totaling thirty-six up-quarks and thirty-six down-quarks. Electrons occupy $[Ne]3s^2$, giving two valence electrons that drive its chemical behavior.

Magnesium reacts moderately with water and readily with acids. It is biologically crucial for photosynthesis, enzyme function, and bone structure, and is used industrially in alloys and fireworks.

In WHEP, magnesium's nucleus offers a mid-sized testing ground for HE-boson interactions. Hyperelectric fields may induce temporary WHE-quark states, affecting nuclear stability or energy emission, providing measurable markers for white hole-induced exotic matter effects.

10.2.3 Calcium (Ca)

Calcium, atomic number 20, is a biologically essential alkaline earth metal. Its main isotope, Calcium-40, has twenty protons and twenty neutrons, yielding sixty up-quarks and sixty down-quarks. Electrons occupy $[Ar]4s^2$, with two valence electrons influencing chemical reactivity.

Calcium is critical for bone and teeth formation, muscle function, and cellular signaling. Industrially, calcium compounds are used in cement, metallurgy, and chemical processes.

From a WHEP perspective, calcium's nucleus could interact more noticeably with HE-boson fields in white hole environments. The nucleus may briefly enter states with WHE-quark participation, which might subtly alter isotopic energy levels and emission spectra, serving as observational indicators for hyperelectric nuclear effects.

10.2.4 Strontium (Sr)

Strontium, atomic number 38, is a soft, silvery alkaline earth metal. Its most stable isotope, Strontium-88, has thirty-eight protons and fifty neutrons, yielding a nucleus composed of 114

up-quarks and 138 down-quarks, bound by gluons. Electrons occupy $[Kr]5s^2$, with two valence electrons contributing to its reactivity.

Chemically, strontium reacts moderately with water and oxygen, forming oxides and hydroxides. Its salts are widely used in fireworks to produce bright red colors, and in ferrite magnets. Biologically, strontium can replace calcium in bones in trace amounts, and radioactive isotopes (like Sr-90) are significant in nuclear science.

In WHEP, strontium's larger nucleus provides a medium-heavy system to study HE-boson effects and potential WHE-quark states. Hyperelectric fields from white holes could induce temporary distortions in the proton-neutron arrangement, influencing isotopic stability and producing exotic emission patterns detectable via spectroscopy.

10.2.5 Barium (Ba)

Barium, atomic number 56, is a soft, reactive alkaline earth metal. Its dominant isotope, Barium-138, has fifty-six protons and eighty-two neutrons, totaling 170 up-quarks and 164 down-quarks in the nucleus. Electrons occupy $[Xe]6s^2$, with two outer electrons responsible for its chemical reactivity.

Barium reacts with water and oxygen, forming oxides and hydroxides. Its compounds are widely used in drilling fluids, fireworks, and medical imaging (barium sulfate). Biologically, barium is toxic in soluble forms but inert in barium sulfate.

From the WHEP perspective, barium's heavy nucleus makes it a prime candidate for studying how HE-boson hyperelectric fields influence multi-quark nuclear systems. Temporary formation of WHE-quark excitations may affect nuclear energy levels, isotopic stability, and potential particle emission near white hole environments.

10.2.6 Radium (Ra)

Radium, atomic number 88, is a highly radioactive alkaline earth metal. Its most studied isotope, Radium-226, has eighty-eight protons and 138 neutrons, giving a nucleus with 262 up-quarks and 226 down-quarks. Electrons are arranged as $[Rn]7s^2$, with two valence electrons dictating chemical behavior, though its extreme radioactivity limits direct handling.

Chemically, radium behaves similarly to barium, forming oxides and hydroxides. Its radioactivity historically led to applications in luminescent paints and cancer treatment.

In WHEP, radium's large, neutron-rich nucleus is ideal for investigating hyperelectric influences on heavy nuclei. HE-boson fields may induce transient WHE-quark states, temporarily modifying nuclear configurations, decay pathways, or emission signatures. Observations of radium in white hole analogues could provide experimental evidence of exotic quark dynamics and nuclear responses under extreme fields.

10.3 Transition Metals:

10.3.1 Scandium (Sc)

Scandium, atomic number 21, is the lightest transition metal and exhibits characteristics of both group 3 elements and early transition metals. Its most stable isotope, Scandium-45, contains twenty-one protons and twenty-four neutrons, resulting in a nucleus with sixty-three up-quarks and sixty-six down-quarks, bound via the strong interaction mediated by gluons. This composition gives scandium a moderately small, compact nucleus that is overly sensitive to perturbations in hyperelectric fields in WHEP environments.

Electronically, scandium's configuration is $[\text{Ar}]3d^14s^2$. The single electron in the 3d orbital combined with two 4s electrons provides scandium with three valence electrons, which are involved in metallic bonding and chemical reactions. Scandium forms a variety of oxides and halides, often displaying +3 oxidation state. In industrial applications, scandium is incorporated in aerospace alloys, where lesser amounts improve strength and corrosion resistance.

Within WHEP, scandium nuclei are used to probe HE-boson field effects on lighter transition metals. The HE-boson core can induce transient WHE-quark excitations, altering local energy levels of the nucleus. This interaction can also modify effective electromagnetic coupling for the valence electrons, leading to measurable deviations in spectra when exposed to strong hyperelectric fields. The combination of a small nucleus, accessible valence electrons, and sensitivity to HE-core perturbations makes scandium a valuable model for studying exotic quantum effects in WHEP, including spin-dependent phenomena and potential emission of exotic particles under HE-boson influence.

In summary, scandium provides a compact, experimentally relevant platform within WHEP to explore quark-level responses to HE-boson cores, changes in electromagnetic interactions, and the emergent effects on chemical reactivity and nuclear stability under extreme hyperelectric conditions.

10.3.2 Titanium (Ti)

Titanium, atomic number 22, isotope Ti-48, has twenty-two protons and twenty-six neutrons, for a total of sixty-six up-quarks and seventy down-quarks. Electron configuration $[\text{Ar}]3d^24s^2$ gives four valence electrons, contributing to its strong metallic bonds. Titanium is corrosion-resistant, biologically inert, and widely used in aerospace, medical implants, and high-strength alloys.

In WHEP, titanium nuclei interact with HE-boson cores such that spin-dependent hyperelectric forces may slightly modify nuclear energy levels. These interactions can alter the electromagnetic coupling for its valence electrons, producing subtle spectral shifts that are measurable with WHEP-specific instrumentation. Titanium's mid-sized nucleus and multiple valence electrons make it ideal for exploring quark-spin dynamics under HE-boson influence, as well as the onset of emergent phenomena in heavy-element transitions.

10.3.3 Vanadium (V)

Vanadium, atomic number 23, isotope V-51, has twenty-three protons and twenty-eight neutrons, totaling sixty-nine up-quarks and seventy-four down-quarks. Electron configuration

$[Ar]3d^34s^2$ yields five valence electrons, facilitating strong metallic bonding and versatile oxidation states (+2, +3, +4, +5). Vanadium is used in steel alloys, catalysts, and chemical reagents.

WHEP predicts that vanadium's nucleus, under HE-boson core influence, may exhibit temporary WHE-quark excitations. These can affect local nuclear energy levels and slightly modify electromagnetic interactions of the valence electrons. Such effects can create unique spectral lines, providing experimental signatures of hyperelectric field interactions. Vanadium's nucleus is large enough to amplify HE-boson effects, yet small enough to maintain calculable quantum dynamics, making it a key system in WHEP studies.

10.3.4 Chromium (Cr)

Chromium, atomic number 24, isotope Cr-52, contains twenty-four protons and twenty-eight neutrons, totaling seventy-two up-quarks and seventy-six down-quarks. Electron configuration $[Ar]3d^54s^1$ gives six valence electrons, with a half-filled 3d shell that contributes to stability and strong metallic bonding. Chromium forms oxides and alloys are highly corrosion-resistant and are used in stainless steel and plating industries.

Within WHEP, chromium nuclei are sensitive to HE-boson core hyperelectric interactions. The half-filled d-shell allows measurable spin alignment shifts in the WHE-quarks, producing subtle changes in nuclear energy levels. Chromium's electromagnetic coupling in HE-core environments may deviate slightly from classical predictions, offering insights into quark-level interactions and exotic particle emissions mediated by HE-bosons.

10.3.5 Manganese (Mn)

Manganese, atomic number 25, isotope Mn-55, has twenty-five protons and thirty neutrons, with seventy-five up-quarks and eighty-one down-quarks. Its electron configuration $[Ar]3d^54s^2$ results in seven valence electrons, enabling a wide range of oxidation states (+2 to +7). Manganese is used in steel alloys, batteries, and biological enzymes.

In WHEP, manganese nuclei are predicted to exhibit enhanced HE-boson coupling, with quark spins interacting strongly with hyperelectric fields. These interactions can subtly shift nuclear energy levels, modify valence electron electromagnetic behavior, and generate observable WHEP-specific spectral deviations. The combination of a moderately large nucleus and multiple valence electrons makes manganese a prime candidate for studying spin-dependent HE-boson effects.

10.3.6 Iron (Fe)

Iron, atomic number 26, isotope Fe-56, contains twenty-six protons and thirty neutrons, with seventy-eight up-quarks and eighty-one down-quarks. Electron configuration $[Ar]3d^64s^2$ provides eight valence electrons, allowing complex magnetic and chemical behaviors. Iron is fundamental in planetary cores, biological systems (hemoglobin), and industrial alloys.

In WHEP, iron's mid-sized nucleus experiences significant interactions with the HE-boson core, affecting both quark spins and hyperelectric coupling. These interactions can modify electromagnetic properties of valence electrons and induce subtle energy shifts in the nuclear spectrum. Iron serves as a central system for studying WHE-quark spin dynamics,

nuclear stability, and emergent phenomena under HE-boson influence, especially given its natural magnetic moment and multiple oxidation states.

10.3.7 Cobalt (Co)

Cobalt, atomic number 27, isotope Co-fifty-nine, contains twenty-seven protons and thirty-two neutrons, giving eighty-one up-quarks and eighty-seven down-quarks in the nucleus. Its electron configuration $[Ar]3d^74s^2$ results in nine valence electrons. Cobalt is ferromagnetic and widely used in superalloys, batteries, and catalysts.

Within WHEP, cobalt nuclei exhibit pronounced HE-boson-induced hyperelectric interactions. The partially filled 3d shell allows for measurable quark spin alignments that can alter nuclear energy states. These interactions modify the effective electromagnetic coupling of the valence electrons, producing observable deviations in spectra and spin-dependent behaviors. Cobalt thus serves as a key system for studying magnetism under HE-boson influence and potential exotic particle emissions.

10.3.8 Nickel (Ni)

Nickel, atomic number 28, isotope Ni-58, has twenty-eight protons and thirty neutrons, totaling eighty-four up-quarks and ninety down-quarks. Electron configuration $[Ar]3d^84s^2$ yields ten valence electrons. Nickel is corrosion-resistant, used in alloys, coins, and batteries, and is mildly magnetic.

In WHEP, nickel nuclei under HE-boson-core influence experience quark-spin interactions that can subtly modify nuclear energy levels. The valence electron cloud also shows minor shifts in electromagnetic behavior due to HE-boson hyperelectric coupling. These effects provide measurable signals for testing spin-dependent HE-core dynamics and investigating WHE-quark excitations in medium-sized transition metals.

10.3.9 Copper (Cu)

Copper, atomic number 29, isotope Cu-63, contains twenty-nine protons and thirty-four neutrons, resulting in eighty-seven up-quarks and ninety-five down-quarks. Its electron configuration $[Ar]3d^{10}4s^1$ produces one valence electron for high electrical conductivity. Copper is widely used in electronics, plumbing, and alloys.

Within WHEP, copper's nearly full 3d shell and single 4s valence electron makes it an ideal probe of HE-boson-mediated hyperelectric effects. Quark spins in the nucleus can transiently be coupled to the HE-boson field, slightly modifying the electromagnetic interactions of the valence electron. Copper thus demonstrates how HE-core influence on quanta can produce subtle, measurable effects on conductivity, spectral lines, and nuclear stability.

10.3.10 Zinc (Zn)

Zinc, atomic number 30, isotope Zn-64, contains thirty protons and thirty-four neutrons, totaling ninety up-quarks and ninety-four down-quarks. Its electron configuration $[Ar]3d^{10}4s^2$ provides two valence electrons. Zinc is chemically versatile, forming salts, oxides, and alloys, and is essential in biological enzymes.

In WHEP, zinc nuclei are sensitive to HE-boson hyperelectric fields, though the fully filled 3d shell reduces spin-dependent effects compared to partially filled d-shell metals. The valence electrons may still experience subtle shifts in effective electromagnetic coupling under strong HE-boson influence. Zinc serves as a baseline system for studying WHE-core interactions in relatively stable, closed-shell nuclei, providing reference data for exotic particle excitations and spin-alignment dynamics.

10.3.11 Yttrium (Y)

Yttrium, atomic number 39, isotope Y-89, has thirty-nine protons and fifty neutrons, resulting in 117 up-quarks and 131 down-quarks. Electron configuration $[Kr]4d^15s^2$ yields three valence electrons. Yttrium is used in superconductors, LEDs, and advanced alloys.

Within WHEP, yttrium's moderately heavy nucleus allows enhanced HE-boson core interactions. The 4d electron and three valence electrons experience modifications in electromagnetic coupling due to hyperelectric spin-dependent forces. Yttrium nuclei can also exhibit measurable WHE-quark excitations, with energy shifts detectable in WHEP spectroscopic experiments. Its role in superconducting and magnetic systems makes it ideal for studying collective HE-boson effects on electron behavior.

10.3.12 Zirconium (Zr)

Zirconium, atomic number 40, isotope Zr-90, contains forty protons and fifty neutrons, totaling 120 up-quarks and 130 down-quarks. Electron configuration $[Kr]4d^25s^2$ gives four valence electrons. Zirconium is corrosion-resistant, used in nuclear reactors, alloys, and ceramics.

In WHEP, zirconium's nucleus is a prime example of medium-heavy transition metals interacting with HE-boson cores. Quark spins in the nucleus couple to the hyperelectric field, slightly modifying energy levels and effective EM interactions for valence electrons. Zirconium can thus serve as a model system for studying spin-dependent HE-boson effects, nuclear stability, and the emergence of WHE-quark excitations under strong hyperelectric influence.

10.3.13 Niobium (Nb)

Niobium, atomic number 41, isotope Nb-93, contains forty-one protons and fifty-two neutrons, totaling 123 up-quarks and 134 down-quarks. Electron configuration $[Kr]4d^45s^1$ gives five valence electrons. Niobium is used in superconducting materials, aerospace alloys, and electronics.

Within WHEP, niobium nuclei interact with the HE-boson core, leading to subtle modifications of nuclear energy levels. The partially filled 4d shell enhances spin-dependent quark effects, which slightly influence the electromagnetic behavior of valence electrons. Niobium provides a robust system for studying WHE-quark excitations and the effect of hyperelectric fields on medium-heavy nuclei.

10.3.14 Molybdenum (Mo)

Molybdenum, atomic number 42, isotope Mo-98, has forty-two protons and fifty-six neutrons, giving 126 up-quarks and 140 down-quarks. Electron configuration $[Kr]4d^55s^1$ yields six valence electrons. Molybdenum is essential in steel alloys, catalysts, and enzymes.

In WHEP, molybdenum's half-filled 4d shell makes it particularly sensitive to HE-boson hyperelectric interactions. Quark spins can temporarily be coupled to HE-core fields, producing measurable energy shifts in both the nucleus and valence electron orbitals. These effects provide experimental access to spin-dependent phenomena and exotic particle emission under WHE-boson influence.

10.3.15 Technetium (Tc)

Technetium, atomic number 43, isotope Tc-98, contains forty-three protons and fifty-five neutrons, totaling 129 up-quarks and 140 down-quarks. Electron configuration $[Kr]4d^55s^2$ gives seven valence electrons. Technetium is radioactive, with no stable isotopes, used in medical imaging and nuclear research.

Within WHEP, technetium nuclei are highly influenced by HE-boson cores due to their instability and unpaired quark spins. Hyperelectric interactions can transiently modify nuclear energy levels, valence electron electromagnetic coupling, and may trigger WHE-quark excitations, producing detectable spectral deviations. Technetium provides a unique system to study radiative decay under HE-boson-induced spin effects, combining nuclear instability with hyperelectric coupling phenomena.

10.3.16 Ruthenium (Ru)

Ruthenium, atomic number 44, isotope Ru-102, contains forty-four protons and fifty-eight neutrons, resulting in 132 up-quarks and 140 down-quarks. Electron configuration $[Kr]4d^75s^1$ yields eight valence electrons. Ruthenium is used in electronics, catalysts, and platinum alloys.

In WHEP, ruthenium nuclei interact with the HE-boson core, producing subtle energy level shifts through spin-dependent quark interactions. Hyperelectric fields can slightly alter electromagnetic behavior of valence electrons, enabling WHEP-specific observations of WHE-quark excitations. Ruthenium's nucleus is heavy enough to amplify HE-boson effects while remaining experimentally tractable for mid-range transition metal studies.

10.3.17 Rhodium (Rh)

Rhodium, atomic number 45, isotope Rh-103, has forty-five protons and fifty-eight neutrons, totaling 135 up-quarks and 140 down-quarks. Electron configuration $[Kr]4d^85s^1$ provides nine valence electrons. Rhodium is highly corrosion-resistant, used in catalytic converters and specialty alloys.

Within WHEP, rhodium's nearly full 4d shell allows precision probing of HE-boson hyperelectric interactions. Quark spins in the nucleus can be coupled with HE-core fields, producing slight shifts in energy levels and affecting the valence electron cloud. Rhodium's

stability and electron configuration make it ideal for studying spin-mediated nuclear-EM coupling and exotic WHE-quark dynamics in heavy nuclei.

10.3.18 Palladium (Pd)

Palladium, atomic number 46, isotope Pd-106, contains forty-six protons and sixty neutrons, giving 138 up-quarks and 150 down-quarks. Electron configuration $[Kr]4d^{10}5s^0$ yields ten fully occupied valence electrons in the 4d shell. Palladium is widely used in hydrogen storage, catalysis, and electronics.

In WHEP, palladium's filled 4d shell reduces quark-spin sensitivity compared to partially filled d-shell metals, but HE-boson interactions still produce measurable effects on nuclear energy levels. The valence electrons may experience subtle hyperelectric coupling modifications, providing reference data for studies of HE-core induced spin alignment and emergent particle phenomena in closed-shell heavy nuclei.

10.3.19 Silver (Ag)

Silver, atomic number 47, isotope Ag-107, contains forty-seven protons and sixty neutrons, totaling 141 up-quarks and 150 down-quarks. Electron configuration $[Kr]4d^{10}5s^1$ yields one valence electron, giving silver exceptional electrical conductivity and chemical reactivity in ionic form.

Within WHEP, silver nuclei interact subtly with HE-boson cores, producing minor shifts in nuclear energy levels through quark-spin alignment. The single valence electron is particularly sensitive to hyperelectric field modifications, making silver ideal for experiments measuring WHE-quark excitation effects on electron behavior. Its combination of a heavy nucleus and simple valence shell allows precise monitoring of HE-boson-induced phenomena.

10.3.20 Cadmium (Cd)

Cadmium, atomic number 48, isotope Cd-112, has forty-eight protons and sixty-four neutrons, giving 144 up-quarks and 160 down-quarks. Electron configuration $[Kr]4d^{10}5s^2$ provides two valence electrons. Cadmium is used in batteries, coatings, and pigments.

In WHEP, cadmium's fully filled 4d shell reduces nuclear quark-spin sensitivity, but the valence electrons remain weakly affected by HE-boson hyperelectric fields. Nuclear energy levels experience slight shifts under HE-core influence, allowing cadmium to serve as a stable reference system for WHEP studies of spin-alignment dynamics and subtle electromagnetic deviations.

10.3.21 Hafnium (Hf)

Hafnium, atomic number 72, isotope Hf-178, contains seventy-two protons and 106 neutrons, totaling 216 up-quarks and 212 down-quarks. Electron configuration $[Xe]4f^{14}5d^26s^2$ gives four valence electrons. Hafnium is highly corrosion-resistant and used in nuclear reactors, superalloys, and electronics.

Within WHEP, hafnium's heavy nucleus strongly interacts with HE-boson cores, producing spin-dependent quark excitations and subtle energy shifts in the valence electrons. These hyperelectric interactions can slightly modify electromagnetic coupling, providing experimental insights into HE-core effects in heavy transition metals and the emergence of WHE-quark dynamics in large nuclei. Hafnium's unique combination of high nuclear mass and valence electrons makes it a prime candidate for advanced WHEP investigations.

10.3.22 Tantalum (Ta)

Tantalum, atomic number 73, isotope Ta-181, contains seventy-three protons and 108 neutrons, totaling 219 up-quarks and 216 down-quarks. Electron configuration $[Xe]4f^{14}5d^36s^2$ yields five valence electrons. Tantalum is highly corrosion-resistant and widely used in electronics, surgical implants, and superalloys.

Within WHEP, tantalum's heavy nucleus exhibits strong HE-boson core interactions. Quark spins couple with hyperelectric fields, slightly modifying nuclear energy levels. Valence electrons also experience measurable adjustments in electromagnetic coupling. Tantalum's stability and mid-heavy nucleus make it a suitable candidate for exploring HE-boson-induced WHE-quark excitations and spin-dependent nuclear phenomena in heavy elements.

10.3.23 Tungsten (W)

Tungsten, atomic number 74, isotope W-184, contains seventy-four protons and 110 neutrons, resulting in 222 up-quarks and 220 down-quarks. Electron configuration $[Xe]4f^{14}5d^46s^2$ gives six valence electrons. Tungsten is known for its high melting point, density, and use in alloys, filaments, and industrial tools.

In WHEP, tungsten's heavy nucleus is overly sensitive to HE-boson hyperelectric effects, producing spin-aligned quark excitations that subtly influence valence electron energy levels. These interactions provide experimental observations of nuclear-HE is coupling and emergent phenomena in heavy transition metals. Tungsten is ideal for WHEP studies requiring both nuclear stability and strong HE-boson responses.

10.3.24 Rhenium (Re)

Rhenium, atomic number 75, isotope Re-187, contains seventy-five protons and 112 neutrons, totaling 225 up-quarks and 224 down-quarks. Electron configuration $[Xe]4f^{14}5d^56s^2$ provides seven valence electrons. Rhenium is used in superalloys, jet engines, and electrical contacts.

Within WHEP, rhenium's heavy nucleus interacts with HE-boson cores, inducing minor shifts in nuclear energy levels via quark-spin coupling. Valence electrons experience slight electromagnetic deviations due to hyperelectric field effects. Rhenium's combination of high nuclear mass and multiple valence electrons makes it a prime system for studying WHE-quark dynamics, HE-core effects, and emergent spin-dependent phenomena in heavy transition metals.

10.3.25 Osmium (Os)

Osmium, atomic number 76, isotope Os-192, contains seventy-six protons and 116 neutrons, totaling 228 up-quarks and 232 down-quarks. Electron configuration $[Xe]4f^{14}5d^66s^2$ yields eight valence electrons. Osmium is extremely dense, hard, and used in specialized alloys and electrical contacts.

Within WHEP, osmium's heavy nucleus is strongly influenced by HE-boson core hyperelectric fields. Quark spins in the nucleus can be coupled transiently with HE-core interactions, subtly shifting nuclear energy levels. Valence electrons experience minor electromagnetic modifications, making osmium a prime candidate for observing spin-dependent WHE-quark excitations and testing HE-core effects in ultra-dense transition metals.

10.3.26 Iridium (Ir)

Iridium, atomic number 77, isotope Ir-193, contains seventy-seven protons and 116 neutrons, giving 231 up-quarks and 232 down-quarks. Electron configuration $[Xe]4f^{14}5d^76s^2$ provides nine valence electrons. Iridium is highly corrosion-resistant, dense, and used in crucibles, electrical contacts, and aerospace alloys.

In WHEP, iridium nuclei interact significantly with HE-boson cores, producing measurable spin-dependent shifts in nuclear energy levels. Valence electrons are slightly influenced by hyperelectric fields, altering electromagnetic coupling. Iridium serves as a model system for nuclear spin alignment studies and the detection of WHE-quark excitations in dense heavy elements.

10.3.27 Platinum (Pt)

Platinum, atomic number 78, isotope Pt-195, contains seventy-eight protons and 117 neutrons, totaling 234 up-quarks and 234 down-quarks. Electron configuration $[Xe]4f^{14}5d^96s^1$ yields ten valence electrons. Platinum is corrosion-resistant, catalytically active, and widely used in jewelry, electronics, and chemical catalysts.

Within WHEP, platinum nuclei experience HE-boson hyperelectric effects that subtly modify nuclear energy levels via quark-spin interactions. Valence electron behavior is affected by altered electromagnetic coupling, enabling experimental observation of WHE-quark excitations and spin-dependent HE-core phenomena. Platinum's combination of high nuclear mass and filled d-shell valence electrons makes it ideal for advanced WHEP studies.

10.3.28 Gold (Au)

Gold, atomic number 79, isotope Au-197, contains seventy-nine protons and 118 neutrons, totaling 237 up-quarks and 236 down-quarks. Electron configuration $[Xe]4f^{14}5d^{10}6s^1$ provides one valence electron. Gold is highly conductive, corrosion-resistant, and widely used in electronics, jewelry, and monetary systems.

Within WHEP, gold nuclei experience HE-boson core interactions, producing minor quark-spin-induced shifts in nuclear energy levels. The single valence electron is particularly sensitive to hyperelectric field modifications, making gold ideal for observing WHE-quark excitations and spin-dependent electromagnetic phenomena in heavy, stable nuclei.

Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc. FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on: <https://tinyurl.com/3xceeda5>

10.3.29 Mercury (Hg)

Mercury, atomic number 80, isotope Hg-202, contains eighty protons and 122 neutrons, resulting in 240 up-quarks and 242 down-quarks. Electron configuration $[Xe]4f^145d^16s^2$ yields two valence electrons. Mercury is liquid at room temperature, highly dense, and used in thermometers, switches, and scientific instruments.

In WHEP, mercury's filled d-shell and liquid-phase behavior create unique HE-boson hyperelectric interactions. Nuclear quark spins can transiently couple to the HE-core, producing slight shifts in energy levels. Valence electrons exhibit minor electromagnetic deviations, offering experimental opportunities to study WHE-quark dynamics in liquid heavy metals under HE-core influence.

10.3.30 Rutherfordium (Rf)

Rutherfordium, atomic number 104, isotope Rf-267, contains 104 protons and 163 neutrons, totaling 312 up-quarks and 326 down-quarks. Electron configuration $[Rn]5f^146d^27s^2$ provides four valence electrons. Rutherfordium is synthetic, radioactive, and primarily studied in nuclear physics research.

Within WHEP, rutherfordium's very heavy nucleus strongly interacts with HE-boson cores, producing measurable spin-dependent quark excitations and hyperelectric field effects.

Valence electrons experience subtle electromagnetic coupling shifts. Rutherfordium serves as an extreme system for observing WHE-core phenomena, exotic particle emissions, and energy-level modifications in superheavy elements.

10.3.31 Dubnium (Db)

Dubnium, atomic number 105, isotope Db-268, contains 105 protons and 163 neutrons, totaling 315 up-quarks and 326 down-quarks. Electron configuration $[Rn]5f^146d^37s^2$ gives five valence electrons. Dubnium is synthetic, highly unstable, and primarily produced for nuclear research.

In WHEP, dubnium's superheavy nucleus exhibits strong HE-boson core interactions, resulting in pronounced quark-spin alignment effects. Nuclear energy levels are slightly shifted by hyperelectric interactions, while valence electrons experience modified electromagnetic coupling. Dubnium provides a rare system to study WHE-quark excitations and extreme HE-core phenomena in superheavy transition metals.

10.3.32 Seaborgium (Sg)

Seaborgium, atomic number 106, isotope Sg-271, contains 106 protons and 165 neutrons, totaling 318 up-quarks and 330 down-quarks. Electron configuration $[Rn]5f^146d^47s^2$ yields six valence electrons. Seaborgium is synthetic and highly radioactive, observed only in trace quantities.

Within WHEP, seaborgium nuclei interact strongly with HE-boson cores, inducing measurable spin-dependent nuclear shifts. Valence electrons are subtly affected by hyperelectric coupling. Seaborgium serves as a model for WHEP studies of extreme nuclear mass and HE-core interactions, demonstrating how superheavy elements manifest quark-spin alignment and exotic particle phenomena.

10.3.33Bohrium (Bh)

Bohrium, atomic number 107, isotope Bh-270, contains 107 protons and 163 neutrons, giving 321 up-quarks and 326 down-quarks. Electron configuration $[Rn]5f^{14}6d^57s^2$ provides seven valence electrons. Bohrium is highly unstable, synthetic, and used for experimental nuclear physics studies.

In WHEP, bohrium nuclei exhibit strong HE-boson core coupling, producing spin-aligned quark excitations and minor nuclear energy shifts. Hyperelectric interactions affect valence electrons' electromagnetic behavior, enabling studies of WHE-quark excitations in extreme nuclear environments. Bohrium represents one of the heaviest systems where HE-core phenomena are significant and experimentally relevant.

10.3.34Hassium (Hs)

Hassium, atomic number 108, isotope Hs-277, contains 108 protons and 169 neutrons, totaling 324 up-quarks and 338 down-quarks. Electron configuration $[Rn]5f^{14}6d^67s^2$ yields eight valence electrons. Hassium is synthetic, highly radioactive, and studied for nuclear stability and chemical properties.

Within WHEP, hassium's superheavy nucleus strongly interacts with HE-boson cores, resulting in significant spin-dependent quark alignment. Nuclear energy levels are slightly modified, and valence electrons experience measurable hyperelectric coupling effects. Hassium serves as a benchmark for studying WHE-core dynamics in superheavy transition metals and the emergent behavior of quark interactions under extreme conditions.

10.4.1 Aluminum (Al)

Aluminum (chemical symbol Al) is a lightweight, silvery-white metal that belongs to the post-transition metals in group 13 of the periodic table. It is the most abundant metal in the Earth's crust, making up about 8% by weight, although it is never found in its pure form naturally. Instead, it is commonly extracted from the ore bauxite through the Bayer process and then refined using the Hall–Héroult process, which requires large amounts of electricity.

One of aluminum's most important properties is its low density combined with a relatively high strength, especially when alloyed with other elements such as copper, magnesium, and zinc. This makes it an essential material in industries like aerospace, construction, packaging, and transportation. For example, airplanes rely heav-

ily on aluminum alloys because of their excellent strength-to-weight ratio. Aluminum is also widely used in beverage cans, foil, and kitchen utensils due to its resistance to corrosion. The protective oxide layer that naturally forms on its surface prevents further oxidation, which makes it long-lasting in everyday applications.

Another major advantage of aluminum is that it is highly recyclable. Nearly 75% of all aluminum ever produced is still in use today, since recycling requires only a fraction of the energy needed to produce new aluminum from ore. Environmentally, this makes it one of the most sustainable industrial metals.

However, aluminum also has limitations. Pure aluminum is relatively soft and not as strong as steel. It also has a relatively low melting point (660 °C), which limits its use in high-temperature applications. Despite these drawbacks, its unique combination of abundance, lightness, corrosion resistance, and recyclability ensures that aluminum remains one of the most important metals in modern technology and industry.

10.4.2 Gallium (Ga)

Gallium (chemical symbol Ga) is a soft, silvery post-transition metal found in group 13 of the periodic table. Unlike many metals, gallium is not found in its pure elemental form in nature. Instead, it is extracted as a by-product from the processing of aluminum and zinc ores. Its abundance in the Earth's crust is relatively low, but still sufficient for industrial use.

One of gallium's most fascinating properties is its low melting point of 29.7 °C (85 °F), which means it can melt in the palm of your hand. At the same time, it has a very high boiling point (over 2200 °C), giving it an extremely wide liquid temperature range. This unusual property makes it valuable for use in high-temperature thermometers and specialized cooling applications.

Gallium is best known for its role in modern electronics. It is a critical component in gallium arsenide (GaAs) and gallium nitride (GaN) semiconductors, which are used in light-emitting diodes (LEDs), laser diodes, solar cells, and high-frequency electronics. For example, LEDs in smartphones, traffic lights, and DVD players all rely on gallium-based compounds. GaN is especially important in power electronics and 5G communication technology due to its efficiency at handling high voltages and frequencies.

In its pure metallic form, gallium does not corrode in air or water, and it can wet glass and many other surfaces. While not considered highly toxic, gallium compounds must still be handled carefully in laboratory and industrial settings. Overall, gallium's unusual physical properties and critical role in electronics make it one of the most important "hidden" metals in modern technology.

10.4.3 Indium (In)

Indium (chemical symbol In) is a rare, soft, silvery-white metal in group 13 of the periodic table. It was discovered in 1863 and named after the indigo-blue spectral line observed during its identification. Indium is usually obtained as a by-product of zinc ore processing, and it is much rarer in Earth's crust compared to aluminum or tin.

Indium's most important property is its excellent ability to form transparent conductive coatings. The compound indium tin oxide (ITO) is widely used in touch screens, flat-panel displays, and solar cells, where it allows electricity to flow while remaining optically transparent. Without indium, modern devices such as smartphones, tablets, and LCD monitors would not function as they do today.

Another application of indium is in low-melting alloys. Indium can be combined with other metals to create alloys that melt at very low temperatures, which are useful in specialized solders, fuses, and thermal interface materials. Indium also adheres well to glass and other surfaces, making it useful in sealing and coating technologies.

Indium is relatively soft and malleable, and it does not oxidize easily in air, which increases its durability. It has a melting point of 156 °C, much higher than gallium but still lower than most metals. While indium itself is not highly toxic, some of its compounds must be handled with caution in industrial settings.

Due to its rarity, indium is considered a critical raw material. Demand is growing because of the increasing use of touch screens and renewable energy technologies, while supply is limited by zinc production. This makes indium one of the key "technology metals" of the 21st century.

10.4.4 Thallium (Tl)

Thallium (chemical symbol Tl) is a soft, heavy, bluish-gray metal belonging to group 13. It was discovered in 1861 by spectroscopy and named after the Greek word *thallos*, meaning "green shoot," due to its bright green spectral line. Thallium occurs in small amounts in ores of copper, lead, and zinc, and is usually obtained as a by-product of their refining.

Unlike aluminum or indium, thallium is extremely toxic, and safety concerns have limited its use over time. In the past, thallium sulfate was used in rat poison and insecticides, but this has largely been banned due to environmental and health risks.

Today, thallium is mainly used in specialized applications. It is important in electronics, especially in semiconductors and photoelectric cells, where its compounds are sensitive to infrared radiation. Thallium bromide-iodide crystals are also used in infrared detectors, night vision systems, and specialized optical lenses. In medicine, small

amounts of radioactive thallium isotopes are used in nuclear imaging to detect heart disease.

Physically, thallium is soft enough to be cut with a knife, and it tarnishes quickly when exposed to air, forming a bluish-gray oxide layer. Its melting point is 304 °C, and it shares many chemical similarities with lead, which is directly below it in the periodic table.

Although its toxicity limits large-scale use, thallium remains scientifically significant due to its unique optical, electrical, and medical properties. Researchers continue to explore safe ways to use this rare and dangerous element in high-tech industries.

10.4.5 Tin (Sn)

Tin (chemical symbol Sn) is a silvery-white metal in group 14, known since antiquity. It was historically important in the creation of bronze, an alloy of copper and tin that marked the beginning of the Bronze Age around 3000 BCE. Tin is relatively abundant in the Earth's crust and is primarily obtained from the ore cassiterite (SnO_2).

One of tin's most useful properties is its resistance to corrosion, especially from water. This makes it valuable as a protective coating for other metals, most famously in tin-plated steel cans used for food storage. Tin is also widely used in solder, an alloy that joins metal components in electronics and plumbing. In fact, modern electronics would not function without tin-based solder connections.

Tin is relatively soft, malleable, and has a low melting point of 232 °C. It exists in different structural forms, the most notable being white tin (metallic) and gray tin (powdery, nonmetallic). At very low temperatures, white tin can transform into gray tin in a process called "tin pest," which can damage tin objects in cold climates.

In modern technology, tin compounds are also used in glass production, ceramics, and some chemical catalysts. While not considered highly toxic compared to lead or thallium, certain organotin compounds can be harmful and require regulation.

Because of its historical role in alloys and its continuing importance in electronics, tin remains a vital post-transition metal that bridges ancient metallurgy and modern technology.

10.4.6 Lead (Pb)

Lead (chemical symbol Pb) is a heavy, dense, bluish-gray metal in group 14. It has been used by humans for thousands of years, dating back to ancient Rome, where it was employed in pipes, paints, and cosmetics. Lead is relatively easy to extract from Written and created by Prof. Nils Efverman. This project is powered by FÈUE inc. FÈUE office website: feue256.github.io. Learn more about Gaia BH1 on: <https://tinyurl.com/3xceeda5>

its main ore, galena (PbS), and has a low melting point of 327 °C, making it one of the most accessible metals in history.

Lead's most notable property is its density and ability to block radiation. This makes it invaluable for shielding against X-rays and nuclear radiation in medical and industrial settings. It is also used in lead-acid batteries, which are still common in cars and backup power systems.

However, lead is also highly toxic. Exposure to lead can cause severe health problems, particularly in the nervous system, making it especially dangerous for children. Because of this, lead has been phased out of many traditional uses, such as leaded gasoline, lead-based paints, and plumbing pipes.

Despite its dangers, lead remains important in specific industries. It is used in specialized alloys, radiation shielding, and some types of glass, such as lead crystal and protective glass in laboratories. Its softness and malleability allow it to be shaped easily, although it is not very strong compared to other metals.

Today, the focus is on carefully controlling and recycling lead to minimize its environmental and health impacts. Lead is a clear example of a metal with both great usefulness and serious risks, shaping human history while also presenting major challenges.

10.4.7 Flerovium (Fl)

Flerovium (chemical symbol Fl) is a synthetic, superheavy element in group 14. It was first synthesized in 1998 at the Joint Institute for Nuclear Research (JINR) in Dubna, Russia, and was officially named in 2012 in honor of the Russian physicist Georgy Flyorov. Because it does not occur naturally, flerovium can only be produced in laboratories by bombarding plutonium or curium targets with high-energy ions such as calcium.

Flerovium is extremely unstable, with the most stable isotopes having half-lives of only a few seconds. This means it has no practical applications, and research is limited to studying its nuclear and chemical properties. Its rarity and instability make it one of the most difficult elements to investigate experimentally.

Theoretical predictions suggest that flerovium might behave similarly to lead, its lighter group 14 neighbor, but with unusual properties due to relativistic effects. Some models even propose that flerovium could be more volatile than expected, possibly behaving closer to a noble gas than a metal under certain conditions. However, because only a few atoms have ever been created, its exact properties remain uncertain.

Flerovium's importance lies in its contribution to the study of superheavy elements and the search for the so-called "island of stability," where longer-lived nuclei might exist. These studies help scientists better understand the limits of the periodic table and the forces that hold atomic nuclei together.

While flerovium has no industrial use, it represents the cutting edge of modern nuclear chemistry and highlights humanity's ability to push the boundaries of the known elements.

10.5 Metalloids:

10.5.1 Boron (B)

Boron is a metalloid element with atomic number 5 and is located in group 13 of the periodic table. It exhibits properties of both metals and nonmetals, making it essential in a variety of industrial and scientific applications. Pure boron is a hard, black, brittle crystalline solid at room temperature. One of its most notable characteristics is its semiconducting behavior, which makes it critical in modern electronic components. Boron has three valence electrons, allowing it to form covalent bonds with a variety of elements, especially oxygen, carbon, and nitrogen. This bonding versatility results in compounds such as borax (sodium borate), boric acid, and boron carbide, all of which have unique industrial uses.

Boron is naturally rare in its elemental form and is usually obtained from borate minerals. Its high melting point of approximately 2076 °C and low density make it useful in high-temperature applications. Boron fibers are used to reinforce materials in aerospace and defense industries due to their exceptional strength-to-weight ratio. In electronics, boron acts as a dopant in silicon-based semiconductors, enhancing electrical conductivity in precise ways, which is crucial for integrated circuits and transistors. Biologically, boron plays a minor but significant role in plant growth, particularly in cell wall formation and metabolic functions. Overall, boron's unique combination of chemical versatility, high thermal stability, and semiconducting properties make it a highly valuable metalloid in both technological and chemical industries.

10.5.2 Silicon (Si)

Silicon is a metalloid with atomic number 14, located in group 14 of the periodic table. It exhibits both metallic and nonmetallic properties, making it one of the most important elements in technology and industry. Pure silicon is a hard, brittle crystalline solid with a bluish-grey metallic luster. Its semiconducting properties are highly valued, as it allows precise control of electrical conductivity when doped with other elements, such as boron or phosphorus. Silicon has four valence electrons, which enables it to form strong covalent bonds, resulting in an extensive three-dimensional lattice in its crystalline form. This bonding is also responsible for the high melting point of approximately 1414 °C and its mechanical strength.

Silicon is the second most abundant element in the Earth's crust, primarily found in the form of silicate minerals such as quartz, feldspar, and mica. Industrially, it is used to manufacture glass, ceramics, concrete, and silicones. In electronics, silicon is the foundation of modern semiconductor technology, being the material for integrated circuits, transistors, diodes, and

solar cells. Its ability to form a stable oxide layer (silicon dioxide) is critical in microelectronics, acting as an insulator and protecting components. Additionally, silicon plays a minor biological role, particularly in plant structural integrity. The versatility, abundance, and unique semiconducting properties of silicon make it indispensable in both everyday materials and advanced technological applications.

10.5.3 Germanium (Ge)

Germanium is a metalloid element with atomic number 32, positioned in group 14 of the periodic table. It exhibits properties intermediate between metals and nonmetals, particularly notable for its semiconducting behavior. Germanium is a lustrous, hard, gray-white brittle solid that can form a diamond-like crystalline structure. Its four valence electrons allow it to form covalent bonds with other elements, giving it a tetrahedral crystal lattice similar to silicon. Germanium's semiconductor characteristics are essential in electronics, where it was historically used in early transistors and diodes before silicon largely replaced it. Despite being less abundant than silicon, germanium remains crucial for specialized electronic applications due to its higher electron mobility and effective performance at higher frequencies.

Naturally occurring germanium is usually obtained from sphalerite ores and other minerals in small quantities. It has a melting point of 938 °C and is chemically stable in air but can oxidize at high temperatures. Germanium forms compounds such as germanium dioxide (GeO_2) and germanates, which have industrial uses in optics and electronics. In fiber optics and infrared optics, germanium's transparency to infrared radiation makes it valuable for lenses and windows. It also serves as a component in semiconductor alloys and as a dopant in silicon devices to modify electrical properties. Overall, germanium's combination of semiconducting ability, optical utility, and chemical stability secures its role in advanced technology, particularly in high-speed electronics and infrared optical systems.

10.5.4 Arsenic (As)

Arsenic is a metalloid with atomic number 33, located in group 15 of the periodic table. It exhibits properties of both metals and nonmetals, making it versatile in industrial and electronic applications. Pure arsenic exists in several allotropes, with gray arsenic being the most stable form at room temperature. Gray arsenic has a metallic luster, is brittle, and conducts electricity, displaying semiconducting behavior. Its five valence electrons allow it to form covalent bonds in compounds with metals and nonmetals, giving rise to a wide range of chemical species such as arsenides, arsenates, and organoarsenic compounds. The element's toxicity is well-known, and it has been historically used in pesticides, wood preservatives, and alloys, although usage is now heavily restricted due to health concerns.

In electronics, arsenic is an important dopant in semiconductors such as gallium arsenide (GaAs), which is used in high-speed and optoelectronic devices, including LEDs, laser diodes, and microwave-frequency integrated circuits. Arsenic compounds also play a role in alloying metals to improve hardness and corrosion resistance. Despite its toxicity, arsenic occurs naturally in the Earth's crust, primarily in minerals like realgar and orpiment. Industrially, arsenic is handled with extreme care due to its carcinogenic nature. Its combination of semiconducting properties, chemical versatility, and historical significance in materials science demonstrates why arsenic is an important metalloid in both technology and chemistry.

10.5.5 Antimony (Sb)

Antimony is a metalloid with atomic number 51, located in group 15 of the periodic table. It exhibits both metallic and nonmetallic properties, appearing as a silvery, lustrous, brittle solid at room temperature. Its five valence electrons enable it to form covalent and ionic compounds, contributing to its chemical versatility. Antimony has been known since ancient times and is used in alloys, flame retardants, and electronics. It has a relatively low melting point of 630 °C compared with other metalloids and is chemically stable in air, forming a thin oxide layer that protects it from further oxidation.

Industrial applications of antimony are diverse. When alloyed with lead, it improves hardness and mechanical strength, which is especially important in batteries, bullets, and bearings. Antimony trioxide is widely used as a flame retardant in plastics, textiles, and coatings, making it crucial in fire safety materials. In electronics, antimony acts as a dopant in semiconductors, particularly in diodes and infrared detectors. Naturally occurring antimony is found in the mineral stibnite (Sb_2S_3), and although relatively rare, its unique combination of properties makes it valuable in metallurgical, chemical, and electronic industries. Antimony's role as a metalloid lies in its ability to bridge metallic conductivity and chemical reactivity, which makes it indispensable in certain technological applications.

10.5.6 Tellurium (Te)

Tellurium is a metalloid with atomic number 52, located in group 16 of the periodic table. It exhibits characteristics of both metals and nonmetals, appearing as a silvery-white, brittle crystalline solid. Tellurium has six valence electrons, allowing it to form covalent and some ionic compounds. Its semiconducting properties, combined with its chemical reactivity, make it useful in a variety of industrial and technological applications. Tellurium has a melting point of 449 °C and is chemically stable in air, forming a thin oxide layer when exposed to oxygen. It occurs naturally in the Earth's crust, primarily in combination with gold, silver, and other metals in tellurides.

Industrially, tellurium is widely used in alloys to improve hardness, machinability, and resistance to corrosion, particularly in lead and copper alloys. Its semiconducting properties are exploited in thermoelectric devices, which convert temperature differences into electrical energy, and in cadmium telluride (CdTe) solar cells, which are an important form of thin-film photovoltaic technology. Tellurium is also employed in the production of certain electronic components, including diodes and infrared detectors. Its combination of metalloid properties, electrical conductivity, and chemical versatility makes tellurium an important element in electronics, energy, and metallurgy. While relatively rare, tellurium's applications in advanced materials highlight its significance in modern industry and technology.

10.5.7 Polonium (Po)

Polonium is a metalloid with atomic number 84, located in group 16 of the periodic table. It is a rare and highly radioactive element, discovered by Marie and Pierre Curie in 1898. Polonium exhibits both metallic and nonmetallic properties, appearing as a silvery-gray, volatile, and brittle solid. Its radioactivity dominates its chemical behavior, as it undergoes alpha decay, producing significant heat and radiation. Polonium has six valence electrons, allowing it to form compounds with halogens, oxygen, and metals, although these compounds are less

stable due to its intense radioactivity. Its melting point is approximately 254 °C, and it is chemically reactive, particularly with oxygen and halogens, forming polonides and oxides.

Polonium occurs in trace amounts in uranium ores and is typically produced artificially by neutron irradiation of bismuth. Due to its extreme radioactivity, polonium has very limited applications, primarily in scientific research, as a heat source in space satellites, and in static eliminators used in industrial settings. Its highly toxic and radioactive nature means that it must be handled with specialized equipment and strict safety protocols. Polonium exemplifies the properties of a metalloid that straddles the boundary between metals and nonmetals, while its radioactivity introduces unique chemical and physical challenges. Despite its rarity and dangers, polonium has provided significant insight into nuclear physics and the behavior of heavy metalloids.

17. The Theory of Everything

The Theory of Everything

Foundations, Applications and Corrections to General Relativity

Michael Scott Peck

Original: July 24th, 2012 | Final: May 20th, 2013 | Copyright 2013 | [Contact](#)

Abstract: Corrections to general relativity are derived from classical theory and applied to the standard model. The perspective offered is the conceptual inverse of Einstein's theory, where particles exist as localized fields. These vacuum fields undergo affine transformations that are locally invariant with respect to the space-time metric. It is demonstrated that the proper vacuum solution to the Einstein-Maxwell field equations is the limit of the single particle vacuum field solution. The existence of event horizon within Einstein's field equations is linked to the application of point-like sources in the local field theory. With vacuum field theory, it is observed that event horizon can no longer form without infinite classical energy. Gravitational waves are also discussed relative to the use of point-like sources in Einstein's field equations and similar geometric field theories. Methods for determining the space-time metric of any object on a per particle basis are provided. The continuous model of the universe is further introduced, where the solutions to several grand cosmological problems are discussed. It is demonstrated that an asymptotically flat universe will appear linear with respect to local observers. The inferred accelerated expansion is an illusion due to local geodesics deflecting towards the center of an asymptotically flat, linear universe. With recent constraints on the abundance of faint blue galaxies and observed evolution, Λ CDM is found to be off in galactic number densities by $70\% \pm 15\%$ and $104\% \pm 25\%$ at $0.5z$ and $1.0z$ respectively. These galaxies are also observed to be similar to local disk and irregular populations, where Λ CDM underestimates their size by $200\% - 300\%$ prior to $0.7z$. This implies that an expanding model predicts the incorrect shape of the universe, which induces systematic lensing errors. After eliminating all viable explanations, an expanding universe is conclusively ruled out. The purposed model however agrees with all observations by applying only classical assumptions. The shape of the universe for example supports a central core, which is responsible for the cosmic background radiation. It is further argued that Einstein's field equations are incompatible with such universe due to predictions of event horizon.

Preface

Due to the diversity of subjects discussed, this page is meant to provide an overview of the paper. The two theories included herein are referred to as vacuum field theory and the continuous model of the universe. These deeply interrelated theories are necessary for complete consistency between general relativity and cosmology. The cosmological aspects are further applied to rule out various theories of general relativity. The foundations of vacuum field theory arise from three postulates with respect to a unified field theory. These postulates however are only introduced for additional insight, as vacuum field theory can be derived from classical laws of physics. The first is a $1/r$ gravitational potential for any particle, e.g. an electron or proton. The second is Einstein's equivalence principle, where a particle in one local frame will be identical to itself in any other local frame.

Rather than particles being point-like sources, it is argued that they instead exist as localized fields throughout chapters 1 and 2. The foundations of vacuum field theory can be viewed in terms of waves travelling through a relative medium. The medium is an energy density with respect to the localized field interpretation of particles. It is demonstrated that the gravitational and electric potential of a charged, non-composite particle are directly proportional to vacuum energy density. After applying the Lorentz transformation to $1/r$ fields, deceleration of a charged particle is directly related to a change in vacuum energy density in the form of bremsstrahlung. The field dynamics within classical electrodynamics are mimicking those of vacuum field theory. It is therefore possible to formulate theories with point-like sources that agree with observations. The underlying vacuum energy density however defines a locally isotropic space-time metric for general relativity. It would therefore be incorrect to treat the space-time metric as an additional medium for the continuum limit of point-like sources to influence. By knowing how each

particle's field varies due to the background field induced by all others in consideration, the effective space-time metric can be determined for objects on a per-particle basis. With the per-particle method provided in chapter 2, it is observed that conical singularities or event horizon can no longer form without infinite energy. The application of point-like sources and coupling to space-time metric is responsible for the predictions of event horizon and gravitational waves in modern general relativity. Direct proof for the localized nature of particles will therefore arise from a null detection of gravitational waves with direct methods. Although a null result would invalidate the coupling of point-like sources to a space-time metric, the cosmological aspects already rule out any theory that allows event horizon from finite energy.

The cosmological model is the central discovery of this paper, where it is demonstrated that the universe is asymptotically flat. In other words, the inferred accelerated expansion is an illusion due to local geodesics deflecting towards the center of the universe. With redshift arising from relative motion and gravitational potential, the observed state of the universe can only be fit by accelerated expansion or an asymptotically flat shape. All observations are further in agreement with a linear, asymptotically flat universe as discussed throughout chapter 3. These include galactic number densities, angular size versus the absolute magnitude of faint blue galaxies and time-dependence. Although Hawking radiation is theorized to exist with respect to event horizon, the 3000 K temperature of the cosmic background radiation would require the core to be many orders of magnitude less massive than the Moon. Countless galaxies and clusters are however continuously flowing towards the center of the universe. The central core must therefore be more massive than any local object, i.e. the observed cosmic background radiation offers direct proof against the existence of event horizon.

1 Foundations

Throughout the history of modern physics, many attempts at developing a viable unified theory have been made. These attempts have diverse underlying principles, most lacking physical interpretation. Without providing the entire unified field theory, it is possible to reformulate general relativity with three fundamental postulates. These postulates are derived from classical principles, which are further discussed relative to the standard model and Einstein field equations (EFE). Methods are derived that allow the effective space-time metric of any object to be determined. These require a revision of general relativity for several reasons, which are discussed throughout the first and second chapters.

(1) EFEs are based upon a continuum limit of point-like sources, which act locally on the space-time metric. The metric is in return mimicking the localized nature of particles; i.e. similar to classical electrodynamics, space-time acts as a medium for waves. However, the actual field general relativity depicts is the underlying vacuum energy density. This includes contributions from classical and semi-classical fields, although only the electromagnetic field is thoroughly discussed herein. It is argued that the principle of locality is invalid and particles exist as localized field rather than point-like objects.

(2) Special relativity demands that a particle's field will deform from variations in relative motion. General relativity should be restricted to similar mechanisms. When multiple particles interact, the field of each is deformed due to its locally invariant nature with respect to the space-time metric.

(3) All massive particles are known to display electromagnetic behavior; however, EFEs decouple mass from the electromagnetic field. Therefore, the Schwarzschild solution cannot represent realistic objects, as even neutrons display non-zero magnetic dipole moments and

electromagnetic form factors.

Postulate I: Classical forces are mathematical constructs, approximating the time dependence of vacuum fields. At this introductory level of vacuum field theory, the focus is directed at both classical and free field force(s). Classical force is a time-dependent variation that acts upon a point-like particle. Regardless if the discussion is general relativity or electrodynamics, force determines the time-dependence of momentum and position. Some theories also produce abstract fields that are related to position and momentum. For example, quantum mechanics provides probabilistic wave functions of an underlying semi-classical system.

In a free field theory, force refers to the action at each point in space; this is not necessarily in a classical sense. For example, say a field existed that represented a single electron. Regardless of the underlying complexity, the electron will have a classical location in space. In addition to the finite energy density at the particle's classical position, all other points in space will have finite energy density. The introduction of another electron displaced from the original would further vary the underlying field at all points in space. The force(s) between these two localized fields arise from the infinitesimal action of the effective field at each point in space. Particles are therefore localized entities displaying action at a distance.

References to quantization throughout this paper refer to reducing the localized nature of particles to point-like objects. Classical forces are then applied to determine the time dependence of position and momentum in quantum systems. This is

achieved by applying the Lorentz transformation or space-time metric to a scalar field; the scalar field is related to vacuum energy density. The vacuum energy density of a single particle can further be approximated with classical theory, i.e. $1/r$.

Postulate II: All particles consist of localized vacuum fields. This is relevant to the concept of field-particle duality, where all particles display decaying fields and point-like structure. Under the most fundamental considerations, objects would cease to exist if matter was not localized. This

The complex component of equation (1) takes a similar form (4).

$$(A^2 + k^2)^{-} v \rightarrow = 0 \quad (4)$$

Solutions to (4) can be determined from the scalar component via (5).

concept can further be extended to what it means 15

for a field to be localized. Localization requires that

$v \rightarrow_{lm} = \frac{A\theta}{k}_{lm}$

the underlying energy is self-reinforced, i.e. any stable vacuum field will not dissipate over time. For this to be plausible, at least two forms of field energy must exist. A complex scalar-vector field is defined below; however, this is inadequate for an interacting theory. Additional degrees of freedom are instead required for time-dependent evolution. The main objective of this paper is to bridge the gap between electrodynamics and general relativity. These additional degrees of freedom can therefore be ignored by applying classical forces.

Vacuum energy density is related to a complex scalar-vector field that is conserved throughout interactions.

Conservation of vacuum energy can be achieved by introducing the continuity equations (6) and (7). These relate to an underlying geometric structure after reducing the additional degrees of freedom.

The equations however do not depict the correct time dependence of vacuum fields, which can instead be approximated with classical

for the far-field of massive particles to decay as

$\frac{1}{r} \frac{d\phi}{dr}$ in Planck units; this is later discussed in section (1.3). With the wave function = $\phi +$

theory. All that is required after the quantization of a localized field is the Lorentz transformation and assumption of $1/r$ gravitational potentials. The application of a single scalar-vector field is important due to the quantization process. It is therefore assumed that ϕ remains constant at the classical position of a massive, non-composite particle. These variables are further related to a scalar invariant and motion of a point contained within ³.

$$v \rightarrow -A\phi$$

complex Helmholtz equation (1) is needed. It is

$$\frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \nabla^2 \phi + A^2 \phi$$

superficially similar to the Schrodinger equation of a free particle, but does not depict probability.

$$(A^2 + k^2)$$

$$= \frac{1}{c^2} \nabla^2 \phi + \frac{1}{c^2} A^2 \phi$$

Equation (1) can be divided into both real (2) and imaginary (4) parts.

$$(A^2 + k^2)\phi = 0$$

In spherical coordinates, solutions to (2) involve spherical harmonics and Bessel functions (3).

$$\phi_{lm} = j_l(kr) Y_{lm}(\theta, \varphi) \quad (3)$$

The complex Hamiltonian density defines the vacuum energy density at any point in space with a quaternion norm (8).

$$\frac{1}{c^2} \nabla^2 \phi + \frac{1}{c^2} A^2 \phi = \frac{1}{c^2} \nabla^2 \phi + \frac{1}{c^2} (\nabla \phi)^2 + \frac{1}{c^2} A^2 \phi$$

The linear wave solutions can be quantized with equation (9), i.e. the point of maximum field energy depicts the classical energy.

$$E = \frac{1}{c} \int \phi \nabla \phi \cdot d\mathbf{r} \quad (9)$$

Postulate III: The vacuum field is the result of transforming the non-linear geometric degrees of freedom. A crucial metaphysical aspect of matter is usually overlooked in modern physics, i.e. what do particles physically consist of? Initial attempts tried to attribute a physical substance to matter, or material upon space that formed particles. However, this perspective is plagued by cyclic reasoning, i.e. if such substance existed, what would be the physical essence of it? Indeed this reasoning is no different from the modern concept of fields. For example, electromagnetic fields are mathematical constructs created in abstract to understand the universe at the quantized level. At any scale however, one fundamental property of the universe is undeniable; i.e. space itself.

In the classical perspective space is a rigid, time-independent structure that quantized mechanics is founded upon. Switching to the more abstract view of general relativity, the properties of space vary from the Euclidean model. It becomes possible to deform space, varying the location of a continuum of points in a smooth manner. However, Einstein's view of relativity is incompatible with quantum mechanics and the cosmological model discussed in chapter 3. He also applied the only physical property of the universe to a single classical force, i.e. gravity. It is demonstrated in section (1.4) that general relativity can be reinterpreted as a tool for quantization. Einstein's perspective is therefore the conceptual inverse of vacuum field theory. This opens a profound path to unification, as an underlying geometric structure can be used to depict all classical forces. Therefore, unification no longer refers to the energy scale where classical forces merge into one, but instead the manifestation of all forces from a single unified field. The essence of matter can now be attributed to something that is physically real rather than a mathematical construct.

Assuming tensors and/or a geometric foundation are capable of fulfilling the first two postulates, the necessity for additional degrees of freedom is clear. The vacuum field represents the energy of an underlying geometric structure, although the actual structure is beyond this paper's scope. Regardless, acknowledging its existence offers an intuitive explanation for the universe. It is trivial that a geometric structure should be time-dependent if it does exist. Therefore, in the most general sense matter is nothing more than fluctuations of space itself. These are much smaller than the macroscopic world, as vacuum field theory indicates structure at the Planck scale. Without assuming Planck scale fluctuations of space are responsible for fields and matter, there is literally no other way of writing a unified field theory. For example, the standard model applies several scalar-vector fields to complete symmetries and fill gaps; however, they are solely mathematical constructs.

As earlier theories developed, the original aether became a resistive medium throughout space rather than mysterious substance that formed particles. This transition was the product of the corpuscle theory of light, attributed to Newton. It was later argued against with the Michelson-Morison experiment, which tested for a variation in the speed of light relative to the local motion of Earth. This concept of anisotropy is flawed, which had been pointed out by Hendrik Lorentz^[A]. As an object's momentum varies in a local frame, the field is transformed in such a way that any anisotropic affects cancel. Motion is instead relative to the vacuum field of all other particles and aether only becomes conceptually crucial for a single particle universe. In other words, the scalar vacuum energy density creates a relative medium upon space, which must further be applied to determine the effective space-time metric.

10.5.7.1.1. Space

When describing the dynamics of particles in a gravitational field, general relativity is a useful theory. However, the theory can be interpreted in two unique perspectives. The first is the mainstream view, where the four-dimensional manifold is to be taken literally; i.e. the coupled entity of space-time physically exists. Gravitational acceleration within this perspective is not induced by the curvature of space, but instead the curvature of time. The curvature of spatial components only varies the path of particles in motion or under classical force. Regardless, the physical existence of space-time is crucial for not only Einstein's interpretation, but also the validity of mainstream astronomy and cosmological models. The undermining of this perspective however originates from the predictions of singularities and gravitational waves. The use of quantized point sources in the geometric Laplace equation (EFEs) produces these artifacts. Particles instead exist as localized fields rather than point-like sources and must be treated as so.

The second view insists that the universe does not physically exist as a four-dimensional space-time manifold. Space and time should instead be treated as two independent entities; i.e. space is depicted by a classical manifold mapped to physical locations, while time is a manifestation of relativity or event comparison. Within this perspective, the vacuum field of each particle is relative to a Euclidean reference space ($y_{\mu\nu}$), which can be arbitrarily chosen. Upon this reference space or frame, particles exist as localized scalar-vector fields. Affine transformations are applied to the field of each particle rather than varying the metric of space, in agreement with the principles of special relativity and electrodynamics. Any reference space is further held static so that it does not allow metric or gravitational waves.

In differential geometry, a metric is defined that maps all points (ξ) bound to a manifold in space to a curvilinear coordinate system (x), or vice-versa. This can be defined as an infinitesimal variation in distance between two points in space, with respect to the original configuration. A two dimensional example is depicted in figure 1.1. It is always valid to vary the points on a manifold as long as they never overlap; this is referred to as a Riemann manifold. This deformation is possible due to the infinitesimal property of nature, i.e. there exist an infinite number of infinitesimal intervals between two points in space.

In another perspective or the one previously argued for is the deformation of a ruler, which consists of a linear lattice of atoms. These atoms contain many quantized particles, although each is relatively localized at a single point in space. When vacuum energy density increases, each particle's field must remain locally invariant with respect to the space-time metric. The space-time metric is therefore encoding the deformations or affine transformations experienced by localized fields, which can further be treated as point-like objects. In a situation where the gravitational potential is increasing with respect to time, distance from the perspective of the ruler remains constant. However, an observer in the reference space will note that length contraction of the ruler has taken place. Only the lattice of atoms becomes deformed with respect to the reference state rather than space itself.

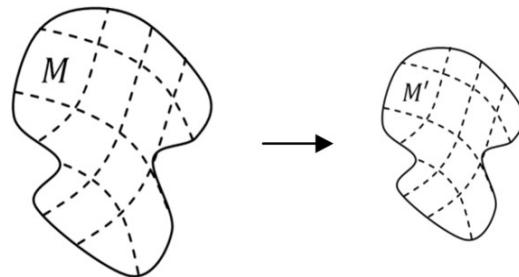


Figure 1.1. A smoothly connected manifold (M) is transformed from an initial state (M) to a final state (M').

10.5.7.1.2. Time

Quantization or the transformation of a localized field into a point-like object allows for the creation of systems. A system exists in a finite region of space and may contain many quantized particles, usually coupled to kinematic equations. Therefore, systems are time-dependent and obey classical energy conservation, i.e. Lagrangian formulations. Within classical mechanics, the trajectory of a particle through a potential can be determined with the Lagrangian equation (10); where L is the kinetic energy minus potential.

$$\frac{d}{dt} \left(\frac{\partial L}{\partial q_i} \right) - \frac{\partial L}{\partial q_i} = 0 \quad (10)$$

The 1-dimensional Lagrangian can be written as (11), where V is the classical potential.

variables evolve at a decreased rate. When quantum mechanics is discussed, it is demonstrated that the intrinsic spin rate of an electron also decreases as vacuum energy density increases. The Dirac equation is further related to relativistic field dynamics, discussed throughout section (1.6). It is clear that time is more fundamental than the kinematics of point particles, i.e. time is a scalar quantity that depicts change in abstract variables.

Under certain conditions, it is possible for time to be undefined. If no energy exists, then there is nothing to compare and space at all scales becomes Euclidean ($y_{\mu\nu}$). The other extreme is a region of infinite vacuum energy density such as an event horizon or conical singularity. Time in this situation is undefined because the underlying field(s) cannot evolve. At this limit, quantum mechanics fails under classical considerations since the observables

$$1 \quad \frac{1}{2} m_0 \dot{x}^2 - V(x) \quad (11)$$

become static. Einstein's field equations allow such

$$L = \frac{1}{2} m_0 \dot{x}^2$$

In this perspective, time is a comparative scalar related to a change in position with respect to some constant rate of observation. It allows for a coupling between quantized energy and the rate at which quantized fields move through space.

Relativity introduces additional complexity as the rate of observation can vary between different scenarios. Both special and general relativity

require a Lorentz scalar, where an increase in vacuum energy density forces a field to evolve at a decreased rate. This time dependence of classical variables is seen throughout various phenomenon including decay rates, classical kinematics and intrinsic spin. For example, it is experimentally known that an unstable particle has a relatively longer half-life when moving with respect to a background field. Particle decay depends upon internal degrees of freedom, which cannot be explained by classical

mechanics. The longer half-life is instead modeled by relativity, where all

anomalies due to the use of point-like sources in a local geometric field theory. This results in the non-linearity between vacuum energy density and space-time metric. The only way to produce infinite vacuum energy density in vacuum field theory is with infinite classical energy, which is impossible.

EFEs are more abstract with respect to time, or space-time. For example, time-dependent variations in the stress-energy tensor can produce gravitational waves. By applying quantized mass and momentum in

sufficiency of ch. These localized fields, in which some scenarios allow quantized variables to be transformed into geometric waves. General relativity is however depicting the

The to el a
d m
of e
p w
oi o
nt r
- k
li t
ke o
so d
ur et
ce e
s r
be m
co i
m n
es e
ge t
o h
m e
et ti
ri m
c e
in d
na e
tu p
re e

vacuum energy density of field(s) responsible for classical forces. Time in a more general sense is therefore a comparative scalar between the various variables of a quantized system or underlying field theory.

10.5.7.1.3. Energy

Effective energy is defined from quantized rest mass and momentum (12).

$$E = \sqrt{(\text{pc})^2 + (m_0 c^2)^2} \quad (12)$$

Returning to the classical wave-like equation (1), effective energy can also be written as a quaternion norm. With a particle's scalar-vector field (), quantized energy is related to the vacuum energy density (13). From the third axiom, this density is

related to the underlying geometric degrees of freedom. The goal is to transform the non-linear geometric structure into a scalar-vector field that is linearly proportional to classical energy.

$$\nabla \quad \underline{\quad}$$

With the second axiom and the expected inverse distance far-field, a spherically symmetric solution is possible. This is derived from the classical wave-like equation and is only meant to approximate the field's envelope. After applying spherical Bessel functions, two linear wave solutions are found for non-composite particles in Planck units (16). These solutions have an E_o/r far-field with energy density at the classical position proportional to E_o .

$$\underline{\sin(r)^2} \quad \underline{\sin(r)} \quad \underline{\cos(r)^2}$$

$$\Psi(r) = E_o \sqrt{(\pm \frac{1}{r}) + (\pm \frac{1}{r^2} \pm \frac{1}{r})} \quad (16)$$

The approximate wavelength that corresponds to an electron or positron is therefore 2π in Planck units.

M SI^{*} units however
(13) wavelength is $2\pi l_p$,
where Planck length is
defined as (17).

In order to
quantize the
 $\underline{\Delta}$
vacuum field,
similar methods are
applied with respect
to the linear wave
approach (9); i.e.
applying a Dirac
delta defines
quantized energy
(14).

$$\underline{l} \cong 1.616199 \cdot 10^{-35} \text{ m} \quad (17)$$

E^∞ As previously stated,
(14) the scalar-vector
notation

f is not adequate for
time-dependent
evolution; i.e.

Ψ (16) finds no real
application beyond
approximating

r
)
 δ
(
 r
)
 d
 r
 r
 θ

* Meters in the International System of Units (SI) is the base unit of length. It is defined as the distance traveled by light in a vacuum in 1/299,792,458 of a second.

The wave function can be written as (15), where quantized mass depicts the scalar field and momentum replaces the vector component. Scalar mass no longer exists solely at the particle's center, as it is a fundamental part of all field solutions.

$$(0) / c^2 m_0 + i \bar{c} \bar{p} \rightarrow \quad (15)$$

With the quaternion norm used to define vacuum energy, it is possible for scalar mass to be negative. Relative to the Dirac equation, the charge conjugate is applied to ensure only positive mass exists. Negative scalar mass however always results in positive vacuum energy density, which depicts the gravitational force. Thus ∇ will always be positive

the far-field vacuum energy envelope. In general, solutions for actual electrons and positrons are only approximated by equation (1) and the resulting linear solutions (16). All that is required by the purposed postulates is for vacuum energy density to be indefinitely localized in space, creating a stable $1/r$ far-field. Planck units are used due to a relation between the space-time metric and vacuum energy density. This effectively sets $G_o = \hbar_o = c_o = k_e = k_B = 1$; i.e. the far-field gravity-electric potential of an electron is defined as (18), where σ is the charge to mass ratio. Classical energy variations from this potential are now directly proportional to

variations in vacuum energy and the quantized scalar field will have the same sign as charge for non-composite massive particles.

density or $\nabla \cong E_o/r$

$$U = \nabla^{(1 \pm \sigma)} \quad (18) \\ : \sigma = \frac{q_p}{m_p} \\ 15.15612(63)$$

Energy (12) can be reformulated by introducing the Lorentz scalar defined by (19), which is relative to a local frame of reference or space-time metric.

Newtonian energy principles (25) can be derived by equating $\langle g \rangle = \langle \cdot \rangle$. The general Lorentz scalar is defined relative to $1_{\mu\nu}$, while the velocity within γ is with respect to $g_{\mu\nu}$.

$$\langle \cdot \rangle = \sqrt{1 + (\nabla \cdot)^2} = \frac{\langle \cdot \rangle}{c} \quad (19)$$

$$\frac{1}{c} \frac{\nabla \cdot}{\nabla^2} \frac{E}{\Delta E} \quad (20)$$

$$\sqrt{1 - \frac{c}{v}} \frac{E_o}{E_o - \frac{\Delta E}{\Delta t}}$$

The Lorentz factor scales rest energy, resulting in the effective quantized energy (20).

Special relativity defines force as the change in proper momentum with respect to metric time

(20)

(26). The following notation will be used for common variables; $u = dx'/dt$ is proper velocity,

The Lorentz factor (19) is simply the ratio of energy to rest energy, and has a range from 1 to ∞ . For agreement between general frame, a

$v = dx'/dt$ is metric velocity and $w = dx/dt$.

$$f \rightarrow v \rightarrow +\gamma m a \rightarrow \frac{dp}{dt} \rightarrow \frac{\gamma^3}{m_0} v_a$$

scalar field $\langle g \rangle$ is introduced.

Generalization of $\langle g \rangle$ is

c_0

achieved by looking for a function similar to $\langle \cdot \rangle$ with a range from 1 to ∞ . This is accomplished by first defining the reference vacuum energy

$$\frac{dt'}{2^o} \quad (26)$$

density (21), which any observer will consistently measure as constant. Δ plays the role of E_o as defined in (20), which is a product of quantized energy

being proportional to vacuum energy density.

G_o

The equations can be simplified by considering an object moving at escape velocity along

c the gradient of ∇

4 . If ∇ is a

single static

(particle, the escape velocity

1 (27) relative to the particle's field is derived

$)$ via $(= (g$.

Taking the limit of (27) as

$r \rightarrow \infty$ is ; $\bar{v} \rightarrow$;

$= 0$, while the limit as $M \rightarrow \infty$

results in ; $\bar{v} \rightarrow$;

$= c_o$. When

transforming

(27) to the frame relative to a distant observer

or Δ , the

velocity as $M \rightarrow$

∞ is ; $\bar{w} \rightarrow$; =

0 implying

infinite

vacuum

energy density. This is

a consequence of $\nabla \rightarrow$

∞

rather than $M \rightarrow \infty$.

The net vacuum energy density is defined as (22).

There also exists a simple relation (23)

between Δ and Δ

similar to $E = E_o$.

$$(22) \quad \frac{\Delta}{\Delta} = \frac{1 - \frac{2G}{r}}{1 - \frac{2G}{r}}$$

(27)

$$\frac{\Delta}{\Delta} = \frac{1 + \frac{2G}{r}}{1 + \frac{2G}{r}}$$

$$\frac{\Delta}{\Delta} = \frac{\gamma_g}{\Delta}$$

Acceleration is derived by differentiating (27) with

The domain can also be extended from 0 to ∞ when considering reference frames within a local field. However, all applications within this paper use a reference frame where the local source is removed

$(r \rightarrow \infty)$, or as ∇

$\rightarrow 0$. To ensure equivalence as previously purposed, the correct equation meeting

all
re
qu
ire
m
en
ts
is
(2
4).

$$\frac{\frac{1}{\sqrt{1 + \frac{2G}{r}}}}{a} = \frac{\frac{G_o M}{r}}{\frac{dt^2}{r \gamma_g}}$$

Acceleration relative to the space-time metric at escape velocity is therefore equal to (29).

10.5.7.1.4. Continuum Mechanics

A localized field can be represented as a group of sections or wave fronts, where the space-time metric is applied for the quantization process. This

Applying the contravariant and covariant metrics together results in any underlying field to be transformed back to the $y_{\mu\nu}$ state; this is due to the equivalence principle and equation (30).

reference to μ
quantization is $=$
not with respect δ
to quantum v
gravity, but
instead the

transformation of
localized fields into
point-like objects.
From the perspective
being argued within
this paper, the proper
view is that of
localized fields
deforming due to the
presence of other
localized fields.
Classical mechanics
are therefore replaced
by continuum
mechanics and
resulting metric(s). For
example, a finite
manifold (M) can be
equipped with an
arbitrary metric ($g^{\mu\nu}$).
The metric will then
undergo affine
transformations as
depicted in figure
1.2. For simplicity,
each manifold is
smoothly connected
and semi-rigid.
Manifolds of this type
are useful for
describing vacuum
fields, although the
concept is further

complicated by intrinsic
spin. Relative to vacuum
field theory, each particle
exists as an independent
manifold ($g^{\mu\nu}$) that is
equipped with a scalar
field. The scalar field or
vacuum energy density is
in return relative to the
preferred reference frame.
Each localized vacuum
field also has a reference
state (0) with respect to the
space-time metric. With
special relativity and
classical electrodynamics,
it is known that additional
energy or momentum
induces length contraction
of the underlying field(s).
This can be extended to
general relativity by
applying the spatial
components of the
contravariant metric tensor.
Figure 1.3 demonstrates
how general relativity
allows for the quantization
of a localized field.

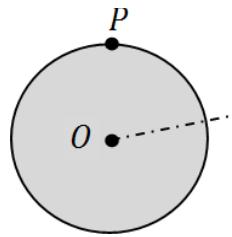


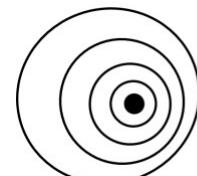
Figure 1.2. A finite 2-dimensional manifold is depicted in an initial (O) and final state (O'). Several affine transformations are applied including translation, rotation and deformation.

Transforming each field in this manner allows for quantization, where the velocity of each point upon the field is equivalent in both direction and magnitude. The dynamics of the field can therefore be reduced to the point-like particle perspective.

Without quantization, each point along a particle's manifold will travel at various velocities. This will further require continuum mechanics to determine the proper translation, rotation and deformation.

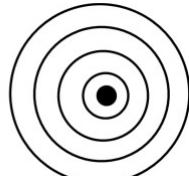
A particle's manifold will vary depending upon the background vacuum energy density that it

resides in. Under realistic considerations, the particle also influences the gravitational potential generated by the background field. If the effective energy and proper velocity of each particle is known, then it is possible to determine the effective space-



time metric of any object; this is discussed in chapter 2. If a particle is moving with respect to a local field, both special and general transformations must be

applied. From (30), it was shown how wave fronts become equivalent to the $y_{\mu\nu}$ frame when space-time is deformed. Any Lorentz boost (38) must be applied relative to this configuration and then mapped from $g_{\mu\nu} \rightarrow y_{\mu\nu}$. This ensures that all field transformations are invariant for local observers; i.e. any stationary observer



will view space relative to the space-time metric.

Figure 1.3. Sections of two manifolds $g^{\mu\nu}$ are depicted relative to a specific reference frame. (Left) The frame is with respect to the Euclidean frame $\eta_{\mu\nu}$. (Right) The frame is with respect to the deformed space-time metric $g_{\mu\nu}$.

Assuming a particle's classical position is with respect to the preferred reference frame ($y_{\mu\nu}$), r_n denotes an independent coordinate system for each. The notation can be simplified by introducing ∇^{mn} , which is the non-effective field from particle m to a point in the n^{th} coordinate system. The effective field from particle m at the same point relative to the n^{th} coordinate system is ∇^{mn} . Equation (31) is based upon the equivalence principle, where each particle is relative to the effective background field or space-time metric.

$$\nabla$$

$$\nabla$$

If a point or region of infinite vacuum energy density (∇) exists, (g also becomes infinite. The previous mapping (33) fails at boundaries of infinite vacuum energy density or points beyond them. For example, if the field of an external particle is calculated, any radial lines at or beyond the singularity will be mapped to the event horizon. The wave fronts or sections also become non-continuous, violating the assumption of smoothly connected vacuum field manifolds. It is known from classical electrodynamics, QED and QCD that particles exist as localized fields. Due to these fields

$$- (r) \quad \{ \begin{matrix} m = n : 0 \\ m \neq n : \nabla_m(r_n) \end{matrix} \quad (31)$$

following the metric of space-time, they

$$\text{become } \begin{matrix} mn \\ n \end{matrix} \quad m \neq n : \nabla_m(r_n)$$

The effective field of particle n due to all other particles is defined by equation (32),

restricted to any boundary of infinite vacuum energy. Since these fields are in return responsible for all classical forces including gravity, a black

$$- (r) = - \left(\begin{matrix} r^7 \\ n \\ n \\ n \end{matrix} \right) \quad (32)$$

where r^7_n is determined by integration (33).

with event horizon display external fields. Finite black holes are predicted to exist with

$$\begin{aligned} & \int_{r_7}^{r_n} \nabla_m(r_n) dr_n \\ &= \Sigma \nabla \end{aligned}$$

(33) $r_e^s p_e^c c_t^u t_f$

o ield theory,
although their
surfaces must
have finite
vacuum energy
density.
Therefore, black
holes should not
only
demonstrate

Relative to the Euclidean line element (dS), the transformed radial coordinate is defined as (34).

$$r_n^r = \int_{r^-}^{r^+} ds = f \frac{ds}{dr}$$

(34) energy density from finite classical energy. From

the methods
herein, the effective field of any
object

The line element can be written in terms of the metric tensor (35).

$$(ds)^2 = g_{\mu\nu} dx^\mu dx^\nu$$

(35)

If only a single non-composite particle existed, ($g = 1$) and the field would be in the original configuration with

can now be determined down to the Planck scale. With the advent of QCD and resulting states of dense quark matter, it is now possible to model the finite fields of quark stars and black

would exhibit the E_o/r of event horizon and agreement with classical theory. The maximum value of this localized field is proportional to E at the classical position.

holes. The name given to black holes remains valid since they demonstrate near perfect black body spectrums and immense gravitational fields. Energy will escape over time due to relativistic jets and free field far-field envelope in radiation. The non-existence

singularities is later discussed with respect to the cosmic background radiation and observed shape of the universe.

10.5.7.1.5. Electromagnetic Fields

Relativistic electrodynamics provides additional insight into how the vacuum field of a particle varies due to relative velocity. As the momentum of a uniformly charged particle increases, the electric field lines and magnitude loss isotropy. The electric field in terms of the particle's classical position is given by (36)^[B], where $\theta = 0^\circ$ is parallel with $\vec{v} \rightarrow$.

$$(1 - \zeta_c$$

For a massive non-composite particle, vacuum energy density is not proportional to the electric field. The electric potential can instead be related to vacuum energy density as demonstrated by equation (18). Returning to the foundations of vacuum field theory, there is a distinction between quantized energy and vacuum energy density. This is because quantized energy is with respect to the amount of vacuum energy density at the classical position. It is

$$E_{\text{q}} = \frac{k}{r} \frac{\gamma^2 \sin^2(\theta)}{1 - \zeta_c} \quad (36)$$

important to notice however that the gravitational potential is both equal to vacuum energy density and the variations in classical energy due to a

The magnitude of the electric field is therefore (37). potential. Since quantized energy has only two

$$E_q = \frac{k}{r} \frac{\gamma^2 \cos^2(\theta) + \sin^2(\theta)}{1 - \zeta_c} \quad (37)$$

components, the classical dynamics of a point-like particle can therefore be retained. The remaining

$$\gamma$$

$$e r^2$$

As the particle's momentum increases, the electric field is compressed in the direction of motion. The magnitude of the field perpendicular to motion also increases, while the electric field

tangent to the particle's trajectory weakens. As the limit of $v \rightarrow c$ is approached the field becomes compressed into a cylindrical plane of infinite vacuum energy density, depicted in figure 1.4. These field dynamics with

respect to the field dynamics
 Lorentz are derived by
 transformation applying a
 crucial to Lorentz
 transformation
 to the vacuum
 field of a
 particle.

The Lorentz transformation Λ (38)^[B] allows mathematically defined objects to be transformed in space-time. The objects in this case are individual manifolds equipped with a scalar field, which depicts the vacuum energy density of each particle.

the application of point-like sources in classical

△

e

1

e

c

t

1

0

d

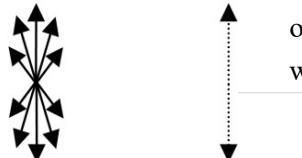
y

11

e				
v				
er				
d				
o				
es				
n				
ot				
in				
di				
c				
at				
e				
th				
at	γ	$-\gamma\beta_i$	$-\gamma\beta_j$	$-\gamma\beta_k$
th	$F_{-\gamma\beta_i}$	$(1 + \beta_{ii})$	β_{ij}	β_{ik}
e	$l_{-\gamma\beta_j}$	β_{ij}	$(1 + \beta_{jj})$	β_{jk}
u	$L_{-\gamma\beta_k}$	β_{ik}	β_{jk}	$(1 + \beta_{kk})$
se				
of				
p				
oi				
nt				
-				
li				
k				
e				
s				

o
ur
c
es
in
ot
h
er
fi
el
d

theories is valid.
Where $\beta_{\mu\nu}$ is defined as (39).



(38)
)]

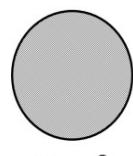
8

$\mu\nu$

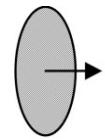
β^2

$\bar{v} \rightarrow$

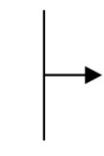
In
general
any
vector
can
transfor
med
(40)
includin
g the
electro
magneti
c 4-
potentia
l.



$v = 0$



$v < c_0$



$v = c_0$

R'
= Λ
 R
(4
o)

Two
consecutive
Lorentz boost
can be
determined with
matrix
multiplication
(41).

$\Lambda(v^-_1 +$
 $v^-_2) =$
 $\Lambda(v^-$
 $_1)\Lambda(v^-_2)$
(41)

Figure 1.4. (Top) The electric field of an electron is provided in several states: (Bottom) An electron's manifold undergoing length contraction in the direction of motion.

With respect to the electromagnetic nature of matter, vacuum field theory would be incomplete without discussing photons. From the approximate field of a non-composite massive particle (16), vacuum energy density is determined from the classical gravitational potential and linear wave-like

Each photon or localized packet of electromagnetic energy remains localized due to the reinforcement of vacuum energy density. Assuming the symmetry of a photon is cylindrical, Bessel functions of the first kind (42) are applied perpendicular to the direction of propagation. equation (1). This says nothing about how the ∞

$(-1)^m$
 x_{2m+n}
 Electromagnetic
 Field relates back to
 the unified

$$\left\{ \begin{array}{c} J_x \\ = \\ \Sigma \\ C \end{array} \right\} \quad \boxed{4}$$

scalar-vector field.

There is also a
 distinction

$$\frac{m! \Gamma m + n}{+ 1}$$

$$\sum_{m=0}^2$$

between individually localized photons and free electromagnetic energy. Classically, an electro-magnetic field can be described in terms of a superposition of waves. However, there is no guarantee that these are individual packets of vacuum energy. Unlike the electric field, superimposed vacuum fields do not display field interference; i.e. positive and negative electric field contributions will result in no electric field, while vacuum energy density is always positive. It is therefore possible to use Fourier series to create a super-

position of many waves, although these will not represent actual particles. Two processes in nature provide insight into the distinction between free electromagnetic energy and localized photon (or quantized particle); these are electron-positron annihilation and bremsstrahlung.

Electron-positron annihilation demonstrates the

If a linear wave equation is used similar to (1), the gradient of the scalar component must have the opposite sign of the vector field. This will essentially require for the central region of the field to have quantized mass.

Therefore, the linear-wave approximations cannot be taken literally. An actual photon will have no scalar mass at the classical position,

particle nature of the electromagnetic field, where $a = E \sqrt{(\pm J_0)^2 + (\pm J_1)^2} \approx \frac{4}{3} \sqrt{r}$

particle and its anti-particle produce gamma rays after colliding. The localized scalar

and will therefore be anti-symmetric with respect to the vacuum scalar field. The wave equation is instead applied to approximate the far-field envelope under consideration of vacuum energy conservation.

The exact nature of the underlying vacuum field is therefore irrelevant, as only vacuum energy density is required with respect to general relativity.

Applying the linear wave approximation (1), a photons field perpendicular to propagation is (43).

$$\underline{E}_0$$

$$(r, \theta, z) = E_0 e^{-|z|} \sqrt{(J_0)^2 + (J_1)^2} \approx \frac{4}{3} r e^{-|z|}$$

fields of the electron and positron cancel, creating two or more massless photon. Assuming conservation

of energy, the resulting photons will split quantized energy. Since the total vacuum energy of a single electron is infinite, only the conservation of classical energy can be considered. The gamma rays resulting from the annihilation process also remain localized in space indefinitely. This has implications for the CMBR, or black body spectrum observed in all

In order to determine the 3-dimensional vacuum energy density of a single photon, the envelope in the z-direction is required. Without a rigorous method capable of providing the exact vacuum field of a photon, the linear wave

approximation must once again be applied. The vacuum field of an individually localized photon in Planck units is therefore (44), where coordinates are with respect to the particle's classical position.

$$\underline{E}_0$$

directions of local space.

$$\nabla (r, \theta, z) = E_0 e^{-|z|} \sqrt{(J_0)^2 + (J_1)^2} \approx \frac{4}{3} r e^{-|z|}$$

(4)

$$e^{-|z|}$$

-

Bremsstrahlung is the second case of electromagnetic energy, where a charged particle passing close to another emits braking radiation. Unlike the previous case, electromagnetic energy is radiated as a free field over a range of frequencies. The free

Relative to quantized energy, the Lorentz scalar is the only free variable. Therefore γ will assume two states, i.e. prior to bremsstrahlung ($(_1)$) and after ($(_2)$). The variation of vacuum energy density between these states can be derived by subtracting

field should be treated as an independent manifold. ∇

with respect to other localized fields and particles.

In other words, the free electromagnetic

∇ from (γ_1) , i.e. equation (48).

E_O

$$\Delta_{EM} = ($$

locally invariant with respect to the space-time

$$\sqrt{(_1 \cos \theta)^2 + (\sin \theta)^2}$$

metric. This emission of electromagnetic energy in the framework of point-

like sources is similar to the gravitational waves in general relativity. In general, theories that couple fields to point-like sources will

generate waves upon the respective medium due to variations in a quantized source.

The emission of electromagnetic energy can

further be related to a variation in vacuum energy density. For the non-relativistic case, the radial

The initial and final states from (47) are expanded via Taylor expansion at $\zeta = 1$ resulting in (49).

$$1 + (\gamma - 1) \sin^2(\theta) -$$

Ignoring higher order contributions, equation (49) reduces to a dipole field approximation (49) (50).

$$\Delta \frac{\nabla^{EM}}{V} \approx \frac{E_0}{r} \quad (50)$$

$$\begin{matrix} \Delta \\ \{ \\ s \\ \} \\ \theta \\ r \end{matrix}$$

$\frac{S}{q} \rightarrow \frac{f}{\theta}$ transformation derived from classical

$$\frac{4\pi r}{\theta} \text{ electrodynamics is therefore directly related to a}$$

The electromagnetic energy radiated per unit solid angle is therefore (46)^[B].

fundamental scalar field.

The electromagnetic field cannot be easily related to vacuum energy, since

$$\frac{dW}{dt d\Omega} = \frac{q^2 a^2}{4\pi c^3 \sin^2(\theta)} \quad (46)$$

The first order approximation of bremsstrahlung

indicates that a charged particle will emit a free field electromagnetic dipole. It is important to realize that electromagnetic energy in this situation is not localized, but

continuous over a range of angles and frequencies. It is distinct from the annihilation case, where two photons are emitted at unique angles in order to preserve quantized energy. With results from chapter 2, the effective vacuum far-field of a moving electron is defined by (47).

only the electric potential is proportional to ∇ . Therefore, the r^2 in (45) is expected, while vacuum energy decays proportional to the inverse distance. The vacuum energy emitted in terms of non-relativistic motion becomes (51).

In consideration of a unified field theory, all particles would exist upon a single field in space. However, vacuum field theory depicts each localized field as a deformable manifold relative to a reference space ($y_{\mu\nu}$). In this perspective, vacuum energy density (∇) radiated due to bremsstrahlung is physically detached from the electron's manifold into a free field described by Maxwell's equations.

$$\nabla = \frac{E_0 \gamma}{\sqrt{(x')^2 + (y')^2 + (z')^2}} \quad (47)$$

$$a(51)$$

10.5.7.1.6. Quantum Mechanics

Quantum mechanics was introduced by Erwin Schrodinger, who had initially attempted to create a relativistic theory. Due to the many difficulties related to the relativistic form, the time-dependent Schrodinger equation (52)^[C] was instead published.

$$ik \cdot = H = (V - \frac{k^2}{2m} A^2) \quad (52)$$

In consideration of a classical potential such as the electric field produced by a proton, the probability of an electron being detected at any given position is $|\psi|^2$. This should not be confused with the wave-function attributed to vacuum field theory, which depicts a Hamiltonian density. The wave function of quantum mechanics can be interpreted in various ways. Vacuum field theory agrees with the path-integral approach to quantum mechanics, where

two super positioned plane-waves, so (53) remains true for all photons.

$$E = kw = hf = hc/\lambda \quad (53)$$

The photon's luminal field can be carried over to fermions. By applying spacetime algebra, it is observed that the field of an electron orbits the spin-plane at the speed of light. If this is true, then there must be kinematic effects due to the coupling between light-like field dynamics and space-like trajectories. The classical structure of the electron is discussed at the end of this section, for now the Dirac equation (54)^[E] is examined for its connection to relativistic field dynamics.

$$ik \cdot = H = (ca \cdot p^\wedge + \beta m_e c^2) \quad (54)$$

This can be rewritten in a more intuitive way since $m_e c^2$ is actually related to the intrinsic spin (55).

each particle has a
classical location in
space. It
attributes no physical
meaning to the
quantum wave

$$\begin{aligned} i & \& m_e \\ \wedge p & = \\ \{ & \cdot \\ + & Q \end{aligned} \quad (55)$$

other than probability.
This is known as the
minimalist perspective or
ensemble interpretation
attributed to Max Born^[D].

In order to comprehend fermion spin and mass, the field of spin 1 particles must be initially discussed. Photons are the most fundamental spin 1

c & t k
k

Setting the rest energy
of the electron equal
to the spin angular
frequency (53) results
in a spin radius of
(56); this is the
reduced Compton
wavelength. The
radius (r_e) is constant
relative to the metric
of space, tracking a
set of points along the
field.

particle and can be
either polarized or
non-polarized. For

a circularly
polarized
photon, the

$r_e = k/m_e c$

spin state is either $\pm k$. To simplify the problem, the photon will be reduced to a plane-wave that has a helix shaped electric field. Relative to a massive particle located at a fixed point in space, the propagating EM field will appear to spin around a fixed axis. The electric field however is actually traversing space at the speed of light perpendicular to the spin plane. The field is therefore not spinning with respect to the reference frame. The quantized energy of a photon can be written with respect to

the perceived angular frequency or wavelength by

equation (53). Circular

(56)

to

k
 $k = \sum_k \alpha_k$

In consideration of the quantization process used within relativity, the objective is to demonstrate that tracking a single point upon the electron's field satisfies the equations of motion at the classical position. To simplify the motion of the field, a local orthogonal coordinate system (e_1, e_2, e_3) is defined relative to the classical position.

Historically, Schrodinger was the first to apply the Heisenberg picture in order to determine the time dependence of the position operator (57).

$$\frac{d\mathbf{r}}{dt} = \mathbf{p}$$

polarization
is identical

After integrating (57) twice with respect to time, the position operator becomes (58).

Spin arises as a bi-vector defined as (63)^[F], where $\gamma_2\gamma_1$ is the spin plane.

$$\begin{array}{c} \text{x} & \text{(o)} & \text{t} \\ (t &) &) \\) & & = \\ = & & x \\ x & p \\ \downarrow & k \\ & i \\ & s \\ & v \end{array} \quad \begin{array}{c} \text{k} \\ \text{s}_k \\ \text{i} \\ \text{e} \\ \text{v} \end{array} \quad (63)$$

$$\begin{array}{c} - & k & k & ikc & cp \\ - & o & o & R\gamma_2\gamma_1R & \\ 2 & + & 2H(\alpha_k(k)(e^{-2ik\nu}-1) & \text{For the free} \\ & & H & \text{wave solution to the Dirac equation, the} \\ & & - & \text{Dirac rotor becomes (64)^[F].} \end{array}$$

The last term is the complex quantum oscillation known as zitterbewegung; it is complex due the connection with spinors.

Furthermore, the first two terms provide the classical trajectory of the particle, which is the average zitterbewegung path. Applying the Heisenberg picture, particle motion is combined

with a non-classical rotation of the field and cannot

$$\frac{\partial \mathbf{R}}{\partial t} = \frac{2mc^2}{e} \mathbf{R} \cdot \mathbf{\Omega} = \frac{1}{e} (64) e_1 e_2$$

The time dependence of the local coordinates is related to the angular velocity bi-vector via (65)^[F].

de_μ

be directly interpreted in the classical sense. From (58), the zitterbewegung angular frequency and

On the spin plane $\mu = 1, 2$ resulting in (66), where t is relative to metric time.

$$\begin{array}{c} = \\ \pm \\ \Omega \\ \cdot \\ e \\ \mu \end{array} \quad (65)$$

$$\begin{array}{c} . \\ e \\ \mu \end{array} \quad (66)$$

radius
are
(5
9)

$$\begin{aligned}
 & 2H \quad c \quad : = \pm \quad _1 \quad _2 \quad _1 \\
 & 2m \quad k \quad \left\{ \begin{array}{l} s \\ \bar{s} \end{array} \right\} \quad dt(k) \\
 & \frac{w_{zwa}}{c} \quad \frac{r_{zwa}}{= \quad \frac{e}{k}} \\
 & \frac{e}{k} \\
 & \frac{d}{e} = \pm \quad 2m_e c^2 \\
 & \frac{2}{2} \\
 & w_o \\
 & 2m_e c = \\
 & \pm \\
 & \frac{2}{2} \\
 & m \\
 & e \\
 & c \\
 & \frac{2}{2} \\
 & e \\
 & e \\
 & e
 \end{aligned}$$

Spinors in general require two rotations in order to
 $\frac{dt}{dt}$

1 2 2

return to the initial state. Picking half the classical radius upon the spin-plane remains a valid option. It is claimed that the field has an

angular velocity equivalent to the speed of light; therefore, the angular frequency also doubles. Choosing the classical radius (56) simplifies the situation

since it returns to the original state after a single rotation; i.e. the relativistic spin period is defined as (60).

These equations define the time dependence of the local coordinate system attached to the electron's classical position. Since an electron moving at the speed of light violates relativity (19), intrinsic spin and zitterbewegung must be field related.

Comparing the Heisenberg approach to the geometric algebra derivation, the only self-consistent interpretation of the Dirac equation is mechanical in nature. The Heisenberg picture

demonstrates that the position operator is following a complex, light-like trajectory.

$$\frac{h}{E_0} \approx \frac{eC^2}{c} = \text{phy} \quad (60)$$

E
o
m
o
v
=

To develop a mathematical model of the local coordinate system and spin,

spacetime algebra is applied. Geometric (spacetime) algebra allows the geometric product to be

defined as (61)^[F].

demonstrates
that the
electron has a
classical
velocity (v) and
an attached
coordinate
system at the
local position.
In addition, a
multivector
rotation is an
active
transformation,
which acts on
a field

$$uv = u \cdot v + u A v \quad (61)$$

independent of the reference coordinate system.

The
original
space-
like
geodesi
cs
therefor
e be

The orthogonal
reference vectors
are related to the
initial set by
Lorentz spinors
(62)^[F], i.e. $SL(2,$
 $C)$.

modified so
that the field
always follows
time-like
geodesics.
This in return
allows the
relativistic

(62)

dynamics of a
localized field to be
reduced to a point in
space-time.

The classical structure of electrons/positrons is required to further the localized field interpretation of the Dirac equation. Quantum theory hides the localized nature of particles through Lorentz transformations. Similar to general relativity, a

Measuring the magnetic field along the spin-plane (at r_e) results in a magnetic dipole moment that is twice the Bohr magneton (72). If the zitterbewegung radius is used instead, the dipole moment becomes equivalent to the Bohr magneton. localized field can

be reduced to a point like object. μ

Since the Dirac equation is $B = 2m_e$

Lorentz invariant, knowing the trajectory of any point along the field allows all others to be determined. Therefore, it is assumed that the active transformation applied to the spin-plane carries over to all other points along the field. According to classical electrodynamics, which is implied by quantum theory via minimal coupling, a moving electric field will produce a magnetic field equal to (67).

It is obvious that the field generated by a spinning electron is not a true magnetic dipole. This is irrelevant until the hyperfine structure, where the nucleus interacts with the electron's far-field. The relativistic electric field also deforms with respect to the appropriate Lorentz transformations, while vacuum energy density must also be included.

Ignoring higher order effects, the approximate evolution of a spin $\frac{1}{2}$ quantum system can be

$$\vec{B}^{-\rightarrow} = \frac{E_0}{c^2} \left(\vec{v} \rightarrow \times \vec{E}^\wedge \right)$$

The active

(67) described by relativistic field dynamics. Returning

transformation must be light-like acting on the entire field of the electron;

i.e. the magnetic field becomes (68).

to the spin bi-vector, the equation can be expanded with the geometric product (73)^[G].

$$S = i s v = i(s \cdot v + s A v) \quad (73)$$

B — The Hodge dual (74) allows for an identity between the wedge product and cross product.

$$(68) \quad s A v = * (s \times v) = i(s \times v)$$

cr

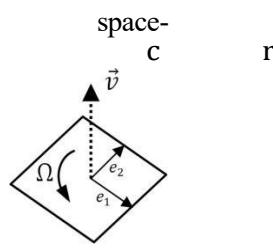
The classic definition for a magnetic dipole is (69), where m is the dipole moment.

$$B = \frac{k_m m}{r^3} \begin{pmatrix} \cos(\varphi) \hat{i} \\ \sin(\varphi) \hat{j} \\ \varphi \hat{k} \end{pmatrix} \quad (69)$$

Since it is claimed that the Dirac equation is specific to a single point on the spin-plane, equations (68, 69) are combined resulting in (70).

$$k_e q_e = \frac{k_m m}{r}$$

(70)



P
l
u
g
g
i
n
g

B
—
=
k_e
q_e
(68)
)
cr

The spin bi-vector (75) is composed of a real scalar

S = isv[cos(β) + i n sin(β)] $\quad (75)$

Figure 1.5 demonstrates how quantum mechanics reduces a field's relativistic

spin to a single point in

s

h
e
c
a
s
s

i
c
a
l
r
a
d
i

u
s

(
5
6
)

a
s

a
d s
e p
f i
i n
n e
d m
i a
n g
t n
h e
e t
c i

D d
i i
r p
a o
c l
e e

q m
u o
a m
t e
i n
o t
n e

r q
e u
s a
u l
l l
t t
s o

i (

n 7

1

)

$$\begin{matrix} m \\ = \\ m_e \\ = \\ q \\ f \end{matrix}$$

(71)

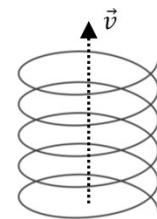
μ

e

e

e

Figure 1.5. For
electron
moving in a
straight line
without external
field, $\beta = \infty$
to
ensure the fields
velocity
remains co.



10.5.7.1.7. Applying Vacuum Fields

Vacuum field theory bridges the gap between general relativity and electrodynamics. It is argued within the previous sections that particles exist as localized field(s). This is contrary to Einstein's field equations, which depict particles as point-like sources. In perspective, EFEs take the quantized attributes of localized fields and couples them to a geometric field equation. The principle of locality is essential to Einstein's interpretation of general relativity, i.e. it allows point-like sources to have a local effect on the surrounding space-time metric. Quantum mechanics on the other hand demonstrates that local hidden variable theories are invalid via Bell's theorem. By applying vacuum field theory, the Dirac equation can be interpreted as a quantized field theory. The non-local hidden variables become visible once spacetime algebra is applied, which reveals the underlying relativistic field dynamics. Ignoring the significance of localized fields rather than point-like sources will result in theories that allow gravitational waves. For example, both EFEs and the Brans-Dicke theory allow these waves. Gravitational waves however have not been ruled out experimentally, although one has never been directly detected. Gravitational waves and the current probability of non-existence are further discussed throughout section (2.2).

The Lorentz transformation is a crucial aspect of quantum field theory or classical electrodynamics. Section (1.5) demonstrated that an accelerating charged particle emits vacuum energy density. This is related to the emission of electromagnetic energy in terms of bremsstrahlung. Therefore, application of locality in classical electrodynamics is valid because the electric potential is directly related to vacuum energy density. Application of point-like sources in quantum mechanics is also valid when considering the process of quantization. This allows

for hidden non-local variables in agreement with Bell's theorem and action at a distance. Quantum mechanics itself is based upon quantized variables such as rest mass, position and momentum. Spin only complicates the situation by offsetting the point of quantization from the classical position, i.e. the electric field will always travel at the speed of light. Satisfying the equations of motion for a point along the field will automatically solve all others.

It is difficult to define vacuum energy density in terms of the additional fields within QFT and the standard model. However, the electric potential is proportional to vacuum energy density with the far-field approximation. The mass of a single fermion is also directly proportional to an underlying scalar field. The energy density at the classical position of a particle is crucial in terms of the quantization process. It allows a localized entity to be reduced to a point-like object, where classical mechanics can be applied. Vacuum energy density can also exist without the presence of classical fields. Neutrons for example are massive compared to electrons, but demonstrate minimal electromagnetic properties. Although the charge of negative and positive quarks can cancel, the underlying vacuum energy density must be conserved. The unified field theory would therefore require classical quantized forces to be abandoned. This is not required to arrive at a theory that predicts the outcome of any experiment. For example, it is always possible to include additional factors into mathematical models for agreement with observations. However, this does not mean the resulting theory will depict what is actually taking place. It will also be difficult if not impossible to arrive at an exact formulation connecting general relativity and the standard model. With vacuum field theory, far-field approximations can instead be applied in order to arrive at a perturbative theory of everything.

The Dirac equation written in covariant form is defined as (76).

$$(\gamma^\mu \partial_\mu + i \frac{m_0 c}{k}) = 0 \quad (76)$$

Since Planck units offer a natural scale for vacuum fields, the equations for the remainder of this section are written with $k = c = G = 1$. The metric relative to vacuum field theory in accordance

In order to include general relativity or external fields into equation (81), $\partial_\mu \rightarrow D_a$ is the covariant derivative (77) with respect to the local frame.

$$\underline{m_0 c}$$

$$\begin{matrix} (i) \\ \gamma \\ D \\ a \end{matrix}$$

$$\underline{k}$$

$$\bar{0}$$

with section (1.5) is isotropic, i.e. it must be defined by a single scalar field (ϕ_g).

Therefore, the vierbein

defined in (78) is directly related to the effective vacuum energy density (84).

$$e^a = \gamma = 1 + \nabla$$

$$(84)$$

-

μ

g

The anholonomic Dirac matrices (γ^a) are related to the Dirac matrices (78) by a vierbein field.

Neglecting any sub-structure of the nucleus or self-interactions, the static field is approximately (85).

$$\gamma^a = \frac{\bar{e}_\mu{}^a \gamma^\mu}{\nabla^p} \quad (78) \quad (85)$$

$$\nabla^p \approx$$

$$r$$

The vierbein field is related to the metric tensor via

The Dirac equation with an external field is (86).

$$g_{\mu\nu} = \frac{\partial E^a}{b} \frac{\partial \bar{E}^b}{\mu} \quad (79)$$

$$a \quad b$$

$$\frac{v}{y} \quad [i\gamma^\mu + (\partial_\mu +$$

$$w)$$

1 $\frac{ah}{\varrho}$

4 w

$y_{ab} - ieA_\mu \partial^\mu x^\nu$	$= 0$	(86)	respects that more exact pre-derivatives are covariant for variations from the metric of space and any external electro-magnetic field defined upon it.	D	$\frac{-m_e c}{\hbar}$	$= 0$	(82)
Equation (78) finds its origins from the commutator of the Dirac matrices (80).	In order to ensure that the field remains the light-like, the spin connection (87) ^[H] must be introduced into the covariant derivative.	The Christoffel symbols ($\Gamma^{\mu}_{\nu\lambda}$) are further derived from vacuum energy density and resulting space-time metric.	fered reference to the metric of space and any external electro-magnetic field defined upon it.	ions that are μ μ k	electron. This usually involves solving continuous fractions by iteration discussed in section (2.6).	Since the inverse distance is an approximation for the far-field, it is also necessary to apply an energy cut-off when the radius is 1 in Planck units. This ensures that vacuum energy density does not surpass the maximum value depicted by quantized classical energy.	
[γ _a , , γ _b] = 2y _a In order to relate the local tetrad frame to the metric, the general commutator is defined as (81).	framed measured symbols ($\Gamma^{\mu}_{\nu\lambda}$) are further derived from vacuum energy density and resulting space-time metric.	frame measured (...) be co lized s to be ac co un te d fo r, i.e. .th e nu cl eu s an d (i e a γ _μ	twice measured in direction vi due all by located calibrated ld s to be ac co un te d fo r, i.e. .th e nu cl eu s an d (i e a γ _μ	1	$D_\mu = \partial_\mu + \frac{1}{4} w_\mu^{\alpha\beta} \epsilon^{\alpha\beta\gamma} \Gamma^\nu_\gamma$		(83)
$[e^a \gamma^a, e^b \gamma^b] = 2g^{\mu\nu}$	$w^{ab} = e^a \partial_b + e^b \partial_a + e^a e^b \Gamma^{\nu}_{\nu ab}$	(81)					
The space-time indices can be raised or lowered by applying $g^{\mu\nu}$ or $g_{\mu\nu}$ respectively; i.e. (88).							
The local frame is attached to the electron's classical position and remains light-like. Therefore, the Dirac	$e_\mu = \frac{1}{2} [e_\mu \gamma^a, e_\nu \gamma^b] e_\mu$	equation with					(88)

10.5.7.1.8. The Standard Model

The standard model is an extension of quantum field theory, which is based upon classical electrodynamics and special relativity. It includes several

The neutral components are included within the interaction Lagrangian density (91)^[1], along with the running coupling constants. In order to ensure that $SU(2)_L$ invariance is not violated, a current (J^Y) other fields such as the electroweak and Higgs, is added which preserves the symmetry.

which model weak interactions and mass.

Many of

the previous principles from the Dirac equation

$$L^{(W)} = \frac{J^Y B^\mu}{ig J^3 W} \quad (91)$$

$$\frac{3\mu}{i}$$

carry over to quantum electrodynamics (QED) and the standard model. The Lagrangian density (89)^[1] of QED for example consists of the Dirac equation and classical electromagnetic contributions.

$$int \quad \mu$$

$$2^\mu$$

The $SU(2)_L \times U(1)_Y$ gauge group (92)^[1] shares similarities with the 2-dimensional rotors applied in space-time algebra.

$$QED \quad - \quad \frac{L}{4} \quad F_{\mu\nu} F^{\mu\nu} \quad L \quad L$$

$$= \frac{1}{e^{i\theta_a \sigma^a}} \frac{1}{e^{i\beta^a \tau^a}} \quad (89)$$

$$(92)$$

QED is formulated with classical fields coupled to spinning light-like manifolds.

In this perspective,

the mass term offers no additional insight beyond

semi-classical Lagrangian dynamics. Due to this,

extensions of quantum theory fail to explain the

In canonical form, even sub-algebra solutions (R^+)

physical essence behind classical fields and mass. This

$$= \frac{1}{2} (1 + \gamma^5) = \frac{1}{2} (1 - \gamma^5) \quad (93)$$

$$= \frac{1}{2} \gamma^5 \quad (93)$$

however does not make the theory useless, as experiments can only

m_e
 e_a

sure quantized variables including rest mass, position and momentum.

The electromagnetic field is also closely related to vacuum field theory and depicts time dependence of quantized charged particles.

The transition from QED to a more general theory requires the addition of neutral currents and

to the Dirac equation for \tilde{G}^0 are (94)^[G]. (94)

$$= \sqrt{\rho e i \beta} \overline{R}^2$$

The connection between (92) and (94) arises because the group of 2D rotors and unitary group U(1) are locally isomorphic. However, there are two unique copies of U(1): U(1)_{EM} and U(1)_Y with generators defined as (95)^[I] respectively.

Y

$$\begin{array}{ccc} : & : & Y \\ Q & (95) & \\ = & & \\ 2 & & \\ + & & \\ I_3 & & \end{array}$$

weak interactions. The standard model unifies weak

interactions and QED with electroweak theory, defined by the SU(2)_L \times U(1)_Y gauge group. Neutral vector bosons A_μ (photon) and Z_μ (Z^0 mass eigenstates) are

coupling constants of SU(2)_L and U(1)_Y (g and g' respectively).

The second component of (92) is locally isomorphic to SO(3), although requires two complete rotations in order to return to the original state. The Pauli matrices define the axis of rotation, while θ_a is the gauge parameter; i.e. the amount of rotation on each spin axis. Since the lie algebra of SU(2) and SO(3) are isomorphic, the general rotor in R^3 is defined as (96). Thus $i\sigma^a$ are the infinitesimal generators of SU(2), similar to T_a .

$$\begin{array}{ccc} A_\mu & & (90) \\ \cos \theta_W & & \\ \sin \theta_W & & \\ B_\mu & & \end{array}$$

$$\begin{bmatrix} Z_\mu \\ \underline{ } \\ \sin \theta_W \\ \cos \theta_W \\ W^3 \end{bmatrix} = R = \cos \frac{\theta}{2} + T_3 n \sin \frac{\theta}{2}$$

Returning to the conical solution of the Dirac equation, the beta factor encodes the angle between the spin-plane and velocity. Furthermore, the rotor determines the rotation of the field with respect to spin-coordinates. Combining these properties with the $SU(2)_L \times U(1)_F$ gauge demonstrates the degrees of freedom for the underlying field. Either the spin-plane to velocity angle is varied, or an active 3-dimensional rotation is applied with respect to a rotational-axis. Similar to the Dirac equation, the process of quantization is crucial to understanding these transformations. It allows a localized field to be treated as a point-like object. Spin for example

The Higgs field was introduced in 1962 by Philip Anderson to compensate for the lack of mass for gauge bosons within the standard model. The relativistic model was further developed in 1964 by independent groups who were awarded the Nobel Prize. The additional field predicted the existence of a Higgs boson, which gives mass to other particles. The mass of the Higgs boson can be theoretically determined from the mass of the top quark and w -boson. Earlier measurements of these particles predicted a Higgs boson with an expected mass of 85^{+54}_{-34} GeV; however, recent world averaged values (March 2012)^[K] of the top quark and W-boson vary allows the quantization process to take place from the classical position. This allows an active 3-

model therefore predicts that the mass of the Higgs boson $7-9$ with ranges 123 theoretical from GeV several running coupling constants determined experimentally^[J]. Any unified field theory should have zero free variables except for the fundamental constants (c_0 , G_0 , \hbar , k_e), which depict underlying properties of space. For example, the electron rest mass should be determined from the only stable non-composite solution to the unified field theory. Quantized mass takes the particular value due to the non-linear nature of the field, i.e. there is only one stable value. The standard model in original form however does not attribute mass to fermions and other

¹⁹ free parameters and

massive particles. An attempt to resolve this absence requires an additional field and resulting scalar particle, i.e. the Higgs field and boson respectively.

methods. These values are relative to the top quark having a bare mass of 173.2 ± 0.9 GeV and W-boson of 80.399 ± 0.023 GeV.

Direct methods of detecting the Higgs boson initially began at CERN with LEP2.

Preliminary results from data collected over the year 2000 claimed that four LEP2 experiments were consistent to the 2.9 sigma level (“1.4 in 1000 chance of statistical fluctuation”) of a 115 GeV Higgs boson^[L]. This was reported in November 2000, based upon an excess of events over the theoretical background rates. Results were published in December 2001

conserving the previously predicted mass, although the combined probability had decreased to $2.4 \cdot 10^{-3}^{[M]}$. From the four individual experiments, ALEPH provided the most significant results. The excess of events over background was initially placed at 3.4σ , which was further reduced to 3.2σ

in the final report. The more recent large hadron collider began operations in November 2009, which also contains multiple experiments for detecting the Higgs boson. Initial results were released in July 2011 with respect to ATLAS and CMS, showing an excess of events around 144 GeV^[N]. This was compatible with a Higgs boson at the 2.9 sigma

level. An article published in nature around mid-august of the same year later revised the confidence to a sigma of 2.0^[O]. Results that are more recent were published in December of 2011, where ATLAS had a signal at 126 GeV with 3.6 sigma; CMS showed an excess of events around 124 GeV with 2.6 sigma^[P]. An additional weak signal was detected by both experiments around 119 GeV with a 2.1 sigma. Figure 1.6 depicts these recent results and theoretical mass as derived from the top quark and W-boson.

Between LEP2 and LHC, there seems to be a disagreement with the observed excess of events. The results of ALEPH demonstrated a sigma that is relatively close to ATLAS. If the Higgs boson exists in accordance to the standard model, both results cannot be correct. This raises concerns over the understanding of background processes and their contribution to excess events. There also lacks a single region that demonstrates a Higgs boson signal, i.e. mass has varied between individual experiments and runs. This of course can be explained by lack of data, both by the results of LEP2 and preliminary results at LHC. Relative to theoretical predictions, there exists a large margin

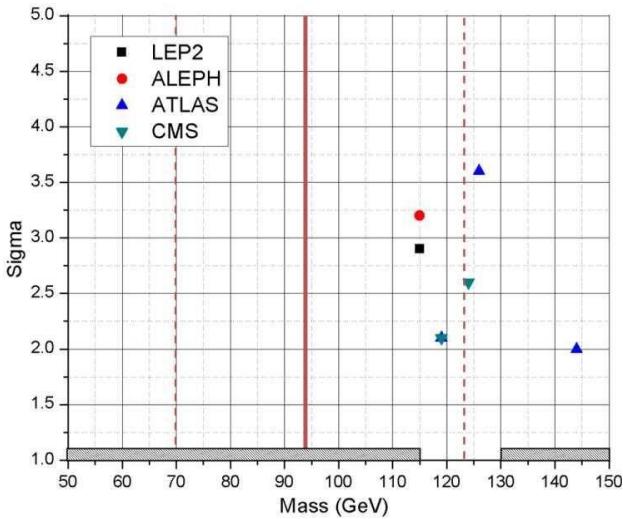


Figure 1.6. The thick red line is the theoretical Higgs mass determined from the top quark and W-boson; dashed lines are error boundaries as of March 2012 for the 68% confidence level^[K]. The hatched area at the bottom depicts regions that have already been ruled out by other experiments.

between experimental results and the preferred mass range. Due to the uncertainty of the top quark's mass, recent claims of a Higgs-like particle between 4.5 to 5.0 sigma (125 GeV)^[Q] are borderline acceptable up to the 68% confidence level^[K]. However, only decay products of the Higgs boson are being directly detected, which coincide with the decay products of other known particles. Even at 6 sigma in agreement with the standard model, there lacks explanation for the physical essence of mass and additional deflection a particle experiences in external fields.

From vacuum field theory, the classical energy (97) of a single fermion is proportional to the point of maximum vacuum energy density. It is this central point of a quantized particle that depicts the kinematics of the entire localized field. This is in accordance with the process of quantization and affine transformations previously applied.

$$E = \sqrt{(pc)^2 + (m_0 c^2)^2} \quad (97)$$

It is also predicted that massive particles such as electrons will have symmetric scalar fields, while massless photons consist of anti-symmetric scalar fields. This symmetry allows massive particles to have a finite amount of scalar mass at the classical position. Massive particles will therefore require momentum to move through an external field due to a localized scalar field rather than a massless scalar-vector field. Mass in QED is not well defined because the Dirac equation uses it to quantize spin. This relativistic spin is in return balanced with the quantized velocity so that all points upon the spin plane move at the local speed of light. The unified field theory should instead reduce to a single scalar-vector field (98), where mass is a localized scalar field ($\pm\emptyset$) depicting matter and anti-matter for each pair of fundamental particles.

$$\nabla(x) = \sqrt{-g} = \sqrt{(\emptyset)^2 + (v)^2} \quad (98)$$

2. Relativity and Differential Geometry

Vacuum field theory requires a single scalar field determined from quantized variables and affine transformations. From this scalar field, it is possible to define the space-time metric similar to Einstein's field equations. Identical mathematical tools are required for either field theory, i.e. differential geometry and Riemann manifolds are necessary for quantizing a field's motion. Vacuum field theory also explains the mechanism behind gravitational force. Variations in time dependence at each point in space forces a quantized field to accelerate. Similar to section (1.4), applying the space-time metric to the particle's manifold ensures the underlying field's time dependence resembles the initial $y_{\mu\nu}$ configuration. Differential geometry and Riemann manifolds are therefore indispensable

With respect to classical electrodynamics, a charged particle will have an effective electric field defined by (99). The velocity must be relative to the metric of space-time, or background vacuum energy density due to all other particles.

$$\mathbf{E}^{\rightarrow} = \mathbf{r} \frac{\gamma}{(\gamma^2 \cos^2(\theta) + \sin^2(\theta))^{3/2}} \quad (99)$$

The electromagnetic field does not easily transform to a particle's vacuum energy density. However, the electric potential is proportional to ∇ with respect to charged particles. Since the electric field lines are parallel with r^{\wedge} , the electric field is proportional to the radial derivative of the effective vacuum energy density (100); where σ is the charge to mass ratio and Planck units are applied.

t
re
la
ti
vi
ty
,re
g
ar
dl
es
s
of
th
e
u
n
d
er
ly
in
g
fi
el
d
th
e
or
y.

$$\vec{E} = -\sigma \left(\frac{\&}{r} \right) \hat{r}$$

A crucial modification to the theory of general relativity is the coupling between point-like sources and the corresponding space-time metric. If this coupling is poor, artifacts will appear under certain

scenarios; i.e. gravitational waves and singularities. After discussing the correct metric from vacuum field theory, it becomes clear that EFEs are using the manifold of space-time in disguise of a localized field. The coupling between EFEs and Maxwell's

equations is also poor, i.e. the contributions to the space-time metric are incorrect. The space-time metric should instead include all vacuum energy components, i.e. the electric and neutral fields. EFEs instead decouple these fields from quantized

$$) \hat{r} \quad (100)$$

Assuming that the field is only compressed relative to the direction of motion, the effective vacuum field is (101).

$$- \frac{E}{1 - \frac{1}{r}} \quad (101)$$

$$\sqrt{(\gamma \cos^2(\theta) + \sin^2(\theta))^2}$$

The partial derivative of the transformed field (101) is therefore (102).

$$\frac{\partial}{\partial r} \frac{-E}{1 - \frac{1}{r}} \quad (102)$$

$$\partial_r \frac{r^3 (\gamma^2 \cos^2(\theta) + \sin^2(\theta))}{r^3}$$

Equation (100) is equivalent to the effective electric field of a moving charged particle (103).

mass, depicting them as separate entities. The

$$- \frac{\nabla \cdot \vec{E}}{V} \quad (103)$$

&r
r
2
2
2
3

Schwarzschild solution in return cannot represent $(\gamma \cos \theta + \sin \theta)$

realistic objects since all massive particles display some electromagnetic component. Attempting to produce a proper field solution via the Einstein-Maxwell equations is also incorrect.

Vacuum energy density is therefore directly related to the electric potential of a non-composite massive particle. This connection allows localized fields to be treated as point-like sources in electrodynamics.

10.5.7.2.1. General Relativity

The foundational aspects of general relativity discussed over the previous sections are sufficient for understanding the coupling between EFEs and classical electrodynamics. The Maxwell-Einstein field equations are introduced (104)^[R], where $T^{\mu\nu}$ is the electromagnetic stress-energy tensor.

μ

contribute to vacuum energy density with respect to a charged particle; i.e. the EM components are already included in (24). The Reissner-Nordstrom metric for a single particle therefore reduces to the derived vacuum far-field of a charged particle after proper field contributions are considered.

Returning to electromagnetic fields, quantized conservation laws can be written as (108)^[B]. The

$$\frac{1}{2} \frac{\partial^{\mu\nu}}{\partial x^{\alpha}\partial x^{\beta}} R = \frac{2G\varrho}{c^2\mu} \quad (104)$$

) is
electromagnetic
stress-energy
tensor ($T^{\mu\nu}$)

$\partial_{\alpha} \partial_{\beta}$

again related to classical mechanics. The Reissner-Nordstrom metric (105) is a static

solution to the Maxwell-Einstein field equations. It defines

the gravitational field around a charged, non-rotating, spherically symmetric object.

$$\partial_{\nu} T^{\mu\nu} + y^{\mu/\nu} f_{/\nu} = 0 \quad (108)$$

The field tensor (109)^[B] is derived from the electromagnetic four-potential, it is therefore related to quantized lagrangian dynamics and should not

$$g^{\mu\nu} \equiv \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix} \quad \gamma^{\mu\nu} \equiv \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

contribute to the vacuum energy density. Once again, there is a distinction between vacuum energy density and quantized energy due to classical force.

Where (γ_g is (106)^[S] and the Schwarzschild radius $F^{\mu\nu} = \partial^{\mu}A^{\nu} - \partial^{\nu}A^{\mu}$)

(109)

is defined as $r_c = 2E/c^2$.

The four-potential (110)^[B] is proportional to the scalar and vector potentials.

$$\begin{aligned} \gamma' &= \frac{1}{1 - \frac{r_c}{r}} \\ &= \frac{1}{1 - \frac{g r_c}{c^2 r}} + \left(\frac{k}{r} \right) \end{aligned}$$

$$\phi = \frac{1}{c} \int_0^r A(r') dr'$$

This is the limit as $\Delta \rightarrow \infty$ relative to the vacuum field definition (107), after the proper electric field contribution (18, 100) is taken into consideration.

Einstein's field equations apply quantized densities in order to determine the curvature and metric of space-time. The quantized source term on

the right side of EFEs (111) allows for mass and momentum

$$\gamma_g = \frac{1}{1 - \frac{r_c}{c^2 r}} \rightarrow \frac{1}{\Delta}$$

applied in a continuum limit. The left

$$\begin{aligned} &= \frac{\nabla^2 \phi}{\Delta} \\ &= \frac{1}{r} \frac{\nabla^2 \phi}{r} \end{aligned}$$

It is no coincidence that equations (106, 107) have similar

single entity. A simple case would be that of an electron, which is spherically symmetric with respect to any co-moving frame. According to the second axiom, the conservation of

field or space-time metric.

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G \rho}{c^4} T_{\mu\nu}$$

With the foundations offered from vacuum field theory, it is clear that Einstein's field equations

vacuum energy creates an E/r far-field via the contain several flaws. Up to this point, the three linear wave-equation reasons initially offered for a revision of general relativity have been discussed.

form. Vacuum field theory unifies mass and external fields into a

The first was the coupling of point-like sources to the space-time metric. The metric further mimics the underlying vacuum field from which matter originates. Similar to electrodynamics, the coupling of a localized field to a point-like source allows for waves upon the relative medium; i.e. the classical

The third flaw is the incorrect coupling between electrodynamics and classical gravitational field. From section (2), it was argued that the electric field is a component of the underlying vacuum field. With the conservation of vacuum energy density or classical theory, a non-composite electromagnetic field or space-time metric of

charged particle should display an E_0/r far-field. general relativity. The

vacuum field of a single non- composite particle however is proportional to its

The Maxwell-Einstein vacuum solution ($T_{\mu\nu} = 0$) gives a squared Lorentz scalar of (112).

electric potential.

When the electric potential of a particle varies relative to some background field, so

$$\frac{\nabla}{\Delta} = \frac{\gamma'}{g^2} = \frac{1}{k_e Q} \quad (112)$$

does the underlying vacuum energy. This was previously discussed relative to bremsstrahlung, where this variation was directly related to the variation in vacuum energy density and Lorentz transformation. Since the space-time metric is

$$1 - 2 \frac{1}{r} + \left(\frac{1}{\Delta} \right)$$

For a single particle, $T_{\mu\nu} = 0$ everywhere except at the particle's classical position. Therefore, (112) is essentially the far-field solution of a single charged particle. However, the Lorentz scalar does not obey

defined solely by vacuum energy density, it is

the E_0/r law as derived under the assumption of incorrect to include where the electric field of a single particle as a separate entity.

The second flaw is similar to the first,

vacuum energy conservation. According to vacuum field theory, the correct particles must exist as localized fields rather than point-like objects. Similar to relativistic electro-dynamics, the field of a particle becomes deformed when the background vacuum energy density varies. Therefore, the metric of space-time is of mathematical origin and plays no role in the physical structure of space. An observer will always view space relative to the metric of space-time.

squared Lorentz scalar for a single particle reduces to (113).

$$\frac{\gamma}{\Delta} = \left(1 + \frac{1}{\Delta} \right)^2 \quad (113)$$

(113) is the limit of equation (112) as $\Delta \rightarrow \Delta$ and reduces to the Maxwell-Einstein vacuum solution after proper coupling of the electromagnetic field via equations (18, 100).

$$\frac{\gamma}{\Delta} = \left(1 - \frac{1}{\Delta} \right)^2 = \frac{\gamma}{\Delta} + \frac{\gamma}{\Delta} \quad (113)$$

However, space does not deform at the large-scale structure as posited by Einstein. This also pertains to gravitational waves, as it is claimed by the third

The limit produces a singular

y when =

Δ , while

postulate that matter is Planck scale fluctuations of space. It is contradictory to allow space-time waves

from the quantized variables of vacuum fields, when vacuum fields are a representation of Planck

scale waves.

$$\text{Furthermore, if } \frac{\nabla}{\Delta} \text{ were responsible for } \left(1 - \frac{\nabla}{\Delta} \right)^2 = \left(1 + \frac{\nabla}{\Delta} \right)^2 \quad (114)$$

classical force, objects with event horizon would lack external gravitational fields; this is contradictory to EFEs vacuum solutions.

(1)
1
5

10.5.7.2.2. Gravitational Waves

The existence of gravitational waves is crucial to the validity of EFEs. From a theoretical standpoint, variations in certain stress-energy moments allow quantized variables to transform into metric waves. If these waves do not exist, then the conservation of energy is clearly violated within EFEs. Indirect evidence for gravitational waves comes from the binary system PSR B1913+16. This system consists of orbiting neutron stars that emit pulsed radio signals at nearly constant periods. After measuring these pulses, their arrival was observed to oscillate over a period of about 7.75 hours. Additional observations allowed for the change in epoch of periastron to be measured, which agreed with the predictions of general relativity to within 0.2%^[T]. However, it is the assumption that EFEs are correct which defines the plausible attributes of the system. It is possible for PSR B1913+16 to have parameters that vary from EFE solutions, i.e. these observations only provide indirect evidence for the existence of gravitational waves. Complications arise from uncertainties in the structure of neutron stars, their effective field, orbital parameters and classical energy flux.

Direct evidence by physically measuring the distortion due to gravitational waves appears to be the only valid option for proving their existence. Several experiments have been conducted over the previous 52 years; however, only LIGO, GEO600 and VIRGO are discussed due to precession. The probability of detecting a gravitational wave from a BH-BH event is approximated from table 2.2.

The theoretical event rates are required for determining the probability of gravitational waves existing. They have varied drastically over the previous 15 years as demonstrated by table 2.3. The running length of each experiment is also provided in table 2.4.

TABLE 2.3. Theoretical event rates

ID	Experiment	Event Rate (yr ⁻¹)		
		NS-NS	NS-BH	BH-BH
A*	LIGO I	0.03	0.25	0.19
	LIGO IE	50	400	400
B*	LIGO I	100	30	500
	LIGO I	20	1	2
	LIGO I	[2 · 10 ⁻⁴ , 0.2]	[7 · 10 ⁻⁵ , 0.1]	[2 · 10 ⁻⁴ , 0.5]
	LIGO II	[0.4, 400]	[0.2, 300]	[0.4, 1000]
	LIGO I	0.01	0.02	4.9
	LIGO II	45.1	85.8	21,425
DII	LIGO I	0.002	0.01	0.05
	LIGO II	9.5	42.8	242
	LIGO I	[0.015, 0.15]	-	[0.01, 1.7]
	LIGO IE	[0.15, 1.5]	-	[0.11, 18]
	LIGO II	[20, 200]	-	[16, 270]
F	LIGO I	0.05	0.02	0.8
	LIGO II	[60, 500]	80	2,000
	Virgo I	[0.002, 0.04]	-	-
	LIGO IE	[0.02, 0.4]	-	-
	Virgo+	[0.25, 5]	-	-
H	LIGO I	[0.008, 0.13]	-	-
	LIGO II	[40.2, 310.9]	-	-
	Virgo+	0.003	[0.01, 0.02]	[0.07, 0.08]
	Virgo II	[3.0, 3.6]	[12, 19]	[35, 92]
I	LIGO S5	0.004	0.02	[0.08, 0.09]
	LIGO S6	[0.008, 0.009]	[0.03, 0.04]	[0.17, 0.21]

* indicates older theoretical models.

TABLE 2.4. Experiment runtime

Experiment	Run	Days	Run-Time	NS-NS Range
TABLE 2.2. Theoretical BH-BH detection rates				S1 4.8 S2 0.08 Mpc 0.3 Mpc
ID	Source	Published	BH-	18.0
BH Detection Rates (yr⁻¹)			LIGO	54.6 ^[AD]
I				
A	[U] 1999			S3 13.2 5.0 Mpc
	0.19 ^(b) ,			S4 18.6 8.6 Mpc
	400 ^{(a)(b)}			
B	[V] 2, 500 ^(b)	2007		S5 12.0 Mpc
C	[W] [2 · 10 ⁻⁴ , 0.5]	2010		365 7 3 X D [A] E
	LIGO IE			

S6

365

~50 Mpc

D	[X]	2010	<i>[0.05, 4.9]</i>
E	[Y]	2008	<i>[0.01, 1.7], [0.11, 18]^(a)</i>
F	[Z]	2011	<i>0.8</i>
G	[AA]	2009	<i>-</i>
H	[AB]	2004	<i>-</i>
I	[AC]		

2012
[0.08,
0.17]^(a)

Note: (a) indicates LIGO IE and (b) are older models.

Note: The more

recent runs of GEO600 are included since the detection rates are similar to the earlier LIGO I runs.

Virgo I	VSR1 Mpc	111	111 ^[AF]	12.4
Virgo+	VSR2 VSR3	98 61.3	159.3 ^[AG]	16.8 Mpc ~50 Mpc
GEO600	S4 (S1)	28.8	370.8 ^[AH]	~LIGO I
	S5 (S2)	342+		~LIGO I

The probability of gravitational waves not existing is compared to flipping a loaded coin. When a normal coin is flipped, the probability of it landing tails is 50%. This is equivalent to measuring for gravitational waves over the period required for a single event, and having a 50% chance of detecting one. For a loaded coin, the result will always be tails regardless of N. As the coin is flipped N amount of times, the probability of the coin being loaded increases if the results are always tails. Therefore, the probability that the coin is loaded is equivalent to that of gravitational waves not existing. This probability is defined as (116), where N is the expected events per total period.

There exist several orders of magnitude between individual models. The majority of this variation is due to the merger rate and density of BH-BH events. Excluding the older models, the remaining models are grouped together in table 2.6. With respect to old models, it is clear from the number of events expected that gravitational waves could not exist ($\sigma > 6$). The new models decrease expected rates by three to four orders of magnitude. However, two of the new models also indicate a $\sigma > 6$ for the max event limit, with a third at $\sigma = 3.7$. The current data does not allow for a definitive answer for whether gravitational waves

exist, although it does begin to raise doubts. The expected rates are also highly dependent upon the

$$P(\text{Null}) = 1 - 2^{-N} \quad (116)$$

theoretical model. Assuming that these do not vary

Table 2.5 depicts the probability of gravitational waves not existing for each experiment and model.

TABLE 2.5. Theoretical detection rates by experiment

drastically in the future, the next generation of detectors should be capable of bringing all models to $\sigma > 6$. For example, conservative estimates of

LIGO I	A*	0.07	-	4.76%
	B*	94	-	~100.00%
	B	3.4	-	90.8%
	C	0.12	0.0049%	7.96%
	DI	0.74	-	40.0%
	DII	0.009	-	0.64%
LIGO IE	F	0.13	-	8.63%
	A*	1700	-	~100.00%
	S5	0.114	6.96%	7.60%
Virgo+	I	0.259	13.43%	16.4%
	A†	0.045	2.48%	3.07%
		0.48	-	28.2%

Advanced LIGO (LIGO II) project hundreds of events per year. Advanced LIGO and Advanced Virgo are expected to begin operations in 2014, so direct proof will require three to four years as of 2013.

B n (range of
e t L frequencies.
y e I LISA is not
o r S expected to be
n f A in operation
d e) until after
r 2020;
t o w however, it
h m i will be
e e l capable of
s t l detecting
e e massive BH-
, r p BH events if
S r gravitational
t p o waves exist.
h a v Due to the net
e c i BH-BH mass,
e d the waves
L A e generated
a n would be
s t a several orders
e e of magnitude
r n n larger than
n e other sources.

I a w **TABLE 2.6. Approximate combined probability**

	Models	Net Max Events	Min	Max	Max
	B,I,I,B[†]	27.2	-	~100.00%	> 6
	C,I,C[†]	1.35	27.1%	60.8%	1.7
	D,I,I,D[†]	6.17	28.8%	98.6%	3.7
	E,DII,I,E[†]	41.3	37.0%	~100.00%	> 6
	F,I,I,F[†]	1.43	-	62.9%	1.8

B^{**}	* indicates old model
640	† indicates LIGO I statistics were applied to the latest runs of GEO600.
-	
~100.0	
0%	
B[†]	
23.4	
-	
~100.0	
0%	
GEO600	
(S4/S5)	
C[†]	
DI[†]	
DII[†]	
F[†]	
0.88	
-	
45.8%	

10.5.7.2.3. Single Particle Metric

For a massive non-composite particle, the field should be approximately (E_0/r) in the co-moving frame. Since the metric must be isotropic, the solution in spherical coordinates is (117).

$$\gamma^{-2} \quad 0 \quad 0 \quad 0$$

∇

The Christoffel symbols are determined for an isotropic, spherical vacuum field in table 2.7. To compare the radial acceleration as derived from the geodesic equations to the algebraic results (29), the motion of a particle can be restricted to the radial direction (123); i.e. θ' and φ' are 0.

$$\begin{aligned}
 & \frac{F^g}{I} \quad \frac{t^2 G_0}{M} \quad \frac{t' r''}{r c_0} \\
 & g_{22} = \frac{(1)}{\gamma^2} \quad \left. \right) \\
 & \frac{0}{uv} \quad \frac{G_0}{M} \quad \frac{G}{r} \\
 & 0 \quad \frac{r''}{r^2} = \frac{1}{\gamma^2} \quad \frac{t'}{r} + \frac{1}{M} \quad \frac{r'}{r^2} \\
 & 0 \\
 & \left(\begin{array}{c} \gamma \\ \gamma \\ \gamma \\ \gamma \end{array} \right) \\
 & L \quad 0 \quad 0 \quad 0 \quad (\gamma r)^{-5} \quad (r c_0)^2
 \end{aligned}$$

The Lorentz scalar (118) is defined as usual, where the far-field approximation is applied.

$G_0 M$

For a massive particle moving at escape velocity relative to the metric, r' is replaced by (27); t' is defined as (124), where $k = 1$

$$\gamma_g = 1 + \frac{G_0 M}{r c^2} \cong 1 + \frac{1}{\Delta} \quad \text{and } (\gamma_g)$$

(118)

Δ

{124}

$$t' = k \gamma^2$$

$_0$

g

For comparison, the Lorentz scalar relative to the

The acceleration at escape velocity is (125) or (29).

Schwarzschild

metric is written

as (119).

$$\int \frac{dt'^2}{r^2}$$

$$d^2 r^7 a =$$

$$\frac{1}{\gamma}$$

—

$$= \frac{G_0 M}{\sqrt{1 - 2 \frac{G_0 M}{r^2}}} r^2 c$$

TABLE 2.7.
Christoffel
symbols
between
theories

$$\Gamma^\mu_{\nu\rho} = \frac{\partial x^\mu}{\partial x^\nu} \frac{\partial x^\rho}{\partial x^\sigma} \Gamma^\sigma_{\nu\rho}$$

$$\Gamma^1_1 = \frac{c^2}{r^2} \Gamma^1_1 \frac{\partial \theta}{\partial r}$$

$$\frac{r}{\gamma} \quad \frac{r}{g}$$

θ

θ

γ

γ

(121) Γ^3, Γ^3

1

$$\text{Single Particle } \theta' = -2\Gamma^2 \theta' r - \Gamma^2 \varphi'^2$$

- 13 31

$$\gamma g r$$

12 33

r

Schwarzschild

$$d \quad \frac{\partial r'}{\partial r} \frac{\partial \varphi'}{\partial r} \frac{\partial \theta}{\partial r}$$

1 r -

$$\frac{-2\Gamma^3}{r} ,$$

$\sin^2 \theta$

$$\frac{2}{r} \varphi$$

$r \bar{\Delta}$

$$\frac{3}{r} ,$$

si

01

n

10

2

$$\gamma g c$$

0

g (

13 23

Γ_{33}

γ_g γ_g^2

γ_g^2

The geodesic equations become

relative³² to the metric after applying

the following relations

(122)³³ $\cot \theta$

$$\frac{d}{x'} \frac{d}{d} \frac{dd^2 x'}{2} / \frac{r \gamma^4 \bar{\Delta}}{(g)} \frac{\Gamma^2}{-\sin \theta \cos \theta} - \sin \theta \cos \theta$$

$$\frac{x'}{x} \frac{x'_2 \gamma^3}{dt^2} - \frac{1}{2} \quad \text{Note: Single particle refers to massive and non-composite.}$$

: : : : : :)

$$\frac{d}{r} \frac{d}{t} \frac{d}{r}$$

$$-1 \quad t \quad r$$

$$-7 \quad 2$$

$r \bar{\Delta}$

The geodesic equations are applied to determine the motion of a single particle inside the previous

$$\frac{G_0 M}{r^2} \approx \frac{1}{\gamma^2} \nabla^2$$

potential.
Using space-like convention ($- + +$), the space-time interval is defined as (120).

$$(ds)^2 = \gamma \frac{dx^\mu dx^\nu}{c^2} \quad (120)$$

Expanding the geodesic equations results in the proper acceleration for each component (121), $(x^0)^2$ must

be replaced with

12 21

1

$$\frac{\gamma}{\gamma^2 (rc_0)} \frac{1}{g}$$

$$\frac{1}{(rc_0)^2}$$

$$\frac{1}{g}$$

10.5.7.2.4. Gravitational Force and Potential

From sections (1.3, 2.3), it was observed that the gravitational potential is dependent upon a variation in the effective background vacuum energy density. For static fields, classical energy conservation does not depend upon the path taken between two points. The proper force (126) upon a particle is therefore similar to the relativistic Newtonian perspective. If a particle's geodesic path follows the gradient of ∇ , the problem is always reduced to the covariant

derivative of the scalar field; i.e. the vector $\mathbf{g} \rightarrow$ will point in the direction ∇A .

$$\overline{\Delta} \quad \nabla$$

$$-\frac{\partial}{\partial \Delta} \nabla$$

transformed vacuum fields of an anisotropic and isotropic metric. Since a realistic gravitational field will complicate the field dynamics, a Cartesian coordinate system will instead be applied so that $(r, \theta, \phi) \rightarrow (x, y, z)$. The potential along the x-axis is anisotropic in the sense that it only increases in a single direction (x^a), similar to the Schwarzschild metric (\hat{r}). It is difficult to consider a test particle initially placed at an infinite distance from the field. The problem is simplified by introducing an artificial potential as depicted in figure 2.1.

A massive particle placed where the local field is zero ($y_{\mu\nu}$) has a far-field approximated by (128);

$$\frac{\mathbf{f}}{\mathbf{A}} = \frac{\mathbf{E}}{o} \quad (126)$$

$$\frac{-}{g} \frac{-}{\nabla} \quad \text{this is equivalent to } r \rightarrow \infty \text{ for a spherical metric.}$$

$$\rightarrow$$

$$\frac{1}{U}$$

$$=$$

$$-$$

$$A$$

$$-$$

$$g$$

$$\rightarrow$$

$$1$$

The gravitational potential energy (U) is derived by

$$(128)$$

equating $\gamma_G = \gamma$ as applied in section (1.3); the potential is therefore defined as (127).

$$\Phi = -\frac{n}{c^2} \quad (127)$$

As the test particle moves into the gravitational potential, it gains quantized energy.

This variation in energy is due to an influx of vacuum

energy

density rather than a net increase (129).

Vacuum field theory requires that the metric of

E₁

space-time is isotropic, while

EEFs are anisotropic. The

choice of

isotropy or

anisotropy is

crucial for

motion as observed

by a distant

observer. This is

not locally

detectable since any

observer ($g^{\mu\nu}$)

deforms with

respect to the space-

time metric ($g_{\mu\nu}$).

Under the

assumption that

vacuum energy

density is

conserved, the

vacuum far-field

energy of a non-

composite particle

should at most

remain constant or

decrease when

moving through an

external field.

Conservation of

classical energy

does not always

ensure conservation

of vacuum energy

density. The

problem is

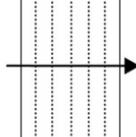
complicated since

actual particles will

$$- (r, \theta) = \frac{r_n \sqrt{(\gamma \cos \theta)^2 + (\sin \theta)^2}}{1})$$

This means that the particle has

transformation is similar to the

		
$\eta_{\mu\nu} (y_g = 1)$	Φ	$\eta_{\mu\nu} (y_g \neq 1)$

gained quantized energy relative to a stationary observer; however, relative to $y_{\mu\nu}$ vacuum energy must be conserved. Since the anisotropic

special relativistic version, it can be directly substituted into (129).

Figure 2.1. Sections of a localized field are illustrated relative to an initial and final state. The fields relative to the preferred reference frame ($\eta_{\mu\nu}$) are described by the space-time metric and Lorentz transformation.

The resulting effective field (130) is once again relative to $y_{\mu\nu}$ or the preferred reference frame.

$$\Psi(r, \theta) = \frac{E_0}{r_n} - \sqrt{(\gamma \gamma_g \cos \theta)^2 + \sin^2 \theta}$$
(130)

For an isotropic metric with identical setup, the field relative to the reference frame ($y_{\mu\nu}$) is defined by (137).—

$$\Psi(r, \theta) = \frac{E_0}{r_n} - \sqrt{(\gamma \gamma_g \cos \theta)^2 + \sin^2 \theta}$$

$$\Psi(r, \gamma)$$

$$(\gamma \cos \theta)^2$$

²

$n \ g \ n$

+ (si)

Since the test particle will be traveling at escape velocity, ($\gamma = \gamma_g$ can be applied). After reorienting

the coordinate system, the Cartesian equivalent is defined as (131) with motion along the \mathbf{z} -axis.

At escape velocity, the same ($\gamma = \gamma_g$) relation can be applied resulting in (138).

$$\Psi^{(x \rightarrow)} = \frac{E_0}{r_n}$$
(138)

$$\Psi^{(x \rightarrow)} = \frac{E_0}{\sqrt{(x)^2 + (y)^2 + (z)^2}}$$

After integration, the shell energy for the isotropic metric is (139).
 $\sqrt{x^2 + y^2 + (yz)^2} (\gamma_x + \gamma_y)$

Equation (131) can further be put in spherical coordinates

(132), allowing a shells net energy to be compared between configurations.

$$-\frac{\pi(A^2 - B^2)}{LN(\gamma + \sigma)} = \frac{\pi(A^2 - B^2)}{LN(\gamma - \sigma)} \quad (139)$$

Where σ is defined as
(140).

$$\frac{E}{\gamma^2} \frac{s}{n^{1/2}} \frac{p}{\rho} = \sigma \sqrt{\gamma^2 - 1} \quad (140)$$

$$\bar{\Psi} = \sigma [(\gamma^2 - 1) + (\gamma \cos(\varphi))^2] \quad (132)$$

The factor between the initial configuration (135) and isotropic case is therefore (141).

Volumetric integration between two radii results in shell energy for the anisotropic case (133),

$$\frac{1}{k} = \frac{1}{[LN(\gamma - \sigma) - LN(\gamma + \sigma)]} \quad (133)$$

$$-\frac{\pi\gamma(A^2 - B^2)}{\bar{\Psi}^{\text{shell}}} = \frac{\pi\gamma(A^2 - B^2)}{2\sigma} [LN(\gamma^2 - \alpha) - LN(\gamma^2 + \alpha)] \quad (133)$$

Considering vacuum energy density conservation,

where α is defined as (134).

$$\alpha = \frac{1}{\sqrt{\gamma^4 - 1}} \quad (134)$$

The original field configuration (129) reduces to

(135) after solving for shell energy.

$$\frac{B}{A} = -\frac{\pi\gamma(A^2 - B^2)}{2\sigma} [LN(\gamma^2 + \alpha) - LN(\gamma^2 - \alpha)] \quad (136)$$

the isotropic and anisotropic cases are compared. As α is rearranged, the peak energy of the field must also increase proportional to the classical quantized energy. There must be an influx of field energy into the central region, which is assumed smooth and finite. For field energy to be conserved, the far-field energy must be equivalent

$$\text{For comparison, } \frac{\bar{\Psi}^{\text{shell}}}{\bar{\Psi}} = 2\pi(B^2 - A^2) \quad (135)$$

variation in field energy for any shell less than the original configuration can be defined with a Plotting factor between the two configurations.

Therefore, the original field is multiplied by a function of the Lorentz scalar, defined as (136) for the anisotropic case.

$$k = \frac{1}{[LN(\gamma^2 + \alpha) - LN(\gamma^2 - \alpha)]} \quad (136)$$

Anisotropic space-time metrics therefore cannot conserve vacuum energy density.

10.5.7.2.5. Arbitrary Space-Time Metric

In order to determine the space-time metric from quantized fields, vacuum field theory must be applied. A coordinate basis is chosen so that $g_{\mu\nu}$ is diagonal with each term proportional to the general Lorentz scalar (g_0). From the gradient of $\nabla(x)$, the fundamental coordinate line is defined at each point in space with the base vector \hat{e}_1 . If the gradient is zero, then the space is locally flat ($y_{\mu\nu}$).

From this notation, the two vectors (r_u, r_v) are not necessarily of unit length. Vacuum field theory is however locally isotropic, i.e. the space-time metric is determined from a single scalar field or vacuum energy density. The previous vectors are therefore directly proportional to the curvilinear basis, i.e. $r_a = (g)e_a$. The first fundamental form is also equivalent to the metric tensor, i.e. each component is the scalar product ($g_{\mu\nu} = g_{\mu} \cdot g_{\nu}$). For example, if $\nabla \cong E_0/r$ then each point would

The variables become $E = r_u \cdot r_u, F = r_u \cdot r_v, G = r_v \cdot r_v$. This can be extended to three dimensions, where the first fundamental form (145)^[AF] is represented in quadratic form.

perpendicular to \hat{e}_1 . For other cases, the remaining

$$dx^T A_{11}$$

$$A_{21}$$

$$A_{31}$$

dx orthogonal base vectors must be determined by the

$$I(dx, dy, dz) = [dy] [A_{12}$$

$$A_{22}$$

$$A_{32}] [dy]$$

(145) following methods.

$$dz$$

$$A_{13}$$

$$A_{23}$$

$$A_{33}$$

dz

Depending on choice of coordinates the field perpendicular to \hat{e}_1 can vary, requiring that at least one other principle direction exists. Each aligns with the planes of maximum and minimum curvature relative to constant ∇ surfaces. The additional bases are therefore eigenvectors of the shape operator (142), which is defined by the first $A\nabla$ and second fundamental forms.

1

$$W = \frac{EG}{I} \left[\frac{LG - MF}{2,33} + \frac{MG - NF}{3,23} + \frac{1,21 - 2,11}{NE - MF} \right] = \frac{(142)}{F^2}$$

The first and second fundamental forms relative to the tangent plane of each surface $R^2 \in R^3$ is (143).

The components ($A_{\mu\nu} = r_\mu \cdot r_\nu$) can be defined in terms of partial derivatives of ∇ (146)^[AF], where $(dx \rightarrow dx^1)$, $(dy \rightarrow dx^2)$, $(dz \rightarrow dx^3)$ and partial derivatives are written as $\nabla^i \nabla_{jk} = i,jk$.

: $A_{11} = 2,31 - 3,21$
: $A_{22} = 3,12 - 1,32$
: $A_{33} = 1,23 - 2,13$
: $A_{21} = A_{12} = (\frac{1}{2})(3,11 - 1,31 + 2,32 - 3,22)$
(146)

: $I = \begin{bmatrix} E & F \\ F & J \end{bmatrix}$: $II = \begin{bmatrix} L & M \\ M & J \end{bmatrix}$ (143)
other principle directions to be determined. From
 $F \quad G$
 $M \quad N$
If two unique principle directions exist beyond the initial gradient, vectors tangent to coordinate lines are (r_u, r_v) . Since the first base vector's (\hat{e}_1) direction is determined by the normalized gradient of ∇ , n is the normal to the surface (144)^[AF].
the two dimensional II (143), the variables can be defined as $L = r_{uu} \cdot n$, $M = r_{uv} \cdot n$, $N = r_{vv} \cdot n$, or as a tensor $B_{\mu\nu} = r_{\mu\nu} \cdot n$. Using the following method, the second fundamental form can be written in terms of partial derivatives of the vacuum energy density scalar field (147).
 $n = \frac{A}{\|r\|} \times \frac{r_u}{r} \times \frac{r_v}{r} \times \frac{r_u}{r} \times \frac{r_v}{r}$
(144) $II(dx, dy, dz) = \nabla^x dx + \nabla^y dy + \nabla^z dz$ (147)

$$: A_{32} = A_{23} = (\frac{1}{2}) (1,22 - 2,12 + 3,13 - 1,33)$$

The second fundamental form is now introduced, which when combined with the first allows the

For an implicit surface where $\nabla^2(x, y, z) = 0$, both fundamental forms are equal to zero. If the partial derivative of the field with respect to a given component is non-zero, that component can be solved for within $\text{II}(dx, dy, dz)$ and substituted into $\text{I}(dx, dy, dz)$, arriving at the third fundamental form

coordinate lines of the metric tensor (g_{uv}). To transform from the curvilinear basis \hat{e}_v to $g \rightarrow$ requires a tensor so that (152) is true. Since each component is already aligned, F is diagonally symmetric and varies only in magnitude.

(III). The idea is to factor the new expression so $g \rightarrow = F_v \hat{e}_v$
(152)

that it is quadratic with respect to the two remaining components. If $\nabla^2 G = 0$, then (148)^[AF] becomes a function of (dx, dy) .

$$\text{III}(dx, dy) = U(dx)^2 + Vdxdy + W(dy)^2 = 0 \quad (148)$$

The discriminant of this equation is defined as $\Delta = V^2 - 4UW$, which can be used to determine the remaining principle directions (149)^[AF].

$$(-V \pm \sqrt{\Delta}) \frac{\partial}{\partial z}$$

Due to the parameterization of effective vacuum fields with metric distance, the transformation must be isotropic. In comparison to the relativistic case where the field is compressed in a single direction, the presence of background vacuum energy warps a particles manifold equivalently in all directions at each point. Without this feature, the perceived space-time metric cannot be attributed to a relative medium upon space, further induced by vacuum energy density or a scalar field. From the isotropic nature of general field

$$T = [\quad \quad \quad] \quad (149)$$

$$2U_z \quad \quad \quad$$

$$\nabla \quad \quad \quad$$

$$\text{nature of general field} \quad \quad \quad$$

$$\text{transformations, the previous cases} \quad \quad \quad$$

$$(V \pm \sqrt{\Delta}) \frac{\partial}{\partial x} - 2U \frac{\partial}{\partial y} \quad \quad \quad$$

$$\text{cases} \quad \quad \quad$$

$$\text{where } \nabla \quad \quad \quad$$

$$_z = 0 \text{ by} \quad \quad \quad$$

These vectors can be solved for in other

$$y \quad \quad \quad$$

$$z = 0 \text{ by} \quad \quad \quad$$

lic permutation of the components^[AF]. The components of $A_{\mu\nu}$ remain

constant, although U, V and W must each be recalculated and the components of T rearranged. For example, if ∇^x G = 0 were true instead, the correct principle directions are defined as (150), i.e. $(x \rightarrow y), (y \rightarrow z), (z \rightarrow x)$.

$$(V \pm \sqrt{\Delta}) \nabla^y - 2U \nabla^z$$

$$L \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - T = \begin{bmatrix} \gamma_g \\ 0 \\ 2U \nabla^x \end{bmatrix} \quad (150)$$

The complete curvilinear basis is therefore defined as (151).

tensor F can be written as (153). (1)
53
)

The tensor (153) is extended to a Minkowski space via (154).

$$\gamma^{-1} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \gamma_g & 0 & 0 \\ 0 & 0 & \gamma_g & 0 \\ 0 & 0 & 0 & \gamma_g \end{bmatrix} - \quad (154)$$

In general, the initial coordinate system used for the gradient and partial derivative F = [0 γ_g 0 0] = $\gamma_g g_{\mu\nu}$ e effective resulting curvilinear coordinates will also consist of an orthogonal basis, which follows the

To determine the proper space-time for any energy density must first be calculated; see section (2.6). The curvilinear basis is then determined for time dependence.

Therefore, an arbitrary space-time metric (155) can be defined relative to the curvilinear basis.

$$:\hat{e} = \begin{bmatrix} 0 & \gamma_g & 0 \\ \nabla & \nabla \end{bmatrix} : \hat{e} = \begin{bmatrix} 0 & T_1 \\ T_1 & g_{\mu\nu} \end{bmatrix} : \hat{e} = \begin{bmatrix} T_2 \\ g_{\mu\nu} \end{bmatrix} \quad (151) \quad (155)$$

$${}^1 \quad ;\overline{\nabla \nabla}; \quad {}^2 \quad ;T_1; \quad {}^3 \quad ;T_2; \quad L \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \gamma^2 \end{bmatrix} \quad (155)$$

10.5.7.2.6. Numerical Methods

Since solutions are based upon each individual particles influence on local fields, realistic objects must be numerically determined. Each particle further consists of a localized field relative to a

If both particles have equivalent rest mass ($E_0 = E_n$), the effective field of each is (161); where d is the distance between particles.

$$\bar{\nabla} (r) = \frac{2E_0}{d^n \sqrt[4]{1 + \frac{E_0}{r^4}}}$$

preferred reference frame ($y_{\mu\nu}$).

Therefore, the field of a single

particle does not follow the metric

induced by its own vacuum field. It instead follows the background vacuum field depicted by all other particles and free fields. This is applied to determine the effective space-time metric of any object using quantized variables. For first-order methods, the field of each particle will undergo isotropic deformations.

The magnitude of these transformations will be proportional to the general Lorentz scalar at the classical position.

Therefore, each particle has a unique general Lorentz scalar (156) due to the effective field of every other particle; note that ($r_n = 0$) is equivalent to (d_n).

Assuming the gradient of the local vacuum field is small, the first order

approximation is useful for determining the field of large objects. However, systems with more than two particles are more difficult to deal with. The objective is to develop an iterative numerical method that is equivalent to the algebraic results.

The constant isotropic deformation is retained for dust solutions.

Assuming each particle is stationary relative to the effective metric ($g_{\mu\nu}$), the algorithm for two particles is the following.

1. Start with the

$$\begin{aligned} & \text{constant,} \\ & \text{non-} \\ & \left(\frac{1}{m} \sum_{n=1}^N \frac{\nabla}{\Delta m} (\mathbf{r}_n \neq 0) \right) = 0 \end{aligned}$$

of both particles ; i.e.
 E_1/r_1 and
 E_2/r_2 .

(156) 2. Determine $(g_{,1})$ and $(g_{,2})$ from (156), relative to the non-effective fields defined in

1.

3. Iterate

by
updati

b ng
effecti

ve
fields

define
d as

$E_1/(r_1)$

$(g_{,1})$
and

$E_2/(r_2)$

4. Determine $(g_{,1})$ and $(g_{,2})$ relative to the

$$\begin{aligned} : \bar{\nabla} & : \bar{E} \\ \left(\frac{1}{r_1} \right) & \left(\frac{1}{r_2} \right) \\) & = \end{aligned} \quad (157)$$

$$= \frac{1}{r_1 \gamma_{g,2}} \bar{\nabla} \frac{1 + \frac{E_1}{r_1}}{1 + \frac{E_2}{r_2}}$$

$$r_2 \gamma_{g,1}$$

In Planck units, the Lorentz scalars (156) can be written as continuous fractions (158).

$$(=$$

effective field

5. Loop back to 3.

This can also be calculated by hand, where the first few

$$g_{,1} = \frac{d_1 - \frac{E_2 - 1}{1 + d_2}}{r_1 - \frac{E_1}{1 + d_1}} \quad : \bar{\nabla}^1 = \frac{r_1}{r_2} \Delta$$

Setting $a = E_1/d_1$ and $b = E_2/d_2$ results in (159).

$$\oint \frac{E_1}{r_1} \frac{1}{1 + \frac{E_2}{r_2}} \Delta \quad (162)$$

$$\begin{aligned} & x & I2 : \bar{\nabla}^1 = r_1 \frac{1}{1 + \frac{E_2}{r_2}} \Delta \\ & \frac{1}{a} & \frac{1}{x} \\ & 1 & \Delta \\ & x & \Delta \\ & & 1 \\ & & + \\ & & \frac{E_1}{d_1} \end{aligned} \quad (159)$$

The effective Lorentz scalar is therefore (160).

$$g_{,1} = \frac{(\sqrt{a^2 + 2a} - 2ab + (1+b)^2 + a - b + 1)}{2(1+b)^2}$$

The previous iterative method is equivalent to solving the continuous fractions of equation (156) and can be extended to N particles.

$$(160)$$

Expanding equation (156) by hand matches the iterative method for three particles, although the equations become relatively large. The effective Lorentz scalars will always be linear combinations

```

// STEP 1 //
DOUBLE RE[N];           // Rest Energy
DOUBLE EF[N];           // Effective Energy
DOUBLE P[N][3];         // Position
DOUBLE G_C = Δ          // Reference Energy
DOUBLE lorentz_scalar[N];

// STEP 2 //
FOR(INT A = 0, A < N, A++) {
    y_temp = 0;
    FOR(INT B = 0, B < N, B++) {
        { IF(A != B) {
            dx = P[A][0] - P[B][0];
            dy = P[A][1] - P[B][1];
            dz = P[A][2] - P[B][2];
            D = sqrt(dx*dx + dy*dy + dz*dz);
            y_temp += RE[B]/D; }
        }
    lorentz_scalar[A] = 1+y_temp/G_C;
    // STEP 3 //
    EF[A] = RE[A]/lorentz_scalar[A];
}

// STEPS 4-5 //
WHILE(KEEP_ITERATING == TRUE)
{ FOR(INT A = 0, A < N, A++) {
    y_temp = 0;

    FOR(INT B = 0, B < N, B++) {
        { IF(A != B) {
            dx = P[A][0] - P[B][0];
            dy = P[A][1] - P[B][1];
            dz = P[A][2] - P[B][2];
            D = sqrt(dx*dx + dy*dy + dz*dz);
            y_temp += EF[B]/D; }
        }
    EF[A] = RE[A]/(1+y_temp/G_C);
}

```

The exact vacuum field can be determined by applying the proper, active transformation (163) to each particle.

$$r_n \text{ of nested, continuous fractions.}$$

This structure is $\frac{1}{\Delta} \mathfrak{f} \left(\sum_{m=1}^N \Psi^{mn}(r_n) \right)$

preserved by applying the iterative method outlined

$$r'_n = r_n + \frac{dr_n}{\Psi^{mn}(r_n)}$$

in figure 2.2 for any amount of particles.
o

W
h
e
n
t
h
e
f
i
r
s
t
-
o
r
d
e
r
a
p
p
r
o
x
i
m
a
t
i
o

n was
defined,
(
held
constan
t; i.e.
the
transfor
med
radius
(164
all
directio
ns
became
isotropi
c.

r_n
(16
4)

θ
field
be obtained

T
h
e

far-field
approxi
mation
is
instead -
applied,
which
is in
agreem

roximat Now that the
 ion fields are finite
 howeve at all points, it
 r is possible to
 naturall integrate along
 y has the effective
 singular field of each.
 ities, so For an exact
 each solution,
 particle memory
 's field requirements
 must drastically
 have a increase since
 cut-off the effective
 (165 field of each
 when particle must
 m_n be known at
 surpass each point in
 es the space. This is
 maximum because each
 m particle's field
 quantiz is relative to
 ed the background
 field of all
 value
 classica

other particles and free fields. This can be optimized by assuming many particles exist, so that the influence one has on the others is negligible. Therefore, only one effective field is defined based upon the contribution from all particles; i.e. a continuum approximation is made similar to EFEs. Identical numerical methods can be applied with respect to figure 2.2, where the initial configuration converges towards the effective field.

classica
 l
 energy.
 -
 :
 -
 mn
 ;
 -
 mn
 (165)
 -
 mn
 {
 : >
 ax ∇
 m ∇
 n
 -
 -
 -
 -
 ∇
 ∇
 ∇

Figure 2.2. C-code for calculating the effective vacuum field of composite objects with first order methods.

2.7. Relativistic Pressure and Bulk Flow

Temperature is a scalar quantity that depicts a system's internal kinematic energy. For a system at equilibrium, the energy distribution of individual particles is related to temperature via the Maxwell-Juttner equation (166)^[AG].

$$\frac{e^{-\gamma\theta}}{\text{_____}}$$

The number density depicts the amount of particles per unit volume. Normalizing this so there is only one particle per finite volume (V) allows the metric to be determined. For objects at equilibrium, each particle will be confined to its own respective volume. The one-dimensional force on the plane of another particle's volume is defined by (172).

$$f(\gamma) = \gamma \sqrt{\gamma^2 - 1} \quad (166)$$

$$\frac{m_o c^2 \{ \gamma^2 - 1 \}}{1} E_o$$

$$= \frac{\theta K_z}{1/\theta}$$

$$f = \frac{(\gamma - 1)}{(\gamma + 1)}$$

$$(172)$$

$K_2(z)$ is a modified
Bessel function of
the second

kind (167) and $\theta = k_B T / E_o$.

Proper pressure assumes that the particle has an

equal probability of hitting the other two walls.

$$\int_{-\infty}^{\infty} \frac{K_2(z)}{dx} dz$$

The average Lorentz scalar for the Maxwell-Juttner equation (166) is calculated via (168).

$$\frac{P}{E} = \frac{m}{\gamma v} \frac{E}{V}$$

$$f_\infty = \frac{-3V}{E}$$

$$avg = \frac{1}{d} \int d\Omega$$

$$f_\infty$$

(167) the proper pressure is (173).

e.

Temperature is related to pressure by an averaged Lorentz scalar (267), determined

f e beta factor ($\beta = v/c_o$) and proper energy density (ϵ). m Pressure itself is only dependent upon average h

For the classical limit, (168) can be approximated by Taylor expansion resulting in (169).

particle energy and the volume attributed to each.

With the averaged Lorentz scalar, the effective field due to pressure can be determined. A massive

(169) particle moving with respect to a field will have a vacuum field defined by (174).

$$\gamma_{avg} = \frac{k_B T}{E_0}$$

This reduces to the classical relation between average kinematic energy and temperature (170).

$$\frac{E_0 \gamma}{1} = \frac{1}{(\sqrt{(\gamma^2 - r \cos \theta)^2 + r^2 \sin^2 \theta})}$$

Since each particle's velocity has an arbitrary

$$E_k \cong \frac{1}{2} k_B T \quad (170)$$

direction, the field must be averaged over $d\Omega$. As

By applying the Maxwell-Juttner equation, the average kinematic energy of particles does not

remain proportional to temperature. Therefore, the kinematic energy must be related to the proper force or pressure instead. When dealing with pressure at the atomic scale, a particle that collides with a perpendicular wall will experience a change in momentum via (171).

with section (2.4), the coordinates are reoriented so that the correct integral is (175).

$$\begin{aligned} E &= \frac{\gamma}{\sin(\varphi)} \\ -\varrho &= \frac{1}{d\theta d\varphi} \\ a_{v^4} &= \frac{7}{5} \\ g^r &= \frac{5}{\sqrt{(\gamma \cos \varphi)^2 + \sin^2 \varphi}} \end{aligned}$$

Integrating over $\theta: [0, 2\pi]$ and $\varphi: [0, \pi]$ provides the averaged field for each particle (176) with respect to the statistical distribution of velocities

(168).

$$\Delta p = 2m c \sqrt{\gamma^2 - 1}$$

(171)

$$\frac{E_0 v}{T} \ln(\gamma^2 + \sqrt{\gamma^4 - 1}) - \ln(\gamma^2 - \sqrt{\gamma^4 - 1})$$

$$\frac{x}{I} = \left(\frac{\theta}{\sqrt{\gamma}} \right) \quad (176)$$

$$\frac{v}{I} = 2$$

Reducing an object to its individual particles allows the effective field to be approximated with the iterative methods discussed in section (2.6). Determining the effective field of a gas under bulk

Since the choice of u_o is arbitrary, it will be set parallel to the bulk flow vector field ($\vec{v} \rightarrow$). Therefore, the proper velocity of each particle under the (α, β) parameterization is (180).

flow requires a vector field ($\vec{v} \rightarrow$) relative to the space-

flow requires a vector field (\vec{v}) relative to the space-

time metric. Considering a system of particles under

$$\frac{\bar{w} \rightarrow}{y(u)} = \frac{u}{[-S \ C_\beta] + \dots} \quad (180)$$

ball-flow-th-

bulk flow, the effective field of each particle must

be relative to the From the

effective field of all others. Bulk flow is therefore the transportation of kinematic energy density along the space-time metric.

The

averaged distribution
of velocities from
(166) must

averaged distribution
of velocities from
(166) must

be properly added to the bulk flow ($\bar{v} \rightarrow$). This is $\gamma(w)$

$$\gamma(w)$$

equivalent to applying a Lorentz boost in arbitrary

However, the coordinates within (181) are with directions of \vec{u} . Under the assumption that bulk

respect to the direction $w \rightarrow$. These must be mapped

flow moves freely in the forward direction, the problem can always be reduced to an addition of velocities

Proper velocity addition relative to the

angle between a particle's velocity (\vec{u}) and bulk flow (\vec{v}) is (177)^[B].

back to the reference frame, where the bulk flow $\bar{v} \rightarrow$ is defined. This can be accomplished.

d by finding the parameters of the mapping from the direction of $\vec{w} \rightarrow$ to the z-axis via (182).

$$w_x = \frac{C_2}{S} + \frac{C_1}{S - C_1}$$

b
a b
b b
a b

$$\begin{aligned} \text{v} & \quad \frac{1}{[w_y]} = [\bar{S}_b C_b (\bar{C}_a - \\ 1) & \quad S_b^2 + C_a C_b^2 \\ & \quad - \bar{S}_a C_b] [\begin{matrix} 0 \end{matrix}] \\ (182) & \end{aligned}$$

The first step is determining the final velocity

relative to the metric after applying (177) for all x

$$\begin{aligned} & C^2 + CS^2 \\ & SC(C-1) \\ & -SS \\ & X \end{aligned}$$

directions of $\bar{u} \rightarrow$. This can be accomplished with an

$$\begin{aligned}[y] &= [S_b C_b (C_a - 1) \\ &\quad S_b^2 + C_a C_b^2 \\ &\quad - S_a C_b] [y] \\ (183) \end{aligned}$$

active rotation derived from the Rodrigues' rotation

$$\begin{array}{l} z' \\ S_a S_b \\ S_a C_b \end{array}$$

C_a

z

formula (178)^[AH]. \vec{K}^{\rightarrow} is the axis of rotation and

Inserting the solutions from (183) into the primed
 $M = KK^T$; the trigonometric functions are written as $\sin(x) \rightarrow S_x$ and $\cos(x) \rightarrow C_x$ to shorten notation. The coordinates of (181) provides $\nabla(r, \theta, \varphi, \alpha, \beta)$. The last step involves integrating over r .

$$R = S \begin{bmatrix} 0 & -K_3 & K \\ K & 0 & - \\ K_1^2 & M + I & M \end{bmatrix} \quad (178)$$

By restricting the axis of rotation to the x-y plane

$$-\Psi^{avg}(\vec{r}, \theta, \varphi) = \frac{\int_{\alpha, \beta}^f d\alpha d\beta}{4\pi (184)_o} (r, \theta, \varphi)$$

and applying a 2-dimensional rotation to $\vec{K} \rightarrow, (177)$

can be mapped to the unit sphere (179).

$$C_{\beta^2} + C \; S_{\beta^2}$$

$$S_\beta C_\beta(C - 1) - S S$$

The problem can also be viewed as a super-position of an infinite number of configurations relative to

$$R = [S_\beta C_\beta (C - 1) \\ S_\beta^2 + C(C - 2) \\ -S C_\beta] \\ S S_\beta$$

relative to each ($\vec{u} \rightarrow$). With
ical point picking, a finite
{ amount of these can
1 be rotated in a 3-
7 dimensional space
9) and averaged.

3. The Universe

The big bang theory is currently the most widely accepted cosmological model, with the 2011 Nobel Prize awarded for the discovery of accelerated expansion^[BM]. The model however contains several anomalies, unexplained observations and various non-classical assumptions. These aspects can be resolved by abandoning an expanding model in favor of one that is simultaneously expanding and contracting, i.e. a steady state. Current observations are already sufficient for ruling out an expanding universe. Difficulty of arriving at such conclusion arises from the recent acceptance of non-classical assumptions and lack of theoretical constraints. Dark energy for example is not predicted by the standard model and cannot be directly detected. It is widely assumed dark energy exists solely because it allows an expanding model to fit redshift versus distance modulus. The inferred expansion however is an illusion from the local deflection of geodesics, which produces a nearly spherical projection.

With insight from recent observations, aspects that conclusively rule out an expanding universe can be focused upon. Two characteristics that stand out are incorrect predictions of large-scale curvature and the perspective of time versus redshift; these are discussed throughout sections (3.3, 3.5). It is proven that the observed abundance of faint blue galaxies is due to Λ CDM's incorrect predictions for the curvature of the universe. Additional constraints allow all explanations for the 2 — 3x abundance of faint blue galaxies to be ruled out. These range from evolution of the local luminosity function to drastic mergers. Λ CDM further underestimates the size of the faint blue galaxies by 2 — 5x relative to their angular size versus luminosity. Number densities of weak MgII absorbing galaxies in section (3.6) are also in agreement with the prior conclusion. These incorrect predictions by Λ CDM result in systematic lensing errors as discussed in section (3.4).

Since all explanations can be ruled out relative to Λ CDM, the faint blue galaxy abundance is proof rather than evidence. Although proof exists against Λ CDM, there also exists strong evidence against an expanding universe. The purposed theory predicts for distant galaxies to be older than local ones. An expanding model predicts the opposite, which is contrary to observations. For example, galaxies are observed to cool with increasing redshift. Distant quasars contain relatively higher FeII:MgII ratios, depicting increased metallicity with redshift. There are many other firmly grounded observations not compatible with an expanding universe such as the in-fall velocity of the Bullet cluster. For clarity, the first half of chapter 3 will focus on the foundations of the continuous model. The remaining sections discuss the various proofs against Λ CDM, including a statistical comparison between models.

The new cosmological theory only requires the standard model and corrections to general relativity herein. From this short introduction alone, the new model is superior with respect to Occam's razor. In other words, the simplest theory that agrees with all observations is the correct theory. Similar to initial motivation behind an expanding model, the shape of the universe can be fit with a single constant. The inferred accelerated expansion is nothing more than local geodesics deflecting towards the center of an asymptotically flat, linear universe. Dark energy is therefore not required to explain redshift versus distance modulus. The trend is better fit by distant galaxies falling into an asymptotically flat universe, depicted by relativistic redshift and gravitational acceleration. The cosmic background radiation must therefore originate from the central region of an asymptotically flat universe. Classical assumptions insist that this black body radiation is emitted from a central core, which is not compatible with theories that predict event horizon such as EFEs.

3.1. The Big Bang Theory

Georges Lemaître, a Belgian priest, was the first to propose the big bang theory, originally named “hypothesis of the primeval atom”^[AK]. Although Lemaître was the first to discover the “Hubble constant”, it was named after Edwin Hubble. Hubble’s observations in 1929 also showed a linear relationship between the distance and redshift of local galaxies^[AL]. The Hubble constant and linear trend provided the initial motivation behind an expanding model, where redshift is attributed to the recession velocity of local galaxies. According to the Hubble model, the relation between redshift and metric distance is (185).

$$\frac{c}{H_0} \frac{(z+1)^2 - 1}{z^2} = v$$

The initial big bang model is only valid under Hubble’s and Lemaître’s limited observations, i.e. for local galaxies with redshift below $z \approx 0.15$. It is clear that beyond this point, an expanding universe depicted by the big bang theory would need to be accelerating. Assuming the universe began as a point of infinite energy density that consequently erupted into an expanding sphere of energy, there are two plausible scenarios for recession velocity. For a homogenous universe, any initial acceleration from pressure or bulk flow should be constrained to relatively high redshift. Therefore, the first scenario requires that the mass of the universe is large enough to collapse back onto itself. The second assumes the kinematic energy imparted to matter

$$\frac{d}{H_0} = \frac{(z+1)^2 - 1}{z^2} \quad (185)$$

from a big bang event is
large enough to continue
expansion at a constant or
decreasing rate. Neither

Due to the inverse square law, metric distance is related to distance modulus (μ) by equation (186).

$$d^7 = \frac{10^{51}}{\mu^2} \quad (186)$$

Combining equations (185, 186) and subtracting the result from observations, the disagreement becomes apparent. Figure 3.1 provides the error relative to the initial big bang model.

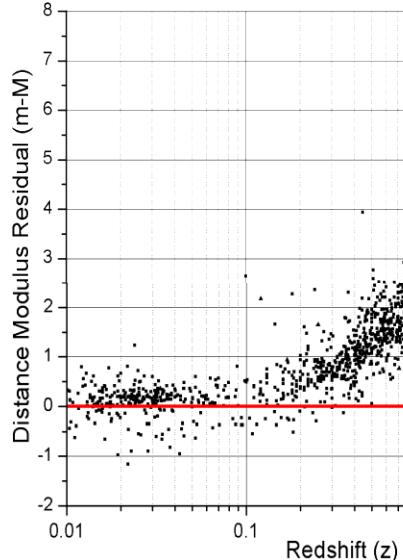


Figure 3.1. Data is from the NED database^[AJ] and linear trend from equation (185) with $H_0 = 73.8 \text{ km s}^{-1} \text{ Mpc}^{-1}$ ^[AI].

possibility fits the illusion of observations, requiring acceleration since the ad-hoc introduction of dark energy.

From the previous chapters, several constraints have been placed upon vacuum field theory. These include the conservation of vacuum energy density, its connection to the space-time metric and the localized nature of particles. With these additional aspects, it is clear dark energy has characteristics similar to vacuum energy density. For example, dark energy in Λ CDM does not force matter to become repulsive; it instead acts to expand the space-time metric. This only provides

the illusion of acceleration since all observers view the universe relative to the space-time metric. With respect to only experimentally confirmed contributions to an object's redshift, there exist two explanations for current observations.

Either the universe consists almost entirely of undetectable energy and matter causing accelerated expansion, or this improperly inferred expansion is an illusion due to the universe being asymptotically flat. It is therefore important to distinguish between the angular scales and time-dependence of each model.

3.2. Redshift and Distance Modulus

The redshift of distant objects can be described in terms of relativistic redshift due to a variation in gravitational potential. Vacuum field theory is not required for determining redshift versus distance; however, it is necessary for large-scale curvature. The relation ($=$) is applied to determine the average relative velocity (187) that is induced from a change in vacuum energy density between source and local observer.

$$\frac{(\nabla)}{\Delta} = \frac{\overline{\Delta}}{\Delta}$$

$$= \frac{\overline{\Delta}}{\Delta}$$

$$= \frac{\overline{\nabla(\nabla + 2\Delta)}}{\Delta}$$

From numerical methods, it is observed that an asymptotically flat universe will generate a field that appears nearly linear with respect to metric distance or local observers. When this is integrated to provide a plot of metric distance versus vacuum energy density, the field becomes approximately linear. Applying equation (191) to the spectral redshift of distant SNIa/GRB demonstrates a linear trend as depicted in figure 3.4. Prior to discussing actual data from SNIa and GRB observations, an ensemble of asymptotically flat universes (192) is

$$v = \frac{-c_0}{r} \sqrt{\nabla(\nabla + 2\Delta)} \quad (192)$$

(187) provided in figure 3.2.

$$= \frac{+1}{\nabla}$$

The Doppler effect requires for any relative velocity

to result in a redshift (188).

$$z_1 = \frac{1 + \beta}{1 - \beta} \quad (188)$$

Plugging equation (187) into (188) results in the relativistic redshift (189) from a change in ∇ .

$$z = \frac{1}{\sqrt{1 + \frac{\nabla}{\Delta}}} \quad (189)$$

T
h
e
r
e

i n
s s
a i
l n
s v
o a
c c
a u
g u
e m
n e
e n
r e
a r
l g
r y
r d
e e
d n
s s
h i
i t
f y
t ,
d i
u .
e e
t .
o t
v h
a e
r e
i s
a e
t c
i o
o n

d b
c o
o t
m h
p c
o o
n m
e p
n o
t n
i e
s n
d t
e s
f r
i e
n s
e u
d l
b t
y s

(i
1 n
9 e
0 t
) h
. e

$$z_2 = \frac{\nabla}{\Delta}$$

(190) f
f

S e
u c
m t
m i
i v
n e
g

r
e
d
s
h
i
f
t

(
1
9
1
)

f
o
r

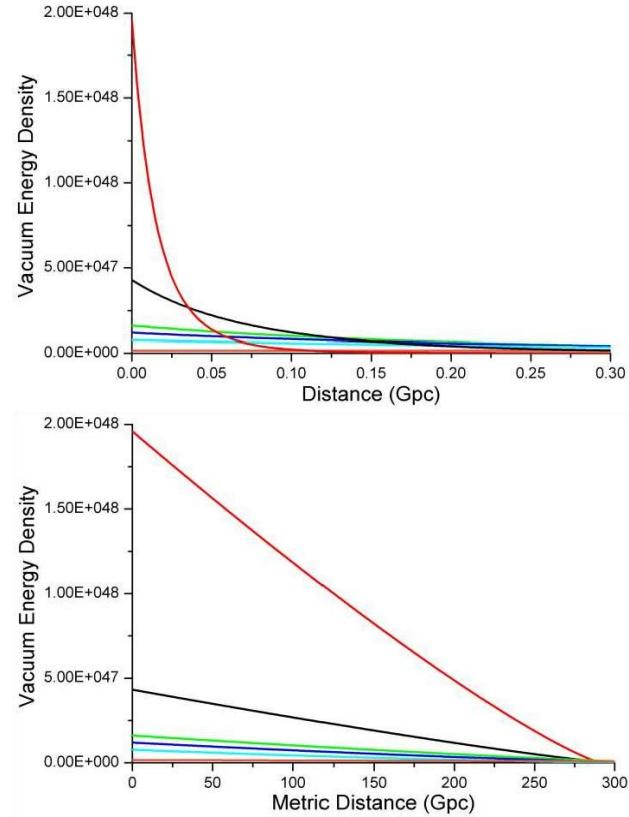
d
i
s
t
a
n
t

g
a
l
a
x
i
e
s

a
n
d

c
1

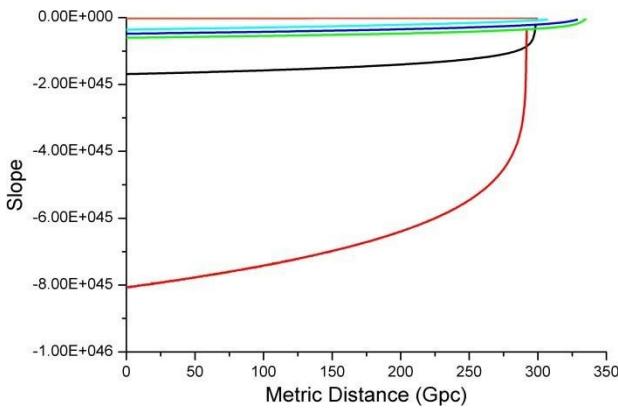
$$\overline{\Delta}$$



strong evidence for an asymptotically flat universe, but also justifies the linearity of vacuum field theory.

at approximately 300 Gpc. (Top) Ensemble of asymptotically flat universes relative to the preferred reference frame and (Bottom) relative to local observers or metric distance.

The derivative of figure 3.2 with respect to metric distance provides a more accurate representation of the linear variations in figure 3.3. This linear trend is crucial for explaining dark energy or the illusion of accelerated expansion. From the NED database, metric distance to each object is determined from distance modulus. The previous net redshift relation (191) is then applied to the data, resulting in figure 3.4. The best fitting trend is with respect to data beyond 0.15z, which results in a constant slope of



From the linear slope relative to the space-time metric, an upper limit on metric or luminosity distance to the central core can be determined. The slope may slightly vary as the core is approached; however, this is not noticeable within currently observable distances (< 100 Gpc). From analyzing the redshift of the core's black body spectrum, the vacuum energy density at the surface is determined with respect to the local space. In other words, the observed spectrum is shifted until it matches the

$S_0 = 3.248 \pm 0.047 \cdot 10^{42} (\text{kg m s}^{-2})/\text{Gpc}$ and y -spectral distribution at emission. The core is found to have a redshift of $z \approx 1089$ with a surface temperature around 3000K^[AN]. Since this form of redshift is solely due to variations in vacuum energy density, the change in ∇ can be determined from equation (193).

$$\nabla = \Delta^z$$

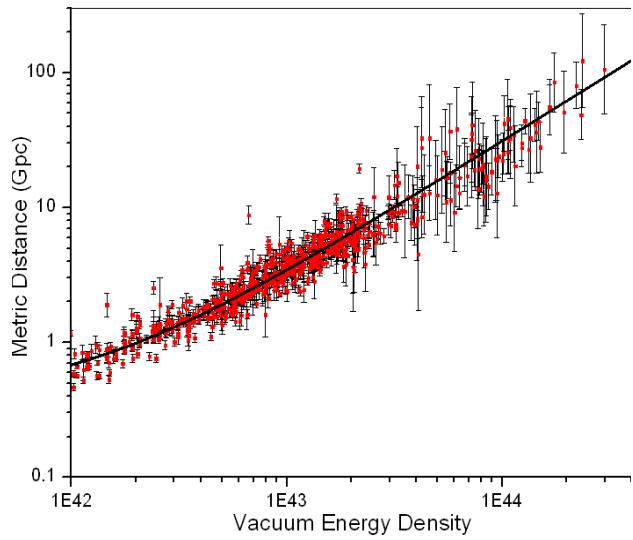
(193)

Therefore, the variation in vacuum energy density from Earth to the core's surface is approximately $1.315 \cdot 10^{47} (\text{kg} \cdot \text{m} \cdot \text{s}^{-2})$. With the best fitting linear slope (all data beyond $z > 0.15$), the maximum metric distance to the central core is determined from

Figure 3.3. The slope of each function has origin at the center of the universe. Note that the deviations are minimal when the observer is inside the localized universe, becoming less linear

equation

$$S_0 d^\gamma = \nabla = \Delta z \quad (194)$$



different form. However, they all produce nearly constant slopes relative to a local observer or metric distance.

Figure 3.4.

Logarithmic plot of metric distance versus vacuum energy density with respect to Earth.

$$\nabla^{(r)} = (S_0 d^\gamma + \nabla^o + \Delta) e^{-\frac{r}{R_0}} \Delta \quad (195)$$

Relative to an asymptotically flat universe and observed cosmic background radiation (CMBR), it is clear that something large has or had existed prior to the present. If distant galaxies and clusters are falling into an asymptotically flat universe, then there must be a mechanism that transports matter from the central region outwards. A situation similar to that of an explosion is not completely out of the question. However, a mechanism already exists that can replace spherical expansion. Relativistic jets emanating from black holes are not well understood, but have been observed from numerous sources with varying intensity and duration^[AM]. The first step in producing the model that agrees with all observations is assuming the local space emerged from such jet. In this perspective, distant galaxies and clusters fall back into the equatorial regions of the central core at relativistic speeds. Due to the conservation of momentum, in falling matter is ejected at the polar regions in the form of dense quark matter. This quark matter further decays into hot, x-ray emitting gas commonly seen in young clusters and galaxies. The local jet is later discussed in section (3.8) with respect to the dark flow and cleaned CMBR image.

Black holes that emit polar jets are known to exist after such events. When considering finite black holes, the best approximations available are QCD and vacuum field theory. If the CMBR is to be taken as black body radiation from a massive but cooled object, then the surface must be finite. One could argue that Hawking radiation is already theorized to be emitted from the surface of non-finite black holes. However, temperature in this perspective is inversely proportional to mass. For an Einstein black hole to emit a 3000K blackbody temperature, it would need to be several orders of magnitude less massive than the Moon. This is clearly impossible with respect to a central core, as countless black holes exist locally that are much more massive. Although the various proofs against

an expanding universe have not yet been discussed, they would clearly nullify current interpretations of the CMBR. In other words, the CMBR cannot be due to a period of recombination. It is instead classical black body radiation emitted from the central core. In an asymptotically flat universe, geodesics will begin to deflect from the local space as distance increases. After sufficient distances, the majority of local geodesics will turn towards the center of the universe. Although the projection is not perfectly spherical, it creates the illusion of accelerated expansion. The CMBR is therefore projected onto all local directions of space, as it originates from the center of the universe.

Putting all of the pieces together, a self-consistent model of the universe emerges in figure 3.5. For any steady state model to be valid, the universe must conserve energy and act as a perpetual machine. Other models similar to the cyclic big bang inherently describe the universe as such. However, they suffer from incorrect large-scale curvature similar to Λ CDM. This is later discussed relative to the size and number densities of distant galaxies. Galactic merger times and properties of distant clusters also insist that the universe is in a steady state. In other words, there is a constant flow of matter from the central core to the outer regions, which further flows back to the central region. This requires for distant galaxies and clusters to be older than local populations when observed from Earth.

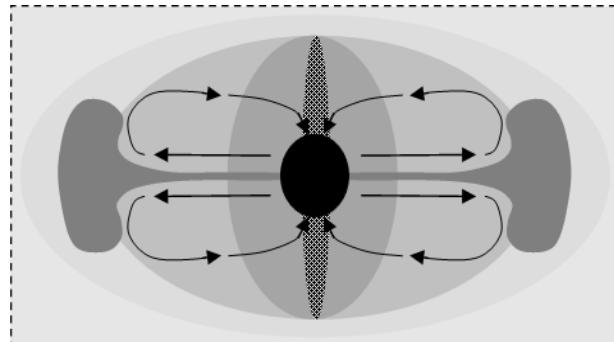


Figure 3.5. A cross section of an asymptotically flat universe in a steady state. The structure takes the form of a Y_{2^0} spherical harmonic with two polar jets and annihilation boundary between hemispheres.

3.3. Revised Galaxy Evolution

As inferred from section (3.2), distant galaxies are older than local populations. This is contrary to an expanding model, where objects are predicted to be younger as redshift increases. Distant galaxies and clusters with respect to Earth should therefore contain higher fractions of cold baryonic matter, increased star formation rates and high metallicity. Relative to the local space, galaxies and clusters display characteristics that are evident of an origin from hot, x-ray emitting gas. The x-ray emitting gas is the product of decaying quark matter as inferred from its connection to the dark flow. Both Λ CDM and the purposed model are similar in the sense that the local space emerged from dense quark matter. Relative to an expanding model, it is expected that high redshift clusters and galaxies are hotter than similar local populations; observations however depict the exact opposite. For example, local x-ray emitting clusters transform into lyman-alpha blobs beyond 2z. To temporary resolve this problem, it is usually assumed that drastic major mergers take place and heat up the intergalactic medium. However, it is illogical to have two cool clusters with high metallicity merge into a single hot cluster with low metallicity.

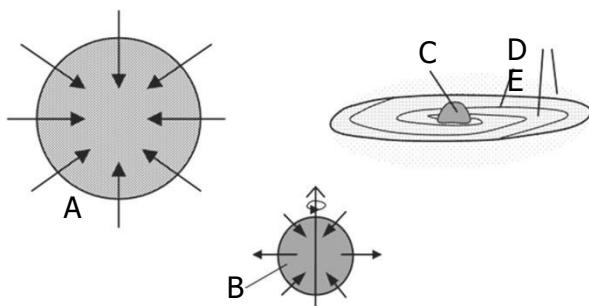
As the local jet of quark matter begins to cool and expand, it decays into a dense non-metallic gas. The oldest stars known to date are metal poor, indicating that they formed some time after this phase. All elements heavier than helium are usually produced through nuclear fusion. The following population II stars are abundant in both globular clusters and elliptical galaxies^[AP]. Most elliptic galaxies contain only population II stars and large amounts of x-ray emitting gas^[AT]. The source of x-rays (0.5 — 1.5 keV) is thermal bremsstrahlung due to hot ionized gas (5 to > 15 keV)^[AZ]. Young stars often undergo supernova after billions of years, enriching the surrounding medium with metallic elements. Around this point, the x-ray emitting gas

originating from the core's relativistic jet begins to drastically cool. Metal-rich population I stars then form in the dense but cooler regions of galaxies and nebulas. Beyond Earth's current position in the flow, baryonic matter is observed to become increasingly colder and more metallic. Due to the abundance of preferred fuel^[AQ], galaxies between a redshift of 0.5z — 3z will demonstrate intense star formation^[AR]. Late-type galaxies commonly have regions of active star formation, which is favored due to cold, dense baryonic matter^{[AQ][AU]}. The cold interstellar gas required to ignite these galaxies was recently observed. A letter to Nature states that galaxies at redshift of 1.2z and 2.3z consist of 34% and 44% cold baryonic matter respectively, which is 3 — 10x more than local late-type galaxies^[AQ]. Beyond the overly abundant blue field galaxies, red and ultra-red galaxies dominate^{[AV][AW][AX]}. The red and ultra-red colors are an indication of abundant dust or mature star populations^[AX], both being characteristics of older galactic populations. Several of the distant red galaxies within the Hubble deep field image are also undergoing mergers, which is consistent with half of normal galaxies experiencing a major merger by 0.75z.

The previous overview of star formation does not drastically differ from current models. It is instead the evolution of galaxies that must be heavily revised. Although the purposed model and Λ CDM both insist the local universe originated from dense quark matter, their perspective of time versus redshift are opposite. This will vary the inferred evolution of galaxies with respect to observations. For example, predicted time-scales drastically differ between models. To explain galactic formation with respect to Λ CDM, large amounts of dark matter are required. These processes should instead occur over 50 — 100 Gyr, which is why excessive amounts of dark matter are required. Time-dependence is later reinforced by comparing simulated and observed merger times; discussed in section (3.5).

Relative to the purposed model or Λ CDM, the initial environment will consist of decaying quark matter. Galactic formation will therefore take place in a hot, non-metallic gas. With any classical gas, the system will move towards equilibrium with respect to density and pressure. As thermal pressure is overcome by gravity, regions will begin to collapse. Due to the conservation of momentum, any radial collapse will be transferred into angular momentum. The properties of galaxies also depict an evolution from early to late-type; i.e. an older galaxy will be more metallic, contain vast amounts of cold baryonic matter and be less symmetric in shape. This transition from early-type galaxies into late-type is depicted in figure 3.6.

It is commonly debated whether disk galaxies merge to form ellipticals^[AP]. However, mergers are insignificant with respect to galactic evolution prior to 1z. A large elliptical galaxy was also discovered to contain a rotating disk of x-ray emitting gas^[AS]. In addition, elliptic galaxies at 0.5z are observed to be on average rotating faster than those in the local space^[AY]. Distributions of star populations further agree with the purposed model of evolution^[BA]. Due to flawed foundations however, various modern theories must be discarded. This includes drastic merger rates and dark matter, as the cooling of x-ray emitting gas and unstable rotational curves are

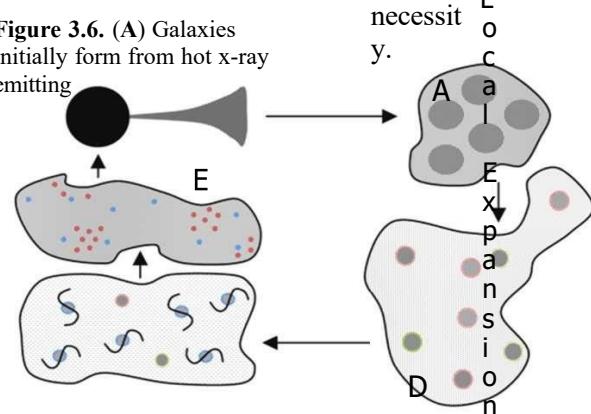


Approximately 60% of major galaxies are disk within the local supercluster, while the remaining are mostly elliptical. From a redshift of $0.5z - 2.0z$ intense blue galaxies dominate, which continuously transform into blue irregulars beyond $1z^{[AR]}$. It is important to notice however that galactic evolution is minimal for those involved in the faint blue galaxy problem, i.e. ones in the range of $0.3 - 0.7z$. This includes luminosity, color and size variations. Application of modern literature must also proceed with caution, as old calculations do not take into account the proper curvature of the universe. For example, a constant number density of elliptic

galaxies out to $1z$ with respect to Λ CDM^[BH] would indicate a relative decrease after considering proper large-scale curvature.

Another aspect of galactic evolution arises from the statistically significant dark flow. With the bulk of galaxies and clusters falling towards the center of the universe, there must be another local flow that replenishes these populations. From WMAP, small variations in the cosmic background radiation were measured and analyzed^[AZ]. These variations are due to scattering from clusters of galaxies containing large amounts of x-ray emitting gas. It is clear that hot, x-ray emitting galaxies and clusters should be younger and therefore closer to the local jet. The dark flow is therefore not only expected by the new the mechanisms behind model, but a necessit y.

Figure 3.6. (A) Galaxies initially form from hot x-ray emitting



C T m
Present i e

gas, which begins to collapse after sufficient cooling. **(B)** Early galaxies obtain a preferred axis of rotation due to local or global gravitational fields. Young metal poor stars form in the bulge due to preferred density. **(C)** Young metal rich stars later form in the remaining bulge **(D, E)** while older populations are transported outwards due to unstable rotational curves.

Figure 3.7. **(A)** Central core with jet consisting of hot, dense quark matter. **(B, C)** After the quark matter decays into hot x-ray emitting gas, low metallicity clusters, galaxies and stars begin to form. **(C, D)** Figure 3.6 overviews galactic evolution. **(E)** Increased merger fractions, metallicity and cold baryonic matter with respect to the local space.

The last part of this section will focus on deriving the time dependence of distant galaxies and clusters. Since all objects not in the dark flow will be falling into an asymptotically flat universe, the redshift equation can be applied to determine kinematics. The goal is to approximate the total amount of proper time a galaxy experiences while following a geodesic originating from Earth's present position, i.e. the reference frame is relative

From the initial position, a change in vacuum energy density can be made relative to the averaged start of flow at $d^7 \cong 0.375$ Gpc. After normalizing units ($\Delta = 1$), the y-intercept can be negated by applying $\nabla = 0.02683 \cdot D'$. The normalized redshift equation (199) is then applied to relate an object's spectral redshift to a variation in metric distance D' with respect to Earth.

$$\text{to Earth. Metric } z_{\text{net}} = \sqrt{\frac{9}{\nabla(\nabla + 2)}} \quad (19)$$

distance (196) is defined

from the slope of the universe and y-intercept, although only⁽¹⁹⁾₆₎ the slope depicts the actual shape.

Since galactic evolution is relative to proper time, proper velocity must be applied (200).⁽²⁰⁾_o

dx'

$$d^7 = (3.0788 \cdot 10^{-43}) \nabla + 0.3748$$

Solving equation (196) in terms of vacuum energy density results in equation (197).

$$u = \frac{dr}{d\nabla} = v(\nabla + 1) = c_o \sqrt{\nabla(\nabla + 2)}$$

With both (199, 200), the proper velocity and

$$\nabla = (3.248 \cdot 10^{42}) d^7 - 1.217 \cdot 10^{42} \quad (197)$$

To determine the time dependence of non-local

Plugging equation (197) into the redshift equation (191) provides the relation between redshift and metric distance. Directional dependence for local redshift however must also be considered, where the y-intercept provides the average distance to the start of flow towards central core.

redshift are coupled to metric distance (D') as the only free variable. The proper velocity versus redshift provides all information necessary in order to determine the duration of proper time a distant object has

experienced. Proper velocity is relative to the amount of metric distance traveled with respect to a moving objects perspective of time. The averaged proper velocity over metric distance D' is
(201). Carrying out the integral of equation (201)

galaxies, the change in redshift with respect to and applying the relation $u_{avg} = \Delta x' / \Delta r$ provides

proper time must be determined. Each distant galaxy or cluster relative to Earth took r amount of proper time to arrive at the position where currently

observed light was emitted. Therefore, the time it took for a light ray to travel from source to observer is not necessary relative to a steady state model. Directional dependence for local redshift ($<0.2z$) can also be considered by varying y_o from 0 to

0.54 Gpc. For all directions relative

the galaxy has experienced without considering directional dependence.

$$u_{avg} = \frac{1}{D} \int_{D'}^D \frac{dt}{dD'}$$

It is also important to realize that younger galaxies and clusters exist with respect to the
(198)

$\Delta t = (3.248 \cdot 10^{42}) D'$ will

∇

follow the previously derived time dependence.

to Earth, the average change in vacuum energy density between the start of flow (d'_o) and finite amount of metric distance (D') is defined as

(198).

t_h^e

t_o^o

t_a^a

t_i^i

t_m^m

t_e^e

t_a^a

t_c^c

local jet, observed in the form of the dark flow. Neither Hubble's law nor the new redshift equation are capable of modeling this since it has a separate origin.

Therefore, the error with respect to the relative age of local objects becomes large due to this uncertainty. However, the majority of objects

(198)

will

Due to matter emerging from and falling back towards the central core at relativistic speeds, there must be a turning point where relative motion is minimal. The CMBR dipole moment provides a velocity of $627 \pm 22 \text{ km s}^{-1}$ for the local group^[AZ]. This is approximately 0.209% the speed of light, which is both insignificant and expected for regions that are distant from the central core. If the metric distance from Earth's current position to the central core is about 40.6 Tpc, the most distant 98% of metric space would account for only 7.8% of all proper time experienced. The most distant objects currently observable are approximately 140 billion years older than the local group ($z \approx 8.0z$). For comparison, the Sun would take about 100 billion years to consume all of its hydrogen fuel. Although this would not actually occur, it is clear that the depletion of interstellar hydrogen occurs over long time-scales. For example, UDFy-38135539 is a distant galaxy observed in the Hubble Deep Field and demonstrates strong lyman-alpha emission^[BB]. The light being emitted by the object is passing through dust that has been reionized beyond what can be explained by an expanding model. The presence of neutral hydrogen gas would only be plausible if the universe is cooling as redshift increases. In addition, the galaxies redshift ($8.55z$) would correspond to a proper age of about 144 Gyr with respect to the local group.

Since baryonic matter is already observed to be cooling from the local region up to $2.3z$, evidence supports the aging of galaxies with increasing redshift. This abundance of cold baryonic matter further induces intense star formation, which is observed for distant late-type galaxies ($0.5z$ to $3z$). These redshift correspond to proper times between 28 and 92 Gyr with respect to local populations. Therefore, the epoch of intense star formation is in agreement with the continuous model and expected conditions. Minimal evolution however occurs prior to $0.7z$ (35.7 Gyr), which is later discussed through-

out section (3.5). With respect to proper time, Λ CDM predicts that the current age of the universe is $13.75 \pm 0.13 \text{ Gyr}^{[\text{AO}]}$. Since Earth's position is in proximity to the CMBR dipole turning point, the total proper time experienced from the core's jet back to the surface is twice that of future proper time. From figure 3.8, the amount of proper time experienced up to the surface of the central core with respect to Earth is approximately 230 Gyr. The total time experienced by an observer in the bulk of flow is therefore $460 \pm 100 \text{ Gyr}$ relative to a complete cycle external to the core.

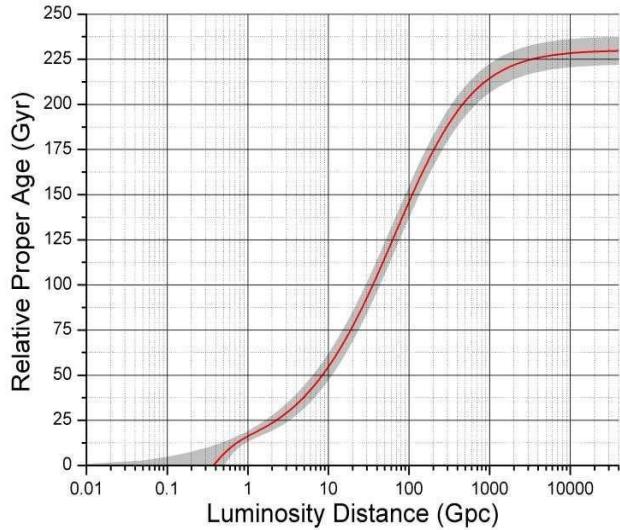


Figure 3.8. Proper age (r) versus luminosity distance (D_L). Error is derived from uncertainty in slope, with the y-intercept ranging from 0 Gpc to 0.54 Gpc .

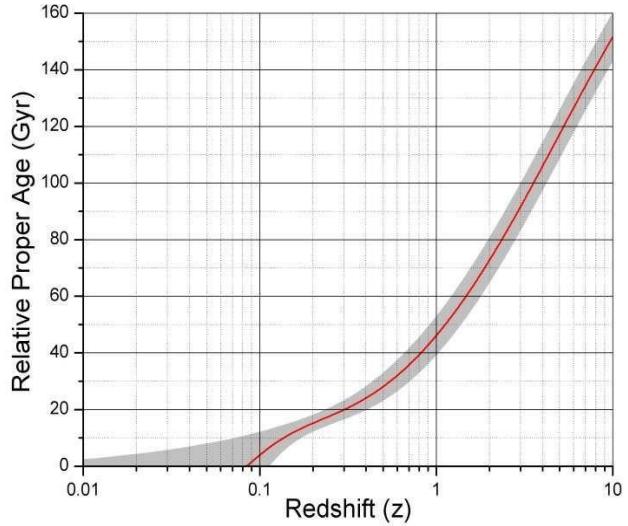


Figure 3.9. Proper age versus effective redshift with respect to Earth; constraints are identical to figure 3.8.

3.4. Angular Scales and Weak Lensing

Although many characteristics can differentiate between models, the curvature of the universe is the only one that offers conclusive proof with current observations. The angular scale versus redshift for Λ CDM was obtained from recent constraints^[BN]. The angular scale or scale-factor ($\text{kpc}''/\text{arcsec}''$) of an asymptotically flat universe is derived from the following method. The distance to an object is determined from its luminosity or metric distance. Considering that local geodesics begin to curve towards the center of the universe at relatively short distances, the projection of distant space will appear almost spherical relative to Earth. The only missing factor is the variations in metric volume induced by the vacuum field of an asymptotically flat universe. From vacuum field theory, the space-time metric is locally isotropic and defined by a scalar field. The circumference of a sphere projected from distant space (202) can therefore be scaled by (g) relative to Earth's perspective.

Plotting the scale-factor ratio between models in figure 3.11 provides several important constraints. This variation is key to ruling out an expanding universe. Although the models are opposite in several aspects, acceptance of poorly constrained hypothesis makes it difficult to rule out Λ CDM. For example, the amount of cold baryonic matter is observed to drastically increase from the local space to $1.2z$ and $2.3z$. With only classical assumptions, galaxies should cool as they age. Although mergers can heat up ISM or ICM, such drastic increase in cold baryonic matter with redshift should be taken as strong evidence against Λ CDM. The processes that occur in the cooling or heating of normal galaxies however are poorly constrained. Ruling out an expanding universe therefore requires aspects that are fully constrained. This is why the angular scales are crucial, as Λ CDM cannot explain the $2 - 3x$ excess of faint blue field galaxies or the disagreement between their size and luminosity. Incorrect predictions of large-scale curvature also induce systematic lensing errors, which currently

$$C = 2\pi d' \gamma_g \quad (202)$$

provide the only direct evidence for dark matter.

Figure 3.10 provides a comparison between Λ CDM and the continuous model.

with
 Λ CDM in black. Due to local directional dependence, Hubble

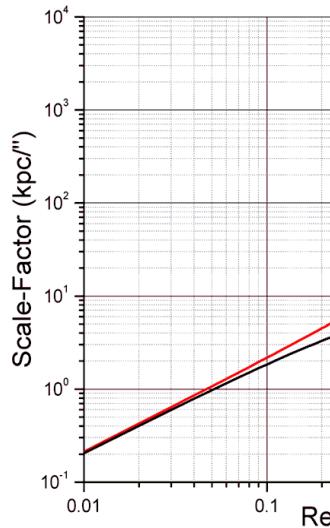


Figure 3.10. The continuous model is depicted in red

The continuous model on the other hand does not require the non-classical assumptions of dark matter and dark energy.

expansion with $H_0 = 68.7$ is applied for redshift below z .

The continuous model however is in agreement with local redshift when applying the full range of y_0 .

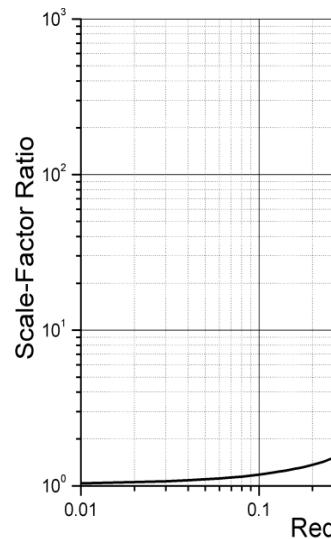


Figure 3.11. Scale-factor ratio between the continuous model

and Λ CDM with respect to redshift. A $5x$ disagreement exists by $1z$ further increasing to $100x$ around $5z$.

The disagreement between each models scale-factor allows distant gravitational lenses to appear stronger than expected from only visible matter. With respect to an expanding universe, there are three distance scales. These consist of angular diameter distance (D_A), comoving distance (D_C) and luminosity distance (D_L). The angular diameter distance corresponds to the visual size of an object at a given redshift, written as (203)^[BO].

With respect to dark matter, the only evidence for its existence is gravitational lensing from distant clusters. Although proof that the universe is not expanding is not discussed until section (3.5), large variations between scale-factors will clearly induce systematic lensing errors. In general, an expanding universe will overestimate distant lens efficiency with respect to visible matter. Dark matter is also self-contradictory with observations. For example,

$$\frac{D}{z} = \frac{1}{\Omega_m} \int_0^z dz' \quad \text{the TrainWreck cluster has lumps of dark matter}$$

$$(203) \quad \text{that coincide with both galaxies and ICM}^{[BP]}.$$

$$= \frac{H_0(1+z)}{\sqrt{\Omega_m(1+z)^3 + \Omega_A}}$$

The remaining distances are related to bolometric luminosity and flux via (204)^[BO].

$$D_L = \sqrt{\frac{L}{4\pi F}} = (1+z)D_A = (1+z)^2 D_C \quad (204)$$

lensing errors are involved, i.e. only the

visible

$$L = 4\pi F$$

$$c$$

$$A$$

flat, steady state projection.

The continuous model does not require comoving distances since the universe is in a steady state. With respect to distant galaxies and clusters falling into an asymptotically flat universe, the angular diameter distance will vary from that of Λ CDM.

Distant objects in an expanding universe appear

larger with respect to a

) matter existed to begin with.

Properties of the Bullet cluster with respect to

Λ CDM are also

contradictive, where it is claimed to be proof of dark matter^[BQ]. At

the same time, the existence of this cluster is not compatible with

Λ CDM^[BR].

Observations that are incompatible with a theory cannot be seen as proof for a specific aspect of it. The remaining

reason dark matter has been inferred are the unstable rotational

curves of galaxies.

Assuming the virial theorem is valid in this case originates from an ad-hoc attempt at forcing

observations to agree with an expanding model. This assumption is flawed as

inferred from observed galactic evolution. The formation of

$$D_A = D_L \left(1 + \frac{y_0}{L_g} \right)^{\frac{4}{3}}$$

The slope (S_0) was discussed in section (3.2), with the y-intercept (y_0) providing the average distance to the start of flow towards central core. Redshift becomes directionally dependent due to the local deflection of geodesics; on average, this distance is about 0.375 Gpc. The right side of equation (205) should therefore have D_A set equal to D_L prior to y_0 .

galaxies does not begin from reionized plumes of hydrogen gas anchored to dark matter, but instead hot x-ray emitting gas. Without large amounts of dark matter, galaxies could not have formed within

the time-scales predicted by an expanding model. From section 3.3, the age of the local space is well over 100 Gyr more than what an expanding model predicts. This disagreement increases with redshift, where time is viewed in the reverse of actuality. The inferred existence of dark matter therefore originates from improper foundations.

A spherically symmetric lens can be applied to compare the continuous model and Λ CDM. The inferred magnification of a lens is dependent upon the angular diameter distances (D_A). As previously stated, current observations support the conclusion of systematic lensing errors, which arises from the incorrect shape of the universe. Since vacuum field theory simplifies to EFE's weak field limit, the formulation is equivalent. Therefore, the equation for a spherically symmetric lens is (206)^[BS].

Although the disagreement between models is relatively small prior to 1z, the systematic error becomes apparent from moderate to high redshift. The ratio of distances provides a single variable (k) that directly scales the mass in equation (209). This indicates that ratios between each models k is equal to the inferred abundance of mass. Figure 3.12 depicts this abundance relative to several spherical lenses. The majority of lensed sources range from 3 – 5z, where the disagreement becomes apparent. For example, the Train Wreck cluster would be

$$\frac{\beta_{LS}}{D_{LS}} = \theta - \frac{\beta}{D_{OL} D_{OL}^2 c^2 \theta^2} \quad (206) \quad \text{inferred to have 49\% more mass at 5z than actually}$$

$$k \frac{4G_0 M}{c^2 \theta^2 \beta^2} = \theta -$$

θ exists with respect to Λ CDM. The Bullet cluster

The free variables are D_{OL} , D_{LS} and y ; these depict the distance from observer to lens, observer to source and the horizontal offset of the source respectively. The remaining variables represent the angular offset of the source (β) and image (θ). The magnification due to a spherical lens is therefore determined by equation (207).

$$\frac{\theta}{d\theta}$$

would instead be inferred to have 94% more mass at identical redshift.

As previously discussed, the initial motivation behind dark matter originates from applying the virial theorem to disk galaxies. Unstable rotational curves and cooling of x-ray emitting gas however are the mechanisms behind galactic evolution. This is reinforced by observations of entropy, angular

$$\mu = \beta d\beta \quad (207)$$

size,
number densities and stellar populations. To

— —

a
v
o
i
d
c
i
r

c
reaso
ng,
only m
lensing
data p
should a

From the initial conditions, $D_{OS} = D_{OL} + D_{LS}$ and β is determined from equation (208).

$$\beta = \arctan\left(\frac{y}{D_{OS}}\right)$$

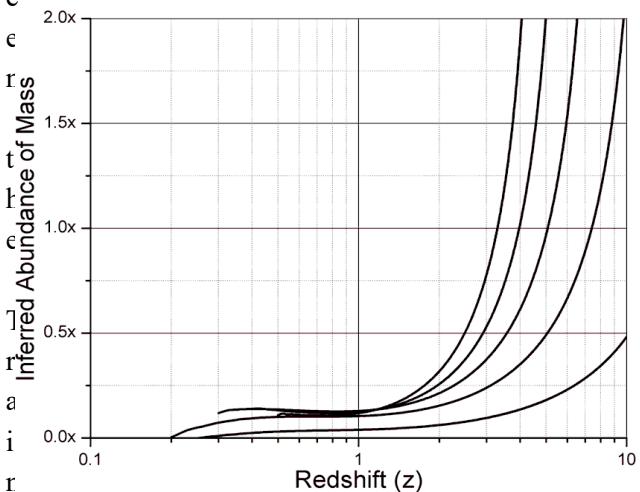
Solving the lens equation (206) results in (209). (208)

$$\begin{aligned} 1 & \\ \sqrt{\beta^2} & \\ + k & \\ \frac{16G_o}{M} & \\ \theta &= c \\ &=) \\ 2 & \\ (\beta & \\ \pm & \end{aligned} \quad (209) \quad \text{---}$$

be compared to the visible baryonic matter within clusters. Section (3.5) further demonstrates that an expanding model or Λ CDM predicts incorrect large-scale curvature.

r
i
s
o
n

b
e
t
w
e



w
r
e
c
k

c
l
u
s
t
e
r

(
0
.2
0
1
z
)

T
a
b
l
e

3
.1

p
r
o
v
i
d
e
s

a

a
n
d

B
u
l
e
t

c
1
u
s
t

2
9
6
z
)

T
a
b
l
e
-
3
-
1
-
k
-
f
a
c
t
o
r
-
b
e
t
w
e
e
n
-
c
o

n
t
i
n
u
o
u
s
-
m
o
d
e
l
-
a
n
d
-
 Δ
C
D
M

	k (0.201z)	k (0.296z)	k (0.201z)	k (0.296z)
0.500z	0.7424	0.3487	0.6749	0.3537
1.000z	0.9574	0.5637	0.8673	0.4996
2.000z	1.036	0.6423	0.8976	0.5299
3.000z	1.055	0.6614	0.8506	0.4829
4.000z	1.063	0.6691	0.7856	0.4179
5.000z	1.067	0.6730	0.7140	0.3464

Figure 3.12. Spherical lenses are plotted from top to bottom at $0.5z$, $0.4z$, $0.3z$, $0.2z$ and $0.1z$ respectively. The abundance is relative to the continuous model versus Λ CDM.

3.5. The Faint Blue Galaxy Problem

The abundance of faint blue galaxies (FBG) up to moderate redshift is known as a grand cosmological problem^[BT]. With observed merger fractions and the angular size of these galaxies, the problem can be resolved with only classical assumptions. From the LDSS deep redshift survey, an 2x abundance exists up to $M_B = 22.5$ with respect to no evolution^[BU]. The survey provides an average redshift of $0.32z$ at $M_B = 21.8$. More distant surveys indicate an 2 — 3x abundance within the limits of $M_B = [22.5, 24]$ ^[BV]. The majority of recent studies also focus on the B- band, where most of the FBGs with M_B ranging from 23 to 24 exist prior to $1.0z$ ^[BW]. With high- resolution imaging, distant FBGs are found to be consistent with local disk and irregular galaxies^[BX]. On average, this dataset provides a similar redshift versus M_B of 21.9 at $< 0.34z >$. The observed OII widths of distant FBGs also indicate intense star formation across the entire disk^[BX]. Many of the local FBGs ($< 0.3z$) however are dwarfs. To the contrary, distant FBGs are not dwarf galaxies but instead intermediate disks and irregulars. Applying the purposed model with FBG surveys produces the various plots in figure 3.13^{[BU][BZ]}.

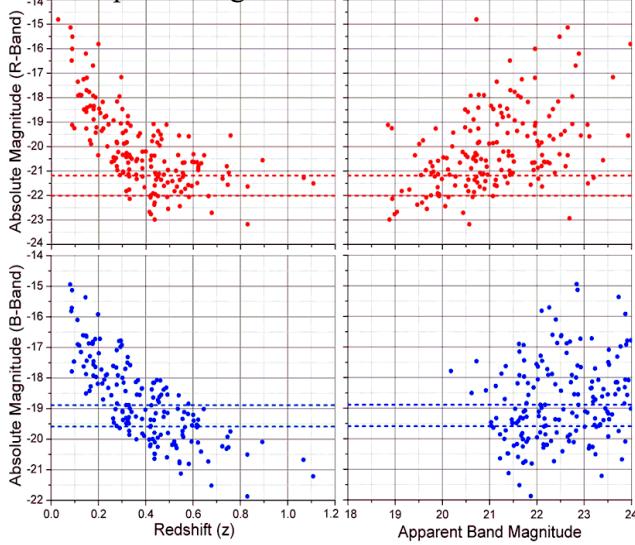


Figure 3.13. Applying a $B - V$ filter to local surveys results in disk and irregular galaxies on average having m_B ranging from

Due to the various attempts at explaining the abundance of field galaxies, each will be discussed in detail. The local FBG abundance is completely compatible with no evolution with respect to their redshift distribution^[CA]. Many have purposed that either drastic mergers or evolution of the luminosity function must take place. However, number counts in the K, R and B bands rule out evolution in the faint end of the luminosity function^[CA]. Additional studies have also concluded that any evolution at the bright end of the luminosity function must be minimal below $0.5z$ ^[BY]. Therefore, the observed 2x abundance placed around $0.5z$ ($22.5 M_B$) cannot be explained by evolution of the luminosity function. Furthermore, recent constraints on merger fractions limit the total amount of major mergers in these regions to $< 30\%$ by $0.5z$. This would in return only reduce the 2x abundance to 1.7x. The FBG anomaly is known as a grand cosmological problem because there is no self-consistent way to explain the abundance with respect to Λ CDM^[BT]. The local abundance of FBGs is crucial for ruling out an expanding model since the various aspects are well constrained. When the observed evolution of local blue galaxies is applied to the continuous model however, the 2x abundance is in perfect agreement with predictions.

The FBG problem at moderate redshift becomes problematic as they display properties of normal sized disk and irregular galaxies. However, Λ CDM predicts that they are 2 — 5x smaller than similar local populations. This has lead to distant FBGs being improperly inferred to as dwarf galaxies. The luminosity function however does not drastically evolve, so this cannot be true. Taking the average of absolute magnitudes from $0.3z$ to $1.0z$ results in $< m_B > = -19.51$ and $< m_R > = -20.96$. This further supports distant FBGs being common late-type galaxies. From high-resolution imaging of FBGs, there is no evidence for an abundance of dwarfs^[BX] -18.89 to -19.59 ; m_R in comparison ranges from -21.19 to

undergoing intense star formation

. Observations

—22.01. After combining several faint blue galaxy surveys, it is observed that many of the FBGs between $0.3 z$ and $0.7z$ are either normal disks or irregulars.

clearly rule
out the
purposed
solutions to

the Λ CDM
faint blue
galaxy problem.

Several surveys allow the size of FBGs to be compared to their absolute B-band magnitudes. The available data is limited, with the majority lacking reliable redshift. Edge detection with difference of gaussians was instead applied to high resolution FBGs^[BX]. This provides major-axis diameters with respect to absolute B-band magnitude as depicted in figure 3.14. Absolute magnitude is related to metric distance derived from spectroscopically confirmed redshift. As previously stated, the distant FBGs are not dwarfs. Bolometric limitations force galaxies with faint absolute magnitudes to be closer, while the remaining are more distant. This is observed in figure 3.14, where disagreement between models increases as m_B decreases. Half of the FBGs are observed to have mild to moderate star formation, usually across the entire disk. The rest range from common young to late-type disk, some of which are very blue or bulge dominated^[BX]. Λ CDM is below even the most extreme cases from the local space. Combined with the lack of luminosity evolution, an expanding universe is ruled out due to incorrect predictions of large-scale curvature. Observations instead insist that the universe is asymptotically flat.

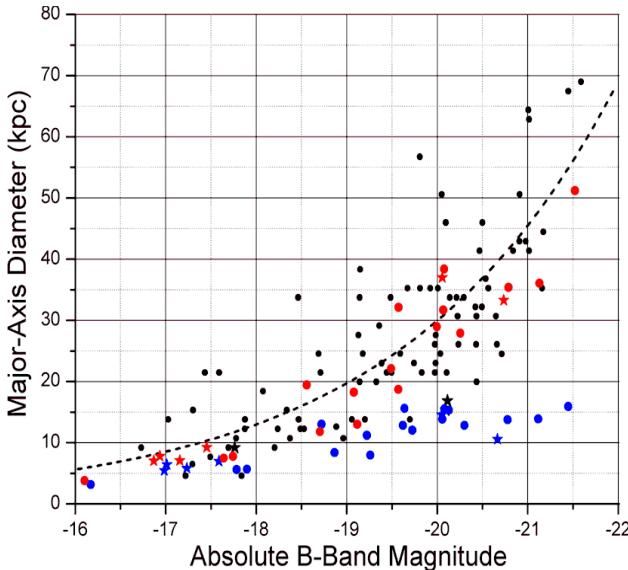


Figure 3.14. Black stars depict several local starburst galaxies, with the remaining circles being normal spiral, dwarf and irregulars. Red are the high resolution FBGs with respect to the continuous model, while blue is with respect to Λ CDM. The dashed trend line is relative to local galaxies with the standard logarithmic fit^[CB]: $m_B = -5.5 \cdot \text{LOG}(A) - 11.9$

Further insight can be obtained from the most massive objects visible at various redshift ($< 7z$). Locally these are hot x-ray emitting clusters, while the more distant populations consist of reionized hydrogen known as lyman-alpha blobs^{[CC][CD]}. As predicted by the continuous model, distant clusters are older with respect to local clusters. The angular diameter of 148 clusters was further measured to demonstrate that the curvature of an asymptotically flat universe agrees with observations. The diameter of each is measured with respect to x-ray emissions up to 3σ above background rates. For more distant clusters, the extent of lyman-alpha emission was instead applied. Figure 3.15 provides a plot of these clusters and expected size assuming no evolution. The majority of objects were observed by Chandra ACIS-I^[CE] and the XMM cluster survey^[CF]. Since the angular scale was previously verified with FBG luminosity out to $0.7z$ in figure 3.14, there must be minimal change in cluster size up to this redshift. Mergers in figure 3.15 are depicted by large errors, where uncertainty provides the minor to major axis diameters. A major merger will peak at 30% above initial angular diameter, with an error around 25%.

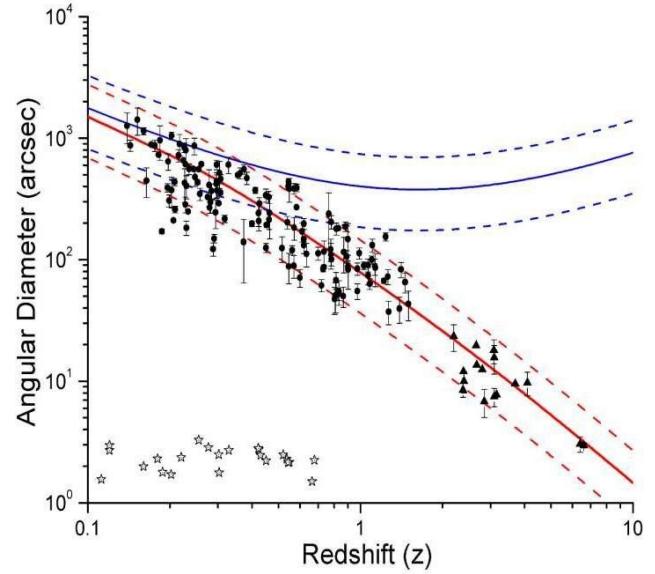


Figure 3.15. 148 massive clusters with diameters determined by averaging the minor and major axis of each. Red is relative to the continuous model with a cluster size of $3.25^{+1.75}_{-1.75} \text{ Mpc}$ and blue is with respect to Λ CDM. Circles depict x-ray emitting clusters, while triangles have lyman-alpha emission. Stars at the bottom represent the FBGs from figure 3.14.

The FBG problem becomes more apparent when observed merger fractions are examined. Several recent studies focus on calculating merger fractions by surveying distant galaxies. For moderate redshift ($< 1.2z$), galaxies with low to intermediate mass are inferred to have merger fractions of 5 – 10%^[CG]. From a survey of massive galaxies, the merger fraction ranges from 0.03 to 0.14^[CH]. Another finds a morphological merger fraction less than 6% for massive disk galaxies prior to 1z^[CJ]. An in-depth survey focusing on both active and prior mergers allows uncertainty in merger duration to be ignored. It is observed that major mergers such as those between two medium-sized disks occur once on average by 1.4z^[CJ]. Minor mergers between satellite galaxies and their host are about three times as abundant. However, these do not explain the excess of intermediate sized FBGs. In other words, the FBGs are not satellite galaxies. Distant FBGs are fully consistent with common late-type galaxies, with local populations being dwarfs similar to NGC 4214 or NGC 1310. Without drastic merger rates, there are no remaining explanations with respect to an expanding model. The purposed model however predicts for the excess to exist due to the curvature of the universe.

Merger fractions beyond 1z increase in response to the time-scales involved, i.e. these objects are at least 45 Gyr older than the local group. Red and ultra-red galaxies are found at moderate redshift. The Hubble Ultra Deep Field (HUDF) shows many of these galaxies undergoing mergers, resulting in deformed galaxies with tails or multiple cores^[AW]. Products of major mergers are not consistent with local ellipticals, explaining why these red galaxies are fitted with extended star formation histories and abundant dust^[CK]. From the HUDF, a peak merger fraction of 30% occurs around 2z with massive galaxies^[CL]. High merger fractions of 40% to 50% are observed beyond 2.5z, where the objects are consistent with Lyman-break galaxies^[CM].

Comparing proper time between 0z and 1z, the continuous model has an additional 38.3 ± 6.6 Gyr. Current studies based upon Λ CDM should therefore have observed merger times 4.1 — 5.8x quicker than expected from simulations. These simulations rely on fundamental physics, making it difficult to explain how the process would be occurring at five times the expected rate. Many also underestimate merger times due to the inclusion of dark matter. There is however no proof or direct evidence for dark matter, at least not in any exotic forms. Direct attempts at locally detecting dark matter have also failed^{[CN][CO]}. Instead, dark matter is the result of systematic lensing errors and improper foundations. Without dark matter, the expected mass of galaxies and clusters will decrease. Therefore, merger times from prior simulations are likely underestimated.

From one study, the first pass on average occurs at 0.72 Gyr for Sbc galaxies. Max separation occurs on average by 1.20 Gyr, while galaxies merge at < 1.88 Gyr $>$. After < 2.88 Gyr $>$ have passed the galaxy is considered to be a merger remnant^[CP]. With the merger fractions applied on the next page, the one merger per galaxy by 1.4z fits with an average merger time of 4.0 Gyr. Both values can be compared to inferred merger times with respect to Λ CDM. Several estimates for $< r_{oac} >$ range from 0.2 Gyr to 1.0 Gyr^[CQ], which is in disagreement by 2.9x to 17.5x with simulations. A more precise ratio can be obtained from table 3.2. Taking the average of several surveys results in an average merger time of 0.65 Gyr, which results in a 4.4x to 5.4x disagreement with numerical models.

Table 3.2. Merger times of close galaxy pairs

Reference ^[CQ]	$< z >$	$< r_{obs} >_{S08}$	$< r_{obs} >_{Co6}$
Patton & Atfield	0.05	0.36	0.37
Lin et al.	0.79	0.63	0.63
de Ravel et al.	0.72	1.61	1.50
Kartaltepe et al.	0.70	0.33	0.35
Bundy et al.	0.83	0.32	0.35
Average	0.62	0.65	0.64

Several sets of merger fractions are available from recent literature. Some of these could however be overestimated for several reasons. For example, mature disk or irregular galaxies can demonstrate several areas of intense star formation, which may be improperly interpreted as remnant cores from a previous merger^[CR]. The application of maximum likelihood techniques also tends to overestimate merger fractions^[CI]. A rough estimate is obtained by assuming one major merger per galaxy at 1.4z. Since merger times relative to Λ CDM are underestimated, the averaged value from simulations is applied (2.88 Gyr). Time dependence is with respect to the purposed model as discussed in section (3.3). The fractional merger rate is defined by equation (210), depicting the fraction of galaxies completing a major merger per proper merger time.

$$\underline{f_m}$$

With proper time as derived from the continuous model, local merger fractions are easily constrained below 0.04. For example, the fractional merger rate (R_t) derived from a constant merger rate provides merger fractions (f_m) ranging from 0.030 to 0.040. Distant close pair merger fractions however range from 0.066 to beyond 0.10 throughout the various surveys. Local close pair surveys provide merger fractions between 0.005 and 0.02. Therefore, the approximation of $R_t = 0.0120$ is overestimated for redshift below 0.7z. Even after applying this value with respect to Λ CDM, the 2x abundance of FBGs at 0.5z only decreases by 28.6%. If the number density at 0.5z is 200 galaxies per metric volume, the final amount of galaxies would decrease to 143 by 0.0z. Λ CDM or an expanding model is therefore off by 43% when overestimating major mergers prior to 0.5z.

$$R_c = \langle r_m \rangle \quad (210)$$

Merger fractions above 0.04 for the local space are clearly too high. The survey that is found to be

For a given redshift, the change in total galaxies is related to the fractional merger rate and number density by equation (211).

$$\frac{dN}{dr} \bar{\bar{R}}_c \bar{\bar{N}}$$

the most consistent focuses on galaxy pairs with $M < -19.8$ ^[CQ]; this limit coincides with common

Sb/Sc disk galaxies. The projected radius for these close pairs ranges from 5 to 20 kpc (Kartaltepe et al. 2007).

With respect to the average merger time obtained from simulations^[CP], the ensemble of Sbc

Solving equation (211) for number density results in exponential decay (212).

mergers has the majority of runs starting at 11 kpc. A few runs have

much initial radius
greater ranging

$$N(r) = N e^{-Rr^c} \quad (212)$$

from 44 to 50 kpc. Some of the merger times in

o table 3.2 also

Assuming the merger rate is constant relative to proper time, R_t is determined with respect to one major merger per galaxy by 1.4z; this results in $R_t = 0.012 \pm 0.002$ Gyr⁻¹. The 2.88 Gyr merger time translates to a constant merger fraction of 0.035 relative to the continuous model. If the 4.00 Gyr value is applied, the constant merger rate

apply such large distances (de Ravel et al. 2009), indicating that the ratio of expected versus observed merger time is consistent with prior parameters.

Extrapolating data from Kartaltepe et al. (2007) provides table 3.3. These fractions are applied with proper time to determine the redshift dependence of major mergers.

Table 3.3. Merger fractions with respect to redshift

varies to 0.048. However, redshift and time are not directly proportional. For example, about 64% of proper time prior to 1.4z occurs before 0.7z. The merger fraction with respect to redshift would therefore be approximately < 0.035 prior to 0.7z

and greater than beyond.

Relative to the local space, merger fractions as determined from morphology require that 3% of local galaxies are merger remnants^[CS]. Averaging several local pair studies places the merger fraction at 0.018, while Kartaltepe et al. (2007) provide 0.007. Relative to the purposed model, one merger on average at 1.4z should also include mergers that have already occurred prior to the local space. This requires that $3\% \pm 1.8\%$ of the local population has already undergone mergers, which is insignificant. To include merger fractions beyond 1z, the data was extrapolated with an average between several surveys. These vary between 0.19 and 0.22 at 2.5z, with a maximum fraction for any reference being

0.30 at 2.0z. Applying these additional constraints with respect to the continuous model provides the relative amount of galaxies for a particular redshift in figure 3.16. The one major merger per galaxy occurs at a redshift of 1.2z instead of the previously referenced 1.4z. The 4.4 — 5.4x disagreement with expected merger times makes it difficult to plot the amount of galaxies with respect to Λ CDM. The normalized distribution of proper time between models however is nearly proportional at low z. The 3x abundance around 0.8 — 1.0z would therefore decrease by 32%, i.e. Λ CDM is off by $104\% \pm 25\%$ in effective number density. The 2x abundance at 0.5z indicates an error of $70\% \pm 15\%$.

The ratio of scale-factors can be applied to the normalized number density in order to determine the expected abundance of FBG, depicted in figure 3.17. The 2x abundance at 0.5z is in agreement with prior constraints including lack of drastic merger fractions and minimal luminosity evolution. The abundance peaks beyond the observed 2 – 3x disagreement relative to $M_B = [22.5, 24]$. Apparent magnitudes of FBGs versus redshift were further predicted from the absolute B-band distribution of local blue galaxies. With the continuous model and observed mergers, 22.5 M_B and 24.0 M_B correspond on average to 0.54z and 1.01z respectively.

Although observations such as baryonic matter cooling with increasing redshift can be blamed on various hypothesis, there is no answer for the 2x to 3x abundance of FBGs. Combined with luminosity characteristics, they should be nearly the same size as local late-types. However, the sizes inferred from Λ CDM are 2 — 5x smaller than would be expected. The nearly equal disagreement between both of these aspects must be due to improper curvature of the universe. Occam's razor alone would support this conclusion; however, all viable explanations have also been ruled out.

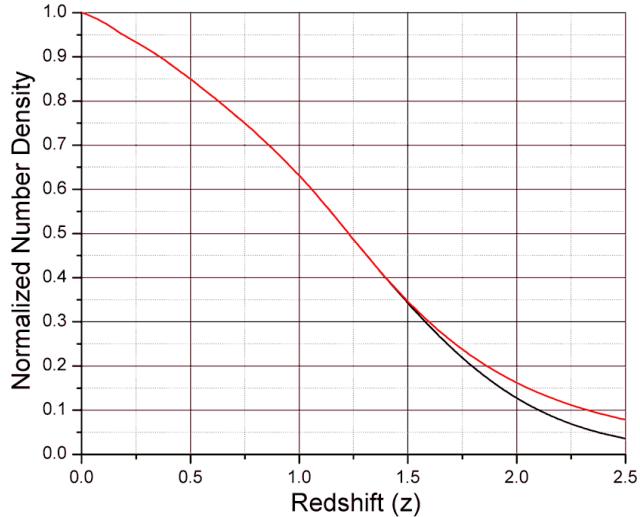


Figure 3.16. Number density of galaxies with mergers only.

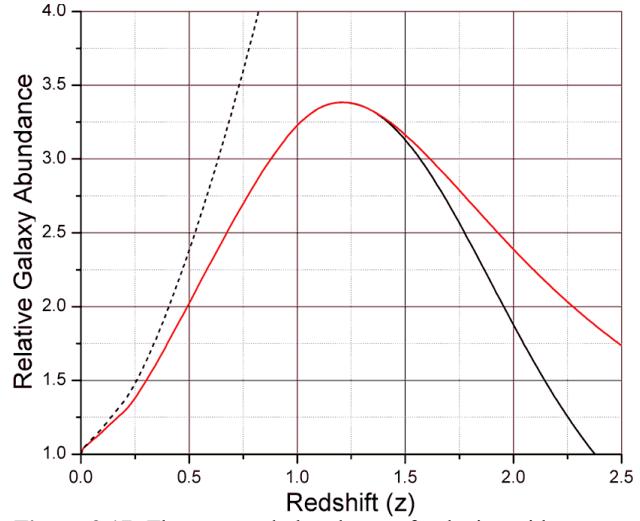


Figure 3.17. The expected abundance of galaxies with respect to no evolution Λ CDM is represented with a dashed line. Red corresponds to the continuous model with the merger fractions provided in table 3.3 and $\langle r \rangle = 2.88 \text{ Gyr}$. Black is maximum merger fractions extrapolated beyond 1.3z^[CM].

3.6. Metallicity

Nuclear entropy is also useful for differentiating between the purposed model and Λ CDM. There are several properties of galaxies that can determine the redshift dependent evolution of metallicity. The focus therefore changes from that of large-scale curvature back to time dependence versus redshift.

Λ CDM or an expanding model requires galactic age to decrease with distance, while the continuous model predicts the opposite. Time-scales involved were also found to agree with the purposed model after comparing simulations to observed merger times. FeII:MgII ratios and variations in magnesium within galaxies further provide strong evidence for the continuous model. Since the models provide opposite predictions for time dependence, evidence for the purposed model is evidence against Λ CDM.

The evolution of metallicity can be inferred from galactic morphology, which trends from early-type in the local space to late-type for the more distant galaxies. Early-type galaxies consist of ellipticals and lenticulars, usually in clusters containing large amounts of hot, x-ray emitting intercluster medium (ICM). Clusters such as Abell 1367 demonstrate how galaxies contained within regions of hot ICM lack active star formation, i.e. they are considered to be passive^[CT]. The majority of galaxies on the perimeter of the ICM region are either active or starburst. Similar clusters such as Abell 1656 are common in the local space, containing primarily early-type galaxies and a few late-type populations. The observed epoch of intense star formation is from 0.5z to 3z, in agreement with x-ray emitting gas inhibiting stellar formation. The ISM is also about 3.8x cooler by 1.2z relative to local galaxies. It was previously discussed how lyman-alpha blobs have dimensions similar to local x-ray emitting clusters. An expanding model however provides the wrong dimensions of these objects by several orders of magnitude. The lyman-alpha blobs are instead massive clusters that have cooled over ≥ 50 Gyr. All of these observations agree with the continuous model and purposed revisions to galactic evolution.

Local merger fractions are too insignificant to play a role in the thermal or nuclear entropy of galaxies prior to 1.0z. Even with a major merger, the nuclear entropy of a galaxy cannot drastically vary. Early-type galaxies for example contain an abundance of population II stars, which are metal poor. Middle-aged spiral galaxies on the other hand contain a mixture of population II and I stars. These galaxies are not metal poor, containing dust and much less x-ray emitting gas. A collision between two disk galaxies may increase the temperature of ISM; however, it will not reverse nuclear entropy or eliminate prior population I stars. There is also no clear transition from disk to elliptical galaxies in surveys. NGC 6240 is a good example of a major merger between two disks. It is similar to the major mergers occurring with red and ultra-red galaxies in the HUDF. These red populations have abundant dust, which is not similar to the x-ray emitting gas in local elliptical galaxies. They instead have very luminous cores with tails or peculiar shapes. Time-dependence of galactic entropy clearly disagrees with Λ CDM when considering observations.

From moderate to distant redshift (2 – 4z), the purposed model predicts 60 – 110 Gyr of evolution from the local space. Therefore, dense regions of intense star formation should begin to deplete primordial hydrogen and helium. This however does not imply that regions without star formation will become metal rich, i.e. distant galaxies or clusters are usually embedded within regions of reionized hydrogen. Galactic star formation will also depict the evolution of ISM. Type II supernova events are associated with metal poor stars, where the final Fe:Mg ratio is about 1.65x. Type Ia events are more metallic, producing Fe:Mg ratios far above type II events (393x). The size and degeneracy of stars also play roles in the type of supernova. For example, type Ia events are inferred to occur from massive stars and are much more energetic than other types. The location of SNIa are also consistent with the continuous model, which uses them to track the flow towards central core.

The iron and magnesium concentrations in ISM play a crucial role in determining the relative age of distant galaxies. Nuclear entropy naturally favors the production of iron over long time-scales, with magnesium slightly lower with respect to nuclear potential. Both are created from nuclear fusion and should be subjected to similar environments. For example, any mechanism besides nuclear fusion that varies FeII will proportionally vary MgII. Since rotational curves are observed to be unstable, it is unlikely that supermassive black holes at the center of galaxies will substantially vary the surrounding metallicity. Therefore, distant galaxies are expected to have increased metallicity and higher FeII:MgII ratios relative to similar local populations.

Prior to discussing FeII:MgII ratios, evidence of increasing metallicity can be inferred through other methods. The absorption of distant sources by local galaxies demonstrates an abundance of MgII. These galaxies are observed from the local space up to about $0.9z^{[CU]}$. Weak MgII absorbers are further observed in abundance from 0.7 to 2.2z, while vanishing beyond $2.7z^{[CV]}$. With observed merger fractions and continuous model, the relative amount of weak MgII absorbers is depicted in figure 3.18.

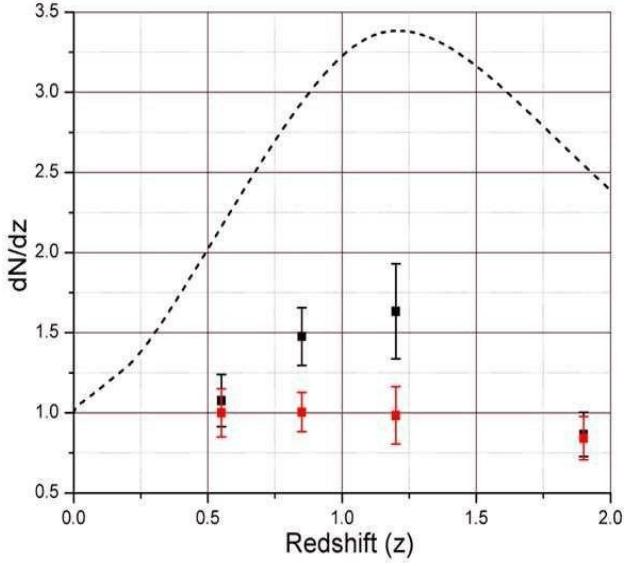


Figure 3.18. The dashed line represents the inferred abundance with mergers from section (3.5). Black indicates the averaged dN/dz from Λ CDM surveys and the continuous model with

The spectrum of quasars also provides valuable information relevant to metallicity. Quasars are some of the most luminous and distant objects in the observable universe. Similar to SNIa, quasars display characteristics of highly degenerate matter. These however are related to active galactic nuclei, which harbor supermassive black holes. Relative to the continuous model, highly degenerate objects are older and therefore in the flow towards central core. Several studies were combined in figure 3.19 to determine the evolution of FeII:MgII with respect to redshift^{[CW][CX][CY][CZ]}. The high ratios observed at 6z are indicative of galaxies that have already undergone intense star formation^[CZ]. For the ISM to be enriched with an abundance of these elements, several generations of stars must have undergone supernovas. Relative to Λ CDM, the proper age of the universe is less than 1 Gyr at 6z. Considering that the region of intense star formation is observed from $0.5z$ to $3z$, Λ CDM does not fit observations. Other observations such as increasing cold baryonic matter with redshift, number densities, galactic evolution and the dark flow agree with galactic age increasing with redshift. An expanding universe on the other hand would violate several fundamental laws of physics including the second law of thermodynamics and nuclear entropy.

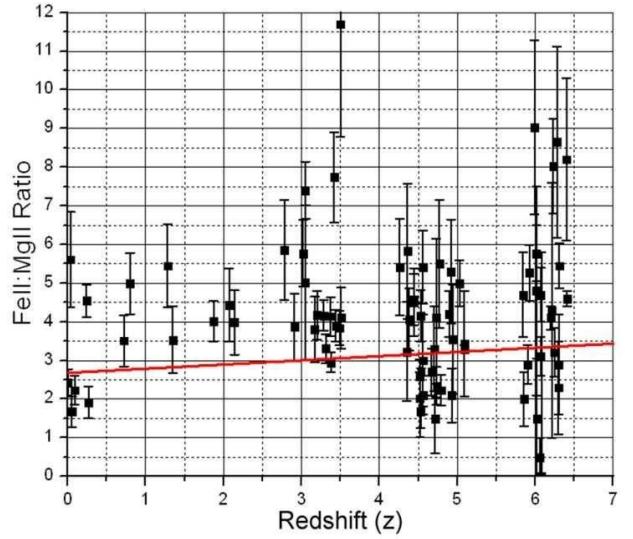


Figure 3.19. High FeII:MgII ratios at extreme redshift are evolution in this case indicative of older galaxies. A is in red. refers to the statistically significant trend is dN/dz number density multiplied by the proper geometric cross section^[CV].

also observed with a slope of 0.108 ± 0.03 .

3.7. Statistical Analysis

Although an expanding universe is conclusively ruled out from incorrect predictions of large-scale curvature, it is possible to compare redshift versus distance modulus between Λ CDM and the purposed model. Distinction between models at low redshift arises in the form of dispersion due to directional dependence. This is predicted by the continuous model from varying y_0 over its complete range (0 Gpc to > 0.54 Gpc). Relative to section (3.2), the slope of the universe and average y-intercept are determined from type Ia supernova (SNIa) and gamma ray burst (GRB). SNIa are superior for determining cosmological distances due to their nearly uniform properties. GRB are less reliable, but can still be used to constrain redshift versus luminosity distance. SNIa and GRBs both display characteristics of highly degenerate matter, which is an indication of relatively older galaxies. These events will therefore be statistically more abundant along the flow towards central core, where the continuous model predicts for objects to be older than the local group.

With respect to Occam's razor, the purposed model fits the shape of the universe with a single constant (S_0). Λ CDM usually needs two constants in the form of dark energy and matter. Dark energy however cannot be directly detected and has no connection to the standard model. It is inferred to exist solely because it allows an expanding model to match observations. Although these non-classical modifications fit redshift versus distance modulus, they fail to agree with the observed shape of the universe. Occam's razor is therefore a necessity for arriving at the proper theory, as anyone can force a model to agree with observations by introducing purely mathematical constructs. Failure to reach parsimony and over-reliance on confirmation rather than refutation are dangerous practices for this reason. From these aspects alone, the models cannot be put on equal footing. The continuous model contains the least amount of free variables and non-classical assumptions.

Λ CDM and the continuous model have similar redshift versus distance modulus predictions from 0.5z to about 10z. Disagreement between models in this region peaks at 0.25μ around 2.5z, making it difficult to differentiate between the two. Λ CDM is constrained by an interpretation of the CMBR and baryon acoustic oscillation data^[BN]. These provide $H_0 = 70.4^{+1.3}_{-1.4}$, $\Omega_b = 0.0456 \pm 0.0035$, $\Omega_c = 0.222 \pm 0.026$ and $\Omega_A = 0.728^{+0.010}_{-0.016}$. Figure 3.20 depicts an ensemble within these limits for $\Omega_m = \Omega_b + \Omega_c$ and $\Omega_A = 1 - \Omega_m$. The purposed model applies the best fitting slope and y-intercept, including uncertainties previously provided in section (3.2). Since SNIa are considered standard candles, observations from 0.1z to 0.5z can easily rule out an expanding universe. The disagreement for local redshift once again arises from directional dependence due to an asymptotically flat universe. The majority of events beyond 1.0z consist of GRB, which suffer from circularity problems. In other words, most methods require that a prior cosmological model be selected in order to determine the luminosity distance to GRB sources. Attempts to avoid this problem apply various relations between GRB parameters and extrapolating SNIa data. However, some of these methods are also found to be model dependent and should not be used to independently determine cosmological distances.

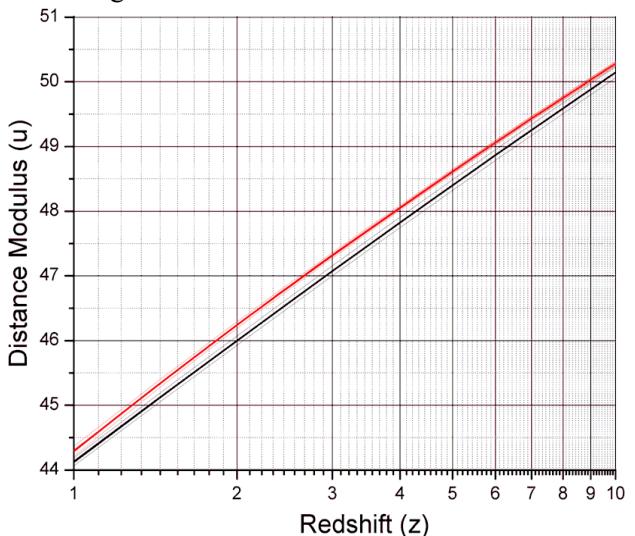


Figure 3.20. Red is the continuous model within limits of error, while black is Λ CDM within limits of error.

A more detailed comparison between models can be achieved from the distance modulus residual with respect to the continuous model. However, it is determined that the dataset obtained from NED redshift independent calculations was contaminated with respect to a few GRBs. The problem arises due to the lack of standardization for GRB sources and inclusion of fiducial data; this is apparent in figure 3.21 from 2.0z to 5.0z. Prior to 1.0z, uncertainty is too small to be fit with homogeneous expansion. For example, a portion of data is outside of either models best fit by more than $\pm 0.50\mu$. However, the trends for the continuous model in figure 3.21 only include uncertainty in slope and average distance to the start of flow towards central core. Considering the shortest path instead begins at 0.0 Gpc and others at distances greater than y_0 , the continuous model explains this dispersion relatively well. The purposed model is also centered on the bulk of available SNIa data prior to 1z indicating a superior fit. To the contrary, homogeneous expansion does not fit observations from 0.1z to 0.5z. Extrapolating SNIa data under the assumption of Λ CDM will therefore produce incorrect predictions for GRBs beyond available SNIa data.

Table 3.4. Average SNIa/GRB error versus redshift

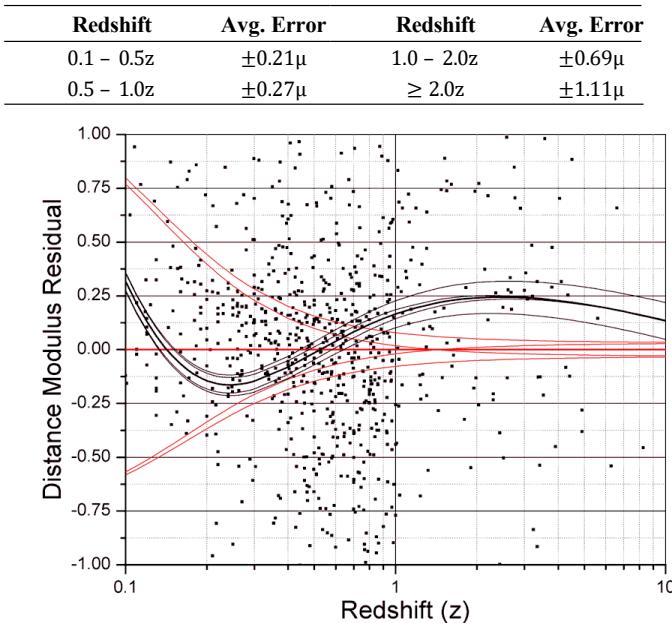


Figure 3.21. A plot of the combined SNIa/GRB dataset used to determine the slope of the universe and average y-intercept.

Although the dataset applied from the NED database is claimed to be redshift independent, some of the data is fiducial; i.e. metric distance is determined assuming the big bang model is correct. For example, there are several data points in figure 3.21 that are anomalously clustered around the Λ CDM trend from 2.0 – 5.0z. To demonstrate this, the residual is plotted in figure 3.22 with respect to several studies. As would be expected from an average error of $\pm 1.11\mu$ beyond 2.0z, the anomaly no longer exists. After applying solely SNIa events for all available redshift, the slope ($3.216 \cdot 10^{42}$) and average y-intercept (0.313 Gpc) are still within previous limits. Several contaminated data points are related to improper methods. GRB 051109A for example contains 3 of 8 values from 2009MNRAS. Removing these from the average increases distance modulus to 46.426μ , with the continuous model predicting 46.429μ . Issues with these references also include the circularity problem and use of the Amati relation. The Amati relation is found to suffer from selection effects and should not be used to probe distance^[DD]. This is applied in 2010JCAP, although the data is in better agreement with SNIa. 2009MNRAS extrapolates only 55% of SNIa data with fiducial methods. This systematically offsets the 2009MNRAS GRBs with respect to the more accurate SNIa dataset between redshift of 0.015z and 1.55z.

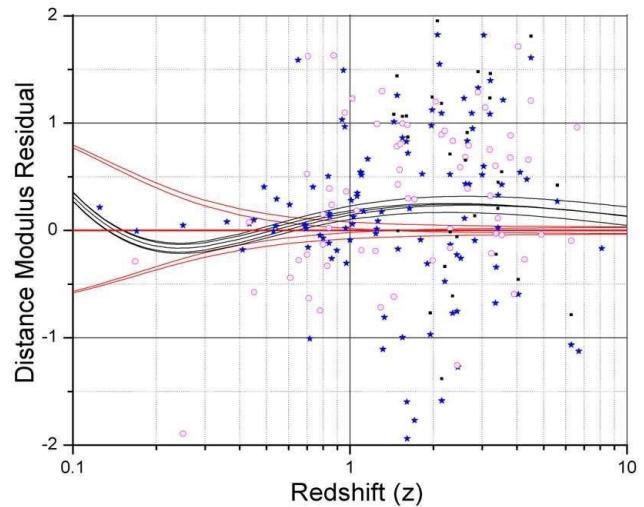


Figure 3.22. Several sets of gamma ray burst are plotted from 2010JCAP^[DA], 2009MNRAS^[DB] and 2009EPJC^[DC].

3.8. The Cosmic Background Radiation

The big bang theory claims that the CMBR is due to a period of recombination after the initial creation of space-time. At this point, the universe cooled until space became transparent to free photons. The big bang theory also claims that the observed blackbody radiation in all directions of space is not a free electromagnetic field, but instead localized packets of electromagnetic energy. To understand why this is not true, the source of blackbodies must be understood. All finite objects with a temperature will emit a spectrum of radiation that peaks at a given wavelength. When an object emits this blackbody spectrum, it is due to the internal kinematic energy or temperature. The free field emitted from massive objects therefore obeys a statistical distribution of internal energy, which is released at the surface boundary. Converting the observed CMBR temperature as depicted by figure 3.23 into the relative value at emission, the core's surface temperature is 3000 K. In comparison, the Sun's surface has a temperature of about 5778 K, indicating that the core likely consists of dense quark matter. Relative to an Einstein black hole with event horizon, the surface will theoretically emit black body radiation in the form of Hawking radiation. The temperature is inversely proportional to mass with a coefficient of $6.1686 \cdot 10^{-8} M_{\odot} \cdot K$. A 3000 K central core with respect to an Einstein black hole would have a mass of $2.056 \cdot 10^{-11} M_{\odot}$, compared to $3.694 \cdot 10^{-8} M_{\odot}$ for the Moon.

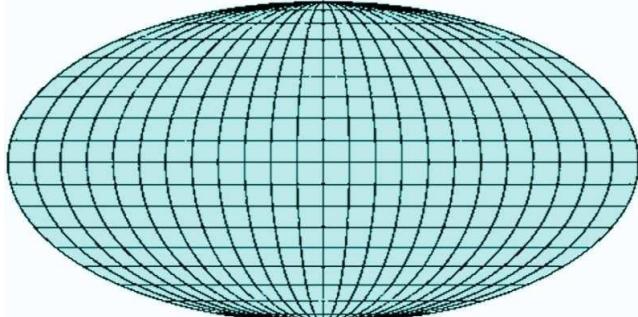


Figure 3.23. The locally observed cosmic background radiation with a temperature of approximately $2.725K$ after redshift.

The CMBR shows peculiarities such as a dipole moment, large-scale bulk flows and additional fluctuations from interaction with external matter. The CMBR temperature prior to subtracting the average value is $2.72548 \pm 0.00057 K^{[BC]}$. Figure 3.24 depicts the dipole moment observed after subtracting the average temperature; figure 3.25 is the CMBR with the dipole subtracted. The dipole moment is due to Earth's motion relative to the CMBR source or the core's surface. Since a spherical body will emit blackbody radiation at a nearly constant z or radius, only the Doppler Effect applies. Relative to the source of the CMBR, the solar system is moving at $369.0 \pm 2.5 km \cdot s^{-1}$ towards $(l, b) = (264.26^{\circ} \pm 0.33^{\circ}, 48.22^{\circ} \pm 0.13^{\circ})^{[AZ]}$.

Taking into account the local group's motion, this velocity becomes $627 \pm 22 km \cdot s^{-1}$ towards the direction $(l, b) = (276^{\circ} \pm 3^{\circ}, 30^{\circ} \pm 3^{\circ})^{[AZ]}$.

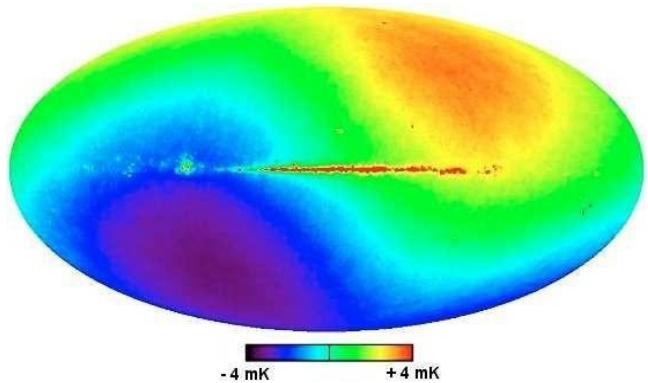


Figure 3.24. After subtracting the average temperature from the sample measured by COBE, the dipole dominates. **Image credited to NASA/WMAP Science Team^[BD]**

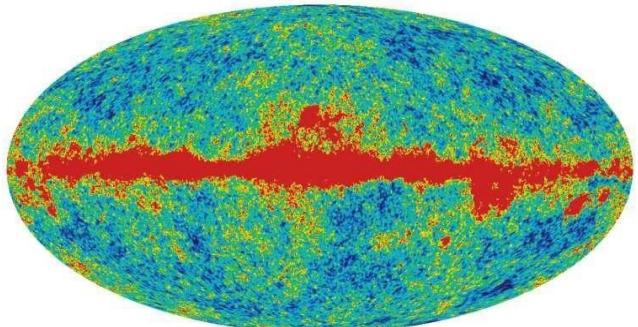


Figure 3.25. After subtracting the dipole moment from figure 3.24, the remaining fluctuations in the CMBR occur externally from the core. Image is provided from WMAP (2003). **Image credited to NASA/WMAP Science Team^[BD]**

Remaining fluctuations in the CMBR are from the Milky Way and scattering of electromagnetic by matter external to the core's surface. Figure 3.26 is the CMBR after removing local foreground sources; it depicts two hot stripes and a central cold patch. The source of the CMBR itself should be at a nearly constant temperature of 3000K. Variations within the cleaned CMBR image instead occur between the foreground and background due to scattering from the x-ray emitting gas of massive clusters. The relativistic charged particles boost the black body spectrum to higher energy levels. This increases the observed black body temperature from deep blue to green and red. Analysis of x-ray emitting clusters has also shown a statistically significant bulk flow extending from the local group to ~ 0.77 Gpc^[AZ]. The velocity is estimated to be [600, 1000] km s⁻¹ from the thermal S-Z effect^[AZ]; however, free field radiation undergoes Thomson scattering. Although many of the directions in local space lead back to the central core, there must logically be a flow of younger galaxies and clusters into Earth's present region. This is necessary to remain consistent with the foundations used in deriving non-local redshift, i.e. the universe is in a steady state. With the medium of galaxies and clusters progressing from hot x-ray emitting gas into cold metallic dust, the dark flow should consist of relatively younger clusters.

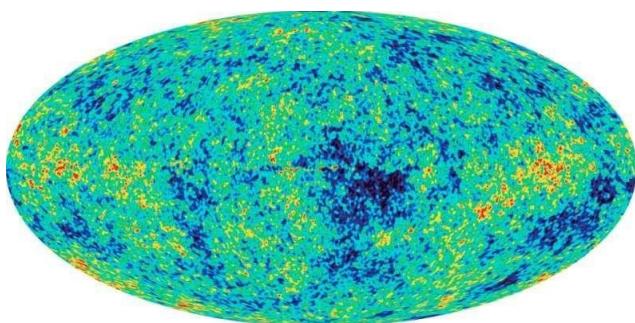


Figure 3.26. The cleaned CMBR is observed to contain a large hot strip originating from a central cold patch. An annihilation boundary or hot ring is surrounding the central cold patch. The base of the local jet should be visible, with the hot strip to the right being a continuation of the dark flow at extreme redshift. Image credited to NASA/WMAP Science Team^[BD]

Conventional theory attributes the CMBR to an epoch of recombination, where photon decoupling takes place. This explanation is only valid for an expanding universe, which has a specific shape and angular scale. Section (3.5) however proved that both galaxy number densities and angular scales are incompatible with an expanding universe. This is easily observed from 0.3z to 0.7z, becoming more drastic beyond. Since an expanding model can be conclusively ruled out, current foundations for the CMBR are invalid. It was further demonstrated in sections (3.5, 3.6) that the purposed model fits the correct shape of the universe with respect to number densities, angular size and several other aspects.

The purposed or continuous model requires two polar jets originating from a central core in order to explain current observations of entropy. If the core acts as a mechanism for baryon asymmetry, an annihilation boundary should also be observed between hemispheres. The inferred local jet is depicted in figure 3.27. Anomalies may distort the cleaned image, possibly beyond use in some regions. The elliptic plane for example runs through the far right side of the local jet, which on average becomes cooler than the surrounding areas. It also correlates with zodiacal dust and several features such as quadrupole/octupole alignment and the cold strips or “fingers” in the southern hemisphere^[BI].

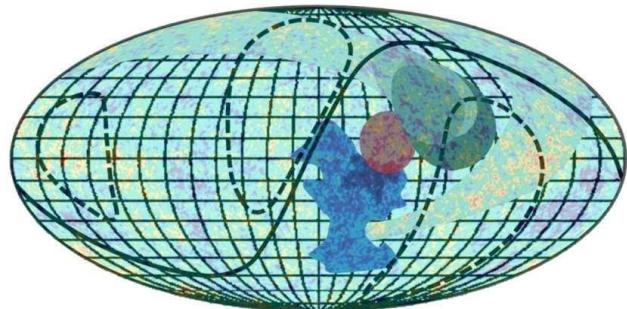


Figure 3.27. The dark flow is depicted in shades of gray, where lighter shades are more distant^[BK]. Since SNIa are used to measure the flow towards central core, an observed SNIa bulk flow is in red^[BL]. Hemispherical power asymmetry^[BJ] could be due to the local jet, where temperature and resolution variations are more extreme closer to the central region.

3.9. Baryon Asymmetry

Section (3.5) demonstrated that an expanding universe is incompatible with observations. The previous sections also provide sufficient evidence for an asymptotically flat universe with central core. It is therefore important to discuss the consequences of having an always existent, steady state universe in terms of entropy and stability. The concept of an absolute beginning of reference time is flawed. Reference time in this perspective is relative to a space-time with no vacuum energy. For example, it is known that metric distance becomes infinite as an event horizon is approached with respect to the preferred reference frame. Therefore, nothing can reach the surface boundary within a finite amount of reference time. Relative to proper time, an observer falling into an event horizon would do so in a finite amount. This is an illusion since the observer will cease to evolve as the surface is approached. In comparison to a photon, the problem can still be defined in terms of metric distance. Since the photon will be travelling along a null geodesic, it must travel an infinite amount of metric distance prior to reaching the surface.

It is possible to take the limit of the inferred state of an expanding universe as $t \rightarrow -\infty$. With either Einstein's field equations or vacuum field theory, matter will converge at a single region in space until it is infinitesimally close to forming an event horizon. With respect to the preferred reference frame, a discontinuity forms. For any event horizon, time is undefined due to infinite vacuum energy density. From the other direction, the collapsing system cannot form into a conical singularity within a finite amount of reference time. This discontinuity between time prior to a big bang scenario and formation of event horizon indicates that an initial singularity could have never existed with respect to the reference frame. On the other hand, one could argue that the universe began at some infinite time

ago as an object infinitesimally close to a conical singularity; i.e. the interpretation of accelerated expansion is still viable. If this were true, Hubble's law would remain valid for all redshift rather than just the local. In addition, this perspective requires galaxies to become older as redshift increases, which is contrary to several recent observations. The most important of these is the increase in cold baryonic matter with redshift. Even beyond this, an expanding theory requires several non-classical assumptions such as dark matter and dark energy. It fails to explain observations of galactic evolution, number densities or the size of distant galaxies and clusters. Therefore, Λ CDM or an expanding model should be abandoned.

An always-existent universe has other properties based upon boundary conditions. If at some time prior to the present the universe was unstable and energy could escape, the instability would have existed at some prior point in time. With respect to a quantum system, there will be a distribution of possible events occurring over a finite period. If the system has a finite probability distribution, then it is impossible for any instability to have not occurred prior to a finite time before present. In other words, any instability must have existed for an infinite period and according to probability, any finite distribution requires that the universe is stable at $t \rightarrow -\infty$. This also relates to the null existence of event horizon. For example, all observations agree with a central core existing in the present universe. If black holes had event horizon as predicted by EFEs, the universe would clearly not exist in the current state. As matter approaches a central event horizon, some would be captured while the rest is ejected. Over an infinite amount of reference time prior to the present, the central event horizon would capture all matter within the universe. The CMBR temperature further provides direct proof against the existence of event horizon and Hawking radiation.

In order for an always existent universe to be in its present form, the laws of thermodynamics must be missing something. The last piece of the puzzle was previously beyond comprehension due to the belief in an expanding universe and event horizon. Since all massive objects must have finite vacuum fields, the gravitational force aids in completing the thermodynamic cycle. The entire process can be viewed as beginning from the surface of the central core in the form of a dense relativistic jet. From this point to Earth's present position in the universe, the usual thermodynamic principles apply. Dense quark matter for example will decay into x-ray emitting gas. Radiation emitted over this transition follows geodesics back to the central core, ensuring the universe is stable. However, an asymptotically flat universe by definition will have finite vacuum energy density at all points in space. The only requirement is that the universe remains localized for an infinite period, i.e. it exists in a steady state. Beyond Earth's present position, population I stars are abundant due to increased metallicity and cold baryonic matter. The bulk flow continues to move towards entropy as galaxies gain momentum falling into the center of the universe. The missing piece is where matter falls back into the central core and momentum is conserved via two polar jets. The universe therefore exists in an anisotropic state of entropy as depicted by figures 3.5 and 3.7.

The laws of thermodynamics demand that the entropy of a closed system can never be reversed without external energy. From figure 3.5, it is clear that entropy is constant for all time relative to the preferred reference frame. Therefore, the laws of thermodynamics are not violated since the universe acts as a perpetual machine. The actual mechanism that creates relativistic jets is speculated upon with QCD and modern MHD simulations. Compared to the local region of space that Earth currently resides in, the jets emanating from the central core should

be extremely large. With respect to the amount of galaxies and clusters falling into the central core, it likely contains the mass of millions or billions of galaxies and clusters. The center of the core should therefore exist in a dense, color superconducting state. The central region of the core is assumed to consist of top, bottom and charm quarks due to sheer size. With respect to modern theory, compact stars are already predicted to exist in a non-CFL color superconducting state^[BF]. The layer directly adjacent to the core's central region will also likely exist in a superfluid state (CSL, Planar, A/Polar). The center could possibly rotate, further inducing a magnetic field from the London moment.

Comparing the bare mass of quarks, the core should have layers depicted by chemical potentials. For example, the up and down quarks exist within the 2 — 15 Mev range. The strange quark has a bare mass between 100 — 300 Mev, while the remaining quarks exist from 1000 Mev and beyond. Baryon asymmetry is speculated to originate from the strange quark/anti-quark layer. Considering that the core acts as a perpetual machine, the energy needed to produce the jets must be provided by the inflow alone. Due to the asymmetric shape of the universe, matter enters the equatorial regions and is funneled inwards. As density increases, the quarks at each radius become more massive until a strange diquark layer is reached. This type of condensate should be favored due to the gap between quark masses and single quark flavor. The CSL, planar or A/polar single flavor states also demonstrate superfluidity and the Meissner effect^[BE]. As matter approaches the center, the magnetic field energy density begins to surpass kinematic energy density^[AM]. At this point, it is expected that the inflow stops moving with the bulk of material, effectively producing a toroidal magnetic field^[AM]. Due to lorentz forces, the $\langle ss \rangle$ condensates are accelerated in one direction with $\langle s\bar{s} \rangle$ in the opposite.

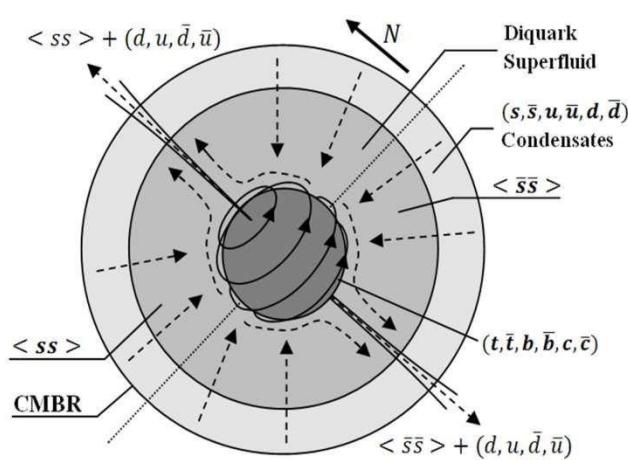
Putting things into perspective, the structure of the core should be similar to figure 3.28. After applying the methods discussed in section (3.3) to derive proper time, matter approaching the surface of the core will be traveling at 99.999988% the speed of light. In consideration of momentum conservation and the internal magnetic field, the rate of inflow alone is capable of explaining the energy behind each jet. The bulk of material ejected should originate from the strange diquark layer due to its insulating property. As the quark matter begins to cool into mesons and baryons, the jet pointing in the direction of T N will consist of K^0 , K^- , π^\pm , π^0 , p^+ , n^0 , A^0 , Σ^\pm , Σ^0 , $\bar{\nu}^-$ and $\bar{\nu}^0$; although not exclusively. These pairings occur naturally since baryons containing a mixture of quarks and anti-quarks have not been observed experimentally. However, the anti-quarks must still be paired with something external to the surface. The resulting mesons decay into photons, electrons and neutrinos; table 3.5 provides the most common decay modes.

The amount of material ejected from each shell of the core depicts the ratio of electrons and neutrinos to baryonic matter, although the composition of

The decay of unstable baryons in the northern hemisphere of the universe also results in products that are observed in abundance locally. Table 3.6 provides the common decay modes for the main constituents. This process creates an abundance of protons and electrons (p^+ , e^-) in the northern hemisphere, with an abundance of anti-matter (p^- , e^+) in the other. It is concluded that Earth currently resides in the northern hemisphere beyond the point where the relativistic proton, electron and neutrino gas has cooled. For any steady state model in the current universe, a perpetual machine and explanation for the CMBR are required. With the corrections to general relativity, event horizon are no longer possible. Therefore, an asymptotically flat universe containing a central core would emit a 3000 K black body spectrum as observed. The purposed configuration further explains the hot ring surrounding the central cold spot in the cleaned CMBR image, the mechanics behind a steady state and origin of the dark flow.

Table 3.5. Common decay modes of mesons and fermions^[BG].

Particle(s)	Decay Mode(s)	Particle(s)	Decay Mode(s)
each layer must also be known.		n^-	$(\mu^- + v_\mu),$ $v_\mu),$
		n^+	$(e^+ + v_e)$
		$(^-$	$(^-$
		u	u
		d	d
		$)$	$)$
			$+ v_e)$
			$-$
		n^0	$(\frac{uu - dd}{\sqrt{2}}) - \bar{\Delta}^0 + \gamma$
			$K^0 (ds)$
		$K^- (us)$	$(\mu^- + v_\mu),$ $(n^- \pm n^0),$ $(e^- + v_e + v_\mu)$
			μ^- $e^- + v_\mu +$
			$v_\mu e^+$
			$+ e^-$



**Table
3.6.**
Common
decay
modes
baryons
GJ.

Baryon**Decay
Mode(s)****Baryon****Decay
Mode(s)** p^+

Stable

 Σ^0 $A^0 +$

$n^0 (udd)$ $p^+ + e^- + \bar{\nu}_e$
 $\Sigma^- (dds)$ $n^0 + n^-$

$A^0 (uds)$ $(p^+ + n^-),$
 $\Sigma^+ (uus)$ $(p^+ + n^0),$
 $(n^- + n^-)$
 $(n^- + n^-)$

o o

o +

$\bar{\nu}^0 (uss)$ $A^0 + n^0$
 $\bar{\nu}^- (dss)$ $A^0 + n^-$

Figure 3.28. A simplified section of a central core that would induce baryon asymmetry. The focus is placed upon the strange diquark layer, which is theorized to exist due to the gap in mass between quarks. In a steady state model, the core must have a constant inflow and outflow. This process creates an abundance of (p^+, e^-) in one hemisphere and (p^-, e^+) in the other.

References

- [A] Edmund, Whittaker. “**A History of the Theories of Aether and Electricity from the Age of Descrates to the Close of the Nineteenth Century**”. Dublin, Ireland: Hodges, Figgis. 1910
- [B] Jackson, John. “**Classical Electrodynamics**”. New York: Wiley, 1999.
- [C] Griffiths, David. “**Introduction to Quantum Mechanics**”. Upper Saddle River: Pearson Prentice Hall, 2005.
- [D] Born, Max. “**The Statistical Interpretation of Quantum Mechanics**”. Nobel Lecture, December 11, 1954. Link ([^](#))
- [E] Dirac, Paul. “**Theory of Electrons and Positrons**”. Nobel Lecture, December 12, 1933. Link ([^](#))
- [F] Hestenes, David. “**The Zitterbewegung Interpretation of Quantum Mechanics**”. 1990. Link ([^](#))
- [G] Hestenes, David. “**Real Dirac Theory**”. 1996. Link ([^](#))
- [H] Pollock, M.D. “**The Dirac Equation in Curved Space-Time**”. March 11, 2010. Link ([^](#))
- [I] Gaberdiel, Matthias; Ridder, Aude. “**Quantum Field Theory II**”. July 15, 2011. Link ([^](#))
- [J] Ellis, John. “**Standard Model of Particle Physics**”. Link ([^](#))
- [K] Grünewald, Martin. “**The LEP Electroweak Working Group**”. Link ([^](#))
- [L] “**End of the Line for LEP**”. November 8, 2000. Link ([^](#))
- [M] “**God particle may not exist**”. December 6, 2001. Link ([^](#))
- [N] “**Higgs hunting at 144GeV**”. July 27, 2011. Link ([^](#))
- [O] “**Higgs signal sinks from view**”. August 22, 2011. Link ([^](#))
- [P] “**Detectors home in on Higgs boson**”. December 13, 2011. Link ([^](#))
- [Q] “**Physicists find new particle, but is it the Higgs?**”. July 3, 2012. Link ([^](#))
- [R] Mammadov, Gulmammad. “**Reissner-Nordstrom Metric**”. May 4, 2009. Link ([^](#))
- [S] Stoica, Ovidiu-Cristinel. “**Analytic Reissner-Nordstrom Singularity**”. April 19, 2012. Link ([^](#))
- [T] Taylor, Joseph. “**Binary Pulsars and Relativistic Gravity**”, Nobel Lecture, December 8, 1993. Link ([^](#)); Huse, Russell. “**The Discovery of the Binary Pulsar**”, Nobel Lecture, December 8, 1993. Link ([^](#))
- [U] Brady, Patrick. “**Astrophysical Sources of Gravitational Radiation**”. 1999. Link ([^](#))
- [V] Belczynski, Chris. “**Double Compact Object Mergers: Short- hard GRBs and Gravitational-wave Signals**”. 2007. Link ([^](#))
- [W] Dent, Thomas. “**Searching for Inspiring and Merging Binaries in LIGO-Virgo Data**”. 2011. Link ([^](#))
- [X] Belczynski, Chris. “**Double Black Holes: Recent Observations and Predictions**”. March 24, 2011. Link ([^](#))
- [Y] Brown, Duncan. “**Searches for Gravitational Waves from the Inspiral of Binary Neutron Stars and Black Holes**”. April 13, 2008. Link ([^](#))
- [Z] Gondek-Rosinska, Dorota. “**Perspective of Discovering Gravitational Waves from Astrophysical Sources**”. 2011. Link ([^](#))
- [AA] Nikhef, Jo van den Brand. “**Gravitational Wave Detection**”. June 3, 2009. Link ([^](#))
- [AB] Coward; Lilley; Howell; Burman; Blair. “**The ‘Probability Event Horizon’ for Neutron Star Merger Detection with Advanced LIGO**”. 2004. Link ([^](#))

[AC] Corvino; Ferrari; Marassi; Schneider. “**Compact Binaries Detection Rates from Gravitational Wave Interferometers: Comparison of Different Procedures**”. June 18, 2012. Link [\(▲\)](#)

[AD] Waldman, Samuel. “**Status of LIGO at the Start of the Fifth Science Run**”. June 2006. Link [\(▲\)](#)

[AE] The LIGO Scientific Collaboration and The Virgo Collaboration. “**Sensitivity Achieved by the LIGO and Virgo Gravitational Wave Detectors during LIGO's Sixth and Virgo's Second and Third Science Runs**”. March 15, 2012. Link [\(▲\)](#)

[AF] Rolland, L. “**The Status of Virgo**”. January 2009. Link [\(▲\)](#)

[AG] Fafone, Viviana. “**VIRGO: Where We Come From, Where We Are Going**”. June 7, 2011. Link [\(▲\)](#)

[AH] Strain, Ken. “**The Status of GEO600**”. July 2007. Link [\(▲\)](#)

[AF] Che, Wujun; Paul, Jean-Claude; Zhang, Xiaopeng. “**Lines of Curvature and Umbilical Points for Implicit Surfaces**”. April 16, 2007. Link [\(▲\)](#)

[AG] Kremer; Devecchi. “**Thermodynamics and Kinetic Theory of Relativistic Gases in 2-D Cosmological Models**”. February 7, 2002. Link [\(▲\)](#)

[AH] Belongie, Serge. “**Rodrigues' Rotation Formula**”. Mathworld, Wolfram. Link [\(▲\)](#)

[AI] Riess; Macri; Casertano; Lampeitl; Ferguson; Filippenko; Jha; Li; Chornock. “**A 3% Solution: Determining the Hubble Constant with the Hubble Space Telescope and Wide Field Camera**”. March 10, 2011. Link [\(▲\)](#)

[AJ] This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Link [\(▲\)](#)

[AK] Gamow, George. “**The Creation of the Universe**”. Mineola, N.Y: Dover Publications, 2004.

[AL] Hubble, Edwin. “**A Relation Between Distance and Radial Velocity Among Extra-Galactic Nebulae**”. January 17, 1929. Link [\(▲\)](#)

[AM] Mirabel, Felix; Rodriguez, Luis. “**Sources of Relativistic Jets in the Galaxy**”. February 4, 1999. Link [\(▲\)](#)

[AN] Riess; Strolger; Casertano; Ferguson; Mobasher; Gold; Challis; Filippenko; Jha; Li; Tonry; Foley; Kirshner; Dickinson; MacDonald; Eisenstein; Livio; Younger; Xu; Dahlen; Stern. “**New Hubble Space Telescope Discoveries of Type Ia Supernovae at $z \geq 1$: Narrowing Constraints on the Early Behavior of Dark Energy**”. December 20, 2006. Link [\(▲\)](#)

[AO] Jarosik; Bennett; Dunkley; Gold; Greason; Halpern; Hill; Hinshaw; Kogut; Komatsu; Larson; Limon; Meyer; Nolta; Odegard; Page; Smith; Spergel; Tucker; Weiland; Wollack; Wright. “**Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP1) Observations: Sky Maps, Systematic Errors, and Basic Results**”. January 26, 2010. Link [\(▲\)](#)

[AP] “**Encyclopedia of astronomy and astrophysics**”. Bristol Philadelphia London New York: Institute of Physics Pub. Nature Pub. Group, 2001.

[AQ] Tacconi; Genzel; Neri; Cox; Cooper; Shapiro; Bolatto; Bouché; Bournaud; Burkert; Combes; Comerford; Davis; Förster Schreiber; García-Burillo; Gracia-Carpio; Lutz;

Naab; Omont; Shapley; Sternberg; Weiner. “**High molecular gas fractions in normal massive star-forming galaxies in the young Universe**”. December 22, 2009. Link [\(▲\)](#)

- [AR] Driver, Simon. "Hubble Deep Fever: A faint galaxy diagnosis". February 26, 1998. Link ([^](#))
- [AS] Ohio University. "Discovery Of Giant X-Ray Disk Sheds Light On Elliptical Galaxies". December 19, 2002. Link ([^](#))
- [AT] Sarazin, Craig; "X-ray Emission from Elliptical Galaxies". December 5, 1996. Link ([^](#))
- [AU] Dekel; Birnboim; Engel; Freundlich; Goerdt; Mumcuoglu; Neistein; Pichon; Teyssier; Zinger. "Cold streams in early massive hot haloes as the main mode of galaxy formation". January 16, 2009. Link ([^](#))
- [AV] Harvard-Smithsonian Center for Astrophysics. "Strange new 'species' of ultra-red galaxy discovered". ScienceDaily, December 1, 2011. Web. 14 Jul. 2012. Link ([^](#))
- [AW] NASA. "Hubble Finds Hundreds of Young Galaxies in Early Universe". April 2006. Link ([^](#))
- [AX] Smithsonian Astrophysical Observatory. "Mysterious Red Galaxies". December 9, 2011. Link ([^](#))
- [AY] Marel, Roeland; Dokkum, Pieter. "Dynamic models of elliptical galaxies in z=0.5 clusters: I. Data-model comparison and evolution of galaxy rotation". November 17, 2006. Link ([^](#))
- [AZ] Kashlinsky; Atrio-Barandela; Kocevski; Ebeling. "A measurement of large-scale peculiar velocities of clusters of galaxies: technical details". February 4, 2009. Link ([^](#))
- [BA] Clayton, D. (1983). "Principles of Stellar Evolution and Nucleosynthesis". Chicago: University of Chicago Press.
- [BB] Lehnert; Nesvadba; Cuby; Swinbank; Morris; Clement; Evans; Bremer; Basa. "Spectroscopic Confirmation of a Galaxy at Redshift z=8.6". October 20, 2010. Link ([^](#))
- [BC] Fixsen. "The Temperature of the Cosmic Microwave Background". November 30, 2009. Link ([^](#))
- [BD] Images are credited to NASA/WMAP Science Team. "WMAP Calibration". Link ([^](#)). "Three-Year WMAP View of Early Universe". Link ([^](#)). "WMAP Resolves the Universe". Link ([^](#))
- [BE] Shovkovy, Igor. "Transport Properties of Stellar Quark Matter". September 25, 2008. Link ([^](#))
- [BF] Bowers, Jeffrey. "Color Superconducting Phases of Cold Dense Quark Matter". 1998. Link ([^](#))
- [BG] J. Beringer (Particle Data Group), J. Phys. D86, 010001 (2012). Link ([^](#))
- [BH] Totani, Tomonori; Yoshii, Yuzuru. "Does the Number Density of Elliptical Galaxies Change at z < 1?". May 20th, 1998. Link ([^](#))
- [BI] Bennett, C. L.; Hill, R. S.; Hinshaw, G.; Larson, D.; Smith, K. M.; Dunkley, J.; Gold, B.; Halpern, M.; Jarosik, N.; Kogut, A.; Komatsu, E.; Limon, M.; Meyer, S. S.; Nolta, M. R.; Odegard, N.; Page, L.; Spergel, D. N.; Tucker, G. S.; Weiland, J. L.; Wollack, E.; Wright, E. L. "Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Are There Cosmic Microwave Background Anomalies?". January 3rd, 2011. Link ([^](#))
- [BJ] Hoftuft, J.; Eriksen, H. K.; Banday, A. J.; Gorski, K. M.; Hansen, F. K.; Lilje, P. B.. "Increasing Evidence for Hemispherical Power Asymmetry in the Five-Year WMAP Data". April 20th, 2009. Link ([^](#))
- [BK] NASA/Goddard/A. Kashlinsky, et al. "Mysterious Cosmic 'Dark Flow' Tracked Deeper into Universe". Link ([^](#))
- [BL] Turnbull, Stephen J.; Hudson, Michael J.; Feldman, Hume A.; Hicken, Malcolm; Kirshner, Robert P.; Watkins, Richard. "Cosmic flows in the nearby universe from Type Ia Supernovae". November 7th, 2011. Link ([^](#))
- [BM] "The Nobel Prize in Physics 2011". Nobelprize.org. Accessed November 4th, 2012. Link ([^](#))
- [BN] Jarosik, N.; Bennett, C. L.; Dunkley, J.; Gold, B.; Greason, M. R.; Halpern, M.; Hill, R. S.; Hinshaw, G.; Kogut, A.; Komatsu, E.; Larson, D.; Limon, M.; Meyer, S. S.; Nolta, M. R.; Odegard, N.; Page, L.; Smith, K. M.; Spergel, D. N.; Tucker, G. S.; Weiland, J. L.; Wollack, E.; Wright, E. L.. "Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors and Basic Results". January 26th, 2010. Link ([^](#))
- [BO] Hogg, David. "Distance Measures in Cosmology". December 16th, 2000. Link ([^](#))
- [BP] Mahdavi, Andisheh; Hoekstra, Henk; Babul, Arif; Balam, David; Capak, Peter. "A Dark Core in Abell 520". February 10th, 2007. Link ([^](#))
- [BQ] NASA's Chandra X-ray Observatory. "NASA Finds Direct Proof of Dark Matter". August 21st, 2006. Link ([^](#))
- [BR] Lee, Joung hun; Komatsu, Eiichiro. "Bullet Cluster: A Challenge to Λ CDM Cosmology". May 22nd, 2010. Link ([^](#))
- [BS] Narayan, Ramesh; Bartelmann, Matthias. "Lectures on Gravitational Lensing". 1995. Link ([^](#))
- [BT] He, Ping; Zhang, Yuan-Zhong. "Modelling the Number Counts of Early-Type Galaxies by Pure Luminosity Evolution". February 17th, 1998. Link ([^](#))
- [BU] Colless, Matthew; Ellis, Richard; Taylor, Keith; Hook, Richard. "The LDSS Deep Redshift Survey". November 14th, 1989. Link ([^](#))
- [BV] Driver, Simon; Couch, Warrick. "The Inferred Redshift Distribution of the Faint Blue Galaxy Excess". July 20th, 1996. Link ([^](#))
- [BW] Roche, N.; Ratnatunga, K.; Griffiths, R. E.; Im, M. "Angular Sizes of the Faint Blue Galaxies". July 15th, 1996. Link ([^](#))
- [BX] Colless, M.; Schade, D.; Broadhurst, T. J.; Ellis, R. S. "High-Resolution Imaging of Faint Blue Galaxies". September 13th, 1993. Link ([^](#))
- [BY] Broadhurst, T. J.; Ellis, R. S.; Shanks, T. "The Durham/Anglo-Australian Telescope Faint Galaxy Redshift Survey". July 14th, 1988. Link ([^](#))
- [BZ] Glazebrook, K.; Ellis, R.; Colless, M.; Broadhurst, T.; Allington-Smith, J.; Tanvir, N. "A Faint Galaxy Redshift Survey to B = 24". March 8th, 1995. Link ([^](#))
- [CA] Loveday, Jon. "The Local Space Density of Dwarf Galaxies". May 19th, 1997. Link ([^](#))
- [CB] "Correlation Between Absolute Magnitude and Diameter". Link ([^](#)). HOLMBERG, E., Ark. Astr. 3, 387 = Uppsala astr. Obs. Medd., No. 148 (1964).
- [CC] Weijmans, Anne-Marie; Bower, Richard G.; Geach, James E.; Swinbank, A. Mark; Wilman, R. J.; Zeeuw, P. T. de; Morris, Simon L. "Dissecting the Lyman- α emission halo of LAB1". November 18th, 2009. Link ([^](#))
- [CD] Ouchi, Masami; Ono, Yoshiaki; Egami, Eiichi; Saito, Tomoki; Oguri, Masamune; McCarthy, Patrick J.; Farrah, Duncan; Kashikawa, Nobunari; Momcheva, Ivelina; Shimasaku, Kazuhiro; Nakanishi, Kouichiro; Furusawa, Hisanori; Akiyama, Masayuki; Dunlop, James S.; Mortier, Angela M. J.; Okamura, Sadanori; Hayashi, Masao; Cirasuolo, Michele; Dressler, Alan; Iye, Masanori; Jarvis, Matt. J.; Kodama, Tadayuki; Martin, Crystal L.; McLure, Ross J.; Ohta, Kouji; Yamada, Toru; Yoshida, Michitoshi. "Discovery of a Giant Ly- α Emitter Near the Reionization Epoch". February 21st, 2009. Link ([^](#))

[CE] Mehrtens, Nicola; Romer, A. Kathy; Hilton, Matt; Lloyd-Davies, E. J.; Miller, Christopher J.; Stanford, S. A.; Hosmer, Mark; Hoyle, Ben; Collins, Chris A.; Liddle, Andrew R.; Viana, Pedro T. P.; Nichol, Robert C.; Stott, John P.; Dubois, E. Naomi; Kay, Scott T.; Sahlén, Martin; Young, Owain; Short, C. J.; Christodoulou, L.; Watson, William A.; Davidson, Michael; Harrison, Craig D.; Baruah, Leon; Smith, Mathew; Burke, Claire; Mayers, Julian A.; Deadman, Paul-James; Rooney, Philip J.; Edmondson, Edward M.; West, Michael; Campbell, Heather C.; Edge, Alastair C.; Mann, Robert G.; Sabirli, Kivanc; Wake, David; Benoist, Christophe; da Costa, Luiz; Maia, Marcio A. G.; Ogando, Ricardo. **"The XMM Cluster Survey: optical analysis methodology and the first data release"**. June 2012.

[CF] The scientific results reported in this article are based in part on observations made by the Chandra X-ray Observatory

[CG] Hopkins, Philip F.; Croton, Darren; Bundy, Kevin; Khochfar, Sadegh; Bosch, Frank van den; Somerville, Rachel S.; Wetzel, Andrew; Keres, Dusan; Hernquist, Lars; Stewart, Kyle; Younger, Joshua D.; Genel, Shy; Ma, Chung-Pei. **"Mergers in Λ CDM: Uncertainties in Theoretical Predictions and Interpretations of the Merger Rate"**. June 2010. Link ([^](#))

[CH] Conselice, Christopher J.; Yang, Cui; Bluck, Asa F. L. **"The Structures of Distant Galaxies - III: The Merger History of over 20,000 Massive Galaxies at $z < 1.2$ "**. December 17th, 2008. Link ([^](#))

[CI] Lopez-Sanjuan, Carlos; Balcells, Marc; Perez-Gonzalez, Pablo G.; Barro, Guillermo; Garcia-Dabo, Cesar Enrique; Gallego, Jesus; Zamorano, Jaime. **"The Galaxy Major Merger Fraction to $z \sim 1$ "**. February 23rd, 2009. Link ([^](#))

[CJ] Lotz, Jennifer; Jonsson, Patrik; T.J. Cox; Croton, Darren; Primack, Joel; Somerville, Rachel; Stewart, Kyle. **"Astronomers Pin Down Galaxy Collision Rate"**. October 27th, 2011. Link ([^](#))

[CK] Toft, S.; Dokkum, P. van; Franx, M.; Thompson, R. I.; Illingworth, G. D.; Bouwens, R. J.; Kriek, M. **"Distant Red Galaxies in the Hubble Ultra Deep Field"**. November 16th, 2004. Link ([^](#))

[CL] Conselice, C. J.; Bluck, A.F.L.; Ravindranath, S.; Mortlock, A.; Koekemoer, A.; Buitrago, F.; Grützbauch, R.; Penny, S. **"The Tumultuous Formation of the Hubble Sequence at $z > 1$ Examined with HST/WFC3 Observations of the Hubble Ultra Deep Field"**. May 12th, 2011. Link ([^](#))

[CM] Conselice, Christopher J.; Bershadsky, Matthew A.; Dickinson, Mark; Papovich, Casey. **"A Direct Measurement of Major Galaxy Mergers at $z \leq 3$ "**. June 5th, 2003. Link ([^](#))

[CN] XENON100 Collaboration: Aprile, E.; Arisaka, K.; Arneodo, F.; Askin, A.; Baudis, L.; Behrens, A.; Bokeloh, K.; Brown, E.; Bruch, T.; Bruno, G.; Cardoso, J. M. R.; Chen, W.T.; Choi, B.; Cline, D.; Duchovni, E.; Fattori, S.; Ferella, A. D.; Gao, F.; Giboni, K.L.; Gross, E.; Kish, A.; Lam, C. W.; Lamblin, J.; Lang, R. F.; Levy, C.; Lim, K. E.; Lin, Q.; Lindemann, S.; Lindner, M.; Lopes, J. A. M.; Lung, K.; Undagoitia, T.; Marrodan; Mei, Y.; Fernandez, A. J. Melgarejo; Ni, K.; Oberlack, U.; Orrigo, S. E. A.; Pantic, E.; Persiani, R.; Plante, G.; Ribeiro, A. C. C.; Santorelli, R.; Santos, J. M. F. dos; Sartorelli, G.; Schumann, M.; Selvi, M.; Shagin, P.; Simgen, H.; Teymourian, A.; Thers, D.; Vitells, O.; Wang, H.; Weber, M.; Weinheimer, C. **"Dark Matter Results from 100 Live Days of XENON100 Data"**. September 7th, 2011. Link ([^](#))

[CO] XENON100 Collaboration. **"Dark Matter Results from 225 Live Days of XENON100 Data"**. July 25th, 2012. Link ([^](#))

[CP] Lotz, Jennifer M.; Jonsson, Patrik; Cox, T.J.; Primack, Joel R. **"Galaxy Merger Morphologies and Timescales from Simulations of Equal-Mass Gas-Rich Disk Mergers"**. May 8th, 2008. Link ([^](#))

[CQ] Lotz, Jennifer; Jonsson, Patrik; T.J. Cox; Croton, Darren; Primack, Joel; Somerville, Rachel; Stewart, Kyle. **"The Major and Minor Galaxy Merger Rates at $z < 1.5$ "**. August 10th, 2011. Link ([^](#))

[CR] Hsieh, B. C.; Yee, H. K. C.; Lin, H.; Gladders, M. D.; Gilbank, D. G. **"Pair Analysis of Field Galaxies from the Red-Sequence Cluster Survey"**. April 10th, 2008. Link ([^](#))

[CS] Xu, C. Kevin. **"NIR/Optical Selected Local Mergers - Spatial Density and sSFR Enhancement"**. May 4th, 2012. Link ([^](#))

[CT] Edwards, Louise O.V.; Fadda, Dario. **"A Multi-Wavelength Analysis of Spitzer Selected Coma Cluster Galaxies: Star Formation Rates and Masses"**. September 14th, 2011. Link ([^](#))

[CU] Churchill, Chris. **"Mg II Absorbers: An On-line Review 'Paper'"**. December 1999. Link ([^](#))

[CV] Evans, Jessica L.; Churchill, Christopher W.; Murphy, Michael T. **"The Redshift Distribution of Intervening Weak MgII Quasar Absorbers and a Curious Dependence on Quasar Luminosity"**. Draft: July 3rd, 2012. Link ([^](#))

[CW] Dietrich, M.; Hamann, F.; Appenzeller, I.; Vestergaard, M. **"FeII/MgII Emission-Line Ratio in High-Redshift Quasars"**. June 27th, 2003. Link ([^](#))

[CX] Rosa, Gisella De; Decarli, Roberto; Walter, Fabian; Fan, Xiaohui; Jiang, Linhua; Kurk, Jaron; Pasquali, Anna; Rix, Hans-Walter. **"Evidence for Non-Evolving FeII/MgII Ratios in Rapidly Accreting Z~6 QSOs"**. June 23rd, 2007. Link ([^](#))

[CY] Jiang, Linhua; Fan, Xiaohui; Vestergaard, Marianne; Kurk, Jaron D.; Walter, Fabian; Kelly, Brandon C.; Strauss, Michael A. **"Gemini Near-Infrared Spectroscopy of Luminous z~6 Quasars: Chemical Abundances, Black Hole Masses and MgII Absorption"**. July 11th, 2007. Link ([^](#))

[CZ] Maiolino, R.; Juarez, Y.; Mujica, R.; Nagar, N.; Oliva, E. **"Early Star Formation Traced by the Highest Redshift Quasars"**. September 8th, 2003. Link ([^](#))

[DA] Wei, Hao. **"Observational Constraints on Cosmological Models with the Updated Long Gamma-Ray Bursts"**. August 16th, 2010. Link ([^](#))

[DB] Cardone, V.F.; Capozziello, S.; Dainotti, M.G. **"An Updated Gamma Ray Bursts Hubble Diagram"**. July 27th, 2009. Link ([^](#))

[DC] Wei, Hao; Zhang, Shuang Nan. **"Reconstructing the Cosmic Expansion History up to Redshift $z = 6.29$ with the Calibrated Gamma-Ray Bursts"**. August 31st, 2009. Link ([^](#))

[DD] Collazzi, Andrew C.; Schaefer, Bradley E.; Goldstein, Adam; Preece, Robert D. **"A Significant Problem with Using the Amati Relation for Cosmological Purposes"**. December 19th, 2011. Link ([^](#))

18. Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program

Galactic stellar populations and planets^{Y,YY,YYY}

V. Zh. Adibekyan¹, S. G.
Sousa^{1,2}, N. C. Santos^{1,3}, E.
Delgado Mena¹, J. I. González
Hernández^{2,4}, G. Israelian^{2,4},

M.
Mayor⁵,
and G.
Khachat
ryan^{1,3}

¹
C
e
n
t
r
o
n
d
e

A
s
t
r
o
f
í
s
i
c
a
d
a

U
n
i
v
e
r
s
i
d
a
d
e
d
o

P

o
r
t
o
,
R
u
a
d
a
s
E
s
t
r
e
l
a
s
,
4
1
5
0
-
7
6
2
P
o
r
t
o
,
P
o
r
t
u
g
a
l
e
-
m
a
i
l
:
V
a
r
d
a
n
-
A
d
i
b
e
k
y
a
n
@
a
s
t
r
o
-
u
p
-p

² Instituto de Astrofísica de
Canarias, 38200 La Laguna,
Tenerife, Spain
³ Departamento de Física e
Astronomia, Faculdade de Ciências
da Universidade do Porto, Portugal
⁴ Departamento de Astrofísica,
Universidad de La Laguna, 38206
La Laguna, Tenerife, Spain
⁵

O
bs
er
va
to
ir
e
de
G
en
èv
e,
U
ni
ve
rsi
té
de
G
en
èv
e,
51
C
h.
de
s
M
ail
let
es
,,
12
90
Sa
uv
er
ny
,S
wi

t
z
e
r
l
a
n
d
R
e
c
e
i
v
e
d
1
2
A
p
r
i
1
2
0
1
2
/

A
c
c
e
p
t
e
d
1
0
J

ABSTRACT

Context. We performed a uniform and detailed abundance analysis of 12 refractory elements (Na, Mg, Al, Si, Ca, Ti, Cr, Ni, Co, Sc, Mn, and V) for a sample of 1111 FGK dwarf stars from the HARPS GTO planet search program. Of these stars, 109 are known to harbor giant planetary companions and 26 stars are exclusively hosting Neptunians and super-Earths.

Aims. The two main goals of this paper are to investigate whether there are any differences between the elemental abundance trends for stars of different stellar populations and to characterize the planet host and non-host samples in terms of their [X/H]. The extensive study of this sample, focused on the abundance differences between stars with and without planets will be presented in a parallel paper.

Methods. The equivalent widths of spectral lines were automatically measured from HARPS spectra with the ARES code. The abundances of the chemical elements were determined using an LTE abundance analysis relative to the Sun, with the 2010 revised version of the spectral synthesis code MOOG and a grid of Kurucz ATLAS9 atmospheres. To separate the Galactic stellar populations we applied both a purely kinematical approach and a chemical method.

Results. We found that the chemically separated (based on the Mg, Si, and Ti abundances) thin- and thick disks are also chemically disjunct for Al, Sc, Co, and Ca. Some bifurcation might also exist for Na, V, Ni, and Mn, but there is no clear boundary of their [X/Fe] ratios. We confirm that an overabundance in giant-planet host stars is clear for all studied elements. We also confirm that stars hosting only Neptunian-

like planets may be easier to detect around stars with similar metallicities than around non-planet hosts, although for some elements (particularly α-elements) the lower limit of [X/H] is very abrupt.

Key words. stars: abundances – planetary systems – stars: fundamental parameters – Galaxy:

disk – solar neighborhood – stars: kinematics and dynamics

1. Introduction

(ESO runs ID 72.C-0488, 082.C-0212,

High-precision radial velocity measurements resulted in the detection of the first extra-solar planetary system surrounding a main-sequence star similar to our own in 1995 (Mayor & Queloz 1995).

Observational progress in extra-solar planet detection and characterization is now moving rapidly on several fronts.

More than 750 planetary companions have already been found orbiting late-type stars¹. The total number of planet-harboring systems that are found using Doppler technique is approaching 500.

Figure A.1 is available in electronic form at

<http://www.aanda.org>

<http://exoplanet.eu/>

^x Based on observations collected at the La Silla Paranal Observatory, ESO (Chile) with the HARPS spectrograph at the 3.6-m telescope

A strong input for this number was made by several dedicated planet-search programs that systematically monitor the sky. Among these programs, the HARPS planet search program made a special contribution. The high spectral resolution and most importantly the long-term stability of the HARPS spectrograph (Mayor et al. 2003) allowed discovering a fairly large number of new planets, including the large majority of the known planets with masses near the mass of Neptune or below (e.g. Santos et al. 2004b; Lovis et al. 2006; Mayor et al. 2009, 2011).

Shortly after the discovery of the first extra-solar planet, Gonzalez (1998), based on a small sample of eight planet-host stars (PHS), suggested that PHSs tend to be metal-rich compared with the nearby field FGK stars that are known to host no-planet. The metal-rich nature of the PHSs have been confirmed in subsequent papers (e.g. Gonzalez et al. 2001; Santos et al. 2001, 2003, 2004a, 2005; Laws et al. 2003; Fischer & Valenti 2005; Gilli et al. 2006; Udry et al. 2006; Ecuillon et al. 2007; Sousa et al. 2008; Neves et al. 2009; Johnson et al. 2010; Kang et al. 2011). This tendency for giant planets that orbit

A
r
t
i
c
l
e
p
u
b
li
s
h
e
d
b
y
E
D
P
S
c
i
e
n
c
e
s

metal-rich stars strongly supports the core-accretion model of planet formation (e.g. Pollack et al. 1996). This implies that core accretion (Ida & Lin 2004; Mordasini et al. 2009) and not disk-instability (Boss 1997) is the main working mechanism for the formation of giant planets. Interestingly, recent studies show that Neptune and super-Earth-class planets may easier form in a low-metal-content environment (e.g. Udry et al. 2006; Sousa et al. 2008, 2011a; Ghezzi et al. 2010; Mayor et al. 2011; Buchhave et al. 2012).

Most spectroscopic studies are in general limited to small samples of a few hundred comparison stars and less than one hundred PHSs at most, and only a few studies have been based on samples as large as 1000 stars (e.g. Gazzano et al. 2010; Gazzano 2011; Petigura & Marcy 2011). In order to minimize the errors, one needs to have large and homogeneous samples with reliable measurements of their chemical features.

In this paper, we present a uniform spectroscopic analysis of 1111 FGK dwarfs observed within the context of the HARPS GTO planet search program. The paper is organized as follows: in Sect. 2, we introduce the sample used in this work. The method of the chemical abundance determination and analysis will be explained in Sect. 3. This section also includes discussion of the uncertainties and errors in our methodology as well as a comparison of our results with the literature. The calculation of the galactic space velocity data and the selection of different populations of stars, based on their kinematic and chemical properties, are presented in Sect. 4. A discussion of the [X/H] abundances of the exoplanet hosts can be found in Sect. 5. The main conclusions of the paper are finally addressed in Sect. 6. The extensive and full investigation of this sample, focused on the abundance difference between stars with and without planets will be presented in a parallel paper (Adibekyan et al. 2012).

2. Sample description and stellar parameters

The sample used in this work consists of 1111 FGK stars observed in the context of the HARPS GTO programs. It is a combination of three HARPS subsamples hereafter called HARPS-1 (Mayor et al. 2003), HARPS-2 (Lo Curto et al. 2010), and HARPS-4 (Santos et al. 2011). Note that the HARPS-2 planet search program is the complement of the previously started CORALIE survey (Udry et al. 2000) to fainter magnitudes and to a larger volume. The stars were selected to be suitable for radial velocity surveys. They are slowly rotating and non-evolved solar-type dwarfs with spectral type between F2 and M0 that do not show a high level of chromospheric activity either.

The individual spectra of each star were reduced using the HARPS pipeline and then combined with IRAF² after correcting for its radial velocity. The final spectra have a resolution of $R \sim 110\,000$ and signal-to-noise ratio (S/N) ranging from ~ 20 to ~ 2000 , depending on the amount and quality of the original spectra. Fifty-five percent of the spectra have an S/N higher than 200, about 16% of stars have an S/N lower than 100, and less than 1% of the stars have an S/N lower than 40.

Precise stellar parameters for the entire sample were deter-

mined by Sousa et al. (2008, 2011a,b) using the same spectra as we did for this study. We refer the reader to these papers for details. The authors used a set of FeI and FeII lines whose

equivalent widths (EW) were measured using the ARES³ code (automatic routine for line equivalent widths in stellar spectra – Sousa et al. 2007)⁴. Assuming ionization and excitation equilibrium, the parameters were derived through an iterative process until the slope of the relation between the abundances given by individual FeI lines and both the excitation potential (X_i) and reduced equivalent width ($\log EW/\lambda$) were zero, and until the FeI and FeII lines provided the same average abundance. The spectroscopic analysis was completed assuming local thermodynamic equilibrium (LTE) with a grid of Kurucz atmosphere models (Kurucz et al. 1993), and making use of a recent version of the MOOG⁵ radiative transfer code (Sneden 1973). The typical precision uncertainties for the atmospheric parameters are of about 30 K for T_{eff} , 0.06 dex for $\log g$, 0.08 km s^{-1} for ξ_t , and

0.03 dex for [Fe/H]. We note that there are four stars in common between HARPS-1 and HARPS-4, and 14 stars between the HARPS-2 and HARPS-4 subsamples.

The stars in the sample have derived effective temperatures from 4487 K to 7212 K, but very few stars have temperatures that are very different from those of the Sun (there are e.g. only 12 stars with $T_{\text{eff}} > 6500$ K). The metallicities of the stars range from -1.39 to 0.55 dex and have surface gravities from 2.68 to 4.96 dex (again there are very few “outliers”, only five stars with $\log g < 3.8$ dex).

As already noted before, HARPS has contributed very much to the present high number of known planetary systems. Recently, Mayor et al. (2011) reported on the results of an eight-year HARPS survey with a statistical analysis of the planet and host samples. Simultaneously, they presented the list of newly discovered planets. We included these data when we updated the original GTO (Guaranteed Time Observations) catalog using data from the extra-solar planets encyclopedia⁶. The total number of PHSs in the current sample is now 135, of which 26 are super-Earths and Neptune-mass (the mass of the heaviest planet is less than $30 M_{\oplus}$) planet hosts (hereafter NH).

3. Abundance analysis and uncertainties

Elemental abundances for 12 elements (Na, Mg, Al, Si, Ca, Ti, Cr, Ni, Co, Sc, Mn, and V) were determined using an LTE analysis with the Sun as reference point with the 2010 revised version of the spectral synthesis code MOOG (Sneden 1973) and a grid of Kurucz ATLAS9 plane-parallel model atmospheres (Kurucz et al. 1993). The reference abundances used in the analysis were taken from Anders & Grevesse (1989). The line list and atomic parameters of Neves et al. (2009) were used, adding the CaI line at $\lambda 5260.39$ (excitation energy of the lower energy level $X_1 = 2.52$, and oscillator strength $\log g f = -1.836$) and excluding five NiI lines ($\lambda 4811.99, \lambda 4946.04, \lambda 4995.66, \lambda 5392.33$, and $\lambda 5638.75$), two SiI lines ($\lambda 5517.54$ and $\lambda 5797.87$), two TiII lines ($\lambda 4657.20$ and $\lambda 4708.67$) and five TiI lines ($\lambda 4656.47, \lambda 5064.06, \lambda 5113.44, \lambda 5219.70$, and $\lambda 5490.16$). These lines were excluded because the [X/Fe] abundance ratios determined by them showed significant trends with effective temperature (see also Neves et al. 2009, for details of the lines selection). The EWs were automatically measured with the ARES code. The

³ The ARES code can be downloaded at <http://www.astro.up.pt/sousasag/ares>

⁴ The EWs of the lines for the entire sample is available at the CDS.

⁵ The source code of MOOG2010 can be downloaded at <http://www.as.utexas.edu/~chris/moog.html>

⁶ <http://exoplanet.eu/>

² IRAF is distributed by National Optical Astronomy Observatories, operated by the Association of Universities for Research in Astronomy, Inc., under contract with the National Science Foundation, USA.

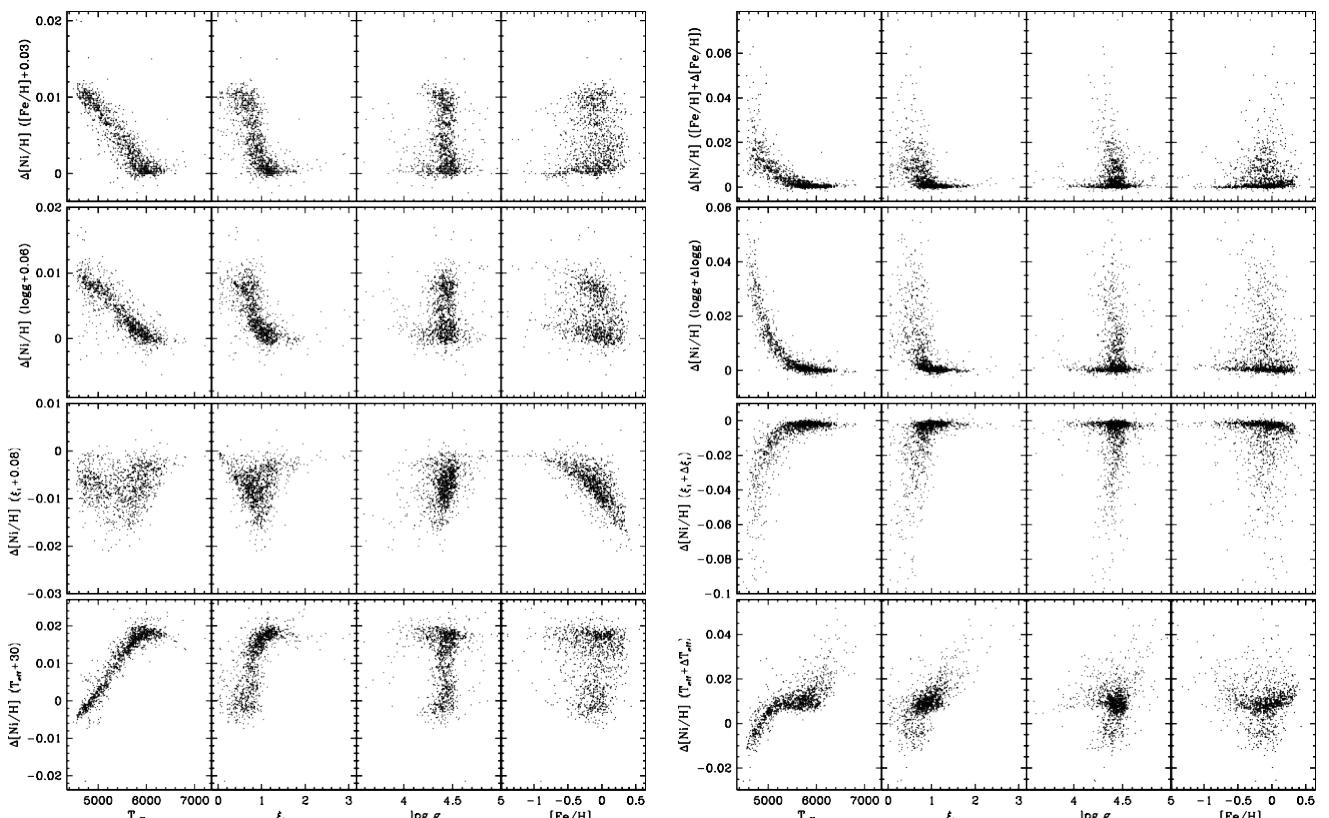


Fig. 1. Ni abundance sensitivity to the stellar parameter variations as a function of model atmosphere parameters. *Left* – The variation of the atmospheric parameters are the same for all stars and are equal to the typical errors. *Right* – The variation of the atmospheric parameters are equal to their one-sigma errors taken for each star individually.

input parameters for ARES were calculated following the procedure discussed in Sousa et al. (2011b).

The final abundance for each star and element was calculated to be the average value of the abundances given by all lines detected in a given star and element. Individual lines for a given star and element with a line dispersion more than a factor of two higher than the rms were excluded. In this way we avoided the errors caused by bad pixels, cosmic rays, or other unknown effects.

3.1. Uncertainties

Since the abundances were determined via the measurement of EWs and using already determined stellar parameters, the errors might still come from the EW measurements, from the errors in the atomic parameters, and from the uncertainties of the atmospheric parameters that were used to make an atmosphere model. In addition to the above-mentioned errors, one should add systematic errors that can occur due to NLTE or granulation (3D) effects. To minimize the errors, it is very important to use high-quality data and as many lines as possible for each element.

It is hard to define the contribution of each error source on the abundance results separately, but we can examine the sensitivity of the abundances to the stellar parameters and test the reliability of our results by comparing the abundances with those obtained in the literature.

First, to study the sensitivity trends of the abundances to the variation of the stellar parameters in general, we performed numerical tests with variations in the model parameters by a constant value similar to their typical errors: $\Delta T_{\text{eff}} = \pm 30$ K,

Table 1. Average of the stellar parameters for the subsamples with different T_{eff} .

	T_{eff} (K)	$\log g$ (dex)	[Fe/H] (dex)	ξ_t (km s $^{-1}$)
low T_{eff}	4934	4.4	-0.17	0.59
Solar	5769	4.39	-0.12	1.01
high T_{eff}	6440	4.51	-0.05	1.7

$\Delta \log g = \pm 0.06$ dex, $\Delta \xi_t = \pm 0.08$ km s $^{-1}$, and $\Delta[\text{Fe}/\text{H}] = \pm 0.03$ dex. Then we calculated the abundance differences between the values obtained with and without varying the parameter. The maxima from the plus and minus cases were taken. A thorough investigation of this experiment shows that the picture is quite complicated. Changing one of the parameters will increase or decrease the abundance of a certain element depending on the stellar parameters. For example, in Fig. 1 (left panel) one can see that a variation of T_{eff} by +30 K may change the Ni abundance from about +0.02 to -0.01 dex, depending on the T_{eff} of the stars. Despite the complex picture, it can be observed that in general, the sensitivity of all element abundances to the stellar parameters also depends on the effective temperature. Following this correlation, we grouped our sample stars into three temperature groups: “low T_{eff} ” stars – stars with $T_{\text{eff}} < 5277$ K, “solar” – stars with $T_{\text{eff}} = T_s \pm 500$ K, and “high T_{eff} ” – stars with $T_{\text{eff}} > 6277$ K. The average of the stellar parameters for the aforementioned groups are presented in Table 1.

The results obtained from the test for three groups of stars are displayed in Table 2. Table 2 shows that neutral species are generally more sensitive to changes in effective temperature. For

Table 2. Abundance sensitivities of the studied elements to changes of ± 100 K in T_{eff} , ± 0.2 dex in $\log g$ and [Fe/H], ± 0.5 km s $^{-1}$ in ξ_t .

	FeI	NaI	MgI	AlI	SiI	CaI	ScI	ScII
$\Delta T_{\text{eff}} = \pm 30$ K								
low T_{eff}	± 0.00	± 0.02	± 0.01	± 0.02	∓ 0.01	± 0.03	± 0.04	∓ 0.00
solar	± 0.01	± 0.01	± 0.01	± 0.01	± 0.00	± 0.02	± 0.02	± 0.00
high T_{eff}	± 0.01	± 0.02	± 0.02	± 0.00				
$\Delta [\text{Fe/H}] = \pm 0.03$ dex								
low T_{eff}	—	∓ 0.00	± 0.00	± 0.00	± 0.01	± 0.00	± 0.00	± 0.01
solar	—	± 0.00	± 0.01					
high T_{eff}	—	± 0.00						
$\Delta \log g = \pm 0.06$ dex								
low T_{eff}	∓ 0.01	∓ 0.01	∓ 0.01	∓ 0.01	± 0.01	∓ 0.02	∓ 0.01	± 0.02
solar	∓ 0.01	∓ 0.00	∓ 0.01	∓ 0.00	± 0.00	∓ 0.01	∓ 0.00	± 0.02
high T_{eff}	∓ 0.00	± 0.02						
$\Delta \xi_t = \pm 0.08$ km s $^{-1}$								
low T_{eff}	∓ 0.01	∓ 0.00	∓ 0.00	∓ 0.00	∓ 0.00	∓ 0.01	∓ 0.01	∓ 0.01
solar	∓ 0.01	∓ 0.00	∓ 0.00	∓ 0.00	∓ 0.00	∓ 0.01	∓ 0.00	∓ 0.01
high T_{eff}	∓ 0.00	∓ 0.01	∓ 0.00	∓ 0.01				
	TiI	TiII	VI	CrI	CrII	MnI	CoI	NiI
$\Delta T_{\text{eff}} = \pm 30$ K								
low T_{eff}	± 0.04	∓ 0.00	± 0.04	± 0.03	∓ 0.02	± 0.02	± 0.00	± 0.00
solar	± 0.03	± 0.00	± 0.03	± 0.02	∓ 0.00	± 0.02	± 0.02	± 0.01
high T_{eff}	± 0.02	± 0.00	± 0.02	± 0.02	∓ 0.00	± 0.02	± 0.02	± 0.02
$\Delta [\text{Fe/H}] = \pm 0.03$ dex								
low T_{eff}	± 0.00	± 0.01	± 0.00	± 0.00	± 0.01	± 0.01	± 0.01	± 0.01
solar	± 0.00	± 0.01	± 0.00					
high T_{eff}	± 0.00							
$\Delta \log g = \pm 0.06$ dex								
low T_{eff}	∓ 0.01	± 0.02	∓ 0.01	± 0.01	∓ 0.02	∓ 0.02	± 0.01	± 0.01
solar	∓ 0.00	± 0.02	∓ 0.00	± 0.00	∓ 0.02	∓ 0.01	± 0.00	± 0.00
high T_{eff}	∓ 0.00	± 0.02	∓ 0.00	± 0.00	∓ 0.02	∓ 0.00	∓ 0.00	± 0.00
$\Delta \xi_t = \pm 0.08$ km s $^{-1}$								
low T_{eff}	∓ 0.02	∓ 0.01	∓ 0.02	∓ 0.01				
solar	∓ 0.01	∓ 0.01	∓ 0.00	∓ 0.01	∓ 0.02	∓ 0.02	∓ 0.00	∓ 0.01
high T_{eff}	∓ 0.00	∓ 0.01	∓ 0.00	∓ 0.00	∓ 0.02	∓ 0.01	∓ 0.00	∓ 0.00

gravity variations, the neutral species were hardly affected, and the variations become noticeable only for stars with low T_{eff} , but the ionized species constantly varied by the same amount independently of the effective temperature. The ions are also more sensitive to metallicity changes than the neutral elements, although the sensitivity is not as significant as that for either T_{eff} and $\log g$. Finally, microturbulence variations led to only very small changes in most abundances (because many species are represented only by weak lines) and only few species are an exception.

Table 2 gives an overview of the elemental abundances variation with the variation of the stellar parameters, but not the uncertainties induced by the errors in the stellar parameters for our sample. The spectroscopic stellar parameters and metallicities were derived based on the equivalent widths of the FeI and FeII weak lines by imposing excitation and ionization equilibrium assuming LTE (e.g. Sousa et al. 2011b, and references therein). The errors obtained for the stars are typically very small, especially for stars similar to the Sun. This comes directly from the method itself because a differential analysis is performed with the Sun as reference. Stars that are significantly cooler or hotter than the Sun have larger intrinsic errors. To estimate the scale errors induced by uncertainties in the model atmosphere parameters, we varied the model parameters by an amount of their one-sigma errors available for each star and then we again divided our sample stars into three temperature groups as presented above. The average errors in the T_{eff} are 70, 24, and 45 K for cool, Sun-like, and hot star groups, respectively. The average

errors in $\log g$ are 0.15, 0.03, and 0.05, in ξ_t – 0.3, 0.04, and 0.08, and in [Fe/H] –0.04, 0.02, and 0.03 for the three groups, respectively. The right panel of Fig. 1 shows an example of the abundance variations with the variation of the stellar parameters against model parameters for Ni. From the figure it becomes clear that for stars with atmospheric parameters close to those of the Sun the uncertainties induced by the errors in the stellar parameters are very small (except for $\log g$). This is because both the abundance and stellar parameters are determined using an analysis with the Sun as reference point.

We evaluated the errors in the abundances of all elements [X/H], adding quadratically the line-to-line scatter errors and errors induced by uncertainties in the model atmosphere param-

eters. The line-to-line scatter errors were estimated as σ/\sqrt{N} , where σ is the standard deviation of N measurements (unfortunately, for some elements we were only able to select two or three lines). The average of σ/\sqrt{N} and [X/H] errors for the three grouped stars are presented in Table 3. The table shows that the σ/\sqrt{N} errors constitute the main part of the $\sigma[X/H]$ total errors

for the stars with $T_{\text{eff}} = T > 500$ K. The atmospheric parameters were obtained from the FeI and FeII lines by iterating until the correlation coefficients between $\log S(\text{FeI})$ and X_i , and between $\log S(\text{FeI})$ and $\log(W\lambda/\lambda)$ were zero, and the mean abundance given by FeI and FeII lines were the same (e.g. Santos et al. 2004a; Sousa et al. 2008). This means that the parameters are interrelated, i.e., variation of one parameter will influence others. Hence, the total error could be slightly higher due to the

Table 3. The average error for the element abundances [X/H], and abundance ratios [X/Fe].

Elem	Low T_{eff}			solar			High T_{eff}		
	\sqrt{N}	$\sigma[\text{X}/\text{H}]$	$\sigma[\text{X}/\text{Fe}]$	\sqrt{N}	$\sigma[\text{X}/\text{H}]$	$\sigma[\text{X}/\text{Fe}]$	\sqrt{N}	$\sigma[\text{X}/\text{H}]$	$\sigma[\text{X}/\text{Fe}]$
NaI	0.05	0.09	0.08	0.02	0.02	0.02	0.06	0.07	0.06
MgI	0.07	0.08	0.07	0.03	0.04	0.03	0.05	0.06	0.05
AlI	0.03	0.07	0.06	0.02	0.03	0.02	0.08	0.09	0.08
SiI	0.02	0.05	0.06	0.01	0.01	0.01	0.02	0.03	0.02
CaI	0.03	0.10	0.09	0.01	0.02	0.01	0.02	0.04	0.02
ScI	0.11	0.16	0.13	0.03	0.04	0.03	0.04	0.14	0.06
ScII	0.04	0.08	0.08	0.02	0.03	0.03	0.04	0.05	0.04
TiI	0.02	0.12	0.10	0.01	0.02	0.01	0.02	0.04	0.02
TiII	0.05	0.09	0.09	0.02	0.03	0.03	0.03	0.04	0.05
VI	0.07	0.16	0.14	0.02	0.03	0.02	0.05	0.07	0.05
CrI	0.02	0.08	0.07	0.01	0.02	0.01	0.02	0.03	0.02
CrII	0.07	0.11	0.10	0.03	0.03	0.03	0.03	0.05	0.05
MnI	0.05	0.10	0.08	0.03	0.04	0.03	0.04	0.05	0.04
CoI	0.03	0.06	0.04	0.02	0.03	0.02	0.05	0.06	0.05
NiI	0.01	0.04	0.03	0.00	0.01	0.01	0.01	0.03	0.02

described covariance terms (e.g. Johnson et al. 2002; Cayrel et al. 2004; Lai et al. 2008).

The errors in the abundance ratios, [X/Fe], were determined taking into account the differences between the sensitivities of the resulting abundance ratios to changes in the assumed atmospheric parameters and the dispersion of the abundances from individual lines of each X element. Table 2 shows that, in general, the model changes (variation of stellar parameters) induce similar effects in the abundances of different elements and Fe, so that they partially cancel out in the ratio [X/Fe]. The average error for the element abundances [X/H] and abundance ratios [X/Fe] are presented in Table 3.

3.2. Testing the validity of the stellar parameters

As stated before, the chemical abundances of the elements were derived by completing an LTE abundance analysis with the Sun as reference point using EW measurements. To check the validity limit of the adopted methodology in terms of stellar parameter ranges, and to test the stellar parameters themselves, we tested our results in a variety of ways. First we calculated the slopes of the derived abundances of the considered lines as a function of the excitation potential (EP) of the NiI lines. We chose nickel because its lines cover a wide range of EPs. In this way, we verified whether the excitation equilibrium enforced on the FeI lines of every star was applicable to other species. In Fig. 2, we plot the slopes of EP obtained for each star against the stellar parameters. The figure shows that there are no discernible trends of EP with $\log g$ and $[\text{Fe}/\text{H}]$, but there is a trend with T_{eff} and ξ_t (the ξ_t trend is just noticeable): cooler stars with $T_{\text{eff}} \leq 5000$ K, which also have low microturbulence velocities, have a systematic bias away from the expected values.

Then, in Fig. 3 we plot the [CrI/CrII] and [TiI/TiII] as a function of the stellar parameters to ensure that the ionization equilibrium enforced on the FeII lines (Sousa et al. 2008) is acceptable to other elements. The figure shows that the aforementioned ratios gradually increase with decreasing T_{eff} . Finally, plotting our abundance values of [X/Fe] as a function of the stellar parameters, we detect a significant trend for the T_{eff} plot, which is presented in Fig. 4. As seen in Fig. 4, Co and Al show a systematic trend with T_{eff} in all temperature ranges and TiI, ScI, V, CrII, and Na show a trend with T_{eff} in the low-temperature domain. The higher effective temperatures of the elements from

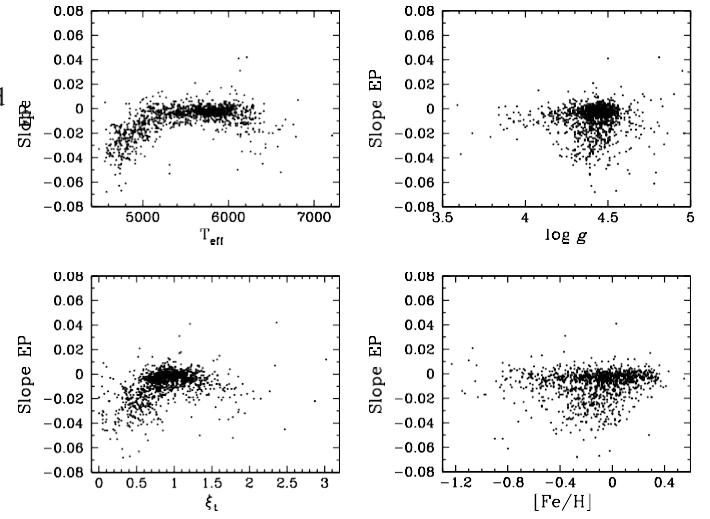


Fig. 2. Excitation potential slopes as a function of stellar atmospheric parameters for Ni.

which the trends appear are 4900 K for CrII, 5000 K for NaI and TiI, 5100 K for ScI and 5300 K for VI; these values are also indicated by vertical dotted lines in Fig. 4. As can be seen in Table 2, the elements and ions are very sensitive to the effective temperature, and the overestimation of the T_{eff} in the low-temperature domain might drift away from the expected abundance values. Similar trends for different elements with T_{eff} have been already noted in the literature (see e.g. Valenti & Fischer 2005; Preston et al. 2006; Gilli et al. 2006; Lai et al. 2008; Neves et al. 2009; Suda et al. 2011). As discussed in Neves et al. (2009), abundances of the cooler stars might have been overestimated due to the stronger line blending and also because the computed $\log g$ values may be inadequate for these stars. The unexpected trends may also be connected to either deviations from excitation or ionization equilibrium, or to problems associated with the differential analysis. Finally, a possible explanation for the observed trends with T_{eff} could be an incorrect T-T relationship in the adopted model atmospheres (Lai et al. 2008). While this effect on the derived [Fe/H] abundances can be compensated for by adjusting the value of the microturbulence, this does not apply to other elements.

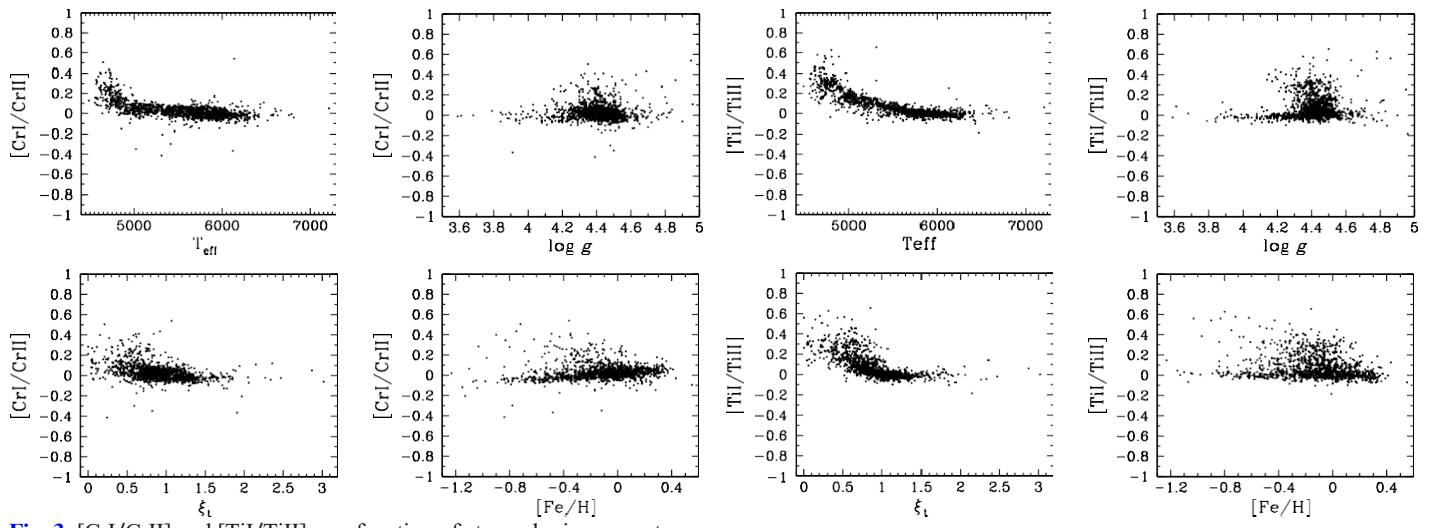


Fig. 3. $[CrI/CrII]$ and $[TiI/TiIII]$ as a function of atmospheric parameters.

This discussion indicates that, the observed trends are probably not an effect of stellar evolution, and uncertainties in atmospheric models are the dominant effect in measurements. Therefore, we chose to remove the T_{eff} trends for these elements. We fitted the data by a cubic polynomial and adding a constant

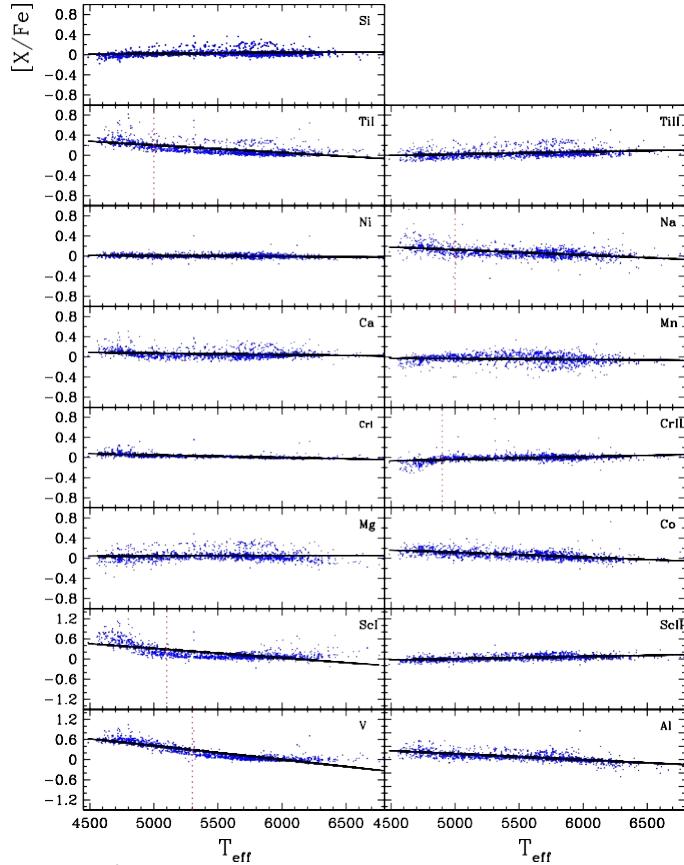


Fig. 4. $[X/Fe]$ vs. T_{eff} plots. The blue dots represent the stars of the sample. The black solid lines depict the linear fits of the data. The vertical purple dotted lines indicate the “cutoff” temperature when $[X/Fe]$ starts to show a systematic trend with T_{eff} . Each element is identified in the *upper right corner* of the respective plot.

term, chosen so that the correction is zero at solar temperature. The constant term was added because simply subtracting the cubic would force the mean [X/Fe] to zero, which is an unphysical situation. A similar approach has already been applied in previous studies (see e.g. Valenti & Fischer 2005; Petigura & Marcy 2011). A sample of our results for ten stars is presented in Table 4. We present the [X/H] values before and after correction for the T_{eff} trends. The complete results are available at the CDS.

3.3. Comparison with previous studies

As a final check of our method and analysis, we compare our derived abundances with those obtained by Bensby et al. (2005), Valenti & Fischer (2005), Gilli et al. (2006), and Takeda (2007) for stars in common with this paper. Although we have 451 stars in common with Neves et al. (2009) and 270 with Delgado Mena et al. (2010), we do not present a comparisons of the abundances, because the methods, atomic data, and the line list are almost the same. Very small differences observed for individual stars and elements during the comparison with these papers can be explained

with the small differences in the line list (see the beginning of Sect. 3) and moreover for some stars we used new spectra with higher S/N compared to those used in Neves et al. (2009). We note that the comparison was performed after removing the T_{eff} trends. The results are presented in Fig. 5. As can be seen, except for the paper by Gilli et al. (2006), our results agree very well with these previous studies which lends a certain reliability to our results. Figure 5 shows that there are systematic discrepancies with Gilli et al. (2006) for most of the elements. We note that Gilli et al. (2006) also observed systematic trends with T_{eff} for some at lower effective temperatures, but they did not correct their [X/Fe] abundance ratios. Our analysis shows that the higher discrepancies show stars with $T_{\text{eff}} < 5000$ K. Unfortunately, we do not have cool stars ($T_{\text{eff}} < 5000$ K) in common with Bensby et al. (2005), and Takeda (2007) to test an agreement (or disagreement) at low temperatures, but we have 15 cool stars in common with Valenti & Fischer (2005), whose abundance results agree very well with those achieved in this work. We note that Valenti & Fischer (2005) also observed abundance trends with T_{eff} for some elements, and as mentioned before, they chose to remove the spurious trends. The observed discrepancies with

Table 4. Sample table of the derived abundances of the elements, rms, and number of measured lines for each star.

Star	...	[TiI/H]	rms	n	[TiI/H] _{corr*}	[TiI/H]	rms	n	[VI/H]	rms	n	...
...
HD 109409	...	0.33	0.03	23	0.34	0.36	0.05	6	0.38	0.02	7	...
HD 109423	...	0.05	0.06	23	-0.06	-0.07	0.07	6	0.24	0.18	8	...
HD 109684	...	-0.27	0.05	24	-0.26	-0.24	0.03	6	-0.33	0.03	8	...
HD 109723	...	-0.01	0.06	25	-0.02	-0.10	0.03	6	0.00	0.06	8	...
HD 109988	...	0.23	0.05	23	0.14	0.15	0.07	6	0.53	0.20	8	...
HD 110291	...	0.03	0.05	23	-0.01	-0.03	0.05	6	0.08	0.07	8	...
HD 110557	...	0.04	0.04	23	-0.03	-0.05	0.04	6	0.17	0.15	8	...
HD 110619	...	-0.30	0.03	23	-0.31	-0.35	0.02	6	-0.35	0.01	8	...
HD 110668	...	0.16	0.05	23	0.16	0.22	0.01	5	0.22	0.04	8	...
HD 111031	...	0.30	0.03	24	0.30	0.29	0.03	6	0.35	0.02	7	...
...

Notes. (*) The [X/H] abundances after correction for the T_{eff} trends. The full table is available at the CDS.

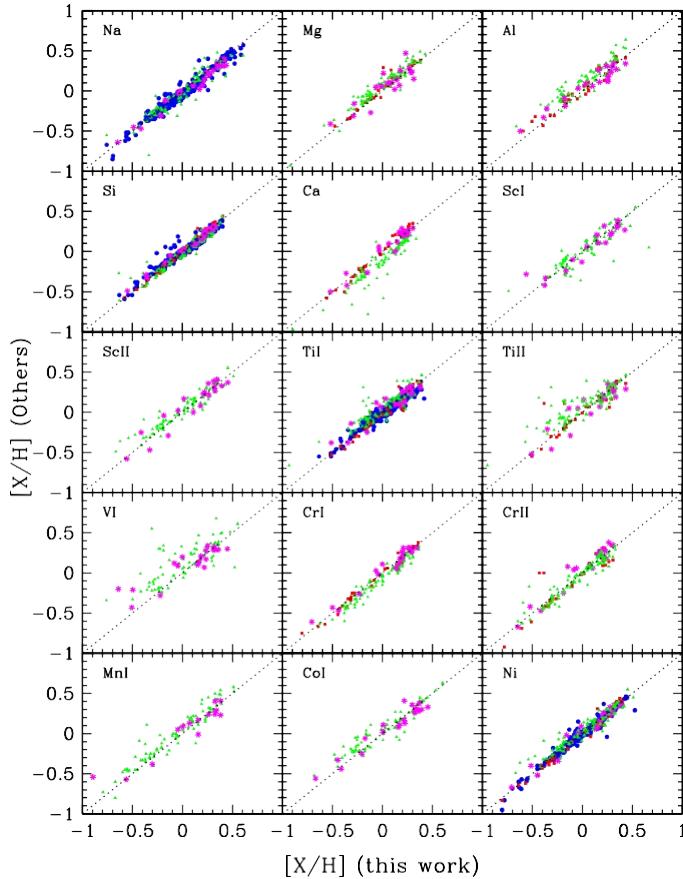


Fig. 5. Comparison of our abundance to those derived in other studies: Bensby et al. (2005) (red squares), Valenti & Fischer (2005) (blue dots), Gilli et al. (2006) (green triangles), and Takeda (2007) (magenta asterisks). The element label is located at the upper left corner of each plot.

Gilli et al. (2006) and at the same time perfect agreement with other studies confirm that applying corrections to remove the observed trends with T_{eff} is a correct approach.

4. Kinematics, chemistry, and stellar populations

The Milky Way (MW) has a composite structure with several stellar subsystems. The main three stellar populations of the MW in the solar neighborhood are the thin disk, thick disk, and the halo, although most of the stars belong to the thin disk.

These populations have different kinematic and chemical properties. Generally, the thick disk is composed of relatively old (e.g. Bensby et al. 2005; Adibekyan et al. 2011), metal-poor and α -enhanced (Fuhrmann 1998; Prochaska et al. 2000; Feltzing et al. 2003; Mishenina et al. 2004; Reddy et al. 2006; Haywood 2008b; Lee et al. 2011) stars that move in Galactic orbits with a large-scale height and long-scale length (Robin et al. 1996; Buser et al. 2001; Juric' et al. 2008). However, recent analyses of the geometric decompositions of the Galactic disk based on the elemental-abundance selection of the sample stars yielded strikingly different results (see e.g., Bovy et al. 2012a,b; Liu & van de Ven 2012). The latter authors found a chemo-orbital evidence that the thicker component of the MW disk is not distinct from the thin component (the MW has no thick disk – Bovy et al. 2012a), which can be explained by smooth internal evolution through radial migration (Liu & van de Ven 2012). The exceptions are the old metal-poor stars with different orbital properties that could be part of a distinct thick-disk component formed through an external mechanism (Liu & van de Ven 2012).

The first and important step in developing an understanding of the differences between the thin (thinner) and the thick (thicker) disks is to find an accurate and reliable method of assigning a star to a certain population. There is no obvious predetermined way to identify purely thick or thin disk stars in the solar neighborhood. The main essential ways of distinguishing local thick and thin disk stars are a purely kinematical approach (e.g. Bensby et al. 2003, hereafter B03, 2005; Reddy et al. 2006), a purely chemical method (e.g. Navarro et al. 2011; Adibekyan et al. 2011), and looking at a combination of kinematics, metallicities, and stellar ages (e.g. Fuhrmann 1998; Haywood 2008a).

Although the kinematic selection is a much more common method than the chemical approach, the chemical distinction of the disks can be more useful and reliable, because chemistry is a relatively more stable property of a star than the spatial positions and kinematics. In this section we present the adopted methods to separate stars into different stellar populations on the basis of their chemistry and kinematics.

4.1. Kinematical separation

To separate different stellar population by their kinematics, we computed Galactic space velocities for the stars. The space velocity components (UVW) were derived with respect to the local standard of rest, adopting the standard solar motion (U_S , V_S , W_S) = (11.1, 12.24, 7.25) km s⁻¹ of Schönrich et al. (2010). The main source of the parallaxes and proper motions were the

Table 5. Sample table of the Galactic space velocity components and the probabilities to assign the stellar population to which each star belongs.

Star	U_{LSR}	V_{LSR}	W_{LSR}	B03			R03		
				P_{thick}	P_{thin}	P_{halo}	group	P_{thick}	P_{thin}
...
HD 104800	122	-131	-57	0.91	0.00	0.09	thick	0.84	0.00
HD 104982	72	-10	-38	0.28	0.72	0.00	thin	0.13	0.87
HD 105004	-34	-225	-76	0.00	0.00	1.00	halo	0.06	0.00
HD 105671	-21	0	-8	0.01	0.99	0.00	thin	0.01	0.99
HD 105779	-29	-37	7	0.03	0.97	0.00	thin	0.02	0.98
HD 105837	23	13	41	0.15	0.85	0.00	thin	0.08	0.92
HD 105938	35	0	13	0.02	0.98	0.00	thin	0.01	0.99
HD 106116	-107	8	39	0.76	0.24	0.00	thick	0.29	0.71
HD 106275	18	-69	11	0.38	0.62	0.00	trans	0.11	0.89
HD 104006	-21	-188	1	0.48	0.00	0.52	trans	0.69	0.00
...

Notes. The full table is available at the CDS.

updated version of the Hipparcos catalog (van Leeuwen 2007). Data for eight stars with unavailable Hipparcos information were taken from the TYCHO Reference Catalog (Hog et al. 1998). The parallaxes with errors larger than 10%, (which is true for less than 5% of the stars in the sample) were redetermined following the procedure described in Sousa et al. (2011b). The percentage of stars with inaccurate proper motions (errors larger than 10%) is less than 8%. We did not perform a quality selection of them, because these errors in general do not change their membership to a certain population. The radial velocities were obtained from the HARPS spectra (courtesy of the HARPS GTO team). Combining the measurement errors in the parallaxes, proper motions, and radial velocities, the resulting average errors in the U , V , and W velocities are of about 1 km s⁻¹.

The selection of the thin disk, thick disk, and halo stars was completed using the method described in Reddy et al. (2006). This assumes that the sample is a mixture of the three populations and each population follows a Gaussian distribution of random velocities in each component (Schwarzschild 1907). Here, we adopted the mean values (asymmetric drift) and dispersion in the Gaussian distribution (characteristic velocity dispersion), and the population fractions were taken from B03 and Robin et al. (2003, hereafter R03; see also Ojha et al. 1996; Soubiran et al. 2003). We considered that a probability in excess of 70% suffices to assign a star to the concrete population. All remaining stars with a probability of less than 70% were included in a transition population. A sample of the probabilities calculated for each star according to B03 and R03, as well as Galactic space velocity components used in their calculation, are presented in Table 5. The complete results are available at the CDS.

According to the B03 criteria, among the 1111 stars in our sample, we have 964 stars from the thin disk, 78 from the thick disk, 58 are considered to be transition stars that do not belong to any group, and only 11 star belong to the halo. Adopting the criteria from R03 gives 1016 thin disk stars, 49 thick disk stars, 36 transition stars, and 10 stars belonging to the halo. We note that the B03 criteria approximately translate into the R03 criteria if $P_{\text{thick}} > 50\%$ for a star to belong to the thick disk (Reddy et al. 2006). The distribution of stars of our sample in the Toomre diagram is shown in Fig. 6 using both the R03 and B03 criteria.

4.2. Chemical separation

As mentioned above, in addition to the difference in their kinematics and ages, the thin- and thick disk stars are also different

in their α content at a given metallicity ([Fe/H]). This dichotomy in the chemical evolution allows one to separate different stellar populations.

Adibekyan et al. (2011) showed that the stars of our sample fall into two populations, clearly separated in terms of $[\alpha/\text{Fe}]$ (“ α ” refers to the average abundance of Mg, Si, and Ti) up to super-solar metallicities. We recall that Ca was not included in the α index, because at solar metallicities the $[\text{Ca}/\text{Fe}]$ trend differs from that of other α -elements. In turn, high- α stars were also separated into two families with a gap in both $[\alpha/\text{Fe}]$ ($[\alpha/\text{Fe}] \approx 0.17$) and metallicity ($[\text{Fe}/\text{H}] \approx -0.2$) distributions. This showed that the metal-rich high- α stars (h α m α r) and metal-poor high- α (thick disk) stars are on average older than chemically defined thin disk stars (low- α stars). At the same time h α m α r stars have kinematics and orbits similar to the thin disk stars.

Although in Adibekyan et al. (2011) we established a cutoff temperature for TiI because of the observed trend with T_{eff} for the [TiI/Fe] ratio (here we removed these trends, which are also observed for some other elements, see Sect. 3.2), the chemical separation of the stellar population was based on the stars with effective temperatures close to the Sun by ± 300 K. In this paper we used the chemical separation described in Adibekyan et al. (2011), i.e., thin disk, thick disk, and h α m α r stars.

The $[\alpha/\text{Fe}]$ versus $[\text{Fe}/\text{H}]$ plot for the sample stars is depicted in Fig. 7. The blue triangles refer to the thick disk, red circles to the thin disk. The green asterisks and the black crosses refer to the transition stars between thin-thick and thick-halo, respectively. Magenta squares represent the stars belonging to the halo. For the kinematical separation in the top panel we used the criteria from R03, and in the bottom panel the stars are separated according to the B03 criteria. The black dashed curve separates the stars with high- and low- α content. Clearly, the kinematically selected samples of thick- and thin disk stars are both well mixed, judging by their $[\alpha/\text{Fe}]$. The chemically separated thin disk contains several kinematically hot stars that are classified as thick disk stars. Using the R03 criteria almost “cleans” the thin disk from the kinematically selected thick disk stars, but produces a high “contamination” of the chemically selected thick disk by stars with thin disk kinematics. This mixing and contamination must in part result from the fact that the assignment to the thin or the thick disk is based on probability, but the main reason could be that the stars in the local neighborhood have different birth radii and reached the solar neighborhood because of their eccentric orbits or via radial migration (e.g. Haywood 2008b; Schönrich & Binney 2009).

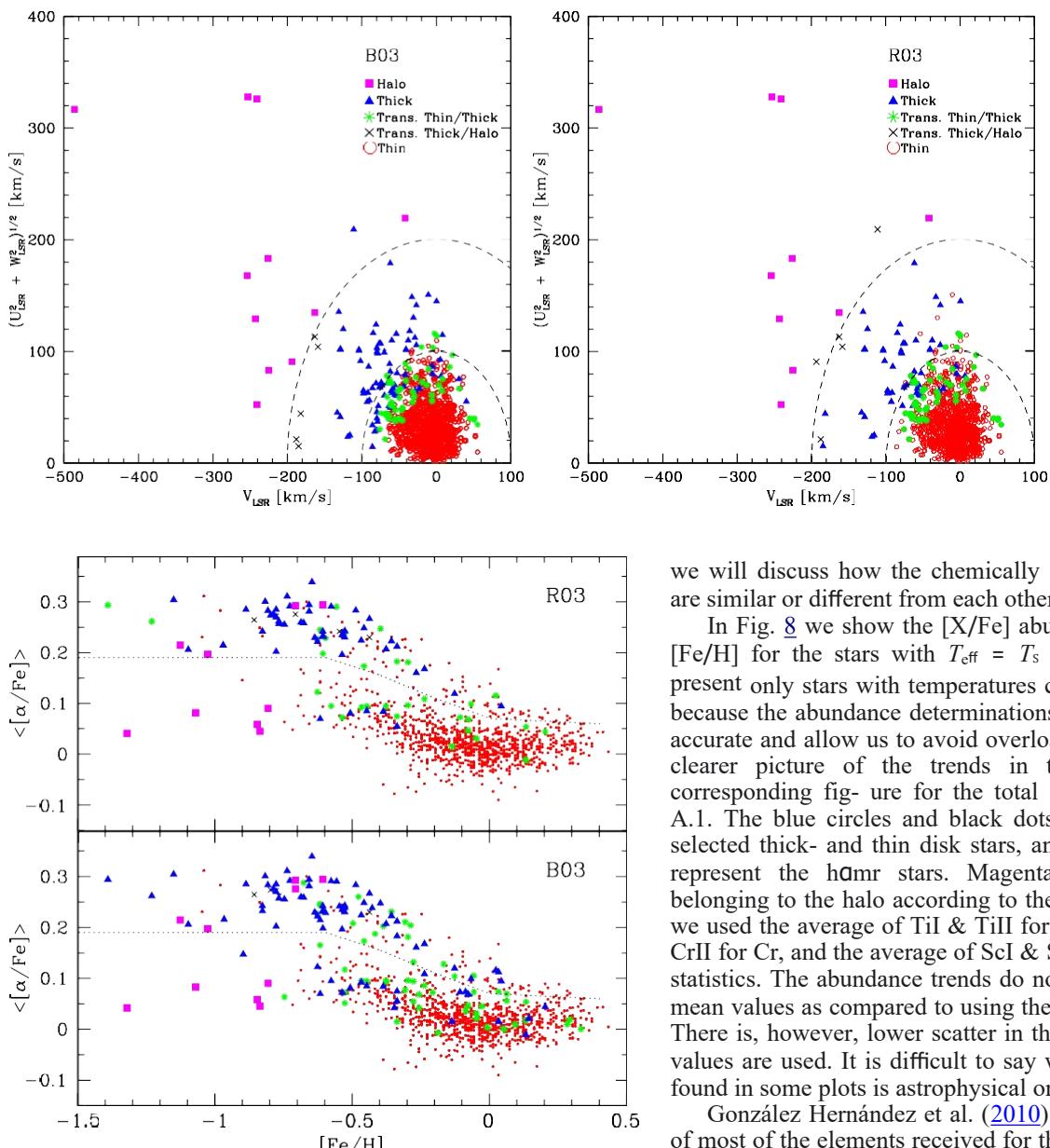


Fig. 7. Abundance ratios $\langle \alpha/\text{Fe} \rangle$ vs. $[\text{Fe}/\text{H}]$ for the total sample. The blue triangles refer to the thick disk, red circles to the thin disk. The green asterisks and the black crosses refer to the transition stars between thin-thick and thick-halo, respectively. Magenta squares represent the stars belonging to the halo. The black dashed curve separates the stars with high- and low- α content.

4.3. The $[\text{X}/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$: the thin- and thick disks

Low-mass stars have long lifetimes and their envelopes have preserved much of their original chemical composition. Studying FGK dwarfs is very useful because they contain information about the history of the evolution of chemical abundances in the Galaxy. The $[\text{X}/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ is traditionally used to study the Galactic chemical evolution because iron is a good chronological indicator of nucleosynthesis. In this paper we will not describe the $[\text{X}/\text{Fe}]$ abundance trends relative to Fe because they are discussed in Neves et al. (2009), whose sample consisted of about half the number of our stars. Neves et al. (2009) also performed a detailed analysis of the $[\text{X}/\text{Fe}]$ distributions of the kinematically separated stellar populations. In this subsection

Fig. 6. Toomre diagram for the entire sample. The left and right panels show the separation of the stellar groups according to the B03 and R03 criteria, respectively. The symbols are explained in the figure.

we will discuss how the chemically separated stellar families are similar or different from each other in terms of their $[\text{X}/\text{Fe}]$.

In Fig. 8 we show the $[\text{X}/\text{Fe}]$ abundance trends relative to $[\text{Fe}/\text{H}]$ for the stars with $T_{\text{eff}} = T_s \pm 300$ K. In the plot we present only stars with temperatures close to those of the Sun, because the abundance determinations for these stars are more accurate and allow us to avoid overloading the plot to obtain a clearer picture of the trends in the stellar groups. The corresponding figure for the total sample is shown in Fig. A.1. The blue circles and black dots refer to the chemically selected thick- and thin disk stars, and the red filled triangles represent the hQmr stars. Magenta squares are the stars belonging to the halo according to their kinematics. In the plot we used the average of TiI & TiII for Ti, the average of CrI & CrII for Cr, and the average of ScI & ScII for Sc to increase the statistics. The abundance trends do not change when using the mean values as compared to using the different ions separately. There is, however, lower scatter in the plots when the average values are used. It is difficult to say whether the heavy scatter found in some plots is astrophysical or due to errors.

González Hernández et al. (2010) noted the low dispersion of most of the elements received for their sample of solar twins and analogs with $S/N > 350$. To understand if the higher scatter found in this work is due to the quality of the data, we created a sample of solar analogs with the same stellar parameters as described in González Hernández et al. (2010). Then we divided the sample into two subsamples with $S/N > 400$ and $S/N < 150$. In general, we found similar dispersions for the two subsamples, comparable with those found in González Hernández et al. (2010).

Figure 8 shows that in addition to the Mg, Si, and Ti (on which our chemical separation is based), the thin- and thick disks are chemically different for Al, Sc, Co, and Ca. There are some hints that the two disks have different Na, V, Ni, and Mn ratios, but there is no clear boundary of their $[\text{X}/\text{Fe}]$ ratios. The only element for which the thin and the thick disks have the same $[\text{X}/\text{Fe}]$ values is Cr. A similar result was obtained in Neves et al. (2009), who separated the thin- and thick disks according to the kinematical features of the stars.

Inspection of Fig. 8 shows that the α -enhanced families, separated from the thick disk by the α -element and Fe content, show different $[\text{X}/\text{Fe}]$ trends with metallicity for different elements. As can be seen, at metallicities above solar the thin disk stars show a rise in the $[\text{Al}/\text{Fe}]$, $[\text{Sc}/\text{Fe}]$, $[\text{V}/\text{Fe}]$, $[\text{Ni}/\text{Fe}]$, $[\text{Co}/\text{Fe}]$, and

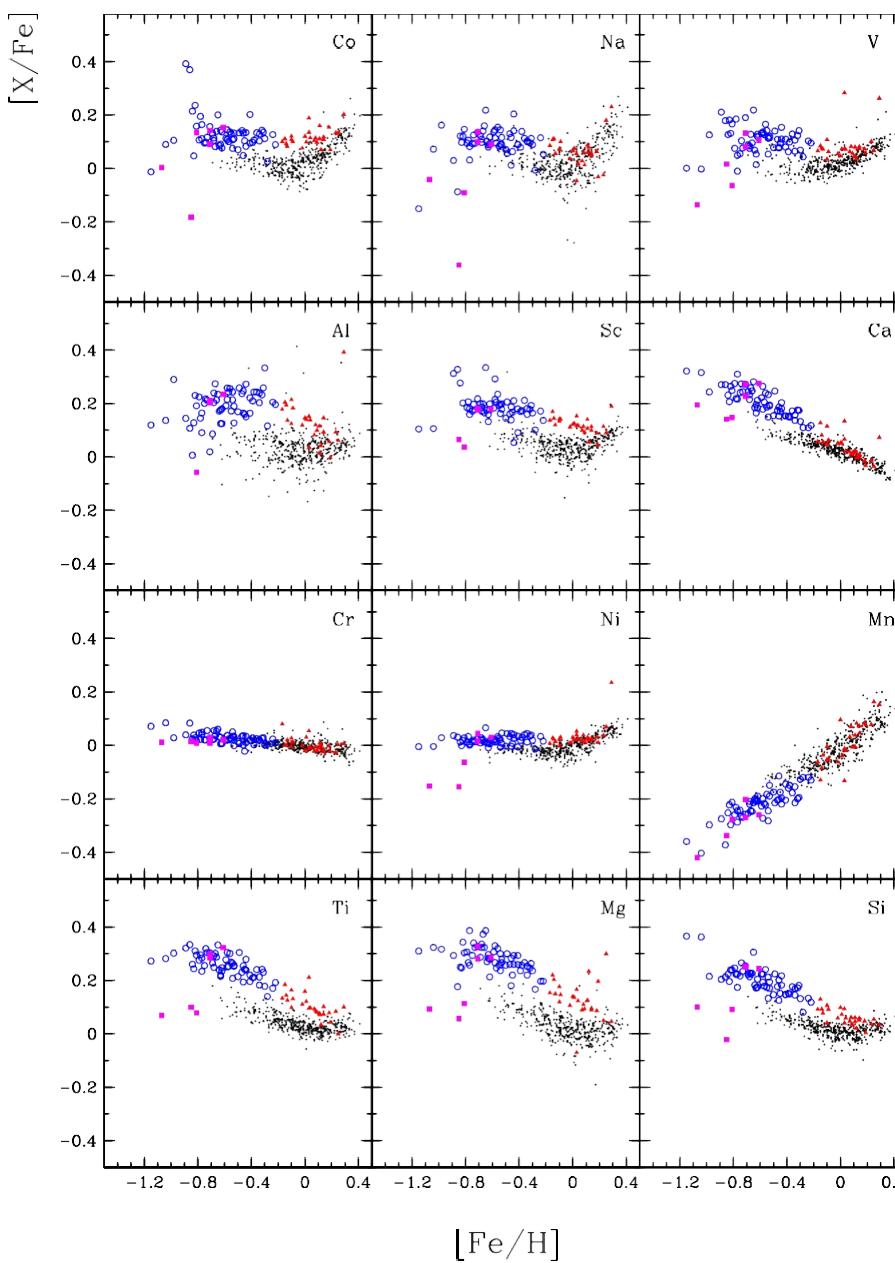


Fig. 8. Abundance ratios $[X/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ for the stars with $T_{\text{eff}} = T_s \pm 300$ K. The blue circles and black dots refer to the chemically selected thick- and thin disk stars, and the red filled triangles are the h α mr stars. Each element is identified in the *upper right corner* of the respective plot. Magenta squares represent the stars belonging to the halo according to their kinematics. The total sample is shown in Fig. A.1.

$[\text{Na}/\text{Fe}]$ (for the last two elements the rise is more pronounced and steeper), while the $[\text{Sc}/\text{Fe}]$, $[\text{Co}/\text{Fe}]$, $[\text{Ni}/\text{Fe}]$, and $[\text{V}/\text{Fe}]$ trends for the h α mr stars are essentially flat; moreover, for the $[\text{Na}/\text{Fe}]$ and $[\text{Al}/\text{Fe}]$ we observe a downward trend. It is interesting to see that the h α mr group stars are mixed with the thin disk stars in the $[\text{Ca}/\text{Fe}]$ vs. $[\text{Fe}/\text{H}]$ plot, while the thick- and thin disks are separated well.

Adibekyan et al. (2011), studying the orbital properties and α -element abundances of these stars, have put

forward the idea that this group of stars may have originated from the inner Galactic disk. Nevertheless, their origin and exact nature still remains to be clarified.

5. $[\text{X}/\text{H}]$ of planet-host stars

As stated before, in a separate paper we will focus on the the abundance differences between the stars with and without

planets. In this section we will briefly describe the sample of planet-host and non-host stars in terms of their [X/H].

The strong correlation is now well established between the rate of giant planets and host star metallicity. In turn, as noted before, recent studies showed that Neptune and super-Earth class planet hosts have a different metallicity distribution compared to those with giant gaseous planets. Although in this study we used relatively few PHSs (109 hosts of giants, and 26 hosts of only Neptune masses and below), this number is sufficient to observe whether there are

any discernible differences in the abundances of stars without planets and planets with different masses. The [X/H] distribution histograms for planet- and non-planet hosts are depicted in Fig. 9. The stars with giant planets, without planets, and the stars hosting exclusively Neptunes and super-Earths are represented by a dashed red, dotted black, and shaded blue line, respectively.

As expected, we observe a clear metallicity excess for Jovian hosts (JH) in all spaces, which agrees well with previ-

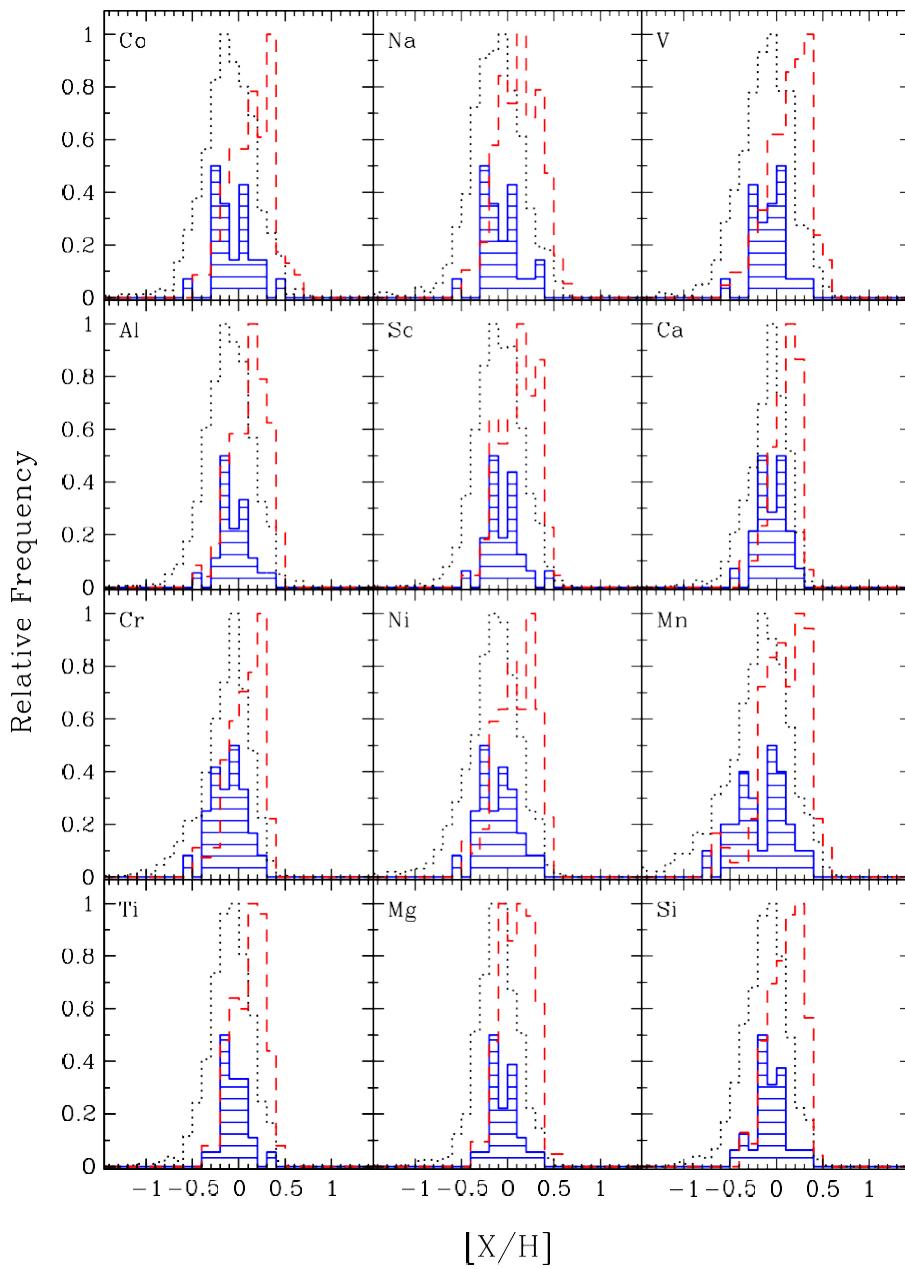


Fig. 9. $[X/H]$ distribution of the different elements. The stars with giant planets and without planets are represented by a red dashed and black dotted lines, respectively. The stars that exclusively host Neptunian and super-Earth planets are represented by a shaded blue. The Neptunian and super-Earth distribution was set smaller for clarity. The element label is located at the upper left corner of each plot.

ous similar studies for refractory elements (e.g. Bodaghee et al. 2003; Gilli et al. 2006; Takeda et al. 2007; Neves et al. 2009; Kang et al. 2011) and for iron (e.g. Gonzalez et al. 2001; Santos et al. 2001, 2003, 2004a, 2005; Fischer & Valenti 2005; Bond et al. 2006, 2008; Johnson et al. 2010). As already noted in the literature (e.g. Gilli et al. 2006; Neves et al. 2009), in most histograms the distributions of the abundances in JHs are not symmetrical: the distribution increases with $[X/H]$ to a maximum value and afterward abruptly drops. The observed cutoff might suggest that this is the metallicity limit of solar neighborhood stars (e.g. Santos et al. 2003), since most of the planet hosts are at the high-metallicity end of the sample.

As can be seen from Fig. 9, the $[X/H]$ distribution of 26 NHs in general repeats the distribution of stars without planets for all elements we studied (except that the distributions start very abruptly from the metal-poor side, probably indicating the minimum amount of some metals required to form them). This result confirms the “metal-poor” nature (e.g. Udry & Santos 2007; Sousa et al. 2008, 2011a; Mayor et al. 2011) of low-mass planet

hosts, when extended to elements other than iron. The average values of $[X/H]$ for three groups of stars, along with their rms dispersion, the number of stars used in their determination, and the difference of averages between Neptunian and Jovian hosts and stars without planets are listed in Table 6. These differences range from 0.17 (CaI) to 0.28 (MnI) for JHs and from 0.01 (CrII) to 0.09 (MgI) for NHs. These values agree well with those obtained by Neves et al. (2009) for the sample of 451 FGK stars.

Figure 10 illustrates the fraction of stars with Neptune-like and gaseous giant planets as a function of $[X/H]$. For each bin (the size of each bin is 0.1 dex), we divided the number of planet-bearing stars by the total number of stars in the bin. For all elements studied, we observe a continuous increase in the percentage of JHs as a function of increasing $[X/H]$. This result agrees with the previous findings of other authors e.g. Santos et al. (2001), Fischer & Valenti (2005), and Neves et al. (2009) for $[\text{Fe}/\text{H}]$ and Petigura & Marcy (2011) for $[\text{O}/\text{H}]$, $[\text{C}/\text{H}]$ and $[\text{Fe}/\text{H}]$. Petigura & Marcy (2011), noting the small-number statistics, reported a hint of possible plateau or turnover at the

Table 6. Average abundances [X/H] for stars without planets, with giant planets, and stars that exclusively host Neptunians, along with their rms dispersion, the number of stars used in their determination, and the difference of averages between Neptunian and Jovian hosts and stars without planets.

Species X	Jovian hosts			Neptunian hosts			Non-planet hosts			Difference of Averages	
	[X/H]]	σ	N	[X/H]]	σ	N	[X/H]]	σ	N	Jovian – Non-hosts	Neptunian – Non-hosts
NaI	0.12	0.23	109	-0.06	0.21	26	-0.12	0.30	975	0.24	0.06
MgI	0.10	0.18	109	-0.02	0.14	26	-0.11	0.23	976	0.21	0.09
All	0.10	0.19	109	-0.04	0.16	26	-0.11	0.25	969	0.21	0.07
SiI	0.10	0.18	109	-0.07	0.17	26	-0.12	0.24	976	0.22	0.05
CaI	0.08	0.15	109	-0.05	0.15	26	-0.09	0.21	976	0.17	0.04
ScI	0.14	0.21	109	-0.01	0.18	26	-0.08	0.25	947	0.22	0.07
ScII	0.11	0.20	109	-0.06	0.19	26	-0.12	0.27	976	0.23	0.06
TiI	0.11	0.17	109	-0.03	0.14	26	-0.08	0.22	976	0.19	0.05
TiII	0.10	0.18	109	-0.07	0.16	26	-0.11	0.23	976	0.21	0.04
VI	0.13	0.22	109	-0.06	0.19	26	-0.11	0.28	973	0.24	0.05
CrI	0.08	0.19	109	-0.09	0.18	26	-0.13	0.26	976	0.21	0.04
CrII	0.05	0.18	109	-0.14	0.18	26	-0.15	0.26	976	0.20	0.01
MnI	0.08	0.25	109	-0.17	0.27	26	-0.20	0.35	976	0.28	0.03
CoI	0.13	0.23	109	-0.06	0.20	26	-0.11	0.28	975	0.24	0.05
NiI	0.09	0.21	109	-0.10	0.21	26	-0.16	0.30	976	0.25	0.06
FeI	0.07	0.19	109	-0.12	0.2	26	-0.16	0.27	976	0.23	0.04
$\log g$	4.37	0.14	109	4.39	0.08	26	4.41	0.15	976	-0.04	-0.02
ξ	1.01	0.25	109	0.81	0.23	26	0.88	0.37	976	0.13	-0.07
T_{eff}	5656	412	109	5442	359	26	5490	502	976	166	-48

Notes. The four bottom rows list the average stellar parameters of the three aforementioned groups, taken from Sousa et al. (2008, 2011a,b).

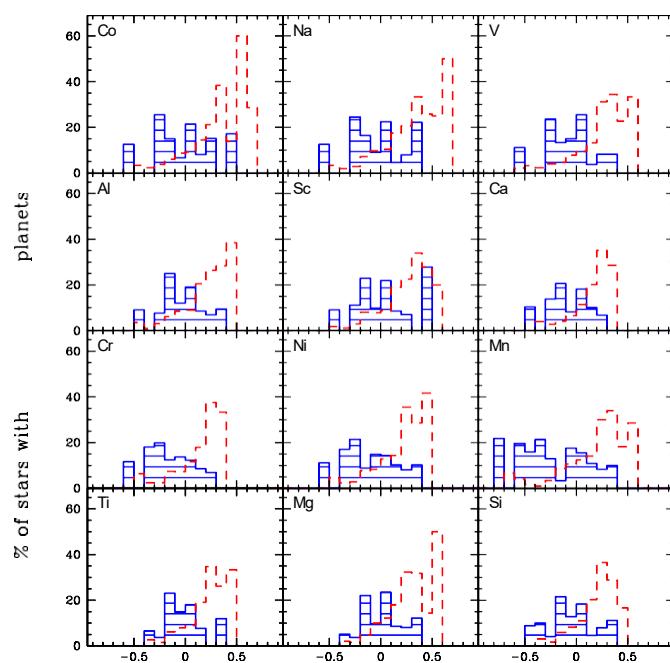


Fig. 10. Percentage of stars with giant (dashed red) and exclusively Neptunian and super-Earth (shaded blue) planets as a function of [X/H]. The Neptunian and super-Earth distribution was multiplied by 5 for clarity. Each element is identified in the upper left corner of the respective plot.

highest abundance bins for [C/H] and [Fe/H]. For our sample stars it is also possible to observe a small plateau or even turnover for some elements (Si, Ca, Sc, V, Cr, and Mn), but we also should note that at the highest abundance bins the number of stars sometimes does not exceed 4–5 stars.

For the fraction of low-mass planets hosts we do not observe any increasing or decreasing trends with [X/H] abundances. The distributions of the percentage of NHs are in general symmetric around the mean values listed in Table 6, which are on average less than solar abundance values by about 0.05 dex. These observations agree with the previous results for [Fe/H] (e.g. Sousa et al. 2008; Ghezzi et al. 2010; Mayor et al. 2011).

When we consider the possible dependence of planet formation on chemical composition, Gonzalez (2009) recommended to use a so-called refractory index “Ref”, which quantifies the mass abundances of refractory elements (Mg, Si and Fe) important for planet formation, rather than [Fe/H]. The importance of this index increases in the Fe-poor region, when one compares statistics of planets around the thin disk and thick disk stars. The left panel of Fig. 11 illustrates the [Fe/H] and [Ref/H] distribution histograms for planet and non-planet host stars. The fraction of stars with planets of different mass as a function of [Fe/H] and [Ref/H] are presented in the right panels. Clearly, the distributions of the three subsamples are shifted toward the higher “metallicities” in the [Ref/H] histograms, compared to their distributions in the [Fe/H]. This shift in the redistribution for planet-host stars is higher at lower metallicities, indicating their high [a/Fe] values. We again observe turnover at the highest abundance bins for [Fe/H] and [Ref/H].

The four bottom rows in Table 6 list the average stellar parameters of the three groups. It shows that hosts of low-mass planets on average have the same effective temperature as non-host stars. Interestingly, JJs are hotter by about 170 K than their non-host counterparts. The planet-search surveys are usually based on volume-limited samples, but the criteria to “cut” the sample were usually also based on the $B - V$ color. Our sample stars mostly have $B - V$ colors from 0.5 to 1.2. The top panel in Fig. 12 shows our sample stars in the [Fe/H] against T_{eff} (note that one star with $T_{\text{eff}} = 7212$ K is not presented in the plot). The dotted lines represent the approximate lower and upper limits in $B - V$ ($B - V = 0.5$ and 1.2). The lines were constructed using

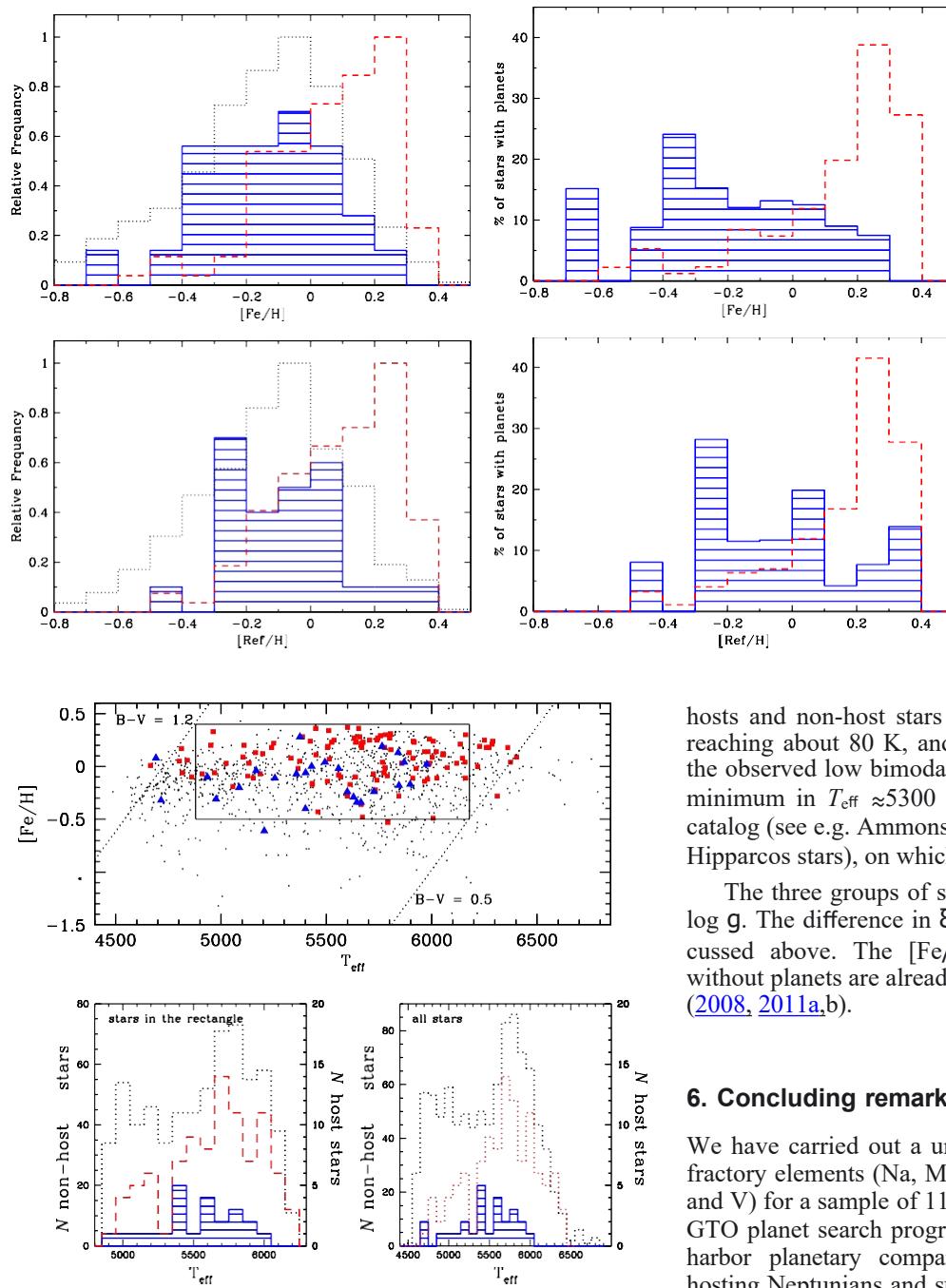


Fig. 12. Metallicity as a function of the effective temperature for stars with Jupiters (red circles), with Neptunes (blue triangles) and comparison sample stars (black crosses). The dotted line represents the approximate lower and upper limits in $B-V$ ($B-V = 0.5$ and 1.2). The T_{eff} distributions of all sample stars without planets (black dotted) and stars hosting Jovians (red dashed) and only Neptunians (shaded blue) are presented in the right bottom panel, and the distributions of the stars in the rectangle are shown in the left bottom.

the calibration equation from Sousa et al. (2008). Evidently, we missed stars with “low” $[Fe/H]$ and “high” T_{eff} in our sample, as well as “high” $[Fe/H]$ objects with “low” T_{eff} . To avoid these biases in $[Fe/H]$ and T_{eff} , we cut our sample in $[Fe/H]$ and in T_{eff} , as shown in Fig. 12. The T_{eff} distributions of all sample stars without planets and stars hosting Jovians and only Neptunians are presented in the right bottom panel of Fig. 12, and the distributions of the same groups of stars lying in the “cut rectangle” are shown in the left. The difference of average T_{eff} s of Jupiter

Fig. 11. *Left* – $[Fe/H]$ and $[Ref/H]$ distribution of the sample stars. The distribution lines for Jovian, Neptunian/super-Earth host and non-host stars are the same as in Fig. 9. The Neptunian and super-Earth distribution was set smaller for clarity. *Right* – Percentage of stars with giant (red dashed) and exclusively Neptunian and super-Earth (shaded blue) planets as a function of $[Fe/H]$ and $[Ref/H]$. The Neptunian and super-Earth distribution was multiplied by 5 for better visibility.

hosts and non-host stars in the rectangle has now decreased, reaching about 80 K, and for NHs about 50 K. We note that the observed low bimodality in T_{eff} for all three groups (with a minimum in $T_{\text{eff}} \approx 5300$ K) are inherited from the Hipparcos catalog (see e.g. Ammons et al. 2006, for the T_{eff} distribution of Hipparcos stars), on which the HARPS sample is based.

The three groups of stars have on average almost the same log g . The difference in ξ_t reflects the difference in T_{eff} , as discussed above. The $[Fe/H]$ distributions of stars with and without planets are already extensively discussed in Sousa et al. (2008, 2011a,b).

6. Concluding remarks

We have carried out a uniform abundance analysis for 12 refractory elements (Na, Mg, Al, Si, Ca, Ti, Cr, Ni, Co, Sc, Mn, and V) for a sample of 1111 FGK dwarf stars from the HARPS GTO planet search program. Of these stars, 135 are known to harbor planetary companions (26 of them are exclusively hosting Neptunians and super-Earth planets) and the remaining 976 stars do not have any known orbiting planet. The precise spectroscopic parameters for the entire sample were derived by Sousa et al. (2008, 2011a,b) in the same manner and from the same spectra as were used in the present study.

We discussed the possible sources of uncertainties and errors in our methodology in detail, and also we compared our results with those presented in other works to ensure consistency and reliability in our analysis. The large size of our sample allowed us to characterize and remove systematic abundance trends for some elements with T_{eff} .

To separate Galactic stellar populations, we applied both purely kinematical approach and chemical method. We showed that both kinematically selected thin- and thick disks are “contaminated”. The main reason of this “contamination” could be the fact that the stars in the local neighborhood have different birth radii and reached the Solar Neighborhood due to their eccentric orbits or via radial migration (e.g. Schonrich & Binney 2009).

Inspection of [X/Fe] against [Fe/H] plots suggests us that chemically separated thin- and thick disks, in addition to the Mg, Si, and Ti, are also different for Al, Sc, Co, and Ca. Some bifurcation might also exist for Na, V, Ni, and Mn, but there is no clear boundary of their [X/Fe] ratios. We observed no abundance difference between the thin- and thick disks for chromium. We found that the metal-poor α-enhanced stars and their metal-rich counterparts show different [X/Fe] trends with metallicity for different elements.

We confirmed that an overabundance in giant-planet host stars is clear for all studied elements, which lends strong support to the core-accretion model of planet formation (e.g. Pollack et al. 1996). We also confirmed that stars hosting only Neptunian-like planets may be easier to detect around stars with similar metallicity than non-planet hosts, although for some elements (particularly α-elements) we observed an abrupt lower limit of [X/H], which may indicate that these elements are important in their formation. The maximum abundance difference between Neptunian-like planet hosts and non-host stars is observed for Mg ([Mg/H] ≈ 0.09 dex).

Acknowledgements. This work was supported by the European Research Council/European Community under the FP7 through Starting Grant agreement number 239953. N.C.S. also acknowledges the support from Fundação para a Ciência e a Tecnologia (FCT) through program Ciência 2007 funded by FCT/MCTES (Portugal) and POPH/FSE (EC), and in the form of grant reference PTDC/CTE-AST/098528/2008. V.Zh.A., S.G.S. and E.D.M are supported by grants SFRH/BPD/70574/2010, SFRH/BPD/47611/2008 and SFRH/BPD/76606/2011 from FCT (Portugal), respectively. J.I.G.H. and G.I. acknowledge financial support from the Spanish Ministry project MICINN AYA2011-29060 and J.I.G.H. also from the Spanish Ministry of Science and Innovation (MICINN) under the 2009 Juan de la Cierva Programme. G.Kh. would like to acknowledge the support from the CAUP-11/2011-BI fellowship. We thank the anonymous referee for its useful comments and Astrid Peter for the help concerning English.

References

- Adibekyan, V. Zh., Santos, N. C., Sousa, S. G., & Israelián, G. 2011, A&A, 535, L11
- Adibekyan, V. Zh., Santos, N. C., Sousa, S. G., et al. 2012, A&A, 543, A89
- Ammons, S. M., Robinson, S. E., Strader, J., et al. 2006, ApJ, 638, 1004
- Anders, E., & Grevesse, N. 1989, Geochim. Cosmochim. Acta, 53, 197
- Bensby, T., Feltzing, S., & Lundström, I. 2003, A&A, 410, 527
- Bensby, T., Feltzing, S., Lundström, I., & Ilyin, I. 2005, A&A, 433, 185
- Bodaghee, A., Santos, N. C., Israelián, G., & Mayor, M. 2003, A&A, 404, 715
- Bond, J. C., Tinney, C. G., Butler, R. P., et al. 2006, MNRAS, 370, 163
- Bond, J. C., Lauretta, D. S., Tinney, C. G., et al. 2008, ApJ, 682, 1234
- Boss, A. P. 1997, Science, 276, 1836
- Bovy, J., Rix, H.-W., Liu, C., et al. 2012a, ApJ, 753, 148
- Bovy, J., Rix, H.-W., & Hogg, D. W. 2012b, ApJ, 751, 131
- Buchhave, L., Latham, D. W., Johansen, A., et al. 2012, Nature, 486, 375
- Buser, R., Rong, J., & Karaali, S. 1999, A&A, 348, 98
- Cayrel, R., Depagne, E., Spite, M., et al. 2004, A&A, 416, 1117
- Delgado Mena, E., Israelián, G., González Hernández, J. I., et al. 2010, ApJ, 725, 2349
- Ecuivillon, A., Israelián, G., Pont, F., Santos, N. C., & Mayor, M. 2007, A&A, 461, 171
- Feltzing, S., Bensby, T., & Lundström, I. 2003, A&A, 397, 1
- Fischer, D. A., & Valenti, J. 2005, ApJ, 622, 1102
- Fuhrmann, K. 1998, A&A, 338, 161
- Gazzano, J. 2011, Ph.D. Thesis, Marseille
- Gazzano, J., de Laverny, P., Deleuil, M., et al. 2010, A&A, 523, A91
- Ghezzi, L., Cunha, K., Smith, V. V., et al. 2010, ApJ, 720, 1290
- Gilli, G., Israelián, G., Ecuivillon, A., Santos, N. C., & Mayor, M. 2006, A&A, 449, 723
- Gonzalez, G. 1998, A&A, 334, 221
- Gonzalez, G. 2009, MNRAS, 399, L103
- Gonzalez, G., Laws, C., Tyagi, S., & Reddy, B. E. 2001, AJ, 121, 432
- González Hernández, J. I., Israelián, G., Santos, N. C., et al. 2010, ApJ, 720, 1592
- Haywood, M. 2008a, A&A, 482, 673
- Haywood, M. 2008b, MNRAS, 388, 1175
- Hog, E., Kuzmin, A., Bastian, U., et al. 1998, A&A, 335, 65
- Ida, S., & Lin, D. N. C. 2004, ApJ, 616, 567
- Johnson, J. A. 2002, ApJS, 139, 219
- Johnson, J. A., Aller, K. M., Howard, A. W., & Crepp, J. R. 2010, PASJ, 122, 905
- Juric', M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864
- Kang, W., Lee, S. G., & Kim, K. M. 2011, ApJ, 736, 87
- Kurucz, R. 1993, ATLAS9 Stellar Atmosphere Programs and 2 km s⁻¹ grid, Kurucz CD-ROM No. 13, Cambridge, Mass.: Smithsonian Astrophysical Observatory, 13
- Lai, K. D., Bolte, M., Johnson, J. A., et al. 2008, ApJ, 681, 1524
- Laws, C., Gonzalez, G., Walker, K. M., et al. 2003, AJ, 125, 2664
- Lee, Y. S., Beers, T. C., An, D., et al. 2011, ApJ, 738, 187
- Liu, C., & van de Ven, G. 2012, MNRAS, submitted [arXiv:1201.1635]
- Lo Curto, G., Mayor, M., Benz, W., et al. 2010, A&A, 512, A48
- Lovis, C., Mayor, M., Pepe, F., et al. 2006, Nature, 441, 305
- Mayor, M., & Queloz, D. 1995, Nature, 378, 355
- Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
- Mayor, M., Bonfils, X., Forveille, T., et al. 2009, A&A, 507, 487
- Mayor, M., Marmier, M., Lovis, C., et al. 2011, A&A, submitted [arXiv:1109.2497]
- Mishenina, T. V., Soubiran, C., Kovtyukh, V. V., & Korotin, S. A. 2004, A&A, 418, 551
- Mordasini, C., Alibert, Y., & Benz, W. 2009, A&A, 501, 1139
- Navarro, J. F., Abadi, M. G., Venn, K. A., et al. 2011, MNRAS, 412, 1203
- Neves, V., Santos, N. C., Sousa, S. G., Correia, A. C. M., & Israelián, G. 2009, A&A, 497, 563
- Ojha, D. K., Bienaymé, O., Robin, A. C., Creze, M., & Mohan, V. 1996, A&A, 311, 456
- Petigura, E. A., & Marcy, G. W. 2011, ApJ, 735, 41
- Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Icarus, 124, 62
- Preston, G. W., Sneden, C., Thompson, I. B., Shectman, S. A., & Burley, G. S. 2006, AJ, 132, 85
- Prochaska, J. X., Naumov, S. O., Carney, B. W., McWilliam, A., & Wolfe, A. M. 2000, AJ, 120, 2513
- Reddy, B. E., Lambert, D. L., & Allende Prieto, C. 2006, MNRAS, 367, 1329
- Robin, A. C., Haywood, M., Crézé, M., Ojha, D. K., & Bienaymé, O. 1996, A&A, 305, 125
- Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. 2003, A&A, 409, 523
- Santos, N. C., Israelián, G., & Mayor, M. 2001, A&A, 373, 1019
- Santos, N. C., Israelián, G., Mayor, M., Rebolo, R., & Udry, S. 2003, A&A, 398, 363
- Santos, N. C., Israelián, G., & Mayor, M. 2004a, A&A, 415, 1153
- Santos, N. C., Bouchy, F., Mayor, M., et al. 2004b, A&A, 426, L19
- Santos, N. C., Israelián, G., Mayor, M., et al. 2005, A&A, 437, 1127
- Santos, N. C., Mayor, M., Bonfils, X., et al. 2011, A&A, 526, A112
- Schwarzschild, K. 1907, Göttingen Nachr., 614
- Schönrich, R., & Binney, J. 2009, MNRAS, 396, 203
- Schönrich, R., Binney, J., & Dehnen, W. 2010, MNRAS, 403, 1829
- Sneden, C. 1973, Ph.D. Thesis, Univ. of Texas
- Soubiran, C., Bienaymé, O., & Siebert, A. 2003, A&A, 398, 141
- Sousa, S. G., Santos, N. C., Israelián, G., et al. 2007, A&A, 469, 783
- Sousa, S. G., Santos, N. C., Mayor, M., et al. 2008, A&A, 487, 373
- Sousa, S. G., Santos, N. C., Israelián, G., et al. 2011a, A&A, 533, A141
- Sousa, S. G., Santos, N. C., Israelián, G., et al. 2011b, A&A, 526, A99
- Suda, T., Yamada, S., Katsuta, Y., et al. 2011, MNRAS, 412, 843
- Takeda, Y. 2007, PASJ, 59, 335
- Udry, S., Mayor, M., Queloz, D., Naef, D., & Santos, N. 2000, Conf. Proc.: The VLT opening symposium, 571
- Udry, S., Mayor, M., Benz, W., et al. 2006, A&A, 447, 361
- Valenti, J. A., & Fischer, D. A. 2005, VizieR Online Data Catalog, 215, 90141
- van Leeuwen, F. 2007, A&A, 474, 653

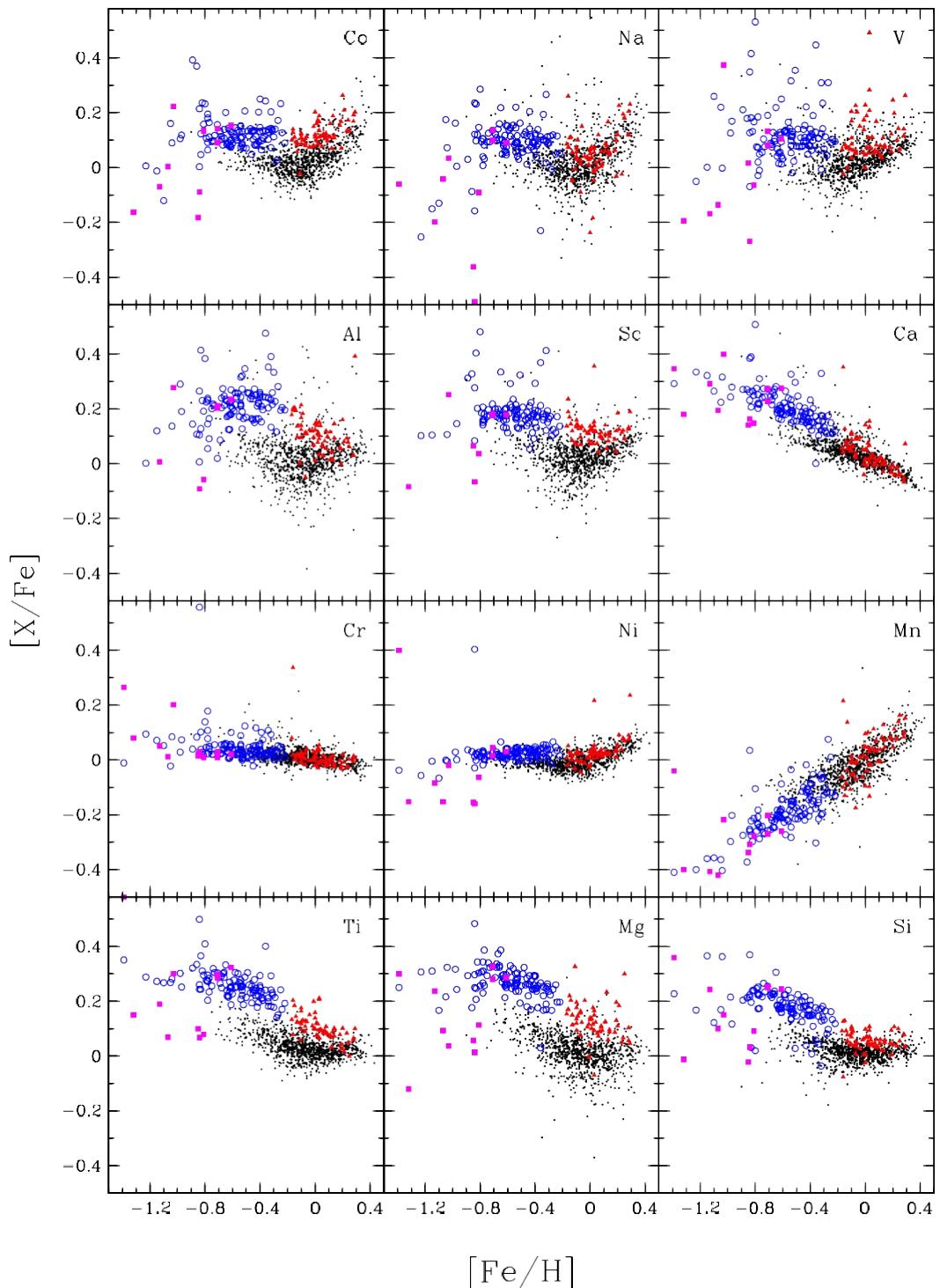


Fig. A.1. Same as Fig. 8 but for the whole sample.

19. Quantum Field Theory

University of Cambridge Part III Mathematical Tripos

Dr David Tong

*Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OWA, UK*

<http://www.damtp.cam.ac.uk/user/tong/qft.html>
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

- M. Peskin and D. Schroeder, *An Introduction to Quantum Field Theory*

This is a very clear and comprehensive book, covering everything in this course at the right level. It will also cover everything in the “Advanced Quantum Field Theory” course, much of the “Standard Model” course, and will serve you well if you go on to do research. To a large extent, our course will follow the first section of this book.

There is a vast array of further Quantum Field Theory texts, many of them with redeeming features. Here I mention a few very different ones.

- S. Weinberg, *The Quantum Theory of Fields, Vol 1*

This is the first in a three volume series by one of the masters of quantum field theory. It takes a unique route to through the subject, focussing initially on particles rather than fields. The second volume covers material lectured in “AQFT”.

- L. Ryder, *Quantum Field Theory*

This elementary text has a nice discussion of much of the material in this course.

- A. Zee, *Quantum Field Theory in a Nutshell*

This is charming book, where emphasis is placed on physical understanding and the author isn’t afraid to hide the ugly truth when necessary. It contains many gems.

- M Srednicki, *Quantum Field Theory*

A very clear and well written introduction to the subject. Both this book and Zee’s focus on the path integral approach, rather than canonical quantization that we develop in this course.

There are also resources available on the web. Some particularly good ones are listed on the course webpage: <http://www.damtp.cam.ac.uk/user/tong/qft.html>

Acknowledgements

These lecture notes are far from original. My primary contribution has been to borrow, steal and assimilate the best discussions and explanations I could find from the vast literature on the subject. I inherited the course from Nick Manton, whose notes form the backbone of the lectures. I have also relied heavily on the sources listed at the beginning, most notably the book by Peskin and Schroeder. In several places, for example the discussion of scalar Yukawa theory, I followed the lectures of Sidney Coleman, using the notes written by Brian Hill and a beautiful abridged version of these notes due to Michael Luke.

My thanks to the many who helped in various ways during the preparation of this course, including Joe Conlon, Nick Dorey, Marie Ericsson, Eyo Ita, Ian Drummond, Jerome Gauntlett, Matt Headrick, Ron Horgan, Nick Manton, Hugh Osborn and Jenni Smillie. My thanks also to the students for their sharp questions and sharp eyes in spotting typos. I am supported by the Royal Society.

1. Introduction

"There are no real one-particle systems in nature, not even few-particle systems. The existence of virtual pairs and of pair fluctuations shows that the days of fixed particle numbers are over."

Viki Weisskopf

The concept of wave-particle duality tells us that the properties of electrons and photons are fundamentally very similar. Despite obvious differences in their mass and charge, under the right circumstances both suffer wave-like diffraction and both can pack a particle-like punch.

Yet the appearance of these objects in classical physics is very different. Electrons and other matter particles are postulated to be elementary constituents of Nature. In contrast, light is a derived concept: it arises as a ripple of the electromagnetic field. If photons and particles are truly to be placed on equal footing, how should we reconcile this difference in the quantum world? Should we view the particle as fundamental, with the electromagnetic field arising only in some classical limit from a collection of quantum photons? Or should we instead view the field as fundamental, with the photon appearing only when we correctly treat the field in a manner consistent with quantum theory? And, if this latter view is correct, should we also introduce an "electron field", whose ripples give rise to particles with mass and charge? But why then didn't Faraday, Maxwell and other classical physicists find it useful to introduce the concept of matter fields, analogous to the electromagnetic field?

The purpose of this course is to answer these questions. We shall see that the second viewpoint above is the most useful: the field is primary and particles are derived concepts, appearing only after quantization. We will show how photons arise from the quantization of the electromagnetic field and how massive, charged particles such as electrons arise from the quantization of matter fields. We will learn that in order to describe the fundamental laws of Nature, we must not only introduce electron fields, but also quark fields, neutrino fields, gluon fields, W and Z-boson fields, Higgs fields and a whole slew of others. There is a field associated to each type of fundamental particle that appears in Nature.

Why Quantum Field Theory?

In classical physics, the primary reason for introducing the concept of the field is to construct laws of Nature that are *local*. The old laws of Coulomb and Newton involve "action at a distance". This means that the force felt by an electron (or planet) changes

immediately if a distant proton (or star) moves. This situation is philosophically unsatisfactory. More importantly, it is also experimentally wrong. The field theories of Maxwell and Einstein remedy the situation, with all interactions mediated in a local fashion by the field.

The requirement of locality remains a strong motivation for studying field theories in the quantum world. However, there are further reasons for treating the quantum field as fundamental¹. Here I'll give two answers to the question: Why quantum field theory?

Answer 1: Because the combination of quantum mechanics and special relativity implies that particle number is not conserved.

Particles are not indestructible objects, made at the beginning of the universe and here for good. They can be created and destroyed. They are, in fact, mostly ephemeral and fleeting. This experimentally verified fact was first predicted by Dirac who understood how relativity implies the necessity of anti-particles. An extreme demonstration of particle creation is shown in the picture, which comes from the Relativistic Heavy Ion Collider (RHIC) at Brookhaven, Long Island. This machine crashes gold nuclei together, each containing 197 nucleons. The resulting explosion contains up to 10,000 particles, captured here in all their beauty by the STAR detector.

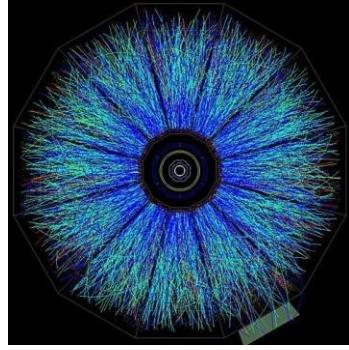


Figure 1:

We will review Dirac's argument for anti-particles later in this course, together with the better understanding that we get from viewing particles in the framework of quantum field theory. For now, we'll quickly sketch the circumstances in which we expect the number of particles to change. Consider a particle of mass m trapped in a box of size L . Heisenberg tells us that the uncertainty in the momentum is $\Delta p \geq k/L$. In a relativistic setting, momentum and energy are on an equivalent footing, so we should also have an uncertainty in the energy of order $\Delta E \geq kc/L$. However, when the uncertainty in the energy exceeds $\Delta E = 2mc^2$, then we cross the barrier to pop particle anti-particle pairs out of the vacuum. We learn that particle-anti-particle pairs are expected to be important when a particle of mass m is localized within a distance of order

$$\lambda = \frac{k}{mc}$$

¹A concise review of the underlying principles and major successes of quantum field theory can be found in the article by Frank Wilczek, <http://arxiv.org/abs/hep-th/9803075>

At distances shorter than this, there is a high probability that we will detect particle-anti-particle pairs swarming around the original particle that we put in. The distance λ is called the *Compton wavelength*. It is always smaller than the de Broglie wavelength $\lambda_{\text{dB}} = h/|p\rightarrow|$. If you like, the de Broglie wavelength is the distance at which the wavelike nature of particles becomes apparent; the Compton wavelength is the distance at which the concept of a single pointlike particle breaks down completely.

The presence of a multitude of particles and antiparticles at short distances tells us that any attempt to write down a relativistic version of the one-particle Schrödinger equation (or, indeed, an equation for any *fixed* number of particles) is doomed to failure. There is no mechanism in standard non-relativistic quantum mechanics to deal with changes in the particle number. Indeed, any attempt to naively construct 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 are the common issues that arise). In each case, this failure is telling us that once we enter the relativistic regime we need a new formalism in order to treat states with an unspecified number of particles. This formalism is quantum field theory (QFT).

Answer 2: Because all particles of the same type are the same

This sound rather dumb. But it's not! What I mean by this is that two electrons are identical in every way, regardless of where they came from and what they've been through. The same is true of every other fundamental particle. Let me illustrate this through a rather prosaic story. Suppose we capture a proton from a cosmic ray which we identify as coming from a supernova lying 8 billion lightyears away. We compare this proton with one freshly minted in a particle accelerator here on Earth. And the two are exactly the same! How is this possible? Why aren't there errors in proton production? How can two objects, manufactured so far apart in space and time, be identical in all respects? One explanation that might be offered is that there's a sea of proton "stuff" filling the universe and when we make a proton we somehow dip our hand into this stuff and from it mould a proton. Then it's not surprising that protons produced in different parts of the universe are identical: they're made of the same stuff. It turns out that this is roughly what happens. The "stuff" is the proton field or, if you look closely enough, the quark and gluon fields.

In fact, there's more to this tale. Being the "same" in the quantum world is not like being the "same" in the classical world: quantum particles that are the same are truly indistinguishable. Swapping two particles around leaves the state completely unchanged — apart from a possible minus sign. This minus sign determines the statistics of the particle. In quantum mechanics you have to put these statistics in by hand and,

to agree with experiment, should choose Bose statistics (no minus sign) for integer spin particles, and Fermi statistics (yes minus sign) for half-integer spin particles. In quantum field theory, this relationship between spin and statistics is not something that you have to put in by hand. Rather, it is a consequence of the framework.

What is Quantum Field Theory?

Having told you why QFT is necessary, I should really tell you what it is. The clue is in the name: it is the quantization of a classical field, the most familiar example of which is the electromagnetic field. In standard quantum mechanics, we're taught to take the classical degrees of freedom and promote them to operators acting on a Hilbert space. The rules for quantizing a field are no different. Thus the basic degrees of freedom in quantum field theory are *operator valued functions of space and time*. This means that we are dealing with an infinite number of degrees of freedom — at least one for every point in space. This infinity will come back to bite on several occasions.

It will turn out that the possible interactions in quantum field theory are governed by a few basic principles: locality, symmetry and renormalization group flow (the decoupling of short distance phenomena from physics at larger scales). These ideas make QFT a very robust framework: given a set of fields there is very often an almost unique way to couple them together.

What is Quantum Field Theory Good For?

The answer is: almost everything. As I have stressed above, for any relativistic system it is a necessity. But it is also a very useful tool in non-relativistic systems with many particles. Quantum field theory has had a major impact in condensed matter, high-energy physics, cosmology, quantum gravity and pure mathematics. It is literally the language in which the laws of Nature are written.

1.1 Units and Scales

Nature presents us with three fundamental dimensionful constants; the speed of light c , Planck's constant (divided by 2π) k and Newton's constant G . They have dimensions

$$\begin{aligned}[c] & [c] = LT^{-1} \\ & [k] = L^2 MT^{-1} \\ & [G] = L^3 M^{-1} T^{-2} \end{aligned}$$

Throughout this course we will work with “natural” units, defined by

$$c = k = 1 \tag{0.1}$$

which allows us to express all dimensionful quantities in terms of a single scale which we choose to be mass or, equivalently, energy (since $E = mc^2$ has become $E = m$). The usual choice of energy unit is eV , the electron volt or, more often $GeV = 10^9 eV$ or $TeV = 10^{12} eV$. To convert the unit of energy back to a unit of length or time, we need to insert the relevant powers of c and k . For example, the length scale λ associated to a mass m is the Compton wavelength

$$\lambda = \frac{k}{mc}$$

With this conversion factor, the electron mass $m_e = 10^6 eV$ translates to a length scale $\lambda_e \sim 10^{-12} m$. (The Compton wavelength is also defined with an extra factor of 2π : $\lambda = 2\pi k/mc$.)

Throughout this course we will refer to the *dimension* of a quantity, meaning the mass dimension. If X has dimensions of $(\text{mass})^d$ we will write $[X] = d$. In particular, the surviving natural quantity G has dimensions $[G] = -2$ and defines a mass scale,

$$G = \frac{kc}{M_p^2} = \frac{1}{M_p^2} \quad (0.2)$$

where $M_p \approx 10^{19} GeV$ is the *Planck scale*. It corresponds to a length $l_p \approx 10^{-33} cm$. The Planck scale is thought to be the smallest length scale that makes sense: beyond this quantum gravity effects become important and it's no longer clear that the concept of spacetime makes sense. The largest length scale we can talk of is the size of the cosmological horizon, roughly $10^{60} l_p$.

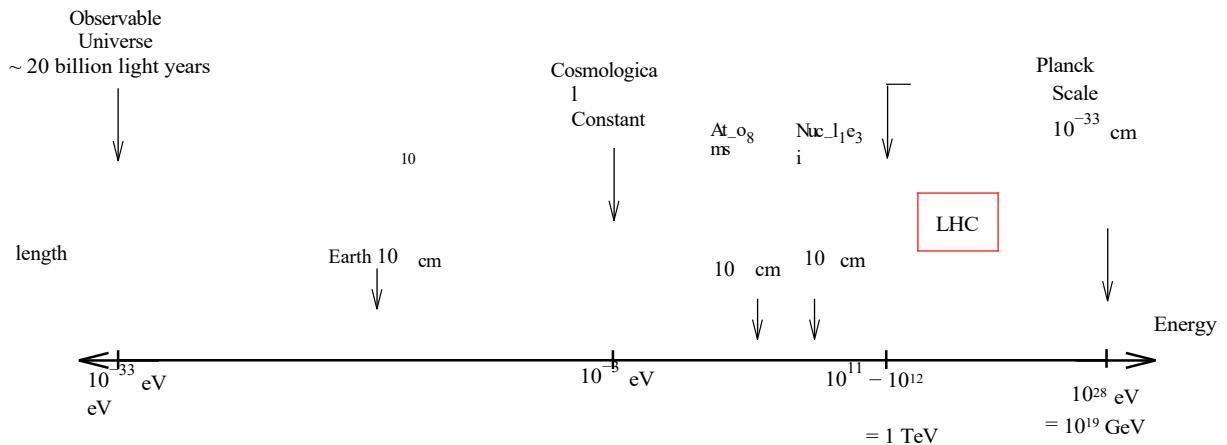


Figure 2: Energy and Distance Scales in the Universe

Some useful scales in the universe are shown in the figure. This is a logarithmic plot, with energy increasing to the right and, correspondingly, length increasing to the left. The smallest and largest scales known are shown on the figure, together with other

relevant energy scales. The standard model of particle physics is expected to hold up to about the TeV . This is precisely the regime that is currently being probed by the Large Hadron Collider (LHC) at CERN. There is a general belief that the framework of quantum field theory will continue to hold to energy scales only slightly below the Planck scale — for example, there are experimental hints that the coupling constants of electromagnetism, and the weak and strong forces unify at around 10^{18} GeV .

For comparison, the rough masses of some elementary (and not so elementary) particles are shown in the table,

Particle	Mass
Neutrinos	$\sim 10^{-2} \text{ eV}$
Electron	0.5 MeV
Muon	100 MeV
Pions	140 MeV
Proton, Neutron	1 GeV
Tau	2 GeV
W,Z Bosons	80-90 GeV
Higgs Boson	125 GeV

1. Classical Field Theory

In this first section we will discuss various aspects of classical fields. We will cover only the bare minimum ground necessary before turning to the quantum theory, and will return to classical field theory at several later stages in the course when we need to introduce new ideas.

1.1 The Dynamics of Fields

A *field* is a quantity defined at every point of space and time ($\rightarrow x, t$). While classical particle mechanics deals with a finite number of generalized coordinates $q_a(t)$, indexed by a label a , in field theory we are interested in the dynamics of fields

$$\varphi_a(\rightarrow x, t) \quad (1.1)$$

where both a and $\rightarrow x$ are considered as labels. Thus we are dealing with a system with an infinite number of degrees of freedom — at least one for each point $\rightarrow x$ in space. Notice that the concept of position has been relegated from a dynamical variable in particle mechanics to a mere label in field theory.

An Example: The Electromagnetic Field

The most familiar examples of fields from classical physics are the electric and magnetic fields, $E^\rightarrow(\rightarrow x, t)$ and $B^\rightarrow(\rightarrow x, t)$. Both of these fields are spatial 3-vectors. In a more sophisticated treatment of electromagnetism, we derive these two 3-vectors from a single 4-component field $A^\mu(\rightarrow x, t) = (\varphi, A^\rightarrow)$ where $\mu = 0, 1, 2, 3$ shows that this field is a vector in *spacetime*. The electric and magnetic fields are given by

$$E^\rightarrow = -\nabla\varphi - \frac{\partial A^\rightarrow}{\partial t} \quad \text{and} \quad B^\rightarrow = \nabla \times A^\rightarrow \quad (1.2)$$

which ensure that two of Maxwell's equations, $\nabla \cdot B^\rightarrow = 0$ and $d B^\rightarrow /dt = -\nabla \times E^\rightarrow$, hold immediately as identities.

The Lagrangian

The dynamics of the field is governed by a Lagrangian which is a function of $\varphi(\rightarrow x, t)$, $\dot{\varphi}(\rightarrow x, t)$ and $\nabla\varphi(\rightarrow x, t)$. In all the systems we study in this course, the Lagrangian is of the form,

$$L(t) = \int d^3x L(\varphi_a, \partial_\mu \varphi_a) \quad (1.3)$$

where the official name for L is the *Lagrangian density*, although everyone simply calls it the Lagrangian. The action is,

$$S = \int_{t_1}^{t_2} dt \int d^3x L = \int d^4x L \quad (1.4)$$

Recall that in particle mechanics L depends on q and \dot{q} , but not \ddot{q} . In field theory we similarly restrict to Lagrangians L depending on φ and $\partial_\mu \varphi$, and not $\partial_\mu \partial_\nu \varphi$. In principle, there's nothing to stop L depending on $\nabla\varphi$, $\nabla^2\varphi$, $\nabla^3\varphi$, etc. However, with an eye to later Lorentz invariance, we will only consider Lagrangians depending on $\nabla\varphi$ and not higher derivatives. Also we will not consider Lagrangians with explicit dependence on x^μ ; all such dependence only comes through φ and its derivatives.

We can determine the equations of motion by the principle of least action. We vary the path, keeping the end points fixed and require $\delta S = 0$,

$$\begin{aligned} \delta S &= \int d^4x \frac{\partial L}{\partial \varphi_a} \delta \varphi_a + \frac{\partial L}{\partial (\partial_\mu \varphi_a)} \delta (\partial_\mu \varphi_a) \\ &= \int d^4x \left[\frac{\partial \varphi_a}{\partial \varphi_a} - \partial_\mu \frac{\partial (\partial_\mu \varphi_a)}{\partial (\partial_\mu \varphi_a)} \right] \delta \varphi_a + \partial_\mu \frac{\partial L}{\partial (\partial_\mu \varphi_a)} \delta \varphi_a \end{aligned} \quad (1.5)$$

The last term is a total derivative and vanishes for any $\delta \varphi_a(\rightarrow x, t)$ that decays at spatial infinity and obeys $\delta \varphi_a(\rightarrow x, t_1) = \delta \varphi_a(\rightarrow x, t_2) = 0$. Requiring $\delta S = 0$ for all such paths yields the Euler-Lagrange equations of motion for the fields φ_a ,

$$\partial_\mu \frac{\partial L}{\partial (\partial_\mu \varphi_a)} = 0 \quad (1.6)$$

1.1.1 An Example: The Klein-Gordon Equation

Consider the Lagrangian for a real scalar field $\varphi(\rightarrow x, t)$,

$$\begin{aligned} L &= \frac{1}{2} \eta^{\mu\nu} \partial_\mu \varphi \partial_\nu \varphi - \frac{1}{2} m^2 \varphi^2 \\ &= \frac{1}{2} \varphi^2 - \frac{1}{2} (\nabla \varphi)^2 - \frac{1}{2} m^2 \varphi^2 \end{aligned} \quad (1.7)$$

where we are using the Minkowski space metric

$$\eta^{\mu\nu} = \eta_{\mu\nu} = \begin{smallmatrix} +1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{smallmatrix} \quad (1.8)$$

Comparing (1.7) to the usual expression for the Lagrangian $L = T - V$, we identify the kinetic energy of the field as

$$T = \int d^3x \frac{1}{2} \dot{\varphi}^2 \quad (1.9)$$

and the potential energy of the field as

$$V = \int d^3x \frac{1}{2} (\nabla\varphi)^2 + \frac{1}{2} m^2 \varphi^2 \quad (1.10)$$

The first term in this expression is called the gradient energy, while the phrase “potential energy”, or just “potential”, is usually reserved for the last term.

To determine the equations of motion arising from (1.7), we compute

$$\frac{\partial L}{\partial\varphi} = -m^2\varphi \quad \text{and} \quad \frac{\partial L}{\partial(\partial_\mu\varphi)} = \partial^\mu\varphi \equiv (\dot{\varphi}, -\nabla\varphi) \quad (1.11)$$

The Euler-Lagrange equation is then

$$\ddot{\varphi} - \nabla^2\varphi + m^2\varphi = 0 \quad (1.12)$$

which we can write in relativistic form as

$$\partial_\mu\partial^\mu\varphi + m^2\varphi = 0 \quad (1.13)$$

This is the *Klein-Gordon Equation*. The Laplacian in Minkowski space is sometimes denoted by \Box . In this notation, the Klein-Gordon equation reads $\Box\varphi + m^2\varphi = 0$.

An obvious generalization of the Klein-Gordon equation comes from considering the Lagrangian with arbitrary potential $V(\varphi)$,

$$L = \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - V(\varphi) \Rightarrow \partial_\mu\partial^\mu\varphi + \frac{\partial V}{\partial\varphi} = 0 \quad (1.14)$$

1.1.2 Another Example: First Order Lagrangians

We could also consider a Lagrangian that is linear in time derivatives, rather than quadratic. Take a complex scalar field ψ whose dynamics is defined by the real Lagrangian

$$L = \frac{i}{2}(\psi^\wedge\dot{\psi} - \dot{\psi}^\wedge\psi) - \nabla\psi^\wedge \cdot \nabla\psi - m\psi^\wedge\psi \quad (1.15)$$

We can determine the equations of motion by treating ψ and ψ^\wedge as independent objects, so that

$$\frac{\partial L}{\partial\psi^\wedge} = \frac{i}{2}\dot{\psi} - m\psi \quad \text{and} \quad \frac{\partial L}{\partial\dot{\psi}^\wedge} = -\frac{i}{2}\psi \quad \text{and} \quad \frac{\partial L}{\partial\nabla\psi^\wedge} = -\nabla\psi \quad (1.16)$$

This gives us the equation of motion

$$\frac{\partial\psi}{\partial t} = -\nabla^2\psi + m\psi \quad (1.17)$$

This looks very much like the Schrödinger equation. Except it isn't! Or, at least, the interpretation of this equation is very different: the field ψ is a classical field with none of the probability interpretation of the wavefunction. We'll come back to this point in Section 2.8.

The initial data required on a Cauchy surface differs for the two examples above. When $L \sim \varphi^2$, both φ and $\dot{\varphi}$ must be specified to determine the future evolution; however when $L \sim \psi^\mu \psi_\mu$, only ψ and $\dot{\psi}^\mu$ are needed.

1.1.3 A Final Example: Maxwell's Equations

We may derive Maxwell's equations in the vacuum from the Lagrangian,

$$L = -\frac{1}{2} (\partial_\mu A_\nu) (\partial^\mu A^\nu) + \frac{1}{2} (\partial_\mu A^\mu)^2 \quad (1.18)$$

Notice the funny minus signs! This is to ensure that the kinetic terms for A_i are positive using the Minkowski space metric (1.8), so $L_t \sim \frac{1}{2} A^2$. The Lagrangian (1.18) has no kinetic term A^2 for A_0 . We will see the consequences of this in Section 6. To see that

Maxwell's equations indeed follow from (1.18), we compute

$$\frac{\partial L}{\partial(\partial_\mu A_\nu)} = -\partial^\mu A^\nu + (\partial_\rho A^\rho) \eta^{\mu\nu} \quad (1.19)$$

from which we may derive the equations of motion,

$$\partial_\mu \frac{\partial L}{\partial(\partial_\mu A_\nu)} = -\partial^2 A^\nu + \partial^\nu (\partial_\rho A^\rho) = -\partial_\mu (\partial^\mu A^\nu - \partial^\nu A^\mu) \equiv -\partial_\mu F^{\mu\nu} \quad (1.20)$$

where the *field strength* is defined by $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. You can check using (1.2) that this reproduces the remaining two Maxwell's equations in a vacuum: $\nabla \cdot E^\rightarrow = 0$ and $\partial E^\rightarrow / \partial t = \nabla \times B^\rightarrow$. Using the notation of the field strength, we may rewrite the

Maxwell Lagrangian (up to an integration by parts) in the compact form

$$L = -\frac{1}{4} \int_{\mu\nu} F_{\mu\nu} F \quad (1.21)$$

1.1.4 Locality, Locality, Locality

In each of the examples above, the Lagrangian is *local*. This means that there are no terms in the Lagrangian coupling $\varphi(\rightarrow x, t)$ directly to $\varphi(\rightarrow y, t)$ with $\rightarrow x \neq \rightarrow y$. For example, there are no terms that look like

$$L = \int d^3x d^3y \varphi(\rightarrow x) \varphi(\rightarrow y) \quad (1.22)$$

A priori, there's no reason for this. After all, $\rightarrow x$ is merely a label, and we're quite happy to couple other labels together (for example, the term $\partial_3 A_0 \partial_0 A_3$ in the Maxwell Lagrangian couples the $\mu = 0$ field to the $\mu = 3$ field). But the closest we get for the $\rightarrow x$ label is a coupling between $\varphi(\rightarrow x)$ and $\varphi(\rightarrow x + \delta \rightarrow x)$ through the gradient term $(\nabla \varphi)^2$. This property of locality is, as far as we know, a key feature of *all* theories of Nature. Indeed, one of the main reasons for introducing field theories in classical physics is to implement locality. In this course, we will only consider local Lagrangians.

1.2 Lorentz Invariance

The laws of Nature are relativistic, and one of the main motivations to develop quantum field theory is to reconcile quantum mechanics with special relativity. To this end, we want to construct field theories in which space and time are placed on an equal footing and the theory is invariant under Lorentz transformations,

$$x^\mu \longrightarrow (x^r)^\mu = \Lambda^\mu_\nu x^\nu \quad (1.23)$$

where Λ^μ_ν satisfies

$$\sum_\sigma \eta^{\sigma\tau} \Lambda^\nu_\tau = \eta^{\mu\nu} \quad (1.24)$$

For example, a rotation by ϑ about the x^3 -axis, and a boost by $v < 1$ along the x^1 -axis are respectively described by the Lorentz transformations

$$\Lambda^\nu = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \vartheta & -\sin \vartheta & 0 \\ 0 & \sin \vartheta & \cos \vartheta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \Lambda^\mu = \begin{bmatrix} \gamma & -\gamma v & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1.25)$$

with $\gamma = \sqrt{1 - v^2}$. The Lorentz transformations form a Lie group under matrix multiplication. You'll learn more about this in the "Symmetries and Particle Physics" course.

The Lorentz transformations have a *representation* on the fields. The simplest example is the scalar field which, under the Lorentz transformation $x \rightarrow \Lambda x$, transforms as

$$\varphi(x) \rightarrow \varphi^r(x) = \varphi(\Lambda^{-1}x) \quad (1.26)$$

The inverse Λ^{-1} appears in the argument because we are dealing with an *active* transformation in which the field is truly shifted. To see why this means that the inverse appears, it will suffice to consider a non-relativistic example such as a temperature field. Suppose we start with an initial field $\varphi(\rightarrow x)$ which has a hotspot at, say, $\rightarrow x = (1, 0, 0)$. After a rotation $\rightarrow x \rightarrow R \rightarrow x$ about the z-axis, the new field $\varphi^r(\rightarrow x)$ will have the hotspot at

$\rightarrow x = (0, 1, 0)$. If we want to express $\varphi^r(\rightarrow x)$ in terms of the old field φ , we need to place ourselves at $\rightarrow x = (0, 1, 0)$ and ask what the old field looked like where we've come from at $R^{-1}(0, 1, 0) = (1, 0, 0)$. This R^{-1} is the origin of the inverse transformation. (If we were instead dealing with a passive transformation in which we relabel our choice of coordinates, we would have instead $\varphi(x) \rightarrow \varphi^r(x) = \varphi(\Lambda x)$).

The definition of a Lorentz invariant theory is that if $\varphi(x)$ solves the equations of motion then $\varphi(\Lambda^{-1}x)$ also solves the equations of motion. We can ensure that this property holds by requiring that the action is Lorentz invariant. Let's look at our examples:

Example 1: The Klein-Gordon Equation

For a real scalar field we have $\varphi(x) \rightarrow \varphi'(x) = \varphi(\Lambda^{-1}x)$. The derivative of the scalar field transforms as a vector, meaning

$$(\partial_\mu \varphi)(x) \rightarrow (\Lambda^{-1})^\nu_\mu (\partial_\nu \varphi)(y)$$

where $y = \Lambda^{-1}x$. This means that the derivative terms in the Lagrangian density transform as

$$\begin{aligned} L_{\text{deriv}}(x) &= \partial_\mu \varphi(x) \partial_\nu \varphi(x) \eta^{\mu\nu} \longrightarrow (\Lambda^{-1})^\rho_\mu (\partial_\rho \varphi)(y) (\Lambda^{-1})^\sigma_\nu (\partial_\sigma \varphi)(y) \eta^{\mu\nu} \\ &= (\partial_\rho \varphi)(y) (\partial_\sigma \varphi)(y) \eta^{\rho\sigma} \\ &= L_{\text{deriv}}(y) \end{aligned} \quad (1.27)$$

The potential terms transform in the same way, with $\varphi^2(x) \rightarrow \varphi^2(y)$. Putting this all together, we find that the action is indeed invariant under Lorentz transformations,

$$S = \int d^4x L(x) \longrightarrow \int d^4x L(y) = \int d^4y L(y) = S \quad (1.28)$$

where, in the last step, we need the fact that we don't pick up a Jacobian factor when we change integration variables from d^4x to d^4y . This follows because $\det \Lambda = 1$. (At least for Lorentz transformations connected to the identity which, for now, is all we deal with).

Example 2: First Order Dynamics

In the first-order Lagrangian (1.15), space and time are not on the same footing. (L is linear in time derivatives, but quadratic in spatial derivatives). The theory is not Lorentz invariant.

In practice, it's easy to see if the action is Lorentz invariant: just make sure all the Lorentz indices $\mu = 0, 1, 2, 3$ are contracted with Lorentz invariant objects, such as the metric $\eta_{\mu\nu}$. Other Lorentz invariant objects you can use include the totally antisymmetric tensor $\epsilon_{\mu\nu\rho\sigma}$ and the matrices γ_μ that we will introduce when we come to discuss spinors in Section 4.

Example 3: Maxwell's Equations

Under a Lorentz transformation $A^\mu(x) \rightarrow \Lambda^\mu_\nu A^\nu(\Lambda^{-1}x)$. You can check that Maxwell's Lagrangian (1.21) is indeed invariant. Of course, historically electrodynamics was the first Lorentz invariant theory to be discovered: it was found even before the concept of Lorentz invariance.

1.3 Symmetries

The role of symmetries in field theory is possibly even more important than in particle mechanics. There are Lorentz symmetries, internal symmetries, gauge symmetries, supersymmetries.... We start here by recasting Noether's theorem in a field theoretic framework.

1.3.1 Noether's Theorem

Every continuous symmetry of the Lagrangian gives rise to a conserved *current* $j^\mu(x)$ such that the equations of motion imply

$$\partial_\mu j^\mu = 0 \quad (1.29)$$

or, in other words, $\partial j^0 / \partial t + \nabla \cdot \vec{j} = 0$.

A Comment: A conserved current implies a conserved charge Q , defined as

$$Q = \int_{\mathbf{R}^3} d^3x j^0 \quad (1.30)$$

which one can immediately see by taking the time derivative,

$$\frac{dQ}{dt} = \int_{\mathbf{R}^3} d^3x \frac{\partial j^0}{\partial t} = - \int_{\mathbf{R}^3} d^3x \nabla \cdot \vec{j} = 0 \quad (1.31)$$

assuming that $\vec{j} \rightarrow 0$ sufficiently quickly as $|\vec{x}| \rightarrow \infty$. However, the existence of a current is a much stronger statement than the existence of a conserved charge because it implies that charge is conserved *locally*. To see this, we can define the charge in a finite volume V ,

$$Q_V = \int_V d^3x j^0 \quad (1.32)$$

Repeating the analysis above, we find that

$$\frac{dQ_V}{dt} = - \int_V d^3x \nabla \cdot \vec{j} = - \int_A \vec{j} \cdot d\vec{S} \quad (1.33)$$

where A is the area bounding V and we have used Stokes' theorem. This equation means that any charge leaving V must be accounted for by a flow of the current 3-vector \vec{j} out of the volume. This kind of local conservation of charge holds in any local field theory.

Proof of Noether's Theorem: We'll prove the theorem by working infinitesimally. We may always do this if we have a continuous symmetry. We say that the transformation

$$\delta\varphi_a(x) = X_a(\varphi) \quad (1.34)$$

is a symmetry if the Lagrangian changes by a total derivative,

$$\delta L = \partial_\mu F^\mu \quad (1.35)$$

for some set of functions $F^\mu(\varphi)$. To derive Noether's theorem, we first consider making an *arbitrary* transformation of the fields $\delta\varphi_a$. Then

$$\begin{aligned} \delta L &= \frac{\partial L}{\partial \varphi_a} \delta\varphi_a + \frac{\partial L}{\partial (\partial_\mu \varphi_a)} \partial_\mu (\delta\varphi_a) \\ &= \frac{\partial L}{\partial \varphi_a} - \partial^\mu \partial_\mu (\delta\varphi_a) + \delta\varphi_a \partial_\mu \frac{\partial L}{\partial (\partial_\mu \varphi_a)} \end{aligned} \quad (1.36)$$

When the equations of motion are satisfied, the term in square brackets vanishes. So we're left with

$$\delta L = \partial_\mu \frac{\partial L}{\partial (\partial_\mu \varphi_a)} \delta\varphi_a \quad (1.37)$$

But for the symmetry transformation $\delta\varphi_a = X_a(\varphi)$, we have by definition $\delta L = \partial_\mu F^\mu$. Equating this expression with (1.37) gives us the result

$$\partial_\mu j^\mu = 0 \quad \text{with} \quad j^\mu = \frac{\partial L}{\partial (\partial_\mu \varphi_a)} X_a(\varphi) - F^\mu(\varphi) \quad (1.38)$$

1.3.2 An Example: Translations and the Energy-Momentum Tensor

Recall that in classical particle mechanics, invariance under spatial translations gives rise to the conservation of momentum, while invariance under time translations is responsible for the conservation of energy. We will now see something similar in field theories. Consider the infinitesimal translation

$$x^\nu \rightarrow x^\nu - \epsilon^\nu \quad \Rightarrow \quad \varphi_a(x) \rightarrow \varphi_a(x) + \epsilon^\nu \partial_\nu \varphi_a(x) \quad (1.39)$$

(where the sign in the field transformation is plus, instead of minus, because we're doing an active, as opposed to passive, transformation). Similarly, once we substitute a specific field configuration $\varphi(x)$ into the Lagrangian, the Lagrangian itself also transforms as

$$L(x) \rightarrow L(x) + \epsilon^v \partial_v L(x) \quad (1.40)$$

Since the change in the Lagrangian is a total derivative, we may invoke Noether's theorem which gives us four conserved currents $(j^\mu)_v$, one for each of the translations ϵ^v with $v = 0, 1, 2, 3$,

$$(j^\mu)_v = \frac{\partial L}{\partial(\partial_\mu \varphi_a)} \partial_v \varphi_a - \delta_v^\mu L \equiv T_v^\mu \quad (1.41)$$

T_v^μ is called the *energy-momentum tensor*. It satisfies

$$\partial_\mu T^\mu = 0 \quad (1.42)$$

The four conserved quantities are given by

$$E = \int d^3x T^{00} \quad \text{and} \quad P^i = \int d^3x T^{0i} \quad (1.43)$$

where E is the total energy of the field configuration, while P^i is the total momentum of the field configuration.

An Example of the Energy-Momentum Tensor

Consider the simplest scalar field theory with Lagrangian (1.7). From the above discussion, we can compute

$$T^{\mu\nu} = \partial^\mu \varphi \partial^\nu \varphi - \eta^{\mu\nu} L \quad (1.44)$$

One can verify using the equation of motion for φ that this expression indeed satisfies $\partial_\mu T^{\mu\nu} = 0$. For this example, the conserved energy and momentum are given by

$$E = \int d^3x \frac{1}{2} \dot{\varphi}^2 + \frac{1}{2} (\nabla \varphi)^2 + \frac{1}{2} m^2 \varphi^2 \quad (1.45)$$

$$P^i = \int d^3x \varphi \cdot \partial^i \varphi \quad (1.46)$$

Notice that for this example, $T^{\mu\nu}$ came out symmetric, so that $T^{\mu\nu} = T^{\nu\mu}$. This won't always be the case. Nevertheless, there is typically a way to massage the energy-momentum tensor of any theory into a symmetric form by adding an extra term

$$\Theta^{\mu\nu} = T^{\mu\nu} + \partial_\rho \Gamma^{\mu\nu\rho} \quad (1.47)$$

where $\Gamma^{\mu\nu\rho}$ is some function of the fields that is anti-symmetric in the first two indices so $\Gamma^{\mu\nu\rho} = -\Gamma^{\nu\mu\rho}$. This guarantees that $\partial_\mu \partial_\rho \Gamma^{\mu\nu\rho} = 0$ so that the new energy-momentum tensor is also a conserved current.

A Cute Trick

One reason that you may want a symmetric energy-momentum tensor is to make contact with general relativity: such an object sits on the right-hand side of Einstein's field equations. In fact this observation provides a quick and easy way to determine a symmetric energy-momentum tensor. Firstly consider coupling the theory to a curved background spacetime, introducing an arbitrary metric $g_{\mu\nu}(x)$ in place of $\eta_{\mu\nu}$, and replacing the kinetic terms with suitable covariant derivatives using "minimal coupling". Then a symmetric energy momentum tensor in the flat space theory is given by

$$\Theta^{\mu\nu} = -\frac{2}{-g} \frac{\partial(\sqrt{-g}L)}{\partial g_{\mu\nu}} \quad (1.48)$$

$g_{\mu\nu} = \eta_{\mu\nu}$

It should be noted however that this trick requires a little more care when working with spinors.

1.3.3 Another Example: Lorentz Transformations and Angular Momentum

In classical particle mechanics, rotational invariance gave rise to conservation of angular momentum. What is the analogy in field theory? Moreover, we now have further Lorentz transformations, namely boosts. What conserved quantity do they correspond to? To answer these questions, we first need the infinitesimal form of the Lorentz transformations

$$\Lambda^\mu_v = \delta^\mu_v + \omega^\mu_v \quad (1.49)$$

where ω^μ_v is infinitesimal. The condition (1.24) for Λ to be a Lorentz transformation becomes

$$\begin{aligned} (\delta^\mu_\sigma + \omega^\mu_\sigma)(\delta^\nu_\tau + \omega^\nu_\tau) \eta^{\sigma\tau} &= \eta^{\mu\nu} \\ \Rightarrow \quad \omega^{\mu\nu} + \omega^{\nu\mu} &= 0 \\ \omega^{\nu\mu} & \end{aligned} \quad (1.50)$$

So the infinitesimal form $\omega^{\mu\nu}$ of the Lorentz transformation must be an anti-symmetric matrix. As a check, the number of different 4×4 anti-symmetric matrices is $4 \times 3/2 = 6$, which agrees with the number of different Lorentz transformations (3 rotations + 3 boosts). Now the transformation on a scalar field is given by

$$\begin{aligned} \varphi(x) \rightarrow \varphi^r(x) &= \varphi(\Lambda^{-1}x) \\ &= \varphi(x^\mu - \omega^\mu_v x^v) \\ &= \varphi(x^\mu) - \omega^\mu_v x^v \partial_\mu \varphi(x) \end{aligned} \quad (1.51)$$

from which we see that

$$\delta\varphi = -\omega^\mu x^\nu \partial_\mu \varphi \quad (1.52)$$

By the same argument, the Lagrangian density transforms as

$$\delta L = -\omega^\mu x^\nu \partial_\mu L = -\partial_\mu (\omega^\mu x^\nu L) \quad (1.53)$$

where the last equality follows because $\omega^\mu_\mu = 0$ due to anti-symmetry. Once again, the Lagrangian changes by a total derivative so we may apply Noether's theorem (now with $F^\mu = -\omega^\mu x^\nu L$) to find the conserved current

$$\begin{aligned} j^\mu &= -\frac{\partial L}{\partial(\partial_\mu \varphi)} x^\rho \partial_\rho \varphi + \omega^\mu x^\nu L \\ &= -\omega^\rho \frac{\partial L}{\partial(\partial_\mu \varphi)} x^\nu \partial_\rho \varphi - \delta^\mu_\rho x^\nu L = -\omega^\rho x^\nu T^\mu_\rho \end{aligned} \quad (1.54)$$

Unlike in the previous example, I've left the infinitesimal choice of ω^μ_ν in the expression for this current. But really, we should strip it out to give six different currents, i.e. one for each choice of ω^μ_ν . We can write them as

$$(J^\mu)^{\rho\sigma} = x^\rho T^{\mu\sigma} - x^\sigma T^{\mu\rho} \quad (1.55)$$

which satisfy $\partial_\mu (J^\mu)^{\rho\sigma} = 0$ and give rise to 6 conserved charges. For $\rho, \sigma = 1, 2, 3$, the Lorentz transformation is a rotation and the three conserved charges give the total angular momentum of the field.

$$Q^{ij} = \int d^3x (x^i T^{0j} - x^j T^{0i}) \quad (1.56)$$

But what about the boosts? In this case, the conserved charges are

$$Q^{0i} = \int d^3x (x^0 T^{0i} - x^i T^{00}) \quad (1.57)$$

The fact that these are conserved tells us that

$$\begin{aligned} 0 &= dQ^{0i} = \int_3 \partial_{0i} \partial T^{0i} - \int_3 \frac{d}{dt} \left(\int_3 d x^i T^{00} \right) \\ &= \frac{d}{dt} \left(\int_3 d x^i T^{00} \right) + t \frac{d}{dt} \left(\int_3 d x^i T^{00} \right) \\ &= P^i + t \frac{dP^i}{dt} - \frac{d}{dt} \left(\int_3 d x^i T^{00} \right) \end{aligned} \quad (1.58)$$

But we know that P^i is conserved, so $dP^i/dt = 0$, leaving us with the following consequence of invariance under boosts:

$$\frac{d}{dt} \left(\int d^3x x^i T^{00} \right) = \text{constant} \quad (1.59)$$

This is the statement that the center of energy of the field travels with a constant velocity. It's kind of like a field theoretic version of Newton's first law but, rather surprisingly, appearing here as a conservation law.

1.3.4 Internal Symmetries

The above two examples involved transformations of spacetime, as well as transformations of the field. An *internal symmetry* is one that only involves a transformation of the fields and acts the same at every point in spacetime. The simplest example occurs for a complex scalar field $\psi(x) = (\varphi_1(x) + i\varphi_2(x))/\sqrt{2}$. We can build a real Lagrangian by

$$L = \partial_\mu \psi^\dagger \partial^\mu \psi - V(|\psi|^2) \quad (1.60)$$

where the potential is a general polynomial in $|\psi|^2 = \psi^\dagger \psi$. To find the equations of motion, we could expand ψ in terms of φ_1 and φ_2 and work as before. However, it's easier (and equivalent) to treat ψ and ψ^\dagger as independent variables and vary the action with respect to both of them. For example, varying with respect to ψ^\dagger leads to the equation of motion

$$\frac{\partial_\mu \partial^\mu \psi + \frac{\partial V(\psi^\dagger)}{\partial \psi} = 0}{\partial \psi^\dagger} \quad (1.61)$$

The Lagrangian has a continuous symmetry which rotates φ_1 and φ_2 or, equivalently, rotates the phase of ψ :

$$\psi \rightarrow e^{i\alpha} \psi \quad \text{or} \quad \delta\psi = i\alpha\psi \quad (1.62)$$

where the latter equation holds with α infinitesimal. The Lagrangian remains invariant under this change: $\delta L = 0$. The associated conserved current is

$$j^\mu = i(\partial^\mu \psi^\dagger)\psi - i\psi^\dagger(\partial^\mu \psi) \quad (1.63)$$

We will later see that the conserved charges arising from currents of this type have the interpretation of electric charge or particle number (for example, baryon or lepton number).

Non-Abelian Internal Symmetries

Consider a theory involving N scalar fields φ_a , all with the same mass and the Lagrangian

$$L = \frac{1}{2} \sum_{a=1}^N \partial_\mu \varphi_a \partial^\mu \varphi_a - \frac{1}{2} \sum_{a=1}^N m \varphi_a^2 - g \sum_{a=1}^N \varphi_a^2 \quad ! \quad (1.64)$$

In this case the Lagrangian is invariant under the non-Abelian symmetry group $G = SO(N)$. (Actually $O(N)$ in this case). One can construct theories from complex fields in a similar manner that are invariant under an $SU(N)$ symmetry group. Non-Abelian symmetries of this type are often referred to as *global* symmetries to distinguish them from the “local gauge” symmetries that you will meet later. Isospin is an example of such a symmetry, albeit realized only approximately in Nature.

Another Cute Trick

There is a quick method to determine the conserved current associated to an internal symmetry $\delta\varphi = \alpha\varphi$ for which the Lagrangian is invariant. Here, α is a constant real number. (We may generalize the discussion easily to a non-Abelian internal symmetry for which α becomes a matrix). Now consider performing the transformation but where α depends on spacetime: $\alpha = \alpha(x)$. The action is no longer invariant. However, the change must be of the form

$$\delta L = (\partial_\mu \alpha) h^\mu(\varphi) \quad (1.65)$$

since we know that $\delta L = 0$ when α is constant. The change in the action is therefore

$$\delta S = \int d^4x \delta L = - \int d^4x \alpha(x) \partial_\mu h^\mu \quad (1.66)$$

which means that when the equations of motion are satisfied (so $\delta S = 0$ for all variations, including $\delta\varphi = \alpha(x)\varphi$) we have

$$\partial_\mu h^\mu = 0 \quad (1.67)$$

We see that we can identify the function $h^\mu = j^\mu$ as the conserved current. This way of viewing things emphasizes that it is the derivative terms, not the potential terms, in the action that contribute to the current. (The potential terms are invariant even when $\alpha = \alpha(x)$).

1.4 The Hamiltonian Formalism

The link between the Lagrangian formalism and the quantum theory goes via the path integral. In this course we will not discuss path integral methods, and focus instead on canonical quantization. For this we need the Hamiltonian formalism of field theory. We start by defining the *momentum* $\pi^a(x)$ conjugate to $\varphi_a(x)$,

$$\pi^a(x) = \frac{\partial \mathcal{L}}{\partial \dot{\varphi}_a} \quad (1.68)$$

The conjugate momentum $\pi^a(x)$ is a function of x , just like the field $\varphi_a(x)$ itself. It is not to be confused with the total momentum P^i defined in (1.43) which is a single number characterizing the whole field configuration. The *Hamiltonian density* is given by

$$H = \pi^a(x) \dot{\varphi}_a(x) - L(x) \quad (1.69)$$

where, as in classical mechanics, we eliminate $\dot{\varphi}_a(x)$ in favour of $\pi^a(x)$ everywhere in H . The Hamiltonian is then simply

$$H = \int d^3x H \quad (1.70)$$

An Example: A Real Scalar Field

For the Lagrangian

$$L = \frac{1}{2}\dot{\varphi}^2 - \frac{1}{2}(\nabla\varphi)^2 - V(\varphi) \quad (1.71)$$

the momentum is given by $\pi = \dot{\varphi}$, which gives us the Hamiltonian,

$$H = \int d^3x \left[\frac{1}{2}\pi^2 + \frac{1}{2}(\nabla\varphi)^2 + V(\varphi) \right] \quad (1.72)$$

Notice that the Hamiltonian agrees with the definition of the total energy (1.45) that we get from applying Noether's theorem for time translation invariance.

In the Lagrangian formalism, Lorentz invariance is clear for all to see since the action is invariant under Lorentz transformations. In contrast, the Hamiltonian formalism is *not* manifestly Lorentz invariant: we have picked a preferred time. For example, the equations of motion for $\varphi(x) = \varphi(\rightarrow x, t)$ arise from Hamilton's equations,

$$\varphi'(\rightarrow x, t) = \frac{\partial H}{\partial \pi(\rightarrow x, t)} \quad \text{and} \quad \pi'(\rightarrow x, t) = -\frac{\partial H}{\partial \varphi(\rightarrow x, t)} \quad (1.73)$$

which, unlike the Euler-Lagrange equations (1.6), do not look Lorentz invariant. Nevertheless, even though the Hamiltonian framework doesn't *look* Lorentz invariant, the physics must remain unchanged. If we start from a relativistic theory, all final answers must be Lorentz invariant even if it's not manifest at intermediate steps. We will pause at several points along the quantum route to check that this is indeed the case.

2. Free Fields

"The career of a young theoretical physicist consists of treating the harmonic oscillator in ever-increasing levels of abstraction."

Sidney

Coleman

2.1 Canonical Quantization

In quantum mechanics, canonical quantization is a recipe that takes us from the Hamiltonian formalism of classical dynamics to the quantum theory. The recipe tells us to take the generalized coordinates q_a and their conjugate momenta p^a and promote them to operators. The Poisson bracket structure of classical mechanics morphs into the structure of commutation relations between operators, so that, in units with $\hbar = 1$,

$$\begin{aligned}[q_a, q_b] &= [p^a, p^b] = 0 \\ [q_a, p^b] &= i \delta_a^b\end{aligned}\tag{2.1}$$

In field theory we do the same, now for the field $\varphi_a(\rightarrow x)$ and its momentum conjugate $\pi^b(\rightarrow x)$. Thus a *quantum field* is an operator valued function of space obeying the commutation relations

$$\begin{aligned}[\varphi_a(\rightarrow x), \varphi_b(\rightarrow y)] &= [\pi^a(\rightarrow x), \pi^b(\rightarrow y)] = 0 \\ [\varphi_a(\rightarrow x), \pi^b(\rightarrow y)] &= i \delta^{(3)}(\rightarrow x - \rightarrow y_a) \delta^b\end{aligned}\tag{2.2}$$

Note that we've lost all track of Lorentz invariance since we have separated space $\rightarrow x$ and time t . We are working in the Schrödinger picture so that the operators $\varphi_a(\rightarrow x)$ and $\pi^a(\rightarrow x)$ do not depend on time at all — only on space. All time dependence sits in the states $|\psi\rangle$ which evolve by the usual Schrödinger equation

$$i \frac{d|\psi\rangle}{dt} = H |\psi\rangle\tag{2.3}$$

We aren't doing anything different from usual quantum mechanics; we're merely applying the old formalism to fields. Be warned however that the notation $|\psi\rangle$ for the state is deceptively simple: if you were to write the wavefunction in quantum field theory, it would be a *functional*, that is a function of every possible configuration of the field φ .

The typical information we want to know about a quantum theory is the spectrum of the Hamiltonian H . In quantum field theories, this is usually *very hard*. One reason for this is that we have an infinite number of degrees of freedom — at least one for every point $\rightarrow x$ in space. However, for certain theories — known as *free theories* — we can find a way to write the dynamics such that each degree of freedom evolves independently

from all the others. Free field theories typically have Lagrangians which are quadratic in the fields, so that the equations of motion are linear. For example, the simplest relativistic free theory is the classical Klein-Gordon (KG) equation for a real scalar field $\varphi(\rightarrow x, t)$,

$$\partial_\mu \partial^\mu \varphi + m^2 \varphi = 0 \quad (2.4)$$

To exhibit the coordinates in which the degrees of freedom decouple from each other, we need only take the Fourier transform,

$$\varphi(\rightarrow x, t) = \int \frac{d^3 p}{(2\pi)^3} e^{ip \cdot \rightarrow x} \varphi(\rightarrow p, t) \quad (2.5)$$

Then $\varphi(\rightarrow p, t)$
satisfies

$$\frac{\partial^2}{\partial t^2} + (p^2 + m^2) \varphi(\rightarrow p, t) = 0 \quad (2.6)$$

Thus, for each value of $p \rightarrow$, $\varphi(p \rightarrow, t)$ solves the equation of a harmonic oscillator vibrating at frequency

$$\omega_{p \rightarrow} = \sqrt{p^2 + m^2} \quad (2.7)$$

We learn that the most general solution to the KG equation is a linear superposition of simple harmonic oscillators, each vibrating at a different frequency with a different amplitude. To quantize $\varphi(\rightarrow x, t)$ we must simply quantize this infinite number of harmonic oscillators. Let's recall how to do this.

2.1.1 The Simple Harmonic Oscillator

Consider the quantum mechanical Hamiltonian

$$H = \frac{1}{2} p^2 + \frac{1}{2} \omega^2 q^2 \quad (2.8)$$

with the canonical commutation relations $[q, p] = i$. To find the spectrum we define the creation and annihilation operators (also known as raising/lowering operators, or sometimes ladder operators)

$$q = \frac{i\omega}{2} q + \frac{\sqrt{\omega}}{2} p \quad , \quad a^\dagger = \frac{i\omega}{\sqrt{2}} q - \frac{\sqrt{\omega}}{2} p \quad (2.9)$$

which can be easily inverted to give

$$q = \frac{1}{\sqrt{2\omega}} (a + a^\dagger) \quad , \quad p = -\frac{\sqrt{\omega}}{2} (a - a^\dagger) \quad (2.10)$$

Substituting into the above expressions we find

$$[a, a^\dagger] = 1 \quad (2.11)$$

while the Hamiltonian is given by

$$\begin{aligned} H &= \frac{1}{2}\omega(aa^\dagger + a^\dagger a) \\ &= \omega(a^\dagger a + \frac{1}{2}) \end{aligned} \quad (2.12)$$

One can easily confirm that the commutators between the Hamiltonian and the creation and annihilation operators are given by

$$[H, a^\dagger] = \omega a^\dagger \quad \text{and} \quad [H, a] = -\omega a \quad (2.13)$$

These relations ensure that a and a^\dagger take us between energy eigenstates. Let $|E\rangle$ be an eigenstate with energy E , so that $H|E\rangle = E|E\rangle$. Then we can construct more eigenstates by acting with a and a^\dagger ,

$$Ha^\dagger|E\rangle = (E + \omega)a^\dagger|E\rangle, \quad Ha|E\rangle = (E - \omega)a|E\rangle \quad (2.14)$$

So we find that the system has a ladder of states with energies

$$\dots, E - \omega, E, E + \omega, E + 2\omega, \dots \quad (2.15)$$

If the energy is bounded below, there must be a *ground state* $|0\rangle$ which satisfies $a|0\rangle = 0$. This has ground state energy (also known as zero point energy),

$$H|0\rangle = \frac{1}{2}\omega|0\rangle \quad (2.16)$$

Excited states then arise from repeated application of a^\dagger ,

$$|n\rangle = (a^\dagger)^n|0\rangle \quad \text{with} \quad H|n\rangle = (n + \frac{1}{2})\omega|n\rangle \quad (2.17)$$

where I've ignored the normalization of these states so, $\langle n|n\rangle = 1$.

2.2 The Free Scalar Field

We now apply the quantization of the harmonic oscillator to the free scalar field. We write φ and π as a linear sum of an infinite number of creation and annihilation operators a_p^\dagger and $a_{p\rightarrow}$, indexed by the 3-momentum $p\rightarrow$,

$$\varphi(\rightarrow x) = \int \frac{d^3 p}{(2\pi)^3} \frac{\sqrt{1}}{2\omega_{p\rightarrow}} \frac{h}{\underline{\epsilon}} a_{p\rightarrow} e^{ip\rightarrow \cdot \rightarrow x} + a_{p\rightarrow}^\dagger e^{-ip\rightarrow \cdot \rightarrow x} \quad (2.18)$$

$$\pi(\rightarrow x) = \frac{h}{(2\pi)^3} \frac{\omega_{p\rightarrow}}{(-i)} \frac{a_{p\rightarrow}}{2} e^{ip\rightarrow \cdot \rightarrow x} - a_{p\rightarrow}^\dagger e^{-ip\rightarrow \cdot \rightarrow x} \quad (2.19)$$

Claim: The commutation relations for φ and π are equivalent to the following commutation relations for $a_{\vec{p} \rightarrow}$ and $a_{\vec{p} \rightarrow}^\dagger$

$$\begin{aligned} [\varphi(\vec{x}), \varphi(\vec{y})] &= [\pi(\vec{x}), \pi(\vec{y})] = 0 & [a_{\vec{p}' \rightarrow} a_{\vec{q} \rightarrow}] &= [a_{\vec{p} \rightarrow}^\dagger a_{\vec{q} \rightarrow}] = 0 \\ & [\varphi(\vec{x}), \pi(\vec{y})] = i\delta^{(3)}(\vec{x} - \vec{y}) & \Leftrightarrow & \\ & [a_{\vec{p} \rightarrow}, a_{\vec{q} \rightarrow}^\dagger] = (2\pi)^3 \delta^{(3)}(\vec{p} - \vec{q}) & & \end{aligned} \quad (2.20)$$

Proof: We'll show this just one way. Assume that $[a_{\vec{p} \rightarrow}, a_{\vec{p} \rightarrow}^\dagger] = (2\pi)^3 \delta^{(3)}(\vec{p} - \vec{q})$. Then

$$\begin{aligned} [\varphi(\vec{x}), \pi(\vec{y})] &= \int \frac{d^3 p d^3 q}{(2\pi)^6} \frac{(-i)}{2} \frac{\omega_{\vec{q}}}{\omega_{\vec{p}}} - [a_{\vec{p} \rightarrow}, a_{\vec{q} \rightarrow}^\dagger] e^{i\vec{p} \cdot (\vec{x} - \vec{y})} - [a_{\vec{q} \rightarrow}^\dagger, a_{\vec{p} \rightarrow}] e^{-i\vec{p} \cdot (\vec{x} + \vec{y})} \\ &= \frac{1}{(2\pi)^3} \frac{1}{2} - e^{i\vec{p} \cdot (\vec{x} - \vec{y})} - e^{i\vec{p} \cdot (\vec{y} - \vec{x})} \\ &= i\delta^{(3)}(\vec{x} - \vec{y}) \end{aligned} \quad (2.21)$$

The Hamiltonian

Let's now compute the Hamiltonian in terms of $a_{\vec{p} \rightarrow}$ and $a_{\vec{p} \rightarrow}^\dagger$. We have

$$\begin{aligned} H &= \frac{1}{2} \int d^3 x \pi^2 + (\nabla \varphi)^2 + m^2 \varphi^2 \\ &= \frac{1}{2} \frac{\int d^3 x d^3 p}{d^3 q} \frac{\sqrt{\omega_{\vec{p} \rightarrow} \omega_{\vec{q} \rightarrow}}}{\int d^3 p d^3 q} \frac{i\vec{p} \cdot \vec{x} + -i\vec{p} \cdot \vec{x} + i\vec{q} \cdot \vec{x} + -i\vec{q} \cdot \vec{x}}{2} (a_{\vec{p} \rightarrow} e^\dagger - a_{\vec{p} \rightarrow} e) (a_{\vec{q} \rightarrow} e^\dagger - a_{\vec{q} \rightarrow} e) \\ &\stackrel{(2\pi)^6}{=} \frac{1}{2} \frac{\sqrt{\frac{1}{2 \omega_{\vec{p} \rightarrow} \omega_{\vec{q} \rightarrow}}} (i\vec{p} \cdot \vec{x} - i\vec{p} \cdot \vec{x}^\dagger e^{-i\vec{p} \cdot \vec{x}}) \cdot (i\vec{q} \cdot \vec{x} - i\vec{q} \cdot \vec{x}^\dagger e^{-i\vec{q} \cdot \vec{x}})}{\int d^3 p} \\ &\quad + \frac{\sqrt{m^2}}{2 \omega} (a_{\vec{p} \rightarrow} e^\dagger + a_{\vec{p} \rightarrow}^\dagger e) (a_{\vec{q} \rightarrow} e^\dagger + a_{\vec{q} \rightarrow}^\dagger e) \\ &= \frac{1}{4} \frac{\int d^3 p \frac{h}{\omega_{\vec{p} \rightarrow}} (-\omega_{\vec{p} \rightarrow} + \vec{p} \cdot \vec{p} + m_2) (a_{\vec{p} \rightarrow} a_{-\vec{p} \rightarrow} + a_{\vec{p} \rightarrow}^\dagger a_{-\vec{p} \rightarrow}) + (\omega_{\vec{p} \rightarrow} + \vec{p} \cdot \vec{p} + m) (a_{\vec{p} \rightarrow} a_{\vec{p} \rightarrow}^\dagger + a_{\vec{p} \rightarrow} a_{\vec{p} \rightarrow})}{(2\pi)^3 \omega_{\vec{p} \rightarrow}} \end{aligned}$$

where in the second line we've used the expressions for φ and π given in (2.18) and (2.19); to get to the third line we've integrated over $d^3 x$ to get delta-functions $\delta^{(3)}(\vec{p} \pm \vec{q})$ which, in turn, allow us to perform the $d^3 q$ integral. Now using the expression for the frequency $\omega^2 = \vec{p}^2 + m^2$, the first term vanishes and we're left with

$$\begin{aligned} H &= \frac{1}{2} \frac{\int d^3 p \frac{h}{\omega_{\vec{p} \rightarrow}} a^\dagger + a^\dagger a}{(2\pi)^3} \vec{p} \cdot \vec{p} \\ &= \frac{1}{a} \frac{\int d^3 p \frac{h}{\omega_{\vec{p} \rightarrow}} a^\dagger}{(2\pi)^3} \frac{\frac{1}{2}(2\pi)^3 \delta^{(3)}(0)}{\vec{p} \cdot \vec{p}} \end{aligned} \quad (2.22)$$

Hmmmm. We've found a delta-function, evaluated at zero where it has its infinite spike. Moreover, the integral over $\omega_{p\rightarrow}$ diverges at large p . What to do? Let's start by looking at the ground state where this infinity first becomes apparent.

2.3 The Vacuum

Following our procedure for the harmonic oscillator, let's define the vacuum $|0\rangle$ by insisting that it is annihilated by *all* $a_{p\rightarrow}$,

$$a_{p\rightarrow}|0\rangle = 0 \quad \forall p\rightarrow \quad (2.23)$$

With this definition, the energy E_0 of the ground state comes from the second term in (2.22),

$$\stackrel{H}{=} |0\rangle \equiv E |0\rangle - \int_0^{\infty} d^3p \frac{1}{2} \omega_{p\rightarrow} \delta^{(3)}(0) |0\rangle = \infty |0\rangle \quad (2.24)$$

The subject of quantum field theory is rife with infinities. Each tells us something important, usually that we're doing something wrong, or asking the wrong question. Let's take some time to explore where this infinity comes from and how we should deal with it.

In fact there are two different ∞ 's lurking in the expression (2.24). The first arises because space is infinitely large. (Infinities of this type are often referred to as *infra-red divergences* although in this case the ∞ is so simple that it barely deserves this name). To extract out this infinity, let's consider putting the theory in a box with sides of length L . We impose periodic boundary conditions on the field. Then, taking the limit where $L \rightarrow \infty$, we get

$$\lim_{L \rightarrow \infty} (2\pi)^3 \delta^{(3)}(0) = \int_{-L/2}^{L/2} d^3x e^{i\mathbf{x}\cdot\mathbf{p}\rightarrow} = \lim_{L \rightarrow \infty} \int_{-L/2}^{L/2} d^3x = V \quad (2.25)$$

where V is the volume of the box. So the $\delta(0)$ divergence arises because we're computing the total energy, rather than the energy density E_0 . To find E_0 we can simply divide by the volume,

$$E_0 = \frac{E_0}{V} = \frac{\int d^3p}{(2\pi)^3} \frac{1}{2} \omega_{p\rightarrow} \quad (2.26)$$

which is still infinite. We recognize it as the sum of ground state energies for each harmonic oscillator. But $E_0 \rightarrow \infty$ due to the $|p\rightarrow| \rightarrow \infty$ limit of the integral. This is a high frequency — or short distance — infinity known as an *ultra-violet divergence*. This divergence arises because of our hubris. We've assumed that our theory is valid to arbitrarily short distance scales, corresponding to arbitrarily high energies. This is clearly absurd. The integral should be cut-off at high momentum in order to reflect the fact that our theory is likely to break down in some way.

We can deal with the infinity in (2.24) in a more practical way. In physics we're only interested in energy differences. There's no way to measure E_0 directly, so we can simply redefine the Hamiltonian by subtracting off this infinity,

$$H = \frac{d^3 p}{(2\pi)^3} \omega_{p \rightarrow} a^\dagger a \quad (2.27)$$

so that, with this new definition, $H |0\rangle = 0$. In fact, the difference between this Hamiltonian and the previous one is merely an ordering ambiguity in moving from the classical theory to the quantum theory. For example, if we defined the Hamiltonian of the harmonic oscillator to be $H = (1/2)(\omega q - ip)(\omega q + ip)$, which is classically the same as our original choice, then upon quantization it would naturally give $H = \omega a^\dagger a$ as in (2.27). This type of ordering ambiguity arises a lot in field theories. We'll come across a number of ways of dealing with it. The method that we've used above is called *normal ordering*.

Definition: We write the *normal ordered* string of operators $\varphi_1(\rightarrow x_1) \dots \varphi_n(\rightarrow x_n)$ as

$$:\varphi_1(\rightarrow x_1) \dots \varphi_n(\rightarrow x_n): \quad (2.28)$$

It is defined to be the usual product with all annihilation operators $a_{p \rightarrow}$ placed to the right. So, for the Hamiltonian, we could write (2.27) as

$$:H := \frac{d^3 p}{(2\pi)^3} \omega_{p \rightarrow} a^\dagger a \quad (2.29)$$

In the remainder of this section, we will normal order all operators in this manner.

2.3.1 The Cosmological Constant

Above I wrote "there's no way to measure E_0 directly". There is a BIG caveat here: gravity is supposed to see everything! The sum of all the zero point energies should contribute to the stress-energy tensor that appears on the right-hand side of Einstein's equations. We expect them to appear as a *cosmological constant* $\Lambda = E_0/V$,

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = -8\pi G T_{\mu\nu} + \Lambda g_{\mu\nu} \quad (2.30)$$

Current observation suggests that 70% of the energy density in the universe has the properties of a cosmological constant with $\Lambda \sim (10^{-3} eV)^4$. This is much smaller than other scales in particle physics. In particular, the Standard Model is valid at least up to $10^{12} eV$. Why don't the zero point energies of these fields contribute to Λ ? Or, if they do, what cancels them to such high accuracy? This is the cosmological constant problem. No one knows the answer!

2.3.2 The Casimir Effect

"I mentioned my results to Niels Bohr, during a walk. That is nice, he said, that is something new... and he mumbled something about zero-point energy."

Hendrik Casimir

Using the normal ordering prescription we can happily set $E_0 = 0$, while chanting the mantra that only energy differences can be measured. But we should be careful, for there is a situation where differences in the energy of vacuum fluctuations themselves can be measured.

To regulate the infra-red divergences, we'll make the x^1 direction periodic, with size L , and impose periodic boundary conditions such that

$$\varphi(\rightarrow x) = \varphi(\rightarrow x + L\rightarrow n) \quad (2.31)$$

with $\rightarrow n = (1, 0, 0)$. We'll leave y and z alone, but remember that we should compute all physical quantities per unit area A . We insert two reflecting plates, separated by a distance $d \ll L$ in the x^1 direction. The plates are such that they impose $\varphi(x) = 0$ at the position of the plates. The presence of these plates affects the Fourier decomposition of the field and, in particular, means that the momentum of the field inside the

plates is quantized as

$$p \rightarrow = \frac{n\pi}{d}, p_y, p_z \quad n \in \mathbf{Z}^+ \quad (2.32)$$

For a *massless* scalar field, the ground state energy between the plates is

$$\underline{E(d)} = \int_{-\infty}^{\infty} dp_y dp_z$$

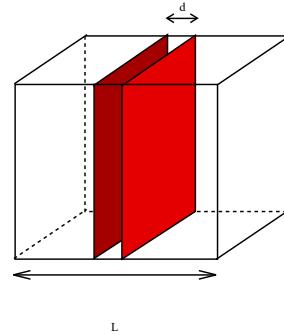


Figure 3:

$$A_{n=1} = \frac{1}{(2\pi)^2} \frac{1}{2} \sqrt{\frac{n\pi}{d}}^2 + p_y^2 + p_z^2 \quad (2.33)$$

while the energy outside the plates is $E(L - d)$. The total energy is therefore

$$E = E(d) + E(L - d) \quad (2.34)$$

which – at least naively – depends on d . If this naive guess is true, it would mean that there is a force on the plates due to the fluctuations of the vacuum. This is the Casimir force, first predicted in 1948 and observed 10 years later. In the real world, the effect is due to the vacuum fluctuations of the electromagnetic field, with the boundary conditions imposed by conducting plates. Here we model this effect with a scalar.

But there's a problem. E is infinite! What to do? The problem comes from the arbitrarily high momentum modes. We could regulate this in a number of different ways. Physically one could argue that any real plate cannot reflect waves of arbitrarily high frequency: at some point, things begin to leak. Mathematically, we want to find a way to neglect modes of momentum $p \gg a^{-1}$ for some distance scale $a \ll d$, known as the ultra-violet (UV) cut-off. One way to do this is to change the integral (2.33) to,

$$\frac{E(d)}{A} = \sum_{n=1}^{\infty} \int \frac{dp_y dp_z}{(2\pi)^2} \frac{1}{2} \frac{\Gamma(n\pi/2)}{d} e^{-\frac{(p_y + p_z)^2}{a^2}} \quad (2.35)$$

which has the property that as $a \rightarrow 0$, we regain the full, infinite, expression (2.33). However (2.35) is finite, and gives us something we can easily work with. Of course, we made it finite in a rather ad-hoc manner and we better make sure that any physical quantity we calculate doesn't depend on the UV cut-off a , otherwise it's not something we can really trust.

The integral (2.35) is do-able, but a little complicated. It's a lot simpler if we look at the problem in $d = 1 + 1$ dimensions, rather than $d = 3 + 1$ dimensions. We'll find that all the same physics is at play. Now the energy is given by

$$E_{1+1}(d) = \frac{\pi}{2d} \sum_{n=1}^{\infty} n \quad (2.36)$$

We now regulate this sum by introducing the UV cutoff a introduced above. This renders the expression finite, allowing us to start manipulating it thus,

$$\begin{aligned} E_{1+1}(d) &\rightarrow \frac{\pi}{2d} \sum_{n=1}^{\infty} n e^{-an\pi/d} \\ &= \frac{1}{2} \frac{\partial}{\partial a} \sum_{n=1}^{\infty} e^{-an\pi/d} \\ &= -\frac{1}{2} \frac{\partial}{\partial a} \frac{1}{1 - e^{-a\pi/d}} \\ &= \frac{2d}{2\pi a^2} \frac{e^{a\pi/d}}{24d} + O(a^2) \end{aligned} \quad (2.37)$$

where, in the last line, we've used the fact that $a \ll d$. We can now compute the full energy,

$$E_{1+1} = E_{1+1}(d) + E_{1+1}(L-d) = \frac{L}{2\pi a^2} - \frac{\pi}{24} \frac{1}{d} + \frac{1}{L-d} + O(a^2) \quad (2.38)$$

This is still infinite in the limit $a \rightarrow 0$, which is to be expected. However, the force is given by

$$\frac{\partial E_{1+1}}{\partial d} = \frac{\pi}{24d^2} + \dots \quad (2.39)$$

where the \dots include terms of size d/L and a/d . The key point is that as we remove both the regulators, and take $a \rightarrow 0$ and $L \rightarrow \infty$, the force between the plates remains finite. This is the Casimir force².

If we ploughed through the analogous calculation in $d = 3 + 1$ dimensions, and performed the integral (2.35), we would find the result

$$\frac{1}{A} \frac{\partial E}{\partial d} = \frac{\pi^2}{480d^4} \quad (2.40)$$

The true Casimir force is twice as large as this, due to the two polarization states of the photon.

2.4 Particles

Having dealt with the vacuum, we can now turn to the excitations of the field. It's easy to verify that

$$[H, a_{\vec{p}}^\dagger] = \omega_{\vec{p}} a_{\vec{p}}^\dagger \quad \text{and} \quad [H, a_{\vec{p}}] = -\omega_{\vec{p}} a_{\vec{p}} \quad (2.41)$$

which means that, just as for the harmonic oscillator, we can construct energy eigenstates by acting on the vacuum $|0\rangle$ with a^\dagger . Let

$$|p\rangle = {}_{\vec{p}} a^\dagger |0\rangle \quad (2.42)$$

This state has energy

$$H|p\rangle = \omega_{\vec{p}} |p\rangle \quad \text{with} \quad \omega_{\vec{p}}^2 = \vec{p}^2 + m^2 \quad (2.43)$$

But we recognize this as the relativistic dispersion relation for a particle of mass m and 3-momentum \vec{p} ,

$$E_{\vec{p}}^2 = \vec{p}^2 + m^2 \quad (2.44)$$

²The number 24 that appears in the denominator of the one-dimensional Casimir force plays a more famous role in string theory: the same calculation in that context is the reason the bosonic string lives in $26 = 24 + 2$ spacetime dimensions. (The +2 comes from the fact the string itself is extended in one space and one time dimension). You will need to attend next term's "String Theory" course to see what on earth this has to do with the Casimir force.

We interpret the state $| p \rightarrow \rangle$ as the momentum eigenstate of a single particle of mass m . To stress this, from now on we'll write $E_{p \rightarrow}$ everywhere instead of $\omega_{p \rightarrow}$. Let's check this particle interpretation by studying the other quantum numbers of $| p \rightarrow \rangle$. We may take the classical total momentum P^{\rightarrow} given in (1.46) and turn it into an operator. After normal ordering, it becomes

$$P^{\rightarrow} = \frac{\int}{a} d^3x \pi \nabla^{\rightarrow} \varphi = \frac{d^3 p}{(2\pi)^3} p^{\rightarrow} a^{\dagger} \quad (2.45)$$

Acting on our state $| p \rightarrow \rangle$ with P^{\rightarrow} , we learn that it is indeed an eigenstate,

$$P^{\rightarrow} | p \rightarrow \rangle = p^{\rightarrow} | p \rightarrow \rangle \quad (2.46)$$

telling us that the state $| p \rightarrow \rangle$ has momentum p^{\rightarrow} . Another property of $| p \rightarrow \rangle$ that we can study is its angular momentum. Once again, we may take the classical expression for the total angular momentum of the field (1.55) and turn it into an operator,

$$J^i = \epsilon^{ijk} \int d^3x (J^0)^{jk} \quad (2.47)$$

It's not hard to show that acting on the one-particle state with zero momentum, $J^i | p \rightarrow = 0 \rangle = 0$, which we interpret as telling us that the particle carries no internal angular momentum. In other words, quantizing a scalar field gives rise to a spin 0 particle.

Multi-Particle States, Bosonic Statistics and Fock Space

We can create multi-particle states by acting multiple times with a^{\dagger} 's. We interpret the state in which n a^{\dagger} 's act on the vacuum as an n -particle state,

$$| p_1 \rightarrow, \dots, p_n \rightarrow \rangle = a_{p_1 \rightarrow}^{\dagger} \dots a_{p_n \rightarrow}^{\dagger} | 0 \rangle \quad (2.48)$$

Because all the a^{\dagger} 's commute among themselves, the state is symmetric under exchange of any two particles. For example,

$$| p \rightarrow, q \rightarrow \rangle = | q \rightarrow, p \rightarrow \rangle \quad (2.49)$$

This means that the particles are *bosons*.

The full Hilbert space of our theory is spanned by acting on the vacuum with all possible combinations of a^{\dagger} 's,

$$| 0 \rangle, a_{p \rightarrow}^{\dagger} | 0 \rangle, a_{p \rightarrow}^{\dagger} a_{q \rightarrow}^{\dagger} | 0 \rangle, a_{p \rightarrow}^{\dagger} a_{q \rightarrow}^{\dagger} a_{r \rightarrow}^{\dagger} | 0 \rangle \dots \quad (2.50)$$

This space is known as a *Fock space*. The Fock space is simply the sum of the n -particle Hilbert spaces, for all $n \geq 0$. There is a useful operator which counts the number of particles in a given state in the Fock space. It is called the *number operator* N

$$N = \frac{d^3 p}{(2\pi)^3} a^\dagger_{\vec{p}} a_{\vec{p}} \quad (2.51)$$

and satisfies $N |p_1, \dots, p_n\rangle = n |p_1, \dots, p_n\rangle$. The number operator commutes with the Hamiltonian, $[N, H] = 0$, ensuring that particle number is conserved. This means that we can place ourselves in the n -particle sector, and stay there. This is a property of free theories, but will no longer be true when we consider interactions: interactions create and destroy particles, taking us between the different sectors in the Fock space.

Operator Valued Distributions

Although we're referring to the states $|p\rangle$ as "particles", they're not localized in space in any way — they are momentum eigenstates. Recall that in quantum mechanics the position and momentum eigenstates are not good elements of the Hilbert space since they are not normalizable (they normalize to delta-functions). Similarly, in quantum field theory neither the operators $\varphi(\vec{x})$, nor $a_{\vec{p}}$ are good operators acting on the Fock

space. This is because they don't produce normalizable states. For example,

$$\langle 0 | a_{\vec{p}}^\dagger a_{\vec{p}} | 0 \rangle = \langle p | p \rangle = (2\pi)^3 \delta(0) \text{ and } \langle 0 | \varphi(\vec{x}) \varphi(\vec{x}) | 0 \rangle = \langle \vec{x} | \vec{x} \rangle = \delta(0) \quad (2.52)$$

They are operator valued distributions, rather than functions. This means that although $\varphi(\vec{x})$ has a well defined vacuum expectation value, $\langle 0 | \varphi(\vec{x}) | 0 \rangle = 0$, the fluctuations of the operator at a fixed point are infinite, $\langle 0 | \varphi(\vec{x}) \varphi(\vec{x}) | 0 \rangle = \infty$. We can construct well defined operators by smearing these distributions over space. For example, we can create a wavepacket

$$|\phi\rangle = \int \frac{d^3 p}{(2\pi)^3} e^{-i\vec{p}\cdot\vec{x}} \phi(p) |p\rangle \quad (2.53)$$

which is partially localized in both position and momentum space. (A typical state might be described by the Gaussian $\phi(p) = \exp(-p^2/2m^2)$).

2.4.1 Relativistic Normalization

We have defined the vacuum $|0\rangle$ which we normalize as $\langle 0 | 0 \rangle = 1$. The one-particle states $|p\rangle = a_{\vec{p}}^\dagger |0\rangle$ then satisfy

$$\langle p | \vec{q} \rangle = (2\pi)^3 \delta^{(3)}(p - \vec{q}) \quad (2.54)$$

But is this Lorentz invariant? It's not obvious because we only have 3-vectors. What could go wrong? Suppose we have a Lorentz transformation

$$p^\mu \rightarrow (p^r)^\mu = \Lambda^\mu_\nu p^\nu \quad (2.55)$$

such that the 3-vector transforms as $p \rightarrow \rightarrow p^r$. In the quantum theory, it would be preferable if the two states are related by a unitary transformation,

$$|p\rangle \rightarrow |p^r\rangle = U(\Lambda) |p\rangle \quad (2.56)$$

This would mean that the normalizations of $|p\rangle$ and $|p^r\rangle$ are the same whenever p and p^r are related by a Lorentz transformation. But we haven't been at all careful with normalizations. In general, we could get

$$|p\rangle \rightarrow \lambda(p, p^r) |p^r\rangle \quad (2.57)$$

for some unknown function $\lambda(p, p^r)$. How do we figure this out? The trick is to look at an object which we know is Lorentz invariant. One such object is the identity operator on one-particle states (which is really the projection operator onto one-particle states). With the normalization (2.54) we know this is given by

$$1 = \int \frac{d^3 p}{(2\pi)^3} |p\rangle \langle p| \quad (2.58)$$

This operator is Lorentz invariant, but it consists of two terms: the measure $d^3 p$ and the projector $|p\rangle \langle p|$. Are these individually Lorentz invariant? In fact the answer is no.

Claim The Lorentz invariant measure is,

$$\int \frac{d^3 p}{2E_p} \quad (2.59)$$

Proof: $\int d^4 p$ is obviously Lorentz invariant. And the relativistic dispersion relation for a massive particle,

$$p_\mu p^\mu = m^2 \Rightarrow p_0^2 = E^2 = p^2 + m^2 \quad (2.60)$$

is also Lorentz invariant. Solving for p_0 , there are two branches of solutions: $p_0 = \pm E_p$. But the choice of branch is another Lorentz invariant concept. So piecing everything together, the following combination must be Lorentz invariant,

$$\int \frac{d^4 p \delta(p_0^2 - p^2 - m^2)}{2p_0} \quad (2.61)$$

which completes the proof.

From this result we can figure out everything else. For example, the Lorentz invariant δ -function for 3-vectors is

$$2E_{p\rightarrow} \delta^{(3)}(\vec{p} \rightarrow \vec{q}) \quad (2.62)$$

which follows because

$$\int \frac{d^3p}{2E_{p\rightarrow}} 2E_{p\rightarrow} \delta^{(3)}(\vec{p} \rightarrow \vec{q}) = 1 \quad (2.63)$$

So finally we learn that the relativistically normalized momentum states are given by

$$|p\rangle = \sqrt{\frac{1}{2E_{p\rightarrow}}} |p\rightarrow\rangle = \sqrt{\frac{1}{2E_{p\rightarrow}}} a^\dagger |0\rangle \quad (2.64)$$

Notice that our notation is rather subtle: the relativistically normalized momentum state $|p\rangle$ differs from $|\vec{p}\rangle$ by the factor $\sqrt{\frac{1}{2E_{p\rightarrow}}}$. These states now satisfy

$$\langle p| q\rangle = (2\pi)^3 \frac{1}{2E_{p\rightarrow}} \delta^{(3)}(\vec{p} \rightarrow \vec{q}) \quad (2.65)$$

Finally, we can rewrite the identity on one-particle states as

$$1 = \frac{d^3p}{(2\pi)^3} \frac{1}{2E_{p\rightarrow}} |p\rangle \langle p| \quad (2.66)$$

Some texts also define relativistically normalized creation operators by $a^\dagger(p) = \sqrt{\frac{1}{2E_{p\rightarrow}}} a^\dagger$. We won't make use of this notation here.

2.5 Complex Scalar Fields

Consider a complex scalar field $\psi(x)$ with Lagrangian

$$L = \partial_\mu \psi^\dagger \partial^\mu \psi - M^2 \psi^\dagger \psi \quad (2.67)$$

Notice that, in contrast to the Lagrangian (1.7) for a real scalar field, there is no factor of $1/2$ in front of the Lagrangian for a complex scalar field. If we write ψ in terms of real scalar fields by $\psi = (\varphi_1 + i\varphi_2)/\sqrt{2}$, we get the factor of $1/2$ coming from the $1/\sqrt{2}$'s. The equations of motion are

$$\begin{aligned} \partial_\mu \partial^\mu \psi + M^2 \psi &= 0 \\ \partial_\mu \partial^\mu \psi^\dagger + M^2 \psi^\dagger &= 0 \end{aligned} \quad (2.68)$$

where the second equation is the complex conjugate of the first. We expand the complex field operator as a sum of plane waves as

$$\begin{aligned} \psi &= \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{\frac{2E_{p\rightarrow}}{(2\pi)^3}}} b_{p\rightarrow} e^{+ip\cdot x} + c_{p\rightarrow}^\dagger e^{-ip\cdot x} \\ \psi^\dagger &= \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{\frac{2E_{p\rightarrow}}{(2\pi)^3}}} b_{p\rightarrow}^\dagger e^{-ip\cdot x} + c_{p\rightarrow} e^{+ip\cdot x} \end{aligned} \quad (2.69)$$

Since the classical field ψ is not real, the corresponding quantum field ψ is not hermitian. This is the reason that we have different operators b and c^\dagger appearing in the positive and negative frequency parts. The classical field momentum is $\pi = \partial L / \partial \dot{\psi} = \dot{\psi}$. We also turn this into a quantum operator field which we write as,

$$\begin{aligned}\pi &= i \int \frac{d^3 p}{(2\pi)^3} \frac{E_{p \rightarrow}}{2\pi} b_{p \rightarrow} e^{-ip \cdot x} - c_{p \rightarrow} e^{+ip \cdot x} \\ \pi^\dagger &= \int \frac{d^3 p}{(2\pi)^3} (-i) \frac{E_{p \rightarrow}}{2} b_{p \rightarrow} e^{+ip \cdot x} - c_{p \rightarrow}^\dagger e^{-ip \cdot x}\end{aligned}\quad (2.70)$$

The commutation relations between fields and momenta are given by

$$[\psi(\rightarrow x), \pi(\rightarrow y)] = i\delta^{(3)}(\rightarrow x - \rightarrow y) \text{ and } [\psi(\rightarrow x), \pi^\dagger(\rightarrow y)] = 0 \quad (2.71)$$

together with others related by complex conjugation, as well as the usual $[\psi(\rightarrow x), \psi(\rightarrow y)] = [\psi(\rightarrow x), \psi^\dagger(\rightarrow y)] = 0$, etc. One can easily check that these field commutation relations are equivalent to the commutation relations for the operators $b_{p \rightarrow}$ and $c_{p \rightarrow}$,

$$\begin{aligned}[b_{p \rightarrow}, b_{q \rightarrow}^\dagger] &= (2\pi)^3 \delta^{(3)}(\rightarrow p - \rightarrow q) \\ [c_{p \rightarrow}, c_{q \rightarrow}^\dagger] &= (2\pi)^3 \delta^{(3)}(\rightarrow p - \rightarrow q)\end{aligned}\quad (2.72)$$

and

$$[b_{p \rightarrow}, b_{q \rightarrow}] = [c_{p \rightarrow}, c_{q \rightarrow}] = [b_{p \rightarrow}, c_{q \rightarrow}] = [b_{p \rightarrow}, c_{q \rightarrow}^\dagger] = 0 \quad (2.73)$$

In summary, quantizing a complex scalar field gives rise to two creation operators, b_p^\dagger and c_q^\dagger . These have the interpretation of creating two types of particle, both of mass M and both spin zero. They are interpreted as particles and anti-particles. In contrast,

for a real scalar field there is only a single type of particle: for a real scalar field, the particle is its own antiparticle.

Recall that the theory (2.67) has a classical conserved charge

$$Q = i \int d^3x (\psi^\wedge \psi - \psi^\wedge \psi) = i \int d^3x (\pi \psi - \psi^\wedge \pi^\wedge) \quad (2.74)$$

After normal ordering, this becomes the quantum operator

$$Q = \frac{i}{(2\pi)} \int d^3p (c_{p \rightarrow}^\dagger c_{p \rightarrow} - b_{p \rightarrow}^\dagger b_{p \rightarrow}) = N_c - N_b \quad (2.75)$$

so Q counts the number of anti-particles (created by c^\dagger) minus the number of particles (created by b^\dagger). We have $[H, Q] = 0$, ensuring that Q is a conserved quantity in the quantum theory. Of course, in our free field theory this isn't such a big deal because both N_c and N_b are separately conserved. However, we'll soon see that in interacting theories Q survives as a conserved quantity, while N_c and N_b individually do not.

2.6 The Heisenberg Picture

Although we started with a Lorentz invariant Lagrangian, we slowly butchered it as we quantized, introducing a preferred time coordinate t . It's not at all obvious that the theory is still Lorentz invariant after quantization. For example, the operators $\varphi(\rightarrow x)$ depend on space, but not on time. Meanwhile, the one-particle states evolve in time by Schrödinger's equation,

$$\frac{d}{dt} \langle p\rightarrow(t) \rangle = H \langle p\rightarrow(t) \rangle \Rightarrow \langle p\rightarrow(t) \rangle = e^{-iEt} \langle p\rightarrow \rangle \quad (2.76)$$

Things start to look better in the Heisenberg picture where time dependence is assigned to the operators O ,

$$O_H = e^{iHt} O_S e^{-iHt} \quad (2.77)$$

so that

$$\frac{dO_H}{dt} = i[H, O_H] \quad (2.78)$$

where the subscripts S and H tell us whether the operator is in the Schrödinger or Heisenberg picture. In field theory, we drop these subscripts and we will denote the picture by specifying whether the fields depend on space $\varphi(\rightarrow x)$ (the Schrödinger picture) or spacetime $\varphi(\rightarrow x, t) = \varphi(x)$ (the Heisenberg picture).

The operators in the two pictures agree at a fixed time, say, $t = 0$. The commutation relations (2.2) become equal time commutation relations in the Heisenberg picture,

$$\begin{aligned} [\varphi(\rightarrow x, t), \varphi(\rightarrow y, t)] &= [\pi(\rightarrow x, t), \pi(\rightarrow y, t)] = 0 \\ [\varphi(\rightarrow x, t), \pi(\rightarrow y, t)] &= i\delta^{(3)}(\rightarrow x - \rightarrow y) \end{aligned} \quad (2.79)$$

Now that the operator $\varphi(x) = \varphi(\rightarrow x, t)$ depends on time, we can start to study how it evolves. For example, we have

$$\begin{aligned} \dot{\varphi} &= i[H, \varphi] = -[\partial_y \pi(y)^3 + \nabla \varphi(y)^2 + m^2 \varphi(y)^2, \varphi(x)] \\ &= i \int_0^2 d^3y \pi(y) (-i) \delta^{(3)}(\rightarrow y - \rightarrow x) = \pi(x) \end{aligned} \quad (2.80)$$

Meanwhile, the equation of motion for π reads,

$$\dot{\pi} = i[H, \pi] = \frac{i}{2} [d^3y \pi(y)^2 + \nabla \varphi(y)^2 + m^2 \varphi(y)^2, \pi(x)]$$

$$\begin{aligned}
&= - \frac{i}{2} \int d^3y (\nabla [\varphi(y), \pi(x)]) \nabla \varphi(y) + \nabla \varphi(y) \nabla [\varphi(y), \pi(x)] \\
&\quad + 2im^2 \varphi(y) \delta^{(3)}(\rightarrow x - \rightarrow y) \\
&= - \int d^3y \nabla_y \delta^{(3)}(\rightarrow x - \rightarrow y) \nabla_y \varphi(y) - m^2 \varphi(x) \\
&= \nabla^2 \varphi - m^2 \varphi
\end{aligned} \tag{2.81}$$

where we've included the subscript y on ∇_y when there may be some confusion about which argument the derivative is acting on. To reach the last line, we've simply integrated by parts. Putting (2.80) and (2.81) together we find that the field operator φ satisfies the Klein-Gordon equation

$$\partial_\mu \partial^\mu \varphi + m^2 \varphi = 0 \tag{2.82}$$

Things are beginning to look more relativistic. We can write the Fourier expansion of $\varphi(x)$ by using the definition (2.77) and noting,

$$e^{iHt} a_{p\rightarrow} e^{-iHt} = e^{-iE_{p\rightarrow} t} a_{p\rightarrow} \quad \text{and} \quad e^{iHt} a_{p\rightarrow}^\dagger e^{-iHt} = e^{+iE_{p\rightarrow} t} a_{p\rightarrow}^\dagger \tag{2.83}$$

which follows from the commutation relations $[H, a_{p\rightarrow}] = -E_{p\rightarrow} a_{p\rightarrow}$ and $[H, a_{p\rightarrow}^\dagger] = +E_{p\rightarrow} a_{p\rightarrow}^\dagger$.

This then gives,

$$\varphi(\rightarrow x, t) = \frac{\int d^3p}{(2\pi)^3} \frac{1}{\sqrt{2E_{p\rightarrow}}} a_{p\rightarrow} e^{-ip \cdot x} + a_{p\rightarrow}^\dagger e^{+ip \cdot x} \tag{2.84}$$

which looks very similar to the previous expansion (2.18) except that the exponent is now written in terms of 4-vectors, $p \cdot x = E_{p\rightarrow} t - p_{\rightarrow} \cdot \rightarrow x$. (Note also that a sign has flipped in the exponent due to our Minkowski metric contraction). It's simple to check that (2.84) indeed satisfies the Klein-Gordon equation (2.82).

2.6.1 Causality

We're approaching something Lorentz invariant in the Heisenberg picture, where $\varphi(x)$ now satisfies the Klein-Gordon equation. But there's still a hint of non-Lorentz invariance because φ and π satisfy *equal time* commutation relations,

$$[\varphi(\rightarrow x, t), \pi(\rightarrow y, t)] = i\delta^{(3)}(\rightarrow x - \rightarrow y) \tag{2.85}$$

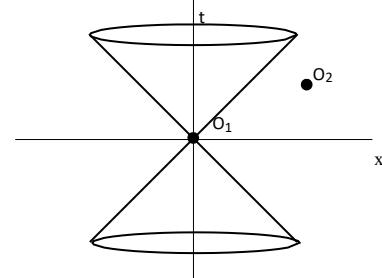


Figure 4:

But what about arbitrary spacetime separations? In particular, for our theory to be *causal*, we must require that all spacelike separated operators commute,

$$[O_1(x), O_2(y)] = 0 \quad \forall (x - y)^2 < 0 \quad (2.86)$$

This ensures that a measurement at x cannot affect a measurement at y when x and y are not causally connected. Does our theory satisfy this crucial property? Let's define

$$\Delta(x - y) = [\varphi(x), \varphi(y)] \quad (2.87)$$

The objects on the right-hand side of this expression are operators. However, it's easy to check by direct substitution that the left-hand side is simply a c-number function with the integral expression

$$\Delta(x - y) = \int \frac{-d^3 p}{(2\pi)^3} \frac{1}{2E_p} e^{-ip \cdot (x-y)} - e^{ip \cdot (x-y)} \quad (2.88)$$

\Rightarrow

What do we know about this function?

- It's Lorentz invariant, thanks to the appearance of the Lorentz invariant measure $d^3 p / 2E_{p\rightarrow}$ that we introduced in (2.59).
- It doesn't vanish for timelike separation. For example, taking $x - y = (t, 0, 0, 0)$ gives $[\varphi(\rightarrow x, 0), \varphi(\rightarrow x, t)] \sim e^{-imt} - e^{+imt}$.
- It vanishes for space-like separations. This follows by noting that $\Delta(x - y) = 0$ at equal times for all $(x - y)^2 = -(\rightarrow x - \rightarrow y)^2 < 0$, which we can see explicitly by writing

$$[\varphi(\rightarrow x, t), \varphi(\rightarrow y, t)] = \int \frac{d^3 p}{(2\pi)^3} \frac{1}{\sqrt{\rightarrow p^2 + m^2}} e^{i\vec{p} \cdot (\rightarrow x - \rightarrow y)} - e^{-i\vec{p} \cdot (\rightarrow x - \rightarrow y)} \quad (2.89)$$

and noticing that we can flip the sign of \vec{p} in the last exponent as it is an integration variable. But since $\Delta(x - y)$ is Lorentz invariant, it can only depend on $(x - y)^2$ and must therefore vanish for all $(x - y)^2 < 0$.

We therefore learn that our theory is indeed causal with commutators vanishing outside the lightcone. This property will continue to hold in interacting theories; indeed, it is usually given as one of the axioms of local quantum field theories. I should mention however that the fact that $[\varphi(x), \varphi(y)]$ is a c-number function, rather than an operator, is a property of free fields only.

2.7 Propagators

We could ask a different question to probe the causal structure of the theory. Prepare a particle at spacetime point y . What is the amplitude to find it at point x ? We can calculate this:

$$\begin{aligned} \langle 0 | \varphi(x) \varphi(y) | 0 \rangle &= \int \frac{d^3 p}{(2\pi)^3} \frac{d^3 p'}{(2\pi)^3} \sqrt{\frac{1}{4E_p E_{p'}}} \langle 0 | a_{p \rightarrow}^\dagger a_{p' \rightarrow} | 0 \rangle e^{-ip \cdot x + ip' \cdot y} \\ &= \frac{1}{2E} \theta^{-(x-y)} \equiv D(x-y) \end{aligned} \quad (2.90)$$

The function $D(x-y)$ is called the *propagator*. For spacelike separations, $(x-y)^2 < 0$, one can show that $D(x-y)$ decays like

$$D(x-y) \sim e^{-m|x-y|} \quad (2.91)$$

So it decays exponentially quickly outside the lightcone but, nonetheless, is non-vanishing! The quantum field appears to leak out of the lightcone. Yet we've just seen that spacelike measurements commute and the theory is causal. How do we reconcile these two facts? We can rewrite the calculation (2.89) as

$$[\varphi(x), \varphi(y)] = D(x-y) - D(y-x) = 0 \text{ if } (x-y)^2 < 0 \quad (2.92)$$

There are words you can drape around this calculation. When $(x-y)^2 < 0$, there is no Lorentz invariant way to order events. If a particle can travel in a spacelike direction from $x \rightarrow y$, it can just as easily travel from $y \rightarrow x$. In any measurement, the amplitudes for these two events cancel.

With a complex scalar field, it is more interesting. We can look at the equation $[\psi(x), \psi^\dagger(y)] = 0$ outside the lightcone. The interpretation now is that the amplitude for the particle to propagate from $x \rightarrow y$ cancels the amplitude for the *antiparticle* to travel from $y \rightarrow x$. In fact, this interpretation is also there for a real scalar field because the particle is its own antiparticle.

2.7.1 The Feynman Propagator

As we will see shortly, one of the most important quantities in interacting field theory is the *Feynman propagator*,

$$\Delta_F(x-y) = \langle 0 | T\varphi(x)\varphi(y) | 0 \rangle = \begin{cases} D(x-y) & x^0 > y^0 \\ D(y-x) & y^0 > x^0 \end{cases} \quad (2.93)$$

where T stands for time ordering, placing all operators evaluated at later times to the left so,

$$T\varphi(x)\varphi(y) = \begin{cases} \varphi(x)\varphi(y) & x^0 > y^0 \\ \varphi(y)\varphi(x) & y^0 > x^0 \end{cases} \quad (2.94)$$

Claim: There is a useful way of writing the Feynman propagator in terms of a 4-momentum integral that shows that it is explicitly Lorentz invariant

$$\Delta_F(x - y) = \frac{d^4 p}{(2\pi)^4} \frac{i}{p^2 - m^2} e^{-ip \cdot (x-y)} \quad (2.95)$$

Notice that this is the first time in this course that we've integrated over 4-momentum. Until now, we integrated only over 3-momentum, with p^0 fixed by the mass-shell condition to be $p^0 = E_{p\rightarrow}$. In the expression (2.95) for Δ_F , we have no such condition on p^0 . However, as it stands this integral is ill-defined because, for each value of $p\rightarrow$, the denominator $p^2 - m^2 = (p^0)^2 - p\rightarrow^2 - m^2$ produces a pole when $p^0 = \pm E_{p\rightarrow} = \pm \sqrt{p\rightarrow^2 + m^2}$.

We need a prescription for avoiding these singularities in the p_0 integral. To get the Feynman propagator, we must choose the contour to be

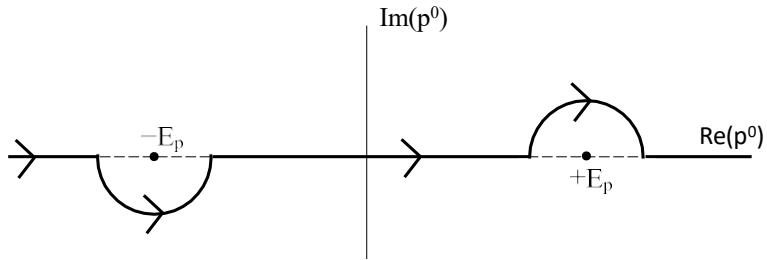


Figure 5: The contour for the Feynman propagator.

Proof:

$$\frac{1}{p^2 - m^2} = \frac{1}{(p^0)^2 - E_{p\rightarrow}^2} = \frac{1}{(p^0 - E_{p\rightarrow})(p^0 + E_{p\rightarrow})} \quad (2.96)$$

so the residue of the pole at $p^0 = \pm E_{p\rightarrow}$ is $\pm 1/2E_{p\rightarrow}$. When $x^0 > y^0$, we close the contour in the lower half plane, where $p^0 \rightarrow -i\infty$, ensuring that the integrand vanishes since $e^{-ip^0(x^0-y^0)} \rightarrow 0$. The integral over p^0 then picks up the residue at $p^0 = +E_{p\rightarrow}$ which is $-2\pi i/2E_{p\rightarrow}$ where the minus sign arises because we took a clockwise contour. Hence when $x^0 > y^0$ we have

$$\Delta_F(x - y) = \int \frac{d^3 p}{(2\pi)^4} \frac{-2\pi i}{2E_{p\rightarrow}} i e^{-iE_{p\rightarrow}(x^0-y^0)+ip\cdot(x-y)}$$

$$= \frac{d^3 p}{(2\pi)^3} \frac{1}{2E} e^{-ip \cdot (x-y)} = D(x-y) \quad (2.97)$$

which is indeed the Feynman propagator for $x^0 > y^0$. In contrast, when $y^0 > x^0$, we close the contour in an anti-clockwise direction in the upper half plane to get,

$$\frac{\Delta}{\pi^4} \left(2 \int_0^\infty i \frac{e^{+iE_p(x^0-y)}}{p^2 - (2E)^2} \right) =$$

$$\int d^3 p \frac{1}{2E} e^{-ip \cdot (y-x)} = D(y-x) \quad (2.98)$$

where to go to from the second line to the third, we have flipped the sign of $p \rightarrow$ which is valid since we integrate over $d^3 p$ and all other quantities depend only on $p \rightarrow^2$. Once again we reproduce the Feynman propagator.

Instead of specifying the contour, it is standard to write

the Feynman propagator as

$$\frac{d p}{(2\pi)^4} \frac{i e}{p^2 - m^2 + i\epsilon} = D_F(x-y) \quad (2.99)$$

$i\epsilon$

(2.99)

with $\epsilon > 0$, and infinitesimal. This has the effect of shifting the poles slightly off the real axis, so the integral along the real p^0 axis is equivalent to the contour shown in Figure 5. This way of writing the propagator is, for obvious reasons, called the “ $i\epsilon$ prescription”.

**Figure
6:**

2.7.2 Green's Functions

There is another avatar of the propagator: it is a Green's function for the Klein-Gordon operator. If we stay away from the singularities, we have

$$\begin{aligned}
 (\partial_t - \nabla^2 + m^2) \Delta_F(x-y) &= \frac{i}{(2\pi)^4 p^2 \int d^4 p} e^{-ip \cdot (x-y)} \\
 &= -i \frac{(2\pi)^4}{p^2} \frac{e^{-ip \cdot (x-y)}}{4\pi^2} \\
 &= -i \frac{\delta^{(4)}(x-y)}{4\pi^2}
 \end{aligned}
 \tag{2.100}$$

Note that we didn't make use of the contour anywhere in this derivation. For some purposes it is also useful to pick other contours which also give rise to Green's functions.

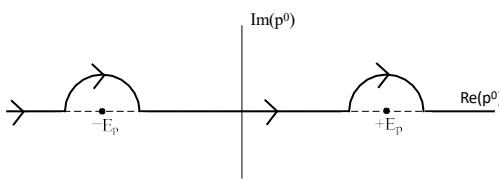


Figure 7: The retarded contour

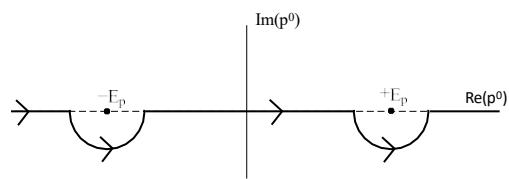


Figure 8: The advanced contour

For example, the *retarded* Green's function $\Delta_R(x - y)$ is defined by the contour shown in Figure 7 which has the property

$$\Delta_R(x - y) = \begin{cases} D(x - y) - D(y - x) & x^0 > y^0 \\ 0 & y^0 > x^0 \end{cases} \quad (2.101)$$

The retarded Green's function is useful in classical field theory if we know the initial value of some field configuration and want to figure out what it evolves into in the presence of a source, meaning that we want to know the solution to the inhomogeneous Klein-Gordon equation,

$$\partial_\mu \partial^\mu \varphi + m^2 \varphi = J(x) \quad (2.102)$$

for some fixed background function $J(x)$. Similarly, one can define the *advanced* Green's function $\Delta_A(x - y)$ which vanishes when $y^0 < x^0$, which is useful if we know the end point of a field configuration and want to figure out where it came from. Given that next term's course is called "Advanced Quantum Field Theory", there is an obvious name for the current course. But it got shot down in the staff meeting. In the quantum theory, we will see that the Feynman Green's function is most relevant.

2.8 Non-Relativistic Fields

Let's return to our classical complex scalar field obeying the Klein-Gordon equation. We'll decompose the field as

$$\psi(\rightarrow x, t) = e^{-imt} \tilde{\psi}(\rightarrow x, t) \quad (2.103)$$

Then the KG-equation reads

$$\partial_t^2 \psi - \nabla^2 \psi + m^2 \psi = e^{-imt} \left(\frac{\tilde{\psi}}{\psi} - 2im\tilde{\psi} - \nabla^2 \right) \psi = 0 \quad (2.104)$$

with the m^2 term cancelled by the time derivatives. The non-relativistic limit of a

particle is $| p \rightarrow -m \rangle$. Let's look at what this does to our field. After a Fourier

this is equivalent to saying that $|\tilde{\psi}| m |\tilde{\psi}|$. In this limit, we drop the term with two time derivatives and the KG equation becomes,

$$i\frac{\partial \tilde{\psi}}{\partial t} = \frac{1}{2m} \nabla^2 \tilde{\psi} \quad (2.105)$$

This looks very similar to the Schrödinger equation for a non-relativistic free particle of mass m . Except it doesn't have any probability interpretation — it's simply a classical field evolving through an equation that's first order in time derivatives.

We wrote down a Lagrangian in section [1.1.2](#) which gives rise to field equations which are first order in time derivatives. In fact, we can derive this from the relativistic Lagrangian for a scalar field by again taking the limit $\partial_t \tilde{\psi} \rightarrow m\psi$. After losing the tilde, so $\tilde{\psi} \rightarrow \psi$, the non-relativistic Lagrangian becomes

$$L = +i\psi \dot{\psi} - \frac{1}{2m} \nabla \psi \cdot \nabla \psi \quad (2.106)$$

where we've divided by $1/2m$. This Lagrangian has a conserved current arising from the internal symmetry $\psi \rightarrow e^{i\alpha} \psi$. The current has time and space components

$$j^\mu = -\psi^\lambda \psi_\lambda, \frac{i}{2m} (\nabla^\lambda \psi_\lambda - \psi_\lambda \nabla^\lambda) \quad (2.107)$$

To move to the Hamiltonian formalism we compute the momentum

$$\pi = \frac{\partial L}{\partial \dot{\psi}} = i\psi^\lambda \quad (2.108)$$

This means that the momentum conjugate to ψ is $i\psi^\lambda$. The momentum does not depend on time derivatives at all! This looks a little disconcerting but it's fully consistent for a theory which is first order in time derivatives. In order to determine the full trajectory of the field, we need only specify ψ and ψ^λ at time $t = 0$: no time derivatives on the initial slice are required.

Since the Lagrangian already contains a “ pq ” term (instead of the more familiar $\frac{1}{2}pq$ term), the time derivatives drop out when we compute the Hamiltonian. We get,

$$H = \frac{1}{2m} \nabla \psi^\lambda \nabla \psi_\lambda \quad (2.109)$$

To quantize we impose (in the Schrödinger picture) the canonical commutation relations

$$[\psi(\vec{x}), \psi(\vec{y})] = [\psi^\dagger(\vec{x}), \psi^\dagger(\vec{y})] = 0$$

$$[\psi(\vec{x}), \psi^\dagger(\vec{y})] = \delta^{(3)}(\vec{x} - \vec{y}) \quad (2.110)$$

We may expand $\psi(\rightarrow x)$ as a Fourier transform

$$\psi(\rightarrow x) = \frac{d^3 p}{(2\pi)^3} \sum_{\vec{p}} a_{\vec{p}} e^{i\vec{p}\cdot\vec{x}} \quad (2.111)$$

where the commutation relations (2.110) require

$$[a_{\vec{p}}, a_{\vec{q}}^\dagger] = (2\pi)^3 \delta^{(3)}(\vec{p} - \vec{q}) \quad (2.112)$$

The vacuum satisfies $|0\rangle = 0$, and the excitations are $a_{\vec{p}}^\dagger \dots a_{\vec{p}_n}^\dagger |0\rangle$. The one-particle states have energy

$$H | \vec{p} \rangle = \frac{\vec{p}^2}{2m} | \vec{p} \rangle \quad (2.113)$$

which is the non-relativistic dispersion relation. We conclude that quantizing the first order Lagrangian (2.106) gives rise to non-relativistic particles of mass m . Some comments:

- We have a complex field but only a single type of particle. The anti-particle is not in the spectrum. The existence of anti-particles is a consequence of relativity.
- A related fact is that the conserved charge $Q = \int d^3x : \psi^\dagger \psi :$ is the particle number. This remains conserved even if we include interactions in the Lagrangian of the form

$$\Delta L = V(\psi^\dagger \psi) \quad (2.114)$$

So in non-relativistic theories, particle number is conserved. It is only with relativity, and the appearance of anti-particles, that particle number can change.

- There is no non-relativistic limit of a real scalar field. In the relativistic theory, the particles are their own anti-particles, and there can be no way to construct a multi-particle theory that conserves particle number.

2.8.1 Recovering Quantum Mechanics

In quantum mechanics, we talk about the position and momentum operators \mathbf{x}^\rightarrow and \mathbf{P}^\rightarrow . In quantum field theory, position is relegated to a label. How do we get back to quantum mechanics? We already have the operator for the total momentum of the field

$$\mathbf{P}^\rightarrow = \frac{d^3 p}{(2\pi)^3} \sum_{\vec{p}} \vec{p} a_{\vec{p}}^\dagger a_{\vec{p}} \quad (2.115)$$

which, on one-particle states, gives $P^\rightarrow | p \rightarrow \rangle = p \rightarrow | p \rightarrow \rangle$. It's also easy to construct the position operator. Let's work in the non-relativistic limit. Then the operator

$$\psi^\dagger(\rightarrow x) = \frac{d^3 p}{(2\pi)^3} \int \frac{a^\dagger_{p \rightarrow} - e}{\rightarrow x - p \rightarrow} \quad (2.116)$$

creates a particle with δ -function localization at $\rightarrow x$. We write $| \rightarrow x \rangle = \psi^\dagger(\rightarrow x) | 0 \rangle$. A natural position operator is then

$$X^\rightarrow = \int d^3 x \rightarrow x \psi^\dagger(\rightarrow x) \psi(\rightarrow x) \quad (2.117)$$

so that $X^\rightarrow | \rightarrow x \rangle = \rightarrow x | \rightarrow x \rangle$.

Let's now construct a state $|\phi\rangle$ by taking superpositions of one-particle states $|\rightarrow x\rangle$,

$$|\phi\rangle = \int d^3 x \phi(\rightarrow x) |\rightarrow x\rangle \quad (2.118)$$

The function $\phi(\rightarrow x)$ is what we would usually call the Schrödinger wavefunction (in the position representation). Let's make sure that it indeed satisfies all the right properties.

Firstly, it's clear that acting with the position operator X^\rightarrow has the right action of $\phi(\rightarrow x)$,

$$X^i |\phi\rangle = \int d^3 x x^i \phi(\rightarrow x) |\rightarrow x\rangle \quad (2.119)$$

but what about the momentum operator P^\rightarrow ? We will now show that

$$P^i |\phi\rangle = \int d^3 x -i \frac{\partial \phi}{\partial x^i} |\rightarrow x\rangle \quad (2.120)$$

which tells us that P^i acts as the familiar derivative on wavefunctions $|\phi\rangle$. To see that this is the case, we write

$$\begin{aligned} P^i |\phi\rangle &= \int \frac{d^3 x d^3 p}{(2\pi)^3} \frac{a^\dagger_{p \rightarrow} - a_{p \rightarrow}}{p \cdot a^\dagger_{p \rightarrow} - a_{p \rightarrow}} \phi(\rightarrow x) \psi^\dagger(\rightarrow x) | 0 \rangle \\ &\stackrel{!}{=} \int \frac{(2\pi)^3}{(2\pi)^3} \frac{d^3 x d^3 p}{p \cdot a^\dagger_{p \rightarrow} - a_{p \rightarrow}} e^{-ip \cdot \rightarrow x} \phi(\rightarrow x) | 0 \rangle \end{aligned} \quad (2.121)$$

where we've used the relationship $[a_{p \rightarrow}, \psi^\dagger(\rightarrow x)] = e^{-ip \cdot \rightarrow x}$ which can be easily checked. Proceeding with our calculation, we have

$$\begin{aligned} P^i |\phi\rangle &= \int d^3 x d^3 p \left[i \frac{\partial}{\partial p} e^{-ip \cdot \rightarrow x} \phi(\rightarrow x) \right] | 0 \rangle \\ &\stackrel{!}{=} \int \frac{(2\pi)^3}{(2\pi)^3} \frac{d^3 x d^3 p}{p \cdot a^\dagger_{p \rightarrow} - a_{p \rightarrow}} \frac{\partial \phi}{\partial x^i} e^{-ip \cdot \rightarrow x} | 0 \rangle \\ &= \frac{(2\pi)^3}{(2\pi)^3} \frac{e^{-ip \cdot \rightarrow x}}{\partial x^i} \frac{-i}{p \cdot a^\dagger_{p \rightarrow}} | 0 \rangle \\ &= \int d^3 x -i \frac{\partial \phi}{\partial x^i} |\rightarrow x\rangle \end{aligned} \quad (2.122)$$

which confirms (2.120). So we learn that when acting on one-particle states, the operators X^j and P^j act as position and momentum operators in quantum mechanics, with

$[X^i, P^j] |\phi\rangle = i\delta^{ij} |\phi\rangle$. But what about dynamics? How does the wavefunction $\phi(\rightarrow x, t)$ change in time? The Hamiltonian (2.109) can be rewritten as

$$H = \int_{-\infty}^{\infty} \frac{1}{2m} \nabla \psi^\dagger \nabla \psi - \int_{-\infty}^{\infty} \frac{i\hbar p^2}{2m} + \int_{-\infty}^{\infty} \frac{d^3p}{(2\pi)^3} \frac{p^2}{2m}$$

so we find that

$$i\frac{\partial \phi}{\partial t} = -\frac{1}{2m} \nabla^2 \phi \quad (2.124)$$

But this is the same equation obeyed by the original field (2.105)! Except this time, it really is the Schrödinger equation, complete with the usual probabilistic interpretation for the wavefunction ϕ . Note in particular that the conserved charge arising from the

Noether current (2.107) is $Q = \int d^3x |\phi(\rightarrow x)|^2$ which is the total probability.

Historically, the fact that the equation for the classical field (2.105) and the one-particle wavefunction (2.124) coincide caused some confusion. It was thought that perhaps we are quantizing the wavefunction itself and the resulting name “second quantization” is still sometimes used today to mean quantum field theory. It’s important to stress that, despite the name, we’re not quantizing anything twice! We simply quantize a classical field once. Nonetheless, in practice it’s useful to know that if we treat the one-particle Schrödinger equation as the equation for a quantum field then it will give the correct generalization to multi-particle states.

Interactions

Often in quantum mechanics, we’re interested in particles moving in some fixed background potential $V(\rightarrow x)$. This can be easily incorporated into field theory by working with a Lagrangian with explicit $\rightarrow x$ dependence,

$$L = i\psi^\dagger \dot{\psi} - \frac{1}{2m} \nabla \psi^\dagger \nabla \psi - V(\rightarrow x) \psi^\dagger \psi \quad (2.125)$$

Note that this Lagrangian doesn’t respect translational symmetry and we won’t have the associated energy-momentum tensor. While such Lagrangians are useful in condensed matter physics, we rarely (or never) come across them in high-energy physics, where all equations obey translational (and Lorentz) invariance.

One can also consider interactions *between* particles. Obviously these are only important for n particle states with $n \geq 2$. We therefore expect them to arise from additions to the Lagrangian of the form

$$\Delta L = \psi^\wedge(\rightarrow x) \psi^\wedge(\rightarrow x) \psi(\rightarrow x) \psi(\rightarrow x) \quad (2.126)$$

which, in the quantum theory, is an operator which destroys two particles before creating two new ones. Such terms in the Lagrangian will indeed lead to inter-particle forces, both in the non-relativistic and relativistic setting. In the next section we explore these types of interaction in detail for relativistic theories.

3. Interacting Fields

The free field theories that we've discussed so far are very special: we can determine their spectrum, but nothing interesting then happens. They have particle excitations, but these particles don't interact with each other.

Here we'll start to examine more complicated theories that include interaction terms. These will take the form of higher order terms in the Lagrangian. We'll start by asking what kind of *small* perturbations we can add to the theory. For example, consider the Lagrangian for a real scalar field,

$$L = \frac{1}{2} \partial_\mu \varphi \partial^\mu \varphi - \frac{1}{2} m^2 \varphi^2 - \sum_{n \geq 3} \frac{\lambda_n}{n!} \varphi^n \quad (3.1)$$

The coefficients λ_n are called *coupling constants*. What restrictions do we have on λ_n to ensure that the additional terms are small perturbations? You might think that we need simply make " $\lambda_n = 1$ ". But this isn't quite right. To see why this is the case, let's do some dimensional analysis. Firstly, note that the action has dimensions of angular momentum or, equivalently, the same dimensions as k . Since we've set $k = 1$, using the convention described in the introduction, we have $[S] = 0$. With $S = \int d^4x L$, and $[d^4x] = 4$, the Lagrangian density must therefore have

$$[L] = 4 \quad (3.2)$$

What does this mean for the Lagrangian (3.1)? Since $[\partial_\mu] = 1$, we can read off the mass dimensions of all the factors to find,

$$[\varphi] = 1 , \quad [m] = 1 , \quad [\lambda_n] = 4 - n \quad (3.3)$$

So now we see why we can't simply say we need $\lambda_n = 1$, because this statement only makes sense for dimensionless quantities. The various terms, parameterized by λ_n , fall into three different categories

- $[\lambda_3] = 1$: For this term, the dimensionless parameter is λ_3/E , where E has dimensions of mass. Typically in quantum field theory, E is the energy scale of the process of interest. This means that $\lambda_3 \varphi^3/3!$ is a small perturbation at high energies $E \gg \lambda_3$, but a large perturbation at low energies $E \ll \lambda_3$. Terms that we add to the Lagrangian with this behavior are called *relevant* because they're most relevant at low energies (which, after all, is where most of the physics we see lies). In a relativistic theory, $E > m$, so we can always make this perturbation small by taking $\lambda_3 \ll m$.

- $[\lambda_4] = 0$: this term is small if $\lambda_4 \ll 1$. Such perturbations are called *marginal*.
- $[\lambda_n] < 0$ for $n \geq 5$: The dimensionless parameter is $(\lambda_n E)^{\frac{4}{n-4}}$, which is small at low-energies and large at high energies. Such perturbations are called *irrelevant*.

As you'll see later, it is typically impossible to avoid high energy processes in quantum field theory. (We've already seen a glimpse of this in computing the vacuum energy). This means that we might expect problems with irrelevant operators. Indeed, these lead to "non-renormalizable" field theories in which one cannot make sense of the infinities at arbitrarily high energies. This doesn't necessarily mean that the theory is useless; just that it is incomplete at some energy scale.

Let me note however that the naive assignment of relevant, marginal and irrelevant is not always fixed in stone: quantum corrections can sometimes change the character of an operator.

An Important Aside: Why QFT is Simple

Typically in a quantum field theory, only the relevant and marginal couplings are important. This is basically because, as we've seen above, the irrelevant couplings become small at low-energies. This is a huge help: of the infinite number of interaction terms that we could write down, only a handful are actually needed (just two in the case of the real scalar field described above).

Let's look at this a little more. Suppose that we some day discover the true superduper "theory of everything unimportant" that describes the world at very high energy scales, say the GUT scale, or the Planck scale. Whatever this scale is, let's call it Λ . It is an energy scale, so $[\Lambda] = 1$. Now we want to understand the laws of physics down at our puny energy scale $E \ll \Lambda$. Let's further suppose that down at the energy scale E , the laws of physics are described by a real scalar field. (They're not of course: they're described by non-Abelian gauge fields and fermions, but the same argument applies in that case so bear with me). This scalar field will have some complicated interaction terms (3.1), where the precise form is dictated by all the stuff that's going on in the high energy superduper theory. What are these interactions? Well, we could write our dimensionful coupling constants λ_n in terms of dimensionless couplings g_n , multiplied by a suitable power of the relevant scale Λ ,

$$\underline{g_n}$$

$$\lambda_n = \underline{g_n}^{\frac{4}{n-4}} \quad (3.4)$$

The exact values of dimensionless couplings $\underline{g_n}$ depend on the details of the high-energy superduper theory, but typically one expects them to be of order 1: $g_n \sim O(1)$. This

means that for experiments at small energies $E \ll \Lambda$, the interaction terms of the form φ^n with $n > 4$ will be suppressed by powers of $(E/\Lambda)^{n-4}$. This is usually a suppression by many orders of magnitude. (e.g. for the energies E explored at the LHC, $E/M_{\text{Pl}} \sim 10^{-16}$). It is this simple argument, based on dimensional analysis, that ensures that we need only focus on the first few terms in the interaction: those which are relevant and marginal. It also means that if we only have access to low-energy experiments (which we do!), it's going to be very difficult to figure out the high energy theory (which it is!), because its effects are highly diluted except for the relevant and marginal interactions. The discussion given above is a poor man's version of the ideas of *effective field theory* and *Wilson's renormalization group*, about which you can learn more in the "Statistical Field Theory" course.

Examples of Weakly Coupled Theories

In this course we'll study only weakly coupled field theories i.e. ones that can truly be considered as small perturbations of the free field theory at all energies. In this section, we'll look at two types of interactions

1) φ^4 theory:

$$L = \frac{1}{2} \partial_\mu \varphi \partial^\mu \varphi - \frac{1}{2m} \varphi^2 - \frac{\lambda}{4!} \varphi^4 \quad (3.5)$$

with $\lambda \ll 1$. We can get a hint for what the effects of this extra term will be. Expanding out φ^4 in terms of $a_{p \rightarrow}^\dagger$ and $a_{p \rightarrow}$, we see a sum of interactions that look like

$$a_{p \rightarrow}^\dagger a_{p \rightarrow}^\dagger a_{p \rightarrow}^\dagger a_{p \rightarrow} + \text{etc.} \quad (3.6)$$

These will create and destroy particles. This suggests that the φ^4 Lagrangian describes a theory in which particle number is not conserved. Indeed, we could check that the number operator N now satisfies $[H, N] \neq 0$.

2) Scalar Yukawa Theory

$$L = \partial_\mu \psi^\dagger \partial^\mu \psi + \frac{1}{2} \partial_\mu \varphi \partial^\mu \varphi - M \psi^\dagger \psi - \frac{1}{2m} \varphi^2 - g \psi^\dagger \psi \varphi \quad (3.7)$$

with $g \ll M, m$. This theory couples a complex scalar ψ to a real scalar φ . While the individual particle numbers of ψ and φ are no longer conserved, we do still have a symmetry rotating the phase of ψ , ensuring the existence of the charge Q defined in (2.75) such that $[Q, H] = 0$. This means that the number of ψ particles minus the number of ψ^\dagger anti-particles is conserved. It is common practice to denote the anti-particle as ψ^\dagger .

The scalar Yukawa theory has a slightly worrying aspect: the potential has a stable local minimum at $\varphi = \psi = 0$, but is unbounded below for large enough $-g\varphi$. This means we shouldn't try to push this theory too far.

A Comment on Strongly Coupled Field Theories

In this course we restrict attention to weakly coupled field theories where we can use perturbative techniques. The study of strongly coupled field theories is much more difficult, and one of the major research areas in theoretical physics. For example, some of the amazing things that can happen include

- **Charge Fractionalization:** Although electrons have electric charge 1, under the right conditions the elementary excitations in a solid have fractional charge $1/N$ (where $N \in 2\mathbb{Z} + 1$). For example, this occurs in the fractional quantum Hall effect.
- **Confinement:** The elementary excitations of quantum chromodynamics (QCD) are quarks. But they *never* appear on their own, only in groups of three (in a baryon) or with an anti-quark (in a meson). They are confined.
- **Emergent Space:** There are field theories in four dimensions which at strong coupling become quantum gravity theories in ten dimensions! The strong coupling effects cause the excitations to act as if they're gravitons moving in higher dimensions. This is quite extraordinary and still poorly understood. It's called the AdS/CFT correspondence.

3.1 The Interaction Picture

There's a useful viewpoint in quantum mechanics to describe situations where we have small perturbations to a well-understood Hamiltonian. Let's return to the familiar ground of quantum mechanics with a finite number of degrees of freedom for a moment. In the Schrödinger picture, the states evolve as

$$i \frac{d|\psi\rangle_s}{dt} = H |\psi\rangle_s \quad (3.8)$$

while the operators O_s are independent of time.

In contrast, in the Heisenberg picture the states are fixed and the operators change in time

$$\begin{aligned} O_H(t) &= e^{iHt} O_s e^{-iHt} \\ |\psi\rangle_H &= e^{iHt} |\psi\rangle_s \end{aligned} \quad (3.9)$$

The *interaction picture* is a hybrid of the two. We split the Hamiltonian up as

$$H = H_0 + H_{\text{int}} \quad (3.10)$$

The time dependence of operators is governed by H_0 , while the time dependence of states is governed by H_{int} . Although the split into H_0 and H_{int} is arbitrary, it's useful when H_0 is soluble (for example, when H_0 is the Hamiltonian for a free field theory). The states and operators in the interaction picture will be denoted by a subscript I and are given by,

$$\begin{aligned} |\psi(t)\rangle_I &= e^{iH_0 t} |\psi(t)\rangle_S \\ O_I(t) &= e^{iH_0 t} O_S e^{-iH_0 t} \end{aligned} \quad (3.11)$$

This last equation also applies to H_{int} , which is time dependent. The interaction Hamiltonian in the interaction picture is,

$$H_I \equiv (H_{\text{int}})_I = e^{iH_0 t} (H_{\text{int}})_S e^{-iH_0 t} \quad (3.12)$$

The Schrödinger equation for states in the interaction picture can be derived starting from the Schrödinger picture

$$\begin{aligned} i \frac{d|\psi\rangle_S}{dt} &= H_S |\psi\rangle \quad \Rightarrow \quad i \frac{d}{dt} e^{\frac{-iH_0 t}{\hbar}} |\psi\rangle_I = (H_0 + H_{\text{int}})_S e^{\frac{-iH_0 t}{\hbar}} |\psi\rangle_I \\ &\Rightarrow i \frac{d}{dt} \frac{e^{\frac{-iH_0 t}{\hbar}} |\psi\rangle_I}{e^{\frac{-iH_0 t}{\hbar}}} = (H_{\text{int}})_S |\psi\rangle_I \end{aligned} \quad (3.13)$$

So we learn that

$$i \frac{d|\psi\rangle}{dt} = H_I(t) |\psi\rangle_I \quad (3.14)$$

3.1.1 Dyson's Formula

"Well, Birmingham has much the best theoretical physicist to work with, Peierls; Bristol has much the best experimental physicist, Powell; Cambridge has some excellent architecture. You can make your choice."

Oppenheimer's advice to Dyson on which university position to accept.

We want to solve (3.14). Let's write the solution as

$$|\psi(t)\rangle_I = U(t, t_0) |\psi(t_0)\rangle_I \quad (3.15)$$

where $U(t, t_0)$ is a unitary time evolution operator such that $U(t_1, t_2)U(t_2, t_3) = U(t_1, t_3)$ and $U(t, t) = 1$. Then the interaction picture Schrödinger equation (3.14) requires that

$$i \frac{dU}{dt} = H_l(t) U \quad (3.16)$$

If H_l were a function, then we could simply solve this by

$$U(t, t_0) = \exp \left(-i \int_{t_0}^t H_l(t') dt' \right) \quad (3.17)$$

But there's a problem. Our Hamiltonian H_l is an operator, and we have ordering issues. Let's see why this causes trouble. The exponential of an operator is defined in terms of the expansion,

$$\exp \left(-i \int_{t_0}^t H_l(t') dt' \right) = 1 - i \int_{t_0}^t H_l(t') dt' + \frac{(-i)^2}{2} \int_{t_0}^t \int_{t'}^t H_l(t') H_l(t'') dt' dt'' + \dots \quad (3.18)$$

But when we try to differentiate this with respect to t , we find that the quadratic term gives us

$$-\frac{1}{2} \int_{t_0}^t \frac{d}{dt'} H_l(t') = H_l(t) - \frac{1}{2} \int_{t_0}^t \frac{d}{dt'} H_l(t') \quad (3.19)$$

Now the second term here looks good, since it will give part of the $H_l(t)U$ that we need on the right-hand side of (3.16). But the first term is no good since the $H_l(t)$ sits on the wrong side of the integral term, and we can't commute it through because $[H_l(t'), H_l(t)] \neq 0$ when $t' \neq t$. So what's the way around this?

Claim: The solution to (3.16) is given by *Dyson's Formula*. (Essentially first figured out by Dirac, although the compact notation is due to Dyson).

$$U(t, t_0) = T \exp \left(-i \int_{t_0}^t H_l(t') dt' \right) \quad (3.20)$$

where T stands for *time ordering* where operators evaluated at later times are placed to the left

$$T(O_1(t_1) O_2(t_2)) = \begin{cases} O_1(t_1) O_2(t_2) & t_1 > t_2 \\ O_2(t_2) O_1(t_1) & t_2 > t_1 \end{cases} \quad (3.21)$$

Expanding out the expression (3.20), we now have

$$\begin{aligned} U(t, t_0) &= 1 - i \int_{t_0}^t dt' H_l(t') + \frac{(-i)^2}{2} \int_{t_0}^t \int_{t'}^t dt' dt'' H_l(t') H_l(t'') \\ &\quad + \frac{(-i)^3}{3!} \int_{t_0}^t \int_{t'}^t \int_{t''}^t dt' dt'' dt''' H_l(t') H_l(t'') H_l(t''') \dots + \dots \end{aligned}$$

Actually these last two terms double up since

$$\begin{aligned}
 & \int_{t_0}^t \int_{t'} dt^{rr} H_l(t^{rr}) H_l(t^r) = \int_{t_0}^t \int_{t'}^t dt^r H_l(t^{rr}) H_l(t^r) \\
 & \quad = \int_{t_0}^t dt^{rr} H_l(t^r) H_l(t^{rr})
 \end{aligned} \tag{3.22}$$

where the range of integration in the first expression is over $t^{rr} \geq t^r$, while in the second expression it is $t^r \leq t^{rr}$ which is, of course, the same thing. The final expression is the same as the second expression by a simple relabelling. This means that we can write

$$U(t, t_0) = 1 - i \int_{t_0}^t dt^r H_l(t^r) + (-i)^2 \int_{t_0}^t \int_{t'}^t dt^{rr} H_l(t^r) H_l(t^{rr}) + \dots \tag{3.23}$$

Proof: The proof of Dyson's formula is simpler than explaining what all the notation means! Firstly observe that under the T sign, all operators commute (since their order is already fixed by the T sign). Thus

$$\begin{aligned}
 i \frac{\partial}{\partial t} T \exp \left[-i \int_{t_0}^t dt^r H_l(t^r) \right] &= T \exp \left[-i \int_{t_0}^t dt^r H_l(t^r) \right] \\
 &= H_l(t) T \exp \left[-i \int_{t_0}^t dt^r H_l(t^r) \right]
 \end{aligned} \tag{3.24}$$

since t , being the upper limit of the integral, is the latest time so $H_l(t)$ can be pulled out to the left.

Before moving on, I should confess that Dyson's formula is rather formal. It is typically very hard to compute time ordered exponentials in practice. The power of the formula comes from the expansion which is valid when H_l is small and is very easily computed.

3.2 A First Look at Scattering

Let us now apply the interaction picture to field theory, starting with the interaction Hamiltonian for our scalar Yukawa theory,

$$\int d^3x \psi^\dagger \psi \varphi \tag{3.25}$$

Unlike the free theories discussed in Section 2, this interaction doesn't conserve particle number, allowing particles of one type to morph into others. To see why this is, we use

the interaction picture and follow the evolution of the state: $|\psi(t)\rangle = U(t, t_0) |\psi(t_0)\rangle$, where $U(t, t_0)$ is given by Dyson's formula (3.20) which is an expansion in powers of H_{int} . But H_{int} contains creation and annihilation operators for each type of particle. In particular,

- $\varphi \sim a + a^\dagger$: This operator can create or destroy φ particles. Let's call them *mesons*.
- $\psi \sim b + c^\dagger$: This operator can destroy ψ particles through b , and create anti-particles through c^\dagger . Let's call these particles *nucleons*. Of course, in reality nucleons are spin 1/2 particles, and don't arise from the quantization of a scalar field. But we'll treat our scalar Yukawa theory as a toy model for nucleons interacting with mesons.
- $\psi^\dagger \sim b^\dagger + c$: This operator can create nucleons through b^\dagger , and destroy anti-nucleons through c .

Importantly, $Q = N_c - N_b$ remains conserved in the presence of H_{int} . At first order in perturbation theory, we find terms in H_{int} like $c^\dagger b^\dagger a$. This kills a meson, producing a nucleon-anti-nucleon pair. It will contribute to meson decay $\varphi \rightarrow \psi \bar{\psi}$.

At second order in perturbation theory, we'll have more complicated terms in $(H_{\text{int}})^2$, for example $(c^\dagger b^\dagger a)(cba^\dagger)$. This term will give contributions to scattering processes $\psi \bar{\psi} \rightarrow \varphi \rightarrow \psi \bar{\psi}$. The rest of this section is devoted to computing the quantum amplitudes for these processes to occur.

To calculate amplitudes we make an important, and slightly dodgy, assumption:

Initial and final states are eigenstates of the free theory

This means that we take the initial state $|i\rangle$ at $t \rightarrow -\infty$, and the final state $|f\rangle$ at $t \rightarrow +\infty$, to be eigenstates of the free Hamiltonian H_0 . At some level, this sounds plausible: at $t \rightarrow -\infty$, the particles in a scattering process are far separated and don't feel the effects of each other. Furthermore, we intuitively expect these states to be eigenstates of the individual number operators N , which commute with H_0 , but not H_{int} . As the particles approach each other, they interact briefly, before departing again, each going on its own merry way. The amplitude to go from $|i\rangle$ to $|f\rangle$ is

$$\lim_{t_\pm \rightarrow \pm\infty} \langle f | U(t_+, t_-) | i \rangle \equiv \langle f | S | i \rangle \quad (3.26)$$

where the unitary operator S is known as the S-matrix. (S is for scattering). There are a number of reasons why the assumption of non-interacting initial and final states is shaky:

- Obviously we can't cope with bound states. For example, this formalism can't describe the scattering of an electron and proton which collide, bind, and leave as a Hydrogen atom. It's possible to circumvent this objection since it turns out that bound states show up as poles in the S-matrix.
- More importantly, a single particle, a long way from its neighbors, is never alone in field theory. This is true even in classical electrodynamics, where the electron sources the electromagnetic field from which it can never escape. In quantum electrodynamics (QED), a related fact is that there is a cloud of *virtual* photons surrounding the electron. This line of thought gets us into the issues of renormalization — more on this next term in the “AQFT” course. Nevertheless, motivated by this problem, after developing scattering theory using the assumption of non-interacting asymptotic states, we'll mention a better way.

3.2.1 An Example: Meson Decay

Consider the relativistically normalized initial and final states,

$$\begin{aligned} |i\rangle &= \sqrt{2E_{\vec{p}} \rightarrow} a_{\vec{p}}^\dagger |0\rangle \\ |f\rangle &= \sqrt{4E_{\vec{q}} \rightarrow E_{\vec{q}_1 \rightarrow 2}} b_{\vec{q}_1}^\dagger c_{\vec{q}_2}^\dagger |0\rangle \end{aligned} \quad (3.27)$$

The initial state contains a single meson of momentum p ; the final state contains a nucleon-anti-nucleon pair of momentum q_1 and q_2 . We may compute the amplitude for the decay of a meson to a nucleon-anti-nucleon pair. To leading order in g , it is

$$\langle f | S | i \rangle = -ig \langle f | \int d^4x \psi^\dagger(x) \psi(x) \varphi(x) | i \rangle \quad (3.28)$$

Let's go slowly. We first expand out $\varphi \sim a + a^\dagger$ using (2.84). (Remember that the φ in this formula is in the interaction picture, which is the same as the Heisenberg picture of the free theory). The a piece will turn $|i\rangle$ into something proportional to $|0\rangle$, while the a^\dagger piece will turn $|i\rangle$ into a two meson state. But the two meson state will have zero overlap with $|f\rangle$, and there's nothing in the ψ and ψ^\dagger operators that lie between them to change this fact. So we have

$$\begin{aligned} \langle f | S | i \rangle &= -ig \langle f | \int d^4x \psi^\dagger(x) \psi(x) \int \frac{d^3k}{(2\pi)^3} \frac{\sqrt{2E_{\vec{k}}}}{2E_{\vec{k}}} a_{\vec{k}}^\dagger a_{\vec{k}} e^{-ik \cdot x} | 0 \rangle \\ &= -ig \langle f | \int d^4x \psi^\dagger(x) \psi(x) e^{-ip \cdot x} | 0 \rangle \end{aligned} \quad (3.29)$$

where, in the second line, we've commuted $a_{\vec{k}}$ past $a_{\vec{p}}$, picking up a $\delta^{(3)}(\vec{p} \rightarrow \vec{k})$ delta-function which kills the d^3k integral. We now similarly expand out $\psi \sim b + b^\dagger$ and

$\psi^\dagger \sim b^\dagger + c$. To get non-zero overlap with $\langle f |$, only the b^\dagger and c^\dagger contribute, for they create the nucleon and anti-nucleon from $|0\rangle$. We then have

$$\begin{aligned} \langle f | S | i \rangle &= -ig \langle 0 | \frac{d^4x d^3k_1 d^3k_2}{(2\pi)^6} \cancel{E} \cancel{E} c_{q \rightarrow \frac{1}{2} \vec{k}_1}^\dagger b_{q \rightarrow \frac{1}{2} \vec{k}_2}^\dagger c_{q \rightarrow \frac{1}{2} \vec{k}_1}^\dagger b_{q \rightarrow \frac{1}{2} \vec{k}_2}^\dagger |0\rangle e^{i(\vec{k}_1 + \vec{k}_2 - \vec{p}) \cdot \vec{x}} \\ &= -ig (2\pi)^4 \delta^{(4)}(q_1 + q_2 - p) \end{aligned} \quad (3.30)$$

and so we get our first quantum field theory amplitude.

Notice that the δ -function puts constraints on the possible decays. In particular, the decay only happens at all if $m \geq 2M$. To see this, we may always boost ourselves to a reference frame where the meson is stationary, so $p = (m, 0, 0, 0)$. Then the delta function imposes momentum conservation, telling us that $\vec{q}_1 = -\vec{q}_2$ and $m = \sqrt{2M^2 + |\vec{q}|^2}$.

Later you will learn how to turn this quantum amplitude into something more physical, namely the lifetime of the meson. The reason this is a little tricky is that we must square the amplitude to get the probability for decay, which means we get the square of a δ -function. We'll explain how to deal with this in Section 3.6 below, and again in next term's "Standard Model" course.

3.3 Wick's Theorem

From Dyson's formula, we want to compute quantities like $\langle f | T \{ H_1(x_1) \dots H_l(x_n) \} | i \rangle$, where $|i\rangle$ and $|f\rangle$ are eigenstates of the free theory. The ordering of the operators is fixed by T , time ordering. However, since the H 's contain certain creation and annihilation operators, our life will be much simpler if we can start to move all annihilation operators to the right where they can start killing things in $|i\rangle$. Recall that this is the definition of normal ordering. Wick's theorem tells us how to go from time ordered products to normal ordered products.

3.3.1 An Example: Recovering the Propagator

Let's start simple. Consider a real scalar field which we decompose in the Heisenberg picture as

$$\text{where } \varphi(x) = \varphi^+(x) + \varphi^-(x) \quad (3.31)$$

$$\begin{aligned} \varphi^+(x) &= \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{\frac{2E}{\cancel{p}}}} e^{-ip \cdot x} \\ &\quad - \int \frac{(2\pi)^3}{d^3p} \frac{a}{\sqrt{\frac{2E}{\cancel{p}}}} e^{+ip \cdot x} \\ \varphi^-(x) &= \frac{e}{(2\pi)^3} \sqrt{2E} a_p \end{aligned} \quad (3.32)$$

where the \pm signs on φ^\pm make little sense, but apparently you have Pauli and Heisenberg to blame. (They come about because $\varphi^+ \sim e^{-iEt}$, which is sometimes called the positive frequency piece, while $\varphi^- \sim e^{+iEt}$ is the negative frequency piece). Then choosing $x^0 > y^0$, we have

$$\begin{aligned} T\varphi(x)\varphi(y) &= \varphi(x)\varphi(y) \\ &= (\varphi^+(x) + \varphi^-(x))(\varphi^+(y) + \varphi^-(y)) \\ &= \varphi^+(x)\varphi^+(y) + \varphi^-(x)\varphi^+(y) + \varphi^-(y)\varphi^+(x) + [\varphi^+(x), \varphi^-(y)] + \varphi^-(x)\varphi^-(y) \end{aligned} \quad (3.33)$$

where the last line is normal ordered, and for our troubles we have picked up the extra term $D(x-y) = [\varphi^+(x), \varphi^-(y)]$ which is the propagator we met in (2.90). So for $x^0 > y^0$ we have

$$T\varphi(x)\varphi(y) =: \varphi(x)\varphi(y) : + D(x-y) \quad (3.34)$$

Meanwhile, for $y^0 > x^0$, we may repeat the calculation to find

$$T\varphi(x)\varphi(y) =: \varphi(x)\varphi(y) : + D(y-x) \quad (3.35)$$

So putting this together, we have the final expression

$$T\varphi(x)\varphi(y) =: \varphi(x)\varphi(y) : + \Delta_F(x-y) \quad (3.36)$$

where $\Delta_F(x-y)$ is the Feynman propagator defined in (2.93), for which we have the integral representation

$$\Delta_F(x-y) = \int \frac{d^4 k}{(2\pi)^4} \frac{i e^{ik \cdot (x-y)}}{k^2 - m^2 + i\epsilon} \quad (3.37)$$

Let me reiterate a comment from Section 2: although $T\varphi(x)\varphi(y)$ and $: \varphi(x)\varphi(y) :$ are both operators, the difference between them is a c-number function, $\Delta_F(x-y)$.

Definition: We define the *contraction* of a pair of fields in a string of operators $\dots \varphi(x_1) \dots \varphi(x_2) \dots$ to mean replacing those operators with the Feynman propagator, leaving all other operators untouched. We use the notation,

$$\dots \overbrace{\varphi(x_1)}^x \dots \overbrace{\varphi(x_2)}^y \dots \quad (3.38)$$

to denote contraction. So, for example,

$$\overbrace{\varphi(x)\varphi(y)}^x = \Delta_F(x-y) \quad (3.39)$$

A similar discussion holds for complex scalar fields. We have

$$T\psi(x)\psi^\dagger(y) =: \psi(x)\psi^\dagger(y) : + \Delta_F(x - y) \quad (3.40)$$

prompting us to define the contraction

$$\overline{\psi(x)\psi^\dagger(y)} = \Delta_F(x - y) \quad \text{and} \quad \overline{\psi(x)\psi(y)} = \overline{\psi^\dagger(x)\psi^\dagger(y)} = 0 \quad (3.41)$$

3.3.2 Wick's Theorem

For any collection of fields $\varphi_1 = \varphi(x_1)$, $\varphi_2 = \varphi(x_2)$, etc, we have

$$T(\varphi_1 \dots \varphi_n) =: \varphi_1 \dots \varphi_n : + : \text{all possible contractions} : \quad (3.42)$$

To see what the last part of this equation means, let's look at an example. For $n = 4$, the equation reads

$$\begin{aligned} T(\varphi_1\varphi_2\varphi_3\varphi_4) &= : \varphi_1\varphi_2\varphi_3\varphi_4 : + \overline{\varphi_1}\varphi_2 : \varphi_3\varphi_4 : + \varphi_1\overline{\varphi_3} : \varphi_2\varphi_4 : + \text{four similar terms} \\ &\quad + \overline{\varphi_1}\overline{\varphi_2} \varphi_3\varphi_4 + \varphi_1\overline{\varphi_3} \varphi_2\varphi_4 + \varphi_1\overline{\varphi_4} \varphi_2\varphi_3 \end{aligned} \quad (3.43)$$

Proof: The proof of Wick's theorem proceeds by induction and a little thought. It's true for $n = 2$. Suppose it's true for $\varphi_2 \dots \varphi_n$ and now add φ_1 . We'll take $x_1^0 > x_k^0$ for all $k = 2, \dots, n$. Then we can pull φ_1 out to the left of the time ordered product, writing

$$T(\varphi_1\varphi_2 \dots \varphi_n) = (\varphi_1^+ + \varphi_1^-) \left(: \varphi_2 \dots \varphi_n : + : \text{contractions} : \right) \quad (3.44)$$

The φ_1^- term stays where it is since it is already normal ordered. But in order to write the right-hand side as a normal ordered product, the φ_1^+ term has to make its way past the crowd of φ_k^- operators. Each time it moves past φ_k^- , we pick up a factor of \overline{x}_k . $\varphi_1\varphi_k = \Delta_F(x_1 - x_k)$ from the commutator. (Try it!)

3.3.3 An Example: Nucleon Scattering

Let's look at $\psi\psi \rightarrow \psi\psi$ scattering. We have the initial and final states

$$\begin{aligned} |i\rangle &= \sqrt{2E_{p_1}} \sqrt{2E_{p_2}} b_1^\dagger b_2^\dagger |0\rangle \equiv |p_1, p_2\rangle \\ |f\rangle &= q_1 \overline{2E_{p_1}} q_2 \overline{2E_{p_2}} b_1^\dagger b_2^\dagger |0\rangle \equiv |p_1, p_2\rangle \end{aligned} \quad (3.45)$$

We can then look at the expansion of $\langle f | S | i \rangle$. In fact, we really want to calculate $\langle f | S - 1 | i \rangle$ since we're not interested in situations where no scattering occurs. At order g^2 we have the term

$$\frac{(-ig)^2}{2} \frac{d^4x_1 d^4x_2}{\psi^\dagger(x_1)\psi(x_1)\psi^\dagger(x_2)\psi(x_2)} T \psi^\dagger(x_1)\psi(x_1)\varphi(x_1)\psi^\dagger(x_2)\psi(x_2) \quad (3.46)$$

Now, using Wick's theorem we see there is a piece in the string of operators which looks like

$$:\psi^\dagger(x_1)\psi(x_1)\psi^\dagger(x_2)\psi(x_2): \quad \overset{x}{\overbrace{\varphi(x_1)}} \quad \overset{x}{\overbrace{\varphi(x_2)}} \quad (3.47)$$

which will contribute to the scattering because the two ψ fields annihilate the ψ particles, while the two ψ^\dagger fields create ψ particles. Any other way of ordering the ψ and ψ^\dagger fields will give zero contribution. This means that we have

$$\begin{aligned} & \langle p_1^r, p_2^r | : \overset{r}{\psi}(x_1) \overset{r}{\psi}(x_1) \overset{\dagger}{\psi}(x_2) \overset{\dagger}{\psi}(x_2) : | p_1, p_2 \rangle \\ &= \langle p_1^r, p_2^r | \psi^\dagger(x_1) \psi^\dagger(x_2) | 0 \rangle \langle 0 | \psi(x_1) \psi(x_2) | p_1, p_2 \rangle \\ &= e^{ip'_1 \cdot x_1 + ip'_2 \cdot x_2} + e^{ip'_1 \cdot x_2 + ip'_2 \cdot x_1} - e^{-ip_1 \cdot x_1 - ip_2 \cdot x_2} + e^{-ip_1 \cdot x_2 - ip_2 \cdot x_1} \\ &= e^{ix_1 \cdot (p'_1 - p_1) + ix_2 \cdot (p'_2 - p_2)} + e^{ix_1 \cdot (p'_2 - p_1) + ix_2 \cdot (p'_1 - p_2)} + (x_1 \leftrightarrow x_2) \end{aligned} \quad (3.48)$$

where, in going to the third line, we've used the fact that for relativistically normalized states,

$$\langle 0 | \psi(x) | p \rangle = e^{-ip \cdot x} \quad (3.49)$$

Now let's insert this into (3.46), to get the expression for $\langle f | S | i \rangle$ at order g^2 ,

$$\frac{(-ig)^2}{2} \int d^4x_1 d^4x_2 \left[e^{i \dots} + e^{i \dots} + (x_1 \leftrightarrow x_2) \right] \int \frac{d^4k}{(2\pi)^4} \frac{i e^{ik \cdot (x_1 - x_2)}}{k^2 - m^2 + i\epsilon} \quad (3.50)$$

where the expression in square brackets is (3.48), while the final integral is the φ propagator which comes from the contraction in (3.47). Now the $(x_1 \leftrightarrow x_2)$ terms double up with the others to cancel the factor of $1/2$ out front. Meanwhile, the x_1 and x_2 integrals give delta-functions. We're left with the expression

$$\begin{aligned} & (-ig)^2 \frac{i(2\pi)^8}{(2\pi)^4 k^2 - m^2 + i\epsilon} \delta^{(4)}(p^r_1 - p^r_2 + k) \delta^{(4)}(p^r_2 - p^r_1 - k) \\ &+ \delta^{(4)}(p^r_2 - p_1 + k) \delta^{(4)}(p^r_1 - p_2 - k) \end{aligned} \quad (3.51)$$

Finally, we can trivially do the d^4k integral using the delta-functions to get

$$i(-ig)^2 \frac{1}{(p_1 - p_2 - m^2 + i\epsilon)^2} + \frac{1}{(p_1 - p_2 - m^2 + i\epsilon)^2} \frac{(2\pi)^4 \delta^{(4)}(p_1 + p_2 - p^r_1 - p^r_2)}{1^2 2^1 1^2}$$

In fact, for this process we may drop the $+i\epsilon$ terms since the denominator is never zero. To see this, we can go to the center of mass frame, where $p_1 = -p_2$ and, by

momentum conservation, $|p \rightarrow_1| = |p_1 \rightarrow^r|$. This ensures that the 4-momentum of the meson is $k = (0, p \rightarrow - p_1^r)$, so $k^2 < 0$. We therefore have the end result,

$$i(-ig)^2 \frac{1}{(p_1 - p_1^r)^2 - m^2} + \frac{1}{(p_1 - p_2^r)^2 - m^2} (2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_1^r - p_2^r) \quad (3.52)$$

We will see another, much simpler way to reproduce this result shortly using Feynman diagrams. This will also shed light on the physical interpretation.

This calculation is also relevant for other scattering processes, such as $\psi^- \psi^- \rightarrow \psi^- \psi^-$, $\psi^- \psi^- \rightarrow \psi^- \psi^-$. Each of these comes from the term (3.48) in Wick's theorem. However, we will never find a term that contributes to scattering $\psi \psi \rightarrow \psi^- \psi^-$, for this would violate the conservation of Q charge.

Another Example: Meson-Nucleon Scattering

If we want to compute $\psi \varphi \rightarrow \psi \varphi$ scattering at order g^2 , we would need to pick out the term

$$: \psi^\dagger(x_1)\varphi(x_1)\psi(x_2)\varphi(x_2) : \overset{x}{\overbrace{\psi(x_1)\psi^\dagger(x_2)}} \quad (3.53)$$

and a similar term with ψ and ψ^\dagger exchanged. Once more, this term also contributes to similar scattering processes, including $\psi^- \varphi \rightarrow \psi^- \varphi$ and $\varphi \varphi \rightarrow \psi^- \psi^-$.

3.4 Feynman Diagrams

“Like the silicon chips of more recent years, the Feynman diagram was bringing computation to the masses.”

Julian Schwinger

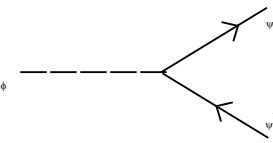
As the above example demonstrates, to actually compute scattering amplitudes using Wick's theorem is rather tedious. There's a much better way. It requires drawing pretty pictures. These pictures represent the expansion of $\langle f | S | i \rangle$ and we will learn how to associate numbers (or at least integrals) to them. These pictures are called *Feynman diagrams*.

The object that we really want to compute is $\langle f | S - 1 | i \rangle$, since we're not interested in processes where no scattering occurs. The various terms in the perturbative expansion can be represented pictorially as follows

- Draw an external line for each particle in the initial state $|i\rangle$ and each particle in the final state $|f\rangle$. We'll choose dotted lines for mesons, and solid lines for nucleons. Assign a directed momentum p to each line. Further, add an arrow to

solid lines to denote its charge; we'll choose an incoming (outgoing) arrow in the initial state for ψ ($\bar{\psi}$). We choose the reverse convention for the final state, where an outgoing arrow denotes ψ .

- Join the external lines together with trivalent vertices



Each such diagram you can draw is in 1-1 correspondence with the terms in the expansion of $\langle f | S - 1 | i \rangle$.

3.4.1 Feynman Rules

To each diagram we associate a number, using the *Feynman rules*

- Add a momentum k to each internal line
- To each vertex, write down a factor of

$$(-ig) (2\pi)^4 \delta^{(4)}(\sum_i k_i) \quad (3.54)$$

where $\sum_i k_i$ is the sum of all momenta flowing *into* the vertex.

- For each internal dotted line, corresponding to a φ particle with momentum k , we write down a factor of

$$\int \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2 + i\epsilon} \quad (3.55)$$

We include the same factor for solid internal ψ lines, with m replaced by the nucleon mass M .

3.5 Examples of Scattering Amplitudes

Let's apply the Feynman rules to compute the amplitudes for various processes. We start with something familiar:

Nucleon Scattering Revisited

Let's look at how this works for the $\psi\psi \rightarrow \psi\psi$ scattering at order g^2 . We can write down the two simplest diagrams contributing to this process. They are shown in Figure 9.

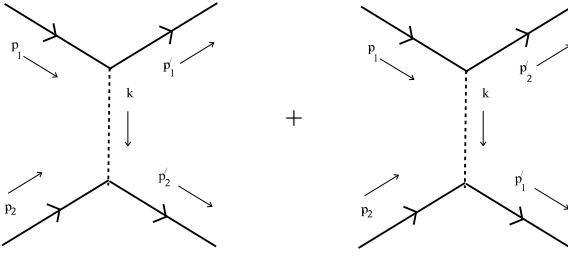


Figure 9: The two lowest order Feynman diagrams for nucleon scattering.

Applying the Feynman rules to these diagrams, we get

$$i(-ig)^2 \frac{1}{(p_1 - p_1^r)^2 - m^2} + \frac{1}{(p_1 - p_2^r)^2 - m^2} (2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_1^r - p_2^r) \quad (3.56)$$

which agrees with the calculation (3.51) that we performed earlier. There is a nice physical interpretation of these diagrams. We talk, rather loosely, of the nucleons exchanging a meson which, in the first diagram, has momentum $k = (p_1 - p_1^r) = (p_1^r - p_2)$.

$\overline{\text{---}}_2$

This meson doesn't satisfy the usual energy dispersion relation, because $k^2 \neq m^2$: the meson is called a *virtual particle* and is said to be *off-shell* (or, sometimes, off mass-shell). Heuristically, it can't live long enough for its energy to be measured to great accuracy. In contrast, the momentum on the external, nucleon legs all satisfy $p^2 = M^2$, the mass of the nucleon. They are *on-shell*. One final note: the addition of the two diagrams above ensures that the particles satisfy Bose statistics.

There are also more complicated diagrams which will contribute to the scattering process at higher orders. For example, we have the two diagrams shown in Figures 10 and 11, and similar diagrams with p_1^r and p_2^r exchanged. Using the Feynman rules, each of these diagrams translates into an integral that we will not attempt to calculate here. And so we go on, with increasingly complicated diagrams, all appearing at higher order in the coupling constant g .

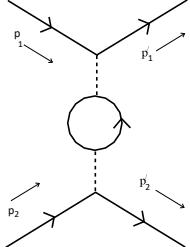


Figure 10: A contribution at $O(g^4)$.

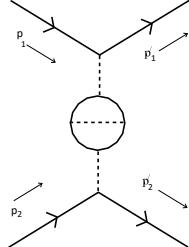


Figure 11: A contribution at $O(g^6)$

Amplitudes

Our final result for the nucleon scattering amplitude $\langle f | S - 1 | i \rangle$ at order g^2 was

$$i(-ig)^2 \frac{1}{(p_1 - p_1' - m^2)} + \frac{1}{(p_1 - p_2' - m^2)} \frac{(2\pi)^4 \delta^{(4)}(p_1 + p_2 - p_1' - p_2')}{1 \quad 2 \quad 1 \quad 2}$$

The δ -function follows from the conservation of 4-momentum which, in turn, follows from spacetime translational invariance. It is common to all S-matrix elements. We will define the amplitude A_{fi} by stripping off this momentum-conserving delta-function,

$$\langle f | S - 1 | i \rangle = i A_{fi} (2\pi)^4 \delta^{(4)}(p_F - p_i) \quad (3.57)$$

where p_i (p_F) is the sum of the initial (final) 4-momenta, and the factor of i out front is a convention which is there to match non-relativistic quantum mechanics. We can now refine our Feynman rules to compute the amplitude iA_{fi} itself:

- Draw all possible diagrams with appropriate external legs and impose 4-momentum conservation at each vertex.
- Write down a factor of $(-ig)$ at each vertex.
- For each internal line, write down the propagator
- Integrate over momentum k flowing through each loop $d^4k/(2\pi)^4$.

This last step deserves a short explanation. The diagrams we've computed so far have no loops. They are *tree level* diagrams. It's not hard to convince yourself that in tree diagrams, momentum conservation at each vertex is sufficient to determine the momentum flowing through each internal line. For diagrams with loops, such as those shown in Figures 10 and 11, this is no longer the case.

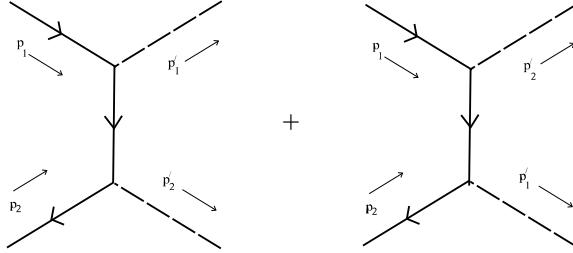


Figure 12: The two lowest order Feynman diagrams for nucleon to meson scattering.

Nucleon to Meson Scattering

Let's now look at the amplitude for a nucleon-anti-nucleon pair to annihilate into a pair of mesons: $\psi \bar{\psi} \rightarrow \varphi \varphi$. The simplest Feynman diagrams for this process are shown in Figure 12 where the virtual particle in these diagrams is now the nucleon ψ rather than the meson φ . This fact is reflected in the denominator of the amplitudes which are given by

$$iA = (-ig)^2 \frac{i}{(p_1 - p^r)^2 - M^2} + \frac{i}{(p_1 - p^r)^2 - M^2} \quad (3.58)$$

As in (3.52), we've dropped the $i\epsilon$ from the propagators as the denominator never vanishes.

Nucleon-Anti-Nucleon Scattering

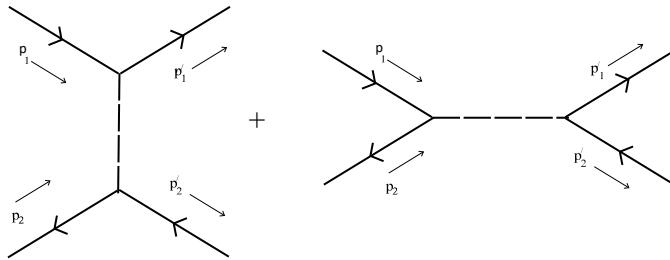


Figure 13: The two lowest order Feynman diagrams for nucleon-anti-nucleon scattering.

For the scattering of a nucleon and an anti-nucleon, $\psi \bar{\psi} \rightarrow \psi \bar{\psi}$, the Feynman diagrams are a little different. At lowest order, they are given by the diagrams of Figure 13. It is a simple matter to write down the amplitude using the Feynman rules,

$$iA = (-ig)^2 \frac{i}{(p_1 - p^r)^2 - m^2} + \frac{i}{(p_1 + p_2)^2 - m^2 + i\epsilon} \quad (3.59)$$

Notice that the momentum dependence in the second term is different from that of nucleon-nucleon scattering (3.56), reflecting the different Feynman diagram that contributes to the process. In the center of mass frame, $p \rightarrow_1 = -p \rightarrow_2$, the denominator of the second term is $4(M^2 + p \rightarrow_1^2) - m^2$. If $m < 2M$, then this

term never vanishes and we may drop the $i\epsilon$. In contrast, if $m > 2M$, then the amplitude corresponding to the second diagram diverges at some value of $p \rightarrow$. In this case it turns out that we may also neglect the $i\epsilon$ term, although for a different reason: the meson is unstable when $m > 2M$, a result we derived in (3.30). When correctly treated, this instability adds a finite imaginary piece to the denominator which

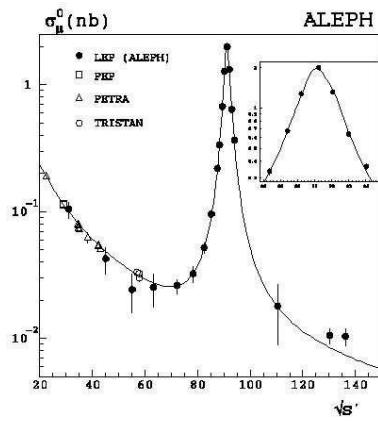
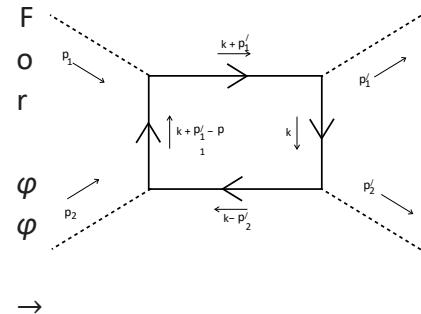


Figure 14:

overwhelms the $i\epsilon$. Nonetheless, the increase in the scattering amplitude which we see in the second diagram when $4(M^2 + p \rightarrow^2) = m^2$ is what allows us to discover new particles: they appear as a resonance in the cross section. For example, the Figure 14 shows the cross-section (roughly the amplitude squared) plotted vertically for $e^+e^- \rightarrow \mu^+\mu^-$ scattering from the ALEPH experiment in CERN. The horizontal axis shows the center of mass energy. The curve rises sharply around 91 GeV, the mass of the Z-boson.

Meson Scattering



→
 φ
 φ
,
t
h

the simplest diagram we can write down has a single loop, and momentum conservation at each vertex is no longer sufficient to determine every momentum passing through the diagram. We choose to assign the single undetermined momentum k to the right-hand propagator. All other momenta are then determined. The amplitude corresponding to the diagram shown in the figure is

Figure 15:

$$\begin{aligned}
 & \left(-\frac{\int}{ig)^4} \frac{1}{(k^2 - M^2 + i\epsilon)((k + p^r)^2 - M_1^2 + i\epsilon)} \right. \\
 & d^4 k \\
 & (2\pi)^4 \times \frac{1}{((k + p^r_1 - p_1)^2 - M^2 + i\epsilon)((k - p^r_2)^2 - M^2 + i\epsilon)} \\
 & \left. \int \frac{M^2 + i\epsilon}{4 \quad 8} \right)
 \end{aligned}$$

These integrals can be tricky. For large k , this integral goes as $d k/k$, which is at least convergent as $k \rightarrow \infty$. But this won't always be the case!

3.5.1 Mandelstam Variables

We see that in many of the amplitudes above — in particular those that include the exchange of just a single particle — the same combinations of momenta are appearing frequently in the denominators. There are standard names for various sums and differences of momenta: they are known as *Mandelstam variables*. They are

$$\begin{aligned}s &= (p_1 + p_2)^2 = (p_1^r + p_2^r)^2 \\t &= (p_1 - p_2^r)^2 = (p_2 - p_1^r)^2 \\u &= (p_1 - p_2^r)^2 = (p_2 - p_1^r)^2\end{aligned}\tag{3.60}$$

where, as in the examples above, p_1 and p_2 are the momenta of the two initial particles, and p_1^r and p_2^r are the momenta of the final two particles. We can define these variables whether the particles involved in the scattering are the same or different. To get a feel for what these variables mean, let's assume all four particles are the same. We sit in the center of mass frame, so that the initial two particles have four-momenta

$$p_1 = (E, 0, 0, p) \text{ and } p_2 = (E, 0, 0, -p)\tag{3.61}$$

The particles then scatter at some angle ϑ and leave with momenta

$$p_1^r = (E, 0, p \sin \vartheta, p \cos \vartheta) \text{ and } p_2^r = (E, 0, -p \sin \vartheta, -p \cos \vartheta)\tag{3.62}$$

Then from the above definitions, we have that

$$s = 4E^2 \quad \text{and} \quad t = -2p^2(1 - \cos \vartheta) \quad \text{and} \quad u = -2p^2(1 + \cos \vartheta)\tag{3.63}$$

The variable s measures the total center of mass energy of the collision, while the variables t and u are measures of the momentum exchanged between particles. (They are basically equivalent, just with the outgoing particles swapped around). Now the amplitudes that involve exchange of a single particle can be written simply in terms of the Mandelstam variables. For example, for nucleon-nucleon scattering, the amplitude (3.56) is schematically $A \sim (t - m^2)^{-1} + (u - m^2)^{-1}$. For the nucleon-anti-nucleon scattering, the amplitude (3.59) is $A \sim (t - m^2)^{-1} + (s - m^2)^{-1}$. We say that the first case involves “t-channel” and “u-channel” diagrams. Meanwhile the nucleon-anti-nucleon scattering is said to involve “t-channel” and “s-channel” diagrams. (The first diagram indeed includes a vertex that looks like the letter “T”).

Note that there is a relationship between the Mandelstam variables. When all the masses are the same we have $s + t + u = 4M^2$. When the masses of all 4 particles differ,

$$\text{this becomes } s + t + u = \sum_i M_i^2.$$

3.5.2 The Yukawa Potential

So far we've computed the quantum amplitudes for various scattering processes. But these quantities are a little abstract. In Section 3.6 below (and again in next term's "Standard Model" course) we'll see how to turn amplitudes into measurable quantities such as cross-sections, or the lifetimes of unstable particles. Here we'll instead show how to translate the amplitude (3.52) for nucleon scattering into something familiar from Newtonian mechanics: a potential, or force, between the particles.

Let's start by asking a simple question in classical field theory that will turn out to be relevant. Suppose that we have a fixed δ -function source for a real scalar field φ , that persists for all time. What is the profile of $\varphi(\vec{x})$? To answer this, we must solve the static Klein-Gordon equation,

$$-\nabla^2 \varphi + m^2 \varphi = \delta^{(3)}(\vec{x}) \quad (3.64)$$

We can solve this using the Fourier transform,

$$\varphi(\vec{x}) = \int \frac{d^3 k}{(2\pi)^3} e^{i \vec{k} \cdot \vec{x}} \tilde{\varphi}(\vec{k}) \quad (3.65)$$

Plugging this into (3.64) tells us that $(\vec{k}^2 + m^2) \tilde{\varphi}(\vec{k}) = 1$, giving us the solution

$$\varphi(\vec{x}) = \frac{\int d^3 k}{(2\pi)^3} \frac{e^{i \vec{k} \cdot \vec{x}}}{\vec{k}^2 + m^2} \quad (3.66)$$

Let's now do this integral. Changing to polar coordinates, and writing $\vec{k} \cdot \vec{x} = kr \cos \theta$, we have

$$\begin{aligned} \varphi(\vec{x}) &= \frac{1}{(2\pi)^2} \int_0^\infty dk \frac{k_2}{k^2 + m^2} \frac{2 \sin kr}{kr} \\ &= \frac{1}{(2\pi)^2 r} \int_0^{+\infty} dk \frac{k \sin kr}{\frac{kr}{m^2} +} \\ &= \frac{1}{2\pi r} \operatorname{Re} \int_{-\infty}^{+\infty} dk \frac{ke^{ikr}}{2\pi i k^2 +} \end{aligned} \quad (3.67)$$

We compute this last integral by closing the contour in the upper half plane $k \rightarrow +i\infty$, picking up the pole at $k = +im$. This gives

$$\varphi(\vec{x}) = \frac{1}{4\pi r} e^{-mr} \quad (3.68)$$

The field dies off exponentially quickly at distances $1/m$, the Compton wavelength of the meson.

Now we understand the profile of the φ field, what does this have to do with the force between ψ particles? We do very similar calculations to that above in electrostatics where a charged particle acts as a δ -function source for the gauge potential: $-\nabla^2 A_0 = \delta^{(3)}(\vec{r})$, which is solved by $A_0 = 1/4\pi r$. The profile for A_0 then acts as the potential energy for another charged (test) particle moving in this background. Can we give the same interpretation to our scalar field? In other words, is there a classical limit of the scalar Yukawa theory where the ψ particles act as δ -function sources for φ , creating the profile (3.68)? And, if so, is this profile then felt as a static potential? The answer is essentially yes, at least in the limit $M \gg m$. But the correct way to describe the potential felt by the ψ particles is not to talk about classical fields at all, but instead work directly with the quantum amplitudes.

Our strategy is to compare the nucleon scattering amplitude (3.52) to the corresponding amplitude in non-relativistic quantum mechanics for two particles interacting through a potential. To make this comparison, we should first take the non-relativistic limit of (3.52). Let's work in the center of mass frame, with $p \rightarrow \equiv p \rightarrow_1 = -p \rightarrow_2$ and $p \rightarrow' \equiv p \rightarrow_1' = -p \rightarrow_2'$. The non-relativistic limit means $|p \rightarrow| \ll M$ which, by momentum

conservation, ensures that $|p \rightarrow'| \ll M$. In fact one can check that, for this particular example, this limit doesn't change the scattering amplitude (3.52): it's given by

$$iA = +ig^2 \frac{1}{(p \rightarrow - p \rightarrow')^2 + m^2} + \frac{1}{(p \rightarrow + p \rightarrow')^2 + m^2} \quad (3.69)$$

How do we compare this to scattering in quantum mechanics? Consider two particles, separated by a distance \vec{r} , interacting through a potential $U(\vec{r})$. In non-relativistic quantum mechanics, the amplitude for the particles to scatter from momentum states $\pm p \rightarrow$ into momentum states $\pm p \rightarrow'$ can be computed in perturbation theory, using the techniques described in Section 3.1. To leading order, known in this context as the Born approximation, the amplitude is given by

$$\langle p \rightarrow' | U(\vec{r}) | p \rightarrow \rangle = -i \int d^3r U(r) e^{-i(p \rightarrow - p \rightarrow') \cdot \vec{r}} \quad (3.70)$$

There's a relative factor of $(2M)^2$ that arises in comparing the quantum field theory amplitude A to $\langle p \rightarrow' | U(\vec{r}) | p \rightarrow \rangle$, that can be traced to the relativistic normalization of the states $|p_1, p_2\rangle$. (It is also necessary to get the dimensions of the potential to work out correctly). Including this factor, and equating the expressions for the two amplitudes, we get

$$\int d^3r U(\vec{r}) e^{-i(p \rightarrow - p \rightarrow') \cdot \vec{r}} = \frac{-\lambda_2}{(p \rightarrow - p \rightarrow')^2 + m^2} \quad (3.71)$$

where we've introduced the dimensionless parameter $\lambda = g/2M$. We can trivially invert this to find,

$$U(\rightarrow r) = -\frac{\lambda^2}{(2\pi)^3} \int \frac{d^3 p}{p^2 + m^2} e^{ip \cdot \rightarrow r} \quad (3.72)$$

But this is exactly the integral (3.66) we just did in the classical theory. We have

$$U(\rightarrow r) = \frac{-\lambda^2}{4\pi r} e^{-mr} \quad (3.73)$$

This is the *Yukawa potential*. The force has a range $1/m$, the Compton wavelength of the exchanged particle. The minus sign tells us that the potential is attractive.

Notice that quantum field theory has given us an entirely new perspective on the nature of forces between particles. Rather than being a fundamental concept, the force arises from the virtual exchange of other particles, in this case the meson. In Section 6 of these lectures, we will see how the Coulomb force arises from quantum field theory due to the exchange of virtual photons.

We could repeat the calculation for nucleon-anti-nucleon scattering. The amplitude from field theory is given in (3.59). The first term in this expression gives the same result as for nucleon-nucleon scattering *with the same sign*. The second term vanishes in the non-relativistic limit (it is an example of an interaction that doesn't have a simple Newtonian interpretation). There is no longer a factor of $1/2$ in (3.70), because the incoming/outgoing particles are not identical, so we learn that the potential between a nucleon and anti-nucleon is again given by (3.73). This reveals a key feature of forces arising due to the exchange of scalars: they are universally attractive. Notice that this is different from forces due to the exchange of a spin 1 particle — such as electromagnetism — where the sign flips when we change the charge. However, for forces due to the exchange of a spin 2 particle — i.e. gravity — the force is again universally attractive.

3.5.3 φ^4 Theory

Let's briefly look at the Feynman rules and scattering amplitudes for the interaction Hamiltonian

$$H_{\text{int}} = \frac{\lambda}{4!} \varphi^4 \quad (3.74)$$

The theory now has a single interaction vertex, which comes with a factor of $(-i\lambda)$, while the other Feynman rules remain the same. Note that we assign $(-i\lambda)$ to the

vertex rather than ($-i\lambda/4!$). To see why this is, we can look at $\varphi\varphi \rightarrow \varphi\varphi$ scattering, which has its lowest contribution at order λ , with the term

$$\frac{-i\lambda}{4!} \langle p^r_1, p^r_2 | : \varphi(x)\varphi(x)\varphi(x)\varphi(x) : | p_1, p_2 \rangle \quad (3.75)$$

Any one of the fields can do the job of annihilation or creation. This gives $4!$ different contractions, which cancels the $1/4!$ sitting out front.

Feynman diagrams in the φ^4 theory sometimes come with extra combinatoric factors (typically 2 or 4) which are known as symmetry factors that one must take into account. For more details, see the book by Peskin and Schroeder.

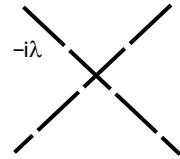


Figure 16:

Using the Feynman rules, the scattering amplitude for $\varphi\varphi \rightarrow \varphi\varphi$ is simply $iA = -i\lambda$. Note that it doesn't depend on the angle at which the outgoing particles emerge: in φ^4 theory the leading order two-particle scattering occurs with equal probability in all directions. Translating this into a potential between two mesons, we have

$$U(\vec{r}) = \frac{\lambda}{(2m)^2} \int \frac{d^3p}{(2\pi)^3} e^{+ip\cdot\vec{r}} = \frac{\lambda}{(2m)^2} \delta^{(3)}(\vec{r}) \quad (3.76)$$

So scattering in φ^4 theory is due to a δ -function potential. The particles don't know what hit them until it's over.

3.5.4 Connected Diagrams and Amputated Diagrams

We've seen how one can compute scattering amplitudes by writing down all Feynman diagrams and assigning integrals to them using the Feynman rules. In fact, there are a couple of caveats about what Feynman diagrams you should write down. Both of these caveats are related to the assumption we made earlier that "initial and final states are eigenstates of the free theory" which, as we mentioned at the time, is not strictly accurate. The two caveats which go some way towards ameliorating the problem are the following

- We consider only connected Feynman diagrams, where every part of the diagram is connected to at least one external line. As we shall see shortly, this will be related to the fact that the vacuum $|0\rangle$ of the free theory is not the true vacuum $|\Omega\rangle$ of the interacting theory. An example of a diagram that is not connected is shown in Figure 17.

- We do not consider diagrams with loops on external lines, for example the diagram shown in the Figure 18. We will not explain how to take these into account in this course, but you will discuss them next term. They are related to the fact that the one-particle states of the free theory are not the same as the one-particle states of the interacting theory. In particular, correctly dealing with these diagrams will account for the fact that particles in interacting quantum field theories are never alone, but surrounded by a cloud of virtual particles. We will refer to diagrams in which all loops on external legs have been cut-off as “amputated”.

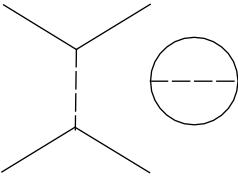


Figure 17: A disconnected diagram.

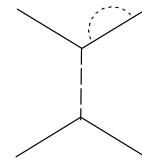


Figure 18: An un-amputated diagram

3.6 What We Measure: Cross Sections and Decay Rates

So far we've learnt to compute the quantum amplitudes for particles decaying or scattering. As usual in quantum theory, the probabilities for things to happen are the (modulus) square of the quantum amplitudes. In this section we will compute these probabilities, known as decay rates and cross sections. One small subtlety here is that the S-matrix elements $\langle f | S - 1 | i \rangle$ all come with a factor of $(2\pi)^4 \delta^{(4)}(p_F - p_I)$, so we end up with the square of a delta-function. As we will now see, this comes from the fact that we're working in an infinite space.

3.6.1 Fermi's Golden Rule

Let's start with something familiar and recall how to derive Fermi's golden rule from Dyson's formula. For two energy eigenstates $|m\rangle$ and $|n\rangle$, with $E_m = E_n$, we have to leading order in the interaction,

$$\begin{aligned}
 \langle m | U(t) | n \rangle &= -i \langle m | \int_0^t dt' H_{\text{I}}(t') | n \rangle \\
 &= -i \langle m | H_{\text{int}} | n \rangle \int_0^\infty dt' e^{i\omega t'} \\
 &= -\langle m | H_{\text{int}} | n \rangle \frac{e^{i\omega t} - 1}{\omega}
 \end{aligned} \tag{3.77}$$

where $\omega = E_m - E_n$. This gives us the probability for the transition from $|n\rangle$ to $|m\rangle$ in time t , as

$$P_{n \rightarrow m}(t) = |\langle m | U(t) | n \rangle|^2 = 2|\langle m | H_{\text{int}} | n \rangle|^2 \frac{1 - \cos \omega t}{\omega^2} \quad (3.78)$$

The function in brackets is plotted in Figure 19 for fixed t . We see that in time t , most transitions happen in a region between energy eigenstates separated by $\Delta E = 2\pi/t$. As $t \rightarrow \infty$, the function in the figure starts to approach a delta-function. To find the normalization, we can calculate

$$\begin{aligned} & \int_{-\infty}^{+\infty} d\omega \frac{1 - \cos \omega t}{\omega^2} = \pi t \\ \Rightarrow & \frac{1 - \cos \omega t}{\omega^2} \rightarrow \pi t \delta(\omega) \quad \text{as } t \rightarrow \infty \end{aligned}$$

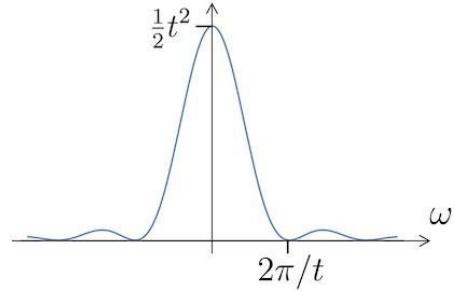


Figure 19:

Consider now a transition to a cluster of states with density $\rho(E)$. In the limit $t \rightarrow \infty$, we get the transition probability

$$\begin{aligned} P_{n \rightarrow m} &= \int_m^{\infty} dE \rho(E_m) \frac{2|\langle m | H_{\text{int}} | n \rangle|^2}{H} \frac{1 - \cos \omega t}{\omega^2} \\ &\rightarrow 2\pi |\langle m | H_{\text{int}} | n \rangle|^2 \rho(E_n) t \end{aligned} \quad (3.79)$$

which gives a constant probability for the transition per unit time for states around the same energy $E_n \sim E_m = E$.

$$\dot{P}_{n \rightarrow m} = 2\pi |\langle m | H_{\text{int}} | n \rangle|^2 \rho(E) \quad (3.80)$$

This is Fermi's Golden Rule.

In the above derivation, we were fairly careful with taking the limit as $t \rightarrow \infty$. Suppose we were a little sloppier, and first chose to compute the amplitude for the state $|n\rangle$ at $t \rightarrow -\infty$ to transition to the state $|m\rangle$ at $t \rightarrow +\infty$. Then we get

$$\int_{t=-\infty}^{t=+\infty} -i \langle m | H_l(t) | n \rangle = -i \langle m | H_{\text{int}} | n \rangle 2\pi \delta(\omega) \quad (3.81)$$

Now when squaring the amplitude to get the probability, we run into the problem of the square of the delta-function: $P_{n \rightarrow m} = |\langle m | H_{\text{int}} | n \rangle|^2 (2\pi)^2 \delta(\omega)^2$. Tracking through the previous computations, we realize that the extra infinity is coming because $P_{m \rightarrow n}$

is the probability for the transition to happen in infinite time $t \rightarrow \infty$. We can write the delta-functions as

$$(2\pi)^2 \delta(\omega)^2 = (2\pi) \delta(\omega) T \quad (3.82)$$

where T is shorthand for $t \rightarrow \infty$ (we used a very similar trick when looking at the vacuum energy in (2.25)). We now divide out by this power of T to get the transition probability per unit time,

$$P_{n \rightarrow m} = 2\pi | \langle m | H_{\text{int}} | n \rangle |^2 \delta(\omega) \quad (3.83)$$

which, after integrating over the density of final states, gives us back Fermi's Golden rule. The reason that we've stressed this point is because, in our field theory calculations, we've computed the amplitudes in the same way as (3.81), and the square of the $\delta^{(4)}$ -functions will just be re-interpreted as spacetime volume factors.

3.6.2 Decay Rates

Let's now look at the probability for a single particle $|i\rangle$ of momentum p_i (i=initial) to decay into some number of particles $|f\rangle$ with momentum p_f and total momentum

$p_F = \sum_i p_i$. This is given by

$$P = \frac{|\langle f | S | i \rangle|^2}{\langle f | f \rangle \langle i | i \rangle} \quad (3.84)$$

Our states obey the relativistic normalization formula (2.65),

$$\langle i | i \rangle = (2\pi)^3 2E_{p \rightarrow i} \delta^{(3)}(0) = 2E_{p \rightarrow i} V \quad (3.85)$$

where we have replaced $\delta^{(3)}(0)$ by the volume of 3-space. Similarly,

$$\langle f | f \rangle = \sum_{\text{final states}} \frac{1}{2E_{p \rightarrow f} V} \quad (3.86)$$

If we place our initial particle at rest, so $p_{p \rightarrow i} = 0$ and $E_{p \rightarrow i} = m$, we get the probability for decay

$$P = \frac{|A_{fi}|^2}{2mV} \frac{(2\pi)^3}{(2\pi)^4} \frac{(p_i - p_f) V T}{\sum_{\text{final states}} \frac{1}{2E_{p \rightarrow f} V}} \quad (3.87)$$

where, as in the second derivation of Fermi's Golden Rule, we've exchanged one of the delta-functions for the volume of spacetime: $(2\pi)^4 \delta^{(4)}(0) = V T$. The amplitudes A_{fi} are, of course, exactly what we've been computing. (For example, in (3.30), we saw

that $A = -g$ for a single meson decaying into two nucleons). We can now divide out by T to get the transition function per unit time. But we still have to worry about summing over all final states. There are two steps: the first is to integrate over all possible momenta of the final particles: $\int d^3 p_f / (2\pi)^3$. The factors of spatial volume V in this measure cancel those in (3.87), while the factors of $1/2E_{p \rightarrow i}$ in (3.87) conspire to produce the Lorentz invariant measure for 3-momentum integrals. The result is an expression for the density of final states given by the Lorentz invariant measure

$$d\Pi = \frac{1}{(2\pi)^4} \frac{(p_F - p_i)}{\delta} \sum_{\text{final states}} \frac{d^3 p_i}{(2\pi)^3} \frac{1}{2E} \quad (3.88)$$

The second step is to sum over all final states with different numbers (and possibly types) of particles. This gives us our final expression for the decay probability per unit time, $\Gamma = P$.

$$\Gamma = \frac{1}{2m} \sum_{\text{final states}} |A_{fi}|^2 d\Pi \quad (3.89)$$

Γ is called the width of the particle. It is equal to the reciprocal of the half-life $\tau = 1/\Gamma$.

3.6.3 Cross Sections

Collide two beams of particles. Sometimes the particles will hit and bounce off each other; sometimes they will pass right through. The fraction of the time that they collide is called the *cross section* and is denoted by σ . If the incoming flux F is defined to be the number of incoming particles per area per unit time, then the total number of scattering events N per unit time is given by,

$$N = F\sigma \quad (3.90)$$

We would like to calculate σ from quantum field theory. In fact, we can calculate a more sensitive quantity $d\sigma$ known as the *differential cross section* which is the probability for a given scattering process to occur in the solid angle (ϑ, φ) . More precisely

$$d\sigma = \frac{\text{Differential Probability}}{\text{Unit Time} \times \text{Unit Flux}} = \frac{1}{4E_1 E_2 V F} |A_{fi}|^2 d\Pi \quad (3.91)$$

where we've used the expression for probability per unit time that we computed in the previous subsection. E_1 and E_2 are the energies of the incoming particles. We now need an expression for the unit flux. For simplicity, let's sit in the center of mass frame of the collision. We've been considering just a single particle per spatial volume V ,

meaning that the flux is given in terms of the 3-velocities $\rightarrow v_i$ as $F = |\rightarrow v_1 - \rightarrow v_2|/V$.

This then gives,

$$d\sigma = \frac{1}{4E_1 E_2} \frac{1}{|\rightarrow v_1 - \rightarrow v_2|} |A_{fi}|^2 d\Pi \quad (3.92)$$

If you want to write this in terms of momentum, then recall from your course on special relativity that the 3-velocities $\rightarrow v_i$ are related to the momenta by $\rightarrow v = p/\sqrt{m(1-v^2)} = \rightarrow p/p^0$.

Equation (3.92) is our final expression relating the S-matrix to the differential cross section. You may now take your favorite scattering amplitude, and compute the probability for particles to fly out at your favorite angles. This will involve doing the integral over the phase space of final states, with measure $d\Pi$. Notice that different scattering amplitudes have different momentum dependence and will result in different angular dependence in scattering amplitudes. For example, in φ^4 theory the amplitude for tree level scattering was simply $A = -\lambda$. This results in isotropic scattering. In contrast,

for nucleon-nucleon scattering we have schematically $A \sim (t - m^2)^{-1} + (u - m^2)^{-1}$. This gives rise to angular dependence in the differential cross-section, which follows from the fact that, for example, $t = -2|\rightarrow p|^2(1 - \cos\vartheta)$, where ϑ is the angle between the incoming and outgoing particles.

3.7 Green's Functions

So far we've learnt to compute scattering amplitudes. These are nice and physical (well – they're directly related to cross-sections and decay rates which are physical) but there are many questions we want to ask in quantum field theory that aren't directly related to scattering experiments. For example, we might want to compute the viscosity of the quark gluon plasma, or the optical conductivity in a tentative model of strange metals, or figure out the non-Gaussianity of density perturbations arising in the CMB from novel models of inflation. All of these questions are answered in the framework of quantum field theory by computing elementary objects known as *correlation functions*. In this section we will briefly define correlation functions, explain how to compute them using Feynman diagrams, and then relate them back to scattering amplitudes. We'll leave the relationship to other physical phenomena to other courses.

We'll denote the true vacuum of the interacting theory as $|\Omega\rangle$. We'll normalize H such that

$$H |\Omega\rangle = 0 \quad (3.93)$$

and $\langle \Omega | \Omega \rangle = 1$. Note that this is different from the state we've called $|0\rangle$ which is the vacuum of the free theory and satisfies $H_0 |0\rangle = 0$. Define

$$G^{(n)}(x_1, \dots, x_n) = \langle \Omega | T \varphi_H(x_1) \dots \varphi_H(x_n) | \Omega \rangle \quad (3.94)$$

where φ_H is φ in the Heisenberg picture of the full theory, rather than the interaction picture that we've been dealing with so far. The $G^{(n)}$ are called correlation functions, or *Green's functions*. There are a number of different ways of looking at these objects which tie together nicely. Let's start by asking how to compute $G^{(n)}$ using Feynman diagrams. We prove the following result

Claim: We use the notation $\varphi_1 = \varphi(x_1)$, and write φ_{1H} to denote the field in the Heisenberg picture, and φ_{1I} to denote the field in the interaction picture. Then

$$\frac{G^{(n)}(x_1, \dots, x_n)}{\varphi} = \langle \Omega | T \varphi_{1I} \dots \varphi_{nI} S | \Omega \rangle \quad (3.95)$$

$1 \quad n \quad 1H$

$$= \frac{\langle 0 | T \varphi_{1I} \dots \varphi_{nI} S | 0 \rangle}{\langle 0 | S | 0 \rangle}$$

where the operators on the right-hand side are evaluated on $|0\rangle$, the vacuum of the free theory.

Proof: Take $t_1 > t_2 > \dots > t_n$. Then we can drop the T and write the numerator of the right-hand side as

$$\langle 0 | U_I(+\infty, t_1) \varphi_{1I} U(t_1, t_2) \varphi_{2I} \dots \varphi_{nI} U_I(t_n, -\infty) | 0 \rangle$$

We'll use the factors of $U_I(t_k, t_{k+1}) = T \exp(-i \frac{t_{k+1}}{t_k} H)$ to convert each of the φ_i into φ_H and we choose operators in the two pictures to be equal at some arbitrary time t_0 . Then we can write

$$\begin{aligned} \langle 0 | U_I(+\infty, t_1) \varphi_{1I} U(t_1, t_2) \varphi_{2I} \dots \varphi_{nI} U_I(t_n, -\infty) | 0 \rangle \\ = \langle 0 | U_I(+\infty, t_0) \varphi_{1H} \dots \varphi_{nH} U_I(t_0, -\infty) | 0 \rangle \end{aligned}$$

Now let's deal with the two remaining $U(t_0, \pm\infty)$ at either end of the string of operators. Consider an arbitrary state $|\Psi\rangle$ and look at

$$\langle \Psi | U(t, -\infty) | 0 \rangle = \langle \Psi | U(t, -\infty) | 0 \rangle \quad (3.96)$$

where $U(t, -\infty)$ is the Schrödinger evolution operator, and the equality above follows because $H_0 |0\rangle = 0$. Now insert a complete set of states, which we take to be energy eigenstates of $H = H_0 + H_{int}$,

$$\begin{aligned} \langle \Psi | U(t, -\infty) | 0 \rangle &= \langle \Psi | U(t, -\infty) | \Omega \rangle \langle \Omega | \sum_{n=0}^{\infty} |n\rangle \langle n| | 0 \rangle \\ &+ \langle \Psi | \Omega \rangle \langle \Omega | 0 \rangle \lim_{t' \rightarrow -\infty} \sum_{n=0}^{\infty} e^{iE_n(t' - t)} \langle \Psi | n \rangle \langle n | 0 \rangle \quad (3.97) \end{aligned}$$

But the last term vanishes. This follows from the Riemann-Lebesgue lemma which says that for any well-behaved function

$$\lim_{\mu \rightarrow \infty} \int_a^b dx f(x) e^{i\mu x} = 0 \quad (3.98)$$

\sum

Why is this relevant? The point is that the \int_n in (3.97) is really an integral dn , because all states are part of a continuum due to the momentum. (There is a caveat here: we want the vacuum $|\Omega\rangle$ to be special, so that it sits on its own, away from the continuum of the integral. This means that we must be working in a theory with a mass gap – i.e. with no massless particles). So the Riemann-Lebesgue lemma gives us

$$\lim_{t' \rightarrow -\infty} \langle \Psi | U(t, t') | 0 \rangle = \langle \Psi | \Omega \rangle \langle \Omega | 0 \rangle \quad (3.99)$$

(Notice that to derive this result, Peskin and Schroeder instead send $t \rightarrow -\infty$ in a slightly imaginary direction, which also does the job). We now apply the formula (3.99), to the top and bottom of the right-hand side of (3.95) to find

$$\frac{\langle 0 | \Omega \rangle \langle \Omega | T\varphi_1 \dots \varphi_n | \Omega \rangle \langle \Omega | 0 \rangle}{\langle 0 | \Omega \rangle \langle \Omega | \Omega \rangle \langle \Omega | 0 \rangle} \quad (3.100)$$

which, using the normalization $\langle \Omega | \Omega \rangle = 1$, gives us the left-hand side, completing the proof.

3.7.1 Connected Diagrams and Vacuum Bubbles

We're getting closer to our goal of computing the Green's functions $G^{(n)}$ since we can compute both $\langle 0 | T\varphi_1(x_1) \dots \varphi_n(x_n) S | 0 \rangle$ and $\langle 0 | S | 0 \rangle$ using the same methods we developed for S-matrix elements; namely Dyson's formula and Wick's theorem or, alternatively, Feynman diagrams. But what about dividing one by the other? What's that all about? In fact, it has a simple interpretation. For the following discussion, we will work in φ^4 theory. Since there is no ambiguity in the different types of lines in Feynman diagrams, we will represent the φ particles as solid lines, rather than the dashed lines that we used previously. Then we have the diagrammatic expansion for $\langle 0 | S | 0 \rangle$.

$$\langle 0 | S | 0 \rangle = 1 + \text{Diagram } 1 + \left(\text{Diagram } 2 + \text{Diagram } 3 + \text{Diagram } 4 \right) + \dots \quad (3.101)$$

These diagrams are called vacuum bubbles. The combinatoric factors (as well as the symmetry factors) associated with each diagram are such that the whole series sums

to an exponential,

$$\langle 0 | S | 0 \rangle = \exp \left(\text{ } \text{ } + \text{ } \text{ } + \text{ } \text{ } + \dots \right) \quad (3.102)$$

So the amplitude for the vacuum of the free theory to evolve into itself is $\langle 0 | S | 0 \rangle = \exp(\text{all distinct vacuum bubbles})$. A similar combinatoric simplification occurs for generic correlation functions. Remarkably, the vacuum diagrams all add up to give the same exponential. With a little thought one can show that

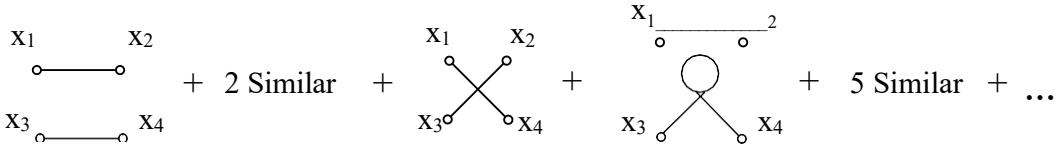
$$\langle 0 | T\varphi_1 \dots \varphi_n S | 0 \rangle = \sum_{\text{connected diagrams}} \langle 0 | S | 0 \rangle \quad (3.103)$$

where “connected” means that every part of the diagram is connected to at least one of the external legs. The upshot of all this is that dividing by $\langle 0 | S | 0 \rangle$ has a very nice interpretation in terms of Feynman diagrams: we need only consider the connected Feynman diagrams, and don’t have to worry about the vacuum bubbles. Combining this with (3.95), we learn that the Green’s functions $G^{(n)}(x_1 \dots, x_n)$ can be calculated by summing over all connected Feynman diagrams,

$$\langle \Omega | T\varphi_H(x_1) \dots \varphi_H(x_n) | \Omega \rangle = \sum_{\text{Connected Feynman Graphs}} \langle \Omega | \dots | \Omega \rangle \quad (3.104)$$

An Example: The Four-Point Correlator: $\langle \Omega | T\varphi_H(x_1) \dots \varphi_H(x_4) | \Omega \rangle$

As a simple example, let’s look at the four-point correlation function in φ^4 theory. The sum of connected Feynman diagrams is given by,



All of these are connected diagrams, even though they don’t look that connected! The point is that a connected diagram is defined by the requirement that every line is joined to an external leg. An example of a diagram that is not connected is shown in the figure. As we have seen, such diagrams are taken care of in shifting the vacuum from $|0\rangle$ to $|\Omega\rangle$.

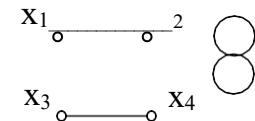


Figure 20:

Feynman Rules

The Feynman diagrams that we need to calculate for the Green’s functions depend on x_1, \dots, x_n . This is rather different than the Feynman diagrams that we calculated for

the S-matrix elements, where we were working primarily with momentum eigenstates, and ended up integrating over all of space. However, it's rather simple to adapt the Feynman rules that we had earlier in momentum space to compute $G^{(n)}(x_1, \dots, x_n)$. For φ^4 theory, we have

- Draw n external points x_1, \dots, x_n , connected by the usual propagators and vertices. Assign a spacetime position y to the end of each line.
- For each line $x \xrightarrow{ } y$ from x to y write down a factor of the Feynman propagator $\Delta_F(x - y)$.
- For each vertex \times_y at position y , write down a factor of $-i\lambda \int d^4y$.

3.7.2 From Green's Functions to S-Matrices

Having described how to compute correlation functions using Feynman diagrams, let's now relate them back to the S-matrix elements that we already calculated. The first step is to perform the Fourier transform,

$$\tilde{G}^{(n)}(p_1, \dots, p_n) = \int_{i=1}^n d^4x_i e^{-ip_i x_i} G^{(n)}(x_1, \dots, x_n) \quad (3.105)$$

These are very closely related to the S-matrix elements that we've computed above. The difference is that the Feynman rules for $G^{(n)}(x_1, \dots, x_n)$, effectively include propagators Δ_F for the external legs, as well as the internal legs. A related fact is that the 4-momenta assigned to the external legs is arbitrary: they are not on-shell. Both of these problems are easily remedied to allow us to return to the S-matrix elements: we need to simply cancel off the propagators on the external legs, and place their momentum back on shell. We have

$$\langle p_r, \dots, p_n | S - 1 | p_1, \dots, p_r \rangle = \sum_{i=1}^{n+n'} \sum_{j=1}^{2} \frac{Y_{n'}(p_i - m)}{p_i^2 - m^2} Y_n(p_j - m) \times \tilde{G}^{(n+n')}(p_1, \dots, p_r, p_j, \dots, p_n) \quad (3.106)$$

Each of the factors $(p^2 - m^2)$ vanishes once the momenta are placed on-shell. This means that we only get a non-zero answer for diagrams contributing to $G^{(n)}(x_1, \dots, x_n)$ which have propagators for each external leg.

So what's the point of all of this? We've understood that ignoring the unconnected diagrams is related to shifting to the true vacuum $|\Omega\rangle$. But other than that, introducing the Green's functions seems like a lot of bother for little reward. The important point

is that this provides a framework in which to deal with the true particle states in the interacting theory through renormalization. Indeed, the formula (3.106), suitably interpreted, remains true even in the interacting theory, taking into account the swarm of virtual particles surrounding asymptotic states. This is the correct way to consider scattering. In this context, (3.106) is known as the LSZ reduction formula. You will derive it properly next term.

4. The Dirac Equation

"A great deal more was hidden in the Dirac equation than the author had expected when he wrote it down in 1928. Dirac himself remarked in one of his talks that his equation was more intelligent than its author. It should be added, however, that it was Dirac who found most of the additional insights."

Weisskopf on Dirac

So far we've only discussed scalar fields such that under a Lorentz transformation $x^\mu \rightarrow (x^r)^\mu = \Lambda^\mu{}_\nu x_\nu$, the field transforms as

$$\varphi(x) \rightarrow \varphi^r(x) = \varphi(\Lambda^{-1}x) \quad (4.1)$$

We have seen that quantization of such fields gives rise to spin 0 particles. But most particles in Nature have an intrinsic angular momentum, or spin. These arise naturally in field theory by considering fields which themselves transform non-trivially under the Lorentz group. In this section we will describe the Dirac equation, whose quantization gives rise to fermionic spin 1/2 particles. To motivate the Dirac equation, we will start by studying the appropriate representation of the Lorentz group.

A familiar example of a field which transforms non-trivially under the Lorentz group is the vector field $A_\mu(x)$ of electromagnetism,

$$A^\mu(x) \rightarrow \Lambda^\mu{}_\nu A^\nu(\Lambda^{-1}x) \quad (4.2)$$

We'll deal with this in Section 6. (It comes with its own problems!). In general, a field can transform as

$$\varphi^a(x) \rightarrow D[\Lambda]^a{}_b \varphi^b(\Lambda^{-1}x) \quad (4.3)$$

where the matrices $D[\Lambda]$ form a *representation* of the Lorentz group, meaning that

$$D[\Lambda_1]D[\Lambda_2] = D[\Lambda_1\Lambda_2] \quad (4.4)$$

and $D[\Lambda^{-1}] = D[\Lambda]^{-1}$ and $D[1] = 1$. How do we find the different representations? Typically, we look at infinitesimal transformations of the Lorentz group and study the resulting Lie algebra. If we write,

$$\Lambda^\mu{}_\nu = \delta^\mu{}_\nu + \omega^\mu{}_\nu \quad (4.5)$$

for infinitesimal ω , then the condition for a Lorentz transformation $\Lambda^\mu{}_\sigma \Lambda^\nu{}_\rho \eta^{\sigma\rho} = \eta^{\mu\nu}$ becomes the requirement that ω is anti-symmetric:

$$\omega^{\mu\nu} + \omega^{\nu\mu} = 0 \quad (4.6)$$

Note that an antisymmetric 4×4 matrix has $4 \times 3/2 = 6$ independent components, which agrees with the 6 transformations of the Lorentz group: 3 rotations and 3 boosts. It's going to be useful to introduce a basis of these six 4×4 anti-symmetric matrices. We could call them $(M^A)^{\mu\nu}$, with $A = 1, \dots, 6$. But in fact it's better for us (although initially a little confusing) to replace the single index A with a pair of antisymmetric indices $[\rho\sigma]$, where $\rho, \sigma = 0, \dots, 3$, so we call our matrices $(M^{\rho\sigma})^\mu_\nu$. The antisymmetry on the ρ and σ indices means that, for example, $M^{01} = -M^{10}$, etc, so that ρ and σ again label six different matrices. Of course, the matrices are also antisymmetric on the $\mu\nu$ indices because they are, after all, antisymmetric matrices. With this notation in place, we can write a basis of six 4×4 antisymmetric matrices as

$$(M^{\rho\sigma})^{\mu\nu} = \eta^{\rho\mu} \eta^{\sigma\nu} - \eta^{\sigma\mu} \eta^{\rho\nu} \quad (4.7)$$

where the indices μ and ν are those of the 4×4 matrix, while ρ and σ denote which basis element we're dealing with. If we use these matrices for anything practical (for example, if we want to multiply them together, or act on some field) we will typically need to lower one index, so we have

$$\delta_{\rho}^{\mu} \delta_{\sigma}^{\nu} = \eta^{\rho\mu} \delta^{\sigma} - \eta^{\sigma\mu} \delta^{\rho} \quad (4.8)$$

Since we lowered the index with the Minkowski metric, we pick up various minus signs which means that when written in this form, the matrices are no longer necessarily antisymmetric. Two examples of these basis matrices are,

$$(M^{01})^{\mu}_\nu = \begin{matrix} & \begin{smallmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{smallmatrix} \end{matrix} \quad \text{and} \quad (M^{12})^{\mu}_\nu = \begin{matrix} & \begin{smallmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{smallmatrix} \end{matrix} \quad (4.9)$$

The first, M^{01} , generates boosts in the x^1 direction. It is real and symmetric. The second, M^{12} , generates rotations in the (x^1, x^2) -plane. It is real and antisymmetric. We can now write any ω_v^μ as a linear combination of the $M^{\rho\sigma}$,

$$\omega_v^\mu = \frac{1}{2} \Omega_{\rho\sigma} (M^{\rho\sigma})^\mu_\nu \quad (4.10)$$

where $\Omega_{\rho\sigma}$ are just six numbers (again antisymmetric in the indices) that tell us what Lorentz transformation we're doing. The six basis matrices $M^{\rho\sigma}$ are called the *generators* of the Lorentz transformations. The generators obey the Lorentz Lie algebra relations,

$$[M^{\rho\sigma}, M^{\tau\nu}] = \eta^{\sigma\tau} M^{\rho\nu} - \eta^{\rho\tau} M^{\sigma\nu} + \eta^{\rho\nu} M^{\sigma\tau} - \eta^{\sigma\nu} M^{\rho\tau} \quad (4.11)$$

where we have suppressed the matrix indices. A finite Lorentz transformation can then be expressed as the exponential

$$\Lambda = \exp \frac{1}{2} \Omega_{\rho\sigma} M^{\rho\sigma} \quad (4.12)$$

Let me stress again what each of these objects are: the $M^{\rho\sigma}$ are six 4×4 basis elements of the Lorentz Lie algebra; the $\Omega_{\rho\sigma}$ are six numbers telling us what kind of Lorentz transformation we're doing (for example, they say things like rotate by $\vartheta = \pi/7$ about the x^3 -direction and run at speed $v = 0.2$ in the x^1 direction).

4.1 The Spinor Representation

We're interested in finding other matrices which satisfy the Lorentz algebra commutation relations (4.11). We will construct the spinor representation. To do this, we start by defining something which, at first sight, has nothing to do with the Lorentz group. It is the *Clifford algebra*,

$$\{\gamma^\mu, \gamma^\nu\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2\eta^{\mu\nu} \mathbf{1} \quad (4.13)$$

where γ^μ , with $\mu = 0, 1, 2, 3$, are a set of four matrices and the $\mathbf{1}$ on the right-hand side denotes the unit matrix. This means that we must find four matrices such that

$$\gamma^\mu \gamma^\nu = -\gamma^\nu \gamma^\mu \quad \text{when } \mu \neq \nu \quad (4.14)$$

and

$$(\gamma^0)^2 = 1 , \quad (\gamma^i)^2 = -1 \quad i = 1, 2, 3 \quad (4.15)$$

It's not hard to convince yourself that there are no representations of the Clifford algebra using 2×2 or 3×3 matrices. The simplest representation of the Clifford algebra is in terms of 4×4 matrices. There are many such examples of 4×4 matrices which obey (4.13). For example, we may take

$$\gamma^0 = \begin{matrix} & & & \\ & 0 & 1 & \\ & 1 & 0 & \end{matrix}, \quad \gamma^i = \begin{matrix} & & & \\ & 0 & \sigma^i & \\ & -\sigma^i & 0 & \end{matrix} \quad (4.16)$$

where each element is itself a 2×2 matrix, with the σ^i the Pauli matrices

$$\sigma^1 = \begin{matrix} & & & \\ & 0 & 1 & \\ & 1 & 0 & \end{matrix}, \quad \sigma^2 = \begin{matrix} & & & \\ & 0 & -i & \\ & i & 0 & \end{matrix}, \quad \sigma^3 = \begin{matrix} & & & \\ & 1 & 0 & \\ & 0 & -1 & \end{matrix} \quad (4.17)$$

which themselves satisfy $\{\sigma^i, \sigma^j\} = 2\delta^{ij}$.

One can construct many other representations of the Clifford algebra by taking $V \gamma^\mu V^{-1}$ for any invertible matrix V . However, up to this equivalence, it turns out that there is a unique irreducible representation of the Clifford algebra. The matrices (4.16) provide one example, known as the *Weyl* or *chiral representation* (for reasons that will soon become clear). We will soon restrict ourselves further, and consider only representations of the Clifford algebra that are related to the chiral representation by a unitary transformation V .

So what does the Clifford algebra have to do with the Lorentz group? Consider the commutator of two γ^μ ,

$$S^{\rho\sigma} = \frac{1}{2} [\gamma^\rho, \gamma^\sigma] = \begin{array}{ccccc} & 0 & & & \\ \sigma & & & & \\ & \frac{1}{2} \gamma^\rho \gamma^\sigma & & \rho \neq \sigma & \\ 4 & & & & 2 \end{array} = \frac{1}{2} \gamma^\rho \gamma^\sigma - \frac{1}{2} \eta^{\rho\sigma} \quad (4.18)$$

Let's see what properties these matrices have:

Claim 4.1: $[S^{\mu\nu}, \gamma^\rho] = \gamma^\mu \eta^{\nu\rho} - \gamma^\nu \eta^{\rho\mu}$

Proof: When $\mu \neq \nu$ we have

$$\begin{aligned} [S^{\mu\nu}, \gamma^\rho] &= \frac{1}{2} [\gamma^\mu \gamma^\nu, \gamma^\rho] \\ &= \frac{1}{2} \gamma^\mu \gamma^\nu \gamma^\rho - \frac{1}{2} \gamma^\rho \gamma^\mu \gamma^\nu \\ &= \frac{1}{2} \gamma^\mu \{\gamma^\nu, \gamma^\rho\} - \frac{1}{2} \gamma^\mu \gamma^\rho \gamma^\nu - \frac{1}{2} \{\gamma^\rho, \gamma^\mu\} \gamma^\nu + \frac{1}{2} \gamma^\mu \gamma^\rho \gamma^\nu \\ &= \gamma^\mu \eta^{\nu\rho} - \gamma^\nu \eta^{\rho\mu} \end{aligned}$$

Claim 4.2: The matrices $S^{\mu\nu}$ form a representation of the Lorentz algebra (4.11), meaning

$$[S^{\mu\nu}, S^{\rho\sigma}] = \eta^{\nu\rho} S^{\mu\sigma} - \eta^{\mu\rho} S^{\nu\sigma} + \eta^{\mu\sigma} S^{\nu\rho} - \eta^{\nu\sigma} S^{\mu\rho} \quad (4.19)$$

Proof: Taking $\rho \neq \sigma$, and using Claim 4.1 above, we have

$$\begin{aligned} [S^{\mu\nu}, S^{\rho\sigma}] &= \frac{1}{2} [S^{\mu\nu}, \gamma^\rho \gamma^\sigma] \\ &= \frac{1}{2} [S^{\mu\nu}, \gamma^\rho] \gamma^\sigma + \frac{1}{2} \gamma^\rho [S^{\mu\nu}, \gamma^\sigma] \\ &= \frac{1}{2} \gamma^\mu \gamma^\sigma \eta^{\nu\rho} - \frac{1}{2} \gamma^\nu \gamma^\sigma \eta^{\rho\mu} + \frac{1}{2} \gamma^\rho \gamma^\mu \eta^{\nu\sigma} - \frac{1}{2} \gamma^\rho \gamma^\nu \eta^{\sigma\mu} \end{aligned} \quad (4.20)$$

Now using the expression (4.18) to write $\gamma^\mu \gamma^\sigma = 2S^{\mu\sigma} + \eta^{\mu\sigma}$, we have

$$[S^{\mu\nu}, S^{\rho\sigma}] = S^{\mu\sigma} \eta^{\nu\rho} - S^{\nu\sigma} \eta^{\rho\mu} + S^{\rho\mu} \eta^{\nu\sigma} - S^{\rho\nu} \eta^{\sigma\mu} \quad (4.21)$$

which is our desired expression.

4.1.1 Spinors

The $S^{\mu\nu}$ are 4×4 matrices, because the γ^μ are 4×4 matrices. So far we haven't given an index name to the rows and columns of these matrices: we're going to call them $\alpha, \beta = 1, 2, 3, 4$.

We need a field for the matrices $(S^{\mu\nu})^\alpha{}_\beta$ to act upon. We introduce the Dirac *spinor* field $\psi^\alpha(x)$, an object with four complex components labelled by $\alpha = 1, 2, 3, 4$. Under Lorentz transformations, we have

$$\psi^\alpha(x) \rightarrow S[\Lambda]^\alpha{}_\beta \psi^\beta(\Lambda^{-1}x) \quad (4.22)$$

where

$$\Lambda = \exp \frac{1}{2} \Omega_{\rho\sigma} M^{\rho\sigma} \quad (4.23)$$

$$S[\Lambda] = \exp \frac{1}{2} \Omega_{\rho\sigma} S^{\rho\sigma} \quad (4.24)$$

Although the basis of generators $M^{\rho\sigma}$ and $S^{\rho\sigma}$ are different, we use the same six numbers $\Omega_{\rho\sigma}$ in both Λ and $S[\Lambda]$: this ensures that we're doing the same Lorentz transformation on x and ψ . Note that we denote both the generator $S^{\rho\sigma}$ and the full Lorentz transformation $S[\Lambda]$ as "S". To avoid confusion, the latter will always come with the square brackets $[\Lambda]$.

Both Λ and $S[\Lambda]$ are 4×4 matrices. So how can we be sure that the spinor representation is something new, and isn't equivalent to the familiar representation $\Lambda^\mu{}_\nu$? To see that the two representations are truly different, let's look at some specific transformations.

Rotations

$$S^{ij} = \frac{1}{2} \begin{matrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{matrix} \begin{matrix} 0 & \sigma^j \\ -\sigma^j & 0 \end{matrix} = \frac{i}{2} \epsilon^{ijk} \begin{matrix} \sigma^k & 0 \\ 0 & \sigma^k \end{matrix} \quad (\text{for } i \neq j) \quad (4.25)$$

If we write the rotation parameters as $\Omega_{ij} = -\epsilon_{ijk}\phi^k$ (meaning $\Omega_{12} = -\phi^3$, etc) then the rotation matrix becomes

$$S[\Lambda] = \exp \frac{1}{2} \Omega_{\rho\sigma} S^{\rho\sigma} = \begin{matrix} e^{i\phi_3 \cdot \sigma/2} & 0 \\ 0 & e^{i\phi_3 \cdot \sigma/2} \end{matrix} \quad (4.26)$$

where we need to remember that $\Omega_{12} = -\Omega_{21} = -\phi^3$ when following factors of 2.

Consider now a rotation by 2π about, say, the x^3 -axis. This is achieved by $\phi \rightarrow = (0, 0,$

and the spinor rotation matrix becomes,

$$S[\Lambda] = \begin{vmatrix} e^{+i\pi\sigma^3} & 0 \\ 0 & e^{+i\pi\sigma^3} \end{vmatrix} = -1 \quad (4.27)$$

Therefore under a 2π rotation

Therefore under a 2π rotation

$$\psi^\alpha(x) \rightarrow -\psi^\alpha(x)$$

(4.28)

which is definitely not what happens to a vector! To check that we haven't been cheating with factors of 2, let's see how a vector would transform under a rotation by $\phi \rightarrow = (0, 0, \phi^3)$. We have

$$\Lambda^1 \rho\sigma = \text{exp}^{\frac{1}{2}\rho\sigma} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \phi^3 & 0 \\ -\phi^3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (4.29)$$

So when we rotate a vector by $\phi^3 = 2\pi$, we learn that $\Lambda = 1$ as you would expect. So

$S[\Lambda]$ is definitely a different representation from the familiar vector representation $\Lambda^\mu{}_\nu$.

$$\begin{array}{c}
 \text{B} \\
 \text{o} \\
 \text{o} \\
 \text{s} \\
 \text{t} \\
 \text{s}
 \end{array}
 \begin{array}{c}
 ! & ! & ! \\
 \frac{1}{\sigma^i} & 1 & - \\
 \xi_0 & 1 & =\sigma^i \\
 = & & 0
 \end{array}
 \begin{array}{c}
 (\\
 4 \\
 . \\
 3 \\
 0 \\
)
 \end{array}
 \begin{array}{c}
 2 & 1 & 0 \\
 -\sigma^i & 0
 \end{array}$$

$$\begin{matrix} 2 & & 0 \\ & \sigma^i \end{matrix}$$

Writing the boost parameter as $\Omega_{i0} = -\Omega_{0i} = \chi_i$, we have

$$S[\Lambda] = \begin{matrix} e^{i\vec{\chi} \cdot \vec{\sigma}/2} & \\ 0 & \end{matrix} \quad (4.31)$$

0

$$\begin{matrix} e^{-i\vec{\chi} \cdot \vec{\sigma}/2} & \\ \cdot & / \\ 2 & \end{matrix}$$

Representations of the Lorentz Group are not Unitary

Note that for rotations given in (4.26), $S[\Lambda]$ is unitary, satisfying $S[\Lambda]^\dagger S[\Lambda] = 1$. But for boosts given in (4.31), $S[\Lambda]$ is not unitary. In fact, there are *no* finite dimensional unitary representations of the Lorentz group. We have demonstrated this explicitly for the spinor representation using the chiral representation (4.16) of the Clifford algebra. We can get a feel for why it is true for a spinor representation constructed from any representation of the Clifford algebra.

Recall that

$$S[\Lambda] = \exp \begin{matrix} 1\Omega \\ \rho\sigma \\ S \\ \rho\sigma \end{matrix} \quad (4.32)$$

so the representation is unitary if $S^{\mu\nu}$ are anti-hermitian, i.e. $(S^{\mu\nu})^\dagger = -S^{\mu\nu}$. But we have

$$(S^{\mu\nu})^\dagger = -\frac{1}{2} [(\gamma^\mu)^\dagger, (\gamma^\nu)^\dagger]$$

(4.33)

which can be anti-hermitian if all γ^μ are hermitian or all are anti-hermitian. However, we can never arrange for this to happen since

$$\begin{aligned} (\gamma^0)^2 &= 1 \Rightarrow \text{Real Eigenvalues} \\ (\gamma^i)^2 &= -1 \Rightarrow \text{Imaginary Eigenvalues} \end{aligned} \quad (4.34)$$

So we could pick γ^0 to be hermitian, but we can only pick γ^i to be anti-hermitian. Indeed, in the chiral representation (4.16), the matrices have this property: $(\gamma^0)^\dagger = \gamma^0$ and $(\gamma^i)^\dagger = -\gamma^i$. In general there is no way to pick γ^μ such that $S^{\mu\nu}$ are anti-hermitian.

4.2 Constructing an Action

We now have a new field to work with, the Dirac spinor ψ . We would like to construct a Lorentz invariant equation of motion. We do this by constructing a Lorentz invariant action.

We will start in a naive way which won't work, but will give us a clue how to proceed. Define

$$\psi^\dagger(x) = (\psi^\wedge)^\top(x) \quad (4.35)$$

which is the usual adjoint of a multi-component object. We could then try to form a Lorentz scalar by taking the product $\psi^\dagger \psi$, with the spinor indices summed over. Let's see how this transforms under Lorentz transformations,

$$\begin{aligned} \psi(x) &\rightarrow S[\Lambda] \psi(\Lambda^{-1}x) \\ \psi^\dagger(x) &\rightarrow \psi^\dagger(\Lambda^{-1}x) S[\Lambda]^\dagger \end{aligned} \quad (4.36)$$

So $\psi^\dagger(x)\psi(x) \rightarrow \psi^\dagger(\Lambda^{-1}x)S[\Lambda]^\dagger S[\Lambda]\psi(\Lambda^{-1}x)$. But, as we have seen, for some Lorentz transformation $S[\Lambda]^\dagger S[\Lambda] \neq 1$ since the representation is not unitary. This means that $\psi^\dagger \psi$ isn't going to do it for us: it doesn't have any nice transformation under the Lorentz group, and certainly isn't a scalar. But now we see why it fails, we can also see how to proceed. Let's pick a representation of the Clifford algebra which, like the chiral representation (4.16), satisfies $(\gamma^0)^\dagger = \gamma^0$ and $(\gamma^i)^\dagger = -\gamma^i$. Then for all $\mu = 0, 1, 2, 3$ we have

$$\gamma^0 \gamma^\mu \gamma^0 = (\gamma^\mu)^\dagger \quad (4.37)$$

which, in turn, means that

$$(S^{\mu\nu})^\dagger = \frac{1}{2} [(\gamma^\nu)^\dagger, (\gamma^\mu)^\dagger] = -\gamma^0 S^{\mu\nu} \gamma^0 \quad (4.38)$$

so that

$$S[\Lambda]^\dagger = \exp -\frac{1}{2} \Omega_{\rho\sigma} (S^{\rho\sigma})^\dagger = \gamma^0 S[\Lambda]^{-1} \gamma^0 \quad (4.39)$$

With this in mind, we now define the *Dirac adjoint*

$$\bar{\psi}(x) = \psi^\dagger(x) \gamma^0 \quad (4.40)$$

Let's now see what Lorentz covariant objects we can form out of a Dirac spinor ψ and its adjoint $\bar{\psi}$.

Claim 4.3: $\bar{\psi} \bar{\psi}$ is a Lorentz scalar.

Proof: Under a Lorentz transformation,

$$\begin{aligned} \bar{\psi}(x) \bar{\psi}(x) &= \bar{\psi}^\dagger(x) \gamma^0 \\ \psi(x) &\rightarrow \bar{\psi}^\dagger(\Lambda^{-1}x) S[\Lambda] \gamma^0 S[\Lambda]^\dagger \psi(\Lambda^{-1}x) \\ &= \bar{\psi}^\dagger(\Lambda^{-1}x) \gamma^0 \psi(\Lambda^{-1}x) \\ &= \bar{\psi}(\Lambda^{-1}x) \bar{\psi}(\Lambda^{-1}x) \end{aligned} \quad (4.41)$$

which is indeed the transformation law for a Lorentz scalar.

Claim 4.4: $\bar{\psi} \gamma^\mu \psi$ is a Lorentz vector, which means that

$$\bar{\psi}(x) \gamma^\mu \psi(x) \rightarrow \Lambda^\mu_\nu \bar{\psi}(\Lambda^{-1}x) \gamma^\nu \psi(\Lambda^{-1}x) \quad (4.42)$$

This equation means that we can treat the $\mu = 0, 1, 2, 3$ index on the γ^μ matrices as a true vector index. In particular we can form Lorentz scalars by contracting it with other Lorentz indices.

Proof: Suppressing the x argument, under a Lorentz transformation we have,

$$\bar{\psi} \gamma^\mu \psi \rightarrow \bar{\psi} S[\Lambda]^{-1} \gamma^\mu S[\Lambda] \psi \quad (4.43)$$

If $\bar{\psi} \gamma^\mu \psi$ is to transform as a vector, we must have

$$S[\Lambda]^{-1} \gamma^\mu S[\Lambda] = \Lambda^\mu_\nu \gamma^\nu \quad (4.44)$$

We'll now show this. We work infinitesimally, so that

$$\Lambda = \exp -\frac{1}{2} \Omega_{\rho\sigma} M^{\rho\sigma} \approx 1 + \frac{1}{2} \Omega_{\rho\sigma} M^{\rho\sigma} + \dots \quad (4.45)$$

$$S[\Lambda] = \exp -\frac{1}{2} \Omega_{\rho\sigma} S^{\rho\sigma} \approx 1 + \frac{1}{2} \Omega_{\rho\sigma} S^{\rho\sigma} + \dots \quad (4.46)$$

so the requirement (4.44) becomes

$$-[S^{\rho\sigma}, \gamma^\mu] = (M^{\rho\sigma})_\nu^\mu \gamma^\nu \quad (4.47)$$

where we've suppressed the α, β indices on γ^μ and $S^{\mu\nu}$, but otherwise left all other indices explicit. In fact equation (4.47) follows from Claim 4.1 where we showed that $[S^{\rho\sigma}, \gamma^\mu] = \gamma^\rho \eta^{\sigma\mu} - \gamma^\sigma \eta^{\mu\rho}$. To see this, we write the right-hand side of (4.47) by expanding out M ,

$$\begin{aligned} (M^{\rho\sigma})^\mu \gamma^\nu &= (\eta^{\rho\mu} \delta_\nu^\sigma - \eta^{\sigma\mu} \delta_\nu^\rho) \gamma^\nu \\ &= \eta^{\rho\mu} \gamma^\sigma - \eta^{\sigma\mu} \gamma^\rho \end{aligned} \quad (4.48)$$

which means that the proof follows if we can show

$$\begin{aligned} -[S^{\rho\sigma}, \gamma^\mu] &= \eta^{\rho\mu} \gamma^\sigma - \eta^{\sigma\mu} \\ \gamma^\rho \end{aligned} \quad (4.49)$$

which is exactly what we proved in Claim 4.1.

Claim 4.5: $\bar{\psi} \gamma^\mu \gamma^\nu \psi$ transforms as a Lorentz tensor. More precisely, the symmetric part is a Lorentz scalar, proportional to $\eta^{\mu\nu} \bar{\psi} \psi$, while the antisymmetric part is a Lorentz tensor, proportional to $\bar{\psi} S^{\mu\nu} \psi$.

Proof: As above.

We are now armed with three bilinears of the Dirac field, $\bar{\psi} \psi$, $\bar{\psi} \gamma^\mu \psi$ and $\bar{\psi} \gamma^\mu \gamma^\nu \psi$, each of which transforms covariantly under the Lorentz group. We can try to build a Lorentz invariant action from these. In fact, we need only the first two. We choose

$$S = \int d^4x \bar{\psi}(x) (i\gamma^\mu \partial_\mu - m) \psi(x) \quad (4.50)$$

This is the Dirac action. The factor of "i" is there to make the action real; upon complex conjugation, it cancels a minus sign that comes from integration by parts. (Said another way, it's there for the same reason that the Hermitian momentum operator $-i\nabla$ in quantum mechanics has a factor i). As we will see in the next section, after quantization this theory describes particles and anti-particles of mass $|m|$ and spin 1/2. Notice that the Lagrangian is first order, rather than the second order Lagrangians we were working with for scalar fields. Also, the mass appears in the Lagrangian as m , which can be positive or negative.

4.3 The Dirac Equation

The equation of motion follows from the action (4.50) by varying with respect to ψ and $\bar{\psi}$ independently. Varying with respect to ψ , we have

$$(i\gamma^\mu \partial_\mu - m) \psi = 0 \quad (4.51)$$

This is the *Dirac equation*. It's completely gorgeous. Varying with respect to $\bar{\psi}$ gives the conjugate equation

$$i\partial_\mu \bar{\psi}^\dagger \gamma^\mu + m \bar{\psi}^\dagger = 0 \quad (4.52)$$

The Dirac equation is first order in derivatives, yet miraculously Lorentz invariant. If we tried to write down a first order equation of motion for a scalar field, it would look like $v^\mu \partial_\mu \varphi = \dots$, which necessarily includes a privileged vector in spacetime v^μ and is not Lorentz invariant. However, for spinor fields, the magic of the γ^μ matrices means that the Dirac Lagrangian is Lorentz invariant.

The Dirac equation mixes up different components of ψ through the matrices γ^μ . However, each individual component itself solves the Klein-Gordon equation. To see this, write

$$(i\gamma^\nu \partial_\nu + m)(i\gamma^\mu \partial_\mu - m) \psi = -\gamma^\mu \gamma^\nu \partial_\mu \partial_\nu + m^2 \psi = 0 \quad (4.53)$$

But $\gamma^\mu \gamma^\nu \partial_\mu \partial_\nu = \frac{1}{2}\{\gamma^\mu, \gamma^\nu\} \partial_\mu \partial_\nu = \partial_\mu \partial^\mu$, so we get

$$-(\partial_\mu \partial^\mu + m^2)\psi = 0 \quad (4.54)$$

where this last equation has no γ^μ matrices, and so applies to each component ψ^α , with $\alpha = 1, 2, 3, 4$.

The Slash

Let's introduce some useful notation. We will often come across 4-vectors contracted with γ^μ matrices. We write

$$A_\mu \gamma^\mu \equiv A/ \quad (4.55)$$

so the Dirac equation reads

$$(i \partial/ - m) \psi = 0 \quad (4.56)$$

4.4 Chiral Spinors

When we've needed an explicit form of the γ^μ matrices, we've used the chiral representation

$$\gamma^0 = \begin{matrix} & 1 \\ 0 & \end{matrix}, \quad \gamma^i = \begin{matrix} & \sigma^i \\ 0 & -\sigma^i \end{matrix} \quad (4.57)$$

In this representation, the spinor rotation transformation $S[\Lambda_{\text{rot}}]$ and boost transformation $S[\Lambda_{\text{boost}}]$ were computed in (4.26) and (4.31). Both are block diagonal,

$$S[\Lambda_{\text{rot}}] = \begin{matrix} e^{i\varphi \cdot \rightarrow \sigma / 2} & 0 \\ 0 & e^{i\varphi \cdot \rightarrow \sigma / 2} \end{matrix} \quad \text{and} \quad S[\Lambda_{\text{boost}}] = \begin{matrix} e^{i\chi \cdot \rightarrow \sigma / 2} & 0 \\ 0 & e^{-i\chi \cdot \rightarrow \sigma / 2} \end{matrix} \quad (4.58)$$

This means that the Dirac spinor representation of the Lorentz group is *reducible*. It decomposes into two irreducible representations, acting only on two-component spinors u_\pm which, in the chiral representation, are defined by

$$\psi = \begin{matrix} u^+ \\ u^- \end{matrix} \quad (4.59)$$

The two-component objects u_\pm are called *Weyl spinors* or *chiral spinors*. They transform in the same way under rotations,

$$u_\pm \rightarrow e^{i\varphi \cdot \rightarrow \sigma / 2} u_\pm \quad (4.60)$$

but oppositely under boosts,

$$u_\pm \rightarrow e^{\pm i\chi \cdot \rightarrow \sigma / 2} u_\pm \quad (4.61)$$

In group theory language, u_+ is in the $(\frac{1}{2}, 0)$ representation of the Lorentz group, while u_- is in the $(0, \frac{1}{2})$ representation. The Dirac spinor ψ lies in the $(\frac{1}{2}, 0) \oplus (0, \frac{1}{2})$ representation. (Strictly speaking, the spinor is a representation of the double cover of the Lorentz group $SL(2, \mathbb{C})$).

4.4.1 The Weyl Equation

Let's see what becomes of the Dirac Lagrangian under the decomposition (4.59) into Weyl spinors. We have

$$L = \bar{\psi} (i \partial/\! - m) \psi = i u^+ \sigma^\mu \partial_\mu u_- + i u^- \sigma_-^\mu \partial_\mu u_+ - m(u_+^\dagger u_- + u_-^\dagger u_+) = 0 \quad (4.62)$$

where we have introduced some new notation for the Pauli matrices with a $\mu = 0, 1, 2, 3$ index,

$$\sigma^\mu = (1, \sigma^i) \quad \text{and} \quad \sigma^{-\mu} = (1, -\sigma^i) \quad (4.63)$$

From (4.62), we see that a massive fermion requires both u_+ and u_- , since they couple through the mass term. However, a massless fermion can be described by u_+ (or u_-) alone, with the equation of motion

$$\begin{aligned} i\sigma^{-\mu}\partial_\mu u_+ &= 0 \\ \text{or} \quad i\sigma^\mu\partial_\mu u_- &= 0 \end{aligned} \quad (4.64)$$

These are the *Weyl equations*.

Degrees of Freedom

Let me comment here on the degrees of freedom in a spinor. The Dirac fermion has 4 complex components = 8 real components. How do we count degrees of freedom? In classical mechanics, the number of degrees of freedom of a system is equal to the dimension of the configuration space or, equivalently, half the dimension of the phase space. In field theory we have an infinite number of degrees of freedom, but it makes sense to count the number of degrees of freedom per spatial point: this should at least be finite. For example, in this sense a real scalar field φ has a single degree of freedom. At the quantum level, this translates to the fact that it gives rise to a single type of particle. A classical complex scalar field has two degrees of freedom, corresponding to the particle and the anti-particle in the quantum theory.

But what about a Dirac spinor? One might think that there are 8 degrees of freedom. But this isn't right. Crucially, and in contrast to the scalar field, the equation of motion is first order rather than second order. In particular, for the Dirac Lagrangian, the momentum conjugate to the spinor ψ is given by

$$\pi_\psi = \partial L / \partial \dot{\psi} = i\psi^\dagger \quad (4.65)$$

It is not proportional to the time derivative of ψ . This means that the phase space for a spinor is therefore parameterized by ψ and ψ^\dagger , while for a scalar it is parameterized by φ and $\pi = \dot{\varphi}$. So the *phase space* of the Dirac spinor ψ has 8 real dimensions and correspondingly the number of real degrees of freedom is 4. We will see in the next section that, in the quantum theory, this counting manifests itself as two degrees of freedom (spin up and down) for the particle, and a further two for the anti-particle.

A similar counting for the Weyl fermion tells us that it has two degrees of freedom.

4.4.2 γ^5

The Lorentz group matrices $S[\Lambda]$ came out to be block diagonal in (4.58) because we chose the specific representation (4.57). In fact, this is why the representation (4.57) is called the chiral representation: it's because the decomposition of the Dirac spinor ψ is simply given by (4.59). But what happens if we choose a different representation γ^μ of the Clifford algebra, so that

$$\gamma^\mu \rightarrow U\gamma^\mu U^{-1} \quad \text{and} \quad \psi \rightarrow U\psi \quad ? \quad (4.66)$$

Now $S[\Lambda]$ will not be block diagonal. Is there an invariant way to define chiral spinors? We can do this by introducing the “fifth” gamma-matrix

$$\gamma^5 = -i\gamma^0\gamma^1\gamma^2\gamma^3 \quad (4.67)$$

You can check that this matrix satisfies

$$\{\gamma^5, \gamma^\mu\} = 0 \quad \text{and} \quad (\gamma^5)^2 = +1 \quad (4.68)$$

The reason that this is called γ^5 is because the set of matrices $\tilde{\gamma}^A = (\gamma^\mu, i\gamma^5)$, with $A = 0, 1, 2, 3, 4$ satisfy the $d = 4 + 1$ Clifford algebra $\{\tilde{\gamma}^A, \tilde{\gamma}^B\} = 2\eta^{AB}$. (You might think that γ^4 would be a better name! But γ^5 is the one everyone chooses - it's a more sensible name in Euclidean space, where $A = 1, 2, 3, 4, 5$). You can also check that $[S_{\mu\nu}, \gamma^5] = 0$, which means that γ^5 is a scalar under rotations and boosts. Since $(\gamma^5)^2 = 1$, this means we may form the Lorentz invariant projection operators

$$P_\pm = \frac{1}{2} (1 \pm \gamma^5) \quad (4.69)$$

such that $P_+^2 = P_+$ and $P_-^2 = P_-$ and $P_+P_- = 0$. One can check that for the chiral representation (4.57),

!

$$\gamma^5 = \begin{matrix} 1 & 0 \\ 0 & -1 \end{matrix} \quad (4.70)$$

from which we see that the operators P_\pm project onto the Weyl spinors u_\pm . However, for an arbitrary representation of the Clifford algebra, we may use γ^5 to define the chiral spinors,

$$\psi_\pm = P_\pm \psi \quad (4.71)$$

which form the irreducible representations of the Lorentz group. ψ_+ is often called a “left-handed” spinor, while ψ_- is “right-handed”. The name comes from the way the spin precesses as a massless fermion moves: we'll see this in Section 4.7.2.

4.4.3 Parity

The spinors ψ_{\pm} are related to each other by *parity*. Let's pause to define this concept. The Lorentz group is defined by $x^{\mu} \rightarrow \Lambda_{\nu}^{\mu} x^{\nu}$ such that

$$\underset{\nu}{\Lambda^{\mu}} \underset{\sigma}{\Lambda^{\rho}} \eta^{\nu\sigma} = \eta^{\mu\rho} \quad (4.72)$$

So far we have only considered transformations Λ which are continuously connected to the identity; these are the ones which have an infinitesimal form. However there are also two discrete symmetries which are part of the Lorentz group. They are

$$\begin{aligned} \text{Time Reversal } T : x^0 &\rightarrow -x^0 ; x^i \rightarrow x^i \\ \text{Parity } P : x^0 &\rightarrow x^0 ; x^i \rightarrow -x^i \end{aligned} \quad (4.73)$$

We won't discuss time reversal too much in this course. (It turns out to be represented by an anti-unitary transformation on states. See, for example the book by Peskin and Schroeder). But parity has an important role to play in the standard model and, in particular, the theory of the weak interaction.

Under parity, the left and right-handed spinors are exchanged. This follows from the transformation of the spinors under the Lorentz group. In the chiral representation, we saw that the rotation (4.60) and boost (4.61) transformations for the Weyl spinors u_{\pm} are

$$\begin{array}{ccc} u_{\pm} & \xrightarrow[\pm]{e^{i\varphi \gamma^0 + \sigma/2}} & u^t \\ & & \end{array} \quad \text{and} \quad \begin{array}{ccc} u & \xrightarrow[\pm]{e^{\pm \gamma^0 \cdot \vec{x}/2}} & u^{\text{ost}} \\ & & \end{array} \quad (4.74)$$

Under parity, rotations don't change sign. But boosts do flip sign. This confirms that parity exchanges right-handed and left-handed spinors, $P : u_{\pm} \rightarrow u_{\mp}$, or in the notation $\psi_{\pm} = \frac{1}{2}(1 \pm \gamma^5)\psi$, we have

$$P : \psi_{\pm}(\rightarrow x, t) \rightarrow \psi_{\mp}(-\rightarrow x, t) \quad (4.75)$$

Using this knowledge of how chiral spinors transform, and the fact that $P^2 = 1$, we see that the action of parity on the Dirac spinor itself can be written as

$$P : \psi(\rightarrow x, t) \rightarrow \gamma^0 \psi(-\rightarrow x, t) \quad (4.76)$$

Notice that if $\psi(\rightarrow x, t)$ satisfies the Dirac equation, then the parity transformed spinor

$\gamma^0 \psi(-\rightarrow x, t)$ also satisfies the Dirac equation, meaning

$$(i\gamma^0 \partial_t + i\gamma^i \partial_i - m)\gamma^0 \psi(-\rightarrow x, t) = \gamma^0(i\gamma^0 \partial_t - i\gamma^i \partial_i - m)\psi(-\rightarrow x, t) = 0 \quad (4.77)$$

where the extra minus sign from passing γ^0 through γ^i is compensated by the derivative acting on $-\rightarrow x$ instead of $+\rightarrow x$.

4.4.4 Chiral Interactions

Let's now look at how our interaction terms change under parity. We can look at each of our spinor bilinears from which we built the action,

$$P : \bar{\psi} \psi (\rightarrow x, t) \rightarrow \bar{\psi} \psi (-\rightarrow x, t) \quad (4.78)$$

which is the transformation of a scalar. For the vector $\bar{\psi} \gamma^\mu \psi$, we can look at the temporal and spatial components separately,

$$\begin{aligned} P : \bar{\psi} \gamma^0 \psi (\rightarrow x, t) &\rightarrow \bar{\psi} \gamma^0 \psi (-\rightarrow x, t) \\ P : \bar{\psi} \gamma^i \psi (\rightarrow x, t) &\rightarrow \bar{\psi} \gamma^0 \gamma^i \gamma^0 \psi (-\rightarrow x, t) = -\bar{\psi} \gamma^i \psi (-\rightarrow x, t) \end{aligned} \quad (4.79)$$

which tells us that $\bar{\psi} \gamma^\mu \psi$ transforms as a vector, with the spatial part changing sign. You can also check that $\bar{\psi} S^{\mu\nu} \psi$ transforms as a suitable tensor.

However, now we've discovered the existence of γ^5 , we can form another Lorentz scalar and another Lorentz vector,

$$\bar{\psi} \gamma^5 \psi \text{ and } \bar{\psi} \gamma^5 \gamma^\mu \psi \quad (4.80)$$

How do these transform under parity? We can check:

$$P : \bar{\psi} \gamma^5 \psi (\rightarrow x, t) \rightarrow \bar{\psi} \gamma^0 \gamma^5 \gamma^0 \psi (-\rightarrow x, t) = -\bar{\psi} \gamma^5 \psi (-\rightarrow x, t) \quad (4.81)$$

$$\begin{aligned} P : \bar{\psi} \gamma_5 \gamma_\mu \psi (\rightarrow x, t) &\rightarrow \bar{\psi} \gamma_0 \gamma_5 \gamma_\mu \gamma_0 \psi (-\rightarrow x, t) = -\bar{\psi} \gamma^5 \gamma^0 \psi (-\rightarrow x, t) \quad \mu = 0 \\ &= +\bar{\psi} \gamma^5 \gamma^i \psi (-\rightarrow x, t) \quad \mu = i \end{aligned}$$

which means that $\bar{\psi} \gamma^5 \psi$ transforms as a *pseudoscalar*, while $\bar{\psi} \gamma^5 \gamma^\mu \psi$ transforms as an *axial vector*. To summarize, we have the following spinor bilinears,

$$\begin{aligned} \bar{\psi} \psi &: \text{scalar} \\ \bar{\psi} \gamma^\mu \psi &: \text{vector} \\ \bar{\psi} S^{\mu\nu} \psi &: \text{tensor} \\ \bar{\psi} \gamma^5 \psi &: \text{pseudoscalar} \\ \bar{\psi} \gamma^5 \gamma^\mu \psi &: \text{axial vector} \end{aligned} \quad (4.82)$$

The total number of bilinears is $1 + 4 + (4 \times 3/2) + 4 + 1 = 16$ which is all we could hope for from a 4-component object.

We're now armed with new terms involving γ^5 that we can start to add to our Lagrangian to construct new theories. Typically such terms will break parity invariance of the theory, although this is not always true. (For example, the term $\varphi\psi^\dagger\gamma^5\psi$ doesn't break parity if φ is itself a pseudoscalar). Nature makes use of these parity violating interactions by using γ^5 in the weak force. A theory which treats ψ_\pm on an equal footing is called a *vector-like theory*. A theory in which ψ_+ and ψ_- appear differently is called a *chiral theory*.

4.5 Majorana Fermions

Our spinor ψ^α is a complex object. It has to be because the representation $S[\Lambda]$ is typically also complex. This means that if we were to try to make ψ real, for example by imposing $\psi = \psi^\dagger$, then it wouldn't stay that way once we make a Lorentz transformation. However, there is a way to impose a reality condition on the Dirac spinor ψ . To motivate this possibility, it's simplest to look at a novel basis for the Clifford algebra, known as the *Majorana basis*.

$$\gamma^0 = \begin{matrix} & 0 & \sigma^2 \\ 0 & \sigma^2 & 0 \end{matrix}, \quad \gamma^1 = \begin{matrix} & i\sigma^3 & 0 \\ 0 & 0 & i\sigma^3 \end{matrix}, \quad \gamma^2 = \begin{matrix} & 0 & -\sigma^2 \\ \sigma^2 & 0 & 0 \end{matrix}, \quad \gamma^3 = \begin{matrix} & -i\sigma^1 & 0 \\ 0 & 0 & -i\sigma^1 \end{matrix}$$

These matrices satisfy the Clifford algebra. What is special about them is that they are all pure imaginary $(\gamma^\mu)^\dagger = -\gamma^\mu$. This means that the generators of the Lorentz group $S^{\mu\nu} = \frac{1}{2} [\gamma^\mu, \gamma^\nu]$, and hence the matrices $S[\Lambda]$ are real. So with this basis of the Clifford algebra, we can work with a real spinor simply by imposing the condition,

$$\psi = \psi^\dagger \tag{4.83}$$

which is preserved under Lorentz transformation. Such spinors are called *Majorana spinors*.

So what's the story if we use a general basis for the Clifford algebra? We'll ask only that the basis satisfies $(\gamma^0)^\dagger = \gamma^0$ and $(\gamma^i)^\dagger = -\gamma^i$. We then define the *charge conjugate* of a Dirac spinor ψ as

$$\psi^{(c)} = C\psi^\dagger \tag{4.84}$$

Here C is a 4×4 matrix satisfying

$$C^\dagger C = 1 \quad \text{and} \quad C^\dagger \gamma^\mu C = -(\gamma^\mu)^\dagger \tag{4.85}$$

Let's firstly check that (4.84) is a good definition, meaning that $\psi^{(c)}$ transforms nicely under a Lorentz transformation. We have

$$\psi^{(c)} \rightarrow CS[\Lambda]^\dagger \psi^\dagger = S[\Lambda]C\psi^\dagger = S[\Lambda]\psi^{(c)} \tag{4.86}$$

where we've made use of the properties (4.85) in taking the matrix C through $S[\Lambda]^\wedge$. In fact, not only does $\psi^{(c)}$ transform nicely under the Lorentz group, but if ψ satisfies the Dirac equation, then $\psi^{(c)}$ does too. This follows from,

$$\begin{aligned}(i\partial/\! - m)\psi = 0 &\Rightarrow (-i\partial/\! - m)\psi^\wedge = 0 \\ &\Rightarrow C(-i\partial/\! - m)\psi^\wedge = (+i\partial/\! - m)\psi^{(c)} = 0\end{aligned}$$

Finally, we can now impose the Lorentz invariant reality condition on the Dirac spinor, to yield a Majorana spinor,

$$\psi^{(c)} = \psi \tag{4.87}$$

After quantization, the Majorana spinor gives rise to a fermion that is its own anti-particle. This is exactly the same as in the case of scalar fields, where we've seen that a real scalar field gives rise to a spin 0 boson that is its own anti-particle. (Be aware: In many texts an extra factor of γ^0 is absorbed into the definition of C).

So what is this matrix C ? Well, for a given representation of the Clifford algebra, it is something that we can find fairly easily. In the Majorana basis, where the gamma matrices are pure imaginary, we have simply $C_{\text{Maj}} = 1$ and the Majorana condition $\psi = \psi^{(c)}$ becomes $\psi = \psi^\wedge$. In the chiral basis (4.16), only γ^2 is imaginary, and we may take $C_{\text{chiral}} = i\gamma^2 = \begin{pmatrix} 0 & i\sigma^2 \\ -i\sigma^2 & 0 \end{pmatrix}$. (The matrix $i\sigma^2$ that appears here is simply the anti-symmetric matrix $\epsilon^{\alpha\beta}$). It is interesting to see how the Majorana condition (4.87) looks in terms of the decomposition into left and right handed Weyl spinors (4.59). Plugging in the various definitions, we find that $u_+ = i\sigma^2 u^\wedge$ and $u_- = -i\sigma^2 u^\wedge$. In other words, a Majorana spinor can be written in terms of Weyl spinors as

$$\psi = \begin{matrix} u_+ \\ -i\sigma^2 u_- \end{matrix}^\wedge \tag{4.88}$$

Notice that it's not possible to impose the Majorana condition $\psi = \psi^{(c)}$ at the same time as the Weyl condition ($u_- = 0$ or $u_+ = 0$). Instead the Majorana condition relates u_- and u_+ .

An Aside: Spinors in Different Dimensions: The ability to impose Majorana or Weyl conditions on Dirac spinors depends on both the dimension and the signature of spacetime. One can always impose the Weyl condition on a spinor in even dimensional Minkowski space, basically because you can always build a suitable “ γ^5 ” projection matrix by multiplying together all the other γ -matrices. The pattern for when the Majorana condition can be imposed is a little more sporadic. Interestingly, although the Majorana condition and Weyl condition cannot be imposed simultaneously in four dimensions, you can do this in Minkowski spacetimes of dimension 2, 10, 18,

4.6 Symmetries and Conserved Currents

The Dirac Lagrangian enjoys a number of symmetries. Here we list them and compute the associated conserved currents.

Spacetime Translations

Under spacetime translations the spinor transforms as

$$\delta\psi = \epsilon^\mu \partial_\mu \psi \quad (4.89)$$

The Lagrangian depends on $\partial_\mu \psi$, but not $\partial_\mu \bar{\psi}$, so the standard formula (1.41) gives us the energy-momentum tensor

$$T^{\mu\nu} = i\bar{\psi} \gamma^\mu \partial^\nu \psi - \eta^{\mu\nu} L \quad (4.90)$$

Since a current is conserved only when the equations of motion are obeyed, we don't lose anything by imposing the equations of motion already on $T^{\mu\nu}$. In the case of a scalar field this didn't really buy us anything because the equations of motion are second order in derivatives, while the energy-momentum is typically first order. However, for a spinor field the equations of motion are first order: $(i\partial/\partial t - m)\psi = 0$. This means we can set $L = 0$ in $T^{\mu\nu}$, leaving

$$T^{\mu\nu} = i\bar{\psi} \gamma^\mu \partial^\nu \psi \quad (4.91)$$

In particular, we have the total energy

$$E = \int d^3x T^{00} = \int d^3x i\bar{\psi} \gamma^0 \psi = \int d^3x \psi^\dagger \gamma^0 (-i\gamma^i \partial_i + m) \psi \quad (4.92)$$

where, in the last equality, we have again used the equations of motion.

Lorentz Transformations

Under an infinitesimal Lorentz transformation, the Dirac spinor transforms as (4.22) which, in infinitesimal form, reads

$$\delta\psi^\alpha = -\omega_v^\mu x^\nu \partial_\mu \psi^\alpha + \frac{1}{2} \Omega_{\rho\sigma} (S^{\rho\sigma})^\alpha_\beta \psi^\beta \quad (4.93)$$

where, following (4.10), we have $\omega^\mu_v = \frac{1}{2} \Omega_{\rho\sigma} (M^{\rho\sigma})^\mu_v$, and $M^{\rho\sigma}$ are the generators of the Lorentz algebra given by (4.8)

$$(M^{\rho\sigma})^\mu_v = \eta^{\rho\mu} \delta^\sigma_v - \eta^{\sigma\mu} \delta^\rho_v \quad (4.94)$$

which, after direct substitution, tells us that $\omega^{\mu\nu} = \Omega^{\mu\nu}$. So we get

$$\delta\psi^\alpha = -\omega^{\mu\nu} x_\nu \partial_\mu \psi^\alpha - \frac{1}{2}(S_{\mu\nu})^\alpha_\beta \psi^\beta \quad (4.95)$$

The conserved current arising from Lorentz transformations now follows from the same calculation we saw for the scalar field (1.54) with two differences: firstly, as we saw above, the spinor equations of motion set $L = 0$; secondly, we pick up an extra piece in the current from the second term in (4.95). We have

$$(J^\mu)^{\rho\sigma} = x^\rho T^{\mu\sigma} - x^\sigma T^{\mu\rho} + i\bar{\psi} \gamma^\mu S^{\rho\sigma} \psi \quad (4.96)$$

After quantization, when $(J^\mu)^{\rho\sigma}$ is turned into an operator, this extra term will be responsible for providing the single particle states with internal angular momentum, telling us that the quantization of a Dirac spinor gives rise to a particle carrying spin 1/2.

Internal Vector Symmetry

The Dirac Lagrangian is invariant under rotating the phase of the spinor, $\psi \rightarrow e^{-i\alpha}\psi$. This gives rise to the current

$$j_V^\mu = \bar{\psi} \gamma^\mu \psi \quad (4.97)$$

where “V” stands for *vector*, reflecting the fact that the left and right-handed components ψ_\pm transform in the same way under this symmetry. We can easily check that j_V^μ is conserved under the equations of motion,

$$\partial_\mu j_V^\mu = (\partial_\mu \bar{\psi}) \gamma^\mu \psi + \bar{\psi} \gamma^\mu (\partial_\mu \psi) = im \bar{\psi} \gamma^\mu \psi - im \bar{\psi} \gamma^\mu \psi = 0 \quad (4.98)$$

where, in the last equality, we have used the equations of motion $i\partial/\psi = m\psi$ and $i\partial_\mu \bar{\psi} \gamma^\mu = -m \bar{\psi}$. The conserved quantity arising from this symmetry is

$$Q = \int d^3x \bar{\psi} \gamma^0 \psi = \int d^3x \bar{\psi}^\dagger \psi \quad (4.99)$$

We will see shortly that this has the interpretation of electric charge, or particle number, for fermions.

Axial Symmetry

When $m = 0$, the Dirac Lagrangian admits an extra internal symmetry which rotates left and right-handed fermions in opposite directions,

$$\psi \rightarrow e^{i\alpha\gamma^5} \psi \quad \text{and} \quad \bar{\psi} \rightarrow \bar{\psi} e^{i\alpha\gamma^5} \quad (4.100)$$

Here the second transformation follows from the first after noting that $e^{-i\alpha\gamma^5}\gamma^0 = \gamma^0 e^{+i\alpha\gamma^5}$. This gives the conserved current,

$$j_A^\mu = \bar{\psi} \gamma^\mu \gamma^5 \psi \quad (4.101)$$

where A is for “axial” since j_A^μ is an axial vector. This is conserved only when $m = 0$. Indeed, with the full Dirac Lagrangian we may compute

$$\partial_\mu j_A^\mu = (\partial_\mu \bar{\psi}) \gamma^\mu \gamma^5 \psi + \bar{\psi} \gamma^\mu \gamma^5 \partial_\mu \psi = 2im \bar{\psi} \gamma^5 \psi \quad (4.102)$$

which vanishes only for $m = 0$. However, in the quantum theory things become more interesting for the axial current. When the theory is coupled to gauge fields (in a manner we will discuss in Section 6), the axial transformation remains a symmetry of the classical Lagrangian. But it doesn’t survive the quantization process. It is the archetypal example of an *anomaly*: a symmetry of the classical theory that is not preserved in the quantum theory.

4.7 Plane Wave Solutions

Let’s now study the solutions to the Dirac equation

$$(i\gamma^\mu \partial_\mu - m)\psi = 0 \quad (4.103)$$

We start by making a simple ansatz:

$$\psi = u(p) e^{-ip \cdot x} \quad (4.104)$$

where $u(p)$ is a four-component spinor, independent of spacetime x which, as the notation suggests, can depend on the 3-momentum p . The Dirac equation then becomes

$$\begin{aligned} (\gamma^\mu p_\mu - m)u(p) &= -m \frac{p_\mu \sigma^\mu}{p_\mu \sigma^- \mu - m} u(p) \\ &= 0 \end{aligned} \quad (4.105)$$

where we’re again using the definition,

$$\sigma^\mu = (1, \sigma^i) \quad \text{and} \quad \sigma^- \mu = (1, -\sigma^i) \quad (4.106)$$

Claim: The solution to (4.105) is

$$u(p) = \frac{\sqrt{p \cdot \sigma \xi}}{\sqrt{p \cdot \sigma^- \xi}} \quad (4.107)$$

for any 2-component spinor ξ which we will normalize to $\xi^\dagger \xi = 1$.

Proof: Let's write $u(p\rightarrow)^T = (u_1, u_2)$. Then equation (4.105) reads

$$(p \cdot \sigma) u_2 = m u_1 \quad \text{and} \quad (p \cdot \sigma^-) u_1 = m u_2 \quad (4.108)$$

Either one of these equations implies the other, a fact which follows from the identity $(p \cdot \sigma)(p \cdot \sigma^-) = p^2 - p_i p_j \sigma^i \sigma^j = p_0^2 - p_i p_j \delta^{ij} = p_\mu p^\mu = m^2$. To start with, let's try the ansatz $u_1 = (p \cdot \sigma) \xi^r$ for some spinor ξ^r . Then the second equation in (4.108) immediately tells us that $u_2 = m \xi^r$. So we learn that any spinor of the form

$$u(p\rightarrow) = A \frac{(p \cdot \sigma) \xi^r}{m \xi^r} \quad ! \quad (4.109)$$

with constant A is a solution to (4.105). To make this more symmetric, we choose $A = 1/m$ and $\xi^r = \sqrt{p \cdot \sigma^-} \xi$ with constant ξ . Then $u_1 = (p \cdot \sigma) \sqrt{p \cdot \sigma^-} \xi = m \sqrt{p \cdot \sigma} \xi$. So we get the promised result (4.107)

Negative Frequency Solutions

We get further solutions to the Dirac equation from the ansatz

$$\psi = v(p\rightarrow) e^{i p \cdot x} \quad (4.110)$$

Solutions of the form (4.104), which oscillate in time as $\psi \sim e^{-iEt}$, are called positive frequency solutions. If we compute the energy of these solutions using (4.92), we find that it is positive. Those of the form (4.110), which oscillate as $\psi \sim e^{+iEt}$, are negative frequency solutions. Now if we compute the energy using (4.92), it is negative.

The Dirac equation requires that the 4-component spinor $v(p\rightarrow)$ satisfies

$$(\gamma^\mu p_\mu + m)v(p\rightarrow) = \frac{m}{p_\mu \sigma^-} \frac{p_\mu \sigma^\mu}{m} v(p\rightarrow) = 0 \quad ! \quad (4.111)$$

which is solved by

$$v(p\rightarrow) = \frac{\sqrt{p \cdot \sigma^-}}{-p \cdot \sigma^-} \eta \quad ! \quad (4.112)$$

for some 2-component spinor η which we take to be constant and normalized to $\eta^\dagger \eta = 1$.

4.7.1 Some Examples

Consider the positive frequency solution with mass m and 3-momentum $\vec{p} = 0$,

$$u(\vec{p}) = \frac{\sqrt{m}}{\xi} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \quad (4.113)$$

where ξ is any 2-component spinor. Spatial rotations of the field act on ξ by (4.26),

$$\xi \rightarrow e^{i\vec{\varphi} \cdot \vec{\sigma}/2} \xi \quad (4.114)$$

The 2-component spinor ξ defines the *spin* of the field. This should be familiar from quantum mechanics. A field with spin up (down) along a given direction is described by the eigenvector of the corresponding Pauli matrix with eigenvalue +1 (-1 respectively). For example, $\xi^T = (1, 0)$ describes a field with spin up along the z-axis. After quantization, this will become the spin of the associated particle. In the rest of this section, we'll indulge in an abuse of terminology and refer to the classical solutions to the Dirac equations as "particles", even though they have no such interpretation before quantization.

Consider now boosting the particle with spin $\xi^T = (1, 0)$ along the x^3 direction, with $p^\mu = (E, 0, 0, p^3)$. The solution to the Dirac equation becomes

$$u(\vec{p}) = \frac{\sqrt{p \cdot \sigma}}{\sqrt{\sigma^2 - p^2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \frac{\sqrt{E - p^3}}{E + p^3} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (4.115)$$

In fact, this expression also makes sense for a massless field, for which $E = p^3$. (We picked the normalization (4.107) for the solutions so that this would be the case). For a massless particle we have

$$u(\vec{p}) = \frac{\sqrt{-\sigma^2}}{2E} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad (4.116)$$

Similarly, for a boosted solution of the spin down $\xi^T = (0, 1)$ field, we have

$$u(\vec{p}) = \frac{\sqrt{p \cdot \sigma}}{\sqrt{p^2 - \sigma^2}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \frac{\sqrt{E + p^3}}{\sqrt{E - p^3}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \xrightarrow{m \rightarrow 0} \frac{\sqrt{-\sigma^2}}{2E} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad (4.117)$$

4.7.2 Helicity

The helicity operator is the projection of the angular momentum along the direction of momentum,

$$h = \frac{^i \epsilon}{ij k} \frac{p^{\hat{i}} S^{jk}}{^1 p^{\hat{i}}} = \frac{!}{\frac{1}{2}} \frac{\sigma^i \cdot \mathbf{0}}{0 \sigma^i} \quad (4.118)$$

where S^{ij} is the rotation generator given in (4.25). The massless field with spin $\xi^T = (1, 0)$ in (4.116) has helicity $h = 1/2$: we say that it is *right-handed*. Meanwhile, the field (4.117) has helicity $h = -1/2$: it is *left-handed*.

4.7.3 Some Useful Formulae: Inner and Outer Products

There are a number of identities that will be very useful in the following section, regarding the inner (and outer) products of the spinors $u(p\rightarrow)$ and $v(p\rightarrow)$. It's firstly convenient to introduce a basis ξ^s and η^s , $s = 1, 2$ for the two-component spinors such that

$$\xi^r \xi^s = \delta^{rs} \quad \text{and} \quad \eta^r \eta^s = \delta^{rs} \quad (4.119)$$

for example,

$$\xi^1 = \begin{matrix} 1 & ! \\ 0 & \end{matrix} \quad \text{and} \quad \xi^2 = \begin{matrix} 0 & ! \\ 1 & \end{matrix} \quad (4.120)$$

and similarly for η^s . Let's deal first with the positive frequency plane waves. The two independent solutions are now written as

$$u^s(p\rightarrow) = \sqrt{\frac{p \cdot \sigma}{p \cdot \sigma^-}} \xi^s \quad (4.121)$$

We can take the inner product of four-component spinors in two different ways: either as $u^\dagger \cdot u$, or as $\bar{u} \cdot u$. Of course, only the latter will be Lorentz invariant, but it turns out

that the former is needed when we come to quantize the theory. Here we state both:

$$\begin{aligned} u^r \dagger (p\rightarrow) \cdot u^s (p\rightarrow) &= \xi^r \dagger \sqrt{\frac{p \cdot \sigma}{p \cdot \sigma^-}} \xi^s \sqrt{\frac{p \cdot \sigma}{p \cdot \sigma^-}} \xi^s \dagger \\ &= \xi^r \dagger p \cdot \sigma \xi^s + \xi^r \dagger p \cdot \sigma^- \xi^s = 2 \xi^r \dagger p_0 \xi^s = 2 p_0 \delta^{rs} \end{aligned} \quad (4.122)$$

while the Lorentz invariant inner product is

$$\begin{aligned} u^{-r} (p\rightarrow) \cdot u^s (p\rightarrow) &= \xi^r \dagger \sqrt{\frac{p \cdot \sigma}{p \cdot \sigma^-}} \frac{0 \ 1}{\xi^s} \sqrt{\frac{p \cdot \sigma}{p \cdot \sigma^-}} \xi^s \dagger \\ &= 2 m \delta^{rs} \end{aligned} \quad (4.123)$$

We have analogous results for the negative frequency solutions, which we may write as

$$v^s(p\rightarrow) = \frac{\sqrt{\frac{p \cdot \sigma}{\eta}} \eta^s}{-\vec{p} \cdot \vec{\sigma} \eta^s} \quad ! \quad \text{with} \quad v^{r\dagger}(p\rightarrow) \cdot v^s(p\rightarrow) = 2p_0\delta^{rs} \text{ and } v^{-r}(p\rightarrow) \cdot v^s(p\rightarrow) = -2m\delta^{rs} \quad (4.124)$$

We can also compute the inner product between u and v . We have

$$u^{-r}(p\rightarrow) \cdot v^s(p\rightarrow) = \xi^{r\dagger} p \cdot \sigma, \xi^{r\dagger} p \cdot \sigma^- \frac{\sqrt{\frac{p \cdot \sigma}{\eta}} \eta^s}{-\vec{p} \cdot \vec{\sigma} \eta^s} \\ = \xi^{r\dagger} (p \cdot \sigma^-)(p \cdot \sigma) \eta^s - \xi^{r\dagger} (p \cdot \sigma^-)(p \cdot \sigma) \eta^s = 0 \quad (4.125)$$

and similarly, $v^{-r}(p\rightarrow) \cdot u^s(p\rightarrow) = 0$. However, when we come to $u^r \cdot v$, it is a slightly different combination that has nice properties (and this same combination appears when we quantize the theory). We look at $u^{r\dagger}(p\rightarrow) \cdot v^s(-p\rightarrow)$, with the 3-momentum in the spinor v taking the opposite sign. Defining the 4-momentum $(p^r)^\mu = (p^0, -\vec{p})$, we

have

$$u^{r\dagger}(p\rightarrow) \cdot v^s(-p\rightarrow) = \xi^{r\dagger} \frac{\sqrt{\frac{p \cdot \sigma}{\eta}} \eta^s}{\vec{p} \cdot \vec{\sigma}^-} - \xi^{r\dagger} \frac{\sqrt{\frac{p^r \cdot \sigma^-}{\eta}} \eta^s}{\sqrt{\frac{p^r \cdot \sigma^-}{\eta}}} \\ = \xi^{r\dagger} (p \cdot \sigma)(p^r \cdot \sigma) \eta^s - \xi^{r\dagger} (p \cdot \sigma^-)(p^r \cdot \sigma^-) \eta^s \quad (4.126)$$

Now the terms under the square-root are given by $(p \cdot \sigma)(p^r \cdot \sigma) = (p_0 + p_i \sigma^i)(p_0 - p_i \sigma^i) = p^2 - p^2 = m^2$. The same expression holds for $(p \cdot \sigma^-)(p^r \cdot \sigma^-)$, and the two terms cancel. We learn

$$u^{r\dagger}(p\rightarrow) \cdot v^s(-p\rightarrow) = v^{r\dagger}(p\rightarrow) \cdot u^s(-p\rightarrow) = 0 \quad (4.127)$$

Outer Products

There's one last spinor identity that we need before we turn to the quantum theory. It is:

Claim

:

$$\sum_{s=1}^2 u^s(p\rightarrow) u^{-s}(p\rightarrow) = p/+ m \quad (4.128)$$

$s=1$

where the two spinors are not now contracted, but instead placed back to back to give a 4×4 matrix. Also,

$$\sum_{s=1}^2 v^s(p\rightarrow) v^{-s}(p\rightarrow) = p/- m \quad (4.129)$$

$s=1$

Proof:

$$\begin{aligned}
 & \sum^2 \sqrt{p \cdot \sigma} \xi^s ! \\
 &= \sum_{s=1}^2 \frac{u^s(p)}{u^{-s}(p)} \sigma^s, \\
 &= \sum_{s=1}^2 \frac{\sqrt{\frac{p \cdot \sigma}{p}}}{\xi^s} \sigma^s
 \end{aligned} \tag{4.130}$$

$B_s \xi^s \xi^{s\dagger} = \mathbf{1}$, the 2×2 unit matrix, which then gives us

t

$$! \quad \Sigma \tag{4.131}$$

$$\begin{matrix} u \\ (\\ \rightarrow \\ u \\ s \\ p \\) \end{matrix}$$

m

$$\begin{matrix} p \\ \sigma \end{matrix}$$

$$\sum_{s=1}^2 p \cdot \sigma^s m$$

which is the desired result. A similar proof works for $v^s(p) v^{-s}(p)$.

5. Quantizing the Dirac Field

We would now like to quantize the Dirac Lagrangian,

$$L = \bar{\psi}(x) i \partial/\!\!— m \psi(x) \quad (5.1)$$

We will proceed naively and treat ψ as we did the scalar field. But we'll see that things go wrong and we will have to reconsider how to quantize this theory.

5.1 A Glimpse at the Spin-Statistics Theorem

We start in the usual way and define the momentum,

$$\pi = \frac{\partial L}{\partial \dot{\psi}} = i \bar{\psi} \gamma^0 = i \psi^\dagger \quad (5.2)$$

For the Dirac Lagrangian, the momentum conjugate to ψ is $i\psi^\dagger$. It does not involve the time derivative of ψ . This is as it should be for an equation of motion that is first order in time, rather than second order. This is because we need only specify ψ and ψ^\dagger on an initial time slice to determine the full evolution.

To quantize the theory, we promote the field ψ and its momentum ψ^\dagger to operators, satisfying the canonical commutation relations, which read

$$[\psi_\alpha(\rightarrow x), \psi_\beta(\rightarrow y)] = [\psi_\alpha^\dagger(\rightarrow x), \psi_\beta^\dagger(\rightarrow y)] = 0$$

$$[\psi_\alpha(\rightarrow x), \psi_\beta^\dagger(\rightarrow y)] = \delta_{\alpha\beta} \delta^{(3)}(\rightarrow x - \rightarrow y) \quad (5.3)$$

It's this step that we'll soon have to reconsider.

Since we're dealing with a free theory, where any classical solution is a sum of plane waves, we may write the quantum operators as

$$\psi(\rightarrow x) = \sum_{s=1}^2 \frac{d^3 p}{(2\pi)^3} \frac{1}{2E_p} \frac{\hbar}{\sqrt{2}} b_{p\rightarrow} u(p\rightarrow) e^i + c_{p\rightarrow} v(p\rightarrow) e^{-i}$$

$$+ \sum_{s=1}^2 \frac{d^3 p}{(2\pi)^3} \frac{1}{2E_p} \frac{\hbar}{\sqrt{2}} b_{p\rightarrow}^\dagger u(p\rightarrow) e^{-i} + c_{p\rightarrow}^\dagger v(p\rightarrow) e^i$$

$$\psi(\rightarrow x) = \sum_{s=1}^2 \frac{(2\pi)^3}{2E_p} b_{p\rightarrow} u(p\rightarrow) e^i + c_{p\rightarrow} v(p\rightarrow) e^{-i} \quad (5.4)$$

$$(2\pi)^3$$

where the operators $b_{p\rightarrow}^\dagger$ create particles associated to the spinors $u^s(p\rightarrow)$, while $c_{p\rightarrow}^\dagger$ create particles associated to $v^s(p\rightarrow)$. As with the scalars, the commutation relations of the fields imply commutation relations for the annihilation and creation operators.

Claim: The field commutation relations (5.3) are equivalent to

$$[b_{\vec{p}}^r b_{\vec{q}}^s] = (2\pi)^3 \delta^{rs} \delta^{(3)}(\vec{p} - \vec{q})$$

$$[c_{\vec{p}}^r c_{\vec{q}}^s] = -(2\pi)^3 \delta^{rs} \delta^{(3)}(\vec{p} - \vec{q}) \quad (5.5)$$

with all other commutators vanishing.

Note the strange minus sign in the $[c, c^\dagger]$ term. This means that we can't define the ground state $|0\rangle$ as something annihilated by $c_{\vec{p}}^r |0\rangle = 0$, because then the excited states $c_{\vec{p}}^{s\dagger} |0\rangle$ would have negative norm. To avoid this, we will have to flip the interpretation of c and c^\dagger , with the vacuum defined by $c^s |0\rangle = 0$ and the excited states by $c^r |0\rangle$.

This, as we will see, will be our undoing.

Proof: Let's show that the $[b, b^\dagger]$ and $[c, c^\dagger]$ commutators reproduce the field commutators (5.3),

$$\begin{aligned} & \sum_{r,s} \frac{d^3 p}{(2\pi)} \frac{d^3 q}{(2\pi)} \frac{1}{4E_p E_q} [b_{\vec{p}}^r, b_{\vec{q}}^{s\dagger}] u(\vec{p}) u(\vec{q}) e^{-i(\vec{x} \cdot \vec{p} - \vec{y} \cdot \vec{q})} \\ & + [c_{\vec{p}}^{r\dagger} c_{\vec{q}}^s] v^r(\vec{p}) v^s(\vec{q}) e^{-i(\vec{x} \cdot \vec{p} - \vec{y} \cdot \vec{q})} \\ & = \sum_s \frac{d^3 p}{(2\pi)^3} \frac{1}{2E} u^s(\vec{p}) u^{-s}(\vec{p}) \gamma^0 e^{i\vec{p} \cdot (\vec{x} - \vec{y})} + v^s(\vec{p}) v^{-s}(\vec{p}) \gamma^0 e^{-i\vec{p} \cdot (\vec{x} - \vec{y})} \end{aligned} \quad (5.6)$$

At this stage we use the outer product formulae (4.128) and (4.129) which tell us $\sum_s u^s(\vec{p}) u^{-s}(\vec{p}) = p/+$ and $\sum_s v^s(\vec{p}) v^{-s}(\vec{p}) = p/-m$, so that m and

$$\begin{aligned} [\psi(\vec{x}), \psi^\dagger(\vec{y})] &= \frac{d^3 p}{(2\pi)^3 2E} (p/+m) \gamma^0 e^{i\vec{p} \cdot (\vec{x} - \vec{y})} + (p/-m) \gamma^0 e^{-i\vec{p} \cdot (\vec{x} - \vec{y})} \\ &= \frac{d^3 p}{(2\pi)^3 2E} \frac{1}{p_i \gamma^i} (p_0 \gamma^0 + \sum_i m) \gamma^0 + (p_0 \gamma^0 - p_i \gamma^i - m) \gamma^0 e^{i\vec{p} \cdot (\vec{x} - \vec{y})} \end{aligned}$$

where, in the second term, we've changed $p \rightarrow -p$ under the integration sign.

Now, using $p_0 = E_{\vec{p}}$ we have

$$\begin{aligned} [\psi(\vec{x}), \psi^\dagger(\vec{y})] &= \int e^{i\vec{p} \cdot (\vec{x} - \vec{y})} = \delta^{(3)}(\vec{x} - \vec{y}) \\ &\frac{d^3 p}{(2\pi)^3} \end{aligned} \quad (5.7)$$

as promised. Notice that it's a little tricky in the middle there, making sure that the $p_i \gamma^i$ terms cancel. This was the reason we needed the minus sign in the $[c, c^\dagger]$ commutator terms in (5.5).

5.1.1 The Hamiltonian

To proceed, let's construct the Hamiltonian for the theory. Using the momentum $\pi = i\psi^\dagger$, we have

$$H = \pi\psi^\dagger - L = \psi^\dagger (-i\gamma^i \partial_i + m)\psi \quad (5.8)$$

which means that $H = \int d^3x H$ agrees with the conserved energy computed using Noether's theorem (4.92). We now wish to turn the Hamiltonian into an operator. Let's firstly look at

$$(-i\gamma^i \partial_i + m)\psi = \frac{\int d^3p}{(2\pi)^3} \frac{1}{\sqrt{2E}} b_{p\rightarrow}^s (-i\gamma^i p_i + m) u^s(p\rightarrow) e^{+ip\cdot x} + c_{p\rightarrow}^s (\gamma^i p_i + m) v^s(p\rightarrow) e^{-ip\cdot x} \quad \mathbf{i}$$

where, for once we've left the sum over $s = 1, 2$ implicit. There's a small subtlety with the minus signs in deriving this equation that arises from the use of the Minkowski metric in contracting indices, so that $p\cdot x \equiv x^i p_i = -x^i p_i$. Now we use the defining equations for the spinors $u^s(p\rightarrow)$ and $v^s(p\rightarrow)$ given in (4.105) and (4.111), to replace

$$(-i\gamma^i p_i + m) u^s(p\rightarrow) = \gamma^0 p_0 u^s(p\rightarrow) \text{ and } (\gamma^i p_i + m) v^s(p\rightarrow) = -\gamma^0 p_0 v^s(p\rightarrow) \quad (5.9)$$

so we can write

$$(-i\gamma^i \partial_i + m)\psi = \frac{\int d^3p}{(2\pi)^3} \frac{1}{\sqrt{2}\gamma^0} b_{p\rightarrow}^s u(p\rightarrow) e^{+ip\cdot x} - c_{p\rightarrow}^s v(p\rightarrow) e^{-ip\cdot x} \quad \mathbf{i} \quad (5.10)$$

We now use this to write the operator Hamiltonian

$$\begin{aligned} H &= \int d^3x \psi^\dagger \gamma^0 (-i\gamma^i \partial_i + m)\psi \\ &= \frac{\int d^3x d^3p d^3q}{(2\pi)^6} \frac{1}{4E_q} b_{p\rightarrow}^r b_{q\rightarrow}^s u(\rightarrow q) e^{+iq\cdot x} c_{p\rightarrow}^r v(\rightarrow q) e^{-iq\cdot x} \\ &= \frac{\int d^3p}{(2\pi)^3} \frac{1}{2} h_{p\rightarrow}^r b_{p\rightarrow}^s u(\rightarrow p) e^{+ip\cdot x} - c_{p\rightarrow}^s v(\rightarrow p) e^{-ip\cdot x} \\ &= \frac{1}{(2\pi)^3} \left[b_{p\rightarrow}^{r\dagger} b_{p\rightarrow}^s [u(\rightarrow p) \cdot u(\rightarrow p)] - c_{p\rightarrow}^{r\dagger} c_{p\rightarrow}^s [v(\rightarrow p) \cdot v(\rightarrow p)] \right. \\ &\quad \left. - b_{p\rightarrow}^{r\dagger} c_{p\rightarrow}^s [u(\rightarrow p) \cdot v(-\rightarrow p)] + c_{p\rightarrow}^{r\dagger} b_{p\rightarrow}^s [v(\rightarrow p) \cdot u(-\rightarrow p)] \right] \end{aligned} \quad \mathbf{i}$$

where, in the last two terms we have relabelled $p\rightarrow \rightarrow -p\rightarrow$. We now use our inner product formulae (4.122), (4.124) and (4.127) which read

$$u^r(p\rightarrow)^\dagger \cdot u^s(p\rightarrow) = v^r(p\rightarrow)^\dagger \cdot v^s(p\rightarrow) = 2p_0 \delta^{rs} \quad \text{and} \quad u^r(\rightarrow p)^\dagger \cdot v^s(-\rightarrow p) = v^r(\rightarrow p)^\dagger \cdot u^s(-\rightarrow p) = 0$$

giving us

$$\int \frac{1}{(2\pi)^3} F_{p \rightarrow} b^s^\dagger b^s - c^s c^s = \frac{E_{p \rightarrow}}{(2\pi)^3 \delta^{(3)}(0)} b^s^\dagger b^s - \frac{c^s c^s}{(2\pi)^3 \delta^{(3)}(0)} \quad (5.11)$$
$$H = \frac{d^3 p}{(2\pi)^3} \quad (5.12)$$

The
 $\delta^{(3)}$
term
is
famili
ar
and
easily
dealt
with
by
norm
al
order
ing.
The
 $b^\dagger b$
term
is
famili
ar
and
we
can
chec
k
that
 b^\dagger
creat
e
positi
ve
ener
gy
state
s as
expe
cted,

$[H, b]$	positive energy states,	states of lower and lower energy by continually	that we miss d.
$p \rightarrow$			
$p \rightarrow$			
The minus sign in front of the $c^\dagger c$ term should make us nervous. If we think of c^\dagger as creation operators then there's no problem since , using the commutation relation (5.5), we still find that c^\dagger creates	However, as we noted after (5.5), these states have negative norm. To have a sensible Hilbert space, we need to interpret c as the creation operator. But then the Hamiltonian is not bounded below because	producing c particles. As the English would say, it's all gone a bit Pete Tong. (No relation). Since the above calculation was a little tricky, you might think that it's possible to rescue the theory to get the minus signs to work out right. You can play around with different things, but you'll always find this minus sign cropping up somewhere. And, in fact, it's telling us something important	The key piece of physi cs that we miss ed is that spin 1/2 particles are fermions, meaning that they obey Fermi -Dirac statistics with the quantum state pickin g up a minus sign upon the
	$[H, c^s] = -E_{p \rightarrow} c^s$		
	This is a disaster. Taken seriously it would tell us that we could tumble to		

inter spin
chan fields
ge must be
of quantize
any d as
two fermions
parti . Any
cles. attempt
This to do
fact otherwis
is e will
emb lead to
edde an
d inconsist
into ency,
the such as
struc the
ture unboun
of ded
relati Hamilto
vistic nian we
quan saw in
tum (5.12).
field
theo
ry:
the
spin-
statis
tics
theo
rem
says
that
integ
er
spin
fields
must
be
quan
tized
as
boso
ns,
whil
e
half-
integ
er

So how do we go about quantizing a field as a fermion? Recall that when we quantized the scalar field, the resulting particles obeyed bosonic statistics because the creation and annihilation operators satisfied the commutation relations,

$$[a_{p\rightarrow}^{\dagger}, a_{q\rightarrow}^{\dagger}] = 0 \Rightarrow a_{p\rightarrow}^{\dagger} a_{q\rightarrow}^{\dagger} |0\rangle \equiv |p\rightarrow, q\rightarrow\rangle = |\rightarrow q, p\rightarrow\rangle \quad (5.13)$$

To have states obeying fermionic statistics, we need anti-commutation relations, $\{A, B\} \equiv AB + BA$. Rather than (5.3), we will ask that the spinor fields satisfy

$$\begin{aligned} \{\psi_{\alpha}(\rightarrow x), \psi_{\beta}(\rightarrow y)\} &= \{\psi_{\alpha}^{\dagger}(\rightarrow x), \psi_{\beta}^{\dagger}(\rightarrow y)\} = 0 \\ \{\psi_{\alpha}(\rightarrow x), \psi^{\dagger}(\rightarrow y) &= \delta_{\alpha\beta} \delta^{(3)}(\rightarrow x - \rightarrow y) \\ \} &\qquad\qquad\qquad (\Rightarrow) \\ &\qquad\qquad\qquad \beta \end{aligned} \quad (5.14)$$

We still have the expansion (5.4) of ψ and ψ^{\dagger} in terms of b, b^{\dagger}, c and c^{\dagger} . But now the same proof that led us to (5.5) tells us that

$$\begin{aligned} \{b_{p\rightarrow}^r, b_{q\rightarrow}^s\} &= (2\pi) \delta^{(3)}(p\rightarrow - q\rightarrow) \\ \{c_{p\rightarrow}, c_{q\rightarrow}\} &= (2\pi) \delta^{(3)}(p\rightarrow - q\rightarrow) \end{aligned} \quad (5.15)$$

with all other *anti-commutators* vanishing,

$$\{b_{p\rightarrow}^r, b_{q\rightarrow}^s\} = \{c_{p\rightarrow}, c_{q\rightarrow}\} = \{b_{p\rightarrow}, c_{q\rightarrow}\} = \{b_{p\rightarrow}, c_{q\rightarrow}\} = \dots = 0 \quad (5.16)$$

The calculation of the Hamiltonian proceeds as before, all the way through to the penultimate line (5.11). At that stage, we get

$$\begin{aligned} H_b^s &= \int d^3p \frac{h}{(2\pi)^3} E_{p\rightarrow} b_{p\rightarrow}^s - c^s c^{s\dagger} \mathbf{i} \\ \bar{\tau}_b^s &= \int d^3p \frac{h}{(2\pi)^3} E_{p\rightarrow} b_{p\rightarrow}^s + c_{p\rightarrow}^{s\dagger} \bar{c}_{p\rightarrow}^s \frac{(2\pi)^3 \delta^{(3)}}{(0)} \mathbf{i} \end{aligned} \quad (5.17)$$

The anti-commutators have saved us from the indignity of a Hamiltonian unbounded below. Note that when normal ordering the Hamiltonian we now throw away a negative contribution $-(2\pi)^3 \delta^{(3)}(0)$. In principle, this could partially cancel the positive contribution from bosonic fields. Cosmological constant problem anyone?!

5.2.1 Fermi-Dirac Statistics

Just as in the bosonic case, we define the vacuum $|0\rangle$ to satisfy,

$$b_{p\rightarrow}^s |0\rangle = c^s |0\rangle = 0 \quad (5.18)$$

Although b and c obey anti-commutation relations, the Hamiltonian (5.17) has nice commutation relations with them. You can check that

$$\begin{aligned} [H, b^r] &= -E_{p \rightarrow} b^r \quad \text{and} \quad [H, b^{r\dagger}] = E_{p \rightarrow} b^{r\dagger} \\ [H, c^r_{p \rightarrow}] &= -E_{p \rightarrow} c^r_{p \rightarrow} \quad \text{and} \quad [H, c^{r\dagger}_{p \rightarrow}] = E_{p \rightarrow} c^{r\dagger}_{p \rightarrow} \end{aligned} \quad (5.19)$$

This means that we can again construct a tower of energy eigenstates by acting on the vacuum by $b^{r\dagger}$ and $c^{r\dagger}$ to create particles and antiparticles, just as in the bosonic case.

For example, we have the one-particle states

$$|p \rightarrow, r\rangle = b^{r\dagger} |0\rangle \quad (5.20)$$

The two particle states now satisfy

$$|p \rightarrow_1, r_1; p \rightarrow_2, r_2\rangle \equiv b^{r_1\dagger}_{p \rightarrow_1} b^{r_2\dagger}_{p \rightarrow_2} |0\rangle = -|p \rightarrow_2, r_2; p \rightarrow_1, r_1\rangle \quad (5.21)$$

confirming that the particles do indeed obey Fermi-Dirac statistics. In particular, we have the Pauli-Exclusion principle $|p \rightarrow, r; p \rightarrow, r\rangle = 0$. Finally, if we wanted to be sure about the spin of the particle, we could act with the angular momentum operator (4.96) to confirm that a stationary particle $|p \rightarrow = 0, r\rangle$ does indeed carry intrinsic angular momentum 1/2 as expected.

5.3 Dirac's Hole Interpretation

"In this attempt, the success seems to have been on the side of Dirac rather than logic"

Pauli on Dirac

Let's pause our discussion to make a small historical detour. Dirac originally viewed his equation as a relativistic version of the Schrödinger equation, with ψ interpreted as the wavefunction for a single particle with spin. To reinforce this interpretation, he wrote $(i/\partial - m)\psi = 0$ as

$$\frac{\partial \psi}{i \partial t} = -i \alpha \cdot \vec{\nabla} \psi + m\beta \psi \equiv \hat{H} \psi \quad (5.22)$$

where $\alpha \cdot \vec{\nabla} = -\gamma_0 \vec{\gamma} \cdot \vec{\nabla}$ and $\beta = \gamma_0$. Here the operator \hat{H} is interpreted as the one-particle

Hamiltonian. This is a very different viewpoint from the one we now have, where ψ is a classical field that should be quantized. In Dirac's view, the Hamiltonian of the system is \hat{H} defined above, while for us the Hamiltonian is the field operator (5.17).

Let's see where Dirac's viewpoint leads.

With the interpretation of ψ as a single-particle wavefunction, the plane-wave solutions (4.104) and (4.110) to the Dirac equation are thought of as energy eigenstates, with

$$\begin{aligned} \psi &= u(p^\rightarrow) e^{-ip \cdot x} & \Rightarrow i \frac{\partial \psi}{\partial p^\rightarrow} &= E_p \psi \\ \psi &= v(p^\rightarrow) e^{+ip \cdot x} & \Rightarrow i \frac{\partial \psi}{\partial t} &= -E_p \psi \end{aligned} \quad (5.23)$$

which look like positive and negative energy solutions. The spectrum is once again unbounded below; there are states $v(p^\rightarrow)$ with arbitrarily low energy $-E_p$. At first glance this is disastrous, just like the unbounded field theory Hamiltonian (5.12). Dirac postulated an ingenious solution to this problem: since the electrons are fermions (a fact which is put in by hand to Dirac's theory) they obey the Pauli-exclusion principle. So we could simply stipulate that in the true vacuum of the universe, all the negative energy states are filled. Only the positive energy states are accessible. These filled negative energy states are referred to as the *Dirac sea*. Although you might worry about the infinite negative charge of the vacuum, Dirac argued that only charge differences would be observable (a trick reminiscent of the normal ordering prescription we used for field operators).

Having avoided disaster by floating on an infinite sea comprised of occupied negative energy states, Dirac realized that his theory made a shocking prediction. Suppose that a negative energy state is excited to a positive energy state, leaving behind a hole. The hole would have all the properties of the electron, except it would carry positive charge. After flirting with the idea that it may be the proton, Dirac finally concluded that the hole is a new particle: the positron. Moreover, when a positron comes across an electron, the two can annihilate. Dirac had predicted anti-matter, one of the greatest achievements of theoretical physics. It took only a couple of years before the positron was discovered experimentally in 1932.

Although Dirac's physical insight led him to the right answer, we now understand that the interpretation of the Dirac spinor as a single-particle wavefunction is not really correct. For example, Dirac's argument for anti-matter relies crucially on the particles being fermions while, as we have seen already in this course, anti-particles exist for both fermions and bosons. What we really learn from Dirac's analysis is that there is no consistent way to interpret the Dirac equation as describing a single particle. It is instead to be thought of as a classical field which has only positive energy solutions because the Hamiltonian (4.92) is positive definite. Quantization of this field then gives rise to both particle and anti-particle excitations.

This from Julian Schwinger:

"Until now, everyone thought that the Dirac equation referred directly to physical particles. Now, in field theory, we recognize that the equations refer to a sublevel. Experimentally we are concerned with particles, yet the old equations describe fields.... When you begin with field equations, you operate on a level where the particles are not there from the start. It is when you solve the field equations that you see the emergence of particles."

5.4 Propagators

Let's now move to the Heisenberg picture. We define the spinors $\psi(\rightarrow x, t)$ at every point in spacetime such that they satisfy the operator equation

$$\frac{\partial \psi}{\partial t} = i[H, \psi] \quad (5.24)$$

We solve this by the expansion

$$\begin{aligned} \sum_s \int \frac{d^3 p}{h} \frac{1}{s s - ip \cdot x} & \stackrel{\textbf{i}}{=} b_{p \rightarrow} u(p \rightarrow) e + c_{p \rightarrow} v(p \rightarrow) e \\ \psi(x) = \frac{(2\pi)^3}{2E} \sqrt{\frac{1}{h}} & \stackrel{\textbf{i}}{=} b_{p \rightarrow} u(p \rightarrow) e + c_{p \rightarrow} v(p \rightarrow) e \\ \psi^\dagger(x) = \sum_{s=1}^s \int \frac{d^3 p}{2E_p} \frac{1}{s \dagger s + ip \cdot x} & \stackrel{\textbf{i}}{=} b_{p \rightarrow} u(p \rightarrow) e + c_{p \rightarrow} v(p \rightarrow) e \\ & \stackrel{(2\pi)^3}{=} \end{aligned} \quad (5.25)$$

Let's now look at the anti-commutators of these fields. We define the fermionic propagator to be

$$iS_{\alpha\beta} = \{\psi_\alpha(x), \psi_\beta^\dagger(y)\} \quad (5.26)$$

In what follows we will often drop the indices and simply write $iS(x-y) = \{\psi(x), \psi^\dagger(y)\}$, but you should remember that $S(x-y)$ is a 4×4 matrix. Inserting the expansion (5.25), we have

$$\begin{aligned} iS(x-y) &= \int d^3 p d^3 q \frac{1}{\{b_s^\dagger, b_r\}} \frac{h}{s r - i(p \cdot x - q \cdot y)} \\ &\quad \frac{(2\pi)^6}{4E_p E_q} \frac{u(p \rightarrow) u^\dagger(\rightarrow q) e}{\stackrel{\textbf{i}}{=}} \\ &\quad \rightarrow \\ &= \int \frac{d^3 p}{(2\pi)^3 2E} \frac{1}{d^3 p} \frac{u(s p \rightarrow) u^\dagger(s \rightarrow p) e^{-ip \cdot (x-y)}}{(2\pi)^3 2E} \\ &\quad + \{c_{p \rightarrow}^s, c_{q \rightarrow}^r\} v^s(\rightarrow p) v^\dagger(r \rightarrow q) e^{+i(p \cdot x - q \cdot y)} \end{aligned}$$

$$\frac{1}{(p/+m)e^{-ip\cdot(x-)}} + \frac{1}{(p/-m)e^{+ip\cdot(x-)}} \quad (5.27)$$

$$(2\pi)^3 \, 2E$$

where to reach the final line we have used the outer product formulae (4.128) and (4.129). We can then write

$$iS(x - y) = (i \partial_x + m)(D(x - y) - D(y - x)) \quad (5.28)$$

in terms of the propagator for a real scalar field $D(x - y)$ which, recall, can be written as (2.90)

$$D(x - y) = \frac{\int d^3 p}{(2\pi)^3} \frac{1}{2E_p} e^{-ip \cdot (x-y)} \quad (5.29)$$

Some comments:

- For spacelike separated points $(x - y)^2 < 0$, we have already seen that $D(x - y) - D(y - x) = 0$. In the bosonic theory, we made a big deal of this since it ensured that

$$[\varphi(x), \varphi(y)] = 0 \quad (x - y)^2 < 0 \quad (5.30)$$

outside the lightcone, which we trumpeted as proof that our theory was causal. However, for fermions we now have

$$\{\psi_\alpha(x), \psi_\beta(y)\} = 0 \quad (x - y)^2 < 0 \quad (5.31)$$

outside the lightcone. What happened to our precious causality? The best that we can say is that all our observables are bilinear in fermions, for example the Hamiltonian (5.17). These still commute outside the lightcone. The theory remains causal as long as fermionic operators are not observable. If you think this is a little weak, remember that no one has ever seen a physical measuring apparatus come back to minus itself when you rotate by 360 degrees!

- At least away from singularities, the propagator satisfies

$$(i \partial_x - m)S(x - y) = 0 \quad (5.32)$$

which follows from the fact that $(\partial_x^2 + m^2)D(x - y) = 0$ using the mass shell condition $p^2 = m^2$.

5.5 The Feynman Propagator

By a similar calculation to that above, we can determine the vacuum expectation value,

$$\begin{aligned} \langle 0 | \psi_\alpha(x) \bar{\psi}_\beta(y) | 0 \rangle &= \frac{\int d^3 p}{(2\pi)^3} \frac{1}{2E_p} (p/+ m)_{\alpha\beta} e^{-ip \cdot (x-y)} \\ &= \frac{\int d^3 p}{(2\pi)^3} \frac{1}{2E_p} (p/- m)_{\alpha\beta} e^{+ip \cdot (x-y)} \end{aligned} \quad (5.33)$$

$$(2\pi)^3 2E$$

We now define the Feynman propagator $S_F(x - y)$, which is again a 4×4 matrix, as the time ordered product,

$$S_F(x - y) = \langle 0 | T\psi(x)\psi^\dagger(y) | 0 \rangle \equiv \begin{cases} \langle 0 | \psi(x)\psi^\dagger(y) | 0 \rangle & x^0 > y^0 \\ \langle 0 | -\psi^\dagger(y)\psi(x) | 0 \rangle & y^0 > x^0 \end{cases} \quad (5.34)$$

Notice the minus sign! It is necessary for Lorentz invariance. When $(x-y)^2 < 0$, there is no invariant way to determine whether $x^0 > y^0$ or $y^0 > x^0$. In this case the minus sign is necessary to make the two definitions agree since $\{\psi(x), \psi^\dagger(y)\} = 0$ outside the lightcone.

We have the 4-momentum integral representation for the Feynman propagator,

$$S_F(x - y) = i \int \frac{d^4 p}{(2\pi)^4} e^{-ip \cdot (x-y)} \frac{\gamma \cdot p + m}{p^2 - m^2 + i\epsilon} \quad (5.35)$$

which satisfies $(i\partial_x - m)S_F(x - y) = i\delta^{(4)}(x - y)$, so that S_F is a Green's function for the Dirac operator.

The minus sign that we see in (5.34) also occurs for any string of operators inside a time ordered product $T(\dots)$. While bosonic operators commute inside T , fermionic operators anti-commute. We have this same behaviour for normal ordered products as well, with fermionic operators obeying $:\psi_1\psi_2: = - :\psi_2\psi_1::$. With the understanding that all fermionic operators anti-commute inside T and ::, Wick's theorem proceeds just as in the bosonic case. We define the contraction

$$\overline{\psi(x)\psi^\dagger(y)} = T(\psi(x)\psi^\dagger(y)) - :\psi(x)\psi^\dagger(y): = S_F(x - y) \quad (5.36)$$

5.6 Yukawa Theory

The interaction between a Dirac fermion of mass m and a real scalar field of mass μ is governed by the Yukawa theory,

$$L = \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - \frac{1}{2}\mu^2\varphi^2 + \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - \lambda\varphi\bar{\psi}\psi \quad (5.37)$$

which is the proper version of the baby scalar Yukawa theory we looked at in Section 3. Couplings of this type appear in the standard model, between fermions and the Higgs boson. In that context, the fermions can be leptons (such as the electron) or quarks.

Yukawa originally proposed an interaction of this type as an effective theory of nuclear forces. With an eye to this, we will again refer to the φ particles as mesons, and the ψ particles as nucleons. Except, this time, the nucleons have spin. (This is still not a particularly realistic theory of nucleon interactions, not least because we're omitting isospin. Moreover, in Nature the relevant mesons are pions which are pseudoscalars, so a coupling of the form $\varphi\bar{\psi}\gamma^5\psi$ would be more appropriate. We'll turn to this briefly in Section 5.7.3).

Note the dimensions of the various fields. We still have $[\varphi] = 1$, but the kinetic terms require that $[\psi] = 3/2$. Thus, unlike in the case with only scalars, the coupling is dimensionless: $[\lambda] = 0$.

We'll proceed as we did in Section 3, firstly computing the amplitude of a particular scattering process then, with that calculation as a guide, writing down the Feynman rules for the theory. We start with:

5.6.1 An Example: Putting Spin on Nucleon Scattering

Let's study $\psi\psi \rightarrow \psi\psi$ scattering. This is the same calculation we performed in Section (3.3.3) except now the fermions have spin. Our initial and final states are

$$\begin{aligned} |i\rangle &= \sqrt{\frac{4E_p E_q}{4E_{p'} E_{q'}}} b_{sp}^{\dagger} b_{rq}^{\dagger} |0\rangle \equiv |p\rightarrow, s; q\rightarrow, r\rangle \\ |f\rangle &= \sqrt{\frac{4E_p E_q}{4E_{p'} E_{q'}}} b_{p\rightarrow} b_{q\rightarrow} |0\rangle \equiv |p\rightarrow, r; q\rightarrow, r\rangle \end{aligned} \quad (5.38)$$

We need to be a little cautious about minus signs, because the b^{\dagger} 's now anti-commute. In particular, we should be careful when we take the adjoint. We have

$$\langle f | = \sqrt{\frac{4E_p E_q}{4E_{p'} E_{q'}}} \langle 0 | b_{q\rightarrow}^{\dagger} b_{p\rightarrow}^{\dagger} \quad (5.39)$$

We want to calculate the order λ^2 terms from the S-matrix element $\langle f | S - 1 | i \rangle$.

$$\frac{(-i\lambda)^2}{2} \int d^4x_1 d^4x_2 T \left[\bar{\psi}(x_1)\psi(x_1)\varphi(x_1) \right. \left. - \bar{\psi}(x_2)\psi(x_2)\varphi(x_2) \right] \quad (5.40)$$

where, as usual, all fields are in the interaction picture. Just as in the bosonic calculation, the contribution to nucleon scattering comes from the contraction

$$:\bar{\psi}(x_1)\psi(x_1)\bar{\psi}(x_2)\psi(x_2):\varphi(x_1)\varphi(x_2) \quad (5.41)$$

We just have to be careful about how the spinor indices are contracted. Let's start by looking at how the fermionic operators act on $|i\rangle$. We expand out the ψ fields, leaving the $\bar{\psi}$ fields alone for now. We may ignore the c^{\dagger} pieces in ψ since they give no contribution at order λ^2 . We have

$$:\bar{\psi}(x_1)\psi(x_1)\bar{\psi}(x_2)\psi(x_2): \underset{p\rightarrow q\rightarrow}{b^{s\dagger} b^{r\dagger}} |0\rangle = \int \frac{d^3k_1 d^3k_2}{(2\pi)^6} \underset{1}{[\bar{\psi}(x_1) \cdot u^m(k_1)]} \underset{1}{[\bar{\psi}(x_2) \cdot u^n(k_2)]} e^{-ik_1 \cdot x_1 - ik_2 \cdot x_2} \frac{4E_{k_1} E_{k_2}}{b_{k_1}^m b_{k_2}^n b_{p\rightarrow}^{\dagger} b_{q\rightarrow}^{\dagger}} |0\rangle \quad (5.42)$$

where we've used square brackets $[\cdot]$ to show how the spinor indices are contracted. The minus sign that sits out front came from moving $\psi(x_1)$ past $\psi^\dagger(x_2)$. Now anti-commuting the b 's past the b^\dagger 's, we get

$$= \frac{1}{3} \frac{\sqrt{-}}{\overline{E_{p \rightarrow} E_{q \rightarrow}}} [\psi_1^\dagger(x_1) \cdot u^r(\rightarrow q)] [\psi_2^\dagger(x_2) \cdot u^s(p \rightarrow)] e^{-ip \cdot x_2 - iq \cdot x_1} \\ - [\psi_1^\dagger(x_1) \cdot u^s(p \rightarrow)] [\psi_2^\dagger(x_2) \cdot u^r(\rightarrow q)] e^{-ip \cdot x_1 - iq \cdot x_2} |0\rangle \quad (5.43)$$

Note, in particular, the relative minus sign that appears between these two terms. Now let's see what happens when we hit this with $\langle f |$. We look at

$$\langle 0 | b_{q \rightarrow} b_{p \rightarrow} [\psi(x_1) \cdot u(\rightarrow q)] [\psi(x_2) \cdot u(p \rightarrow)] \frac{e^{+ip' \cdot x_1 + iq' \cdot x_2}}{\sqrt{\frac{2}{\overline{[u]}}}} (p \rightarrow) \cdot u(\rightarrow q)] [u^\dagger(p \rightarrow)] \\ - \frac{e^{+ip' \cdot x_2 + iq' \cdot x_1}}{2 \sqrt{\overline{E_{p \rightarrow} E_{q \rightarrow}}}} [u^\dagger(p \rightarrow) \cdot u(\rightarrow q)] [u^\dagger(p \rightarrow)] \cdot u(s \rightarrow)$$

The $[\psi^\dagger(x_1) \cdot u^s(p \rightarrow)] [\psi^\dagger(x_2) \cdot u^r(\rightarrow q)]$ term in (5.43) does up with this, cancelling the factor of $1/2$ in front of (5.40). Meanwhile, the $1/\sqrt{E}$ terms cancel the relativistic state normalization. Putting everything together, we have the following expression for

$$\langle f | S - 1 | i \rangle \\ = \frac{(-i\lambda)^2}{(2\pi)^4} \int \frac{d^4x^1 d^4x^2 d^4k}{k^2 - \mu^2 + i\epsilon} ie^{ik \cdot (x_1 - x_2)} [u^\dagger(s \rightarrow) \cdot s \rightarrow r \rightarrow r \cdot t^r(\rightarrow q)] e^{+ix_1 \cdot (q' - q) - x_2 \cdot (p' - p)} \\ (p \rightarrow) \cdot u(p \rightarrow)] [u^\dagger(r \rightarrow) \cdot r \rightarrow r \cdot t^r(\rightarrow q)] e^{+ix_1 \cdot (p' - q) + ix_2 \cdot (q' - p)} \\ - [u^\dagger(p \rightarrow) \cdot u(p \rightarrow)] [u^\dagger(r \rightarrow) \cdot r \rightarrow r \cdot t^r(\rightarrow q)] e^{+ix_1 \cdot (p' - q) + ix_2 \cdot (q' - p)}$$

where we've put the φ propagator back in. Performing the integrals over x_1 and x_2 , this becomes,

$$\int_k \frac{(2\pi)^4 i (-i\lambda)^2}{k^2 - \mu^2 + i\epsilon} [u^\dagger(s \rightarrow) \cdot u(p \rightarrow)] [u^\dagger(r \rightarrow) \cdot u(r \rightarrow)] \delta^{(4)}(q^r - q + k) \delta^{(4)}(p^r - p - k) \\ (\rightarrow q) \\ - [u^\dagger(s \rightarrow) \cdot u(p \rightarrow)] [u^\dagger(r \rightarrow) \cdot u(r \rightarrow)] \delta^{(4)}(p_r - q + k) \delta^{(4)}(q_r - p - k) \\ - (\quad) - (\quad)$$

And we're almost there! Finally, writing the S-matrix element in terms of the amplitude in the usual way, $\langle f | S - 1 | i \rangle = iA(2\pi)^4 \delta^{(4)}(p + q - p^r - q^r)$, we have

$$A = \frac{[u^\dagger(s \rightarrow) \cdot u(p \rightarrow)] (\rightarrow q) \cdot u^\dagger(r \rightarrow) [u^\dagger(r \rightarrow) \cdot u(p \rightarrow)]}{(p^r - p)^2 - \mu^2 + i\epsilon} - \frac{[u^\dagger(s \rightarrow) \cdot u(p \rightarrow)] (\rightarrow q) \cdot u^\dagger(r \rightarrow) [u^\dagger(r \rightarrow) \cdot u(p \rightarrow)]}{(q^r - p)^2 - \mu^2 + i\epsilon}$$

which is our final answer for the amplitude.

5.7 Feynman Rules for Fermions

It's important to bear in mind that the calculation we just did kind of blows. Thankfully the Feynman rules will once again encapsulate the combinatoric complexities and make life easier for us. The rules to compute amplitudes are the following

- To each incoming fermion with momentum p and spin r , we associate a spinor $u^r(p \rightarrow)$. For outgoing fermions we associate $u^{-r}(p \rightarrow)$.

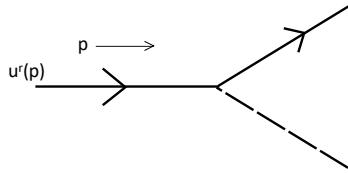


Figure 21: An incoming fermion

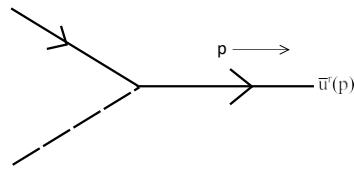


Figure 22: An outgoing fermion

- To each incoming anti-fermion with momentum p and spin r , we associate a spinor $\bar{v}^r(p \rightarrow)$. For outgoing anti-fermions we associate $v^r(p \rightarrow)$.

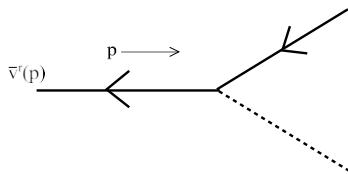


Figure 23: An incoming anti-fermion

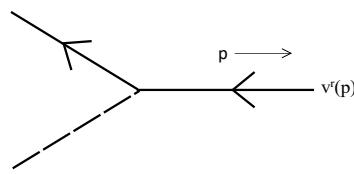


Figure 24: An outgoing anti-fermion

- Each vertex gets a factor of $-i\lambda$.
- Each internal line gets a factor of the relevant propagator.

	$\frac{i}{p^2 - \mu^2 + i\epsilon}$	for scalars
	$\frac{i}{p^2 - m^2 + i\epsilon}$	for fermions

(5.44)

The arrows on the fermion lines must flow consistently through the diagram (this ensures fermion number conservation). Note that the fermionic propagator is a 4×4 matrix. The matrix indices are contracted at each vertex, either with further propagators, or with external spinors u, u^-, v or v^- .

- Impose momentum conservation at each vertex, and integrate over undetermined loop momenta.
- Add extra minus signs for statistics. Some examples will be given below.

5.7.1 Examples

Let's run through the same examples we did for the scalar Yukawa theory. Firstly, we have

Nucleon Scattering

For the example we worked out previously, the two lowest order Feynman diagrams are shown in Figure 25. We've drawn the second Feynman diagram with the legs crossed

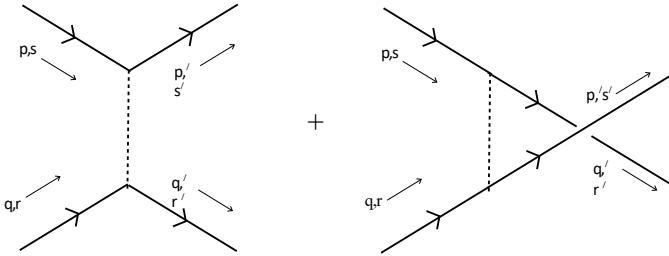


Figure 25: The two Feynman diagrams for nucleon scattering

to emphasize the fact that it picks up a minus sign due to statistics. (Note that the way the legs point in the Feynman diagram doesn't tell us the direction in which the particles leave the scattering event: the momentum label does that. The two diagrams above are different because the incoming legs are attached to different outgoing legs). Using the Feynman rules we can read off the amplitude.

$$A = \frac{i\lambda}{(-i\lambda)^2} \frac{[u^- (p^+ \cdot u^-) (\not{p} \rightarrow q^-) \cdot u^-] + [u^- (p^+ \cdot u^-) (\not{q} \rightarrow q^-) \cdot u^-]}{(p - p^r)^2 - \mu^2} \quad (5.45)$$

The denominators in each term are due to the meson propagator, with the momentum determined by conservation at each vertex. This agrees with the amplitude we computed earlier using Wick's theorem.

Nucleon to Meson Scattering

Let's now look at $\psi \bar{\psi} \rightarrow \varphi \varphi$. The two lowest order Feynman diagrams are shown in Figure 26. Applying the Feynman rules, we have

$$A = (-i\lambda)^2 \frac{v^- (r \rightarrow q^-) [\gamma^\mu (p_\mu - p^r) + m] u^s(p^+) + v^- r (\not{q} \rightarrow q^-) [\gamma^\mu (p_\mu - q^r) + m] u^s(p^+)}{(p - p^r)^2 - m^2}$$

Since the internal line is now a fermion, the propagator contains $\gamma_\mu (p_\mu - p^r_\mu) + m$ factors. This is a 4×4 matrix which sits on the top, sandwiched between the two external spinors. Now the exchange statistics applies to the final meson states. These are bosons and, correspondingly, there is no relative minus sign between the two diagrams.

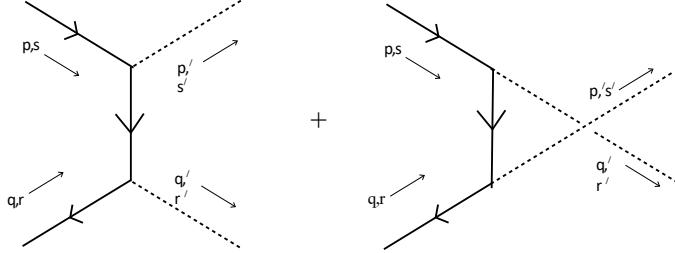


Figure 26: The two Feynman diagrams for nucleon to meson scattering

Nucleon-Anti-Nucleon Scattering

For $\psi \bar{\psi} \rightarrow \psi \bar{\psi}$, the two lowest order Feynman diagrams are of two distinct types, just like in the bosonic case. They are shown in Figure 27.

The corresponding amplitude is given by,

$$A = (-i\lambda) \frac{[u^- (\vec{p} \rightarrow) \cdot u^- (\vec{q} \rightarrow) \cdot v^- (\vec{p} \rightarrow)] [v^- (\vec{q} \rightarrow) \cdot u^- (\vec{p} \rightarrow) \cdot v^- (\vec{q} \rightarrow)]}{(p - p')^2 - \mu^2} + \frac{[v^- (\vec{q} \rightarrow) \cdot u^- (\vec{p} \rightarrow) \cdot v^- (\vec{q} \rightarrow)] [u^- (\vec{p} \rightarrow) \cdot v^- (\vec{q} \rightarrow)]}{(p + q)^2 - \mu^2 + i\epsilon} \quad (5.46)$$

As in the bosonic diagrams, there is again the difference in the momentum dependence in the denominator. But now the difference in the diagrams is also reflected in the spinor contractions in the numerator.

More subtle are the minus signs. The fermionic statistics mean that the first diagram has an extra minus sign relative to the $\psi\psi$ scattering of Figure 25. Since this minus sign will be important when we come to figure out whether the Yukawa force is attractive or repulsive, let's go back to basics and see where it comes from. The initial and final states for this scattering process are

$$|i\rangle = \sqrt{\frac{4E_p E_q}{s'}} b_{p\rightarrow}^{\dagger} c_{q\rightarrow}^{\dagger} |0\rangle \equiv |\vec{p} \rightarrow, s; \vec{q}, r\rangle$$

$$|f\rangle = \sqrt{\frac{4E_p E_q}{s'}} b_{p\rightarrow'}^{\dagger} c_{q\rightarrow'}^{\dagger} |0\rangle \equiv |\vec{p} \rightarrow', s; \vec{q}, r\rangle \quad (5.47)$$

The ordering of b^\dagger and c^\dagger in these states is crucial and reflects the scattering $\psi \bar{\psi} \rightarrow \psi \bar{\psi}$, as opposed to $\psi \bar{\psi} \rightarrow \psi \bar{\psi}$ which would differ by a minus sign. The first diagram in Figure 27 comes from the term in the perturbative expansion,

$$\langle f | : \psi^-(x_1) \psi(x_1) \bar{\psi}^-(x_2) \bar{\psi}(x_2) : b_{p\rightarrow}^{\dagger} c_{q\rightarrow}^{\dagger} |0\rangle \sim \langle f | [v^{-m}(\vec{k}_1) \cdot \psi(x_1)] [\bar{\psi}^-(x_2) \cdot u^n(\vec{k}_2)] c_m^m b_n^{\dagger} |0\rangle$$

1 2

where we've neglected a bunch of objects in this equation like $d^4 k_i$ and exponential factors because we only want to keep track of the minus signs. Moving the annihilation operators past the creation operators, we have

$$+ \langle f | [v^{-r}(\vec{q}) \cdot \psi(x_1)] [\bar{\psi}^-(x_2) \cdot u^s(\vec{p} \rightarrow)] |0\rangle \quad (5.48)$$

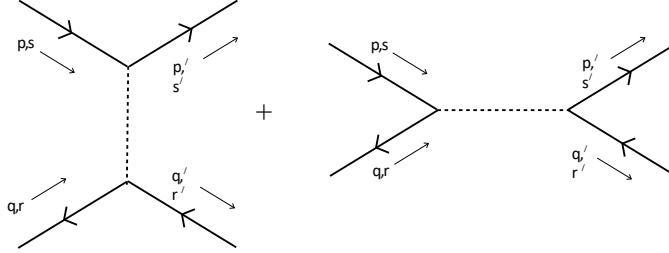


Figure 27: The two Feynman diagrams for nucleon-anti-nucleon scattering

Repeating the process by expanding out the $\psi(x_1)$ and $\bar{\psi}(x_2)$ fields and moving them to the left to annihilate $\langle f |$, we have

$$\langle 0 | \bar{s}' m^\dagger n^\dagger r' (\not{v}^m \not{q}) [\not{u}^{-n} (\not{p}) \cdot \not{u}^s (\not{p})] | 0 \rangle \sim -[\not{v}^{-r} \not{q}] [\not{u}^{-s} (\not{p}) \cdot \not{u}^s (\not{p})]$$

$c_{q \rightarrow b'} \rightarrow \not{v}^{-r} (\not{q})$ 1 2

b_1

where the minus sign has appeared from anti-commuting $c^{m\dagger}$ past $b_{p \rightarrow s'}^s$. This is the overall minus sign found in (5.46). One can also follow similar contractions to compute the second diagram in Figure 27.

Meson Scattering

Finally, we can also compute the scattering of $\varphi\varphi \rightarrow \varphi\varphi$ which, as in the bosonic case, picks up its leading contribution at one-loop. The amplitude for the diagram shown in the figure is

$$iA = -(-i\lambda) \int_4 \frac{k/+ m}{d^4 k} \frac{k/+ p/r + m}{(2\pi)^4} \text{Tr} \frac{(k^2 - m^2 + i\epsilon) ((k + p_1^r)^2 - m^2 + i\epsilon)}{k/+ p/r - p_1^r + m} \frac{k/- p/r + m}{\int^2} \times \frac{1}{((k + p_1^r - p_1)^2 - m^2 + i\epsilon)} \frac{2}{((k - p_2^r)^2 - m^2 + i\epsilon)}$$

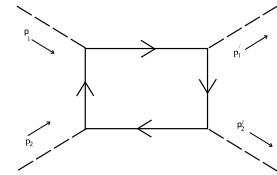
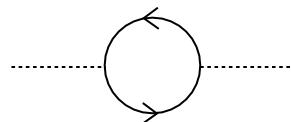


Figure 28:

Notice that the high momentum limit of the integral is $d^4 k/k^4$, which is no longer finite, but diverges logarithmically. You will have to wait until next term to make sense of this integral.

There's an overall minus sign sitting in front of this amplitude. This is a generic feature of diagrams with fermions running in loops: each fermionic loop in a diagram gives rise to an extra minus sign. We can see this rather simply in the diagram



which involves the expression

$$\overline{\psi_\alpha(x)} \overline{\psi_\beta(y)} \psi_\beta(y) \psi_\alpha(x) = -\psi_\beta(y) \psi_\alpha(x) \overline{\psi_\alpha(x)} \overline{\psi_\beta(y)} \\ = -\text{Tr}(S_F(y-x) S_F(x-y))$$

After passing the fermion fields through each other, a minus sign appears, sitting in front of the two propagators.

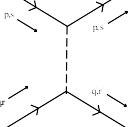
5.7.2 The Yukawa Potential Revisited

We saw in Section 3.5.2, that the exchange of a real scalar particle gives rise to a universally attractive Yukawa potential between two spin zero particles. Does the same hold for the spin 1/2 particles?

Recall that the strategy to compute the potential is to take the non-relativistic limit of the scattering amplitude, and compare with the analogous result from quantum mechanics. Our new amplitude now also includes the spinor degrees of freedom $u(p)$ and $v(p)$. In the non-relativistic limit, $p \rightarrow (m, p)$, and

$$u(p) = \frac{p \cdot \sigma \xi}{\sqrt{p \cdot \sigma \xi}} \rightarrow \frac{\sqrt{m} \xi}{m} \\ v(p) = \frac{\sqrt{p \cdot \sigma^- \xi}}{-p \cdot \sigma^- \xi} \rightarrow \frac{\sqrt{m} \xi}{-\xi} \quad (5.49)$$

In this limit, the spinor contractions in the amplitude for $\psi\psi \rightarrow \psi\psi$ scattering (5.45) become $u^s \bar{u}^s \cdot u^s \bar{u}^s = 2m\delta^{ss}$ and the amplitude is



$$= -i(-i\lambda)^2 (2m) \frac{\delta^{s s} \delta^{r r}}{(p \rightarrow \rightarrow p')^2 + \mu^2} - \frac{\delta^{s' r} \delta^{r' s}}{(p \rightarrow \rightarrow q')^2 + \mu^2} \quad (5.50)$$

The δ symbols tell us that spin is conserved in the non-relativistic limit, while the momentum dependence is the same as in the bosonic case, telling us that once again the particles feel an attractive Yukawa potential,

$$U(r) = -\frac{\lambda^2 e^{-\mu r}}{4\pi r} \quad (5.51)$$

Repeating the calculation for $\psi \bar{\psi} \rightarrow \psi \bar{\psi}$, there are two minus signs which cancel each other. The first is the extra overall minus sign in the scattering amplitude (5.46),

due to the fermionic nature of the particles. The second minus sign comes from the non-relativistic limit of the spinor contraction for anti-particles in (5.46), which is $v^{-s} \cdot v^s = -2m\delta^{ss}$. These two signs cancel, giving us once again an attractive Yukawa potential (5.51).

5.7.3 Pseudo-Scalar Coupling

Rather than the standard Yukawa coupling, we could instead consider

$$L_{Yuk} = -\lambda \varphi \bar{\psi} \gamma^5 \psi \quad (5.52)$$

This still preserves parity if φ is a pseudoscalar, i.e.

$$P : \varphi(\rightarrow x, t) \rightarrow -\varphi(-\rightarrow x, t) \quad (5.53)$$

We can compute in this theory very simply: the Feynman rule for the interaction vertex is now changed to a factor of $-i\lambda\gamma^5$. For example, the Feynman diagrams for $\psi\psi \rightarrow \psi\psi$ scattering are again given by Figure 25, with the amplitude now

$$A = (-i\lambda)^2 \frac{[u^{-s}(p \rightarrow) \gamma u(p)]^\dagger [u^{-r}(\rightarrow q) \gamma u(q)]^\dagger}{(p - p^r)^2 - \mu^2} - \frac{[u^{-s}(p \rightarrow) \gamma u(\rightarrow)] q [u^{-r}(\rightarrow q) \gamma u(p)]}{(p - q^r)^2 - \mu^2}$$

We could again try to take the non-relativistic limit for this amplitude. But this time, things work a little differently. Using the expressions for the spinors (5.49), we have $\gamma u \rightarrow 0$ in the non-relativistic limit. To find the non-relativistic amplitude, u^-

we must go to next to leading order. One can easily check that $u^{-s}(p \rightarrow) \gamma^5 u^s(p \rightarrow) \rightarrow m \xi^s T(p \rightarrow - p \rightarrow) \cdot \sigma \xi^s$. So, in the non-relativistic limit, the leading order amplitude arising from pseudoscalar exchange is given by a spin-spin coupling,

$$\rightarrow +im(-i\lambda)^2 \frac{[\xi^s T(p \rightarrow - p \rightarrow) \cdot \sigma \xi^s] [\xi^r T(p \rightarrow - p \rightarrow) \cdot \sigma \xi^r]}{(p \rightarrow - p \rightarrow)^2 + \mu^2} \quad (5.54)$$

6. Quantum Electrodynamics

In this section we finally get to quantum electrodynamics (QED), the theory of light interacting with charged matter. Our path to quantization will be as before: we start with the free theory of the electromagnetic field and see how the quantum theory gives rise to a photon with two polarization states. We then describe how to couple the photon to fermions and to bosons.

6.1 Maxwell's Equations

The Lagrangian for Maxwell's equations in the absence of any sources is simply

$$L = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \quad (6.1)$$

where the field strength is defined by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (6.2)$$

The equations of motion which follow from this Lagrangian are

$$\partial_\mu \frac{\partial L}{\partial (\partial_\mu A_\nu)} = -\partial_\mu F^{\mu\nu} = 0 \quad (6.3)$$

Meanwhile, from the definition of $F_{\mu\nu}$, the field strength also satisfies the Bianchi identity

$$\partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} = 0 \quad (6.4)$$

To make contact with the form of Maxwell's equations you learn about in high school, we need some 3-vector notation. If we define $A^\mu = (\varphi, \mathbf{A}^\rightarrow)$, then the electric field \mathbf{E}^\rightarrow and magnetic field \mathbf{B}^\rightarrow are defined by

$$\mathbf{E}^\rightarrow = -\nabla\varphi - \frac{\partial \mathbf{A}^\rightarrow}{\partial t} \quad \text{and} \quad \mathbf{B}^\rightarrow = \nabla \times \mathbf{A}^\rightarrow \quad (6.5)$$

which, in terms of $F_{\mu\nu}$, becomes

$$F_{\mu\nu} = \begin{matrix} & & & \\ & 0 & E_x & E_y & E_z \\ & -E_x & 0 & -B_z & B_y \\ & -E_y & B_z & 0 & -B_x \\ & -E_z & -B_y & B_x & 0 \end{matrix} \quad (6.6)$$

The Bianchi identity (6.4) then gives two of Maxwell's equations,

$$\nabla \cdot \mathbf{B}^\rightarrow = 0 \quad \text{and} \quad \frac{\partial \mathbf{B}^\rightarrow}{\partial t} = -\nabla \times \mathbf{E}^\rightarrow \quad (6.7)$$

These remain true even in the presence of electric sources. Meanwhile, the equations of motion give the remaining two Maxwell equations,

$$\nabla \cdot \mathbf{E}^{\rightarrow} = 0 \quad \text{and} \quad \frac{\partial \mathbf{E}^{\rightarrow}}{\partial t} = \nabla \times \mathbf{B}^{\rightarrow} \quad (6.8)$$

As we will see shortly, in the presence of charged matter these equations pick up extra terms on the right-hand side.

6.1.1 Gauge Symmetry

The massless vector field A_{μ} has 4 components, which would naively seem to tell us that the gauge field has 4 degrees of freedom. Yet we know that the photon has only two degrees of freedom which we call its polarization states. How are we going to resolve this discrepancy? There are two related comments which will ensure that quantizing the gauge field A_{μ} gives rise to 2 degrees of freedom, rather than 4.

- The field A_0 has no kinetic term A_0 in the Lagrangian: it is not dynamical. This means that if we are given some initial data A_i and A_0 at a time t_0 , then the field A_0 is fully determined by the equation of motion $\nabla \cdot \mathbf{E}^{\rightarrow} = 0$ which, expanding out, reads

$$\nabla^2 A_0 + \nabla \cdot \frac{\partial \mathbf{A}^{\rightarrow}}{\partial t} = 0 \quad (6.9)$$

This has the solution

$$A_0(\vec{x}) = \int d^3x' \frac{(\nabla' \cdot \partial \mathbf{A}^{\rightarrow} / \partial t)}{(\vec{x}' - \vec{x})^r} \frac{4\pi}{r} \quad (6.10)$$

So A_0 is not independent: we don't get to specify A_0 on the initial time slice. It looks like we have only 3 degrees of freedom in A_{μ} rather than 4. But this is still one too many.

- The Lagrangian (6.3) has a *very* large symmetry group, acting on the vector potential as

$$A_{\mu}(x) \rightarrow A_{\mu}(x) + \partial_{\mu}\lambda(x) \quad (6.11)$$

for any function $\lambda(x)$. We'll ask only that $\lambda(x)$ dies off suitably quickly at

spatial

$\vec{x} \rightarrow \infty$. We call this a *gauge symmetry*. The field strength is invariant under the gauge symmetry:

$$F_{\mu\nu} \rightarrow \partial_{\mu}(A_{\nu} + \partial_{\nu}\lambda) - \partial_{\nu}(A_{\mu} + \partial_{\mu}\lambda) = F_{\mu\nu} \quad (6.12)$$

So what are we to make of this? We have a theory with an infinite number of symmetries, one for each function $\lambda(x)$. Previously we only encountered symmetries which act the same at all points in spacetime, for example $\psi \rightarrow e^{i\alpha}\psi$ for a complex scalar field. Noether's theorem told us that these symmetries give rise to conservation laws. Do we now have an infinite number of conservation laws?

The answer is no! Gauge symmetries have a very different interpretation than the global symmetries that we make use of in Noether's theorem. While the latter take a physical state to another physical state with the same properties, the gauge symmetry is to be viewed as a redundancy in our description. That is, two states related by a gauge symmetry are to be identified: they are the same physical state. (There is a small caveat to this statement which is explained in Section 6.3.1). One way to see that this interpretation is necessary is to notice that Maxwell's equations are not sufficient to specify the evolution of A_μ . The equations read,

$$[\eta_{\mu\nu}(\partial^\rho\partial_\rho) - \partial_\mu\partial_\nu] A^\nu = 0 \quad (6.13)$$

But the operator $[\eta_{\mu\nu}(\partial^\rho\partial_\rho) - \partial_\mu\partial_\nu]$ is not invertible: it annihilates any function of the form $\partial_\mu\lambda$. This means that given any initial data, we have no way to uniquely determine A_μ at a later time since we can't distinguish between A_μ and $A_\mu + \partial_\mu\lambda$. This would be problematic if we thought that A_μ is a physical object. However, if we're happy to identify A_μ and $A_\mu + \partial_\mu\lambda$ as corresponding to the same physical state, then our problems disappear.

Since gauge invariance is a redundancy of the system, we might try to formulate the theory purely in terms of the local, physical, gauge invariant objects $E^{\vec{r}}$ and $B^{\vec{r}}$. This

is fine for the free classical theory: Maxwell's equations were, after all, first written in terms of $E^{\vec{r}}$ and $B^{\vec{r}}$. But it is

not possible to describe certain quantum phenomena, such as the Aharonov-Bohm effect, without using the gauge potential A_μ . We will see shortly that we also require the gauge potential to describe classically charged fields. To describe Nature, it appears that we have to introduce quantities A_μ that we can never measure.

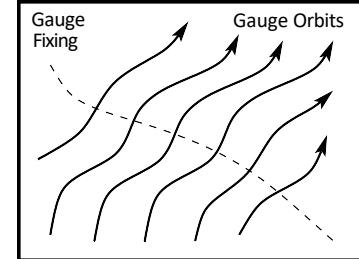


Figure 29:

The picture that emerges for the theory of electromagnetism is of an enlarged phase space, foliated by gauge orbits as shown in the figure. All states that lie along a given

line can be reached by a gauge transformation and are identified. To make progress, we pick a representative from each gauge orbit. It doesn't matter which representative we pick — after all, they're all physically equivalent. But we should make sure that we pick a "good" gauge, in which we cut the orbits.

Different representative configurations of a physical state are called different *gauges*. There are many possibilities, some of which will be more useful in different situations. Picking a gauge is rather like picking coordinates that are adapted to a particular problem. Moreover, different gauges often reveal slightly different aspects of a problem. Here we'll look at two different gauges:

- **Lorentz Gauge:** $\partial_\mu A^\mu = 0$

To see that we can always pick a representative configuration satisfying $\partial_\mu A^\mu = 0$, suppose that we're handed a gauge field ${}_\mu A^r$ satisfying $\partial_\mu (A^r)^\mu = f(x)$. Then we choose $A_\mu \not\equiv A^r + \partial_\mu \lambda$, where

$$\partial_\mu \partial^\mu \lambda = -f \quad (6.14)$$

This equation always has a solution. In fact this condition doesn't pick a unique representative from the gauge orbit. We're always free to make further gauge transformations with $\partial_\mu \partial^\mu \lambda = 0$, which also has non-trivial solutions. As the name suggests, the Lorentz gauge³ has the advantage that it is Lorentz invariant.

- **Coulomb Gauge:** $\nabla \cdot A^\rightarrow = 0$

We can make use of the residual gauge transformations in Lorentz gauge to pick $\nabla \cdot A^\rightarrow = 0$. (The argument is the same as before). Since A_0 is fixed by (6.10), we have as a consequence

$$A_0 = 0 \quad (6.15)$$

(This equation will no longer hold in Coulomb gauge in the presence of charged matter). Coulomb gauge breaks Lorentz invariance, so may not be ideal for some purposes. However, it is very useful to exhibit the physical degrees of freedom: the 3 components of A^\rightarrow satisfy a single constraint: $\nabla \cdot A^\rightarrow = 0$, leaving behind just 2 degrees of freedom. These will be identified with the two polarization states of the photon. Coulomb gauge is sometimes called radiation gauge.

³Named after Lorenz who had the misfortune to be one letter away from greatness.

6.2 The Quantization of the Electromagnetic Field

In the following we shall quantize free Maxwell theory twice: once in Coulomb gauge, and again in Lorentz gauge. We'll ultimately get the same answers and, along the way, see that each method comes with its own subtleties.

The first of these subtleties is common to both methods and comes when computing the momentum π^μ conjugate to A_μ ,

$$\begin{aligned}\pi^0 &= \frac{\partial L}{\partial A^0} = 0 \\ \underline{\frac{\partial L}{\partial A^i}} &= -F^{0i} \equiv E^i \quad (6.16)\end{aligned}$$

so the momentum π^0 conjugate to A_0 vanishes. This is the mathematical consequence of the statement we made above: A_0 is not a dynamical field. Meanwhile, the momentum conjugate to A_i is our old friend, the electric field. We can compute the Hamiltonian,

$$\begin{aligned}H &= \int d^3x \pi^i A_i - L \\ &= \int d^3x \frac{1}{2} \vec{E} \cdot \vec{E} + \frac{1}{2} \vec{B} \cdot \vec{B} - A_0 (\nabla \cdot \vec{E}) \quad (6.17)\end{aligned}$$

So A_0 acts as a Lagrange multiplier which imposes Gauss' law

$$\nabla \cdot \vec{E} = 0 \quad (6.18)$$

which is now a constraint on the system in which \vec{A} are the physical degrees of freedom. Let's now see how to treat this system using different gauge fixing conditions.

6.2.1 Coulomb Gauge

In Coulomb gauge, the equation of motion for \vec{A} is

$$\partial_\mu \partial^\mu \vec{A} = 0 \quad (6.19)$$

which we can solve in the usual way,

$$\vec{A} = \frac{d^3 p}{e^{\vec{p} \cdot \vec{x}} (2\pi)^3} \vec{\xi}(p) \quad (6.20)$$

with $p^2 = |\vec{p}|^2$. The constraint $\nabla \cdot \vec{A} = 0$ tells us that $\vec{\xi}$ must satisfy

$$\vec{\xi} \cdot \vec{p} = 0 \quad (6.21)$$

which means that $\xi \rightarrow$ is perpendicular to the direction of motion $p \rightarrow$. We can pick $\xi \rightarrow$ ($p \rightarrow$) to be a linear combination of two orthonormal vectors $\epsilon_r \rightarrow$, $r = 1, 2$, each of which satisfies

$$\epsilon_r(p \rightarrow) \cdot p \rightarrow = 0 \text{ and}$$

$$\epsilon_r(p \rightarrow) \cdot \epsilon_s(p \rightarrow) = \delta_{rs} \quad r, s =$$

1, 2

(6.22) These two vectors correspond

to the two polarization states of the photon. It's worth pointing out that you can't consistently pick a continuous basis of polarization vectors for every value of $p \rightarrow$ because you can't comb the hair on a sphere. But this topological fact doesn't cause any complications in computing QED scattering processes.

To quantize we turn the Poisson brackets into commutators. Naively we would write

$$\begin{aligned} [A_i(\rightarrow x), A_j(\rightarrow y)] &= [E^i(\rightarrow x), E^j(\rightarrow y)] = 0 \\ [A_i(\rightarrow x), E^j(\rightarrow y)] &= i\delta^j \delta_i^{(3)}(\rightarrow x - \rightarrow y) \end{aligned} \quad (6.23)$$

But this can't quite be right, because it's not consistent with the constraints. We still want to have $\nabla \cdot A \rightarrow = \nabla \cdot E \rightarrow = 0$, now imposed on the operators. But from the commutator relations above, we see

$$[\nabla \cdot A \rightarrow(\rightarrow x), \nabla \cdot E \rightarrow(\rightarrow y)] = i\nabla^2 \delta^{(3)}(\rightarrow x - \rightarrow y) \neq 0 \quad (6.24)$$

What's going on? In imposing the commutator relations (6.23) we haven't correctly taken into account the constraints. In fact, this is a problem already in the classical theory, where the Poisson bracket structure is already altered⁴. The correct Poisson bracket structure leads to an alteration of the last commutation relation,

$$[A_i(\rightarrow x), E_j(\rightarrow y)] = i \delta_{ij} - \frac{\partial_i \partial_j}{\nabla^2} \delta^{(3)}(\rightarrow x - \rightarrow y) \quad (6.25)$$

To see that this is now consistent with the constraints, we can rewrite the right-hand side of the commutator in momentum space,

$$[A_i(\rightarrow x), E_j(\rightarrow y)] = i \int \frac{d^3 p}{(2\pi)^3} \delta_{ij} - \frac{p_i p_j}{|\rightarrow p|} e^{i\rightarrow p \cdot (\rightarrow x - \rightarrow y)} \quad (6.26)$$

which is now consistent with the constraints, for example

$$[\partial_i A_i(\rightarrow x), E_j(\rightarrow y)] = i \int \frac{d^3 p}{(2\pi)^3} \delta_{ij} - \frac{p_i p_j}{|\rightarrow p|} i p_i e^{i\rightarrow p \cdot (\rightarrow x - \rightarrow y)} = 0 \quad (6.27)$$

⁴For a nice discussion of the classical and quantum dynamics of constrained systems, see the small book by Paul Dirac, "Lectures on Quantum Mechanics"

We now write \vec{A}^\rightarrow in the usual mode expansion,

$$\begin{aligned}\vec{A}(\rightarrow x) &= \int \frac{d^3 p}{(2\pi)^3} \frac{1}{2|\vec{p}|} \sum_{r=1}^{\infty} \sum_{\text{h}} \frac{h}{r \cdot i p \rightarrow \cdot \rightarrow x} \frac{i}{r \dagger -i p \rightarrow \cdot \rightarrow x} \\ &\quad \times \sum_{\text{h}} \rightarrow \epsilon_r(\rightarrow p) a_p^r e + a_p^r e^\dagger \\ E^\rightarrow(\rightarrow x) &= \frac{(-i)}{(2\pi)^3} \frac{|\vec{p}|}{2} \sum_{r=1}^{\infty} \rightarrow \epsilon_r(\rightarrow p) a_p^r e^{ip \rightarrow \cdot \rightarrow x} - a_p^r e^\dagger e^{-ip \rightarrow \cdot \rightarrow x} \end{aligned}\quad (6.28)$$

where, as before, the polarization vectors satisfy

$$\begin{aligned}\rightarrow \epsilon_r(p \rightarrow) \cdot p \rightarrow &= 0 \quad \text{and} \\ \rightarrow \epsilon_s(p \rightarrow) &= \delta_{rs}\end{aligned}\quad (6.29)$$

It is not hard to show that the commutation relations (6.25) are equivalent to the usual commutation relations for the creation and annihilation operators,

$$\begin{aligned}[a_p^r, a_q^s] &= [a_p^r \dagger, a_q^s \dagger] = 0 \\ [a_p^r, a_q^s \dagger] &= (2\pi)^3 \delta^{rs} \delta^{(3)}(p \rightarrow - q \rightarrow)\end{aligned}\quad (6.30)$$

where, in deriving this, we need the completeness relation for the polarization vectors,

$$\sum_{r=1}^{\infty} \sum_{i,j} \frac{\epsilon_i(p \rightarrow) \epsilon_j(p \rightarrow)}{|\vec{p}|^2} = \delta^{ij}\quad (6.31)$$

You can easily check that this equation is true by acting on both sides with a basis of vectors $(\rightarrow \epsilon_1(p \rightarrow), \rightarrow \epsilon_2(p \rightarrow), p \rightarrow)$.

We derive the Hamiltonian by substituting (6.28) into (6.17). The last term vanishes in Coulomb gauge. After normal ordering, and playing around with $\rightarrow \epsilon_r$ polarization vectors, we get the simple expression

$$H = \frac{1}{(2\pi)^3} \sum_{r=1}^{\infty} \frac{d^3 p}{|\vec{p}|} a_p^r a_p^r \dagger\quad (6.32)$$

The Coulomb gauge has the advantage that the physical degrees of freedom are manifest. However, we've lost all semblance of Lorentz invariance. One place where this manifests itself is in the propagator for the fields $A_i(x)$ (in the Heisenberg picture). In Coulomb gauge the propagator reads

$$D_{ij}(x - y) \equiv \langle 0 | T A_i(x) A_j(y) | 0 \rangle = \frac{1}{(2\pi)^4} \frac{d^4 p}{p^2 + i\epsilon} \frac{i}{\delta_{ij} - \frac{p_i p_j}{|p|^2}} e^{-ip \cdot (x-y)}\quad (6.33)$$

The tr superscript on the propagator refers to the “transverse” part of the photon. When we turn to the interacting theory, we will have to fight to massage this propagator into something a little nicer.

6.2.2 Lorentz Gauge

We could try to work in a Lorentz invariant fashion by imposing the Lorentz gauge condition $\partial_\mu A^\mu = 0$. The equations of motion that follow from the action are then

$$\partial_\mu \partial^\mu A^\nu = 0 \quad (6.34)$$

Our approach to implementing Lorentz gauge will be a little different from the method we used in Coulomb gauge. We choose to change the theory so that (6.34) arises directly through the equations of motion. We can achieve this by taking the Lagrangian

$$L = \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2} (\partial_\mu A^\mu)^2 \quad (6.35)$$

The equations of motion coming from this action are

$$\partial_\mu F^{\mu\nu} + \partial^\nu (\partial_\mu A^\mu) = \partial_\mu \partial^\mu A^\nu = 0 \quad (6.36)$$

(In fact, we could be a little more general than this, and consider the Lagrangian

$$L = \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{1}{2\alpha} (\partial_\mu A^\mu)^2 \quad (6.37)$$

with arbitrary α and reach similar conclusions. The quantization of the theory is independent of α and, rather confusingly, different choices of α are sometimes also referred to as different “gauges”. We will use $\alpha = 1$, which is called “Feynman gauge”. The other common choice, $\alpha = 0$, is called “Landau gauge”.)

Our plan will be to quantize the theory (6.36), and only later impose the constraint $\partial_\mu A^\mu = 0$ in a suitable manner on the Hilbert space of the theory. As we'll see, we will also have to deal with the residual gauge symmetry of this theory which will prove a little tricky. At first, we can proceed very easily, because both π^0 and π^i are dynamical:

$$\begin{aligned} \dot{\pi}^0 &= \frac{\partial L}{\partial \dot{A}_0} = -\partial_\mu A^\mu \\ \dot{\pi}^i &= \frac{\partial L}{\partial \dot{A}^i} = \partial^i A^0 - \dot{A}^i \end{aligned} \quad (6.38)$$

Turning these classical fields into operators, we can simply impose the usual commutation relations,

$$\begin{aligned} [A_\mu(\vec{x}), A_\nu(\vec{y})] &= [\pi^\mu(\vec{x}), \pi^\nu(\vec{y})] = 0 \\ [A_\mu(\vec{x}), \pi_\nu(\vec{y})] &= i\eta_{\mu\nu} \delta^{(3)}(\vec{x} - \vec{y}) \end{aligned} \quad (6.39)$$

and we can make the usual expansion in terms of creation and annihilation operators and 4 polarization vectors $(\epsilon_\mu)^\lambda$, with $\lambda = 0, 1, 2, 3$.

$$A_\mu(\rightarrow x) = \frac{\int d^3 p}{(2\pi)^3} \frac{1}{|p|} \sum_{\lambda=0}^3 \epsilon_\mu(p^\lambda) a_{p^\lambda} e^+ + a_{p^\lambda} e^-$$

$$\pi^\mu(\rightarrow x) = \frac{\int d^3 p}{(2\pi)^3} \frac{|p|}{2} \sum_{\lambda=0}^3 \epsilon_\mu(p^\lambda) a_{p^\lambda} e^{ip \cdot x} - a_{p^\lambda} e^{-ip \cdot x}$$
(6.40)

Note that the momentum π^μ comes with a factor of $(+i)$, rather than the familiar $(-i)$ that we've seen so far. This can be traced to the fact that the momentum (6.38) for the classical fields takes the form $\pi^\mu = -A^\mu + \dots$. In the Heisenberg picture, it becomes clear that this descends to $(+i)$ in the definition of momentum.

There are now four polarization 4-vectors $\epsilon^\lambda(p^\lambda)$, instead of the two polarization 3-vectors that we met in the Coulomb gauge. Of these four 4-vectors, we pick ϵ^0 to be timelike, while $\epsilon^{1,2,3}$ are spacelike. We pick the normalization

$$\epsilon^\lambda \cdot \epsilon^{\lambda'} = \eta^{\lambda\lambda'} \quad (6.41)$$

which also means that

$$(\epsilon_\mu)^\lambda (\epsilon_\nu)^{\lambda'} \eta_{\lambda\lambda'} = \eta_{\mu\nu} \quad (6.42)$$

The polarization vectors depend on the photon 4-momentum $p = (|p|, p^\lambda)$. Of the two spacelike polarizations, we will choose ϵ^1 and ϵ^2 to lie transverse to the momentum:

$$\epsilon^1 \cdot p = \epsilon^2 \cdot p = 0 \quad (6.43)$$

The third vector ϵ^3 is the longitudinal polarization. For example, if the momentum lies along the x^3 direction, so $p \sim (1, 0, 0, 1)$, then

$$\epsilon^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \epsilon^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \epsilon^3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad (6.44)$$

For other 4-momenta, the polarization vectors are the appropriate Lorentz transformations of these vectors, since (6.43) are Lorentz invariant.

We do our usual trick, and translate the field commutation relations (6.39) into those for creation and annihilation operators. We find $[a_\mu^\lambda, a_{\mu'}^{\lambda'}] = [a_{\mu'}^{\lambda'}, a_{\mu'}^{\lambda'}] = 0$ and

$$[a_{\mu'}^\lambda, a_{\mu'}^{\lambda'}] = -\eta^{\lambda\lambda'} (2\pi)^3 \delta^{(3)}(p^\lambda - q^\lambda) \quad (6.45)$$

The minus signs here are odd to say the least! For spacelike $\lambda = 1, 2, 3$, everything looks fine,

$$[a_{p\rightarrow}^{\lambda}, a^{\lambda'}^\dagger] = \delta^{\lambda\lambda'} (2\pi)^3 \delta^{(3)}(p\rightarrow - \rightarrow q) \quad (\text{6.46})$$

But for the timelike annihilation and creation operators, we have

$$[a_{p\rightarrow}^0, a^0^\dagger] = -(2\pi)^3 \delta^{(3)}(p\rightarrow - \rightarrow q) \quad (\text{6.47})$$

This is very odd! To see just how strange this is, we take the Lorentz invariant vacuum $|0\rangle$ defined by

$$a_p^\lambda |0\rangle = 0 \quad (\text{6.48})$$

Then we can create one-particle states in the usual way,

$$|p\rightarrow, \lambda\rangle = a_p^\lambda |0\rangle \quad (\text{6.49})$$

For spacelike polarization states, $\lambda = 1, 2, 3$, all seems well. But for the timelike polarization $\lambda = 0$, the state $|p\rightarrow, 0\rangle$ has negative norm,

$$\langle p\rightarrow, 0 | \rightarrow q, 0 \rangle = \langle 0 | a_{p\rightarrow}^0 a_{q\rightarrow}^0 | 0 \rangle = -(2\pi)^3 \delta^{(3)}(p\rightarrow - \rightarrow q) \quad (\text{6.50})$$

Wtf? That's very very strange. A Hilbert space with negative norm means negative probabilities which makes no sense at all. We can trace this negative norm back to the wrong sign of the kinetic term for A_0 in our original Lagrangian: $L = +\frac{1}{2} \partial_\mu A^\mu - \frac{1}{2} A^0 \dot{A}^0 + \dots$

At this point we should remember our constraint equation, $\partial_\mu A^\mu = 0$, which, until now, we've not imposed on our theory. This is going to come to our rescue. We will see that it will remove the timelike, negative norm states, and cut the physical polarizations down to two. We work in the Heisenberg picture, so that

$$\partial_\mu A^\mu = 0 \quad (\text{6.51})$$

makes sense as an operator equation. Then we could try implementing the constraint in the quantum theory in a number of different ways. Let's look at a number of increasingly weak ways to do this

- We could ask that $\partial_\mu A^\mu = 0$ is imposed as an equation on operators. But this can't possibly work because the commutation relations (6.39) won't be obeyed for $\pi^0 = -\partial_\mu A^\mu$. We need some weaker condition.

- We could try to impose the condition on the Hilbert space instead of directly on the operators. After all, that's where the trouble lies! We could imagine that there's some way to split the Hilbert space up into good states $|\Psi\rangle$ and bad states that somehow decouple from the system. With luck, our bad states will include the weird negative norm states that we're so disgusted by. But how can we define the good states? One idea is to impose

$$\partial_\mu A^\mu |\Psi\rangle = 0 \quad (6.52)$$

on all good, physical states $|\Psi\rangle$. But this can't work either! Again, the condition is too strong. For example, suppose we decompose $A_\mu(x) = A^+(x) + A^-(x)$ with

$$A_\mu(x) = \frac{\int d^3p}{(2\pi)^3} \sum_3 \lambda \lambda - ip \cdot x \epsilon_\mu a_p \rightarrow e$$

$$A_\mu(x) = \frac{\int d^3p}{(2\pi)^3} \sum_3 \lambda \lambda + ip \cdot x \epsilon_\mu a_p \rightarrow e \quad (6.53)$$

Then, on the vacuum $A_\mu^+ |0\rangle = 0$ automatically, but $\partial^\mu A_\mu^- |0\rangle \neq 1$. So not even the vacuum is a physical state if we use (6.52) as our constraint

- Our final attempt will be the correct one. In order to keep the vacuum as a good physical state, we can ask that physical states $|\Psi\rangle$ are defined by

$$\partial^\mu A_\mu^+ |\Psi\rangle = 0 \quad (6.54)$$

This ensures that

$$\langle \Psi | \partial_\mu A^\mu | \Psi \rangle = 0 \quad (6.55)$$

so that the operator $\partial_\mu A^\mu$ has vanishing matrix elements between physical states. Equation (6.54) is known as the *Gupta-Bleuler* condition. The linearity of the constraint means that the physical states $|\Psi\rangle$ span a physical Hilbert space H_{phys} .

So what does the physical Hilbert space H_{phys} look like? And, in particular, have we rid ourselves of those nasty negative norm states so that H_{phys} has a positive definite inner product defined on it? The answer is actually no, but almost!

Let's consider a basis of states for the Fock space. We can decompose any element of this basis as $|\Psi\rangle = |\psi_T\rangle |\varphi\rangle$, where $|\psi_T\rangle$ contains only transverse photons, created by

$a_{p\rightarrow}^{1,2\dagger}$, while $|\varphi\rangle$ contains the timelike photons created by $a^0\dagger$ and longitudinal photons created by $a^3\dagger$. The Gupta-Bleuler condition (6.54) requires

$$(a_{p\rightarrow}^3 - a^0) |\varphi\rangle = 0 \quad (6.56)$$

This means that the physical states must contain combinations of timelike and longitudinal photons. Whenever the state contains a timelike photon of momentum $p\rightarrow$, it must also contain a longitudinal photon with the same momentum. In general $|\varphi\rangle$ will be a linear combination of states $|\varphi_n\rangle$ containing n pairs of timelike and longitudinal photons, which we can write as

$$|\varphi\rangle = \sum_{n=0}^{\infty} C_n |\varphi_n\rangle \quad (6.57)$$

where $|\varphi_0\rangle = |0\rangle$ is simply the vacuum. It's not hard to show that although the condition (6.56) does indeed decouple the negative norm states, all the remaining states involving timelike and longitudinal photons have zero norm

$$\langle \varphi_m | \varphi_n \rangle = \delta_{m0}\delta_{n0} \quad (6.58)$$

This means that the inner product on H_{phys} is positive semi-definite. Which is an improvement. But we still need to deal with all these zero norm states.

The way we cope with the zero norm states is to treat them as gauge equivalent to the vacuum. Two states that differ only in their timelike and longitudinal photon content, $|\varphi_n\rangle$ with $n \geq 1$ are said to be physically equivalent. We can think of the gauge symmetry of the classical theory as descending to the Hilbert space of the quantum theory. Of course, we can't just stipulate that two states are physically identical unless they give the same expectation value for all physical observables. We can check that this is true for the Hamiltonian, which can be easily computed to be

$$H = \int \frac{d^3p}{(2\pi)^3} |\rightarrow p| \sum_{i=1}^3 a_i^\dagger a_i - a^0 \dagger a^0 \quad p\rightarrow \quad p\rightarrow \quad p\rightarrow \quad ! \quad (6.59)$$

But the condition (6.56) ensures that $\langle \Psi | a^3\dagger a^3 | \Psi \rangle = \langle \Psi | a^0\dagger a^0 | \Psi \rangle$ so that the contributions from the timelike and longitudinal photons cancel amongst themselves in the Hamiltonian. This also renders the Hamiltonian positive definite, leaving us just with the contribution from the transverse photons as we would expect.

In general, one can show that the expectation values of all gauge invariant operators evaluated on physical states are independent of the coefficients C_n in (6.57).

Propagators

Finally, it's a simple matter to compute the propagator in Lorentz gauge. It is given by

$$\langle 0 | T A_\mu(x) A_\nu(y) | 0 \rangle = \frac{\int d^4 p}{(2\pi)^4} \frac{-i\eta_{\mu\nu}}{p^2 + i\epsilon} e^{-ip \cdot (x-y)} \quad (6.60)$$

This is a lot nicer than the propagator we found in Coulomb gauge: in particular, it's Lorentz invariant. We could also return to the Lagrangian (6.37). Had we pushed through the calculation with arbitrary coefficient α , we would find the propagator,

$$\langle 0 | T A_\mu(x) A_\nu(y) | 0 \rangle = \frac{\int d^4 p}{(2\pi)^4} \frac{-i}{p^2 + i\epsilon} (\eta_{\mu\nu} + (\alpha - 1) \frac{\mu_\nu}{p^2}) e^{-ip \cdot (x-y)} \quad (6.61)$$

6.3 Coupling to Matter

Let's now build an interacting theory of light and matter. We want to write down a Lagrangian which couples A_μ to some matter fields, either scalars or spinors. For example, we could write something like

$$L = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - j^\mu A_\mu \quad (6.62)$$

where j^μ is some function of the matter fields. The equations of motion read

$$\partial_\mu F^{\mu\nu} = j^\nu \quad (6.63)$$

so, for consistency, we require

$$\partial_\mu j^\mu = 0 \quad (6.64)$$

In other words, j^μ must be a conserved current. But we've got lots of those! Let's look at how we can couple two of them to electromagnetism.

6.3.1 Coupling to Fermions

The Dirac Lagrangian

$$L = \bar{\psi} (i \partial/\! - m) \psi \quad (6.65)$$

has an internal symmetry $\psi \rightarrow e^{-i\alpha} \psi$ and $\bar{\psi} \rightarrow e^{+i\alpha} \bar{\psi}$, with $\alpha \in \mathbb{R}$. This gives rise to the conserved current $j_\nu^\mu = \bar{\psi} \gamma^\mu \psi$. So we could look at the theory of electromagnetism coupled to fermions, with the Lagrangian,

$$L = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar{\psi} (i \partial/\! - m) \psi - e \bar{\psi} \gamma^\mu A_\mu \psi \quad (6.66)$$

where we've introduced a coupling constant e . For the free Maxwell theory, we have seen that the existence of a gauge symmetry was crucial in order to cut down the physical degrees of freedom to the requisite 2. Does our interacting theory above still have a gauge symmetry? The answer is yes. To see this, let's rewrite the Lagrangian as

$$L = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(iD/\!\!- m)\psi \quad (6.67)$$

where $D_\mu\psi = \partial_\mu\psi + ieA_\mu\psi$ is called the *covariant derivative*. This Lagrangian is invariant under gauge transformations which act as

$$A_\mu \rightarrow A_\mu + \partial_\mu\lambda \quad \text{and} \quad \psi \rightarrow e^{-ie\lambda}\psi \quad (6.68)$$

for an arbitrary function $\lambda(x)$. The tricky term is the derivative acting on ψ , since this will also hit the $e^{-ie\lambda}$ piece after the transformation. To see that all is well, let's look at how the covariant derivative transforms. We have

$$\begin{aligned} D_\mu\psi &= \partial_\mu\psi + ieA_\mu\psi \\ &\rightarrow \partial_\mu(e^{-ie\lambda}\psi) + ie(A_\mu + \partial_\mu\lambda)(e^{-ie\lambda}\psi) \\ &= e^{-ie\lambda}D_\mu\psi \end{aligned} \quad (6.69)$$

so the covariant derivative has the nice property that it merely picks up a phase under the gauge transformation, with the derivative of $e^{-ie\lambda}$ cancelling the transformation of the gauge field. This ensures that the whole Lagrangian is invariant, since $\psi \rightarrow e^{+ie\lambda(x)}\psi$.

Electric Charge

The coupling e has the interpretation of the electric charge of the ψ particle. This follows from the equations of motion of classical electromagnetism $\partial_\mu F^{\mu\nu} = j^\nu$: we know that the j^0 component is the charge density. We therefore have the total charge Q given by

$$Q = e \int d^3x \bar{\psi}(\rightarrow x)\gamma^0\psi(\rightarrow x) \quad (6.70)$$

After treating this as a quantum equation, we have

$$Q = e \sum_s \frac{(b^s)^\dagger b^s - c^s)^\dagger c^s)}{(2\pi)^3 \int p \rightarrow p \rightarrow p \rightarrow p} \quad (6.71)$$

which is the number of particles, minus the number of antiparticles. Note that the particle and the anti-particle are required by the formalism to have opposite electric

charge. For QED, the theory of light interacting with electrons, the electric charge is usually written in terms of the dimensionless ratio α , known as the fine structure constant

$$\alpha = \frac{e^2}{4\pi k c} \approx \frac{1}{137} \quad (6.72)$$

Setting $k = c = 1$, we have $e = \sqrt{4\pi\alpha} \approx 0.3$.

There's a small subtlety here that's worth elaborating on. I stressed that there's a radical difference between the interpretation of a global symmetry and a gauge symmetry. The former takes you from one physical state to another with the same properties and results in a conserved current through Noether's theorem. The latter is a redundancy in our description of the system. Yet in electromagnetism, the gauge symmetry $\psi \rightarrow e^{+ie\lambda(x)}\psi$ seems to lead to a conservation law, namely the conservation of electric charge. This is because among the infinite number of gauge symmetries parameterized by a function $\lambda(x)$, there is also a single global symmetry: that with $\lambda(x) = \text{constant}$. This is a true symmetry of the system, meaning that it takes us to another physical state. More generally, the subset of global symmetries from among the gauge symmetries are those for which $\lambda(x) \rightarrow \alpha = \text{constant}$ as $x \rightarrow \infty$. These take us from one physical state to another.

Finally, let's check that the 4×4 matrix C that we introduced in Section 4.5 really deserves the name "charge conjugation matrix". If we take the complex conjugation of the Dirac equation, we have

$$(i\gamma^\mu \partial_\mu - e\gamma^\mu A_\mu - m)\psi = 0 \Rightarrow (-i(\gamma^\mu)^\wedge \partial_\mu - e(\gamma^\mu)^\wedge A_\mu - m)\psi^\wedge = 0$$

Now using the defining equation $C^\dagger \gamma^\mu C = -(\gamma^\mu)^\wedge$, and the definition $\psi^{(c)} = C\psi^\wedge$, we see that the charge conjugate spinor $\psi^{(c)}$ satisfies

$$(i\gamma^\mu \partial_\mu + e\gamma^\mu A_\mu - m)\psi^{(c)} = 0 \quad (6.73)$$

So we see that the charge conjugate spinor $\psi^{(c)}$ satisfies the Dirac equation, but with charge $-e$ instead of $+e$.

6.3.2 Coupling to Scalars

For a real scalar field, we have no suitable conserved current. This means that we can't couple a real scalar field to a gauge field.

Let's now consider a complex scalar field ϕ . (For this section, I'll depart from our previous notation and call the scalar field ϕ to avoid confusing it with the spinor). We have a symmetry $\phi \rightarrow e^{-i\alpha}\phi$. We could try to couple the associated current to the gauge field,

$$L_{\text{int}} = -i((\partial_\mu \phi)^\wedge \phi - \phi^\wedge \partial_\mu \phi) A^\mu \quad (6.74)$$

But this doesn't work because

- The theory is no longer gauge invariant
- The current j^μ that we coupled to A_μ depends on $\partial_\mu \phi$. This means that if we try to compute the current associated to the symmetry, it will now pick up a contribution from the $j^\mu A_\mu$ term. So the whole procedure wasn't consistent.

We solve both of these problems simultaneously by remembering the covariant derivative. In this scalar theory, the combination

$$D_\mu \phi = \partial_\mu \phi + ieA_\mu \phi \quad (6.75)$$

again transforms as $D_\mu \phi \rightarrow e^{-ie\lambda} D_\mu \phi$ under a gauge transformation $A_\mu \rightarrow A_\mu + \partial_\mu \lambda$ and $\phi \rightarrow e^{-ie\lambda} \phi$. This means that we can construct a gauge invariant action for a charged scalar field coupled to a photon simply by promoting all derivatives to covariant derivatives

$$L = \frac{1}{4} F^{\mu\nu} F^{\mu\nu} + D_\mu \phi^\wedge D^\mu \phi - m^2 |\phi|^2 \quad (6.76)$$

In general, this trick works for any theory. If we have a $U(1)$ symmetry that we wish to couple to a gauge field, we may do so by replacing all derivatives by suitable covariant derivatives. This procedure is known as *minimal coupling*.

6.4 QED

Let's now work out the Feynman rules for the full theory of quantum electrodynamics (QED) – the theory of electrons interacting with light. The Lagrangian is

$$\underline{L} = \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar{\psi} (i D^\mu - m) \psi \quad (6.77)$$

where $D_\mu = \partial_\mu + ieA_\mu$.

The route we take now depends on the gauge choice. If we worked in Lorentz gauge previously, then we can jump straight to Section 6.5 where the Feynman rules for QED are written down. If, however, we worked in Coulomb gauge, then we still have a bit of work in front of us in order to massage the photon propagator into something Lorentz invariant. We will now do that.

In Coulomb gauge $\nabla \cdot \mathbf{A}^\rightarrow = 0$, the equation of motion arising from varying A_0 is now

$$\int_0 -\nabla^2 A_0 = e \psi^\dagger \psi \equiv ej \quad (6.78)$$

which has the solution

$$A_0(\vec{x}, t) = \frac{e}{4\pi} \int \frac{d^3x' j^0(\vec{x}', t)}{|\vec{x} - \vec{x}'|} \quad (6.79)$$

In Coulomb gauge we can rewrite the Maxwell part of the Lagrangian as

$$\begin{aligned} L_{\text{Maxwell}} &= \int d^3x \frac{1}{2} \dot{\mathbf{E}}^\rightarrow \cdot \frac{1}{2} \mathbf{B}^\rightarrow \cdot \\ &= \int d^3x \frac{1}{2} (\dot{\mathbf{A}}^\rightarrow + \nabla A_0) \cdot \frac{1}{2} \mathbf{B}^\rightarrow \cdot \\ &= \int d^3x \frac{1}{2} \dot{\mathbf{A}}^\rightarrow \cdot \frac{1}{2} \mathbf{B}^\rightarrow + \frac{1}{2} (\nabla A_0) \cdot \frac{1}{2} \mathbf{B}^\rightarrow \cdot \end{aligned} \quad (6.80)$$

where the cross-term has vanished using $\nabla \cdot \mathbf{A}^\rightarrow = 0$. After integrating the second term by parts and inserting the equation for A_0 , we have

$$L_{\text{Maxwell}} = \int d^3x \frac{1}{2} \dot{\mathbf{A}}^\rightarrow \cdot \frac{1}{2} \mathbf{B}^\rightarrow + \frac{e^2}{2} \int d^3r \frac{j_0(\vec{x}) j_0(\vec{x}')}{4\pi |\vec{x} - \vec{x}'|} \quad (6.81)$$

We find ourselves with a nonlocal term in the action. This is exactly the type of interaction that we boasted in Section 1.1.4 never arises in Nature! It appears here as an artifact of working in Coulomb gauge: it does not mean that the theory of QED is nonlocal. For example, it wouldn't appear if we worked in Lorentz gauge.

We now compute the Hamiltonian. Changing notation slightly from previous chapters, we have the conjugate momenta,

$$\begin{aligned} \Pi^\rightarrow &= \frac{\partial \underline{L}}{\partial \dot{\mathbf{A}}} = \mathbf{A}^\rightarrow \\ \pi &= \frac{\partial \underline{L}}{\partial \dot{\psi}} = i\psi^\dagger \\ \psi &= \frac{\partial \underline{L}}{\partial \dot{\psi}} \end{aligned} \quad (6.82)$$

which gives us the Hamiltonian

$$H = \int d^3x \left[\frac{1}{2} \dot{\mathbf{A}}^\rightarrow \cdot \frac{1}{2} \mathbf{B}^\rightarrow + \psi(-i\gamma \partial_i + m)\psi - ej \cdot \mathbf{A}^\rightarrow + \frac{e^2}{2} \int d^3r \frac{j_0(\vec{x}) j_0(\vec{x}')}{4\pi |\vec{x} - \vec{x}'|} \right]$$

where $\vec{j} = \psi^\dagger \vec{\gamma} \psi$ and $j^0 =$

6.4.1 Naive Feynman Rules

We want to determine the Feynman rules for this theory. For fermions, the rules are the same as those given in Section 5. The new pieces are:

- We denote the photon by a wavy line. Each end of the line comes with an $i, j = 1, 2, 3$ index telling us the component of A^{μ} . We calculated the transverse photon propagator in (6.33): it is  and contributes $D^{\text{tr}} = \frac{i}{p_i p_j} \delta^{ij} = \frac{ij}{p^2 + i\epsilon} \delta^{ij} |p|^{-2}$

- The vertex  contributes $-ie\gamma^i$. The index on γ^i contracts with the index on the photon line.
- The non-local interaction which, in position space, is given by  contributes a factor of $\frac{i(e\gamma^0)^2 \delta(x^0 - y^0)}{4\pi|x-y|}$

These Feynman rules are rather messy. This is the price we've paid for working in Coulomb gauge. We'll now show that we can massage these expressions into something much more simple and Lorentz invariant. Let's start with the offending instantaneous interaction. Since it comes from the A_0 component of the gauge field, we could try to redefine the propagator to include a D_{00} piece which will capture this term. In fact, it fits quite nicely in this form: if we look in momentum space, we have

$$\frac{\delta(x^0 - y^0)}{4\pi|x-y|} = \frac{\int d^4p \frac{e^{ip \cdot (x-y)}}{(2\pi)^4}}{|p|^{-2}} \quad (6.83)$$

so we can combine the non-local interaction with the transverse photon propagator by defining a new photon propagator

$$D_{\mu\nu}(p) = \begin{cases} \frac{i}{|p|^2} & \mu, \nu = 0 \\ \frac{\delta^{ij} - p_i p_j}{p^2 + i\epsilon} & \mu = i \neq 0, \nu = j \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (6.84)$$

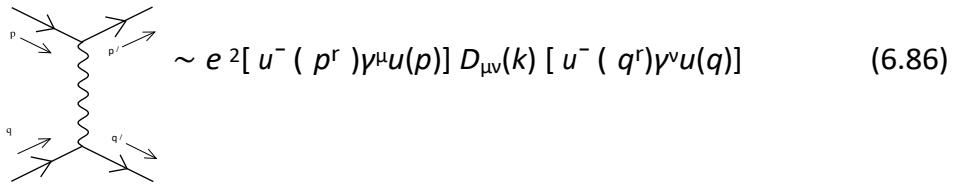
With this propagator, the wavy photon line now carries a $\mu, \nu = 0, 1, 2, 3$ index, with the extra $\mu = 0$ component taking care of the instantaneous interaction. We now need to change our vertex slightly: the $-ie\gamma^i$ above gets replaced by $-ie\gamma^\mu$ which correctly accounts for the $(e\gamma^0)^2$ piece in the instantaneous interaction.

The D_{00} piece of the propagator doesn't look a whole lot different from the transverse photon propagator. But wouldn't it be nice if they were both part of something more symmetric! In fact, they are. We have the following:

Claim: We can replace the propagator $D_{\mu\nu}(p)$ with the simpler, Lorentz invariant propagator

$$D_{\mu\nu}(p) = \frac{i\eta_{\mu\nu}}{p^2} \quad (6.85)$$

Proof: There is a general proof using current conservation. Here we'll be more pedestrian and show that we can do this for certain Feynman diagrams. In particular, we focus on a particular tree-level diagram that contributes to $e^-e^- \rightarrow e^-e^-$ scattering,



where $k = p - p^r = q^r - q$. Recall that $u(p \rightarrow)$ satisfies the equation

$$(p/ - m)u(p \rightarrow) = 0 \quad (6.87)$$

Let's define the spinor contractions $\alpha^\mu = u^- (p \rightarrow)^r \gamma^\mu u(p \rightarrow)$ and $\beta^\nu = u^- (\rightarrow q^r) \gamma^\nu u(\rightarrow q)$. Then since $k = p - p^r = q^r - q$, we have

$$k_\mu \alpha^\mu = u^- (p \rightarrow)^r (p/ - p/) u(p \rightarrow) = u^- (p \rightarrow)^r (m - m) u(\rightarrow p) = 0 \quad (6.88)$$

and, similarly, $k_\nu \beta^\nu = 0$. Using this fact, the diagram can be written as

$$\begin{aligned} i \alpha^\mu D_{\mu\nu} \beta^\nu &= i \frac{\alpha \cdot \beta}{k^2} \frac{(\alpha \cdot k)(\beta \cdot k)}{k^2 + k_0^2 |k|^2} \frac{\alpha^0 \beta^0}{|k|^2} \\ &= i \frac{\alpha \cdot \beta}{k^2} \frac{k^2 + k_0^2 |k|^2}{k^2 + k_0^2 |k|^2} \frac{\alpha^0 \beta^0}{|k|^2} \\ &= i \frac{\alpha \cdot \beta}{k^2} \frac{1}{|k|^2} \frac{(k^2 - k_0^2) \alpha^0 \beta^0}{|k|^2} \\ &= -\frac{i}{k^2} \alpha \cdot \beta = \alpha^\mu - \frac{i \eta_{\mu\nu}}{k^2} \beta^\nu \end{aligned} \quad ! \quad (6.89)$$

which is the claimed result. You can similarly check that the same substitution is legal in the diagram

$$\sim e [v^-(\rightarrow q)]^\mu u(\rightarrow p) D_{\mu\nu}(k) [u^-(\rightarrow p)]^\nu v(\rightarrow q) \quad (6.90)$$

In fact, although we won't show it here, it's a general fact that in every Feynman diagram we may use the very nice, Lorentz invariant propagator $D_{\mu\nu} = -i\eta_{\mu\nu}/p^2$.

Note: This is the propagator we found when quantizing in Lorentz gauge (using the Feynman gauge parameter). In general, quantizing the Lagrangian (6.37) in Lorentz gauge, we have the propagator

$$D_{\mu\nu} = \frac{i}{p} \eta_{\mu\nu} + (\alpha - 1) \frac{p_\mu p_\nu}{p^2} \quad (6.91)$$

Using similar arguments to those given above, you can show that the $p_\mu p_\nu/p^2$ term cancels in all diagrams. For example, in the following diagrams the $p_\mu p_\nu$ piece of the propagator contributes as

$$\sim u^-(p_r) \gamma^\mu u(p) k_\mu = u^-(p_r)(p/ - p_r^r) u(p) = 0$$

$$\sim v^-(p)^\mu u(q) k_\mu = u^-(p)(p_r^r + q/ r) u(q) = 0 \quad (6.92)$$

6.5 Feynman Rules

Finally, we have the Feynman rules for QED. For vertices and internal lines, we write

- Vertex:
- Photon Propagator:
- Fermion Propagator:

For external lines in the diagram, we attach

- Photons: We add a polarization vector $\epsilon_{in}^\mu/\epsilon_{out}^\mu$ for incoming/outgoing photons. In Coulomb gauge, $\epsilon^0 = 0$ and $\epsilon \cdot p = 0$.
- Fermions: We add a spinor $u(p^r)/u^r(p)$ for incoming/outgoing fermions. We add a spinor $v^{-r}(p)/v^r(p)$ for incoming/outgoing anti-fermions.

6.5.1 Charged Scalars

"Pauli asked me to calculate the cross section for pair creation of scalar particles by photons. It was only a short time after Bethe and Heitler had solved the same problem for electrons and positrons. I met Bethe in Copenhagen at a conference and asked him to tell me how he did the calculations. I also inquired how long it would take to perform this task; he answered, "It would take me three days, but you will need about three weeks." He was right, as usual; furthermore, the published cross sections were wrong by a factor of four."

Viki Weisskopf

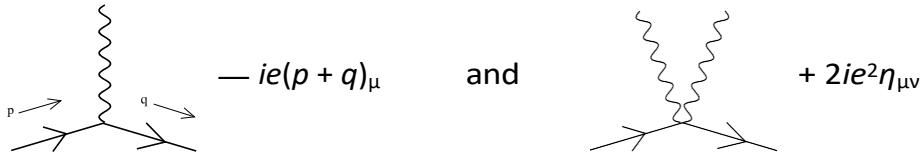
The interaction terms in the Lagrangian for charged scalars come from the covariant derivative terms,

$$L = D_\mu \psi^\dagger D^\mu \psi = \partial_\mu \psi^\dagger \partial^\mu \psi - ie A_\mu (\psi^\dagger \partial^\mu \psi - \psi \partial^\mu \psi^\dagger) + e^2 A_\mu A^\mu \psi^\dagger \psi$$

(6.93) This gives rise to two interaction vertices. But the cubic vertex is something we

haven't

seen before: it contains kinetic terms. How do these appear in the Feynman rules? After a Fourier transform, the derivative term means that the interaction is stronger for fermions with higher momentum, so we include a momentum factor in the Feynman rule. There is also a second, "seagull" graph. The two Feynman rules are



The factor of two in the seagull diagram arises because of the two identical particles appearing in the vertex. (It's the same reason that the $1/4!$ didn't appear in the Feynman rules for φ^4 theory).

6.6 Scattering in QED

Let's now calculate some amplitudes for various processes in quantum electrodynamics, with a photon coupled to a single fermion. We will consider the analogous set of processes that we saw in Section 3 and Section 5. We have

Electron Scattering

Electron scattering $e^-e^- \rightarrow e^-e^-$ is described by the two leading order Feynman diagrams, given by

$$= -i(-ie)^2 \frac{[u^-(p \rightarrow) \gamma u(p \rightarrow)] [u^-(\rightarrow q) \gamma_\mu] (p^r - p)^2}{(p \rightarrow)^2}$$

$$\frac{[u^-(p \rightarrow) \gamma u(q)] \rightarrow [u^-(\rightarrow q^r) \gamma_\mu u^s]}{(p - q^r)^2}$$

The overall $-i$ comes from the $-i\eta_{\mu\nu}$ in the propagator, which contract the indices on the γ -matrices (remember that it's really positive for $\mu, \nu = 1, 2, 3$).

Electron Positron Annihilation

Let's now look at $e^-e^+ \rightarrow 2\gamma$, two gamma rays. The two lowest order Feynman diagrams are,

$$= i(-ie)^2 v^{-r} \frac{\gamma_\mu (p^f - p^{r'} + m) \gamma}{(p - p^{r'})^2 - m^2}$$

$$+ \frac{\gamma (p^f - q^{r'} + m) \gamma}{(p - q^{r'})^2 - m^2} \frac{v^r r^\mu r^r u^s(p \rightarrow) (p \rightarrow) \epsilon_2(\rightarrow q)}{\epsilon_1}$$

Electron Positron Scattering

For $e^-e^+ \rightarrow e^-e^+$ scattering (sometimes known as Bhabha scattering) the two lowest order Feynman diagrams are

$$= -i(-ie)^2 \frac{[u^-(p \rightarrow) \gamma u^s(p \rightarrow)] [v^-(\rightarrow q) \gamma_\mu v(\rightarrow q)] (p^r - p^{r'})^2}{(p + q)^2}$$

$$+ \frac{[v^-(q \rightarrow) \gamma u(p \rightarrow)] [u^-(p \rightarrow) \gamma_\mu v(\rightarrow q)]}{(p + q)^2}$$

Compton Scattering

The scattering of photons (in particular x-rays) off electrons $e^- \gamma \rightarrow e^- \gamma$ is known as Compton scattering. Historically, the change in wavelength of the photon in the

scattering process was one of the conclusive pieces of evidence that light could behave as a particle. The amplitude is given by,

$$= i(-ie) u^r (p')^2 - m^2 + \frac{\gamma_\mu (p/ + q/ + m) \gamma_\nu}{(p+q)^2 - m^2} \frac{\gamma_\nu (p/ - q/ + m) \gamma_{\mu_s}}{(p-q')^2 - m^2} u^s(p') \epsilon_{in}^\mu \epsilon_{out}^\nu$$

This amplitude vanishes for longitudinal photons. For example, suppose $\epsilon_{in} \sim q$. Then, using momentum conservation $p + q = p' + q'$, we may write the amplitude as

$$\begin{aligned} iA_r &= i(-ie)^2 \frac{u^r (p')^2 - m^2}{\epsilon_{out}^\nu (p')^2 - m^2} \frac{(p/ + q/ + m)}{(p+q)^2 - m^2} \frac{q/(p'^r - q/ + m)}{(p^r - q')^2 - m^2} u^s(p') \\ &= i(-ie) u^r (p') \epsilon_{out}^\nu u^s(p') \frac{2p \cdot q}{(p+q)^2 - m^2} \end{aligned} \quad (6.94)$$

where, in going to the second line, we've performed some γ -matrix manipulations, together with the spinor equations $(p/ - m)u(p)$ and $u^r(p) (p'^r - m) = 0$. We now recall the fact that q is a null vector, while $p^2 = (p')^2 = m^2$ since the external legs are on mass-shell. This means that the two denominators in the amplitude read $(p+q)^2 - m^2 = 2p \cdot q$ and $(p^r - q')^2 - m^2 = -2p^r \cdot q$. This ensures that the combined amplitude vanishes for longitudinal photons as promised. A similar result holds when $\epsilon_{out} \sim q^r$.

Photon Scattering

In QED, photons no longer pass through each other unimpeded. At one-loop, there is a diagram which leads to photon scattering. Although naively logarithmically divergent, the diagram is actually rendered finite by gauge invariance.

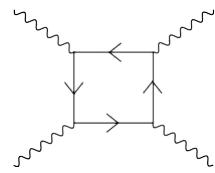


Figure 30:

Adding Muons

Adding a second fermion into the mix, which we could identify as a muon, new processes become possible. For example, we can now have processes such as $e^- \mu^- \rightarrow e^- \mu^-$ scattering, and $e^+ e^-$ annihilation into a muon anti-muon pair. Using our standard notation of p and q for incoming momenta, and p^r and q^r for outgoing

momenta, we have the amplitudes given by

$$\sim \frac{1}{(p - p^r)^2} \quad \text{and} \quad \sim \frac{1}{(p + q)^2} \quad (6.95)$$

6.6.1 The Coulomb Potential

We've come a long way. We've understood how to compute quantum amplitudes in a large array of field theories. To end this course, we use our newfound knowledge to rederive a result you learnt in kindergarten: Coulomb's law.

To do this, we repeat our calculation that led us to the Yukawa force in Sections [3.5.2](#) and [5.7.2](#). We start by looking at $e^-e^- \rightarrow e^-e^-$ scattering. We have

$$= -i(-ie) \frac{2 [u^-(p \rightarrow r) \gamma^\mu u(\rightarrow p)] [u^-(\rightarrow q) \gamma_\mu u(\rightarrow q)]}{(p^r - p)^2} \quad (6.96)$$

Following [\(5.49\)](#), the non-relativistic limit of the spinor is $u(p) \rightarrow \sqrt{\xi} \frac{!}{m \xi}$. This

means that the γ^0 piece of the interaction gives a term $u^- s (p \rightarrow) \gamma^0 u^r (\rightarrow q) \rightarrow 2m\delta^{rs}$, while the spatial γ^i , $i = 1, 2, 3$ pieces vanish in the non-relativistic limit: $u^- s (p \rightarrow) \gamma^i u^r (\rightarrow q) \rightarrow 0$. Comparing the scattering amplitude in this limit to that of non-relativistic quantum mechanics, we have the effective potential between two electrons given by,

$$U(\rightarrow r) = +e^2 \int \frac{d^3 p}{(2\pi)^3} \frac{e^{ip \rightarrow \cdot \rightarrow r}}{|p \rightarrow|^2} = +\frac{e^2}{4\pi r} \quad (6.97)$$

We find the familiar repulsive Coulomb potential. We can trace the minus sign that gives a repulsive potential to the fact that only the A_0 component of the intermediate propagator $\sim -i\eta_{\mu\nu}$ contributes in the non-relativistic limit.

For $e^-e^+ \rightarrow e^-e^+$ scattering, the amplitude is

$$= +i(-ie) \frac{2 [u^-(p \rightarrow r) \gamma^\mu u(\rightarrow p)] [v^-(\rightarrow q) \gamma_\mu v(\rightarrow q)]}{(p^r - p)^2} \quad (6.98)$$

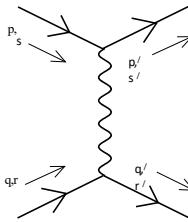
The overall + sign comes from treating the fermions correctly: we saw the same minus sign when studying scattering in Yukawa theory. The difference now comes from looking at the non-relativistic limit. We have $v^- \gamma^0 v \rightarrow 2m$, giving us the potential between opposite charges,

$$U(\rightarrow r) = -e^2 \frac{e^2}{(2\pi)^3 |p^\rightarrow|^2} = -\frac{e^2}{4\pi r} \quad (6.99)$$

Reassuringly, we find an attractive force between an electron and positron. The difference from the calculation of the Yukawa force comes again from the zeroth component of the gauge field, this time in the guise of the γ^0 sandwiched between $v^- \gamma^0 v \rightarrow 2m$, rather than the $v^- v \rightarrow -2m$ that we saw in the Yukawa case.

The Coulomb Potential for Scalars

There are many minus signs in the above calculation which somewhat obscure the crucial one which gives rise to the repulsive force. A careful study reveals the offending sign to be that which sits in front of the A_0 piece of the photon propagator $-in_{\mu\nu}/p^2$. Note that with our signature (+—), the propagating A_i have the correct sign, while A_0 comes with the wrong sign. This is simpler to see in the case of scalar QED, where we don't have to worry about the gamma matrices. From the Feynman rules of Section 6.5.1, we have the non-relativistic limit of scalar $e^- e^-$ scattering,



$$= -i\eta_{\mu\nu}(-ie)^2 \frac{(p + p^r)^\mu (q + q^r)_\nu}{(p^r - p)^2} \rightarrow -i(-ie)^2 \frac{(2m)^2}{-(p^\rightarrow - p^{r\rightarrow})^2}$$

where the non-relativistic limit in the numerator involves $(p+p^r) \cdot (q+q^r) \approx (p+p^r)^0 (q+q^r)_0 \approx (2m)^2$ and is responsible for selecting the A_0 part of the photon propagator rather than the A_i piece. This shows that the Coulomb potential for spin 0 particles of the

same charge is again repulsive, just as it is for fermions. For $e^- e^+$ scattering, the amplitude picks up an extra minus sign because the arrows on the legs of the Feynman rules in Section 6.5.1 are correlated with the momentum arrows. Flipping the arrows on one pair of legs in the amplitude introduces the relevant minus sign to ensure that the non-relativistic potential between $e^- e^+$ is attractive as expected.

6.7 Afterword

In this course, we have laid the foundational framework for quantum field theory. Most of the developments that we've seen were already in place by the middle of the 1930s, pioneered by people such as Jordan, Dirac, Heisenberg, Pauli and Weisskopf⁵.

Yet by the end of the 1930s, physicists were ready to give up on quantum field theory. The difficulty lies in the next terms in perturbation theory. These are the terms that correspond to Feynman diagrams with loops in them, which we have scrupulously avoided computing in this course. The reason we've avoided them is because they typically give infinity! And, after ten years of trying, and failing, to make sense of this, the general feeling was that one should do something else. This from Dirac in 1937,

Because of its extreme complexity, most physicists will be glad to see the end of QED

But the leading figures of the day gave up too soon. It took a new generation of postwar physicists — people like Schwinger, Feynman, Tomonaga and Dyson — to return to quantum field theory and tame the infinities. The story of how they did that will be told in next term's course.

⁵For more details on the history of quantum field theory, see the excellent book “QED and the Men who Made it” by Sam Schweber.

Chapter 2

Basic Set Theory

A set is a Many that allows itself to be thought of as a One.

- Georg Cantor

This chapter introduces set theory, mathematical induction, and formalizes the notion of mathematical functions. The material is mostly elementary. For those of you new to abstract mathematics elementary does not mean *simple* (though much of the material is fairly simple). Rather, elementary means that the material requires very little previous education to understand it. Elementary material can be quite challenging and some of the material in this chapter, if not exactly rocket science, may require that you adjust your point of view to understand it. The single most powerful technique in mathematics is to adjust your point of view until the problem you are trying to solve becomes simple.

Another point at which this material may diverge from your previous experience is that it will require proof. In standard introductory classes in algebra, trigonometry, and calculus there is currently very little emphasis on the discipline of *proof*. Proof is, however, the central tool of mathematics. This text is for a course that is a students formal introduction to tools and methods of proof.

2.1 Set Theory

A *set* is a collection of distinct objects. This means that $\{1, 2, 3\}$ is a set but $\{1, 1, 3\}$ is not because 1 appears twice in the second collection. The second collection is called a *multiset*. Sets are often specified with curly brace notation. The set of even integers

can be written:

$$\{2n : n \text{ is an integer}\}$$

The opening and closing curly braces denote a set, $2n$ specifies the members of the set, the colon says “such that” or “where” and everything following the colon are conditions that explain or refine the membership. All correct mathematics can be spoken in English. The set definition above is spoken “The set of twice n where n is an integer”.

The only problem with this definition is that we do not yet have a formal definition of the integers. The integers are the set of whole numbers, both positive and negative: $\{0, \pm 1, \pm 2, \pm 3, \dots\}$. We now introduce the operations used to manipulate sets, using the opportunity to practice curly brace notation.

Definition 2.1 *The empty set is a set containing no objects. It is written as a pair of curly braces with nothing inside {} or by using the symbol \emptyset .*

As we shall see, the empty set is a handy object. It is also quite strange. The set of all humans that weigh at least eight tons, for example, is the empty set. Sets whose definition contains a contradiction or impossibility are often empty.

Definition 2.2 *The set membership symbol \in is used to say that an object is a member of a set. It has a partner symbol \notin which is used to say an object is not in a set.*

Definition 2.3 *We say two sets are equal if they have exactly the same members.*

Example 2.1 If

$$S = \{1, 2, 3\}$$

then $3 \in S$ and $4 \notin S$. The set membership symbol is often used in defining operations that manipulate sets. The set

$$T = \{2, 3, 1\}$$

is equal to S because they have the same members: 1, 2, and 3. While we usually list the members of a set in a “standard” order (if one is available) there is no requirement to do so and sets are indifferent to the order in which their members are listed.

Definition 2.4 The cardinality of a set is its size. For a finite set, the cardinality of a set is the number of members it contains. In symbolic notation the size of a set S is written $|S|$. We will deal with the idea of the cardinality of an infinite set later.

Example 2.2 Set cardinality

For the set $S = \{1, 2, 3\}$ we show cardinality by writing $|S| = 3$

We now move on to a number of *operations* on sets. You are already familiar with several operations on numbers such as addition, multiplication, and negation.

Definition 2.5 The intersection of two sets S and T is the collection of all objects that are in both sets. It is written $S \cap T$. Using curly brace notation

$$S \cap T = \{x : (x \in S) \text{ and } (x \in T)\}$$

The symbol *and* in the above definition is an example of a Boolean or logical operation. It is only true when both the propositions it joins are also true. It has a symbolic equivalent \wedge . This lets us write the formal definition of intersection more compactly:

$$S \cap T = \{x : (x \in S) \wedge (x \in T)\}$$

Example 2.3 Intersections of sets

Suppose $S = \{1, 2, 3, 5\}$,
 $T = \{1, 3, 4, 5\}$, and $U = \{2, 3, 4, 5\}$.
Then:

$$S \cap T = \{1, 3, 5\},$$

$$S \cap U = \{2, 3, 5\}, \text{ and}$$

$$T \cap U = \{3, 4, 5\}$$

Definition 2.6 If A and B are sets and $A \cap B = \emptyset$ then we say that A and B are disjoint, or disjoint sets.

Definition 2.7 The union of two sets S and T is the collection of all objects that are in either set. It is written $S \cup T$. Using curly brace notion

$$S \cup T = \{x : (x \in S) \text{ or } (x \in T)\}$$

The symbol *or* is another Boolean operation, one that is true if either of the propositions it joins are true. Its symbolic equivalent is \vee which lets us re-write the definition of union as:

$$S \cup T = \{x : (x \in S) \vee (x \in T)\}$$

Example 2.4 Unions of sets.

Suppose $S = \{1, 2, 3\}$, $T = \{1, 3, 5\}$, and $U = \{2, 3, 4, 5\}$.

Then:

$$S \cup T = \{1, 2, 3, 5\},$$

$$S \cup U = \{1, 2, 3, 4, 5\}, \text{ and}$$

$$T \cup U = \{1, 2, 3, 4, 5\}$$

When performing set theoretic computations, you should declare the domain in which you are working. In set theory this is done by declaring a universal set.

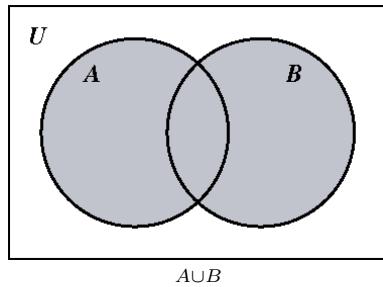
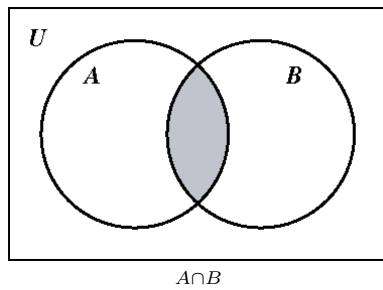
Definition 2.8 The universal set, at least for a given collection of set theoretic computations, is the set of all possible objects.

If we declare our universal set to be the integers then $\{\frac{1}{2}, \frac{2}{3}\}$ is not a well defined set because the objects used to define it are not members of the universal set. The symbols $\{\frac{1}{2}, \frac{2}{3}\}$ do define a set if a universal set that includes $\frac{1}{2}$ and $\frac{2}{3}$ is chosen. The problem arises from the fact that neither of these numbers are integers. The universal set is commonly written \mathcal{U} . Now that we have the idea of declaring a universal set we can define another operation on sets.

2.1.1 Venn Diagrams

A Venn diagram is a way of depicting the relationship between sets. Each set is shown as a circle and circles overlap if the sets intersect.

Example 2.5 The following are Venn diagrams for the intersection and union of two sets. The shaded parts of the diagrams are the intersections and unions respectively.



Notice that the rectangle containing the diagram is labeled with a U representing the universal set.

Definition 2.9 The **compliment** of a set S is the collection of objects in the universal set that are not in S . The compliment is written S^c . In curly brace notation

$$S^c = \{x : (x \in \mathcal{U}) \wedge (x \notin S)\}$$

or more compactly as

$$S^c = \{x : x \notin S\}$$

however it should be apparent that the compliment of a set always depends on which universal set is chosen.

There is also a Boolean symbol associated with the complementation operation: the *not* operation. The

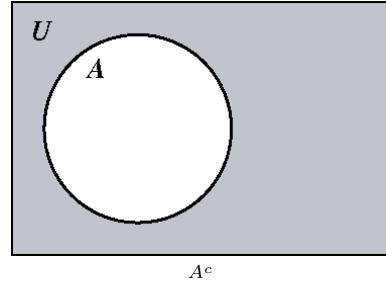
notation for not is \neg . There is not much savings in space as the definition of compliment becomes

$$S^c = \{x : \neg(x \in S)\}$$

Example 2.6 Set Compliments

- (i) Let the universal set be the integers. Then the compliment of the even integers is the odd integers.
- (ii) Let the universal set be $\{1, 2, 3, 4, 5\}$, then the compliment of $S = \{1, 2, 3\}$ is $S^c = \{4, 5\}$ while the compliment of $T = \{1, 3, 5\}$ is $T^c = \{2, 4\}$.
- (iii) Let the universal set be the letters $\{a, e, i, o, u, y\}$. Then $\{y\}^c = \{a, e, i, o, u\}$.

The Venn diagram for A^c is



We now have enough set-theory operators to use them to define more operators quickly. We will continue to give English and symbolic definitions.

Definition 2.10 The **difference** of two sets S and T is the collection of objects in S that are not in T . The difference is written $S - T$. In curly brace notation

$$S - T = \{x : x \in (S \cap (T^c))\},$$

or alternately

$$S - T = \{x : (x \in S) \wedge (x \notin T)\}$$

Notice how intersection and complementation can be used together to create the difference operation and that the definition can be rephrased to use Boolean operations. There is a set of rules that reduces the number of parenthesis required. These are called **operator precedence rules**.

- (i) Other things being equal, operations are performed left-to-right.
- (ii) Operations between parenthesis are done first, starting with the innermost of nested parenthesis.
- (iii) All compliments are computed next.
- (iv) All intersections are done next.
- (v) All unions are performed next.
- (vi) Tests of set membership and computations, equality or inequality are performed last.

Special operations like the set difference or the symmetric difference, defined below, are not included in the precedence rules and thus always use parenthesis.

Example 2.7 Operator precedence

Since complementation is done before intersection the symbolic definition of the difference of sets can be rewritten:

$$S - T = \{x : x \in S \cap T^c\}$$

If we were to take the set operations

$$A \cup B \cap C^c$$

and put in the parenthesis we would get

$$(A \cup (B \cap (C^c)))$$

Definition 2.11 *The symmetric difference of two sets S and T is the set of objects that are in one and only one of the sets. The symmetric difference is written $S \Delta T$. In curly brace notation:*

$$S \Delta T = \{(S - T) \cup (T - S)\}$$

Example 2.8 Symmetric differences

Let S be the set of non-negative multiples of two that are no more than twenty four. Let T be the non-negative multiples of three that are no more than twenty four. Then

$$S \Delta T = \{2, 3, 4, 8, 9, 10, 14, 15, 16, 20, 21, 22\}$$

Another way to think about this is that we need numbers that are positive multiples of 2 or 3 (but not both) that are no more than 24.

Another important tool for working with sets is the ability to compare them. We have already defined what it means for two sets to be equal, and so by implication what it means for them to be unequal. We now define another comparator for sets.

Definition 2.12 *For two sets S and T we say that S is a subset of T if each element of S is also an element of T. In formal notation $S \subseteq T$ if for all $x \in S$ we have $x \in T$.*

If $S \subseteq T$ then we also say T contains S which can be written $T \supseteq S$. If $S \subseteq T$ and $S \neq T$ then we write $S \subset T$ and we say S is a *proper* subset of T.

Example 2.9 Subsets

If $A = \{a, b, c\}$ then A has eight different subsets:

\emptyset	$\{a\}$	$\{b\}$	$\{c\}$
$\{a, b\}$	$\{a, c\}$	$\{b, c\}$	$\{a, b, c\}$

Notice that $A \subseteq A$ and in fact each set is a subset of itself. The empty set \emptyset is a subset of every set.

We are now ready to prove our first proposition. Some new notation is required and we must introduce an important piece of mathematical culture. If we say “A if and only if B” then we mean that either A and B are both true or they are both false in any given circumstance. For example: “an integer x is even if and only if it is a multiple of 2”. The phrase “if and only if” is used to establish *logical equivalence*. Mathematically, “A if and only if B” is a way of stating that A and B are simply different ways of saying the same thing. The phrase “if and only if” is abbreviated iff and is represented symbolically as the double arrow \Leftrightarrow . Proving an iff statement is done by independently demonstrating that each may be deduced from the other.

Proposition 2.1 *Two sets are equal if and only if each is a subset of the other. In symbolic notation:*

$$(A = B) \Leftrightarrow (A \subseteq B) \wedge (B \subseteq A)$$

Proof:

Let the two sets in question be A and B. Begin by assuming that $A = B$. We know that every set is

a subset of itself so $A \subseteq A$. Since $A = B$ we may substitute into this expression on the left and obtain $B \subseteq A$. Similarly we may substitute on the right and obtain $A \subseteq B$. We have thus demonstrated that if $A = B$ then A and B are both subsets of each other, giving us the first half of the iff.

Assume now that $A \subseteq B$ and $B \subseteq A$. Then the definition of subset tells us that any element of A is an element of B . Similarly any element of B is an element of A . This means that A and B have the same elements which satisfies the definition of set equality. We deduce $A = B$ and we have the second half of the iff. \square

A note on mathematical grammar: the symbol \square indicates the end of a proof. On a paper turned in by a student it is usually taken to mean “I think the proof ends here”. Any proof should have a \square to indicate its end. The student should also note the lack of calculations in the above proof. If a proof cannot be read back in (sometimes overly formal) English then it is probably incorrect. Mathematical symbols should be used for the sake of brevity or clarity, not to obscure meaning.

Proposition 2.2 De Morgan’s Laws Suppose that S and T are sets. DeMorgan’s Laws state that

- (i) $(S \cup T)^c = S^c \cap T^c$, and
- (ii) $(S \cap T)^c = S^c \cup T^c$.

Proof:

Let $x \in (S \cup T)^c$; then x is not a member of S or T . Since x is not a member of S we see that $x \in S^c$. Similarly $x \in T^c$. Since x is a member of both these sets we see that $x \in S^c \cap T^c$ and we see that $(S \cup T)^c \subseteq S^c \cap T^c$. Let $y \in S^c \cap T^c$. Then the definition of intersection tells us that $y \in S^c$ and $y \in T^c$. This in turn lets us deduce that y is not a member of $S \cup T$, since it is not in either set, and so we see that $y \in (S \cup T)^c$. This demonstrates that $S^c \cap T^c \subseteq (S \cup T)^c$. Applying Proposition 2.1 we get that $(S \cup T)^c = S^c \cap T^c$ and we have proven part (i). The proof of part (ii) is left as an exercise. \square

In order to prove a mathematical statement you must prove it is always true. In order to disprove a mathematical statement you need only find a single instance

where it is false. It is thus possible for a false mathematical statement to be “true most of the time”. In the next chapter we will develop the theory of prime numbers. For now we will assume the reader has a modest familiarity with the primes. The statement “Prime numbers are odd” is false once, because 2 is a prime number. All the other prime numbers are odd. The statement is a false one. This very strict definition of what makes a statement true is a convention in mathematics. We call 2 a *counter example*. It is thus necessary to find only one counter-example to demonstrate a statement is (mathematically) false.

Example 2.10 Disproof by counter example

Prove that the statement $A \cup B = A \cap B$ is false.

Let $A = \{1, 2\}$ and $B = \{3, 4\}$. Then $A \cap B = \emptyset$ while $A \cup B = \{1, 2, 3, 4\}$. The sets A and B form a counter-example to the statement.

Problems

Problem 2.1 Which of the following are sets? Assume that a proper universal set has been chosen and answer by listing the names of the collections of objects that are sets. Warning: at least one of these items has an answer that, while likely, is not 100% certain.

- (i) $A = \{2, 3, 5, 7, 11, 13, 19\}$
- (ii) $B = \{A, E, I, O, U\}$
- (iii) $C = \{\sqrt{x} : x < 0\}$
- (iv) $D = \{1, 2, A, 5, B, Q, 1, V\}$
- (v) E is the list of first names of people in the 1972 phone book in Lawrence Kansas in the order they appear in the book. There were more than 35,000 people in Lawrence that year.
- (vi) F is a list of the weight, to the nearest kilogram, of all people that were in Canada at any time in 2007.
- (vii) G is a list of all weights, to the nearest kilogram, that at least one person in Canada had in 2007.

Problem 2.2 Suppose that we have the set $U = \{n : 0 \leq n < 100\}$ of whole numbers as our universal set. Let P be the prime numbers in U , let E be the even numbers in U , and let $F = \{1, 2, 3, 5, 8, 13, 21, 34, 55, 89\}$. Describe the following sets either by listing them or with a careful English sentence.

- (i) E^c ,
- (ii) $P \cap F$,
- (iii) $P \cap E$,
- (iv) $F \cap E \cup F \cap E^c$, and
- (v) $F \cup F^c$.

Problem 2.3 Suppose that we take the universal set U to be the integers. Let S be the even integers, let T be the integers that can be obtained by tripling any one integer and adding one to it, and let V be the set of numbers that are whole multiples of both two and three.

- (i) Write S , T , and V using symbolic notation.
- (ii) Compute $S \cap T$, $S \cap V$ and $T \cap V$ and give symbolic representations that do not use the symbols S , T , or V on the right hand side of the equals sign.

Problem 2.4 Compute the cardinality of the following sets. You may use other texts or the internet.

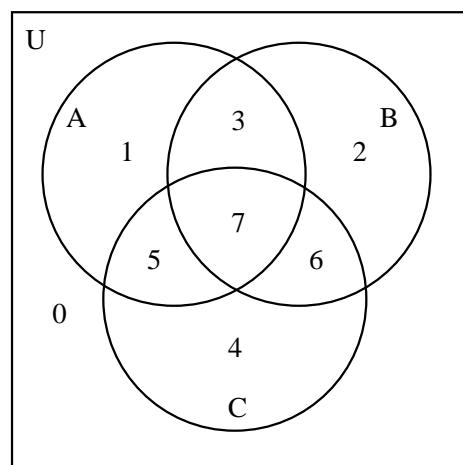
- (i) Two digit positive odd integers.
- (ii) Elements present in a sucrose molecule.
- (iii) Isotopes of hydrogen that are not radioactive.
- (iv) Planets orbiting the same star as the planet you are standing on that have moons. Assume that Pluto is a minor planet.
- (v) Elements with seven electrons in their valence shell. Remember that Ununoctium was discovered in 2002 so be sure to use a relatively recent reference.
- (vi) Subsets of $S = \{a, b, c, d\}$ with cardinality 2.
- (vii) Prime numbers whose base-ten digits sum to ten. Be careful, some have three digits.

Problem 2.5 Find an example of an infinite set that has a finite complement, be sure to state the universal set.

Problem 2.6 Find an example of an infinite set that has an infinite complement, be sure to state the universal set.

Problem 2.7 Add parenthesis to each of the following expressions that enforce the operator precedence rules as in Example 2.7. Notice that the first three describe sets while the last returns a logical value (true or false).

- (i) $A \cup B \cup C \cup D$
 - (ii) $A \cup B \cap C \cup D$
 - (iii) $A^c \cap B^c \cup C$
 - (iv) $A \cup B = A \cap C$
- Problem 2.8** Give the Venn diagrams for the following sets.
- (i) $A - B$ (ii) $B - A$ (iii) $A^c \cap B$
 - (iv) $A \Delta B$ (v) $(A \Delta B)^c$ (vi) $A^c \cup B^c$



Problem 2.9 Examine the Venn diagram above. Notice that every combination of sets has a unique number in common. Construct a similar collection of four sets.

Problem 2.10 Read Problem 2.9. Can a system of sets of this sort be constructed for any number of sets? Explain your reasoning.

Problem 2.11 Suppose we take the universal set to be the set of non-negative integers. Let E be the set of even numbers, O be the set of odd numbers and $F = \{0, 1, 2, 3, 5, 8, 13, 21, 34, 89, 144, \dots\}$ be the set of Fibonacci numbers. The Fibonacci sequence is $0, 1, 1, 2, 3, 5, 8, \dots$ in which the next term is obtained by adding the previous two.

- (i) Prove that the intersection of F with E and O are both infinite.
- (ii) Make a Venn diagram for the sets E , F , and O , and explain why this is a Mickey-Mouse problem.

Problem 2.12 A binary operation \odot is commutative if $x \odot y = y \odot x$. An example of a commutative operation is multiplication. Subtraction is non-commutative. Determine, with proof, if union, intersection, set difference, and symmetric difference are commutative.

Problem 2.13 An identity for an operation \odot is an object i so that, for all objects x , $i \odot x = x \odot i = x$. Find, with proof, identities for the operations set union and set intersection.

Problem 2.14 Prove part (ii) of Proposition 2.2.

Problem 2.15 Prove that

$$A \cup (B \cup C) = (A \cup B) \cup C$$

Problem 2.16 Prove that

$$A \cap (B \cap C) = (A \cap B) \cap C$$

Problem 2.17 Prove that

$$A \Delta (B \Delta C) = (A \Delta B) \Delta C$$

Problem 2.18 Disprove that

$$A \Delta (B \cup C) = (A \Delta B) \cup C$$

Problem 2.19 Consider the set $S = \{1, 2, 3, 4\}$. For each $k = 0, 1, \dots, 4$ how many k element subsets does S have?

Problem 2.20 Suppose we have a set S with $n \geq 0$ elements. Find a formula for the number of different subsets of S that have k elements.

Problem 2.21 For finite sets S and T , prove

$$|S \cup T| = |S| + |T| - |S \cap T|$$

2.2 Mathematical Induction

Mathematical induction is a technique used in proving mathematical assertions. The basic idea of induction is that we prove that a statement is true in one case and then also prove that if it is true in a given case it is true in the next case. This then permits the cases for which the statement is true to cascade from the initial true case. We will start with the mathematical foundations of induction.

We assume that the reader is familiar with the symbols $<$, $>$, \leq and \geq . From this point on we will denote the set of integers by the symbol \mathbb{Z} . The non-negative integers are called the *natural numbers*. The symbol for the set of natural numbers is \mathbb{N} . Any mathematical system rests on a foundation of axioms. Axioms are things that we simply assume to be true. We will assume the truth of the following principle, adopting it as an axiom.

The well-ordering principle: Every non-empty set of natural numbers contains a smallest element.

The well ordering principle is an axiom that agrees with the common sense of most people familiar with the natural numbers. An empty set does not contain a smallest member because it contains no members at all. As soon as we have a set of natural numbers with some members then we can order those members in the usual fashion. Having ordered them, one will be smallest. This intuition agreeing with this latter claim depends strongly on the fact the integers are “whole numbers” spaced out in increments of one. To see why this is important consider the smallest positive distance. If such a distance existed, we could cut it in half to obtain a smaller distance - the quantity contradicts its own existence. The well-ordering principle can be used to prove the correctness of induction.

Theorem 2.1 Mathematical Induction I Suppose that $P(n)$ is a proposition that it either true or false for any given natural numbers n . If

(i) $P(0)$ is true and,

(ii) when $P(n)$ is true so is $P(n+1)$

Then we may deduce that $P(n)$ is true for any natural number.

Proof:

Assume that (i) and (ii) are both true statements. Let S be the set of all natural numbers for which $P(n)$ is false. If S is empty then we are done, so assume that S is not empty. Then, by the well ordering principle, S has a least member m . By (i) above $m \neq 0$ and so $m - 1$ is a natural number. Since m is the smallest member of S it follows that $P(m-1)$ is true. But this means, by (ii) above, that $P(m)$ is true. We have a contradiction and so our assumption that $S \neq \emptyset$ must be wrong. We deduce S is empty and that as a consequence $P(n)$ is true for all $n \in \mathbb{N}$. \square

The technique used in the above proof is called *proof by contradiction*. We start by assuming the logical opposite of what we want to prove, in this case that there is some m for which $P(m)$ is false, and from that assumption we derive an impossibility. If an assumption can be used to demonstrate an impossibility then it is false and its logical opposite is true.

A nice problem on which to demonstrate mathematical induction is counting how many subsets a finite set has.

Proposition 2.3 **Subset counting.** A set S with n elements has 2^n subsets.

Proof:

First we check that the proposition is true when $n = 0$. The empty set has exactly one subset: itself. Since $2^0 = 1$ the proposition is true for $n = 0$. We now assume the proposition is true for some n . Suppose that S is a set with $n + 1$ members and that $x \in S$. Then $S - \{x\}$ (the set difference of S and a set $\{x\}$ containing only x) is a set of n elements and so, by the assumption, has 2^n subsets. Now every subset of S either contains x or it fails to. Every subset of S that does not contain x is a subset of $S - \{x\}$ and so there are 2^n such subsets of S . Every subset of S that contains x may be obtained in exactly one way from one that does not by taking the union with $\{x\}$. This means that the number of subsets of S containing or failing to contain x are equal. This means there are 2^n subsets of S containing x . The total number of subsets of S is thus $2^n + 2^n = 2^{n+1}$. So if we assume the proposition is true for n we can demonstrate that it is also true for $n + 1$. It follows by mathematical

induction that the proposition is true for all natural numbers. \square

The set of all subsets of a given set is itself an important object and so has a name.

Definition 2.13 The set of all subsets of a set S is called the **powerset** of S . The notation for the powerset of S is $\mathcal{P}(S)$.

This definition permits us to rephrase Proposition 2.3 as follows: the power set of a set of n elements has size 2^n .

Theorem 2.1 lets us prove propositions that are true on the natural numbers, starting at zero. A small modification of induction can be used to prove statements that are true only for those $n \geq k$ for any integer k . All that is needed is to use induction on a proposition $Q(n - k)$ where $Q(n - k)$ is logically equivalent to $P(n)$. If $Q(n - k)$ is true for $n - k \geq 0$ then $P(n)$ is true for $n \geq k$ and we have the modified induction. The practical difference is that we start with k instead of zero.

Example 2.11 Prove that $n^2 \geq 2n$ for all $n \geq 2$.

Notice that $2^2 = 4 = 2 \times 2$ so the proposition is true when $n = 2$. We next assume that $P(n)$ is true for some n and we compute:

$$\begin{aligned} n^2 &\geq 2n \\ n^2 + 2n + 1 &\geq 2n + 2n + 1 \\ (n+1)^2 &\geq 2n + 2n + 1 \\ (n+1)^2 &\geq 2n + 1 + 1 \\ (n+1)^2 &\geq 2n + 2 \\ (n+1)^2 &\geq 2(n+1) \end{aligned}$$

To move from the third step to the fourth step we use the fact that $2n > 1$ when $n \geq 2$. The last step is $P(n+1)$, which means we have deduced $P(n+1)$ from $P(n)$. Using the modified form of induction we have proved that $n^2 \geq 2n$ for all $n \geq 2$.

It is possible to formalize the procedure for using mathematical induction into a three-part process. Once we have a proposition $P(n)$,

- (i) First demonstrate a *base case* by directly demonstrating $P(k)$,
- (ii) Next make the *induction hypothesis* that $P(n)$ is true for some n ,
- (iii) Finally, starting with the assumption that $P(n)$ is true, demonstrate $P(n+1)$.

These steps permit us to deduce that $P(n)$ is true for all $n \geq k$.

Example 2.12 Using induction, prove

$$1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$$

In this case $P(n)$ is the statement

$$1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$$

Base case: $1 = \frac{1}{2}1(1+1)$, so $P(1)$ is true. **Induction hypothesis:** for some n ,

$$1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$$

Compute:

$$\begin{aligned} 1 + 2 + \cdots + (n+1) &= 1 + 2 + \cdots + n + (n+1) \\ &= \frac{1}{2}n(n+1) + (n+1) \\ &= \frac{1}{2}(n(n+1) + 2(n+1)) \\ &= \frac{1}{2}(n^2 + 3n + 2) \\ &= \frac{1}{2}(n+1)(n+2) \\ &= \frac{1}{2}(n+1)((n+1)+1) \end{aligned}$$

and so we have shown that if $P(n)$ is true then so is $P(n+1)$. We have thus proven that $P(n)$ is true for all $n \geq 1$ by mathematical induction.

We now introduce *sigma notation* which makes problems like the one worked in Example 2.12 easier to state and manipulate. The symbol \sum is used to add

up lists of numbers. If we wished to sum some formula $f(i)$ over a range from a to b , that is to say $a \leq i \leq b$, then we write :

$$\sum_{i=a}^b f(i)$$

On the other hand if S is a set of numbers and we want to add up $f(s)$ for all $s \in S$ we write:

$$\sum_{s \in S} f(s)$$

The result proved in Example 2.12 may be stated in the following form using sigma notation.

$$\sum_{i=1}^n i = \frac{1}{2}n(n+1)$$

Proposition 2.4 Suppose that c is a constant and that $f(i)$ and $g(i)$ are formulas. Then

- (i) $\sum_{i=a}^b (f(i) + g(i)) = \sum_{i=a}^b f(i) + \sum_{i=a}^b g(i)$
- (ii) $\sum_{i=a}^b (f(i) - g(i)) = \sum_{i=a}^b f(i) - \sum_{i=a}^b g(i)$
- (iii) $\sum_{i=a}^b c \cdot f(i) = c \cdot \sum_{i=a}^b f(i)$.

Proof:

Part (i) and (ii) are both simply the associative law for addition: $a + (b+c) = (a+b)+c$ applied many times. Part (iii) is a similar multiple application of the distributive law $ca + cb = c(a+b)$. \square

The sigma notation lets us work with indefinitely long (and even infinite) sums. There are other similar notations. If A_1, A_2, \dots, A_n are sets then the intersection or union of all these sets can be written:

$$\begin{aligned} \bigcap_{i=1}^n A_i \\ \bigcup_{i=1}^n A_i \end{aligned}$$

Similarly if $f(i)$ is a formula on the integers then

$$\prod_{i=1}^n f(i)$$

is the notation for computing the product $f(1) \cdot f(2) \cdot \dots \cdot f(n)$. This notation is called **Pi** notation.

Definition 2.14 When we solve an expression involving \sum to obtain a formula that does not use \sum or "... " as in Example 2.12 then we say we have found a **closed form** for the expression. Example 2.12 finds a closed form for $\sum_{i=1}^n i$.

At this point we introduce a famous mathematical sequence in order to create an arena for practicing proofs by induction.

Definition 2.15 The **Fibonacci numbers** are defined as follows. $f_1 = f_2 = 1$ and, for $n \geq 3$, $f_n = f_{n-1} + f_{n-2}$.

Example 2.13 The Fibonacci numbers with four or fewer digits are: $f_1 = 1$, $f_2 = 1$, $f_3 = 2$, $f_4 = 3$, $f_5 = 5$, $f_6 = 8$, $f_7 = 13$, $f_8 = 21$, $f_9 = 34$, $f_{10} = 55$, $f_{11} = 89$, $f_{12} = 144$, $f_{13} = 233$, $f_{14} = 377$, $f_{15} = 610$, $f_{16} = 987$, $f_{17} = 1597$, $f_{18} = 2584$, $f_{19} = 4181$, and $f_{20} = 6765$.

Example 2.14 Prove that the Fibonacci number f_{3n} is even.

Solution:

Notice that $f_3 = 2$ and so the proposition is true when $n = 1$. Assume that the proposition is true for some $n \geq 1$. Then:

$$f_{3(n+1)} = f_{3n+3} \quad (2.1)$$

$$= f_{3n+2} + f_{3n+1} \quad (2.2)$$

$$= f_{3n+1} + f_{3n} + f_{3n+1} \quad (2.3)$$

$$= 2 \cdot f_{3n+1} + f_{3n} \quad (2.4)$$

but this suffices because f_{3n} is even by the induction hypothesis while $2 \cdot f_{3n+1}$ is also even. The sum is thus even and so $f_{3(n+1)}$ is even. It follows by induction that f_{3n} is even for all n . \square

Problems

Problem 2.22 Suppose that $S = \{a, b, c\}$. Compute and list explicitly the members of the powerset, $\mathcal{P}(S)$.

Problem 2.23 Prove that for a finite set X that

$$|X| \leq |\mathcal{P}(X)|$$

Problem 2.24 Suppose that $X \subseteq Y$ with $|Y| = n$ and $|X| = m$. Compute the number of subsets of Y that contain X .

Problem 2.25 Compute the following sums.

$$(i) \sum_{i=1}^{20} i,$$

$$(ii) \sum_{i=10}^{30} i, \text{ and}$$

$$(iii) \sum_{i=-20}^{21} i.$$

Problem 2.26 Using mathematical induction, prove the following formulas.

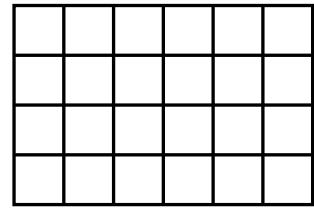
$$(i) \sum_{i=1}^n 1 = n,$$

$$(ii) \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}, \text{ and}$$

$$(iii) \sum_{i=1}^n i^3 = \frac{n^2(n+1)^2}{4}.$$

Problem 2.27 If $f(i)$ and $g(i)$ are formulas and c and d are constants prove that

$$\sum_{i=a}^b (c \cdot f(i) + d \cdot g(i)) = c \cdot \sum_{i=a}^b f(i) + d \cdot \sum_{i=a}^b g(i)$$



Problem 2.28 Suppose you want to break an $n \times m$ chocolate bar, like the 6×4 example shown above, into pieces corresponding to the small squares shown. What is the minimum number of breaks you can make? Prove your answer is correct.

Problem 2.29 Prove by induction that the sum of the first n odd numbers equals n^2 .

Problem 2.30 Compute the sum of the first n positive even numbers.

Problem 2.31 Find a closed form for

$$\sum_{i=1}^n i^2 + 3i + 5$$

Problem 2.32 Let $f(n, 3)$ be the number of subsets of $\{1, 2, \dots, n\}$ of size 3. Using induction, prove that $f(n, 3) = \frac{1}{6}n(n-1)(n-2)$.

Problem 2.33 Suppose that we have sets X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n such that $X_i \subseteq Y_i$. Prove that the intersection of all the X_i is a subset of the intersection of all the Y_i :

$$\bigcap_{i=1}^n X_i \subseteq \bigcap_{i=1}^n Y_i$$

Problem 2.34 Suppose that S_1, S_2, \dots, S_n are sets. Prove the following generalization of DeMorgan's laws:

$$(i) (\bigcap_{i=1}^n S_i)^c = \bigcup_{i=1}^n S_i^c, \text{ and}$$

$$(ii) (\bigcup_{i=1}^n S_i)^c = \bigcap_{i=1}^n S_i^c.$$

Problem 2.35 Prove by induction that the Fibonacci number f_{4n} is a multiple of 3.

Problem 2.36 Prove that if r is a real number $r \neq 1$ and $r \neq 0$ then

$$\sum_{i=0}^n r^i = \frac{1 - r^{n+1}}{1 - r}$$

Problem 2.37 Prove by induction that the Fibonacci number f_{5n} is a multiple of 5.

Problem 2.38 Prove by induction that the Fibonacci number f_n has the value

$$f_n = \frac{\sqrt{5}}{5} \cdot \left(\frac{1 + \sqrt{5}}{2} \right)^n - \frac{\sqrt{5}}{5} \cdot \left(\frac{1 - \sqrt{5}}{2} \right)^n$$

Problem 2.39 Prove that for sufficiently large n the Fibonacci number f_n is the integer closest to

$$\frac{\sqrt{5}}{5} \left(\frac{1 + \sqrt{5}}{2} \right)^n$$

and compute the exact value of f_{30} . Show your work (i.e. don't look the result up on the net).

Problem 2.40 Prove that $\frac{n(n-1)(n-2)(n-3)}{24}$ is a whole number for any whole number n .

Problem 2.41 Consider the statement "All cars are the same color." and the following "proof".

Proof:

We will prove for $n \geq 1$ that for any set of n cars all the cars in the set have the same color.

- *Base Case:* $n=1$ If there is only one car then clearly there is only one color the car can be.
- *Inductive Hypothesis:* Assume that for any set of n cars there is only one color.
- *Inductive step:* Look at any set of $n + 1$ cars. Number them: 1, 2, 3, ..., $n, n + 1$. Consider the sets $\{1, 2, 3, \dots, n\}$ and $\{2, 3, 4, \dots, n + 1\}$. Each is a set of only n cars, therefore for each set there is only one color. But the n^{th} car is in both sets so the color of the cars in the first set must be the same as the color of the cars in the second set. Therefore there must be only one color among all $n + 1$ cars.
- The proof follows by induction. \square

What are the problems with this proof?

2.3 Functions

In this section we will define functions and extend much of our ability to work with sets to infinite sets. There are a number of different types of functions and so this section contains a great deal of terminology.

Recall that two finite sets are the same size if they contain the same number of elements. It is possible to make this idea formal by using functions and, once the notion is formally defined, it can be applied to infinite sets.

Definition 2.16 An ordered pair is a collection of two elements with the added property that one element comes first and one element comes second. The set containing only x and y (for $x \neq y$) is written $\{x, y\}$. The ordered pair containing x and y with x first is written (x, y) . Notice that while $\{x, x\}$ is not a well defined set, (x, x) is a well defined ordered pair because the two copies of x are different by virtue of coming first and second.

The reason for defining ordered pairs at this point is that it permits us to make an important formal definition that pervades the rest of mathematics.

Definition 2.17 A function f with domain S and range T is a set of ordered pairs (s, t) with first element from S and second element from T that has the property that every element of S appears exactly once as the first element in some ordered pair. We write $f : S \rightarrow T$ for such a function.

Example 2.15 Suppose that $A = \{a, b, c\}$ and $B = \{0, 1\}$ then

$$f = \{(a, 0), (b, 1), (c, 0)\}$$

is a function from A to B . The function $f : A \rightarrow B$ can also be specified by saying $f(a) = 0$, $f(b) = 1$ and $f(c) = 0$.

The set of ordered pairs $\{(a, 0), (b, 1)\}$ is not a function from A to B because c is not the first coordinate of any ordered pair. The set of ordered pairs $\{(a, 0), (a, 1), (b, 0), (c, 0)\}$ is not a function from A to B because a appears as the first coordinate of two different ordered pairs.

In calculus you may have learned the *vertical line rule* that states that the graph of a function may not intersect a vertical line at more than one point. This corresponds to requiring that each point in the domain of the function appear in only one ordered pair. In set theory, all functions are required to state their domain and range when they are defined. In calculus functions had a domain that was a subset of the real numbers and you were sometimes required to identify the subset.

Example 2.16 This example contrasts the way functions were treated in a typical calculus course with the way we treat them in set theory.

Calculus: find the domain of the function

$$f(x) = \sqrt{x}$$

Since we know that the square root function exists only for non-negative real numbers the domain is $\{x : x \geq 0\}$.

Set theory: the function $f = \sqrt{x}$ from the non-negative real numbers to the real numbers is the set

of ordered pairs $\{(r^2, r) : r \geq 0\}$. This function is well defined because each non-negative real number is the square of some positive real number.

The major contrasts between functions in calculus and functions in set theory are:

- (i) The domain of functions in calculus are often specified only by implication (you have to know how all the functions used work) and are almost always a subset of the real numbers. The domain in set theory must be explicitly specified and may be any set at all.
- (ii) Functions in calculus typically had graphs that you could draw and look at. Geometric intuition driven by the graphs plays a major role in our understanding of functions. Functions in set theory are seldom graphed and often don't have a graph.

A point of similarity between calculus and set theory is that the range of the function is not explicitly specified. When we have a function $f : S \rightarrow T$ then the range of f is a subset of T .

Definition 2.18 If f is a function then we denote the domain of f by $\text{dom}(f)$ and the range of f by $\text{rng}(f)$

Example 2.17 Suppose that $f(n) : \mathbb{N} \rightarrow \mathbb{N}$ is defined by $f(n) = 2n$. Then the domain and range of f are the integers: $\text{dom}(f) = \text{rng}(f) = \mathbb{N}$. If we specify the ordered pairs of f we get

$$f = \{(n, 2n) : n \in \mathbb{N}\}$$

There are actually two definitions of range that are used in mathematics. The definition we are using, the set from which second coordinates of ordered pairs in a function are drawn, is also the definition typically using in computer science. The other definition is the set of second coordinates that actually appear in ordered pairs. This set, which we will define formally later, is the *image* of the function. To make matters even worse the set we are calling the range of a function is also called the *co-domain*. We include these confusing terminological notes for students that may try and look up supplemental material.

Definition 2.19 Let X , Y , and Z be sets. The **composition** of two functions $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ is a function $h : X \rightarrow Z$ for which $h(x) = g(f(x))$ for all $x \in X$. We write $g \circ f$ for the composition of g with f .

The definition of the composition of two functions requires a little checking to make sure it makes sense. Since *every* point must appear as a first coordinate of an ordered pair in a function, every result of applying f to an element of X is an element of Y to which g can be applied. This means that h is a well-defined set of ordered pairs. Notice that the order of composition is important - if the sets X , Y , and Z are distinct there is only one order in which composition even makes sense.

Example 2.18 Suppose that $f : \mathbb{N} \rightarrow \mathbb{N}$ is given by $f(n) = 2n$ while $g : \mathbb{N} \rightarrow \mathbb{N}$ is given by $g(n) = n + 4$. Then

$$(g \circ f)(n) = 2n + 4$$

while

$$(f \circ g)(n) = 2(n + 4) = 2n + 8$$

We now start a series of definitions that divide functions into a number of classes. We will arrive at a point where we can determine if the mapping of a function is reversible, if there is a function that exactly reverses the action of a given function.

Definition 2.20 A function $f : S \rightarrow T$ is **injective** or **one-to-one** if no element of T (no second coordinate) appears in more than one ordered pair. Such a function is called an **injection**.

Example 2.19 The function $f : \mathbb{N} \rightarrow \mathbb{N}$ given by $f(n) = 2n$ is an injection. The ordered pairs of f are $(n, 2n)$ and so any number that appears as a second coordinate does so once.

The function $g : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $g(n) = n^2$ is not an injection. To see this notice that g contains the ordered pairs $(1, 1)$ and $(-1, 1)$ so that 1 appears twice as the second coordinate of an ordered pair.

Definition 2.21 A function $f : S \rightarrow T$ is **surjective** or **onto** if every element of T appears in an ordered pair. Surjective functions are called **surjections**.

We use the symbol \mathbb{R} for the real numbers. We also assume familiarity with interval notation for contiguous subsets of the reals. For real numbers $a \leq b$

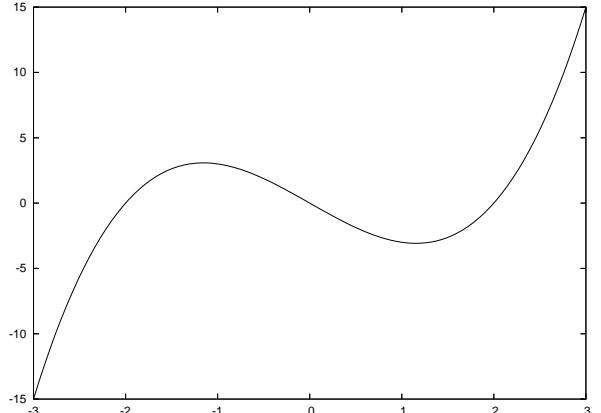
(a, b)	is	$\{x : a < x < b\}$
$(a, b]$	is	$\{x : a < x \leq b\}$
$[a, b)$	is	$\{x : a \leq x < b\}$
$[a, b]$	is	$\{x : a \leq x \leq b\}$

Example 2.20 The function $f : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $f(n) = 5 - n$ is a surjection. If we set $m = 5 - n$ then $n = 5 - m$. This means that if we want to find some n so that $f(n)$ is, for example, 8, then $5 - 8 = -3$ and we see that $f(-3) = 8$. This demonstrates that all m have some n so that $f(n) = m$, showing that all m appear as the second coordinate of an ordered pair in f .

The function $g : \mathbb{R} \rightarrow \mathbb{R}$ given by $g(x) = \frac{x^2}{1+x^2}$ is not a surjection because $-1 < g(x) < 1$ for all $x \in \mathbb{R}$.

Definition 2.22 A function that is both surjective and injective is said to be **bijective**. Bijective functions are called **bijections**.

Example 2.21 The function $f : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $f(n) = n$ is a bijection. All of its ordered pairs have the same first and second coordinate. This function is called the **identity function**.



The function $g : \mathbb{R} \rightarrow \mathbb{R}$ given by $g(x) = x^3 - 4x$ is not a bijection. It is not too hard to show that it is a surjection, but it fails to be an injection. The portion of the graph shown above demonstrates that $g(x)$ takes on the same value more than once. This means that

some numbers appear twice as second coordinates of ordered pairs in g . We can use the graph because g is a function from the real numbers to the real numbers.

For a function $f : S \rightarrow T$ to be a bijection every element of S appears in an ordered pair as the first member of an ordered pair and every element of T appears in an ordered pair as the second member of an ordered pair. Another way to view a bijection is as a matching of the elements of S and T so that every element of S is paired with an element of T . For finite sets this is clearly only possible if the sets are the same size and, in fact, this is the formal definition of “same size” for sets.

Definition 2.23 Two sets S and T are defined to be the same size or to have equal cardinality if there is a bijection $f : S \rightarrow T$.

Example 2.22 The sets $A = \{a, b, c\}$ and $Z = \{1, 2, 3\}$ are the same size. This is obvious because they have the same number of elements, $|A| = |Z| = 3$ but we can construct an explicit bijection

$$f = \{(a, 3), (b, 1), (c, 2)\}$$

with each member of A appearing once as a first coordinate and each member of B appearing once as a second coordinate. This bijection is a witness that A and B are the same size.

Let E be the set of even integers. Then the function

$$g : \mathbb{Z} \rightarrow E$$

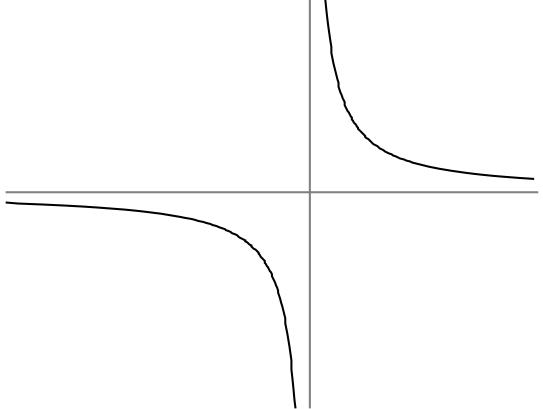
in which $g(n) = 2n$ is a bijection. Notice that each integers can be put into g and that each even integer has exactly one integer that can be doubled to make it. The existence of g is a witness that the set of integers and the set of even integers are the same size. This may seem a bit bizarre because the set $\mathbb{Z} - E$ is the infinite set of odd integers. In fact one hallmark of an infinite set is that it can be the same size as a proper subset. This also means we now have an equality set for sizes of infinite sets. We will do a good deal more with this in Chapter 3.

Bijections have another nice property: they can be unambiguously reversed.

Definition 2.24 The inverse of a function $f : S \rightarrow T$ is a function $g : T \rightarrow S$ so that for all $x \in S$, $g(f(x)) = x$ and for all $y \in T$, $f(g(y)) = y$.

If a function f has an inverse we use the notation f^{-1} for that inverse. Since an exponent of -1 also means reciprocal in some circumstances this can be a bit confusing. The notational confusion is resolved by considering context. So long as we keep firmly in mind that functions are sets of ordered pairs it is easy to prove the proposition/definition that follows after the next example.

Example 2.23 If E is the set of even integers then the bijection $f(n) = 2n$ from \mathbb{Z} to E has the inverse $f^{-1} : E \rightarrow \mathbb{Z}$ given by $g(2n) = n$. Notice that defining the rule for g as depending on the argument $2n$ seamlessly incorporates the fact that the domain of g is the even integers.



If $g(x) = \frac{x}{x-1}$, shown above with its asymptotes $x = 1$ and $y = 1$ then f is a function from the set $H = \mathbb{R} - \{1\}$ to itself. The function was chosen to have asymptotes at equal x and y values; this is a bit unusual. The function g is a bijection. Notice that the graph intersects any horizontal or vertical line in at most one point. Every value except $x = 1$ may be put into g meaning that g is a function on H . Since the vertical asymptote goes off to ∞ in both directions, all values in H come out of g . This demonstrates g is a bijection. This means that it has an inverse which we now compute using a standard

technique from calculus classes.

$$\begin{aligned}y &= \frac{x}{x-1} \\y(x-1) &= x \\xy - y &= x \\xy - x &= y \\x(y-1) &= y \\x &= \frac{y}{y-1}\end{aligned}$$

which tells us that $g^{-1}(x) = \frac{x}{x-1}$ so $g = g^{-1}$: the function is its own inverse.

Proposition 2.5 A function has an inverse if and only if it is a bijection.

Proof:

Suppose that $f : S \rightarrow T$ is a bijection. Then if $g : T \rightarrow S$ has ordered pairs that are the exact reverse of those given by f it is obvious that for all $x \in S$, $g(f(x)) = x$, likewise that for all $y \in T$, $f(g(y)) = y$. We have that bijections possess inverses. It remains to show that non-bijections do not have inverses.

If $f : S \rightarrow T$ is not a bijection then either it is not a surjection or it is not an injection. If f is not a surjection then there is some $t \in T$ that appears in no ordered pair of f . This means that no matter what $g(t)$ is, $f(g(t)) \neq t$ and we fail to have an inverse. If, on the other hand, $f : S \rightarrow T$ is a surjection but fails to be an injection then for some distinct $a, b \in S$ we have that $f(a) = t = f(b)$. For $g : T \rightarrow S$ to be an inverse of f we would need $g(t) = a$ and $g(t) = b$, forcing t to appear as the first coordinate of two ordered pairs in g and so rendering g a non-function. We thus have that non-bijections do not have inverses. \square

The type of inverse we are discussing above is a *two-sided inverse*. The functions f and f^{-1} are mutually inverses of one another. It is possible to find a function that is a one-way inverse of a function so that $f(g(x)) = x$ but $g(f(x))$ is not even defined. These are called *one-sided inverses*.

Note on mathematical grammar: Recall that when two notions, such as “bijection” and “has an inverse” are equivalent we use the phrase “if and only if” (abbreviated iff) to phrase a proposition declaring that the notions are equivalent. A proposition that A iff

B is proven by first assuming A and deducing B and then separately assuming B and deducing A . The formal symbol for A iff B is $A \Leftrightarrow B$. Likewise we have symbols for the ability to deduce B given A , $A \Rightarrow B$ and vice-versa $B \Rightarrow A$. These symbols are spoken “A implies B” and “B implies A” respectively.

Proposition 2.6 Suppose that X , Y , and Z are sets. If $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are bijections then so is $g \circ f : X \rightarrow Z$.

Proof: this proof is left as an exercise.

Definition 2.25 Suppose that $f : A \rightarrow B$ is a function. The **image of A in B** is the subset of B made of elements that appear as the second element of ordered pairs in f . Colloquially the image of f is the set of elements of B hit by f . We use the notation $Im(f)$ for images. In other words $Im(f) = \{f(a) : a \in A\}$.

Example 2.24 If $f : \mathbb{N} \rightarrow \mathbb{N}$ is given by the rule $f(n) = 3n$ then the set $T = \{0, 3, 6, \dots\}$ of natural numbers that are multiples of three is the image of f . Notation: $Im(f) = T$.

If $g : \mathbb{R} \rightarrow \mathbb{R}$ given by $g(x) = x^2$ then

$$Im(g) = \{y : y \geq 0, y \in \mathbb{R}\}$$

There is a name for the set of all ordered pairs drawn from two sets.

Definition 2.26 If A and B are sets then the set of all ordered pairs with the first element from A and the second from B is called the **Cartesian Product of A and B** .

The notation for the Cartesian product of A and B is $A \times B$. using curly brace notation:

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Example 2.25 If $A = \{1, 2\}$ and $B = \{x, y\}$ then

$$A \times B = \{(1, x), (1, y), (2, x), (2, y)\}$$

The **Cartesian plane** is an example of a Cartesian product of the real numbers with themselves: $\mathbb{R} \times \mathbb{R}$.

2.3.1 Permutations

In this section we will look at a very useful sort of function, bijections of finite sets.

Definition 2.27 A **permutation** is a bijection of a finite set with itself. Likewise a bijection of a finite set X with itself is called a **permutation of X** .

Example 2.26 Let $A = \{a, b, c\}$ then the possible permutations of A consist of the following six functions:

$$\begin{array}{ll} \{(a,a)(b,b)(c,c)\} & \{(a,a)(b,c)(c,b)\} \\ \{(a,b)(b,a)(c,c)\} & \{(a,b)(b,c)(c,a)\} \\ \{(a,c)(b,a)(c,b)\} & \{(a,c)(b,b)(c,a)\} \end{array}$$

Notice that the number of permutations of three objects does not depend on the identity of those objects. In fact there are always six permutations of any set of three objects. We now define a handy function that uses a rather odd notation. The method of showing permutations in Example 2.26, explicit listing of ordered pairs, is a bit cumbersome.

Definition 2.28 Assume that we have agreed on an order, e.g. a, b, c , for the members of a set $X = \{a, b, c\}$. Then **one-line notation** for a permutation f consists of listing the first coordinate of the ordered pairs in the agreed on order. The table in Example 2.26 would become:

$$\begin{array}{ll} \text{abc} & \text{acb} \\ \text{bac} & \text{bca} \\ \text{cab} & \text{cba} \end{array}$$

in one line notation. Notice the saving of space.

Definition 2.29 The **factorial** of a natural number n is the product

$$n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1 = \prod_{i=1}^n i$$

with the convention that the factorial of 0 is 1. We denote the factorial of n as $n!$, spoken "n factorial".

Example 2.27 Here are the first few factorials:

n	0	1	2	3	4	5	6	7
$n!$	1	1	2	6	24	120	720	5040

Proposition 2.7 The number of permutations of a finite set with n elements is $n!$.

Proof: this proof is left as an exercise.

Notice that one implication of Proposition 2.6 is that the composition of two permutations is a permutation. This means that the set of permutations of a set is *closed* under functional composition.

Definition 2.30 A **fixed point** of a function $f : S \rightarrow S$ is any $x \in S$ such that $f(x) = x$. We say that **f fixes x**.

Problems

Problem 2.42 Suppose for finite sets A and B that $f : A \rightarrow B$ is an injective function. Prove that

$$|B| \geq |A|$$

Problem 2.43 Suppose that for finite sets A and B that $f : A \rightarrow B$ is a surjective function. Prove that $|A| \geq |B|$.

Problem 2.44 Using functions from the integers to the integers give an example of

- (i) A function that is an injection but not a surjection.
- (ii) A function that is a surjection but not an injection.
- (iii) A function that is neither an injection nor a surjection.
- (iv) A bijection that is not the identity function.

Problem 2.45 For each of the following functions from the real numbers to the real numbers say if the function is surjective or injective. It may be neither.

- (i) $f(x) = x^2$ (ii) $g(x) = x^3$
- (iii) $h(x) = \begin{cases} \sqrt{x} & x \geq 0 \\ -\sqrt{-x} & x < 0 \end{cases}$

Interlude

The Collatz Conjecture

One of the most interesting features of mathematics is that it is possible to phrase problems in a few lines that turn out to be incredibly hard. The Collatz conjecture was first posed in 1937 by Lothar Collatz. Define the function f from the natural numbers to the natural numbers with the rule

$$f(n) = \begin{cases} 3n + 1 & n \text{ odd} \\ \frac{n}{2} & n \text{ even} \end{cases}$$

Collatz' conjecture is that if you apply f repeatedly to a positive integer then the resulting sequence of numbers eventually arrives at one. If we start with 17, for example, the result of repeatedly applying f is:

$$\begin{aligned} f(17) &= 52, f(52) = 26, f(26) = 13, f(13) = 40, f(40) = 20, f(20) = 10, \\ f(10) &= 5, f(5) = 16, f(16) = 8, f(8) = 4, f(4) = 2, f(2) = 1 \end{aligned}$$

The sequences of numbers generated by repeatedly applying f to a natural number are called *hailstone sequences* with the collapse of the value when a large power of 2 appears being analogous to the impact of a hailstone. If we start with the number 27 then 111 steps are required to reach one and the largest intermediate number is 9232. This quite irregular behavior of the sequence is not at all apparent in the original phrasing of the problem.

The Collatz conjecture has been checked for numbers up to 5×2^{61} (about 5.764×10^{18}) by using a variety of computational tricks. It has not, however, been proven or disproven. The very simple statement of the problem causes mathematicians to underestimate the difficulty of the problem. At one point a mathematician suggested that the problem might have been developed by the Russians as a way to slow American mathematical research. This was after several of his colleagues spent months working on the problem without obtaining results.

A simple (but incorrect) argument suggests that hailstone sequences ought to grow indefinitely. Half of all numbers are odd, half are even. The function f slightly more than triples odd numbers and divides even numbers in half. Thus, on average, f increases the value of numbers. The problem is this: half of all even numbers are multiples of four and so are divided in half twice. One-quarter of all even numbers are multiples of eight and so get divided in half three times, and so on. The net effect of factors that are powers of two is to defeat the simple argument that f grows “on average”.

Problem 2.46 True or false (and explain): The function $f(x) = \frac{x-1}{x+1}$ is a bijection from the real numbers to the real numbers.

Problem 2.47 Find a function that is an injection of the integers into the even integers that does not appear in any of the examples in this chapter.

Problem 2.48 Suppose that $B \subset A$ and that there exists a bijection $f : A \rightarrow B$. What may be reasonably deduced about the set A ?

Problem 2.49 Suppose that A and B are finite sets. Prove that $|A \times B| = |A| \cdot |B|$.

Problem 2.50 Suppose that we define $h : \mathbb{N} \rightarrow \mathbb{N}$ as follows. If n is even then $h(n) = n/2$ but if n is odd then $h(n) = 3n + 1$. Determine if h is a (i) surjection or (ii) injection.

Problem 2.51 Prove proposition 2.6.

Problem 2.52 Prove or disprove: the composition of injections is an injection.

Problem 2.53 Prove or disprove: the composition of surjections is a surjection.

Problem 2.54 Prove proposition 2.7.

Problem 2.55 List all permutations of

$$X = \{1, 2, 3, 4\}$$

using one-line notation.

Problem 2.56 Suppose that X is a set and that f , g , and h are permutations of X . Prove that the equation $f \circ g = h$ has a solution g for any given permutations f and h .

Problem 2.57 Examine the permutation f of $Q = \{a, b, c, d, e\}$ which is **bcaed** in one line notation. If we create the series $f, f \circ f, f \circ (f \circ f), \dots$ does the identity function, **abcde**, ever appear in the series? If so, what is its first appearance? If not, why not?

Problem 2.58 If f is a permutation of a finite set, prove that the sequence $f, f \circ f, f \circ (f \circ f), \dots$ must contain repeated elements.

Problem 2.59 Suppose that X and Y are finite sets and that $|X| = |Y| = n$. Prove that there are $n!$ bijections of X with Y .

Problem 2.60 Suppose that X and Y are sets with $|X| = n$, $|Y| = m$. Count the number of functions from X to Y .

Problem 2.61 Suppose that X and Y are sets with $|X| = n$, $|Y| = m$ for $m > n$. Count the number of injections of X into Y .

Problem 2.62 For a finite set S with a subset T prove that the permutations of S that have all members of T as fixed points form a set that is closed under functional composition.

Problem 2.63 Compute the number of permutations of a set S with n members that fix at least $m < n$ points.

Problem 2.64 Using any technique at all, estimate the fraction of permutations of an n -element set that have no fixed points. This problem is intended as an exploration.

Problem 2.65 Let X be a finite set with $|X| = n$. Let $C = X \times X$. How many subsets of C have the property that every element of X appears once as a first coordinate of some ordered pair and once as a second coordinate of some ordered pair?

Problem 2.66 An alternate version of Sigma (\sum) and Pi (\prod) notation works by using a set as an index. So if $S = \{1, 3, 5, 7\}$ then

$$\sum_{s \in S} s = 16 \text{ and } \prod_{s \in S} s = 105$$

Given all the material so far, give and defend reasonable values for the sum and product of an empty set.

Problem 2.67 Suppose that $f_\alpha : [0, 1] \rightarrow [0, 1]$ for $-1 < \alpha < \infty$ is given by

$$f_\alpha(x) = \frac{(\alpha + 1)x}{\alpha x + 1},$$

prove that f_α is a bijection.

Problem 2.68 Find, to five decimals accuracy:

$$\ln(200!)$$

Explain how you obtained the answer.

2.4 $\infty + 1$

We conclude the chapter with a brief section that demonstrates a strange thing that can be accomplished with set notation. We choose to represent the natural numbers $0, 1, 2, \dots$ by sets that contain the number of elements counted by the corresponding natural number. We also choose to do so as simply as possible, using only curly braces and commas. Given this the numbers and their corresponding sets are:

$$\begin{aligned} 0 &: \{\} \\ 1 &: \{\{\}\} = \{0\} \\ 2 &: \{\{\}, \{\{\}\}\} = \{0, 1\} \\ 3 &: \{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}\} = \{0, 1, 2\} \\ 4 &: \{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}, \{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}\} \\ &\quad = \{0, 1, 2, 3\} \end{aligned}$$

The trick for the above representation is this. Zero is represented by the empty set. One is represented by the set of the only thing we have constructed - zero, represented as the empty set. Similarly the representation of two is the set of the representation of zero and one (the empty set and the set of the empty set). This representation is incredibly inefficient but it uses a very small number of symbols. This representation also has a useful property. As always, we will start with a definition.

Definition 2.31 *The minimal set representation of the natural numbers is constructed as follows:*

- (i) *Let 0 be represented by the empty set.*
- (ii) *For $n > 0$ let n be represented by the set $\{0, 1, \dots, n - 1\}$.*

The shorthand $\{0, 1\}$ for $\{\{\}, \{\{\}\}\}$ is called the *simplified notation* for the minimal set representation. We now give the useful property of the minimal set representation.

Proposition 2.8 $n + 1 = n \cup \{n\}$

Proof:

This follows directly from Definition 2.31 by considering the set difference of the representations of n and $n - 1$. \square

The definition says that any set of the representations of consecutive natural numbers, starting at zero, is

the representation of the next natural number. This permits us to conclude that the set of all natural numbers

$$\{0, 1, 2, \dots\}$$

fits the definition of a natural number. Which natural number is it? It is easy to see, in the minimal set representation, that for natural numbers m and n , $m < n$ implies that the representation of m is a subset of the representation of n . Every finite natural number is a subset of the set of all natural numbers and so we conclude that $\{0, 1, 2, \dots\}$ is an infinite natural number. The set notation thus permits us to construct an infinite number.

The set consisting of the representations of all finite natural numbers is an infinite natural number. The number has been given the name ω , the lower-case omega. In addition to being a letter omega traditionally also means “the last”. The number ω comes after all the finite natural numbers. If we now apply Proposition 2.8 we see that

$$\omega \cup \{\omega\} = \omega + 1$$

This means that we can add one to an infinite number. Is the resulting number $\omega + 1$ a different number from ω ? It turns out the answer is “yes”, because the representations of these numbers are different as sets. The representation of ω contains no infinite sets while the representation of $\omega + 1$ contains one.

Problems

Problem 2.69 *Find the representation for 5 using the curly-brace-and-comma notation.*

Problem 2.70 *Give the minimal set representation of $\omega + 2$ using the simplified notation.*

Problem 2.71 *Suppose that $n > m$ are natural numbers and that S is the minimal set representation of n while T is the minimal set representation of m . Is the representation of $n - m$ a member of the set difference $S - T$?*

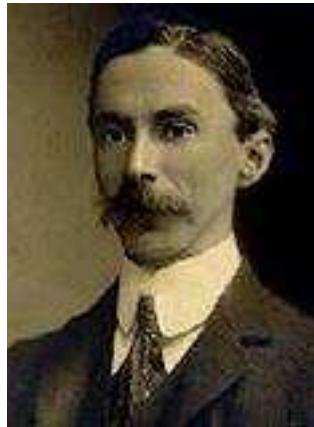
Problem 2.72 *Give a formula, as a function of n , for the number of times that the symbol $\{$ appears in the representation of n .*

Problem 2.73 *Prove or disprove: there are an infinite number of distinct infinite numbers.*

Interlude

Russell's Paradox

Bertrand Arthur William Russell, 3rd Earl Russell, OM, FRS (18 May 1872–2 February 1970), commonly known as simply Bertrand Russell, was a British philosopher, logician, mathematician, historian, religious skeptic, social reformer, socialist and pacifist. Although he spent the majority of his life in England, he was born in Wales, where he also died.



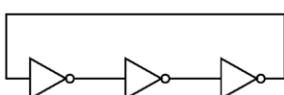
Let Q be the set of all sets that do not contain themselves as a member. Consider the question: “Does Q contain itself?” If the answer to this question is no then Q , by definition must contain itself. If, however, Q contains itself then it is by definition unable to contain itself. This rather annoying contradiction, constructed by Russell, had a rather amusing side effect.

Friedrich Frege had just finished the second of a three volume set of works called the *Basic Laws of Arithmetic* that was supposed to remove all intuition from mathematics and place it on a purely logical basis. Russell wrote Frege, explaining his paradox. Frege added an appendix to his second volume that attempted to avoid Russell's paradox. The third volume was never published.

It is possible to resolve Russell's paradox by being much more careful about what objects may be defined to be sets; the *category* of all sets that do not contain themselves gives rise to no contradiction (it does give rise to an entire field of mathematics, category theory). The key to resolving the paradox from a set theoretic perspective is that one cannot assume that, for every property, there is a set of all things satisfying that property. This is a reason why it is important that a set is properly defined. Another consequence of Russell's paradox is a warning that self-referential statements are both potentially interesting and fairly dangerous, at least on the intellectual plane.

The original phrasing of Russell's paradox was in terms of normal and abnormal sets. A set is *normal* if it fails to contain itself and abnormal otherwise. Consider the set of all normal sets. If this set is abnormal, it contains itself but by definition the set contains only normal sets and hence it is itself normal. The normality of this set forces the set to contain itself, which makes it abnormal. This is simply a rephrasing of the original contradiction.

Puzzle: what does the circuit below have to do with Russell's paradox and what use is it?



String Theory

University of Cambridge Part III Mathematical Tripos

Dr David Tong

*Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OWA, UK*

<http://www.damtp.cam.ac.uk/user/tong/string.html>
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

- J. Polchinski, *String Theory*

This two volume work is the standard introduction to the subject. Our lectures will more or less follow the path laid down in volume one covering the bosonic string. The book contains explanations and descriptions of many details that have been deliberately (and, I suspect, at times inadvertently) swept under a very large rug in these lectures. Volume two covers the superstring.

- M. Green, J. Schwarz and E. Witten, *Superstring Theory*

Another two volume set. It is now over 20 years old and takes a slightly old-fashioned route through the subject, with no explicit mention of conformal field theory. However, it does contain much good material and the explanations are uniformly excellent. Volume one is most relevant for these lectures.

- B. Zwiebach, *A First Course in String Theory*

This book grew out of a course given to undergraduates who had no previous exposure to general relativity or quantum field theory. It has wonderful pedagogical discussions of the basics of lightcone quantization. More surprisingly, it also has some very clear descriptions of several advanced topics, even though it misses out all the bits in between.

- P. Di Francesco, P. Mathieu and D. Sénéchal, *Conformal Field Theory*

This big yellow book is affectionately known as the yellow pages. It's a great way to learn conformal field theory. At first glance, it comes across as slightly daunting because it's big. (And yellow). But you soon realise that it's big because it starts at the beginning and provides detailed explanations at every step. The material necessary for this course can be found in chapters 5 and 6.

Further References: “*String Theory and M-Theory*” by Becker, Becker and Schwarz and “*String Theory in a Nutshell*” (it’s a big nutshell) by Kiritsis both deal with the bosonic string fairly quickly, but include more advanced topics that may be of interest. The book “*D-Branes*” by Johnson has lively and clear discussions about the many joys of D-branes. Links to several excellent online resources, including video lectures by Shiraz Minwalla, are listed on the course webpage.

Contents

0. Introduction	1
0.1 Quantum Gravity	3
1. The Relativistic String	9
1.1 The Relativistic Point Particle	9
1.1.1 Quantization	11
1.1.2 Ein Einbein	13
1.2 The Nambu-Goto Action	14
1.2.1 Symmetries of the Nambu-Goto Action	17
1.2.2 Equations of Motion	18
1.3 The Polyakov Action	18
1.3.1 Symmetries of the Polyakov Action	20
1.3.2 Fixing a Gauge	22
1.4 Mode Expansions	25
1.4.1 The Constraints Revisited	26
2. The Quantum String	28
2.1 A Lightning Look at Covariant Quantization	28
2.1.1 Ghosts	30
2.1.2 Constraints	30
2.2 Lightcone Quantization	32
2.2.1 Lightcone Gauge	33
2.2.2 Quantization	36
2.3 The String Spectrum	40
2.3.1 The Tachyon	40
2.3.2 The First Excited States	41
2.3.3 Higher Excited States	45
2.4 Lorentz Invariance Revisited	46
2.5 A Nod to the Superstring	48
3. Open Strings and D-Branes	50
3.1 Quantization	53
3.1.1 The Ground State	54
3.1.2 First Excited States: A World of Light	55

3.1.3	Higher Excited States and Regge Trajectories	56
3.1.4	Another Nod to the Superstring	56
3.2	Brane Dynamics: The Dirac Action	57
3.3	Multiple Branes: A World of Glue	59
4.	Introducing Conformal Field Theory	61
4.0.1	Euclidean Space	62
4.0.2	The Holomorphy of Conformal Transformations	63
4.1	Classical Aspects	63
4.1.1	The Stress-Energy Tensor	64
4.1.2	Noether Currents	66
4.1.3	An Example: The Free Scalar Field	67
4.2	Quantum Aspects	68
4.2.1	Operator Product Expansion	68
4.2.2	Ward Identities	70
4.2.3	Primary Operators	73
4.3	An Example: The Free Scalar Field	77
4.3.1	The Propagator	77
4.3.2	An Aside: No Goldstone Bosons in Two Dimensions	79
4.3.3	The Stress-Energy Tensor and Primary Operators	80
4.4	The Central Charge	82
4.4.1	c is for Casimir	85
4.4.2	The Weyl Anomaly	86
4.4.3	c is for Cardy	89
4.4.4	c has a Theorem	91
4.5	The Virasoro Algebra	94
4.5.1	Radial Quantization	94
4.5.2	The Virasoro Algebra	97
4.5.3	Representations of the Virasoro Algebra	99
4.5.4	Consequences of Unitarity	100
4.6	The State-Operator Map	101
4.6.1	Some Simple Consequences	104
4.6.2	Our Favourite Example: The Free Scalar Field	105
4.7	Brief Comments on Conformal Field Theories with Boundaries	108
5.	The Polyakov Path Integral and Ghosts	110
5.1	The Path Integral	110
5.1.1	The Faddeev-Popov Method	111

5.1.2	The Faddeev-Popov Determinant	114
5.1.3	Ghosts	115
5.2	The Ghost CFT	116
5.3	The Critical “Dimension” of String Theory	119
5.3.1	The Usual Nod to the Superstring	120
5.3.2	An Aside: Non-Critical Strings	121
5.4	States and Vertex Operators	122
5.4.1	An Example: Closed Strings in Flat Space	124
5.4.2	An Example: Open Strings in Flat Space	125
5.4.3	More General CFTs	126
6.	String Interactions	127
6.1	What to Compute?	127
6.1.1	Summing Over Topologies	129
6.2	Closed String Amplitudes at Tree Level	132
6.2.1	Remnant Gauge Symmetry: $SL(2, \mathbb{C})$	132
6.2.2	The Virasoro-Shapiro Amplitude	134
6.2.3	Lessons to Learn	137
6.3	Open String Scattering	141
6.3.1	The Veneziano Amplitude	143
6.3.2	The Tension of D-Branes	144
6.4	One-Loop Amplitudes	145
6.4.1	The Moduli Space of the Torus	145
6.4.2	The One-Loop Partition Function	148
6.4.3	Interpreting the String Partition Function	151
6.4.4	So is String Theory Finite?	154
6.4.5	Beyond Perturbation Theory?	155
6.5	Appendix: Games with Integrals and Gamma Functions	156
7.	Low Energy Effective Actions	159
7.1	Einstein’s Equations	160
7.1.1	The Beta Function	161
7.1.2	Ricci Flow	165
7.2	Other Couplings	165
7.2.1	Charged Strings and the B field	165
7.2.2	The Dilaton	167
7.2.3	Beta Functions	169
7.3	The Low-Energy Effective Action	169

7.3.1	String Frame and Einstein Frame	170
7.3.2	Corrections to Einstein's Equations	172
7.3.3	Nodding Once More to the Superstring	173
7.4	Some Simple Solutions	175
7.4.1	Compactifications	176
7.4.2	The String Itself	177
7.4.3	Magnetic Branes	179
7.4.4	Moving Away from the Critical Dimension	182
7.4.5	The Elephant in the Room: The Tachyon	185
7.5	D-Branes Revisited: Background Gauge Fields	185
7.5.1	The Beta Function	186
7.5.2	The Born-Infeld Action	189
7.6	The DBI Action	190
7.6.1	Coupling to Closed String Fields	191
7.7	The Yang-Mills Action	193
7.7.1	D-Branes in Type II Superstring Theories	197
8.	Compactification and T-Duality	199
8.1	The View from Spacetime	199
8.1.1	Moving around the Circle	201
8.2	The View from the Worldsheet	202
8.2.1	Massless States	204
8.2.2	Charged Fields	204
8.2.3	Enhanced Gauge Symmetry	205
8.3	Why Big Circles are the Same as Small Circles	206
8.3.1	A Path Integral Derivation of T-Duality	208
8.3.2	T-Duality for Open Strings	209
8.3.3	T-Duality for Superstrings	210
8.3.4	Mirror Symmetry	210
8.4	Epilogue	211

Acknowledgements

These lectures are aimed at beginning graduate students. They assume a working knowledge of quantum field theory and general relativity. The lectures were given over one semester and are based broadly on Volume one of the book by Joe Polchinski. I inherited the course from Michael Green whose notes were extremely useful. I also benefited enormously from the insightful and entertaining video lectures by Shiraz Minwalla.

I'm grateful to Anirban Basu, Niklas Beisert, Joe Bhaseen, Diego Correa, Nick Dorey, Michael Green, Anshuman Maharana, Malcolm Perry and Martin Schnabl for discussions and help with various aspects of these notes. I'm also grateful to the students, especially Carlos Guedes, for their excellent questions and superhuman typo-spotting abilities. Finally, my thanks to Alex Considine for infinite patience and understanding over the weeks these notes were written. I am supported by the Royal Society.

0. Introduction

String theory is an ambitious project. It purports to be an all-encompassing theory of the universe, unifying the forces of nature, including gravity, in a single quantum mechanical framework.

The premise of string theory is that, at the fundamental level, matter does not consist of point-particles but rather of tiny loops of string. From this slightly absurd beginning, the laws of physics emerge. General relativity, electromagnetism and Yang-Mills gauge theories all appear in a surprising fashion. However, they come with baggage. String theory gives rise to a host of other ingredients, most strikingly extra spatial dimensions of the universe beyond the three that we have observed. The purpose of this course is to understand these statements in detail.

These lectures differ from most other courses that you will take in a physics degree. String theory is speculative science. There is no experimental evidence that string theory is the correct description of our world and scant hope that hard evidence will arise in the near future. Moreover, string theory is very much a work in progress and certain aspects of the theory are far from understood. Unresolved issues abound and it seems likely that the final formulation has yet to be written. For these reasons, I'll begin this introduction by suggesting some answers to the question: Why study string theory?

Reason 1. String theory is a theory of quantum gravity

String theory unifies Einstein's theory of general relativity with quantum mechanics. Moreover, it does so in a manner that retains the explicit connection with both quantum theory and the low-energy description of spacetime.

But quantum gravity contains many puzzles, both technical and conceptual. What does spacetime look like at the shortest distance scales? How can we understand physics if the causal structure fluctuates quantum mechanically? Is the big bang truly the beginning of time? Do singularities that arise in black holes really signify the end of time? What is the microscopic origin of black hole entropy and what is it telling us? What is the resolution to the information paradox? Some of these issues will be reviewed later in this introduction.

Whether or not string theory is the true description of reality, it offers a framework in which one can begin to explore these issues. For some questions, string theory has given very impressive and compelling answers. For others, string theory has been almost silent.

Reason 2. String theory may be *the theory of quantum gravity*

With broad brush, string theory looks like an extremely good candidate to describe the real world. At low-energies it naturally gives rise to general relativity, gauge theories, scalar fields and chiral fermions. In other words, it contains all the ingredients that make up our universe. It also gives the only presently credible explanation for the value of the cosmological constant although, in fairness, I should add that the explanation is so distasteful to some that the community is rather amusingly split between whether this is a good thing or a bad thing. Moreover, string theory incorporates several ideas which do not yet have experimental evidence but which are considered to be likely candidates for physics beyond the standard model. Prime examples are supersymmetry and axions.

However, while the broad brush picture looks good, the finer details have yet to be painted. String theory does not provide unique predictions for low-energy physics but instead offers a bewildering array of possibilities, mostly dependent on what is hidden in those extra dimensions. Partly, this problem is inherent to any theory of quantum gravity: as we'll review shortly, it's a long way down from the Planck scale to the domestic energy scales explored at the LHC. Using quantum gravity to extract predictions for particle physics is akin to using QCD to extract predictions for how coffee makers work. But the mere fact that it's hard is little comfort if we're looking for convincing evidence that string theory describes the world in which we live.

While string theory cannot at present offer falsifiable predictions, it has nonetheless inspired new and imaginative proposals for solving outstanding problems in particle physics and cosmology. There are scenarios in which string theory might reveal itself in forthcoming experiments. Perhaps we'll find extra dimensions at the LHC, perhaps we'll see a network of fundamental strings stretched across the sky, or perhaps we'll detect some feature of non-Gaussianity in the CMB that is characteristic of D-branes at work during inflation. My personal feeling however is that each of these is a long shot and we may not know whether string theory is right or wrong within our lifetimes. Of course, the history of physics is littered with naysayers, wrongly suggesting that various theories will never be testable. With luck, I'll be one of them.

Reason 3. String theory provides new perspectives on gauge theories

String theory was born from attempts to understand the strong force. Almost forty years later, this remains one of the prime motivations for the subject. String theory provides tools with which to analyze down-to-earth aspects of quantum field theory that are far removed from high-falutin' ideas about gravity and black holes.

Of immediate relevance to this course are the pedagogical reasons to invest time in string theory. At heart, it is the study of conformal field theory and gauge symmetry. The techniques that we'll learn are not isolated to string theory, but apply to countless systems which have direct application to real world physics.

On a deeper level, string theory provides new and very surprising methods to understand aspects of quantum gauge theories. Of these, the most startling is the *AdS/CFT correspondence*, first conjectured by Juan Maldacena, which gives a relationship between strongly coupled quantum field theories and gravity in higher dimensions. These ideas have been applied in areas ranging from nuclear physics to condensed matter physics and have provided qualitative (and arguably quantitative) insights into strongly coupled phenomena.

Reason 4. String theory provides new results in mathematics

For the past 250 years, the close relationship between mathematics and physics has been almost a one-way street: physicists borrowed many things from mathematicians but, with a few noticeable exceptions, gave little back. In recent times, that has changed. Ideas and techniques from string theory and quantum field theory have been employed to give new “proofs” and, perhaps more importantly, suggest new directions and insights in mathematics. The most well known of these is *mirror symmetry*, a relationship between topologically different Calabi-Yau manifolds.

The four reasons described above also crudely characterize the string theory community: there are “relativists” and “phenomenologists” and “field theorists” and “mathematicians”. Of course, the lines between these different sub-disciplines are not fixed and one of the great attractions of string theory is its ability to bring together people working in different areas — from cosmology to condensed matter to pure mathematics — and provide a framework in which they can profitably communicate. In my opinion, it is this cross-fertilization between fields which is the greatest strength of string theory.

0.1 Quantum Gravity

This is a starter course in string theory. Our focus will be on the perturbative approach to the bosonic string and, in particular, why this gives a consistent theory of quantum gravity. Before we leap into this, it is probably best to say a few words about quantum gravity itself. Like why it's hard. And why it's important. (And why it's not).

The Einstein Hilbert action is given by

$$S_{EH} = \frac{1}{16\pi G_N} \int d^4x \sqrt{-g} \mathcal{R}$$

Newton's constant G_N can be written as

$$8\pi G_N = \frac{\hbar c}{M_{pl}^2}$$

Throughout these lectures we work in units with $\hbar = c = 1$. The Planck mass M_{pl} defines an energy scale

$$M_{pl} \approx 2 \times 10^{18} \text{ GeV} .$$

(This is sometimes referred to as the reduced Planck mass, to distinguish it from the scale without the factor of 8π , namely $\sqrt{1/G_N} \approx 1 \times 10^{19}$ GeV).

There are a couple of simple lessons that we can already take from this. The first is that the relevant coupling in the quantum theory is $1/M_{pl}$. To see that this is indeed the case from the perspective of the action, we consider small perturbations around flat Minkowski space,

$$g_{\mu\nu} = \eta_{\mu\nu} + \frac{1}{M_{pl}} h_{\mu\nu}$$

The factor of $1/M_{pl}$ is there to ensure that when we expand out the Einstein-Hilbert action, the kinetic term for h is canonically normalized, meaning that it comes with no powers of M_{pl} . This then gives the kind of theory that you met in your first course on quantum field theory, albeit with an infinite series of interaction terms,

$$S_{EH} = \int d^4x (\partial h)^2 + \frac{1}{M_{pl}} h (\partial h)^2 + \frac{1}{M_{pl}^2} h^2 (\partial h)^2 + \dots$$

Each of these terms is schematic: if you were to do this explicitly, you would find a mess of indices contracted in different ways. We see that the interactions are suppressed by powers of M_{pl} . This means that quantum perturbation theory is an expansion in the dimensionless ratio E^2/M_{pl}^2 , where E is the energy associated to the process of interest. We learn that gravity is weak, and therefore under control, at low-energies. But gravitational interactions become strong as the energy involved approaches the Planck scale. In the language of the renormalization group, couplings of this type are known as *irrelevant*.

The second lesson to take away is that the Planck scale M_{pl} is very very large. The LHC will probe the electroweak scale, $M_{EW} \sim 10^3$ GeV. The ratio is $M_{EW}/M_{pl} \sim 10^{-15}$. For this reason, quantum gravity will not affect your daily life, even if your daily life involves the study of the most extreme observable conditions in the universe.

Gravity is Non-Renormalizable

Quantum field theories with irrelevant couplings are typically ill-behaved at high-energies, rendering the theory ill-defined. Gravity is no exception. Theories of this type are called *non-renormalizable*, which means that the divergences that appear in the Feynman diagram expansion cannot be absorbed by a finite number of counterterms. In pure Einstein gravity, the symmetries of the theory are enough to ensure that the one-loop S-matrix is finite. The first divergence occurs at two-loops and requires the introduction of a counterterm of the form,

$$\Gamma \sim \frac{1}{\epsilon} \frac{1}{M_{pl}^4} \int d^4x \sqrt{-g} \mathcal{R}^{\mu\nu}{}_{\rho\sigma} \mathcal{R}^{\rho\sigma}{}_{\lambda\kappa} \mathcal{R}^{\lambda\kappa}{}_{\mu\nu}$$

with $\epsilon = 4 - D$. All indications point towards the fact that this is the first in an infinite number of necessary counterterms.

Coupling gravity to matter requires an interaction term of the form,

$$S_{int} = \int d^4x \frac{1}{M_{pl}} h_{\mu\nu} T^{\mu\nu} + \mathcal{O}(h^2)$$

This makes the situation marginally worse, with the first divergence now appearing at one-loop. The Feynman diagram in the figure shows particle scattering through the exchange of two gravitons. When the momentum k running in the loop is large, the diagram is badly divergent: it scales as

$$\frac{1}{M_{pl}^4} \int^\infty d^4k$$

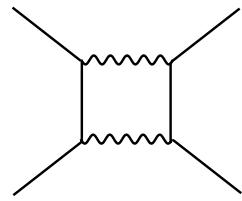


Figure 1:

Non-renormalizable theories are commonplace in the history of physics, the most commonly cited example being Fermi's theory of the weak interaction. The first thing to say about them is that they are far from useless! Non-renormalizable theories are typically viewed as *effective* field theories, valid only up to some energy scale Λ . One deals with the divergences by simply admitting ignorance beyond this scale and treating Λ as a UV cut-off on any momentum integral. In this way, we get results which are valid to an accuracy of E/Λ (perhaps raised to some power). In the case of the weak interaction, Fermi's theory accurately predicts physics up to an energy scale of $\sqrt{1/G_F} \sim 100$ GeV. In the case of quantum gravity, Einstein's theory works to an accuracy of $(E/M_{pl})^2$.

However, non-renormalizable theories are typically unable to describe physics at their cut-off scale Λ or beyond. This is because they are missing the true ultra-violet degrees of freedom which tame the high-energy behaviour. In the case of the weak force, these new degrees of freedom are the W and Z bosons. We would like to know what missing degrees of freedom are needed to complete gravity.

Singularities

Only a particle physicist would phrase all questions about the universe in terms of scattering amplitudes. In general relativity we typically think about the geometry as a whole, rather than bastardizing the Einstein-Hilbert action and discussing perturbations around flat space. In this language, the question of high-energy physics turns into one of short distance physics. Classical general relativity is not to be trusted in regions where the curvature of spacetime approaches the Planck scale and ultimately becomes singular. A quantum theory of gravity should resolve these singularities.

The question of spacetime singularities is morally equivalent to that of high-energy scattering. Both probe the ultra-violet nature of gravity. A spacetime geometry is made of a coherent collection of gravitons, just as the electric and magnetic fields in a laser are made from a collection of photons. The short distance structure of spacetime is governed – after Fourier transform – by high momentum gravitons. Understanding spacetime singularities and high-energy scattering are different sides of the same coin.

There are two situations in general relativity where singularity theorems tell us that the curvature of spacetime gets large: at the big bang and in the center of a black hole. These provide two of the biggest challenges to any putative theory of quantum gravity.

Gravity is Subtle

It is often said that general relativity contains the seeds of its own destruction. The theory is unable to predict physics at the Planck scale and freely admits to it. Problems such as non-renormalizability and singularities are, in a Rumsfeldian sense, known unknowns. However, the full story is more complicated and subtle. On the one hand, the issue of non-renormalizability may not quite be the crisis that it first appears. On the other hand, some aspects of quantum gravity suggest that general relativity isn't as honest about its own failings as is usually advertised. The theory hosts a number of unknown unknowns, things that we didn't even know that we didn't know. We won't have a whole lot to say about these issues in this course, but you should be aware of them. Here I mention only a few salient points.

Firstly, there is a key difference between Fermi’s theory of the weak interaction and gravity. Fermi’s theory was unable to provide predictions for any scattering process at energies above $\sqrt{1/G_F}$. In contrast, if we scatter two objects at extremely high-energies in gravity — say, at energies $E \gg M_{pl}$ — then we know exactly what will happen: we form a big black hole. We don’t need quantum gravity to tell us this. Classical general relativity is sufficient. If we restrict attention to scattering, the crisis of non-renormalizability is not problematic at ultra-high energies. It’s troublesome only within a window of energies around the Planck scale.

Similar caveats hold for singularities. If you are foolish enough to jump into a black hole, then you’re on your own: without a theory of quantum gravity, no one can tell you what fate lies in store at the singularity. Yet, if you are smart and stay outside of the black hole, you’ll be hard pushed to see any effects of quantum gravity. This is because Nature has conspired to hide Planck scale curvatures from our inquisitive eyes. In the case of black holes this is achieved through cosmic censorship which is a conjecture in classical general relativity that says singularities are hidden behind horizons. In the case of the big bang, it is achieved through inflation, washing away any traces from the very early universe. Nature appears to shield us from the effects of quantum gravity, whether in high-energy scattering or in singularities. I think it’s fair to say that no one knows if this conspiracy is pointing at something deep, or is merely inconvenient for scientists trying to probe the Planck scale.

While horizons may protect us from the worst excesses of singularities, they come with problems of their own. These are the unknown unknowns: difficulties that arise when curvatures are small and general relativity says “trust me”. The entropy of black holes and the associated paradox of information loss strongly suggest that local quantum field theory breaks down at macroscopic distance scales. Attempts to formulate quantum gravity in de Sitter space, or in the presence of eternal inflation, hint at similar difficulties. Ideas of holography, black hole complimentarity and the AdS/CFT correspondence all point towards non-local effects and the emergence of spacetime. These are the deep puzzles of quantum gravity and their relationship to the ultra-violet properties of gravity is unclear.

As a final thought, let me mention the one observation that has an outside chance of being related to quantum gravity: the cosmological constant. With an energy scale of $\Lambda \sim 10^{-3}$ eV it appears to have little to do with ultra-violet physics. If it does have its origins in a theory of quantum gravity, it must either be due to some subtle “unknown unknown”, or because it is explained away as an environmental quantity as in string theory.

Is the Time Ripe?

Our current understanding of physics, embodied in the standard model, is valid up to energy scales of 10^3 GeV. This is 15 orders of magnitude away from the Planck scale. Why do we think the time is now ripe to tackle quantum gravity? Surely we are like the ancient Greeks arguing about atomism. Why on earth do we believe that we've developed the right tools to even address the question?

The honest answer, I think, is hubris.

However, there is mild circumstantial evidence that the framework of quantum field theory might hold all the way to the Planck scale without anything very dramatic happening in between. The main argument is unification. The three coupling constants of Nature run logarithmically, meeting miraculously at the GUT energy scale of 10^{15} GeV. Just slightly later, the fourth force of Nature, gravity, joins them. While not overwhelming, this does provide a hint that perhaps quantum field theory can be taken seriously at these ridiculous scales.

Historically I suspect this was what convinced large parts of the community that it was ok to speak about processes at 10^{18} GeV.

Finally, perhaps the most compelling argument for studying physics at the Planck scale is that string theory *does* provide a consistent unified quantum theory of gravity and the other forces. Given that we have this theory sitting in our laps, it would be foolish not to explore its consequences. The purpose of these lecture notes is to begin this journey.

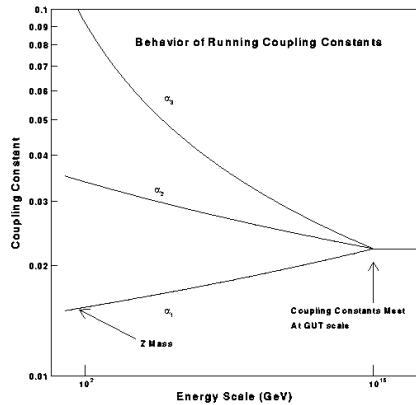


Figure 2:

1. The Relativistic String

All lecture courses on string theory start with a discussion of the point particle. Ours is no exception. We'll take a flying tour through the physics of the relativistic point particle and extract a couple of important lessons that we'll take with us as we move onto string theory.

1.1 The Relativistic Point Particle

We want to write down the Lagrangian describing a relativistic particle of mass m . In anticipation of string theory, we'll consider D -dimensional Minkowski space $\mathbf{R}^{1,D-1}$. Throughout these notes, we work with signature

$$\eta_{\mu\nu} = \text{diag}(-1, +1, +1, \dots, +1)$$

Note that this is the opposite signature to my quantum field theory notes.

If we fix a frame with coordinates $X^\mu = (t, \vec{x})$ the action is simple:

$$S = -m \int dt \sqrt{1 - \dot{\vec{x}} \cdot \dot{\vec{x}}} . \quad (1.1)$$

To see that this is correct we can compute the momentum \vec{p} , conjugate to \vec{x} , and the energy E which is equal to the Hamiltonian,

$$\vec{p} = \frac{m \dot{\vec{x}}}{\sqrt{1 - \dot{\vec{x}} \cdot \dot{\vec{x}}}} , \quad E = \sqrt{m^2 + \vec{p}^2} ,$$

both of which should be familiar from courses on special relativity.

Although the Lagrangian (1.1) is correct, it's not fully satisfactory. The reason is that time t and space \vec{x} play very different roles in this Lagrangian. The position \vec{x} is a dynamical degree of freedom. In contrast, time t is merely a parameter providing a label for the position. Yet Lorentz transformations are supposed to mix up t and \vec{x} and such symmetries are not completely obvious in (1.1). Can we find a new Lagrangian in which time and space are on equal footing?

One possibility is to treat both time and space as labels. This leads us to the concept of field theory. However, in this course we will be more interested in the other possibility: we will promote time to a dynamical degree of freedom. At first glance, this may appear odd: the number of degrees of freedom is one of the crudest ways we have to characterize a system. We shouldn't be able to add more degrees of freedom

at will without fundamentally changing the system that we're talking about. Another way of saying this is that the particle has the option to move in space, but it doesn't have the option to move in time. It *has* to move in time. So we somehow need a way to promote time to a degree of freedom without it really being a true dynamical degree of freedom! How do we do this? The answer, as we will now show, is gauge symmetry.

Consider the action,

$$S = -m \int d\tau \sqrt{-\dot{X}^\mu \dot{X}^\nu \eta_{\mu\nu}}, \quad (1.2)$$

where $\mu = 0, \dots, D-1$ and $\dot{X}^\mu = dX^\mu/d\tau$. We've introduced a new parameter τ which labels the position along the worldline of the particle as shown by the dashed lines in the figure. This action has a simple interpretation: it is just the proper time $\int ds$ along the worldline.

Naively it looks as if we now have D physical degrees of freedom rather than $D-1$ because, as promised, the time direction $X^0 \equiv t$ is among our dynamical variables: $X^0 = X^0(\tau)$. However, this is an illusion. To see why, we need to note that the action (1.2) has a very important property: reparameterization invariance. This means that we can pick a different parameter $\tilde{\tau}$ on the worldline, related to τ by any monotonic function

$$\tilde{\tau} = \tilde{\tau}(\tau).$$

Let's check that the action is invariant under transformations of this type. The integration measure in the action changes as $d\tau = d\tilde{\tau} |d\tau/d\tilde{\tau}|$. Meanwhile, the velocities change as $dX^\mu/d\tau = (dX^\mu/d\tilde{\tau})(d\tilde{\tau}/d\tau)$. Putting this together, we see that the action can just as well be written in the $\tilde{\tau}$ reparameterization,

$$S = -m \int d\tilde{\tau} \sqrt{-\frac{dX^\mu}{d\tilde{\tau}} \frac{dX^\nu}{d\tilde{\tau}} \eta_{\mu\nu}}.$$

The upshot of this is that not all D degrees of freedom X^μ are physical. For example, suppose you find a solution to this system, so that you know how X^0 changes with τ and how X^1 changes with τ and so on. Not all of that information is meaningful because τ itself is not meaningful. In particular, we could use our reparameterization invariance to simply set

$$\tau = X^0(\tau) \equiv t \quad (1.3)$$

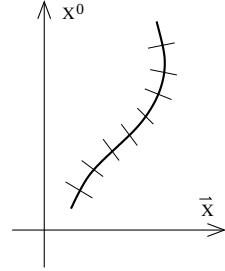


Figure 3:

If we plug this choice into the action (1.2) then we recover our initial action (1.1). The reparameterization invariance is a *gauge symmetry* of the system. Like all gauge symmetries, it's not really a symmetry at all. Rather, it is a redundancy in our description. In the present case, it means that although we seem to have D degrees of freedom X^μ , one of them is fake.

The fact that one of the degrees of freedom is a fake also shows up if we look at the momenta,

$$p_\mu = \frac{\partial L}{\partial \dot{X}^\mu} = \frac{m \dot{X}^\nu \eta_{\mu\nu}}{\sqrt{-\dot{X}^\lambda \dot{X}^\rho \eta_{\lambda\rho}}} \quad (1.4)$$

These momenta aren't all independent. They satisfy

$$p_\mu p^\mu + m^2 = 0 \quad (1.5)$$

This is a constraint on the system. It is, of course, the mass-shell constraint for a relativistic particle of mass m . From the worldline perspective, it tells us that the particle isn't allowed to sit still in Minkowski space: at the very least, it had better keep moving in a timelike direction with $(p^0)^2 \geq m^2$.

One advantage of the action (1.2) is that the Poincaré symmetry of the particle is now manifest, appearing as a global symmetry on the worldline

$$X^\mu \rightarrow \Lambda^\mu_\nu X^\nu + c^\mu \quad (1.6)$$

where Λ is a Lorentz transformation satisfying $\Lambda^\mu_\nu \eta^{\nu\rho} \Lambda^\sigma_\rho = \eta^{\mu\sigma}$, while c^μ corresponds to a constant translation. We have made all the symmetries manifest at the price of introducing a gauge symmetry into our system. A similar gauge symmetry will arise in the relativistic string and much of this course will be devoted to understanding its consequences.

1.1.1 Quantization

It's a trivial matter to quantize this action. We introduce a wavefunction $\Psi(X)$. This satisfies the usual Schrödinger equation,

$$i \frac{\partial \Psi}{\partial \tau} = H \Psi .$$

But, computing the Hamiltonian $H = \dot{X}^\mu p_\mu - L$, we find that it vanishes: $H = 0$. This shouldn't be surprising. It is simply telling us that the wavefunction doesn't depend on

τ . Since the wavefunction is something physical while, as we have seen, τ is not, this is to be expected. Note that this doesn't mean that time has dropped out of the problem. On the contrary, in this relativistic context, time X^0 is an operator, just like the spatial coordinates \vec{x} . This means that the wavefunction Ψ is immediately a function of space and time. It is not like a static state in quantum mechanics, but more akin to the fully integrated solution to the non-relativistic Schrödinger equation.

The classical system has a constraint given by (1.5). In the quantum theory, we impose this constraint as an operator equation on the wavefunction, namely $(p^\mu p_\mu + m^2)\Psi = 0$. Using the usual representation of the momentum operator $p_\mu = -i\partial/\partial X^\mu$, we recognize this constraint as the Klein-Gordon equation

$$\left(-\frac{\partial}{\partial X^\mu} \frac{\partial}{\partial X^\nu} \eta^{\mu\nu} + m^2 \right) \Psi(X) = 0 \quad (1.7)$$

Although this equation is familiar from field theory, it's important to realize that the interpretation is somewhat different. In relativistic field theory, the Klein-Gordon equation is the equation of motion obeyed by a scalar field. In relativistic quantum mechanics, it is the equation obeyed by the wavefunction. In the early days of field theory, the fact that these two equations are the same led people to think one should view the wavefunction as a classical field and quantize it a second time. This isn't correct, but nonetheless the language has stuck and it is common to talk about the point particle perspective as "first quantization" and the field theory perspective as "second quantization".

So far we've considered only a free point particle. How can we introduce interactions into this framework? We would have to first decide which interactions are allowed: perhaps the particle can split into two; perhaps it can fuse with other particles? Obviously, there is a huge range of options for us to choose from. We would then assign amplitudes for these processes to happen. There would be certain restrictions coming from the requirement of unitarity which, among other things, would lead to the necessity of anti-particles. We could draw diagrams associated to the different interactions — an example is given in the figure — and in this manner we would slowly build up the Feynman diagram expansion that is familiar from field theory. In fact, this was pretty much the way Feynman himself approached the topic of QED. However, in practice we rarely construct particle interactions in this way because the field theory framework provides a much better way of looking at things. In contrast, this way of building up interactions is exactly what we will later do for strings.

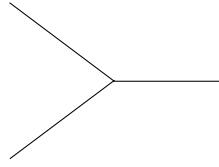


Figure 4:

1.1.2 Ein Einbein

There is another action that describes the relativistic point particle. We introduce yet another field on the worldline, $e(\tau)$, and write

$$S = \frac{1}{2} \int d\tau \left(e^{-1} \dot{X}^2 - em^2 \right) , \quad (1.8)$$

where we've used the notation $\dot{X}^2 = \dot{X}^\mu \dot{X}^\nu \eta_{\mu\nu}$. For the rest of these lectures, terms like X^2 will always mean an implicit contraction with the spacetime Minkowski metric.

This form of the action makes it look as if we have coupled the worldline theory to 1d gravity, with the field $e(\tau)$ acting as an einbein (in the sense of vierbeins that are introduced in general relativity). To see this, note that we could change notation and write this action in the more suggestive form

$$S = -\frac{1}{2} \int d\tau \sqrt{-g_{\tau\tau}} \left(g^{\tau\tau} \dot{X}^2 + m^2 \right) . \quad (1.9)$$

where $g_{\tau\tau} = (g^{\tau\tau})^{-1}$ is the metric on the worldline and $e = \sqrt{-g_{\tau\tau}}$

Although our action appears to have one more degree of freedom, e , it can be easily checked that it has the same equations of motion as (1.2). The reason for this is that e is completely fixed by its equation of motion, $\dot{X}^2 + e^2 m^2 = 0$. Substituting this into the action (1.8) recovers (1.2)

The action (1.8) has a couple of advantages over (1.2). Firstly, it works for massless particles with $m = 0$. Secondly, the absence of the annoying square root means that it's easier to quantize in a path integral framework.

The action (1.8) retains invariance under reparameterizations which are now written in a form that looks more like general relativity. For transformations parameterized by an infinitesimal η , we have

$$\tau \rightarrow \tilde{\tau} = \tau - \eta(\tau) , \quad \delta e = \frac{d}{d\tau}(\eta(\tau)e) , \quad \delta X^\mu = \frac{dX^\mu}{d\tau} \eta(\tau) \quad (1.10)$$

The einbein e transforms as a density on the worldline, while each of the coordinates X^μ transforms as a worldline scalar.

1.2 The Nambu-Goto Action

A particle sweeps out a worldline in Minkowski space. A string sweeps out a *worldsheet*. We'll parameterize this worldsheet by one timelike coordinate τ , and one spacelike coordinate σ . In this section we'll focus on closed strings and take σ to be periodic, with range

$$\sigma \in [0, 2\pi) . \quad (1.11)$$

We will sometimes package the two worldsheet coordinates together as $\sigma^\alpha = (\tau, \sigma)$, $\alpha = 0, 1$. Then the string sweeps out a surface in spacetime which defines a map from the worldsheet to Minkowski space, $X^\mu(\sigma, \tau)$ with $\mu = 0, \dots, D - 1$. For closed strings, we require

$$X^\mu(\sigma, \tau) = X^\mu(\sigma + 2\pi, \tau) .$$

In this context, spacetime is sometimes referred to as the *target space* to distinguish it from the worldsheet.

We need an action that describes the dynamics of this string. The key property that we will ask for is that nothing depends on the coordinates σ^α that we choose on the worldsheet. In other words, the string action should be reparameterization invariant. What kind of action does the trick? Well, for the point particle the action was proportional to the length of the worldline. The obvious generalization is that the action for the string should be proportional to the area, A , of the worldsheet. This is certainly a property that is characteristic of the worldsheet itself, rather than any choice of parameterization.

How do we find the area A in terms of the coordinates $X^\mu(\sigma, \tau)$? The worldsheet is a curved surface embedded in spacetime. The induced metric, $\gamma_{\alpha\beta}$, on this surface is the pull-back of the flat metric on Minkowski space,

$$\gamma_{\alpha\beta} = \frac{\partial X^\mu}{\partial \sigma^\alpha} \frac{\partial X^\nu}{\partial \sigma^\beta} \eta_{\mu\nu} . \quad (1.12)$$

Then the action which is proportional to the area of the worldsheet is given by,

$$S = -T \int d^2\sigma \sqrt{-\det \gamma} . \quad (1.13)$$

Here T is a constant of proportionality. We will see shortly that it is the *tension* of the string, meaning the mass per unit length.

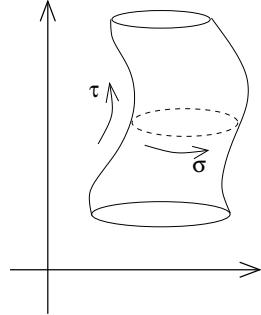


Figure 5:

We can write this action a little more explicitly. The pull-back of the metric is given by,

$$\gamma_{\alpha\beta} = \begin{pmatrix} \dot{X}^2 & \dot{X} \cdot X' \\ \dot{X} \cdot X' & X'^2 \end{pmatrix} .$$

where $\dot{X}^\mu = \partial X^\mu / \partial \tau$ and $X^{\mu'} = \partial X^\mu / \partial \sigma$. The action then takes the form,

$$S = -T \int d^2\sigma \sqrt{-(\dot{X})^2 (X')^2 + (\dot{X} \cdot X')^2} . \quad (1.14)$$

This is the *Nambu-Goto* action for a relativistic string.

Action = Area: A Check

If you're unfamiliar with differential geometry, the argument about the pull-back of the metric may be a bit slick. Thankfully, there's a more pedestrian way to see that the action (1.14) is equal to the area swept out by the worldsheet. It's slightly simpler to make this argument for a surface embedded in Euclidean space rather than Minkowski space. We choose some parameterization of the sheet in terms of τ and σ , as drawn in the figure, and we write the coordinates of Euclidean space as $\vec{X}(\sigma, \tau)$. We'll compute the area of the infinitesimal shaded region. The vectors tangent to the boundary are,

$$\vec{dl}_1 = \frac{\partial \vec{X}}{\partial \sigma} , \quad \vec{dl}_2 = \frac{\partial \vec{X}}{\partial \tau} .$$

If the angle between these two vectors is θ , then the area is then given by

$$ds^2 = |\vec{dl}_1| |\vec{dl}_2| \sin \theta = \sqrt{dl_1^2 dl_2^2 (1 - \cos^2 \theta)} = \sqrt{dl_1^2 dl_2^2 - (\vec{dl}_1 \cdot \vec{dl}_2)^2} \quad (1.15)$$

which indeed takes the form of the integrand of (1.14).

Tension and Dimension

Let's now see that T has the physical interpretation of tension. We write Minkowski coordinates as $X^\mu = (t, \vec{x})$. We work in a gauge with $X^0 \equiv t = R\tau$, where R is a constant that is needed to balance up dimensions (see below) and will drop out at the end of the argument. Consider a snapshot of a string configuration at a time when

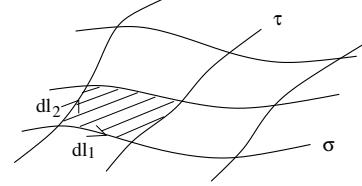


Figure 6:

$d\vec{x}/d\tau = 0$ so that the instantaneous kinetic energy vanishes. Evaluating the action for a time dt gives

$$S = -T \int d\tau d\sigma R \sqrt{(d\vec{x}/d\sigma)^2} = -T \int dt (\text{spatial length of string}) . \quad (1.16)$$

But, when the kinetic energy vanishes, the action is proportional to the time integral of the potential energy,

$$\text{potential energy} = T \times (\text{spatial length of string}) .$$

So T is indeed the energy per unit length as claimed. We learn that the string acts rather like an elastic band and its energy increases linearly with length. (This is different from the elastic bands you're used to which obey Hooke's law where energy increased quadratically with length). To minimize its potential energy, the string will want to shrink to zero size. We'll see that when we include quantum effects this can't happen because of the usual zero point energies.

There is a slightly annoying way of writing the tension that has its origin in ancient history, but is commonly used today

$$T = \frac{1}{2\pi\alpha'} \quad (1.17)$$

where α' is pronounced “alpha-prime”. In the language of our ancestors, α' is referred to as the “universal Regge slope”. We'll explain why later in this course.

At this point, it's worth pointing out some conventions that we have, until now, left implicit. The spacetime coordinates have dimension $[X] = -1$. In contrast, the worldsheet coordinates are taken to be dimensionless, $[\sigma] = 0$. (This can be seen in our identification $\sigma \equiv \sigma + 2\pi$). The tension is equal to the mass per unit length and has dimension $[T] = 2$. Obviously this means that $[\alpha'] = -2$. We can therefore associate a length scale, l_s , by

$$\alpha' = l_s^2 \quad (1.18)$$

The *string scale* l_s is the natural length that appears in string theory. In fact, in a certain sense (that we will make more precise later in the course) this length scale is the only parameter of the theory.

Actual Strings vs. Fundamental Strings

There are several situations in Nature where string-like objects arise. Prime examples include magnetic flux tubes in superconductors and chromo-electric flux tubes in QCD. Cosmic strings, a popular speculation in cosmology, are similar objects, stretched across the sky. In each of these situations, there are typically two length scales associated to the string: the tension, T and the width of the string, L . For all these objects, the dynamics is governed by the Nambu-Goto action as long as the curvature of the string is much greater than L . (In the case of superconductors, one should work with a suitable non-relativistic version of the Nambu-Goto action).

However, in each of these other cases, the Nambu-Goto action is not the end of the story. There will typically be additional terms in the action that depend on the width of the string. The form of these terms is not universal, but often includes a *rigidity* piece of form $L \int K^2$, where K is the extrinsic curvature of the worldsheet. Other terms could be added to describe fluctuations in the width of the string.

The string scale, l_s , or equivalently the tension, T , depends on the kind of string that we're considering. For example, if we're interested in QCD flux tubes then we would take

$$T \sim (1 \text{ GeV})^2 \quad (1.19)$$

In this course we will consider *fundamental strings* which have zero width. What this means in practice is that we take the Nambu-Goto action as the complete description for all configurations of the string. These strings will have relevance to quantum gravity and the tension of the string is taken to be much larger, typically an order of magnitude or so below the Planck scale.

$$T \lesssim M_{pl}^2 = (10^{18} \text{ GeV})^2 \quad (1.20)$$

However, I should point out that when we try to view string theory as a fundamental theory of quantum gravity, we don't really know what value T should take. As we will see later in this course, it depends on many other aspects, most notably the string coupling and the volume of the extra dimensions.

1.2.1 Symmetries of the Nambu-Goto Action

The Nambu-Goto action has two types of symmetry, each of a different nature.

- Poincaré invariance of the spacetime (1.6). This is a global symmetry from the perspective of the worldsheet, meaning that the parameters Λ^μ_ν and c^μ which label

the symmetry transformation are constants and do not depend on worldsheet coordinates σ^α .

- Reparameterization invariance, $\sigma^\alpha \rightarrow \tilde{\sigma}^\alpha(\sigma)$. As for the point particle, this is a gauge symmetry. It reflects the fact that we have a redundancy in our description because the worldsheet coordinates σ^α have no physical meaning.

1.2.2 Equations of Motion

To derive the equations of motion for the Nambu-Goto string, we first introduce the momenta which we call Π because there will be countless other quantities that we want to call p later,

$$\begin{aligned}\Pi_\mu^\tau &= \frac{\partial \mathcal{L}}{\partial \dot{X}^\mu} = -T \frac{(\dot{X} \cdot X') X'_\mu - (X'^2) \dot{X}_\mu}{\sqrt{(\dot{X} \cdot X')^2 - \dot{X}^2 X'^2}} . \\ \Pi_\mu^\sigma &= \frac{\partial \mathcal{L}}{\partial X'^\mu} = -T \frac{(\dot{X} \cdot X') \dot{X}_\mu - (\dot{X}^2) X'_\mu}{\sqrt{(\dot{X} \cdot X')^2 - \dot{X}^2 X'^2}} .\end{aligned}$$

The equations of motion are then given by,

$$\frac{\partial \Pi_\mu^\tau}{\partial \tau} + \frac{\partial \Pi_\mu^\sigma}{\partial \sigma} = 0$$

These look like nasty, non-linear equations. In fact, there's a slightly nicer way to write these equations, starting from the earlier action (1.13). Recall that the variation of a determinant is $\delta \sqrt{-\gamma} = \frac{1}{2} \sqrt{-\gamma} \gamma^{\alpha\beta} \delta \gamma_{\alpha\beta}$. Using the definition of the pull-back metric $\gamma_{\alpha\beta}$, this gives rise to the equations of motion

$$\partial_\alpha (\sqrt{-\det \gamma} \gamma^{\alpha\beta} \partial_\beta X^\mu) = 0 , \quad (1.21)$$

Although this notation makes the equations look a little nicer, we're kidding ourselves. Written in terms of X^μ , they are still the same equations. Still nasty.

1.3 The Polyakov Action

The square-root in the Nambu-Goto action means that it's rather difficult to quantize using path integral techniques. However, there is another form of the string action which is classically equivalent to the Nambu-Goto action. It eliminates the square root at the expense of introducing another field,

$$S = -\frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{-g} g^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu \eta_{\mu\nu} \quad (1.22)$$

where $g \equiv \det g$. This is the *Polyakov* action. (Polyakov didn't discover the action, but he understood how to work with it in the path integral and for this reason it carries his name. The path integral treatment of this action will be the subject of Chapter 5).

The new field is $g_{\alpha\beta}$. It is a dynamical metric on the worldsheet. From the perspective of the worldsheet, the Polyakov action is a bunch of scalar fields X coupled to 2d gravity.

The equation of motion for X^μ is

$$\partial_\alpha(\sqrt{-g}g^{\alpha\beta}\partial_\beta X^\mu) = 0 , \quad (1.23)$$

which coincides with the equation of motion (1.21) from the Nambu-Goto action, except that $g_{\alpha\beta}$ is now an independent variable which is fixed by its own equation of motion. To determine this, we vary the action (remembering again that $\delta\sqrt{-g} = -\frac{1}{2}\sqrt{-g}g_{\alpha\beta}\delta g^{\alpha\beta} = +\frac{1}{2}\sqrt{-g}g^{\alpha\beta}\delta g_{\alpha\beta}$),

$$\delta S = -\frac{T}{2} \int d^2\sigma \delta g^{\alpha\beta} (\sqrt{-g} \partial_\alpha X^\mu \partial_\beta X^\nu - \frac{1}{2}\sqrt{-g} g_{\alpha\beta} g^{\rho\sigma} \partial_\rho X^\mu \partial_\sigma X^\nu) \eta_{\mu\nu} = 0 . \quad (1.24)$$

The worldsheet metric is therefore given by,

$$g_{\alpha\beta} = 2f(\sigma) \partial_\alpha X \cdot \partial_\beta X , \quad (1.25)$$

where the function $f(\sigma)$ is given by,

$$f^{-1} = g^{\rho\sigma} \partial_\rho X \cdot \partial_\sigma X$$

A comment on the potentially ambiguous notation: here, and below, any function $f(\sigma)$ is always short-hand for $f(\sigma, \tau)$: it in no way implies that f depends only on the spatial worldsheet coordinate.

We see that $g_{\alpha\beta}$ isn't quite the same as the pull-back metric $\gamma_{\alpha\beta}$ defined in equation (1.12); the two differ by the conformal factor f . However, this doesn't matter because, rather remarkably, f drops out of the equation of motion (1.23). This is because the $\sqrt{-g}$ term scales as f , while the inverse metric $g^{\alpha\beta}$ scales as f^{-1} and the two pieces cancel. We therefore see that Nambu-Goto and the Polyakov actions result in the same equation of motion for X .

In fact, we can see more directly that the Nambu-Goto and Polyakov actions coincide. We may replace $g_{\alpha\beta}$ in the Polyakov action (1.22) with its equation of motion $g_{\alpha\beta} = 2f\gamma_{\alpha\beta}$. The factor of f also drops out of the action for the same reason that it dropped out of the equation of motion. In this manner, we recover the Nambu-Goto action (1.13).

1.3.1 Symmetries of the Polyakov Action

The fact that the presence of the factor $f(\sigma, \tau)$ in (1.25) didn't actually affect the equations of motion for X^μ reflects the existence of an extra symmetry which the Polyakov action enjoys. Let's look more closely at this. Firstly, the Polyakov action still has the two symmetries of the Nambu-Goto action,

- Poincaré invariance. This is a global symmetry on the worldsheet.

$$X^\mu \rightarrow \Lambda^\mu_\nu X^\nu + c^\mu .$$

- Reparameterization invariance, also known as diffeomorphisms. This is a gauge symmetry on the worldsheet. We may redefine the worldsheet coordinates as $\sigma^\alpha \rightarrow \tilde{\sigma}^\alpha(\sigma)$. The fields X^μ transform as worldsheet scalars, while $g_{\alpha\beta}$ transforms in the manner appropriate for a 2d metric.

$$\begin{aligned} X^\mu(\sigma) &\rightarrow \tilde{X}^\mu(\tilde{\sigma}) = X^\mu(\sigma) \\ g_{\alpha\beta}(\sigma) &\rightarrow \tilde{g}_{\alpha\beta}(\tilde{\sigma}) = \frac{\partial \sigma^\gamma}{\partial \tilde{\sigma}^\alpha} \frac{\partial \sigma^\delta}{\partial \tilde{\sigma}^\beta} g_{\gamma\delta}(\sigma) \end{aligned}$$

It will sometimes be useful to work infinitesimally. If we make the coordinate change $\sigma^\alpha \rightarrow \tilde{\sigma}^\alpha = \sigma^\alpha - \eta^\alpha(\sigma)$, for some small η . The transformations of the fields then become,

$$\begin{aligned} \delta X^\mu(\sigma) &= \eta^\alpha \partial_\alpha X^\mu \\ \delta g_{\alpha\beta}(\sigma) &= \nabla_\alpha \eta_\beta + \nabla_\beta \eta_\alpha \end{aligned}$$

where the covariant derivative is defined by $\nabla_\alpha \eta_\beta = \partial_\alpha \eta_\beta - \Gamma_{\alpha\beta}^\sigma \eta_\sigma$ with the Levi-Civita connection associated to the worldsheet metric given by the usual expression,

$$\Gamma_{\alpha\beta}^\sigma = \frac{1}{2} g^{\sigma\rho} (\partial_\alpha g_{\beta\rho} + \partial_\beta g_{\rho\alpha} - \partial_\rho g_{\alpha\beta})$$

Together with these familiar symmetries, there is also a new symmetry which is novel to the Polyakov action. It is called *Weyl invariance*.

- Weyl Invariance. Under this symmetry, $X^\mu(\sigma) \rightarrow X^\mu(\sigma)$, while the metric changes as

$$g_{\alpha\beta}(\sigma) \rightarrow \Omega^2(\sigma) g_{\alpha\beta}(\sigma) . \quad (1.26)$$

Or, infinitesimally, we can write $\Omega^2(\sigma) = e^{2\phi(\sigma)}$ for small ϕ so that

$$\delta g_{\alpha\beta}(\sigma) = 2\phi(\sigma) g_{\alpha\beta}(\sigma) .$$

It is simple to see that the Polyakov action is invariant under this transformation: the factor of Ω^2 drops out just as the factor of f did in equation (1.25), canceling between $\sqrt{-g}$ and the inverse metric $g^{\alpha\beta}$. This is a gauge symmetry of the string, as seen by the fact that the parameter Ω depends on the worldsheet coordinates σ . This means that two metrics which are related by a Weyl transformation (1.26) are to be considered as the same physical state.

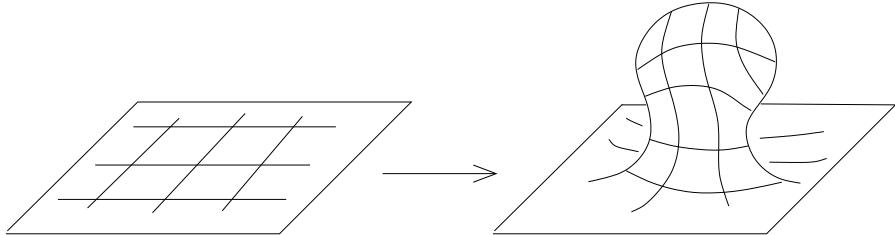


Figure 7: An example of a Weyl transformation

How should we think of Weyl invariance? It is not a coordinate change. Instead it is the invariance of the theory under a local change of scale which preserves the angles between all lines. For example the two worldsheet metrics shown in the figure are viewed by the Polyakov string as equivalent. This is rather surprising! And, as you might imagine, theories with this property are extremely rare. It should be clear from the discussion above that the property of Weyl invariance is special to two dimensions, for only there does the scaling factor coming from the determinant $\sqrt{-g}$ cancel that coming from the inverse metric. But even in two dimensions, if we wish to keep Weyl invariance then we are strictly limited in the kind of interactions that can be added to the action. For example, we would not be allowed a potential term for the worldsheet scalars of the form,

$$\int d^2\sigma \sqrt{-g} V(X) .$$

These break Weyl invariance. Nor can we add a worldsheet cosmological constant term,

$$\mu \int d^2\sigma \sqrt{-g} .$$

This too breaks Weyl invariance. We will see later in this course that the requirement of Weyl invariance becomes even more stringent in the quantum theory. We will also see what kind of interactions terms can be added to the worldsheet. Indeed, much of this course can be thought of as the study of theories with Weyl invariance.

1.3.2 Fixing a Gauge

As we have seen, the equation of motion (1.23) looks pretty nasty. However, we can use the redundancy inherent in the gauge symmetry to choose coordinates in which they simplify. Let's think about what we can do with the gauge symmetry.

Firstly, we have two reparameterizations to play with. The worldsheet metric has three independent components. This means that we expect to be able to set any two of the metric components to a value of our choosing. We will choose to make the metric locally conformally flat, meaning

$$g_{\alpha\beta} = e^{2\phi} \eta_{\alpha\beta}, \quad (1.27)$$

where $\phi(\sigma, \tau)$ is some function on the worldsheet. You can check that this is possible by writing down the change of the metric under a coordinate transformation and seeing that the differential equations which result from the condition (1.27) have solutions, at least locally. Choosing a metric of the form (1.27) is known as *conformal gauge*.

We have only used reparameterization invariance to get to the metric (1.27). We still have Weyl transformations to play with. Clearly, we can use these to remove the last independent component of the metric and set $\phi = 0$ such that,

$$g_{\alpha\beta} = \eta_{\alpha\beta}. \quad (1.28)$$

We end up with the flat metric on the worldsheet in Minkowski coordinates.

A Diversion: How to make a metric flat

The fact that we can use Weyl invariance to make any two-dimensional metric flat is an important result. Let's take a quick diversion from our main discussion to see a different proof that isn't tied to the choice of Minkowski coordinates on the worldsheet. We'll work in 2d Euclidean space to avoid annoying minus signs. Consider two metrics related by a Weyl transformation, $g'_{\alpha\beta} = e^{2\phi} g_{\alpha\beta}$. One can check that the Ricci scalars of the two metrics are related by,

$$\sqrt{g'} R' = \sqrt{g}(R - 2\nabla^2 \phi). \quad (1.29)$$

We can therefore pick a ϕ such that the new metric has vanishing Ricci scalar, $R' = 0$, simply by solving this differential equation for ϕ . However, in two dimensions (but not in higher dimensions) a vanishing Ricci scalar implies a flat metric. The reason is simply that there aren't too many indices to play with. In particular, symmetry of the Riemann tensor in two dimensions means that it must take the form,

$$R_{\alpha\beta\gamma\delta} = \frac{R}{2}(g_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}g_{\beta\gamma}).$$

So $R' = 0$ is enough to ensure that $R'_{\alpha\beta\gamma\delta} = 0$, which means that the manifold is flat. In equation (1.28), we've further used reparameterization invariance to pick coordinates in which the flat metric is the Minkowski metric.

The equations of motion and the stress-energy tensor

With the choice of the flat metric (1.28), the Polyakov action simplifies tremendously and becomes the theory of D free scalar fields. (In fact, this simplification happens in any conformal gauge).

$$S = -\frac{1}{4\pi\alpha'} \int d^2\sigma \partial_\alpha X \cdot \partial^\alpha X , \quad (1.30)$$

and the equations of motion for X^μ reduce to the free wave equation,

$$\partial_\alpha \partial^\alpha X^\mu = 0 . \quad (1.31)$$

Now that looks too good to be true! Are the horrible equations (1.23) really equivalent to a free wave equation? Well, not quite. There is something that we've forgotten: we picked a choice of gauge for the metric $g_{\alpha\beta}$. But we must still make sure that the equation of motion for $g_{\alpha\beta}$ is satisfied. In fact, the variation of the action with respect to the metric gives rise to a rather special quantity: it is the stress-energy tensor, $T_{\alpha\beta}$. With a particular choice of normalization convention, we define the stress-energy tensor to be

$$T_{\alpha\beta} = -\frac{2}{T} \frac{1}{\sqrt{-g}} \frac{\partial S}{\partial g^{\alpha\beta}} .$$

We varied the Polyakov action with respect to $g_{\alpha\beta}$ in (1.24). When we set $g_{\alpha\beta} = \eta_{\alpha\beta}$ we get

$$T_{\alpha\beta} = \partial_\alpha X \cdot \partial_\beta X - \frac{1}{2}\eta_{\alpha\beta}\eta^{\rho\sigma}\partial_\rho X \cdot \partial_\sigma X . \quad (1.32)$$

The equation of motion associated to the metric $g_{\alpha\beta}$ is simply $T_{\alpha\beta} = 0$. Or, more explicitly,

$$\begin{aligned} T_{01} &= \dot{X} \cdot X' = 0 \\ T_{00} = T_{11} &= \frac{1}{2}(\dot{X}^2 + X'^2) = 0 . \end{aligned} \quad (1.33)$$

We therefore learn that the equations of motion of the string are the free wave equations (1.31) subject to the two constraints (1.33) arising from the equation of motion $T_{\alpha\beta} = 0$.

Getting a feel for the constraints

Let's try to get some intuition for these constraints. There is a simple meaning of the first constraint in (1.33): we must choose our parameterization such that lines of constant σ are perpendicular to the lines of constant τ , as shown in the figure.

But we can do better. To gain more physical insight, we need to make use of the fact that we haven't quite exhausted our gauge symmetry. We will discuss this more in Section 2.2, but for now one can check that there is enough remnant gauge symmetry to allow us to go to static gauge,

$$X^0 \equiv t = R\tau ,$$

so that $(X^0)' = 0$ and $\dot{X}^0 = R$, where R is a constant that is needed on dimensional grounds. The interpretation of this constant will become clear shortly. Then, writing $X^\mu = (t, \vec{x})$, the equation of motion for spatial components is the free wave equation,

$$\ddot{\vec{x}} - \vec{x}'' = 0$$

while the constraints become

$$\begin{aligned} \dot{\vec{x}} \cdot \vec{x}' &= 0 \\ \dot{\vec{x}}^2 + \vec{x}'^2 &= R^2 \end{aligned} \tag{1.34}$$

The first constraint tells us that the motion of the string must be perpendicular to the string itself. In other words, the physical modes of the string are transverse oscillations. There is no longitudinal mode. We'll also see this again in Section 2.2.

From the second constraint, we can understand the meaning of the constant R : it is related to the length of the string when $\dot{\vec{x}} = 0$,

$$\int d\sigma \sqrt{(d\vec{x}/d\sigma)^2} = 2\pi R .$$

Of course, if we have a stretched string with $\dot{\vec{x}} = 0$ at one moment of time, then it won't stay like that for long. It will contract under its own tension. As this happens, the second constraint equation relates the length of the string to the instantaneous velocity of the string.

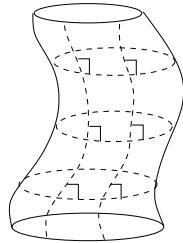


Figure 8:

1.4 Mode Expansions

Let's look at the equations of motion and constraints more closely. The equations of motion (1.31) are easily solved. We introduce lightcone coordinates on the worldsheet,

$$\sigma^\pm = \tau \pm \sigma ,$$

in terms of which the equations of motion simply read

$$\partial_+ \partial_- X^\mu = 0$$

The most general solution is,

$$X^\mu(\sigma, \tau) = X_L^\mu(\sigma^+) + X_R^\mu(\sigma^-)$$

for arbitrary functions X_L^μ and X_R^μ . These describe left-moving and right-moving waves respectively. Of course the solution must still obey both the constraints (1.33) as well as the periodicity condition,

$$X^\mu(\sigma, \tau) = X^\mu(\sigma + 2\pi, \tau) . \quad (1.35)$$

The most general, periodic solution can be expanded in Fourier modes,

$$\begin{aligned} X_L^\mu(\sigma^+) &= \tfrac{1}{2}x^\mu + \tfrac{1}{2}\alpha' p^\mu \sigma^+ + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \tilde{\alpha}_n^\mu e^{-in\sigma^+} , \\ X_R^\mu(\sigma^-) &= \tfrac{1}{2}x^\mu + \tfrac{1}{2}\alpha' p^\mu \sigma^- + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^\mu e^{-in\sigma^-} . \end{aligned} \quad (1.36)$$

This mode expansion will be very important when we come to the quantum theory. Let's make a few simple comments here.

- Various normalizations in this expression, such as the α' and factor of $1/n$ have been chosen for later convenience.
- X_L and X_R do not individually satisfy the periodicity condition (1.35) due to the terms linear in σ^\pm . However, the sum of them is invariant under $\sigma \rightarrow \sigma + 2\pi$ as required.
- The variables x^μ and p^μ are the position and momentum of the center of mass of the string. This can be checked, for example, by studying the Noether currents arising from the spacetime translation symmetry $X^\mu \rightarrow X^\mu + c^\mu$. One finds that the conserved charge is indeed p^μ .
- Reality of X^μ requires that the coefficients of the Fourier modes, α_n^μ and $\tilde{\alpha}_n^\mu$, obey

$$\alpha_n^\mu = (\alpha_{-n}^\mu)^* , \quad \tilde{\alpha}_n^\mu = (\tilde{\alpha}_{-n}^\mu)^* . \quad (1.37)$$

1.4.1 The Constraints Revisited

We still have to impose the two constraints (1.33). In the worldsheet lightcone coordinates σ^\pm , these become,

$$(\partial_+ X)^2 = (\partial_- X)^2 = 0 . \quad (1.38)$$

These equations give constraints on the momenta p^μ and the Fourier modes α_n^μ and $\tilde{\alpha}_n^\mu$. To see what these are, let's look at

$$\begin{aligned} \partial_- X^\mu &= \partial_- X_R^\mu = \frac{\alpha'}{2} p^\mu + \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \alpha_n^\mu e^{-in\sigma^-} \\ &= \sqrt{\frac{\alpha'}{2}} \sum_n \alpha_n^\mu e^{-in\sigma^-} \end{aligned}$$

where in the second line the sum is over all $n \in \mathbf{Z}$ and we have defined α_0^μ to be

$$\alpha_0^\mu \equiv \sqrt{\frac{\alpha'}{2}} p^\mu .$$

The constraint (1.38) can then be written as

$$\begin{aligned} (\partial_- X)^2 &= \frac{\alpha'}{2} \sum_{m,p} \alpha_m \cdot \alpha_p e^{-i(m+p)\sigma^-} \\ &= \frac{\alpha'}{2} \sum_{m,n} \alpha_m \cdot \alpha_{n-m} e^{-in\sigma^-} \\ &\equiv \alpha' \sum_n L_n e^{-in\sigma^-} = 0 . \end{aligned}$$

where we have defined the sum of oscillator modes,

$$L_n = \frac{1}{2} \sum_m \alpha_{n-m} \cdot \alpha_m . \quad (1.39)$$

We can also do the same for the left-moving modes, where we again define an analogous sum of operator modes,

$$\tilde{L}_n = \frac{1}{2} \sum_m \tilde{\alpha}_{n-m} \cdot \tilde{\alpha}_m . \quad (1.40)$$

with the zero mode defined to be,

$$\tilde{\alpha}_0^\mu \equiv \sqrt{\frac{\alpha'}{2}} p^\mu .$$

The fact that $\tilde{\alpha}_0^\mu = \alpha_0^\mu$ looks innocuous but is a key point to remember when we come to quantize the string. The L_n and \tilde{L}_n are the Fourier modes of the constraints. Any classical solution of the string of the form (1.36) must further obey the infinite number of constraints,

$$L_n = \tilde{L}_n = 0 \quad n \in \mathbf{Z} .$$

We'll meet these objects L_n and \tilde{L}_n again in a more general context when we come to discuss conformal field theory.

The constraints arising from L_0 and \tilde{L}_0 have a rather special interpretation. This is because they include the square of the spacetime momentum p^μ . But, the square of the spacetime momentum is an important quantity in Minkowski space: it is the square of the rest mass of a particle,

$$p_\mu p^\mu = -M^2 .$$

So the L_0 and \tilde{L}_0 constraints tell us the effective mass of a string in terms of the excited oscillator modes, namely

$$M^2 = \frac{4}{\alpha'} \sum_{n>0} \alpha_n \cdot \alpha_{-n} = \frac{4}{\alpha'} \sum_{n>0} \tilde{\alpha}_n \cdot \tilde{\alpha}_{-n} \tag{1.41}$$

Because both α_0^μ and $\tilde{\alpha}_0^\mu$ are equal to $\sqrt{\alpha'/2} p^\mu$, we have two expressions for the invariant mass: one in terms of right-moving oscillators α_n^μ and one in terms of left-moving oscillators $\tilde{\alpha}_n^\mu$. And these two terms must be equal to each other. This is known as *level matching*. It will play an important role in the next section where we turn to the quantum theory.

2. The Quantum String

Our goal in this section is to quantize the string. We have seen that the string action involves a gauge symmetry and whenever we wish to quantize a gauge theory we're presented with a number of different ways in which we can proceed. If we're working in the canonical formalism, this usually boils down to one of two choices:

- We could first quantize the system and then subsequently impose the constraints that arise from gauge fixing as operator equations on the physical states of the system. For example, in QED this is the Gupta-Bleuler method of quantization that we use in Lorentz gauge. In string theory it consists of treating all fields X^μ , including time X^0 , as operators and imposing the constraint equations (1.33) on the states. This is usually called covariant quantization.
- The alternative method is to first solve all of the constraints of the system to determine the space of physically distinct classical solutions. We then quantize these physical solutions. For example, in QED, this is the way we proceed in Coulomb gauge. Later in this chapter, we will see a simple way to solve the constraints of the free string.

Of course, if we do everything correctly, the two methods should agree. Usually, each presents a slightly different challenge and offers a different viewpoint.

In these lectures, we'll take a brief look at the first method of covariant quantization. However, at the slightest sign of difficulties, we'll bail! It will be useful enough to see where the problems lie. We'll then push forward with the second method described above which is known as lightcone quantization in string theory. Although we'll succeed in pushing quantization through to the end, our derivations will be a little cheap and unsatisfactory in places. In Section 5 we'll return to all these issues, armed with more sophisticated techniques from conformal field theory.

2.1 A Lightning Look at Covariant Quantization

We wish to quantize D free scalar fields X^μ whose dynamics is governed by the action (1.30). We subsequently wish to impose the constraints

$$\dot{X} \cdot X' = \dot{X}^2 + X'^2 = 0 . \quad (2.1)$$

The first step is easy. We promote X^μ and their conjugate momenta $\Pi_\mu = (1/2\pi\alpha')\dot{X}_\mu$ to operator valued fields obeying the canonical equal-time commutation relations,

$$[X^\mu(\sigma, \tau), \Pi_\nu(\sigma', \tau)] = i\delta(\sigma - \sigma')\delta_\nu^\mu ,$$
$$[X^\mu(\sigma, \tau), X^\nu(\sigma', \tau)] = [\Pi_\mu(\sigma, \tau), \Pi_\nu(\sigma', \tau)] = 0 .$$

We translate these into commutation relations for the Fourier modes x^μ , p^μ , α_n^μ and $\tilde{\alpha}_n^\mu$. Using the mode expansion (1.36) we find

$$[x^\mu, p_\nu] = i\delta_\nu^\mu \quad \text{and} \quad [\alpha_n^\mu, \alpha_m^\nu] = [\tilde{\alpha}_n^\mu, \tilde{\alpha}_m^\nu] = n \eta^{\mu\nu} \delta_{n+m,0}, \quad (2.2)$$

with all others zero. The commutation relations for x^μ and p^μ are expected for operators governing the position and momentum of the center of mass of the string. The commutation relations of α_n^μ and $\tilde{\alpha}_n^\mu$ are those of harmonic oscillator creation and annihilation operators in disguise. And the disguise isn't that good. We just need to define (ignoring the μ index for now)

$$a_n = \frac{\alpha_n}{\sqrt{n}} \quad , \quad a_n^\dagger = \frac{\alpha_{-n}}{\sqrt{n}} \quad n > 0 \quad (2.3)$$

Then (2.2) gives the familiar $[a_n, a_m^\dagger] = \delta_{mn}$. So each scalar field gives rise to two infinite towers of creation and annihilation operators, with α_n acting as a rescaled annihilation operator for $n > 0$ and as a creation operator for $n < 0$. There are two towers because we have right-moving modes α_n and left-moving modes $\tilde{\alpha}_n$.

With these commutation relations in hand we can now start building the Fock space of our theory. We introduce a vacuum state of the string $|0\rangle$, defined to obey

$$\alpha_n^\mu |0\rangle = \tilde{\alpha}_n^\mu |0\rangle = 0 \quad \text{for } n > 0 \quad (2.4)$$

The vacuum state of string theory has a different interpretation from the analogous object in field theory. This is not the vacuum state of spacetime. It is instead the vacuum state of a single string. This is reflected in the fact that the operators x^μ and p^μ give extra structure to the vacuum. The true ground state of the string is $|0\rangle$, tensored with a spatial wavefunction $\Psi(x)$. Alternatively, if we work in momentum space, the vacuum carries another quantum number, p^μ , which is the eigenvalue of the momentum operator. We should therefore write the vacuum as $|0;p\rangle$, which still obeys (2.4), but now also

$$\hat{p}^\mu |0;p\rangle = p^\mu |0;p\rangle \quad (2.5)$$

where (for the only time in these lecture notes) we've put a hat on the momentum operator \hat{p}^μ on the left-hand side of this equation to distinguish it from the eigenvalue p^μ on the right-hand side.

We can now start to build up the Fock space by acting with creation operators α_n^μ and $\tilde{\alpha}_n^\mu$ with $n < 0$. A generic state comes from acting with any number of these creation operators on the vacuum,

$$(\alpha_{-1}^{\mu_1})^{n_{\mu_1}} (\alpha_{-2}^{\mu_2})^{n_{\mu_2}} \dots (\tilde{\alpha}_{-1}^{\nu_1})^{n_{\nu_1}} (\tilde{\alpha}_{-2}^{\nu_2})^{n_{\nu_2}} \dots |0;p\rangle$$

Each state in the Fock space is a different excited state of the string. Each has the interpretation of a different species of particle in spacetime. We'll see exactly what particles they are shortly. But for now, notice that because there's an infinite number of ways to excite a string there are an infinite number of different species of particles in this theory.

2.1.1 Ghosts

There's a problem with the Fock space that we've constructed: it doesn't have positive norm. The reason for this is that one of the scalar fields, X^0 , comes with the wrong sign kinetic term in the action (1.30). From the perspective of the commutation relations, this issue raises its head in presence of the spacetime Minkowski metric in the expression

$$[\alpha_n^\mu, \alpha_m^\nu] = n \eta^{\mu\nu} \delta_{n,m} .$$

This gives rise to the offending negative norm states, which come with an odd number of timelike oscillators excited, for example

$$\langle p'; 0 | \alpha_1^0 \alpha_{-1}^0 | 0; p \rangle \sim -\delta^D(p - p')$$

This is the first problem that arises in the covariant approach to quantization. States with negative norm are referred to as *ghosts*. To make sense of the theory, we have to make sure that they can't be produced in any physical processes. Of course, this problem is familiar from attempts to quantize QED in Lorentz gauge. In that case, gauge symmetry rides to the rescue since the ghosts are removed by imposing the gauge fixing constraint. We must hope that the same happens in string theory.

2.1.2 Constraints

Although we won't push through with this programme at the present time, let us briefly look at what kind of constraints we have in string theory. In terms of Fourier modes, the classical constraints can be written as $L_n = \tilde{L}_n = 0$, where

$$L_n = \frac{1}{2} \sum_m \alpha_{n-m} \cdot \alpha_m$$

and similar for \tilde{L}_n . As in the Gupta-Bleuler quantization of QED, we don't impose all of these as operator equations on the Hilbert space. Instead we only require that the operators L_n and \tilde{L}_n have vanishing matrix elements when sandwiched between two physical states $|\text{phys}\rangle$ and $|\text{phys}'\rangle$,

$$\langle \text{phys}' | L_n | \text{phys} \rangle = \langle \text{phys}' | \tilde{L}_n | \text{phys} \rangle = 0$$

Because $L_n^\dagger = L_{-n}$, it is therefore sufficient to require

$$L_n|\text{phys}\rangle = \tilde{L}_n|\text{phys}\rangle = 0 \quad \text{for } n > 0 \quad (2.6)$$

However, we still haven't explained how to impose the constraints L_0 and \tilde{L}_0 . And these present a problem that doesn't arise in the case of QED. The problem is that, unlike for L_n with $n \neq 0$, the operator L_0 is not uniquely defined when we pass to the quantum theory. There is an operator ordering ambiguity arising from the commutation relations (2.2). Commuting the α_n^μ operators past each other in L_0 gives rise to extra constant terms.

Question: How do we know what order to put the α_n^μ operators in the quantum operator L_0 ? Or the $\tilde{\alpha}_n^\mu$ operators in \tilde{L}_0 ?

Answer: We don't! Yet. Naively it looks as if each different choice will define a different theory when we impose the constraints. To make this ambiguity manifest, for now let's just pick a choice of ordering. We define the quantum operators to be normal ordered, with the annihilation operators α_n^i , $n > 0$, moved to the right,

$$L_0 = \sum_{m=1}^{\infty} \alpha_{-m} \cdot \alpha_m + \frac{1}{2} \alpha_0^2 \quad , \quad \tilde{L}_0 = \sum_{m=1}^{\infty} \tilde{\alpha}_{-m} \cdot \tilde{\alpha}_m + \frac{1}{2} \tilde{\alpha}_0^2$$

Then the ambiguity rears its head in the different constraint equations that we could impose, namely

$$(L_0 - a)|\text{phys}\rangle = (\tilde{L}_0 - a)|\text{phys}\rangle = 0 \quad (2.7)$$

for some constant a .

As we saw classically, the operators L_0 and \tilde{L}_0 play an important role in determining the spectrum of the string because they include a term quadratic in the momentum $\alpha_0^\mu = \tilde{\alpha}_0^\mu = \sqrt{\alpha'/2} p^\mu$. Combining the expression (1.41) with our constraint equation for L_0 and \tilde{L}_0 , we find the spectrum of the string is given by,

$$M^2 = \frac{4}{\alpha'} \left(-a + \sum_{m=1}^{\infty} \alpha_{-m} \cdot \alpha_m \right) = \frac{4}{\alpha'} \left(-a + \sum_{m=1}^{\infty} \tilde{\alpha}_{-m} \cdot \tilde{\alpha}_m \right)$$

We learn therefore that the undetermined constant a has a direct physical effect: it changes the mass spectrum of the string. In the quantum theory, the sums over α_n^μ modes are related to the number operators for the harmonic oscillator: they count the number of excited modes of the string. The level matching in the quantum theory tells us that the number of left-moving modes must equal the number of right-moving modes.

Ultimately, we will find that the need to decouple the ghosts forces us to make a unique choice for the constant a . (Spoiler alert: it turns out to be $a = 1$). In fact, the requirement that there are no ghosts is much stronger than this. It also restricts the number of scalar fields that we have in the theory. (Another spoiler: $D = 26$). If you're interested in how this works in covariant formulation then you can read about it in the book by Green, Schwarz and Witten. Instead, we'll show how to quantize the string and derive these values for a and D in lightcone gauge. However, after a trip through the world of conformal field theory, we'll come back to these ideas in a context which is closer to the covariant approach.

2.2 Lightcone Quantization

We will now take the second path described at the beginning of this section. We will try to find a parameterization of all classical solutions of the string. This is equivalent to finding the classical phase space of the theory. We do this by solving the constraints (2.1) in the classical theory, leaving behind only the physical degrees of freedom.

Recall that we fixed the gauge to set the worldsheet metric to

$$g_{\alpha\beta} = \eta_{\alpha\beta} .$$

However, this isn't the end of our gauge freedom. There still remain gauge transformations which preserve this choice of metric. In particular, any coordinate transformation $\sigma \rightarrow \tilde{\sigma}(\sigma)$ which changes the metric by

$$\eta_{\alpha\beta} \rightarrow \Omega^2(\sigma)\eta_{\alpha\beta} , \quad (2.8)$$

can be undone by a Weyl transformation. What are these coordinate transformations? It's simplest to answer this using lightcone coordinates on the worldsheet,

$$\sigma^\pm = \tau \pm \sigma , \quad (2.9)$$

where the flat metric on the worldsheet takes the form,

$$ds^2 = -d\sigma^+ d\sigma^-$$

In these coordinates, it's clear that any transformation of the form

$$\sigma^+ \rightarrow \tilde{\sigma}^+(\sigma^+) \quad , \quad \sigma^- \rightarrow \tilde{\sigma}^-(\sigma^-) , \quad (2.10)$$

simply multiplies the flat metric by an overall factor (2.8) and so can be undone by a compensating Weyl transformation. Some quick comments on this surviving gauge symmetry:

- Recall that in Section 1.3.2 we used the argument that 3 gauge invariances (2 reparameterizations + 1 Weyl) could be used to fix 3 components of the worldsheet metric $g_{\alpha\beta}$. What happened to this argument? Why do we still have some gauge symmetry left? The reason is that $\tilde{\sigma}^\pm$ are functions of just a single variable, not two. So we did fix nearly all our gauge symmetries. What is left is a set of measure zero amongst the full gauge symmetry that we started with.
- The remaining reparameterization invariance (2.10) has an important physical implication. Recall that the solutions to the equations of motion are of the form $X_L^\mu(\sigma^+) + X_R^\mu(\sigma^-)$ which looks like $2D$ functions worth of solutions. Of course, we still have the constraints which, in terms of σ^\pm , read

$$(\partial_+ X)^2 = (\partial_- X)^2 = 0 , \quad (2.11)$$

which seems to bring the number down to $2(D - 1)$ functions. But the reparameterization invariance (2.10) tells us that even some of these are fake since we can always change what we mean by σ^\pm . The physical solutions of the string are therefore actually described by $2(D - 2)$ functions. But this counting has a nice interpretation: the degrees of freedom describe the *transverse* fluctuations of the string.

- The above comment reaches the same conclusion as the discussion in Section 1.3.2. There, in an attempt to get some feel for the constraints, we claimed that we could go to static gauge $X^0 = R\tau$ for some dimensionful parameter R . It is easy to check that this is simple to do using reparameterizations of the form (2.10). However, to solve the string constraints in full, it turns out that static gauge is not that useful. Rather we will use something called “lightcone gauge”.

2.2.1 Lightcone Gauge

We would like to gauge fix the remaining reparameterization invariance (2.10). The best way to do this is called lightcone gauge. In counterpoint to the worldsheet lightcone coordinates (2.9), we introduce the spacetime lightcone coordinates,

$$X^\pm = \sqrt{\frac{1}{2}}(X^0 \pm X^{D-1}) . \quad (2.12)$$

Note that this choice picks out a particular time direction and a particular spatial direction. It means that any calculations that we do involving X^\pm will not be manifestly Lorentz invariant. You might think that we needn’t really worry about this. We could try to make the following argument: “The equations may not *look* Lorentz invariant

but, since we started from a Lorentz invariant theory, at the end of the day any physical process is guaranteed to obey this symmetry". Right?! Well, unfortunately not. One of the more interesting and subtle aspects of quantum field theory is the possibility of anomalies: these are symmetries of the classical theory that do not survive the journey of quantization. When we come to the quantum theory, if our equations don't look Lorentz invariant then there's a real possibility that it's because the underlying physics actually isn't Lorentz invariant. Later we will need to spend some time figuring out under what circumstances our quantum theory keeps the classical Lorentz symmetry.

In lightcone coordinates, the spacetime Minkowski metric reads

$$ds^2 = -2dX^+dX^- + \sum_{i=1}^{D-2} dX^i dX^i$$

This means that indices are raised and lowered with $A_+ = -A^-$ and $A_- = -A^+$ and $A_i = A^i$. The product of spacetime vectors reads $A \cdot B = -A^+B^- - A^-B^+ + A^iB^i$.

Let's look at the solution to the equation of motion for X^+ . It reads,

$$X^+ = X_L^+(\sigma^+) + X_R^+(\sigma^-) .$$

We now gauge fix. We use our freedom of reparameterization invariance to choose coordinates such that

$$X_L^+ = \frac{1}{2}x^+ + \frac{1}{2}\alpha'p^+\sigma^+ , \quad X_R^+ = \frac{1}{2}x^+ + \frac{1}{2}\alpha'p^+\sigma^- .$$

You might think that we could go further and eliminate p^+ and x^+ but this isn't possible because we don't quite have the full freedom of reparameterization invariance since all functions should remain periodic in σ . The upshot of this choice of gauge is that

$$X^+ = x^+ + \alpha'p^+\tau . \quad (2.13)$$

This is *lightcone gauge*. Notice that, as long as $p^+ \neq 0$, we can always shift x^+ by a shift in τ .

There's something a little disconcerting about the choice (2.13). We've identified a timelike worldsheet coordinate with a null spacetime coordinate. Nonetheless, as you can see from the figure, it seems to be a good parameterization of the worldsheet. One could imagine that the parameterization might break if the string is actually massless and travels in the X^- direction, with $p^+ = 0$. But otherwise, all should be fine.

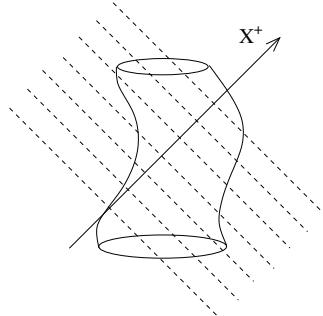


Figure 9:

Solving for X^-

The choice (2.13) does the job of fixing the reparameterization invariance (2.10). As we will now see, it also renders the constraint equations trivial. The first thing that we have to worry about is the possibility of extra constraints arising from this new choice of gauge fixing. This can be checked by looking at the equation of motion for X^+ ,

$$\partial_+ \partial_- X^- = 0$$

But we can solve this by the usual ansatz,

$$X^- = X_L^-(\sigma^+) + X_R^-(\sigma^-) .$$

We're still left with all the other constraints (2.11). Here we see the real benefit of working in lightcone gauge (which is actually what makes quantization possible at all): X^- is completely determined by these constraints. For example, the first of these reads

$$2\partial_+ X^- \partial_+ X^+ = \sum_{i=1}^{D-2} \partial_+ X^i \partial_+ X^i \quad (2.14)$$

which, using (2.13), simply becomes

$$\partial_+ X_L^- = \frac{1}{\alpha' p^+} \sum_{i=1}^{D-2} \partial_+ X^i \partial_+ X^i . \quad (2.15)$$

Similarly,

$$\partial_- X_R^- = \frac{1}{\alpha' p^+} \sum_{i=1}^{D-2} \partial_- X^i \partial_- X^i . \quad (2.16)$$

So, up to an integration constant, the function $X^-(\sigma^+, \sigma^-)$ is completely determined in terms of the other fields. If we write the usual mode expansion for $X_{L/R}^-$

$$\begin{aligned} X_L^-(\sigma^+) &= \frac{1}{2} x^- + \frac{1}{2} \alpha' p^- \sigma^+ + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \tilde{\alpha}_n^- e^{-in\sigma^+} , \\ X_R^-(\sigma^-) &= \frac{1}{2} x^- + \frac{1}{2} \alpha' p^- \sigma^- + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^- e^{-in\sigma^-} . \end{aligned}$$

then x^- is the undetermined integration constant, while p^- , α_n^- and $\tilde{\alpha}_n^-$ are all fixed by the constraints (2.15) and (2.16). For example, the oscillator modes α_n^- are given by,

$$\alpha_n^- = \sqrt{\frac{1}{2\alpha'} \frac{1}{p^+}} \sum_{m=-\infty}^{+\infty} \sum_{i=1}^{D-2} \alpha_{n-m}^i \alpha_m^i , \quad (2.17)$$

A special case of this is the $\alpha_0^- = \sqrt{\alpha'/2} p^-$ equation, which reads

$$\frac{\alpha' p^-}{2} = \frac{1}{2p^+} \sum_{i=1}^{D-2} \left(\frac{1}{2} \alpha' p^i p^i + \sum_{n \neq 0} \alpha_n^i \alpha_{-n}^i \right) . \quad (2.18)$$

We also get another equation for p^- from the $\tilde{\alpha}_0^-$ equation arising from (2.15)

$$\frac{\alpha' p^-}{2} = \frac{1}{2p^+} \sum_{i=1}^{D-2} \left(\frac{1}{2} \alpha' p^i p^i + \sum_{n \neq 0} \tilde{\alpha}_n^i \tilde{\alpha}_{-n}^i \right) . \quad (2.19)$$

From these two equations, we can reconstruct the old, classical, level matching conditions (1.41). But now with a difference:

$$M^2 = 2p^+ p^- - \sum_{i=1}^{D-2} p^i p^i = \frac{4}{\alpha'} \sum_{i=1}^{D-2} \sum_{n>0} \alpha_{-n}^i \alpha_n^i = \frac{4}{\alpha'} \sum_{i=1}^{D-2} \sum_{n>0} \tilde{\alpha}_{-n}^i \tilde{\alpha}_n^i . \quad (2.20)$$

The difference is that now the sum is over oscillators α^i and $\tilde{\alpha}^i$ only, with $i = 1, \dots, D-2$. We'll refer to these as *transverse* oscillators. Note that the string isn't necessarily living in the X^0 - X^{D-1} plane, so these aren't literally the transverse excitations of the string. Nonetheless, if we specify the α^i then all other oscillator modes are determined. In this sense, they are the physical excitation of the string.

Let's summarize the state of play so far. The most general classical solution is described in terms of $2(D-2)$ transverse oscillator modes α_n^i and $\tilde{\alpha}_n^i$, together with a number of zero modes describing the center of mass and momentum of the string: x^i, p^i, p^+ and x^- . But x^+ can be absorbed by a shift of τ in (2.13) and p^- is constrained to obey (2.18) and (2.19). In fact, p^- can be thought of as (proportional to) the lightcone Hamiltonian. Indeed, we know that p^- generates translations in x^+ , but this is equivalent to shifts in τ .

2.2.2 Quantization

Having identified the physical degrees of freedom, let's now quantize. We want to impose commutation relations. Some of these are easy:

$$\begin{aligned} [x^i, p^j] &= i\delta^{ij} , \quad [x^-, p^+] = -i \\ [\alpha_n^i, \alpha_m^j] &= [\tilde{\alpha}_n^i, \tilde{\alpha}_m^j] = n\delta^{ij}\delta_{n+m,0} . \end{aligned} \quad (2.21)$$

all of which follow from the commutation relations (2.2) that we saw in covariant quantization¹.

What to do with x^+ and p^- ? We could implement p^- as the Hamiltonian acting on states. In fact, it will prove slightly more elegant (but equivalent) if we promote both x^+ and p^- to operators with the expected commutation relation,

$$[x^+, p^-] = -i . \quad (2.22)$$

This is morally equivalent to writing $[t, H] = -i$ in non-relativistic quantum mechanics, which is true on a formal level. In the present context, it means that we can once again choose states to be eigenstates of p^μ , with $\mu = 0, \dots, D$, but the constraints (2.18) and (2.19) must still be imposed as operator equations on the physical states. We'll come to this shortly.

The Hilbert space of states is very similar to that described in covariant quantization: we define a vacuum state, $|0; p\rangle$ such that

$$\hat{p}^\mu |0; p\rangle = p^\mu |0; p\rangle , \quad \alpha_n^i |0; p\rangle = \tilde{\alpha}_n^i |0; p\rangle = 0 \quad \text{for } n > 0 \quad (2.23)$$

and we build a Fock space by acting with the creation operators α_{-n}^i and $\tilde{\alpha}_{-n}^i$ with $n > 0$. The difference with the covariant quantization is that we only act with transverse oscillators which carry a spatial index $i = 1, \dots, D-2$. For this reason, the Hilbert space is, by construction, positive definite. We don't have to worry about ghosts.

¹**Mea Culpa:** We're not really supposed to do this. The whole point of the approach that we're taking is to quantize just the physical degrees of freedom. The resulting commutation relations are not, in general, inherited from the larger theory that we started with simply by closing our eyes and forgetting about all the other fields that we've gauge fixed. We can see the problem by looking at (2.17), where α_n^- is determined in terms of α_n^i . This means that the commutation relations for α_n^i might be infected by those of α_n^- which could potentially give rise to extra terms. The correct procedure to deal with this is to figure out the Poisson bracket structure of the physical degrees of freedom in the classical theory. Or, in fancier language, the symplectic form on the phase space which schematically looks like

$$\omega \sim \int d\sigma \ - d\dot{X}^+ \wedge dX^- - d\dot{X}^- \wedge dX^+ + 2d\dot{X}^i \wedge dX^i ,$$

The reason that the commutation relations (2.21) do not get infected is because the α^- terms in the symplectic form come multiplying X^+ . Yet X^+ is given in (2.13). It has no oscillator modes. That means that the symplectic form doesn't pick up the Fourier modes of X^- and so doesn't receive any corrections from α_n^- . The upshot of this is that the naive commutation relations (2.21) are actually right.

The Constraints

Because p^- is not an independent variable in our theory, we must impose the constraints (2.18) and (2.19) by hand as operator equations which define the physical states. In the classical theory, we saw that these constraints are equivalent to mass-shell conditions (2.20).

But there's a problem when we go to the quantum theory. It's the same problem that we saw in covariant quantization: there's an ordering ambiguity in the sum over oscillator modes on the right-hand side of (2.20). If we choose all operators to be normal ordered then this ambiguity reveals itself in an overall constant, a , which we have not yet determined. The final result for the mass of states in lightcone gauge is:

$$M^2 = \frac{4}{\alpha'} \left(\sum_{i=1}^{D-2} \sum_{n>0} \alpha_{-n}^i \alpha_n^i - a \right) = \frac{4}{\alpha'} \left(\sum_{i=1}^{D-2} \sum_{n>0} \tilde{\alpha}_{-n}^i \tilde{\alpha}_n^i - a \right)$$

Since we'll use this formula quite a lot in what follows, it's useful to introduce quantities related to the number operators of the harmonic oscillator,

$$N = \sum_{i=1}^{D-2} \sum_{n>0} \alpha_{-n}^i \alpha_n^i , \quad \tilde{N} = \sum_{i=1}^{D-2} \sum_{n>0} \tilde{\alpha}_{-n}^i \tilde{\alpha}_n^i . \quad (2.24)$$

These are not quite number operators because of the factor of $1/\sqrt{n}$ in (2.3). The value of N and \tilde{N} is often called the level. Which, if nothing else, means that the name “level matching” makes sense. We now have

$$M^2 = \frac{4}{\alpha'} (N - a) = \frac{4}{\alpha'} (\tilde{N} - a) . \quad (2.25)$$

How are we going to fix a ? Later in the course we'll see the correct way to do it. For now, I'm just going to give you a quick and dirty derivation.

The Casimir Energy

“I told him that the sum of an infinite no. of terms of the series: $1 + 2 + 3 + 4 + \dots = -\frac{1}{12}$ under my theory. If I tell you this you will at once point out to me the lunatic asylum as my goal.”

Ramanujan, in a letter to G.H.Hardy.

What follows is a heuristic derivation of the normal ordering constant a . Suppose that we didn't notice that there was any ordering ambiguity and instead took the naive classical result directly over to the quantum theory, that is

$$\frac{1}{2} \sum_{n \neq 0} \alpha_{-n}^i \alpha_n^i = \frac{1}{2} \sum_{n<0} \alpha_{-n}^i \alpha_n^i + \frac{1}{2} \sum_{n>0} \alpha_{-n}^i \alpha_n^i .$$

where we've left the sum over $i = 1, \dots, D - 2$ implicit. We'll now try to put this in normal ordered form, with the annihilation operators α_n^i with $n > 0$ on the right-hand side. It's the first term that needs changing. We get

$$\frac{1}{2} \sum_{n<0} [\alpha_n^i \alpha_{-n}^i - n(D-2)] + \frac{1}{2} \sum_{n>0} \alpha_{-n}^i \alpha_n^i = \sum_{n>0} \alpha_{-n}^i \alpha_n^i + \frac{D-2}{2} \sum_{n>0} n .$$

The final term clearly diverges. But it at least seems to have a physical interpretation: it is the sum of zero point energies of an infinite number of harmonic oscillators. In fact, we came across exactly the same type of term in the course on quantum field theory where we learnt that, despite the divergence, one can still extract interesting physics from this. This is the physics of the Casimir force.

Let's recall the steps that we took to derive the Casimir force. Firstly, we introduced an ultra-violet cut-off $\epsilon \ll 1$, probably muttering some words about no physical plates being able to withstand very high energy quanta. Unfortunately, those words are no longer available to us in string theory, but let's proceed regardless. We replace the divergent sum over integers by the expression,

$$\begin{aligned} \sum_{n=1}^{\infty} n &\longrightarrow \sum_{n=1}^{\infty} n e^{-\epsilon n} = -\frac{\partial}{\partial \epsilon} \sum_{n=1}^{\infty} e^{-\epsilon n} \\ &= -\frac{\partial}{\partial \epsilon} (1 - e^{-\epsilon})^{-1} \\ &= \frac{1}{\epsilon^2} - \frac{1}{12} + \mathcal{O}(\epsilon) \end{aligned}$$

Obviously the $1/\epsilon^2$ piece diverges as $\epsilon \rightarrow 0$. This term should be renormalized away. In fact, this is necessary to preserve the Weyl invariance of the Polyakov action since it contributes to a cosmological constant on the worldsheet. After this renormalization, we're left with the wonderful answer, first intuited by Ramanujan

$$\sum_{n=1}^{\infty} n = -\frac{1}{12} .$$

While heuristic, this argument does predict the correct physical Casimir energy measured in one-dimensional systems. For example, this effect is seen in simulations of quantum spin chains.

What does this mean for our string? It means that we should take the unknown constant a in the mass formula (2.25) to be,

$$M^2 = \frac{4}{\alpha'} \left(N - \frac{D-2}{24} \right) = \frac{4}{\alpha'} \left(\tilde{N} - \frac{D-2}{24} \right) . \quad (2.26)$$

This is the formula that we will use to determine the spectrum of the string.

Zeta Function Regularization

I appreciate that the preceding argument is not totally convincing. We could spend some time making it more robust at this stage, but it's best if we wait until later in the course when we will have the tools of conformal field theory at our disposal. We will eventually revisit this issue and provide a respectable derivation of the Casimir energy in Section 4.4.1. For now I merely offer an even less convincing argument, known as zeta-function regularization.

The zeta-function is defined, for $\text{Re}(s) > 1$, by the sum

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} .$$

But $\zeta(s)$ has a unique analytic continuation to all values of s . In particular,

$$\zeta(-1) = -\frac{1}{12} .$$

Good? Good. This argument is famously unconvincing the first time you meet it! But it's actually a very useful trick for getting the right answer.

2.3 The String Spectrum

Finally, we're in a position to analyze the spectrum of a single, free string.

2.3.1 The Tachyon

Let's start with the ground state $|0; p\rangle$ defined in (2.23). With no oscillators excited, the mass formula (2.26) gives

$$M^2 = -\frac{1}{\alpha'} \frac{D-2}{6} . \quad (2.27)$$

But that's a little odd. It's a negative mass-squared. Such particles are called *tachyons*.

In fact, tachyons aren't quite as pathological as you might think. If you've heard of these objects before, it's probably in the context of special relativity where they're strange beasts which always travel faster than the speed of light. But that's not the right interpretation. Rather we should think more in the language of quantum field theory. Suppose that we have a field in spacetime — let's call it $T(X)$ — whose quanta will give rise to this particle. The mass-squared of the particle is simply the quadratic term in the action, or

$$M^2 = \left. \frac{\partial^2 V(T)}{\partial T^2} \right|_{T=0}$$

So the negative mass-squared in (2.27) is telling us that we're expanding around a maximum of the potential for the tachyon field as shown in the figure. Note that from this perspective, the Higgs field in the standard model at $H = 0$ is also a tachyon.

The fact that string theory turns out to sit at an unstable point in the tachyon field is unfortunate. The natural question is whether the potential has a good minimum elsewhere, as shown in the figure to the right. No one knows the answer to this! Naive attempts to understand this don't work. We know that around $T = 0$, the leading order contribution to the potential is negative and quadratic. But there are further terms that we can compute using techniques that we'll describe in Section 6. An expansion of the tachyon potential around $T = 0$ looks like

$$V(T) = \frac{1}{2}M^2T^2 + c_3T^3 + c_4T^4 + \dots$$

It turns out that the T^3 term in the potential does give rise to a minimum. But the T^4 term destabilizes it again. Moreover, the T field starts to mix with other scalar fields in the theory that we will come across soon. The ultimate fate of the tachyon in the bosonic string is not yet understood.

The tachyon is a problem for the bosonic string. It may well be that this theory makes no sense — or, at the very least, has no time-independent stable solutions. Or perhaps we just haven't worked out how to correctly deal with the tachyon. Either way, the problem does not arise when we introduce fermions on the worldsheet and study the superstring. This will involve several further technicalities which we won't get into in this course. Instead, our time will be put to better use if we continue to study the bosonic string since all the lessons that we learn will carry over directly to the superstring. However, one should be aware that the problem of the unstable vacuum will continue to haunt us throughout this course.

Although we won't describe it in detail, at several times along our journey we'll make an aside about how calculations work out for the superstring.

2.3.2 The First Excited States

We now look at the first excited states. If we act with a creation operator α_{-1}^j , then the level matching condition (2.25) tells us that we also need to act with a $\tilde{\alpha}_{-1}^i$ operator. This gives us $(D - 2)^2$ particle states,

$$\tilde{\alpha}_{-1}^i \alpha_{-1}^j |0; p\rangle , \quad (2.28)$$

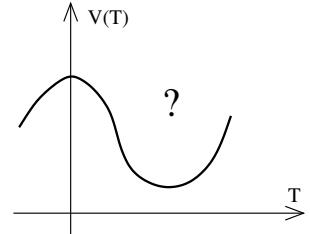


Figure 11:

each of which has mass

$$M^2 = \frac{4}{\alpha'} \left(1 - \frac{D-2}{24} \right) .$$

But now we seem to have a problem. Our states have space indices $i, j = 1, \dots, D-2$. The operators α^i and $\tilde{\alpha}^i$ each transform in the vector representation of $SO(D-2) \subset SO(1, D-1)$ which is manifest in lightcone gauge. But ultimately we want these states to fit into some representation of the full Lorentz $SO(1, D-1)$ group. That looks as if it's going to be hard to arrange. This is the first manifestation of the comment that we made after equation (2.12): it's tricky to see Lorentz invariance in lightcone gauge.

To proceed, let's recall Wigner's classification of representations of the Poincaré group. We start by looking at massive particles in $\mathbf{R}^{1,D-1}$. After going to the rest frame of the particle by setting $p^\mu = (p, 0, \dots, 0)$, we can watch how any internal indices transform under the little group $SO(D-1)$ of spatial rotations. The upshot of this is that any massive particle must form a representation of $SO(D-1)$. But the particles described by (2.28) have $(D-2)^2$ states. There's no way to package these states into a representation of $SO(D-1)$ and this means that there's no way that the first excited states of the string can form a massive representation of the D -dimensional Poincaré group.

It looks like we're in trouble. Thankfully, there's a way out. If the states are massless, then we can't go to the rest frame. The best that we can do is choose a spacetime momentum for the particle of the form $p^\mu = (p, 0, \dots, 0, p)$. In this case, the particles fill out a representation of the little group $SO(D-2)$. This means that massless particles get away with having fewer internal states than massive particles. For example, in four dimensions the photon has two polarization states, but a massive spin-1 particle must have three.

The first excited states (2.28) happily sit in a representation of $SO(D-2)$. We learn that if we want the quantum theory to preserve the $SO(1, D-1)$ Lorentz symmetry that we started with, then these states will have to be massless. And this is only the case if the dimension of spacetime is

$$D = 26 .$$

This is our first derivation of the critical dimension of the bosonic string.

Moreover, we've found that our theory contains a bunch of massless particles. And massless particles are interesting because they give rise to long range forces. Let's look

more closely at what massless particles the string has given us. The states (2.28) transform in the $\mathbf{24} \otimes \mathbf{24}$ representation of $SO(24)$. These decompose into three irreducible representations:

$$\text{traceless symmetric} \oplus \text{anti-symmetric} \oplus \text{singlet} (= \text{trace})$$

To each of these modes, we associate a massless field in spacetime such that the string oscillation can be identified with a quantum of these fields. The fields are:

$$G_{\mu\nu}(X) , \quad B_{\mu\nu}(X) , \quad \Phi(X) \tag{2.29}$$

Of these, the first is the most interesting and we shall have more to say momentarily. The second is an anti-symmetric tensor field which is usually called the anti-symmetric tensor field. It also goes by the names of the “Kalb-Ramond field” or, in the language of differential geometry, the “2-form”. The scalar field is called the *dilaton*. These three massless fields are common to all string theories. We’ll learn more about the role these fields play later in the course.

The particle in the symmetric traceless representation of $SO(24)$ is particularly interesting. This is a massless spin 2 particle. However, there are general arguments, due originally to Feynman and Weinberg, that *any* theory of interacting massless spin two particles must be equivalent to general relativity². We should therefore identify the field $G_{\mu\nu}(X)$ with the metric of spacetime. Let’s pause briefly to review the thrust of these arguments.

Why Massless Spin 2 = General Relativity

Let’s call the spacetime metric $G_{\mu\nu}(X)$. We can expand around flat space by writing

$$G_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}(X) .$$

Then the Einstein-Hilbert action has an expansion in powers of h . If we truncate to quadratic order, we simply have a free theory which we may merrily quantize in the usual canonical fashion: we promote $h_{\mu\nu}$ to an operator and introduce the associated creation and annihilation operators $a_{\mu\nu}$ and $a_{\mu\nu}^\dagger$. This way of looking at gravity is anathema to those raised in the geometrical world of general relativity. But from a particle physics language it is very standard: it is simply the quantization of a massless spin 2 field, $h_{\mu\nu}$.

²A very readable description of this can be found in the first few chapters of the Feynman Lectures on Gravitation.

However, even on this simple level, there is a problem due to the indefinite signature of the spacetime Minkowski metric. The canonical quantization relations of the creation and annihilation operators are schematically of the form,

$$[a_{\mu\nu}, a_{\rho\sigma}^\dagger] \sim \eta_{\mu\rho}\eta_{\nu\sigma} + \eta_{\mu\sigma}\eta_{\nu\rho}$$

But this will lead to a Hilbert space with negative norm states coming from acting with time-like creation operators. For example, the one-graviton state of the form,

$$a_{0i}^\dagger |0\rangle \tag{2.30}$$

suffers from a negative norm. This should be becoming familiar by now: it is the usual problem that we run into if we try to covariantly quantize a gauge theory. And, indeed, general relativity is a gauge theory. The gauge transformations are diffeomorphisms. We would hope that this saves the theory of quantum gravity from these negative norm states.

Let's look a little more closely at what the gauge symmetry looks like for small fluctuations $h_{\mu\nu}$. We've butchered the Einstein-Hilbert action and left only terms quadratic in h . Including all the index contractions, we find

$$S_{EH} = \frac{M_{pl}^2}{2} \int d^4x \left[\partial_\mu h^\rho_\rho \partial_\nu h^{\mu\nu} - \partial^\rho h^{\mu\nu} \partial_\mu h_{\rho\nu} + \frac{1}{2} \partial_\rho h_{\mu\nu} \partial^\rho h^{\mu\nu} - \frac{1}{2} \partial_\mu h^\nu_\nu \partial^\mu h^\rho_\rho \right] + \dots$$

One can check that this truncated action is invariant under the gauge symmetry,

$$h_{\mu\nu} \longrightarrow h_{\mu\nu} + \partial_\mu \xi_\nu + \partial_\nu \xi_\mu \tag{2.31}$$

for any function $\xi_\mu(X)$. The gauge symmetry is the remnant of diffeomorphism invariance, restricted to small deviations away from flat space. With this gauge invariance in hand one can show that, just like QED, the negative norm states decouple from all physical processes.

To summarize, theories of massless spin 2 fields only make sense if there is a gauge symmetry to remove the negative norm states. In general relativity, this gauge symmetry descends from diffeomorphism invariance. The argument of Feynman and Weinberg now runs this logic in reverse. It goes as follows: suppose that we have a massless, spin 2 particle. Then, at the linearized level, it must be invariant under the gauge symmetry (2.31) in order to eliminate the negative norm states. Moreover, this symmetry must survive when interaction terms are introduced. But the only way to do this is to ensure that the resulting theory obeys diffeomorphism invariance. That means the theory of any interacting, massless spin 2 particle is Einstein gravity, perhaps supplemented by higher derivative terms.

We haven't yet shown that string theory includes interactions for $h_{\mu\nu}$ but we will come to this later in the course. More importantly, we will also explicitly see how Einstein's field equations arise directly in string theory.

A Comment on Spacetime Gauge Invariance

We've surreptitiously put $\mu, \nu = 0, \dots, 25$ indices on the spacetime fields, rather than $i, j = 1, \dots, 24$. The reason we're allowed to do this is because both $G_{\mu\nu}$ and $B_{\mu\nu}$ enjoy a spacetime gauge symmetry which allows us to eliminate appropriate modes. Indeed, this is exactly the gauge symmetry (2.31) that entered the discussion above. It isn't possible to see these spacetime gauge symmetries from the lightcone formalism of the string since, by construction, we find only the physical states (although, by consistency alone, the gauge symmetries must be there). One of the main advantages of pushing through with the covariant calculation is that it does allow us to see how the spacetime gauge symmetry emerges from the string worldsheet. Details can be found in Green, Schwarz and Witten. We'll also briefly return to this issue in Section 5.

2.3.3 Higher Excited States

We rescued the Lorentz invariance of the first excited states by choosing $D = 26$ to ensure that they are massless. But now we've used this trick once, we still have to worry about all the other excited states. These also carry indices that take the range $i, j = 1, \dots, D - 2 = 24$ and, from the mass formula (2.26), they will all be massive and so must form representations of $SO(D - 1)$. It looks like we're in trouble again.

Let's examine the string at level $N = \tilde{N} = 2$. In the right-moving sector, we now have two different states: $\alpha_{-1}^i \alpha_{-1}^j |0\rangle$ and $\alpha_{-2}^i |0\rangle$. The same is true for the left-moving sector, meaning that the total set of states at level 2 is (in notation that is hopefully obvious, but probably technically wrong)

$$(\alpha_{-1}^i \alpha_{-1}^j \oplus \alpha_{-2}^i) \otimes (\tilde{\alpha}_{-1}^i \tilde{\alpha}_{-1}^j \oplus \tilde{\alpha}_{-2}^i) |0; p\rangle .$$

These states have mass $M^2 = 4/\alpha'$. How many states do we have? In the left-moving sector, we have,

$$\frac{1}{2}(D - 2)(D - 1) + (D - 2) = \frac{1}{2}D(D - 1) - 1 .$$

But, remarkably, that does fit nicely into a representation of $SO(D - 1)$, namely the traceless symmetric tensor representation.

In fact, one can show that all excited states of the string fit nicely into $SO(D - 1)$ representations. The only consistency requirement that we need for Lorentz invariance is to fix up the first excited states: $D = 26$.

Note that if we are interested in a fundamental theory of quantum gravity, then all these excited states will have masses close to the Planck scale so are unlikely to be observable in particle physics experiments. Nonetheless, as we shall see when we come to discuss scattering amplitudes, it is the presence of this infinite tower of states that tames the ultra-violet behaviour of gravity.

2.4 Lorentz Invariance Revisited

The previous discussion allowed us to derive both the critical dimension and the spectrum of string theory in the quickest fashion. But the derivation creaks a little in places. The calculation of the Casimir energy is unsatisfactory the first time one sees it. Similarly, the explanation of the need for massless particles at the first excited level is correct, but seems rather cheap considering the huge importance that we're placing on the result.

As I've mentioned a few times already, we'll shortly do better and gain some physical insight into these issues, in particular the critical dimension. But here I would just like to briefly sketch how one can be a little more rigorous within the framework of lightcone quantization. The question, as we've seen, is whether one preserves spacetime Lorentz symmetry when we quantize in lightcone gauge. We can examine this more closely.

Firstly, let's go back to the action for free scalar fields (1.30) before we imposed lightcone gauge fixing. Here the full Poincaré symmetry was manifest: it appears as a global symmetry on the worldsheet,

$$X^\mu \rightarrow \Lambda^\mu{}_\nu X^\nu + c^\mu \quad (2.32)$$

But recall that in field theory, global symmetries give rise to Noether currents and their associated conserved charges. What are the Noether currents associated to this Poincaré transformation? We can start with the translations $X^\mu \rightarrow X^\mu + c^\mu$. A quick computation shows that the current is,

$$P_\mu^\alpha = T\partial^\alpha X_\mu \quad (2.33)$$

which is indeed a conserved current since $\partial_\alpha P_\mu^\alpha = 0$ is simply the equation of motion. Similarly, we can compute the $\frac{1}{2}D(D-1)$ currents associated to Lorentz transformations. They are,

$$J_{\mu\nu}^\alpha = P_\mu^\alpha X_\nu - P_\nu^\alpha X_\mu$$

It's not hard to check that $\partial_\alpha J_{\mu\nu}^\alpha = 0$ when the equations of motion are obeyed.

The conserved charges arising from this current are given by $M_{\mu\nu} = \int d\sigma J_{\mu\nu}^\tau$. Using the mode expansion (1.36) for X^μ , these can be written as

$$\begin{aligned}\mathcal{M}^{\mu\nu} &= (p^\mu x^\nu - p^\nu x^\mu) - i \sum_{n=1}^{\infty} \frac{1}{n} (\alpha_{-n}^\nu \alpha_n^\mu - \alpha_{-n}^\mu \alpha_n^\nu) - i \sum_{n=1}^{\infty} \frac{1}{n} (\tilde{\alpha}_{-n}^\nu \tilde{\alpha}_n^\mu - \tilde{\alpha}_{-n}^\mu \tilde{\alpha}_n^\nu) \\ &\equiv l^{\mu\nu} + S^{\mu\nu} + \tilde{S}^{\mu\nu}\end{aligned}$$

The first piece, $l^{\mu\nu}$, is the orbital angular momentum of the string while the remaining pieces $S^{\mu\nu}$ and $\tilde{S}^{\mu\nu}$ tell us the angular momentum due to excited oscillator modes. Classically, these obey the Poisson brackets of the Lorentz algebra. Moreover, if we quantize in the covariant approach, the corresponding operators obey the commutation relations of the Lorentz Lie algebra, namely

$$[\mathcal{M}^{\rho\sigma}, \mathcal{M}^{\tau\nu}] = \eta^{\sigma\tau} \mathcal{M}^{\rho\nu} - \eta^{\rho\tau} \mathcal{M}^{\sigma\nu} + \eta^{\rho\nu} \mathcal{M}^{\sigma\tau} - \eta^{\sigma\nu} \mathcal{M}^{\rho\tau}$$

However, things aren't so easy in lightcone gauge. Lorentz invariance is not guaranteed and, in general, is not there. The right way to go about looking for it is to make sure that the Lorentz algebra above is reproduced by the generators $\mathcal{M}^{\mu\nu}$. It turns out that the smoking gun lies in the commutation relation,

$$[\mathcal{M}^{i-}, \mathcal{M}^{j-}] = 0$$

Does this equation hold in lightcone gauge? The problem is that it involves the operators p^- and α_n^- , both of which are fixed by (2.17) and (2.18) in terms of the other operators. So the task is to compute this commutation relation $[\mathcal{M}^{i-}, \mathcal{M}^{j-}]$, given the commutation relations (2.21) for the physical degrees of freedom, and check that it vanishes. To do this, we re-instate the ordering ambiguity a and the number of spacetime dimension D as arbitrary variables and proceed.

The part involving orbital angular momenta l^{i-} is fairly straightforward. (Actually, there's a small subtlety because we must first make sure that the operator $l^{\mu\nu}$ is Hermitian by replacing $x^\mu p^\nu$ with $\frac{1}{2}(x^\mu p^\nu + p^\nu x^\mu)$). The real difficulty comes from computing the commutation relations $[S^{i-}, S^{j-}]$. This is messy³. After a tedious computation, one finds,

$$[\mathcal{M}^{i-}, \mathcal{M}^{j-}] = \frac{2}{(p^+)^2} \sum_{n>0} \left(\left[\frac{D-2}{24} - 1 \right] n + \frac{1}{n} \left[a - \frac{D-2}{24} \right] \right) (\alpha_{-n}^i \alpha_n^j - \alpha_{-n}^j \alpha_n^i) + (\alpha \leftrightarrow \tilde{\alpha})$$

³The original, classic, paper where lightcone quantization was first implemented is Goddard, Goldstone, Rebbo and Thorn “*Quantum Dynamics of a Massless Relativistic String*”, Nucl. Phys. B56 (1973). A pedestrian walkthrough of this calculation can be found in the lecture notes by Gleb Arutyunov. A link is given on the course webpage.

The right-hand side does not, in general, vanish. We learn that the relativistic string can only be quantized in flat Minkowski space if we pick,

$$D = 26 \quad \text{and} \quad a = 1.$$

2.5 A Nod to the Superstring

We won't provide details of the superstring in this course, but will pause occasionally to make some pertinent comments. Although what follows is nothing more than a list of facts, it will hopefully be helpful in orienting you when you do come to study this material.

The key difference between the bosonic string and the superstring is the addition of fermionic modes on its worldsheet. The resulting worldsheet theory is supersymmetric. (At least in the so-called Neveu-Schwarz-Ramond formalism). Hence the name “superstring”. Applying the kind of quantization procedure we've discussed in this section, one finds the following results:

- The critical dimension of the superstring is $D = 10$.
- There is no tachyon in the spectrum.
- The massless bosonic fields $G_{\mu\nu}$, $B_{\mu\nu}$ and Φ are all part of the spectrum of the superstring. In this context, $B_{\mu\nu}$ is sometimes referred to as the Neveu-Schwarz 2-form. There are also massless spacetime fermions, as well as further massless bosonic fields. As we now discuss, the exact form of these extra bosonic fields depends on exactly what superstring theory we consider.

While the bosonic string is unique, there are a number of discrete choices that one can make when adding fermions to the worldsheet. This gives rise to a handful of different perturbative superstring theories. (Although later developments reveal that they are actually all part of the same framework which sometimes goes by the name of *M-theory*). The most important of these discrete options is whether we add fermions in both the left-moving and right-moving sectors of the string, or whether we choose the fermions to move only in one direction, usually taken to be right-moving. This gives rise to two different classes of string theory.

- Type II strings have both left and right-moving worldsheet fermions. The resulting spacetime theory in $D = 10$ dimensions has $\mathcal{N} = 2$ supersymmetry, which means 32 supercharges.
- Heterotic strings have just right-moving fermions. The resulting spacetime theory has $\mathcal{N} = 1$ supersymmetry, or 16 supercharges.

In each of these cases, there is then one further discrete choice that we can make. This leaves us with four superstring theories. In each case, the massless bosonic fields include $G_{\mu\nu}$, $B_{\mu\nu}$ and Φ together with a number of extra fields. These are:

- **Type IIA:** In the type II theories, the extra massless bosonic excitations of the string are referred to as *Ramond-Ramond* fields. For Type IIA, they are a 1-form C_μ and a 3-form $C_{\mu\nu\rho}$. Each of these is to be thought of as a gauge field. The gauge invariant information lies in the field strengths which take the form $F = dC$.
- **Type IIB:** The Ramond-Ramond gauge fields consist of a scalar C , a 2-form $C_{\mu\nu}$ and a 4-form $C_{\mu\nu\rho\sigma}$. The 4-form is restricted to have a self-dual field strength: $F_5 = {}^*F_5$. (Actually, this statement is almost true...we'll look a little closer at this in Section 7.3.3).
- **Heterotic $SO(32)$:** The heterotic strings do not have Ramond-Ramond fields. Instead, each comes with a non-Abelian gauge field in spacetime. The heterotic strings are named after the gauge group. For example, the Heterotic $SO(32)$ string gives rise to an $SO(32)$ Yang-Mills theory in ten dimensions.
- **Heterotic $E_8 \times E_8$:** The clue is in the name. This string gives rise to an $E_8 \times E_8$ Yang-Mills field in ten-dimensions.

It is sometimes said that there are five perturbative superstring theories in ten dimensions. Here we've only mentioned four. The remaining theory is called Type I and includes open strings moving in flat ten dimensional space as well as closed strings. We'll mention it in passing in the following section.

3. Open Strings and D-Branes

In this section we discuss the dynamics of open strings. Clearly their distinguishing feature is the existence of two end points. Our goal is to understand the effect of these end points. The spatial coordinate of the string is parameterized by

$$\sigma \in [0, \pi] .$$

The dynamics of a generic point on a string is governed by local physics. This means that a generic point has no idea if it is part of a closed string or an open string. The dynamics of an open string must therefore still be described by the Polyakov action. But this must now be supplemented by something else: boundary conditions to tell us how the end points move. To see this, let's look at the Polyakov action in conformal gauge

$$S = -\frac{1}{4\pi\alpha'} \int d^2\sigma \partial_\alpha X \cdot \partial^\alpha X .$$

As usual, we derive the equations of motion by finding the extrema of the action. This involves an integration by parts. Let's consider the string evolving from some initial configuration at $\tau = \tau_i$ to some final configuration at $\tau = \tau_f$:

$$\begin{aligned} \delta S &= -\frac{1}{2\pi\alpha'} \int_{\tau_i}^{\tau_f} d\tau \int_0^\pi d\sigma \partial_\alpha X \cdot \partial^\alpha \delta X \\ &= \frac{1}{2\pi\alpha'} \int d^2\sigma (\partial^\alpha \partial_\alpha X) \cdot \delta X + \text{total derivative} \end{aligned}$$

For an open string the total derivative picks up the boundary contributions

$$\frac{1}{2\pi\alpha'} \left[\int_0^\pi d\sigma \dot{X} \cdot \delta X \right]_{\tau=\tau_i}^{\tau=\tau_f} - \frac{1}{2\pi\alpha'} \left[\int_{\tau_i}^{\tau_f} d\tau X' \cdot \delta X \right]_{\sigma=0}^{\sigma=\pi}$$

The first term is the kind that we always get when using the principle of least action. The equations of motion are derived by requiring that $\delta X^\mu = 0$ at $\tau = \tau_i$ and τ_f and so it vanishes. However, the second term is novel. In order for it too to vanish, we require

$$\partial_\sigma X^\mu \delta X_\mu = 0 \quad \text{at } \sigma = 0, \pi$$

There are two different types of boundary conditions that we can impose to satisfy this:

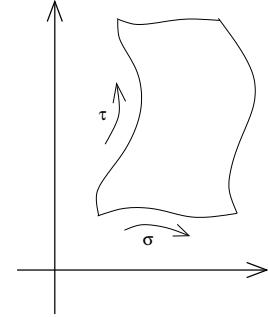


Figure 12:

- Neumann boundary conditions.

$$\partial_\sigma X^\mu = 0 \quad \text{at } \sigma = 0, \pi \quad (3.1)$$

Because there is no restriction on δX^μ , this condition allows the end of the string to move freely. To see the consequences of this, it's useful to repeat what we did for the closed string and work in static gauge with $X^0 \equiv t = R\tau$, for some dimensionful constant R . Then, as in equations (1.34), the constraints read

$$\dot{\vec{x}} \cdot \vec{x}' = 0 \quad \text{and} \quad \dot{\vec{x}}^2 + \vec{x}'^2 = R^2$$

But at the end points of the string, $\vec{x}' = 0$. So the second equation tells us that $|d\vec{x}/dt| = 1$. Or, in other words, the end point of the string moves at the speed of light.

- Dirichlet boundary conditions

$$\delta X^\mu = 0 \quad \text{at } \sigma = 0, \pi \quad (3.2)$$

This means that the end points of the string lie at some constant position, $X^\mu = c^\mu$, in space.

At first sight, Dirichlet boundary conditions may seem a little odd. Why on earth would the strings be fixed at some point c^μ ? What is special about that point? Historically people were pretty hung up about this and Dirichlet boundary conditions were rarely considered until the mid-1990s. Then everything changed due to an insight of Polchinski...

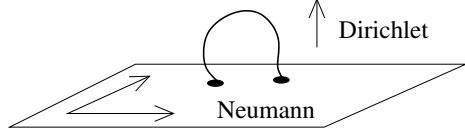


Figure 13:

Let's consider Dirichlet boundary conditions for some coordinates and Neumann for the others. This means that at both end points of the string, we have

$$\begin{aligned} \partial_\sigma X^a &= 0 && \text{for } a = 0, \dots, p \\ X^I &= c^I && \text{for } I = p+1, \dots, D-1 \end{aligned} \quad (3.3)$$

This fixes the end-points of the string to lie in a $(p+1)$ -dimensional hypersurface in spacetime such that the $SO(1, D-1)$ Lorentz group is broken to,

$$SO(1, D-1) \rightarrow SO(1, p) \times SO(D-p-1).$$

This hypersurface is called a *D-brane* or, when we want to specify its dimension, a *D_p-brane*. Here D stands for Dirichlet, while p is the number of spatial dimensions of the brane. So, in this language, a D0-brane is a particle; a D1-brane is itself a string; a D2-brane a membrane and so on. The brane sits at specific positions c^I in the transverse space. But what is the interpretation of this hypersurface?

It turns out that the D-brane hypersurface should be thought of as a new, dynamical object in its own right. This is a conceptual leap that is far from obvious. Indeed, it took decades for people to fully appreciate this fact. String theory is not just a theory of strings: it also contains higher dimensional branes. In Section 7.5 we will see how these D-branes develop a life of their own. Some comments:

- We've defined D-branes that are infinite in space. However, we could just as well define finite D-branes by specifying closed surfaces on which the string can end.
- There are many situations where we want to describe strings that have Neumann boundary conditions in all directions, meaning that the string is free to move throughout spacetime. It's best to understand this in terms of a space-filling D-brane. No Dirichlet conditions means D-branes are everywhere!
- The D p -brane described above always has Neumann boundary conditions in the X^0 direction. What would it mean to have Dirichlet conditions for X^0 ? Obviously this is a little weird since the object is now localized at a fixed point in time. But there is an interpretation of such an object: it is an *instanton*. This "D-instanton" is usually referred to as a D(-1)-brane. It is related to tunneling effects in the quantum theory.

Mode Expansion

We take the usual mode expansion for the string, with $X^\mu = X_L^\mu(\sigma^+) + X_R^\mu(\sigma^-)$ and

$$\begin{aligned} X_L^\mu(\sigma^+) &= \tfrac{1}{2}x^\mu + \alpha' p^\mu \sigma^+ + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \tilde{\alpha}_n^\mu e^{-in\sigma^+}, \\ X_R^\mu(\sigma^-) &= \tfrac{1}{2}x^\mu + \alpha' p^\mu \sigma^- + i\sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^\mu e^{-in\sigma^-}. \end{aligned} \quad (3.4)$$

The boundary conditions impose relations on the modes of the string. They are easily checked to be:

- Neumann boundary conditions, $\partial_\sigma X^a = 0$, at the end points require that

$$\alpha_n^a = \tilde{\alpha}_n^a \quad (3.5)$$

- Dirichlet boundary conditions, $X^I = c^I$, at the end points require that

$$x^I = c^I \quad , \quad p^I = 0 \quad , \quad \alpha_n^I = -\tilde{\alpha}_n^I$$

So for both boundary conditions, we only have one set of oscillators, say α_n . The $\tilde{\alpha}_n$ are then determined by the boundary conditions.

It's worth pointing out that there is a factor of 2 difference in the p^μ term between the open string (3.4) and the closed string (1.36). This is to ensure that p^μ for the open string retains the interpretation of the spacetime momentum of the string when $\sigma \in [0, \pi]$. To see this, one needs to check the Noether current associated to translations of X^μ on the worldsheet: it was given in (2.33). The conserved charge is then

$$P^\mu = \int_0^\pi d\sigma (P^\tau)^\mu = \frac{1}{2\pi\alpha'} \int_0^\pi d\sigma \dot{X}^\mu = p^\mu$$

as advertised. Note that we've needed to use the Neumann conditions (3.5) to ensure that the Fourier modes don't contribute to this integral.

3.1 Quantization

To quantize, we promote the fields x^a and p^a and α_n^μ to operators. The other elements in the mode expansion are fixed by the boundary conditions. An obvious, but important, point is that the position and momentum degrees of freedom, x^a and p^a , have a spacetime index that takes values $a = 0, \dots, p$. This means that the spatial wavefunctions only depend on the coordinates of the brane not the whole spacetime. Said another, quantizing an open string gives rise to states which are restricted to lie on the brane.

To determine the spectrum, it is again simplest to work in lightcone gauge. The spacetime lightcone coordinate is chosen to lie within the brane,

$$X^\pm = \sqrt{\frac{1}{2}} (X^0 \pm X^p)$$

Quantization now proceeds in the same manner as for the closed string until we arrive at the mass formula for states which is a sum over the transverse modes of the string.

$$M^2 = \frac{1}{\alpha'} \left(\sum_{i=1}^{p-1} \sum_{n>0} \alpha_{-n}^i \alpha_n^i + \sum_{i=p+1}^{D-1} \sum_{n>0} \alpha_{-n}^i \alpha_n^i - a \right)$$

The first sum is over modes parallel to the brane, the second over modes perpendicular to the brane. It's worth commenting on the differences with the closed string formula. Firstly, there is an overall factor of 4 difference. This can be traced to the lack of the factor of 1/2 in front of p^μ in the mode expansion that we discussed above. Secondly, there is a sum only over α modes. The $\tilde{\alpha}$ modes are not independent because of the boundary conditions.

Open and Closed

In the mass formula, we have once again left the normal ordering constant a ambiguous. As in the closed string case, requiring the Lorentz symmetry of the quantum theory — this time the reduced symmetry $SO(1, p) \times SO(D - p - 1)$ — forces us to choose

$$D = 26 \quad \text{and} \quad a = 1 .$$

These are the same values that we found for the closed string. This reflects an important fact: the open string and closed string are not different theories. They are both different states inside the same theory.

More precisely, theories of open strings necessarily contain closed strings. This is because, once we consider interactions, an open string can join to form a closed string as shown in the figure. We'll look at interactions in Section 6. The question of whether this works the other way — meaning whether closed string theories require open strings — is a little more involved and is cleanest to state in the context of the superstring. For type II superstrings, the open strings and D-branes are necessary ingredients. For heterotic superstrings, there appear to be no open strings and no D-branes. For the bosonic theory, it seems likely that the open strings are a necessary ingredient although I don't know of a killer argument. But since we're not sure whether the theory exists due to the presence of the tachyon, the point is probably moot. In the remainder of these lectures, we'll view the bosonic string in the same manner as the type II string and assume that the theory includes both closed strings and open strings with their associated D-branes.

3.1.1 The Ground State

The ground state is defined by

$$\alpha_n^i |0; p\rangle = 0 \quad n > 0$$

The spatial index now runs over $i = 1, \dots, p-1, p+1, \dots, D-1$. The ground state has mass

$$M^2 = -\frac{1}{\alpha'}$$

It is again tachyonic. Its mass is half that of the closed string tachyon. As we commented above, this time the tachyon is confined to the brane. In contrast to the closed string tachyon, the open string tachyon is now fairly well understood and its potential

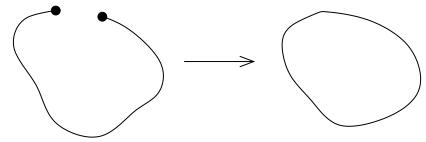


Figure 14:

is of the form shown in the figure. The interpretation is that the brane is unstable. It will decay, much like a resonance state in field theory. It does this by dissolving into closed string modes. The end point of this process – corresponding to the minimum at $T > 0$ in the figure – is simply a state with no D-brane. The difference between the value of the potential at the minimum and at $T = 0$ is the tension of the D-brane.

Notice that although there is a minimum of the potential at $T > 0$, it is not a global minimum. The potential seems to drop off without bound to the left. This is still not well understood. There are suggestions that it is related in some way to the closed string tachyon.

3.1.2 First Excited States: A World of Light

The first excited states are massless. They fall into two classes:

- Oscillators longitudinal to the brane,

$$\alpha_{-1}^a |0; p\rangle \quad a = 1, \dots, p-1$$

The spacetime indices a lie within the brane so this state transforms under the $SO(1, p)$ Lorentz group. It is a spin 1 particle on the brane or, in other words, it is a photon. We introduce a gauge field A_a with $a = 0, \dots, p$ lying on the brane whose quanta are identified with this photon.

- Oscillators transverse to the brane,

$$\alpha_{-1}^I |0; p\rangle \quad I = p+1, \dots, D-1$$

These states are scalars under the $SO(1, p)$ Lorentz group of the brane. They can be thought of as arising from scalar fields ϕ^I living on the brane. These scalars have a nice interpretation: they are fluctuations of the brane in the transverse directions. This is our first hint that the D-brane is a dynamical object. Note that although the ϕ^I are scalar fields under the $SO(1, p)$ Lorentz group of the brane, they do transform as a vector under the $SO(D-p-1)$ rotation group transverse to the brane. This appears as a global symmetry on the brane worldvolume.

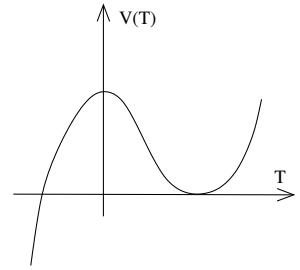


Figure 15:

3.1.3 Higher Excited States and Regge Trajectories

At level N , the mass of the string state is

$$M^2 = \frac{1}{\alpha'}(N - 1)$$

The maximal spin of these states arises from the symmetric tensor. It is

$$J_{max} = N = \alpha' M^2 + 1$$

Plotting the spin vs. the mass-squared, we find straight lines. These are usually called *Regge trajectories*. (Or sometimes Chew-Frautschi trajectories). They are seen in Nature in both the spectrum of mesons and baryons. Some examples involving ρ -mesons are shown in the figure. These stringy Regge trajectories suggest a naive cartoon picture of mesons as two rotating quarks connected by a confining flux tube.

The value of the string tension required to match the hadron spectrum of QCD is $T \sim 1$ GeV. This relationship between the strong interaction and the open string was one of the original motivations for the development of string theory and it is from here that the parameter α' gets its (admittedly rarely used) name “Regge slope”. In these enlightened modern times, the connection between the open string and quarks lives on in the AdS/CFT correspondence.

3.1.4 Another Nod to the Superstring

Just as supersymmetry eliminates the closed string tachyon, so it removes the open string tachyon. Open strings are an ingredient of the type II string theories. The possible D-branes are

- Type IIA string theory has stable D p -branes with p even.
- Type IIB string theory has stable D p -branes with p odd.

The most important reason that D-branes are stable in the type II string theories is that they are charged under the Ramond-Ramond fields. (This was actually Polchinski’s insight that made people take D-branes seriously). However, type II string theories also contain unstable branes, with p odd in type IIA and p even in type IIB.

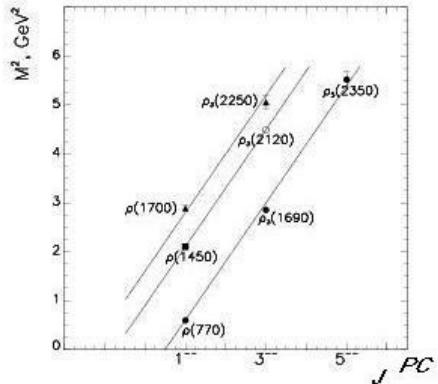


Figure 16:

The fifth string theory (which was actually the first to be discovered) is called Type I. Unlike the other string theories, it contains both open and closed strings moving in flat ten-dimensional Lorentz-invariant spacetime. It can be thought of as the Type IIB theory with a bunch of space-filling D9-branes, together with something called an orientifold plane. You can read about this in Polchinski.

As we mentioned above, the heterotic string doesn't have (finite energy) D-branes. This is due to an inconsistency in any attempt to reflect left-moving modes into right-moving modes.

3.2 Brane Dynamics: The Dirac Action

We have introduced D-branes as fixed boundary conditions for the open string. However, we've already seen a hint that these objects are dynamical in their own right, since the massless scalar excitations ϕ^I have a natural interpretation as transverse fluctuations of the brane. Indeed, if a theory includes both open strings and closed strings, then the D-branes have to be dynamical because there can be no rigid objects in a theory of gravity. The dynamical nature of D-branes will become clearer as the course progresses.

But any dynamical object should have an action which describes how it moves. Moreover, after our discussion in Section 1, we already know what this is! On grounds of Lorentz invariance and reparameterization invariance alone, the action must be a higher dimensional extension of the Nambu-Goto action. This is

$$S_{Dp} = -T_p \int d^{p+1}\xi \sqrt{-\det \gamma} \quad (3.6)$$

where T_p is the tension of the Dp-brane which we will determine later, while ξ^a , $a = 0, \dots, p$, are the worldvolume coordinates of the brane. γ_{ab} is the pull back of the spacetime metric onto the worldvolume,

$$\gamma_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} \eta_{\mu\nu} .$$

This is called the *Dirac action*. It was first written down by Dirac for a membrane some time before Nambu and Goto rediscovered it in the context of the string.

To make contact with the fields ϕ^I , we can use the reparameterization invariance of the Dirac action to go to static gauge. For an infinite, flat Dp-brane we can choose

$$X^a = \xi^a \quad a = 0, \dots, p .$$

The dynamical transverse coordinates are then identified with the fluctuations ϕ^I through

$$X^I(\xi) = 2\pi\alpha' \phi^I(\xi) \quad I = p + 1, \dots, D - 1$$

However, the Dirac action can't be the whole story. It describes the transverse fluctuations of the D-brane, but has nothing to say about the $U(1)$ gauge field A_μ which lives on the D-brane. There must be some action which describes how this gauge field moves as well. We will return to this in Section 7.

What's Special About Strings?

We could try to quantize the Dirac action (3.6) for a D-brane in the same manner that we quantized the action for the string. Is this possible? The answer, at present, is no. There appear to be both technical and conceptual obstacles. The technical issue is just that it's hard. Weyl invariance was one of our chief weapons in attacking the string, but it doesn't hold for higher dimensional objects.

The conceptual issue is that quantizing a membrane, or higher dimensional object, would not give rise to a discrete spectrum of states which have the interpretation of particles. In this way, they appear to be fundamentally different from the string.

Let's get some intuition for why this is the case. The energy of a string is proportional to its length. This ensures that strings behave more or less like familiar elastic bands. What about D2-branes? Now the energy is proportional to the area. In the back of your mind, you might be thinking of a rubber-like sheet. But membranes, and higher dimensional objects, governed by the Dirac action don't behave as household rubber sheets. They are more flexible. This is because a membrane can form many different shapes with the same area. For example, a tubular membrane of length L and radius $1/L$ has the same area for all values of L ; short and stubby, or long and thin. This means that long thin spikes can develop on a membrane at no extra cost of energy. In particular, objects connected by long thin tubes have the same energy, regardless of their separation. After quantization, this property gives rise to a continuous spectrum of states. A quantum membrane, or higher dimensional object, does not have the single particle interpretation that we saw for the string. The expectation is that the quantum membrane should describe multi-particle states.

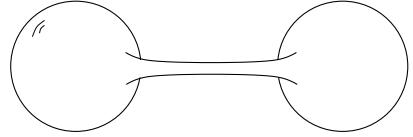


Figure 17:

3.3 Multiple Branes: A World of Glue

Consider two parallel D p -branes. An open string now has options. It could either end on the same brane, or stretch between the two branes. Let's consider the string that stretches between the two. It obeys

$$X^I(0, \tau) = c^I \quad \text{and} \quad X^I(\pi, \tau) = d^I$$

where c^I and d^I are the positions of the two branes. In terms of the mode expansion, this requires

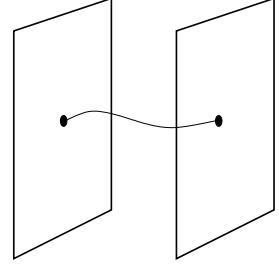


Figure 18:

$$X^I = c^I + \frac{(d^I - c^I)\sigma}{\pi} + \text{oscillator modes}$$

The classical constraints then read

$$\partial_+ X \cdot \partial_+ X = \alpha'^2 p^2 + \frac{|\vec{d} - \vec{c}|^2}{4\pi^2} + \text{oscillator modes} = 0$$

which means the classical mass-shell condition is

$$M^2 = \frac{|\vec{d} - \vec{c}|^2}{(2\pi\alpha')^2} + \text{oscillator modes}$$

The extra term has an obvious interpretation: it is the mass of a classical string stretched between the two branes. The quantization of this string proceeds as before. After we include the normal ordering constant, the ground state of this string is only tachyonic if $|\vec{d} - \vec{c}|^2 < 4\pi^2\alpha'$. Or in other words, the ground state is tachyonic if the branes approach to a sub-stringy distance.

There is an obvious generalization of this to the case of N parallel branes. Each end point of the string has N possible places on which to end. We can label each end point with a number $m, n = 1, \dots, N$ which tell us which brane it ends on. This label is sometimes referred to as a *Chan-Paton factor*.

Consider now the situation where all branes lie at the same position in spacetime. Each end point can lie on one of N different branes, giving N^2 possibilities in total. Each of these strings has the mass spectrum of an open string, meaning that there are now N^2 different particles of each type. It's natural to arrange the associated fields to sit inside $N \times N$ Hermitian matrices. We then have the open string tachyon T_n^m and the massless fields

$$(\phi^I)_n^m , \quad (A_a)_n^m \tag{3.7}$$

Here the components of the matrix tell us which string the field came from. Diagonal components arise from strings which have both ends on the same brane.

The gauge field A_a is particularly interesting. Written in this way, it looks like a $U(N)$ gauge connection. We will later see that this is indeed the case. One can show that as N branes coincide, the $U(1)^N$ gauge symmetry of the branes is enhanced to $U(N)$. The scalar fields ϕ^I transform in the adjoint of this symmetry.

4. Introducing Conformal Field Theory

The purpose of this section is to get comfortable with the basic language of two dimensional conformal field theory⁴. This is a topic which has many applications outside of string theory, most notably in statistical physics where it offers a description of critical phenomena. Moreover, it turns out that conformal field theories in two dimensions provide rare examples of interacting, yet exactly solvable, quantum field theories. In recent years, attention has focussed on conformal field theories in higher dimensions due to their role in the AdS/CFT correspondence.

A *conformal transformation* is a change of coordinates $\sigma^\alpha \rightarrow \tilde{\sigma}^\alpha(\sigma)$ such that the metric changes by

$$g_{\alpha\beta}(\sigma) \rightarrow \Omega^2(\sigma) g_{\alpha\beta}(\sigma) \quad (4.1)$$

A *conformal field theory* (CFT) is a field theory which is invariant under these transformations. This means that the physics of the theory looks the same at all length scales. Conformal field theories care about angles, but not about distances.

A transformation of the form (4.1) has a different interpretation depending on whether we are considering a fixed background metric $g_{\alpha\beta}$, or a dynamical background metric. When the metric is dynamical, the transformation is a diffeomorphism; this is a gauge symmetry. When the background is fixed, the transformation should be thought of as an honest, physical symmetry, taking the point σ^α to point $\tilde{\sigma}^\alpha$. This is now a global symmetry with the corresponding conserved currents.

In the context of string theory in the Polyakov formalism, the metric is dynamical and the transformations (4.1) are residual gauge transformations: diffeomorphisms which can be undone by a Weyl transformation.

In contrast, in this section we will be primarily interested in theories defined on fixed backgrounds. Apart from a few noticeable exceptions, we will usually take this background to be flat. This is the situation that we are used to when studying quantum field theory.

⁴Much of the material covered in this section was first described in the ground breaking paper by Belavin, Polyakov and Zamalodchikov, “*Infinite Conformal Symmetry in Two-Dimensional Quantum Field Theory*”, Nucl. Phys. B241 (1984). The application to string theory was explained by Friedan, Martinec and Shenker in “*Conformal Invariance, Supersymmetry and String Theory*”, Nucl. Phys. B271 (1986). The canonical reference for learning conformal field theory is the excellent review by Ginsparg. A link can be found on the course webpage.

Of course, we can alternate between thinking of theories as defined on fixed or fluctuating backgrounds. Any theory of 2d gravity which enjoys both diffeomorphism and Weyl invariance will reduce to a conformally invariant theory when the background metric is fixed. Similarly, any conformally invariant theory can be coupled to 2d gravity where it will give rise to a classical theory which enjoys both diffeomorphism and Weyl invariance. Notice the caveat “classical”! In some sense, the whole point of this course is to understand when this last statement also holds at the quantum level.

Even though conformal field theories are a subset of quantum field theories, the language used to describe them is a little different. This is partly out of necessity. Invariance under the transformation (4.1) can only hold if the theory has no preferred length scale. But this means that there can be nothing in the theory like a mass or a Compton wavelength. In other words, conformal field theories only support massless excitations. The questions that we ask are not those of particles and S-matrices. Instead we will be concerned with correlation functions and the behaviour of different operators under conformal transformations.

4.0.1 Euclidean Space

Although we’re ultimately interested in Minkowski signature worldsheets, it will be much simpler and elegant if we work instead with Euclidean worldsheets. There’s no funny business here — everything we do could also be formulated in Minkowski space.

The Euclidean worldsheet coordinates are $(\sigma^1, \sigma^2) = (\sigma^1, i\sigma^0)$ and it will prove useful to form the complex coordinates,

$$z = \sigma^1 + i\sigma^2 \quad \text{and} \quad \bar{z} = \sigma^1 - i\sigma^2$$

which are the Euclidean analogue of the lightcone coordinates. Motivated by this analogy, it is common to refer to holomorphic functions as “left-moving” and anti-holomorphic functions as “right-moving”.

The holomorphic derivatives are

$$\partial_z \equiv \partial = \frac{1}{2}(\partial_1 - i\partial_2) \quad \text{and} \quad \partial_{\bar{z}} \equiv \bar{\partial} = \frac{1}{2}(\partial_1 + i\partial_2)$$

These obey $\partial z = \bar{\partial} \bar{z} = 1$ and $\partial \bar{z} = \bar{\partial} z = 0$. We will usually work in flat Euclidean space, with metric

$$ds^2 = (d\sigma^1)^2 + (d\sigma^2)^2 = dz d\bar{z} \tag{4.2}$$

In components, this flat metric reads

$$g_{zz} = g_{\bar{z}\bar{z}} = 0 \quad \text{and} \quad g_{z\bar{z}} = \frac{1}{2}$$

With this convention, the measure factor is $dzd\bar{z} = 2d\sigma^1 d\sigma^2$. We define the delta-function such that $\int d^2z \delta(z, \bar{z}) = 1$. Notice that because we also have $\int d^2\sigma \delta(\sigma) = 1$, this means that there is a factor of 2 difference between the two delta functions. Vectors naturally have their indices up: $v^z = (v^1 + iv^2)$ and $v^{\bar{z}} = (v^1 - iv^2)$. When indices are down, the vectors are $v_z = \frac{1}{2}(v^1 - iv^2)$ and $v_{\bar{z}} = \frac{1}{2}(v^1 + iv^2)$.

4.0.2 The Holomorphy of Conformal Transformations

In the complex Euclidean coordinates z and \bar{z} , conformal transformations of flat space are simple: they are any holomorphic change of coordinates,

$$z \rightarrow z' = f(z) \quad \text{and} \quad \bar{z} \rightarrow \bar{z}' = \bar{f}(\bar{z})$$

Under this transformation, $ds^2 = dzd\bar{z} \rightarrow |df/dz|^2 dzd\bar{z}$, which indeed takes the form (4.1). Note that we have an infinite number of conformal transformations — in fact, a whole functions worth $f(z)$. This is special to conformal field theories in two dimensions. In higher dimensions, the space of conformal transformations is a finite dimensional group. For theories defined on $\mathbf{R}^{p,q}$, the conformal group is $SO(p+1, q+1)$ when $p+q > 2$.

A couple of particularly simple and important examples of 2d conformal transformations are

- $z \rightarrow z + a$: This is a translation.
- $z \rightarrow \zeta z$: This is a rotation for $|\zeta| = 1$ and a scale transformation (also known as a *dilatation*) for real $\zeta \neq 1$.

For many purposes, it's simplest to treat z and \bar{z} as independent variables. In doing this, we're really extending the worldsheet from \mathbf{R}^2 to \mathbf{C}^2 . This will allow us to make use of various theorems from complex methods. However, at the end of the day we should remember that we're really sitting on the real slice $\mathbf{R}^2 \subset \mathbf{C}^2$ defined by $\bar{z} = z^*$.

4.1 Classical Aspects

We start by deriving some properties of classical theories which are invariant under conformal transformations (4.1).

4.1.1 The Stress-Energy Tensor

One of the most important objects in any field theory is the *stress-energy tensor* (also known as the energy-momentum tensor). This is defined in the usual way as the matrix of conserved currents which arise from translational invariance,

$$\delta\sigma^\alpha = \epsilon^\alpha .$$

In flat spacetime, a translation is a special case of a conformal transformation.

There's a cute way to derive the stress-energy tensor in any theory. Suppose for the moment that we are in flat space $g_{\alpha\beta} = \eta_{\alpha\beta}$. Recall that we can usually derive conserved currents by promoting the constant parameter ϵ that appears in the symmetry to a function of the spacetime coordinates. The change in the action must then be of the form,

$$\delta S = \int d^2\sigma J^\alpha \partial_\alpha \epsilon \tag{4.3}$$

for some function of the fields, J^α . This ensures that the variation of the action vanishes when ϵ is constant, which is of course the definition of a symmetry. But when the equations of motion are satisfied, we must have $\delta S = 0$ for all variations $\epsilon(\sigma)$, not just constant ϵ . This means that when the equations of motion are obeyed, J^α must satisfy

$$\partial_\alpha J^\alpha = 0$$

The function J^α is our conserved current.

Let's see how this works for translational invariance. If we promote ϵ to a function of the worldsheet variables, the change of the action must be of the form (4.3). But what is J^α ? At this point we do the cute thing. Consider the same theory, but now coupled to a dynamical background metric $g_{\alpha\beta}(\sigma)$. In other words, coupled to gravity. Then we could view the transformation

$$\delta\sigma^\alpha = \epsilon^\alpha(\sigma)$$

as a diffeomorphism and we know that the theory is invariant as long as we make the corresponding change to the metric

$$\delta g_{\alpha\beta} = \partial_\alpha \epsilon_\beta + \partial_\beta \epsilon_\alpha .$$

This means that if we just make the transformation of the coordinates in our original theory, then the change in the action must be the opposite of what we get if we just

transform the metric. (Because doing both together leaves the action invariant). So we have

$$\delta S = - \int d^2\sigma \frac{\partial S}{\partial g_{\alpha\beta}} \delta g_{\alpha\beta} = -2 \int d^2\sigma \frac{\partial S}{\partial g_{\alpha\beta}} \partial_\alpha \epsilon_\beta$$

Note that $\partial S/\partial g_{\alpha\beta}$ in this expression is really a functional derivatives but we won't be careful about using notation to indicate this. We now have the conserved current arising from translational invariance. We will add a normalization constant which is standard in string theory (although not necessarily in other areas) and define the stress-energy tensor to be

$$T_{\alpha\beta} = -\frac{4\pi}{\sqrt{g}} \frac{\partial S}{\partial g^{\alpha\beta}} \quad (4.4)$$

If we have a flat worldsheet, we evaluate $T_{\alpha\beta}$ on $g_{\alpha\beta} = \delta_{\alpha\beta}$ and the resulting expression obeys $\partial^\alpha T_{\alpha\beta} = 0$. If we're working on a curved worldsheet, then the energy-momentum tensor is covariantly conserved, $\nabla^\alpha T_{\alpha\beta} = 0$.

The Stress-Energy Tensor is Traceless

In conformal theories, $T_{\alpha\beta}$ has a very important property: its trace vanishes. To see this, let's vary the action with respect to a scale transformation which is a special case of a conformal transformation,

$$\delta g_{\alpha\beta} = \epsilon g_{\alpha\beta} \quad (4.5)$$

Then we have

$$\delta S = \int d^2\sigma \frac{\partial S}{\partial g_{\alpha\beta}} \delta g_{\alpha\beta} = -\frac{1}{4\pi} \int d^2\sigma \sqrt{g} \epsilon T_\alpha^\alpha$$

But this must vanish in a conformal theory because scaling transformations are a symmetry. So

$$T_\alpha^\alpha = 0$$

This is the key feature of a conformal field theory in any dimension. Many theories have this feature at the classical level, including Maxwell theory and Yang-Mills theory in four-dimensions. However, it is much harder to preserve at the quantum level. (The weight of the world rests on the fact that Yang-Mills theory fails to be conformal at the quantum level). Technically the difficulty arises due to the need to introduce a scale when regulating the theories. Here we will be interested in two-dimensional theories

which succeed in preserving the conformal symmetry at the quantum level.

Looking Ahead: Even when the conformal invariance survives in a 2d quantum theory, the vanishing trace $T_\alpha^\alpha = 0$ will only turn out to hold in flat space. We will derive this result in section 4.4.2.

The Stress-Tensor in Complex Coordinates

In complex coordinates, $z = \sigma^1 + i\sigma^2$, the vanishing of the trace $T_\alpha^\alpha = 0$ becomes

$$T_{z\bar{z}} = 0$$

Meanwhile, the conservation equation $\partial_\alpha T^{\alpha\beta} = 0$ becomes $\partial T^{zz} = \bar{\partial} T^{\bar{z}\bar{z}} = 0$. Or, lowering the indices on T ,

$$\bar{\partial} T_{zz} = 0 \quad \text{and} \quad \partial T_{\bar{z}\bar{z}} = 0$$

In other words, $T_{zz} = T_{zz}(z)$ is a holomorphic function while $T_{\bar{z}\bar{z}} = T_{\bar{z}\bar{z}}(\bar{z})$ is an anti-holomorphic function. We will often use the simplified notation

$$T_{zz}(z) \equiv T(z) \quad \text{and} \quad T_{\bar{z}\bar{z}}(\bar{z}) \equiv \bar{T}(\bar{z})$$

4.1.2 Noether Currents

The stress-energy tensor $T_{\alpha\beta}$ provides the Noether currents for translations. What are the currents associated to the other conformal transformations? Consider the infinitesimal change,

$$z' = z + \epsilon(z) \quad , \quad \bar{z}' = \bar{z} + \bar{\epsilon}(\bar{z})$$

where, making contact with the two examples above, constant ϵ corresponds to a translation while $\epsilon(z) \sim z$ corresponds to a rotation and dilatation. To compute the current, we'll use the same trick that we saw before: we promote the parameter ϵ to depend on the worldsheet coordinates. But it's already a function of half of the worldsheet coordinates, so this now means $\epsilon(z) \rightarrow \epsilon(z, \bar{z})$. Then we can compute the change in the action, again using the fact that we can make a compensating change in the metric,

$$\begin{aligned} \delta S &= - \int d^2\sigma \frac{\partial S}{\partial g^{\alpha\beta}} \delta g^{\alpha\beta} \\ &= \frac{1}{2\pi} \int d^2\sigma T_{\alpha\beta} (\partial^\alpha \delta\sigma^\beta) \\ &= \frac{1}{2\pi} \int d^2z \frac{1}{2} [T_{zz}(\partial^z \delta z) + T_{\bar{z}\bar{z}}(\partial^{\bar{z}} \delta \bar{z})] \\ &= \frac{1}{2\pi} \int d^2z [T_{zz} \partial_{\bar{z}} \epsilon + T_{\bar{z}\bar{z}} \partial_z \bar{\epsilon}] \end{aligned} \tag{4.6}$$

Firstly note that if ϵ is holomorphic and $\bar{\epsilon}$ is anti-holomorphic, then we immediately have $\delta S = 0$. This, of course, is the statement that we have a symmetry on our hands. (You may wonder where in the above derivation we used the fact that the theory was conformal. It lies in the transition to the third line where we needed $T_{z\bar{z}} = 0$).

At this stage, let's use the trick of treating z and \bar{z} as independent variables. We look at separate currents that come from shifts in z and shifts \bar{z} . Let's first look at the symmetry

$$\delta z = \epsilon(z) , \quad \delta \bar{z} = 0$$

We can read off the conserved current from (4.6) by using the standard trick of letting the small parameter depend on position. Since $\epsilon(z)$ already depends on position, this means promoting $\epsilon \rightarrow \epsilon(z)f(\bar{z})$ for some function f and then looking at the $\bar{\partial}f$ terms in (4.6). This gives us the current

$$J^z = 0 \quad \text{and} \quad J^{\bar{z}} = T_{zz}(z) \epsilon(z) \equiv T(z) \epsilon(z) \quad (4.7)$$

Importantly, we find that the current itself is also holomorphic. We can check that this is indeed a conserved current: it should satisfy $\partial_\alpha J^\alpha = \partial_z J^z + \partial_{\bar{z}} J^{\bar{z}} = 0$. But in fact it does so with room to spare: it satisfies the much stronger condition $\partial_{\bar{z}} J^{\bar{z}} = 0$.

Similarly, we can look at transformations $\delta \bar{z} = \bar{\epsilon}(\bar{z})$ with $\delta z = 0$. We get the anti-holomorphic current \bar{J} ,

$$\bar{J}^z = \bar{T}(\bar{z}) \bar{\epsilon}(\bar{z}) \quad \text{and} \quad \bar{J}^{\bar{z}} = 0 \quad (4.8)$$

4.1.3 An Example: The Free Scalar Field

Let's illustrate some of these ideas about classical conformal theories with the free scalar field,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \partial_\alpha X \partial^\alpha X$$

Notice that there's no overall minus sign, in contrast to our earlier action (1.30). That's because we're now working with a Euclidean worldsheet metric. The theory of a free scalar field is, of course, dead easy. We can compute anything we like in this theory. Nonetheless, it will still exhibit enough structure to provide an example of all the abstract concepts that we will come across in CFT. For this reason, the free scalar field will prove a good companion throughout this part of the lectures.

Firstly, let's just check that this free scalar field is actually conformal. In particular, we can look at rescaling $\sigma^\alpha \rightarrow \lambda\sigma^\alpha$. If we view this in the sense of an active transformation, the coordinates remain fixed but the value of the field at point σ gets moved to point $\lambda\sigma$. This means,

$$X(\sigma) \rightarrow X(\lambda^{-1}\sigma) \quad \text{and} \quad \frac{\partial X(\sigma)}{\partial \sigma^\alpha} \rightarrow \frac{\partial X(\lambda^{-1}\sigma)}{\partial \sigma^\alpha} = \frac{1}{\lambda} \frac{\partial X(\tilde{\sigma})}{\partial \tilde{\sigma}}$$

where we've defined $\tilde{\sigma} = \lambda^{-1}\sigma$. The factor of λ^{-2} coming from the two derivatives in the Lagrangian then cancels the Jacobian factor from the measure $d^2\sigma = \lambda^2 d^2\tilde{\sigma}$, leaving the action invariant. Note that any polynomial interaction term for X would break conformal invariance.

The stress-energy tensor for this theory is defined using (4.4),

$$T_{\alpha\beta} = -\frac{1}{\alpha'} \left(\partial_\alpha X \partial_\beta X - \frac{1}{2} \delta_{\alpha\beta} (\partial X)^2 \right), \quad (4.9)$$

which indeed satisfies $T_\alpha^\alpha = 0$ as it should. The stress-energy tensor looks much simpler in complex coordinates. It is simple to check that $T_{z\bar{z}} = 0$ while

$$T = -\frac{1}{\alpha'} \partial X \partial X \quad \text{and} \quad \bar{T} = -\frac{1}{\alpha'} \bar{\partial} X \bar{\partial} X$$

The equation of motion for X is $\partial\bar{\partial}X = 0$. The general classical solution decomposes as,

$$X(z, \bar{z}) = X(z) + \bar{X}(\bar{z})$$

When evaluated on this solution, T and \bar{T} become holomorphic and anti-holomorphic functions respectively.

4.2 Quantum Aspects

So far our discussion has been entirely classical. We now turn to the quantum theory. The first concept that we want to discuss is actually a feature of any quantum field theory. But it really comes into its own in the context of CFT: it is the *operator product expansion*.

4.2.1 Operator Product Expansion

Let's first describe what we mean by a *local* operator in a CFT. We will also refer to these objects as *fields*. There is a slight difference in terminology between CFTs and more general quantum field theories. Usually in quantum field theory, one reserves the

term ‘‘field’’ for the objects ϕ which sit in the action and are integrated over in the path integral. In contrast, in CFT the term ‘‘field’’ refers to any local expression that we can write down. This includes ϕ , but also includes derivatives $\partial^n\phi$ or composite operators such as $e^{i\phi}$. All of these are thought of as different fields in a CFT. It should be clear from this that the set of all ‘‘fields’’ in a CFT is always infinite even though, if you were used to working with quantum field theory, you would talk about only a finite number of fundamental objects ϕ . Obviously, this is nothing to be scared about. It’s just a change of language: it doesn’t mean that our theory got harder.

We now define the *operator product expansion* (OPE). It is a statement about what happens as local operators approach each other. The idea is that two local operators inserted at nearby points can be closely approximated by a string of operators at one of these points. Let’s denote all the local operators of the CFT by \mathcal{O}_i , where i runs over the set of all operators. Then the OPE is

$$\mathcal{O}_i(z, \bar{z}) \mathcal{O}_j(w, \bar{w}) = \sum_k C_{ij}^k(z - w, \bar{z} - \bar{w}) \mathcal{O}_k(w, \bar{w}) \quad (4.10)$$

Here $C_{ij}^k(z - w, \bar{z} - \bar{w})$ are a set of functions which, on grounds of translational invariance, depend only on the separation between the two operators. We will write a lot of operator equations of the form (4.10) and it’s important to clarify exactly what they mean: they are always to be understood as statements which hold as operator insertions inside time-ordered correlation functions,

$$\langle \mathcal{O}_i(z, \bar{z}) \mathcal{O}_j(w, \bar{w}) \dots \rangle = \sum_k C_{ij}^k(z - w, \bar{z} - \bar{w}) \langle \mathcal{O}_k(w, \bar{w}) \dots \rangle$$

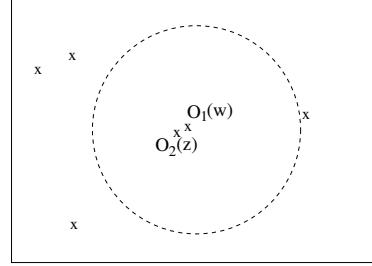


Figure 19:

where the \dots can be any other operator insertions that we choose. Obviously it would be tedious to continually write $\langle \dots \rangle$. So we don’t. But it’s always implicitly there. There are further caveats about the OPE that are worth stressing

- The correlation functions are always assumed to be time-ordered. (Or something similar that we will discuss in Section 4.5.1). This means that as far as the OPE is concerned, everything commutes since the ordering of operators is determined inside the correlation function anyway. So we must have $\mathcal{O}_i(z, \bar{z}) \mathcal{O}_j(w, \bar{w}) = \mathcal{O}_j(w, \bar{w}) \mathcal{O}_i(z, \bar{z})$. (There is a caveat here: if the operators are Grassmann objects, then they pick up an extra minus sign when commuted, even inside time-ordered products).

- The other operator insertions in the correlation function (denoted \dots above) are arbitrary. *Except* they should be at a distance large compared to $|z - w|$. It turns out — rather remarkably — that in a CFT the OPEs are exact statements and have a radius of convergence equal to the distance to the nearest other insertion. We will return to this in Section 4.6. The radius of convergence is denoted in the figure by the dotted line.
- The OPEs have singular behaviour as $z \rightarrow w$. In fact, this singular behaviour will really be the only thing we care about! It will turn out to contain the same information as commutation relations, as well as telling us how operators transform under symmetries. Indeed, in many equations we will simply write the singular terms in the OPE and denote the non-singular terms as $+ \dots$.

4.2.2 Ward Identities

The spirit of Noether's theorem in quantum field theories is captured by operator equations known as *Ward Identities*. Here we derive the Ward identities associated to conformal invariance. We start by considering a general theory with a symmetry. Later we will restrict to conformal symmetries.

Games with Path Integrals

We'll take this opportunity to get comfortable with some basic techniques using path integrals. Schematically, the path integral takes the form

$$Z = \int \mathcal{D}\phi e^{-S[\phi]}$$

where ϕ collectively denote all the fields (in the path integral sense...not the CFT sense!). A symmetry of the quantum theory is such that an infinitesimal transformation

$$\phi' = \phi + \epsilon\delta\phi$$

leaves both the action *and* the measure invariant,

$$S[\phi'] = S[\phi] \quad \text{and} \quad \mathcal{D}\phi' = \mathcal{D}\phi$$

(In fact, we only really need the combination $\mathcal{D}\phi e^{-S[\phi]}$ to be invariant but this subtlety won't matter in this course). We use the same trick that we employed earlier in the classical theory and promote $\epsilon \rightarrow \epsilon(\sigma)$. Then, typically, neither the action nor the measure are invariant but, to leading order in ϵ , the change has to be proportional to

$\partial\epsilon$. We have

$$\begin{aligned} Z &\longrightarrow \int \mathcal{D}\phi' \exp(-S[\phi']) \\ &= \int \mathcal{D}\phi \exp\left(-S[\phi] - \frac{1}{2\pi} \int J^\alpha \partial_\alpha \epsilon\right) \\ &= \int \mathcal{D}\phi e^{-S[\phi]} \left(1 - \frac{1}{2\pi} \int J^\alpha \partial_\alpha \epsilon\right) \end{aligned}$$

where the factor of $1/2\pi$ is merely a convention and \int is shorthand for $\int d^2\sigma \sqrt{g}$. Notice that the current J^α may now also have contributions from the measure transformation as well as the action.

Now comes the clever step. Although the integrand has changed, the actual value of the partition function can't have changed at all. After all, we just redefined a dummy integration variable ϕ . So the expression above must be equal to the original Z . Or, in other words,

$$\int \mathcal{D}\phi e^{-S[\phi]} \left(\int J^\alpha \partial_\alpha \epsilon \right) = 0$$

Moreover, this must hold for all ϵ . This gives us the quantum version of Noether's theorem: the vacuum expectation value of the divergence of the current vanishes:

$$\langle \partial_\alpha J^\alpha \rangle = 0 .$$

We can repeat these tricks of this sort to derive some stronger statements. Let's see what happens when we have other insertions in the path integral. The time-ordered correlation function is given by

$$\langle \mathcal{O}_1(\sigma_1) \dots \mathcal{O}_n(\sigma_n) \rangle = \frac{1}{Z} \int \mathcal{D}\phi e^{-S[\phi]} \mathcal{O}_1(\sigma_1) \dots \mathcal{O}_n(\sigma_n)$$

We can think of these as operators inserted at particular points on the plane as shown in the figure. As we described above, the operators \mathcal{O}_i are any general expressions that we can form from the ϕ fields. Under the symmetry of interest, the operator will change in some way, say

$$\mathcal{O}_i \rightarrow \mathcal{O}_i + \epsilon \delta \mathcal{O}_i$$

We once again promote $\epsilon \rightarrow \epsilon(\sigma)$. As our first pass, let's pick a choice of $\epsilon(\sigma)$ which only has support away from the operator insertions as shown in the Figure 20. Then,

$$\delta \mathcal{O}_i(\sigma_i) = 0$$

and the above derivation goes through in exactly the same way to give

$$\langle \partial_\alpha J^\alpha(\sigma) \mathcal{O}_1(\sigma_1) \dots \mathcal{O}_n(\sigma_n) \rangle = 0 \quad \text{for } \sigma \neq \sigma_i$$

Because this holds for any operator insertions away from σ , from the discussion in Section 4.2.1 we are entitled to write the operator equation

$$\partial_\alpha J^\alpha = 0$$

But what if there are operator insertions that lie at the same point as J^α ? In other words, what happens as σ approaches one of the insertion points? The resulting formulae are called Ward identities. To derive these, let's take $\epsilon(\sigma)$ to have support in some region that includes the point σ_1 , but not the other points as shown in Figure 21. The simplest choice is just to take $\epsilon(\sigma)$ to be constant inside the shaded region and zero outside. Now using the same procedure as before, we find that the original correlation function is equal to,

$$\frac{1}{Z} \int \mathcal{D}\phi e^{-S[\phi]} \left(1 - \frac{1}{2\pi} \int J^\alpha \partial_\alpha \epsilon \right) (\mathcal{O}_1 + \epsilon \delta \mathcal{O}_1) \mathcal{O}_2 \dots \mathcal{O}_n$$

Working to leading order in ϵ , this gives

$$-\frac{1}{2\pi} \int_\epsilon \partial_\alpha \langle J^\alpha(\sigma) \mathcal{O}_1(\sigma_1) \dots \rangle = \langle \delta \mathcal{O}_1(\sigma_1) \dots \rangle \quad (4.11)$$

where the integral on the left-hand-side is only over the region of non-zero ϵ . This is the *Ward Identity*.

Ward Identities for Conformal Transformations

Ward identities (4.11) hold for any symmetries. Let's now see what they give when applied to conformal transformations. There are two further steps needed in the derivation. The first simply comes from the fact that we're working in two dimensions and we can use Stokes' theorem to convert the integral on the left-hand-side of (4.11) to a line integral around the boundary. Let \hat{n}^α be the unit vector normal to the boundary. For any vector J^α , we have

$$\int_\epsilon \partial_\alpha J^\alpha = \oint_{\partial\epsilon} J_\alpha \hat{n}^\alpha = \oint_{\partial\epsilon} (J_1 d\sigma^2 - J_2 d\sigma^1) = -i \oint_{\partial\epsilon} (J_z dz - J_{\bar{z}} d\bar{z})$$

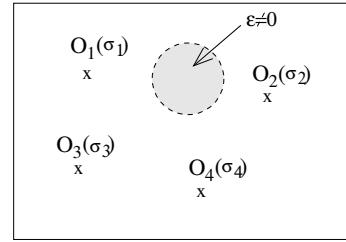


Figure 20:

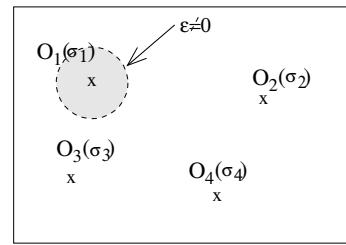


Figure 21:

where we have written the expression both in Cartesian coordinates σ^α and complex coordinates on the plane. As described in Section 4.0.1, the complex components of the vector with indices down are defined as $J_z = \frac{1}{2}(J_1 - iJ_2)$ and $J_{\bar{z}} = \frac{1}{2}(J_1 + iJ_2)$. So, applying this to the Ward identity (4.11), we find for two dimensional theories

$$\frac{i}{2\pi} \oint_{\partial\epsilon} dz \langle J_z(z, \bar{z}) \mathcal{O}_1(\sigma_1) \dots \rangle - \frac{i}{2\pi} \oint_{\partial\epsilon} d\bar{z} \langle J_{\bar{z}}(z, \bar{z}) \mathcal{O}_1(\sigma_1) \dots \rangle = \langle \delta \mathcal{O}_1(\sigma_1) \dots \rangle$$

So far our derivation holds for any conserved current J in two dimensions. At this stage we specialize to the currents that arise from conformal transformations (4.7) and (4.8). Here something nice happens because J_z is holomorphic while $J_{\bar{z}}$ is anti-holomorphic. This means that the contour integral simply picks up the residue,

$$\frac{i}{2\pi} \oint_{\partial\epsilon} dz J_z(z) \mathcal{O}_1(\sigma_1) = -\text{Res}[J_z \mathcal{O}_1]$$

where this means the residue in the OPE between the two operators,

$$J_z(z) \mathcal{O}_1(w, \bar{w}) = \dots + \frac{\text{Res}[J_z \mathcal{O}_1(w, \bar{w})]}{z - w} + \dots$$

So we find a rather nice way of writing the Ward identities for conformal transformations. If we again view z and \bar{z} as independent variables, the Ward identities split into two pieces. From the change $\delta z = \epsilon(z)$, we get

$$\delta \mathcal{O}_1(\sigma_1) = -\text{Res}[J_z(z) \mathcal{O}_1(\sigma_1)] = -\text{Res}[\epsilon(z) T(z) \mathcal{O}_1(\sigma_1)] \quad (4.12)$$

where, in the second equality, we have used the expression for the conformal current (4.7). Meanwhile, from the change $\delta \bar{z} = \bar{\epsilon}(\bar{z})$, we have

$$\delta \mathcal{O}_1(\sigma_1) = -\text{Res}[\bar{J}_{\bar{z}}(\bar{z}) \mathcal{O}_1(\sigma_1)] = -\text{Res}[\bar{\epsilon}(\bar{z}) \bar{T}(\bar{z}) \mathcal{O}_1(\sigma_1)]$$

where the minus sign comes from the fact that the $\oint d\bar{z}$ boundary integral is taken in the opposite direction.

This result means that if we know the OPE between an operator and the stress-tensors $T(z)$ and $\bar{T}(\bar{z})$, then we immediately know how the operator transforms under conformal symmetry. Or, standing this on its head, if we know how an operator transforms then we know at least some part of its OPE with T and \bar{T} .

4.2.3 Primary Operators

The Ward identity allows us to start piecing together some OPEs by looking at how operators transform under conformal symmetries. Although we don't yet know the

action of general conformal symmetries, we can start to make progress by looking at the two simplest examples.

Translations: If $\delta z = \epsilon$, a constant, then all operators transform as

$$\mathcal{O}(z - \epsilon) = \mathcal{O}(z) - \epsilon \partial \mathcal{O}(z) + \dots$$

The Noether current for translations is the stress-energy tensor T . The Ward identity in the form (4.12) tells us that the OPE of T with any operator \mathcal{O} must be of the form,

$$T(z) \mathcal{O}(w, \bar{w}) = \dots + \frac{\partial \mathcal{O}(w, \bar{w})}{z - w} + \dots \quad (4.13)$$

Similarly, the OPE with \bar{T} is

$$\bar{T}(\bar{z}) \mathcal{O}(w, \bar{w}) = \dots + \frac{\bar{\partial} \mathcal{O}(w, \bar{w})}{\bar{z} - \bar{w}} + \dots \quad (4.14)$$

Rotations and Scaling: The transformation

$$z \rightarrow z + \epsilon z \quad \text{and} \quad \bar{z} \rightarrow \bar{z} + \bar{\epsilon} \bar{z} \quad (4.15)$$

describes rotation for ϵ purely imaginary and scaling (dilatation) for ϵ real. Not all operators have good transformation properties under these actions. This is entirely analogous to the statement in quantum mechanics that not all states transform nicely under the Hamiltonian H and angular momentum operator L . However, in quantum mechanics we know that the eigenstates of H and L can be chosen as a basis of the Hilbert space provided, of course, that $[H, L] = 0$.

The same statement holds for operators in a CFT: we can choose a basis of local operators that have good transformation properties under rotations and dilatations. In fact, we will see in Section 4.6 that the statement about local operators actually follows from the statement about states.

Definition: An operator \mathcal{O} is said to have *weight* (h, \tilde{h}) if, under $\delta z = \epsilon z$ and $\delta \bar{z} = \bar{\epsilon} \bar{z}$, \mathcal{O} transforms as

$$\delta \mathcal{O} = -\epsilon(h\mathcal{O} + z \partial \mathcal{O}) - \bar{\epsilon}(\tilde{h}\mathcal{O} + \bar{z} \bar{\partial} \mathcal{O}) \quad (4.16)$$

The terms $\partial \mathcal{O}$ in this expression would be there for any operator. They simply come from expanding $\mathcal{O}(z - \epsilon z, \bar{z} - \bar{\epsilon} \bar{z})$. The terms $h\mathcal{O}$ and $\tilde{h}\mathcal{O}$ are special to operators which are eigenstates of dilatations and rotations. Some comments:

- Both h and \tilde{h} are real numbers. In a unitary CFT, all operators have $h, \tilde{h} \geq 0$. We will prove this in Section 4.5.4.
- The weights are not as unfamiliar as they appear. They simply tell us how operators transform under rotations and scalings. But we already have names for these concepts from undergraduate days. The eigenvalue under rotation is usually called the *spin*, s , and is given in terms of the weights as

$$s = h - \tilde{h}$$

Meanwhile, the *scaling dimension* Δ of an operator is

$$\Delta = h + \tilde{h}$$

- To motivate these definitions, it's worth recalling how rotations and scale transformations act on the underlying coordinates. Rotations are implemented by the operator

$$L = -i(\sigma^1 \partial_2 - \sigma^2 \partial_1) = z\partial - \bar{z}\bar{\partial}$$

while the dilation operator D which gives rise to scalings is

$$D = \sigma^\alpha \partial_\alpha = z\partial + \bar{z}\bar{\partial}$$

- The scaling dimension is nothing more than the familiar “dimension” that we usually associate to fields and operators by dimensional analysis. For example, worldsheet derivatives always increase the dimension of an operator by one: $\Delta[\partial] = +1$. The tricky part is that the naive dimension that fields have in the classical theory is not necessarily the same as the dimension in the quantum theory.

Let's compare the transformation law (4.16) with the Ward identity (4.12). The Noether current arising from rotations and scaling $\delta z = \epsilon z$ was given in (4.7): it is $J(z) = zT(z)$. This means that the residue of the $J\mathcal{O}$ OPE will determine the $1/z^2$ term in the $T\mathcal{O}$ OPE. Similar arguments hold, of course, for $\delta\bar{z} = \bar{\epsilon}\bar{z}$ and \bar{T} . So, the upshot of this is that, for an operator \mathcal{O} with weight (h, \tilde{h}) , the OPE with T and \bar{T} takes the form

$$\begin{aligned} T(z)\mathcal{O}(w, \bar{w}) &= \dots + h \frac{\mathcal{O}(w, \bar{w})}{(z-w)^2} + \frac{\partial\mathcal{O}(w, \bar{w})}{z-w} + \dots \\ \bar{T}(\bar{z})\mathcal{O}(w, \bar{w}) &= \dots + \tilde{h} \frac{\mathcal{O}(w, \bar{w})}{(\bar{z}-\bar{w})^2} + \frac{\bar{\partial}\mathcal{O}(w, \bar{w})}{\bar{z}-\bar{w}} + \dots \end{aligned}$$

Primary Operators

A *primary* operator is one whose OPE with T and \bar{T} truncates at order $(z - w)^{-2}$ or order $(\bar{z} - \bar{w})^{-2}$ respectively. There are no higher singularities:

$$T(z) \mathcal{O}(w, \bar{w}) = h \frac{\mathcal{O}(w, \bar{w})}{(z - w)^2} + \frac{\partial \mathcal{O}(w, \bar{w})}{z - w} + \text{non-singular}$$

$$\bar{T}(\bar{z}) \mathcal{O}(w, \bar{w}) = \tilde{h} \frac{\mathcal{O}(w, \bar{w})}{(\bar{z} - \bar{w})^2} + \frac{\bar{\partial} \mathcal{O}(w, \bar{w})}{\bar{z} - \bar{w}} + \text{non-singular}$$

Since we now know all singularities in the $T\mathcal{O}$ OPE, we can reconstruct the transformation under all conformal transformations. The importance of primary operators is that they have particularly simple transformation properties. Focussing on $\delta z = \epsilon(z)$, we have

$$\begin{aligned} \delta \mathcal{O}(w, \bar{w}) &= -\text{Res} [\epsilon(z) T(z) \mathcal{O}(w, \bar{w})] \\ &= -\text{Res} \left[\epsilon(z) \left(h \frac{\mathcal{O}(w, \bar{w})}{(z - w)^2} + \frac{\partial \mathcal{O}(w, \bar{w})}{z - w} + \dots \right) \right] \end{aligned}$$

We want to look at smooth conformal transformations and so require that $\epsilon(z)$ itself has no singularities at $z = w$. We can then Taylor expand

$$\epsilon(z) = \epsilon(w) + \epsilon'(w)(z - w) + \dots$$

We learn that the infinitesimal change of a primary operator under a general conformal transformation $\delta z = \epsilon(z)$ is

$$\delta \mathcal{O}(w, \bar{w}) = -h\epsilon'(w) \mathcal{O}(w, \bar{w}) - \epsilon(w) \partial \mathcal{O}(w, \bar{w}) \quad (4.17)$$

There is a similar expression for the anti-holomorphic transformations $\delta \bar{z} = \bar{\epsilon}(\bar{z})$.

Equation (4.17) holds for infinitesimal conformal transformations. It is a simple matter to integrate up to find how primary operators change under a finite conformal transformation,

$$z \rightarrow \tilde{z}(z) \quad \text{and} \quad \bar{z} \rightarrow \bar{\tilde{z}}(\bar{z})$$

The general transformation of a primary operator is given by

$$\mathcal{O}(z, \bar{z}) \rightarrow \tilde{\mathcal{O}}(\tilde{z}, \bar{\tilde{z}}) = \left(\frac{\partial \tilde{z}}{\partial z} \right)^{-h} \left(\frac{\partial \bar{\tilde{z}}}{\partial \bar{z}} \right)^{-\tilde{h}} \mathcal{O}(z, \bar{z}) \quad (4.18)$$

It will turn out that one of the main objects of interest in a CFT is the spectrum of weights (h, \tilde{h}) of primary fields. This will be equivalent to computing the particle mass spectrum in a quantum field theory. In the context of statistical mechanics, the weights of primary operators are the critical exponents.

4.3 An Example: The Free Scalar Field

Let's look at how all of this works for the free scalar field. We'll start by familiarizing ourselves with some techniques using the path integral. The action is,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \partial_\alpha X \partial^\alpha X \quad (4.19)$$

The classical equation of motion is $\partial^2 X = 0$. Let's start by seeing how to derive the analogous statement in the quantum theory using the path integral. The key fact that we'll need is that the integral of a total derivative vanishes in the path integral just as it does in an ordinary integral. From this we have,

$$0 = \int \mathcal{D}X \frac{\delta}{\delta X(\sigma)} e^{-S} = \int \mathcal{D}X e^{-S} \left[\frac{1}{2\pi\alpha'} \partial^2 X(\sigma) \right]$$

But this is nothing more than the Ehrenfest theorem which states that expectation values of operators obey the classical equations of motion,

$$\langle \partial^2 X(\sigma) \rangle = 0$$

4.3.1 The Propagator

The next thing that we want to do is compute the propagator for X . We could do this using canonical quantization, but it will be useful to again see how it works using the path integral. This time we look at,

$$0 = \int \mathcal{D}X \frac{\delta}{\delta X(\sigma')} [e^{-S} X(\sigma')] = \int \mathcal{D}X e^{-S} \left[\frac{1}{2\pi\alpha'} \partial^2 X(\sigma) X(\sigma') + \delta(\sigma - \sigma') \right]$$

So this time we learn that

$$\langle \partial^2 X(\sigma) X(\sigma') \rangle = -2\pi\alpha' \delta(\sigma - \sigma') \quad (4.20)$$

Note that if we'd computed this in the canonical approach, we would have found the same answer: the δ -function arises in this calculation because all correlation functions are time-ordered.

We can now treat (4.20) as a differential equation for the propagator $\langle X(\sigma) X(\sigma') \rangle$. To solve this equation, we need the following standard result

$$\partial^2 \ln(\sigma - \sigma')^2 = 4\pi\delta(\sigma - \sigma') \quad (4.21)$$

Since this is important, let's just quickly check that it's true. It's a simple application of Stokes' theorem. Set $\sigma' = 0$ and integrate over $\int d^2\sigma$. We obviously get 4π from the right-hand-side. The left-hand-side gives

$$\int d^2\sigma \partial^2 \ln(\sigma_1^2 + \sigma_2^2) = \int d^2\sigma \partial^\alpha \left(\frac{2\sigma_\alpha}{\sigma_1^2 + \sigma_2^2} \right) = 2 \oint \frac{(\sigma_1 d\sigma^2 - \sigma_2 d\sigma^1)}{\sigma_1^2 + \sigma_2^2}$$

Switching to polar coordinates $\sigma_1 + i\sigma_2 = re^{i\theta}$, we can rewrite this expression as

$$2 \int \frac{r^2 d\theta}{r^2} = 4\pi$$

confirming (4.21). Applying this result to our equation (4.20), we get the propagator of a free scalar in two-dimensions,

$$\langle X(\sigma)X(\sigma') \rangle = -\frac{\alpha'}{2} \ln(\sigma - \sigma')^2$$

The propagator has a singularity as $\sigma \rightarrow \sigma'$. This is an ultra-violet divergence and is common to all field theories. It also has a singularity as $|\sigma - \sigma'| \rightarrow \infty$. This is telling us something important that we will mention in Section 4.3.2.

Finally, we could repeat our trick of looking at total derivatives in the path integral, now with other operator insertions $\mathcal{O}_1(\sigma_1), \dots, \mathcal{O}_n(\sigma_n)$ in the path integral. As long as $\sigma, \sigma' \neq \sigma_i$, then the whole analysis goes through as before. But this is exactly our criterion to write the operator product equation,

$$X(\sigma)X(\sigma') = -\frac{\alpha'}{2} \ln(\sigma - \sigma')^2 + \dots \quad (4.22)$$

We can also write this in complex coordinates. The classical equation of motion $\partial\bar{\partial}X = 0$ allows us to split the operator X into left-moving and right-moving pieces,

$$X(z, \bar{z}) = X(z) + \bar{X}(\bar{z})$$

We'll focus just on the left-moving piece. This has the operator product expansion,

$$X(z)X(w) = -\frac{\alpha'}{2} \ln(z - w) + \dots$$

The logarithm means that $X(z)$ doesn't have any nice properties under the conformal transformations. For this reason, the "fundamental field" X is not really the object of interest in this theory! However, we can look at the derivative of X . This has a rather nice looking OPE,

$$\partial X(z) \partial X(w) = -\frac{\alpha'}{2} \frac{1}{(z - w)^2} + \text{non-singular} \quad (4.23)$$

4.3.2 An Aside: No Goldstone Bosons in Two Dimensions

The infra-red divergence in the propagator has an important physical implication. Let's start by pointing out one of the big differences between quantum mechanics and quantum field theory in $d = 3 + 1$ dimensions. Since the language used to describe these two theories is rather different, you may not even be aware that this difference exists.

Consider the quantum mechanics of a particle on a line. This is a $d = 0 + 1$ dimensional theory of a free scalar field X . Let's prepare the particle in some localized state – say a Gaussian wavefunction $\Psi(X) \sim \exp(-X^2/L^2)$. What then happens? The wavefunction starts to spread out. And the spreading doesn't stop. In fact, the would-be ground state of the system is a uniform wavefunction of infinite width, which isn't a state in the Hilbert space because it is non-normalizable.

Let's now compare this to the situation of a free scalar field X in a $d = 3 + 1$ dimensional field theory. Now we think of this as a scalar without potential. The physics is very different: the theory has an infinite number of ground states, determined by the expectation value $\langle X \rangle$. Small fluctuations around this vacuum are massless: they are Goldstone bosons for broken translational invariance $X \rightarrow X + c$.

We see that the physics is very different in field theories in $d = 0 + 1$ and $d = 3 + 1$ dimensions. The wavefunction spreads along flat directions in quantum mechanics, but not in higher dimensional field theories. But what happens in $d = 1 + 1$ and $d = 2 + 1$ dimensions? It turns out that field theories in $d = 1 + 1$ dimensions are more like quantum mechanics: the wavefunction spreads. Theories in $d = 2 + 1$ dimensions and higher exhibit the opposite behaviour: they have Goldstone bosons. The place to see this is the propagator. In d spacetime dimensions, it takes the form

$$\langle X(r) X(0) \rangle \sim \begin{cases} 1/r^{d-2} & d \neq 2 \\ \ln r & d = 2 \end{cases}$$

which diverges at large r only for $d = 1$ and $d = 2$. If we perturb the vacuum slightly by inserting the operator $X(0)$, this correlation function tells us how this perturbation falls off with distance. The infra-red divergence in low dimensions is telling us that the wavefunction wants to spread.

The spreading of the wavefunction in low dimensions means that there is no spontaneous symmetry breaking and no Goldstone bosons. It is usually referred to as the Coleman-Mermin-Wagner theorem. Note, however, that it certainly doesn't prohibit massless excitations in two dimensions: it only prohibits Goldstone-like massless excitations.

4.3.3 The Stress-Energy Tensor and Primary Operators

We want to compute the OPE of T with other operators. Firstly, what is T ? We computed it in the classical theory in (4.9). It is,

$$T = -\frac{1}{\alpha'} \partial X \partial X \quad (4.24)$$

But we need to be careful about what this means in the quantum theory. It involves the product of two operators defined at the same point and this is bound to mean divergences if we just treat it naively. In canonical quantization, we would be tempted to normal order by putting all annihilation operators to the right. This guarantees that the vacuum has zero energy. Here we do something that is basically equivalent, but without reference to creation and annihilation operators. We write

$$T = -\frac{1}{\alpha'} : \partial X \partial X : \equiv -\frac{1}{\alpha'} \lim_{z \rightarrow w} (\partial X(z) \partial X(w) - \langle \partial X(z) \partial X(w) \rangle) \quad (4.25)$$

which, by construction, has $\langle T \rangle = 0$.

With this definition of T , let's start to compute the OPEs to determine the primary fields in the theory.

Claim 1: ∂X is a primary field with weight $h = 1$ and $\tilde{h} = 0$.

Proof: We need to figure out how to take products of normal ordered operators

$$T(z) \partial X(w) = -\frac{1}{\alpha'} : \partial X(z) \partial X(z) : \partial X(w)$$

The operators on the left-hand side are time-ordered (because all operator expressions of this type are taken to live inside time-ordered correlation functions). In contrast, the right-hand side is a product of normal-ordered operators. But we know how to change normal ordered products into time ordered products: this is the content of Wick's theorem. Although we have defined normal ordering in (4.25) without reference to creation and annihilation operators, Wick's theorem still holds. We must sum over all possible contractions of pairs of operators, where the term “contraction” means that we replace the pair by the propagator,

$$\overbrace{\partial X(z) \partial X(w)}^{} = -\frac{\alpha'}{2} \frac{1}{(z-w)^2}$$

Using this, we have

$$T(z) \partial X(w) = -\frac{2}{\alpha'} \partial X(z) \left(-\frac{\alpha'}{2} \frac{1}{(z-w)^2} + \text{non-singular} \right)$$

Here the “non-singular” piece includes the totally normal ordered term $:T(z)\partial X(w):$. It is only the singular part that interests us. Continuing, we have

$$T(z)\partial X(w) = \frac{\partial X(z)}{(z-w)^2} + \dots = \frac{\partial X(w)}{(z-w)^2} + \frac{\partial^2 X(w)}{z-w} + \dots$$

This is indeed the OPE for a primary operator of weight $h = 1$. \square

Note that higher derivatives $\partial^n X$ are not primary for $n > 1$. For example, $\partial^2 X$ has weight $(h, \tilde{h}) = (2, 0)$, but is not a primary operator, as we see from the OPE,

$$T(z)\partial^2 X(w) = \partial_w \left[\frac{\partial X(w)}{(z-w)^2} + \dots \right] = \frac{2\partial X(w)}{(z-w)^3} + \frac{2\partial^2 X(w)}{(z-w)^2} + \dots$$

The fact that the field $\partial^n X$ has weight $(h, \tilde{h}) = (n, 0)$ fits our natural intuition: each derivative provides spin $s = 1$ and dimension $\Delta = 1$, while the field X does not appear to be contributing, presumably reflecting the fact that it has naive, classical dimension zero. However, in the quantum theory, it is not correct to say that X has vanishing dimension: it has an ill-defined dimension due to the logarithmic behaviour of its OPE (4.22). This is responsible for the following, more surprising, result

Claim 2: The field $:e^{ikX}:$ is primary with weight $h = \tilde{h} = \alpha'k^2/4$.

This result is not what we would guess from the classical theory⁵. Indeed, it’s obvious that it has a quantum origin because the weight is proportional to α' , which sits outside the action in the same place that \hbar would (if we hadn’t set it to one). Note also that this means that the spectrum of the free scalar field is continuous. This is related to the fact that the range of X is non-compact. Generally, CFTs will have a discrete spectrum.

Proof: Let’s first compute the OPE with ∂X . We have

$$\begin{aligned} \partial X(z) :e^{ikX(w)}: &= \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \partial X(z) :X(w)^n: \\ &= \sum_{n=1}^{\infty} \frac{(ik)^n}{(n-1)!} :X(w)^{n-1}: \left(-\frac{\alpha'}{2} \frac{1}{z-w} \right) + \dots \\ &= -\frac{i\alpha'k}{2} \frac{:e^{ikX(w)}:}{z-w} + \dots \end{aligned} \tag{4.26}$$

⁵We could, however, guess it with a little knowledge of renormalisation. Indeed, we previously derived this result in the lectures on [Statistical Field Theory](#) where we computed RG flows in the Sine-Gordon model; see Section 4.4.3 of those lectures.

From this, we can compute the OPE with T .

$$\begin{aligned} T(z) : e^{ikX(w)} : &= -\frac{1}{\alpha'} : \partial X(z) \partial X(z) : : e^{ikX(w)} : \\ &= \frac{\alpha' k^2}{4} \frac{: e^{ikX(w)} :}{(z-w)^2} + ik \frac{: \partial X(z) e^{ikX(w)} :}{z-w} + \dots \end{aligned}$$

where the first term comes from two contractions, while the second term comes from a single contraction. Replacing ∂_z by ∂_w in the final term we get

$$T(z) : e^{ikX(w)} : = \frac{\alpha' k^2}{4} \frac{: e^{ikX(w)} :}{(z-w)^2} + \frac{\partial_w : e^{ikX(w)} :}{z-w} + \dots \quad (4.27)$$

showing that $: e^{ikX(w)} :$ is indeed primary. We will encounter this operator frequently later, but will choose to simplify notation and drop the normal ordering colons. Normal ordering will just be assumed from now on. \square .

Finally, lets check to see the OPE of T with itself. This is again just an exercise in Wick contractions.

$$\begin{aligned} T(z) T(w) &= \frac{1}{\alpha'^2} : \partial X(z) \partial X(z) : : \partial X(w) \partial X(w) : \\ &= \frac{2}{\alpha'^2} \left(-\frac{\alpha'}{2} \frac{1}{(z-w)^2} \right)^2 - \frac{4}{\alpha'^2} \frac{\alpha'}{2} \frac{: \partial X(z) \partial X(w) :}{(z-w)^2} + \dots \end{aligned}$$

The factor of 2 in front of the first term comes from the two ways of performing two contractions; the factor of 4 in the second term comes from the number of ways of performing a single contraction. Continuing,

$$\begin{aligned} T(z) T(w) &= \frac{1/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} - \frac{2}{\alpha'} \frac{\partial^2 X(w) \partial X(w)}{z-w} + \dots \\ &= \frac{1/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots \end{aligned} \quad (4.28)$$

We learn that T is *not* a primary operator in the theory of a single free scalar field. It is an operator of weight $(h, \tilde{h}) = (2, 0)$, but it fails the primary test on account of the $(z-w)^{-4}$ term. In fact, this property of the stress energy tensor a general feature of all CFTs which we now explore in more detail.

4.4 The Central Charge

In any CFT, the most prominent example of an operator which is not primary is the stress-energy tensor itself.

For the free scalar field, we have already seen that T is an operator of weight $(h, \tilde{h}) = (2, 0)$. This remains true in any CFT. The reason for this is simple: $T_{\alpha\beta}$ has dimension $\Delta = 2$ because we obtain the energy by integrating over space. It has spin $s = 2$ because it is a symmetric 2-tensor. But these two pieces of information are equivalent to the statement that T has weight $(2, 0)$. Similarly, \bar{T} has weight $(0, 2)$. This means that the TT OPE takes the form,

$$T(z)T(w) = \dots + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots$$

and similar for $\bar{T}\bar{T}$. What other terms could we have in this expansion? Since each term has dimension $\Delta = 4$, any operators that appear on the right-hand-side must be of the form

$$\frac{\mathcal{O}_n}{(z-w)^n} \quad (4.29)$$

where $\Delta[\mathcal{O}_n] = 4 - n$. But, in a unitary CFT there are no operators with $h, \tilde{h} < 0$. (We will prove this shortly). So the most singular term that we can have is of order $(z-w)^{-4}$. Such a term must be multiplied by a constant. We write,

$$T(z)T(w) = \frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots$$

and, similarly,

$$\bar{T}(\bar{z})\bar{T}(\bar{w}) = \frac{\tilde{c}/2}{(\bar{z}-\bar{w})^4} + \frac{2\bar{T}(\bar{w})}{(\bar{z}-\bar{w})^2} + \frac{\bar{\partial}\bar{T}(\bar{w})}{\bar{z}-\bar{w}} + \dots$$

The constants c and \tilde{c} are called the *central charges*. (Sometimes they are referred to as left-moving and right-moving central charges). They are perhaps the most important numbers characterizing the CFT. We can already get some intuition for the information contained in these two numbers. Looking back at the free scalar field (4.28) we see that it has $c = \tilde{c} = 1$. If we instead considered D non-interacting free scalar fields, we would get $c = \tilde{c} = D$. This gives us a hint: c and \tilde{c} are somehow measuring the number of degrees of freedom in the CFT. This is true in a deep sense! However, be warned: c is not necessarily an integer.

Before moving on, it's worth pausing to explain why we didn't include a $(z-w)^{-3}$ term in the TT OPE. The reason is that the OPE must obey $T(z)T(w) = T(w)T(z)$ because, as explained previously, these operator equations are all taken to hold inside time-ordered correlation functions. So the quick answer is that a $(z-w)^{-3}$ term would

not be invariant under $z \leftrightarrow w$. However, you may wonder how the $(z - w)^{-1}$ term manages to satisfy this property. Let's see how this works:

$$T(w) T(z) = \frac{c/2}{(z-w)^4} + \frac{2T(z)}{(z-w)^2} + \frac{\partial T(z)}{w-z} + \dots$$

Now we can Taylor expand $T(z) = T(w) + (z-w)\partial T(w) + \dots$ and $\partial T(z) = \partial T(w) + \dots$. Using this in the above expression, we find

$$T(w) T(z) = \frac{c/2}{(z-w)^4} + \frac{2T(w) + 2(z-w)\partial T(w)}{(z-w)^2} - \frac{\partial T(w)}{z-w} + \dots = T(z) T(w)$$

This trick of Taylor expanding saves the $(z - w)^{-1}$ term. It wouldn't work for the $(z - w)^{-3}$ term.

The Transformation of Energy

So T is not primary unless $c = 0$. And we will see shortly that all theories have $c > 0$. What does this mean for the transformation of T ?

$$\begin{aligned}\delta T(w) &= -\text{Res} [\epsilon(z) T(z) T(w)] \\ &= -\text{Res} \left[\epsilon(z) \left(\frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots \right) \right]\end{aligned}$$

If $\epsilon(z)$ contains no singular terms, we can expand

$$\epsilon(z) = \epsilon(w) + \epsilon'(w)(z-w) + \frac{1}{2}\epsilon''(w)(z-w)^2 + \frac{1}{6}\epsilon'''(w)(z-w)^3 + \dots$$

from which we find

$$\delta T(w) = -\epsilon(w) \partial T(w) - 2\epsilon'(w) T(w) - \frac{c}{12}\epsilon'''(w) (z-w)^3 \quad (4.30)$$

This is the infinitesimal version. We would like to know what becomes of T under the finite conformal transformation $z \rightarrow \tilde{z}(z)$. The answer turns out to be

$$\tilde{T}(\tilde{z}) = \left(\frac{\partial \tilde{z}}{\partial z} \right)^{-2} \left[T(z) - \frac{c}{12} S(\tilde{z}, z) \right] \quad (4.31)$$

where $S(\tilde{z}, z)$ is known as the *Schwarzian* and is defined by

$$S(\tilde{z}, z) = \left(\frac{\partial^3 \tilde{z}}{\partial z^3} \right) \left(\frac{\partial \tilde{z}}{\partial z} \right)^{-1} - \frac{3}{2} \left(\frac{\partial^2 \tilde{z}}{\partial z^2} \right)^2 \left(\frac{\partial \tilde{z}}{\partial z} \right)^{-2} \quad (4.32)$$

It is simple to check that the Schwarzian has the right infinitesimal form to give (4.30). Its key property is that it preserves the group structure of successive conformal transformations.

4.4.1 c is for Casimir

Note that the extra term in the transformation (4.31) of T does not depend on T itself. In particular, it will be the same evaluated on all states. It only affects the constant term — or zero mode — in the energy. In other words, it is the Casimir energy of the system.

Let's look at an example that will prove to be useful later for the string. Consider the Euclidean cylinder, parameterized by

$$w = \sigma + i\tau \quad , \quad \sigma \in [0, 2\pi)$$

We can make a conformal transformation from the cylinder to the complex plane by

$$z = e^{-iw}$$

The fact that the cylinder and the plane are related by a conformal map means that if we understand a given CFT on the cylinder, then we immediately understand it on the plane. And vice-versa. Notice that constant time slices on the cylinder are mapped to circles of constant radius. The origin, $z = 0$, is the distant past, $\tau \rightarrow -\infty$.

What becomes of T under this transformation? The Schwarzian can be easily calculated to be $S(z, w) = 1/2$. So we find,

$$T_{\text{cylinder}}(w) = -z^2 T_{\text{plane}}(z) + \frac{c}{24} \quad (4.33)$$

Suppose that the ground state energy vanishes when the theory is defined on the plane: $\langle T_{\text{plane}} \rangle = 0$. What happens on the cylinder? We want to look at the Hamiltonian, which is defined by

$$H \equiv \int d\sigma T_{\tau\tau} = - \int d\sigma (T_{ww} + \bar{T}_{\bar{w}\bar{w}})$$

The conformal transformation then tells us that the ground state energy on the cylinder is

$$E = -\frac{2\pi(c + \tilde{c})}{24} \quad (4.34)$$

This is indeed the (negative) Casimir energy on a cylinder. For a free scalar field, we have $c = \tilde{c} = 1$ and the energy density $E/2\pi = -1/12$. This is the same result that we got in Section 2.2.2, but this time with no funny business where we throw out infinities.

An Application: The Lüscher Term

If we're looking at a physical system, the cylinder will have a radius L . In this case, the Casimir energy is given by $E = -2\pi(c + \tilde{c})/24L$. There is an application of this to QCD-like theories. Consider two quarks in a confining theory, separated by a distance L . If the tension of the confining flux tube is T , then the string will be stable as long as $TL \lesssim m$, the mass of the lightest quark. The energy of the stretched string as a function of L is given by

$$E(L) = TL + a - \frac{\pi c}{24L} + \dots$$

Here a is an undetermined constant, while c counts the number of degrees of freedom of the QCD flux tube. (There is no analog of \tilde{c} here because of the reflecting boundary conditions at the end of the string). If the string has no internal degrees of freedom, then $c = 2$ for the two transverse fluctuations. This contribution to the string energy is known as the *Lüscher term*.

4.4.2 The Weyl Anomaly

There is another way in which the central charge affects the stress-energy tensor. Recall that in the classical theory, one of the defining features of a CFT was the vanishing of the trace of the stress tensor,

$$T_{\alpha}^{\alpha} = 0$$

However, things are more subtle in the quantum theory. While $\langle T_{\alpha}^{\alpha} \rangle$ indeed vanishes in flat space, it will no longer be true if we place the theory on a curved background. The purpose of this section is to show that

$$\langle T_{\alpha}^{\alpha} \rangle = -\frac{c}{12}R \tag{4.35}$$

where R is the Ricci scalar of the 2d worldsheet. Before we derive this formula, some quick comments:

- Equation (4.35) holds for any state in the theory — not just the vacuum. This reflects the fact that it comes from regulating short distant divergences in the theory. But, at short distances all finite energy states look basically the same.
- Because $\langle T_{\alpha}^{\alpha} \rangle$ is the same for any state it must be equal to something that depends only on the background metric. This something should be local and must be dimension 2. The only candidate is the Ricci scalar R . For this reason, the formula $\langle T_{\alpha}^{\alpha} \rangle \sim R$ is the most general possibility. The only question is: what is the coefficient. And, in particular, is it non-zero?

- By a suitable choice of coordinates, we can always put any 2d metric in the form $g_{\alpha\beta} = e^{2\omega}\delta_{\alpha\beta}$. In these coordinates, the Ricci scalar is given by

$$R = -2e^{-2\omega}\partial^2\omega \quad (4.36)$$

which depends explicitly on the function ω . Equation (4.35) is then telling us that any conformal theory with $c \neq 0$ has at least one physical observable, $\langle T_\alpha^\alpha \rangle$, which takes different values on backgrounds related by a Weyl transformation ω . This result is referred to as the *Weyl anomaly*, or sometimes as the trace anomaly.

- There is also a Weyl anomaly for conformal field theories in higher dimensions. For example, 4d CFTs are characterized by two numbers, a and c , which appear as coefficients in the Weyl anomaly,

$$\langle T_\mu^\mu \rangle_{4d} = \frac{c}{16\pi^2} C_{\rho\sigma\kappa\lambda} C^{\rho\sigma\kappa\lambda} - \frac{a}{16\pi^2} \tilde{R}_{\rho\sigma\kappa\lambda} \tilde{R}^{\rho\sigma\kappa\lambda}$$

where C is the Weyl tensor and \tilde{R} is the dual of the Riemann tensor.

- Equation (4.35) involves only the left-moving central charge c . You might wonder what's special about the left-moving sector. The answer, of course, is nothing. We also have

$$\langle T_\alpha^\alpha \rangle = -\frac{\tilde{c}}{12} R$$

In flat space, conformal field theories with different c and \tilde{c} are perfectly acceptable. However, if we wish these theories to be consistent in fixed, curved backgrounds, then we require $c = \tilde{c}$. This is an example of a *gravitational anomaly*.

- The fact that Weyl invariance requires $c = 0$ will prove crucial in string theory. We shall return to this in Chapter 5.

We will now prove the Weyl anomaly formula (4.35). Firstly, we need to derive an intermediate formula: the $T_{z\bar{z}} T_{w\bar{w}}$ OPE. Of course, in the classical theory we found that conformal invariance requires $T_{z\bar{z}} = 0$. We will now show that it's a little more subtle in the quantum theory.

Our starting point is the equation for energy conservation,

$$\partial T_{z\bar{z}} = -\bar{\partial} T_{z\bar{z}}$$

Using this, we can express our desired OPE in terms of the familiar TT OPE,

$$\partial_z T_{z\bar{z}}(z, \bar{z}) \partial_w T_{w\bar{w}}(w, \bar{w}) = \bar{\partial}_{\bar{z}} T_{z\bar{z}}(z, \bar{z}) \bar{\partial}_{\bar{w}} T_{w\bar{w}}(w, \bar{w}) = \bar{\partial}_{\bar{z}} \bar{\partial}_{\bar{w}} \left[\frac{c/2}{(z-w)^4} + \dots \right] \quad (4.37)$$

Now you might think that the right-hand-side just vanishes: after all, it is an anti-holomorphic derivative $\bar{\partial}$ of a holomorphic quantity. But we shouldn't be so cavalier because there is a singularity at $z = w$. For example, consider the following equation,

$$\bar{\partial}_{\bar{z}} \partial_z \ln |z - w|^2 = \bar{\partial}_{\bar{z}} \frac{1}{z - w} = 2\pi \delta(z - w, \bar{z} - \bar{w}) \quad (4.38)$$

We proved this statement after equation (4.21). (The factor of 2 difference from (4.21) can be traced to the conventions we defined for complex coordinates in Section 4.0.1). Looking at the intermediate step in (4.38), we again have an anti-holomorphic derivative of a holomorphic function and you might be tempted to say that this also vanishes. But you'd be wrong: subtle things happen because of the singularity and equation (4.38) tells us that the function $1/z$ secretly depends on \bar{z} . (This should really be understood as a statement about distributions, with the delta function integrated against arbitrary test functions). Using this result, we can write

$$\bar{\partial}_{\bar{z}} \bar{\partial}_{\bar{w}} \frac{1}{(z - w)^4} = \frac{1}{6} \bar{\partial}_{\bar{z}} \bar{\partial}_{\bar{w}} \left(\partial_z^2 \partial_w \frac{1}{z - w} \right) = \frac{\pi}{3} \partial_z^2 \partial_w \bar{\partial}_{\bar{w}} \delta(z - w, \bar{z} - \bar{w})$$

Inserting this into the correlation function (4.37) and stripping off the $\partial_z \partial_w$ derivatives on both sides, we end up with what we want,

$$T_{z\bar{z}}(z, \bar{z}) T_{w\bar{w}}(w, \bar{w}) = \frac{c\pi}{6} \partial_z \bar{\partial}_{\bar{w}} \delta(z - w, \bar{z} - \bar{w}) \quad (4.39)$$

So the OPE of $T_{z\bar{z}}$ and $T_{w\bar{w}}$ almost vanishes, but there's some strange singular behaviour going on as $z \rightarrow w$. This is usually referred to as a contact term between operators and, as we have shown, it is needed to ensure the conservation of energy-momentum. We will now see that this contact term is responsible for the Weyl anomaly.

We assume that $\langle T_{\alpha}^{\alpha} \rangle = 0$ in flat space. Our goal is to derive an expression for $\langle T_{\alpha}^{\alpha} \rangle$ close to flat space. Firstly, consider the change of $\langle T_{\alpha}^{\alpha} \rangle$ under a general shift of the metric $\delta g_{\alpha\beta}$. Using the definition of the energy-momentum tensor (4.4), we have

$$\begin{aligned} \delta \langle T_{\alpha}^{\alpha}(\sigma) \rangle &= \delta \int \mathcal{D}\phi e^{-S} T_{\alpha}^{\alpha}(\sigma) \\ &= \frac{1}{4\pi} \int \mathcal{D}\phi e^{-S} \left(T_{\alpha}^{\alpha}(\sigma) \int d^2\sigma' \sqrt{g} \delta g^{\beta\gamma} T_{\beta\gamma}(\sigma') \right) \end{aligned}$$

If we now restrict to a Weyl transformation, the change to a flat metric is $\delta g_{\alpha\beta} = 2\omega \delta_{\alpha\beta}$, so the change in the inverse metric is $\delta g^{\alpha\beta} = -2\omega \delta^{\alpha\beta}$. This gives

$$\delta \langle T_{\alpha}^{\alpha}(\sigma) \rangle = -\frac{1}{2\pi} \int \mathcal{D}\phi e^{-S} \left(T_{\alpha}^{\alpha}(\sigma) \int d^2\sigma' \omega(\sigma') T_{\beta}^{\beta}(\sigma') \right) \quad (4.40)$$

Now we see why the OPE (4.39) determines the Weyl anomaly. We need to change between complex coordinates and Cartesian coordinates, keeping track of factors of 2. We have

$$T_\alpha^\alpha(\sigma) T_\beta^\beta(\sigma') = 16 T_{z\bar{z}}(z, \bar{z}) T_{w\bar{w}}(w, \bar{w})$$

Meanwhile, using the conventions laid down in 4.0.1, we have $8\partial_z \bar{\partial}_{\bar{w}} \delta(z - w, \bar{z} - \bar{w}) = -\partial^2 \delta(\sigma - \sigma')$. This gives us the OPE in Cartesian coordinates

$$T_\alpha^\alpha(\sigma) T_\beta^\beta(\sigma') = -\frac{c\pi}{3} \partial^2 \delta(\sigma - \sigma')$$

We now plug this into (4.40) and integrate by parts to move the two derivatives onto the conformal factor ω . We're left with,

$$\delta \langle T_\alpha^\alpha \rangle = \frac{c}{6} \partial^2 \omega \Rightarrow \langle T_\alpha^\alpha \rangle = -\frac{c}{12} R$$

where, to get to the final step, we've used (4.36) and, since we're working infinitesimally, we can replace $e^{-2\omega} \approx 1$. This completes the proof of the Weyl anomaly, at least for spaces infinitesimally close to flat space. The fact that R remains on the right-hand-side for general 2d surfaces follows simply from the comments after equation (4.35), most pertinently the need for the expression to be reparameterization invariant.

4.4.3 c is for Cardy

The Casimir effect and the Weyl anomaly have a similar smell. In both, the central charge provides an extra contribution to the energy. We now demonstrate a different avatar of the central charge: it tells us the density of high energy states.

We will study conformal field theory on a Euclidean torus. We'll keep our normalization $\sigma \in [0, 2\pi)$, but now we also take τ to be periodic, lying in the range

$$\tau \in [0, \beta)$$

The partition function of a theory with periodic Euclidean time has a very natural interpretation: it is related to the free energy of the theory at temperature $T = 1/\beta$.

$$Z[\beta] = \text{Tr } e^{-\beta H} = e^{-\beta F} \tag{4.41}$$

At very low temperatures, $\beta \rightarrow \infty$, the free energy is dominated by the lowest energy state. All other states are exponentially suppressed. But we saw in 4.4.1 that the vacuum state on the cylinder has Casimir energy $H = -c/12$. In the limit of low temperature, the partition function is therefore approximated by

$$Z \rightarrow e^{c\beta/12} \quad \text{as } \beta \rightarrow \infty \tag{4.42}$$

Now comes the trick. In Euclidean space, both directions of the torus are on equal footing. We're perfectly at liberty to decide that σ is “time” and τ is “space”. This can't change the value of the partition function. So let's make the swap. To compare to our original partition function, we want the spatial direction to have range $[0, 2\pi]$. Happily, due to the conformal nature of our theory, we arrange this through the scaling

$$\tau \rightarrow \frac{2\pi}{\beta} \tau \quad , \quad \sigma \rightarrow \frac{2\pi}{\beta} \sigma$$

Now we're back where we started, but with the temporal direction taking values in $\sigma \in [0, 4\pi^2/\beta]$. This tells us that the high-temperature and low-temperature partition functions are related,

$$Z[4\pi^2/\beta] = Z[\beta]$$

This is called modular invariance. We'll come across it again in Section 6.4. Writing $\beta' = 4\pi^2/\beta$, this tells us the very high temperature behaviour of the partition function

$$Z[\beta'] \rightarrow e^{c\pi^2/3\beta'} \quad \text{as } \beta' \rightarrow 0$$

But the very high temperature limit of the partition function is sampling all states in the theory. On entropic grounds, this sampling is dominated by the high energy states. So this computation is telling us how many high energy states there are.

To see this more explicitly, let's do some elementary manipulations in statistical mechanics. Any system has a density of states $\rho(E) = e^{S(E)}$, where $S(E)$ is the entropy. The free energy is given by

$$e^{-\beta F} = \int dE \rho(E) e^{-\beta E} = \int dE e^{S(E)-\beta E}$$

In two dimensions, all systems have an entropy which scales at large energy as

$$S(E) \rightarrow N\sqrt{E} \tag{4.43}$$

The coefficient N counts the number of degrees of freedom. The fact that $S \sim \sqrt{E}$ is equivalent to the fact that $F \sim T^2$, as befits an energy density in a theory with one

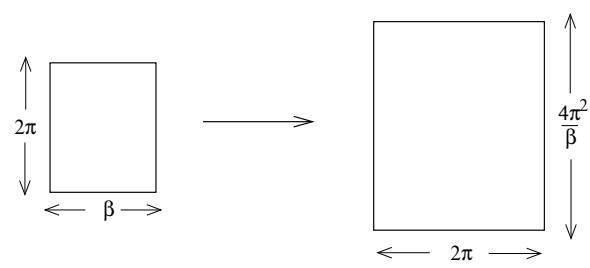


Figure 23:

spatial dimension. To see this, we need only approximate the integral by the saddle point $S'(E_\star) = \beta$. From (4.43), this gives us the free energy

$$F \sim N^2 T^2$$

We can now make the statement about the central charge more explicit. In a conformal field theory, the entropy of high energy states is given by

$$S(E) \sim \sqrt{cE}$$

This is *Cardy's formula*. A more careful analysis of the coefficients shows that the high energy density of states scales as

$$S(E) \rightarrow 2\pi \sqrt{\frac{c}{6} \left(ER - \frac{c}{24} \right)} \quad (4.44)$$

where the offset is the Casimir energy (4.34) that we derived previously. This is the contribution from left-movers. There is a similar contribution from right-movers, depending on \tilde{c} .

4.4.4 c has a Theorem

The connection between the central charge and the degrees of freedom in a theory is given further weight by a result of Zamalodchikov, known as the *c-theorem*. The idea of the c-theorem is to stand back and look at the space of all theories and the renormalization group (RG) flows between them.

Conformal field theories are special. They are the fixed points of the renormalization group, looking the same at all length scales. One can consider perturbing a conformal field theory by adding an extra term to the action,

$$S \rightarrow S + \alpha \int d^2\sigma \mathcal{O}(\sigma)$$

Here \mathcal{O} is a local operator of the theory, while α is some coefficient. These perturbations fall into three classes, depending on the dimension Δ of \mathcal{O} .

- $\Delta < 2$: In this case, α has positive dimension: $[\alpha] = 2 - \delta$. Such deformations are called *relevant* because they are important in the infra-red. RG flow takes us away from our original CFT. We only stop flowing when we hit a new CFT (which could be trivial with $c = 0$).
- $\Delta = 2$: The constant α is dimensionless. Such deformations are called *marginal*. The deformed theory defines a new CFT.

- $\Delta > 2$: The constant α has negative dimension. These deformations are irrelevant. The infra-red physics is still described by the original CFT. But the ultra-violet physics is altered.

We expect information is lost as we flow from an ultra-violet theory to the infra-red. The c-theorem makes this intuition precise. The theorem exhibits a function c on the space of all theories which monotonically decreases along RG flows. At the fixed points, c coincides with the central charge of the CFT.

A Thermodynamic Proof of the c-Theorem

There are a number of different proofs of the c-theorem. Here we give one that is particularly physical. The basic idea is to heat up the system to a finite temperature T and compute the speed of sound. The c-theorem follows from the requirement that the speed of sound does not exceed the speed of light (which, in our conventions, is simply 1). I should warn you that the style of argument in this section is somewhat different from the rest of these lectures. But, if nothing else, it reminds you that just because you're learning string theory, you shouldn't neglect basic physics!

Let's first start with a CFT. For simplicity, we assume that $c = \tilde{c}$. Then, from (4.44), we have the asymptotic behaviour

$$S(E) \rightarrow 4\pi \sqrt{\frac{cER}{6}}$$

where we have dropped the $c/24$ offset, and the overall coefficient is 4π rather than 2π because we are including both left- and right-moving sectors. To compare with familiar, thermodynamic formulae we write this in terms of the spatial volume $V = 2\pi R$, so

$$S(E) \rightarrow 4\pi \sqrt{\frac{\pi cEV}{3}}$$

Now, the temperature is defined to be

$$\frac{1}{T} = \frac{\partial S}{\partial E} = 2\pi \sqrt{\frac{\pi cV}{3E}} \quad \Rightarrow \quad \sqrt{E} = 2\pi T \sqrt{\frac{\pi cV}{3}}$$

From this, we can compute the entropy of a CFT as a function of temperature, rather than as a function of energy

$$S(T) = \frac{8\pi^3 c V T}{3} \quad \Rightarrow \quad s(T) = \frac{8\pi^3 c}{3} T \tag{4.45}$$

where $s = S/V$ is the entropy density.

Now we'll consider a more general situation. We'll flow from some CFT in the UV with central charge c_{UV} to another CFT in the IR with central charge c_{IR} . It may be that the final theory is gapped – meaning that everything is massless – in which case $c_{IR} = 0$. Our goal is to prove that, regardless of the flow, we always have $c_{UV} \geq c_{IR}$ (with equality if there is no flow at all). To achieve this, we need to play around with some thermodynamic identities. In particular, we need to following result

Claim:

$$s = \left. \frac{\partial P}{\partial T} \right|_V \quad (4.46)$$

with P the pressure.

Proof: Given the energy $E = E(S, V)$, the first law of thermodynamics tells us

$$dE = TdS - PdV$$

The free energy is then defined as $F(T, V) = E - TS$ and obeys

$$dF = -SdT - PdV \quad (4.47)$$

But the free energy is extensive and this means that it must, in fact, be proportional to V since this is the only extensive quantity that it can depend on. So

$$F(T, V) = -P(T)V$$

From this we learn that

$$dF = -\frac{\partial P}{\partial T}VdT - PdV$$

Comparing to (4.47) gives us the claimed result (4.46). □

Finally, we recall that the speed of sound in a system is given by (see, for example, the lectures on [Fluid Mechanics](#))

$$c_s^2 = \frac{dP}{d\epsilon}$$

where $\epsilon = E/V$ is the energy density. At fixed volume, we have

$$dE = TdS \quad \Rightarrow \quad d\epsilon = Tds$$

All of which means that we can express the speed of sound as

$$c_s^2 = \frac{1}{T} \frac{dP}{ds} = \frac{1}{T} \frac{dP}{dT} \frac{dT}{ds} = \frac{s}{T} \frac{dT}{ds} = \frac{d \log T}{d \log s}$$

This is the key result that we need. Now we define a thermal *c-function*

$$\chi = \frac{s}{T}$$

As we've seen in (4.45), when we have a CFT the function χ is proportional to the central charge: $\chi = 8\pi^3 c/3$. If we flow from a CFT in the UV, with central charge c_{UV} , to a different CFT in the IR with central charge c_{IR} , then χ will interpolate between these two values (multiplied by $8\pi^3/3$) as we vary the temperature. To prove the c-theorem, we need to show that as we decrease the temperature, and so excite lower energy degrees of freedom, the function χ necessarily decreases. We do this by relating χ to the speed of sound,

$$\frac{1}{c_s^2} = \frac{d \log s}{d \log T} = \frac{d \log(\chi T)}{d \log T} = 1 + \frac{d \log \chi}{d \log T}$$

By causality, we must have $c_s^2 \leq 1$ (with equality when we have a CFT) and so

$$\frac{d \log \chi}{d \log T} \geq 0 \quad \Rightarrow \quad \frac{d\chi}{dT} \geq 0$$

But this is what we wanted. We learn that we necessarily have $c_{UV} \geq c_{IR}$. This is the c-theorem.

4.5 The Virasoro Algebra

So far our discussion has been limited to the operators of the CFT. We haven't said anything about states. We now remedy this. We start by taking a closer look at the map between the cylinder and the plane.

4.5.1 Radial Quantization

To discuss states in a quantum field theory we need to think about where they live and how they evolve. For example, consider a two dimensional quantum field theory defined on the plane. Traditionally, when quantizing this theory, we parameterize the plane by Cartesian coordinates (t, x) which we'll call "time" and "space". The states live on spatial slices. The Hamiltonian generates time translations and hence governs the evolution of states.

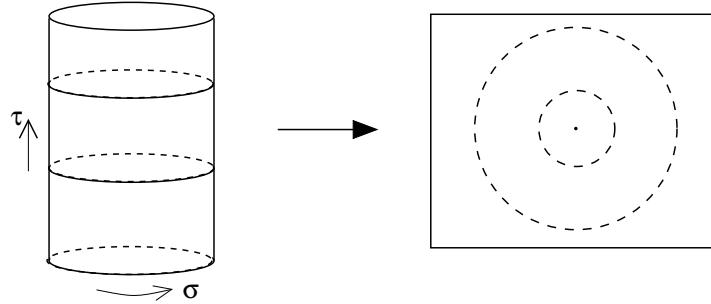


Figure 25: The map from the cylinder to the plane.

However, the map between the cylinder and the plane suggests a different way to quantize a CFT on the plane. The complex coordinate on the cylinder is taken to be ω , while the coordinate on the plane is z . They are related by,

$$\omega = \sigma + i\tau \quad , \quad z = e^{-i\omega}$$

On the cylinder, states live on spatial slices of constant σ and evolve by the Hamiltonian,

$$H = \partial_\tau$$

After the map to the plane, the Hamiltonian becomes the dilatation operator

$$D = z\partial_z + \bar{z}\bar{\partial}_z$$

If we want the states on the plane to remember their cylindrical roots, they should live on circles of constant radius. Their evolution is governed by the dilatation operator D . This approach to a theory is known as *radial quantization*.

Usually in a quantum field theory, we're interested in time-ordered correlation functions. Time ordering on the cylinder becomes radial ordering on the plane. Operators in correlation functions are ordered so that those inserted at larger radial distance are moved to the left.

Virasoro Generators

Let's look at what becomes of the stress tensor $T(z)$ evaluated on the plane. On the cylinder, we would decompose T in a Fourier expansion.

$$T_{\text{cylinder}}(w) = - \sum_{m=-\infty}^{\infty} L_m e^{imw} + \frac{c}{24}$$

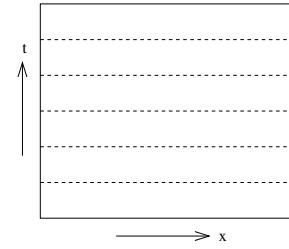


Figure 24:

After the transformation (4.33) to the plane, this becomes the Laurent expansion

$$T(z) = \sum_{m=-\infty}^{\infty} \frac{L_m}{z^{m+2}}$$

As always, a similar statement holds for the right-moving sector

$$\bar{T}(\bar{z}) = \sum_{m=-\infty}^{\infty} \frac{\tilde{L}_m}{\bar{z}^{m+2}}$$

We can invert these expressions to get L_m in terms of $T(z)$. We need to take a suitable contour integral

$$L_n = \frac{1}{2\pi i} \oint dz z^{n+1} T(z) \quad , \quad \tilde{L}_n = \frac{1}{2\pi i} \oint d\bar{z} \bar{z}^{n+1} \bar{T}(\bar{z}) \quad (4.48)$$

where, if we just want L_n or \tilde{L}_n , we must make sure that there are no other insertions inside the contour.

In radial quantization, L_n is the conserved charge associated to the conformal transformation $\delta z = z^{n+1}$. To see this, recall that the corresponding Noether current, given in (4.7), is $J(z) = z^{n+1}T(z)$. Moreover, the contour integral $\oint dz$ maps to the integral around spatial slices on the cylinder. This tells us that L_n is the conserved charge where “conserved” means that it is constant under time evolution on the cylinder, or under radial evolution on the plane. Similarly, \tilde{L}_n is the conserved charge associated to the conformal transformation $\delta\bar{z} = \bar{z}^{n+1}$.

When we go to the quantum theory, conserved charges become generators for the transformation. Thus the operators L_n and \tilde{L}_n generate the conformal transformations $\delta z = z^{n+1}$ and $\delta\bar{z} = \bar{z}^{n+1}$. They are known as the *Virasoro* generators. In particular, our two favorite conformal transformations are

- L_{-1} and \tilde{L}_{-1} generate translations in the plane.
- L_0 and \tilde{L}_0 generate scaling and rotations.

The Hamiltonian of the system — which measures the energy of states on the cylinder — is mapped into the dilatation operator on the plane. When acting on states of the theory, this operator is represented as

$$D = L_0 + \tilde{L}_0$$

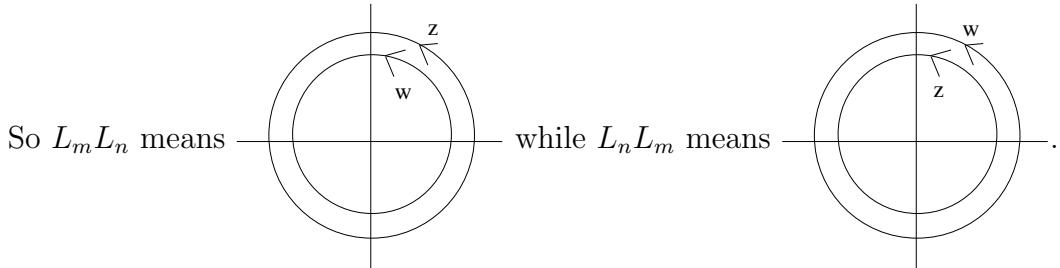
4.5.2 The Virasoro Algebra

If we have some number of conserved charges, the first thing that we should do is compute their algebra. Representations of this algebra then classify the states of the theory. (For example, think angular momentum in the hydrogen atom). For conformal symmetry, we want to determine the algebra obeyed by the L_n generators. It's a nice fact that the commutation relations are actually encoded TT OPE. Let's see how this works.

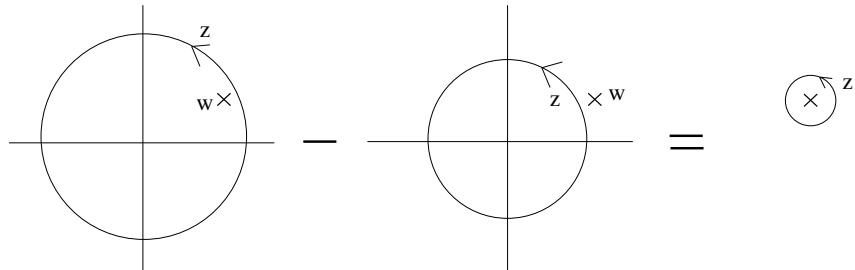
We want to compute $[L_m, L_n]$. Let's write L_m as a contour integral over $\oint dz$ and L_n as a contour integral over $\oint dw$. (Note: both z and w denote coordinates on the complex plane now). The commutator is

$$[L_m, L_n] = \left(\oint \frac{dz}{2\pi i} \oint \frac{dw}{2\pi i} - \oint \frac{dw}{2\pi i} \oint \frac{dz}{2\pi i} \right) z^{m+1} w^{n+1} T(z) T(w)$$

What does this actually mean?! We need to remember that all operator equations are to be viewed as living inside time-ordered correlation functions. Except, now we're working on the z -plane, this statement has transmuted into radially ordered correlation functions: outies to the left, innies to the right.



The trick to computing the commutator is to first fix w and do the $\oint dz$ integrations. The resulting contour is,



In other words, we do the z -integration around a fixed point w , to get

$$[L_m, L_n] = \oint \frac{dw}{2\pi i} \oint_w \frac{dz}{2\pi i} z^{m+1} w^{n+1} T(z) T(w)$$

$$= \oint \frac{dw}{2\pi i} \text{Res} \left[z^{m+1} w^{n+1} \left(\frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots \right) \right]$$

To compute the residue at $z = w$, we first need to Taylor expand z^{m+1} about the point w ,

$$\begin{aligned} z^{m+1} &= w^{m+1} + (m+1)w^m(z-w) + \frac{1}{2}m(m+1)w^{m-1}(z-w)^2 \\ &\quad + \frac{1}{6}m(m^2-1)w^{m-2}(z-w)^3 + \dots \end{aligned}$$

The residue then picks up a contribution from each of the three terms,

$$[L_m, L_n] = \oint \frac{dw}{2\pi i} w^{n+1} \left[w^{m+1} \partial T(w) + 2(m+1)w^m T(w) + \frac{c}{12}m(m^2-1)w^{m-2} \right]$$

To proceed, it is simplest to integrate the first term by parts. Then we do the w -integral. But for both the first two terms, the resulting integral is of the form (4.48) and gives us L_{m+n} . For the third term, we pick up the pole. The end result is

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}$$

This is the *Virasoro algebra*. It's quite famous. The \tilde{L}_n 's satisfy exactly the same algebra, but with c replaced by \tilde{c} . Of course, $[L_n, \tilde{L}_m] = 0$. The appearance of c as an extra term in the Virasoro algebra is the reason it is called the “central charge”. In general, a central charge is an extra term in an algebra that commutes with everything else.

Conformal = Diffeo + Weyl

We can build some intuition for the Virasoro algebra. We know that the L_n 's generate conformal transformations $\delta z = z^{n+1}$. Let's consider something closely related: a coordinate transformation $\delta z = z^{n+1}$. These are generated by the vector fields

$$l_n = z^{n+1} \partial_z \tag{4.49}$$

But it's a simple matter to compute their commutation relations:

$$[l_n, l_m] = (m-n)l_{m+n}$$

So this is giving us the first part of the Virasoro algebra. But what about the central term? The key point to remember is that, as we stressed at the beginning of this chapter, a conformal transformation is not just a reparameterization of the coordinates: it is a reparameterization, followed by a compensating Weyl rescaling. The central term in the Virasoro algebra is due to the Weyl rescaling.

4.5.3 Representations of the Virasoro Algebra

With the algebra of conserved charges at hand, we can now start to see how the conformal symmetry classifies the states into representations.

Suppose that we have some state $|\psi\rangle$ that is an eigenstate of L_0 and \tilde{L}_0 .

$$L_0 |\psi\rangle = h |\psi\rangle \quad , \quad \tilde{L}_0 |\psi\rangle = \tilde{h} |\psi\rangle$$

Back on the cylinder, this corresponds to some state with energy

$$\frac{E}{2\pi} = h + \tilde{h} - \frac{c + \tilde{c}}{24}$$

For this reason, we'll refer to the eigenvalues h and \tilde{h} as the energy of the state. By acting with the L_n operators, we can get further states with eigenvalues

$$L_0 L_n |\psi\rangle = (L_n L_0 - n L_n) |\psi\rangle = (h - n) L_n |\psi\rangle$$

This tells us that L_n are raising and lowering operators depending on the sign of n . When $n > 0$, L_n lowers the energy of the state and L_{-n} raises the energy of the state. If the spectrum is to be bounded below, there must be some states which are annihilated by all L_n and \tilde{L}_n for $n > 0$. Such states are called *primary*. They obey

$$L_n |\psi\rangle = \tilde{L}_n |\psi\rangle = 0 \quad \text{for all } n > 0$$

In the language of representation theory, they are also called highest weight states. They are the states of lowest energy.

Representations of the Virasoro algebra can now be built by acting on the primary states with raising operators L_{-n} with $n > 0$. Obviously this results in an infinite tower of states. All states obtained in this way are called *descendants*. From an initial primary state $|\psi\rangle$, the tower fans out...

$$\begin{aligned} & |\psi\rangle \\ & L_{-1} |\psi\rangle \\ & L_{-1}^2 |\psi\rangle , L_{-2} |\psi\rangle \\ & L_{-1}^3 |\psi\rangle , L_{-1} L_{-2} |\psi\rangle , L_{-3} |\psi\rangle \end{aligned}$$

The whole set of states is called a *Verma module*. They are the irreducible representations of the Virasoro algebra. This means that if we know the spectrum of primary states, then we know the spectrum of the whole theory.

Some comments:

- The vacuum state $|0\rangle$ has $h = 0$. This state obeys

$$L_n |0\rangle = 0 \quad \text{for all } n \geq -1 \quad (4.50)$$

Note that this state preserves the maximum number of symmetries: like all primary states, it is annihilated by L_n with $n > 0$, but it is also annihilated by L_0 and L_{-1} . This fits with our intuition that the vacuum state should be invariant under as many symmetries as possible. You might think that we could go further and require that the vacuum state obeys $L_n |0\rangle = 0$ for all n . But that isn't consistent with the central charge term in Virasoro algebra. The requirements (4.50) are the best we can do.

- This discussion should be ringing bells. We saw something very similar in the covariant quantization of the string, where we imposed conditions (2.6) as constraints. We will see the connection between the primary states and the spectrum of the string in Section 5.
- There's a subtlety that you should be aware of: the states in the Verma module are not necessarily all independent. It could be that some linear combination of the states vanishes. This linear combination is known as a null state. The existence of null states depends on the values of h and c . For example, suppose that we are in a theory in which the central charge is $c = 2h(5 - 8h)/(2h + 1)$, where h is the energy of a primary state $|\psi\rangle$. Then it is simple to check that the following combination has vanishing norm:

$$L_{-2} |\psi\rangle - \frac{3}{2(2h + 1)} L_{-1}^2 |\psi\rangle \quad (4.51)$$

- There is a close relationship between the primary states and the primary operators defined in Section 4.2.3. In fact, the energies h and \tilde{h} of primary states will turn out to be exactly the weights of primary operators in the theory. This connection will be described in Section 4.6.

4.5.4 Consequences of Unitarity

There is one physical requirement that a theory must obey which we have so far neglected to mention: *unitarity*. This is the statement that probabilities are conserved when we are in Minkowski signature spacetime. Unitarity follows immediately if we have a Hermitian Hamiltonian which governs time evolution. But so far our discussion has been somewhat algebraic and we've not enforced this condition. Let's do so now.

We retrace our footsteps back to the Euclidean cylinder and then back again to the Minkowski cylinder where we can ask questions about time evolution. Here the Hamiltonian density takes the form

$$\mathcal{H} = T_{ww} + T_{\bar{w}\bar{w}} = \sum_n L_n e^{-in\sigma^+} + \tilde{L}_n e^{-in\sigma^-}$$

So for the Hamiltonian to be Hermitian, we require

$$L_n = L_{-n}^\dagger$$

This requirement imposes some strong constraints on the structure of CFTs. Here we look at a couple of trivial, but important, constraints that arise due to unitarity and the requirement that the physical Hilbert space does not contain negative norm states.

- $h \geq 0$: This fact follows from looking at the norm,

$$|L_{-1}|\psi\rangle|^2 = \langle\psi|L_{+1}L_{-1}|\psi\rangle = \langle\psi|[L_{+1}, L_{-1}]|\psi\rangle = 2h\langle\psi|\psi\rangle \geq 0$$

The only state with $h = 0$ is the vacuum state $|0\rangle$.

- $c > 0$: To see this, we can look at

$$|L_{-n}|0\rangle|^2 = \langle 0|[L_n, L_{-n}]|0\rangle = \frac{c}{12}n(n^2 - 1) \geq 0 \quad (4.52)$$

So $c \geq 0$. If $c = 0$, the only state in the vacuum module is the vacuum itself. It turns out that, in fact, the only state in the whole theory is the vacuum itself. Any non-trivial CFT has $c > 0$.

There are many more requirements of this kind that constrain the theory. In fact, it turns out that for CFTs with $c < 1$ these requirements are enough to classify and solve all theories.

4.6 The State-Operator Map

In this section we describe one particularly important aspect of conformal field theories: a map between states and local operators.

Firstly, let's get some perspective. In a typical quantum field theory, the states and local operators are very different objects. While local operators live at a point in spacetime, the states live over an entire spatial slice. This is most clear if we write down a Schrödinger-style wavefunction. In field theory, this object is actually a wavefunctional, $\Psi[\phi(\sigma)]$, describing the probability for every field configuration $\phi(\sigma)$ at each point σ in space (but at a fixed time).

Given that states and local operators are such very different beasts, it's a little surprising that in a CFT there is an isomorphism between them: it's called the state-operator map. The key point is that the distant past in the cylinder gets mapped to a single point $z = 0$ in the complex plane. So specifying a state on the cylinder in the far past is equivalent to specifying a local disturbance at the origin.

To make this precise, we need to recall how to write down wavefunctions using path integrals. Different states are computed by putting different boundary conditions on the functional integral. Let's start by returning to quantum mechanics and reviewing a few simple facts. The propagator for a particle to move from position x_i at time τ_i to position x_f at time τ_f is given by

$$G(x_f, x_i) = \int_{x(\tau_i)=x_i}^{x(\tau_f)=x_f} \mathcal{D}x e^{iS}$$

This means that if our system starts off in some state described by the wavefunction $\psi_i(x_i)$ at time τ_i then (ignoring the overall normalization) it evolves to the state

$$\psi_f(x_f, \tau_f) = \int dx_i G(x_f, x_i) \psi_i(x_i, \tau_i)$$

There are two lessons to take from this. Firstly, to determine the value of the wavefunction at a given point x_f , we evaluate the path integral restricting to paths which satisfy $x(\tau_f) = x_f$. Secondly, the initial state $\psi_i(x_i)$ acts as a weighting factor for the integral over initial boundary conditions.

Let's now write down the same formula in a field theory, where we're dealing with wavefunctionals. We'll work with the Euclidean path integral on the cylinder. If we start with some state $\Psi_i[\phi_i(\sigma)]$ at time τ_i , then it will evolve to the state

$$\Psi_f[\phi_f(\sigma), \tau_f] = \int \mathcal{D}\phi_i \int_{\phi(\tau_i)=\phi_i}^{\phi(\tau_f)=\phi_f} \mathcal{D}\phi e^{-S[\phi]} \Psi_i[\phi_i(\sigma), \tau_i]$$

How do we write a similar expression for states after the map to the complex plane? Now the states are defined on circles of constant radius, say $|z| = r$, and evolution is governed by the dilatation operator. Suppose the initial state is defined at $|z| = r_i$. In the path integral, we integrate over all fields with fixed boundary conditions $\phi(r_i) = \phi_i$ and $\phi(r_f) = \phi_f$ on the two edges of the annulus shown in the figure,

$$\Psi_f[\phi_f(\sigma), r_f] = \int \mathcal{D}\phi_i \int_{\phi(r_i)=\phi_i}^{\phi(r_f)=\phi_f} \mathcal{D}\phi e^{-S[\phi]} \Psi_i[\phi_i(\sigma), r_i]$$

This is the traditional way to define a state in field theory, albeit with a slight twist because we're working in radial quantization. We see that the effect of the initial state is to change the weighting of the path integral over the inner ring at $|z| = r_i$.

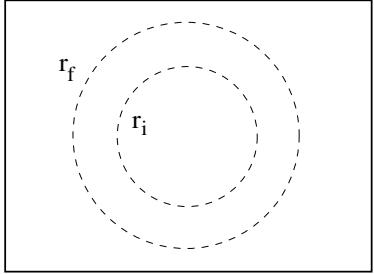


Figure 26:

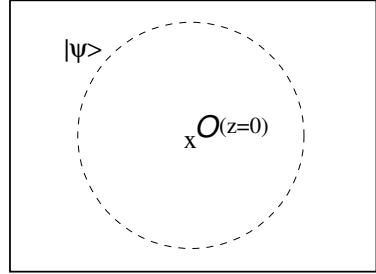


Figure 27:

Let's now see what happens as we take the initial state back to the far past and, ultimately, to $z = 0$? We must now integrate over the whole disc $|z| \leq r_f$, rather than the annulus. The only effect of the initial state is now to change the weighting of the path integral at the point $z = 0$. But that's exactly what we mean by a local operator inserted at that point. This means that each local operator $\mathcal{O}(z = 0)$ defines a different state in the theory,

$$\Psi[\phi_f; r] = \int^{\phi(r)=\phi_f} \mathcal{D}\phi e^{-S[\phi]} \mathcal{O}(z = 0)$$

We're now integrating over all field configurations within the disc, including all possible values of the field at $z = 0$, which is analogous to integrating over the boundary conditions $\int \mathcal{D}\phi_i$ on the inner circle.

- The state-operator map is only true in conformal field theories where we can map the cylinder to the plane. It also holds in conformal field theories in higher dimensions (where $\mathbf{R} \times \mathbf{S}^{D-1}$ can be mapped to the plane \mathbf{R}^D). In non-conformal field theories, a typical local operator creates many different states.
- The state-operator map does not say that the number of states in the theory is equal to the number of operators: this is never true. It does say that the states are in one-to-one correspondence with the *local* operators.
- You might think that you've seen something like this before. In the canonical quantization of free fields, we create states in a Fock space by acting with creation operators. That's *not* what's going on here! The creation operators are just about as far from local operators as you can get. They are the Fourier transforms of local operators.
- There's a special state that we can create this way: the vacuum. This arises by inserting the identity operator $\mathbf{1}$ into the path integral. Back in the cylinder

picture, this just means that we propagate the state back to time $\tau = -\infty$ which is a standard trick used in the Euclidean path integral to project out all but the ground state. For this reason the vacuum is sometimes referred to, in operator notation, as $|1\rangle$.

4.6.1 Some Simple Consequences

Let's use the state-operator map to wrap up a few loose ends that have arisen in our study of conformal field theory.

Firstly, we've defined two objects that we've called “primary”: states and operators. The state-operator map relates the two. Consider the state $|\mathcal{O}\rangle$, built from inserting a primary operator \mathcal{O} into the path integral at $z = 0$. We can look at,

$$\begin{aligned} L_n |\mathcal{O}\rangle &= \oint \frac{dz}{2\pi i} z^{n+1} T(z) \mathcal{O}(z=0) \\ &= \oint \frac{dz}{2\pi i} z^{n+1} \left(\frac{h\mathcal{O}}{z^2} + \frac{\partial\mathcal{O}}{z} + \dots \right) \end{aligned} \quad (4.53)$$

You may wonder what became of the path integral $\int \mathcal{D}\phi e^{-S[\phi]}$ in this expression. The answer is that it's still implicitly there. Remember that operator expressions such as (4.48) are always taken to hold inside correlation functions. But putting an operator in the correlation function is the same thing as putting it in the path integral, weighted with $e^{-S[\phi]}$.

From (4.53) we can see the effect of various generators on states

- $L_{-1} |\mathcal{O}\rangle = |\partial\mathcal{O}\rangle$: In fact, this is true for all operators, not just primary ones. It is expected since L_{-1} is the translation generator.
- $L_0 |\mathcal{O}\rangle = h |\mathcal{O}\rangle$: This is true of any operator with well defined transformation under scaling.
- $L_n |\mathcal{O}\rangle = 0$ for all $n > 0$. This is true only of primary operators \mathcal{O} . Moreover, it is our requirement for $|\mathcal{O}\rangle$ to be a primary state.

This has an important consequence. We stated earlier that one of the most important things to compute in a CFT is the spectrum of weights of primary operators. This seems like a slightly obscure thing to do. But now we see that it has a much more direct, physical meaning. It is the spectrum of energy and angular momentum of states of the theory defined on the cylinder.

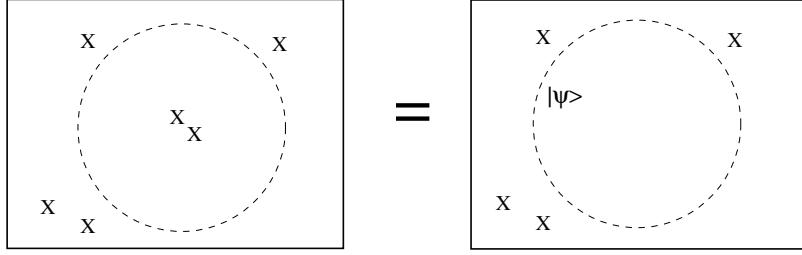


Figure 28:

Another loose end: when defining operators which carry specific weight, we made the statement that we could always work in a basis of operators which have specified eigenvalues under D and L . This follows immediately from the statement that we can always find a basis of eigenstates of H and L on the cylinder.

Finally, we can use this idea of the state-operator map to understand why the OPE works so well in conformal field theories. Suppose that we're interested in some correlation function, with operator insertions as shown in the figure. The statement of the OPE is that we can replace the two inner operators by a sum of operators at $z = 0$, *independent* of what's going on outside of the dotted line. As an operator statement, that sounds rather surprising. But this follows by computing the path integral up to the dotted line, by which point the only effect of the two operators is to determine what state we have. This provides us a way of understanding why the OPE is exact in CFTs, with a radius of convergence equal to the next-nearest insertion.

4.6.2 Our Favourite Example: The Free Scalar Field

Let's illustrate the state-operator map by returning yet again to the free scalar field. On a Euclidean cylinder, we have the mode expansion

$$X(w, \bar{w}) = x + \alpha' p \tau + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} (\alpha_n e^{inw} + \tilde{\alpha}_n e^{in\bar{w}})$$

where we retain the requirement of reality in Minkowski space, which gave us $\alpha_n^* = \alpha_{-n}$ and $\tilde{\alpha}_n^* = \tilde{\alpha}_{-n}$. We saw in Section 4.3 that X does not have good conformal properties. Before transforming to the $z = e^{-iw}$ plane, we should work with the primary field on the cylinder,

$$\partial_w X(w, \bar{w}) = -\sqrt{\frac{\alpha'}{2}} \sum_n \alpha_n e^{inw} \quad \text{with } \alpha_0 \equiv i \sqrt{\frac{\alpha'}{2}} p$$

Since ∂X is a primary field of weight $h = 1$, its transformation to the plane is given by (4.18) and reads

$$\partial_z X(z) = \left(\frac{\partial z}{\partial w} \right)^{-1} \partial_w X(w) = -i \sqrt{\frac{\alpha'}{2}} \sum_n \frac{\alpha_n}{z^{n+1}}$$

and similar for $\bar{\partial}X$. Inverting this gives an equation for α_n as a contour integral,

$$\alpha_n = i \sqrt{\frac{2}{\alpha'}} \oint \frac{dz}{2\pi i} z^n \partial X(z) \quad (4.54)$$

Just as the TT OPE allowed us to determine the $[L_m, L_n]$ commutation relations in the previous section, so the $\partial X \partial X$ OPE contains the information about the $[\alpha_m, \alpha_n]$ commutation relations. The calculation is straightforward,

$$\begin{aligned} [\alpha_m, \alpha_n] &= -\frac{2}{\alpha'} \left(\oint \frac{dz}{2\pi i} \oint \frac{dw}{2\pi i} - \oint \frac{dw}{2\pi i} \oint \frac{dz}{2\pi i} \right) z^m w^n \partial X(z) \partial X(w) \\ &= -\frac{2}{\alpha'} \oint \frac{dw}{2\pi i} \text{Res}_{z=w} \left[z^m w^n \left(\frac{-\alpha'/2}{(z-w)^2} + \dots \right) \right] \\ &= m \oint \frac{dw}{2\pi i} w^{m+n-1} = m \delta_{m+n,0} \end{aligned}$$

where, in going from the second to third line, we have Taylor expanded z around w . Hearteningly, the final result agrees with the commutation relation (2.2) that we derived in string theory using canonical quantization.

The State-Operator Map for the Free Scalar Field

Let's now look at the map between states and local operators. We know from canonical quantization that the Fock space is defined by acting with creation operators α_{-m} with $m > 0$ on the vacuum $|0\rangle$. The vacuum state itself obeys $\alpha_m |0\rangle = 0$ for $m > 0$. Finally, there is also the zero mode $\alpha_0 \sim p$ which provides all states with another quantum number. A general state is given by

$$\prod_{m=1}^{\infty} \alpha_{-m}^{k_m} |0; p\rangle$$

Let's try and recover these states by inserting operators into the path integral. Our first task is to check whether the vacuum state is indeed equivalent to the insertion of the identity operator. In other words, is the ground state wavefunctional of the theory on the circle $|z| = r$ really given by

$$\Psi_0[X_f] = \int^{X_f(r)} \mathcal{D}X e^{-S[X]} \quad ? \quad (4.55)$$

We want to check that this satisfies the definition of the vacuum state, namely $\alpha_m|0\rangle = 0$ for $m > 0$. How do we act on the wavefunctional with an operator? We should still integrate over all field configurations $X(z, \bar{z})$, subject to the boundary conditions at $X(|z| = r) = X_f$. But now we should insert the contour integral (4.54) at some $|w| < r$ (because, after all, the state is only going to vanish after we've hit it with α_m , not before!). So we look at

$$\alpha_m \Psi_0[X_f] = \int^{X_f} \mathcal{D}X e^{-S[X]} \oint \frac{dw}{2\pi i} w^m \partial X(w)$$

The path integral is weighted by the action (4.19) for a free scalar field. If a given configuration diverges somewhere inside the disc $|z| < r$, then the action also diverges. This ensures that only smooth functions $\partial X(z)$, which have no singularity inside the disc, contribute. But for such functions we have

$$\oint \frac{dw}{2\pi i} w^m \partial X(w) = 0 \quad \text{for all } m \geq 0$$

So the state (4.55) is indeed the vacuum state. In fact, since α_0 also annihilates this state, it is identified as the vacuum state with vanishing momentum.

What about the excited states of the theory?

Claim: $\alpha_{-m}|0\rangle = |\partial^m X\rangle$. By which we mean that the state $\alpha_{-m}|0\rangle$ can be built from the path integral,

$$\alpha_{-m}|0\rangle = \int \mathcal{D}X e^{-S[X]} \partial^m X(z=0) \tag{4.56}$$

Proof: We can check this by acting on $|\partial^m X\rangle$ with the annihilation operators α_n .

$$\alpha_n |\partial^m X\rangle \sim \int^{X_f(r)} \mathcal{D}X e^{-S[X]} \oint \frac{dw}{2\pi i} w^n \partial X(w) \partial^m X(z=0)$$

We can focus on the operator insertions and use the OPE (4.23). We drop the path integral and just focus on the operator equation (because, after all, operator equations only make sense in correlation functions which is the same thing as in path integrals). We have

$$\oint \frac{dw}{2\pi i} w^n \partial_z^{m-1} \frac{1}{(w-z)^2} \Big|_{z=0} = m! \oint \frac{dw}{2\pi i} w^{n-m-1} = 0 \quad \text{unless } m = n$$

This confirms that the state (4.56) has the right properties. \square

Finally, we should worry about the zero mode, or momentum $\alpha_0 \sim p$. It is simple to show using the techniques above (together with the OPE (4.26)) that the momentum of a state arises by the insertion of the primary operator e^{ipX} . For example,

$$|0; p\rangle \sim \int \mathcal{D}X e^{-S[X]} e^{ipX(z=0)} .$$

4.7 Brief Comments on Conformal Field Theories with Boundaries

The open string lives on the infinite strip with spatial coordinate $\sigma \in [0, \pi]$. Here we make just a few brief comments on the corresponding conformal field theories.

As before, we can define the complex coordinate $w = \sigma + i\tau$ and make the conformal map

$$z = e^{-iw}$$

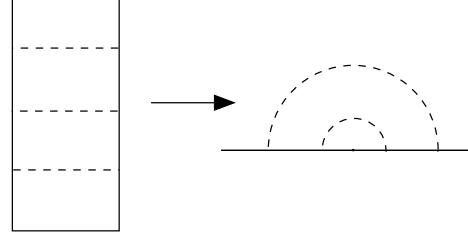


Figure 29:

This time the map takes us to the upper-half plane: $\text{Im}z \geq 0$. The end points of the string are mapped to the real axis, $\text{Im}z = 0$.

Much of our previous discussion goes through as before. But now we need to take care of boundary conditions at $\text{Im}z = 0$. Let's first look at $T_{\alpha\beta}$. Recall that the stress-energy tensor exists because of translational invariance. We still have translational invariance in the direction parallel to the boundary — let's call the associated tangent vector t^α . But translational invariance is broken perpendicular to the boundary — we call the normal vector n^α . The upshot of this is that $T_{\alpha\beta}t^\beta$ remains a conserved current.

To implement Neumann boundary conditions, we insist that none of the current flows out of the boundary. The condition is

$$T_{\alpha\beta}n^\alpha t^\beta = 0 \quad \text{at } \text{Im}z = 0$$

In complex coordinates, this becomes

$$T_{zz} = T_{\bar{z}\bar{z}} \quad \text{at } \text{Im}z = 0$$

There's a simple way to implement this: we extend the definition of T_{zz} from the upper-half plane to the whole complex plane by defining

$$T_{zz}(z) = T_{\bar{z}\bar{z}}(\bar{z})$$

For the closed string we had both functions T and \bar{T} in the whole plane. But for the open string, we have just one of these – say, T , — in the whole plane. This contains the same information as both T and \bar{T} in the upper-half plane. It’s simpler to work in the whole plane and focus just on T . Correspondingly, we now have just a single set of Virasoro generators,

$$L_n = \oint \frac{dz}{2\pi i} z^{n+1} T_{zz}(z)$$

There is no independent \tilde{L}_n for the open string.

A similar doubling trick works when computing the propagator for the free scalar field. The scalar field $X(z, \bar{z})$ is only defined in the upper-half plane. Suppose we want to implement Neumann boundary conditions. Then the propagator is defined by

$$\langle X(z, \bar{z}) X(w, \bar{w}) \rangle = G(z, \bar{z}; w, \bar{w})$$

which obeys $\partial^2 G = -2\pi\alpha' \delta(z - w, \bar{z} - \bar{w})$ subject to the boundary condition

$$\partial_\sigma G(z, \bar{z}; w, \bar{w})|_{\sigma=0} = 0$$

But we solve problems like this in our electrodynamics courses. A useful way of proceeding is to introduce an “image charge” in the lower-half plane. We now let $X(z, \bar{z})$ vary over the whole complex plane with its dynamics governed by the propagator

$$G(z, \bar{z}; w, \bar{w}) = -\frac{\alpha'}{2} \ln |z - w|^2 - \frac{\alpha'}{2} \ln |z - \bar{w}|^2 \quad (4.57)$$

Much of the remaining discussion of CFTs carries forward with only minor differences. However, there is one point that is simple but worth stressing because it will be of importance later. This concerns the state-operator map. Recall the logic that leads us to this idea: we consider a state at fixed time on the strip and propagate it back to past infinity $\tau \rightarrow -\infty$. After the map to the half-plane, past infinity is again the origin. But now the origin lies on the boundary. We learn that the state-operator map relates states to local operators defined on the boundary.

This fact ensures that theories on a strip have fewer states than those on the cylinder. For example, for a free scalar field, Neumann boundary conditions require $\partial X = \bar{\partial}X$ at $\text{Im}z = 0$. (This follows from the requirement that $\partial_\sigma X = 0$ at $\sigma = 0, \pi$ on the strip). On the cylinder, the operators ∂X and $\bar{\partial}X$ give rise to different states; on the strip they give rise to the same state. This, of course, mirrors what we’ve seen for the quantization of the open string where boundary conditions mean that we have only half the oscillator modes to play with.

5. The Polyakov Path Integral and Ghosts

At the beginning of the last chapter, we stressed that there are two very different interpretations of conformal symmetry depending on whether we're thinking of a fixed 2d background or a dynamical 2d background. In applications to statistical physics, the background is fixed and conformal symmetry is a global symmetry. In contrast, in string theory the background is dynamical. Conformal symmetry is a gauge symmetry, a remnant of diffeomorphism invariance and Weyl invariance.

But gauge symmetries are not symmetries at all. They are redundancies in our description of the system. As such, we can't afford to lose them and it is imperative that they don't suffer an anomaly in the quantum theory. At worst, theories with gauge anomalies make no sense. (For example, Yang-Mills theory coupled to only left-handed fundamental fermions is a nonsensical theory for this reason). At best, it may be possible to recover the quantum theory, but it almost certainly has nothing to do with the theory that you started with.

Piecing together some results from the previous chapter, it looks like we're in trouble. We saw that the Weyl symmetry is anomalous since the expectation value of the stress-energy tensor takes different values on backgrounds related by a Weyl symmetry:

$$\langle T_{\alpha}^{\alpha} \rangle = -\frac{c}{12} R$$

On fixed backgrounds, that's merely interesting. On dynamical backgrounds, it's fatal. What can we do? It seems that the only way out is to ensure that our theory has $c = 0$. But we've already seen that $c > 0$ for all non-trivial, unitary CFTs. We seem to have reached an impasse. In this section we will discover the loophole. It turns out that we do indeed require $c = 0$, but there's a way to achieve this that makes sense.

5.1 The Path Integral

In Euclidean space the Polyakov action is given by,

$$S_{\text{Poly}} = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} g^{\alpha\beta} \partial_{\alpha} X^{\mu} \partial_{\beta} X^{\nu} \delta_{\mu\nu}$$

From now on, our analysis of the string will be in terms of the path integral⁶. We integrate over all embedding coordinates X^{μ} and all worldsheet metrics $g_{\alpha\beta}$. Schematically,

⁶The analysis of the string path integral was first performed by Polyakov in “*Quantum geometry of bosonic strings*,” Phys. Lett. B **103**, 207 (1981). The paper weighs in at a whopping 4 pages. As a follow-up, he took another 2.5 pages to analyze the superstring in “*Quantum geometry of fermionic strings*,” Phys. Lett. B **103**, 211 (1981).

the path integral is given by,

$$Z = \frac{1}{\text{Vol}} \int \mathcal{D}g \mathcal{D}X e^{-S_{\text{Poly}}[X,g]}$$

The “Vol” term is all-important. It refers to the fact that we shouldn’t be integrating over all field configurations, but only those physically distinct configurations not related by diffeomorphisms and Weyl symmetries. Since the path integral, as written, sums over all fields, the “Vol” term means that we need to divide out by the volume of the gauge action on field space.

To make the situation more explicit, we need to split the integration over all field configurations into two pieces: those corresponding to physically distinct configurations — schematically depicted as the dotted line in the figure — and those corresponding to gauge transformations — which are shown as solid lines. Dividing by “Vol” simply removes the piece of the partition function which comes from integrating along the solid-line gauge orbits.

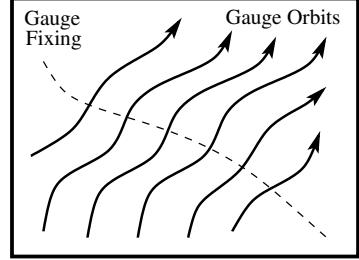


Figure 30:

In an ordinary integral, if we change coordinates then we pick up a Jacobian factor for our troubles. The path integral is no different. We want to decompose our integration variables into physical fields and gauge orbits. The tricky part is to figure out what Jacobian we get. Thankfully, there is a standard method to determine the Jacobian, first introduced by Faddeev and Popov. This method works for all gauge symmetries, including Yang-Mills and you will also learn about it in the “Advanced Quantum Field Theory” course.

5.1.1 The Faddeev-Popov Method

We have two gauge symmetries: diffeomorphisms and Weyl transformations. We will schematically denote both of these by ζ . The change of the metric under a general gauge transformation is $g \rightarrow g^\zeta$. This is shorthand for,

$$g_{\alpha\beta}(\sigma) \longrightarrow g_{\alpha\beta}^\zeta(\sigma') = e^{2\omega(\sigma)} \frac{\partial\sigma^\gamma}{\partial\sigma'^\alpha} \frac{\partial\sigma^\delta}{\partial\sigma'^\beta} g_{\gamma\delta}(\sigma)$$

In two dimensions these gauge symmetries allow us to put the metric into any form that we like — say, \hat{g} . This is called the fiducial metric and will represent our choice of gauge fixing. Two caveats:

- Firstly, it’s not true that we can put any 2d metric into the form \hat{g} of our choosing. This is only true locally. Globally, it remains true if the worldsheet has the

topology of a cylinder or a sphere, but not for higher genus surfaces. We'll revisit this issue in Section 6.

- Secondly, fixing the metric locally to \hat{g} does not fix all the gauge symmetries. We still have the conformal symmetries to deal with. We'll revisit this in the Section 6 as well.

Our goal is to only integrate over physically inequivalent configurations. To achieve this, first consider the integral over the gauge orbit of \hat{g} . For some value of the gauge transformation ζ , the configuration g^ζ will coincide with our original metric g . We can put a delta-function in the integral to get

$$\int \mathcal{D}\zeta \delta(g - g^\zeta) = \Delta_{FP}^{-1}[g] \quad (5.1)$$

This integral isn't equal to one because we need to take into account the Jacobian factor. This is analogous to the statement that $\int dx \delta(f(x)) = 1/|f'|$, evaluated at points where $f(x) = 0$. In the above equation, we have written this Jacobian factor as Δ_{FP}^{-1} . The inverse of this, namely Δ_{FP} , is called the *Faddeev-Popov determinant*. We will evaluate it explicitly shortly. Some comments:

- This whole procedure is rather formal and runs into the usual difficulties with trying to define the path integral. Just as for Yang-Mills theory, we will find that it results in sensible answers.
- We will assume that our gauge fixing is good, meaning that the dotted line in the previous figure cuts through each physically distinct configuration exactly once. Equivalently, the integral over gauge transformations $\mathcal{D}\zeta$ clicks exactly once with the delta-function and we don't have to worry about discrete ambiguities (known as Gribov copies in QCD).
- The measure is taken to be the analogue of the Haar measure for Lie groups, invariant under left and right actions

$$\mathcal{D}\zeta = \mathcal{D}(\zeta'\zeta) = \mathcal{D}(\zeta\zeta')$$

When gauge fixing in Yang-Mills theory, the first thing we do is prove that the Faddeev-Popov determinant Δ_{FP} is gauge invariant. However, our route here is a little more subtle. As we've stressed above, the Weyl anomaly means that our original theory actually fails to be gauge invariant. We will see that the Faddeev-Popov determinant also fails but can, in certain circumstances, cancel the original failure leaving behind a well-defined theory.

The Faddeev-Popov procedure starts by inserting a factor of unity into the path integral, in the guise of

$$1 = \Delta_{FP}[g] \int \mathcal{D}\zeta \delta(g - \hat{g}^\zeta)$$

We'll call the resulting path integral expression $Z[\hat{g}]$ since it depends on the choice of fiducial metric \hat{g} . The first thing we do is use the $\delta(g - \hat{g}^\zeta)$ delta-function to do the integral over metrics,

$$\begin{aligned} Z[\hat{g}] &= \frac{1}{\text{Vol}} \int \mathcal{D}\zeta \mathcal{D}X \mathcal{D}g \Delta_{FP}[g] \delta(g - \hat{g}^\zeta) e^{-S_{\text{Poly}}[X,g]} \\ &= \frac{1}{\text{Vol}} \int \mathcal{D}\zeta \mathcal{D}X \Delta_{FP}[\hat{g}^\zeta] e^{-S_{\text{Poly}}[X,\hat{g}^\zeta]} \end{aligned} \quad (5.2)$$

At this stage the integrand depends on \hat{g}^ζ , where ζ is shorthand for a diffeoemorphism and Weyl transformation. Everything in the equation is invariant under diffeomorphisms, but Weyl transformations are another matter. We know that quantum theory $\int \mathcal{D}X e^{-S_{\text{Poly}}}$ suffers a Weyl anomaly. The action S_{Poly} is invariant under Weyl rescalings, so the subtlety must come from the measure. Meanwhile, anticipating what's to come, we will find a similar issue with the Faddeev-Popov determinant Δ_{FP} .

If, however, we find ourselves in the fortunate situation where the problems cancel then things would work out nicely. In that situation, everything on the right-hand side of (5.2) would be conspire to be invariant under both diffeomorphisms and Weyl transformations and we could write

$$Z[\hat{g}] = \frac{1}{\text{Vol}} \int \mathcal{D}\zeta \mathcal{D}X \Delta_{FP}[\hat{g}] e^{-S_{\text{Poly}}[X,\hat{g}]}$$

But now, nothing depends on the gauge transformation ζ . Indeed, this is precisely the integration over the gauge orbits that we wanted to isolate and it cancels the “Vol” factor sitting outside. We're left with

$$Z[\hat{g}] = \int \mathcal{D}X \Delta_{FP}[\hat{g}] e^{-S_{\text{Poly}}[X,\hat{g}]} \quad (5.3)$$

This is the integral over physically distinct configurations — the dotted line in the previous figure. We see that the Faddeev-Popov determinant is precisely the Jacobian factor that we need.

Clearly the above discussion only flies if we find ourselves in a situation in which the theory (5.2) is genuinely Weyl invariant. Our next task is to understand when this happens which means that we need to figure out what becomes of Δ_{FP} when we do a Weyl transformation.

5.1.2 The Faddeev-Popov Determinant

We still need to compute $\Delta_{FP}[\hat{g}]$. It's defined in (5.1). Let's look at gauge transformations ζ which are close to the identity. In this case, the delta-function $\delta(g - \hat{g}^\zeta)$ is going to be non-zero when the metric g is close to the fiducial metric \hat{g} . In fact, it will be sufficient to look at the delta-function $\delta(\hat{g} - \hat{g}^\zeta)$, which is only non-zero when $\zeta = 0$. We take an infinitesimal Weyl transformation parameterized by $\omega(\sigma)$ and an infinitesimal diffeomorphism $\delta\sigma^\alpha = v^\alpha(\sigma)$. The change in the metric is

$$\delta\hat{g}_{\alpha\beta} = 2\omega\hat{g}_{\alpha\beta} + \nabla_\alpha v_\beta + \nabla_\beta v_\alpha$$

Plugging this into the delta-function, the expression for the Faddeev-Popov determinant becomes

$$\Delta_{FP}^{-1}[\hat{g}] = \int \mathcal{D}\omega \mathcal{D}v \delta(2\omega\hat{g}_{\alpha\beta} + \nabla_\alpha v_\beta + \nabla_\beta v_\alpha) \quad (5.4)$$

where we've replaced the integral $\mathcal{D}\zeta$ over the gauge group with the integral $\mathcal{D}\omega \mathcal{D}v$ over the Lie algebra of group since we're near the identity. (We also suppress the subscript on v_α in the measure factor to keep things looking tidy).

At this stage it's useful to represent the delta-function in its integral, Fourier form. For a single delta-function, this is $\delta(x) = \int dp \exp(2\pi i p x)$. But the delta-function in (5.4) is actually a delta-functional: it restricts a whole function. Correspondingly, the integral representation is in terms of a functional integral,

$$\Delta_{FP}^{-1}[\hat{g}] = \int \mathcal{D}\omega \mathcal{D}v \mathcal{D}\beta \exp \left(2\pi i \int d^2\sigma \sqrt{\hat{g}} \beta^{\alpha\beta} [2\omega\hat{g}_{\alpha\beta} + \nabla_\alpha v_\beta + \nabla_\beta v_\alpha] \right)$$

where $\beta^{\alpha\beta}$ is a symmetric 2-tensor on the worldsheet.

We now simply do the $\int \mathcal{D}\omega$ integral. It doesn't come with any derivatives, so it merely acts as a Lagrange multiplier, setting

$$\beta^{\alpha\beta} \hat{g}_{\alpha\beta} = 0$$

In other words, after performing the ω integral, $\beta^{\alpha\beta}$ is symmetric and traceless. We'll take this to be the definition of $\beta^{\alpha\beta}$ from now on. So, finally we have

$$\Delta_{FP}^{-1}[\hat{g}] = \int \mathcal{D}v \mathcal{D}\beta \exp \left(4\pi i \int d^2\sigma \sqrt{\hat{g}} \beta^{\alpha\beta} \nabla_\alpha v_\beta \right)$$

5.1.3 Ghosts

The previous manipulations give us an expression for Δ_{FP}^{-1} . But we want to invert it to get Δ_{FP} . Thankfully, there's a simple way to achieve this. Because the integrand is quadratic in v and β , we know that the integral computes the inverse determinant of the operator ∇_α . (Strictly speaking, it computes the inverse determinant of the projection of ∇_α onto symmetric, traceless tensors. This observation is important because it means the relevant operator is a square matrix which is necessary to talk about a determinant). But we also know how to write down an expression for the determinant Δ_{FP} , instead of its inverse, in terms of path integrals: we simply need to replace the commuting integration variables with anti-commuting fields,

$$\begin{aligned}\beta_{\alpha\beta} &\longrightarrow b_{\alpha\beta} \\ v^\alpha &\longrightarrow c^\alpha\end{aligned}$$

where b and c are both Grassmann-valued fields (i.e. anti-commuting). They are known as *ghost fields*. This gives us our final expression for the Faddeev-Popov determinant,

$$\Delta_{FP}[g] = \int \mathcal{D}b \mathcal{D}c \exp[iS_{\text{ghost}}]$$

where the ghost action is defined to be

$$S_{\text{ghost}} = \frac{1}{2\pi} \int d^2\sigma \sqrt{g} b_{\alpha\beta} \nabla^\alpha c^\beta \quad (5.5)$$

and we have chosen to rescale the b and c fields at this last step to get a factor of $1/2\pi$ sitting in front of the action. (This only changes the normalization of the partition function which doesn't matter). Rotating back to Euclidean space, the factor of i disappears. The expression for the full partition function (5.3) is

$$Z[\hat{g}] = \int \mathcal{D}X \mathcal{D}b \mathcal{D}c \exp(-S_{\text{Poly}}[X, \hat{g}] - S_{\text{ghost}}[b, c, \hat{g}])$$

Something lovely has happened. Although the ghost fields were introduced as some auxiliary constructs, they now appear on the same footing as the dynamical fields X . We learn that gauge fixing comes with a price: our theory has extra ghost fields.

The role of these ghost fields is to cancel the unphysical gauge degrees of freedom, leaving only the $D - 2$ transverse modes of X^μ . Unlike lightcone quantization, they achieve this in a way which preserves Lorentz invariance.

Simplifying the Ghost Action

The ghost action (5.5) looks fairly simple. But it looks even simpler if we work in conformal gauge,

$$\hat{g}_{\alpha\beta} = e^{2\omega} \delta_{\alpha\beta}$$

The determinant is $\sqrt{\hat{g}} = e^{2\omega}$. Recall that in complex coordinates, the measure is $d^2\sigma = \frac{1}{2}d^2z$, while we can lower the index on the covariant derivative using $\nabla^z = g^{z\bar{z}}\nabla_{\bar{z}} = 2e^{-2\omega}\nabla_{\bar{z}}$. We have

$$S_{\text{ghost}} = \frac{1}{2\pi} \int d^2z (b_{zz}\nabla_{\bar{z}}c^z + b_{\bar{z}\bar{z}}\nabla_z c^{\bar{z}})$$

In deriving this, remember that there is no field $b_{z\bar{z}}$ because $b_{\alpha\beta}$ is traceless. Now comes the nice part: the covariant derivatives are actually just ordinary derivatives. To see why this is the case, look at

$$\nabla_{\bar{z}}c^z = \partial_{\bar{z}}c^z + \Gamma_{\bar{z}\alpha}^z c^\alpha$$

But the Christoffel symbols are given by

$$\Gamma_{\bar{z}\alpha}^z = \frac{1}{2}g^{z\bar{z}}(\partial_{\bar{z}}g_{\alpha\bar{z}} + \partial_\alpha g_{\bar{z}\bar{z}} - \partial_{\bar{z}}g_{\bar{z}\alpha}) = 0 \quad \text{for } \alpha = z, \bar{z}$$

So in conformal gauge, the ghost action factorizes into two free theories,

$$S_{\text{ghost}} = \frac{1}{2\pi} \int d^2z b_{zz} \partial_{\bar{z}}c^z + b_{\bar{z}\bar{z}} \partial_z c^{\bar{z}}$$

The action doesn't depend on the conformal factor ω . In other words, it is Weyl invariant without any need to change b and c : these are therefore both neutral under Weyl transformations.

(It's worth pointing out that $b_{\alpha\beta}$ and c^α are neutral under Weyl transformations. But if we raise or lower these indices, then the fields pick up factors of the metric. So $b^{\alpha\beta}$ and c_α would not be neutral under Weyl transformations).

5.2 The Ghost CFT

Fixing the Weyl and diffeomorphism gauge symmetries has left us with two new dynamical ghost fields, b and c . Both are Grassmann (i.e. anti-commuting) variables. Their dynamics is governed by a CFT. Define

$$\begin{aligned} b &= b_{zz} & , & \bar{b} = b_{\bar{z}\bar{z}} \\ c &= c^z & , & \bar{c} = c^{\bar{z}} \end{aligned}$$

The ghost action is given by

$$S_{\text{ghost}} = \frac{1}{2\pi} \int d^2z \ (b \bar{\partial}c + \bar{b} \partial\bar{c})$$

Which gives the equations of motion

$$\bar{\partial}b = \partial\bar{b} = \bar{\partial}c = \partial\bar{c} = 0$$

So we see that b and c are holomorphic fields, while \bar{b} and \bar{c} are anti-holomorphic.

Before moving onto quantization, there's one last bit of information we need from the classical theory: the stress tensor for the bc ghosts. The calculation is a little bit fiddly. We use the general definition of the stress tensor (4.4), which requires us to return to the theory (5.5) on a general background and vary the metric $g^{\alpha\beta}$. The complications are twofold. Firstly, we pick up a contribution from the Christoffel symbol that is lurking inside the covariant derivative ∇^α . Secondly, we must also remember that $b_{\alpha\beta}$ is traceless. But this is a condition which itself depends on the metric: $b_{\alpha\beta}g^{\alpha\beta} = 0$. To account for this we should add a Lagrange multiplier to the action imposing tracelessness. After correctly varying the metric, we may safely retreat back to flat space where the end result is rather simple. We have $T_{z\bar{z}} = 0$, as we must for any conformal theory. Meanwhile, the holomorphic and anti-holomorphic parts of the stress tensor are given by,

$$T = 2(\partial c)b + c\partial b \quad , \quad \bar{T} = 2(\bar{\partial}\bar{c})\bar{b} + \bar{c}\bar{\partial}\bar{b}. \quad (5.6)$$

Operator Product Expansions

We can compute the OPEs of these fields using the standard path integral techniques that we employed in the last chapter. In what follows, we'll just focus on the holomorphic piece of the CFT. We have, for example,

$$0 = \int \mathcal{D}b \mathcal{D}c \frac{\delta}{\delta b(\sigma)} [e^{-S_{\text{ghost}}} b(\sigma')] = \int \mathcal{D}b \mathcal{D}c e^{-S_{\text{ghost}}} \left[-\frac{1}{2\pi} \bar{\partial}c(\sigma) b(\sigma') + \delta(\sigma - \sigma') \right]$$

which tells us that

$$\bar{\partial}c(\sigma) b(\sigma') = 2\pi \delta(\sigma - \sigma')$$

Similarly, looking at $\delta/\delta c(\sigma)$ gives

$$\bar{\partial}b(\sigma) c(\sigma') = 2\pi \delta(\sigma - \sigma')$$

We can integrate both of these equations using our favorite formula $\bar{\partial}(1/z) = 2\pi\delta(z, \bar{z})$. We learn that the OPEs between fields are given by

$$b(z)c(w) = \frac{1}{z-w} + \dots$$

$$c(w)b(z) = \frac{1}{w-z} + \dots$$

In fact the second equation follows from the first equation and Fermi statistics. The OPEs of $b(z)b(w)$ and $c(z)c(w)$ have no singular parts. They vanish as $z \rightarrow w$.

Finally, we need the stress tensor of the theory. After normal ordering, it is given by

$$T(z) = 2 : \partial c(z) b(z) : + : c(z) \partial b(z) :$$

We will shortly see that with this choice, b and c carry appropriate weights for tensor fields which are neutral under Weyl rescaling.

Primary Fields

We will now show that both b and c are primary fields, with weights $h = 2$ and $h = -1$ respectively. Let's start by looking at c . The OPE with the stress tensor is

$$T(z)c(w) = 2 : \partial c(z) b(z) : c(w) + : c(z) \partial b(z) : c(w)$$

$$= \frac{2\partial c(z)}{z-w} - \frac{c(z)}{(z-w)^2} + \dots = -\frac{c(w)}{(z-w)^2} + \frac{\partial c(w)}{z-w} + \dots$$

confirming that c has weight -1 . When taking the OPE with b , we need to be a little more careful with minus signs. We get

$$T(z)b(w) = 2 : \partial c(z) b(z) : b(w) + : c(z) \partial b(z) : b(w)$$

$$= -2b(z) \left(\frac{-1}{(z-w)^2} \right) - \frac{\partial b(z)}{z-w} = \frac{2b(w)}{(z-w)^2} + \frac{\partial b(w)}{z-w} + \dots$$

showing that b has weight 2 . As we've pointed out a number of times, conformal = diffeo + Weyl. We mentioned earlier that the fields b and c are neutral under Weyl transformations. This is reflected in their weights, which are due solely to diffeomorphisms as dictated by their index structure: b_{zz} and c^z .

The Central Charge

Finally, we can compute the TT OPE to determine the central charge of the bc ghost system.

$$T(z)T(w) = 4 : \partial c(z) b(z) : : \partial c(w) b(w) : + 2 : \partial c(z) b(z) : : c(w) \partial b(w) :$$

$$+ 2 : c(z) \partial b(z) : : \partial c(w) b(w) : + : c(z) \partial b(z) : : c(w) \partial b(w) :$$

For each of these terms, making two contractions gives a $(z-w)^{-4}$ contribution to the OPE. There are also two ways to make a single contraction. These give $(z-w)^{-1}$ or $(z-w)^{-2}$ or $(z-w)^{-3}$ contributions depending on what the derivatives hit. The end result is

$$\begin{aligned} T(z)T(w) = & \frac{-4}{(z-w)^4} + \frac{4 : \partial c(z)b(w) :}{(z-w)^2} - \frac{4 : b(z)\partial c(w) :}{(z-w)^2} \\ & - \frac{4}{(z-w)^4} + \frac{2 : \partial c(z)\partial b(w) :}{z-w} - \frac{4 : b(z)c(w) :}{(z-w)^3} \\ & - \frac{4}{(z-w)^4} - \frac{4 : c(z)b(w) :}{(z-w)^3} + \frac{2 : \partial b(z)\partial c(w) :}{z-w} \\ & - \frac{1}{(z-w)^4} - \frac{: c(z)\partial b(w) :}{(z-w)^2} + \frac{\partial b(z)c(w) :}{(z-w)^2} + \dots \end{aligned}$$

After some Taylor expansions to turn $f(z)$ functions into $f(w)$ functions, together with a little collecting of terms, this can be written as,

$$T(z)T(w) = \frac{-13}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots$$

The first thing to notice is that it indeed has the form expected of TT OPE. The second, and most important, thing to notice is the central charge of the bc ghost system: it is

$$c = -26$$

5.3 The Critical “Dimension” of String Theory

Let’s put the pieces together. We’ve learnt that gauge fixing the diffeomorphisms and Weyl gauge symmetries results in the introduction of ghosts which contribute central charge $c = -26$. We’ve also learnt that the Weyl symmetry is anomalous unless $c = 0$. Since the Weyl symmetry is a gauge symmetry, it’s crucial that we keep it. We’re forced to add exactly the right degrees of freedom to the string to cancel the contribution from the ghosts.

The simplest possibility is to add D free scalar fields. Each of these contributes $c = 1$ to the central charge, so the whole procedure is only consistent if we pick

$$D = 26$$

This agrees with the result we found in Chapter 2: it is the critical dimension of string theory.

However, there's no reason that we have to work with free scalar fields. The consistency requirement is merely that the degrees of freedom of the string are described by a CFT with $c = 26$. Any CFT will do. Each such CFT describes a different background in which a string can propagate. If you like, the space of CFTs with $c = 26$ can be thought of as the space of classical solutions of string theory.

We learn that the “critical dimension” of string theory is something of a misnomer: it is really a “critical central charge”. Only for rather special CFTs can this central charge be thought of as a spacetime dimension.

For example, if we wish to describe strings moving in 4d Minkowski space, we can take $D = 4$ free scalars (one of which will be timelike) together with some other $c = 22$ CFT. This CFT may have a geometrical interpretation, or it may be something more abstract. The CFT with $c = 22$ is sometimes called the “internal sector” of the theory. It is what we really mean when we talk about the “extra hidden dimensions of string theory”. We'll see some examples of CFTs describing curved spaces in Section 7.

There's one final subtlety: we need to be careful with the transition back to Minkowski space. After all, we want one of the directions of the CFT, X^0 , to have the wrong sign kinetic term. One safe way to do this is to keep X^0 as a free scalar field, with the remaining degrees of freedom described by some $c = 25$ CFT. This doesn't seem quite satisfactory though since it doesn't allow for spacetimes which evolve in time — and, of course, these are certainly necessary if we wish to understand early universe cosmology. There are still some technical obstacles to understanding the worldsheet of the string in time-dependent backgrounds. To make progress, and discuss string cosmology, we usually bi-pass this issue by working with the low-energy effective action which we will derive in Section 7.

5.3.1 The Usual Nod to the Superstring

The superstring has another gauge symmetry on the worldsheet: supersymmetry. This gives rise to more ghosts, the so-called $\beta\gamma$ system, which turns out to have central charge +11. Consistency then requires that the degrees of freedom of the string have central charge $c = 26 - 11 = 15$.

However, now the CFTs must themselves be invariant under supersymmetry, which means that bosons come matched with fermions. If we add D bosons, then we also need to add D fermions. A free boson has $c = 1$, while a free fermion has $c = 1/2$. So, the total number of free bosons that we should add is $D(1 + 1/2) = 15$, giving us the

critical dimension of the superstring:

$$D = 10$$

5.3.2 An Aside: Non-Critical Strings

Although it's a slight departure from our main narrative, it's worth pausing to mention what Polyakov actually did in his four page paper. His main focus was not critical strings, with $D = 26$, but rather *non-critical* strings with $D \neq 26$. From the discussion above, we know that these suffer from a Weyl anomaly. But it turns out that there is a way to make sense of the situation.

The starting point is to abandon Weyl invariance from the beginning. We start with D free scalar fields coupled to a dynamical worldsheet metric $g_{\alpha\beta}$. (More generally, we could have any CFT). We still want to keep reparameterization invariance, but now we ignore the constraints of Weyl invariance. Of course, it seems likely that this isn't going to have too much to do with the Nambu-Goto string, but let's proceed anyway. Without Weyl invariance, there is one extra term that it is natural to add to the 2d theory: a worldsheet cosmological constant μ ,

$$S_{\text{non-critical}} = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} (g^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X_\mu + \mu)$$

Our goal will be to understand how the partition function changes under a Weyl rescaling. There will be two contributions: one from the explicit μ dependence and one from the Weyl anomaly. Consider two metrics related by a Weyl transformation

$$\hat{g}_{\alpha\beta} = e^{2\omega} g_{\alpha\beta}$$

As we vary ω , the partition function $Z[\hat{g}]$ changes as

$$\begin{aligned} \frac{1}{Z} \frac{\partial Z}{\partial \omega} &= \frac{1}{Z} \int \mathcal{D}X e^{-S} \left(-\frac{\partial S}{\partial \hat{g}_{\alpha\beta}} \frac{\partial \hat{g}_{\alpha\beta}}{\partial \omega} \right) \\ &= \frac{1}{Z} \int \mathcal{D}X e^{-S} \left(-\frac{1}{2\pi} \sqrt{\hat{g}} T^\alpha_\alpha \right) \\ &= \frac{c}{24\pi} \sqrt{\hat{g}} \hat{R} - \frac{1}{2\pi\alpha'} \mu e^{2\omega} \\ &= \frac{c}{24\pi} \sqrt{g} (R - 2\nabla^2 \omega) - \frac{1}{2\pi\alpha'} \mu e^{2\omega} \end{aligned}$$

where, in the last two lines, we used the Weyl anomaly (4.35) and the relationship between Ricci curvatures (1.29). The central charge appearing in these formulae includes the contribution from the ghosts,

$$c = D - 26$$

We can now just treat this as a differential equation for the partition function Z and solve. This allows us to express the partition function $Z[\hat{g}]$, defined on one worldsheet metric, in terms of $Z[g]$, defined on another. The relationship is,

$$Z[\hat{g}] = Z[g] \exp \left[-\frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} \left(2\mu e^{2\omega} - \frac{c\alpha'}{6} (g^{\alpha\beta} \partial_\alpha \omega \partial^\beta \omega + R\omega) \right) \right]$$

We see that the scaling mode ω inherits a kinetic term. It now appears as a new dynamical scalar field in the theory. It is often called the Liouville field on account of the exponential potential term multiplying μ . Solving this theory is quite hard⁷. Notice also that our new scalar field ω appears in the final term multiplying the Ricci scalar R . We will describe the significance of this in Section 7.2.1. We'll also see another derivation of this kind of Lagrangian in Section 7.4.4.

5.4 States and Vertex Operators

In Chapter 2 we determined the spectrum of the string in flat space. What is the spectrum for a general string background? The theory consists of the b and c ghosts, together with a $c = 26$ CFT. At first glance, it seems that we have a greatly enlarged Hilbert space since we can act with creation operators from all fields, including the ghosts. However, as you might expect, not all of these states will be physical. After correctly accounting for the gauge symmetry, only some subset survives.

The elegant method to determine the physical Hilbert space in a gauge fixed action with ghosts is known as *BRST quantization*. You will learn about it in the “Advanced Quantum Field Theory” course where you will apply it to Yang-Mills theory. Although a correct construction of the string spectrum employs the BRST method, we won’t describe it here for lack of time. A very clear description of the general method and its application to the string can be found in Section 4.2 of Polchinski’s book.

Instead, we will make do with a poor man’s attempt to determine the spectrum of the string. Our strategy is to simply pretend that the ghosts aren’t there and focus on the states created by the fields of the matter CFT (i.e. the X^μ fields if we’re talking about flat space). As we’ll explain in the next section, if we’re only interested in tree-level scattering amplitudes then this will suffice.

To illustrate how to compute the spectrum of the string, let’s go back to flat $D = 26$ dimensional Minkowski space and the discussion of covariant quantization in Section

⁷A good review can be found Seiberg’s article “*Notes on Quantum Liouville Theory and Quantum Gravity*”, Prog. Theor. Phys. Supl. 102 (1990) 319.

[2.1](#). We found that physical states $|\Psi\rangle$ are subject to the Virasoro constraints [\(2.6\)](#) and [\(2.7\)](#) which read

$$\begin{aligned} L_n |\Psi\rangle &= 0 && \text{for } n > 0 \\ L_0 |\Psi\rangle &= a |\Psi\rangle \end{aligned}$$

and similar for \tilde{L}_n ,

$$\begin{aligned} \tilde{L}_n |\Psi\rangle &= 0 && \text{for } n > 0 \\ \tilde{L}_0 |\Psi\rangle &= \tilde{a} |\Psi\rangle \end{aligned}$$

where we have, just briefly, allowed for the possibility of different normal ordering coefficients a and \tilde{a} for the left- and right-moving sectors. But there's a name for states in a conformal field theory obeying these requirements: they are primary states of weight (a, \tilde{a}) .

So how do we fix the normal ordering ambiguities a and \tilde{a} ? A simple way is to first replace the states with operator insertions on the worldsheet using the state-operator map: $|\Psi\rangle \rightarrow \mathcal{O}$. But we have a further requirement on the operators \mathcal{O} : gauge invariance. There are two gauge symmetries: reparameterization invariance and Weyl symmetry. Both restrict the possible states.

Let's start by considering reparameterization invariance. In the last section, we happily placed operators at specific points on the worldsheet. But in a theory with a dynamical metric, this doesn't give rise to a diffeomorphism invariant operator. To make an object that is invariant under reparameterizations of the worldsheet coordinates, we should integrate over the whole worldsheet. Our operator insertions (in conformal gauge) are therefore of the form,

$$V \sim \int d^2z \mathcal{O} \tag{5.7}$$

Here the \sim sign reflects the fact that we've dropped an overall normalization constant which we'll return to in the next section.

Integrating over the worldsheet takes care of diffeomorphisms. But what about Weyl symmetries? The measure d^2z has weight $(-1, -1)$ under rescaling. To compensate, the operator \mathcal{O} must have weight $(+1, +1)$. This is how we fix the normal ordering ambiguity: we require $a = \tilde{a} = 1$. Note that this agrees with the normal ordering coefficient $a = 1$ that we derived in lightcone quantization in Chapter 2.

This, then, is the rather rough derivation of the string spectrum. The physical states are the primary states of the CFT with weight $(+1, +1)$. The operators (5.7) associated to these states are called *vertex operators*.

5.4.1 An Example: Closed Strings in Flat Space

Let's use this new language to rederive the spectrum of the closed string in flat space. We start with the ground state of the string, which was previously identified as a tachyon. As we saw in Section 4, the vacuum of a CFT is associated to the identity operator. But we also have the zero modes. We can give the string momentum p^μ by acting with the operator $e^{ip \cdot X}$. The vertex operator associated to the ground state of the string is therefore

$$V_{\text{tachyon}} \sim \int d^2z : e^{ip \cdot X} : \quad (5.8)$$

In Section 4.3.3, we showed that the operator $e^{ip \cdot X}$ is primary with weight $h = \tilde{h} = \alpha' p^2 / 4$. But Weyl invariance requires that the operator has weight $(+1, +1)$. This is only true if the mass of the state is

$$M^2 \equiv -p^2 = -\frac{4}{\alpha'}$$

This is precisely the mass of the tachyon that we saw in Section 2.

Let's now look at the first excited states. In covariant quantization, these are of the form $\zeta_{\mu\nu} \alpha_{-1}^\mu \tilde{\alpha}_{-1}^\nu |0; p\rangle$, where $\zeta_{\mu\nu}$ is a constant tensor that determines the type of state, together with its polarization. (Recall: traceless symmetric $\zeta_{\mu\nu}$ corresponds to the graviton, anti-symmetric $\zeta_{\mu\nu}$ corresponds to the $B_{\mu\nu}$ field and the trace of $\zeta_{\mu\nu}$ corresponds to the scalar known as the dilaton). From (4.56), the vertex operator associated to this state is,

$$V_{\text{excited}} \sim \int d^2z : e^{ip \cdot X} \partial X^\mu \bar{\partial} X^\nu : \zeta_{\mu\nu} \quad (5.9)$$

where ∂X^μ gives us a α_{-1}^μ excitation, while $\bar{\partial} X^\mu$ gives a $\tilde{\alpha}_{-1}^\mu$ excitation. It's easy to check that the weight of this operator is $h = \tilde{h} = 1 + \alpha' p^2 / 4$. Weyl invariance therefore requires that

$$p^2 = 0$$

confirming that the first excited states of the string are indeed massless. However, we still need to check that the operator in (5.9) is actually primary. We know that ∂X is

primary and we know that $e^{ip \cdot X}$ is primary, but now we want to consider them both sitting together inside the normal ordering. This means that there are extra terms in the Wick contraction which give rise to $1/(z-w)^3$ terms in the OPE, potentially ruining the primacy of our operator. One such term arises from a double contraction, one of which includes the $e^{ip \cdot X}$ operator. This gives rise to an offending term proportional to $p^\mu \zeta_{\mu\nu}$. The same kind of contraction with \bar{T} gives rise to a term proportional to $p^\nu \zeta_{\nu\mu}$. In order for these terms to vanish, the polarization tensor must satisfy

$$p^\mu \zeta_{\mu\nu} = p^\nu \zeta_{\mu\nu} = 0$$

which is precisely the transverse polarization condition expected for a massless particle.

5.4.2 An Example: Open Strings in Flat Space

As explained in Section 4.7, vertex operators for the open-string are inserted on the boundary $\partial\mathcal{M}$ of the worldsheet. We still need to ensure that these operators are diffeomorphism invariant which is achieved by integrating over $\partial\mathcal{M}$. The vertex operator for the open string tachyon is

$$V_{\text{tachyon}} \sim \int_{\partial\mathcal{M}} ds : e^{ip \cdot X} :$$

We need to figure out the dimension of the boundary operator $: e^{ip \cdot X} :$. It's not the same as for the closed string. The reason is due to presence of the image charge in the propagator (4.57) for a free scalar field on a space with boundary. This propagator appears in the Wick contractions in the OPEs and affects the weights. Let's see why this is the case. Firstly, we look at a single scalar field X ,

$$\begin{aligned} \partial X(z) : e^{ipX(w, \bar{w})} : &= \sum_{n=1}^{\infty} \frac{(ip)^n}{(n-1)!} : X(w, \bar{w})^{n-1} : \left(-\frac{\alpha'}{2} \frac{1}{z-w} - \frac{\alpha'}{2} \frac{1}{z-\bar{w}} \right) + \dots \\ &= -\frac{i\alpha' p}{2} : e^{ipX(w, \bar{w})} : \left(\frac{1}{z-w} + \frac{1}{z-\bar{w}} \right) + \dots \end{aligned}$$

With this result, we can now compute the OPE with T ,

$$T(z) : e^{ipX(w, \bar{w})} : = \frac{\alpha' p^2}{4} : e^{ipX} : \left(\frac{1}{z-w} + \frac{1}{z-\bar{w}} \right)^2 + \dots$$

When the operator $: e^{ipX(w, \bar{w})} :$ is placed on the boundary $w = \bar{w}$, this becomes

$$T(z) : e^{ipX(w, \bar{w})} : = \frac{\alpha' p^2 : e^{ipX(w, \bar{w})} :}{(z-w)^2} + \dots$$

This tells us that the boundary operator $: e^{ip \cdot X} :$ is indeed primary, with weight $\alpha' p^2$.

For the open string, Weyl invariance requires that operators have weight +1 in order to cancel the scaling dimension of -1 coming from the boundary integral $\int ds$. So the mass of the open string ground state is

$$M^2 \equiv -p^2 = -\frac{1}{\alpha'}$$

in agreement with the mass of the open string tachyon computed in Section 3.

The vertex operator for the photon is

$$V_{\text{photon}} \sim \int_{\partial\mathcal{M}} ds \zeta_a : \partial X^a e^{ip \cdot X} : \quad (5.10)$$

where the index $a = 0, \dots, p$ now runs only over those directions with Neumann boundary conditions that lie parallel to the brane worldvolume. The requirement that this is a primary operator gives $p^a \zeta_a = 0$, while Weyl invariance tells us that $p^2 = 0$. This is the expected behaviour for the momentum and polarization of a photon.

5.4.3 More General CFTs

Let's now consider a string propagating in four-dimensional Minkowski space \mathcal{M}_4 , together with some internal CFT with $c = 22$. Then any primary operator of the internal CFT with weight (h, h) can be assigned momentum p^μ , for $\mu = 0, 1, 2, 3$ by dressing the operator with $e^{ip \cdot X}$. In order to get a primary operator of weight $(+1, +1)$ as required, we must have

$$\frac{\alpha' p^2}{4} = 1 - h$$

We see that the mass spectrum of closed string states is given by

$$M^2 = \frac{4}{\alpha'}(h - 1)$$

where h runs over the spectrum of primary operators of the internal CFT. Some comments:

- Relevant operators in the internal CFT have $h < 1$ and give rise to tachyons in the spectrum. Marginal operators, with $h = 1$, give massless particles. And irrelevant operators result in massive states.
- Notice that requiring the vertex operators to be Weyl invariant determines the mass formula for the state. We say that the vertex operators are “on-shell”, in the same sense that external legs of Feynman diagrams are on-shell. We will have more to say about this in the next section.

6. String Interactions

So far, despite considerable effort, we've only discussed the free string. We now wish to consider interactions. If we take the analogy with quantum field theory as our guide, then we might be led to think that interactions require us to add various non-linear terms to the action. However, this isn't the case. Any attempt to add extra non-linear terms for the string won't be consistent with our precious gauge symmetries. Instead, rather remarkably, all the information about interacting strings is already contained in the free theory described by the Polyakov action. (Actually, this statement is almost true).

To see that this is at least feasible, try to draw a cartoon picture of two strings interacting. It looks something like the worldsheet shown in the figure. The worldsheet is smooth. In Feynman diagrams in quantum field theory, information about interactions is inserted at vertices, where different lines meet. Here there are no such points. Locally, every part of the diagram looks like a free propagating string. Only globally do we see that the diagram describes interactions.

6.1 What to Compute?

If the information about string interactions is already contained in the Polyakov action, let's go ahead and compute something! But what should we compute? One obvious thing to try is the probability for a particular configuration of strings at an early time to evolve into a new configuration at some later time. For example, we could try to compute the amplitude associated to the diagram above, stipulating fixed curves for the string ends.

No one knows how to do this. Moreover, there are words that we can drape around this failure that suggests this isn't really a sensible thing to compute. I'll now try to explain these words. Let's start by returning to the familiar framework of quantum field theory in a fixed background. There the basic objects that we can compute are correlation functions,

$$\langle \phi(x_1) \dots \phi(x_n) \rangle \tag{6.1}$$

After a Fourier transform, these describe Feynman diagrams in which the external legs carry arbitrary momenta. For this reason, they are referred to as *off-shell*. To get the scattering amplitudes, we simply need to put the external legs on-shell (and perform a few other little tricks captured in the LSZ reduction formula).



Figure 31:

The discussion above needs amendment if we turn on gravity. Gravity is a gauge theory and the gauge symmetries are diffeomorphisms. In a gauge theory, only gauge invariant observables make sense. But the correlation function (6.1) is not gauge invariant because its value changes under a diffeomorphism which maps the points x_i to another point. This emphasizes an important fact: there are no local off-shell gauge invariant observables in a theory of gravity.

There is another way to say this. We know, by causality, that space-like separated operators should commute in a quantum field theory. But in gravity the question of whether operators are space-like separated becomes a dynamical issue and the causal structure can fluctuate due to quantum effects. This provides another reason why we are unable to define local gauge invariant observables in any theory of quantum gravity.

Let's now return to string theory. Computing the evolution of string configurations for a finite time is analogous to computing off-shell correlation functions in QFT. But string theory is a theory of gravity so such things probably don't make sense. For this reason, we retreat from attempting to compute correlation functions, back to the S-matrix.

The String S-Matrix

The object that we can compute in string theory is the S-matrix. This is obtained by taking the points in the correlation function to infinity: $x_i \rightarrow \infty$. This is acceptable because, just like in the case of QED, the redundancy of the system consists of those gauge transformations which die off asymptotically. Said another way, points on the boundary don't fluctuate in quantum gravity. (Such fluctuations would be over an infinite volume of space and are suppressed due to their infinite action).

So what we're really going to calculate is a diagram of the type shown in the figure, where all external legs are taken to infinity. Each of these legs can be placed in a different state of the free string and assigned some spacetime momentum p_i . The resulting expression is the string *S-matrix*.

Using the state-operator map, we know that each of these states at infinity is equivalent to the insertion of an appropriate vertex operator on the worldsheet. Therefore, to compute this S-matrix element we use a conformal transformation to bring each of these infinite legs to a finite distance. The end result is a worldsheet with the topology

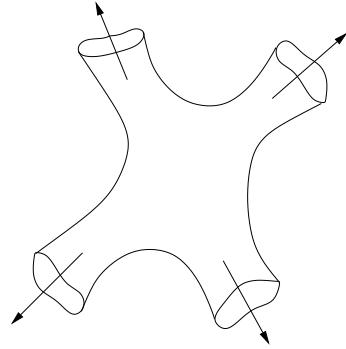


Figure 32:

of the sphere, dotted with vertex operators where the legs used to be. However, we already saw in the previous section that the constraint of Weyl invariance meant that vertex operators are necessarily on-shell. Technically, this is the reason that we can only compute on-shell correlation functions in string theory.

6.1.1 Summing Over Topologies

The Polyakov path integral instructs us to sum over all metrics. But what about worldsheets of different topologies? In fact, we should also sum over these. It is this sum that gives the perturbative expansion of string theory. The scattering of two strings receives contributions from worldsheets of the form

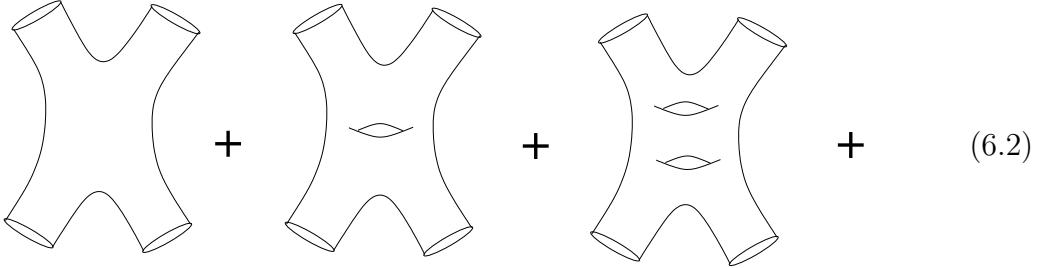


Figure 33:

The only thing that we need to know is how to weight these different worldsheets. Thankfully, there is a very natural coupling on the string that we have yet to consider and this will do the job. We augment the Polyakov action by

$$S_{\text{string}} = S_{\text{Poly}} + \lambda \chi \quad (6.3)$$

Here λ is simply a real number, while χ is given by an integral over the (Euclidean) worldsheet

$$\chi = \frac{1}{4\pi} \int d^2\sigma \sqrt{g} R \quad (6.4)$$

where R is the Ricci scalar of the worldsheet metric. This looks like the Einstein-Hilbert term for gravity on the worldsheet. It is simple to check that it is invariant under reparameterizations and Weyl transformations.

In four-dimensions, the Einstein-Hilbert term makes gravity dynamical. But life is very different in 2d. Indeed, we've already seen that all the components of the metric can be gauged away so there are no propagating degrees of freedom associated to $g_{\alpha\beta}$. So, in two-dimensions, the term (6.4) doesn't make gravity dynamical: in fact, classically, it doesn't do anything at all!

The reason for this is that χ is a topological invariant. This means that it doesn't actually depend on the metric $g_{\alpha\beta}$ at all – it depends only on the topology of the worldsheet. (More precisely, χ only depends on those global properties of the metric which themselves depend on the topology of the worldsheet). This is the content of the Gauss-Bonnet theorem: the integral of the Ricci scalar R over the worldsheet gives an integer, χ , known as the Euler number of the worldsheet. For a worldsheet without boundary (i.e. for the closed string) χ counts the number of handles h on the worldsheet. It is given by,

$$\chi = 2 - 2h = 2(1 - g) \quad (6.5)$$

where g is called the *genus* of the surface. The simplest examples are shown in the figure. The sphere has $g = 0$ and $\chi = 2$; the torus has $g = 1$ and $\chi = 0$. For higher $g > 1$, the Euler character χ is negative.

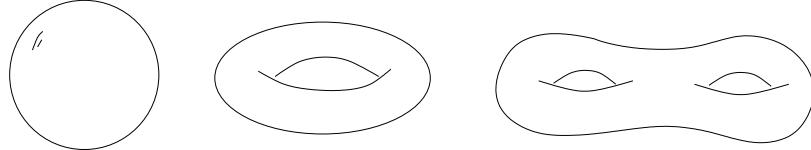


Figure 34: Examples of increasingly poorly drawn Riemann surfaces with $\chi = 2, 0$ and -2 .

Now we see that the number λ — or, more precisely, e^λ — plays the role of the string coupling. The integral over worldsheets is weighted by,

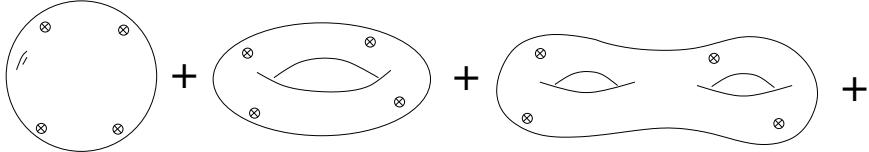
$$\sum_{\substack{\text{topologies} \\ \text{metrics}}} e^{-S_{\text{string}}} \sim \sum_{\text{topologies}} e^{-2\lambda(1-g)} \int \mathcal{D}X \mathcal{D}g e^{-S_{\text{Poly}}}$$

For $e^\lambda \ll 1$, we have a good perturbative expansion in which we sum over all topologies. (In fact, it is an asymptotic expansion, just as in quantum field theory). It is standard to define the string coupling constant as

$$g_s = e^\lambda$$

After a conformal map, tree-level scattering corresponds to a worldsheet with the topology of a sphere: the amplitudes are proportional to $1/g_s^2$. One-loop scattering corresponds to toroidal worldsheets and, with our normalization, have no power of g_s . (Although, obviously, these are suppressed by g_s^2 relative to tree-level processes). The end

result is that the sum over worldsheets in (6.2) becomes a sum over Riemann surfaces of increasing genus, with vertex operators inserted for the initial and final states,



The Riemann surface of genus g is weighted by

$$(g_s^2)^{g-1}$$

While it may look like we've introduced a new parameter g_s into the theory and added the coupling (6.3) by hand, we will later see why this coupling is a necessary part of the theory and provide an interpretation for g_s .

Scattering Amplitudes

We now have all the information that we need to explain how to compute string scattering amplitudes. Suppose that we want to compute the S-matrix for m states: we will label them as Λ_i and assign them spacetime momenta p_i . Each has a corresponding vertex operator $V_{\Lambda_i}(p_i)$. The S-matrix element is then computed by evaluating the correlation function in the 2d conformal field theory, with insertions of the vertex operators.

$$\mathcal{A}^{(m)}(\Lambda_i, p_i) = \sum_{\text{topologies}} g_s^{-\chi} \frac{1}{\text{Vol}} \int \mathcal{D}X \mathcal{D}g e^{-S_{\text{Poly}}} \prod_{i=1}^m V_{\Lambda_i}(p_i)$$

This is a rather peculiar equation. We are interpreting the correlation functions of a two-dimensional theory as the S-matrix for a theory in $D = 26$ dimensions!

To properly compute the correlation function, we should introduce the b and c ghosts that we saw in the last chapter and treat them carefully. However, if we're only interested in tree-level amplitudes, then we can proceed naively and ignore the ghosts. The reason can be seen in the ghost action (5.5) where we see that the ghosts couple only to the worldsheet metric, not to the other worldsheet fields. This means that if our gauge fixing procedure fixes the worldsheet metric completely — which it does for worldsheets with the topology of a sphere — then we can forget about the ghosts. (At least, we can forget about them as soon as we've made sure that the Weyl anomaly cancels). However, as we'll explain in 6.4, for higher genus worldsheets, the gauge fixing does not fix the metric completely and there are residual dynamical modes of the metric, known as moduli, which couple the ghosts and matter fields. This is analogous to the statement in field theory that we only need to worry about ghosts running in loops.

6.2 Closed String Amplitudes at Tree Level

The tree-level scattering amplitude is given by the correlation function of the 2d theory, evaluated on the sphere,

$$\mathcal{A}^{(m)} = \frac{1}{g_s^2} \frac{1}{\text{Vol}} \int \mathcal{D}X \mathcal{D}g e^{-S_{\text{Poly}}} \prod_{i=1}^m V_{\Lambda_i}(p_i)$$

where $V_{\Lambda_i}(p_i)$ are the vertex operators associated to the states.

We want to integrate over all metrics on the sphere. At first glance that sounds rather daunting but, of course, we have the gauge symmetries of diffeomorphisms and Weyl transformations at our disposal. Any metric on the sphere is conformally equivalent to the flat metric on the plane. For example, the round metric on the sphere of radius R can be written as

$$ds^2 = \frac{4R^2}{(1 + |z|^2)^2} dz d\bar{z}$$

which is manifestly conformally equivalent to the plane, supplemented by the point at infinity. The conformal map from the sphere to the plane is the stereographic projection depicted in the diagram. The south pole of the sphere is mapped to the origin; the north pole is mapped to the point at infinity. Therefore, instead of integrating over all metrics, we may gauge fix diffeomorphisms and Weyl transformations to leave ourselves with the seemingly easier task of computing correlation functions on the plane.

6.2.1 Remnant Gauge Symmetry: $\text{SL}(2, \mathbb{C})$

There's a subtlety. And it's a subtlety that we've seen before: there is a residual gauge symmetry. It is the conformal group, arising from diffeomorphisms which can be undone by Weyl transformations. As we saw in Section 4, there are an infinite number of such conformal transformations. It looks like we have a whole lot of gauge fixing still to do.

However, global issues actually mean that there's less remnant gauge symmetry than you might think. In Section 4, we only looked at infinitesimal conformal transformations, generated by the Virasoro operators L_n , $n \in \mathbb{Z}$. We did not examine whether these transformations are well-defined and invertible over all of space. Let's take a

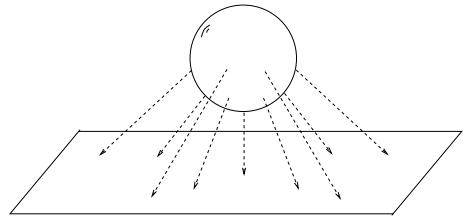


Figure 35:

look at this. Recall that the coordinate changes associated to L_n are generated by the vector fields (4.49),

$$l_n = z^{n+1} \partial_z$$

which result in the shift $\delta z = \epsilon z^{n+1}$. This is non-singular at $z = 0$ only for $n \geq -1$. If we restrict to smooth maps, that gets rid of half the transformations right away. But, since we're ultimately interested in the sphere, we now also need to worry about the point at $z = \infty$ which, in stereographic projection, is just the north pole of the sphere. To do this, it's useful to work with the coordinate

$$u = \frac{1}{z}$$

The generators of coordinate transformations for the u coordinate are

$$l_n = z^{n+1} \partial_z = \frac{1}{u^{n+1}} \frac{\partial u}{\partial z} \partial_u = -u^{1-n} \partial_u$$

which is non-singular at $u = 0$ only for $n \leq 1$.

Combining these two results, the only generators of the conformal group that are non-singular over the whole Riemann sphere are l_{-1} , l_0 and l_1 which act infinitesimally as

$$\begin{aligned} l_{-1} : z &\rightarrow z + \epsilon \\ l_0 : z &\rightarrow (1 + \epsilon)z \\ l_1 : z &\rightarrow (1 + \epsilon z)z \end{aligned}$$

The global version of these transformations is

$$\begin{aligned} l_{-1} : z &\rightarrow z + \alpha \\ l_0 : z &\rightarrow \lambda z \\ l_1 : z &\rightarrow \frac{z}{1 - \beta z} \end{aligned}$$

which can be combined to give the general transformation

$$z \rightarrow \frac{az + b}{cz + d} \tag{6.6}$$

with a, b, c and $d \in \mathbf{C}$. We have four complex parameters, but we've only got three transformations. What happened? Well, one transformation is fake because an overall

scaling of the parameters doesn't change z . By such a rescaling, we can always insist that the parameters obey

$$ad - bc = 1$$

The transformations (6.6) subject to this constraint have the group structure $SL(2; \mathbf{C})$, which is the group of 2×2 complex matrices with unit determinant. In fact, since the transformation is blind to a flip in sign of all the parameters, the actual group of global conformal transformations is $SL(2; \mathbf{C})/\mathbf{Z}_2$, which is sometimes written as $PSL(2; \mathbf{C})$. (This \mathbf{Z}_2 subtlety won't be important for us in what follows).

The remnant global transformations on the sphere are known as *conformal Killing vectors* and the group $SL(2; \mathbf{C})/\mathbf{Z}_2$ is the *conformal Killing group*. This group allows us to take any three points on the plane and move them to three other points of our choosing. We will shortly make use of this fact to gauge fix, but for now we leave the $SL(2; \mathbf{C})$ symmetry intact.

6.2.2 The Virasoro-Shapiro Amplitude

We will now compute the S-matrix for closed string tachyons. You might think that this is the least interesting thing to compute: after all, we're ultimately interested in the superstring which doesn't have tachyons. This is true, but it turns out that tachyon scattering is much simpler than everything else, mainly because we don't have a plethora of extra indices on the states to worry about. Moreover, the lessons that we will learn from tachyon scattering hold for the scattering of other states as well.

The m -point tachyon scattering amplitude is given by the flat space correlation function

$$\mathcal{A}^{(m)}(p_1, \dots, p_m) = \frac{1}{g_s^2} \frac{1}{\text{Vol}(SL(2; \mathbf{C}))} \int \mathcal{D}X e^{-S_{\text{Poly}}} \prod_{i=1}^m V(p_i)$$

where the tachyon vertex operator is given by,

$$V(p_i) = g_s \int d^2 z e^{ip_i \cdot X} \equiv g_s \int d^2 z \hat{V}(z, p_i) \quad (6.7)$$

Note that, in contrast to (5.8), we've added an appropriate normalization factor to the vertex operator. Heuristically, this reflects the fact that the operator is associated to the addition of a closed string mode. A rigorous derivation of this normalization can be found in Polchinski.

The amplitude can therefore be written as,

$$\mathcal{A}^{(m)}(p_1, \dots, p_m) = \frac{g_s^{m-2}}{\text{Vol}(SL(2; \mathbf{C}))} \int \prod_{i=1}^m d^2 z_i \langle \hat{V}(z_1, p_1) \dots \hat{V}(z_m, p_m) \rangle$$

where the expectation value $\langle \dots \rangle$ is computed using the gauge fixed Polyakov action. But the gauge fixed Polyakov action is simply a free theory and our correlation function is something eminently computable: a Gaussian integral,

$$\langle \hat{V}(z_1, p_1) \dots \hat{V}(z_m, p_m) \rangle = \int \mathcal{D}X \exp \left(-\frac{1}{2\pi\alpha'} \int d^2 z \partial X \cdot \bar{\partial} X \right) \exp \left(i \sum_{i=1}^m p_i \cdot X(z_i, \bar{z}_i) \right)$$

The normalization in front of the Polyakov action is now $1/2\pi\alpha'$ instead of $1/4\pi\alpha'$ because we're working with complex coordinates and we need to remember that $\partial_\alpha \partial^\alpha = 4\partial \bar{\partial}$ and $d^2 z = 2d^2 \sigma$.

The Gaussian Integral

We certainly know how to compute Gaussian integrals. Let's go slow. Consider the following general integral,

$$\int \mathcal{D}X \exp \left(\int d^2 z \frac{1}{2\pi\alpha'} X \cdot \partial \bar{\partial} X + i J \cdot X \right) \sim \exp \left(\frac{\pi\alpha'}{2} \int d^2 z d^2 z' J(z, \bar{z}) \frac{1}{\partial \bar{\partial}} J(z', \bar{z}') \right)$$

Here the \sim symbol reflects the fact that we've dropped a whole lot of irrelevant normalization terms, including $\det^{-1/2}(-\partial \bar{\partial})$. The inverse operator $1/\partial \bar{\partial}$ on the right-hand-side of this equation is shorthand for the propagator $G(z, z')$ which solves

$$\partial \bar{\partial} G(z, \bar{z}; z', \bar{z}') = \delta(z - z', \bar{z} - \bar{z}')$$

As we've seen several times before, in two dimensions this propagator is given by

$$G(z, \bar{z}; z', \bar{z}') = \frac{1}{2\pi} \ln |z - z'|^2$$

Back to the Scattering Amplitude

Comparing our scattering amplitude with this general expression, we need to take the source J to be

$$J(z, \bar{z}) = \sum_{i=1}^m p_i \delta(z - z_i, \bar{z} - \bar{z}_i)$$

Inserting this into the Gaussian integral gives us an expression for the amplitude

$$\mathcal{A}^{(m)} \sim \frac{g_s^{m-2}}{\text{Vol}(SL(2; \mathbf{C}))} \int \prod_{i=1}^m d^2 z_i \exp \left(\frac{\alpha'}{2} \sum_{j,l} p_j \cdot p_l \ln |z_j - z_l| \right)$$

The terms with $j = l$ seem to be problematic. In fact, they should just be left out. This follows from correctly implementing normal ordering and leaves us with

$$\mathcal{A}^{(m)} \sim \frac{g_s^{m-2}}{\text{Vol}(SL(2; \mathbf{C}))} \int \prod_{i=1}^m d^2 z_i \prod_{j < l} |z_j - z_l|^{\alpha' p_j \cdot p_l} \quad (6.8)$$

Actually, there's something that we missed. (Isn't there always!). We certainly expect scattering in flat space to obey momentum conservation, so there should be a $\delta^{(26)}(\sum_{i=1}^m p_i)$ in the amplitude. But where is it? We missed it because we were a little too quick in computing the Gaussian integral. The operator $\partial\bar{\partial}$ annihilates the zero mode, x^μ , in the mode expansion. This means that its inverse, $1/\partial\bar{\partial}$, is not well-defined. But it's easy to deal with this by treating the zero mode separately. The derivatives ∂^2 don't see x^μ , but the source J does. Integrating over the zero mode in the path integral gives us our delta function

$$\int dx \exp(i \sum_{i=1}^m p_i \cdot x) \sim \delta^{26}(\sum_{i=1}^m p_i)$$

So, our final result for the amplitude is

$$\mathcal{A}^{(m)} \sim \frac{g_s^{m-2}}{\text{Vol}(SL(2; \mathbf{C}))} \delta^{26}(\sum_i p_i) \int \prod_{i=1}^m d^2 z_i \prod_{j < l} |z_j - z_l|^{\alpha' p_j \cdot p_l} \quad (6.9)$$

The Four-Point Amplitude

We will compute only the four-point amplitude for two-to-two scattering of tachyons. The $\text{Vol}(SL(2; \mathbf{C}))$ factor is there to remind us that we still have a remnant gauge symmetry floating around. Let's now fix this. As we mentioned before, it provides enough freedom for us to take any three points on the plane and move them to any other three points. We will make use of this to set

$$z_1 = \infty \quad , \quad z_2 = 0 \quad , \quad z_3 = z \quad , \quad z_4 = 1$$

Inserting this into the amplitude (6.9), we find ourselves with just a single integral to evaluate,

$$\mathcal{A}^{(4)} \sim g_s^2 \delta^{26}(\sum_i p_i) \int d^2 z |z|^{\alpha' p_2 \cdot p_3} |1-z|^{\alpha' p_3 \cdot p_4} \quad (6.10)$$

(There is also an overall factor of $|z_1|^4$, but this just gets absorbed into an overall normalization constant). We still need to do the integral. It can be evaluated exactly in terms of gamma functions. We relegate the proof to Appendix 6.5, where we show that

$$\int d^2 z |z|^{2a-2} |1-z|^{2b-2} = \frac{2\pi\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(1-a)\Gamma(1-b)\Gamma(1-c)} \quad (6.11)$$

where $a + b + c = 1$.

Four-point scattering amplitudes are typically expressed in terms of Mandelstam variables. We choose p_1 and p_2 to be incoming momenta and p_3 and p_4 to be outgoing momenta, as shown in the figure. We then define

$$s = -(p_1 + p_2)^2 \quad , \quad t = -(p_1 + p_3)^2 \quad , \quad u = -(p_1 + p_4)^2$$

These obey

$$s + t + u = - \sum_i p_i^2 = \sum_i M_i^2 = -\frac{16}{\alpha'}$$

where, in the last equality, we've inserted the value of the tachyon mass (2.27). Writing the scattering amplitude (6.10) in terms of Mandelstam variables, we have our final answer

$$\mathcal{A}^{(4)} \sim g_s^2 \delta^{26}(\sum_i p_i) \frac{\Gamma(-1 - \alpha's/4)\Gamma(-1 - \alpha't/4)\Gamma(-1 - \alpha'u/4)}{\Gamma(2 + \alpha's/4)\Gamma(2 + \alpha't/4)\Gamma(2 + \alpha'u/4)} \quad (6.12)$$

This is the *Virasoro-Shapiro amplitude* governing tachyon scattering in the closed bosonic string.

Remarkably, the Virasoro-Shapiro amplitude was almost the first equation of string theory! (That honour actually goes to the Veneziano amplitude which is the analogous expression for open string tachyons and will be derived in Section 6.3.1). These amplitudes were written down long before people knew that they had anything to do with strings: they simply exhibited some interesting and surprising properties. It took several years of work to realise that they actually describe the scattering of strings. We will now start to tease apart the Virasoro-Shapiro amplitude to see some of the properties that got people hooked many years ago.

6.2.3 Lessons to Learn

So what's the physics lying behind the scattering amplitude (6.12)? Obviously it is symmetric in s , t and u . That is already surprising and we'll return to it shortly. But we'll start by fixing t and looking at the properties of the amplitude as we vary s .

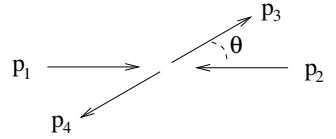
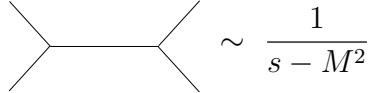


Figure 36:

The first thing to notice is that $\mathcal{A}^{(4)}$ has poles. Lots of poles. They come from the factor of $\Gamma(-1 - \alpha's/4)$ in the numerator. The first of these poles appears when

$$-1 - \frac{\alpha's}{4} = 0 \quad \Rightarrow \quad s = -\frac{4}{\alpha'}$$

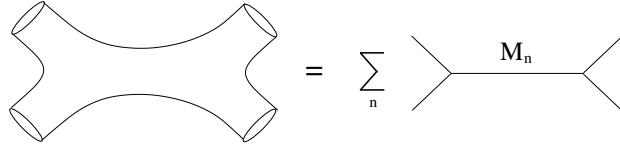
But that's the mass of the tachyon! It means that, for s close to $-4/\alpha'$, the amplitude has the form of a familiar scattering amplitude in quantum field theory with a cubic vertex,



$$\sim \frac{1}{s - M^2}$$

where M is the mass of the exchanged particle, in this case the tachyon.

Other poles in the amplitude occur at $s = 4(n - 1)/\alpha'$ with $n \in \mathbf{Z}^+$. This is precisely the mass formula for the higher states of the closed string. What we're learning is that the string amplitude is summing up an infinite number of tree-level field theory diagrams,



where the exchanged particles are all the different states of the free string.

In fact, there's more information about the spectrum of states hidden within these amplitudes. We can look at the residues of the poles at $s = 4(n - 1)/\alpha'$, for $n = 0, 1, \dots$. These residues are rather complicated functions of t , but the highest power of momentum that appears for each pole is

$$\mathcal{A}^{(4)} \sim \sum_{n=0}^{\infty} \frac{t^{2n}}{s - M_n^2} \tag{6.13}$$

The power of the momentum is telling us the highest spin of the particle states at level n . To see why this is, consider a field corresponding to a spin J particle. It has a whole bunch of Lorentz indices, $\chi_{\mu_1 \dots \mu_J}$. In a cubic interaction, each of these must be soaked up by derivatives. So we have J derivatives at each vertex, contributing powers of (momentum) 2J to the numerator of the Feynman diagram. Comparing with the string scattering amplitude, we see that the highest spin particle at level n has $J = 2n$. This is indeed the result that we saw from the canonical quantization of the string in Section 2.

Finally, the amplitude (6.12) has a property that is very different from amplitudes in field theory. Above, we framed our discussion by keeping t fixed and expanding in s . We could just have well done the opposite: fix s and look at poles in t . Now the string amplitude has the interpretation of an infinite number of t -channel scattering amplitudes, one for each state of the string

Usually in field theory, we sum up both s -channel and t -channel scattering amplitudes. Not so in string theory. The sum over an infinite number of s -channel amplitudes can be reinterpreted as an infinite sum of t -channel amplitudes. We don't include both: that would be overcounting. (Similar statements hold for u). The fact that the same amplitude can be written as a sum over s -channel poles *or* a sum over t -channel poles is sometimes referred to as "duality". (A much overused word). In the early days, before it was known that string theory was a theory of strings, the subject inherited its name from this duality property of amplitudes: it was called the *dual resonance model*.

High Energy Scattering

Let's use this amplitude to see what happens when we collide strings at high energies. There are different regimes that we could look at. The most illuminating is $s, t \rightarrow \infty$, with s/t held fixed. In this limit, all the exchanged momenta become large. It corresponds to high-energy scattering with the angle θ between incoming and outgoing particles kept fixed. To see this consider, for example, massless particles (our amplitude is really for tachyons, but the same considerations hold). We take the incoming and outgoing momenta to be

$$p_1 = \frac{\sqrt{s}}{2}(1, 1, 0, \dots), \quad p_2 = \frac{\sqrt{s}}{2}(1, -1, 0, \dots)$$

$$p_3 = \frac{\sqrt{s}}{2}(1, \cos \theta, \sin \theta, \dots), \quad p_4 = \frac{\sqrt{s}}{2}(1, -\cos \theta, -\sin \theta, \dots)$$

Then we see explicitly that $s \rightarrow \infty$ and $t \rightarrow \infty$ with the ratio s/t fixed also keeps the scattering angle θ fixed.

We can evaluate the scattering amplitude $\mathcal{A}^{(4)}$ in this limit by using $\Gamma(x) \sim \exp(x \ln x)$. We send $s \rightarrow \infty$ avoiding the poles. (We can achieve this by sending $s \rightarrow \infty$ in a slightly imaginary direction. Ultimately this is valid because all the higher string states are actually unstable in the interacting theory which will shift their poles off the real axis once taken into account). It is simple to check that the amplitude drops off exponentially quickly at high energies,

$$\mathcal{A}^{(4)} \sim g_s^2 \delta^{26} \left(\sum_i p_i \right) \exp \left(-\frac{\alpha'}{2} (s \ln s + t \ln t + u \ln u) \right) \quad \text{as } s \rightarrow \infty \quad (6.14)$$

The exponential fall-off seen in (6.14) is much faster than the amplitude of any field theory which, at best, fall off with power-law decay at high energies and, at worse, diverge. For example, consider the individual terms (6.13) corresponding to the amplitude for s -channel processes involving the exchange of particles with spin $2n$. We see that the exchange of a spin 2 particle results in a divergence in this limit. This is reflecting something you already know about gravity: the dimensionless coupling is $G_N E^2$ (in four-dimensions) which becomes large for large energies. The exchange of higher spin particles gives rise to even worse divergences. If we were to truncate the infinite sum (6.13) at any finite n , the whole thing would diverge. But infinite sums can do things that finite sums can't and the final behaviour of the amplitude (6.14) is much softer than any of the individual terms. The infinite number of particles in string theory conspire to render finite any divergence arising from an individual particle species.

Phrased in terms of the s -channel exchange of particles, the high-energy behaviour of string theory seems somewhat miraculous. But there is another viewpoint where it's all very obvious. The power-law behaviour of scattering amplitudes is characteristic of point-like charges. But, of course, the string isn't a point-like object. It is extended and fuzzy at length scales comparable to $\sqrt{\alpha'}$. This is the reason the amplitude has such soft high-energy behaviour. Indeed, this idea that smooth extended objects give rise to scattering amplitudes that decay exponentially at high energies is something that you've seen before in non-relativistic quantum mechanics. Consider, for example, the scattering of a particle off a Gaussian potential. In the Born approximation, the differential cross-section is just given by the Fourier transform which is again a Gaussian, now decaying exponentially for large momentum.

It's often said that theories of quantum gravity should have a “minimum length”, sometimes taken to be the Planck scale. This is roughly true in string theory, although not in any crude simple manner. Rather, the minimum length reveals itself in different

ways depending on which question is being asked. The above discussion highlights one example of this: strings can't probe distance scales shorter than $l_s = \sqrt{\alpha'}$ simply because they are themselves fuzzy at this scale. It turns out that D-branes are much better probes of sub-stringy physics and provide a different view on the short distance structure of spacetime. We will also see another manifestation of the minimal length scale of string theory in Section 8.3.

Graviton Scattering

Although we've derived the result (6.14) for tachyons, all tree-level amplitudes have this soft fall-off at high-energies. Most notably, this includes graviton scattering. As we noted above, this is in sharp contrast to general relativity for which tree-level scattering amplitudes diverge at high-energies. This is the first place to see that UV problems of general relativity might have a good chance of being cured in string theory.

Using the techniques described in this section, one can compute m -point tree-level amplitudes for graviton scattering. If we restrict attention to low-energies (i.e. much smaller than $1/\sqrt{\alpha'}$), one can show that these coincide with the amplitudes derived from the Einstein-Hilbert action in $D = 26$ dimensions

$$S = \frac{1}{2\kappa^2} \int d^{26}X \sqrt{-G} \mathcal{R}$$

where \mathcal{R} is the $D = 26$ Ricci scalar (not to be confused with the worldsheet Ricci scalar which we call R). The gravitational coupling, κ^2 is related to Newton's constant in 26 dimensions. It plays no role for pure gravity, but is important when we couple to matter. We'll see shortly that it's given by

$$\kappa^2 \approx g_s^2(\alpha')^{12}$$

We won't explicitly compute graviton scattering amplitudes in this course, partly because they're fairly messy and partly because building up the Einstein-Hilbert action from m -particle scattering is hardly the best way to look at general relativity. Instead, we shall derive the Einstein-Hilbert action in a much better fashion in Section 7.

6.3 Open String Scattering

So far our discussion has been entirely about closed strings. There is a very similar story for open strings. We again compute S-matrix elements. Conformal symmetry now maps tree-level scattering to the disc, with vertex operators inserted on the boundary of the disc.

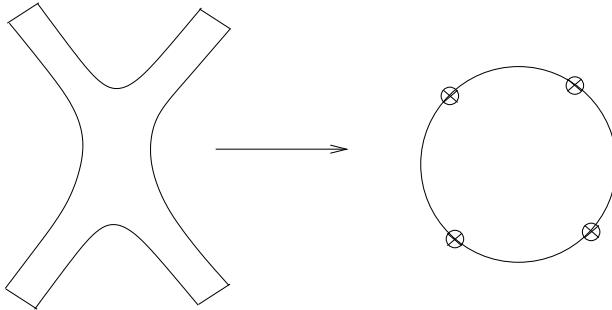


Figure 37: The conformal map from the open string worldsheet to the disc.

For the open string, the string coupling constant that we add to the Polyakov action requires the addition of a boundary term to make it well defined,

$$\chi = \frac{1}{4\pi} \int_{\mathcal{M}} d^2\sigma \sqrt{g} R + \frac{1}{2\pi} \int_{\partial\mathcal{M}} ds k \quad (6.15)$$

where k is the geodesic curvature of the boundary. To define it, we introduce two unit vectors on the worldsheet: t^α is tangential to the boundary, while n^α is normal and points outward from the boundary. The geodesic curvature is defined as

$$k = -t^\alpha n_\beta \nabla_\alpha t^\beta$$

Boundary terms of the type seen in (6.15) are also needed in general relativity for manifolds with boundaries: in that context, they are referred to as Gibbons-Hawking terms.

The Gauss-Bonnet theorem has an extension to surfaces with boundary. For surfaces with h handles and b boundaries, the Euler character is given by

$$\chi = 2 - 2h - b$$

Some examples are shown in Figure 38. The expansion for open-string scattering consists of adding consecutive boundaries to the worldsheet. The disc is weighted by $1/g_s$; the annulus has no factor of g_s and so on. We see that the open string coupling is related to the closed string coupling by

$$g_{\text{open}}^2 = g_s \quad (6.16)$$

One of the key steps in computing closed string scattering amplitudes was the implementation of the conformal Killing group, which was defined as the surviving gauge

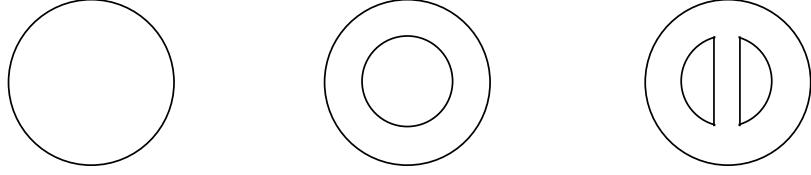


Figure 38: Riemann surfaces with boundary with $\chi = 1, 0$ and -1 .

symmetry with a global action on the sphere. For the open string, there is again a residual gauge symmetry. If we think in terms of the upper-half plane, the boundary is $\text{Im}z = 0$. The conformal Killing group is composed of transformations

$$z \rightarrow \frac{az + b}{cz + d}$$

again with the requirement that $ad - bc = 1$. This time there is one further condition: the boundary $\text{Im}z = 0$ must be mapped onto itself. This requires $a, b, c, d \in \mathbf{R}$. The resulting conformal Killing group is $SL(2; \mathbf{R})/\mathbf{Z}_2$.

6.3.1 The Veneziano Amplitude

Since vertex operators now live on the boundary, they have a fixed ordering. In computing a scattering amplitude, we must sum over all orderings. Let's look again at the 4-point amplitude for tachyon scattering. The vertex operator is

$$V(p_i) = \sqrt{g_s} \int dx e^{ip_i \cdot X}$$

where the integral $\int dx$ is now over the boundary and $p^2 = 1/\alpha'$ is the on-shell condition for an open-string tachyon. The normalization $\sqrt{g_s}$ is that appropriate for the insertion of an open-string mode, reflecting (6.16).

Going through the same steps as for the closed string, we find that the amplitude is given by

$$\mathcal{A}^{(4)} \sim \frac{g_s}{\text{Vol}(SL(2; \mathbf{R}))} \delta^{26}(\sum_i p_i) \int \prod_{i=1}^4 dx_i \prod_{j < l} |x_j - x_l|^{2\alpha' p_j \cdot p_l} \quad (6.17)$$

Note that there's a factor of 2 in the exponent, differing from the closed string expression (6.8). This comes about because the boundary propagator (4.57) has an extra factor of 2 due to the image charge.

We now use the $SL(2; \mathbf{R})$ residual gauge symmetry to fix three points on the boundary. We choose a particular ordering and set $x_1 = 0$, $x_2 = x$, $x_3 = 1$ and $x_4 \rightarrow \infty$. The only free insertion point is $x_2 = x$ but, because of the restriction of operator ordering, this must lie in the interval $x \in [0, 1]$. The interesting part of the integral is then given by

$$\mathcal{A}^{(4)} \sim g_s \int_0^1 dx |x|^{2\alpha' p_1 \cdot p_2} |1-x|^{2\alpha' p_2 \cdot p_3}$$

This integral is well known: as shown in Appendix 6.5, it is the Euler beta function

$$B(a, b) = \int_0^1 dx x^{a-1} (1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

After summing over the different orderings of vertex operators, the end result for the amplitude for open string tachyon scattering is,

$$\mathcal{A}^{(4)} \sim g_s [B(-\alpha's - 1, -\alpha't - 1) + B(-\alpha's - 1, -\alpha'u - 1) + B(-\alpha't - 1, -\alpha'u - 1)]$$

This is the famous *Veneziano Amplitude*, first postulated in 1968 to capture some observed features of the strong interactions. This was before the advent of QCD and before it was realised that the amplitude arises from a string.

The open string scattering amplitude contains the same features that we saw for the closed string. For example, it has poles at

$$s = \frac{n-1}{\alpha'} \quad n = 0, 1, 2, \dots$$

which we recognize as the spectrum of the open string.

6.3.2 The Tension of D-Branes

Recall that we introduced D-branes as surfaces in space on which strings can end. At the time, I promised that we would eventually discover that these D-branes are dynamical objects in their own right. We'll look at this more closely in the next section, but for now we can do a simple computation to determine the tension of D-branes.

The tension T_p of a Dp -brane is defined as the energy per spatial volume. It has dimension $[T_p] = p+1$. The tension is telling us the magnitude of the coupling between the brane and gravity. Or, in our new language, the strength of the interaction between a closed string state and an open string. The simplest such diagram is shown in the figure, with a graviton vertex operator inserted. Although we won't compute this

diagram completely, we can figure out its most important property just by looking at it: it has the topology of a disc, so is proportional to $1/g_s$. Adding powers of α' to get the dimension right, the tension of a D p -brane must scale as

$$T_p \sim \frac{1}{l_s^{p+1}} \frac{1}{g_s} \quad (6.18)$$

where the string length is defined as $l_s = \sqrt{\alpha'}$. The $1/g_s$ scaling of the tension is one of the key characteristic features of a D-brane.

I should confess that there's a lot swept under the carpet in the above discussion, not least the question of the correct normalization of the vertex operators and the difference between the string frame and the Einstein frame (which we will discuss shortly). Nonetheless, the end result (6.18) is correct. For a fuller discussion, see Section 8.7 of Polchinski.

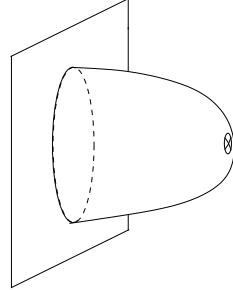


Figure 39:

6.4 One-Loop Amplitudes

We now return to the closed string to discuss one-loop effects. As we saw above, this corresponds to a worldsheet with the topology of a torus. We need to integrate over all metrics on the torus.

For tree-level processes, we used diffeomorphisms and Weyl transformations to map an arbitrary metric on the sphere to the flat metric on the plane. This time, we use these transformations to map an arbitrary metric on the torus to the flat metric on the torus. But there's a new subtlety that arises: not all flat metrics on the torus are equivalent.

6.4.1 The Moduli Space of the Torus

Let's spell out what we mean by this. We can construct a torus by identifying a region in the complex z -plane as shown in the figure. In general, this identification depends on a single complex parameter, $\tau \in \mathbf{C}$.

$$z \equiv z + 2\pi \quad \text{and} \quad z \equiv z + 2\pi\tau$$

Do not confuse τ with the Minkowski worldsheet time: we left that behind way back in Section 3. Everything here is Euclidean worldsheet and τ is just a parameter telling us how skewed the torus is. The flat metric on the torus is now simply

$$ds^2 = dz d\bar{z}$$

subject to the identifications above.

A general metric on a torus can always be transformed to a flat metric for some value of τ . But the question that interests us is whether two tori, parameterized by different τ , are conformally equivalent. In general, the answer is no. The space of conformally inequivalent tori, parameterized by τ , is called the *moduli space* \mathcal{M} .

However, there are some values of τ that do correspond to the same torus. In particular, there are a couple of obvious ways in which we can change τ without changing the torus. They go by the names of the S and T transformations:

- $T : \tau \rightarrow \tau + 1$: This clearly gives rise to the same torus, because the identification is now

$$z \equiv z + 2\pi \quad \text{and} \quad z \equiv z + 2\pi(\tau + 1) \equiv z + 2\pi\tau$$

- $S : \tau \rightarrow -1/\tau$: This simply flips the sides of the torus. For example, if $\tau = ia$ is purely imaginary, then this transformation maps $\tau \rightarrow i/a$, which can then be undone by a scaling.

It turns out that these two changes S and T are the only ones that keep the torus intact. They are sometimes called *modular transformations*. A general modular transformations is constructed from combinations of S and T and takes the form,

$$\tau \rightarrow \frac{a\tau + b}{c\tau + d} \quad \text{with } ad - bc = 1 \tag{6.19}$$

where a, b, c and $d \in \mathbf{Z}$. This is the group $SL(2, \mathbf{Z})$. (In fact, we have our usual \mathbf{Z}_2 identification and the group is actually $PSL(2, \mathbf{Z}) = SL(2; \mathbf{Z})/\mathbf{Z}_2$). The moduli space \mathcal{M} of the torus is given by

$$\mathcal{M} \cong \mathbf{C}/SL(2; \mathbf{Z})$$

What does this space look like? Using $T : \tau \rightarrow \tau + 1$, we can always shift τ until it lies within the interval

$$\operatorname{Re} \tau \in [-\frac{1}{2}, +\frac{1}{2}]$$

where the edges of the interval are identified. Meanwhile, $S : \tau \rightarrow -1/\tau$ inverts the

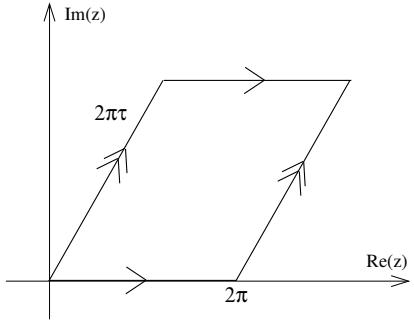


Figure 40:

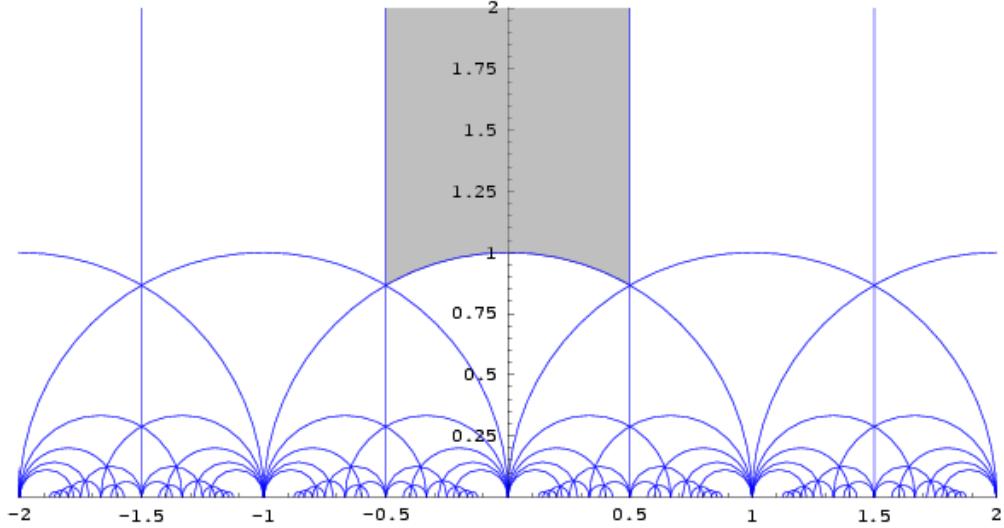


Figure 41: The fundamental domain.

modulus $|\tau|$, so we can use this to map a point inside the circle $|\tau| < 1$ to a point outside $|\tau| > 1$. One can show that by successive combinations of S and T , it is possible to map any point to lie within the shaded region shown in the figure, defined by

$$|\tau| \geq 1 \quad \text{and} \quad \operatorname{Re} \tau \in [-\frac{1}{2}, +\frac{1}{2}]$$

This is referred to as the *fundamental domain* of $SL(2; \mathbf{Z})$.

We could have just as easily chosen one of the other fundamental domains shown in the figure. But the shaded region is the standard one.

Integrating over the Moduli Space

In string theory we're invited to sum over all metrics. After gauge fixing diffeomorphisms and Weyl invariance, we still need to integrate over all inequivalent tori. In other words, we integrate over the fundamental domain. The $SL(2; \mathbf{Z})$ invariant measure over the fundamental domain is

$$\int \frac{d^2\tau}{(\operatorname{Im} \tau)^2}$$

To see that this is $SL(2; \mathbf{Z})$ invariant, note that under a general transformation of the form (6.19) we have

$$d^2\tau \rightarrow \frac{d^2\tau}{|c\tau + d|^4} \quad \text{and} \quad \operatorname{Im} \tau \rightarrow \frac{\operatorname{Im} \tau}{|c\tau + d|^2}$$

There's some physics lurking within these rather mathematical statements. The integration over the fundamental domain in string theory is analogous to the loop integral over momentum in quantum field theory. Consider the square tori defined by $\text{Re } \tau = 0$. The tori with $\text{Im } \tau \rightarrow \infty$ are squashed and chubby. They correspond to the infra-red region of loop momenta in a Feynman diagram. Those with $\text{Im } \tau \rightarrow 0$ are long and thin. Those correspond to the ultra-violet limit of loop momenta in a Feynman diagram. Yet, as we have seen, we should not integrate over these UV regions of the loop since the fundamental domain does not stretch down that far. Or, more precisely, the thin tori are mapped to chubby tori. This corresponds to the fact that any putative UV divergence of string theory can always be reinterpreted as an IR divergence. This is the second manifestation of the well-behaved UV nature of string theory. We will see this more explicitly in the example of Section 6.4.2.

Finally, when computing a loop amplitude in string theory, we still need to worry about the residual gauge symmetry that is left unfixed after the map to the flat torus. In the case of tree-level amplitudes on the sphere, this residual gauge symmetry was due to the conformal Killing group $SL(2; \mathbf{C})$. For the torus, the conformal Killing group is generated by the obvious generators ∂_z and $\bar{\partial}_{\bar{z}}$. It is $U(1) \times U(1)$.

Higher Genus Surfaces

The moduli space \mathcal{M}_g of the Riemann surface of genus $g > 1$ can be shown to have dimension,

$$\dim \mathcal{M}_g = 3g - 3$$

There are no conformal Killing vectors when $g > 1$. These facts can be demonstrated as an application of the Riemann-Roch theorem. For more details, see section 5.2 of Polchinski, or sections 3.3 and 8.2 of Green, Schwarz and Witten.

6.4.2 The One-Loop Partition Function

We won't compute any one-loop scattering amplitudes in string theory. Instead, we will look at something a little simpler: the one-loop vacuum to vacuum amplitude. A Euclidean worldsheet with periodic time has the interpretation of a finite temperature partition function for the theory defined on a cylinder. In $D = 26$ dimensional spacetime, it is related to the cosmological constant in bosonic string theory.

Consider firstly the partition function of a theory on a square torus, with $\text{Re } \tau = 0$. Compactifying Euclidean time, with period $(\text{Im } \tau)$ is equivalent to putting the theory at temperature $T = 1/(\text{Im } \tau)$,

$$Z[\tau] = \text{Tr } e^{-2\pi(\text{Im } \tau)H}$$

where the Tr is over all states in the theory. For any CFT defined on a cylinder, the Hamiltonian given by

$$H = L_0 + \tilde{L}_0 - \frac{c + \tilde{c}}{24}$$

where the final term is the Casimir energy computed in Section 4.4.1.

What then is the interpretation of the vacuum amplitude computed on a torus with $\text{Re } \tau \neq 0$? From the diagram, we see that the effect of such a skewed torus is to translate a given point around the cylinder by $\text{Re } \tau$. But we know which operator implements such a translation: it is $\exp(2\pi i(\text{Re } \tau)P)$, where P is the momentum operator on the cylinder. After the map to the plane, this becomes the rotation operator

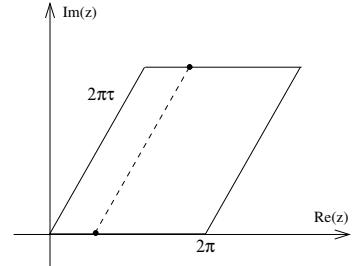


Figure 42:

$$P = L_0 - \tilde{L}_0$$

So the vacuum amplitude on the torus has the interpretation of the sum over all states in the theory, weighted by

$$Z[\tau] = \text{Tr } e^{-2\pi(\text{Im } \tau)(L_0 + \tilde{L}_0)} e^{-2\pi i(\text{Re } \tau)(L_0 - \tilde{L}_0)} e^{2\pi(\text{Im } \tau)(c + \tilde{c})/24}$$

We define

$$q = e^{2\pi i \tau} \quad , \quad \bar{q} = e^{-2\pi i \bar{\tau}}$$

The partition function can then be written in slick notation as

$$Z[\tau] = \text{Tr } q^{L_0 - c/24} \bar{q}^{\tilde{L}_0 - \tilde{c}/24}$$

Let's compute this for the free string. We know that each scalar field X decomposes into a zero mode and an infinite number harmonic oscillator modes α_{-n} which create states of energy n . We'll deal with the zero mode shortly but, for now, we focus on the oscillators. Acting d times with the operator α_{-n} creates states with energy dn . This gives a contribution to $\text{Tr } q^{L_0}$ of the form

$$\sum_{d=0}^{\infty} q^{nd} = \frac{1}{1 - q^n}$$

But the Fock space of a single scalar field is built by acting with oscillator modes $n \in \mathbf{Z}^+$. Including the central charge, $c = 1$, the contribution from the oscillator modes of a single scalar field is therefore

$$\text{Tr } q^{L_0 - c/24} = \frac{1}{q^{1/24}} \prod_{n=1}^{\infty} \frac{1}{1 - q^n}$$

There is a similar expression from the $\bar{q}^{\tilde{L}_0 - \tilde{c}/24}$ sector. We're still left with the contribution from the zero mode p of the scalar field. The contribution to the energy H of the state on the worldsheet is

$$\frac{1}{4\pi\alpha'} \int d\sigma (\alpha' p)^2 = \frac{1}{2} \alpha' p^2$$

The trace in the partition function requires us to sum over all states, which gives

$$\int \frac{dp}{2\pi} e^{-\pi\alpha'(\text{Im }\tau)p^2} \sim \frac{1}{\sqrt{\alpha'\text{Im }\tau}}$$

So, including both the zero mode and oscillators, we get the partition function for a single free scalar field,

$$Z_{\text{scalar}}[\tau] \sim \frac{1}{\sqrt{\alpha'\text{Im }\tau}} \frac{1}{(q\bar{q})^{1/24}} \prod_{n=1}^{\infty} \frac{1}{1 - q^n} \prod_{n=1}^{\infty} \frac{1}{1 - \bar{q}^n} \quad (6.20)$$

where I haven't been careful to keep track of constant factors.

To build the string partition function, we should really work in covariant quantization and include the ghost fields. Here we'll cheat and work in lightcone gauge. This is dodgy because, if we do it honestly, much of the physics gets pushed to the $p^+ = 0$ limit of the lightcone momentum where the gauge choice breaks down. So instead we'll do it dishonestly.

In lightcone gauge, we have 24 oscillator modes. But we have 26 zero modes. (You may worry that we still have to impose level matching...this is the dishonest part of the calculation. We'll see partly where it comes from shortly). Finally, there's a couple of extra steps. We need to divide by the volume of the conformal Killing group. This is just $U(1) \times U(1)$, acting by translations along the cycles of the torus. The volume is just $\text{Vol} = 4\pi^2 \text{Im } \tau$. Finally, we also need to integrate over the moduli space of the torus. Our final result, neglecting all constant factors, is

$$Z_{\text{string}} = \int d^2\tau \frac{1}{(\text{Im } \tau)} \frac{1}{(\alpha'\text{Im } \tau)^{13}} \frac{1}{q\bar{q}} \left(\prod_{n=1}^{\infty} \frac{1}{1 - q^n} \right)^{24} \left(\prod_{n=1}^{\infty} \frac{1}{1 - \bar{q}^n} \right)^{24} \quad (6.21)$$

Modular Invariance

The function appearing in the partition function for the scalar field has a name: it is the inverse of the Dedekind eta function

$$\eta(q) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n)$$

It was studied in the 1800s by mathematicians interested in the properties of functions under modular transformations $T : \tau \rightarrow \tau + 1$ and $S : \tau \rightarrow -1/\tau$. The eta-function satisfies the identities

$$\eta(\tau + 1) = e^{2\pi i/24} \eta(\tau) \quad \text{and} \quad \eta(-1/\tau) = \sqrt{-i\tau} \eta(\tau)$$

These two statements ensure that the scalar partition function (6.20) is a modular invariant function. Of course, that kinda had to be true: it follows from the underlying physics.

Written in terms of η , the string partition function (6.21) takes the form

$$Z_{\text{string}} = \int \frac{d^2\tau}{(\text{Im } \tau)^2} \left(\frac{1}{\sqrt{\text{Im } \tau}} \frac{1}{\eta(q)} \frac{1}{\bar{\eta}(\bar{q})} \right)^{24}$$

Both the measure and the integrand, are individually modular invariant.

6.4.3 Interpreting the String Partition Function

It's probably not immediately obvious what the string partition function (6.21) is telling us. Let's spend some time trying to understand it in terms of some simpler concepts.

We know that the free string describes an infinite number of particles with mass $m_n^2 = 4(n-1)/\alpha'$, $n = 0, 1, \dots$. The string partition function should just be a sum over vacuum loops of each of these particles. We'll now show that it almost has this interpretation.

Firstly, let's figure out what the contribution from a single particle would be? We'll consider a free massive scalar field ϕ in D dimensions. The partition function is given by,

$$\begin{aligned} Z &= \int \mathcal{D}\phi \exp \left(-\frac{1}{2} \int d^D x \phi (-\partial^2 + m^2) \phi \right) \\ &\sim \det^{-1/2} (-\partial^2 + m^2) \\ &= \exp \left(\frac{1}{2} \int \frac{d^D p}{(2\pi)^D} \ln(p^2 + m^2) \right) \end{aligned}$$

This is the partition function of a field theory. It contains vacuum loops for all numbers of particles. To compare to the string partition function, we want the vacuum amplitude for just a single particle. But that's easy to extract. We write the field theory partition function as,

$$Z = \exp(Z_1) = \sum_{n=0}^{\infty} \frac{Z_1^n}{n!}$$

Each term in the sum corresponds to n particles propagating in a vacuum loop, with the $n!$ factor taking care of Bosonic statistics. So the vacuum amplitude for a single, free massive particle is simply

$$Z_1 = \frac{1}{2} \int \frac{d^D p}{(2\pi)^D} \ln(p^2 + m^2)$$

Clearly this diverges in the UV range of the integral, $p \rightarrow \infty$. There's a nice way to rewrite this integral using something known as Schwinger parameterization. We make use of the identity

$$\int_0^\infty dl e^{-xl} = \frac{1}{x} \quad \Rightarrow \quad \int_0^\infty dl \frac{e^{-xl}}{l} = -\ln x$$

We then write the single particle partition function as

$$Z_1 = \int \frac{d^D p}{(2\pi)^D} \int_0^\infty \frac{dl}{2l} e^{-(p^2+m^2)l} \tag{6.22}$$

It's worth mentioning that there's another way to see that this is the single particle partition function that is a little closer in spirit to the method we used in string theory. We could start with the einbein form of the relativistic particle action (1.8). After fixing the gauge to $e = 1$, the exponent in (6.22) is the energy of the particle traversing a loop of length l . The integration measure dl/l sums over all possible sizes of loops.

We can happily perform the $\int d^D p$ integral in (6.22). Ignoring numerical factors, we have

$$Z_1 = \int_0^\infty dl \frac{1}{l^{1+D/2}} e^{-m^2 l} \tag{6.23}$$

Note that the UV divergence as $p \rightarrow \infty$ has metamorphosised into a divergence associated to small loops as $l \rightarrow 0$.

Equation (6.23) gives the answer for a single particle of mass m . In string theory, we expect contributions from an infinite number of species of particles of mass m_n . Specializing to $D = 26$, we expect the partition function to be

$$Z = \int_0^\infty dl \frac{1}{l^{14}} \sum_{n=0}^{\infty} e^{-m_n^2 l}$$

But we know that the mass spectrum of the free string: it is given in terms of the L_0 and \tilde{L}_0 operators by

$$m^2 = \frac{4}{\alpha'}(L_0 - 1) = \frac{4}{\alpha'}(\tilde{L}_0 - 1) = \frac{2}{\alpha'}(L_0 + \tilde{L}_0 - 2)$$

subject to the constraint of level matching, $L_0 = \tilde{L}_0$. It's easy to impose level matching: we simply throw in a Kronecker delta in its integral representation,

$$\frac{1}{2\pi} \int_{-1/2}^{+1/2} ds e^{2\pi i s(L_0 - \tilde{L}_0)} = \delta_{L_0, \tilde{L}_0} \quad (6.24)$$

Replacing the sum over species, with the trace over the spectrum of states subject to level matching, the partition function becomes,

$$Z = \int_0^\infty dl \frac{1}{l^{14}} \int_{-1/2}^{+1/2} ds \text{Tr} e^{2\pi i s(L_0 - \tilde{L}_0)} e^{-2(L_0 + \tilde{L}_0 - 2)l/\alpha'} \quad (6.25)$$

We again use the definition $q = \exp(2\pi i \tau)$, but this time the complex parameter τ is a combination of the length of the loop l and the auxiliary variable that we introduced to impose level matching,

$$\tau = s + \frac{2li}{\alpha'}$$

The trace over the spectrum of the string once gives the eta-functions, just as it did before. We're left with the result for the partition function,

$$Z_{\text{string}} = \int \frac{d^2\tau}{(\text{Im } \tau)^2} \left(\frac{1}{\sqrt{\text{Im } \tau}} \frac{1}{\eta(q)} \frac{1}{\bar{\eta}(\bar{q})} \right)^{24}$$

But this is exactly the same expression that we saw before. With a difference! In fact, the difference is hidden in the notation: it is the range of integration for $d^2\tau$ which can be found in the original expressions (6.23) and (6.24). $\text{Re } \tau$ runs over the same interval $[-\frac{1}{2}, +\frac{1}{2}]$ that we saw in string theory. As is clear from this discussion, it is this integral which implements level matching. The difference comes in the range of $\text{Im } \tau$ which, in this naive analysis, runs over $[0, \infty)$. This is in stark contrast to string theory where we only integrate over the fundamental domain.

This highlights our previous statement: the potential UV divergences in field theory are encountered in the region $\text{Im } \tau \sim l \rightarrow 0$. In the above analysis, this corresponds to particles traversing small loops. But this region is simply absent in the correct string theory computation. It is mapped, by modular invariance, to the infra-red region of large loops.

It is often said that in the $g_s \rightarrow 0$ limit string theory becomes a theory of an infinite number of free particles. This is true of the spectrum. But this calculation shows that it's not really true when we compute loops because the modular invariance means that we integrate over a different range of momenta in string theory than in a naive field theory approach.

So what happens in the infra-red region of our partition function? The easiest place to see it is in the $l \rightarrow \infty$ limit of the integral (6.25). We see that the integral is dominated by the lightest state which, for the bosonic string is the tachyon. This has $m^2 = -4/\alpha'$, or $(L_0 + \tilde{L}_0 - 2) = -2$. This gives a contribution to the partition function of,

$$\int^\infty \frac{dl}{l^{14}} e^{+4l/\alpha'}$$

which clearly diverges. This IR divergence of the one-loop partition function is another manifestation of tachyonic trouble. In the superstring, there is no tachyon and the IR region is well-behaved.

6.4.4 So is String Theory Finite?

The honest answer is that we don't know. The UV finiteness that we saw above holds for all one-loop amplitudes. This means, in particular, that we have a one-loop finite theory of gravity interacting with matter in higher dimensions. This is already remarkable.

There is more good news: One can show that UV finiteness continues to hold at the two-loops. And, for the superstring, state-of-the-art techniques using the “pure-spinor” formalism show that certain objects remain finite up to five-loops. Moreover, the exponential suppression (6.14) that we saw when all momentum exchanges are large continues to hold for all amplitudes.

However, no general statement of finiteness has been proven. The danger lurks in the singular points in the integration over Riemann surfaces of genus 3 and higher.

6.4.5 Beyond Perturbation Theory?

From the discussion in this section, it should be clear that string perturbation theory is entirely analogous to the Feynman diagram expansion in field theory. Just as in field theory, one can show that the expansion in g_s is asymptotic. This means that the series does not converge, but we can nonetheless make sense of it.

However, we know that there are many phenomena in quantum field theory that aren't captured by Feynman diagrams. These include confinement in the strongly coupled regime and instantons and solitons in the weakly coupled regime. Does this mean that we are missing similarly interesting phenomena in string theory? The answer is almost certainly yes! In this section, I'll very briefly allude to a couple of more advanced topics which allow us to go beyond the perturbative expansion in string theory. The goal is not really to teach you these things, but merely to familiarize you with some words.

One way to proceed is to keep quantum field theory as our guide and try to build a non-perturbative definition of string theory in terms of a path integral. We've already seen that the Polyakov path integral over worldsheets is equivalent to Feynman diagrams. So we need to go one step further. What does this mean? Recall that in QFT, a field creates a particle. In string theory, we are now looking for a field which creates a loop of string. We should have a different field for each configuration of the string. In other words, our field should itself be a function of a function: $\Phi(X^\mu(\sigma))$. Needless to say, this is quite a complicated object. If we were brave, we could then consider the path integral for this field,

$$Z = \int \mathcal{D}\Phi e^{iS[\Phi(X(\sigma))]}$$

for some suitable action $S[\Phi]$. The idea is that this path integral should reproduce the perturbative string expansion and, furthermore, defines a non-perturbative completion of the theory. This line of ideas is known as *string field theory*. It should be clear that this is one step further in the development: particles \rightarrow fields \rightarrow string fields. Or, in more historical language, if field theory is “second quantization”, then string field theory is “third quantization”.

String field theory has been fairly successful for the open string and some interesting non-perturbative results have been obtained in this manner. However, for the closed string this approach has been much less useful. It is usually thought that there are deep reasons behind the failure of closed string field theory, related to issues that we mentioned at the beginning of this section: there are no off-shell quantities in a theory

of gravity. Moreover, we mentioned in Section 4 that a theory of interacting open strings necessarily includes closed strings, so somehow the open string field theory should already contain gravity and closed strings. Quite how this comes about is still poorly understood.

There are other ways to get a handle on non-perturbative aspects of string theory using the low-energy effective action (we will describe what the “low-energy effective action” is in the next section). Typically these techniques rely on supersymmetry to provide a window into the strongly coupled regime and so work only for the superstring. These methods have been extremely successful and any course on superstring theory would be devoted to explaining various aspects of such as dualities and M-theory.

Finally, in asymptotically AdS spacetimes, the AdS/CFT correspondence gives a non-perturbative definition of string theory and quantum gravity in the bulk in terms of Yang-Mills theory, or something similar, on the boundary. In some sense, the boundary field theory is a “string field theory”.

6.5 Appendix: Games with Integrals and Gamma Functions

The gamma function is defined by the integral representation

$$\Gamma(z) = \int_0^\infty dt t^{z-1} e^{-t} \quad (6.26)$$

which converges if $\text{Re}z > 0$. It has a unique analytic expression to the whole z -plane. The absolute value of the gamma function over the z -plane is shown in the figure.

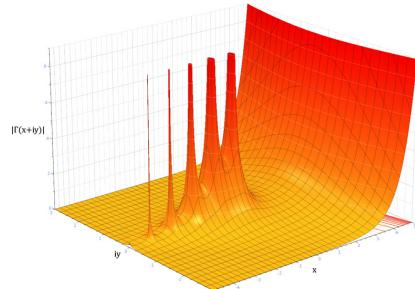


Figure 43:

The gamma function has a couple of important properties. Firstly, it can be thought of as the analytic continuation of the factorial function for positive integers, meaning

$$\Gamma(n) = (n-1)! \quad n \in \mathbf{Z}^+$$

Secondly, $\Gamma(z)$ has poles at non-positive integers. More precisely when $z \approx -n$, with $n = 0, 1, \dots$, there is the expansion

$$\Gamma(z) \approx \frac{1}{z+n} \frac{(-1)^n}{n!}$$

The Euler Beta Function

The Euler beta function is defined for $x, y \in \mathbf{C}$ by

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

It has the integral representation

$$B(x, y) = \int_0^1 dt t^{x-1} (1-t)^{y-1} \quad (6.27)$$

Let's prove this statement. We start by looking at

$$\Gamma(x)\Gamma(y) = \int_0^\infty du \int_0^\infty dv e^{-u} u^{x-1} e^{-v} v^{y-1}$$

We write $u = a^2$ and $v = b^2$ so the integral becomes

$$\begin{aligned} \Gamma(x)\Gamma(y) &= 4 \int_0^\infty da \int_0^\infty db e^{-(a^2+b^2)} a^{2x-1} b^{2y-1} \\ &= \int_{-\infty}^\infty da \int_{-\infty}^\infty db e^{-(a^2+b^2)} |a|^{2x-1} |b|^{2y-1} \end{aligned}$$

We now change coordinates once more, this time to polar $a = r \cos \theta$ and $b = r \sin \theta$. We get

$$\begin{aligned} \Gamma(x)\Gamma(y) &= \int_0^\infty r dr e^{-r^2} r^{2x+2y-2} \int_0^{2\pi} d\theta |\cos \theta|^{2x-1} |\sin \theta|^{2y-1} \\ &= \frac{1}{2} \Gamma(x+y) \times 4 \int_0^{\pi/2} d\theta (\cos \theta)^{2x-1} (\sin \theta)^{2y-1} \\ &= \Gamma(x+y) \int_0^1 dt (1-t)^{y-1} t^{x-1} \end{aligned}$$

where, in the final line, we made the substitution $t = \cos^2 \theta$. This completes the proof.

The Virasoro-Shapiro Amplitude

In the closed string computation, we came across the integral

$$C(a, b) = \int d^2 z |z|^{2a-2} |1-z|^{2b-2}$$

We will now evaluate this and show that it is given by (6.11). We start by using a trick. We can write

$$|z|^{2a-2} = \frac{1}{\Gamma(1-a)} \int_0^\infty dt t^{-a} e^{-|z|^2 t}$$

which follows from the definition (6.26) of the gamma function. Similarly, we can write

$$|1-z|^{2b-2} = \frac{1}{\Gamma(1-b)} \int_0^\infty du u^{-b} e^{-|1-z|^2 u}$$

We decompose the complex coordinate $z = x + iy$, so that the measure of the integral is $d^2z = 2dxdy$. We can then write the integral $C(a, b)$ as

$$\begin{aligned} C(a, b) &= \int \frac{d^2z \, du \, dt}{\Gamma(1-a)\Gamma(1-b)} t^{-a} u^{-b} e^{-|z|^2 t} e^{-|1-z|^2 u} \\ &= 2 \int \frac{dx \, dy \, du \, dt}{\Gamma(1-a)\Gamma(1-b)} t^{-a} u^{-b} e^{-(t+u)(x^2+y^2)+2xu-u} \\ &= 2 \int \frac{dx \, dy \, du \, dt}{\Gamma(1-a)\Gamma(1-b)} t^{-a} u^{-b} \exp \left(-(t+u) \left[\left(x - \frac{u}{t+u} \right)^2 + y^2 \right] - u + \frac{u^2}{t+u} \right) \end{aligned}$$

Now we do the $dxdy$ integral which is simply Gaussian. We find

$$C(a, b) = \frac{2\pi}{\Gamma(1-a)\Gamma(1-b)} \int_0^\infty du \, dt \frac{t^{-a} u^{-b}}{t+u} e^{-tu/(t+u)}$$

Finally, we make a change of variables. We write $t = \alpha\beta$ and $u = (1-\beta)\alpha$. In order for t and u to take values in the range $[0, \infty)$, we require $\alpha \in [0, \infty)$ and $\beta \in [0, 1]$. Taking into account the Jacobian arising from this transformation, which is simply α , the integral becomes

$$C(a, b) = \frac{2\pi}{\Gamma(1-a)\Gamma(1-b)} \int d\alpha \, d\beta \frac{\alpha^{1-a-b}}{\alpha} \beta^{-a} (1-\beta)^{-b} e^{-\alpha\beta(1-\beta)}$$

But we recognize the integral over $d\alpha$: it is simply

$$\int_0^\infty d\alpha \alpha^{-a-b} e^{-\beta\alpha(1-\beta)} = [\beta(1-\beta)]^{a+b-1} \Gamma(1-a-b)$$

We write $c = 1 - a - b$. Finally, we're left with

$$C(a, b) = \frac{2\pi\Gamma(c)}{\Gamma(1-a)\Gamma(1-b)} \int_0^1 d\beta (1-\beta)^{a-1} \beta^{b-1}$$

But the final integral is the Euler beta function (6.27). This gives us our promised result,

$$C(a, b) = \frac{2\pi\Gamma(a)\Gamma(b)\Gamma(c)}{\Gamma(1-a)\Gamma(1-b)\Gamma(1-c)}$$

7. Low Energy Effective Actions

So far, we've only discussed strings propagating in flat spacetime. In this section we will consider strings propagating in different backgrounds. This is equivalent to having different CFTs on the worldsheet of the string.

There is an obvious generalization of the Polyakov action to describe a string moving in curved spacetime,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} g^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu G_{\mu\nu}(X) \quad (7.1)$$

Here $g_{\alpha\beta}$ is again the worldsheet metric. This action describes a map from the worldsheet of the string into a spacetime with metric $G_{\mu\nu}(X)$. (Despite its name, this metric is not to be confused with the Einstein tensor which we won't have need for in this lecture notes).

Actions of the form (7.1) are known as *non-linear sigma models*. (This strange name has its roots in the history of pions). In this context, the D -dimensional spacetime is sometimes called the *target space*. Theories of this type are important in many aspects of physics, from QCD to condensed matter.

Although it's obvious that (7.1) describes strings moving in curved spacetime, there's something a little fishy about just writing it down. The problem is that the quantization of the closed string already gave us a graviton. If we want to build up some background metric $G_{\mu\nu}(X)$, it should be constructed from these gravitons, in much the same manner that a laser beam is made from the underlying photons. How do we see that the metric in (7.1) has anything to do with the gravitons that arise from the quantization of the string?

The answer lies in the use of vertex operators. Let's expand the metric as a small fluctuation around flat space

$$G_{\mu\nu}(X) = \delta_{\mu\nu} + h_{\mu\nu}(X)$$

Then the partition function that we build from the action (7.1) is related to the partition function for a string in flat space by

$$Z = \int \mathcal{D}X \mathcal{D}g e^{-S_{\text{Poly}} - V} = \int \mathcal{D}X \mathcal{D}g e^{-S_{\text{Poly}}} (1 - V + \frac{1}{2}V^2 + \dots)$$

where S_{Poly} is the action for the string in flat space given in (1.22) and V is the expression

$$V = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} g^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu h_{\mu\nu}(X) \quad (7.2)$$

But we've seen this before: it's the vertex operator associated to the graviton state of the string! For a plane wave, corresponding to a graviton with polarization given by the symmetric, traceless tensor $\zeta_{\mu\nu}$ and momentum p^μ , the fluctuation is given by

$$h_{\mu\nu}(X) = \zeta_{\mu\nu} e^{ip \cdot X}$$

With this choice, the expression (7.2) agrees with the vertex operator (5.9). But in general, we could take any linear superposition of plane waves to build up a general fluctuation $h_{\mu\nu}(X)$.

We know that inserting a single copy of V in the path integral corresponds to the introduction of a single graviton state. Inserting e^V in the path integral corresponds to a coherent state of gravitons, changing the metric from $\delta_{\mu\nu}$ to $\delta_{\mu\nu} + h_{\mu\nu}$. In this way we see that the background curved metric of (7.1) is indeed built of the quantized gravitons that we first met back in Section 2.

7.1 Einstein's Equations

In conformal gauge, the Polyakov action in flat space reduces to a free theory. This fact was extremely useful, allowing us to compute the spectrum of the theory. But on a curved background, it is no longer the case. In conformal gauge, the worldsheet theory is described by an interacting two-dimensional field theory,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma G_{\mu\nu}(X) \partial_\alpha X^\mu \partial^\alpha X^\nu \quad (7.3)$$

To understand these interactions in more detail, let's expand around a classical solution which we take to simply be a string sitting at a point \bar{x}^μ .

$$X^\mu(\sigma) = \bar{x}^\mu + \sqrt{\alpha'} Y^\mu(\sigma)$$

Here Y^μ are the dynamical fluctuations about the point which we assume to be small. The factor of $\sqrt{\alpha'}$ is there for dimensional reasons: since $[X] = -1$, we have $[Y] = 0$ and statements like $Y \ll 1$ make sense. Expanding the Lagrangian gives

$$G_{\mu\nu}(X) \partial X^\mu \partial X^\nu = \alpha' \left[G_{\mu\nu}(\bar{x}) + \sqrt{\alpha'} G_{\mu\nu,\omega}(\bar{x}) Y^\omega + \frac{\alpha'}{2} G_{\mu\nu,\omega\rho}(\bar{x}) Y^\omega Y^\rho + \dots \right] \partial Y^\mu \partial Y^\nu$$

Each of the coefficients $G_{\mu\nu,\dots}$ in the Taylor expansion are coupling constants for the interactions of the fluctuations Y^μ . The theory has an infinite number of coupling constants and they are nicely packaged into the function $G_{\mu\nu}(X)$.

We want to know when this field theory is weakly coupled. Obviously this requires the whole infinite set of coupling constants to be small. Let's try to characterize this in a crude manner. Suppose that the target space has characteristic radius of curvature r_c , meaning schematically that

$$\frac{\partial G}{\partial X} \sim \frac{1}{r_c}$$

The radius of curvature is a length scale, so $[r_c] = -1$. From the expansion of the metric, we see that the effective dimensionless coupling is given by

$$\frac{\sqrt{\alpha'}}{r_c} \tag{7.4}$$

This means that we can use perturbation theory to study the CFT (7.3) if the spacetime metric only varies on scales much greater than $\sqrt{\alpha'}$. The perturbation series in $\sqrt{\alpha'}/r_c$ is usually called the α' -expansion to distinguish it from the g_s expansion that we saw in the previous section. Typically a quantity computed in string theory is given by a double perturbation expansion: one in α' and one in g_s .

If there are regions of spacetime where the radius of curvature becomes comparable to the string length scale, $r_c \sim \sqrt{\alpha'}$, then the worldsheet CFT is strongly coupled and we will need to develop new methods to solve it. Notice that strong coupling in α' is hard, but the problem is at least well-defined in terms of the worldsheet path integral. This is qualitatively different to the question of strong coupling in g_s for which, as discussed in Section 6.4.5, we're really lacking a good definition of what the problem even means.

7.1.1 The Beta Function

Classically, the theory defined by (7.3) is conformally invariant. But this is not necessarily true in the quantum theory. To regulate divergences we will have to introduce a UV cut-off and, typically, after renormalization, physical quantities depend on the scale of a given process μ . If this is the case, the theory is no longer conformally invariant. There are plenty of theories which classically possess scale invariance which is broken quantum mechanically. The most famous of these is Yang-Mills.

As we've discussed several times, in string theory conformal invariance is a gauge symmetry and we can't afford to lose it. Our goal in this section is to understand the circumstances under which (7.3) retains conformal invariance at the quantum level.

The object which describes how couplings depend on a scale μ is called the β -function. Since we have a functions worth of couplings, we should really be talking about a β -functional, schematically of the form

$$\beta_{\mu\nu}(G) \sim \mu \frac{\partial G_{\mu\nu}(X; \mu)}{\partial \mu}$$

The quantum theory will be conformally invariant only if

$$\beta_{\mu\nu}(G) = 0$$

We now compute this for the non-linear sigma model at one-loop. Our strategy will be to isolate the UV divergence of the theory and figure out what kind of counterterm we should add. The beta-function will vanish if this counterterm vanishes.

The analysis is greatly simplified by a cunning choice of coordinates. Around any point \bar{x} , we can always pick Riemann normal coordinates such that the expansion in $X^\mu = \bar{x}^\mu + \sqrt{\alpha'} Y^\mu$ gives

$$G_{\mu\nu}(X) = \delta_{\mu\nu} - \frac{\alpha'}{3} \mathcal{R}_{\mu\lambda\nu\kappa}(\bar{x}) Y^\lambda Y^\kappa + \mathcal{O}(Y^3)$$

To quartic order in the fluctuations, the action becomes

$$S = \frac{1}{4\pi} \int d^2\sigma \partial Y^\mu \partial Y^\nu \delta_{\mu\nu} - \frac{\alpha'}{3} \mathcal{R}_{\mu\lambda\nu\kappa} Y^\lambda Y^\kappa \partial Y^\mu \partial Y^\nu$$

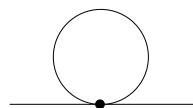
We can now treat this as an interacting quantum field theory in two dimensions. The quartic interaction gives a vertex with the Feynman rule,



$$\sim \mathcal{R}_{\mu\lambda\nu\kappa} (k^\mu \cdot k^\nu)$$

where k_α^μ is the 2d momentum ($\alpha = 1, 2$ is a worldsheet index) for the scalar field Y^μ . It sits in the Feynman rules because we are talking about derivative interactions.

Now we've reduced the problem to a simple interacting quantum field theory, we can compute the β -function using whatever method we like. The divergence in the theory comes from the one-loop diagram



It's actually simplest to think about this diagram in position space. The propagator for a scalar particle is

$$\langle Y^\lambda(\sigma)Y^\kappa(\sigma') \rangle = -\frac{1}{2} \delta^{\lambda\kappa} \ln |\sigma - \sigma'|^2$$

For the scalar field running in the loop, the beginning and end point coincide. The propagator diverges as $\sigma \rightarrow \sigma'$, which is simply reflecting the UV divergence that we would see in the momentum integral around the loop.

To isolate this divergence, we choose to work with dimensional regularization, with $d = 2 + \epsilon$. The propagator then becomes,

$$\begin{aligned} \langle Y^\lambda(\sigma)Y^\kappa(\sigma') \rangle &= 2\pi \delta^{\lambda\kappa} \int \frac{d^{2+\epsilon} k}{(2\pi)^{2+\epsilon}} \frac{e^{ik \cdot (\sigma - \sigma')}}{k^2} \\ &\longrightarrow \frac{\delta^{\lambda\kappa}}{\epsilon} \quad \text{as } \sigma \rightarrow \sigma' \end{aligned}$$

The necessary counterterm for this divergence can be determined simply by replacing $Y^\lambda Y^\kappa$ in the action with $\langle Y^\lambda Y^\kappa \rangle$. To subtract the $1/\epsilon$ term, we add the counterterm

$$\mathcal{R}_{\mu\lambda\nu\kappa} Y^\lambda Y^\kappa \partial Y^\mu \partial Y^\nu \rightarrow \mathcal{R}_{\mu\lambda\nu\kappa} Y^\lambda Y^\kappa \partial Y^\mu \partial Y^\nu - \frac{1}{\epsilon} \mathcal{R}_{\mu\nu} \partial Y^\mu \partial Y^\nu$$

One can check that this can be absorbed by a wavefunction renormalization $Y^\mu \rightarrow Y^\mu + (\alpha'/6\epsilon) \mathcal{R}_\nu^\mu Y^\nu$, together with the renormalization of the coupling constant which, in our theory, is the metric $G_{\mu\nu}$. We require,

$$G_{\mu\nu} \rightarrow G_{\mu\nu} + \frac{\alpha'}{\epsilon} \mathcal{R}_{\mu\nu} \tag{7.5}$$

From this we learn the beta function of the theory and the condition for conformal invariance. It is

$$\beta_{\mu\nu}(G) = \alpha' \mathcal{R}_{\mu\nu} = 0 \tag{7.6}$$

This is a magical result! The requirement for the sigma-model to be conformally invariant is that the target space must be Ricci flat: $\mathcal{R}_{\mu\nu} = 0$. Or, in other words, the background spacetime in which the string moves must obey the vacuum Einstein equations! We see that the equations of general relativity also describe the renormalization group flow of 2d sigma models.

There are several more magical things just around the corner, but it's worth pausing to make a few diverse comments.

Beta Functions and Weyl Invariance

The above calculation effectively studies the breakdown of conformal invariance in the CFT (7.3) on a flat worldsheet. We know that this should be the same thing as the breakdown of Weyl invariance on a curved worldsheet. Since this is such an important result, let's see how it works from this other perspective. We can consider the worldsheet metric

$$g_{\alpha\beta} = e^{2\phi} \delta_{\alpha\beta}$$

Then, in dimensional regularization, the theory is not Weyl invariant in $d = 2 + \epsilon$ dimensions because the contribution from \sqrt{g} does not quite cancel that from the inverse metric $g^{\alpha\beta}$. The action is

$$\begin{aligned} S &= \frac{1}{4\pi\alpha'} \int d^{2+\epsilon}\sigma e^{\phi\epsilon} \partial_\alpha X^\mu \partial^\alpha X^\nu G_{\mu\nu}(X) \\ &\approx \frac{1}{4\pi\alpha'} \int d^{2+\epsilon}\sigma (1 + \phi\epsilon) \partial_\alpha X^\mu \partial^\alpha X^\nu G_{\mu\nu}(X) \end{aligned}$$

where, in this expression, the $\alpha = 1, 2$ index is now raised and lowered with $\delta_{\alpha\beta}$. If we replace $G_{\mu\nu}$ in this expression with the renormalized metric (7.5), we see that there's a term involving ϕ which remains even as $\epsilon \rightarrow 0$,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \partial_\alpha X^\mu \partial^\alpha X^\nu [G_{\mu\nu}(X) + \alpha'\phi \mathcal{R}_{\mu\nu}(X)]$$

This indicates a breakdown of Weyl invariance. Indeed, we can look at our usual diagnostic for Weyl invariance, namely the vanishing of T_α^α . In conformal gauge, this is given by

$$T_{\alpha\beta} = +\frac{4\pi}{\sqrt{g}} \frac{\partial S}{\partial g^{\alpha\beta}} = -2\pi \frac{\partial S}{\partial \phi} \delta_{\alpha\beta} \Rightarrow T_\alpha^\alpha = -\frac{1}{2} \mathcal{R}_{\mu\nu} \partial X^\mu \partial X^\nu$$

In this way of looking at things, we define the β -function to be the coefficient in front of $\partial X \partial X$, namely

$$T_\alpha^\alpha = -\frac{1}{2\alpha'} \beta_{\mu\nu} \partial X^\mu \partial X^\nu$$

Again, we have the result

$$\beta_{\mu\nu} = \alpha' \mathcal{R}_{\mu\nu}$$

7.1.2 Ricci Flow

In string theory we only care about conformal theories with Ricci flat metrics. (And generalizations of this result that we will discuss shortly). However, in other areas of physics and mathematics, the RG flow itself is important. It is usually called Ricci flow,

$$\mu \frac{\partial G_{\mu\nu}}{\partial \mu} = \alpha' \mathcal{R}_{\mu\nu} \quad (7.7)$$

which dictates how the metric changes with scale μ .

As an illustrative and simple example, consider the target space S^2 with radius r . This is an important model in condensed matter physics where it describes the low-energy limit of a one-dimensional Heisenberg spin chain. It is sometimes called the $O(3)$ sigma-model. Because the sphere is a symmetric space, the only effect of the RG flow is to make the radius scale dependent: $r = r(\mu)$. The beta function is given by

$$\mu \frac{\partial r^2}{\partial \mu} = \frac{\alpha'}{2\pi}$$

Hence r gets large as we go towards the UV and small towards the IR. Since the coupling is $1/r$, this means that the non-linear sigma model with S^2 target space is asymptotically free. At low energies, the theory is strongly coupled and perturbative calculations — such as this one-loop beta function — are no longer trusted. In particular, one can show that the S^2 sigma-model develops a mass gap in the IR.

The idea of Ricci flow (7.7) was recently used by Perelman to prove the Poincaré conjecture. In fact, Perelman used a slightly generalized version of Ricci flow which we will see shortly. In the language of string theory, he introduced the dilaton field.

7.2 Other Couplings

We've understood how strings couple to a background spacetime metric. But what about the other modes of the string? In Section 2, we saw that a closed string has further massless states which are associated to the anti-symmetric tensor $B_{\mu\nu}$ and the dilaton Φ . We will now see how the string reacts if these fields are turned on in spacetime.

7.2.1 Charged Strings and the B field

Let's start by looking at how strings couple to the anti-symmetric field $B_{\mu\nu}$. We discussed the vertex operator associated to this state in Section 5.4.1. It is given in

(5.9) and takes the same form as the graviton vertex operator, but with $\zeta_{\mu\nu}$ anti-symmetric. It is a simple matter to exponentiate this, to get an expression for how strings propagate in background $B_{\mu\nu}$ field. We'll keep the curved metric $G_{\mu\nu}$ as well to get the general action,

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} \left(G_{\mu\nu}(X) \partial_\alpha X^\mu \partial_\beta X^\nu g^{\alpha\beta} + i B_{\mu\nu}(X) \partial_\alpha X^\mu \partial_\beta X^\nu \epsilon^{\alpha\beta} \right) \quad (7.8)$$

Where $\epsilon^{\alpha\beta}$ is the anti-symmetric 2-tensor, normalized such that $\sqrt{g}\epsilon^{12} = +1$. (The factor of i is there in the action because we're in Euclidean space and this new term has a single “time” derivative). The action retains invariance under worldsheet reparameterizations and Weyl rescaling.

So what is the interpretation of this new term? We will now show that we should think of the field $B_{\mu\nu}$ as analogous to the gauge potential A_μ in electromagnetism. The action (7.8) is telling us that the string is “electrically charged” under $B_{\mu\nu}$.

Gauge Potentials

We'll take a short detour to remind ourselves about some pertinent facts in electromagnetism. Let's start by returning to a point particle. We know that a charged point particle couples to a background gauge potential A_μ through the addition of a worldline term to the action,

$$\int d\tau A_\mu(X) \dot{X}^\mu . \quad (7.9)$$

If this relativistic form looks a little unfamiliar, we can deconstruct it by working in static gauge with $X^0 \equiv t = \tau$, where it reads

$$\int dt A_0(X) + A_i(X) \dot{X}^i ,$$

which should now be recognizable as the Lagrangian that gives rise to the Coulomb and Lorentz force laws for a charged particle.

So what is the generalization of this kind of coupling for a string? First note that (7.9) has an interesting geometrical structure. It is the pull-back of the one-form $A = A_\mu dX^\mu$ in spacetime onto the worldline of the particle. This works because A is a one-form and the worldline is one-dimensional. Since the worldsheet of the string is two-dimensional, the analogous coupling should be to a two-form in spacetime. This is an anti-symmetric

tensor field with two indices, $B_{\mu\nu}$. The pull-back of $B_{\mu\nu}$ onto the worldsheet gives the interaction,

$$\int d^2\sigma \ B_{\mu\nu}(X) \partial_\alpha X^\mu \partial_\beta X^\nu \epsilon^{\alpha\beta} . \quad (7.10)$$

This is precisely the form of the interaction we found in (7.8).

The point particle coupling (7.9) is invariant under gauge transformations of the background field $A_\mu \rightarrow A_\mu + \partial_\mu \alpha$. This follows because the Lagrangian changes by a total derivative. There is a similar statement for the two-form $B_{\mu\nu}$. The spacetime gauge symmetry is,

$$B_{\mu\nu} \rightarrow B_{\mu\nu} + \partial_\mu C_\nu - \partial_\nu C_\mu \quad (7.11)$$

under which the Lagrangian (7.10) changes by a total derivative.

In electromagnetism, one can construct the gauge invariant electric and magnetic fields which are packaged in the two-form field strength $F = dA$. Similarly, for $B_{\mu\nu}$, the gauge invariant field strength $H = dB$ is a three-form,

$$H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu} .$$

This 3-form H is sometimes known as the *torsion*. It plays the same role as torsion in general relativity, providing an anti-symmetric component to the affine connection.

7.2.2 The Dilaton

Let's now figure out how the string couples to a background dilaton field $\Phi(X)$. This is more subtle. A naive construction of the vertex operator is not primary and one must work a little harder. The correct derivation of the vertex operators can be found in Polchinski. Here I will simply give the coupling and explain some important features.

The action of a string moving in a background involving profiles for the massless fields $G_{\mu\nu}$, $B_{\mu\nu}$ and $\Phi(X)$ is given by

$$S = \frac{1}{4\pi\alpha'} \int d^2\sigma \sqrt{g} \ (G_{\mu\nu}(X) \partial_\alpha X^\mu \partial_\beta X^\nu g^{\alpha\beta} + iB_{\mu\nu}(X) \partial_\alpha X^\mu \partial_\beta X^\nu \epsilon^{\alpha\beta} + \alpha' \Phi(X) R^{(2)}) \quad (7.12)$$

where $R^{(2)}$ is the two-dimensional Ricci scalar of the worldsheet. (Up until now, we've always denoted this simply as R but we'll introduce the superscript from hereon to distinguish the worldsheet Ricci scalar from the spacetime Ricci scalar).

The coupling to the dilaton is surprising for several reasons. Firstly, we see that the term in the action vanishes on a flat worldsheet, $R^{(2)} = 0$. This is one of the reasons that it's a little trickier to determine this coupling using vertex operators.

However, the most surprising thing about the coupling to the dilaton is that it *does not* respect Weyl invariance! Since a large part of this course has been about understanding the implications of Weyl invariance, why on earth are we willing to throw it away now?! The answer, of course, is that we're not. Although the dilaton coupling does violate Weyl invariance, there is a way to restore it. We will explain this shortly. But firstly, let's discuss one crucially important implication of the dilaton coupling (7.12).

The Dilaton and the String Coupling

There is an exception to the statement that the classical coupling to the dilaton violates Weyl invariance. This arises when the dilaton is constant. For example, suppose

$$\Phi(X) = \lambda , \text{ a constant}$$

Then the dilaton coupling reduces to something that we've seen before: it is

$$S_{\text{dilaton}} = \lambda \chi$$

where χ is the Euler character of the worldsheet that we introduced in (6.4). This tells us something important: the constant mode of the dilaton, $\langle \Phi \rangle$ determines the string coupling constant. This constant mode is usually taken to be the asymptotic value of the dilaton,

$$\Phi_0 = \lim_{X \rightarrow \infty} \Phi(X) \tag{7.13}$$

The string coupling is then given by

$$g_s = e^{\Phi_0} \tag{7.14}$$

So the string coupling is not an independent parameter of string theory: it is the expectation value of a field. This means that, just like the spacetime metric $G_{\mu\nu}$ (or, indeed, like the Higgs vev) it can be determined dynamically.

We've already seen that our perturbative expansion around flat space is valid as long as $g_s \ll 1$. But now we have a stronger requirement: we can only trust perturbation theory if the string is localized in regions of space where $e^{\Phi(X)} \ll 1$ for all X . If the string ventures into regions where $e^{\Phi(X)}$ is of order 1, then we will need to use techniques that don't rely on string perturbation theory as described in Section 6.4.5.

7.2.3 Beta Functions

We now return to understanding how we can get away with the violation of Weyl invariance in the dilaton coupling (7.12). The key to this is to notice the presence of α' in front of the dilaton coupling. It's there simply on dimensional grounds. (The other two terms in the action both come with derivatives $[\partial X] = -1$, so don't need any powers of α').

However, recall that α' also plays the role of the loop-expansion parameter (7.4) in the non-linear sigma model. This means that the classical lack of Weyl invariance in the dilaton coupling can be compensated by a one-loop contribution arising from the couplings to $G_{\mu\nu}$ and $B_{\mu\nu}$.

To see this explicitly, one can compute the beta-functions for the two-dimensional field theory (7.12). In the presence of the dilaton coupling, it's best to look at the breakdown of Weyl invariance as seen by $\langle T_\alpha^\alpha \rangle$. There are three different kinds of contribution that the stress-tensor can receive, related to the three different spacetime fields. Correspondingly, we define three different beta functions,

$$\langle T_\alpha^\alpha \rangle = -\frac{1}{2\alpha'}\beta_{\mu\nu}(G) g^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu - \frac{i}{2\alpha'}\beta_{\mu\nu}(B) \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu - \frac{1}{2}\beta(\Phi)R^{(2)} \quad (7.15)$$

We will not provide the details of the one-loop beta function computations. We merely state the results⁸,

$$\begin{aligned} \beta_{\mu\nu}(G) &= \alpha' \mathcal{R}_{\mu\nu} + 2\alpha' \nabla_\mu \nabla_\nu \Phi - \frac{\alpha'}{4} H_{\mu\lambda\kappa} H_\nu^{\lambda\kappa} \\ \beta_{\mu\nu}(B) &= -\frac{\alpha'}{2} \nabla^\lambda H_{\lambda\mu\nu} + \alpha' \nabla^\lambda \Phi H_{\lambda\mu\nu} \\ \beta(\Phi) &= -\frac{\alpha'}{2} \nabla^2 \Phi + \alpha' \nabla_\mu \Phi \nabla^\mu \Phi - \frac{\alpha'}{24} H_{\mu\nu\lambda} H^{\mu\nu\lambda} \end{aligned}$$

A consistent background of string theory must preserve Weyl invariance, which now requires $\beta_{\mu\nu}(G) = \beta_{\mu\nu}(B) = \beta(\Phi) = 0$.

7.3 The Low-Energy Effective Action

The equations $\beta_{\mu\nu}(G) = \beta_{\mu\nu}(B) = \beta(\Phi) = 0$ can be viewed as the equations of motion for the background in which the string propagates. We now change our perspective: we

⁸The relationship between the beta function and Einstein's equations was first shown by Friedan in his 1980 PhD thesis. A readable account of the full beta functions can be found in the paper by Callan, Friedan, Martinec and Perry “*Strings in Background Fields*”, Nucl. Phys. B262 (1985) 593. The full calculational details can be found in TASI lecture notes by Callan and Thorlacius which can be downloaded from the course webpage.

look for a $D = 26$ dimensional spacetime action which reproduces these beta-function equations as the equations of motion. This is the *low-energy effective action* of the bosonic string,

$$S = \frac{1}{2\kappa_0^2} \int d^{26}X \sqrt{-G} e^{-2\Phi} \left(\mathcal{R} - \frac{1}{12} H_{\mu\nu\lambda} H^{\mu\nu\lambda} + 4\partial_\mu\Phi \partial^\mu\Phi \right) \quad (7.16)$$

where we have taken the liberty of Wick rotating back to Minkowski space for this expression. Here the overall constant involving κ_0 is not fixed by the field equations but can be determined by coupling these equations to a suitable source as described, for example, in 7.4.2. On dimensional grounds alone, it scales as $\kappa_0^2 \sim l_s^{24}$ where $\alpha' = l_s^2$.

Varying the action with respect to the three fields can be shown to yield the beta functions thus,

$$\begin{aligned} \delta S = \frac{1}{2\kappa_0^2\alpha'} \int d^{26}X \sqrt{-G} e^{-2\Phi} & (\delta G_{\mu\nu} \beta^{\mu\nu}(G) - \delta B_{\mu\nu} \beta^{\mu\nu}(B) \\ & -(2\delta\Phi + \frac{1}{2}G^{\mu\nu}\delta G_{\mu\nu})(\beta_\lambda^\lambda(G) - 4\beta(\Phi)) \end{aligned}$$

Equation (7.16) governs the low-energy dynamics of the spacetime fields. The caveat “low-energy” refers to the fact that we only worked with the one-loop beta functions which requires large spacetime curvature.

Something rather remarkable has happened here. We started, long ago, by looking at how a single string moves in flat space. Yet, on grounds of consistency alone, we’re led to the action (7.16) governing how spacetime and other fields fluctuate in $D = 26$ dimensions. It feels like the tail just wagged the dog. That tiny string is seriously high-maintenance: its requirements are so stringent that they govern the way the whole universe moves.

You may also have noticed that we now have two different methods to compute the scattering of gravitons in string theory. The first is in terms of scattering amplitudes that we discussed in Section 6. The second is by looking at the dynamics encoded in the low-energy effective action (7.16). Consistency requires that these two approaches agree. They do.

7.3.1 String Frame and Einstein Frame

The action (7.16) isn’t quite of the familiar Einstein-Hilbert form because of that strange factor of $e^{-2\Phi}$ that’s sitting out front. This factor simply reflects the fact that the action has been computed at tree level in string perturbation theory and, as we saw in Section 6, such terms typically scale as $1/g_s^2$.

It's also worth pointing out that the kinetic terms for Φ in (7.16) seem to have the wrong sign. However, it's not clear that we should be worried about this because, again, the factor of $e^{-2\Phi}$ sits out front meaning that the kinetic terms are not canonically normalized anyway.

To put the action in more familiar form, we can make a field redefinition. Firstly, it's useful to distinguish between the constant part of the dilaton, Φ_0 , and the part that varies which we call $\tilde{\Phi}$. We defined the constant part in (7.13); it is related to the string coupling constant. The varying part is simply given by

$$\tilde{\Phi} = \Phi - \Phi_0 \quad (7.17)$$

In D dimensions, we define a new metric $\tilde{G}_{\mu\nu}$ as a combination of the old metric and the dilaton,

$$\tilde{G}_{\mu\nu}(X) = e^{-4\tilde{\Phi}/(D-2)} G_{\mu\nu}(X) \quad (7.18)$$

Note that this isn't to be thought of as a coordinate transformation or symmetry of the action. It's merely a relabeling, a mixing-up, of the fields in the theory. We could make such redefinitions in any field theory. Typically, we choose not to because the fields already have canonical kinetic terms. The point of the transformation (7.18) is to get the fields in (7.16) to have canonical kinetic terms as well.

The new metric (7.18) is related to the old by a conformal rescaling. One can check that two metrics related by a general conformal transformation $\tilde{G}_{\mu\nu} = e^{2\omega} G_{\mu\nu}$, have Ricci scalars related by

$$\tilde{\mathcal{R}} = e^{-2\omega} (\mathcal{R} - 2(D-1)\nabla^2\omega - (D-2)(D-1)\partial_\mu\omega\partial^\mu\omega)$$

(We used a particular version of this earlier in the course when considering $D=2$ conformal transformations). With the choice $\omega = -2\tilde{\Phi}/(D-2)$ in (7.18), and restricting back to $D=26$, the action (7.16) becomes

$$S = \frac{1}{2\kappa^2} \int d^{26}X \sqrt{-\tilde{G}} \left(\tilde{\mathcal{R}} - \frac{1}{12} e^{-\tilde{\Phi}/3} H_{\mu\nu\lambda} H^{\mu\nu\lambda} - \frac{1}{6} \partial_\mu \tilde{\Phi} \partial^\mu \tilde{\Phi} \right) \quad (7.19)$$

The kinetic terms for $\tilde{\Phi}$ are now canonical and come with the right sign. Notice that there is no potential term for the dilaton and therefore nothing that dynamically sets its expectation value in the bosonic string. However, there do exist backgrounds of the superstring in which a potential for the dilaton develops, fixing the string coupling constant.

The gravitational part of the action takes the standard Einstein-Hilbert form. The gravitational coupling is given by

$$\kappa^2 = \kappa_0^2 e^{2\Phi_0} \sim l_s^{24} g_s^2 \quad (7.20)$$

The coefficient in front of Einstein-Hilbert term is usually identified with Newton's constant

$$8\pi G_N = \kappa^2$$

Note, however, that this is Newton's constant in $D = 26$ dimensions: it will differ from Newton's constant measured in a four-dimensional world. From Newton's constant, we define the $D = 26$ Planck length $8\pi G_N = l_p^{24}$ and Planck mass $M_p = l_p^{-1}$. (With the factor of 8π sitting there, this is usually called the reduced Planck mass). Comparing to (7.20), we see that weak string coupling, $g_s \ll 1$, provides a parameteric separation between the Planck scale and the string scale,

$$g_s \ll 1 \Rightarrow l_p \ll l_s$$

Often the mysteries of gravitational physics are associated with the length scale l_p . We understand string theory best when $g_s \ll 1$ where much of stringy physics occurs at $l_s \gg l_p$ and can be disentangled from strong coupling effects in gravity.

The original metric $G_{\mu\nu}$ is usually called the *string metric* or *sigma-model metric*. It is the metric that strings see, as reflected in the action (7.1). In contrast, $\tilde{G}_{\mu\nu}$ is called the *Einstein metric*. Of course, the two actions (7.16) and (7.19) describe the same physics: we have simply chosen to package the fields in a different way in each. The choice of metric — $G_{\mu\nu}$ or $\tilde{G}_{\mu\nu}$ — is usually referred to as a choice of *frame*: string frame, or Einstein frame.

The possibility of defining two metrics really arises because we have a massless scalar field Φ in the game. Whenever such a field exists, there's nothing to stop us measuring distances in different ways by including Φ in our ruler. Said another way, massless scalar fields give rise to long range attractive forces which can mix with gravitational forces and violate the principle of equivalence. Ultimately, if we want to connect to Nature, we need to find a way to make Φ massive. Such mechanisms exist in the context of the superstring.

7.3.2 Corrections to Einstein's Equations

Now that we know how Einstein's equations arise from string theory, we can start to try to understand new physics. For example, what are the quantum corrections to Einstein's equations?

On general grounds, we expect these corrections to kick in when the curvature r_c of spacetime becomes comparable to the string length scale $\sqrt{\alpha'}$. But that dovetails very nicely with the discussion above where we saw that the perturbative expansion parameter for the non-linear sigma model is α'/r_c^2 . Computing the next loop correction to the beta function will result in corrections to Einstein's equations!

If we ignore H and Φ , the 2-loop sigma-model beta function can be easily computed and results in the α' correction to Einstein's equations:

$$\beta_{\mu\nu} = \alpha' \mathcal{R}_{\mu\nu} + \frac{1}{2} \alpha'^2 \mathcal{R}_{\mu\lambda\rho\sigma} \mathcal{R}_{\nu}^{\lambda\rho\sigma} + \dots = 0$$

Such two loop corrections also appear in the heterotic superstring. However, they are absent for the type II string theories, with the first corrections appearing at 4-loops from the perspective of the sigma-model.

String Loop Corrections

Perturbative string theory has an α' expansion and g_s expansion. We still have to discuss the latter. Here an interesting subtlety arises. The sigma-model beta functions arise from regulating the UV divergences of the worldsheet. Yet the g_s expansion cares only about the topology of the string. How can the UV divergences care about the global nature of the worldsheet. Or, equivalently, how can the higher-loop corrections to the beta-functions give anything interesting?

The resolution to this puzzle is to remember that, when computing higher g_s corrections, we have to integrate over the moduli space of Riemann surfaces. But this moduli space will include some tricky points where the Riemann surface degenerates. (For example, one cycle of the torus may pinch off). At these points, the UV divergences suddenly do care about global topology and this results in the g_s corrections to the low-energy effective action.

7.3.3 Nodding Once More to the Superstring

In section 2.5, we described the massless bosonic content for the four superstring theories: Heterotic $SO(32)$, Heterotic $E_8 \times E_8$, Type IIA and Type IIB. Each of them contains the fields $G_{\mu\nu}$, $B_{\mu\nu}$ and Φ that appear in the bosonic string, together with a collection of further massless fields. For each, the low-energy effective action describes the dynamics of these fields in $D = 10$ dimensional spacetime. It naturally splits up into three pieces,

$$S_{\text{superstring}} = S_1 + S_2 + S_{\text{fermi}}$$

Here S_{fermi} describes the interactions of the spacetime fermions. We won't describe these here. But we will briefly describe the low-energy bosonic action $S_1 + S_2$ for each of these four superstring theories.

S_1 is essentially the same for all theories and is given by the action we found for the bosonic string in string frame (7.16). We'll start to use form notation and denote $H_{\mu\nu\lambda}$ simply as H_3 , where the subscript tells us the degree of the form. Then the action reads

$$S_1 = \frac{1}{2\kappa_0^2} \int d^{10}X \sqrt{-G} e^{-2\Phi} \left(\mathcal{R} - \frac{1}{2} |\tilde{H}_3|^2 + 4\partial_\mu\Phi \partial^\mu\Phi \right) \quad (7.21)$$

There is one small difference, which is that the field \tilde{H}_3 that appears here for the heterotic string is not quite the same as the original H_3 ; we'll explain this further shortly.

The second part of the action, S_2 , describes the dynamics of the extra fields which are specific to each different theory. We'll now go through the four theories in turn, explaining S_2 in each case.

- **Type IIA:** For this theory, \tilde{H}_3 appearing in (7.21) is $H_3 = dB_2$, just as we saw in the bosonic string. In Section 2.5, we described the extra bosonic fields of the Type IIA theory: they consist of a 1-form C_1 and a 3-form C_3 . The dynamics of these fields is governed by the so-called Ramond-Ramond part of the action and is written in form notation as,

$$S_2 = -\frac{1}{4\kappa_0^2} \int d^{10}X \left[\sqrt{-G} \left(|F_2|^2 + |\tilde{F}_4|^2 \right) + B_2 \wedge F_4 \wedge F_4 \right]$$

Here the field strengths are given by $F_2 = dC_1$ and $F_4 = dC_3$, while the object that appears in the kinetic terms is $\tilde{F}_4 = F_4 - C_1 \wedge H_3$. Notice that the final term in the action does not depend on the metric: it is referred to as a *Chern-Simons* term.

- **Type IIB:** Again, $\tilde{H}_3 \equiv H_3$. The extra bosonic fields are now a scalar C_0 , a 2-form C_2 and a 4-form C_4 . Their action is given by

$$S_2 = -\frac{1}{4\kappa_0^2} \int d^{10}X \left[\sqrt{-G} \left(|F_1|^2 + |\tilde{F}_3|^2 + \frac{1}{2} |\tilde{F}_5|^2 \right) + C_4 \wedge H_3 \wedge F_3 \right]$$

where $F_1 = dC_0$, $F_3 = dC_2$ and $F_5 = dC_4$. Once again, the kinetic terms involve more complicated combinations of the forms: they are $\tilde{F}_3 = F_3 - C_0 \wedge H_3$ and

$\tilde{F}_5 = F_5 - \frac{1}{2}C_2 \wedge H_3 + \frac{1}{2}B_2 \wedge F_3$. However, for type IIB string theory, there is one extra requirement on these fields that cannot be implemented in any simple way in terms of a Lagrangian: \tilde{F}_5 must be self-dual

$$\tilde{F}_5 = {}^* \tilde{F}_5$$

Strictly speaking, one should say that the low-energy dynamics of type IIB theory is governed by the equations of motion that we get from the action, supplemented with this self-duality requirement.

- **Heterotic:** Both heterotic theories have just one further massless bosonic ingredient: a non-Abelian gauge field strength F_2 , with gauge group $SO(32)$ or $E_8 \times E_8$. The dynamics of this field is simply the Yang-Mills action in ten dimensions,

$$S_2 = \frac{\alpha'}{8\kappa_0^2} \int d^{10}X \sqrt{-G} \text{Tr} |F_2|^2$$

The one remaining subtlety is to explain what \tilde{H}_3 means in (7.21): it is defined as $\tilde{H}_3 = dB_2 - \alpha' \omega_3/4$ where ω_3 is the Chern-Simons three form constructed from the non-Abelian gauge field A_1

$$\omega_3 = \text{Tr} \left(A_1 \wedge dA_1 + \frac{2}{3} A_1 \wedge A_1 \wedge A_1 \right)$$

The presence of this strange looking combination of forms sitting in the kinetic terms is tied up with one of the most intricate and interesting aspects of the heterotic string, known as anomaly cancelation.

The actions that we have written down here probably look a little arbitrary. But they have very important properties. In particular, the full action $S_{\text{superstring}}$ of each of the Type II theories is invariant under $\mathcal{N} = 2$ spacetime supersymmetry. (That means 32 supercharges). They are the unique actions with this property. Similarly, the heterotic superstring actions are invariant under $\mathcal{N} = 1$ supersymmetry and, crucially, do not suffer from anomalies. The second book by Polchinski is a good place to start learning more about these ideas.

7.4 Some Simple Solutions

The spacetime equations of motion,

$$\beta_{\mu\nu}(G) = \beta_{\mu\nu}(B) = \beta(\Phi) = 0$$

have many solutions. This is part of the story of vacuum selection in string theory. What solution, if any, describes the world we see around us? Do we expect this putative

solution to have other special properties, or is it just a random choice from the many possibilities? The answer is that we don't really know, but there is currently no known principle which uniquely selects a solution which looks like our world — with the gauge groups, matter content and values of fundamental constants that we observe — from the many other possibilities. Of course, these questions should really be asked in the context of the superstring where a greater understanding of various non-perturbative effects such as D-branes and fluxes leads to an even greater array of possible solutions.

Here we won't discuss these problems. Instead, we'll just discuss a few simple solutions that are well known. The first plays a role when trying to make contact with the real world, while the value of the others lies mostly in trying to better understand the structure of string theory.

7.4.1 Compactifications

We've seen that the bosonic string likes to live in $D = 26$ dimensions. But we don't. Or, more precisely, we only observe three macroscopically large spatial dimensions. How do we reconcile these statements?

Since string theory is a theory of gravity, there's nothing to stop extra dimensions of the universe from curling up. Indeed, under certain circumstances, this may be required dynamically. Here we exhibit some simple solutions of the low-energy effective action which have this property. We set $H_{\mu\nu\rho} = 0$ and Φ to a constant. Then we are simply searching for Ricci flat backgrounds obeying $\mathcal{R}_{\mu\nu} = 0$. There are solutions where the metric is a direct product of metrics on the space

$$\mathbf{R}^{1,3} \times \mathbf{X} \quad (7.22)$$

where \mathbf{X} is a compact 22-dimensional Ricci-flat manifold.

The simplest such manifold is just $\mathbf{X} = \mathbf{T}^{22}$, the torus endowed with a flat metric. But there are a whole host of other possibilities. Compact, complex manifolds that admit such Ricci-flat metrics are called *Calabi-Yau* manifolds. (Strictly speaking, Calabi-Yau manifolds are complex manifolds with vanishing first Chern class. Yau's theorem guarantees the existence of a unique Ricci flat metric on these spaces).

The idea that there may be extra, compact directions in the universe was considered long before string theory and goes by the name of *Kaluza-Klein compactification*. If the characteristic length scale L of the space \mathbf{X} is small enough then the presence of these extra dimensions would not have been observed in experiment. The standard model of particle physics has been accurately tested to energies of a TeV or so, meaning that

if the standard model particles can roam around \mathbf{X} , then the length scale must be $L \lesssim (\text{TeV})^{-1} \sim 10^{-16} \text{ cm}$.

However, one can cook up scenarios in which the standard model is stuck somewhere in these extra dimensions (for example, it may be localized on a D-brane). Under these circumstances, the constraints become much weaker because we would rely on gravitational experiments to detect extra dimensions. Present bounds require only $L \lesssim 10^{-5} \text{ cm}$.

Consider the Einstein-Hilbert term in the low-energy effective action. If we are interested only in the dynamics of the 4d metric on $\mathbf{R}^{1,3}$, this is given by

$$S_{EH} = \frac{1}{2\kappa^2} \int d^{26}X \sqrt{-\tilde{G}} \tilde{\mathcal{R}} = \frac{\text{Vol}(\mathbf{X})}{2\kappa^2} \int d^4X \sqrt{-G_{4d}} \mathcal{R}_{4d}$$

(There are various moduli of the internal manifold \mathbf{X} that are being neglected here). From this equation, we learn that effective 4d Newton constant is given in terms of 26d Newton constant by,

$$8\pi G_N^{4d} = \frac{\kappa^2}{\text{Vol}(\mathbf{X})}$$

Rewriting this in terms of the 4d Planck scale, we have $l_p^{(4d)} \sim g_s l_s^{12} / \sqrt{\text{Vol}(\mathbf{X})}$. To trust this whole analysis, we require $g_s \ll 1$ and all length scales of the internal space to be bigger than l_s . This ensures that $l_p^{(4d)} < l_s$. Although the 4d Planck length is ludicrously small, $l_p^{(4d)} \sim 10^{-33} \text{ cm}$, it may be that we don't have to probe to this distance to uncover UV gravitational physics. The back-of-the-envelope calculation above shows that the string scale l_s could be much larger, enhanced by the volume of extra dimensions.

7.4.2 The String Itself

We've seen that quantizing small loops of string gives rise to the graviton and $B_{\mu\nu}$ field. Yet, from the sigma model action (7.12), we also know that the string is charged under the $B_{\mu\nu}$. Moreover, the string has tension, which ensures that it also acts as a source for the metric $G_{\mu\nu}$. So what does the back-reaction of the string look like? Or, said another way: what is the sigma-model describing a string moving in the background of another string?

Consider an infinite, static, straight string stretched in the X^1 direction. We can solve for the background fields by coupling the equations of motion to a delta-function string

source. This is the same kind of calculation that we're used to in electromagnetism. The resulting spacetime fields are given by

$$ds^2 = f(r)^{-1} (-dt^2 + dX_1^2) + \sum_{i=2}^{25} dX_i^2 \\ B = (f(r)^{-1} - 1) dt \wedge dX_1 \quad , \quad e^{2\Phi} = f(r)^{-1} \quad (7.23)$$

The function $f(r)$ depends only on the transverse direction $r^2 = \sum_{i=2}^{25} X_i^2$ and is given by

$$f(r) = 1 + \frac{g_s^2 N l_s^{22}}{r^{22}}$$

Here N is some constant which we will shortly demonstrate counts the number of strings which source the background. The string length scale in the solutions is $l_s = \sqrt{\alpha'}$. The function $f(r)$ has the property that it is harmonic in the space transverse to the string, meaning that it satisfies $\nabla_{\mathbf{R}^{24}}^2 f(r) = 0$ except at $r = 0$.

Let's compute the B -field charge of this solution. We do exactly what we do in electromagnetism: we integrate the total flux through a sphere which surrounds the object. The string lies along the X^1 direction so the transverse space is \mathbf{R}^{24} . We can consider a sphere \mathbf{S}^{23} at the boundary of this transverse space. We should be integrating the flux over this sphere. But what is the expression for the flux?

To see what we should do, let's look at the action for $H_{\mu\nu\rho}$ in the presence of a string source. We will use form notation since this is much cleaner and refer to $H_{\mu\nu\rho}$ simply as H_3 . Schematically, the action takes the form

$$\frac{1}{g_s^2} \int_{\mathbf{R}^{26}} H_3 \wedge {}^*H_3 + \int_{\mathbf{R}^2} B_2 = \frac{1}{g_s^2} \int_{\mathbf{R}^{26}} H_3 \wedge {}^*H_3 + g_s^2 B_2 \wedge \delta(\omega)$$

Here $\delta(\omega)$ is a delta-function source with support on the 2d worldsheet of the string. The equation of motion is

$$d{}^*H_3 \sim g_s^2 \delta(\omega)$$

From this we learn that to compute the charge of a single string we need to integrate

$$\frac{1}{g_s^2} \int_{\mathbf{S}^{23}} {}^*H_3 = 1$$

After these general comments, we now return to our solution (7.23). The above discussion was schematic and no attention was paid to factors of 2 and π . Keeping in this spirit, the flux of the solution (7.23) can be checked to be

$$\frac{1}{g_s^2} \int_{\mathbf{S}^{23}} {}^*H_3 = N$$

This is telling us that the solution (7.23) describes the background sourced by N coincident, parallel fundamental strings. Another way to check this is to compute the ADM mass per unit length of the solution: it is $NT \sim N/\alpha'$ as expected.

Note as far as the low-energy effective action is concerned, there is nothing that insists $N \in \mathbf{Z}$. This is analogous to the statement that nothing in classical Maxwell theory requires e to be quantized. However, in string theory, as in QED, we know the underlying sources of the microscopic theory and N must indeed take integer values.

Finally, notice that as $r \rightarrow 0$, the solution becomes singular. It is not to be trusted in this regime where higher order α' corrections become important.

7.4.3 Magnetic Branes

We've already seen that string theory is not just a theory of strings; there are also D-branes, defined as surfaces on which strings can end. We'll have much more to say about D-branes in Section 7.5. Here, we will consider a third kind of object that exists in string theory. It is again a brane — meaning that it is extended in some number of spacetime directions — but it is not a D-brane because the open string cannot end there. In these lectures we will call it the *magnetic brane*.

Electric and Magnetic Charges

You're probably not used to talking about magnetically charged objects in electromagnetism. Indeed, in undergraduate courses we usually don't get much further than pointing out that $\nabla \cdot \mathbf{B} = 0$ does not allow point-like magnetic charges. However, in the context of quantum field theory, much of the interesting behaviour often boils down to understanding how magnetic charges behave. And the same is true of string theory. Because this may be unfamiliar, let's take a minute to discuss the basics.

In electromagnetism in $d = 3 + 1$ dimensions, we measure electric charge q by integrating the electric field \vec{E} over a sphere \mathbf{S}^2 that surrounds the particle,

$$q = \int_{\mathbf{S}^2} \vec{E} \cdot d\vec{S} = \int_{\mathbf{S}^2} {}^*F_2 \quad (7.24)$$

In the second equality we have introduced the notation of differential forms that we also used in the previous example to discuss the string solutions.

Suppose now that a particle carries magnetic charge g . This can be measured by integrating the magnetic field \vec{B} over the same sphere. This means

$$g = \int_{\mathbf{S}^2} \vec{B} \cdot d\vec{S} = \int_{\mathbf{S}^2} F_2 \quad (7.25)$$

In $d = 3+1$ dimensions, both electrically and magnetically charged objects are particles. But this is not always true in any dimension! The reason that it holds in $4d$ is because both the field strength F_2 and the dual field strength *F_2 are 2-forms. Clearly, this is rather special to four dimensions.

In general, suppose that we have a p -brane that is electrically charged under a suitable gauge field. As we discussed in Section 7.2.1, a $(p+1)$ -dimensional object naturally couples to a $(p+1)$ -form gauge potential C_{p+1} through,

$$\mu \int_W C_{p+1}$$

where μ is the charge of the object, while W is the worldvolume of the brane. The $(p+1)$ -form gauge potential has a $(p+2)$ -form field strength

$$G_{p+2} = dC_{p+1}$$

To measure the electric charge of the p -brane, we need to integrate the field strength over a sphere that completely surrounds the object. A p -brane in D -dimensions has a transverse space \mathbf{R}^{D-p-1} . We can integrate the flux over the sphere at infinity, which is \mathbf{S}^{D-p-2} . And, indeed, the counting works out nicely because, in D dimensions, the dual field strength is a $(D-p-2)$ -form, ${}^*G_{p+2} = \tilde{G}_{D-p-2}$, which we can happily integrate over the sphere to find the charge sitting inside,

$$q = \int_{\mathbf{S}^{D-p-2}} {}^*G_{p+2}$$

This equation is the generalized version of (7.24)

Now let's think about magnetic charges. The generalized version of (7.25) suggest that we should compute the magnetic charge by integrating G_{p+2} over a sphere \mathbf{S}^{p+2} . What kind of object sits inside this sphere to emit the magnetic charge? Doing the sums backwards, we see that it should be a $(D-p-4)$ -brane.

We can write down the coupling between the $(D-p-4)$ -brane and the field strength. To do so, we first need to introduce the magnetic gauge potential defined by

$${}^*G_{p+2} = \tilde{G}_{D-p-2} = d\tilde{C}_{D-p-3} \tag{7.26}$$

We can then add the magnetic coupling to the worldvolume \tilde{W} of a $(D-p-4)$ -brane simply by writing

$$\tilde{\mu} \int_{\tilde{W}} \tilde{C}_{D-p-3}$$

where $\tilde{\mu}$ is the magnetic charge. Note that it's typically not possible to write down a Lagrangian that includes both magnetically charged object and electrically charged objects at the same time. This would need us to include both C_{p+1} and \tilde{C}_{D-p-3} in the Lagrangian, but these are not independent fields: they're related by the rather complicated differential equations (7.26).

The Magnetic Brane in Bosonic String Theory

After these generalities, let's see what it means for the bosonic string. The fundamental string is a 1-brane and, as we saw in Section 7.2.1, carries electric charge under the 2-form B . The appropriate object carrying magnetic charge under B is therefore a $(D - p - 4) = (26 - 1 - 4) = 21$ -brane.

To stress a point: neither the fundamental string, nor the magnetic 21-brane are D-branes. They are not surfaces where strings can end. We are calling them *branes* only because they are extended objects.

The magnetic 21-brane of the bosonic string can be found as a solution to the low-energy equations of motion. The solution can be written in terms of the dual potential \tilde{B}_{22} such that $d\tilde{B}_{22} = {}^*dB_2$. It is

$$\begin{aligned} ds^2 &= \left(-dt^2 + \sum_{i=1}^{21} dX_i^2 \right) + h(r) (dX_{22}^2 + \dots + dX_{25}^2) \\ \tilde{B}_{22} &= (1 - h(r)^{-2}) dt \wedge dX_1 \wedge \dots \wedge dX_{21} \\ e^{2\Phi} &= h(r) \end{aligned} \tag{7.27}$$

The function $h(r)$ depends only on the radial direction in \mathbf{R}^4 transverse to the brane: $r^2 = \sum_{i=22}^{25} X_i^2$. It is a harmonic function in \mathbf{R}^4 , given by

$$h(r) = 1 + \frac{Nl_s^2}{r^2}$$

The role of this function in the metric (7.27) is to warp the transverse \mathbf{R}^4 directions. Distances get larger as you approach the brane and the origin, $r = 0$, is at infinite distance.

It can be checked that the solution carried N units of magnetic charge and has tension

$$T \sim \frac{N}{l_s^{22}} \frac{1}{g_s^2}$$

Let's summarize how the tension of different objects scale in string theory. The powers of $\alpha' = l_s^2$ are entirely fixed on dimensional grounds. (Recall that the tension is mass per spatial volume, so the tension of a p -brane has $[T_p] = p + 1$). More interesting is the dependence on the string coupling g_s . The tension of the fundamental string does not depend on g_s , while the magnetic brane scales as $1/g_s^2$. This kind of $1/g^2$ behaviour is typical of solitons in field theories. The D-branes sit between the two: their tension scales as $1/g_s$. Objects with this behaviour are somewhat rarer (although not unheard of) in field theory.

In the perturbative limit, $g_s \rightarrow 0$, both D-branes and magnetic branes are heavy. The coupling of an object with tension T to gravity is governed by $T\kappa^2$ where the gravitational coupling scales as $\kappa \sim g_s^2$ (7.20). This means that in the weak coupling limit, the gravitational backreaction of the string and D-branes can be neglected. However, the coupling of the magnetic brane to gravity is always of order one.

The Magnetic Brane in Superstring Theory

Superstring theories also have a brane magnetically charged under B . It is a $(D - p - 4) = (10 - 1 - 4) = 5$ -brane and is usually referred to as the NS5-brane. The solution in the transverse \mathbf{R}^4 again takes the form (7.27).

The NS5-brane exists in both type II and heterotic string. In many ways it is more mysterious than D-branes and its low-energy effective dynamics is still poorly understood. It is closely related to the 5-brane of M-theory.

7.4.4 Moving Away from the Critical Dimension

The beta function equations provide a new view on the critical dimension $D = 26$ of the bosonic string. To see this, let's look more closely at the dilaton beta function $\beta(\Phi)$ defined in (7.15): it takes the same form as the Weyl anomaly that we discussed back in Section 4.4.2. This means that if we consider a string propagating in $D \neq 26$ then the Weyl anomaly simply arises as the leading order term in the dilaton beta function. So let's relax the requirement of the critical dimension. The equations of motion arising from $\beta_{\mu\nu}(G)$ and $\beta_{\mu\nu}(B)$ are unchanged, while the dilaton beta function equation becomes

$$\beta(\Phi) = \frac{D - 26}{6} - \frac{\alpha'}{2} \nabla^2 \Phi + \alpha' \nabla_\mu \Phi \nabla^\mu \Phi - \frac{\alpha'}{24} H_{\mu\nu\lambda} H^{\mu\nu\lambda} = 0 \quad (7.28)$$

The low-energy effective action in string frame picks up an extra term which looks like a run-away potential for Φ ,

$$S = \frac{1}{2\kappa_0^2} \int d^D X \sqrt{-G} e^{-2\Phi} \left(\mathcal{R} - \frac{1}{12} H_{\mu\nu\lambda} H^{\mu\nu\lambda} + 4\partial_\mu \Phi \partial^\mu \Phi - \frac{2(D - 26)}{3\alpha'} \right)$$

This sounds quite exciting. Can we really get string theory living in $D = 4$ dimensions so easily? Well, yes and no. Firstly, with this extra potential term, flat D -dimensional Minkowski space no longer solves the equations of motion. This is in agreement with the analysis in Section 2 where we showed that full Lorentz invariance was preserved only in $D = 26$.

Another, technical, problem with solving the string equations of motion this way is that we're playing tree-level term off against a one-loop term. But if tree-level and one-loop terms are comparable, then typically all higher loop contributions will be as well and it is likely that we can't trust our analysis.

The Linear Dilaton CFT

In fact, there is one simple solution to (7.28) which we can trust. It is the solution to

$$\partial_\mu \Phi \partial^\mu \Phi = \frac{26 - D}{6\alpha'}$$

Recall that we're working in signature $(-, +, +, \dots)$, meaning that Φ takes a spacelike profile if $D < 26$ and a timelike profile if $D > 26$,

$$\begin{aligned} \Phi &= \sqrt{\frac{26 - D}{6\alpha'}} X^1 & D < 26 \\ \Phi &= \sqrt{\frac{D - 26}{6\alpha'}} X^0 & D > 26 \end{aligned}$$

This gives a dilaton which is linear in one direction. This can be compared to the study of the path integral for non-critical strings that we saw in 5.3.2. There are two ways of seeing the same physics.

The reason that we can trust this solution is that there is an exact CFT underlying it which we can analyze to all orders in α' . It's called, for obvious reasons, the *linear dilaton CFT*. Let's now look at this in more detail.

Firstly, consider the worldsheet action associated to the dilaton coupling. For now we'll consider an arbitrary dilaton profile $\Phi(X)$,

$$S_{\text{dilaton}} = \frac{1}{4\pi} \int d^2\sigma \sqrt{g} \Phi(X) R^{(2)} \quad (7.29)$$

Although this term vanishes on a flat worldsheet, it nonetheless changes the stress-energy tensor $T_{\alpha\beta}$ because this is defined as

$$T_{\alpha\beta} = -4\pi \left. \frac{\partial S}{\partial g^{\alpha\beta}} \right|_{g_{\alpha\beta}=\delta_{\alpha\beta}}$$

The variation of (7.29) is straightforward. Indeed, the term is akin to the Einstein-Hilbert term in general relativity but things are simpler in 2d because, for example $R_{\alpha\beta} = \frac{1}{2} g_{\alpha\beta} R$. We have

$$\delta(\sqrt{g} g^{\alpha\beta} R_{\alpha\beta}) = \sqrt{g} g^{\alpha\beta} \delta R_{\alpha\beta} = \sqrt{g} \nabla^\alpha v_\alpha$$

where

$$v_\alpha = \nabla^\beta \delta g_{\alpha\beta} - g^{\gamma\delta} \nabla_\alpha \delta g_{\gamma\delta}$$

Using this, the variation of the dilaton term in the action is given by

$$\delta S_{\text{dilaton}} = \frac{1}{4\pi} \int d^2\sigma \sqrt{g} (\nabla^\alpha \nabla^\beta \Phi - \nabla^2 \Phi g^{\alpha\beta}) \delta g_{\alpha\beta}$$

which, restricting to flat space $g_{\alpha\beta} = \delta_{\alpha\beta}$, finally gives us the stress-energy tensor of a theory with dilaton coupling

$$T_{\alpha\beta}^{\text{dilaton}} = -\partial_\alpha \partial_\beta \Phi + \partial^2 \Phi \delta_{\alpha\beta}$$

Note that this stress tensor is not traceless. This is to be expected because, as we described above, the dilaton coupling is not Weyl invariant at tree-level. In complex coordinates, the stress tensor is

$$T^{\text{dilaton}} = -\partial^2 \Phi \quad , \quad \bar{T}^{\text{dilaton}} = -\bar{\partial}^2 \Phi$$

Linear Dilaton OPE

The stress tensor above holds for any dilaton profile $\Phi(X)$. Let's now restrict to a linear dilaton profile for a single scalar field X ,

$$\Phi = QX$$

where Q is some constant. We also include the standard kinetic terms for D scalar fields, of which X is a chosen one, giving the stress tensor

$$T = -\frac{1}{\alpha'} : \partial X \partial X : -Q \partial^2 X$$

It is a simple matter to compute the TT OPE using the techniques described in Section 4. We find,

$$T(z) T(w) = \frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w} + \dots$$

where the central charge of the theory is given by

$$c = D + 6\alpha' Q^2$$

Note that Q^2 can be positive or negative depending on the whether we have a timelike or spacelike linear dilaton. In this way, we see explicitly how a linear dilaton gradient can absorb central charge.

7.4.5 The Elephant in the Room: The Tachyon

We've been waxing lyrical about the details of solutions to the low-energy effective action, all the while ignoring the most important, relevant field of them all: the tachyon. Since our vacuum is unstable, this is a little like describing all the beautiful pictures we could paint if only that damn paintbrush would balance, unaided, on its tip.

Of course, the main reason for discussing these solutions is that they all carry directly over to the superstring where the tachyon is absent. Nonetheless, it's interesting to ask what happens if the tachyon is turned on. Its vertex operator is simply

$$V_{\text{tachyon}} \sim \int d^2\sigma \sqrt{g} e^{ip \cdot X}$$

where $p^2 = 4/\alpha'$. Piecing together a general tachyon profile $V(X)$ from these Fourier modes and exponentiating, results in a potential on the worldsheet of the string

$$S_{\text{potential}} = \int d^2\sigma \sqrt{g} \alpha' V(X)$$

This is a relevant operator for the worldsheet CFT. Whenever such a relevant operator turns on, we should follow the RG flow to the infra-red until we land on another CFT. The c-theorem tells us that $c_{IR} < c_{UV}$, but in string theory we always require $c = 26$. The deficit, at least initially, is soaked up by the dilaton in the manner described above. The end point of the tachyon RG flow for the bosonic string is not understood. It may be that there is no end point and the bosonic string simply doesn't make sense once the tachyon is turned on. Or perhaps we haven't yet understood the true ground state of the bosonic string.

7.5 D-Branes Revisited: Background Gauge Fields

Understanding the constraints of conformal invariance on the closed string backgrounds led us to Einstein's equations and the low-energy effective action in spacetime. Now we would like to do the same for the open string. We want to understand the restrictions that consistency places on the dynamics of D-branes.

We saw in Section 3 that there are two types of massless modes that arise from the quantization of an open string: scalars, corresponding to the fluctuation of the D-brane, and a $U(1)$ gauge field. We will ignore the scalar fluctuations for now, but will return to them later. We focus initially on the dynamics of a gauge field A_a , $a = 0, \dots, p$ living on a Dp-brane

The first question that we ask is: how does the end of the string react to a background gauge field? To answer this, we need to look at the vertex operator associated to the photon. It was given in (5.10)

$$V_{\text{photon}} \sim \int_{\partial\mathcal{M}} d\tau \zeta_a \partial^\tau X^a e^{ip \cdot X}$$

which is Weyl invariant and primary only if $p^2 = 0$ and $p^a \zeta_a = 0$. Exponentiating this vertex operator, as described at the beginning of Section 7, gives the coupling of the open string to a general background gauge field $A_a(X)$,

$$S_{\text{end-point}} = \int_{\partial\mathcal{M}} d\tau A_a(X) \frac{dX^a}{d\tau}$$

But this is a very familiar coupling — we've already mentioned it in (7.9). It is telling us that the end of the string is charged under the background gauge field A_a on the brane.

7.5.1 The Beta Function

We can now perform the same type of beta function calculation that we saw for the closed string⁹. To do this, it's useful to first use conformal invariance to map the open string worldsheet to the Euclidean upper-half plane as we described in Section 4.7. The action describing an open string propagating in flat space, with its ends subject to a background gauge field on the D-brane splits up into two pieces

$$S = S_{\text{Neumann}} + S_{\text{Dirichlet}}$$

where S_{Neumann} describes the fluctuations parallel to the Dp-brane and is given by

$$S_{\text{Neumann}} = \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \partial^\alpha X^a \partial_\alpha X^b \delta_{ab} + i \int_{\partial\mathcal{M}} d\tau A_a(X) \dot{X}^a \quad (7.30)$$

Here $a, b = 0, \dots, p$. The extra factor of i arises because we are in Euclidean space. Meanwhile, the fields transverse to the brane have Dirichlet boundary conditions and take range $I = p+1, \dots, D-1$. Their dynamics is given by

$$S_{\text{Dirichlet}} = \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \partial^\alpha X^I \partial_\alpha X^J \delta_{IJ}$$

⁹We'll be fairly explicit here, but if you want to see more details then the best place to look is the original paper by Abouelsaood, Callan, Nappi and Yost, “*Open Strings in Background Gauge Fields*”, Nucl. Phys. B280 (1987) 599.

The action $S_{\text{Dirichlet}}$ describes free fields and doesn't play any role in the computation of the beta-function. The interesting part is S_{Neumann} which, for non-zero $A_a(X)$, is an interacting quantum field theory with boundary. Our task is to compute the beta function associated to the coupling $A_a(X)$. We use the same kind of technique that we earlier applied to the closed string. We expand the fields $X^a(\sigma)$ as

$$X^a(\sigma) = \bar{x}^a(\sigma) + \sqrt{\alpha'} Y^a(\sigma)$$

where $\bar{x}^a(\sigma)$ is taken to be some fixed background which obeys the classical equations of motion,

$$\partial^2 \bar{x}^a = 0$$

(In the analogous calculation for the closed string we chose the special case of \bar{x}^a constant. Here we are more general). However, we also need to impose boundary conditions for this classical solution. In the absence of the gauge field A_a , we require Neumann boundary conditions $\partial_\sigma X^a = 0$ at $\sigma = 0$. However, the presence of the gauge field changes this. Varying the full action (7.30) shows that the relevant boundary condition is supplemented by an extra term,

$$\partial_\sigma \bar{x}^a + 2\pi\alpha' i F^{ab} \partial_\tau \bar{x}_b = 0 \quad \text{at } \sigma = 0 \quad (7.31)$$

where the F_{ab} is the field strength

$$F_{ab}(X) = \frac{\partial A_b}{\partial X^a} - \frac{\partial A_a}{\partial X^b} \equiv \partial_a A_b - \partial_b A_a$$

The fields $Y^a(\sigma)$ are the fluctuations which are taken to be small. Again, the presence of $\sqrt{\alpha'}$ in the expansion ensures that Y^a are dimensionless. Expanding the action S_{Neumann} (which we'll just call S from now on) to second order in fluctuations gives,

$$\begin{aligned} S[\bar{x} + \sqrt{\alpha'} Y] &= S[\bar{x}] + \frac{1}{4\pi} \int_{\mathcal{M}} d^2\sigma \partial Y^a \partial Y^b \delta_{ab} \\ &\quad + i\alpha' \int_{\partial\mathcal{M}} d\tau \left(\partial_a A_b Y^a \dot{Y}^b + \frac{1}{2} \partial_a \partial_b A_c Y^a Y^b \dot{x}^c \right) + \dots \end{aligned}$$

where all expressions involving the background gauge fields are now evaluated on the classical solution \bar{x} . We can rearrange the boundary terms by splitting the first term up into two halves and integrating one of these pieces by parts,

$$\int d\tau (\partial_a A_b) Y^a \dot{Y}^b = \frac{1}{2} \int d\tau \partial_a A_b Y^a \dot{Y}^b - \partial_a A_b \dot{Y}^a Y^b - \partial_c \partial_a A_b Y^a Y^b \dot{x}^c$$

Combining this with the second term means that we can write all interactions in terms of the gauge invariant field strength F_{ab} ,

$$S[\bar{x} + \sqrt{\alpha'} Y] = S[\bar{x}] + \frac{1}{4\pi} \int_{\mathcal{M}} d^2\sigma \partial Y^a \partial Y^b \delta_{ab} + \frac{i\alpha'}{2} \int_{\partial\mathcal{M}} d\tau \left(F_{ab} Y^a \dot{Y}^b + \partial_b F_{ac} Y^a Y^b \dot{x}^c \right) + \dots \quad (7.32)$$

where the $+ \dots$ refer to the higher terms in the expansion which come with higher derivatives of F_{ab} , accompanied by powers of α' . We can neglect them for the purposes of computing the one-loop beta function.

The Propagator

This Lagrangian describes our interacting boundary theory to leading order. We can now use this to compute the beta function. Firstly, we should determine where possible divergences arise. The offending term is the last one in (7.32). This will lead to a divergence when the fluctuation fields Y^a are contracted with their propagator

$$\langle Y^a(z, \bar{z}) Y^b(w, \bar{w}) \rangle = G^{ab}(z, \bar{z}; w, \bar{w})$$

We should be used to these free field Green's functions by now. The propagator satisfies

$$\partial \bar{\partial} G^{ab}(z, \bar{z}) = -2\pi \delta^{ab} \delta(z, \bar{z}) \quad (7.33)$$

in the upper half plane. But now there's a subtlety. The Y^a fields need to satisfy a boundary condition at $\text{Im } z = 0$ and this should be reflected in the boundary condition for the propagator. We discussed this briefly for Neumann boundary conditions in Section 4.7. But we've also seen that the background field strength shifts the Neumann boundary conditions to (7.31). Correspondingly, the propagator $G(z, \bar{z}; w, \bar{w})$ must now satisfy

$$\partial_\sigma G^{ab}(z, \bar{z}; w, \bar{w}) + 2\pi\alpha' i F_c^a \partial_\tau G^{cb}(z, \bar{z}; w, \bar{w}) = 0 \quad \text{at } \sigma = 0 \quad (7.34)$$

In Section 4.7, we showed how Neumann boundary conditions could be imposed by considering an image charge in the lower half plane. A similar method works here. We extend $G^{ab} \equiv G^{ab}(z, \bar{z}; w, \bar{w})$ to the entire complex plane. The solution to (7.33) subject to (7.34) is given by

$$G^{ab} = -\delta^{ab} \ln |z - w| - \frac{1}{2} \left(\frac{1 - 2\pi\alpha' F}{1 + 2\pi\alpha' F} \right)^{ab} \ln(z - \bar{w}) - \frac{1}{2} \left(\frac{1 + 2\pi\alpha' F}{1 - 2\pi\alpha' F} \right)^{ab} \ln(\bar{z} - w)$$

The Counterterm and Beta Function

Let's now return to the interacting theory (7.32) and see what counterterm is needed to remove the divergence. Since all interactions take place on the boundary, we should evaluate our propagator on the boundary, which means $z = \bar{z}$ and $w = \bar{w}$. In this case, all the logarithms become the same and, in the limit that $z \rightarrow w$, gives the leading divergence $\ln|z - w| \rightarrow \epsilon^{-1}$. We learn that the UV divergence takes the form,

$$-\frac{1}{\epsilon} \left[\delta^{ab} + \frac{1}{2} \left(\frac{1 - 2\pi\alpha' F}{1 + 2\pi\alpha' F} \right)^{ab} + \frac{1}{2} \left(\frac{1 + 2\pi\alpha' F}{1 - 2\pi\alpha' F} \right)^{ab} \right] = -\frac{2}{\epsilon} \left(\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right)^{ab}$$

It's now easy to determine the necessary counterterm. We simply replace $Y^a Y^b$ in the final term with $\langle Y^a Y^b \rangle$. This yields

$$-\frac{i2\pi\alpha'^2}{\epsilon} \int_{\partial\mathcal{M}} d\tau \partial_b F_{ac} \left[\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right]^{ab} \dot{x}^c$$

For the open string theory to retain conformal invariance, we need the associated beta function to vanish. This gives us the condition on the field strength F_{ab} : it must satisfy the equation

$$\partial_b F_{ac} \left[\frac{1}{1 - 4\pi^2\alpha'^2 F^2} \right]^{ab} = 0 \quad (7.35)$$

This is our final equation governing the equations of motion that F_{ab} must satisfy to provide a consistent background for open string propagation.

7.5.2 The Born-Infeld Action

Equation (7.35) probably doesn't look too familiar! Following the path we took for the closed string, we wish to write down an action whose equations of motion coincide with (7.35). The relevant action was actually constructed many decades ago as a non-linear alternative to Maxwell theory: it goes by the name of the *Born-Infeld action*:

$$S = -T_p \int d^{p+1}\xi \sqrt{-\det(\eta_{ab} + 2\pi\alpha' F_{ab})} \quad (7.36)$$

Here ξ are the worldvolume coordinates on the brane and T_p is the tension of the D p -brane (which, since it multiplies the action, doesn't affect the equations of motion). The gauge potential is to be thought of as a function of the worldvolume coordinates: $A_a = A_a(\xi)$. It actually takes a little work to show that the equations of motion that we derive from this action coincide with the vanishing of the beta function (7.35). Some hints on how to proceed are provided on Example Sheet 4.

For small field strengths, $F_{ab} \ll 1/\alpha'$, the action (7.36) coincides with Maxwell's action. To see this, we need simply expand to get

$$S = -T_p \int d^{p+1}\xi \left(1 + \frac{(2\pi\alpha')^2}{4} F_{ab}F^{ab} + \dots \right)$$

The leading order term, quadratic in field strengths, is the Maxwell action. Terms with higher powers of F_{ab} are suppressed by powers of α' .

So, for small field strengths, the dynamics of the gauge field on a D-brane is governed by Maxwell's equations. However, as the electric and magnetic field strengths increase and become of order $1/\alpha'$, non-linear corrections to the dynamics kick in and are captured by the Born-Infeld action.

The Born-Infeld action arises from the one-loop beta function. It is the exact result for constant field strengths. If we want to understand the dynamics of gauge fields with large gradients, ∂F , then we will have determine the higher loop contributions to the beta function.

7.6 The DBI Action

We've understood that the dynamics of gauge fields on the brane is governed by the Born-Infeld action. But what about the fluctuations of the brane itself. We looked at this briefly in Section 3.2 and suggested, on general grounds, that the action should take the Dirac form (3.6). It would be nice to show this directly by considering the beta function equations for the scalar fields ϕ^I on the brane. Turning these on corresponds to considering boundary conditions where the brane is bent. It is indeed possible to compute something along the lines of beta-function equations and to show directly that the fluctuations of the brane are governed by the Dirac action¹⁰.

More generally, one could consider both the dynamics of the gauge field and the fluctuation of the brane. This is governed by a mixture of the Dirac action and the Born-Infeld action which is usually referred to as the *DBI action*,

$$S_{DBI} = -T_p \int d^{p+1}\xi \sqrt{-\det(\gamma_{ab} + 2\pi\alpha' F_{ab})}$$

As in Section (3.2), γ_{ab} is the pull-back of the spacetime metric onto the worldvolume,

$$\gamma_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} \eta_{\mu\nu}$$

¹⁰A readable discussion of this calculation can be found in the original paper by Leigh, *Dirac-Born-Infeld Action from Dirichlet Sigma Model*, Mod. Phys. Lett. A4: 2767 (1989).

The new dynamical fields in this action are the embedding coordinates $X^\mu(\xi)$, with $\mu = 0, \dots, D - 1$. This appears to be D new degrees of freedom while we expect only $D - p - 1$ transverse physical degrees of freedom. The resolution to this should be familiar by now: the DBI action enjoys a reparameterization invariance which removes the longitudinal fluctuations of the brane.

We can use this reparameterization invariance to work in static gauge. For an infinite, flat D p -brane, it is useful to set

$$X^a = \xi^a \quad a = 0, \dots, p$$

so that the pull-back metric depends only on the transverse fluctuations X^I ,

$$\gamma_{ab} = \eta_{ab} + \frac{\partial X^I}{\partial \xi^a} \frac{\partial X^J}{\partial \xi^b} \delta_{IJ}$$

If we are interested in situations with small field strengths F_{ab} and small derivatives $\partial_a X$, then we can expand the DBI action to leading order. We have

$$S = -(2\pi\alpha')^2 T_p \int d^{p+1}\xi \left(\frac{1}{4} F_{ab} F^{ab} + \frac{1}{2} \partial_a \phi^I \partial^a \phi^I + \dots \right)$$

where we have rescaled the positions to define the scalar fields $\phi^I = X^I / 2\pi\alpha'$. We have also dropped an overall constant term in the action. This is simply free Maxwell theory coupled to free massless scalar fields ϕ^I . The higher order terms that we have dropped are all suppressed by powers of α' .

7.6.1 Coupling to Closed String Fields

The DBI action describes the low-energy dynamics of a D p -brane in flat space. We could now ask how the motion of the D-brane is affected if it moves in a background created by closed string modes $G_{\mu\nu}$, $B_{\mu\nu}$ and Φ . Rather than derive this, we'll simply write down the answer and then justify each term in turn. The answer is:

$$S_{DBI} = -T_p \int d^{p+1}\xi e^{-\tilde{\Phi}} \sqrt{-\det(\gamma_{ab} + 2\pi\alpha' F_{ab} + B_{ab})}$$

Let's start with the coupling to the background metric $G_{\mu\nu}$. It's actually hidden in the notation in this expression: it appears in the pull-back metric γ_{ab} which is now given by

$$\gamma_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} G_{\mu\nu}$$

It should be clear that this is indeed the natural place for it to sit.

Next up is the dilaton. As in (7.17), we have decomposed the dilaton into a constant piece and a varying piece: $\Phi = \Phi_0 + \tilde{\Phi}$. The constant piece governs the asymptotic string coupling, $g_s = e^{\Phi_0}$, and is implicitly sitting in front of the action because the tension of the D-brane scales as

$$T_p \sim 1/g_s$$

This, then, explains the factor of $e^{-\tilde{\Phi}}$ in front of the action: it simply reunites the varying part of the dilaton with the constant piece. Physically, it's telling us that the tension of the D-brane depends on the local value of the dilaton field, rather than its asymptotic value. If the dilaton varies, the effective string coupling at a point X in spacetime is given by $g_s^{eff} = e^{\Phi(X)} = g_s e^{\tilde{\Phi}(X)}$. This, in turn, changes the tension of the D-brane. It can lower its tension by moving to regions with larger g_s^{eff} .

Finally, let's turn to the $B_{\mu\nu}$ field. This is a 2-form in spacetime. The function B_{ab} appearing in the DBI action is the pull-back to the worldvolume

$$B_{ab} = \frac{\partial X^\mu}{\partial \xi^a} \frac{\partial X^\nu}{\partial \xi^b} B_{\mu\nu}$$

Its appearance in the DBI action is actually required on grounds of gauge invariance alone. This can be seen by considering an open string, moving in the presence of both a background $B_{\mu\nu}(X)$ in spacetime and a background $A_a(X)$ on the worldvolume of a brane. The relevant terms on the string worldsheet are

$$\frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu B_{\mu\nu} + \int_{\partial\mathcal{M}} d\tau A_a \dot{X}^a$$

Under a spacetime gauge transformation

$$B_{\mu\nu} \rightarrow B_{\mu\nu} + \partial_\mu C_\nu - \partial_\nu C_\mu \quad (7.37)$$

the first term changes by a total derivative. This is fine for a closed string, but it doesn't leave the action invariant for an open string because we pick up the boundary term. Let's quickly look at what we get in more detail. Under the gauge transformation (7.37), we have

$$\begin{aligned} S_B &= \frac{1}{4\pi\alpha'} \int_{\mathcal{M}} d^2\sigma \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu B_{\mu\nu} \\ &\longrightarrow S_B + \frac{1}{2\pi\alpha'} \int_{\mathcal{M}} d\sigma d\tau \epsilon^{\alpha\beta} \partial_\alpha X^\mu \partial_\beta X^\nu \partial_\mu C_\nu \\ &= S_B + \frac{1}{2\pi\alpha'} \int_{\mathcal{M}} d\sigma d\tau \epsilon^{\alpha\beta} \partial_\alpha (\partial_\beta X^\nu C_\nu) \\ &= S_B + \frac{1}{2\pi\alpha'} \int_{\partial\mathcal{M}} d\tau \dot{X}^\nu C_\nu = S_B + \frac{1}{2\pi\alpha'} \int_{\partial\mathcal{M}} d\tau \dot{X}^a C_a \end{aligned}$$

where, in the last line, we have replaced the sum over all directions X^ν with the sum over those directions obeying Neumann boundary conditions X^a , since $\dot{X}^I = 0$ at the end-points for any directions with Dirichlet boundary conditions.

The result of this short calculation is to see that the string action is not invariant under (7.37). To restore this spacetime gauge invariance, this boundary contribution must be canceled by an appropriate shift of A_a in the second term,

$$A_a \rightarrow A_a - \frac{1}{2\pi\alpha'} C_a \quad (7.38)$$

Note that this is not the usual kind of gauge transformation that we consider in electrodynamics. In particular, the field strength F_{ab} is not invariant. Rather, the gauge invariant combination under (7.37) and (7.38) is

$$B_{ab} + 2\pi\alpha' F_{ab}$$

This is the reason that this combination must appear in the DBI action. This is also related to an important physical effect. We have already seen that the string in spacetime is charged under $B_{\mu\nu}$. But we've also seen that the end of the string is charged under the gauge field A_a on the D-brane. This means that the open string deposits B charge on the brane, where it is converted into A charge. The fact that the gauge invariant field strength involves a combination of both F_{ab} and B_{ab} is related to this interplay of charges.

7.7 The Yang-Mills Action

Finally, let's consider the case of N coincident D-branes. We discussed this in Section 3.3 where we showed that the massless fields on the brane could be naturally packaged as $N \times N$ Hermitian matrices, with the element of the matrix telling us which brane the end points terminate on. The gauge field then takes the form

$$(A_a)_n^m$$

with $a = 0, \dots, p$ and $m, n = 1, \dots, N$. Written this way, it looks rather like a $U(N)$ gauge connection. Indeed, this is the correct interpretation. But how do we see this? Why is the gauge field describing a $U(N)$ gauge symmetry rather than, say, $U(1)^{N^2}$?

The quickest way to see that coincident branes give rise to a $U(N)$ gauge symmetry is to recall that the end point of the string is charged under the $U(1)$ gauge field that inhabits the brane it's ending on. Let's illustrate this with the simplest example. Suppose that we have two branes. The diagonal components $(A_a)_1^1$ and $(A_a)_2^2$ arise

from strings which begin and end on the same brane. Each is a $U(1)$ gauge field. What about the off-diagonal terms $(A_a)^1{}_2$ and $(A_a)^2{}_1$? These come from strings stretched between the two branes. They are again massless gauge bosons, but they are charged under the two original $U(1)$ symmetries; they carry charge $(+1, -1)$ and $(-1, +1)$ respectively. But this is precisely the structure of a $U(2)$ gauge theory, with the off-diagonal terms playing a role similar to W-bosons. In fact, the only way to make sense of massless, charged spin 1 particles is through non-Abelian gauge symmetry.

So the massless excitations of N coincident branes are a $U(N)$ gauge field $(A_a)^m{}_n$, together with scalars $(\phi^I)^m{}_n$ which transform in the adjoint representation of the $U(N)$ gauge group. We saw in Section 3 that the diagonal components $(\phi^I)^m{}_m$ have the interpretation of the transverse fluctuations of the m^{th} brane. Can we now write down an action describing the interactions of these fields?

In fact, there are several subtleties in writing down a non-Abelian generalization of the DBI action and such an action is not known (if, indeed, it makes sense at all). However, we can make progress by considering the low-energy limit, corresponding to small field strengths. The field strength in question is now the appropriate non-Abelian expression which, neglecting the matrix indices, reads

$$F_{ab} = \partial_a A_b - \partial_b A_a + i[A_a, A_b]$$

The low-energy action describing the dynamics of N coincident Dp-branes can be shown to be (neglecting an overall constant term),

$$S = -(2\pi\alpha')^2 T_p \int d^{p+1}\xi \text{ Tr} \left(\frac{1}{4} F_{ab} F^{ab} + \frac{1}{2} \mathcal{D}_a \phi^I \mathcal{D}^a \phi^I - \frac{1}{4} \sum_{I \neq J} [\phi^I, \phi^J]^2 \right) \quad (7.39)$$

We recognize the first term as the $U(N)$ Yang-Mills action. The coefficient in front of the Yang-Mills action is the coupling constant $1/g_{YM}^2$. For a Dp-brane, this is given by $\alpha'^2 T_p$, or

$$g_{YM}^2 \sim l_s^{p-3} g_s$$

The kinetic term for ϕ^I simply reflects the fact that these fields transform in the adjoint representation of the gauge group,

$$\mathcal{D}_a \phi^I = \partial_a \phi^I + i[A_a, \phi^I]$$

We won't derive this action in these lectures: the first two terms basically follow from gauge invariance alone. The potential term is harder to see directly: the quick ways to derive it use T-duality or, in the case of the superstring, supersymmetry.

A flat, infinite D p -brane breaks the Lorentz group of spacetime to

$$S(1, D - 1) \rightarrow SO(1, p) \times SO(D - p - 1) \quad (7.40)$$

This unbroken group descends to the worldvolume of the D-brane where it classifies all low-energy excitations of the D-brane. The $SO(1, p)$ is simply the Lorentz group of the D-brane worldvolume. The $SO(D - p - 1)$ is a global symmetry of the D-brane theory, rotating the scalar fields ϕ^I .

The potential term in (7.39) is particularly interesting,

$$V = -\frac{1}{4} \sum_{I \neq J} \text{Tr} [\phi^I, \phi^J]^2$$

The potential is positive semi-definite. We can look at the fields that can be turned on at no cost of energy, $V = 0$. This requires that all ϕ^I commute which means that, after a suitable gauge transformation, they take the diagonal form,

$$\phi^I = \begin{pmatrix} \phi_1^I & & \\ & \ddots & \\ & & \phi_N^I \end{pmatrix} \quad (7.41)$$

The diagonal component ϕ_n^I describes the position of the n^{th} brane in transverse space \mathbf{R}^{D-p-1} . We still need to get the dimensions right. The scalar fields have dimension $[\phi] = 1$. The relationship to the position in space (which we mentioned before in 3.2) is

$$\vec{X}_n = 2\pi\alpha' \vec{\phi}_n \quad (7.42)$$

where we've swapped to vector notation to replace the I index.

The eigenvalues ϕ_n^I are not quite gauge invariant: there is a residual gauge symmetry — the Weyl group of $U(N)$ — which leaves ϕ^I in the form (7.41) but permutes the entries by S_N , the permutation group of N elements. But this has a very natural interpretation: it is simply telling us that the D-branes are indistinguishable objects.

When all branes are separated, the vacuum expectation value (7.41) breaks the gauge group from $U(N) \rightarrow U(1)^N$. The W-bosons gain a mass M_W through the Higgs mechanism. Let's compute this mass. We'll consider a $U(2)$ theory and we'll separate

the two D-branes in the direction $X^D \equiv X$. This means that we turn on a vacuum expectation value for $\phi^D = \phi$, which we write as

$$\phi = \begin{pmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{pmatrix} \quad (7.43)$$

The values of ϕ_1 and ϕ_2 are the positions of the first and second brane. Or, more precisely, we need to multiply by the conversion factor $2\pi\alpha'$ as in (7.42) to get the position X_m of the $m = 1^{\text{st}}, 2^{\text{nd}}$ brane,

Let's compute the mass of the W-boson from the Yang-Mills action (7.39). It comes from the covariant derivative terms $\mathcal{D}\phi$. We expand out the gauge field as

$$A_a = \begin{pmatrix} A_a^{11} & W_a \\ W_a^\dagger & A_a^{22} \end{pmatrix}$$

with A^{11} and A^{22} describing the two $U(1)$ gauge fields and W the W-boson. The mass of the W-boson comes from the $[A_a, \phi]$ term inside the covariant derivative which, using the expectation value (7.43), is given by

$$\frac{1}{2} \text{Tr} [A_a, \phi]^2 = -(\phi_2 - \phi_1)^2 |W_a|^2$$

This gives us the mass of the W-boson: it is

$$M_W^2 = (\phi_2 - \phi_1)^2 = T^2 |X_2 - X_1|^2$$

where $T = 1/2\pi\alpha'$ is the tension of the string. But this has a very natural interpretation. It is precisely the mass of a string stretched between the two D-branes as shown in the figure above. We see that D-branes provide a natural geometric interpretation of the Higgs mechanism using adjoint scalars.

Notice that when branes are well separated, and the strings that stretch between them are heavy, their positions are described by the diagonal elements of the matrix given in (7.41). However, as the branes come closer together, these stretched strings become light and are important for the dynamics of the branes. Now the positions of the branes should be described by the full $N \times N$ matrices, including the off-diagonal elements. In this manner, D-branes begin to see space as something non-commutative at short distances.

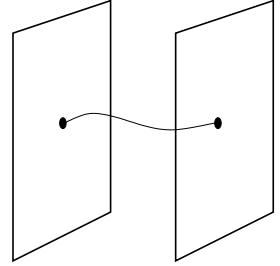


Figure 44:

In general, we can consider N D-branes located at positions \vec{X}_m , $m = 1, \dots, N$ in transverse space. The string stretched between the m^{th} and n^{th} brane has mass

$$M_W = |\vec{\phi}_n - \vec{\phi}_m| = T|\vec{X}_n - \vec{X}_m|$$

which again coincides with the mass of the appropriate W-boson computed using (7.39).

7.7.1 D-Branes in Type II Superstring Theories

As we mentioned previously, D-branes are ingredients of the Type II superstring theories. Type IIA has Dp -branes with p even, while Type IIB is home to Dp -branes with p odd. The D-branes have a very important property in these theories: they preserve half the supersymmetries.

Let's take a moment to explain what this means. We'll start by returning to the Lorentz group $SO(1, D - 1)$ now, of course, with $D = 10$. We've already seen that an infinite, flat Dp -brane is not invariant under the full Lorentz group, but only the subgroup (7.40). If we act with either $SO(1, p)$ or $SO(D - p - 1)$ then the D-brane solution remains invariant. We say that these symmetries are preserved by the solution.

However, the role of the preserved symmetries doesn't stop there. The next step is to consider small excitations of the D-brane. These must fit into representations of the preserved symmetry group (7.40). This ensures that the low-energy dynamics of the D-brane must be governed by a theory which is invariant under (7.40) and we have indeed seen that the Lagrangian (7.39) has $SO(1, p)$ as a Lorentz group and $SO(D - p - 1)$ as a global symmetry group which rotates the scalar fields.

Now let's return to supersymmetry. The Type II string theories enjoy a lot of supersymmetry: 32 supercharges in total. The infinite, flat D-branes are invariant under half of these; if we act with one half of the supersymmetry generators, the D-brane solutions don't change. Objects that have this property are often referred to as *BPS* states. Just as with the Lorentz group, these unbroken symmetries descend to the worldvolume of the D-brane. This means that the low-energy dynamics of the D-branes is described by a theory which is itself invariant under 16 supersymmetries.

There is a unique class of theories with 16 supersymmetries and a non-Abelian gauge field and matter in the adjoint representation. This class is known as maximally supersymmetric Yang-Mills theory and the bosonic part of the action is given by (7.39). Supersymmetry is realized only after the addition of fermionic fields which also live on the brane. These theories describe the low-energy dynamics of multiple D-branes.

As an illustrative example, consider D3-branes in the Type IIB theory. The theory describing N D-branes is $U(N)$ Yang-Mills with 16 supercharges, usually referred to as $U(N) \mathcal{N} = 4$ super-Yang-Mills. The bosonic part of the action is given by (7.39), where there are $D - p - 1 = 6$ scalar fields ϕ^I in the adjoint representation of the gauge group. These are augmented with four Weyl fermions, also in the adjoint representation.

8. Compactification and T-Duality

In this section, we will consider the simplest compactification of the bosonic string: a background spacetime of the form

$$\mathbf{R}^{1,24} \times \mathbf{S}^1 \quad (8.1)$$

The circle is taken to have radius R , so that the coordinate on \mathbf{S}^1 has periodicity

$$X^{25} \equiv X^{25} + 2\pi R$$

We will initially be interested in the physics at length scales $\gg R$ where motion on the \mathbf{S}^1 can be ignored. Our goal is to understand what physics looks like to an observer living in the non-compact $\mathbf{R}^{1,24}$ Minkowski space. This general idea goes by the name of *Kaluza-Klein compactification*. We will view this compactification in two ways: firstly from the perspective of the spacetime low-energy effective action and secondly from the perspective of the string worldsheet.

8.1 The View from Spacetime

Let's start with the low-energy effective action. Looking at length scales $\gg R$ means that we will take all fields to be independent of X^{25} : they are instead functions only on the non-compact $\mathbf{R}^{1,24}$.

Consider the metric in Einstein frame. This decomposes into three different fields on $\mathbf{R}^{1,24}$: a metric $\tilde{G}_{\mu\nu}$, a vector A_μ and a scalar σ which we package into the $D = 26$ dimensional metric as

$$ds^2 = \tilde{G}_{\mu\nu} dX^\mu dX^\nu + e^{2\sigma} (dX^{25} + A_\mu dX^\mu)^2 \quad (8.2)$$

Here all the indices run over the non-compact directions $\mu, \nu = 0, \dots, 24$ only.

The vector field A_μ is an honest gauge field, with the gauge symmetry descending from diffeomorphisms in $D = 26$ dimensions. To see this recall that under the transformation $\delta X^\mu = V^\mu(X)$, the metric transforms as

$$\delta G_{\mu\nu} = \nabla_\mu \Lambda_\nu + \nabla_\nu \Lambda_\mu$$

This means that diffeomorphisms of the compact direction, $\delta X^{25} = \Lambda(X^\mu)$, turn into gauge transformations of A_μ ,

$$\delta A_\mu = \partial_\mu \Lambda$$

We'd like to know how the fields $G_{\mu\nu}$, A_μ and σ interact. To determine this, we simply insert the ansatz (8.2) into the $D = 26$ Einstein-Hilbert action. The $D = 26$ Ricci scalar $\mathcal{R}^{(26)}$ is given by

$$\mathcal{R}^{(26)} = \mathcal{R} - 2e^{-\sigma}\nabla^2 e^\sigma - \frac{1}{4}e^{2\sigma}F_{\mu\nu}F^{\mu\nu}$$

where \mathcal{R} in this formula now refers to the $D = 25$ Ricci scalar. The action governing the dynamics becomes

$$S = \frac{1}{2\kappa^2} \int d^{26}X \sqrt{-\tilde{G}^{(26)}} \mathcal{R}^{(26)} = \frac{2\pi R}{2\kappa^2} \int d^{25}X \sqrt{-\tilde{G}} e^\sigma \left(\mathcal{R} - \frac{1}{4}e^{2\sigma}F_{\mu\nu}F^{\mu\nu} + \partial_\mu\sigma\partial^\mu\sigma \right)$$

The dimensional reduction of Einstein gravity in D dimensions gives Einstein gravity in $D - 1$ dimensions, coupled to a $U(1)$ gauge theory and a single massless scalar. This illustrates the original idea of Kaluza and Klein, with Maxwell theory arising naturally from higher-dimensional gravity.

The gravitational action above is not quite of the Einstein-Hilbert form. We need to again change frames, absorbing the scalar σ in the same manner as we absorbed the dilaton in Section 7.3.1. Moreover, just as for the dilaton, there is no potential dictating the vacuum expectation value of σ . Changing the vev of σ corresponds to changing R , so this is telling us that nothing in the gravitational action fixes the radius R of the compact circle. This is a problem common to all Kaluza-Klein compactifications¹¹: there are always massless scalar fields, corresponding to the volume of the internal space as well as other deformations. Massless scalar fields, such as the dilaton Φ or the volume σ , are usually referred to as *moduli*.

If we want this type of Kaluza-Klein compactification to describe our universe — where we don't see massless scalar fields — we need to find a way to “fix the moduli”. This means that we need a mechanism which gives rise to a potential for the scalar fields, making them heavy and dynamically fixing their vacuum expectation value. Such mechanisms exist in the context of the superstring.

Let's now also look at the Kaluza-Klein reduction of the other fields in the low-energy effective action. The dilaton is easy: a scalar in D dimensions reduces to a scalar in $D - 1$ dimensions. The anti-symmetric 2-form has more structure: it reduces to a 2-form $B_{\mu\nu}$, together with a vector field $\tilde{A}_\mu = B_{\mu 25}$.

¹¹The description of compactification on more general manifolds is a beautiful story involving aspects of differential geometry and topology. This story is told in the second volume of Green, Schwarz and Witten.

In summary, the low-energy physics of the bosonic string in $D - 1$ dimensions consists of a metric $G_{\mu\nu}$, two $U(1)$ gauge fields A_μ and \tilde{A}_μ and two massless scalars Φ and σ .

8.1.1 Moving around the Circle

In the above discussion, we assumed that all fields are independent of the periodic direction X^{25} . Let's now look at what happens if we relax this constraint. It's simplest to see the resulting physics if we look at the scalar field Φ where we don't have to worry about cluttering equations with indices. In general, we can expand this field in Fourier modes around the circle

$$\Phi(X^\mu; X^{25}) = \sum_{n=-\infty}^{\infty} \Phi_n(X^\mu) e^{inX^{25}/R}$$

where reality requires $\Phi_n^* = \Phi_{-n}$. Ignoring the coupling to gravity for now, the kinetic terms for this scalar are

$$\int d^{26}X \partial_\mu \Phi \partial^\mu \Phi + (\partial_{25} \Phi)^2 = 2\pi R \int d^{25}X \sum_{n=-\infty}^{\infty} \left(\partial_\mu \Phi_n \partial^\mu \Phi_{-n} + \frac{n^2}{R^2} |\Phi_n|^2 \right)$$

This simple Fourier decomposition is telling us something very important: a single scalar field on $\mathbf{R}^{1,D-1} \times \mathbf{S}^1$ splits into an infinite number of scalar fields on $\mathbf{R}^{1,D-2}$, indexed by the integer n . These have mass

$$M_n^2 = \frac{n^2}{R^2} \tag{8.3}$$

For R small, all particles are heavy except for the massless zero mode $n = 0$. The heavy particles are typically called Kaluza-Klein (KK) modes and can be ignored if we're probing energies $\ll 1/R$ or, equivalently, distance scales $\gg R$.

There is one further interesting property of the KK modes Φ_n with $n \neq 0$: they are charged under the gauge field A_μ arising from the metric. The simplest way to see this is to look at the appropriate gauge transformation which, from the spacetime perspective, is the diffeomorphism $X^{25} \rightarrow X^{25} + \Lambda(X^\mu)$. Clearly, this shifts the KK modes

$$\Phi_n \rightarrow \exp\left(\frac{in\Lambda}{R}\right) \Phi_n$$

This tells us that the n^{th} KK mode has charge n/R . In fact, one usually rescales the gauge field to $A'_\mu = A_\mu/R$, under which the charge of the KK mode Φ_n is simply $n \in \mathbf{Z}$.

8.2 The View from the Worldsheet

We now consider the Kaluza-Klein reduction from the perspective of the string. We want to study a string moving in the background $\mathbf{R}^{1,24} \times \mathbf{S}^1$. There are two ways in which the compact circle changes the string dynamics.

The first effect of the circle is that the spatial momentum, p , of the string in the circle direction can no longer take any value, but is quantized in integer units

$$p^{25} = \frac{n}{R} \quad n \in \mathbf{Z}$$

The simplest way to see this is simply to require that the string wavefunction, which includes the factor $e^{ip \cdot X}$, is single valued.

The second effect is that we can allow more general boundary conditions for the mode expansion of X . As we move around the string, we no longer need $X(\sigma + 2\pi) = X(\sigma)$, but can relax this to

$$X^{25}(\sigma + 2\pi) = X^{25}(\sigma) + 2\pi m R \quad m \in \mathbf{Z}$$

The integer m tells us how many times the string winds around \mathbf{S}^1 . It is usually simply called the *winding number*.

Let's now follow the familiar path that we described in Section 2 to study the spectrum of the string on the spacetime (8.1). We start by considering only the periodic field X^{25} , highlighting the differences with our previous treatment. The mode expansion of X^{25} is now given by

$$X^{25}(\sigma, \tau) = x^{25} + \frac{\alpha' n}{R} \tau + m R \sigma + \text{oscillator modes}$$

which incorporates both the quantized momentum and the possibility of a winding number. Before splitting $X^{25}(\sigma, \tau)$ into right-moving and left-moving parts, it will be useful to introduce the quantities

$$p_L = \frac{n}{R} + \frac{m R}{\alpha'} \quad , \quad p_R = \frac{n}{R} - \frac{m R}{\alpha'} \tag{8.4}$$

Then we have $X^{25}(\sigma, \tau) = X_L^{25}(\sigma^+) + X_R^{25}(\sigma^-)$, where

$$\begin{aligned} X_L^{25}(\sigma^+) &= \frac{1}{2} x^{25} + \frac{1}{2} \alpha' p_L \sigma^+ + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \tilde{\alpha}_n^{25} e^{-in\sigma^+} , \\ X_R^{25}(\sigma^-) &= \frac{1}{2} x^{25} + \frac{1}{2} \alpha' p_R \sigma^- + i \sqrt{\frac{\alpha'}{2}} \sum_{n \neq 0} \frac{1}{n} \alpha_n^{25} e^{-in\sigma^-} \end{aligned}$$

This differs from the mode expansion (1.36) only in the terms p_L and p_R . The mode expansion for all the other scalar fields on flat space $\mathbf{R}^{1,24}$ remains unchanged and we don't write them explicitly.

Let's think about what the spectrum of this theory looks like to an observer living in $D = 25$ non-compact directions. Each particle state will be described by a momentum p^μ with $\mu = 0, \dots, 24$. The mass of the particle is

$$M^2 = - \sum_{\mu=0}^{24} p_\mu p^\mu$$

As before, the mass of these particles is fixed in terms of the oscillator modes of the string by the L_0 and \tilde{L}_0 equations. These now read

$$M^2 = p_L^2 + \frac{4}{\alpha'} (\tilde{N} - 1) = p_R^2 + \frac{4}{\alpha'} (N - 1)$$

where N and \tilde{N} are the levels, defined in lightcone quantization by (2.24). (One should take the lightcone coordinate inside $\mathbf{R}^{1,24}$ rather than along the \mathbf{S}^1). The factors of -1 are the necessary normal ordering coefficients that we've seen in several guises in this course.

These equations differ from (2.25) by the presence of the momentum and winding terms around \mathbf{S}^1 on the right-hand side. In particular, level matching no longer tells us that $N = \tilde{N}$, but instead

$$N - \tilde{N} = nm \tag{8.5}$$

Expanding out the mass formula, we have

$$M^2 = \frac{n^2}{R^2} + \frac{m^2 R^2}{\alpha'^2} + \frac{2}{\alpha'} (N + \tilde{N} - 2) \tag{8.6}$$

The new terms in this formula have a simple interpretation. The first term tells us that a string with $n > 0$ units of momentum around the circle gains a contribution to its mass of n/R . This agrees with the result (8.3) that we found from studying the KK reduction of the spacetime theory. The second term is even easier to understand: a string which winds $m > 0$ times around the circle picks up a contribution $2\pi m R T$ to its mass, where $T = 1/2\pi\alpha'$ is the tension of the string.

8.2.1 Massless States

We now restrict attention to the massless states in $\mathbf{R}^{1,24}$. This can be achieved in the mass formula (8.6) by looking at states with zero momentum $n = 0$ and zero winding $m = 0$, obeying the level matching condition $N = \tilde{N} = 1$. The possibilities are

- $\alpha_{-1}^\mu \tilde{\alpha}_{-1}^\nu |0; p\rangle$: Under the $SO(1, 24)$ Lorentz group, these states decompose into a metric $G_{\mu\nu}$, an anti-symmetric tensor $B_{\mu\nu}$ and a scalar Φ .
- $\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} |0; p\rangle$ and $\alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu |0; p\rangle$: These are two vector fields. We can identify the sum of these $(\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} + \alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu) |0; p\rangle$ with the vector field A_μ coming from the metric and the difference $(\alpha_{-1}^\mu \tilde{\alpha}_{-1}^{25} - \alpha_{-1}^{25} \tilde{\alpha}_{-1}^\mu) |0; p\rangle$ with the vector field \tilde{A}_μ coming from the anti-symmetric field.
- $\alpha_{-1}^{25} \tilde{\alpha}_{-1}^{25} |0; p\rangle$: This is another scalar. It is identified with the scalar σ associated to the radius of \mathbf{S}^1 .

We see that the massless spectrum of the string coincides with the massless spectrum associated with the Kaluza-Klein reduction of the previous section.

8.2.2 Charged Fields

One can also check that the KK modes with $n \neq 0$ have charge n under the gauge field A_μ . We can determine the charge of a state under a given $U(1)$ by computing the 3-point function in which two legs correspond to the state of interest, while the third is the appropriate photon. We have two photons, with vertex operators given by,

$$V_\pm(p) \sim \int d^2 z \, \zeta_\mu (\partial X^\mu \bar{\partial} \bar{X}^{25} \pm \partial X^{25} \bar{\partial} \bar{X}^\mu) e^{ip \cdot X}$$

where $+$ corresponds to A_μ and $-$ to \tilde{A}_μ and we haven't been careful about the overall normalization. Meanwhile, any state can be assigned momentum n and winding m by dressing the operator with the factor $e^{ip_L X^{25}(z) + ip_R \bar{X}^{25}(\bar{z})}$. As always, it's simplest to work with the momentum and winding modes of the tachyon, whose vertex operators are of the form

$$V_{m,n}(p) \sim \int d^2 z \, e^{ip \cdot X} e^{ip_L X^{25} + ip_R \bar{X}^{25}}$$

The charge of a state is the coefficient in front of the 3-point coupling of the field and the photon,

$$\langle V_\pm(p_1) V_{m,n}(p_2) V_{-m,-n}(p_3) \rangle \sim \delta^{25}(\sum_i p_i) \zeta_\mu (p_2^\mu - p_3^\mu) (p_L \pm p_R)$$

The first few factors are merely kinematical. The interesting information is in the last factor. It is telling us that under A_μ , fields have charge $p_L + p_R \sim n/R$. This is in agreement with the Kaluza-Klein analysis that we saw before. However, it's also telling us something new: under \tilde{A}_μ , fields have charge $p_L - p_R \sim mR/\alpha'$. In other words, winding modes are charged under the gauge field that arises from the reduction of $B_{\mu\nu}$. This is not surprising: winding modes correspond to strings wrapping the circle and we saw in Section 7 that strings are electrically charged under $B_{\mu\nu}$.

8.2.3 Enhanced Gauge Symmetry

With a circle in the game, there are other ways to build massless states that don't require us to work at level $N = \tilde{N} = 1$. For example, we can set $N = \tilde{N} = 0$ and look at winding modes $m \neq 0$. The level matching condition (8.5) requires $n = 0$ and the mass of the states is

$$M^2 = \left(\frac{mR}{\alpha'} \right)^2 - \frac{4}{\alpha'}$$

and states can be massless whenever the radius takes special values $R^2 = 4\alpha'/m^2$ with $m \in \mathbf{Z}$. Similarly, we can set the winding to zero $m = 0$ and consider the KK modes of the tachyon which have mass

$$M^2 = \frac{n^2}{R^2} - \frac{4}{\alpha'}$$

which become massless when $R^2 = n^2\alpha'/4$.

However, the richest spectrum of massless states occurs when the radius takes a very special value, namely

$$R = \sqrt{\alpha'}$$

Solutions to the level matching condition (8.5) with $M^2 = 0$ are now given by

- $N = \tilde{N} = 1$ with $m = n = 0$. These give the states described above: a metric, two $U(1)$ gauge fields and two neutral scalars.
- $N = \tilde{N} = 0$ with $n = \pm 2$ and $m = 0$. These are KK modes of the tachyon field. They are scalars in spacetime with charges $(\pm 2, 0)$ under the $U(1) \times U(1)$ gauge symmetry.
- $N = \tilde{N} = 0$ with $n = 0$ and $m = \pm 2$. This is a winding mode of the tachyon field. They are scalars in spacetime with charges $(0, \pm 2)$ under $U(1) \times U(1)$.

- $N = 1$ and $\tilde{N} = 0$ with $n = m = \pm 1$. These are two new spin 1 fields, $\alpha_{-1}^\mu |0; p\rangle$. They carry charge $(\pm 1, \pm 1)$ under the two $U(1) \times U(1)$.
- $N = 1$ and $\tilde{N} = 0$ with $n = -m = \pm 1$. These are a further two spin 1 fields, $\tilde{\alpha}_{-1}^\mu |0; p\rangle$, with charge $(\pm 1, \mp 1)$ under $U(1) \times U(1)$.

How do we interpret these new massless states? Let's firstly look at the spin 1 fields. These are charged under $U(1) \times U(1)$. As we mentioned in Section 7.7, the only way to make sense of charged massless spin 1 fields is in terms of a non-Abelian gauge symmetry. Looking at the charges, we see that at the critical radius $R = \sqrt{\alpha'}$, the theory develops an enhanced gauge symmetry

$$U(1) \times U(1) \rightarrow SU(2) \times SU(2)$$

The massless scalars from the $N = \tilde{N} = 0$ now join with the previous scalars to form adjoint representations of this new symmetry. We move away from the critical radius by changing the vacuum expectation value for σ . This breaks the gauge group back to the Cartan subalgebra by the Higgs mechanism.

From the discussion above, it's clear that this mechanism for generating non-Abelian gauge symmetries relies on the existence of the tachyon. For this reason, this mechanism doesn't work in Type II superstring theories. However, it turns out that it does work in the heterotic string, even though it has no tachyon in its spectrum.

8.3 Why Big Circles are the Same as Small Circles

The formula (8.6) has a rather remarkable property: it is invariant under the exchange

$$R \leftrightarrow \frac{\alpha'}{R} \tag{8.7}$$

if, at the same time, we swap the quantum numbers

$$m \leftrightarrow n \tag{8.8}$$

This means that a string moving on a circle of radius R has the same spectrum as a string moving on a circle of radius α'/R . It achieves this feat by exchanging what it means to wind with that it means to move.

As the radius of the circle becomes large, $R \rightarrow \infty$, the winding modes become very heavy with mass $\sim R/\alpha'$ and are irrelevant for the low-energy dynamics. But the momentum modes become very light, $M \sim 1/R$, and, in the strict limit form a continuum. From the perspective of the energy spectrum, this continuum of energy states is exactly what we mean by the existence of a non-compact direction in space.

In the other limit, $R \rightarrow 0$, the momentum modes become heavy and can be ignored: it takes way too much energy to get anything to move on the \mathbf{S}^1 . In contrast, the winding modes become light and start to form a continuum. The resulting energy spectrum looks as if another dimension of space is opening up!

The equivalence of the string spectrum on circles of radii R and α'/R extends to the full conformal field theory and hence to string interactions. Strings are unable to tell the difference between circles that are very large and circles that are very small. This striking statement has a rubbish name: it is called *T-duality*.

This provides another mechanism in which string theory exhibits a minimum length scale: as you shrink a circle to smaller and smaller sizes, at $R = \sqrt{\alpha'}$, the theory acts as if the circle is growing again, with winding modes playing the role of momentum modes.

The New Direction in Spacetime

So how do we describe this strange new spatial direction that opens up as $R \rightarrow 0$? Under the exchange (8.7) and (8.8), we see that p_L and p_R transform as

$$p_L \rightarrow p_L , \quad p_R \rightarrow -p_R$$

Motivated by this, we define a new scalar field,

$$Y^{25} = X_L^{25}(\sigma^+) - X_R^{25}(\sigma^-)$$

It is simple to check that in the CFT for a free, compact scalar field all OPEs of Y^{25} coincide with the OPEs of X^{25} . This is sufficient to ensure that all interactions defined in the CFT are the same.

We can write the new spatial direction Y directly in terms of the old field X , without first doing the split into left and right-moving pieces. From the definition of Y , one can check that $\partial_\tau X = \partial_\sigma Y$ and $\partial_\sigma X = \partial_\tau Y$. We can write this in a unified way as

$$\partial_\alpha X = \epsilon_{\alpha\beta} \partial^\beta Y \tag{8.9}$$

where $\epsilon_{\alpha\beta}$ is the antisymmetric matrix with $\epsilon_{\tau\sigma} = -\epsilon_{\sigma\tau} = +1$. (The minus sign from $\epsilon_{\sigma\tau}$ in the above equation is canceled by another from the Minkowski worldsheet metric when we lower the index on ∂^β).

The Shift of the Dilaton

The dilaton, or string coupling, also transforms under T-duality. Here we won't derive this in detail, but just give a plausible explanation for why it's the case. The main idea is that a scientist living in a stringy world shouldn't be able to do any experiments that distinguish between a compact circle of radius R and one of radius α'/R . But the first place you would look is simply the low-energy effective action which, working in Einstein frame, contains terms like

$$\frac{2\pi R}{2l_s^{24}g_s^2} \int d^{25}X \sqrt{-\tilde{G}} e^\sigma \mathcal{R} + \dots$$

A scientist cannot tell the difference between R and $\tilde{R} = \alpha'/R$ only if the value of the dilaton is also ambiguous so that the term in front of the action remains invariant: i.e. $R/g_s^2 = \tilde{R}/\tilde{g}_s^2$. This means that, under T-duality, the dilaton must shift so that the coupling constant becomes

$$g_s \rightarrow \tilde{g}_s = \frac{\sqrt{\alpha'} g_s}{R} \quad (8.10)$$

8.3.1 A Path Integral Derivation of T-Duality

There's a simple way to see T-duality of the quantum theory using the path integral. We'll consider just a single periodic scalar field $X \equiv X + 2\pi R$ on the worldsheet. It's useful to change normalization and write $X = R\varphi$, so that the field φ has periodicity 2π . The radius R of the circle now sits in front of the action,

$$S[\varphi] = \frac{R^2}{4\pi\alpha'} \int d^2\sigma \partial_\alpha \varphi \partial^\alpha \varphi \quad (8.11)$$

The Euclidean partition function for this theory is $Z = \int \mathcal{D}\varphi e^{-S[\varphi]}$. We will now play around with this partition function and show that we can rewrite it in terms of new variables that describe the T-dual circle.

The theory (8.11) has a simple shift symmetry $\varphi \rightarrow \varphi + \lambda$. The first step is to make this symmetry local by introducing a gauge field A_α on the worldsheet which transforms as $A_\alpha \rightarrow A_\alpha - \partial_\alpha \lambda$. We then replace the ordinary derivatives with covariant derivatives

$$\partial_\alpha \varphi \rightarrow \mathcal{D}_\alpha \varphi = \partial_\alpha \varphi + A_\alpha$$

This changes our theory. However, we can return to the original theory by adding a new field, θ which couples as

$$S[\varphi, \theta, A] = \frac{R^2}{4\pi\alpha'} \int d^2\sigma \mathcal{D}_\alpha \varphi \mathcal{D}^\alpha \varphi + \frac{i}{2\pi} \int d^2\sigma \theta \epsilon^{\alpha\beta} \partial_\alpha A_\beta \quad (8.12)$$

The new field θ acts as a Lagrange multiplier. Integrating out θ sets $\epsilon^{\alpha\beta}\partial_\alpha A_\beta = 0$. If the worldsheet is topologically \mathbf{R}^2 , then this condition ensures that A_α is pure gauge which, in turn, means that we can pick a gauge such that $A_\alpha = 0$. The quantum theory described by (8.12) is then equivalent to that given by (8.11).

Of course, if the worldsheet is topologically \mathbf{R}^2 then we're missing the interesting physics associated to strings winding around φ . On a non-trivial worldsheet, the condition $\epsilon^{\alpha\beta}\partial_\alpha A_\beta = 0$ does not mean that A_α is pure gauge. Instead, the gauge field can have non-trivial holonomy around the cycles of the worldsheet. One can show that these holonomies are gauge trivial if θ has periodicity 2π . In this case, the partition function defined by (8.12),

$$Z = \frac{1}{\text{Vol}} \int \mathcal{D}\varphi \mathcal{D}\theta \mathcal{D}A e^{-S[\varphi, \theta, A]}$$

is equivalent to the partition function constructed from (8.11) for worldsheets of any topology.

At this stage, we make use of a clever and ubiquitous trick: we reverse the order of integration. We start by integrating out φ which we can do by simply fixing the gauge symmetry so that $\varphi = 0$. The path integral then becomes

$$Z = \int \mathcal{D}\theta \mathcal{D}A \exp \left(-\frac{R^2}{4\pi\alpha'} \int d^2\sigma A_\alpha A^\alpha + \frac{i}{2\pi} \int d^2\sigma \epsilon^{\alpha\beta} (\partial_\alpha \theta) A_\beta \right)$$

where we have also taken the opportunity to integrate the last term by parts. We can now complete the procedure and integrate out A_α . We get

$$Z = \int \mathcal{D}\theta \exp \left(-\frac{\tilde{R}^2}{4\pi\alpha'} \int d^2\sigma \partial_\alpha \theta \partial^\alpha \theta \right)$$

with $\tilde{R} = \alpha'/R$ the radius of the T-dual circle. In the final integration, we threw away the overall factor in the path integral, which is proportional to $\sqrt{\alpha'}/R$. A more careful treatment shows that this gives rise to the appropriate shift in the dilaton (8.10).

8.3.2 T-Duality for Open Strings

What happens to open strings and D-branes under T-duality? Suppose firstly that we compactify a circle in direction X transverse to the brane. This means that X has Dirichlet boundary conditions

$$X = \text{const} \Rightarrow \partial_\tau X^{25} = 0 \quad \text{at } \sigma = 0, \pi$$

But what happens in the T-dual direction Y ? From the definition (8.9) we learn that the new direction has Neumann boundary conditions,

$$\partial_\sigma Y = 0 \quad \text{at } \sigma = 0, \pi$$

We see that T-duality exchanges Neumann and Dirichlet boundary conditions. If we dualize a circle transverse to a Dp -brane, then it turns into a $D(p+1)$ -brane.

The same argument also works in reverse. We can start with a Dp -brane wrapped around the circle direction X , so that the string has Neumann boundary conditions. After T-duality, (8.9) changes these to Dirichlet boundary conditions and the Dp -brane turns into a $D(p-1)$ -brane, localized at some point on the circle Y .

In fact, this was how D-branes were originally discovered: by following the fate of open strings under T-duality.

8.3.3 T-Duality for Superstrings

To finish, let's nod one final time towards the superstring. It turns out that the ten-dimensional superstring theories are not invariant under T-duality. Instead, they map into each other. More precisely, Type IIA and IIB transform into each other under T-duality. This means that Type IIA string theory on a circle of radius R is equivalent to Type IIB string theory on a circle of radius α'/R . This dovetails with the transformation of D-branes, since type IIA has Dp -branes with p even, while IIB has p odd. Similarly, the two heterotic strings transform into each other under T-duality.

8.3.4 Mirror Symmetry

The essence of T-duality is that strings get confused. Their extended nature means that they're unable to tell the difference between big circles and small circles. We can ask whether this confusion extends to more complicated manifolds. The answer is yes. The fact that strings can see different manifolds as the same is known as *mirror symmetry*.

Mirror symmetry is cleanest to state in the context of the Type II superstring, although similar behaviour also holds for the heterotic strings. The simplest example is when the worldsheet of the string is governed by a superconformal non-linear sigma-model with target space given by some Calabi-Yau manifold \mathbf{X} . The claim of mirror symmetry is that this CFT is identical to the CFT describing the string moving on a different Calabi-Yau manifold \mathbf{Y} . The topology of \mathbf{X} and \mathbf{Y} is not the same. Their Hodge diamonds are the mirror of each other; hence the name. The subject of mirror symmetry is an active area of research in geometry and provides a good example of the impact of string theory on mathematics.

8.4 Epilogue

We are now at the end of this introductory course on string theory. We began by trying to make sense of the quantum theory of a relativistic string moving in flat space. It is, admittedly, an odd place to start. But from then on we had no choices to make. The relativistic string leads us ineluctably to conformal field theory, to higher dimensions of spacetime, to Einstein’s theory of gravity at low-energies, to good UV behaviour at high-energies and to Yang-Mills theories living on branes. There are few stories in theoretical physics where such meagre input gives rise to such a rich structure.

This journey continues. There is one further ingredient that it is necessary to add: supersymmetry. Even this is in some sense not a choice, but is necessary to remove the troublesome tachyon that plagued these lectures. From there we may again blindly follow where the string leads, through anomalies (and the lack thereof) in ten dimensions, to dualities and M-theory in eleven dimensions, to mirror symmetry and moduli stabilization and black hole entropy counting and holography and the miraculous AdS/CFT correspondence.

However, the journey is far from complete. There is much about string theory that remains to be understood. This is true both of the mathematical structure of the theory and of its relationship to the world that we observe. The problems that we alluded to in Section 6.4.5 are real. Non-perturbative completions of string theory are only known in spacetimes which are asymptotically anti-de Sitter, but cosmological observations suggest that our home is not among these. In attempts to make contact with the standard models of particle physics and cosmology, we typically return to the old idea of Kaluza-Klein compactifications. Is this the right approach? Or are we missing some important and subtle conceptual ingredient? Or is the existence of this remarkable mathematical structure called string theory merely a red-herring that has nothing to do with the real world?

In the years immediately after its birth, no one knew that string theory was a theory of strings. It seems very possible that we’re currently in a similar situation. When the theory is better understood, it may have little to do with strings. We are certainly still some way from answering the simple question: what is string theory really?