

TASI Lectures: (No) Global Symmetries to Axion Physics

Matthew Reece*

*Department of Physics, Harvard University,
17 Oxford St, Cambridge, MA 02138, USA*

E-mail: mreece@g.harvard.edu

These notes are an expanded version of lectures given at the 2022 TASI summer school in Boulder, Colorado. One goal of these lecture notes is to (partially) bridge the gap between what one learns in typical introductory quantum field theory classes and what one needs to understand to follow modern developments in particle theory beyond the Standard Model. Topics covered include global and gauge symmetries, charge quantization, instantons, chiral anomalies, the Strong CP problem, axion models from 4d and from higher dimensions, the expected absence of global symmetries in quantum gravity, and some phenomenological implications thereof. If these topics seem to be at best loosely related, I hope that reading the notes will convince you otherwise. Recent developments in quantum field theory have shown that ordinary gauge theories exhibit a much wider range of (generalized) global symmetries than were previously understood, while recent work in quantum gravity has provided sharper arguments that global symmetries should not exist. This sets up an interesting tension, the resolution of which can have implications for particle physics in the real world. Axion physics is one setting in which these ideas can guide phenomenology. The TASI audience comprised particle phenomenology students whom I hoped to convince of the importance of learning more about quantum field theory and quantum gravity. These notes may also be of interest to formal theory readers seeking closer connections to real-world particle physics.

*TASI 2022 – “Ten Years After the Higgs Discovery: Particle Physics Now and Future,”
June 6 - July 1 2022
Boulder, CO USA*

*Speaker

Contents

1	Introduction	4
1.1	Outline	6
	Part One: Global and Gauge Symmetries, Charge Quantization	9
2	Differential forms: notation	9
2.1	What is a differential form?	9
2.2	The exterior derivative	10
2.3	Integration and Stokes's theorem	12
2.4	The Hodge star	13
2.5	Examples of actions in differential form notation	14
3	Ordinary global symmetries; currents and charges	15
3.1	The symmetry group $U(1)$	15
3.2	$U(1)$ charges	16
3.3	$U(1)$ conservation laws and currents	16
3.4	Symmetry operators	19
4	$U(1)$ gauge theory	21
4.1	Basic definitions; charge quantization	21
4.2	Gauge fields	22
4.3	Wilson loops, quantized magnetic flux	23
4.4	Brief aside on topology	25
4.5	Configurations with magnetic flux	26
4.5.1	Example 1: flux on a torus	26
4.5.2	Example 2: the Dirac monopole	26
4.5.3	Flux in integral cohomology	27
4.6	Electric flux quantization; the Witten effect	28
4.7	Canonically normalized gauge fields and the Ward identity	30
5	Kinetic mixing of $U(1)$s	31
	Part Two: Instantons and Chiral Anomalies	33
6	$U(1)$ fields with nonzero $\int F \wedge F$	33
7	Non-abelian gauge fields	34
8	Instantons	35
8.1	The BPST instanton solution and its properties	35
8.2	Instantons in the path integral	37
8.3	Comments on instantons	38

9 The chiral ABJ and 't Hooft anomalies	38
9.1 Introductory remarks	38
9.2 The particle production calculation	40
9.2.1 The $(1+1)$ d anomaly	40
9.2.2 Landau levels and the $(3+1)$ d anomaly	42
9.3 The triangle anomaly calculation	44
9.4 Fujikawa's path integral calculation	46
9.5 The takeaway message	46
9.6 One-loop exactness	47
9.7 Further remarks on anomalies and QCD	47
10 Axion couplings to gauge fields	48
11 Chern-Simons terms	50
11.1 Chern-Simons terms in 3d gauge theory	50
11.2 Chern-Simons masses and Stueckelberg masses	53
11.2.1 A magnetic photon mass	55
11.3 Higher Chern-Simons terms	56
Part Three: The Strong CP Problem and Axion Models	57
12 The Strong CP problem and the axion solution	57
12.1 The problem and proposed solutions	57
12.2 Axions for Strong CP: the core idea	58
12.3 Axions for Strong CP: further details	61
12.3.1 The axion-photon coupling from pion mixing	61
12.4 Terminology: pseudo-Nambu-Goldstone bosons, QCD axions, and ALPs	62
13 Classic 4d axion models; axion quality problem	63
13.1 The KSVZ model	63
13.2 DFSZ model	66
13.3 Axion quality problem	68
14 Axions from higher-dimensional gauge fields	69
14.1 Basic ingredients for extra-dimensional axions	69
14.2 Quality problem for extra-dimensional axions	71
14.3 Brief comments on higher dimensions	74
15 Perspective: axions as gauge fields	74
16 Axions in cosmology	76
Part Four: No Global Symmetries in Quantum Gravity	78
17 Proton stability: symmetry, or not?	78

18 Black holes: mini-review	81
18.1 Classical black hole solutions	81
18.2 Black hole thermodynamics	83
19 Black holes versus continuous global symmetries	85
20 Generalized global symmetries	88
20.1 p -form global symmetries	88
20.2 Breaking of p -form global symmetries	89
20.2.1 Explicit breaking	89
20.2.2 Spontaneous breaking	91
21 Global symmetries versus quantum gravity	91
22 Weak Gravity Conjecture	93
23 Extensions and refinements of the Weak Gravity Conjectures	96
23.1 Multiple gauge fields	96
23.2 Cutoffs take the form of towers of particles	97
23.3 The WGC for p -form gauge fields	101
Part Five: Phenomenological Insights from Quantum Gravity	103
24 Charge quantization	103
25 Weak Gravity Conjecture and phenomenology	104
25.1 The photon mass	104
25.2 Can a light $B - L$ gauge boson exist?	106
26 Neutrino masses	107
27 The Strong CP Problem in quantum gravity	109
27.1 The WGC for axions	110
27.2 The WGC and axion strings	110
27.3 The expected axion	113
28 Closing remarks	114

1. Introduction

These notes are extended version of lectures I gave at the TASI 2022 summer school. They cover aspects of symmetries in quantum field theory and quantum gravity, and applications to particle physics. The central phenomenological topic of these lectures is the physics of axions

and their role in solving the Strong CP problem. I emphasize some topics that are not commonly emphasized in phenomenological treatments, related to charge quantization, the axion-gauge field coupling as a special case of a Chern-Simons term, and the viewpoint that an axion is a kind of gauge field (and indeed originates, in many theories, as a mode of a higher-dimensional gauge field).

There are a number of review articles on the Strong CP problem and axion physics that get right to the details, without the lengthy review of QFT and symmetries I have here. There are reviews of generalized symmetries and the absence of such symmetries in quantum gravity, with little to say about axions. Why am I combining these topics?

The TASI audience in 2022 comprised mostly phenomenologically-oriented students. I think that it is important that particle theorists, even those who are most oriented toward experiments, have a thorough understanding of conceptual aspects of quantum field theory. Part of my aim with these lectures, then, is to try to equip hep-ph physicists with more of a common language to understand recent developments in hep-th. I hope that the lecture notes will also be of interest to hep-th readers who want to learn more about phenomenological applications.

The landscape of high-energy theory looks rather different today than it did when I was a student at TASI, in 2006. At that time, my sense is that there was a higher degree of uniformity of interests in the field. The LHC was due to turn on in the near future, and there was widespread interest in how electroweak symmetry was broken, whether the Higgs boson really existed, and whether superpartners might be found near the TeV scale. Even a number of string theorists whose careers had mostly been focused on formal topics were briefly entranced by collider physics. Since the LHC confirmed that a Standard Model-like Higgs boson exists, but has (so far) found no evidence of physics beyond the Standard Model, the field has fragmented. Astroparticle physics and cosmology have become more central, and a wider range of smaller-scale precision or high-intensity experiments are being pursued. This is healthy, as we have no way of knowing where the first discovery of new physics might arise. On the other hand, this wider range of phenomenological activity sometimes means that today's hep-ph students learn less quantum field theory than their predecessors did. Sometimes, this leads to misguided theoretical claims, and sometimes these even lead to misguided experimental proposals. It is important that we continue to ground particle theory in a solid understanding of quantum field theory.

Axion physics is ideally situated at the intersection of particle phenomenology, fundamental theory, and experiment. Axions have a longstanding bottom-up phenomenological motivation as a solution of the Strong CP problem, and naturally provide a compelling dark matter candidate. Axion dark matter will be a major target of experimental investigation in the coming decade. Furthermore, the theory of axions is very rich. They are intimately connected to the chiral anomaly. Their couplings to gauge fields are a type of Chern-Simons term. Axions and axion-like particles appear ubiquitously in string theory constructions of particle physics models, and there are reasons to think this isn't a "lamppost effect," but that their existence is required by deep principles of quantum gravity. These are all positive reasons to invest some time exploring axion physics and how it connects to various important threads running through modern quantum field theory and quantum gravity. From a different viewpoint, we should situate axion physics within this larger field theoretic context because it helps us get the physics right. In some cases, axion physics can be quite subtle, and it is important that as theorists we do not mislead experimentalists into searching for effects that are simply theoretically inconsistent.

Most importantly, I think that this is an area where the best is yet to come. I am optimistic that the future will bring experimental discoveries and deeper theoretical understanding that will allow us to infer aspects of physics at very high energy scales from such experiments. I hope that these lecture notes will set some readers on a path to participate in such exciting future developments.

1.1 Outline

These lecture notes are divided into five parts. There were only four in-person lectures at TASI. Parts four and five of these notes contain a significantly expanded treatment of the material from the fourth lecture. Throughout the notes, there are more details and added topics that could not be included in the lectures due to time constraints.

Below is a brief outline of the topics of the five parts, together with suggested reading for those wanting an alternative introduction or more depth on the topic. Many more references are given throughout the text of these notes, although this is not a comprehensive review article and does not aim to cite or survey the entirety of the literature on these topics.

- **Part One.** In the first part, we will discuss global and gauge symmetries, focusing especially on the group $U(1)$ and the quantization of charge and flux. Before diving into the physics, we provide a brief review of the notation of differential forms. Although symmetries are discussed in standard quantum field theory textbooks, many of the aspects treated here, including symmetry operators and charge quantization, are usually not. (We return to the topic of generalized symmetries and symmetry operators in Part Four.)

Further reading: Our viewpoint on symmetries is based on “Generalized Global Symmetries” by Gaiotto, Kapustin, Seiberg, and Willett, especially sections 2 through 4.2 [1]. Although many of the detailed examples discussed in that paper draw on formal QFT topics that may be unfamiliar to a particle phenomenology readership, some of the exposition of the main ideas should be readable. A very recent pedagogical review of higher-form symmetries is [2] by Pedro R. S. Gomes. Another recent review of generalized global symmetries is [3] by John McGreevy. Although it focuses on applications to condensed matter physics, it uses the language of quantum field theory and should be accessible to particle theorists. A recent Snowmass white paper [4] gives a (very) brief introduction to this topic and extensive pointers to recent related literature.

- **Part Two.** The second part begins with an exposition of the quantization of instanton number, i.e., of the integral of $\text{tr}(F \wedge F)$. This includes a discussion of the classical BPST instanton solution to the Yang-Mills equations. We then turn to the chiral anomaly, explain the conceptual difference between ABJ and ’t Hooft anomalies, and discuss how the path integral measure changes when fermion fields are rephased. We explain a conclusion that is important for axion physics, namely that the couplings of a periodic scalar field to $F \wedge F$ are quantized (an example of a Chern-Simons term). This section concludes with a rapid overview of Chern-Simons terms in general.

Further reading: Instantons (and other semiclassical physics) are discussed extensively in textbooks by Shifman [5] and by Erick Weinberg [6]. For the chiral anomaly, see sections

19.1 to 19.4 of the textbook by Peskin and Schroeder [7], chapter 20 of the recent textbook by Fradkin [8], and TASI lectures on anomalies by Jeff Harvey [9].

- **Part Three.** The third part is an introduction to the Strong CP problem and models of axions. We will see how the $\theta(x)\text{tr}(G \wedge G)$ coupling solves the Strong CP problem, and how it can arise in different types of models, including classic 4d axion models (KSVZ and DFSZ) and in qualitatively different models in which the axion arises as a mode of a higher-dimensional gauge field. In my opinion, the latter class of models is strongly favored both by bottom-up considerations (as a solution to the axion quality problem) and top-down considerations (both from examples in string theory, and from quantum gravity arguments covered in later parts). Further reading: There are earlier TASI lectures on the Strong CP problem by Michael Dine [10] and on the Strong CP problem and axions by Anson Hook [11]. Outside the TASI context, a very good review article is [12]. To fully understand the derivation of the axion potential, you need to know about the chiral Lagrangian for pion physics, which is reviewed in (for example) [13–15]. Ben Safdi’s lectures from this TASI school also discuss axions, with an emphasis on axion dark matter and its consequences for astrophysics and cosmology [16]. I do not explore in much detail the cosmology of axions (see [17] for more) or experimental searches for them (see [18, 19]).

- **Part Four.** The fourth part introduces the concept that there are no fundamental global symmetries. This is a longstanding idea about quantum gravity, which has also served as an aesthetic guide to model-building in quantum field theory: it is better to find an accidental global symmetry (one which simply cannot be violated by low-dimension operators) than to postulate a nearly unbroken global symmetry without further explanation. In recent years, arguments against global symmetries in quantum gravity have been put on a stronger foundation, and in some cases we can quantify the extent to which we expect a symmetry to be broken. The Weak Gravity Conjecture (WGC) is one statement along these lines, which has itself been sharpened into more useful conjectures. Recent years have also seen significant generalizations of the concept of global symmetry in quantum field theory. Combined with the expectation that such symmetries are absent in quantum gravity, this sets up a tension: how does quantum gravity manage to break (or gauge) all the candidate symmetries?

Many of the topics discussed in this part fall under the rubric of the “Swampland program,” which searches for universal features of quantum gravity theories and obstructions to embedding effective field theories in quantum gravity [20]. A global symmetry is one such obstruction, an idea that was widely explored for decades before the term “Swampland” was coined. In these lectures I will stick to what I view as the most well-established parts of the Swampland program.

Further reading: The Harlow-Ooguri holographic argument against global symmetries is a long paper, but along the way reviews many important aspects of physics related to these lectures in great detail [21]. (You might at least read the short version, [22], though it omits much of the useful pedagogical content.) For the Weak Gravity Conjecture, you should read the original paper [23]. More recent work on the WGC is thoroughly reviewed in a recent article that I wrote with Daniel Harlow, Ben Heidenreich, and Tom Rudelius [24].

Readers wishing to learn about other aspects of the Swampland program beyond those discussed here may be interested in the reviews [25, 26].

- **Part Five.** The fifth part explores the consequences of no global symmetries for particle physics. The quantum gravity perspective can shed light on various questions, such as whether the photon has a mass, and potential origins of neutrino masses. In my opinion, the Strong CP problem is currently an especially exciting place to look for a confrontation between quantum gravity principles and phenomenology. I discuss an emerging picture of why axion-like fields have a crucial role to play in quantum gravity, which could inform the search for axions and axion-like particles in the real world.

Further reading: There is a recent Snowmass white paper on phenomenological implications of quantum gravity that I wrote with Patrick Draper and Isabel Garcia Garcia [27].

David Tong’s lecture notes on gauge theory [28] are excellent and have substantial overlap with a number of topics discussed here, though in a different enough order that it’s hard to point to specific places to dip into them. I highly recommend finding the time to read them all.

Acknowledgments

First, I thank the scientific organizers of the 2022 TASI summer school (JiJi Fan, Stefania Gori, and Lian-Tao Wang) for inviting me to give these lectures and providing the impetus for writing up these notes, and the TASI local organizers (Tom DeGrand, Oliver DeWolfe, and Ethan Neil) for providing an excellent venue and ensuring that the school ran smoothly. I also thank the TASI students, too many to name individually, who asked many insightful questions and stopped me whenever I wasn’t being clear or making sense. This has been a great help in figuring out what to say, and how to say it, in this written form of the lectures. Some of the material in these notes has similarly been tested on colloquium audiences at the University of Chicago, Caltech, and the IFT in Madrid, who have provided useful feedback. I thank Daniel Aloni, Liam Fitzpatrick, Ben Heidenreich, Jake McNamara, and Tom Rudelius for feedback on the draft of these lecture notes.

I’ve learned much of the material I present here over the years from many collaborators with whom I’ve worked on axion physics, the Weak Gravity Conjecture, and the absence of global symmetries in quantum gravity. In alphabetical order: Prateek Agrawal, Manuel Buen-Abad, JiJi Fan, Katie Fraser, Daniel Harlow, Ben Heidenreich, Jake McNamara, Miguel Montero, Tom Rudelius, John Stout, Chen Sun, Irene Valenzuela, and Lian-Tao Wang. Others who have taught me about some of the material discussed here include Nima Arkani-Hamed, Tom Banks, Cliff Cheung, Kiwoon Choi, Clay Córdova, Patrick Draper, Thomas Dumitrescu, Isabel Garcia Garcia, Grant Remmen, Shu-Heng Shao, Cumrun Vafa, and Sasha Zhiboedov. I’m sure that I’m unintentionally omitting others who deserve thanks. These lecture notes are my particular way of deconstructing and re-assembling a large body of collective knowledge, and I hope that others will mine them for their own pedagogical writing in the future.

My work is partially supported by the DOE grant DE-SC0013607, the Alfred P. Sloan Foundation Grant No. G-2019-12504, and the NASA Grant 80NSSC20K0506.

Part One: Global and Gauge Symmetries, Charge Quantization

2. Differential forms: notation

Readers who are comfortable with the concept and notation of differential forms should be able to freely skip this section (but may want to glance at the actions in §2.5 to be sure they are familiar).

In studying gauge theory, it is useful to use the notation of *differential forms*, which I will briefly review in this section. Differential forms are objects that can be integrated over p -dimensional manifolds. Essentially, a differential p -form is a p -index antisymmetric tensor. However, many formulas are simpler and clearer to think about when written in the language of differential forms instead of tensors. For example, this notation can make it easier to think about whether terms in an action are topological, meaning that they do not depend on the metric tensor that we use to measure distances. Topological terms are essential to axion physics, which is one reason why we will make extensive use of differential form notation in these lectures.

2.1 What is a differential form?

Let's start with one example to illustrate the general point. A gauge field A_μ can be integrated along the worldline γ of a charged particle (i.e., its path through spacetime). This integral might be written as $\int_\gamma A_\mu dx^\mu$. What does this mean? One approach is to parametrize the curve γ in terms of a function from a parameter λ to spacetime, $x^\mu(\lambda)$. Then we could compute $\int_\gamma A_\mu dx^\mu$ as $\int_{-\infty}^{\infty} d\lambda \frac{dx^\mu}{d\lambda} A_\mu$. However, because the answer is independent of the specific choice of parametrization (we could replace λ by a monotonic function $\mu(\lambda)$, and get the same result from an integral over μ), we write the integral simply as $\int_\gamma A_\mu dx^\mu$. Differential form notation takes this a step further, writing the integral simply as $\int_\gamma A$, where the (dimensionless) object A is a *differential 1-form*, defined as $A_\mu dx^\mu$. It packages the field to be integrated together with the differential dx^μ that tells us about the integration measure.

A second example is a gauge field strength $F_{\mu\nu}$, an antisymmetric rank-2 tensor that we can integrate over a *surface* to calculate a flux. In this case, we define the differential form F (again, a dimensionless object) as

$$F = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu. \quad (1)$$

The *wedge product* \wedge is an antisymmetrized tensor product,

$$dx^\mu \wedge dx^\nu = dx^\mu \otimes dx^\nu - dx^\nu \otimes dx^\mu. \quad (2)$$

The fact that a given term, like $dx^1 \otimes dx^2$, appears twice here (in different orderings) accounts for the normalization factor of $1/2$ in the definition of F relative to $F_{\mu\nu}$. Why should a differential form be *antisymmetric*? It's because the area element on a surface involves each of the *independent* coordinates, e.g., $dx dy$ in the plane; integrating an object $dx dx$ is meaningless. The antisymmetrization ensures that a p -dimensional integrand really involves p *different* directions. Rephrasing our field strength example may help: think about calculating a magnetic flux through a surface. You may have encountered this in introductory physics in the form of an integral of a vector \vec{B} dotted into a unit normal vector to a surface, $\int \vec{B} \cdot \hat{n} dS$, with dS an area element on the surface. However, the

antisymmetrization was lurking here in the fact that to obtain a vector like \vec{B} from the field strength F , we use an antisymmetric symbol, $B^k = \frac{1}{2}\epsilon^{ijk}F_{ij}$. The combination $\epsilon^{ijk}\hat{n}_k dS$ is then the area element $dx^i \wedge dx^j$.

A differential form of degree p , or a *differential p -form* (or even just *p -form*, for short) is a simple generalization of this idea, with more indices and correspondingly more differentials:

$$\omega_p = \frac{1}{p!} \omega_{\mu_1 \dots \mu_p} dx^{\mu_1} \wedge dx^{\mu_2} \wedge \dots \wedge dx^{\mu_p}, \quad (3)$$

where $\omega_{\mu_1 \dots \mu_p}$ are components of a p -index antisymmetric tensor. The subscript p in the name ω_p is just a reminder that this is a p -form; it is not an index. Here the wedge product of the forms is a signed sum over permutations,

$$dx^{\mu_1} \wedge dx^{\mu_2} \wedge \dots \wedge dx^{\mu_p} = \sum_{\sigma \in S_p} \epsilon(\sigma) dx^{\sigma(\mu_1)} \otimes dx^{\sigma(\mu_2)} \otimes \dots \otimes dx^{\sigma(\mu_p)}, \quad (4)$$

where σ denotes a permutation of the p integers μ_1, \dots, μ_p and $\epsilon(\sigma)$ is its *sign*, i.e., 1 if it can be accomplished by an even number of swaps of two integers and -1 if it involves an odd number of swaps. If this abstract notation about permutations is unfamiliar, one example might help to illustrate the idea:

$$dx \wedge dy \wedge dz = dx \otimes dy \otimes dz - dx \otimes dz \otimes dy + dz \otimes dx \otimes dy \quad (5)$$

$$- dz \otimes dy \otimes dx + dy \otimes dz \otimes dx - dy \otimes dx \otimes dz. \quad (6)$$

The $1/p!$ in (3) is a normalization factor correcting for the counting of the different permutations in which the same differential elements can appear. It's worth explicitly mentioning one special case, for clarity: a differential 0-form is the degenerate case without any differentials dx ; in other words, it's just a function $f(x)$.

The wedge product of differential forms, in general, is determined by applying linearity together with the definition (4) of the wedge product of the differential elements. Notice that changing the order of a wedge product depends on the degree of the forms. For instance, two 1-forms A_1 and B_1 have $A_1 \wedge B_1 = -B_1 \wedge A_1$, but a 1-form A_1 and a 2-form F_2 have $A_1 \wedge F_2 = F_2 \wedge A_1$. The reason is that, in the latter case, we have differentials appearing in the form $dx^\mu \wedge (dx^\nu \wedge dx^\rho) = -dx^\nu \wedge dx^\mu \wedge dx^\rho = +(dx^\nu \wedge dx^\rho) \wedge dx^\mu$. In general, any p -form with *even* p will commute with other forms under the wedge product, because we move each differential past an even number of others and do not acquire a net sign. This can be summarized with the rule: given a p -form ω_p and a q -form η_q , we have

$$\omega_p \wedge \eta_q = (-1)^{pq} \eta_q \wedge \omega_p. \quad (7)$$

This holds because $(-1)^{pq}$ is equal to -1 if and only if *both* p and q are odd.

2.2 The exterior derivative

The exterior derivative d is an operation that maps p -forms to $(p+1)$ -forms. Starting with a 0-form, the components of the exterior derivative are those of the *gradient*:

$$df = \frac{\partial f}{\partial x^\mu} dx^\mu, \quad (8)$$

with the sum over μ implicit as usual. The exterior derivative of forms of higher degree follows a similar pattern: given a p -form ω as in (3), its exterior derivative is the $(p + 1)$ -form

$$d\omega = \frac{\partial \omega_{\mu_1 \dots \mu_p}}{\partial x^\mu} dx^\mu \wedge dx^{\mu_1} \wedge dx^{\mu_2} \wedge \dots \wedge dx^{\mu_p}. \quad (9)$$

In words, $d\omega$ is an antisymmetrized derivative whose components consist of all of the derivatives of the components of ω .

In particular, if $A = A_\mu dx^\mu$, the exterior derivative is

$$dA = \frac{\partial A_\nu}{\partial x^\mu} dx^\mu \wedge dx^\nu = F, \quad (10)$$

where F , as in (1), has components $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. This is just the usual field strength of a vector field.

Exercise: In 3-dimensional space, we can use the antisymmetric symbol ϵ^{ijk} to convert a 2-form like F to a vector. Check that the resulting object can be identified with the *curl* $\vec{\nabla} \times \vec{A}$.

The exterior derivative obeys a version of the usual product rule for derivatives, but potentially with an extra minus sign due to its antisymmetric property:

$$d(\omega_p \wedge \eta_q) = (d\omega_p) \wedge \eta_q + (-1)^p \omega_p \wedge (d\eta_q). \quad (11)$$

In particular, you should be careful about this sign when integrating by parts to derive equations of motion!

The antisymmetry property further implies that $d^2 = 0$, i.e.,

$$d(d\omega_p) = 0, \quad (12)$$

for any p -form ω_p . Whenever $d\omega_p = 0$, we say that ω_p is a *closed* p -form. We will see below that such forms should be thought of as *conserved currents*. When there exists a $(p - 1)$ -form η_{p-1} such that $\omega_p = d\eta_{p-1}$, we say that ω_p is an *exact* p -form. Because of (12), all exact forms are closed, but not all closed forms are exact. The space of closed forms modulo exact forms is known as *de Rham cohomology*, and is a useful tool for analyzing the topology of the spaces on which differential forms are defined [29].

We urge readers to be somewhat cautious in interpreting (12) in physics: sometimes the “forms” that we encounter are not true forms when they are gauge-dependent quantities, and as a result this equation fails to hold. For example, given a vector potential A , we have a field strength $F = dA$; the above equation leads us to expect that $dF = 0$. This result, known as the *Bianchi identity*, can be violated in physics when the gauge field becomes ill-defined at a singular point, like the core of a magnetic monopole. Furthermore, despite the equation $F = dA$, we should not think of F as an exact form in general. The reason is that there is no globally-defined 1-form A ; we must patch together definitions of A in different coordinate charts that are related by nontrivial gauge transformations on overlaps. As a result, A is not really a 1-form (formally, it is a connection on a $U(1)$ principal bundle); F , however, is an honest 2-form that is globally defined (because electromagnetic fields, unlike gauge potentials, are physical observables). A classic example of this, the Dirac magnetic monopole, is discussed below in §4.5.2.

2.3 Integration and Stokes's theorem

Given a differential form ω_p , one can integrate it over a closed p -dimensional surface. In fact, there is a small subtlety here: we have to pick an *oriented* surface. For example, if we want to integrate the 1-form $x dx$ over the manifold $(0, 1) \subset \mathbb{R}$, we can choose the conventional orientation and calculate (as in introductory calculus!) that $\int_0^1 x dx = \frac{1}{2}$, or we can choose the opposite orientation and calculate $\int_1^0 x dx = -\frac{1}{2}$. The results of integration with different orientations differ by a sign.

Let me make a brief aside here, which will not be relevant for the remainder of these lectures but which actually does have physical importance. The need for an orientation may bother you, and it should. For instance, the surface area of a Möbius strip is a perfectly well-defined quantity, even though the space is non-orientable. Why should we need an orientation to compute an integral? In fact, we don't. We can integrate an object known variously as a differential *pseudo-form* [30] or a *twisted* differential form [29] (more verbosely, a form twisted by the orientation line bundle). You have encountered such objects in physics before, even if the terminology is new to you. An example is a pseudoscalar. It is defined only up to a sign; when swapping the orientation of our coordinates (as in a parity transformation), we also change the sign of the field. Pseudo-forms can be integrated over non-orientable manifolds by picking a set of coordinate charts. On any given coordinate chart, we choose an orientation. We can choose whichever one we like, because if we change our choice, the pseudo-form we are integrating changes sign so that the integral does not. Then we patch together the integrals on different charts using a partition of unity (i.e., a set of functions defined on overlapping charts such that they add up to one everywhere when summed over all the charts). From this, we see that the volume element of a p -dimensional space is most naturally a *pseudo-p-form*, rather than an ordinary p -form.

A quantum field theory that can be defined without reference to an orientation is said to have *parity* as a global symmetry. Parity can also be a gauge symmetry, in the context of quantum gravity, which means that the path integral includes a sum over both orientable and non-orientable spacetimes. For more on this topic, see [31].

Differential forms allow a simple statement of the generalized Stokes's theorem. Given a p -dimensional manifold M with boundary ∂M (the notation ∂ is commonly used in mathematics for boundaries, not just for derivatives!), and a $(p-1)$ -form ω_{p-1} ,

$$\int_{\partial M} \omega_{p-1} = \int_M d\omega_{p-1}. \quad (13)$$

(A little fine print: M has an orientation and this *induces* an orientation on ∂M , so that the two sides of the equation are computed with compatible orientations.) This one statement encodes many familiar statements in vector calculus: Green's theorem, Stokes's theorem, the divergence theorem, and even Cauchy's integral formula can all be understood as special cases.

Exercise: work out how each of the aforementioned classic theorems can be understood in terms of the generalized Stokes's theorem.

Notice that our discussion of integration has not referred to a metric on the manifold that we integrate over. Given a differential form, you can integrate it without needing to be told anything about a metric. This is one reason why differential forms play a prominent role in the study of topological effects in field theory. Of course, in order to find the components of a differential form in the first place, you might need to use the metric. For example, a d -dimensional manifold equipped with a metric tensor $g_{\mu\nu} dx^\mu \otimes dx^\nu$ (which is a *symmetric* tensor and hence *not* a differential form!) has a *volume form*,

$$\text{vol}_d = \begin{cases} \sqrt{+\det g} dx^1 \wedge \cdots \wedge dx^d, & \text{Euclidean;} \\ \sqrt{-\det g} dx^0 \wedge \cdots \wedge dx^{d-1}, & \text{Minkowski.} \end{cases} \quad (14)$$

We will also write simply $|g|$ for $|\det g|$. In either case, we will also use the shorthand $\text{vol}_d = \sqrt{|g|} d^d x$. This is the familiar measure we integrate against when defining actions in d -dimensional spacetime. The d -form vol_d is sometimes referred to as the Levi-Civita tensor. (In general, a d -form on a d -dimensional manifold is sometimes called a “top form.”)

Note that in the Minkowski case we always write the time coordinate x^0 first. This is a choice of orientation, with the convention that $\int \sqrt{|g|} d^d x$ is always positive.

2.4 The Hodge star

The *Hodge star* is an operation that, on d -dimensional spaces (or spacetimes), takes a p -form to a $(d-p)$ -form. It is defined only for manifolds equipped with a metric tensor, so it is geometric, rather than purely topological. In terms of the antisymmetric tensor components, you can think of this operation as essentially contracting a tensor with the Levi-Civita tensor with all d indices raised. In terms of the differentials appearing in a given form, you can think of the Hodge star as replacing all those that appear in a given term with all the *others*; e.g., in flat 3-dimensional space, dx is replaced by $dy \wedge dz$.

Specifically, the Hodge star is defined by:

$$\star(dx^{\mu_1} \wedge \cdots \wedge dx^{\mu_p}) = \frac{1}{(d-p)!} \sqrt{|g|} \epsilon_{\nu_1 \cdots \nu_d} g^{\mu_1 \nu_1} \cdots g^{\mu_p \nu_p} dx^{\nu_{p+1}} \cdots dx^{\nu_d}, \quad (15)$$

where ϵ is a fully antisymmetric symbol with

$$\begin{aligned} \epsilon_{12 \cdots d} &\equiv +1 \quad (\text{Euclidean}), \\ \epsilon_{01 \cdots (d-1)} &\equiv +1 \quad (\text{Minkowski}). \end{aligned} \quad (16)$$

With this sign convention, the Hodge star operation gives a very compact expression for the volume form (14):

$$\text{vol}_d = \star 1. \quad (17)$$

As is clear from the definition, the Hodge star operation depends explicitly on the metric.

The differential form $\star \omega_p$ is sometimes referred to as the “Hodge dual” of ω_p . From the action of the Hodge star on differential elements, it is clear that acting twice with the Hodge star approximately returns us to what we started with. Working this out more carefully, one finds that if ω_p is a p -form in a d -dimensional space with m minus signs in the metric signature, we have

$$\star(\star \omega_p) = (-1)^{p(d-p)+m} \omega_p. \quad (18)$$

Exercise: In 3-dimensional space, given a 1-form V , explain how to combine exterior derivatives and Hodge stars to write a differential form expression equivalent to the *divergence* $\vec{\nabla} \cdot \vec{V}$.

2.5 Examples of actions in differential form notation

Let's compare some familiar physics quantities in tensor index notation and in differential form notation. The action for a real scalar field (in mostly-minus metric signature) is given by

$$\int d^d x \sqrt{|g|} \frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi = \frac{1}{2} \int d\phi \wedge \star d\phi. \quad (19)$$

The first thing to note is that the familiar measure $d^d x \sqrt{|g|}$ is *not* written on the right-hand side, because it's built into the definitions of the differential forms that appear there! If you're used to writing every integral sign with a $d^d x$ following it, when you work with differential forms you have to get used to the fact that it doesn't show up. The next thing to notice is that the expression on the left has a metric tensor $g^{\mu\nu}$ used to raise an index so that the two scalar gradients can be dotted into each other. In the expression on the right, we don't see this. Instead, what has happened is that $\star d\phi$ is now a 3-index tensor constructed by contracting $\partial_\nu \phi$ with an *antisymmetric* ϵ -symbol. This is then combined with $d\phi$ using the *antisymmetric* wedge product. These two antisymmetrizations effectively cancel each other out, and lead to the contraction we're used to.

Exercise: Verify the equivalence of the two sides in (19), and the other two examples (20) and (24) below.

Our next example is a gauge field kinetic term. We will write this in a commonly used, but non-canonical, normalization where the gauge coupling e appears in front of the kinetic term in a factor of $1/e^2$. This normalization of the gauge field is especially convenient for discussing charge quantization, as we will do below. With this normalization, $F_{\mu\nu}$ always has mass dimension 2 (whereas F , the differential form, is dimensionless) and e^2 has mass dimension $4 - d$. We have:

$$\int d^d x \sqrt{|g|} \left(-\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu} \right) = \int \left(-\frac{1}{2e^2} F \wedge \star F \right). \quad (20)$$

Similar to the scalar action, we again accomplish the dotting together of indices through a combination of a wedge product and a Hodge star. Notice that the left-hand side has a factor of 4 where the right-hand side has a factor of 2. In general, for a p -index antisymmetric tensor field strength, the tensor index notation would have an extra factor of $1/p!$ relative to the differential form notation, which always has simply a factor of $1/2$.

The final example will be one of the major players in the lectures below: the theta angle. In four dimensions, the Hodge dual of a 2-form field strength is again a 2-form. We can write it in components, which are often denoted $\tilde{F}_{\mu\nu}$:

$$\star F = \frac{1}{2} \tilde{F}_{\mu\nu} dx^\mu \wedge dx^\nu, \quad \text{where} \quad \tilde{F}_{\mu\nu} = \frac{1}{2} \sqrt{|g|} \epsilon_{\mu\nu\rho\sigma} F^{\rho\sigma}. \quad (21)$$

When F is the electromagnetic field strength, this operation is electric-magnetic duality. We should be careful when raising indices: the ϵ symbol by itself is *not* a tensor; the Levi-Civita tensor is the combination $\sqrt{|g|}\epsilon_{\mu\nu\rho\sigma}$. The proper statement about raising indices is that

$$\sqrt{|g|}\epsilon_{\mu\nu\rho\sigma}g^{\mu\alpha}g^{\nu\beta}g^{\rho\gamma}g^{\sigma\delta} = \frac{(-1)^m}{\sqrt{|g|}}\epsilon^{\alpha\beta\gamma\delta}, \quad (22)$$

where m is the number of minus signs in the metric signature and ϵ is as in (16). Hence, in Minkowski space we have

$$\tilde{F}^{\mu\nu} = -\frac{1}{2\sqrt{|g|}}\epsilon^{\mu\nu\rho\sigma}F_{\rho\sigma}. \quad (23)$$

The “ θ term” that is written in tensor index notation by dotting F into \tilde{F} can thus be seen, in differential forms notation, to involve simply the wedge product of F with itself:

$$\int d^4x \sqrt{|g|} \frac{\theta}{16\pi^2} F_{\mu\nu} \tilde{F}^{\mu\nu} = - \int \frac{\theta}{8\pi^2} F \wedge F. \quad (24)$$

(We have chosen the sign for Minkowski space here; in Euclidean signature, there will also be an important factor of i from Wick rotation.) Here θ may be a constant or a pseudoscalar field; both cases will appear in the lectures below. In this case, the differential form notation reveals that the somewhat complicated-looking term appearing on the left is, in fact, *topological*: the right-hand side depends on θ and the 2-form F but has no dependence whatsoever on the metric tensor! (Among other things, this means that the term does not contribute to the stress-energy tensor, which we compute by varying the action with respect to the metric.) The relative simplicity of this term in differential form notation is one of the reasons we will make extensive use of this formalism below.

3. Ordinary global symmetries; currents and charges

3.1 The symmetry group $U(1)$

A $U(1)$ global symmetry is one associated with *counting* some kind of stuff, e.g., the number of particles of some type. A familiar example is a theory of a complex scalar field with a potential that depends only on the magnitude of the scalar:

$$\frac{1}{\sqrt{|g|}}\mathcal{L} = \partial_\mu\phi^\dagger\partial^\mu\phi - m^2\phi^\dagger\phi - \frac{\lambda}{4}(\phi^\dagger\phi)^2. \quad (25)$$

This theory is invariant under the symmetry transformation

$$\phi(x) \mapsto e^{i\alpha}\phi(x), \quad (26)$$

with constant α . Notice that this symmetry operation only depends on $\alpha \bmod 2\pi$. This is the defining feature of $U(1)$.

The group $U(1)$ is simply the group of complex numbers of unit magnitude, i.e., numbers of the form $e^{i\alpha}$ for $\alpha \in \mathbb{R}$. The group operation is multiplication of these numbers:

$$e^{i\alpha} \cdot e^{i\beta} = e^{i(\alpha+\beta)}. \quad (27)$$

This is a *group* because it has an identity element (the number 1), every element has an inverse ($e^{i\alpha} \cdot e^{-i\alpha} = 1$), and the multiplication law is associative.

There is a *different*, but closely related, group: \mathbb{R} , the real numbers viewed as a group under addition. Locally these groups look the same. They have the same Lie algebra, $\mathfrak{u}(1)$, which is also the group of real numbers under addition. (Lie algebras are equipped with another operation, the commutator, but because these groups are abelian the commutators in $\mathfrak{u}(1)$ are all zero.) To make the relationship precise, we can map an element $\alpha \in \mathbb{R}$ to the element $e^{i\alpha} \in U(1)$. This is a *group homomorphism*, i.e., it commutes with the group operation. However, it is a many-to-one map: $U(1)$ consists of the real numbers modulo addition of integer multiples of 2π . Notice that we could, via this homomorphism, say that the Lagrangian (25) admits an \mathbb{R} global symmetry. However, in physics it is more natural to study theories with $U(1)$ symmetries; we will discuss some reasons for this below.

3.2 $U(1)$ charges

Consider a field Φ charged under a $U(1)$ symmetry, with some charge q . This means that a group element $g = e^{i\alpha} \in U(1)$ acts on Φ according to the rule

$$\Phi(x) \mapsto e^{iq\alpha} \Phi(x). \quad (28)$$

Now, we have to remember that the group element $e^{i\alpha}$ is *the same* as the group element $e^{i(\alpha+2n\pi)}$, for integer n . The right hand side of the above rule for how Φ transforms has to be well-defined, which means that we need

$$e^{iq\alpha} \Phi(x) = e^{iq(\alpha+2n\pi)} \Phi(x), \quad \forall n \in \mathbb{Z}. \quad (29)$$

This means that we need *quantized* charges!

$$q \in \mathbb{Z}. \quad (30)$$

In mathematical terminology, the *representations* of the group $U(1)$ are labeled by integers.

This is a very basic mathematical fact, but you can find many books or papers by particle physicists that deny its validity, so let me put it in a box to emphasize it:

The allowed charges under a $U(1)$ symmetry are quantized. They must be integers.

If someone tells you otherwise, it means that they are really thinking of an \mathbb{R} symmetry. One sometimes encounters the opinion that in physics we should only label symmetry groups by their Lie algebra. But as we will see, the global structure of the group has important implications, so it is useful to be precise about our language and label a theory by the group that acts on the fields.

3.3 $U(1)$ conservation laws and currents

Let's first recall what the conservation of any locally conserved “stuff” (e.g., mass of a fluid, or charge) looks like. In introductory physics classes you encountered many of these equations, which

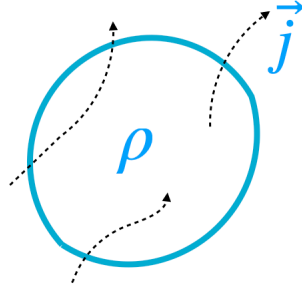


Figure 1: A U(1) conservation law relates a density ρ of some quantity inside a region to the flux through the boundary of the region, as determined by the current \vec{j} .

say that if a quantity is conserved then the rate of change of the amount of stuff inside some volume is given by the integrated flux of the stuff through the surface bounding the volume:

$$\frac{dQ}{dt} = - \oint \vec{\Phi} \cdot d\vec{S}. \quad (31)$$

Such an equation has a *differential* formulation in terms of a density ρ of stuff and a current \vec{j} describing the motion of the stuff, generally known as a *continuity equation*:

$$\frac{\partial \rho(t, \vec{x})}{\partial t} = -\vec{\nabla} \cdot \vec{j}(t, \vec{x}) \quad (32)$$

See Fig. 1. In the relativistic context, ρ can be viewed as the time component of a 4-vector whose spatial components are \vec{j} , and these conservation equations take on the particularly simple form $\partial_\mu j^\mu = 0$.

A Lagrangian field theory with a U(1) global symmetry will have such a conserved current j^μ . This is a special case of *Noether's theorem*, which tells us that any continuous global symmetry in a Lagrangian theory gives rise to a conserved current. (This applies to internal symmetries; the story for spacetime symmetries is a little more subtle, and we won't have a need to go into it here.) Noether's theorem has a straightforward derivation. Consider a field variation that *would* be a symmetry if ϵ is constant:

$$\phi(x) \mapsto \phi(x) + \epsilon(x)\xi(x). \quad (33)$$

By definition of a symmetry, the action doesn't change if ϵ is constant, which means that the change in the action due to our field variation must come from *derivatives* of $\epsilon(x)$, i.e.,

$$S[\phi] \mapsto S[\phi] + \int d^d x \sqrt{|g|} j^\mu(x) \partial_\mu \epsilon(x), \quad (34)$$

for *some* quantity $j^\mu(x)$. In other words, we vary the action and simply *define* $j^\mu(x)$ to be whatever appears multiplying the derivative of ϵ (with a factor of $\sqrt{|g|}$ pulled out). Now, we recall that for a Lagrangian theory, a field configuration that solves the equations of motion of the theory is one for which $\delta S = 0$ for *any* small variation of the fields. The only way that we can have $\delta S = 0$ for the particular variation of the action (34) is to have $\partial_\mu (\sqrt{|g|} j^\mu) = 0$, which is therefore an equation of motion obeyed by the fields. The reason that we don't include the factor of $\sqrt{|g|}$ in the definition of

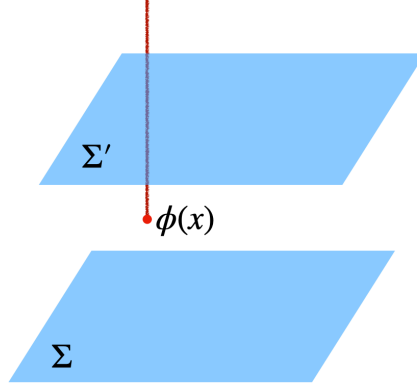


Figure 2: We can measure the global U(1) charge Q along two different spatial slices, Σ and Σ' . An insertion of an operator carrying global charge q , like $\phi(x)$, at an intermediate time will change the result: $Q(\Sigma') - Q(\Sigma) = q$. One way to think about this is that the local operator $\phi(x)$ can create a particle of charge q , which exists on the surface Σ' (its worldline, depicted as a fuzzy red line emanating from the operator insertion, passes through the surface).

j^μ is that our chosen definition gives us the *covariant* form of the current conservation law for j^μ in curved spacetime:¹

$$\nabla_\mu j^\mu = \frac{1}{\sqrt{|g|}} \partial_\mu (\sqrt{|g|} j^\mu) = 0. \quad (35)$$

We emphasize that this is *not* an identity that follows from the form of j^μ for *any* field configuration; it holds only for those specific field configurations that solve the classical equations of motion. In particular, in the quantum theory, this will not be true of every field configuration we sum over in the path integral, only the saddle-point configurations.

Exercise: following the logic above, verify that the Noether current associated with the U(1) symmetry of (25) is

$$j^\mu = i \left[(\partial^\mu \phi^\dagger) \phi - \phi^\dagger (\partial^\mu \phi) \right]. \quad (36)$$

Given a conserved current $j^\mu(x)$, we can find the total charge in the space by integrating the charge density $\rho = j^0$ over all of space. In d spacetime dimensions, we have

$$Q = \int d^{d-1}x \sqrt{|g|} j^0(x). \quad (37)$$

The charge is a conserved quantity according to the equations of motion, but if we evaluate a correlation function involving insertions of charged operators $\phi(t, \vec{x})$ at some time, these can change the charge, because they effectively insert a worldline of a charged particle (see Fig. 2).

The fact that we integrate the charge density over a $(d-1)$ -dimensional slice of spacetime is a hint that thinking of currents as vectors is sometimes not optimal. The object that can naturally be

¹If you haven't seen this before, or if your general relativity is rusty, you can verify the identity for Christoffel symbols with two indices contracted, $\Gamma_{\mu\nu}^\mu = \frac{1}{\sqrt{|g|}} \partial_\nu \sqrt{|g|}$, from which this form of $\nabla_\mu j^\mu$ follows.

integrated over a $(d-1)$ -dimensional manifold is a $(d-1)$ -form. How is this related to the current we have discussed? We can view the current as defining a 1-form, $j = j_\mu dx^\mu$. The Hodge dual of this one form,

$$J = \star j, \quad (38)$$

is a $(d-1)$ -form. The charge, then, is simply

$$Q = \int_\Sigma J. \quad (39)$$

This is the most mathematically convenient way to formulate a conserved current. The conservation law is simply the statement that this $(d-1)$ -form current is *closed*:

$$dJ = 0. \quad (40)$$

This is such a useful perspective to keep in mind that I frame it in another box:

Conserved currents are often best thought of as closed differential forms, $dJ = 0$.

Exercise: convince yourself that the equations $\nabla_\mu j^\mu = 0$ and $dJ = 0$ are equivalent.

3.4 Symmetry operators

An important part of the modern perspective on global symmetries is that global symmetries *act locally*. This is already evident in the fact that local operators transform under these symmetries, though we usually consider the action of the symmetry on all local operators everywhere in spacetime at once. However, there is in fact a family of *surface operators* that implement the action of global symmetries within a limited region of spacetime [1]. These are called *symmetry operators* or sometimes *charge operators* (we prefer the former, since the latter *sounds* almost like “charged operators,” but means something different). We can construct these operators in the Euclidean theory, since we will treat space and time directions on an equal footing in their construction instead of only integrating currents over fixed-time spatial slices. Essentially, the idea is to take the two spatial slices shown in Fig. 2 and bend them around to meet each other, surrounding the operator insertion $\phi(x)$.

The symmetry operators for a $U(1)$ global symmetry are labeled by a closed $(d-1)$ -dimensional manifold Σ and an element $e^{i\alpha} \in U(1)$ (or simply by the phase α , understood to be defined mod 2π). These operators are defined by exponentiating the integrated current over Σ , with coefficient α :

$$U(\Sigma, \alpha) \equiv \exp \left(i\alpha \int_\Sigma J \right). \quad (41)$$

See Fig. 3. If the surface Σ surrounds an operator insertion of a charged local operator $\phi(x)$ of charge q , and no other charged operators, then the symmetry operator inserted in a correlation function is equivalent to rephasing the charged operator. In other words, there is an OPE, $U(\Sigma, \alpha)\phi(x) = e^{i\alpha q}\phi(x)$. This is illustrated in Fig. 3.

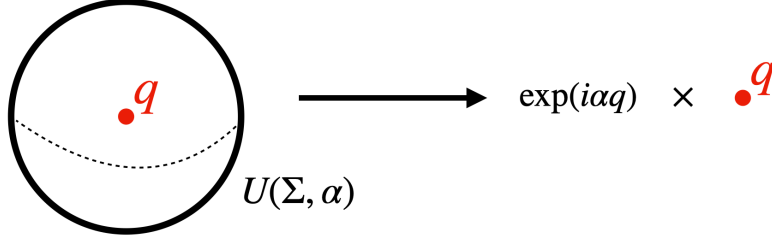


Figure 3: For a global $U(1)$ symmetry, a *charge operator* or *symmetry operator* supported on a 3-dimensional slice Σ of spacetime measures the charge q of a pointlike operator, via this Ward identity.

The symmetry operators are *topological operators*, meaning that when inserted in correlation functions, the surface Σ can be deformed arbitrarily without changing the answer, *provided* that no charged operators cross through Σ as it is deformed. Why are the operators topological? Suppose that we deform the surface Σ to a different surface Σ' . Then the operators are related by

$$U(\Sigma', \alpha) = \exp\left(i\alpha \int_{\Sigma - \Sigma'} J\right) U(\Sigma, \alpha). \quad (42)$$

Fill in the interior of the region swept out by deforming Σ to Σ' and call it M (see Fig. 4). If no operators were crossed by the surface as we deformed it, we have $\partial M = \Sigma - \Sigma'$. Then by Stokes's theorem, $\int_{\Sigma - \Sigma'} J = \int_M dJ$. The latter is zero by current conservation, which tells us that $U(\Sigma', \alpha)$ and $U(\Sigma, \alpha)$ are equivalent. However, a local operator insertion is a singularity: if we insert a charged operator $\phi(x)$ at a point that lies inside M , then $\partial M = (\Sigma - \Sigma') \cup S_x^{d-1}$, where S_x^{d-1} is a little sphere around the singular point with the operator insertion. This means that correlation functions of $U(\Sigma, \alpha)$ and $U(\Sigma', \alpha)$ will differ precisely due to localized contributions from the operator insertions in the region M in between. This is just a description of the usual Ward identity in terms of symmetry operators in the path integral. Physically, the picture is that $\phi(x)$ *creates* a charged state, as in Fig. 2, whose propagation crosses Σ' but not Σ .

There is a nice discussion in section 2 of David Simmons-Duffin's TASI lectures on the conformal bootstrap [32] on how topological operators implement symmetries, albeit focused on the case of spacetime symmetries.

The formalism of symmetry operators is very powerful. For example, a discrete symmetry group like \mathbb{Z}_N has no associated conserved current. It *does*, however, have a family of symmetry operators, $U(\Sigma, \alpha)$ where now the phase α takes on only the discrete values $\{0, 2\pi/N, 4\pi/N, \dots, 2\pi(N-1)/N\}$. More generally, *any* global symmetry group G comes with a family of surface operators $U(\Sigma, g)$ associated with $(d-1)$ -dimensional surfaces Σ and group elements $g \in G$. Inserted in correlation functions, these act on local G -charged operators surrounded by Σ by transforming these operators according to the appropriate *representation* of G . (In the general case, this will not be simply a rephasing of the operator, but will replace an operator with a linear combination of other operators).

A variant of this formalism allows us to describe generalized global symmetries that act on extended operators. For instance, a 1-form global symmetry acts on 1-dimensional operators like Wilson loops; the charged *objects* would be strings, not point particles. There are a family of

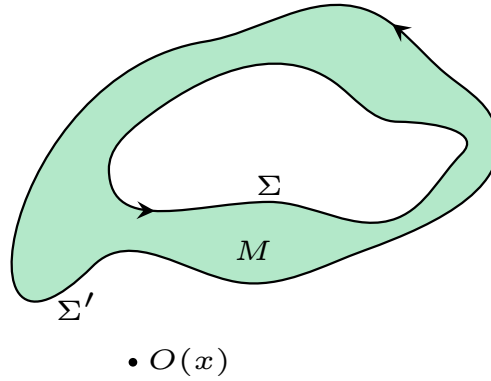


Figure 4: Illustration of a topological operator (in 2d, for convenience). If we insert a symmetry operator $U(\Sigma, g)$ or a symmetry operator $U(\Sigma', g)$, their correlation functions will be identical provided that all charged operator insertions like $O(x)$ are outside the (shaded green) region M swept out when deforming Σ into Σ' .

symmetry operators now labeled by $(d-2)$ -dimensional surfaces, which can link with 1-dimensional loops. In general, p -form generalized global symmetries act on p -dimensional charged operators and are implemented by $(d-p-1)$ -dimensional surface operators. We will discuss such generalized symmetries, known as higher-form symmetries, in §20. This is part of a modern understanding of symmetries that dramatically enlarges the range of possible symmetries in quantum field theory and condensed matter theory, which I believe will play a large role in all areas of theoretical physics, including particle phenomenology, in the future. These include not only higher-form symmetries, but more complicated cases known as higher-group symmetries and non-invertible symmetries. It is well worth your effort to invest time in learning more.

4. U(1) gauge theory

“U(1) is very deep.” – Lian-Tao Wang

4.1 Basic definitions; charge quantization

Gauge symmetries have a very different interpretation than global symmetries. A global symmetry maps one physical state to a different physical state. A gauge symmetry, by contrast, is not really a symmetry at all: it is a *redundancy* in our description of the theory. It allows us to describe precisely the same physical state in different ways. This conceptual distinction has many important physical implications. For instance, when a continuous global symmetry is spontaneously broken, we can excite a long wavelength mode that transitions from one vacuum state to another. This is a Nambu-Goldstone boson. The same is not true in gauge theory, because there are no different states to transition between! This is why the Nambu-Goldstone boson disappears from the spectrum in the Higgs mechanism. As we will see later in these lectures, gauge symmetries are perfectly consistent with quantum gravity, whereas global symmetries are not.

As we discussed above, the group $U(1)$ consists of phases $\exp(i\alpha)$ under multiplication, and as such is distinguished from the group \mathbb{R} of real numbers under addition by the 2π periodicity of

α . These groups are *locally* the same, and both have the Lie algebra $\mathfrak{u}(1) \cong \mathbb{R}$. However, they are *globally* different, and have different physics. In particular, $U(1)$ gauge theory has quantized charge, and admits magnetic monopoles. On the other hand, \mathbb{R} gauge theory does not have quantized charge, and forbids magnetic monopoles. We will derive these statements shortly, at least at a somewhat heuristic level. Readers seeking a more explicit mathematical treatment, still aimed a broad physics audience, can refer to [33] or [34].

When we carry out a gauge transformation, we choose a gauge group element $g(x)$ for every point x in spacetime, and specify an action of $g(x)$ on our field. For $U(1)$ gauge theory, a field $\psi(x)$ of charge q transforms under the gauge transformation $g(x) = \exp(iq\alpha(x))$ according to

$$\psi(x) \mapsto \psi^g(x) \equiv \exp(iq\alpha(x))\psi(x). \quad (43)$$

Because $\alpha(x)$ is only defined modulo 2π , this expression only makes sense if $\exp(2\pi iq) = 1$, i.e., if $q \in \mathbb{Z}$ is an integer. Thus, *by definition*, $U(1)$ gauge charge is quantized, as we already emphasized for $U(1)$ global charge.

By contrast, for the gauge group \mathbb{R} , we can consider a choice of gauge transformation $g(x) = \exp(iq\alpha(x))$ with $\alpha(x) \in \mathbb{R}$, together with a gauge transformation rule of precisely the same form, $\psi(x) \mapsto \exp(iq\alpha(x))\psi(x)$. This time, α is simply a real number, and this expression is always well-defined. Hence, charges in \mathbb{R} gauge theory need not be quantized: any $q \in \mathbb{R}$ is allowed.

4.2 Gauge fields

In either $U(1)$ or \mathbb{R} gauge theory, we introduce a gauge field $A_\mu(x)$ that transforms under gauge transformations via

$$A_\mu \mapsto A_\mu - ie^{-i\alpha(x)} \partial_\mu e^{i\alpha(x)} = A_\mu + \partial_\mu \alpha(x). \quad (44)$$

In the $U(1)$ case, the gauge transformation is defined by the $U(1)$ element $e^{i\alpha(x)}$; that is, $\alpha(x)$ is only defined mod 2π . This allows for winding of $e^{i\alpha(x)}$ around a circle, which will play an important role below.

The gauge-invariant field strength $F_{\mu\nu}(x)$ is defined by

$$F_{\mu\nu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x). \quad (45)$$

In the language of differential forms that we reviewed in §2, $A(x) = A_\mu(x) dx^\mu$ is a 1-form and $F(x) = \frac{1}{2} F_{\mu\nu}(x) dx^\mu \wedge dx^\nu$ is a 2-form. The gauge transformation is $A \mapsto A + d\alpha$, and the gauge invariance of F is an automatic consequence of the general mathematical fact that $d^2 = 0$ (as a consequence of antisymmetry).

In fact, we have been a bit too hasty in calling A a 1-form. In general, the gauge field $A(x)$ need not be well-defined over all of spacetime. Spacetime can be covered with multiple coordinate patches, and we define fields separately on each patch. We'll give an explicit example below in §4.5.2 when we discuss the Dirac monopole, which should make this idea clear, but let me state the general abstract formalism once for completeness. If we have two overlapping patches U and V with associated gauge fields $A_U(x)$ and $A_V(x)$, the fields only have to agree with each other up to a gauge transformation. That is, there should be a gauge transformation $g_{U \rightarrow V}(x) = \exp(i\alpha_{U \rightarrow V}(x))$ on the overlap $U \cap V$ under which A_U maps to A_V as in (44). Furthermore, if we have *three*

overlapping regions U , V , and W , we need a compatibility condition on the triple overlap $U \cap V \cap W$: $g_{V \rightarrow W}(x) \cdot g_{U \rightarrow V}(x) = g_{U \rightarrow W}(x)$. (The mathematical jargon for this is a “cocycle condition.”) It turns out—and here I will just point you to the more mathematical literature cited above, rather than trying to give an argument—that we can stop at triple overlaps; we don’t have to worry further about quadruple overlaps and so on.

The collection of coordinate charts U, V, \dots together with choices of gauge field A_U, A_V, \dots on the charts and gauge group elements $g_{U \rightarrow V}, \dots$ on the pairwise overlaps determines what is known as a $U(1)$ *gauge bundle with connection* on spacetime (also known as a principal $U(1)$ bundle with connection). The gauge field A is known as the *connection*: like the metric connection in general relativity (represented by the Christoffel symbols), it tells us how to parallel transport particles around loops, this time for charged particles. The field strength F is also called the *curvature* of the connection A , in much the same way that the metric curvature (measured by the Riemann tensor) comes from derivatives of the Christoffel symbols.

The definition of gauge bundles that we have given applies to any gauge group G , not just $U(1)$. The only difference is that the gauge transformations like $g_{U \rightarrow V}$ take values in G , and the gauge fields like A_U take values in the Lie algebra of G . Even a *discrete* group G , like \mathbb{Z}_N , can be used to define a gauge theory in this way. In that case, the connection is trivial, so we only need the overlap transformations $g_{U \rightarrow V}$, which are locally constant. Discrete gauge theories tend to get little attention in introductions to quantum field theory for particle physicists, especially as theories in their own right rather than remnants of a higgsed continuous gauge theory. For a little more detailed introduction to discrete gauge theory, I refer you to §2 of [31], where Jake McNamara and I recently tried to give a clear pedagogical summary.

4.3 Wilson loops, quantized magnetic flux

In abelian gauge theory, we can define a family of gauge-invariant *Wilson loops* associated with closed loops γ in spacetime and charges q ,

$$W_q(\gamma) \equiv \exp \left(i q \oint_{\gamma} A_{\mu} dx^{\mu} \right) = \exp \left(i q \oint_{\gamma} A \right). \quad (46)$$

As written, this is a classical expression, which can be inserted in a path integral to compute a correlation function. If the curve γ extends in the time direction, we can think of $W_q(\gamma)$ as inserting a very heavy particle of charge q with worldline γ , too heavy to move, which acts as a static probe of the theory. When inserted along a spatial slice in a path integral, the Wilson loop can also be thought of as an operator acting on Hilbert space. In the general case, the definition of $W_q(\gamma)$ requires a *path-ordering* of operators along the curve γ , sometimes written with a P in front of the exponential. This is just like the time-ordered exponential you have encountered in quantum mechanics.

For gauge transformations by a well-defined 0-form α , the Wilson loop $W_q(\gamma)$ is clearly invariant under $A \mapsto A + d\alpha$, because $\oint_{\gamma} d\alpha = 0$, using Stokes’ theorem and the fact that γ is a closed loop without boundary. For \mathbb{R} gauge theory, this establishes the gauge invariance of the Wilson loop, for *any* charge q . For $U(1)$ gauge theory, we must be a bit more careful (recall the note of caution at the end of §2.2) about the case where α itself is not well-defined but $e^{i\alpha}$ and $d\alpha$ are. In other words, we must also consider shifts of A by a general flat connection or, equivalently,

cases in which $\alpha(x)$ has a nontrivial winding number around the circle, $\alpha(2\pi) = \alpha(0) + 2\pi n$. In this case, we have

$$W_q(\gamma) \mapsto W_q(\gamma) \exp[iq(\alpha(2\pi) - \alpha(0))] = W_q(\gamma) \exp(iq2\pi n) = W_q(\gamma). \quad (47)$$

We see that only $q \in \mathbb{Z}$ is a consistent charge assignment for Wilson loops in $U(1)$ gauge theory, consistent with our earlier remark that charge is quantized in $U(1)$ gauge theory.

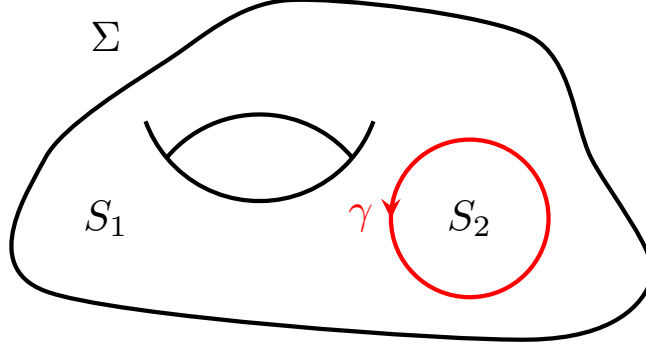


Figure 5: Illustration of the magnetic flux quantization argument discussed in the text: we divide a two-dimensional surface Σ along a curve γ into two oppositely-oriented regions S_1 and S_2 . The Wilson loop over γ is well-defined, which implies that the surface operators $\exp\left(iq \int F\right)$ over S_1 and $-S_2$ agree, which implies the flux of F through all of Σ is quantized.

Consider a Wilson loop in either $U(1)$ or \mathbb{R} gauge theory. If the closed loop γ is the boundary of a two-dimensional surface S , a relationship denoted $\partial S = \gamma$, then we can use Stokes's theorem to write the Wilson loop observable in terms of the gauge-invariant field strength F integrated over S :

$$W_q(\gamma) = \exp\left(iq \oint_{\partial S} A\right) = \exp\left(iq \int_S F\right). \quad (48)$$

We have to be a bit careful about this: Stokes's theorem works when A is a well-defined 1-form in the region of interest, but in general, gauge fields need not be single-valued. However, we can carry out a gauge transformation so that the “Dirac string” where A is poorly defined (see §4.5.2) is outside the region S . Suppose that we consider two *different* surfaces, S_1 and S_2 , both bounded by γ . Because the integral of F over both surfaces corresponds to the same Wilson loop, we can combine the two to form a trivial operator. That is, we make a closed surface Σ , without boundary, by combining S_1 with the orientation-reversed surface $-S_2$: $\Sigma = S_1 \cup (-S_2)$, as depicted in Fig. 5. Thus, we have

$$\exp\left(iq \int_{\Sigma} F\right) = \exp\left(iq \int_{S_1} F\right) \exp\left(-iq \int_{S_2} F\right) = \exp\left(iq \oint_{\partial S} A\right) \exp\left(-iq \oint_{\partial S} A\right) = 1. \quad (49)$$

Conversely, given any closed surface Σ , we can draw a closed loop $\gamma \subset \Sigma$ dividing Σ up into two regions and use this argument to conclude that $\exp\left(iq \int_{\Sigma} F\right) = 1$ for any allowed charge q . (When there is nonzero flux, we have to be a bit careful, defining each integral in a gauge where the Dirac string does not go through the surface of interest.) From this we learn that:

- In \mathbb{R} gauge theory, where q can take on any real value, we must have $\int_{\Sigma} F = 0$ for any closed two-dimensional surface Σ . In other words, *magnetic flux vanishes* in \mathbb{R} gauge theory. The theory forbids magnetic monopoles.
- In $U(1)$ gauge theory, we require $q \int_{\Sigma} F \in 2\pi\mathbb{Z}$ for any integer q . This is only possible if *magnetic flux is quantized*, i.e.,

$$\frac{1}{2\pi} \int_{\Sigma} F \in \mathbb{Z}. \quad (50)$$

These facts together form the statement of “Dirac quantization.” It is commonly said that the existence of magnetic monopoles requires electric charge to be quantized. Here we have approached this from the other direction: beginning by specifying our gauge group as $U(1)$, we learn that electric charge is quantized and magnetic monopoles are allowed. On the other hand, non-quantized electric charge requires a gauge group of \mathbb{R} , which we have found to be incompatible with magnetic monopoles.

Importantly, the quantization condition (50) is a property of any $U(1)$ bundle, meaning that it is obeyed by *every field configuration* that is summed over in the path integral. Electric flux quantization holds in a quite different way, as we will see below.

4.4 Brief aside on topology

Before I give an example of nonzero magnetic flux by using a space with a nontrivial topology, a brief remark is in order. The modern viewpoint on quantum field theory is that a QFT should have the ability to be defined not just in Minkowski space, but on arbitrary spacetime manifolds (or at least, those within some very general class). Formal quantum field theorists tend to take this as a given, and it has also become a commonplace in condensed matter theory, where many interesting phases of matter can be classified by the ground states that they have on spaces of nontrivial topology. However, the assumption that we can analyze QFTs by studying them on nontrivial spaces seems to be viewed with some skepticism (and occasional outright hostility) by some particle physicists. I think that the simplest justification that I can give you is that we live in a world with gravity, and we believe that a quantum theory of gravity is described (at least in a semiclassical limit) by a path integral that sums over different spacetime manifolds (including those with different topologies). Thus, any QFT that we eventually hope to couple with gravity should be compatible with spacetimes of nontrivial topology. Another justification is that conclusions that we can draw by studying QFT on nontrivial spacetimes often coincide with conclusions that we can draw from QFTs in the presence of interesting dynamical objects like magnetic monopoles or cosmic strings. Which objects are allowed (even as static probes of the theory, like Wilson lines) depends on the global structure of the gauge group, e.g., $SU(2)$ versus $SO(3)$ (and the classification can even depend on additional discrete data [35]). As we will discuss later, it is expected that in quantum gravity, all charged objects allowed by a gauge group actually exist, so these global choices have real physical implications. From this viewpoint, the claim (sometimes found in textbooks) that only projective representations matter in quantum field theory is too glib; it discards a useful mathematical tool for distinguishing between different theories.

Without further apology, I will proceed to analyze the structure of $U(1)$ gauge theory by placing it on topologically nontrivial spacetimes.

4.5 Configurations with magnetic flux

4.5.1 Example 1: flux on a torus

As a first example of U(1) magnetic flux, we consider a case where there is no magnetically charged object, but a magnetic flux arises because our theory is defined on a spacetime with nontrivial topology. Suppose our spacetime contains two periodic directions forming a torus T , parametrized by (x_1, x_2) where $x_i \cong x_i + 2\pi r_i$ ($i = 1, 2$). Then a field configuration with nonzero flux over the torus is given by a constant field strength F_{12} ,

$$\frac{1}{2\pi} \int_T F = \frac{1}{2\pi} \int_0^{2\pi r_1} dx_1 \int_0^{2\pi r_2} dx_2 F_{12} = 2\pi r_1 r_2 F_{12}. \quad (51)$$

Based on our general reasoning above, we know that this field configuration is allowed only if F_{12} is an integer multiple of $\frac{1}{2\pi r_1 r_2}$. To understand why only these values of F_{12} are allowed, let us try to construct a gauge field A whose field strength is F . Because $F_{12} = \partial_1 A_2 - \partial_2 A_1$, we can consider a case where A_2 increases linearly with x_1 , i.e., $A = F_{12} x_1 dx_2$. This is not a well-defined *function* because x_1 is only defined modulo $2\pi r_1$. However, it does define a *connection* if the difference between the original A and a transformed A with $x_1 \mapsto x_1 + 2\pi r_1$ is *gauge equivalent* to zero, i.e., if we can write

$$2\pi r_1 F_{12} dx_2 = -ie^{-i\alpha(x)} de^{i\alpha(x)} \quad (52)$$

for some U(1)-valued function $e^{i\alpha(x)}$. To accomplish this, we consider gauge transformations that wind around the x_2 direction, i.e.,

$$\exp(i\alpha(x)) = \exp\left(in \frac{x_2}{r_2}\right), \quad n \in \mathbb{Z}. \quad (53)$$

Comparing this to (52), we see that we have a sensible connection A in the case that

$$2\pi r_1 F_{12} dx_2 = \frac{n}{r_2} dx_2 \quad \Rightarrow \quad F_{12} = \frac{n}{2\pi r_1 r_2}. \quad (54)$$

This establishes that our field configuration obeys (50).

For \mathbb{R} gauge theory, $\alpha(x)$ lives on the real line, not on a circle. As a result, winding configurations like (53) do not exist, because no well-defined choice of $\alpha(x)$ can satisfy the equation. As a result, only $F_{12} = 0$ is permitted.

4.5.2 Example 2: the Dirac monopole

A magnetic monopole is an object that carries magnetic charge. If we integrate the magnetic flux around a monopole of magnetic charge $q_m \in \mathbb{Z}$, we should obtain $\int_{S^2} F = 2\pi q_m$. If we locate our magnetic monopole at the origin, one choice of F that achieves the correct magnetic flux is proportional to the volume form on S^2 , namely

$$F = \frac{1}{2} q_m \text{vol}_{S^2} = \frac{1}{2} q_m \sin \theta d\theta \wedge d\phi. \quad (55)$$

We could try to integrate this to find a gauge field A such that $F = dA$, but it turns out that no global solution works. For instance, you might try to write $A = -\frac{1}{2} q_m \phi \sin \theta d\theta$ or $A = -\frac{1}{2} q_m \cos \theta d\phi$.

However, neither of these are actually well-defined 1-forms on the whole space. The problem is that ϕ is not globally defined, and not single-valued where it is defined. In particular, ϕ degenerates along the z -axis where $\theta = 0, \pi$. This doesn't cause any problem for F , because the prefactor $\sin \theta$ is 0 precisely where the $d\phi$ factor stops making sense. However, it is a problem for A .

What we can do is define two *different* choices of A , each of which is valid in part of our space, and which are gauge-equivalent in the region where both are valid. Specifically, we define:

$$\begin{aligned} A_N &= \frac{1}{2} q_m (+1 - \cos \theta) d\phi, & \theta &\neq \pi. \\ A_S &= \frac{1}{2} q_m (-1 - \cos \theta) d\phi, & \theta &\neq 0. \\ A_N - A_S &= q_m d\phi = -ie^{-iq_m \phi} d(e^{iq_m \phi}), & \theta &\neq \{0, \pi\}. \end{aligned} \quad (56)$$

This gauge field configuration is known as the Dirac monopole. We have $dA_{N,S} = F$ wherever they are defined. When we pick a particular gauge, the locus where the gauge field is not valid in that gauge is known as the “Dirac string.” It is not a physical object, just an artifact of a choice of gauge. The “northern” gauge field A_N is valid everywhere except a Dirac string at the “south pole” at $\theta = \pi$; the “southern” gauge field A_S is valid everywhere except a Dirac string at the “north pole” at $\theta = 0$. In particular, $A_N \rightarrow 0$ at the north pole and $A_S \rightarrow 0$ at the south pole, so they are well-defined at these points even though $d\phi$ isn't. Everywhere that both are valid, they are related by a gauge transformation as in (44). As in the torus example from §4.5.1, the gauge field transformation has nontrivial *winding*, this time in the azimuthal angle ϕ . Unlike the case of the torus, the spacetime itself does not have any 1-cycles for the gauge transformation to wind around. However, the region on which the gauge fields overlap does: it is a slice of spacetime with the z -axis removed, allowing winding around the axis. The integer winding number q_m is the same as the number of units of magnetic charge carried by the monopole.

The Dirac monopole is a singular field configuration, in the sense that the energy density stored in the magnetic field diverges at the location of the monopole. This is nothing special about monopoles; it is really the same problem as the classical self-energy puzzle for the electron. However, as we will discuss in §22, the expected resolution of the puzzle is different. For the electron, the self-energy puzzle is resolved by quantum mechanics. For monopoles, we expect that it is resolved by the monopole having new physical structure inside a core region. A famous example is the *'t Hooft–Polyakov monopole*, which is a classical solution in the theory of an $SU(2)$ gauge field higgsed to $U(1)$ by the VEV of an adjoint scalar. In this case, the monopole core radius is of order m_W^{-1} ; the solution behaves like the Dirac monopole (56) at larger radii, but involves the full set of $SU(2)$ fields inside the core. The classical solution has finite energy. However, it is important to realize that this is just one example of how a magnetic monopole can arise. Not every $U(1)$ gauge theory has a non-abelian UV completion. There are other known examples where magnetic monopoles are really fundamental objects, like D-branes, and EFT breaks down completely inside the radius of their core.

4.5.3 Flux in integral cohomology

When we do not consider magnetic monopoles, we have $dF = 0$, i.e., F is a *closed* differential form. This means that F has an image in the de Rham cohomology of our spacetime manifold M ,

$[F] \in H_{\text{dR}}^2(M, \mathbb{R})$, which consists of closed forms modulo *exact* forms (those with $F = dA$ with A a well-defined 1-form). In fact, the quantization condition (50) implies that $\frac{1}{2\pi}F$ is a representative of a class in the *integral* cohomology,

$$\left[\frac{F}{2\pi} \right] \in H^2(M, \mathbb{Z}). \quad (57)$$

This will be important below when we discuss quantization of axion couplings. When we *do* have magnetic monopoles in the theory, we could consider M to be our spacetime manifold with monopole worldlines excised, and these statements will still apply. (However, there is also more to say in cases with degrees of freedom localized on monopole worldlines.)

4.6 Electric flux quantization; the Witten effect

We have seen that electric charge is quantized in U(1) gauge theory. This also leads to a quantization of electric flux. To discuss electric flux quantization, we have to include the kinetic term of the gauge field and its coupling to an electric current j_{el} . We can write this action in two equivalent forms, one with tensor index notation and one with differential form notation. As we discussed in the case of global symmetries, it is actually most natural to define a current as a $(d-1)$ -form, hence the 3-form J_{el} in 4d U(1) gauge theory. Thus we have:²

$$\begin{aligned} S &= \int d^4x \sqrt{-g} \left(-\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu} - A_\mu j_{\text{el}}^\mu \right) \\ &= \int \left(-\frac{1}{2e^2} F \wedge \star F - A \wedge J_{\text{el}} \right). \end{aligned} \quad (58)$$

Note that the gauge field here is not canonically normalized, due to the $1/e^2$ in front of the kinetic term. We have been using this non-canonical normalization all along, because it makes formulas like the Wilson line simple. Charge is manifestly quantized in integer units in this normalization. One can easily translate between these normalizations; the canonical gauge field $\hat{A} = A/e$.

From the action (58), we can derive the equation of motion for A :

$$\frac{1}{e^2} d\star F = J_{\text{el}}. \quad (59)$$

This is just the familiar Maxwell equation (in curved spacetime) $\frac{1}{e^2} \nabla^\mu F_{\mu\nu} = j_\nu$, written in the language of differential forms. An immediate consequence of this equation is that, in QED, given a 3-manifold Ω in spacetime we have

$$Q(\Omega) = \int_\Omega J_{\text{el}} = \frac{1}{e^2} \int_\Omega d\star F = \frac{1}{e^2} \int_{\partial\Omega} \star F. \quad (60)$$

In particular, if Ω is a closed manifold (i.e., compact and without boundary), then $Q(\Omega) = 0$. This is just Gauss's law: on a compact space without boundary, there can be no net electric charge, because all field lines must end somewhere. Notice that Gauss's law follows from the fact that J_{el} is not just a closed form (i.e., a conserved current), but an *exact* one: it is d of another form. Generic

²In general, we may have to amend our definition of the current to generate "seagull terms" like the $A_\mu A^\mu |\phi|^2$ term that is familiar in scalar QED.

closed forms define global symmetries, but exact forms define *gauged* symmetries. This is worth emphasizing.

A conserved current J that is not just closed but *exact*, $J = dK$, is the hallmark of a symmetry that has been *gauged*. There is a corresponding Gauss's law constraint.

Another consequence is the quantization of electric flux. Suppose that Σ is a closed surface in spacetime that does not intersect any charged objects or nonzero currents, which bounds a 3-manifold Ω .³ Then (60) implies

$$\frac{1}{e^2} \int_{\Sigma} \star F \in \mathbb{Z}. \quad (61)$$

Notice that, unlike *magnetic* flux quantization (50), electric flux quantization depends on the spacetime metric (via the Hodge star) and on the prefactor $1/e^2$ for the gauge kinetic term. A related, important remark is that electric flux quantization, (61), is a statement that holds *for field configurations that obey the equations of motion of the theory*, as evidenced by the prominent role played by Maxwell's equation (59) in our discussion above. It is *not* a valid statement about arbitrary configurations that are summed over in the path integral. This is important, because otherwise one would have been able to view the gauge field kinetic term as a topological invariant, formed from the wedge product of the electric and magnetic flux densities. Instead, quantization of $\int F$ is topological but quantization of $\int \star F$ is dynamical: it holds only for those special field configurations which are saddle points of the path integral. In theories with electric-magnetic duality, like free Maxwell theory, these roles can be interchanged. It is important to realize, however, that to define the path integral in any duality frame, one must specify what U(1) gauge bundles are being summed over, which will always impose a topological constraint (50) on that U(1) field strength's fluxes in every field configuration in the path integral.

In fact, electric flux quantization in general depends on the detailed form of the Lagrangian. Given the coupling $A \wedge J$, we can insert a quantized charge $J = q \delta^{(3)}(\vec{x})$ and then work out which flux it sources. For example, if we added terms proportional to $(F_{\mu\nu} F^{\mu\nu})^2$ or $\bar{\Psi} \sigma^{\mu\nu} \Psi F_{\mu\nu}$ to our Lagrangian, then the quantized quantity on the left-hand side of (61) would be corrected. You might recall that the electric field in free Maxwell theory is the canonical conjugate of the gauge field. This generalizes to the statement that the “electric flux” whose integral is quantized is $\delta \mathcal{L} / \delta F$. A useful example is provided in the case of two U(1) gauge fields with kinetic mixing, as discussed in §5 below. Another important example arises when we add a θ term (24) to the theory. In this case, the variation of \mathcal{L} with respect to F acquires a new term, and so the electric flux quantization condition takes the form

$$\int_{\Sigma} \left(\frac{1}{e^2} \star F + \frac{\theta}{4\pi^2} F \right) \in \mathbb{Z}. \quad (62)$$

This fact is known as the Witten effect [36]. It implies that if we have a magnetic monopole of magnetic charge $p \in \mathbb{Z}$, which sources a magnetic flux $\frac{1}{2\pi} \int_{\Sigma} F = p$, then the electric charge of the

³Unlike for magnetic flux quantization, we will not discuss the case of homologically nontrivial Ω here. The general story will appear later, in §20, when we discuss the 1-form electric symmetry.

monopole, defined as $Q = \int_{\Sigma} \frac{1}{e^2} \star F$, is necessarily nonzero for generic θ :

$$Q = n - p \frac{\theta}{2\pi}, \quad n \in \mathbb{Z}. \quad (63)$$

The minimal electric charge is, in general, fractional. This does not violate charge quantization, in the sense that there is still a discrete set of possible charge assignments labeled by a lattice of integers $(n, p) \in \mathbb{Z}^2$. It does mean that the quantity we usually refer to as electric charge—what you would infer if you placed an electron near the object and measured the Coulomb force—no longer takes only integer values, because it is an irrational combination (63) of the integers n and p .

Another simple derivation of the Witten effect is given by considering the theory in a spatially varying background $\theta(x)$ that gradually turns on at some radius away from a magnetic monopole, and solving the modified Maxwell's equations, then taking a limit where θ becomes constant. You can find this in [37] or the lecture notes [28].

4.7 Canonically normalized gauge fields and the Ward identity

In the normalization in which we are working, J_{el} and A have normalizations that are fixed by topology: J_{el} is normalized so that the charges obtained by integrating it are integers, and A is normalized so that it transforms as (44) under $U(1)$ gauge transformations for which $\alpha(x) \cong \alpha(x) + 2\pi$. Thus, quantum corrections can only change the coefficient $\frac{1}{e^2}$ in front of the kinetic term, and indeed they do. The familiar QED beta function shows up in this kind of shift. At one loop,

$$S \mapsto \int \left[-\frac{1}{2} \left(\frac{1}{e^2(\Lambda)} + \frac{b}{8\pi^2} \log \frac{\Lambda}{\mu} \right) F \wedge \star F - A \wedge J_{\text{el}} \right], \quad (64)$$

where we identify the quantity in parentheses as the running coupling $1/e^2(\mu)$:

$$\frac{1}{e^2(\mu)} = \frac{1}{e^2(\Lambda)} + \frac{b}{8\pi^2} \log \frac{\Lambda}{\mu}. \quad (65)$$

Here b is a beta function coefficient, equal to $\frac{2}{3} \sum_i q_i^2$ in a theory with a collection of Dirac fermions of charge q_i .

This way of understanding the running of the electromagnetic coupling is one advantage of working in a non-canonical normalization where charge quantization is manifest. Compare the textbook approach: one introduces separate rescaling factors for the kinetic term $F_{\mu\nu}^2$, the kinetic term of charged fields, and the coupling term $A_\mu J_{\text{el}}^\mu$. It seems not at all obvious that the renormalization of the three-point electron-positron-photon vertex should have anything to do with the vacuum polarization diagram that renormalizes the photon two-point function! However, eventually, one finds that the Ward identity guarantees that renormalization of the gauge coupling e is completely determined by the renormalization of the photon kinetic term. This is often argued by complicated diagrammatic analysis. When we work in the normalization with manifest $U(1)$ charge quantization, it is just obvious: there is no coupling in front of the $A \wedge J_{\text{el}}$ term, only in front of the $F \wedge \star F$ term, so the latter is the only thing that can run! After working this out, we are then always free to go back to canonical normalization.

5. Kinetic mixing of U(1)s

In recent years, “millicharged particles” have been frequently considered in the study of dark matter or physics beyond the Standard Model more generally. In this context, the “milli-” prefix just means “very small,” not specifically 10^{-3} . I have told you that U(1) charge is quantized, so how could we have a millicharge? It arises in theories with multiple gauge fields that kinetically mix with each other [38, 39], and is perfectly compatible with charge quantization.

Consider a theory with gauge group $U(1)_A \times U(1)_B$, where the group elements for the gauge transformations are parametrized by $e^{i\alpha(x)}$ and $e^{i\beta(x)}$ respectively. We label the gauge fields A and B . A field ψ can transform in the (q_A, q_B) representation, with

$$\psi(x) \mapsto e^{i[q_A \alpha(x) + q_B \beta(x)]} \psi(x). \quad (66)$$

From this it is clear that *both* $q_A, q_B \in \mathbb{Z}$ are quantized, since α and β are only defined modulo 2π .

The Lagrangian for the A and B gauge fields can have a kinetic mixing parameter κ , and currents J_A, J_B coupled to the gauge fields:

$$-\frac{1}{2e_A^2} F_A \wedge \star F_A - \frac{\kappa}{e_A e_B} F_A \wedge \star F_B - \frac{1}{2e_B^2} F_B \wedge \star F_B - A \wedge J_A - B \wedge J_B \quad (67)$$

In this basis charge conservation is completely manifest, $\int_{\Sigma_3} J_{A,B} \in \mathbb{Z}$, and each gauge field couples to a quantized charge. In particular, we have the usual magnetic flux quantization conditions:

$$\frac{1}{2\pi} \int_{\Sigma_2} F_A \in \mathbb{Z}, \quad \frac{1}{2\pi} \int_{\Sigma_2} F_B \in \mathbb{Z}. \quad (68)$$

Notice that these hold independently for each U(1) gauge group.

Nonetheless, a particle charged under A and a particle charged under B can scatter through the kinetic mixing. There is a propagator that connects an A vertex on one end to a B vertex on the other. In this sense, the scattering amplitude of an electron e^- charged under A and a dark fermion f charged under B will *appear* as if f carries a small charge under A . One manifestation of this is that the *electric* flux quantization conditions are nonstandard. In particular, a particle charged under A will source a B field, despite having no direct coupling. The equations of motion tell us that

$$\int \left(\frac{1}{e_A^2} \star F_A + \frac{\kappa}{e_A e_B} \star F_B \right) \in \mathbb{Z}, \quad \int \left(\frac{1}{e_B^2} \star F_B + \frac{\kappa}{e_A e_B} \star F_A \right) \in \mathbb{Z}. \quad (69)$$

This makes clear that if we insert a particle of charge q under B , sourcing a flux $q = \int \frac{1}{e_B^2} \star F_B$, we will inevitably also source a small flux of A , namely (to order κ)

$$\int \frac{1}{e_A^2} \star F_A = - \int \frac{\kappa}{e_A e_B} \star F_B = - \frac{\kappa e_B}{e_A} q. \quad (70)$$

This is the “millicharge” that B -charged particles carry under the photon A . (Notice the close similarity to the discussion of the Witten effect (62), although that case was a mixing of magnetic and electric charge under a single U(1), rather than of electric charges under two different U(1)s.)

Another way to see that particles carrying different charges interact through the kinetic mixing is to do a field redefinition. We generally choose to maintain the definition of A as the field to which

the electron couples. However, we can redefine $B \mapsto B - \frac{\kappa e_B}{e_A} A$, which cancels the kinetic mixing term. This choice gives familiar-looking Feynman rules where each particle has an independent propagator. However, it also means that the gauge fields in the new basis no longer couple to quantized charges: we now have a coupling $A \wedge J_B$ with an irrational coefficient (the millicharge). Another consequence is that the *magnetic* flux quantization conditions now take on an unusual form: instead of $\frac{1}{2\pi} \int F_B \in \mathbb{Z}$, we now have $\frac{1}{2\pi} \int \left(F_B - \frac{\kappa e_B}{e_A} F_A \right) \in \mathbb{Z}$. There is nothing wrong with such a basis. It is well-suited for perturbative calculations, but it does make questions of charge quantization more obscure.

Exercise: Consider a case where the second photon, B , is massive. This could happen through the Higgs mechanism, but for the current problem you can simply add an explicit $B_\mu B^\mu$ term (which you can write in Stueckelberg form to make it manifestly gauge invariant, if you like). Explain how a field redefinition can simultaneously diagonalize the kinetic and mass terms. Investigate the charge and flux quantization conditions in the new frame. This is the well-studied case of a “dark photon” particle, often studied as a potential mediator between the Standard Model and dark matter.

Notice that, although the preferred choice of basis might shift, the physics of the case of massless and massive B is completely continuous as the mass is taken to zero. (At least one popular review article about dark photons is highly misleading on this point.)

To summarize, in a theory with multiple $U(1)$ gauge groups, the charges (both magnetic and electric) are always quantized in a lattice. However, particles carrying different charge can scatter with each other through Coulomb interactions mediated by the off-diagonal kinetic terms. Depending on the physical question that one wants to ask, one basis or another might be better suited, and in a given basis it might appear that there can be a non-quantized millicharge. None of this changes the underlying group theoretic fact of charge quantization.

Part Two: Instantons and Chiral Anomalies

6. U(1) fields with nonzero $\int F \wedge F$

We saw above that U(1) gauge theory obeys a flux quantization condition: $\frac{1}{2\pi} \int F \in \mathbb{Z}$. Much of our following discussion will be about axion fields that couple to $F \wedge F$, where a generalized “squared-flux quantization” condition will play an important role, both for U(1) gauge fields and for SU(N) gauge fields. A full mathematical treatment of these conditions is beyond the scope of these lectures, but can be found in textbooks that discuss characteristic classes, e.g., [34, 40]. Rather than giving a general proof, let’s look at an example that happens to give us the right answer.

In §4.5.1, we showed that in a theory where two spatial dimensions form a torus, configurations exist with flux n , given by $F = \frac{n}{2\pi r_1 r_2} dx_1 \wedge dx_2$. Then it is clear that we can define a configuration on a four-dimensional torus that has a nonzero $\int F \wedge F$ simply by taking

$$F = \frac{n}{2\pi r_1 r_2} dx_1 \wedge dx_2 + \frac{m}{2\pi r_3 r_4} dx_3 \wedge dx_4, \quad n, m \in \mathbb{Z}. \quad (71)$$

Indeed, direct computation shows that

$$F \wedge F = \frac{nm}{2\pi^2 r_1 r_2 r_3 r_4} dx_1 \wedge dx_2 \wedge dx_3 \wedge dx_4, \quad (72)$$

and hence

$$\int F \wedge F = (2\pi)^4 r_1 r_2 r_3 r_4 \frac{nm}{2\pi^2 r_1 r_2 r_3 r_4} = 8\pi^2 nm. \quad (73)$$

Thus we see that, for this class of field configurations, the smallest possible value of $\int F \wedge F$ is $8\pi^2$, and the attainable values are all integer multiples of this.

It turns out that, even though we focused on a particular example, this conclusion is the correct one for our purposes. The full mathematical story is somewhat more subtle. The quantization of magnetic flux (57) implies that

$$\left[\frac{F}{2\pi} \right] \smile \left[\frac{F}{2\pi} \right] \in H^4(M, \mathbb{Z}), \quad (74)$$

and hence that

$$\frac{1}{4\pi^2} \int_M F \wedge F \in \mathbb{Z}, \quad \text{for any closed } M. \quad (75)$$

Thus, the base unit of $\int F \wedge F$ is, in complete generality, $4\pi^2$ rather than $8\pi^2$. Any example, as in our torus context, that constructs M as a product of two two-manifolds with flux will lead to $\frac{1}{4\pi^2} \int F \wedge F$ being an *even* integer. However, it is possible to define a gauge field configuration on the complex projective space \mathbb{CP}^2 (which does not have such a product form) for which $\frac{1}{4\pi^2} \int F \wedge F = 1$. On the other hand, \mathbb{CP}^2 does not admit a spin structure, meaning that we cannot consistently define a quantum field theory with spinor fields on this manifold.

It is a mathematical fact that, for any *spin* 4-manifold, the integral of the square of the magnetic flux is an even integer, or equivalently

$$\frac{1}{8\pi^2} \int_M F \wedge F \in \mathbb{Z}, \quad \text{for any closed, spin } M. \quad (76)$$

Since we will exclusively be interested in theories with fermions, as in the real world, this is the quantization condition that will be relevant for us.

There are no $U(1)$ gauge field configurations on S^4 for which $\int F \wedge F \neq 0$; the structure (74) implies that we only find such configurations on spaces with 2-cycles. As we will see in section §8, the story is different for the gauge group $SU(N)$.

7. Non-abelian gauge fields

We will not give a detailed review of non-abelian gauge theory in these notes, as it is covered adequately in many quantum field theory textbooks. It is useful, however, for us to give a very quick summary of how some of the standard formulas look when written in differential form notation, and in particular we should see how differential form notation meshes with familiar non-abelian matrix notation. A (continuous) nonabelian gauge group has a set of generators T^a whose commutation relations define the structure constants f^{abc} :

$$[T^a, T^b] = i f^{abc} T^c. \quad (77)$$

The vector space of real linear combinations of these generators, together with the commutation relations, defines a Lie algebra. We define an appropriate field strength F that transforms linearly under a gauge transformation, $F \mapsto g F g^{-1}$, with components

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + f^{abc} A_\mu^b A_\nu^c. \quad (78)$$

Our gauge field A is (locally) a 1-form taking values in the Lie algebra,

$$A \equiv A_\mu dx^\mu \equiv A_\mu^a T^a dx^\mu. \quad (79)$$

The field strength F is then a Lie algebra-valued 2-form,

$$F = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu = \frac{1}{2} (\partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu]) dx^\mu \wedge dx^\nu = dA - iA \wedge A. \quad (80)$$

The final step gives the most convenient differential form notation for the field strength, which implicitly encodes the matrix algebra because

$$A \wedge A = A_\mu^a T^a A_\nu^b T^b dx^\mu \wedge dx^\nu = \frac{1}{2} A_\mu^a A_\nu^b i f^{abc} T^c dx^\mu \wedge dx^\nu. \quad (81)$$

The standard kinetic term for a non-abelian gauge theory is

$$\int \left(-\frac{1}{g^2} \text{tr}(F \wedge \star F) \right), \quad (82)$$

where for $SU(N)$ gauge theory we conventionally write the generators T^a in the fundamental representation normalized such that $\text{tr}(T^a T^b) = \frac{1}{2} \delta^{ab}$. (It is because of this $1/2$ that we simply have $-1/g^2$ in (82), rather than the $-1/(2e^2)$ factor in the $U(1)$ case (20).)

In working with non-abelian gauge theories, we frequently encounter traces. When one has differential forms appearing inside a trace, one must be careful with minus signs when applying

standard identities like the cyclic property of the trace. For example, if ω and η are respectively Lie algebra-valued p - and q -forms, then we have

$$\text{tr}(\omega \wedge \eta) = (-1)^{pq} \text{tr}(\eta \wedge \omega). \quad (83)$$

This combines the familiar graded commutativity of the differential forms with the familiar cyclic property of the trace. Simple applications are:

$$\begin{aligned} \text{tr}(A \wedge F) &= \text{tr}(F \wedge A), \\ \text{tr}(A \wedge A) &= -\text{tr}(A \wedge A) = 0, \\ \text{tr}(A \wedge A \wedge A \wedge A) &= 0. \end{aligned} \quad (84)$$

This last identity will make an appearance below, in §11.

8. Instantons

“You can’t eat an instanton.” – John Stout

8.1 The BPST instanton solution and its properties

We have just constructed field configurations with nonzero $F \wedge F$ in a $U(1)$ gauge theory. This was relatively straightforward, once we allow ourselves to work on a space of nontrivial topology, because we can just exploit the winding of $U(1)$ gauge transformations around a circle. For $SU(N)$ gauge theory, there is a similar story, but the topology that we exploit is a bit different. In this case, it is possible to construct a classical solution to the Euclidean Yang-Mills equations of motion on S^4 , which is *localized* in Euclidean spacetime and has a nonzero $\int \text{tr}(F \wedge F)$. This is known as the BPST (Belavin, Polyakov, Schwarz, Tyupkin) instanton solution [41]. For the gauge group $SU(2)$, an explicit solution for a family of such solutions is given that depends on five real parameters: a four-vector of positions x_0^μ and a “size modulus” ρ . (When a classical solution depends on parameters in this way, they are variously referred to as “zero modes,” “moduli,” or “collective coordinates.”) Namely,

$$A_\mu^a(x) = \frac{2\eta_{\mu\nu}^a (x - x_0)^\nu}{(x - x_0)^2 + \rho^2}. \quad (85)$$

Here μ is a (Euclidean) spatial index and a is an $SU(2)$ adjoint gauge index ($a \in \{1, 2, 3\}$). Here $\eta_{\mu\nu}^a$ is known as the “’t Hooft symbol” [42] and is defined as

$$\eta_{\mu\nu}^a = \begin{cases} \epsilon^{a\mu\nu}, & \mu, \nu \in \{1, 2, 3\} \\ -\delta^{a\nu}, & \mu = 0 \\ +\delta^{a\mu}, & \nu = 0 \\ 0, & \mu = \nu = 0. \end{cases} \quad (86)$$

The first thing you might notice is that we have written this as a solution in flat Euclidean space \mathbb{R}^4 . However, it can be thought of as a solution on the compact space S^4 , with a point at infinity added to \mathbb{R}^4 . The reason is that this solution decays away quickly at $|x| \rightarrow \infty$, as $F \sim |x|^{-4}$, so the

field strength goes to zero at the extra point at infinity (where A_μ^a asymptotes to a pure gauge field configuration).

This solution has the special property that it is *self-dual*, i.e.,

$$F_{\mu\nu}^a = \tilde{F}_{\mu\nu}^a. \quad (87)$$

There is a different solution which is *anti-self-dual*, $F_{\mu\nu}^a = -\tilde{F}_{\mu\nu}^a$. It looks exactly the same except that the 't Hooft symbol is replaced by a different symbol $\bar{\eta}_{\mu\nu}^a$, which exchanges the role of μ and ν in (86). The 't Hooft symbol may seem mysterious. Some rough intuition for where it comes from is the following: the Euclidean theory has an $SO(4)$ rotational symmetry. There is a double cover of $SO(4)$ by the group $SU(2) \times SU(2)$. One can search for a classical solution that breaks the product of *one* spatial $SU(2)$ and the gauge $SU(2)$ to the diagonal. This is loosely what is going on with the 't Hooft symbol, which mixes up spatial and internal $SU(2)$ indices.⁴

The self-dual instanton solution has two key properties that will be at the core of the physics of axions that we will discuss soon. First, it has a nontrivial integral $\int_{S^4} \text{tr}(F \wedge F)$:

$$\int_{S^4} \text{tr}(F \wedge F) \Big|_{\text{inst.}} = 8\pi^2. \quad (88)$$

The anti-self-dual instanton solution has $-8\pi^2$ for this integral. In fact, one can show that *any* $SU(N)$ gauge field configuration on any closed 4-manifold has a quantized integral of $\text{tr}(F \wedge F)$ which is an integer multiple of this base unit:

$$\frac{1}{8\pi^2} \int_{\Sigma} \text{tr}(F \wedge F) \in \mathbb{Z}. \quad (89)$$

Like the quantization of magnetic flux, this is a property of *any field configuration* that we sum over in the path integral, not just of solutions to the equations of motion. We call this integer the *instanton number* of the field configuration. (It is also sometimes called the “topological charge,” though this is a bit more vague because it could have other meanings in different contexts.)

The second key property is that the instanton solution has a large Euclidean action, diverging in the limit $g \rightarrow 0$:

$$\int_{S^4} \left(\frac{1}{g^2} \text{tr}(F \wedge \star F) \right) \Big|_{\text{inst.}} = \frac{8\pi^2}{g^2}. \quad (90)$$

This shows that the instanton is a non-perturbative effect: the *canonically* normalized gauge field solution scales as $1/g$. (Notice that the Euclidean expression for the kinetic term has the opposite sign from the Minkowski action in (82).)

We can understand the special role of self-dual field configurations in this story using a Bogomol'nyi trick. This is a rewriting of the action as a sum of a perfect square and a topological invariant, which is useful in studying a number of interesting topological objects in quantum field theory. In this context, the Euclidean action can be rewritten as

$$\int \frac{1}{g^2} \text{tr}(F \wedge \star F) = \int \frac{1}{g^2} \left\{ \frac{1}{2} \text{tr} [(F \mp \star F) \wedge \star (F \mp \star F)] \pm \text{tr}(F \wedge F) \right\}. \quad (91)$$

⁴A reader might wonder why I am discussing this in terms of $SU(2)$ rather than $SO(3)$; the reason is that $SU(N)/\mathbb{Z}_N$ has a different quantization of instanton number than $SU(N)$, allowing “fractional instantons.” This is related to all sorts of interesting physics that I will omit to keep these lecture notes to a vaguely manageable length.

The Bogomol'nyi trick makes it obvious that a self-dual field is always a solution to the full Yang-Mills equations of motion. This is because $\int \text{tr}(F \wedge F)$ is a topological invariant, so a small variation away from a self-dual field cannot change the second term in (91). On the other hand, the first term (with the upper sign choice) vanishes for a self-dual field, and is positive otherwise, so it can only increase when the field is varied. As a result, a self-dual field is necessarily a local minimum of the Euclidean action. The same holds for an anti-self-dual field, with the lower sign choice. Because self-dual and anti-self-dual field configurations have positive and negative instanton number n respectively, we see that they all have $S = 8\pi^2|n|/g^2$.

8.2 Instantons in the path integral

The path integral for non-abelian gauge theory sums over all field configurations in spacetime. Because the instanton number n in (89) is a topological invariant, one can decompose the path integral into a discrete sum over $n \in \mathbb{Z}$ together with a continuous integral over topologically trivial differences in field configurations for every n . The sum over nontrivial topologies is necessary: restricting to the $n = 0$ sector is actually inconsistent with locality and unitarity. For example, a spacetime that pinches off into two separate spacetimes can have a total instanton number zero but have nonzero instanton numbers in its constituent parts. Thus, compatibility with basic axioms of QFT requires a sum over topological sectors. However, it was appreciated in recent years that it is possible to restrict the sum to only values of n that are multiples of a fixed nonzero integer $p \in \mathbb{Z}$ [43]. Such theories have identical local physics, but differ in the correlation functions of nonlocal operators; in particular, they have a \mathbb{Z}_p 3-form symmetry [44] (see §20 for this terminology, although not this application).

When instantons were first discovered, there were hopes that they might help to analytically understand the strongly coupled physics of confinement, because they contribute calculable non-perturbative effects in the path integral [45]. One can treat them with semiclassical methods: the instanton solutions are saddle points in the path integral, and we can integrate over their collective coordinates or moduli, like the instanton position and size. Unfortunately, these methods do not take us very far. The semiclassical method works well when the instanton action $8\pi^2/g^2$ is large, i.e., when g is small, and so we should expect that in the infrared, where a non-abelian gauge theory confines because g becomes large, semiclassical methods will also fail. In fact, when one computes the measure for integrating over the instanton size parameter ρ , one finds that it modifies the instanton measure in the path integral for $\text{SU}(N)$ Yang-Mills theory with a prefactor [5, 6, 46]

$$\left(\frac{8\pi^2}{g^2}\right)^{2N} \frac{1}{\rho^5} (M\rho)^{11N/3} \exp\left(-\frac{8\pi^2}{g^2}\right). \quad (92)$$

Here M is a UV regulator mass scale, and we should understand the coupling g to be the running coupling evaluated at the scale M . The exponent $11N/3$ should look familiar: it is the one-loop beta function coefficient in Yang-Mills theory. This is no accident, as it allows the M -dependent prefactor to be combined with the exponential of the instanton action to form

$$(M\rho)^{11N/3} \exp\left(-\frac{8\pi^2}{g(M)^2}\right) = \exp\left(-\frac{8\pi^2}{g(M)^2} + \frac{11N}{3} \log(M\rho)\right) = \exp\left(-\frac{8\pi^2}{g(\rho^{-1})^2}\right). \quad (93)$$

In other words, the measure for integrating over the size modulus effectively replaces the coupling g evaluated at the UV scale M in the instanton action with the coupling evaluated at the scale

ρ^{-1} . For large values of ρ , this running coupling becomes large, and the calculation breaks down. In particular, the integral over ρ with this measure diverges. We say that a “dilute instanton gas approximation” is valid for small ρ , but not for large ρ . This limits the phenomenological utility of semiclassical instanton calculations in QCD. We will see a consequence when we discuss the axion solution to the Strong CP problem in §12.2: there is a loose sense in which the axion potential is generated “by instantons” (as is often said colloquially), which is that axions couple to the instanton number density $\text{tr}(F \wedge F)$. However, the axion potential cannot be calculated using semiclassical instanton methods, and is instead obtained from a conceptually different approach using the chiral Lagrangian.

8.3 Comments on instantons

The sage words of my collaborator John Stout at the opening of this section, “you can’t eat an instanton,” refer to the fact that an instanton is not an object that exists in real time. It is localized in Euclidean time as well as in space. It has no worldline. It lives at one *instant*—hence the name, due to ’t Hooft. (BPST called it a “pseudoparticle,” but ’t Hooft’s term became the standard one.) However, in more than four dimensions, BPST instantons *do* become ordinary dynamical objects. For example, in a 5d theory, we can consider the solution (85) as a function of four spatial coordinates, and just tack on extra time coordinate that the solution does not depend on. This now describes a kind of *particle*, of size ρ , localized at x_0 and staying put for all time. Such a particle is an honest, dynamical, solitonic object in the theory. If you lived in 5d, you *could* eat an instanton. In 6d a BPST instanton is a string; in 7d it is a membrane with two spatial dimensions (a “2-brane”); and so on. (In 10d it is a 5-brane, which might ring a bell for any of you who have studied string theory; in fact, NS5 branes are intimately related to BPST instantons in heterotic string theory.) Here we should mention an annoying feature of language: in *any* number of spacetime dimensions, the word “instanton” can be used to refer to a localized solution of the Euclidean equations of motion: a zero-dimensional object, localized in spacetime. On the other hand, the term “BPST instanton” or “Yang-Mills instanton” (or even just “instanton”) can refer to the specific solution (85), which is a $(d - 4)$ -dimensional object. This is one of those ambiguities in language that you just have to learn to resolve from contextual clues (or directly asking someone which they mean).

9. The chiral ABJ and ’t Hooft anomalies

I suggest that even readers who are thoroughly familiar with the chiral anomaly read 9.5, which summarizes the key way in which we will be making use of the anomaly to discuss axion physics in the subsequent sections.

9.1 Introductory remarks

In this section we will briefly review the *chiral anomaly*, in two forms: the Adler–Bell–Jackiw or ABJ anomaly (explicit symmetry breaking) [47, 48] and the ’t Hooft anomaly (obstruction to gauging) [49]. There are several different QFT calculations that reveal these anomalies, which require defining a regulator and using it carefully to obtain a regulator-independent physical result in the end. These calculations are somewhat subtle and are discussed in great detail in standard textbooks, so I will not reproduce them in depth here. In particular, the first four sections of chapter

19 of the textbook by Peskin and Schroeder [7] contain several calculations of the chiral anomaly from different viewpoints. I strongly encourage you to work through these calculations carefully. Here, I will simply summarize these standard arguments and highlight some conceptual aspects of these anomalies. I will also outline a derivation by Nielsen and Ninomiya, which is less of a standard textbook treatment, and which gives a picture of the anomaly in terms of a physical process of particle production [50].

We will study these anomalies in two different, related QFTs. The first is simply a theory of two free, massless Weyl fermions,

$$\frac{1}{\sqrt{|g|}} \mathcal{L} = \psi^\dagger i \not{\partial} \psi + \bar{\psi}^\dagger i \not{\partial} \bar{\psi}, \quad (94)$$

with $\not{\partial} \equiv \bar{\sigma}^\mu \partial_\mu$. Notice that here ψ and $\bar{\psi}$ are just *names* of two *different* left-handed Weyl fermions; I use the notation ψ^\dagger for the hermitian conjugate of ψ , so don't interpret the bar over $\bar{\psi}$ as a conjugate. This theory has a $U(1)_L \times U(1)_R$ global symmetry, where the two $U(1)$ factors rotate the two fermions *independently*: $\psi \mapsto e^{i\alpha} \psi$ and $\bar{\psi} \mapsto e^{-i\beta} \bar{\psi}$. In other words, the charges (q_L, q_R) of ψ and $\bar{\psi}$ under this group are $(1, 0)$ and $(0, -1)$, respectively. There is a $U(1)$ *diagonal* subgroup of this symmetry, denoted $U(1)_V$, given by the elements where $\beta = \alpha$. Under $U(1)_V$, ψ has charge 1 and $\bar{\psi}$ has charge -1 . The V subscript here stands for “vector-like”; if we package ψ and $\bar{\psi}^\dagger$ into a single Dirac fermion Ψ , this would simply be the rotation $\Psi \mapsto e^{i\alpha} \Psi$ that would be preserved by a Dirac mass term $m \bar{\Psi} \Psi$ (where here $\bar{\Psi}$ *does* mean the Dirac conjugate of Ψ). (The idea of packaging the two fields into a single Dirac fermion is also the origin of our labels “L” and “R,” as well as the sign convention we put on β , but from the point of the theory (94) taken on its own merits, which after all describes two completely unrelated and non-interacting fields, these conventions seem like idiosyncratic historical artifacts.) These symmetries have corresponding Noether currents

$$j_L^\mu = \psi^\dagger \bar{\sigma}^\mu \psi, \quad j_R^\mu = -\bar{\psi}^\dagger \bar{\sigma}^\mu \bar{\psi}, \quad j_V^\mu = j_L^\mu + j_R^\mu. \quad (95)$$

We will see that this theory has an ‘t Hooft *anomaly*: although $U(1)_L \times U(1)_R$ is a perfectly good global symmetry of the quantum field theory, the full group *cannot be gauged*.

In the second theory, we also have a $U(1)$ gauge field, which has gauged only part of our original symmetry group, namely $U(1)_V$:

$$\frac{1}{\sqrt{|g|}} \mathcal{L} = -\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu} + \psi^\dagger i \not{D} \psi + \bar{\psi}^\dagger i \not{D} \bar{\psi}, \quad (96)$$

where the covariant derivatives are

$$\not{D} \psi \equiv \bar{\sigma}^\mu (\partial_\mu - i A_\mu) \psi, \quad \not{D} \bar{\psi} \equiv \bar{\sigma}^\mu (\partial_\mu + i A_\mu) \bar{\psi}. \quad (97)$$

Taken as a classical Lagrangian, this theory has a global symmetry that acts as $U(1)_L$, i.e., it acts as $\psi \mapsto e^{i\alpha} \psi$ but does not touch $\bar{\psi}$. In the quantum field theory, this is not a symmetry at all! Its associated Noether current is not actually conserved. This is an example of an *ABJ anomaly*, when a symmetry of the classical action is not a symmetry of the quantum field theory. (We could say the same thing about $U(1)_R$.)

These two different types of anomalies are clearly related to each other. The ABJ anomaly, in a sense, triggers the 't Hooft anomaly. If we tried to gauge the full $U(1)_L \times U(1)_R$ symmetry of (94), we could do so by first gauging the subgroup $U(1)_V$. We would not encounter an obstruction. But if we then tried to gauge the rest of the group, we would be faced with the ABJ anomaly of (96): the current is not conserved, and so we can't couple a gauge field to it. The 't Hooft anomaly of (94) can alternatively be thought of as a *precursor* to the ABJ anomaly of (96).

9.2 The particle production calculation

The first argument I want to present for the chiral anomaly is arguably the most physical. It computes how particles are produced in the presence of a time-dependent gauge field background, and how the resulting particle production violates chiral symmetry. It was first published (as far as I am aware) by Nielsen and Ninomiya [50], and is discussed in a recent textbook by Fradkin [8]. A nice treatment of it can also be found in lectures by David B. Kaplan [51].

9.2.1 The (1 + 1)d anomaly

For this calculation it is easiest to explain the (1 + 1)d case first; this is an interesting result in its own right, as well as a key input to the (3 + 1)d calculation. Our goal is to study the behavior of chiral fermions in the presence of a background electric field. In (1 + 1)d, space is a line, and the analogue of chirality is whether the fermion is left-moving or right-moving. A left-moving Weyl fermion ψ_- and right-moving Weyl fermion ψ_+ coupled to a background gauge field A_μ have an action

$$\int d^2x \sqrt{|g|} \left(i\psi_+^\dagger D_- \psi_+ + i\psi_-^\dagger D_+ \psi_- \right), \quad (98)$$

where

$$D_\pm = (\partial_t \pm \partial_x) + i(A_t \pm A_x). \quad (99)$$

In particular, if we turn off the gauge field A , we have equations of motion $\partial_+ \psi_- = 0$ and $\partial_- \psi_+ = 0$, which are solved by allowing ψ_- to be an arbitrary function of $t + x$ and ψ_+ to be an arbitrary function of $t - x$ (hence, left- and right-moving respectively). There is an important difference between (1 + 1)d chirality and (3 + 1)d chirality: in (3 + 1)d, the conjugate ψ^\dagger of a left-handed Weyl fermion is a right-handed Weyl fermion, whereas in (1 + 1)d, the conjugate ψ_-^\dagger of a left-mover is still a left-mover.

The classically conserved 1-form currents in this theory are

$$\begin{aligned} J_+ &= \sqrt{|g|} \left(+\psi_+^\dagger \psi_+ dt + \psi_+^\dagger \psi_+ dx \right), \\ J_- &= \sqrt{|g|} \left(-\psi_-^\dagger \psi_- dt + \psi_-^\dagger \psi_- dx \right). \end{aligned} \quad (100)$$

The gauge field A couples as $A \wedge (J_+ + J_-)$.

Using the ansatz $\psi \propto \exp(-i\omega t + ipx)$, the equations of motion $\partial_t \psi_+ = \partial_x \psi_+$ and $\partial_t \psi_- = -\partial_x \psi_-$ indicate that the right- and left-moving fermions have dispersion relation $\omega(p) = -p$ and $\omega(p) = +p$, respectively. It is useful to think in terms of the ‘‘Dirac sea’’ picture, so that initially all negative-frequency states, corresponding to $p \geq 0$ for ψ_+ and $p \leq 0$ for ψ_- , are filled. In general, I tend to think of the Dirac sea as a historical artifact that confuses more than it explains, but in this particular

context it seems to be the simplest way to discuss the physics. If you are uncomfortable with this, you should be able to restate all of the following in terms of Bogoliubov coefficients and particle creation.

Now, consider the theory in a background electric field $F = E(t) dt \wedge dx$, where $E(t)$ is adiabatically turned on at some time, and subsequently adiabatically turned off. In the presence of the electric field, the charged particles accelerate: their momentum changes according to $dp/dt = E(t)$. For concreteness, let's suppose that $p_E \equiv \int E(t) dt \geq 0$. For ψ_- , we initially had populated all states with $p \leq 0$. Each of these states increases in momentum, so at the time the electric field is switched off, we have populated a *larger* range of states, up to p_E . On the other hand, for ψ_+ , we initially had populated all states with $p \geq 0$. Each of these states increases in p , so at the time the electric field is switched off, we have populated a *smaller* range of states, $p \geq p_E$. The result is a net depletion of right-moving states and a net increase in the population of left-moving states, i.e., a creation of a chiral asymmetry. This is depicted in Fig. 6.

This story sounds a bit odd, because I've phrased it as a population of particle states that individually shift in momentum, which doesn't *sound* like it should be creating or destroying any particles. What makes it work, as Kaplan explains [51], is a sort of "Hilbert hotel": the states in the Dirac sea extend all the way out to negative infinity in ω , so for every ψ_- state we have populated by moving a formerly negative-energy state to positive energy, there is another to take its place.

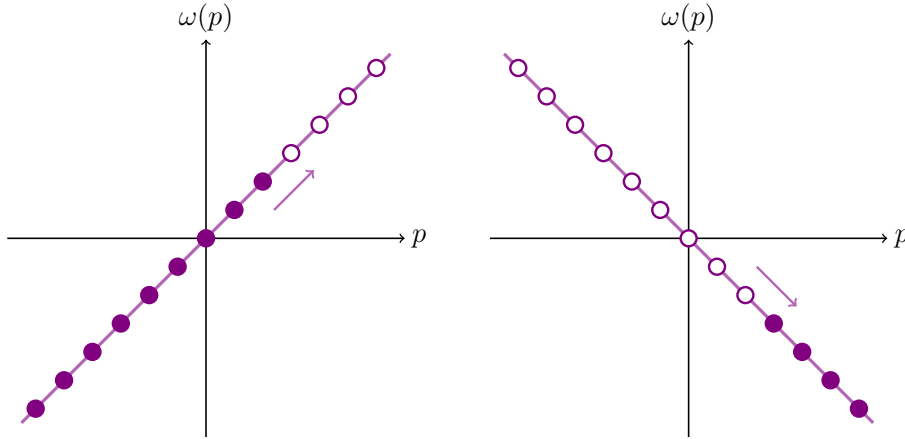


Figure 6: Fermi surface for left- and right-handed fermion modes in the $(1+1)d$ theory [50]. Shaded circles are populated levels and white circles are unpopulated levels. An applied electric field increases the population of left-moving states and decreases the population of right-moving states, as indicated by the arrows.

How many particles have we actually created? To answer this question it is useful to temporarily introduce an infrared cutoff. We imagine that our particles live on a circle of length L , so that the available momentum modes are discretized: $p_k = 2\pi k/L$, for $k \in \mathbb{Z}$. Then the number of new left-moving particles we have introduced is given by

$$n_- = \frac{p_E}{p_1} = \frac{L}{2\pi} \int E(t) dt = \frac{1}{2\pi} \int F, \quad (101)$$

where the final integral is over all of spacetime. We see that this result is actually independent of our infrared cutoff, which we can now dispense with. Eq. (101) corresponds to the local violation

of current conservation (and the analogous equation for the left-movers),

$$dJ_{\pm} = \mp \frac{1}{2\pi} F. \quad (102)$$

This is the anomaly equation in $(1+1)d$.

In the above discussion, we have characterized the gauge field A as a *background* gauge field, an external classical source that drives quantum particle creation. In this interpretation, equation (102) corresponds to an 't Hooft anomaly. This is an obstruction to gauging: we cannot gauge the full $U(1) \times U(1)$ symmetry generated by the currents J_+ and J_- . If we do choose to gauge the diagonal combination $J_+ + J_-$ with a *dynamical* gauge field A , then (102) is an operator equation that tells us that the individual currents J_+ and J_- are simply not conserved in the gauge theory. In that case, we characterize this as an ABJ anomaly.

Note that the integral of the right-hand side of (102) over a closed Euclidean spacetime is always (for every field configuration in the path integral) an integer, by (50). The anomaly equation implies that, when the equation of motion is satisfied, the flux is actually *zero* because it is the integral of a total derivative. Structurally, this is similar to what we said about the electric charge in light of Maxwell's equation (59): one might say that in the $(1+1)d$ theory with chiral fermions, there is now a “Gauss law constraint” imposing that the net electromagnetic flux through a closed Euclidean spacetime is zero. We will have more to say along similar lines when discussing axions and instantons in the next lecture.

9.2.2 Landau levels and the $(3+1)d$ anomaly

Now, continuing to follow Nielsen and Ninomiya, we consider an analogous calculation in $(3+1)d$. The anomaly involves $F \wedge F \propto \vec{E} \cdot \vec{B}$, so we want to turn on parallel electric and magnetic fields to see the effect. We begin by considering fermions in a uniform magnetic field along the z -direction, which we achieve by taking

$$A = Bx \, dy, \quad B > 0. \quad (103)$$

We can solve for the states of a massless charged Weyl fermion in such a background. This is a familiar problem of *Landau levels*.

Specifically, we aim to solve the Dirac equation for a left-handed Weyl fermion ψ of charge q , $\bar{\sigma}^{\mu} D_{\mu} \psi = 0$. If we can solve the auxiliary equation

$$\bar{\sigma}^{\mu} D_{\mu} (\sigma^{\nu} D_{\nu} \phi) = 0, \quad (104)$$

then we can obtain a solution as $\psi = \sigma^{\nu} D_{\nu} \phi$. The auxiliary equation takes a useful form assuming the ansatz $\phi(t, \vec{x}) = e^{-i\omega t} \phi(\vec{x})$ and a time-independent magnetic field, namely,

$$\omega^2 \phi(\vec{x}) = \left[(\vec{p} - q\vec{A})^2 - 2q\vec{B} \cdot \vec{S} \right] \phi(\vec{x}), \quad (105)$$

where $\vec{p} = -i\vec{\nabla}$ and $\vec{S} = \frac{1}{2}\vec{\sigma}$ is the spin operator. This has precisely the same form as the non-relativistic Schrödinger equation for a charged particle in a magnetic field, with the exception that the eigenvalue is ω^2 rather than simply ω . For the choice (103) of gauge field, the operators p_y, p_z ,

and S_z all commute with the Hamiltonian. The solution proceeds in the familiar way (see, e.g., [52]): we can replace p_y by its eigenvalue and then complete the square to obtain harmonic-oscillator solutions.

The ω^2 eigenvalues are then determined by a continuous momentum eigenvalue p_z , an integer n , and an S_z eigenvalue $s_z = \pm 1/2$, with corresponding dispersion relation

$$\omega_n(p_z, s_z)^2 = p_z^2 + (2n + 1)|q|B - 2qBs_z. \quad (106)$$

The corresponding solutions depend on p_y , but the eigenvalues do not (i.e., there is an infinite degeneracy for each one). For $n > 0$, or $n = 0$ and $s_z = -\frac{1}{2} \text{sgn}(q)$, the solutions to this equation form a hyperbola in the (p_z, ω) plane with two branches, one where $\omega > 0$ and one where $\omega < 0$. In the Dirac sea, the $\omega < 0$ branch is completely filled. Turning on an electric field will shift states within such a band, but will not create any new particles. The $\omega > 0$ branch is completely empty, and will remain so when the electric field is turned on. Thus, none of these solutions is relevant for the chiral anomaly.

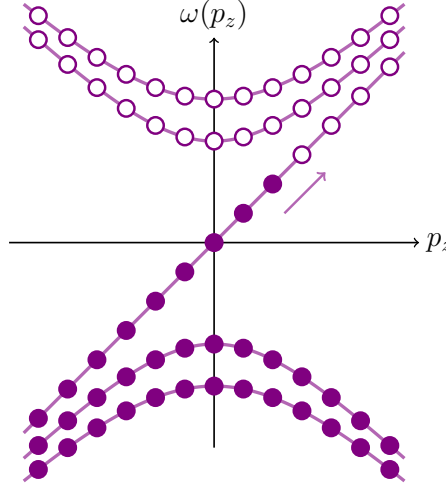


Figure 7: Fermi surface for a $q < 0$ fermion in the $(3+1)d$ theory [50]. The hyperbolic branches of the Landau levels (106) are either fully populated or empty, and the states within these branches simply rearrange under an applied electric field. Only the linear $n = 0$ branch (107) changes in the presence of an applied field, effectively reducing to the $(1+1)d$ problem.

The special case that remains is $n = 0, s_z = +\frac{1}{2} \text{sgn}(q)$, for which the dispersion relation becomes simply $\omega^2 = p_z^2$. In this case, the hyperbola degenerates to the two intersecting lines, $\omega(p_z) = \pm p_z$. However, we should be a bit more careful: these are the solutions to (105) for $\phi(\vec{x})$. We must act on them with $\sigma^\nu D_\nu$ to get ψ . We find that if $q > 0$ and $\omega = p_z$, or if $q < 0$ and $\omega = -p_z$, we obtain $\psi = 0$. Thus, this solution to (105) doesn't correspond to an actual state; instead of two intersecting lines, we have only a single line, as depicted in Fig. 7. Summarizing, the nontrivial solutions of interest are:

$$\begin{aligned} q > 0 : \quad n = 0, \quad s_z = +\frac{1}{2}, \quad \omega = -p_z; \\ q < 0 : \quad n = 0, \quad s_z = -\frac{1}{2}, \quad \omega = +p_z. \end{aligned} \quad (107)$$

Now this looks just like the $(1+1)$ d case! Our $(3+1)$ d fermions $\psi, \bar{\psi}$ in (94) with opposite charges $+1$ and -1 behave like $(1+1)$ d right- and left-moving fermions, respectively, albeit with a huge degeneracy associated with p_y . If we adiabatically turn on an electric field in the z -direction, it will create particles according to (101).

Exercise. Fill in the details of the derivation of (106) and (107). Carefully check what happens in the degenerate case (107): substitute the solutions back into the $(3+1)$ d action and show that the modes behave in the $t-z$ space the way we expect $(1+1)$ d fermions to behave.

The remaining detail to understand is the density of states: how do we relate the $(1+1)$ d particle creation rate to a $(3+1)$ d particle creation rate? For a given set of labels (p_z, n, s_z) for eigenvalues in (106), there is a large degeneracy corresponding to the other, continuous, quantum number p_y . Precisely as in the familiar non-relativistic context, the density of states corresponds to $\frac{1}{2\pi}B$ times the transverse area. As a result, we have

$$\frac{dn_R}{dt} = \frac{1}{4\pi^2} \int d^3x EB \quad (108)$$

which corresponds to the $(3+1)$ d anomaly equation

$$dJ_R = \frac{1}{8\pi^2} F \wedge F, \quad \text{and similarly} \quad dJ_L = -\frac{1}{8\pi^2} F \wedge F. \quad (109)$$

This is the chiral anomaly, from a physical perspective: in the presence of parallel electric and magnetic fields, charged fermions of one chirality are created and those of the opposite chirality are destroyed. When we consider A as a background gauge field, (109) shows that there is an 't Hooft anomaly, and we cannot gauge the full $U(1) \times U(1)$ symmetry generated by J_L and J_R . When A is a dynamical gauge field coupling to the diagonal $U(1)$ symmetry, this becomes an ABJ anomaly, and J_L and J_R are simply not conserved.

Now that we have derived the chiral anomaly in a physically transparent manner, let's quickly summarize the two more common textbook derivations in the next two subsections.

9.3 The triangle anomaly calculation

The three-point function of $U(1)$ currents, computed via a loop of fermions, is the classic calculation that revealed the existence of the chiral anomaly. Because it can be found in so many places, I will not give a detailed treatment here. That shouldn't be taken as an indication that the details are unimportant. This is a calculation that every particle physicist should work through, carefully, at least once. I especially recommend the treatment in terms of Weyl fermions found in [53].

The same triangle diagram calculation sheds light on multiple anomalies. We can compute a three-point function of currents, each of which is bilinear in fermions, from a triangle loop with a current insertion at each vertex. We can also replace some of these external currents with gauge bosons. In Fig. 8 we show such diagrams. We can interpret this as a calculation in which two of the external currents have been gauged, so there are now external gauge boson particles. This calculation reveals the ABJ anomaly, an explicit breaking of the remaining current j_L^μ in the

gauge theory. On the other hand, if we think of the external gauge bosons as merely *background* gauge fields, the calculation reveals a non-conservation of the remaining current in the presence of certain classical backgrounds. This is an 't Hooft anomaly: it is an obstruction to gauging, but the current remains conserved in general spacetime backgrounds as long as we don't turn on nontrivial background gauge fields as well. (You might wonder: what if we consider a case with two external currents, and a single gauge field? The answer is more subtle, involving what is known as a 2-group symmetry [54].)

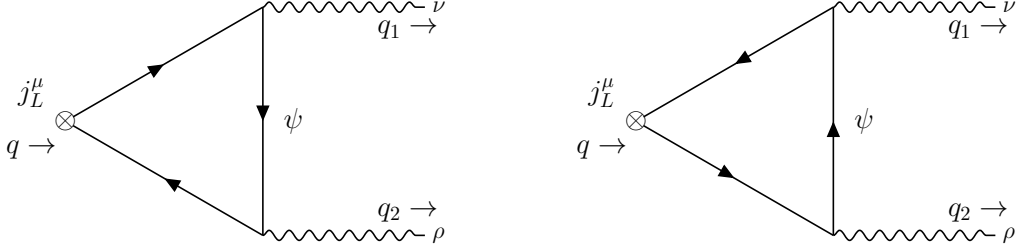


Figure 8: Triangle diagrams for the chiral anomaly: here, the ABJ anomaly between the chiral current $j_L^\mu = \psi^\dagger \bar{\sigma}^\mu \psi$ and electromagnetism, induced by a loop of the chiral fermion ψ . The two diagrams differ by the orientation of the fermion loop in the triangle (or, equivalently, by crossing the two exterior photons).

In evaluating the triangle diagrams in Fig. 8, naively we have

$$\begin{aligned} i\mathcal{M}^{\mu\nu\rho} = & -e^2 \int \frac{d^4k}{(2\pi)^4} \text{tr} \left[(-i\bar{\sigma}^\mu) \frac{i}{\not{k} - q_2} (-i\bar{\sigma}^\rho) \frac{i}{\not{k}} (-i\bar{\sigma}^\nu) \frac{i}{\not{k} + q_1} \right] \\ & - e^2 \int \frac{d^4k}{(2\pi)^4} \text{tr} \left[(-i\bar{\sigma}^\mu) \frac{i}{\not{k} - q_1} (-i\bar{\sigma}^\nu) \frac{i}{\not{k}} (-i\bar{\sigma}^\rho) \frac{i}{\not{k} + q_2} \right]. \end{aligned} \quad (110)$$

We are using notation $1/\not{k} \equiv (k_\mu \sigma^\mu)/k^2$. If current conservation holds, one should be able to contract $\mathcal{M}^{\mu\nu\rho}$ with q_1^ν , q_2^ρ , or $q^\mu = (q_1 + q_2)^\mu$ and find zero, according to the Ward identity. However, one must be careful with this procedure. The integrals in (110) are linearly divergent. Such integrals are notoriously ill-defined. Under a shift $k_\mu \mapsto k_\mu + k_\mu^{(0)}$ in the integration variable, the integral acquires a term proportional to the constant shift vector $k_\mu^{(0)}$. We can carry out such shifts separately in the two terms in (110). These constant shifts, then, must be chosen according to some physical principle to give a well-defined value to $\mathcal{M}^{\mu\nu\rho}$. Making this choice correctly is the sort of fiddly apparent regulator-dependence that usually shows up, in one way or another, in calculations of the chiral anomaly.

Because we have taken the momenta q_1 and q_2 to correspond to external photons, consistency of QED requires that the Ward identity hold for these. One can then show that there is a unique consistent choice in the shifts of the loop momentum variables, which leads to the equation

$$(q_1 + q_2)_\mu \mathcal{M}^{\mu\nu\rho} = -\frac{e^2}{4\pi^2} \epsilon^{\nu\rho\alpha\beta} (q_1)_\alpha (q_2)_\beta. \quad (111)$$

This reveals a non-conservation of the chiral current j_L^μ that, when translated into an operator equation, precisely agrees with (109).

The triangle anomaly calculation has the virtue of being easily adapted to general symmetry currents, not just U(1). If we insert symmetry generators T_R^a , T_R^b , and T_R^c for a fermion in a group

representation \mathbf{R} at the three vertices of the triangle, we find that the corresponding matrix element is proportional to a group-theoretic invariant that is symmetric in a , b , and c :

$$D^{abc}(\mathbf{R}) \equiv \frac{1}{2} \text{tr} \left(\{T_{\mathbf{R}}^a, T_{\mathbf{R}}^b\} T_{\mathbf{R}}^c \right). \quad (112)$$

The only simple Lie groups for which there are representations \mathbf{R} with nonzero $D^{abc}(\mathbf{R})$ are the $\text{SU}(N)$ groups with $N > 2$. Furthermore, it turns out that independent of a , b , and c , there is a proportionality $D^{abc}(\mathbf{R}) = A(\mathbf{R}) D^{abc}(\square)$, where \square denotes the fundamental representation of $\text{SU}(N)$. The relative factor $A(\mathbf{R})$ is known as the *anomaly coefficient* of the representation \mathbf{R} . Consistency of gauge theory then requires that all $\text{SU}(N)^3$ anomalies, $\text{U}(1)^3$ anomalies, and $\text{U}(1)\text{SU}(N)^2$ anomalies vanish. In fact, there is a further related criterion arising from a possible *gravitational* anomaly of a $\text{U}(1)$ current, in which dJ can be proportional to $\text{tr}(R \wedge R)$ where R is the Riemann curvature 2-form.

9.4 Fujikawa's path integral calculation

Another way to find the anomaly is to consider the path integral measure. A path integral for the theory (96) has the form

$$\int \mathcal{D}A \mathcal{D}\bar{\psi} \mathcal{D}\psi e^{iS[A, \psi, \bar{\psi}]}. \quad (113)$$

A classical symmetry operation on the fields, by definition, leaves the last $\exp(iS)$ factor invariant. Thus, the only way that a symmetry of the classical theory can fail to be a symmetry of the quantum theory is if it does not leave the *measure* invariant. This is what Fujikawa calculated, by carefully regulating the path integral measure [55, 56].

We will only very briefly sketch the idea. We expand the field $\psi(x)$ in a set of orthonormal modes, $\psi(x) = \sum_m a_m \phi_m(x)$, where the $\phi_m(x)$ which are eigenstates of the Dirac operator, $i\not{D}\phi_m = \lambda_m \phi_m$. Then $\mathcal{D}\psi$ is $\prod_m da_m$. The hard work is entirely in understanding how to properly regulate this product, and I encourage you to seek out either the original literature or textbook accounts to understand the details. It turns out (carefully regulating) that $\prod_m da_m$ has a nontrivial Jacobian under the field redefinition $\psi(x) \mapsto e^{i\alpha(x)}\psi(x)$. The end result is that under this field redefinition, the path integral measure changes according to

$$\mathcal{D}\psi \mapsto \mathcal{D}\psi e^{i \int \frac{1}{8\pi^2} \alpha(x) F(x) \wedge F(x)}. \quad (114)$$

Effectively, this has added a new term to the action, where $\alpha(x)$ couples to $F \wedge F$.

I have stated this argument for the case where we have a dynamical gauge field in the theory, where it reproduces the ABJ anomaly. But we could also do the calculation without integrating $\mathcal{D}A$, but keeping A in the action as a fixed classical background field sourcing the fermion currents. The final conclusion is the same: the path integral measure changes in a way that is sensitive to the background field. This is the hallmark of an 't Hooft anomaly.

9.5 The takeaway message

The chiral anomaly will play a *central* role in our explanation of axion physics, so let me emphasize the one major result that we will need for both $\text{U}(1)$ and $\text{SU}(N)$ gauge groups.

A field redefinition on a Weyl fermion ψ ,

$$\psi(x) \mapsto e^{i\alpha(x)}\psi(x), \quad (115)$$

not only produces the obvious changes from plugging this into the path integral (e.g., rephasing mass or Yukawa terms, generating a derivative coupling $\psi^\dagger \bar{\sigma}^\mu \psi \partial_\mu \alpha$ from the kinetic term), but *also*, via the chiral anomaly, adds a term to the action of the form

$$\int 2I(\mathbf{R}_\psi) \frac{1}{8\pi^2} \alpha(x) \text{tr}(F \wedge F) \quad (116)$$

for every $SU(N)$ gauge group ψ is charged under, where \mathbf{R}_ψ is the representation of ψ under the group and $I(\mathbf{R}_\psi)$ is the Dynkin index of that representation (e.g., $1/2$ for the fundamental representation of $SU(N)$), as well as a term

$$\int q^2 \frac{1}{8\pi^2} \alpha(x) F \wedge F \quad (117)$$

for every $U(1)$ group ψ is charged under, with q the charge of ψ under that group.

9.6 One-loop exactness

An important fact about the anomaly is that it is *one-loop exact*: the operator equation (109) for current non-conservation receives no corrections beyond one loop. One way to see this is to consider the case $\alpha = 2\pi$ in (116) or (117). Clearly this is a trivial rephasing, doing nothing to the fermion, but it seems that it adds a nontrivial contribution to the action. How is this possible? It makes sense only if e^{iS} is unchanged by this addition to S . This follows from the quantization conditions (89) and (76) that we have discussed previously. If we allowed for a more general function of the gauge coupling modifying the relationship between the left- and right-hand sides of (109), it would no longer be true that a 2π rephasing of the fermion would leave e^{iS} invariant. This tells us that there are no perturbative corrections to (109) involving a series in the gauge couplings.

A related fact is that instantons contribute to amplitudes that violate chiral symmetries, converting some fermion modes into other fermion modes. In particular, fermions have zero modes in an instanton background, which lead to an effective interaction known as the 't Hooft vertex, which involves one mode for every chiral fermion. This is a direct consequence of (109): an instanton has a nonzero integral of $\frac{1}{8\pi^2} F \wedge F$, which in turn corresponds to nonzero integrals of dJ_R and dJ_L , i.e., to a net change in the number of right- and left-handed fermions. The 't Hooft vertex is simply the effective operator that encapsulates the associated multi-fermion interaction. Again, this makes it clear that the instanton effect is one-loop exact: if we had a generic coefficient between the left and right sides of (109), an instanton would correspond to a *fractional* change in the number of fermion modes, and no sensible interpretation in terms of a chirality-violating process could be associated with it.

9.7 Further remarks on anomalies and QCD

The one-loop exactness of chiral anomalies is a consequence of their topological nature. This makes anomalies an extremely powerful tool for analyzing nonperturbative properties of quantum field theory. An example is 't Hooft anomaly matching: given a set of symmetry currents in a

theory that have an 't Hooft anomaly, this anomaly will be preserved under RG flows that do not explicitly break any of the symmetries [49]. In QCD (with massless quarks), the nonabelian flavor symmetry currents for $SU(N_f)_L \times SU(N_f)_R$ have 't Hooft anomalies. This implies that confinement cannot fully gap the spectrum: there should be massless composite states in the infrared that match these anomalies. These could potentially involve new composite chiral fermions, or they can involve massless Nambu-Goldstone bosons (the pions) with Wess-Zumino-Witten interactions that match the 't Hooft anomalies from the UV theory. With some additional assumptions, one can argue that only the latter possibility is consistent, so the theory must spontaneously break $SU(N_f)_L \times SU(N_f)_R \rightarrow SU(N_f)_V$ [57, 58]. This is consistent with experiment: the pions, kaons, and η meson appear to have the properties of pseudo-Nambu-Goldstone bosons parametrizing the coset $[SU(3)_L \times SU(3)_R]/SU(3)_V$.

Another important application of the chiral anomaly in QCD is the resolution of the U(1) puzzle. Naively, one would expect an additional light pseudo-Nambu-Goldstone boson for the broken $U(1)_A$ symmetry, but this is not what we see: the η' is the candidate, but it is much heavier than one would expect if it were a pseudo-Nambu-Goldstone boson. This is because the $U(1)_A$ symmetry has an ABJ anomaly with the strong interactions themselves, so strong dynamics badly breaks the would-be symmetry and produces a large mass for the η' .

10. Axion couplings to gauge fields

The next part of these lecture notes will be all about the phenomenology of fields known as *axions*. One important piece of axion physics follows from our discussions of the quantization laws (76) and (89), so we explain it here. For now, the term “axion” refers to a compact scalar field $\theta(x)$. By “compact” we mean that the field values live on a circle, which is to say that the value of $\theta(x)$ is only defined modulo 2π . This means that $\theta(x) \mapsto \theta(x) + 2\pi$ is a *gauge redundancy* of our theory.

This gauge redundancy places strong limitations on the Lagrangian for our theory. We can write arbitrary interactions involving $\partial_\mu \theta$, because this is gauge-invariant. We can also write interactions involving 2π -periodic functions like $\cos(k\theta)$ (with $k \in \mathbb{Z}$), which can give rise to a mass for the axion. (As we will see in the next part, axion model-building primarily revolves around finding ways to explain why such terms can naturally have tiny coefficients.) What we cannot do is write terms that involve θ without a derivative acting on it and without packing it in a periodic function. However, there is one remarkable exception, which is the key to all of axion physics. This is a term of the form $\theta F \wedge F$ (or $\theta \text{tr}(F \wedge F)$, for nonabelian gauge fields).

Let F be a $U(1)$ gauge field strength, and consider the Euclidean action

$$S_E = i c \int \theta F \wedge F. \quad (118)$$

Here c is a real coefficient; the factor of i in front arises from Wick rotation from Minkowski signature to Euclidean signature, because $F \wedge F$ has a single time-index in it. (Such factors of i generally accompany terms that, in tensor index notation, are written with an odd number of Levi-Civita tensors.) This Euclidean action is manifestly not gauge-invariant. Under the operation

$\theta \mapsto \theta + 2\pi$, we have

$$S_E \mapsto S'_E = S_E + 2\pi i c \int F \wedge F. \quad (119)$$

In fact, this is not a fatal problem for the theory. In order for calculations of physical quantities to be gauge invariant, it is not necessary for S_E to be gauge invariant. The requirement is that the path integral measure be gauge invariant, and hence that $\exp[-S_E] = \exp[-S'_E]$. In other words, we require

$$\exp \left[-2\pi i c \int F \wedge F \right] = 1. \quad (120)$$

Equivalently, for any F , we must have $c \int F \wedge F \in \mathbb{Z}$. In \mathbb{R} gauge theory, we have $\int F \wedge F = 0$ and this is automatically satisfied for any value of c . In $U(1)$ gauge theory, however, $\int F \wedge F$ can take on nonzero values, as in (76). This immediately implies that c is quantized. Considering the smallest possible value of $\int F \wedge F$, the condition (120) becomes

$$\exp \left[2\pi i \cdot c \cdot 8\pi^2 \right] = 1 \quad \Rightarrow \quad c = \frac{k}{8\pi^2}, \quad k \in \mathbb{Z}. \quad (121)$$

In other words, we have learned that consistently defining the theory in the presence of general $U(1)$ gauge bundles requires that the axion coupling to $U(1)$ gauge fields be an integer multiple of $\frac{1}{8\pi^2}$.

The story is precisely the same for $SU(N)$ gauge fields, due to the quantization of instanton number (89). Reverting to Minkowski signature, we can summarize by saying that the coupling of an axion θ to a $U(1)$ gauge field strength F and an $SU(N)$ gauge field strength G is

$$S = \frac{k_F}{8\pi^2} \int \theta F \wedge F + \frac{k_G}{8\pi^2} \int \theta \text{tr}(G \wedge G), \quad k_F, k_G \in \mathbb{Z}. \quad (122)$$

Such a coupling of an axion to $SU(3)$ gauge fields is the crucial element in the axion solution to the Strong CP problem, discussed in detail in the next part of these lecture notes.

This argument relies on one assumption that has gone unstated so far: in (119) we took the gauge field strength F to be invariant under the gauge transformation $\theta \mapsto \theta + 2\pi$. This is not a completely innocuous assumption. For example, in free $U(1)$ gauge theory, there is an $SL(2, \mathbb{Z})$ duality group whose action mixes up F and \tilde{F} . As a consequence of the Witten effect (62), the magnetic gauge field strength shifts when $\theta \mapsto \theta + 2\pi$. One could imagine a theory in which the shift of θ is accompanied by such a duality operation that changes the *electric* field strength, so that $F \wedge F$ itself would not be invariant, and our argument above would not apply. It turns out that such a loophole does not apply to real-world electromagnetism, because it would require that the electron obtain a magnetic monopole charge in the presence of an axion background, which would be a phenomenological disaster.⁵

One last consequence of the Witten effect that I would like to mention here is that loops of magnetic monopoles can generate a mass for the axion. In order for a magnetic monopole to acquire an electric charge in an axion background, as in (62), the monopole must have a dyon collective coordinate. (Such a mode was originally found for the 't Hooft-Polyakov monopole [59], but it must exist for any monopole in a theory with an axion.) This is a quantum mechanical degree of

⁵I thank Ben Heidenreich and Jake McNamara for clarifying discussions on this topic, more details of which will be published elsewhere.

freedom living on the monopole worldline that takes the form of a particle on a circle, which has a quantized momentum. Giving momentum to this degree of freedom is equivalent to giving the magnetic monopole an electric charge. This compact degree of freedom can wind around a loop of monopole worldline. This can be viewed as a $U(1)$ instanton, closely analogous to the $SU(N)$ instantons we have discussed. One can integrate out monopole worldlines in the path integral to see that such a $U(1)$ instanton generates a periodic potential for the axion [60]. (Related physics is also reviewed in [61, 62].)

It is also possible to have additional axion couplings to gauge fields via higher-dimension operators that are explicitly gauge invariant, starting with $(\Box\theta) F \wedge F$. However, via equations of motion, this is suppressed by the mass of the axion squared, and in most cases will be extremely subdominant. An important exception is the QCD axion's coupling to photons via mixing with the pion, which is large because the axion and pion mass arise through the same strong dynamics. We will discuss this further below, in §12.3.1.

Let us sum this up, now quoting the result in canonical normalization. To canonically normalize, we assume the axion kinetic term has the form $\frac{1}{2}f^2 d\theta \wedge \star d\theta$, where the numerical factor f is known as the “axion decay constant.” The gauge field is canonically normalized with a factor of the gauge coupling. This implies:

The leading coupling of a light axion to gauge fields is *quantized* in integer multiples of a loop factor divided by the axion decay constant, $g^2/(8\pi^2 f)$, in canonical normalization. Corrections are necessarily suppressed by the square of the axion mass.

11. Chern-Simons terms

The couplings of a compact scalar to gauge fields that we have just described are an example of a larger class of interactions known as *Chern-Simons terms*. The characteristic feature of a Chern-Simons term in an action is that it is not gauge-invariant, but that $\exp[iS]$ is gauge-invariant, and so the path integral is well-defined. Chern-Simons terms are topological and have a quantized coefficient. Here we discuss some other examples. This section lies somewhat outside the main line of reasoning of these lectures and can be skipped by readers eager to get to axion physics, though we will refer back to some parts of it in later sections.

11.1 Chern-Simons terms in 3d gauge theory

Perhaps the most familiar and well-studied examples of Chern-Simons terms are those appearing in 3d gauge theories. A Chern-Simons term for a $U(1)$ gauge theory takes the form

$$S_{\text{CS}}^{U(1)} = \frac{k}{4\pi} \int_M A \wedge dA, \quad (123)$$

and for a non-abelian gauge theory takes the form

$$S_{\text{CS}}^{SU(N)} = \frac{k}{4\pi} \int_M \text{tr} \left(A \wedge dA - i \frac{2}{3} A \wedge A \wedge A \right). \quad (124)$$

If we view the integrands of these 3d Chern-Simons terms from the 4d viewpoint, their derivatives are the familiar instanton number densities that we have discussed above. In the $U(1)$ case we obviously have $d(A \wedge dA) = dA \wedge dA = F \wedge F$. In the nonabelian case, we have

$$\begin{aligned} d \left[\text{tr} \left(A \wedge dA - \frac{2}{3} i A \wedge A \wedge A \right) \right] &= \text{tr} (dA \wedge dA - 2i dA \wedge A \wedge A) \\ &= \text{tr} (dA \wedge dA - 2i dA \wedge A \wedge A - A \wedge A \wedge A \wedge A) \\ &= \text{tr} (F \wedge F), \end{aligned} \quad (125)$$

where we have made repeated use of the graded cyclic property of the trace reviewed in §7 and the last step uses the formula $F = dA - iA \wedge A$.

The Chern-Simons term has a quantized coefficient, which is a consequence of its lack of gauge invariance. At first glance, you might think that under a gauge transformation (in the $U(1)$ case)

$$A \wedge dA \mapsto (A + d\alpha) \wedge dA = A \wedge dA + d(\alpha dA), \quad (126)$$

so the integrand in the action shifts by a total derivative and the action is invariant. As usual, the subtleties are all about topology: what if the gauge transformation α winds around a circle, or if A is not single-valued? Giving a proper treatment of this subject is beyond the scope of these lecture notes, but we can give a partial argument based on the connection to $F \wedge F$ terms in one higher dimension. Specifically, suppose that we want to make sense of (123) on a 3d manifold M , despite its lack of gauge invariance. What we really need, to define a path integral, is not the action itself but $\exp(iS)$. Suppose that M is the boundary of a 4d manifold X and that we can extend our gauge field configuration over X .⁶ Then we can replace our apparently ill-defined formula (123) with a manifestly well-defined, gauge-invariant formula in 4d:

$$\exp(iS_{\text{CS}}^{U(1)}) = \exp\left(i \frac{k}{4\pi} \int_X F \wedge F\right). \quad (127)$$

This is motivated by Stokes's theorem, which suggests $\int_X F \wedge F = \int_X d(A \wedge F) = \int_M A \wedge F$, even though this isn't as simple as it looks since A may not be single-valued. In any case, if we try to define the path integral directly via (127) instead of (123), we can ask whether we would get the same answer for a different choice of 4-manifold X' that *also* has boundary M . Now, because X and X' have the same boundary, we can glue X to the orientation-reversal $-X'$ to make a *closed* 4-manifold. (If this argument is sounding familiar, it's essentially the flux-quantization argument that we gave in §4.3, run backwards. In fact, a Wilson loop can be thought of as a 1d Chern-Simons term.) We know that the integral of $F \wedge F$ over any closed 4-manifold is $8\pi^2 n$ for $n \in \mathbb{Z}$. Thus we have

$$\exp\left(i \frac{k}{4\pi} \int_{X \cup (-X')} F \wedge F\right) = \exp(2\pi i k n), \quad (128)$$

⁶This is not an innocuous assumption; a d -dimensional manifold that can be written as the boundary of a $(d+1)$ -dimensional manifold is said to be “null-bordant.” It is a nontrivial fact that all closed 3-manifolds are null-bordant, but this is special to 3 dimensions. Even when the spacetime is null-bordant, it may not be possible to extend the gauge bundle over the higher-dimensional manifold. Discussing this properly would require a significant enlargement of these lectures. See [63].

which is always 1 if

$$k \in \mathbb{Z}. \quad (129)$$

Precisely the same argument goes through for the case of $SU(N)$ and the non-abelian Chern-Simons term (124). Thus, we conclude (up to all the mathematical subtleties I've swept under the rug along the way!) that the coefficient of 3d Chern-Simons terms is quantized in units of $\frac{1}{4\pi}$ in order for us to obtain a well-defined path integral. You can consider winding gauge transformations on simple spacetimes like $S^1 \times S^2$ to gain more confidence in this conclusion.

This argument follows very similar logic to the argument that led us to conclude that axion couplings to gauge fields are quantized: in both cases, we have an action that is not gauge invariant, but $\exp(iS)$ is gauge invariant for specific choices of coupling.

Chern-Simons terms in 3d have important physical implications. For example, they lead to an effective photon mass, as you can derive in the following exercise.

Exercise. Consider a 3d $U(1)$ gauge theory with both a Maxwell kinetic term and a Chern-Simons term,

$$S = \int \left(-\frac{1}{2e^2} F \wedge \star F + \frac{k}{4\pi} A \wedge dA \right). \quad (130)$$

Derive the equation of motion for the gauge field, and show that it implies that the field strength obeys a massive wave equation,

$$(\square + m_A^2) F^{\mu\nu} = 0, \quad (131)$$

for some constant m_A . What is m_A in terms of the parameters e and k ? How many independent, propagating polarization states does a massless gauge field (with $k = 0$) have in $2 + 1$ dimensions? What about the gauge field with a Chern-Simons mass, $k \neq 0$? Explain your answers, and comment on how this differs from the familiar $(3 + 1)$ -dimensional case.

Chern-Simons terms can also be related to θ terms in one lower dimension, as you can see by example in the following exercise.

Exercise. Now suppose that we compactify the theory from the previous exercise on a circle $x^3 \cong x^3 + 2\pi L$, and consider an ansatz where we turn on a constant $A_3 \neq 0$ so that

$$\varphi \equiv \int_0^{2\pi L} dx^3 A_3 \neq 0. \quad (132)$$

Show that in the dimensionally reduced $(1 + 1)$ -dimensional theory, the gauge field has a 2d theta term,

$$\frac{\theta}{2\pi} \int F. \quad (133)$$

How is θ related to φ ? Explain how the periodicity of the 2d coefficient θ is compatible with the 3d origin of the term.

Our discussion of how the Chern-Simons term can be well-defined crucially assumed that the 3-manifold M had no boundary. On a 3-manifold with boundary, even topologically trivial

gauge transformations as in (126) do not give rise to a gauge-invariant action, as we acquire a term $\int_{\partial M} \frac{k}{4\pi} \alpha dA$ when carrying out the gauge transformation $A \mapsto A + d\alpha$. However, this can be compensated if there is an *anomalous* 2d gauge theory living on the boundary ∂M . This is a common property of Chern-Simons theories, sometimes referred to as *anomaly inflow* [64]: consistency of the theory on spaces containing boundaries or other defects requires the existence of localized, charged degrees of freedom on the defect. A familiar example arises in condensed matter physics, where quantum Hall systems are described by $(2+1)$ d Chern-Simons effective theories in the bulk and admit charged *edge modes* that provide an anomalous boundary theory that cancels the non-gauge-invariant terms from the bulk.

11.2 Chern-Simons masses and Stueckelberg masses

In 4d gauge theory, we can't add a photon mass with a term of the form $A \wedge dA$ as in 3d. However, there is a very similar type of photon mass term, sometimes called a “BF term.” In this case, we add a new 2-form $U(1)$ gauge field B to the theory. That is, there is an antisymmetric tensor field $B_{\mu\nu}$, with $B = \frac{1}{2} B_{\mu\nu} dx^\mu \wedge dx^\nu$ having a gauge invariance $B \mapsto B + d\lambda$, where λ is a 1-form. This is a $U(1)$ gauge symmetry in the sense that it obeys quantization laws similar to those for ordinary $U(1)$ gauge fields; in particular, the field strength $H = dB$ has quantized flux when integrated over any closed 3-manifold Ω ,

$$\frac{1}{2\pi} \int_{\Omega} H \in \mathbb{Z}, \quad (134)$$

directly parallel to the usual magnetic flux quantization (50). Similarly, B is invariant under large or “winding” gauge transformations, in the sense that if $[\omega] \in H^2(M, \mathbb{Z})$ is a class in integral cohomology, there is an extended gauge invariance under $B \mapsto B + \omega$ even though ω can only *locally* be written as $d\lambda$. Such higher-form gauge fields are ubiquitous in string theory and quantum field theories in more than four spacetime dimensions, but they can also be useful in 4d as well. In particular, the following action describes a massive photon field:

$$S = \int \left(-\frac{1}{2e^2} F \wedge \star F - \frac{1}{2g^2} H \wedge \star H + \frac{k}{2\pi} B \wedge F \right). \quad (135)$$

Here g has dimensions of mass and can be thought of as the coupling constant of the B field, and the last term is a Chern-Simons term that is only well-defined when $k \in \mathbb{Z}$. You can see that the photon is massive in essentially the same way that you derived (131) in the exercise above; in this case, its mass is proportional to k , e , and g .

There is a different way to formulate the theory of a photon with a BF term mass as a theory with a Stueckelberg mass. This reflects a more general way of recasting Chern-Simons terms as Stueckelberg terms. This is very useful physics to familiarize yourself with, although it will play only a small role in the remainder of these lectures. We begin with the idea of a *Hodge dual* to a gauge field, which is a sort of generalization of electric-magnetic duality. A free p -form $U(1)$ gauge field C_p with action

$$\int \left(-\frac{1}{2e_p^2} dC_p \wedge \star dC_p \right) \quad (136)$$

can be dualized to a $(d - p - 2)$ -form gauge field \tilde{C}_{d-p-2} , via the map

$$\frac{1}{2\pi} d\tilde{C}_{d-p-2} = \frac{1}{e_p^2} \star dC_p. \quad (137)$$

The dual theory has action

$$\int \left(-\frac{1}{2\tilde{e}_p^2} d\tilde{C}_{d-p-2} \wedge \star d\tilde{C}_{d-p-2} \right) \quad (138)$$

with the gauge coupling $\tilde{e}_{d-p-2} = 2\pi/e_p$, as in the usual Dirac quantization formula relating electric and magnetic couplings. One can also derive that if C_p is a $U(1)$ gauge field (with the associated flux quantization condition on its field strength), then so is \tilde{C}_{d-p-2} , with the duality trading a magnetic flux quantization condition like (50) for an electric flux quantization condition like (61). This can be derived explicitly at the level of the path integral by introducing auxiliary Lagrange multiplier fields and then integrating out the original gauge field; see, e.g., [65, 66]. As an example, in 4d, the Hodge dual of a 2-form gauge field B is a 0-form field θ with quantized fluxes $\int_C d\theta = 2\pi n$ ($n \in \mathbb{Z}$) around closed curves. This is just a periodic scalar field, $\theta \cong \theta + 2\pi$.

In the presence of a Chern-Simons term as in (135), the Hodge dualization procedure is more subtle. One would like to write $\frac{1}{2\pi} d\theta = \frac{1}{g^2} \star H$. However, (135) implies the equation of motion

$$d\left(\frac{1}{g^2} \star H\right) = \frac{k}{2\pi} F. \quad (139)$$

Because $\star H$ isn't closed, we can't locally write it as d of some quantity θ . Instead, what we can do is rewrite the above equation (locally) as

$$d\left(\frac{1}{g^2} \star H - \frac{k}{2\pi} A\right) = 0. \quad (140)$$

We can thus identify the quantity in parentheses with $\frac{1}{2\pi} d\theta$, or in other words, we have

$$\frac{1}{2\pi} (d\theta + kA) = \frac{1}{g^2} \star H, \quad (141)$$

with a compact scalar $\theta \cong \theta + 2\pi$. The twist is that only the combination $d\theta + kA$ is gauge invariant. This means that θ must be a Stueckelberg field, which *shifts* under a gauge transformation of A :

$$A \mapsto A + d\alpha, \quad \theta \mapsto \theta - k\alpha. \quad (142)$$

The compact scalar field θ , dual to B , is eaten to provide a mass for A . The action dual to (135) is the standard Stueckelberg action:

$$S = \int \left(-\frac{1}{2e^2} F \wedge \star F - \frac{1}{2} f^2 (d\theta + kA) \wedge \star (d\theta + kA) \right), \quad (143)$$

with $f = g/(2\pi)$. This formulation of a massive gauge field is likely more familiar to you than the BF term, but they are completely equivalent. Notice that it no longer has a Chern-Simons term! The ability to trade a Chern-Simons term in one formulation of a theory for a Stueckelberg term in a dual formulation, in which one gauge field shifts under the gauge transformation of a *different*

gauge field, is quite general. In fact, there is another such action lurking in this example as well. We could have started with (135) and dualized A to a magnetic 1-form gauge field \tilde{A} , obtaining a theory in which the B field has a standard kinetic term but $\tilde{A} \mapsto \tilde{A} + k\lambda$ when $B \mapsto B + d\lambda$, where the mass term takes the form $|d\tilde{A} - kB|^2$, which we could think of as the B field eating the magnetic photon. (For a more extended review of this theory including its global symmetry properties and the possible addition of objects carrying various electric and magnetic charges, see [61].)

11.2.1 A magnetic photon mass

The possibility of a mass for the photon is often studied in the real world, and is highly constrained experimentally. We will discuss the implications of quantum gravity for this possibility below, in §25.1. Recently attention was drawn to the possibility that the photon could have a *magnetic* mass [67]. One way to formulate this theory is in terms of a magnetic dual photon \tilde{A} , with a Stueckelberg mass term of the form $|d\theta + k\tilde{A}|^2$. Attempts to treat both standard electric photon couplings to A and terms involving the magnetic dual \tilde{A} within the same action are cumbersome.

The discussion above suggests an equivalent formulation, which avoids the need to refer to the magnetic gauge field \tilde{A} .⁷ Instead, we introduce a 2-form gauge field B with field strength $H = dB$, with an action of the form

$$S = \int \left(-\frac{1}{2g^2} H \wedge \star H - \frac{1}{2e^2} (dA - kB) \wedge \star (dA - kB) \right). \quad (144)$$

This is a dual description of a photon with magnetic mass kg/e : the usual gauge field A is eaten by B . It has the crucial feature that there are two nontrivial gauge transformations acting on the gauge field A :

$$A \mapsto A + d\alpha, \quad \text{and} \quad B \mapsto B + d\lambda, \quad A \mapsto A + k\lambda. \quad (145)$$

The novel gauge invariance, with A shifting under a B gauge transformation, spoils our ability to couple particles electrically to A in the standard way. However, this is exactly what we expect a magnetic mass to do! A magnetic mass confines electrically charged particles, meaning that they come with strings attached: we can write a gauge invariant coupling of A to a particle worldline C only if C is the boundary of a string worldsheet Σ , with

$$S = q \left(\int_C A - \int_\Sigma kB \right). \quad (146)$$

These strings will generally have a tension, which we might expect to be of order g^2 but which can only be determined within a UV completion of the theory.

This formulation is useful for highlighting what the invariant physical challenge of modeling a magnetic photon mass is. The formulation in [67] makes it appear that the challenge is in simultaneously keeping track of electric and magnetic vector potentials describing the same underlying photon field. Instead, we see that the real challenge is to study a theory of dynamical strings coupled to all of the electrically charged particles we know. There's an important general principle here that is mostly orthogonal to the topics of these lectures, but let's highlight it anyway:

⁷I thank Ben Heidenreich for a discussion on this topic.

Higgsing an electric gauge theory confines the dual magnetic charges, and vice versa.

11.3 Higher Chern-Simons terms

So far we have seen a few different examples of Chern-Simons terms: $A \wedge F$ in 3d, $B \wedge F$ in 4d, and $\theta F \wedge F$ in 4d. The general pattern is that these terms are not gauge invariant: they involve several gauge field strengths and a single additional gauge field with no derivative acting on it. The gauge fields are potentially higher degree p -forms (like the 2-form B in 4d) or even a 0-form gauge field, i.e., a periodic scalar θ . This pattern continues in higher dimensions. For example, in a 5d U(1) gauge theory we might have a term $\frac{k}{4\pi^2} \int A \wedge dA \wedge dA$, or if we have a U(1) gauge field A and an SU(N) gauge field with field strength G , we can have a 5d Chern-Simons term $\frac{k}{8\pi^2} \int A \wedge \text{tr}(G \wedge G)$. Such Chern-Simons terms always have a quantized coefficient $k \in \mathbb{Z}$ (when normalized appropriately). It is also true in general that if we lift a Chern-Simons term to one higher dimension and take d of it, we obtain a *theta* term in the higher-dimensional theory, which has a periodic coefficient θ . We have seen this with terms of the form $F \wedge F$ or $\text{tr}(F \wedge F)$, but other examples include $\frac{1}{2\pi} \int F$ in 2d U(1) gauge theory, $F \wedge F \wedge F$ terms in 6d, or even $d\sigma \wedge F$ terms in 3d where $\sigma \cong \sigma + 2\pi$ is a compact scalar.

Along the lines that we saw in §11.2, general Chern-Simons terms involving a U(1) p -form gauge field can have a dual formulation as Stueckelberg terms. This is due to non-conservation of electric flux in their presence, much as we saw in (139) for the case of a BF term. To give another example, in a 4d theory with a $\frac{1}{8\pi^2} \theta F \wedge F$ term, the compact scalar θ can be dualized to a 2-form gauge field B . However, our original theory has an equation of motion

$$f^2 d\star d\theta = -\frac{1}{8\pi^2} F \wedge F, \quad (147)$$

which in the dual picture is a “modified Bianchi identity” requiring that we introduce the dual 2-form via

$$\frac{1}{2\pi} dB = f^2 \star d\theta + \frac{1}{8\pi^2} A \wedge F. \quad (148)$$

The kinetic term for B in the dual formulation then takes the form

$$\int \left[-\frac{1}{2g^2} \left(dB - \frac{1}{4\pi} A \wedge F \right) \wedge \star \left(dB - \frac{1}{4\pi} A \wedge F \right) \right], \quad (149)$$

which is gauge invariant because B shifts as $B \mapsto B + \frac{1}{4\pi} \alpha F$ under the gauge transformation $A \mapsto A + d\alpha$. In much the same way that an ordinary Chern-Simons term in 3d requires edge modes on boundaries, this structure in the 4d theory requires chiral charged modes to exist on “axion strings,” the objects that are charged under B [64]. This is another example of the general concept of anomaly inflow, which we will discuss in more detail below.

Part Three: The Strong CP Problem and Axion Models

12. The Strong CP problem and the axion solution

12.1 The problem and proposed solutions

The Standard Model can have a θ term for the $SU(3)_C$ gluon field, whose field strength I will denote G to distinguish it from the generic field strengths F that I have been referring to all along. This term takes the form

$$\frac{\theta}{8\pi^2} \int \text{tr}(G \wedge G). \quad (150)$$

This is a CP-violating term. When quarks and gluons confine into hadrons at low energies, this term potentially has a variety of effects on hadron physics. Physical implications include a θ -dependent vacuum energy, which is minimized at $\theta = 0$ (this is the Vafa–Witten theorem [68]); CP-violating pion-nucleon couplings, like $\pi^0 \bar{N} N$ or $(\bar{N} \tau^i N) \pi^i$ where τ^i is an $SU(2)$ (global) isospin generator; and a CP-violating neutron electric dipole moment,

$$-i \frac{d_n}{2} \bar{N} \sigma_{\mu\nu} \gamma^5 N F^{\mu\nu}. \quad (151)$$

This last effect has proven to be the easiest to constrain experimentally. The current bound is [69]

$$|d_n| \leq 1.8 \times 10^{-26} e \text{ cm} \quad (90\% \text{ C.L.}). \quad (152)$$

We expect, just from dimensional analysis, a neutron EDM of order θ times the size of the neutron. At a cartoon level, this is simply because the neutron is made up of three valence quarks, two down quarks of charge $-1/3$ and one up quark of charge $2/3$, and a generic such configuration has an electric dipole moment, as sketched in Fig. 9. To have no EDM at all, for this cartoon neutron, would require an unlikely configuration with the two down quarks precisely lined up on opposite sides of the up quark. (For a more extended discussion, including an interesting analogy to the CO_2 molecule, see [11]; another analogy involving a pool table, elaborate almost to the point of absurdity, can be found in [70].) A slightly more detailed estimate leads us to expect that

$$d_n \sim 10^{-16} \theta e \text{ cm}. \quad (153)$$

Comparing expectations with data, we learn that

$$|\theta| \lesssim 10^{-10}. \quad (154)$$

Why is this number so small? The only symmetry that it violates is CP, so one explanation could be that our universe is CP-symmetric. However, we know that to be false. The CKM matrix has an order-one CP-violating phase. So we would like to have a better explanation. The puzzle of small θ is known as the Strong CP Problem.

There are essentially three leading ideas that have been proposed to explain the Strong CP Problem. Namely,

- A massless up quark. If the up quark has no mass, we can rephase one chirality of the up quark, e.g., $u_L(x) \mapsto e^{-i\theta} u_L(x)$, to remove the θ term from the theory (due to the chiral

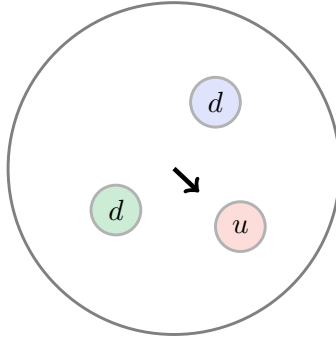


Figure 9: A cartoon of a neutron, containing two down quarks and an up quark at generic positions in the interior. This configuration has a sizable electric dipole moment, as indicated by the arrow.

anomaly; see §9.5). The reason we focus on the *up* quark, rather than any of the others, is simply that empirically we know it is the lightest quark, so the closest to being massless. However, despite occasional attempts to resuscitate it, expert consensus is that this scenario has been ruled out by a combination of precision data and theory. (See, e.g., [71–74].)

- **CP** is a fundamental symmetry of nature, which is spontaneously broken. The breaking must somehow generate a large CKM phase but *not* a large θ . This requires some clever model-building. The paradigmatic example is the Nelson–Barr model [75, 76], invented by Ann Nelson when she was a graduate student. A related class of models uses a generalized parity symmetry (which can exchange some fields with others, in addition to acting on spacetime) instead of CP (e.g., [77]). Unfortunately, a discussion of such models goes well beyond what I have time to discuss in these lectures.
- **Axions!** This solution, by far the most popular, is the subject of much of the remainder of these lectures. My current belief is that the existence of an axion field with at least approximately the properties needed to solve Strong CP is likely to be a *requirement* of a consistent theory of quantum gravity, for reasons we will come to later in these lectures. Thus, I find this to be one of the most promising scenarios for physics beyond the Standard Model.

12.2 Axions for Strong CP: the core idea

The core idea of using axions to solve the Strong CP problem emerged from a set of papers by Peccei, Quinn, Weinberg, and Wilczek [78–81].⁸ The *only* necessary ingredients for this solution are a light, dynamical scalar field $\theta(x)$ —the axion—which is periodic, $\theta \cong \theta + 2\pi$, and has an approximate continuous shift symmetry $\theta \mapsto \theta + \text{const.}$ (like a Nambu-Goldstone boson) that is (to very good approximation) *only* broken by a coupling to gluons:

$$S_{\text{axion}} = \int d^4x \sqrt{|g|} \frac{1}{2} f^2 \partial_\mu \theta \partial^\mu \theta + \frac{N}{8\pi^2} \int \theta(x) \text{tr}(G \wedge G). \quad (155)$$

⁸Wilczek named the particle the “axion” after a laundry detergent, presumably motivated by its relationship to axial symmetries, and because it “cleans up” the Strong CP problem. Weinberg, on the other hand, originally named the particle the “higglet.”

There can also be a similar $\theta F \wedge F$ coupling to photons, which is harmless for the Strong CP problem and potentially useful for experimental detection; in terms of the axion potential generated—via magnetic monopole loops—this is a very subleading effect to the QCD-generated potential we will shortly discuss. I will refer to the number f as the axion decay constant, but be aware that some literature will instead refer to f/N with that term. (We will shortly see why the combination f/N is relevant.)⁹

Let's (re)emphasize some important theoretical points, already discussed around (122) above. First, $\theta \mapsto \theta + 2\pi$ is a *gauge* symmetry; these are two different labels for the same field configuration. Gauge invariant operators include $\partial_\mu \theta(x)$ or $e^{in\theta(x)}$ (with $n \in \mathbb{Z}$), but θ itself is *not* gauge-invariant. As a result, the action S_{axion} is not gauge invariant because the $\theta \text{tr}(G \wedge G)$ term above changes when we shift θ by a multiple of 2π . However, what must be well-defined for the path integral to make sense is only $\exp(iS_{\text{axion}})$, which transforms to $\exp[iS_{\text{axion}} + 2\pi i N \frac{1}{8\pi^2} \int \text{tr}(G \wedge G)]$. This is the same as $\exp(iS_{\text{axion}})$ because of the quantization of instanton number, (89), provided $N \in \mathbb{Z}$.

Why does this model solve the Strong CP problem? QCD dynamics generates a potential for θ which is minimized at the CP-preserving value $\theta = 0$. This follows from general principles [68]. However, we can actually go further. At high energies, $E \gg \Lambda_{\text{QCD}}$, small instantons (with size $\rho \sim E^{-1}$) generate calculable semiclassical contributions to the effective potential, proportional to $e^{-8\pi^2/g^2(E)} (e^{iN\theta} + \text{c.c.})$. Recall that the QCD scale comes from RG running with a beta function coefficient $b = \frac{11}{3}N_c - \frac{2}{3}N_f$, so it is proportional to $e^{-8\pi^2/(bg^2)}$ with $b > 1$, and the small-instanton contributions are expected to be much smaller than contributions from the QCD scale.

At low energies, $E \lesssim \Lambda_{\text{QCD}}$, a description of the dynamics in terms of gluons and quarks is no longer valid; we should describe the axion's interactions with hadrons. Remarkably, we can obtain such an accurate description using the chiral Lagrangian! Here is a very quick review of the chiral Lagrangian, which exploits the fact that the only light composite states in QCD are the pseudo-Nambu-Goldstone bosons associated with spontaneous chiral symmetry breaking, which take values in the coset $[\text{SU}(N_f)_L \times \text{SU}(N_f)_R]/\text{SU}(N_f)_V$. Thus, the low-energy EFT must be a sigma model living on this coset, which is accomplished by introducing the pseudo-Nambu-Goldstone boson fields π^i , packaged into a matrix

$$U(x) = \exp[2i\pi^a(x)T^a/f_\pi], \quad (156)$$

with T^a the hermitian generators of $\text{SU}(N_f)$. We use the normalization

$$f_\pi \approx 92 \text{ MeV}. \quad (157)$$

The matrix-valued field $U(x)$ transforms under the flavor symmetry as $U(x) \mapsto LU(x)R^\dagger$ with $L \in \text{SU}(N_f)_L, R \in \text{SU}(N_f)_R$. One then builds up the chiral Lagrangian from flavor-symmetry invariants, like the kinetic term $\frac{1}{4}f_\pi^2 \text{tr}[(D_\mu U)^\dagger D^\mu U]$. The quark masses explicitly break $\text{SU}(N_f)_L \times \text{SU}(N_f)_R$, and are incorporated as spurions in the form of a matrix M that transforms as

$$M \mapsto LMR^\dagger, \quad (158)$$

⁹The term “decay constant” is an artifact of history. The analogous quantity f_π in pion physics, which can be read off from the matrix element between an axial current and a one-pion state, has long been known as the pion decay constant. It should not be confused with the decay *width*. The charged pion decay width to leptons is proportional to f_π^2 , for example. On the other hand, the axion decay width turns out to scale as $1/f^2$, for fixed axion mass.

which at leading order can only appear in the chiral Lagrangian via a term

$$\mu^3 \text{tr} [MU^\dagger + M^\dagger U], \quad (159)$$

where μ is a dimensionful scale related to the scale of chiral symmetry breaking, $\mu^3 \sim \langle q\bar{q} \rangle$. Expanding out this term, we learn that pion masses scale like the square root of quark masses. If any of this is unfamiliar, I strongly encourage you to learn more [13–15]. The chiral Lagrangian is a key piece of Standard Model physics, and a beautiful example of how the logic of effective field theory and symmetry allows us to match aspects of UV physics to IR physics even across a strongly-coupled boundary region we don't understand. It is also vital to understanding the form of the axion potential, as we will now see.

At first glance, it might seem hopeless to match the gluonic operator $\text{tr}(G \wedge G)$ onto a theory of effective interactions of pions and kaons, which after all are more closely related to quarks than to gluons. But in fact, there is a clever trick for doing so. We can remove the axion coupling to gluons using the by-now familiar trick of rephasing a fermion field and exploiting the chiral anomaly, i.e., we perform a field redefinition like $q_L \mapsto q_L e^{iN\theta}$. This removes the axion–gluon coupling, produces a derivative coupling of the axion to the quark (not very important for our purposes at the moment), and changes the quark mass term:

$$mq_L \bar{q}_R + \text{h.c.} \mapsto me^{iN\theta} q_L \bar{q}_R + \text{h.c.} \quad (160)$$

We now have a shift-symmetry violating coupling of the axion to quarks (at leading order in small field fluctuations, this is a Yukawa coupling). This is useful, because we know that quark mass terms appear in the chiral Lagrangian via the term (159). Because of our field redefinition, the matrix M is now not a constant but has $\theta(x)$ dependence inside it! We can expand this out and obtain an axion potential, which takes the form [82, 83]

$$V_{\text{QCD}}(\theta) = -m_\pi^2 f_\pi^2 \sqrt{1 - \frac{4m_u m_d}{(m_u + m_d)^2} \sin^2 \frac{N\theta}{2}}. \quad (161)$$

Equation (161) is a good approximation to the axion potential generated by strong dynamics. There can be terms in the chiral Lagrangian involving higher powers of M , but because the quark masses are small, they should give small corrections. Expanding in small θ , and recalling the factor of f^2 in the axion kinetic term to properly normalize the field, this translates into an axion mass

$$m_a = \frac{m_\pi f_\pi}{f/N} \sqrt{\frac{m_u m_d}{2(m_u + m_d)^2}} \approx 6 \mu\text{eV} \frac{10^{12} \text{ GeV}}{f/N}. \quad (162)$$

Notice that this vanishes when a quark mass goes to zero, consistent with the massless up quark solution to Strong CP that we discussed earlier, when the value of θ has no physical effect. I've chosen a reference value of 10^{12} GeV for f/N because this turns out to be about the right value for axions to constitute all of the dark matter in the universe, in the simplest axion cosmology scenario. Thus, we expect the axion particle to be quite light, but also very weakly interacting, because all of its couplings are suppressed by the high energy scale f . Let's highlight the conclusion:

The strong dynamics of QCD generates a potential for the axion, which relaxes it to the value that cancels the θ term, explaining why we do not see a nonzero neutron EDM. The axion mass is of order $m_\pi f_\pi/f$. The axion is very light and very weakly coupled when f is a UV scale.

Exercise. Fill in the details, i.e., derive the potential (161) by carrying out an axion-dependent rephasing of a quark field and matching to the chiral Lagrangian.

12.3 Axions for Strong CP: further details

I have streamlined the above discussion somewhat, in the interest of stating the core points about axion physics concisely. Here are a few further details and clarifications.

Above, I have phrased the whole discussion in terms of the θ term (150). However, we know that due to the chiral anomaly such a term can be altered by a fermion field redefinition (see §9.5). We used this to shift θ entirely into the quark mass matrix to compute the axion potential. However, in general, we might have *both* a θ term and a phase of the quark mass matrix M . The physically meaningful combination is

$$\bar{\theta} = \theta - \arg \det M. \quad (163)$$

The chiral anomaly allows us to shift θ and $\arg \det M$ by the same constant, keeping $\bar{\theta}$ unchanged. Thus, the whole discussion surrounding (155) should be understood as one in which we have initially rephased the fermions to set $\arg \det M = 0$, and absorbed the corresponding constant shift in the θ term into the definition of our axion field $\theta(x)$.

We discussed the axion mass in terms of the chiral Lagrangian, which includes the pseudo-Nambu-Goldstone bosons of spontaneous chiral symmetry breaking. These are the pion fields, and (if we are interested in subleading effects) also the kaon and η fields. The remaining pseudoscalar boson, the η' , obtains a significantly larger mass because of the $U(1)_A$ anomaly. Nonetheless, one can include it and its mixing with the pions and the axion by keeping the extra $U(N_f)$ generator in the definition of the field $U(x)$ in the chiral Lagrangian (i.e., not requiring that $\det U = 1$, but allowing it to have a phase), and giving the extra mode a mass via a $(\log \det U)^2$ term in the chiral Lagrangian. In this case, $\log \det U$ shifts under a chiral rephasing of the quarks. You can read more in [82, 84].

12.3.1 The axion-photon coupling from pion mixing

The theory above the QCD scale (but below the electroweak scale) can contain a coupling

$$\frac{E}{8\pi^2} \int \theta F \wedge F, \quad (164)$$

where F is the field strength of electromagnetism, normalized so that the charge of the electron is -1 as usual. Here E is quantized, as discussed around (122). However, it is not necessarily quantized in integer units, because in our conventional normalization electric charge can be as small as $1/3$. This is a somewhat subtle point. Even if we assume that there are no particles in nature with

hypercharge smaller than $1/6$, the Standard Model gauge group is still ambiguous: it takes the form $[\text{SU}(3)_C \times \text{SU}(2)_L \times \text{U}(1)_Y]/\Gamma$ where Γ can be trivial, \mathbb{Z}_2 , \mathbb{Z}_3 , or \mathbb{Z}_6 . This is because, given the Standard Model matter content alone, the centers of the nonabelian factors act on the fields in the same way as a hypercharge transformation. An extensive discussion can be found in [85]. If $\Gamma = \mathbb{Z}_3$ or \mathbb{Z}_6 , the QED θ angle in its conventional normalization (24) has period 2π and so we must have $E \in \mathbb{Z}$. On the other hand, if $\Gamma = \mathbf{1}$ or \mathbb{Z}_2 then the QED θ angle has period $2\pi/9$. However, we've normalized the *axion* field in (164) to have period 2π , so E is allowed to be any integer multiple of $1/9$. Currently we have no experimental evidence either way, so the most general statement we can make is (assuming the smallest hypercharge really is $1/6$) that $9E \in \mathbb{Z}$.

In any theory with the coupling (155), the axion will couple to photons, *whether or not* it has a quantized coupling to electromagnetism of the form (164). The reason is that, because of its coupling to gluons, the axion inevitably mixes with the π^0 meson [86–89], and the π^0 couples to photons. Superficially, this coupling seems to violate the quantization condition that we derived on topological grounds in §10. It does not, because this coupling is proportional to the axion mass, which is to say it is really a coupling of the form $(\Box\theta)F \wedge F$ or (using equations of motion) $\sin(\theta)F \wedge F$. Ordinarily, we expect that such contributions are highly subdominant, because they are proportional to m_a^2 and the axion mass is small. This case is an exception; since the axion mass and the pion mass are both generated at the QCD scale, we have $m_a^2 f^2 \sim m_\pi^2 f_\pi^2$, and it turns out that the coefficient of the axion coupling to photons induced by mixing with the pion is of order $(m_a^2 f^2)/(m_\pi^2 f^2)$, so it gives an $O(1)$ modification. For a more detailed discussion on this point, see Appendix A of [90].

The end result is that, below the QCD scale, the axion has an effective coupling to photons of the form [83, 89]

$$\frac{1}{8\pi^2} \left(E - \frac{2}{3} N \frac{4m_d + m_u}{m_d + m_u} \right) \int \theta F \wedge F, \quad (165)$$

up to subleading terms (e.g., those suppressed by the kaon mass). What one encounters most often in the phenomenological literature is a coupling $g_{a\gamma\gamma}$ written in terms of the canonically normalized axion field $\hat{\theta} = f\theta$ and the canonically normalized photon field strength \hat{F} , of the form

$$\begin{aligned} \mathcal{L}_{a\gamma\gamma} &= -\sqrt{|g|} \frac{1}{4} g_{a\gamma\gamma} \hat{\theta} \hat{F}_{\mu\nu} \tilde{F}^{\mu\nu}, \\ g_{a\gamma\gamma} &= \frac{\alpha}{2\pi f/N} \left(\frac{E}{N} - 1.92(4) \right). \end{aligned} \quad (166)$$

Here α is the fine structure constant. The factor f/N is often simply called f_a , as it is the combination appearing in the coupling to gluons. The numerical value $-1.92(4)$ is a relatively recent estimate from [83], including subdominant terms beyond those in (165).

12.4 Terminology: pseudo-Nambu-Goldstone bosons, QCD axions, and ALPs

Before looking at detailed models of axions that solve the Strong CP problem, I want to make a brief aside about terminology. A pseudo-Nambu-Goldstone boson (PNGB) arises whenever an approximate continuous global symmetry is broken. If the global symmetry is *compact* (as it usually is), then the PNGB will parametrize a compact field space. For instance, pion fields in QCD-like theories take values in the compact coset manifold $[\text{SU}(N_f)_L \times \text{SU}(N_f)_R]/\text{SU}(N_f)_V$.

When an approximate $U(1)$ global symmetry is spontaneously broken, the field space is a circle, so the PNGB is a periodic scalar field, $\theta \cong \theta + 2\pi$. It is common for such a periodic scalar field to be referred to as an axion, or axion-like particle, even outside of the context of axion solutions of the strong CP problem. Some people prefer to reserve the word “axion” for the original context of a periodic scalar coupled to $\text{tr}(G \wedge G)$ and solving the Strong CP problem. Those who favor the strict use of the word “axion” favor the term “axion-like particle” or “ALP” for similar fields that do not interact with QCD and solve the Strong CP problem.¹⁰ On the other hand, people who use the term “axion” more broadly will often say “QCD axion” to refer to the original case. I usually fall in the latter camp, although in this part of the lecture notes I will simply write “axion” since the context is the Strong CP problem.

Any periodic scalar field can be thought of as a PNGB, if only in the trivial sense that the low-energy theory contains the shift symmetry current $\partial_\mu \theta$ (or $\star d\theta$) that creates a single particle from the vacuum. In the original Peccei-Quinn scenario, as well as in other models we will review shortly, the QCD axion explicitly arose as a PNGB for a spontaneously broken approximate global $U(1)$ “PQ” symmetry. However, we will also discuss models in which the axion arises from a higher-dimensional gauge field, where there is no 4d PQ symmetry to break, so in such models axions are only PNGBs in the trivial sense.

In my usage, the term “PNGB” would generally be associated with pseudo-Nambu-Goldstone bosons for approximate symmetries broken by generic operators, whereas the term “axion” mostly refers to special cases where couplings of the form $\theta F \wedge F$ or $\theta \text{tr}(F \wedge F)$ are the *dominant* sources of breaking of the scalar field’s shift symmetry. In any case, you will quickly learn to infer from context what someone means by the term “axion.”

13. Classic 4d axion models; axion quality problem

13.1 The KSVZ model

The KSVZ (Kim [91]; Shifman, Vainshtein, Zakharov [92]) axion model is the simplest, most canonical model of a QCD axion. It consists of a complex scalar ϕ and two new fermion fields, Q and \bar{Q} , transforming in the $\mathbf{3}$ and $\bar{\mathbf{3}}$ of $SU(3)_C$ respectively.

$$\frac{\mathcal{L}_{\text{KSVZ}}}{\sqrt{-g}} = \frac{1}{2} \partial_\mu \phi^* \partial^\mu \phi + i Q^\dagger \bar{\sigma}^\mu D_\mu Q + i \bar{Q}^\dagger \bar{\sigma}^\mu D_\mu \bar{Q} + (y \phi Q \bar{Q} + \text{h.c.}) - V(\phi^* \phi). \quad (167)$$

(As in §9, the bar does not denote complex conjugation; Q and \bar{Q} are *independent* Weyl fermion fields and the bar is just part of the name of the field \bar{Q} .) In general, Q and \bar{Q} can transform under $SU(2)_L \times U(1)_Y$ as well, but the choice of representation may be model-dependent; this choice, of course, determines the form of the covariant derivatives D_μ . The model in which Q and \bar{Q} are neutral under electroweak interactions is the minimal model that you will often see labeled simply “KSVZ” on plots of constraints on the axion parameter space.¹¹ This Lagrangian, as written, has a

¹⁰I have also seen “ALF” for “axion-like field,” though US readers of a certain age will associate this more strongly with the Alien Life Form from a late-1980s sitcom.

¹¹If Q and \bar{Q} have no hypercharge, they cannot decay to any Standard Model state due to the fractional hypercharge assignments of the ordinary quarks. This poses a potential cosmological problem, although the Q or \bar{Q} particles would

classical global (0-form) symmetry $U(1)_{PQ} \times U(1)_Q$ acting as

$$\begin{aligned} U(1)_{PQ} : \quad \phi &\mapsto e^{i\alpha} \phi, \quad Q \mapsto e^{-i\alpha} Q; \\ U(1)_Q : \quad Q &\mapsto e^{i\beta} Q, \quad \bar{Q} \mapsto e^{-i\beta} \bar{Q}. \end{aligned}$$

In a more complete model $U(1)_Q$ might be subsumed into a larger baryon number symmetry, or simply broken explicitly. It is $U(1)_{PQ}$ that will concern us here. We assume that $V(\phi^* \phi)$ has a symmetry-breaking form, so that $\langle \phi \rangle \neq 0$ and the $U(1)_{PQ}$ symmetry is spontaneously broken at the potential's minimum. Thus, at the classical level, there is a $U(1)$ Nambu-Goldstone boson θ , which is the Peccei-Quinn axion mode in this theory, parametrizing a circular vacuum manifold.

The expectation value of ϕ determines the axion decay constant. There is a corresponding massive radial mode of ϕ , moving up the hill away from the minimum, which is sometimes called the “saxion” field or the *scalar partner* of the axion. (The terminology “saxion” has been mostly used in supersymmetric theories, where one adds an ‘s’ at the beginning of the name of a fermion to refer to its scalar superpartner, e.g., squarks and sleptons; here, we are referring instead to a scalar partner of a *different* (pseudo)scalar, and the terminology need not be limited to supersymmetric theories.) Summing up,

$$\phi(x) = \frac{1}{\sqrt{2}} \left(\underbrace{f}_{\text{decay constant}} + \underbrace{s(x)}_{\text{saxion}} \right) \exp(i \underbrace{\theta(x)}_{\text{axion}}). \quad (168)$$

In this parametrization, it is manifest that $\theta(x)$ is a *periodic* scalar field, defined only modulo 2π shifts. In more formal language, we can say that in the low-energy effective theory (below the mass scale of the saxion), there is an *emergent* \mathbb{Z} gauge symmetry, $\theta \cong \theta + 2\pi n$. This symmetry breaks down at the origin of field space, $\phi = 0$, where θ is no longer well-defined (not even mod 2π). However, that point is not accessible within the low-energy theory because $s(x)$ is massive.

Inspecting (167), we see that the fields Q and \bar{Q} pair up into a massive Dirac fermion of mass

$$m_Q = \frac{1}{\sqrt{2}} y f. \quad (169)$$

while the saxion gets a mass from the potential. If the potential has a simple quartic form, $V(\phi^* \phi) = \lambda_\phi (\phi^* \phi - f^2/2)^2$, then the saxion mass is

$$m_s = \sqrt{2\lambda_\phi} f. \quad (170)$$

At energies $\Lambda_{QCD} \ll E \ll f$, we can integrate out the heavy fields Q , \bar{Q} , and s to obtain an EFT of the axion $\theta(x)$ coupled to the Standard Model. However, there is a small complication. When we substitute the ansatz (168) into (167), we see that the mass term for the heavy quarks is θ -dependent:

$$m_Q e^{i\theta(x)} Q \bar{Q}(x) + m_Q e^{-i\theta(x)} Q^\dagger \bar{Q}^\dagger(x). \quad (171)$$

mostly annihilate away and any surviving asymmetric population must be bound into heavy hadrons. The stability of $Y = 0$ quarks is related to the question of the global structure of the Standard Model gauge group mentioned above in §12.3.1. Finding a color triplet particle with zero hypercharge would imply that the \mathbb{Z}_6 quotient is not the true global structure of the gauge group.

We cannot simply set $Q = \bar{Q} = 0$ in the Lagrangian due to this θ -dependence. We have to integrate out the quarks more carefully. This can be done by computing loop diagrams, but a more efficient way is to use the chiral anomaly. We can eliminate $\theta(x)$ from (171) by carrying out a field redefinition,

$$Q(x) \mapsto e^{-i\theta(x)} Q(x). \quad (172)$$

Substituting this field redefinition into (167), we find no non-derivative interaction between θ and the heavy fields, which we can then integrate out by simply setting them to zero.¹² However, due to the chiral anomaly, our field redefinition changes the path integral measure in a way that corresponds to adding a new term to the action, of the form (116):

$$-\frac{1}{8\pi^2} \int \theta(x) \text{tr}(G \wedge G), \quad (173)$$

where we have used the Dynkin index $I_2(\mathbf{3}) = 1/2$. This coupling reflects the existence of a chiral $U(1)_{\text{PQ}}\text{-SU}(3)_C^2$ anomaly. (This is an ABJ anomaly: QCD explicitly breaks the would-be global symmetry $U(1)_{\text{PQ}}$, and in particular we cannot gauge it.) Thus we see that θ couples to gluons, and our effective theory has the form (155) that we have previously argued solves the strong CP problem (with $N = -1$). In the case that Q and \bar{Q} carry electroweak charges, there would be additional interactions of θ with the electroweak gauge fields.

The PQ symmetry is *not* an accidental symmetry of $\mathcal{L}_{\text{KSVZ}}$.¹³ Gauge invariance does not forbid relevant terms like $M_Q Q \bar{Q} + \text{h.c.}$ or $b_\phi \phi^2 + \text{h.c.}$ (where M_Q, b_ϕ are parameters of positive mass dimension), which would explicitly break the symmetry. This is an important point to which we will return later: we never expect global symmetries to be fundamental, so the theory must have more structure in order to explain why these terms are suppressed.

Exercise: convince yourself that we can still talk about an EFT of an axion, even with (small) explicit breaking of Peccei-Quinn symmetry (beyond the ABJ anomaly). More explicitly: for a sufficiently small $b_\phi \phi^2$ term, show that the parametrization (168) is still sensible. We can integrate out the mode $s(x)$ and write an effective theory of $\theta(x)$. However, now $\theta(x)$ has a potential. What condition is required for this to be subdominant to the potential generated by QCD dynamics?

In fact, the PQ symmetry is not even a symmetry of the *quantum* theory with Lagrangian $\mathcal{L}_{\text{KSVZ}}$, due to its ABJ anomaly with $SU(3)_C$. It is an explicitly broken symmetry. This makes it even harder to see why the Lagrangian should have respected the symmetry in the first place.¹⁴ Because of the ABJ anomaly, the axion field θ is not a true Nambu-Goldstone boson but a *pseudo*-Nambu-Goldstone boson. As such, we expect it to acquire a mass; the coupling $\theta \text{tr}(G \wedge G)$ indeed generates a mass, as discussed in §12.2.

¹²We do find a derivative interaction of the axion with the heavy quarks, $Q^\dagger \overleftrightarrow{\sigma}^\mu Q \partial_\mu \theta$. This respects a continuous shift symmetry of θ . When we integrate out the heavy quarks, this leads to derivative self-interactions of the axion, with effects that vanish at small momentum and have negligible impact on axion phenomenology.

¹³An accidental symmetry is one that can only be broken by irrelevant operators; see §17.

¹⁴Interestingly, it has recently been shown that in some cases, a type of generalized *non-invertible* symmetry remains even after an ABJ anomaly [93, 94], though this is not the case for the mixed anomaly with QCD discussed here.

13.2 DFSZ model

The DFSZ model (Zhitnitsky [95]; Dine, Fischler, Srednicki [96]) is the second classic model of a weakly-coupled axion. In this model, the Standard Model is extended to a two Higgs doublet model (2HDM), and the PQ symmetry acts on the Higgs fields as well as on a heavy complex scalar ϕ . Because the Higgs fields carry PQ charge and have Yukawa couplings to SM fermions, the fermion fields must also carry PQ charge. Thus, DFSZ models are more complicated than KSVZ models.

As an example, consider a Type II 2HDM, in which the two Higgs doublets H_u and H_d couple to Standard Model fermions via

$$y_u H_u q \bar{u} + y_d H_d q \bar{d} + y_e H_d \ell \bar{e} + \text{h.c.} \quad (174)$$

This structure arises in the MSSM, though we can also consider it outside the context of supersymmetry. We take the fields H_u and H_d to have PQ charge +1, and the fields \bar{u} , \bar{d} , and \bar{e} to have PQ charge -1. We also take the complex scalar ϕ to have PQ charge +1. This allows for a quartic coupling between the Higgs bosons and the ϕ field,

$$\lambda_{ud\phi} H_u H_d \phi^{\dagger 2} + \text{h.c.} \quad (175)$$

Again, we assume that the field ϕ has a potential leading to a VEV as in (168), with $f \gg v_u, v_d$. From this we immediately see that we either require $\lambda_{ud\phi} \lesssim v^2/f^2$, or the model must have some fine tuning to separate the weak scale from the scale of the axion decay constant. Of course, this is not unique to the DFSZ model; quite generally, extensions of the Standard Model involving heavy mass scales generate electroweak fine-tuning problems. In the KSVZ model, however, the corrections to the Higgs mass arise only at higher loops, whereas in the DFSZ case (175) is an integral part of the model and the problem is already apparent at tree level.

As in the KSVZ case, we would like to integrate out the radial mode of the heavy field $\phi(x)$ and treat its phase $\theta(x)$ as a pseudo-Nambu-Goldstone boson that survives in the low-energy theory. However, due to the coupling (175), the phase $\theta(x)$ appears in an effective Higgs mass term below the scale f . We can perform a field redefinition to rephase the Higgs fields and eliminate this term, but then the phase will appear in the Yukawa couplings (174). Finally, we can rephase Standard Model fermions to eliminate the phase in the Yukawa couplings, but this generates a coupling of $\theta(x)$ to gluons via the chiral anomaly (as in §9.5) as well as derivative couplings of Standard Model fermions to the axion, of the form

$$\frac{v_u^2}{v_u^2 + v_d^2} (\partial_\mu \theta) \bar{u}^\dagger \bar{\sigma}^\mu u + \frac{v_d^2}{v_u^2 + v_d^2} (\partial_\mu \theta) \bar{d}^\dagger \bar{\sigma}^\mu d + \dots \quad (176)$$

From the experimental viewpoint, then, an important distinction between KSVZ models and DFSZ models is that the axion couples more strongly to Standard Model fermions in DFSZ models. Note that the derivative coupling to fermions cannot induce a non-derivative coupling to gauge fields like photons, because the former preserves a continuous shift symmetry, whereas the latter preserves only a discrete shift of θ . As we saw in §10, the non-derivative couplings have quantized coefficients for topological reasons, so they cannot be generated by non-quantized couplings like (176).

A full discussion of the couplings in the DFSZ model is not very useful for the more conceptual points I want to focus on in this note, but I do want to make a few brief comments on how to analyze such models. Much of the literature on DFSZ models carries out field redefinitions on Standard Model fermions of the form $\psi(x) \mapsto e^{i\alpha\theta(x)}\psi(x)$. If $\alpha \notin \mathbb{Z}$, this is not a mathematically sensible operation. Relatedly, many of these papers also refer to $U(1)$ charges that are irrational numbers. Despite such intermediate steps that (strictly speaking) make no sense, they tend to get the right answers. My collaborators and I tried to explain how to do these calculations carefully in [97], a paper written in the context of a particular experimental anomaly but one that (I hope) may be a useful reference outside the context of the anomaly. In any case, it's a good exercise for you to work through all the details of the DFSZ model for yourself.

Exercise (somewhat open-ended): Explore the EFT obtained by integrating out heavy fields in the DFSZ model. Notice that you have some choices along the way. For example, you might choose to eliminate the phase in (175) by carrying out a field redefinition only on H_u , or alternatively only on H_d . This would then lead to different field redefinitions on fermions to eliminate phases in (174). Understand why the results are physically equivalent despite such arbitrary choices.

In the DFSZ model, the axion acquires a coupling to photons due to the chiral anomaly associated with the charged Standard Model fermions. This is unlike the (minimal) KSVZ model, where the PQ-charged quarks are electrically neutral. However, people tend to overstate the importance of this difference: one could consider a modified KSVZ model where Q and \bar{Q} have hypercharge, so the size of the axion-photon coupling in the two models is not a real structural difference between them. In any case, it turns out that the contribution to the axion-photon coupling from the anomaly in the DFSZ model is $8/3$, but this must be combined with the contribution from the axion-pion mixing discussed in §12.3.1. The result is that in this model, we have, following (166),

$$g_{a\gamma\gamma} \propto \frac{E}{N} - 1.92 = \frac{8}{3} - 1.92 \approx 0.75, \quad (177)$$

which is significantly smaller (in absolute value) than the pion mixing contribution alone. Thus, the (minimal) DFSZ model predicts a smaller axion-photon coupling than the (minimal) KSVZ model, and for this reason it is often taken as a target for ambitious experiments that wish to probe the full range of possible axion-photon couplings. (Of course, one could easily write down another model where $E/N = 2$, and then would have much more of a challenge!)

The value $E/N = 8/3$ appearing in the minimal DFSZ model is also characteristic of a large class of GUT models, in which the Standard Model gauge group embeds in an $SU(5)$ subgroup of the GUT gauge group and an axion couples to $\text{tr}(F \wedge F)$ for the full GUT group. You can find an extensive discussion of the phenomenology of axions in GUTs in the recent paper [98]. Note that a DFSZ model need not be a GUT model: the defining feature of DFSZ is Higgs fields that carry PQ charge, independent of the details of far-UV physics. Similarly, not every GUT model with an axion is a DFSZ model: models with no 4d Peccei-Quinn symmetry at all, along the lines we will discuss in §14, could be GUT models.

For all of these reasons, I don't like the practice of labeling experimental plots with "KSVZ" and "DFSZ" lines. I would prefer to see labels like " $E/N = 0$ " and " $E/N = 8/3$ " that don't bias

the interpretation toward specific UV completions, but there is enough inertia behind the current practice that it's unlikely to change.

13.3 Axion quality problem

Throughout the above discussion we've been assuming that we can impose the global PQ symmetry to restrict the terms that we write in the Lagrangian. As we already hinted in §13.1, there are two problems with this. The first is that $U(1)_{\text{PQ}}$ has an ABJ anomaly, which means that it is not a symmetry of the quantum theory at all. The second is that even without the anomaly, it is a global symmetry, and (as we will discuss extensively in §19 and §21 below), we do not expect global symmetries to ever exist in theories of quantum gravity.

This is a severe problem, because explicit PQ-breaking terms can completely spoil the solution to the Strong CP problem! For example, we can consider the KSVZ scenario with an added Peccei-Quinn-violating (“PQV”) term in the Lagrangian of the form

$$\frac{1}{\sqrt{-g}} \mathcal{L}_{\text{PQV}} = \frac{c}{M_{\text{Pl}}^{n-4}} \phi^n + \text{h.c.}, \quad (178)$$

where the coefficient c in general can have a complex phase φ ,

$$c = |c| e^{i\varphi}. \quad (179)$$

Expanding around a ϕ VEV as in (168), this becomes an effective axion potential,

$$V_{\text{PQV}}(\theta) = \frac{|c|}{M_{\text{Pl}}^{n-4}} \left(\frac{f}{\sqrt{2}} \right)^n \left[e^{i(\varphi+n\theta)} + e^{-i(\varphi+n\theta)} \right] = 2|c|M_{\text{Pl}}^4 \left(\frac{f}{\sqrt{2}M_{\text{Pl}}} \right)^n \cos(n\theta + \varphi). \quad (180)$$

If CP is not a fundamental symmetry, there is no reason for φ to be a small phase. As a result, such an effective potential term can shift the minimum of $V(\theta)$ away from $\theta = 0$. However, experimentally, we know that $|\theta| \lesssim 10^{-10}$, so either the magnitude of $V_{\text{PQV}}(\theta)$ or the phase φ must be extremely small.

This problem is known as the *axion quality problem*, and it is extremely severe for simple models of KSVZ or DFSZ type. The QCD axion potential (161) is naturally exponentially small, because the overall size of the potential is set by the QCD scale, which arises from dimensional transmutation. By contrast, the PQV contribution (180) is suppressed only by a *power* of f/M_{Pl} . In order for a power-law suppressed potential to be many orders of magnitude smaller than an exponentially suppressed potential, we require the power to be large. For example, for $|c| = 1$ and $f = 10^{12}$ GeV we compute

$$V_{\text{PQV}}(\theta) \approx \begin{cases} (250 \text{ TeV})^4 \cos(8\theta + \varphi), & n = 8, \\ (73 \text{ MeV})^4 \cos(12\theta + \varphi), & n = 12, \end{cases} \quad (181)$$

showing that even moderately large values of n would give contributions that overwhelm the QCD axion potential without further suppression in the coefficient or the phase. Indeed, for $O(1)$ values of $|c|$ and φ we need $n \geq 14$ to avoid spoiling the solution of the Strong CP problem.

The axion quality problem, then, requires that we either forbid (using an exact gauge symmetry) or strongly (exponentially) suppress *many* dangerous operators so that the desired low-energy axion

EFT (155) dominates the dynamics [99–102]. Notice that the first priority is to forbid *relevant* operators, which are a major hazard in typical axion models without additional gauge symmetries. But because the problem is so severe, even highly irrelevant operators can be dangerous.

Within 4d axion models, a solution to the axion quality problem necessarily requires a theory with extended gauge symmetry beyond the Standard Model. One of the simplest approaches is to invoke a discrete gauge symmetry, e.g., a \mathbb{Z}_k subgroup of $U(1)_{\text{PQ}}$. For such a subgroup to be non-anomalous, given that $U(1)_{\text{PQ}}$ is anomalous, we require that the constant N in (155) is a multiple of k . This is potentially a viable solution, though the values of k that are required are awkwardly large. Another approach is to consider a model of a composite axion, replacing the complex Peccei-Quinn field ϕ in the basic axion models with a composite operator of larger scaling dimension. If this operator dimension is high enough, the number of low-dimension operators we must forbid in the Lagrangian can be much smaller than in the simplest models. One can also combine discrete gauge symmetries and compositeness.

Over the years, a great deal of ingenuity has been applied in constructing 4d axion models that evade the quality problem. However, in my opinion, the best solution is to abandon models with a 4d Peccei-Quinn symmetry entirely, as discussed in the next section.

14. Axions from higher-dimensional gauge fields

A particularly elegant solution to the axion quality problem is to dispense with the whole idea of spontaneously breaking a 4d $U(1)_{\text{PQ}}$ symmetry, and instead to derive the basic low-energy axion action (155) from an entirely different UV starting point: a higher-dimensional gauge theory with a Chern-Simons term. Zero modes of a higher-dimensional gauge field can automatically have the key features of axions, with *exponentially good* control of the axion quality problem.

14.1 Basic ingredients for extra-dimensional axions

This idea originated (shortly after the KSVZ and DFSZ papers) in string theory examples, where the gauge field in question was a higher p -form field [103–105]. To illustrate the core idea, we will present a simpler example arising from an ordinary 1-form gauge field in a 5d context (see [106, 107] and especially [108] for similar phenomenological models). We consider a 5d theory of a $U(1)$ gauge field A compactified to 4d on a circle S^1 with coordinate $x^5 \cong x^5 + 2\pi R$. We identify the 4d axion θ as a Kaluza-Klein zero mode of A . That is:

$$\theta(x) = \int_{S^1} A(x) = \int_0^{2\pi R} A_5(x, x^5) dx^5. \quad (182)$$

This is a periodic variable, $\theta \cong \theta + 2\pi$. To see this, consider a winding gauge transformation of A in 5d, with the $U(1)$ element $g(x, x^5) = \exp(ix^5/R)$. Then $A \mapsto A + d\alpha$ where $\alpha(x^5) = \frac{x^5}{R}$ is not single-valued. Under such a gauge transformation, $\int_{S^1} A \mapsto \int_{S^1} A + 2\pi$. This shows that θ is not a well-defined (gauge invariant) quantity, but it is defined modulo 2π , and in particular $e^{i\theta}$ is well-defined. This discussion should be familiar: it's precisely how we described Wilson loops in §4.3. Another way to explain the origin of the 4d axion field is that the gauge invariant Wilson loop around the 5d circle over the 4d point x , $W(S_x^1)$, is an element of $U(1)$. We identify this element as $e^{i\theta(x)}$ to define the 4d periodic scalar $\theta(x)$.

Given that a 4d periodic scalar exists, our next task is to identify the origin of the two pieces of the action (155): the kinetic term, from which we can read off the decay constant f , and the coupling to gluons that is crucial for solving the Strong CP problem. The kinetic term arises directly from the 5d kinetic term for a gauge field, where the gauge coupling squared e_5^2 has units of length,

$$\int_{M \times S^1} -\frac{1}{2e_5^2} F \wedge \star_{5d} F, \quad (183)$$

where M is the 4d spacetime manifold. This becomes, with the ansatz $A = \theta dx^5 / (2\pi R)$,

$$\int_M \frac{1}{2} f^2 d\theta \wedge \star d\theta, \quad f^2 = \frac{1}{2\pi R e_5^2}. \quad (184)$$

Thus, we see that a small f can be achieved when the compactification radius is large compared to the 5d length scale e_5^2 . The second ingredient is a 5d Chern-Simons term, of the general type discussed in §11. For this, we have to extend the Standard Model gauge fields over the extra dimension, so we suppose that there are $SU(3)$ gluon fields G also propagating in 5d, with coupling

$$\frac{N}{8\pi^2} \int_{M \times S^1} A \wedge \text{tr}(G \wedge G). \quad (185)$$

In 4d this directly becomes the usual axion $\theta \text{tr}(G \wedge G)$ Chern-Simons term.

Because the gluon fields propagate in the extra dimensions, the 4d gluon kinetic term arises when the 5d gluon kinetic term

$$\int_{M \times S^1} -\frac{1}{2g_5^2} \text{tr}(G \wedge \star_{5d} G) \quad (186)$$

becomes, with the simple ansatz that G is independent of the 5th dimension,

$$\int_M -\frac{1}{2g^2} \text{tr}(G \wedge \star G), \quad \text{where} \quad \frac{1}{g^2} = \frac{2\pi R}{g_5^2}. \quad (187)$$

Thus, we see that the smallness of the 4d $SU(3)_C$ gauge coupling (at the compactification energy, where we match 5d to 4d) can potentially be explained by a relatively large volume of the internal dimension.

One major difference between such a higher-dimensional axion and the familiar 4d models is that we have not mentioned a $U(1)_{PQ}$ symmetry at all! In particular, the higher-dimensional gauge group is *not* the same thing as the Peccei-Quinn symmetry group. In these models there is no Peccei-Quinn symmetry breaking phase transition. Is the axion still a pseudo-Nambu-Goldstone boson? Yes, but only in the trivial sense mentioned in §12.4: there is an approximately conserved current that, when acting on the vacuum, produces a single-particle state of the axion. This current is simply $\partial_\mu \theta$ itself (or in differential form notation, $\star d\theta$), that is, the shift symmetry current. The sense in which the axion is a pseudo-Nambu-Goldstone boson in these theories is the same sense in which the photon is a pseudo-Nambu-Goldstone boson, which we will discuss later in §20.2.2. The Peccei-Quinn symmetry is just the 4d remnant of the electric 1-form symmetry of the 5d theory.

The case of a 5d compactification on a circle is somewhat unappealing, because it also produces a massless $U(1)$ gauge boson (the Kaluza-Klein gauge field) in 4d. It is also unclear where 4d

chiral fermions would originate in this construction. For a more realistic model, we could take two approaches. One is to compactify the 5d theory on an interval, rather than a circle. However, in this case we will not necessarily find a corresponding massless axion in 4d, depending on the boundary conditions. The other approach, which is realized in a large collection of string theory examples, is to consider a geometry with multiple extra dimensions and obtain the axion by reducing a p -form gauge field on a p -dimensional cycle within the extra dimensions. In this case, there is always a (perturbatively) massless mode for every such cycle. We will comment on this case further in §14.3. For now, let's continue to investigate the 5d toy model on a circle, as it provides the simplest setting in which to learn some important qualitative lessons.

14.2 Quality problem for extra-dimensional axions

Because A is a gauge field, it is exactly massless in the 5d theory. So there is no immediate mystery of why we don't have a large bare potential for θ . (One could wonder about a Stueckelberg or BF -type mass for A , but these depend on an integer coefficient, so it is perfectly consistent to set any such coupling to zero and forget about it.) In general, 5d couplings that depend on the field strength $F = dA$ become 4d terms depending on $d\theta$, which do not generate an axion potential. Thus, the only terms that can possibly matter for the axion quality problem are 5d terms that depend *non-derivatively* on A . In gauge theories, such terms are very highly constrained. One such term is the 5d Chern-Simons term (185), which becomes the 4d axion coupling to gluons, and is crucial for generating the potential V_{QCD} . We could also potentially have 5d Chern-Simons couplings of A to different gauge fields (e.g., hidden sector gluons). These would generate a 4d potential in a similar manner to QCD. In particular, such a potential would be exponentially small, and so it would not be surprising for such contributions to be small enough, relative to V_{QCD} , to not spoil the Strong CP solution.

The other 5d source of couplings depending non-derivatively on A is the existence of 5d fields charged under the gauge field. In other words, the 5d covariant derivative for a field with $U(1)$ charge $q \in \mathbb{Z}$, which is $D = d + iqA$, leads to θ -dependent terms in 4d. The axion field lives in A_5 , so we specifically look at the 5th component:

$$D_5 = \underbrace{\partial_5}_{i \frac{n}{R}, \text{ KK number}} + iq \underbrace{A_5}_{\frac{\theta}{2\pi R}}. \quad (188)$$

A 5d field with mass m_{5d} becomes an infinite tower of 4d fields with $n \in \mathbb{Z}$ units of Kaluza-Klein momentum around the circle, and an axion-dependent mass,

$$m_n^2 = m_{5d}^2 + \frac{1}{R^2} \left(n + \frac{q\theta}{2\pi} \right)^2. \quad (189)$$

This is a manifestly non-derivative coupling of the axion. We also see that the mass of a given KK mode m_n is *not* a periodic function of θ , even though our theory is supposed to be gauge invariant under $\theta \mapsto \theta + 2\pi$. It *is*, but in a nontrivial way: the mass of an individual mode is not invariant, but the mass of the mode with number $n - q$ shifts to match the mass that the mode with number n originally had. The infinite tower rearranges itself so that the full spectrum is invariant under a 2π shift of θ . This phenomenon, illustrated in Fig. 10, is generally referred to as *monodromy*. The

simplest example, which you may have encountered before, is the quantum mechanical problem of a particle on a ring.

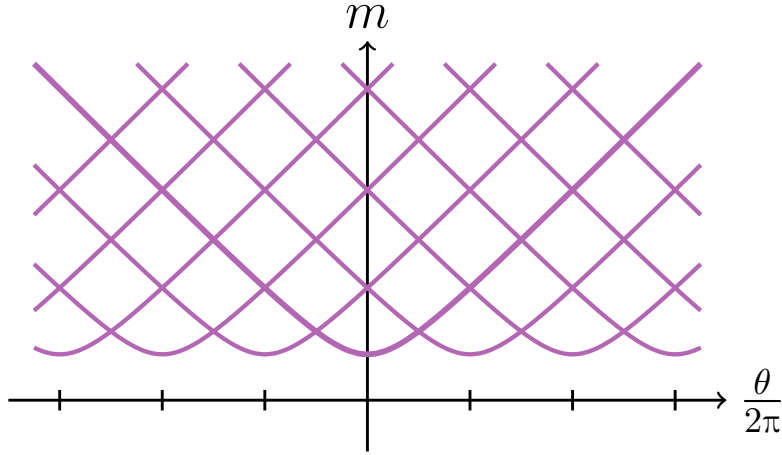


Figure 10: Spectrum of KK mode masses (189) as a function of θ , illustrating monodromy. The $n = 0$ mode is drawn with a thicker line; other curves are for different values of n . The spectrum of masses is 2π periodic, but the mass of an individual mode is not.

Once we have a set of modes with mass depending on the value of θ , quantum effects will necessarily translate this into a θ -dependent potential energy. This type of calculation is often referred to as a Coleman-Weinberg potential [109]: we think of $\theta(x)$ as a classical background in which the various KK modes are propagating, and simply sum up their vacuum bubbles in this background. When we integrate out the KK modes of a given 5d field, we have to sum over all n . This sum should give us a periodic function of θ , because of the monodromy phenomenon. One mathematical trick for making this periodicity manifest is Poisson resummation. This is a frequently useful trick, so it is worth your time to become familiar with it by working through some exercises.

Exercise (Poisson summation formula): This formula asserts that the sum of a (sufficiently well-behaved) function at integer values is the same as the sum of its Fourier transform at integer values, i.e.,

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{k=-\infty}^{\infty} \hat{f}(k), \quad \text{where} \quad \hat{f}(k) \equiv \int_{-\infty}^{\infty} f(x) e^{-2\pi i k x} dx. \quad (190)$$

To see why this is true, notice that $F(x) \equiv \sum_{n=-\infty}^{\infty} f(x+n)$ is a periodic function, so it can be expanded in a Fourier series. The usual formula for the coefficients in a Fourier series leads almost directly to the desired identity; fill in the details.

Exercise: Use the Poisson summation formula and the general formula for a Gaussian integral to show that

$$\sum_{n=-\infty}^{\infty} e^{-An^2 - Bn - C} = \sqrt{\frac{\pi}{A}} \sum_{k=-\infty}^{\infty} e^{-\tilde{A}k^2 - \tilde{B}k - \tilde{C}}, \quad (191)$$

where the coefficients are related as

$$\tilde{A} = \frac{\pi^2}{A}, \quad \tilde{B} = -\frac{i\pi B}{A}, \quad \tilde{C} = C - \frac{B^2}{4A}. \quad (192)$$

In our context, the Poisson summation trick converts a sum over KK mode number n to a sum over *winding* number $w \in \mathbb{Z}$. There is a particularly nice semiclassical understanding of this in the limit when the 5d mass of a charged field is large compared to the KK scale, $2\pi m_{5d}R \gg 1$. Instead of thinking of a path integral over the charged field, we can imagine the 5d particles as heavy objects with a 1-dimensional worldline γ tracing their path through spacetime, and we can sum over these worldlines in the path integral. Attached to the worldline is an action for the particle, which for a Euclidean (spacelike) worldline is

$$S(\gamma) = \int_{\gamma} m_{5d} d\tau + iq \int_{\gamma} A, \quad (193)$$

where τ is the proper length along the worldline. The sum over worldlines includes a sum over topological sectors where γ wraps w times around the 5d circle. In particular, there are saddle-point contributions to the path integral for each such topological sector, and we can estimate that the sum of contributions from winding numbers $\pm w$ scales as

$$e^{-S(\gamma)} + e^{-S(-\gamma)} \sim e^{-2\pi w m_{5d}R} (e^{-iqw \int_{S^1} A} + \text{c.c.}) \sim e^{-2\pi w m_{5d}R} \cos(wq\theta). \quad (194)$$

This heuristic expectation is correct, and in fact the sums can be done analytically (see, e.g., [107, 110–112]).

We have learned that a periodic potential for θ is generated when we integrate out KK modes of 5d charged particles, and this potential is *exponentially small* when the extra dimensions are large. The intuition behind this exponential is that any effect that can generate an axion potential must somehow know the correct periodicity of θ ; as we saw in (184), the period of the canonically normalized 4d axion is sensitive to the size of the compactification, $f^2 \propto 1/R$. Thus, effects generating an axion potential are somehow *non-local* in the extra dimensions. This makes sense, as local effects would exist already in 5d, but we don't expect a 5d gauge field to become massive merely by coupling to charged particles. In the semiclassical worldline picture, this nonlocality is very explicit: the worldline of the charged particle must stretch all the way around the cycle we integrated over to get the axion. Propagation of massive particles over long distances always comes at a cost that is exponentially small in mass times distance, and this is precisely what we see. Such wrapped worldlines are sometimes referred to as “Euclidean worldline instantons,” and we can think of the exponent $2\pi w m R$ as the action of such an instanton.

From (194), we see that the axion quality problem can be solved by an extra-dimensional axion provided that all the 5d charged particles are sufficiently heavy enough compared to the compactification radius. We are now comparing the exponentially small QCD scale to the exponentially small worldline instanton effect, rather than to a power-law suppression as in the KSVZ and DFSZ scenarios. We can summarize this:

Extra-dimensional axion scenarios effectively take the log of the axion quality problem, making it much milder than in conventional 4d scenarios.

14.3 Brief comments on higher dimensions

Although we have focused on a 5d example, the principles apply to a much larger class of theories. We consider a p -form gauge field C_p , which can be integrated over a p -dimensional closed manifold $\Sigma^{(p)}$ without boundary (a “cycle”) in the extra dimensions:

$$\theta = \int_{\Sigma^{(p)}} C_p. \quad (195)$$

More precisely, we take an ansatz $C_p(x, y) = \theta(x)\omega(y)$, where $\omega(y)$ is a harmonic p -form in the extra dimensions. That is, ω is a p -form living in the extra dimensional manifold Y which is both closed ($d\omega = 0$) and co-closed ($d\star_Y\omega = 0$). This ansatz leads to a massless 4d field θ . It is periodic because of generalized “winding” gauge transformations of C_p around the cycle $\Sigma^{(p)}$. The axion coupling to gluons arises from a Chern-Simons coupling $\frac{N}{8\pi^2} \int_{M \times \Sigma^{(p)}} C_p \wedge \text{tr}(G \wedge G)$. Finally, we again obtain exponentially small contributions to the axion potential from objects charged under C_p . The difference is that this object now has a p -dimensional worldvolume. Such an object is conventionally called a $(p-1)$ -brane, because it has $(p-1)$ spatial dimensions and one time dimension. However, the axion potential arises from a *Euclidean* brane where all p worldvolume dimensions are spatial and are wrapped on the cycle $\Sigma^{(p)}$. In this case, the semiclassical sum over winding is much easier to understand than the sum over Kaluza-Klein modes. It tells us that the axion potential is proportional to a factor of $\exp(-\mathcal{T} \text{Vol}(\Sigma^{(p)}))$, where \mathcal{T} is the brane’s tension. Again, the quality problem is potentially solved when the extra dimensions are large compared to the tension scale of the branes. As in (187), the smallness of the Standard Model gauge couplings in the UV can be explained if the cycle $\Sigma^{(p)}$ has large volume compared to fundamental scales, which makes it very plausible that the Euclidean brane instanton effects are small.

Importantly, we can have more than p extra dimensions in total, as long as there is a p -dimensional cycle within the extra dimensions. For example, in Type IIB string theory, there are six extra dimensions but one might consider an axion field arising from C_4 . The Standard Model gluons would then live not in the full ten dimensions, but on an 8-dimensional submanifold $M \times \Sigma^{(4)}$. More precisely, they would live on a localized object wrapping that submanifold: a stack of D7-branes. More generally, axion-like fields are ubiquitous in string theory constructions of 4d gauge theories that are at least vaguely Standard Model-like in the sense that they contain gauge fields and chiral matter. Such axions are widely studied; see, for instance, [89, 103–105, 113–116].

15. Perspective: axions as gauge fields

Extra-dimensional axions are modes of gauge fields in the extra dimensions, but there is a more general sense in which *any* axion with couplings of the form (155) can be thought of as a type of gauge field. First, the axion field itself, like a gauge field A , is not single-valued; it has a gauge redundancy, $\theta \cong \theta + 2\pi$, and “Wilson point operators” like $\exp(i\theta)$ are well-defined in much the

same way that Wilson loop operators are for ordinary gauge fields. Furthermore, the axion has an equation of motion

$$d(f^2 \star d\theta) = \frac{N}{8\pi^2} \text{tr}(G \wedge G). \quad (196)$$

If we integrate both sides of this equation over all of (a closed) spacetime, we learn that

$$0 = N \cdot N_{\text{inst}}, \quad (197)$$

where N_{inst} is the total instanton number. This is a type of ‘‘Gauss’s law constraint’’: the axion equation of motion enforces that the net instanton number on spacetime is zero. The logic is completely parallel to the usual derivation of Gauss’s law for electric charge on a closed *space*, derived from Maxwell’s equation (59). In general, U(1) gauge theories with p -form gauge fields lead to Gauss’s law constraints on a charge that is evaluated over a $(d - p)$ -dimensional slice of spacetime. From this perspective, it is reasonable to say that an axion θ is a 0-form gauge field, which serves to gauge the instanton number charge $\int \frac{1}{8\pi^2} \text{tr}(G \wedge G)$.

The axion quality problem is related to the possibility of additional terms on the right-hand side of (196). Such terms would tend to spoil the interpretation of the axion as a gauge field. On the other hand, in Maxwell’s equations we have various currents on the right-hand side (the electron current, the muon current, and so on). These currents are not independently conserved, due to processes that convert one kind of charge into another (e.g., a muon decay transfers the electric charge to the electron). In models where the axion is a higher-dimensional gauge field, there may be additional terms on the right-hand side corresponding to other objects carrying gauge charge (e.g., Euclidean D-brane instantons), but the objects carrying such charges can be continuously deformed into gauge theory instantons [117, 118], so there is ultimately only a single conserved charge. Thus, such models can maintain the interpretation of the axion as a gauge field for some generalization of instanton number, whereas a 4d model with generic Peccei-Quinn violating operators may not admit such an interpretation.

In higher dimensions, instanton number becomes a perfectly ordinary symmetry charge. For instance, for a 5d gauge theory, $\text{tr}(G \wedge G)$ is still a 4-form, so one can integrate it over a spatial slice. Instantons are now charged particles: there is a BPST instanton solution in 5d which is just the 4d solution, taken to be independent of time, and interpreted as a kind of solitonic particle with a worldline extended through time, as we mentioned in §8.3. Furthermore, in higher dimensions it becomes a well-defined question to ask if $d[\text{tr}(G \wedge G)] = 0$, i.e., if instanton number is conserved. (In 4d, it is trivially true that d of *any* 4-form is zero, so this is not a very meaningful question.) In higher dimensions, the answer is yes:

$$d[\text{tr}(G \wedge G)] = 2 \text{tr}(dG \wedge G) = 2 \text{tr}(DG \wedge G) = 0, \quad (198)$$

where DG is the *covariant* derivative and the final step follows from the nonabelian Bianchi identity $DG = 0$. This computation plays a role in what is known as ‘‘Chern-Weil theory,’’ which is the mathematical framework that relates abstract topological invariants called characteristic classes with concrete formulas built out of differential forms, like $\int \frac{1}{8\pi^2} \text{tr}(G \wedge G)$. We learn from (198) that instanton number *is* a U(1) global symmetry charge in 5d. My collaborators and I gave such symmetries, with conserved currents built out of gauge field strengths, the name *Chern-Weil*

global symmetries [61].¹⁵ This symmetry can be gauged by a Chern-Simons coupling of the form $C \wedge \text{tr}(G \wedge G)$, where C is an ordinary $U(1)$ gauge field. Such Chern-Simons couplings are ubiquitous in known quantum gravity theories, likely for fundamental reasons [61, 125]. The 4d analogue of C is an axion, and as we saw in §14, one can obtain such 4d theories by dimensional reduction of the higher-dimensional theory.

As we will discuss more extensively in subsequent parts of these lecture notes, quantum gravity theories do not admit global symmetries. There are reasons to believe that this applies to even generalized symmetries like instanton number in four dimensions. However, quantum gravity is perfectly consistent with gauge symmetries. This gives the axion a reason for being: it gauges instanton number, and thus eliminates a would-be generalized global symmetry that is incompatible with quantum gravity. I believe that this is the underlying explanation for why axion fields are so ubiquitous in string theory constructions of gauge theories like the Standard Model. (Though this is not a complete argument: one should ask when it is possible for the symmetry to simply be broken, rather than gauged.)

Summarizing, there are multiple reasons why I find extra-dimensional axions more compelling than 4d axion models:

Extra-dimensional axion scenarios have a strong phenomenological motivation (solving the axion quality problem) and a strong motivation from deeper principles (gauging instanton number). They also appear ubiquitously in string theory compactifications.

16. Axions in cosmology

Apart from the Strong CP problem and other theoretical motivations, one reason that axions have attracted a great deal of attention is that they are a very natural dark matter candidate. A light scalar field, in the early universe, can be frozen at a value away from the minimum of its potential (“misaligned”). When the Hubble expansion rate drops below its mass, it can begin to coherently oscillate. A coherently oscillating scalar field is essentially the same thing as a collection of massive particles at rest [126]. This is the misalignment mechanism for dark matter production, and it was appreciated at an early stage that it could give rise to QCD axion dark matter [127–129].

Here, I want to very briefly mention a couple of important issues in axion cosmology. For a more complete discussion and references to the literature, see the review article [17].

The cosmology of the axion is conventionally divided into the “pre-inflationary scenario” and the “post-inflationary scenario,” depending on whether the Peccei-Quinn phase transition occurs before (or during) inflation, or after inflation. One point that I would emphasize is that in my preferred axion models, those where the axion arises from a higher-dimensional gauge field as in §14, there is no Peccei-Quinn phase transition. The axion is not a 4d pseudo-Nambu-Goldstone boson, it is intrinsically a compact scalar, and there is no point in field space at which the symmetry is restored. As a result, these models are expected to have a cosmology of the “pre-inflationary axion” type, although the terminology is a bit misleading.

¹⁵Examples were previously discussed in various contexts, e.g., [119–124].

The pre-inflationary and post-inflationary axion scenarios each suffer from potential cosmological problems, but they are different problems. In the post-inflationary axion scenario, during the Peccei-Quinn phase transition the complex scalar ϕ rolls off its potential in different directions in different parts of the universe. Because the initial value of θ is randomized over the universe, one obtains a sharp prediction for the amount of axion dark matter for a given decay constant (and a given assumption about the thermal expansion history of the universe). The scrambling of initial values of θ at the PQ phase transition leads to the formation of topological defects called axion strings, that is, dynamical objects around which the field θ winds from 0 to 2π . At later times, during the QCD phase transition, the axion acquires a periodic potential and domain walls can form separating different minima. In this context, the integer N in (155) is very important. Across a domain wall, the value of θ changes by the range from one minimum of the potential to the next, which in this case is $2\pi/|N|$. If $|N| = 1$, a domain wall can end on an axion string. The axion strings that formed at the PQ phase transition fill the universe, domain walls tend to form attached to strings, and the network of strings and walls destroys itself. However, if $|N| > 1$, then one has to put $|N|$ domain walls together to achieve a total winding of 2π . A network with multiple domain walls ending on a single string cannot tear itself apart. The domain walls overwhelm the energy density of the universe, and we have a cosmological disaster. As a result, in the post-inflationary axion scenario, a major challenge is to achieve a model where $|N| = 1$.

In the pre-inflationary scenario, on the other hand, the axion is a dynamical mode during inflation and the radial mode is effectively heavy during inflation. As a result, at the end of inflation, it is a reasonable approximation to assume that θ has the same value everywhere in the universe. No domain walls form, because the field is homogeneous. In this case, there are two potential cosmological problems. The first is that, if the initial value of θ is too far from its minimum, then the theory could predict too much dark matter. However, unlike in the post-inflationary case, we can always decrease the predicted amount of dark matter simply by assuming that θ begins close to its minimum (i.e., at the cost of some tuning). A more difficult problem to avoid is the axion isocurvature problem: the axion field fluctuates, during inflation, by an amount $\delta\theta = \frac{1}{f}\delta\hat{\theta} \sim \frac{H}{2\pi f}$. These fluctuations imply that the energy density stored in the axion field has small inhomogeneities across the universe, in a way that is decoupled from the adiabatic density perturbations sourced by the inflaton. Observations place powerful constraints on such isocurvature perturbations in the dark matter density. In the simplest cosmology, this implies that the Hubble scale during inflation must be very low relative to f . Non-standard cosmologies can ameliorate this problem.

There is a large literature on axion cosmology, and saying more would take us outside the scope of these lectures. What I would like to emphasize, which is *not* commonly said in the literature, is that the extra-dimensional axion models that are most compelling from the theoretical viewpoint are necessarily of the pre-inflationary type. In this case, we have no axion domain wall problem and the value of N is not a concern. However, the axion isocurvature problem can be rather severe in such models, and may point to a need for a non-minimal cosmology.

Part Four: No Global Symmetries in Quantum Gravity

17. Proton stability: symmetry, or not?

Global symmetries are very useful tools for characterizing quantum field theories. However, the modern view is that we do not expect global symmetries to ever truly exist in theories that describe nature. We often find very good *approximate* global symmetries, but they are not exact. Below, I will review some arguments for why we expect that global symmetries do not exist in quantum theories that include gravity. First, though, it is worth recalling that particle physics as we know it does not require us to invoke any fundamental global symmetries. I will illustrate this point by discussing how a theorist seventy years ago might have thought about the stability of the proton, and contrast this with the modern perspective that emerged after the Standard Model was established.

Imagine that you were a physicist in the 1950s. A number of particles were known. The positron had been discovered by Carl Anderson in 1932. In 1930, Pauli postulated the existence of the neutrino in a desperate attempt to patch up energy conservation in beta decay of nuclei.¹⁶ The neutron was discovered by Chadwick in 1932 and Fermi explained how its decay, $n \rightarrow p e^- \bar{\nu}$, could be accounted for with a four-fermion interaction Lagrangian, of the schematic form¹⁷

$$\mathcal{L}_\beta \sim G_F \bar{n} p e^- \bar{\nu} + \text{h.c.}, \quad (199)$$

where the Fermi constant G_F has mass dimension -2 . The neutral pion was discovered in 1950 (a few years after the charged pion), having been predicted by Yukawa as a mediator of the strong force in 1935.

All of this sets the stage for a basic question: why is the proton stable, when the neutron is unstable? The neutron decays in a way that conserves electric charge and angular momentum. So could the proton, via the decay $p \rightarrow e^+ \pi^0$. We could even write down an interaction term in our Lagrangian that would allow it,

$$\mathcal{L}_{\text{dec}} = y_p p e^- \pi^0 + \text{h.c.} \quad (200)$$

Observations tell us that the proton is extremely stable. There are enormous numbers of protons in our vicinity and we don't observe them decaying, so the proton lifetime τ_p (if indeed it decays at all) must be much longer than the age of the universe. In fact, experiments now tell us that $\tau_p \gtrsim 10^{34}$ yrs [130]. A back of the envelope estimate tells us that this requires $y_p \lesssim 10^{-32}$. This is an extraordinarily tiny number! A physicist in the 1950s would not have had such a stringent bound on the proton lifetime, but they would still have concluded that y_p must be many orders of magnitude less than one. Why is it so small?

A clever theorist then might have observed¹⁸ that a possible answer would lie in *global symmetries*. Fermi's beta decay Lagrangian (199) is compatible with two global symmetries, namely

¹⁶In fact, he called it the “neutron”; it was only after the particle we now know as the neutron was later discovered that Fermi used the name “neutrino” for Pauli's lighter neutral particle, as a joke. Of course, the name stuck, and now the Italian diminutive “-ino” is ubiquitous in particle theory.

¹⁷“Schematic” because I'm omitting the projection onto left-handed fermions, which historically came later.

¹⁸And presumably did; a full historical review is outside my scope.

“baryon number” and “lepton number”:

$$U(1)_B : p \mapsto e^{i\alpha} p, \quad n \mapsto e^{i\alpha} n;$$

$$U(1)_L : e^- \mapsto e^{i\beta} e^-, \quad \nu \mapsto e^{i\beta} \nu.$$

On the other hand, the proton decay Lagrangian (200) explicitly violates these symmetries (though it does preserve the subgroup with $\alpha = -\beta$, known as $U(1)_{B-L}$). Thus, one could explain the lack of observation of proton decay by postulating that baryon number is an exact global symmetry of the universe. In that case, $y_p = 0$, and there is no longer a puzzle. This understanding is depicted in Fig. 11.

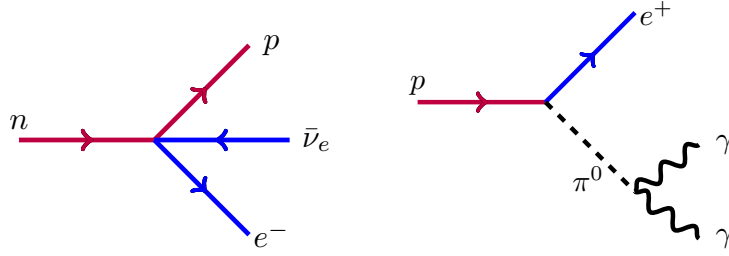


Figure 11: How a 1950s physicist might have thought about (left) neutron decay, preserving baryon number symmetry (purple lines) and lepton number symmetry (blue lines); (right) proton decay, violating both baryon and lepton number symmetries, and hence forbidden.

This was a completely reasonable expectation several decades ago, but it turned out to be the wrong way to think about proton decay. Instead, by the mid-1970s it was established that protons were in fact *composite objects*, built out of three quarks. In the fundamental theory, the operator that destroys a proton really looks like uud (with appropriately antisymmetrized $SU(3)_C$ indices). This turns out not to fundamentally change the structure of the beta decay Lagrangian, because beta decay involves a single down quark turning into an up quark, i.e., it has the schematic form $G_F \bar{d} u e^- \bar{\nu}$ (generated by integrating out a W boson, and with only the left-handed chiral fermions in the interaction). Because the beta decay interaction is a dimension six operator in effective theories both above and below the QCD scale, there is no substantial change in how we understand the parametric prediction for the neutron lifetime. However, the compositeness of the proton *does* fundamentally change the way we think about proton decay. To obtain a description of proton decay in terms of quarks, the interaction (200) should be replaced by something of the form:

$$\mathcal{L}_{\text{dec}} = \frac{1}{M_{\text{dec}}^2} u u d e^- + \text{h.c.} \quad (201)$$

Now two quarks in the proton can scatter into a positron and an antiquark, with the outgoing antiquark and the remaining initial quark rearranging themselves into a pion. This process is depicted in Fig. 12. We can still use our old interaction Lagrangian (200) as an effective theory for the proton decay process; however, by dimensional analysis, the matching between its parameter y_p and the parameter M_{dec} in (201) is schematically

$$y_p \sim \frac{\Lambda_{\text{QCD}}^2}{M_{\text{dec}}^2} \sim 10^{-32} \left(\frac{10^{15} \text{ GeV}}{M_{\text{dec}}} \right)^2. \quad (202)$$

This suggests a completely different possible explanation for why the proton is so stable: there is no symmetry at all, but the symmetry-violating interactions are irrelevant at low energies. Symmetry violation could be mediated by very heavy particles, with mass larger than 10^{15} GeV [131]. This is relatively close to the Planck scale, so perhaps this is a fundamental process in a hypothetical theory incorporating gravity. At such high energies, the amplitude for baryon number violating processes could become large: the symmetry is *badly* broken. The insight that was missing for our hypothetical 1950s theorist was that the proton and pion were not elementary fields; compositeness can make the tiny coupling $y_p \sim 10^{-32}$ much less concerning. Or rather, to state the case more carefully, it translates the question of why y_p is small into the question of why Λ_{QCD} is small—a question that was answered by the discovery of asymptotic freedom and dimensional transmutation, which produces an exponentially small Λ_{QCD} from a mildly small coupling g_s at high energies.

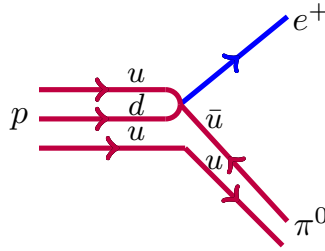


Figure 12: Proton decay after the discovery of quarks: the process must involve a four-fermion interaction, potentially suppressed by a very large mass scale. There is no need to postulate an exact symmetry to explain why we do not observe the process.

What the discovery of the full structure of the Standard Model made clear was that baryon number could be an *accidental* symmetry. We say that a theory has an accidental symmetry when the lowest-dimension gauge-invariant operator that violates a symmetry is irrelevant. This is the case for baryon number in the Standard Model, which is first violated by dimension six operators like $uude$ and $qqq\ell$ [132]. It is also the case for lepton number in the Standard Model, though this is violated already at dimension five, by operators of the form $(h\ell)^2$, which correspond to Majorana neutrino masses at low energies. We do not know that neutrino masses take this form, but it is very plausible that they do; again, they are small parameters in the low-energy theory, but the perspective of accidental symmetries can render their smallness much less mysterious. The modern perspective on global symmetries is that we expect them to arise in the way that baryon and lepton number symmetries arise in the Standard Model: not as exact properties of the UV theory, but as emergent properties of the IR theory, enforced by gauge invariance. Indeed, we generally expect that they are not even approximate symmetries in the UV. Good approximate symmetries can emerge naturally from a complete lack of symmetry, simply by dimensional analysis, when the symmetry is accidental.

There is more to the story of baryon number violation, but the discussion above captures the key insight we will build on below. The additional important physics related to baryon number is that ABJ anomalies lead to nonperturbative violation of both baryon and lepton number in the Standard Model (but preserve $U(1)_{B-L}$), even *without* adding any irrelevant operators. Interestingly, the anomaly leads to processes that violate baryon number only by multiples of three units. Because of this, it is still a consistent possibility that our universe has a *discrete* gauge symmetry that stabilizes

the proton, but would allow three protons to collectively annihilate into leptons. Discrete gauge symmetry is an interesting topic that I do not have time to do justice to in these lectures.

In the real world, we sometimes encounter approximate symmetries that are *not* accidental. A good example is isospin, the symmetry of the strong interactions that rotates up and down quarks into each other. This is explicitly (but weakly) broken by electromagnetism, because the up and down quarks have different electric charge. It is also explicitly broken by the unequal up and down quark masses. This is said to be a *soft* breaking: a small breaking by relevant operators, which does not introduce divergent corrections to quantities in the IR. The fact that such a non-accidental, approximate global symmetry arises in the real world tells us that such symmetries might also arise in physics beyond the Standard Model. They can't be forbidden by any general principle. Still, we would like to understand them better. In the full Standard Model, the fact that the up and down quark masses are not arbitrarily big is partially explained by the fact that they arise from the Higgs vev that breaks electroweak symmetry (though why *that* is small is a great mystery). They are further suppressed by small Yukawa couplings. We expect the smallness of these couplings to be explained by some deeper principle: perhaps by the higgsing of some new gauge symmetry (like a discrete flavor symmetry) in the UV, or perhaps by localization of the left- and right-handed fermions at different places in extra dimensions. This example highlights that approximate but not accidental symmetries are allowed, but are also unsatisfying, and may be important clues to deeper structure in a theory.

Many model builders use *technical naturalness* as a guiding principle: parameters in a Lagrangian are allowed to be very small if a symmetry is restored when they are set to zero. Technical naturalness is important, as it tells us that quantum corrections are under control. However, it is far from sufficient. If the explanation of a small parameter invokes a very good approximate global symmetry, then the approximate global symmetry itself cries out for a deeper explanation. Technical naturalness can be a useful guidepost in searching for a theory, but it is far from the end of the story. Ultimately, we would like to have a satisfying explanation of small parameters in terms of $O(1)$ inputs or deeper principles.

18. Black holes: mini-review

The most convincing arguments about universal properties of quantum gravity rely on semi-classical properties of black holes, since these results can be derived simply by combining general relativity and quantum field theory, rather than relying on a specific theory of quantum gravity like string theory. A complete review of black holes in effective field theory would go far outside the scope of these lecture notes, so instead we will take a brief tour of key aspects of black hole physics. For a more thorough introduction to black hole physics, I recommend [133].

18.1 Classical black hole solutions

The most prominent feature of a black hole is its *horizon*, the boundary through which signals from inside the black hole cannot escape. The horizon is more important than the *singularity*, which lies deep in the interior of the black hole. One might wonder if the singularity, where the curvature becomes infinite, is somehow resolved by unknown short-distance physics. On the other hand, for a large enough black hole, the horizon is a region of small curvature, within the domain where we

expect effective field theory to be reliable. Thus, the interesting properties of black holes are to a large extent associated with the horizon.

The simplest black hole solution in general relativity is the neutral, spherically symmetric, static *Schwarzschild* black hole. The Schwarzschild metric takes the form

$$ds^2 = -f(r) dt^2 + \frac{1}{f(r)} dr^2 + r^2 d\Omega^2, \quad (203)$$

where for a four-dimensional black hole $d\Omega^2 \equiv d\theta^2 + \sin^2 \theta d\phi^2$ is the unit metric on the 2-sphere and

$$f(r) \equiv 1 - \frac{2GM}{r}. \quad (204)$$

Here G is Newton's constant and M is the black hole mass. In units where $\hbar = c = 1$, Newton's constant is the same as $(8\pi M_{\text{Pl}}^2)^{-1}$ where $M_{\text{Pl}} \approx 2.4 \times 10^{18}$ GeV is the Planck mass. The coordinates (θ, ϕ) parametrize the angular directions, forming an S^2 that shrinks to zero size at the singularity $r = 0$. This is not just a coordinate singularity. For example, the Kretschmann scalar, a curvature invariant defined by $R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}$, blows up at $r \rightarrow 0$. The black hole *horizon* is located at $r = 2GM$, where $f(r) = 0$. At this point, the signs in front of the dt^2 and dr^2 terms in the metric flip, so that time and space exchange roles at the horizon. This corresponds to the fact that light rays inside the horizon can never propagate to the exterior. The Schwarzschild metric is an approximate description of a realistic black hole which forms by collapse and eventually evaporates; Fig. 13 shows a Penrose diagram depiction of the spacetime of such a black hole.

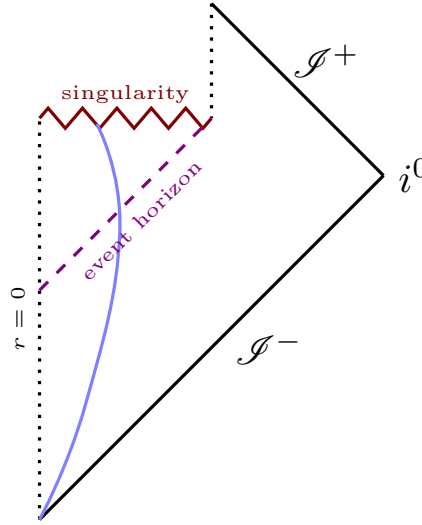


Figure 13: Penrose diagram for a black hole formed by collapse of infalling matter (blue curve) and then evaporating. Time runs vertically and (radial) space horizontally, with an implicit angular two-sphere over every point in the diagram. The dashed purple line is the event horizon, where $f(r) = 0$, which is a lightlike boundary. The red jagged line is the spacelike singularity at $r = 0$. Dotted lines are coordinate singularities at $r = 0$, not physical singularities.

A black hole can carry electric or magnetic charge under a $U(1)$ gauge field. For the case of electric charge, this corresponds to the Reissner-Nordström solution, for which the metric has the

same form (203) where

$$f(r) \equiv 1 - \frac{2GM}{r} + \frac{G\tilde{Q}^2}{r^2}, \quad (205)$$

and there is also a U(1) gauge field

$$A_0 = \frac{\tilde{Q}}{r}. \quad (206)$$

(Sometimes instead a gauge is chosen where a constant is subtracted in order to take $A_0 = 0$ at a special radius $r = r_+$, to be defined shortly.) The charge \tilde{Q} is normalized according to a commonly-used convention in GR, related to a U(1) integer charge Q with gauge coupling e as:

$$\tilde{Q}^2 = \frac{e^2}{4\pi} Q^2. \quad (207)$$

The Reissner-Nordström black hole has *two* horizons where $f(r) = 0$, the outer horizon at r_+ and the inner horizon at r_- , where

$$r_{\pm} = GM \pm \sqrt{G^2 M^2 - G\tilde{Q}^2}. \quad (208)$$

This expression does not make sense for arbitrary (M, \tilde{Q}) . In particular, for a given charge there is a *minimum mass*:

$$M^2 \geq M_{\text{ext}}(Q)^2 \equiv \frac{1}{G} \tilde{Q}^2 = 2e^2 Q^2 M_{\text{Pl}}^2. \quad (209)$$

Intuitively, this is because a charged black hole has a nonzero electric field outside the horizon, since $\partial_r A_0 \neq 0$. The electric field stores energy, which adds to the rest mass of the black hole, preventing it from being arbitrarily small. The bound (209) is known as the “extremality bound.” A charged black hole with a mass that saturates this bound is referred to as an extremal black hole. A black hole that satisfies the bound without saturating it is referred to as subextremal. An object that violates this bound is said to be superextremal. Black hole solutions, at least to the classical theory with a standard two-derivative action, are never superextremal.

In the presence of higher derivative terms in the effective action, black hole solutions are modified, and the extremality bound generally changes. (The exception is in when extremal black holes are BPS, in theories with sufficient supersymmetry.) In general, the extremal mass-to-charge ratio obtains a series of corrections that can be written in inverse powers of the charge. If we define the extremality bound as the limiting case of asymptotically large black holes, these corrections can allow a finite-mass black hole to slightly violate the bound [134].

Exercise: Find the D -dimensional generalization of the Schwarzschild and Reissner-Nordström solutions, and of the extremality bound.

18.2 Black hole thermodynamics

In the early 1970s, Bekenstein and Hawking began to study black holes semiclassically, that is, incorporating quantum effects. Important early insights into the quantum properties of black holes came from studying quantum field theory on a fixed black hole background spacetime (effectively, taking the strict $G \rightarrow 0$ limit) and by considering the Euclidean path integral for gravity at finite

temperature. There is a rich and fascinating story to tell here, but in these notes I will only summarize a few key conclusions.

Famously, black holes were found to be thermodynamic objects, with an associated entropy and temperature [135–138]. For a black hole of mass M and (integer) charge Q , we define the *extremality ratio*

$$\xi \equiv \frac{2e^2 Q^2 M_{\text{Pl}}^2}{M^2}. \quad (210)$$

In this convention, $\xi = 1$ for an extremal black hole, and $0 \leq \xi \leq 1$ for any black hole. Bekenstein found that the *entropy* of a black hole is proportional to the area of its (outer) horizon, namely

$$\begin{aligned} S &= \frac{A_+}{4G} = 8\pi^2 r_+^2 M_{\text{Pl}}^2 \\ &= \frac{M^2}{4M_{\text{Pl}}^2} \left(1 + \sqrt{1 - \xi} - \frac{1}{2}\xi \right). \end{aligned} \quad (211)$$

This entropy allows black holes to obey a Generalized Second Law, which is to say that the sum of this entropy and conventional thermodynamic entropy always increases. Hawking showed that black holes emit blackbody radiation with temperature [139]

$$\begin{aligned} T &= \frac{1}{4\pi} \left(\frac{1}{r_+} - \frac{G\tilde{Q}^2}{r_+^3} \right) \\ &= \frac{4M_{\text{Pl}}^2}{M} \frac{\sqrt{1 - \xi}}{(1 + \sqrt{1 - \xi})^2}. \end{aligned} \quad (212)$$

Hawking radiation is thermal, but for a charged black hole the electric potential at the horizon serves as *chemical potential* favoring discharge,

$$\mu = \frac{\tilde{Q}}{\sqrt{4\pi}r_+} = \frac{\sqrt{2\xi}}{1 + \sqrt{1 - \xi}} M_{\text{Pl}} \text{sgn}(Q). \quad (213)$$

This is a chemical potential in the usual thermodynamic sense that it distorts the Boltzmann factor for emitted particles, so that the probability of emitting a particle of energy E and charge q is proportional to $\exp[-(E - \mu eq)/T]$. In particular, the black hole is *more likely* to emit radiation of the same-sign charge as the black hole, so that the magnitude of its charge decreases over time. In terms of an intuitive heuristic that Hawking radiation can be thought of as the production of particle/antiparticle pairs, one of which falls into the black hole horizon and one of which escapes, what is happening is that the electric field pushes the same-sign particle away from the black hole so that it escapes, while the opposite-sign particle is attracted and falls in, decreasing the charge of the black hole.

Let us emphasize a few qualitative properties of these equations. The black hole entropy is proportional to the horizon area, and remains nonzero for an extremal black hole. The black hole *temperature*, far from extremality, is of order the inverse radius of the horizon. In particular, large black holes are very cold and radiate slowly. However, at extremality, $T \rightarrow 0$. Thus, exactly extremal black holes do not emit Hawking radiation. However, if charged particles are present in

the theory, the black hole can discharge by Schwinger pair production of particles in the electric field outside the horizon [140]. For a clear recent discussion that treats Hawking radiation and Schwinger pair production together, see [141].

Exercise: Find the D -dimensional generalization of the formulas for black hole entropy and temperature.

Exercise: Calculate the lifetime of a four-dimensional Schwarzschild black hole of mass M to evaporate via Hawking radiation. (Hint: neglecting “greybody factors” that describe how particles escape from the near-horizon region to the asymptotic region, you can assume that radiation is emitted from the horizon according to the Stefan-Boltzmann law.)

19. Black holes versus continuous global symmetries

From the time that Hawking radiation was discovered, it was used to argue that black hole physics could forbid global symmetry charges [142]. The general idea is that global symmetry charge, unlike gauge symmetry charge, is invisible to semiclassical black hole physics. If you throw an electron into a black hole, you can tell that it has acquired an electric charge because you can measure the electric field outside the black hole. As discussed above, the electric charge of a black hole also distorts the spectrum of its Hawking radiation, serving as a chemical potential. Global symmetry charge is entirely different. For example, $B - L$ is a non-anomalous global symmetry of the Standard Model with right-handed neutrinos, if we do not turn on Majorana neutrino masses or other explicit symmetry violation. Thus, it is a potential global symmetry of nature. Neutrons carry $B - L$ charge but not electric charge, so you could imagine throwing a large number of neutrons into a black hole to give it $B - L$ charge. However, no measurement that you can do outside the black hole would tell you that it carries this charge. For a global symmetry, there is no electric field to measure, and there is correspondingly no chemical potential, so the Hawking radiation will not care about the black hole’s global charge. It will continue to emit just as many neutrons as antineutrons, or neutrinos as antineutrinos. This led many people to strongly suspect that black holes violate global symmetry charges: they will simply eat such charges and then forget about them. This argument is not entirely convincing, however, because one could always imagine that the black hole does remember its global charge, but in a way that is invisible to semiclassical physics like Hawking radiation. Perhaps black holes with different $B - L$ charges are simply different states in the theory, which happen to appear the same to external observers.

A modern reformulation of the argument by Banks and Seiberg presents a sharp contradiction between this possibility and entropy bounds [143]. It is believed that for any bounded region of spacetime, there is a maximum possible entropy in quantum gravity, i.e., a finite number of possible different microstates that can exist in this region, corresponding to the entropy of a black hole that fills the region. This idea goes back to Bekenstein [144] and was given a precise covariant formulation by Bousso [145]. Banks and Seiberg pointed out that the existence of any continuous

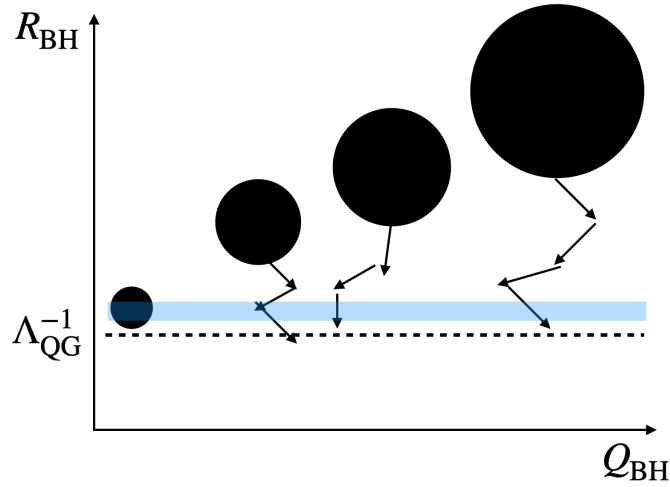


Figure 14: The Banks-Seiberg argument against a global $U(1)$ symmetry. By creating large black holes of arbitrarily large global charge Q_{BH} , then letting them evaporate down to the radius Λ_{QG}^{-1} where we stop trusting Hawking’s semiclassical calculation, we can construct infinitely many different states that all fit in a region of size slightly larger than Λ_{QG}^{-1} (blue band), in contradiction to any entropy bound.

global symmetry—let’s discuss $U(1)$, for simplicity, but the key point is that a continuous symmetry has infinitely many representations—would actually lead to an *infinite* entropy, and hence be in conflict with any such entropy bound. The argument is simple. We can construct black holes with arbitrarily large global symmetry charge Q_{BH} , just by throwing many particles with global charge (e.g., neutrons, in our hypothetical example above) into a black hole. After constructing a black hole of large global charge, simply wait. Hawking’s calculation tells us that, as long as we trust semiclassical effective field theory, the black hole will shrink down to smaller size. Its global symmetry charge could go up or down, in a random walk, because the black hole is just as likely to emit particles with either sign of this charge. Importantly, it is not preferentially driven to zero, because there is no chemical potential. Thus, if we start with a large black hole of large charge and wait, we will obtain a small black hole of large charge, with radius somewhat above Λ_{QG}^{-1} , the scale at which we stop trusting Hawking’s calculation. We can carry out this process for states of charge as large as we like, and in this way we can obtain *infinitely* many different black hole states fitting in a region of fixed size—in blatant violation of the entropy bound. This is illustrated in Fig. 14. This argument relies only on the assumption that we can trust Hawking’s calculation of the evolution of the global symmetry charge as the black hole evaporates. One may doubt this; the black hole information problem famously tells us that effective field theory can get the answers to certain questions about black holes badly wrong. However, unlike in the black hole information context, we are not asking a subtle question here about entanglement among many Hawking quanta. We are simply asking about the global charge of the black hole—a one-point function of a simple operator. If effective field theory fails to answer such a simple question, there is no reason to trust that it gets anything right, about the temperature or lifetime or any other basic property of the black hole. Rather than discard the elaborate, self-reinforcing structure of black hole thermodynamics

as developed over the last fifty years, it seems more plausible that continuous global symmetries simply can't exist in theories with black holes. Let's highlight this key point:

Black hole thermodynamics is incompatible with continuous global symmetries.

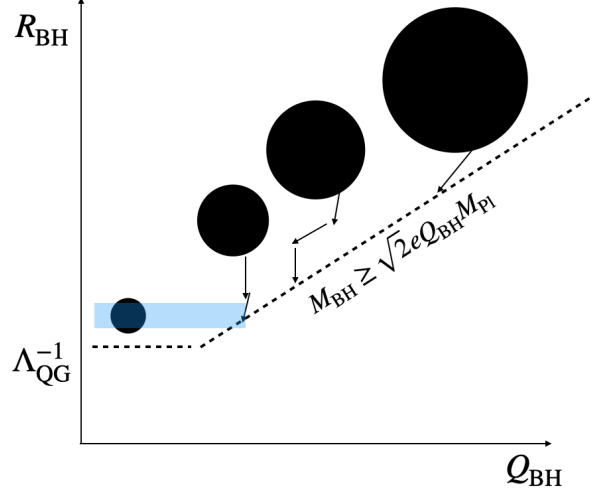


Figure 15: The Banks-Seiberg argument in the case of gauge charge. The extremality bound now prohibits the existence of small black holes of large charge. As a result, in the $U(1)$ case, there is only a finite entropy in a region of a given size. However, for a noncompact gauge group like \mathbb{R} , charge becomes a continuous quantity and there is still infinite entropy.

The Banks-Seiberg argument does not rule out theories with $U(1)$ *gauge* symmetry. In this case, because the gauge charge is associated with an electric field outside the black hole, an evaporating black hole preferentially sheds its charge as it evaporates. Eventually it either fully discharges or it hits the extremality bound (209), where it could get stuck if there are no sufficiently light particles for it to decay to by Schwinger pair production. In any case, the extremality bound imposes an upper limit on the charge of a black hole that can fit in a region of a given size, as indicated by the truncated blue band in Fig. 15. The quantization of $U(1)$ charge implies that there are only finitely many states in this band, so there is no immediate contradiction with the existence of an entropy bound.¹⁹ However, if the gauge group is \mathbb{R} , rather than $U(1)$, then there *are* infinitely many different possible charges—a continuum of them, in fact, which we could densely fill by starting with any two particles with mutually irrational charges, e.g., 1 and $\sqrt{2}$. In this way, the argument rules out not only continuous global symmetries in quantum gravity, but also noncompact, continuous *gauge* symmetries.

¹⁹One might try to argue for a bound by comparing the finite number of states in this region with a *quantitative* entropy bound. The number of states in the region is proportional to $Q_{\max} \sim 1/e$, shifting the entropy by $\log |e|$, thus requiring precise quantitative control over corrections to the leading black hole entropy calculation and the derivation of the bound. Despite some efforts by various groups, to the best of my knowledge no convincing derivation of the Weak Gravity Conjecture (or similar result) has been given along these lines.

Black hole thermodynamics is incompatible with noncompact, continuous gauge symmetries. In particular, abelian gauge groups in quantum gravity are always $U(1)$, never \mathbb{R} .

20. Generalized global symmetries

20.1 p -form global symmetries

We discussed ordinary global symmetries in §3. They acted on local operators living at a point (0-dimensional) or charged particles with 1-dimensional worldlines. They were associated with a charge Q measured on a spatial slice. More generally, the charge can be measured on any codimension-1 slice Σ through spacetime, e.g., a 3-manifold in 4d spacetime. (The “codimension” of a submanifold is just the number of dimensions *transverse* to the submanifold, i.e., the codimension of a p -manifold in d -dimensional spacetime is $d - p$.) An ordinary global symmetry comes with a family of symmetry operators $U(\Sigma, g)$ associated to a given symmetry group element g and a closed 3-manifold Σ . For a $U(1)$ symmetry, there is a 3-form current J which is conserved, $dJ = 0$, and the charge is simply $\int_{\Sigma} J \in \mathbb{Z}$. Such a symmetry can be gauged by a 1-form gauge field A with a coupling $A \wedge J$.

There are many familiar structures in quantum field theory that are almost completely parallel except that the dimensions are different. For example: in a $U(1)$ gauge theory without magnetic monopoles, we have $dF = 0$ (the Bianchi identity), and we also have a related integrated “charge” over a closed 2-dimensional manifold Σ , which is just the magnetic flux $\int_{\Sigma} \frac{1}{2\pi} F \in \mathbb{Z}$. This is measured by a topological operator that depends on an angle α and a choice of 2-manifold Σ , $U(\Sigma, \alpha) \equiv \exp\left(i\alpha \int_{\Sigma} F\right)$. Notice that this manifold always has two dimensions, no matter what d is, because F is always a 2-form. We say that such an operator generates a $(d - 3)$ -form global symmetry.

Similarly, in a $U(1)$ gauge theory without electrically charged particles, Maxwell’s equations tell us that $d\star F = 0$. This is related to the conservation of electric flux, which we measure by integrating over a closed $(d - 2)$ -dimensional manifold, $\int \frac{1}{e^2} \star F \in \mathbb{Z}$. Again this is associated with a topological operator. It generates a 1-form global symmetry.

Thus, we can have generalized conservation laws of the form $dJ_p = 0$, and associated topological operators $U(\Sigma_p, \alpha) = \exp(i\alpha \int_{\Sigma_p} J_p)$. In general, we call this a $(d - p - 1)$ -form global symmetry. We could gauge such a symmetry by adding a $(d - p)$ -form gauge field with a coupling $C_{d-p} \wedge J_p$, which respects a gauge symmetry of the form $C_{d-p} \mapsto C_{d-p} + d\lambda_{d-p-1}$. To complete the analogy, we should identify objects that are charged under this symmetry. In the gauge case, we should be able to integrate C_{d-p} over the worldvolume of such an object, just as in the usual case we integrate A over a particle’s worldline. Thus, we expect that a current J_p is related to an object with $(d - p - 1)$ spatial dimensions and one time dimension, which can be created by an operator extended along $(d - p - 1)$ spatial dimensions.

For example, in the case of $U(1)$ gauge theory with no electrically charged particles, we had a conserved current $\frac{1}{e^2} \star F$, which is a $(d - 2)$ -form. Thus we expect that the operators carrying the associated charge have dimension $d - (d - 2) - 1 = 1$. These must be line operators that have an associated electric flux. In fact, we already know about such operators: they are just the Wilson

lines that we discussed in §4.3. We refer to this symmetry as the “electric 1-form symmetry” of electromagnetism. Similarly, the symmetry with conserved current $\frac{1}{2\pi}F$ is the “magnetic $(d-3)$ -form symmetry” of electromagnetism, and the charged operators are ’t Hooft operators, which are the magnetic analogue of Wilson lines. In the case $d=4$, these are both 1-form symmetries, the Wilson and ’t Hooft operators are both lines, and in fact there is a more general set of “dyonic” lines carrying both electric and magnetic flux and (in the theory of a free $U(1)$ gauge field) related to each other by $SL(2, \mathbb{Z})$ duality transformations.

Let’s summarize: a p -form generalized global symmetry in a d -dimensional QFT is one that [1]

- Acts on operators extended along p dimensions, or equivalently dynamical objects with $(p+1)$ dimensional worldvolumes in spacetime.
- Is generated by symmetry operators $U(\Sigma, g)$ living on $(d-p-1)$ -dimensional slices Σ of spacetime. These are *topological* operators: we can deform the manifold Σ without changing any correlation functions except when Σ crosses the worldvolume of a charged operator.
- In the special case that the symmetry group is $U(1)$, the symmetry is generated by a local conserved current operator which is a $(d-p-1)$ -form obeying $dJ_{d-p-1} = 0$. (Be aware that in some of the literature, what I call J_{d-p-1} would be referred to as $\star j_{p+1}$.)
- Can be gauged by a $(p+1)$ -form gauge field. (Unless the symmetry has an ’t Hooft anomaly, which prohibits gauging.)

The most familiar, ordinary symmetries are 0-form global symmetries.

20.2 Breaking of p -form global symmetries

20.2.1 Explicit breaking

Just for convenience, I will now specialize to $d=4$. We have seen that a $U(1)$ gauge theory without any electrically or magnetically charged particles has a $U(1)_{\text{el}} \times U(1)_{\text{mag}}$ 1-form global symmetry, acting on Wilson and ’t Hooft lines. However, these symmetries should be explicitly broken if charged objects exist, because the underlying conservation laws are no longer true:

$$d\left(\frac{1}{e^2} \star F\right) = J_{\text{el}}, \quad d\left(\frac{1}{2\pi} F\right) = J_{\text{mag}}. \quad (214)$$

These equations imply that the operators measuring electric and magnetic flux are no longer topological in a theory with charged objects.

Let’s see how this works more explicitly in the case of a Wilson line $W_q(\gamma)$ that carries charge under the 1-form electric symmetry, which is generated by an operator $U(\Sigma, \alpha)$. We will take the Wilson line to be supported on the timelike curve that sits at the origin of space, and the flux to be measured on a 2-sphere $\Sigma = S^2$ around the origin at some fixed time. Then we have an OPE

$$U(S^2, \alpha) W_q(\gamma) = e^{iq\alpha} W_q(\gamma). \quad (215)$$

This is completely parallel to the case of a 0-form symmetry depicted in Fig. 3, and holds independent of what time t we locate the 2-sphere at. However, if we add a charged operator ϕ of charge q to

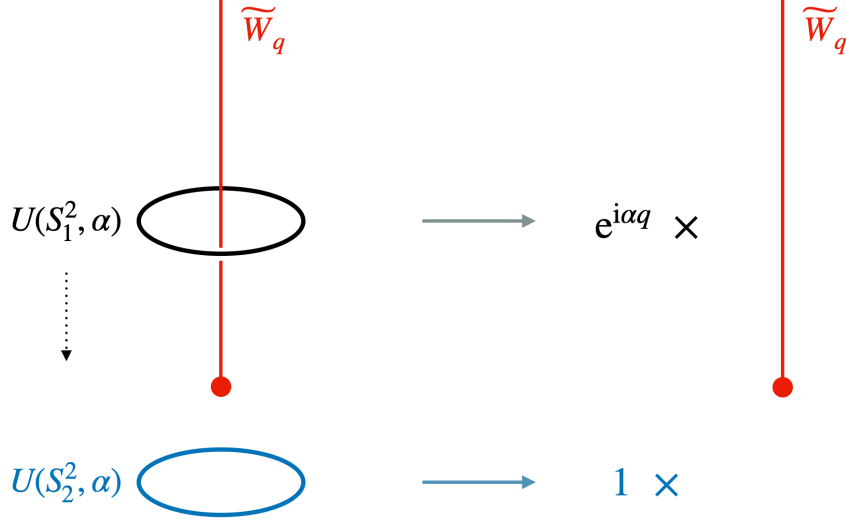


Figure 16: Endability of a Wilson line causes the 1-form electric symmetry operator $U(\Sigma, \alpha)$ to no longer be topological. Left: symmetry operator inserted on two different spheres, S_1^2 and S_2^2 . Right: corresponding OPE. The symmetry operator on S_1^2 sees an electric flux associated with charge q , and multiplies the Wilson line by a phase $e^{iq\alpha}$. However, the symmetry operator on S_2^2 does not see a flux, and acts trivially. Hence, the electric 1-form symmetry is broken.

the theory, then we can now consider a new operator which is the Wilson line ending at time $t = 0$ on a ϕ insertion, as depicted in Figure 16. That is, we consider the Wilson operator with endpoint,

$$\widetilde{W}_q \equiv \exp \left[iq \int_{t=0}^{\infty} A(t, 0) \right] \phi^*(0, 0). \quad (216)$$

Now, if we insert the symmetry operator $U(S^2, \alpha)$ at $t > 0$ (as indicated by S_1^2 in the figure), it surrounds the Wilson line and acts with a nontrivial phase $\exp(iq\alpha)$, but if we insert the symmetry operator at $t < 0$ (as indicated by S_2^2 in the figure), it surrounds nothing at all and acts trivially. Thus, the *endability* of the Wilson line on a charged operator causes the symmetry operator to *no longer be topological*. Only topological operators can be thought of as obeying conservation laws. Said differently, acting on the vacuum with the local charged operator ϕ at some time creates an electric flux at subsequent times that was not present at earlier times, explicitly violating the conservation of electric flux.

In quantum gravity, we expect all global symmetries to be broken or gauged. To fully break the electric and magnetic 1-form global symmetries of U(1) gauge theory, we need particles (or collections of particles) with all possible electric and magnetic charge to exist. We say that such a theory has a *complete spectrum*. It is a longstanding conjecture (the Completeness Hypothesis) that quantum gravity theories must have a complete spectrum [146]. Here we see that, in the case of U(1) gauge theory, the Completeness Hypothesis is implied by the absence of generalized global symmetries. It turns out that this does *not* generalize to all possible gauge groups: one can have an incomplete spectrum but no p -form generalized global symmetry [21]. However, if we broaden our notion of symmetry even more, we can revive the link between breaking of global symmetries and

completeness. A gauge theory with an incomplete spectrum always contains topological operators, which can be said to generate a “non-invertible global symmetry” [147, 148]. Such topological operators obey a more complicated fusion algebra, rather than group multiplication law as for standard symmetries.

We saw that the breaking of a 1-form symmetry was associated with the endability of an operator linked by the symmetry operator. More generally, we expect that in quantum gravity, all extended operators should either be endable, or should themselves be the boundary of a different operator. The latter is the case where an operator is not itself gauge invariant (e.g., the local operator creating an electron is the end of a Wilson loop operator, without which it is not well-defined).

20.2.2 Spontaneous breaking

A p -form global symmetry can also be spontaneously broken. For a 0-form symmetry, this happens when a charged local operator has a vacuum expectation value, $\langle O \rangle \neq 0$. For a 1-form global symmetry, the expected behaviors of the expectation value of a charged loop operator are that it scales with the exponential of minus the perimeter or the area of the loop. For area-law scaling, the expectation value of the loop operator goes rapidly to zero as the loop size increases. If it has the milder perimeter scaling, we say that the 1-form global symmetry is spontaneously broken. In this case, the operator can be redefined by adding a counterterm along the loop, so that it has a nonzero expectation value even for arbitrarily big loops [1, 149, 150].

A concrete example is the expectation value of a Wilson loop or ’t Hooft loop in QED. When we are in the phase where the photon is massless, the electric (or magnetic) 1-form symmetry is spontaneously broken. We can think of the photon as a Nambu-Goldstone boson for either spontaneously broken 1-form symmetry! For an ordinary Nambu-Goldstone boson, we expect that the spontaneously broken symmetry current can create a 1-particle Nambu-Goldstone boson state from the vacuum, $\langle \Omega | J_\mu(x) | p \rangle = p_\mu e^{ip \cdot x} f$ for some constant f . An analogous statement holds for the photon, in terms of the electric or magnetic 1-form symmetry current, e.g.,

$$\langle \Omega | F_{\mu\nu} | p, \epsilon \rangle \propto \epsilon_{[\mu} p_{\nu]} e^{ip \cdot x}. \quad (217)$$

It may seem somewhat mysterious to think of the photon as a Nambu-Goldstone boson of the 1-form symmetry, because this symmetry is explicitly broken! If there are both electrically and magnetically charged particles in the universe, *both* 1-form symmetries are explicitly broken. When we explicitly break a 0-form symmetry, we are used to the fact that the pseudo-Nambu-Goldstone boson acquires a mass. But the photon is (as far as we know) exactly massless! This is something special about the explicit breaking of higher p -form symmetries, where $p > 0$: a pseudo-Nambu-Goldstone boson remains massless. This point is discussed further in [3, 151]. (The enhanced robustness of higher p -form symmetries is closely related to the exponential improvement in the axion quality problem when the axion arises from a higher-dimensional gauge field, as discussed in §14.2.)

21. Global symmetries versus quantum gravity

In §19 we saw that black hole thermodynamics provides arguments against continuous global symmetries. A much stronger statement is believed to be true: quantum gravity theories do not have

any global symmetries. This includes ordinary (0-form) discrete global symmetries, but also p -form global symmetries (continuous or discrete) and even less familiar examples like non-invertible (or categorical) symmetries. In all of these cases, there are symmetry operators that implement the symmetry, which are *topological*. A very general expectation is that quantum gravity does not admit topological operators. Heuristically, this is because the gravitational path integral sums over all spacetimes, including those of nontrivial topology, and admits topology-changing transitions. Topology is not expected to be an invariant property in quantum gravity, and so it should not be possible to construct well-defined operators that correspond to topological invariants.

The state of the art in *proving*, from a well-defined starting point, that quantum gravity does not admit global symmetries is a holographic argument due to Harlow and Ooguri [21, 22]. This argument only applies in asymptotically AdS spacetimes, and exploits the fact that quantum gravity in such a spacetime is equivalent to a conformal field theory defined on the conformal boundary of spacetime. Reviewing the argument in detail is beyond the scope of these lectures, but I will give a very brief summary of the main idea. It relies on two key facts. The first is entanglement wedge reconstruction: an operator acting in a boundary region R of the CFT can only influence a limited part of the bulk of AdS, the corresponding entanglement wedge W_R . The second is splittability: a symmetry operator on the boundary can be factored into a product of symmetry operators associated with a collection of subregions R_i . By breaking apart a global symmetry operator $U(R, g)$ on the boundary into a product of $U(R_i, g)$ over sufficiently many small regions R_i , we can make their domain of influence in the bulk, the union of the W_{R_i} , as small as we like. This shows that operators deep in the bulk cannot carry charge under the global symmetry—a contradiction, since a global symmetry of the type we discussed in §3 acts locally via topological operators that we should be able to pull out to the boundary. This argument rules out p -form global symmetries (continuous or discrete) in d -dimensional asymptotically AdS quantum gravity, for $0 \leq p \leq d - 2$ (the case $p = d - 2$ is slightly subtle, and less rigorous than the rest).

The Harlow and Ooguri argument also establishes that the gauge group in asymptotically AdS quantum gravity must be compact, and that objects (or collections of objects) should exist in the bulk theory transforming in all representations of the gauge group (i.e., the Completeness Hypothesis).

Although this argument relied on AdS/CFT, there is a widespread expectation that the conclusions apply much more broadly to any realistic theory of quantum gravity. The caveat “realistic” is important here, as a number of theories of quantum gravity in low dimensions lack many features of higher-dimensional quantum gravity; they may have global symmetries, but they also do not have unitary black hole evaporation, for example [152].

These modern arguments supplement a number of older arguments. For example, in perturbative string theory, Banks and Dixon showed that any putative continuous global symmetry actually is a gauge symmetry [153]: given a global symmetry current in spacetime, one can construct a worldsheet vertex operator that creates a gauge field in spacetime. In AdS/CFT, a global symmetry of the boundary CFT is a gauge symmetry of the bulk [154]. Other arguments, related to wormholes, have both early incarnations [155–158] and very recent (and sharper) ones [159–162].

Gauge symmetries are common in quantum gravity, unlike global symmetries. However, we should remember that a global symmetry maps one state of the theory to another, but a gauge symmetry is a redundancy of our description, not really a symmetry at all. Let’s sum this up with two crucial slogans to remember.

There are no symmetries in quantum gravity except for gauge symmetries.
Gauge symmetries are not symmetries.

Any would-be global symmetry in quantum gravity must be either explicitly broken or gauged (a spontaneously broken symmetry is still a symmetry). Saying that global symmetries do not exist in quantum gravity is all well and good, but it's not very useful for doing real-world physics. As we discussed in §17, for particle physics we often care about *approximate* symmetries, like baryon or lepton number, or the approximate shift symmetry of an axion field or of an inflaton in the early universe. The arguments that we have discussed so far tell us that these symmetries cannot be exact, but to have a really *useful* statement about quantum gravity we need to be able to quantify how much a symmetry should be broken. This has been the focus of significant attention in recent years, which we will now review.

22. Weak Gravity Conjecture

One way to obtain an approximate global symmetry is simply to have a very weakly coupled gauge theory. If we could start with QED and dial the electric coupling constant e to be extremely small, then photons would almost decouple and the $U(1)$ phase invariance of the electron would behave very much like a global symmetry. The principle that quantum gravity does not admit global symmetries then suggests that something should prevent us from taking the strict $e \rightarrow 0$ limit of a gauge coupling. The Weak Gravity Conjecture (WGC) is a quantitative statement of what happens in this limit. More precisely, there is in fact a family of related conjectures, some of which are not very constraining and some of which can have powerful implications for phenomenology. In this section we will discuss the original Electric and Magnetic WGCs, before turning to more powerful related conjectures in the next section.

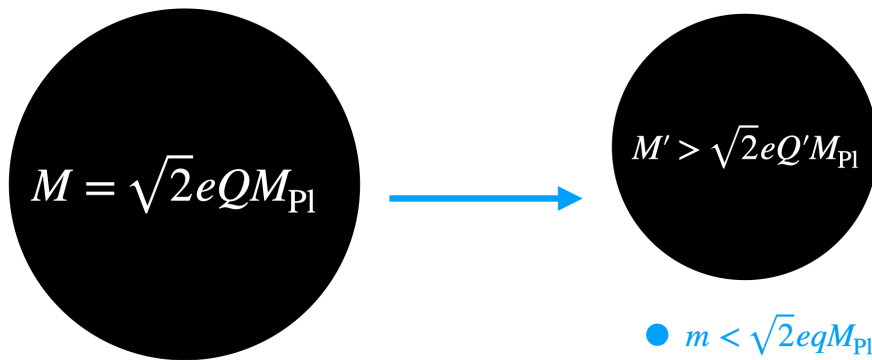


Figure 17: An extremal black hole can decay to a WGC-obeying, or superextremal, particle (in blue, with $m < \sqrt{2}eQM_{\text{Pl}}$) and a subextremal black hole.

The WGC story begins with the *Electric* WGC, proposed by Arkani-Hamed, Motl, Nicolis, and Vafa (henceforth “AMNV”) in 2006 [23]. This is simply the statement that a charged particle

exists that is not too heavy: given a $U(1)$ gauge field in a theory of quantum gravity, there exists a particle of nonzero electric charge q with a mass m obeying the inequality

$$m < \sqrt{2}e|q|M_{\text{Pl}}. \quad (218)$$

Here e is the $U(1)$ gauge coupling, and we recognize that this inequality is the *opposite* of the black hole extremality bound (209). One way to understand this condition is that it is a necessary kinematic requirement for extremal black holes to be able to shed their charge. As depicted in Fig. 17, an extremal black hole with $M = \sqrt{2}eQM_{\text{Pl}}$ can decay (by Schwinger pair production) to a WGC-satisfying particle and a subextremal black hole with $M' > \sqrt{2}eQ'M_{\text{Pl}}$. Thus, one way to understand the WGC is as a statement that black holes never get “stuck” with large charge that they can’t get rid of. However, note that if they *did*, nothing would be obviously wrong with this, as we explained in Fig. 15 and surrounding discussion. This is why the WGC remains a conjecture, although versions of it have been proved in various contexts and with various additional assumptions.

Notice that for the one $U(1)$ gauge theory that we know of in nature, the WGC is satisfied with many orders of magnitude to spare: for the electron, we have

$$\frac{m_e}{\sqrt{2}eM_{\text{Pl}}} \approx 5 \times 10^{-22}. \quad (219)$$

It would have been more exciting if this was a close call, but at least the conjecture passes this test.

The Electric WGC by itself doesn’t tell us that anything particularly bad or dramatic happens as we try to create a global symmetry by sending $e \rightarrow 0$. There must be some charged particle in the theory with a mass that tends to zero in this limit, but that seems innocuous. In particular, the WGC is a “there exists” statement, not a “for all” statement; only one particle needs to become light, for the minimal version of the conjecture.

A more useful constraint on weakly-coupled gauge theory arises from the *Magnetic* WGC, which was also explained in the original AMNV paper. The Magnetic WGC is just the WGC applied to magnetic charge. Due to Dirac quantization, when an electric field couples with strength e , the dual magnetic field couples with strength $2\pi/e$. Thus, we can just repeat (218) but now conclude that there must exist a *magnetic monopole* with (integer) magnetic charge q_M , coupling strength $2\pi/e$, and mass m_M :

$$m_M < \sqrt{2}\frac{2\pi}{e}|q_M|M_{\text{Pl}}. \quad (220)$$

So far this is essentially just a relabeling of the Electric WGC, but the crucial physics insight was that electrically charged objects (with weak coupling $e \ll 1$) and magnetically charged objects (with strong coupling $2\pi/e \gg 1$) are qualitatively different.

The key difference is in the classical versus quantum (Compton) radii of the objects. Recall from classical electrodynamics that the *classical radius* of an electron is the radius at which the classical self-energy stored in the electric field is equal to the electron’s mass, i.e., it is a radius R_C at which

$$\int_{R_C}^{\infty} r^2 dr d\Omega \left(\frac{eq}{4\pi r^2} \right)^2 = m_e. \quad (221)$$

The integral is divergent at short distances, proportional to $1/R_C$. Solving for R_C , we see that

$$R_C = \frac{\alpha q^2}{m_e}, \quad (222)$$

where $\alpha = \frac{e^2}{4\pi}$ is the fine structure constant. By comparison, the Compton radius of the object, which we will denote by R_Q with Q for “quantum,” is

$$R_Q = \frac{1}{m_e}. \quad (223)$$

Because the electric interaction is weakly coupled ($\alpha \ll 1$), we see that $R_C \ll R_Q$ (for $q \sim O(1)$). Thus, quantum mechanics resolved the classical self-energy puzzle of electromagnetism. Classically, one could argue that at the radius R_C , new physics should arise that explains why the electron mass is not extraordinarily large due to the energy stored in the electric field. In quantum mechanics, however, we see that the electron already behaves as a “fuzzy” object at the *much larger* distance scale $R_Q = \frac{1}{\alpha} R_C$, associated with virtual electron-positron pairs that screen the effective charge in QED. (This is perhaps the most dramatic example among the many places in physics where, in modern language, one could say that a “naturalness argument” successfully predicted new physics.)

Now consider the case of magnetically charged objects. Again, we define the Compton radius simply as the inverse mass, $R_Q = 1/m_M$. However, the classical radius now comes from integrating the energy stored in the *magnetic* field. This amounts to sending $e \rightarrow 2\pi/e$ in the electric formula, so we have

$$R_C = \frac{\pi q_M^2}{e^2 m_M} = \frac{q_M^2}{4\alpha} R_Q. \quad (224)$$

In other words, for magnetically charged objects, the classical radius is *much larger* than the quantum radius! (At least for the integer charge $q_M \sim O(1)$.) The behavior of R_Q and R_C in the electric and magnetic cases is illustrated in Fig. 18. In the magnetic case, we can no longer argue that quantum fuzziness will save us from a naturalness problem. Instead, we expect that new physics will come in at distances of order R_C , corresponding to an *ultraviolet cutoff*

$$\Lambda_{U(1)} \lesssim R_C^{-1} = \frac{e^2}{\pi q_M^2} m_M. \quad (225)$$

The details of this new physics will be model-dependent. In the classic example of the 't Hooft–Polyakov monopole, for instance, this is the scale at which we see that the gauge group is $SU(2)$. In the case of the Kaluza-Klein monopole, this is the scale of the extra dimensions. In any case, we expect that at the energy scale $\Lambda_{U(1)}$, the physics will no longer be described by a 4d $U(1)$ gauge theory.

The conclusion (225) relating the UV cutoff of the theory to the mass of a magnetic monopole is purely a statement about effective field theory, unrelated to gravity. However, if we combine it with the Magnetic WGC (220), we obtain a constraint relating the UV cutoff to the gauge coupling and the Planck scale:

$$\Lambda_{U(1)} \lesssim e M_{\text{Pl}}. \quad (226)$$

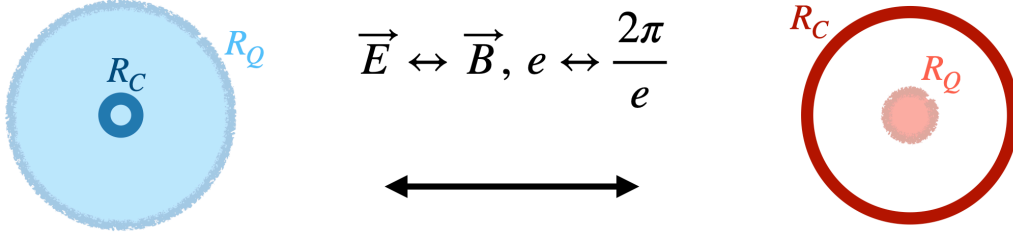


Figure 18: Illustration of the classical radius R_C and quantum (Compton) radius R_Q of electrically (left, blue) and magnetically (right, red) charged objects.

This is now a much more useful statement than the Electric WGC! It doesn't just tell us that a charged particle becomes light when we send $e \rightarrow 0$, it says that the EFT description in terms of a 4d U(1) gauge theory breaks down in this limit.

Magnetic Weak Gravity Conjecture Cutoff: a weakly-coupled U(1) gauge theory with small coupling e breaks down at or below the energy $e M_{\text{Pl}}$.

Exercise: Find the condition for the classical radius R_C of a magnetically charged object to be larger than its Schwarzschild radius. Why is it reasonable to expect that some magnetically charged object exists that obeys this condition? (See discussions in [23, 163].)

23. Extensions and refinements of the Weak Gravity Conjectures

The minimal electric WGC has little direct relevance for phenomenology. The magnetic WGC is a more interesting statement, because it tells us about when an EFT can break down at weak coupling. The last several years have led to a much sharper understanding of how and why such a breakdown of EFT can occur, due to *towers* of charged particles. Our main goal in this section is to summarize these new developments, which have enhanced the potential relevance of the WGC to phenomenological models. However, we will first turn to a relatively technical comment on the extension of the WGC to theories with multiple U(1) gauge fields, because an analysis of such a case is one of the arguments supporting the new qualitative picture of the WGC and towers.

23.1 Multiple gauge fields

In a theory with multiple U(1) gauge groups, an extremal black hole can carry a combination of all the U(1) charges. In particular, given a collection of n gauge fields each normalized with integer magnetic flux $\frac{1}{2\pi} \int F^i \in \mathbb{Z}$, with a kinetic term allowing for general mixing (recall §5)

$$\int \left(-\frac{1}{2} K_{ij} F^i \wedge \star F^j \right), \quad (227)$$

the black hole extremality condition for a 4d black hole carrying integer charges Q_i under the gauge fields A^i becomes

$$M^2 \geq M_{\text{ext}}(Q)^2 \equiv 2K^{ij}Q_iQ_jM_{\text{Pl}}^2, \quad (228)$$

where K^{ij} is the inverse matrix of K_{ij} (the analogue of e^2 in (209)). One way to extend the minimal electric WGC (218) is to generalize the kinematic criterion it captures: what is a necessary condition for all extremal black holes, with any combination of charges, to be able to shed their charge by emitting light charged particles? The answer is often referred to as the convex hull criterion [164]. A particle with mass m and charges $q_i \in \mathbb{Z}$ is represented by the charge-to-mass vector $z_i = \sqrt{2}M_{\text{Pl}}q_i/m$. The extremality bound (228) tells us that there is a region at small z , $K^{ij}z_iz_j \leq 1$, that can be occupied by black holes. The convex hull condition says that there must exist a collection of particles in the theory with z vectors whose convex hull, as measured by the metric K^{ij} , contains the black hole region.

One could imagine the convex hull condition being satisfied by a finite number of particles, but it could also be satisfied by an infinite number of states of different charges that hug the exterior of the black hole region. A general statement, valid in this limit where infinitely many different particles are relevant, is: for every direction \hat{q} in the space of possible $U(1)^n$ charges allowed by Dirac quantization, there is a superextremal multiparticle state with $z \propto \hat{q}$. By a superextremal multiparticle state, we simply mean a collection of particles whose total-charge to total-mass ratio vector z_i lies outside the black hole region. This formulation is also valid in the case when massless scalar fields affect the form of the black hole solution, which changes the quantitative form of (228) but not the qualitative existence of a black hole region in the space of z vectors.

23.2 Cutoffs take the form of towers of particles

The magnetic WGC (226) suggests that, in quantum gravity, weak coupling for a gauge theory comes at the cost of a low UV cutoff, below the Planck scale. But what happens at this UV cutoff? If, for example, it only requires us to embed a $U(1)$ gauge theory in a weakly-coupled non-abelian gauge theory, this would be a very mild form of cutoff and would not necessarily impose an interesting constraint for phenomenology.

A large body of work in the last several years has built up a compelling picture for a much more substantial sort of cutoff, involving a *tower* of charged particles [165–168]. Specifically:

For a gauge theory with a weak coupling e in quantum gravity, there is an *infinite* tower of charged particles of different charge, *each* obeying the electric WGC bound $m < \sqrt{2}eqM_{\text{Pl}}$. In the non-abelian case, we have an infinite tower of particles in different representations of the gauge group obeying the WGC with respect to the Cartan $U(1)$ subgroups.

One of the sharpest such statements consistent with known evidence, the Sublattice WGC, holds that there is a sublattice of the charge lattice of the theory on which a WGC-obeying particle exists at every site [166, 167]. That is, given any charge vector \vec{q} in the charge lattice, there is a small integer n (the *coarseness* of the sublattice) such that there is a superextremal particle of charge $n\vec{q}$. In known examples, n does not exceed 3.

Towers of particles are associated with UV cutoffs in two different ways. There is an obvious weak sense that an EFT breaks down if it fails to include massive particles that exist. But there is a more dramatic sense that an infinite tower of particles eventually leads to a complete failure of EFT altogether: all of these particles couple to gauge fields and to gravity, and run in loops, and eventually drive the whole theory to strong coupling. In particular, a gravitational theory with N weakly coupled light particle species in 4d is expected to break down by an energy scale $\Lambda_{\text{QG}} \lesssim M_{\text{Pl}}/\sqrt{N}$, sometimes referred to as the “species bound” or “species scale” [169–171]. One familiar example arises from extra dimensions that are large compared to the 4d Planck length; in this case, the higher-dimensional Planck mass M_D is parametrically below the 4d Planck mass, precisely by a factor of the square root of the volume in higher-dimensional Planck units, i.e., the square root of the number of Kaluza-Klein modes below the cutoff. You can see this just by dimensional reduction of the higher-dimensional Einstein-Hilbert term:

$$\int_{M_4 \times X} d^D x \sqrt{|g|} M_D^{D-2} \mathcal{R} \mapsto \int_{M_4} d^4 x \sqrt{|g|} M_{\text{Pl}}^2 \mathcal{R} \quad \text{where} \quad M_{\text{Pl}}^2 = M_D^{D-2} \text{Vol}(X). \quad (229)$$

One way to define the quantum gravity cutoff Λ_{QG} is to say that Λ_{QG}^{-1} is the smallest radius that a black hole admitting a good semiclassical description can have. One argument for the species bound is that a black hole with radius Λ_{QG}^{-1} can emit Hawking radiation quanta of all N light species. Unless $N \lesssim M_{\text{Pl}}^2/\Lambda_{\text{QG}}^2$, the lifetime of the black hole would be smaller than its radius, a clear breakdown of the semiclassical description [171].

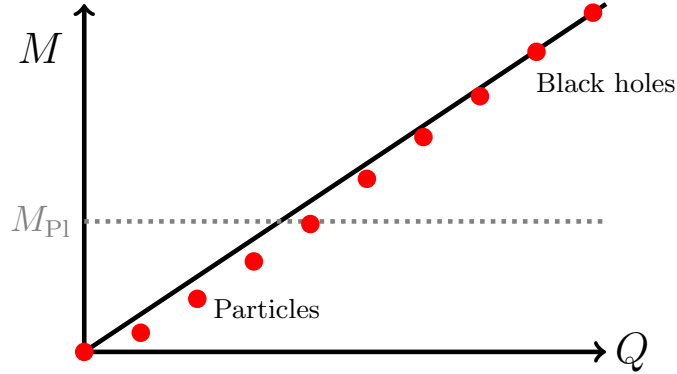


Figure 19: Mass and charge spectrum of a tower of particles (red dots). The diagonal line indicates the black hole extremality bound, which is derived from the two-derivative action. Any state lying below this line obeys the Weak Gravity Conjecture. At large (M, Q) , there are states lying along the line, but they can approach it from below if higher-derivative corrections have appropriate signs. At small M , below the Planck scale, the states are interpreted as light particles.

Formulations of the WGC in which a tower of charged particles exists beginning at the scale eM_{Pl} thus imply a powerful statement about how local quantum field theory breaks down at high energies, eventually requiring a full theory of quantum gravity to understand physics at the scale Λ_{QG} . Let’s briefly summarize several of the independent arguments pointing in this direction:

Interpolation from particles to black holes. Extremal black holes are *already* an infinite tower of states, and even the smallest semiclassical black hole has a very large entropy if $\Lambda_{\text{QG}} \ll M_{\text{Pl}}$.

This suggests that there should be many more charged states, at lower masses, that continuously interpolate into black hole states, as depicted in Fig. 19. For small black holes, higher derivative corrections modify the extremality bound, and there are reasons to think they do so in the direction of smaller mass (see, e.g., [134, 172–175]; this is now a large body of research).

Consistency under dimensional reduction. If we reduce a theory with gravity from D dimensions to $d = D - 1$ dimensions by compactifying on a circle, $x^D \cong x^D + 2\pi R$, then we can see by matching the Einstein-Hilbert term that the lower-dimensional Planck scale M_d is related to the higher-dimensional Planck scale M_D via

$$M_d^{d-2} = 2\pi R M_D^{D-2}. \quad (230)$$

Similarly, if a $U(1)$ gauge theory in D dimensions has coupling e_D and reduces to a $U(1)$ gauge theory in d dimensions with coupling e_d , we match the gauge kinetic term to find

$$\frac{1}{e_d^2} = \frac{2\pi R}{e_D^2}. \quad (231)$$

From these relations, we see that *parametrically* a WGC bound of the form $m^2 \leq \gamma_D e_D^2 q^2 M_D^{D-2}$ is maintained under dimensional reduction. The prefactor γ_D , read off from the black hole extremality condition, turns out to depend not only on D but on how scalars couple to the gauge field. Under dimensional reduction there is always a scalar radion that affects the answer in precisely the right way that the extremality bound in D dimensions reduces to the extremality bound in d dimensions. This can be thought of as a consistency check of the WGC: we might hope that a good conjecture about quantum gravity would hold independently of the number of spacetime dimensions.

However, the full story is more subtle than this. When we dimensionally reduce the theory on a circle, we also obtain a new $U(1)$ gauge boson, the Kaluza-Klein gauge boson arising from the $g_{\mu D}$ components of the metric tensor. The graviton has a tower of charged Kaluza-Klein excitations (i.e., modes with different amounts of momentum in the x^D direction), which saturate the WGC for the Kaluza-Klein $U(1)$. If we have a charged particle in D dimensions, it gives rise to a tower of particles in d dimensions that all carry the original charge as well as any number of units of additional Kaluza-Klein charge. If the D -dimensional particle obeys the WGC, all of these particles lie outside the black hole region in d dimensions. However, their convex hull does not necessarily contain the black hole region. In this sense, the WGC (in its convex hull formulation for multiple $U(1)$ factors) is *not* automatically preserved under dimensional reduction [165]. However, if we start with an infinite tower of particles of different charge in D dimensions, each obeying the WGC, their infinite set of towers of d -dimensional Kaluza-Klein excitations will satisfy the convex hull condition. This was an early motivation for postulating a strong form of the conjecture, requiring infinite towers of WGC-obeying particles.

String theory examples; modular invariance. String theory provides us with a wealth of examples of quantum gravity theories (or vacua of a single theory, depending on your perspective). The masses of various charged particles are sometimes calculable within string vacua, generally in two cases: when the theory is weakly coupled, or when the masses of states saturate a BPS bound due to supersymmetry. Whenever we are able to carry out checks in such examples, we find that there are infinite towers of charged particles that obey the WGC. In the perturbative regime, this can be proven quite generally with a string worldsheet argument based on the principle of modular

invariance [166, 167]. More recently, similar arguments have been extended to a much larger class of string compactifications known as F-theory vacua, which are essentially the strongly coupled limit of Type IIB string theory. Although F-theory is generically strongly coupled, an individual gauge group within an F-theory vacuum may be weakly coupled, and in this case spectra are again calculable, and again examples have towers of particles obeying the WGC [176–178]. Other interesting checks have been carried out for BPS states, providing examples that satisfy the tower form of the WGC away from the weakly-coupled limit [179, 180].

Emergence of weak coupling. Above, I mentioned the quantum gravity cutoff Λ_{QG} associated with the species bound, where large numbers of particles drive gravity to strong coupling. We expect that, at energies near Λ_{QG} , there are no approximate global symmetries whatsoever. In particular, we expect that there are no weakly coupled gauge theories at such a scale. Instead, weak coupling should always “emerge” only in the infrared [181]. Weak coupling is associated with a large coefficient $1/e^2$ for a gauge field kinetic term. One way to achieve such a coefficient is to integrate out a tower of charged particles, each of which contributes to this coefficient. Heuristically, then (ignoring constant factors and logarithms), we expect a relationship like (in 4d, using the standard QED beta function (65))

$$\frac{1}{e^2} \sim \sum_{i|m_i < \Lambda_{\text{gauge}}} q_i^2 = N(\Lambda_{\text{gauge}}) \langle q^2 \rangle_{\Lambda_{\text{gauge}}}, \quad (232)$$

where the sum is over all particles i with mass below the gauge theory’s UV cutoff Λ_{gauge} , with associated charge q_i . In the last step, $N(\Lambda_{\text{gauge}})$ refers to the number of particles below the energy Λ_{gauge} and $\langle q^2 \rangle_{\Lambda_{\text{gauge}}}$ refers to the average charge-squared of these particles. The assumption that the gauge theory is weakly coupled at the quantum gravity scale tells us that $\Lambda_{\text{gauge}} \lesssim \Lambda_{\text{QG}}$. From this together with the species bound, we derive that

$$\Lambda_{\text{gauge}}^2 \lesssim e^2 \langle q^2 \rangle_{\Lambda_{\text{gauge}}} M_{\text{Pl}}^2. \quad (233)$$

This is an interesting WGC-like statement that shows that the basic emergence principle implies that the *average* particle below the gauge theory’s UV cutoff obeys the WGC bound [182].

A variety of similar statements can be derived in arbitrary numbers of dimensions, and for non-abelian gauge groups. These emergence arguments also connect closely to another idea known as the Swampland Distance Conjecture (SDC) [183], which posits that in asymptotic limits of scalar field moduli spaces in quantum gravity, there is always an infinite tower of states whose mass $m(\phi)$ goes to zero exponentially, at least as fast as $\exp(-\lambda d(\phi)/M_{\text{Pl}})$ with the geodesic distance $d(\phi)$ in field space (from some fixed reference point) measured in Planck units (with λ expected to be $O(1)$; recent arguments have sharpened this to $\lambda \geq 1/\sqrt{d-2}$ [184]). One version of the emergence argument assumes that the kinetic term of ϕ is itself generated (or at least dominated) by loops of the particles in the tower, which can be shown to imply that the tower masses are exponential in the canonically normalized scalar field [185, 186]; see also [187, 188]. Limits where a gauge coupling $g \rightarrow 0$ in quantum gravity are expected to be infinite-distance limits in moduli spaces, in which case the SDC tower and the WGC tower may be one and the same, although sometimes there are multiple towers becoming light and the WGC tower is not the lightest.

Strong breaking of global symmetries at the cutoff. As discussed in §20.2.1, the existence of electrically charged particles explicitly breaks a 1-form global symmetry that acts on Wilson line operators in free $U(1)$ gauge theory. We expect that all global symmetries are broken in quantum gravity, which implies that some electrically charged particles should exist. But in fact, we have even stronger expectations: all global symmetries should be *badly* broken at the cutoff Λ_{QG} . If a gauge theory is weakly coupled, a single charged particle (like the electron) does not *badly* break the 1-form symmetry. To quantify this, one can consider how “close to topological” the symmetry operator is [189]. An unbroken symmetry has an associated topological symmetry operator. For a mildly broken symmetry, enlarging the surface on which the operator is inserted will *slowly* change a correlation function. For a badly broken symmetry, it will *rapidly* change a correlation function. It turns out that the deviation of the 1-form symmetry operator from being topological is precisely measured by the beta function of the gauge theory induced by charged particles running in loops. A tower of charged particles can badly break the 1-form symmetry, by driving the theory to strong coupling by the cutoff Λ_{QG} . This calculation then becomes essentially the same as the emergence calculation discussed above. Thus, towers of WGC-obeying charged particles are expected to appear in order to ensure that the 1-form electric symmetry is badly broken in the UV.

Before closing the discussion on towers of particles, I want to highlight one more recent development that proposes a much sharper picture of what these towers are. The Emergent String Conjecture [190], formulated by Lee, Lerche, and Weigand based on evidence in F-theory [176–178, 191], holds that weak-coupling limits in quantum gravity come in only two forms: decompactification limits and emergent string limits. In the former case, a tower of light states arises from Kaluza-Klein modes. In the latter, it arises from excitations of a string that becomes asymptotically tensionless at weak coupling. One reason why this is a very promising idea for phenomenology is that particles carrying Kaluza-Klein charge are never chiral. For many particle physics applications, we are interested in weakly coupled gauge groups coupled to chiral matter. The Emergent String Conjecture suggests that these are always described by low-tension strings. In the case of a Kaluza-Klein tower, a 4d gauge theory with coupling g can have a quantum gravity cutoff as high as $\Lambda_{\text{QG}} \lesssim g^{1/3} M_{\text{Pl}}$. However, we expect that in the case of a stringy tower, we have a quantum gravity cutoff near the WGC scale itself, $\Lambda_{\text{QG}} \lesssim g M_{\text{Pl}}$. By providing a reason to focus on the latter case in many applications, the Emergent String Conjecture greatly strengthens the power of the WGC (at the cost of relying on stronger assumptions). We will have more to say about implications for phenomenology in the following part.

23.3 The WGC for p -form gauge fields

In the discussion above I have focused on the WGC for ordinary (1-form) gauge fields, but AMNV also formulated it for general p -form gauge fields. In this case, charged objects (called “branes,” in general, especially when $p \geq 3$) have p -dimensional worldvolume, and instead of being characterized by a mass these have a tension \mathcal{T}_p with mass dimension p . A p -form gauge field has a coupling e_p of mass dimension $p + 1 - D/2$, in D -dimensional spacetime. The WGC then says that there should exist a charged object whose tension obeys the inequality

$$\mathcal{T}_p^2 \leq \gamma e_p^2 Q^2 M_{\text{Pl}}^{D-2}, \quad (234)$$

where γ is a constant factor derived from the extremality bound for charged black branes (often $O(1)$, but not necessarily when there are scalar forces that are much stronger than gravity).

In some cases, the charged objects (strings or branes) may arise as solitons within an effective field theory. An example arises in the abelian Higgs model: in the limit where the Higgs mode is heavy, we can integrate it out and write the theory in terms of a BF mass (as in §11.2). The B field dual to the Higgs phase is a 2-form gauge field, and the WGC tells us that there should exist a charged string with tension below a scale of order vM_{Pl} . This string is the familiar ANO (Abrikosov-Nielsen-Olesen) vortex. In its core, the Higgs VEV goes to zero. Such a solitonic charged object has a tension that can be derived within effective field theory (in this case, $\mathcal{T} \sim v^2$), in which case the WGC bound is usually the simple constraint that all VEVs in the theory are below the Planck scale. By contrast, a fundamental string or brane that is not describable as a soliton within some EFT has a core that probes UV physics. In such a case, the tension of the object is generally a strong cutoff on the theory, in the sense that it is of order Λ_{QG} or higher.

Part Five: Phenomenological Insights from Quantum Gravity

Now that we've seen some examples of how quantum gravity might constrain effective quantum field theories, let's try to take a further step and connect these ideas to real-world particle phenomenology and potential experiments. I will allow myself somewhat more speculation in this part of the notes than in earlier parts. On the other hand, I will try to stick close to ideas that are relatively well-established, such as conjectures that are known to hold in a large collection of quantum gravity theories. It is possible that some of these conjectures can be falsified theoretically, with explicit counterexamples in string theory vacua. This would be very interesting. (A conjecture about quantum gravity can only be shown to be false with an explicit, consistent quantum gravity theory, a basic fact about logic that is sometimes overlooked by authors who present EFT counterexamples.)

24. Charge quantization

A simple black hole argument tells us that gauge groups in quantum gravity should be compact, as discussed in §19. This means that abelian, continuous gauge groups should be made up of products of $U(1)$ factors, not \mathbb{R} , and hence electric charge should be quantized. In the Standard Model, the $U(1)_Y$ charge assignments are fixed by anomaly cancellation arguments once the $SU(3)_C \times SU(2)_L$ representations are chosen,²⁰ so we don't need to invoke quantum gravity to argue for why the charges are all multiples of a base unit. On the other hand, beyond the Standard Model, we could ask whether we might expect to see irrational charges, and quantum gravity gives us a clear answer. As explained in §5, small, non-quantized effective “millicharges” can arise by kinetic mixing of the hypercharge gauge boson with a massless dark photon. This is a well-studied scenario in particle phenomenology and cosmology. Importantly, the dark photon will mediate self-interactions of the millicharged particles that are much stronger than their interaction through ordinary photons. In some cases, millicharged particles have been studied *without* including a dark photon, and hence omitting such strong self-interactions. For irrational values of millicharge, this is inconsistent with quantum gravity.

Of course, charge quantization in quantum gravity does leave open the alternative that the basic unit of $U(1)_Y$ charge is much smaller than the smallest value we see in the Standard Model (conventionally normalized to be $1/6$). Perhaps it is, for example, $1/6000$, and the Standard Model fields all have charge that is a multiple of 1000 in terms of the base unit, *without* violating charge quantization. This is a logical possibility, albeit one that seems implausible. However, the search for quantum gravity theories in which light fields have large integer $U(1)$ charges has recently turned up interesting examples [195], so this remains an active area of investigation that may yet have surprises in store for us.

²⁰There is some fine print: this argument assumes that all three generations have the same hypercharge assignments, and even then there is one other viable solution in which the right-handed up and down quarks have equal and opposite hypercharge and the other fermions are hypercharge-neutral. See §22.4 of [192], or early discussion in [193, 194].

25. Weak Gravity Conjecture and phenomenology

25.1 The photon mass

In effective field theory, it is perfectly consistent to add a mass term for an abelian spin-1 boson. For example, we could add a mass term of the form $A_\mu A^\mu$ for the photon (which could arise by matching to a mass term for the hypercharge spin-1 boson in the full Standard Model). Doing this for *non-abelian* spin-1 fields is problematic: scattering amplitudes for the longitudinal modes of gluons of mass m interacting through the non-abelian self-interaction with coupling g give rise to amplitudes that grow with energy, signaling that the theory breaks down at or before a cutoff energy $E_{\text{max}} \sim m/g$. This argument for the W and Z bosons of the Standard Model led to the LHC “no-lose theorem” [196]. The Higgs mechanism provides extra terms in the scattering amplitude that remedy the problem. In the abelian case, however, no Higgs mechanism is needed because the dangerous self-interaction is absent. Nothing goes wrong whatsoever! This is counterintuitive, because adding a mass produces a discontinuous change in the number of degrees of freedom: the photon would have three propagating polarization states, instead of two. However, all interactions of the longitudinal mode turn off in the limit as the mass goes to zero.

Why, then, should we expect the Standard Model photon to be massless? A common answer is “gauge invariance,” but this is poor logic. A massless spin-1 boson must have an associated gauge redundancy to remove the longitudinal mode. A *massive* spin-1 boson is perfectly healthy without any such redundancy, so there is no need to include a gauge symmetry. One can do so if one wishes, writing the mass term in the Stueckelberg form $(A_\mu - \partial_\mu \theta)^2$ where θ shifts under a gauge transformation of A , but this is essentially putting an extra degree of freedom in and then promptly taking it right back out.

We know, in the real world, that if the photon has a mass it is a very tiny one. One of the easiest bounds to understand comes from Fast Radio Bursts: they emit radio waves over a range of frequencies, which arrive on Earth at about the same time. If the photon has a mass, the lower-frequency modes will travel more slowly than the higher-frequency modes, and a burst traveling over a very long distance would arrive on Earth with a noticeable delay between modes of different frequencies. Radio waves are useful for this test because they have relatively low frequency, meaning that a small photon mass has a larger effect on their speed than it would for, say, a gamma ray. Examining such a signal, one can conclude that [197, 198]

$$m_\gamma \lesssim 10^{-14} \text{ eV}. \quad (235)$$

This is not the strongest photon mass bound that you can find in the literature, it’s just one of the easiest to understand and the least sensitive to any assumptions about modeling magnetic fields in the Solar System or the galaxy. And it is, compared to the mass of any known massive particle, a very stringent bound. Given our discussion above, one might think that an effective field theorist should think that, despite this stringent bound, if we keep doing more and more precise measurements we will find someday that the photon mass is nonzero. Just like the cosmological constant or the neutrino mass, it is not forbidden by any deep principle, so it will turn out to be not exactly zero.

Really?

Surely not!

I think that any reasonable physicist *should* expect that the photon mass really is zero—*exactly* zero!—but it is surprisingly hard to give an airtight argument for this. My take on this question is that effective field theory is not sufficiently powerful to prove that the photon mass is zero, but that quantum gravity is [199]. Here is my argument, which is not completely free of loopholes, but which I find compelling.

The first case to discuss is the Higgs mechanism. We could imagine a new scalar field ϕ , which is a singlet of $SU(3)_C$ and $SU(2)_L$ with nonzero hypercharge. In order to give a very tiny mass to the photon, consistent with (235), it must have either a very small charge or a very small VEV or both. If it has a large charge, we would see its effects; for instance, the particle could be directly produced in scattering experiments, and it would affect the running of the fine structure constant. So the only way to make this work is the kind of scenario that we called “implausible” in §24, in which the fundamental unit of hypercharge is much, much smaller than the smallest hypercharge of any Standard Model fermion. This means that the QED coupling constant that we call e should really be thought of as $e_0 N$, where N is a very large integer and $e_0 \ll 1$. At energies well above m_ϕ , where the photon mass is negligible, ϕ will essentially behave like a millicharged particle, and experimental constraints tell us that $N \gtrsim 10^{14}$ for $m_\phi \lesssim \text{keV}$ (see, e.g., [200]). We could have, for example, $m_\phi \sim \langle \phi \rangle \sim 1 \text{ eV}$ and $N \sim 10^{15}$, which would give $m_\gamma \sim e_0 \langle \phi \rangle \sim 10^{-15} \text{ eV}$. There are at least two things that should make us uncomfortable here. One is that the scalar ϕ is extremely light, posing a severe fine-tuning or naturalness problem. Even *graviton* loops would produce a large correction to m_ϕ unless the cutoff is below about $\sqrt{m_\phi M_{\text{Pl}}} \sim \text{MeV}$. Of course, naturalness problems are not sharp no-go statements. The second source of discomfort is that we are postulating that the electron charge, in units of the fundamental electric charge, is on the order of 10^{15} , which seems absurd (but which I don’t currently know a completely convincing quantum gravity argument against). Furthermore, the magnetic WGC tells us that there would be a UV cutoff on the validity of the $U(1)$ gauge theory description, at a scale $\Lambda_{U(1)} \lesssim e_0 M_{\text{Pl}} \sim 1 \text{ TeV}$. This raises a concern that there could be trouble with LHC constraints (and we could use stronger bounds on the photon mass, like that arising from Jupiter’s electric field, to find a sharper problem, albeit with more assumptions to examine). However, we really have to know more about what happens at the scale $\Lambda_{U(1)}$ to see if this is a problem. If there is simply a tower of extremely weakly interacting particles coupling to the Standard Model with strength e_0 , these could easily have escaped experimental notice. The Tower or Sublattice WGCs tell us that the fundamental quantum gravity cutoff in such a scenario is below $e_0^{1/3} M_{\text{Pl}} \sim 10^{13} \text{ GeV}$, which is not an obvious problem. However, the Emergent String Conjecture would suggest that, because there are chiral fermions charged under the photon, the $e_0 \rightarrow 0$ limit is not a decompactification limit but a tensionless string limit, and so $\Lambda_{\text{QG}} \lesssim e_0 M_{\text{Pl}} \sim 1 \text{ TeV}$.

Suppose, then, that we give up on the Higgs mechanism and simply write down a bare (Proca or Stueckelberg) mass for the hypercharge boson. What could go wrong? I claim that in the quantum gravity context, we really should think of such a massive spin-1 boson as a genuine $U(1)$ gauge field that has eaten a (necessarily) compact scalar boson θ of charge q . This is because, if we consider black holes with radius much less than the Compton radius of the photon, the arguments of §19 should still apply to good approximation, and we recover our conclusions about compactness. Thus, the longitudinal mode of the photon should be a compact boson, not a noncompact one. As

discussed in §11.2, such a theory can be reinterpreted as a theory of a 2-form gauge field B that is dual to the compact boson θ , in the sense that $\frac{1}{2\pi} dB = f^2 \star d\theta$. In this interpretation, the photon mass is due to a $\frac{1}{2\pi} B \wedge F$ term. Now, we apply the Weak Gravity Conjecture to both the photon (with fundamental coupling e_0) and to the B field (with coupling $2\pi f$). The photon mass is $e_0 q f / (2\pi)$. Thus, a small photon mass requires either $e_0 \ll 1$ (similar to the discussion in the Higgs case) or a small f (or both). However, now a new problem arises for small f : the WGC for p -form gauge fields (234) tells us that there are strings charged under the B field with tension $\mathcal{T} \lesssim 2\pi f M_{\text{Pl}}$. In the Higgs case, these strings are ANO vortices, solitonic solutions where the Higgs VEV goes to zero in the string core. However, in the Stueckelberg case, by definition there is no Higgs field and the string core is not describable within effective field theory. As a defining feature of a scenario where the UV completion is *fundamentally* of Stueckelberg type, rather than just a limit of the Higgs scenario, the string should be a fundamental object whose core probes deep UV physics, not resolvable within any EFT description. In this case, the tension of the string is at or above the quantum gravity cutoff. In other words, for a given UV cutoff Λ_{QG} , we obtain a *lower* bound on f from the WGC:

$$\sqrt{2\pi\mathcal{T}} \gtrsim \Lambda_{\text{QG}} \quad \Rightarrow \quad f \gtrsim \frac{\Lambda_{\text{QG}}^2}{4\pi^2 M_{\text{Pl}}}. \quad (236)$$

Now, very conservatively, we know that Λ_{QG} is at least above the TeV scale, which tells us that $f \gtrsim 10^{-5}$ eV. This is not nearly small enough to explain (235) without also invoking a huge integer N , as in the Higgs case. Thus, quantum gravity strongly disfavors the possibility that the photon has a nonzero mass. I expect that, in fact, it completely rules it out, though the argument is not rigorous (and in particular, we would have to close the large-integer loophole [201]).

Similar conclusions apply to dark photons. In fact, there are interesting scenarios where string theory compactifications contain dark photons with Stueckelberg masses [202, 203], and the relationship between their mass and the UV cutoff is consistent with this discussion. These conclusions also apply to the magnetic photon mass discussed in §11.2.1.

25.2 Can a light B – L gauge boson exist?

The Standard Model with right-handed neutrinos admits a $U(1)_{\text{B-L}}$ global symmetry. By right-handed neutrinos, I mean fermions that do not carry any charges under the Standard Model gauge group. They can, potentially, have Dirac masses that pair them up with the ordinary neutrinos inside the lepton doublets, via the Higgs. However, in this context they should *not* have a Majorana mass term (which is allowed by all the gauge symmetries), because that would break the $U(1)_{\text{B-L}}$ symmetry. Beyond that, I assume that all renormalizable terms of the Standard Model are included, but only those non-renormalizable terms consistent with the B – L symmetry are allowed. This symmetry is non-anomalous, both in the sense that it has no ABJ anomaly with the Standard Model (which is to say, it is a good global symmetry of the quantum theory) and it has no 't Hooft anomaly (so that it can be gauged).

It is thus natural to ask: is $U(1)_{\text{B-L}}$ a gauge symmetry of our universe?

It is certainly possible that it is a *higgsed* gauge symmetry, perhaps with an $O(1)$ coupling, provided that it is broken above the TeV scale. But if it is a massless gauge symmetry, it mediates a new long-range force that acts on ordinary matter. Experimental constraints are very strong: $e_{\text{B-L}} \lesssim 10^{-24}$ [204, 205]. The associated WGC scale is $e_{\text{B-L}} M_{\text{Pl}} \lesssim \text{keV}$. The minimal (electric)

WGC can be satisfied by neutrinos. Even the form of the WGC calling for a *tower* of $B - L$ charged particles at the keV scale, as discussed in §23.2, is not obviously ruled out, because these particles would interact extremely weakly with ordinary matter. The scale at which the tower drives gravity to strong coupling could be as high as $e_{B-L}^{1/3} M_{Pl} \lesssim 10^{10}$ GeV, safely out of reach of current experiments.

However, the Emergent String Conjecture is more powerful. It tells us that we should expect that an extraordinarily weak e_{B-L} would arise only in one of two limits: a decompactification limit, with a tower of Kaluza-Klein modes, or an emergent string limit. The former would have a higher quantum gravity cutoff, but is incompatible with $B - L$, because the Standard Model contains chiral fermions carrying $B - L$ charge. Kaluza-Klein modes are never chiral. This leaves the emergent string case, but in that case, we expect the fundamental cutoff to be $\Lambda_{QG} \sim e_{B-L} M_{Pl}$. This is in clear contradiction with the validity of local EFT well above the keV scale.

This example highlights the utility of sharpened conjectures that have emerged in recent years. It is important to put them on a more solid theoretical footing, so that we can argue that phenomenologically viable scenarios, like an ultralight $B - L$ gauge boson, are in fact incompatible with the principles of quantum gravity as we currently understand them. Conversely, experiments that pursue such scenarios are useful, because they could falsify proposed principles of quantum gravity.

26. Neutrino masses

The origin of neutrino masses is not yet known. One possibility is that, in the absence of new degrees of freedom, they arise from a dimension five operator:

$$\frac{1}{M} c_{ij} (h \cdot \ell_i) (h \cdot \ell_j) + \text{h.c.} \quad (237)$$

Here M is a mass scale, c_{ij} a dimensionless matrix, and the i and j indices label generations. We have (in unitary gauge)

$$h = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + h^0) \end{pmatrix}, \quad \ell_i = \begin{pmatrix} \nu_i \\ \ell_i^- \end{pmatrix}, \quad (238)$$

and $h \cdot \ell$ denotes the antisymmetric contraction of $SU(2)_L$ indices, so this term becomes

$$\frac{1}{2M} c_{ij} (v + h^0)^2 \nu_i \nu_j + \text{h.c.}, \quad (239)$$

which is a Majorana mass term for the neutrinos. We don't know the absolute neutrino mass scale, but atmospheric neutrino oscillations tell us about a difference in squared masses, $\Delta m^2 \approx 2.4 \times 10^{-3} \text{ eV}^2$, which implies that at least one neutrino mass is larger than 0.05 eV. If we take $c \lesssim 1$ this corresponds to a mass scale $M \lesssim v^2 / (0.05 \text{ eV}) \sim 10^{16} \text{ GeV}$. This is the well-known *seesaw scale* for Majorana neutrino masses. The physics at the scale M could consist of heavy singlet fermions which themselves have Majorana masses, which is the standard seesaw mechanism. However, this is just one possibility, not a requirement.

The Majorana mass scenario requires no special symmetries, and the associated scale M is below the Planck scale. If the c_{ij} are small (perhaps for whatever—currently unknown—reason many of the Standard Model Yukawa couplings are small), the scale could be even lower. None

of this runs into any tension with anything that we know about quantum gravity. In many ways, Majorana neutrinos are the simplest option. This scenario is also testable through neutrinoless double beta decay experiments, though the favored parameter range will not be in reach of near-future experiments.

The other scenario is that neutrinos have *Dirac* masses: in this case, there are additional light degrees of freedom in the form of fermions $\tilde{\nu}_i$ that are neutral under the Standard Model gauge group, with mass terms arising from Yukawa couplings of the form

$$(y_\nu)_{ij} (h \cdot \ell_i) \tilde{\nu}_j + \text{h.c.}, \quad (240)$$

directly analogous to the up-type quark masses in the Standard Model. These terms preserve a $U(1)_L$ “lepton number” symmetry, $\ell_i \mapsto e^{-i\alpha} \ell_i$, $\tilde{\nu} \mapsto e^{+i\alpha} \tilde{\nu}$. This symmetry has an ABJ anomaly, but it can be extended to the non-anomalous $B - L$ symmetry. For this mechanism to explain the data, the parameters y_ν must be much smaller than any other Yukawa couplings in the Standard Model. However, the known Yukawas already span several orders of magnitude for unknown reasons (maybe flavor symmetries, maybe locality in extra dimensions, or some combination thereof). Perhaps a full understanding of this mechanism would also explain why the y_ν are small. (Another option is to consider $\tilde{\nu}_j$ to be fundamentally a higher dimension operator, e.g., to realize these neutrinos as composite states.)

This scenario requires more care to embed in quantum gravity. In particular, it requires a symmetry to forbid large Majorana masses for the singlet fermions $\tilde{\nu}$, of the form

$$M_{ij} \tilde{\nu}_i \tilde{\nu}_j + \text{h.c.} \quad (241)$$

In quantum gravity, we can’t invoke a global symmetry, so the absence of this term should somehow be explained as either a direct consequence or a side effect of gauge symmetry. The obvious candidate is to gauge $B - L$. As discussed in §25.2, a massless $B - L$ gauge boson is highly constrained by data, and incompatible with plausible forms of the WGC. One possibility is that it is gauged but higgsed at some high energy scale. If it is higgsed to nothing, or to a \mathbb{Z}_2 subgroup, then the terms (241) can be generated below the scale of higgsing. However, if these terms are sufficiently small, the neutrinos can be “pseudo-Dirac,” i.e., they can have mostly Dirac masses with small Majorana terms that slightly split each Dirac fermion into two nearly-degenerate Majorana mass eigenstates. However, in this case there are experimental constraints that require that the pseudo-Dirac mass terms be Dirac to *extremely* high approximation [206], which strongly restricts the possible means of generating Majorana terms.

A simpler scenario is to gauge a \mathbb{Z}_k subgroup of $B - L$, with $k > 2$. (Gauging the non-anomalous discrete subgroup of $B + L$ is also an option.) This forbids the Majorana mass terms (241), without producing any obvious pathology. Thus, if experiments indicate that neutrino masses are Dirac, a very plausible explanation would be that there is a new discrete gauge symmetry in nature. This would be a profound discovery, as it differs from the known ingredients in the Standard Model. However, it is difficult to decisively test experimentally. Discrete gauge groups in quantum gravity imply the existence of cosmic twist strings, which have Aharonov-Bohm interactions with charged particles. Such strings must exist as dynamical objects in the theory to avoid a generalized 2-form global symmetry [148]. If the \mathbb{Z}_k gauge symmetry is fundamental,

rather than a remnant of a higgsed $U(1)$, there is no obvious cosmological mechanism to produce the strings, so there is no straightforward observational constraint.

Thus, I leave you with a challenge: if neutrino masses are Dirac (or pseudo-Dirac to very good approximation), quantum gravity leads us to expect that there is a new gauge symmetry of nature beyond the Standard Model. Is there a way to confront this expectation with experiment? In the meantime, my bet is on the Majorana scenario (though not with great confidence).

27. The Strong CP Problem in quantum gravity

In my opinion, the Strong CP problem that we introduced in §12 is currently the best place to look for an interface between ideas from quantum gravity and experimental particle physics. Solutions to this problem are all closely related to symmetries. One approach assumes that CP or some other generalized parity symmetry is a symmetry of nature [75, 76]. Because these are *spacetime* symmetries, the only way that they can be gauged is via quantum gravity. The symmetry breaking phase transition produces exactly stable domain walls [31], which must be inflated away for a viable cosmology. Such CP- or parity-based models are interesting, but in my opinion, less plausible than the axion solution.

Both the massless up quark solution and the axion solution to the Strong CP problem can be thought of as gauging the instanton number symmetry of QCD, in the sense discussed in §15: the equations of motion lead to a Gauss law constraint that sets the net instanton number in spacetime to zero. In the language of §20, we can think of instanton number as a kind of (-1) -form symmetry charge: it must be integrated over *all* dimensions of spacetime, rather than over a slice of positive codimension. Such a putative symmetry is not as well-understood as higher p -form symmetries, but there are reasons to think that (-1) -form global symmetries are also forbidden in quantum gravity [61, 207], and must be either broken (meaning that there must be some field configurations interpolating between those with different instanton number, which obviously requires a UV completion beyond the original gauge theory) or gauged. Gauging with a chiral fermion current, as in the massless up quark solution, seems unlikely to be realized in quantum gravity. It requires a chiral symmetry that is broken *only* by the ABJ anomaly. Such a broken symmetry cannot be gauged, so it is difficult to see why there would not be generic symmetry-violating terms in the action.²¹ On the other hand, the axion can literally *be* a gauge field. We have many examples in string theory in which 4d axions arise from higher-dimensional gauge fields. It has long been appreciated that axions are ubiquitous in string theory [103]. What we can add to that, from a more modern perspective, is that they appear to be there not just because we are looking under a lamppost of (for example) highly supersymmetric vacua, but because they have an important job to do: the axions eliminate the would-be (-1) -form global instanton number symmetry by gauging it. This gives us a reason to be optimistic that they will still exist even in less familiar corners of the quantum gravity landscape. When we think of axions as a type of gauge field, we are naturally led to ask what the Weak Gravity Conjecture has to say about axion physics.

²¹An anomalous chiral symmetry can be gauged by a massive gauge field, as in the Green-Schwarz mechanism. In this case, the gauge fields eat an axion-like field to obtain its mass. Examples of this type in string theory often nonetheless have a *different* axion in the low-energy EFT [208–211].

27.1 The WGC for axions

In §23.3, we discussed the generalization of the WGC to p -form gauge fields. In §15, we argued that axions should be thought of as a type of 0-form gauge field. This suggests that the WGC can be applied to axions. In fact, this application was already discussed in the original AMNV paper proposing the WGC [23]. The analogy is that the gauge coupling e_p becomes the inverse decay constant $1/f$, the objects charged under the axion are instantons, and the analogue of the mass or tension is just the *action* S_{inst} of the instanton. Thus, the axion WGC is the statement that an instanton should exist with

$$S_{\text{inst}} \lesssim \frac{M_{\text{Pl}}}{f}. \quad (242)$$

The prefactor that turns \lesssim into a sharp inequality depends, as usual, on the extremality bound, and is expected to be $O(1)$ if all scalar forces have strength comparable to gravity. In the absence of a saxion mode, there are no non-singular gravitational instanton solutions analogous to extremal black holes, so it is unclear if we can assign a precise constant prefactor at all. However, in cases with a saxion mode there is an analogue of the extremality bound for instantons [212–214], so the analogy can be made precise. In many cases, the axion WGC is just the dimensional reduction of a more well-behaved p -form WGC in a higher dimensional theory. (See discussions in [165, 215].)

The axion WGC can be directly applied to QCD, where we know that there are BPST instantons with $S_{\text{inst}} = 8\pi^2/g^2$. The axion WGC says that *some* instanton should obey the bound, but we expect that this should apply to the conventional gauge theory instantons. In particular, if the axion is to solve the Strong CP problem, there should not be other contributions to the axion potential that will dominate over the QCD contribution. Thus, a reasonable expectation for a QCD axion is that the axion WGC implies that

$$f \lesssim \frac{g^2}{8\pi^2} M_{\text{Pl}} \sim 10^{16} \text{ GeV}. \quad (243)$$

This is a nontrivial prediction about axion physics from the WGC! It disfavors axions with very high decay constant. Importantly, the range $f \lesssim 10^{16} \text{ GeV}$ of QCD axion decay constants can realistically be targeted by experiments in the foreseeable future [19].

Note that there is some ambiguity in the interpretation of (243) because the coupling g runs and becomes large in the infrared. We do not currently have a sharp enough understanding of the axion WGC to be very precise about this, but we expect that the right interpretation is that g should be evaluated at a UV scale, such as f . In the numerical estimate, we have used the GUT value $\alpha_{\text{GUT}} \approx 1/25$. Because the running is slow in the UV, the precise details are not very important.

27.2 The WGC and axion strings

The natural next step after considering the WGC for axions is to consider the *magnetic* WGC for axions, or equivalently, the WGC for the 2-form gauge field B that is Hodge dual to the axion, with $f^2 \star d\theta \equiv \frac{1}{2\pi} dB$. (Recall the discussion in §11.2; there is also a subtlety related to the modified Bianchi identity, to which we will return in a moment.) The objects charged under B are strings: these are the *axion strings* around which θ winds. The general p -form WGC becomes, in this case,

$$\mathcal{T} \lesssim \gamma 2\pi f M_{\text{Pl}}, \quad (244)$$

for some prefactor γ which, as usual, we would read off from the extremality bound for black string solutions, and which we expect to be $O(1)$ unless there are scalar forces that are much stronger than gravity. There are several subtleties for the case of axion strings. The tension of an axion string in 4d has a logarithmic IR divergence associated with the energy cost of the winding of θ at large distances. We interpret the scale \mathcal{T} in (244) as the tension of the core of the string, not this long-distance contribution. Gravitational backreaction is also important: cosmic strings in 4d (or codimension-2 objects in any number of dimensions) cause a *deficit angle* in space. If their tension becomes larger than the Planck scale, they would effectively eat up the entire space, so that no static string solution exists. In this case, there are interesting time-dependent string solutions, explored for example in [216]. For our current purposes, in light of (243), we are interested in values of f well below the Planck scale and we will assume that gravitational backreaction is small.

If we combine (243) and (244), we obtain a bound on the axion string tension that depends on the gauge coupling and the Planck scale:

$$\mathcal{T} \lesssim \gamma \frac{g^2}{4\pi} M_{\text{Pl}}^2. \quad (245)$$

We expect massive string modes (closed loops of string, interpreted as particles) to have mass at the scale $\sqrt{2\pi\mathcal{T}} \lesssim g M_{\text{Pl}}$. But this bounds their mass in terms of the WGC scale for the gauge theory, $g M_{\text{Pl}}$!

We believe that this is not a coincidence [217]. The reason lies in the subtlety we encountered when constructing the B field dual to the axion when the axion couples to gauge fields (148). We saw that to consistently reproduce the modified Bianchi identity $d\star d\theta \propto \text{tr}(G \wedge G)$, it is necessary that the gauge field B shifts under a gauge transformation of the ordinary gauge field A . This implies that we cannot simply write a coupling $\int_{\Sigma} B$ on the axion string worldsheet Σ : it would not be gauge invariant under an A gauge transformation. This requires charged modes to exist on the axion string worldsheet, an example of the general phenomenon of anomaly inflow. In fact, this was one of the original examples [64].

Following [64], let's explain this physics again, from a slightly different perspective. Around the axion string, we have $\oint d\theta = 2\pi$. In particular, θ cannot be well-defined in the core of the string. This motivates rewriting the axion coupling to gauge fields by integrating by parts:

$$\frac{N}{8\pi^2} \int \theta \text{tr}(G \wedge G) = -\frac{N}{2\pi} \int d\theta \wedge J_{\text{CS}}, \quad (246)$$

with the Chern-Simons current (recall §11.1)

$$J_{\text{CS}} = \frac{1}{4\pi} \text{tr} \left(A \wedge dA - i \frac{2}{3} A \wedge A \wedge A \right). \quad (247)$$

The form (246) of the term that is most convenient in the presence of an axion string is not gauge invariant under A gauge transformations! In fact, if $A \mapsto A + d\alpha$, we have

$$J_{\text{CS}} \mapsto J_{\text{CS}} + \frac{1}{4\pi} \text{tr}(\alpha G). \quad (248)$$

This makes the action (246) ill-defined on its own, but the string worldsheet action can *also* not be gauge invariant, in such a way that the sum of (246) and the string worldsheet action *is* invariant.

The way that this can happen is that the $(1+1)$ d string worldsheet hosts an *anomalous* gauge theory with chiral, charge-carrying excitations. Figure 20 depicts a closed axion string loop with such a charge-carrying mode excited.

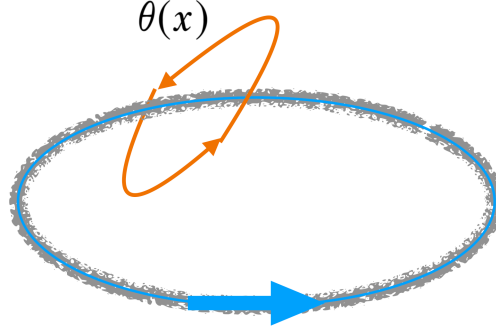


Figure 20: Sketch of a closed axion string (fuzzy gray loop) with circulating electric charge under an ordinary gauge group. The orange loop indicates that θ winds from 0 to 2π when circling the string. The blue arrow indicates the direction of propagation of electric charge along the string.

To summarize, if an axion couples to gauge fields, axion strings can always admit charge under the gauge field. The combination of the axion WGC bound on the decay constant (243) and the magnetic axion WGC bound on the string tension (244) then tells us, as in (245), that the axion string modes obey the WGC for the *ordinary* gauge field (at least up to $O(1)$ prefactors; see [218]). Such a “mixing” of WGC bounds among different gauge fields that couple via Chern-Simons terms is quite general [217, 219].

This leads us to conclude:

Axion strings are WGC towers. When a gauge field couples to an axion, the tower of charged particles satisfying the WGC for the gauge field can arise from closed loops of axion string with charge circulating around the string.

As discussed in §23.3, there is an important distinction between *solitonic* axion strings, which arise in conventional 4d axion theories like the KSVZ model and have a core described within effective field theory, and *fundamental* axion strings with cores that probe UV physics. In string theory, the latter often arise as either the fundamental F-string of string theory or as D-branes wrapped on internal cycles. It is the latter type of string that we expect to give rise to a tower of states corresponding to a fundamental UV cutoff on the theory.

A theory with a very light axion has a very good approximate global symmetry in the infrared: the shift symmetry of the axion. We expect that in quantum gravity all global symmetries are badly broken in the ultraviolet; for an axion arising as a mode of a higher-dimensional gauge field, the symmetry breaking is expected to be via towers of charged objects, as discussed in §23.2. One can ask: how good can an approximate global symmetry be in the infrared? It has been proposed that any symmetry will be broken by effects of order $\exp(-S_{\text{BH}})$ where S_{BH} is the entropy of the smallest semiclassical black hole (see [220, 221] and an interesting recent semiclassical calculation [162]). If this black hole has radius Λ_{QG}^{-1} , this estimate corresponds to $\exp(-8\pi^2 M_{\text{Pl}}^2 / \Lambda_{\text{QG}}^2)$. Comparing

this to the suppression of an axion potential generated by gauge theory instantons, $\exp(-8\pi^2/g^2)$, we see that the instanton effects are dominant provided $\Lambda_{\text{QG}} \leq g M_{\text{Pl}}$, consistent with a cutoff at the scale where the WGC tower appears.

27.3 The expected axion

I will now briefly summarize some expectations that I have about axions, based on work in progress that will be published elsewhere. In this subsection, I will allow myself to venture further out on a limb than I have elsewhere in these lecture notes.

We have argued that *if* an axion exists in a theory of quantum gravity, the WGC implies an upper bound on its decay constant and a corresponding upper bound on the tension of axion strings, and that the axion strings may supply the WGC tower of charged states. We have also pointed out that the axion can be thought of as gauging the (-1) -form instanton number global symmetry, and that there are reasons to expect that quantum gravity requires that this symmetry be either gauged or broken. However, this leaves open the possibility that the symmetry is simply *broken*, without any axion field in the effective theory.

There are a number of examples of quantum gravity theories in which we find gauge fields that do not couple to an associated axion. These gauge fields include the graviphoton field in rigid Calabi-Yau compactifications [222] and Kaluza-Klein gauge fields (which may couple to axions in various ways, but not the standard $\theta F \wedge F$ coupling [61, 217, 223]). There are cases where an axion coupled to a gauge field exists, but has a large tree-level mass of monodromy type [224], i.e., a θdC_3 coupling of the general BF form discussed in §11.2. This allows an instanton to dissolve into dC_3 flux, which is tantamount to breaking the (-1) -form instanton number symmetry. Thus, one might conclude that some gauge fields in quantum gravity theories couple to light axions, and others do not, so a general argument will not shed light on whether or not we expect a light axion coupled to gluons to exist in our universe.

However, the examples of which I am aware in which one finds gauge fields without a light axion are *also* examples lacking light charged matter. For example, in Kaluza-Klein compactifications, fields with Kaluza-Klein charge necessarily have a mass at the cutoff scale at which the 4d EFT breaks down. In the rigid Calabi-Yau case, the charged objects are D3 branes wrapped on the holomorphic 3-cycle of the Calabi-Yau, with mass around the 4d Planck scale. Such examples leave open the conjecture that quantum gravity theories with gauge fields coupled to light charged matter must have a gauged, rather than broken, (-1) -form instanton number symmetry, with a corresponding light axion field. One heuristic reason for thinking this might be true is that if the instanton number symmetry is broken, we expect that it should be badly broken at the cutoff scale. This means that an instanton configuration of size Λ_{QG}^{-1} should be easily deformed into configurations with no instanton number at all. However, if there are light charged fermions in the theory, such instanton field configurations have long-range fermionic zero modes attached, as discussed in §9.6. These act to stabilize the instanton configuration against short-range deformations that could destroy it.

This perspective also resonates with the Emergent String Conjecture. For asymptotically weak-coupling limits, it states that the tower of light modes is either a set of Kaluza-Klein modes or a set of string modes. The Kaluza-Klein case is incompatible with light charged matter; the string case resembles an axion string (see other related comments on axion strings in [225, 226]). What

I am suggesting here is stronger, in that my remarks are not restricted to asymptotic limits. This is likely necessary to make contact with phenomenology: the Standard Model gauge couplings are $O(1)$ numbers, suggesting that at least the moduli that control these couplings are far from any asymptotic regime.

The suggestion that gauge theories with light charged matter fields require a light axion is an example of a compelling phenomenological claim that may be derivable from quantum gravity. Any UV completion of the Standard Model that respects some plausible general principles, such as the absence of generalized global symmetries, may be required to contain a light axion field, with the crucial interaction with gluons that makes it at least a strong candidate for solving the Strong CP problem. This suggestion is consistent (in a nontrivial way) with a large body of evidence, and amenable to further study. A counterexample of a string compactification with gauge fields and light charged matter but no light axion would immediately imply that I'm on the wrong track, unless it has other unusual features that could motivate a refined conjecture.

My expectation, based on my current understanding of quantum gravity, is that a light axion field with a $\theta \text{tr}(G \wedge G)$ coupling exists in our universe. The axion is likely to arise from a higher-dimensional gauge field. Furthermore, the axion decay constant should be near the fundamental UV cutoff of the theory, with the axion string as a fundamental object rather than a solitonic one. More specifically, I expect that

$$\Lambda_{\text{QG}} \lesssim \frac{4\pi^2}{g} f, \quad (249)$$

with f the axion decay constant (which is experimentally measurable) and g the QCD coupling evaluated at high energies. This is a stronger statement than (245), based on the expectation that if f is far below the upper bound in (242) with $O(1)$ coefficient, the axion string tension will be correspondingly far below the upper bound (244) with $O(1)$ coefficient, because the prefactors in both cases become small in the same limit of strong scalar forces. There are several assumptions underlying these claims, which I will present in more detail elsewhere. What I hope that readers will take away is that there is at least the potential to extract nontrivial expectations about phenomenology from our current knowledge of quantum gravity. These are not rigorous theorems, but they rest on plausible general assumptions, and if they are falsified we will learn that quantum gravity in our universe behaves rather differently from the examples of consistent quantum gravity theories that we know so far.

28. Closing remarks

This section of these lecture notes might age more rapidly than the rest, but I would like to give some big-picture thoughts about the current state of particle physics, and how I would situate the topics discussed in these notes within that broader context.

Before 2012, a large fraction of the particle physics community agreed on one urgent goal: understand the mechanism of electroweak symmetry breaking. Since the LHC discovered a particle with all the expected properties of the Higgs boson, completing the Standard Model, the priorities have become less clear. We lack a theory of dark matter and of the matter-antimatter asymmetry in the universe. The Standard Model itself has many unexplained small parameters. These remain

important problems, but the community's confidence that the answers will be discovered by current or near-future experiments has decreased. In this environment, what should we do?

Of course, opinions vary widely, but here are some of mine. On the experimental front, from the bottom up, there are several experiments that have the potential to push constraints on physics beyond the Standard Model into the many-TeV range. These include searches for electric dipole moments [227] and for charged lepton flavor violation (e.g., $\mu \rightarrow e$ conversion when scattering on a nucleus) [228]. These have in common that the Standard Model prediction is many orders of magnitude below current sensitivity, so there is a large territory in which any signal at all would constitute a definitive discovery of new physics. The reason they have such high sensitivity is that they search for new violation of symmetries that are broken in only very mild ways by known physics.

In my opinion, it should be an urgent priority for the field to build a new high-energy collider, with reach well beyond that of the LHC, whether this is a hadron collider, a muon collider, or (if technology develops fast enough) a high-energy linear e^+e^- collider. If we put this off until several decades in the future, both the expertise and the enthusiasm for high-energy colliders that currently exists in the community will dwindle, and will be difficult to recreate. But it may be that the community will not rally behind such a high-energy machine until we get a clear signal from somewhere else, like an EDM, that there is new physics not too far away.

Efforts to study particle physics using astrophysics and cosmology have greatly expanded in recent years. It is possible that we will learn dramatically different information from such probes of new physics than we could learn from any terrestrial experiment. For example, a measurement of a nonzero tensor to scalar ratio r in the CMB would tell us that the Hubble scale during inflation was large (say, 10^{13} GeV), providing one anchor at a very high energy scale that could be the entry point for learning more (e.g., through non-Gaussianities) about heavy particles that could have existed during inflation. We should enthusiastically pursue such well-motivated opportunities to learn about high energy physics from clues left behind from the early universe.

While there are a number of such well-motivated experimental and observational directions to pursue, recent years have also seen an explosion of phenomenological and experimental fishing expeditions, searching for models that have no particular theory motivation simply because we can. There is nothing wrong with this, to the extent that it can be done affordably and we might stumble across something unanticipated and exciting. But I think that we must also continue, from the theory side, to seek a better understanding of where we should *expect* to find new physics. The hubristic claims of some theorists in advance of the LHC have cast such pursuits in a bad light, but we should not forget that often in the history of physics, reasoning about theory from general principles has led to conclusions that were confirmed by experiments decades later.

Developments in quantum gravity provide an often overlooked tool for thinking about particle physics. Quantum gravity is extraordinarily difficult to connect with experiment, because it seems to offer no guarantee of new physics below the Planck scale, and there seem to be an enormous number of consistent quantum gravity theories (or vacua of one theory). But recent years have shown us that some of the oldest posited principles of quantum gravity, such as the absence of global symmetries, have much richer implications than anyone had previously anticipated. We should push such principles as far as we can, to try to find sharp confrontations between their predictions and experiment or observation. My current belief is that the Strong CP problem offers

the most promising arena for bridging the gap between such principles and the real world. I think that it should be a high priority of the theory community to assess to what extent the existence of a light axion follows from general principles of quantum gravity, together with our existing knowledge about particle physics in our universe. More generally, axions and neutrinos with Majorana masses are two of the obvious places in particle physics where a very high energy scale appears in the denominator of a mass, potentially linking light particles with physics at the quantum gravity cutoff scale. This is a way that, even in standard quantum field theory, ultraviolet energies can leave behind clues in the far infrared. We should make the most of such opportunities.

I am optimistic that the next two decades will see a convergence between pure theory and dramatic experimental results. I hope that these lecture notes will provide some young theorists with useful tools to contribute to this effort.

References

- [1] D. Gaiotto, A. Kapustin, N. Seiberg, and B. Willett, “Generalized Global Symmetries,” *JHEP* **02** (2015) 172, [arXiv:1412.5148 \[hep-th\]](#).
- [2] P. R. S. Gomes, “An Introduction to Higher-Form Symmetries,” [arXiv:2303.01817 \[hep-th\]](#).
- [3] J. McGreevy, “Generalized Symmetries in Condensed Matter,” *Annu. Rev. Condens. Matter Phys.* **14** (2023) 57–82, [arXiv:2204.03045 \[cond-mat.str-el\]](#).
- [4] C. Cordova, T. T. Dumitrescu, K. Intriligator, and S.-H. Shao, “Snowmass White Paper: Generalized Symmetries in Quantum Field Theory and Beyond,” in *2022 Snowmass Summer Study*. 5, 2022. [arXiv:2205.09545 \[hep-th\]](#).
- [5] M. Shifman, *Advanced Topics in Quantum Field Theory*. Cambridge University Press, 4, 2022.
- [6] E. J. Weinberg, *Classical solutions in quantum field theory: Solitons and Instantons in High Energy Physics*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 9, 2012.
- [7] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*. Addison-Wesley, Reading, USA, 1995.
- [8] E. Fradkin, *Quantum Field Theory: An Integrated Approach*. Princeton University Press, 2021.
- [9] J. A. Harvey, “TASI 2003 lectures on anomalies,” 9, 2005. [arXiv:hep-th/0509097](#).
- [10] M. Dine, “TASI lectures on the strong CP problem,” in *Theoretical Advanced Study Institute in Elementary Particle Physics (TASI 2000): Flavor Physics for the Millennium*, pp. 349–369. 6, 2000. [arXiv:hep-ph/0011376](#).
- [11] A. Hook, “TASI Lectures on the Strong CP Problem and Axions,” *PoS TASI2018* (2019) 004, [arXiv:1812.02669 \[hep-ph\]](#).
- [12] J. E. Kim and G. Carosi, “Axions and the Strong CP Problem,” *Rev. Mod. Phys.* **82** (2010) 557–602, [arXiv:0807.3125 \[hep-ph\]](#). [Erratum: *Rev.Mod.Phys.* 91, 049902 (2019)].
- [13] H. Georgi, *Weak Interactions*. <https://sites.harvard.edu/hgeorgi/files/2022/04/weak.pdf>.
- [14] H. Leutwyler, “Principles of chiral perturbation theory,” in *Hadrons 94 Workshop*. 5, 1994. [arXiv:hep-ph/9406283](#).
- [15] A. V. Manohar, “Effective field theories,” *Lect. Notes Phys.* **479** (1997) 311–362, [arXiv:hep-ph/9606222](#).

- [16] B. R. Safdi, “TASI Lectures on the Particle Physics and Astrophysics of Dark Matter,” 3, 2023. [arXiv:2303.02169 \[hep-ph\]](#).
- [17] D. J. E. Marsh, “Axion Cosmology,” *Phys. Rept.* **643** (2016) 1–79, [arXiv:1510.07633 \[astro-ph.CO\]](#).
- [18] P. W. Graham, I. G. Irastorza, S. K. Lamoreaux, A. Lindner, and K. A. van Bibber, “Experimental Searches for the Axion and Axion-Like Particles,” *Ann. Rev. Nucl. Part. Sci.* **65** (2015) 485–514, [arXiv:1602.00039 \[hep-ex\]](#).
- [19] J. Jaeckel, G. Rybka, and L. Winslow, “Report of the Topical Group on Wave Dark Matter for Snowmass 2021,” [arXiv:2209.08125 \[hep-ph\]](#).
- [20] C. Vafa, “The String landscape and the swampland,” [arXiv:hep-th/0509212](#).
- [21] D. Harlow and H. Ooguri, “Symmetries in quantum field theory and quantum gravity,” *Commun. Math. Phys.* **383** no. 3, (2021) 1669–1804, [arXiv:1810.05338 \[hep-th\]](#).
- [22] D. Harlow and H. Ooguri, “Constraints on Symmetries from Holography,” *Phys. Rev. Lett.* **122** no. 19, (2019) 191601, [arXiv:1810.05337 \[hep-th\]](#).
- [23] N. Arkani-Hamed, L. Motl, A. Nicolis, and C. Vafa, “The String landscape, black holes and gravity as the weakest force,” *JHEP* **06** (2007) 060, [arXiv:hep-th/0601001](#).
- [24] D. Harlow, B. Heidenreich, M. Reece, and T. Rudelius, “The Weak Gravity Conjecture: A Review,” [arXiv:2201.08380 \[hep-th\]](#).
- [25] E. Palti, “The Swampland: Introduction and Review,” *Fortsch. Phys.* **67** no. 6, (2019) 1900037, [arXiv:1903.06239 \[hep-th\]](#).
- [26] N. B. Agmon, A. Bedroya, M. J. Kang, and C. Vafa, “Lectures on the string landscape and the Swampland,” [arXiv:2212.06187 \[hep-th\]](#).
- [27] P. Draper, I. G. Garcia, and M. Reece, “Snowmass White Paper: Implications of Quantum Gravity for Particle Physics,” in 2022 *Snowmass Summer Study*. 3, 2022. [arXiv:2203.07624 \[hep-ph\]](#).
- [28] D. Tong, “Gauge Theory,” <https://www.damtp.cam.ac.uk/user/tong/gaugetheory/gt.pdf>. Lecture notes.
- [29] R. Bott and L. W. Tu, *Differential Forms in Algebraic Topology*. Graduate Texts in Mathematics. Springer New York, NY, 1982.
- [30] T. Frankel, *The Geometry of Physics: An Introduction*. Cambridge University Press, 3rd ed., 2011.
- [31] J. McNamara and M. Reece, “Reflections on Parity Breaking,” [arXiv:2212.00039 \[hep-th\]](#).

- [32] D. Simmons-Duffin, “The Conformal Bootstrap,” in *Theoretical Advanced Study Institute in Elementary Particle Physics: New Frontiers in Fields and Strings*, pp. 1–74. 2017. [arXiv:1602.07982 \[hep-th\]](#).
- [33] O. Alvarez, “Topological Quantization and Cohomology,” *Commun. Math. Phys.* **100** (1985) 279.
- [34] M. Nakahara, *Geometry, Topology and Physics*. Taylor & Francis, 2003.
- [35] O. Aharony, N. Seiberg, and Y. Tachikawa, “Reading between the lines of four-dimensional gauge theories,” *JHEP* **08** (2013) 115, [arXiv:1305.0318 \[hep-th\]](#).
- [36] E. Witten, “Dyons of Charge $e\theta/2\pi$,” *Phys. Lett. B* **86** (1979) 283–287.
- [37] F. Wilczek, “Two Applications of Axion Electrodynamics,” *Phys. Rev. Lett.* **58** (1987) 1799.
- [38] P. Galison and A. Manohar, “Two Z’s or Not Two Z’s?,” *Phys. Lett. B* **136** (1984) 279–283.
- [39] B. Holdom, “Two U(1)’s and Epsilon Charge Shifts,” *Phys. Lett. B* **166** (1986) 196–198.
- [40] J. Milnor and J. Stasheff, *Characteristic Classes*. Annals of Mathematics Studies. Princeton University Press, 1974.
- [41] A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Y. S. Tyupkin, “Pseudoparticle Solutions of the Yang-Mills Equations,” *Phys. Lett. B* **59** (1975) 85–87.
- [42] G. ’t Hooft, “Computation of the Quantum Effects Due to a Four-Dimensional Pseudoparticle,” *Phys. Rev. D* **14** (1976) 3432–3450. [Erratum: *Phys. Rev. D* **18**, 2199 (1978)].
- [43] N. Seiberg, “Modifying the Sum Over Topological Sectors and Constraints on Supergravity,” *JHEP* **07** (2010) 070, [arXiv:1005.0002 \[hep-th\]](#).
- [44] Y. Tanizaki and M. Ünsal, “Modified instanton sum in QCD and higher-groups,” *JHEP* **03** (2020) 123, [arXiv:1912.01033 \[hep-th\]](#).
- [45] C. G. Callan, Jr., R. F. Dashen, and D. J. Gross, “The Structure of the Gauge Theory Vacuum,” *Phys. Lett. B* **63** (1976) 334–340.
- [46] T. Schäfer and E. V. Shuryak, “Instantons in QCD,” *Rev. Mod. Phys.* **70** (1998) 323–426, [arXiv:hep-ph/9610451](#).
- [47] S. L. Adler, “Axial vector vertex in spinor electrodynamics,” *Phys. Rev.* **177** (1969) 2426–2438.
- [48] J. S. Bell and R. Jackiw, “A PCAC puzzle: $\pi^0 \rightarrow \gamma\gamma$ in the σ model,” *Nuovo Cim. A* **60** (1969) 47–61.
- [49] G. ’t Hooft, “Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking,” *NATO Sci. Ser. B* **59** (1980) 135–157.

- [50] H. B. Nielsen and M. Ninomiya, “Adler-Bell-Jackiw Anomaly and Weyl Fermions in Crystal,” *Phys. Lett. B* **130** (1983) 389–396.
- [51] D. B. Kaplan, “Chiral Symmetry and Lattice Fermions,” in *Les Houches Summer School: Session 93: Modern perspectives in lattice QCD: Quantum field theory and high performance computing*, pp. 223–272. 12, 2009. [arXiv:0912.2560 \[hep-lat\]](#).
- [52] S. Weinberg, *Lectures on Quantum Mechanics*. Cambridge University Press, 2015.
- [53] H. K. Dreiner, H. E. Haber, and S. P. Martin, “Two-component spinor techniques and Feynman rules for quantum field theory and supersymmetry,” *Phys. Rept.* **494** (2010) 1–196, [arXiv:0812.1594 \[hep-ph\]](#).
- [54] C. Córdova, T. T. Dumitrescu, and K. Intriligator, “Exploring 2-Group Global Symmetries,” *JHEP* **02** (2019) 184, [arXiv:1802.04790 \[hep-th\]](#).
- [55] K. Fujikawa, “Path Integral Measure for Gauge Invariant Fermion Theories,” *Phys. Rev. Lett.* **42** (1979) 1195–1198.
- [56] K. Fujikawa, “Path Integral for Gauge Theories with Fermions,” *Phys. Rev. D* **21** (1980) 2848. [Erratum: *Phys.Rev.D* 22, 1499 (1980)].
- [57] Y. Frishman, A. Schwimmer, T. Banks, and S. Yankielowicz, “The Axial Anomaly and the Bound State Spectrum in Confining Theories,” *Nucl. Phys. B* **177** (1981) 157–171.
- [58] S. R. Coleman and E. Witten, “Chiral Symmetry Breakdown in Large N Chromodynamics,” *Phys. Rev. Lett.* **45** (1980) 100.
- [59] R. Jackiw, “Charge and Mass Spectrum of Quantum Solitons,” *Gauge Theories and Modern Field Theory. Proceedings: Northeastern University, Boston, Sep 26-27, 1975* (R. L. Arnowitt and P. Nath, eds.), MIT Press, (1976) 377–401. Note: In this talk, Jackiw reported work with Goldstone that established the existence of the dyon collective coordinate.
- [60] J. Fan, K. Fraser, M. Reece, and J. Stout, “Axion Mass from Magnetic Monopole Loops,” *Phys. Rev. Lett.* **127** no. 13, (2021) 131602, [arXiv:2105.09950 \[hep-ph\]](#).
- [61] B. Heidenreich, J. McNamara, M. Montero, M. Reece, T. Rudelius, and I. Valenzuela, “Chern-Weil global symmetries and how quantum gravity avoids them,” *JHEP* **11** (2021) 053, [arXiv:2012.00009 \[hep-th\]](#).
- [62] Y. Choi, H. T. Lam, and S.-H. Shao, “Non-invertible Gauss Law and Axions,” [arXiv:2212.04499 \[hep-th\]](#).
- [63] R. Dijkgraaf and E. Witten, “Topological Gauge Theories and Group Cohomology,” *Commun. Math. Phys.* **129** (1990) 393.
- [64] C. G. Callan, Jr. and J. A. Harvey, “Anomalies and Fermion Zero Modes on Strings and Domain Walls,” *Nucl. Phys. B* **250** (1985) 427–436.

- [65] E. Witten, “On S duality in Abelian gauge theory,” *Selecta Math.* **1** (1995) 383, [arXiv:hep-th/9505186](#).
- [66] P. Deligne, P. Etingof, D. S. Freed, L. C. Jeffrey, D. Kazhdan, J. W. Morgan, D. R. Morrison, and E. Witten, eds., *Quantum fields and strings: A course for mathematicians. Vol. 1, 2.* 1999.
- [67] A. Hook and J. Huang, “A Mass for the Dual Photon,” [arXiv:2210.00015 \[hep-ph\]](#).
- [68] C. Vafa and E. Witten, “Parity Conservation in QCD,” *Phys. Rev. Lett.* **53** (1984) 535.
- [69] C. Abel *et al.*, “Measurement of the Permanent Electric Dipole Moment of the Neutron,” *Phys. Rev. Lett.* **124** no. 8, (2020) 081803, [arXiv:2001.11966 \[hep-ex\]](#).
- [70] P. Sikivie, “The Pool table analogy to axion physics,” [arXiv:hep-ph/9506229](#).
- [71] D. B. Kaplan and A. V. Manohar, “Current Mass Ratios of the Light Quarks,” *Phys. Rev. Lett.* **56** (1986) 2004.
- [72] T. Banks, Y. Nir, and N. Seiberg, “Missing (up) mass, accidental anomalous symmetries, and the strong CP problem,” in *2nd IFT Workshop on Yukawa Couplings and the Origins of Mass*, pp. 26–41. 2, 1994. [arXiv:hep-ph/9403203](#).
- [73] **Flavour Lattice Averaging Group** Collaboration, S. Aoki *et al.*, “FLAG Review 2019: Flavour Lattice Averaging Group (FLAG),” *Eur. Phys. J. C* **80** no. 2, (2020) 113, [arXiv:1902.08191 \[hep-lat\]](#).
- [74] C. Alexandrou, J. Finkenrath, L. Funcke, K. Jansen, B. Kostrzewa, F. Pittler, and C. Urbach, “Ruling Out the Massless Up-Quark Solution to the Strong **CP** Problem by Computing the Topological Mass Contribution with Lattice QCD,” *Phys. Rev. Lett.* **125** no. 23, (2020) 232001, [arXiv:2002.07802 \[hep-lat\]](#).
- [75] A. E. Nelson, “Naturally Weak CP Violation,” *Phys. Lett. B* **136** (1984) 387–391.
- [76] S. M. Barr, “Solving the Strong CP Problem Without the Peccei-Quinn Symmetry,” *Phys. Rev. Lett.* **53** (1984) 329.
- [77] K. S. Babu and R. N. Mohapatra, “A Solution to the Strong CP Problem Without an Axion,” *Phys. Rev. D* **41** (1990) 1286.
- [78] R. D. Peccei and H. R. Quinn, “Constraints Imposed by CP Conservation in the Presence of Instantons,” *Phys. Rev.* **D16** (1977) 1791–1797.
- [79] R. D. Peccei and H. R. Quinn, “CP Conservation in the Presence of Instantons,” *Phys. Rev. Lett.* **38** (1977) 1440–1443.
- [80] S. Weinberg, “A New Light Boson?,” *Phys. Rev. Lett.* **40** (1978) 223–226.
- [81] F. Wilczek, “Problem of Strong P and T Invariance in the Presence of Instantons,” *Phys. Rev. Lett.* **40** (1978) 279–282.

- [82] P. Di Vecchia and G. Veneziano, “Chiral Dynamics in the Large N Limit,” *Nucl. Phys. B* **171** (1980) 253–272.
- [83] G. Grilli di Cortona, E. Hardy, J. Pardo Vega, and G. Villadoro, “The QCD axion, precisely,” *JHEP* **01** (2016) 034, [arXiv:1511.02867 \[hep-ph\]](#).
- [84] E. Witten, “Large N Chiral Dynamics,” *Annals Phys.* **128** (1980) 363.
- [85] D. Tong, “Line Operators in the Standard Model,” *JHEP* **07** (2017) 104, [arXiv:1705.01853 \[hep-th\]](#).
- [86] D. B. Kaplan, “Opening the Axion Window,” *Nucl. Phys. B* **260** (1985) 215–226.
- [87] M. Srednicki, “Axion Couplings to Matter. 1. CP Conserving Parts,” *Nucl. Phys. B* **260** (1985) 689–700.
- [88] H. Georgi, D. B. Kaplan, and L. Randall, “Manifesting the Invisible Axion at Low-energies,” *Phys. Lett. B* **169** (1986) 73–78.
- [89] P. Svrcek and E. Witten, “Axions In String Theory,” *JHEP* **06** (2006) 051, [arXiv:hep-th/0605206](#).
- [90] P. Agrawal, J. Fan, M. Reece, and L.-T. Wang, “Experimental Targets for Photon Couplings of the QCD Axion,” *JHEP* **02** (2018) 006, [arXiv:1709.06085 \[hep-ph\]](#).
- [91] J. E. Kim, “Weak Interaction Singlet and Strong CP Invariance,” *Phys. Rev. Lett.* **43** (1979) 103.
- [92] M. A. Shifman, A. I. Vainshtein, and V. I. Zakharov, “Can Confinement Ensure Natural CP Invariance of Strong Interactions?,” *Nucl. Phys. B* **166** (1980) 493–506.
- [93] Y. Choi, H. T. Lam, and S.-H. Shao, “Noninvertible Global Symmetries in the Standard Model,” *Phys. Rev. Lett.* **129** no. 16, (2022) 161601, [arXiv:2205.05086 \[hep-th\]](#).
- [94] C. Cordova and K. Ohmori, “Noninvertible Chiral Symmetry and Exponential Hierarchies,” *Phys. Rev. X* **13** no. 1, (2023) 011034, [arXiv:2205.06243 \[hep-th\]](#).
- [95] A. R. Zhitnitsky, “On Possible Suppression of the Axion Hadron Interactions. (In Russian),” *Sov. J. Nucl. Phys.* **31** (1980) 260. [*Yad. Fiz.* 31,497(1980)].
- [96] M. Dine, W. Fischler, and M. Srednicki, “A Simple Solution to the Strong CP Problem with a Harmless Axion,” *Phys. Lett.* **104B** (1981) 199–202.
- [97] M. A. Buen-Abad, J. Fan, M. Reece, and C. Sun, “Challenges for an axion explanation of the muon $g - 2$ measurement,” *JHEP* **09** (2021) 101, [arXiv:2104.03267 \[hep-ph\]](#).
- [98] P. Agrawal, M. Nee, and M. Reig, “Axion couplings in grand unified theories,” *JHEP* **10** (2022) 141, [arXiv:2206.07053 \[hep-ph\]](#).

- [99] S. M. Barr and D. Seckel, “Planck scale corrections to axion models,” *Phys. Rev. D* **46** (1992) 539–549.
- [100] M. Kamionkowski and J. March-Russell, “Planck scale physics and the Peccei-Quinn mechanism,” *Phys. Lett. B* **282** (1992) 137–141, [arXiv:hep-th/9202003](#).
- [101] R. Holman, S. D. H. Hsu, T. W. Kephart, E. W. Kolb, R. Watkins, and L. M. Widrow, “Solutions to the strong CP problem in a world with gravity,” *Phys. Lett. B* **282** (1992) 132–136, [arXiv:hep-ph/9203206](#).
- [102] L. Randall, “Composite axion models and Planck scale physics,” *Phys. Lett. B* **284** (1992) 77–80.
- [103] E. Witten, “Some Properties of O(32) Superstrings,” *Phys. Lett. B* **149** (1984) 351–356.
- [104] K. Choi and J. E. Kim, “Harmful Axions in Superstring Models,” *Phys. Lett. B* **154** (1985) 393. [Erratum: *Phys.Lett.B* 156, 452 (1985)].
- [105] S. M. Barr, “Harmless Axions in Superstring Theories,” *Phys. Lett. B* **158** (1985) 397–400.
- [106] H.-C. Cheng and D. E. Kaplan, “Axions and a gauged Peccei-Quinn symmetry,” [arXiv:hep-ph/0103346](#).
- [107] N. Arkani-Hamed, H.-C. Cheng, P. Creminelli, and L. Randall, “Extra natural inflation,” *Phys. Rev. Lett.* **90** (2003) 221302, [arXiv:hep-th/0301218](#).
- [108] K. Choi, “A QCD axion from higher dimensional gauge field,” *Phys. Rev. Lett.* **92** (2004) 101602, [arXiv:hep-ph/0308024](#).
- [109] S. R. Coleman and E. J. Weinberg, “Radiative Corrections as the Origin of Spontaneous Symmetry Breaking,” *Phys. Rev. D* **7** (1973) 1888–1910.
- [110] Y. Hosotani, “Dynamical Mass Generation by Compact Extra Dimensions,” *Phys. Lett. B* **126** (1983) 309–313.
- [111] H.-C. Cheng, K. T. Matchev, and M. Schmaltz, “Radiative corrections to Kaluza-Klein masses,” *Phys. Rev. D* **66** (2002) 036005, [arXiv:hep-ph/0204342](#).
- [112] N. Arkani-Hamed, S. Dubovsky, A. Nicolis, and G. Villadoro, “Quantum Horizons of the Standard Model Landscape,” *JHEP* **06** (2007) 078, [arXiv:hep-th/0703067](#).
- [113] J. P. Conlon, “The QCD axion and moduli stabilisation,” *JHEP* **05** (2006) 078, [arXiv:hep-th/0602233](#).
- [114] A. Arvanitaki, S. Dimopoulos, S. Dubovsky, N. Kaloper, and J. March-Russell, “String Axiverse,” *Phys. Rev. D* **81** (2010) 123530, [arXiv:0905.4720](#) [[hep-th](#)].
- [115] I. Broeckel, M. Cicoli, A. Maharana, K. Singh, and K. Sinha, “Moduli stabilisation and the statistics of axion physics in the landscape,” *JHEP* **08** (2021) 059, [arXiv:2105.02889](#) [[hep-th](#)]. [Addendum: *JHEP* 01, 191 (2022)].

- [116] M. Demirtas, N. Gendler, C. Long, L. McAllister, and J. Moritz, “PQ Axiverse,” [arXiv:2112.04503 \[hep-th\]](#).
- [117] E. Witten, “Small instantons in string theory,” *Nucl. Phys. B* **460** (1996) 541–559, [arXiv:hep-th/9511030](#).
- [118] M. R. Douglas, “Branes within branes,” *NATO Sci. Ser. C* **520** (1999) 267–275, [arXiv:hep-th/9512077](#).
- [119] N. Lambert, C. Papageorgakis, and M. Schmidt-Sommerfeld, “Instanton Operators in Five-Dimensional Gauge Theories,” *JHEP* **03** (2015) 019, [arXiv:1412.2789 \[hep-th\]](#).
- [120] Y. Tachikawa, “Instanton operators and symmetry enhancement in 5d supersymmetric gauge theories,” *PTEP* **2015** no. 4, (2015) 043B06, [arXiv:1501.01031 \[hep-th\]](#).
- [121] P. Benetti Genolini and L. Tizzano, “Instantons, symmetries and anomalies in five dimensions,” *JHEP* **04** (2021) 188, [arXiv:2009.07873 \[hep-th\]](#).
- [122] F. Apruzzi, M. Dierigl, and L. Lin, “The fate of discrete 1-form symmetries in 6d,” *SciPost Phys.* **12** no. 2, (2022) 047, [arXiv:2008.09117 \[hep-th\]](#).
- [123] L. Bhardwaj and S. Schäfer-Nameki, “Higher-form symmetries of 6d and 5d theories,” *JHEP* **02** (2021) 159, [arXiv:2008.09600 \[hep-th\]](#).
- [124] C. Cordova, T. T. Dumitrescu, and K. Intriligator, “2-Group Global Symmetries and Anomalies in Six-Dimensional Quantum Field Theories,” *JHEP* **04** (2021) 252, [arXiv:2009.00138 \[hep-th\]](#).
- [125] M. Montero, A. M. Uranga, and I. Valenzuela, “A Chern-Simons Pandemic,” *JHEP* **07** (2017) 123, [arXiv:1702.06147 \[hep-th\]](#).
- [126] M. S. Turner, “Coherent Scalar Field Oscillations in an Expanding Universe,” *Phys. Rev. D* **28** (1983) 1243.
- [127] J. Preskill, M. B. Wise, and F. Wilczek, “Cosmology of the Invisible Axion,” *Phys. Lett.* **120B** (1983) 127–132.
- [128] M. Dine and W. Fischler, “The Not So Harmless Axion,” *Phys. Lett.* **120B** (1983) 137–141.
- [129] L. F. Abbott and P. Sikivie, “A Cosmological Bound on the Invisible Axion,” *Phys. Lett.* **120B** (1983) 133–136.
- [130] **Super-Kamiokande** Collaboration, A. Takenaka *et al.*, “Search for proton decay via $p \rightarrow e^+\pi^0$ and $p \rightarrow \mu^+\pi^0$ with an enlarged fiducial volume in Super-Kamiokande I-IV,” *Phys. Rev. D* **102** no. 11, (2020) 112011, [arXiv:2010.16098 \[hep-ex\]](#).
- [131] H. Georgi and S. L. Glashow, “Unity of All Elementary Particle Forces,” *Phys. Rev. Lett.* **32** (1974) 438–441.

- [132] S. Weinberg, “Baryon and Lepton Nonconserving Processes,” *Phys. Rev. Lett.* **43** (1979) 1566–1570.
- [133] D. Harlow, “Jerusalem Lectures on Black Holes and Quantum Information,” *Rev. Mod. Phys.* **88** (2016) 015002, [arXiv:1409.1231 \[hep-th\]](#).
- [134] Y. Kats, L. Motl, and M. Padi, “Higher-order corrections to mass-charge relation of extremal black holes,” *JHEP* **12** (2007) 068, [arXiv:hep-th/0606100](#).
- [135] J. D. Bekenstein, “Black holes and entropy,” *Phys. Rev. D* **7** (1973) 2333–2346.
- [136] J. M. Bardeen, B. Carter, and S. W. Hawking, “The Four laws of black hole mechanics,” *Commun. Math. Phys.* **31** (1973) 161–170.
- [137] J. D. Bekenstein, “Generalized second law of thermodynamics in black hole physics,” *Phys. Rev. D* **9** (1974) 3292–3300.
- [138] S. W. Hawking, “Black hole explosions,” *Nature* **248** (1974) 30–31.
- [139] S. Hawking, “Particle Creation by Black Holes,” *Commun. Math. Phys.* **43** (1975) 199–220. [Erratum: *Commun. Math. Phys.* 46, 206 (1976)].
- [140] G. W. Gibbons, “Vacuum Polarization and the Spontaneous Loss of Charge by Black Holes,” *Commun. Math. Phys.* **44** (1975) 245–264.
- [141] G. Johnson, “Tunnelling of Charged Particles from Black Holes,” *JHEP* **03** (2020) 038, [arXiv:1911.12379 \[hep-th\]](#).
- [142] Y. B. Zeldovich, “A New Type of Radioactive Decay: Gravitational Annihilation of Baryons,” *Phys. Lett. A* **59** (1976) 254.
- [143] T. Banks and N. Seiberg, “Symmetries and Strings in Field Theory and Gravity,” *Phys. Rev. D* **83** (2011) 084019, [arXiv:1011.5120 \[hep-th\]](#).
- [144] J. D. Bekenstein, “A Universal Upper Bound on the Entropy to Energy Ratio for Bounded Systems,” *Phys. Rev. D* **23** (1981) 287.
- [145] R. Bousso, “A Covariant entropy conjecture,” *JHEP* **07** (1999) 004, [arXiv:hep-th/9905177](#).
- [146] J. Polchinski, “Monopoles, duality, and string theory,” *Int. J. Mod. Phys. A* **19S1** (2004) 145–156, [arXiv:hep-th/0304042](#).
- [147] T. Rudelius and S.-H. Shao, “Topological Operators and Completeness of Spectrum in Discrete Gauge Theories,” *JHEP* **12** (2020) 172, [arXiv:2006.10052 \[hep-th\]](#).
- [148] B. Heidenreich, J. McNamara, M. Montero, M. Reece, T. Rudelius, and I. Valenzuela, “Non-invertible global symmetries and completeness of the spectrum,” *JHEP* **09** (2021) 203, [arXiv:2104.07036 \[hep-th\]](#).

- [149] E. Lake, “Higher-form symmetries and spontaneous symmetry breaking,” [arXiv:1802.07747 \[hep-th\]](#).
- [150] D. M. Hofman and N. Iqbal, “Goldstone modes and photonization for higher form symmetries,” *SciPost Phys.* **6** no. 1, (2019) 006, [arXiv:1802.09512 \[hep-th\]](#).
- [151] N. Iqbal and J. McGreevy, “Mean string field theory: Landau-Ginzburg theory for 1-form symmetries,” [arXiv:2106.12610 \[hep-th\]](#).
- [152] D. Harlow and E. Shaghoulian, “Global symmetry, Euclidean gravity, and the black hole information problem,” *JHEP* **04** (2021) 175, [arXiv:2010.10539 \[hep-th\]](#).
- [153] T. Banks and L. J. Dixon, “Constraints on String Vacua with Space-Time Supersymmetry,” *Nucl. Phys. B* **307** (1988) 93–108.
- [154] E. Witten, “Anti-de Sitter space and holography,” *Adv. Theor. Math. Phys.* **2** (1998) 253–291, [arXiv:hep-th/9802150](#).
- [155] S. B. Giddings and A. Strominger, “Loss of Incoherence and Determination of Coupling Constants in Quantum Gravity,” *Nucl. Phys. B* **307** (1988) 854–866.
- [156] L. F. Abbott and M. B. Wise, “Wormholes and Global Symmetries,” *Nucl. Phys. B* **325** (1989) 687–704.
- [157] S. R. Coleman and K.-M. Lee, “Wormholes Made Without Massless Matter Fields,” *Nucl. Phys. B* **329** (1990) 387–409.
- [158] R. Kallosh, A. D. Linde, D. A. Linde, and L. Susskind, “Gravity and global symmetries,” *Phys. Rev. D* **52** (1995) 912–935, [arXiv:hep-th/9502069](#).
- [159] Y. Chen and H. W. Lin, “Signatures of global symmetry violation in relative entropies and replica wormholes,” *JHEP* **03** (2021) 040, [arXiv:2011.06005 \[hep-th\]](#).
- [160] P.-S. Hsin, L. V. Iliesiu, and Z. Yang, “A violation of global symmetries from replica wormholes and the fate of black hole remnants,” *Class. Quant. Grav.* **38** no. 19, (2021) 194004, [arXiv:2011.09444 \[hep-th\]](#).
- [161] A. Belin, J. De Boer, P. Nayak, and J. Sonner, “Charged eigenstate thermalization, Euclidean wormholes and global symmetries in quantum gravity,” *SciPost Phys.* **12** no. 2, (2022) 059, [arXiv:2012.07875 \[hep-th\]](#).
- [162] I. Bah, Y. Chen, and J. Maldacena, “Estimating global charge violating amplitudes from wormholes,” [arXiv:2212.08668 \[hep-th\]](#).
- [163] A. de la Fuente, P. Saraswat, and R. Sundrum, “Natural Inflation and Quantum Gravity,” *Phys. Rev. Lett.* **114** no. 15, (2015) 151303, [arXiv:1412.3457 \[hep-th\]](#).
- [164] C. Cheung and G. N. Remmen, “Naturalness and the Weak Gravity Conjecture,” *Phys. Rev. Lett.* **113** (2014) 051601, [arXiv:1402.2287 \[hep-ph\]](#).

- [165] B. Heidenreich, M. Reece, and T. Rudelius, “Sharpening the Weak Gravity Conjecture with Dimensional Reduction,” *JHEP* **02** (2016) 140, [arXiv:1509.06374 \[hep-th\]](#).
- [166] B. Heidenreich, M. Reece, and T. Rudelius, “Evidence for a sublattice weak gravity conjecture,” *JHEP* **08** (2017) 025, [arXiv:1606.08437 \[hep-th\]](#).
- [167] M. Montero, G. Shiu, and P. Soler, “The Weak Gravity Conjecture in three dimensions,” *JHEP* **10** (2016) 159, [arXiv:1606.08438 \[hep-th\]](#).
- [168] S. Andriolo, D. Junghans, T. Noumi, and G. Shiu, “A Tower Weak Gravity Conjecture from Infrared Consistency,” *Fortsch. Phys.* **66** no. 5, (2018) 1800020, [arXiv:1802.04287 \[hep-th\]](#).
- [169] G. Veneziano, “Large N bounds on, and compositeness limit of, gauge and gravitational interactions,” *JHEP* **06** (2002) 051, [arXiv:hep-th/0110129](#).
- [170] G. Dvali, “Black Holes and Large N Species Solution to the Hierarchy Problem,” *Fortsch. Phys.* **58** (2010) 528–536, [arXiv:0706.2050 \[hep-th\]](#).
- [171] G. Dvali and M. Redi, “Black Hole Bound on the Number of Species and Quantum Gravity at LHC,” *Phys. Rev. D* **77** (2008) 045027, [arXiv:0710.4344 \[hep-th\]](#).
- [172] C. Cheung, J. Liu, and G. N. Remmen, “Proof of the Weak Gravity Conjecture from Black Hole Entropy,” *JHEP* **10** (2018) 004, [arXiv:1801.08546 \[hep-th\]](#).
- [173] Y. Hamada, T. Noumi, and G. Shiu, “Weak Gravity Conjecture from Unitarity and Causality,” *Phys. Rev. Lett.* **123** no. 5, (2019) 051601, [arXiv:1810.03637 \[hep-th\]](#).
- [174] A. M. Charles, “The Weak Gravity Conjecture, RG Flows, and Supersymmetry,” [arXiv:1906.07734 \[hep-th\]](#).
- [175] N. Arkani-Hamed, Y.-t. Huang, J.-Y. Liu, and G. N. Remmen, “Causality, unitarity, and the weak gravity conjecture,” *JHEP* **03** (2022) 083, [arXiv:2109.13937 \[hep-th\]](#).
- [176] S.-J. Lee, W. Lerche, and T. Weigand, “Tensionless Strings and the Weak Gravity Conjecture,” *JHEP* **10** (2018) 164, [arXiv:1808.05958 \[hep-th\]](#).
- [177] S.-J. Lee, W. Lerche, and T. Weigand, “Modular Fluxes, Elliptic Genera, and Weak Gravity Conjectures in Four Dimensions,” *JHEP* **08** (2019) 104, [arXiv:1901.08065 \[hep-th\]](#).
- [178] S.-J. Lee, W. Lerche, and T. Weigand, “Emergent strings, duality and weak coupling limits for two-form fields,” *JHEP* **02** (2022) 096, [arXiv:1904.06344 \[hep-th\]](#).
- [179] M. Alim, B. Heidenreich, and T. Rudelius, “The Weak Gravity Conjecture and BPS Particles,” *Fortsch. Phys.* **69** no. 11-12, (2021) 2100125, [arXiv:2108.08309 \[hep-th\]](#).
- [180] N. Gendler, B. Heidenreich, L. McAllister, J. Moritz, and T. Rudelius, “Moduli Space Reconstruction and Weak Gravity,” [arXiv:2212.10573 \[hep-th\]](#).

- [181] D. Harlow, “Wormholes, Emergent Gauge Fields, and the Weak Gravity Conjecture,” *JHEP* **01** (2016) 122, [arXiv:1510.07911 \[hep-th\]](#).
- [182] B. Heidenreich, M. Reece, and T. Rudelius, “The Weak Gravity Conjecture and Emergence from an Ultraviolet Cutoff,” *Eur. Phys. J. C* **78** no. 4, (2018) 337, [arXiv:1712.01868 \[hep-th\]](#).
- [183] H. Ooguri and C. Vafa, “On the Geometry of the String Landscape and the Swampland,” *Nucl. Phys. B* **766** (2007) 21–33, [arXiv:hep-th/0605264](#).
- [184] M. Etheredge, B. Heidenreich, S. Kaya, Y. Qiu, and T. Rudelius, “Sharpening the Distance Conjecture in diverse dimensions,” *JHEP* **12** (2022) 114, [arXiv:2206.04063 \[hep-th\]](#).
- [185] B. Heidenreich, M. Reece, and T. Rudelius, “Emergence of Weak Coupling at Large Distance in Quantum Gravity,” *Phys. Rev. Lett.* **121** no. 5, (2018) 051601, [arXiv:1802.08698 \[hep-th\]](#).
- [186] T. W. Grimm, E. Palti, and I. Valenzuela, “Infinite Distances in Field Space and Massless Towers of States,” *JHEP* **08** (2018) 143, [arXiv:1802.08264 \[hep-th\]](#).
- [187] J. Stout, “Infinite Distance Limits and Information Theory,” [arXiv:2106.11313 \[hep-th\]](#).
- [188] J. Stout, “Infinite Distances and Factorization,” [arXiv:2208.08444 \[hep-th\]](#).
- [189] C. Cordova, K. Ohmori, and T. Rudelius, “Generalized symmetry breaking scales and weak gravity conjectures,” *JHEP* **11** (2022) 154, [arXiv:2202.05866 \[hep-th\]](#).
- [190] S.-J. Lee, W. Lerche, and T. Weigand, “Emergent strings from infinite distance limits,” *JHEP* **02** (2022) 190, [arXiv:1910.01135 \[hep-th\]](#).
- [191] R. Álvarez-García, D. Kläwer, and T. Weigand, “Membrane limits in quantum gravity,” *Phys. Rev. D* **105** no. 6, (2022) 066024, [arXiv:2112.09136 \[hep-th\]](#).
- [192] S. Weinberg, *The quantum theory of fields. Vol. 2: Modern applications*. Cambridge University Press, 8, 2013.
- [193] C. Q. Geng and R. E. Marshak, “Uniqueness of Quark and Lepton Representations in the Standard Model From the Anomalies Viewpoint,” *Phys. Rev. D* **39** (1989) 693.
- [194] J. A. Minahan, P. Ramond, and R. C. Warner, “A Comment on Anomaly Cancellation in the Standard Model,” *Phys. Rev. D* **41** (1990) 715.
- [195] S. Y. Li and W. Taylor, “Large U(1) charges from flux breaking in 4D F-theory models,” *JHEP* **02** (2023) 186, [arXiv:2211.11768 \[hep-th\]](#).
- [196] B. W. Lee, C. Quigg, and H. B. Thacker, “Weak Interactions at Very High-Energies: The Role of the Higgs Boson Mass,” *Phys. Rev. D* **16** (1977) 1519.

- [197] X.-F. Wu, S.-B. Zhang, H. Gao, J.-J. Wei, Y.-C. Zou, W.-H. Lei, B. Zhang, Z.-G. Dai, and P. Mészáros, “Constraints on the Photon Mass with Fast Radio Bursts,” *Astrophys. J. Lett.* **822** no. 1, (2016) L15, [arXiv:1602.07835 \[astro-ph.HE\]](#).
- [198] L. Bonetti, J. Ellis, N. E. Mavromatos, A. S. Sakharov, E. K. G. Sarkisyan-Grinbaum, and A. D. A. M. Spallicci, “Photon Mass Limits from Fast Radio Bursts,” *Phys. Lett. B* **757** (2016) 548–552, [arXiv:1602.09135 \[astro-ph.HE\]](#).
- [199] M. Reece, “Photon Masses in the Landscape and the Swampland,” *JHEP* **07** (2019) 181, [arXiv:1808.09966 \[hep-th\]](#).
- [200] S. Davidson, S. Hannestad, and G. Raffelt, “Updated bounds on millicharged particles,” *JHEP* **05** (2000) 003, [arXiv:hep-ph/0001179](#).
- [201] N. Craig and I. Garcia Garcia, “Rescuing Massive Photons from the Swampland,” *JHEP* **11** (2018) 067, [arXiv:1810.05647 \[hep-th\]](#).
- [202] M. Goodsell, J. Jaeckel, J. Redondo, and A. Ringwald, “Naturally Light Hidden Photons in LARGE Volume String Compactifications,” *JHEP* **11** (2009) 027, [arXiv:0909.0515 \[hep-ph\]](#).
- [203] M. Cicoli, M. Goodsell, J. Jaeckel, and A. Ringwald, “Testing String Vacua in the Lab: From a Hidden CMB to Dark Forces in Flux Compactifications,” *JHEP* **07** (2011) 114, [arXiv:1103.3705 \[hep-th\]](#).
- [204] T. A. Wagner, S. Schlamminger, J. H. Gundlach, and E. G. Adelberger, “Torsion-balance tests of the weak equivalence principle,” *Class. Quant. Grav.* **29** (2012) 184002, [arXiv:1207.2442 \[gr-qc\]](#).
- [205] J. Heeck, “Unbroken B – L symmetry,” *Phys. Lett. B* **739** (2014) 256–262, [arXiv:1408.6845 \[hep-ph\]](#).
- [206] A. de Gouvea, W.-C. Huang, and J. Jenkins, “Pseudo-Dirac Neutrinos in the New Standard Model,” *Phys. Rev. D* **80** (2009) 073007, [arXiv:0906.1611 \[hep-ph\]](#).
- [207] J. McNamara and C. Vafa, “Cobordism Classes and the Swampland,” [arXiv:1909.10355 \[hep-th\]](#).
- [208] J. E. Kim, “The Strong CP Problem in Orbifold Compactifications and an $SU(3) \times SU(2) \times U(1)^n$ Model,” *Phys. Lett. B* **207** (1988) 434–440.
- [209] G. Honecker and W. Staessens, “On axionic dark matter in Type IIA string theory,” *Fortsch. Phys.* **62** (2014) 115–151, [arXiv:1312.4517 \[hep-th\]](#).
- [210] K. Choi, K. S. Jeong, and M.-S. Seo, “String theoretic QCD axions in the light of PLANCK and BICEP2,” *JHEP* **07** (2014) 092, [arXiv:1404.3880 \[hep-th\]](#).
- [211] E. I. Buchbinder, A. Constantin, and A. Lukas, “Heterotic QCD axion,” *Phys. Rev. D* **91** no. 4, (2015) 046010, [arXiv:1412.8696 \[hep-th\]](#).

- [212] M. Gutperle and W. Sabra, “Instantons and wormholes in Minkowski and (A)dS spaces,” *Nucl. Phys. B* **647** (2002) 344–356, [arXiv:hep-th/0206153](#).
- [213] E. Bergshoeff, A. Collinucci, U. Gran, D. Roest, and S. Vandoren, “Non-extremal instantons and wormholes in string theory,” *Fortsch. Phys.* **53** (2005) 990–996, [arXiv:hep-th/0412183](#).
- [214] E. Bergshoeff, A. Collinucci, U. Gran, D. Roest, and S. Vandoren, “Non-extremal D-instantons,” *JHEP* **10** (2004) 031, [arXiv:hep-th/0406038](#).
- [215] A. Hebecker, P. Mangat, S. Theisen, and L. T. Witkowski, “Can Gravitational Instantons Really Constrain Axion Inflation?,” *JHEP* **02** (2017) 097, [arXiv:1607.06814](#) [[hep-th](#)].
- [216] M. J. Dolan, P. Draper, J. Kozaczk, and H. Patel, “Transplanckian Censorship and Global Cosmic Strings,” *JHEP* **04** (2017) 133, [arXiv:1701.05572](#) [[hep-th](#)].
- [217] B. Heidenreich, M. Reece, and T. Rudelius, “The Weak Gravity Conjecture and axion strings,” *JHEP* **11** (2021) 004, [arXiv:2108.11383](#) [[hep-th](#)].
- [218] C. F. Cota, A. Mininno, T. Weigand, and M. Wiesner, “The asymptotic Weak Gravity Conjecture for open strings,” *JHEP* **11** (2022) 058, [arXiv:2208.00009](#) [[hep-th](#)].
- [219] S. Kaya and T. Rudelius, “Higher-group symmetries and weak gravity conjecture mixing,” *JHEP* **07** (2022) 040, [arXiv:2202.04655](#) [[hep-th](#)].
- [220] S. Fichtel and P. Saraswat, “Approximate Symmetries and Gravity,” *JHEP* **01** (2020) 088, [arXiv:1909.02002](#) [[hep-th](#)].
- [221] T. Daus, A. Hebecker, S. Leonhardt, and J. March-Russell, “Towards a Swampland Global Symmetry Conjecture using weak gravity,” *Nucl. Phys. B* **960** (2020) 115167, [arXiv:2002.02456](#) [[hep-th](#)].
- [222] S. Cecotti and C. Vafa, “Theta-problem and the String Swampland,” [arXiv:1808.03483](#) [[hep-th](#)].
- [223] T. W. Grimm, S. Lanza, and T. van Vuren, “Global symmetry-breaking and generalized theta-terms in Type IIB EFTs,” [arXiv:2211.11769](#) [[hep-th](#)].
- [224] E. Silverstein and A. Westphal, “Monodromy in the CMB: Gravity Waves and String Inflation,” *Phys. Rev. D* **78** (2008) 106003, [arXiv:0803.3085](#) [[hep-th](#)].
- [225] S. Lanza, F. Marchesano, L. Martucci, and I. Valenzuela, “Swampland Conjectures for Strings and Membranes,” *JHEP* **02** (2021) 006, [arXiv:2006.15154](#) [[hep-th](#)].
- [226] S. Lanza, F. Marchesano, L. Martucci, and I. Valenzuela, “The EFT stringy viewpoint on large distances,” *JHEP* **09** (2021) 197, [arXiv:2104.05726](#) [[hep-th](#)].
- [227] R. Alarcon *et al.*, “Electric dipole moments and the search for new physics,” in *Snowmass 2021*. 3, 2022. [arXiv:2203.08103](#) [[hep-ph](#)].

- [228] A. Baldini *et al.*, “A submission to the 2020 update of the European Strategy for Particle Physics on behalf of the COMET, MEG, Mu2e and Mu3e collaborations,”
[arXiv:1812.06540](#) [hep-ex].