Investigations on Black Holes, Cosmic Censorship, and Scalar Field Dark Matter Cosmology

by

James Wheeler

Department of Physics
Duke University

Date: _____

Approved:

_____
Hubert Bray, Supervisor

_____
Paul Aspinwall

_____
Roxanne Springer

_____
Michael Troxel

_____
Daniel Scolnic

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Physics
in the Graduate School of Duke University
2023

Abstract

Investigations on Black Holes, Cosmic Censorship, and Scalar Field Dark Matter Cosmology

by

James Wheeler

Department of Physics
Duke University

Date: _____

Approved:

_____
Hubert Bray, Supervisor

_____
Paul Aspinwall

_____
Roxanne Springer

_____
Michael Troxel

_____
Daniel Scolnic

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Physics in the Graduate School of Duke University
2023

Abstract

Einstein's General Theory of Relativity sits among the pillars of modern physics as the means by which we describe the universe across an enormous range of scales. This theory has furnished our most robust understanding of the origins of the universe, the dynamics of astronomical objects, and the fundamental structure of space and time. For all of general relativity's successes, however, a wide array of deep questions remain. Its sophisticated mathematical structure renders foundational questions surrounding the extent to which the theory is well-posed difficult to answer (and indeed, difficult to ask), and consistent systematic discrepancies between the universe's dynamics and what the theory leads us to expect given our knowledge of the structure of matter leave us puzzling over which of general relativity and particle physics is more incomplete.

This thesis seeks to explore a small cross-section of the fundamental challenges faced by general relativity through two distinct avenues. The first is an investigation of the cosmological properties of scalar field dark matter, often informed by the fact that it may arise through a minor geometric adjustment to the core structure of the theory. The novel cosmological phenomena under consideration primarily include a dark-matter dominated regime in the early universe and a modification to the standard gravitational redshift, and we generally find that (though they are not ruled out) there is little compelling evidence for either amongst the empirical probes considered herein, namely the anisotropies in the cosmic microwave background radiation as

measured by the Planck collaboration and a six-year time-domain survey of spectra across many astronomical sources completed by the Anglo-Australian Telescope. The second is a reflection on both the challenge and posing of the Weak Cosmic Censorship Conjecture, the problem of whether singularities in general relativity must generically reside within black holes. We demonstrate that violating singularities are generic within a particular class of spherically symmetric spacetimes, the Vaidya spacetimes, and this reflection leads us to the development of a novel characterization of the phenomenon of black holes, utilized to formulate a more comprehensive rigorous statement of weak cosmic censorship.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to offer thanks to a great many people for their encouragement, support, and guidance through the this arduous process. Each of my parents, John and Cheryl Wheeler, and my siblings, John, Anna, Jeremiah, Joshua, Sarah, Joseph, and Rachael, has offered unwavering support and many weekends of respite over these six years, which I've greatly appreciated.

I would like to thank the many professors who have guided me as a student in their courses– I have learned a great deal both from the curated content of their lectures and materials as well as their examples as academics. I would like to thank Dr. Ronen Plesser and Dr. Michael Troxel for taking continued interest in my work and offering constructive comments, thoughts, and support over the past few years. I would like to thank Dr. Paul Aspinwall for his thoughtful questions, comments, and instruction across many interactions, as well as his repeated assistance in sorting out my funding and instruction in the math department. I would like to thank my committee at large for their careful evaluation of this thesis and my achievements. I would like to thank Dr. Arya Roy for helping me to grow extensively as an instructor through several years of working together in a myriad of introductory physics courses.

I would like to thank my advisor, Dr. Hubert Bray, for many hundreds of hours of thoughtful discussion on all manner of topics, academic and otherwise, in our five years of working together. His authoritative perspective and guidance have been utterly indispensable to my development as both a physicist and mathematician, as

an academic seeking to tread the line between these intimately connected subjects, and as a human hoping to positively impact my domain. For all the challenges that this inter-departmental endeavor has posed, I can hardly imagine a more fitting or effective advisor. I would like to thank our broader group, including Dr. Ben Hamm, Dr. Sven Hirsch, Dr. Yiyue Zhang, Dr. Demetre Kazaras, and Dr. Marcus Khuri, for their constructive collaboration and many insightful thoughts and probing questions surrounding my work, even though it was often disjoint from their own.

Finally, I would like to thank my loving partner, Willa Papanikolas, for her steadfast and stalwart support and encouragement, especially in these past few months. I am ever grateful for the buttress she provided whenever my stamina waned. Of course, I would be remiss to omit the essential contributions of both Callie and Luna, who presided over the compilation of much of this thesis.

1

Introduction

1.1 Conventions and Notation

This thesis will work extensively in the domains of general relativity and differential geometry. While we will assume a basic familiarity with these domains, here we establish our notational conventions and briefly review some core concepts. While a couple of geometric terms are more explicitly defined as they arise in Section 1.2, we overview some of their relevant features here in 1.1.2.

1.1.1 Units

We will universally utilize natural units wherein $\hbar = c = G = k_b = 1$ for simplicity of presentation and manipulation of equations. For example, we will frequently work with the mass parameter m associated to a scalar field, most commonly referred to in eV, and we would like to consider this equivalent to the associated time- and distance-frequencies $\frac{m}{\hbar}$ and $\frac{m}{\hbar c}$. While this system of units in principle allows for nearly all quantities of interest in this work to be unambiguously expressed without dimension, we will often refer to them as multiples of common quantities appropriate to the situation at hand (1 eV, 1 year, 1 lightyear, etc.) to provide a sense of scale.

1.1.2 Geometry

The bulk of this work will be concerned with manipulations taking place on smooth manifolds, generally denoted by M and given dimension n, in both coordinate-based and coordinate-invariant contexts. A useful reference for the geometrical parlance we employ is O'Neil's *Semi-Riemannian Geometry With Applications to Relativity* [101], with secondarily recommend references including do Carmo [44] and Wald [137].

Tangent Vectors and Vector Fields

To begin, a coordinate chart $\phi : U \to V$, with $U \subset M$ and $V \subset \mathbb{R}^n$, around a point $p \in U$ induces a description of the manifold's tangent space T_pM in terms of tangent vectors to $\mathbb{R}^n$ in $T_{\phi(p)}\mathbb{R}^n \cong \mathbb{R}^n$. In particular, denoting the canonical basis of $\mathbb{R}^n$ by $\{e_i\}_{i=1}^n$, we will refer to the associated basis vectors of T_pM induced by push forward under ϕ^{-1} as ∂_i or $\frac{\partial}{\partial x^i}$ interchangeably, explicitly defined by

$$\partial_i = \frac{\partial}{\partial x^i} := (\phi^{-1})_*(e_i). \tag{1.1}$$

The associated dual basis elements for T_x^*M are denoted dx^i, defined by

$$dx^i(\partial_j) = \delta^i_j. \tag{1.2}$$

These allow the description of an arbitrary tangent vector $v \in T_pM$ or dual vector $\omega \in T_p^*M$ in terms of its *coordinate components*

$$v = v^i\partial_i, \qquad \omega = \omega_i dx^i, \tag{1.3}$$

where the Einstein summation convention is employed, as it will be throughout this work.

A tangent vector $v \in T_pM$ provides a notion of a directional derivative at p (in the direction of v) applicable to real-valued functions, returning a number $v(f)$ for

a function $f : M \to \mathbb{R}$. This may be computed in coordinates according to

$$v(f) = v^i \partial_i(f), \tag{1.4}$$

where $\partial_i f$ is defined as the usual partial derivative of $f \circ \phi^{-1}$, i.e. the function expressed in terms of the $\mathbb{R}^n$ coordinates. In fact, a tangent vector is entirely characterized by its action on all possible functions.

A *vector field* $V \in \Gamma(TM)$ is a section of the tangent bundle, or a smooth association to each $p \in M$ a tangent vector $V\big|_p \in T_pM$, and may also be described locally via coordinate components: $V = V^k \partial_k$, where now each V^k is a function on the domain of the coordinate chart. We denote the set of vector fields on M by $\mathfrak{X}(M)$. A vector field may also act on functions, now returning another function $V(f)$ defined by $(V(f))(p) = V\big|_p(f)$. The *Lie bracket* $[X, Y]$ between two vector fields $X, Y \in \mathfrak{X}(M)$ is defined as that vector field whose action on functions is given by

$$[X, Y](f) := X(Y(f)) - Y(X(f)). \tag{1.5}$$

General Tensors

More generally, a (p, q)-*tensor* at $x \in M$ is a multilinear map

$$(T_xM)^{\otimes q} \to (T_xM)^{\otimes p}, \tag{1.6}$$

or equivalently an element of $(T_xM)^{\otimes p} \otimes (T_x^*M)^{\otimes q}$. A (p, q)-tensor field (often referred to as simply a "tensor") T may also be described via coordinate component functions according to

$$T = T^{i_1 \cdots i_p}_{j_1 \cdots j_q} \cdot \partial_{i_1} \otimes \cdots \otimes \partial_{i_p} \otimes dx^{j_1} \otimes \cdots \otimes dx^{j_q}, \tag{1.7}$$

or equivalently

$$T(\partial_{j_1}, \cdots, \partial_{j_q}) = T^{i_1 \cdots i_p}_{j_1 \cdots j_q} \cdot \partial_{i_1} \otimes \cdots \otimes \partial_{i_p}. \tag{1.8}$$

A useful shorthand when working with tensor components is to refer to their coordinate partial derivatives via the addition of lower indices, demarcated from the tensorial indices via a comma:

$$T^{i_1\cdots i_p}_{j_1\cdots j_q,k} := \partial_k \left(T^{i_1\cdots i_p}_{j_1\cdots j_q} \right). \tag{1.9}$$

In any event, we will often prefer a more mathematical parlance, whereby we work with a tensor T directly rather than its coordinate component functions $T^{i_1\cdots i_p}_{j_1\cdots j_q}$ whenever possible. We primarily encounter tensors with $p = 0$ or $p = 1$, in which case T takes in q vectors and returns either a number or a vector (or a function or vector field, if we have put vector fields into a tensor field).

Metrics

Our smooth manifolds will often be endowed with a *metric*, generally denoted g, a symmetric and nondegenerate $(0, 2)$-tensor field. We are most concerned with *Lorentzian* metrics, which we take to have signature $(-1, 1, 1, 1)$. We often use the inner product notation $\langle \cdot, \cdot \rangle$ for the action of the metric, so that

$$\langle X, Y \rangle := g(X, Y) \tag{1.10}$$

(for vector fields X and Y). This notation can also be more generally applied to two tensors of the same type, indicating the number obtained upon metrically contracting all corresponding indices. A metric is sometimes specified via the notation ds^2, e.g.

$$\begin{aligned} & ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 \\ \iff \quad & g = -dt \otimes dt + dx \otimes dx + dy \otimes dy + dz \otimes dz. \end{aligned} \tag{1.11}$$

In coordinates, it is sometimes useful to think of the metric components $g_{ij} = g(\partial_i, \partial_j)$ as comprising a matrix– the metric with raised indices g^{ij} is then defined as the entries in the inverse matrix to (g_{ij}). The notation $|g|$ refers to the determinate of this metric matrix in the coordinate system at hand. A metric induces a canonical

isomorphism between T_pM and T_p^*M, so that a tangent vector $v \in T_pM$ has an associated *metric dual* covector $v^* \in T_p^*M$ defined by $v^*(w) := \langle v, w \rangle$, and similarly in reverse: $\langle \omega^*, v \rangle := \omega(v)$ for $\omega \in T_p^*M$. More generally, this can be used to adjust the type of a (p, q)-tensor to a (k, m)-tensor with $p + q = k + m$. This is referred to as "lowering" or "raising" indices, taken quite literally in coordinates, wherein the raising or lowering of an index is achieved by contracting it with g_{ij} or g^{ij}.

Connections

Coordinate-invariant differentiation of tensors is achieved via an *affine connection* (Definition 2, typically abbreviated to just "connection") $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$, where $\nabla_X Y$ represents the rate of change of Y in the direction of X. While a connection takes in two vector fields and returns another, it is not a tensor because the map is not well-defined point-wise: $(\nabla_X Y)\big|_p$ depends on the behavior of Y in a neighborhood of p rather than just $Y\big|_p$, as expected of a rate of change. Though nominally restricted to vector fields, a connection can be uniquely extended (via a Leibniz rule with respect to the tensor product and invariance under contraction) to act on general tensors as well, giving meaning to $\nabla_X T$ for any tensor field T.

Though not a tensor, connections still admit a coordinate description via the *Christoffel symbols*, defined according to

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^{\ k} \partial_k. \tag{1.12}$$

Alternatively, in the presence of a metric we may lower the k index:

$$\Gamma_{ijk} := \langle \nabla_{\partial_i} \partial_j, \partial_k \rangle. \tag{1.13}$$

In coordinates, the action of the connection takes the form

$$\begin{aligned}\nabla_X Y &= \nabla_X(Y^j\partial_j) = X(Y^j)\partial_j + Y^j\nabla_X\partial_j \\ &= X^i Y^j{}_{,i}\partial_j + X^i Y^j \Gamma_{ij}{}^k \partial_k \\ &= \left[X^i Y^k{}_{,i} + X^i Y^j \Gamma_{ij}{}^k\right]\partial_k. \end{aligned} \tag{1.14}$$

When both inputs are the tangent vector to a curve $\gamma : I \to M$ (with $I \subset \mathbb{R}$ an interval), the condition that $\nabla_{\dot\gamma}\dot\gamma = 0$ yields the *geodesic equation*

$$\ddot\gamma^k + \dot\gamma^i\dot\gamma^j\Gamma_{ij}{}^k = 0. \tag{1.15}$$

A curve satisfying this is dubbed a *geodesic*– it is non-accelerating according to the coordinate-invariant notion of acceleration induced by ∇.

In the presence of a metric, a connection of particular significance is the *Levi-Civita connection* (Definition 3). We denote this particular connection by $\overline{\nabla}$, and its Christoffel symbols by $\overline{\Gamma}$. This connection is determined entirely by the metric, and it can be described in coordinates by

$$\overline{\Gamma}_{ijk} = \frac{1}{2}\left[g_{ik,j} + g_{jk,i} - g_{ij,k}\right]. \tag{1.16}$$

This is the unique connections which is both *metric compatible*, meaning $\overline{\nabla}_X g = 0$, and *torsion-free*, meaning $\overline{\nabla}_X Y - \overline{\nabla}_Y X = [X, Y]$.

Curvature

With the Levi-Civita connection in hand, one can construct the *curvature* of the metric. The most general object in this vein is the *Riemann curvature tensor* Rm, given as a $(1, 3)$-tensor by

$$\mathrm{Rm}(X, Y)Z = \overline{\nabla}_X\overline{\nabla}_Y Z - \overline{\nabla}_Y\overline{\nabla}_X Z - \overline{\nabla}_{[X,Y]}Z. \tag{1.17}$$

It is sometimes also useful to consider this a $(0, 4)$-tensor given by

$$\mathrm{Rm}(X, Y, Z, W) = \langle \mathrm{Rm}(X, Y)Z, W\rangle. \tag{1.18}$$

In particular, from this description we see that contracting over the first and third inputs leaves us with a $(0,2)$-tensor, the *Ricci tensor* Ric:

$$\mathrm{Ric}(X,Y) := g^{ij}\mathrm{Rm}(\partial_i, X, \partial_j, Y). \tag{1.19}$$

The *scalar curvature* R, a function on M, is what remains upon contracting over the final two indices:

$$R := g^{ij}\mathrm{Ric}(\partial_i, \partial_j). \tag{1.20}$$

The *Einstein curvature tensor* G is now the $(0,2)$-tensor defined according to

$$G(X,Y) := \mathrm{Ric}(X,Y) - \frac{R}{2}g(X,Y) \tag{1.21}$$

While one can construct similar tensors from any connection, and we will work with other connections, our curvatures are always built out of the Levi-Civita connection.

Differential Forms and Integration

A class of tensors of interest is comprised of the fully antisymmetric $(0,k)$-tensor fields. These are the *differential k-forms*, denoted $\Omega^k(M) := \Gamma(\Lambda^k(T^*M))$. Pointwise, they are elements of the kth exterior power of the cotangent space $\Lambda^k(T_p^*M)$, which is formally spanned by objects of the form

$$\omega_1 \wedge \cdots \wedge \omega_k, \tag{1.22}$$

where each ω_i is a dual vector (so $\omega_i \in T_p^*(M)$) and the associative and bilinear *wedge product* $\wedge$ is subject to the relation $\omega_i \wedge \omega_j = -\omega_j \wedge \omega_i$. Taking k-forms together over all k, these make up the space of *differential forms* $\Omega(M) := \oplus_{k=1}^n \Omega^k(M)$. Note that antisymmetry ensures that n-forms have the highest admissible degree.

Differential forms are set apart largely because they admit a coordinate-invariant notion of differentiation even in the absence of a connection or metric, induced only

by the smooth structure of M, which extends the differential df of a function f. This is the *exterior derivative* $d : \Omega^k(M) \to \Omega^{k+1}(M)$, given in coordinates by

$$\begin{aligned} d\left[f dx^{i_1} \wedge \cdots \wedge dx^{i_k}\right] &= df \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k} \\ &= \frac{\partial f}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k} \end{aligned} \tag{1.23}$$

and extended linearly.

Differential forms also furnish the means of integration on an (oriented) manifold: in general, the naturally integrable objects are not functions, but top-dimensional forms, elements of $\Omega^n(M)$ (sometimes called *volume forms* for this reason). One recovers the usual idea of naturally integrating functions in the presence of a metric, which induces a preferred n-form corresponding to the metric's notion of volume. This metric volume form, often denoted dV (somewhat unfortunately, as this does not mean to suggest it is d of an $(n-1)$-form V), is given in coordinates by

$$dV = \sqrt{|g|} dx^1 \wedge \cdots \wedge dx^n. \tag{1.24}$$

A (compactly supported) function f is now naturally integrated by integrating the volume form $f dV$.

Finally, in the presence of a metric one may naturally relate $\Omega^k(M)$ to $\Omega^{n-k}(M)$ via the *Hodge star* operation $\star : \Omega^k(M) \to \Omega^{n-k}(M)$, defined by

$$(\star\omega)(v_1, \cdots, v_{n-k}) := dV(\omega^*, v_1, \cdots, v_{n-k}). \tag{1.25}$$

Here, ω^* is meant as the $(k, 0)$-tensor obtained by dualizing ω in every slot.

1.2 A Case for and Review of General Relativity

Einstein's General Theory of Relativity has persisted, in the century since its development, as our most successful and robust fundamental description for the structure

and dynamics of the universe on macroscopic scales. We will be inspecting this theory closely and considering possible adjustments, so it is worthwhile to review its essential features and motivations. In this spirit, here we attempt to build up the core features of general relativity from immediate and intuitive features of the world around us, assuming the reader has a basic familiarity with special relativity and the mathematical content of differential geometry.

1.2.1 The Universe as a Manifold

The starting point of general relativity is the stipulation that we may fruitfully model the universe as a 4-dimensional manifold:

Definition 1. *A Hausdorff, second-countable topological space M is called an n-dimensional manifold provided that for each point $p \in M$ there exists a homeomorphism $\phi : U \to V$ between an open set $U \subset M$ containing p and an open set $V \subset \mathbb{R}^n$. The triple (ϕ, U, V) is called a coordinate chart.*

Let us reflect on this stipulation. First, the most fundamental mathematical identification we shall make, even prior to the above, is that we consider "the universe" to be, as a point-set, the set comprised of all possible *events*– unique combinations of space and time at which physical phenomena may take place. We shall equivalently refer to the universe as "spacetime", and it is the backdrop on which physics is to be described. It is apparently the case in our everyday experience that these spacetime events have a reasonable notion of nearness among them, in some sense. That is, we can meaningfully consider events in our surroundings to be close by, and even make sense of a sequence of events limiting to another– the events at which my high-precision watch reads 12:00 and $\frac{1}{n}$ seconds limit to the event at which it reads 12:00 exactly in the $n \to \infty$ limit, under a continuum perspective standard of classical physics. The existence of such notions is what it means to say that the universe is

a topological space, with the qualifiers of "Hausdorff" and "second-countable" being mathematical niceties which rule out pathologies in these notions. Further, we can intuitively see how an observer, armed with (say) a stopwatch and meter stick, can readily parameterize her immediate surroundings in spacetime as combinations of three spatial coordinates and a time coordinate, which together capture this idea of nearness (via the topology of $\mathbb{R}^4$). Such parameterizations precisely provide the coordinate charts of Definition 1, provided only that this most basic description of one's surroundings may, in principle, be carried out at any event, as all experience points to. These considerations amount to the conclusion that the concept of a 4-dimensional manifold apparently provides an utterly reasonable characterization of the classical universe.

What more can be immediately said about the structure of this manifold? Our intent is to model physics, and a defining feature of physics has been its success in modeling dynamics by differential equations, so it is natural to ask that our backdrop can consistently accommodate calculus. Mathematically, this requires M to be sufficiently *smooth*, meaning that the transition function $\phi_2 \circ \phi_1^{-1}$ between any two coordinate charts (with overlapping domains, $U_1 \cap U_2 \neq \varnothing$) is sufficiently smooth as a map between open sets in $\mathbb{R}^4$, ensuring that calculus done in one coordinate chart can be meaningfully translated into any other. For simplicity, in this work we shall take "sufficiently smooth" to mean C^∞, though the core features of general relativity remain unchanged when only requiring transition functions to be C^2, and some formulations make sense under even weaker conditions.

These are the essential features of the universe described as a manifold and their intuitive roots. General relativity has loftier goals than introducing some language, however: we seek to model the physical phenomenon of gravity, and this will require a reflection on what this phenomenon is.

1.2.2 What is Gravity?

We first remark that this titular question is neither philosophical nor mathematical: we mean to review what gravity is empirically. What is it that we see in the physical world that we refer to as gravity, and so hope to explain with general relativity? The simplest answer, of course, is the fact that a ball falls back down when thrown up, but there is plenty more to remark upon. In particular, that the ball falls back down at a different rate when thrown at different distances from the Earth, or if thrown on the Moon, and that it is evidently the case that the same pattern by which these rates vary applies to the rates at which the celestial bodies of the solar system fall around each other– gravity is a seemingly universal phenomenon.

Moreover, in each of the above scenarios, it seems to be that case that any two balls whatsoever, bouncy or bowling or planetary, fall in the same manner (modulo air resistance, buoyancy, and other complications mediated by electromagnetism). This is not restricted to spheres, of course: any shape will do. This is formally encoded in one of Einstein's most essential insights, the *(weak) equivalence principle*, now verified at the level of one part in 10^{15} by the MICROSCOPE collaboration [134], which asserts that the influence of gravity is locally indistinguishable from one's reference frame being accelerated. That is, in a small enough neighborhood of any event $p \in M$ furnished with the coordinates of a Newtonian inertial observer, any test particle in free fall (i.e. subject to no electromagnetic forces) accelerates at essentially the same rate, so that passing to the coordinates of an appropriately accelerated reference frame (e.g. that of an observer also in free fall through p) will render all such objects moving at essentially constant velocity. We shall refer to these coordinate systems as *inertial*, distinguished from the "Newtonian inertial observer" referred to above which would say that gravity induces an acceleration.

What makes gravity empirically different from acceleration, then? Why is it

counted among the four fundamental interactions of matter if it can be transformed away? Because of the need for the qualifiers of "locally", "small enough", and "essentially": the equivalence principle states what occurs in a particular limit, and it is the deviation from this limiting behavior that characterizes gravity and distinguishes it from acceleration. These deviations become globally apparent in observing that distant observers must accelerate differently to locally transform gravity away. Observers at the North Pole and the equator, for example, each must accelerate towards the center of the Earth, but these are different directions in their shared Newtonian inertial coordinates, so no single transformation nullifies gravity everywhere. We may succinctly summarize the empirical phenomenon of gravity, then, as the fact that observers in free fall do not see each other as traveling at exactly constant velocity, and observe that gravity is quantified by the rate of apparent relative acceleration. While this stresses the effect of gravity on matter, we also comment that it is among gravity's most recognizable features that the magnitude and orientation of these effects are in turn strongly correlated with the presence of matter.

1.2.3 Gravity and Geometry

Though a local statement, the equivalence principle has an important global consequence hinted at earlier: the trajectory of an object under the influence of (only) gravity is independent of that object's specific structure. Shape, mass, charge, etc. are all immaterial, provided only that the object is small compared to the length scale over which the deviations discussed above emerge. This is in sharp contrast with the behavior of electromagnetism, in which objects' trajectories are distinctly influenced by both mass and charge. One interpretation of this situation is that there is a family of "preferred" trajectories associated to gravity through the spacetime manifold– for each point $p \in M$ and initial velocity at p, there is a curve which any test particle with these initial conditions would traverse.

Curves and Connections

Among the key ideas of general relativity is that one should identify this family of preferred curves as a geometric feature of the spacetime, that these curves should in some geometric sense be the "default" curves in M. Particularly given that the equivalence principle indicates one can always find local inertial coordinates in which these curves appear very nearly non-accelerating, we might achieve this by asking for a sense in which these curves are exactly non-accelerating according to the structure of M, in a manner independent of coordinates. As coordinate chart accelerations are nebulous and incongruous, identifying acceleration on a manifold requires the invocation of some geometric structure. This is accomplished with a *connection*:

Definition 2. *An affine connection on a smooth manifold M is a bilinear map*

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M),$$

written $(X, Y) \mapsto \nabla_X Y$, which satisfies for each $f \in C^\infty(M)$

(i) $\nabla_{fX} Y = f \nabla_X Y$

(ii) $\nabla_X fY = X(f)Y + f\nabla_X Y$.

The vector field output of $\nabla_X Y$ here, sometimes referred to as the *covariant derivative*, is meant to mathematically represent "the rate of change of Y in the direction of X"– for example, the standard connection in Euclidean $\mathbb{R}^n$ is the directional derivative $\nabla_X Y = (X \cdot \vec{\nabla})Y$, with $\cdot$ the usual dot product and $\vec{\nabla}$ the gradient. Given a curve through spacetime $\gamma : I \to M$ with $I \subset \mathbb{R}$ an interval, a connection allows us to identify the acceleration of γ as being $\nabla_{\gamma'(s)}\gamma'(s)$, the rate of change of the velocity in its own direction (along the curve). The curves which do not accelerate with respect to ∇, its *geodesics*, are then those curves which satisfy $\nabla_{\gamma'(s)}\gamma'(s) = 0$.

The apparently emerging postulate, then, is that the preferred trajectories of gravity through spacetime should be the geodesics of some appropriate connection.

There are in general, however, infinitely many possible choices of connection on a given manifold, so we will need some prescription by which to choose the connection appropriate to modeling gravity in general relativity. Conspicuously missing so far has been any mention of what constraint special relativity places upon the spacetime manifold M, and an inspection of this will provide a crucial hint as to how we might identify a natural choice of connection.

The Inevitibility of the Metric

Whatever the structure of M, it must accommodate special relativity locally. In particular, special relativity dictates that the inertial coordinate systems based at an event $p \in M$, the coordinates of observers in free fall through p and in which free falling objects have effectively constant velocity (corresponding to a straight line trajectory in a spacetime diagram), are mutually related via Lorentz transformations. More pointedly, this ensures that between these coordinate systems the *spacetime interval*

$$\Delta s^2 = -\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2,$$

measured between p and another point in the neighborhood, should be conserved, at least in the limit that each of the changes in this expression is small. We may restate this as saying that the inner product

$$\begin{pmatrix} t & x & y & z \end{pmatrix} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix}$$

is preserved in this limit. Moreover, since Lorentz transformations are linear, the inner product represented by this same matrix between tangent vectors $v \in T_pM$ at p will also be preserved. That is to say, at each point $p \in M$ we may find a $(0,2)$-tensor g_p which is diagonalized in any inertial coordinate system about p and

defined therein by

$$g_p(v,w) = g_p(v^i\partial_i, w^j\partial_j) = v^i w^j g_p(\partial_i, \partial_j) = -v^0w^0 + v^1w^1 + v^2w^2 + v^3w^3.$$

The structure of this tensor ensures that it is a *Lorentzian metric*, simply meaning that it is symmetric and nondegenerate with signature $(1,3)$. We assume that g_p varies smoothly with p, yielding a metric tensor field g on M.

The scenario in which we have found ourselves, then, is that special relativistic considerations apparently endow the spacetime manifold M with a Lorentzian metric g with the property that it is diagonalized at the base point in any inertial coordinate system, precisely the coordinates in which the preferred curves of gravity through this point are (very nearly) straight lines. This relation provides a strong suggestion that the metric tensor is somehow affiliated with gravity and its preferred curves. It is now commonplace to, inspired by this suggestion, put forward this core postulate of general relativity:

Postulate 1. *All gravitational phenomena on the spacetime manifold M are encoded within and mediated by the metric tensor g.*

Armed with this postulate, we return to the problem of identifying a connection with which to geometrically identify the preferred curves of gravity as geodesics. It is among the seminal results of the theory of differential geometry, often dubbed the fundamental theorem of (pseudo-) Riemannian geometry, that a metric tensor on a manifold uniquely determines a naturally associated connection:

Definition 3. *The Levi-Civita connection $\overline{\nabla}$ associated to a pseudo-Riemannian manifold (M,g) is the unique affine connection on M which is (for all vector fields $X,Y,Z \in \mathfrak{X}(M)$) both*

(i) Metric compatible: $X(\langle Y,Z\rangle) = \langle \overline{\nabla}_X Y, Z\rangle + \langle Y, \overline{\nabla}_X Z\rangle$, *and*

(ii) Torsion free: $\overline{\nabla}_X Y - \overline{\nabla}_Y X = [X, Y]$.

The first of these conditions ensures that geodesics of $\overline{\nabla}$ have constant speed, and in particular that an initially timelike geodesic remains timelike ("an object in free fall will never surpass the speed of light"). The second is motivated more geometrically than physically: it ensures the action of $\overline{\nabla}$ on differential forms reduces to the exterior derivative d when antisymmetrized. If the metric is to entirely encode the influence of gravity, and if gravity's preferred trajectories are to be the geodesics of a connection, then it is natural to adopt the Levi-Civita connection $\overline{\nabla}$– this is the adoption made by general relativity.

Curvature and Gravity

Having constructed and motivated from basic principles the core geometric structures present in general relativity, it remains to tie these structures directly to the empirical phenomenon of gravity arrived at in the previous subsection: the apparent relative acceleration between nearby objects in free fall. This requires one more geometric tool, which is constructed out of the Levi-Civita connection $\overline{\nabla}$.

Definition 4. *The Riemann curvature tensor Rm associated to a pseudo-Riemannian manifold* (M, g) *is given, as a* $(1, 3)$ *tensor, by*

$$Rm(X, Y)Z = \overline{\nabla}_X \overline{\nabla}_Y Z - \overline{\nabla}_Y \overline{\nabla}_X Z - \overline{\nabla}_{[X,Y]} Z.$$

The odd use of parentheses is simply a widespread notational convention (tailored to mathematicians who fancy this object an "endomorphism-valued 2-form"). This object most evidently encodes the failure of distinct covariant derivatives to commute, but it also exactly encodes what we've described as the phenomenon of gravity.

Given a timelike geodesic congruence, i.e. a one-parameter family of geodesics $s \mapsto \gamma_s$ (with each γ_s a timelike geodesic of $\overline{\nabla}$) such that the map $(s, t) \mapsto \gamma_s(t)$ from

an open set in $\mathbb{R}^2$ into M is smooth, one can define the vector fields $X := \frac{\partial}{\partial s}\gamma_s(t)$ and $T := \frac{\partial}{\partial t}\gamma_s(t)$ on this map's image. It is automatically the case that $[X, T] = 0$ (by virtue of these vector fields being pushforwards of the coordinate vector fields $\frac{\partial}{\partial s}$ and $\frac{\partial}{\partial t}$ on $\mathbb{R}^2$), and by parameterizing each γ_s by proper time and appropriately shifting the starting point of each geodesic, one can arrange that X is everywhere orthogonal to T [137]. Physically, this congruence is simply a smooth family of nearby free fall trajectories, and the spacelike vector X now has the property that $|X|\Delta s$ is the distance between equally progressed points on the two infinitesimally separated trajectories γ_s and $\gamma_{s+\Delta s}$. The relative velocity V between these trajectories is the rate of change of $X\Delta s$ as one progresses along the geodesics, i.e. $V/\Delta s = \overline{\nabla}_T X$, and the relative acceleration A hence satisfies $A/\Delta s = \overline{\nabla}_T \overline{\nabla}_T X$. Using the geodesic condition $\overline{\nabla}_T T = 0$, that $[X, T] = 0$, and the torsion-free property of $\overline{\nabla}$, we find

$$\begin{aligned} A/\Delta s &= \overline{\nabla}_T \overline{\nabla}_T X = \overline{\nabla}_T \overline{\nabla}_X T \\ &= \overline{\nabla}_T \overline{\nabla}_X T - \overline{\nabla}_X \overline{\nabla}_T T - \overline{\nabla}_{[X,T]} T \\ &= R(T, X)T \end{aligned} \tag{1.26}$$

Now supposing that X is also normalized to $|X| = 1$, this result indicates that the Riemann curvature tensor output $R(T, X)T$ is precisely the relative acceleration per unit distance of separation between infinitesimally shifted free fall trajectories traveling in the direction of T. It is for this reason that curvature is such a crucial concept, and it is in this way that the geometric structure of general relativity purports to capture the phenomenon of gravity.

A Superfluous Postulate

Leading up to Postulate 1, we commented that the fact that g is diagonalized at the base point of any inertial coordinate system (implied only by the special relativistic limit together with the equivalence principle) provided a strong suggestion

that the metric tensor is affiliated with gravity, inspiring us to adopt the Levi-Civita connection as a means of characterizing gravitational trajectories. Here we remark that, under a reasonable mathematical formalization of this fact, this is not merely a suggestion, but an implication, rendering the a priori assumption of Postulate 1 unnecessary for this purpose. That is: given that we model spacetime as a smooth manifold M, the equivalence principle together with local (C^1-)compatibility with special relativity are sufficient to deduce that M is endowed with a metric g for which the preferred curves of gravity, when parameterized by proper time, are the timelike geodesics of its Levi-Civita connection $\overline{\nabla}$ (and hence that the curvature of g necessarily characterizes the empirical effects of gravity).

Let us establish the aforementioned formalization. We have already discussed the metric's existence, and we've said that special relativity indicates that inertial coordinate systems about $p \in M$, wherein the preferred trajectories of gravity (induced by the equivalence principle) are nearly straight lines, have the property that the metric tensor g is diagonalized at p, $(g_{ij})\big|_p = \text{diag}(-1, 1, 1, 1)$. Perhaps the most conservative mathematical interpretation of the curves being "nearly straight lines" here, in a sense which should be exact at p itself, is that their coordinate acceleration is 0 at p. Under this interpretation, the result follows provided only that the reduction to special relativity $(g_{ij})\big|_p = \text{diag}(-1, 1, 1, 1)$ holds to first order, due to the following straightforward fact of Lorentzian geometry:

Lemma 1. *Let (M,g) be a pseudo-Riemannian manifold. Suppose a unit-speed timelike curve $\gamma : I \to M$ has the property that about each point p in the image of γ there exists a coordinate system in which (i) γ has zero coordinate acceleration at p and (ii) g is diagonalized to first order at p, meaning that $(g_{ij})\big|_p = diag(-1, 1, 1, 1)$ and $g_{ij,k}\big|_p = 0$ for all i, j, k. Then γ is a geodesic of $\overline{\nabla}$.*

The converse of this result is also true (by the existence of normal coordinates), but

that is not what we wish to stress in this statement.

Proof. Writing out the acceleration $\overline{\nabla}_{\dot\gamma}\dot\gamma$ (as in (1.14)) in the hypothesized coordinates about some point $p \in M$, we have

$$\overline{\nabla}_{\dot\gamma}\dot\gamma = \left(\ddot\gamma^k + \overline{\Gamma}_{ij}^{\ k}\dot\gamma^i\dot\gamma^j\right)\partial_k. \tag{1.27}$$

Dot derivatives here are evaluated with respect to the curve's parameter, proper time. The formula (1.16) for the Levi-Civita Christoffel symbol allows us to deduce from $g_{ij,k}\big|_p = 0$ that $\overline{\Gamma}_{ij}^{\ k}\big|_p = 0$, so (1.27) becomes $\overline{\nabla}_{\dot\gamma}\dot\gamma\big|_p = (\ddot\gamma^k\partial_k)\big|_p$.

The coordinate velocity of γ is comprised of the rates of change of its coordinate components γ^k with respect to the 0th coordinate t, i.e. $v^k = \frac{d\gamma^k}{dt} = \frac{\dot\gamma^k}{\dot\gamma^0}$. That the coordinate acceleration is 0 at p means that

$$\begin{aligned} 0 = \frac{dv^k}{dt}\bigg|_p &= \frac{1}{\dot\gamma^0}\left[\frac{\ddot\gamma^k}{\dot\gamma^0} - \frac{\ddot\gamma^0}{(\dot\gamma^0)^2}\dot\gamma^k\right]\bigg|_p \\ \implies \quad \ddot\gamma^k\big|_p &= \frac{\ddot\gamma^0}{\dot\gamma^0}\dot\gamma^k\big|_p \end{aligned} \tag{1.28}$$

On the other hand, differentiating the unit-speed condition $-1 = g_{ij}\dot\gamma^i\dot\gamma^j$ at p, using that $g_{ij,k}\big|_p = 0$, yields

$$\dot\gamma^0\ddot\gamma^0\big|_p = (\dot\gamma^1\ddot\gamma^1 + \dot\gamma^2\ddot\gamma^2 + \dot\gamma^3\ddot\gamma^3)\big|_p. \tag{1.29}$$

Substituting in (1.28) for $k = 1, 2, 3$ now gives

$$\begin{aligned} \dot\gamma^0\ddot\gamma^0\big|_p &= \frac{\ddot\gamma^0}{\dot\gamma^0}\cdot((\dot\gamma^1)^2 + (\dot\gamma^2)^2 + (\dot\gamma^3)^2)\big|_p \\ \implies \quad 0 &= \ddot\gamma^0\cdot(-(\dot\gamma^0)^2 + (\dot\gamma^1)^2 + (\dot\gamma^2)^2 + (\dot\gamma^3)^2)\big|_p \\ &= -\ddot\gamma^0\big|_p \end{aligned} \tag{1.30}$$

By (1.28), this implies $\ddot\gamma^k\big|_p = 0$, and so (1.27) becomes $\overline{\nabla}_{\dot\gamma}\dot\gamma\big|_p = 0$. As this can be carried through at any $p \in M$ by hypothesis, this ensures that γ is a geodesic of $\overline{\nabla}$.

□

1.2.4 Geometry and Matter

We have established that the core geometric object of interest for describing the effects of gravity in general relativity is the spacetime metric g, which determines the trajectories of test particles via the Levi-Civita geodesic equation and their relative deviations via the curvature. Indeed, regardless of whether the perspective of general relativity reflects the "true" nature of gravity and spacetime, we have argued that it is a mathematical fact, given only very physically reasonable assumptions, that this geometric framework models gravity's influence well.

What has not been established is how one should discern the appropriate metric with which to model gravity. As noted in closing Section 1.2.2, the effects of gravity are apparently closely tied with the presence of matter, and to capture this within our geometric framework requires some dynamic coupling between geometry and matter. This cannot be discerned so directly from considering only such highly-constrained phenomena as the weak equivalence principle and the special relativistic limit, but a compelling argument for the coupling adopted by general relativity can be deduced from an action principle, a standard means of characterizing equations of motions across many fields of physics.

The Einstein Equation

The *Einstein-Hilbert action* of general relativity is a functional of the Lorentzian metric g on M given by

$$S[g] = \int_U R \, dV, \tag{1.31}$$

where R and dV are (respectively, see Section 1.1.2) the scalar curvature and volume form associated to g and U is any open set with compact closure in M. This is to be the primary geometric contribution to the action of general relativity, and naturally so since R is perhaps the simplest entirely geometric scalar available. Indeed, results

of Lovelock [93], Cartan [21], and Weyl [138] (accessibly reproduced in [110]) indicate that this is, up to a multiplicative constant, the most general coordinate-invariant action in g which is quadratic in the metric's derivatives (up to a divergence).

To identify those metrics which extremize this action, we consider a smooth one-parameter family of metrics $s \mapsto g_s$ such that the variations $g_s - g$ are compactly supported in U (i.e. do not affect the boundary) and ask that $s = 0$ is a critical point of $s \mapsto S[g_s]$:

$$0 = \frac{dS}{ds}\bigg|_{s=0} = \int_U \left[\dot{R}\, dV + R\, d\dot{V}\right], \tag{1.32}$$

where dots indicate s-derivatives evaluated at 0. Denoting $g := g_0$ and $h := \dot{g}_s$, we first note that in any coordinate system the Jacobi formula[1] for the derivative of a determinant yields

$$\frac{d}{ds}\left[\sqrt{|g_s|}\right]\bigg|_{s=0} = \frac{1}{2\sqrt{|g|}}\frac{d|g|}{ds}\bigg|_{s=0} = \frac{\sqrt{|g|}}{2}\langle g, h\rangle. \tag{1.33}$$

Hence the rate of change of the volume form is

$$\begin{aligned} d\dot{V} &= \frac{d}{ds}\left[\sqrt{|g|}\right]\bigg|_{s=0} dx^1 \wedge \cdots \wedge dx^n \\ &= \frac{\langle g, h\rangle}{2}\sqrt{|g|}dx^1 \wedge \cdots \wedge dx^n \\ &= \frac{\langle g, h\rangle}{2}dV. \end{aligned} \tag{1.34}$$

Meanwhile, noting that $R = \langle \mathrm{Ric}, g\rangle$, we have that

$$\dot{R} = \langle \dot{\mathrm{Ric}}, g\rangle - \langle \mathrm{Ric}, h\rangle \tag{1.35}$$

(the negative sign is due to g having raised indices in this contraction). The first term $\langle \dot{\mathrm{Ric}}, g\rangle$ turns out to be a divergence, so it does not contribute to (1.32). With

[1] $\frac{d}{ds}\left[\det(A(s))\right] = \det(A)\mathrm{tr}(A^{-1}\dot{A})$

these evaluations, (1.32) becomes

$$0 = \dot{S} = \int_U \left\langle -\mathrm{Ric} + \frac{R}{2} g, h \right\rangle dV = -\int_U \langle G, h \rangle \, dV, \tag{1.36}$$

where $G = \mathrm{Ric} - \frac{R}{2} g$ is the Einstein tensor. That this should be true for every variation (and hence every symmetric $(0,2)$-tensor h) on every U then requires that

$$G = 0 \tag{1.37}$$

at the critical g, known as the *vacuum Einstein equation*. It is fairly straightforward to see by the above that adding a constant to the action in the form

$$S[g] = \int_U [R - 2\Lambda] \, dV \tag{1.38}$$

modifies the result to the vacuum Einstein equation with *cosmological constant*,

$$G + \Lambda g = 0. \tag{1.39}$$

(1.38) is again a natural action to consider, being the most general coordinate-invariant action, up to a multiplicative constant, in g which is up to and including quadratic in the metric's derivatives.

In general, the full dynamical action will contain both geometric and matter terms. General relativity posits that this takes the general form

$$S[g, \dots] = \int_U [c_1(R - 2\Lambda) + \mathcal{L}_m] \, dV \tag{1.40}$$

for some *matter Lagrangian density* $\mathcal{L}_m$ constructed out of all pertinent matter fields (the referents of the ellipsis) and the metric. Varying the geometric component $[R - 2\Lambda]dV$ proceeds as before, and the variation of $\mathcal{L}_m dV$ with respect to g necessarily gives rise to a $(0,2)$-tensor (which may be taken to be symmetric) in the equation

of motion (1.39), defined to be the *stress-energy tensor* T of the matter. This yields the full Einstein equation with cosmological constant,

$$\boxed{G + \Lambda g = 8\pi T}, \tag{1.41}$$

where c_1 has been chosen to give the factor of 8π leading to Newtonian gravity in the weak-field limit. In principle, this equation must be supplemented with additional equations of motions for each of the contributing matter fields, obtained by varying (1.40) with respect to these fields.

The stress-energy tensor T apparently fully encodes the gravitationally relevant features of the spacetime's matter content. Physically, the evaluation of $T(u, v)$ for two timelike unit tangent vectors $u, v \in T_pM$ is meant to represent the flux of energy-momentum density in the direction of v as measured by an observer traveling in the direction of u– in particular, $T(u, u)$ is simply the energy density as measured by this observer. Equivalently, dualizing one slot to obtain a $(1, 1)$-tensor, the vector $-T(u)$ is the total four-momentum density seen by this observer. As one may find a basis of timelike vectors to the tangent space, this completely characterizes T.

A Scalar Field Source

Of interest both as a precursor to further considerations in this work as well as a minimal example demonstrating explicit matter in (1.40) is the adoption of a *scalar field* source. This is arguably the simplest type of matter one could incorporate, characterized entirely by a function $\phi : M \to \mathbb{R}$. The associated Lagrangian density is typically taken to be of the form

$$\mathcal{L}_m = -\frac{1}{2}|d\phi|^2 - V(\phi), \tag{1.42}$$

where $V : \mathbb{R} \to \mathbb{R}$ is the *potential* function. Again considering a variation g_s of the metric and writing $|d\phi|^2 = \langle d\phi \otimes d\phi, g\rangle$, we see directly that $\dot{\mathcal{L}}_m = \frac{1}{2}\langle d\phi \otimes d\phi, h\rangle$

(notice $V(\phi)$ does not involve the metric), so that

$$\frac{d}{ds}\left[\mathcal{L}_m dV\right]\big|_{s=0} = \dot{\mathcal{L}}_m dV + \mathcal{L}_m d\dot{V} = \frac{1}{2}\left\langle d\phi \otimes d\phi + \mathcal{L}_m g, h\right\rangle. \tag{1.43}$$

The associated stress-energy tensor is then evidently

$$T = \frac{1}{2} d\phi \otimes d\phi - \frac{1}{2}\left(\frac{1}{2}|d\phi|^2 + V(\phi)\right) g. \tag{1.44}$$

We are primarily interested in the case of a *free* scalar field, for which $V(\phi) = \frac{1}{2}m^2\phi^2$, with *mass parameter* m. We also choose to rescale ϕ relative to (1.44) and identify

$$T = 2\frac{d\phi \otimes d\phi}{m^2} - \left(\frac{|d\phi|^2}{m^2} + \phi^2\right) g. \tag{1.45}$$

To obtain the additional equation of motion to couple with (1.41), we must vary $\mathcal{L}_m$ with respect to ϕ. As with the variation of g, we consider a compactly supported variation $s \mapsto \phi_s$ and denote $\phi = \phi_0$, $\psi = \dot{\phi}_s$, asking

$$0 = \dot{S} = \int_U \dot{\mathcal{L}}_m dV. \tag{1.46}$$

Noting $|d\phi_s|^2 = \langle \nabla\phi_s, \nabla\phi_s\rangle$, we see that

$$\dot{\mathcal{L}}_m = -\langle \nabla\phi, \nabla\psi\rangle - V'(\phi)\psi = -\mathrm{div}(\psi\nabla\phi) + \left[\Box\phi - V'(\phi)\right]\psi. \tag{1.47}$$

As the divergence term does not contribute to (1.46), the requirement that $\dot{S} = 0$ for all variations, and hence all choices of ψ, is equivalent to $\Box\phi = V'(\phi)$. For the free case of interest, this is the *Klein-Gordon* equation

$$\Box\phi = m^2\phi. \tag{1.48}$$

Note this equation is linear in ϕ, and so it applies irrespective of the rescaling mentioned above. Equations (1.41), (1.45), and (1.48) taken together, forming the *Einstein Klein-Gordon system*, now completely characterize general relativity with a free scalar field source.

1.2.5 A More General Connection

In this work, we will have occasion to entertain the notion that standard general relativity should be modified by relaxing Postulate 1 to utilize a connection ∇ other than the Levi-Civita connection $\overline{\nabla}$ (apparently meaning, by Lemma 1 and surrounding discussion, that the reduction to special relativity cannot be achieved at the C^1 level if the geodesics of ∇ and $\overline{\nabla}$ do not agree). Here we briefly introduce some tools relevant to such a pursuit.

As we have seen, regardless of what additional or alternative structures are present, the spacetime metric g is an indispensable tool necessitated by the reduction to special relativity. Hence, it provides an important reference point with respect to which we will describe a general connection. It is a standard result in differential geometry that the difference between any two connections, and in particular between our new connection ∇ and the Levi-Civita connection $\overline{\nabla}$, can be encoded in a (0,3) tensor termed the *difference tensor*,

$$D(X,Y,Z) := \langle \nabla_X Y - \overline{\nabla}_X Y, Z \rangle. \tag{1.49}$$

Note this has the coordinate description $D_{ijk} = \Gamma_{ijk} - \overline{\Gamma}_{ijk}$. Given the two defining features of the Levi-Civita connection in Definition 3, it is fruitful to decompose the difference tensor into two subcomponents which describe the extent to which each of these conditions fails. These are the *metric compatibility tensor*

$$\begin{aligned} M_c(X,Y,Z) &:= \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle - Z\left(\langle X, Y \rangle\right) \\ &= D(Z,X,Y) + D(Z,Y,X) \end{aligned} \tag{1.50}$$

(distinguished from the spacetime M via the subscript) and the *torsion tensor*

$$\begin{aligned} T(X,Y,Z) &:= \langle \nabla_X Y - \nabla_Y X - [X,Y], Z \rangle \\ &= D(X,Y,Z) - D(Y,X,Z). \end{aligned} \tag{1.51}$$

We note that T is antisymmetric in X and Y while M_c is symmetric in the same. Together, these entirely characterize D (and hence ∇) according to

$$D(X,Y,Z) = \frac{1}{2}\big[T(X,Y,Z) - T(Y,Z,X) + T(Z,X,Y) \\ -M_c(X,Y,Z) + M_c(Y,Z,X) + M_c(Z,X,Y)\big]. \tag{1.52}$$

We shall use this decomposition in Chapter 2 when varying ∇ in a generalized action.

1.3 Modern Cosmology

Cosmology is the study of the dynamics and structure of the universe on the largest scales– what are the governing principles by which the universe as a whole has evolved? We will be interested in the influence of an adjustment to general relativity on these scales, and so here we review the relevant essential features of modern cosmology.

1.3.1 The Essential Framework

At the foundation of modern cosmology is the *cosmological principle*, the hypothesis that the universe and its matter content are, in an averaged sense over large scales, homogeneous and isotropic. In the context of general relativity, this is most simply realized via a spacetime manifold $(\mathbb{R}_+ \times \Sigma, g)$ with g constructed out of a homogeneous Riemannian metric h on the 3-manifold Σ according to $g = -dt^2 + a(t)^2 h$. In light of observational constraints on the magnitude of the (constant) sectional curvature of h [2], we will restrict to the case that h is flat, requiring that (Σ, h) be a quotient of $\mathbb{R}^3$; in terms of natural (x_1, x_2, x_3) coordinates on $\mathbb{R}^3$, we then have

$$g = -dt^2 + a(t)^2(dx_1^2 + dx_2^2 + dx_3^2). \tag{1.53}$$

This is a geometric ansatz consistent with homogeneity and isotropy, the (flat) *Friedmann–Lemaître–Robertson–Walker (FLRW) spacetime*. The constant-time slices $\Sigma_t :=$

$\{t\} \times \Sigma$ represent the spatial universe, and the *scale factor* $a(t)$ describes the stretching of the spatial metric on Σ_t, known colloquially as the expansion of the universe. Objects following the tangent vector field ∂_t are described as *comoving* and are effectively at rest with respect to the larger universe (this motion is geodesic).

For the matter content in such a spacetime to be isotropic it is required that the stress-energy tensor T, thought of as a $(1,1)$-tensor and hence an endomorphism of the tangent space T_pM at each point $p \in M$, both restricts to a multiple of the identity operator on the spatial subspace $T_p\Sigma_t = \text{span}(\partial_1, \partial_2, \partial_3)$ and preserves the time direction ∂_t, so $T(\partial_t) \propto \partial_t$– a violation of either of these would allow the identification of a preferred direction on Σ_t in the matter profile. For the matter content to be homogeneous, the above constants of proportionality cannot vary with p when translating on the spatial slices Σ_t, meaning they are solely functions of t. These proportionality constants are the comoving *pressure* P in the spatial components and the (negative of) comoving *energy density* ρ in the time component, so we've deduced that T must assume the form

$$T^i{}_j = \begin{pmatrix} -\rho & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & 0 & 0 & P \end{pmatrix} \tag{1.54}$$

in the natural coordinate system associated to our geometric ansatz.

Under these identifications and introducing the Hubble parameter $H := \frac{\dot{a}}{a}$, the Einstein equation (1.41) reduces to the following system:

$$\dot{\rho} = -3H(\rho + P), \tag{1.55}$$

$$H^2 = \frac{8\pi}{3}\rho. \tag{1.56}$$

(1.56) is precisely one of the two *Friedmann equations*, while (1.55) is a simple manipulation of the other known as the *fluid equation*. In general, ρ and P here are the

sums of the energy densities and pressures of all types of matter– baryonic matter, dark matter, radiation, neutrinos, and dark energy–, each of which has its own ρ_i and P_i. In the simplest treatments neglecting energy transfer between the types (an excellent approximation once the universe cools sufficiently that all massive particles become nonrelativistic), equation (1.55) holds for each type individually. Given n species of matter, then, we have $n+1$ equations in $2n+1$ unknowns ($\{P_i\}$, $\{\rho_i\}$, and a), so an additional characterizing constraint for each species (as usual) is sufficient to close the system.

A common scenario is that an energy component satisfies the *equation of state* $P = w\rho$ for a constant w, with cases of interest including $w = 0$ for baryonic matter or cold dark matter, $w = 1/3$ for radiation and relativistic matter, and $w = -1$ for Λ dark energy. In this case, equation (1.55) reads $\dot{\rho} = -3H(1+w)\rho$, yielding

$$\rho \propto a^{-3(1+w)}. \tag{1.57}$$

This characteristic scale-factor dependence for each species means that the history of the universe's energy budget can largely be parameterized by present-day values (set at $a = 1$). Denoting the present-day Hubble parameter $H_0 := H\big|_{a=1}$ and defining the present *critical density* of the universe to be $\rho_{\text{crit}} := \frac{3H_0^2}{8\pi}$, cosmologists generally describe the contribution of each matter species in terms of its present-day fraction of ρ_{crit}, written $\Omega_i := \frac{\rho_i}{\rho_{\text{crit}}}\big|_{a=1}$. In a universe comprised of cold matter and radiation, for example, this allows one to write (1.56) as

$$H^2 = H_0^2\left[\frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_\Lambda\right]. \tag{1.58}$$

In the standard cosmological model, $\Omega_m = \Omega_b + \Omega_d$ includes both baryonic matter and any cold dark matter and $\Omega_r = \Omega_\gamma + \Omega_n$ includes both photons and neutrinos.

The varying exponents in (1.57) also mean that different components will dominate the universe's energy density at different times. The radiation and matter

densities, for example, behave according to $\rho_r = \Omega_r \rho_{\text{crit}} a^{-4}$ and $\rho_m = \Omega_m \rho_{\text{crit}} a^{-3}$, so the universe switches from being radiation-dominated to matter-dominated at the time of *matter-radiation equality*, when $a = a_{eq} := \frac{\Omega_r}{\Omega_m}$. During an era dominated by a species with equation of state w, the time evolution of a can be solved for exactly via (1.56), yielding

$$a(t) \propto (t - t_0)^{\frac{2}{3(1+w)}} \tag{1.59}$$

(excepting the $w = -1$ case for Λ dark-energy domination). As the scale factor generally grows by orders of magnitude during the era dominated by a given matter component, t_0 quickly becomes insignificant, and this yields the general rough approximation that $H \sim \frac{1}{t}$, up to the order 1 coefficient $\frac{2}{3(1+w)}$. Tracing time backwards, however, (1.59) also leads one to the conclusion that $a(t) \to 0$ at some finite time in the past, definitionally shifted to have occurred at $t = 0$– this is easily affirmed more rigorously via an inequality on H, provided the matter content doesn't fundamentally change and (1.56) remains valid (both of these are eventually suspect, of course, in the extreme conditions sufficiently close to $a = 0$). This is dubbed the *initial singularity* of theory, leading to the terminology of the *big bang*.

1.3.2 Observables

The study of astronomy has identified a myriad of observables which can be used to probe and constrain our understanding of the universe's structure on cosmological scales. We review a few of these of particular interest to this work.

Redshift

At the base of a great many cosmological conclusions is the phenomenon that various emission and absorption lines present in the spectra of astronomical sources are shifted, present at different wavelengths than they appear when produced in rest frame terrestrial experiments. The identification of these shifts (and related spectral

features like line broadening) are crucial to our understanding of the local and global universe alike. Locally, they are the primary means by which we deduce relative velocities and rotational rates of nearby stars and galaxies, allowing us to deduce stellar rotation curves and velocity dispersions. On larger scales, these doppler contributions generally become dwarfed by the *cosmological redshift*.

Empirically, cosmological redshift is the phenomenon that more distant objects rather consistently demonstrate successively higher redshifts, known as *Hubble's law*, established in the late 1920s by Edwin Hubble [69] and Georges Lemaître [89]. Theoretically, in the context of the FLRW spacetime, this is a result of the geometry of the metric ansatz (1.53), being a specific instance of the wider phenomenon of *gravitational redshift*. In general, if light traverses a null geodesic trajectory $\gamma : I \to M$, then its four-momentum is proportional to the tangent vector $\dot{\gamma}(s)$, so the energy (and hence frequency) of the light as measured by an observer with timelike tangent vector field T is proportional to $\langle T, \dot{\gamma}(s)\rangle$. If two observers with timelike tangent vector fields T_1 and T_2 measure the frequency of this light at two points $\gamma(s_1)$ and $\gamma(s_2)$ along γ, then, they observe a redshift between them given by

$$1 + z := \frac{\lambda_2}{\lambda_1} = \frac{f_1}{f_2} = \frac{\langle T_1, \dot{\gamma}(s_1)\rangle}{\langle T_2, \dot{\gamma}(s_2)\rangle} \tag{1.60}$$

If we take $T_1 = \partial_t$ and $T_2 = \partial_t$ in the geometry of (1.53), we obtain the gravitational redshift seen by comoving observers. Assuming that both we and the astronomical source of interest have small *peculiar velocities* relative to the universe at large, and that the cosmological averaging implicit in the invocation of (1.53) faithfully encodes redshift influences, this should reflect the observed redshift of distant sources in practice. To compute $\langle \partial_t, \dot{\gamma}\rangle = -\dot{\gamma}^0$, we observer that since $\dot{\gamma}$ is null

$$\begin{aligned} 0 = \langle \dot{\gamma}, \dot{\gamma}\rangle &= -(\dot{\gamma}^0)^2 + a^2\left[(\dot{\gamma}^1)^2 + (\dot{\gamma}^2)^2 + (\dot{\gamma}^3)^2\right] \\ \implies \dot{\gamma}^0 &= a\sqrt{(\dot{\gamma}^1)^2 + (\dot{\gamma}^2)^2 + (\dot{\gamma}^3)^2}. \end{aligned} \tag{1.61}$$

Further noting that $\langle \partial_i, \dot{\gamma} \rangle = a^2 \dot{\gamma}^i$ is preserved along γ since ∂_i is a Killing vector field for $i = 1, 2, 3$, we find that

$$a\dot{\gamma}^0 = \sqrt{\langle \partial_1, \dot{\gamma} \rangle^2 + \langle \partial_2, \dot{\gamma} \rangle^2 + \langle \partial_3, \dot{\gamma} \rangle^2} \tag{1.62}$$

is preserved along γ. Equation (1.60) therefore becomes

$$1 + z = \frac{\langle \partial_t, \dot{\gamma}(s_1) \rangle}{\partial_t, \dot{\gamma}(s_2) \rangle} = \frac{\dot{\gamma}^0(s_1)}{\dot{\gamma}^0(s_2)} = \frac{a_2}{a_1}, \tag{1.63}$$

indicating that the cosmological redshift is precisely given by the ratio of the scale factors between observations, consistent with the idea of the expanding universe "stretching" the wavelength.

The structure of the redshift-distance relation provided the most definitive evidence in early 20th century that the universe is not static, but instead expanding. In 1998, the extension of the cosmic distance ladder to include type IA supernovae allowed this same relation to provide evidence that the universe's expansion is accelerating [109, 115]. This relation is the primary means by which the present-day Hubble parameter H_0 is measured directly in the local universe.

The Cosmic Microwave Background

The Cosmic Microwave Background (CMB) radiation provides some of the most stringent tests and parameter constraints available for cosmology. The energy density scaling (1.57) together with the expectation that $a \to 0$ in the past indicate that earlier in the history of the big bang model, the universe was much denser and thereby much hotter. Early enough, the thermal energy of atomic nuclei would have been sufficient to consistently ionize them, so that the universe would have been a thermalized sea of ions opaque to radiation due to Compton scattering. At some point between then and now, the temperature of this sea dropped sufficiently that neutral atoms could form in an event known as *recombination*, allowing radiation to

essentially free-stream thereafter. This leads to the expectation that the redshifted signal from this time of last scattering may still be present today, given the paucity of the intergalactic medium.

The accidental discovery of this signal in 1965 [108], as well as the later confirmation that it exhibits the most precise blackbody spectrum observed in nature [96] with temperature $T_0 \approx 2.726$ K ≈ 0.235 meV, provided some of the most definitive evidence for the big bang model. The CMB is persistent in time and almost perfectly uniform across all observation directions, not associated to any particular astronomical source. Beyond its mere existence, the true weight of the CMB in constraining cosmology stems from its minuscule anisotropies, variations in temperature (and polarization) on the order of one part in 10^5, as the statistics of these fluctuations encodes a wealth of information.

The CMB temperature deviation as a function of observed direction $\tilde{\Theta}(\hat{\mathbf{n}}) := \frac{\Delta T}{T}$ defines a map $S^2 \to \mathbb{R}$, and the specific quantity of interest is the correlation between fluctuations in different directions $\langle \tilde{\Theta}(\hat{\mathbf{n}}) \tilde{\Theta}(\hat{\mathbf{n}}') \rangle$. Strictly speaking, this correlation is theoretically to be carried out over many realizations of quantum mechanical seed fluctuations, i.e. over many instantiations of the universe. Of course, we can only probe one such instantiation, but the expected isotropy of the seed statistics indicates that the correlation should depend only on the relative angle between $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$, so one averages over all pairs of directions with the same relative angle to obtain an estimate for $\langle \tilde{\Theta}(\hat{\mathbf{n}}) \tilde{\Theta}(\hat{\mathbf{n}}') \rangle$ as a function of $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}'$. This function can now be expanded in Legendre polynomials (or $m = 0$ spherical harmonics):

$$\langle \tilde{\Theta}(\hat{\mathbf{n}}) \tilde{\Theta}(\hat{\mathbf{n}}') \rangle = \sum_{l=0}^{\infty} \frac{2l+1}{4\pi} C_l P_l(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}'), \tag{1.64}$$

with the coefficients C_l comprising the *angular power spectrum* of the CMB temperature fluctuations. The intrinsic limitation imposed by the fact that we only have

one universe to probe, ultimately meaning that there is less data available to inform lower multipoles, is known as *cosmic variance*. $\tilde{\Theta}(\hat{\mathbf{n}})$ has been measured and utilized to estimate C_l up to $l \sim 2500$ most precisely by the Planck collaboration, with the results being used to ascertain best-fit values for the suite of cosmological parameters (see Section 1.3.4). We detail the theoretical prediction of C_l in Section 3.2, and succinct summaries of the phenomenology can be found in [67, 68]

Primordial Light Element Abundances

Recombination occurred when the temperature of the cosmic primordial soup passed well below the binding energy of hydrogen $B_H \approx 13.6$ eV, so that the photon bath no longer had enough energy to re-ionize spontaneously combining electrons and protons. Long before this, the radiation temperature would have been substantially higher, varying in proportion to the inverse of the scale factor due to the redshift relation (1.63), eventually high enough that photons would have also dissociated the protons and neutrons in nuclei. In the progression forward from the initial singularity, then, there was a time at which a sea of neutrons and protons coalesced into nuclei, an event known as *Big Bang Nucleosynthesis* (BBN), beginning once the temperature passed well below the binding energy of deuterium $B_D \approx 2.22$ MeV. After a short time, however, the universe expanded sufficiently that nucleons are too diffuse and cool for further synthesis to occur, leading the elemental abundances to effectively *freeze out* by $T \sim 0.01$ MeV. These abundances then remain fixed until stars provide another venue for nucleosynthesis quite a bit later on.

Assuming that nuclear physics proceeds according to the same fundamental principles and cross-sections as found in terrestrial experiments (across the same energy scales), and that BBN occurs while the universe is strongly radiation-dominated, the ultimate freeze out abundances are fixed by the single input $\Omega_b h^2$, with the *reduced Hubble parameter* h defined by $H_0 = 100h\ \frac{\text{km/s}}{\text{Mpc}}$. In the context of BBN, this input is

typically described by the *baryon-to-photon ratio*

$$\eta_b := \frac{n_b}{n_\gamma} = \frac{\pi^2 \Omega_b \rho_{\text{crit}}}{2\zeta(3) T_0^3 m_p} \approx 5.5 \cdot 10^{-10} \cdot (50 \Omega_b h^2), \tag{1.65}$$

where n_b and n_γ are the number densities of baryons and photons, T_0 is the present-day CMB temperature, m_p is the proton mass, and ζ is the Riemann zeta function. For reasonable values of this parameter, the only isotopes produced in any appreciable amount are hydrogen, deuterium, ^{3}He, ^{4}He, and ^{7}Li, and the abundance for each of these is readily calculable via numerical analysis of a coupled system of Boltzmann equations incorporating all relevant nuclear reactions (discussed in more detail in Section 3.1).

These abundances can also be probed observationally. The elemental makeup of a great variety of astrophysical sources is discernible via spectroscopy, and it has long been recognized that hydrogen and helium are by far the most abundant elements in stars, the interstellar and intergalactic media, and the universe at large. A natural quantity for characterizing and categorizing astrophysical sources, then, is their *metallicity*, the mass fraction accounted for by all elements other than hydrogen and helium. As stars process hydrogen and helium into heavier elements through stellar nucleosynthesis, metallicity is generally indicative of the extent of stellar astration in an environment, i.e. how much stellar processes have adjusted the distribution of elements from primordial conditions. By investigating high redshift, metal-poor environments, astronomers can infer the relative elemental abundances at the earliest stages of the universe's evolution.

1.3.3 Dark Matter

While general relativity has had great success in describing both detailed features and broad strokes of cosmological evolution, it often does so at the cost of invoking matter for which we have seen no terrestrial indications. This comes in two flavors,

the first being *dark energy*, which is at least phenomenologically accommodated in general relativity via the cosmological constant Λ. Dark energy is needed within this framework to explain the universe's accelerating expansion and to simultaneously allow the flat universe preferred by CMB anisotropies and the matter contribution $\Omega_m \sim 0.3$ indicated by galaxy surveys [1, 33]. While it may be that this is simply a fundamental feature of Einstein's equation (1.41), physicists continue to probe whether observed dynamics might be better explained with a novel form of matter contributing to T, such as a fluid with equation of state w slightly different than -1.

The second flavor is *dark matter*. Across a wide array of observational data, evidence abounds for a consistent discrepancy between the observed gravitational dynamics of the universe, across all scales much larger than the solar system, and what is expected based upon the amount of baryonic matter detected or inferred, generally pointing to a matter deficit. The possibility of such a deficit has been on the minds of physicists for more than a century: based on his measurements of the velocity dispersion of stars in the Milky Way, Lord Kelvin explicitly posited as early as 1904 that "many of our stars, perhaps a great majority of them, may be dark bodies" [80]. It wasn't until a few decades later, however, that this hypothesis began to be taken seriously by astronomers, largely beginning with Francis Zwicky's deduction in 1933 that the velocity dispersion of galaxies in the Coma Cluster was much higher than could be virially supported by the mass suggested by the cluster's luminosity [146]. By the late 1970s, the flatness of galaxy rotation curves had been firmly established, yielding similar conclusions [118] (see Figure 1.1). Beyond these relatively local considerations, models of cosmological structure evolution were also found to require more gravity than can be supplied by baryons in order for galaxies in bulk to form and cluster as observed [1, 33]. Modern observations and analyses, as in gravitational lensing signatures and CMB anisotropies, have continued to affirm that both cosmological and galactic dynamics are consistent with a matter deficit

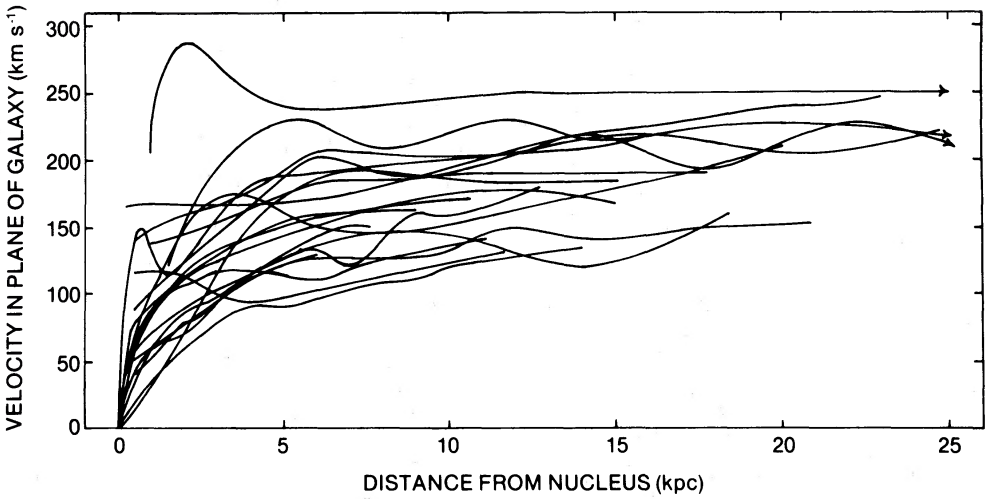

FIGURE 1.1: Superimposed galactic rotation curves for some 21 Sc galaxies, as determined and compiled by Rubin, Ford, and Thonnard in 1980 [118][2]. At large radii beyond the luminous matter, one expects $v \propto 1/\sqrt{r}$– that none of these curves exhibit this eventually decreasing behavior demonstrates the problem of dark matter

[2, 24, 25, 31, 73, 88].

Though the need for a theoretical explanation for this apparent deficit was empirically well-established by the 1970s, a compelling resolution still remains elusive today. A great many explanatory hypotheses have been put forward, and these can be broadly categorized as either specifying the nature of the unseen matter or modifying the theory of gravity so that no supplementary matter is needed. Perhaps the most popular among theories modifying gravity is the theory of Modified Newtonian Dynamics (MOND), which phenomenologically alters the law of Newtonian Gravitation at sufficiently small accelerations. Among the astrophysical community's primary objections to modified gravity is the 2006 analysis of the bullet cluster [31] (Figure 1.2), a pair of recently-collided galaxy clusters for which lensing data suggests that the apparently missing mass separated considerably from the visible matter in the course of the collision. Generally speaking, one might expect that a simply implemented theory of modified gravity would predict that the apparently missing mass

[2] DOI 10.1086/158003. ©AAS. Reproduced with permission.

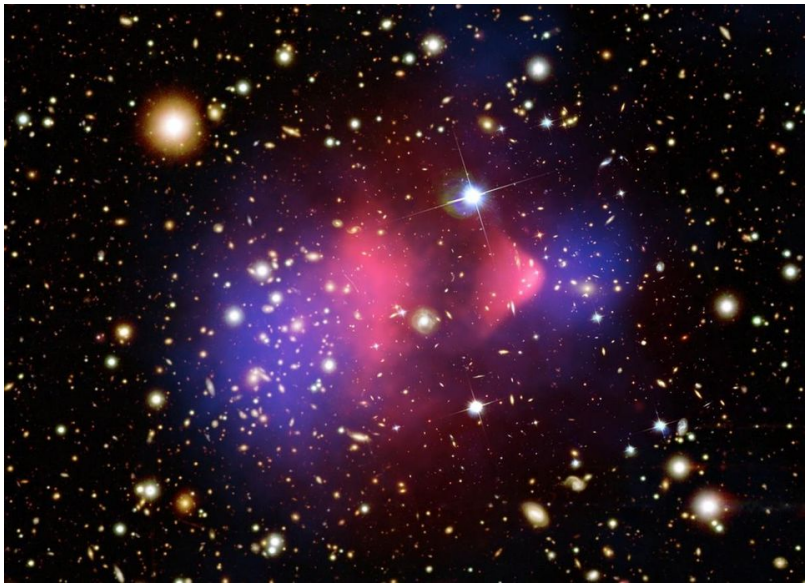

FIGURE 1.2: The distribution of baryonic versus apparent matter in the Bullet Cluster. The pink overlay is in the x-ray spectrum, indicating the location of the bulk of the baryonic matter present in the intracluster medium, while the blue overlay is the apparently gravitating mass distribution inferred via gravitational lensing. Credit: X-ray: NASA/CXC/CfA/M.Markevitch et al.; Optical: NASA/STScI; Magellan/U.Arizona/D.Clowe et al.; Lensing Map: NASA/STScI; ESO WFI; Magellan/U.Arizona/D.Clowe et al.

closely tracks the visible matter. The bullet cluster showed that such tracking does not occur universally, seen by many as sufficient reason to focus attention on the hypothesis of additional matter.

Dark Matter Candidates

Perhaps the most significant cosmological constraint on (the bulk of) dark matter is that it must be *cold* for most of the universe's history. That is, its cosmological pressure is negligible compared to its energy density (so it has equation of state $w \sim 0$), as for a gas in thermal equilibrium at a temperature much smaller than the particle mass. This is required by the observed scales of galaxy formation and clustering. Following the gas analogy (not an analogy at all in particle models of dark matter), negligible pressure corresponds to gas particles moving nonrelativistically, so the cold constraint heuristically ensures that the particles do not stream too quickly

to be gravitationally bound into overdensities on scales as small as galaxies when initialized with seed perturbations at a magnitude set by CMB anisotropies.

Among the leading models of cold dark matter are Weakly Interacting Massive Particles (WIMPs) sufficiently heavy to be cold. WIMPs do very well cosmologically, naturally yielding the appropriate relic abundance and accommodating galaxy formation and clustering, and they have long been viewed as the prototypical cold dark matter candidate. By the early 2000s, however, WIMPs were found to have some shortcomings on sub-galactic scales, with simulations generally overproducing small-scale structure compared to observations. Among these, for example, is the *cusp-core problem*: many-body simulations of interacting WIMP clusters consistently develop sharp, cuspy density profiles at the centers of galaxies, while observations tend to favor smooth density cores [40, 99]. Other such issues included the *missing satellites problem*, a discrepancy between the number of observed and predicted persisting satellite galaxies orbiting a host [82], and the *too-big-to-fail problem*, a prediction by WIMP simulations of dark matter subhalos with higher central densities than observed in Milky Way satellites [16]. It has been argued that these issues may well be resolved by incorporating more sophisticated baryonic physics into simulations or by more careful analysis of observational data [81, 103], but the community's unbridled confidence in WIMPs has waned nonetheless, not least due to their repeated failure to manifest in collider experiments.

Another class of models for cold dark matter which has received much attention is broadly categorized as Scalar Field Dark Matter (SFDM, with other relevant monikers including "fuzzy" or "wave" dark matter), characterized by dark matter's being classically well-modeled by a scalar field ϕ coupled to the Einstein equation [66, 97], as discussed in Section 1.2.4. SFDM is often considered with an extremely light mass parameter on the order $m \sim 10^{-22}$ eV since this naturally suppresses structure formation on small scales (with an absolute minimum set by the Compton

length scale $\hbar c/m$), and it was hoped this might address the challenges faced by WIMPs. A variety of recent observational constraints, however, seem to favor larger masses [4, 23, 39, 41, 64, 124]: these tend to roughly prefer $m \gtrsim 10^{-21}$ eV, though a more recent work extends as far as $m \geqslant 3 \cdot 10^{-19}$ eV [38]. Whether or not SFDM can ultimately resolve problems surrounding small-scale structure (should they exist) given these developments, it remains at least as viable a candidate as WIMPs.

The most common means of fundamentally motivating SFDM is the invocation of an ultralight axion-like particle [22, 70, 95], though for many purposes SFDM can be investigated purely phenomenologically via its coupling to the Einstein equation since dark matter is an empirically classical phenomenon. In 2010, Bray showed that one can also motivate SFDM entirely classically via a natural modification to the geometry of general relativity, namely by allowing a nontrivial connection to contribute to the action [19] . In Chapters 2 and 3, we investigate some potential implications of this geometric adjustment for the cosmological observables discussed in Section 1.3.2.

1.3.4 ΛCDM

Λ Cold Dark Matter (ΛCDM), widely known as the *standard model of cosmology*, is an accounting of the cosmological history of the universe in strong agreement with observations on many counts, particularly tailored towards those observables discussed in Section 1.3.2. ΛCDM is situated in the framework of general relativity and posits that the cosmological universe is, on average, well-described by a flat FLRW spacetime containing baryons, radiation, neutrinos, dark matter, and dark energy. It describes dark energy via the cosmological constant Λ, but it does not invoke an explicit underlying model for cold dark matter– it is treated in an effective manner as a perfect fluid with equation of state $w = 0$, for many purposes only distinguished from regular matter in that its dynamics are not coupled to photons.

The picture painted by ΛCDM is as follows. The very early universe (no later than $z \sim 10^7$) was comprised of scalar adiabatic perturbations to each of the matter species on the flat FLRW background characterized by its baryonic matter fraction Ω_b, dark matter fraction Ω_d, and the present-day Hubble parameter H_0, with perturbations having *fluctuation amplitude* A_s and distributed across scales according to the *spectral index* n_s. These perturbations propagate according to the Einstein equation in a geometry perturbed from (1.53). The early universe is radiation dominated, with the energy density determined by the present-day CMB temperature $T_0 = 2.7255$ K and the effective number of neutrinos $N_{\text{eff}} = 3.046$. The sum of the neutrino masses is set to $\Sigma m_\nu = 0.06$ eV, so these contribute to radiation early on but eventually become nonrelativistic. Matter comes to dominate the energy density at $a_{\text{eq}} = \Omega_r/\Omega_m$, at which point structure growth begins within dark matter perturbations. Baryons, distributed amongst nuclei according to the outcome of radiation-dominated BBN, are delayed from following suit until decoupling from radiation upon recombination at $z \sim 1100$ (set by T_0, $\Omega_b h^2$, and B_H and marginally influenced by the primordial helium mass fraction Y_{p}), when the state of perturbations is imprinted in the CMB as anisotropies. The universe remains fairly homogeneous and neutral until structure growth proceeds sufficiently that baryon overdensities coalesce into stars and galaxies. Eventually, the stellar luminosity is sufficient to largely reionize the sparse intergalactic medium, an event known as *reionization* which is quantified by its *optical depth* τ_{re} (physically, $e^{-\tau_{re}}$ is the probability that a photon streams without scattering between reionization and the present). The ionized medium now weakly recouples to the free-streaming CMB, but scattering is inefficient due to the paucity of ions. Near the present day, the matter density diffuses below the scale of the cosmological constant, initiating the era of dark energy domination.

Out of this picture, then, emerge the six free *cosmological parameters* upon which ΛCDM is based (and which are not definitively set by independent observations)–

see Table 1.1. These provide all of the inputs necessary for ΛCDM to predict a wide range of cosmological phenomena, including the large-scale structure of galaxies encoded in the *matter power spectrum*, the redshift-distance relation, the CMB temperature and polarization anisotropies, and the primordial light element abundances. The angular power spectrum of the CMB temperature anisotropy alone, in fact, is sufficient to highly constrain these parameters, and ΛCDM's great success has stemmed from these preferred values' generally remaining robust upon consideration of additional data. Particularly compelling for the big bang picture at large, for example, is that the CMB's preferred baryon fraction falls squarely within the range required by comparison of (entirely independent) BBN analysis with observed light element abundances, namely ^{4}He and deuterium. Reported in Table 1.1 are constraints provided by a slew of early-universe data, primarily derived from the CMB, as analyzed by the Planck collaboration [2].

Tensions in ΛCDM

While ΛCDM has historically been probed via early-universe phenomena, a variety of local universe measurements have gained increasing constraining power over the past two decades, yielding impressive alignment in some cases and growing discrepancies in others. Perhaps the most significant and consistent discrepancy is known as the

Table 1.1: **Cosmological Parameters.** The suite of six cosmological parameters completely characterizing the ΛCDM model of cosmology, together with associated 68% confidence constraints reported by the Planck collaboration [2].

Description	**Parameter**	**Planck Constraint**
Baryon fraction	Ω_b	$\Omega_b h^2 = 0.02242 \pm 0.00014$
Dark matter fraction	Ω_d	$\Omega_d h^2 = 0.11933 \pm 0.00091$
Present hubble parameter	H_0	$h = 0.6766 \pm 0.0042$
Fluctuation amplitude	A_s	$\ln(10^{10} A_s) = 3.040 \pm 0.016$
Spectral index	n_s	$n_s = 0.9626 \pm 0.0057$
Optical depth to reionization	τ_{re}	$\tau_{re} = 0.0522 \pm 0.0080$

Hubble tension, a disagreement in the preferred value of the present Hubble parameter H_0 between the local redshift-distance relation amongst cepheid-calibrated Type IA supernovae, found to require $h = 0.732 \pm 0.013$ by the SH0ES collaboration [114], and the above tabulated value reported by Planck. While each of these is the most precise measurement within its class (early- versus late-universe), they lie within a trend of similar results establishing the discrepancy to at least 4σ. Extensive review of the Hubble tension and the myriad of observational constraints can be found in [43, 77].

An older and less extreme concern is the *lithium problem* of BBN. BBN analysis indicates that while almost all of the universe's neutrons, whose abundance froze out after the weak interaction became inefficient at $T \sim 1$ MeV, ultimately wound up in ^{4}He nuclei, a handful went into traces of primordial deuterium, ^{3}He, and ^{7}Li. The primordial ^{3}He abundance has not yet been reliably empirically determined, but the relic abundances of ^{4}He and deuterium (relative to hydrogen) are relatively well-established [55],

$$Y_{\rm p} = 0.245 \pm 0.003, \tag{1.66}$$

$$\left.{\rm D/H}\right|_{\rm p} = (25.47 \pm 0.25) \times 10^{-6}. \tag{1.67}$$

The ^{4}He abundance is reported via its mass fraction $Y_{\rm p}$, while other abundances are reported via their number densities relative to hydrogen. As mentioned above, these are in excellent agreement with the value of $\eta_b \propto \Omega_b h^2$ suggested by CMB anisotropies. Meanwhile, best observational estimates for the relic abundance of ^{7}Li presently return [55]

$$\left.{}^7{\rm Li/H}\right|_{\rm p} = (1.6 \pm 0.3) \times 10^{-10}, \tag{1.68}$$

some 3.5 times smaller than the predicted value of $(5.623 \pm 0.247) \times 10^{-10}$ [113]. While it is encouraging that the prediction and observation are on the same order of magnitude (note the wide spread in the magnitudes of (1.66), (1.67), and (1.68)),

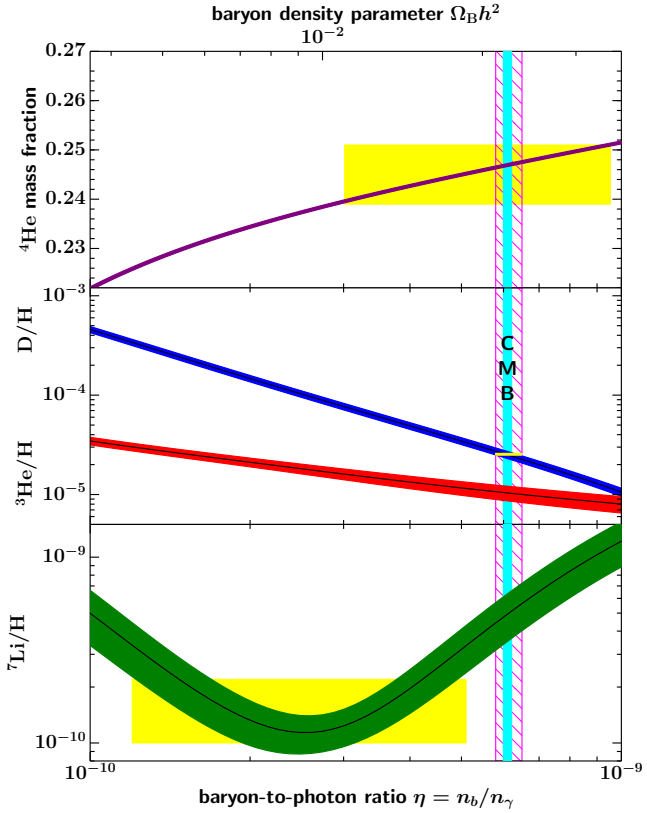

FIGURE 1.3: Concordance in light element abundances as a function of η_b. Colored bands around the black curves indicate the predictions' 95% confidence intervals; yellow boxes indicate the region of agreement between prediction and measurement of each nuclide abundance (excepting ^{3}He). The vertical blue region spans the η_b range in agreement with the CMB; the wider pink region spans the range in which both ^{4}He and deuterium abundances are consistent with measurements (both to 95% confidence). Image generated by the Particle Data Group [55].

the confidence levels are decidedly incompatible. See Figure 1.3.

It's possible the lithium discrepancy is due to observational challenges in obtaining the primordial value. ^{7}Li abundances are most easily determined from metal-poor stars, and these exhibit unexplained large scatter at very low metallicities [Fe/H] < -3.0, inhibiting extrapolation to 0 metallicity (the primordial situation). It is seen as unlikely that this could fully resolve the issue, however, because stellar mechanisms for depleting ^{7}Li aren't thought to be efficient enough to reduce the expected primordial abundance down to the value (1.68) measured within moder-

ately metallic stars [55]. Further, the nuclear reaction rates informing the predicted value are well enough constrained that it is unlikely that the lithium problem will be resolved through nuclear physics [113]. New physics hypotheses proposed to resolve the lithium problem abound, but none has yet emerged as thoroughly convincing. Further current discussion may be found in [15].

1.3.5 Free Scalar Fields in Cosmology

Here we discuss the behavior of a free scalar field ϕ coupled to the metric (1.53). As discussed in Section 1.2.4, we take ϕ to have the associated stress-energy tensor T_ϕ given by (1.45) and satisfy the Klein-Gordon equation (1.48). First we establish the conditions under which T_ϕ is homogeneous and isotropic, i.e. has the form (1.54). Noting that the metric g becomes the identity map upon dualizing a slot to obtain a $(1,1)$-tensor, we need only impose homogeneity and isotropy on $\nabla\phi \otimes d\phi$. Since the image of $\nabla\phi \otimes d\phi$ is at most one-dimensional (being the span of $\nabla\phi$), the only multiple of the identity map it can restrict to on the spatial subspace $\mathrm{span}(\partial_1, \partial_2, \partial_3)$ is 0. That is, it must be that $0 = d\phi(\partial_i) = \partial_i(\phi)$ for $i = 1, 2, 3$, so ϕ is solely a function of time t.

That ϕ depends only on time implies $d\phi = \dot{\phi}\, dt$ and $\nabla\phi = -\dot{\phi}\, \partial_t$, so $|d\phi|^2 = -\dot{\phi}^2$ and the scalar field's energy density ρ_ϕ and pressure p_ϕ are evidently

$$\rho_\phi = 2\frac{\dot{\phi}^2}{m^2} + \left(\frac{|d\phi|^2}{m^2} + \phi^2\right) = \frac{\dot{\phi}^2}{m^2} + \phi^2, \tag{1.69}$$

$$p_\phi = \qquad - \left(\frac{|d\phi|^2}{m^2} + \phi^2\right) = \frac{\dot{\phi}^2}{m^2} - \phi^2. \tag{1.70}$$

We may also now write out the Klein-Gordon equation $\Box\phi = m^2\phi$, which becomes

$$\ddot{\phi} + 3H\dot{\phi} + m^2\phi = 0. \tag{1.71}$$

These equations are coupled to the broader cosmology through the Friedmann equations involving the total ρ and p. Defining $\psi := \frac{\dot{\phi}}{m}$, we will think of the second-order

ODE (1.71) in ϕ as a coupled system of two first-order ODEs in ϕ and ψ,

$$\begin{aligned}\dot{\phi} &= \quad m\psi \\ \dot{\psi} &= -m\phi - 3H\psi. \end{aligned} \tag{1.72}$$

The energy density and pressure are now $\psi^2 \pm \phi^2$, with ψ giving the kinetic term and ϕ the potential. There will be two regimes of interest in the dynamics of (1.72), $m \ll H$ and $m \gg H$. Noting that $H \to \infty$ at the initial singularity and decreases monotonically thereafter in standard cosmologies, this partitions the evolution into at most two eras, with the transition roughly occurring at $t \sim 1/m$.

In the remainder of this subsection, ρ will refer to $\rho_\phi = \psi^2 + \phi^2$. We observe

$$\dot{\rho} = 2\psi(m\phi + \dot{\psi}) = -6H\psi^2 \tag{1.73}$$

(notice this is equivalent to (1.55)). We consider the quantity θ defined by $\sin(\theta) := \frac{\phi}{\sqrt{\rho}}$ and $\cos(\theta) := \frac{\psi}{\sqrt{\rho}}$, measuring the angle in the ψ-ϕ plane, and find

$$\begin{aligned}\sin(2\theta)\dot{\theta} = \frac{d}{dt}\left(\sin^2(\theta)\right) &= \frac{2\dot{\phi}\phi\rho - \phi^2\dot{\rho}}{\rho^2} \\ &= 2m\frac{\phi\psi}{\rho} + 6H\frac{\phi^2\psi^2}{\rho^2} \\ &= m\sin(2\theta) + \frac{3}{2}H\sin^2(2\theta), \end{aligned} \tag{1.74}$$

which implies

$$\dot{\theta} = m + \frac{3}{2}H\sin(2\theta). \tag{1.75}$$

This result will allow us to understand the qualitative behavior of the system (1.72).

In the $m \gg H$ regime, the second term in (1.75) is negligible, so the ψ-ϕ angle θ increases at the constant rate of m, corresponding to sinusoidal oscillations in $\phi/\sqrt{\rho}$ and $\psi/\sqrt{\rho}$ at frequency m. Moreover, rewriting (1.73) as

$$\frac{d}{dt}\ln(\rho) = -6H\cos^2(\theta) \tag{1.76}$$

and replacing $\cos^2(\theta) = 1/2$ via a time-averaging over the oscillation period $\frac{2\pi}{m}$ (this is reasonable because it is much shorter than the timescale $1/H$ of variations in both ρ and H), we find $\frac{d}{dt}\ln(\rho) \approx -3H$, leading to $\rho \propto a^{-3}$, as expected of a cold matter component ($w = 0$). This is consistent with the observation that the effective equation of state $p/\rho = \cos^2(\theta) - \sin^2(\theta) = \cos(2\theta)$ time-averages to 0.

In the $m \ll H$ regime, the first term in (1.75) is negligible for the most significant dynamical features, so we are interested in the system $\dot{\theta} = \frac{3}{2}H\sin(2\theta)$. In particular, this system has a stable equilibrium at $\theta = \pi/2$ and an unstable equilibrium at $\theta = 0$. We generally expect, then, that $\theta \to 0$ at very early times and $\theta \to \pi/2$ at late times (with distance from the equilibrium proportional to $a^{\pm 3}$ in these limits). Equation (1.76) then indicates that $\rho \propto a^{-6}$ at early times ($\theta \to 0$), a period known in the literature as *kination* [42, 76, 119], and $\rho \sim$ constant at late times ($\theta \to \pi/2$), consistent with the corresponding equations of state $w \to \pm 1$.

The most general picture, then, is the following. At the earliest times, $\psi \gg \phi$ and $\psi^2 \approx \rho \propto a^{-6}$, so $\psi \propto a^{-3}$ while ϕ varies comparatively little. Eventually, the decay in ψ leads to $\psi < \phi$ and ρ levels off to the value of ϕ^2, remaining nearly constant (contributing like dark energy) until $H \sim m$ and oscillations begin. At this point, $\rho \propto a^{-3}$ and the scalar field contributes like cold dark matter. Even in the intermediate dark energy phase, ψ continues to decay like a^{-3} until becoming comparable to $\frac{m}{H}\phi$, at which point further analysis of (1.72) indicates that the next-order correction is $(\psi + \frac{m}{5H}\phi) \propto a^{-3}$.

The duration, or even existence, of each of these eras depends on the initial value of θ (set at some very early time, for our purposes between inflation and nucleosynthesis) and the value of m. In particular, the $\rho \sim$ constant dark energy era can either not be present at all if θ starts sufficiently close to $\pi/2$, or extend back arbitrarily close to the singularity if θ starts very near to 0, effectively eliminating

the $\rho \propto a^{-6}$ era. The later dark matter phase is eliminated if $m \lesssim H_0$, meaning that the transition to oscillations would not occur before the present day. The axion-like particle perspective on SFDM typically suggests that axions are produced with low momentum [117], meaning that $\theta \sim 0$ even initially, which eliminates the $\rho \propto a^{-6}$ era in favor of the dark energy phase. From a geometric perspective on SFDM given in Chapter 2, however, such a choice appears arbitrary, and so it is natural to ask what observable implications allowing the $\rho \propto a^{-6}$ kination era might have, particularly as it might inform the tensions in ΛCDM discussed in Section 1.3.4. We investigate possible implications in Chapter 3.

2

Geometric Scalar Field Dark Matter and Oscillating Redshifts

In 2010, Bray [19] showed that one can motivate SFDM entirely classically via a natural modification to the geometry of general relativity, namely by allowing a nontrivial connection to contribute to the action. In this chapter, we investigate some potential implications of this geometric adjustment when one treats the nontrivial connection as having physical content beyond its implications for the Euler-Lagrange equations. We find that treating Bray's axioms broadly can lead to a rather distinct prediction for the behavior of gravitational redshifts in the presence of variations of the scalar field ϕ, encapsulated in equation (2.43). In particular, this could have readily evaluable implications for the time evolution of cosmological redshifts, quantities which, as one of the primary observables by which we probe and characterize the universe, have undergone a large degree of empirical scrutiny.

This scrutiny notwithstanding, the time evolution of redshifts of fixed sources has not been thoroughly investigated empirically. This is largely due to the fact that the expected rate of change (within ΛCDM) of the redshift of a source at fixed comoving distance is comparable (for redshifts $z \leqslant 10$) to the Hubble parameter in order of

magnitude,

$$H = h \cdot \frac{100 \text{ km/s}}{\text{Mpc}} \approx h \cdot \frac{10^{-10}}{\text{yr}}. \tag{2.1}$$

Though theorists have toyed with the idea of detecting this for at least half a century [120], it has remained hopelessly outside the reach of direct measurement on reasonable timescales, perhaps until very recently [8, 87, 92]. Indeed, it is only recently that surveys have begun collecting high-quality spectroscopy data for fixed sources repeatedly over many years, though this has largely been done with an interest in the reverberation mapping of Active Galactic Nuclei (AGNs) rather than redshift evolution [111, 131, 145]. We utilize the catalogued data of one such recently-completed survey, the Australian Dark Energy Survery (OzDES) [91, 145], to investigate the time evolution in the redshifts of 1457 distinct sources in order to assess the empirical standing of the predictions of the geometric model of SFDM considered herein, ultimately placing tentative constraints on the parameters of the theory in (2.52).

This chapter, which primarily presents the content of my work in [144], is organized as follows. The following section presents and discusses the theory under consideration, deriving the general redshift adjustment as well as its specialization to the cosmological context. Section 2.2 describes the data set under scrutiny and formulates the computational problem of and approach to extracting redshifts from the catalogued spectroscopy data. Section 2.3 presents the results of our analysis, largely contained in Figures 2.3-2.5, and the constraints derived on the theory. Section 2.4 reflects on the work and puts forward some concluding remarks.

2.1 Geometric SFDM Theory

2.1.1 A Geometric Picture of Scalar Field Dark Matter

Recall that the Einstein-Hilbert action S of general relativity is a functional of the Lorentzian metric g on M given by

$$S[g] = \int_U R - 2\Lambda \, dV. \tag{2.2}$$

As reviewed in Section 1.2.4, this action characterizes vacuum general relativity (with cosmological constant) in that requiring g to a be a critical point of S for every U is equivalent to the vacuum Einstein equation $G + \Lambda g = 0$. Furthermore, this provides a particularly compelling picture due to a result of Lovelock (building on Cartan and Weyl) indicating that S is, in fact, the unique coordinate-invariant action up to quadratic in the derivatives of g. This latter description might then be aptly thought of as an axiom which one could adopt to characterize general relativity. Bray [19] demonstrated in 2010 that minimally relaxing this axiom to incorporate a general connection ∇ of the spacetime manifold– by allowing $S[g, \nabla]$ to be quadratic in $g_{ij,k}$ and the connection coefficients Γ_{ijk} and their derivatives $\Gamma_{ijk,l}$– generically leads to the inclusion of a massive free scalar field source term in the Einstein equation, thereby providing a geometric motivation for considering a scalar field as a potential dark matter candidate.

Let us review this result. As discussed in Section 1.2.5, a general connection ∇ on a semi-Riemannian manifold (M, g) can be entirely characterized by its difference tensor D with respect to the Levi-Civita Connection $\overline{\nabla}$,

$$D(X, Y, Z) = \langle \nabla_X Y, Z \rangle - \langle \overline{\nabla}_X Y, Z \rangle, \tag{2.3}$$

so the variation of an appropriate action $S[g, \nabla]$ with respect to ∇ can be carried out by varying D. The difference tensor can itself be understood in terms of the

metric compatibility tensor M_c and the torsion tensor T. Together, these entirely characterize D, and hence ∇, according to

$$D(X,Y,Z) = \frac{1}{2}\big[T(X,Y,Z) - T(Y,Z,X) + T(Z,X,Y) \\ -M_c(X,Y,Z) + M_c(Y,Z,X) + M_c(Z,X,Y)\big]. \tag{2.4}$$

Varying ∇, then, is equivalent to independently varying M_c and T.

In accordance with Bray's axiom, we wish to consider actions $S[g,\nabla]$ which are quadratic in the metric derivatives $g_{ij,k}$ as well as the connection coefficients Γ_{ijk} and their derivatives $\Gamma_{ijk,l}$, extending the similar axiomatization of the Einstein-Hilbert action in terms of g alone. Due to the obstruction that squares of derivatives of the form $\overline{\nabla}D$ (or ∇D) violate the axiom by including terms quadratic in metric second derivatives, Bray conjectured ([19], Conjecture 1) that the most general means of introducing a squared derivative of D in the action (so as to obtain nontrivial second order equations of motion for D) in keeping with the axiom was through terms of the form $|d\omega|^2$, where ω is the fully antisymmetric part of D (which now no longer has any metric terms). That is, ω is a 3-form satisfying

$$\omega(X,Y,Z) = \frac{1}{6}\left[T(X,Y,Z) + T(Y,Z,X) + T(Z,X,Y)\right]. \tag{2.5}$$

If this is to be the only means of introducing derivatives of D, the remaining contribution of D to $S[g,\nabla]$ is quadratic in D itself.

The Simplest Case

At this point, Bray restricts to the representative simplest case that D is entirely described by its antisymmetric part, $D = \omega$, for illustrative purposes. This results in the action

$$S[g,\nabla] = \int_U \left[R - 2\Lambda - 2c_1|d\omega|^2 - 2c_2|\omega|^2\right] dV. \tag{2.6}$$

For the purpose of carrying out the variations, it is convenient to recast the role of ω in terms of the vector field $w := (\star\omega)^*$. This turns (2.6) into

$$S[g, \nabla] = \int_U \left[R - 2\Lambda + 2c_1(\text{div } w)^2 + 2c_2|w|^2\right] dV. \tag{2.7}$$

The variation of (g, ∇) in this action is equivalent to varying (g, w), so this will be our approach. As in Section 1.2.4, considering any compactly supported (in U) one-parameter variation of w given by $s \mapsto w(s)$ with $\dot{w} := \frac{d}{ds}\big|_{s=0} w(s)$, that $w(0)$ (abbreviated to w) is at a critical point of S requires

$$\begin{aligned} 0 = \frac{d}{ds}\bigg|_{s=0} S &= \int_U \left[4c_1(\text{div } w)(\text{div } \dot{w}) + 4c_2\langle w, \dot{w}\rangle\right] dV \\ &= \int_U \langle -4c_1\overline{\nabla}(\text{div } w) + 4c_2 w, \dot{w}\rangle \, dV, \end{aligned} \tag{2.8}$$

where we've utilized the divergence theorem and dispensed with the boundary term due to the variation's being compactly supported in U. This must hold for every U and every possible variation vector field $\dot{w}$, allowing us to deduce

$$4c_1\overline{\nabla}(\text{div } w) = 4c_2 w. \tag{2.9}$$

We'll assume $c_1 \neq 0$ so as to obtain a nontrivial theory for w. Taking the divergence of both sides of this equation and defining $m^2 := c_2/c_1$ and $\phi := \sqrt{c_1}(\text{div } w)$ (any here coefficient will do at the moment, but $\sqrt{c_1}$ yields the standard stress-energy tensor later) leads us to the Klein Gordon equation,

$$\Box\phi = m^2\phi. \tag{2.10}$$

Though we've obtained the Klein-Gordon equation with clever identifications of m and ϕ, it remains to show that variation with respect to g yields a contribution equivalent to the free scalar field stress-energy tensor (1.45) under these same clever identifications. From our work in Section 1.2.4, we see that varying g in the action

(2.7) will yield the Einstein equation with a contribution effectively equivalent to matter with Lagrangian density

$$\mathcal{L}_{\text{eff}} = 2c_1(\text{div } w)^2 + 2c_2|w|^2. \tag{2.11}$$

We consider a variation $s \mapsto g_s$ with $g := g_0$ and $h = \dot{g}_s := \frac{d}{ds}\big|_{s=0} g_s$ and seek to compute

$$\frac{d}{ds}\left[\mathcal{L}_{\text{eff}} dV\right]\big|_{s=0} = \dot{\mathcal{L}}_{\text{eff}} dV + \mathcal{L}_{\text{eff}} d\dot{V} = \left[\dot{\mathcal{L}}_{\text{eff}} + \frac{1}{2}\langle \mathcal{L}_{\text{eff}} g, h\rangle\right] dV. \tag{2.12}$$

To compute $\dot{\mathcal{L}}_{\text{eff}}$, we first note that $|w|^2 = g_s(w,w)$, so the second term in (2.11) yields $2c_2 h(w,w) = 2c_2\langle w^* \otimes w^*, h\rangle$ upon differentiation. The first term requires a bit more work. Noting div $w := \text{tr}(v \mapsto \overline{\nabla}_v w)$, we can write the divergence in coordinates as

$$\text{div } w = \partial_i w^i + w^j \overline{\Gamma}_{ij}^{\ i}. \tag{2.13}$$

The first term $\partial_i w^i$, being independent of g_s, does not contribute to the derivative, so we need only understand $\dot{\overline{\Gamma}}_{ij}^{\ i}$. From the formula (1.16) for $\overline{\Gamma}_{ijk}$, we find

$$\dot{\overline{\Gamma}}_{ij}^{\ k} = \frac{1}{2} g^{kl}[h_{jl,i} + h_{il,j} - h_{ij,l}] - \frac{1}{2} h^{kl}[g_{jl,i} + g_{il,j} - g_{ij,l}]. \tag{2.14}$$

Summing over $k = i$, the summations on the right become symmetric in i and l, yielding a considerable simplification to

$$\begin{aligned} \dot{\overline{\Gamma}}_{ij}^{\ i} &= \frac{1}{2} g^{il} h_{il,j} - \frac{1}{2} h^{il} g_{il,j} \\ &= \frac{1}{2}\partial_j \left[g^{il} h_{il}\right] \\ &= \frac{1}{2}\partial_j(\langle g, h\rangle). \end{aligned} \tag{2.15}$$

Combining this with (2.13), we obtain the coordinate-free result

$$\frac{d}{ds}[\text{div } w]\big|_{s=0} = \frac{1}{2}\langle w, \overline{\nabla}\langle g, h\rangle\rangle. \tag{2.16}$$

Hence we may write

$$\begin{aligned}\frac{d}{ds}[(\text{div } w)^2]\big|_{s=0} &= \langle (\text{div } w)w, \overline{\nabla}\langle g, h\rangle\rangle \\ &= \text{div}[(\text{div } w)\langle g, h\rangle w] - (\text{div } w)^2\langle g, h\rangle \\ &\quad - \langle w, \overline{\nabla}(\text{div } w)\rangle\langle g, h\rangle. \end{aligned} \tag{2.17}$$

Discarding the first term as an overall divergence, we've found

$$\int_U \dot{\mathcal{L}}_{\text{eff}} dV = \int_U \Big\langle 2c_2 w^* \otimes w^* - 2c_1 \left[(\text{div } w)^2 + \langle w, \overline{\nabla}(\text{div } w)\rangle\right] g, h \Big\rangle \, dV. \tag{2.18}$$

Combining this with $\int_U \mathcal{L}_{\text{eff}} d\dot{V}$ as in (2.12), then, we deduce the form of the effective stress-energy tensor contributing to the Einstein equation to be

$$T_{\text{eff}} = 2c_2 w^* \otimes w^* + c_1 \left[\frac{c_2}{c_1}|w|^2 - (\text{div } w)^2 - 2\langle w, \overline{\nabla}(\text{div } w)\rangle\right] g. \tag{2.19}$$

Having completed the variation over g, we are free to use the identity (2.9) obtained from the variation over w and substitute in terms of $\phi = \sqrt{c_1}(\text{div } w)$ and $m^2 = c_2/c_1$. These together indicate $\overline{\nabla}\phi = \sqrt{c_1} m^2 w$ (and hence $d\phi = \sqrt{c_1} m^2 w^*$), turning the above into

$$T_{\text{eff}} = 2\frac{d\phi \otimes d\phi}{m^2} - \left[\frac{|d\phi|^2}{m^2} + \phi^2\right] g. \tag{2.20}$$

This is precisely our standard form for the stress-energy tensor of a free scalar field (1.45). Together with the fact that this choice of ϕ satisfies the Klein-Gordon equation, this means that the fully geometric action (2.6) is entirely equivalent to the Einstein equation in the presence of a free scalar field matter source.

Notice that taking the hodge star of $d\phi = \sqrt{c_1} m^2 w^*$ allows us to identify ω, and thereby D, in terms of ϕ, so we may write

$$D(X, Y, Z) = \frac{1}{\sqrt{c_1} m^2}(\star d\phi)(X, Y, Z). \tag{2.21}$$

The quantity $\sqrt{c_1}m^2$ is apparently an additional free parameter of the theory relating the scalar field to the geometry of the connection.

A More General Action

If one is to take this geometric picture seriously as a framework giving rise to dark matter, we should consider the question of what the connection ∇ indicates physically: what is the physical distinction between this theory and one incorporating the Levi-Civita connection? The most natural hypothesis is that ∇ provides the geodesics along which test particles and light propagate, according to the coordinate geodesic equation (1.15), copied below:

$$\ddot{\gamma}^k + \Gamma_{ij}{}^k \dot{\gamma}^i \dot{\gamma}^j = 0. \tag{2.22}$$

In the case of the simplest nontrivial admissible action (2.6) considered above, observe that D was restricted to be entirely antisymmetric at the outset. This ensures that the geodesics of ∇ and $\overline{\nabla}$ are, in fact, *identical* by the same antisymmetry in $\Gamma_{ijk} - \overline{\Gamma}_{ijk}$. That is, this simplest incarnation yields standard general relativity with the only primary modification being the addition of an effective scalar field source, even upon imbuing the modified connection with physical significance (though more general parallel propagation would be adjusted).

While it demonstrates well the generic emergence of an effective scalar field, the action (2.6) is but one choice of many in keeping with Bray's axiom and conjecture, and different choices will have their own version of the connection relation (2.21) coupling to the Einstein Klein-Gordon system. If the content of the connection is physical, then, these different choices will yield different physics. In the interest of exploring the range of physical phenomena this geometric picture might give rise to, then, here we would like to consider the next-simplest case. Adopting the hypothesis that ∇ manifests physically in the determination of geodesic trajectories of test

particles, we're led to the expectation that ∇ should be metric compatible (so $M_c = 0$), as this is the only geometrically natural means of enforcing the special relativistic constraint that geodesics preserve timelike or null behavior. Under this expectation, (2.4) implies that ∇ is entirely characterized by the torsion tensor T.

The simplest extension of the previously considered case of a fully antisymmetric T is to now allow T to have, in addition to its fully antisymmetric part, a nontrivial trace (due to its definitional antisymmetry in the second and third slots, it can only have one), which we describe in terms of the 1-form α given in coordinates by

$$\alpha_j = T_{ij}{}^{i}. \tag{2.23}$$

When $M_c = 0$ and T is entirely characterized by this trace form and its antisymmetric part 2ω, we may write the difference tensor as

$$D(X,Y,Z) = \omega(X,Y,Z) + \frac{1}{3}\left[\alpha(Y)\langle X,Z\rangle - \alpha(Z)\langle X,Y\rangle\right], \tag{2.24}$$

and the most general associated action becomes

$$S[g,\nabla] = \int_U \left[R - 2\Lambda - 2c_1|d\omega|^2 - 2c_2|\omega|^2 - 2c_3|\alpha|^2 + 2c_4\langle\star\omega,\alpha\rangle\right] dV, \tag{2.25}$$

where $\star$ again denotes the hodge star operation and the constants c_i are parameters of the theory. As before, it is convenient to recast the roles of ω and α in terms of the vector fields $w := (\star\omega)^*$ and $v := \alpha^*$ turning (2.25) into

$$S[g,\nabla] = \int_U \left[R - 2\Lambda + 2c_1(\text{div } w)^2 + 2c_2|w|^2 - 2c_3|v|^2 + 2c_4\langle w,v\rangle\right] dV. \tag{2.26}$$

As before, we vary over each of v and w (independently) in place of ∇. The easier of these is v– considering any one-parameter variation of v given by $s \mapsto v(s)$ with $\dot{v} := \frac{d}{ds}\big|_{s=0} v(s)$, that $v := v(0)$ is at a critical point of S requires

$$0 = \frac{d}{ds}\bigg|_{s=0} S = \int_U \langle -4c_3 v + 2c_4 w, \dot{v} \rangle dV$$

for every choice of variation, and hence for every U and every possible variational vector field $\dot{v}$ (compactly supported in U). This requires the relation

$$2c_3 v = c_4 w \tag{2.27}$$

to hold at a critical configuration of v and w. The same procedure for varying w yields

$$\begin{aligned} 0 &= \int_U \left[4c_1 (\text{div } w)(\text{div } \dot{w}) + \langle 4c_2 w + 2c_4 v, \dot{w} \rangle \right] dV \\ &= \int_U \langle -4c_1 \overline{\nabla}(\text{div } w) + 4c_2 w + 2c_4 v, \dot{w} \rangle dV, \end{aligned}$$

utilizing the divergence theorem and dispensing with the boundary term as usual. Hence a critical configuration must also satisfy

$$2c_1 \overline{\nabla}(\text{div } w) = 2c_2 w + c_4 v = \left(2c_2 + \frac{c_4^2}{2c_3} \right) w. \tag{2.28}$$

Taking the divergence of both sides of this equation and defining $m^2 := \frac{1}{2c_1}\left(2c_2 + \frac{c_4^2}{2c_3}\right)$ and $\phi := \sqrt{c_1}(\text{div } w)$ leads us to the Klein Gordon equation,

$$\Box \phi = m^2 \phi. \tag{2.29}$$

Having identified our effective scalar field ϕ and mass parameter m, we are in a position to unravel our equations to obtain the form of the connection. Taking the metric dual of (2.28) yields

$$d\phi = \sqrt{c_1} m^2 (\star \omega), \tag{2.30}$$

while further taking the hodge star gives

$$(\star d\phi) = \sqrt{c_1} m^2 \omega. \tag{2.31}$$

This identifies the first term on the righthand side of (2.24) in terms of ϕ, and we may similarly identify the latter terms by substituting (2.30) into the metric dual of (2.27), obtaining

$$\alpha = \frac{c_4}{2c_3}(\star\omega) = 3Cd\phi, \tag{2.32}$$

where we've set $C := \frac{c_4}{6c_3\sqrt{c_1}m^2}$. Putting (2.31) and (2.32) into (2.24), then, yields

$$D(X,Y,Z) = \frac{1}{\sqrt{c_1}m^2}(\star d\phi)(X,Y,Z) + C\left[d\phi(Y)\langle X,Z\rangle - d\phi(Z)\langle X,Y\rangle\right]. \tag{2.33}$$

This completes the variation with respect to ∇. Equation 2.33 characterizes the connection's novel features in terms of ϕ. Though we also have the Klein-Gordon equation, to fully close the system we must, as before, also vary with respect to g, utilizing the effective Lagrangian density

$$\mathcal{L}_{\text{eff}} = c_1(\text{div } w)^2 + c_2|w|^2 - c_3|v|^2 + c_4\langle w, v\rangle. \tag{2.34}$$

This again results in the Einstein equation with the free scalar field effective stress-energy tensor source (1.45), given the above identifications of ϕ and m. As this proceeds very similarly to the prior action (2.6), we do not repeat the computation. In a similar vein to the free parameters m and $\sqrt{c_1}m^2$, this theory contains the additional parameter C which couples ϕ to new behavior in geodesics, with the case $C = 0$ reducing to the simpler theory of (2.6).

2.1.2 A General Adjustment to Gravitational Redshift

Let us investigate the implications of (2.33) for geodesics. We first restate this result explicitly in terms of the connections:

$$\langle\nabla_X Y, Z\rangle = \langle\overline{\nabla}_X Y, Z\rangle + (*d\phi)(X,Y,Z) + C\left[Y(\phi)\langle X,Z\rangle + Z(\phi)\langle X,Y\rangle\right] \tag{2.35}$$

When evaluating whether a given curve through spacetime is a geodesic, one is interested in $\nabla_T T$ with T the tangent vector field to the curve, for which the hodge

star term above is null by antisymmetry:

$$\langle \nabla_T T, Z\rangle = \langle \overline{\nabla}_T T, Z\rangle + C\left[T(\phi)\langle T, Z\rangle + Z(\phi)|T|^2\right]. \tag{2.36}$$

Observing $Z(\phi) = \langle \nabla\phi, Z\rangle$ (recall that $\nabla\phi = \overline{\nabla}\phi = \text{grad }\phi$ is constructed out of the metric g independently of the connection), we notice that the entire righthand side may be written in the form $\langle \cdot\, , Z\rangle$, and so nondegeneracy of the metric allows us to deduce

$$\nabla_T T = \overline{\nabla}_T T + C\left[T(\phi)T + |T|^2\nabla\phi\right]. \tag{2.37}$$

Supposing that T is the tangent vector field to a geodesic of the Levi-Civita Connection $\overline{\nabla}$ (so that $\overline{\nabla}_T T = 0$), then, we've found that

$$\nabla_T T = C\left[T(\phi)T + |T|^2\nabla\phi\right]. \tag{2.38}$$

In general, this equation means that a geodesic of $\overline{\nabla}$ is no longer a geodesic of ∇, since the righthand side is not universally 0 so long as $C\nabla\phi \neq 0$, meaning that the *variation* of the dark matter scalar field ϕ can impact the trajectories of test particles beyond its usual gravitational influence mediated by the metric. Considering the particular case of a null geodesic to $\overline{\nabla}$ to understand implications for light, the result further reduces to

$$\nabla_T T = CT(\phi)T. \tag{2.39}$$

In this final case, that $\nabla_T T$ is parallel to T means that the trajectory giving rise to T is still that of a geodesic, but its geodesic *parameterization* has changed. This parameterization is what determines the gravitational redshift of light following the trajectory in question, leading us to the potential for an easily observable signal in redshifts. Let us compute the general adjustment to the gravitational redshift in this theory before specializing to the standard FLRW cosmology.

If $\gamma : I \to M$ (for some interval $I \subset \mathbb{R}$) is a null geodesic of $\overline{\nabla}$, we wish to compute how γ should be reparameterized according to a reparameterizing function

$s \mapsto \tau(s)$ to obtain a geodesic $\tilde{\gamma}(s) := \gamma(\tau(s))$ of ∇. Then $\tilde{\gamma}'(s) = \tau'(s)\gamma'(\tau(s))$, and we find

$$\begin{aligned}
\nabla_{\tilde{\gamma}'(s)}\tilde{\gamma}'(s) &= \nabla_{\tilde{\gamma}'(s)}\left[\tau'(s)\gamma'(\tau(s))\right] \\
&= \left[\nabla_{\tilde{\gamma}'(s)}\tau'(s)\right]\gamma'(\tau(s)) + \tau'(s)\left[\nabla_{\tilde{\gamma}'(s)}\gamma'(\tau(s))\right] \\
&= \tau''(s)\gamma'(\tau(s)) + (\tau'(s))^2\nabla_{\gamma'(\tau(s))}\gamma'(\tau(s)) \\
&= \left[\tau''(s) + C(\tau'(s))^2\gamma'(\tau(s))[\phi]\right]\gamma'(\tau(s)) \\
&= \left[\tau''(s) + C\tau'(s)(\phi\circ\tilde{\gamma})'(s)\right]\gamma'(\tau(s)),
\end{aligned} \tag{2.40}$$

where we have used (2.39) to replace $\nabla_{\gamma'(\tau(s))}\gamma'(\tau(s))$ as well as that the action of $\tilde{\gamma}'(s)$ on ϕ results in $(\phi\circ\tilde{\gamma})'(s)$ by definition of the action of a tangent vector on a function. Hence, requiring that this be 0 so that $\tilde{\gamma}$ is a geodesic of ∇ leads us to an ODE for $\tau(s)$:

$$\tau''(s) + C\tau'(s)(\phi\circ\tilde{\gamma})'(s) = 0. \tag{2.41}$$

The general solution satisfies

$$\tau'(s) = Ke^{-C\phi(\tilde{\gamma}(s))}, \tag{2.42}$$

with $K \in \mathbb{R}$ an arbitrary constant.

If observers at the points $p_1 = \tilde{\gamma}(s_1)$ and $p_2 = \tilde{\gamma}(s_2)$ following worldlines with tangent vectors T_1 and T_2 (in the cosmological case that follows, T_1 and T_2 are both the tangent vector field $\frac{\partial}{\partial t}$ to comoving observers) measure the frequency of a light ray propagating along $\tilde{\gamma}$, they measure frequencies proportional to $\tilde{\omega}_i = \langle T_i, \tilde{\gamma}'(s_i)\rangle$,

meaning that between them they observe a redshift

$$\begin{aligned} 1+\tilde{z} = \frac{\tilde{\omega}_1}{\tilde{\omega}_2} &= \frac{\langle T_1, \tilde{\gamma}'(s_1)\rangle}{\langle T_2, \tilde{\gamma}'(s_2)\rangle} \\ &= \frac{\tau'(s_1)}{\tau'(s_2)} \cdot \frac{\langle T_1, \gamma'(\tau_1)\rangle}{\langle T_2, \gamma'(\tau_2)\rangle} \\ &= e^{C[\phi(p_2)-\phi(p_1)]} \cdot \frac{\omega_1}{\omega_2} \end{aligned}$$

$$\boxed{1+\tilde{z} = (1+z)e^{C[\phi(p_2)-\phi(p_1)]},} \tag{2.43}$$

where quantities with a tilde correspond to light propagating along $\tilde{\gamma}$ (in accordance with ∇) and quantities without a tilde correspond to light propagating along γ (in accordance with $\overline{\nabla}$). Equation (2.43) is finally the general adjustment to the gravitational redshift expected within this geometric framework for scalar field dark matter, assuming the next-to-simplest admissible action $S[g, \nabla]$ yielding metric compatibility (equation (2.25)) and that the nontrivial connection ∇ manifests physically in the trajectories of test particles. It indicates that the redshift expected under ∇ is, in general, that expected under $\overline{\nabla}$ modulated by the change in the value of the scalar field ϕ between observation and emission, with the degree of modulation set by the free parameter C of the theory (which evidently has units, under $\hbar = c = 1$, of inverse energy squared, inverse to those of ϕ).

2.1.3 Redshift Adjustments in Cosmology

With the general result in hand, we now specialize to the standard cosmological model, reviewed in Section 1.3.5, of a spatially flat FLRW spacetime $M = \mathbb{R} \times \Sigma$ on which the metric locally takes the form

$$g = -dt^2 + a(t)^2 \left[dx^2 + dy^2 + dz^2\right], \tag{2.44}$$

coupled to a scalar field ϕ (and other standard matter components) through the Einstein and Klein-Gordon equations. As we've seen, $\phi = \phi(t)$ is purely a function

of t, and the Klein Gordon equation takes the form of a damped oscillator equation

$$\ddot{\phi} + 3H\dot{\phi} + m^2\phi = 0, \tag{2.45}$$

where $H = \frac{\dot{a}}{a}$ is the Hubble parameter. The mass parameter m directly takes the role of the oscillator's (angular) frequency, while the damping term $3H\dot{\phi}$ leads the amplitude to decay (once $H \lesssim m$) proportionally to $a^{-3/2} = (1+z)^{3/2}$ as the universe expands.

Under the usual approximation of cosmological averaging for the purposes of understanding redshifts of distant sources, then, we expect that the difference $\phi(p_2) - \phi(p_1)$ in (2.43) relevant to the time-varying observed redshift $\tilde{z}$ of a source at fixed comoving distance corresponding to a standard redshift z has two distinct oscillating components: the oscillation of ϕ at observation (the point p_2) at frequency m and the oscillation of ϕ at emission (the point p_1) at the redshifted frequency $\frac{m}{1+z}$. The latter frequency is shifted precisely by the standard cosmological factor $1 + z = \frac{a(t_2)}{a(t_1)}$ because it is purely due to the universe's expansion, not geodesic parameterization– the distance between two light pulses emitted by the source at subsequent crests of ϕ expands by this factor by the time they reach the observer. The logarithm of (2.43),

$$\ln(1 + \tilde{z}) = \ln(1 + z) + C\left[\phi(p_2) - \phi(p_1)\right], \tag{2.46}$$

indicates that these oscillatory frequencies m and $\frac{m}{1+z}$ should appear directly in the quantity $\ln(1 + \tilde{z})$, potentially making this signal easy to pick out via Fourier techniques applied to $\ln(1+\tilde{z})$. Moreover, the amplitudes of these oscillations, while not set absolutely due to the unconstrained parameter C, should be correlated in a specific way due to the $a^{-3/2}$ decay of ϕ– more distant sources at fixed comoving distance should exhibit larger oscillations in a directly quantifiable manner.

The above characteristics can be well-captured by modeling the repeated measuring, over laboratory time t, of the observed redshift $\tilde{z}(t)$ of an object at fixed co-

moving distance by making the identifications $\phi(p_2) \sim \sin(mt)$, where we've shifted $t = 0$ to eliminate any phase and suppressed the present-day amplitude, and $\phi(p_1) \sim (1+z)^{3/2} \sin(\frac{mt}{1+z} - \delta)$, where δ is a phase shift arising due to the time delay between emission and observation, set by the precise distance to the source. As an order of magnitude estimate, $\delta \sim mD$, with D the comoving distance, so that this shift is sensitive to variations in distance on the order of $\frac{2\pi}{m}$. Inserting these identifications into (2.46) yields the qualitative expectation

$$\ln\left(\frac{1+\tilde{z}}{1+z}\right) \propto \sin(mt) - (1+z)^{3/2} \sin\left(\frac{mt}{1+z} - \delta\right). \tag{2.47}$$

In regards to the timescales of these oscillations, we express the mass parameter m in units of 10^{-22} eV as m_{22} and observe that

$$m = m_{22} \cdot 10^{-22} \text{ eV} \approx \frac{2\pi m_{22}}{(1.3 \text{ yrs})} \tag{2.48}$$

(recall we've set $\hbar = 1$), so that the frequency m corresponds to an oscillatory period of about $\frac{1.3}{m_{22}}$ years. Since observational constraints largely point to $m_{22} \gtrsim 1$, we conclude that typical treatments of a cosmological scalar field as a viable primary dark matter candidate would lead to redshift oscillations in the theory developed here with period on the order of ~ 1 year or shorter as well as larger-amplitude oscillations at a redshifted period $1 + z$ times longer. That these oscillations might occur on terrestrial timescales is a remarkable feature allowing the possibility of a comparatively simple means of detecting a signal from this instantiation of geometric scalar field dark matter.

Before turning to some preliminary analysis of redshift data, we reflect on how one would expect this signal to emerge in practice. We first observe that, though the oscillation amplitude discussed above should increase proportionally to $(1+z)^{3/2}$ as we look at more distant sources, this does not mean that we should expect exorbitantly

large oscillations in the logarithm of the CMB temperature (some $1100^{3/2} \sim 3.6 \times 10^4$ times larger than any present oscillations), the most distant source we can observe, even over long timescales. This is because the CMB is not emitted at a fixed comoving distance, but rather at a fixed (range of) time, so that $\phi(p_1)$, the scalar field at emission (appropriately averaged over emission times according to the recombination visibility function), does not change as we repeatedly observe the CMB.

A separate consideration arises for spatially extended sources, those larger than a few times $\frac{2\pi}{m} \sim \frac{1}{m_{22}}$ lyr. Light received from such sources at a given observation time would have been emitted over a range of emission times spanning several periods of the oscillation in $\phi(p_1)$, washing out this contribution to $\ln(1+\tilde{z})$ (while perhaps broadening spectral peaks)– in (2.47), this amounts to summing many different spectra with an effective continuum of values of δ that span a range much larger than 2π, so that the upward and downward shifts due to the second term in (2.47) largely negate each other. Such extended sources, of course, are generally all that can be made out at even mildly high redshifts ($z \gtrsim 0.1$), likely nullifying the $(1+z)^{3/2}$ growth in practical observations. The only likely exceptions to this nullification are supernovae redshifts, though these are more difficult to monitor given their short lifespan. On the other hand, since the second term in (2.47) is expected to wash out for extended sources, oscillations in such sources would be entirely due to conditions at the point of observation– that is, they should be *coherent* across all such sources, giving a powerful means of testing our theory.

In all cases the oscillation in $\phi(p_2)$, the scalar field at observation, should remain present, provided only that observations' exposure times are much shorter than $2\pi/m$. At small redshifts ($z \lesssim 0.1$), however, the amplitude of this oscillation (for a compact source) becomes sensitive to the source's precise distance due to the potential for both constructive and destructive interference between $\phi(p_1)$ and $\phi(p_2)$,

or the two sinusoids in (2.47). Indeed, standard trigonometric manipulations[1] yield that in the limit $z << 1$, (2.47) effectively becomes

$$\ln\left(\frac{1+\tilde{z}}{1+z}\right) \propto 2\sin\left(\frac{z}{2+2z}mt+\frac{\delta}{2}\right)\cos\left(mt-\frac{\delta}{2}\right), \tag{2.49}$$

wherein the cosine term gives the expected oscillation at frequency m, but modulated by the much more slowly-varying sine term setting the amplitude in a manner highly sensitive to the value of δ. For $m_{22} \gtrsim 1$, this amplitude is sensitive to moving a source on the scale of a lightyear or less, meaning that amplitudes of the oscillations in low-redshift sources would be expected to be somewhat haphazardly distributed even at effectively fixed z. The factor of 2 here means that the maximum amplitude is twice that expected from the $\sin(mt)$ term alone in (2.47), arising from potentially constructive interference. We comment that, at the larger end of the redshifts for which (2.49) still gives a qualitatively correct picture ($z \sim 0.2-0.3$, though it would again be difficult to observe a compact source at such values), it becomes feasible that one might be able to observe both the frequency m oscillations as well as their modulation on reasonable timescales.

From this investigation, then, we take away that in an aggregate view of many redshift variations across many sources, this model leads us to expect those at low redshifts to have oscillation amplitudes scattered between zero and a maximum value set by the parameter C and the present-day amplitude of ϕ, while higher redshift objects, generally being well beyond a lightyear in spatial extent, should exhibit oscillations which consistently attain about half this maximum amplitude. Moreover, oscillations of higher redshifts should be collectively coherent at frequency m. We

[1]
$$A\sin(x)+B\sin(y) = (A+B)\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)+(A-B)\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)$$

remark that these conclusions are all made operating under the assumption that (2.43) may be reasonably applied using the cosmologically averaged geometry of (2.44). Though such assumptions have largely born out well in the standard cosmology, they merit further consideration in this modification, particularly given the large discrepancy generally expected between ϕ and its cosmological average at the points of emission and observation (being in galaxies) and the dependence of some features of this discussion on sub-lightyear scales. While an interesting problem, the resolution to this question is beyond the scope of this work, and we will simply assess whether the averaged predictions have any empirical support. Perhaps the best interpretation of this subsection's discussion is that it provides a heuristic motivation for seeking these signals rather than a robust prediction that they must occur precisely as described.

2.2 Seeking Redshift Variance in OzDES

To make a preliminary assessment as to whether the patterns discussed above are present in extant observational data, we make use of the Australian Dark Energy Survey's (OzDES) second data release [91], which catalogues high-quality redshift and spectroscopy data of some 30,000 sources up to redshift $z \sim 4$, with the highest priority sources being active transients, active galactic nuclei, and supernovae host galaxies. Each source in the catalogue was observed multiple times over the survey's duration from 2013 to 2019, annually between August and January, until the desired quality of redshift was obtainable from that source's stacked spectrum, an appropriately weighted average of all observations of the source of interest. Only this single, aggregate redshift was obtained and reported for each source in the catalogue, though the data release contained the individual spectra for each of their $\sim 375,000$ observations. The catalogue also contained data associated to observations of some 10,000 additional sources to which a redshift could not be confidently assigned, which

we do not consider (in particular, we only considered sources with a redshift quality flag, assigned by OzDES, of at least 3).

As the patterns we seek to evaluate are in the time variation of the redshift $\tilde{z}(t)$ of individual objects, we need to assign a redshift to the individual observations' spectra rather than just each object's stacked spectrum. To have hope of extracting any meaningful representation of periodicity, we require many individual observations for each object we consider, so we restricted to those sources which were observed at least 30 separate times, reducing our data set to 1,457 sources with a total of 98,370 individual observations. For each source, we take the stacked spectrum's redshift reported by OzDES to represent the standard cosmological redshift z, as the averaging process should largely nullify the oscillations in (2.47), provided they occur over the data's 6 year timescale (we should obtain a null result otherwise). As we are ultimately interested in the relative quantity $\ln((1+\tilde{z})/(1+z))$, we use each object's stacked spectrum as a baseline from which we ascertain a relative shift for each observation via template matching techniques.

2.2.1 Identifying Redshift

Though discerning an optimal relative shift may seem like a straightforward task, some care must be taken to do this robustly. We first consider that if the unredshifted "true" spectrum is $f(\lambda)$, then the stacked spectrum is expected to be $g(\lambda) := f((1+z)\lambda)$, and the observed spectrum is expected to be $h(\lambda) := f((1+\tilde{z})\lambda)$. Describing the relative shift via $\alpha := \frac{1+\tilde{z}}{1+z}$, this means we expect $h(\lambda) = g(\alpha\lambda)$, and our computational task is to extract α from the data of h and g reported by OzDES. This will be made simpler with a logarithmic change of variables to turn the multiplicative shift by α into a linear shift by $\ln(\alpha)$. That is, defining $s := \ln(\lambda)$ and re-expressing the spectra as $\bar{h}(s) := h(\lambda(s)) = h(e^s)$ and $\bar{g}(s) := g(\lambda(s)) = g(e^s)$, the identity $h(\lambda) = g(\alpha\lambda)$ translates into

$$\bar{h}(s) = h(e^s) = g(\alpha e^s) = g(e^{s+\ln(\alpha)}) = \bar{g}(s + \ln(\alpha)).$$

Our task is now to extract $\ln(\alpha)$, precisely the quantity in which our theory predicts oscillations, as the horizontal translation between the graphs of $\bar{h}$ and $\bar{g}$.

This is complicated in practice by random variations in noise, differing bulk atmospheric effects across the various observations, and the fact that such effects additionally mean that the spectra could not be consistently calibrated. Indeed, the OzDES documentation[2] indicates: "The spectra are not flux calibrated, not even in a relative sense. This is due to fibre positioning errors, chormatic [*sic*] aberrations from the 2dF corrector, and seeing." Our prescription for identifying the relative shift must therefore make the graphs of $\bar{h}$ and $\bar{g}$ most similar in an appropriate sense in light of these complications. A familiar tool for achieving this in general is the *cross-correlation* between $\bar{h}$ and $\bar{g}$:

$$(\bar{h} \star \bar{g})(\tau) := \int_{\mathbb{R}} \bar{h}(s)\bar{g}(s + \tau)ds, \tag{2.50}$$

a measure of the overlap between the graphs of $\bar{h}$ and the translational shift of $\bar{g}$ by τ to the left. Note that the bars here are part of the function notation, not complex conjugation– all quantities are real. The value of τ which maximizes $\bar{h} \star \bar{g}$ would then be that which optimizes this overlap, providing a natural choice of $\ln(\alpha)$. A nice feature of (2.50) is that the maximizing value of τ is not affected by either vertical shifts or rescalings of either $\bar{f}$ or $\bar{g}$, so that concerns of callibration would be largely immaterial if we could actually work with this quantity.

A practical complication to working with (2.50), however, is that one cannot observe the spectrum over all wavelengths– the spectra with which we are working span about 3700Å-8900Å–, so the integral in (2.50) must be truncated to $\int_a^b \bar{h}(s)\bar{g}(s+$

[2] https://docs.datacentral.org.au/ozdes/overview/dr2/

$\tau)ds$ for some appropriate a and b. Unfortunately, the adjustment imparted to the integral by adding a constant to $\bar{h}$ is now a function of τ, so that the maximizing value of τ is no longer independent of vertical shifts. Moreover, the truncation can bias the maximal τ away from optimal alignment towards those shifts which move larger values of $\bar{g}$ into the integration range. Hence we must modify (2.50) beyond simply truncating.

The bias due to the changing magnitude of $\bar{g}$ over $[a, b]$ can be countered by appropriately normalizing. Setting $\bar{g}_\tau(s) := \bar{g}(s + \tau)$ and noting that $\int_a^b \bar{h}(s)\bar{g}(s + \tau)ds = \int_a^b \bar{h}(s)\bar{g}_\tau(s)ds$ is precisely the $L^2([a, b])$ inner product between $\bar{h}$ and $\bar{g}_\tau$, the most natural normalizing procedure would seem to be dividing by the L^2-norm. This is only strictly necessary for $\bar{g}_\tau$, as $\|\bar{g}_\tau\|$ depends on τ while $\|\bar{h}\|$ does not, but we also normalize $\bar{h}$ because it yields a universally meaningful quantity that can be used to compare the degree of correlation across different observations:

$$C_N(\tau) = \frac{\langle \bar{h}, \bar{g}_\tau \rangle}{\|\bar{h}\|\|\bar{g}_\tau\|} = \frac{\int_a^b \bar{h}(s)\bar{g}_\tau(s)ds}{\sqrt{\int_a^b (\bar{h}(s))^2 ds \cdot \int_a^b (\bar{g}_\tau(s))^2 ds}}. \tag{2.51}$$

This is the *normalized cross-correlation*, a commonplace tool in the evaluation of redshifts [79, 85, 129, 133]. The reasonableness of this quantity as a measurement of the similarity between $\bar{h}$ and $\bar{g}_\tau$ is supported by the Cauchy-Schwartz inequality for $L^2([a, b])$, which indicates that (2.51) has magnitude at most 1, and further that its magnitude is equal to 1 if and only if $\bar{g}_\tau$ is (almost everywhere) a constant multiple of $\bar{h}$, which is almost precisely what we'd like to detect.

The qualifier of "almost" is used because the final concern to address is the potential need for a vertical shift to align the spectra. One approach to addressing concerns of this nature is to subtract away the average values of $\bar{h}$ and $\bar{g}_\tau$ before computing (2.51)– that is, working with $\bar{h} - \frac{1}{b-a}\int_a^b \bar{h}(s)ds$ instead of $\bar{h}$, and similarly for $\bar{g}_\tau$. This is because the average behavior is very much susceptible to calibration

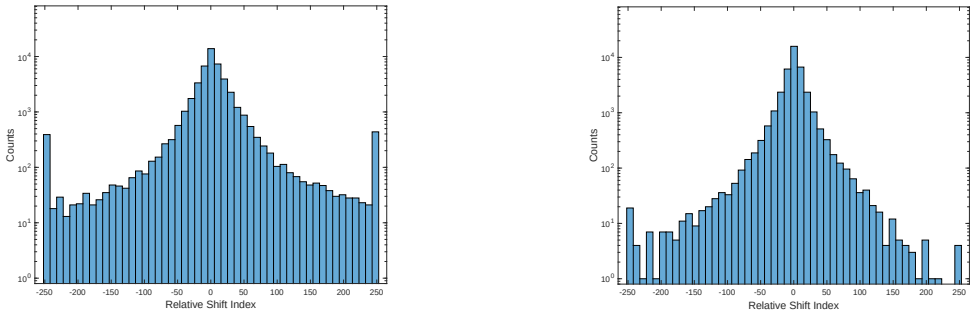

FIGURE 2.1: Incidence rates of relative shifts (after applying the cuts discussed in Section 2.2.2), measured in terms of the 250 steps between 0 and the extreme shifts $\pm\tau_{\max}$, in the cases that (left) constant averages and (right) large-scale Gaussian-weighted moving averages are removed. Note the logarithmic scaling of the vertical axis. The strategy on the right is subject to significantly less of the variation presumably induced by both incompatible calibrations and intrinsically varying continua between exposures.

concerns, and the spectral features by which redshifts are primarily identified are the variations on top of this average behavior anyway. These concerns are true of the average behavior more broadly than that contained in the average values, particularly since AGNs, which make up the bulk of our sources, exhibit varying spectral continua (this was a large part of what OzDES hoped to monitor, after all). Indeed, we've found in a number of cases that an appreciable bias can remain if we only subtract constant averages, leading to much wider variation in $\ln(\alpha)$ (see Figures 2.1 and 2.2). Hence, we subtract a broader characterization of the average behavior, specified below, before evaluating (2.51). Having done this, we identify $\ln(\alpha) = \ln((1+\tilde{z})/(1+z))$ as the optimal value of τ, that which maximizes $C_N(\tau)$.

2.2.2 Programmatic Procedure

The spectra catalogued by OzDES are reported on a wavelength range centered at $\lambda_c = 6295$Å in 5000 steps of width of about 1Å. To carry out the optimization discussed above, we maximized $C_N(\tau)$ among 501 values of τ spanning the range between $\pm\tau_{\max}$, where $\tau_{\max}$ is defined so as to increase the central wavelength λ_c by 50 steps, or $\tau_{\max} \sim 0.0082$– an initial evaluation allowed for shifting λ_c by 200 steps,

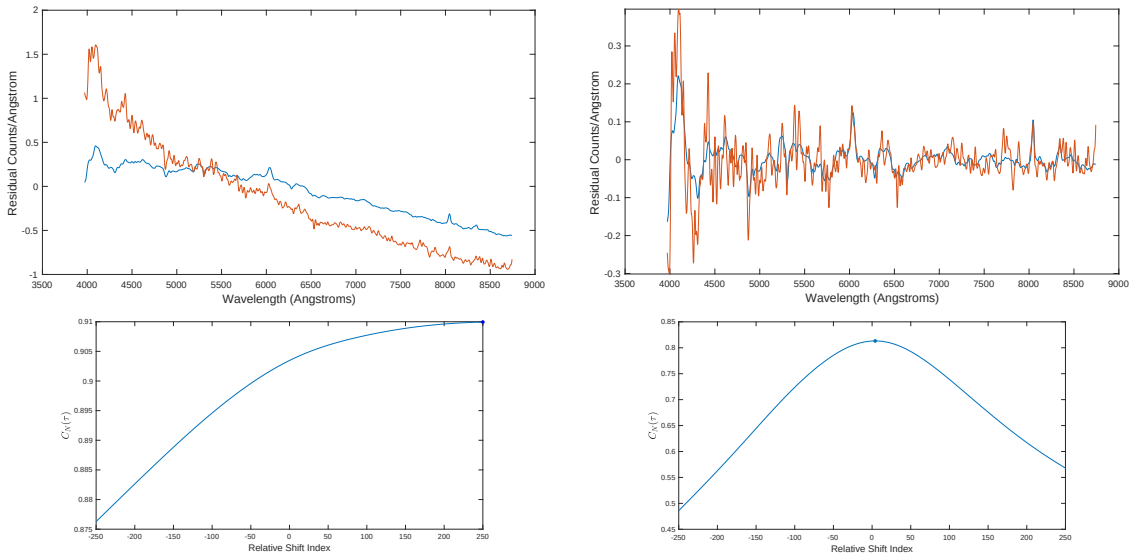

FIGURE 2.2: Top row: residuals of an observation (red) and stacked (blue) spectrum for a AGN source (named SVA1_COADD-2971194281 by OzDES), upon subtraction of (left) constant averages and (right) large-scale Gaussian-weighted moving averages. On the left, one can see the change in calibration and/or continuum between the stacked and observation spectrum, while these are removed on the right, primarily leaving the closely aligned spectral features.
Bottom row: Plots of $C_N(\tau)$ for each of the pairs of spectra in the top row, with maxima emboldened. The maximum occurs at the largest probed shift $\tau_{\max}$ on the left and very near the minimal shift on the right.

but no reasonably-confident shifts fell outside 50. The integration range utilized in computing (2.51) was truncated by 230 steps at the lower end and 171 steps at the higher end so as to leave a buffer region from which data could be shifted into the range as τ is varied, yielding the window 3971Å-8743Å. All integrals needed in (2.51) were computed via the trapezoid rule. All analysis was done with the 35-value Gaussian-weighted moving averages of the observed and stacked spectra h and g (using MATLAB's smoothdata function) to smooth out noise fluctuations occurring on the scale of several angstroms and mollify artifacts which yield large spikes in h, such as cosmic ray residuals [91]. This smoothing is also how we characterized the average behavior to remove, identified as the Gaussian-weighted moving average over 1000 values.

Beyond smoothing, we applied a number of qualitative cuts to the data to address

concerns surrounding poor data quality, reducing our effective data set. We did not consider observations for which the optimal correlation was poor, defined as the maximal $C_N(\tau)$ being less than 0.5, as we took this to mean that spectral features were not strong enough to identify the redshift with confidence. Following [91], we further eliminated those observations which occurred during poor atmospheric conditions, evaluated via the catalogued zero points in the red and blue arms: we required both zero points to be greater than 30, with at least one greater than 31. The OzDES team also visually inspected most of their spectra, identifying a number of recurring spectral artifacts and recording them under the 'QC' keyword in the FITS files– we have ignored all observations which did not receive a flag of 'ok'. Finally, we have removed those observations whose spectra exhibited exorbitant spikes, defined as occurring when the sum of the 25 largest values of $|h|$ was more than 15% of the sum of all values of $|h|$ (after smoothing and subtraction of average behavior), as such spikes exert undue influence on $C_N(\tau)$.

As with any such procedure, the schema outlined here may well still be subject to some pathologies, and it will not perfectly capture the appropriate shift in every case, but we maintain that it should be sufficiently robust to capture consistent trends across a wide array of data.

2.3 Empirical Results

Applying all of the cuts discussed in the previous section leaves us with some 38,575 observations to which we've been able to assign a relative redshift with reasonable confidence, and these are associated to 902 sources which have at least ten admissible relative redshift values remaining. We first investigate the average magnitude of redshift deviations for each of these 902 sources, plotted against each source's baseline (stacked) redshift– see Figure 2.3. To construct each point, then, we average the values of $|\ln(\alpha)|$ across the admissible observations associated to a given source.

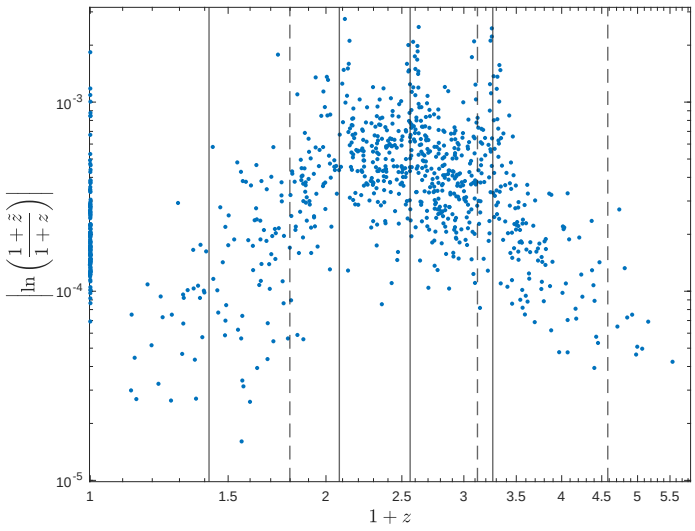

FIGURE 2.3: Mean redshift variation versus baseline redshift. Each point indicates the variation in the relative redshifts across individual exposure spectra for a single source. Also shown by vertical lines are the redshifts at which pertinent emission lines are shifted into (solid lines) and out of (dashed lines) the probed wavelength range. From left to right, these lines correspond to: MgII, H-β, CIII], CIV, MgII, Ly-α, and CIII].

Irrespective of whether the data for the source in question is sufficiently fine to resolve periodic behavior in its redshift variations, the mean value of $|\ln(\alpha)|$ should, in aggregate across many sources, be a meaningful indicator of the amplitudes of oscillations in accordance with (2.47) that occur on timescales of a few years or less. Also overlain on this plot are vertical lines indicating the redshifts at which the most commonly strong emission lines enter or exit our integration interval.

We first note that the cluster of points around $z \approx 0$ in Figure 2.3 corresponds to F stars used for throughput calibration, and these demonstrate a wide array of redshift variations with no readily apparent pattern. The primary feature among the remaining points, which generally correspond to AGNs and supernovae hosts, is the broad arch shape, indicating that redshifts in the range $1 \lesssim z \lesssim 2.5$ have relatively high variance, while both higher and lower values of z yield significantly lower variance, decreasing as z gets farther from this range, eventually reaching the rough scale of our minimum step size, $\frac{\tau_{\max}}{250} \sim 3.3 \cdot 10^{-5}$. Also of note are the three

apparent "columns" exhibiting particularly high variance around $z \sim 1.1, 1.6$, and 2.2, which roughly line up with certain emission lines (CIII], CIV, MgII, and Ly-α) transitioning into or out of the integration domain.

While the haphazard assortment of variances among F stars is not at odds with the general discussion of low z in Section 2.1.3 (nor is it especially strong evidence in favor of it), the same cannot be said for the arch structure present throughout $z > 0.1$– in particular, we do not see the redshift variance level off at and beyond $z \gtrsim 0.1$, as (2.47) led us to expect. Instead, the lines demarcating changes in spectral features paint the picture that these features are the dominant drivers of the variation in our identified shifts, suggesting that this variation is uncertainty inherent to our technique rather than an indication of intrinsically varying redshifts. The variance predicted by (2.46), then, apparently cannot contribute above the $\sim 10^{-4}$ level: this constrains the parameters m and C of the theory of Section 2.1 to satisfy at least one of the order of magnitude constraints

$$m \lesssim 10^{-23} \text{ eV} \quad \textbf{or} \quad m \gtrsim 10^{-18} \text{ eV} \quad \textbf{or} \quad C \lesssim 10 \text{ eV}^{-2}, \tag{2.52}$$

corresponding to oscillations being either too slow to observe on the timescale of our data, too fast to be resolved given the typical instrument exposure time of 40 minutes, or too small to be detectable over the fluctuations induced by spectral features. Note that the constraint for C assumes $\Omega_\phi \sim 0.25$ in a cosmology with Hubble parameter $H \sim 70 \frac{\text{km}}{\text{s·Mpc}}$ (each in order of magnitude). As discussed at the end of Section 2.1.3, these constraints are subject to the caveat that they've assumed the legitimacy of cosmological averaging. More generally, the lack of signal detected here may instead translate to constraints on C based on galactic dark matter densities rather than the cosmological average density, or constraints on m based on the timescales of soliton or "quasiparticle" periods in the Milky Way (generally much longer than $2\pi/m$) [56] rather than the cosmological oscillation timescale– these are more convoluted threads

to follow, and we will not attempt to do so in this work.

We now turn to whether there is any evidence for periodic behavior in $\ln(\alpha)$ as a function of time. Even if, as discussed above, the apparent variations in $\ln(\alpha)$ are largely due to limitations of our technique and the structure of the spectra, their behavior over time could still conceivably encode a preferred frequency extractable via Fourier techniques. No single source has enough observations for this to be done very meaningfully, but the conclusion of Section 2.1.3 that the oscillations of sources at $z \gtrsim 0.1$ should be coherent means that we may probe for an underlying frequency using data from all such sources at once.

In the interest of broadly exploring the available data, we perform this probe with two different data sets: one comprised of all 33,727 observations at $z > 0.1$ to which we've assigned a relative redshift, the other comprised of the subset of 6922 observations which are further constrained to not lie in the range $0.7 < z < 2.5$. This latter restriction is informed by Figure 2.3, which leads us to expect that we should acquire a better signal to noise ratio by excluding these intermediate redshifts. Spectrally, these are the observations which have either the H-β or the Ly-α emission peak squarely contained within their wavelength windows. In Figure 2.4, we plot the relative redshift measure $\ln(\alpha)$ in each of these data sets against the observation timestamps, measured in days since December 31st, 2013, and in Figure 2.5 we plot

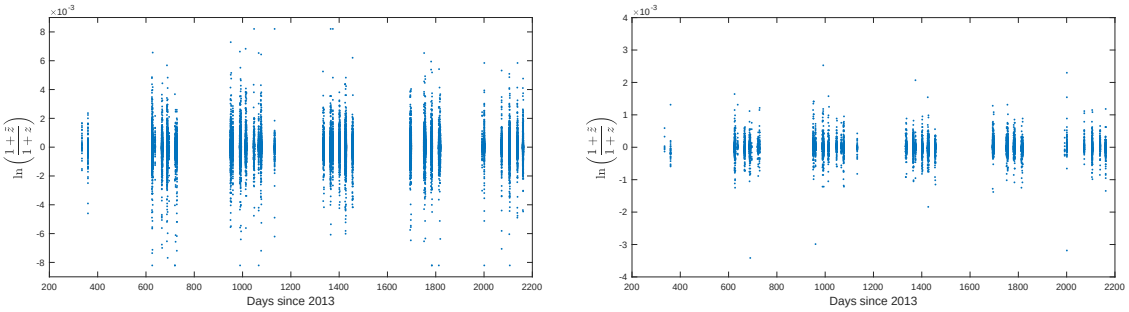

FIGURE 2.4: Relative redshifts plotted over time across all observations with (left) sources with $z > 0.1$ and (right) sources with either $0.1 < z < 0.7$ or $z > 2.5$.

the nonuniform discrete Fourier transforms of each of these up to a frequency of 100 yrs^{-1} (computed with MATLAB's nufft function).

The most notable features in Figure 2.4 are the gaps, both between the yearly observation schedules and between the observation runs in clusters generally a few weeks apart. A zoomed in view would reveal further gaps between the nightly observations in each run, and on all of these scales points are scattered vertically into columns– many points even exist at identical timestamps, as spectroscopy data was taken for many sources simultaneously in each exposure. Though the scatter is much less in the reduced data set on the right (note the vertical axis limits), these qualitative features are present in both. While some columns do appear vertically higher or lower than others at a glance, neither plot exhibits any glaringly obvious periodicity on the whole, though it is difficult to be definitive given the gaps.

Absent any visually obvious periodicity, we turn to the discrete Fourier transforms shown in Figure 2.5. In keeping with our take away from Figure 2.4, here we see that there are no peaks set strongly apart from the noise. This is especially so in the case of the larger, noisier data set of the left plot, where the two largest peaks at $f \approx 0.51\ \text{yrs}^{-1}$ and $f \approx 17.23\ \text{yrs}^{-1}$ are accompanied by several other peaks of similar height (though we observe that nearly all sizable peaks beyond 20 yrs^{-1} seem to be harmonics of the latter). In the smaller data set on the right, the two largest peaks at $f \approx 1.02\ \text{yrs}^{-1}$ and $f \approx 26.35\ \text{yrs}^{-1}$ are marginally more distinguished, but still not exceedingly so. Of course, the data's generally being taken at 1 year intervals means the peaks at 0.51 and 1.02 yrs^{-1} are somewhat suspect, and observing 17.23 $\text{yrs}^{-1} \approx \frac{0.99}{3\ \text{wks}}$ and 26.35 $\text{yrs}^{-1} \approx \frac{1.01}{2\ \text{wks}}$ renders these frequencies suspicious as well. Such patterns continue beyond the plotted range (e.g., peaks appear at $f \approx 365\ \text{yrs}^{-1}$ as well). In any event, this analysis weakly brings out some frequencies of potential interest, but a data set that's more complete in the time domain would be helpful to making definitive conclusions. Again, the absence of a strongly preferred frequency

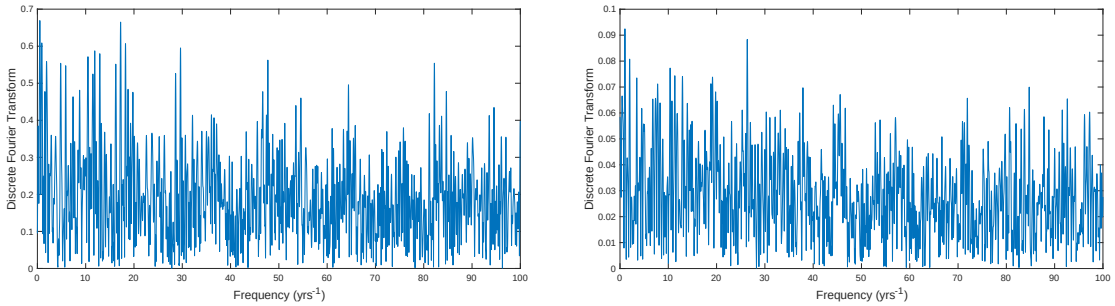

FIGURE 2.5: Magnitudes of the nonuniform discrete Fourier transforms of the two time series of Figure 2.4. The horizontal axis records linear frequency (as opposed to angular).

may either mean that m is incompatible with oscillations on these timescales or that C is sufficiently small that they are obfuscated by the noise, as in (2.52).

2.4 Conclusions

Here we have developed a novel theoretical prediction made by a particular instantiation of the geometric model for scalar field dark matter broadly described by Bray [19], and we have explored its implications for one of the most important cosmological observables, redshifts of distant sources. This pursuit has led us, through equations (2.43) and (2.47), to several features of the theory which readily lend themselves to empirical verification (given the reasonability of cosmological averaging), including broadly coherent oscillations, at the frequency of the scalar field's mass parameter m, in the time evolution of redshifts of sources at mildly high baseline redshifts ($z \gtrsim 0.1$), as well as the dependence of such oscillations' amplitudes on baseline redshift across all values, depending on the spatial scale of the source of interest.

To investigate these predictions, we've drawn on the observations made by the Anglo-Australian Telescope for OzDES, which catalogued spectroscopy data for many thousands of sources observed several times each over the course of six years. By maximizing the cross-correlation (2.51) with the source's stacked spectrum, we have associated to each observation of interest a shift $\alpha = \frac{1+\tilde{z}}{1+z}$ relative to the source's

baseline redshift z reported by OzDES. Comparing the behavior of α across the catalogue, in its dependence on both time and baseline redshift z, to the present theory's predictions, we have not found any compelling evidence that these predictions are born out in empirical data. While this result does not rule out the geometric model for dark matter under consideration, the absence of a signal at the levels of redshift variance probed here have led us to the tentative order of magnitude constraints of equation (2.52) on the free parameters m and C of the theory. At the very least, the analysis culminating in Figure 2.3 is the first to our knowledge establishing the consistent empirical stability of cosmological redshifts over timescales of several years, conservatively at the level of one part (of $1+z$) in a thousand (still some six orders of magnitude too coarse to probe the standard model's order of variation from (2.1)).

While the investigation carried through here has yielded null results, the relative ease of potentially obtaining a positive identification of dark matter through these methods means it remains of interest to probe the time and redshift evolution of α in both more sensitive and complete cosmological data sets. Perhaps the largest shortcoming of the data set utilized here was its discreteness, seen in the large gaps present in Figure 2.4– the 33,727 observation spectra represented there were collected over only 92 nights, and the average individual source we considered only had admissible observations from 14 separate nights across the six years. A more continuous observation schedule would improve both the confidence in the amplitude of α and the ability of Fourier techniques to pick out an underlying frequency. Beyond this, the spatial compactness of supernovae means their oscillations may have larger amplitudes by up to an order of magnitude at higher redshifts (as discussed in Section 2.1.3), so these can also provide a route to improvement. The Time-Domain Extra-Galactic Survey (TiDES) is an upcoming cosmological survey with the capacity to provided a more complete data set informing constraints on this theory with higher frequency observations and smaller seasonal gaps [91, 131].

3

Scalar Field Dark Matter Cosmology

In Section 2.1.1, we detailed Bray's mild generalization of the formalism of general relativity which yields a geometric model of SFDM as an artifact of a nontrivial connection, making and investigating a definitive prediction emergent from one instantiation of the model. While we did not find evidence of this instantiation in the time evolution of cosmological redshifts in the previous chapter, restricted regions of the parameter space and simpler variants more similar to standard general relativity remain plenty viable. Within such models, standard initial conditions for a cosmological scalar field inspired by low-momentum production of ultralight axions lose motivation. In this chapter, we explore the cosmological phenomena supported by relaxing these conditions and considering a kination era of dark matter domination with $\rho_\phi \propto a^{-6}$ in the very early universe. In particular, we investigate impacts on the outcome of big bang nucleosynthesis, the angular power spectrum of CMB temperature anisotropies, and structure growth.

3.1 BBN in SFDM Cosmology

Among the most influential predictions of the standard ΛCDM cosmology are the primordial light element abundances– specifically the abundances of deuterium, ^{3}He, ^{4}He, and ^{7}Li relative to Hydrogen– some minutes after the big bang. Due to the divergence of temperature in the $a \to 0$ limit, early enough on all standard model particles are relativistic and in thermal equilibrium, forming a radiation bath. As the universe expands and the temperature drops, particle-antiparticle pairs annihilate, quarks and gluons combine into baryonic bound states, and most of the remaining massive particles become non-relativistic. These particles' number densities are then exponentially suppressed by the Boltzmann factor $e^{-m/T}$ as long as they remain thermodynamically coupled to the radiation bath via electromagnetic and weak interactions.

As the universe continues to expand, however, the interactions coupling various particles and nuclides to the radiation become too slow compared to the universe's expansion, leading to abundances "freezing out". In particular, weak interactions effectively stop at approximately $T \sim 1$ MeV (about 1 second after the Big Bang), so that the relative abundance of neutrons to protons freezes out. Around $T \sim 0.1$ MeV, these protons and neutrons begin synthesizing into deuterium in appreciable amounts, opening the floodgate for larger nuclides to be built. A large network of coupled Boltzmann equations, one for each possible nuclear reaction between the various nuclides, then determines how nuclide abundances evolve over time in the process of Big Bang Nucleosynthesis. Several minutes after the Big Bang, the rates of the involved nuclear reactions slow compared to the universe's expansion rate, so that the abundances of the various nuclides level off. These steady state values, the primordial abundances, are then the abundances of the various nuclides that persist until matter coalesces into stars some hundreds of millions of years later. We seek

to investigate how the primordial abundances might be influenced by SFDM.

3.1.1 ODEs and Methods

The primordial abundances depend crucially on the time evolution of the scale factor $a(t)$ of the homogeneous and isotropic universe, as the timing of freeze out processes is determined by a comparison between the expansion rate $H(t)$ and the reaction rates. Since SFDM cosmology allows for ρ_ϕ to dominate the total energy density during the $\rho_\phi \propto a^{-6}$ era, it may qualitatively change the evolution of $a(t)$ during nucleosynthesis, so we are led to ask how this alteration to the scale factor can adjust primordial abundance predictions. To answer this question, we modified a pre-existing BBN code, PArthENoPE (Public Algorithm Evaluating the Nucleosynthesis of Primordial Elements), built upon the foundation of earlier codes by Kawano [78] and Wagoner [135]. In the following, we follow the notation and treatment of the works published accompanying the PArthENoPE code [112, 34].

The system of ODEs determining the time evolution of nucleosynthesis with SFDM is comprised of the the Klein-Gordon equation (1.71), the Friedman equations (1.55) and (1.56), the conservation of baryon number, charge neutrality, and the collection of Boltzmann equations for each of the tracked nuclides. Note that equation (1.55) will now only apply to the total energy density and total pressure

$$\rho = \rho_b + \rho_\gamma + \rho_e + \rho_\nu + \rho_\phi, \tag{3.1}$$

$$p = p_b + p_\gamma + p_e + p_\nu + p_\phi, \tag{3.2}$$

as there is exchange of energy between matter sources. Denoting by n_b the baryon number density, $X_i := \frac{n_i}{n_b}$ and Z_i the relative abundance and proton number of the ith nuclide type, and $\varphi_e := \frac{\mu_e}{T}$ with μ_e the (temperature-dependent) electron chemical

potential, the latter three equations are, respectively,

$$\frac{\dot{n}_b}{n_b} = -3H \tag{3.3}$$

$$n_b \sum_j Z_j X_j = n_{e^-} - n_{e^+} = T^3 \hat{L}\left(\frac{m_e}{T}, \varphi_e\right) \tag{3.4}$$

$$\dot{X}_i = \sum_{j,k,l} N_i \left(\Gamma_{kl \to ij} \frac{X_l^{N_l} X_k^{N_k}}{N_l! N_k!} - i, j \leftrightarrow k, l \right). \tag{3.5}$$

Here, the quantity $\Gamma_{kl \to ij}$ is the (temperature-dependent) rate, determinable from measured cross-sections, for the reaction yielding the ith and jth nuclides with the kth and lth nuclides as reactants. The positive integer N_i is the number of the ith nuclide produced in the reaction (similarly for N_j, N_k, and N_l, though the latter two are numbers consumed). Per the Fermi-Dirac distribution, the function $\hat{L}$ appearing in (3.4) describing the number density difference between electrons and positrons is

$$\hat{L}(\xi, \omega) = \frac{1}{\pi^2} \int_\xi^\infty \left(\frac{\zeta \sqrt{\zeta^2 - \xi^2}}{e^{\zeta - \omega} + 1} - \frac{\zeta \sqrt{\zeta^2 - \xi^2}}{e^{\zeta + \omega} + 1} \right) d\zeta.$$

The baryon energy density ρ_b and pressure p_b are computed via n_b, the relative abundances X_i, the nuclide binding energies, and the photon temperature T, while those of electrons, neutrinos, and photons are computed using their momentum space distribution functions appropriate to their being in thermodynamic equilibrium. Hence, if we model a network of reactions among N_{nuc} nuclides, the $5 + N_{\text{nuc}}$ equations in the above system have $5 + N_{\text{nuc}}$ unknowns: $\phi, H, n_b, T, \varphi_e$, and the relative abundances X_i.

We take initial conditions at $T = 10$ MeV, before the weak interaction has shut off, so that all relative abundances (most importantly those of neutrons and protons) are at their equilibrium values. The initial n_b is determined by the baryon-to-photon ratio η_b and the current CMB temperature. The initial H and φ_e are set by equations (1.56) and (3.4), respectively.

The initial values for ϕ and $\dot{\phi}$ are constrained by the requirement that ρ_ϕ evolves to the current observed dark matter density, measured by *Planck* to satisfy $\Omega_{\rm d}h^2 = 0.1200 \pm 0.0012$ [2], but there remains a degree of freedom in the choice of initial $\phi, \dot{\phi}$ that amounts to choosing the temperature T_t at which the ρ_ϕ-dominated era ends, i.e. at which $\rho_\phi = \rho_\gamma + \rho_e + \rho_\nu$. There is, in principle, the additional free parameter of the SFDM mass m, but we expect it to have very little effect on the abundances once the transition temperature T_t and the current-day ρ_ϕ have been set, as these parameters largely determine the temperature dependence of ρ_ϕ. It warrants mentioning that [3] also sought to evaluate the influence of SFDM on BBN (as part of a broader investigation of general scalar fields in cosmology), but it seems they arbitrarily constrained their range of T_t values such that ρ_ϕ was always negligible during BBN.

We modified the PArthENoPE code, which solves a more numerically convenient recasting of the system of equations (1.55), (1.56), and (3.3)-(3.5) detailed in [112], to include the dynamics of a scalar field through equation (1.71) and its contribution of ρ_ϕ, p_ϕ to the energy density and pressure. Our quoted results will use a network of the $N_{\rm nuc} = 9$ lightest nuclides, related via some 40 nuclear reactions. See [128] for a detailed review of the nuclear reactions important to BBN and [113] for a discussion of subtle corrections to the simplest implementations of BBN.

3.1.2 Abundance Results

Our numerical results for the light element abundances in the presence of SFDM are presented above in Figure 3.1. The leftmost column, $T_t = 5$ MeV, is essentially the standard BBN scenario, as the SFDM energy density is effectively negligible compared to radiation by the time nucleosynthesis-relevant processes begin around $T \sim 1$ MeV (when the neutron-to-proton ratio n/p freezes out) in this case. As one decreases the transition temperature, SFDM becomes more and more important to

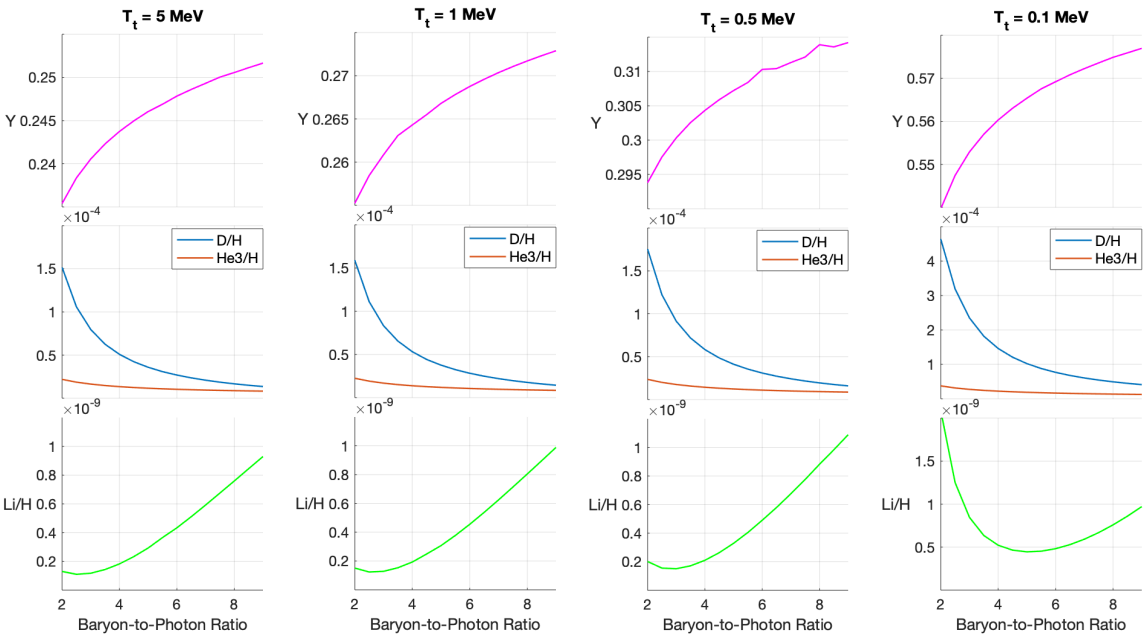

FIGURE 3.1: Predicted light element primordial abundances from BBN with SFDM for various values of the transition temperature T_t, plotted versus $10^{10}\eta_b$. Notice that the scale of the y-axis in the top row (reporting the Helium-to-Hydrogen mass ration Y) changes from column to column, while the scales in the bottom two rows only change in the last column.

the scale factor evolution during nucleosynthesis.

The most notable effect of SFDM's becoming more significant is an increase in the predicted abundances across the board, though this increase is slight for lithium, deuterium, and ^{3}He in the $T_t = 1$ MeV and $T_t = 0.5$ MeV cases. Worse than the fact that this pushes the Lithium prediction further from its measured value, however, even in these cases the perturbation to the energy density already pushes the ^{4}He mass fraction Y_p squarely out of concordance with the empirical constraints reported in equation (1.66) and Table 1.1 (for $\Omega_b h^2 \propto \eta_b$). Even if we relax the latter requirement that η_b remain compatible with CMB anisotropies in ΛCDM, our results indicate that one can no longer simultaneously recover both the ^{4}He and deuterium (1.67) abundances. Indeed, a cursory review of the first and second rows in Figure 3.1 indicates that decreasing T_t requires a smaller η_b to match Y_p and a larger η_b

to match $\mathrm{D/H}|_p$ (with both diverging from the value preferred by the CMB). As both of these are empirically constrained much more confidently than $\mathrm{Li/H}|_p$, this is decidedly undesirable.

In the $T_t = 0.1$ MeV case, SFDM has non-negligible energy density during the nucleosynthesis processes themselves (rather than just the n/p freeze out process), and this makes a somewhat dramatic difference in our results: now all of the predicted abundances are shifted up by an appreciable amount (less so only for ^{3}He). The ^{4}He mass fraction obtained in the CMB η_b range is now more than double the measured value, and the deuterium prediction is also well outside of the range compatible with equation (1.67). Meanwhile, the lithium situation has only gotten worse.

3.1.3 BBN Conclusions

Generally speaking, we've found that the first-order adjustments due to extending the $\rho_\phi \propto a^{-6}$ era of SFDM Cosmology appear to do little to alleviate the lithium problem of BBN, and even slightly more dramatic adjustments seem starkly incompatible with observations. Indeed, allowing ρ_ϕ to be nontrivial during nucleosynthesis primarily serves to increase the matter density and hence the Hubble parameter H by equation (1.56), which should heuristically have the effect that freeze outs occur earlier (at higher temperatures) while ρ_ϕ is large. In particular, if $T_t \lessapprox 1$ MeV this causes n/p to freeze out while the equilibrium value is larger; this increase in available neutrons for nucleosynthesis should increase the production of nuclides (especially ^{4}He, the primary neutron receptacle), tending to shift the predicted ^{7}Li abundance upward away from the measured value. By further pushing down T_t to $T_t \lessapprox 0.1$ MeV, one allows the nontrivial ρ_ϕ to influence the dynamics of the nucleosynthesis reactions themselves, but this both further increases n/p and increases the likelihood that the adjustment destroys the successful predictions of the deuterium and ^{4}He abundances. Our numerical results seem to reinforce these heuristic objections.

These results are subject to the caveat that the PArthENoPE code has the electron, neutrino, and photon energy densities hard-coded as functions of temperature based upon their distribution function dynamics in a radiation-dominated background (this is reasonable in standard BBN because the Baryon energy density is negligible compared to these radiation sources, so their dynamics may be solved independently), so that the neutrino and photon energy densities used are not quite correct for the SFDM scenario. There is good theoretical reason to expect that the error in these quantities is rather low after electron-positron annihilation completes (simply shifting the effective number of neutrino species N_{eff} from 3.046 closer to 3) by around $T \sim 0.1$ MeV, so during nucleosynthesis itself. Further, the adjustment to dynamics is negligible before T_t because these contributions to the energy density are then subdominant. These considerations lead us to expect that the computations put forward here are qualitatively correct, plenty sufficient for the conclusion that SFDM does not seem to provide a viable resolution to the lithium problem. Indeed, our results agree with the comparable work in [32], published after this section's work was completed ahead of my preliminary examination in 2020, as well as the earlier, more qualitative constraint arrived at in [42].

Finally, we emphasize that, irrespective of the magnitude of the above correction, the results presented here do not indicate that SFDM is in tension with abundance measurements– it only indicates that the transition temperature T_t must satisfy $T_t \gtrapprox 5$ MeV. So long as this is the case (as in axion-like models, which effectively take $T_t \to \infty$), SFDM does not change the predictions from standard BBN.

3.2 Perturbative Cosmology

We wish to investigate possible influences of a $\rho_\phi \propto a^{-6}$ era on perturbative cosmology, namely CMB temperature anisotropies. These are apparently deviations from homogeneity and isotropy, so their analysis requires a framework which deviates

from the average FLRW geometry of (1.53). This section reviews this framework, making connections to SFDM as appropriate, and establishes our notational conventions within it. The review of standard material largely follows the excellent and comprehensive lecture notes of Daniel Baumann [11], though we offer a personalized perspective at times.

3.2.1 Perturbation Variables: Metric and Matter

As a starting point, we will find it convenient to treat the averaged background geometry in *conformal time* $\eta(t) := \int_0^t \frac{1}{a(t')} dt'$, with respect to which the metric becomes manifestly conformal to the flat metric:

$$ds^2 = a(\eta)^2 \left[-d\eta^2 + dx_1^2 + dx_2^2 + dx_3^2 \right]. \tag{3.6}$$

Note that $\eta(t)$ is precisely the comoving radius of the *particle horizon* at time t, i.e. the maximum distance light could have travelled between the initial singularity and time t, scaled to the present-day ($a = 1$) notion of distance. η thereby sets the length scale of causality at each point, making it particularly well-suited to cosmology– we will work almost exclusively in terms of η over t in the remainder of this chapter. To avoid ambiguity, derivatives with respect to η will be denoted with a prime, while those with respect to t will be denoted with a dot– these are related by a factor of $\frac{d\eta}{dt} = 1/a$. For example, we will make frequent use of the *conformal Hubble parameter* $\mathcal{H} := a'/a = \dot{a}$. With respect to η, the Friedmann equations become

$$\bar{\rho}' = -3\mathcal{H}(\bar{\rho} + \bar{p}), \tag{3.7}$$

$$\mathcal{H}^2 = \frac{8\pi a^2}{3} \bar{\rho}, \tag{3.8}$$

where we have used bars to indicate that these are the background quantities.

To analyze deviations from the FLRW background, we work within the *Newtonian gauge*, defined (for *scalar* perturbations) by the perturbed metric ansatz

$$ds^2 = a(\eta)^2 \left[-(1 + 2\Psi) d\eta^2 + (1 - 2\Phi) \left(dx_1^2 + dx_2^2 + dx_3^2 \right) \right], \tag{3.9}$$

where the *potentials* Ψ and Φ are functions on M (i.e. functions of both η and $\vec{\mathbf{x}}$) taken to satisfy $\Psi, \Phi \ll 1$. The former is often considered analogous to the Newtonian gravitational potential, and the latter is thought of as a local adjustment to the spatial scale factor. Of course, a general perturbation admits $d\eta \otimes dx^i$ and $dx^i \otimes dx^j$ cross terms, described by *vector* and *tensor* perturbations, but the influence of these on the observables of interest here is expected to be suppressed relative to that of scalar modes, so these are not relevant to our qualitative assessment of the impact of SFDM.

To couple to the perturbed metric (3.9), we must also consider perturbations in the matter, described in the stress-energy tensor T, on top of the average given by (1.54). At first order in these perturbations, we may take T to have the form

$$T^i{}_j = \begin{pmatrix} -\bar{\rho} - \delta\rho & \vec{\mathbf{q}}^T \\ -\vec{\mathbf{q}} & (\bar{P} + \delta P)\mathbf{I}_3 + \mathbf{\Pi} \end{pmatrix}, \tag{3.10}$$

where $\delta\rho$ and δP are the perturbations to the energy density and pressure, $\vec{\mathbf{q}}$ is a 3 dimensional column vector representing the matter's *momentum density*, $\mathbf{I}_3$ is the 3×3 identity matrix, and $\mathbf{\Pi}$ is a traceless and symmetric 3×3 matrix representing *anisotropic stress*. These quantities are emboldened to emphasize that they are not vector or tensor fields on M as defined in Section 1.1, but simply maps $M \to \mathbb{R}^3$ or $M \to \mathbb{R}^{3\times 3}$. For scalar perturbations, we may further take $\vec{\mathbf{q}}$ and $\mathbf{\Pi}$ to be characterized by functions $v, \sigma : M \to \mathbb{R}$ according to

$$\vec{\mathbf{q}} = (\bar{\rho} + \bar{P})\vec{\nabla} v, \tag{3.11}$$

$$\mathbf{\Pi}_{ij} = (\bar{\rho} + \bar{P}) \left(\partial_i \partial_j - \frac{1}{3}\delta_{ij}\mathbf{\Delta} \right) \sigma. \tag{3.12}$$

Here $\vec{\nabla}$ is the usual Euclidean spatial gradient (*not* a spacetime connection) in the spatial coordinates, and $\mathbf{\Delta} = \vec{\nabla} \cdot \vec{\nabla}$ is the Euclidean spatial laplacian. $\vec{\mathbf{v}} = \vec{\nabla} v$ is

the matter's *bulk velocity*. All quantities other than $\bar{\rho}$ and $\bar{P}$ in this description are perturbative, and hence are thought of as small by comparison. We note that it is often useful to work with the relative density perturbation $\delta := \delta\rho/\rho$.

The structure of (3.10), as well as the broad discussion of the previous paragraph, is applicable to each component of the universe's matter content: baryonic matter, dark matter, photons, and neutrinos (dark energy modeled as a cosmological constant maintains its background form (1.54) exactly). As usual, the stress-energy tensors of these components simply add together, but the perturbation variables $\delta\rho_i, \delta P_i, \sigma_i$, and v_i for each component must be evolved according to appropriate equations of motion coupled to the geometry of (3.9). For a matter species modeled as a perfect fluid with equation of state w that is maintained even upon perturbation, one has $\delta P = w\delta\rho$ and $\sigma = 0$, reducing the number of unknowns– unfortunately, this is only tenable for baryons (or standard CDM), with $w = 0$. For photons, neutrinos, and SFDM, further care must be taken.

With both geometry and matter characterized in (3.9) and (3.10), one can compute the constraints imposed by the Einstein equation (1.41). We are interested in the evolution of perturbation variables at first order, so we neglect terms quadratic in perturbations. This standard (if tedious) computation yields [11]

$$8\pi a^2(\bar{\rho} + \bar{P})\sigma = \Phi - \Psi \tag{3.13}$$

$$4\pi a^2(\bar{\rho} + \bar{P})v = -\Phi' - \mathcal{H}\Psi \tag{3.14}$$

$$4\pi a^2\delta\rho = \boldsymbol{\Delta}\Phi - 3\mathcal{H}(\Phi' + \mathcal{H}\Psi) \tag{3.15}$$

$$4\pi a^2\delta P = \Phi'' + \mathcal{H}(\Psi' + 2\Phi') + (2\mathcal{H}' + \mathcal{H}^2)\Psi + \frac{2}{3}\boldsymbol{\Delta}(\Psi - \Phi), \tag{3.16}$$

where each of these holds for the total stress energy tensor. Note this means in (3.13) that $(\bar{\rho} + \bar{P})\sigma = \sum_i(\bar{\rho}_i + \bar{P}_i)\sigma_i$, where the sum is taken over all matter species, and similarly for (3.14).

Given the size of the observable spatial universe and the range of scales of interest,

it is not reasonable to directly numerically integrate these equations as they are to study how perturbative structure grows and propagates. Instead, the evolution of perturbations is more naturally carried out in the Fourier domain by taking the spatial Fourier transform of all perturbative quantities, e.g. writing

$$\Phi(\eta, \vec{\mathbf{x}}) = \int_{\mathbb{R}^3} \frac{d^3k}{(2\pi)^{3/2}} \Phi(\eta, \vec{\mathbf{k}}) e^{i\vec{\mathbf{k}}\cdot\vec{\mathbf{x}}} \tag{3.17}$$

and so on. In an abuse of notation, we generally do not symbolically distinguish between a quantity and its Fourier transform. In an abuse of mathematics, we gloss over technicalities surrounding integrability and existence of the transform. In any event, the linearity of all of our equations in the perturbative quantities means that they may be translated to the Fourier domain upon the minimal substitution $\boldsymbol{\Delta} \rightarrow -k^2$. As all evolution equations will depend only upon k rather than $\vec{\mathbf{k}}$, for numerical integration purposes this allows us to trade the discretization of a 3-dimensional grid of $\vec{\mathbf{x}}$ values for a 1-dimensional grid of k values. Beyond being numerically convenient, this translation also allows us to definitively probe the expected structure of the universe at particular length scales.

A glance back at equations (3.13)-(3.16) indicates that there is frequent occasion to compare the magnitudes of $k = |\vec{\mathbf{k}}|$ and $\mathcal{H}$. While k inversely specifies the perturbative (comoving) length scale under consideration, $\mathcal{H} \sim \frac{1}{\eta}$ similarly encodes the (comoving) size of the particle horizon, so that their comparison physically indicates how the considered length scale measures against the reach of causal propagation. Scales satisfying $k \ll \mathcal{H}$ are said to lie *outside* the horizon, while scales with $k \gg \mathcal{H}$ are *inside* the horizon. These limits often yield qualitatively different dynamics based upon which terms in the equations of motion are dominant– indeed, scales outside the horizon generally exhibit stagnant dynamics, evolving very little. Equation (3.8) indicates how $\mathcal{H}$ varies over time, in particular that it has squarely decreased through-

out most of the history of the universe (up until recently, as dark energy has begun to dominate). This means that many scales initially outside the horizon eventually entered it and began evolving. This description applies to all scales of cosmological interest, and the timing of when a scale entered the horizon is among the most significant features informing how it has evolved.

While equations (3.13)-(3.16) provide a good bit of information, they still must be supplemented by equations of motion for the matter components. For a noninteracting perfect fluid with specified equation of state w and *sound speed* $c_s^2 := \frac{\delta P}{\delta \rho}$, for example, this is furnished entirely by the conservation condition div $T = 0$ on its stress-energy tensor. While this cannot quite be immediately done for baryons due to their tight coupling to photons, it is how canonical ΛCDM treats dark matter, using $w = c_s^2 = 0$. For SFDM, we will generally need to consider the perturbed Klein-Gordon equation.

3.2.2 SFDM Perturbations

A cosmological perturbation of SFDM may be described by decomposing $\phi = \bar{\phi} + \delta\phi$, where $\bar{\phi} = \bar{\phi}(\eta)$ is the background scalar field as determined by its coupling to the average FLRW geometry, discussed in Section 1.3.5, and $\delta\phi$ is a small perturbation which may depend on $\vec{\mathbf{x}}$ as well as η. As before, we will identify ψ according to $\dot{\phi} = m\psi$, or equivalently $\phi' = am\psi$, meaning we also have

$$\delta\psi' = am\delta\phi. \tag{3.18}$$

We first identify the perturbation variables associated to ϕ, requiring that we match the general scalar field stress energy tensor (1.45) to the form (3.10). We reproduce (1.45) for convenience:

$$T = 2\frac{d\phi \otimes d\phi}{m^2} - \left(\frac{|d\phi|^2}{m^2} + \phi^2\right) g. \tag{3.19}$$

At first order, we find

$$\begin{aligned}\frac{|d\phi|^2}{m^2} &= \frac{g^{ij}\partial_i\phi\partial_j\phi}{m^2} \\ &\approx -\frac{(\phi')^2}{a^2m^2(1+2\Psi)} \\ &\approx -\psi^2(1-2\Psi) \\ &\approx -\bar{\psi}^2 - 2\bar{\psi}\delta\psi + 2\bar{\psi}^2\Psi. \end{aligned} \tag{3.20}$$

Note that terms containing $\partial_i\phi\partial_j\phi$ with $i, j > 0$ are second order and hence discarded. This yields

$$\begin{aligned}T^0{}_0 &= 2\phi'\frac{g^{0k}\partial_k\phi}{m^2} - \left(\frac{|d\phi|^2}{m^2} + \phi^2\right) \\ &\approx \frac{|d\phi|^2}{m^2} - \phi^2 \\ &\approx -\bar{\psi}^2 - \bar{\phi}^2 - 2\bar{\psi}\delta\psi - 2\bar{\phi}\delta\phi + 2\bar{\psi}^2\Psi, \end{aligned} \tag{3.21}$$

and for $i, j > 0$,

$$\begin{aligned}T^i{}_j &= 2\frac{g^{ik}\partial_k\phi\partial_j\phi}{m^2} - \left(\frac{|d\phi|^2}{m^2} + \phi^2\right)\delta^i_j \\ &\approx -\left(\frac{|d\phi|^2}{m^2} + \phi^2\right)\delta^i_j \\ &\approx \left[\bar{\psi}^2 - \bar{\phi}^2 + 2\bar{\psi}\delta\psi - 2\bar{\phi}\delta\phi - 2\bar{\psi}^2\Psi\right]\delta^i_j. \end{aligned} \tag{3.22}$$

Lastly, we similarly find (for $i > 0$)

$$T^i{}_0 = 2\phi'\frac{g^{ik}\partial_k\phi}{m^2} \approx 2\frac{\phi'\partial_i\phi}{a^2m^2} \approx \frac{2\bar{\psi}}{am}\partial_i(\delta\phi). \tag{3.23}$$

Putting these together, we identify the perturbation variables as

$$\delta\rho_\phi = 2\left[\bar{\psi}\delta\psi + \bar{\phi}\delta\phi - \bar{\psi}^2\Psi\right] \tag{3.24}$$

$$\delta P_\phi = 2\left[\bar{\psi}\delta\psi - \bar{\phi}\delta\phi - \bar{\psi}^2\Psi\right] \tag{3.25}$$

$$v_\phi = -\frac{\delta\phi}{am\bar{\psi}} = -\frac{\delta\phi}{\bar{\phi}'} \tag{3.26}$$

$$\sigma_\phi = 0. \tag{3.27}$$

These identifications allow us to couple $\delta\phi$ and $\delta\psi$ to the perturbative geometry through (3.13)-(3.16). To also evolve them, it remains to expand the Klein-Gordon equation to first order. Noting that $|g| = a^8(1-2\Phi)^3(1+2\Psi)$ in the Newtonian gauge coordinates, we have

$$\begin{aligned}
\Box\phi &= \frac{1}{\sqrt{|g|}}\partial_i(\sqrt{|g|}g^{ij}\partial_j\phi) \\
&= \frac{-\left(a^2\frac{(1-2\Phi)^{3/2}}{\sqrt{1+2\Psi}}\phi'\right)' + a^2\vec{\nabla}\cdot\left(\sqrt{(1-2\Phi)(1+2\Psi)}\vec{\nabla}\phi\right)}{a^4(1-2\Phi)^{3/2}\sqrt{1+2\Psi}} \\
&\approx -\frac{\phi'' + 2\mathcal{H}\phi'}{a^2(1+2\Psi)} - \left(\frac{(1-2\Phi)^{3/2}}{\sqrt{1+2\Psi}}\right)'\frac{\phi'}{a^2} + \frac{\boldsymbol{\Delta}\phi}{a^2} \\
&\approx \frac{1}{a^2}\left[-(1-2\Psi)(\bar{\phi}'' + 2\mathcal{H}\bar{\phi}') - (\delta\phi'' + 2\mathcal{H}\delta\phi') + (3\Phi' + \Psi')\bar{\phi}' + \boldsymbol{\Delta}(\delta\phi)\right]
\end{aligned} \tag{3.28}$$

Setting up the Klein-Gordon equation $\Box\phi = m^2\phi$ and cancelling the zeroth order terms by the background equation of motion $\bar{\phi}'' + 2\mathcal{H}\bar{\phi}' + a^2m^2\bar{\phi} = 0$ then yields

$$\delta\phi'' + 2\mathcal{H}\delta\phi' + a^2m^2\delta\phi = (3\Phi' + \Psi')\bar{\phi}' - 2a^2m^2\bar{\phi}\Psi + \boldsymbol{\Delta}(\delta\phi). \tag{3.29}$$

Substituting in terms of $\bar{\psi}$ and $\delta\psi$ now gives, in Fourier space,

$$\delta\psi' + 3\mathcal{H}\delta\psi + \frac{a^2m^2 + k^2}{am}\delta\phi = (3\Phi' + \Psi')\bar{\psi} - 2am\bar{\phi}\Psi \tag{3.30}$$

This evolution equation (3.30) (and $\delta\phi' = am\delta\psi$), together with the perturbation variables (3.24)-(3.27), now completely characterizes the evolution of SFDM perturbations and their coupling to the broader evolution.

While the above characterizes SFDM perturbations in principle, a challenge of working with SFDM in cosmology is that the late-time oscillations in $\bar{\phi}$, occurring with frequency m (expected to correspond to timescales on the order of years or shorter, as discussed in Chapter 2), are immensely faster than all other cosmological timescales for the bulk of the universe's history. This makes it impractically computationally expensive to treat the evolution of SFDM exactly indefinitely. In practice, the evolution of SFDM is often averaged over the oscillation period once $H \ll m$, and it is then treated as an effective fluid with equation of state $w = 0$ [64, 95]. Since dark matter should be noninteracting, the conservation condition div $T = 0$ is then sufficient to close the system of perturbation variables for the effective SFDM fluid, provided one also knows its sound speed $c_s^2 = \frac{\delta P}{\delta \rho}$. Utilizing an oscillating ansatz for both ϕ and $\delta\phi$ in the $H \ll m$ regime, one can extract the time-average, scale-dependent sound speed for the effective SFDM fluid. This has been discussed in the literature a number of times [45, 66, 72, 95], resulting in

$$c_s^2 = \frac{\langle \delta P_\phi \rangle}{\langle \delta \rho_\phi \rangle} = \left(1 + \left(\frac{2ma}{k}\right)^2\right)^{-1}. \tag{3.31}$$

With the effective sound speed known, the conservation condition on the effective dark matter stress-energy tensor becomes (we use the subscript d as opposed to ϕ in reference to averaged quantities)

$$\delta_d' + 3\mathcal{H}c_s^2\delta_d = k^2 v_d + 3\Phi', \tag{3.32}$$

$$v_d' + \mathcal{H}v_d = -c_s^2\delta_d - \Psi. \tag{3.33}$$

Equations (3.31)-(3.33) now replace (3.24)-(3.26) and (3.30) to characterize the averaged role of SFDM in the perturbative evolution.

3.2.3 Photons, Neutrinos, and Baryons

We now seek to characterize the evolution of perturbations to the remaining matter components: photons, neutrinos, and baryons. To describe the CMB, we would also like to make contact with a precise notion of the photon temperature. Both of these requires describing the above matter components according to their *distribution functions* $f_i(\eta, \vec{\mathbf{x}}, \vec{\mathbf{p}})$ which returns the the phase-space number density of particles (of the ith type) at the point $(\eta, \vec{\mathbf{x}}) \in M$ with 3-momentum $\vec{\mathbf{p}}$. This means, for example, that the usual number densities are given at each point by

$$n_i(\eta, \vec{\mathbf{x}}) = \int_{\mathbb{R}^3} d^3p \; f_i(\eta, \vec{\mathbf{x}}, \vec{\mathbf{p}}). \tag{3.34}$$

Free-streaming Photons

We drop the subscript i and work with the photon distribution function f. Mathematically, this can be thought of as a function on the null tangent bundle [20], $f : \mathcal{N}M \to \mathbb{R}$, with $f(q, v)$ indicating the relative number of particles with position and tangent vector specified by $(q, v) \in \mathcal{N}M$. As with the usual tangent bundle, metric compatibility ensures that geodesic propagation defines a natural flow on $\mathcal{N}M$, i.e. a map $\mathbb{R} \times \mathcal{N}M \to \mathcal{N}M$ (neglecting completeness concerns) defined by $(s, (q_0, v_0)) \mapsto (\gamma_0(s), \dot{\gamma}_0(s))$, with $\gamma_0 : \mathbb{R} \to M$ the geodesic with initial condition (q_0, v_0). Particles follow this flow in the absence of interactions, and one can show that this gives rise to a variant of Liouville's theorem: the distribution function f is preserved under the flow. Denoting the tangent vector field on $\mathcal{N}M$ associated to this flow by $\frac{d}{ds}$, this can be succinctly written as

$$\frac{df}{ds} = 0. \tag{3.35}$$

Of course, for (3.35) to be true one requires an appropriate identification of the 3-momentum $\vec{\mathbf{p}}$, something not naturally and generally available in general relativity.

Under our simple metric ansatz (3.9), however, a working choice is found by measuring with respect the orthonormal frame obtained by normalizing our coordinate basis vectors. That is, if a photon traverses a null geodesic $\gamma(s)$, then we identify its energy E and 3-momentum $\vec{\mathbf{p}}$ in terms of the coordinate components $(\eta, \vec{\mathbf{x}})$ of γ according to

$$\frac{d\eta}{ds} = \frac{E}{\sqrt{-g_{00}}}, \qquad \frac{d\vec{\mathbf{x}}}{ds} = \frac{\vec{\mathbf{p}}}{\sqrt{g_{ss}}}, \tag{3.36}$$

where $g_{00} = -a^2(1+2\Psi)$ and $g_{ss} = a^2(1-2\Phi)$ in our case. That γ is null requires $E = |\vec{\mathbf{p}}|$ (in a Euclidean sense), so we may write $\vec{\mathbf{p}} = E\hat{\mathbf{p}}$ for a (Euclidean) unit vector $\hat{\mathbf{p}}$. It is convenient to further identify the *comoving energy* $\epsilon := Ea$, which would be a conserved quantity in the background geometry (discussed in Section 1.3.2).

With these identifications, we now consider the distribution function $f(\eta, \vec{\mathbf{x}}, \epsilon, \hat{\mathbf{p}})$ to have inputs conformal time, spatial position, comoving energy, and propagation direction– this information entirely specifies a point on $\mathcal{N}M$. Note that in the homogeneous and isotropic background, the averaged distribution function $\bar{f}$ must be independent of $\vec{\mathbf{x}}$ and $\hat{\mathbf{p}}$. Parameterizing the geodesic flow by conformal time η, we may write (3.35) in terms of the *total derivative* $\frac{df}{d\eta}$ along the flow, schematically obtaining

$$\begin{aligned} 0 &= \frac{df}{d\eta} \\ &= \frac{\partial f}{\partial \eta} + \frac{\partial f}{\partial \vec{\mathbf{x}}} \cdot \frac{d\vec{\mathbf{x}}}{d\eta} + \frac{\partial f}{\partial \epsilon}\frac{d\epsilon}{d\eta} + \frac{\partial f}{\partial \hat{\mathbf{p}}} \cdot \frac{d\hat{\mathbf{p}}}{d\eta} \end{aligned} \tag{3.37}$$

We'll write this at first order in the perturbations. As indicated above, the quantities $\partial f/\partial \vec{\mathbf{x}}$ and $\partial f/\partial \hat{\mathbf{p}}$ are already at least first order, so we may use the zeroth order identities $d\hat{\mathbf{x}}/d\eta = \hat{\mathbf{p}}$ (from (3.36) with $g_{ss}, -g_{00} \to a^2$) and $d\hat{\mathbf{p}}/d\eta = 0$ (since ∂_i for $i > 0$ are killing in the background geometry). Further observing that $d\epsilon/d\eta$ is at

least first order, the above becomes

$$0 = \frac{\partial f}{\partial \eta} + \hat{\mathbf{p}} \cdot \vec{\nabla} f + \frac{\partial \bar{f}}{\partial \epsilon} \frac{d\epsilon}{d\eta}. \tag{3.38}$$

We wish to translate this into an equation for perturbations to the photon temperature. Note that at zeroth order, (3.38) simply says $0 = \partial \bar{f}/\partial \eta$, meaning that $\bar{f}$ depends only on ϵ. Since the photons are very nearly in thermodynamic equilibrium (as evidenced by the nearly perfect blackbody spectrum of the CMB), however, $\bar{f}$ should be given by the Bose-Einstein distribution at a background temperate $\bar{T}$:

$$\bar{f}(\epsilon) \propto \frac{1}{e^{E/\bar{T}} - 1} = \frac{1}{e^{\epsilon/a\bar{T}} - 1}. \tag{3.39}$$

That this is independent of η requires that $\bar{T} \propto 1/a$, as expected of the expanding universe. We now describe the perturbed distribution function f according to this same distribution with a perturbed temperature $T(\eta, \vec{\mathbf{x}}, \hat{\mathbf{p}})$, in particular via a relative temperature perturbation $\Theta(\eta, \vec{\mathbf{x}}, \hat{\mathbf{p}}) := (T - \bar{T})/\bar{T}$ (so that $\Theta \ll 1$):

$$f(\eta, \vec{\mathbf{x}}, \epsilon, \hat{\mathbf{p}}) \propto \left[e^{\epsilon/aT} - 1\right]^{-1} = \left[e^{\epsilon/a\bar{T}(1+\Theta)} - 1\right]^{-1}. \tag{3.40}$$

At first order, this may be written

$$f(\eta, \vec{\mathbf{x}}, \epsilon, \hat{\mathbf{p}}) \propto \left[e^{\epsilon(1-\Theta)/a\bar{T}} - 1\right]^{-1}, \tag{3.41}$$

so we have

$$\begin{aligned} f(\eta, \vec{\mathbf{x}}, \epsilon, \hat{\mathbf{p}}) &\approx \bar{f}((1-\Theta)\epsilon) = \bar{f}(\epsilon - \Theta\epsilon) \\ &\approx \bar{f}(\epsilon) - \frac{d\bar{f}}{d\epsilon} \cdot \Theta\epsilon \\ &= \bar{f}(\epsilon) - \frac{d\bar{f}}{d\ln(\epsilon)}\Theta \end{aligned} \tag{3.42}$$

With this identification, (3.38) becomes

$$0 = \frac{d\bar{f}}{d\ln(\epsilon)}\left[-\frac{\partial\Theta}{\partial\eta} - \hat{\mathbf{p}}\cdot\vec{\nabla}\Theta + \frac{d\ln(\epsilon)}{d\eta}\right] \tag{3.43}$$

By reversing prior reasoning for f, at first order we may combine the first two terms in brackets into the total derivative of Θ along the geodesic flow:

$$0 = \frac{d\bar{f}}{d\ln(\epsilon)}\left[\frac{d\ln(\epsilon)}{d\eta} - \frac{d\Theta}{d\eta}\right]. \tag{3.44}$$

The emergent picture from this discussion, then, is that photon perturbations can be characterized by a temperature field which varies both across points in spacetime $(\eta, \vec{\mathbf{x}})$ and across directions of propagation $\hat{\mathbf{p}}$, and the dynamics of this temperature field are encoded in (3.44) in the absence of scattering. In particular, this equation says that the temperature fluctuations in this scenario are precisely governed by the gravitational redshift of the comoving photon energy ϵ under null geodesic propagation. This redshift can be computed in a straightforward manner via the geodesic equation in our perturbed geometry, whereby one finds (to first order)

$$\frac{d\ln(\epsilon)}{d\eta} \approx \Phi' - \hat{\mathbf{p}}\cdot\vec{\nabla}\Psi. \tag{3.45}$$

The prime here denotes the usual partial η derivative of the potential function $\Phi : M \to \mathbb{R}$, while $\frac{d}{d\eta}$ on the LHS is the total derivative along the geodesic flow.

Thomson Scattering and the Multipoles of Madness

Of course, the objective of CMB analysis is to understand how perturbations were imprinted in the temperature variations at the surface of last scattering during recombination, so the free-streaming equation (3.44) is not entirely sufficient. The primary interaction of interest is Thomson scattering of photons off of electrons (recombination occurs once photon energies are lower than $B_H \sim 13.6$ eV $\ll m_e$, so

the Thomson limit is appropriate), which modify the distribution function f, and hence Θ, by momentum exchange. Denoting the Thomson scattering rate (scattering events per photon per unit conformal time) $\Gamma := a\bar{n}_e\sigma_T$, where $\bar{n}_e$ is the free electron number density and σ_T is the *Thomson cross section*, a known constant, one finds in analyzing this momentum exchange at first order that (3.35) is adjusted to [11]

$$\frac{df}{d\eta} \approx \frac{d\bar{f}}{d\ln(\epsilon)}\Gamma\left[\Theta - \Theta_0 - \hat{\mathbf{p}}\cdot\vec{\mathbf{v}}_b\right]. \tag{3.46}$$

Here $\Theta_0(\eta,\vec{\mathbf{x}}) := \int_{S^2}\frac{d\hat{\mathbf{p}}}{4\pi}\,\Theta(\eta,\vec{\mathbf{x}},\hat{\mathbf{p}})$ is the isotropic *temperature monopole* and $\vec{\mathbf{v}}_b$ is the electron bulk velocity (defined in Section 3.2.1), the same as that of baryons overall due to their tight coupling via electromagnetic interactions. Putting together equations (3.44) and (3.45) with the source term (3.46), one finds the complete first order evolution equation for (scalar) photon temperature perturbations:

$$\frac{d\Theta}{d\eta} = \Phi' - \hat{\mathbf{p}}\cdot\vec{\nabla}\Psi - \Gamma\left[\Theta - \Theta_0 - \hat{\mathbf{p}}\cdot\vec{\mathbf{v}}_b\right] \tag{3.47}$$

Observing that Γ diverges like a^{-2} as $a \to 0$, at early times the *tight coupling limit* enforces that the term in brackets goes to zero, i.e. that $\Theta \to \Theta_0 + \hat{\mathbf{p}}\cdot\vec{\mathbf{v}}_b$. This apparently indicates that, initially, the directional temperature perturbation is entirely due to a doppler shift from the local electron rest frame, so Thomson scattering coupled photons and baryons to have the same bulk velocity. To extract the standard perturbation variables and their evolution from (3.47), it is fruitful to pass to the Fourier domain (so $\Theta(\eta,\vec{\mathbf{x}},\hat{\mathbf{p}}) \to \Theta(\eta,\vec{\mathbf{k}},\hat{\mathbf{p}})$), wherein this becomes

$$\Theta' + ik\mu\Theta = \Phi' - ik\mu\Psi - \Gamma\left[\Theta - \Theta_0 - ik\mu v_b\right], \tag{3.48}$$

where we've set $\mu := \hat{\mathbf{k}}\cdot\hat{\mathbf{p}}$ (and $k = |\vec{\mathbf{k}}|$). We can take this a step further by observing that for each η and $\vec{\mathbf{k}}$, the map on S^2 given by $\hat{\mathbf{p}} \mapsto \Theta(\eta,\vec{\mathbf{k}},\hat{\mathbf{p}})$ may be expanded in spherical harmonics Y_{lm}. Noting that the evolution equation (3.48) for each Fourier

mode of Θ depends only on $\hat{\mathbf{p}}$ through μ, as well as that the initial tight-coupling constraint $\Theta \to \Theta_0 + ik\mu v_b$ also only depends on $\hat{\mathbf{p}}$ through μ, apparently this restriction always holds. The spherical harmonic expansion thus has no azimuthal ($m \neq 0$) terms with respect to the $\hat{\mathbf{k}}$ axis and may be expressed in terms of Legendre polynomials $P_l(\mu)$:

$$\Theta(\eta, \vec{\mathbf{k}}, \hat{\mathbf{p}}) = \sum_{l=0}^{\infty} (-i)^l \Theta_l(\eta, \vec{\mathbf{k}}) P_l(\mu). \tag{3.49}$$

The expansion coefficients $\Theta_l(\eta, \vec{\mathbf{k}})$ comprise the *multipole moments* of the temperature perturbation, extractable via integration against $P_l(\mu)$:

$$\Theta_l(\eta, \vec{\mathbf{k}}) = (2l+1) i^l \int_{S^2} \frac{d\hat{\mathbf{p}}}{4\pi} \Theta(\eta, \vec{\mathbf{k}}, \hat{\mathbf{p}}) P_l(\hat{\mathbf{p}} \cdot \hat{\mathbf{k}}). \tag{3.50}$$

Note the agreement with our prior definition of the monopole Θ_0.

The multipole expansion is particularly pertinent because it quite directly encodes the standard scalar perturbation variables δ, v, and σ. Indeed, identifying for each $(\eta, \vec{\mathbf{x}}) \in M$ and 3-momentum $\vec{\mathbf{p}} \in \mathbb{R}^3$ the associated 4-momentum tangent vector $P \in T_{(\eta,\vec{\mathbf{x}})} M$ (that is, the initial tangent vector to the associated null geodesic, with components (3.36)), one can construct from the distribution function f the total photon stress-energy tensor by simply adding the contributions from each photon:

$$T^{\mu}{}_{\nu} = \int_{\mathbb{R}^3} \frac{d^3 p}{E(\vec{\mathbf{p}})} f(\eta, \vec{\mathbf{x}}, \vec{\mathbf{p}}) P^{\mu} P_{\nu}. \tag{3.51}$$

Recalling the association (3.42) of f to Θ and comparing to the standard form (3.10) of the perturbed stress-energy tensor, one finds [11]

$$\delta_\gamma = 4\Theta_0, \qquad k v_\gamma = -\Theta_1, \qquad k^2 \sigma_\gamma = -\frac{3}{5}\Theta_2. \tag{3.52}$$

This matching also allows one to verify that $\delta P_\gamma = \delta\rho_\gamma/3$, similarly to the background density and pressure.

Finally one can obtain the collection of coupled evolution equations for the multipole moments by integrating (3.48) against $P_l(\mu)$:

$$\Theta_0' + \frac{k}{3}\Theta_1 = \Phi', \tag{3.53}$$

$$\Theta_1' + \frac{2}{5}k\Theta_2 - k(\Theta_0 + \Psi) = -\Gamma\left[\Theta_1 + kv_b\right], \tag{3.54}$$

$$\Theta_l' + k\left[\frac{l+1}{2l+3}\Theta_{l+1} - \frac{l}{2l-2}\Theta_{l-1}\right] = -\Gamma\Theta_l, \tag{3.55}$$

where the final equation holds for $l \geqslant 2$. These three equations together with the identifications (3.52) completely characterize the evolution of photon perturbations and their coupling to the broader dynamics. Since (3.53)-(3.55) comprise an infinite ladder of coupled equations, numerically integrating perturbations in practice requires truncating, i.e. neglecting multipoles beyond some maximum. Compare this limitation to that inherent in the direct utilization of (3.48), sans multipoles, which would require discretizing the S^2 of $\hat{\mathbf{p}}$ values and integrating for each direction in the discretization (with all directions still coupled through the monopole Θ_0). In any event, this truncation is not so much a concern for photons: observe that in the tight coupling limit $\Gamma \gg k$, (3.55) ensures that higher multipoles are suppressed, as it indicates that Θ_l will rapidly decay if ever $|\Theta_l| \gg k|\Theta_{l-1}|/\Gamma$ (for $l \geqslant 2$). This limit generally breaks down right around recombination, as the free electron number density $\bar{n}_e$ sharply drops and scattering becomes inefficient.

Neutrinos and Baryons

The equations of motion for neutrino and baryon perturbations follow quite directly from the work already done for photons. In particular, as neutrinos are highly relativistic for most of the history of the universe (indeed, we approximate them as massless in our numerics), their perturbative dynamics are identical to those of photons, except that they have no coupling to baryons. For many purposes, then,

Neutrinos are to photons as (standard ΛCDM) dark matter is to baryons– they satisfy the same equations sans interactions. Denoting the neutrino temperature perturbation by $\mathcal{N}$, the characterizing equations for neutrinos are therefore precisely equations (3.52) and (3.53)-(3.55) under the substitutions $\gamma \to n$ (in subscripts), $\Theta \to \mathcal{N}$, and $\Gamma \to 0$. That is,

$$\delta_n = 4\mathcal{N}_0, \qquad kv_n = -\mathcal{N}_1, \qquad k^2\sigma_n = -\frac{3}{5}\mathcal{N}_2, \tag{3.56}$$

and

$$\mathcal{N}_0' + \frac{k}{3}\mathcal{N}_1 = \Phi', \tag{3.57}$$

$$\mathcal{N}_1' + \frac{2}{5}k\mathcal{N}_2 - k(\mathcal{N}_0 + \Psi) = 0, \tag{3.58}$$

$$\mathcal{N}_l' + k\left[\frac{l+1}{2l+3}\mathcal{N}_{l+1} - \frac{l}{2l-2}\mathcal{N}_{l-1}\right] = 0, \tag{3.59}$$

Of course, neutrinos are not subject to any tight-coupling limit, and so generally one should include higher multipoles earlier.

While one could approach baryons from scratch via their distribution function as we have photons, it is a good deal simpler to use the results we've accumulated for photons and take advantage of the fact that, since photons and baryons only exchange energy between each other (and not with either neutrinos or dark matter), the combined stress-energy tensor of photons and baryons $T_b + T_\gamma$ is conserved. Setting this combined tensor's divergence to zero, one finds the equations of motion for the baryon perturbation variables,

$$\delta_b' = k^2 v_b + 3\Phi' \tag{3.60}$$

$$v_b' + \mathcal{H}v_b = -\Psi - \frac{\Gamma}{R}\left(\frac{\Theta_1}{k} + v_b\right), \tag{3.61}$$

where $R := \frac{3\bar{\rho}_b}{4\bar{\rho}_\gamma}$ encodes the relative weight of photons versus baryons in their coupled fluid. Supplemented by the perfect fluid conditions that $c_s^2 = w\ (= 0)$ and $\sigma_b = 0$, this

pair of equations completely characterizes the evolution of baryon perturbations and their coupling to the broader dynamics, completing the suite of evolution equations for all perturbations. Comparing these to the effective fluid description of SFDM (after averaging), (3.32) and (3.33), we see that the difference is only that SFDM also has pressure terms (indicated by c_s^2) but no photon coupling Γ term.

3.2.4 Initial Conditions

To numerically evolve the suite of equations discussed in the previous subsections requires explicit conditions by which each perturbation variable is initialized. Initial conditions are set in the superhorizon limit $k \ll \mathcal{H}$ (for all scales of physical interest), well before dynamics begin once $k \sim \mathcal{H}$.

Adiabatic Initial Conditions

We shall take perturbations to be initialized in an *adiabatic* manner at the earliest times, as preferred by empirical constraints [2]. The underlying idea of adiabatic initial conditions is that the universe's variations from the FLRW background are characterized entirely at each point $(t, \vec{\mathbf{x}}) \in M$ by being ahead or behind of the background evolution by a slight shift in comoving time $\delta t(t, \vec{\mathbf{x}})$. This idea is motivated by inflation, wherein the inflaton field decays into standard model particles (thereby initializing the standard big bang evolution) based upon when its value crosses a certain threshold, and quantum mechanical variations in the field lead it to cross this threshold at slightly different times across the spatial universe. As the standard evolution is initialized at different times, different regions are ahead or behind the average. Moreover, the relative delay $\delta t/t$ is precisely identified with the Newtonian potential Ψ. Indeed, if a local comoving observer's proper time τ is shifted from that of the background geometry, $\tau(t) = t + \delta t$, then $d\tau^2 \approx (1 + \delta t/t)^2 dt^2 \approx (1 + 2\delta t/t) dt^2$ to first order, if one assumes $\delta t/t$ is slowly varying. This near-linearity of the shift is

heuristically expected from gravitational time dilation due to the change in density.

More quantitatively, adiabatic initial conditions require that the perturbed energy density ρ_i of the ith matter components is given to first order by

$$\rho_i(t_0, \vec{\mathbf{x}}) = \bar{\rho}_i(t_0 + \delta t) \approx \bar{\rho}_i(t_0) + \dot{\bar{\rho}}_i(t_0)\delta t, \tag{3.62}$$

Equation (1.55) now indicates the relative density perturbation δ_i is

$$\delta_i = \frac{\dot{\bar{\rho}}_i}{\rho_i}\delta t = -3H(1 + w_i)\delta t = -3tH(1 + w_i)\Psi. \tag{3.63}$$

The pairwise ratios δ_i/δ_j between matter components are then particularly simple:

$$\frac{\delta_i}{\delta_j} = \frac{1 + w_i}{1 + w_j}. \tag{3.64}$$

To determine all energy perturbations, then, it suffices to know one. These can all be grounded neatly to Ψ if the background is dominated by a matter component with equation of state $\bar{w}$, as (1.59) then implies $H = \frac{2}{3(1+\bar{w})}\frac{1}{t}$, yielding

$$\delta_i = -2\left(\frac{1 + w_i}{1 + \bar{w}}\right)\Psi. \tag{3.65}$$

In particular, the dominant matter component has relative density perturbation $\delta = -2\Psi$. In light of the Friedmann equation (3.8), this is consistent with (3.15) in the superhorizon limit $k \ll \mathcal{H}$ only if $\Phi' = 0$.

Initial conditions are set long before recombination, so that the tight-coupling limit ensures $v_b = v_\gamma = -\Theta_1/k$ and $\Theta_l = 0$ for $l \geqslant 2$. The factor of $R \propto 1/a$ in (3.61) means that baryons respond to the former constraint much more readily than do photons, so photons' inter-multipole couplings largely drive the dynamics of Θ_1 while baryons simply follow along, at least until around matter-radiation equality. A self-consistent and numerically stable initialization of the neutrino multipoles $\mathcal{N}_l$

is obtained via the ansatz $\mathcal{N}_l = A_l(k\eta)^l$. Since $k\eta \ll 1$ (equivalent to $k \ll \mathcal{H}$), lower multipoles are significantly larger and (3.59) reduces to

$$\begin{aligned} \mathcal{N}_l' &\approx \frac{kl}{2l-2}\mathcal{N}_{l-1} \\ lA_l k^l \eta^{l-1} &= \frac{lA_{l-1}}{2l-2} k^l \eta^{l-1} \\ \Longrightarrow \qquad A_l &= \frac{A_{l-1}}{2l-2} \\ \Longrightarrow \qquad A_l &= \frac{A_1}{2^{l-1}(l-1)!}, \end{aligned} \tag{3.66}$$

for $l \geqslant 2$. Assuming that Ψ is slowly varying, one can complete this procedure by observing that (3.58) then indicates

$$\begin{aligned} \mathcal{N}_1' &\approx k(\mathcal{N}_0 + \Psi) \\ \Longrightarrow \qquad A_1 &= (\mathcal{N}_0 + \Psi) \\ &= \frac{1+3\bar{w}}{3(1+\bar{w})}\Psi, \end{aligned} \tag{3.67}$$

where $\bar{w}$ is again the background equation of state at initialization and we've used (3.56) and (3.65). This last manipulation is applicable to Θ_1 as well, in fact, so that $\Theta_1 \approx \mathcal{N}_1 = A_1 k\eta$.

To both check that the assumption $\Psi \sim$ constant is consistent and anchor Ψ (and hence all of the above quantities) to Φ, we make use of (3.13). Since $\Theta_2 = 0$ due to the tight coupling limit, neutrinos are the only source of anisotropic stress with

$$\sigma_n = -\frac{3}{5}\frac{\mathcal{N}_2}{k^2} = -\frac{3}{5}\frac{A_1}{2}\eta^2 = -\frac{1+3\bar{w}}{10(1+\bar{w})}\eta^2\Psi. \tag{3.68}$$

With the anisotropic stress identified, (3.13) becomes

$$\Psi - \Phi = -\frac{32\pi}{3}a^2\bar{\rho}_n\sigma_n = \frac{16\pi}{15}\left(\frac{1+3\bar{w}}{1+\bar{w}}\right)\frac{\eta^2}{a^2}\Omega_n\rho_{\text{crit}}\Psi, \tag{3.69}$$

where ρ_{crit} is the present-day critical density. Specializing to our case of interest, we consider an early universe which is primarily a combination of $\rho_\phi \propto a^{-6}$ SFDM and radiation with $\rho_r = \rho_\gamma + \rho_n = \frac{\Omega_r \rho_{\text{crit}}}{a^4}$. In this scenario, one can solve (3.7) for η in terms of a. Denoting the scale factor at which the early-universe SFDM to radiation transition occurs as a_C (another means of describing T_t from Section 3.1), one finds

$$\eta = \frac{\sqrt{a^2 + a_C^2} - a_C}{H_0 \sqrt{\Omega_r}}. \tag{3.70}$$

At large scale factors $a \gg a_C$, the universe is radiation dominated and $\eta \to a/H_0\sqrt{\Omega_r}$. At small scale factors, the universe is dominated by SFDM and $\eta \to \frac{a^2}{2a_C H_0 \sqrt{\Omega_r}}$. Using $H_0^2 = 8\pi\rho_{\text{crit}}/3$, The prior result now becomes

$$\Psi - \Phi = \frac{2}{5}\left(\frac{1+3\bar{w}}{1+\bar{w}}\right)\frac{\Omega_n}{\Omega_r}\left(\frac{\sqrt{a^2+a_C^2}-a_C}{a}\right)^2 \Psi \tag{3.71}$$

$$(a \gg a_C, \text{ radiation domination}) \to \frac{3\Omega_n}{5\Omega_r}\Psi \tag{3.72}$$

$$(a \ll a_C, \quad \text{SFDM domination}) \to \frac{\Omega_n}{5\Omega_r}\left(\frac{a}{a_C}\right)^2 \Psi. \tag{3.73}$$

We may solve for Ψ in each case. In the radiation case, Ψ is constant (as desired) and obtained as a multiple of Φ, namely $\Psi = 5\Omega_r\Phi/(5\Omega_r - 3\Omega_n)$. In the SFDM case, the difference $\Psi - \Phi$ is highly suppressed, so that $\Psi \approx \Phi$ is effectively constant. For reference, recall that

$$\frac{\Omega_n}{\Omega_\gamma} = \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\text{eff}} \approx 0.6918 \tag{3.74}$$

in the standard model.

Lastly, we address adiabatically initializing SFDM perturbations. This is done straightforwardly along the same lines as δ_i above:

$$\phi(t_0, \vec{\mathbf{x}}) = \bar{\phi}(t_0 + \delta t) \approx \bar{\phi}(t_0) + \dot{\bar{\phi}}(t_0)\delta t, \tag{3.75}$$

so that

$$\delta\phi = \dot{\bar{\phi}}\delta t = mt\bar{\psi}\Psi. \tag{3.76}$$

Dividing this with (3.63) for δ_b (say) yields

$$\delta\phi = \dot{\bar{\phi}}\delta t = -\frac{m\delta_b}{3H}\bar{\psi} = \frac{2m}{3H}\frac{\bar{\psi}\Psi}{1+\bar{w}}. \tag{3.77}$$

Meanwhile, recall that the perturbation variable $\delta\psi$ was defined to be $\dot{\delta\phi}/m$, and

$$\begin{aligned}
\dot{\delta\phi} = \ddot{\bar{\phi}}\delta t + \dot{\bar{\phi}}\dot{\delta t} &= -(m^2\bar{\phi} + 3Hm\bar{\psi})\delta t + m\bar{\psi}\Psi \\
\implies \qquad \delta\psi &= -(m\bar{\phi} + 3H\bar{\psi})\delta t + \bar{\psi}\Psi \\
&= \left[(\bar{w}-1)\bar{\psi} - \frac{2m}{3H}\bar{\phi}\right]\frac{\Psi}{1+\bar{w}},
\end{aligned} \tag{3.78}$$

where we've used the background Klein-Gordon equation (1.72). Note that substituting these into (3.24) yields

$$\begin{aligned}
\delta\rho_\phi &= 2\left[\bar{\psi}\delta\psi + \bar{\phi}\delta\phi - \bar{\psi}^2\Psi\right] \\
&= 2\bar{\psi}^2\left[\frac{\bar{w}-1}{1+\bar{w}} - 1\right]\Psi \\
&= -2\left(\frac{\bar{\rho}_\phi + \bar{P}_\phi}{1+\bar{w}}\right)\Psi,
\end{aligned} \tag{3.79}$$

consistent with (3.65).

Seed Fluctuations

We finally have but one degree of freedom remaining among all of our perturbation variables: each of $\delta_b, \delta_\gamma, \delta_n, \Theta_l, \mathcal{N}_l, v_b, \delta\phi, \delta\psi$, and Ψ has been related proportionally to Φ. Since our first order analysis is entirely linear in all of these, varying the initial value of Φ simply rescales the entire solution. Practically speaking, then, initializing $\Phi = 1$ for each scale k to be integrated schematically yields a *transfer function* $T(k)$ which multiplicatively maps input fluctuations, as they manifest in Φ as a function

of $\vec{\mathbf{k}}$, to the output perturbation variables. Such transfer functions depend only on k because the same is true of all of our evolution equations.

The initial perturbations are thought to be seeded quantum mechanically via variations in the inflaton field, as briefly discussed above, and the adiabatic imprint left by these is most directly encoded in the *comoving curvature perturbation* $\mathcal{R}$, defined according to

$$\mathcal{R} = -\Phi + \mathcal{H}v. \tag{3.80}$$

Crucially, one can show that this quantity is preserved on superhorizon scales $k \ll \mathcal{H}$ [11], allowing us to the consider the initial imprint in $\mathcal{R}$ solely as a function of $\vec{\mathbf{k}}$, independently of our initial η. Moreover, in this limit $\mathcal{R}$ can be related proportionally to Φ, using (3.14), (3.15), and (3.65), when the background has equation of state $\bar{w}$:

$$\begin{aligned}
\mathcal{R} &= -\Phi - \frac{\mathcal{H}(\Phi' + \mathcal{H}\Psi)}{4\pi a^2 \bar{\rho}(1+\bar{w})} \\
&\approx -\Phi + \frac{\delta\rho}{3\bar{\rho}(1+\bar{w})} \\
&\approx -\Phi - \frac{2}{3(1+\bar{w})}\Psi.
\end{aligned} \tag{3.81}$$

According to this relation, the initial perturbation $\mathcal{R}_0(\vec{\mathbf{k}})$ determines the initial Φ in each Fourier mode, tying the residual inflationary imprint firmly to the entire suite of perturbation variables.

Of course, the probabilistic nature of quantum mechanical fluctuations means that inflation cannot predict precisely what $\mathcal{R}_0(\vec{\mathbf{k}})$ should be– we can only expect to predict its statistics. Assuming that the statistics of $\mathcal{R}_0(\vec{\mathbf{x}})$ are homogeneous and isotropic, in particular that the spatial two-point correlation function $\langle \mathcal{R}_0(\vec{\mathbf{x}})\mathcal{R}_0^*(\vec{\mathbf{x}}')\rangle$ (with angle brackets now indicating a quantum mechanical expectation value) is a function of only $|\vec{\mathbf{x}} - \vec{\mathbf{x}}'|$, then one finds that the analogous expectation quantity in

Fourier space may be written in the form

$$\langle \mathcal{R}_0(\vec{\mathbf{k}})\mathcal{R}_0^*(\vec{\mathbf{k}}')\rangle = \frac{2\pi^2}{k^3}\Delta^2_{\mathcal{R}}(k)\delta(\vec{\mathbf{k}}-\vec{\mathbf{k}}'), \tag{3.82}$$

where $\Delta^2_{\mathcal{R}}(k)$ is the *dimensionless power spectrum* of $\mathcal{R}_0$. A reasonable a priori guess is that this initial power spectrum should be *scale-invariant*, meaning $\Delta^2_{\mathcal{R}}(k) = A_s$ is constant. Inflationary models generally bear out this guess [11], though they tend to favor a slight deviation from scale-invariance, allowed for in ΛCDM by taking the power law ansatz

$$\Delta^2_{\mathcal{R}}(k) = A_s\left(\frac{k}{k_*}\right)^{n_s-1}. \tag{3.83}$$

Here A_s and n_s are, respectively, the *fluctuation amplitude* and *spectral index*, two of ΛCDM's six base cosmological parameters (see Table 1.1). We follow the Planck collaboration in setting the reference scale $k_* = 0.05$ Mpc^{-1} [2].

3.2.5 CMB Temperature Anisotropies

Having fully characterized the initialization and evolution of linear perturbations on top of an FLRW background cosmology, it remains to tie the perturbation variables to the observed temperature anisotropies in the CMB, discussed briefly in Section 1.3.2. Ultimately, we are interested in the multipole expansion of the relative photon temperature deviation $\tilde{\Theta}(\hat{\mathbf{n}}) := \Theta(\eta_0, \vec{\mathbf{x}}_0, -\hat{\mathbf{n}})$ at the present day η_0, at the Earth's location $\vec{\mathbf{x}}_0$, as a function of observation direction $\hat{\mathbf{n}}$ (note the propagation direction $\hat{\mathbf{p}}$ is $-\hat{\mathbf{n}}$). While we carried out a similar expansion in Section 3.2.3, note that we crucially took the Fourier transform first, which picked out a preferred direction $\hat{\mathbf{k}}$ about which we could expect azimuthal symmetry. What's more, given the quantum nature of their ultimate initialization (along with plenty of other complications), we again cannot expect to be able to predict $\tilde{\Theta}(\hat{\mathbf{n}})$ (or its multipoles) directly.

As with the seed perturbations encoded in $\mathcal{R}_0(\vec{\mathbf{k}})$, then, we take an interest in the two-point correlation in $\tilde{\Theta}$, namely the expectation value (taken over the initial quantum mechanical fluctuations) $\langle\tilde{\Theta}(\hat{\mathbf{n}})\tilde{\Theta}(\hat{\mathbf{n}}')\rangle$. Given the statistical homogeneity and isotropy of the seed perturbations, it is expected (and can be verified) that this depends only on the relative angle between $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$, so that one can expand

$$\langle\tilde{\Theta}(\hat{\mathbf{n}})\tilde{\Theta}(\hat{\mathbf{n}}')\rangle = \sum_{l=0}^{\infty} \frac{2l+1}{4\pi} C_l P_l(\hat{\mathbf{n}}\cdot\hat{\mathbf{n}}'). \tag{3.84}$$

Our objective is to compute the coefficients C_l, which comprise the CMB's (temperature) *angular power spectrum*, in terms of the evolved perturbation variables. We first relate this multipole expansion to that performed earlier, (3.49): defining the *temperature transfer functions*

$$\bar{\Theta}_l(k) := \frac{\Theta_l(\eta_0, \vec{\mathbf{k}})}{\mathcal{R}_0(\vec{\mathbf{k}})}, \tag{3.85}$$

one can write

$$\tilde{\Theta}(\hat{\mathbf{n}}) = \int_{\mathbb{R}^3} \frac{d^3k}{(2\pi)^{3/2}} e^{i\vec{\mathbf{k}}\cdot\vec{\mathbf{x}}_0} \Theta(\eta_0, \vec{\mathbf{k}}, -\hat{\mathbf{n}}) \tag{3.86}$$

$$= \int_{\mathbb{R}^3} \frac{d^3k}{(2\pi)^{3/2}} e^{i\vec{\mathbf{k}}\cdot\vec{\mathbf{x}}_0} \sum_{l=0}^{\infty} (-i)^l \Theta_l(\eta_0, \vec{\mathbf{k}}) P_l(-\hat{\mathbf{n}}\cdot\hat{\mathbf{k}}) \tag{3.87}$$

$$= \int_{\mathbb{R}^3} \frac{d^3k}{(2\pi)^{3/2}} e^{i\vec{\mathbf{k}}\cdot\vec{\mathbf{x}}_0} \sum_{l=0}^{\infty} i^l \bar{\Theta}_l(k) \mathcal{R}_0(\vec{\mathbf{k}}) P_l(\hat{\mathbf{n}}\cdot\hat{\mathbf{k}}). \tag{3.88}$$

Substituting this into the expectation value $\langle\tilde{\Theta}(\hat{\mathbf{n}})\tilde{\Theta}(\hat{\mathbf{n}}')\rangle$ and utilizing (3.82) allows one to extract C_l after standard manipulations:

$$\boxed{C_l = \frac{4\pi}{(2l+1)^2} \int_0^{\infty} \frac{dk}{k} \bar{\Theta}_l^2(k) \Delta_{\mathcal{R}}^2(k)} \tag{3.89}$$

The computation of the CMB power spectrum, then, amounts to the computation of the transfer functions $\bar{\Theta}_l$, which are in principle determinable by evolving

the suite of perturbation variables adiabatically initialized with $\Phi = 1$. As Planck measures up to $l = 2500$ (and the Atacama Cosmology Telescope reports sparser data out to $l \sim 4000$ [24]), proceeding in this direct manner out to the multipoles probed by observation is computationally prohibitive. Indeed, this would require the integration of at least some ~ 2500 coupled equations in the hierarchy specified by (3.55), repeated for each Fourier mode needed to reasonably estimate (3.89). As each multipole is coupled to its nearest neighbors, none of these could be omitted.

An alternative approach to computing $\bar{\Theta}_l(k)$ is offered by only explicitly coupling the first few multipoles to the evolution of the perturbation variables and separately computing each $\bar{\Theta}_l$ of interest from the results. To do so requires integration of the photon Boltzmann equation (3.47), reproduced (with slight adjustment) below:

$$\frac{d\Theta}{d\eta} = \Phi' + \Psi' - \frac{d\Psi}{d\eta} - \Gamma\left[\Theta - \Theta_0 - \hat{\mathbf{p}} \cdot \vec{\mathbf{v}}_b\right]. \tag{3.90}$$

Recall that here $\frac{d}{d\eta}$ is the total derivative along the geodesic flow, so this expression lends itself to integrating along the past line of sight of an observer situated at $(\eta_0, \vec{\mathbf{x}}_0)$ and looking in the direction $\hat{\mathbf{n}}$ to characterize $\tilde{\Theta}(\hat{\mathbf{n}}) = \Theta(\eta_0, \vec{\mathbf{x}}_0, -\hat{\mathbf{n}})$. Indeed, it was the splitting of the total derivative of Θ in (3.48) which coupled adjacent multipoles, so we will need to integrate along the flow to avoid this.

It is helpful to introduce the *optical depth* function $\tau(\eta)$

$$\tau(\eta) := \int_{\eta}^{\eta_0} \Gamma(\eta')d\eta', \tag{3.91}$$

which has the physical interpretation that $e^{-\tau}$ is the probability that a photon emitted at conformal time η propagates to the present η_0 without scattering. Further introducing the *visibility function* $g(\eta)$,

$$g := \frac{d}{d\eta}\left[e^{-\tau}\right] = -\tau' e^{-\tau} = \Gamma e^{-\tau} \tag{3.92}$$

(the probability density of last scattering at time η), one may rewrite (3.90) as

$$\frac{d}{d\eta}\left[e^{-\tau}(\Theta+\Psi)\right] = e^{-\tau}(\Phi'+\Psi') + g\left[\Theta_0+\Psi+\hat{\mathbf{p}}\cdot\vec{\mathbf{v}}_b\right]. \tag{3.93}$$

We define the right hand side of this equation to be $S(\eta,\vec{\mathbf{x}},\hat{\mathbf{p}})$ and note that the null geodesic path passing through $(\eta_0,\vec{\mathbf{x}}_0)$ from the $\hat{\mathbf{n}}$ direction is simply (to zeroth order) $\eta \mapsto (\eta,\vec{\mathbf{x}}_0+(\eta_0-\eta)\hat{\mathbf{n}})$. Integrating (3.93) along this path, from the initial singularity to the present, then yields

$$\Theta(\eta_0,\vec{\mathbf{x}}_0,-\hat{\mathbf{n}})+\Psi(\eta_0,\vec{\mathbf{x}}_0) = \int_0^{\eta_0} S\left(\eta',\vec{\mathbf{x}}_0+(\eta_0-\eta')\hat{\mathbf{n}},-\hat{\mathbf{n}}\right)d\eta'. \tag{3.94}$$

Here we've used that $\tau(\eta_0)=0$ and $\tau(0)\to\infty$, due to the divergence of $\Gamma\propto 1/a^2$ in this limit.

Defining $\chi(\eta) := \eta_0-\eta$, writing the above integrand $S\left(\eta',\vec{\mathbf{x}}_0+\chi(\eta')\hat{\mathbf{n}},-\hat{\mathbf{n}}\right)$ in terms of the Fourier transform of S,

$$S\left(\eta',\vec{\mathbf{x}}_0+\chi(\eta')\hat{\mathbf{n}},-\hat{\mathbf{n}}\right) = \int_{\mathbb{R}^3}\frac{d^3k}{(2\pi)^{3/2}}S(\eta',\vec{\mathbf{k}},-\hat{\mathbf{n}})e^{i\vec{\mathbf{k}}\cdot(\vec{\mathbf{x}}_0+\chi(\eta')\hat{\mathbf{n}})}, \tag{3.95}$$

and utilizing the Rayleigh plane wave expansion

$$e^{i\chi\vec{\mathbf{k}}\cdot\hat{\mathbf{n}}} = \sum_{l=0}^{\infty}(-i)^l(2l+1)j_l(k\chi)P_l(\hat{\mathbf{k}}\cdot\hat{\mathbf{n}}) \tag{3.96}$$

(j_l is the lth spherical Bessel function) to manifest the multipoles, one can at last extract $\Theta_l(\eta_0,\vec{\mathbf{k}})$ entirely in terms of other perturbation variables:

$$\boxed{\frac{\Theta_l(\eta_0,\vec{\mathbf{k}})}{(2l+1)} = \int_0^{\eta_0}\left[g(\Theta_0+\Psi)j_l(k\chi)-gkv_bj_l'(k\chi)+e^{-\tau}(\Phi'+\Psi')j_l(k\chi)\right]d\eta} \tag{3.97}$$

The transfer functions $\bar{\Theta}_l(k)$ may now be determined by evolving perturbations, initialized with $\Phi=1$, through η_0 and computing the above integral. This may be

done for any subset of l values desired without any need to evolve the entire hierarchy of multipoles. For intuition, it is worth noting that g is very nearly a delta function and $e^{-\tau}$ very nearly a Heaviside step function, each centered around the time of recombination $z \sim 1100$– see Figure 3.3 below.

3.3 SFDM and Cosmological Perturbations

3.3.1 Programmatic Procedure

To assess the impact of SFDM in the above perturbative framework, we have written a MATLAB CMB code which evolves the system of coupled ODEs including (3.13), (3.16), (3.18), (3.30), (3.60), (3.61), (3.53)-(3.55), and (3.57)-(3.59) for the perturbation variables $\delta_b, \delta_\gamma, \delta_n, \Theta_l, \mathcal{N}_l, v_b, \delta\phi, \delta\psi, \Psi$, and Φ, adiabatically initialized with $\Phi = 1$, for a range of k values spanning $k_{eq}/1000$ to $30k_{eq}$ (see Figure 3.2), where k_{eq} is the scale which crosses the horizon at matter-radiation equality

$$k_{eq} := \mathcal{H}_{eq} = \sqrt{2}H_0\frac{\Omega_m}{\sqrt{\Omega_r}} = \frac{\sqrt{3}\Omega_m H_0^2}{\sqrt{4\pi\Omega_r\rho_{\text{crit}}}} \approx \frac{1}{91\text{ Mpc}}\left(\frac{\Omega_m h^2}{0.15}\right). \tag{3.98}$$

Note that only two of the four conditions from the Einstein equation (3.13)-(3.16) are needed to integrate the system, as the matter equations already enforce the conservation condition div $T = 0$ (indeed, many were derived in this way). Our code then proceeds to numerically compute the integrals (3.97) and (3.89) for some 200 l values, up to $l = 2500$, to obtain the CMB temperature angular power spectrum. Of course, prior to this we evolve the background geometry for the scale factor $a(\eta)$ and scalar field variables $\bar{\phi}$ and $\bar{\psi}$ from the Friedmann equation (3.7) coupled with (1.72), treating each of baryons, photons, and (massless) neutrinos via their constant equations of state. We neglect background transfer of energy between matter and radiation at the earliest scale factors (circa nucleosynthesis and prior, $a \lesssim 10^{-10}$) when we have occasion to integrate in that domain. Our code takes as inputs 5

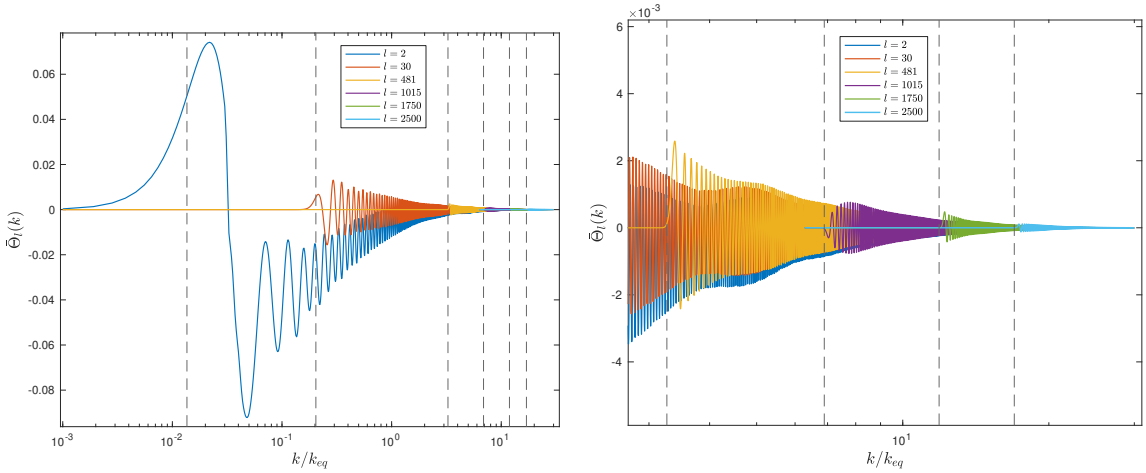

FIGURE 3.2: The temperature transfer functions $\bar{\Theta}_l(k)$ computed according to (3.97) for a variety of l values, demonstrating the sufficiency of our chosen range of k. Dashed lines show the k values expected to begin contributing to each $\bar{\Theta}_l(k)$ based upon the value of $j_l(k\chi)$ at recombination, namely $k = l/\eta_0$. The right figure simply zooms in on the left. Computed with Planck's central cosmological parameters, $m = 10^{-22}$ eV, and $a_C = 10^{-8}$.

of the cosmological parameters– Ω_b, Ω_d, H_0, n_s, and τ_{re} (see Table 1.1)– as well as two parameters associated to SFDM: the mass m and the scale factor a_C at which $\rho_\phi \propto a^{-6}$ overtakes radiation in the early universe. The last cosmological parameter, A_s, is simply an overall multiplicative scale.

The evolution of perturbations requires, through the Thomson scattering rate $\Gamma = a\bar{n}_e\sigma_T$ and associated quantities τ and g, one to know the ionization history of the universe, namely the free electron number density $\bar{n}_e$ as a function of time. Aside from the overall expansion dilution, the two major events informing this quantity's dynamics are recombination at $z \sim 1100$ and reionization at $z_{re} \sim 7.7$ (this is just an alternative quantification of τ_{re}). We implement recombination in a relatively simple manner, equivalent to the original version of the recfast code described by Seager, Sasselov, and Scott [127]. That is, we evolve a simple system of three ODEs for the hydrogen and (first) helium ionization fractions together with the matter temperature– the structure of these ODEs is inspired by effective 3-level atom treatments, but coefficients are fit to match a detailed calculation simultaneously evolving

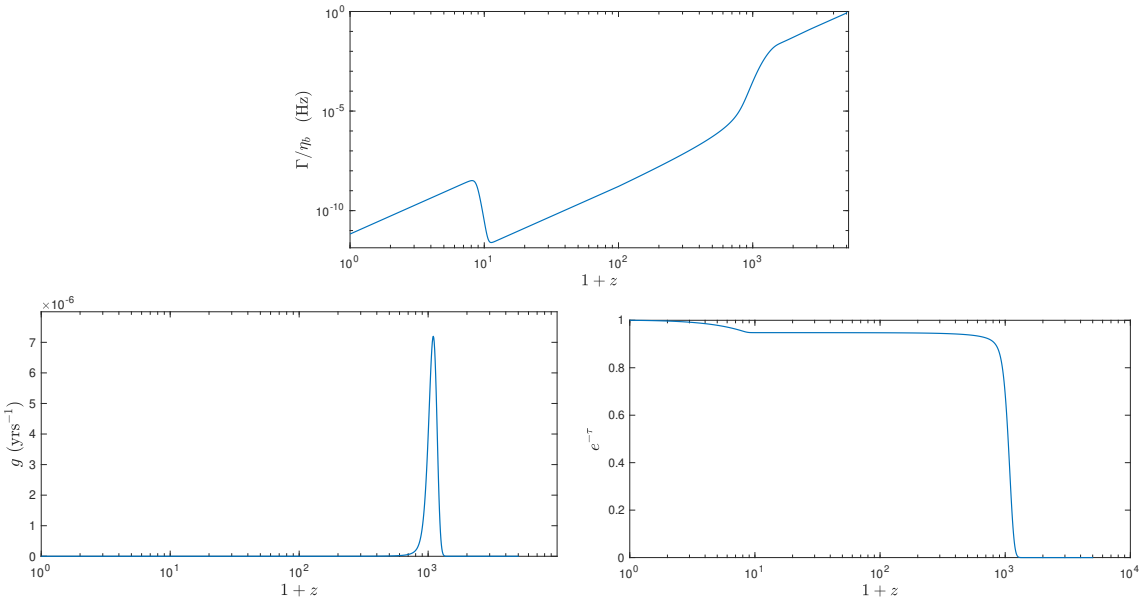

FIGURE 3.3: Ionization history functions Γ, $e^{-\tau}$, and g, against redshift $1+z$. The scattering rate Γ is given in collisions per (comoving) second per baryon– note the sharp adjustments at reionization and recombination. The visibility function g, given in probability density per (comoving) year, is sharply peaked around recombination. The plateau in $e^{-\tau}$ exhibits the optical depth to reionization τ_{re}. Computed for Planck's central cosmological parameters.

the coupled occupations of 300 atomic energy levels. We implement reionization via a standard tanh interpolation centered at z_{re} [2, 90]. See Figure 3.3

As discussed briefly in Section 3.2.2, the oscillations in SFDM beginning once $H \lesssim m$ must be averaged out to render the numerics feasible. Assuming oscillations initiate in the radiation-dominated regime (necessary for SFDM to support structure formation, one of the essential roles of dark matter), they begin at the scale factor

$$a_{\rm osc} := \sqrt{\frac{H_0}{m}}\Omega_r^{1/4} \approx \frac{3.7 \cdot 10^{-7}}{\sqrt{m_{22}}}, \tag{3.99}$$

and we begin averaging in both the background and perturbative evolutions once we have progressed through 15 periods, i.e. up to $t_{15} = 30\pi/m$ (or $\eta_{15} = 2\sqrt{15\pi}/ma_{\rm osc}$). In practice, we also wait until this frequency m is significantly faster than any simultaneous oscillations in the averaged δ_d, discussed in Section 3.3.3 below. As this occurs well before recombination (when the tight coupling limit relaxes) for the pa-

rameter space of interest, we only evolve the photon multipoles Θ_0 and Θ_1 prior to this averaging. Afterwards, we include Θ_2 and Θ_3 before truncating– these provide the most substantive contribution to the larger evolution.

A similar high frequency conundrum arises in a separate, unrelated manner for particularly high values of k. This is most easily seen by considering the hierarchy of neutrino (or photon, after decoupling) multipole equations (3.57)-(3.59) and neglecting the potentials Φ and Ψ. Again truncating beyond $l = 3$, these become

$$\mathcal{N}_0' = -\frac{1}{3}k\mathcal{N}_1 \tag{3.100}$$

$$\mathcal{N}_1' = k\mathcal{N}_0 - \frac{2}{5}k\mathcal{N}_2 \tag{3.101}$$

$$\mathcal{N}_2' = k\mathcal{N}_1 - \frac{3}{7}k\mathcal{N}_3 \tag{3.102}$$

$$\mathcal{N}_3' = \frac{3}{4}k\mathcal{N}_2 \tag{3.103}$$

By triply differentiating (3.100), successively substituting each of (3.101)-(3.103), and then using these again to express the result entirely in terms of $\mathcal{N}_0$, e.g. using $k^2\mathcal{N}_2 = \frac{5}{2}(k^2\mathcal{N}_0 - k\mathcal{N}_1') = \frac{5}{2}(k^2\mathcal{N}_0 + 3\mathcal{N}_0'')$, one obtains an oscillator equation for $\mathcal{N}_0$:

$$3\mathcal{N}_0'''' + \frac{443k^2}{140}\mathcal{N}_0'' + \frac{9k^4}{28}\mathcal{N}_0 = 0. \tag{3.104}$$

Utilizing the ansatz $\mathcal{N}_0 = e^{i\omega k\eta}$, one expects sinusoidal oscillations in $\mathcal{N}_0$ (with higher multipoles necessarily following suit) at frequencies ωk given by roots to the equation

$$3\omega^4 - \frac{443}{140}\omega^2 + \frac{9}{28} = 0. \tag{3.105}$$

The four roots are real, $\omega \approx \pm 0.33745, \pm 0.96999$. The same qualitative procedure carries through regardless of how many multipoles one includes before truncating, though the polynomial in ω becomes higher degree and the precise roots change.

While we have neglected the potentials entirely in the above, it is straightforward to see that precisely the same analysis applies with $\mathcal{N}_0 \to \mathcal{N}_0 + \Psi$ if one takes the potentials to be slowly varying, which is true in the matter-dominated regime once $c_s^2 \to 0$ and SFDM behaves like standard CDM. Even if the potentials are not slowly varying, once k enters the horizon ($k \gg \mathcal{H}$) the results (3.13), (3.15), and (3.16) couple a (damped) oscillator equation for Φ, at similar frequencies, to the multipoles $\mathcal{N}_l$ which drives the system. A similar conclusion holds for the photon multipoles Θ_l, except that multipoles above Θ_1 are suppressed prior to decoupling– this yields $\omega = 1/\sqrt{3}$, though baryons complicate things a bit between matter-radiation equality and recombination. Indeed, it is these oscillations in Θ_0, evaluated at the time of recombination for each k, which imprint in the multipole transfer functions $\bar{\Theta}_l(k)$ through the integral (3.97) (recall g is very nearly a delta function) and determine the angular power spectrum's broad structure.

In any event, multipole oscillations at conformal frequency $\sim k$ definitively emerge once $k\eta \gtrsim 1$, or equivalently once the scale enters the horizon. At k_{eq}, the history of the universe includes some (45 Glyr)/2π(91 Mpc) ~ 24 such periods, and higher values of k accumulate proportionally more. We average out these oscillations once we've evolved through 50 periods, we are well into the matter-dominated regime (rendering these subdominant), and they are much faster than any oscillations in δ_d. In practice, we average over the smaller frequency found above, $\sim 0.337k$. After averaging, we set both Θ_0 and $\mathcal{N}_0$ to their average value $-\Psi$ and $\Theta_l = \mathcal{N}_l = 0$ for higher multipoles. Note that these oscillations occur with a particular conformal frequency, being periodic in η, while the background SFDM oscillations occur at a fixed comoving frequency, being periodic in t, meaning the latter always eventually outpaces the former. See Figure 3.4.

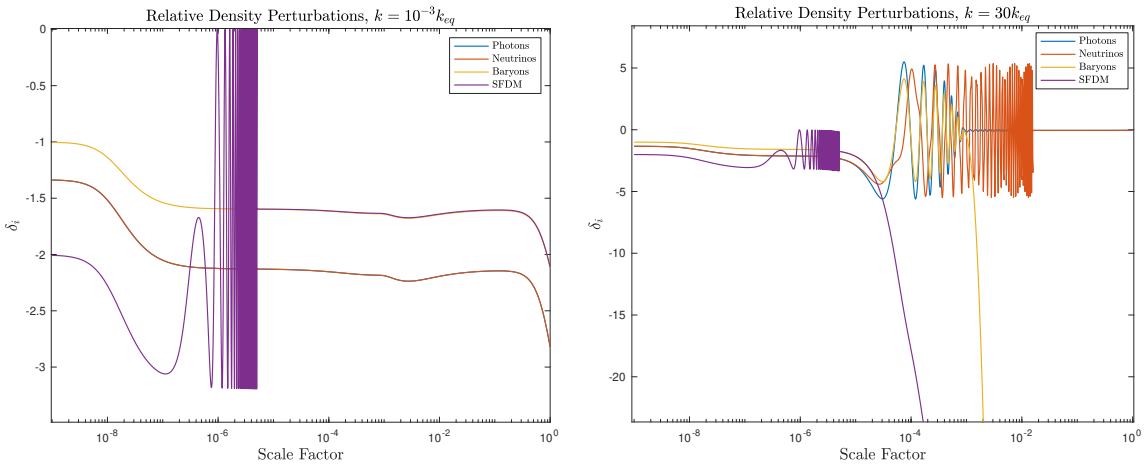

FIGURE 3.4: Evolution of each matter component's relative density perturbation $\delta_i = \delta\rho_i/\bar{\rho}_i$ at the extreme values of k relevant to CMB computations. Computed for central Planck parameters and $m = 10^{-22}$ eV, $a_C = 10^{-8}$. On the left, k remains outside the horizon indefinitely, so few dynamics are present. Observe the oscillatory behavior in SFDM averaging to behave identically to baryons; photons and neutrinos are indistinguishable. On the right, k enters the horizon at roughly $a \sim 7 \cdot 10^{-6}$, and one can see a number of phenomena. SFDM oscillations are averaged early on; its perturbation grows upon entering the horizon. Baryons are delayed by their coupling to photons, with which they undergo damped oscillations until decoupling at recombination. Neutrinos oscillate freely at the combination of frequencies given by (3.105), seen by mild beats; this is eventually averaged out. In both cases, initialization in the ρ_ϕ^{-6} regime has minimal impact.

3.3.2 Mitigating Error Growth

A recurring and inhibiting challenge in this project has been the tendency of errors in the ODEs to grow, apparently according to a power law in the scale factor. This is particularly problematic when solving from very small scale factors so as to accommodate the implementation of a $\rho_\phi \propto a^{-6}$ regime prior to radiation domination. Regardless of our initial precision, numerical error inevitably seeds displacements that can grow to be problematic. Indeed, eventually the error can accumulate enough that it has the effective impact of additional matter, driving structure growth too early and qualitatively changing various results.

Given the apparent "anomalous matter" nature of the error's manifestation, we made use of the energy component (3.15) of the perturbed Einstein equation to quantify its magnitude. Recall that since the slew of perturbative evolution equations

discussed in Section 3.2 are not all independent, we did not need to directly integrate this equation, instead favoring (3.16) to treat the dynamics of Φ. Given that (3.15) is logically implied by the ODEs we did solve, however, its deviation from holding provided a useful measure of numerical error. In particular, setting $L := 4\pi a^2 \delta\rho$ and $R := \boldsymbol{\Delta}\Phi - 3\mathcal{H}(\Phi' + \mathcal{H}\Psi)$, we defined the *relative error measure* $\mathcal{E}$ to be the absolute difference $|R - L|$ divided by the absolute average $|R + L|/2$, so

$$\mathcal{E} := 2\left|\frac{R - L}{R + L}\right|. \tag{3.106}$$

While $\mathcal{E}$ can't be claimed to be the literal solution error (relative difference from the true solution for perturbations), it is nevertheless a useful metric, particularly of whether numerical artifacts have introduced a substantial effective matter source.

To mitigate the growth in $\mathcal{E}$, we made use of the fact that the true evolution of the perturbation variables traverses a level set of the quantity $E(\eta, \vec{\mathbf{z}}) := R - L$, thought of us as a function of η and the vector $\vec{\mathbf{z}}$ of perturbation variables (δ_i, Θ_l,

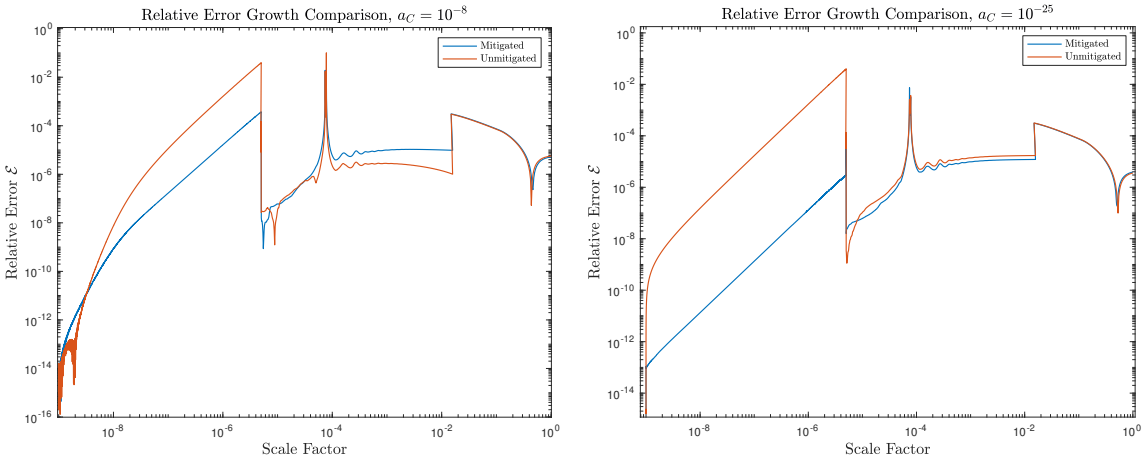

FIGURE 3.5: Evolution of the relative error measure $\mathcal{E}$ under identical conditions when including (blue) and not including (red) the mitigating term in (3.109). Computed for central Planck parameters, $m = 10^{-22}$ eV, and $a_C = 10^{-8}$ (left) and $a_C = 10^{-25}$ (right). In both cases, the standard evolution in red accumulates error above the percent level, while the adjusted evolution in blue remains below about 10^{-4} (excepting the short-lived spike at matter-radiation equality). The two jumps occur when averaging oscillations.

$\mathcal{N}_l$, etc.). In particular, our ODE system

$$\frac{d\vec{\mathbf{z}}}{d\eta} = \vec{\mathbf{D}}(\eta, \vec{\mathbf{z}}) \tag{3.107}$$

defines a flow on the space of $(\eta, \vec{\mathbf{z}})$ along which the total derivative of E is null,

$$0 = \frac{dE}{d\eta} = \frac{\partial E}{\partial \eta} + \vec{\mathbf{D}} \cdot \vec{\nabla} E, \tag{3.108}$$

at least when already on the zero level set of E.

The identity (3.108) is evidently not always true in our numerics, however, whether due to finite numerical precision or to the fact that we had already deviated from the zero level set. To counter this, we force the total derivative of E to be zero by projecting out the unwanted component of $\vec{\mathbf{D}}$, that orthogonal to the level set. That is, instead of setting $d\vec{\mathbf{z}}/d\eta$ according to (3.107), we evaluate each of $\vec{\mathbf{D}}(\eta, \vec{\mathbf{z}})$, $\frac{\partial E}{\partial \eta}$, and $\vec{\nabla} E$ numerically and set

$$\frac{d\vec{\mathbf{z}}}{d\eta} = \vec{\mathbf{D}}(\eta, \vec{\mathbf{z}}) - \left(\frac{\partial E}{\partial \eta} + \vec{\mathbf{D}} \cdot \vec{\nabla} E \right) \frac{\vec{\nabla} E}{|\vec{\nabla} E|^2}. \tag{3.109}$$

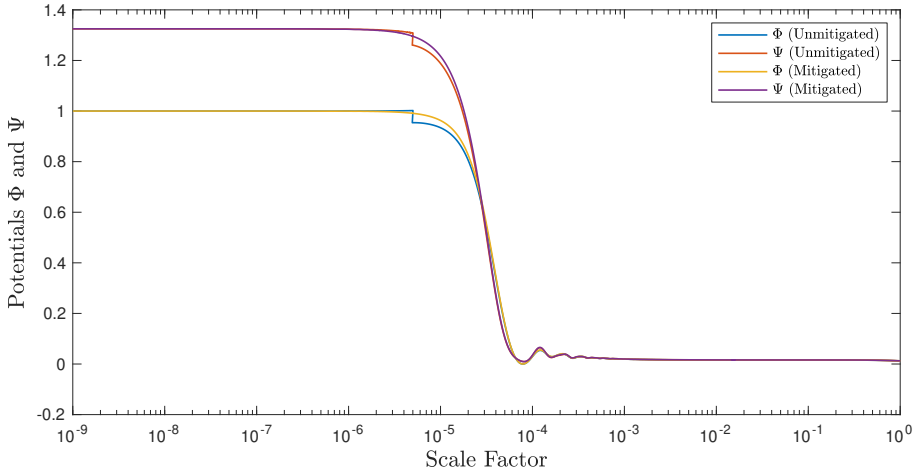

FIGURE 3.6: The Newtonian potentials associated to the $a_C = 10^{-25}$ evolution with relative error shown in Figure 3.5. The percent-level error accumulated leads to the visible discontinuity in the averaging process (blue and red), compared to the smooth transition seen when error reduction is included (yellow and purple).

In practice, we only do this at early times, prior to averaging out SFDM oscillations, as this is when the bulk of error growth occurs. This mitigation technique does not eliminate the growth of $\mathcal{E}$ entirely, though it often sufficiently inhibits it that we have reasonable confidence in the solutions. See Figures 3.5 and 3.6.

3.3.3 SFDM Structure Growth and Suppression

The essential features of SFDM's role in structure growth emerge after oscillations have been averaged out, when dynamics are governed by (3.32) and (3.33), with the sound speed c_s given in (3.31). In the subhorizon limit $k \gg \mathcal{H}$, we may approximate (3.32) as $\delta_d' \approx k^2 v_d$, which one can combine with (3.33) to yield a second-order equation for the relative dark matter density perturbation δ_d:

$$\delta_d'' + \mathcal{H}\delta_d' + k^2(c_s^2\delta_d + \Psi) = 0. \tag{3.110}$$

Further, in the matter-dominated era the anisotropic stress $(\bar{\rho} + \bar{P})\sigma$ is negligible, so that equations (3.13) and (3.15) together yield that $k^2\Psi \approx k^2\Phi \approx -4\pi a^2\delta\rho \approx -4\pi a^2\bar{\rho}_d\delta_d$, with which we may rewrite the above as

$$\delta_d'' + \mathcal{H}\delta_d' + (c_s^2k^2 - 4\pi a^2\bar{\rho}_d)\delta_d = 0. \tag{3.111}$$

This standard equation makes manifest the *implicit Jeans scale* $k_I := \sqrt{4\pi a^2\bar{\rho}_d/c_s^2} \sim \mathcal{H}/c_s$ above which (i.e. for $k \ll k_I$) δ_d grows as it would for standard CDM, which is subject to the same equations with $c_s = 0$. Below the implicit Jeans scale ($k \gg k_I$), however, (3.111) indicates that δ_d oscillates and growth is suppressed– modulo the damping term, this is precisely a wave equation with speed c_s. Note that this is an oscillation in the *averaged* relative SFDM density perturbation, entirely separate from the underlying background oscillations at comoving frequency m; we do not average out the m oscillation until it is much faster than this one (until $am > 10c_sk$). This suppression is taken to mean that gravitational collapse is inhibited by SFDM pressure on small scales, and this is the means by which SFDM functionally suppresses

small-scale structure. The intuitive physical significance of k_I is that it approximately encodes the *sound horizon* $\mathcal{H}/c_s$ of the effective fluid, evidently meaning that perturbations may grow across scales that are causally, but not acoustically, connected: $\mathcal{H} \ll k \ll \mathcal{H}/c_s$ (recall that our analysis has invoked the subhorizon limit).

Of course, as it stands $k_I = k_I(\eta, k)$ depends on both time and the scale k of interest (since c_s depends on k). A more transparent comparator is the *Jeans scale* k_J, a function of time only, defined as that unique scale for which $k_J = k_I(\eta, k_J)$ at each η. k_J has the property that $k > k_J(\eta) \iff k > k_I(\eta, k)$, so that it may be used to more simply determine whether a given scale is oscillating or undergoing growth. The defining condition of k_J is

$$k_J = k_I(\eta, k_J) = \sqrt{4\pi a^2 \bar{\rho}_d \left(1 + \left(\frac{2ma}{k_J}\right)^2\right)}, \tag{3.112}$$

which may be restated as a quadratic in k_J^2. The unique real, positive solution is

$$k_J = \sqrt{4\pi a^2 \bar{\rho}_d \left(1 + \sqrt{1 + 4m^2/\pi\bar{\rho}_d}\right)} \tag{3.113}$$

Noting that $\pi\bar{\rho}_d < \pi\bar{\rho} = 3H^2/8$ and $m/H \gg 1$ since we are operating in the oscillatory regime of ϕ, the above may be approximated as

$$\frac{k_J}{a} \approx \sqrt{8m}(\pi\bar{\rho}_d)^{1/4} \sim \sqrt{mH}. \tag{3.114}$$

Since $k_J \propto a^{1/4}$ is increasing, we see that an inequality $k < k_J$ is conserved, and hence that structure growth persists once it begins at a given scale. Moreover, structure is able to grow on smaller and smaller scales as the universe evolves.

We can understand the evolution of a given scale, with fixed k, as follows. The evolution is essentially stagnant until the scale enters the horizon, at a time such that $k \sim \mathcal{H}$. If this occurs during the matter-dominated regime, the final approximation in

(3.114) holds well (to order of magnitude), and we have $k_J \sim a\sqrt{mH} \gg aH = \mathcal{H} \sim k$, so $k \ll k_J$ and structure growth begins immediately upon horizon crossing, as it would for standard CDM. On any scale which enters the horizon during the matter-dominated regime, then, δ_d evolves as it would for standard CDM. Smaller scales crossing the horizon earlier, during the radiation-dominated era, remain essentially stagnant until matter-radiation equality (with δ_d growing at most like $\ln(a)$ in the meantime, as for standard CDM) at η_{eq}, and subsequent behavior depends on the comparison of k to $k_J|_{eq}$. That is, if $k \gtrsim k_J|_{eq}$, δ_d will oscillate for a time once the matter-dominated era begins, delaying the onset of structure growth at this scale. This implies that $k_J|_{eq} \sim a_{eq}\sqrt{mH_{eq}}$ acts as a cutoff scale below which SFDM structure growth is suppressed relative to standard CDM, as is well-known. See Figure 3.7.

Utilizing the cutoff scale

$$k_J|_{eq} \approx \frac{\sqrt{m_{22}}}{80 \text{ kpc}} \tag{3.115}$$

(note that this estimate is independent of the cosmological parameters), one may readily obtain a lower bound on m by asking that standard CDM structure growth is not impacted on scales for which it agrees with observations. Such considerations are the primary source of constraints on SFDM, and the balance between this and reducing small-scale structure was the original motivation for suggesting $m_{22} \sim 1$. We note that the absolute minimal requirement for the above picture, that growth is largely characterized by the cutoff scale $k_J|_{eq}$, to hold is that matter-radiation equality occurs during the oscillatory regime, i.e. that

$$m \gg H_{eq} = \sqrt{2}H_0\frac{\Omega_m^2}{\Omega_r^{3/2}} \approx (2.5 \cdot 10^{-28} \text{ eV})\left(\frac{\Omega_m h^2}{0.15}\right)^2. \tag{3.116}$$

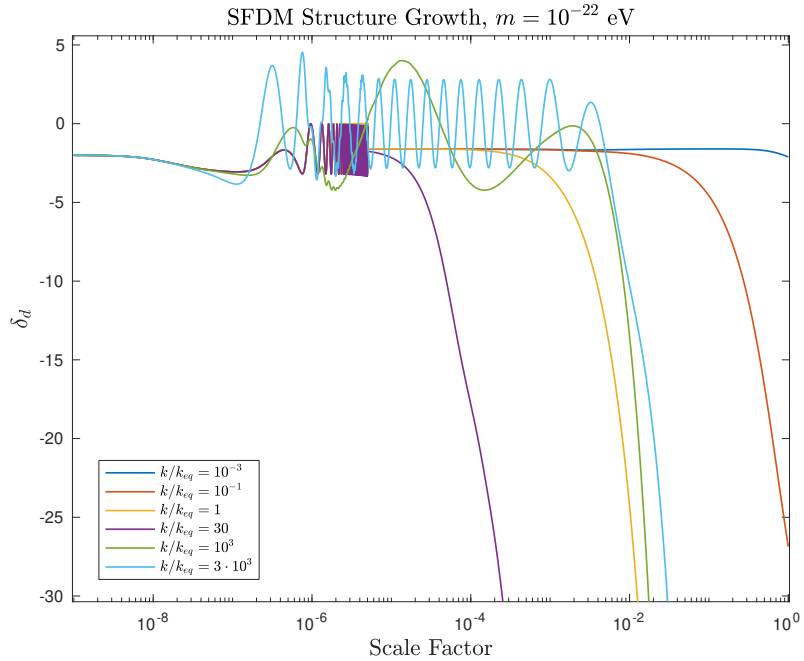

FIGURE 3.7: Evolution of the SFDM relative density perturbation δ_d across a range of scales. Computed for central Planck parameters and $m = 10^{-22}$ eV, $a_C = 10^{-8}$, so that $k_J|_{eq} \approx 10^3 k_{eq}$ (cf. (3.98) and (3.115)). Large scales that do not enter the Horizon do not grow, and as the scale decreases (k increases) we see more growth. This continues until a few times k_{eq}, at which point growth remains essentially fixed for standard CDM. Below the Jeans cutoff, however, we see oscillations in SFDM perturbations which delay and suppress growth. Note that all cases shown here have the underlying frequency m averaged out at the same point.

3.3.4 CMB Results

We show our computation of the CMB temperature angular power spectrum in Figure 3.8, computed with Planck's central cosmological parameters (see Table 1.1), $m = 10^{-22}$ eV, and $a_C = 10^{-25}$. The immensely small a_C effectively means that we've suppressed the $\rho_\phi \propto a^{-6}$ regime entirely (we initialized at $a_i = 10^{-9}$), so this should agree with the standard cosmology in the same manner as axions. We've overlaid Planck's binned power spectrum observations for qualitative comparison. While the agreement is not exact since we have not attempted a state of the art implementation– sources of error include a comparatively simple treatment of recombination and taking massless neutrinos–, we deem it fair enough to reasonably infer qualitative impacts of SFDM.

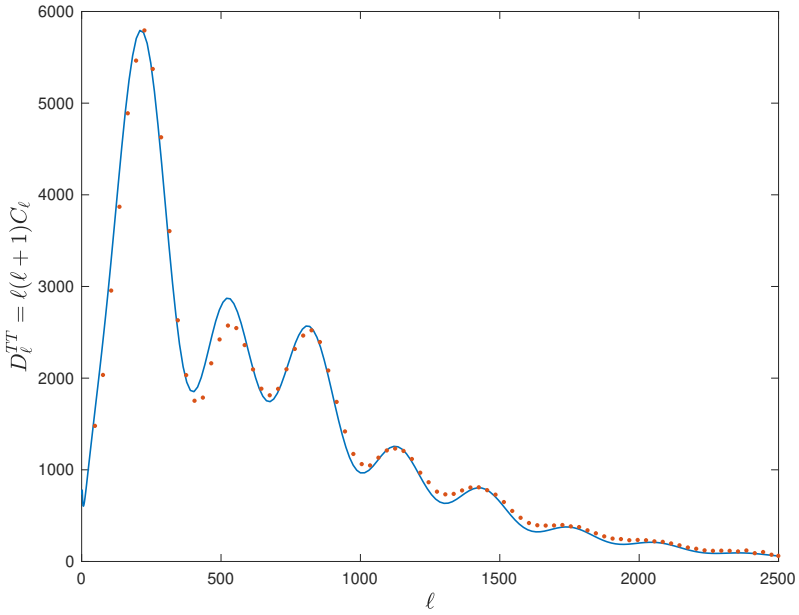

FIGURE 3.8: Our computation of the CMB temperature angular power spectrum (blue), compared with Planck's binned power spectrum observations (red) [2], scaled to have the same maximum. Computed with central Planck parameters, $m = 10^{-22}$ eV, and $a_C = 10^{-25}$. The result with $a_C = 10^{-8}$ (not plotted) is indistinguishable.

In Figure 3.8, we have not plotted the power spectrum with $a_C = 10^{-8}$ because it is virtually indistinguishable from that shown. That is, we've found that adiabatically initializing in the $\rho_\phi \propto a^{-6}$ regime has minimal impact on CMB anisotropies at this mass scale and with a cosmological parameter suite near the current standard in ΛCDM. While this could plausibly cease to be true upon wide variation in the full space of the six cosmological parameters, the result seems robust within constraints set by presently emerging independent, local-universe observations [1]. The apparent primary reason that anisotropies are impartial to the initial background equation of state is the stability of the collection of adiabatic initial conditions discussed in Section 3.2.4 under a phase change. That is, on superhorizon scales we find that adiabatic initial conditions beginning when $\rho_\phi \propto a^{-6}$ remain adiabatic upon the transition to radiation domination, meaning that the evolution then proceeds identically to how it would if the universe had always been radiation dominated. This

concordance is demonstrated in Figure 3.9.

Of course, the qualifier *on superhorizon scales* is critical, as these are generally characterized by a lack of dynamics. A necessary condition for the observed concordance, in the CMB computation and in the potentials in Figure 3.9, is that the scales under consideration do not enter the horizon until well after SFDM-radiation equality at a_C. These scales are essentially set via the condition $1 \lesssim k\eta_0 \lesssim l_{\max}$ by the properties of the spherical bessel functions $j_l(k\chi)$ in (3.97). Noting that one roughly has $1/\eta_0 \propto \sqrt{\Omega_m h^2}$ (with an additional weak dependence on Ω_Λ/Ω_m), we assume the matter content does not vary much from the standard $\Omega_m h^2 \sim 0.15$ and assess the conditions under which these fixed scales, the smallest having $k_{\max} \approx l_{\max}/\eta_0 \approx l_{\max}/(10\text{ Gpc})$ enter the horizon prior to a_C.

The scale crossing the horizon at SFDM-radiation equality is

$$k_{\text{SFeq}} = \mathcal{H}_{\text{SFeq}} = \frac{H_0\sqrt{2\Omega_r}}{a_C} \approx \frac{1}{3.3\text{ kpc}} \cdot \frac{1}{10^8 a_C}, \tag{3.117}$$

independently of any cosmological parameters or the mass parameter m. However, the mass parameter does determine the maximal possible value of a_C, as the $\rho_\phi \propto a^{-6}$

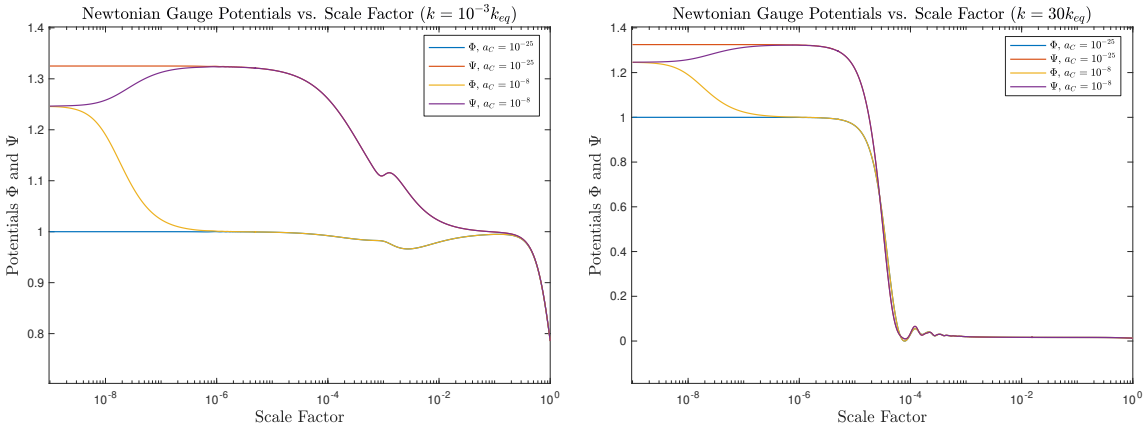

FIGURE 3.9: Evolution of Newtonian gauge potentials initiated in a $\rho_\phi \propto a^{-6}$ regime (yellow and purple) compared to those initiated when radiation dominated (blue and red), at the extreme scales relevant to CMB computations. Computed for central Planck parameters and $m = 10^{-22}$ eV, $a_C = 10^{-8}$ and $a_C = 10^{-25}$. In the former case, the potentials quickly converge to be identical to the latter case once radiation domination begins.

regime must give way to oscillations once $H \lesssim m$, or by the scale factor $a_{\rm osc}$ given in (3.99). The maximal value A_C is determined by assuming this transition is immediate, with no intermediate dark energy behavior of the scalar field. One can then write

$$\bar{\rho}_\phi \approx \frac{\Omega_d \rho_{\rm crit}}{a_{\rm osc}^3} \cdot \left(\frac{a_{\rm osc}}{a}\right)^6, \tag{3.118}$$

and A_C is now given by setting $\bar{\rho}_r = \bar{\rho}_\phi$, giving

$$A_C = \sqrt{\frac{\Omega_d}{\Omega_r}} a_{\rm osc}^{3/2} \approx (1.2 \cdot 10^{-8}) \cdot m_{22}^{-3/4} \cdot \sqrt{\frac{\Omega_d h^2}{0.12}} \tag{3.119}$$

Putting this together with (3.117), the largest possible scale which crosses the horizon during the SFDM-dominated regime with a given mass parameter is therefore $\sim (4 \text{ kpc}) m_{22}^{-3/4}$. At $m_{22} = 1$, then, this would require $l_{\rm max} \gtrsim 10^6$ to be observable in the CMB, or sub-arcsecond resolution at minimum, utterly infeasible at present.

From the other direction, for the current resolution at $l_{\rm max} \sim 2500$ to have observable $\rho_\phi \propto 10^{-6}$ effects would require $m_{22} \lesssim 10^{-4}$, pushing $A_C \sim 10^{-5}$ and encountering fatal constraints surrounding structure growth, as discussed in the previous section. Such small masses already manifest much larger effects in the CMB related to growth suppression that have been ruled out [64, 65]. Note that the limitations discussed here carry over quite directly to modifications to structure growth and the matter power spectrum, adjustments to which would also require that scales of interest enter the horizon during SFDM domination.

Heuristic Proof of Adiabatic Stability

One can see in each of our evolution equations for δ_i, namely (3.53), (3.57), (3.60), that δ_i' is given by terms of order k, which only induce changes on the timescale $\eta \sim 1/k$ of horizon entry, combined with $3(1 + w_i)\Phi'$. Indeed, one can also generally find that the energy component of the conservation condition div $T = 0$ applied to

an individual matter component (applicable if the component is noninteracting, e.g. for SFDM) indicates [11]

$$\delta_i' + 3\mathcal{H}\left(\frac{\delta P_i}{\delta\rho_i} - \frac{\bar{P}_i}{\bar{\rho}_i}\right)\delta_i = \left(1 + \frac{\bar{P}_i}{\bar{\rho}_i}\right)\left(k^2 v_i + 3\Phi'\right). \tag{3.120}$$

If $\delta P_i/\delta\rho_i = w_i$, then, on superhorizon scales one has $\delta_i' \approx 3(1+w_i)\Phi'$.

This condition is met exactly for each of baryons, photons, and neutrinos, and it can be verified for adiabatically initialized SFDM in the relevant regime. Indeed, recalling the form of $\delta\rho_\phi$ and δP_ϕ in (3.24)-(3.25) and the initializations (3.77)-(3.78), we see that

$$\bar{\phi}\delta\phi = \frac{2m}{3H}\frac{\bar{\phi}\bar{\psi}}{1+\bar{w}}\Psi, \qquad \bar{\psi}\delta\psi = \left[\frac{\bar{w}-1}{1+\bar{w}}\bar{\psi}^2 - \frac{2m}{3H}\frac{\bar{\phi}\bar{\psi}}{1+\bar{w}}\right]\Psi. \tag{3.121}$$

If we initialize in the $\rho_\phi \propto a^{-6}$ regime (characterized by $w_\phi \to 1$), then, the term $\bar{\phi}\delta\phi$ is suppressed relative to the others in $\delta\rho_\phi$ and δP_ϕ, namely $\bar{\psi}\delta\psi$ and $\bar{\psi}^2\Psi$, by factors of $\frac{\bar{\phi}}{\bar{\psi}}$ and $\frac{m}{H}$, both of which are much smaller than 1. This suppression leads to $\delta\rho_\phi \approx \delta P_\phi$, so that $\delta P_\phi/\delta\rho_\phi \approx 1 \approx w_\phi$, as desired.

We therefore have $\delta_i' \approx 3(1+w_i)\Phi'$ on superhorizon scales for all matter components, so that all components maintain their relative contributions according to (3.64). Moreover, the change $\Delta\delta_i$ in δ_i on such scales through a phase change in the background is given by $3(1+w_i)\Delta\Phi$. But the conservation of $\mathcal{R}$ on superhorizon scales means, according to (3.81), that $\Delta\Phi = -\frac{2}{3}\Delta\left[\frac{\Psi}{1+\bar{w}}\right]$, so we find

$$\Delta\delta_i = \Delta\left[-2\left(\frac{1+w_i}{1+\bar{w}}\right)\Psi\right]. \tag{3.122}$$

That is, the adiabatic condition (3.65) is preserved through the background phase change. This condition propagates the appropriate corresponding adjustments through the Θ_l and $\mathcal{N}_l$ multipole hierarchies as discussed in Section 3.2.4. Such a phase

change thus has no impact on the ultimate evolution of perturbations and any late-universe observables, provided that it occurs before all relevant scales enter the horizon. We have seen here that this follows if the additional matter component satisfies $\delta P_i/\delta\rho_i = w_i$. This identity need not be satisfied in general– indeed, that it is *not* satisfied in the effective SFDM fluid at later times is why SFDM suppresses small scale structure–, but that it is for adiabatically initialized SFDM in the $\rho_\phi \propto a^{-6}$ regime seems to be the underlying reason that this regime had no detectable impact.

3.3.5 Conclusions

Both numerical and analytical results presented here have indicated that the adiabatic initialization of cosmological perturbations are sufficiently stable on superhorizon scales that beginning in an SFDM-dominated $\rho_\phi \propto a^{-6}$ regime has essentially no impact on the ultimate evolution of perturbations. As all scales remotely relevant to presently measurable multipoles remain superhorizon well beyond SFDM-radiation equality for the mass range admissible by structure growth constraints, we've thereby found that any impacts on the CMB from an initial SFDM-dominated phase are squarely outside of experimental reach. This both means that such a phase is entirely admissible by the observational constraints considered herein, as well as that they exhibit very little promise for addressing any outstanding tensions in cosmology, e.g. the Hubble tension.

Even so, it is of academic interest to investigate what qualitative impacts may occur in these inaccessible regimes, and a wide range of the parameter space has not yet been thoroughly investigated. Indeed, we have primarily been limited to working near the standard cosmology for standard values of m. It remains a remote possibility that more dramatic adjustments, either in m and a_C or the cosmological parameters, can reduce initial deviations from admissibility (i.e. partially reverse over-suppression of small-scale structure at small m), though a full recovery to concordance seems

especially unlikely given the rather decisive constraints on a_C imposed by BBN, discussed in Section 3.1. Presently limiting factors to further exploration include that smaller scales are both more numerically expensive and require earlier initializations, exacerbating concerns of error growth discussed in Section 3.3.2.

4

Naked Singularities in Vaidya Spacetimes

In this chapter and the next, we consider a wholly different challenge than dark matter and cosmology. Though our gears shift, we remain affixed on the objective of investigating the fundamental structure and content of the theory of general relativity. This chapter presents the content of my work in [142].

Indeed, among the largest outstanding problems in theoretical general relativity are the Cosmic Censorship Conjectures, which essentially posit that the theory is well-posed despite the effectively assured phenomenon of singularities [58, 102, 104, 123]. These are a pair of conjectures, termed *weak* and *strong* (somewhat unfortunately, as technical formulations are independent of each other– see, for example, Chapter 12 of Wald [137]), both originally due to Penrose [105, 106]. The weak version's heuristic content is that *dynamical singularities in general relativity are generically not visible to observers at infinity*, while the strong version's heuristic content is that *dynamical singularities in general relativity are generically not visible to any observer*. Singularities in violation of the weak version are dubbed *globally naked*, while those in violation of the strong version are dubbed *locally naked*.

While challenges in arriving at general proofs of comprehensive and compelling

formulations of these conjectures abound, among the largest is nailing down precisely what *generic* is to mean in this context. Some such caveat is certainly required, as several examples of spacetimes containing naked singularities have been constructed [26, 27, 75, 84, 116]. Arriving at the appropriate notion of genericness will doubtless require a thorough understanding of the extant examples. Towards this effort, we demonstrate here that one class of examples, the incoming Vaidya spacetimes, feature globally naked singularities which are much more prolific than previously stressed in the literature.

It is well known that the incoming Vaidya spacetimes, perhaps the simplest possible models for the dynamical formation of a black hole, can also exhibit the dynamical formation of naked singularities [12, 54, 75, 84, 86], i.e. naked singularities to the future of a complete spacelike hypersurface (a nonsingular "instant" in time). While the occurrence of the locally naked case has been sufficiently characterized in these sources, it seems the globally naked case has not been thoroughly explored outside of the highly restricted self-similar subclass and a few related examples. We treat both cases, as well as the singularity's curvature strength, in a uniform manner with simple ODE comparison techniques (distinguished from the ansatz approach of, say, Kuroda [84], and the node analysis of Joshi and Dwivedi [75]).

4.1 Vaidya Spacetimes

We consider the spacetime $M = (\mathbb{R} \times \mathbb{R}^3) \backslash \{(v, 0, 0, 0) \,|\, v \geqslant 0\}$ with metric given by

$$g = -\left(1 - \frac{2m(v)}{r}\right) dv^2 + dv \otimes dr + dr \otimes dv + r^2 d\sigma^2, \tag{4.1}$$

where $d\sigma^2$ is the metric on the unit sphere $S^2 \subset \mathbb{R}^3$, r is the radial coordinate on $\mathbb{R}^3$, and $m : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ is nondecreasing, continuous, piecewise C^1, and satisfies $m(v) = 0$ for $v \leqslant 0$, $m(v) > 0$ for $v > 0$. That is, the mass parameter $m(v)$ (this is, e.g.,

the Hawking mass of spheres of constant r and v) increases continuously from 0 beginning at $v = 0$, with continuous derivative at all but finitely many points, at which left- and right-handed limits of $m'(v)$ exist and are equal to the left- and right-handed derivatives $m'_{\pm}(v)$. The Einstein tensor has a single nonzero component in these coordinates– it can be computed to be

$$G = \frac{2m'(v)}{r^2} dv^2, \tag{4.2}$$

from which we easily deduce that the Dominant Energy Condition (DEC) is satisfied since $m'(v) \geqslant 0$. Strictly speaking, this observation is only entirely unambiguous when $m(v)$ is globally C^2, but the results which follow hold for the piecewise C^1 case.

For $v \leqslant 0$, the metric (4.1) is precisely the Minkowski metric in ingoing null-radial coordinates. If $m(v)$ levels off to some fixed value m_0 at some $v_0 > 0$, i.e. if $m(v) = m_0$ for $v \geqslant v_0$, then in this region (4.1) is precisely the Schwarzschild metric of mass m_0 in ingoing Eddington-Finkelstein coordinates, so this is the sense in which (M, g) is a simple model of the formation of a Schwarzschild black hole due to matter falling into the origin along the null geodesics of constant $0 < v < v_0$. While this is a useful image to keep in mind, the analysis will not require $m(v)$ to level off (so there need not be an exactly Schwarzschild region). See Figure 4.1 for a schematic Penrose diagram of two contrasting cases. There and in what follows, we suppress the S^2 coordinates.

4.2 Conditions for Naked Singularities

A subclass of examples which has received much attention is the self-similar case [46, 47], that of taking $m(v) = \frac{m_0}{v_0} v$ for $v \in [0, v_0]$ and $m(v) = m_0$ for $v \geqslant v_0$, and it is well understood that in this scenario a globally naked singularity develops at $(v, r) = (0, 0)$ provided that $\frac{m_0}{v_0} \leqslant \frac{1}{16}$. In fact, it is true that a globally naked singularity develops

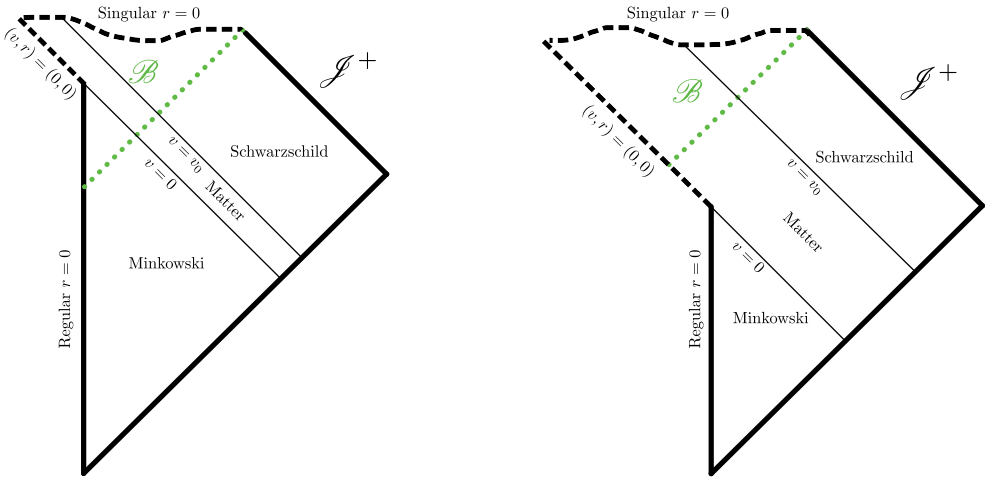

FIGURE 4.1: Schematic Penrose diagrams for two Vaidya spacetimes of interest, with smooth mass functions which grow at different rates and level off. The Minkowski, Schwarzschild, and matter-containing regions are labelled in each case, separated by the lines of $v = 0$ and $v = v_0$. The dashed lines mark the singularities, while the dotted line marks the event horizon, above which is the black hole region $\mathscr{B}$. When the mass function increases more slowly (on the right), the $(v, r) = (0, 0)$ singularity, which is stretched into a diagonal line in null-null coordinates, becomes globally naked. See Propositions 2 and 4.

under the significantly more general constraint that $\sup_{v>0} \frac{m(v)}{v} < \frac{1}{16}$, regardless of the form of m. To show this, we simply investigate the ODE for outgoing radial null curves (which will necessarily be pre-geodesics): parameterizing a curve γ by v in (v, r)-coordinates, i.e. taking $\gamma(v) = (v, r(v))$, we find that $\gamma'(v) = \partial_v + r'(v)\partial_r$ is null if and only if

$$r'(v) = \frac{1}{2} - \frac{m(v)}{r(v)}. \tag{4.3}$$

We would like to identify solutions to this ODE which terminate to the past at $r = 0$ as $v \to 0$ (from above). Observe that the Picard-Lindelöf theorem ensures existence and uniqueness on $(v, r) \in \mathbb{R} \times \mathbb{R}_+$, and maximal solutions either exist for all $v \in \mathbb{R}$ or have $r \to 0$ at finite v.

Consider the curve $v \mapsto (v, \alpha(v))$ with $\alpha(v) := 2m(v)$ (this is the apparent hori-

zon), and define $\Delta_2 := r(v) - \alpha(v)$ for a solution $r(v)$ to the ODE (4.3). Then we have

$$\Delta_2'(v) = r'(v) - 2m'(v) \leqslant r'(v) = \frac{1}{2} - \frac{m(v)}{r(v)} = \frac{\Delta_2(v)}{2r(v)},$$

so Δ_2 will remain negative if it is ever negative. This is the familiar statement that the apparent horizon "traps" the outgoing null curves. Hence if $r(v_0) \geqslant 2m(v_0) = \alpha(v_0)$ at some $v_0 > 0$, we must have had $r(v) \geqslant \alpha(v)$ for $v < v_0$.

We investigate when we can obtain similar comparisons via the curves associated to $\beta_k(v) := km(v)$ with $k > 0$. Define $\Delta_k(v) := r(v) - \beta_k(v)$ and consider

$$\begin{aligned}
\Delta_k'(v) = r'(v) - km'(v) &= \frac{1}{2} - \frac{m(v)}{r(v)} - km'(v) \\
&= \left[\frac{1}{2} - \frac{1}{k} - km'(v)\right] + \frac{1}{k} - \frac{m(v)}{r(v)} \\
&= k\left[\frac{1}{2k} - \frac{1}{k^2} - m'(v)\right] + \frac{\Delta_k(v)}{kr(v)},
\end{aligned} \tag{4.4}$$

so $\Delta_k(v)$ will have a conserved sign provided that the term in brackets has fixed sign. This will be the key tool used in the following results. Defining $f : \mathbb{R}_+ \to \mathbb{R}$ by

$$f(k) := \frac{1}{2k} - \frac{1}{k^2},$$

we see that the significance of the value $\frac{1}{16}$ is that it is the maximum of f, achieved at $k = 4$, so controlling the comparison of $m'(v)$ to $\frac{1}{16}$ allows us to control the sign of Δ_k.

We first treat the locally naked case.

4.2.1 Locally Naked Singularities

Proposition 2. *(Locally Naked) Consider the incoming Vaidya spacetime as above characterized by a nondecreasing mass function $m(v)$ which is continuous and piecewise C^1.*

(i) If there exists a $v_0 > 0$ such that $m'(v) \leqslant \frac{1}{16}$ on $(0, v_0]$, then there exists a one-parameter family (modulo S^2) of outgoing radial null geodesics which terminate to the past at $(v, r) \to (0, 0)$.

(ii) If there exists a $v_0 > 0$ and $\varepsilon > 0$ such that $m'(v) \geqslant \frac{1}{16} + \varepsilon$ on $(0, v_0]$, then no causal curve terminates to the past at $(v, r) \to (0, 0)$.

Note the hypothesis for *(i)* is ensured if $m'_+(0) < \frac{1}{16}$, while the hypothesis for *(ii)* is equivalent to $m'_+(0) > \frac{1}{16}$. It is already known by other techniques that the singularity is locally naked if and only if $m'_+(0) \leqslant \frac{1}{16}$ [12, 75], but we include this statement and proof for use in and comparison to the globally naked result, and because the proof method makes trivial an observation regarding the curvature strength of the singularity, leading us to the next result. In the globally C^2 case, of course, $m'_+(0) = m'_-(0) = 0$, yielding the interesting observation that *smoothness favors a locally naked singularity.*

Proof. (i): Suppose $m'(v) \leqslant \frac{1}{16}$ on $(0, v_0]$. By the intermediate value theorem, we may find a $k_+ \in [4, \infty)$ such that $f(k_+) = \sup_{0<v\leqslant v_0} m'(v)$. Consider a solution $r(v)$ to the ODE (4.3) with initial value $0 < r(v_0) \leqslant k_+ m(v_0)$. By (4.4) and the hypothesis,

$$\Delta'_{k_+}(v) \geqslant \frac{\Delta_{k_+}(v)}{k_+ r(v)}$$

on $(0, v_0]$, so Δ_{k_+} will remain positive on this domain if it is ever positive. Since $\Delta_{k_+}(v_0)$ is nonpositive by the initial condition, it must therefore be nonpositive for $0 < v < v_0$.

Since $m(v_0) > 0$, we may find a $0 < k_0 \leqslant 2$ such that $r(v_0) > k_0 m(v_0) = \beta_{k_0}(v_0)$, and (4.4) yields

$$\Delta'_{k_0}(v) \leqslant \frac{\Delta_{k_0}(v)}{k_0 r(v)},$$

so Δ_{k_0} remains negative if it is ever negative. Hence Δ_{k_0} is nonnegative for $0 < v < v_0$. Putting these comparisons together, we've found

$$k_0 m(v) \leqslant r(v) \leqslant k_+ m(v) \tag{4.5}$$

for $0 < v < v_0$, which ensures $\lim_{v \to 0^+} r(v) = 0$, as desired.

(ii): As in the proof of *(i)*, given any solution $r(v)$ to (4.3), regardless of initial condition, we may find a $0 < k_0 \leqslant 2$ such that $\Delta_{k_0} \geqslant 0$ for $v > 0$ sufficiently small, which ensures that $(v, r(v))$ cannot terminate to the past at $(v, 0)$ for any $v > 0$. Since solutions also cannot terminate to the future at $(v, 0)$ for any $v \leqslant 0$, this further ensures that any maximal solution to (4.3) has a value at some point in $(0, v_0]$, so initializing in this range of v treats all outgoing radial null geodesics.

Now, suppose $m'(v) \geqslant \frac{1}{16} + \varepsilon$ on $(0, v_0]$. We first note that since (4.3) immediately yields that $r'(v) < \frac{1}{2}$, any solution which terminates to the past at $(0, 0)$ must always satisfy the crude bound $r(v) < \frac{v}{2}$. We show that every solution somewhere violates this bound under our hypothesis. In particular, the hypothesis indicates that $m(v) > \frac{v}{16}$ on $(0, v_0]$, so it suffices to show that every solution satisfies $r(v) \geqslant 8m(v)$ for some $v \in (0, v_0]$.

Consider a solution $r(v)$ satisfying $r(v_*) < 8m(v_*)$ at some $v_* \in (0, v_0]$. We define the quantity $k(v)$ by $r(v) = k(v)m(v)$, so the initial condition stipulates $0 < k(v_*) < 8$, and we wish to show $k(v) \geqslant 8$ at some $v \in (0, v_0]$. Differentiating this definition yields

$$\frac{1}{2} - \frac{1}{k(v)} = r'(v) = k(v)m'(v) + k'(v)m(v),$$

which implies $k(v)$ satisfies

$$\begin{aligned} k'(v) &= \frac{f(k(v)) - m'(v)}{m(v)} k(v) \\ &\leqslant -\frac{\varepsilon}{m(v)} k(v) \end{aligned} \tag{4.6}$$

on $(0, v_0]$. The piecewise C^1 condition ensures we may also bound $m(v) \leqslant Cv$ on $(0, v_0]$, for some $C > 0$. Combining this with the above inequality gives

$$k'(v) \leqslant -\frac{\varepsilon}{C}\frac{k(v)}{v}$$

on $(0, v_0]$, and an application of Gronwall's inequality now yields

$$k(v) \geqslant k(v_*)\left(\frac{v_*}{v}\right)^{\frac{\varepsilon}{C}}$$

for $v \in (0, v_*]$. This implies $k(v) \to \infty$ as $v \to 0$, and in particular that there is a $v \in (0, v_*]$ at which $k(v) \geqslant 8$. This demonstrates that there are no outward radial null geodesics which terminate to the past at $(0, 0)$.

If there were a radial causal curve $\gamma(s) := (v(s), r(s))$ which terminated to the past at $(0, 0)$, then $\partial(J^+(\gamma))$ would be generated by outward radial null geodesics doing the same, so this cannot occur. Since projecting a non-radial such causal curve to the (v, r) coordinates yields a radial causal curve of the above form, these also cannot occur.

□

4.2.2 Curvature Strength

We observe that it is a trivial consequence of the proof of Proposition 2*(i)* that the Kretschmann scalar $K := \|\mathrm{Rm}\|^2 = \frac{48m(v)^2}{r^6}$ diverges along all identified null geodesics as they limit to $(0, 0)$, by virtue of the comparison $r(v) \leqslant k_+ m(v)$ in (4.5) for $0 < v \leqslant v_0$. We can show something stronger:

Proposition 3. *Consider the incoming Vaidya spacetime as above characterized by a nondecreasing mass function $m(v)$ which is continuous and piecewise C^1. There can exist at most one (modulo S^2) outgoing radial null geodesic which terminates to the past at $(v, r) \to (0, 0)$ and along which the Kretschmann scalar $K = \frac{48m(v)^2}{r^6}$ does not diverge.*

While such divergences were already known for the case $m'_+(0) > 0$ [75], it seems they were not known in general for $m'_+(0) = 0$. This significantly generalizes an explicit computation done by Joshi and Dwivedi [75] which showed such divergences for a particular form of the mass function with $m'_+(0) = 0$ ($m(v) \sim v^n$, $n > 1$). The present result indicates that the naked singularity in question *always* has divergent curvature associated with it, regardless of the form of $m(v)$. Importantly, this rules out the possibility of objecting to the physical significance of previously known results by maintaining that the singularity's strength was an artifact of a lack of regularity in the chosen mass function. Curvatures diverge no matter how smooth one takes $m(v)$.

Proof. We consider a solution $r(v)$ to (4.3) satisfying $\lim_{v\to 0^+} r(v) = 0$, and as in the previous proof we consider the quantity $k(v)$ defined by $r(v) = k(v)m(v)$. We note that, along the curve of interest,

$$K(v) = \frac{48m(v)^2}{r(v)^6} = \frac{48}{r(v)^4 k(v)^2},$$

so $\lim_{v\to 0^+} K(v) = \infty$ provided that $k(v)$ is bounded as $v \to 0$. It is not difficult to see from the bound $r(v) < \frac{v}{2}$ that $k(v)$ is bounded near $v = 0$ if $m'_+(0) > 0$ (meaning K always diverges in this scenario), so assume $m'_+(0) = 0$. It can be seen in this case that if $k(v)$ is unbounded near $v = 0$, then in fact $\lim_{v\to 0^+} k(v) = \infty$: if not, there is an $C > 0$ such that $4 < k(v_n) < C$ along some positive sequence $(v_n)_{n=1}^{\infty}$ with $v_n \to 0$, so eventually $\sup_{0<v<v_n} m'(v) < f(C) < f(k(v_n))$, which implies by (4.6) that $k(v) \leqslant k(v_n)$ for $v \in (0, v_n]$, showing that $k(v)$ is bounded near $v = 0$.

We now assume there exist two such solutions $r_1(v)$, $r_2(v)$ to (4.3), satisfying $\lim_{v\to 0^+} r_i(v) = 0$ and $\lim_{v\to 0^+} k_i(v) = \infty$ for $i = 1, 2$, and show they must be equal. In particular, for a fixed $\varepsilon > 0$ there is a $v_0 > 0$ such that $r_i(v) > (4 + \varepsilon)m(v)$ for $v \in (0, v_0]$– this is all that will be needed from the $k_i(v) \to \infty$ hypothesis to deduce

the equality. We restrict attention to the interval $(0, v_0]$. Setting $\lambda := \frac{1}{2} - \frac{1}{4+\varepsilon}$, we have by (4.3) that $\lambda v < r_i(v) < \frac{v}{2}$, so we further have that $\frac{r_2}{r_1} > 2\lambda$. Observe

$$\begin{aligned}(r_1 - r_2)' &= \frac{m}{r_2} - \frac{m}{r_1} = \frac{m}{r_2}\left(1 - \frac{r_2}{r_1}\right) \\ &< \frac{1 - \frac{r_2}{r_1}}{4+\varepsilon}. \end{aligned} \tag{4.7}$$

Setting $C_0 := \frac{1-2\lambda}{4+\varepsilon} = \frac{2}{(4+\varepsilon)^2}$, then, (4.7) together with the aforementioned bound on $\frac{r_2}{r_1}$ indicates

$$r_1(v) - r_2(v) < C_0 v.$$

Dividing through by r_1 and rearranging, however, yields the new bound $\frac{r_2}{r_1} > 1 - \frac{C_0}{\lambda}$. Combining this new bound with (4.7) now yields

$$r_1(v) - r_2(v) < C_1 v,$$

where $C_1 := \frac{C_0}{\lambda(4+\varepsilon)} = \frac{2}{2+\varepsilon} C_0$. Iterating this procedure, we deduce for each $n \in \mathbb{Z}_+$ that

$$r_1(v) - r_2(v) < C_n v,$$

where $C_n := \left(\frac{2}{2+\varepsilon}\right)^n C_0$. Since $\lim_{n\to\infty} C_n = 0$ (and by symmetry between r_1 and r_2), this implies $r_1(v) = r_2(v)$, as desired.

□

We learn from this result and its proof that, while in the case $0 < m'_+(0) \leqslant \frac{1}{16}$ every outgoing radial null geodesic terminating to the past at $(v, r) \to (0, 0)$ has bounded $k(v)$ (and hence divergent $K(v)$) near $v = 0$, when $m'_+(0) = 0$ there is exactly one (modulo S^2) outgoing radial null geodesic terminating to the past at $(v, r) \to (0, 0)$ along which $k(v)$ is unbounded near $v = 0$, and this curve necessarily generates the Cauchy horizon associated to a complete spacelike hypersurface in M. This curve's $K(v)$ will behave in boundedness like $\frac{m(v)}{v^3}$ near $v = 0$, which

may be bounded or unbounded depending upon the precise mass function (it is bounded for m sufficiently smooth). All others in the continuum of outgoing radial null geodesics emanating from $(0,0)$ will have bounded $k(v)$ (in fact, they will have $\lim_{v\to 0^+} k(v) = 2$), and hence divergent $K(v)$, near $v = 0$.

We now treat the globally naked subcase.

4.2.3 Globally Naked Singularities

Proposition 4. *(Globally Naked) Consider the incoming Vaidya spacetime as above characterized by a nondecreasing mass function $m(v)$ which is continuous and piecewise C^1.*

(i) If $\sup_{v>0} \frac{m(v)}{v} < \frac{1}{16}$, then there exists a one-parameter family (modulo S^2) of outgoing radial null geodesics which both terminate to the past at $(v,r) \to (0,0)$ and reach $r \to \infty$ to the future.

(ii) If there exists a $v_0 > 0$ at which $\frac{m(v_0)}{v_0} \geqslant \frac{1}{4}$, then no causal curve both terminates to the past at $(v,r) \to (0,0)$ and reaches $r \to \infty$ to the future.

We remark that the hypothesis of *(i)* essentially stipulates that the *average* growth rate of the mass function (between 0 and v) is always less than $\frac{1}{16}$, while that of *(ii)* stipulates that the average growth is at some point at least $\frac{1}{4}$. The heuristic physical takeaway from part *(i)* of Propositions 2 and 4, then, is that if the in-falling mass accumulates at the origin *slowly* enough, a signal from the initial singularity has time to escape before enough mass gathers to trap it. Part *(ii)* of these propositions indicates that these naked singularities *require* slow accumulation.

Proof. (i): Choose $C > 0$ with $\sup_{v>0} \frac{m(v)}{v} < C < \frac{1}{16}$. Notice that, in particular, $m(v) < Cv$ for $v > 0$. We first show the conclusion holds for $\widetilde{m}(v) = Cv$ (a result

already well-known by explicit computation, but which we show here for completeness), then compare. Quantities below with a tilde are defined according to the mass function $\widetilde{m}(v)$ instead of $m(v)$.

Fix $v_0 > 0$. By the Intermediate Value Theorem, we may find $k_\pm$ with $k_- \in (2, 4)$ and $k_+ \in (4, \infty)$ such that $f(k_\pm) = C$. Following the proof of Proposition 2*(i)*, we consider a solution $\tilde{r}(v)$ to (4.3) satisfying $k_-\widetilde{m}(v_0) < \tilde{r}(v_0) \leqslant k_+\widetilde{m}(v_0)$. (4.4) yields

$$\widetilde{\Delta}'_{k_\pm}(v) = \frac{\widetilde{\Delta}_{k_\pm}(v)}{k_\pm \tilde{r}(v)},$$

so $\widetilde{\Delta}_{k_-}(v)$ remains positive for all $v > v_0$ since it was positive initially, while $\widetilde{\Delta}_{k_+}(v)$ is nonpositive for all $0 < v < v_0$ since it was nonpositive initially. That is,

$$\tilde{r}(v) > k_-\widetilde{m}(v) = k_- Cv, \quad v \in [v_0, \infty)$$
$$\tilde{r}(v) \leqslant k_+\widetilde{m}(v) = k_+ Cv, \quad v \in (0, v_0]$$

These inequalities imply $\tilde{r}(v)$ will terminate to the past at $(0, 0)$ and reach $r \to \infty$ to the future.

Now, for each such $\tilde{r}(v)$ consider a solution $r(v)$ to (4.3) (now with $m(v)$) satisfying $r(v_0) = \tilde{r}(v_0)$. Defining $\Delta(v) := r(v) - \tilde{r}(v)$, we have

$$\begin{aligned}
\Delta'(v) &= r'(v) - \tilde{r}'(v) \\
&= \frac{\widetilde{m}(v)}{\tilde{r}(v)} - \frac{m(v)}{r(v)} \\
&> Cv\left[\frac{1}{\tilde{r}(v)} - \frac{1}{r(v)}\right] \\
&= \frac{Cv}{r(v)\tilde{r}(v)}\Delta(v).
\end{aligned}$$

Hence Δ remains positive if it is ever positive, and if Δ is ever 0, it must immediately become and remain positive. These constraints ensure that, since $\Delta(v_0) = 0$, we have

$\Delta(v) > 0$ for $v > v_0$ and $\Delta(v) < 0$ for $v < v_0$, showing that $\lim_{v\to 0^+} r(v) = 0$ and $\lim_{v\to\infty} r(v) = \infty$ from the same limits for $\tilde{r}(v)$.

(ii): Suppose $m(v_0) \geqslant \frac{v_0}{4}$ for some $v_0 > 0$. Consider a solution $r(v)$ to (4.3), and observe that if it is to reach $r \to \infty$, we certainly must have $r(v_0) \geqslant 2m(v_0)$ (if not, as usual we have $\Delta_2 < 0$ for $v > v_0$, which implies by (4.3) that $r'(v)$ is negative for $v > v_0$). Hence a solution which escapes to infinity satisfies

$$r(v_0) \geqslant 2m(v_0) \geqslant \frac{v_0}{2},$$

which implies $r(v)$ cannot terminate to the past at $(0, 0)$, as discussed in the proof of Proposition 2*(ii)*. Also as discussed there, this rules out any causal curve with the same properties.

□

The argument of (i) for $\widetilde{m}(v) = Cv$ carries through nearly exactly for any $m(v)$ with $\sup_{v>0} m'(v) = C < \frac{1}{16}$, but the less restrictive hypothesis of $\sup_{v>0} \frac{m(v)}{v} < \frac{1}{16}$ is accommodated by comparing to the linear case. One may observe that we only needed the strict inequality in $\sup_v \frac{m(v)}{v} < \frac{1}{16}$ for $v > v_0$, so one could relax this to an inclusive inequality on $v < v_0$. If one stipulates that the mass function should level off, for example, one may relax the hypothesis to $\sup_v \frac{m(v)}{v} \leqslant \frac{1}{16}$.

One may also observe that the hypotheses of (i) and (ii) above are, apart from the discrepancy between the values $\frac{1}{16}$ and $\frac{1}{4}$, very nearly logical negations of each other. This might lead us to hope that we may be able to close the gap between these values and obtain a complete characterization of globally naked singularities in these spacetimes. Marginal improvements as in the previous paragraph notwithstanding, this is unfortunately not possible in a direct manner: neither value can be improved, and one can either have or not have globally naked singularities in the gap. Regarding (i), for any $\varepsilon > 0$ there exist mass functions with $\sup_{v>0} \frac{m(v)}{v} < \frac{1}{16} + \varepsilon$ which have no causal curves *either* terminating at $(0, 0)$ to the past *or* reaching $r \to \infty$ to the future–

this is seen in the linear case $m(v) = Cv$ with $\frac{1}{16} < C < \frac{1}{16} + \varepsilon$. Regarding (ii), for any $\varepsilon > 0$ there exist mass functions possessing a $v_0 > 0$ at which $\frac{m(v_0)}{v_0} \geqslant \frac{1}{4} - \varepsilon$ while still having outgoing radial null geodesics which both terminate to the past at $(0, 0)$ and reach $r \to \infty$ to the future– though some technical details must be worked through to show this, the underlying idea is to consider an $m(v)$ which closely approximates the step function $v_0(\frac{1}{4} - \frac{\varepsilon}{2})H(v - v_0)$, where H is the Heaviside step function, while satisfying our regularity and structural constraints.

These considerations apparently indicate that Proposition 4 is optimal inasmuch as the range of $\frac{m(v)}{v}$ alone can readily characterize globally naked singularities. Further results in this vein must therefore be structurally different, perhaps hypothesizing a constraint on the duration that $\frac{m(v)}{v}$ exceeds $\frac{1}{16}$ (notice that step functions minimize this duration subject to the monotonicity condition). In any case, Proposition 4 is sufficient for the physical observations of interest in this work.

4.3 Conclusions

Proposition 4*(i)* implies that the singularity at $(v, r) \to (0, 0)$ is visible from infinity for an extended period provided, in particular, that the open condition $\sup_{v>0} m'(v) < \frac{1}{16}$ is met, while Proposition 3 implies that this singularity is always physically strong in some sense. A reasonable topology one might put on the collection of incoming Vaidya spacetimes is that induced by putting them in bijection with the set of admissible $m(v)$ endowed with the C^1-type norm

$$\|m\| = \sup_{v>0} \left[|m'_-(v)| + |m'_+(v)| \right].$$

Notice that this topology has the minimal fineness in m, and therefore in the initial data induced on some complete spacelike hypersurface, required to capture "closeness" in G (given (4.2)), and hence the implicit matter distribution. In such a

reasonable topology, the subset of Vaidya spacetimes exhibiting globally naked singularities has nonempty interior. We note that, since $m(v)$ is not required to level off, this observation is true both in and outside of the asymptotically flat context: the hypothesis of Proposition 4*(i)* can be met even while $\lim_{v\to\infty} m(v) = \infty$.

These observations are significant because they add to the small list of examples wherein naked singularities are apparently generic (at least, within spherical symmetry), this conclusion being made tractable due to the Vaidya spacetimes' relative simplicity– compared, say, to Tolman-Bondi models. Though the genericness is within the class of Vaidya spacetimes, varying within this class does not seem meaningfully more restrictive than varying initial data in a manner consistent with a particular "fundamental" matter field, in regards to degrees of freedom. One hopes for the sake of potential implications on weak cosmic censorship, of course, that this genericness is destroyed by perturbing outside of spherical symmetry. While the physical heuristic that slow accumulation allows an escaping signal certainly seems independent of spherical symmetry, it may well be that the singularity's forming early enough for it to be the source of such a signal depends critically on spherical symmetry's generically allowing the perfect focusing of matter shells at the origin.

While the Vaidya spacetimes are clearly in violation of the physical spirit of the Weak Cosmic Censorship Conjecture, they are apparently outside the technical scope of current rigorous formulations of the conjecture as an initial value problem (see, e.g. [29, 137]), as they do not a priori stipulate a specific matter field coupling to the Einstein equation to yield a PDE governing how the spacetime can be evolved from initial data. Since these spacetimes satisfy the DEC, however, and since the complete suite of fundamental physical matter fields is not known, one might argue they have just as much claim to exhibiting physical behavior as do toy matter models arising from a Lagrangian.

Of course, though the denomination is somewhat nebulous, we should not dismiss

out of hand that there is a well-reasoned school of thought that one should restrict entirely to "fundamental" matter models when formulating cosmic censorship so as to filter out singularities that may arise purely out of approximations and idealizations of matter (e.g. pressureless dust). This perspective can at least be traced to Eardley and Smarr [48] (who credit an unpublished report by Hawking), and further discussion can be found in Chapter 12 of Wald [137]. Much of the discussion on this topic predates Christodoulou's work [27, 28] on the scalar field case demonstrating that a "generic" qualifier is unavoidable, however, which opened up the question of whether both this *and* a restriction to fundamental matter is necessary. That is, it is worthwhile to recognize the logical distinction between weak censorship formulations stating

(i) Spacetimes subject to the DEC do not admit globally naked singularities.

(ii) Spacetimes containing only fundamental matter do not admit globally naked singularities.

(iii) Spacetimes subject to the DEC do not *generically* admit globally naked singularities.

(iv) Spacetimes containing only fundamental matter do not *generically* admit globally naked singularities.

While examples demonstrated very early on that (i) could not be true, Christodoulou's work in the 1990s demonstrated that (ii) is almost certainly not true either, though (iii) or (iv) may well be. Knowing which, if either, of these options ends up being viable would provide significant insight into the structure and content of general relativity. To put the present work into the context of this conversation, our results provide some of the strongest evidence that we are aware of (but still far from definitive due to the restriction to spherical symmetry) that (iii) may not be true, while

making no direct implications about (iv) (though hints surrounding the structure of naked singularities gleaned herein may be helpful in understanding (iv)).

In any event, it is clear from the manner in which our (and many others') analysis has proceeded that the physical evaluation of cosmic censorship in examples is often most naturally done in regards to the totality of a spacetime and its structure, rather than through initial data and evolutions thereof. It is among the primary objectives of the following Chapter to put forward a more comprehensive technical formulation of weak cosmic censorship which is better aligned with this perspective, and hence more naturally accommodates the physically objectionable behavior demonstrated here– that is, a technical formulation better tuned to handling the broadest interpretation of version (iii) above.

5

Defining Black Holes and Posing Weak Censorship

As discussed in the previous chapter, the problem of cosmic censorship is central to ensuring that classical general relativity makes sense at its core. While singularities apparently cannot be avoided, it is crucial to establish how pervasive they truly are, and in particular to discern whether one can reasonably expect to be able solve Einstein's equation in domains of interest for "most" sets of physically reasonable initial data. Weak cosmic censorship heuristically asks whether general relativistic evolution can at least usually be fully propagated, in a manner free of singularities, outside of the pathological interiors of black holes. While this is a simple enough sentiment to express while waving one's hands, it has been difficult to make it precise enough that it is rigorously meaningful while still capturing the full spirit and breadth of the question. Central to this problem is making precise, in a versatile enough setting, the concept of a black hole in the first place. It is our objective in this final chapter to put forward a rigorous characterization of black holes suited to this task. The parlance of this chapter will be careful mathematical definitions, deductions, and argumentation rather than equations and computation. This chapter largely reproduces my work in [143].

The phenomenon of black holes is among the most surprising and intriguing features of Einstein's General Theory of Relativity– these causally self-contained regions of spacetime captivate the minds of laypeople and seasoned physicists alike. Among their most widely-discussed features in both of these audiences are the singularities that black holes are generally said to contain. It is a curious situation, then, that the standard definition of what physicists formally mean by the term "black hole" in classical General Relativity, i.e. the complement of the causal past of an idealized "future null infinity", neither makes reference to nor is known to guarantee the existence of singularities. It is further curious that while the feature that makes the study of black holes so pertinent is their seemingly generic nature, this conventional definition is anything but. Indeed, while cosmological investigations lead us to believe the universe at large is definitively not asymptotically flat given the presence of dark energy [2, 115, 109] it continues to be the case that the most standard and robust classical definition of black holes hinges on asymptotic flatness. Nevertheless, this standard approach and existing variations upon it have proven remarkably fruitful and insightful, so we by no means wish to supplant them: rather, it is the objective of this chapter to provide an additional and complementary perspective on defining black holes which ameliorates the above curiosities.

Though both black holes and singularities in general relativity have a storied history, the programs of research devoted to them diverged somewhat following the seminal result associating them, Penrose's widely celebrated Incompleteness Theorem [104], for which he was partially awarded the 2020 Nobel Prize in Physics. The phenomenon of singularities in general relativity has been investigated in its own right, independently of any black holes to which they may be associated, for over half a century through the development of various so-called "boundary constructions". This course of study seemingly began in 1960 with Szekeres' *On the Singularities of a Riemannian Manifold* [132], which has been succeeded by various constructions

seeking to associate to a general (time-oriented) spacetime manifold (M, g) a topological space $\overline{M}$ consisting of M together with a collection of boundary points ∂M, some or all of which are identified as singularities. Among the best-known and most influential such constructions include the geodesic boundary (g-boundary) of Geroch [49], the bundle boundary (b-boundary) of Schmidt [122], and the causal boundary (c-boundary) of Geroch, Kronheimer, and Penrose [53]. We shall be particularly interested in the abstract boundary (a-boundary) of Scott and Szekeres [125].

While the various boundary constructions get rather technically involved, they generally achieve a somewhat simple goal: by constructing an appropriately physical topology on $\overline{M}$, they identify what it means for a point in the spacetime manifold M to be "near" a singularity. Crucially, they do this in a manner that is entirely intrinsic to M– no additional structure on the spacetime manifold need be invoked in order to make the identification. In this sense, programs for identifying "where" singularities lie in General Relativity are available for completely general spacetimes. This will be a key ingredient in our definition of a black hole for a general spacetime.

Meanwhile, the study of black holes and their features has been principally treated, or at least motivated, via their original characterization, due to Penrose [107], as that subset of spacetime which cannot be seen from infinity, provided one restricts spacetime to be of a form in which one has a particularly well-behaved notion of infinity. This is discussed at length in classic texts [60, 137] and reviewed below. This has been plenty sufficient for many observational and numerical modeling investigations, as well as theoretically useful to proving many important and well-known properties of black holes, e.g. thermodynamic properties [9], in particular Hawking's area theorem [59], and the Riemannian Penrose inequality [18, 71]. Beyond the aesthetic appeal of having a consistent definition which makes unequivocal sense in the context of cosmology and beyond, it is of great theoretical interest to ask how such properties may or may not extend to a broader description of black

holes, as well as to investigate what insight might be gained into significant open questions when formulated under such a description.

Indeed, we use our definitions to put forward a more comprehensive formulation of weak cosmic censorship in Section 5.5, a problem to which our perspective on black holes is uniquely suited due to its intimate ties to singularities. We argue there that such a reformulation is needed to handle naked singularities that cannot be readily accommodated via the standard IVP formulation. Such naked singularities include those present in the Vaidya spacetimes, demonstrated in the previous chapter (and in [142]) to be significantly more generic than previously discussed in the literature.

The problem of defining black holes is one of many facets, of course: many different formal characterizations have been put forward in different subfields of physics, and even more numerous heuristic characterizations are used in practice. For a discussion of the history of this topic in the various subfields of physics, see Curiel's essay [35]. We work strictly within the domain of classical general relativity, wherein alternatives to the standard paradigm are generally based upon local convergence properties of geodesics and suggesting one identify black holes by a new type of horizon loosely modeled after apparent horizons (prototypes were provided by Krolak [83], Hayward [61, 62], and Ashtekar et al. [6]). These are generally conceived with practical numerical and semi-classical thermodynamic concerns in mind. While we comment on these very briefly in Section 5.1, see [7, 100] for reviews of the status of such horizon-based approaches.

This chapter will be organized as follows. In Section 5.1, we give a brief review of the standard characterization of black holes (following Hawking and Ellis [60]). In Section 5.2, we motivate and put forward the key novel definitions of interest, given that one has a notion of "where" singularities are (e.g. via a boundary construction), leading to the proposed definition of black holes. We illustrate these definitions in several examples highlighting both similarities and differences from the classic defi-

nition. In Section 5.3, we briefly weigh the merits of various boundary constructions before committing to the a-boundary (reviewed in the appendix) and using it to formalize singular neighborhoods. In Section 5.4, various results surrounding the proposed notion of black hole are presented, proven, and discussed. In Section 5.5, we discuss application of the formalism to weak cosmic censorship.

5.1 The Classic Perspective

The classic formalization of the notion of a black hole depends on a hierarchy of technical mathematical structure building up to formalizing *asymptotic flatness* of the spacetime (M, g), meant to encode that the spacetime is comprised of an isolated or bounded system of matter. The precise details of this hierarchy are subject to some variation (see, for example, [107] versus [60] versus [137], or [37] for a more modern perspective), but the broad strokes are essentially the same. We present the hierarchy as described in Hawking and Ellis [60], so definitions in this section are taken from there. At the base is an *asymptotically empty and simple* spacetime:

Definition 5. *A time- and space-orientable spacetime $(\widetilde{M}, \tilde{g})$ is asymptotically empty and simple if there exists a strongly causal spacetime $(\widehat{M}, \hat{g})$ and a conformal embedding $\phi : \widetilde{M} \to \widehat{M}$, under which $\overline{\phi(\widetilde{M})}$ is a submanifold with smooth boundary $\partial\phi(\widetilde{M})$, satisfying:*

(1) There is a sufficiently smooth function $\Omega : \widehat{M} \to \mathbb{R}$ satisfying $\Omega > 0$ on $\phi(\widetilde{M})$ such that $(\Omega \circ \phi)^2 \tilde{g} = \phi^* \hat{g}$.

(2) $\Omega = 0$ and $d\Omega \neq 0$ on $\partial\phi(\widetilde{M})$.

(3) Every inextendible null geodesic in $\widetilde{M}$ has two endpoints on $\partial\phi(\widetilde{M})$.

(4) There exists an open set $\widehat{U} \subset \widehat{M}$ containing $\partial\phi(\widetilde{M})$ such that $\widetilde{Ric} = 0$ on $\widetilde{U} := \phi^{-1}(\widehat{U})$.

Condition (1) indicates that $(\widehat{M}, \hat{g})$ encodes the global causal structure of $(\widetilde{M}, \tilde{g})$; condition (2) ensures that $\partial\phi(\widetilde{M})$ is indeed at "infinity" with respect to $\widetilde{M}$ in the sense that any null geodesics in $\widetilde{M}$ approaching $\partial\phi(\widetilde{M})$ must have infinite affine parameter; condition (3) ensures that $\partial\phi(\widetilde{M})$ includes the entirety of infinity; condition (4) ensures, given the Einstein equation, that the matter under consideration is bounded away from infinity. For more discussion of these conditions and their motivations and implications, see [60]. Perhaps most importantly, they imply that $\partial\phi(\widetilde{M})$ is a null hypersurface in $\widehat{M}$ comprised of two disconnected pieces, *future null infinity* $\mathscr{J}^+$ and *past null infinity* $\mathscr{J}^-$, the collections of future and past (respectively) endpoints in $\widehat{M}$ of inextendible null geodesics in $\widetilde{M}$. It was among the first successes of boundary constructions that they demonstrated that the closure of $\phi(\widetilde{M})$ in $\widehat{M}$ is unique, e.g. independent of the choice of ϕ and $\widehat{M}$ [49, 121], so that the structure at infinity really is intrinsically meaningful to $\widetilde{M}$ in this setting.

While effective at capturing many desired qualities, this definition is apparently too restrictive in that it manifestly requires $\widetilde{M}$ to be null geodesically complete, and so cannot be applied directly to the standard black hole solutions (e.g Schwarzschild, Reissner-Nordstrom, Kerr), or any spacetimes including trapped surfaces and satisfying the hypotheses of Penrose's Incompleteness theorem. The standard remedy is to put forward the notion of a *weakly asymptotically simple and empty* spacetime:

Definition 6. *A spacetime (M, g) is said to be weakly asymptotically empty and simple if there is an open set $U \subset M$ and an asymptotically empty and simple spacetime $(\widetilde{M}, \tilde{g})$ with an open neighborhood $\widehat{U}$ of $\partial\phi(\widetilde{M})$ in $\widehat{M}$ such that (U, g) is isometric to $(\widetilde{U}, \tilde{g})$, where $\widetilde{U} := \phi^{-1}(\widehat{U})$.*

This definition simply describes a spacetime which looks weakly asymptotically empty and simple in some region including an apparently full notion of infinity. Any

choice of such a $U \subset M$ and associated asymptotically empty and simple spacetime $(\widetilde{M}, \tilde{g})$ endows M with a future null infinity $\mathscr{J}^+ \subset \widehat{M}$, whose causal relation to M is unambiguous in that we can make sense of the causal past of $\mathscr{J}^+$ in M. To be painfully explicit, denoting by $\psi : U \to \widetilde{U}$ the isometry referred to in the most recent definition, this relation is

$$J^-(\mathscr{J}^+) := J^-\left((\phi \circ \psi)^{-1}\left(\widehat{J}^-(\mathscr{J}^+)\right)\right),$$

where $\widehat{J}^-(\cdot)$ refers to taking the causal past in $\widehat{M}$. There *is* ambiguity, however, implicit in the choice of U and $(\widetilde{M}, \tilde{g})$ above. To make sense of $J^-(\mathscr{J}^+)$, then, one must fix this choice.

At this point, the core structures are in place to allow one to define a black hole in the classic sense. It is simply that subset of a weakly asymptotically empty and simple spacetime which cannot "escape to infinity". That is,

Definition 7. *In a weakly asymptotically empty and simple spacetime (M, g), the (classic) black hole region with respect to a future null infinity $\mathscr{J}^+$ is defined as*

$$\mathscr{B}_c := M \backslash J^-(\mathscr{J}^+).$$

The term "black hole" now typically refers to the intersection of $\mathscr{B}_c$ with a space-like hypersurface $\Sigma \subset M$. Though perfectly meaningful and aligned with the physical criteria of what a black hole should be, this definition is somewhat pre-emptive in that one typically restricts a bit more the class of spacetimes within which one identifies $\mathscr{B}_c$ so as to render provable various celebrated results on black hole properties and dynamics. In particular, these often require working within the following domain:

Definition 8. *A weakly asymptotically empty and simple spacetime (M, g) is said to be future asymptotically predictable provided that there exists an edgeless acausal subset $\mathscr{S} \subset M$ such that $\mathscr{J}^+$ is in the closure of $D^+(\mathscr{S})$, the future Cauchy development of $\mathscr{S}$, in $\widehat{M}$.*

This condition essentially encodes that there is a three-dimensional spacelike hypersurface in M from which $\mathscr{J}^+$ can be evolved. This is interpreted as meaning that there are no non-initial naked singularities, singularities visible from $\mathscr{J}^+$ to the future of this hypersurface. This condition is generally considered reasonable, then, since it is seen as tantamount to assuming the veracity of the Weak Cosmic Censorship Conjecture. With it (or perhaps with a slight strengthening), one may prove [60] that closed trapped surfaces, outer trapped surfaces, and apparent horizons in $D^+(\mathscr{S})$ always lie in $\mathscr{B}_c$, and that black hole boundaries increase in area over time. This is then the basis upon which many take trapped surfaces and variations on the concept (marginally trapped surfaces, apparent horizons, minimal surfaces in time-symmetric initial data, etc.) to be quasi-local stand-ins for the concept of a black hole when working in particular contexts, e.g. numerical relativity or initial data formulations of the Penrose inequality. Indeed, the extant attempts at generalizing the term "black hole" outside of asymptotically flat contexts [7, 13, 61, 83] are built around these ideas.

The most obvious limitation of Definition 7 is its dependence on the hierarchy of structure built up over Definitions 5, 6, and 8. These are significant constraints on (M, g) invoking considerable extrinsic structure which, while inarguably useful for investigations both numerical and theoretical into the properties of approximately isolated systems in general relativity, seem well beyond what should be required to characterize the intuitive concept of a black hole as a region of no escape, and are manifestly inapplicable to cosmological situations which cannot be well-approximated as isolated (e.g. primordial black hole formation). We take the perspective that this concept should be characterizable in a general spacetime in a completely intrinsic manner. There is also implicit ambiguity in Definition 7 inherited from that in the notion of $\mathscr{J}^+$ induced by Definition 6, in that there may well be multiple distinct neighborhoods $U \subset M$ isometric to a neighborhood of infinity

in some $\widetilde{M}$. This is not at all unfamiliar, as it is an immediate feature of the maximal extensions of the most standard black hole spacetimes that there are multiple distinct future null infinities, and so associated to each of them there is a distinct black hole region (though they may partially overlap, as in the maximally extended Schwarzschild case). "The" classic black hole region $\mathscr{B}_c$ of a given spacetime is then a matter of perspective.

While, as previously mentioned, some more broad approaches to defining black holes have been put forward, this broadening has generally come at the cost of losing direct association to a core feature of the intuitive physical description: alternative approaches are largely based on using local convergence properties of geodesics to identify new types of horizons [6, 61, 62, 83], which may not in general enjoy the "trapping" features with respect to infinity guaranteed in the future asymptotically predictable setting. Sets identified via these conditions simply need not bound a region that is meaningfully "small", as one might like. Indeed, the trapped surfaces in the maximally extended Kerr spacetime have the property that their causal future always includes some $\mathscr{J}^+$ (infinitely many, in fact– see Figure 5.7), so it is certainly not the case that trapped surfaces bound a region that is "trapped" in any universal sense. Without the structure of asymptotic flatness to pick out an infinity with respect to which they are "trapped", then, these surfaces and their generalizations lose their direct link to a significant component of the black hole concept. This is further exemplified in non-singular constructions conforming to some such definitions (but not Definition 7), such as Hayward's [63], wherein every timelike curve exits the region identified as a black hole. Depending on one's objective, this may not be so dire a cost, particularly from a semi-classical perspective seeking to accommodate escaping Hawking radiation (as in Hayward's case). Moreover, these convergence characterizations have certainly been valuable to numerically identifying black holes,

investigating primordial black holes in practice [57], and deciding how to best formulate black hole thermodynamics. There are contexts within which they fall short, however – in particular, they are not suited for the purely classical questions of weak cosmic censorship–, and so we feel something is still missing in the effort to fully, intuitively, and rigorously characterize black holes in a general setting.

5.2 What Makes a Black Hole?

5.2.1 A Dual Perspective

Having seen the standard construction which we hope to build upon, we should now step back and reflect: what is the essence of a black hole which it captured, and so which any extension must capture as well? The heuristic layman's definition, of course, is that a black hole is *a region of spacetime in which gravity is so strong that light cannot escape*– this is perhaps the most crucial and recognizable feature of a black hole. By itself, though, this is clearly not sufficient as a technical definition, as it is true directly by definition of any future set whatsoever (a set $F \subset M$ such that $J^+(F) = F$) that light cannot escape from it. Such sets include $J^+(p)$ for any $p \in M$, $M \backslash J^-(p)$ for any $p \in M$, or even the entirety of M itself; these clearly should not be called black holes in general. There is another feature of a black hole which a complete definition should capture, however: a black hole is "small" in some appropriate sense. Taking all of M most glaringly demonstrates the need for this restriction, as it is not so insightful to note that light cannot escape the whole of spacetime. This is only a meaningful constraint when the light is somehow bounded.

How can we mathematically capture the idea of a black hole's boundedness? The standard construction's answer to this question is to appeal to a notion of infinity: the light of a standard black hole is "bounded" in the sense that it cannot reach the infinity provided by the structure of asymptotic flatness. We hope to find an answer to this question even when asymptotic flatness cannot provide such an

infinity. Extant attempts are content with a local stand-in for boundedness through geodesic convergence, while we seek to preserve the global outlook. To do so, we shall take a perspective somewhat dual to the standard one– we ask not which light cannot reach infinity, but which light *must* reach a singularity. Indeed, if light approaching the edge of spacetime is not to reach any notion of infinity, then a singularity seems the only alternative.

The topologist's favorite characterization of "small" or "bounded" sets is, of course, compactness. In the context of manifolds, which are topologically sequential, compact sets are simply those in which every sequence of points has a convergent subsequence. This nicely describes that a compact set is "small" in the sense that an infinite collection of points cannot spread out– they must tend to gather around at least one point in order to all fit in the set. This would be a fine characterization of "small", except that it utterly fails to capture the smallness of black holes. Indeed, in the prototypical case of the Schwarzschild spacetime, one may take any sequence of points in the black hole region along which $r \to 0$ monotonically, and this can have no convergent subsequence. This is demonstrated in the Penrose diagram of the Schwarzschild spacetime in Figure 5.1: while the sequence depicted in red apparently converges to a point on the $r = 0$ line of singularity, this point is not in the spacetime, so the sequence has no subsequence which converges to a point in the spacetime, and hence the closed set A which contains the sequence is non-compact. Being a subset of the black hole, however, A must be "small" in whatever sense the black hole is.

This apparent failure of compactness can be remedied. What we notice is that the failure of a sequence in the depicted set A to have a convergent subsequence can *only* happen in this way– problematic sequences *must* approach the $r = 0$ singularity. So that we might convey the core content and pictorial intuition of our construction undistracted by technicalities, let us for the moment sweep under the rug the problem of identifying "where" the singularities are, to be returned to in Section 5.3. That

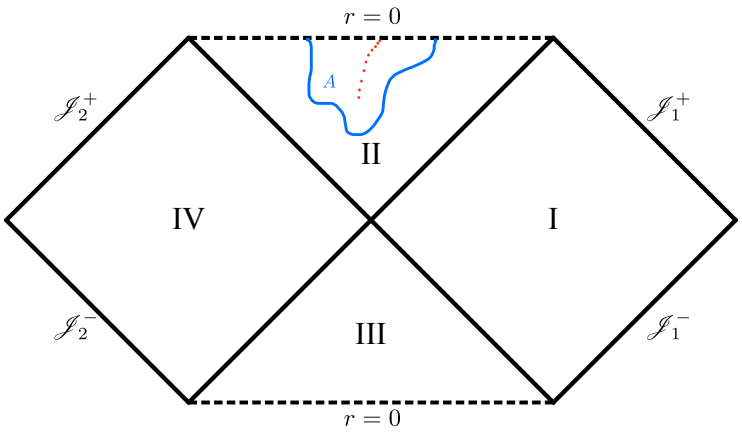

FIGURE 5.1: Schwarzschild Penrose diagram depicting a sequence in the black hole, region II, with no convergent subsequence. Also shown is a closed set A in region II containing the sequence which should be "small" in whatever sense the black hole is.

is, let us assume that we are gifted the collection $\mathscr{U}$ of open subsets of M which are deemed to envelop all singularities of M, in the sense that any sequence in M which "approaches a singularity", whatever that might mean, must eventually enter and remain inside each $U \in \mathscr{U}$. These open sets will be called *singular neighborhoods*.

As some examples, the family $\mathscr{U}$ of singular neighborhoods should always include the entire manifold, i.e. $M \in \mathscr{U}$, as well as the complement of any compact set. It is closed under union with arbitrary open sets and finite intersection. In a "non-singular" spacetime (again, whatever that might mean– though it should certainly include Minkowski space), $\mathscr{U}$ should simply be the entire topology on M, as the condition of approaching a singularity is then null. Some examples for the Schwarzschild spacetime are depicted in Figure 5.2.

Armed with the collection of singular neighborhoods, we are prepared to introduce our means of characterizing the "smallness" of the set A depicted in Figure 5.1:

Definition 9. *Let (M,g) be a spacetime manifold. A closed set $A \subset M$ is called singularly compact if $A \backslash U$ is compact for every singular neighborhood $U \in \mathscr{U}$.*

This definition is saying, then, that singularly compact sets are precisely those that

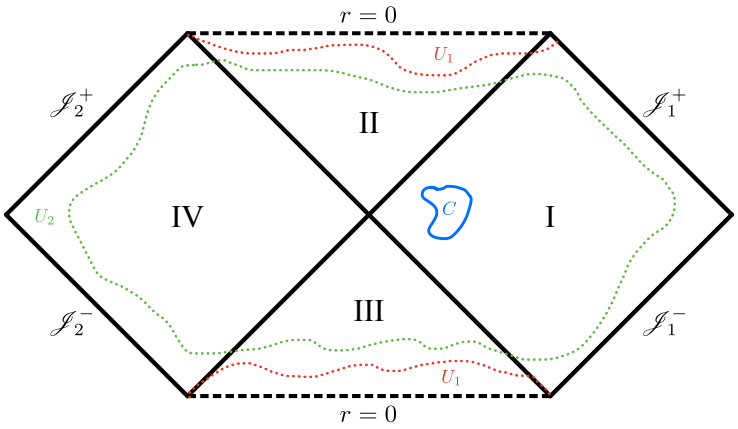

FIGURE 5.2: Examples of singular neighborhoods in the Schwarzschild spacetime. U_1 is the region above the top and below the bottom red dotted curves. U_2 is the exterior of the closed green dotted curve. The complement of the compact set C is also a singular neighborhood. All of these "surround" the singular $r = 0$ lines by as much as possible from within M. If one removes any of these from the set A of Figure 5.1, it becomes compact.

are compact outside of singular behavior, i.e. except for the possibility of sequences approaching singularities. Indeed, it is immediate from the definition that a closed set A is singularly compact iff every sequence in A with no convergent subsequence eventually enters and remains in every singular neighborhood. The family of singularly compact sets is closed under intersections and finite unions, and closed subsets of singularly compact sets are singularly compact; compact sets are trivially singularly compact. In a "nonsingular" spacetime, the singularly compact sets are precisely the compact sets. Physically, singularly compact sets are simply "finite" or "small" sets that may be arbitrarily close to any singularities.

Definition 10. *Let $\mathscr{F}$ be the family of singularly compact future sets, i.e. singularly compact sets $A \subset M$ satisfying $J^+(A) = A$. The black region $\mathscr{B} \subset M$ is given by*

$$\mathscr{B} := \bigcup_{A \in \mathscr{F}} A.$$

A connected component of the intersection of $\mathscr{B}$ with a spacelike hypersurface is called a black hole.

This definition is the proposed formal characterization of black holes, applicable to any spacetime, which we wish to put forward. It stipulates that a point $p \in M$ is in $\mathscr{B}$ if and only if it lies in a singularly compact future set, a "small" set from which no causal signal can escape. One could, of course, define a white region $\mathscr{W}$ and its associated white holes in a completely analogous manner by replacing future sets with past sets everywhere, but we will not treat this explicitly. Notice that this definition makes no reference to a particular choice of region at infinity, as is implicit in Definition 7 via its invocation of $\mathscr{J}^+$, so these two definitions are certainly not identifying the same subset of M (when Definition 7 is applicable). Instead, Definition 10 hopes to identify those points which would be said to be in a black hole with respect to *any* portion of infinity, which any observer should agree would be in a black hole. We explore this distinction further in the examples to follow. First, a simple consequence of the definition:

Lemma 5. *For any $p \in M$, $p \in \mathscr{B}$ iff $\overline{J^+(p)}$ is singularly compact.*

Proof. If $p \in \mathscr{B}$, $p \in A$ for some singularly compact set A satisfying $J^+(A) = A$, and hence $J^+(p) \subset A$. Since A is closed, $\overline{J^+(p)} \subset A$, so for every singular neighborhood U, $\overline{J^+(p)} \backslash U$ is a closed subset of the compact set $A \backslash U$, hence is compact, showing $\overline{J^+(p)}$ is singularly compact.

Conversely, if $\overline{J^+(p)}$ is singularly compact, then since $J^+(\overline{J^+(p)}) = \overline{J^+(p)}$, we have $\overline{J^+(p)} \in \mathscr{F}$, so $p \in \overline{J^+(p)} \subset \mathscr{B}$.

□

This straightforward lemma indicates that a point lies in $\mathscr{B}$ iff its causal future is "small" in the sense of singular compactness, so that the question of a point's being in the black region is a question about its causal future. This means that the characterization for a point $p \in M$ is somewhat localizable, in that if one modifies the spacetime only outside of $\overline{J^+(p)}$, it should not change whether p is deemed to be

in $\mathscr{B}$. If one perturbs or adjusts the Schwarzschild spacetime in the $r > 2m$ region in any way (even destroying asymptotic flatness), for example, the black piece of the $r < 2m$ region should still be deemed to be in a black hole. Precise results of this nature can only be formally proven, of course, once one has precisely described singular neighborhoods. We detail one way of doing so in Section 5.3, but first we explore the intuition and heuristics of Definition 10 in some standard spacetimes.

5.2.2 Examples

We discuss the application of Definition 10 to some standard black hole spacetimes, temporarily identifying singular neighborhoods in a heuristic fashion for conceptual clarity. In each of the following, "singularities" are simply identified via inextendible, incomplete geodesics in maximal spacetimes, given a point-set and topological structure via standard coordinate charts in which these geodesics have endpoints. We present these examples prior to providing a more precise, coordinate-invariant formulation of singular neighborhoods in Section 5.3 to emphasize that there are multiple ways one might attempt to rigorously characterize singular neighborhoods, and that the singular neighborhoods depicted in Figures 5.2 through 5.7 should be taken as the underlying ideal of what we would *like* the term to mean.

As a zeroth example, we comment that a trivial class of examples are the aforementioned "nonsingular" spacetimes, which admit no singularly compact future sets, and hence have empty black region $\mathscr{B}$, under mild causality constraints (say, a distinguishing condition– see Lemma 10).

Example 1. *M = the Schwarzschild spacetime.*

Using Lemma 5, we would like to identify those points in the Schwarzschild spacetime which are in the black region $\mathscr{B}$. We ask, then, which points $p \in M$ have the property that their causal future is singularly compact. This analysis is

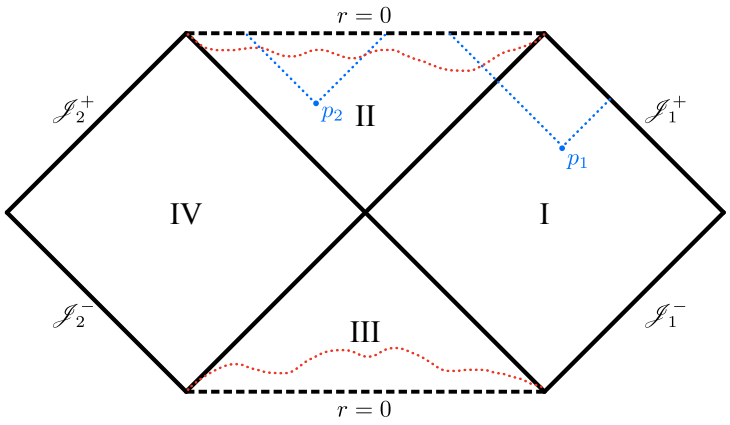

FIGURE 5.3: Identifying the black region in the Schwarzschild spacetime. The red dotted curves form the boundary of a typical singular neighborhood U (similar to U_1 in the previous figure), while the blue dotted lines are the boundaries to the causal futures of the points p_1 and p_2. These causal futures are then everything above the boundaries.

represented pictorially in Figure 5.3: we consider points p_1 in region I and p_2 in region II (the subset commonly identified as the black hole). The critical observation to make is that upon removing from $J^+(p_2) = \overline{J^+(p_2)}$ any singular neighborhood (such as U, demarcated by the red dotted curves) it clearly becomes compact, being closed and bounded in the plane of the Penrose diagram. On the other hand, the singular neighborhood U is an example of one for which $J^+(p_1)\backslash U$ is noncompact– even upon the removal of U, there remain sequences in $J^+(p_1)$ which limit to the future null infinity $\mathscr{I}_1^+$, which is not in the spacetime.

These observations demonstrate that $p_2 \in \mathscr{B}$, while $p_1 \notin \mathscr{B}$. Clearly the same reasoning as for p_2 applies to any point in region II, while the same reasoning as for p_1 applies to any point in either asymptotically flat region (I and IV) or the white hole region (III). Hence region II is a subset of $\mathscr{B}$, while any point outside of region II is not in $\mathscr{B}$. The boundary of region II in M is then identified as the event horizon $H = \partial\mathscr{B}$ as expected, and points in this boundary are not included in $\mathscr{B}$ according to how we have drawn U (not necessarily enveloping the timelike infinities).

This picture aligns with expectations quite well, but it is worth distinguishing

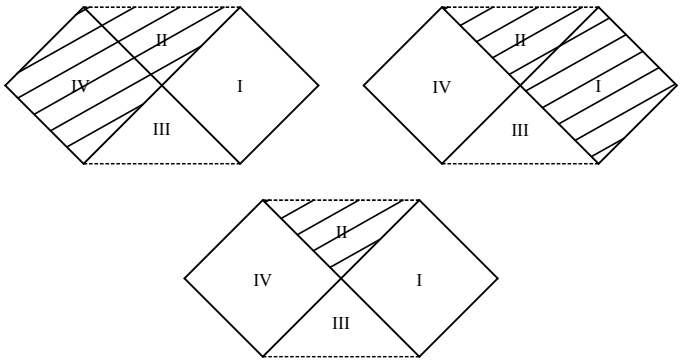

FIGURE 5.4: The two possible classic black hole regions $\mathscr{B}_c$, shown shaded, in the Schwarzschild spacetime (top), contrasted with the single black region $\mathscr{B}$ of Definition 10 (bottom).

it from the standard paradigm. The sequence of definitions in Section 5.1 yields that there are two distinct classic black hole regions $\mathscr{B}_c$, according to whether one chooses $\mathscr{J}_1^+$ or $\mathscr{J}_2^+$ as the preferred notion of infinity. With respect to $\mathscr{J}_1^+$, the black hole region is comprised of regions II and IV; with respect to $\mathscr{J}_2^+$, it is comprised of regions II and I. In contrast, Definition 10 identifies a single black region $\mathscr{B}$, comprised of region II alone– the (interior of the) intersection of the two classic black hole regions. See Figure 5.4. Of course, dual statements hold for white holes.

□

Example 2. *M = the deSitter Schwarzschild spacetime.*

We study the deSitter Schwarzschild metric, which minimalistically models a black hole situated within an expanding universe (effectively being Schwarzschild plus a positive cosmological constant). In general, this can be done atop an immense variety of maximally extended topologies for M [98]– shown in Figure 5.5 is the Penrose diagram for the "largest" choice, the universal cover to all others typically considered, with points p_1 and p_2 in the regions II_c and II_b, respectively. The subscript c denotes those regions near "cosmological" infinity, outside the cosmological horizons centered around the Schwarzschild-like regions II_b and III_w. A first observation is that this spacetime is definitively not asymptotically flat due to the presence of

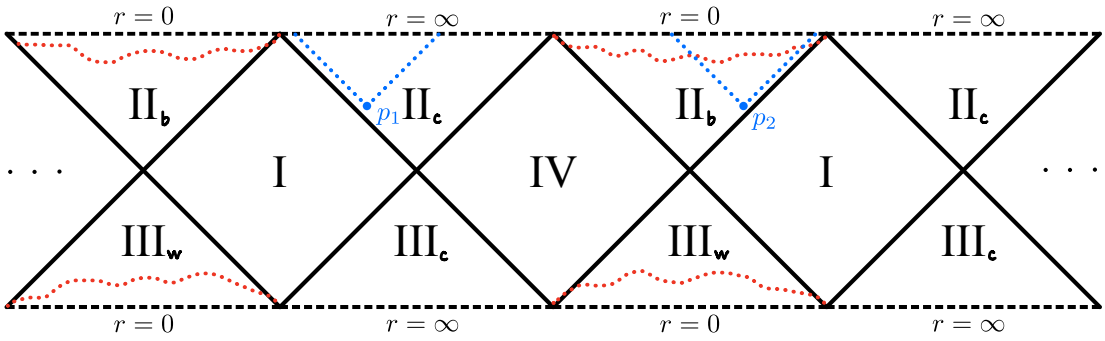

FIGURE 5.5: Identifying the black region in the deSitter Schwarzschild spacetime.

the cosmological constant, which manifests in the Penrose diagram via the boundary apparently at "infinity" being spacelike instead of null. While the conceptual picture laid out in Section 5.1 for identifying $\mathscr{B}_c$ may be applied in spirit since there seem to be natural candidates for pieces of "future infinity" in this spacetime, the technical details may not, and standard results surrounding $\mathscr{B}_c$ therefore cannot be directly applied. Proceeding heuristically yields many choices of what one might mean by the black hole region $\mathscr{B}_c$, one for each component of $r = \infty$, and each instantiation of $\mathscr{B}_c$ contains every other cosmological infinity. See Figure 5.6.

The identification of $\mathscr{B}$, meanwhile, follows the Schwarzschild analysis rather directly. Indeed, looking at the causal future of p_2, it is still the case that it becomes compact upon removing the typical singular neighborhood shown in red, so $p_2 \in \mathscr{B}$. Removing the same from $J^+(p_1)$, there remain sequences in $J^+(p_1)$ which approach the $r = \infty$ boundary, so $p_1 \notin \mathscr{B}$. As before, these reasonings can easily be extended to find that the regions of type II_b are all contained in $\mathscr{B}$, while the regions of type

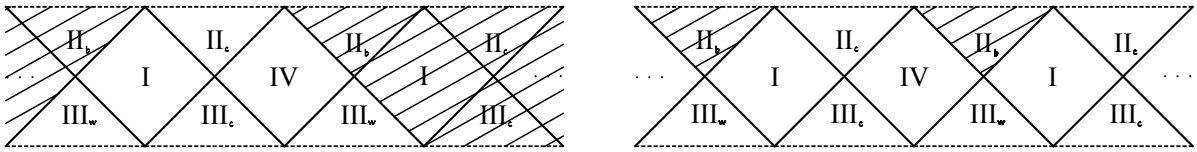

FIGURE 5.6: A classic black hole region $\mathscr{B}_c$ in the deSitter Schwarzschild spacetime with respect to a particular component of "infinity" (left), contrasted with the black region $\mathscr{B}$ (right).

I, II_c, III_c, III_w, and IV are all in the complement of $\mathscr{B}$. Hence, $\mathscr{B}$ is precisely the union of all regions of type II_b, as one would expect.

Definition 10, then, is naturally able to accommodate black holes to which the standard paradigm does not directly apply and give the expected results, with no global structure required. These considerations are robust to standard quotienting of Figure 5.5 down to any finite number of black holes (e.g. quotienting by the group generated by a translation identifying regions of a common type, or by certain horizontal reflections composed with the S^2 antipodal map, as in [98]).

□

Example 3. *M = the Kerr spacetime.*

Having seen a successful deviation of Definition 10 from the standard paradigm with the deSitter Schwarzschild spacetime, we now turn to one which may be seen as questionable in the maximally extended Kerr spacetime. See the usual (subextremal) Penrose diagram of a $\theta = \frac{\pi}{2}$ slice in Figure 5.7. Depicted there is a point p in a region of type II, with the boundary of its causal future shown as blue dotted lines. We notice that, though p is in the classic black hole region $\mathscr{B}_c$ with respect to $\mathscr{I}_1^+$ (shown on the right) and $\mathscr{I}_2^+$, it is *not* with respect to $\mathscr{I}_3^+$ and $\mathscr{I}_4^+$. In fact, it is the case in maximally extended Kerr that every point in M is in the past of the future null infinity associated to some asymptotically flat region, in contrast to region II in Schwarzschild (Figure 5.3) and regions II_b in deSitter Schwarzschild (Figure 5.5). Since $\mathscr{B}$ is heuristically supposed to be the intersection of all possible $\mathscr{B}_c$, this means that $\mathscr{B}$ should be empty!

Indeed, looking at the causal future of p (as for any other point in M), we see that upon the removal of the typical singular neighborhood shown (red dotted curves), there still remains a sequence limiting to a sufficiently "late" $\mathscr{I}^+$. This indicates $J^+(p)$ is not singularly compact, and so $p \notin \mathscr{B}$, and $\mathscr{B}$ is empty. That is, there

is no black region in the Kerr spacetime according Definition 10. This should not be entirely surprising, as there is apparently no universal sense in which any classic black hole region $\mathscr{B}_c$ is small– it always contains the entirety of infinitely many asymptotically flat regions.

Should one conclude, then, that Definition 10 is incompatible with the notion of a rotating black hole? Given that the prevailing hope amongst physicists (encoded

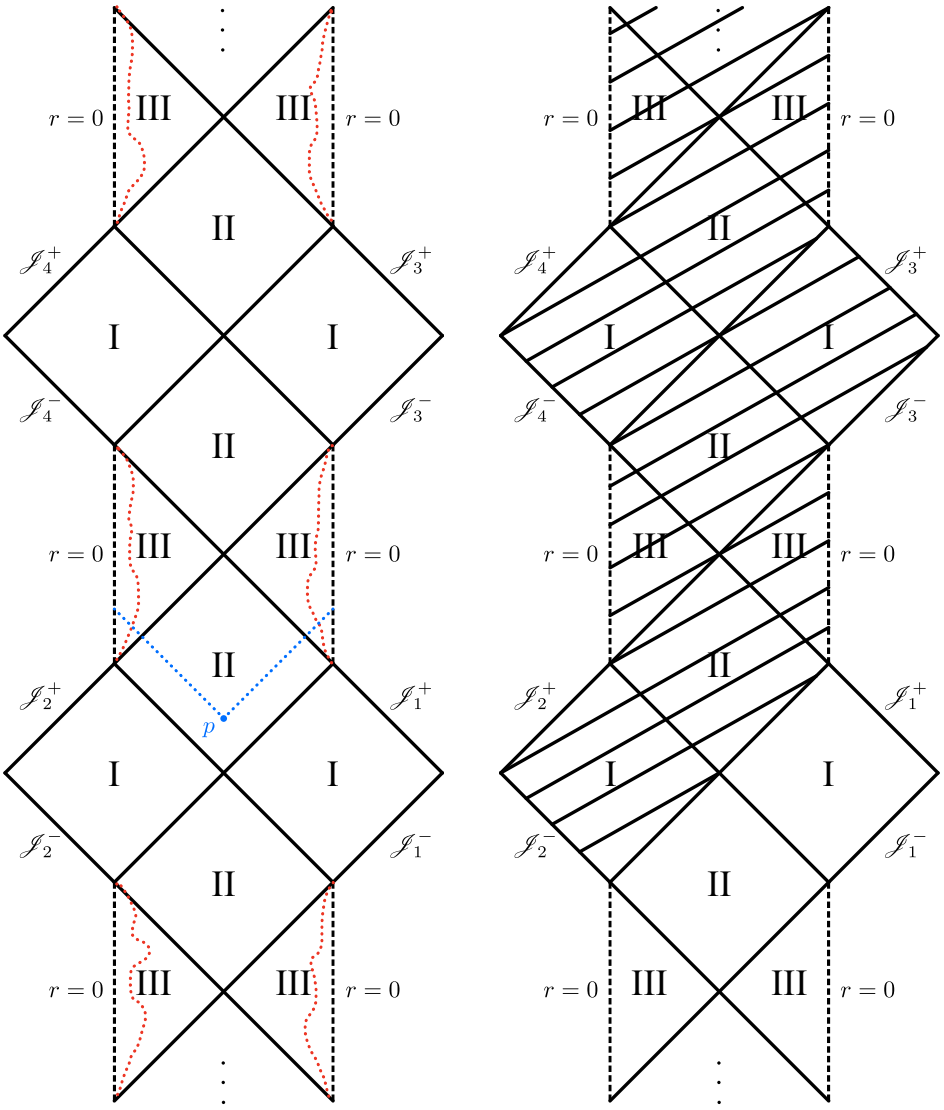

FIGURE 5.7: Identifying the black region in the Kerr spacetime. The classic black region $\mathscr{B}_c$ with respect to $\mathscr{I}_1^+$ is shown on the right.

in Penrose's Strong Cosmic Censorship Conjecture [29, 36, 94, 106]) is that the full structure of a maximal, dynamically forming black hole spacetime, rotating or no, is more similar to Schwarzschild's Penrose diagram than to Kerr's (in particular, it is expected to be globally hyperbolic), one would still expect such a spacetime to have a nonempty black region $\mathscr{B}$. That is, Definition 10 should be able to characterize generic physical rotating black hole solutions, given strong cosmic censorship, even though it deems Kerr itself to have an empty black region. This is not too unreasonable a caveat given that the classic black hole region $\mathscr{B}_c$ of Section 5.1 is typically only even defined at all under the assumption of weak cosmic censorship. In any event, this is an inescapable peculiarity of taking the intuitive description "light cannot escape" entirely seriously: there simply is no global sense in which this true for any portion of Kerr without invoking a preferred infinity, which we cannot easily do under a program that remains applicable outside of asymptotic flatness. It is up to the reader to decide whether this captures the physical phenomenon they wish to consider in a given scenario. We would argue at the very least that there exist some considerations, including the framing of weak cosmic censorship, under which Definition 10 captures precisely what one would want irrespective of the peculiarity encountered here.

□

5.3 Defining Singular Neighborhoods

5.3.1 Avoiding Pathologies

At the core of our proposed characterization of black holes is a notion of "where" the singularities in a spacetime (M, g) lie through the concept of a singular neighborhood. One means of obtaining this notion is heuristically looking at the structure of ∂M in particularly attractive coordinate charts or embeddings (such as Penrose diagrams), wherein it seems intuitively clear where the manifest singularities are, and adopting

the embedding's or coordinates' topology. This is precisely the program carried out in the examples of Section 5.2, and it is similar to the route one typically takes in trying to employ the intuition of the standard black hole paradigm where it isn't technically applicable (e.g. discussing the past of infinity in deSitter-Schwarzschild). This may well give reasonable, desirable, and intuitive notions in many examples, and as such is not a universally incorrect approach. However, it is neither entirely generalizable nor invariant: the conclusions one makes may well depend upon the choice of coordinates or embedding, and it is a rather tricky prospect to characterize a "correct" or "optimal" choice, and even trickier to establish one's existence.

We take the perspective that singular neighborhoods should entirely be properties of (M, g), with no extrinsic structure or constricting hypotheses required to make sense of them (and hence of black holes). The boundary constructions mentioned at the beginning of this chapter seemingly provide a natural means of defining singular neighborhoods in accordance with this perspective: they require nothing more than M being a smooth manifold and g being a sufficiently smooth Lorentzian metric (often even time orientability isn't strictly required). Singularities are subtle, however: a critical topological shortcoming that arises in many such constructions, including the b- and g-boundaries, is that the singularities in $\partial M \subset \overline{M}$ are not guaranteed to be Hausdorff-separated from points in M. Indeed, Geroch, Can-Bin, and Wald [52] demonstrated that any construction sufficiently similar to the g-boundary will have this pathology, and it seems this result was one of the primary reasons that the community's interest in boundary constructions waned after the 1970's.

In the language of singular neighborhoods, this pathology means that an intuitive stipulation we put forward previously, that singular neighborhoods include all complements of compact sets, is not necessarily satisfied when using such constructions to identify singular neighborhoods. This is undesirable, as compact sets are in a sense bounded in the interior of M, and so should not be arbitrarily close to the

boundary ∂M. While the construction of black holes put forward in the previous section could, in principle, be carried out in spite of this pathology, we consider it to violate the physical intuition underlying the definition. For this reason, we consider the later-developed a-boundary, or abstract boundary, to be the most natural means of identifying singular neighborhoods, as its topological structure manifestly separates interior and boundary points.

5.3.2 Singular Neighborhoods Via the Abstract Boundary

In this and the following section, we draw heavily upon abstract boundary concepts, terminology, and notation to provide and work with a definite meaning of singular neighborhoods. A streamlined review of all we will need, as well as references to various works on the abstract boundary, is provided in Appendix A. Assuming familiarity with that material, we proceed to denote by $\mathcal{S}_p(M) \subset \mathcal{B}(M) \subset \overline{M}$ the collection of pure singularity abstract boundary points, and by $\mathcal{I}(M) \subset \mathcal{B}(M) \subset \overline{M}$ the collection of abstract boundary points at infinity.

Definition 11. *An open set $U \subset M$ will be called a singular neighborhood provided that every pure singularity abstract boundary point is strongly attached to U. Equivalently, provided that $U = \widetilde{U} \cap M$ for some neighborhood $\widetilde{U} \subset \overline{M}$ of $\mathcal{S}_p(M)$ in $\overline{M}$ endowed with $\mathcal{T}_{sap}(M)$.*

Given the results and discussion of the appendix, this captures precisely that U "surrounds" any pure singularities as much as possible from within M, and so agrees with the intuition and heuristic descriptions put forward in Section 5.2. Indeed, the following corollary to Proposition 17 demonstrates how this definition yields the picture of singular neighborhoods presented in the figures of that section:

Corollary 6. *If $U \subset M$ is a singular neighborhood and $\phi : M \to \widehat{M}$ is an envelopment, then there exists an open set $\widehat{U} \subset \widehat{M}$ containing all pure singularity boundary*

points in $\widehat{M}$ *with* $\widehat{U} \cap \phi(M) = \phi(U)$.

Proof. By definition, we may find an open neighborhood $\widetilde{U} \subset \overline{M}$ of $\mathcal{S}_p(M)$ satisfying $\widetilde{U} \cap M = U$. By Proposition 17, $V := \overline{\pi}^{-1}(\widetilde{U})$ is open in $\overline{\phi(M)}$, so there is an open set $\widehat{U} \subset \widehat{M}$ satisfying $\widehat{U} \cap \overline{\phi(M)} = V$. By definition of $\overline{\pi}$ on $\phi(M)$, we find

$$\widehat{U} \cap \phi(M) = V \cap \phi(M) = \phi(\widetilde{U} \cap M) = \phi(U).$$

Furthermore, $V \subset \widehat{U}$ contains $\overline{\pi}^{-1}(\mathcal{S}_p(M))$, which is precisely the collection of pure singularity boundary points in $\widehat{M}$ by definition of $\overline{\pi}$ on $\partial\phi(M)$.

□

This definition, in furnishing us with a rigorous description of the collection $\mathscr{U}$ of singular neighborhoods, now puts us in a position to prove some expected features of the other objects defined in Section 5.2, namely singularly compact sets and the black region. First, we firmly establish the "closeness" of singular neighborhoods to singularities with a straightforward topological result.

Proposition 7. *A subset* $S \subset M$ *which meets every singular neighborhood has a pure singularity as an accumulation point in* $\overline{M}$.

Proof. We prove the contrapositive. Take $S \subset M$, and suppose it has no pure singularity as an accumulation point in $\overline{M}$. Then for each $s \in \mathcal{S}_p(M) \subset \mathcal{B}(M)$, we may find an open set $\widetilde{U}_s \subset \overline{M}$ of s such that $S \cap \widetilde{U}_s$ is at most finite. Since $s \in \mathcal{B}(M)$ is T_2-separated from points in M, we may in fact take $S \cap \widetilde{U}_s$ to be empty. Then

$$\widetilde{U} := \bigcup_{s \in \mathcal{S}_p(M)} \widetilde{U}_s$$

is an open set in $\overline{M}$, from which S is disjoint, containing every pure singularity. By definition, then, $\widetilde{U} \cap M$ is a singular neighborhood which S does not meet.

□

This ensures, for example, that any non-compact subset $A \subset M$ which is singularly compact, with respect to the singular neighborhoods induced by the abstract boundary, must have a pure singularity as an accumulation point in $\overline{M}$, in alignment with expectations. Indeed, it ensures that *every* sequence in A without accumulation points in M must have a pure singularity as an accumulation point in $\overline{M}$. This makes precise our claims made in Section 5.2 regarding "nonsingular" spacetimes (in particular, that their singularly compact sets are precisely their compact sets), which we can now describe as spacetimes which admit no pure singularity abstract boundary points. When working with geodesics as the b.p.p. family of curves $\mathcal{C}$ (Definition 16), for example, this includes all geodesically complete spacetimes. We'll find further use for Proposition 7 in the subsequent section. Another result of great interest to the physical merit of the singular neighborhoods induced by the abstract boundary can currently only be put forward as a conjecture:

Conjecture 1. *Under physically relevant choices of curve families $\mathcal{C}$ on a smooth Lorentzian manifold (M, g) (e.g. geodesics with affine parameter, C^1 curves with generalized affine parameter, the causal subfamilies of either of these, etc.), points in $\mathcal{I}(M)$ and $\mathcal{S}_p(M)$ are T_2-separated from each other.*

This conjecture would ensure that abstract boundary points at infinity cannot be topologically embroiled with pure singularities– in particular, it is equivalent to the claim that there cannot exist a sequence in M which limits to both a pure singularity and a point at infinity. It is easily seen to be true that such points are T_1-separated, as T_1-separation is equivalent to neither of the points covering the other, but T_2-separation, while a reasonable conjecture, is not so immediately obvious. A proof of this conjecture would go a long way toward establishing the physical reasonability of utilizing the abstract boundary to identify singular neighborhoods, as seen in its corollary, Conjecture 2, below.

5.4 Characterizing the Black Region $\mathscr{B}$

We have now provided a precise notion of a singular neighborhood and therefore made precise the black region of Definition 10. As seen in the appendix, however, the technical machinery required to do so was rather nontrivial, and so it appears a difficult task to navigate this machinery and characterize $\mathscr{B}$ in a given example, in principle requiring understanding the structure of every possible envelopment $\phi : M \to \widehat{M}$. In practice, one should be able to carry out a process very similar to the heuristic approach taken in the examples of Section 5.2. In this section, we both establish and conjecture results to this effect, as well as results affirming the intuitive features expected of $\mathscr{B}$. We begin by recalling some useful, well-known lemmas.

Lemma 8. *The following conditions on a curve $\sigma : [0, b) \to M$, with $0 < b \leqslant \infty$, are equivalent:*

(i) For any compact set $K \subset M$, σ eventually leaves and never returns to K.

(ii) For any sequence $(t_n)_{n=1}^{\infty} \subset [0, b)$ satisfying $t_n \to b$, $(\sigma(t_n))$ does not converge in M.

Proof. For $(i) \implies (ii)$, suppose $\sigma(t_n) \to p \in M$. Then given any compact neighborhood $K \subset M$ of p, $\sigma(t_n)$ is eventually always in K, in contradiction to (i).

For $(ii) \implies (i)$, note that if (i) does not hold, then there must be a compact $K \subset M$ and a sequence $(t_n)_{n=1}^{\infty} \subset [0, b)$ with $t_n \to b$ such that $\sigma(t_n) \in K$, so that $(\sigma(t_n))$ must have a convergent subsequence, meaning (ii) does not hold.

□

Recall that the converse to condition (i) above is commonly referred to in the literature by saying σ is *partially imprisoned* in some compact $K \subset M$. A more restrictive condition is that σ be *totally imprisoned* in some compact $K \subset M$, meaning σ eventually enters and remains inside K. Any curve satisfying the above conditions,

then, cannot be even partially imprisoned. When M is strongly causal, the following ensures that we may utilize these properties for any inextendible causal curve:

Lemma 9. *Let $\sigma : [0, b) \to M$ be an inextendible causal curve. If strong causality is not violated on the closure of the image of σ, then σ has property (ii) (and hence (i)) in the preceding lemma. That is, σ cannot be partially imprisoned.*

Proof. Suppose there exists a sequence $(t_n)_{n=1}^{\infty} \subset [0, b)$ such that $t_n \to b$ and $\sigma(t_n) \to p \in M$. Since σ is inextendible, there must exist a sequence (s_n) satisfying $s_n \to b$ such that $(\sigma(s_n))$ does not converge to p, so there is a neighborhood U of p such that $(\sigma(s_n))$ has a subsequence not meeting U. By passing to appropriate monotonic subsequences, we may assume $t_n \leqslant s_n \leqslant t_{n+1}$ and $(\sigma(s_n))$ does not meet U. For any neighborhood $V \subset U$ of p, however, $(\sigma(t_n))$ eventually enters and remains inside V, so that the causal curves $\sigma|_{[t_n, t_{n+1}]}$ violate strong causality at p. This is a contradiction since p is in the closure of the image of σ.

□

A similar result can be obtained by weakening both the hypothesis and conclusion. We refer the reader to [60] for proof:

Lemma 10 ([60], Proposition 6.4.8)**.** *Let $\sigma : [0, b) \to M$ be an inextendible causal curve. If either the past or future distinguishing conditions holds on a compact set $K \subset M$, then σ cannot be totally imprisoned in K.*

□

In particular, these results affirm that standard compactness has no hope of describing causally well-behaved black holes, as they indicate that any nonempty future set cannot be compact in a causally distinguishing spacetime (since any such set contains inextendible causal curves). With them, however, we may establish the singular nature of $\mathscr{B}$ under such causality conditions.

Proposition 11. *If an inextendible, future-directed causal curve $\sigma : [0, b) \to M$ which meets the black region $\mathscr{B}$ is not totally imprisoned, then σ has a pure singularity accumulation point in $\overline{M}$.*

Proof. Let $\sigma : [0, b) \to M$ be an inextendible, future-directed causal curve meeting $\mathscr{B}$, i.e. meeting some singularly compact set $A \subset M$ satisfying $J^+(A) = A$ (so σ remains in A), and suppose σ is not totally imprisoned. For each singular neighborhood $U \in \mathscr{U}$, $A \backslash U$ is compact. Since σ is not totally imprisoned, then, there exists a sequence $(t_n) \subset [0, b)$ with $t_n \to b$ such that $\sigma(t_n)$ is not in $A \backslash U$, so we must have $\sigma(t_n) \in U$. In particular, σ meets every singular neighborhood. Taking S to be the image of σ in Proposition 7 completes the proof.

□

Strengthening the hypothesis by switching total imprisonment to partial imprisonment, we can get a more compelling result:

Proposition 12. *If an inextendible, future-directed causal curve $\sigma : [0, b) \to M$ which meets the black region $\mathscr{B}$ is not partially imprisoned, then every sequence along the end of σ has a pure singularity accumulation point in $\overline{M}$.*

Proof. Let $\sigma : [0, b) \to M$ be an inextendible, future-directed causal curve meeting $\mathscr{B}$, i.e. meeting some singularly compact set $A \subset M$ satisfying $J^+(A) = A$ (so σ remains in A), and suppose σ is not partially imprisoned. For each singular neighborhood $U \in \mathscr{U}$, $A \backslash U$ is compact. Since σ is not partially imprisoned, then, there is a $t_U \in [0, b)$ such that $\sigma(t) \notin A \backslash U$, and hence $\sigma(t) \in U$, for $t > t_U$. Thus, given any sequence $(t_n)_{n=1}^{\infty} \subset [0, b)$ with $t_n \to b$, $(\sigma(t_n))$ meets every singular neighborhood $U \in \mathscr{U}$, and the result follows from Proposition 7.

□

Combining either of these with the previous lemmas, then, yields the desired singular nature of $\mathscr{B}$. We only state the strongly causal conclusion in the following,

but it should be clear that the same statement is true under the distinguishing condition upon the removal of the phrase "sequence along the end of an" from the theorem.

Theorem 13. *Let (M, g) be strongly causal. If $p \in \mathscr{B}$, then every sequence along the end of an inextendible, future-directed causal curve through p has a pure singularity as an accumulation point in $\overline{M}$.*

□

It is in this sense that the black region $\mathscr{B}$ satisfies the heuristic description of being in the "past Cauchy development of the singularities": any future-directed causal curve emanating from a point in the black region must approach a pure singularity in the abstract boundary, under reasonable causality restrictions. It should be expected, of course, that such conditions are needed, as causality conditions are what put structure on objects like $J^+(p)$ for some $p \in M$, and $\mathscr{B}$ is built using such causal objects. The converse to this theorem is not strictly true– see Figure 5.8.

A partial converse to the heuristic, that any point in the "past Cauchy development of the singularities" should indeed lie in the black region, can be obtained, however. We provide this in the following two theorems. The first is a conceptual result to this effect at the level of the full structure of the abstract boundary, which both supplies this heuristic and links the present notion of a black hole to the standard perspective.

Theorem 14. *Suppose (M, g) is globally hyperbolic, take the b.p.p. family of curves $\mathcal{C}$ to be C^1 curves with generalized affine parameter, and take $p \in M$. If there are no points at infinity strongly attached to $I^+(p)$, then $I^+(p) \subset \mathscr{B}$ (and hence $p \in \overline{\mathscr{B}}$).*

Proof. Take $\Sigma \subset M$ to be a Cauchy hypersurface for M. It is a classic result originally due to Geroch [51] that there exists a homeomorphism, strengthened to a

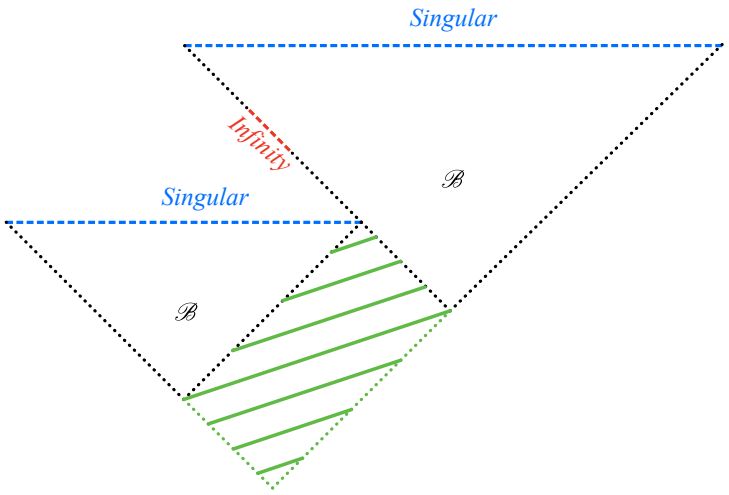

FIGURE 5.8: A schematic counterexample to the converse of Theorem 13. The dashed lines are excised from the spacetime, which has the causal structure of 2D Minkowski space. Under the envelopment of the figure, the excised points are taken to be pure singularities (blue dashed lines) and points at infinity (red dashed line). The black dotted lines are null curves, some of which comprise the boundary of $\mathscr{B}$. The green shaded region is the set of points which satisfy the consequent of Theorem 13 but not the antecedent. These points are not in $\mathscr{B}$ due to the presence of the points at infinity, but every inextendible, future-directed causal curve through them must approach a pure singularity. This example is not explicit, though we believe one can construct such a spacetime by excising the dashed lines from Minkowski space and introducing a conformal factor which appropriately diverges to 0 or ∞ on them.

diffeomorphism by Bernal and Sanchez [14], $\phi : M \to \mathbb{R} \times \Sigma$ satisfying

(i) $\Sigma_t := \phi^{-1}(\{t\} \times \Sigma)$ is a Cauchy hypersurface in M for each $t \in \mathbb{R}$.

(ii) For each $x \in \Sigma$, $t > s \implies \phi^{-1}(t, x) \in I^+(\phi^{-1}(s, x))$.

(iii) The "time function" $\tau : M \to \mathbb{R}$, defined as $\tau := \pi_1 \circ \phi$ with $\pi_1 : \mathbb{R} \times \Sigma \to \mathbb{R}$ the projection onto the first coordinate, is strictly increasing on future-directed causal curves.

Denoting for $t \in \mathbb{R}$

$$\Sigma_t^- := D^-(\Sigma_t) = \bigcup_{s \leqslant t} \Sigma_s,$$

it is also a classic result that $J^+(p) \cap \Sigma_t^-$ is compact ([101], Lemma 14.40).

Similarly to τ, define the smooth map $\rho : M \to \Sigma$ by $\rho := \pi_2 \circ \phi$, where π_2 is the projection onto the second coordinate, and set

$$C_t := \rho\left(I^+(p) \cap \Sigma_t\right) \subset \Sigma.$$

By property (ii), $s < t \implies C_s \subset C_t$, which together with property (iii) implies that C_t is path-connected. Indeed, if $q \in I^+(p)$, there exists a timelike curve $\sigma : [0,1] \to M$ from $\sigma(0) = p$ to $\sigma(1) = q$, and $\rho \circ \sigma$ is a continuous path from $\rho(p)$ to $\rho(q)$ in $C_{\tau(q)}$, since $\rho \circ \sigma(s) \in C_{\tau(\sigma(s))} \subset C_{\tau(q)}$ for each $s \in (0,1]$. This shows every point in C_t, for $t > \tau(p)$, can be connected via a path in C_t to $\rho(p)$. For $t \leqslant \tau(p)$, C_t is empty.

Consider a sequence $(p_n)_{n=1}^{\infty} \subset I^+(p)$ with no accumulation points in M. Since $J^+(p) \cap \Sigma_t^-$ is compact, for each $t \in \mathbb{R}$ we must eventually have $\tau(p_n) > t$. Hence we may assume, perhaps upon passing to a subsequence, that $t_n := \tau(p_n)$ is strictly increasing with $t_n \to \infty$. We can now construct a curve γ from p_1 to p_2 by using condition (ii) to follow a future-directed timelike curve from $p_1 \in \Sigma_{t_1}$ to $\phi^{-1}(t_2, \rho(p_1)) \in \Sigma_{t_2}$, and then use the path-connectedness of C_{t_2} to travel within $I^+(p) \cap \Sigma_{t_2}$ to p_2. Iterating this procedure, we extend γ to a curve through the sequence (p_n), entirely contained in $I^+(p)$, on which $\tau \to \infty$ and is non-decreasing. The last condition ensures that γ has no accumulation points in M, as any sequence along the end of γ has $\tau \to \infty$. Smoothing out γ under these conditions, the endpoint theorem for curves [126][1] ensures γ, and hence (p_n), limits to an approachable abstract boundary point $[p_\infty] \in \mathcal{B}(M)$ which is strongly attached to $I^+(p)$. By hypothesis, $[p_\infty]$ can neither be nor cover a point at infinity, and hence it must be a pure singularity. $[p_\infty]$ is then strongly attached to every singular neighborhood by definition, so (p_n) must meet every singular neighborhood.

[1] The *endpoint theorem for curves* refers to the result that any smooth, non-self-intersecting curve in M without accumulation points limits to some abstract boundary point in $\overline{M}$. It is a trivial consequence of the proof that if the curve can be constrained to lie entirely in some open set, the provided endpoint can be chosen to be strongly attached to that open set.

Now, take $p_0 \in I^+(p)$, and fix a singular neighborhood $U \subset M$. The above argument demonstrates that any sequence $(p_n)_{n=1}^{\infty} \subset I^+(p)\backslash U$ must have an accumulation point in M, and hence that $\overline{I^+(p)\backslash U}$ is compact. Since M is globally hyperbolic, it is causally simple, so that $J^+(p_0)$ is closed. Hence $J^+(p_0)\backslash U \subset I^+(p)\backslash U$ is a closed subset of $\overline{I^+(p)\backslash U}$, so it is compact. This demonstrates $J^+(p_0) = \overline{J^+(p_0)}$ is singularly compact, and therefore that $p_0 \in \mathscr{B}$ by Lemma 5.

□

Noting that $I^+(p)$ having no points at infinity strongly attached is equivalent to saying that every abstract boundary point strongly attached to $I^+(p)$ is a pure singularity, as well as that the endpoint theorem for curves ensures that an inextendible causal curve contained in an open set $V \subset M$, with M strongly causal, limits to an abstract boundary point in $\overline{M}$ strongly attached to V, the hypotheses of this theorem ensure that any inextendible causal curve starting at $p_0 \in I^+(p)$ must limit to a pure singularity, i.e. that p_0 is in the "past Cauchy development of the singularities". Under this hypothesis, the theorem establishes that $p_0 \in \mathscr{B}$, as claimed in the heuristic.

Moreover, Theorem 14 indicates that $\mathscr{B}$, when formalized via the abstract boundary, is closely related to the complement of the past of infinity, the more standard means of defining a black hole. This is a testament both to the reasonable nature of $\mathscr{B}$ and to the naturalness of the structure of the abstract boundary. Recall that the original definition of $\mathscr{B}$, Definition 10, makes no mention of any concept of infinity, nor even requires that one exists– the abstract boundary formalism furnished both the concept of infinity in a general spacetime and tied that concept to the black region. A priori, there was no reason that such a link must hold.

The weakness of Theorem 14 is twofold, however. The first weakness is the somewhat restrictive causality condition it requires. We expect a similar result is true in, say, a causally simple setting, but a proof remains elusive. The second is

in the difficulty of applying Theorem 14 in practice, as making a conclusion about all possible abstract boundary points strongly attached to an open set seems rather difficult to do. Fortunately, the next result (though it says much less about how $\mathscr{B}$ is related to the full abstract boundary structure) ameliorates both of these concerns by providing a more practical means of identifying points in $\mathscr{B}$, free of causality constraints or a restriction on $\mathcal{C}$.

Theorem 15. *If there exists an envelopment $\phi : M \to \widehat{M}$ under which $\overline{\phi(I^+(p))} = \overline{\phi\left(\overline{J^+(p)}\right)}$ is compact and every boundary point in $\widehat{M}$ attached to $I^+(p)$ is a pure singularity, then $p \in \mathscr{B}$.*

We provide two proofs of distinctly different flavors, one sequential and one more directly wielding the topological machinery of the abstract boundary.

Proof 1. Let $U \subset M$ be a singular neighborhood, and consider a sequence $(p_n)_{n=1}^{\infty} \subset \overline{J^+(p)}\backslash U$. Since $\overline{\phi(I^+(p))}$ is compact, $(\phi(p_n))$ has an accumulation point $q \in \overline{\phi(I^+(p))} \subset \overline{\phi(M)}$. Suppose that $q \in \partial\phi(M)$. Being attached to $I^+(p)$, q would then be a pure singularity by hypothesis, and by Definition 11, q would be strongly attached to U. This is a contradiction, however, since (p_n) does not meet U. Thus we must have that $q \in \phi(M)$, and so $\phi^{-1}(q)$ is an accumulation point of (p_n) in M. Since $\overline{J^+(p)}\backslash U$ is closed, this shows $\overline{J^+(p)}\backslash U$ is compact.

This demonstrates that $\overline{J^+(p)}$ is singularly compact, and therefore that $p \in \mathscr{B}$ by Lemma 5.

□

Proof 2. Let $U \subset M$ be a singular neighborhood. By Corollary 6, there exists an open set $\widehat{U} \subset \widehat{M}$ containing all pure singularities in $\widehat{M}$ and satisfying $\widehat{U} \cap \phi(M) = \phi(U)$. Since every boundary point in $\widehat{M}$ attached to $I^+(p)$ is a pure singularity, we

have

$$\overline{\phi(I^+(p))}\backslash\phi\left(\overline{J^+(p)}\right) = \overline{\phi(I^+(p))} \cap \partial\phi(M) \subset \widehat{U}.$$

Hence we find

$$\overline{\phi(I^+(p))}\backslash\widehat{U} = \phi\left(\overline{J^+(p)}\right)\backslash\phi(U) = \phi\left(\overline{J^+(p)}\backslash U\right).$$

Since $\widehat{U}$ is open and $\overline{\phi(I^+(p))}$ is compact by hypothesis, $\overline{\phi(I^+(p))}\backslash\widehat{U}$ is compact, and thus so is $\overline{J^+(p)}\backslash U$ by the above equality. This demonstrates $\overline{J^+(p)}$ is singularly compact.

□

This theorem provides the formal justification, when defining things precisely via the abstract boundary formalism, for taking the heuristic approach presented in the examples of Section 5.2 to identifying points in $\mathscr{B}$. It is by far the most useful result to identifying points in $\mathscr{B}$ in practice, in that it reduces the problem of investigating all possible abstract boundary points (a seemingly intractable task) to investigating only those boundary points arising in a particular envelopment.

The other side of identifying the subset $\mathscr{B}$, namely of identifying those points in M which are *not* in $\mathscr{B}$, is more subtle. An approach is provided, in principle, by Theorem 13: if one could find a sequence along an inextendible, future-directed causal curve through a point $p \in M$ which does *not* have a pure singularity as a limit point, then $p \notin \mathscr{B}$. It is difficult to establish in practice, however, that a sequence in M does not have a pure singularity as a limit point under *any* envelopment without the topological constraint of Conjecture 1. To formalize this, we state the following conjecture which follows readily from Conjecture 1, and which would more firmly tie the present notion of a black hole to the standard perspective.

Conjecture 2. *If $p \in M$ has a point at infinity attached to $I^+(p)$, then $p \notin \mathscr{B}$.*

Proof (provided Conjecture 1). Suppose $p \in \mathscr{B}$ has a point at infinity $[q] \in \overline{M}$ attached to $I^+(p)$. Then by definition there is a sequence of points $(p_n) \subset I^+(p) \subset \overline{J^+(p)}$ such that $p_n \to [q]$ in $\overline{M}$. Since $[q]$ is T_2-separated from points in M, (p_n) has no accumulation points in M, so it enters and remains inside of every singular neighborhood by the singular compactness of $\overline{J^+(p)}$. By Proposition 7, (p_n) has a pure singularity $[s] \in \overline{M}$ as an accumulation point, and hence passing to an appropriate subsequence yields both $p_n \to [q]$ and $p_n \to [s]$, in contradiction to Conjecture 1.

□

5.5 Weak Cosmic Censorship

Having formally defined a general notion of black hole applicable to any maximal spacetime, we turn again to the important outstanding question of weak cosmic censorship. The incompleteness theorems of Hawking and Penrose [58, 104], early investigations of dynamical collapse by Oppenheider and Schneider [102], and Schoen and Yau's demonstration of the dynamical nature of trapped surfaces [123] have convinced physicists that singularities are a generic feature of general relativity which must be wrangled with. This is not so surprising, as it was known and expected that general relativity must give way to quantum corrections in extreme situations. On its face, however, it may spell disaster for the predictive power of the theory in that singularities, by definition, cannot be evolved through. Points to the future of a singularity would necessarily be causally influenced in a manner that could not be modeled within the context of general relativity, and hence the structure and dynamics of these points could not be predicted by it. The generic emergence of singularities essentially means, then, that classical physics is not even nominally self-contained.

This situation might not be so disastrous, however, if the futures of singularities could be said to be "small" in some sense. That is, this is not so concerning for

the conceptual value of the theory so long as things are solvable sufficiently far from singularities. The Weak Cosmic Censorship Conjecture is the hope that this is the case, with its heuristic content being that *singularities in general relativity are generically hidden behind black holes.* This would mean that, at least generically, the futures of singularities are contained in the "small" interiors of black holes, and hence they do not pose a significant obstruction to general relativity's self-consistently modeling the universe at large. Singularities in violation are dubbed *naked* [2].

5.5.1 Posing Weak Censorship

Historically, this conjecture has been formulated as an initial value problem, a statement about spacetimes which evolve from initial data stipulated on a sufficiently nice (e.g. complete, asymptotically flat) Riemannian 3-manifold. This was classically formalized [60, 137] by positing that such a spacetime, the maximal Cauchy development of the initial data in question, generically turns out to be future asymptotically predictable (recall Definition 8). More modern formulations [29, 116] are framed in terms of the "completeness of future null infinity", which, while still requiring the notion of asymptotic flatness, avoids dependence on an explicit conformal embedding. The seminal results in this domain are Christodoulou's demonstration for a real scalar field matter source in spherical symmetry [28] and Christodoulou and Kleinerman's demonstration for vacuum perturbations of Minkowski initial data [30]. Review and discussion of this topic can be found in [29, 74, 130, 136].

A significant challenge to achieving a comprehensive result to this effect is the differing characteristics of various matter fields when coupled to the Einstein equation, hence results are generally restricted either to vacuum or particular matter fields. Moreover, it is not entirely clear what the "correct" complete set of physical

[2] In the previous chapter, we distinguished between *locally* and *globally* naked singularities, respectively being violations of strong and weak cosmic censorship. In this chapter, we have described and exclusively refer to globally naked singularities.

matter fields is, so the goalpost for a physically compelling resolution is difficult to set. Moreover, the typical formalization requires asymptotic flatness, though the physical significance of the conjecture still persists outside of this context. While it was originally conceived with isolated gravitational collapse in mind, the spirit of weak cosmic censorship is not only a physical problem for isolated systems – this is manifest in the non-asymptotically flat cosmological structure of the universe at large, as well as in the consideration of singularities that may have arisen very early in the universe's history, in the cosmic primordial soup.

We would like, then, to put forward a global formulation of this conjecture which depends critically neither on the particular choice of matter fields (perhaps only, say, energy condition restrictions on curvature) nor the assumption of asymptotic flatness to make sense. The following is a somewhat broad conjecture in this direction:

Conjecture 3. *(Global Weak Cosmic Censorship) In a generic, maximal, physically admissible spacetime* (M, g) *which admits a complete space-like hypersurface* Σ*, there exists a singular neighborhood* $U \in \mathscr{U}$ *such that* $U \cap D^+(\Sigma) \subset \mathscr{B}$*.*

Setting aside the hypotheses, this simply states that if a spacetime contains some "full" instant of time Σ at which there are no singularities, then no naked singularities will develop in that instant's future domain of dependence $D^+(\Sigma)$. Indeed, the condition's negation is that $D^+(\Sigma)\backslash\mathscr{B}$ meets every singular neighborhood, which should mean (by Proposition 7 when formalizing via the abstract boundary) that $D^+(\Sigma)\backslash\mathscr{B}$ is arbitrarily close to a singularity– that is, a singularity outside of the black region $\mathscr{B}$ develops within $D^+(\Sigma)$, precisely what weak cosmic censorship should avoid. This formulation in terms of some complete Σ is necessary to rule out spacetimes which "always" have a singularity in some sense, such as negative mass Schwarzschild, while still allowing for the initial singularity in cosmological models. We've dubbed our description "global" due to its working in terms of the full structure of a spacetime,

rather than an initial data set to be evolved.

A helpful exercise to grok our terminology is observing that Conjecture 3 does *not* claim that singularities are generic. Recall from Section 5.2 that in a nonsingular spacetime, the collection of singular neighborhoods $\mathscr{U}$ is the entire topology (this agrees with the approach of Definition 11: if there are no pure singularity abstract boundary points, then "all" of them are strongly attached to every open set in M), and in particular includes the empty set. Explicitly, then, the stipulation of Conjecture 3 is satisfied in a nonsingular spacetime by taking U to be the empty set.

We now discuss the hypotheses. At this level of generality, the conjecture has several avenues along which it is subject to variation, as is familiar in dealings of cosmic censorship. The first is quite standard and common among all formulations: the relevant notion of "generic" is rather flexible so long as it is reasonable. One might define a topology on the collection of appropriate spacetimes or on the space of appropriate metrics on a particular smooth manifold, and show the set of spacetimes or metrics satisfying the stated condition is open and its complement has empty interior. Alternatively, one might define a measure on the same spaces, and show that the set satisfying the condition has full measure. Some such genericness clause is necessary, as it is not difficult to write down maximal spacetimes which do not satisfy the conjectured condition while appearing physically admissible (see Chapter 4 and Figure 5.9 below).

The term "maximal" is similarly subject to interpretation, depending upon the metric regularity class of interest. The metric's being C^2, time-orientable, and satisfying a causality condition are common physical invocations, though PDE approaches to IVP formulations lend themselves to weakening the C^2 condition to only requiring that the Christoffel symbols be in L^2_{loc}. Note that maximality is required for the claim to be meaningful, as it ensures that singular neighborhoods contain all relevant data– clearly they cannot identify singular behavior which has been omitted from

the spacetime by fiat (e.g., taking M to be the region $r > m$ in the Schwarzschild spacetime). One might assume something akin to strong cosmic censorship as a hypothesis to trivialize this condition.

"Physically admissible", also a standard fudge factor in cosmic censorship, should be taken as a substitute for the ad hoc restriction that one's matter fields admit a nice PDE description, with the hope being that, say, the dominant energy condition is all that is needed. It may well be that the conjecture cannot be proven true without taking this phrase to mean that the spacetime is a solution of Einstein's equation with a particular set of matter fields, but since it is unclear what the full suite of physical matter fields should be, it is desirable to have a formulation of the conjecture which at least allows for the possibility that this is not required.

Finally, as we have discussed at length, one might take several different approaches to rigorously defining singular neighborhoods, so this is another choice one can make in rendering Conjecture 3 a specific claim subject to proof. We have focused on one such framework through the abstract boundary, though there may well exist reasonable alternatives. Even within the context of the abstract boundary, one has some freedom in choosing the physically relevant b.p.p. family of curves one uses to identify singularities. While C^1 curves of generalized affine parameter are the most comprehensive and compelling in view of Geroch's work [50][3], geodesics are a natural starting point.

5.5.2 Revisiting Vaidya

We conclude this section by recalling the conclusions of Chapter 4, in particular Proposition 4*(i)*, to demonstrate both the physical need for a formulation of weak cosmic censorship in the vein of Conjecture 3 as well as the challenge one must

[3] Geroch constructed here a geodesically complete spacetime which contains an inextendible timelike curve of finite proper time and bounded acceleration, demonstrating that geodesic completeness is not entirely sufficient to rule out physical pathologies associated to singularities.

overcome in hoping to prove such a result. The Vaidya spacetimes exhibiting naked singularities are clearly in violation of the physical spirit of weak cosmic censorship, genericness aside. The standard IVP formulation, however, cannot consider them as such; they are outside of its scope because they do not utilize a specific matter field characterized by a PDE coupled to the Einstein equation. That is: given a complete, asymptotically flat initial data set obtained from a complete space-like hypersurface Σ within one of these spacetimes, it is not clear what PDE one might use to evolve from it $D(\Sigma)$, so this evolution cannot be readily posed as an IVP. As they satisfy the dominant energy condition, however, one might argue that the Vaidya spacetimes have just as much a claim to exhibiting physical behavior as do toy matter models that happen to arise from a Lagrangian. In this sense, it is arguable that the IVP formulation of weak cosmic censorship, while certainly deeply significant, is not entirely sufficient to capture the physical spirit of the conjecture. We've therefore put forward Conjecture 3 as an attempt at doing so.

It behooves us, then, to comment on how the naked singularities of Vaidya spacetimes indeed violate the condition of Conjecture 3. A complete space-like hypersurface Σ passes through the regular $r = 0$ axis in the Minkowski region, say at some $v_\Sigma < 0$ (see Figure 5.9). Each point $(v, 0)$ satisfying $v_\Sigma < v < 0$ is in $D^+(\Sigma)$, but the entire regular $r = 0,\ v < 0$ axis is in the past of any of the escaping null geodesics found in Proposition 4, so these points are not in $\mathscr{B}$ (either heuristically in the spirit of Section 5.2 or within the formal structure of the abstract boundary, given Conjecture 2) whenever these escaping null geodesics exist. In this scenario, then, any singular neighborhood, which includes an open set around $(0, 0)$ in the topology of the envelopment into $\mathbb{R} \times \mathbb{R}^3$ manifest in the Vaidya coordinates (by Corollary 6, within the abstract boundary), must intersect $D^+(\Sigma)\backslash\mathscr{B}$, contrary to the stipulation of Conjecture 3. In contrast, when no null geodesics escape from $(0, 0)$ to $r \to \infty$, one can find a singular neighborhood entirely contained in $\mathscr{B}$. Importantly, these

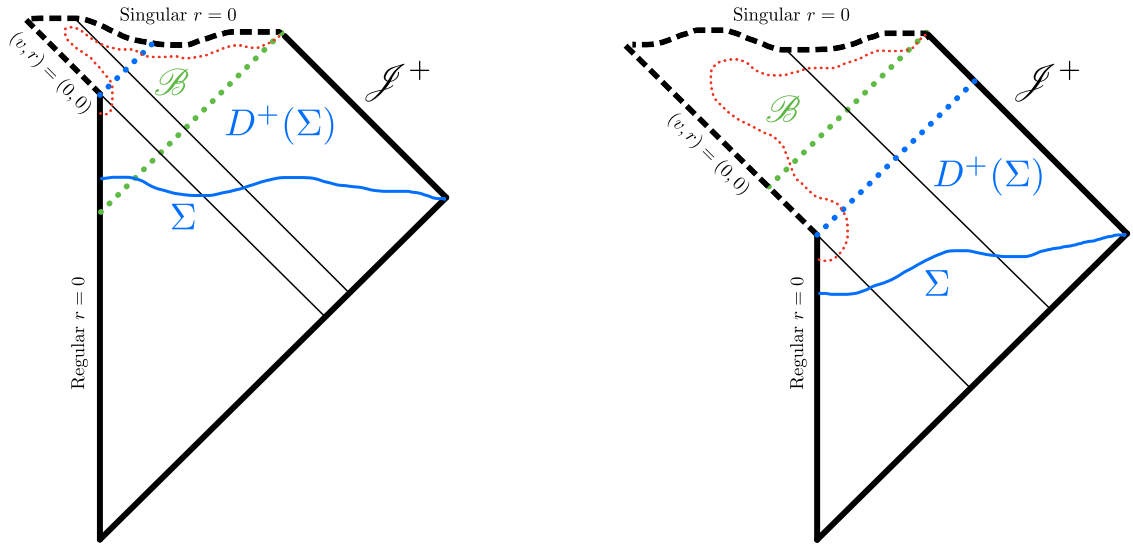

FIGURE 5.9: The Penrose diagrams of Figure 4.1, with a complete space-like hypersurface Σ added in each (in blue) to demonstrate the condition of Conjecture 3. $D^+(\Sigma)$ is bounded by Σ below and the blue dotted lines above. The red dotted curves mark the boundaries of typical singular neighborhoods. In the left case (with no naked singularitiy), the singular neighborhood shown has the property that its intersection with $D^+(\Sigma)$ is contained in $\mathscr{B}$; in the right case, any singular neighborhood intersects $D^+(\Sigma)\backslash\mathscr{B}$, indicating singular behavior outside of $\mathscr{B}$.

conclusions carry over to non-asymptotically flat case, when $\lim_{v\to\infty} m(v) = \infty$.

5.6 Conclusion

Through the definitions of Section 5.2, we have provided a novel program for formalizing the notion of a black hole in an arbitrary maximal spacetime (M, g), with no extrinsic structure or constricting hypotheses required, in a sense dual to the standard program of identifying that subset of M which cannot be seen from infinity. We feel that this new program captures well important intuitive features a classical black hole should exhibit and nicely complements the existing treatments which are either much more limited in scope or hone in on other features of interest. Indeed, having provided one means of following through this program in a complete fashion, utilizing the tool of the abstract boundary to identify singular neighborhoods

in Definition 11, we have been able to arrive at several results pointing towards the naturalness of these definitions.

As the abstract boundary is as unwieldy a tool as it is powerful, however, questions still remain. Perhaps the largest immediately outstanding question, Conjecture 1, probes the physical reasonability of utilizing the abstract boundary formalism for the purpose of identifying black holes by association with its immediate corollary in Conjecture 2. Other questions surrounding the relationship of the proposed notion of black holes to other, established intuitions include the link between $\mathscr{B}$ and trapped surfaces. While Penrose's Incompleteness Theorem ensures that trapped surfaces indicate null geodesic incompleteness, evidently the incompleteness need not be so severe as to yield a nonempty $\mathscr{B}$ in a maximal extension (as seen in Kerr). Would some natural additional hypotheses, such as the spacetime's remaining globally hyperbolic under maximal extension (i.e. strong cosmic censorship), lead trapped surfaces to demand a nonempty $\mathscr{B}$?

Beyond the philosophical appeal of providing a completely general characterization of the important physical concept of a global black hole unburdened by the constraint of asymptotic flatness, perhaps the most significant application of the program lies in its yielding a means to put forward a more comprehensive formulation of the Weak Cosmic Censorship Conjecture in Conjecture 3. As seen in our explorations of Vaidya spacetimes, this formulation is able to consider physically objectionable phenomena surrounding naked singularities that current IVP formulations cannot, both in and out of the asymptotically flat context. It is our hope that this will prove useful in rigorously illuminating the physical content of the General Theory of Relativity.

6

Conclusions

In this thesis, we have investigated a few foundational issues at the heart of general relativity, centered on some of the largest outstanding questions, both theoretical and empirical, for our foremost theory of gravity. Our primary objectives have been to derive and assess some potential observable imprints of geometric instantiations of scalar field dark matter (Chapters 2 and 3) as well as to advance our understanding of the status and formal content of the weak cosmic censorship conjecture (Chapters 4 and 5), with the latter leading to the auxiliary objective of laying out an alternative rigorous characterization of classical black holes which respects their global features without the need for any constricting hypotheses on the asymptotic structure of spacetime (Chapter 5). While we have largely succeeded in fulfilling these objectives, the underlying questions of the nature of dark matter and the veracity of weak censorship are far from resolved, and each of the lines of analysis carried out herein leaves open salient questions for future work.

In Chapter 2, we invoked the next-to-simplest geometric model of SFDM (under the axioms of Bray [19]) in order to yield a nontrivial manifestation of the modified geometry in geodesics, leading to the novel prediction of a general adjustment to the gravitational redshift in this theory, equation (2.43). Noting that this may well lead to a readily apparent signal in the time evolution of the cosmological redshift of a fixed source, in the form of oscillations on terrestrial timescales, we investigated whether the recent OzDES redshift survey by the Anglo-Australian Telescope encoded any such signal. We found that any underlying time variations in redshift are masked by noise roughly at the level of one part in 10^3, tentatively constraining the parameters of the geometric theory in question in equation (2.52). At the level of scrutiny currently available in cosmological surveys, then, it would seem that there is no empirical evidence for this particular geometric model over standard general relativity.

In Chapter 3, we investigated the impact of an early era of kination, wherein SFDM dominates the universe's energy density while varying like $\rho_\phi \propto a^{-6}$, on both the primordial light element abundances emergent from BBN and the present-day observable CMB temperature anisotropies. The former analysis found that the empirically observed light element abundances require radiation domination during the era of BBN, yielding that the transition temperature T_t between kination and radiation domination must satisfy $T_t \gtrsim 5$ MeV, or equivalently that the transition scale factor a_C must satisfy $a_C \lesssim 10^{-10}$. The latter analysis found both numerically and analytically that the predicted CMB anisotropies are robust, at (presently) remotely probeable angular scales, to kination characterized by any a_C admissible under reasonable values of the mass parameter $m \gtrsim 10^{-23}$ eV given the myriad of constraints. The observable of CMB temperature anisotropies thus cannot presently distinguish between kination and radiation domination, while BBN provides a meaningful con-

straint. While SFDM remains a viable model of dark matter, then, it is apparently not able to mollify either the Hubble tension or lithium problem.

Black Holes and Weak Censorship

In Chapter 4, we scrutinized the structure of general incoming Vaidya spacetimes with zero initial mass, a collection of simplistic spherically symmetry models for the dynamical formation of a Schwarzschild black hole, in accordance with standard energy conditions, from regular initial data. The particularly simple "self-similar" subclass was known to exhibit globally naked singularities, but the full collection of Vaidya spacetimes is much richer than this. In Propositions 3 and 4, we proved results that much more broadly characterize the emergence of globally naked singularities in this context, finding that they are in fact generic (in a physically appropriate sense) within this collection of spacetimes. These results provide some of the strongest evidence available that energy conditions alone may not be sufficient to rule out the generic evolution of globally naked singularities, though it must be kept in mind that the evidence is not at all definitive due to the Vaidya spacetimes' being restricted to spherical symmetry.

Motivated by the results of Chapter 4, and in particular that the standard rigorous formulation of the weak cosmic censorship conjecture does not readily accommodate the Vaidya examples in the full breadth of their identified singular behavior (which is not restricted to the asymptotically flat context), in Chapter 5 we sought to provide a more general formulation of weak cosmic censorship. This was achieved by putting forward a novel, more general characterization of the phenomenon of black holes that captures their essential global properties without invoking asymptotic flatness, or indeed any hypotheses on spacetime beyond standard causality conditions. The construction utilizes the intimate relation between black holes and singularities, the technical details of which we explored and characterized through both examples and

formal results. The core heuristic of our construction is that a black hole might be reasonably identified as the past Cauchy development of the set of singularities.

Future Work

Many avenues for further exploration of these projects remain. First, the continuing improvement of cosmological data sets means that we may soon probe smaller amplitudes and both shorter and longer timescales in the evolution of the cosmological redshift of fixed sources, allowing for significant improvement on the constraints obtained in Chapter 2. While this remains somewhat of a long shot, the comparative simplicity of possibly obtaining a positive identification of the nature of dark matter through such analyses makes this a compelling prospect for future investigation. Moreover, the investigations of Chapter 3 did not utilize any explicitly geometric features of our model for SFDM– it was compatible with any underlying model of SFDM allowing kination initial conditions. An interesting prospect for further study is the imprint that explicitly geometric features may have. In the simplest case that geometric SFDM does not impact geodesics, there is no such imprint in primordial abundances or temperature anisotropies, but even here there would be an adjustment to more general parallel propagation which may well be relevant for polarization modes in CMB anisotropies. Beyond this, there remains a wide range of the parameter space in our present investigations into temperature anisotropies that we have not yet explored due to its either being ruled out or presently outside of empirical reach. It remains of academic interest to understand the qualitative impact of kination on CMB modes in this range, particularly as empirical techniques improve.

As weak cosmic censorship remains an open problem, of course there persist questions to think about in this domain as well. While our results have largely characterized globally naked singularities in Vaidya spacetimes, it is of great interest to

reflect further on how they might be generalized to perturbations outside of spherical symmetry, either in general or within larger classes of spacetimes that include Vaidya as a subclass. Many such larger classes exists in the literature, offering natural threads to follow. In a related vein, our broader characterization of black holes invites many new questions surrounding its relation to the more standard approach. While the largest such question is spelled out in Conjectures 1 and 2, another pressing question is how quasi-local horizons might be related to $\mathscr{B}$. In any event, it is our hope that the developments obtained herein serve to illuminate both the formal and physical content of general relativity.

Appendix A

Review of the Abstract Boundary

Here we provide a brief overview of the abstract boundary construction. For a more detailed treatment, including a litany of examples addressing the various types of boundary points, we refer the reader to the seminal paper on the subject by Scott and Szekeres [125], as well as the later paper by Barry and Scott introducing the a-boundary's strongly attached point topology [10]. We will follow the crucial features of those discussions closely, with some details and proofs omitted. Further resources on this topic include [5, 139, 140, 141], and references therein.

A.1 The Essential Definitions

The core construction of the abstract boundary may be carried out for any smooth manifold. The motivating idea that one should keep in mind is that the abstract boundary attempts to make rigorous and entirely chart-invariant the standard heuristic chart-based approach to identifying and analyzing singularities described at the beginning of sections 5.2 and 5.3. All manifolds considered herein will be smooth, connected, Hausdorff, paracompact, without boundary, and have the same dimension

n (with the case $n = 4$ being of particular interest). The idea is to consider boundary points of M under all possible smooth embeddings, quotienting by an equivalence relation capturing when boundary points arising in different embeddings are "the same" from the perspective of M. To that end, we establish some standard notation. All definitions given in this and the following subsection were originally put forward in [125].

Definition 12. *An envelopment is a smooth embedding $\phi : M \to \widehat{M}$, where M and $\widehat{M}$ are smooth manifolds of the same dimension.*

As the dimensions of M and $\widehat{M}$ are the same, $\phi(M)$ is an open submanifold of $\widehat{M}$. We will be interested in extracting information from the topological closure $\overline{\phi(M)} \subset \widehat{M}$ that is invariant under the choice of envelopment, keeping M fixed.

Definition 13. *Given an envelopment $\phi : M \to \widehat{M}$, a boundary set B is a non-empty subset of $\partial\phi(M) \subset \widehat{M}$.*

We must establish when such boundary sets are equivalent.

Definition 14. *Given boundary sets B, B' under different envelopments $\phi : M \to \widehat{M}$, $\phi' : M \to \widehat{M}'$, we say that B covers B', written $B \rhd B'$, if for every neighborhood $N \subset \widehat{M}$ of B, there exists a neighborhood $N' \subset \widehat{M}'$ of B' such that*

$$\phi \circ \phi'^{-1}(N' \cap \phi'(M)) \subset N.$$

The covering relation is depicted in Figure A.1. This relation captures when B contains all topological information present in B', as it is shown in [125] (Theorem 19) that $B \rhd B'$ iff every sequence $(p_n)_{n=1}^{\infty} \subset M$ with the property that $\phi'(p_n)$ has a limit point in B' also satisfies that $\phi(p_n)$ has a limit point in B. We now say that boundary sets B and B' are *equivalent*, written $B \sim B'$, if both $B' \rhd B$ and $B \rhd B'$.

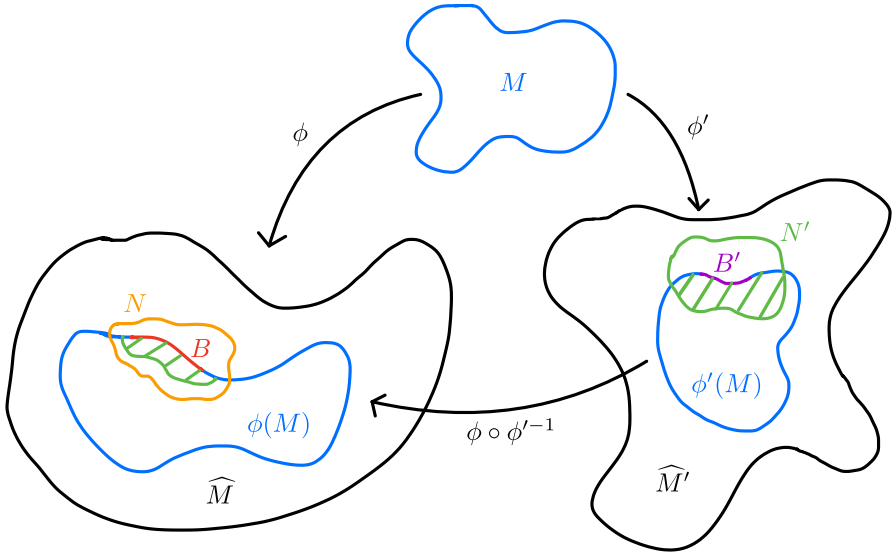

FIGURE A.1: A depiction of the covering relation of Definition 14. Given any neighborhood $N \subset \widehat{M}$ of B, we may find a neighborhood $N' \subset \widehat{M}'$ of B' "contained" within N under the mapping $\phi \circ \phi'^{-1}$ between the envelopments.

It is not difficult to see that this establishes an equivalence relation on the collection of boundary sets, finally leading to:

Definition 15. *An abstract boundary set is an equivalence class of boundary sets, denoted* $[B]$. *An abstract boundary point is an abstract boundary set which has a singleton* $\{p\}$ *as a representative boundary set in some envelopment. The abstract boundary* $\mathcal{B}(M)$ *is the set of all abstract boundary points.*

A point of notation: we often refer to an abstract boundary point $[\{p\}]$ with representative boundary point $p \in \widehat{M}$ under some envelopment by the shorthand $[p]$.

It should be kept in mind that the primary utility of the abstract boundary does not lie in its capacity to provide at once an easily-conceivable global visualization of a spacetime together with its boundary (as some previous constructions sought to do)– it is far too complicated an object for that. Instead, it provides a formalism for describing when boundary features are invariant under one's choice of envelopment, i.e. are intrinsic to M. Global visualizations of spacetimes are often achieved via

envelopments (e.g. in Penrose diagrams and global coordinate charts), and so it is of interest to know when features inferred from these visualizations are intrinsic to M. Clearly from the definitions, features which can be attributed to an abstract boundary set irrespective of its representative, as opposed to a particular boundary set, are manifestly independent of one's choice of envelopment.

Before proceeding on to the classification of abstract boundary points, we illustrate the definitions with the single example for which a complete and nontrivial (if M is compact, it admits no envelopments and hence has empty abstract boundary) description of the abstract boundary of a smooth manifold as a point-set is tractable to obtain.

Example 4. *The abstract boundary of $M = \mathbb{R}$. See Figure A.2.*

We begin by considering that any envelopment sends M either to an interval of the form $(a, b) \subset \mathbb{R}$, $-\infty \leqslant a < b \leqslant \infty$, or to a connected open strict subset of S^1, e.g. $e^{i(c,d)} \subset S^1 \subset \mathbb{C}$, $0 < d - c \leqslant 2\pi$. Boundary sets in the former type of envelopment take the form $\{a\}, \{b\}$, or $\{a, b\}$ (any of these containing an ∞ would be omitted), while boundary sets in the latter type take the form $\{e^{ic}\}, \{e^{id}\}$, or $\{e^{ic}, e^{id}\}$ (in the special case $d - c = 2\pi$, these three are all the same).

For two envelopments $\phi, \widetilde{\phi} : M \to \mathbb{R}$, the diffeomorphism $\phi \circ \widetilde{\phi}^{-1} : (\tilde{a}, \tilde{b}) \to (a, b)$ is either increasing or decreasing. It is straightforward to see that in the former case $\{a\} \sim \{\tilde{a}\}$ and $\{b\} \sim \{\tilde{b}\}$, while in the latter case $\{a\} \sim \{\tilde{b}\}$ and $\{b\} \sim \{\tilde{a}\}$, provided that the objects on both sides of the $\sim$ in each case are indeed boundary sets, i.e. are finite. Hence the envelopments into $\mathbb{R}$ give rise to exactly two abstract boundary points, which we will denote $[L]$ and $[R]$, indicating that they are the equivalence classes of the left and right endpoints of $\phi_0(M)$ under some particular chosen envelopment ϕ_0 into the unit interval $(a_0, b_0) = (0, 1)$.

It is similarly straightforward to observe that any envelopment $\psi : M \to S^1$ with

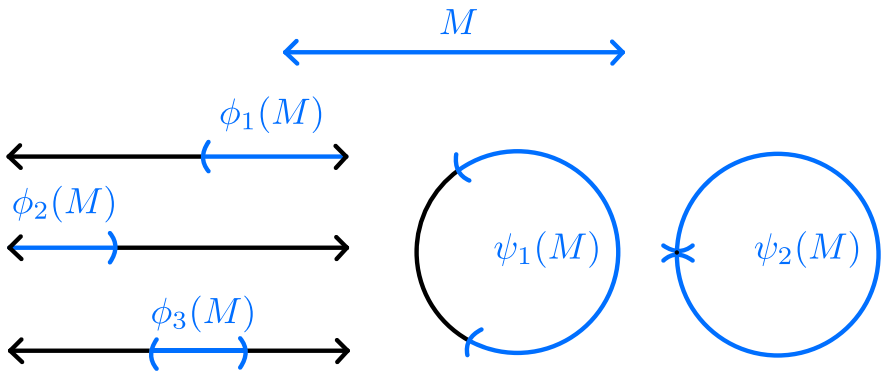

FIGURE A.2: The possible types of envelopments of $M = \mathbb{R}$ which give rise to boundary points. M and its various images are displayed in blue, while the rest of the enveloping manifold $\widehat{M}$ is shown in black in each case. Types ϕ_1 through ϕ_3 and ψ_1 give rise to representatives of $[L]$ and $[R]$, while type ψ_2 gives rise to a representative of $[LR]$.

the property that $d - c < 2\pi$ has that the pair $\{e^{ic}\}, \{e^{id}\}$ are equivalent to the pair $[L]$, $[R]$, so that these do not give any new abstract boundary points. However, when $d - c = 2\pi$, then $e^{ic} = e^{id}$, and in fact we now find that $\{e^{ic}\} \sim \{a_0, b_0\}$, so that $[\{a_0, b_0\}] = [\{e^{ic}\}]$ is also an abstract boundary point which we will denote by $[LR]$. Since we have exhausted all possible singleton boundary sets, the entirety of the abstract boundary of $M = \mathbb{R}$ is given by

$$\mathcal{B}(\mathbb{R}) = \{[L], [R], [LR]\}.$$

□

One can appreciate that the next-simplest nontrivial case of $M = \mathbb{R}^2$ is already utterly impractical to describe completely by observing that M is diffeomorphic to the unit disc D^2, and by identifying the points in the bounding S^1 according to a fundamental polygon of any closed surface S whatsoever, one can obtain an evelopment of M into S. All of these will yield different boundary topologies of $\partial\phi(M)$ and contain boundary points equivalent to various subsets of the original S^1 boundary, yielding a very large collection of distinct abstract boundary points. Even when working only with boundary points induced by envelopments of M into $\mathbb{R}^2$,

the Riemann mapping theorem ensures the possibilities are vast, as it yields such an envelopment with $\phi(M)$ any non-empty simply connected open set in the plane.

A.2 Classifying Abstract Boundary Points

Thus far, the construction has used nothing more than the smooth manifold structure of M, and we have managed to generate the set $\mathcal{B}(M)$ of abstract boundary points. We would now like to use the spacetime structure of (M, g) to classify the abstract boundary points according to a scheme which identifies any singularities. To avoid complications distracting from the physical application in mind, we will assume that (M, g) is maximally extended subject to the desired regularity in g and any other physical requirements (e.g. time-orientability or other minimal causality condition). In the language of the broader classification detailed in [125], this amounts to assuming there exist no *regular* boundary points. This considerably simplifies the forms of Definitions 18, 19, and 20 below as compared to their statements in [125].

As is standard in the theory of singularities, we must distinguish singular boundary points from those at "infinity" or otherwise by identifying singularities as being reachable with bounded parameter via a physically meaningful family of curves. The minimum requirements of the curve family are as follows:

Definition 16. *A family $\mathcal{C}$ of C^1 parameterised curves $\gamma : [a, b) \to M$ $(a < b \leqslant \infty)$ is said to have the bounded parameter property (b.p.p.) provided that*

(i) For any $p \in M$, there exists a $\gamma \in \mathcal{C}$ passing through p.

(ii) For any $\gamma \in \mathcal{C}$, any subcurve $\gamma|_{[a',b')}$ (where $[a', b') \subset [a, b)$) is also in $\mathcal{C}$.

(iii) For any $\gamma, \tilde{\gamma} \in \mathcal{C}$ related by an increasing change in parameter, either both $b, \tilde{b} < \infty$ ($\gamma, \tilde{\gamma}$ have bounded parameter) or $b = \tilde{b} = \infty$ ($\gamma, \tilde{\gamma}$ have unbounded parameter).

These conditions simply ensure that the family $\mathcal{C}$ has sufficiently many curves to arguably characterize boundary behavior, and that it provides unambiguous notions of "bounded" and "unbounded" distance along the curves. Families of particular interest that satisfy this definition include the collection of geodesics with affine parameter, the collection of C^1 curves with generalized affine parameter, the causal or future-directed causal subfamilies of each of these, and the collection of C^1 timelike curves with proper time parameter. The b.p.p. family of curves with which one chooses to work is the means by which the spacetime metric characterizes the upcoming classification– notice that in all of the curve families just listed, the metric determines the parameterization, and hence whether a given curve has bounded or unbounded parameter.

Definition 17. *Given an envelopment $\phi : M \to \widehat{M}$ and a b.p.p. family of curves $\mathcal{C}$, a boundary set $B \subset \partial\phi(M)$ is $\mathcal{C}$-approachable, or just approachable, if there exists a $\gamma \in \mathcal{C}$ such that $\phi \circ \gamma$ has a limit point in B.*

It is a direct corollary of the sequential remark made following Definition 14 that if $B \rhd B'$ and B' is approachable, then B is approachable as well (via the same curve γ). Hence the property of approachability passes to the abstract boundary: if $B \sim B'$, then B is approachable iff B' is approachable, and it is unambiguous to say that the abstract boundary set $[B]$ is approachable. Approachability is a basic requirement of being able to proceed with the classification, as approaching curves are the means by which boundary points are measured to be at "finite" or "infinite" distance, and this is precisely what determines whether the boundary point is deemed singular or not:

Definition 18. *Given an envelopment $\phi : M \to \widehat{M}$ (M maximally extended) and a b.p.p. family of curves $\mathcal{C}$, an approachable boundary point $p \in \widehat{M}$ will be called a point at infinity if every $\gamma \in \mathcal{C}$ approaching p has unbounded parameter.*

Definition 19. *Given an envelopment $\phi : M \to \widehat{M}$ (M maximally extended) and a b.p.p. family of curves $\mathcal{C}$, an approachable boundary point $p \in \widehat{M}$ will be called singular or a singularity if there exists a $\gamma \in \mathcal{C}$ approaching p with bounded parameter.*

The properties of being a point at infinity or a singularity again clearly pass to the abstract boundary, so that we may say the abstract boundary point $[p]$ is a point at infinity or singularity in the same manner. This nominally achieves the desired classification, but the immense flexibility of envelopments means that we need one final distinction to extract "pure" singular behavior:

Definition 20. *A singular boundary point will be called mixed or directional if it covers a point at infinity. Otherwise, it will be called a pure singularity.*

Mixed singularities are unaviodable byproducts of the generality of the abstract boundary construction. They apparently contain both singular and nonsingular behavior, and we'd like to filter out the latter in our identification of singular neigh-

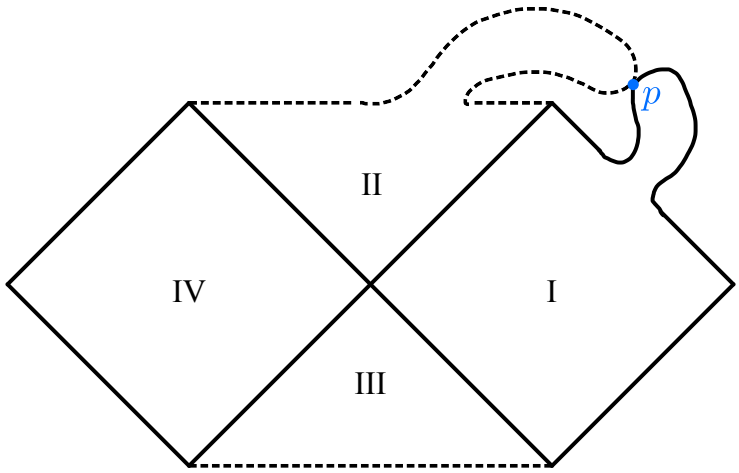

FIGURE A.3: An envelopment of the Schwarzschild spacetime giving rise to a mixed singularity. The usual envelopment of the Penrose diagram is locally modified around a portion of the $r = 0$ boundary to region II and a portion of $\mathscr{I}^+$ for region I, stretching the boundaries until they touch at a point. The Schwarzschild spacetime can still be smoothly (not conformally) embedded without issue into the interior of this region, and the boundary point p of tangency is a mixed singularity. Thus if Definition 11 were based around general singularities instead of pure singularities, then singular neighborhoods would apparently have to surround $\mathscr{I}^+$ as well, which is undesirable.

borhoods. See Figure A.3 for an example with the Schwarzschild spacetime demonstrating the need for this distinction, particularly as it impacts the notion of singular neighborhood given in Definition 11. In any event, Definition 20 completes the classification scheme of abstract boundary points for a maximally extended spacetime. Within this scheme, every abstract boundary point is exactly one of the following, given a choice of b.p.p. curve family: unapproachable, a point at infinity, a mixed singularity, or a pure singularity.

It is worth briefly commenting on how one might hope to identify a pure singularity $p \in \widehat{M}$, as it in principle requires checking the status of boundary points in every other envelopment and their covering relations with p. The easiest way to identify a pure singularity via the data available in a single envelopment is to observe that if *every* $\gamma \in \mathcal{C}$ approaching p has bounded parameter, then p must be a pure singularity. This is simply because if p covered a point at infinity q, then any curve γ approaching q would also have to approach p, and this curve would have to have unbounded parameter. This is the easiest way to identify, say, the $r = 0$ points in a Schwarzschild Penrose diagram as necessarily being pure singularities with respect to physically relevant curve families.

A.3 The Strongly Attached Point Topology

We now turn to the problem of identifying a natural topology on the set $\overline{M} := M \cup \mathcal{B}(M)$, which is furnished readily by the available topological structure of the envelopments giving rise to abstract boundary sets. The definitions in this section were originally put forward in [10]. We first define two terms capturing what it means for a boundary set to be "near" an open set in M– the former of these we state only to simplify the statement of some results in Section 5.4, while the latter is critical to identifying singular neighborhoods.

Definition 21. *Given an open set $U \subset M$ and an envelopment $\phi : M \to \widehat{M}$, a boundary set $B \subset \partial\phi(M)$ is said to be attached to U if $B \cap \overline{\phi(U)} \neq \varnothing$.*

Definition 22. *Given an open set $U \subset M$ and an envelopment $\phi : M \to \widehat{M}$, a boundary set $B \subset \partial\phi(M)$ is said to be strongly attached to U if there exists an open set $N \subset \widehat{M}$ containing B such that*

$$N \cap \phi(M) \subset \phi(U).$$

See Figure A.4. This latter definition captures that U "surrounds" B as much as possible from within M, in the sense that if a sequence $(p_n)_{n=1}^{\infty} \subset M$ has the property that $(\phi(p_n))$ limits to a point in B, then (p_n) eventually enters and remains inside U. What's more, the collection of open sets in M to which B is strongly attached fully characterizes the topological association of B to M, as seen in the near-converse to the above: if (p_n) eventually enters and remains inside every open set $U \subset M$ to which B is strongly attached, then $(\phi(p_n))$ has a limit point in B.

It is a fairly immediate exercise in definition chasing to verify that if the boundary set B is strongly attached to $U \subset M$ and $B \rhd B'$, then B' is also strongly attached to U. Hence the property of being strongly attached to a given open set passes to

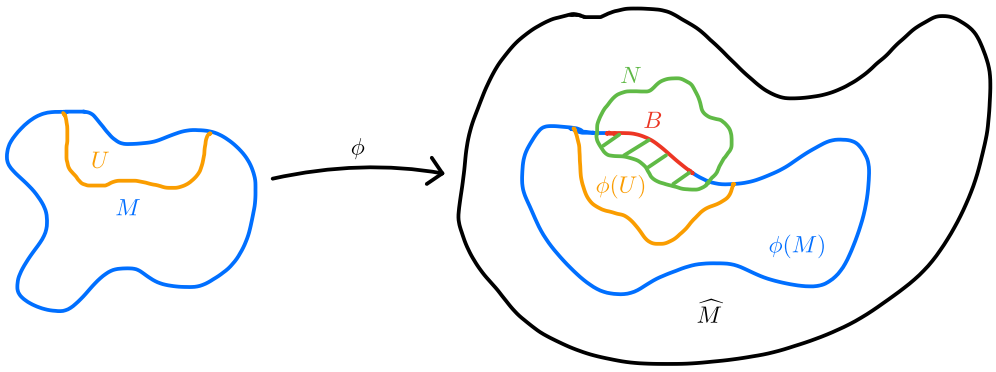

FIGURE A.4: A depiction of Definition 22. We may find an open neighborhood $N \subset \widehat{M}$ around the boundary set $B \subset \partial\phi(M) \subset \widehat{M}$ which has intersection with $\phi(M)$ entirely contained within $\phi(U)$. B, and every boundary point therein, is strongly attached to U.

the abstract boundary, and it is unambiguous to say that the abstract boundary set $[B]$ is strongly attached to U. Denoting by $\mathcal{B}_U \subset \mathcal{B}(M)$ the set of abstract boundary points strongly attached to U, we may now characterize the desired topology on $\overline{M}$:

Definition 23. *The strongly attached point topology* $\mathcal{T}_{sap}(M)$ *on* $\overline{M} = M \cup \mathcal{B}(M)$ *is the topology with basis*

$$\mathcal{W} := \{U \cup \mathcal{B}_U \mid U \subset M \text{ is open}\}.$$

It is straightforward to see that $\mathcal{W}$ covers $\overline{M}$ and that $\mathcal{B}_{U_1} \cap \mathcal{B}_{U_2} = \mathcal{B}_{U_1 \cap U_2}$ ([10], Lemmas 17 and 22), so that $\mathcal{W}$ is indeed a legitimate basis to a topology. This topology inherits from the sequential observations made following Definition 22 the property that, given a boundary point $p \in \partial\phi(M)$ under the envelopment $\phi : M \to \widehat{M}$, a sequence $(p_n)_{n=1}^{\infty} \subset M \subset \overline{M}$ limits to the abstract boundary point $[p] \in \overline{M}$ iff $(\phi(p_n))$ limits to $p \in \widehat{M}$. As it can be shown that $\mathcal{T}_{sap}(M)$ is first countable ([10], Proposition 46), and hence sequential, this indicates the strong association between the topological structure of $\mathcal{T}_{sap}(M)$ and that of the various envelopments.

Additional features of interest of $\mathcal{T}_{sap}(M)$ are that it induces the standard topology back onto M, $M \subset \overline{M}$ is open, and the closure of M is indeed all of $\overline{M}$. Hence $\mathcal{B}(M)$ is the topological boundary ∂M of M in $\overline{M}$ under $\mathcal{T}_{sap}(M)$, as one would hope of a reasonable topology on $\overline{M}$. While $\overline{M}$ is *not* Hausdorff, this is a feature rather than a bug of $\mathcal{T}_{sap}(M)$: the failure of Hausdorff separation between distinct abstract boundary points encodes when boundary points in different envelopments are topologically entangled, e.g. when one covers the other but not vice versa, or when a sequence in M may limit to both simultaneously under their respective envelopments. As seen in the discussion following Example 4 above, there is immense degeneracy in distinct abstract boundary points covering others, and such points' topological relations to M are intertwined in a complicated manner, so it is useful and expected that $\mathcal{T}_{sap}$ should encode this in topological closeness. Also see Example

5 below. The important separation property for our purposes is that points in M *are* Hausdorff separated from points in $\mathcal{B}(M)$. For proof and additional discussion of these features and more, we again refer the reader to [10].

While we generally refer to [10] for proofs of a-boundary topological features, we will now provide one example of how such arguments go by extending a result in [10] indicating the close topological relationship between $\mathcal{T}_{sap}(M)$ and envelopments. Given an envelopment $\phi : M \to \widehat{M}$, we denote by $\sigma_\phi \subset \mathcal{B}(M)$ the set of abstract boundary points with singleton representatives in $\partial\phi(M)$, e.g.

$$\sigma_\phi := \{[p] \in \mathcal{B}(M) \mid p \in \partial\phi(M)\}.$$

Since distinct boundary points in the same envelopment cannot cover each other due to the Hausdorff property of $\widehat{M}$, clearly σ_ϕ is in bijection with $\partial\phi(M)$ under the natural quotient map $\pi : \partial\phi(M) \to \sigma_\phi$ given by $\pi(p) = [p]$.

Proposition 16. *When $\sigma_\phi \subset \mathcal{B}(M)$ and $\partial\phi(M)$ are given the subspace topologies induced by $\mathcal{T}_{sap}(M)$ and $\widehat{M}$ respectively, the quotient map $\pi : \partial\phi(M) \to \sigma_\phi$ is a homeomorphism.*

Proof. Let $O \subset \sigma_\phi$ be open in the subspace topology, so that $O = \widetilde{U} \cap \sigma_\phi$ for some open $\widetilde{U} \subset \overline{M}$. WLOG, we may assume $\widetilde{U}$ is in the basis, i.e. $\widetilde{U} = U \cup \mathcal{B}_U$ for some open $U \subset M$, and hence that $O = \mathcal{B}_U \cap \sigma_\phi$. Take $p \in \pi^{-1}(O)$, so that $p \in \partial\phi(M)$ such that $[p] \in O \subset \mathcal{B}_U$, and hence p is strongly attached to U. Then there exists a neighborhood $N \subset \widehat{M}$ of p such that $N \cap \phi(M) \subset \phi(U)$. Of course, every point $q \in N \cap \partial\phi(M)$ is also strongly attached to U, as N serves equally well as the required neighborhood of q in Definition 22. That is, $\pi(N \cap \partial\phi(M)) \subset \mathcal{B}_U \cap \sigma_\phi = O$, and hence $N \cap \partial\phi(M)$ is a neighborhood of p entirely contained in $\pi^{-1}(O)$, showing $\pi^{-1}(O)$ is open. This shows π is continuous.

Now let $O \subset \partial\phi(M)$ be open in the subspace topology, so that $O = N \cap \partial\phi(M)$ for some open $N \subset \widehat{M}$. Fix $p \in O$, and observe that since $\widehat{M}$ is topologically regular, there exists an open neighborhood N' of p with $\overline{N'} \subset N$. Setting $U := \phi^{-1}(N' \cap \phi(M))$ and noting p is clearly strongly attached to U, we have that $\mathcal{B}_U \cap \sigma_\phi = (U \cup \mathcal{B}_U) \cap \sigma_\phi$ is an open neighborhood of $[p] = \pi(p)$ in σ_ϕ. Noting that $q \in \partial\phi(M)$ being strongly attached to U implies that $q \in \overline{\phi(U)}$, we have that

$$\pi^{-1}(\mathcal{B}_U \cap \sigma_\phi) \subset \overline{\phi(U)} \cap \partial\phi(M) \subset \overline{N'} \cap \partial\phi(M) \subset N \cap \partial\phi(M) = O.$$

We therefore have $\mathcal{B}_U \cap \sigma_\phi \subset \pi(O)$, so that $\mathcal{B}_U \cap \sigma_\phi$ is a neighborhood of $\pi(p)$ entirely contained in $\pi(O)$, showing $\pi(O)$ is open and hence that π is an open map.

□

This strengthens Proposition 52 in [10], which invoked an additional, somewhat complicated hypothesis (given there as Condition 51) to prove an equivalent result. Extracting the intuitive information from this result requires a slight rephrasing. Given an envelopment $\phi : M \to \widehat{M}$, define $\overline{\pi} : \overline{\phi(M)} \to \overline{M}$ by

$$\overline{\pi}(p) = \begin{cases} \phi^{-1}(p) & p \in \phi(M) \\ [p] & p \in \partial\phi(M). \end{cases}$$

This map just identifies points in $\overline{\phi(M)} \subset \widehat{M}$ with points in $\overline{M}$ in the most direct way possible. Essentially the same proof as above carries through to demonstrate the following:

Proposition 17. *When $\overline{\phi(M)}$ is given the subspace topology induced by $\widehat{M}$, the map $\overline{\pi} : \overline{\phi(M)} \to \overline{M}$ is a topological embedding, i.e. a homeomorphism onto its image.*

□

This essentially says that the natural identification of the closure of M within an envelopment with the corresponding points in $\overline{M}$ preserves all topological features

of the envelopment. The abstract boundary endowed with the strongly attached point topology, then, naturally contains within it all possible closures of M within envelopments, complete with their topological structure.

We once again conclude by demonstrating the definitions via the single example for which complete, explicit, and nontrivial computation is tractable.

Example 5. *The strongly attached point topology on $\mathcal{B}(M)$ for $M = \mathbb{R}$.*

The subspace topology $\mathcal{T}^{\mathcal{B}}_{sap}(\mathbb{R})$ on $\mathcal{B}(\mathbb{R})$ induced by $\mathcal{T}_{sap}(\mathbb{R})$ on $\overline{\mathbb{R}}$ has basis given by

$$\mathcal{W}^{\mathcal{B}} := \{\mathcal{B}_U \mid U \subset \mathbb{R} \text{ is open}\}.$$

It is straightforward to see that if an open set $U \subset \mathbb{R}$ contains an interval of the form (a, ∞), then $[R]$ is strongly attached to U; if it contains an interval of the form $(-\infty, b)$, then then $[L]$ is strongly attached to U; and if it contains intervals of both forms, then all of $[L], [R]$, and $[LR]$ are strongly attached to U. Hence we have

$$\mathcal{W}^{\mathcal{B}} = \{\ \{[L]\}, \{[R]\}, \{[L], [R], [LR]\}\ \},$$

and this yields

$$\mathcal{T}^{\mathcal{B}}_{sap}(\mathbb{R}) = \{\ \varnothing, \{[L]\}, \{[R]\}, \{[L], [R]\}, \{[L], [R], [LR]\}\ \}.$$

This topology is T_0 but not T_1. That every neighborhood of $[LR]$ includes both $[L]$ and $[R]$ encodes topologically that $[LR] \rhd [L]$ and $[LR] \rhd [R]$.

Bibliography

[1] TMC Abbott, M Aguena, A Alarcon, S Allam, O Alves, A Amon, F Andrade-Oliveira, J Annis, S Avila, D Bacon, et al. Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2):023520, 2022.

[2] Nabila Aghanim, Yashar Akrami, Mark Ashdown, J Aumont, C Baccigalupi, M Ballardini, AJ Banday, RB Barreiro, N Bartolo, S Basak, et al. Planck 2018 results-vi. cosmological parameters. *Astronomy & Astrophysics*, 641:A6, 2020.

[3] A Arbey and J-F Coupechoux. Cosmological scalar fields and big-bang nucleosynthesis. *Journal of Cosmology and Astroparticle Physics*, 2019(11):038, 2019.

[4] Eric Armengaud, Nathalie Palanque-Delabrouille, Christophe Yèche, David JE Marsh, and Julien Baur. Constraining the mass of light bosonic dark matter using sdss lyman-α forest. *Monthly Notices of the Royal Astronomical Society*, 471(4):4606–4614, 2017.

[5] Michael Ashley. Singularity theorems and the abstract boundary construction. 2002.

[6] Abhay Ashtekar, Christopher Beetle, Olaf Dreyer, Stephen Fairhurst, Badri Krishnan, Jerzy Lewandowski, and Jacek Wiśniewski. Generic isolated horizons and their applications. *Physical Review Letters*, 85(17):3564, 2000.

[7] Abhay Ashtekar and Badri Krishnan. Isolated and dynamical horizons and their applications. *Living Reviews in Relativity*, 7(1):1–91, 2004.

[8] A Balbi and C Quercellini. The time evolution of cosmological redshift as a test of dark energy. *Monthly Notices of the Royal Astronomical Society*, 382(4):1623–1629, 2007.

[9] James M Bardeen, Brandon Carter, and Stephen W Hawking. The four laws of black hole mechanics. *Communications in mathematical physics*, 31(2):161–170, 1973.

[10] Richard A Barry and Susan M Scott. The strongly attached point topology of the abstract boundary for space-time. *Classical and Quantum Gravity*, 31(12):125004, 2014.

[11] Daniel Baumann. *Cosmology*. Cambridge University Press, 2022.

[12] Ingemar Bengtsson. Spherical symmetry and black holes. Lecture Notes, Stockholm University, 2012.

[13] Ingemar Bengtsson and Jose MM Senovilla. Region with trapped surfaces in spherical symmetry, its core, and their boundaries. *Physical Review D*, 83(4):044012, 2011.

[14] Antonio N Bernal and Miguel Sánchez. On smooth cauchy hypersurfaces and geroch's splitting theorem. *arXiv preprint gr-qc/0306108*, 2003.

[15] Carlos A Bertulani, Francis W Hall, and Benjami I Santoyo. Big bang nucleosynthesis as a probe of new physics. *arXiv preprint arXiv:2210.04071*, 2022.

[16] Michael Boylan-Kolchin, James S Bullock, and Manoj Kaplinghat. Too big to fail? the puzzling darkness of massive milky way subhaloes. *Monthly Notices of the Royal Astronomical Society: Letters*, 415(1):L40–L44, 2011.

[17] Hubert Bray, Benjamin Hamm, Sven Hirsch, James Wheeler, and Yiyue Zhang. Flatly foliated relativity. *Pure and Applied Mathematics Quarterly*, 15(2):707–747, December 2019.

[18] Hubert L Bray. Proof of the riemannian penrose inequality using the positive mass theorem. *Journal of Differential Geometry*, 59(2):177–267, 2001.

[19] Hubert L Bray. On dark matter, spiral galaxies, and the axioms of general relativity. *Geometric analysis, mathematical relativity, and nonlinear partial differential equations*, 599:1–64, 2010.

[20] Andrew James Bruce. On the bundle of null cones. *arXiv preprint arXiv:2204.11645*, 2022.

[21] Elie Cartan. Sur les équations de la gravitation d'einstein. *Journal de Mathématiques pures et appliquées*, 1:141–203, 1922.

[22] Francesca Chadha-Day, John Ellis, and David JE Marsh. Axion dark matter: What is it and why now? *Science advances*, 8(8):eabj3618, 2022.

[23] Barry T Chiang, Hsi-Yu Schive, Tzihong Chiueh, et al. Soliton oscillations and revised constraints from eridanus ii of fuzzy dark matter. *Physical Review D*, 103(10):103019, 2021.

[24] Steve K Choi, Matthew Hasselfield, Shuay-Pwu Patty Ho, Brian Koopman, Marius Lungu, Maximilian H Abitbol, Graeme E Addison, Peter AR Ade, Simone Aiola, David Alonso, et al. The atacama cosmology telescope: a measurement of the cosmic microwave background power spectra at 98 and 150 ghz. *Journal of Cosmology and Astroparticle Physics*, 2020(12):045, 2020.

[25] R Chown, Y Omori, K Aylor, BA Benson, LE Bleem, JE Carlstrom, CL Chang, HM Cho, TM Crawford, AT Crites, et al. Maps of the southern millimeter-wave sky from combined 2500 deg2 spt-sz and planck temperature data. *The Astrophysical Journal Supplement Series*, 239(1):10, 2018.

[26] Demetrios Christodoulou. Violation of cosmic censorship in the gravitational collapse of a dust cloud. *Communications in Mathematical Physics*, 93(2):171–195, 1984.

[27] Demetrios Christodoulou. Examples of naked singularity formation in the gravitational collapse of a scalar field. *Annals of Mathematics*, 140(3):607–653, 1994.

[28] Demetrios Christodoulou. The instability of naked singularities in the gravitational collapse of a scalar field. *Annals of Mathematics*, 149(1):183–217, 1999.

[29] Demetrios Christodoulou. On the global initial value problem and the issue of singularities. *Classical and Quantum Gravity*, 16(12A):A23, 1999.

[30] Demetrios Christodoulou and Sergiu Klainerman. The global nonlinear stability of the minkowski space. *Séminaire Équations aux dérivées partielles (Polytechnique) dit aussi" Séminaire Goulaouic-Schwartz"*, pages 1–29, 1993.

[31] Douglas Clowe, Maruša Bradač, Anthony H Gonzalez, Maxim Markevitch, Scott W Randall, Christine Jones, and Dennis Zaritsky. A direct empirical proof of the existence of dark matter. *The Astrophysical Journal Letters*, 648(2):L109, 2006.

[32] Raymond T Co, David Dunsky, Nicolas Fernandez, Akshay Ghalsasi, Lawrence J Hall, Keisuke Harigaya, and Jessie Shelton. Gravitational wave and cmb probes of axion kination. *Journal of High Energy Physics*, 2022(9):1–55, 2022.

[33] Matthew Colless, Gavin Dalton, Steve Maddox, Will Sutherland, Peder Norberg, Shaun Cole, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Chris Collins, et al. The 2df galaxy redshift survey: spectra and redshifts. *Monthly Notices of the Royal Astronomical Society*, 328(4):1039–1063, 2001.

[34] R Consiglio, PF de Salas, Giuseppe Mangano, Gennaro Miele, S Pastor, and O Pisanti. Parthenope reloaded. *Computer Physics Communications*, 233:237–242, 2018.

[35] Erik Curiel. The many definitions of a black hole. *Nature Astronomy*, 3(1):27–34, 2019.

[36] Mihalis Dafermos. Black holes without spacelike singularities. *Communications in Mathematical Physics*, 332(2):729–757, 2014.

[37] Mihalis Dafermos and Igor Rodnianski. Lectures on black holes and linear waves. *Clay Math. Proc*, 17:97–205, 2013.

[38] Neal Dalal and Andrey Kravtsov. Excluding fuzzy dark matter with sizes and stellar kinematics of ultrafaint dwarf galaxies. *Physical Review D*, 106(6):063517, 2022.

[39] Hooman Davoudiasl and Peter B Denton. Ultralight boson dark matter and event horizon telescope observations of m 87. *Physical review letters*, 123(2):021102, 2019.

[40] WJG De Blok. The core-cusp problem. *Advances in Astronomy*, 2010, 2010.

[41] Mona Dentler, David JE Marsh, Renée Hložek, Alex Laguë, Keir K Rogers, and Daniel Grin. Fuzzy dark matter and the dark energy survey year 1 data. *Monthly Notices of the Royal Astronomical Society*, 515(4):5646–5664, 2022.

[42] Francesco D'Eramo, Nicolas Fernandez, and Stefano Profumo. When the universe expands too fast: relentless dark matter. *Journal of Cosmology and Astroparticle Physics*, 2017(05):012, 2017.

[43] Eleonora Di Valentino, Olga Mena, Supriya Pan, Luca Visinelli, Weiqiang Yang, Alessandro Melchiorri, David F Mota, Adam G Riess, and Joseph Silk. In the realm of the hubble tension—a review of solutions. *Classical and Quantum Gravity*, 38(15):153001, 2021.

[44] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992.

[45] Xiaolong Du. *Structure formation with ultralight axion dark matter*. PhD thesis, Georg-August-Universität Göttingen, 2018.

[46] IH Dwivedi and PS Joshi. On the nature of naked singularities in Vaidya spacetimes. *Classical and Quantum Gravity*, 6(11):1599, 1989.

[47] IH Dwivedi and PS Joshi. On the nature of naked singularities in Vaidya spacetimes: II. *Classical and Quantum Gravity*, 8(7):1339, 1991.

[48] Douglas M Eardley and Larry Smarr. Time functions in numerical relativity: marginally bound dust collapse. *Physical Review D*, 19(8):2239, 1979.

[49] Robert Geroch. Local characterization of singularities in general relativity. *Journal of Mathematical Physics*, 9(3):450–465, 1968.

[50] Robert Geroch. What is a singularity in general relativity? *Annals of Physics*, 48(3):526–540, 1968.

[51] Robert Geroch. Domain of dependence. *Journal of Mathematical Physics*, 11(2):437–449, 1970.

[52] Robert Geroch, Liang Can-bin, and Robert M Wald. Singular boundaries of space–times. *Journal of Mathematical Physics*, 23(3):432–435, 1982.

[53] Robert Geroch, EH Kronheimer, and Roger Penrose. Ideal points in space-time. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 327(1571):545–567, 1972.

[54] SG Ghosh and Naresh Dadhich. Naked singularities in higher dimensional Vaidya space-times. *Physical Review D*, 64(4):047501, 2001.

[55] Particle Data Group, RL Workman, VD Burkert, V Crede, E Klempt, U Thoma, L Tiator, K Agashe, G Aielli, BC Allanach, et al. Review of particle physics. *Progress of theoretical and experimental physics*, 2022(8):083C01, 2022.

[56] Benjamin Hamm. *Scalar Field Wave Dark Matter and Galactic Halos*. PhD thesis, Duke University, 2021.

[57] Tomohiro Harada, Chul-Moon Yoo, and Kazunori Kohri. Threshold of primordial black hole formation. *Physical Review D*, 88(8):084051, 2013.

[58] Stephen W Hawking. Occurrence of singularities in open universes. *Physical Review Letters*, 15(17):689, 1965.

[59] Stephen W Hawking. Black holes in general relativity. *Communications in Mathematical Physics*, 25(2):152–166, 1972.

[60] Stephen W Hawking and George Francis Rayner Ellis. *The large scale structure of space-time*, volume 1. Cambridge university press, 1973.

[61] Sean A Hayward. General laws of black-hole dynamics. *Physical Review D*, 49(12):6467, 1994.

[62] Sean A Hayward. Black holes: New horizons. *arXiv preprint gr-qc/0008071*, 2000.

[63] Sean A Hayward. Formation and evaporation of nonsingular black holes. *Physical review letters*, 96(3):031103, 2006.

[64] Renée Hlozek, Daniel Grin, David JE Marsh, and Pedro G Ferreira. A search for ultralight axions using precision cosmological data. *Physical Review D*, 91(10):103512, 2015.

[65] Renée Hložek, David JE Marsh, and Daniel Grin. Using the full power of the cosmic microwave background to probe axion dark matter. *Monthly Notices of the Royal Astronomical Society*, 476(3):3063–3085, 2018.

[66] Wayne Hu, Rennan Barkana, and Andrei Gruzinov. Fuzzy cold dark matter: the wave properties of ultralight particles. *Physical Review Letters*, 85(6):1158, 2000.

[67] Wayne Hu and Scott Dodelson. Cosmic microwave background anisotropies. *Annual Review of Astronomy and Astrophysics*, 40(1):171–216, 2002.

[68] Wayne Hu, Naoshi Sugiyama, and Joseph Silk. The physics of microwave background anisotropies. *Nature*, 386(6620):37–43, 1997.

[69] Edwin Hubble. A relation between distance and radial velocity among extragalactic nebulae. *Proceedings of the national academy of sciences*, 15(3):168–173, 1929.

[70] Lam Hui, Jeremiah P Ostriker, Scott Tremaine, and Edward Witten. Ultralight scalars as cosmological dark matter. *Physical Review D*, 95(4):043541, 2017.

[71] Gerhard Huisken and Tom Ilmanen. The inverse mean curvature flow and the riemannian penrose inequality. *Journal of Differential Geometry*, 59(3):353–437, 2001.

[72] Jai-chan Hwang and Hyerim Noh. Axion as a cold dark matter candidate. *Physics Letters B*, 680(1):1–3, 2009.

[73] Myungkook James Jee, Holland C Ford, Garth D Illingworth, Richard L White, TJ Broadhurst, DA Coe, GR Meurer, Arjen van der Wel, N Benitez, John P Blakeslee, et al. Discovery of a ringlike dark matter structure in the core of the galaxy cluster cl 0024+ 17. *The Astrophysical Journal*, 661(2):728, 2007.

[74] Pankaj S Joshi. Gravitational collapse: the story so far. *Pramana*, 55(4):529–544, 2000.

[75] Pankaj S Joshi and IH Dwivedi. Strong curvature naked singularities in non-self-similar gravitational collapse. *General relativity and gravitation*, 24(2):129–137, 1992.

[76] Michael Joyce. Electroweak baryogenesis and the expansion rate of the universe. *Physical Review D*, 55(4):1875, 1997.

[77] Marc Kamionkowski and Adam G Riess. The hubble tension and early dark energy. *arXiv preprint arXiv:2211.04492*, 2022.

[78] Lawrence Kawano. Let's go: Early universe 2. primordial nucleosynthesis the computer way. 1992.

[79] Daniel D Kelson, Paul Martini, and JS Mulchaey. Optimal measurements of redshifts using the weighted cross-correlation, 2003.

[80] William Thomson Baron Kelvin. *Baltimore lectures on molecular dynamics and the wave theory of light*. CJ Clay and Sons, 1904.

[81] Stacy Y Kim, Annika HG Peter, and Jonathan R Hargis. there is no missing satellites problem. *arXiv preprint arXiv:1711.06267*, 2017.

[82] Anatoly Klypin, Andrey V Kravtsov, Octavio Valenzuela, and Francisco Prada. Where are the missing galactic satellites? *The Astrophysical Journal*, 522(1):82, 1999.

[83] Andrzej Królak. Definitions of black holes without use of the boundary at infinity. *General Relativity and Gravitation*, 14(8):793–801, 1982.

[84] Yuhji Kuroda. Naked singularities in the Vaidya spacetime. *Progress of theoretical physics*, 72(1):63–72, 1984.

[85] Michael J Kurtz and Douglas J Mink. Rvsao 2.0: Digital redshifts and radial velocities. *Publications of the Astronomical Society of the Pacific*, 110(750):934, 1998.

[86] Kayll Lake. Naked singularities in gravitational collapse which is not self-similar. *Physical Review D*, 43(4):1416, 1991.

[87] Kayll Lake. Testing the λ cold dark matter model (and more) with the time evolution of the redshift. *Physical Review D*, 76(6):063508, 2007.

[88] D Larson, MR Nolta, M Halpern, RS Hill, N Odegard, et al. Nine-year wilkinson microwave anisotropy probe (wmap) observations: cosmological parameter results. *The Astrophysical Journal Supplement Series*, 208(2):19, 2013.

[89] Georges Lemaître. Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles, A47, p. 49-59*, 47:49–59, 1927.

[90] Antony Lewis. Cosmological parameters from wmap 5-year temperature maps. *Physical Review D*, 78(2):023002, 2008.

[91] Chris Lidman, BE Tucker, TM Davis, SA Uddin, J Asorey, K Bolejko, D Brout, J Calcino, D Carollo, A Carr, et al. Ozdes multi-object fibre spectroscopy for the dark energy survey: results and second data release. *Monthly Notices of the Royal Astronomical Society*, 496(1):19–35, 2020.

[92] Abraham Loeb. Direct measurement of cosmological parameters from the cosmic deceleration of extragalactic objects. *The Astrophysical Journal*, 499(2):L111, 1998.

[93] David Lovelock. The einstein tensor and its generalizations. *Journal of Mathematical Physics*, 12(3):498–501, 1971.

[94] Jonathan Luk and Sung-Jin Oh. Strong cosmic censorship in spherical symmetry for two-ended asymptotically flat initial data ii: the exterior of the black hole region. *Annals of PDE*, 5(1):1–194, 2019.

[95] David JE Marsh. Axion cosmology. *Physics Reports*, 643:1–79, 2016.

[96] John C Mather, ES Cheng, RE Eplee Jr, RB Isaacman, SS Meyer, RA Shafer, R Weiss, EL Wright, CL Bennett, NW Boggess, et al. A preliminary measurement of the cosmic microwave background spectrum by the cosmic background explorer (cobe) satellite. *Astrophysical Journal, Part 2-Letters (ISSN 0004-637X), vol. 354, May 10, 1990, p. L37-L40.*, 354:L37–L40, 1990.

[97] Tonatiuh Matos, Alberto Vázquez-González, and Juan Magana. $\varphi 2$ as dark matter. *Monthly Notices of the Royal Astronomical Society*, 393(4):1359–1369, 2009.

[98] Brett McInnes. De sitter and schwarzschild-de sitter according to schwarzschild and de sitter. *Journal of High Energy Physics*, 2003(09):009, 2003.

[99] Ben Moore. Evidence against dissipation-less dark matter from observations of galaxy haloes. *Nature*, 370(6491):629–631, 1994.

[100] Alex B Nielsen. Black holes and black hole thermodynamics without event horizons. *General Relativity and Gravitation*, 41(7):1539–1584, 2009.

[101] Barrett O'neill. *Semi-Riemannian geometry with applications to relativity*. Academic press, 1983.

[102] J Robert Oppenheimer and Hartland Snyder. On continued gravitational contraction. *Physical Review*, 56(5):455, 1939.

[103] Jeremiah P Ostriker, Ena Choi, Anthony Chow, and Kundan Guha. Mind the gap: Is the too big to fail problem resolved? *The Astrophysical Journal*, 885(1):97, 2019.

[104] Roger Penrose. Gravitational collapse and space-time singularities. *Physical Review Letters*, 14(3):57, 1965.

[105] Roger Penrose. Gravitational collapse: The role of general relativity. *Nuovo Cimento Rivista Serie*, 1:252, 1969.

[106] Roger Penrose. Singularities and time-asymmetry. In *General Relativity : An Einstein Centenary Survey*, chapter 12, pages 581–683. Cambridge University Press, 1979.

[107] Roger Penrose. Republication of: Conformal treatment of infinity. *General Relativity and Gravitation*, 43(3):901–922, 2011.

[108] Arno A Penzias and Robert Woodrow Wilson. A measurement of excess antenna temperature at 4080 mc/s. *Astrophysical Journal, vol. 142, p. 419-421*, 142:419–421, 1965.

[109] Saul Perlmutter, Goldhaber Aldering, Gerson Goldhaber, RA Knop, Peter Nugent, Patricia G Castro, Susana Deustua, Sebastien Fabbro, Ariel Goobar, Donald E Groom, et al. Measurements of ω and λ from 42 high-redshift supernovae. *The Astrophysical Journal*, 517(2):565, 1999.

[110] Peter Pesic and Stephen P Boughn. The weyl–cartan theorem and the naturalness of general relativity. *European journal of physics*, 24(3):261, 2003.

[111] Bradley M Peterson. Reverberation mapping of active galactic nuclei. *Publications of the Astronomical Society of the Pacific*, 105(685):247, 1993.

[112] O Pisanti, A Cirillo, Salvatore Esposito, F Iocco, Giuseppe Mangano, Gennaro Miele, and PD Serpico. Parthenope: Public algorithm evaluating the nucleosynthesis of primordial elements. *Computer Physics Communications*, 178(12):956–971, 2008.

[113] Cyril Pitrou, Alain Coc, Jean-Philippe Uzan, and Elisabeth Vangioni. Precision big bang nucleosynthesis with improved helium-4 predictions. *Physics Reports*, 754:1–66, 2018.

[114] Adam G Riess, Stefano Casertano, Wenlong Yuan, J Bradley Bowers, Lucas Macri, Joel C Zinn, and Dan Scolnic. Cosmic distances calibrated to 1% precision with gaia edr3 parallaxes and hubble space telescope photometry of 75 milky way cepheids confirm tension with λcdm. *The Astrophysical Journal Letters*, 908(1):L6, 2021.

[115] Adam G Riess, Alexei V Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M Garnavich, Ron L Gilliland, Craig J Hogan, Saurabh Jha,

Robert P Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The astronomical journal*, 116(3):1009, 1998.

[116] Igor Rodnianski and Yakov Shlapentokh-Rothman. Naked singularities for the Einstein vacuum equations: The exterior solution. *arXiv preprint arXiv:1912.08478*, 2019.

[117] VA Rubakov. Cosmology and dark matter. *arXiv preprint arXiv:1912.04727*, 2019.

[118] Vera C Rubin, W Kent Ford Jr, and Norbert Thonnard. Rotational properties of 21 sc galaxies with a large range of luminosities and radii, from ngc 4605/r= 4kpc/to ugc 2885/r= 122 kpc. *Astrophysical Journal, Part 1, vol. 238, June 1, 1980, p. 471-487.*, 238:471–487, 1980.

[119] Pierre Salati. Quintessence and the relic density of neutralinos. *Physics Letters B*, 571(3-4):121–131, 2003.

[120] Allan Sandage. The change of redshift and apparent luminosity of galaxies due to the deceleration of selected expanding universes. *The Astrophysical Journal*, 136:319, 1962.

[121] Bernd Schmidt. On the uniqueness of boundaries at infinity of asymptotically flat spacetimes. *Classical and Quantum Gravity*, 8(8):1491, 1991.

[122] BG Schmidt. A new definition of singular points in general relativity. *General relativity and gravitation*, 1(3):269–280, 1971.

[123] Richard Schoen and S-T Yau. The existence of a black hole due to condensation of matter. *Communications in Mathematical Physics*, 90(4):575–579, 1983.

[124] Katelin Schutz. Subhalo mass function and ultralight bosonic dark matter. *Physical Review D*, 101(12):123026, 2020.

[125] Susan M Scott and Peter Szekeres. The abstract boundary—a new approach to singularities of manifolds. *Journal of Geometry and Physics*, 13(3):223–253, 1994.

[126] Susan M Scott and Ben E Whale. The endpoint theorem. *Classical and Quantum Gravity*, 38(6):065012, 2021.

[127] Sara Seager, Dimitar D Sasselov, and Douglas Scott. A new calculation of the recombination epoch. *The Astrophysical Journal*, 523(1):L1, 1999.

[128] Pasquale Dario Serpico, S Esposito, F Iocco, G Mangano, G Miele, and O Pisanti. Nuclear reaction network for primordial nucleosynthesis: a detailed analysis of rates, uncertainties and light nuclei yields. *Journal of Cosmology and Astroparticle Physics*, 2004(12):010, 2004.

[129] Susan M Simkin. Measurements of velocity dispersions and doppler shifts from digitized optical spectra. *Astronomy and Astrophysics*, 31:129, 1974.

[130] Tejinder Pal Singh. Gravitational collapse and cosmic censorship. *arXiv preprint gr-qc/9606016*, 1996.

[131] Elizabeth Swann, Mark Sullivan, Jonathan Carrick, Sebastian Hoenig, Isobel Hook, Rubina Kotak, Kate Maguire, Richard McMahon, Robert Nichol, and Stephen Smartt. 4most consortium survey 10: The time-domain extragalactic survey (tides). *The Messenger*, 175:58–61, 2019.

[132] György Szekeres. On the singularities of a riemannian manifold. *Publicationes Mathematicae Debrecen 7*, 7:285, 1960.

[133] John Tonry and Marc Davis. A survey of galaxy redshifts. i-data reduction techniques. *Astronomical Journal, vol. 84, Oct. 1979, p. 1511-1525.*, 84:1511–1525, 1979.

[134] Pierre Touboul, Gilles Métris, Manuel Rodrigues, Joel Bergé, Alain Robert, Quentin Baghi, Yves André, Judicaël Bedouet, Damien Boulanger, Stefanie Bremer, et al. M i c r o s c o p e mission: Final results of the test of the equivalence principle. *Physical review letters*, 129(12):121102, 2022.

[135] Robert V Wagoner. Big-bang nucleosynthesis revisited. *The Astrophysical Journal*, 179:343–360, 1973.

[136] Robert M Wald. Gravitational collapse and cosmic censorship. In *Black holes, gravitational radiation and the universe*, pages 69–86. Springer, 1999.

[137] Robert M Wald. *General relativity*. University of Chicago press, 2010.

[138] Hermann Weyl. *Space–time–matter*. Dutton, 1922.

[139] BE Whale. The dependence of the abstract boundary classification on a set of curves i: An algebra of sets on bounded parameter property satisfying sets of curves. *arXiv preprint arXiv:1001.5091*, 2010.

[140] BE Whale. The dependence of the abstract boundary classification on a set of curves ii: How the classification changes when the bounded parameter property satisfying set of curves changes. *arXiv preprint arXiv:1201.6414*, 2012.

[141] Ben Edward Whale. Foundations of and applications for the abstract boundary construction for space-time. 2010.

[142] James Wheeler. Generic naked singularities in vaidya spacetimes. *Classical and Quantum Gravity*, 39(19):197001, September 2022.

[143] James Wheeler. On the definition of black holes: Bridging the gap between black holes and singularities. *arXiv preprint arXiv:2205.12942*, 2022.

[144] James Wheeler. Seeking dark signals in oscillating redshifts: Exploring geometric scalar field dark matter. *arXiv preprint arXiv:2302.08753*, 2023.

[145] Fang Yuan, C Lidman, Tamara M Davis, M Childress, FB Abdalla, M Banerji, E Buckley-Geer, A Carnero Rosell, D Carollo, FJ Castander, et al. Ozdes multifibre spectroscopy for the dark energy survey: first-year operation and results. *Monthly Notices of the Royal Astronomical Society*, 452(3):3047–3063, 2015.

[146] Fritz Zwicky. Republication of: The redshift of extragalactic nebulae. *General Relativity and Gravitation*, 41(1):207–224, 2009.

Biography

James Cyrus Wheeler hails from the mountains of Lansing, North Carolina, where he grew up with his family of seven siblings. He graduated from Baylor School as a boarding student in Chattanooga, Tennessee in 2013. He attended UNC Chapel Hill from 2013 through 2017, obtaining a B.S. in physics as well as a B.S. in mathematics. From 2017 through 2023, he studied foundational aspects of general relativity under the advisement of Dr. Hubert Bray to obtain his PhD in Physics. He published several research articles during this time on a variety of topics central to general relativity [17, 142, 143, 144], most of which are discussed thoroughly in this work. He also taught in some capacity nearly every semester and summer session while at Duke, for courses in both the Physics and Mathematics Departments. James will continue his academic career as a Postdoctoral Assistant Professor in the Mathematics Department at the University of Michigan, Ann Arbor beginning in the Fall of 2023.