Is the Universe As Large As It Can Be? A Study of Spacetime Possibility

JB Manchak

For June

Contents

Preface

This book concerns the modal structure of spacetime within the context of Einstein's general relativity. The aim is to expose a rich set of philosophical issues somewhat informally and from a bird's eye view. No familiarity with general relativity is presupposed. A large number of examples (worked out in coordinates) and corresponding diagrams (over 130) will play a central role in illustrating the ideas involved. One of the intended readers is a non-expert. She is perhaps a philosophy graduate student just getting started with an interest in better understanding cosmic possibility delimited by Einstein's theory. Under her belt, she has little more than some basic set theory and a passing acquaintance with calculus and vector spaces. The other intended reader is the expert. She will find several interconnecting lines of current research mapped out and a clear thesis defended. Many of the results have appeared in print before but there is also a good deal of new material here.

The investigation will focus on the "maximality" property of spacetime – the requirement that (a model of) the universe must be "as large as it can be." The maximality of spacetime is something of a dogma within the context of generality relativity. In the background, it is often presupposed by practitioners without comment in order to go on to investigate a number of other foundational topics (e.g. determinism). Here and there, one does sometimes find a few sentences by way of justification. A passage from John Earman (1995, p. 32) gives sense of the literature.

> Metaphysical considerations suggest that to be a serious candidate for describing actuality, a spacetime should be maximal. For example, for the Creative Force to actualize a proper subpart of a larger spacetime would seem to be a violation of Leibniz's principles of sufficient reason and plenitude. If one adopts the image of spacetime as being generated or built up as time passes then the dynamical version of the principle of sufficient reason

> would ask why the Creative Force would stop building if it is possible to continue.

The Leibnizian reasoning here may seem compelling – even obvious. But under the surface, one finds a surprisingly weak foundation for the dogma of spacetime maximality. First, there are definitional puzzles: what does it even mean to say that the universe is as large as it can be? The modal notion of spacetime maximality is defined relative to a background collection of possible universes. But what is a "possible universe" within the context of general relativity? Much depends on this deep, murky question. Second, the dogma comes with an epistemological problem. Spacetime maximality is not the sort of property that can be empirically observed. Various forms of induction do not help the situation. So how can one ever be in position to know that the universe is as large as it can be? Third, metaphysical tensions abound. Leibnizian principles are problematic in some contexts: it is not always possible for the Creative Force to build a universe to be as large as it can be. Moreover, the dogma can clash with other cherished metaphysical demands. One example is the requirement that spacetime maximality be a property of all "nearby" possible universes.

In what follows, I do not defend or oppose the case for spacetime maximality. Instead, I will simply put forward a thesis that nonetheless runs counter to the prevailing orthodoxy: it is not at all clear that the universe is a large as it can be. Keeping track of the evidence for and against spacetime maximality will be the primary undertaking. In addition, I will explore the some of the philosophical consequences if the dogma were to hold and also if it were to fail. The inspiration for the project can be traced back to three main sources:

(i) Appendix B from the paper "Singularities" by Bob Geroch (1970b). There, it is emphasized that a suitable definition of a "singularity" within the context of general relativity must depend crucially on the definition of spacetime maximality. Geroch also suggests that the latter definition is far from clear given its sensitivity to a background collection of possible universes. To help clarify the situation, a number of "important and unsolved problems" are communicated – many of which remain open more than a half century later (Geroch, 1970b, p. 276). Essentially, the present work is an attempt to follow up on several of these lines of inquiry as best I can.

(ii) The "dirty open secret" discussed by John Earman (1995) in his book *Bangs, Crunches, Whimpers, Shrieks*. Attention is drawn to the fact that

those wishing to secure determinism for general relativity employ a type of circular reasoning: indeterminism is only avoided by excluding certain types of spacetime "holes" by fiat. A similar point applies to spacetime maximality which is also imposed by fiat but at a more fundamental level. I have endeavored to expose this even dirtier open secret in a variety of ways.

(iii) A table presented by David Malament (1977b) in his "Observationally Indistinguishable Space-Times" paper. The 13 rows are first-order spacetime properties and the three columns are second-order properties of collections of spacetimes (these concern whether certain relations hold among elements in the collection). Each "yes" or "no" entry corresponds to the result of one of 13x3=39 precise questions under consideration. This visual helps to illuminate a nuanced situation from many angles at once. Here, I have tried to adapt this idea to the topic of spacetime maximality; the entire book can be summarized in similar table on the final page. Although most of the 20x6=120 questions under consideration have been settled, there are also dozens of "?" entries. My hope is that, by drawing attention to these issues, even more progress can be made by others.

A number of people have helped this project along. David Malament has been an extraordinary sounding board for many years now and his suggestions and corrections have guided the inquiry significantly. I am very grateful to him. Thomas Barrett, Bob Geroch, and Jim Weatherall also provided useful comments on an earlier draft. Jeff Barrett, Gordon Belot, Eddy Chen, Erik Curiel, Juliusz Doboszewski, John Earman, Sam Fletcher, Hans Halvorson, Sergey Krasnikov, Ettore Minguzzi, Bryan Roberts, and Jan Sbierski deserve thanks as well for enlightening discussions and feedback on earlier work. I am fortunate to have such mentors and colleagues. Above all, I appreciate Meka and June for their love and support along the way. Midas too.

Introduction

It is helpful to think of (standard) general relativity as a particular collection of models of spacetime on the largest possible scale. Each model represents a possible universe compatible with the theory. Just as it would be a mistake to conclude that the earth is flat simply because it seems to be that way in one's immediate vicinity, so too would it be a mistake to conclude that the local structure of the universe mirrors its global structure. As George Ellis (2007, p. 1231) has put it: "The situation is like that of an ant surveying the world from the top of a sand dune in the Sahara desert. Her world model will be a world composed only of sand dunes – despite the existence of cities, oceans, forests, tundra, mountains, and so on beyond her horizon." For this reason, global spacetime structure within the context of general relativity can be quite permissive with respect to physical possibility. Pathologies such as "time travel" and spacetime "holes" of various kinds are not ruled out a priori. After all, how is one to "identify the space-times which are too pathological to be of physical interest unless he has at least examined the possibilities which can arise?" (Geroch, 1971b, p. 72)

Focusing on global spacetime structure will allow us to investigate the more foundational aspects of general relativity. For the most part, we will step away from the details of local physics in order to examine a number of qualitative features. Some of these have received a great deal of philosophical attention over the years such as the causal and singular structures of spacetime (Earman, 1995; Malament, 2012). These features will certainly figure prominently in what follows. But the central focus of this book will concern a somewhat less explored aspect of global structure of spacetime: its modal structure. Before jumping in, it might be useful to provide a brief sketch of the style of work carried out within the study of global structure more generally. Doing so will help to illuminate the distinctive nature of our exploration of the modality of spacetime.

As a mathematical subject, global structure is somewhat unusual in the sense that it is not characterized by a small number of key theorems from which corollaries easily flow. Instead, one finds "a large number of true statements, all of about equal utility" (Geroch and Horowitz, 1979, p. 214). In practice, this means that a type of careful collecting activity is encouraged; propositions which have limited significance when taken in isolation can be bundled together to shed light on deep questions. An example may be useful here. Consider a collection $\mathscr{U}$ of models of (standard) general relativity. Each element in the collection represents a possible universe. Naturally, any spacetime property corresponds to a particular subcollection of $\mathscr{U}$. Much of the work in global structure proceeds by first restricting attention to some collection $\mathscr{P} \subset \mathscr{U}$ of "physically reasonable" universes in order to tame the unruly background possibility space. The limited context then allows one to show that a certain spacetime property of interest must hold. For example, the famous "singularity theorems" of Hawking and Penrose (1970) come out as propositions of the form $\mathscr{P} \subset \mathscr{S}$ where the collection $\mathscr{S} \subset \mathscr{U}$ represents the property of being "singular" in some precise sense.

Let's now consider the modal structure of spacetime. Usually, spacetime properties are defined without reference to the background possibility space $\mathscr{U}$. Examples of such properties include those concerning the "causal" structure of spacetime as well as the local distribution and flow of matter in the form of "energy" conditions. There are a few exceptions, however. These are modal properties and spacetime "maximality" – the requirement that the universe be "as large as it can be" – is the paradigm example. A curious tension arises with respect to modal properties that is almost never appreciated in the literature. On the one hand, practitioners often pare down the background possibility space by restricting attention to some physically reasonable collection $\mathscr{P} \subset \mathscr{U}$ as in the singularity theorems mentioned above. On the other hand, they do not correspondingly define the modal properties of spacetime (e.g. maximality) relative to the reduced collection $\mathscr{P}$; the physically unreasonable background possibility space $\mathscr{U}$ is used instead (Hawking and Ellis, 1973; Wald, 1984). It is not clear why this is done, perhaps for reasons of simplicity. (As we shall see, non-standard definitions call for a great deal of book keeping.) But as a result of this mismatch, one can easily find situations where a possible universe in $\mathscr{P}$ is both (i) maximal relative to the collection $\mathscr{P}$ and yet (ii) not maximal relative to the collection $\mathscr{U}$. If standard practice is followed and spacetime maximality is assumed to hold under the usual definition (relative to the collection $\mathscr{U}$), then such a uni-

verse is ruled out on physically unreasonable grounds! Given the situation, it would seem appropriate to initiate a consistent, systematic exploration of spacetime maximality under various choices of background possibility spaces $\mathscr{P} \subset \mathscr{U}$. This, in a nutshell, is what this book is all about.

The presentation is divided into two parts. In Part I, the vast possibility space $\mathscr{U}$ is surveyed. Chapter 1 investigates the basic structure of spacetime. In order to help visualize this entity, diagrams are introduced first. Attention is then given to the possible universe of Minkowski – the setting for Einstein's special relativity. Within this context, any spacetime event can be characterized as a point in a four-dimensional Cartesian coordinate system. Its shape is therefore a higher dimensional generalization of the Euclidean plane. Moreover, the universe is "flat" in the sense that there is no spacetime curvature present. One simple way to get a grip on the universes permitted by general relativity is to keep the requirement of flatness in place but explore other spacetime shapes. Chapter 2 does this with a study of spacetime "manifolds" of various kinds. Chapter 3, brings curvature into the mix and a variety of spacetime "metrics" are considered. The connection between curvature and the distribution of flow of matter is also discussed.

Chapter 4 introduces the transfer of vectors between manifolds and the key notion of an "isometry" which is an isomorphism between possible universes. This will allow for a definition of invariant spacetime properties and a distinction between "global" and "local" varieties. Stepping back, the first four chapters build, from the ground up, a good portion of basic differential geometry in a reader friendly way: manifolds, metrics, and isometries. Other foundational topics such as derivative operators and arbitrary tensors are not rigorously defined but these are discussed informally. Various global properties become the focus of the next few chapters.

Chapter 5 concerns the causal structure of the universe and a hierarchy of spacetime properties is central. The lowest level rules out a type of "time travel" in which an event may causally influence itself. The highest level ensures that the universe is "deterministic" in the sense that physical situation at any one instant depends entirely on the physical situation at any other.

Chapter 6 considers spacetime "holes" which signal a type of incompleteness present in the universe. The singularity theorems are discussed and key examples such as black holes and the big bang are reviewed. A number of modal properties are then introduced to help classify spacetime holes. Spacetime maximality is the most basic among them.

Chapter 7 examines a hierarchy of spacetime asymmetry properties. The

highest level is characterized by the "Heraclitus" demand that no distinct pair of spacetime events have the same local structure. The Heraclitus asymmetry property proves useful in the extended investigation of spacetime maximality to follow.

Part I provides dozens of ways to slice and dice the possibility space $\mathscr{U}$ into subcollections $\mathscr{P} \subset \mathscr{U}$ of physical significance. In Part II, the modal property of spacetime maximality is investigated relative to these various subcollections.

Chapter 8 considers a plurality of definitions of spacetime maximality. For any $\mathscr{P} \subseteq \mathscr{U}$, a universe in the subcollection counts as $\mathscr{P}$-maximal if it is "as large as it can be" relative $\mathscr{P}$. It was conjectured by Geroch (1970b) that some reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$ are such that a universe is $\mathscr{P}$-maximal if and only if it is $\mathscr{U}$-maximal. Although some important cases remain open, it is shown that the conjecture fails in almost all contexts.

Chapter 9 explores the principles of Leibnizian metaphysics used to motivate the imposition of spacetime maximality. The linchpin for such a position is a theorem due to Geroch (1970b): any universe in $\mathscr{U}$ is either maximal or can be extended to be maximal. Although analogous statements remain true for some reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$, one finds that others come out as false. In such cases, Leibnizian metaphysics faces significant difficulties in getting off the ground.

Chapter 10 takes up the question of whether, by using empirical observations and some form of induction, one can come to know that one's universe is maximal. Building on the results of Malament (1977b), we show that the prospects are unsurprisingly dismal no matter what background possibility space is considered. In an appendix, we also explore the epistemological situation in general relativity if the dogma of spacetime maximality were to hold with respect to Heraclitus asymmetry property. Curiously, we find a sense in which one can know which universe one inhabits within this context. We also explore of a type of "meta-maximality" which requires reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$ to be "as large as they can be" with respect to some second-order property.

Chapter 11 concerns the "stability" of spacetime maximality. It is often assumed that any physically significant property of spacetime must be stable – it must hold in all "nearby" universes (Hawking and Ellis, 1973). Although some limited stability results are available, $\mathscr{P}$-maximality turns out to be unstable for some collections $\mathscr{P} \subset \mathscr{U}$. Moreover, we emphasize that there are many open questions including the stability of the standard notion of

$\mathscr{U}$-maximality.

Chapter 12 investigates the notion of "determinism" within the context of general relativity which is connected to several different notions of spacetime maximality. A celebrated theorem due to Choquet-Bruhat and Geroch (1969) is considered which captures a sense in which general relativity is a deterministic theory: there exists a unique "maximal development" spacetime associated with any initial data. For uniqueness to hold, the result must presupposes a type of dynamical spacetime maximality relative to the collection $\mathscr{U}$. We emphasize that statements analogous to the Choquet-Bruhat and Geroch (1969) theorem can be false relative to various reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$. This poses a problem for the dynamical justification for the dogma of spacetime maximality. We also note that any such justification also depends crucially on the "cosmic censorship" conjecture of Penrose (1979) which states that, relative to some collection $\mathscr{P} \subset \mathscr{U}$ of "physically reasonable" universes, any initial data (assumed to maximal in an appropriate sense) must evolve uniquely into a $\mathscr{P}$-maximal universe. But given the many questions that remain unsettled concerning cosmic censorship, the status of the dogma of spacetime maximality also remains murky.

Chapter 13 concerns the maximality properties of certain "branching" (i.e. non-Hausdorff) universes. Because such models of general relativity are non-standard, the background possibility space $\mathscr{U}$ is expanded in various ways rather than reduced. After a brief walk on the wild side, we explore and extend the foundational work of Clarke (1976) to show senses in which some branching universes can be domesticated so as to be compatible with the maximality dogma.

Chapter 14 provides a comprehensive review of the book. From the work primarily done in Part I, we identify twenty different possibility spaces $\mathscr{P}$ of physical interest. These concern the first-order local, causal, asymmetry, and branching properties of spacetime. In Part II, we identify six second-order conditions on such possibility spaces $\mathscr{P}$ that, if satisfied, speak in favor of the $\mathscr{P}$-maximality of spacetime. These second-order conditions mirror foundational results and conjectures of standard general relativity such as: the Geroch (1970b) theorem showing the existence of maximal spacetimes, the Choquet-Bruhat and Geroch (1969) theorem concerning determinism, and the cosmic censorship conjecture of Penrose (1979). Twenty first-order properties times six second-order properties gives 120 precise questions to consider. The questions are not all of equal importance but the end tally does give a sense a the balance of evidence for and against the maximality

of spacetime. Of the 120 question, 25 results speak in favor spacetime maximality, 56 speak against it, and 39 are still open. It is not at all clear that the universe is as large as it can be.

Part I

Possible Universes

Chapter 1

Spacetime

1.1 Introduction

In this chapter, we introduce the basics of relativistic spacetime. We start by conveying some of its qualitative features via a number diagrams. These pictures will help us to visualize the universe and its happenings from a four-dimensional point of view. After we get accustomed to the new spacetime framework, the focus then turns to the possible universe of Minkowski – the setting for Einstein's special relativity. An elementary mathematical formalism is gently introduced somewhat informally. Using a number of examples, we explore the curious way in which both "time" and "space" are understood to be relative to the observer. We also highlight the "metric" structure of Minkowski spacetime which is observer independent.

1.2 Diagrams

One can think of spacetime as a collection of events with some additional structure that specifies how the events are related. One's birth and one's death are events. The moon landing is also an event. But July 20, 1969 is not an event. And the moon is not an event. Experience seems to tell us that any event can be characterized by four numbers: one temporal coordinate t and three spatial coordinates x, y, z. Accordingly, the local structure of spacetime resembles a four-dimensional Cartesian coordinate system. Diagrams can help us "see" this spacetime structure. Because our brains are not good at visualizing four-dimensional entities, we will need to suppress one or two

spatial dimensions when representing spacetime diagrammatically. Consider a spacetime diagram of the moon landing for example (see Figure 1.1).

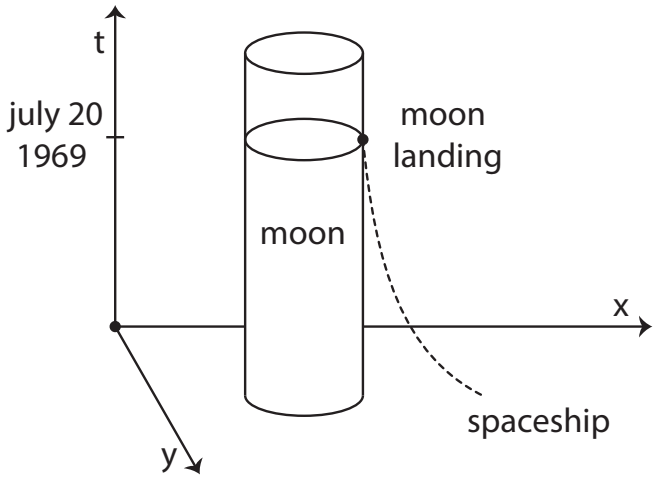

Figure 1.1: A spacetime diagram of the moon landing event. The world-line of the spaceship and the world-tube of the moon are depicted.

Following a long tradition, the time axis t is vertical with the up arrow pointing in the future direction. Two spatial dimensions x and y are also depicted. Because the spatial dimension z is suppressed, the moon at a given time is a two-dimensional disk instead of a three-dimensional sphere. We can think of each of the moon-at-a-time disks as being stacked like cards along the t axis. The result is that the moon is a three-dimensional tube in the diagram. Now for the spaceship. At any particular time, it is represented as a point (given that it is so much smaller than the moon). When we stack all of the spaceship-at-a-time points, the result is the smooth curve. We say the future-directed path an object takes through spacetime is its **world-line** (as in the case of the spaceship) or a "world-tube" (as in the case of the moon). We see that as time gets closer to July 20, 1969, the world-line of the spaceship gets closer to the world-tube of the moon. The event of the moon landing is represented as the dot on the outside surface of the world-tube of the moon.

Consider another example: a race between a tortoise and the hare (see Figure 1.2). The race starts at the same event at the lower, left-hand side of the spacetime diagram. The runners then move toward the finish line along the x axis. At any instant, the finish line is a one-dimensional string but,

over time, the collection of all such strings form a two-dimensional "world-sheet" in the diagram. During the race, three events are depicted along the world-line of the hare and corresponding velocity vectors are represented by the three dotted arrows. At the first such event early in the race, the hare is moving very fast which is indicated by a near horizontal arrow. Roughly halfway through, he decides to take a nap and stops moving altogether. This is represented by the vertical second arrow. After waking from the nap, the hare once again moves quickly. The near horizontal third arrow once again captures this state of affairs. Meanwhile, the tortoise begins the race at a slow and steady pace. An event is depicted along his world-line exactly where he passes the sleeping hare. At this point, the velocity of the tortoise is represented by an arrow which is neither vertical nor nearly horizontal – it is somewhere in between. Because this particular velocity is maintained all along the world-line, the tortoise reaches the finish line just ahead of the rushing hare. It is important to appreciate that the tortoise and the hare both move in a straight line in space. The curve in the world-line of the hare simply corresponds to his changing velocity through time. On the other hand, the world-line of the tortoise appears straight because of his constant velocity throughout.

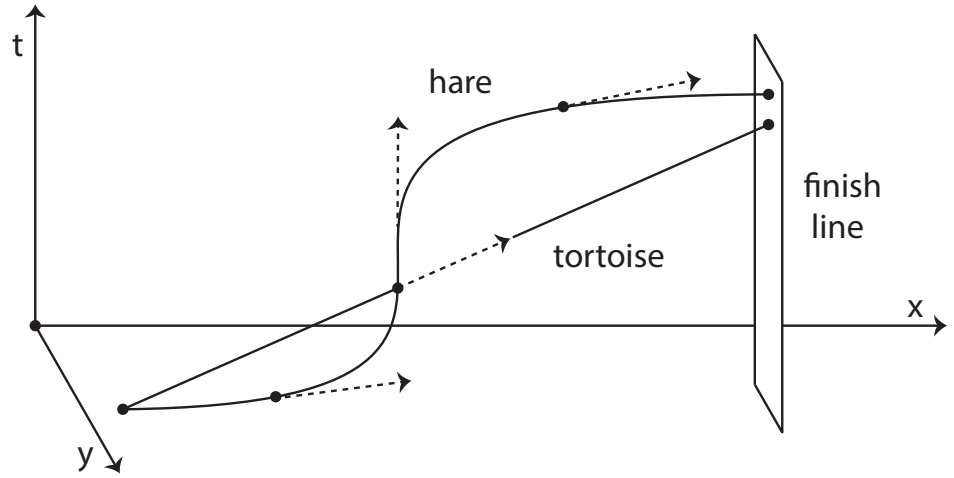

Figure 1.2: A race between the tortoise and the hare. The world-lines of the runners along with a few velocity vectors are depicted.

Spacetime diagrams take some getting used to. But with a little practice, one can develop the ability view the universe and its happenings from a dif-

ferent angle. A door opening is depicted in Figure 1.3. Or consider another example: what is the spacetime diagram of the entire human race? After a bit of thought, it becomes clear that it must look be something like an enormous tree. The world-line of each person is a little branch of the tree that is joined to some other branch – the world-line of the person's biological mother – at the event of the person's birth. At any given time, humanity is a collection of disconnected bodies in three-dimensional space. In four-dimensional spacetime, however, humanity is a single entity. From this perspective, the aphorism "we are all connected, man" is not a mere metaphor.

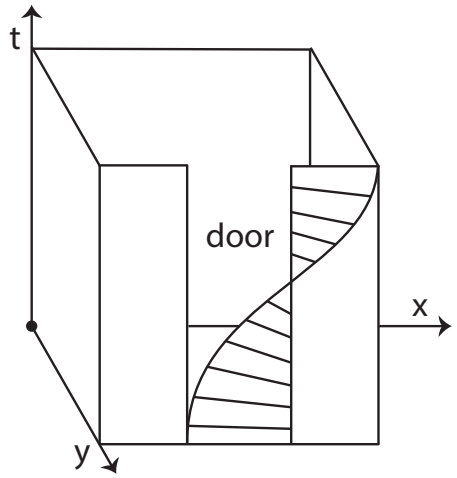

Figure 1.3: A door opens.

Our next spacetime diagram example will help in the transition to the Minkowski universe. Consider a strike of lightning (see Figure 1.4). Just after the event, light propagates radially outward in all spatial directions. The uniform speed of light has the effect of producing a cone shape in spacetime. The thundering sound of the strike creates a similar structure. The "sound cone" fits inside the "light cone" because the speed of light is so much faster than that of sound. If you are nearby, you will see the lightning before hearing the thunder. How much time passes between these two event will depend on how far away you are from the strike; the further away, the more time will separate the two events.

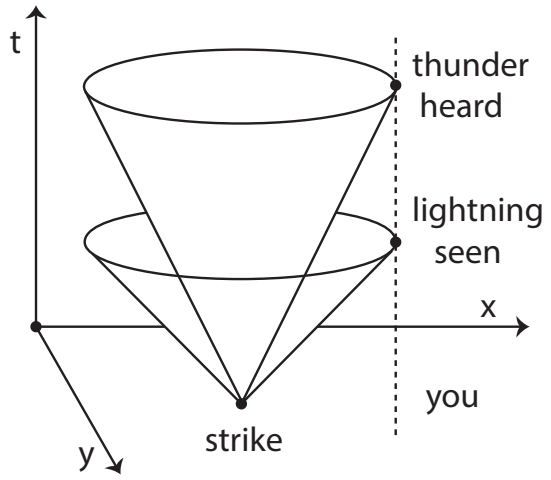

Figure 1.4: The event of a lightning strike along with the associated light and sound cones. You see lighting and hear thunder at the indicated events.

1.3 Minkowski Universe

Einstein's special relativity came in 1905. A few year later, the theory was given a geometric formulation by Hermann Minkowski (1908). Let us now consider one (standard) way to present this possible universe. As before, any spacetime event can be characterized as some point $p = (t, x, y, z)$ in a four-dimensional Cartesian coordinate system. The shape of the Minkowski universe is therefore $\mathbb{R}^4$ which is just a higher dimensional generalization of the Euclidean plane.

The idea that "nothing can travel faster than light" is central in relativity theory. Photons travel at the speed of light while objects with non-zero mass must travel more slowly. At each event in $\mathbb{R}^4$, there is a double "light cone" structure that marks this cosmic speed limit; one lobe corresponds to the future and the other to the past. We have already explored the future lobe of the light cone in the lightning strike example just considered (recall Figure 1.4). From the event of the strike, a massive object can only travel to spacetime points found inside the light cone region depicted. This is the future lobe. Similarly, the past lobe of the light cone (not shown in the diagram) extends in the other direction and demarcates the region of spacetime from which a massive object could have traveled to reach the lightning strike.

Now consider the spacetime diagram given in Figure 1.5. A spaceship travels from event p to event q and then on to event r. At each of these points, the velocity vector of the spaceship (dotted arrow) is found inside the light cone. We call such vectors **timelike**. In addition, we see a photon traveling from event q to event s. The velocity vector of the photon is found on the boundary of the light cone at events q and s. We call such vectors **null**. Notice that null vectors are depicted as arrows with a 45 degree angle: light travels one unit of distance per one unit of time. In what follows, we will stick to using units of years and light years. Finally, vectors which fall neither in nor on the boundary of the light cone we call **spacelike**. Velocity vectors can never be spacelike and so no such vectors are depicted in the diagram. But this notion will be useful later on in defining other geometrical objects of physical significance. Note that since the spaceship at p is contained in the past light cone of s, it would have been possible for it to travel from the former event to the later. But at q it is too late for the spaceship to travel to s; only a photon is fast enough to do this. Finally, nothing – no spaceship, not even light itself – can travel from r to s (or vice versa).

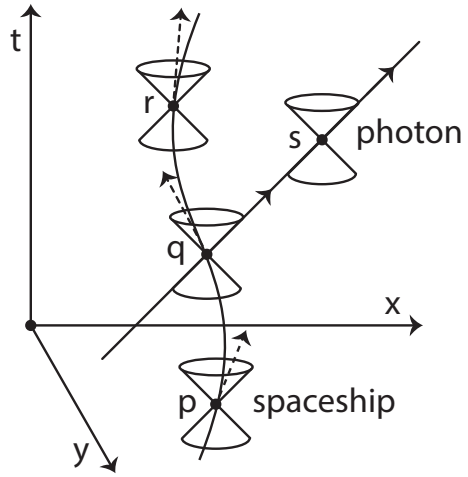

Figure 1.5: Light cone structures at various events. The world-lines of a spaceship and a photon are depicted along with their velocity vectors.

Any **curve** in Minkowski spacetime (e.g. a world-line) is a smooth function $\lambda : I \to \mathbb{R}^4$ where I is some interval of the real numbers such as $(0, 1)$ or $\mathbb{R}$ itself. Each point $s \in I$ is mapped to an event $\lambda(s) \in \mathbb{R}^4$ in Minkowski spacetime. A **tangent vector** $v = [v_t, v_x, v_y, v_z]$ is associated with each

event on $\lambda(s)$ which tracks the direction of the curve in spacetime at that location. Here square brackets are used to distinguish vectors from points since both are members of $\mathbb{R}^4$. We are already familiar with tangent vectors in the case of world-lines: they are just the velocity vectors. Suppose there is a spaceship whose world-line is given by the curve $\lambda : \mathbb{R} \to \mathbb{R}^4$ defined by $\lambda(s) = (s, \sin(s)/2, 0, 0)$. The parameter time s coincides with t. As $s = t$ passes, the position in the x direction oscillates between $-1/2$ and $1/2$ while the position in the y and z directions does not change. The motion is somewhat similar to that of the spaceship in Figure 1.5.

One can easily calculate the tangent vector at each point. We first break $\lambda(s)$ down into four component functions $\lambda_t(s) = s$, $\lambda_x(s) = \sin(s)/2$, $\lambda_y(s) = \lambda_z(s) = 0$. We then use basic calculus to differentiate each of these component functions with respect to s to find $\lambda_t'(s) = 1$, $\lambda_x'(s) = \cos(s)/2$, and $\lambda_y'(s) = \lambda_z'(s) = 0$. Here, the prime symbol indicates that the function has been differentiated in accordance with standard calculus notation. We then slot these derivatives in as components of a vector $[\lambda_t'(s), \lambda_x'(s), \lambda_y'(s), \lambda_z'(s)] = [1, \cos(s)/2, 0, 0]$. This is the tangent vector of our curve as a function of s which will be often be written $\lambda'(s)$. Since we are using units of years and light years, we see that $\lambda'(s)$ is found inside the light cone at each point and is therefore timelike. In the natural way, a curve is timelike if all of its tangent vectors are timelike and similarly for null and spacelike curves. So $\lambda(s)$ counts as a timelike curve.

We say a curve in spacetime is a **geodesic** if it is "as straight as possible" in a sense we will explore a later on. An observer along a geodesic world-line experiences no acceleration. In the standard presentation of Minkowski spacetime that we have been using, the images of geodesics appear as straight lines or portions thereof. It turns out there is a simple formula for the "length" of timelike and null geodesics. In the timelike case, this length represents the elapsed time measured by a clock along the geodesic. The formula to determine length amounts to a slight tweak of the Pythagorean theorem. Now consider a timelike or null geodesic $\lambda : I \to \mathbb{R}^4$ in the Minkowski universe which runs from the event $p = (p_t, p_x, p_y, p_z)$ to the event $q = (q_t, q_x, q_y, q_z)$. Let $\Delta t = p_t - q_t$ and similarly for Δx, Δy, and Δz. The **length** of the geodesic λ is denoted by $\|\lambda\|$ and is given by the following formula which we will call the **(Minkowskian) interval**.

$$\|\lambda\|^2 = \Delta t^2 - \frac{1}{c^2}(\Delta x^2 + \Delta y^2 + \Delta z^2)$$

Here, c is the speed of light. Since we are using units of years and light years, we have $c = 1$. But because of the $1/c^2$ term, the units of $\|\lambda\|$ come out as years. When a geodesic is timelike, we call its length **elapsed time**.

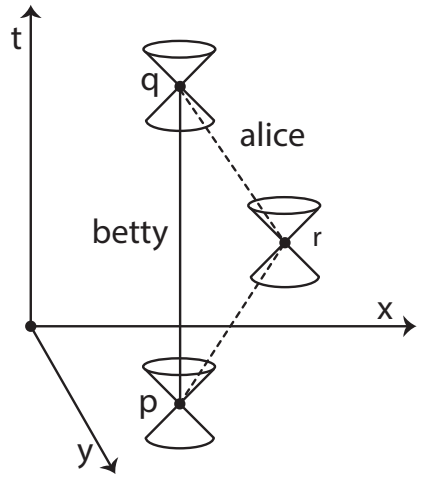

Figure 1.6: The world-lines of Alice and Betty from p to q. The elapsed times are 6 and 10 years respectively.

1.4 Relative Time

A mind bending consequence of interval formula this: the elapsed time between two events depends on the path taken between them through spacetime. Let us take a closer look (see Figure 1.6). Consider twins Alice and Betty at the event $p = (0, 5, 5, 0)$. Suppose they want to attend a tea party at event $q = (10, 5, 5, 0)$. Betty travels along the timelike geodesic λ_{pq} which runs from p to q. Using the interval formula, we find that this geodesic has an elapsed time of $\|\lambda_{pq}\| = 10$ years. This follows since $c = 1$ light year/year, $\Delta t = 10$ years, and $\Delta x = \Delta y = \Delta z = 0$ light years. Betty's watch has measured ten years between events p and q. This makes intuitive sense and seems to accords with our everyday experiences. But Alice has found another way to travel from p to q. She goes from p to the event $r = (5, 9, 5, 0)$ along the timelike geodesic λ_{pr}. She turns around and goes from r to q along the timelike geodesic λ_{rq}. What is $\|\lambda_{pr}\|$? We know $\Delta t = 5$ years, $\Delta t = 4$ light years, and $\Delta y = \Delta z = 0$ light years. Using the interval formula, we find

that $\|\lambda_{pr}\| = 3$ years. In a similar way, we find that $\|\lambda_{rq}\| = 3$ years. So zig-zagging from p to r to q, Alice's watch has measured only 6 years in total. She is now four years younger than Betty when the twins meet up at the tea party at q. Curious!

An even stranger state of affairs arises when we examine the behavior of null geodesics. Consider again the lightning strike example (recall Figure 1.4). What is the length along the geodesic running from the event of the strike to the event at which the observer sees the lightning? From the diagram, we see that $\Delta y = \Delta z = 0$. Since $c = 1$ light year per year, this means that whatever Δt and Δx happen to be (this is not indicated in the diagram), it must be the case that $\Delta t = \Delta x$. It follows from the interval formula that the spacetime length from the event of the strike to the event at which the observer sees the lightning must be zero. Curiouser!

Useful as it is, the interval formula is limited in that it only applies to timelike and null geodesics. How can one determine the elapsed time along a smooth but curvy world-line? A Minkowskian "metric" is similar to the interval formula but assigns a length to vectors at each event instead of geodesics. In order to determine the length of a smooth timelike or null curve, the lengths of the velocity vectors at each point are "added up" along the curve using integral calculus. Let's explore this idea in more detail. The **Minkowskian metric** η assigns a real number $\eta(v, w)$ to any pair of vectors $v = [v_t, v_x, v_y, v_z]$ and $w = [w_t, w_x, w_y, w_z]$ at any point according to the following rule.

$$\eta(v, w) = v_t w_t - v_x w_x - v_y w_y - v_z w_z$$

Here we have suppressed a $1/c^2$ term since we are working in units where $c = 1$. The (squared) **length** $\|v\|$ of a vector v is just $\eta(v, v)$. So, for example, the length of the vector $v = [3, 1, 2, 0]$ at any point is $\|v\| = \eta(v, v) = 3^2 - 1^2 - 2^2 - 0^2 = 4$. One finds that any timelike vector has positive length; any null vector has zero length; and any spacelike vector has negative length. In this way, the light cone structure at each event is encoded in the metric η. To give a sense things, a few timelike vectors of unit length are depicted in Figure 1.7. Such vectors form a hyperboloid structure at each point.

Let $\lambda : I \to \mathbb{R}^4$ be any timelike or null curve. For each $s \in I$, the tangent vector at $\lambda(s)$ is $\lambda'(s)$. One can integrate the quantity $\sqrt{\|\lambda'(s)\|}$ along the curve to calculate its total length (or elapsed time in the case of timelike curves) which is denoted $\|\lambda\|$. Let's again consider an example of the Alice-Betty twin effect but now only make use of smooth timelike curves

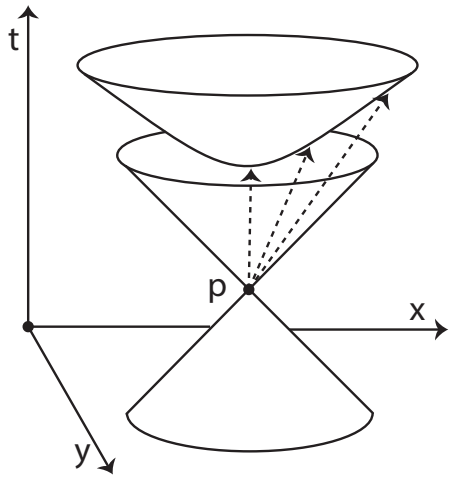

Figure 1.7: At p, the Minkowskian metric η assigns a unit length to all of the timelike vectors depicted. Such vectors form a hyperboloid structure.

– no zig-zags allowed!

Before we get going, it will prove useful to have a basic understanding of the smooth hyperbolic sine and cosine functions $\sinh : \mathbb{R} \to \mathbb{R}$ and $\cosh : \mathbb{R} \to \mathbb{R}$ respectively (see Figure 1.8). Two remarkable properties will be used often in what follows: (i) for any $x \in \mathbb{R}$, we have $\cosh^2(x) - \sinh^2(x) = 1$ and (ii) the derivative of $\sinh(x)$ is $\cosh(x)$ and the derivative of $\cosh(x)$ is $\sinh(x)$. From the diagram, we also see that $\sinh(-x) = -\sinh(x)$ and $\cosh(-x) = \cosh(x)$.

Now consider Alice. Let her world-line be the (non-geodesic) timelike curve $\lambda : (-3, 3) \to \mathbb{R}^4$ given by $\lambda(s) = (\sinh(s), \cosh(s), 0, 0)$ (see Figure 1.9). Differentiating each component of $\lambda(s)$ with respect to s gives the velocity vector $\lambda'(s) = [\cosh(s), \sinh(s), 0, 0]$ at every point along the curve. The (squared) length $\|\lambda'(s)\|$ is given by $\eta(\lambda'(s), \lambda'(s)) = \cosh^2(s) - \sinh^2(s) = 1$. We then integrate $\sqrt{\|\lambda'(s)\|} = 1$ from $s = -3$ to $s = 3$ to find the Alice's elapsed time is $\|\lambda\| = 6$ years. Notice that the curve starts at the point $p = (\sinh(-3), \cosh(-3), 0, 0)$ and ends at the point $q = (\sinh(3), \cosh(3), 0, 0)$. One can get a qualitative grip on things by noting that $\sinh(3) \approx \cosh(3) \approx 10$. If Betty's world-line is the geodesic γ from p to q, her elapsed time is easy to calculate using the interval formula: $\|\gamma\|^2 = [\sinh(-3) - \sinh(3)]^2 - [\cosh(-3) - \cosh(3)]^2 - 0 - 0$. It follows that $\|\gamma\| = 2\sinh(3) \approx 20$ years as one might expect. In the original situation,

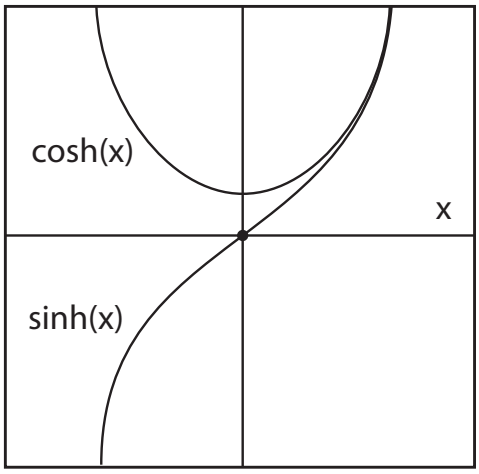

Figure 1.8: The functions $\sinh(x)$ and $\cosh(x)$ have two remarkable properties: (i) $\cosh^2(x) - \sinh^2(x) = 1$ and (ii) the derivative of $\sinh(x)$ is $\cosh(x)$ and the derivative of $\cosh(x)$ is $\sinh(x)$.

Alice experiences a radical and instantaneous change in velocity when turning around half-way through. The second iteration here is similar but we have effectively "smoothed out" this acceleration at the turnaround point.

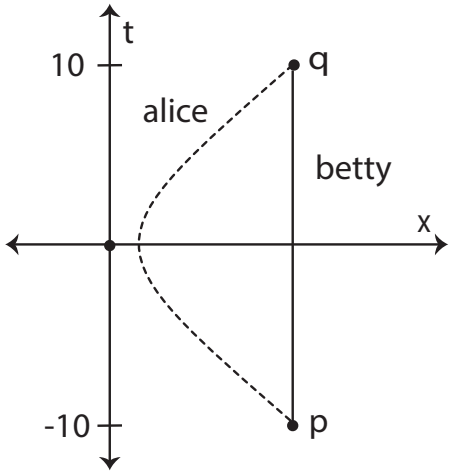

Figure 1.9: The world-lines of Alice and Betty from p to q. The elapsed times are 6 years and $2\sinh(3) \approx 20$ years respectively.

Consider a general result along these lines. Let p and q be any events in Minkowski spacetime that can be connected by a timelike world-line. Among all smooth timelike curves connecting the events, the geodesic between them has the longest elapsed time. Moreover, for every $\epsilon > 0$, there is a timelike curve connecting the events with an elapsed time less than ϵ. A trip from p to q can take less than a year. Or less than a second. One can see this must be true since one can approximate a zig-zag null geodesic with zero length arbitrarily closely with a timelike curve.

1.5 Relative Space

We have seen a sense in which the notion of "time" is relative to the observer. Now let us briefly consider a sense in which "space" is as well. In addition to determining the lengths of vectors, the metric η also keeps track of the angles between them. Given vectors v and w at a point, if it is the case that $\eta(v, w) = 0$, then we say the vectors are **orthogonal**. It is immediate that a null vector is orthogonal to any scalar multiple of itself. Naturally, the timelike vector $[1, 0, 0, 0]$ pointing in the t direction is orthogonal to scalar multiples of the vectors $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, and $[0, 0, 0, 1]$ that point in the x, y, and z directions respectively.

Let $\lambda : I \to \mathbb{R}$ be a timelike geodesic with tangent vector v at event p. Let q be any other event. We say p and q are **simultaneous** relative to λ if the geodesic from p to q has a tangent vector w at p that is orthogonal to v. A timelike geodesic λ through event p determines an associated **simultaneity slice**: the collection of all events simultaneous with p relative to λ. Suppose your world-line is the geodesic λ with tangent $v = [1, 0, 0, 0]$ at event $p = (2, 2, 2, 0)$. Intuitively, we find that the simultaneity slice relative to you at event p is the three-dimensional surface given by the constraint $t = 2$. This region is depicted in Figure 1.10 which contains the event $q = (2, 0, 2, 0)$. You judge p and q to be simultaneous.

Now consider your friend's world-line which is the geodesic γ that also runs through event p but with tangent vector $w = [2, 1, 0, 0]$. What is the simultaneity slice at p relative to γ? To get a sense of things, let's first find a non-zero vector $u = [u_t, u_x, u_y, u_z]$ at p which is orthogonal to w. The condition $\eta(w, u) = 0$ implies that $2u_t = u_x$. So we see, for example, that the vector $u = [1, 2, 0, 0]$ at p is orthogonal to $w = [2, 1, 0, 0]$ there. Next, we can find a geodesic through p with tangent vector u there. Starting from

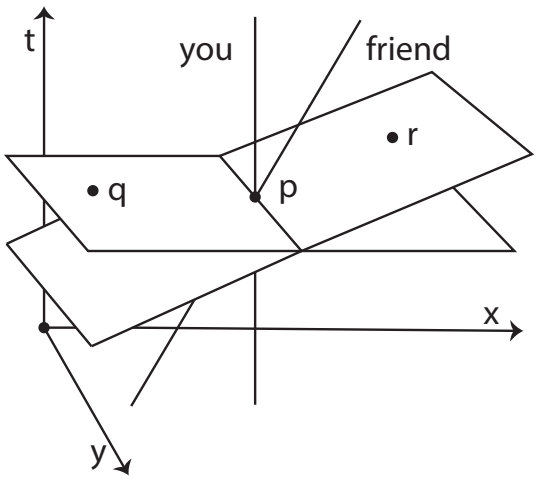

Figure 1.10: Simultaneity slices at p relative to both you and your friend. You judge p and q to be simultaneous; she judges p and r to be so.

the point $p = (2, 2, 2, 0)$ and using the vector $u = [1, 2, 0, 0]$ as a guide, we move up one unit in the t direction and over two units in the x direction to arrive at the point $r = (3, 4, 2, 0)$. This means the geodesic from p to r has the features we are after: it runs through p with tangent vector u there. It follows from all of this that the event r is orthogonal to p relative to your friends's world-line γ (see Figure 1.10). From here, it is not too difficult to see that extending the geodesic from p to r out as far as it can go gives the line $t = (x + 2)/2$. Any point on this line is orthogonal to p according to your friend. Indeed, since the y and z components of the vector u are zero, we see that any point on the three-dimensional surface given by the constraint $t = (x + 2)/2$ is orthogonal to p according to your friend. This is the simultaneity slice at p relative to γ.

Stepping back, we have now seen senses in which both "time" and "space" are relative to the observer. In contrast, we emphasize here that the spacetime metric is observer independent. There are no disagreements concerning the light cone and geodesic structures. This means that the speed of light and the notion of acceleration is the same for everyone. We will explore these ideas in greater depth in Chapter 4 when we discuss the "symmetries" and "invariant properties" of spacetime.

1.6 Conclusion

We have come a long way already. We have introduced the notion of four-dimensional spacetime and explored some of the basic features of special relativity. This includes the curious ways in which "time" and "space" are observer dependent. Now that we have a better sense of the spacetime metric and its invariant nature, we are in a position to characterize formally the standard presentation of **Minkowski spacetime** that we have been working with. It is an ordered pair $(\mathbb{R}^4, \eta)$ where $\mathbb{R}^4$ is the background collection of spacetime events (with an associated topological structure we will explore in the next chapter) and η is the Minkowskian metric defined at each event. In what follows, we continue to learn more about the properties of Minkowski spacetime (e.g. its symmetries). We will also investigate ways of generalizing this structure so as to generate the wide variety of spacetimes permitted by general relativity.

Chapter 2

Shapes

2.1 Introduction

As we have seen, the Minkowski universe $(\mathbb{R}^4, \eta)$ is a bit strange. Even so, it is the vanilla model of general relativity. One way to start exploring some of the other flavors is to retain the Minkowskian metric η but consider some spacetime shapes other than $\mathbb{R}^4$. We start with some basic topology which will allow us to precisely characterize various shapes in the most general setting. We then work our way to the formal notion of a "spacetime manifold" which has the local structure of $\mathbb{R}^4$ and is smooth in the appropriate sense. We also define vectors within this context. This will allow us to consider the Minkowskian metric η on spacetime manifolds other than $\mathbb{R}^4$. We close with a focus on a particular example – a locally Minkowskian model of general relativity with a cylindrical shape.

2.2 Topology

Let X be any set whatsoever. We can endow X with a shape by associating with it a collection τ of subsets of X that satisfy certain properties. We say τ is a **topology** on X if (i) both the empty set $\varnothing$ and X itself are in τ, (ii) an arbitrary (finite or infinite) union of members of τ is also in τ, and (ii) a finite intersection of members of τ is also in τ. If τ is a topology for a set X, the ordered pair (X, τ) is called a **topological space** and elements of τ are the **open** subsets sets of X. A set $C \subseteq X$ is **closed** if its complement $X - C$ is open. An (open) **neighborhood** of a point $p \in X$ is an open set

$O \subseteq X$ such that $p \in O$. Intuitively, a neighborhood of a point p is a set which contains points in X that are "close" to it.

It is easy to find a topology for any set X. One can let $\tau = \{\varnothing, X\}$ for example. This is the **trivial** topology on X. One can also let τ be the collection of all subsets of X (i.e. the power set of X). This is the **discrete** topology on X. Usually, topologies of interest are in between these two extremes. We will consider a number of examples as we go along. To help us, here we introduce two ways of making new topologies from old ones. Given a topological space (X, τ), the **subspace** topology for a set $A \subset X$ is the collection of all subsets $A \cap O \subseteq O$ where $O \subseteq X$ is an open set in τ. If (X, τ) and (Y, σ) are topological spaces, the **product** topology on the set $X \times Y$ is the collection of all subsets of $X \times Y$ which can be expressed as unions of sets of the form $O \times U$ with $O \in \tau$ and $U \in \sigma$. We will consider many examples of these topologies in what follows.

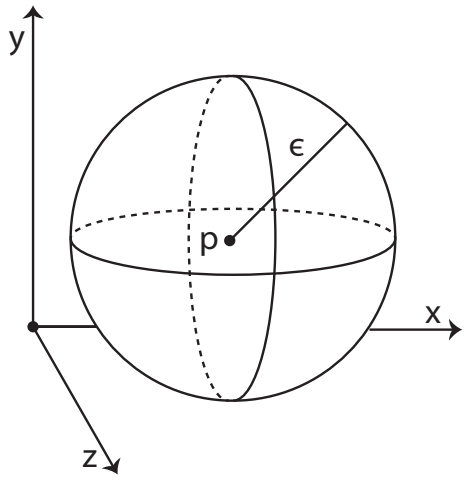

Figure 2.1: A ball centered at the point p with radius ϵ.

The shape of the Minkowski universe is given by the set $\mathbb{R}^4$ with its "standard" topology. Consider the set $\mathbb{R}^n$ where n is any positive integer. Let $p = (p_1, ..., p_n)$ be any point in $\mathbb{R}^n$ and let ϵ be any positive real number. An (open) **ball** $B \subset \mathbb{R}^n$ with radius ϵ centered at the point p is defined as the collection of all points $(x_1, ..., x_n) \in \mathbb{R}^n$ such that the quantity $\sqrt{(p_1 - x_1)^2 + ... + (p_n - x_n)^2}$ is less than ϵ. A diagram of a ball in $\mathbb{R}^3$ centered at the point p with radius ϵ is depicted in Figure 2.1. A one-dimensional ball with radius ϵ at a point p in $\mathbb{R}$ is just the interval $(p - \epsilon, p + \epsilon)$. An

open subset O of $\mathbb{R}^n$ is one such that, for each point $p \in O$, there is a ball B centered at p with some radius ϵ (however small) such that $B \subseteq O$. The **standard** topology on $\mathbb{R}^n$ is the collection of all of these open subsets of $\mathbb{R}^n$. One can verify that this collection satisfies conditions (i)-(iii) in the topology definition. We will assume the standard topology on $\mathbb{R}^n$ throughout and suppress explicit reference to it.

Before moving on, we note a few basic definitions here that will be needed later on. (Beyond the basics, the reader is referred to Willard (1970). For some fun practice with the basics, see Geroch (2013).) Let $(X, \mathscr{T})$ be any topological space and let $A \subseteq X$. The **closure** of A is the intersection of all closed sets containing A. The **interior** of A is the union of all open sets contained in A. The **boundary** A is defined as the closure of A with the interior removed. The closure of A is always closed; A is a subset of the closure of A; and the two sets are identical if A is closed. Similarly, the interior of A is always open; the interior of A is a subset of A; and the two sets are identical if A is open. Finally, the boundary of A is always closed and the closure of A is equal to the union of boundary of A and the interior of A. As a simple example, consider the set $A = (-1, 1]$ in $\mathbb{R}$ with its standard topology. We see that the closure of A is $\{-1\} \cup A$, the interior of A is $A - \{1\}$, and the boundary of A is $\{-1, 1\}$.

What are the spacetime shapes compatible with general relativity? These are the spacetime "manifolds" which are topological spaces having a "local structure" of $\mathbb{R}^n$ that are also "smooth" in the appropriate sense. One usually requires that spacetime manifolds are also "connected" in an intuitive way and satisfy the "Hausdorff" condition ensuring that distinct events are properly separated from each other. We will slowly build up to all of these ideas in what follows.

2.3 Continuity

Consider a pair of topological spaces (X, τ) and (Y, σ) and let $f : X \to Y$ be any function. For any subset $A \subset X$, we define its **image** $f[A]$ as the set of all points $f(p) \in Y$ such that $p \in A$. Similarly, for any set $A \subset Y$, we define its **preimage** $f^{-1}[A]$ as the set of all points $p \in X$ such that $f(p) \in A$. We say the function $f : X \to Y$ is **continuous** if, for each open set $O \subseteq Y$, its preimage $f^{-1}[O]$ is an open subset in X. To get a grip on this notion of continuity, consider an example timelike curve $\lambda : \mathbb{R} \to \mathbb{R}^4$ in

Minkowski spacetime. Define it by setting $\lambda(s) = (s, 1, 1, 0)$ for all $s < 2$ and $\lambda(s) = (s, 3, 1, 0)$ for $s \geq 2$ (see Figure 2.2). As one would expect, the curve is not continuous. Consider the ball B centered on the point $(2, 3, 1, 0)$ with unit radius. Since any ball is necessarily open, B is open. But the preimage $\lambda^{-1}[B]$ is the interval $[2, 3)$ in $\mathbb{R}$ and this interval is not open in since there is no one-dimensional ball (open interval) around the point $s = 2$ which fits inside the interval $[2, 3)$. Since B is open in $\mathbb{R}^3$ but $\lambda^{-1}[B]$ is not open in $\mathbb{R}$, the curve is not continuous.

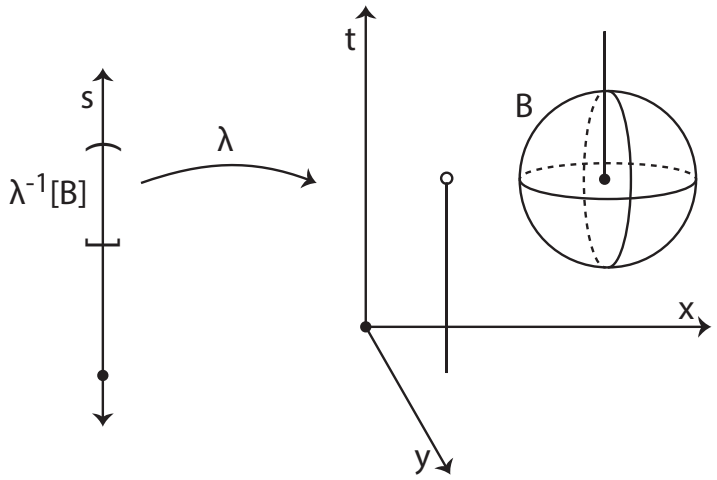

Figure 2.2: A timelike world-line λ in the Minkowski universe. It is not continuous since the open ball B has a preimage $\lambda^{-1}[B]$ that is not open.

A continuous function $f : X \to Y$ from a topological space (X, τ) to a topological space (Y, σ) need not preserve basic topological notions. For example, an open set $O \subseteq X$ may not have an open image $f[O]$ in Y. Just consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1$ for all $x \in \mathbb{R}$. It must be continuous since any open set $O \subseteq \mathbb{R}$ must either contain 1 or not. If so, $f^{-1}[O] = \mathbb{R}$ which is open; if not, $f^{-1}[O] = \varnothing$ which is also open. But the continuous function f does not map open sets to open sets. To see this, just consider that $\mathbb{R}$ is open while $f[\mathbb{R}] = \{1\}$ isn't.

But stepping back, there is one basic topological notion that is preserved under continuous functions: that being a "compact" subset. This property is extremely important and captures an abstract sense of the "finiteness" or "boundedness" of a topological space. Let (X, τ) be a topological space and let $A \subseteq X$. A collection $\{O_i\}$ of open sets is an **open cover** for A if the

union of all of the O_i contains A as a subset. An open **subcover** of A is a subcollection of $\{O_i\}$ which is also an open cover of A. We say A is **compact** if every open cover of A has a subcover with only a finite number of elements.

In $\mathbb{R}^n$, one can show that a set A is compact if and only if it is (i) closed and (ii) "bounded" in the sense that $A \subset B$ for some ball B. This is the Heine-Borel theorem. Let's work through a pair of simple examples to better understand the role of conditions (i) and (ii). First, let A be the interval $(0, 1)$ in $\mathbb{R}$ that fails to be closed but is bounded since it is contained in the unit ball B centered at 0, i.e. the open interval open interval $(-1, 1)$. Now consider the open cover $\{O_i\}$ defined by setting O_i to the open interval $(1/(1+i), 1)$ for all positive integers i. Any finite subcollection of $\{O_i\}$ does not cover A which shows it is not compact. Now consider a second example: the closed set $A = \mathbb{R}$ which fails to be bounded. Let $\{O_i\}$ be the open cover defined by setting O_i to the open interval $(i, i+2)$ for all integers i. Any finite subcollection of $\{O_i\}$ does not cover A which shows it is not compact.

We now come to the foundational result mentioned above concerning the preservation of compactness by continuous functions. Let (X, τ) and (Y, σ) be topological spaces and let $f : X \to Y$ be a continuous function. If $A \subset X$ is compact, then the image $f[A]$ is also compact. Using the Heine-Borel theorem, we also have a useful corollary: If $f : X \to \mathbb{R}$ is a continuous function on a topological space (X, τ) and $A \subseteq X$ is compact, then $f[A]$ is closed and bounded. It follows that there will be points $a, b \in \mathbb{R}$ such that $f[A]$ is contained in the closed interval $[a, b]$ with $f(p) = a$ and $f(q) = b$ for some points $p, q \in A$ (see Figure 2.3). Another useful result is that the product $X \times Y$ of compact topological spaces (X, τ) and (Y, σ) must also be compact in the product topology.

We are now ready to spell out what it means to say that two topological spaces have the same structure. The topological structures (X, τ) and (Y, σ) are **homeomorphic** if there is a bijection (one-to-one and onto function) $f : X \to Y$ such that both f and its inverse f^{-1} are continuous. Such a bijection is called a **homeomorphism**. One finds that $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^m$ if and only if $n = m$. A somewhat counterintuitive result is this: an open ball in $\mathbb{R}^n$ with the subspace topology is homeomorphic to $\mathbb{R}^n$ itself. For example, consider open interval $(-1, 1)$ which is a one-dimensional ball in $\mathbb{R}$ centered at 0 with radius 1. Take this interval as a topological space in its own right by endowing it with the subspace topology from $\mathbb{R}$. The function $f : (-1, 1) \to \mathbb{R}$ defined by $f(x) = x/(1 - x^2)$ counts as a homeomorphism. The function "stretches" the interval to infinity without changing its topological

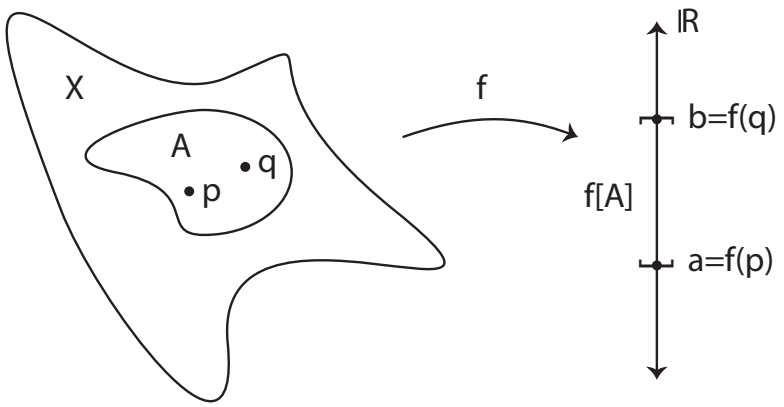

Figure 2.3: The continuous function f maps the compact set $A \in X$ to its image $f[A] \in \mathbb{R}$. This image is contained in the closed interval $[a, b]$ with $f(p) = a$ and $f(q) = b$ for some points $p, q \in A$.

structure.

2.4 Manifolds

We say a topological space (M, τ) is **locally** $\mathbb{R}^n$ if each point $p \in M$ has a neighborhood $O \subseteq M$ that is homeomorphic to some open subset of $\mathbb{R}^n$. It is trivial that $\mathbb{R}^n$ is locally $\mathbb{R}^n$. A sphere is a simple but non-trivial example. For any positive integer n, we define the n-dimensional **sphere** S^n with radius ϵ centered at the point $p = (p_1, ..., , p_{n+1})$ in $\mathbb{R}^{n+1}$ to be the collection of all points $(x_1, ..., x_{n+1})$ such that $\sqrt{(p_1 - x_1)^2 + ... + (p_{n+1} - x_{n+1})^2} = \epsilon$. As one would expect, we see that a sphere S^n is just the boundary of some open ball in $\mathbb{R}^{n+1}$. Unless otherwise flagged, we take S^n to be centered at the origin $\mathbb{R}^{n+1}$ with radius $\epsilon = 1$.

The **standard** topology on S^n (assumed throughout) is the subspace topology induced from $\mathbb{R}^{n+1}$. The sphere S^n is compact in this topology for all n. The unit sphere S^2 centered at the point $p = (2, 2, 2)$ is depicted in Figure 2.4. It is worth mentioning the points such as p that are contained inside the sphere are not part of the sphere itself. The "eastern hemisphere" O is the $x > 2$ region of the sphere which counts as an open set. The function

f defined by taking any point $(x, y, z) \in O$ and projecting it to the point $(y, z) \in \mathbb{R}^2$ is a homeomorphism. One can verify that that the image $f[O]$ is just the unit ball in $\mathbb{R}^2$ centered at the point $(2, 2)$ and is therefore open. So the hemisphere and the ball have the same structure. Since we can cover the surface of S^2 with hemispheres such as O, we find that the sphere S^2 is locally $\mathbb{R}^2$. We now consider the notion of "smoothness" on a locally $\mathbb{R}^n$ topological space.

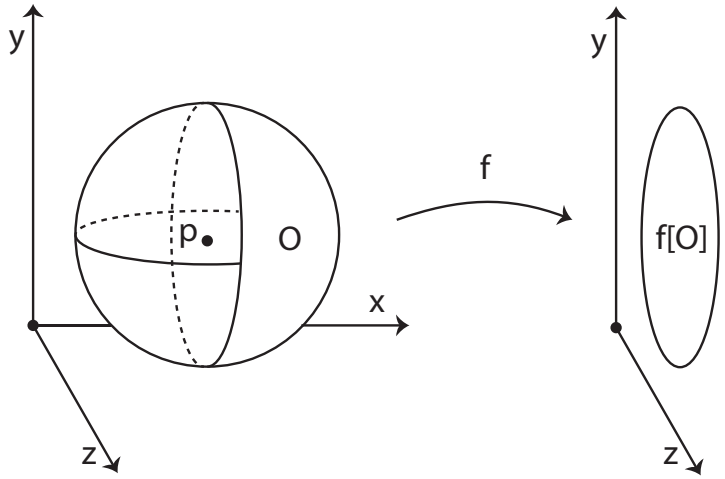

Figure 2.4: The sphere centered at p. The homeomorphism f projects the open hemisphere $O \subset S^2$ to an open ball $f[O] \subset \mathbb{R}^2$. The surface of the sphere can be covered by such hemispheres showing that it is locally $\mathbb{R}^2$.

Let (M, τ) be a topological space. A n-dimensional **chart** (or coordinate patch) on M is an ordered pair (U, φ) where U is an open subset of M and φ is a homeomorphism from U to an open subset of $\mathbb{R}^n$. In the sphere example just given, we see that (O, f) is a chart on S^2 (recall Figure 2.4). We say that any pair of n-dimensional charts (U, φ) and (V, ψ) on M are **compatible** if either $U \cap V = \varnothing$ or both of the composed "transition" maps $\varphi \circ \psi^{-1} : \psi[U \cap V] \to \mathbb{R}^n$ and $\psi \circ \varphi^{-1} : \varphi[U \cap V] \to \mathbb{R}^n$ are smooth. Since each of the composed maps is just a function from an open set of $\mathbb{R}^n$ to some other open set of $\mathbb{R}^n$ we know what "smoothness" means here: having continuous partial derivatives of all orders. It might be useful to consider an example of compatible charts.

Consider the unit circle (one-dimensional sphere) $S \subset \mathbb{R}^2$ that is centered at the origin $(0, 0)$ (see Figure 2.5). Let the sets U and V be, respectively,

the $x > 0$ and $y > 0$ portions of the circle which are open in S. The function φ defined by projecting any point $(x, y) \in U$ to $y \in \mathbb{R}$ is a homeomorphism. So (U, φ) is a one-dimensional chart on S. Similarly, the function ψ defined by projecting any point $(x, y) \in V$ to $x \in \mathbb{R}$ is a homeomorphism. So (V, ψ) is also a one-dimensional chart on S. Are the two charts compatible? Since $U \cap V$ is non-empty (dotted line in the diagram), we must check the transition maps. Consider $\psi \circ \varphi^{-1} : \varphi[U \cap V] \to \mathbb{R}$ depicted in the diagram. Let p be any point in $\varphi[U \cap V] = (0, 1)$. The inverse function φ^{-1} must send this point p to the point $(\sqrt{1 - p^2}, p) \in U \cap V$. And ψ sends the point $(\sqrt{1 - p^2}, p) \in U \cap V$ to the point $\sqrt{1 - p^2} \in \mathbb{R}$. So we have $\varphi \circ \psi^{-1}(p) = \sqrt{1 - p^2}$ for all $p \in \varphi[U \cap V] = (0, 1)$ which is a smooth function. The other transition map is handled similarly and so the two charts are compatible.

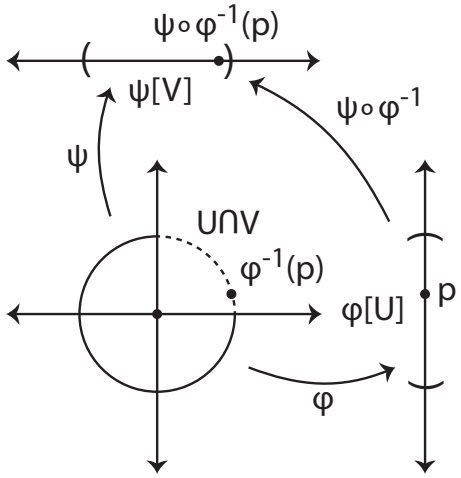

Figure 2.5: The unit circle S centered at the origin. The charts (U, φ) and (V, ψ) are depicted with overlapping domain $U \cap V$. A smooth transition map sends the point $p \in \mathbb{R}$ to the point $\psi \circ \varphi^{-1}(p) \in \mathbb{R}$.

A topological space (M, τ) has an **atlas** (of dimension n) if there is a collection $\mathcal{C}$ of n-dimensional charts on M that cover it in the sense that for each $p \in M$, there is a chart $(U, \varphi) \in \mathcal{C}$ such that $p \in U$. If an atlas is such that any pair of its charts are compatible, it is **smooth**. A smooth atlas is **maximal** if there is no chart compatible with all charts in the atlas that is not already found in the atlas. A topological space (M, τ) is a (smooth) n-dimensional **manifold** if it has a maximal smooth atlas of dimension n. Of course, all n-dimensional manifolds are locally $\mathbb{R}^n$ topological spaces. In

dimensions three or less, each locally $\mathbb{R}^n$ topological space admits exactly one maximal smooth atlas. But in higher dimensions, things are more complicated. A locally $\mathbb{R}^n$ topological space need not admit a maximal smooth atlas at all; an example in ten dimensions was first given by Kervaire (1960). And a locally $\mathbb{R}^n$ topological space can also admit more than one maximal smooth atlas; an example in seven dimensions was first given by Milnor (1956).

When no confusion arises, we will drop explicit reference to the topology τ and maximal smooth atlas $\mathcal{C}$ of a manifold M. It will be useful to note a couple of basic facts here. First, the result of excising any closed proper subset from a manifold is also a manifold. Let M be a manifold with C a closed proper subset of M. The set $M - C$ with the subspace topology needs a maximal smooth atlas. It is given by the collection of all charts (U, φ) in the maximal smooth atlas of M for which $U \subseteq M - C$. Second, one can show that the product of any two manifolds is itself a manifold. Let M and N be an m and n dimensional manifolds respectively. The set $M \times N$ with the product topology needs a smooth maximal atlas. Suppose (U_M, φ_M) and (U_N, φ_N) are any charts in the maximal smooth atlases for M and N respectively. One can define a chart (U, φ) for $M \times N$ by letting $U = U_M \times U_N$ and letting φ map any point $(p, q) \in U$ to the point $(p_1, ..., p_m, q_1, ..., q_n)$ in $\mathbb{R}^{m+n}$ where $\varphi_M(p) = (p_1, .., p_m)$ and $\varphi_N(q) = (q_1, .., q_n)$. The collection of all such charts (U, φ) is a smooth atlas for $M \times N$. Adding to this collection all charts that are compatible with all charts in the collection defines a maximal smooth atlas for $M \times N$.

Let M be an manifold. A scalar function $f : M \to \mathbb{R}$ is **smooth** if, for all charts (U, φ) in the maximal smooth atlas of M, the function $f \circ \varphi^{-1} : \varphi[U] \to \mathbb{R}$ is smooth. The composed map is just a function from an open portion of $\mathbb{R}^n$ to $\mathbb{R}$ and the notion of smoothness is clear in that context. We will soon use this notion of smooth scalar functions on manifolds to characterize vectors in a general way. Before doing so, it will be helpful to define what it means to say that a map between a pair of manifolds is smooth. This idea gives rise to a natural formulation of a smooth curve $\lambda : I \to M$ on a manifold M – as long as we require that the interval $I \subseteq \mathbb{R}$ is open and connected. In that case, I is the result of removing a closed set from the manifold $\mathbb{R}$. And as we have mentioned above, this means I inherits a natural maximal smooth atlas from M and therefore can be thought of as a manifold in its own right. We will not always require the interval I to be an open, connected interval. It will be useful, for example to consider closed curves of the form $\lambda : [0, 1] \to M$ so as to allow for curve endpoints. There is a way to rigorously

define smooth curves within this context but we will not consider it here (see Lee (2013) for details).

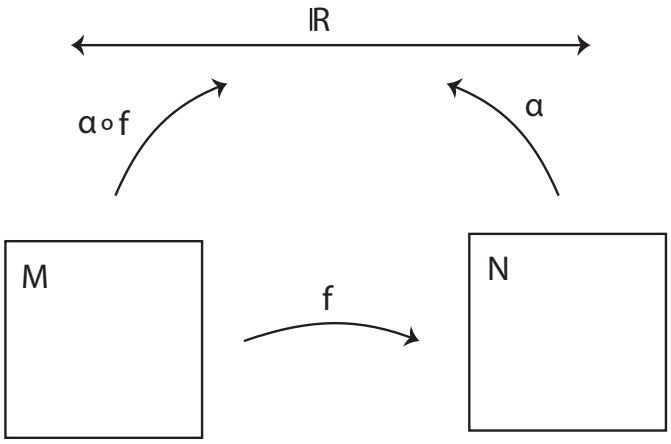

Figure 2.6: The map f from the manifold M to the manifold N is smooth if, for all smooth scalar functions $\alpha : N \to \mathbb{R}$, the composed map $\alpha \circ f : M \to \mathbb{R}$ is smooth.

Let $f : M \to N$ be a map from the manifold M to the manifold N. We say f is **smooth** if, for all smooth scalar functions $\alpha : N \to \mathbb{R}$, the composed map $\alpha \circ f : M \to \mathbb{R}$ is smooth (see Figure 2.6). From this definition, we see that if M is a manifold and $\lambda : I \to M$ is a curve, then the curve is smooth if $\alpha \circ \lambda : I \to \mathbb{R}$ is smooth in the familiar sense. A second equivalent definition of a smooth map between manifolds is sometimes useful: the map $f : M \to N$ between manifolds M and N is smooth if and only if for any $p \in M$ one can find a chart (U, ψ) containing p and a chart (V, ψ) containing $f(p) \in N$ such that (i) $f[U] \subseteq V$ and (ii) the composed map $\psi \circ f \circ \varphi^{-1} : \varphi[U] \to \mathbb{R}^n$ is smooth where n is the dimension of N.

2.5 Vectors

When working with the manifold $M = \mathbb{R}^n$, the notion of a vector at point is clear. At each $p \in M$, one associates a copy of the vector space $\mathbb{R}^n$ whose elements are thought to be based at p. Denote this vector space by V_p. A point $p = (p_1, ..., p_n)$ is an element of $M = \mathbb{R}^n$ while a vector $v = [v_1, ..., v_n]$

at p is an element of $V_p = \mathbb{R}^n$. We now explore the notion of "vectors" with respect to arbitrary manifolds.

Let M be an n-dimensional manifold M and let $p \in M$. Let $\mathfrak{F}$ be the collection of all smooth functions $\alpha : M \to \mathbb{R}$. A **vector** at p is a function $v : \mathfrak{F} \to \mathbb{R}$ such that for all $\alpha, \beta \in \mathfrak{F}$ the following are satisfied: (i) $v(\alpha+\beta) = v(\alpha) + v(\beta)$, (ii) $v(\alpha\beta) = \alpha(p)v(\beta) + \beta(p)v(\alpha)$ and (iii) $v(\alpha) = 0$ if α is a constant function, i.e. $\alpha(q)$ is the same real number for all $q \in M$. The collection V_p of all vectors at p has a natural vector space structure. One can, for example, add the vectors v and w by setting $(v + w)(\alpha) = v(\alpha) + w(\alpha)$. This vector space V_p is n-dimensional and is called the **tangent space** of p.

The characterization of vectors here is quite abstract. Let's try to connect it up to some familiar notions. Let (U, φ) be a chart containing p where $(x_1, ...x_n)$ are the coordinates of $\mathbb{R}^n$. For each $i = 1, ..., n$, let $X_i : \mathfrak{F} \to \mathbb{R}$ be defined as follows: for all $\alpha \in \mathfrak{F}$, the quantity $X_i(\alpha)$ is just the partial derivative $\frac{\partial}{\partial x_i}(\alpha \circ \varphi^{-1})$ evaluated at $\varphi(p)$. This definition makes sense because the composition $\alpha \circ \varphi^{-1}$ is just a smooth scalar function on a portion of $\mathbb{R}^n$ containing the point $\varphi(p)$. So we can take partial derivatives in each of the x_i directions. It turns out one can express any vector v at p as linear combination of the X_i functions: for some real numbers $v_1, ..., v_n$, we have $v(\alpha) = v_1X_1(\alpha) + ... + v_nX_n(\alpha)$ for all $\alpha \in \mathfrak{F}$. So relative to a given chart (U, φ), the vector v can be represented simply by its coordinate components which we will put in square brackets as before: $v = [v_1, ..., v_n]$.

Now consider any smooth curve $\lambda : I \to M$ such that $\lambda(s_0) = p$ for some $s_0 \in I$. The **tangent vector** of λ at p is the vector $v : \mathfrak{F} \to \mathbb{R}$ defined as follows: for all $\alpha \in \mathfrak{F}$, the quantity $v(\alpha)$ is the derivative of the composed map $\alpha \circ \lambda : I \to \mathbb{R}$ evaluated at $s = s_0$. Here, the number $v(\alpha)$ can be thought of as the rate of change of α at p in the direction of the curve.

Now let's think about what the tangent vector looks like in particular coordinates. Let (U, φ) be a chart where $(x_1, ...x_n)$ are the coordinates of $\mathbb{R}^n$. For each $i = 1, ..., n$, one has a natural projection $x_i : \mathbb{R}^n \to \mathbb{R}$ that maps the point $(x_1, ...x_n) \in \mathbb{R}^n$ to its ith coordinate. Associated with the chart (U, φ) are the **coordinate maps** $u_i, ..., u_n : U \to \mathbb{R}$ defined as follows: for all $q \in U$, the quantity $u_i(q)$ is just $(x_i \circ \varphi)(q)$ for $i = 1, ..., n$. This means that $\varphi(q) = (u_1(q), ..., u_n(q))$. The tangent vector of λ at the point $\lambda(s)$ in coordinate components comes out as as $[(u_1 \circ \lambda)'(s), ..., (u_n \circ \lambda)'(s)]$ where the prime indicates differentiation. For this reason, we shall often denote the tangent vector by $\lambda'(s)$ as before.

An example might be useful to bring things down to earth. Let M be the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. Here, we allow the coordinate $\theta \in S$ to take on all values of $\mathbb{R}$ but we identify each θ with $\theta + 2\pi n$ for all integers n. In these coordinates, at least two charts are needed to cover M. Consider a curve $\lambda : \mathbb{R} \to M$ defined by $\lambda(s) = (s, s^2)$. Suppose we wanted to find an expression for its tangent vector v at the point $p = (1, 1)$. Given our choice of coordinates, there must be a chart (U, φ) containing p where U is the $0 < \theta < \pi$ region of M and φ is a map from U to a subset of $\mathbb{R}^2$ defined by $\varphi(t, \theta) = (t, \theta)$. This chart does not contain all of the curve but this is okay; all that matters is that $p \in U$. Naturally, the coordinate maps $u_t, u_\theta : U \to \mathbb{R}$ are given by $u_t(t, \theta) = t$ and $u_\theta(t, \theta) = \theta$. If we let the composed maps $u_t \circ \lambda = s$ and $u_\theta \circ \lambda = s^2$ be denoted λ_t and λ_θ to accord with our earlier notation, we find that the tangent vector of λ in coordinate components comes out as as $\lambda'(s) = [\lambda_t'(s), \lambda_\theta'(s)] = [1, 2s]$. Since $s = 1$ at p, we see that the tangent vector there is $v = [1, 2]$ (see Figure 2.7).

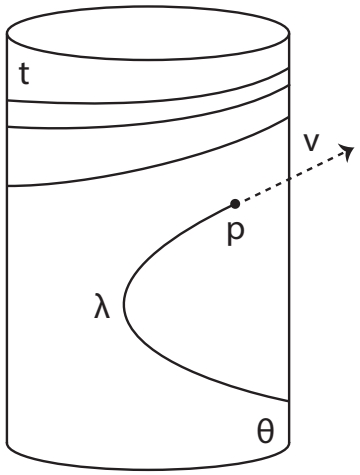

Figure 2.7: The tangent vector $v = [1, 2]$ to the curve λ at point $p = (1, 1)$.

2.6 Spacetime Shapes

We now wish to restrict attention to manifolds that are suitable for representing spacetime shapes. A pair of topological conditions are usually imposed: it must be connected and Hausdorff. We now turn to these notions. We say a

topological space (X, τ) is **connected** if there do not exist disjoint open sets $O_1, O_2 \subset X$ such that $O_1 \cup O_2 = X$. This idea here is quite intuitive and one can verify that $\mathbb{R}^n$ and S^n are connected for all n. Now consider $X = \mathbb{R} - \{0\}$ with the subspace topology induced from $\mathbb{R}$. It counts as a manifold but fails to be connected since X is the union of the disjoint open intervals $(-\infty, 0)$ and $(0, \infty)$. There is no "connection" between these parts of X. One finds that the product of any two connected topological spaces is itself connected. Moreover, if (X, τ) to (Y, σ) are topological spaces, then a continuous function $f : X \to Y$ will preserve connectedness: if (X, τ) is connected then so is $f[X] \subseteq Y$ in the subspace topology. In particular, this means that for any connected manifold M and any continuous function $f : M \to \mathbb{R}$, the image $f[M]$ is a connected interval of $\mathbb{R}$.

We say a topological space (X, τ) is **Hausdorff** if, for any distinct points $p, q \in X$, there exist disjoint neighborhoods of the points. The Hausdorff condition ensures that distinct points can be properly "separated" no matter how close they are. One can verify that $\mathbb{R}^n$ and S^n are Hausdorff for all n. Consider any distinct points $p, q \in \mathbb{R}$ for example. If we let $\epsilon = |p - q|/2$, then the intervals $(p - \epsilon, p + \epsilon)$ and $(q - \epsilon, q + \epsilon)$ are disjoint open balls containing p and q respectively. A non-Hausdorff topological space called the "branching line" is depicted in Figure 2.8. Consider two copies of the real line: $X = \mathbb{R}$ and $Y = \mathbb{R}$. Identify the points $x \in X$ and $y \in Y$ if such points are both negative and $x = y$. The resulting structure is a manifold but fails to be Hausdorff since any neighborhoods of the distinct points $p = 0 \in X$ and $q = 0 \in Y$ overlap. For example, the interval $U = (-1, 1)$ in X and $V = (-2, 2)$ in Y are neighborhoods of p and q respectively. But because of the identifications for all negative points, one finds that U and V have a non-empty intersection (dotted line in the diagram). If a topological space is Hausdorff, then (i) a compact set must be closed and (ii) a closed set contained in a compact set must itself be compact. The product of any two Hausdorff topological spaces is itself Hausdorff. Moreover, the result of excising any closed proper subset from a Hausdorff topological space is also Hausdorff.

We have finally arrived at the class of spacetime shapes that will be the primary focus in what follows: connected, Hausdorff manifolds. All standard general relativistic spacetimes will necessarily have an underlying manifold of this type. But as we will see in the next chapter, some connected, Hausdorff manifolds fail to admit a spacetime metric and are therefore not suitable for representing spacetime. (In Chapter 13, we will consider a non-standard

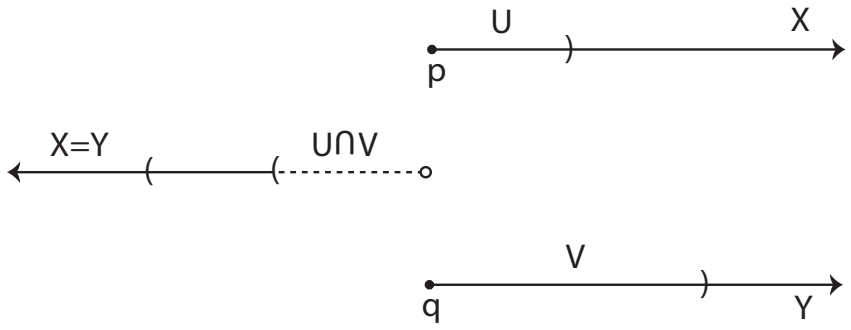

Figure 2.8: The negative portions of two copies of the real line $X = \mathbb{R}$ and $Y = \mathbb{R}$ are identified. The distinct points p and q have respective neighborhoods U and V with non-empty intersection.

context of general relativity in which the Hausdorff condition is dropped.) A large number of such shapes can be constructed by taking products of $\mathbb{R}^n$ and S^n and excising closed sets. For example, consider the manifold $\mathbb{R}^2$ in (t, x) coordinates and let C be the closed set $\{p\}$ for any point $p \in \mathbb{R}^2$. Then $M = \mathbb{R}^2 - C$ counts as a connected, Hausdorff manifold. Now endow M with a two-dimensional version of the Minkowskian metric η: for any vectors $v = [v_t, v_x]$ and $w = [w_t, w_x]$, we let $\eta(v, w) = v_t w_t - v_x w_x$. The resulting structure (M, η) qualifies as a possible universe compatible with general relativity (see Figure 2.9). But it does not seem to be "as large as it can be" because of the "missing" point p. Such examples will play an important role in our study of spacetime modality in what follows.

Let's now consider another locally Minkowskian spacetime with a cylindrical manifold. Unlike Minkowski spacetime itself, this possible universe allows for some spacetime events to be connected by more than one timelike geodesic. This gives rise to a third iteration of the Alice-Betty twin effect. But this time no acceleration is needed for Alice to arrive at the tea party younger much younger than Betty. Let M be the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. As before, we allow the coordinate $\theta \in S$ to take on all values of $\mathbb{R}$ but we identify each θ with $\theta + 2\pi n$ for all integers n. As we have seen, a vector v at any point $p \in M$ can be expressed in term of the coordinate

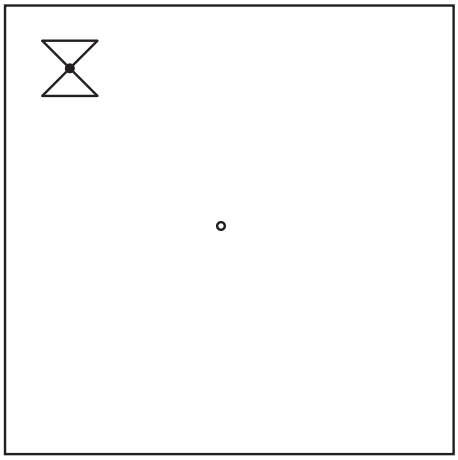

Figure 2.9: The result of removing a point from Minkowski spacetime is also a spacetime.

components $[v_t, v_\theta]$. A two-dimensional version of the Minkowskian metric η can now be defined on M: for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$, we let $\eta(v, w) = v_t w_t - v_\theta w_\theta$. We see that this spacetime (M, η) is a type of "rolled up" Minkowski universe in two dimensions.

Consider the events $p = (0, 0)$ and $q = (5\pi, 0)$ and Alice's world-line $\lambda : [0, 5\pi] \to M$ that connects them which is defined by $\lambda(s) = (s, 4s/5)$. This is a timelike geodesic that moves around the cylinder twice (see Figure 2.10). Differentiating the components of $\lambda(s)$ we find that Alice's the tangent vector is $\lambda'(s) = [1, 4/5]$. So extending the definitions and notion used in our study of Minkowski spacetime, we find that the (squared) length of this tangent vector is $\|\lambda'(s)\| = \eta(\lambda'(s), \lambda'(s)) = 1 - 16/25 = 9/25$. Integrating $\sqrt{\|\lambda'(s)\|} = 3/5$ from $s = 0$ to $s = 5\pi$, we find that Alice's elapsed time is $\|\lambda\| = 3\pi$ years. Betty's world-line is the timelike geodesic $\gamma : [0, 5\pi] \to M$ which also runs from p to q which is defined by setting $\gamma(s) = (s, 0)$. Here, Betty's tangent vector $\gamma'(s)$ is $[1, 0]$ at every point. So $\|\gamma'(s)\| = 1$. Integrating $\sqrt{\|\gamma'(s)\|} = 1$ from $s = 0$ to $s = 5\pi$, we find that $\|\gamma\| = 5\pi$ years. Alice is more than 6 years younger than Betty when they meet up and no acceleration was needed to achieve the effect.

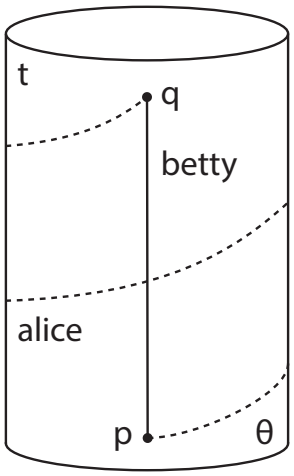

Figure 2.10: The geodesic world-lines of Alice and Betty from p to q. The elapsed times are 3π and 5π years respectively.

2.7 Conclusion

The shape of spacetime within the context of general relativity must be a connected, Hausdorff manifold. Spacetime manifolds have the local structure of the manifold $\mathbb{R}^n$ but are permitted to have different global features. In this chapter, we have given a given a formal characterization of such objects. Because spacetime manifolds are smooth in the appropriate sense, we were able to give quite general definitions of a variety of notions that are familiar in the context of $\mathbb{R}^n$ such as smooth scalar fields and vectors at a point (including tangent vectors associated with smooth curves). This allowed us to consider a generalized notion of the Minkowskian metric η on spacetime manifolds other than $\mathbb{R}^4$. A pair of examples provided a first look at general relativistic spacetimes other than the Minkowski universe.

Chapter 3

Curvatures

3.1 Introduction

The spacetimes of general relativity we have considered so far have all been of the form (M, η) where M is a connected, Hausdorff manifold and η is the Minkowskian metric defined at each event in M. Even when the shape of M is different from $\mathbb{R}^n$, such a model is "flat" with respect to its metric structure. In what follows, we will explore other metrics permitted by general relativity that allow for spacetime "curvatures" of various kinds. We begin with an informal discussion to better understand the notion of geodesics. We then get rigorous and formally define smooth vector fields on a given manifold. This allows for a precise characterization of the main idea of the chapter: a (smooth) general relativistic spacetime metric. We take a look at pair of examples: the non-flat de Sitter and anti-de Sitter metrics. We return to an informal discussion of curvy spacetime and its relationship to the distribution and flow of matter given by Einstein's equation. We close by considering a few "energy conditions" connected with Einstein's equation that constrain the local spacetime structure in different ways. For details concerning all of the informal discussions, we refer the reader to the foundational texts of Wald (1984) and Malament (2012).

3.2 Geodesics

In Minkowski spacetime $(\mathbb{R}^4, \eta)$, there is a natural way of "connecting up" the tangent spaces V_p and V_q at distinct points p and q. A vector $v =$

$[v_t, v_x, v_y, v_z]$ in V_p is "the same" as a vector $w = [w_t, w_x, w_y, w_z]$ in V_q if all of the components are identical. But how does one compare vectors at distinct points p and q on the sphere S^2? (See Figure 3.1.) One needs a way of "transporting" a vector from one of the points to a vector at another. A **derivative operator** ∇ specifies how this is to be done relative to any smooth curve $\lambda : I \to M$ which runs from p to q. It is a type of standard by which one can keep track of how a vector (or any geometrical structure on M) "changes" along the curve. If one starts with a vector v at p, one can **parallel transport** it to the point q by requiring that it stay constant with respect to ∇ along the way. Geodesic curves are special in the sense that if one parallel transports its tangent vector at any point p on the curve to any other point q along the curve, the result is identical to the curve's tangent vector at q. In this way, geodesics are "self parallel" or "as straight as possible" relative to ∇.

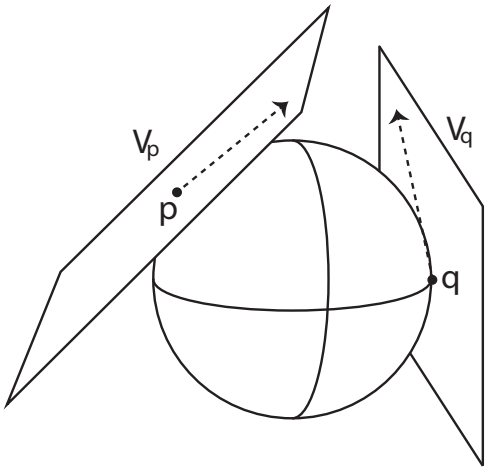

Figure 3.1: Vectors at points p and q on the sphere are elements of different tangent spaces V_p and V_q.

There are infinitely many distinct derivative operators ∇ that one can define on any manifold M. In some contexts, one choice may be more natural than others. For example, one usually defines ∇ on the sphere S^2 so that the geodesics come out as (portions of) the great circles that divide the sphere in two equal hemispheres (e.g. the equator on the earth). Consider Figure 3.2. We see a geodesic λ on the sphere go from the point p on the "equator" through the "north pole" to another point q on the equator. Because λ is a

geodesic, if one parallel transports its tangent vector v at p to the point q along the curve, the result w is just the tangent vector of λ of at that point. Also depicted in the diagram is another geodesic γ that runs from p to q as well but this time along the equator. Note that if we parallel transport the vector v at p along γ, then the result u at q is quite different from w. When parallel transport is path dependent in this way, the derivative operator is **curved**. Otherwise, it is **flat**.

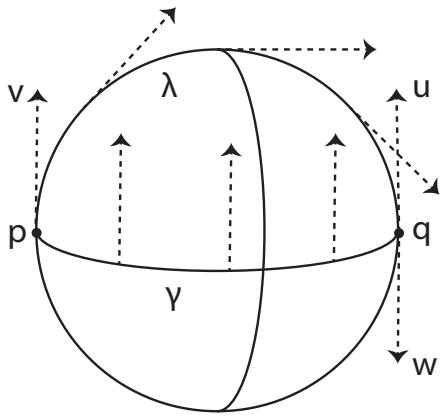

Figure 3.2: When the vector v at point p on the sphere is parallel transported to q along one curve, it yields w; along another, it yields u.

Consider again the Minkowski universe $(\mathbb{R}^4, \eta)$. There are many derivative operators ∇ one could define on $\mathbb{R}^4$. But there is only one such that the metric η constant with respect to it. Relative to that choice, if the metric is parallel transported along any curve whatsoever, it will not change. An equivalent way to put the point: if vectors v and w at p are parallel transported to the vectors v' and w' at point q along any curve whatsoever, then $\eta(v, w) = \eta(v', w')$. In the Minkowski universe, the unique derivative operator ∇ that makes η constant is flat (see Figure 3.3). We have been implicitly using this derivative operator all along when considering the geodesics of that universe.

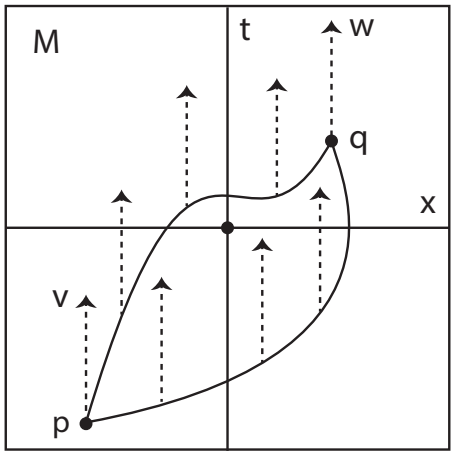

Figure 3.3: When the vector v at event p in the Minkowski universe is parallel transported to q along any curve, it always yields the same vector w.

3.3 Metrics

Let M be a manifold. A **vector field** over M a function v that assigns to every point $p \in M$ a vector $v_p \in V_p$. For any smooth function $\alpha : M \to \mathbb{R}$, can one construct the scalar function $v(\alpha) : M \to \mathbb{R}$ defined by $v(\alpha)(p) = v_p(\alpha)$ for all $p \in M$. We say the vector field v is **smooth** if this function $v(\alpha)$ is smooth for all smooth functions α. For example, on $M = \mathbb{R}^2$ in (t, x) coordinates, one can consider the smooth vector field v defined by assigning the vector $v_p = [t/2, x]$ at each point $p = (t, x)$ in M (see Figure 3.4). Recall what this means: the vector v_p at p assigns a number to each smooth function $\alpha : M \to \mathbb{R}$. What is the number? Let's say $\alpha(t, x) = tx^2$. So at the point $p = (t, x)$ the vector $v_p = [t/2, x]$ assigns the function α the number $v_p(\alpha) = (t/2)(\partial_t \alpha) + (x)(\partial_x \alpha)$ evaluated at p. (Recall the discussion in Section 2.5.) This comes out as $v_p(\alpha) = (t/2)(x^2) + (x)(2tx) = 5tx^2/2$ at $p = (t, x)$. Thus, at the point $p = (3, 2)$ the vector $v_p = [3/2, 2]$ assigns the function α the number $v_p(\alpha) = 30$.

Stepping back, we see the scalar function $v(\alpha) : M \to \mathbb{R}$ defined by $v(\alpha)(p) = v_p(\alpha)$ for all $p \in M$ is just $5tx^2/2$ which is smooth. One can verify that this would be the case no matter what function α were chosen and so the vector field v is smooth. It is of some interest to explore which smooth vector fields can be defined on which manifolds. For example, the "hairy

ball" theorem of Brouwer states that the sphere S^n admits a non-vanishing continuous vector field if and only if n is odd. Intuitively, one can try to "comb" a hairy two-dimensional sphere but one always finds this impossible as a "cowlick" must appear somewhere.

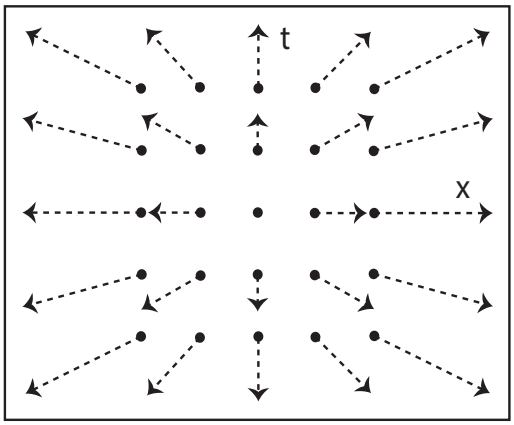

Figure 3.4: The smooth vector field $v = [t/2, x]$ on M.

A **metric** at $p \in M$ is a function $g_p : V_p \times V_p \to \mathbb{R}$ that is everywhere (i) symmetric in the sense that $g_p(v_p, w_p) = g_p(w_p, v_p)$ for all vectors $v_p, w_p \in V_p$ and (ii) non-degenerate in the sense that if $g_p(v_p, w_p) = 0$ for all $v_p \in V_p$, then w_p is the zero vector. In the natural way, one can define a **metric** g on all of M by assigning a metric g_p at each point $p \in M$. We say a metric g on M is **smooth** if, for any smooth vector fields v and w on M, the function $g(v, w) : M \to \mathbb{R}$ defined by $g(v, w)(p) = g_p(v_p, w_p)$ is smooth. For any metric g on M, there is a unique derivative operator ∇ on M that is **compatible** with g in the sense that g is constant with respect to ∇. So the metric g encodes (via its compatible derivative operator ∇) a natural geodesic structure on M.

Let g be a smooth metric on a manifold M. Naturally, the (squared) **length** $\|v\|$ of a vector v at a given point in M is given by $g(v, v)$. We find that at every point $p \in M$, there will be a basis of vectors $v_1, ..., v_n \in V_p$ that is **orthonormal** in the sense that (i) $g(v_i, v_j) = 0$ if $i \neq j$ and (ii) the length $\|v_i\|$ of any vector v_i is either 1 or -1. A metric is **Riemannian** if, at any point, the length of all such basis vectors is 1. One can show that the manifolds $\mathbb{R}^n$ and S^n admit a Riemannian metric for all n. Let

g be any Riemannian metric on a manifold M and let $\lambda : I \to M$ be a smooth curve with tangent vector $\lambda'(s)$. Because the length $\|\lambda'(s)\|$ will be non-negative, we can integrate the quantity $\sqrt{\|\lambda'(s)\|}$ along the curve to calculate its **length** $\|\lambda\|$. A familiar Riemannian metric is the **Euclidean** metric e on $\mathbb{R}^n$ which in standard $(x_1, ..., x_n)$ coordinates defined as follows: at any point $p \in \mathbb{R}^n$, and for any vectors $v = [v_1, ..., v_n]$ and $w = [w_1, ..., w_n]$ at the point, set $e(v, w) = v_1w_1 + ... + v_nw_n$. Like the Minkowskian metric η, the derivative operator compatible with the Euclidean metric e is flat.

A smooth metric g on the n-dimensional manifold M is **Lorentzian** if $n \geq 2$ and, at any point, there is an orthonormal basis such that one basis vector has length 1 while $n - 1$ basis vectors have length -1. Of course, the Minkowski metric η we have been working with is Lorentzian. Let us now define, in a quite general way, many of the notions we have already introduced within the familiar but limited context of Minkowski spacetime.

Let g be any Lorentzian metric on a manifold M. A vector at a point $p \in M$ with positive length is **timelike**; a zero length vector is **null**; a negative length vector is **spacelike**. Thus, at any point $p \in M$, we recover a light cone structure. A smooth curve $\lambda : I \to M$ is also called timelike, null, or spacelike in the natural way. If λ is timelike, we can integrate the quantity $\sqrt{\|\lambda'(s)\|}$ along the curve to calculate its **elapsed time**. If λ is spacelike or null, we can integrate the quantity $\sqrt{-\|\lambda'(s)\|}$ along the curve to calculate its **length** $\|\lambda\|$. Of course, the length of any null curve must be zero. The elapsed times and lengths along several geodesics in Minkowski spacetime are depicted in Figure 3.5. One finds the familiar Pythagorean theorem on the $t = 0$ surface and a type of Minkowskian variation on the $y = 0$ surface.

We are finally ready to introduce the definition of a (standard) **spacetime** according to general relativity: an ordered pair (M, g) where M is an n-dimensional (for $n \geq 2$), connected, Hausdorff, (smooth) manifold and g is a (smooth) Lorentzian metric on M. Although much of this chapter is presented informally, we emphasize here that the definition of a relativistic spacetime has been precisely formulated. We have carefully built, from the ground up, the notions of a spacetime manifold and a Lorentzian metric.

Some connected, Hausdorff manifolds fail to admit a Lorentzian metric. For example, the fact that there cannot be a continuous non-vanishing vector field on the sphere S^n for even n implies that no Lorentzian metric cannot be defined on such manifolds. But S^n and $\mathbb{R}^m$ admit a Lorentzian metric for all odd $n \geq 2$ and all $m \geq 2$. If a connected, Hausdorff manifold does admit

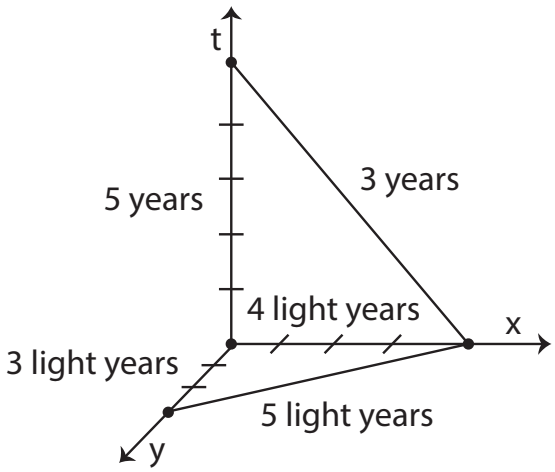

Figure 3.5: The elapsed times and lengths along several geodesics in Minkowski spacetime.

a Lorentzian metric, then it must also satisfy a useful topological property ensuring that it is not "too big." Consider a topological space (X, τ). A set $\sigma \subseteq \tau$ is a **basis** if every open set $O \in \tau$ can be expressed as a union of sets in σ. A topological space (X, τ) is **second countable** if it has a countable basis. One can show that $\mathbb{R}^n$ and S^n are second countable for all n. In the case of $\mathbb{R}^n$, a countable basis can be found by taking the collection of all open balls $B \subset \mathbb{R}$ with rational radius ϵ centered at the point $p = (p_1, ..., p_n)$ for which $p_1, ..., p_n$ are all rational. A simple example of a topological space which is not second countable is $\mathbb{R}$ with the discrete topology. We introduced this topological definition to state the following result: any standard general relativistic spacetime necessarily has a second countable manifold. (As we will see in Chapter 13, dropping the Hausdorff condition will permit non-standard spacetimes that fail to be second countable.)

3.4 Curvy Spacetime

In what follows, let $\mathscr{U}$ be the collection of all spacetimes. Let's take a look at some members of this collection with non-trivial spacetime curvature. Let M be the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. As before, we allow the coordinate $\theta \in S$ to take on all values of $\mathbb{R}$ but we identify each θ with $\theta + 2\pi n$ for all

integers n. Now consider a metric g defined on M as follows: at each point $(t, \theta) \in M$ and for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $g(v, w) = v_t w_t - v_\theta w_\theta \cosh^2(t)$. One can verify that a given metric is smooth by checking to make sure that any scalar functions used to define it are themselves smooth. In this case, $\cosh(t)$ is smooth and so we know g is also. (Recall our discussion of the hyperbolic sine and cosine functions in Section 1.4.)

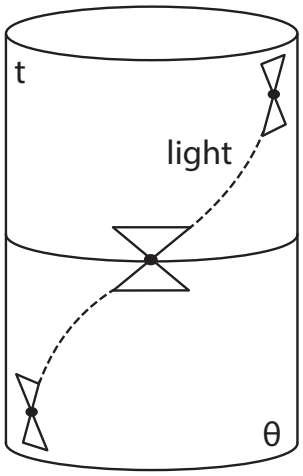

Figure 3.6: The de Sitter universe with null geodesic depicted.

The pair (M, g) is a two-dimensional version of **de Sitter spacetime** introduced and named after Willem de Sitter (1917). Associated with the metric g is a compatible derivative operator ∇ defined on M. But in contrast to the flat derivative operator compatible with the Minkowskian metric η, the derivative operator ∇ compatible with g is curved. The non-trivial curvature present gives rise to peculiar geodesic structure. One can better understand things by considering the behavior of light which travels along null geodesics (see Figure 3.6). At any point $(t, \theta) \in M$ a null vector $v = [v_t, v_\theta]$ must have zero length: $\|v\| = 0$. But since $\|v\| = g(v, v)$, this means that $v_t^2 = v_\theta^2 \cosh^2(t)$. The function $\cosh(t)$ has a value of 1 at $t = 0$ and increases rapidly as t increases in absolute value. This means that the light cones will be at a 45 degree angle at $t = 0$ and rapidly narrow as t increases in absolute value. An example null geodesic λ through the point $(0, 0)$ is depicted in the diagram. As $t \to \infty$, the curve λ approaches but never reaches $\theta = \pi/2$; similarly, as $t \to -\infty$, it approaches $\theta = -\pi/2$.

In any spacetime (M, g), the derivative operator ∇ compatible with metric g on can be used to define the **Riemann curvature** $\mathcal{R}$ on the manifold M. With various combinations of $\mathcal{R}$, g, and ∇, one can construct an infinite collection of geometric objects, each of which keeps tracks of some type of invariant "curvature" on M. Of these infinitely many notions of curvature, we will focus primarily on two in what follows.

The **Einstein curvature** G on M is, like the metric g, a smooth assignment of a real number to pairs of vectors $v, w \in V_p$ at each point $p \in M$. The Einstein curvature plays an important role in the famous "Einstein's equation" relating the curvature and matter content of the universe that we will consider in a moment. The other type of curvature we want to consider is the **Ricci curvature** R on M which is a smooth scalar function $R : M \to \mathbb{R}$. For certain "maximally symmetric" spacetimes, the Ricci curvature is constant on M and, moreover, its value determines the local structure of the spacetime completely. Spacetimes of this kind come in three varieties: (i) those with $R = 0$ are locally structured like the Minkowski spacetime, (ii) those with $R > 0$ are locally structured like de Sitter spacetime (in the version presented here $R = 2$), and (iii) those with $R < 0$ are locally structured like "anti-de Sitter spacetime" we will discuss now.

Let $M = \mathbb{R}^2$ in (t, x) coordinates. Now consider a metric g defined on M as follows: at each point $(t, x) \in M$ and for any vectors $v = [v_t, v_x]$ and $w = [w_t, w_x]$ at the point, let $g(v, w) = v_t w_t \cosh^2(x) - v_\theta w_\theta$. **Anti-de Sitter spacetime** is the pair (M, g). The associated derivative operator ∇ is curved (here $R = -2$) but the behavior of geodesics is very different from the de Sitter universe. Let us again consider the behavior of light. At any point $(t, x) \in M$ a null vector $v = [v_t, v_x]$ must have zero length: $\|v\| = 0$. Since $\|v\| = g(v, v)$, this means that $v_t^2 \cosh^2(t) = v_x^2$. So the light cones will be at a 45 degree angle at $x = 0$ and rapidly widen as x increases in absolute value. An example null geodesic λ through the point $(0, 0)$ is depicted in the diagram. As $x \to \infty$, the curve λ approaches but never reaches $t = \pi/2$; similarly, as $x \to -\infty$, it approaches $t = -\pi/2$. Timelike curves can also exhibit a similar behavior. Perhaps an explicit example may be useful to work through.

Let's consider a wizard whose world-line is the timelike curve $\gamma : \mathbb{R} \to M$ defined by the function $\gamma(s) = (\sqrt{2}\arctan(\sinh(s)), s)$ where arctan is the inverse tangent function (see Figure 3.7). Do not worry too much about the ugly function $\sqrt{2}\arctan(\sinh(s))$; it is carefully chosen to make calculations easy in a moment. One can verify that as $s = x \to \infty$, the curve λ approaches

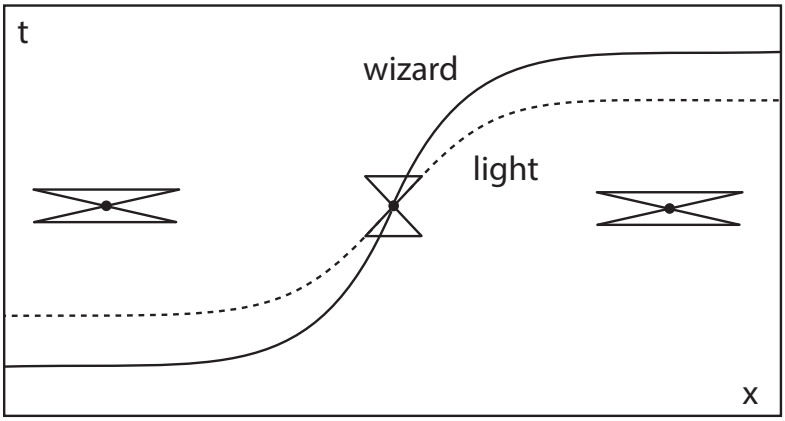

Figure 3.7: The anti-de Sitter universe with a null geodesic and a timelike curve depicted.

but never reaches $t = \sqrt{2}\pi/2$. This follows since, as $s \to \infty$, we have $\sinh(s) \to \infty$ and therefore $\arctan(\sinh(s)) \to \pi/2$. Similarly as $s = x \to -\infty$, the curve approaches $t = -\sqrt{2}\pi/2$. Differentiating the components of $\gamma(s)$ we find that the velocity vector is $\gamma'(s) = [\sqrt{2}/\cosh(s), 1]$. Things are reducing nicely already. What is $\|\gamma'(s)\|$? It comes out as $g(\gamma'(s), \gamma'(s)) = 2[\cosh(x)/\cosh(s)]^2 - 1^2$. Because $s = x$, things now reduce magically and we find that $\|\gamma'(s)\| = 1$. Integrating $\sqrt{\|\gamma'(s)\|} = 1$ from $s = -\infty$ to $s = \infty$, we find the elapsed time is $\|\gamma\| = \infty$. At any event, the wizard has "always existed" in the sense that his elapsed time is infinite in the past direction. And yet the he "never existed" before $t = -\sqrt{2}\pi/2$ after which he seems to have appeared out of thin air. The causal structure of this peculiar situation will be explored later on (see Section 5.5).

3.5 Einstein's Equation

We turn now to the connection between curvature and matter. Let (M, g) be a spacetime. One can consider an **energy momentum tensor** T on M which, like the metric g, amounts to a smooth assignment of a real number to pairs of vectors $v, w \in V_p$ at each point $p \in M$. For an observer at p with tangent vector v, the quantity $T(v, v)$ represents the energy density of

matter as determined by the observer at the point. Recall that the metric g determines (via its unique compatible derivative operator ∇) the Riemann curvature $\mathcal{R}$ on M which encodes every imaginable notion of "curvature" at each point on M. Of all such notions, the Einstein curvature G is special. To see why, consider an observer at point p with tangent vector v. The real number $G(v, v)$ represents a particular type of curvature as measured by the observer at p. By some sort of cosmic coincidence, the observer always finds that the following relationship holds at every spacetime event: $G = 8\pi T$. This is **Einstein's equation** which will be assumed in what follows.

Under Einstein's equation, we can associate with each spacetime (M, g) a unique energy momentum tensor T: the metric g determines G which determines T. So the geometry of spacetime determines the distribution and flow of matter. What about the other direction? Although T determines G via Einstein's equation, one cannot determine g from G in general. To better understand this asymmetry, let's consider an example. We say that a spacetime (M, g) is a **vacuum solution** to Einstein's equation if its associated energy momentum tensor T is such that $T(v, w) = 0$ for all vectors v and w at any point $p \in M$. One can show that any flat spacetime (e.g. Minkowski) is a vacuum solution. But in two dimensions, any spacetime is a vacuum solution (Fletcher et al., 2018). This includes the non-flat de Sitter and anti-de Sitter universes we have already considered. There are also examples of non-flat vacuum universes in four dimensions that we will consider later on. So, the distribution and flow of matter does not determine the geometry of spacetime.

Without any constraints on the energy momentum tensor T, any spacetime (M, g) counts as a type of "solution" to Einstein's equation. A given metric g on M will always give rise to some T or other via Einstein's equation. But the distribution and flow of matter represented by that T may not be "physically reasonable" in various senses. So one often uses a variety of local "energy conditions" to constrain things (Curiel, 2016). Let's take an informal look at a few. The **weak energy condition** requires that for any timelike vector v at any point $p \in M$, we have $T(v, v) \geq 0$. This requires that the energy density of matter as determined by an observer with tangent v is never negative. The **strong energy condition** is satisfied when a certain effective energy density as determined by any observer is never negative. This requires that "gravitation is attractive" in some sense. The weak and strong energy conditions are independent in the sense that neither implies the other. The **dominant energy condition** can be thought of as prohibit-

ing the flow of matter in a spacelike direction. The dominant and strong energy conditions are independent but dominant does imply weak. All three conditions do imply the **null energy condition** which requires that, for any null vector v at any point $p \in M$, we have $T(v, v) \geq 0$. This doesn't have much significance physically but is a simple condition to work with that is useful to have around as a minimal non-trivial constraint: if it were to fail, all three of the other energy conditions would also fail. On the other extreme, the condition of being a vacuum solution is very strong as it implies all four energy conditions.

Later on, we will explore the modal structure of spacetime in terms of background "possibility spaces" consisting of various collections of spacetime models. For this reason, it will be useful to think of the energy conditions as subcollections of the collection $\mathscr{U}$ of all spacetimes. Let $(NEC), (WEC), (SEC), (DEC) \subset \mathscr{U}$ be the collections of all spacetimes satisfying, respectively, the null, weak, strong, and dominant energy conditions. Let $(Vac) \subset \mathscr{U}$ be the collection of vacuum solutions. We can sum up the relationships between all the conditions as follows (see Figure 3.8).

$$(Vac) \subset (DEC) \subset (WEC) \subset (NEC)$$
$$(Vac) \subset (SEC) \subset (NEC)$$

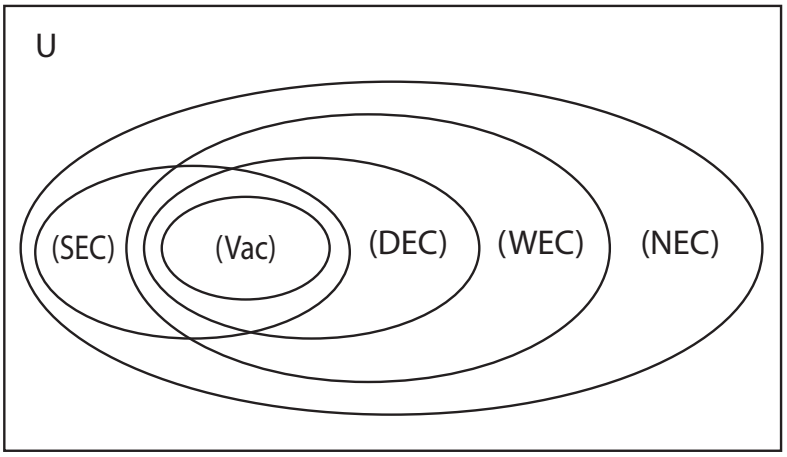

Figure 3.8: Various local spacetime properties concerning the distribution and flow of matter.

3.6 Conclusion

We have just discussed a number of topics in an informal way. Examples include derivative operators, parallel transport, geodesics, various types of spacetime curvature, Einstein's equation, and a number of the energy conditions. But the centerpiece of the chapter was a precise definition of spacetime according to (standard) general relativity: an ordered pair (M, g) where M is an n-dimensional (for $n \geq 2$), connected, Hausdorff, (smooth) manifold and g is a (smooth) Lorentzian metric on M. We also introduced the associated collection $\mathscr{U}$ of all such spacetimes and studied some example members with non-trivial curvature. This collection $\mathscr{U}$ represents the standard background "possibility space" used to define various of modal properties of spacetime (e.g. maximality) that will be central in what follows.

Chapter 4

Isomorphisms

4.1 Introduction

Isomorphisms are structure preserving maps between a pair of mathematical objects. Isomorphic objects are the "same" with resect to all of the relevant structure. We have already considered homeomorphisms which are isomorphisms with respect to topological spaces. In what follows, we will first consider "diffeomorphisms" which are isomorphisms between manifolds. Diffeomorphism allow for the smooth transfer of vectors between manifolds of the same structure. This vector transfer gives us the resources to define "isometries" which are the isomorphisms between spacetimes. This key definition will be used to define many foundational notions from here on out. We close by exploring two of them: the symmetries and invariant properties of spacetime.

4.2 Diffeomorphisms

We already know what it means to say that a map $f : M \to N$ from a manifold M to a manifold N is smooth. We used this definition to define smooth scalar and vector fields on a given manifold as well as smooth curves on a manifold. In all of these definitions, one has a map between manifolds that is smooth in one direction but not necessarily the other. We now consider maps between manifolds that are smooth in both directions. We say the map $f : M \to N$ is a **diffeomorphism** if it is a bijection (one-to-one and onto) and both f and its inverse f^{-1} are smooth. In this case, we say the manifolds

M and N are **diffeomorphic**. In the same way that homoemorphic topological spaces have the same topological structure, diffeomorphic manifolds have the same manifold structure. Consider a simple example (see Figure 4.1).

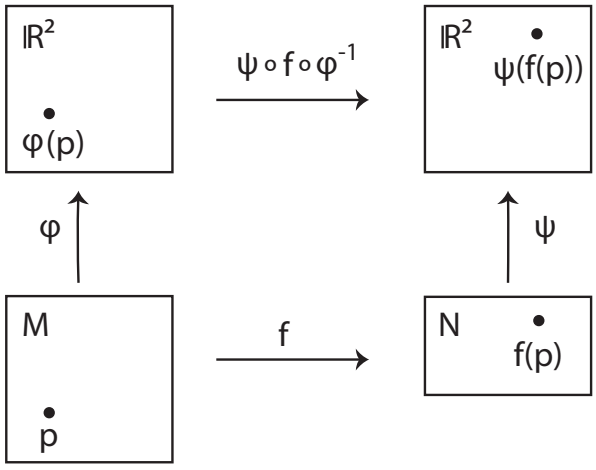

Figure 4.1: The function $f : M \to N$ along with charts (M, φ) and (N, ψ). Since the composed map $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ is smooth, so is f.

Let M be $\mathbb{R}^2$ in (t, x) coordinates. Let N be $\mathbb{R}^+ \times \mathbb{R}$ in (τ, χ) coordinates where $\mathbb{R}^+$ is the open interval $(0, \infty)$ in $\mathbb{R}$. The map $f : M \to N$ defined by $f(t, x) = (\exp(t), 2x)$ is a diffeomorphism. Let us verify this.

Recall that a map $f : M \to N$ between manifolds M and N is smooth if and only if for any $p \in M$ one can find a chart (U, ψ) containing p and a chart (V, ψ) containing $f(p) \in N$ such that (i) $f[U] \subseteq V$ and (ii) the composed map $\psi \circ f \circ \varphi^{-1} : \varphi[U] \to \mathbb{R}^n$ is smooth where n is the dimension of N. Consider any point $p \in M$. There is a global chart (M, φ) containing p where $\varphi : M \to \mathbb{R}^2$ is just the identity $\varphi(t, x) = (t, x)$. Similarly, there is a global chart (N, ψ) containing $f(p)$ where $\psi : N \to \mathbb{R}^2$ is just the inclusion map $\psi(\tau, \chi) = (\tau, \chi)$. Since $f[M] = N$, we see that condition (i) is satisfied. The composed map $\psi \circ f \circ \varphi^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ is just the function that sends $(t, x) \in \mathbb{R}^2$ to the point $(\exp(t), 2x) \in \mathbb{R}^2$ which is smooth. So condition (ii) is satisfied and we see that f is smooth. The other direction where $f^{-1} : N \to M$ is defined by $f^{-1}(\tau, \chi) = (\ln(\tau), \chi/2)$ is handled similarly. So f is a diffeomrophism. Although M and N are different manifolds, they share all of the same structure. We will return to this example in a moment.

4.3 Vector Transfer

One can use a diffeomorphism $f : M \to N$ to transfer a vector v at $p \in M$ to the "push forward" vector $f_*(v)$ at $f(p) \in N$. Similarly, we can also "pull back" a vector w at $q \in N$ to the vector $f^*(w)$ at $f^{-1}(q) \in M$. Let's build up to this idea step by step.

Let $f : M \to N$ be a diffeomorphism. Consider a smooth function $\beta : N \to \mathbb{R}$. It is easy to see that one can use f to pull back β to M by considering the composed function $\beta \circ f : M \to \mathbb{R}$. In a similar way, one can use f to push forward a smooth function $\alpha : M \to \mathbb{R}$ to N by considering the function $\alpha \circ f^{-1} : N \to \mathbb{R}$. Now for vectors.

Let v be a vector at $p \in M$. For each smooth function $\alpha : M \to \mathbb{R}$, we know that v determines a smooth function $v(\alpha) : M \to \mathbb{R}$. We can use f to to **push forward** the vector v at p to the vector $f_*(v)$ at $f(p)$ by setting $f_*(v)(\beta) = v(\beta \circ f)$ for all smooth functions $\beta : N \to \mathbb{R}$. So the vector $f_*(v)$ at $f(p)$ assigns to any function β just what the vector v assigns to the pulled back function $\beta \circ f$ at p. In a similar way, we can use f to to **pull back** a vector w at $q \in N$ to the vector $f^*(w)$ at $f^{-1}(q)$ by setting $f^*(w)(\alpha) = w(\alpha \circ f^{-1})$ for all smooth functions $\alpha : M \to \mathbb{R}$. So the vector $f^*(w)$ at $f^{-1}(q)$ assigns to any function α just what the vector w assigns to the pushed forward function $\alpha \circ f^{-1}$ at q. On can extend these definitions in the natural way to apply to smooth vector fields v on M and w on N. For example, one can push forward a vector field on M by pushing forward the vector at each point $p \in M$.

At each point $p \in M$, the push forward map f_* is a function of from the tangent space V_p to the tangent space $V_{f(p)}$. One can think of this function as the "derivative" of the diffeomorphism f at the point p. If one considers coordinates for M and N, then one can express the push forward as a function of a number of partial derivatives. Even in low dimensions, things can get a bit hairy. But the payoff is worth it. After we better understand the idea of transferring vectors between diffeomorphic manifolds, the definition of an isometry is one simple step away. And that notion is at the center of philosophical discussions of symmetry, structure, modality, and more. A few concrete examples will help us get a grip on things. Let's start with the push forward map in one dimension.

Suppose M is $\mathbb{R}$ in x coordinates. Let's consider three different diffeomorphisms from M to itself: a, b, and c. The first diffeomorphism is the identity map $a : M \to M$ defined by $a(x) = x$ for all $x \in M$. Now let $v = [v_x]$ be a

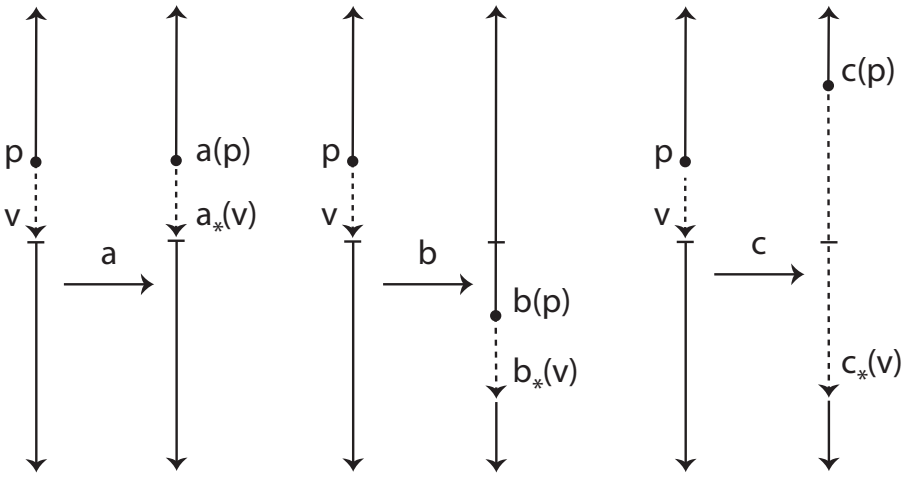

Figure 4.2: The diffeomorphisms a, b, and c all push forward the vector $v = [-1]$ at a point $x = 1$ but do so in different ways.

vector at a point $p \in M$. (Even if one dimension, we continue to use square brackets to distinguish vectors from points.) Where do we map v to at the point $a(p) = p$? To itself of course: we push forward v at p to the vector $a_*(v) = v$ at $a(p) = p$. For example, the vector $v = [-1]$ at the point $p = 1$ is mapped to the vector $a_*(v) = [-1]$ at the point $a(p) = 1$ (see Figure 4.2).

The second diffeomorphism is just a tad more interesting. The map $b : M \to M$ is defined by $b(x) = x - 2$ for all $x \in M$. Here, we shift each point two units in the negative direction. Let $v = [v_x]$ be a vector at a point $p \in M$. Where do we map v to at the point $b(p) = p - 2$? Even though, we have "moved" the points, the vector v at p and the push forward vector $b_*(v) = v$ at $b(p) = p - 2$ are "the same" in the sense that their coordinate components are identical. For example, the vector $v = [-1]$ at the point $p = 1$ is mapped to the vector $b_*(v) = [-1]$ at the point $b(1) = -1$ (see again Figure 4.2).

Now for the third diffeomorphism $c : M \to M$ is defined by $c(x) = x^3 + x$ for all $x \in M$. One can show that this map is smooth with a smooth inverse. (Why couldn't we use the function x^3 instead?) Here, we are "stretching" the manifold. Let $v = [v_x]$ be a vector at a point $p \in M$. Where do we map v to at the point $c(p) = p^3 + p$? This is trickier. A general formula will be given in a moment. But what we do is this: we take the derivative of $c(x)$ with respect to x, evaluate it at p, and then multiply it by the vector component v_x. Whatever this number is becomes the component of the push

forward vector at the point $c(p)$. Since the derivative of $c(x)$ with respect to x comes out as $c'(x) = 3x^2 + 1$, we push forward v at p to the vector $c_*(v) = [(3p^2+1)v_x]$ at $c(p) = p^3 + p$. As we stretch the manifold, we stretch the vectors on it as well. For example, the vector $v = [-1]$ at the point $p = 1$ is mapped to the vector $c_*(v) = [(3(1^2) + 1)(-1)] = [-4]$ at the point $c(1) = 2$ (see again Figure 4.2).

Let's try two-dimensions. Now four partial derivatives are required to determine the push forward. (In general, one must keep track of n^2 partial derivatives when working in n dimensions.) Returning to our example in the previous section, let M be $\mathbb{R}^2$ in (t, x) coordinates and let N be $\mathbb{R}^+ \times \mathbb{R}$ in (τ, χ) coordinates. As we have seen, the map $f : M \to N$ defined by $f(t, x) = (\exp(t), 2x)$ is a diffeomorphism. Now let $v = [v_t, v_x]$ be a vector at a point p in M. What is the push forward vector $f_*(v)$ at the point $f(p)$? It will be useful to separate $f(t, x) = (\exp(t), 2x)$ into its τ and χ components. Let $f_\tau(t, x) = \exp(t)$ and $f_\chi(t, x) = 2x$ so that $f(t, x) = (f_\tau(t, x), f_\chi(t, x))$. Various partial derivatives of these functions can be arranged into the **Jacobian matrix** below.

$$\begin{bmatrix} \partial_t f_\tau & \partial_x f_\tau \\ \partial_t f_\chi & \partial_x f_\chi \end{bmatrix} = \begin{bmatrix} \exp(t) & 0 \\ 0 & 2 \end{bmatrix}$$

The push forward vector $f_*(v)$ is the result of matrix multiplication of the vector $v = [v_t, v_x]$ by the Jacobian matrix. This comes out as the following.

$$[v_t \partial_t f_\tau + v_x \partial_x f_\tau, v_t \partial_t f_\chi + v_x \partial_x f_\chi] = [v_t \exp(t), 2v_x]$$

Suppose one has the vector $v = [-2, 3]$ at the point $p = (1, -2)$ in M. So $f_*(v) = [-2\exp(1), 6]$ at $f(p) = (\exp(1), -4)$ in N (see Figure 4.3).

In the example just given, two partial derivatives vanish. Let's try another example where things aren't so simple. Let $M = \mathbb{R}^+ \times S$ be the half cylinder in (z, θ) coordinates. Let $N = \mathbb{R}^2 - \{(0, 0)\}$ be the punctured plane in (x, y) coordinates. The map $f : M \to N$ defined by $f(z, \theta) = (z\cos\theta, z\sin\theta)$ is a diffeomorphism. Suppose one has a vector $v = [v_z, v_\theta]$ at a point p in M. Where do we map v to at the point $f(p)$? We first separate $f(z, \theta) = (z\cos\theta, z\sin\theta)$ into its x and y components. Let $f_x(z, \theta) = z\cos\theta$ and $f_y(z, \theta) = z\sin\theta$ so that $f(z, \theta) = (f_x(z, \theta), f_y(z, \theta))$. The Jacobian matrix in this case is the following.

$$\begin{bmatrix} \partial_z f_x & \partial_\theta f_x \\ \partial_z f_y & \partial_\theta f_y \end{bmatrix} = \begin{bmatrix} \cos\theta & -z\sin\theta \\ \sin\theta & z\cos\theta \end{bmatrix}$$

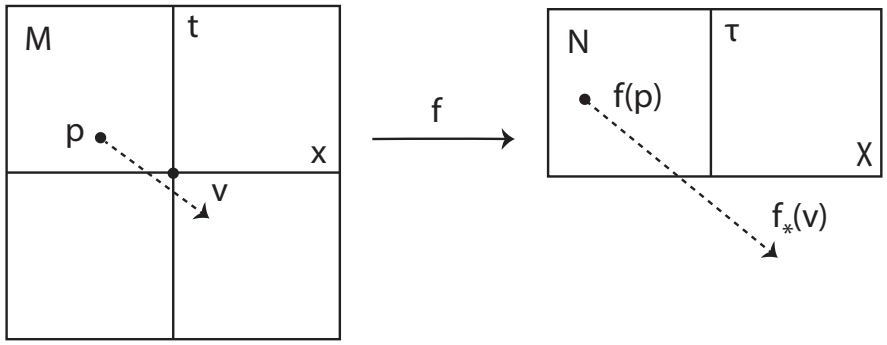

Figure 4.3: The diffeomorphism $f : M \to N$ pushes forward a vector $v = [-2, 3]$ at $p = (1, -2)$ to the vector $f_*(v) = [-2\exp(1), 6]$ at $f(p) = (\exp(1), -4)$.

The push forward vector $f_*(v)$ is the result of matrix multiplication of the vector $v = [v_z, v_\theta]$ by the Jacobian matrix. This comes out as the following.

$$[v_z\partial_z f_x + v_\theta\partial_\theta f_x, v_z\partial_z f_y + v_\theta\partial_\theta f_y] = [v_z\cos\theta - v_\theta z\sin\theta, v_z\sin\theta + v_\theta z\cos\theta]$$

Suppose one has the vector $v = [-4, -1]$ at the point $p = (3, 0)$ in M. So $f_*(v) = [(-4)(1) - (-1)(3)(0), (-4)(0) + (-1)(3)(1)] = [-4, -3]$ at $f(p) = (3, 0)$ (see Figure 4.4). We will return to this example again soon.

Given a diffeomorphism $f : M \to N$, we now know how to push forward a vector v at $p \in M$ to a vector $f_*(v)$ at $f(p) \in N$. One can also use f to pull back any vector w at $f(p) \in N$ to a vector $f^*(w)$ at $p \in M$. We do this by pushing forward w using the inverse of f and setting $f^*(w) = f_*^{-1}(w)$. It is a basic fact that if we push forward a vector with f and then pull it back with f, the result is just the vector we started with, i.e. $f^*(f_*(v)) = v$ for any vector v at any point in M. Similarly, if we pull back a vector with f and then push it forward with f, the result is just the vector we started with, i.e. $f_*(f^*(w)) = w$ for any vector w at any point in N.

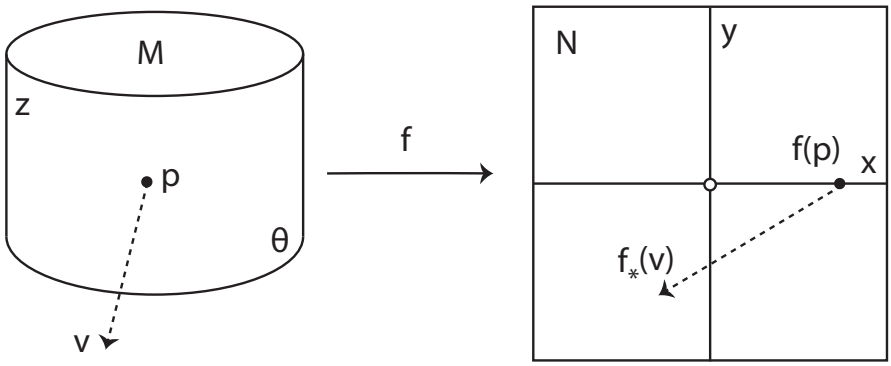

Figure 4.4: The diffeomorphism $f : M \to N$ pushes forward a vector $v = [-4, -1]$ at $p = (3, 0)$ to the vector $f_*(v) = [-4, -3]$ at $f(p) = (3, 0)$.

4.4 Isometries

Let M and N be manifolds such that there is a diffeomorphism $f : M \to N$. Suppose M and N are endowed with smooth metrics g and h respectively. Just as we can use f to transfer vectors between at any point $p \in M$ and the corresponding point $f(p) \in N$, we can also use it to transfer metrics as well. We use f to **pull back** the metric h at $f(p)$ to the metric $f^*(h)$ at p. How do we do this? Let v and w be any vectors at the point p. We need to find a suitable number to assign to the quantity $f^*(h)(v, w)$. Here is how we provide that number. We can push forward the vectors v and w at p to the vectors $f_*(v)$ and $f_*(w)$ at $f(p)$. The metric h then assigns a number $h(f_*(v), f_*(w))$ to these vectors. Whatever this number winds up being, we assign it to the quantity $f^*(h)(v, w)$. In this way, we have now defined a new metric $f^*(h)$ on M. We can also **push forward** the metric g at p to the metric $f_*(g)$ at $f(p)$. We just set $f_*(g)$ to $f^{-1*}(g)$ – the pull back of g using the inverse of f. As with vectors, if we push forward a metric on M with f and then pull it back with f, the result is just the metric we started with, i.e. $f^*(f_*(g)) = g$ for any metric g on M. Similarly, if we pull back a metric with f and then push it forward with f, the result is just the metric we started with, i.e. $f_*(f^*(h)) = h$ for any metric h on N.

We can now define what it means for (M, g) and (N, h) to have the same

structure. We say a diffeomorphism $f : M \to N$ is an **isometry** if, at any point $p \in M$ and any vectors v, w at p, we have $g(v, w) = f^*(h)(v, w)$. This definition makes sense. The metric g assigns a number $g(v, w)$ to the vectors v and w at p. We can use f to pull back the metric h on N to a metric $f^*(h)$ on M. If $f^*(h)$ just is g, then (M, g) to (N, h) share the same metric structure – they are **isometric**. Here's an equivalent way to formulate the definition of isometry. Suppose one uses f to push forward the vectors v and w at p to vectors $f_*(v)$ and $f_*(w)$ at $f(p)$. The metric h then assigns them a number $h(f_*(v), f_*(w))$. This number is the same as the number $g(v, w)$ if and only if (M, g) to (N, h) are isometric. Let's take a look at some examples.

We start with a pair of Riemannian manifolds. Consider again the half cylinder and punctured plane. Let's define a metric g on the half cylinder M as follows: at each point $(z, \theta) \in M$ and for any vectors $v = [v_z, v_\theta]$ and $w = [w_z, w_\theta]$ at the point, let $g(v, w) = v_z w_z + z^2 v_\theta w_\theta$. On the punctured plane N define the Euclidean metric e: at each point in N and for any vectors $v = [v_x, v_y]$ and $w = [w_x, w_y]$ at the point, let $e(v, w) = v_x w_x + v_y w_y$. It is not to difficult to verify that (M, g) and (N, e) are isometric. One can use the diffeomorphism $f : M \to N$ defined by $f(z, \theta) = (z \cos\theta, z \sin\theta)$ that we have already explored above. Try it! Here's a way to see the idea intuitively.

Fix some $k > 0$ and consider a curve $\lambda_k : [0, 2\pi] \to M$ around the cylinder defined by $\lambda(s) = (k, s)$. So $z = k$. What is the length of this curve according to g? Differentiating its components, we find that its tangent vector comes out as $\lambda_k'(s) = [0, 1]$ and so $\|\lambda_k'(s)\| = z^2 = k^2$. Integrating $\sqrt{\|\lambda_k'(s)\|} = k$ from $s = 0$ to $s = 2\pi$ gives $2\pi k$. Thus, the length of the curve $\|\lambda_k\|$ around the cylinder goes to zero as $k \to 0$. Such curves are mapped via the isometry f to curves $f \circ \lambda_k : [0, 2\pi] \to N$. The images are circles of radius k going around the "missing" origin in N like ripples in a pond just after a dropped pebble. According to the metric h, the circumference $\|f \circ \lambda_k\|$ of these ripples also goes to zero as $k \to 0$ (see Figure 4.5). We see that even though (M, g) and (N, e) are presented differently, they represent the same structure.

Let's do one more check to verify that f is an isometry. Earlier, we saw that the vector $v = [-4, -1]$ at $p = (3, 0)$ in M is pushed forward to the vector $f_*(v) = [-4, -3]$ at $f(p) = (3, 0)$ in N (recall Figure 4.4). The (squared) length $\|v\| = v_z^2 + z^2 v_\theta^2$ of $v = [-4, 1]$ according to g is comes out as $\|v\| = (-4)^2 + (3)^2(-1)^2 = 25$. But we see that the (squared) length $\|f_*(v)\| = v_x^2 + v_y^2$ of $f_*(v) = [-4, -3]$ according to h also is comes out as $\|f_*(v)\| = (-4)^2 + (-3)^2 = 25$ just as we would expect.

We now consider isometric spacetimes. We will extend the other example

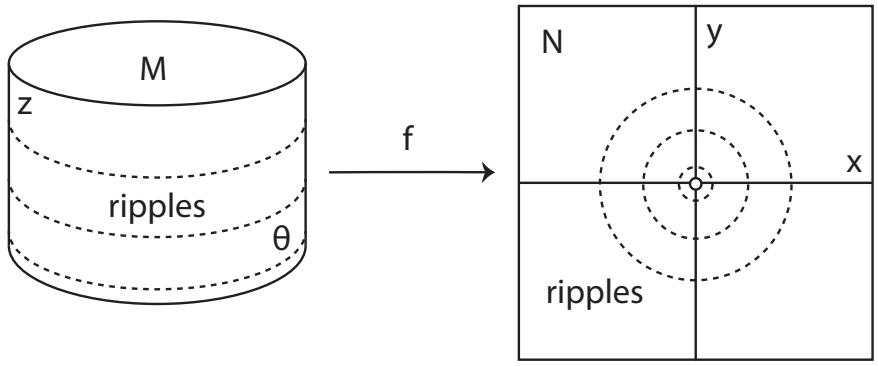

Figure 4.5: The isometry f maps ripples in M to ripples in N.

we have have been working with in the chapter: the diffeomorphic manifolds $M = \mathbb{R}^2$ and $N = \mathbb{R}^+ \times \mathbb{R}$. Let (M, η) be two-dimensional Minkowski spacetime in standard (t, x) coordinates. Let $N = \mathbb{R}^+ \times \mathbb{R}$ be given in (τ, χ) coordinates as before. As we have seen, the map $f : M \to N$ defined by $f(t, x) = (\exp(t), 2x)$ is a diffeomorphism (recall Figure 4.1). We have already explored how to push forward any vector v at point $p \in M$ to the vector $f_*(v)$ at point $f(p) \in N$ (recall Figure 4.3). We now use f to push forward the metric η on M to some metric $f_*(\eta)$ on N so that f is an isometry.

Let v and w be any vectors at the point $f(p)$. We will pull back v and w at $f(p)$ to the vectors $f^*(v)$ and $f^*(w)$ at p. The metric η then assigns a number $\eta(f^*(v), f^*(w))$ to these vectors. Whatever this number winds up being, we will assign it to the quantity $f_*(\eta)(v, w)$. In this way, we can define a metric $f_*(\eta)$ on N. Let's work out an expression for this metric in the current example.

Let $v = [v_\tau, v_\chi]$ and $w = [w_\tau, w_\chi]$ be any vectors at the point $f(p) = (\tau, \chi)$. As we have seen, pulling back these vectors using f is equivalent to pushing them forward using the inverse of f. It is not difficult to see that in our case, the inverse $f^{-1} : N \to M$ comes out as $f^{-1}(\tau, \chi) = (\ln(\tau), \chi/2)$. We now separate $f^{-1}(\tau, \chi) = (\ln(\tau), \chi/2)$ into its t and x components. Let $f_t^{-1}(\tau, \chi) = \ln(\tau)$ and $f_x^{-1}(\tau, \chi) = \chi/2$ so that $f^{-1}(\tau, \chi) = (f_t^{-1}(\tau, \chi), f_x^{-1}(\tau, \chi))$. The

Jacobian matrix is given by the following.

$$\begin{bmatrix} \partial_\tau f_t^{-1} & \partial_\chi f_t^{-1} \\ \partial_\tau f_x^{-1} & \partial_\chi f_x^{-1} \end{bmatrix} = \begin{bmatrix} 1/\tau & 0 \\ 0 & 1/2 \end{bmatrix}$$

The push forward vector $f_*^{-1}(v)$ at $p \in M$ is then result of matrix multiplication of the vector $v = [v_\tau, v_\chi]$ at $f(p) \in N$ by the Jacobian matrix. This comes out as the following.

$$[v_\tau \partial_\tau f_t^{-1} + v_\chi \partial_\chi f_t^{-1}, v_\tau \partial_\tau f_x^{-1} + v_\chi \partial_\chi f_x^{-1}] = [v_\tau/\tau, v_\chi/2]$$

So pushing forward the vectors v and w, we get $f_*^{-1}(v) = [v_\tau/\tau, v_\chi/2]$ and $f_*^{-1}(w) = [w_\tau/\tau, w_\chi/2]$. But as we mentioned, using f^{-1} to push forward vectors v and w is the same as using f to pull them back. So $f^*(v) = [v_\tau/\tau, v_\chi/2]$ and $f^*(w) = [w_\tau/\tau, w_\chi/2]$. Now what number does η assign to the pulled back vectors $f^*(v)$ and $f^*(w)$? It comes out as the quantity $v_\tau w_\tau/\tau^2 - v_\chi w_\chi/4$. This number is what we require $f_*(\eta)$ to assign to v and w. In this way, we have now defined the push forward $f_*(\eta)(v, w) = v_\tau w_\tau/\tau^2 - v_\chi w_\chi/4$ of the metric η. Let's call it h. We see that (N, h) is just Minkowski spacetime (M, η) in disguise. Things "look different" in (N, h) but all the relevant structure is preserved. In what follows, when we speak of "Minkowski spacetime" we are really speaking of any spacetime isometric to the particular presentation $(\mathbb{R}^n, \eta)$ we have been working with.

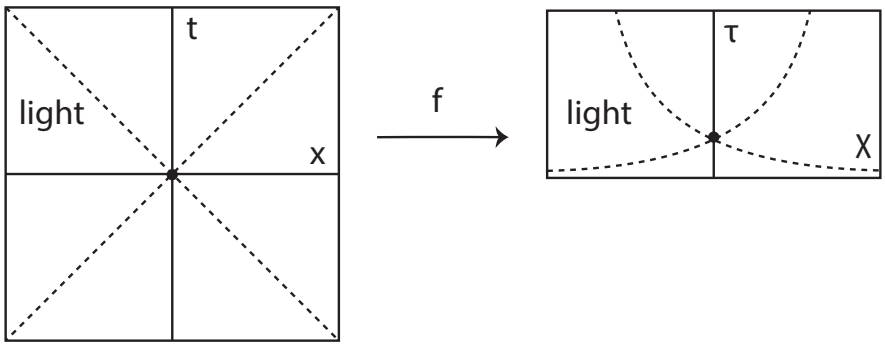

Figure 4.6: The isometry f maps null geodesics in M to null geodesics in N.

We have already seen in our example that the vector $v = [-2, 3]$ at the point $p = (1, 2) \in M$ is pushed forward to the vector $f_*(v) = [-2\exp(1), 6]$ at $f(p) = (\exp(1), -4) \in N$ (recall Figure 4.3). Let's check to make sure that the lengths the vectors v and $f_*(v)$ match up. The metric η assigns v the (squared) length $\|v\| = v_t^2 - v_x^2 = (-2)^2 - 3^2 = -5$. And as we would expect, the metric $h = f_*(\eta)$ assigns the (squared) length $\|f_*(v)\| = v_\tau^2/\tau^2 - v_\chi^2/4 = ((-2\exp(1))^2/\exp(1)^2 - 6^2/4) = 4 - 9 = -5$ as well. We see that the isometry f transfers the spacelike vector v at p to its counterpart spacelike vector $f_*(v)$ at $f(p)$ with the same length.

One can get a better grip on the light cone structure of (N, h) by pushing forward null geodesics in (M, η) via f to null geodesics in (N, h). Consider the null geodesics $\lambda_\pm : \mathbb{R} \to M$ defined by $\lambda_\pm(s) = (s, \pm s)$. We can now compose these geodesics with f to produce the null geodesics $f \circ \lambda_\pm : \mathbb{R} \to M$ defined by $(f \circ \lambda_\pm)(s) = (\exp(s), \pm 2s)$ (see Figure 4.6). We see that the light cones widen rapidly as $\tau \to 0$. Stepping back, one can also verify that for any s, if one pushes forward the tangent vector $\lambda_+'(s)$ of the geodesic λ_+ at $p = \lambda_+(s)$ to the point $f(p)$, the result is just the the tangent vector $(f \circ \lambda_+)'(s)$ of the geodesic $f \circ \lambda_+$ at $f(p)$. A foundational result says that this is true in general for any curve whatsoever. In some cases, this can make pushing forward a vector a piece of cake: no Jacobian matrix needed!

4.5 Symmetries

The (global) **symmetries** of a given mathematical structure are the isomorphisms from the given structure to itself. Manifolds have infinitely many symmetries in this sense. But if we add a metric to a manifold, then the resulting structure can have, at best, only a few non-trivial symmetries and sometimes none at all. We will return to this point in Chapter 7. For now, let's explore the symmetries of the two-dimensional Minkowski universe we have been working with.

Consider (M, η) where $M = \mathbb{R}^2$ in (t, x) coordinates and $\eta(v, w) = v_t w_t - v_x w_x$ for all vectors $v = [v_t, v_x]$ and $w = [w_t, w_x]$ at any point in M. We know the identity map trivially counts as an isometry from (M, η) to itself. Translations and reflections with respect to the t or x coordinates are also isometries. For example, consider the map $f : M \to M$ defined by $f(t, x) = (t + a, x + b)$ where a and b are any real numbers. A vector $v = [v_t, v_x]$ at a point $p = (t, x) \in M$ get pushed forward to the vector $f_*(v) = [v_t, v_x] = v$ at

the point $f(p) = (t + a, x + b)$. So $\eta(v, w)$ for any vectors v and w at p will be the same as $\eta(f_*(v), f_*(w))$ at $f(p)$ which shows that f is an isometry. By choosing the real numbers a and b carefully, one can use the isometry f to map any point $p \in M$ into any other point $q \in M$. So the event "here and now" is just like the event "there and then" because of the symmetries of the Minkowski universe (see Figure 4.7)

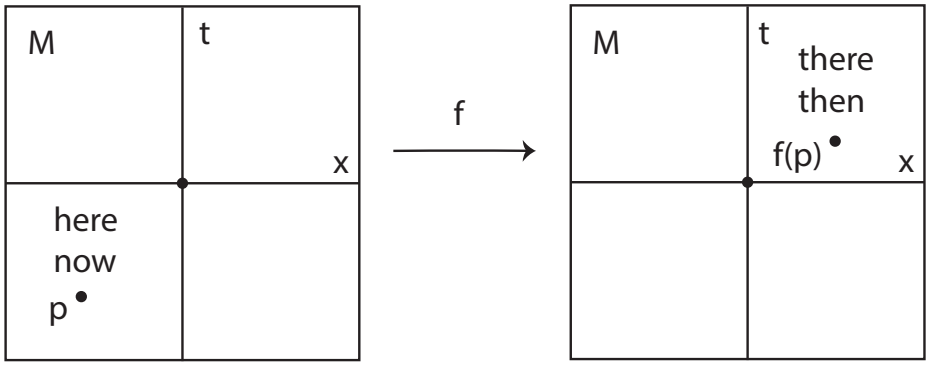

Figure 4.7: The isometry f maps any event "here and now" to any other event "there and then" in the Minkowski universe.

The most interesting symmetry of Minkowski universe is the **Lorentz transformation**. Fix a real number k. We let $\ell : M \to M$ be defined by the following.

$$\ell(t, x) = (t\cosh(k) - x\sinh(k), x\cosh(k) - t\sinh(k))$$

It is not difficult to verify that the map ℓ is an isometry. Try it! To get a better conceptual grip on the significance of the Lorentz transformation, suppose your world-line is the timelike geodesic $\lambda : \mathbb{R} \to M$ given by $\lambda(s) = (s, 0)$ with tangent vector $v = [v_t, v_x] = [1, 0]$. Now suppose your friend's world-line is the timelike geodesic $\gamma : \mathbb{R} \to M$ defined by $\gamma(s) = (s\cosh(k), s\sinh(k))$ with tangent vector $w = [w_t, w_x] = [\cosh(k), \sinh(k)]$ (see Figure 4.8). In the case where $k \neq 0$, it would appear as though your friend is "moving" since $w_x/w_t \neq 0$ while you are "not moving" since $v_x/v_t = 0$. The Lorentz transformation shows that such talk does not make sense. It maps the world-lines of

you and your friend so that from the new perspective, the roles are reversed: it appears as though you are "moving" while your friend is "not moving" (see Figure 4.8). Indeed, under the Lorentz transformation, one can verify that your friend's world-line $\ell \circ \gamma : \mathbb{R} \to M$ comes out as $\ell \circ \gamma(s) = (s, 0)$ which is the same as your world-line from the first perspective. Your friend appears to be "not moving" now as her new velocity vector $\ell_*(w)$ is just $[1, 0]$. On the other hand, we find that your world-line $\ell \circ \lambda : \mathbb{R} \to M$ comes out as $\ell \circ \lambda(s) = (s\cosh(k), -s\sinh(k))$ under the Lorentz transformation. From the new perspective, your velocity vector $\ell_*(v) = [\cosh(k), -\sinh(k)]$ is the mirror image of your friend's original velocity vector w across the $x = 0$ line. In the case where $k \neq 0$, you now appear to be "moving" in the opposite direction that your friend was earlier.

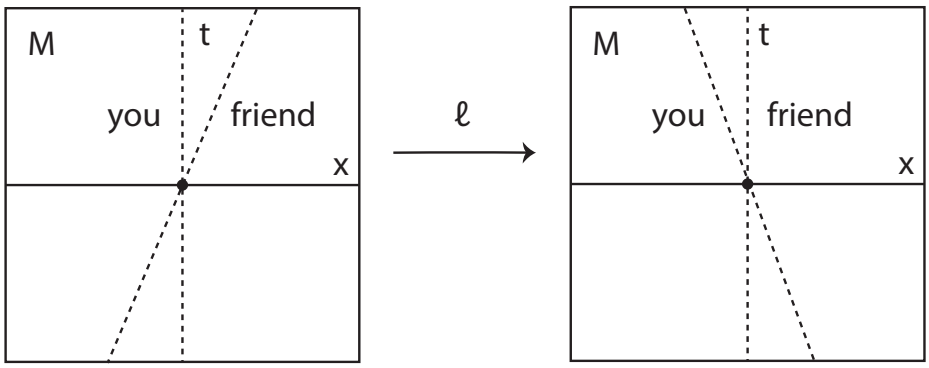

Figure 4.8: The world-lines of you and your friend under a Lorentz transformation.

Stepping back, we see that there is no matter of fact about the "velocity" of you or your friend. Such a notion is not preserved under the symmetries of Minkowski spacetime. But everyone agrees on the speed of light. Consider the null geodesic $\lambda : \mathbb{R} \to M$ defined by $\lambda(s) = (s, s)$. From the first perspective, it has a null tangent vector of $v = [v_t, v_x] = [1, 1]$. So the speed of light must come out as $v_x/v_t = 1$. Under a Lorentz transformation, we find that $\ell \circ \lambda : \mathbb{R} \to M$ comes out as $(\ell \circ \gamma)(s) = (s\cosh(k) - s\sinh(k), s\cosh(k) - s\sinh(k))$. So the tangent vector $(\ell \circ \gamma)'(s)$ becomes the null vector $[\cosh(k) - \sinh(k), \cosh(k) - \sinh(k)]$ which, as we have noted, must be the same as the

push forward $\ell_*(v)$ of the original tangent vector v of the curve $\lambda(s)$. If we let $\ell_*(v)_t$ and $\ell_*(v)_x$ be the coordinate components of this push forward vector such that $\ell_*(v) = [\ell_*(v)_t, \ell_*(v)_x]$, we see that, from the new perspective, the speed of light must also come out as $\ell_*(v)_x/\ell_*(v)_t = 1$.

4.6 Properties

The symmetries of a spacetime can be used to define its invariant properties. Recall that $\mathscr{U}$ is the collection of possible universes: all pairs (M, g) where M is a connected, Hausdorff manifold and g is a smooth Lorentzian metric on M. In a sense, any subcollection $\mathscr{P} \subseteq \mathscr{U}$ can be thought of as a spacetime "property" but we will only be interested in those that are invariant under symmetries. Consider Minkowski spacetime (M, η). We do not want to think of "having $M = \mathbb{R}^n$ as the spacetime manifold" as a property of Minkowski spacetime since it has isometric variants (N, h) where $N \neq \mathbb{R}^n$ (recall Figure 4.6). Let us say that a subcollection $\mathscr{P} \subseteq \mathscr{U}$ is an (invariant) **property** of spacetime if, for any two isometric spacetimes $(M, g), (N, h) \in \mathscr{U}$, we have $(M, g) \in \mathscr{P}$ if and only if $(N, h) \in \mathscr{P}$. On this definition, we see that "having a manifold $\mathbb{R}^n$" is not an invariant property of Minkowski spacetime while "having a manifold diffeomorphic to $\mathbb{R}^n$" does count as an invariant property of Minkowski spacetime.

It will be useful to distinguish "local" and "global" properties. We can do this by considering the "local symmetries" of spacetime. Let (M, g) be a spacetime and let $O \subseteq M$ be any connected, open subset M. The pair (O, g) counts as a spacetime in its own right since O is a connected, Hausdorff manifold. We will call (O, g) a **sub-region** of (M, g). We say that the spacetimes $(M, g), (N, h) \in \mathscr{U}$ are **locally isometric** if, for any event $p \in M$, there is a sub-region (O, g) of (M, g) with $p \in O$ and a sub-region (U, h) of (N, h) such that (O, g) and (U, h) are isometric, and, correspondingly, with the roles of (M, g) and (N, h) reversed. One can show that local isometry, like isometry itself, is an equivalence relation on the collection of possible universes $\mathscr{U}$.

Consider the two-dimensional Minkowski universe (M, g) and any event p. If $N = M - \{p\}$, then (N, g) counts as a sub-region of (M, g). Because of the "missing" point p, the spacetimes (M, g) and (N, g) are not isometric. The manifold M is diffeomorphic to $\mathbb{R}^2$ while N is diffeomorphic to $\mathbb{R}^+ \times S$ (recall Figure 4.4). But (M, g) and (N, g) are locally isometric. Any event

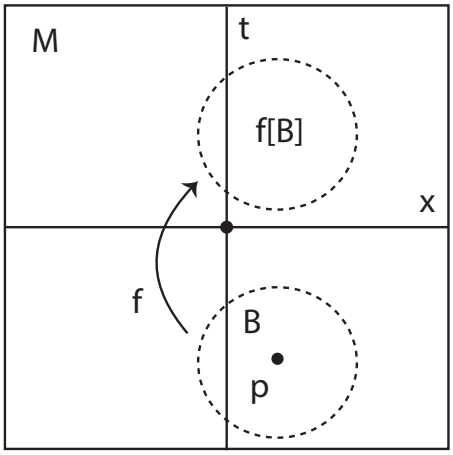

Figure 4.9: The isometry f maps the ball B to a region not containing p.

$q \neq p \in M$ is contained in the sub-region (N, g) of (M, g). So the identity map from N to itself shows that q is contained in a sub-region which is isometric to (N, g). What about event p? For that special case, just consider a unit ball $B \in M$ centered at p and a map that shifts B over sufficiently far away from p. For example, let $f : B \to M$ defined by $f(t, x) = (t + 3, x)$. Since $f[B] \subset N$, this local translation f counts as an isometry (see Figure 4.8). So even a sub-region containing p has an isometric counterpart in (N, h) even though p is "missing" there. We have verified one direction of the local isometry. Now consider any point $r \in N$. There is a sub-universe of (N, g) containing p, namely (N, g) itself, that is isometric to a sub-universe of (M, g). Just take that sub-universe to be (N, g) and consider the identity map from N to N. So (M, g) and (N, g) are locally isometric. Any universe that is locally isometric to Minkowski spacetime we will call **locally Minkowskian**. This counts as an invariant property of spacetime. Moreover, one can show that a spacetime is flat if and only if it is locally Minkowskian.

We say that an (invariant) property $\mathscr{P} \subseteq \mathscr{U}$ is **local** if, for any locally isometric spacetimes $(M, g), (N, h) \in \mathscr{U}$, we have $(M, g) \in \mathscr{P}$ if and only if $(N, h) \in \mathscr{P}$. A property that is not local is **global**. We see that "having a manifold diffeomorphic to $\mathbb{R}^n$" comes out as a global property given the example we just considered. On the other hand, "being flat" comes out as a local property. The properties of "being a vacuum solution of Einstein's equation" and "satisfying an energy condition" (of any type) counts as local.

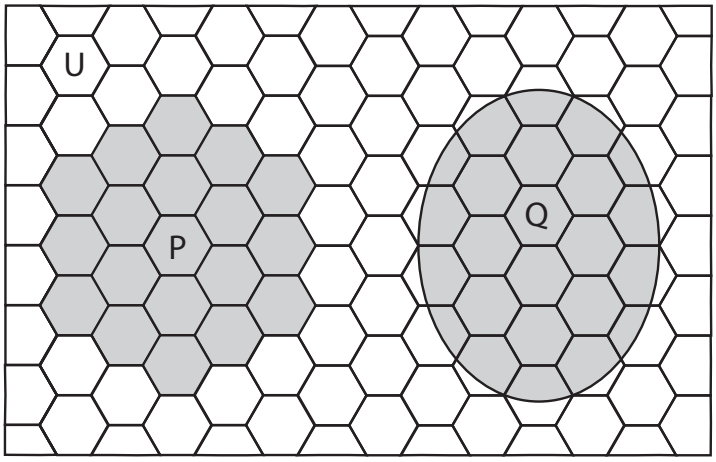

Figure 4.10: Property $\mathscr{P} \subset \mathscr{U}$ is local while property $\mathscr{Q} \subset \mathscr{U}$ is global. Each cell is an equivalence class of locally isometric universes.

Because local isometry is an equivalence relation on $\mathscr{U}$, we can think of a local property as a union of equivalence classes of locally isometric spacetimes. A global property necessarily fails to be such a union (see Figure 4.10). In what follows, a wide variety of global properties will be considered.

4.7 Conclusion

In previous chapters, we have formally built up the key definition of a general relativistic spacetime. This allowed us to construct the collection $\mathscr{U}$ of all spacetimes. In this chapter, we have added another definition that will be central in what follows: isometries. These are isomorphisms (structure preserving maps) between certain elements of $\mathscr{U}$. In order to formulate this definition, we first had to consider the notion of diffeomorphisms, i.e. the isomorphisms with respect to manifolds. A diffeomorphism $f : M \to N$ between the manifolds M and N, gives rise to the associated push forward and pull back maps f_* and f^*. These maps permit one to transfer any vector at any point in one manifold to a corresponding vector at a corresponding point in the other manifold. Formally defining the push forward and pull back maps was a bit hairy. But the return on investment has already been great. Suppose the manifolds M and N are endowed with the metrics g and

h respectively. Then the transfer the vectors between these manifolds can be used to formulate a natural definition: if at any point $p \in M$ and any vectors v, w at p, we have $g(v, w) = h(f_*(v), f_*(w))$, then the diffeomorphism f also counts as an isometry.

With the notion of an isometry in hand, the dividends start rolling in. One can understand the (global) symmetries of a spacetime to be the isometries from that spacetime to itself. We looked at a number of symmetries of Minkowski spacetime including translations, reflections, and the Lorentz transformations. The notion of an isometry also allowed for a precise formulation of the invariant properties of spacetime as well as a useful distinction between local and global varieties of such properties. Additional topics based on the foundational definition of isometry will be explored in Chapters 6 and 7. The former deals with the modal properties of spacetime including the central definition of spacetime maximality. The latter deals with the asymmetry properties of spacetime including local and global types. Before moving to discuss these properties, it will be useful to consider a hierarchy of global properties relating to the causal structure of spacetime.

Chapter 5

Causality

5.1 Introduction

In the Minkowski universe, the causal structure of spacetime is very well-behaved. But the possible shapes and curvatures of spacetime permitted by general relativity permit a wide variety global causal pathologies. For example, it may not be possible to label the two lobes of each light cone as "past" and "future" in a continuous way. Such spacetimes fail to be "time-orientable" and we begin our exploration of the causal structure of spacetime with a look at this property. Under the assumption of time-orientability, we then move to consider a hierarchy of six global causal properties. The lowest two levels of the hierarchy rule out types of "causal loops" in which events can causally influence themselves. The possibility of "time travel" is discussed. The middle two levels of the hierarchy forbid spacetimes that "almost" have causal loops. At these middle levels, the causal structure of spacetime is sufficiently well-behaved to allow for different senses in which the topology of spacetime can be determined from its causal structure. The two highest levels of the hierarchy concern the "stability" of causal structure and a form of causal "dependence" of certain spacetime regions upon others. The latter notion is connected to a kind of causal "determinism" in which the structure of the entire universe depends only on the physical situation at any given instant.

5.2 Orientability

In order to consider the hierarchy of causal conditions in the few sections, we will need a global distinction between the "past" and "future" directions of time. We say that a spacetime (M, g) is **time-orientable** if one can define a continuous timelike vector field on M. If such a vector field exists, then at each point p, the there will be vector v_p pointing in one of two lobes of the light cone. If we label that lobe "the future" at each point and the other lobe "the past" we have the globally defined distinction we are after. Of course, we could switch the labels at each point. This means that if a spacetime is time-orientable, there are two such time **orientations** to choose from. The "problem of the direction of time" concerns the justification of choosing one orientation over another (Callender, 2017). It is an interesting question whether all "physically reasonable" spacetimes must be time-orientable (Bielińska and Read, 2023). Here is an intuitive argument (Geroch and Horowitz, 1979, p. 228-229).

> We observers, in our own local region of spacetime, perceive a preferred future time-direction. Furthermore, there is agreement between different observers as to which time-direction this is. Suppose, then, that one universalizes these local experiences – i.e. one imagines that there could be local observers in all regions of the universe, that each observer would perceive a preferred future time-direction, and that there would be agreement among these observers. One would then conclude that a physically realistic model of our universe must be time-orientable."

All of the spacetimes we have encountered so far have been time-orientable. Let's look at an example which isn't. To construct it, first consider two-dimensional Minkowski spacetime $(\mathbb{R}^2, \eta)$ in standard (t, x) coordinates. This spacetime is time-orientable since it admits a smooth (and therefore continuous) timelike vector field v on M defined by assigning the vector $v_p = [1, 0]$ to each event p. Now let N be the set of points $(t, x) \in \mathbb{R}^2$ such that $-1 < t < 1$ and $0 \leq x \leq 10$. The set N is not a manifold since any point at either of the $x = 0$ and $x = 10$ ends of the strip must fail to have a neighborhood homeomorphic to $\mathbb{R}^2$. But we can construct the manifold M by "gluing" these ends together in curious way. Let E_0 and E_{10} be, respectively, the one-dimensional manifolds consisting of the $x = 0$ and

$x = 10$ portions of N. Any point in either E_0 or E_{10} is characterized by its t coordinate which ranges from -1 to 1. Now consider the diffeomorphism $f : E_0 \to E_{10}$ defined by $f(t) = -t$. Identity the point $p \in E_0$ with the point $f(p) \in E_{10}$ for all $p \in E_0$. Because points are identified in this way, a vector at one end of the strip is identified with a vector at the other end that points in the opposite t direction. Indeed, one can easily verify that any vector $v = [v_t]$ at point p gets pushed forward by the diffeomorphism f to the vector $f_*(v) = [-v_t] = -v$ at the point $f(p)$. Under the identifications, the resulting structure M is a manifold: the famous **Möbius strip** (see Figure 5.1).

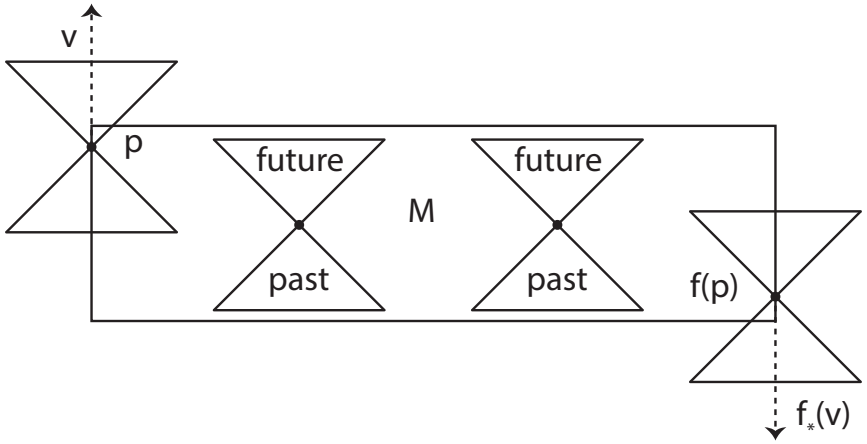

Figure 5.1: The Möbius strip M endowed with the metric η. The vector v at p is identified with the vector $f_*(v)$ at $f(p)$. It is not possible to label all of the lobes of the light cone as "past" and "future" in a continuous way.

At each point on the Möbius strip M, one can define the metric η in the natural way to construct the spacetime (M, η). One can verify there is no problem in doing so under the proposed identifications. Vectors $v_0 = [v_t, v_x]$ and $w_0 = [w_t, w_x]$ at a point $p \in E_0$ are identified with the vectors $v_{10} = [-v_t, v_x]$ and $w_{10} = [-w_t, w_x]$ at the point $f(p) \in E_{10}$. Notice that $\eta(v_0, w_0) = \eta(v_{10}, w_{10})$. But there is a problem in finding a continuous timelike vector field on M which shows that the spacetime fails to be time-orientable. One can label the lobes of the light cone in continuous way in the middle portion of the Möbius strip as in the diagram. But because of the "flip" in t orientation at the ends, there is no way to globally extend the labelling. Because (M, η)

is locally isometric to Minkowski spacetime and but only one of the pair is time-orientable, we see that "being time-orientable" is a global spacetime property.

5.3 Causal Loops

In what follows, we will assume that spacetimes are time-orientable and that a particular orientation has been chosen. Let (M, g) be such a spacetime. A curve $\lambda : I \to M$ is **causal** if its tangent vectors are all either timelike or null. A causal curve is **future-directed** if each tangent vector falls in or on the future lobe of the light cone. We can use the notions of future-directed timelike and causal curves to define a pair of useful relations on the points in the manifold M. For any points $p, q \in M$, we write $p \ll q$ if there is a future-directed timelike curve which has a past endpoint at p and a future endpoint at q. Similarly, we write $p < q$ if there is a future-directed causal curve which has a past endpoint at p and a future endpoint at q. These relations can be used to formulate the four "domains of influence" associated with each event $p \in M$.

$$\begin{aligned} I^+(p) &= \{q \in M : p \ll q\} \\ I^-(p) &= \{q \in M : q \ll p\} \\ J^+(p) &= \{q \in M : p < q\} \\ J^-(p) &= \{q \in M : q < p\} \end{aligned}$$

The set $I^+(p)$ is called the **timelike future** of p and represents the region of spacetime which can be be reached by causal influences emanating from the event p propagating below the speed of light. Similarly, the set $I^-(p)$ is called the **timelike past** of p and represents the region of spacetime which can reach the event p via causal influences propagating below the speed of light. The **causal future** $J^+(p)$ and **causal past** $J^-(p)$ are analogous to $I^+(p)$ and $I^-(p)$ except that now causal influences are permitted to propagate at light speed. It would seem that only events in the causal past of p are empirically accessible from p. An observer has no way of "seeing" events outside this region.

One can generalize these definitions in the natural way so as to apply to sets of events: for any set $S \subseteq M$, define $I^+(S)$ to be the region $\cup\{I^+(p) : p \in S\}$ and similarly for $I^-(S)$, $J^+(S)$, and $J^-(S)$. Clearly, we have $I^+(S) \subseteq J^+(S)$ and $I^-(S) \subseteq J^-(S)$. One can verify that the regions $I^+(S)$ and $I^-(S)$ are always open sets. In Minkowski spacetime, the sets $J^+(p)$ and $J^-(p)$ are closed for any point $p \in M$. But in general, these sets are neither open nor closed. To see this, consider two-dimensional Minkowski spacetime $(\mathbb{R}^2, \eta)$ in standard (t, x) coordinates. Remove the origin point $(0, 0)$ and let and let (M, η) be the resulting spacetime. The point $p = (2, 2)$ is such that $J^-(p)$ is: (i) not open since the event $q = (1, 1)$ is in $J^-(p)$ and yet every neighborhood of q extends outside of $J^-(p)$ and (ii) not closed since the event $r = (-1, -1)$ is not in $J^-(p)$ and yet every neighborhood of r extends inside $J^-(p)$ (see Figure 5.2).

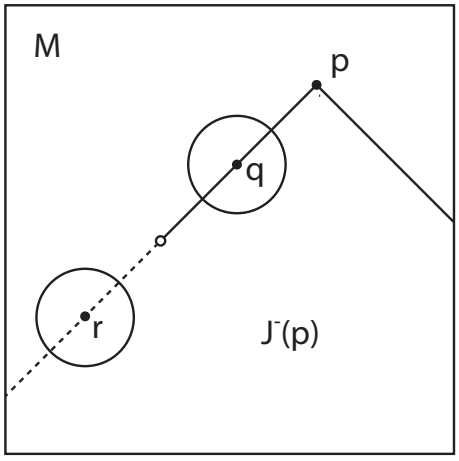

Figure 5.2: The causal past $J^-(p)$ of the p is neither open nor closed. The region contains q but every neighborhood of q extends outside of $J^-(p)$ and r is not in $J^-(p)$ and yet every neighborhood of r extends inside $J^-(p)$.

Consider again an arbitrary time-orientable spacetime (M, g). For any $p \in M$, let $\lambda : \mathbb{R} \to M$ be the smooth curve defined by $\lambda(s) = p$ for all $s \in \mathbb{R}$. The curve counts as a null curve since its tangent vector must be the zero vector for all $s \in \mathbb{R}$. Given the way we have set things up, the curve is both future-directed and past-directed. It follows that $p \in J^+(p)$ and $p \in J^-(p)$. In contrast, since timelike curves cannot have vanishing tangent vectors by definition, we see that $p \in I^+(p)$ and $p \in I^-(p)$ do not hold in general. But

such statements can be true when a type of "time travel" is present in the spacetime (M, g). Let's now explore this notion.

A future-directed causal curve $\lambda : I \to M$ is **closed** if its tangent vectors are nowhere vanishing and there are distinct $s_0, s_1 \in I$ such that $\lambda(s_0) = \lambda(s_1)$. Let's consider an example of a closed timelike curve (CTC). Let M be the product $S \times \mathbb{R}$ in (t, x) coordinates where $0 \leq t \leq 2\pi$ and $t = 0$ is identified with $t = 2\pi$. The spacetime (M, η) is a type of "rolled up" Minkowski universe where the time coordinate t has a circular structure (see Figure 5.3). Now consider an observer, Marty, whose world-line is the CTC $\lambda : [0, 2\pi] \to M$ defined by $\lambda(s) = (s, 0)$. So differentiating the curve components gives a tangent vector of $\lambda'(s) = [1, 0]$ that we will take to be future directed. So $\sqrt{\|\lambda'(s)\|} = 1$ which, when integrated from $s = 0$ to $s = 2\pi$, results in an elapsed time of $\|\lambda\| = 2\pi$ years. Marty begins his journey at the event $p = \lambda(0)$, then travels into the future for 2π years, and finally reaches the end his journey at $\lambda(2\pi)$ which is just the event p where he started. We see that Marty has not really gone "back to the future" but rather "forward to the past."

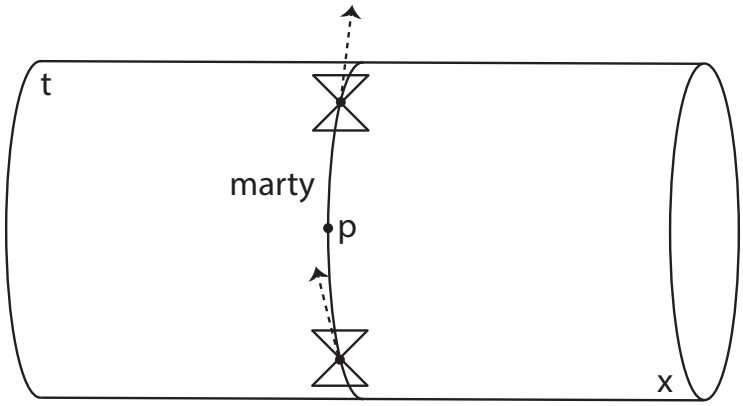

Figure 5.3: Marty begins and ends his journey at the very same event p. His tangent vector points in the future direction at every point along the curve.

Let us say that a spacetime (M, g) without CTCs satisfies the **chronology** condition. One can show that any spacetime with compact manifold (e.g. the torus $S \times S$) must violate chronology (Geroch, 1967). Moreover, for any n-dimensional spacetime (M, g) where $n \geq 3$, there is a spacetime

(M, g') with CTCs (Manchak, 2016c). This means that even a manifold like $\mathbb{R}^4$ admits a metric that fails to satisfy the chronology condition. Perhaps the most famous example is the one due to Gödel (1949). Einstein famously responded to such chronology violating spacetimes as follows: "It will be interesting to weigh whether these are not to be excluded on physical grounds" (Einstein, 1949). To be sure, the physical situation along a CTC will be constrained so as to be consistent. But the possibility of "time travel" remains an open question (Smeenk et al., 2023; Earman et al., 2024).

Let us say that a spacetime without closed causal curves satisfies the **causality** condition. One finds that the causality condition is equivalent to the requirement that for all $p \in M$, it is the case that $J^+(p) \cap J^-(p) = \{p\}$. It is immediate that causality implies chronology. But the two conditions are not equivalent. This can be seen by "rolling up" two-dimensional Minkowski spacetime in one of the null directions. We will do this by considering some non-standard coordinates. Let M and N be two copies of $\mathbb{R}^2$ in (θ, u) and (t, x) coordinates respectively. Let $f : M \to N$ be the diffeomorphism defined by $f(\theta, u) = (\theta + u, \theta - u)$. Now consider the metric η on N given in the usual (t, x) coordinates. We'd like to pull this metric back to define a metric g on M. Let $v = [v_\theta, v_u]$ and $w = [w_\theta, w_u]$ be vectors at a point $p \in M$. We can push forward these vectors via f to the vectors $f_*(v)$ and $f_*(w)$ at the point $f(p)$. We first separate f into its component functions $f_t(\theta, u) = \theta + u$ and $f_x(\theta, u) = \theta - u$ such that $f(\theta, u) = (f_t(\theta, u), f_x(\theta, u))$. The Jacobian matrix comes out as the following.

$$\begin{bmatrix} \partial_\theta f_t & \partial_u f_t \\ \partial_\theta f_x & \partial_u f_x \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

The push forward vector $f_*(v)$ is the result of matrix multiplication of the vector $v = [v_\theta, v_u]$ by the Jacobian matrix. This comes out as the following.

$$[v_\theta \partial_\theta f_t + v_u \partial_u f_t, v_\theta \partial_\theta f_x + v_u \partial_u f_x] = [v_\theta + v_u, v_\theta - v_u]$$

So $f_*(v) = [v_\theta + v_u, v_\theta - v_u]$ and $f_*(w) = [w_\theta + w_u, w_\theta - w_u]$ at $f(p)$. Now what is $\eta(f_*(v), f_*(w))$? After a bit of algebra, we find that it is simply $2(v_\theta w_u + v_u w_\theta)$. So this is the number that the pull back metric $f^*(\eta)$ assigns to v and w at p. Let $g = f^*(\eta)$ be the metric defined in this way at every point in M. By construction we now have an isometry f from (M, g) to (N, η). This makes sense. The vectors $v = [1, 0]$ and $w = [1, 1]$ at $p = (0, 0) \in M$ are such that $\|v\| = 0$ and $\|w\| = 4$ according to g. So v is null and w is

timelike. But $f_*(v) = [1, 1]$ and $f_*(w) = [2, 0]$ at $f(p) = (0, 0) \in N$ are such that $\|f_*(v)\| = 0$ and $\|f_*(w)\| = 4$ according to η. So $f_*(v)$ is null and $f_*(w)$ is timelike (see Figure 5.4).

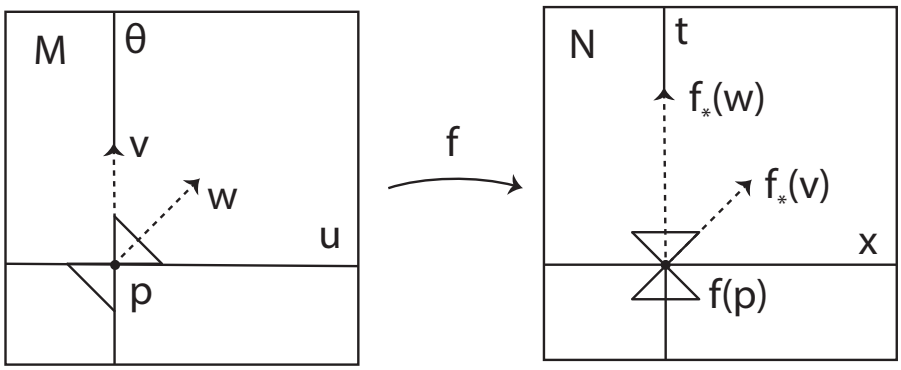

Figure 5.4: The isometry f takes the vectors v and w at $p \in M$ to the vectors $f_*(v)$ and $f_*(w)$ at $f(p) \in N$.

Now consider the $0 \leq \theta \leq 2\pi$ portion of M and identify, for all u, the point $(0, u)$ with $(2\pi, u)$. The resulting manifold is just $S \times \mathbb{R}$. Since the θ coordinate is null, the spacetime $(S \times \mathbb{R}, g)$ has closed null curves. Just consider the curve $\lambda : [0, 2\pi] \to M$ defined by $\lambda(s) = (s, 0)$. We find a (non-vanishing) tangent vector of $\lambda'(s) = [1, 0]$ which, as we have seen, is null according to g. Since $\lambda(0) = \lambda(2\pi)$, the curve is closed and we see that the spacetime violates causality. But one can verify that any future-directed timelike curve must always increase along the u coordinate and thus can never be closed. So the spacetime satisfies chronology.

5.4 Topology from Causality

We now consider a couple of conditions that rule out types of "almost" closed causal curves. For each condition, we will see a sense in which the topology of spacetime can be determined from its causal structure. Let us say that a model (M, g) **distinguishing** if for any distinct $p, q \in M$, it follows that $I^+(p) \neq I^+(q)$ and $I^-(p) \neq I^-(q)$. One can show that every distinguishing

spacetime satisfies causality but not the other way around. Let M be the product $\mathbb{R} \times S$ in (t, θ) coordinates where $0 \leq \theta \leq 2\pi$ and $\theta = 0$ is identified with $\theta = 2\pi$. Now consider the spacetime (M, g) where g is the metric defined by $g(v, w) = v_t w_\theta + v_\theta w_t - t^2 v_\theta w_\theta$ for all vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$. We see that the vector $v = [0, 1]$ has a length of $\|v\| = -t^2$ at any point (t, θ). So v is spacelike except at $t = 0$ where it is null. On the other hand, the vector $w = [1, 0]$ is null at every point in the spacetime. The result is that at $t = 0$ the light cones are at a 45 degree angle but "tipped" to allow for a single closed null curve. As t increases in absolute value, the light cones close up rapidly but in a way that keeps the t direction null. Now remove the point $(0, 0)$ from M and let N be the resulting manifold. Because of the "missing" point, the single closed null curve no longer closes ensuring that (N, g) satisfies causality. But the distinguishability condition is violated since the distinct points $p = (0, 1)$ and $q = (0, 2)$ are such that $I^-(p) = I^-(q)$ which is just the $t < 0$ region of N (see Figure 5.5).

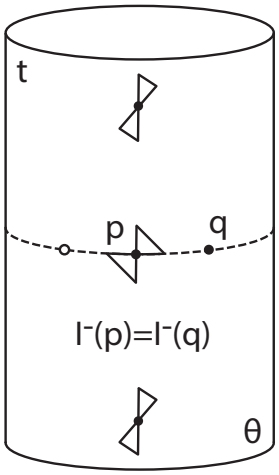

Figure 5.5: The spacetime violates the distinguishing condition since the distinct points p and q share the same timelike past $I^-(p) = I^-(q)$.

Spacetimes that satisfy the distinguishing condition are sufficiently well behaved that there is a sense in which the causal structure determines the topology spacetime. In fact, it even determines all manifold structure. We say a bijection $f : M \to N$ is a **causal isomorphism** between spacetimes (M, g) and (N, h) if, for all $p, q \in M$, the following holds: $p \ll q$ if and only if $f(p) \ll f(q)$. If there is a causal isomorphism between the spacetimes

(M, g) and (N, h), then they share the same causal structure. In general, a causal isomorphism between the spacetimes need not be an isomorphism of another kind of structure (homeomorphism, diffeomorphism, etc). To see this, let (M, g) be Marty's spacetime allowing for time travel. One can show that for any events $p, q \in M$, the $p \ll q$. Now remove a point $r \in M$ and let (N, h) be the resulting spacetime. Again, we find that $p, q \in N$, the $p \ll q$. This means that any bijection $f : M \to N$ whatsoever be must be a causal isomorphism. (We know such a bijection exists since the two manifolds both have continuum many points.) But because of the "missing" event r in (N, h), it doesn't have the same topology as (M, g). Even if were to require the spacetimes (M, g) and (N, h) to be chronological or causal, similar examples could be constructed showing that a causal isomorphism need not preserve other structure. But as we move up the causal hierarchy, everything changes at the level of the distinguishing condition.

Let (M, g) and (N, h) be spacetimes. We say a diffeomorphism $f : M \to N$ is a **conformal isometry** if $f^*(h) = \Omega^2 g$ for some smooth positive function $\Omega : M \to \mathbb{R}$. Here, the function Ω is called a **conformal factor** on M. If spacetimes (M, g) and (N, h) are related by a conformal isometry $f : M \to N$, then the light cone structure as determined by g at any point $p \in M$ is such that, when pushed forward to $f(p) \in N$, it is the same as the light cone structure as determined by h there. A conformal isometry is somewhere between a diffeomorphism and an isometry. All manifold structure is preserved and, in addition, so is the causal structure determined by the metric. But the geodesic structure determined by the metric may not be preserved as it is in an isometry. We are now ready to state a foundational result: a causal isomorphism between distinguishing spacetimes must be a conformal isometry (Malament, 1977a). When attention is restricted to distinguishing spacetimes, information concerning which events are causally related to which others encodes all manifold (including topological) structure.

A spacetime (M, g) satisfies the **strong causality** condition if, for each event $p \in M$ and any neighborhood O of p, there is a smaller neighborhood $U \subset O$ of p such that no future-directed causal curve that begins in U and leaves it, ever returns. One can show that strong causality implies distinguishability. The other direction does not hold. To see this, consider Marty's spacetime (M, η) once more in which $M = S \times \mathbb{R}$ in (t, x) coordinates where $0 \leq t \leq 2\pi$ and $t = 0$ is identified with $t = 2\pi$. Now remove the slits $S_0 : \{(0, x) : x \leq 1\}$ and $S_1 = \{(1, x) : x \geq 0\}$ from M to produce the manifold N. One can show that the spacetime (N, g) satisfies distin-

guishability but not strong causality. Consider the point $p = (1/2, 1/2)$ and an open neighborhood O of p consisting of an open ball centered at p with radius $1/2$. Then every neighborhood U of p that fits inside O will be such that there is a causal curve that starts in U, leaves it, and then returns (see Figure 5.6).

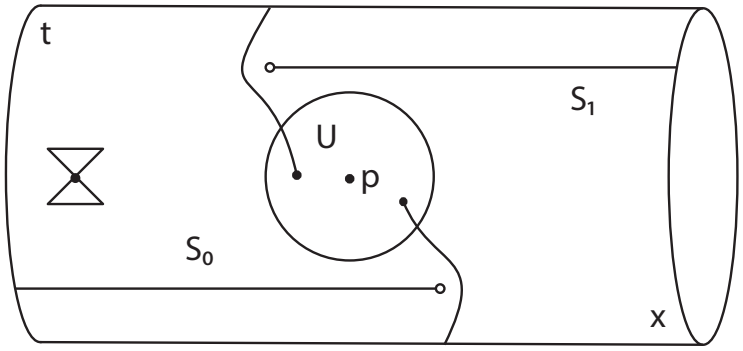

Figure 5.6: The spacetime violates strong causality condition since any sufficiently small neighborhood U of p will be such that there is a future-directed causal curve that starts in U, leaves it, and then returns.

The strong causality condition ensures that certain causal curves cannot be "imprisoned" inside a compact set. Let $\lambda : I \to M$ be a future-directed causal curve in the spacetime (M, g). A point $p \in M$ is a **future endpoint** of λ if, for every neighborhood O of p, there exists a $s_0 \in I$ such that $\lambda(s) \in O$ for all $s > s_0$. Notice that a future endpoint of a causal curve need not be a part of the curve itself. A causal curve is **future inextendible** if it does not have a future endpoint. Similarly, one can define a **past endpoint** and a **past inextendible** curve. A causal curve is **inextendible** if it is both future and past inextendible. Let (M, g) be a spacetime satisfying the strong causality condition and let $\lambda : I \to M$ be a future-directed causal curve such that the image of λ is contained in a compact set $K \subset M$. Then the curve λ has past and future endpoints in K. In other words, a strongly causal spacetime cannot imprison a (past or future) inextendible causal curve in a compact set.

Strongly causal spacetimes have another interesting property connected

to a type of "causal topology" one can define on the manifold. Let (M, g) be a spacetime and, for any $p, q \in M$, define the set $O(p, q)$ to be the intersection $I^+(p) \cap I^-(q)$. Let σ be the collection of all subsets of M that can be expressed as a union of sets of the form $O(p, q)$ for $p, q \in M$. The collection σ counts as a topology on M, called the **Alexandrov topology**, which is always a subset of manifold topology τ. In general, we find $\sigma \neq \tau$. For example, in Marty's spacetime (M, g) we have $I^-(p) = I^+(p) = M$ for any point p. So $O(p, q) = M$ for any $p, q \in M$. The only subsets that can be expressed as a union of sets of the form $O(p, q)$ are M and $\varnothing$ (since it is the union of an empty collection of sets). So σ is the trivial topology which is very different from the manifold topology τ. Notice, for example, that the trivial topology is not Hausdorff. A foundational result is this: a spacetime satisfies strong causality if and only if $\sigma = \tau$ if and only if σ is Hausdorff. So one can explicitly define the topological structure of a strongly causal spacetime from its causal structure.

5.5 Stability and Dependence

A spacetime (M, g) satisfies the **stable causality** condition if there is a smooth function $t : M \to \mathbb{R}$ such that for any distinct points $p, q \in M$, if $p \in J^-(q)$, then $t(p) < t(q)$. The function t is called a **global time function**. So along any future-directed causal curve, the "time" t always increases. Stable causality gets its name because there is a sense in which it is equivalent to the condition that "nearby" spacetimes are chronological. We will come back to this idea in a later chapter. For now, we note that stable causality implies strong causality but not the other way around. Consider again Marty's spacetime (M, η) in which $M = S \times \mathbb{R}$ in (t, x) coordinates where $0 \leq t \leq 2\pi$ and $t = 0$ is identified with $t = 2\pi$. In the last example we removed the slits $S_0 : \{(0, x) : x \leq 1\}$ and $S_1 = \{(1, x) : x \geq 0\}$ from M (recall Figure 5.6). In this example, we remove these slits again as well as the slit $S_2 : \{(2, x) : x \leq 1\}$ from M to produce the manifold N. One can show that the spacetime (N, η) fails to be stably causal. But it does satisfy strong causality: the series of three slits ensures that one can find a sufficiently small neighborhood U of any point p such that no future-directed causal curve that begins in U and leaves it, ever returns (see 5.7).

One can construct similar examples to show that there are actually an infinite number of levels of the causal hierarchy in between strong causality

and stable causality (Carter, 1971). For more on these levels, as well as others not covered in our presentation, we refer the reader to the comprehensive review given by Minguzzi (2019). In what follows, we will make use of this result: if (M, g) is stably causal and C is any closed proper subset of M for which $M - C$ is connected, then (N, g) is a stably causal spacetime where $N = M - C$. To see why, just take any global time function $t : M \to \mathbb{R}$ for (M, g) and restrict its domain to obtain the global time function $t_N : N \to \mathbb{R}$ for (N, g). So Minkowski spacetime with a point removed is stably causal.

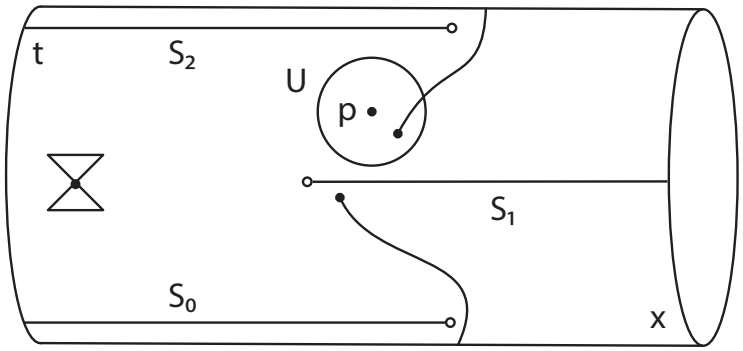

Figure 5.7: The spacetime satisfies the strong causality condition since a sufficiently small neighborhood U of any point p will be such that no future-directed causal curve that begins in U and leaves it, ever returns

Our next condition sits atop the causal hierarchy and ensures a sense of "determinism" in spacetime. We say a spacetime (M, g) is **causally compact** if for any $p, q \in M$, the region $J^+(p) \cap J^-(q)$ is compact. Minkowski spacetime (M, η) is causally compact. Since $J^+(p)$ and $J^-(p)$ are closed for any $p \in M$, we know that $J^+(p) \cap J^-(q)$ is closed for any $p, q \in M$. One can also show that such a set always fits inside some ball in $M = \mathbb{R}^4$ so they are also bounded. Since $J^+(p) \cap J^-(q)$ is both closed and bounded in $M = \mathbb{R}^4$, it must be compact (see Figure 5.8). On the other hand, Minkowski spacetime with a point removed is not causally compact. In such a spacetime, there are events p such that the region $J^-(p)$ is not closed (recall Figure 5.2). One can find some q is this region $J^-(p)$ (just below the "missing point") such that $J^+(p) \cap J^-(q)$ is not closed. So because the manifold is Hausdorff, we

find that $J^+(p) \cap J^-(q)$ is not compact. Let us say that a spacetime satisfies the **global hyperbolicity** condition, if it is causally compact and causal. One can show that any globally hyperbolic spacetime (such as Minkowski spacetime) must be stably causal but not the other way around. Minkowski spacetime with a point removed is not causally compact and therefore not globally hyperbolic. But as we have seen, this spacetime is stably causal.

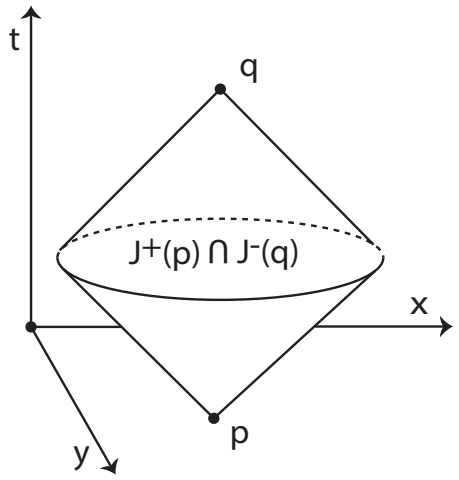

Figure 5.8: In Minowski spacetime, the region $J^+(p) \cap J^-(q)$ is compact for any events p and q.

Older definitions of global hyperbolicity used the strong causality condition in place of the causality condition (Hawking and Ellis, 1973). Eventually, a useful result showed that using the weaker causality condition gives rise to an equivalent definition (Bernal and Sánchez, 2007). It turns out that if attention is restricted to non-compact spacetimes of dimension three or more, the causality condition can be dropped altogether (Hounnonkpe and Minguzzi, 2019). Let's now try to get a better grip on the physical significance of the global hyperbolicity condition.

Let (M, g) be any spacetime and let $S \subseteq M$ be any set. The **future domain of dependence** of S, denoted $D^+(S)$, is the set of points $p \in M$ such that every past-inextendible causal curve through p meets S. The **past domain of dependence** of S, denoted $D^-(S)$, is defined analogously. The **domain of dependence** $D(S)$ of S is the union $D^+(S) \cup D^-(S)$. Since nothing can travel faster than light, any causal influences at each point $p \in D(S)$ must register somewhere on S. Often one restricts attention to sets

$S \subset M$ which are **achronal** in the sense that $I^+(S) \cap S = \varnothing$. Or one might consider cases where $S \subset M$ is a "spacelike surface" is a sense we now make precise.

Let n be the dimension of (M, g). Let S be a manifold of dimension k for $1 \leq k \leq n$. A smooth map $f : S \to M$ is an **embedding** if (i) f is one to one (ii) for all $p \in S$, the push forward map f_* at p is one to one and (iii) the inverse map $f^{-1} : f[S] \to S$ is continuous where $f[S]$ has the subspace topology inherited from M. (Note that we haven't formally defined the push forward map f_* in the case where f is not a diffeomorphism but it is clear how to generalize this notion to the present context of smooth maps.) Condition (i) ensures that the set $f[S]$ does not intersect itself while conditions (ii) and (iii) capture senses in which $f[S]$ does not "almost" intersect itself. One can show that the three conditions are all independent, i.e. one can find examples where any two of the conditions are satisfied while the third is not.

If $S \subseteq M$ and the inclusion map is an embedding, we say that S is an **embedded submanifold** of M. An embedded submanifold $S \subseteq M$ is a **hypersurface** if the dimension k of S is $n - 1$. A hypersurface $S \subseteq M$ is a **spacelike surface** if every curve contained in S is a spacelike curve. A spacelike surface can fail to be an achronal set (e.g. if there are CTCs through the surface). Alternatively, an achronal set $S \subseteq M$ need not be spacelike surface (e.g. it may contain only null curves). In Figure 5.9, a closed, achronal, spacelike surface S is depicted in Minkowski spacetime with a point removed, along with its associated domain of dependence $D(S)$. Because of the "missing" point, there is an inextendible causal curve through p that never registers on S. So $p \notin D(S)$. But any inextendible causal curve through q must meet S at some point $r \in S$ and therefore $q \in D(S)$.

If a spacetime (M, g) has a closed, achronal set $S \subset M$ such that $D(S) = M$, then S is a **Cauchy surface**. One can show a sense in which the physical situation on a Cauchy surface S completely determines the situation at every point in M (Choquet-Bruhat and Geroch, 1969). We will explore this idea in Chapter 12. A foundational result is this: a spacetime (M, g) admits a Cauchy surface S if and only if it is globally hyperbolic (Geroch, 1970a). Another equivalent formulation will introduce a few definitions that will be needed later on. Let (M, g) be a spacetime and let $S \subset M$ be a closed, achronal surface. The **future Cauchy horizon** of S, denoted $H^+(S)$, is defined by taking the closure of $D^+(S)$ and then removing the set $I^-[D^+(S)]$. The **past domain Cauchy horizon** of S, denoted $H^-(S)$, is defined analogously. The **Cauchy horizon** $H(S)$ of S is the union $H^+(S) \cup H^-(S)$. One

can show that $H(S)$ is just the boundary of $D(S)$. If S is nonempty, then $H(S)$ is empty if and only if S is a Cauchy surface. Finally, we note that if a spacetime (M, g) is globally hyperbolic, then one can choose a global time function $t : M \to \mathbb{R}$ such that each surface of constant t is a Cauchy surface $S \subset M$. It follows that the topology of M is just $\mathbb{R} \times S$.

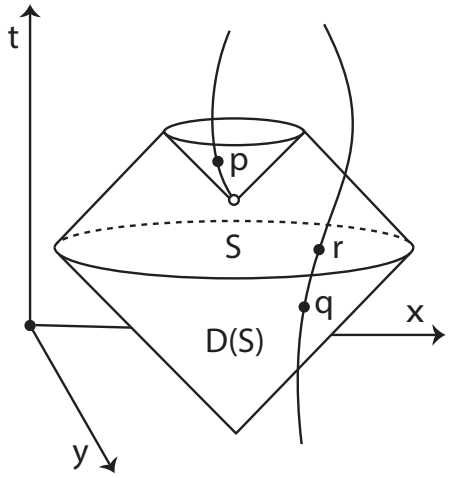

Figure 5.9: Due to the "missing" point, an inextendible causal curve through p never registers on S. So $p \notin D(S)$. But any inextendible causal curve through q must meet S at some point $r \in S$. So $q \in D(S)$.

Global hyperbolicity is a very strong condition. Even so, it has been argued that: "All physically reasonable spacetimes are globally hyperbolic" (Wald, 1984, p. 304). Indeed this is one way of understanding the content of one version of the famous "cosmic censorship" conjecture of Roger Penrose (1979) which rules out "naked singularities" of a certain kind. We will explore this idea later on in Chapters 6 and 12.

We have already seen how removing points from otherwise causally well-behaved spacetime results in a spacetime that is not globally hyperbolic. But such mutilations are not necessary for spacetime to fail to be globally hyperbolic. Marty's time travel spacetime, for example, does not contain a non-empty achronal set so there can be no Cauchy surface. Other examples exist which are causally well-behaved. Consider again two-dimensional anti-de Sitter spacetime (M, g) (recall Section 3.4). Here $M = \mathbb{R}^2$ in (t, x) coordinates and g defined as follows: at each point $(t, x) \in M$ and for any vectors $v = [v_t, v_x]$ and $w = [w_t, w_x]$ at the point, let $g(v, w) = v_t w_t \cosh^2(x) - v_\theta w_\theta$.

The function $f : M \to \mathbb{R}$ defined by $f(t, x) = t$ is a global time function which shows that anti-de Sitter spacetime is stably causal. But it is not globally hyperbolic. Consider, for example, any real number k and the set S defined by $t = k$. One can show that $D(S)$ must be confined to the region $k - \pi/2 < t < k + \pi/2$ and so S is not a Cauchy surface. Recall the wizard whose world-line seemed to appear out of thin air (see Figure 3.7). Such a timelike curve is inextendible and can be chosen so that it is confined to the future of $D(S)$ (see Figure 5.10). So we see a sense in which determinism fails in such a spacetime.

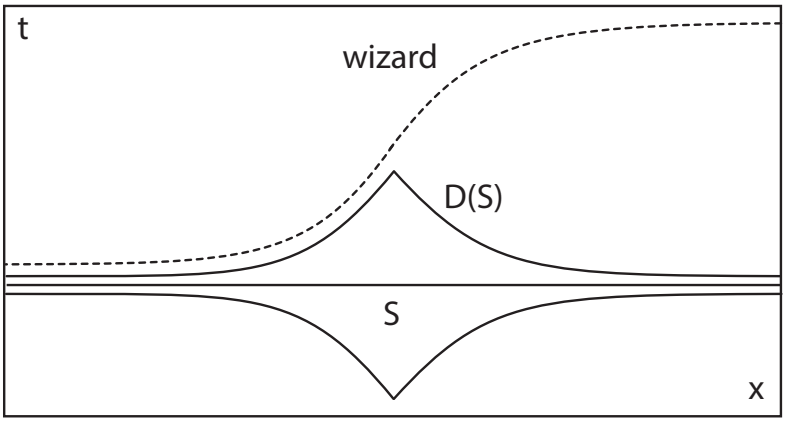

Figure 5.10: Anti-de Sitter spacetime fails to be gloablly hyperbolic. The world-line of a wizard is an inextendible timelike curve that fails to meet S showing $D(S) \neq M$.

5.6 Conclusion

In this chapter, we have considered the causal structure of spacetime. After restricting attention to time-orientable spacetimes, we examined six levels of a causal hierarchy. The lower two levels – chronology and causality – rule out types of causal loops. The middle two levels – distinguishability and strong causality – rule out spacetimes which "almost" have such loops. At these levels, one finds various senses in which the causal structure of spacetime determines its topology. The higher two levels concern "time" on a global

scale. One level – stable causality – requires the existence of such a global time. The other – global hyperbolicity – requires a sense in which there is global time at some instant that can be used to determine the structure of the entire universe.

Just as with the various energy conditions considered in Section 3.5, it will be useful later on to think of the causal hierarchy in terms of subcollections of the collection $\mathscr{U}$ of all spacetimes. Let $(Chron), (Caus), (Dist), (Str), (Stab), (GH) \subset \mathscr{U}$ be the collections of all spacetimes satisfying, respectively, the chronology, causality, distinguishing, strong causality, stable causality, and global hyperbolicity conditions. It is easy to see that each of these properties count as global. This follows since Minkowski spacetime is a member of each collection and Marty's time travel spacetime (which is locally isometric to Minkowski spacetime) is a member of none of them. The hierarchy of causal properties can be summarized as follows (see Figure 5.11).

$$(GH) \subset (Stab) \subset (Str) \subset (Dist) \subset (Caus) \subset (Chron)$$

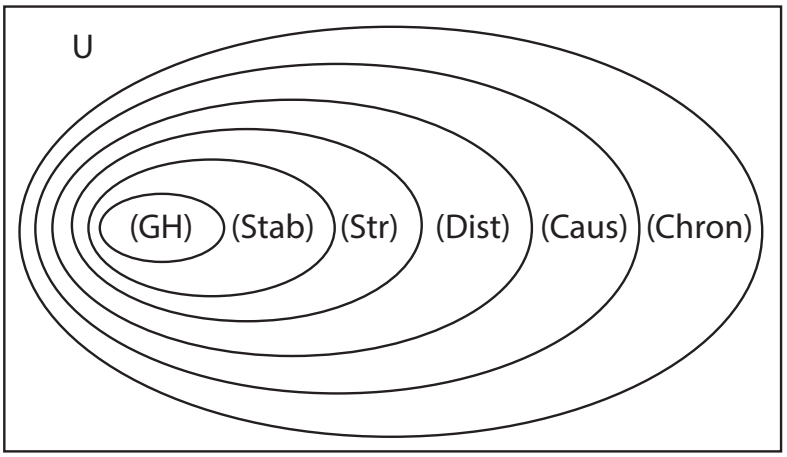

Figure 5.11: The hierarchy of causal properties.

Chapter 6

Holes

6.1 Introduction

Take any spacetime and remove a point from its manifold. The resulting structure is also a spacetime and would seem to be pathological in a variety of senses due to its "missing" point. In what follows, we will explore a hierarchy of "no-hole" conditions that can be used to rule out such examples. We begin with a natural proposal that requires geodesics to be "complete" in the appropriate sense. We find that this route is closed off due to the presence of "singularities" in some physically reasonable spacetimes. We take a look at some examples including a black hole and the big bang. Next, we focus on a series of three weaker modal conditions to rule out spacetime holes: maximality, hole-freeness, and local maximality.

Unlike the energy and causal conditions we have considered so far, the definitions of these modal conditions depend crucially on a background possibility space in the form of some collection of spacetimes. Each of the conditions pinpoint different senses in which spacetime can fail to be "as large as it can be" relative to the chosen background possibility space. In this chapter, we use the standard collection $\mathscr{U}$ in formulating all three definitions. Even within this context, we will see that the justification for the imposition of the conditions is already quite thin. In the second half of the book, we will explore the weakest of the three – spacetime maximality – defined relative to background possibility spaces other than the standard collection $\mathscr{U}$. As we will see, the thin justification for spacetime maximality will get even thinner.

6.2 Geodesic Completeness

Let's start with a natural no-hole condition that concerns the behavior of geodesics. A (smooth) curve $\lambda : I \to M$ in a spacetime (M, g) is **maximal** if there is no curve $\gamma : J \to M$ such that I is a proper subset of J and $\lambda(s) = \gamma(s)$ for all $s \in I$. When attention is restricted to causal geodesics, a curve is maximal if and only if it is inextendible (i.e. has no future or past endpoint). But inextendibility can fail if a maximal causal curve is not a geodesic. In two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates, consider the timelike curve $\lambda : I \to M$ defined by $\lambda(s) = (s + \sqrt{s}, s)$ where I is the interval $(0, \infty)$. The curve is maximal since it cannot be extended in a smooth way through the origin point $p = (0, 0)$. This follows since the tangent vector along the curve is $[1 + 1/(2\sqrt{s}), 1]$ which is undefined at $s = 0$. But λ is not inextendible since the origin point p counts as its past endpoint: for any neighborhood O of p, there exists a $s_0 \in I$ such that $\lambda(s) \in O$ for all $s < s_0$ (see Figure 6.1).

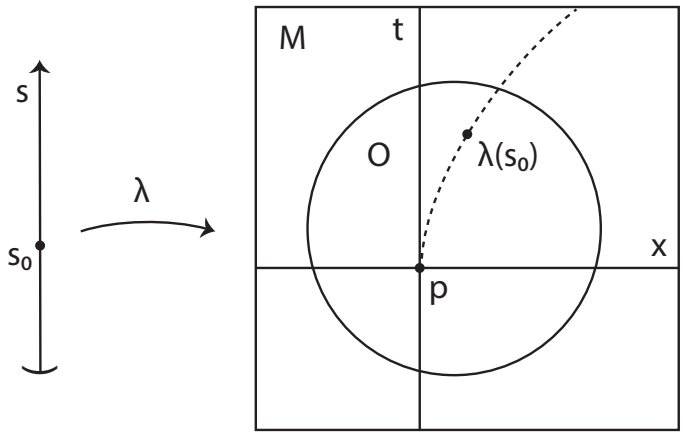

Figure 6.1: The point p is a past endpoint for the curve $\lambda : I \to M$ since, for any neighborhood O of p, there is a $s_0 \in I$ such that $\lambda(s) \in O$ for all $s < s_0$

Now consider a maximal geodesic $\lambda : I \to M$ in a spacetime (M, g). We say it is **incomplete** if $I \neq \mathbb{R}$. Let (N, η) be the $t < 0$ portion of two-dimensional Minkowski spacetime (M, η). Consider the timelike geodesic $\lambda : I \to M$ defined by $\lambda(s) = (s, 0)$ where I is the interval $(-\infty, 0)$. This geodesic is maximal but incomplete. At any event along the curve, the

elapsed time is finite in the future direction. So an observer with such a worldline will have an existence that is "cut short" in a seemingly artificial manner. Let us say that a spacetime with an incomplete geodesic is **geodesically incomplete**. Otherwise it is **geodesically complete**. Insisting on geodesic completeness would seem to be a natural way to rule out "holes" in spacetime. It does exclude virtually all of the problematic examples. But the proposal runs into a couple of major problems.

First, it excludes spacetimes in which holes seemingly cannot exist. In any topological space (X, τ), a point $p \in X$ is an **accumulation point** of an sequence $\{p_n\}$ in X if every open neighborhood of p contains infinitely many points in the sequence. In $\mathbb{R}$, the infinite sequence $\{p_n\}$ defined by $p_n = (n/n + 1)(-1)^n$ for all positive integers n has accumulation points at both -1 and 1. A useful result is the following: if a topological space (X, τ) is second countable, then a set $A \subseteq X$ is compact if and only if every infinite sequence $\{p_n\}$ in A has an accumulation point p in A. Recall that every (standard) spacetime manifold is second countable. We now find something curious: "In a compact spacetime, every sequence of points has an accumulation point, so in a strong intuitive sense, no "holes" can be present. Yet compact spacetimes exist which are geodesically incomplete" (Wald, 1984, p. 215). (We will explore an example of a geodesically incomplete compact spacetime in Chapter 11.)

A second major problem concerns the "singularity theorems" of Hawking and Penrose (1970). In these results, attention is restricted to various collections of "physically reasonable" spacetimes. It is shown that any spacetime in such a collection must be geodesically incomplete. Let's take a closer look.

6.3 Singularities

We start by restricting attention to four-dimensional spacetimes that satisfy (i) chronology, (ii) the strong energy condition, and (iii) the **generic** condition which requires that every causal geodesic encounters a particular type of "effective curvature" at some point (Wald, 1984, p. 227). A flat spacetime fails to satisfy the generic condition but it is thought that a slight perturbation to any spacetime will result in one which satisfies the generic condition. A number of singularity theorems proceed by invoking various "boundary" conditions in addition to (i)-(iii) to ensure geodesic incompleteness. Here is an example. Consider a spacetime (M, g). The **edge** of a closed, achronal set

$S \subset M$ is the collection of points $p \in S$ such that every open neighborhood O of p contains points $q \in I^-(p)$ and $r \in I^+(p)$ and a timelike curve $\lambda : I \to O$ from q to r which fails to intersect S (see Figure 6.2). A **slice** is a closed, achronal set with an empty edge. A spacetime (M, g) with a compact slice represents a "spatially closed" universe and can serve a boundary condition for a singularity theorem: any four-dimensional spacetime satisfying (i)-(iii) with a compact slice must be geodesically incomplete.

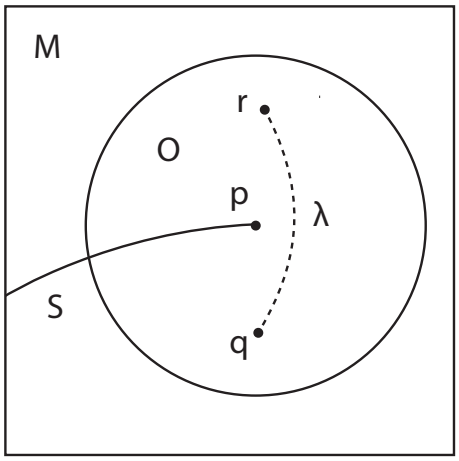

Figure 6.2: The edge of a closed, achronal set $S \subset M$ is the collection of points $p \in S$ such that every open neighborhood O of p contains points $q \in I^-(p)$ and $r \in I^+(p)$ and a timelike curve λ from q to r which fails to intersect S.

Another singularity theorem boundary condition is the existence of a "trapped surface" which forms whenever a sufficiently large amount of matter is contained in a small enough region of spacetime (Schoen and Yau, 1983). The result is a "black hole" structure. To better understand this notion, we now consider a simple example. Consider Minkowski spacetime $(\mathbb{R}^2, \eta)$ in (t, x) coordinates. Let M be the portion of $\mathbb{R}^2$ for which $t^2 - x^2 < 1$. The spacetime (M, η) has the exact same causal structure as a two-dimensional version of the famous Schwarzschild black hole model introduced and named after Karl Schwarzschild (1916). Moreover, because the geodesic structure of (M, η) is sufficiently similar to the Schwarzschild model, we will be able to appreciate the basic features of a black hole within a very simple context. First divide M into four regions: I such that $x > |t|$; II such that $|x| < t$;

III such that $-|x| > t$; and IV such that $x < -|t|$. The "event horizon" is the boundary between these four regions: the union of the $t = x$ and $t = -x$ lines. Geodesics approaching the "singularity" at $t^2 - x^2 = 1$ will be incomplete (see Figure 6.3).

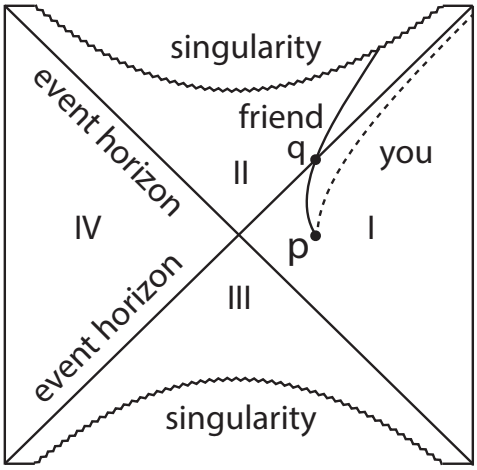

Figure 6.3: From p, you stay forever in region I while you friend crosses the event horizon at q into region II. Any future-inextendible causal curve from an event in region II is destined to hit the "singularity."

Suppose you and a friend are both at event $p = (0, 1)$ in region I. From p it is possible for your world-line to stay within region I and record an infinite elapsed time. Just consider the future-inextendible timelike curve $\lambda : [0, \infty) \to M$ defined by $\lambda(s) = (\sinh(s), \cosh(s))$. Now consider your friend. Suppose her world-line takes her from p to the point $q = (1, 1)$ on the event horizon. Nothing dramatic happens at q. There is no indication that a "point of no return" has been reached by your friend. But every future-directed timelike curve at q must pass through the event horizon and enter region II. And any future-inextendible causal curve from any event in region II is destined to hit the "singularity" at $t^2 - x^2 = 1$. Nothing (not even light) can escape this fate. Meanwhile, since you remain in region I for all time, one can verify that the timelike past of any point along your world-line is contained in regions I and III. So you will never "see" any event in region II. Indeed, it takes you forever just to observe your friend approach q and you never observe this event itself. Tragically, your friend will appear "effectively frozen" for all time (Geroch, 1978, p. 210).

Black holes cannot be observed directly. But the movement of stars near the middle of our Milky Way galaxy seems to indicate that a supermassive black hole exists in the region (Ghez et al., 2000). Given the cosmological data we have collected, we find that our own universe may be best represented by a geodesically incomplete spacetime. So we see that geodesic completeness is a very strong a condition in the sense that it rules out holes that are "physically reasonable" in addition to those that are not. To exclude the latter but not the former, ones needs a more nuanced approach.

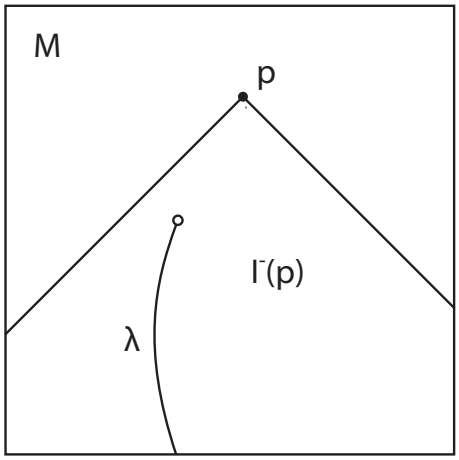

Figure 6.4: The region $I^-(p)$ contains the future-inextendible timelike curve λ. An observer at p can "see" the curve λ fall into the "missing" point.

One influential idea is to formulate a condition to rule out holes using the causal structure of spacetime. Here is one example. Let us say that a spacetime (M, g) has a **naked singularity** if there is a point $p \in M$ such that $I^-(p)$ contains (the image of) a future inextendible timelike curve $\lambda : I \to M$. Intuitively, an observer at p can "see" the curve λ fall into a singularity. This can happen, for example, in Minkowski spacetime with a point removed (see Figure 6.4). On the other hand, the black hole example given above (as well as the Schwarzschild model it is based on) is not nakedly singular. Even if you were to enter region II, you would never be able to witness your friend's final seconds (or vice versa). As Geroch (1978, p. 211) has put it: "The act of "reaching the singularity" is a very personal one."

The "big bang" models of cosmology are also examples of geodesically incomplete spacetimes which are nonetheless free of naked singularities. Let

(M, η) be Minkowski spacetime in (t, x) coordinates and let N be the $t > 0$ portion of M. Now consider the conformal factor $\Omega : N \to \mathbb{R}$ defined by $\Omega(t, x) = t$. The spacetime (N, g) is a big bang model where $g = \Omega^2\eta$. Every maximal timelike geodesic is incomplete in the past direction since it runs into the "singularity" at $t = 0$. One such geodesic is given by the curve $\lambda : (1, \infty) \to N$ where $\lambda(s) = (\ln(s), 0)$. But one can verify that this spacetime does not have a naked singularity. There is no event whose past contains a future inextendible timelike curve (see Figure 6.5).

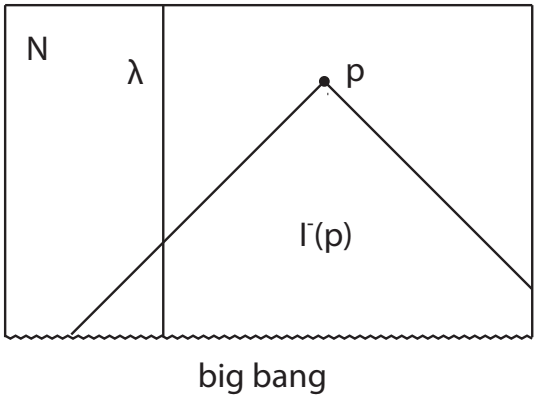

Figure 6.5: The geodesic λ is incomplete in the past direction due to the "big bang." There are no naked singularities since, for any $p \in N$, the region $I^-(p)$ does not contain a future inextendible timelike curve (e.g. λ).

The particular definition of naked singularity we are considering turns out to be equivalent to the non-existence of a Cauchy surface (Earman, 1995, p. 75). We can now see where the Penrose (1979) "cosmic censorship" conjecture gets its name. This conjecture, which is sometimes formulated as the statement "all physically reasonable spacetimes are globally hyperbolic" (Wald, 1984, p. 304), is equivalent to the conjecture that "all physically reasonable spacetimes are free of naked singularities." It is tempting to use the condition of no naked singularities to distinguish between physically reasonable and unreasonable spacetime holes. But this doesn't seem to get to the heart of the matter.

On the one hand, the condition is too strong in the sense that it rules out some spacetimes that it seemingly shouldn't. For example, the rotating

black hole spacetime introduced by Kerr (1963) seems to be physically reasonable but it has a naked singularity. Moreover, anti-de Sitter spacetime is not globally hyperbolic and therefore counts as nakedly singular despite the fact that it is geodesically complete. On the other hand, the condition is too weak in the sense that it fails to rule out examples in which regions of spacetime seem to have been "artificially" removed. To see this, just take two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates and remove the $t \leq 0$ portion of M. Despite the "missing" region, the resulting spacetime is globally hyperbolic and thus free of naked singularities. Since the example has an impeccable causal structure, using that structure alone to appropriately sort varieties of spacetime holes will not work.

There have been other attempts at using the causal structure to sort between physically reasonable and unreasonable spacetime holes but none have been entirely successful (Manchak, 2016a; Doboszewski, 2020). A more fruitful approach is to bring the modal structure of spacetime into the picture. We now look at three conditions of this kind.

6.4 Maximality

Intuitively, what we need is a condition to ensure that "space-time does not arbitrarily stop" (Clarke, 1976, p. 17). The most basic such condition is the requirement of spacetime "maximality" that is the focus of this book. Hawking and Ellis (1973, p. 58) introduce the notion like so: "We have to impose some some condition on our model (M, g) to ensure that it includes all non-singular points of space-time." Another type of justification based on the Leibnizian principles of plenitude and is sufficient reason is given by Geroch (1970b, p. 262): "We may regard [maximality] as a reasonable physical condition to be imposed on models of the universe. (Why, after all, would Nature stop building our universe at M when She could just as well have carried on to build M'?)"

Let us say the spacetime (M, g) has a (proper) **extension** (N, h), if for some proper subset O of N, the spacetime (M, g) is isometric to the spacetime (O, h). A spacetime is **extendible** if it has an extension and **maximal** otherwise. Notice that, unlike all of the spacetime properties considered up to this point, the maximality property is modal in character. Whether or not the spacetime (M, g) has an extension depends crucially on the existence of some other spacetime (N, h). A foundational theorem is this: any spacetime

is either maximal or has a maximal extension (Geroch, 1970b). This result is often used to underpin the Leibnizian justification for spacetime maximality mentioned above (Earman, 1989, p. 161). We will explore the strength of this justification in Part II of the book (Chapters 8, 9, and 12). For now, we will highlight a few basic features of the spacetime maximality definition.

It is often difficult to determine whether a given spacetime is maximal. Consider again the big bang model from above (recall Figure 6.5). This is the spacetime (N, g) where N is the $t > 0$ portion of Minkowski spacetime (M, η) in (t, x) coordinates and $g = t^2\eta$. This spacetime appears maximal since the t^2 term goes to zero as $t \to 0$. It would seem (and it turns out to be true that) one cannot extend through the singularity $t = 0$. But now consider a very similar spacetime (N, h) where the metric h is defined as follows: at each point $(t, x) \in N$ and for any vectors $v = [v_t, v_x]$ and $w = [w_t, w_x]$ at the point, let $h(v, w) = (1/t^4)v_t w_t - v_x w_x$. This spacetime also appears maximal since the $1/t^4$ term blows up as $t \to 0$. Just as before, it would seem that one cannot extend through the singularity $t = 0$. But with a change of coordinates, we can see this turns out to be possible after all.

Consider the the $t > 0$ portion of Minkowski spacetime (N, η) and the diffeomorophism $f : N \to N$ defined by $f(t, x) = (1/t, x)$. We can use f to pull back the metric η on N to the metric $f^*(\eta)$ on N. One can verify (try it!) that $f^*(\eta) = h$ which shows that f is an isometry. This means that Minkowski spacetime (M, η) counts as an extension for (N, h) and thus that (N, h) fails to be maximal. We see that whether a spacetime harbors a "true singularity" or a mere "coordinate singularity" is not always clear. This is true in toy examples like the one just given but also historically in spacetimes with immense physical significance like the Schwarzschild black hole model.

In Part II, we will explore senses in which the maximality may be too strong a condition to impose on spacetime. Here, we highlight an example from Hawking and Ellis (1973) that shows a sense in which it is also too weak. Consider two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates. Remove from N the slits S_1 and S_2 which are defined, respectively, as the set of points $(0, x) \in M$ such that $-2 \leq x \leq -1$ and the set of points $(0, x) \in M$ such that $1 \leq x \leq 2$. Except for the four boundary points $(0, -2)$, $(0, -1)$, $(0, 1)$, and $(0, 2)$, identify the "top edge" of S_1 with the "bottom edge" of S_2 and vice versa (see Figure 6.6). Let (N, g) be the resulting spacetime. Because of what Hawking and Ellis (1973, p. 59) call the "perverse" identifications, one can show that the spacetime is maximal despite that fact that the four boundary points are "missing" from N.

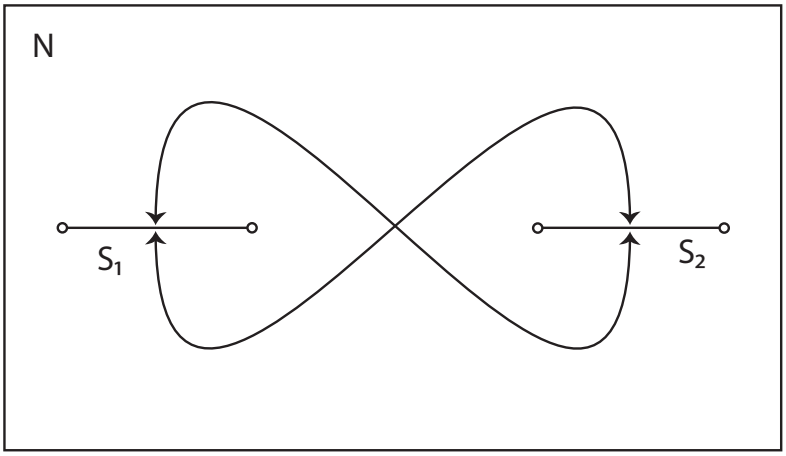

Figure 6.6: The top edge of S_1 is identified with the bottom edge of S_2 and vice versa. The spacetime is maximal despite the four boundary points "missing" from N.

6.5 Hole-Freeness

How does one rule out a spacetime like the perverse example just given? Those in favor of the cosmic censorship conjecture would be quick to point out that the model fails to be globally hyperbolic. But we know that, by itself, the global hyperbolically condition is not strong enough to rule out the "artificial" example of the $t > 0$ portion of Minkowski spacetime. Now we see that spacetime maximality, by itself, is not strong enough to rule out the perverse example just given. Perhaps the conjunction of the global hyperbolicity and maximality conditions will be able to exclude all types physically unreasonable holes. This is an influential route. Indeed, we find that the "statement of cosmic censorship implicitly assumes that the spacetime model is maximal" (Earman 1995, p. 45). But recall how strong the global hyperbolicity condition is (e.g. it rules out the spinning black hole in Kerr spacetime). One wonders if there is a less heavy-handed way to bring together the causal and modal structures of spacetime to deal with the problem of holes. This leads us to the "hole-freeness" property of spacetime.

There a number of definitions of hole-freeness. The first was due to Geroch (1977) whose condition was shown, somewhat surprisingly, to be violated by Minkowski spacetime (Krasnikov, 2009). Revised definitions have been

given by Manchak (2009b) and Minguzzi (2012). Although they are independent notions, hole-freeness and maximality are often assumed together for the same reason that global hyperbolicity and maximality are often assumed together: their conjunction rules out different types of holes that either condition alone fails exclude (Clarke, 1976; Geroch, 1977). In what follows, we will give a simple definition of hole-freeness which presupposes the maximality condition. This will allow us to sidestep some technical difficulties as well as build up a hierarchy of no-hole properties of spacetime. Within the context of maximal spacetimes, the simple definition given below is equivalent to the more general one given by Minguzzi (2012). Indeed, this fact is one of his key results.

Let (M, g) be a spacetime. A set $S \subset M$ fails to be **acausal** if there is a causal curve $\lambda : I \to M$ without vanishing tangent vector and distinct $s_0, s_1 \in I$ such $\lambda(s_0), \lambda(s_1) \in S$. Any acausal set is achronal. But the image of a null geodesic in Minkowski spacetime is achronal but not acausal. It turns out that if an acausal set S is also a slice, then it possesses a particularly nice property: its domain of dependence $D(S)$ must be open (Minguzzi, 2012). We say a maximal spacetime is **hole-free** if, for every acausal slice S and every isometry $f : D(S) \to O$ where $O \subseteq N$ is an open region of a spacetime (N, h) for which $f[S]$ is acausal, we have $f[D(S)] = D(f[S])$. Intuitively, the idea is that a spacetime is not hole-free if there is an acausal slice S whose domain of dependence is not "as large as it can be" in the sense that it can be isometrically embedded into another spacetime (N, h) and extended there. Such a spacetime violates a sense of determinism since it would seem to contain "unpredicted holes developing without reasonable cause" (Ellis and Schmidt, 1977, p. 934). We can see how this works in the case of the perverse example (N, g) from above. Let S be the acausal slice given by $t = -1$. Because of the "missing" boundary points, the domain of dependence $D(S)$ could be larger. There is a natural isometry $f : D(S) \to O$ showing that $D(S)$ is isometric to just a portion $O \subset M$ of Minkwoski spacetime (M, η) while the domain of dependence $f[S]$ is much larger – it is all of M (see Figure 6.7).

One can show that every globally hyperbolic, maximal spacetime must be hole-free (Minguzzi, 2012). In other words, the cosmic censorship conjecture (amended so as to assume maximality) implies a weaker conjecture: all physically reasonable spacetimes are hole-free. But even this weaker conjecture is quite strong and, moreover, its justification is not clear. In what he calls a "dirty open secret," Earman (1995, p. 97-98) remarks on the circular

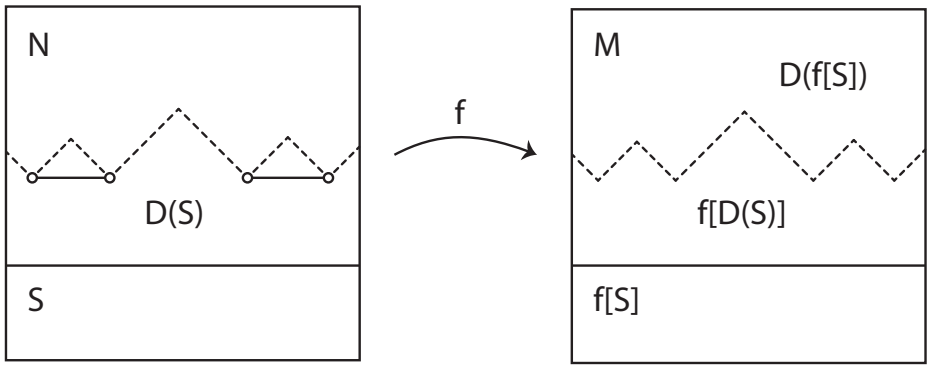

Figure 6.7: The spacetime (N, g) is not hole-free. The domain of dependence $D(S)$ is not "large as it can be" since its image $f[D(S)]$ in Minkowski spacetime (M, g) is a subset of $D(f[S]) = M$.

logic involved in requiring hole-freeness. If the domain of dependence $D(S)$ fails to be "as large as it can be" in some spacetimes, then a type of indeterminism threatens. Determinism can be restored if attention is restricted to hole-free spacetimes. But on what grounds is that not question begging? The perverse example considered above is flat and therefore a vacuum solution to Einstein's equation. After introducing a similar vacuum spacetime with holes, Earman (1995, p. 98) writes:

> "To rule out the above example is to rule out one way Nature might, consistently with all of the known laws of GTR, continue to evolve things [past $D(S)$]. What then is to say that She cannot proceed this way? The most prevalent attitude among general relativists seems to be that fiat is required (see Ellis and Schmidt 1977), otherwise questions about more interesting ways in which determinism can fail are never reached. I implicitly adopted this attitude in the foregoing sections. I am not proud of doing so, but I am no better than my brethren in physics in seeing an alternative to fiat."

Stepping back, we emphasize that there is related "dirty open secret" concerning the maximality condition as well. The $t > 0$ portion of Minkowski

spacetime is a vacuum solution to Einstein's equation. Moreover, it has a well behaved causal structure as it is globally hyperbolic. One can decree that such a spacetime is physically unreasonable. But on what grounds? We will return to this point again in Chapter 12.

6.6 Local Maximality

The hole-freeness condition is not suitable for spacetimes that fail to be causally well-behaved. For example, if one carries out the perverse identifications within Marty's time travel spacetime, the result is not globally hyperbolic (it is not even chronological) but it must be counted as hole-free since there is no acausal slice in such a spacetime. One can rule out such examples without resorting to the heavy-handed cosmic censorship conjecture by invoking a type of "local" maximality condition. As with hole-freeness, there are a number of definitions of this notion – often called "local inextendibility" – that one can find in the literature. An early formulation was given by Hawking and Ellis (1973) that was later shown, somewhat surprisingly, to be violated by Minkowski spacetime (Beem, 1980). In response, variations of another early definition introduced by Clarke (1973) are now often used (Ellis and Schmidt, 1977; Beem et al., 1996). Here, we present a simple formulation of this type where where, in order to both avoid some technical difficulties, we restrict attention to incomplete geodesics rather than the more general class of "b incomplete" curves (Hawking and Ellis, 1973).

Let (M, g) be a spacetime. We say that (M, g) is **locally extendible** if there is an incomplete (and therefore maximal) geodesic $\lambda : I \to M$ contained in an open set $O \subseteq M$ and an isometry $f : O \to U$ where $U \subseteq N$ is an open region of a spacetime (N, h) such that $f \circ \lambda : I \to N$ is not maximal. We say a spacetime is **locally maximal** if it is not locally extendible. Intuitively, the idea is that a spacetime is locally extendible if one can find an incomplete geodesic λ contained in some open set O that is not "as large as it can be" in the sense that it can be isometrically embedded into another spacetime (N, h) and extended there. We can see how this works in the case of the perverse example (N, g) from above (recall Figure 6.6). Let $\lambda : (-\infty, 0) \to N$ be the incomplete timelike geodesic defined by $\gamma(s) = (s, 1)$ which is aimed at one of the four "missing" slit boundary points. Let $O \subset N$ be the $t < 0$ region which contains the image of the curve λ. Finally, let $U \subset M$ be the $t < 0$ region of Minkowski spacetime (M, η). The map $f : O \to U$ defined

by $f(t, x) = (t, x)$ is an isometry and the composed curve $f \circ \lambda : I \to M$ is not maximal since it can be extended through the future endpoint $p = (0, 1)$ (see Figure 6.8).

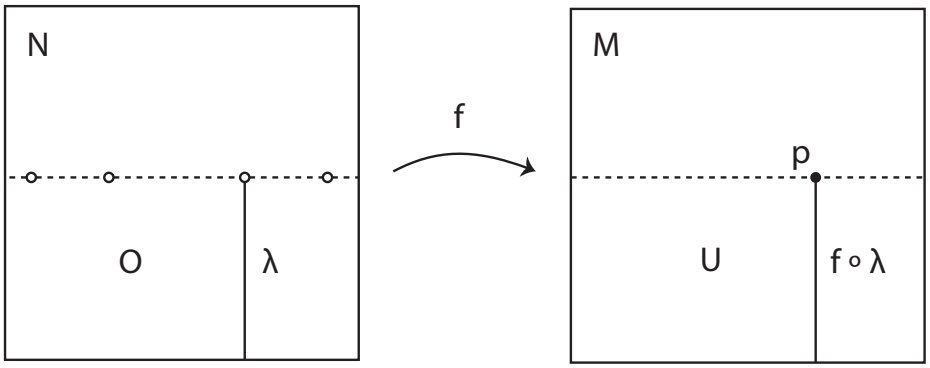

Figure 6.8: The spacetime (N, g) is not locally maximal. The incomplete geodesic λ is not as "large as it can be" since its image $f \circ \lambda$ in Minkowski spacetime (M, g) can be extended through its future endpoint $p = (0, 1)$.

From the way we have set things up, it is immediate that any geodesically complete spacetime is locally maximal. Things do not run in the other direction since the big bang example (recall Figure 6.5) is a locally maximal spacetime but geodesically incomplete. One can show that a spacetime that fails to be hole-free must be locally extendible (Minguzzi, 2012). But the conditions are not equivalent since a perverse version of Marty's time travel spacetime is locally extendible but hole-free as mentioned above.

Here is another example of independent interest (Misner, 1967). Let M be the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. Here, we allow the coordinate $\theta \in S$ to take on all values of $\mathbb{R}$ but we identify each θ with $\theta + 2\pi n$ for all integers n. **Misner spacetime** is the pair (M, g) where the metric g is defined as follows: at each point $(t, \theta) \in M$ and for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $g(v, w) = v_t w_\theta + v_\theta w_t + t v_\theta w_\theta$. One can get a grip on this spacetime by considering the behavior of light traveling along null geodesics. Since there is no $v_t w_t$ term, it is immediate that the vector $[1, 0]$ is null at every point. So one family of null geodesics run along the cylinder and are complete. We also see that the vector $[0, 1]$ is spacelike

for $t < 0$, null for $t = 0$, and timelike for $t > 0$. This means that the light cones open up as t increases. In the $t < 0$ and $t > 0$ regions, another family of null geodesics spiral around the cylinder approaching but never reaching $t = 0$. These geodesics are incomplete. One example in the $t < 0$ region is $\lambda_1 : (-\infty, 0) \to M$ defined by $\lambda_1(s) = (s, -2\ln(-s))$. We also see that there is a closed null geodesic at $t = 0$ and, for any $k > 0$, there is a CTC at $t = k$ (see Figure 6.9).

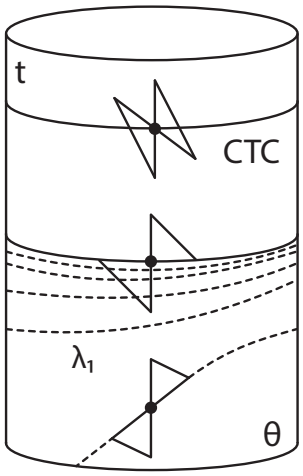

Figure 6.9: In Misner spacetime, an incomplete null geodesic λ_1 spirals around the cylinder approaching but never reaching $t = 0$. There is a closed null geodesic at $t = 0$ and, for any $k > 0$, there is a CTC at $t = k$.

One can show that Misner spacetime is hole-free. To see that it is locally extendible, consider a "reverse twisted" variant (M, h) of Misner spacetime where h is defined as follows: at each point $(t, \theta) \in M$ and for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $h(v, w) = -v_t w_\theta - v_\theta w_t + t v_\theta w_\theta$. One can verify that the diffeomorphism on M which takes the point (t, θ) to the point $(t, -\theta)$ is a reflection isometry from Misner spacetime (M, g) to the reverse twisted variant (M, h). Now consider again the incomplete null geodesic $\lambda_1 : (-\infty, 0) \to M$ in Misner spacetime (M, g) defined above by $\lambda_1(s) = (s, -2\ln(-s))$ which spirals around the cylinder, approaching but never reaching $t = 0$. It is contained in the open set $O \subset M$ given by the $t < 0$ region. One can show that there is an isometry $f : O \to O$ from (O, g) to (O, h) defined by $f(t, \theta) = (t, \theta + 2\ln(-t))$ that "untwists" this geodesic. In the reverse Misner spacetime (M, h), we find that the null

geodesic $f \circ \lambda_1 : (-\infty, 0) \to O$ reduces simply to $f \circ \lambda_1(s) = (s, 0)$ which is not maximal in (M, h) since it can be extended through its future endpoint $p = (0, 0)$ (see Figure 6.10). So Misner spacetime is locally extendible.

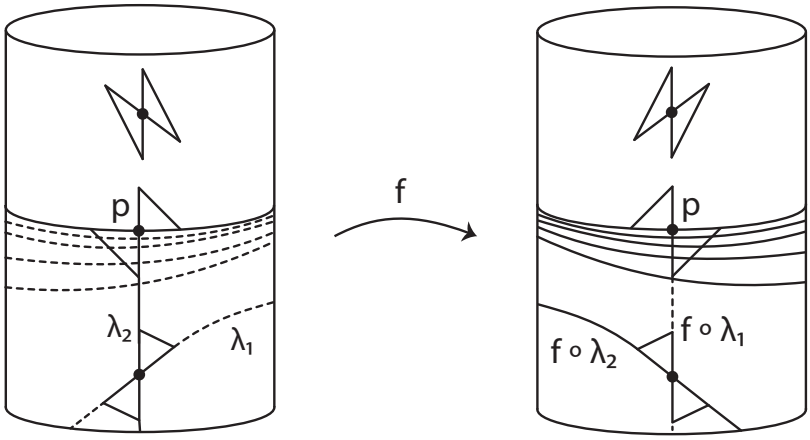

Figure 6.10: The isometry f untwists the incomplete null geodesic λ_1 so that $f \circ \lambda_1$ can be extended through p. The null geodesic λ_2 with future endpoint p is twisted up by f so that it becomes the incomplete null geodesic $f \circ \lambda_2$.

We see that the twisted null geodesic λ_1 in the spacetime (O, g) cannot be extended in (M, g) but can be untwisted and extended in (M, h). But the move to (M, h) has the effect of twisting up other null geodesics that were untwisted in (M, g). For example, consider the null geodesic $\lambda_2 : (-\infty, 0) \to O$ in (M, g) defined by $\gamma_2(s) = (s, 0)$ with future endpoint $p = (0, 0)$. Under the isometry f, this curve becomes the twisted null geodesic $f \circ \gamma_2 : (-\infty, 0) \to O$ in (M, h) defined by $\lambda_2(s) = (s, 2\ln(-s))$ (see Figure 6.10). One cannot untwist both λ_1 and λ_2 in a single extension. At least, one cannot do so while the standard Hausdorff condition is in place. We will explore the possibility of non-standard "branching" spacetimes in Chapter 13.

6.7 Conclusion

In this chapter, we have had our first look at a variety of topics that will become the focus later on when spacetime maximality becomes the focus. We considered four conditions to rule out "holes" in spacetime that form a type

of no-hole hierarchy. We started with a look at the highest level: geodesic completeness. This condition is much too strong given that it rules out physically reasonable models of the universe including various black hole and big bang spacetimes. Indeed, the "singularities" present in these examples seem to be a generic feature of spacetime (Hawking and Penrose, 1970). In light of the situation, we then turned to the three lower levels – all modal conditions that require spacetime to be "as large as it can be" in different senses.

At the lowest level, we have spacetime maximality itself. The perverse example of Hawking and Ellis (1973) showed that spacetime maximality is too weak in the sense that it fails to rule out all spacetimes with holes. (Later on, we will explore senses in which the condition is also too strong.) We then moved up one level in the no-hole hierarchy to consider hole-freeness. This condition requires a sense in which the domain of dependence must be maximal. But the "dirty open secret" discussed by Earman (1995) draws attention to the questionable justification of the hole-freeness condition. Like spacetime maximality itself, the condition seems to be adopted by fiat in order to avoid problems with determinism. We then moved up another level in the no-hole hierarchy to consider the final condition – local maximality – which forbids incomplete geodesics from being extendible in some other spacetime. Since local maximality is stronger than hole-freeness, the justification problems concerning the latter apply to the former as well.

Just as with the various energy and causal conditions, it will be useful later on to think of the no-hole hierarchy in terms of subcollections of the collection $\mathscr{U}$ of all spacetimes. Let $(Max), (HF), (LM), (GC) \subset \mathscr{U}$ be the collections of spacetimes satisfying, respectively, the conditions of maximality, hole-freeness, local maximality, and geodesic completeness. It is easy to see that each of these properties count as global and, except geodesic completeness, they are modal properties whose definition depends on the collection of spacetimes as standardly defined. The hierarchy of these no-hole properties can be summarized as follows (see Figure 6.11).

$$(GC) \subset (LM) \subset (HF) \subset (Max)$$

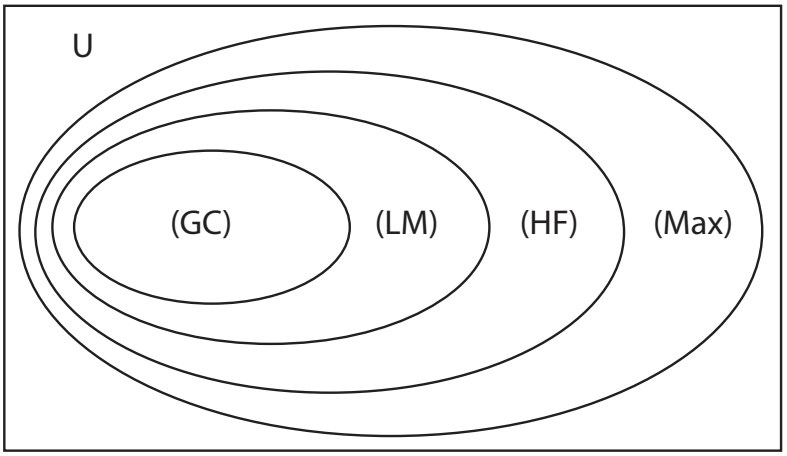

Figure 6.11: A hierarchy of no-hole properties.

Chapter 7

Asymmetries

7.1 Introduction

Recall that the (global) symmetries of a mathematical structure are the isomorphisms from the structure to itself. Just about any example spacetime that one is likely to run across in the literature (and all example spacetimes we have considered so far) has a few different isometries from it to itself. These symmetries allow for nice calculations. But there are reasons to think that spacetimes with non-trivial symmetries are rare among the collection of all spacetimes. In this chapter, we turn our attention to a hierarchy of (global and local) spacetime asymmetry conditions. In Part II, we will see ways in which these conditions can help us better understand aspects of spacetime maximality.

We begin with a discussion concerning the impossibility of global asymmetries on any manifold due to the existence of "hole diffeomorphisms" of a certain kind (Earman and Norton, 1987). In contrast, the additional structure of a metric ensures that a minimal type of spacetime asymmetry called "rigidity" always obtains: no hole diffeomorphism is a spacetime symmetry (Halvorson and Manchak, 2022). We also consider two natural ways to strengthen the rigidity condition. The next two levels of the asymmetry hierarchy concern the "giraffe" condition which requires that the identity map is the only global spacetime symmetry (Barrett et al., 2023). Finally, we consider the "Heraclitus" condition which sits atop the asymmetry hierarchy. This condition is satisfied when any pair of distinct events fail to have even local neighborhoods that are isometric (Manchak and Barrett, 2023).

7.2 Symmetry Holes

As a warm up, and to better appreciate the significance of spacetime asymmetries, let's consider a few different levels of mathematical structures and what the symmetries of each level are like. We start with topological spaces and their associated isomorphisms: homeomorphisms. Let X be any set with either the trivial or discrete topologies. Then any bijection $f : X \to X$ counts as a homeomorphism. One can verify that there will be $n! = (n)(n-1)(n-2)...1$ symmetries in the case of a finite set X with n elements and uncountably many symmetries otherwise. But topologies on X in between these two extremes can allow for highly asymmetric structures.

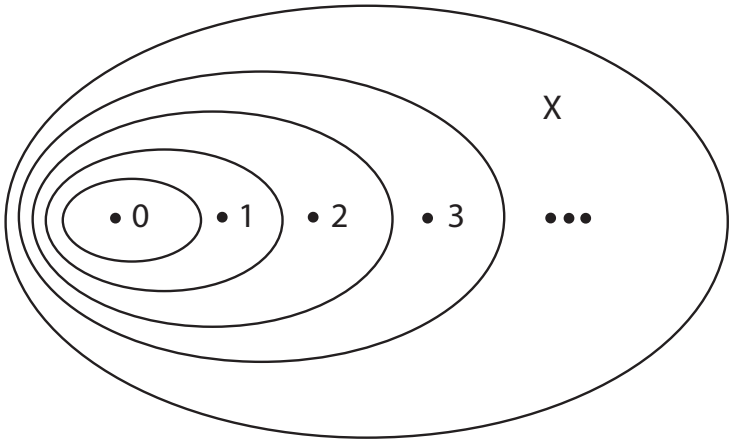

Figure 7.1: The topological space (X, τ) is such that each element $n \in X$ is contained in all but $n + 1$ open sets in τ.

The simplest example is the Sierpiński space (X, σ) where $X = \{0, 1\}$ and $\sigma = \{\varnothing, \{0\}, \{0, 1\}\}$. The identity map is the only homeomorphism from this topological space to itself. The other bijection that exchanges the two elements doesn't respect the topological structure that distinguishes 0 (whose neighborhoods include both $\{0\}$ and X) and 1 (whose only neighborhood is X). One might think this effect obtains because the Sierpiński space has only a small number of elements. But an infinite version has the same properties. Let $X = \{0, 1, 2, ...\}$ and let τ be the set containing $\varnothing$, X, and the subsets $\{0\}$, $\{0, 1\}$, $\{0, 1, 2\}$, and so on. One can verify that (X, τ) is a topological space and that each element $n \in X$ is contained in all but $n + 1$ open sets

in τ (see Figure 7.1). For example, 0 is contained in all open sets but $\varnothing$; 1 is contained in all open sets but $\varnothing$ and $\{0\}$; etc. In this way, one can topologically distinguish any element from any other. So any bijection from X to itself that isn't the identity map will not respect this structure and thus fail to be a homeomorphism.

Now let's now consider manifolds and their associated isomorphisms: diffeomorphisms. Let M be any connected Hausdorff manifold and for any let $p_1, ..., p_n$ and $q_1, ..., q_n$ be any points in M that are all distinct. One can show that M is extremely "non-rigid" in the sense that there is a diffeomorophism $f : M \to M$ such that $f(p_1) = q_1$, ..., $f(p_n) = q_n$ (Geroch, 1969, p. 189). It follows that every manifold has uncountably many symmetries. So in contrast to the situation with topological spaces, we find that asymmetry among manifolds is impossible.

Within this context, it might be useful to consider an influential construction used in discussions of the "hole argument" (Earman and Norton, 1987). Let M be any connected Hausdorff manifold and let $H \subset M$ be such that both it and $M - H$ both contain non-empty open sets. A **hole diffeomorphism** $f : M \to M$ is a diffeomorphism which is (i) not the identity map in the "hole" region H but (ii) acts as the identity map on the restricted domain $M - H$. It turns out that M admits uncountably many hole diffeomorphisms: even fixing the symmetries in the region outside the "hole" will not fix them inside.

Consider a simple example of a hole diffeomorphism on the manifold $M = \mathbb{R}^2$ in (t, x) coordinates where H is the $t > 0$ region of M. Let $f : M \to M$ be defined piecewise: $f(t, x) = (t + \exp(-1/t), x)$ for all points $(t, x) \in H$ and $f(t, x) = (t, x)$ for all points $(t, x) \in M - H$. One can verify that the function f is a diffeomorphism. It maps each point in $M - H$ to itself. So the origin $p = (0, 0)$ gets mapped to $f(p) = p$. But it "stretches" the region H in the positive t direction in such a way that it maps a point $q \in H$ to a point $f(q) \in H$ with a slightly larger t value. For example, it maps the point $q = (1, 0)$ into the point $f(q) = (1 + \exp(-1), 0) \approx (1.37, 0)$ (see Figure 7.2).

7.3 Rigid Spacetime

Let's add some more structure and consider spacetimes and their associated isomorphisms: isometries. We have just seen that there is a hole diffeomorphism $f : M \to M$ with respect to some hole region $H \subset M$. We can use

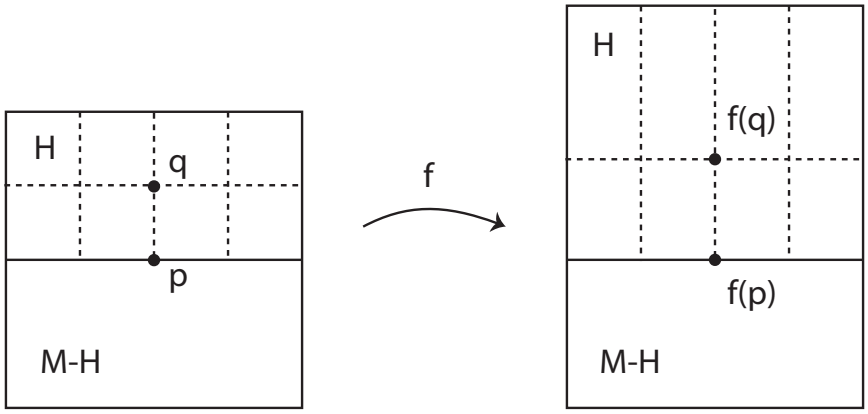

Figure 7.2: The hole diffeomorphism f maps any point in $M - H$ to itself (e.g. the origin p). But it "stretches" the region H. For example, it maps the point $q = (1, 0)$ into the point $f(q) = (1 + \exp(-1), 0) \approx (1.37, 0)$.

this diffeomorphism f to push forward the metric g on M to produce the spacetime $(M, f_*(g))$. The diffeomorphism f certainly counts as an isometry from (M, g) to $(M, f_*(g))$. But we note the following facts which are often not appreciated: (i) the identity map on M fails to be an isometry from (M, g) to $(M, f_*(g))$ (Weatherall, 2018) and (ii) the diffeomorphism f is not an isometry from (M, g) to itself. Fact (ii) shows that f is not a symmetry of the spacetime (M, g) (Halvorson and Manchak, 2022). Let's explore this idea a bit more.

Let's say that a spacetime (M, g) is **rigid** if, for any isometry $f : M \to M$ and any non-empty open set $O \subseteq M$, if f acts as the identity on O, then it is the identity map. A general result due to Geroch (1969) is the following: Let (M, g) and (N, h) be spacetimes, let $p \in M$ and $q \in N$ be points, and let $\{v_n\}$ and $\{w_n\}$ be orthonormal bases of vectors at p and q respectively. Then there is at most one isometry $f : M \to N$ such that $f(p) = q$ and $\{f_*(v_n)\} = \{w_n\}$. From this it follows that any (standard) spacetime must be rigid and thus a hole diffeomorphism is never a spacetime symmetry. (We will see in Chapter 13 that non-standard spacetimes can fail to be rigid if the Hausdorff condition is dropped.) We find that fixing the symmetries of spacetime on an open region, however small, fixes them everywhere. An example may help to illustrate the point.

Consider two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates. Let $f : M \to M$ be the hole diffeomorphism defined above. So H is the $t > 0$ region of M and f is defined piecewise: $f(t, x) = (t + \exp(-1/t), x)$ for all points $(t, x) \in H$ and $f(t, x) = (t, x)$ for all points $(t, x) \in M - H$. We can use f to pull back the metric η on M to produce the spacetime $(M, f^*(\eta))$. What is $f^*(\eta)$? Of course, for any point $p \in M - H$, we have $f^*(\eta) = \eta$. Consider the region H. Let $v = [v_t, v_x]$ and $w = [w_t, w_x]$ be any vectors at any point $q \in H$. One can use f to push these vectors forward. We find that $f_*(v) = [\alpha(t)v_t, v_x]$ and $f_*(w) = [\alpha(t)w_t, w_x]$ where $\alpha(t) = 1 + t^{-2}\exp(-1/t)$. So $\eta(f_*(v), f_*(w)) = \alpha(t)^2 v_t w_t - v_x w_x$. This is the number that the pull back metric $f^*(\eta)$ assigns to v and w. So $\eta \neq f^*(\eta)$ for every point $q \in H$. For example, at the point $q = (1, 0)$, we have $f^*(\eta)(v, w) = \alpha(t)^2 v_t w_t - v_x w_x = (1 + \exp(-1))v_t w_t - v_x w_x$ which is not $\eta(v, w) = v_t w_t - v_x w_x$.

This makes sense. The timelike geodesic $\lambda : [0, 1] \to M$ defined by $\lambda(s) = (s, 0)$ has a tangent vector $\lambda'(s) = [1, 0]$. According to the metric $f^*(\eta)$, the (squared) length $\|\lambda'(s)\|$ of this tangent vector is $\alpha^2(t) = \alpha^2(s)$ at each point $\lambda(s) = (s, 0)$. So integrating $\sqrt{\|\lambda'(s)\|} = \alpha(s)$ from $s = 0$ to $s = 1$ gives an elapsed time of $\|\lambda\| = 1 + \exp(-1) \approx 1.37$ along the curve from the origin p to the point $q = (1, 0)$. On the other hand, the metric η judges this timelike geodesic λ to have a shorter elapsed time of just 1. The fact that the metrics disagree about the elapsed time of λ shows that f is not an isometry from (M, η) to itself. But one can check that pushing forward λ to the curve $f \circ \lambda : [0, 1] \to M$ defined by $f \circ \lambda(s) = (s + \exp(-1/s), 0)$ results in an elapsed time of $\|f \circ \lambda\| = 1 + \exp(-1) \approx 1.37$ as determined by the metric η. So the elapsed time between p and q according to $f^*(\eta)$ is the same as the elapsed time between $f(p) = p$ and $f(q)$ according to η (see Figure 7.3).

Stepping back, we see that spacetime structure allows for asymmetry in a way that manifold structure doesn't: all spacetimes are rigid while all manifolds are not. We have seen that fixing spacetime symmetries in any open region, however small, fixes them everywhere. A natural strengthening of the condition requires that fixing spacetime symmetries at any point fixes them everywhere. We say a spacetime (M, g) is **point rigid** if, for any point $p \in M$, any isometry $f : M \to M$ such that $f(p) = p$ must be the identity map. We know that not all spacetimes are point rigid. Consider two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates. The reflection $f : M \to M$ defined by $f(t, x) = (t, -x)$ is an isometry such that $f(p) = p$

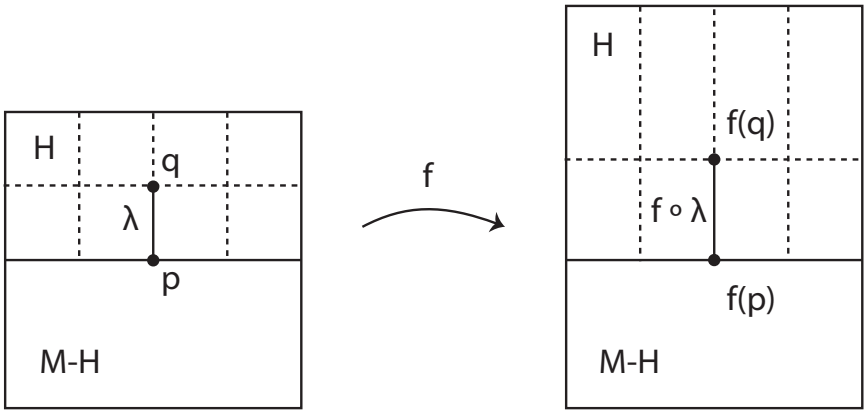

Figure 7.3: The elapsed time along λ from p to q according to the pulled back metric $f^*(\eta)$ is the same as the elapsed time along $f \circ \lambda$ from $f(p)$ to $f(q)$ according to the metric η. In each case, it is $1 + \exp(-1) \approx 1.37$.

for the origin $p = (0, 0)$. Since f is not the identity map, (M, η) is not point rigid. Another natural way to strengthen the rigidity condition is to require that, at least at some points, the spacetime symmetries are completely fixed. We say a spacetime (M, g) has a **fixed point** if, for some point $p \in M$, any isometry $f : M \to M$ is such that $f(p) = p$. As with the point rigid condition, Minkowski spacetime (M, η) shows that not all spacetimes have a fixed point. The time translation $f : M \to M$ defined by $f(t, x) = (t + 1, x)$ is an isometry such that $f(p) \neq p$ for any $p \in M$.

The point rigid and fixed point conditions are independent. Let (M, η) be two-dimensional Minkowski spacetime in (t, x) coordinates. Consider the spacetime (N, η) where $N \subset M$ be the set of points (t, x) such and $t > 0$ and $x^2 < t^2$ (see Figure 7.4). Both the identity map and the reflection $f : N \to N$ defined by $f(t, x) = (t, -x)$ map the point $p = (1, 0)$ into itself. So the spacetime is not point rigid. But one can verify that f and the identity map are the only symmetries of the spacetime. Since p is mapped into itself in both isometries, it is a fixed point for the (N, η).

Now consider an example that is point rigid but has no fixed point. Let (M, η) be two-dimensional Minkowski spacetime as before. For each integer n, excise from M the compact region enclosed by the points $(0, n)$, $(1/2, n)$, and $(0, n + 1/2)$. Let (N, η) be the resulting spacetime (see Figure 7.5. We

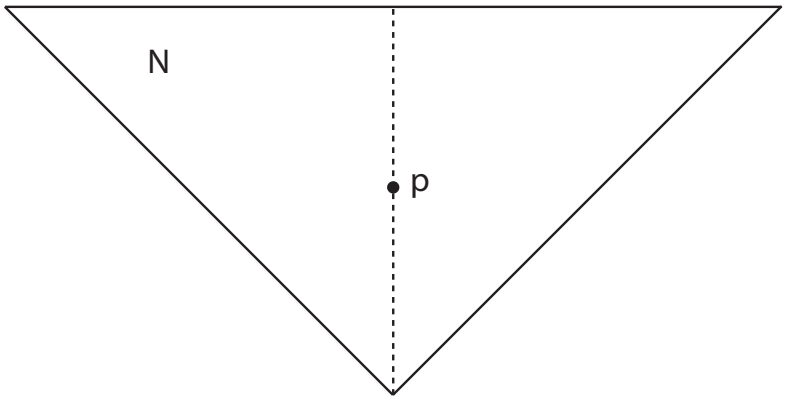

Figure 7.4: The identity map and the isometry defined by $f(t, x) = (t, -x)$ are the only symmetries of the spacetime. In each case, the point $p = (1, 0)$ is mapped to itself. So the spacetime has a fixed point but is not point rigid.

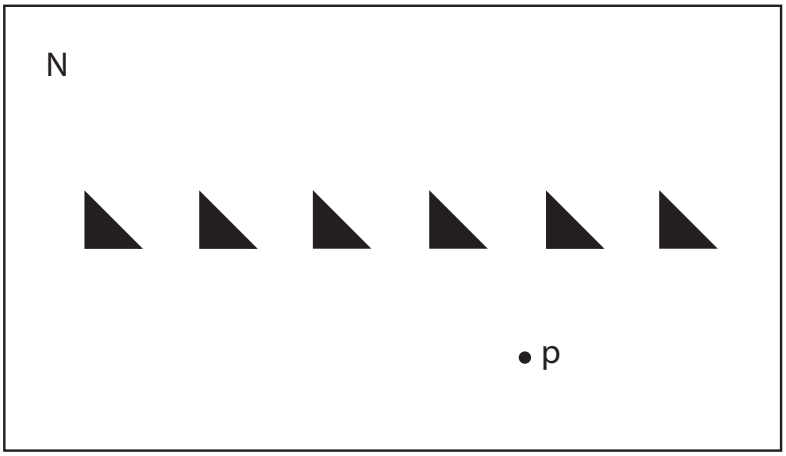

Figure 7.5: For each integer n, there is an isometry defined by $f_n(t, x) = (t, x + n)$. But only the identity map ($n = 0$) takes any point $p \in N$ into itself. So the spacetime is point rigid but has no fixed point.

see that for each integer n, we have an isometry $f_n : N \to N$ defined by $f_n(t, x) = (t, x + n)$. So the spacetime fails to have a fixed point. But it is point rigid. This follows since the f_n isometries just defined are the only symmetries of the spacetime. If $f_n(p) = p$ for any $p \in N$, it must be the case that $n = 0$, i.e. the isometry f_n is the identity map.

7.4 Giraffe Spacetime

All of the spacetimes considered so far have non-trivial symmetries. Let us now consider simple example due to David Malament that does not (Barrett et al., 2023). Let (M, η) be two-dimensional Minkowski spacetime and let $C \subset M$ be a compact set shaped like a (sufficiently asymmetric) giraffe and let $N = M - C$. One can verify that the only symmetry of the spacetime (N, η) is the identity map (see Figure 7.6). The "missing" giraffe region blocks all of the symmetries of Minkowski spacetime. Let us say that a spacetime (M, g) is **giraffe** if the only isometry $f : M \to M$ is the identity map. It is immediate that a giraffe spacetime must have a fixed point and be point rigid. When considered on their own, both the point rigid and the fixed point conditions are strictly weaker than the giraffe condition. But a simple result shows that a spacetime satisfies both of these conditions if and only if it is giraffe (Manchak and Barrett, 2023).

It has been claimed that "everyone knows" giraffe spacetimes are "generic" in some sense (D'Ambra and Gromov, 1991, p. 21). But a general statement is difficult to formulate precisely and a proof remains elusive. Among compact manifolds with Riemannian metric, it has long been known that the giraffe condition is generically satisfied in a natural sense (Ebin, 1968). More recently it has been shown that this is true for compact spacetimes as well (Mounoud, 2015).

Giraffe spacetime have no symmetries in a global sense but they can still be highly symmetric in a "local" sense. Let us say that a spacetime (M, g) is **locally giraffe** if, given any open, connected set $O \subseteq M$, the spacetime (O, g) is giraffe. It is immediate that a locally giraffe spacetime must be giraffe. The example just considered of a compact giraffe region removed from Minkowski spacetime is giraffe but not locally giraffe. Here is a less interesting but more tractable example. Let (M, η) be two-dimensional Minkowski spacetime in (t, x) coordinates. Now consider the spacetime (N, η) where $N \subset M$ be the set of points (t, x) such and $t > 0$, $x > 0$, and $x^2 < t^2$

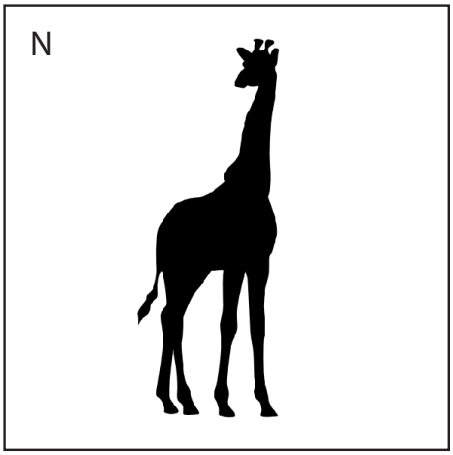

Figure 7.6: Due to the "missing" giraffe region, the only symmetry of the spacetime is the identity map.

(see Figure 7.7). One can verify that the spacetime is giraffe. But consider the (open, connected) ball $O \subset N$ centered at the point $p = (2, 1)$ with radius $1/2$. Let $f : O \to O$ be the isometry defined by $f(t, x) = (t, 2 - x)$ which reflects O about the $x = 1$ line. Since this isometry is not the identity map on O, we see that the spacetime fails to be locally giraffe.

7.5 Heraclitus Spacetime

A locally giraffe spacetime can still have local symmetries of a certain kind. Let us say that a spacetime (M, g) is **Heraclitus** if, for any distinct points $p, q \in M$ and any neighborhoods O_p and O_q of these points, there is no isometry $f : O_p \to O_q$ such that $f(p) = q$ (Manchak and Barrett, 2023). In a Heraclitus spacetime, each event is unlike any other. In this sense, one might say that it is impossible step twice into the same river. One can show that a spacetime (M, g) is Heraclitus if and only if, for any open sets $U, V \subset M$ and any isometry $f : U \to V$, we have $U = V$ and f is the identity map. From this result, it follows easily that any Heraclitus spacetime must be locally giraffe. The other direction does not hold as will be shown a bit later on. First, let's explore an example of a Heraclitus spacetime.

Let (M, η) be two-dimensional Minkowski spacetime in (t, x) coordinates.

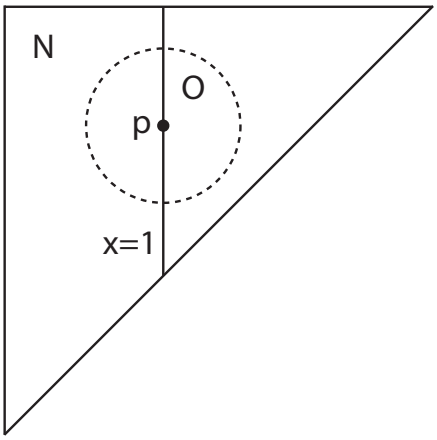

Figure 7.7: The spacetime is giraffe but not locally giraffe. The ball $O \subset N$ centered at $p = (2, 1)$ has an isometry defined by $f(t, x) = (t, 2 - x)$ which reflects O about the $x = 1$ line.

Now consider the giraffe but not locally giraffe spacetime (N, η) that we constructed above where $N \subset M$ is the set of points (t, x) such and $t > 0$, $x > 0$, and $x^2 < t^2$ (recall Figure 7.7). Let $\Omega : N \to \mathbb{R}$ be a conformal factor defined by $\Omega(t, x) = 1/(t^2 + x^2)$. One can show that the spacetime (N, g) is Heraclitus where $g = \Omega^2\eta$. This follows from the peculiar nature of the Ricci scalar curvature $R : N \to \mathbb{R}$ given by $R(t, x) = 8(x^2 - t^2)$. Using the derivative operator ∇ associated with the metric g, one can compute a type of "magnitude of the derivative" of R at each point in N. Using the derivative operator ∇ associated with (N, g), one can differentiate the scalar field R to define a vector v_p at every point in $p \in N$. Let $Q : N \to \mathbb{R}$ be the smooth function defined by $Q(p) = g(v_p, v_p)$. One can show that this scalar curvature function is given by $Q = -32R/\Omega^2$. Consider the points $p, q \in M$ and any neighborhoods O_p and O_q of these points respectively. Suppose there is an isometry $f : O_p \to O_q$ such that $f(p) = q$. Since R and Q are both scalar curvature functions, and since such functions must be preserved under any isometry, we have $R(p) = R(q)$ and $Q(p) = Q(q)$. Since $Q = -32R/\Omega^2$ and $R < 0 < \Omega$ on N, it follows that $\Omega(p) = \Omega(q)$. But because of the way N is truncated, we find that $R(p) = R(q)$ and $\Omega(p) = \Omega(q)$ can only obtain if $p = q$ (see Figure 7.8). So (N, g) is Heraclitus.

Among compact manifolds with Riemannian metric, one can show that

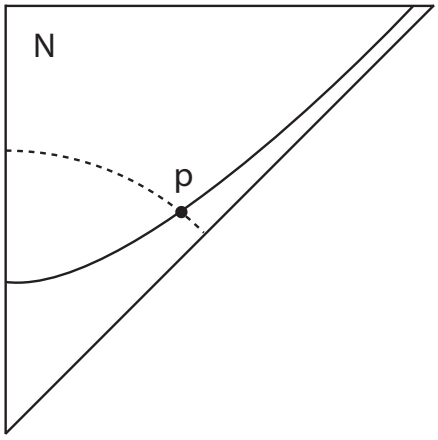

Figure 7.8: The isometry f must map the point p to some point q on the solid line so that $R(p) = R(q)$. But q must also be somewhere on the dotted line so that $\Omega(p) = \Omega(q)$. So $p = q$ and the spacetime is Heraclitus.

the Heraclitus condition is generically satisfied in a natural sense (Sunada, 1985). One wonders if an analogous result holds for compact Heraclitus spacetimes or Heraclitus spacetimes more generally. It turns out that the Heraclitus asymmetry property is sufficiently strong to show a sense in which global spacetime structure is completely determined by local spacetime structure: any pair of locally isometric Heraclitus spacetimes must be isometric. From this it follows easily that there can never be more than one Heraclitus spacetime with the same local properties. One can also show that these local to global uniqueness results fail if the Heraclitus condition is weakened to the local giraffe condition. Consider the following example.

Let (N_i, g_i) for $i = 1, 2, 3$ be three copies of the Heraclitus spacetime just constructed. In copies (N_1, g_1) and (N_2, g_2) cut a slit S^- at $t = 3$ and $1 \leq x \leq 2$. Except for the four boundary points, identify the top edge of this slit in (N_1, g_1) with the bottom edge of the slit in (N_2, g_2) (but not vice versa). In copies (N_2, g_2) and (N_3, g_3) cut a slit S^+ at $t = 4$ and $1 \leq x \leq 2$. Except for the four boundary points, identify the top edge of this slit in (N_2, g_2) with the bottom edge of the slit in (N_3, g_3) (but not vice versa). The resulting spacetime (see Figure 7.9) is both (i) locally giraffe but not Heraclitus and (ii) locally isometric but not isometric to the Heraclitus spacetime (N, g).

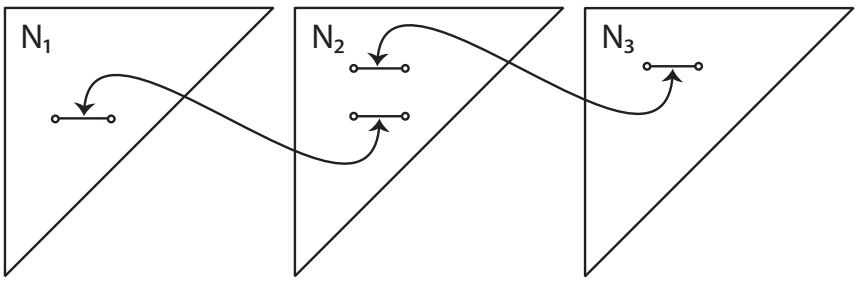

Figure 7.9: Three copies of the Heraclitus spacetime (N, g) are identified as indicated. The resulting spacetime is (i) locally giraffe but not Heraclitus and (ii) locally isometric but not isometric to (N, g).

7.6 Conclusion

In this chapter, we have looked at various conditions forming an asymmetry hierarchy. The lowest level – rigidity – is satisfied by all (standard) spacetimes. The condition ensures that a hole diffeomorphism fails to be a spacetime symmetry. This means that fixing the symmetries of spacetime in an open region – no matter how small – fixes them everywhere. A pair of independent conditions form the next level of the asymmetry hierarchy. The point rigid condition is satisfied when it is that case that that fixing the symmetries of spacetime at a single point fixes them everywhere.The fixed point condition requires the existence of a single point which must be mapped to itself under any symmetry. The conjunction of these two conditions is equivalent to the giraffe condition which is satisfied whenever the identity map is the only spacetime symmetry. Above this level is the local giraffe condition which requires that any open, connected region spacetime is giraffe when considered as a spacetime in its own right. At the highest level of the asymmetry hierarchy is the Heraclitus condition which requires that no distinct points have isometric neighborhoods.

Previous work on compact Riemannian show senses in which versions of all of the asymmetry conditions are generically satisfied (Ebin, 1968; Sunada,

1985). The results also carry over to the context of compact spacetimes and it has been claimed that "everyone knows" that the similar results hold for non-compact spacetimes as well (D'Ambra and Gromov, 1991; Mounoud, 2015). Thus, it would seem that "almost all" physically reasonable spacetimes satisfy even the strongest asymmetry conditions. The situation stands in stark contrast with the highest levels of the causal hierarchy: it is not at all clear, for example, that almost all physically reasonable spacetimes are globally hyperbolic.

Just as with the various energy, causal, and no-hole conditions, it will be useful later on to think of the asymmetry hierarchy in terms of subcollections of the collection $\mathscr{U}$ of all spacetimes. Let $(Rig), (PR), (FP), (Gir), (LG), (Her) \subset \mathscr{U}$ be the collections of spacetimes satisfying, respectively, the point rigid, fixed point, giraffe, locally giraffe, and Heraclitus conditions. It is easy to see that each of these properties count as global. The hierarchy of these asymmetry properties can be summarized as follows (see Figure 7.10).

$$(Her) \subset (LG) \subset (Gir) = (PR) \cap (FP) \subset (PR), (FP) \subset (Rig)$$

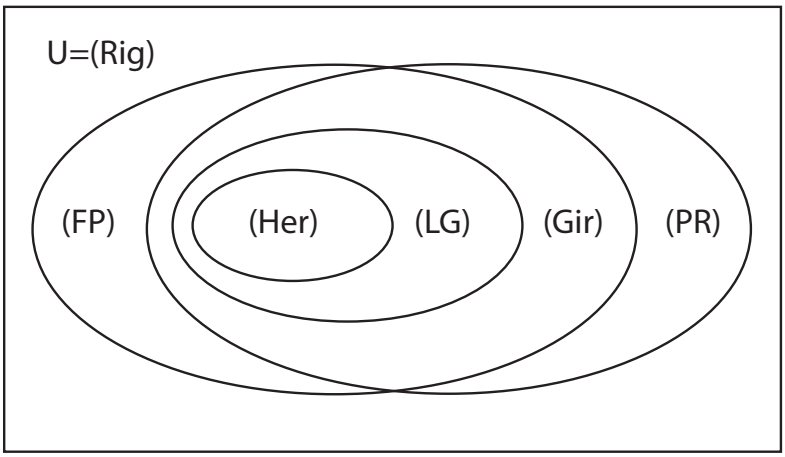

Figure 7.10: A hierarchy of asymmetry properties.

Part II

Maximal Universes

Chapter 8

Meaning

8.1 Introduction

The meaning of spacetime "maximality" depends crucially on a background possibility space in the form of a collection of spacetime models. The standard definition uses the collection $\mathscr{U}$. In Part I, we identified a number of different subcollections $\mathscr{P} \subseteq \mathscr{U}$ of physical significance. Here in Part II, we explore the notion spacetime maximality relative to these subcollections.

In this chapter, we begin by calling into question the significance of the standard definition of spacetime maximality and we introduce the relativized definitions to get a better grip on the situation. We then move to consider some remarks of Geroch (1970b) who conjectures that, at least for some subcollections $\mathscr{P} \subseteq \mathscr{U}$ of physical interest, a type of equivalence holds between the standard and relativized definitions: a spacetime in $\mathscr{P}$ is maximal relative to $\mathscr{P}$ if and only if it is maximal relative to the standard collection $\mathscr{U}$. In the next few sections, we investigate whether this conjecture is true with respect to the various local, causal, and asymmetry properties identified in Part I. Although some important cases remain open, we will show that the conjecture is false for almost all spacetime properties under consideration. This means that in order to understand the notion of spacetime "maximality" in a general, nuanced way, it is not sufficient to study the standard definition. One needs to carefully consider a variety of the relativized definitions as well. This is the task of the remainder of Part II.

8.2 Definitions

As we have seen, the idea that the universe must be "as large as it can be" is something of a dogma within the context of generality relativity. All over the literature, one finds variations of the same decree: "any reasonable space-time should be inextendible" (Clarke, 1993, p. 8). At root, the reasoning behind such a position comes in two parts (Earman, 1989, p. 161):

> "A justification for ignoring space-times that are not [maximal] can be given in two steps. First, it can be shown that any space-time can be extended to a space-time that is maximal...Second, one can argue on PSR [principle of sufficient reason] grounds that there is no good reason for the Creative Force to stop building until the maximal extent is reached, and on grounds of plenitude that the maximal model is better than a truncated submodel."

This justification for spacetime maximality is rarely questioned. When it is, the focus has been almost exclusively on the second step. Let's take a look at a pair of examples. Regarding the grounds of plenitude, some incredulity is expressed by John Norton (2011, p. 173):

> "The principle of plenitude itself is sufficiently implausible that we need to prop it up with anthropomorphic metaphors. We are to imagine a personified Nature in the act of creating spacetime, much as I might be painting my fence on the weekend. Just as I might not want to stop when there is one board remaining unpainted, so Nature is supposedly loath to halt with a cubic mile-year of spacetime still uncreated."

Regarding the principle of sufficient reason, we find that such grounds can be turned on their head to argue against spacetime maximality. Here we have Chris Clarke (1993, p. 9):

> "It can be easily shown that any space-time can in fact be extended until no further extension is possible. At this point the space-time is called maximal, and so we are led to the idea that we need only consider maximal space-times. But this idea is not really as innocuous as it might seem, because of the problem that

> an extension of a space-time, when it exists, cannot usually be determined uniquely...In cases such as these the same principle of sufficient reason would not allow one extension to exist at the expense of another. Perhaps the space-time, like Buridan's ass between two bales of hay, unable to decide which way to go, brings the whole of history to a halt."

Stepping back, it is not too surprising that whenever the maximality condition is questioned (however rarely), the focus has been on the second step of the justification that concerns the Leibnizian principles of plenitude and sufficient reason. After all, step one amounts to a mathematical result: every extendible spacetime has a maximal extension (Geroch, 1970b). How could such a result possibly be questioned? Here, we draw attention to the fact that the physical significance of this mathematical statement depends crucially on a suitable formulation of a "maximal" spacetime. This modal property is defined relative the background collection $\mathscr{U}$ of all (standard) spacetimes. As we have seen, practitioners often pare down the collection $\mathscr{U}$ by restricting attention to various physically reasonable subcollections $\mathscr{P} \subset \mathscr{U}$. But if $\mathscr{U}$ is a physically unreasonable background possibility space, then the significance of the definition of maximality (based on $\mathscr{U}$) is not at all clear. An example may help to illustrate the point.

Consider Misner spacetime (M, g) in (t, θ) coordinates (recall Figure 6.9). As we have seen, the model fails to be chronological since there is a CTC through each point in the $t > 0$ region of M. But consider the spacetime (N, g) where N is the $t < 0$ region of M. Not only does this spacetime not have CTCs, one can show that it is extremely well-behaved causally in the sense of being globally hyperbolic. For each $k < 0$, the $t = k$ slice S_k counts as a Cauchy surface (see Figure 8.1). Moreover, one can show that every extension to (N, g) fails to be globally hyperbolic (Chruściel and Isenberg, 1993). Now suppose, for the sake of argument, that the cosmic censorship conjecture is true: "All physically reasonable spacetimes are globally hyperbolic?" (Wald, 1984, p. 304). It follows that every extension to (N, g) is physically unreasonable. In other words, the $t < 0$ portion of Misner is "as large as it can be" if one restricts attention to the physically reasonable possible extensions. And yet this spacetime is ruled out on maximality grounds under the prevailing dogma.

Given that the collection $\mathscr{U}$ may not adequately capture the notions of physical possibility we are after, its seems natural to explore spacetime

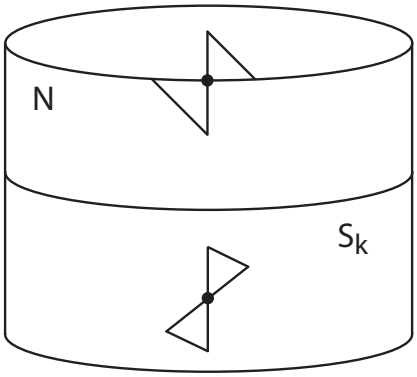

Figure 8.1: The $t < 0$ portion of Misner spacetime is globally hyperbolic. Each $t = k$ slice S_k is a Cauchy surface.

maximality under various choices of background possibility spaces $\mathscr{P} \subset \mathscr{U}$. Indeed, because of examples like the one just given, a move to a plurality of definitions of spacetime maximality is suggested early on by (Geroch, 1970b). For any collection $\mathscr{P} \subseteq \mathscr{U}$, we say a member of $\mathscr{P}$ is a $\mathscr{P}$-**spacetime**. Let us say the $\mathscr{P}$-spacetime (M, g) has a (proper) $\mathscr{P}$-**extension** (N, h), if (N, h) is both a (proper) extension of (M, g) and a $\mathscr{P}$-spacetime. A $\mathscr{P}$-spacetime is $\mathscr{P}$-**extendible** if it has a $\mathscr{P}$-extension and $\mathscr{P}$-**maximal** otherwise.

8.3 Equivalence

We see that for each spacetime property, we have a corresponding definition of spacetime maximality. These alternate definitions give rise to "a number of important and unsolved problems" (Geroch, 1970b, p. 276). Consider a foundational question of meaning: Are there spacetime properties $\mathscr{P} \subseteq \mathscr{U}$ such that the notion of $\mathscr{P}$-maximality is, in some sense, "equivalent" to $\mathscr{U}$-maximality? If so, then it would seem to make no difference whether one considers $\mathscr{P}$-maximality or the standard definition. Perhaps all results established over the decades concerning $\mathscr{U}$-maximality carry over to the context of $\mathscr{P}$-maximality. Consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$ (Geroch, 1970b).

(Equivalence) Any $\mathscr{P}$-spacetime is $\mathscr{P}$-maximal if and only if it is $\mathscr{U}$-maximal.

Given the example of the $t < 0$ portion of Misner spacetime considered above, we see that (Equivalence) is not satisfied by the collection $(GH) \subset \mathscr{U}$ of globally hyperbolic spacetimes. As we have seen, the example is (GH)-maximal but not maximal. On the other hand, it is trivial that the collection $\mathscr{U}$ as well as any subcollection of the collection (Max) of $\mathscr{U}$-maximal spacetimes will satisfy (Equivalence). This includes the collections $(HF), (LM), (GC) \subset \mathscr{U}$ of hole-free, locally maximal, and geodesically complete spacetime respectively. What about the other spacetime properties investigated in Part I?

8.4 Local Properties

First, consider a few local spacetime properties. Geroch has conjectured that the collection $(Vac) \subset \mathscr{U}$ of vacuum solutions of Einstein's equation satisfies (Equivalence). He writes: "While this statement is probably true, no proof is known" (1970, p. 278). It is remarkable that more than fifty years later, a proof (or disproof) of this beautifully simple claim has yet been found. To be sure, settling the question would help to clarify the situation to some degree. But we now emphasize a type of "subcollection problem" concerning the significance of any isolated result of this kind.

Suppose Geroch's conjecture is true and it is the case that the collection (Vac) satisfies (Equivalence). We know that within (Vac) lurk physically unreasonable spacetimes. For example, consider Minkowski spacetime where a closed set spelling out the word "Leibniz" is removed from the manifold (see Figure 8.2). If the "e" and "b" letters are chosen carefully (i.e. without "holes"), then the resulting structure will be a connected spacetime. This spacetime is a member of the collection (Vac) but seems to physically unreasonable in various senses (even if issues of maximality are set aside). Because such spacetimes lurk with within (Vac), this collection does not seem to be a physically reasonable possibility space. So the physical significance of the statement "(Vac) satisfies (Equivalence)" is unclear. To gain clarity, one would like assurance that any subcollection $\mathscr{P} \subseteq (Vac)$ (for example, one

that is more reasonable physically) also satisfies (Equivalence). But such assurance is absent in general – this is the subcollection problem. Each collection of spacetimes, and each of its sub-collections, must be checked independently or some new argument must be introduced for why this is not needed. In the present case, perhaps an example will be useful.

Figure 8.2: A vacuum solution to Einstein's equation. A closed set spelling out the word "Leibniz" has been removed from the manifold.

Let (M, η) be Minkowski spacetime and let (N, η) be Minkowski spacetime with a point removed from M (recall Figure 2.9). Let $\mathscr{P} \subset (Vac)$ be the collection $\{(M, \eta), (N, \eta)\}$. It is not difficult to verify that $\mathscr{P}$ satisfies (Equivalence). We see that (M, g) is both maximal and $\mathscr{P}$-maximal while (N, η) is both extendible and $\mathscr{P}$-extendible. So any $\mathscr{P}$-spacetime is $\mathscr{P}$-maximal if and only if it is maximal. Relative to the collection $\mathscr{P}$, we have a sense in which $\mathscr{P}$-maximality is equivalent to standard maximality. Now let $\mathscr{Q} \subset \mathscr{P}$ be the subcollection $\{(N, \eta)\}$. One can verify that (N, η) is $\mathscr{Q}$-maximal but not maximal. So $\mathscr{Q}$ does not satisfy (Equivalence) (see Figure 8.3). We now see that just because $\mathscr{P}$ satisfies (Equivalence) does not mean that each of its subcollections also does so; each must be checked independently. This is an instance of the subcollection problem that we will encounter often in what follows concerning various second-order modality conditions.

The collections $\mathscr{P}$ and $\mathscr{Q}$ we have just constructed are surely physically unreasonable. These constructions are essentially second-order analogs to

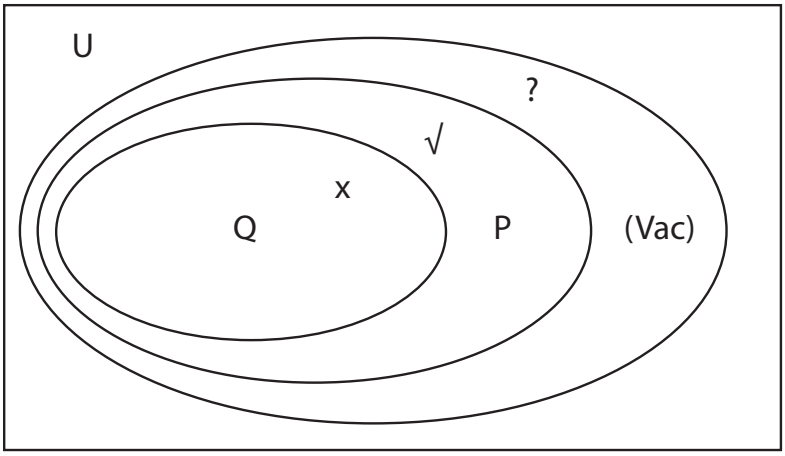

Figure 8.3: It is unknown if (Vac) satisfies (Equivalence). The collection $\mathscr{P}$ does satisfy the condition while one of its subcollections $\mathscr{Q}$ does not.

the "cut and paste" example spacetimes one often finds in the global structure literature. There too, it is acknowledged that the constructed examples are not physically reasonable. Rather, they serve a different purpose as emphasized by Geroch and Horowitz (1979, p. 221):

> "The spacetimes which result from these constructions are, in almost every case, physically unrealistic for various reasons. The point of the construction, however, is not normally to construct physically realistic cosmological models, but rather to demonstrate by means of some example that a certain assertion is false, or that a certain line of argument cannot work."

In the present case, the constructions of $\mathscr{P}$ and $\mathscr{Q}$ show us that the satisfaction of a second-order modal property like (Equivalence) by a spacetime collection does not automatically "transfer down" to its subcollections. This makes it difficult to get a good sense of how common it is that such second-order conditions are satisfied. It is sometimes possible, however, to settle many cases at once. For example, consider the spacetime collections $(NEC), (WEC), (SEC), (DEC) \subset \mathscr{U}$ satisfying, respectively, the null, weak, strong, and dominant, energy conditions. At the very end of his paper, Geroch (1970b, p. 289) wondered about the status of (Equivalence) relative to these collections. Recall that $(DEC) \subset (WEC) \subset (NEC)$ and

$(SEC) \subset (NEC)$. One can use these relations to show the following general result (Manchak, 2021): Any collection $\mathscr{P} \subset \mathscr{U}$ such that either (i) $(DEC) \subseteq \mathscr{P} \subseteq (NEC)$ or (ii) $(SEC) \subseteq \mathscr{P} \subseteq (NEC)$ does not satisfy (Equivalence) (see Figure 8.4).

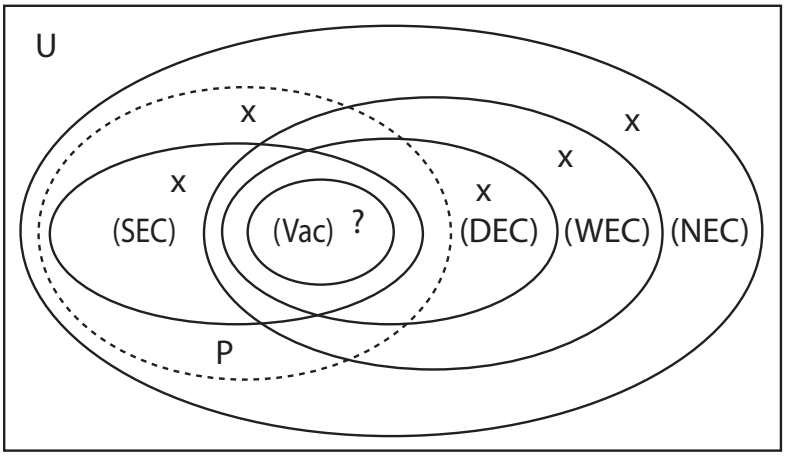

Figure 8.4: Any collection $\mathscr{P} \subset \mathscr{U}$ such that either (i) $(DEC) \subseteq \mathscr{P} \subseteq (NEC)$ or (ii) $(SEC) \subseteq \mathscr{P} \subseteq (NEC)$ does not satisfy (Equivalence). The case of (Vac) is unsettled.

8.5 Causal Properties

Let's now consider causal properties. We know that (Equivalence) is not satisfied by the collection $(GH) \subset \mathscr{U}$ of globally hyperbolic spacetimes. Consider again the spacetime (N, g) which is the $t < 0$ portion of Misner spacetime (recall Figure 8.1). It is (GH)-maximal but not $\mathscr{U}$-maximal. But not only does every extension to (N, g) fail to be globally hyperbolic, every extension must even fail to be distinguishing. This follows since any extension will include a neighborhood O of some event p at $t = 0$. But since there must be a distinct point $q \in O$ also at $t = 0$, we find that $I^-(p) = I^-(q) = N$ (see Figure 8.5). Since (N, g) is globally hyperbolic and therefore distinguishing, we see that the collection $(Dist) \subset \mathscr{U}$ of distinguishing spacetimes does not satisfy (Equivalence). But the example shows much more than this: any

collection $\mathscr{P} \subset \mathscr{U}$ such that $(GH) \subseteq \mathscr{P} \subseteq (Dist)$ must fail to satisfy (Equivalence).

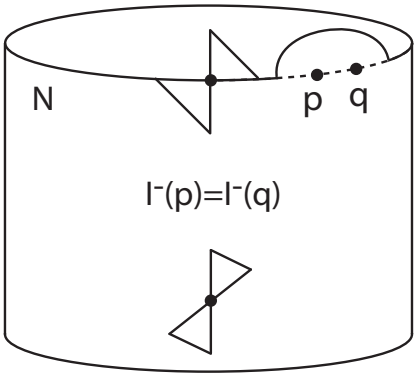

Figure 8.5: Any extension to the $t < 0$ portion of Misner spacetime (N, g) will fail to be distinguishing since it will have a pair of distinct points p, q at $t = 0$ for which $I^-(p) = I^-(q) = N$.

In addition, the collection $(Caus) \subset \mathscr{U}$ of causal spacetimes does not satisfy (Equivalence). To see why, we return to a spacetime that we have already constructed showing that a spacetime can fail to be distinguishing condition and yet still be causal (recall Figure 5.5). Start by letting M be the manifold $\mathbb{R} \times S$ in (t, θ) coordinates. Consider the spacetime (M, g) where the metric g is defined as follows: at each point $(t, \theta) \in M$ and for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $g(v, w) = v_t w_\theta + v_\theta w_t - t^2 v_\theta w_\theta$. In the $t < 0$ region, the causal structure is similar to Misner spacetime: as t increases, the light cones open up and tip over. At $t = 0$, there is a null geodesic in (M, g) just as in Misner spacetime. But in the $t > 0$ region, the spacetimes are very different. In (M, g) the light cones close up as t increases so that no CTCs exist. Now remove the point $(0, 0)$ from M and let N be the resulting manifold. Because of the "missing" point, the single closed null curve no longer closes ensuring that (N, g) satisfies causality (see Figure 8.6)

Given that (M, g) is a maximal spacetime and we have removed a single point $(0, 0)$ to produce (N, g), it would seem that any extension to this spacetime must restore the "missing" point. In other words, in would seem that (M, g) is the only extension (up to isometry) to (N, g). Stepping back,

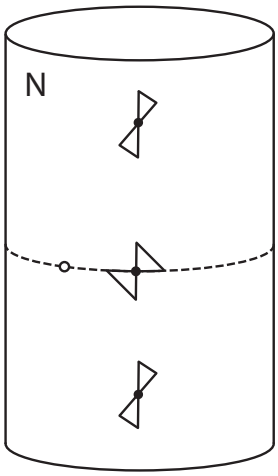

Figure 8.6: The spacetime (N, g) is causal since the "missing" point prevents the null curve from closing.

there is a more general conjecture here. It would seem that for any maximal spacetime (M, g) and any point $p \in M$, there is a unique (up to isometry) extension to the spacetime $(M - \{p\}, g)$: it must be (M, g) itself. Over the years, a number of leading experts were asked about this conjecture in private communication. A consensus emerged that (i) it was true but that (ii) a proof would be difficult to secure. In a recent paper on unique spacetime extensions, the situation was finally clarified by Jan Sbierski (2024, p. 13226): "we answer a question by JB Manchak in the affirmative as to whether the only possible (smooth) extension of an inextendible Lorentzian manifold with one point removed is the restoration of this point." The proof is non-trivial but the result does seem to accord with intuition. It follows from the Sbierski result that the only extension to the causal spacetime (N, g) is the non-causal spacetime (M, g). So (N, g) is $(Caus)$-maximal but not $\mathscr{U}$-maximal. Thus $(Caus)$ does not satisfy (Equivalence). It is worth appreciating that, while $(Caus)$ and $(Dist)$ both fail to satisfy (Equivalence), it unknown if any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Dist) \subseteq \mathscr{P} \subseteq (Caus)$ also fails to satisfy (Equivalence). This is another instance of the subcollection problem discussed above.

The only causal property we have yet to explore is the collection $(Chron) \subset \mathscr{U}$ of chronological spacetimes. Perhaps $(Chron)$ satisfies (Equivalence)? This simple question was posed by Geroch (1970b, p. 278)

and remains open more than fifty years later. But no matter what the outcome, we know that up and down the causal hierarchy, various collections of spacetimes overwhelmingly fail to satisfy (Equivalence) (see Figure 8.7).

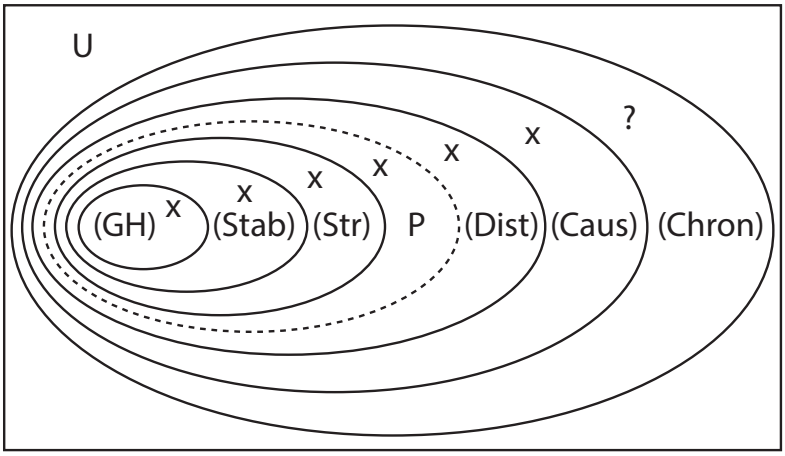

Figure 8.7: Any collection $\mathscr{P} \subset \mathscr{U}$ such that $(GH) \subseteq \mathscr{P} \subseteq (Dist)$ does not satisfy (Equivalence). Neither does $(Caus)$. The case of $(Chron)$ is unknown.

8.6 Asymmetry Properties

We turn now to asymmetry properties. Consider two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates. Remove the null related points $(0, 0)$ and $(1, 1)$ and let the resulting spacetime be (N, η). One can verify that the spacetime is giraffe: the only isometry $f : N \to N$ is the identity map. One can adapt the Sbierski (2024) uniqueness result to show that there are only two possible extensions (up to isometry) to (N, η). One of them replaces one of the two "missing" points; the other replaces both points to recover (M, g). (The result of replacing only $(0, 0)$ is isometric to the result of replacing only $(1, 1)$.) Of course, Minkowski spacetime does not have a fixed point and fails to be point rigid. It turns out this is also true of the other extension as well. To see this, suppose that the point $(1, 1)$ is replaced and the origin $(0, 0)$ remains "missing" (the other case is handled similarly). Then there will be one reflection isometry across the $t = 0$ line and another across the $x = 0$

line. For any point p, one of these reflections will map p into a distinct point. This shows that the extension has no fixed point. Moreover, the reflection isometry across the $t = 0$ line takes the point $q = (0, 1)$ into itself. This shows that the extension is not point rigid (see Figure 8.8).

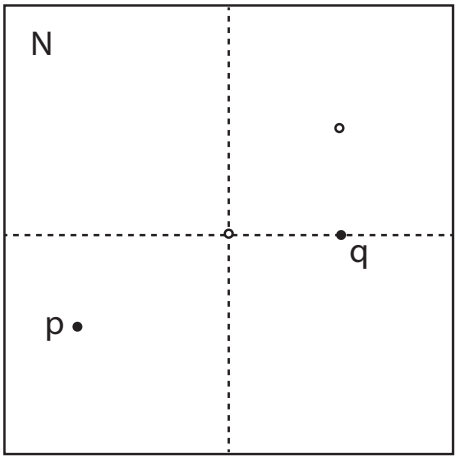

Figure 8.8: A giraffe spacetime with two "missing" points. If $(1, 1)$ is replaced, a reflection across either the $t = 0$ or $x = 0$ will map any point p into a distinct point. The reflection across $t = 0$ maps the point q into itself.

It follows from all of this that for any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (FP) \cup (PR)$ fails to satisfy (Equivalence). What about the collections $(LG), (Her) \subset \mathscr{U}$ of spacetimes satisfying, respectively, the locally giraffe and Heraclitus conditions? As far as we are aware, both questions are open (see Figure 8.9).

8.7 Conclusion

Stepping back, we see (Equivalence) is false relative to almost all of the spacetime properties under consideration. Moreover, none of the properties are known to render (Equivalence) true. A few important cases remain unsettled – most notably the collections (Vac) and $(Chron)$ highlighted by (Geroch, 1970b). But we have also seen that even if positive results were to obtain for these properties, the subcollection problem calls into question the significance of isolated results of this kind. Given the situation, it seems clear

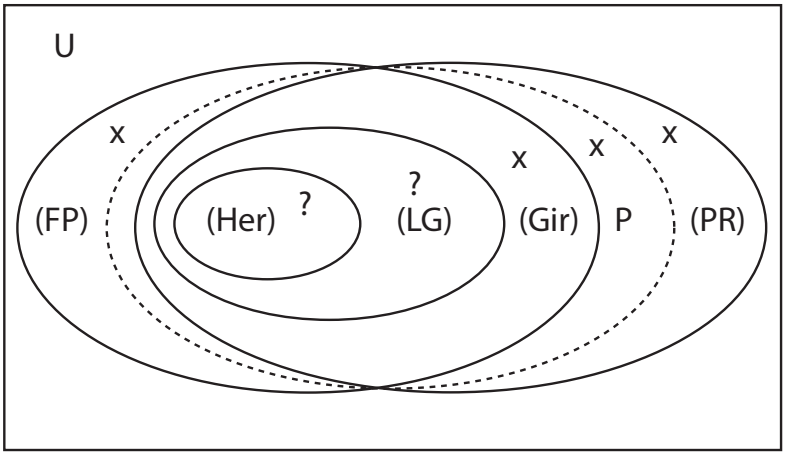

Figure 8.9: Any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (FP) \cup (PR)$ fails to satisfy (Equivalence). The cases of (LG) and (Her) are unknown.

that further investigation is needed to better understand $\mathscr{P}$-maximality with resect to various physically reasonable collections $\mathscr{P} \subset \mathscr{U}$.

Chapter 9

Metaphysics

9.1 Introduction

In the previous chapter, we saw that virtually every physically significant property $\mathscr{P} \subset \mathscr{U}$ does not satisfy (Equivalence). For such collections, a spacetime can be $\mathscr{P}$-maximal and yet extendible in $\mathscr{U}$. Given the state of affairs, one worries that foundational theorems concerning spacetime maximality proved relative the background possibility space $\mathscr{U}$ do not transfer over to its more physically reasonable subcollections. Perhaps the most important such theorem is the statement that any extendible spacetime has a $\mathscr{U}$-maximal extension (Geroch, 1970b). Recall that upon this foundational result rests the metaphysical justification for the spacetime maximality condition via the Leibnizian principles of sufficient reason and plenitude (Earman, 1989, p. 161).

We begin this chapter by introducing a generalized statement of the Geroch (1970b) existence theorem relativized to any collection $\mathscr{P} \subseteq \mathscr{U}$. Next, we review Zorn's lemma which is used to secure the existence theorem within the context of $\mathscr{U}$. Over the next few sections, we then investigate whether this generalized existence statement is true with respect to the various local, causal, and asymmetry properties identified in Part I. We will find that in some cases an analogue to the Geroch (1970b) existence result can be proven with the help of Zorn's lemma. But we emphasize that there are many cases in which Zorn's lemma cannot be applied. We highlight a number of open questions that, if settled, could help clarify the situation. We close with a discussion of the "big bang" property for which the analogue existence result

fails.

9.2 Existence

Let's jump right in. Consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (Existence) Any $\mathscr{P}$-extendible $\mathscr{P}$-spacetime has a $\mathscr{P}$-maximal extension.

If (Existence) were false for any collection $\mathscr{P} \subseteq \mathscr{U}$, then the analog to the Geroch (1970) result would fail relative to $\mathscr{P}$. Thus, within that context, Leibnizian metaphysical justification would face significant difficulties in getting off the ground. We begin our study by noting that (Existence) and (Equivalence) are independent conditions. A number of spacetime properties of interest render the first true but the second false. We will explore many of them soon. For now, we draw attention to the fact that even if (Equivalence) is true for some collection $\mathscr{P} \subseteq \mathscr{U}$, it does not follow that (Existence) is also true for $\mathscr{P}$.

To see this, let (M, g) be the $t < 0$ portion of two-dimensional Minkowski spacetime in (t, x) coordinates. Let $\mathscr{P} = \{(M, g)\}$. We know that (M, g) is extendible in $\mathscr{U}$. But counterintuitively, it also extendible in $\mathscr{P}$ since the spacetime (M, g) extends itself. Let (N, g) be the $t < -1$ portion of (M, g) and let $f : M \to N$ be the isometry defined by $f(t, x) = (t-1, x)$ which shifts all points one unit in the negative t direction (see Figure 9.1). Since (M, g) is isometric to a proper subset of (M, g), the spacetime counts as a (proper) extension of itself. Since (M, g) is both extendible and $\mathscr{P}$-extendible, we see that (Equivalence) is true for $\mathscr{P}$. But (Existence) fails since $\mathscr{P}$ contains no $\mathscr{P}$-maximal spacetimes. So we see that it is not the case (as was the hope mentioned in the previous chapter) that if (Equivalence) were true for some collection $\mathscr{P} \subseteq \mathscr{U}$, then all results established over the decades concerning $\mathscr{U}$-maximality carry over to the context of $\mathscr{P}$-maximality.

Before moving on to explore the (Existence) condition, let's take a brief look at the significance of singleton collections $\mathscr{P} = \{(M, g)\}$ like the one just given. Such collections have been considered in discussions on the metaphysics of laws of nature. For example, Earman (1986) considers a situation

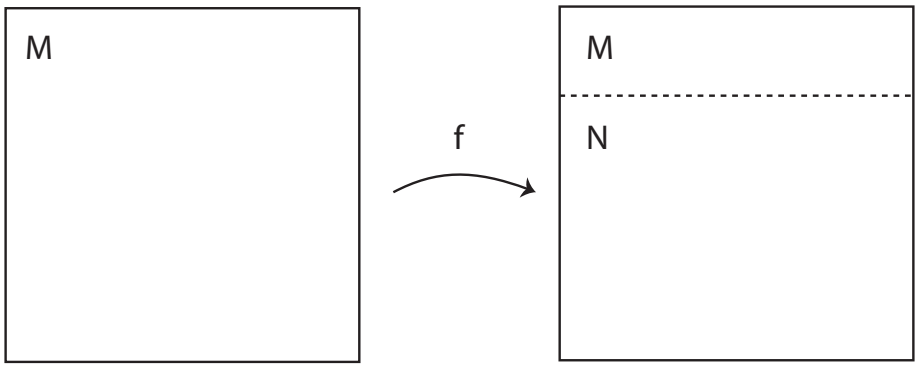

Figure 9.1: The $t < 0$ portion of Minkowski spacetime is isometric to the $t < -1$ portion of Minkowski spacetime. The isometry f maps any point $p = (t, x)$ in M to the point $f(p) = (t - 1, x)$ in N.

in which the actual universe is the only possible universe permitted by the laws. He also remarks that, although Leibniz wished to avoid the "absolute metaphysical fatalism" that would obtain in such a scenario, his principle of sufficient reason seems to push him towards it (Earman, 1986, p. 19). There is currently a renewed interest in this type of "strong determinism" including the study of singleton collections of general relativistic spacetimes (Chen, 2024, p. 56). Here, we note that the example considered above shows that the modal structure of spacetime can be non-trivial for some of these singleton collections. Surprisingly, spacetime can fail to be "as large as it can be" even if the background possibility space has only one element. Of course, the situation only arises because the single spacetime can properly extend itself. One wonders if this consequence can be avoided for spacetimes satisfying some condition of physical interest. Of course, any no-hole condition that implies $\mathscr{U}$-maximality (e.g. geodesic completeness) is trivially sufficient. Using causal properties or local energy properties will not work since the example of $t < 0$ portion of Minkowski spacetime is a globally hyperbolic vacuum solution to Einstein's equation. But looking to spacetime asymmetry properties turns out to be fruitful.

It is not difficult to see that no Heraclitus spacetime can extend itself. Recall that a necessary and sufficient condition for the Heraclitus property

to obtain in a spacetime (M, g) is the following: for any open sets $U, V \subseteq M$ and any isometry $f : U \to V$ we have (i) $U = V$ and (ii) f is the identity map. It is now immediate that a Heraclitus spacetime (M, g) cannot extend itself since in that case there would be some proper subset $N \subset M$ and an isometry $f : M \to N$. Can the Heraclitus condition be weakened while still ensuring that spacetimes do not extend themselves? The case of locally giraffe is not yet clear. But certainly the giraffe condition is not strong enough for these purposes. To see this, let (M, g) be the $t < 0$ portion of Minkowski spacetime as before. Now remove the points $(-1, 0)$ and $(-2, 1)$ and let the resulting spacetime be (N, g). Because of the "missing" points, this spacetime is giraffe. (Removing just one point is not enough since there will be a reflection isometry in that case.) Let O be the set consisting of the $t < -2$ portion of N except for the points $(-3, 0)$ and $(-4, 1)$ (see Figure 9.2). There is an isometry $f : N \to O$ defined by $f(t, x) = (t - 2, x)$. Since (N, g) is isometric to a proper sub-portion (O, g), we find that (N, g) extends itself.

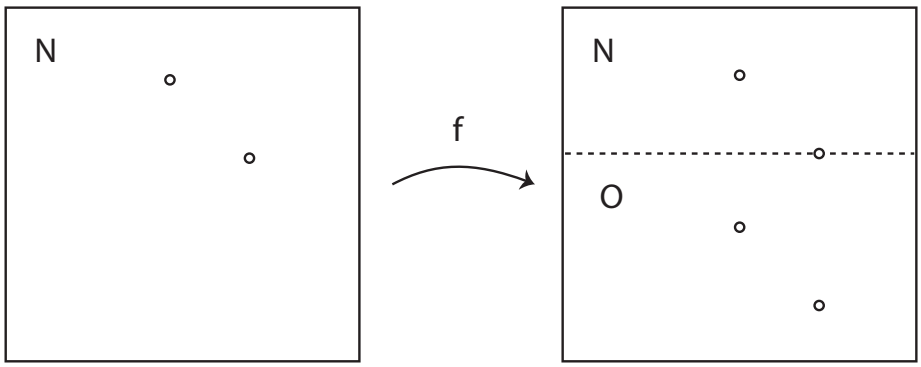

Figure 9.2: (N, g) is the $t < 0$ portion of Minkowski spacetime with the points $(-1, 0)$ and $(-2, 1)$ removed. It is isometric to the proper subset $O \subset N$. The isometry f maps any point $p = (t, x)$ in N to the point $f(p) = (t - 2, x)$ in O.

9.3 Zorn's Lemma

Let us now turn to the question of which collections $\mathscr{P} \subseteq \mathscr{U}$ satisfy the (Existence) condition. In order to do this, we need to review a foundational axiom of set theory: Zorn's lemma. Let S be a set. A relation $\leq$ on S is a **partial order** if, for all $a, b, c \in S$, the following hold: (reflexivity) $a \leq a$, (transitivity) if $a \leq$ and $b \leq c$, then $a \leq c$, and (anti-symmetry) if $a \leq b$ and $b \leq a$, then $a = b$. Consider an example. Let (M, g) be any spacetime and let $\leq$ be the causality relation on M, i.e. for any $p, q \in M$, let $p \leq q$ hold if and only if $p \in J^-(q)$. It is immediate that this relation must satisfy (i) reflexivity and (ii) transitivity. One can verify that (iii) anti-symmetry will also be satisfied if the spacetime (M, g) is causal.

If $\leq$ is a partial ordering on a set S, we say a subset $T \subseteq S$ is **totally ordered** if, for all $a, b \in T$, either $a \leq b$ or $b \leq a$. A totally ordered set will sometimes be called a "chain" in what follows. Let $\leq$ be a partial ordering on S and let T be any subset of S. An **upper bound** for the set T is an element $u \in S$ such that for all $a \in T$, we have $a \leq u$. Note that an upper bound for T need not be a member of T itself. For example, consider a causal spacetime (M, g) and the partial order on M given by the causality relation $\leq$ considered above. For any point $p \in M$, an upper bound for the set $J^-(p)$ is any point $q \in J^+(p)$. This includes the point p itself which is a member of $J^-(p)$. But any other point $q \neq p$ in $J^+(p)$ fails to be in $J^-(p)$ and yet still qualifies as an upper bound for the latter set.

A **maximal element** of a set S partially ordered by the relation $\leq$ is an element $m \in S$ such that for all $c \in S$, if $m \leq c$, then $c = m$. A maximal element in S is one that is not "dominated" by any other member of S. In the example above where (M, g) is a causal spacetime and the partial order on M given by the causality relation $\leq$, there is no maximal element of M. For any point $p \in M$, there will be a point $q \in J^+(p)$ such that $p \leq q$ and $p \neq q$. We note that a maximal element need not be an upper bound for a set S partially ordered by the relation $\leq$. For example, let $S = \{a, b\}$ and let $\leq = \{(a, a), (b, b)\}$. We see that both a and b are maximal elements in S since neither is dominated and yet neither is an upper bound for S.

Zorn's lemma is the following. Let $\leq$ be a partial order on S. If each totally ordered subset $T \subseteq S$ has an upper bound, there is a maximal element of S. Zorn's lemma is equivalent to the axiom of choice relative to standard background set theoretic axioms. (The axiom of choice is defined below with an application.) Zorn's lemma is often used to show the existence of various

mathematical objects. For example, it is invoked in the proof that every vector space has a basis. Zorn's lemma is also central to the Geroch (1970b) result that every extendible spacetime has a maximal extension. Let's explore the proof of this fundamental statement.

We start with an intuitive idea. Let us say that a spacetime (M, g) can be **isometrically embedded** into a spacetime (N, h) if, for some set $O \subseteq N$, there is an isometry $f : M \to O$ from (M, g) to (O, h). It is immediate that a spacetime (M, g) can be isometrically embedded into a spacetime (N, h) if and only if either (N, h) is an extension of (M, g) or the spacetimes are isometric. Consider the collection $\mathscr{U}$ of all spacetimes and let $\leq$ be a relation on $\mathscr{U}$ defined such that, for all $(M, g), (N, h) \in \mathscr{U}$, we have $(M, g) \leq (N, h)$ if (M, g) can be isometrically embedded into a spacetime (N, h). At once we see that the relation $\leq$ is reflexive and transitive. Unfortunately it fails to be anti-symmetric. To see this, just consider a pair of distinct but isometric spacetimes. One could perhaps take Minkowski spacetime (M, η) and use a hole diffeomorphism $f : M \to M$ to pull back the metric η to construct the spacetime $(M, f^*(\eta))$ (recall Figure 7.2). As we have seen, f cannot be an isometry from (M, η) to itself. So $(M, \eta) \neq (M, f^*(\eta))$. But by construction, the hole diffeomorphism f is an isometry from $(M, f^*(\eta))$ to (M, η). Now the problem becomes clear. Since the two spacetimes are isometric, we have $(M, \eta) \leq (M, f^*(\eta))$ and $(M, f^*(\eta)) \leq (M, \eta)$ and yet $(M, \eta) \neq (M, f^*(\eta))$. So the relation is not anti-symmetric and therefore not a partial order.

A natural way to fix things up presents itself. Let $\sim$ be the relation on $\mathscr{U}$ defined such that, for all $(M, g), (N, h) \in \mathscr{U}$, we have $(M, g) \sim (N, h)$ if (M, g) is isometric to (N, h). We know $\sim$ is equivalence relation on $\mathscr{U}$. Let $[(M, g)]$ be the equivalence class of any spacetime $(M, g) \in \mathscr{U}$ and let $\mathscr{U} / \sim$ be the collection of all equivalence classes of all spacetimes. Consider the relation $\leq$ on $\mathscr{U} / \sim$ defined such that, for all $[(M, g)], [(N, h)] \in \mathscr{U} / \sim$, we have $[(M, g)] \leq [(N, h)]$ if any member of $[(M, g)]$ can be isometrically embedded into any member of $[(N, h)]$. It is easy to check that the relation $\leq$ is once again reflexive and transitive. Moreover, the problem from above is now avoided since, if (M, η) and $(M, f^*(\eta))$ are isometrically related by a hole diffeomorphism f, then $[(M, \eta)] = [(M, f^*(\eta))]$. So it would seem that $\leq$ now counts as a partial relation on $\mathscr{U} / \sim$. But it turns out even this is not true. There is a different sort of problem lurking here.

Let (M, g) be the $t < 0$ portion of Minkowski spacetime in (t, x) coordinates. Now remove the point $(-1, 0)$ from M and let the resulting spacetime be (N, g). By construction, (N, g) can be isometrically embedded into (M, g)

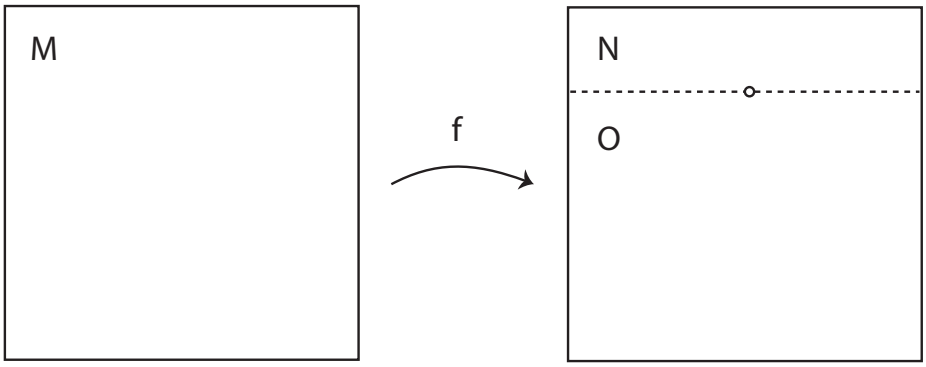

Figure 9.3: The isometry f maps any point $p = (t, x)$ in M to the point $f(p) = (t - 1, x)$ in O. So (N, g) is an extension of (M, g).

using the inclusion map. So $[(N, g)] \leq [(M, g)]$. But (M, g) can also be isometrically embedded into (N, g). To see this, let O be the $t < -1$ portion of N. There is an isometry $f : M \to O$ defined by $f(t, x) = (t - 1, x)$ (see Figure 9.3). So we have both $[(N, g)] \leq [(M, g)]$ and $[(M, g)] \leq [(N, g)]$. But (M, g) and (N, g) are not isometric because of the "missing" point in the latter spacetime. So $[(M, g)] \neq [(N, g)]$ and therefore the relation $\leq$ fails to be anti-symmetric once again.

How can one define an appropriate partial order? The key is to consider not equivalence classes of isometric spacetimes but rather equivalence classes of isometric "framed" spacetimes. We say the triple (M, g, F) is a **framed spacetime** if (M, g) is a spacetime and F is an orthonormal basis of vectors at some point $p \in M$. Let (M, g, F) and (N, h, E) be framed spacetimes. An isometry $f : M \to N$ is a **framed isometry** if the frame E is the result of pushing forward the frame F via f. For any framed spacetimes (M, g, F) and (N, h, F), a **framed embedding** is a map $e : M \to N$ such that if the range is restricted to $e[M] \subseteq N$, then e is a framed isometry. For any collection $\mathscr{P} \subseteq \mathscr{U}$, let $\mathfrak{F}(\mathscr{P})$ be the collection of all framed spacetimes (M, g, F) such that $(M, g) \in \mathscr{P}$. Let $\sim$ be the relation on $\mathfrak{F}(\mathscr{U})$ defined such that, for all $(M, g, F), (N, h, E) \in \mathfrak{F}(\mathscr{U})$, we have $(M, g, F) \sim (N, h, E)$ if there is a framed isometry from (M, g, F) to (N, h, E). One can check that $\sim$ is an equivalence relation on $\mathfrak{F}(\mathscr{U})$. Let $[(M, g, F)]$ be the equivalence class of any

framed spacetime $(M, g, F) \in \mathfrak{F}(\mathscr{U})$ and let $\mathfrak{F}(\mathscr{U})/\sim$ be the collection of all equivalence classes of all framed spacetimes.

Now consider the relation $\leq$ on $\mathfrak{F}(\mathscr{U})/\sim$ defined such that, for all $[(M, g, F)], [(N, h, E)] \in \mathfrak{F}(\mathscr{U})/\sim$, we have $[(M, g, F)] \leq [(N, h, E)]$ if there is a framed embedding of any member of $[(M, g, F)]$ into any member of $[(N, h, E)]$. It is immediate that the relation $\leq$ is reflexive and transitive. We now show that anti-symmetry also holds. Suppose it were the case that both $[(M, g, F)] \leq [(N, h, E)]$ and $[(N, h, E)] \leq [(M, g, F)]$ and yet $[(N, h, E)] \neq [(M, g, F)]$. Since $[(N, h, E)] \neq [(M, g, F)]$, there is no framed isometry from (M, g, F) to (N, h, E). Because both $[(M, g, F)] \leq [(N, h, E)]$ and $[(N, h, E)] \leq [(M, g, F)]$, it follows that there is a proper framed embedding from (M, g, F) into (N, h, E) and vice versa. So there is a proper framed embedding from (M, g, F) into itself. But this cannot be. Recall the general rigidity result due to Geroch (1969) we considered in Section 7.3 which can now be expressed in terms of framed spacetimes: there is at most one framed embedding of a framed spacetime into another. Since the identity map always counts as a framed embedding of a framed spacetime to itself, we know that there can never be a proper framed embedding from a framed spacetime into itself. So we now have a contradiction. We conclude that if both $[(M, g, F)] \leq [(N, h, E)]$ and $[(N, h, E)] \leq [(M, g, F)]$, then $[(N, h, E)] = [(M, g, F)]$. Thus, the relation $\leq$ satisfies anti-symmetry and is therefore a partial order on $\mathfrak{F}(\mathscr{U})/\sim$.

Stepping back, one might wonder about the possibility of defining a partial order on the collection $\mathfrak{F}(\mathscr{U})$ of framed spacetimes rather than the collection $\mathfrak{F}(\mathscr{U})/\sim$ of equivalence classes of framed spacetimes. Indeed, in the original paper Geroch (1970, p. 276) and in subsequent presentations such as Earman (1995, p. 32), equivalence classes are nowhere in sight. However, it is clear that in these presentations the collection $\mathfrak{F}(\mathscr{U})/\sim$ is used implicitly. To see why this must be, consider again Minkowski spacetime (M, η) in (t, x) coordinates and a hole diffeomorphism $f : M \to M$ such that f acts as the identity outside of the $t > 0$ hole and non-trivially inside it. One can pull back the metric η to construct the spacetime $(M, f^*(\eta))$. Now consider the point $p = (-1, 0)$ in each spacetime and the frame F given by the orthonormal basis vectors $[1, 0]$ and $[0, 1]$ at p. By construction, the hole diffeomorphism f is an isometry from $(M, f^*(\eta))$ to (M, η). Moreover, because this isometry acts as the identity outside the hole, we see that f maps p into itself and the frame F into itself (see Figure 9.4). So f counts as a framed embedding of $(M, f^*(\eta), F)$ into (M, η, F). Similarly, the inverse

f^{-1} counts as a framed embedding of (M, η, F) into $(M, f^*(\eta), F)$. Yet by construction $(M, \eta, F) \neq (M, f^*(\eta), F)$. So we see that in order to satisfy the anti-symmetry condition for a partial order, equivalence classes of framed spacetimes must be brought into the picture.

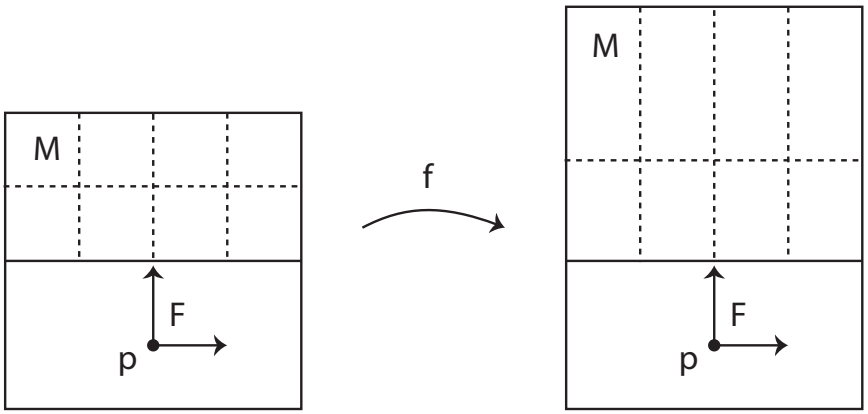

Figure 9.4: The isometry f acts as the identity outside the $t > 0$ hole. So it must map the frame F at the point $p = (-1, 0)$ into itself.

We finally have the desired partial order $\leq$ on the collection $\mathfrak{F}(\mathscr{U})/\sim$ of equivalence classes of framed spacetimes. Let's put it to use. Consider any subset $\mathfrak{T}$ of $\mathfrak{F}(\mathscr{U})/\sim$ that is totally ordered by $\leq$. We now show that $\mathfrak{T}$ has an upper bound in $\mathfrak{F}(\mathscr{U})/\sim$. Let $\{X_i\}$ be the totally ordered collection of equivalence classes of framed spacetimes in $\mathfrak{T}$ which is indexed such that $X_i \leq X_j$ if and only if $i \leq j$. For each equivalence class $X_i \in \mathfrak{T}$, choose a representative framed spacetime $(M_i, g_i, F_i) \in X_i$. To do this, one needs to invoke the **axiom of choice**: for any set S of nonempty sets, there exists function that maps each set $A \in S$ to an element of A. For any framed spacetimes $(M_i, g_i, F_i) \in X_i$ and $(M_j, g_j, F_j) \in X_j$ such that $i \leq j$, there must be a framed embedding $e_{ij} : M_i \to M_j$. From the Geroch (1969) rigidity result, we know that this framed embedding must be unique. We now take a type of "union" of all of the (M_i, g_i, F_i) to construct the framed spacetime (M, g, F) (Hawking and Ellis, 1973, p. 249). We first define the manifold M by taking the union of all the manifolds M_i and then, for any framed spacetimes (M_i, g_i, F_i) and (M_j, g_j, F_j) such that $i \leq j$, we identify any point $p_i \in M_i$ with the point $e_{ij}(p_i) \in M_j$ (see Figure 9.5). Thus, each

point p in the manifold M is really an equivalence class of points from the union of all the manifolds M_i. We will not formally keep track of these equivalence classes of points in order to simplify the presentation. But one can verify that M does count as a (connected, Hausdorff) manifold under this construction.

For each framed spacetime (M_j, g_j, F_j), let $f_i : M_i \to M$ be the function that takes each point $p_i \in M_i$ into its equivalence class in M. This function counts as a diffeomorphism from M_i to $f_i[M_i]$. On each open region $f_i[M_i]$ of M, let the metric g be the push forward $f_{i*}(g_i)$. So $f_i : M_i \to M$ now counts as an isometry from M_i to $f_i[M_i]$. Finally, choose some framed spacetime (M_i, g_i, F_i) and let F be the result of pushing forward the frame F_i using f_i. The resulting structure (M, g, F) is a framed spacetime and, by construction, $X_i \leq [(M, g, F)]$ for all $X_i \in \mathfrak{T}$. So $[(M, g, F)]$ is an upper bound for $\mathfrak{T}$.

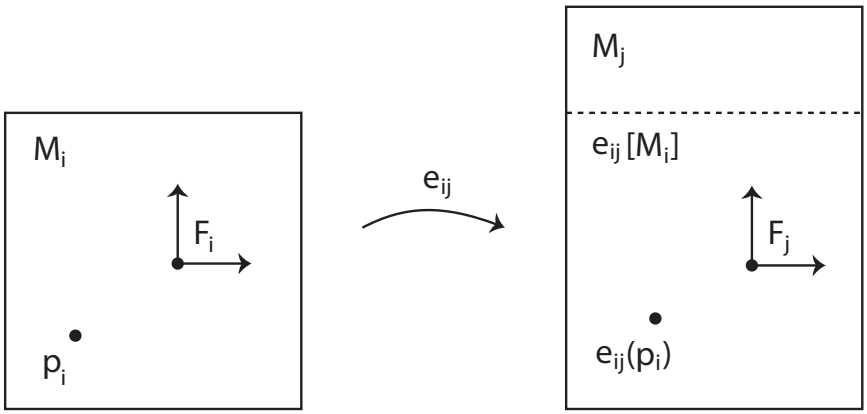

Figure 9.5: For any framed spacetimes (M_i, g_i, F_i) and (M_j, g_j, F_j) such that $i \leq j$, there is a unique framed embedding $e_{ij} : M_i \to M_j$. Each point $p_i \in M_i$ is identified with the point $e_{ij}(p_i) \in M_j$.

We have shown that any subcollection of $\mathfrak{F}(\mathscr{U})/\sim$ that is totally ordered by $\leq$ must have an upper bound in $\mathfrak{F}(\mathscr{U})/\sim$. We are now ready to invoke Zorn's lemma to show that any extendible spacetime has a maximal extension. Let (M, g) be any extendible spacetime. Consider the framed spacetime (M, g, F) where F is any frame at any point $p \in M$. Let $\mathfrak{X} \subset \mathfrak{F}(\mathscr{U})/\sim$ be the collection of all equivalence classes $[(N, h, E)]$ of framed spacetimes in $\mathfrak{F}(\mathscr{U})/\sim$ such that $[(M, g, F)] \leq [(N, h, E)]$. So $\mathfrak{X}$ is the collection of all

equivalence classes of all framed extensions of (M, g, F). Since $\mathfrak{F}(\mathscr{U})/\sim$ is partially ordered by $\leq$, so is $\mathfrak{X}$. Consider any collection of equivalence classes of framed spacetimes in $\mathfrak{X}$ that is totally ordered by $\leq$. We know this collection has an upper bound X in $\mathfrak{F}(\mathscr{U})/\sim$ from the argument given above. So $[(M, g, F)] \leq X$. Because $[(M, g, F)] \leq X$, we know $X \in \mathfrak{X}$ by the definition of $\mathfrak{X}$. So by Zorn's lemma, there is a maximal element $X^* \in \mathfrak{X}$. Let (M^*, g^*, F^*) be any framed spacetime in X^*. It follows that (M^*, g^*, F^*) cannot be properly frame extended be any framed spacetime. So (M^*, g^*) is a $\mathscr{U}$-maximal extension of (M, g).

The application of Zorn's lemma in this context shows that the collection $\mathscr{U}$ of all spacetimes satisfies (Existence). What about various subcollections $\mathscr{P} \subset \mathscr{U}$ of interest? In some cases, one can use Zorn's lemma in an argument that mirrors the one just given. Given a collection $\mathscr{P} \subset \mathscr{U}$, recall that $\mathfrak{F}(\mathscr{P})/\sim$ is the collection of all equivelence classes of all framed spacetimes $(M, g, F) \in \mathfrak{F}(\mathscr{U})$ such that $(M, g) \in \mathscr{P}$. Because $\mathfrak{F}(\mathscr{U})/\sim$ is partially ordered by the relation $\leq$, so is $\mathfrak{F}(\mathscr{P})/\sim$. Let $\mathfrak{T}$ be any subset of $\mathfrak{F}(\mathscr{P})/\sim$ that is totally ordered by $\leq$. If $\mathfrak{T}$ has an upper bound in $\mathfrak{F}(\mathscr{P})/\sim$, then one can invoke Zorn's lemma as before to conclude that $\mathscr{P}$ satisfies (Existence). We shall consider several such examples below. But sometimes there is no upper bound for $\mathfrak{T}$ in $\mathfrak{F}(\mathscr{P})/\sim$. In that case, the use of Zorn's lemma is blocked and it is unclear whether $\mathscr{P}$ satisfies (Existence). A number of examples of this type will also be explored. Finally, we have already seen that (Existence) can be false for some $\mathscr{P}$ – recall the singleton collection consisting of the $t < 0$ portion of Minkowski spacetime. We will highlight that similar situations can also arise in more physically reasonable examples.

9.4 Local Properties

We start with local spacetime properties. Let us say that a property $\mathscr{P} \subset \mathscr{U}$ is **strongly local** if, for any spacetime (M, g) and any open cover $\{O_i\}$ of M, we have $(M, g) \in \mathscr{P}$ if and only if $(O_i, g) \in \mathscr{P}$ for all O_i (Krasnikov, 2014). One can show that any strongly local property must be a local property but not the other way around (Manchak, 2021). Strongly local properties include any of the usual local properties we have been considering: the spacetime collections (NEC), (WEC), (SEC), or (DEC) satisfying, respectively, the null, weak, strong, and dominant energy conditions, or the collection (Vac) of vacuum solutions to Einstein's equation.

Suppose $\mathscr{P} \subset \mathscr{U}$ is any of these strongly local collections. So $\mathfrak{F}(\mathscr{P})/\sim$ is the collection of all equivalence classes of all framed spacetimes $(M, g, F) \in \mathfrak{F}(\mathscr{U})$ such that $(M, g) \in \mathscr{P}$. This collection $\mathfrak{F}(\mathscr{P})/\sim$ is partially ordered by the relation $\leq$. Let $\mathfrak{T}$ be any subset of $\mathfrak{F}(\mathscr{P})/\sim$ that is totally ordered by $\leq$. For each equivalence class $X_i \in \mathfrak{T}$, use the axiom of choice to choose a representative framed spacetime $(M_i, g_i, F_i) \in X_i$. As before, one can construct the framed spacetime (M, g, F) by taking the "union" of all of the (M_i, g_i, F_i). It is immediate that $X_i \leq [(M, g, F)]$ for each equivalence class $X_i \in \mathfrak{T}$. So $[(M, g, F)]$ will count as an upper bound for $\mathfrak{T}$ in $\mathfrak{F}(\mathscr{P})/\sim$ if it can be shown that $(M, g) \in \mathscr{P}$. But this easily follows since we know that each (M_i, g_i) is in $\mathscr{P}$, the collection $\{M_i\}$ is an open cover for M, and $\mathscr{P}$ is a strongly local property.

Since $[(M, g, F)]$ is an upper bound for $\mathfrak{T}$ in $\mathfrak{F}(\mathscr{P})/\sim$, one can invoke Zorn's lemma as before to conclude that $\mathscr{P}$ satisfies (Existence). Thus, (NEC), (WEC), (SEC), (DEC), and (Vac) all satisfy (Existence). On the other hand, we know that arbitrary subcollections of these collections will not necessarily satisfy existence. To see this, recall that the singleton collection $\{(M, \eta)\}$ where (M, η) is the $t < 0$ portion of Minkowski spacetime is both vacuum and fails to satisfy (Existence). Thus, with respect to the local spacetime properties considered here, we have another instance here of the general subcollection problem introduced in the previous chapter; it is unclear to what extent these isolated results can be generalized (see Figure 9.6).

9.5 Causal Properties

Next, we consider causal properties. We start with those for which Zorn's lemma can be applied (Manchak, 2017). Let $\mathscr{P} \subset \mathscr{U}$ be the collection $(Chron)$ of spacetimes satisfying the chronology condition. Once again, consider the collection $\mathfrak{F}(\mathscr{P})/\sim$ partially ordered by the relation $\leq$. Let $\mathfrak{T}$ be any totally ordered subset of $\mathfrak{F}(\mathscr{P})/\sim$. For each equivalence class $X_i \in \mathfrak{T}$, choose a representative framed spacetime $(M_i, g_i, F_i) \in X_i$. As before, one can construct the framed spacetime (M, g, F) by taking the "union" of all of the (M_i, g_i, F_i). The equivalence class $[(M, g, F)]$ will be an upper bound for $\mathfrak{T}$ if it can be shown that (M, g) satisfies chronology. Suppose it does not. Let $\lambda \subset M$ be the image of a CTC. As a topological space (with induced topology from M), λ is compact. Let $\lambda_i = \lambda \cap M_i$ for all i. It follows that

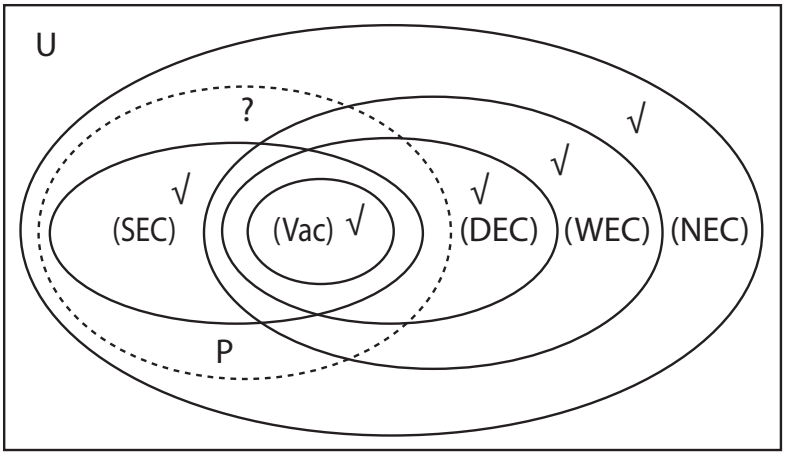

Figure 9.6: The collections (NEC), (WEC), (SEC), (DEC), and (Vac) all satisfy (Existence). But there is no assurance that an arbitrary subcollection $\mathscr{P}$ of any of these collections also satisfies (Existence).

$A = \{\lambda_i\}$ is an open cover for λ. Because λ is a compact, there is a finite subset $A' \subset A$ that is also a cover for λ. The relation $\leq$ can now be used to order the finite elements of A' into a nested sequence of subsets $\lambda_j \subseteq ...\lambda_k$. But this means that $\lambda = \lambda_k$. So (M_k, g_k) has a CTC $\lambda = \lambda_k$: a contradiction. So we may conclude that $(M, g) \in \mathscr{P}$ and thus $[(M, g, F)]$ is an upper bound for $\mathfrak{T}$. Invoking Zorn's lemma in the usual way, it follows that the collection $(Chron)$ satisfies (Existence).

In a completely analogous way, one can show that the collection $(Caus) \subset \mathscr{U}$ of spacetimes satisfying the causality condition also satisfies (Existence). But we emphasize again that the subcollection problem forbids us from concluding that any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Caus) \subset \mathscr{P} \subset (Chron)$ also satisfies (Exisistence). Indeed, it is not difficult to construct such a collection $\mathscr{P}$ which fails to satisfy (Exisistence). Start with Marty's time travel spacetime (M, η) in which $M = S \times \mathbb{R}$ in (t, x) coordinates where $0 \leq t \leq 2\pi$ and $t = 0$ is identified with $t = 2\pi$. Now consider the spacetime (N, η) where N is the $x < 0$ portion of M. In just the same way that the $t < 0$ portion of Minkowski spacetime can (properly) extend itself, so can the spacetime (N, η). It follows that the the collection $\mathscr{P} = (Caus) \cup \{(N, \eta)\}$ will not satisfy (Existence) since there can be no $\mathscr{P}$-maximal extension to (N, η).

It is unknown if the remaining causal properties satisfy (Existence). Con-

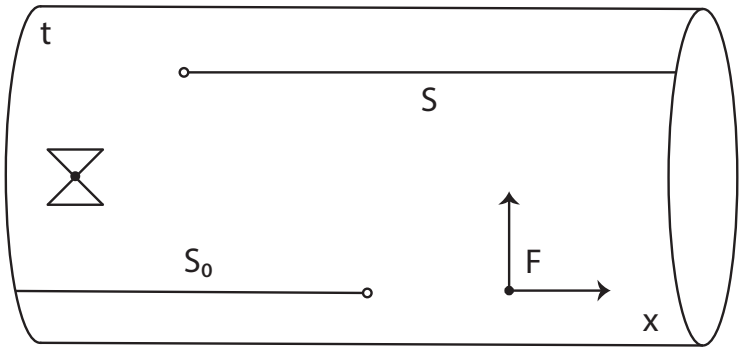

Figure 9.7: The framed spacetime (N, g, F) fails to be strongly causal (and therefore fails to be stably causal).

sider the collections $(Dist), (Str), (Stab), (GH) \subset \mathscr{U}$ of all spacetimes satisfying, respectively, the distinguishing, strong causality, stable causality, and global hyperbolicity conditions. In the cases of $(Dist)$, (Str), and $(Stab)$, one cannot invoke Zorn's lemma. Low (2012) first showed the problem in the case of $(Stab)$. Let's follow his argument for the similar case of (Str).

Let $\mathscr{P} \subset \mathscr{U}$ be the collection (Str) and consider the collection $\mathfrak{F}(\mathscr{P})/\sim$ partially ordered by the relation $\leq$. It is not difficult to construct a totally ordered subset $\mathfrak{T} \subset \mathfrak{F}(\mathscr{P})/\sim$ which has no upper bound. Again, start with Marty's time travel spacetime (M, η) in which $M = S \times \mathbb{R}$ in (t, x) coordinates where $0 \leq t \leq 2\pi$ and $t = 0$ is identified with $t = 2\pi$. For each positive integer i, let (N_i, g_i, F_i) be the framed spacetime constructed by taking (M, η) and removing the slits $S_0 = \{(0, x) : x \leq 1\}$ and $S_i = \{(1, x) : x \geq -1/i\}$ and adding a frame at the point $p = (0, 2)$ consisting of the vectors $[1, 0]$ and $[0, 1]$. Because of the removed slits, each spacetime (N_i, g_i) is strongly causal. But as i increases, the slit S_i becomes smaller as its "edge" approaches the point $(1, 0)$.

For each i, let X_i be the equivalence class $[(N_i, g_i, F_i)]$. Clearly, $\{X_i\} \subset \mathfrak{F}(\mathscr{P})/\sim$ is totally ordered by the relation $\leq$. Now consider the "union" (N, g, F) of all of the (N_i, g_i, F_i). We see that (N, g, F) is just a framed version of Marty's spacetime (M, η) with the slits $S_0 : \{(0, x) : x \leq 1\}$ and $S = \{(1, x) : x \geq 0\}$ removed (see Figure 9.7). We have considered

this example before in our discussion of the causal hierarchy; it was used to show the existence of a spacetime satisfying the distinguishing condition but not the strong causality condition (recall Figure 5.6). So (N, g) fails to be strongly causal and thus $[(N, g, F)]$ fails to be an upper bound for the totally ordered set $\{X_i\}$. Moreover, one can show that any upper bound X for $\{X_i\}$ must be such that $[(N, g, F)] \leq X$. Since any framed extension of (N, g, F) must also fail to be strongly causal, we find that there can be no upper bound for $\{X_i\}$. Because not every totally ordered subset of $\mathfrak{F}(\mathscr{P})/\sim$ has an upper bound, Zorn's lemma cannot be invoked. Thus, it is unclear if $(Stab)$ satisfies (Existence). If so, new proof methods will need to be employed to show this.

A similar situation also arises for the collections $(Dist)$. In an analogous way, one adapts the standard example (M, g) of a causal but not distinguishing spacetime (recall Figure 5.5). For each positive integer i, let (M_i, g_i) be the spacetime constructed by taking (M, g) and removing the slit $S_i = \{(t, 0) : -1/i \leq t \leq 1/i\}$. When properly framed, one creates a sequence of distinguishing framed spacetimes whose "union" (M, g) fails to be distinguishing. The case of (GH) is different in that it not known whether Zorn's lemma can be invoked. Stepping back, we see mostly unsettled questions concerning (Existence) and causal properties (see Figure 9.8).

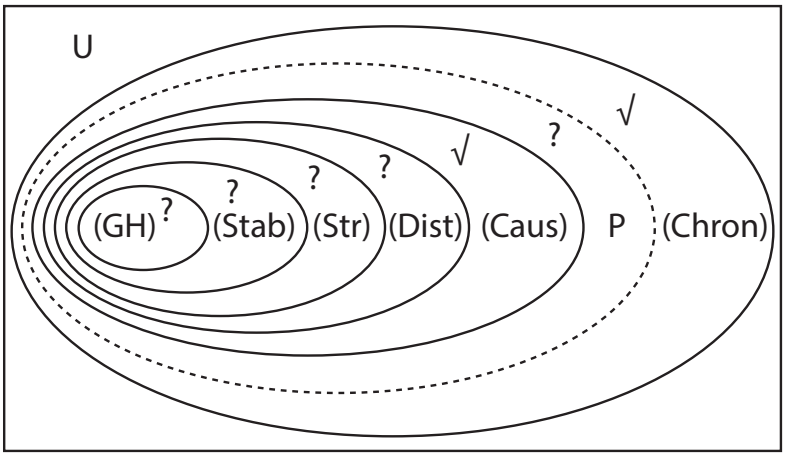

Figure 9.8: The collections $(Chron)$ and $(Caus)$ satisfy (Existence). But there is no assurance that an arbitrary subcollection $\mathscr{P}$ intermediate between these collections also satisfies (Existence). The cases for $(Dist), (Str), (Stab), (GH)$ are all unknown.

9.6 Asymmetry Properties

Next, we explore the (Existence) condition with respect to various asymmetry properties. Let $\mathscr{P} \subset \mathscr{U}$ be the collection (Gir) of all giraffe spacetimes and consider the collection $\mathfrak{F}(\mathscr{P})/\sim$ partially ordered by the relation $\leq$. Start with two-dimensional Minkowski spacetime (M, η) in standard (t, x) coordinates. For each positive integer i, let (N_i, g_i, F_i) be the framed spacetime that results from excising from M the compact triangle region enclosed by the points $(0, 0)$, $(1/i, 0)$, and $(0, 1/i)$ and adding a frame at the point $p = (0, 2)$ consisting of the vectors $[1, 0]$ and $[0, 1]$ (see Figure 9.9).

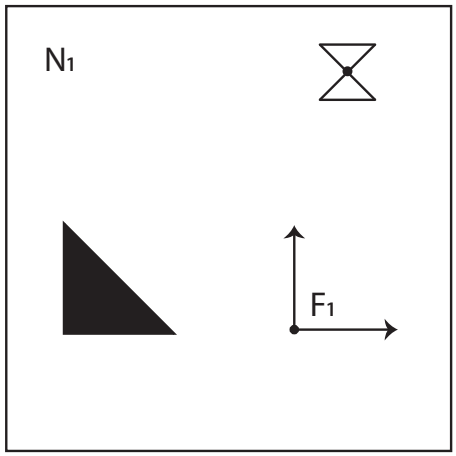

Figure 9.9: The framed spacetime (N_1, g_1, F_1). As as i increases, the removed triangle region becomes smaller as the two vertices $(1/i, 0)$, and $(0, 1/i)$ both approach the third vertex $(0, 0)$.

Because of the "missing" triangle region, each spacetime (N_i, g_i) is giraffe. But as i increases, the removed triangle region becomes smaller as the two vertices $(1/i, 0)$, and $(0, 1/i)$ both approach the third vertex $(0, 0)$. For each i, let X_i be the equivalence class $[(N_i, g_i, F_i)]$. Clearly, $\{X_i\} \subset \mathfrak{F}(\mathscr{P})/\sim$ is totally ordered by the relation $\leq$. Now consider the "union" (N, g, F) of all of the (N_i, g_i, F_i). We see that (N, g, F) is just a framed version of Minkowski spacetime with the origin removed. So (N, g) fails to be giraffe and thus $[(N, g, F)]$ fails to be an upper bound for the totally ordered set $\{X_i\}$. As before, since any upper bound X for $\{X_i\}$ must be such that $[(N, g, F)] \leq X$, we see that any framed spacetime in X must be a framed

extension of (N, g, F). But the only proper framed extension to (N, g, F) is Minkowski spacetime itself which fails to be giraffe. Thus, there can be no upper bound for $\{X_i\}$. Because not every totally ordered subset of $\mathfrak{F}(\mathscr{P})/\sim$ has an upper bound, Zorn's lemma cannot be invoked to show that (Gir) satisfies (Existence). Similar conclusions follow for the collections (PR) and (FP) of all point rigid and fixed point spacetimes respectively. The example given shows this since each (N_i, g_i, F_i) is both point rigid and fixed point and yet the "union" (N, g, F) (as well as any framed extension to it) fails to be both point rigid and fixed point.

What about the (Existence) condition with respect to the collections (LG) and (Her) of, respectively, locally giraffe and Heraclitus spacetimes? The former case is not yet clear. But the latter case is settled: Zorn's lemma can be used to show that (Her) does satisfy the (Existence) condition (Manchak and Barrett, 2024). The argument is simple. Let $\mathscr{P} \subset \mathscr{U}$ be the collection (Her) of all Heraclitus spacetimes and consider the collection $\mathfrak{F}(\mathscr{P})/\sim$ partially ordered by the relation $\leq$. (Although we work with equivalence classes of framed spacetimes as usual here, it turns out that one need not frame Heraclitus spacetimes in order to define a partial order. This follows since one can verify that, unlike the general case, isometric embeddings among Heraclitus spacetimes are unique.)

Let $\mathfrak{T}$ be any totally ordered subset of $\mathfrak{F}(\mathscr{P})/\sim$. For each equivalence class $X_i \in \mathfrak{T}$, choose a representative framed spacetime $(M_i, g_i, F_i) \in X_i$. As before, one can construct the framed spacetime (M, g, F) by taking the "union" of all of the (M_i, g_i, F_i). The equivalence class $[(M, g, F)]$ will be an upper bound for $\mathfrak{T}$ if it can be shown that (M, g, F) is Heraclitus. Suppose not. So there are distinct events $p, q \in M$ with neighborhoods O_p and O_q and an isometry $f : O_p \to O_q$ such that $f(p) = q$. We know there is some framed spacetime (M_k, g_k, F_k) such that $p, q \in M_k$. Let $U_p = O_p \cap M_k$ and let $U_q = f[U_p]$. Although the open set U_p is a subset of M_k by construction, the open set U_q may not be. So consider the open set $V_q = U_q \cap M_k$ and the open set $V_p = f^{-1}[V_q]$. Now it follows that $V_p, V_q \subseteq M_k$. The isometry defined by restricting the domain of f to V_p maps the event p to q which contradicts the Heraclitus property of (M_k, g_k, F_k). So (M, g, F) is Heraclitus. It follows that $[(M, g, F)]$ is an upper bound for $\mathfrak{T}$. Invoking Zorn's lemma in the usual way, we find that the collection (Her) satisfies (Existence).

So we see mostly unsettled questions concerning the (Existence) condition and asymmetry properties. Moreover, we know that Zorn's lemma is of no help for a number asymmetry properties such as the collection (Gir) of all

giraffe spacetimes. The collection (Her) does satisfy (Existence) but we note once again that an instance of the subcollection problem lurks: there is no assurance that an arbitrary subcollection of Heraclitus spacetimes also satisfies (Existence) (see Figure 9.10).

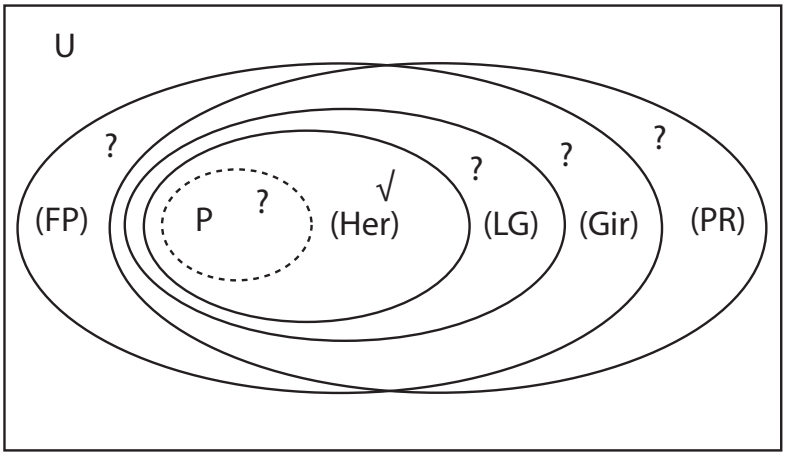

Figure 9.10: The collection (Her) satisfies (Existence). But there is no assurance that an arbitrary subcollection $\mathscr{P} \subset (Her)$ also satisfies (Existence). The cases for $(PR), (FP), (Gir), (LG)$ are all unknown.

9.7 Big Bang Property

Stepping back, it is notable that none of the local, causal, or asymmetry properties we have been considering render (Existence) false. This is good (or, at least not bad) news for those who wish to defend the dogma of spacetime maximality via Leibnizian metaphysics. But it should be emphasized that most of the cases we have examined are unsettled and, of those, most are the unsettled because we know that Zorn's lemma cannot be invoked. We have already seen that (Existence) can be false for some seemingly "artificial" properties (e.g. the singleton collection consisting of the $t < 0$ portion of Minkowski spacetime). We now discuss the possibility of more interesting examples and also the significance of any isolated result concerning the (Existence) condition.

We begin by considering the property of geodesic incompleteness. Given the singularity theorems of Hawking and Penrose (1970), this property seems to be satisfied by some physically reasonable spacetimes. A simple argument shows that the collection $(GI) \subset \mathscr{U}$ of geodesically incomplete spacetimes renders (Existence) true. Consider any geodesically incomplete spacetime (M, g) that is (GI)-extendible. Since $\mathscr{U}$ satisfies (Existence), we know (M, g) has some $\mathscr{U}$-maximal extension (N, h). If $(N, h) \in (GI)$, then we are done. If not, let p be a point in the non-empty region $N - M$ and consider the spacetime $(N - \{p\}, h)$. Either (i) (M, g) and $(N - \{p\}, h)$ are isometric or (ii) the latter spacetime properly extends the former. Clearly $(N - \{p\}, h)$ is geodesically incomplete because of the "missing" point and, in addition, its only extension (up to isometry) is the geodesically complete spacetime (N, h). So $(N - \{p\}, h)$ is (GI)-maximal. Since (M, g) is (GI)-maximal, it follows that(M, g) and $(N - \{p\}, h)$ are not isometric, i.e. (i) cannot hold. So (ii) must hold: $(N - \{p\}, h)$ is a proper extension to (M, g). Because $(N - \{p\}, h)$ is (GI)-maximal and (Existence) must be true for (GI).

Despite the fact that (GI) renders (Existence) true, we know that, because of the subcollection problem, there is no assurance that arbitrary collections $\mathscr{P} \subset (GI)$ also satisfy (Existence). One question of interest is this: Let $(S) \subset (GI)$ be a collection of spacetimes satisfying the assumptions of any one of the singularity theorems. Does such a collection (S) satisfy (Existence)? More work is needed here. Or consider the collection $(BB) \subset (GI)$ of spacetimes with the "big bang" property: every maximal timelike geodesic is incomplete in the past direction. One can show that (BB) does not satisfy the (Existence) condition (Manchak, 2016b).

Start with two-dimensional Minkowski spacetime (M, g) in (t, x) coordinates. For all positive integers i, remove the slits $S_i = \{(-i, x) : x \leq -1/i \text{ or } 1/i \leq x\}$ and let S be the union of all of the S_i. Let $\Omega : M \to \mathbb{R}$ be a conformal factor such that $\Omega = 1$ outside of $D(S)$ but rapidly approaches zero as S is approached along every timelike curve in $D(S)$. The spacetime $(M - S, \Omega^2 g)$ is $\mathscr{U}$-maximal because of the chosen conformal factor. Let p_i be the point $(-i, 0)$ for all positive integers i and let P be the collection of all the p_i. Let (N, h) be the result of taking $(M - S, \Omega^2 g)$ and removing P. The spacetime (N, h) has the big bang property since all maximal timelike geodesics must approach either some slit in S or one of the missing points in P in the past direction (see Figure 9.11).

Any extension to (N, h) will replace some subset of the missing points in P. But in order for an extension to retain the big bang property, an infinite

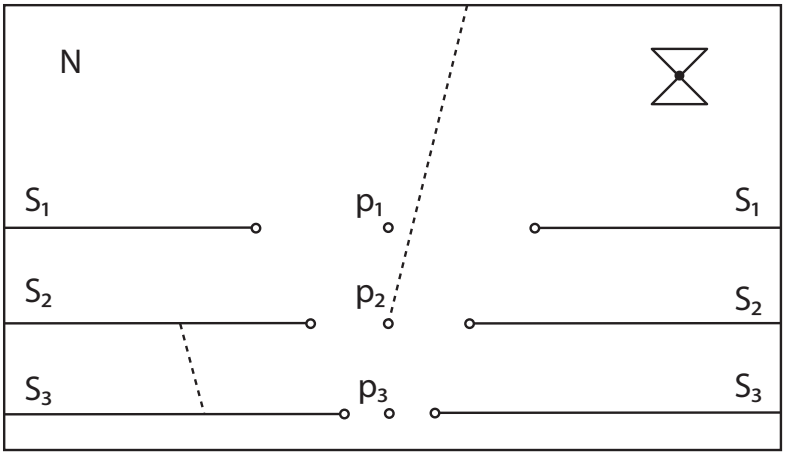

Figure 9.11: The spacetime (N, h) has the big bang property since since all maximal timelike geodesics must approach either some slit S_i or one of the missing points p_i in the past direction. Any big bang extension to (N, h) can also be extended with the big bang property.

number of points in P must remain missing. (If only a finite number of points in P remain missing in the extension, then below the "lowest" such missing point a timelike geodesic along $x = 0$ will be geodesically complete in the past direction.) And since an infinite number of points in P remain missing in any big bang extension to (N, h), one can always extend such an extension even further while still retaining the big bang property by replacing any one of the infinitely many missing points. So there can be no extension to (N, h) which is maximal with respect to the big bang property. So (BB) fails to satisfy (Existence).

Stepping back, one might worry about the significance of the result. After all, the spacetime constructed seems to be outrageously artificial. A natural way to rule out such a mutilated example would be require one of the no-hole spacetime properties we have considered in (e.g. hole-freeness or local maximality). But one must remember that any such property implies $\mathscr{U}$-maximality which implies (BB)-maximality. So invoking a no-hole condition of this kind to secure an (Existence) result is akin to assuming the result itself.

A more promising route is to restrict attention to globally hyperbolic big bang spacetimes. Does $(BB) \cap (GH)$ satisfy (Existence)? Perhaps. But even

so, there remains a question about the significance of such a result given the subcollection problem. Indeed it is not hard to construct a property $\mathscr{P} \subset (BB) \cap (GH)$ which fails to satisfy (Existence). For each positive integer i, let (M_i, g_i) be the $0 < t < i$ region of Minkowski spacetime. If $\mathscr{P} \subset (BB) \cap (GH)$ is the collection of all of the (M_i, g_i), then it renders (Existence) false. Such a property would seem to be artificial but it serves only to demonstrate how difficult it is to get a grip on the big picture with respect to the (Existence) condition; the significance of any isolated result is unclear. Again, we have here an instance of the subcollection problem (see Figure 9.12).

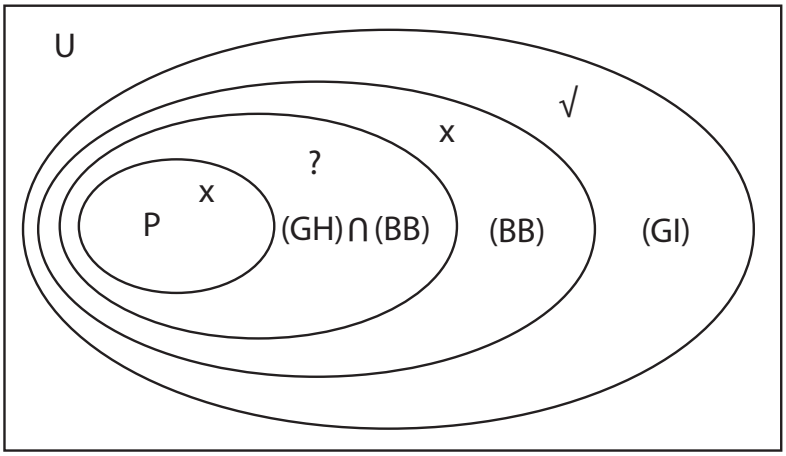

Figure 9.12: An instance of the subcollection problem. The (Existence) condition is true for (GI) but false for one of its subcollections (BB). An even smaller subcollection $(BB) \cap (GH)$ may render (Existence) true. But the condition is false for an even smaller subcollection $\mathscr{P} \subset (BB) \cap (GH)$.

9.8 Conclusion

The metaphysical justification for the spacetime maximality condition seems to depend crucially on the claim that "any space-time can be extended to a space-time that is maximal" (Earman, 1989, p. 161). The foundational result of Geroch (1970b) shows that this claim holds within the standard context: the (Existence) condition is true for the collection $\mathscr{U}$. In this chapter, we

have investigated the extent to which the (Existence) condition is true for various subcollections $\mathscr{P} \subset \mathscr{U}$. We have found that all collections that count as local properties render (Existence) true, e.g. the collection (Vac) of all vacuum solutions. As for causal properties, there are mostly open questions but we have shown that (Existence) is true for the collections $(Chron)$ and $(Caus)$. Moreover, we have seen that if many of these open questions are to be settled, new proof methods will need to be introduced since the usual Zorn's lemma argument is blocked. This is also the situation for almost all asymmetry properties. But (Existence) is true for the collection (Her) of all Heraclitus spacetimes.

We have also highlighted several instances of the sub-collection problem which calls into question the significance of any isolated result. For example, the collection (GI) of all geodesically incomplete spacetimes renders (Existence) true. But (Existence) is false for the subcollection $(BB) \subset (GI)$ of spacetimes with the big bang property. Perhaps restricting attention to the even smaller subcollection $(BB) \cap (GH) \subset (BB)$ of globally hyperbolic big bang models will render (Existence) true again. Even if this is the case, we constructed an even smaller subcollection $\mathscr{P} \subset (BB) \cap (GH)$ for which (Existence) is false. We see just how difficult it is to get a grip on the big picture with respect to the (Existence) condition. Because the significance of any isolated (Existence) result is unclear, the significance of the metaphysical justification for spacetime maximality is also unclear.

Chapter 10

Epistemology

10.1 Introduction

The Leibnizian metaphysical justification for the dogma of spacetime maximality rests on the Geroch (1970b) existence result concerning the possibility space $\mathscr{U}$: every extendible spacetime has a maximal extension. But we have just seen that analogous results are difficult to come by with respect to a number of natural reduced possibility spaces. For the most part, the status of the (Existence) condition remains unsettled relative to the various collections $\mathscr{P} \subset \mathscr{U}$ under consideration. Moreover, the subcollection problem adds another layer of obscurity since any isolated results can only bring limited significance. Given the murky state of the metaphysical justification for spacetime maximality, one wonders about the possibility of empirical justification instead. Perhaps observational data (combined with some form of induction) can somehow allow one to know that spacetime is maximal?

Here, we show that the prospects for an affirmative answer are unsurprisingly dismal. Indeed, an epistemological predicament with respect to spacetime maximality obtains in a wide variety of reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$. Under a modest causality assumption, one finds that for each spacetime (M, g) in $\mathscr{P}$, there exists a non-isometric but "observationally indistinguishable" counterpart spacetime (N, h), also in $\mathscr{P}$, which is not maximal in $\mathscr{P}$. After showing various senses of this type of cosmic underdetermination with respect to spacetime maximality, we highlight one curious exception: the Heraclitus asymmetry property.

In an appendix, we then pivot to investigate the epistemology of spacetime

if the maximality dogma were to hold. We find that maximality with respect to Heraclitus asymmetry allows for a type of uniqueness result: within that context, a pair of spacetimes are observationally indistinguishable if and only if they are isometric. We emphasize a way in which this uniqueness result is quite general in that it is not vulnerable to the subcollection problem. But we also disentangle several different types of underdetermination and show that Heraclitus-maximality is consistent with some of them. We close with a discussion of a type of second-order "meta-maximality" which requires collections of spacetimes to be "as large as they can be" with respect to some second-order property. We show that Zorn's lemma can be used at this higher level to show the existence of various "maximal" collections of spacetimes of this kind.

10.2 Observational Indistinguishability

There are several notions of cosmic underdetermination within general relativity. A few of these will be considered in the appendix at the end of the chapter. For now, we focus a particular definition due to David Malament (1977b) which builds on the ideas of Clark Glymour (1972, 1977). We say a spacetime (M, g) is **observationally indistinguishable** from a spacetime (N, h) if, for each event $p \in M$, there is an event $q \in N$ such that the timelike pasts $I^-(p)$ and $I^-(q)$ are isometric.

To get a grip on this notion, consider a spacetime (M, g) that is observationally indistinguishable from a spacetime (N, h). Because nothing can travel faster than light, any observer at any event $p \in M$ has empirical access only to events in her past light cone, i.e. the region $I^-(p)$. (Here, the timelike past is used instead of the causal past but nothing of consequence turns on the choice. The timelike past is easier to work with since this region is always an open set.) But since $I^-(p)$ and $I^-(q)$ are isometric for some point $q \in N$, the observer at p cannot tell if she is at event p in the spacetime (M, g) or at the event q in the spacetime (N, g). Because this epistemological predicament obtains at every possible event in M, then no observer in (M, g) can be sure that she inhabits the spacetime (M, g) and not (N, h).

To see the definition at work, consider two-dimensional de Sitter spacetime (M, g) (recall Figure 3.6). Here M is the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates where $0 \leq \theta \leq 2\pi$ and $\theta = 0$ is identified with $\theta = 2\pi$. The metric g is defined as follows: at each point $(t, \theta) \in M$ and for any vectors

$v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $g(v, w) = v_t w_t - v_\theta w_\theta \cosh^2(t)$. Now consider an "unrolled" variant of de Sitter spacetime (N, h) where the manifold N is $\mathbb{R}^2$ in (t, x) coordinates and the metric h is defined just like g except in (t, x) rather than (t, θ) coordinates. In each spacetime, the light cones rapidly narrow as the absolute value of t increases. In the case of the spacetime (M, g), this means that for any point $p \in M$, the region $I^-(p)$ must have a θ-width less than 2π. No observer can "see" all the way around the cylinder due to these "observational horizons" (Rindler, 1956). It follows that there will be a corresponding point $q \in N$, such that $I^-(p)$ and $I^-(q)$ are isometric (see Figure 10.1). So (M, g) is observationally indistinguishable from (N, h).

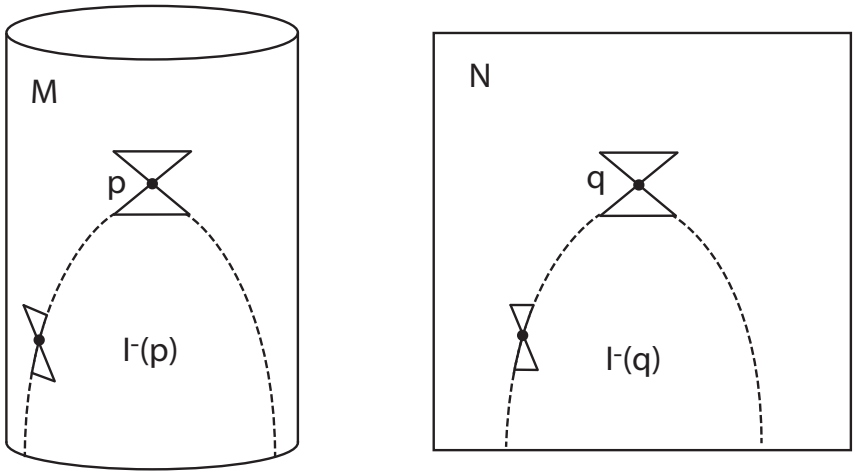

Figure 10.1: The de Sitter spacetime (M, g) and its "unrolled" variant (N, h). For each $p \in M$, there is a $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric.

In the example just given, not only is the spacetime (M, g) observationally indistinguishable from (N, h) but the other direction also holds: (N, h) is observationally indistinguishable from (M, g). In general, however, the situation is not symmetric in this way. Let (M, g) be Minkowski spacetime and let (N, h) be Minkowski spacetime with a point removed. For any event $p \in M$, there is an event $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric – just take q to be any event to the past of the "missing" point. But there are events $r \in N$ such that $I^-(r)$ fails to be isometric to $I^-(p)$ for all points $p \in M$ – just take r to be any event to the future of the missing point (see Figure 10.2). So (M, g) is observationally indistinguishable from (N, h) but

not the other way around: some observers in (N, g) – namely those with the missing point in their timelike past – have the epistemic resources to know that they do not inhabit Minkowski spacetime.

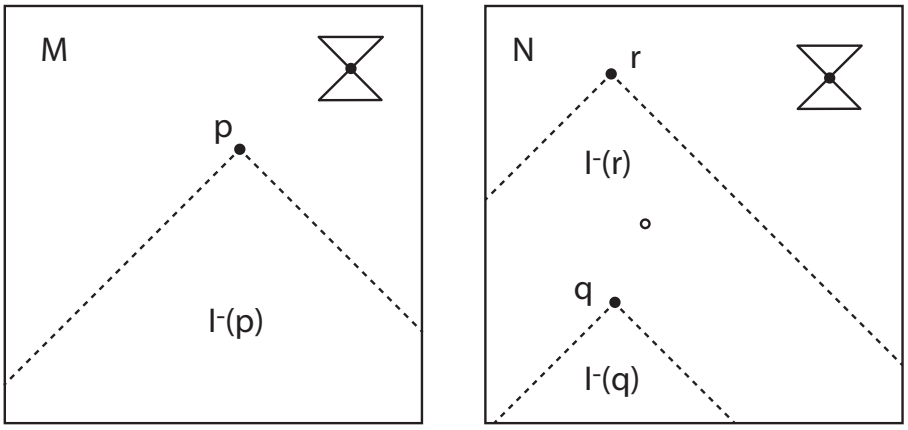

Figure 10.2: For each point $p \in M$, there is a point $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric. But there are some points $r \in N$ such that $I^-(r)$ has no isometric counterpart in M.

The example illustrates additional epistemological problems for observers in Minkowski spacetime. Because (M, g) is $\mathscr{U}$-maximal while (N, h) is not, such observers are not only unable to determine which spacetime they inhabit, they cannot even pin down whether or not their spacetime is maximal under the usual definition. This shows a sense in which collecting empirical data in Minkowski spacetime will never allow observers to "see" the maximality property of their spacetime. The situation may be contrasted with that of some observers in (N, h) who do have the epistemic resources to know that they inhabit a $\mathscr{U}$-extendible spacetime. If (N, h) is observationally indistinguishable from some spacetime (N', h'), then by definition (N', h') must have a region isometric to $I^-(r)$. Because such a region must contain a missing point, (N', h') fails to be $\mathscr{U}$-maximal (it can be properly extended by replacing that missing point). So we find that an observer at $r \in N$ can effectively "see" the $\mathscr{U}$-extendibility property of their spacetime.

We have just seen that observers in some spacetimes (i.e. Minkowski) cannot know that their spacetime is $\mathscr{U}$-maximal while observers in some other spacetimes (i.e. Minkowski with point removed) can know that their

spacetime is $\mathscr{U}$-extendible. Two questions naturally arise. Do there exist spacetimes in which observers can know that their spacetime is $\mathscr{U}$-maximal? Do there exist spacetimes in which observers cannot know that their spacetime is $\mathscr{U}$-extendible? Yes and yes.

For the first question, consider first Marty's time travel spacetime (M, g) in which Minkowski spacetime is "rolled up" in the time direction. Such a spacetime is $\mathscr{U}$-maximal and for any event $p \in M$, we have $I^-(p) = M$. Suppose (M, g) is observationally indistinguishable from a spacetime (N, h). So there is a point $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric. But since $I^-(p) = M$, we see that $I^-(q)$ is isometric to M. If $I^-(q)$ were a proper subset of N, then (M, g) could be properly extended by (N, h) which is impossible since (M, g) is $\mathscr{U}$-maximal. So $I^-(q) = N$ which means that (N, h) is isometric to (M, g). It follows that any observer in (M, g) can know that their spacetime is $\mathscr{U}$-maximal. As for the second question, let (M, g) be the $t < 0$ portion of Minkowski spacetime and let (N, h) be Minkowski spacetime itself. One can easily verify that for any $p \in M$ and $q \in N$ the regions $I^-(p)$ and $I^-(q)$ are isometric. So each spacetime is observationally indistinguishable from the other. Since (M, g) is $\mathscr{U}$-extendible while (N, h) is not, we find that no observer in (M, g) can know that their spacetime is $\mathscr{U}$-extendible.

10.3 Chain Construction

So far, the discussion of observationally indistinguishable spacetimes has operated under the standard background possibility space $\mathscr{U}$. One wonders how the situation changes if reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$ are considered instead. We now turn to a more general investigation. Our focus will be on the possibility of observers knowing that their spacetime is $\mathscr{P}$-maximal relative to various collections $\mathscr{P} \subset \mathscr{U}$. A definition will help to articulate this general question and present the strongest possible results.

Let us say that spacetime (M, g) contains a **god point** if there is an event $p \in M$ such that $I^-(p) = M$. From a god point, an observer can "see" the entirety of spacetime. As we have seen in the example of Marty's time travel spacetime above, if (M, g) is also $\mathscr{U}$-maximal, then there is no possibility of a non-isometric observationally indistinguishable counterpart. So there is a limited sense in which can be spacetime maximality can determined in spacetimes with a god point. But this determination is

necessarily linked with an extreme causal structure which implies a violation of the chronology condition. For this reason, we will now focus on spacetimes without god point. Consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (Observation) There are $\mathscr{P}$-spacetimes without god point that are only observationally indistinguishable from $\mathscr{P}$-spacetimes that are $\mathscr{P}$-maximal.

The condition is formulated so as to be extremely weak. If a collection $\mathscr{P} \subseteq \mathscr{U}$ satisfies (Observation), then merely some – not all – spacetimes (M, g) in the reduced possibility space $\mathscr{P}$ are such that their observers have enough epistemic resources to determine that they inhabit a $\mathscr{P}$-maximal spacetime. As we shall see, even this weak formulation is rarely satisfied by spacetime collections of interest. And if (Observation) is not satisfied by a collection $\mathscr{P}$, then (setting aside spacetimes with god point) every observer in every $\mathscr{P}$-spacetime inherits an cosmic underdetermination problem with respect to $\mathscr{P}$-maximality.

To see why (Observation) is rarely satisfied by a collection $\mathscr{P}$, it proves useful to consider a particular cut and paste construction in which, given a spacetime without god point, can generate a non-isometric observationally indistinguishable counterpart spacetime (Manchak, 2009a). The construction makes precise an informal argument sketch due to Malament (1977b). We start by considering an arbitrary spacetime (M, g). Since any event in M is in the timelike past of some other event in M, it follows that the set $\{I^-(p) : p \in M\}$ is an open cover for M. A general topological result due to Lindelöf is the following: any open cover of a second countable topological space has a countable subcover. Since M is second countable and $\{I^-(p) : p \in M\}$ is an open cover for M, it follows that one can find a countable collection of points $\{p_i\}$ in M such that $\cup\{I^-(p_i)\} = M$. For example, consider Minkowski spacetime in (t, x) coordinates. If $p_i = (i, 0)$ for each positive integer i, then the timelike pasts of all such points cover M (see Figure 10.3).

Now restrict attention to any spacetime (M, g) without god point. Let $\{p_i\}$ be a countable collection of events in M, indexed by the positive integers, such that $\cup\{I^-(p_i)\} = M$. For each p_i, consider two copies of the spacetime (M, g) – call them (N_i^1, h_i^1) and (N_i^2, h_i^2). Since (M, g) fails to have a god

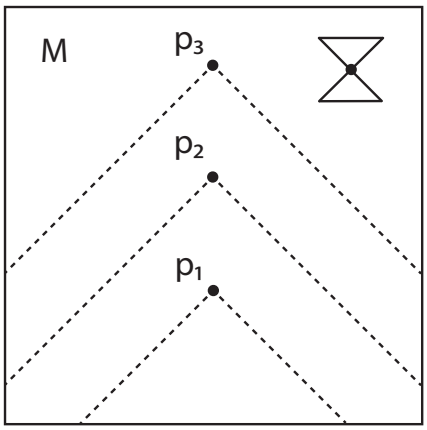

Figure 10.3: The countable collection of points $\{p_i\}$ are such that their time-like pasts cover M.

point, in each of the (N_i^1, h_i^1), one can find an open set O_i outside of $I^-(p_i)$. In each spacetime (N_i^1, h_i^1) for $i > 1$, cut two spacelike slits S_i^- and S_i^+ in the region O_i; in (N_1^1, h_1^1) cut just one slit S_1^+. In each spacetime (N_i^2, h_i^2), cut the slits S_i^+ and S_{i+1}^-. Because of the freedom one has in choosing the slits, it is possible to ensure that S_i^+ and S_{i+1}^- are disjoint in each spacetime (N_i^2, h_i^2). Now, excluding the slit boundary points, identify the top edge of S_i^+ in (N_i^1, h_i^1) with the bottom edge of S_i^+ in (N_i^2, h_i^2). Then identify the top edge of S_{i+1}^- in (N_i^2, h_i^2) with the bottom edge of S_{i+1}^- in (N_{i+1}^1, h_{i+1}^1). Let the resulting "chain" spacetime be called (N, h) (see Figure 10.4).

The mutilated spacetime (N, g) is not isometric to the spacetime (M, g) that we started with. But the chain construction ensures that (M, g) is observationally indistinguishable from the (N, g). This follows since any event $p \in M$ is such that $p \in I^-(p_i)$ for some p_i. So $I^-(p) \subseteq I^-(p_i)$. But by construction, the region $I^-(p_i)$ has an isometric counterpart in the (N_i^1, h_i^1) link of chain (N, h). So there will be a point q in this counterpart region such that $I^-(p)$ and $I^-(q)$ are isometric. It follows that any observer at any event in (M, g) cannot distinguish between that spacetime and (N, h). Now let $q \in N$ be any point in any of the (N_i^2, h_i^2) links. If q is removed from the manifold, the resulting spacetime $(N - \{q\}, h)$ is $\mathscr{U}$-extendible. Since (M, g) is also observationally indistinguishable from this $\mathscr{U}$-extendible spacetime $(N - \{q\}, h)$, we see that (Existence) is false for $\mathscr{U}$.

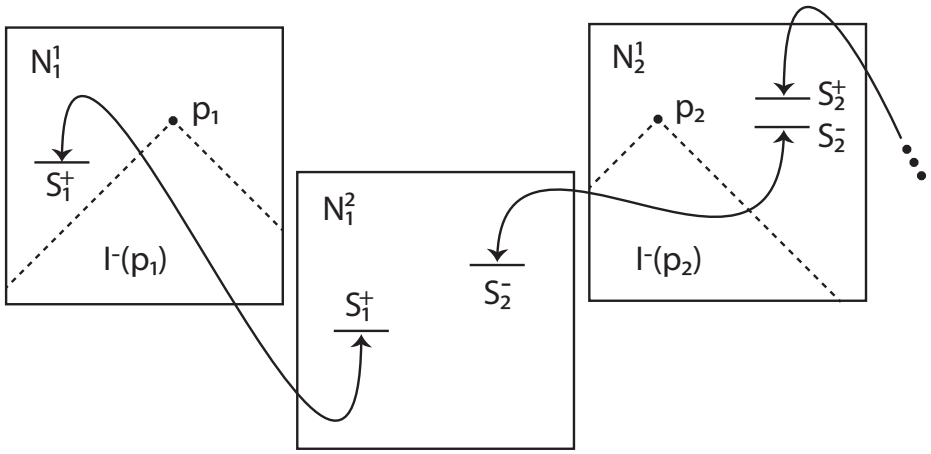

Figure 10.4: The "chain" spacetime (N, h). The top edge of S_i^+ in (N_i^1, h_i^1) is identified with the bottom edge of S_i^+ in (N_i^2, h_i^2) and the top edge of S_{i+1}^- in (N_i^2, h_i^2) is identified with the bottom edge of S_{i+1}^- in (N_{i+1}^1, h_{i+1}^1).

10.4 Local Properties

The chain construction not only ensures that (M, g) is observationally indistinguishable from the non-isometric (N, g), it also preserves an number a properties of (M, g). For example, consider an arbitrary local property $\mathscr{P} \subset \mathscr{U}$ and suppose that (M, g) is a $\mathscr{P}$-spacetime without god point. It is not difficult to verify that the spacetime (N, g) generated via the chain construction is locally isometric to (M, g). So (N, g) is also a $\mathscr{P}$-spacetime. Now, let $q \in N$ be any point in any of the (N_i^2, h_i^2) links. If q is removed from the manifold, the resulting spacetime $(N - \{q\}, h)$ is also locally isometric to (M, g) and therefore also a $\mathscr{P}$-spacetime. Moreover, we find that (M, g) is also observationally indistinguishable from $(N - \{q\}, h)$ and yet the latter spacetime, by construction, is not $\mathscr{P}$-maximal since it can be extended to the $\mathscr{P}$-spacetime (N, g). It follows that any local property $\mathscr{P} \subset \mathscr{U}$ fails to satisfy the (Observation) condition (Manchak, 2011). In particular, all the local properties we have been considering – the collections (NEC), (WEC), (SEC), (DEC), and (Vac) – each render (Observation) false (see Figure 10.5). So because the way the condition (Observation) was formulated (with existential rather than universal quantification), we find that any local property $\mathscr{P}$ inherits a serious underdetermination problem with respect to

spacetime maximality; no observer any $\mathscr{P}$-spacetime without god point has the epistemic resources to determine that their spacetime is $\mathscr{P}$-maximal.

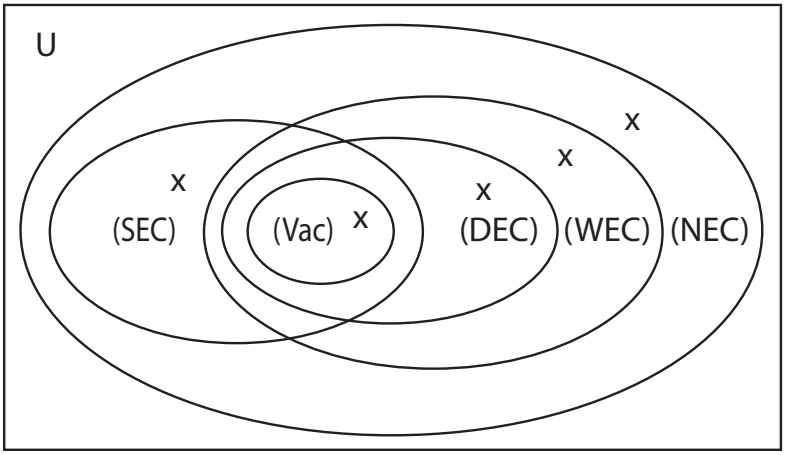

Figure 10.5: The collections (NEC), (WEC), (SEC), (DEC), and (Vac) all fail to satisfy (Observation).

To better understand the significance of the result, we emphasize that cosmologists often employ a type of induction on local spacetime properties whereby "the normal physical laws we determine in our spacetime vicinity are applicable at all other spacetime points" (Ellis, 1975, p. 246). Given that we have yet to empirically observe a violation of, say, the weak energy condition in our local vicinity, the cosmologist extrapolates the finding and assumes the condition holds globally. The result just presented shows that even under this type of local induction, determining that one's spacetime is maximal is not possible. Even so, the spacetimes generated by the chain construction would still seem to be "irrelevant monstrosities by the standards of working cosmologists" (Belot, 2023, p. 147). Perhaps paring down the background possibility space using various global properties can break the cosmic underdetermination? We now highlight that similar results also obtain for almost every global property of interest as well.

10.5 Causal Properties

We start with the chronology condition. Let (M, g) be an arbitrary chronological spacetime which therefore fails to have a god point. Let (N, h) be an observationally indistinguishable counterpart spacetime generated by the chain construction outlined above. It is not difficult to verify that (N, h) too must be chronological. Consider any event $p \in N$ and any future-directed timelike curve λ from p. We know p is located in some link in the chain. We know λ cannot stay within this link and be closed. This follows since each link is a copy of (M, g) which contains no CTCs. But we also know that the timelike curve λ cannot leave the link with p and be closed. Because the slits are spacelike, any future-directed timelike curve leaving one link can never return to it. So λ cannot return to p and therefore (N, h) is chronological. One can remove a point $q \in N$ in any point in any of the (N_i^2, h_i^2) links to produce a spacetime $(N - \{q\}, h)$ which is also chronological and such that (M, g) is observationally indistinguishable from $(N - \{q\}, h)$. But by construction, this latter spacetime is not maximal with respect to the chronology property since it can be extended to (N, g). It follows that the collection $(Chron)$ of chronological spacetimes renders the (Observation) condition false. Analogous arguments can be carried out for the collections $(Caus)$, $(Dist)$, and (Str) of spacetime satisfying, respectively, the causality, distinguishing, and strong causality conditions.

One can also verify that the collection $(Stab)$ of stably causal spacetimes also renders (Observation) false. Let (M, g) be a stably causal spacetime (which therefore fails to have a god point). Let (N, h) be an observationally indistinguishable counterpart spacetime generated by the chain construction outlined above. Since (M, g) is stably causal, it admits a global time function on M. But since each link of the spacetime (N, g) is just a copy of (M, g), one can use the global global time function on M to define a global time function on N in the natural way. So (N, g) is stably causal. As before, one can then remove a point q to produce a stably causal spacetime $(N - \{q\}, h)$ which shows the (Observation) condition is false for $(Stab)$.

Finally, we note that the situation for the collection (GH) of globally hyperbolic spacetimes is open. Given a globally hyperbolic spacetime (M, g), the observationally indistinguishable counterpart spacetime (N, h) generated from the chain construction fails to globally hyperbolic. But this does not necessarily mean that (Observation) is true for (GH). Consider Minkowski spacetime (M, g) for example. It is observationally indistinguishable from

the $t < 0$ portion of Minkowski spacetime (N, h). Moreover, (N, h) is both globally hyperbolic and also fails to be maximal with respect to this property (it can be extended to Minkowski spacetime, for example). But it is not yet clear that one can find such an observationally indistinguishable counterpart for an arbitrary globally hyperbolic spacetime. Stepping back, we see that (Observation) is false for all but one causal property with the case of global hyperbolically still open (see Figure 10.6).

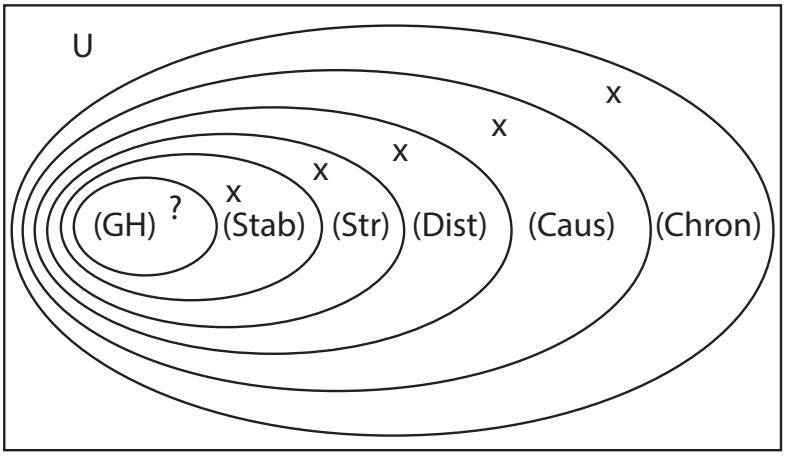

Figure 10.6: The collections $(Chron)$, $(Caus)$, $(Dist)$, (Str), and $(Stab)$ all fail to satisfy (Observation) with the case of (GH) open.

10.6 Asymmetry Properties

Now we consider the asymmetry properties. Let (M, g) be any spacetime without god point and let (N, h) be a spacetime generated by the chain construction. The first link (N_1^1, h_1^1) of this spacetime is unique in that it is linked to only one other link; it has only one slit cut. This ensures that any isometry of (N, h) must map this first link to itself. And since the second link (N_2^1, h_2^1) is the only link attached to this first link, then by continuity considerations, it too must be mapped to itself. And so on. The fact that any isometry must map each link to itself does not necessarily entail that any isometry of (N, h) must be the identity map. For example, consider Minkowski spacetime (M, g) in (t, x) coordinates. If the chain spacetime

(N, h) is constructed such that all spacelike slits are cut so as to be symmetric about the the $x = 0$ line, there will be a non-trivial isometry of (N, h) in which each link is mapped to itself but reflected about $x = 0$ line. In such a case, the spacetime (N, h) will therefore fail to be giraffe. But one can rule out such non-trivial isometries by removing from each link in (N, h) a compact set shaped like a giraffe. If the removal takes place outside the region $I^-(p_i)$ in each of the (N_i^1, h_i^1) links, then (M, g) will be observationally indistinguishable from the resulting giraffe spacetime (see Figure 10.7). Moreover, one can extend this spacetime while maintaining the giraffe property by appropriately (i.e. asymmetrically) replacing a portion of one of the "missing" giraffe regions. Stepping back, this argument shows something quite general: If $\mathscr{P}$ is such that $(Gir) \subseteq \mathscr{P}$, then (Observation) fails to be satisfied by $\mathscr{P}$. So in addition to (Gir) itself, it follows that (Observation) is false for the collections (PR) and (FP) of spacetimes that satisfy, respectively, the point rigid and fixed point asymmetry conditions.

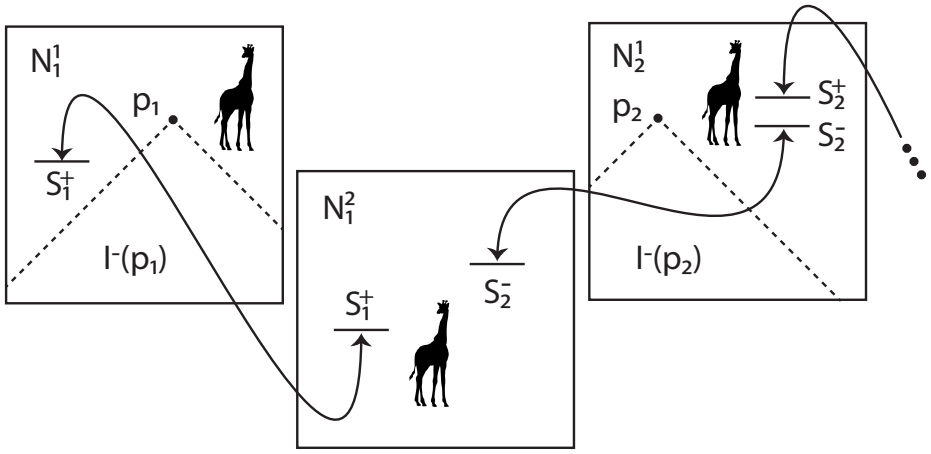

Figure 10.7: Any spacetime without god point is observationally indistinguishable from some (non-isometric) giraffe spacetime.

Given any spacetime without god point, the modified chain construction just considered generates an observationally indistinguishable counterpart spacetime with only trivial isometries. So this construction has the effect of introducing global asymmetries. But it also has the effect of introducing some local symmetries. Since each link in chain spacetime is constructed from a copy of the original spacetime, there will be innumerable local isometries

between any pair of links. So the chain construction necessarily produces a spacetime that fails to have the Heraclitus asymmetry property.

Of course, this does not necessarily mean that the (Observation) condition is true for the collection (Her). Perhaps some other construction could be employed. But that possibility is closed off if one considers spacetimes that are maximal with respect to the Heraclitus property. Indeed this follows from a more general result (Manchak and Barrett, 2024): if (M, g) and (N, h) are Heraclitus spacetimes and (M, g) is observationally indistinguishable from (N, h), then either the two spacetimes are isometric or (N, h) is a proper extension of (M, g).

At its heart, this result follows because isometric embeddings among Heraclitus spacetimes are unique and, for each $p \in M$, the region $I^-(p)$ counts as a Heraclitus spacetime it its own right. Because there is a unique way to embed each region $I^-(p)$ into (N, h), one finds that (M, g) itself can be isometrically embedded into (N, h). (Intuitively, the radical asymmetry of each of the $I^-(p)$ requires that one can only smoothly "glue" all of these regions together in one way, i.e. the way that results in a spacetime isometric to (M, g) itself.) And if this isometric embedding from (M, g) to (N, h) is proper, then (N, h) is a proper extension of (M, g); otherwise, the spacetimes are isometric.

As simple corollary to this result, we see that if (M, g) is maximal with respect to the Heraclitus property and it is observationally indistinguishable from some Heraclitus spacetime (N, g), then (N, g) must be isometric to (M, g) and so must also be Heraclitus-maximal. Thus, if one takes (M, g) to be one that fails to have a god point, we see that the collection (Her) satisfies the (Observation) condition. In the reduced possibility space (Her), there are some spacetimes in which observers do have enough epistemic resources to determine that they inhabit a (Her)-maximal spacetime.

Taking stock, we see that (Observation) is false for any collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P}$. This includes the collections (PR), (FP), and (Gir) itself. In contrast, (Observation) is a true for the collection (Her). Finally, we note that the case concerning the collection (LG) of locally giraffe spacetimes is not yet clear (see Figure 10.8).

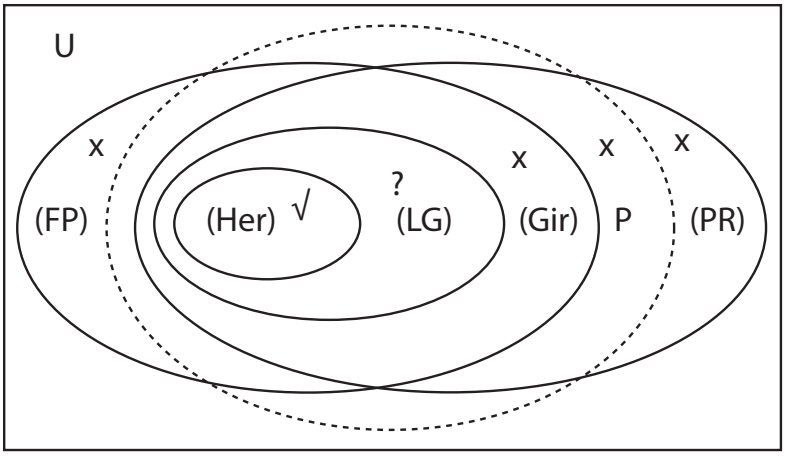

Figure 10.8: (Observation) is false for any collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P}$. This includes the collections (PR), (FP), and (Gir) itself. (Observation) is true for (Her) while the case for (LG) is open.

10.7 Conclusion

Stepping back, we see (Observation) is false relative to every spacetime property $\mathscr{P} \subseteq \mathscr{U}$ under consideration save two: there is an open question with respect to the collection (GH) of all globally hyperbolic spacetimes and (Observation) is true for the collection (Her) of all Heraclitus spacetimes. But we also emphasize again that we are working with an extremely weak sense of (Observation) which, if satisfied by a collection $\mathscr{P}$, only shows that some – not all – spacetimes in $\mathscr{P}$ avoid a cosmic determination problem with respect to $\mathscr{P}$-maximality. We used this formulation so that the negative results as strong as possible. Had we instead formulated the condition (Observation) in the universally quantified way, then even the collections of (GH) and (Her) would fail to satisfy it. Thus, there is a robust sense in which there is a cosmic underdetermination problem with respect to $\mathscr{P}$-maximality for every collection $\mathscr{P} \subseteq \mathscr{U}$ under consideration.

10.8 Appendix: Heraclitus Maximality

We now explore a curious tension between the Heraclitus asymmetry property, the dogma of spacetime maximality, and (some forms of) cosmic underdetermination. The tension will be captured by a no-go result that is quite general in the sense that it is not vulnerable to the subcollection problem. A corollary to the result is this: if the dogma of spacetime maximality were to hold with respect to the Heraclitus asymmetry property, then spacetimes are observationally indistinguishable if and only if they are isometric.

We go on to disentangle several different types of underdetermination and show that Heraclitus-maximality is consistent with some of them. We then consider a type of second-order "meta-maximality" which requires collections of spacetimes to be "as large as they can be" with respect to some second-order property of collections of spacetimes. We apply Zorn's lemma at this higher level to show the existence of various "maximal" collections of spacetimes.

We begin with a statement that captures a relativized form of the dogma of spacetime maximality. Consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (Maximality) Each $\mathscr{P}$-spacetime is $\mathscr{P}$-maximal.

It is immediate that any collection $\mathscr{P} \subseteq \mathscr{U}$ that satisfies (Maximality) must also satisfy (Existence). And as long as the collection $\mathscr{P}$ contains a spacetime without god point, if it satisfies (Maximality), then it must also satisfy (Observation). So we see a clear sense in which our exploration of the (Existence) and (Observation) conditions with respect to various reduced possibility spaces helps to assess the status of relativized dogma; if these conditions fail for a given possibility space, then effectively the relativized dogma fails. We now pivot to consider what follows if the dogma of spacetime maximality is simply assumed to hold for various collections. As we have noted, this is standard practice in the case of the collection $\mathscr{U}$. We will see that it is also fruitful to consider the strength and character of the (Maximality) condition with respect to other collections as well.

Using the notion of observationally indistinguishable spacetimes, one can keep track of universal and existential forms of general underdetermi-

nation (as opposed to underdetermination with respect to the maximality condition). Consider the following (second-order) conditions on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (∀ Underdetermination) Each $\mathscr{P}$-spacetime is observationally indistinguishable from some other non-isometric $\mathscr{P}$-spacetime.

> (∃ Underdetermination) Some $\mathscr{P}$-spacetime is observationally indistinguishable from some other non-isometric $\mathscr{P}$-spacetime.

We see that any non-empty collection $\mathscr{P}$ that satisfies (∀ Underdetermination) must satisfy (∃ Underdetermination). How are these underdetermination conditions related to the (Observation) condition we have been considering? Suppose that (Observation) is satisfied by some collection $\mathscr{P}$. Then there is some $\mathscr{P}$-spacetime (M, g) without god point that is only observationally indistinguishable from $\mathscr{P}$-spacetimes that are $\mathscr{P}$-maximal. But just because observers in (M, g) can determine that their spacetime is $\mathscr{P}$-maximal does not mean that they are in a position to determine that they inhabit a spacetime isometric to (M, g). For example, consider the collection $\mathscr{P}$ consisting of two spacetimes: de Sitter and its "unrolled" variant (recall Figure 10.1). This collection $\mathscr{P}$ satisfies (Observation) and yet it also satisfies (∀ Underdetermination) and hence (∃ Underdetermination). Observers cannot determine which of the two spacetimes they inhabit and yet they can determine that they inhabit a $\mathscr{P}$-maximal spacetime since all $\mathscr{P}$-spacetimes are $\mathscr{P}$-maximal.

We can now state the no-go result (Manchak and Barrett, 2024): any non-empty subcollection $\mathscr{P} \subseteq (Her)$ of Heraclitus spacetimes cannot satisfy both (Maximality) and (∃ Underdetermination). As a corollary it follows that any such subcollection cannot satisfy both (Maximality) and (∀ Underdetermination). So we see that both types of cosmic underdetermination vanish if the relativized dogma holds in any reduced possibility space of Heraclitus spacetime. Let's consider one concrete instantiation of the result.

One can verify that the collection $\mathscr{P} \subseteq (Her)$ of all Heraclitus-maximal spacetimes must satisfy (Maximality). In other words, each Heraclitus-maximal spacetime is also a (Heraclitus-maximal)-maximal spacetime. To

see this, suppose $\mathscr{P}$ fails to satisfy (Maximality). Then there is a $\mathscr{P}$-spacetime (M, g) that can be properly extended by a $\mathscr{P}$-spacetime (N, h). Since both spacetimes are in $\mathscr{P}$, they must both be in (Her). Since (N, h) properly extends (M, g) and both spacetimes are Heraclitus, (M, g) cannot be Heraclitus-maximal. So (M, g) is not in $\mathscr{P}$: a contradiction. So we now have one instantiation of the result: the collection of all Heraclitus-maximal spacetimes (which exists) renders both ($\exists$ Underdetermination) and ($\exists$ Underdetermination) false.

We now emphasize that the tension captured here is quite general in the sense that underdetermination is inconsistent with respect to any subcollection of (Her) that satisfies (Maximality). We have just seen that collection of Heraclitus-maximal spacetimes is one such collection. One can easily verify that any of its subcollections also satisfies (Maximality). But there are many others as well.

Consider any Heraclitus spacetime (M, g) and let N be any connected proper subset of M. The spacetime (N, h) is therefore Heraclitus-extendible. But if $\mathscr{P} = \{(N, g)\}$ then (because Heraclitus spacetimes cannot extend themselves) we see that (N, g) is $\mathscr{P}$-maximal. So $\mathscr{P}$ is a subcollection of Heraclitus spacetimes that satisfies (Maximality). Thus, the no-go result applies which requires that $\mathscr{P}$ renders both underdetermination conditions false. Of course, it is not surprising that there is no underdetermination problem given that $\mathscr{P}$ is a singleton collection. But we will see in a moment a sense in which any collection that satisfies (Maximality) – including any singleton collection – can be "extended" so as to be "as large as it can be" with respect to the (Maximality) property. For now, the point is simply to highlight that the no-go result is not vulnerable to the subcollection problem.

So far, we have considered types of cosmic underdetermination connected to a particular definition of observationally indistinguishable spacetimes. This notion seems to be a "straightforward rendering of conditions under which observers could not determine the spatio-temporal structure of the universe" (Malament, 1977b, p. 69). But although the conditions specified in the definition seem to be sufficient for underdetermination, they do not seem to be necessary. If a spacetime (M, g) has no non-isomorphic observationally indistinguishable counterpart spacetime, then there is a sense in which the collective information that all individuals in the spacetime have together is sufficient to determine which world they inhabit, but that determination may be beyond the observational reach of any one individual in the spacetime. In other words, there is no spacetime event in which all

individuals can bring their collective information together. This suggests a weaker notion of cosmic underdetermination that better captures that epistemological situation of the individual observer (Butterfield, 2014, p. 60). Consider the following (second-order) conditions on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> ($\forall$ Underdetermination*) For each $\mathscr{P}$-spacetime (M, g) and each event $p \in M$, there is a non-isometric $\mathscr{P}$-spacetime (N, h) with event $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric.

> ($\exists$ Underdetermination*) For some $\mathscr{P}$-spacetime (M, g) and each event $p \in M$, there is a non-isometric $\mathscr{P}$-spacetime (N, h) with event $q \in N$ such that $I^-(p)$ and $I^-(q)$ are isometric.

It is immediate that any non-empty collection $\mathscr{P}$ that satisfies ($\forall$ Underdetermination*) must satisfy ($\exists$ Underdetermination*). We also see that each starred condition is implied by its non-starred analogue. Now we show that the implications do not run in the other direction. Indeed, we will show something much stronger: both starred conditions are consistent with collections of Heraclitus spacetimes that satisfy (Maximality). In other words, there is no analogous no-go result that captures a tension between the Heraclitus asymmetry property, the dogma of spacetime maximality, and the starred (weaker) forms of cosmic underdetermination.

To see how this can be, consider again de Sitter spacetime. As we have seen, each individual observer in this spacetime has observational horizons in the sense that she will never "see" some regions of spacetime. The de Sitter model is highly symmetric on a global, matter-averaged scale. But one can imagine a spacetimes with observational horizons similar to de Sitter which are also Heraclitus at a fine-grained scale. One can show that there are collections of such spacetimes that also satisfy the (Maximality) condition (Manchak and Barrett, 2024). We now turn to an example collection.

We have already constructed a Heraclitus spacetime that is conformal to a portion of Minkowski spacetime (recall Figure 7.8). So the causal structure of such a spacetime is, at least locally, the same as that of Minkowski

spacetime. This fact allows us to construct a spacetime from a small portion of Minkowski spacetime and then infer the existence of a Heraclitus spacetime with an identical causal structure. Consider two-dimensional Minkowski spacetime (M, η) in (t, x) coordinates. For each $i = 1, 2, 3, 4$, let S_i be the $t > 0$ region of the timelike past of the point $(2, 2i)$. Let (N, η) be the spacetime consisting of the union of the four S_i regions (see Figure 10.9). It is important to note that for any $i = 1, ..., 4$ any event $p \in S_i$ is such that $I^-(p) \subset S_i$. So there is a sense in which the this spacetime has observational horizons similar to those of de Sitter spacetime.

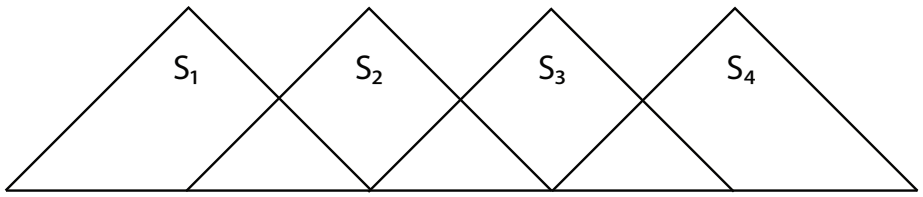

Figure 10.9: The spacetime (N, η) is the union of the regions $S_1, ..., S_4$ in two-dimensional Minkowski spacetime.

Now let (N, h) be a conformally related Heraclitus spacetime with the same causal structure and, for each for $i = 1, 2, 3, 4$, let p_i be the event $(1, 2i)$. For each $i = 1, 2, 3, 4$, let the spacetime (N_i, h_i) be a portion of (N, h) where two of the p_i points have been removed: $N_1 = N - \{p_1, p_2\}$, $N_2 = N - \{p_3, p_4\}$, $N_3 = N - \{p_2, p_3\}$, and $N_4 = N - \{p_1, p_4\}$ (see Figure 10.10). We let $\mathscr{P} = \{(N_1, h_1), ..., (N_4, h_4)\}$ which is, by construction, a subcollection of Heraclitus spacetimes. One can now verify that each $\mathscr{P}$-spacetime is $\mathscr{P}$-maximal. Consider (N_1, h_1) for example. This spacetime cannot be isometrically embedded into (N_2, h_2) or (N_3, h_3) since the event $p_3 \in N_1$ has been removed in those spacetimes. Similarly, (N_1, h_1) cannot be isometrically embedded into (N_4, h_4) since the event $p_4 \in N_1$ has been removed in that spacetime. So (N_1, h_1) is $\mathscr{P}$-maximal. The cases for the other

spacetimes are handled in an analogous way. So $\mathscr{P}$ satisfies the (Maximality) condition.

We also see that the ($\forall$ Underdetermination*) condition is satisfied as well. Consider again the spacetime (N_1, h_1) and any event $p \in N_1$. Since p is in some S_i, we know that $I^-(p) \subset S_i$. If p is in the S_1 region of N_1, then $I^-(p)$ has an isometric counterpart in the S_1 region of N_4; if p is in the S_2 region of N_1, then $I^-(p)$ has an isometric counterpart in the S_2 region of N_3; if p is in the S_3 region of N_1, then $I^-(p)$ has an isometric counterpart in the S_3 region of N_4; if p is in the S_4 region of N_1, then $I^-(p)$ has an isometric counterpart in the S_4 region of N_3. The cases for the other spacetimes in $\mathscr{P}$ are handled in an analogous way. So the ($\forall$ Underdetermination*) condition is satisfied by $\mathscr{P}$. It follows that the ($\exists$ Underdetermination*) is also satisfied by $\mathscr{P}$. Thus, we see that that the general tension between the Heraclitus asymmetry property, the dogma of spacetime maximality, and the unstarred underdetermination conditions does not carry over to the weaker starred variants of the latter conditions.

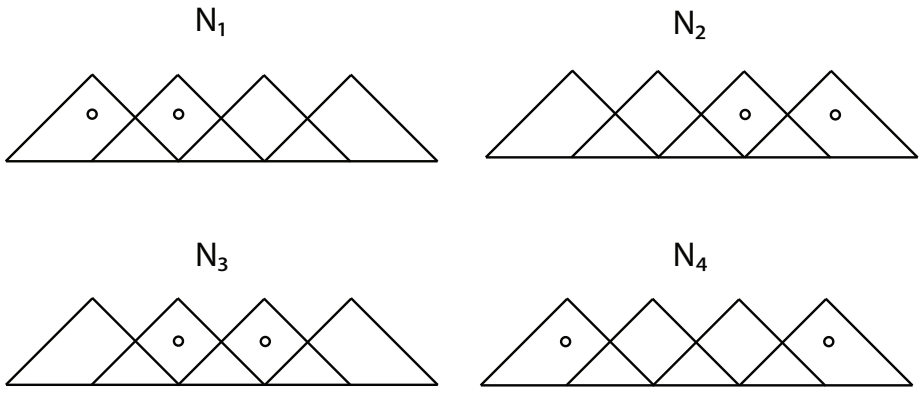

Figure 10.10: The regions $N_1, ..., N_4$ constructed by removing points from N.

One may worry that the collection $\mathscr{P}$ just constructed is artificially small in some sense. But we now introduce a general procedure for "extending" collections via Zorn's lemma so as to be "maximal" with respect to second-order properties of interest. In the present case, we will build a maximal subcollection of Heraclitus spacetimes that satisfies both the (Maximality) and the ($\forall$ Underdetermination*) conditions. We start by considering

the power collection $\mathcal{P}(\mathscr{U})$ – the collection of all subcollections of $\mathscr{U}$. Any second-order property of collections – e.g. the property of satisfying (Maximality) – corresponds to a collection $\mathcal{R} \subseteq \mathcal{P}(\mathscr{U})$ of collections of spacetimes with the property. It is easy to see that the relation $\subseteq$ counts as a partial order on the $\mathcal{P}(\mathscr{U})$. For any collections $\mathscr{P}, \mathscr{Q} \in \mathcal{P}(\mathscr{U})$, we say that $\mathscr{Q}$ is a (not necessarily proper) **extension** of $\mathscr{P}$ if $\mathscr{P} \subseteq \mathscr{Q}$. For any second-order property $\mathcal{R} \subseteq \mathcal{P}(\mathscr{U})$, we say that a collection $\mathscr{P} \in \mathcal{R}$ is **$\mathcal{R}$-maximal** if it has no proper extension in $\mathcal{R}$. Consider the following condition on a collection $\mathcal{R} \subseteq \mathcal{P}(\mathscr{U})$ of collections of spacetimes.

> [Existence] Any collection of spacetimes in $\mathcal{R}$ has an $\mathcal{R}$-maximal extension.

Here, we have used square brackets to distinguish the third order [Existence] condition on collections of collections of spacetimes from the second-order analogue (Existence) condition on collections of spacetimes. A first-order condition concerns spacetime itself, e.g. the chronology condition. First-order conditions give rise to a natural collection of spacetimes: the collection of all spacetimes satisfying the condition. We have often used parentheses and italics when naming such collections, e.g. the collection $(Chron) \subset \mathscr{U}$ of all chronology satisfying spacetimes. In an analogous way, a second-order condition like (Existence) gives rise to a natural collection of collections of spacetimes: the collection of all collections satisfying the second-order condition. We will, in the analogous way, use square brackets and italics when naming such collections of collections. For example, let $[Ex] \subset \mathcal{P}(\mathscr{U})$ be the collection of all collections of spacetimes satisfying the second-order (Existence) condition. In a similar way, define $[Eq], [Ob], [Max] \subset \mathcal{P}(\mathscr{U})$ to be, respectively, the collection of all collections that satisfy the conditions (Equivalence), (Observation), and (Maximality).

It is immediate that $[Eq]$ and $[Ex]$ satisfy the third order [Existence] condition. Consider $[Ex]$ for example. Any collection $\mathscr{P} \in [Ex]$ will have an $[Ex]$-maximal extension – namely the collection $\mathscr{U}$ of all spacetimes. This follows since $\mathscr{U}$ satisfies (Existence) (recall the foundational (Geroch, 1970b) result) and $\mathscr{U}$ counts as an extension to every collection in $\mathscr{P} \in [Ex]$ since $\mathscr{P} \subseteq \mathscr{U}$. So $[Ex]$ satisfies the third order [Existence] condition. The case for $[Eq]$ is similar: since $\mathscr{U}$ trivially satisfies (Equivalence) and this collection

counts as an $[Eq]$-maximal extension to every collection in $\mathscr{P} \in [Eq]$.

A more interesting situation arises when we consider $[Ob]$ and $[Max]$. Consider the latter. Since $\mathscr{U}$ fails to satisfy (Maximality), it cannot be a $[Max]$-maximal extension to a given member of $[Max]$. What about the union $\bigcup[Max] \subset \mathscr{U}$? It is also too big to satisfy (Maximality) and hence cannot be an $[Max]$-maximal extension to a given member of $[Max]$. To see this, consider any Heraclitus spacetime (M, g) and any proper, connected, open set $O \subset M$. The singleton collections $\{(M, g)\}$ and $\{(O, g)\}$ both satisfy (Maximality) since the spacetimes (M, g) and (O, g) are Heraclitus and therefore cannot extend themselves. So these singleton collections are members of $[Max]$ and thus $(M, g), (O, g) \in \bigcup[Max]$. But since (O, g) can be extended by (M, g), we see that (O, g) cannot be $\bigcup[Max]$-maximal. So $\bigcup[Max]$ does not satisfy (Maximality).

To get around the problem, Zorn's lemma can be invoked. Consider any collection $\mathcal{T} \subset [Max]$ totally ordered by the $\subseteq$ relation. We see that the union $\bigcup \mathcal{T} \subset \mathscr{U}$ is an upper bound for $\mathcal{T}$. This union also satisfies (Maximality). If it didn't, there would be spacetimes $(M, g), (N, h) \in \bigcup \mathcal{T}$ such that one is a proper extension of the other. But this cannot be since it means that (M, g) and (N, h) can both be found in some collection in $\mathcal{T}$ and all such collections satisfy (Maximality). From Zorn's lemma, it now follows that any collection in $[Max]$ has an $[Max]$-maximal extension, i.e. it follows that $[Max]$ satisfies [Existence].

This type of higher level Zorn's lemma argument can be used to show that $[Ob]$ also satisfies [Existence]. Another similar example is the following. Let $\mathcal{R} \subseteq \mathcal{P}(\mathscr{U})$ be the collection of all subcollections of Heraclitus spacetimes that satisfy both the (Maximality) condition and the ($\forall$ Underdetermination*) condition. So the collection $\mathscr{P} = \{(N_1, h_1), ..., (N_4, ..., h_4)\}$ constructed above (recall Figure 10.10) counts as one member of $\mathcal{R}$. It is not difficult to verify that Zorn's lemma can be used to show that $\mathcal{R}$ satisfies [Existence]. A similar result holds where the ($\exists$ Underdetermination*) is considered instead of ($\forall$ Underdetermination*). Thus, we have assurance that the collection $\mathscr{P}$ consisting of four elements can be extended to a collection that is "as large as it can be" with respect to the second-order properties of interest. More generally, we find here a useful tool for the construction of maximal possibility spaces (satisfying certain desiderata) given an initial collection of spacetimes.

Chapter 11

Stability

11.1 Introduction

Within the context of general relativity, the "stability" of various spacetime properties has been one important focus of study. Indeed, it has been argued by Hawking and Ellis (1973, p. 197) that "in order to be physically significant, a property of space-time ought to have some form of stability, that is to say, it should be a property of 'nearby' space-times." Geroch (1971a, p. 70) also claims: "It is a general feature of the description of physical systems by mathematics that only conclusions which are stable, in an appropriate sense, are of physical interest." He traces this idea all the way back to Pierre Duhem (1906).

In this chapter, we will investigate the question of whether the property of spacetime maximality (defined relative to various reduced possibility spaces) is stable in an appropriate sense. We will ultimately find that (unlike many other properties of interest) virtually nothing is known about the (in)stability properties of spacetime maximality. This is partly due to technical barriers in defining "stability" in a general way. We consider a workaround definition of this notion and discuss some of its shortcomings. After reviewing some foundational results concerning the stability of causal properties (Hawking, 1969; Geroch, 1970a), we then turn to an investigation of various no-hole spacetime properties to get a better grip on the situation for spacetime maximality. We present a pair of surprising results that show that the properties of geodesic completeness and local maximality are unstable when the background possibility space is $\mathscr{U}$ (Williams, 1984; Beem et al., 1996). This suggests that

perhaps there are similar instability results lurking for spacetime maximality as well. Along these lines, we close by showing that spacetime maximality is unstable for some collections of globally hyperbolic vacuum solutions (Manchak, 2023). This highlights a vexing subcollection problem with respect to the stability of spacetime maximality that is guaranteed to persist no matter what isolated stability results can secured in the future.

11.2 What Is Stability?

In order to make the notion of "stability" precise, ideally one would like to put a suitable topology on the collection $\mathscr{U}$ of all spacetimes. One could then say that a spacetime property $\mathscr{P} \subseteq \mathscr{U}$ is "stable" if each spacetime $(M, g) \in \mathscr{P}$ has a neighborhood $\mathscr{O} \subset \mathscr{U}$ (i.e. a collection of "nearby" spacetimes) such that $\mathscr{O} \subset \mathscr{P}$ (see Figure 11.1). Another way to put it: a property $\mathscr{P} \subseteq \mathscr{U}$ is "stable" if $\mathscr{P}$ is an open set in $\mathscr{U}$.

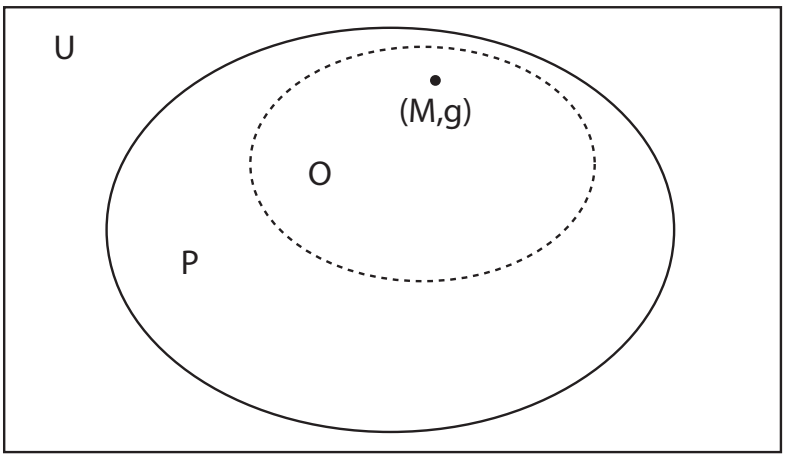

Figure 11.1: A property $\mathscr{P} \subseteq \mathscr{U}$ is stable (relative to a suitable topology on $\mathscr{U}$) if each spacetime $(M, g) \in \mathscr{P}$ has a neighborhood $\mathscr{O} \subseteq \mathscr{P}$.

Unfortunately, no suitable topology on the collection $\mathscr{U}$ has yet been found. Essentially, the problem comes down to technical difficulties in capturing the notion "nearby" spacetimes when the underlying spacetime manifolds are non-diffeomorphic. As a workaround, various topologies have been defined on the collection $\mathscr{L}(M)$ of all spacetimes with the same underlying

manifold M. By far, the most commonly used topologies of this kind are the "C^k fine" topologies which are sometimes called the "C^k open" or "C^k Whitney" topologies (Lerner, 1973; Hawking and Ellis, 1973).

Fix a manifold M that admits a Lorentzian metric and consider the collection $\mathscr{L}(M)$ of spacetimes with underlying manifold M. Let (M, g) and (M, g') be spacetimes. At each point $p \in M$, one would like to calculate a "distance" between the metrics g and g' (the $k = 0$ case) and their kth "derivatives" (for $k = 1, 2, 3, ...$). But this idea runs into a basic problem (Geroch, 1971a, pp. 70-71):

> We have an intuitive idea of what it means to say that "two metrics arc close," but to make this idea precise turns out to be surprisingly difficult. For example, it would not do simply to compare the components of the metrics in some coordinate system, for the difference between the components can, in general, be made either arbitrarily large or arbitrarily small by an appropriate (or inappropriate) choice of coordinates.

One can find a way around this problem by choosing a Riemannian metric h on M to serve as a standard of comparison. This allows one to define a natural **distance** $d(g, g', h, k)$ between the kth derivatives of the metrics g and g' relative to the Riemannian metric h and its associated derivative operator. (Because we have not built up the machinery to express derivative operators and arbitrary tensors in a rigorous way, we cannot give a precise formulation of this distance function here. See Fletcher (2016) for a nice presentation of the details.)

Of course, one does not want the topology on the collection $\mathscr{L}(M)$ to depend on the choice of Riemannian metric h. For any integer $k \geq 0$ consider basis elements of the form $B_k(g, h, \epsilon) = \{(M, g') : \sup_M[d(g, g', h, 0)] < \epsilon, ..., \sup_M[d(g, g', h, k)] < \epsilon\}$ where g and h range over all Lorentzian and Riemannian metrics on M respectively and ϵ ranges over all positive real numbers. A basis element can be thought of as an "open ball" of radius ϵ centered at the point (M, g) relative to h and k. One can then define the C^k **fine topology** on $\mathscr{L}(M)$ to be the collection of all subcollections of $\mathscr{L}(M)$ which can be expressed as a union of the basis elements $B_k(g, h, \epsilon)$.

Any C^k fine topology on $\mathscr{L}(M)$ induces a natural subspace topology on $\mathscr{L}(M) \cap \mathscr{P}$ for any collection $\mathscr{P} \subset \mathscr{U}$. This allows us to formulate a notion of stability relative to a choice of background possibility space. For

all non-empty $\mathscr{Q} \subseteq \mathscr{P} \subseteq \mathscr{U}$, we say the property $\mathscr{Q}$ is C^k **stable** relative to the collection $\mathscr{P}$ if each $\mathscr{Q}$-spacetime (M, g) has a C^k fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M) \cap \mathscr{P}$ such that $\mathscr{O} \subseteq \mathscr{Q}$. It is not difficult to verify that, for all $\mathscr{Q} \subseteq \mathscr{P} \subseteq \mathscr{U}$, if property $\mathscr{Q}$ is C^k stable relative to the collection $\mathscr{P}$, then $\mathscr{Q}$ is C^l stable relative to $\mathscr{P}$ for all $l \geq k$. This follows since, for any $\mathscr{Q}$-spacetime, the collection of C^k fine neighborhoods in $\mathscr{L}(M) \cap \mathscr{P}$ is a subcollection of the C^l fine neighborhoods in $\mathscr{L}(M) \cap \mathscr{P}$.

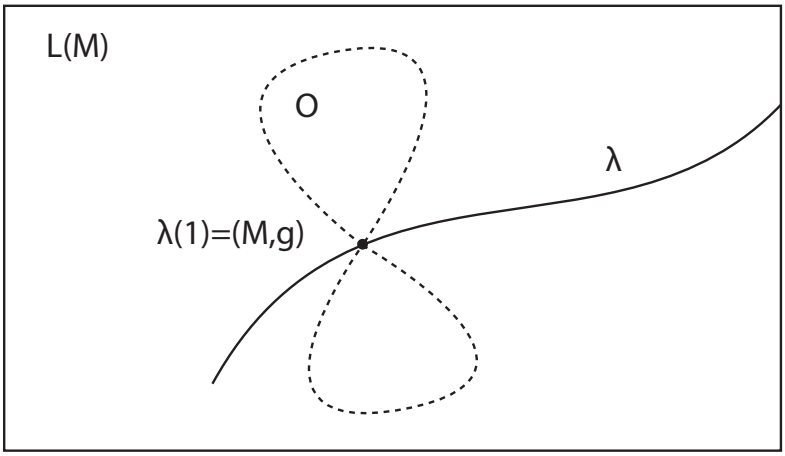

Figure 11.2: There is a C^0 open neighborhood $\mathscr{O} \subset \mathscr{L}(M)$ of the spacetime $\lambda(1) = (M, g)$ that contains no other spacetime on the curve λ.

It has been argued that the C^k fine topologies on $\mathscr{L}(M)$ are too "fine" is the sense that they permit too many open collections. Consider the following example (Geroch, 1971a, p. 71). Let (M, g) be any spacetime for which M is non-compact. Let $\lambda : \mathbb{R}^+ \to \mathscr{L}(M)$ be the curve defined such that $\lambda(s) = (M, sg)$ for all $s \in \mathbb{R}^+$. One finds that this curve is discontinuous at every point when $\mathscr{L}(M)$ carries any of the C^k fine topologies. To see why this must be, consider the point $\lambda(1) = (M, g)$ for example. Since M is non-compact, there is a Riemannian metric h on M whose components approach infinity sufficiently rapidly so that $\sup_M [d(\lambda(1), \lambda(s), h, 0)] = \infty$ for all $s \neq 1$. Thus, any C^0 open neighborhood $\mathscr{O} \subset \mathscr{L}(M)$ of $\lambda(1) = (M, g)$ that is built using such a Riemannian metric h will be too small to contain any of the spacetimes $\lambda(s)$ for $s \neq 1$ (see Figure 11.2). We now see that continuity fails at $s = 1$ since the preimage $\lambda^{-1}[\mathscr{O}]$ of such an open neighborhood is just $\{1\}$ which is not an open set in $\mathbb{R}^+$. Since the image of the curve λ is a

subcollection of $\mathscr{L}(M)$, one can also consider the subspace topologies on the image of λ induced from the C^k fine topologies on $\mathscr{L}(M)$. One finds that all such induced topologies are discrete (i.e. maximally fine). This result is especially troubling when one considers a spacetime (M, g) that is isometric to (M, sg) for all $s \in \mathbb{R}^+$, e.g. when (M, g) is Minkowski spacetime.

The example calls into question the physical significance of the C^k fine topologies. They seem to permit too many open collections to suitably capture, once and for all, what it means for one spacetime to be "close" to another. But as we will see, exploring these topologies can nonetheless shed light on the stability properties of various spacetime properties of interest. Indeed, the fact that the C^k fine topologies have too many open sets means that any instability results that follow are quite significant. This is because the finer the topology, the harder it is to secure such instability results.

11.3 Stable Causality

Early investigations of the C^k fine topologies concerned causal properties. One can show that a spacetime (M, g) admits a global time function if and only if it has a C^0 fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ such that $\mathscr{O} \subset (Chron)$, i.e. each spacetime in the neighborhood $\mathscr{O}$ contains no CTCs (Hawking, 1969; Hawking and Ellis, 1973). So we see that the collection $(Stab) \subset \mathscr{U}$ of all stably causal spacetimes (defined as those with a global time function) is appropriately named. As one would expect, it is also the case that the collection $(Stab)$ is C^k stable relative to the collection $\mathscr{U}$ for all $k \geq 0$ (Beem et al., 1996). On the other hand, we see that any chronological spacetime (M, g) that does not admit a global time function will fail to have a C^0 fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ such that $\mathscr{O} \subseteq (Chron)$ (see Figure 11.3).

It follows that the collection $(Chron)$ of all chronological spacetimes is not C^0 stable relative to the collection $\mathscr{U}$. Analogous arguments give rise to similar instability results for the collections $(Caus), (Dist), (Str) \subset \mathscr{U}$ of all spacetimes satisfying, respectively, the causality, distinguishing, and strong causality conditions. This follows easily since $(Caus), (Dist), (Str) \subset (Chron)$. In contrast, one finds that the collection (GH) of globally hyperbolic spacetimes is C^k stable relative to the collection $\mathscr{U}$ for all $k \geq 0$ (Geroch, 1970a; Beem et al., 1996).

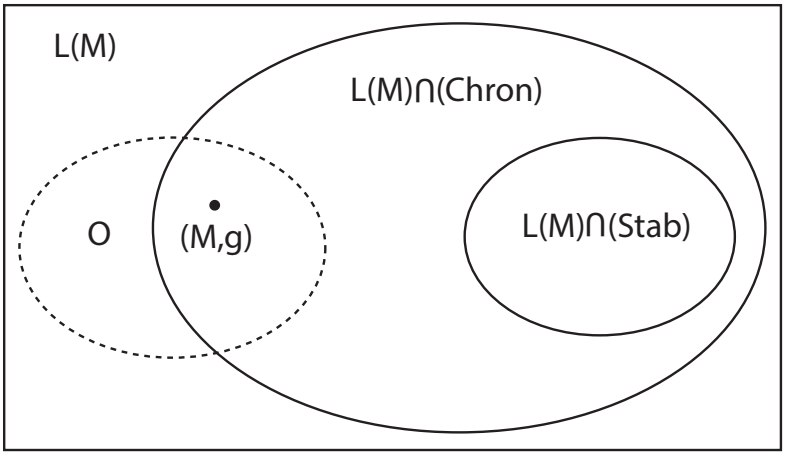

Figure 11.3: Any chronological spacetime (M, g) that is not stably causal will be such that each of its C^0 fine neighborhoods $\mathscr{O} \subseteq \mathscr{L}(M)$ contain spacetimes that violate chronology.

11.4 Subcollection Stability

It is important to appreciate that the stability results just given concerning the collections $(Stab)$ and (GH) are quite robust in the sense that they are not vulnerable to the subcollection problem. Although they are formulated relative to the collection $\mathscr{U}$, analogous results also hold relative to any reduced possibility space $\mathscr{P} \subseteq \mathscr{U}$. Consider, for example, the collection $(Vac) \subset \mathscr{U}$ of vacuum solutions. Because (GH) is C^k stable relative to $\mathscr{U}$, we know that each globally hyperbolic spacetime (M, g) has a C^k fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ such that $\mathscr{O} \subseteq (GH)$. But this means that if (M, g) is also a vacuum solution, then it has a C^k fine neighborhood $\mathscr{N} = \mathscr{O} \cap (Vac)$ in the subspace topology of $\mathscr{L}(M) \cap (Vac)$ such that $\mathscr{N} \subseteq (GH)$ (see Figure 11.4). So the property $(GH) \cap (Vac)$ is C^k stable for all $k \geq 0$ relative to the collection (Vac) of vacuum solutions. More generally, we have the following simple statement concerning the stability of subcollections (Manchak 2023): For all $\mathscr{Q} \subseteq \mathscr{P} \subseteq \mathscr{U}$ and for all integers $k \geq 0$, if the property $\mathscr{Q}$ is C^k stable relative to $\mathscr{P}$, then for any subcollection $\mathscr{R} \subseteq \mathscr{P}$, the property $\mathscr{Q} \cap \mathscr{R}$ is C^k stable relative to $\mathscr{R}$.

Naturally, one wonders about the stability of spacetime maximality relative to various reduced possibility spaces. Consider the following

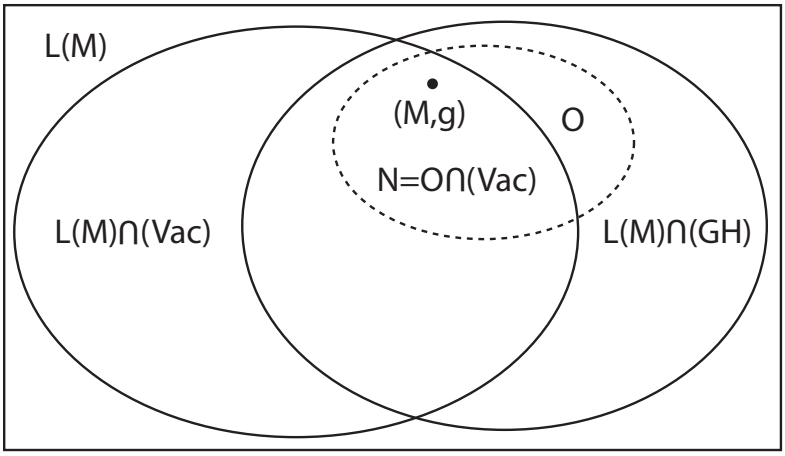

Figure 11.4: Because the globally hyperbolic vacuum solution (M, g) has a C^k fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ such that $\mathscr{O} \subseteq (GH)$, we know that it also has a C^k fine neighborhood $\mathscr{N} = \mathscr{O} \cap (Vac)$ in the subspace topology of $\mathscr{L}(M) \cap (Vac)$ such that $\mathscr{N} \subseteq (GH)$.

(second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (Stability) The collection of $\mathscr{P}$-maximal spacetimes is C^k stable for some $k \geq 0$ relative to $\mathscr{P}$.

Note that the condition is formulated so as to be relatively easy to satisfy. A property need only be C^k stable for some $k \geq 0$. As we have seen, even the C^0 fine topology is already quite fine making C^0 stability results a relatively low bar. So requiring only C^k stability for some $k \geq 0$ or other lowers the bar even further. Not only is it unknown whether the standard collection of spacetimes $\mathscr{U}$ satisfies (Stability), there are also question marks associated with each of the (sixteen) local, causal, and asymmetry properties we have been concerning. We also emphasize that even if (Stability) winds up being true for some collection $\mathscr{P} \subseteq \mathscr{U}$, there is no assurance that it will also be true for arbitrary subcollections $\mathscr{R} \subset \mathscr{P}$. Thus, the situation differs significantly from the case concerning the stability of the arbitrary subcollections of $(Stab)$ and (GH) mentioned above. Such results followed from a more

general statement concerning the stability of subcollections of stable properties. In light of this general statement, how can the (Stability) condition be vulnerable to the subcollection problem?

Suppose that the collection $\mathscr{U}$ satisfies (Stability). So the collection (Max) of $\mathscr{U}$-maximal spacetimes is C^k stable for some $k \geq 0$ relative to $\mathscr{U}$. Now let $\mathscr{P} \subset \mathscr{U}$ be an arbitrary subcollection. It follows from the general statement concerning the stability of subcollections that $\mathscr{P} \cap (Max)$ is C^k stable for some $k \geq 0$ relative to $\mathscr{P}$. But this does not mean that (Stability) is true for $\mathscr{P}$ since $\mathscr{P} \cap (Max)$ and the collection of $\mathscr{P}$-maximal spacetimes are, in general, different collections. We see that it is the modal character of spacetime maximality that makes any stability results involving this property vulnerable to the subcollection problem. We shall return to this point a bit later on with an explicit example. For now, we will we will investigate the stability of various no-hole spacetime properties to get a better grip on the situation for spacetime maximality.

11.5 Geodesic Completeness

We start by considering the collection $(GC) \subset \mathscr{U}$ of geodesically complete spacetimes. In the first edition of *Global Lorentzian Geometry* by Beem and Ehrlich (1981), it was claimed that the collection (GC) was C^k stable relative to $\mathscr{U}$ for all $k \geq 2$. Then came a dramatic turn of events as later recounted by Ehrlich (2006, p. 14):

> That is how matters stood until 1985, when a copy of P. Williams' Ph.D. thesis, "Completeness and its stability on manifolds with connection," was received unexpectedly in the mail. This article revealed that there was a significant gap in the previous arguments for the C^k-stability of geodesic completeness in $\mathrm{Lor}(M)$, and that in fact neither geodesic completeness nor geodesic incompleteness was C^k-stable...From a certain perspective, a good deal of research in global space-time geometry during the next decade can be viewed as trying to understand the more complicated geometry of the space of geodesics, once it was realized that [the claim] failed to be valid.

Williams (1984) was able to show that the collection (CG) is not C^k stable relative to $\mathscr{U}$ for all $k \geq 0$. The result is quite surprising given how fine the

C^k fine topologies are. It also contrasts with the Riemannian case where geodesic completeness is C^k stable for all $k \leq 0$ (Beem and Ehrlich, 1987). Williams proved his remarkable result with a relatively simple example. Start with the manifold $M = \mathbb{R} \times S$ in (x, θ) coordinates. For each integer $i \geq 1$, consider the spacetime (M, g_i) where the metric g_i is defined as follows: at each point $(x, \theta) \in M$ and for any vectors $v = [v_x, v_\theta]$ and $w = [w_x, w_\theta]$ at the point, let $g_i(v, w) = v_x w_\theta + v_\theta w_x + f_i(x) v_\theta w_\theta$ where $f_i(x) = \sin(x)/i$.

One can get a grip on these spacetimes by considering the behavior of light. Since there is no $v_x w_x$ term, it is immediate that the vector $[1, 0]$ is null at every point. So one family of null geodesics run along the cylinder and are complete. We also see that the vector $[0, 1]$ is spacelike for $\pi < x < 0$, null for $x = -\pi, 0, \pi$ and timelike for $0 < x < \pi$. This means that from $x = -\pi$ to $x = \pi$, the light cones close up, open up, and then close up again with this pattern repeating with period 2π. In this, way, the spacetime is essentially a periodic version of Misner spacetime (recall Figure 6.9). Indeed, just as in that spacetime, there are incomplete null geodesics that spiral around the cylinder that approach but never reach $x = 0$ (see Figure 11.5).

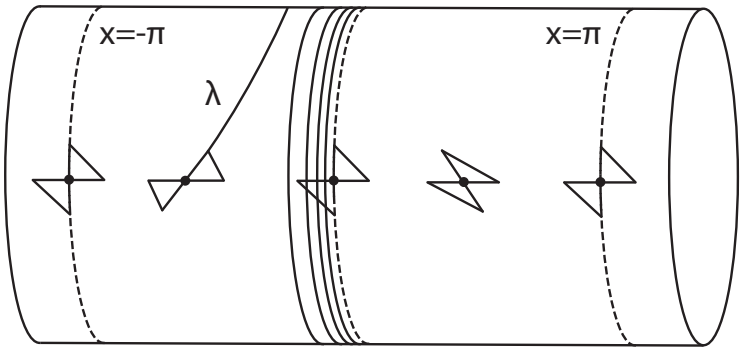

Figure 11.5: From $x = -\pi$ to $x = \pi$, the light cones close up, open up, and then close up again with this pattern repeating with period 2π. There is an incomplete null geodesic λ that spirals around the cylinder approaching but never reaching $x = 0$.

The spacetimes (M, g_i) are such that the third term in g_i (the one with the function $f_i(x)$) approaches zero as the integer i approaches infinity. So

the effect of the opening and closing of the light cones becomes less and less pronounced as i increases. Let (M, g) be the spacetime in which the effect completely vanishes, i.e. let g be the metric on M in which the third term of g_i is dropped completely. This is just two-dimensional Minkwoski spacetime that has been "rolled up" along one null direction. We now "compactify" the spacetimes (M, g) and (M, g_i). Let N be the manifold M where each point (x, θ) is identified with the point $(x + 2\pi n, \theta)$ for all integers n.

Because of the periodicity of the metrics g and g_i on M, one can define the spacetimes (N, g) and (N, g_i) in the natural way. By construction, (N, g) is geodesically complete while (N, g_i) is geodesically incomplete for all $i \geq 1$. Let h be any Riemannian metric on N. Due to the compactness of N, the quantity $\sup_N[d, (g, g_i, h, 0)]$ is bounded and approaches zero as $i \to \infty$. It follows that every C^0 fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(N)$ of (N, g) will contain, for sufficiently large i, the geodesically incomplete spacetime (N, g_i) (see Figure 11.6). Thus, such a neighborhood $\mathscr{O}$ cannot be a subcollection of (GC). In other words, the collection (CG) is not C^0 stable relative to $\mathscr{U}$. Extending the argument shows that analogous results hold for all $k \geq 0$.

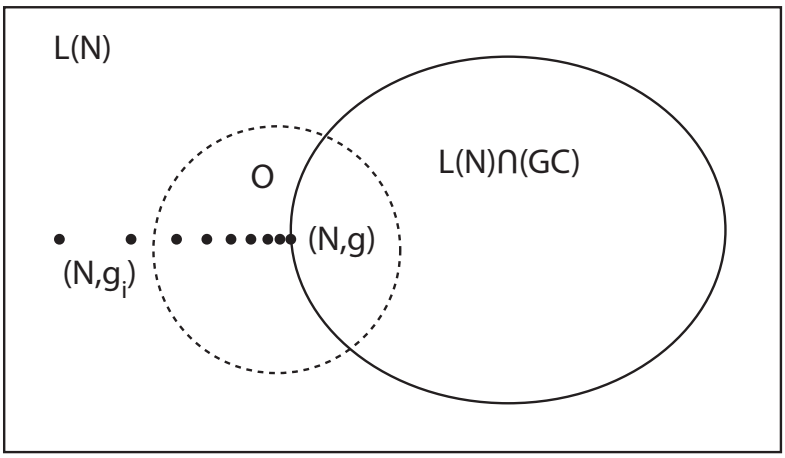

Figure 11.6: Every C^0 fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(N)$ of (N, g) will contain geodesically incomplete spacetimes (N, g_i) for sufficiently large i.

The example given by Williams (1984) sparked a search for sufficient conditions to ensure the C^k stability of geodesic completeness. Immediately, one sees that local conditions are of no help in this regard. The spacetimes (N, g) and (N, g_i) are flat and therefore vacuum solutions. So it follows that

the collection $(Vac) \cap (GC)$ is not C^k stable relative to (Vac) for all $k \geq 0$. Excluding two-dimensional spacetimes will not help secure a stability result either since the example can be generalized to any higher dimension. Looking to causal conditions turns out to be more promising. First, we note that the Williams (1984) instability result can be generalized somewhat (Beem and Ehrlich, 1987, p. 328). Let (M, g) be any geodesically complete spacetime with a closed null geodesic. Then for each $k \geq 0$, every C^k fine neighborhood of (M, g) will contain a geodesically incomplete spacetime. So there is a sense in which the instability of geodesic completeness is a necessary feature of any collection containing certain causally misbehaved spacetimes.

This general result suggests that perhaps restricting attention to some causal property of interest be sufficient for the C^k stability of geodesic completeness. This turns out to be almost true (Beem et al., 1996, p. 270). If (M, g) is a geodesically complete, globally hyperbolic spacetime, there will be a C^1 neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ of (M, g) such that each spacetime in the neighborhood $\mathscr{O}$ is timelike and null geodesically complete. This result is typical of much the work carried out since the Williams (1984) example. Even under extremely strong assumptions (e.g. global hyperbolically) the stability statement is not quite what one would hope for (e.g. the C^0 case is not settled and the C^1 neighborhood $\mathscr{O}$ may nonetheless contain models with incomplete spacelike geodesics).

11.6 Local Maximality

One might think that that instability results concerning geodesic completeness are not representative of the stability properties of other no-hole conditions such as spacetime maximality. After all, geodesic completeness it is an incredibly strong condition. And given the singularity theorems, there is reason to believe that geodesically complete spacetimes are relatively rare among "physically reasonable" reduced possibility spaces. But it turns out that some weaker no-hole conditions also fail to be stable. Indeed, one can use the Williams (1984) example to show that the collection (LM) of locally maximal spacetimes also fails to be C^k stable relative to $\mathscr{U}$ for any $k \geq 0$. The argument follows the one given for a similar result with a slightly different definition of local maximality (Manchak, 2018).

We begin by recalling that $(GC) \subset (LM)$. So we know that the geodesically complete spacetime (N, g) is locally maximal. But for each integer

$i \geq 1$, one can show that the spacetime (N, g_i) is locally extendible. To see this, consider the $-\pi < x < 0$ region $O \subset N$ of the spacetime (N, g_i). As we have seen, this open set will contain an incomplete null geodesic λ that winds around the manifold N ever approaching but never reaching $x = 0$ (recall Figure 11.5). We can extend this null geodesic in the other direction as well so as to be maximal. One finds that it also winds around the manifold ever approaching but never reaching $x = -\pi$. Now let (N, h_i) be the "reverse twisted" isometric variant of the spacetime (N, g_i) in which the light cones tip in the opposite direction. One can find an isometry $f : O \to O$ from (O, g_i) to (O, h_i) in which the null geodesic $f \circ \lambda$ is "unwound" (see Figure 11.7). This geodesic can be then be extended across both $x = -\pi$ and $x = 0$ in the spacetime (N, h_i) showing that (N, g_i) is locally extendible. So for each $k \geq 0$, any C^k neighborhood of the locally maximal (N, g) will contain some locally extendible spacetime (N, g_i) for sufficiently large i. Thus, the collection (LM) fails to be C^k stable relative to $\mathscr{U}$ for all $k \geq 0$.

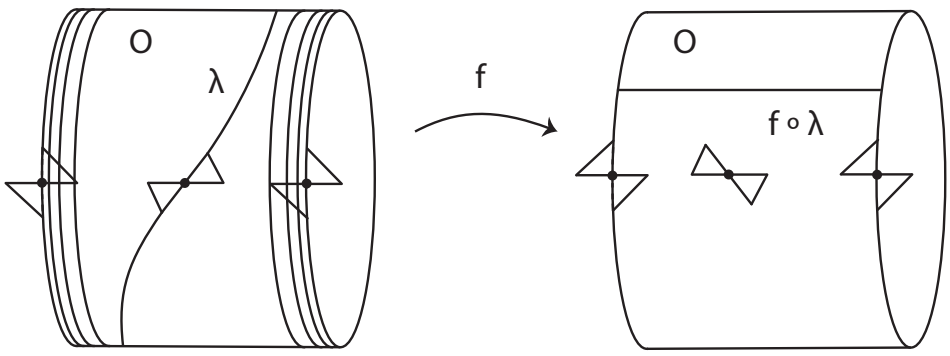

Figure 11.7: The isometry f is such that the null geodesic $f \circ \lambda$ is "unwound."

We have just seen how the Williams (1984) example can be used to show the instability of the no-hole spacetime properties of geodesic completeness and local maximality. But because of the compactness of the spacetimes (N, g) and (N, g_i), the instability results do not carry over to the weaker no-hole condition of spacetime maximality. This follows since any compact spacetime must be $\mathscr{U}$-maximal (O'Neill, 1983, p. 155). So because $\mathscr{U}$-maximality implies $\mathscr{P}$-maximality for any collection $\mathscr{P} \subseteq \mathscr{U}$, we find that

any collection $\mathscr{P}$ of compact spacetimes is a collection of $\mathscr{P}$-maximal spacetimes. So (Stability) must be true for such a collection $\mathscr{P}$.

What else is known concerning the (in)stability of spacetime maximality? Very little. Recall the limited result mentioned above: if (M, g) is a geodesically complete, globally hyperbolic spacetime, there will be a C^1 neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ of (M, g) such that each spacetime in the neighborhood $\mathscr{O}$ is timelike and null geodesically complete. One can show that any spacetime will be $\mathscr{U}$-maximal if it is timelike, null, or spacelike geodesically complete (Beem et al., 1996, p. 220). Combining these two results, we see that if (M, g) is a geodesically complete, globally hyperbolic spacetime, there will be a C^1 neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ of (M, g) such that each spacetime in the neighborhood $\mathscr{O}$ is $\mathscr{U}$-maximal (Beem et al., 1996, p. 270). It is remarkable that this single statement seems to be only known positive result in literature that speaks in favor of the stability of spacetime maximality.

11.7 Subcollection Problem

We close by highlighting a vexing subcollection problem with respect to the (Stability) condition. Consider the collection $(Vac) \cap (GH)$ of globally hyperbolic vacuum solutions. Such a collection contains only incredibly well-behaved spacetimes – locally and globally. Indeed, $(Vac) \cap (GH)$ is a subcollection of any of the local and causal properties we have been considering. It is an open question whether $(Vac) \cap (GH)$ satisfies (Stability). Even if it does, physically unreasonable spacetimes still lurk within this collection. Consider, for example, the $t < 0$ region of Minkwoski spacetime in which notches have been removed that spell out the word "Leibniz" in Morse code (see Figure 11.8). Given the existence of such globally hyperbolic vacuum solutions, one would like some assurance that (Stability) is true for any subcollection $\mathscr{P} \subseteq (Vac) \cap (GH)$. We now show the impossibility of such a result (Manchak, 2023).

We start by constructing a collection $\mathscr{P} \subset (Vac) \cap (GH)$. Consider the smooth bump function $u : [-2, 2] \to \mathbb{R}$ defined by $u(t) = \exp[1/(t^2 - 1)]$ for $-1 < t < 1$ and $u(t) = 0$ otherwise. For each integer $i \geq 1$, we now define a pair of functions $f_i, F_i : [-2, 2] \to \mathbb{R}$. We let $f_i(t) = \sqrt{1 - u(t)/i}$ and we let $F_i(t)$ be the result of integrating the function $f_i(x)$ from $x = 0$ to $x = t$ (see Figure 11.9). It will be useful later on to note the following: $F_i(t)$ has a smooth inverse for all i, $F_1(1) \approx 0.88$, and $F_1(2) = F_1(1) + 1 \approx 1.88$.

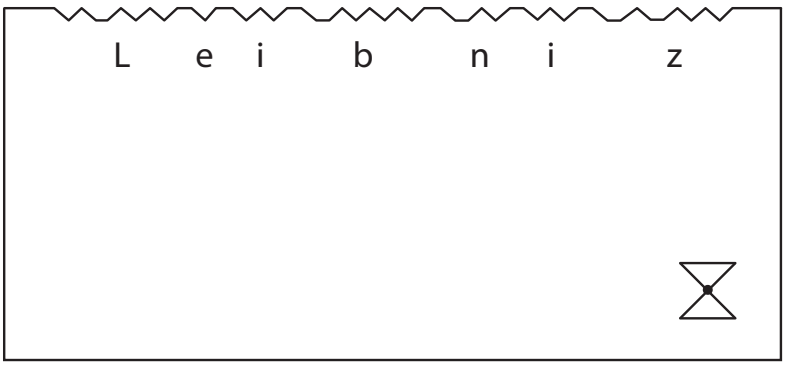

Figure 11.8: A globally hyperbolic vacuum solution in which the removed notches spell out the word "Leibniz" in Morse code.

Now let the manifold M be the $-2 < t < 2$ portion of the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. For each integer $i \geq 1$, let (M, g_i) be the spacetime where g_i is defined as follows: at each point $(t, \theta) \in M$ and for any vectors $v = [v_t, v_\theta]$ and $w = [w_t, w_\theta]$ at the point, let $g_i(v, w) = f_i^2(t) v_t w_t - v_\theta w_\theta$. Let (M, g) be the spacetime where $g_i(v, w) = v_t w_t - v_\theta w_\theta$. Define $\mathscr{P} \subset \mathscr{U}$ be the collection consisting of the spacetimes (M, g) and (M, g_i) for all $i \geq 1$. We claim that (i) $\mathscr{P} \subset (Vac) \cap (GH)$ and (ii) (Stability) is false for the collection $\mathscr{P}$. We will argue for (i) and (ii) in turn.

To show that (i) holds, first note $(M, g) \in (Vac) \cap (GH)$ as it is just the $-2 < t < 2$ portion of two-dimensional Minkowski spacetime "rolled up" in the spacelike direction. What about the rest of the spacetimes in the collection $\mathscr{P}$? It turns out that each (M, g_i) is isometric to a globally hyperbolic portion of (M, g) and hence in the collection $(Vac) \cap (GH)$ as well. Let's verify this. For each $i \geq 1$, let (M_i, g) be the $-F_i(2) < t < F_i(2)$ portion of (M, g). Here, it is helpful to note that $1 < F_i(2) < 2$ for all i (recall $F_1(2) \approx 1.88$) that and $F_i(2)$ approaches 2 as $i \to \infty$. We now claim that for each $i \geq 1$, the spacetime (M, g_i) is isometric to the spacetime (M_i, g). To see this, just consider the isometry $\psi_i : M \to M_i$ defined by $\psi_i(t, \theta) = (F_i(t), \theta)$. When the metric g on M_i is pulled back to the metric $\psi_i^*(g) = g_i$ on M, we see that the light cones are stretched in the region $-1 < t < 1$ so as to match up with those of the metric g_i on M (see Figure 11.10). Since each spacetime

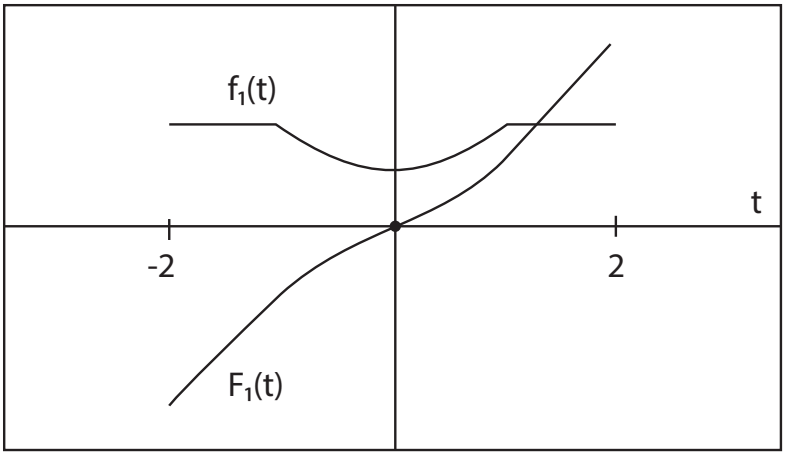

Figure 11.9: The functions $f_1(t)$ and $F_1(t)$.

(M, g_i) is isometric to (M_i, g) (which is just a globally hyperbolic portion of (M, g)) it follows easily that for each i, we have $(M, g_i) \in (Vac) \cap (GH)$

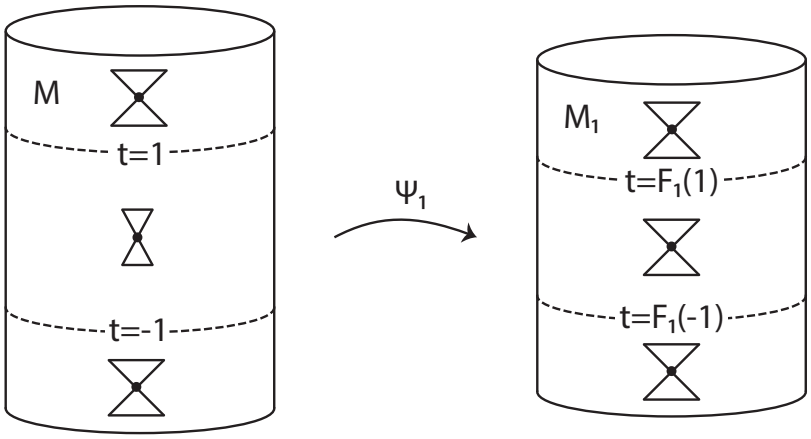

Figure 11.10: The isometry ψ_1. When the metric g on M_1 is pulled back to the metric $\psi_1^*(g) = g_1$ on M, the light cones are stretched in the region $-1 < t < 1$ so as to match up with those of the metric g_1 on M. Recall that $F_1(1) \approx 0.88$.

Now we show (ii) (Stability) is false for the collection $\mathscr{P}$. Since each space-

time (M, g_i) is isometric to (M_i, g) (which is proper sub-portion of (M, g)) we know that each (M, g_i) is $\mathscr{P}$-extendible. We also see that (M, g) is $\mathscr{P}$-maximal since it cannot extend itself. Consider any Riemannian metric h on M. Let N be the compact region of M where $-1 \leq t \leq 1$. By construction, $d(g, g_i, h, 0) = 0$ at each point in $M - N$. On the compact region N region, the quantity $\sup_N[d, (g, g_i, h, 0)]$ is bounded and approaches zero as $i \to \infty$. It follows that every C^0 fine neighborhood $\mathscr{O} \subseteq \mathscr{L}(M)$ of the $\mathscr{P}$-maximal spacetime (M, g) will contain, for sufficiently large i, the $\mathscr{P}$-extendible spacetime (M, g_i). So the collection $\{(M, g)\}$ of $\mathscr{P}$-maximal spacetimes is not C^0 stable relative to $\mathscr{P}$. So the collection $\mathscr{P}$ fails to satisfy the (Stability) collection.

Because $\mathscr{P} \subseteq (Vac) \cap (GH)$, it is difficult to see how one might rule out this collection as "physically unreasonable" without invoking an no-hole condition of some kind. All such conditions are at least as strong as $\mathscr{U}$-maximality which, in turn, implies $\mathscr{P}$-maximality. So invoking a no-hole condition is tantamount to requiring $\mathscr{P}$-maximality itself – the very property under investigation.

11.8 Conclusion

Stepping back, we see that very few results are available concerning the stability of spacetime maximality. It is unknown if the relatively weak (Stability) condition is satisfied by the collection $\mathscr{U}$ or by any of the standard reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$ we have been considering. If the customary line is correct that "in order to be physically significant, a property of space-time ought to have some form of stability, that is to say, it should be a property of 'nearby' space-times" (Hawking and Ellis, 1973, p. 197), then it is not at all clear that spacetime maximality is a physically significant property.

Moreover, the few limited results we do have do not seem to support the dogma of spacetime maximality. We first highlighted a surprising example to due to Williams (1984) showing that the stronger no-hole property of geodesic completeness is not stable relative to the collection $\mathscr{U}$. Next, we used this example to show that the weaker no-hole property of local maximality is also unstable. Both results suggest that perhaps similar instability results hold for even weaker spacetime maximality property as well. Indeed, we have seen that some collections $\mathscr{P} \subseteq (Vac) \cap (GH)$ of globally hyperbolic vacuum solutions fail to satisfy the (Stability) condition (Manchak, 2023).

Such collections cannot be ruled out via the imposition of the usual local or causal properties. And invoking a no-hole condition is tantamount to requiring $\mathscr{P}$-maximality itself – the very property under investigation. It would seem that a disturbing subcollection problem with respect to the stability of spacetime maximality will remain no matter what isolated stability results can secured in the future.

Chapter 12

Determinism

12.1 Introduction

Here we consider the notion of "determinism" within the context of general relativity. A celebrated result due to Choquet-Bruhat and Geroch (1969) captures a sense in which determinism holds: any "initial data set" gives rise to a unique (up to isometry) "development" spacetime. But we will emphasize that the result goes through only after a crucial maximality assumption is made concerning a particular dynamical form of spacetime maximality. The uniqueness clause holds only if one limits attention to "maximal" developments. Moreover, this maximality assumption presupposes that the collection $\mathscr{U}$ is used as a background possibility space. We will draw attention to the fact that analogues of the Choquet-Bruhat and Geroch (1969) statement can be false relative to various reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$. Indeed, we will highlight another instance of the subcollection problem within this context.

We then revisit the related cosmic censorship conjecture of Penrose (1979). The conjecture presupposes two forms of spacetime maximality: the dynamical form utilized in the Choquet-Bruhat and Geroch (1969) result and another which ensures that the initial data set is "as large as it can be" in the appropriate sense. The hope is that when attention is restricted to a certain collections $\mathscr{P} \subset \mathscr{U}$ of "physically reasonable" spacetimes, the two forms of spacetime maximality secure a third form: $\mathscr{P}$-maximality itself. We review an influential formulation of the cosmic censorship conjecture due to Wald (1984) and articulate a generalized variant relative to a choice

of background possibility space $\mathscr{P} \subseteq \mathscr{U}$. Some natural choices render the conjecture either false or open. We also explore the prospect of using asymmetry properties to rule out potential counterexamples to cosmic censorship. We close with a discussion concerning how the notions of determinism and cosmic censorship considered here could be used to justify the dogma of spacetime maximality from a dynamical perspective. We emphasize the limitations of such an approach given the many questions that remain unsettled as well as the lurking subcollection problem.

12.2 Maximal Developments

In what follows, we restrict our discussion of determinism to the context of vacuum solutions where things are relatively simple; an analogous discussion could be carried out in the non-vacuum case. Let (M, g) be any four-dimensional globally hyperbolic vacuum solution and let Σ be any three-dimensional, connected spacelike surface in M. The metric g on M induces a two part initial data set on S: a natural Riemannian **spatial metric** h as well as an associated **extrinsic curvature** π. The latter can be thought of a type of "time derivative" of h and captures how the surface S is embedded in M. Since the spacetime (M, g) is a vacuum solution, we know that h and π must satisfy the appropriate vacuum "constraint equations" on Σ (see Wald 1984, p. 259). Let any triple (Σ, h, π) arising in this way be called an **initial data set**.

We note that initial data sets are usually defined more directly, i.e. without making reference to a background vacuum spacetime. But one can show that the two definitions are equivalent since any initial data set (Σ, h, π) defined in the more direct way always finds a home in some "development" spacetime – a globally hyperbolic vacuum solution (M, g) with Cauchy S and an appropriate diffeomorphism from Σ to S. This "local" existence result concerning developments is the starting point for the work of Choquet-Bruhat and Geroch (1969, p. 331). Their main result shows that not only do developments exist for a given initial data set, but there always exists a "maximal" such development that is unique up to isometry. The notion of maximality considered presupposes the standard background possibility space $\mathscr{U}$. We now work to make precise a generalized version of the Choquet-Bruhat and Geroch (1969) statement that is relativized to a choice of arbitrary reduced collection $\mathscr{P} \subseteq \mathscr{U}$. This will allow us to explore the notion of determinism

in a more nuanced way.

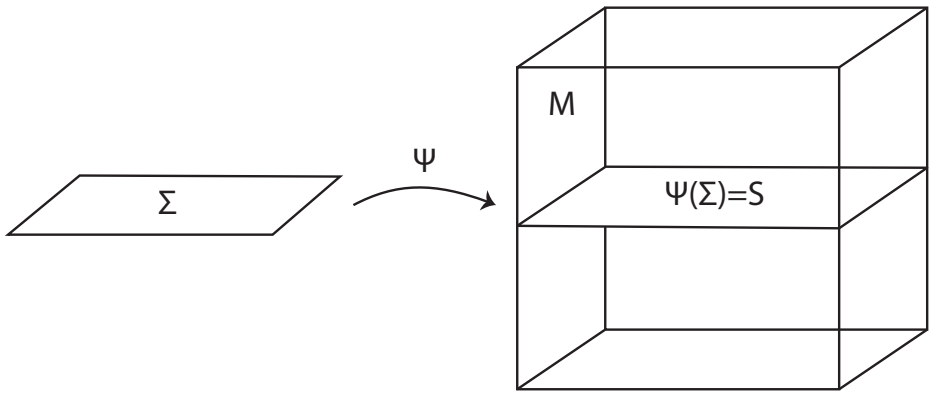

Figure 12.1: The diffeomorphism ψ from the initial data surface Σ to the Cauchy surface S in the $\mathscr{P}$-development spacetime (M, g).

Let (Σ, h, π) be an initial data set. A **$\mathscr{P}$-development** of the initial data set is a triple $((M, g), S, \psi)$ where (M, g) is a globally hyperbolic vacuum $\mathscr{P}$-spacetime, $S \subset M$ is a Cauchy surface, and $\psi : \Sigma \to S$ is a diffeomorphism such that $\psi_*(h)$ and $\psi_*(\pi)$ are, respectively, the spatial metric and extrinsic curvature on S induced from the metric g (see Figure 12.1). For convenience, we will often refer to the spacetime (M, g) as the $\mathscr{P}$-development rather than the triple $((M, g), S, \psi)$.

Whether or not an initial value set (Σ, h, π) has a $\mathscr{P}$-development depends on the collection $\mathscr{P}$. For example, let (M, η) be four-dimensional Minkowski spacetime in standard (t, x, y, z) coordinates. Let Σ be the $t = 0$ region of M and let (Σ, h, π) be the initial data set induced from η. One can show that if a $\mathscr{P}$-development $((N, g), S, \psi)$ exists for (Σ, h, π), then (N, g) must be flat. It follows that if we let $\mathscr{P}$ be the collection of Heraclitus spacetimes (none of which are flat), then (Σ, h, π) has no $\mathscr{P}$-development. This makes sense. The initial data (Σ, h, π) inherited from Minkowski spacetime has non-trivial symmetries in the sense that there is a diffeomorphism $f : \Sigma \to \Sigma$ such that $f_*(h) = h$ and $f_*(\pi) = \pi$ where f is not the identity map. Thus, one would not expect that this initial data with non-trivial symmetries could give rise to a development spacetime with radical Heraclitus asymmetries. Indeed, the example could be considered an instance of "Curie's principle" which states:

"When certain effects show a certain asymmetry, this asymmetry must be found in the causes which gave rise to it" (Curie, 1894, p. 401). Additional results concerning determinism and (a)symmetries will be explored as we go along. (See also Earman (2007) for a nice discussion.)

Given the way we have set things up, any initial value set has a $\mathscr{U}$-development. But $\mathscr{U}$-developments are highly non-unique. Consider again the example initial data set (Σ, h, π) from above arising from Minkowski spacetime (M, η). Of course, $((M, \eta), \Sigma, \psi)$ is a $\mathscr{U}$-development of of this initial data where $\psi : \Sigma \to \Sigma$ is the identity map. But if N is the $-k < t < k$ region of M for any real number $k > 0$, then $((N, \eta), \Sigma, \psi)$ is also a $\mathscr{U}$-development of (Σ, h, π). This follows since the truncated spacetime (N, η) is also a globally hyperbolic vacuum solution with Cauchy surface Σ (see Figure 12.2).

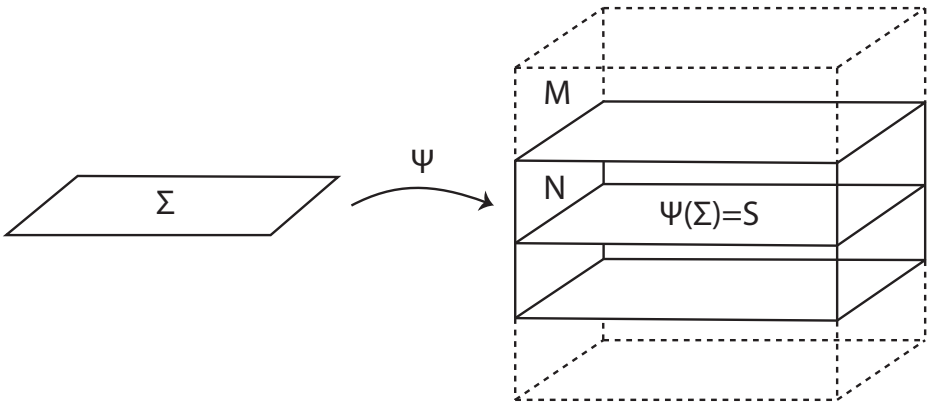

Figure 12.2: The initial data set (Σ, h, π) has both Minkowski spacetime (M, η) and the truncated spacetime (N, η) as $\mathscr{U}$-developments.

Let $((M, g), S, \psi)$ and $((M', g'), S', \psi')$ be $\mathscr{P}$-developments of the same initial data set (Σ, h, π). We say that $((M', g'), S', \psi')$ is a (not necessarily proper) $\mathscr{P}$**-extension** of $((M, g), S, \psi)$ if there is an isometric embedding $f : M \to M'$ such that the composed map $\psi'^{-1} \circ f \circ \psi$ is the identity on Σ (see Figure 12.3). Returning the Minkowski spacetime example from above, we see that the $\mathscr{U}$-development $((M, \eta), \Sigma, \psi)$ is a (proper) $\mathscr{U}$-extension of the $\mathscr{U}$-development $((N, \eta), \Sigma, \psi)$. To see this, just let $f : N \to M$ be the

natural inclusion map and note that, since ψ is the identity map on Σ, so is the composed map $\psi^{-1} \circ f \circ \psi$.

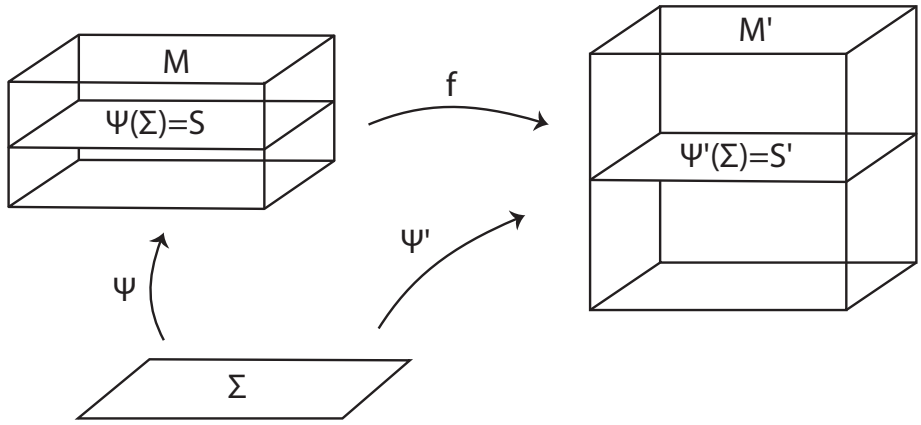

Figure 12.3: If the isometric embedding $f : M \to M'$ is such that the composed map $\psi'^{-1} \circ f \circ \psi$ is the identity on Σ, then the $\mathscr{P}$-development (M', g') is a $\mathscr{P}$-extension of the $\mathscr{P}$-development (M, g).

A $\mathscr{P}$-development $((M', g'), S', \psi')$ of the initial data set (Σ, h, π) is said to be a $\mathscr{P}$-**maximal development** of (Σ, h, π) if it is an $\mathscr{P}$-extension of any other $\mathscr{P}$-development $((M, g), S, \psi)$ of (Σ, h, π). If a $\mathscr{P}$-development $((M, g), S, \psi)$ for some initial data set (Σ, h, π) is such that (M, g) is a $\mathscr{P}$-maximal spacetime, then it is immediate that $((M, g), S, \psi)$ is a $\mathscr{P}$-maximal development of the same initial data set. But the other direction does not hold: $\mathscr{P}$-maximal developments can fail to be $\mathscr{P}$-maximal spacetimes.

To see this, consider again the example of Minkowski spacetime (M, η) from above and let Σ' be the $x^2 + y^2 + z^2 < 1$ portion of the $t = 0$ surface $\Sigma \subset M$. The Riemannian metric and extrinsic curvature on Σ' are inherited from Σ so that (Σ', h, π) counts as is an initial data set. Let $M' \subset M$ be the domain of dependence $D(\Sigma')$ of Σ'. One can show that $((M', \eta), \Sigma', \psi')$ is a $\mathscr{U}$-maximal development of (Σ', h, π) where ψ' is the identity map on Σ'. This follows since the spacetime (M', η) has no $\mathscr{U}$-extension in which Σ' remains is a Cauchy surface. But of course, the spacetime (M', η) is not $\mathscr{U}$-maximal since Minkowski spacetime (M, η) is a $\mathscr{U}$-extension (see Figure 12.4).

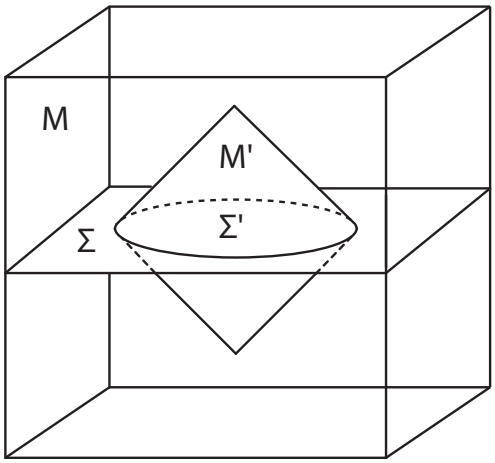

Figure 12.4: The $\mathscr{U}$-maximal development of the initial data set (Σ', h, π) is the spacetime (M', η) which fails to be $\mathscr{U}$-maximal since it can be extended by Minkowski spacetime (M, η).

12.3 Existence and Uniqueness

One might have thought that if an initial data set (Σ, h, π) has a $\mathscr{P}$-maximal development $((M, g), S, \psi)$, then it must be unique. But one can construct innumerable isometric invariants by considering a hole diffeomorphism $f : M \to M$ which acts as the identity map in an open set O containing the Cauchy surface S but does not act as the identity map in the "hole" region $M - O$. The spacetime $(M, f_*(g))$ is then isometric but not identical to (M, g) and one can show that $((M, f_*(g), S, \psi)$ is also a $\mathscr{P}$-maximal development of (Σ, h, π). We are now in a position to make precise a generalized statement of Choquet-Bruhat and Geroch (1969) that captures a sense in which determinism holds relative to some reduced possibility space. Consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$.

> (Determinism) For every $\mathscr{P}$-development of any initial data set, there is a $\mathscr{P}$-maximal development of the same initial data set that is unique up to isometry.

Here, the uniqueness clause can be understood as follows: If $((M, g), S, \psi)$ and $((M', g'), S', \psi')$ are both $\mathscr{P}$-maximal developments of the same initial data set (Σ, h, π), then there is an isometry $f : M \to M'$ such that $f \circ \psi = \psi'$. The result Choquet-Bruhat and Geroch (1969) shows that for every $\mathscr{U}$-development of any initial data set, there is a $\mathscr{U}$-maximal development of the same initial data set that is unique up to isometry. But since, by definition, any $\mathscr{U}$-maximal development $((M, g), S, \psi)$ is such that (M, g) is a globally hyperbolic vacuum solution, the result shows that (Determinism) is true for any $\mathscr{P} \subseteq \mathscr{U}$ which contains the collection $(GH) \cap (Vac)$ of all globally hyperbolic vacuum solutions. It follows that (Determinism) is true for all of the causal and local properties we have been considering. But we note that there is no assurance arbitrary subcollections of $(GH) \cap (Vac)$ will render (Determinism) true (see Figure 12.5). Indeed, we will consider some example subcollections of $(GH) \cap (Vac)$ for which (Determinism) is false in due course.

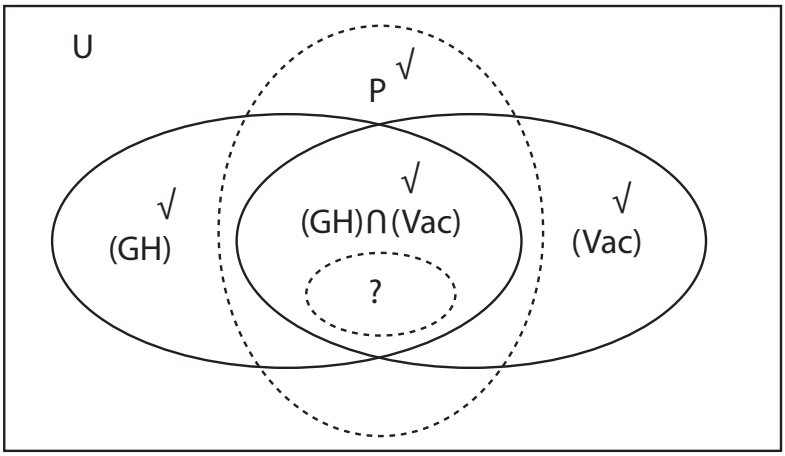

Figure 12.5: Any collection $\mathscr{P} \subseteq \mathscr{U}$ which contains $(GH) \cap (Vac)$ satisfies (Determinism). But there is no assurance that arbitrary subcollections of $(GH) \cap (Vac)$ will also satisfy (Determinism).

The result of Choquet-Bruhat and Geroch (1969) uses Zorn's lemma to build the relevant $\mathscr{U}$-maximal development. But we note that recent work shows that Zorn's lemma is not needed for the result to go through (Wong, 2013; Sbierski, 2016). Here, we give an example of the use of Zorn's lemma in this context by sketching a proof that (Determinism) is true for the collection

$\mathscr{U}$. The sketch closely follows the presentations found in Choquet-Bruhat and Geroch (1969), Hawking and Ellis (1973, p. 249), and Wald (1984, p. 263).

Let $((M, g), S, \psi)$ be any $\mathscr{U}$-development of any initial data set (Σ, h, π). Let $\mathfrak{D}$ be the collection of all $\mathscr{U}$-developments of (Σ, h, π). Let $\sim$ be the relation on $\mathfrak{D}$ defined such that, for all $((M, g), S, \psi), ((M', g'), S', \psi') \in \mathfrak{D}$, we have $((M, g), S, \psi) \sim ((M', g'), S', \psi')$ if there is an isometry $f : M \to M'$ such that $f \circ \psi = \psi'$. One can check that $\sim$ is an equivalence relation on $\mathfrak{D}$. Let $\mathfrak{D}/\sim$ be the collection of all equivalence classes of all $\mathscr{U}$-developments of (Σ, h, π). Now let the relation $\leq$ on $\mathfrak{D}/\sim$ be defined such that, for all $[((M, g), S, \psi)], [((M', g'), S', \psi')] \in \mathfrak{D}/\sim$, we have $[((M, g), S, \psi)] \leq [((M', g'), S', \psi')]$ if any $\mathscr{U}$-development in $[((M', g'), S', \psi')]$ is a $\mathscr{U}$-extension of any $\mathscr{U}$-development in $[((M, g), S, \psi)]$. One can show that this relation $\leq$ is a partial order on $\mathfrak{D}/\sim$ (see Wald 1984, p. 263). Let $\mathfrak{T}$ be a subcollection of $\mathfrak{D}/\sim$ that is totally ordered by $\leq$. We now show that $\mathfrak{T}$ has an upper bound in $\mathfrak{D}/\sim$ and then invoke Zorn's lemma to establish the existence of a $\mathscr{U}$-maximal development of the initial data set (Σ, h, π). We then show that this $\mathscr{U}$-maximal development is unique up to isometry.

For each equivalence class $X_i \in \mathfrak{T}$, use the axiom of choice to choose a representative $\mathscr{U}$-development $((M_i, g_i), S_i, \psi_i)$. For any $\mathscr{U}$-developments $((M_i, g_i), S_i, \psi_i)$ and $((M_j, g_j), S_j, \psi_j)$ in X_i and X_j respectively, if $i \leq j$, there must be an isometric embedding $f_{ij} : M_i \to M_j$ such that the composed map $\psi_j^{-1} \circ f_{ij} \circ \psi_i$ is the identity on Σ. Using the properties of globally hyperbolic spacetimes, one can show that each such embedding f_{ij} must be unique (Hawking and Ellis, 1973, p. 249). These unique isometric embeddings can then be used to form a natural "union" spacetime (M, g) and an associated $\mathscr{U}$-development $((M, g), S, \psi)$ (recall the similar construction outlined in Section 9.3). So the equivalence class $[((M, g), S, \psi)]$ will be an upper bound in $\mathfrak{T}$.

From Zorn's lemma, there must be a maximal element X in $\mathfrak{D}/\sim$. In general, maximal elements need not be unique. But in the present case, X is unique. To see this, suppose there were another maximal element Y in $\mathfrak{D}/\sim$. One could then consider $\mathscr{U}$-developments in X and Y and "patch together" their associated spacetimes to construct another $\mathscr{U}$-development whose equivalence class Z is strictly "larger" than X, i.e. $Z \neq X$ and $X \leq Z$ (Wald, 1984, p. 263). This follows easily once one works to verify that the patched together spacetime satisfies the Hausdorff condition (Choquet-Bruhat and Geroch, 1969, p. 333). The existence of such an "extension" Z of

X violates the maximality of X: a contradiction. So the maximal element X is unique. Let $((M^*, g^*), S^*, \psi^*)$ be any $\mathscr{U}$-development in the unique maximal element $X \in \mathfrak{D}/\sim$. The uniqueness at the level of equivalence classes of developments translates to uniqueness only up to isometry at the level of developments. So we see that the $\mathscr{U}$-maximal development $((M^*, g^*), S^*, \psi^*)$ of the initial data set (Σ, h, π) is unique up to isometry, i.e. any other $\mathscr{U}$-maximal development of (Σ, h, π) must also be a member of the equivalence class X.

12.4 Asymmetry Properties

Here, we explore the (Determinism) condition with respect to asymmetry properties. Let $\mathscr{P} \subset \mathscr{U}$ be any collection such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$ where (PR), (FP), and (Gir) are, respectively, the collections of all point rigid, fixed point, and giraffe spacetimes. We will show that $\mathscr{P}$ must fail to satisfy (Determinism). Consider again the example of Minkowski spacetime (M, η) and the associated initial data surface (Σ, h, π) where Σ is the $t = 0$ portion of M. Now, for each integer $i > 0$, let (M_i, η_i) be constructed as follows. Take Minkowski spacetime (M, η) and remove a three-dimensional compact region shaped like a giraffe on the $t = i$ surface. In the resulting mutilated spacetime, let (M_i, η_i) be the region $D(\Sigma)$ considered as a spacetime is its own right (see Figure 12.6).

By construction, each spacetime (M_i, η_i) is giraffe. Moreover, since each spacetime (M_i, η_i) contains Σ, we see that $((M_i, \eta_i), \Sigma, \psi)$ is a $\mathscr{P}$-development of (Σ, h, π) where ψ is the identity map on Σ. Suppose there were a $\mathscr{P}$-maximal development $((M^*, \eta^*), \Sigma^*, \psi^*)$ of (Σ, h, π). So (M^*, η^*) is a $\mathscr{P}$-spacetime. We show a contradiction. Since $((M^*, \eta^*), \Sigma^*, \psi^*)$ is a $\mathscr{P}$-maximal development, it must be a $\mathscr{P}$-extension of $((M_i, \eta_i), \Sigma, \psi)$ for all integers $i > 0$. Of course, any $\mathscr{P}$-extension must be a $\mathscr{U}$-extension; similarly any $\mathscr{P}$-development must be a $\mathscr{U}$-development. So $((M^*, \eta^*), \Sigma^*, \psi^*)$ must be a $\mathscr{U}$-extension of each of the $\mathscr{U}$-developments $((M_i, \eta_i), \Sigma, \psi)$ for all integers $i > 0$. This can only happen if (M^*, η^*) is isometric to Minkowski spacetime. But Minkowski spacetime is not in the collection $\mathscr{P}$ since it fails to be both point rigid and fixed point. Since (M^*, η^*) is a $\mathscr{P}$-spacetime we have a contradiction. So there is no $\mathscr{P}$-maximal development of the initial data set (Σ, h, π) which shows that the (Determinism) condition is false for the collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$.

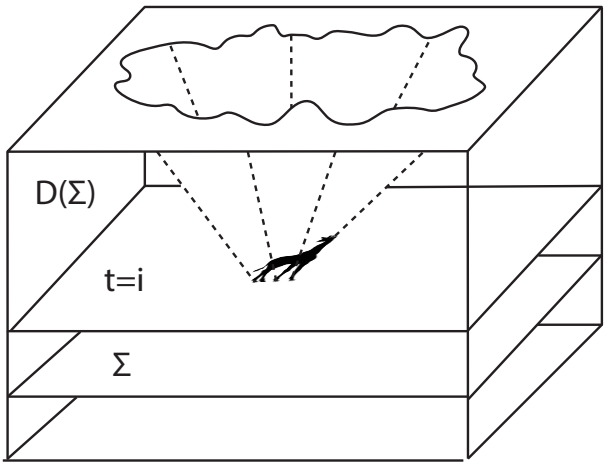

Figure 12.6: A giraffe is removed on the $t = i$ surface in Minkowski spacetime. In the resulting mutilated spacetime, the region $D(\Sigma)$, considered as a spacetime is its own right, is (M_i, η_i).

In the example just given, an asymmetric development fails to have an asymmetric maximal development. One might think that the problem here is that, although the development in the example is asymmetric, the initial data set for this development is not. Indeed, the $t = 0$ surface Σ in Minkowski spacetime has non-trivial symmetries in the sense that there are diffeomorphisms from Σ to itself which preserve the induced Riemannian metric and extrinsic curvature. But we emphasize here that there are other examples in which the initial data set is "giraffe" in the appropriate sense, i.e. no non-trivial global symmetries, and yet the $\mathscr{U}$-maximal development of this initial guaranteed by the Choquet-Bruhat and Geroch (1969) result is not a giraffe spacetime. So even in the standard context, it is not the case that asymmetric initial data gives rise to an asymmetric maximal development.

To see this, let Σ' be a three-dimensional open giraffe shaped region in $t = 0$ surface Σ in Minkowski spacetime (M, η). The Riemannian metric and extrinsic curvature on Σ' are inherited from Σ so that (Σ', h, π) counts as is an initial data set. The giraffe shape of Σ' ensures that the initial data set (Σ', h, π) is free of non-trivial symmetries. But the $\mathscr{U}$-maximal development (M', η) will be the domain of dependence $D(\Sigma')$ region in Minkowski spacetime. And this spacetime (M', η) is not giraffe since there will be a non-trivial reflection isometry $f : M' \to M'$ defined by $f(t, x, y, z) = (-t, x, y, z)$.

This isn't so surprising though. One wouldn't expect asymmetry in three dimensions to necessarily give rise to asymmetry in four dimensions. In contrast, one would expect symmetry in three dimensions to necessarily give rise to symmetry in four dimensions. Indeed, here is a result along these lines. Let (Σ, h, π) be an initial data set which fails to be giraffe in the appropriate sense. The $\mathscr{U}$-maximal development $((M, g), S, \psi)$ of this initial data set is such that the spacetime (M, g) must fail to be giraffe (Friedrich and Rendall, 2000, p. 216-217).

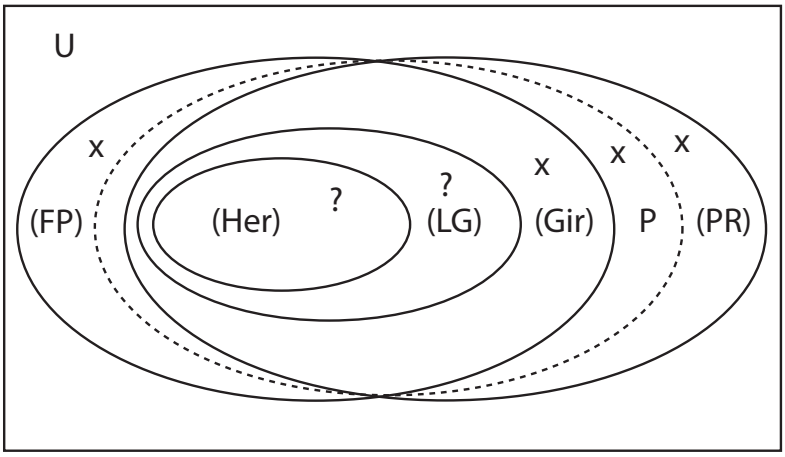

Figure 12.7: (Determinism) is false for any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$. The cases for (LG) and (Her) are open.

We have seen that (Determinism) is false for any collection $\mathscr{P} \subset \mathscr{U}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$ where (PR), (FP), and (Gir) are, respectively, the collections of all point rigid, fixed point, and giraffe spacetimes. What about the other asymmetry conditions? It is an open question whether (Determinism) is satisfied by the collections (LG) and (Her) of all locally giraffe and all Heraclitus spacetimes respectively (see Figure 12.7).

12.5 Dynamic Extendibility

Suppose (Determinism) is true for some collection $\mathscr{P} \subseteq \mathscr{U}$. So we know that for every $\mathscr{P}$-development of any initial data set, there is a $\mathscr{P}$-maximal development of the same initial data set that is unique up to isometry. But as we

have seen, a $\mathscr{P}$-maximal development need not be a $\mathscr{P}$-maximal spacetime. Thus, even after one implicitly assumes a dynamical form of spacetime maximality by limiting attention only to $\mathscr{P}$-maximal developments of initial data sets, the satisfaction of (Determinism) does not amount to a justification for the dogma of spacetime maximality. The cosmic censorship conjecture can be seen as a strictly stronger second-order condition on a collection $\mathscr{P} \subseteq \mathscr{U}$ that, if satisfied, ensures that any $\mathscr{P}$-maximal development of any "suitable" initial data set is always a $\mathscr{P}$-maximal spacetime. So relative to the collection $\mathscr{P}$, a dynamical justification for the dogma of spacetime maximality is established. In this way, much depends on the cosmic censorship conjecture. Let us now work to make precise a general statement.

We have already seen how a "small" initial data set (Σ, h, π) can yield a $\mathscr{U}$-maximal development that is not $\mathscr{U}$-maximal (recall Figure 12.4). One way to rule out such examples is to require that (Σ, h) be geodesically complete as a Riemannian manifold (Wald, 1984, p. 305). This implies that the initial data set (Σ, h, π) is "maximal" in a natural sense: there is no other initial data set (Σ', h', π') with proper subset $O \subset \Sigma'$ such that (Σ, h) and (O, h') are isometric Riemannian manifolds. Unfortunately, even this maximality condition does not ensure that an initial data set is appropriately suitable. Consider Minkowski spacetime (M, g) and let $\Sigma \subset M$ be the "past hyperboloid" given by $t = -\sqrt{x^2 + y^2 + z^2 + 1}$ and let h and π be, respectively, the Riemannian metric and extrinsic curvature on Σ induced from the metric g on M. The resulting initial data set (Σ, h, π) is such that (Σ, h) is a geodesically complete Riemannian manifold. We see that $((N, g), \Sigma, \psi)$ is a $\mathscr{U}$-maximal development of this initial data where ψ is the identity map and (N, g) is the timelike past of the origin in Minkowski spacetime (see Figure 12.8). So (N, g) fails to be a $\mathscr{U}$-maximal spacetime.

In the example just given, any $\mathscr{U}$-extension of the future Cauchy horizon $H^+(\Sigma)$ contains points p such that the closure of $I^-(p) \cap \Sigma$ is non-compact. For example, when extended by Minkowski spacetime (M, g), if $p \in H^+(\Sigma)$ is the origin point, then $I^-(p) \cap \Sigma$ is just the surface Σ which is closed but non-compact (see Figure 12.8). This behavior signals a "poor choice" of initial data set and for this reason and is forbidden by many formulations of the cosmic censorship conjecture (Geroch and Horowitz, 1979; Wald, 1984). We note that such formulations are concerned only with $\mathscr{U}$-extensions across the future Cauchy horizon $H^+(\Sigma)$. This is understandable given the focus of the cosmic censorship conjecture literature on forbidding singularities which are "naked" in the sense that an observer "sees" their formation (recall the

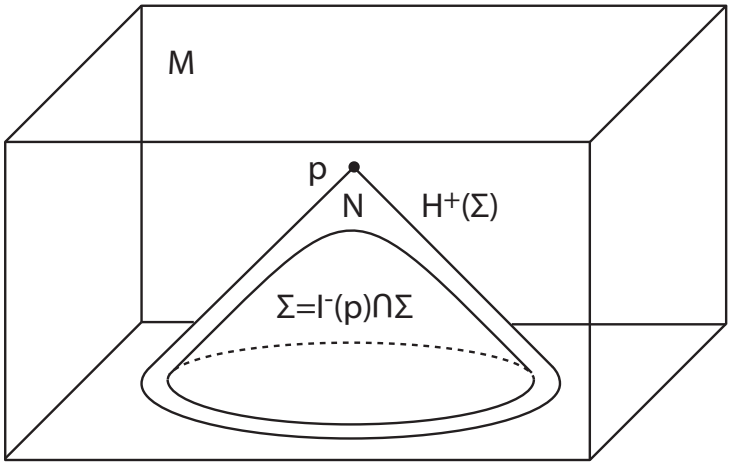

Figure 12.8: The $\mathscr{U}$-maximal development (N, g) of the past hyperboloid Σ. The origin point $p \in H^+(\Sigma)$ in Minkowski spacetime (M, g) is such that $I^-(p) \cap \Sigma = \Sigma$ which is closed but non-compact.

discussion in Section 6.3). But it is important to recognize that there are some initial data sets whose $\mathscr{U}$-maximal development can be extended across $H^-(\Sigma)$ but not $H^+(\Sigma)$. Indeed, just consider the "future" analogue of the past hyperboloid example in which the $\mathscr{U}$-maximal development (N, g) of the future hyperboloid S amounts to the timelike future of the origin in Minkowski spacetime. One can show that $H^+(\Sigma)$ is empty but (N, g) can be extended across $H^-(\Sigma)$. So there is a sense in which the future hyperboloid is also a "poor choice" for an initial data set: in any $\mathscr{U}$-extension, the future Cauchy horizon $H^+(\Sigma)$ contains points p such that the closure of $I^+(p) \cap \Sigma$ is non-compact. For example, when extended by Minkowski spacetime (M, g), if $p \in H^+(\Sigma)$ is the origin point, then $I^+(p) \cap \Sigma$ is just the surface Σ which is closed but non-compact (see Figure 12.9).

12.6 Cosmic Censorship

In light of the past and future hyperboloid examples, we now formulate a version of the cosmic censorship that concerns the extendibility of maximal developments generally – in both the future and past directions. Let us say that an initial data set (Σ, h, π) is **$\mathscr{P}$-suitable** if (i) (Σ, h) is a geodesically

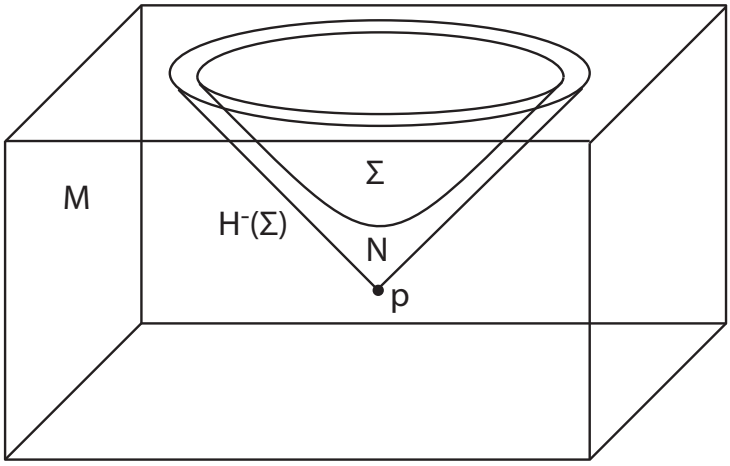

Figure 12.9: The $\mathscr{U}$-maximal development (N, g) of the future hyperboloid Σ is extendible across $H^-(\Sigma)$ but not $H^+(\Sigma)$ since the latter region is empty in any $\mathscr{U}$-extension.

complete Riemannian manifold (ii) there is a $\mathscr{P}$-development of (Σ, h, π) and (iii) for any $\mathscr{P}$-maximal development $((M, g), S, \psi)$ of (Σ, h, π), each $\mathscr{P}$-extension of the spacetime (M, g) is such that, if there is a $p \in H(S)$, then the closures of both $I^-(p) \cap \Sigma$ and $I^+(p) \cap \Sigma$ are compact. Now consider the following (second-order) condition on a spacetime property $\mathscr{P} \subseteq \mathscr{U}$ which is a strengthening of the (Determinsim) condition.

> (Censorship) The collection $\mathscr{P}$ satisfies (Determinism) and, in addition, the $\mathscr{P}$-maximal development of any $\mathscr{P}$-suitable initial data set is a $\mathscr{P}$-maximal spacetime.

We have set things up so that (Censorship) implies (Determinism). If the (Censorship) condition is satisfied by a collection $\mathscr{P} \subseteq \mathscr{U}$, we see that a type of dynamical justification for the $\mathscr{P}$-maximality of spacetime can be given. Suppose a collection $\mathscr{P}$ satisfies (Censorship). Any $\mathscr{P}$-suitable initial data set has a $\mathscr{P}$-development by definition. Because $\mathscr{P}$ satisfies (Censorship), it must satisfy (Determinism). So there is a $\mathscr{P}$-maximal development of this initial data set that is unique up to isometry. And because $\mathscr{P}$ satis-

fies (Censorship), this $\mathscr{P}$-maximal development is a $\mathscr{P}$-maximal spacetime. Finally, one invokes a Leibnizian metaphysics from a dynamical perspective to complete the justification: "If one adopts the image of spacetime as being generated or built up as time passes then the dynamical version of the principle of sufficient reason would ask why the Creative Force would stop building if it is possible to continue" (Earman, 1995, p. 32). If the dynamical justification for spacetime maximality breaks down for a collection $\mathscr{P}$, keeping track of both the (Determinism) and (Censorship) conditions will allow us to pinpoint where this breakdown occurs.

Even in the best case scenario in which a collection $\mathscr{P}$ satisfies both conditions, two forms of spacetime maximality relative to $\mathscr{P}$ are presupposed in order to secure $\mathscr{P}$-maximality: one form which assumes that an initial data set (Σ, h, π) is $\mathscr{P}$-suitable which requires that (Σ, h) is "as large as it can be" as a Riemannian manifold and a second form in which a dynamical version of the principle of sufficient reason is used to select, among all possible $\mathscr{P}$-developments of (Σ, h, π), the unique (up to isometry) $\mathscr{P}$-maximal development. But this brings to mind the "dirty open secret" highlighted by Earman (recall Section 6.5) in which practitioners display a circular sort of logic in presupposing whatever is needed to secure determinism. To rule out a $\mathscr{P}$-development of (Σ, h, π) that fails to be a $\mathscr{P}$-maximal development is to establish by fiat a sense in which spacetime maximality holds (Earman, 1995, p. 98).

We now turn to the question of which collections $\mathscr{P} \subseteq \mathscr{U}$ satisfy (Censorship). It is well known that $\mathscr{U}$ does not. The "Taub-NUT" spacetime provides one such counterexample (Taub, 1951; Newman et al., 1963). Another similar counterexample is given by a four-dimensional version of Misner spacetime (Chruściel and Isenberg, 1993). Investigating the properties of the simpler two-dimensional version suggests why this must be. Let (M, g) be Misner spacetime (recall Section 6.6) and let Σ be the $t = -1$ compact, spacelike surface in M. We see that Σ has a domain of dependence $D(\Sigma)$ which makes up the $t < 0$ region of M (see Figure 12.10). This region $D(\Sigma)$, considered as a spacetime in its own right, is globally hyperbolic with Cauchy surface Σ. Call it (N, g). As we have previously observed, this "bottom half" of Misner spacetime cannot be extended by a globally hyperbolic spacetime. This suggests that the four-dimensional analogue to (N, g) counts as a $\mathscr{U}$-maximal development of the three-dimensional analogue to Σ (also compact) with appropriate initial data. Indeed this is the case. Since Misner spacetime extends (N, g), the latter spacetime is $\mathscr{U}$-extendible. Moreover, one

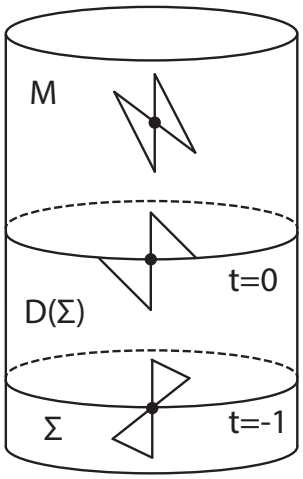

Figure 12.10: In Misner spacetime (M, g), the domain of dependence $D(\Sigma)$ of the compact, spacelike surface Σ at $t = -1$ is just the $t < 0$ region of M.

can verify that any $\mathscr{U}$-extension (N', g') of (N, g) will be such that $H^-(\Sigma)$ is empty and, for any point $p \in H^+(\Sigma)$, the region $I^-(p)$ is just $D(\Sigma) = N$ (see Figure 12.11). So the closure of $I^-(p) \cap \Sigma$ is just Σ which is compact. Similar results holds for in the four-dimensional case showing that the analogue to Σ counts as a $\mathscr{U}$-suitable initial data set. So we see that the collection $\mathscr{U}$ renders the (Censorship) condition false.

So far, we have focused on the cosmic censorship conjecture only within the vacuum context. Even more potential counterexamples arise when matter is brought into the picture. (For nice recent discussions of cosmic censorship, we refer the reader to Landsman (2021); Smeenk and Wüthrich (2021).) Indeed, a number of significant results show senses in which gravitational collapse leads to the formation of a naked singularity (Yodzis et al., 1973; Christodoulou, 1994). But it has been argued that "these examples are extremely special, owing to the fact that spherical symmetry is assumed" (Penrose, 1999, p. 242). The response has been to exclude these seemingly special counterexamples from consideration by moving to various "physically reasonable" reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$. But as we have seen, the question of what counts as a "physically reasonable" collection is a deeply murky one. Earman (1995, p. 80) reminds us that the term "physically unreasonable" should not be "used as an elastic label that can be stretched to include any ad hoc way of discrediting putative counterexamples."

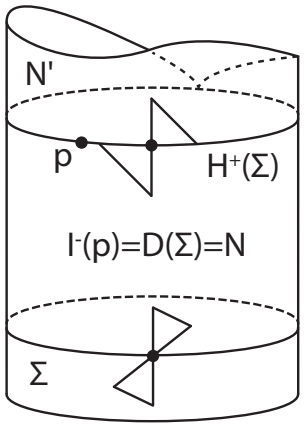

Figure 12.11: Any extension (N', g') of the spacetime (N, g) is such that, for any point $p \in H^+(\Sigma)$, the region $I^-(p)$ is just $D(\Sigma) = N$.

Returning to the vacuum case, the Taub-NUT and four-dimensional Misner examples can be ruled out in different ways. For example, Geroch and Horowitz (1979, p. 288) suggest that we limit attention to initial data sets (Σ, h, π) for which Σ is non-compact. Wald (1984, p. 305) focuses on the fact that all known potential counterexamples seem to violate strong causality on the Cauchy horizon $H^+(\Sigma)$. So his version of the vacuum cosmic censorship conjecture comes out as: the (Censorship) condition is true for the collection $(Str) \subset \mathscr{U}$ of all strongly causal spacetimes. This formulation is a bit conservative since all known potential counterexamples not only violate strong causality on the Cauchy horizon $H^+(\Sigma)$ but also the weaker distinguishing condition. So in the spirit of Wald's approach, one could formulate a general version of the cosmic censorship conjecture as: the (Censorship) condition is true for all $\mathscr{P} \subset \mathscr{U}$ such that $\mathscr{P} \subset (Dist)$.

More than forty years on, Wald's formulation of the cosmic censorship conjecture is still open which highlights its "enduring significance" (Lesourd and Minguzzi, 2022, p. 2). Because there exist extensions to Misner spacetime satisfying the causality condition (one is depicted in Figure 12.11), we know that not only is (Censorship) false for the collection $\mathscr{U}$ but also for any $\mathscr{P} \subset \mathscr{U}$ such that $(Caus) \subseteq \mathscr{P}$. This includes the collection $(Chron)$ of all chronological spacetimes. Finally, we say a word about the collection (GH) of all globally hyperbolic spacetimes. We have already seen a number

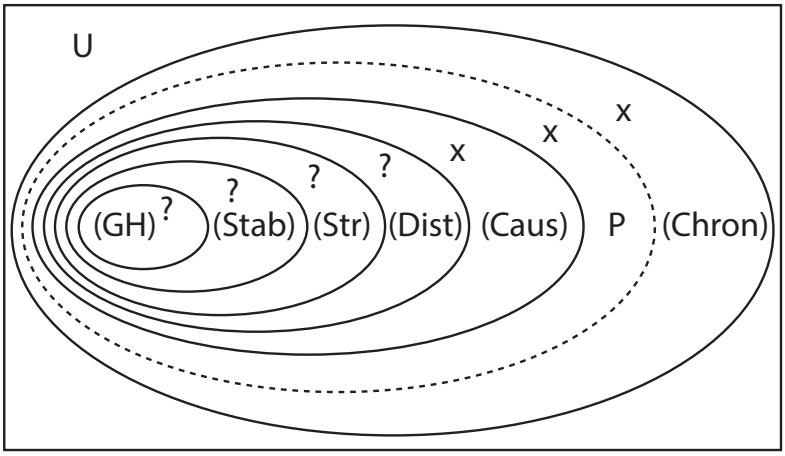

Figure 12.12: Any collection $\mathscr{P} \subseteq \mathscr{U}$ such that $(Caus) \subseteq \mathscr{P}$ renders (Censorship) false. It is unknown if any subcollection of $(Dist)$ satisfies (Censorship).

of examples of maximal developments that can be extended by a globally hyperbolic spacetime (recall Figures 12.4, 12.8, and 12.9). The move to consider only suitable initial data sets excludes all of these examples, and presumably ensures that (Censorship) is true for the collection (GH). But since this does not follow easily, we will consider the question open (see Figure 12.12).

What about local properties? Because the Taub-NUT and Misner examples are members of the collection (Vac) of all vacuum solutions, we see that (Censorship) will be false for any collection $\mathscr{P} \subseteq \mathscr{U}$ such that $(Vac) \subseteq \mathscr{P}$. This includes all the collections defined via the various local energy conditions: (DEC), (SEC), (WEC), (NEC) (see Figure 12.13).

One way to rule out potential counterexamples to cosmic censorship would be to focus on their special status, especially with respect to symmetries. Perhaps there are no "generic" violations to the conjecture? Unfortunately, "it is difficult to give a precise definition of the term "generic" (Wald, 1984, p. 304). But suppose a $\mathscr{U}$-suitable initial set (Σ, h, π) is such that its associated $\mathscr{U}$-maximal developments is $\mathscr{U}$-extendible. It may be that an appropriate perturbation of (Σ, h, π) will produce a $\mathscr{U}$-suitable initial set (Σ, h', π') resulting in an associated $\mathscr{U}$-maximal development that is $\mathscr{U}$-maximal. Indeed, there are some limited results that show a sense in which this is the case for Misner spacetime (Denaro and Dotti, 2015).

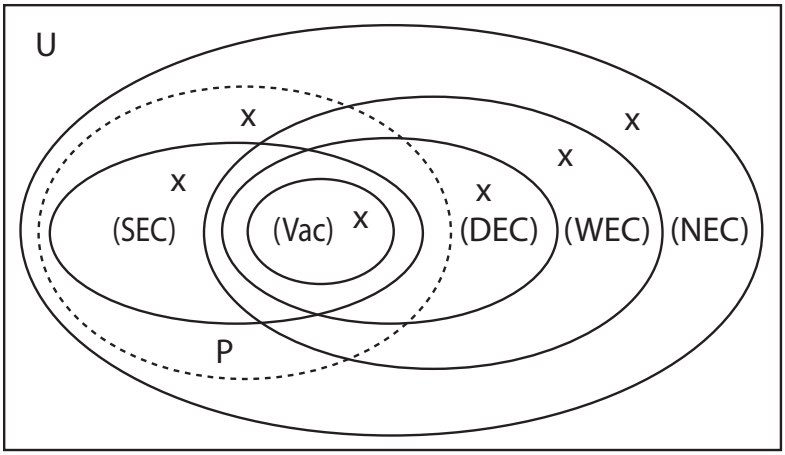

Figure 12.13: Any collection $\mathscr{P} \subseteq \mathscr{U}$ such that $(Vac) \subseteq \mathscr{P}$ renders (Censorship) false. This includes all the collections defined via the various energy conditions: (DEC), (SEC), (WEC), (NEC).

An altogether different – and more general – way of approaching the problem would be to move to a reduced possibility space $\mathscr{P} \subseteq \mathscr{U}$ by requiring some form of spacetime asymmetry. One wonders, for example, if the collection (Her) of all Heraclitus spacetimes satisfies (Censorship). A similar question arises for the collection (LG) of all locally giraffe spacetimes. What about the weaker asymmetry conditions? Recall that (Determinism) is false for any collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$ where (PR), (FP), and (Gir) are the collections of all spacetimes that satisfy, respectively, the point rigid, fixed point, and giraffe asymmetry conditions. Since (Censorship) is a stronger condition than (Determinism), this means that (Censorship) is also false for any collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$ (see Figure 12.14).

12.7 Conclusion

Stepping back, we have seen that dynamical support for a relativized version of the dogma of spacetime maximality must come in two parts: In order to secure $\mathscr{P}$-maximality, a collection $\mathscr{P} \subseteq \mathscr{U}$ must satisfy (Censorship) which presupposes that (Determinism) is already satisfied. None of the

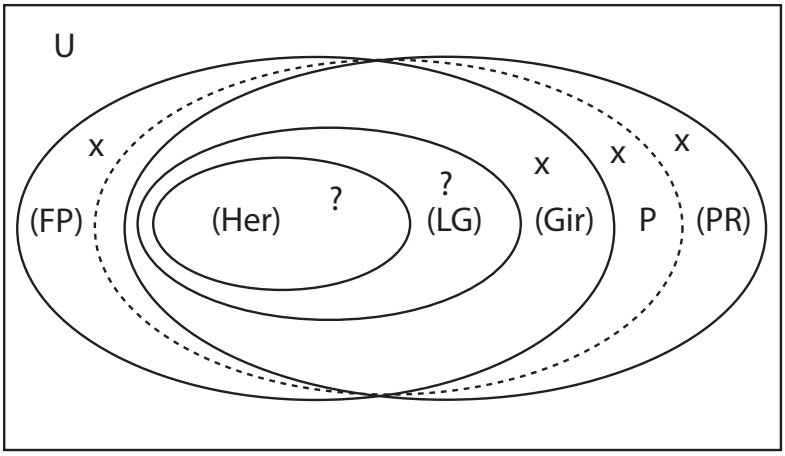

Figure 12.14: It is an open question whether (Censorship) is true for the collections (LG) or (Her). (Censorship) is false for any collection $\mathscr{P}$ such that $(Gir) \subseteq \mathscr{P} \subseteq (PR) \cup (FP)$.

collections under consideration here is known to satisfy (Censorship). The (Determinism) condition is true for all $\mathscr{P} \subseteq \mathscr{U}$ such that $(Vac) \subseteq \mathscr{P}$. This captures a general sense in determinism holds in general relativity. But any such collection $\mathscr{P}$ must necessarily violate (Censorship) since it contains the "physically unreasonable" Taub-NUT and Misner examples. One might try moving to a reduced possibility space $\mathscr{P} \subset (Vac)$. But a version of the subcollection problem then threatens since there are examples of collections $\mathscr{P} \subset (Vac) \cap (GH)$ that fail to satisfy (Determinism) and thus fail to satisfy (Censorship) as well.

Perhaps the best route forward is the one suggested by Wald (1984, p. 305): use a causal condition to find an appropriate "physically reasonable" collection $\mathscr{P}$ that satisfies both (Determinism) and (Censorship). If the causal condition is too weak (e.g. chronology or causality), then (Censorship) is not satisfied. If the the causal condition is too strong (e.g. global hyperbolically), then the significance of (Censorship) becomes trivial: one seeks to show – not assume – that all physically reasonable spacetimes are globally hyperbolic. More promising are the intermediate causal conditions. The collections $(Dist)$, (Str), or $(Stab)$ all satisfy (Determinism) and they may also satisfy (Censorship) as well.

But even if it turns out that one of these collections does satisfy the

(Censorship) condition, spacetime maximality relative to such a collection follows only after a type of circular logic is employed. As we have seen, two forms of spacetime maximality relative to a collection $\mathscr{P}$ are presupposed in order to secure a $\mathscr{P}$-maximality: one form which assumes that an initial data set (Σ, h, π) is $\mathscr{P}$-suitable which requires that (Σ, h) is "as large as it can be" as a Riemannian manifold and a second form in which a dynamical version of the principle of sufficient reason is used to select, among all possible $\mathscr{P}$-developments of (Σ, h, π), the unique $\mathscr{P}$-maximal development. But to rule out a $\mathscr{P}$-development of (Σ, h, π) that fails to be a $\mathscr{P}$-maximal development is "to rule out one way Nature might, consistently with all of the known laws of GTR, continue to evolve things across $H^+(\Sigma)$. What then is to say that She cannot proceed this way?" (Earman, 1995, p. 98).

Chapter 13

Branching

13.1 Introduction

So far, we have followed standard practice by considering the collection $\mathscr{U}$ to be the collection of "all" possible spacetimes. But any number of spacetime conditions could be relaxed so as to move to an expanded possibility space which contains $\mathscr{U}$ as a subcollection. For example, one area of research concerns "spacetimes" (M, g) where the metric g is required to be continuous but not smooth (Dafermos, 2003). Within this context, there is no guarantee that a $\mathscr{U}$-maximal spacetime is also "maximal" relative to the expanded possibility space. But recently, it has been shown that some well-behaved spacetimes (e.g Minkowski, Schwarzchild) do count as maximal even under the more liberal understanding (Sbierski, 2018). A number of maximality questions arise in this framework which will not be explored here. Instead, we shift attention to another common way to expand the collection $\mathscr{U}$: relax the Hausdorff condition to allow for "branching" spacetimes of a certain kind.

In what follows, we will examine non-Hausdorff spacetimes – especially their maximality properties. We will begin with a look at the rationale behind the Hausdorff condition which primarily concerns the preservation of determinism within general relativity (Hajicek, 1971; Earman, 2008). No spacetime is "maximal" if all non-Hausdorff spacetimes are permitted; events can always be pasted onto any given spacetime to construct a larger one. Thus, a collection of not necessarily Hausdorff spacetimes has no hope of satisfying conditions like (Determinism). But we emphasize that if attention is restricted to spacetimes that are permitted to be non-Hausdorff but not

permitted to have a type of "branching curve," then a number of surprising results follow. In particular, we will emphasize that the (Existence) condition is satisfied (Clarke, 1976) and we show that the (Determinism) condition is satisfied as well. Moreover, this expanded possibility space seems to arise "naturally" from considerations of spacetime maximality (Geroch, 1968).

13.2 Why Hausdorff?

Recall that a topological space Hausdorff if there exist disjoint neighborhoods of any distinct points. We have already considered an example of a one-dimensional "branching line" manifold that fails to be Hausdorff (recall Figure 2.8). The example can be easily adapted to construct a non-Hausdorff spacetime. Consider two copies (M_1, η_1) and (M_2, η_2) of two-dimensional Minkowski spacetime in standard (t, x) coordinates. Next, identify the point $(t_1, x_1) \in M_1$ with the point $(t_2, x_2) \in M_2$ if and only if $(t_1, x_1) = (t_2, x_2)$ and $t_1, t_2 < 0$. One can verify that the resulting structure M counts as manifold (Hicks, 1965). A metric η is induced on M in the natural way to produce **branching Minkowski spacetime**. We see that (M, η) is non-Hausdorff since the distinct points $p_1 = (0, x_1)$ and $p_2 = (0, x_2)$ fail to have disjoint neighborhoods if $x_1 = x_2$ (see Figure 13.1).

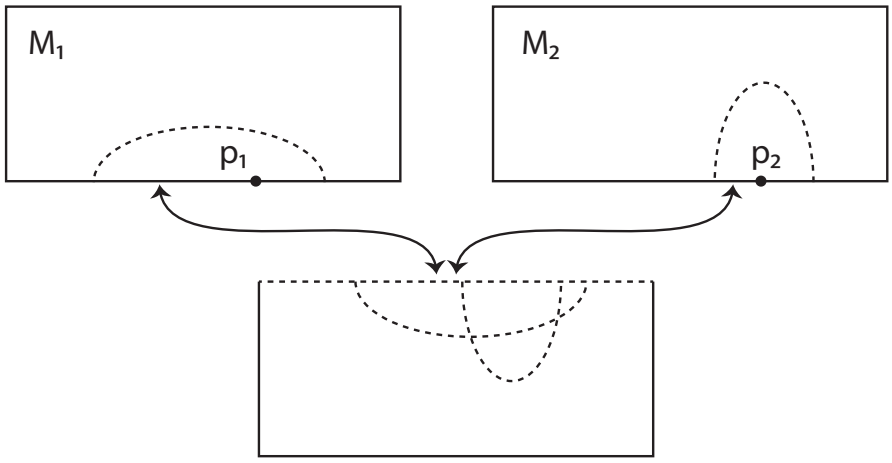

Figure 13.1: Branching Minkowski spacetime is non-Hausdorff since the distinct points p_1 and p_2 fail to have disjoint neighborhoods.

Why is it standard practice to suppose that spacetime is Hausdorff? Most texts provide no justification whatsoever for the condition (Earman, 2008; Luc, 2020). In those texts that do, often only a sentence or two are devoted to the task. We find, for example, the suggestion that a non-Hausdorff spacetime "would perhaps violate what we mean physically by 'distinct events' " (Geroch and Horowitz, 1979, p. 218). But this statement is puzzling since the Hausdorff condition is just one of many separation conditions that one could insist upon. For example, the slightly weaker "T_1" separation condition on a topological space (X, τ) requires that for any point $p \in X$, the set $\{p\}$ is closed. It turns out that all manifolds – Hausdorff or not – satisfy the T_1 condition automatically. Perhaps this is enough to ensure that events are physically distinct? Penrose (1979) has suggested that dropping the Hausdorff condition may actually improve our modeling of spacetime events – especially the physics of time-asymmetry. For him, the pull toward the non-standard picture is sufficiently strong that a mantra must be introduced: "I must therefore return firmly to sanity by repeating to myself three times: 'spacetime is a Hausdorff differentiable manifold; spacetime is a Hausdorff...'!" (Penrose, 1979, p. 595).

Another common reason for excluding non-Hausdorff spacetimes concerns the failure of familiar notions of "determinism" within general relativity (Hajicek, 1971, p. 79). In particular, failing to rule out such spacetimes seems to result in a type of "non-uniqueness of dynamical evolution" (Earman, 2008, p. 201). Ultimately, this indeterminism follows from the innumerable ways in which any spacetime can be "extended" if manifolds are not required to be Hausdorff. We now turn to an investigation of the maximality properties of spacetime within this more permissive context.

Let (NNH) be the collection of all spacetimes that are not necessarily Hausdorff, i.e. spacetimes defined as usual except the Hausdorff requirement is relaxed. In the natural way, we can expand the scope of the second-order maximality conditions we have been considering, e.g. (Determinsm), so as to apply to all collections $\mathscr{P} \subseteq (NNH)$. We begin with an investigation as to whether the collection (NNH) itself satisfies any of these conditions. We find that it generally fails to do so and this captures a sense in which (NNH) has maximality properties quite unlike any spacetime collection we have encountered so far.

13.3 Non-Hausdorff Extendibility

In the natural way, we can generalize the definitions of $\mathscr{P}$-spacetimes, (proper) $\mathscr{P}$-extensions, and $\mathscr{P}$-maximal spacetimes to apply to all collections $\mathscr{P} \subseteq (NNH)$. A number of radical maximality results follow easily from a simple construction that that starts with any spacetime in (NNH) and produces a (NNH)-extension. Let (M, g) be any (NNH)-spacetime (Hausdorff or not) and let p be any point in M. We now construct a spacetime (M', g') which is just like (M, g) except that the point p is "doubled" in a non-Hausdorff way. To do this, just consider two copies (M_1, g_1) and (M_1, g_1) of the spacetime (M, g) and, for any point $q \in M$, let q_1 and q_2 be the associated points in M_1 and M_2 respectively. Now let the non-Hausdorff spacetime (M', g') be the result of identifying q_1 with q_2 for all points $q \neq p$ in M. From this, it follows that any spacetime (M, g) in (NNH) is (NNH)-extendible. So we have a sense in which "there is then no limit to the extent to which additional branches can be grafted onto the space-time" (Clarke, 1976, p. 18).

The fact that there do not exist (NNH)-maximal spacetimes implies that all of the second-order maximality conditions are not satisfied by (NNH) if their scope is appropriately extended. In particular, (Equivalence), (Existence), (Observation), and (Stability) are all easily seen to be false for (NNH). The (Determinism) condition is also not satisfied by (NNH) which implies that (Censorship) is not satisfied as well. To see this, let (M, η) be four-dimensional Minkowski spacetime in standard (t, x, y, z) coordinates. Let (Σ, h, π) be the initial data set induced on the on the $t = -1$ surface Σ and let $((N, g), S, \psi)$ be any (NNH)-development of (Σ, h, π). We see that $((N, g), S, \psi)$ cannot be a (NNH)-maximal development of (Σ, h, π) since it can be properly extended by the (NNH)-development $((N', g'), S, \psi)$ where (N', g') is constructed by taking (N, g) and non-Hausdorffly "doubling" a point p to the future of S as outlined above. One can check that (N', g') counts as a globally hyperbolic spacetime. So (Determinism) is false for (NNH) because of existence problems that arise well before one can even consider the "non-uniqueness of dynamical evolution" (Earman, 2008, p. 201).

We mention here one other sense in which determinism can fail in the non-Hausdorff context. Recall that a spacetime (M, g) is rigid if, for any isometry $f : M \to M$ and any open set $O \subseteq M$, if f acts as the identity on O, then it is the identity map. We have seen that any spacetime in the collection $\mathscr{U}$ counts as rigid (Halvorson and Manchak, 2022). This captures a basic sense

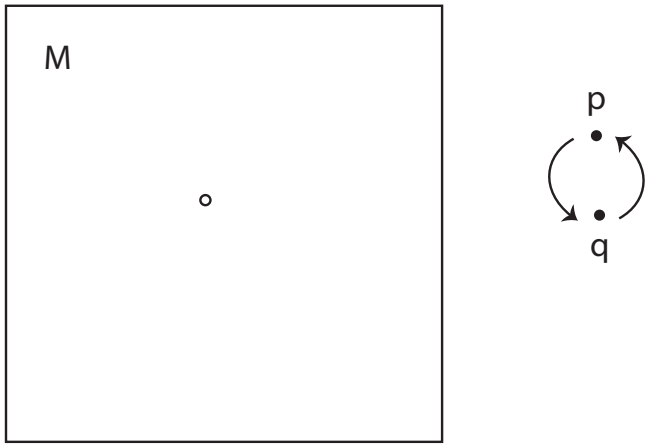

Figure 13.2: The non-Hausdorff spacetime (M, g) fails to be rigid. The map f that acts as the identity on the open region $M - \{p, q\}$ and exchanges the points p and q is a hole-diffeomorphsim and an isometry.

of determinism for the standard collection $\mathscr{U}$ since fixing the symmetries of spacetime in an arbitrary small region fixes them everywhere in a unique way. Another way to put the point: a hole diffeomorphism can never be a spacetime symmetry within the standard context. This all changes when the Hausdorff condition is relaxed. Let (M, g) be any non-Hausdorff spacetime with points $p, q \in M$ that fail to have disjoint neighborhoods. One can show (M, g) fails to be rigid by letting $f : M \to M$ be the map that acts as the identity on the open region $M - \{p, q\}$ and exchanges the points p and q (see Figure 13.2). One can show that this f counts as a hole-diffeomorphsim as well as an isometry (Manchak and Barrett, 2023). So we have another sense in which determinism is satisfied for $\mathscr{U}$ but not for (NNH).

13.4 Bifurcating Curves

We have just seen that the collection (NNH) has some wild extendibility properties. It turns out that there is a natural way to domesticate this collection by moving to a particular subcollection $\mathscr{P} \subset (NNH)$ in which (i) a certain type of non-Hausdorff behavior is still permitted and yet (ii) the maximality properties of $\mathscr{P}$ are very similar to the standard collection $\mathscr{U}$.

Following Hajicek (1971), we say a manifold M has a **bifurcating curve** if there is a pair of (smooth) curves $\lambda_i : [0,1] \to M$ $(i = 1, 2)$ for which $\lambda_1(s) = \gamma_2(s)$ whenever $s \in (0, k)$ and yet $\lambda_1(k) \neq \lambda_2(k)$ for some $k \in (0, 1]$. In the natural way, let us say that a spacetime has a **bifurcating curve** if its underlying manifold does. Any spacetime with a bifurcating curve must necessarily be non-Hausdorff. The branching Minkowski spacetime constructed above is one such example. To see this, take each copy (M_i, η_i) of standard Minkowski spacetime for $i = 1, 2$ and consider the curves $\lambda_i : [0,1] \to M_i$ defined by setting $\lambda(s) = (-1 + 2s, 0)$. When the $t < 0$ regions of (M_1, η_1) and (M_2, η_2) are identified to produce the branching Minkowski spacetime (M, η), we find a bifurcating curve: $\lambda_1(s) = \lambda_2(s)$ for all $s < 1/2$ but $\lambda_1(s) \neq \lambda_2(s)$ for all $s \geq 1/2$ (see Figure 13.3).

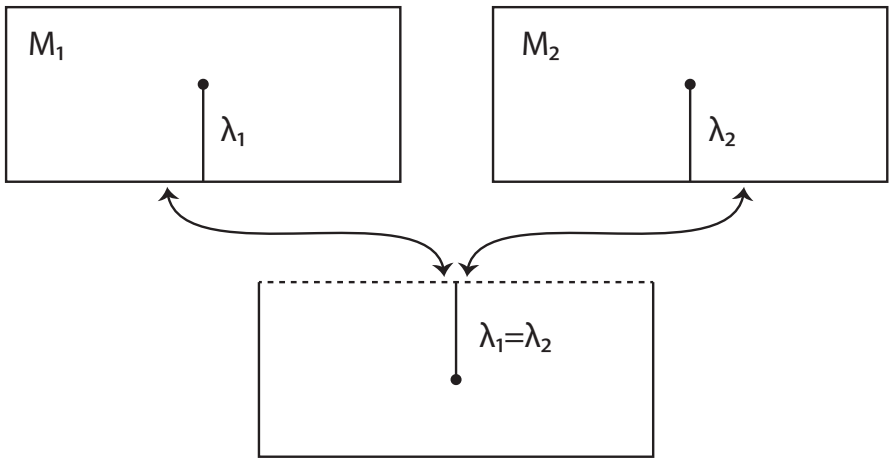

Figure 13.3: The curves curves λ_1 and λ_2 agree in the $t < 0$ portion of branching Minkowski spacetime but diverge thereafter.

Some have suggested that an observer traveling along a birfucating curve "would be very uncomfortable" (Hawking and Ellis, 1973, p. 174). In line with our discussion of determinism in the previous section, others have noted problems for free falling observers: how would they "know which branch of a bifurcating geodesic to follow?" (Earman, 2008, p. 200) Still others maintain that spacetimes with bifurcate curves should not necessarily be considered "too pathological for any physical interpretation" (Miller, 1973, p. 468). In any case, a good deal of problems often associated with non-Hausdorff spacetimes can be eliminated if bifurcating curves are prohibited.

Let $(NBC) \subset (NNH)$ be the collection of all spacetimes with no bifurcating curves. We now show that there are spacetimes in the collection (NBC) but not in the collection $\mathscr{U}$.

We start with Misner spacetime (M_1, g_1) and its "reverse twisted" variant (M_2, g_2) (recall Section 6.6). Here, $M_1 = M_2$ is the cylinder $\mathbb{R} \times S$ in (t, θ) coordinates. Let $O_1 \subset M_1$ and $O_2 \subset M_2$ be the $t < 0$ of regions of each spacetime. Recall that in Misner spacetime (M_1, g_1), one family of complete null geodesics run along the cylinder. In the region O_1 there is another family of incomplete null geodesics that spirals around the cylinder, approaching but never reaching $t = 0$. Similar behavior is exhibited in the reverse twisted Misner variant (M_2, g_2) except that the incomplete null geodesics spiral around the cylinder in the opposite direction. We have seen how there is an isometry $f : O_1 \to O_2$ that maps the twisted null geodesics in O_1 to the untwisted null geodesics in O_2 where they can be extended across $t = 0$. But in the process, this isometry f maps the untwisted null geodesics in O_1 to the twisted null geodesics in O_2 where they cannot be so extended (see Figure 13.4)

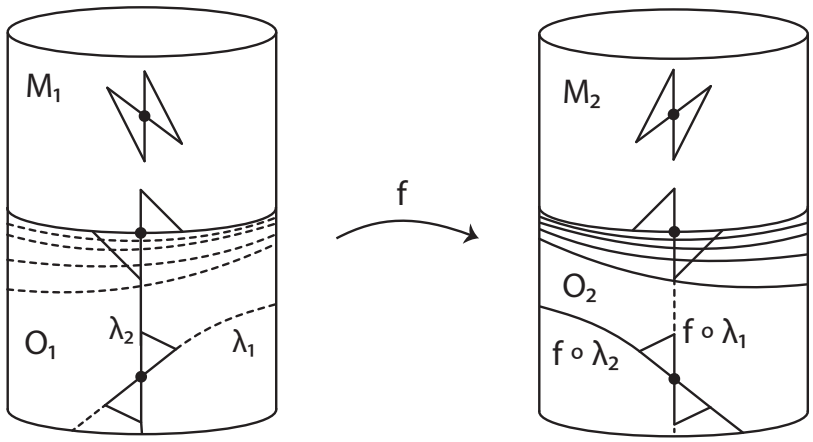

Figure 13.4: The isometry f untwists the null geodesic λ_1 so that the geodesic $f \circ \lambda_1$ can be extended across $t = 0$. But this isometry also maps the null geodesic λ_2 to the twisted geodesic $f \circ \lambda_2$ which cannot be so extended.

If one limits attention to the collection $\mathscr{U}$, there is no way to extend both families of null geodesics across $t = 0$ in either of the Misner spacetime variants. But this can be done if a non-Hausdorff "branching Misner" spacetime is permitted. To construct this model, just identify each point $p \in O_1$ in

Misner spacetime (M_1, g_1) with the point $f(p) \in O_2$ in reverse twisted Misner spacetime (M_2, g_2). The resulting structure (M, g) is non-Hausdorff since one can show that for any point q_1 on the $t = 0$ portion of M_1, there will be a corresponding point q_2 on the $t = 0$ portion of M_2 such that q_1 and q_2 fail to have disjoint neighborhoods (see Figure 13.5). But remarkably, one finds that the branching Misner spacetime has no bifurcating curves (Hawking and Ellis, 1973, p. 174).

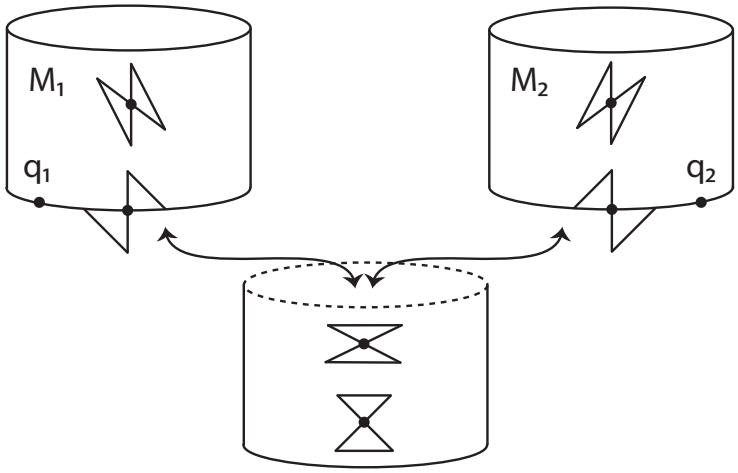

Figure 13.5: The branching Misner spacetime is non Hausdorff since any neighborhoods of the points q_1 and q_2 must overlap in the $t < 0$ region (depicted here in a symmetric way).

13.5 Non-Hausdorff Maximality

Let $(NBC) \subset (NNH)$ be the collection of not necessarily Hausdorff spacetimes that have no bifurcating curves. The non-Hausdorff branching Misner example shows that (NBC) is strictly larger than the collection $\mathscr{U}$. Since (standard) Misner spacetime is $\mathscr{U}$-maximal but can be extended by the branching Misner variant, we see that (Equivalence) is false for the collection (NBC). But as we will now see, the collections (NBC) and $\mathscr{U}$ have quite similar maximality properties in the sense that they give the same verdict for all of the other second-order conditions we have been considering.

We know that some non-Hausdorff spacetimes in the collection (NNH) have underlying manifolds that fail to be second countable. To see this, just consider a version of branching Minkowski spacetime where, instead of just two branches, there are uncountably many. But when attention is restricted to the subcollection (NBC) of spacetimes without bifurcating curves, we find that spacetime manifolds must be necessarily be second countable just as they are in the standard collection $\mathscr{U}$ (Clarke, 1976). This fact (along with Zorn's lemma) allows one to show something remarkable (Clarke, 1976): any (NBC)-extendible spacetime has a (NBC)-maximal extension. This amounts to a significant generalization of the (Geroch, 1970b) existence result that is foundational to the metaphysical justification of spacetime maximality.

We note that just as the $\mathscr{U}$-maximal spacetimes are, in general, highly non-unique, so are (NBC)-maximal spacetimes. For example, Misner spacetime has a wide variety (NBC)-maximal extensions Rieger (2024). One that was first introduced by Geroch (1968, p. 463-464) shows that the branching Misner spacetime considered above is actually (NBC)-extendible. One can construct the (NBC)-maximal extension to this spacetime by pasting in a "top" branch – another copy of the $t < 0$ portion Misner but with opposite time orientation (see Figure 13.6). From the "bottom" branch, a future-directed timelike geodesic can pass through the $t = 0$ boundary of the one of the "side" branches (but not both) depending on which way the bottom branch is twisted up. The geodesic will travel some distance into the side branch before then spiraling around the cylinder back toward the $t = 0$ boundary. If the side branch is then reverse twisted, this spiraling geodesic can again be extended through the $t = 0$ boundary but this time into the top branch. The curious causal structure of this (NBC)-maximal spacetime can perhaps be best understood as the result of removing the origin from two-dimensional Minkowski spacetime and then identifying points in a particular way (Hawking and Ellis, 1973, p. 172-174).

The Clarke (1976) result shows that the collection (NBC) satisfies (Existence). What about the other second-order conditions concerning spacetime maximality? The (Observation) condition comes out as false since the "chain construction" outlined in Section 10.3 can be used to produce, from a given (NBC)-spacetime, an observationally indistinguishable counterpart spacetime that is (NBC)-extendible. Whether the (Stability) condition is satisfied by (NBC) an open question just as it for the collection $\mathscr{U}$. Let us now turn to the (Determinism) and (Censorship) conditions.

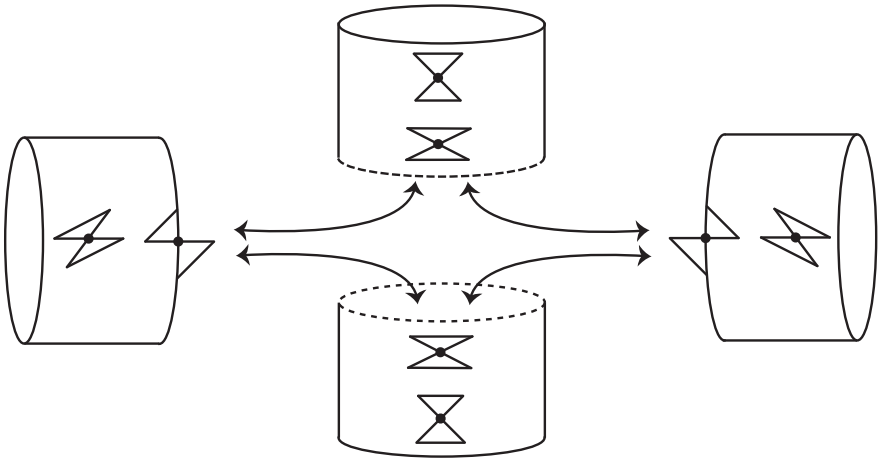

Figure 13.6: A "top" branch is added to branching Misner spacetime to produce an (NBC)-maximal extension.

The branching Minkowski example counts as globally hyperbolic and therefore shows that non-Hausdorff spacetimes in (NNH) can be causally well-behaved. This possibility is closed off when one moves to the subcollection $(NBC) \subset (NNH)$. Indeed, a general result from Clarke (1976) shows that any strongly causal spacetime in (NBC) must be Hausdorff, i.e. it must be in the standard collection $\mathscr{U}$. So we find that the causal misbehavior present in the branching Misner example is representative of all non-Hausdorff spacetimes in (NBC). (One wonders if the result can be generalized: must any distinguishing spacetime in (NBC) also be Hausdorff?)

The Clarke (1976) result can be used to show that (Determinism) is true for (NBC). Let (M, g) be any four-dimensional globally hyperbolic vacuum solution in the collection (NBC) and let Σ be any three-dimensional, connected spacelike surface in M. Since (M, g) is globally hyperbolic, it must be strongly causal. So because this spacetime is in the collection (NBC), we know from the Clarke (1976) result that (M, g) must be Hausdorff. It follows that Σ must also be Hausdorff. Let (Σ, h, π) be the induced initial data set and let $((M', g'), S', \psi')$ be any (NBC)-development of (Σ, h, π). Since $((M', g'), S', \psi')$ is a (NBC)-development, it must be globally hyperbolic and thus strongly causal. So again by the Clarke (1976) result we know that $((M', g'), S', \psi')$ is a $\mathscr{U}$-development of (Σ, h, π). By the Choquet-Bruhat and Geroch (1969) result, there is a $\mathscr{U}$-maximal development $((M'', g''), S'', \psi'')$

of (Σ, h, π) that is unique up to isometry. Since all (NBC)-developments of (Σ, h, π) are $\mathscr{U}$-developments, we know that the $\mathscr{U}$-maximal development $((M'', g''), S'', \psi'')$ cannot be extended by a (NBC)-development, i.e. it is an (NBC)-maximal development for the initial data set (Σ, h, π) that is unique up to isometry. So (Determinism) is true for (NBC) just as it is for $\mathscr{U}$.

Finally, let us consider the (Censorship) condition. As we have seen, the collection $\mathscr{U}$ renders this condition false since it contains the four-dimensional Misner and Taub-NUT spacetimes. Since these spacetimes are in the collection (NBC) as well, we know that the (Censorship) condition is also false for (NBC). Recall the influential formulation of the cosmic censorship conjecture due to Wald (1984) that the collection $(Str) \subset \mathscr{U}$ of all strongly causal spacetimes in $\mathscr{U}$ renders (Censorship) true. Because of the Clarke (1976) result, we see that the collection of all $(Str)* \subset (NBC)$ of all strongly causal spacetimes in (NBC) is such that $(Str)* = (Str)$. So the Wald (1984)formulation of cosmic censorship naturally carries over to the new context.

13.6 Conclusion

In all previous chapters, we have considered the maximality properties of various reduced possibility spaces $\mathscr{P} \subset \mathscr{U}$. Here, for the first time, we have considered enlarged possibility spaces $\mathscr{P} \supset \mathscr{U}$ instead. Two have been the focus: (i) the collection (NNH) which is defined like $\mathscr{U}$ except that the Hausdorff condition on spacetime manifolds is dropped and (ii) the collection (NBC) which is defined like $\mathscr{U}$ except that the Hausdorff condition is replaced with a weaker condition that prohibits bifurcating curves on spacetime manifolds. We have seen that the collection (NNH) of non necessarily Hausdorff spacetimes has maximality properties quite unlike any spacetime collection we have encountered so far. Because any (NNH)-spacetime fails to be (NNH)-maximal, it follows that none of the second-order maximality conditions are satisfied. The fact that conditions like (Existence) and (Determinism) are false for (NNH) gives a sense of the wild nature of this collection. The indeterminism property in particular "has led general relativists to shun non-Hausdorff spacetimes that involve non-Hausdorff branching" (Earman, 2008, p. 200).

But we have also emphasized that non-Hausdorff behavior can be tamed: the move from (NNH) to the collection (NBC) results in maximality prop-

erties very similar to the standard collection $\mathscr{U}$. In particular, the results by Clarke (1976) establish that (Existence) and (Determinism) are both true for (NBC). Moreover, we are now in a position to see how non-Hausdorff behavior can arise naturally from considerations of spacetime maximality within the context of (NBC). Because (Determinism) is true for this collection, we see that initial data associated with the $t < 0$ portion of standard four-dimensional Misner spacetime is guaranteed to have an (NBC)-maximal development (M, g) that is unique up to isometry. This spacetime (M, g) is $\mathscr{U}$-extendible and therefore (NBC)-extendible. Because (Existence) is true for (NBC), we know that (M, g) can be extended to some (NBC)-maximal spacetime (recall Figure 13.6). This spacetime, although non-Hausdorff, can therefore be considered a "natural extension" to (M, g) (Geroch, 1968, p. 465).

Chapter 14

Conclusion

We have noted that work in global spacetime structure is largely an activity of careful collection. Instead of key theorems, the field is characterized by a vast number of smaller results. Such results have limited significance when taken in isolation but can be bundled together to shed light on deep questions. Here, we have engaged in a systematic collection of modal results having to do with the maximality properties of spacetime. From the work primarily done in Part I, we have identified twenty different possibility spaces $\mathscr{P}$ of physical interest. These concern the first-order local, causal, asymmetry, and branching properties of spacetime. The work done in Part II identifies six second-order conditions on such possibility spaces $\mathscr{P}$ that, if satisfied, speak in favor of spacetime $\mathscr{P}$-maximality. These second-order conditions mirror foundational results and conjectures of standard general relativity, e.g. the Geroch (1970b) theorem showing the existence of maximal spacetimes. In this chapter, we collect together what is known and also what is unknown concerning the associated $20 \times 6 = 120$ precise statements concerning the maximality of spacetime.

We begin by listing the six second-order maximality conditions all in one place. We then briefly review their significance and note which of the conditions are satisfied by the collection $\mathscr{U}$ of all standard spacetimes. In the next few sections, we review the basic definitions of the first-order local, causal, asymmetry, and branching properties we have considered. We record which of them satisfy which of the six second-order maximality conditions and note any open questions. A penultimate section reviews a special "subcollection problem" that has reoccured throughout our investigation. The problem shows the various limitations to the following of line of argument: if a given

possibility space $\mathscr{P}$ has certain maximality properties, then those properties automatically "transfer down" to any reduced possibility space $\mathscr{R} \subset \mathscr{P}$. In the final section, we give a general assessment of the results.

14.1 Maximality Conditions

Consider the following six second-order conditions on a collection $\mathscr{P} \subseteq \mathscr{U}$.

(Equivalence) Any $\mathscr{P}$-spacetime is $\mathscr{P}$-maximal if and only if it is $\mathscr{U}$-maximal.

(Existence) Any $\mathscr{P}$-extendible $\mathscr{P}$-spacetime has a $\mathscr{P}$-maximal extension.

(Observation) There are $\mathscr{P}$-spacetimes without god point that are only observationally indistinguishable from $\mathscr{P}$-spacetimes that are $\mathscr{P}$-maximal.

(Stability) The collection of $\mathscr{P}$-maximal spacetimes is C^k stable for some $k \geq 0$ relative to $\mathscr{P}$.

(Determinism) For every $\mathscr{P}$-development of any initial data set, there is a $\mathscr{P}$-maximal development of the same initial data set that is unique up to isometry.

(Censorship) The collection $\mathscr{P}$ satisfies (Determinism) and, in addition, the $\mathscr{P}$-maximal development of any $\mathscr{P}$-suitable initial data set is a $\mathscr{P}$-maximal spacetime.

The (Equivalence) condition captures a sense in which $\mathscr{P}$-maximality is equivalent to $\mathscr{U}$-maximality, i.e. the standard definition of spacetime maximality. It was conjectured by Geroch (1970b) that a number of collections $\mathscr{P}$ satisfy (Equivalence). If so, this would greatly simplify the study of spacetime maximality. On the other hand, if (Equivalence) is false for a collection $\mathscr{P}$, this demonstrates that the modal structure of spacetime works differently

than it does in the standard possibility space $\mathscr{U}$ with respect to spacetime maximality. This means that a careful study of $\mathscr{P}$-maximality must be initiated in order to better understand this property within the possibility space $\mathscr{P}$. Of course, it is trivial that the standard collection $\mathscr{U}$ satisfies (Equivalence).

The (Existence) condition is a generalized statement of the foundational result of Geroch (1970b) which shows that (Existence) is satisfied by the standard collection $\mathscr{U}$. Upon this result rests a general metaphysical justification for the spacetime maximality condition via the Leibnizian principles of sufficient reason and plenitude (Earman, 1989, p. 161). If (Existence) were false for any collection $\mathscr{P}$, this metaphysical justification would face significant difficulties in getting off the ground.

The (Observation) condition captures a weak sense in which observers in some spacetimes in $\mathscr{P}$ can determine that they inhabit a $\mathscr{P}$-maximal spacetime via empirical observations. If (Observation) is not satisfied by a collection $\mathscr{P}$, then (setting aside spacetimes with god point) every observer in every $\mathscr{P}$-spacetime inherits an cosmic underdetermination problem with respect to $\mathscr{P}$-maximality. A conjecture due to Malament (1977b) implies that (Observation) is false for the standard collection $\mathscr{U}$ and this conjecture was later shown to be correct (Manchak, 2009a, 2011).

The (Stability) condition captures a weak sense in which the $\mathscr{P}$-maximality of spacetime is a stable property relative to $\mathscr{P}$. It has been argued that "in order to be physically significant, a property of space-time ought to have some form of stability, that is to say, it should be a property of 'nearby' space-times" (Hawking and Ellis, 1973, p. 197). So if (Stability) is not satisfied by a collection $\mathscr{P}$, then $\mathscr{P}$-maximality would seem to be a physically insignificant property. It is unknown if the standard collection $\mathscr{U}$ satisfies (Stability).

The (Determinism) and (Censorship) conditions provide a type of dynamical justification for the $\mathscr{P}$-maximality of spacetime. The (Determinism) condition is a generalized statement of the foundational result of Choquet-Bruhat and Geroch (1969) which shows that (Determinism) is satisfied by the collection $\mathscr{U}$. If (Determinism) is true for a collection $\mathscr{P}$, then any initial data that can develop into a spacetime in $\mathscr{P}$ must have a "maximal" such development spacetime in $\mathscr{P}$ that is unique up to isometry. This maximal development is not necessarily a $\mathscr{P}$-maximal spacetime but this latter property is guaranteed if the stronger (Censorship) condition is also true for $\mathscr{P}$. The (Censorship) condition is known to be false for the standard collection

$\mathscr{U}$ but the cosmic censorship conjecture of Penrose (1979) holds that the condition will be true for any "physically reasonable" collection $\mathscr{P}$. We see that if either the (Determinism) or (Censorship) were false for a collection $\mathscr{P}$, there would be gaps in the dynamical justification for the $\mathscr{P}$-maximality of spacetime.

14.2 Local Properties

Recall that a spacetime property $\mathscr{P} \subseteq \mathscr{U}$ is **local** if, for any pair of locally isometric spacetimes $(M, g), (M', g') \in \mathscr{U}$, we have $(M, g) \in \mathscr{P}$ if and only if $(M', g') \in \mathscr{P}$. Let (M, g) be a spacetime and let T be its associated energy momentum tensor representing the distribution and flow of matter. A number of local properties of spacetime amount to constraints on T.

The spacetime (M, g) is a **vacuum solution** of Einsten's equation if T vanishes at each point in M. The **weak energy condition** requires that for any timelike vector v at any point $p \in M$, we have $T(v, v) \geq 0$. This requires that the energy density of matter as determined by an observer with tangent v is never negative. The **strong energy condition** is satisfied when a certain effective energy density as determined by any observer is never negative. This requires that "gravitation is attractive" in some sense. The weak and strong energy conditions are independent in the sense that neither implies the other. The **dominant energy condition** can be thought of as prohibiting the flow of matter in a spacelike direction. The dominant and strong energy conditions are independent but dominant does imply weak. All three conditions imply the **null energy condition** which requires that, for any null vector v at any point $p \in M$, we have $T(v, v) \geq 0$. This doesn't have much significance physically but is a simple condition to work with that is useful to have around as a minimal constraint. On the other extreme, the condition of being a vacuum solution is quite strong as it implies all four energy conditions.

Let $(NEC), (WEC), (SEC), (DEC) \subset \mathscr{U}$ be the collections of all spacetimes satisfying, respectively, the null, weak, strong, and dominant energy conditions. Let $(Vac) \subset \mathscr{U}$ be the collection of vacuum solutions. The second-order maximality properties of these first-order local properties of spacetime are summarized in Table 14.1.

	(NEC)	(WEC)	(SEC)	(DEC)	(Vac)
(Equivalence)	X	X	X	X	?
(Existence)	✓	✓	✓	✓	✓
(Observation)	X	X	X	X	X
(Stability)	?	?	?	?	?
(Determinism)	✓	✓	✓	✓	✓
(Censorship)	X	X	X	X	X

Table 14.1: Second-order maximality properties of various first-order local properties of spacetime.

14.3 Causal Properties

Among the global (i.e. non-local) properties of spacetime, the hierarchy causal properties is perhaps most central. Let (M, g) be a time-orientable spacetime. The spacetime is **chronological** if it contains no closed timelike curves (CTCs). This is equivalent to the condition that, for any point $p \in M$, the timelike past $I^-(p)$ does not contain p. The spacetime (M, g) is **causal** if there are no closed causal curves. This means that, for any point $p \in M$, the region $J^+(p) \cap J^-(p)$ is the singleton set $\{p\}$ where $J^+(p)$ and $J^-(p)$ are, respectively, the causal past and future of p. If, for any distinct points $p, q \in M$, the timelike pasts are futures or these points are also distinct, i.e. $I^-(p) \neq I^-(q)$ and $I^+(p) \neq I^+(q)$, then (M, g) is **distinguishing**.

The spacetime (M, g) satisfies **strong causality** if, for each event $p \in M$ and any neighborhood O of p, there is a smaller neighborhood $U \subset O$ of p such that no future-directed causal curve that begins in U and leaves it, ever returns. A spacetime (M, g) satisfies the **stable causality** condition if it admits a global time function, i.e. a smooth function $t : M \to \mathbb{R}$ such that for any distinct points $p, q \in M$, if $p \in J^-(q)$, then $t(p) < t(q)$. This is equivalent to the condition that there is a C^0 fine neighborhood of (M, g) which contains only chronological spacetimes. Finally, the spacetime (M, g) satisfies the **global hyperbolicity** condition, if it is causal and causally compact, i.e. the region $J^+(p) \cap J^-(q)$ is compact for all $p, q \in M$. This is equivalent to the existence of a Cauchy surface, i.e. a closed, achronal set $S \subset M$ such that the domain of dependence $D(S)$ is all of M.

These conditions form a hierarchy of causal properties: global hyperbolicity implies stable causality; stable causality implies strong causality; strong

causality implies distinguishing; distinguishing implies causality; causality implies chronology. None of the implication relations run in the other direction. Let $(Chron), (Caus), (Dist), (Str), (Stab), (GH) \subset \mathscr{U}$ be the collections of spacetimes satisfying, respectively, the chronology, causality, distinguishing, strong causality, stable causality, and global hyperbolicity conditions. The second-order maximality properties of these first-order causal properties of spacetime are summarized in Table 14.2.

	$(Chron)$	$(Caus)$	$(Dist)$	(Str)	$(Stab)$	(GH)
(Equivalence)	?	X	X	X	X	X
(Existence)	✓	✓	?	?	?	?
(Observation)	X	X	X	X	X	?
(Stability)	?	?	?	?	?	?
(Determinism)	✓	✓	✓	✓	✓	✓
(Censorship)	X	X	?	?	?	?

Table 14.2: Second-order maximality properties of various first-order causal properties of spacetime.

14.4 Asymmetry Properties

One would expect that a "generic" spacetime counts as asymmetric various senses. A hierarchy of conditions captures these senses. The conditions turn out to be quite fruitful to consider in discussions of spacetime maximality. Let (M, g) be a spacetime. It is **rigid** if, for any isometry $f : M \to M$ and any non-empty open set $O \subseteq M$, if f acts as the identity on O, then it is the identity map. Every standard spacetime is rigid though violations of rigidity occur in every non-Hausdorff spacetime. The spacetime (M, g) is **point rigid** if, for any point $p \in M$ and any isometry $f : M \to M$, if $f(p) = p$, then f must be identity map. The spacetime has a **fixed point** if, for some point $p \in M$, any isometry $f : M \to M$ is such that $f(p) = p$.

The point rigid and fixed point conditions are independent. They have limited physical significance but prove useful to consider as minimal asymmetry constraints. The conjunction of these two conditions is equivalent to perhaps the most widely considered asymmetry property in the literature: the absence of a non-trivial isometry $f : M \to M$. A spacetime with this property is called **giraffe**. The rigid, point rigid, fixed point, and giraffe

conditions all concern global asymmetries. Two stronger conditions concern local asymmetries. The spacetime (M, g) is **locally giraffe** if given any connected open set $O \subseteq M$, the spacetime (O, g) is giraffe. Any locally giraffe spacetime must be giraffe but nice vice versa. Finally, the spacetime (M, g) is **Heraclitus** if, for any distinct points $p, q \in M$ and any neighborhoods O_p and O_q of these points respectively, there is no isometry $f : O_p \to O_q$ such that $f(p) = q$. In a Heraclitus spacetime, each event is (locally) unlike any other. This condition is equivalent to the requirement that, for any open sets $U, V \subset M$ and any isometry $f : U \to V$, we have $U = V$ and f is the identity map. Any Heraclitus spacetime must be locally giraffe but not vice versa.

Let $(PR), (FP), (Gir), (LG), (Her) \subset \mathscr{U}$ be the collections of spacetimes satisfying, respectively, the point rigid, fixed point, giraffe, locally giraffe, and Heraclitus conditions. The second-order maximality properties of these first-order causal properties of spacetime are summarized in Table 14.3.

	(PR)	(FP)	(Gir)	(LG)	(Her)
(Equivalence)	X	X	X	?	?
(Existence)	?	?	?	?	✓
(Observation)	X	X	X	X	✓
(Stability)	?	?	?	?	?
(Determinism)	X	X	X	?	?
(Censorship)	X	X	X	?	?

Table 14.3: Second-order maximality properties of various first-order asymmetry properties of spacetime.

14.5 Branching Properties

One can also explore the second-order maximality properties of non-standard first-order properties $\mathscr{P}$ that contain $\mathscr{U}$ as a subcollection. One such non-standard property allows for "branching" spacetimes of a certain type. Let (M, g) be a spacetime as defined in the standard way except that the manifold M need not satisfy the usual **Hausdorff** condition, i.e. there exist disjoint neighborhoods of any distinct points $p, q \in M$. A violation of the Hausdorff condition signals the presence of at least two distinct "branches" of a spacetime event. We say a (not necessarily Hausdorff) spacetime (M, g) has a **bifurcating curve** if there is a pair of smooth curves $\lambda_i : [0, 1] \to M$

$(i = 1, 2)$ for which $\lambda_1(s) = \gamma_2(s)$ whenever $s \in (0, k)$ and yet $\lambda_1(k) \neq \lambda_2(k)$ for some $k \in (0, 1]$. If a spacetime has a bifurcating curve, then it is non-Hausdorff but not the other way around.

Let (NNH) be the collection of all spacetimes that are not necessarily Hausdorff and let $(NBC) \subset (NNH)$ be the collection of spacetimes without bifurcating curves. The second-order maximality properties of these first-order branching properties of spacetime are summarized in Table 14.4.

	(NNH)	(NBC)
(Equivalence)	X	X
(Existence)	X	✓
(Observation)	X	X
(Stability)	X	?
(Determinism)	X	✓
(Censorship)	X	X

Table 14.4: Second-order maximality properties of various first-order branching properties of spacetime.

14.6 Subcollection Problem

So far, we have reviewed the second-order maximality properties of 19 first-order properties relating to the local, causal, asymmetry, and branching structures of spacetime. We now consider one final first-order property $(Sub) \subset \mathscr{U}$ concerning the "subcollection problem" that has come up repeatedly throughout our investigation. This problem calls into question the significance of any isolated results obtained so far as well as any that may be secured in the future.

Consider the second-order (Stability) condition as an example. It is unknown whether this condition is satisfied by any of the first-order spacetime properties we have investigated. But suppose (Stability) is true for some spacetime property – say the collection (Vac) of vacuum solutions. This collection (Vac) surely contains "physically unreasonable" spacetimes, e.g. Minkowski spacetime with the word "Leibniz" removed from the spacetime manifold (recall Figure 8.2). So one would like assurance that (Stability) is true not just for the collection (Vac) but also for any reduced possibility $\mathscr{P} \subseteq (Vac)$ including all those that are more "physically reasonable" in some

sense. But we have highlighted that such a general result cannot be. One can artificially construct example subcollections (Vac) that fail to satisfy the (Stability) condition. One could, perhaps, try to be even more restrictive by considering the collection $(Vac) \cap (GH)$ of globally hyperbolic vacuum solutions. Suppose that collection renders (Stability) true. Well, now the subcollection problem manifests itself all over again. The collection $(Vac) \cap (GH)$ surely contains "physically unreasonable" spacetimes, e.g. the $t < 0$ portion of Minkwoski spacetime in which notches have been removed that spell out the word "Leibniz" in Morse code (see Figure 14.1). So one would like assurance that (Stability) is true not just for the collection $(Vac) \cap (GH)$ but also for any reduced possibility space $\mathscr{P} \subseteq (Vac) \cap (GH)$. But we know that such a general result is not possible since we can construct example subcollections of $(Vac) \cap (GH)$ that fail to satisfy the (Stability) condition.

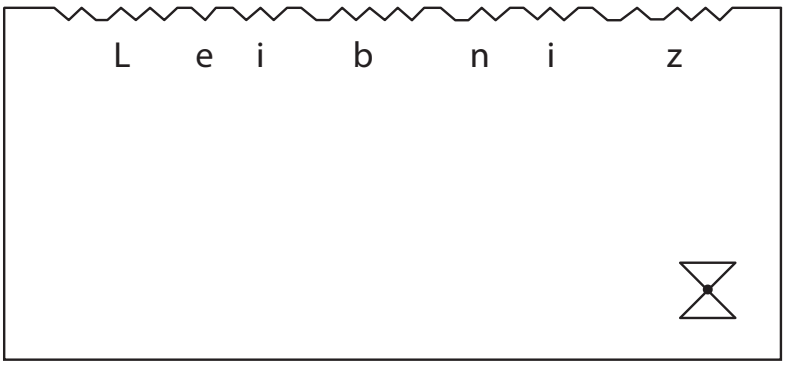

Figure 14.1: A globally hyperbolic vacuum solution in which the removed notches spell out the word "Leibniz" in Morse code.

Stepping back, it would seem that in order to rule out a spacetime like the one in Figure 14.1, one would need to invoke an "no-hole" condition of some kind (e.g. hole-freeness, local maximality, geodesic completeness). All such conditions are at least as strong as $\mathscr{U}$-maximality which, in turn, implies $\mathscr{P}$-maximality. So invoking a no-hole condition is tantamount to requiring $\mathscr{P}$-maximality itself – the very property under investigation. In this way, any future results established concerning (Stability) will come with quite limited significance.

A version of the subcollection problem occurs for each of the six second-order maximality conditions. For each such condition, we have exhibited some subcollection $\mathscr{P} \subset (Vac) \cap (GH)$ which renders the condition false. It is not difficult to show that one can "combine" some of these subcollections to form a single collection $(Sub) \subset (Vac) \cap (GH)$ which renders all six conditions false in one fell swoop. Let $\mathscr{P}_1 \subset (Vac) \cap (GH)$ be the collection mentioned above that shows (Stability) false. This is a collection "rolled up" two-dimensional Minkowski spacetimes which have been truncated in various ways (recall Section 11.7). One can show that $\mathscr{P}_1$ fails to satisfy (Equivalence) and (Observation) as well. Now let $\mathscr{P}_2 \subset (Vac) \cap (GH)$ be the collection of all spacetimes that are the $-k < t < k$ portion of four-dimensional Minkowski spacetime for some positive real number k. This collection that shows (Determinism) false and hence (Censorship) as well. It is easy to see that it doesn't satisfy (Existence) either. Now let $(Sub) \subset (Vac) \cap (GH)$ be the union $\mathscr{P}_1 \cup \mathscr{P}_2$. One can easily verify that this collection (Sub) fails to satisfy all six second-order maximality conditions (see Table 14.5).

	(Sub)
(Equivalence)	X
(Existence)	X
(Observation)	X
(Stability)	X
(Determinism)	X
(Censorship)	X

Table 14.5: Second-order maximality properties of a collection of globally hyperbolic vacuum solutions.

The collection $(Sub) \subset (Vac) \cap (GH)$ surely counts as a "physically unreasonable" possibility space. It amounts to a second-order analog to the "cut and paste" examples that proved to be indispensable tools in the early work on global spacetime structure (Penrose, 1972; Hawking and Ellis, 1973). Within that context, it was widely acknowledged that the constructed examples were not to be considered physically reasonable. Rather, they served another purpose entirely (Geroch and Horowitz, 1979, p. 221):

> "The spacetimes which result from these constructions are, in almost every case, physically unrealistic for various reasons. The

> point of the construction, however, is not normally to construct physically realistic cosmological models, but rather to demonstrate by means of some example that a certain assertion is false, or that a certain line of argument cannot work."

The second-order example $(Sub) \subset (Vac) \cap (GH)$ demonstrates that the following line of argument does not work: if a given possibility space $\mathscr{P}$ has certain maximality properties, then those properties automatically transfer down to any reduced possibility space $\mathscr{R} \subset \mathscr{P}$ (including all those that qualify as "physically reasonable" in some sense).

14.7 Summary of Results

Table 14.6 at the end of this section collects together everything that is known and unknown concerning the six second-order maximality properties of all twenty first-order properties of spacetime. Taken together, the 120 entries seem to indicate a significant lack of clarity with respect to the dogma of spacetime maximality. This lack of clarity comes in three types.

(i) There is a lack of clarity given that 39 precise questions remain unsettled. Of these, most concern the (Stability) condition which ensures that spacetime maximality is a physically significant property in this sense: all spacetimes that are "nearby" a maximal spacetime are also maximal. Very little is known concerning the stability of spacetime maximality. A number of other open questions concern the (Existence) and (Censorship) conditions which are central to the general and dynamical forms of Leibnizian justification for the maximality dogma. One would like to get a better grip on the behavior of various causal and asymmetry properties with respect to these two conditions. The cosmic censorship conjecture of Penrose (1979) has received an enormous amount to attention. Less studied is whether the Geroch (1970b) maximality existence result continues to hold relative to various spacetime properties of interest.

(ii) There is a lack of clarity given that, of the 81 settled questions, 25 results speak in favor of the dogma of spacetime maximality while 56 speak against. Of course, not all questions are all of equal importance. But a murky picture emerges from a number of different angles. Many of the results against the dogma concern the (Equivalence) condition which underscores the need for a careful study of spacetime maximality outside of the standard

context. A pair conjectures of Geroch (1970b) are still open concerning the satisfaction of (Equivalence) by the collection of all vacuum solutions or the collection of all chronological spacetimes. A number of other results against the dogma concern the (Observation) condition showing a generic sense in which observers inherit a cosmic underdetermination problem with respect to spacetime maximality (Malament, 1977b; Manchak, 2009a, 2011).

A person in favor of the dogma might point out that none of these results speak directly to the metaphysical issues that are central to the usual justification for spacetime maximality. But even setting aside the results concerning the (Equivalence) and (Observation) conditions, one still finds a mixed picture. Most properties satisfy the (Determinism) condition showing an analog to the Choquet-Bruhat and Geroch (1969) result. Most properties also satisfy the (Existence) condition showing analogues to the maximality existence result of Geroch (1970b). But most results concerning (Censorship) count against the dogma although there are many open questions including a formulation of the Penrose (1979) cosmic censorship conjecture due to Wald (1984). Recall that (Censorship) is necessary for the dynamical justification for spacetime maximality to go through.

(iii) There is a lack of clarity given that the subcollection problem obscures the significance of the results that we do have. For example, the positive results concerning (Existence) and (Determinism) are all secured relative to possibility spaces that contain "physically unreasonable" spacetimes. The subcollection problem shows that there is no assurance that when one moves to a more appropriate reduced possibility space, similar positive results are maintained. Moreover, this subcollection problem cannot be overcome by restricting attention to highly exclusive local and global properties, i.e. the collection of globally hyperbolic vacuum solutions. Versions of the problem reappear within that context as well. Indeed, one can construct a collection globally hyperbolic vacuum solutions that fails to satisfy all six of the second-order maximally conditions. Invoking a no-hole global spacetime condition is of no help since it is tantamount to imposing spacetime maximality by fiat (Earman, 1995, p. 98).

	(Equivalence)	(Existence)	(Observation)	(Stability)	(Determinism)	(Censorship)
standard collection: $\mathscr{U}$	✓	✓	X	?	✓	X
vacuum solution: (Vac)	?	✓	X	?	✓	X
dominant energy condition: (DEC)	X	✓	X	?	✓	X
strong energy condition: (SEC)	X	✓	X	?	✓	X
weak energy condition: (WEC)	X	✓	X	?	✓	X
null energy condition: (NEC)	X	✓	X	?	✓	X
global hyperbolicity: (GH)	X	?	?	?	✓	?
stable causality: $(Stab)$	X	?	X	?	✓	?
strong causality: (Str)	X	?	X	?	✓	?
distinguishing: $(Dist)$	X	?	X	?	✓	?
causality: $(Caus)$	X	✓	X	?	✓	X
chronology: $(Chron)$	?	✓	X	?	✓	X
Heraclitus: (Her)	?	✓	✓	?	?	?
locally giraffe: (LG)	?	?	X	?	?	?
giraffe: (Gir)	X	?	X	?	X	X
fixed point: (FP)	X	?	X	?	X	X
point rigid: (PR)	X	?	X	?	X	X
no bifurcating curves: (NBC)	X	✓	X	?	✓	X
not necessarily Hausdorff: (NNH)	X	X	X	X	X	X
subcollection: $(Sub) \subset (Vac) \cap (GH)$	X	X	X	X	X	X

Table 14.6: Second-order maximality properties of various first-order properties of spacetime.

Bibliography

Barrett, T., Manchak, J., and Weatherall, J. (2023). On automorphism criteria for comparing amounts of mathematical structure. *Synthese*, 201:191.

Beem, J. (1980). Minkowski space-time is locally extendible. *Communications in Mathematical Physics*, 72:273–275.

Beem, J. and Ehrlich, P. (1981). *Global Lorentzian Geometry*. Marcel Dekker, New York, 1st edition.

Beem, J. and Ehrlich, P. (1987). Geodesic completeness and stability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 102:319–328.

Beem, J., Ehrlich, P., and Easley, K. (1996). *Global Lorentzian Geometry*. Marcel Dekker, New York, 2nd edition.

Belot, G. (2023). *Accelerating Expansion: Philosophy and Physics with a Positive Cosmological Constant*. Oxford University Press, Oxford.

Bernal, A. and Sánchez, M. (2007). Globally hyperbolic spacetimes can be defined as "causal" instead of "strongly causal". *Classical and Quantum Gravity*, 24:745–749.

Bielińska, M. and Read, J. (2023). Testing spacetime orientability. *Foundations of Physics*, 53:8.

Butterfield, J. (2014). On under-determination in cosmology. *Studies in History and Philosophy of Modern Physics*, 46:57–69.

Callender, C. (2017). *What Makes Time Special?* Oxford University Press, Oxford.

Carter, B. (1971). Causal structure in spacetime. *General Relativity and Gravitation*, 1:349–391.

Chen, E. (2024). *Laws of Physics*. Cambridge University Press, Cambridge.

Choquet-Bruhat, Y. and Geroch, R. (1969). Global aspects of the Cauchy problem in general relativity. *Communications in Mathematical Physics*, 14:329–335.

Christodoulou, D. (1994). Examples of naked singularity formation in the gravitational collapse of a scalar field. *Annals of Mathematics*, 140:607–653.

Chruściel, P. and Isenberg, J. (1993). Nonisometric vacuum extensions of vacuum maximal globally hyperbolic spacetimes. *Physical Review D*, 48:1616.

Clarke, C. (1973). Local extensions in singular space-times. *Communications in Mathematical Physics*, 32:205–214.

Clarke, C. (1976). Space-time singularities. *Communications in Mathematical Physics*, 49:17–23.

Clarke, C. (1993). *The Analysis of Space-Time Singularities*. Cambridge University Press, Cambridge.

Curie, P. (1894). Sur la symmétrie des phénomènes physiques: Symmétrie d'un champ electrique et d'un champ magnétique. *Journal de Physique*, 3:393–415.

Curiel, E. (2016). A primer on energy conditions. In D. Lehmkuhl, G. S. and Scholz, E., editors, *Towards a Theory of Spacetime Theories*, pages 43–104. Birkhauser, New York.

Dafermos, M. (2003). Stability and instability of the Cauchy horizon for the spherically symmetric Einstein–Maxwell scalar field equations. *Annals of Mathematics*, 158:875–928.

D'Ambra, G. and Gromov, M. (1991). Lectures on transformation groups: Geometry and dynamics. *Surveys in Differential Geometry*, 1:19–111.

de Sitter, W. (1917). On the curvature of space. *Proceedings of the Royal Academy of Science, Amsterdam*, 19:1217–1225.

Denaro, P. and Dotti, G. (2015). Strong cosmic censorship and Misner spacetime. *Physical Review D*, 92:024017.

Doboszewski, J. (2020). Epistemic holes and determinism in classical general relativity. *British Journal for the Philosophy of Science*, 71:1093–1111.

Duhem, P. (1906). *La Théorie Physique, Son Objet, Sa Structure*. Chevalier et Rivière, Paris.

Earman, J. (1986). *Primer on Determinism*. Reidel, Dordrecht.

Earman, J. (1989). *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time*. MIT Press, Cambridge.

Earman, J. (1995). *Bangs, Crunches, Whimpers, Shrieks: Singularities and Acausalities in Relativistic Spacetime*. Oxford University Press, Oxford.

Earman, J. (2007). Aspects of determinism in modern physics. In Butterfield, J. and Earman, J., editors, *Philosophy of Physics*, pages 1369–1434. Elsevier, Amsterdam.

Earman, J. (2008). Pruning some branches from "branching spacetimes". In Dieks, D., editor, *The Ontology of Spacetime II*, pages 187–205. Elsevier, Amsterdam.

Earman, J. and Norton, J. (1987). What price spacetime substantivalism? the hole story. *British Journal for the Philosophy of Science*, 38:515–525.

Earman, J., Wüthrich, C., and Manchak, J. (2024). Time machines. In Zalta, E. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, URL = https://plato.stanford.edu/archives/sum2024/entries/time-machine/.

Ebin, D. (1968). On the space of Riemannian metrics. *Bulletin of the American Mathematical Society*, 74:1001–1003.

Ehrlich, P. (2006). A personal perspective on global Lorentzian geometry. In Frauendiener, J., Giulini, D., and Perlick, V., editors, *Analytical and Numerical Approaches to Mathematical Relativity*, pages 3–32. Springer-Verlag, Berlin.

Einstein, A. (1949). Einstein's reply to criticisms. In Schilpp, P., editor, *Albert Einstein: Philosopher-Scientist*, pages 665–688. Cambridge University Press, Cambridge.

Ellis, G. (1975). Limits to verification in cosmology. *Quarterly Journal of the Royal Astronomical Society*, 16:245–264.

Ellis, G. (2007). Issues in the philosophy of cosmology. In Butterfield, J. and Earman, J., editors, *Philosophy of Physics*, pages 1183–1285. Elsevier, Amsterdam.

Ellis, G. and Schmidt, B. (1977). Singular space-times. *General Relativity and Gravitation*, 8:915–953.

Fletcher, S. (2016). Similarity, topology, and physical significance in relativity theory. *British Journal for the Philosophy of Science*, 67:365–389.

Fletcher, S., Manchak, J., Schneider, M., and Weatherall, J. (2018). Would two dimensions be world enough for spacetime? *Studies in History and Philosophy of Modern Physics*, 63:100–113.

Friedrich, H. and Rendall, A. (2000). The Cauchy problem for the Einstein equations. In Schmidt, B., editor, *Einstein's Field Equations and their Physical Implications: Selected Essays in Honor of Jürgen Ehlers*, pages 127–223. Springer-Verlag, New York.

Geroch, R. (1967). Topology in general relativity. *Journal of Mathematical Physics*, 8:782–786.

Geroch, R. (1968). Local characterization of singularities in general relativity. *Journal of Mathematical Physics*, 9:450–465.

Geroch, R. (1969). Limits of spacetimes. *Communications in Mathematical Physics*, 13:180–193.

Geroch, R. (1970a). Domain of dependence. *Journal of Mathematical Physics*, 11:437–449.

Geroch, R. (1970b). Singularities. In Carmeli, M., Fickler, S., and Witten, L., editors, *Relativity*, pages 259–291. Plenum Press, New York.

Geroch, R. (1971a). General relativity in the large. *General Relativity and Gravitation*, 2:61–74.

Geroch, R. (1971b). Space-time structure from a global viewpoint. In Sachs, B., editor, *General Relativity and Cosmology*, pages 71–103. Academic Press, New York.

Geroch, R. (1977). Prediction in general relativity. In Earman, J., Glymour, C., and Stachel, J., editors, *Foundations of Space-Time Theories*, volume VIII, pages 81–93. University of Minnesota Press, Minneapolis.

Geroch, R. (1978). *General Relativity From A to B*. University of Chicago Press, Chicago.

Geroch, R. (2013). *Topology, 1978 Lecture Notes*. Minkowski Institute Press, Montreal.

Geroch, R. and Horowitz, G. (1979). Global structure of spacetimes. In Hawkwing, S. and Israel, W., editors, *General Relativity: An Einstein Centenary Survey*, pages 212–293. Cambridge University Press, Cambridge.

Ghez, A., Morris, M., Becklin, E., Tanner, A., and Kremenek, T. (2000). The accelerations of stars orbiting the Milky Way's central black hole. *Nature*, 407:349–351.

Glymour, C. (1972). Topology, cosmology, and convention. *Synthese*, 24:195–218.

Glymour, C. (1977). Indistinguishable space-times and the fundamental group. In Earman, J., Glymour, C., and Stachel, J., editors, *Foundations of Space-Time Theories*, volume VII of *Minnesota Studies in the Philosophy of Science*, pages 50–60. University of Minnesota Press.

Gödel, K. (1949). An example of a new type of cosmological solutions of einstein's field equations of gravitation. *Reviews of Modern Physics*, 21:447–450.

Hajicek, P. (1971). Causality in non-Hausdorff space-times. *Communications in Mathematical Physics*, 21:75–84.

Halvorson, H. and Manchak, J. (2022). Closing the hole argument. *British Journal for the Philosophy of Science*, forthcoming.

Hawking, S. (1969). The existence of cosmic time functions. *Proceedings of the Royal Society A*, 308:433–435.

Hawking, S. and Ellis, G. (1973). *The Large Scale Structure of Space-Time*. Cambridge University Press, Cambridge.

Hawking, S. and Penrose, R. (1970). The singularities of gravitational collapse and cosmology. *Proceedings of the Royal Society A*, 314:529–548.

Hicks, N. (1965). *Notes on Differential Geometry*. Van Nostrand Reinhold Company, New York.

Hounnonkpe, R. and Minguzzi, E. (2019). Globally hyperbolic spacetimes can be defined without the "causal" condition. *Classical and Quantum Gravity*, 36:4109–4129.

Kerr, R. (1963). Gravitational field of a spinning mass as an example of algebraically special metrics. *Physical Review Letters*, 11:237–238.

Kervaire, M. (1960). A manifold which does not admit any differentiable structure. *Commentarii Mathematici Helvetici*, 34:257–270.

Krasnikov, S. (2009). Even the Minkowski space is holed. *Physical Review D*, 79:124041.

Krasnikov, S. (2014). Corrigendum: No time machines in classical general relativity. *Classical and Quantum Gravity*, 31:079503.

Landsman, K. (2021). Singularities, black holes, and cosmic censorship: A tribute to roger penrose. *Foundations of Physics*, 51:42.

Lee, J. (2013). *Introduction to Smooth Manifolds*. Springer, New York.

Lerner, D. (1973). The space of lorentz metrics. *Communications in Mathematical Physics*, 32:19–38.

Lesourd, M. and Minguzzi, E. (2022). Low regularity extensions beyond Cauchy horizons. *Classical and Quantum Gravity*, 39:065007.

Low, R. (2012). Time machines, maximal extensions and zorn's lemma. *Classical and Quantum Gravity*, 29:097001.

Luc, J. (2020). Generalised manifolds as basic objects of general relativity. *Foundations of Physics*, 50:621–643.

Malament, D. (1977a). The class of continuous timelike curves determines the topology of spacetime. *Journal of Mathematical Physics*, 18:1399–1404.

Malament, D. (1977b). Observationally indistinguishable space-times. In Earman, J., Glymour, C., and Stachel, J., editors, *Foundations of Space-Time Theories*, volume VIII of *Minnesota Studies in the Philosophy of Science*, pages 61–80. University of Minnesota Press, Minneapolis.

Malament, D. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*. University of Chicago Press, Chicago.

Manchak, J. (2009a). Can we know the global structure of spacetime? *Studies in History and Philosophy of Modern Physics*, 40:53–56.

Manchak, J. (2009b). Is spacetime hole-free? *General Relativity and Gravitation*, 41:1639–1643.

Manchak, J. (2011). What is a physically reasonable spacetime? *Philosophy of Science*, 78:410–420.

Manchak, J. (2016a). Epistemic "holes" in space-time. *Philosophy of Science*, 83:265–276.

Manchak, J. (2016b). Is the universe as large as it can be? *Erkenntnis*, 81:1341–1344.

Manchak, J. (2016c). On Gödel and the ideality of time. *Philosophy of Science*, 83:1050–1058.

Manchak, J. (2017). On the inextendibility of space-time. *Philosophy of Science*, 84(1215-1225).

Manchak, J. (2018). Some 'no hole' spacetime properties are unstable. *Foundations of Physics*, 48:1539–1545.

Manchak, J. (2021). General relativity as a collection of collections of models. In Madarász, J. and Székely, G., editors, *Hajnal Andréka and István Németi on Unity of Science: from Computing to Relativity Theory Through Algebraic Logic*, pages 409–425. Springer Nature, Cham.

Manchak, J. (2023). On the (in?)stability of spacetime inextendibility. *Philosophy of Science*, 90:1331–1341.

Manchak, J. and Barrett, T. (2023). A hierarchy of spacetime symmetries: Holes to Heraclitus. *British Journal for the Philosophy of Science*, forthcoming.

Manchak, J. and Barrett, T. (2024). Heraclitus-maximal worlds. *Journal of Philosophical Logic*, 53:1519–1536.

Miller, J. (1973). Global analysis of the Kerr–Taub–NUT metric. *Journal of Mathematical Physics*, 14:486–494.

Milnor, J. (1956). On manifolds homeomorphic to the 7-sphere. *Annals of Mathematics*, 64:399–405.

Minguzzi, E. (2012). Causally simple inextendible spacetimes are hole-free. *Journal of Mathematical Physics*, 53:062501.

Minguzzi, E. (2019). Lorentzian causality theory. *Living Reviews in Relativity*, 22:3.

Minkowski, H. (1908). Die grundgleichungen für die elektromagnetischen vorgänge in bewegten körpern. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1908:53–111.

Misner, C. (1967). Taub-NUT space as a counterexample to almost anything. In Ehlers, J., editor, *Relativity Theory and Astrophysics I: Relativity and Cosmology*, pages 160–169. American Mathematical Society, Providence.

Mounoud, P. (2015). Metrics without isometries are generic. *Monatshefte für Mathematik*, 176:603–606.

Newman, E., Tamburino, L., and Unti, T. (1963). Empty-space generalization of the Schwarzschild metric. *Journal of Mathematical Physics*, 4:915–923.

Norton, J. (2011). Observationally indistinguishable spacetimes: A challenge for any inductivist. In Morgan, G., editor, *Philosophy of Science Matters: The Philosophy of Peter Achinstein*, pages 164–176. Oxford University Press, Oxford.

O'Neill, B. (1983). *Semi-Riemannian Geometry with Applications to Relativity*. Academic Press, London.

Penrose, R. (1972). *Techniques of Differential Topology in Relativity*. SIAM, Philadelphia.

Penrose, R. (1979). Singularities and time-asymmetry. In Hawking, S. and Israel, W., editors, *General Relativity: An Einstein Centenary Survey*, pages 581–638. Cambridge University Press, Cambridge.

Penrose, R. (1999). The question of cosmic censorship. *Journal of Astrophysics and Astronomy*, 20:233–248.

Rieger, N. (2024). Topologies of maximally extended non-Hausdorff Misner space. https://arxiv.org/abs/2402.09312.

Rindler, W. (1956). Visual horizons in world-models. *Monthly Notices of the Royal Astronomical Society*, 116:662–677.

Sbierski, J. (2016). On the existence of a maximal Cauchy development for the einstein equations: a dezornification. *Annales Henri Poincaré*, 17:301–329.

Sbierski, J. (2018). The C0-inextendibility of the schwarzschild spacetime and the spacelike diameter in Lorentzian geometry. *Journal of Differential Geometry*, 108:319–378.

Sbierski, J. (2024). Uniqueness and non-uniqueness results for spacetime extensions. *International Mathematics Research Notices*, 20:13221–13254.

Schoen, R. and Yau, S. (1983). The existence of a black hole due to condensation of matter. *Communications in Mathematical Physics*, 90:575–579.

Schwarzschild, K. (1916). Über das gravitationsfeld eines massenpunktes nach der einsteinschen theorie. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 7:189–196.

Smeenk, C., Arntzenius, F., and Maudlin, T. (2023). Time travel and modern physics. In Zalta, E. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, URL = https://plato.stanford.edu/archives/spr2023/entries/time-travel-phys/.

Smeenk, C. and Wüthrich, C. (2021). Determinsim and general relativity. *Philosophy of Science*, 88:638–664.

Sunada, T. (1985). Riemannian coverings and isospectral manifolds. *Annals of Mathematics*, 121:169–186.

Taub, A. (1951). Empty space-times admitting a three parameter group of motions. *Annals of Mathematics*, 53:472–490.

Wald, R. (1984). *General Relativity*. University of Chicago Press, Chicago.

Weatherall, J. (2018). Regarding the 'hole argument'. *British Journal for the Philosophy of Science*, 67:329–350.

Willard, S. (1970). *General Topology*. Addison-Wesley, Reading.

Williams, P. (1984). *Completeness and Its Stability on Manifolds with Connection*. PhD thesis, University of Lancaster.

Wong, W. (2013). A comment on the construction of the maximal globally hyperbolic Cauchy development. *Journal of Mathematical Physics*, 54:113511.

Yodzis, P., Seifert, H., and Müller zum Hagen, H. (1973). On the occurrence of naked singularities in general relativity. *Communications in Mathematical Physics*, 34:135–148.