Causality and Black Holes

Jerić Miloš, Duje

Sveučilište u Splitu
Prirodoslovno-Matematički Fakultet

Causality and Black Holes

Duje Jerić-Miloš

Split, November 2022

Temeljna dokumentacijska kartica

Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Odjel za fiziku
Ruđera Boškovića 33, 21000 Split, Hrvatska

Diplomski rad

Kauzalnost i crne rupe

Duje Jerić-Miloš

Sažetak

U ovom radu je objašnjena osnovna teorija kauzalnosti i crnih rupa. Originalnih rezultata nema, cilj je pedagoške naravi. Glavni teoremi oko kojih je rad napisan su: Birkhoffov teorem, Hawkingov i Penroseov teorem o singularitetu, Gerochov teorem o karakterizaciiji globalno hiperbolnog prostor-vremena te konačno Hawking-Carter-Robinsonov teorem o jedinstvenosti rješenja koja sadrže crne rupe. Pokušao sam dati dokaze svih spomenutih teorema (izuzev teorema o jedinstvenosti, koji je prekompleksan za ovakav rad).

Ključne riječi: Prostor-vrijeme, Birkhoffov teorem, Singularitet, Penroseov teorem, Globalno hiperbolično, Kozmička cenzura,

Rad sadrži: 82 stranica, 2 slike, 0 tablica, 54 literaturna navoda. Izvornik je na engleskom jeziku.

Mentor: prof. dr. sc. Zvonimir Vlah

Ocjenjivači: prof. dr. sc. Marko Kovač, prof. dr. sc. Nikola Godinović

Rad prihvaćen: 6. studenoga, 2022

Basic documentation card

University of Split
Faculty of Science
Physics department
Ruđera Boškovića 33, 21000 Split, Croatia

Diploma thesis

Causality and Black Holes

Duje Jerić-Miloš

Abstract

In this thesis we explain some elementary theory of causality and black holes. No original results are obtained; the goals are pedagogical. The main theorems we shall discuss include: Birkhoff's theorem, Hawking's and Penrose's singularity theorems, Geroch's theorem about globally hyperbolic spacetimes and, finally, Hawking-Carter-Robinson black hole uniqueness theorem (more popularly called the "no-hair theorem"). I have tried to provide self-contained proofs of all the theorems mentioned (except the black hole uniqueness theorem, which is far beyond the scope of this work).

Keywords: Spacetime, Birkhoff's theorem, Singularity, Penrose's theorem, Globally hyperbolic, Cosmic censorship, No-hair theorem

Thesis consists of: 82 pages, 2 figures, 0 tables, 54 references. The original is in English.

Mentor: prof. dr. sc. Zvonimir Vlah

Reviewers: prof. dr. sc. Marko Kovač, prof. dr. sc. Nikola Godinović

Thesis accepted: November 6, 2022

Contents

Introduction

General relativity has, so far, undoubtedly been the most successful theory describing the phenomenon of gravity. Nevertheless, there are still solutions permitted by the theory that one should probably not take too seriously. For example, the interior solution of the Kerr black hole would allow for time travel to both future *and* past. One is therefore tempted to simply discard the interior solution. This, however, is not ideal for the reason that some trajectories in the exterior solution eventually enter the interior. We are thus interested in a maximal solution of sorts, so arbitrarily cutting off a piece of spacetime should not be permitted.

Traditionally, to reach "all possible regions" allowed by the solution, one does an "analytic extension"[1]. It is this process that then leads to parallel universes and regions that generally violate causality. But, it is not clear why one should analytically extend anything - it is the Einstein equations that we are interested in, not analyticity.

A more satisfactory resolution to the problem is provided by the notion of *global hyperbolicity*. This is, roughly speaking, the type of spacetime one can get by evolving some initial data (via the Einstein equations). Here causality holds in its strongest form and no time travel or parallel universes are allowed.

In globally hyperbolic spacetimes, one then proves the singularity theorems which guarantee that some trajectories end in finite time (if certain energy conditions hold). This is relevant to cosmology and black hole physics, as it guarantees that singularities which form under very symmetric conditions remain there even under slight deviations from that symmetry.

It is important to keep in mind that global hyperbolicity is one of the assumptions of the singularity theorems, so trajectories may be extended past the singularities in certain cases, but the extended spacetime then must fail to be globally hyperbolic. The key point is the following: even though some trajectories might exit the globally hyperbolic region (as in e.g. the Kerr solution), Einstein equations tell us nothing about their fate afterwards. It would be nice then (for the theory) if, generically speaking (i.e. for most initial data), the trajectories indeed just end like the singularity theorems tell us they do. All other cases are then to be considered too special to be realistic. This is essentially the philosophy behind the strong cosmic censorship conjecture.

We should mention though, that not all spacetimes a physicist might find useful are covered by globally hyperbolic ones (anti-de Sitter space being one prominent example). Nevertheless, one may take this as the most conservative starting point.

The thesis may be roughly divided into four parts; each culminates in some *Hauptsatz*:

[1]It should immediately be mentioned that extensions are not necessarily unique (so "all possible regions" is really ill defined). Sometimes though, one can find the unique maximal analytic extension of some spacetime, but that doesn't mean it is unique among the smooth ones.

1. First part is concerned with symmetries and culminates in the proof of Birkhoff's theorem. This result shows that spherical symmetry restrict the possible set of solutions drastically.

2. The second part is concerned with causality and global hyperbolicity. Here we prove a characterization of global hyperbolicity due to Geroch.

3. The third part is concerned with incompleteness and singularities, which culminate with Hawking's and Penrose's singularity theorems.

4. The fourth part is concerned with the so-called "no hair theorem", but I have opted to forgo the proofs here, as they tend to get quite involved.

Next, in order to make the main text more accessible and self-contained, let us introduce some preliminary notions having to do with basic differential geometry.

Topology

A **topology** on a set X is a collection τ of its subsets which are closed under unions and finite intersections (i.e. for $U_i \in \tau$, $\bigcup_i U_i \in \tau$ and for $U, V \in \tau$, $U \cap V \in \tau$). One also requires that X and $\emptyset$ be in τ. The idea is to capture a notion of an open set (also called an open neighborhood). Since X is *a priori* only an amorphous collection of points, designating certain sets as open allows us to establish which points are close to each other. The smaller the open set which contains both points, the better we know the relative position of those points.

Consider $\mathbb{R}^n$ with open balls $B_r(x) = \{y \in \mathbb{R}^n \mid |y - x| < r\}$. Taking arbitrary unions of such sets (for different x and r) gives a topology on $\mathbb{R}^n$ - the standard euclidean topology.

To illustrate how topology reflects relative nearness of points, take one point from $B_{1/n}(x)$ for each $n \in \mathbb{N}$. We thus get a sequence x_n with a following property: no matter how small of an open set around x we take, we will always find all but finitely many chosen points in there. We say x_n **converges** to x.

As another example take the notion of a **limit point**: x is a limit point of set A if any open set containing x intersects A. x need not be in A, just very close to it (consider some point of norm 1 and the open ball $B_1(0)$).

One can also put a discrete topology on $\mathbb{R}^n$ by proclaiming *every* subset of $\mathbb{R}^n$ be open. Now, this isn't terribly interesting because sets containing only one point are open as well, so we can perfectly distinguish points. Thus no sequence can converge to x but the eventually constant ones (those equal to x after a certain point) and every limit point of A must be contained within A.

A set is **closed** if it contains all of its limit points. It is not hard to show that closed sets are precisely complements of open sets. One gets the **closure** of A, which we write as $\overline{A}$, by adding all limit points of A to A. **Interior** is the largest open set contained within A, and we may define the **boundary** of A, denoted by ∂A, as the difference between the closure and the interior.

Every subset A of a topological space X becomes a topological space in its own right by considering $U \cap A$ to be open in A for any U open in X.

A map $\varphi : X \to Y$ between two topological spaces is **continuous** if it maps nearby points in X to nearby points in Y. More precisely, if for any (no matter how small) open

neighborhood U around $f(p)$, $f^{-1}(U)$ is open neighborhood of p. A continuous function with a continuous inverse is called a **homeomorphism**.

Compactness and Connectedness

A subset K of a topological space is said to be **compact** if every open cover of K has a finite subcover. Thus, for example, if one can prove some property P in a neighborhood around any point of K, compactness guarantees the existence of a finite cover whose open sets also satisfy P. It is not difficult to show that any infinite set in compact K must have a limit point. Intuitively, infinitely many points in K cannot be spread out homogeneously (they accumulate somewhere). It is also easy to see that a continuous map sends compact sets to compact sets.

A topological space X is said to be disconnected if we can find two disjoint (nonempty) open sets (U, V) whose union is X; if we cannot find such open sets, it is said to be **connected**. Note that U and V are complements of each other so they must be open and closed. Conversely, if the only open and closed sets are the trivial ones (namely X and $\emptyset$), then it is easy to see that such a space must be connected. As an example, any interval in $\mathbb{R}$ is connected.

Continuous function will again map connected sets to connected sets.

A space X is said to be **path connected** if any two points $p, q \in X$ can be connected by some continuous curve $\gamma : [a, b] \to X$. It is readily seen that a path connected set must be connected (the interval is connected, so a separation (U, V) would also separate $\operatorname{Im}\gamma$ if γ goes from $p \in U$ to $q \in V$, giving a contradiction).

For spaces we are interested in (namely manifolds) these two notions of connectedness turn out to actually be equivalent.

Smooth Manifold

By a **smooth manifold** (of dimension n) we mean a space M equipped with **charts**, i.e. bijective mappings $\phi_i : U_i \to \mathbb{R}^n$, where $U_i \subset M$ cover M and $\varphi_i(U_i)$ is an open set in $\mathbb{R}^n$. We can then pull the topology from $\mathbb{R}^n$, by proclaiming each $\varphi_i^{-1}(U)$ to be an open set in M, thereby making each φ_i a continuous map. Furthermore, we require that any transition map $\varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$ be smooth.

This way, if one chart measures a curve $\gamma : [a, b] \to M$ to be smooth (i.e. $\varphi \circ \gamma$ is smooth), then any other chart will do the same. A collection of such charts is (unsurprisingly) called an atlas. We shall require our atlases to be maximal (w.r.t. inclusion).

Charts are usually written as an n-tuple of functions $\phi = (x^1, \ldots, x^n)$, each providing one coordinate to points in $U \subset M$.

We shall generally employ two more restrictions on the global topology of M:

1. M should be a **Hausdorff space**, meaning that any two points $p, q \in M$, $p \neq q$ can be separated by some disjoint open sets $p \in U$, $q \in V$. It is easy to show that this condition guarantees the uniqueness of limits. Note that a euclidean space $\mathbb{R}^n$ is Hausdorff, so a manifold (locally homeomorphic to $\mathbb{R}^n$) will be locally Hausdorff, but will generally fail to be globally Hausdorff. This assumption gets used when proving the existence of maximal flows of vector fields.

2. M should be **second-countable**. This means that there should exist a countable collection of open sets such that any other open set can be obtained as a union of these.

   This property gets used when proving the existence of partitions of unity.

Manifolds generalize surfaces and curves to higher dimensions, but notice that we use an intrinsic definition; we chart the manifold, instead of describing it as an embedded subset of some higher dimensional space.

Tangent Space

A tangent space to a manifold is intuitively obvious; it should generalize a tangent plane at a point of some surface. Things become slightly more complex because we need a way to intrinsically characterize tangent vectors. We do this by considering curves through p:

Let $p \in \mathbb{R}^n$. We say two curves γ and α starting at p ($\gamma(0) = p$ and $\alpha(0) = p$) represent the same direction at p if $\gamma'(0) = \alpha'(0)$ (note that magnitude, i.e. the length of the vector matters as well).

For curves on a manifold M, we say γ and α represent the same direction at p if they represent the same direction on some chart ($(\varphi\circ\gamma)'(0) = (\varphi\circ\alpha)'(0)$). The equivalence class of all curves representing the same direction at p is called a **tangent vector**. Collection of all tangent vectors at p is called a **tangent space** at p, denoted T_pM. One then pulls the vector structure from $\mathbb{R}^n$ and gets a vector space.

Derivative of a Function

Just as with curves, one can check whether a function is smooth by passing to charts: $f : M \to N$ is **smooth** at $p \in M$ if we can find charts φ and ψ around p and $f(p)$ respectively such that $\psi \circ f \circ \varphi^{-1} : \varphi(U) \to \psi(U)$ is smooth (as a map between euclidean spaces). A smooth bijection $f : M \to N$ with a smooth inverse is commonly known as a **diffeomorphism**.

A derivative of a smooth function $f : M \to N$ between smooth manifolds, can now be defined as follows. Since f maps curve γ to curve $f \circ \gamma$, the derivative of f at p should map the tangent vector to γ at p to the tangent vector to $f \circ \gamma$ at $f(p)$. In other words, $df_p : T_pM \to T_{f(p)}N, \quad df_p(\gamma'(0)) = (f \circ \gamma)'(0)$. One then checks this is well defined (i.e. does not depend on the choice of curve γ representing a tangent vector $v \in T_pM$).

In particular, if $\varphi = (x^1, \ldots, x^n)$ is a chart containing p, there is a standard basis on T_pM induced by this chart. Note that, since φ is a bijection, $d\varphi$ is a bijection as well, thus we can find a basis on T_pM which $d\varphi_p : T_pM \to \mathbb{R}^n$ maps to the standard basis $e_i = (0, \ldots, 1, \ldots 0)$. These are coordinate vectors (or coordinate frame) of this chart.

Tangent Bundle

We can now collect the tangent spaces into one object called the **tangent bundle** $TM = \bigcup_{p\in M} T_pM$. It can be shown that TM itself has the structure of a smooth manifold, its charts being induced by the coordinate vectors of some chart on M: if $e_i \in T_pM$ denotes the coordinate vectors and $X_p = \sum_i a^i e_i$, then $\phi(X_p) = (\varphi(p), a^1, \ldots a^n)$ defines a chart for TM.

Vector fields are *sections* of the tangent bundle, i.e. mappings $X : M \to TM$, such that $X_p = X(p) \in T_pM$. The space of all vector fields on M will be denoted by $\Gamma(TM)$, or $\mathfrak{X}(M)$.

Note that a vector field may be identified with a differential operator, namely with $f \mapsto df(X)$ (directional derivative of f in direction X). We often write just Xf. One can even *define* tangent vectors as differential operators on functions obeying linearity and Leibniz rule; but to prove this gives the same construction requires some work. For vector fields X and Y, by XY we shall mean the composition of X and Y as differential operators and by $[X, Y] = XY - YX$ their commutator.

The coordinate vectors of some chart now simply become standard partial derivatives on that chart: ∂_i.

An integral curve of vector field X is a curve γ solving $\gamma'(t) = X_{\gamma(t)}$. A (local) **flow** $\varphi : U \times \mathbb{R} \mapsto M$ of vector field X parametrizes the integral curves by their initial positions: $\varphi_t(p) = \gamma_p(t)$, where γ_p is the integral curve through p.

Dual Space

Recall that a dual of a real vector space V is the space of all of its linear functionals $f : V \to \mathbb{R}$, which we denote by V^*. If e_i is a basis on V, we can define the dual basis $e^i \in V^*$ by $e^i(e_j) = \delta^i_j$, in other words, if $v = \sum_i v^i e_i$, then $e^i(v) = v^i$. We write coordinates of dual vectors in the basis e^i using lower indices: $v = \sum_i v_i e^i \in V^*$. This is the essence of index notation. When the same index appears both as a subscript and a superscript, we shall omit the summation sign. Thus, for example $v = v^i e_i$.

A vector in V may act on V^* as well. Indeed, V^{**} and V are isomorphic: $v(f) = f(v)$.

We can now form the dual bundle $T^*M = \bigcup_{p \in M} T^*_pM$. Sections of this bundle are called covector fields or, more commonly, 1-forms. The space of 1-forms is denoted by $\Gamma(T^*M)$.

Notice that on a chart $\varphi = (x^1, \ldots, x^n)$, dx^i map each $p \in M$ to $dx^i_p : T_pM \to \mathbb{R}$; thus dx^i are 1-forms. It is not difficult to check $dx^i(\partial_j) = \delta^i_j$, so dx^i are dual to the standard coordinate frame.

Metric Tensor

By taking tensor products $TM \otimes TM = \bigcup_{p \in M} T_pM \otimes T_pM$ we get even more objects. A **metric** on M is a smooth mapping $g : M \to TM \otimes TM$ assigning to each point p a scalar product on T_pM. By this we mean a symmetric nondegenerate bilinear map $g_p : T_pM \times T_pM \to \mathbb{R}$. By nondegenerate we mean that if $g(X_p, Y_p) = 0$ for all $X_p \in T_pM$, then $Y_p = 0$.

Non degeneracy can be seen to amount to the map $X_p \mapsto g(X_p, \cdot)$ being an isomorphism between T_pM and T^*_pM (non degeneracy means precisely that this linear map is injective; but V and V^* have the same dimension, so it is bijective).

We shall regularly write ds^2 for the quadratic form $X \mapsto g(X, X)$.

By employing the metric we can lower or raise indices: for a vector field X, define a 1-form $X^\flat = g(X, \cdot)$. The inverse of $\flat : X \mapsto X^\flat$ is denoted by $\sharp$. We then have $X_j = e_j(X^\flat) = X^\flat(e_j) = g(X^i e_i, e_j) = X^i g(e_i, e_j) = X^i g_{ij}$. Thus the matrix g_{ij} maps $(X_1, \ldots, X_n)$ to $(X^1, \ldots, X^n)$. Inverting this relation (and writing g^{ij} for the inverse matrix) we get $g^{ij} X_j = X^i$.

Lorentzian Metric

One can generally prove (via Gram-Schmidt procedure) that any nondegenerate scalar product admits an orthonormal basis, i.e. a basis e_i for which $g(e_i, e_j) = 0$ if $i \neq j$ and $g(e_i, e_i) = \pm 1$. In other words, the matrix of g in this basis is just ± 1 on the diagonal. In fact, a stronger result (so called Silvester's law of inertia) shows that in *any* other orthonormal basis g will have the same number of -1s on the diagonal. Thus the number of -1s is an invariant called the **signature** of the metric (what Silvester called "inertia"). We shall be interested only in metrics of signature $(-, +, \ldots, +)$, i.e. with only one -1. Such metrics are called **Lorentzian**.

For a Lorentzian metric g, we say a vector $X \in T_pM$ is:

1. timelike if $g(X, X) < 0$
2. spacelike if $g(X, X) > 0$
3. null or lightlike if $g(X, X) = 0$

Since in orthonormal coordinates e_i metric g is a diagonal matrix with ± 1 on the diagonal, we see that (in Lorentzian signature) $g(X, X) = -(X^0)^2 + (X^1)^2 + \ldots (X^n)^2$, so $g(X, X) = 0$ defines an hourglass shaped cone in T_pM. Null vectors lie on the cone, while timelike vectors lie (strictly) within the cone. Since the interior of the cone consists of two connected components, we may chose one and label it *future*. Thus future oriented timelike vectors fall into that component and past oriented ones fall into the opposite component.

Physically, particles/observers are constrained to move on trajectories γ whose velocities $\gamma'(t)$ at each t lie within the null cone of $T_{\gamma(t)}M$, i.e. $\gamma'(t)$ are timelike or null and future oriented. This constraint reflects the fact that no signal should travel faster than light (or backwards in time). For a timelike curve $\int_a^b |\gamma'| ds = \int_a^b \sqrt{|g(\gamma', \gamma')|} ds$ is called its proper time. The metric thus provides us with null cones, but also a way of measuring the elapsed time for any given observer (as measured by that observer).

Levi-Civita Connection

Generally one cannot compare a vector $X_p \in T_pM$ and a vector $Y_q \in T_qM$; they do not live in the same vector space. Nevertheless, it would be desirable if we could establish a way of transporting vectors from one tangent space to the other. This is achieved by an **affine connection** (or covariant derivative); this is a mapping $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ satisfying:

1. $\nabla_X Y$ is $\mathcal{F}$-linear in X, a $\mathbb{R}$-linear in Y. In other, words: $\nabla_{fX} Y = f \nabla_X Y$ for all smooth $f : M \to \mathbb{R}$ and $\nabla_X (\alpha Y) = \alpha \nabla_X Y$ for all $\alpha \in \mathbb{R}$, we of course must also have $\nabla_X (Y + Y') = \nabla_X Y + \nabla_X Y'$ and $\nabla_{X+X'} Y = \nabla_X Y + \nabla_{X'} Y$.
2. Leibniz rule holds for ∇ in Y: $\nabla_X (fY) = (Xf)Y + f \nabla_X Y$ for each smooth $f : M \to \mathbb{R}$

We now say Y is parallel if $\nabla_X Y = 0$ for all X.

Consider a vector field along curve $\gamma : [a, b] \to$, i.e. $V : [a, b] \to TM$ such that $V(t) \in T_{\gamma(t)}M$. We cannot directly apply ∇ as V is not defined on M. However, we can

define a unique analogue induced by ∇. Indeed, one shows that there is a unique operator $\frac{D}{dt}$ which is linear, satisfies the Leibniz rule and is also compatible with ∇ in the following sense: if X is a vector field on M, then $X_\gamma(t) = X_{\gamma(t)}$ is a vector field along γ and we must have $\frac{DX_\gamma}{dt}(t) = \nabla_{\gamma'(t)}X$. A field V is parallel along γ if $\frac{DV}{dt} = 0$

Let ∇ be some connection and $\gamma : [a, b] \to M$ a smooth curve on M. For a given $v \in T_{\gamma(a)}M$ there exists a parallel vector field $V : [a, b] \to TM$ along γ with $V(a) = v$. By varying v we get a linear map $\tau : T_{\gamma(a)}M \to T_{\gamma(b)}M$, called the parallel transport map.

For a vector field X we can now show:

$$\frac{DX}{dt} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\tau_\varepsilon^{-1} X(\varepsilon) - X(0) \right)$$

So the intuition is that the covariant derivative looks at how X changes during parallel transport. Generally, we shall write $\nabla_{\gamma'(t)}X$ instead of $\frac{DX}{dt}$.

A celebrated theorem by Riemann states that given a nondegenerate metric g on M, there exists a unique connection ∇, the so-called Riemann or **Levi-Civita connection**, satisfying:

1. ∇ is metrically compatible: $Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y)$ for all $X, Y, Z \in \mathfrak{X}(M)$.

   This is equivalent to parallel transport being an orthogonal transformation (preserving the metric) between tangent spaces.

2. Torsion $T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]$ of ∇ vanishes.

Geodesics

By a geodesic γ we always mean an affinely parameterized one, i.e. the solution to

$$0 = \frac{D\gamma'}{dt} = \nabla_X X,$$

where $X = \gamma'$. Writing this out in some local coordinates, we get:

$$\ddot{y}^k + \sum_{ij} \Gamma_{ij}^k \dot{y}^i \dot{y}^j = 0,$$

where Γ_{ij}^k are Christoffel symbols defined by $\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k$.

Denote by γ_v a geodesic with initial velocity $v \in T_pM$. The **exponential map** at $p \in M$ is a map defined on some open subset of $0 \in T_pM$ by

$$\exp_p(v) = \gamma_v(1).$$

$\exp_p$ sends straight lines in T_pM to geodesics going through p. One actually proves this is a diffeomorphism on a sufficiently small open neighborhood 0. If e_i is a basis on T_pM and r^i are coordinates in that basis, then we may define **normal coordinates**: $x^i = r_i \circ \exp_p^{-1}$.

Curvature Tensors

We define the **Riemann curvature tensor** as

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z.$$

Roughly speaking, if we parallel transport Z along some infinitesimal parallelogram spanned by X, Y, thus going in a loop and returning to the initial tangent space, then $R(X,Y)Z$ is the difference between the transported vector and the original vector Z. More precisely, we have:

$$\lim_{s,t\to 0} \frac{Z_p - TZ_p}{st} = R(X_p, Y_p)Z_p = [R(X,Y)Z](p),$$

where $T = \tilde{\psi}_s^{-1}\tilde{\varphi}_t^{-1}\tilde{\psi}_s\tilde{\varphi}_t$ is the parallel transport along the flows φ_t and ψ_s of vector fields X and Y.

Assume $U \subset M$ is a neighborhood and $e_i \in \mathfrak{X}(U)$ is a local frame on U, i.e. that $(e_i)_p$ form a basis for T_pM at each $p \in U$. We write $R^k_{lij} = (R(e_i, e_j)e_l)^k$, i.e. $R^k_{lij}e_k = R(e_i, e_j)e_l$.

Lowering an index gives $R(W,Z,X,Y) = g(R(X,Y)Z, W)$, i.e. $R_{klij} = g_{mk}R^m_{lij}$.

The **Ricci tensor** is the a contraction of the Riemann tensor: $R_{ij} = R^k_{ikj}$.

Intuitively, the Ricci tensor $\mathrm{Ric}(X,X) = R_{ij}X^iX^j$ measures the difference in volume of a narrow cone of geodesics emanating from p going in the direction X and the volume of the flat (euclidean) cone (in normal coordinates $(x^1, \ldots, x^n)$). More precisely, we have the following formula for the volume form in normal coordinates:

$$\omega = \sqrt{|\det g_{ij}|}dx^1 \wedge \cdots \wedge dx^n = \left(1 - \frac{1}{6}\mathrm{Ric}_p(x,x) + O(|x|^3)\right) dx^1 \wedge \cdots \wedge dx^n.$$

Spacetime

In what follows we will mostly deal with some 4-dimensional connected Lorentzian manifold M with metric g of signature $(-,+,+,+)$ and an induced Levi-Civita connection ∇. We assume further the existence of a global nonvanishing timelike vector field, which gives a smooth choice of future timecone at each T_pM. As usual, we shall refer to M as **spacetime**.

On M matter is represented by some symmetric 2-tensor $T_{\mu\nu}$ called the *stress-energy tensor*. The system evolves according to the **Einstein field equations**:

$$G_{\mu\nu} = 8\pi T_{\mu\nu},$$

where $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$ is the Einstein tensor and $R = R^\mu_\mu$ is the scalar curvature. Alternatively, Einstein field equations may be written as $R_{\mu\nu} = 8\pi\left(T_{\mu\nu} - \frac{1}{2}Tg_{\mu\nu}\right)$, where $T = T^\mu_\mu$. In vacuum $T_{\mu\nu} = 0$, so the Einstein equations reduce to $R_{\mu\nu} = 0$.

Chapter 1

Spacetime Symmetries

Since Einstein's equations are so difficult to solve in general, to make the problem tractable one either linearizes[1] the theory or imposes certain symmetries. Symmetries, in particular, allow us to simplify the metric before solving the Einstein equations. In this chapter we review some elementary notions from group theory, define certain types of spacetime symmetries and prove a result on the splitting of the metric under the action of some (compact connected Lie) group.

1.1 Foliations

An **immersion** is a smooth function $f : M \to N$ whose derivative $df_p : T_pM \to T_{f(p)}N$ is injective at every point $p \in M$. This allows one to embed the tangent space of one manifold into the tangent space of the other. Immersion theorem then guarantees a coordinate system φ around p and ψ around $f(p)$ so that f (or more accurately $\psi \circ f \circ \varphi^{-1}$) has form $(x^1, ..., x^n) \mapsto (x^1, ..., x^n, 0, ..., 0)$ in these coordinates. One can think of f as embedding sufficiently small pieces of M into N, where it will look like an n-dimensional plane in a properly chosen coordinate system. We then say that M is an **immersed submanifold** of N. Of course, one can additionally require that M be globally embedded in N, i.e. that f also be a homeomorphism between M and its image so that, in particular, M inherits the topology of N. In this case we will say that M is an **embedded/regular submanifold** of N (or sometimes simply that M is a submanifold of N).

Definition 1 (Locally trivial collection)**.** The definition given is the same one found in e.g. Lee [15]. Let $\mathcal{F}$ be a collection of immersed k-dimensional submanifolds of M. If every point $p \in M$ has a coordinate neighborhood $(U, \varphi) = (U, x^1, ..., x^n)$ such that every element of $\mathcal{F}$ either doesn't intersect U or intersects it in a (at most) countable union of manifolds of the form $x^{k+1} = c^{k+1}, .., x^n = c^n$ for some constants $c^{k+1}, ..., c^n \in \mathbb{R}$, then we say that $\mathcal{F}$ is **locally trivial**.

In other words, in a appropriately chosen coordinate system, a locally trivial collection $\mathcal{F}$ will look like a stack of k-dimensional planes, all of them parallel to $\mathbb{R}^k \times \{0\}$. One can think of each manifold $F \in \mathcal{F}$ as repeatedly going in and out of the neighborhood U (at most a countable number of times), but looking like a set of (at most countably many) parallel planes on U itself.

[1] In general, one may use higher order perturbation theory.

Definition 2 (Foliation)**.** Let M be a smooth n-dimensional manifold. A foliation partitions M into a (locally trivial) family of connected submanifolds. More precisely, a k-dimensional **foliation** $\mathcal{F}$ is a collection of immersed k-dimensional submanifolds of M (called **leaves**) that are disjoint, connected, locally trivial, and also cover M.

As a trivial example, one can foliate $\mathbb{R}^3$ with planes (take for instance all planes parallel to the XY plane). All 2-dimensional foliations of a 3-dimensional space are locally modeled on this one. Less trivially (but only slightly less so), one can foliate $\mathbb{R}^3 \setminus \{0\}$ with spheres (take all spheres centered at 0). In the next section we will give a completely nontrivial example by foliating a 3-sphere with 2-spheres.

A **distribution** is simply a vector subbundle $D \subset TM$ (a smooth choice of subspace $D_p \subset T_pM$ for each $p \in M$). We say that a distribution D is **integrable** if there exists a submanifold $N \subset M$ which is tangent to D, i.e. for which $D_p = T_pN$ for all $p \in N$. We call N an integral manifold (for D).

The following result gives us a nice characterization of foliations:

Theorem 3 (Frobenius)
Let $D \subset TM$ be a distribution, then D is integrable iff for any two $X, Y \in \Gamma(D)$ (i.e. $X_p, Y_p \in D_p$ for all $p \in M$) we also have $[X, Y] \in \Gamma(D)$. Furthermore, the collection of all maximal connected integral manifolds of D forms a foliation of M.

For a proof see Lee [15]. Conversely, for a foliation $\mathcal{F}$, it is trivial that the collection of all $D_p = T_pF_p$ forms an integrable distribution. Here F_p is the unique leaf which contains $p \in M$.

1.2 Group of Isometries

Let M be a smooth manifold and denote by $\mathrm{diff}(M)$ the set of all diffeomorphisms $M \to M$. $\mathrm{diff}(M)$ obviously forms a group under composition - this group, however, is infinite dimensional. Assume that M comes equipped with a semi-Riemannian metric g (of a given signature), then we can restrict our attention to **isometries**, i.e. diffeomorphisms $\varphi : M \to M$ which preserve the metric $\varphi^* g = g$. Again, collection of all isometries forms a group under composition, which we denote by $I(M)$.

We now recall some terminology from group theory. A **Lie group** G is simply a group with a smooth manifold structure. In particular, this means that the multiplication map $\mu : G \times G \to G,\ (g, h) \mapsto gh$ must be smooth. One can show that the group of isometries $I(M)$ is a Lie group (see Kobayashi [13]).

Generally, a group G can act on a smooth manifold M via diffeomorphisms. In other words, we can embed G into the group of all diffeomorphisms $\mathrm{diff}(M)$ via a (not necessarily injective) homomorphism $G \to \mathrm{diff}(M)$ so that every element $g \in G$ corresponds to a diffeomorphism $g : M \to M$. More specifically, one can require that G act on M via isometries.

Stabilizer (or isotropy group) at p of such a group action is the set of all group elements that fix p:

$$H_p = \{g \in G \mid gp = p\} \subset G.$$

A stabilizer can act on T_pM via derivatives because if ϕ fixes the point $p \in M$, then $d\phi_p : T_pM \to T_pM$.

Orbit at p is the set of all points that can be reached from p by some transformation in G:

$$O_p = \{gp \mid g \in G\} \subset M.$$

For $H \subset G$ we form the set of all left cosets $G/H = \{gH \mid g \in G\}$. Orbit-stabilizer theorem now guarantees that $G/H_p \simeq O_p$, where we have the bijection $j : aH_p \mapsto ap$. An action is called transitive if $O_p = M$ for some (and therefore any) $p \in M$. For a transitive action we then have $G/H_p \simeq M$ and, in fact, can guarantee that j is actually a diffeomorphism (see Lee [15]).

As an application of these ideas, we have the following:

Example 4 (Hopf Fibration)**.** The 3-dimensional rotation group $SO(3)$ of linear operators on $\mathbb{R}^3$ preserves lengths of vectors and therefore (by restriction) acts transitively on the 2-sphere. To fix a point p on the sphere one can only use rotations that have p lying on their axis of rotation and so must rotate in the plane orthogonal to that axis. Therefore, the stabilizer of such action is isomorphic to $SO(2) \simeq S^1$ and we have $S^2 \simeq SO(3)/SO(2)$.

On the other hand, $SO(3)$ has a double cover $SO(3) \simeq SU(2)/\{\pm I\}$ and that double cover acts transitively on S^2 as well (with the same stabilizer). We thus conclude that $SU(2)/SO(2) \simeq S^2$. However, $SU(2)$ is diffeomorphic to the 3-sphere so we get a projection map $\pi : S^3 \to S^2$ whose fiber is diffeomorphic to $SO(2) \simeq S^1$. This partitions S^3 into disjoint 2-spheres and furthermore gives S^3 the structure of a (principal) fiber bundle that locally looks like a product of a piece of S^2 and S^1. This proves that S^3 can indeed be foliated by S^2.

It is often easier to analyze not the Lie group G, but its algebra $\mathfrak{g}$. Let us recall that $\mathfrak{g}$ is simply the tangent space at identity T_eG, which represents infinitesimal transformations (small deviations from the identity).

We should note that every $X_e \in \mathfrak{g}$ generates a vector field on G simply by translating, i.e. $X_g = dl_g X_e$. Here $l_g(h) = gh$ is left multiplication that translates h by g, and so dl_g translates vectors in T_eG to T_gM. This vector field is obviously left-invariant, i.e. $dl_g X_h = X_{gh}$, and is, in fact, the only such vector field that agrees with X_e at the identity. One can thus conclude that $\mathfrak{g}$ can be identified with the space of all left-invariant vector fields and therefore comes equipped with a Lie bracket $[X, Y] = XY - YX$ defined for any two vector fields X, Y.

Integral curves of left-invariant vector fields must run for all times $t \in \mathbb{R}$ and have a specific property: $c(t + s) = c(t)c(s)$. In other words, they are simply homomorphisms between $(\mathbb{R}, +)$ and G and are consequently called 1-parameter groups. Conversely, every 1-parameter group must be an integral curve of some left-invariant field (by uniqueness of solutions to ODE).

We are thus interested not only in isometries themselves, but also in infinitesimal isometries.

Definition 5 (Killing field)**.** A vector field on M is called **Killing** if its local flow $\varphi_t : U \to M$ is an isometry (for all t for which it is defined).

The flow of a complete Killing field is defined on all of M, i.e. $\varphi_t : M \to M$ for all $t \in \mathbb{R}$. Therefore, we get a 1-parameter group in $I(M)$ through the identity $\varphi_0 = \mathrm{id}$. The derivative of this curve at id is precisely the Killing field. We thus conclude that complete Killing fields are in bijective correspondence with the algebra of group $I(M)$.

We finish this section with a couple of comments on the structure of the space of Killing fields. It not too difficult to see that X is Killing iff $\mathcal{L}_X g = 0$. This is precisely because the Lie derivative $\mathcal{L}_X$ is defined by $\mathcal{L}_X g = \lim_{t\to 0} \frac{1}{t}(\varphi_t^* g - g)$. As we have $[\mathcal{L}_X, \mathcal{L}_Y] = \mathcal{L}_{[X,Y]}$, it follows that for X and Y Killing, $[X, Y]$ must be Killing as well. Thus the space of Killing fields is an algebra.

One can relatively easily show that every isometry φ on a connected manifold is determined by $\varphi(p)$ and $d\varphi_p$ at some point $p \in M$ (this is because isometries preserve geodesics). Analogously, a Killing field X (on a connected manifold) is determined by X_p and $(\nabla X)_p$ at some point.

It is not too difficult to show (for a Levi-Civita connection) that $\mathcal{L}_X g = \langle \nabla_A X | B\rangle + \langle A | \nabla_B X\rangle$. Thus X is Killing iff $\nabla X : A \mapsto \nabla_A X$ is antisymmetric with respect to metric $g = \langle \cdot | \cdot \rangle$. This is the so-called Killing equation.

Knowing that $(\nabla X)_p$ must be antisymmetric with respect to metric g, we see that it can have at most $\frac{n(n-1)}{2}$ degrees of freedom. We thus see that the algebra of Killing vector fields (and consequently the group of isometries) can be at most $n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$ dimensional.

1.3 Spacetime Symmetries

We can now define many different kinds of symmetric spacetimes.

Definition 6 (Time symmetries)**.** Let M be a spacetime.

1. We say M is **stationary** if there exists a timelike Killing vector field (Wald [4]). This field gives us (generally only locally defined) time translations. As these time translations are isometries, we interpret this to mean the metric is time independent.

2. In a stationary spacetime, the timelike Killing field X induces the instantaneous rest spaces $D_p = \{v \in T_p M \mid \langle v | X\rangle = 0\}$, which form a distribution $D \subset TM$. If D is integrable in the sense of Frobenius, then we say M is **static** (Wald [4]). Mathematically, this means precisely that M can be foliated into (connected) manifolds all tangential to the distribution D, i.e. orthogonal to X. Physically, this means that the family of observers X can synchronize their clocks in such a way that the space of simultaneity common to these observers is actually a 3-manifold.

We should comment on the definition of static spacetime. The Frobenius integrability condition for D, namely that $[A, B] \in \Gamma(D)$ whenever $A, B \in \Gamma(D)$, can be equivalently stated only in terms of the vector field X. Denote by $\xi = X^\flat = \langle X | \cdot \rangle$ the dual of X. Then we have $\ker \xi = D$, i.e. D consists of precisely those vectors that ξ sends to 0 (as precisely these are orthogonal to X). On the other hand, by $d\xi(A, B) = A\xi(B) - B\xi(A) - \xi([A.B])$ we have $d\xi(A, B) = -\xi([A, B])$ for any $A, B \in \ker \xi = D$. From here one concludes that D is integrable iff $d\xi = 0$ on D.

We call $d\xi$ the curl of X. In particular, we have $\operatorname{curl} X(A, B) = \langle \nabla_A X | B\rangle - \langle \nabla_B X | A\rangle$. One can see this by going to normal coordinates where $\operatorname{curl} X(\partial_i, \partial_j) = \langle \partial_i X | \partial_j\rangle - \langle \partial_j X | \partial_i = \partial_i X_j - \partial_j X_i = dX^\flat(\partial_i, \partial_j)$. Therefore, D is integrable iff $\operatorname{curl} X = 0$ on D so we can say that a static spacetime is one that has an irrotational timelike Killing field.

It is worth noting that "$d\xi = 0$ whenever $\xi = 0$" can be written equivalently as $\xi \wedge d\xi = 0$. To see this simply compute the expression for $\xi \wedge d\xi(A, B, C)$ on some

orthogonal vectors A, B, C, noting that only one of them can be vertical, i.e. not in $\ker \xi$. This kills all but one term (the one that contains the product of ξ on the vertical vector and $d\xi$ on the remaining two horizontal ones). By writing $\xi \wedge d\xi = 0$ out in coordinates we get $0 = \xi_{[\mu}\partial_\nu \xi_{\sigma]} = \xi_{[\mu}\nabla_\nu \xi_{\sigma]}$, where the brackets indicate that we are antisymmetrizing over all three indices. The last equality follows from the fact that Levi-Civita connection $\nabla_k X^i = \partial_k X^i + X^j \Gamma^i_{kj}$ is symmetric ($\Gamma^k_{ij} = \Gamma^k_{ji}$).

Definition 7 (Space symmetries)**.** Let M be a spacetime.

1. Call M **spherically symmetric** if $I(M)$ contains $SO(3)$ as a subgroup and orbits of $SO(3)$ are spacelike 2-spheres (Wald [4]). More generally, one may allow the orbits to be any spacelike 2-manifold Σ (Hawking & Ellis [3]) because one then immediately has that Σ must be locally isometric to a sphere (by having $SO(3)$ in its group of isometries - see theorem 9).

2. We say that M is **axisymmetric** if there exists a 1-parameter group of isometries φ_t whose orbits are closed spacelike curves (topologically speaking these are simply circles). This implies the existence of a spacelike Killing field with closed integral curves. In other words, $I(M)$ contains $SO(2)$, whose orbits are spacelike circles (Wald [4]).

   Some authors (Heusler [10]) require that the fixed point set of $SO(2)$ be non empty as well (so in particular it has an axis of rotation).

   For a spacetime that is both stationary and axisymmetric, it is standard to require that the two Killing fields commute (Wald, [4], Heusler [10]). In particular, this guarantees that their flows will commute as well.

Definition 8 (Maximally symmetric space)**.** We say that a semi-Riemannian manifold M is **maximally symmetric** if its algebra of Killing fields has the highest possible dimension, namely $\frac{n(n+1)}{2}$.

As far as maximally symmetric spacetimes are concerned, one can show the following:

Theorem 9
Let M be a pseudo-Riemannian manifold of dimension n and signature s. Then the following are equivalent:

1. *M is locally maximally symmetric, i.e. every point $p \in M$ has a maximally symmetric neighborhood.*

2. *M is a space of constant curvature. This means that the sectional curvature*

$$K(X,Y) = \frac{\langle R(X,Y)Y|X\rangle}{g(X,X)g(Y,Y) - g(X,Y)^2}$$

   [2] is equal to the same constant $K \in \mathbb{R}$ for any $X, Y \in T_pM$ and any $p \in M$.

3. *There exists a constant K such that the Riemann curvature tensor has the following form in any coordinate system:*

$$R_{ijkl} = K(g_{ik}g_{jl} - g_{il}g_{jk}).$$

[2]Note that here K is defined only for those X, Y for which $g(X,X)g(Y,Y) - g(X,Y)^2 \neq 0$, i.e. which generate a plane Π on which g is non degenerate (has non zero determinant). Also, one can prove that K depends only on this plane Π and not on any specific choice of X and Y that generate Π.

4. M is locally isometric to one of the following spaces:

- *pseudo-euclidean space R^n_s (this is just $\mathbb{R}^n$ with a metric of signature s).*
  If we denote $O_s(n)$ the group of all linear isometries of $\mathbb{R}^n_s$, then any isometry of $\mathbb{R}^n_s$ can be written as a $\tau_x R$, where $R \in O_s(n)$ and $\tau_x : p \mapsto p + x$. As translations and orthogonal transformations do not commute, the isometry group of $\mathbb{R}^n_s$ is $\mathbb{R}^n \rtimes O_s(n)$ - $\mathbb{R}^n$ denoting translations and $\rtimes$ a semidirect product.
- *pseudohyperbolic space $H^n_s(r) = \{p \in \mathbb{R}^{n+1}_{s+1} \mid g(p,p) = -r^2\}$, that has $O_{s+1}(n+1)$ as its group of isometries whenever $s > 0$[3].*
- *pseudosphere $S^n_s(r) = \{p \in \mathbb{R}^{n+1}_s \mid g(p,p) = r^2\}$, that has $O_s(n+1)$ as its group of isometries whenever $s < n$[4].*

In particular, we get the classical theorem that all maximally symmetric Riemannian manifolds are locally isometric either to a Euclidean space, a sphere, or a hyperbolic space. On the other hand, for spaces of Lorentzian signature, we find that every maximally symmetric spacetime must be locally isometric either to Minkowski space ($\mathbb{R}^n_1$), de Sitter space ($dS_n = S^n_1$), or anti-de Sitter space ($AdS_n = H^n_1$).

Proof. Essentially the proof of all these statements can be found in O'Neill [1], if one is willing to work through problem 14 of chapter 9. □

1.4 Splitting of the Metric Under Symmetries

Because the proof is so cute and instructive (and because I've seen so many people butcher it), let us briefly discuss how one would go about decomposing the metric under a symmetry group. This is essentially taken from the lovely paper by B. Schmidt [21].

First we give a preliminary result:

Proposition 10
Assume that a Lie group G of dimension r acts on a smooth manifold M via diffeomorphisms. Assume that the orbits $O_p = \{gp \in M \mid g \in G\}$ of G are all connected smooth manifolds of the same dimension k and assume further that O_p are closed in M. Then the orbits form a foliation of M.

For a compact connected group, we automatically get that the orbits O_p must be compact and connected as well (the group action $G \times M \to M$ is smooth so sends the compact connected set $G \times \{p\}$ to a compact connected set O_p).

Proof. This is a relatively simple application of the Frobenius theorem. We should note that it is not (*a priori*) entirely clear that the orbits form a locally trivial collection.

First we need to prove that $D_p = T_p O_p$ defines a distribution (so that D_p vary smoothly with p). For this it is sufficient to prove that for any sufficiently small neighborhood

[3] H^n_0 has $O^{++}_1(n+1) \cup O^{+-}_1(n+1)$ as its group of isometries, i.e. transformations that preserve time orientation.

[4] S^n_n has $O^{++}_n(n+1) \cup O^{-+}_n(n+1)$ as its group of isometries, i.e. transformations that preserve space orientation.

$U \subset M$ one can find vector fields $X_1, ..., X_k$ on U that constitute a basis on each D_p for all $p \in U$. Indeed, we can choose a basis $e_1, ..., e_r$ for $\mathfrak{g}$ and define

$$(X_i)_p = \frac{d}{dt}\Big|_{t=0}(e^{te_i}p).$$

X_i generate D_p because the action on each orbit is transitive, but are not necessarily linearly independent. Now choose only those X_i which are linearly independent at p and thus form a basis for D_p. It is clear (say by continuity of the determinant) that these must be linearly independent in a sufficiently small neighborhood around p as well.

We therefore see that this is an integrable distribution (orbits are integral manifolds) so by Frobenius the maximal integral manifolds F_p form a foliation $\mathcal{F}$ of M. What remains to be seen is that the orbits *are* the maximal integral manifolds. As each orbit is connected, we have $O_p \subset F_p$ (the maximal integral manifold must contain all connected integral manifolds through p). Once we show that O_p is open and closed in F_p, we are done.

When proving the Frobenius theorem, one proves that $F_p \subset M$ is weakly embedded. This means that every smooth map $N \to M$ whose image lies in F_p is smooth as a map $N \to F_p$ as well. This has two consequences:

1. As orbits are closed in M by assumption, they must be closed in F_p as well.

2. The inclusion $O_p \to F_p$ must be a smooth map and therefore an immersion.

As O_p and F_p have the same dimension, $O_p \to F_p$ is a submersion as well and therefore (by say the submersion theorem) an open map. So $O_p \subset F_p$ must be open and closed, from which it follows that $O_p = F_p$.

□

Now let us further assume that each orbit (i.e. leaf) is an embedded k-dimensional semi-Riemannian submanifold of M. In particular, this means that for every $F \in \mathcal{F}$ and $p \in F$, the metric must be nondegenerate on the subspace $T_pF \subset T_pM$ or, what amounts to the same thing, the tangent space to M at p decomposes as $T_pM = T_pF \oplus T_pF^\perp$. We now prove the two main results of this section:

Proposition 11
If the stabilizer G_p of G leaves no tangent vector in T_pO_p fixed, i.e. $(\forall v \in T_pO_p \setminus \{0\})(\exists g \in G_p)(gv \neq v)$, then one can find a family of $n-k$ dimensional surfaces orthogonal to the orbits.

Proof. Let $T_pO_p^\perp$ be the normal space to O_p at p so that we have $T_pM = T_pO_p \oplus T_pO_p^\perp$. Take an open ball $U \subset T_pM$ to be such that exp is a diffeomorphism on U. Then exp maps $T_pO_p^\perp \cap U$ to some $C \subset M$ diffeomorphically so we get a $(n-k)$-dimensional submanifold C containing precisely all the geodesics going through p and being normal to O_p at p. We now show that C is orthogonal to every orbit that it intersects (not just O_p).

- We first show that G_p fixes C. Note that, in a sufficiently small neighborhood U around point p, M looks like a product $U \cap O_p \times C$ (this being a tubular neighborhood of $U \cap O_p$). Therefore a point $q \neq p$ sufficiently close to p which lies in C cannot lie in O_p. In other words, (at least locally) orbits that go through two different points of C cannot have points in common.

Now G_p fixes every vector in $T_pO_p^\perp$. Otherwise, one could transform two distinct vectors into one another, but then so could one transform the geodesics that these vectors generate (isometries preserve geodesics) thereby giving distinct points in C that lie in the same orbit; a contradiction.

Finally, we see that C must be fixed by G_p because every point on C lies on a geodesic with initial velocity $v \in T_pO_p^\perp$, which is fixed by G_p.

- Let $q \in C$ and $\lambda \in T_qC$. G_p fixes vector λ because it fixes the entirety of C. Decompose λ as $\lambda = \lambda_1 + \lambda_2$, where $\lambda_1 \in T_qO_q$ and $\lambda_2 \in T_qO_q^\perp$. If we can show that $\lambda_1 = 0$, i.e. that $\lambda \in T_qO_q^\perp$, we are done.

  Note that λ_2 is fixed by G_p. Certainly, it is fixed by G_q by the previous argument. On the other hand, $G_q = G_p$ because all elements $g \in G$ that fix p also fix C (and vice-versa of course). As both λ_1 and λ are fixed, $\lambda_1 = \lambda - \lambda_2 \in T_qO_q$ must also be fixed, but by assumption no vector in T_qO_q can be fixed (excluding the 0 of course). We conclude therefore that $\lambda_1 = 0$, i.e. $\lambda = \lambda_2 \in T_qO_q^\perp$.

□

Proposition 12
Assume that the orbits have orthogonal $(n-k)$-dimensional surfaces (as in the previous theorem). Furthermore, assume that in each T_pO_p there exists a basis $B = \{e_1, ..., e_k\}$ on which G_p acts transitively, i.e. $(\forall e_i, e_j \in B)(\exists g \in G_p)(ge_i = e_j)$. Then the orthogonal surfaces map the orbits conformally onto one another (i.e. by following the geodesic in C, one reaches an orbit with a conformally transformed metric).

Proof.

- Let O_p and O_q be two different orbits with p and q sufficiently close. We first show that $C = O_p^\perp$ and O_q intersect in exactly one point.

  We have seen in the proof of the previous theorem that, for p and q sufficiently close, C and O_q can have at most one common point. To show their intersection is not empty, we show that there exists a geodesic normal to O_p intersecting O_q. This is obvious from the fact that orbits form a locally trivial collection - through any point in C there must pass some orbit, but by local triviality, these are precisely all orbits around p.

- Now define $f : p \mapsto f(p)$ where $f(p)$ is the sole element in $O_q \cap O_p^\perp$[5]. We obviously have $gf = fg$ for all $g \in G$ because $g(f(p)) = g(O_q \cap O_p^\perp) = O_q \cap O_{g(p)}^\perp = f(g(p))$, as $gO_q = O_q$ is of course fixed. It is clear now (by chain rule) that this must also hold for the derivative of f (which we denote by the same symbol) when G_p acts on T_pM (via derivatives).

- Take now $e_1, ..., e_k$ to be a basis for T_pO_p on which the isotropy group G_p acts transitively and let $g_i \in G_p$ be given by $g_ie_1 = e_i$. As we have $f(e_i) = f(g_ie_1) = g_if(e_1)$, it is clear that all $f(e_i)$ are of same length (g_i are isometries). Therefore, f maps a basis of unit vectors in T_pO_p to a basis of vectors in $T_{f(p)}O_{f(p)}$ all of which have the same length. In other words, f is a conformal map.

[5]Of course, one does not need axiom of choice for this as one is not really making a choice. We are picking the only possible element from each set. Explicitly, $f(p) = \bigcup O_q \cap O_p^\perp$.

- Moreover, the conformal factor in constant on each O_q because we can compose

$$T_{p'}O_p \xrightarrow{h} T_pO_p \xrightarrow{f} T_{f(p)}O_{f(p)} \xrightarrow{h^{-1}} T_{h^{-1}f(p)}O_{f(p)} = T_{f(p')}O_{f(p)},$$

  where h is an isometry in G that sends $p \in O_p$ to $p' \in O_p$.

□

A couple of comments are in order. Note that in a vector space of *any* signature, one may find a basis e_i all of whose vectors have equal $g(e_i, e_i)$. It is intuitively clear that one can focus all the vectors in one direction (say where the metric is positive $g(v, v) > 0$) and spread the vectors out just enough as to make them linearly independent[6]. Therefore it is possible that the isometry group act transitively on a basis even in a space of indefinite signature (it cannot act transitively on an orthonormal basis of course)

The argument that showed the existence of a normal geodesic from O_p to O_q relies on the fact that the orbits form a locally trivial collection (and is in that case obvious). It is not clear (at least not to me) how one would prove this otherwise. Thus theorem 10 is essential here. This is also not explicitly mentioned in [21] or, for that matter, anywhere in the physics literature (where I looked).

Assuming the conclusions of the previous two theorems are satisfied, the metric can be put in a particularly neat form. First, one can choose a (local) coordinate system (x^α, x^A), where $\alpha = 1, ..., k$ and $A = k + 1, ..., n$ so that $X^A = \text{const.}$ describes the orbits and $x^\alpha = \text{const.}$ describes the space orthogonal to the orbits. Note here that, locally speaking, the space looks like a product $O_p \times C$ so one can take the product chart. In these coordinates the metric has the form:

$$g = B^2(x^A)g_{\alpha\beta}(x^\alpha)dx^\alpha dx^\beta + g_{AB}(x^A)dx^A dx^B,$$

where the first term is a metric on the orbits and the second is a metric on the orthogonal space C. The B^2 is the conformal factor that changes from orbit to orbit. In other words, M is locally isometric to a warped product $O_p \times_B C$.

Finally, applying the above general theorems to a special case of maximally symmetric orbits, we get:

Corollary 13
Let M be a semi-Riemannian manifold and let Lie group G act on M via isometries. Assume that the orbits are k-dimensional semi-Riemannian surfaces that foliate M. Let $\dim G = \frac{k(k+1)}{2}$ so that the orbits are maximally symmetric spaces. Then the assumptions of the previous two theorems are satisfied and the metric locally has the form

$$g = B^2(x^A)g_{\alpha\beta}(x^\alpha)dx^\alpha dx^\beta + g_{AB}(x^A)dx^A dx^B,$$

where $g_{\alpha\beta}(x^\alpha)dx^\alpha dx^\beta$ is a metric of constant curvature on the orbit $x^A = \text{const.}$

Proof. As orbits are maximally symmetric, they are locally either $\mathbb{R}^n_s$, S^n_s or H^n_s. Because, locally speaking, group G of isometries (of orbits) is maximal, the isotropy group G_p is maximal as well and, in particular, the orbits are locally isotropic. In other words, locally

[6]More precisely, take $\mathbb{R}^n$ with standard basis e_i and metric g of signature s. Take $v \in \mathbb{R}^n$ to have $g(v, v) > 0$. Now $v_i = v + \varepsilon e_i$ all have $g(v_i, v_i) > 0$ (for a sufficiently small ε), but are also linearly independent.

speaking, its isotropy group acts transitively on all vectors of the same norm $g(v, v)$ (i.e. the pseudosphere).

This guarantees that no vector can remain fixed under the action of isotropy group and that there exists a basis on which the isotropy group acts transitively (as we have previously commented). It is now obvious that the results from the previous two theorems apply as was to be demonstrated. □

Since spheres are maximally symmetric, it is now relatively straightforward to explicitly find the metric of a spherically symmetric spacetime. Indeed, we see that the metric of such a spacetime must be of the form

$$ds^2 = d\tau^2(t, r) + Y^2(t, r)d\Omega^2(\theta, \phi),$$

where $d\Omega^2(\theta, \phi) = d\theta^2 + \sin^2\theta d\phi^2$ is the metric of the unit sphere and $d\tau^2$ is the indefinite metric (signature $(-, +)$) on the space orthogonal to the spheres parametrized by coordinates t and r. $Y^2 > 0$ is just the conformal factor.

Chapter 2

Examples of Black Hole Spacetimes

Historically, one of the earliest known exact solutions to Einstein vacuum equations was also a solution that contained a black hole. This is the *Schwarzschild metric* describing a spherically symmetric vacuum. Here we prove a stronger result: any spherically symmetric spacetime must be (locally) isometric to Schwarzschild spacetime - a fact known as *Birkhoff's theorem* (so, modulo global topology, Schwarzschild metric is *the* spherically symmetric vacuum solution).

Later we discuss the case of an axially symmetric, i.e. rotating black hole; this is the *Kerr solution*. It is worth noting that an axially symmetric black hole solution was found only relatively late in the game (Kerr 1963), with the help of Petrov types.

2.1 Schwarzschild metric and Birkhoff's Theorem

The **Schwarzschild metric** is given by:

$$ds^2 = -(1 - \frac{2mG}{r})dt^2 + (1 - \frac{2mG}{r})^{-1}dr^2 + r^2 d\Omega^2 \tag{2.1}$$

and is obviously spherically symmetric. Here $r_s = 2mG \geq 0$ (the so-called Schwarzschild radius) is just some number parametrizing the family of metrics. $G = 7.41 \cdot 10^{-28} kg^{-1}m$ is the universal gravitational constant, which guarantees that m will be in units of mass (we are taking the speed of light c to be 1; in SI units $G = 6.67 \cdot 10^{-11} m^3 kg^{-1} s^{-2}$). Note that r coordinate becomes timelike and t coordinate spacelike for $r < 2mG$.

We also notice a break-down of coordinates at $r = 2mG$. This can be remedied by passing to **Eddington-Finkelstein coordinates**. Indeed, if one draws the radial light geodesics, it appears as though they fly off to infinity at $r = 2mG$ and then come back from infinity on the other side ($r < 2mG$). Let us therefore try to find coordinates in which radial lightlike geodesics are straight lines (this way we control the behavior around $r = 2mG$). For a radial light geodesic we have:

$$0 = -(1 - \frac{2mG}{r})dt^2 + (1 - \frac{2mG}{r})^{-1}dr^2.$$

From here it follows that $\frac{dt}{dr} = \pm(1 - \frac{2mG}{r})^{-1}$ ($-$ for ingoing and $+$ for outgoing light geodesics). This has a solution (easily checked by taking the derivative):

$$t = \pm r^* + \text{const},$$

where $r^* = r + 2Gm \ln|\frac{r}{2Gm} - 1|$. We see that the light geodesics would indeed be straight lines if we had r^* instead of r as a coordinate. Thus changing the r coordinate to r^*, the metric becomes:

$$ds^2 = (1 - \frac{2mG}{r})(-dt^2 + (dr^*)^2) + r^2\Omega^2$$

Define $v = t + r^*$, then:

$$ds^2 = -(1 - \frac{2mG}{r})dv^2 + 2dvdr + r^2 d\Omega^2.$$

Note that the determinant of the metric is $-r^4 \sin^2\theta$. We therefore see that the metric is non degenerate at $r = 2mG$ (even though one of its components vanishes there) and the coordinate system has incorporated the event horizon $r = 2mG$.

We finally state the main theorem of this chapter:

Theorem 14 (Birkhoff)
Let M be a spherically symmetric spacetime so that in any case the orbits of $SO(3) \subset I(M)$ are locally isometric to a sphere. Assume M is Ricci flat, i.e. the metric g satisfies the vacuum Einstein equations $R_{\mu\nu} = 0$, where $R_{\mu\nu}$ is the Ricci tensor. Then (in any sufficiently small neighborhood) the metric g must be given by the following expression:

$$ds^2 = -(1 - \frac{2mG}{r})dv^2 + 2dvdr + r^2 d\Omega^2.$$

It is beneficial to prove the theorem in Eddington-Finkelstein coordinates, instead of the original (Schwarzschild) coordinates. Otherwise, we would have to treat the event horizon as an exception and the expression for the metric would then split between the interior and the exterior of the black hole, which complicates the discussion. In this regard we more or less follow [52].

First, the metric $ds^2 = d\tau^2(t, r) + Y^2(t, r)d\Omega^2(\theta, \phi)$ can be further simplified:

Lemma 15
Every spherically symmetric metric can be brought into the following form:

$$ds^2 = F(v, r)dv^2 + 2X(v, r)dvdr + Y^2(v, r)d\Omega^2.$$

The metric is Lorentzian whenever $X \neq 0$ (while there are no restrictions on F).

Proof. Indeed, we generally have $\tau(t, r) = -Adt^2 + 2Bdrdt + Cdr^2$. Note that this can be written as:

$$\tau(t, r) = -A(dt + Ddr)^2 + 2(AD + B)(dT + Ddr)dr,$$

where $C - AD^2 - 2BD = 0$. Now define a 1-form $\omega = dt + Ddr$ and let $\mu \neq 0$ be its integral factor, i.e. $\mu\omega = dv$ for some function $v = v(r, t)$. We then have:

$$\tau(v, r) = -(A/\mu^2)dv^2 + 2\frac{AD + B}{\mu}dvdr,$$

which was to be proven.

The metric $-Adt^2 + 2Bdrdt + Cdr^2$ will be Lorentzian when $AC + B^2 > 0$ (i.e. when it has a negative determinant) so $Fdv^2 + 2Xdvdr$ will be Lorentzian for $X^2 > 0$, i.e. $X \neq 0$. □

Let us comment on the existence of integral factors. Let $\omega = Mdx + Ndy$ be some general 1-form in two dimensions. We then have

$$d\omega = \frac{\partial M}{\partial y} dy \wedge dx + \frac{\partial N}{\partial x} dx \wedge dy \Longrightarrow d\omega = \left(\frac{\partial N}{dx} - \frac{\partial M}{\partial y} \right) dx \wedge dy.$$

We therefore see that $d\omega = 0$ iff $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$. Note that every closed form ($d\omega = 0$) is locally exact ($\omega = dh$) because every sufficiently small neighborhood is diffeomorphic to $\mathbb{R}^n$, whose de Rham cohomologies all vanish (this is the so-called Poincaré lemma). Therefore, on a sufficiently small neighborhood, we have $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$ iff $\omega = dh$ for some function h.

Finally, if ω is not closed, multiply by some integral factor μ. Then $\mu\omega$ will be closed iff $\frac{\partial}{\partial x}(\mu N) = \frac{\partial}{\partial y}(\mu M)$ i.e. iff

$$\partial_y \mu M + \mu \partial_y M = \partial_x \mu N + \mu \partial_x N \Longleftrightarrow (\partial_y \mu) M - (\partial_x \mu) N = \mu(\partial_x N - \partial_y M).$$

This is a linear partial differential equation (of 1st order) so it will always have a (local) solution. There is a geometric proof of this fact (using the language of 1-jet bundles - see Arnol'd [16]).

We conclude that one can always choose an integral factor $\mu \neq 0$ in such a way that $\mu\omega$ is closed (and therefore exact).

Proof of theorem 14. The proof is now relatively simple. Denote by $f' = \partial_r f$ the r derivative, by $\dot{f} = \partial_v f$ the v derivative and compute the Ricci tensor (using e.g. einsteinpy - see appendix B).

- First note that Y can be taken as a coordinate. Indeed, the rr component of the Ricci tensor is:
$$R_{rr} = \frac{2}{XY}(-XY'' + X'Y') = 0.$$
This implies $0 = \frac{XY'' - X'Y'}{X^2} = (\frac{Y'}{X})'$, from which we conclude that either $Y' = 0$ or $X = \xi(v)Y'$ for some function ξ of v. If $Y' = 0$, then we can calculate $R_{\theta\theta} = 1$, which cannot satisfy the vacuum equations. Now from $Y' \neq 0$ we get (locally) $dY \neq 0$ so by submersion theorem, $Y = \text{const}$ are (locally) embedded manifolds, represented as hyperplanes in the adapted coordinate system $Y, x^2, ...x^n$. As $dY = \dot{Y}dv + Y'dr$ (and $X = \xi Y'$), we have:
$$ds^2 = (F - 2\xi\dot{Y})dv^2 + \xi dv dY + Y^2 d\Omega^2$$
Setting $\tilde{F} = (F - 2\xi\dot{Y})$ and $d\tilde{v} = \xi dv$, we get:
$$ds^2 = \tilde{F}dv^2 + 2d\tilde{v}dY + Y^2 d\Omega^2.$$
For convenience we rename the variables once more and write $ds^2 = Fdv^2 + 2dvdr + r^2 d\Omega^2$.

- We now compute the $\theta\theta$ component of the Ricci tensor. Note that $X = 1$ and $Y = r$ so $Y' = 1$, $Y'' = \dot{Y} = 0$. This gives
$$R_{\theta\theta} = F + 1 + rF' = 0.$$

Separating the variables in the equation $F + 1 + rF' = 0$ we get:

$$\frac{dF}{F+1} = -\frac{dr}{r}.$$

Integrating we get $\ln(F+1) = -\ln r + C = \ln(r^{-1}) + C$ where C depends only on v. Taking the exponential:

$$F = mr^{-1} - 1,$$

where $m = e^C > 0$ depends only on v.

Therefore the metric is:

$$(mr^{-1} - 1)dv^2 + 2dvdr + r^2 d\Omega^2$$

and we only need to show that m is constant.

- For this we compute the uu component of the Ricci tensor:

$$R_{uu} = \frac{1}{2}FF'' + \frac{1}{r}FF' + \frac{1}{r}\dot{F} = 0.$$

Substituting $F = mr^{-1} - 1$ into the above expression we get ($F' = -\frac{m}{r^2}$, $F'' = \frac{2m}{r^3}$):

$$0 = \frac{1}{r^4}m^2 - \frac{1}{r^4}m^2 + \frac{\dot{m}}{r^2} = \frac{\dot{m}}{r^2}$$

and it is clear that $\dot{m} = 0$, i.e. $m = \text{const}$.

□

Birkhoff's theorem is only a local result. Globally, one looks for **extensions**. Generally, an extension of a (connected) spacetime M is a (connected) spacetime $\tilde{M}$ for which we can find an isometry $i : M \to \tilde{M}$ and $i(M) \neq M$. In other words, M can be isometrically embedded as an open subset of $\tilde{M}$ (as dimensions of M and $\tilde{M}$ must be the same, i is, in particular, a submersion and therefore an open map). One is most interested, of course, in **maximal** (or inextendible) spacetimes, namely those which do not posses an extension. Let us note here that a given spacetime M may have many different maximal extensions (so we cannot speak of *the* maximal extension)[1]. Birkhoff's theorem can now be stated much more eloquently: every spherically symmetric spacetime which satisfies the vacuum equations is locally isometric to a piece of the maximally extended Schwarzschild spacetime.

By extending the Schwarzschild spacetime to Eddington-Finkelstein coordinates, we found that the apparent singularity at $r = 2MG$ in the Schwarzschild metric disappears. The singularity at $r = 0$, however, does not. Indeed, by calculating the tidal forces, we see that they become infinite as $r \to 0$. Alternatively, one can calculate the so-called Kretschmann scalar $K = R^{abcd}R_{abcd} = \frac{48G^2M^2}{r^6}$.

[1]Indeed, as a counterexample we can take $\mathbb{R}^2$ and the flat torus. Both are complete (and therefore inextendible) as well as flat (so extend the same local geometry), but are not homeomorphic. One can however prove that a simply connected analytic spacetime (under some condition on geodesics) has a unique maximal analytic extension (see Chruściel [6] theorem 4.4.4). This is not terribly useful for e.g. Kerr spacetime as it is not simply connected. Nevertheless one can prove (see Chruściel [6] theorem 7.3.3 and O'Neill [2]) the uniqueness of the maximal Kerr extension and, by extension, the uniqueness of the maximal Schwarzschild extension - the Kruskal spacetime.

2.2 Kerr Spacetime

We now discuss the 2-parameter family of spacetimes describing a rotating black hole. This solution to the Einstein vacuum equations is also important because of the Hawking-Carter-Robinson uniqueness theorem (see chapter 5). A lovely and thorough reference is O'Neill [2].

The only reasonable place to start from is, of course, the metric:

$$ds^2 = \rho^2(\frac{dr^2}{\Delta} + d\theta^2) + (r^2 + a^2)sin^2\theta d\phi^2 - dt^2 + \frac{2mr}{\rho^2}(a\sin^2\theta d\phi - dt)^2, \tag{2.2}$$

where we have defined $\rho^2 = r^2 + a^2\cos^2\theta$ and $\Delta = r^2 - 2mr + a^2$. We should note that $\frac{\Delta}{r^2} = 1 - \frac{2m}{r} + \frac{a^2}{r^2}$ generalizes the Schwarzschild horizon function $r \mapsto 1 - \frac{2m}{r}$. The parameter a can be interpreted as angular momentum (per unit mass). Then $a = 0$ means simply that the black hole isn't spinning (so we get the Schwarzschild metric). We also mention that the electrovac solution to the uniqueness problem is the 3-parameter (m, a, e) **Kerr-Newman** family of metrics. The metric has the same form, but we must take $\Delta = r^2 + a^2 + e^2 - 2mr$. We don't really expect to see black holes with large charge to mass ratio e, though, since such a body would selectively attract particles of the opposite charge. We may therefore neglect the electromagnetic contribution and assume e to be 0.

We can now define several regimes as determined by the roots of the function Δ:

1. $a = 0$, i.e. Schwarzschild spacetime, then Δ has roots 0 and $2m$.
2. $0 < a^2 < m^2$, the **slowly rotating Kerr spacetime**, then Δ has two roots $0 < r_\pm = m \pm \sqrt{m^2 - a^2} < 2m$.
3. $a^2 = m^2$, the **extreme Kerr spacetime**, then $\Delta = (r - m)^2$ has only one root $r = m$.
4. $a^2 > m^2$, the **rapidly rotating Kerr spacetime**, then Δ has no roots.

We call $\{\Delta = 0\}$ the *horizon*. Therefore, in the fast rotating case there is no horizon, in the extreme case there is only one, and in the slow case we get two. It turns out that the exterior horizon is the event horizon and the interior one is a *Cauchy horizon* (which we define in the next chapter). Roughly, the causal structure becomes quite pathological beneath the Cauchy horizon. In a bit more detail, given an initial data outside the black hole, one can predict the future up to the second horizon. We can analytically extend the Kerr solution beyond the Cauchy horizon, but this extension is not unique among all smooth extensions.

The $\Sigma = \{\rho = 0\}$ can be shown to be a real singularity (not just a coordinate one as with $\{\Delta = 0\}$). Notice that the singularity is a ring or, more accurately, the set of all points *on the chart* for which $\rho = 0$, has topology $S^1 \times \mathbb{R}$ (because $\rho = 0$ iff $r = 0$ and $\cos\theta = 0$). This is to be contrasted with the Schwarzschild case where we get a single point or rather a line $\mathbb{R}$.

We now define the following regions (all of which are connected Lorentzian 4-manifolds):

1. For slow Kerr:
   - I: $r > r_+$

- II: $r_- < r < r_+$
- III: $r < r_-$

2. For extreme Kerr:
   - I: $r > m$
   - III: $r < m$
3. Fast Kerr has no horizon so can be regarded as a single region $III = I$ homeomorphic to $(R^2 \times S^2) \setminus \Sigma$.

One can then, for instance, find a maximal analytic extension of Kerr spacetime by gluing these regions in the appropriate way (in particular, the topology of the slow case is most interesting and turns out to be a bundle - see [2]).

As a preparation for the next chapter, we now discuss causality in Kerr spacetime. We shall say M is *causal* if there exist no timelike nor lightlike (i.e. no nonspacelike) curves which are closed. If this condition is broken, then by traveling along such a (timelike) curve, one can travel into the past, or (using lightlike curves) one could send signals into the past. We first examine the exterior region I and region between the two horizons II:

Proposition 16
Blocks I and II are causal.

Proof. See proposition 2.4.6 in [2]. In fact, as we will mention later on, a more general statement holds: these regions are globally hyperbolic (for proof see [54]) □

However, the region below the second horizon III is very pathological. Note that ∂_ϕ becomes timelike in region III near the singularity. Since the integral curves of ∂_ϕ are closed, this region is not causal. The set on which $g_{\phi\phi} < 0$ is usually called Carter time machine.

In fact, causality fails in quite a spectacular manner:

Proposition 17
Block III is vicious, i.e. given any events $p, q \in III$, there exists a timelike curve in III from p to q.

Proof. (for proof see proposition 2.4.7 in [2]) □

From the point of view of Einstein equations, such spacetimes are pathological, so instead of worrying about time paradoxes, let us return to a physically meaningful discussion by explaining how one might extract energy from a Kerr black hole (or any black hole with an ergoregion).

The idea is due to Penrose ([50]) and is known as the **Penrose process**. The region S in blocks I and III where $X = \partial_t$ is spacelike is called the ergosphere. These form enveloping zones around block II that becomes thinner at higher latitudes, leaving the poles uncovered. The idea here is that it is impossible for a particle to follow the integral curves of the field ∂_t, i.e. to remain at rest when viewed from infinity. More precisely, it is impossible for (approximately inertial) observers at infinity, which follow the timelike integral curves of the Killing field $X = \partial_t$, to synchronize their clocks with particles in the ergosphere.

We can now throw a particle from the infinity into the ergosphere. Since the particle follows a geodesic and X is a Killing field, $E = -p^\mu X_\mu > 0$ is a conserved along its worldline, where $p = mu$ is the 4-momentum (tangent to the worldline). Assume now that the particle splits into two particles with momenta p_1 and p_2, where $p = p_1 + p_2$. Since X is spacelike on the ergosphere, we can chose p_1 to be a future-pointing timelike vector such that $E_1 = -p_1^\mu X_\mu < 0$. Then $E_2 = -p_2^\mu X\mu$ must be greater than E. But this means that the second particle can escape to infinity, where it will have more energy than we gave the original particle, so we have effectively extracted energy from the black hole.

Lastly, we should mention that one can discuss relativistic jets and accretion disks in the context of spinning black holes as described by the Kerr solution. This is an interesting topic, which fits mostly in the domain of numerical simulations (and which I do not intend to treat here). The usual approach treats the Kerr metric as a fixed background on which some (charged) fluid is moving as prescribed by the equation of motion $\nabla \cdot T = 0$ and conservation of matter $\nabla \cdot (nu) = 0$, where n is particle number density and u 4-velocity of matter. In the simplest case, we can take the stress-energy tensor $T_{\mu\nu}$ to be the sum of a perfect fluid part $T^{\mu\nu}_{PF} = (\rho + p)u^\mu u\nu + pg^{\mu\nu}$ and an electromagnetic part $T^{\mu\nu}_{EM} F^{\mu\alpha} F^\nu_\alpha - \frac{1}{4} g^{\mu\nu} F^{\alpha\beta} F_{\alpha\beta}$. For more information on these kinds of models see e.g. Gammie et al. [48] and [49].

Chapter 3

Causality and Global Hyperbolicity

Causality theory was instrumental in proving certain theorems about black holes (see chapter 4) and providing a language in which a general theory could then be developed systematically. It is therefore crucial that we introduce some elementary notions of causality theory, of which the most important are: *causal* and *chronological future/past*, *causality conditions* (*causal*, *strongly causal* and especially *globally hyperbolic*), *achronal sets*, *Cauchy surfaces*, *Cauchy developments* and *Cauchy horizons*. As a main result, we prove a characterization theorem for globally hyperbolic spacetimes.

3.1 Causality Conditions

Definition 18.

1. We write $p \ll q$ if there exists a (future-pointing) timelike curve from p to q and say that p and q are **chronologically connected** (they can be experienced in succession by some observer). For a subset $A \subset M$ we write $I^+(A) = \{q \in M \mid (\exists p \in A)(p \ll q)\}$ and call $I^+(A)$ the **chronological future** of A.

2. Similarly, we write $p < q$ if there exists a (future-pointing) causal curve (i.e. lightlike or timelike) from p to q and say p and q are **causally connected** (an observer can send a signal at event p that will be received by another observer at q thereby influencing that observer). As usual we write $p \leq q$ to mean either $p < q$ or $p = q$. For a subset $A \subset M$ we define $J^+(A) = \{q \in M \mid (\exists p \in A)(p \leq q)\}$, which we call the **causal future** of A.

We can, of course, completely analogously define chronological and causal pasts (denoted by $I^-(A)$ and $J^-(A)$ respectively). There are some reasonable restrictions one can impose on chronology and causality on M:

Definition 19. We say that M satisfies the **chronology condition** if M contains no closed timelike curves, so that an observer cannot return to his past. In particular, this means that $\ll$ must be irreflexive ($p \ll q \Longrightarrow \neg(q \ll p)$) so that $(M, \ll)$ is a partially ordered set. Similarly, we say that M satisfies the **causality condition** (or is causal) if there are no closed causal curves, so that an observer cannot signal anyone in the past.

It is interesting to note that compact spacetimes necessarily have closed timelike loops, i.e. violate both the chronology and causality conditions. This is easy to see:

Lemma 20
$I^+(p)$ are open.

Proof. Choose $q \in I^+(p)$. We show that q is contained in an open set, which itself is contained in $I^+(p)$. We first take a timelike curve from p to q and follow it until we reach a point r sufficiently close to q so that q is in a normal neighborhood of r. q can now be reached by a timelike geodesic from r (a straight line from r to q in normal coordinates). The exponential function at r, $\exp_r$, sends a small open cone (focused around a line joining q and r) with vertex at r (i.e. $0 \in T_rM$) diffeomorphically to an open set in M, which contains q. This open set is contained in $I^+(p)$.

□

Proposition 21
Let M be a compact spacetime, then M is not causal.

Proof. $\{I^+(p) \mid p \in M\}$ is an open cover, which, by compactness, must have a finite subcover $I^+(p_1), ..., I^+(p_n)$. We can assume that $I^+(p_1)$ is not contained in any other $I^+(p_i)$ (otherwise we can just discard it and still have an open subcover) so no other $I^+(p_i)$ can contain p_1 (for then we would have $I^+(p_1) \subset I^+(p_i)$). But now we must have $p_1 \in I^+(p_1)$ and so there is a closed timelike loop from p_1 back to p_1. □

One can impose some even stronger conditions on M:

Definition 22.

1. We say M satisfies the **strong causality condition** at p if given any neighborhood $U \subset M$ of p, there is a neighborhood $V \subset U$ of p such that no causal curve intersects V more than once. In other words, any causal curve with endpoints in V lies in V.

2. We say M is **stably causal** if there exists a continuous nonvanishing timelike vector field X such that the metric $g - X^\flat \otimes X^\flat$ contains no closed timelike loops.

Intuitively, the strong causality condition is saying that causal curves which start arbitrarily close to p and leave some neighborhood of p cannot return arbitrarily close to p, i.e. there are no "almost" closed timelike curves at p.

On the other hand, stably causal spacetimes are precisely those whose metric is causal even after a small perturbation. To see this let g be a metric on M, $p \in M$ and X a timelike vector at p, whose dual is $X^\flat = g(X, \cdot)$. Define a new metric on T_pM by $\tilde{g} - X\flat \otimes X^\flat$. The light cone of $\tilde{g}$ is strictly larger than the lightcone of g (every timelike and null vector of g is a timelike vector of $\tilde{g}$). Therefore if we "open up" the lightcone at every point and still do not find a closed timelike curve, we have a stably causal spacetime. There is an alternative characterization of stable causality, which we state, but will not use (or provide proof of):

Proposition 23
M is stably causal iff there exists a globally defined continuous function $f : M \to \mathbb{R}$, which is (strictly) increasing along every future-directed causal curve.

Proof. See Hawking & Ellis [3] (proposition 6.4.9) □

We call such an f a **global time function**.

3.2 Convex Sets

We should first note a couple of things. An open neighborhood C is called *convex* if it is an a normal neighborhood of all of its points. So for any $p \in C$ the exponential map is a diffeomorphism from some open $U_p \subset T_pM$ to C. It is well known that around any $p \in M$ convex neighborhoods exist (see e.g. [14]).

It is useful to introduce the following notation: If γ is a geodesic from $p = \gamma(0)$ and $\gamma(1) = q$, then $\overrightarrow{pq}$ is a vector $\gamma'(0) \in T_pM$. Define a map $\Delta : C \times C \to TM, \quad (p, q) \to \overrightarrow{pq}$. Note that it is the inverse of $E : \Delta(C \times C) \to C \times C, \quad v \mapsto (\pi(v), \exp(v)) = (p, \exp_p(v))$. Since E is a diffeomorphism, Δ is smooth.

Convex neighborhoods have particularly nice causal properties:

Lemma 24
Let C be a convex neighborhood and put $I^{\pm}_C(p) = I^{\pm}(p) \cap C$ (similarly for $J^{\pm}_C(p)$) then:

1. *For $p \neq q$ in C, we have $q \in J^+_C(p)$ iff $\overrightarrow{pq}$ is a future-pointing causal vector. A complete analogue holds for I^+ (replacing the word "causal" with "chronological").*
2. *Closure of $I^{\pm}_C(p)$ in C is $J^{\pm}_C(p)$.*
3. *Causality relation $\leq$ is closed in C: if $p_n \to p$ and $q_n \to q$ $(p_n, q_n, p, q \in C)$, then $q_n \leq p_n$ $(q_n \in J^+_C(p_n))$ holding for all n implies $q \leq p$ $(q \in J^+_C(p))$*
4. *A causal curve in a compact subset of a convex neighborhood C is continuously extendible.*

Note that 1 shows C to be causal. If $p \in J^+_C(q)$ and $q \in J^+_C(p)$, then $\overrightarrow{pq}$ is both past-pointing and future-pointing; a contradiction.

Proof.

- 1 follows because if $q \in J^+_C(p)$, then we can connect p and q with a geodesic starting at p. This geodesic is generated by some unique $\overrightarrow{pq} \in T$, which therefore must be causal. Conversely, if $\overrightarrow{pq} \in T$ is a causal future-pointing vector, then it generates a causal geodesic connecting p and q.
- 2 follows because the closure of $I^{\pm}(p)$ is indeed $J^{\pm}(p)$ in Minkowski space T_pM. On the other hand, exp is a diffomorphism and sends $I^{\pm}(p) \cap U$ to $I^{\pm}_C(p)$ and $J^{\pm}(p) \cap U$ to $J^{\pm}_C(p)$ (where U is the domain of $\exp_p$ on T_pM).
- 3 follows from 1 and the fact that $(p, q) \mapsto \overrightarrow{pq}$ is a continuous map (which therefore must preserve limits).
- Finally, we prove 4. Let γ be a curve contained in a compact set $K \subset C$. We assume γ is defined on $[0, b)$, $b \leq \infty$ and cannot be further extended. In particular, given a sequence of points $s_i \to b$ and $t_i \to b$ we cannot consistently assign a value to b, i.e. $\lim_i \gamma(s_i) \neq \lim_i \gamma(t_i)$. We now show this is false. First note that, since K is compact, we can find a sequence $s_i \to b$ such that $\gamma(s_i)$ converges to some $p \in K$. Let now $t_i \to b$ be a set of points converging to b with $\gamma(t_i) \to q \neq q$. By 3, this must mean that $q \in J^+_C(p)$ and $p \in J^+_C(q)$, which is impossible.

□

Lemma 25
Generally, we have $I^+(J^+(p)) = I^+(p)$.

This means that, if we can reach r by first going from p to q via some causal curve, followed by going from q to r via some chronological curve, then we can reach r from p by a curve that is chronological all the way. We now see that $J^+(p) \subset \overline{I^+(p)}$. Indeed, take $q \in J^+(p)$ and a sequence $q_n \in I^+(q)$ converging to q; then $q_n \in I^+(p)$ so $q \in \overline{I^+(p)}$.

Proof. Let us thus consider the simplest (nontrivial) situation. We have a causal curve from p to q ($q \in J^+(p)$) lying inside a convex set V and then a chronological curve γ from q to r ($r \in I^+(q)$) also lying inside a convex set U. Now we go from r to q via the (past-directed) chronological curve γ. We stop as soon as we are inside set V - at some point $\gamma(s) = x$. Point q is in $J^+(p)$, but, at least inside convex neighborhoods, this means that any neighborhood of q must intersect $I^+(p)$. In particular, $I^-(x)$ for some x slightly in the future of q must intersect $I^+(p)$ and so we can go via some chronological (past-directed) from r to x to, finally, p.

Generally, for $r \in I^+(q)$ and $q \in J^+(p)$, we take first a causal trip from p to q, then a chronological one from q to r. Since the whole trip is some curve $[a, b] \to M$, we may (by compactness) cover this curve by finitely many convex neighborhoods U_0 (around q), U_1 (around some $p_1 \in J^-(q)$), $\ldots$, U_N (around p). The argument then goes the same as before: first take a chronological (past-directed) trip from r to q, but stop when you enter $U_1 \cap I^+(p_1)$ - this is event x_1. Then take a chronological (past-directed) trip from x_1 to p_2, but again stop after entering $U_2 \cap I^+(p_2)$, and so on. Eventually we get to p via some past-directed chronological trip. □

An analogous argument shows that we also have $J^+(I^+(p)) = I^+(p)$. Intuitively, by taking a causal path from p to q to r that starts as a timelike segment (from p to q), we cannot reach any other event other than those already available to timelike paths from p in the first place. In other words, that timelike segment ruins the whole path.

Note the following: if $q \in J^+(p) \setminus I^+(p)$, then any causal curve connecting p with q is a null geodesic. This certainly holds in convex neighborhoods [1], but now we may cover the curve with finitely many convex nieghborhoods and conclude it is a null geodesic (up to reparametrization).

3.3 Limits of Causal Paths: a Couple of Technical Lemmata

We say q is an **endpoint** of curve γ if γ enters and stays inside every neighborhood of q. More precisely, q is a future endpoint if for every open neighborhood U of q we can find some t such that for $s > t$ we have $\gamma(s) \in U$. A past endpoint is defined analogously. γ is future (reps. past) **inextendible** if it has no future (resp. past) endpoint. An inextendible curve $\gamma : (a, b) \to M$ is therefore one for which there exists no limit when t goes to a or b (else one could extend the curve to include that limit as well).

We approximate a causal curve with some discrete points:

[1] Since we cannot reach q by a causal curve from points in $I^+(p)$ (because $J^+(I^+(p)) = I^+(p)$), any curve connecting p with q must lie on the light cone $J^+(p) \setminus I^+(p)$. The light cone is a null hypersurface so a causal curve lying on this hypersurface will be a null geodesic up to reparametrization.

Definition 26. Let $\mathcal{C}$ be a covering of M with convex neighborhoods and γ_n a sequence of future-directed causal curves in M. A limit sequence for γ_n relative to $\mathcal{C}$ is a (finite or infinite) sequence of points $p_0 < p_1 < p_2 < \dots$ in M, which satisfy the following:

- For each p_i we can find a subsequence γ_m, and for each m we can find numbers $t_m^{(0)} < t_m^{(1)} < \dots < t_m^{(i)}$ such that $\lim_{m\to\infty} \gamma_m(t_m^{(j)}) = p_j$ for all $j = 0, \dots i$.
- For each $j < i$, the points p_j, p_{j+1} and the segments $\gamma_m|_{[t_m^{(j)}, t_m^{(j+1)}]}$ are contained in a single $C_j \in \mathcal{C}$
- If the sequence p_i is infinite, it must not converge. If it is finite, it must have more than one point and be maximal in the sense that no sequence that is strictly longer should satisfy the previous two points.

Connecting points p_i with geodesics we get the quasi-limit of γ_n. This is a broken geodesic which can be treated as an approximate limit (the finer the convex covering, the better the approximation).

Lemma 27
Let γ_n be a sequence of future-pointing causal curves which satisfy $\gamma_n(0) \to p$, but γ_n do not accumulate around p (there is a neighborhood of p which contains only finitely many curves γ_n).

Proof. As M is paracompact, $\mathcal{C}$ has a locally finite refinement. In other words, we can find a covering $\mathcal{C}'$ of M with open sets B (we can assume B is so small that $\overline{B}$ is compact), such that each B is contained in some member of $\mathcal{C}$ and any point $x \in M$ has a neighborhood which intersects only finitely many $B \in \mathcal{C}'$. The assumptions of this proposition guarantee that $\mathcal{C}'$ contains (or can be made to contain) a B_0 such that infinitely many curves γ_n start in B_0 and leave $\overline{B}_0$. Call these curves $\gamma_n^{(1)}$.

Let $\gamma_n^{(1)}(t_n)$ be points at which $\gamma_n^{(1)}$ intersect ∂B_0 for the first time. since ∂B_0 is compact, we may pass to a subsequence of $\gamma_n^{(1)}(t_n^{(1)})$, which converges to some $p_1 \in \partial B_0$.

We now choose $B_1 \in \mathcal{C}'$ containing p_1. If again infinitely many $\gamma_n^{(1)}$ leave B_1, we repeat the argument and find a subsequence $\gamma_n^{(2)}$ whose earliest intersections with ∂B_1 converge to some point $p_2 \in \partial B_1$. We repeat this as many times as possible, but with the following condition: if it possible to chose multiple elements of $\mathcal{C}'$ as B_i (i.e. multiple elements contain p_i), then pick the one which has been chosen the fewest times before.

In such a way we have obviously constructed a sequence of points p_i such that $\lim_{m\to\infty} \gamma_m(t_m^{(j)}) = p_j$ and the segments $\gamma_m|_{[t_m^{(j)}, t_m^{(j+1)}]}$ are contained in a single $C_j \in \mathcal{C}$, where C_j is any member of $\mathcal{C}$ that contains B_j. Since relation $\leq$ is closed on C_i (and $\gamma_n^{(i)}(t_n^{(i)}) \leq \gamma_n^{(i+1)}(t_n^{(i+1)})$), we have $p_i \leq p_{i+1}$. But the construction guarantees $p_i \neq p_{i+1}$, so $p_i < p_{i+1}$.

If the sequence p_i is infinite, we must show that it is not convergent. Assume to the contrary that $p_i \to q$ and pick a $b \in \mathcal{C}'$ containing q, then $p_i \in B$ for all but finitely many $i \geq 0$. Since $\mathcal{B}$ has compact closure and $\mathcal{C}'$ is locally finite, only finitely members of $\mathcal{C}'$ intersect B. Thus only finitely many members of $\mathcal{C}'$ cover the tail of the sequence p_j, so one set must have been chosen as B_i for infinitely many i. This is impossible because B was the candidate infinitely many times but was chosen only finitely many times. Indeed, because it contains the tail of the sequence p_i, only finitely many can be in the ∂B.

If, on the other hand, the sequence is some finite set $p = p_0 < \dots < p_k$, then we must show it is maximal, i.e. cannot be extended while satisfying the first two points in

the definition of a limit sequence. In fact, this may not be the case, but the extension can have at most one more point. Since the sequence ends after k steps, only a finite number of curves $\gamma_n^{(k)}$ leave B_k. Let γ_m be those curves trapped in B_k, then by lemma 24 they are extendible. We thus may assume γ_m are defined on $[0, b_m]$ (where $\gamma_m(b_m)$ is the future endpoint of γ_m), but then we can find a subsequence for which $\gamma_m(b_m)$ converge to some $q \in \overline{B}_k$. If $q = p_k$, then $p = p_0 < \cdots < p_k$ is maximal and if $q \neq p$, then $p = p_0 < \ldots p_k < q$ is maximal. □

Lemma 28
Let $K \subset M$ be a compact set on which strong causality holds. If γ is a future-inextendible causal curve starting in K, then γ eventually leaves K and doesn't return. More precisely, there exists $s > 0$ such that $\gamma(t) \notin K$ for all $t \geq s$.

Proof. Aiming for a contradiction, let us assume that the conclusion is false. Then either γ remains in K for all times, or leaves but persistently returns. Assume the domain of γ is $[0, b)$, where $b \leq \infty$.

We thus conclude that there is a sequence of numbers $t_i \in [0, b)$ such that $t_i \to b$ and $\gamma(t_i) \in K$. As K is compact, the sequence $\gamma(t_i)$ has a subsequence which converges to some $p \in K$. As γ is future-inextendible, there must exist a sequence $s_i \to b$ such that $\gamma(s_i)$ does not converge to p (otherwise we would be able to extend γ to include b).

We can suppose that some neighborhood U of p contains no $\alpha(s_j)$; take a subsequence of s_j if needed, which converges to some $q \in K$. Since both s_i and t_i converge to b, we can find subsequences that alternate: $t_1 < s_1 < t_2 < s_2 < \ldots$. But now we see that the curves $\gamma|_{[t_i, t_{i+1}]}$ must exit neighborhood U (to reach $\gamma(s_i)$) and return (to reach $\gamma(t_{i+1})$). These curves are therefore almost closed at p, but this contradicts strong causality on K. □

Lemma 29
Let $K \subset M$ be a compact subset on which strong causality holds. Let γ_n be a sequence of future-pointing causal curves in K for which $\gamma_n(0) \to p$ and $\gamma_n(1) \to q \neq p$. Then there exists a future-pointing broken causal geodesic γ from p to q and a subsequence γ_m of γ_n for which $\lim_{m\to\infty} l(\gamma_n) \leq l(\gamma)$.

Proof. By applying lemma 27 to sequence γ_n, we get a limit sequence p_i starting at p. Assume for the moment p_i to be infinite, then the corresponding quasi-limit, the broken geodesic γ, is a future inextendible causal curve starting at p. It thus must leave K never to return. This means, in particular, that some p_i is not in K, which implies that γ_n leave K, which is contrary to the assumptions of the theorem.

Thus p_i is a finite sequence, which starts at p and ends at $q = \lim_n \gamma_n$. The quasi-limit γ which passes through p_i is thus a broken geodesic from p to q. Now restrict to a particular convex set C_i, where the i-th segments of γ_n live. The length of these segments are bounded by the separation vectors between the $p_m^{(i)} = \gamma_m(t_m^{(i)})$ in C_i (because geodesics give maximal length):

$$l(\gamma_n|_{[s_m^{(i)}, s_m^{(i+1)}]}) \leq |\overrightarrow{p_m^{(i)} p_m^{(i+1)}}|.$$

Summing over i we get:

$$l(\gamma_n) \leq l_n = \sum_{i=0}^{k} |\overrightarrow{p_m^{(i)} p_m^{(i+1)}}|.$$

Since $\overrightarrow{pq}$ depends continuously on (p, q), the norm $|\overrightarrow{pq}|$ does as well. Thus l_n converges to $\sum_{i=0}^{k} |\overrightarrow{p_i p_{i+1}}| = l(\gamma)$. Taking a subsequence if needed, we get the result. □

So far, for a given convex covering, we have managed to find a set of points p_i which give a quasi-limit of some sequence of curves. It turns out one can even find a continuous curve all of whose points behave as p_i do.

To explain this, though, we must extend the notion of being causal or timelike to continuous curves as well. We say a continuous curve γ is causal (resp. timelike), if any two sufficiently close points on γ can be connected by a causal (resp. timelike) smooth curve. More precisely, we say γ is future-directed causal (resp. timelike) if for any $p \in \operatorname{Im}\gamma$ there exists an open neighborhood U of p such that if $\gamma(t_1), \gamma(t_2) \in U$ and $t_1 < t_2$, then there exists a future-directed causal (resp. timelike) smooth curve from $\gamma(t_1)$ to $\gamma(t_2)$.

Thus by allowing continuous curves we do not change our notion of causality (if there exists a continuous causal curve between p and q, then there also exists a smooth one).

Definition 30. A point p is a **limit point** of γ_n if every open neighborhood of p intersects infinitely many γ_n. We say γ is a **limit curve** of γ_n if each $p \in \gamma$ is a limit point.

We now have:

Lemma 31
Let γ_n be a sequence of future-inextendible causal curves with limit point p. Then there exists a causal future-inextendible limit curve γ passing through p.

In fact, from this proposition we get our previous one: Cover the limit curve by a finite set of convex neighborhoods, then chose some set of points on the limit curve within each one of those neighborhoods and connect them with geodesics; this gives a quasi-limit.

Proof. We follow Hawking & Ellis [3], but the argument is quite similar to the proof of 27.

Let C_1 be a convex neighborhood around p with compact closure and $B_q(r) \subset C_1$ a ball (in normal coordinates at p) with radius r and center p. We chose some subsequence γ_m converging to p and, as before, pick a further subsequence converging to some point $x_{11} \geq p$ in $\partial B_p(r)$.

We have thus found a limit point in the future of p as in the proof of lemma 27. Now, we wish for an entire limit curve, rather than just a discrete set of points which may serve as a quasi-limit. Here, instead of going farther into the future (as we did before), we must build the limit points between p and x_{11}.

Indeed, consider balls whose radii are rational multiples of r, i.e. $B_p(ir/j)$, where $i, j \in \mathbb{N} = \{1, 2, 3 \dots\}$ are relatively prime and $i < j$.
At each step $(i, j) = (1,2), (1,3), (2,3), (1,4), (2,4), (3,4), \dots$ we extract a limit $x_{ij} \in \partial B_p(ir/j)$ of some subsequence γ_m. In the next step we extract a further sub-subsequence and so on. We can therefore define a curve $\gamma : [0,1] \cap \mathbb{Q} \to C_1, \quad \frac{i}{j} \mapsto x_{ij}$. It is obvious that for $a, b \in [0,1] \cap \mathbb{Q}$ and $a < b$ we have $\gamma(a) \leq \gamma(b)$, so the curve is causal.

Note that, by the above construction, γ is continuous. Finally, we can take the closure of this curve $\gamma : [0,1] \to C_1$: for $x \in [0,1]$ take some sequence of rational numbers $a_i \to x$ and set $\gamma(x) = \lim_i \gamma(a_i)$. Continuity of γ guarantees that $\gamma(x)$ does not depend on the sequence a_i chosen. It is also obvious that each point on γ is a limit point of γ_n.

We may continue now by choosing some C_2 around $x_{11} = p_1$ and doing the same thing all over again. Continuing inductively for $p_2, p_3 \dots$ we get a limit curve which is inextendible. □

3.4 Achronal Sets

We start off with a couple of definitions:

Definition 32. We call a set $A \subset M$ **achronal** if $p \ll q$ never holds for any $p, q \in A$, i.e. if no two points in A are comparable in the relation $\ll$. Similarly, A is **acausal** if $p < q$ never holds for any $p, q \in A$.

We note here that the closure of an achronal set is achronal. If that were not the case, we would have $p, q \in \overline{A}$ with $p \ll q$, but then $I^-(q)$ is an open neighborhood of p and so must intersect A - let $x \in A \cap I^-(q)$. We now have $q \in I^+(x)$, but then $I^+(x)$ is an open neighborhood around q, which intersects A, contradicting the achronality of A.

Definition 33. For (an achronal) $A \subset M$ we define the **future Cauchy development** $D^+(A)$ as the set of all points $p \in M$ such that every past inextendible causal curve through p meets A. Similarly, define the **past Cauchy development** of A, $D^-(A)$. Then $D(A) = D^+(A) \cup D^-(A)$ is called simply the **Cauchy development** of A.

Intuitively, $D^+(A)$ is the part of causal future of A that is predictable from A (as no past-inextendible causal curve can reach $q \in D^+(A)$ without first going through A).

Definition 34. For a (closed achronal) set A, the **future Cauchy horizon** is $H^+(A) = \overline{D^+(A)} \setminus I^-(D^+(A))$.

In other words, $H^+(A)$ contains precisely those $p \in \overline{D^+(A)}$ for which $I^+(p)$ does not meet $D^+(A)$, i.e. from which we cannot reach (via a timelike curve) any point predictable from A. This separates $D^+(A)$ from the rest of $J^+(A)$. $H^-(A)$ is defined completely analogously.

Note that for an open U, we have $U \subset I^-(U)$ ($p \in U$ implies $q \in I^+(p)$ for some $q \in U$; we see this using normal neighborhoods). This means that the interior of $D^+(A)$ does not intersect $H^+(A)$; in other words $H^+(A)$ is contained in $\partial D^+(A)$.

In fact, we shall see that $H^+(A)$ is more or less the boundary of $D^+(A)$ (excluding A itself).

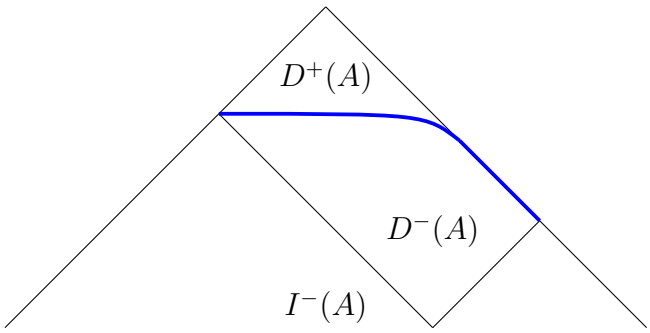

Figure 3.1: Cauchy development of some achronal set A (blue). Note that A is not acausal (a piece of it is contained in the null cone)

Lemma 35
Let $C \subset M$ be a closed set and γ a past-inextendible causal curve starting at p that does not meet C ($\operatorname{Im}\gamma \cap C = \emptyset$). If $q \in I^+(p) \cap M \setminus C$, then there exists a past-inextendible timelike curve starting at q that does not meet C.

Proof. We work inside the open set (submanifold) $M \setminus C$. The idea is to push γ slightly into the future. As γ is past inextendible, we may take its domain to be $(-\infty, 0]$ and that the sequence $\gamma(n)$, $n \in \mathbb{N}$ does not converge. Put $p_0 = q$. Since $\gamma(t)$ is past directed and $p_0 = q \in I^+(p)$, we have $\gamma(1) \leq \gamma(0) \ll p_0$, but we have seen that $I^+(J^+(\gamma(1))) = I^+(\gamma(1))$, so $\gamma(1) \ll p_0$.

We now find a point p_1 between $\gamma(1)$ and p_0. Inductively, we find a sequence p_n such that $\gamma(n) \ll p_n \ll p_{n-1}$. We chose p_n so close to $\gamma(n)$ so that p_n do not converge[2].

Now we just connect the events p_n to get a past inextendible timelike curve α in $M \setminus C$ with $\alpha(0) = p_0 = q$. □

Some authors (e.g. [32]) use timelike curves in the definition of $D^+(A)$. The following result shows that (at least for closed sets) the two definitions are essentially equivalent:

Proposition 36
Let A be a closed achronal set. Then $p \in \overline{D^+(A)}$ iff every past-inextendible timelike curve from p passes through A.

Proof. Assume every past-inextendible timelike curve from q passes through A, then either $q \in I^+(A)$ or $p \in A$. We must show that every open neighborhood of q intersects $D^+(A)$. If $q \in A \subset \overline{D^+(A)}$, there is nothing to prove, so assume $q \in I^+(A)$.

Let $p \in I^-(q) \cap I^+(A)$ and suppose $p \notin D^+(A)$, i.e. suppose some past inextendible causal curve γ starting at p does not meet A. Then lemma 35 gives us an inextendible timelike curve through q that does not meet A. But this contradicts our assumption, so we must have $p \in D^+(A)$. As every open neighborhood of $q \in I^+(A)$ intersects $I^-(q) \cap I^+(A)$, we have that every open neighborhood if q intersects $D^+(A)$, i.e. $q \in \overline{D^+(A)}$.

Conversely, let $q \in \overline{D^+(A)}$. To arrive at a contradiction let us further assume that there exists a past-inextendible timelike curve γ through q which does not meet A. We then have $q \notin A$ and, as A is closed, q has a convex neighborhood C disjoint from A. We now move from q in the past direction on γ to $x \in C$. Then $I^+_C(x)$ contains q and some point $p \in D^+(A)$ (q is assumed to be in the closure). Following a geodesic from p to x and then curve γ, gives a past-inextendible timelike curve through p that does not meet A (contradicting $p \in D^+(A)$). □

Proposition 37
If A is a closed achronal set, then $H^+(A) \cup A = \partial D^+(A)$

Proof. We have seen that in general $H^+(A) \cup A \subset \partial D^+(A)$, so we only need to prove the converse. If $p \in \partial D^+(A) \setminus A \subset \overline{D^+(A)}$, then the preceding proposition gives $p \in I^+(A)$. If we assume $p \notin H^+(A)$, i.e. $p \in I^-(D^+(A))$, then there is a point $q \in I^+(p) \cap D^+(A)$. This however means that $I^+(A) \cap I^-(q)$ is a neighborhood of p contained in $D^+(A)$, contradicting the fact that p is a boundary point. □

Definition 38. The **edge** of an achronal set A is the set of all points $p \in \overline{A}$ such that every neighborhood U of p contains a timelike curve going from $I^-_U(p)$ to $I^+_U(p)$ that does not meet A.

More precisely, for every neighborhood U of p there exist $x \in I^-_U(p)$ and $y \in I^+_U(p)$ and a timelike curve γ from x to y contained in U that does not meet A.

[2]Choose a metric d on M and choose p_n so as to satisfy $d(p_n, \gamma(n)) < \frac{1}{n}$.

Let X be a topological manifold. Recall that $S \subset X$ is a topological hypersurface if around any $p \in S$ we can find a neighborhood U in X and a homeomorphism ϕ from U to an open set in $\mathbb{R}^n$ such that $\phi(U \cap S) = \phi(U) \cap \Pi$, where Π is a hyperplane in $\mathbb{R}^n$.

Proposition 39
An achronal set A is a topological hypersurface iff A and the edge of A are disjoint (that is A does not contain any of its edge points).

Proof. Let A be a topological hypersurface and $p \in A$. We show p cannot be an edge point. Take a neighborhood U so small as to be connected and $U \setminus A$ has 2 components. $I^{\pm}_U(p)$ are open connected sets that do not meet A (as A is achronal), so they must be contained in different components of $U \setminus A$ (this certainly holds if we take U to be convex, by making it smaller if needed). This shows p cannot be an edge point, as any curve in U must cross A going from one component to the other.

Conversely assume A contains none of its edge points. Take $p \in A$ and take some (smooth) coordinate system $\phi = (x^0, x^1, x^2, x^3)$ on U around p with $\partial/\partial x^0$ timelike and future-pointing. We can now find (a possibly smaller) neighborhood V such that $\phi(V)$ has the form $(a - \varepsilon, b + \varepsilon) \times N \subset \mathbb{R}^1 \times \mathbb{R}^3 = \mathbb{R}^4$, where $\{x^0 = a\} \cap V$ is in $I^-_U(p)$ and $\{x^0 = b\} \cap V$ is in $I^+_U(p)$.

For a sufficiently small neighborhood U and $y \in N$, the x^0 curve $t \mapsto \phi^{-1}(t, y)$ ($t \in (a, b)$) must meet A, as p is not an edge point. As A is achronal this meeting point must be unique; we call its x^0 coordinate $f(y)$.

We shall show $f : N \to (a, b)$ is continuous; it then immediately follows that

$$(u^0, u^1, u^2, u^3) = \psi = (x^0 - f(x^1, x^2, x^3), x^1, x^2, x^3)$$

is a homeomorphism which carries $A \cap V$ to $\{u^0 = 0\} \cap \psi(V) \subset \mathbb{R}^4$. Thus, A is a topological surface.

Let $y_n \to y$, it is sufficient to prove $f(y_n) \to f(y)$. Assume to the contrary, $f(y_n)$ does not converge to $f(y)$, then some subsequence $f(y'_m)$ converges to some $r \neq f(y)$ (values of f are bounded). We now see that either time r is (strictly) bigger than $f(y)$, or (strictly) smaller. In other words, $\phi^{-1}(y, r)$ is either in the (chronological) future of $q = \phi^{-1}(y, f(y))$ or in the past, so $\phi^{-1}(y, r)$ lands in the open set $I^-_V(q) \cup I^+_V(q)$. Therefore, the same must be true for some $\phi^{-1}(y_m, f(y_m)) \in A$, contradicting the achronality of A.

In fact, f, i.e. the time coordinate of y, is actually a Lipschitz continuous map of x^1, x^2, x^3, since no two points on A have a timelike separation. Indeed, for any $x = (x^0, \overrightarrow{x})$ and $y = (y^0, \overrightarrow{y})$ on A, x must lie outside the lightcone of y, so the Minkowski norm is $-|x^0 - y^0|^2 + |\overrightarrow{x} - \overrightarrow{y}|^2 \geq 0$, i.e. $|x^0 - y^0| \leq |\overrightarrow{x} - \overrightarrow{y}|$, where the norm on the right is 3D euclidean. □

Corollary 40
An achronal set A is a closed topological hypersurface iff the edge of A is empty.

Proof. If A is a closed hypersurface, then by the preceding proposition A and the edge of A are disjoint, but $A \subset \overline{A} = A$, so the edge is actually empty.

Suppose conversely that the edge is empty, then A is a topological hypersurface by the preceding proposition. We only need to show A is closed, and this follows from the fact that all boundary points in $\overline{A} \setminus A$ must be edge points. Indeed, as $\overline{A}$ is achronal, if $q \in \overline{A} \setminus A$, then no timelike curve through q can meet A. □

We call $F \subset M$ a **future set** if $I^+(F) \subset F$. For example, $J^+(S)$ is a future set for any $S \subset M$. If F is a future set, then $M \setminus F$ is a past set ($I^-(M \setminus F) \subset M \setminus F$).

Corollary 41
Boundary of a future set F is a closed achronal topological surface.

Thus, in particular, $\partial J^{\pm}(S) = \partial I^{\pm}(S)$ is a topological surface.

Proof. Take $p \in \partial F$. If $q \in I^+(p)$, then $I^-(q)$ is an open neighborhood of p so must contain some point in F. Thus, $q \in I^+(F) \subset F$, from which we conclude $I^+(p) \subset F$ and by a completely analogous argument, we also get $I^-(p) \subset M \setminus F$. We thus see $I^+(\partial F)$ and $I^-(\partial F)$ are disjoint, so ∂F must be achronal. We also see that ∂F must have empty edge, since $I^+(p)$ (as an open set) is actually contained in the interior of F, and $I^-(p)$ in the exterior of F. The result thus follows from the previous corollary. □

3.5 Cauchy Surfaces

Definition 42 (Cauchy surface)**.** A **Cauchy surface** is a subset $S \subset M$ that is met exactly once by every maximally extended (i.e. inextendable) timelike curve in M.

Example 43.

1. A most trivial example one can immediately think of is any spacelike plane in Minkowski space.

2. In Schwarzschild spacetime as well, one can immediately (in the exterior solution) see that taking the $t = 0$ slice gives a Cauchy surface. However, it is not entirely clear that one can find a surface spanning the interior as well. One sees that this is the case by considering the maximally extended Kruskal spacetime and taking a $t = 0$ slice ([32]).

3. In Kerr region I, again the $t = const$ are Cauchy surfaces; However, here it is even less obvious that there is a Cauchy surface for the region $I \cup II$. For a construction see [54] proposition C.11. By considering the Cauchy development of such a Cauchy surface, we see that the interior horizon is a Cauchy horizon.

Proposition 44
Cauchy surface S (if it exists) must be a closed achronal topological hypersurface (a 3-manifold) and, in fact, it is met by every inextendible causal curve.

Proof. Immediately from definition we have that S must be achronal and that M must be the disjoint union of $I^+(S)$, S and $I^-(S)$. A timelike curve passing through S must intersect both $I^+(S)$ and $I^-(S)$, so S is the common boundary of the future sets $I^+(S)$ and $I^-(S)$. At this point corollary 41 guarantees that S is a topological hypersurface.

Now, we only need to show that S is met by every inextendible causal curve γ. Assume to the contrary, γ does not meet S. Then either $\gamma(0) \in I^+(S)$ or $\gamma(0) \in I^-(M)$ - we assume $\gamma(0) \in I^+(S)$. Lemma 35 now gives a past-inextendible curve α starting in $I^+(S)$ that does not meet S; a contradiction. □

Note that an achronal set S is a Cauchy surface iff $D(S) = M$. Thus, one can think of $D(A)$ as the largest subset for which A plays the role of a Cauchy surface.

Proposition 45
Any two Cauchy surfaces (if they exist) must be homeomorphic. Additionally, any Cauchy surface of a connected spacetime M is itself connected.

Proof. Since we assume M to be time orientable, let X be a global (nowhere vanishing) timelike vector field on M [3]. Then a maximal integral curve through $p \in M$ of X meets S at a unique point $\psi(p)$. We claim $\psi : M \to S$ is a continuous surjective open mapping leaving S fixed. Indeed, maximal integral curves of X must be inextendible. Let now $\varphi : D \to M$ be the (maximally extended) flow of X. D is an open set in $M \times \mathbb{R}$ and S is a hypersurface in M, so $D_S = (S \times \mathbb{R}) \cap D$ must be a topological hypersurface in D. If we now restrict φ to $\varphi_S : D_S \to M$, then φ_S must be continuous and bijective (S is a Cauchy surface after all).

A continuous bijection need not have a continuous inverse, but recall the following famous result of Brouwer's: if $U \subset \mathbb{R}^n$ is an open subset, $f : U \to \mathbb{R}^n$ an injective continuous map, then $V = f(U)$ is open, and f a homeomorphism between U and V.

Since, φ_S is a mapping between two manifolds of the same dimension, we may apply the above result locally and conclude that φ_S has a continuous inverse, i.e. is a homeomorphism between D_S and M. Note that $\psi(\varphi_S(p,t)) = \psi(\gamma_p(t)) = \gamma_p(0) = p$, where γ_p is the integral curve of X through $p \in S$. This shows $\psi \circ \varphi_S = \pi$, where $\pi : S \times \mathbb{R} \to S$ is the projection mapping. As π is open, continuous and surjective, $\psi = \pi \circ \varphi_S^{-1} : M \to S$ is as well. Clearly $\psi(S) = S$ so S is fixed under ψ. In other words, ψ is a retraction mapping from the manifold M onto S.

In particular, if M is connected, the Cauchy surface must be as well. Finally, let S and Σ be two Cauchy surfaces in M. Then the restrictions $\psi_S|_\Sigma$ and $\psi_\Sigma|_S$ are obviously mutual inverses. □

3.6 Globally Hyperbolic Spacetime

Definition 46. If strong causality condition holds on M and each causal diamond $J(p,q) = J^+(p) \cap J^-(q)$ is compact, we say M is **globally hyperbolic**.

Note that $J(p,q)$ contains all causal curves from p to q. Roughly speaking, the compactness of the causal diamonds $J(p,q)$ then guarantees that no naked singularities can occur (otherwise a "visible hole" in spacetime would make $J(p,q)$ noncompact as one can take a sequence of points in $J(p,q)$ "converging" to the singularity; such a sequence does not have a limit point in $J(p,q)$).

Later we shall see alternative characterizations of global hyperbolicity, but this seems to be the most popular definition and is the one given in O'Neill [1], Hawking & Ellis [3] and Chruściel [6].

In fact, we can replace *strongly causal* by *causal* (or even *distinguishing*) as shown by Bernal and Sánchez in [24]. That it has taken so long to realize this, what amounts to essentially an elementary result, is quite surprising. We therefore give a proof here:

Lemma 47

[3] We should note the following: if M has a nowhere vanishing globally defined vector field X, then there exists a time-orientable Lorentzian metric on M (i.e. a metric which admits a nonvanishing globally defined timelike vector field). Indeed, $g - X^\flat \otimes X^\flat$ is again a Lorentzian metric, but in this metric X becomes timelike.

On the other hand, if M admits a Lorentzian metric, it must have a nonvanishing (globally defined) vector field. Topologically, M has a global nonvanishing vector field iff M is noncompact or M is compact and has Euler characteristic 0. On the other hand this is precisely the obstruction for the existence of a Lorentzian metric (see [1]). Thus, there is no additional obstruction for the existence of a time-orientable Lorentzian metric.

If all causal diamonds $J(p,q) = J^+(p) \cap J^-(q)$ are compact, then $J^\pm(p)$ must be closed for all p.

Note that, when the conclusion of this lemma is satisfied, we have $J^\pm(p) = \overline{I^\pm(p)}$.

Proof. Suppose to the contrary, $J^+(p)$ is not closed, then one can find a sequence $x_n \in J^+(p)$ converging to a point $x \in \overline{J^+(p)} \setminus J^+(p)$. Take some $q \in I^+(x)$, then $I^-(q)$ is an open neighborhood of x. Therefore, for sufficiently large n, we have $x_n \in J^+(p) \cap I^-(q) \subset J^+(p) \cap J^-(q)$. As $J^+(p) \cap J^-(q)$ is compact, the limit of x_n, namely x, should be in $J^+(p) \cap J^-(q)$, but this is not the case as per our choice of x - a contradiction. □

Proposition 48
Let M be a spacetime in which all causal diamonds $J(p,q) = J^+(p) \cap J^-(q)$ are compact (when nonempty). Then the following are equivalent:

1. *M is distinguishing, i.e. $I^\pm(q) \neq I^\pm(p)$ for any $p \neq q$.*
2. *M is causal.*
3. *M is strongly causal.*

Proof. Following [6], we divide the proof into 3 statements:

- First note that any distinguishing spacetime is causal.

  If M were not causal, then we have some closed causal future-directed curve γ. We now take p and q to be two distinct points on this curve. If $x \in I^+(p)$, then we can ride γ from q to p and then some timelike curve from p to x. We therefore have $I^+(p) \subset I^+(q)$. The other inclusion follows by a completely analogous argument, so we get $I^+(q) = I^+(p)$ and M cannot be distinguishing.

- If all causal diamonds $J(p,q) = J^+(p) \cap J^-(q)$ are compact and M is causal, then M must be distinguishing.

  If M were not distinguishing, then there would exist two points $p \neq q$ with $I^+(p) = I^+(q)$. This then gives $q \in \overline{I^+(q)} = \overline{I^+(p)} = \overline{J^+(p)}$ (where the last equality follows from the fact that $\overline{J^+(p)} \subset \overline{I^+(p)}$; the other inclusion is trivial). Now $q \in \overline{J^+(p)} = J^+(p)$ as $J^+(p)$ are closed. By a completely symmetric argument we have $p \in J^+(q)$ as well and, since $p \neq q$, M is not causal.

- Causal spacetime in which all $J^\pm(p)$ are closed must be strongly causal.

  We prove the contrapositive. Suppose that there is a point p at which strong causality is violated. Let C be a convex neighborhood around p. For any open $U \subset C$ around p, we can therefore find a future-directed causal curve γ intersecting U more than once.

  Note that this curve must exit C, i.e. it cannot turn around inside C. Indeed, C has a very nice causal structure, given by normal coordinates, which foliate C. If we only allow curves that completely lie inside of C, then C is strongly causal. This means that the curve γ cannot intersect U more than once, but at the same time remain in C.

  We may therefore construct a sequence of curves γ_i each of which exits C at some $q_i \in \partial C$, enters C again and ends at some point p_i near p, where $p_i \to p$. Taking some subsequence, we can assume that $q_i \to q \in \partial C$.

Taking the limit of some subsequence of γ_i, we get a continuous causal curve from p to q. We thus have $q \in J^+(p)$.

Take some sequence $x_n \in I^-(q)$ (i.e. $q \in I^+(x_n)$) converging to q. We now construct a future-directed causal curve going from x_n to p_i by following some timelike curve from x_n to q_i (for i large enough this is possible since q_i are near q and so will eventually be in $I^+(x_n)$) and then γ_i from q_i to p_i. We now see that $p_i \in J^+(x_n)$ (for i large enough). Since $J^+(x_n)$ is closed, we get $p \in J^+(x_n)$ ($p_i \to p$). In other words, we have $x_n \in J^-(p)$ and as $J^-(p)$ is closed as well, we have $q \in J^-(p)$ ($x_n \to q$).

Finally, note that obviously $q \neq p$ because p is in the interior of C (and q is in the boundary) and, since $q \in J^+(p) \cap J^-(p)$, M cannot be causal.

□

It has only recently been shown by Hounnonkpe and Minguzzi in [26] that, for non-compact spacetimes (which in any case cannot be causal) of dimension > 2, we can even drop the causality condition. The proof of this fact is also given in Chruściel [6].

The crucial fact about globally hyperbolic spacetimes is that they are foliated by Cauchy surfaces:

Theorem 49 (Geroch)
A globally hyperbolic spacetime M must have a Cauchy surface. If fact, it must have a globally defined time function, whose level sets are Cauchy surfaces.

We should note here that if M admits a global time function, then M admits a smooth global time function (see Bernal and Sánchez [23] or [25]). Of course, a smooth f will be increasing along causal curves iff its gradient ∇f is a past-directed timelike vector field.

Proof. As the proof uses some analytic machinery (involving measure theory), we relegate this to appendix B. □

Let $C(p, q)$ be the space of all continuous future-directed causal curves from p to q. We endow $C(p, q)$ with the following topology. Every sufficiently small neighborhood in $C(p, q)$ will contain precisely those curves that are sufficiently close in M. More rigorously, for an open $U \subset M$ we define sets $O_U = \{\alpha \in C(p, q) \mid \operatorname{Im} \alpha \subset U\}$. It is now clear that $O_U \cap O_V = O_{U \cap V}$ so these sets are a topological base (an open set is a union of O_U).

Note that, as M is second-countable, $C(p, q)$ is as well. Indeed, we can find a countable basis for $C(p, q)$ by finding a countable basis U_i for M and just taking O_{U_i}.

Proposition 50
Assume M has a Cauchy surface S, then $C(p, q)$ is compact.

Proof. Since $C(p, q)$ is second-countable, it is sufficient to show that an infinite set of points must have a limit point in $C(p, q)$[4].

Considering the topology on $C(p, q)$, we can use the lemma 31 for this purpose.

[4]Take a second-countable topological space X and a cover U_i of X. We can immediately see that second-countability allows us to take a countable subcover. Now if every infinite set of X has a limit point, we can actually take a finite subcover. Indeed, suppose this is not the case and consider the rising open sets $V_n = \bigcup_i^n U_i$. As there is no finite subcover $X \setminus V_n$ are nonempty so we may choose countably many points $x_n \in X \setminus V_n$. As the set $\{x_n\}$ has a limit point x, we have that V_n for a sufficiently large $n \geq N$ must contain x and therefore all x_n, $n \geq N$; a contradiction.

First consider the case $p, q \in D^-(S)$ (where p and q are in the past to the Cauchy surface S). Let γ_i be a sequence in $C(p,q)$. If we now remove point q from M, we find that γ_i are inextendible curves in $M \setminus \{q\}$ starting at p, so they have a limiting curve γ - this curve is inextendible in $M \setminus \{q\}$ as well. Since no curve γ enters the open set $I^+(S)$, γ doesn't either. Note that γ cannot remain inextendible when we put the point q back. Indeed, since M has a Cauchy surface, an inextendible curve must intersect S and enter $I^+(S)$. Therefore, we may take the point q to be an endpoint of the curve γ. We thus get the desired limit point and $C(p,q)$ is compact.

By a completely analogous argument we get the same result when $p, q \in D^+(S)$.

Finally, the only remaining case is $p \in D^-(S)$ and $q \in I^+(S)$ (p and q are on opposite sides of the Cauchy surface). Take again a sequence of curves γ_n in $C(p,q)$. An identical argument to one before gives us a limit curve γ starting at p and going through S into $I^+(S)$ (though not necessarily ending at q). Choose a point $x \in \operatorname{Im}\gamma \cap I^+(S)$ and take a subsequence $\tilde{\gamma}_m$ so that points on γ between p and x are convergence points. This gives us one half of the desired curve.

To get the other half, invert the argument: consider the sequence of past-inextendible curves $\tilde{\gamma}_m$ in $M \setminus \{p\}$ starting at q. Same as before, we get a limit curve $\tilde{\gamma}$ starting at q and entering $I^-(S)$. Since x was is convergence point of subsequence $\tilde{\gamma}_m$, we find that $\tilde{\gamma}$ must pass through x. Thus, by following γ from p to x and $\tilde{\gamma}$ from x to q, we get the desired limit curve from p to q. □

Proposition 51
If $C(p,q)$ is compact, then $J^+(p) \cap J^-(q)$ is as well.

Proof. We need to prove $J^+(p) \cap J^-(q)$ is compact. To do this, it is sufficient to show that any infinite set of points $x_n \in J^+(p) \cap J^-(q)$ has a limit point in $J^+(p) \cap J^-(q)$. It is obvious how to proceed.

Take a sequence of causal curves γ_n from p to q such that each γ_n passes through x_n. Since $C(p,q)$ is compact, we get a subsequence $\tilde{\gamma}_m$ converging to some continuous causal $\gamma \in C(p,q)$. Since $\operatorname{Im}\gamma$ is compact, we can cover it with finitely many neighborhoods B_i with compact closure (take some small open coordinate balls, then $\overline{B_i}$ is compact) and taking the union of B_i we find an neighborhood U of $\operatorname{Im}\gamma$ with compact closure.

From the definition of topology on $C(p,q)$, it follows that $\operatorname{Im}\gamma_i \in U$ for infinitely many $i \in I \subset \mathbb{N}$. Since $x_i \in \operatorname{Im}\gamma_i$ we get an infinite set of points $\{x_i\}_{i\in I}$ contained in U. This set of points must therefore have a limit point in $\overline{U}$, which is also a limit point of the original set $\{x_n\}_{n\in\mathbb{N}}$. □

Note that if M has a Cauchy surface, then M is causal. Indeed, if we have some closed timelike curve, it can be extended by repeatedly going around, but now this is an inextendible curve and it must therefore intersect each Cauchy surface precisely once. This is of course a contradiction, since the curve is closed.

We can summarize these results as follows:

Theorem 52 (Characterization of Global Hyperbolicity)
Let M be a spacetime. Then the following are equivalent:

1. *M is globally hyperbolic.*
2. *M has a Cauchy surface (this is the definition in Wald [4]).*

3. *M is stably causal and level sets of its time function are Cauchy surfaces. (this is more-or-less the definition in Godinho and Natário [7]).*

4. *M is (strongly) causal and $C(p, q)$ is compact (this seems to be the original definition by Leray).*

Chapter 4

Theorems of Penrose and Hawking

One might expect that the unusual breakdown of the metric and curvature at $r = 0$ in the Schwarzschild solution is somehow due to high degree of symmetry. A similar thing also happens in the highly symmetric FRW spacetime. Celebrated theorems of Penrose and Hawking tell us that this situation is in some sense actually generic.

In proving these results, the outline we follow is roughly that of Godinho and Natário [7]. The central definition is the following:

Definition 53. A spacetime M is called **singular** if it is not geodesically complete, i.e. there exist maximally extended geodesics $\gamma : I \to M$ defined on some open interval $I \neq \mathbb{R}$ (allowed to be infinite).

Intuitively, a singular spacetime has a point beyond which no timelike or lightlike curve can be continued (e.g. $r = 0$ in Schwarzschild coordinates). Though, this "point" (the singularity) is not a part of spacetime so the only way one can test for its presence is to trace out maximally extended geodesics and see if any of them ends abruptly.

Note that we do not mention "curvature blowing up", or "matter density becoming infinite". Certainly, if the curvature blows up, one cannot extend the geodesics past that point, but this behavior is not necessary for a singularity to occur. It is this definition of singularity that is used when proving the singularity theorems.

4.1 Geodesic Congruence

We shall give conditions under which singularities will arise, but to do that we first must discuss geodesic congruence.

Let γ be a timelike geodesic so that the metric is positive definite on the instantaneous rest space $\dot{\gamma}^\perp$. Geodesic deviation (or Jacobi) equation states that $\ddot{J} = R(\dot{\gamma}, J)\dot{\gamma} = F_{\dot{\gamma}}J$, where we define the tidal force operator:

$$F_{\dot{\gamma}} : \dot{\gamma}^\perp \to \dot{\gamma}^\perp, \quad J \mapsto R(\dot{\gamma}, J)\dot{\gamma}$$

We also define $B_{\dot{\gamma}} : \dot{\gamma}^\perp \to \dot{\gamma}^\perp,\ J \mapsto \nabla_J \dot{\gamma}$, which is valid along γ. Note that, if we put $N = \dot{\gamma}$ for the normal vector to the space on which B is defined, then $B : X \mapsto \nabla_X N$. In a local frame B is simply $B^i_j = (\nabla_{\partial_j}\dot{\gamma})^i$. We then have $B_{\dot{\gamma}}J = \nabla_J\dot{\gamma} = \nabla_{\dot{\gamma}}J = \dot{J}$.

By simply substituting B into the the geodesic deviation equation we get

$$F_{\dot{\gamma}}J = (B_{\dot{\gamma}}J)' = \dot{B}_{\dot{\gamma}}J + B_{\dot{\gamma}}\dot{J} = \dot{B}_{\dot{\gamma}}J + B^2_{\dot{\gamma}}J.$$

From this it follows that (on $\dot{\gamma}^\perp$):

$$\dot{B}_{\dot{\gamma}} + B_{\dot{\gamma}}^2 = F_{\dot{\gamma}}.$$

Taking the trace we get the **Raychadhuri equation**.

To get an explicit expression, split the tensor $B = B_{\dot{\gamma}}$ according to symmetries (i.e. irreducible representations) of $SO(3)$ - that is to say, the symmetries of the positive definite metric on $\dot{\gamma}^\perp$:

$$B = \underbrace{\frac{1}{2}(B - B^T)}_{\omega} + \underbrace{\left[\frac{1}{2}(B + B^T) - \frac{1}{3}(\operatorname{tr} B)I\right]}_{\sigma} + \underbrace{\frac{1}{3}(\operatorname{tr} B)I}_{\frac{1}{3}\theta I} = \omega + \sigma + \frac{1}{3}\theta I.$$

Here ω is the antisymmetric part of B, σ is the trace-free symmetric part, and θ is the trace.

Taking the trace, it is not too difficult to show that:

$$\dot{\theta} = \operatorname{tr} \omega\omega^T - \operatorname{tr} \sigma\sigma^T - \frac{1}{3}\theta^2 - \operatorname{Ric}(\dot{\gamma}, \dot{\gamma}).$$

As we generally have $\operatorname{tr} XX^T = X_{ij}X^{ij} \geq 0$, in index notation the formula becomes:

$$\dot{\theta} = \omega_{ij}\omega^{ij} - \sigma_{ij}\sigma^{ij} - \frac{1}{3}\theta^2 - R_{ij}\dot{\gamma}^i\dot{\gamma}^j. \tag{4.1}$$

We now give an intuitive interpretation of B and its trace θ. Recall that for a submanifold $S \subset M$ we can decompose the connection as $\nabla_X Y = (\nabla_X Y)^\parallel + (\nabla_X Y)^\perp$. The second fundamental form is then defined by $II(X, Y) = (\nabla_X Y)^\perp$ for any X and Y tangent to S. If S is a hypersurface with the (unique up to orientation) unit normal vector field N, then we can define a symmetric 2-tensor containing precisely all the information of the 2. fundamental form $K(X, Y) = -\langle II(X, Y)|N\rangle$. For a hypersurface, we therefore do not distinguish between the two and simply call K the second fundamental form (or extrinsic curvature). Let us note that most mathematicians take K to have the opposite sign (e.g. O'Neill [1]); for some reason our convention seems to be prevalent in the physics community (Wald [4], Hawking & Ellis [3], Sean Carroll [19], Baez and Muniain [20]). The associated linear operator L of K is defined by $\langle LX|Y\rangle = K(X, Y)$ (raise one index) and called the shape operator. In fact, one can show the Weingarten formula $L(X) = \nabla_X N$.

Let S be a spacelike hypersurface in M orthogonal to some congruence (i.e. family) of geodesics. Let K be its extrinsic curvature (the second fundamental form). We then have $K_{\mu\nu} = \nabla_\mu N_\nu$, where N is the normal unit vector field on S (N is therefore tangent to the congruence of geodesics). Now it is clear that $K_{\mu\nu}$ is simply the tensor B for this congruence, so its trace is θ. Note that K must be symmetric (i.e. $\omega = 0$) as the second fundamental form is symmetric (an easy consequence of zero torsion). One could also argue that, since the congruence is orthogonal to some surface, it must be irrotational by the Frobenius integrability condition.

Note that K can be written as a Lie derivative. Indeed as already mentioned, for a Levi-Civita connection we have:

$$\mathcal{L}_N g(X, Y) = g(\nabla_X N, Y) + g(\nabla_Y N, X) = K(X, Y) + K(Y, X) = 2K(X, Y),$$

so we see that:

$$K = \frac{1}{2}\mathcal{L}_N g.$$

Now, if h is the metric on S, then h and g agree on all vectors tangent to S. Also, K is only defined on vectors tangent to S so we have $K = \frac{1}{2}\mathcal{L}_N h$.

We now simplify this expression even further by a convenient choice of coordinates

Definition 54. Let $S \subset M$ be a spacelike hypersurface and denote by $N_pS = T_pS^\perp$ the orthogonal complement of T_pS in M. Let $\exp : \mathcal{D} \subset TM \to M$ be the exponential map of M (mapping each $v \in \mathcal{D}$ to $c_v(1)$, where c_v is the geodesic with initial velocity v). We may now restrict exp to the normal bundle $NS = \bigcup_{p\in S} N_pS$ of S to get the **normal exponential map** $\exp : U \subset NS \to M$ ($U = \mathcal{D} \cap NS$) which traces out (timelike) geodesics normal to S.

Definition 55 (Conjugate points)**.** Let $v \in U$ be in the domain of the normal exponential map exp and let $q = \exp(v) \in M$. We say q is **conjugate** to S (or a **focal point** of S) if v is a critical point of exp, i.e. exp is not of full rank at v.

Intuitively, the exponential maps a family of vectors orthogonal to S to a congruence of geodesics covering some open neighborhood of M. When we go too far out, however, the geodesics can converge and exp will then no longer have full rank (the image will become lower dimensional). One can show that q will be conjugate to S iff one can find a nonzero Jacobi field along a geodesic connecting S to q, vanishing at both q and S.

Let S be a spacelike hypersurface and $\varphi = (x^1, x^2, x^3)$ be coordinates in some neighborhood $U \subset S$. Now take a unit normal vector field N on that neighborhood (providing U with orientation in the process) - generally one would take a basis to N_pS at each point $p \in U$. This gives us coordinates on NS, defined by $(t, x^1, x^2, x^3) \mapsto (x, tN_x)$.

Assuming that ($p \in S$) $q = \exp_p(t_0 N_p)$ is not conjugate to S, we may then construct the so-called **synchronized coordinates** in some neighborhood V of q in the following manner. First, as q is not conjugate, exp will be a diffeomorphism on $V \simeq \tilde{V} \subset NS$ if we take V to be small enough. Now it is clear that one can transfer the coordinate system on NS to M via the exp map.

Intuitively, each $r \in V$ lies on some geodesic $r = \exp_x(tN_x)$ and the coordinates of r are (t, x^1, x^2, x^3), where (x^1, x^2, x^3) are coordinates of x on S.

We should note that the metric splits in these coordinates. First, on S we certainly have $g_{0i} = 0$ as in these coordinates $\partial/\partial t$ is normal to S. On the other hand:

$$\partial_t g_{0i} = \partial_t \langle \partial_t | \partial_i \rangle = \langle \partial_t | \nabla_t \partial_i \rangle = \langle \partial_t | \nabla_i \partial_t \rangle = \frac{1}{2}\partial_i \langle \partial_t | \partial_t \rangle = 0,$$

where we have used the fact that torsion vanishes as well that coordinate vector fields commute. We conclude that surfaces of constant t remain orthogonal to the geodesics and the metric must be of the form $g = -dt \otimes dt + h$.

In synchronized coordinates B is the extrinsic curvature for each surface of constant t (in particular for $t = 0$ it is the extrinsic curvature K of S). Now the formula for B simply becomes $B = \frac{1}{2}\mathcal{L}_{\partial_t} h = \frac{1}{2}\partial_t h$.

Finally, it is clear that

$$\theta = \operatorname{tr} B = \frac{1}{2}h^{ij}\partial_t h_{ij} = \frac{1}{2}\operatorname{tr}((h_{ij})^{-1}\partial_t(h_{ij})) = \frac{1}{2}\partial_t \log(\det h_{ij}) = \partial_t \log(\sqrt{\det h_{ij}}),$$

where we have used the fact that for any curve $A : \mathbb{R} \to GL(n)$ the following holds: $\mathrm{tr}(A^{-1}A') = (\log(\det A))'$ [1]. This shows that the expansion gives the variation of the 3-volume element as measured by synchronized observers.

In particular, when the determinant $\det h_{ij}$ vanishes (i.e. the expansion θ blows up), the synchronous coordinate system breaks down (i.e. $\partial_t, \partial_1, \partial_2, \partial_3$ fail to be linearly independent) in which case we must have a conjugate point.

4.2 Time Separation Function

For a causal curve $\gamma : [a, b] \to M$, denote by $l(\gamma) = \int_b^a |\gamma'(t)|dt$ its length. Define the **time separation** of two points $p, q \in M$ by

$$\tau(p, q) = \sup\{l(\gamma) \mid \gamma \text{ is a future-pointing causal curve from } p \text{ to } q\}.$$

Here if $q \notin J^+(p)$ (i.e. $p \not\leq q$), we set $\tau(p, q) = 0$ [2]. Then one has the following result:

Proposition 56
Assume that $p < q$ and that the set $J(p, q) = J^+(p) \cap J^-(q)$ is compact. Assume further that strong causality holds on $J(p, q)$, then there is a causal geodesic of length $\tau(p, q)$ connecting p and q.

Proof. By definition of supremum, we can find future-directed causal curves γ_n from p to q, whose lengths converge to $\tau(p, q)$. These curves are all in $J(p, q)$, so by 29 we can find a broken geodesic α from p to q with length $l(\alpha) \geq \lim_{n\to\infty} l(\gamma_n) = \tau(p, q)$. Since $\tau(p, q)$ is supremum, we actually have $l(\alpha) = \tau(p, q)$. Therefore, $\tau(p, q)$ is finite.

We now show that the breaks on the geodesic must all be trivial. Indeed, first note that on each smooth segment we have consistently either a timelike or a null geodesic. A break cannot go from timelike to null for it would shorten the length of the total curve (this certainly holds on convex neighborhoods). Therefore, either p and q cannot be connected by a timelike curve or p and q can be connected by a timelike broken geodesic. In the former case we have seen that the only curve connecting p and q is a null geodesic, and in the latter case the first variation formula (see the next section) gives the desired result. □

Lemma 57
$\tau : M \times M \to [0, \infty]$ *is lower semi-continuous.*

Proof. For $q \notin I^+(p)$, we have $\tau(p, q) = 0$, and there is nothing to prove. We therefore assume $\tau(p, q) > 0$ and $q \in I^+(p)$. Given a $\varepsilon > 0$ we must find open neighborhoods U of p and V of q such that for all $p' \in U$ and $q' \in V$ we have $\tau(p', q') > \tau(p, q) - \varepsilon$. Take a timelike curve γ from p to q with $l(\gamma) > \tau(p, q) - \varepsilon$. Let C be a convex neighborhood of q and take x to be a point on γ slightly to the past of q in C. Since the length of geodesic segments depends continuously on its endpoints, we can find a neighborhood V of q such that, for any $q' \in V$, the segment $x \to q'$ is causal (we can take V so small as to satisfy

[1] Indeed, the derivative of determinant at X in direction Y is $(d\det)_X(Y) = \det X \,\mathrm{tr}(X^{-1}Y))$. Therefore, a curve going through A with velocity A' will have $\frac{d}{dt}\det A(t) = \det A \,\mathrm{tr}(A^{-1}A')$, i.e. $(\log(\det A))' = \frac{1}{\det A}(\det A)' = \mathrm{tr}(A^{-1}A')$.

[2] One should note that, precisely because we incorporate time orientation, τ will be symmetric only in trivial instances.

$V \subset I^+(x)$) and strictly longer than $l(x \to q) - \varepsilon$. Since the segment $x \to q$ is a geodesic, it must be at least as long as the segment of curve γ going from x to q.

We now do the same thing at the endpoint p, which gives a neighborhood U such that any $p' \in U$ and $q' \in V$ can be joined by a causal curve of length $l > l(\gamma) - 2\varepsilon > \tau(p,q) - 3\varepsilon$. Since ε is arbitrary, this gives the desired result.

If by any chance we have $\tau(p,q) = \infty$, the same argument actually shows that for any $M > 0$, we get neighborhoods U and V such that $\tau(p',q') > M$ for any $p' \in U$ and $q' \in V$. □

In fact, when M is globally hyperbolic, we get a stronger result:

Lemma 58
τ is continuous on $M \times M$ for a globally hyperbolic spacetime M.

Proof. The previous result shows lower semi-continuity and we also know $\tau(p,q)$ is finite in globally hyperbolic spaces. We therefore only need to show upper semi-continuity, i.e. for any $(p,q) \in M \times M$ and $\varepsilon > 0$ there exist an open neighborhoods $U \times V$ such that for any $(p',q') \in U \times V$ we have $\tau(p',q') < \tau(p,q) + \varepsilon$. We assume to the contrary that τ is not upper semi-continuous at some $(p,q) \in M \times M$. We can therefore find $\varepsilon > 0$ and sequences $p_n \to p$ and $q_n \to q$, for which $\tau(p_n,q_n) \geq \tau(p,q) + \varepsilon$ for all n.

As $\tau(p_n,q_n) > 0$, there exists a causal curve γ_n from p_n to q_n satisfying $l(\gamma_n) > \tau(p_n,q_n) - 1/n$. Take some point p^- in the chronological past of p and q^+ in the chronological future of q. By looking only at sufficiently large n, we may assume p_n and q_n are contained in $I^+(p^-)$ and $I^-(q^+)$ respectively. Now γ_n are all in $J(p^-,q^+)$. As $J(p^-,q^+)$ is compact, we apply lemma 29 to get a causal curve α from p to q, which satisfies $l(\alpha) \geq \tau(p,q) + \varepsilon$. This is a contradiction as τ is a supremum. □

Proposition 59
Let M be globally hyperbolic and S a Cauchy surface and $q \in M$. Then $J^+(S) \cap J^-(q)$ is compact.

In particular, $S \cap J^-(q)$ is compact.

Proof. First define $C(S,q)$ to be all continuous future-directed causal curves from S to q. Then, analogously to proposition 50, one can prove $C(S,q)$ is compact. Again, a proof analogous to one used to prove proposition 51 gives the result. □

For subsets $A, B \subset M$ define $\tau(A,B) = \sup\{\tau(a,b) \mid a \in A,\ b \in B\}$. Then we have:

Proposition 60
Let $S \subset M$ be a Cauchy surface. If $q \in D^+(S)$, then there exists a geodesic from S to q of length $\tau(S,q)$. This geodesic must be timelike (except in the trivial case $q \in S$).

Proof. As $J^-(q) \cap S$ is compact and τ continuous on $J^-(q) \cap S$, it must achieve a maximum at some $p \in J^-(q) \cap S$. Obviously we have $\tau(p,q) = \tau(S,q)$, but now proposition 56 gives a geodesic γ from p to q of length $\tau(p,q)$. Let now $q \notin S$, then there certainly exists a timelike curve from S to q, so $\tau(p,q) > 0$ and $q \in I^+(p)$, therefore γ is timelike. □

4.3 Hawking's Singularity Theorem

To prove Hawking's singularity theorem, we must work with a Cauchy surface that is also a smooth manifold (so that second fundamental form and various other quantities are well defined). As we have previously mentioned, a globally hyperbolic space will always have a smooth Cauchy surface, thus we actually require no additional assumptions, other than M being globally hyperbolic. We also must assume a certain condition on the energy tensor:

Definition 61 (Strong energy condition)**.** We say M satisfies the **strong energy condition** if $\mathrm{Ric}(X, X) \geq 0$ for all timelike vector fields $X \in \mathfrak{X}(M)$.

Intuitively this just means that "gravity attracts" as the volume of a free-falling ball of particles will shrink when the condition is satisfied. We can now show:

Lemma 62
Let M be a globally hyperbolic spacetime satisfying the strong energy condition, $S \subset M$ a (smooth) Cauchy hypersurface, and $p \in S$ a point where $\theta = \theta_0 < 0$. Then the geodesic γ through p contains a point conjugate to S, at a distance of at most $-3/\theta_0$ to the future of S (assuming that it can be extended that far).

The lemma simply states that if the surface S is curved in such a way to focus the geodesics at $p \in S$, then the geodesics will eventually meet above p.

Proof. Since the antisymmetric part ω vanishes as the congruence is irrotational, the Raychaudhuri equation becomes:

$$\dot{\theta} + \frac{1}{3}\theta^2 = -\sigma_{ij}\sigma^{ij} - R_{ij}\dot{\gamma}^i\dot{\gamma}^j \leq 0,$$

where by the strong energy condition $R_{ij}\dot{\gamma}^i\dot{\gamma}^j \geq 0$. Integrating the inequality $\dot{\theta} + \frac{1}{3}\theta^2 \leq 0$ (i.e. $\frac{d}{dt}\theta^{-1} \geq \frac{1}{3}$) we get:

$$\frac{1}{\theta} \geq \frac{1}{\theta_0} + \frac{t}{3}.$$

It is clear that θ now must blow up at t no greater than $-3/\theta_0$. □

Let us recall the first and second variation formulae. Let γ_s be a (regular, i.e. non null) family of geodesics - a variation of $\gamma = \gamma_0$. Suppose γ is parametrized by arclength and denote by $\epsilon = \langle\gamma'|\gamma'\rangle = \pm 1$ the sign of γ. Now let $L(s)$ the length of γ_s so $L : \mathbb{R} \to \mathbb{R}$. Finally, we let $V(t) = \frac{d}{ds}\big|_{s=0}\gamma_s(t)$ and $A(t) = \frac{d^2}{ds^2}\big|_{s=0}\gamma_s(t)$ denote the transverse velocity and acceleration and V' and A' their $(\partial/\partial t)$ derivatives.

We then have the first variation formula:

$$L'(0) = -\epsilon \int_b^a \langle\gamma''|V\rangle dt + \epsilon\langle\gamma'|V\rangle\Big|_a^b.$$

If γ is a geodesic ($\gamma'' = 0$), then $L'(0) = \epsilon\langle\gamma'|V\rangle\Big|_a^b$.

For a piecewise variation we let $t_1 < \cdots < t_k$ be the times at which V is not differentiable, and write $\Delta V = V(t_i^+) - V(t_i^-)$ for the jump at t_i. Applying the first variation formula to each smooth segment and summing, gives:

$$L'(0) = -\epsilon \int_b^a \langle \gamma'' | V \rangle dt - \epsilon \sum_i \langle \Delta\gamma(t_i) | V(t_i) \rangle + \epsilon \langle \gamma' | V \rangle \Big|_a^b . \tag{4.2}$$

So a curve which is stationary for piecewise smooth variations with fixed endpoints ($V(a) = V(b) = 0$) must in fact be smooth (i.e. the jumps are trivial).

We also have Synge's formula:

$$L''(0) = \epsilon \int_a^b \left(\langle (V^\perp)' | (V^\perp)' \rangle + \langle R(V, \gamma')V | \gamma' \rangle \right) dt + \epsilon \langle \gamma' | A \rangle \Big|_a^b .$$

Fixing the end points, one gets:

$$L''(0) = \epsilon \int_a^b \left(\langle (V^\perp)' | (V^\perp)' \rangle + \langle R(V, \gamma')V | \gamma' \rangle \right) dt.$$

This can be regarded as a quadratic form on the space of all piecewise differentiable vector fields along γ. The corresponding bilinear form is then the so-called Morse index (or index form):

$$I_\gamma(V, W) = \varepsilon \int_a^b \left(\langle (V^\perp)' | (W^\perp)' \rangle + \langle R(V, \gamma')W | \gamma' \rangle \right) dt.$$

Of course, symmetries of the Riemann tensor guarantee that I_γ is symmetric. If $\Delta V = V(t_i^+) - V(t_i^-)$ is a jump at t_i, then integration by parts then gives:

$$I_\gamma(V, W) = -\epsilon \int_a^b \langle (V^\perp)'' + R(V^\perp, \gamma')\gamma' | W^\perp \rangle dt - \epsilon \sum_i \langle \Delta(V^\perp)' | W^\perp \rangle (t_i). \tag{4.3}$$

Lemma 63
Let M be globally hyperbolic and S a Cauchy surface. Let $\gamma : [a, b] \to M$ be a geodesic from S to $q \in M$ parametrized by arclength (so in particular $\gamma' \neq 0$). If γ maximizes the length $L'(0) = 0$, then γ is orthogonal to S.

Proof. We are trying to find which endpoint $\gamma(a) \in S$ extremizes the length of the geodesic, provided we fix the point $q = \gamma(b)$. Therefore, we take the variation field $V(t)$ to be tangent to S at $\gamma(a)$ and zero at $\gamma(b)$. Then the conclusion is obvious from the first variation formula, when γ is a geodesic: $0 = L'(0) = \epsilon \langle \gamma' | V \rangle \Big|_a^b = \langle \gamma'(a) | V(a) \rangle$. □

It is a well known fact that geodesics do not minimize (or in the case of timelike curves maximize) length past conjugate points. In our case this means:

Lemma 64
Let M be a globally hyperbolic spacetime, S a Cauchy hypersurface, $q \in M$ and γ a timelike geodesic through q orthogonal to S. If there exists a conjugate point between S and q, then γ does not maximize length (among the timelike curves connecting S to q).

We shall give two proofs of this fact.

Proof 1. Certainly, this can be checked by calculating the index form. We follow Sternberg [14] (chapter 8).

It is sufficient to find a vector field X normal to γ that has $I_\gamma(X,X) > 0$, for then γ cannot be a local maximum by the second variation formula.

We are given a geodesic $\gamma : [a,b] \to M$ starting orthogonally at S and ending at q. Let $r = \gamma(t_0)$ be the first conjugate point between S and q along γ. We can thus find a (nonzero) Jacobi field J along the restriction of γ to $[a,t_0]$ vanishing at both $t = a$ and $t = t_0$. We now extend J to a field

$$Y = \begin{cases} J(t), & t \in [a,t_0] \\ 0, & t \in [t_0,b]. \end{cases}$$

At t_0 field Y must have a jump $\Delta Y' \neq 0$ since, if the jump were to be 0, we would have $(\nabla_{\dot{c}} J)(b) = 0$ and $J(b) = 0$ so J would have to be 0.

Plugging Y into the formula for I_γ, we see that the integral vanishes (as Y satisfies the Jacobi equation and is orthogonal to γ). So with $\epsilon = 1$ (γ is timelike) we get $I_\gamma(Y,Z) = -\langle \Delta Y' | Z^\perp \rangle(t_0)$ for any vector field Z along γ. In particular $I_\gamma(Y,Y) = 0$ because $Y(t_0) = 0$. Now take W to be any vector field along γ with $W(t_0) = \Delta Y'(t_0) \neq 0$, then $I(Y,W) < 0$. Now it is clear (from bilinearity and symmetry of I_γ) that for sufficiently small $\delta > 0$ we get:

$$I_\gamma(Y + \delta W, Y + \delta W) = -2\delta I_\gamma(Y,W) - \delta^2 I_\gamma(W,W) < 0,$$

which was to be proven. □

One doesn't have to rely on the second variation formula to prove this result. Indeed, for null geodesics, we cannot apply the variation formulae, so it will be beneficial to see how this is done. Roughly the argument goes as follows:

Proof 2 (heuristic). Let r be the first conjugate point on γ between $p \in S$ and q. We can then use synchronized coordinates around γ in the portion between p and r. Since r is conjugate to S (r is a critical point of exp), we can find another (distinct) geodesic α orthogonal to S with same length as γ (up to r), which approximately (up to first order as far $d\exp$ is concerned) intersects γ at r.

Now, if γ and α really intersected at r (exactly), we could get a broken curve by following α from S to r and then γ from r to q. Since a broken geodesic does not extremize length (twin paradox), this would actually give the result.

However, γ and α do not intersect exactly, so some care is needed. Take some small convex neighborhood C around r and let $x' \in C$ be a point along α between in the chronological past of r (since α comes close to γ this is possible) and w be a point along γ between r and q (thus w is in the chronological future of r). By following α from S to x' and then the unique timelike geodesic from x' to w and then γ from w ro q, we get a curve with strictly greater length than γ. To make this rigorous one can refer to Penrose [8] theorem 7.27. □

Finally, we can now with relative ease show the following:

Theorem 65 (Hawking)
Let M be a globally hyperbolic spacetime satisfying the strong energy condition, and suppose that the expansion satisfies $\theta \leq \theta_0 < 0$ on a Cauchy hypersurface S, then M is singular.

Proof. It is sufficient to show that no future-directed geodesic orthogonal to S can be extended to proper time greater than $\tau_0 = -3/\theta_0$ (to the future of S). Assuming to the contrary that this is not so, we could find a future-directed timelike geodesic α orthogonal to S and parameterized by proper time defined on some interval $[0, \tau_0 + \varepsilon]$ for $\varepsilon > 0$. So let $q = \alpha(\tau_0 + \varepsilon)$, then according to proposition 60 we could find a timelike geodesic γ of maximal length connecting S and q, which by lemma 63 must be orthogonal to S. Thus the proper time of γ necessarily exceeds $\tau_0 + \varepsilon$. Now lemmata 62 and 64 guarantee that γ would develop a conjugate point at a distance of at most τ_0 and that it would then cease to be maximizing beyond this point. This is clearly absurd and the statement follows. □

4.4 Null Hypersurface

To deal with null hypersurfaces, it will be beneficial to first explain null hyperplanes in Lorentzian vector spaces.

Let V be a vector space with a nondegenerate scalar product g. If $W \subset V$ is a vector subspace, we set $W^\perp = \{v \in V \mid g(v, w) = 0 \ (\forall w \in W)\}$. One then shows $\dim W + \dim W^\perp = \dim V = n$. This, however, does *not* mean $V = W \oplus W^\perp$, as W can, for instance, contain $W^\perp$. $V = W \oplus W^\perp$ will hold precisely when g is nondegenerate on W. We say W is **null** if g is degenerate on W. We can also show that, generally, $(W^\perp)^\perp = W$.

If W is a hyperplane, so that $\dim W = \dim V - 1$, then $\dim W^\perp = 1$. We can thus find a vector $N \in V$ that generates $W^\perp$, i.e. $W^\perp = \mathbb{R}N$.

Note that a hyperplane W is null iff N is a null vector, i.e. $g(N, N) = 0$. Indeed, if N is spacelike or timelike, we can use Gram-Schmidt procedure to build an orthogonal basis $N, v_2, ...v_n$ for V. This means that $g(v_i, v_i) \neq 0$ and since v_i span W, we get that W is nondegenerate.

Note also that the normal vector N is null precisely when $W^\perp \subset W$. Indeed, N is orthogonal to W by definition, so if $W^\perp \subset W$, then N must be orthogonal to $W^\perp$ as well, meaning $g(N, N) = 0$ (as $N \in W^\perp$). Conversely, if $g(N, N) = 0$, then in particular $g(N, cN)$ for any $c \in \mathbb{R}$, so N is perpendicular to $W^\perp$, i.e. $N \in (W^\perp)^\perp = W$. This implies $cN \in W$ for all $c \in \mathbb{R}$, i.e. $W^\perp \subset W$.

Since no timelike vector is orthogonal to a (nonzero) null vector, we see that W contains only null and spacelike vectors. Indeed, assume t is timelike $g(t, t) < 0$ (in particular $t \neq 0$), u is null $g(u, u) = 0$ and $g(u, t) = 0$; then a simple calculation shows that $u = 0$. Take an orthonormal basis with a multiple of t as its timelike member (so that $t = (t^0, 0)$). $g(t, u) = 0$ then gives $t^0 u^0 = 0 \Longrightarrow u^0 = 0$ but now $0 = g(u, u) = -(u^0)^2 + |\overrightarrow{u}|^2 = |\overrightarrow{u}|^2$. Since $|\overrightarrow{u}|$ is the euclidean norm, this gives $\overrightarrow{u} = 0$, i.e. $u = (u^0, \overrightarrow{u}) = 0$.

We say a submanifold $S \subset M$ is a null hypersurface if $T_pS \subset T_pM$ are null hyperplanes for all $p \in S$. Such hypersurfaces have the peculiar property that their normal vector fields N are also tangent. At least locally, we can find a nonvanishing future-pointing null smooth vector field N. Now N generates $T_pS^\perp$ (at each $p \in S$ on which N is defined).

In fact, in a time orientable Lorentzian spacetime, the null field N on S is globally defined. To see this, first note that we have a globally defined timelike vector field X; restricting X to S we get a globally defined (on S) transverse vector field.

Thus $b = X^\flat \otimes X^\flat + i^*g$ is a Riemann metric on S (where i^*g is the pullback of g on S via the inclusion). To see b is positive definite, note that $g(Y,Y) > 0$ on spacelike Y and $(X^\flat(Y))^2 > 0$ on null Y. We can now take the dual vector field N in the metric b of the 1-form $X^\flat$ on S and normalize it (in b norm). In other words, N is the unique vector field in TS satisfying $b(N,\cdot) = X^\flat$ and $b(N,N) = X^\flat(N) = 1$. This N field is a null vector field tangent to S. Indeed, computing the b norm gives us $1 = b(N,N) = (X^\flat(N))^2 + g(N,N) = 1 + g(N,N)$, so $g(N,N) = 0$. Since X was nowhere zero and globally defined, so must N be as well.

Even though it is very simple and elegant in hindsight, the idea of taking duals twice (in different metrics) is (at least to the author) entirely nonobvious. I have come across this in [53].

Lemma 66
Integral curves of N are null geodesics (up to parametrization).

Before the proof note that, for a smooth function f, we have $\nabla_\mu\nabla_\nu f = \nabla_\nu\nabla_\mu f$. As is often the case, the index notation here is a bit more subtle than one would like. First, we have $(\nabla f)_\nu = g(\nabla f, \partial_\nu) = df(\partial_\nu) = \partial_\nu f = \nabla_\nu f$. Next, as usual, $\nabla_\mu X_\nu$ means $(\nabla_\mu X)_\nu = g(\nabla_\mu X, \partial_\nu)$, thus $\nabla_\mu\nabla_\nu f = \nabla_\mu(\nabla f)_\nu = (\nabla_\mu(\nabla f))_\nu$. The statement now becomes $(\nabla_\mu(\nabla f))_\nu = (\nabla_\nu(\nabla f))_\mu$, which is not entirely obvious, but holds for any Levi-Civita connection and is not difficult to prove[3].

Proof. Since $g(N,N) = 0$, we have $Xg(N,N)$ for all vector fields X tangent to S. This means that $d(g(N,N))$ annihilates $TS \subset TM$. Thus $\nabla(g(N,N))$ must be normal to S, because $g(\nabla(g(N,N)), X) = d(g(N,N))(X) = 0$ for all X tangent to S. In other words, $\nabla(g(N,N))$ is proportional to N. Let f be a function that locally defines S, i.e. $S = \{f = 0\}$, then df annihilates S, therefore ∇f must be proportional to N.

It is sufficient to prove the theorem for $N = \nabla f$ (as any normal field is proportional to any other normal field). Now we get

$$N \sim \nabla_\mu(g(N,N)) = 2g(\nabla_\mu N, N) = 2N^\nu\nabla_\mu N_\nu = 2N^\nu\nabla_\mu\nabla_\nu f = 2N^\nu\nabla_\nu\nabla_\mu f = 2\nabla_N N_\mu$$

Thus, parallel transporting N along its integral curve again gives some vector that is tangent to the curve. In other words, the integral curve is a geodesic (up to reparametrization).

□

As a corollary, we have that every point $p \in S$ lies on some (unique) future inextendible null geodesic, which in turn lies on S. S is thus foliated by inextendible null geodesics, which are called its **null generators**.

We may thus generalize the notion of a null hypersurface from smooth to topological hypersurfaces, by requiring they be foliated by null geodesics in this manner[4]. This

[3]Indeed, since ∇ is compatible with the metric, we get $Xg(\nabla f, Y) = g(\nabla_X\nabla f, Y) + g(\nabla f, \nabla_X Y)$, and since torsion vanishes $\nabla_X Y - \nabla_Y X = [X,Y]$. Thus, $g(\nabla_X\nabla f, Y) - g(\nabla_Y\nabla f, X) = Xg(\nabla f, Y) - g(\nabla f, \nabla_X Y) - Yg(\nabla f, X) + g(\nabla f, \nabla_Y X) = Xdf(Y) - Ydf(X) - g(\nabla f, \nabla_X Y - \nabla_Y X) = XYf - YXf - g(\nabla f, [X,Y]) = (XY - YX)f - df([X,Y]) = 0$. The special case $X = \partial_\mu$ and $Y = \partial_\nu$ now proves the statement.

[4]Though, a smooth hypersurface S foliated by null geodesics is not necessarily a null hypersurface as previously defined - take for instance a timelike plane in Minkowski space. However, the result does hold for all achronal planes in Minkowski space. Thus, if we require S to be locally achronal, it will hold on S as well.

weakening is necessary when discussing black hole event horizons in full generality (as they need not be smooth).

The following shows that for a closed subset C, the boundary of $J^+(C)$, which is a topological hypersurface, comes quite close to being a null hypersurface:

Proposition 67
Let C be a closed subset of the spacetime M. Then every point $p \in \partial J^+(C)$ (except perhaps $p \in C$) lies on some null geodesic γ, which in turn lies entirely in $\partial J^+(C)$. γ is either inextendible or has a past endpoint in C

Of course, an analogous statement holds for $J^-(C)$ by flipping time orientation.

Proof. Let $p \in \partial J^+(C) = \partial I^+(C)$ and choose a sequence $p_n \in I^+(C)$ which converges to p. For each p_n we find a past directed timelike curve γ_n connecting p_n and some point in C. Since, C is closed, $M \setminus C$ is an open subset of M and therefore a manifold in its own right. On $M \setminus C$ each γ becomes past inextendible, so we may use lemma 31 to obtain a limit curve γ, which passes through p and is past inextendible (in $M \setminus C$). γ being inextendible in $M \setminus C$ means that in M, γ either remains past inextendible, or has a past endpoint in C.

As each point on γ is a limit sequence of points on γ_n (and γ_n lie in $I^+(C)$), γ must lie in $\overline{I^+(C)}$. If any point on γ by any chance landed in $I^+(C)$, we could make the γ timelike by changing it slightly (lemma 25), so we would have $p \in I^+(C)$. We thus see that γ is completely contained in $\partial J^+(C)$. Finally, since $\partial J^+(C)$ is achronal (no two points can be connected by a timelike curve), any curve connecting two points on $\partial J^+(C)$ must be a null geodesic. □

We now turn to analyzing the structure of the tangent space of a smooth null hypersurface S. Since on T_pS the metric g is degenerate, consider the quotient T_pS/N. In T_pS/N we identify $X \in T_pS$ and $X + cN \in T_pS$ for any $c \in \mathbb{R}$. If $[X]$ denotes the equivalence class of X, then can define a metric on T_pS/N by $h([X],[Y]) = g(X,Y)$. This is indeed well defined, as $g(X + aN, Y + bN) = g(X,Y) + ag(N,Y) + bg(X,N) + abg(N,N) = g(X,Y)$.

Note that h must be positive definite on the quotient space T_pS/N, because vectors which live in T_pS, but not in $\mathbb{R}N$ must be spacelike.

Finally, we can introduce the Weingarten map as before and prove the Raychadhuri's equation. We define the null Weingarten map $L_p : T_pS/N \to T_pS/N$ by $L_p([X]) = [\nabla_X N]$. This is well defined as $\nabla_{X+aN} N = \nabla_X N + a\nabla_N N$. But we have seen that $\nabla_N N$ is proportional to N, thus $\nabla_{X+aN} N$ indeed lies in the equivalence class $[\nabla_X N]$.

On the other hand, if we take some other null vector field fN, then $\nabla_X(fN) = f\nabla_X N + (Xf)N$, which is in the equivalence class of $f\nabla_X N$. Thus the L depends on the choice of field N (but only up to scalar multiples).

The second fundamental form is now $K([X],[Y]) = h(L([X]),[Y]) = g(\nabla_X N, Y)$ and the expansion is $\theta = \operatorname{tr} L$.

We must further define the covariant derivative of a field Y along some null curve. This is defined in the obvious way $\nabla_N[Y] = [\nabla_N Y]$, and the usual computation: $\nabla_N(Y + aN) = \nabla_N Y + \nabla_N(aN) = \nabla_N Y + a\nabla_N(N) + (Na)N$ shows that the definition is valid.

Consider now a null curve $\alpha : [a,b] \to S$ and a vector field $V : [a,b] \mapsto TS/N$ along α; we assume V is smooth in the sense $V(s) = [X(s)]$ for some smooth X along α. We then have $\nabla_{\dot\alpha(s)} V = [\nabla_{\dot\alpha(s)} X]$. More briefly, $[X]' = [X']$

Let now L be a Weingarten map for the null vectors α', then its derivative along α is defined in the usual way: $(L')([X]) = (L([X]))' - L([X']) = [\nabla_X N]' - L([X'])$, where X is along α and tangent to S.

We also pull back the tidal operator: $F : T_{\alpha(s)}S/N \to T_{\alpha(s)}S/N$, via $F([X]) = [F(X)] = [R(\alpha'(s), X)\alpha'(s)]$.

Raychadhuri's equation $L' + L^2 = F$ can now be proven with ease (the argument being basically the same as before). Indeed, by scaling N, we may assume that α is a geodesic, i.e. $\nabla_N N = 0$. We work in a neighborhood around $p = \alpha(s)$, and may assume N is extended to a vector field on a neighborhood near p. Furthermore, assume $X \in T_pS$ comes from a vector field commuting with N, so that $\nabla_N X = \nabla_X N$. Then $R(X, N)N = \nabla_X \nabla_N N - \nabla_N \nabla_X N - \nabla_{[X,N]} N = -\nabla_N \nabla_N X$, so we get the usual Jacobi equation $X'' = R(\alpha', X)\alpha'$. From here:

$$\begin{aligned} L'([X]) &= [\nabla_X N]' - L([\nabla_N X]) = [\nabla_N X]' - L([\nabla_X N]) \\ &= [X''] - L(L([X])) = [R(\alpha', X)\alpha'] - L^2([X]). \end{aligned}$$

Taking the trace, as before, we get:

$$\theta' = -\operatorname{Ric}(\alpha', \alpha') - \sigma^2 - \frac{1}{n-1}\theta^2, \tag{4.4}$$

where n is the dimension of S (so $n-1$ is the dimension of T_pS/N) and σ is the traceless symmetric part of L.

4.5 Penrose's Theorem

We now prove Penrose's version of the singularity theorem.

Definition 68 (Trapped surface)**.** Let M be a globally hyperbolic spacetime, S a Cauchy hypersurface with future-pointing unit normal vector field n and let $\Sigma \subset S$ be compact 2-dimensional with unit normal vector field ν in S (thus, for instance, $n \pm \nu$ are lightlike). Σ is said to be **trapped** if the expansions θ^+ and θ^- of the null geodesics with initial velocities $n + \nu$ and $n - \nu$ are both negative everywhere on Σ.

Intuitively, this means that lightlike geodesics converge on both sides of Σ. As an example, take a sphere inside the Schwarzschild event horizon and emit a light impulse on both sides. As light cannot escape from the black hole, both impulses will converge and fall towards the singularity (recall that inside the event horizon r is the timelike coordinate.). Let α_p be null geodesics through $p \in \Sigma$ with initial velocity $n + \nu$, then we may take $\exp : (-\varepsilon, \varepsilon) \times \Sigma \to M, \quad (p, t) \mapsto \alpha_p(t)$, which traces out a null hypersurface in M. Here, instead of synchronous coordinates, exp gives null coordinates.

Example 69. Take a sphere of fixed radius $r < 2GM$ in the Schwarzschild solution 2.1. Since r is timelike in the region $r < 2GM$, this defines a compact 2-dimensional spacelike surface. It is then not difficult to see that light geodesics emanating from such a sphere must evolve to a smaller r coordinate; thus the divergences are negative.

Definition 70 (Null energy condition)**.** We say that M satisfies the **null energy condition** if the Ricci tensor satisfies $\operatorname{Ric}(X, X) \geq$ for all lightlike X.

Lemma 71
Let M be a globally hyperbolic spacetime satisfying the null energy condition, $S \subset M$ a (smooth) Cauchy hypersurface with unit normal n, $\Sigma \subset S$ a compact 2-dimensional manifold with unit normal vector field ν in S. Let $p \in \Sigma$ be a point where $\theta = \theta_0 < 0$. Then the null geodesic α through p with initial velocity $n + \nu$ contains a point conjugate to S, at an affine parameter distance of at most $-2/\theta_0$ to the future of Σ (assuming that it can be extended that far).

Proof. In the case of 4-dim spacetime, we have $\theta' = -\operatorname{Ric}(\alpha', \alpha') - \sigma^2 - \frac{1}{2}\theta^2$. Thus

$$\theta' + \frac{1}{2}\theta^2 = -\operatorname{Ric}(\alpha', \alpha') - \sigma^2 \leq 0,$$

where $\operatorname{Ric}(\alpha', \alpha') \geq 0$ by the null energy condition. Integrating we get, as before,

$$\frac{1}{\theta} \geq \frac{1}{\theta_0} + \frac{t}{2},$$

so θ must blow up at t no greater that $-2/\theta_0$. □

Lemma 72
Let M be a globally hyperbolic spacetime satisfying the null energy condition, $S \subset M$ a (smooth) Cauchy hypersurface with unit normal n, $\Sigma \subset S$ a compact 2-dimensional manifold with unit normal vector field ν in S. Let γ_p be the null geodesic through p with initial velocity $n + \nu$ and $q = \gamma_p(t_0)$ for some t_0. If there exists a conjugate point between p and q, then $q \in I^+(\Sigma)$.

Proof. It is sufficient to show that the existence of a conjugate point guarantees that a non geodesic causal curve connects p and q. Then it is not true that $q \in J^+(p) \setminus I^+(p)$, as in that case only a null geodesic would connect p and q.

To see roughly why the above statement is correct, let r be the first conjugate point between p and q. Since $d\exp$ vanishes at r, up to first order, there is another (distinct) null geodesic α which intersects c at r. We now obtain a piecewise smooth curve (that is not a geodesic) by first following α from p to r and then γ from r to q. Again, some care is needed here as γ and α do not intersect exactly, but the argument is almost identical to the second proof of lemma 63 (where we use null coordinates along γ instead of synchronous coordinates). We thus omit the rigorous proof (it can be found in [8] as theorem 7.27 for instance). Another rigorous proof is given in O'Neill [1] as proposition 10.48. □

Finally, we arrive at:

Theorem 73 (Penrose)
Let M be a connected globally hyperbolic spacetime with a noncompact Cauchy hypersurface S, satisfying the null energy condition. If S contains a trapped surface Σ, then M is singular.

Proof. We roughly follow Godinho and Natário [7] and O'Neill [1].

- Assume M is (future) null complete. The key point then is that, for a compact trapped surface Σ, $\partial I^+(\Sigma)$ must be compact. Indeed, since θ^+ and θ^- are negative everywhere on Σ, there exists a maximum $\theta_0 < 0$ such that $\theta^+, \theta^- \leq \theta_0$. Consider a null geodesic going out from Σ (with initial velocity $n \pm \nu$), then lemma 71 guarantees

that a conjugate point can be found at affine parameter value not exceeding $u_0 = -2/\theta_0$ to the future of Σ. By lemma 72 this means that the null geodesic enters $I^+(\Sigma)$ after affine parameter value of no more than u_0. Consequently, $\partial I^+(\Sigma)$ is a closed subset of the compact set $\exp^+([0, u_0] \times \Sigma) \cup \exp^-([0, u_0] \times \Sigma)$, where $\exp^+$ refers to null geodesics with initial velocity $n + \nu$ and $\exp^-$ to null geodesics with initial velocity $n - \nu$. Note that we use future null completeness in this last step, as we need these null geodesics to go at least as far as u_0.

- Thus, assuming M is null complete, $\partial I^+(\Sigma)$ must be a compact topological hypersurface (by corollary 41). As in 45, let $\psi : M \to S$ be a retraction projecting M to S. We restrict to $\pi : \partial I^+(\Sigma) \to S$. Intuitively, we follow a timelike integral curve (of some nonvanishing vector field) from S until we hit $\partial I^+(\Sigma)$. Once we enter $I^+(\Sigma)$, we must stay within this set, so an integral curve cannot cross the boundary of $I^+(\Sigma)$ more than once.

  Note that π must be injective, because if $\pi(p) = \pi(q)$, then p and q lie on the same integral curve, which intersects $\partial I^+(\Sigma)$. Since the intersection is a unique point, we have $p = q$.

  Immediately from definition (and proposition 45) π is continuous. Since, π is an injective continuous mapping, between topological manifolds of the same dimension, Brouwer's invariance of domain guarantees π is actually a homeomorphism onto some open subset $\operatorname{Im}\pi$ of S. Thus, $\operatorname{Im}\pi$ is open and since $\partial I^+(\Sigma)$ is closed, $\operatorname{Im}\pi$ is closed as well. This means that $\operatorname{Im}\pi$ must be the whole of S, as S is connected.

  This is impossible, because $\operatorname{Im}\pi$ is actually a compact set ($\partial I^+(\Sigma)$ being compact) and S is not compact by assumption. We must conclude M is not null complete, i.e. it is singular on null geodesics.

□

A particular corollary is that the Schwarzschild spacetime will remain singular even if we slightly perturb its initial conditions. One should not forget that global hyperbolicity is in the assumptions. Therefore, the existence of a "singularity" (more precisely incompleteness) should really be thought of primarily as a breakdown of global hyperbolicity, i.e. determinism. In other words, it may so happen that M can be extended, but it cannot be extended as a globally hyperbolic manifold. We shall discuss in more detail the Cauchy problem and time evolution in the next chapter.

Chapter 5

Black Holes in General

In this chapter we give the definition of a black hole and formulate (in modern terms) the cosmic censorship hypotheses. The so-called weak and strong cosmic censorship conjectures were originally formulated by Penrose in 1969 and 1972. respectively. It is important to note that, in their modern formulation, the strong and weak cosmic censorship hypotheses are *independent*. In particular, the strong hypothesis does not imply the weak one, contrary to what one might expect simply based on the names alone.

5.1 Asymptotically Flat Spacetimes

In physics we often use methods such as multipole moments when the system is isolated and we are sufficiently far away. In general relativity we would thus like to formulate what an isolated system is, meaning it is in some sense flat at infinity.

The idea is to embed the spacetime into some larger manifold which will serve as a boundary. This can be demonstrated in the case of Minkowski space. Take the metric in polar coordinates:

$$ds = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)$$

and make a change in coordinates $u = t + r$, $v = t - r$, giving:

$$ds = -dudv + \frac{1}{4}(u - v)^2(d\theta^2 + \sin^2\theta d\phi^2)$$

These new u and v coordinates are null. Now we "compactify" by making the image of u and v bounded (using one's favorite diffeomorphism between $\mathbb{R}$ and some finite interval, say $u' = \tan^{-1} u$, $v' = \tan^{-1} v$). Finally, take $T = u' + v'$ and $R = u' - v'$. In these new coordinates, the metric becomes:

$$ds = \Omega^{-2}\left(-dT^2 + dR^2 + \sin^2 R d\theta^2 + \sin^2 R \sin^2\theta d\phi^2\right),$$

where $\Omega^{-2} = 4(1 + u^2)(1 + v^2)$. We have thus preserved the null cones (by only compactifying along the null directions), but the spacetime is now contained within some compact set. To see this, one must carefully note the ranges of any such coordinates; in this case $-\pi < T + R < \pi$ and $-\pi < T - R < \pi$, as well as $R \geq 0$.

$(R, T) = (0, -\pi)$ gives the lower vertex point, the past timelike infinity i^-. This is a 2-dim surface (a point in (R, T) plane is actually a sphere in Minkowski 4-dim space). Likewise the top vertex i^+ at $(R, T) = (0, \pi)$ is called the future null infinity. i^0 at $(R, T) = (\pi, 0)$ is the spacelike infinity. Finally, we have the 3-dim surfaces $\mathcal{I}^-$ given by

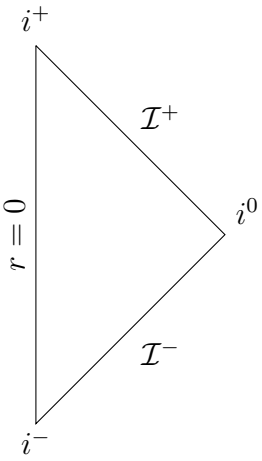

Figure 5.1: Conformal compactification of Minkowski spacetime. The angular coordinates have been suppressed. In $2+1$ dimensions this becomes a double cone. Removing the spacelike infinity i^0 (in this case a circle identified to a point) transforms the boundary into two pointed connected surfaces. Removing the timelike infinities $i^\pm$ these surfaces become smooth.

$T = \pi - R$ for $0 < R < \pi$ (future null infinity) and $\mathcal{I}^+$ given by $T = R - \pi$ for $0 < R < \pi$ (past null infinity).

The null infinities thus have the topology $S^2 \times \mathbb{R}$. We note that the boundary is not smooth at the vertices $i^\pm$ and i^0. In giving a general definition, we shall focus only on the null infinities. Thus the boundary will be smooth, but now the larger manifold need not be compact. This is the reason why we avoid using the term "conformal compactification" when referring to the general construction outlined below. For a unified treatment of null and spatial infinities see Ashtekar and Hansen [29] or Wald [4].

Definition 74 (Conformal equivalence)**.** Let g and $\tilde{g}$ be two Lonrentzian metrics on spacetime M. We say g_1 and g_2 are **conformally equivalent** if there exists a smooth $\Omega : M \to \mathbb{R}$, $\Omega \neq 0$ such that $g_2 = \Omega^2 g_1$. A smooth mapping $f : M \to \tilde{M}$ is **conformal** if the pullback of the metric on $\tilde{M}$ is conformally equivalent to the one on M, i.e. $f^*\tilde{g} = \Omega^2 g$.

Conformal mappings may change the lengths, but preserve the causal structure. In fact, two Lorentzian metrics (scalar products) g_1, g_2 on some vector space V will have the same light cone (the set of all vectors $v \in V$ for which $g(v, v) = 0$) iff $g_1 = Cg_2$ for some constant $C \neq 0$.

Definition 75 (Asymptotically simple, Penrose)**.** We say M is **asymptotically simple** if there exists a manifold $\tilde{M}$ with boundary $\partial\tilde{M}$, on which a Lorentzian metric $\tilde{g}$ is defined and into which M embeds via some diffeomorphism $\varphi : M \to \tilde{M}$ such that:

1. $\tilde{M} \setminus \partial\tilde{M} = \varphi(M)$, i.e. M embeds precisely into the interior of $\tilde{M}$.
2. There exists a smooth function Ω on $\tilde{M}$ such that on $\varphi(M)$ we have $\varphi^*\tilde{g} = \Omega^2 g$, i.e. φ conformally embeds M into $\tilde{M}$.
3. On ∂M we have $\Omega = 0$ and $d\Omega \neq 0$.
4. Every null geodesic in M has two endpoints in ∂M.

If in addition the Ricci tensor vanishes in the neighborhood of ∂M, we say M is **asymptotically empty**. We shall regularly assume that $\tilde{M}$ is strongly causal.

We sometimes say M is a *physical*, while $\tilde{M}$ is an *unphysical* spacetime. $d\Omega \neq 0$ in particular ensures that Ω can be used as a coordinate in a neighborhood of the boundary ∂M. By Taylor expanding in Ω we can then control the falloff towards infinity (∂M) of various physical fields. One can prove that the boundary of an asymptotically simple and empty space has a particularly nice topology:

Proposition 76
In an asymptotically simple and empty space M, the boundary $\mathcal{I} = \partial M$ is a null surface (i.e. its normal vector is null).

Proof. We follow Hawking & Ellis [3].

We first show that ∂M is null. This follows from the fact that Ricci tensor of the unphysical metric $\tilde{g}$ is related to the Ricci tensor of the physical metric g in the following manner:

$$\tilde{R}^{\nu}_{\mu} = \Omega^{-2} R^{\nu}_{\mu} - 2\Omega^{-1}\tilde{g}^{\nu\alpha}\tilde{\nabla}_{\mu}\tilde{\nabla}_{\alpha}\Omega + \left(-\Omega^{-1}\tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}\Omega + 3\Omega^{-2}\tilde{\nabla}_{\alpha}\Omega\tilde{\nabla}_{\beta}\Omega\right)\tilde{g}^{\alpha\beta}\delta^{\nu}_{\mu}$$

This is a standard result one gets when computing how Riemann curvature changes under conformal transformations.

Simplifying we get:

$$\tilde{R} = \Omega^{-2} R - 6\Omega^{-1}\tilde{g}^{\alpha\beta}\tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}\Omega + 3\Omega^{-2}\tilde{g}^{\alpha\beta}\tilde{\nabla}_{\alpha}\Omega\tilde{\nabla}_{\beta}\Omega.$$

Since we require that Ricci curvature be smooth at ∂M (at least C^1), where $\Omega = 0$, we get that (on ∂M) $\tilde{g}^{\alpha\beta}\tilde{\nabla}_{\alpha}\Omega\tilde{\nabla}_{\beta}\Omega = 0$. This means that $\nabla\Omega$ is a null vector and, since $d\Omega \neq 0$, it must be a nonzero null vector, whose integral curves are null generators of ∂M. □

Proposition 77
$\mathcal{I} = \partial M$ has two components:

- *The future null infinity $\mathcal{I}^+$, i.e. the set of future endpoints of light geodesics (the set of boundary points where the future light geodesics are outward-pointing). This means for each point $p \in \mathcal{I}^+$, the past null cone lies in $\tilde{M}$.*
- *The past null infinity $\mathcal{I}^-$, i.e. the set of past endpoints of light geodesics (the set of boundary points where the past light geodesics are outward-pointing). This means for each point $p \in \mathcal{I}^-$, the future null cone lies in $\tilde{M}$.*

Proof.

- To show ∂M has at least two components, we proceed as follows.

  Let $p \in \mathcal{I}$. Since $\mathcal{I}$ is the boundary of $\tilde{M}$, we can find some chart on a neighborhood U (in $\tilde{M}$) around p such that $\mathcal{I} \cap U$ looks like a hyperplane, where $M \cap U$ is on one side of that hyperplane. As $\mathcal{I}$ is null, either $I^+(p)$ or $I^-(p)$ do not intersect M. But if the open set, say, $I^+(p)$ does not intersect the open set M, then the same holds if we take the closure $\overline{I^+(p)}$. Since $J^+(p)$ is contained in that closure, we see that causal curves ending at p must all be future-pointing or all be past-pointing.

In other words, a point in $\mathcal{I}$ can't be both a past and a future endpoint at the same time.

If $p \in \mathcal{I}$ is a future endpoint of some geodesic, then so too must be all points in some neighborhood of p. Indeed, one can could otherwise construct a sequence of curves with past endpoints $p_n \to p$; but then the limit must have a past endpoint at p.

Therefore, we can cover $\mathcal{I}$ with sufficiently small open sets which consist entirely of either future endpoints or past endpoints. U^+ which have future endpoints in $\mathcal{I}$ are mutually disjoint with U^- having past endpoints in $\mathcal{I}$. Taking the union of all U^+ and U^- we get two open sets whose union is ∂M; ∂M is thus not connected.

- For spacetime of dimension > 2 there can be only 2 components (for dimension 2 Minkowski space gives a counterexample; there are two future and two past null infinities). The proof hinges on the fact that in dimension > 2, the space of future null directions at p is connected (this is not the case in 2-dim case as it is a two point set).

  If say $\mathcal{I}^+$ were disconnected consisting of some two components C_1 and C_2, then we will show that one can find a point $p \in M$ through which some future null geodesics go into C_1, and others go into C_2. Now, a null geodesic at p can be identified with its (unique) endpoint in $\mathcal{I}^+$. This mapping $\rho : \gamma_p \mapsto e_{\gamma_p}$ is continuous (if $\alpha_n \to \alpha$ then the endpoints e_n of α_n converge to the endpoint e of α). Thus, the set of null directions at p has at least two components - a contradiction.

  We now show the existence of point p. Assume to the contrary that for all $p \in M$ all future null geodesics go exclusively into one component but not the other. Then we can show that around any $p \in M$ one can find an open set such that any future null geodesic through that set goes into the same component as p.

  Indeed, assume this fails at p, whose future null geodesics go into C_1. Then we can construct a sequence p_n converging to p such that future null geodesics α_n going through p_n all go into C_2. But now take the limit of those geodesics and get a causal curve α going through p and reaching C_2. This causal curve is actually a geodesic, which gives a contradiction.

  To see that the curve indeed is a geodesic, we must show that no two points on α can be connected by a timelike curve. This is true because if $q \in \operatorname{Im}\alpha$ and $q \in I^+(p) \cap C$ is a point in some small convex neighborhood around p, then q being a limit point of the sequence α_n implies that $q_n \in I^+(p) \cap C$ for n sufficiently large, where $q_n \in \operatorname{Im}\alpha_n$. One can then repeat the argument for p as well to conclude $p_n \in J^-(q_n) \cap C$. This is impossible because α_n are null and we can take C so small as to make some α_n not have conjugate points between p_n and q_n. Thus we can cover α with neighborhoods on which α is a null geodesic; so it must be globally a null geodesic (up to reparametrization).

  Thus we can cover M with open sets through which null geodesics go exclusively into one component or the other. But the union of all open sets which go into C_1 is disjoint with the union of all open sets which go into C_2 thereby showing that M is not connected.

□

We should note that, in particular, 4 is too strong as it requires that every null geodesic reach infinity so will generally exclude black holes (which contain bound orbits). For example, the Schwarzschild spacetime does not satisfy this criterion. Asymptotically simple spaces are not just simple in the asymptotic region; in fact we can show:

Proposition 78
An asymptotically simple and empty spacetime must be globally hyperbolic.

One can further show that M will be topologically equal (homeomorphic) to $\mathbb{R}^4$ and $\mathcal{I}^{\pm}$ to $\mathbb{R} \times S^2$ (see Geroch [33] or Hawking & Ellis [3] or the appendix of Penrose [51]).

Proof. We break the proof into two parts:

- M is causally simple, i.e. $J^{\pm}(p)$ are closed. For this we use the fact that every null geodesic has two endpoints in ∂M. Take a sequence of points $q_n \in J^+(p)$ which converge to some q. We show that q must be in $J^+(p)$, i.e. that $J^+(p)$ is closed.

  Indeed, let α_n be some causal curves from p to q_n. Continue these curves all the way up to future null infinity by following some null geodesics, so that they become future inextendible. Therefore, there must exist a future inextendible causal limit curve α, which goes from p and passes through q.

- M is globally hyperbolic. We use results from B. First put a measure $\nu d\mu$ on M with total volume $\int_M \nu d\mu = 1$. Since M is causally simple by the previous point, the functions $V_{\pm}(p) = \int_{J^{\pm}(p)} \nu d\mu$ are continuous. We must prove that V^+ decreases to 0 along any future directed causal curve γ (since we don't yet have global hyperbolicity, lemma 99 doesn't directly apply).

  To do this, consider $F = \bigcap_{p \in \operatorname{Im}\gamma} J^+(p)$; if it is empty for every future-inextendible causal curve, we get out result. Assuming it is not empty, we get that all points $p \in \operatorname{Im}\gamma$ can then be connected via causal curves to some $q \in M$, i.e. $\operatorname{Im}\gamma \subset J^-(q)$.

  This means that γ cannot have an future endpoint in M. Indeed, if $e \in M$ is an endpoint, consider a limit of points $p_n = \gamma(t_n) \to e$ and a sequence of curves α_n going along γ until p_n, then following some causal curve to q and finally some null geodesic to $\mathcal{I}^+$. α_n has a limit curve α passing through e and q, but then γ can be future extended; a contradiction.

  Thus γ goes all the way up to $\mathcal{I}^+$. On the other hand, this means that it must exit $J^-(K)$ for any compact K (as causal curves in $J^-(K)$ do not reach $\mathcal{I}^+$). Now the proof of lemma 99 actually applies, so we get the result. As in appendix B, the level sets of $\tau(p) = \frac{V^-(p)}{V^+(p)}$ are now Cauchy surfaces, so M is globally hyperbolic.

□

We therefore give a weaker definition:

Definition 79 (Weakly asymptotically simple, Penrose)**.** We say M is **weakly asymptotically simple** there is a piece of M isometric to the boundary of an asymptotically simple space. More precisely, if there exists an open set $U \subset M$ and an asymptotically simple space M' with an open neighborhood of the boundary $U' \supset \partial M'$ such that $U' \cap M'$ is isometric to U.

The idea is that a weakly asymptotically simple spacetime possesses the conformal infinity of an asymptotically simple spacetime, but may possess other infinities as well.

5.2 Black Holes

Finally, we can define black holes in general:

Definition 80 (Black hole)**.** Let M be a weakly asymptotically simple spacetime and let $\mathcal{I}^+$ be its future null infinity. If the (causal) past of $\mathcal{I}^+$, $J^-(\mathcal{I}^+)$, does not cover M, then we say M contains a **black hole region** $\mathcal{B} = M \setminus J^-(\mathcal{I}^+)$. We call the topological boundary $\mathcal{H} = \partial\mathcal{B} = (\partial J^-(\mathcal{I}^+)) \cap M$ the **event horizon** of the black hole.

Intuitively, a black hole region is the set of events from which no lightlike geodesic can escape to infinity. Analogously, one may define a white hole region as $\mathcal{W} = M \setminus J^+(\mathcal{I}^-)$.

Actually, we still have a problem with the definition of asymptotic simplicity - the weak asymptotic simplicity is now too weak! In particular, the notion of a weakly asymptotically simple spacetime does not capture the global asymptotic structure of Minkowski spacetime. Consider the following:

Example 81 (Geroch & Horowitz [31])**.** Consider Minkowski space-time with the causal future of the origin removed (i.e., retain the region given, in the usual coordinates, by $t < \sqrt{x^2 + y^2 + z^2}$). This space-time is weakly asymptotically simple.

Thus, Minkowski spacetime with only a portion of its usual boundary (the portion that lies outside outside the causal future of 0) is weakly asymptotically simple.

Now an immediate problem arises: the $\mathcal{I}^+$ in the definition of weakly asymptotically simple space is not unique. It is therefore not clear exactly which $\mathcal{I}^+$ (that makes M into a weakly asymptotically simple space) one should take in the definition of a black hole.

If we require that for at least one $\mathcal{I}^+$ its past $J^-(\mathcal{I}^+)$ does not cover M, then Minkowski space has a black hole. On the other hand, if we require that it hold for all possible $\mathcal{I}^+$ its past $J^-(\mathcal{I}^+)$ does not cover M, then Minkowski space-time with an asymptotic portion of the null cone of the origin removed possesses a black hole.

Therefore, when discussing black holes it is preferable to use the following more restrictive definition:

Definition 82 (Asymptotically flat, Geroch & Horowitz [31])**.** Weakly asymptotically simple spacetime is said to be **asymptotically flat** if in addition its null infinities $\mathcal{I}^{\pm}$ are both topologically $\mathbb{R} \times S^2$ and $\mathcal{I}$ is **complete**[1]. Of course, we shall assume that such a spacetime is asymptotically empty as well (otherwise using the term "flat" doesn't make much sense).

Notice that the previous example does not have a complete $\mathcal{I}$.

So far we have only defined a "black hole region" as a collection of events. To talk about "black hole at time τ", we need some additional assumptions. Following Wald [4], we say:

Definition 83. M is **strongly asymptotically predictable** if in $\tilde{M}$ there exists an open region $\tilde{V}$ with $\overline{M \cap J^-(\mathcal{I}^+)} \subset \tilde{V}$ such that $(\tilde{V}, \tilde{g})$ is globally hyperbolic.

[1] One can formulate completeness in terms of the generators of the null surface $\mathcal{I}$. In particular the normal $\nabla\Omega$ is also tangent to $\mathcal{I}$. Therefore, its integral curves (the generators) are contained in $\mathcal{I}$ and we have seen these must be null geodesics. $\mathcal{I}$ is complete precisely when these geodesics can be extended indefinitely, i.e. when the normal field is complete.

This is essentially the same definition as in Hawking & Ellis [3], if we take into consideration proposition 9.2.3.

In particular, since we now have some family of Cauchy hypersurfaces $S(\tau)$ foliating $\tilde{V}$ and hence the exterior of the Black hole, we can define a black hole at some time τ as a component of the set $B(\tau) = S(\tau) \setminus J^-(\mathcal{I}^+) = \mathcal{B} \cap S(\tau)$. This is just the region of Cauchy surface $S(\tau)$ from which light cannot escape to null infinity.

Note that the assumption of the exterior being globally hyperbolic has certain implications. Namely, there cannot be any "naked" singularities. If $q \in M \cap \tilde{V}$ and if $S(\tau)$ is in the past of q ($q \in J^+(S(\tau))$), then, by virtue of $S(\tau)$ being a Cauchy surface, all inextendible past-directed causal curves must intersect $S(\tau)$. This means no singularities are visible to an observer outside the black hole, i.e. we cannot have any past-directed causal geodesics just ending (a naked singularity in the past of some observer); they must go as far back as there are Cauchy surfaces.

The event horizon $\mathcal{H} = \partial\mathcal{B} = (\partial J^-(\mathcal{I}^+)) \cap M$, being the boundary of a past set, is a closed achronal topological hypersurface. We note first that $I^\pm$ must be closed in $\tilde{M}$, as ∂M is closed, and components of closed sets are closed. Thus (by proposition 67) every point in $\partial J^-(\mathcal{I}^+)$ lies on some null geodesic, which must be contained in $\partial J^-(\mathcal{I}^+)$, which is past inextendible or has a past endpoint in $\mathcal{I}^+$. Since we are working in the physical spacetime M, both cases amount to the same thing.

It must be said, however, that an event horizon is *not* generally a smooth manifold (actually a nowhere C^1 example can be constructed; see [6]). Many authors (Wald [4] and Hawking & Ellis [3]) assume differentiability when proving certain theorems about horizons, in particular, the area theorem. For nowhere differentiable horizons it is not clear that area is even well defined, but in this case it has been shown (see [6] again) that any differentiability assumption can be dispensed with.

5.3 Cosmic Censorship

From a modern point of view, the cosmic censorship hypotheses are conjectures about the nature of maximal Cauchy developments. We therefore first discuss the Cauchy problem in general relativity.

Definition 84 (Cauchy problem)**.** We follow Choquet-Bruhat [5] (as one should) in these matters.

- An **initial data set** is a triple (Σ, h, K), where (Σ, h) is a Riemannian 3-manifold and K a symmetric 2-tensor on Σ.
- A **development** of the initial data (Σ, h, K) is a spacetime M for which there exists an embedding $\varphi : \Sigma \to M$ having the following properties:
  1. The metric h is the pullback of g by φ, $h = \varphi^* h$, or equivalently, if we identify Σ with its image $\varphi(\Sigma)$ in M, then h is simply the induced metric. Since h is Riemannian, $\varphi(\Sigma)$ is spacelike.
  2. The tensor K is the pullback by φ of the second fundamental form (extrinsic curvature) of $\varphi(\Sigma)$ as a submanifold of M.
- We call the development (M, g) **Einsteinian** if the metric g satisfies the Einstein equations. Of course, here the initial data is just a set of initial conditions for

Einstein equations (so is not completely arbitrary and must satisfy any relations imposed by the Einstein equations).

- A development is called globally hyperbolic if M is globally hyperbolic with $\varphi(\Sigma)$ as a Cauchy surface.
- A development is called maximal (or inextendible) if it cannot be extended (i.e. isometrically embedded) to another development. Note this does *not* say that it is inextendible as a Lorentzian manifold.

Theorem 85 (Choquet-Bruhat, Geroch)
For a given initial data set (Σ, h, K) to the vacuum Einstein equations (or for a suitable matter system) there exists a unique (up to isometry) maximal globally hyperbolic Einsteinian development of (Σ, h, K).

Here the fact that the development is *unique* and maximal, guarantees that every other (Einsteinian, globally hyperbolic) development can be isometrically embedded into the maximal one.

Proof. See the original paper by Choquet-Bruhat and Geroch [22] or Ringström [9] (theorem 16.6.) Let us note here that the original proof uses Zorn's lemma, but this can be avoided (i.e. a constructive proof can be given) if one wishes (see Sbierski [28]). □

We can now understand the following modern formulation of the famous strong censorship conjecture:

Conjecture 86 (Strong cosmic censorship)
For "generic" (i.e. not "finely tuned"[2]) initial data for the vacuum equations or for suitable Einstein–matter systems, the maximal Cauchy development is inextendible (as a Lorentzian manifold).

Roughly speaking, this conjecture asserts that general relativity is a deterministic theory in the sense that motions of all observers for all times should be determinable from initial conditions. In particular, one should not be able follow a geodesic outside the maximal development. In fact, if we can extend the maximal development, the extension will usually be severely non-unique. One can therefore think about this conjecture as a statement on uniqueness of the global solution.

Example 87. For a Kerr spacetime 2.2 take some Cauchy slice in the regions $I \cup II$. This slice is taken to be the initial condition for the maximal Cauchy development. The maximal development then generates the Kerr spacetime, but only up to the interior horizon (i.e. Cauchy horizon) of the Kerr black hole. We have seen that in the interior horizon causality breaks down. The solution inside the Cauchy horizon is usually given as

[2]This is not really precisely defined, as a precise notion of "generic" would require a better understanding of the counterexamples. One could interpret this topologically to mean dense (and open) in the set of all possible initial data in some relevant topology. Alternatively, one could interpret it in a probabilistic sense and introduce some probability measure in the space of all possible initial data for which the exceptional set is of measure zero. Though, constructing such a measure (or topology) does not seem to have a bearing on the problem. On a much more elementary level, a Kerr solution is, as we will see, parameterized by two numbers - a and M. Schwarzschild spacetime is then exceptional in that family for it is given by a specific choice of $a = 0$. In particular, it is more symmetric than the rest (by having full spherical symmetry).

a part of the unique analytic extension. But generally there are multiple (smooth) ways one could extend the spacetime beyond the Cauchy horizon and from the perspective of dynamics and Einstein equations, no way is better than the other.

Therefore, the hope is, one shouldn't be able to cross the second horizon in any meaningful sense. Given that no black hole is completely isolated from the rest of the universe, the initial data will not be exactly that of the Kerr solution (say a gravitational wave passes by), so Cauchy horizons should not be stable under generic small perturbations. In fact, there are more direct physical reasons (having to do with the infinite blueshift on the Cauchy horizon) why we believe it to be unstable.

Actually, in [27] Dafermos and Luk show that a C^0 formulation of the strong censorship hypothesis is not true - specifically in the case of Kerr spacetime (under the assumption of stability of the exterior of Kerr spacetime). This, in particular, means that under generic conditions, one can go beyond the Cauchy horizon if one is permitted in using a C^0 metric there.

Though, one may take this formulation to be too strong. If we require Christoffel symbols to be locally square integrable [3] (or metric C^2 for that matter), then it seems the problem is still open. In particular, the Christoffel symbols should not blow up at the Cauchy horizon.

We now turn to the (confusingly named) weak cosmic censorship conjecture. We mention again that the "weak" conjecture is *not* logically weaker; it does not follow from the strong cosmic censorship. For asymptotically flat initial data one can show (see again Geroch & Horowitz [31]) that the maximal evolution is weakly asymptotically simple (and empty). The question remains, though - is the null infinity complete?

Conjecture 88 (Weak cosmic censorship)
For a generic asymptotically flat vacuum initial data, the maximal Cauchy development has a complete null infinity $\mathcal{I}^+$.

This can be stated roughly as "faraway observers live forever".

The conjecture posits that there can be no "naked singularities" visible from the infinity, i.e. all singularities have to be hidden beyond event horizons[4]. The conjecture would fail, for instance, if the Cauchy horizon cuts off the null infinity at some finite parameter value. That the word "generic" is necessary can be seen from the discussion in Christodoulou [34].

5.4 Stationary Black Holes

It turns out that black hole solutions of the Einstein equations have a very restrictive form under some additional hypotheses which guarantee that the black hole has "settled down" (i.e. is stationary). This is the celebrated no hair theorem. We should immediately

[3] With square integrable Christoffel symbols, even though one does not have the Riemann curvature tensor in the usual sense, one can write down the Einstein equations and study their weak solutions. Thus inextendibility here means we can rule out any reasonable notion of weak solution to the Einstein equations.

[4] This is not entirely correct, as one can construct examples where the null infinity is complete, but there is no black hole region. It is really the fact that observers on the null infinity exist for all times that is the essence of the conjecture.

note that the problem has not been completely solved as it rests on certain hard to justify assumptions that have yet to be completely removed.

The no hair theorem can be broken into two pieces - the uniqueness theorem for stationary axisymmetric spacetimes and the rigidity theorem which guarantees axial symmetry.

5.4.1 Uniqueness theorem

We first extend the definition of stationary and axisymmetric spacetime in the asymptotic case:

Definition 89.

- We call M **asymptotically stationary** (or pseudo-stationary by Carter's terminology) if there exists a 1-parameter group of isometries g_t acting on M, whose Killing field X is timelike near $\mathcal{I}^+$ and $\mathcal{I}^-$.
- We will say an asymptotically stationary M is **axisymmetric** if there is a one-parameter cyclic isometry group g_ϕ, $(0 \leq \phi \leq 2\pi)$ of M which commutes with g_t, and whose Killing field Y is spacelike near $\mathcal{I}^+$ and $\mathcal{I}^-$. We further assume that the axis of symmetry is non-empty.

We now define Killing horizons and surface gravity:

Definition 90 (Killing horizon)**. Killing horizon** is a null hypersurface, whose generators are given by some Killing field. It is usually required that the Killing horizon be connected.

Thus on a Killing horizon we can choose the normal vector field to be Killing. Let X be a Killing field and consider the set $\{g(X,X) = 0 \mid X \neq 0\}$, then a Killing horizon is a null hypersurface coinciding with a connected component of that set.

Definition 91 (Surface gravity)**. Surface gravity** κ of a Killing horizon H defined by some Killing field X is given by the formula $d(g(X,X))|_H = -2\kappa X^\flat$.

Since X is normal to the null hypersurface H, we have seen that $d(g(X,X))$ must be proportional to X. Indeed, from $g(X,X) = 0$ we have $Yg(X,X) = 0$, i.e. $d(g(X,X))(Y) = 0$ for all Y tangent to H. On the other hand, the normal space (and therefore its dual) is at each point of H one dimensional.

The terminology stems from the fact that κ measures acceleration of integral curves of the Killing field. This can be seen as follows:

If γ solves the equation $\dot\gamma(t) = X_{\gamma(t)}$, then the acceleration can be found as $a = \frac{D}{dt}\dot\gamma$, so we have:

$$a_\mu = \frac{D}{dt}\dot\gamma_\mu = \nabla_{\dot\gamma}\dot\gamma_\mu = \nabla_X X_\mu = X^\nu \nabla_\nu X_\mu$$

Since X is Killing, $\nabla_\nu X_\mu = -\nabla_\mu X_\nu$. Thus on the horizon:

$$\begin{aligned} a_\mu &= X^\nu \nabla_\nu X_\mu = -X^\nu \nabla_\mu X_\nu = -\frac{1}{2}\nabla_\mu (X^\nu X_\nu) = -\frac{1}{2}\partial_\mu g(X,X) \\ &= -\frac{1}{2} dg(X,X)(\partial_\mu) = \kappa X^\flat(\partial_\mu) = \kappa X_\mu. \end{aligned}$$

By a **nondegenerate** Killing horizon (with Killing field X), we mean that $g(X,X) = X^\mu X_\mu$ has non zero gradient on the Horizon. Equivalently, the surface gravity κ of the horizon must be non zero.

The first theorem on uniqueness of black holes (in the static vacuum case) was due to W. Israel (1967). It gave conditions (not assuming spherical symmetry) under which the only solution was the Schwarzschild spacetime.

In the '70s it became clear that one could obtain a result of similar nature for an axisymmetric spacetime. Indeed, it was the work of Carter in '72 (and later Robinson in '75) that gave birth to the following result:

Theorem 92 (Robinson and Carter uniqueness theorem)
Let M be a (i) strongly asymptotically predictable (ii) stationary axisymmetric spacetime which (iii) satisfies the Einstein vacuum equations. Assume further that the event horizon is a non-degenerate Killing horizon that is topologically a 2-sphere (in particular it is connected). Then M is uniquely specified by two parameters - the mass m and the angular momentum a. More precisely, the family of solutions is the two-parameter Kerr family of metrics 2.2.

Proof. The (general electrovac[5]) problem reduces to a boundary value problem on some 2-manifold (see e.g. the original paper [36] or a more in-depth discussion [35] by Carter). Uniqueness of the solution was then proven only in the vacuum case by Robinson in [39]. However, the divergence identity on which the proof hinges was only understood later by Mazur ([38]), who then gave the proof in the general electrovac case (See e.g. Mazur [37]). Alternatively, the proof is also given in Heusler [10]. □

On the other hand, Hawking had proven (1972) that, assuming analyticity, the spacetime must necessarily be axisymmetric and event horizon must be a Killing horizon.

5.4.2 Rigidity theorem

Let us define the **ergosphere** of a stationary regular predictable spacetime as a region of $\overline{J^-(\mathcal{I}^+)} \cap \overline{J^+(\mathcal{I}^-)}$ on which the Killing field X is spacelike. Intuitively, it is impossible for a particle to follow the integral curves of the field X, i.e. to remain at rest when viewed from infinity.

We now give a description of the Rigidity theorem as can be found in Hawking & Ellis [3].

Definition 93. We say a spacetime M is **regular predictable** if M is strongly asymptotically predictable with Cauchy surface S and the following holds:

1. $S \cap \overline{J^-(\mathcal{I}^+)}$ is homeomorphic to $\mathbb{R}^3 \setminus V$, where V is an open set with compact closure.
2. S is simply connected
3. For large enough τ, $S(\tau) \cap \overline{J^-(\mathcal{I}^+)}$ is contained in $J^+(\mathcal{I}^-)$.

The condition 3 just says that we can actually fall into the black hole; in particular for large enough τ, $\partial B(\tau) \subset J^+(\mathcal{I}^-)$ so that we can "see" the event horizon from past infinity. If 1 and 2 are satisfied, the boundary of the black hole will be compact and connected (Hawking & Ellis [3] proposition 9.2.6). In fact, we have:

[5] We mention that in the electrovac case there are 3 parameters (mass, angular momentum and charge); the solution is given by the so-called Kerr-Newman metric.

Theorem 94
For a stationary regularly predictable M, each component of the event horizon $\partial B(\tau)$ in $J^+(\mathcal{I}^-)$ will be homeomorphic to a 2-sphere.

Proof. See Hawking & Ellis [3] proposition 9.3.2. □

Theorem 95
Let M be a stationary non-static regular, predictable spacetime in which the ergosphere intersects the domain of outer communications $J^-(\mathcal{I}^+) \cap J^+(\mathcal{I}^-)$. Assume further that M is an analytic manifold (and the metric analytic as well), then M is axisymmetric.

Note that the static case is covered by Israel's theorem (and later extensions thereof).

Proof. See Hawking & Ellis [3] theorem 9.3.6. □

One therefore sees that a single black hole in vacuum will eventually settle down into the Kerr family - this is the so-called no hair theorem.

Not so fast!

5.4.3 Issues and later developments

There are a couple of issues with the previous analysis (in particular with theorems 94 and 95). Let us first comment on theorem 94 (i.e. theorem 9.3.2 in [3]). There appears to be a problem with the proof of theorem 9.3.2 as given in [3]. Namely, the argument given does not rule out a toroidal topology. This was more or less settled by Wald and Chruściel in [43].

There appears to have been a hole in the proof of theorem 95 (theorem 9.3.6 in [3]) as well. In fact, as it is currently formulated, theorem 95 is simply wrong - Chruściel has constructed a counterexample in [46]. The problem is in actually globally extending the group of isometries g_t initially (and correctly) defined only near the event horizon. The whole issue has, fortunately, been solved in Chruściel [47], providing us with the desired axial symmetry. See also [44] and [45].

Removing the analyticity assumption is still an open problem. We should mention, though, that there has been some notable progress. For instance, assuming some scalar identity Ionescu and Klainerman [40] prove the rigidity theorem without analyticity. Moreover, in [41] Alexakis, Ionescu and Klainerman prove rigidity holds provided that the spacetime is close to Kerr; this result in particular suggests the conjecture about the nonlinear stability of the Kerr exterior (whose veracity would, as we have mentioned, disprove the C^0 version of the strong cosmic censorship conjecture). Additionally, the same authors prove rigidity for small angular momenta in [42].

Thus, the conclusion that the purported proof of theorem 95 allows us to make is:

Theorem 96 (Hawking Rigidity Theorem)
Under the assumptions of theorem 95 (analyticity being crucial), there exists a Killing field X defined near the event horizon $\mathcal{H}$, such that on $\mathcal{H}$ the integral curves of X coincide with the null generators of $\mathcal{H}$. Simply stated, the event horizon is a Killing horizon.

Conclusion

The main focus of this thesis have been the classical results from the '60s and '70s period. We have covered some basic causality theory, the celebrated singularity (incompleteness) theorems, and equally celebrated "no hair theorem". Notably however, I have not touched upon the subject of black hole thermodynamics, which was first developed geometrically by Bekenstein, Carter, Bardeen, and Hawking (and again cleaned up by Chruściel). Here it became apparent that black hole mechanics had striking similarities to thermodynamics (surface gravity acting as temperature and area of a black hole horizon acting as entropy).

Later Hawking applied QFT to black holes and discovered Hawking radiation, thereby establishing that the link between black holes and thermodynamics is not merely an analogy but something more exact (the temperature of Hawking radiation being exactly proportional to surface gravity). On the other hand, this discovery led to the black hole information paradox, the complete resolution of which is still an ongoing area of research.

Appendix A

Ricci Tensor of a Spherically Symmetric Metric

Assume a metric of the form

$$ds^2 = F(v,r)dv^2 + 2X(v,r)dvdr + Y^2(v,r)d\Omega^2$$

and denote by $f' = \partial_r$ and $\dot{f} = \partial_v f$ the respective partial derivatives. Then we have:

$$\begin{aligned} R_{uu} = \frac{1}{2X^2Y}(&FXYF'' - 2FXY\dot{X}' + 2FXF'Y' - 4FX\dot{X}Y' - FYF'X' \\ &+ 2FY\dot{X}X' - 4X^3\ddot{Y} + 2X^2\dot{F}Y' - 2X^2F'\dot{Y} + 2X^2\dot{X}\dot{Y}) \end{aligned} \tag{A.1}$$

$$R_{rr} = \frac{2}{XY}(-XY'' + X'Y') \tag{A.2}$$

$$R_{\theta\theta} = \frac{1}{X^3}(FXYY'' + FX(Y')^2 - FYX'Y' + X^3 - 2X^2\dot{Y}' - 2X^2\dot{Y}Y' + XYF'Y') \tag{A.3}$$

We also have two more nonzero components $R_{\phi\phi} = R_{\theta\theta}\sin^2\theta$ and $R_{ur} = \frac{1}{2X^2}(-YF'X' + 2Y\dot{X}X' + XYF'' - 2XY\dot{X}' - 4X^2\dot{Y}' + 2XF'\dot{Y})$ for which we have no use. To calculate the components of the Ricci tensor I used einsteinpy. Then the following python code does the job:

```
import sympy
from einsteinpy.symbolic import RicciTensor, RicciScalar, metric
from sympy import Function, Symbol, sin
from einsteinpy.symbolic.predefined import AntiDeSitter


v = Symbol('v')
r = Symbol('r')
theta = Symbol('theta')
phi = Symbol('phi')
F = Function('F')(v,r)
X = Function('X')(v,r)
Y = Function('Y')(v,r)
```

```
arr=[[F, X,0,0], [X,0,0,0], [0,0,Y**2,0], [0,0,0,Y**2*sin(theta)**2]]
syms=[v,r,theta,phi]

g=metric.MetricTensor(arr, syms)
g.tensor() #display metric

Ric = RicciTensor.from_metric(themetric)
Ric.tensor() #display Ricci tensor
```

Appendix B

Proof of Geroch's Theorem

We follow [6]. Let us state once more what is to be proven:

Theorem 97 (Geroch)
A globally hyperbolic spacetime M must have a globally defined time function τ, whose level sets are Cauchy surfaces.

Proof. Let φ_i be a partition of unity on M subordinate to some convex cover. Let h be some complete Riemann metric on M with volume form dV and μ the unique Radon measure induced by the form dV. More precisely, μ is induced (via the Riesz representation theorem - see [17] or [18]) by the positive linear functional $f \mapsto \int_M f dV$ defined on the space of continuous maps $f : M \to \mathbb{R}$ with compact support.

We put $V_i = \int_M \varphi_i d\mu < \infty$ to be the volume of M as measured by the i-th partition of unity. Next, set $\nu = \sum_{i\in\mathbb{N}} \frac{1}{2^i V_i}\varphi_i$, then $\nu d\mu$ is a finite measure: $\int_M \nu d\mu = 1$. Now define $V_\pm(p) = \int_{J^\pm(p)} \nu d\mu$. Since $J^\pm(p)$ contains the open set $I^+(p)$, it is clear that $V \pm (p) > 0$, and since $M \setminus J^\pm(p)$ contains open sets as well, we have $V^\pm(p) < 1$.

We shall later show that, when $J^\pm(p)$ are closed (as in globally hyperbolic spaces), the functions $V^\pm$ are continuous. We also show that V^- tends to 0 along any past directed causal curve, while V^+ tends to 0 along any future-directed causal curve.

Now we set $\tau(p) = \frac{V^-(p)}{V^+(p)}$, then it is clear that τ is continuous. If $\gamma : (a, b) \to M$ is an inextendible future-directed causal curve, then $\lim_{t\to b} \tau(\gamma(t)) = \infty$ and $\lim_{t\to a} \tau(\gamma(t)) = 0$. Thus τ runs from 0 to ∞ on all inextendible future-directed curves γ.

In particular, γ intersects every level set of τ at least once. Furthermore, τ is strictly increasing on future-directed curves, so γ intersects every level set exactly once; the level sets are thus Cauchy surfaces. Indeed, $J^+(q) \subset J^+(p)$ and $J^-(q) \supset J^-(p)$ whenever $p \leq q$, but in globally hyperbolic spaces we have $I^\pm(p) \neq I^\pm(q)$ for $q \neq p$. Now, since $\partial I^\pm(p)$ is a Lipschitz topological hypersurface, it has measure 0 (Lipschitz maps preserve sets of measure 0), so $I^\pm$ and $J^\pm$ have the same measure. Thus, along any future-directed causal curve γ, V^+ must be strictly decreasing and V^- strictly increasing, so τ must be strictly increasing. □

To actually finish the proof we need to show $V^\pm$ are continuous and have the appropriate asymptotic behavior:

Lemma 98
Suppose M is causally simple, i.e. that $J^\pm(p)$ are all closed sets, then $V^\pm$ are continuous.

Proof. Let p_i be some sequence converging to p. We prove $V^{\pm}(p_i) \to V^{\pm}(p)$. Denote by χ_A the characteristic function of set A (equaling 1 on A and 0 outside A). Let q be any point in the past of p: $p \in I^+(q)$ (i.e. $q \in I^-(p)$). Since $I^+(q)$ is a neighborhood of p, for large enough i, all p_i fall into $I^+(q)$ (i.e. q must be in the past of p_i). Thus, for large enough i, $\chi_{I^-(p_i)}(q) = 1 = \chi_{I^-(p)}(q)$. Since $\chi_{I^-(p)}(q) = 0$ for $q \notin I^-(p)$, we get generally $\chi_{I^-(p_i)} \geq \chi_{I^-(p)}$ (on M), so $\liminf_{i\to\infty} \chi_{I^-(p_i)}(q) \geq \chi_{I^-(p)}(q)$ for all $q \in M$. Since $\partial I^{\pm}$ is a topological Lipschitz surface, it has measure 0, so we get:

$$\liminf_{i\to\infty} \chi_{J^-(p_i)}(q) \geq \chi_{J^-(p)}(q),$$

which holds almost everywhere (everywhere except for, perhaps, a set of measure 0).

We now must prove the converse inequality:

$$\limsup_{i\to\infty} \chi_{J^-(p_i)}(q) \leq \chi_{J^-(p)}(q).$$

To establish this, it is sufficient to show that, whenever $\limsup_{i\to\infty} \chi_{J^-(p_i)}(q)$ is 1, $\chi_{J^-(p)}(q)$ must be as well (note that $\limsup_{i\to\infty} \chi_{J^-(p_i)}(q)$ can only take values 1 or 0 as χ can only take those values). Let q be any point for which $\limsup_i \chi_{J^-(p_i)}(q) = 1$, then there exists a subsequence p_j such that $\chi_{J^-(p_j)}(q) > 0$, i.e. $p_j \in J^+(q)$. But since $J^+(q)$ is closed and $p_j \to p$, we also must have $p \in J^+(q)$.

Finally, we have:

$$\liminf_{i\to\infty} \chi_{J^-(p_i)}(q) \geq \limsup_{i\to\infty} \chi_{J^-(p_i)}(q),$$

which holds almost everywhere (a.e.). This means that $\lim_{i\to\infty} \chi_{J^-(p_i)}(q)$ exists a.e. and must be equal to $\chi_{J^-(p)}(q)$ a.e.

Since the (nonnegative) functions $\chi_{J^-(p)}(q)$ are bounded by 1 above and constant functions are integrable in measure $\nu d\mu$, the Lebesgue dominated convergence theorem gives:

$$V^-(p) = \int_M \chi_{J^-(p)} \nu d\mu = \lim_i \int_M \chi_{J^-(p_i)} \nu d\mu = \lim_i V^-(p_i).$$

Continuity of V^+ follows analogously (just change the time orientation everywhere in the argument). □

Lemma 99
Let M be globally hyperbolic. V^- tends to 0 along any past-directed inextendible causal curve $\gamma : [0, b)$. Similarly, V^+ tends to 0 along any future-directed inextendible causal curve.

Proof. Partition the manifold M using some sets X_i with compact closure; thus $\chi_{X_i} + \chi_{X_j} \leq 1$ for $i \neq j$. Using the dominated convergence theorem (twice), we get:

$$\begin{aligned}\lim_{k\to\infty} \sum_{i=k}^{\infty} \int_{X_i} \nu d\mu &= \lim_{k\to\infty} \sum_{i=k}^{\infty} \int_M \chi_{X_i} \nu d\mu = \lim_{k\to\infty} \int_M \sum_{i=k}^{\infty} \chi_{X_i} \nu d\mu \\ &= \lim_{k\to\infty} \int_M \chi_{\bigcup_{i=k}^{\infty} X_i} \nu d\mu = \int_M \lim_{k\to\infty} \chi_{\bigcup_{i=k}^{\infty} X_i} \nu d\mu\end{aligned}$$

But this just means

$$\lim_{k\to\infty}\sum_{i=k}^{\infty}\int_{X_i}\nu d\mu = 0$$

Now $K_n = \overline{\bigcup_{i=0}^{n} X_i}$ are compact sets, but since strong causality holds on each K_n, lemma 28 guarantees that γ must exit each K_n never to return. Since M is globally hyperbolic, this means that for each n, there exists some t_n such that $J^-(\gamma(t))\cap\bigcup_{i=0}^{n} X_i = \emptyset$ for all $t \geq t_n$ (so in particular $\int_{J^-(\gamma(t))\cap\bigcup_{i=0}^{n} X_i}\nu d\mu = 0$). Indeed, for a compact K_n, $J^-(\gamma(t)) \cap J^+(K_n)$ is compact as well, so γ must leave it at some $t'_n \geq t_n$.

This finally gives

$$V^-(\gamma(t)) = \int_{J^-(\gamma(t))\cap\bigcup_{i=n+1}^{\infty} X_i}\nu d\mu \leq \int_{\bigcup_{i=n+1}^{\infty} X_i}\nu d\mu = \sum_{i=n+1}^{\infty}\int_{X_i}\nu d\mu,$$

but the right side can be made as small as we wish, provided we choose n large enough. □

Bibliography

[1] Barrett O'Neill, *Semi-Riemannian Geometry With Applications to Relativity*, Academic Press, 1st edition, 1983.

[2] Barrett O'Neill, *The Geometry of Kerr Black Holes*, Dover Publications, 1st edition, 1995.

[3] S. W. Hawking, G. F. R. Ellis, *The Large Scale Structure of Space-Time*, Cambridge University Press, 1st Edition, 1973.

[4] Robert M. Wald, *General Relativity*, University of Chicago Press, 1st edition, 1984.

[5] Yvonne Choquet-Bruhat, *General Relativity and the Einstein Equations*, Oxford University Press, 1st edition, 2009.

[6] Piotr T. Chruściel, *Geometry of Black Holes*, Oxford University Press, 1st edition, 2020

[7] Leonor Godinho, José Natário, *Riemannian Geometry With Applications to Mechanics and Relativity*, Springer, 1st Edition, 2014.

[8] Roger Penrose, *Techniques of Differential Topology in Relativity*, Society for Industrial and Applied Mathematics, 1st edition, 1972.

[9] Hans Ringström, *The Cauchy Problem in General Relativity*, European Mathematical Society, 1st Edition, 2009.

[10] Markus Heusler, *Black Hole Uniqueness Theorems*, Cambridge University Press, 1st Edition, 1996.

[11] John K. Beem, Paul Ehrlich, Kevin Easley, *Global Lorentzian Geometry*, CRC Press, 2nd edition, 1996.

[12] Charles W. Misner, Kip S. Thorne, John Archibald Wheeler, *Gravitation*, Princeton University Press, 1st edition, 1973.

[13] Shoshichi Kobayashi, *Transformation Groups in Differential Geometry*, Springer, 1st edition, 1972.

[14] Shlomo Sternberg, *Curvature in Mathematics and Physics*, Dover Publications, 1st edition, 2012.

[15] John M. Lee, *Introduction to Smooth Manifolds*, Springer, New York, 2nd edition, 2013.

[16] Vladimir Igorevič Arnol'd, *Lectures on Partial Differential Equations*, Springer-Verlag Berlin Heidelberg, 1st edition, 2004.

[17] Walter Rudin, *Real and Complex Analysis*, McGraw Hill, 3rd edition, 1986.

[18] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley, 2nd edition, 1999.

[19] Sean M. Carroll, *Spacetime and Geometry: An Introduction to General Relativity*, Pearson new international edition, 2014.

[20] John Baez, Javier P Muniain, *Gauge Fields, Knots and Gravity*, World Scientific, 1st edition, 1994.

[21] Bernd Schmidt, *Isometry Groups with Surface-Orthogonal Trajectories*, Zeitschrift für Naturforschung A, Volume 22, Issue 9, pp.1351-1355 (1967). https://doi.org/10.1515/zna-1967-0911

[22] Yvonne Choquet-Bruhat, Robert Geroch, *Global aspects of the Cauchy problem in general relativity*, Communications in Mathematical Physics 14, 329–335 (1969). https://doi.org/10.1007/BF01645389

[23] Antonio N. Bernal, Miguel Sánchez, *Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes*, Communications in Mathematical Physics 257, 43–50 (2005). https://arxiv.org/abs/gr-qc/0401112

[24] Antonio N. Bernal, Miguel Sánchez, *Globally hyperbolic spacetimes can be defined as 'causal' instead of 'strongly causal'*, Class. Quantum Grav. 24 745 (2007). https://arxiv.org/abs/gr-qc/0611138

[25] Antonio N. Bernal, Miguel Sánchez, *On smooth Cauchy hypersurfaces and Geroch's splitting theorem*, Communications in Mathematical Physics 243, 461–470 (2003). https://arxiv.org/abs/gr-qc/0306108

[26] R. A. Hounnonkpe, E. Minguzzi, *Globally hyperbolic spacetimes can be defined without the 'causal' condition*, Class. Quantum. Grav. 36 (2019). https://arxiv.org/abs/1908.11701

[27] Mihalis Dafermos, Jonathan Luk, *The interior of dynamical vacuum black holes I: The C^0-stability of the Kerr Cauchy horizon*, (2017) https://arxiv.org/abs/1710.01722

[28] Jan Sbierski, *On the Existence of a Maximal Cauchy Development for the Einstein Equations: a Dezornification*, Annales Henri Poincaré 17, 301–329 (2016). https://arxiv.org/abs/1309.7591

[29] Abhay Ashtekar, R. O. Hansen, *A unified treatment of null and spatial infinity in general relativity. I. Universal structure, asymptotic symmetries, and conserved quantities at spatial infinity*, Journal of Mathematical Physics 19, 1542 (1978). https://doi.org/10.1063/1.523863

[30] Abhay Ashtekar, *Geometry and Physics of Null Infinity*, Surveys in Differential Geometry (2015). https://arxiv.org/abs/1409.1800v2

[31] Robert Geroch, Gary T. Horowtiz, *Asymptotically Simple Does Not Imply Asymptotically Minkowskian*, Physical Review Letters, 40 (4). 203-206 (1978) doi:10.1103/physrevlett.40.203

[32] Robert Geroch, *Domain of Dependence*, J. Math. Phys. 11, 437 (1970); https://doi.org/10.1063/1.1665157

[33] Robert Geroch, *Space-time structure from a global view point*, in *General Relativity and Cosmology* edited by Rainer K. Sachs, Italian Physical Society (1971)

[34] Demetrios Christodoulou, *On the global initial value problem and the issue of singularities*, Class. Quantum Grav. 16 A23 (1999) https://doi.org/10.1088/0264-9381/16/12A/302

[35] Brandon Carter, *Republication of: Black hole equilibrium states Part II. General theory of stationary black hole states*, General Relativity and Gravitation 42, 653–744 (2010) https://doi.org/10.1007/s10714-009-0920-9

[36] Brandon Carter, *Axisymmetric Black Hole Has Only Two Degrees of Freedom*, Phys. Rev. Lett. 26, 331 (1971) https://doi.org/10.1103/PhysRevLett.26.331

[37] Pawel O. Mazur, *Black Hole Uniqueness Theorems*, Proceedings of the 11th International Conference on General Relativity and Gravitation, ed. M. A. H. MacCallum, Cambridge University Press, Cambridge 1987, pp. 130-157 https://arxiv.org/abs/hep-th/0101012v1

[38] Pawel O. Mazur, *Proof of uniqueness of the Kerr-Newman black hole solution*, J. Phys. A: Math. Gen. 15 3173 (1982). https://doi.org/10.1088/0305-4470/15/10/021

[39] D. C. Robinson, *Uniqueness of the Kerr Black Hole*, Phys. Rev. Lett. 34, 905 (1975). https://doi.org/10.1103/PhysRevLett.34.905

[40] Alexandru D. Ionescu & Sergiu Klainerman, *On the uniqueness of smooth, stationary black holes in vacuum*, Inventiones mathematicae 175, Article number: 35 (2009) https://arxiv.org/abs/0711.0040

[41] Spyros Alexakis, Alexandru D. Ionescu, Sergiu Klainerman, *Uniqueness of smooth stationary black holes in vacuum: small perturbations of the Kerr spaces*, Commun.Math.Phys. 299, 89-127 (2010) https://arxiv.org/abs/0904.0982

[42] Spyros Alexakis, Alexandru D. Ionescu, Sergiu Klainerman, *Rigidity of stationary black holes with small angular momentum on the horizon*, Duke Math. J. 163, no. 14, 2603-2615 (2014) https://arxiv.org/abs/1304.0487v2

[43] Piotr T. Chruściel, Robert M. Wald, *On the topology of stationary black holes*, Class. Quant. Grav. 11, L147-L152, (1994) https://arxiv.org/abs/gr-qc/9410004

[44] Piotr T. Chruściel, João Lopes Costa, *On uniqueness of stationary vacuum black holes*, Astérisque 321, 195-265 (2008) https://arxiv.org/abs/0806.0016

[45] Piotr T. Chruściel, João Lopes Costa, Markus Heusler, *Stationary Black Holes: Uniqueness and Beyond*, Living Rev. Relativity 15, 7 (2012) https://arxiv.org/abs/1205.6112

[46] Piotr T. Chruściel, *On rigidity of analytic black holes*, Commun.Math.Phys. 189, 1-7 (1997) https://arxiv.org/abs/gr-qc/9610011

[47] Piotr T. Chruściel, *Uniqueness of stationary, electro-vacuum black holes revisited*, Helv. Phys. Acta 69, 529-552 (1996) https://arxiv.org/abs/gr-qc/9610010

[48] Charles F. Gammie, Jonathan C. McKinney, Gábor Tóth *HARM: A Numerical Scheme for General Relativistic Magnetohydrodynamics*, Astrophys.J. 589, 444-457 (2003) https://arxiv.org/abs/astro-ph/0301509v1

[49] Sean M. Ressler, Alexander Tchekhovskoy, Eliot Quataert, Mani Chandra, Charles F. Gammie, *Electron Thermodynamics in GRMHD Simulations of Low-Luminosity Black Hole Accretion*, Monthly Notices of the Royal Astronomical Society, Volume 454, Issue 2, Pages 1848–1870 (2015) https://arxiv.org/abs/1509.04717v2

[50] R. Penrose and R. M. Floyd *Extraction of Rotational Energy from a Black Hole*, Nature Physical Science 229, 177–179 (1971) https://doi.org/10.1038/physci229177a0

[51] R. Penrose *Zero rest-mass fields including gravitation: asymptotic behaviour*, Proc. Roy. Soc. 284 A, 159 (1965) https://doi.org/10.1098/rspa.1965.0058

[52] Kristin Schleich, Donald M. Witt, *A simple proof of Birkhoff's theorem for cosmological constant*, J. Math. Phys. 51, 112502 (2010). https://arxiv.org/abs/0908.4110

[53] Manuel Gutiérrez and Benjamín Olea *The Rigging Technique for Null Hypersurfaces*, Axioms (2021) https://doi.org/10.3390/axioms10040284

[54] Christian Gérard, Dietrich Häfner & Michał Wrochna *The Unruh state for massless fermions on Kerr spacetime and its Hadamard property*, (2020) https://arxiv.org/abs/2008.10995