The Singularity Theorems

Amath 875, Fall 2009, Yiannis Loizides

1 Introduction

The main goal of this essay is to explain a proof of the following theorem:

Theorem 1.1 *Let (M, g) be a Lorentzian manifold satisfying the following:*

1. *M is globally hyperbolic.*
2. *$Ric(v, v) \geq 0$ for all timelike vectors v, where Ric is the Ricci tensor.*
3. *There is a spacelike "time slice" S such that the expansion scalar $\theta \geq \theta_0 > 0$ on S.*

Then M is singular.

The meanings of the terms in the theorem will be explained in more detail below. Before diving into the details, I will make some general comments.

I've stated the result essentially as a theorem in differential geometry. General relativity together with observational evidence makes it plausible that this theorem applies to our universe. General relativity tells us that the universe is a 4 dimensional Lorentzian manifold. As we discussed in class the assumption of global hyperbolicity is plausible at least as an approximation to the way the universe appears today. The assumption that the expansion is bounded below by some positive constant on S is also reasonable based on the observation that the universe is expanding. Finally, the Einstein field equations tell us that the condition on the Ricci curvature is equivalent to the so-called *strong energy condition* on the energy-momentum tensor:

$$T_{ab}u^a u^b \geq \frac{1}{2}Tu^a u_a$$

for all timelike vectors u. This condition is satisfied by known forms of matter.

The various singularity theorems, of which the above theorem is the simplest example, were first proved in the late 60s and early 70s by Hawking and Penrose. At the time, there was debate over whether singularities were an essential feature of general relativity, or whether they were merely artifacts of the high degree of symmetry in the known exact solutions, and hence would never occur in any real spacetime. The theorems largely resolved this debate by establishing the existence of singularities under a much broader set of conditions.

There are a number of other reasons why the singularity theorems are of interest. For one, results as general as the singularity theorems for a system of non-linear partial differential equations are certainly very useful. For example, they give some support to the hope that the known exact solutions (which have a high degree of symmetry and are thus "atypical") are to some extent representative of "typical" solutions, at least in terms of key qualitative features. The theorems thus give insight into the space of solutions to the field equations, which is of interest to quantum gravity researchers. The theorems are also of interest in connection with quantum gravity as a kind of challenge: this is what any theory of quantum

gravity must overcome if it is to reproduce general relativity at low energies *and* avoid singularities. The theorems are of interest to pure mathematics as well; for example as the author of [N] notes, in Riemannian geometry the Hopf-Rinow theorem neatly characterizes geodesic completeness, but this theorem fails in Lorentzian geometry.

An outline of the proof is as follows: we will first discuss geodesics and the calculus of variations. The key result will be that geodesics fail to maximize the proper time past a conjugate point. Then we will discuss causal structure, global hyperbolicity and we will see some implications for the topology of M. This will lead to a proof that in a globally hyperbolic spacetime M with timeslice S and $p \in I^+(S)$, there is always a timelike geodesic from S to p that maximizes the proper time. Finally we will show that the conditions on the Ricci curvature and the expansion will guarantee either the existence of conjugate points or of singularities. Since proper time-maximizing timelike curves exist in a globally hyperbolic M, there can be no conjugate points and so M must be singular.

After going through a proof of this theorem, we will touch on the question of so-called "quantum singularities", i.e. which singularities remain problematic when point particles travelling along geodesics are replaced with wavefunctions for quantum particles. This is an exciting approach as it offers insights into quantum gravity. A note on sources: the main sources I will follow in this paper are [N] and [O] for the proof, and [HM] for the introduction to quantum singularities. In almost all the proofs I have filled in details or made changes (although much of the time I don't explicitly say where I am doing so), or in some cases written my own proof. I also used [HE] although far less than the other texts.

2 Calculus of Variations

In this section, we'll consider curves that begin at some (spacelike) surface S and end at a point q in M. The goal will be to understand under what conditions such curves maximize proper time. The natural tool to investigate this is the calculus of variations. Several of the proofs in this section are either fairly standard or are quite typical of the calculus of variations, so we will not give the proofs in full (although I've provided references). Also, we don't want to get too bogged down because some of these results get fairly technical and the central results of interest don't come until the later sections.

Throughout we let $\alpha : [a, b] \to M$ be a piecewise smooth timelike curve with breaks at u_1,...,u_k and we let $x : [a, b] \times (-\delta, \delta) \to M$ be a piecewise smooth variation of α, that is to say $x(u, 0) = \alpha(u)$ for all $u \in [a, b]$, and WLOG x has the same breaks as α (since we can always add more to α). We can suppose α has constant speed $c = \sqrt{|g(\alpha', \alpha')|}$ (not the speed of light), except at the breaks. We would like to compare the length (alternatively proper time) of α with that of $x(u, v_0)$ for v_0 close to zero. To this end, we define

$$L_x(v) := \int_a^b du \sqrt{|g(\partial_u x(u, v), \partial_u x(u, v))|}. \tag{1}$$

We will call the functional $L_x = L[x]$ the *length functional*. We'll now state the first variation formula. When x is a variation, we let $V(u, v) = \partial_v x(u, v)$ which we call the *variation vector field*. We assume that V is continuous and that its covariant derivative in the v direction

is continuous, although covariant derivatives of V in the u direction may only be piece-wise continuous. We'll write $V(0) = V(u, 0)$ for convenience.

Theorem 2.1 *With notation as above, the first variation formula is*

$$L'_x(0) = \frac{1}{c}\int_a^b g(\nabla_u \alpha', V)du + \frac{1}{c}\sum_1^k g\big(\Delta\alpha'(u_i), V(u_i)\big) - \frac{1}{c}g(\alpha', V)\Big|_a^b$$

where $\Delta\alpha'(u_i)$ denotes the change in the tangent to α at the break.

Proof We won't prove this in detail since the result is standard (see for example proposition 10.2 in [O]). The proof is a typical and fairly straightforward calculus of variations-type proof: differentiate under the integral sign in equation (1), perform a few small manipulations using compatibility of the connection and the vanishing of torsion, then integrate by parts. ■

It is well known from differential geometry that smooth geodesics between a pair of points p, q are precisely the critical points of the length functional when the variations are required to fix p, q. More precisely, α is a geodesic from p to q iff for every variation x such that $x(a, v) = p$, $x(b, v) = q$ we have $L'_x(0) = 0$. For geodesics then, we are really more interested in $L''(0)$. Let α now be a piecewise smooth timelike geodesic, and we can suppose that it has constant speed c (except at breaks). We'll write $\nabla_u Y$ for the covariant derivative along the vector field $\frac{\partial}{\partial u} = \partial_u = \partial_u x$, and likewise for v. If Y is some vector field along α, we'll denote by $Y^\perp$ and Y^T the components of Y orthogonal to and parallel to α' respectively. If α is a geodesic and Y is a vector field along α then $Y = Y^\perp + f\alpha'$ for some smooth function f. Using compatibility of the metric,

$$g(\nabla_u(Y^\perp), \alpha') = \partial_u g(Y^\perp, \alpha') - g(Y^\perp, \nabla_u \alpha')$$

and both terms on the right vanish since $Y^\perp$ is orthogonal to α and since α is a geodesic. Again using the fact that α is a geodesic,

$$\nabla_u(Y^T) = \nabla_u(f\alpha') = (\partial_u f)\alpha'$$

and it follows that ∇_u preserves the decomposition $Y = Y^\perp + Y^T$. Thus we can write $\nabla_u Y^T = (\nabla_u Y)^T = \nabla_u(Y^T)$ without ambiguity. Let's now state the formula for the second variation.

Theorem 2.2 *Let α be a geodesic. Then with notation as above the second variation is given by*

$$\begin{aligned} L''_x(0) =& \frac{1}{c}\int_a^b \Big[g(\nabla_u^2 V^\perp + R(V^\perp, \alpha')\alpha', V^\perp)\Big]du + \frac{1}{c}\sum_1^k g\Big(\Delta(\nabla_u V^\perp)(u_i), V^\perp(u_i)\Big) \\ &- \Big[\frac{1}{c}g(\nabla_u V, V) + \frac{1}{c}g(\alpha', \nabla_v V)\Big]_{u=a}^{u=b} \end{aligned}$$

In this expression, R is the Riemann curvature tensor and all terms involving V are to be evaluated at $v = 0$.

Proof Again we won't prove this in detail here as the result is standard (see for example proposition 10.4 in [O]). The proof is once again a typical calculus of variations proof: differentiate equation (1) twice under the integral sign (use proposition 2.1 to speed this up). The Ricci identity allows the order of covariant derivatives to be changed at the cost of producing the curvature term. Compatibility of the metric, the vanishing of the torsion and the fact that α is a geodesic allow the expression to be simplified somewhat, and then integration be parts yields the expression above. ■

Proposition 2.3 *Let V be a smooth vector field along the curve $\alpha : [a, b] \to M$ such that $V(a) = V(b) = 0$, then there is a fixed endpoint variation of α with V as its variation vector field.*

Proof Since $V(u) \in T_{\alpha(u)}M$, we can put $x(u, v) := \exp_{\alpha(u)}(vV(u))$. Then $x(u, 0) = \exp_{\alpha(u)}(0) = \alpha(u)$, $x(a, v) = \exp_{\alpha(a)}(0) = \alpha(a)$, $x(b, v) = \exp_{\alpha(b)}(0) = \alpha(b)$, and $\partial_v x(u, 0) = d(\exp_{\alpha(u)})V(u) = V(u)$. So x is the required fixed endpoint variation. ■

We'll now define a special kind of vector field along a geodesic which can be thought of as "infinitesimal" measures of how nearby geodesics are behaving.

Definition A *Jacobi field* along the geodesic α is a vector field Y satisfying the geodesic deviation equation

$$\nabla_u^2 Y + R(Y, \alpha')\alpha' = 0. \tag{2}$$

Proposition 2.4 *If x is a variation of geodesics (i.e. $x(u, v_0)$ is a geodesic for each $v_0 \in (-\delta, \delta)$) then its variation vector field is a Jacobi field.*

Proof We can use the fact that the connection is torsion free, $[\partial_u, \partial_v] = 0$, and the Ricci identity to write

$$\begin{aligned}
\nabla_u^2(\partial_v x) &= \nabla_u(\nabla_{\partial_u x}\partial_v x) \\
&= \nabla_u(\nabla_{\partial_v x}\partial_u x) \\
&= \nabla_u\nabla_v(\partial_u x) \\
&= \nabla_v\nabla_u(\partial_u x) + R(\partial_u x, \partial_v x)\partial_u x \\
&= -R(\partial_v x, \partial_u x)\partial_u x
\end{aligned}$$

where we've used the fact that $\nabla_u(\partial_u x) = 0$ since x is a variation of geodesics. ■

The situation we will be interested in is the case of timelike geodesics from a spacelike submanifold S of M to a point $q \in M$, so amongst other things, we need to identify which Jacobi fields correspond to variations of geodesics that start on S and end at q. This will take a little work, and will culminate in proposition 2.8.

We define $\Omega(S, q)$ to be the set of all piecewise smooth curves $\alpha : [0, b] \to M$ that start on S and end at q. Any smooth variation in $\Omega(S, q)$ gives rise to variation vector field V with $V(0)$ tangent to S, and $V(b) = 0$. From now on a variation x will refer to a variation in $\Omega(S, q)$.

Proposition 2.5 *Let $\alpha \in \Omega(S, q)$ be timelike, then $L'_x(0) = 0$ for all variations x iff α is a geodesic orthogonal to S.*

Proof Fixed endpoint variations guarantee that α must be a geodesic. Then in the equation for the first variation 2.1, the first term involving the integral must vanish, so we have

$$L'_x(0) = -\frac{1}{c} g(\alpha', V)\Big|_a^b.$$

But $V(b) = 0$, so this vanishes iff $\alpha'(a)$ is orthogonal to $V(a)$. Since any vector $V(a) \in T_{\alpha(a)}S$ can be obtained by a suitable choice of x, it follows that α is orthogonal to S. ■

We review a few definitions: for a vector field V on M and a point $p \in S$, let $norV_p$ and $tanV_p$ denote the components of V_p orthogonal to S and tangent to S respectively. The set of vectors in M orthogonal to S forms a bundle NS called the *normal bundle*. For a bundle F, $\Gamma(F)$ denotes the set of sections of that bundle.

Definition The map $II : \Gamma(TS) \times \Gamma(TS) \to \Gamma(NS)$ given by

$$II(V, W) = nor\nabla_V W$$

is called the *second fundamental form* of $S \subset M$.

Projecting the Levi-Civita connection on M onto S yields the Levi-Civita connection on S, and the second fundamental form gives the difference between the two connections. The connection on M also induces a connection on the normal bundle NS.

Definition The map $\nabla^\perp : \Gamma(TS) \times \Gamma(NS) \to \Gamma(NS)$ given by

$$\nabla^\perp_V W = nor\nabla_V W$$

is called the *normal connection* on NS.

The normal connection allows us to "parallel transport" vectors that are normal to S so that they stay normal to S, i.e. transport a vector W normal to S along a curve β in S such that $\nabla^\perp_{\beta'} W = 0$. Exactly as in the usual case, this is a linear ODE, so it gives a well-defined *normal translation* of W along β.

Definition We define a map $\tilde{II} : \Gamma(TS) \times \Gamma(NS) \to \Gamma(NS)$ given by

$$\tilde{II}(V, W) = tan\nabla_V W$$

so that

$$\nabla_V W = \nabla^\perp_V W + \tilde{II}(V, W)$$

Proposition 2.6 *Let V be tangent to S, and Z, W normal to S, then*

$$g(\tilde{II}(V, Z), W) = -g(II(V, W), Z)$$

Proof If we differentiate the equation $g(Z,W)=0$ and use compatibility of the connection we get $g(\nabla_V Z, W) = -g(Z, \nabla_V W)$. Now W is tangent to S so we can replace $\nabla_V Z$ by its tangential component, which is $\tilde{II}(V,Z)$. Similarly, Z is normal to S, so we can replace $\nabla_V W$ by its normal component, which is $II(V,W)$. ■

We also define an analog of the exponential map.

Definition The *normal exponential map* $\exp^\perp : NS \to M$ is given by

$$\exp^\perp(v) = \gamma_v(1)$$

where γ_v is the unique geodesic starting at the base point of v on S and having initial direction vector v.

As an aside, now is an appropriate time to mention the following result (sometimes called the tubular neighbourhood theorem) from differential geometry, which we shall use in section 4:

Theorem 2.7 *Let S be a submanifold (embedded is needed), then there is some neighbourhood V of $S \subset M$ and some neighbourhood of the zero section $U \subset NS$ such that $\exp^\perp$ is a diffeomorphism from U onto V.*

This result is fairly standard and the proof is rather long and technical so we won't prove it here (see for example proposition 7.26 in [O]).

Following [O],

Proposition 2.8 *A Jacobi field V along a geodesic α which is normal to S is the variation vector field of a variation x of α through geodesics normal to S iff $V(0)$ is tangent to S and $tan\nabla_u V(0) = \tilde{II}(V(0), \alpha'(0))$.*

Proof First suppose x is such a variation. Then we've already proven that its variation vector field is a Jacobi field. Since all the geodesics $u \mapsto x(u,v)$ start on S, $v \mapsto x(0,v)$ is a curve on S, and hence $\partial_v x(0,0)$ is tangent to S. Using the vanishing of torsion

$$\nabla_u \partial_v x(0,0) = \nabla_v \partial_u x(0,0)$$

and taking the tangential component of both sides gives the remaining condition.

For the converse, let V be a Jacobi field along α satisfying the conditions in the statement of the proposition. Since $V(0)$ is tangent to S, let $\beta(v)$ be a curve in S with initial tangent vector $V(0)$. Let $A(v)$ and $B(v)$ be the normal translations of $\alpha'(0)$ and $nor\nabla_u V(0)$ along β respectively, and put

$$Z(v) = A(v) + vB(v).$$

Then we note two properties of Z. First $Z(0) = A(0) = \alpha'(0)$. The second property is $\nabla_v Z(0) = \nabla_u V(0)$, which follows because

$$\nabla_v Z(0) = \nabla_v A(0) + B(0) + 0 \cdot \nabla_v B(0) = \nabla_v A(0) + B(0)$$

Now A is the normal translation of $\alpha'(0)$, so the normal component of its covariant derivative vanishes. Also $B(0) = nor\nabla_u V(0)$, so we have

$$\nabla_v Z(0) = \tilde{II}(\beta'(0), A(0)) + nor\nabla_u V(0) = \tilde{II}(V(0), \alpha'(0)) + nor\nabla_u V(0)$$

and putting in $tan\nabla_u V(0) = \tilde{II}(V(0), \alpha'(0))$ gives the desired result.

Now, define a variation

$$x(u, v) = \exp^{\perp}(uZ(v)).$$

This is clearly a variation through normal geodesics by definition of the normal exponential map. We have $x(u, 0) = \exp^{\perp}(uZ(0)) = \alpha(u)$ by definition of the normal exponential map. Let Y be the Jacobi field corresponding to the variation x. We want to show that $V = Y$, for then we will have constructed a variation corresponding to V. Recall that Jacobi fields satisfy a second order equation, so they are completely specified once two initial conditions are given. So it suffices for us to check that $V(0) = Y(0)$ and $\nabla_u V(0) = \nabla_u Y(0)$. The first follows because Y is tangent to the curve $v \mapsto x(0, v) = \exp^{\perp}(0 \cdot Z(v)) = \beta(v)$ (the vector $0 \cdot Z(v)$ is always zero, but the vector space (fibre) changes as v changes), so $Y(0) = \beta'(0) = V(0)$. For the second condition we use the vanishing of the torsion together with the fact that $d\exp^{\perp}$ is the identity map on vectors normal to S whose base point is on S,

$$\begin{aligned}
\nabla_u Y(0) = \nabla_u \partial_v x(0,0) &= \nabla_v \partial_u x(0,0) \\
&= \nabla_v d\exp^{\perp}(Z(v))\Big|_{v=0} \\
&= \nabla_v Z(0) \\
&= \nabla_u V(0)
\end{aligned}$$

where the last step follows by the second property above. ■

Recall we are interested in curves from a surface S to a point. In this case, let's write the second variation formula, theorem 2.2, in a slightly different form. We note that since $V(b) = 0$,

$$\begin{aligned}
g(\alpha', \nabla_v V)\big|_0^b &= -g(\alpha'(0), \nabla_v V(0)) \\
&= -g(\alpha'(0), nor\nabla_v V(0))
\end{aligned}$$

since $\alpha'(0)$ is orthogonal to S. Recalling the definition of the second fundamental form, we have

$$g(\alpha', \nabla_v V)\big|_0^b = -g(\alpha'(0), II(V(0), V(0)))$$

Definition We define the *index form* I_α of a timelike geodesic α orthogonal to S to be the unique symmetric bilinear form on the set of variation vector fields such that $I_\alpha(V, V) = L_x''(0)$, where x is a variation in $\Omega(S, q)$ with variation vector field V.

Proposition 2.9 *Let $\alpha : [0,1] \to M$ be a unit speed timelike geodesic orthogonal to S as above. Then we have the following explicit formula for the index form:*

$$I_\alpha(V,W) = \int_0^1 g\Big(\nabla_u^2 V + R(V,\alpha')\alpha', W\Big)du + g\Big(V'(0) - \tilde{II}(V(0),\alpha'(0)), W(0)\Big) + \sum_1^k g\Big(\Delta(\nabla_u V(u_i)), W(u_i)\Big)$$

where as above the terms involving V are evaluated at $v = 0$.

Proof It's easy to see that the index form is bilinear, and in fact vanishes if either of V or W is parallel to α. Thus in considering the index form, it suffices to consider variation vector fields that are orthogonal to α. It's easy to see that the formula above indeed satisfies $I_\alpha(V,V) = L_x''(0)$ from theorem 2.2 and the facts we've given about the second fundamental form, etc. It's a little more difficult to see that I is symmetric (the integration by parts referred to in the proof of theorem 2.2 makes the symmetry less obvious; see for example proposition 10.4 in [O]). Nevertheless, the symmetry can be shown using the expression above by "undoing" the integration by parts, i.e. integrate by parts again to change the second covariant derivative to a single covariant derivative. ■

Definition Let $\alpha : [0,b] \to M$ be a geodesic from S to q. We'll say that q is a *conjugate point* if there exists a non-zero Jacobi field J along α such that $J(0)$ is tangent to S, $J(b) = 0$ and the technical condition $tan(\nabla_u J)(0) = \tilde{II}(J(0),\alpha'(0))$ holds (so that the hypotheses of proposition 2.8 are satisfied).

We should think of a conjugate point as a point where "nearby" geodesics converge (or "almost" converge).

Proposition 2.10 *Let $\alpha : [0,1] \to M$ be a timelike geodesic normal to S, with $\alpha(0) = p \in S$. Then the following are equivalent:*

1. *$\alpha(1)$ is a conjugate point of S along α*
2. *There is a non-trivial variation S of α through geodesics normal to S and for which $\partial_v x(1,0) = 0$.*
3. *The map $\exp^\perp$ is singular at $\alpha'(0)$.*

Proof That (1) is equivalent to (2) follows from proposition 2.8. For later results we shall only need that (2) is implied by (3), so we restrict ourselves to proving that. Suppose $d\exp^\perp(y_{\alpha'(0)}) = 0$ where $y \in T_{\alpha'(0)}NP$. We have two cases:

Case 1 $y_{\alpha'(0)}$ is tangent to the fibre, then since the fibre is a vector space, we can identify $y_{\alpha'(0)}$ (by translating to the origin of the fibre) with a vector y in T_pM that is perpendicular to S. Then using the ordinary exponential map, we define a variation

$$x(u,v) := \exp_p(u(\alpha'(0) + vy)).$$

Then clearly $x(u,0) = \exp_p(u\alpha'(0)) = \alpha(u)$ since α is the unique geodesic starting at p with initial velocity $\alpha'(0)$. Also, this is clearly a variation through geodesics perpendicular to S since both $\alpha'(0)$ and y are perpendicular to S and hence so is any linear combination of them. And $\partial_v x(1,0) = (d\exp_p)(y_{\alpha'(0)}) = 0$ as required. Finally, we note that this cannot be a trivial variation through geodesics for then $y \neq 0$ would be a scalar multiple of $\alpha'(0)$ and then the fact that exp is singular would mean that the tangent to α vanishes, which is impossible since α is a geodesic.

Case 2 $y_{\alpha'(0)}$ is not tangent to the fibre so that $d\pi(y_{\alpha'(0)}) \neq 0$, where $\pi : NP \to P$ is the projection map. Let $Z(v)$ be a curve in NP with initial velocity $y_{\alpha'(0)}$. Then in particular $Z(0) = \alpha'(0)$. We define the variation

$$x(u,v) = \exp^{\perp}(uZ(v)).$$

Now $x(u,0) = \exp^{\perp}(uZ(0)) = \exp^{\perp}(u\alpha'(0)) = \alpha(u)$ and $\partial_v x(1,0) = d\exp^{\perp}(Z'(0)) = d\exp^{\perp}(y_{\alpha'(0)}) = 0$ as required. Clearly this is a variation through normal geodesics by definition of the normal exponential map. Finally, at $u = 0$ we have $x(0,v) = \exp^{\perp}(0Z(v))$ which is a curve along S (the *base point* of the zero vector changes as v changes). Its tangent vector at $v = 0$ is the tangent vector of the projection of the curve $Z(v)$ onto S, i.e. $d\pi(y_{\alpha'(0)})$, and since we're assuming this is non-zero, this shows that the variation is non-trivial. ■

Remark Case 1 in the above proof actually isn't relevant for the situation we're considering, since for us S will be 3 dimensional and embedded in 4 dimensional spacetime. So there can be no non-trivial variation through geodesics normal to S that does not change the starting point p on S.

Theorem 2.11 *Let $\alpha \in \Omega(S,q)$ be a timelike geodesic normal to S. Then if there is a conjugate point along α before q, I_α is not semidefinite. In particular, by the definition of the index form, we can find a variation α_v such that $L''(0) > 0$ (while $L'(0) = 0$ since the variation is through normal geodesics). Expanding $L(v)$ via Taylor's theorem gives $L(v) = L(0) + L''(0)v^2/2 + O(v^3)$, so for v sufficiently small, we get a strictly larger length $L(v)$. Summarizing, α fails to maximize proper time past a conjugate point.*

Proof For convenience, we recall the formula for the index form:

$$I_\alpha(V,W) = \int_0^1 g\Big(\nabla_u^2 V + R(V,\alpha')\alpha', W\Big)du + g\Big(V'(0) - \tilde{II}(V(0),\alpha'(0)), W(0)\Big) + \sum_1^k g\Big(\Delta(\nabla_u V(u_i)), W(u_i)\Big)$$

Suppose $\alpha(r)$ is a conjugate point along α between $p = \alpha(0)$ and $q = \alpha(1)$. Let V be a non-trivial Jacobi field along α corresponding to this conjugate point. We recall that this means that:

1. $\nabla_u^2 V + R(V,\alpha')\alpha' = 0$

2. $V(0)$ is tangent to S
3. $tan(\nabla_u V)(0) = \tilde{II}(V(0), \alpha'(0))$
4. $V(r) = 0$

We define a new V to equal the old V between 0 and r, and to equal 0 between r and 1. This is continuous since $V(r) = 0$, but it can't be smooth at r, because if $\nabla_u V(r^-) = 0 = V(r)$, then V would have to be the 0 Jacobi field (a Jacobi field is specified uniquely by two initial conditions), contradicting non-triviality. Since $\nabla_u V(r^+) = 0$, it follows that $\Delta(\nabla_u V(r)) = -\nabla_u V(r^-) \neq 0$. So we define a second vector field W to be any smooth vector field along α tangent to S at 0 and satisfying the technical "boundary" condition (item (3) in the list just above), and such that $W(r) = \pm\Delta(\nabla_u V(r))$ (we can choose the sign later). Then $V + \epsilon W$ will satisfy the conditions of proposition 2.8. We consider

$$I_\alpha(V + \epsilon W, V + \epsilon W) = I_\alpha(V, V) + 2\epsilon I_\alpha(V, W) + O(\epsilon^2) \tag{3}$$

Since V is a Jacobi field, the first property in the list above gives us that the integral term vanishes for both $I_\alpha(V, V)$ and $I_\alpha(V, W)$ (we see at last what a Jacobi field can do for us!). Even though $\nabla_u V(r)$ is non-zero, since $V(r) = 0$ the Δ term in $I_\alpha(V, V)$ will vanish anyway. Now consider

$$\begin{aligned}\nabla_u V(0) - \tilde{II}(V(0), \alpha'(0)) &= \nabla_v \partial_u x(0,0) - \tilde{II}(V(0), \partial_u x(0,0)) \\ &= \nabla_v^\perp \partial_u x(0,0)\end{aligned}$$

which is orthogonal to S by definition of the normal connection. Hence contracting it with either $V(0)$ or $W(0)$, both of which are tangent to S, will yield 0. This shows that the "boundary" terms (those involving $\tilde{II}$) vanish for both $I_\alpha(V, V)$ and $I_\alpha(V, W)$. Using the fact that $W(r) = \Delta(\nabla_u V(r))$, equation (3) thus reduces to

$$\begin{aligned}I_\alpha(V + \epsilon W, V + \epsilon W) &= 2\epsilon I_\alpha(V, W) + O(\epsilon^2) \\ &= \pm 2\epsilon \left|\Delta(\nabla_u V(r))\right|^2 + O(\epsilon^2)\end{aligned}$$

It follows that for ϵ sufficiently small and the correct choice of sign for W, we can make $I_\alpha(V + \epsilon W, V + \epsilon W)$ either positive or negative, so I_α is not semidefinite. ■

3 Causal Structure

Recall from class we defined $I^+(p)$ to be the set of all points reachable from p by a future-directed timelike curve, and similarly for $J^+(p)$ except with "timelike" replaced by "causal". The goal of the next several results will be to establish some of the properties of these sets.

First we recall that every point $p \in M$ has a *convex normal neighbourhood*, i.e. a neighbourhood N on which $\exp_q$ is a diffeomorphism onto N for each $q \in N$, and there is a unique geodesic between each pair of points in the neighbourhood. This is a standard result from differential geometry that we will not prove here. Since neighbourhoods of this type always

exist, we will use them often and usually just refer to them as "normal neighbourhoods" or sometimes "normal geodesic neighbourhoods".

We know that in Minkowski space a future-directed timelike curve passing through the origin cannot escape the future time cone at the origin. Intuitively this says that nothing travels faster than the speed of light. An analog of this should and does hold in general relativity although it is a little trickier to prove. First we recall Gauss' lemma:

Theorem 3.1 *Let $p \in M$ and $0 \neq x \in T_pM$. We identify $T_x(T_pM)$ with T_pM. Then for any $w \in M$*

$$g(d\exp(x), d\exp(w)) = g(x, w)$$

As this is a standard result in differential geometry, we won't prove this here. It follows in particular from this that inside a normal geodesic neighbourhood, given an outgoing timelike geodesic α, it will be orthogonal to the "spheres" (level sets of the distance function, which is well-defined in a normal geodesic neighbourhood) that it crosses. Of course the gradient of the distance function is orthogonal to the level sets as well, and so the two vector fields will be parallel along a given outgoing geodesic.

Proposition 3.2 *Let $\beta : [0, b] \to T_pM$ be piecewise smooth and such that $\alpha = \exp_p \circ \beta$ is timelike. Then β stays in the same timecone in T_pM.*

Proof Since $\alpha'(0)$ is timelike, and since $d\exp$ is the identity map at the origin, $\beta'(0)$ is also timelike. Therefore it follows from Taylor's theorem that $\beta(t)$ is in the same timecone for $0 < t < \epsilon$ for some sufficiently small ϵ. Let r denote the distance function on T_pM given by the restriction of g to T_pM, and write $r(t) = r(\beta(t))$. Then by the remark before the proposition $grad(r)(x)$ is parallel to x, and in fact it is easy to see that $grad(r) = x/|x|$. Then by the definition of gradient, we have

$$\frac{d}{dt}(r \circ \beta) = g(\beta', x/|x|)$$

By Gauss' lemma, the right side equals $g(\alpha', d\exp(x/|x|))$. Now since β stays in the timecone for ϵ sufficiently small, $g(\beta', x/|x|)$ is negative for ϵ sufficiently small, and hence so is $\frac{d}{dt}(r \circ \beta)$. Since $r(t = 0) = 0$, this means that r is initially negative and decreasing along β. In order to leave the timecone, r must become nonnegative, which means that r must increase, i.e. $r'(t)$ must become positive.

But by above we have $r'(t) = g(\alpha', d\exp(x/|x|))$. We know that α is timelike. $d\exp(x)$ is timelike so long as $x = \beta(t)$ is timelike (by Gauss' lemma), which happens iff $r(x) < 0$. Hence $g(\alpha', d\exp(x)) < 0$ so long as $r(x) < 0$. What all of this shows is that $r(t)$ can't possibly increase until after $\beta(t)$ has left the timecone, i.e. until after $r(t)$ has become nonnegative. This is of course impossible, and so β must remain in a single timecone. ■

The author of [O] says that with slight modifications the proof above works also for causal curves and the causal cone, but I have not been able to make this work as of yet (I don't believe that Taylor's theorem can be used to establish the key fact at the beginning that β is initially in one timecone, at least not the straightforward application of Taylor's theorem as in the theorem above). Instead we will get the result above for causal curves from a more complicated result that the author of [O] proves much later in the book.

Lemma 3.3 *Let α be a causal curve, and let x be a variation of α with variation vector field V. Suppose $g(\nabla_u V, \alpha') < 0$. Then for all sufficiently small v_0, the deformed curve $x(u, v_0)$ is timelike.*

Proof This is a straightforward calculation. Since α is causal, $g(\partial_u x(u,0), \partial_u x(u,0)) \leq 0$. Using compatibility of the connection and the fact that the torsion vanishes we have

$$\partial_v\big|_{v=0} g(\partial_u x, \partial_u x) = 2g(\nabla_v \partial_u x, \alpha') = 2g(\nabla_u \partial_v x, \alpha') = 2g(\nabla_u V, \alpha') < 0$$

and the result follows from Taylor's theorem. ■

With this result in hand, our strategy to prove the analog of proposition 3.2 for causal curves will be as follows: null geodesics leaving a point remain in the same causal cone associated to that point (e.g. by definition of the exponential map). For any curve that is not a null geodesic, we will use the lemma to find a small fixed endpoint variation x that makes the curve timelike.

Theorem 3.4 *Let $\alpha : [0,1] \to M$ be a causal curve from p to q that is not a null geodesic. Then there is a timelike curve from p to q that is arbitrarily close to α.*

Remark "Arbitrarily close" is in the sense above, i.e. we can find a fixed endpoint variation x of α such that $x(u, v_0)$ is timelike for all $v_0 > 0$ sufficiently small. If exp is non-singular on α, then we can lift x to T_pM (for sufficiently small v this follows from the inverse function theorem), and then since compact sets are involved, we'll have uniform convergence with respect to the usual Euclidean metric.

Proof First we treat the case in which α is timelike at at least one point. We'll suppose that $\alpha'(1)$ is timelike; any other case is handled in virtually the same way. Since α' is continuous on at least an interval ending at 1 (at most finitely many breaks), it follows that $g(\alpha', \alpha') < -\epsilon$ for $t \in [1-\delta, 1]$. Parallel transport $\alpha'(1)$ along α to obtain a vector field W along α. Since parallel transport preserves "timelike", "null" and "spacelike", W is a timelike vector field, and since α is causal, we have $g(\alpha', W) < 0$ everywhere along the curve. Choose a bump function $f : [0,1] \to \mathbb{R}$ with $f(0) = f(1) = 0$ and $f' > 0$ on $[0, 1-\delta]$, and let $V = fW$ along α. Then since W was obtained by parallel transport, $\nabla_u W = 0$ and so we have

$$g(\nabla_u V, \alpha') = g((\partial_u f)W + f\nabla_u W, \alpha') = (\partial_u f) g(W, \alpha')$$

Thus $g(\nabla_u V, \alpha') < 0$ on $[0, 1-\delta]$. By proposition 2.3 we can find a fixed endpoint variation x having V as its variation vector field, and then the lemma above tells us that $x(u, v_0)$ will be timelike on $[0, 1-\delta]$ for all sufficiently small v_0. Will the curve still be timelike on $[1-\delta, 1]$? Well by Taylor's theorem at any point $u \in [1-\delta, 1]$ we have

$$g(\partial_u x(v_0), \partial_u x(v_0)) = g(\alpha', \alpha') + 2v_0 g(\alpha', V) + O(v_0^2) M(u)$$

where $M(u)$ is some continuous function of u. We know that on $[1-\delta, 1]$, the zeroth order and first order terms in v_0 will be negative. Since $[1-\delta, 1]$ is compact, $M(u)$ is bounded and so by choosing v_0 sufficiently small, we can ensure that $g(\partial_u x(v_0), \partial_u x(v_0))$ is negative.

Slight modifications of this method (which we will skip) work for null curves that are not null geodesics, and for causal curves that have breaks. See [O]. ■

We can now show that the analog of proposition 3.2 holds for causal curves.

Proposition 3.5 *Let $\beta : [0, b] \to T_pM$ be piecewise smooth and such that $\alpha = \exp_p \circ \beta$ is a causal curve from p to q. Suppose also that* exp *is non-singular along β. Then β stays in the same causal cone in T_pM.*

Remark Probably the assumption that exp be nonsingular along β is not necessary, but it will not hamper us when we apply this result in the future.

Proof We know that this will be true for a null geodesic (essentially by definition of the exponential map). For a causal curve that is not a null geodesic, the proposition above tells us that we can find a fixed endpoint variation x of α such that $x(u, v_0)$ is timelike for all sufficiently small $v_0 > 0$. By the inverse function theorem, exp will be non-singular in some neighbourhood of α, and so we can lift $x(u, v)$ to a curve β_v in T_pM for all sufficiently small $v \geq 0$. Since $x(u, v)$ is timelike for small $v > 0$, we can apply proposition 3.2 to get that β_v lies in a single timecone in T_pM. Then since the causal cone in T_pM is the closure of the time cone in T_pM (this is just Minkowski space), and since the curves β_v tend to β pointwise as $v \to 0$, it follows that β stays in a single causal cone in T_pM. ■

With these results, we can now prove the following quite easily:

Proposition 3.6 *Let N be a normal convex neighbourhood of $p \in M$. We can choose N sufficiently small so that its closure is also contained in a normal convex neighbourhood. For convenience, we put $I^+(p, N) := I^+(p) \cap N$, $J^+(p, N) := J^+(p) \cap N$, etc. Then the following are true:*

1. *$q \in I^+(p, N)$ iff there is a future-directed timelike geodesic from p to q.*
2. *$J^+(p, N) = \overline{I^+(p, N)}$*
3. *$q \in J^+(p, N)$ iff there is a future-directed causal geodesic from p to q.*

Proof (1) Since exp is a diffeomorphism onto N, we can lift any timelike curve α from p to q back to T_pM, and by proposition 3.2, this curve stays in the same (future) timecone in T_pM. Thus in particular $\exp_p^{-1}(q)$ is in that timecone, and the straight line from 0 to $\exp_p^{-1}(q)$ in T_pM gets mapped by the exponential map to a timelike geodesic from p to q. The converse is true by definition of $I^+(p, N)$.

(2) Let $\alpha : [0, 1] \to M$ be a causal curve from p to q. If α is not a null geodesic, then a result above showed we could perturb α to get a timelike curve, so that $q \in I^+(p, N)$. If α is a null geodesic, then $\alpha(u) = \exp_p(uX)$ for some null vector X in T_pM. Then we can take a sequence of timelike vectors X_i in $\exp_p^{-1}(N)$ tending to X in T_pM (with respect to the usual topology, i.e. the Euclidean metric), and since $\exp_p$ is a diffeomorphism on N, the points $\exp_p(X_i)$ tend to q. This shows that $\overline{I^+(p, N)} \supset J^+(p, N)$. For the converse, suppose q_i is a sequence in $I^+(p, N)$ converging to some point $q \in \overline{I^+(p, N)} \subset \overline{N}$. Since $\exp_p$ is a diffeomorphism on some neighbourhood of $\overline{N}$, the points $\exp_p^{-1}(q_i)$ converge to a point $\exp_p^{-1}(q)$ in T_pM. Moreover, since the lift of any timelike curve through p to T_pM stays

inside a single timecone, the points $\exp_p^{-1}(q_i)$ are all in the future timecone of T_pM, and thus converge to a point in its closure: the causal cone. Thus $\exp_p^{-1}(q) =: X$ is in the causal cone, and the curve $\exp_p(tX)$ is a causal curve from p to q. This proves $\overline{I^+(p,N)} \subset J^+(p,N)$.

(3) Given a causal curve from p to q: from (2) we have that q is the limit of a sequence of points in $I^+(p,N)$. The causal curve we found in (2) from p to an arbitrary limit of points in $I^+(p,N)$ was in fact a causal geodesic. So we're done. ■

Definition Let $p, q \in M$, we define

1. $p \ll q$ if $q \in I^+(p)$
2. $p \leq q$ if $q \in J^+(p)$

Proposition 3.7 *Let $p \in M$, then*

$$I^+(p) = I^+(I^+(p)) = I^+(J^+(p)) = J^+(I^+(p)) \subset J^+(J^+(p)) = J^+(p)$$

Proof (1) To prove $I^+(p) = I^+(I^+(p))$, suppose $p \ll q \ll r$, then there are timelike curves from p to q and from q to r. Following these two timelike curves in succession gives a timelike curve from p to r. For the reverse inclusion, suppose $p \ll r$ and let α be a timelike curve from p to r. Choose any point q strictly in between p and r along α, then clearly $p \ll q \ll r$.

(2) To prove $I^+(p) = I^+(J^+(p))$, suppose $p \leq q \ll r$. Let α be the causal curve from p to r that follows a causal curve from p to q and then follows a timelike curve from q to r. Since this isn't a null geodesic, theorem 3.4 allows us to deform α slightly to yield a timelike curve. The reverse inclusion is obvious.

(3) The argument that $I^+(p) = J^+(I^+(p))$ is virtually the same as (2), and it is clear that $J^+(I^+(p)) \subset J^+(J^+(p))$.

(4) This leaves $J^+(J^+(p)) = J^+(p)$. Since $p \in J^+(p)$, one inclusion follows. Suppose $p \leq q \leq r$, then there are causal curves from p to q and from q to r. Following these two causal curves in succession gives a causal curve from p to r. ■

Proposition 3.8 *$I^+(p)$ is open.*

Proof Let $q \in I^+(p)$ and let α be a timelike curve from p to q. Let N be an open normal geodesic neighbourhood of q, then since N is an open set, α is eventually inside N. So we can choose a point $r \neq q$ in $N \cap \alpha$ such that α stays inside N on the segment from r to q. Since N is a normal neighbourhood of all of its points, N is a normal neighbourhood of r. Since the segment of α from r to q is timelike, $q \in I^+(r,N)$. By the previous proposition, since the segment of α from p to r is timelike, $I^+(r,N) \subset I^+(p)$. And $I^+(r,N)$ is open, being the image under a diffeomorphism (exp restricted to N) of an open set (the future timecone at r). Thus we've shown that every point in $I^+(p)$ has a neighbourhood contained in $I^+(p)$. ■

We see from these results, and from the proofs of these results that normal geodesic neighbourhoods behave like Minkowski space in many ways and are thus very useful for making certain arguments. For example, as we've seen above and as we'll see more of below, when we're considering a sequence of curves "converging" in some sense to some other curve,

the usual way of handling this is to cover the curve with normal neighbourhoods and then use $\exp^{-1}$ to get back to the more familiar setting of curves converging in $\mathbb{R}^n$.

We will use the same definitions of the future and past domains of dependence $D^+(S)$, $D^-(S)$ of a set S, and the same definition of future/past-inextendible curves as those given in class. With $D(S) = D^+(S) \cup D^+(S)$, we also use the same definition of Cauchy surface as a closed achronal set S such that $D(S) = M$. In order that we are not drawn too far away from the main objective, we will follow [N] and use a slightly different and more convenient definition of globally hyperbolic from the one given in class.

Definition We will say that (M, g) is *globally hyperbolic* if there exists a smooth function $f : M \to \mathbb{R}$ such that $grad(f)$ is a timelike vector field, and the level sets S_c with $c \in Im(f)$ are spacelike Cauchy surfaces. We call f a *(global) time function* and the sets S_c *(global) time slices*.

A standard result from differential geometry (see [BG]) states that since $grad(f)$ is timelike and hence non-zero (i.e. f is *regular*), the level sets of f are embedded hypersurfaces of M. Moreover since $grad(f)$ is orthogonal to its level sets, and since $grad(f)$ is timelike, the hypersurfaces S_c are spacelike.

Following [N], we make a definition

Definition A *simple neighbourhood* $N \subset M$ is a normal convex neighbourhood diffeomorphic to an open ball, and whose boundary is a compact submanifold of a larger normal neighbourhood.

Using the exponential map, it is clear that simple neighbourhoods exist around any point (Start with an open normal neighbourhood N_1, choose $p \in N_1$ and let B be an open ball around 0 in T_pM contained in $\exp_p^{-1}(N_1)$. Finally let N be the image under $\exp_p$ of an open ball around $0 \in T_pM$ with half the radius of B.)

The following lengthy (but essential) argument is in [N], though I have filled in many of the details. The reader is advised to draw a picture in order to keep track of the argument.

Theorem 3.9 *Let M be globally hyperbolic and $S = f^{-1}(0)$ a time slice with $p \in D^+(S)$ and f the global time function. Then $A := J^-(p) \cap D^+(S)$ is compact.*

Proof Since every point in M has a simple neighbourhood, we can cover M with simple neighbourhoods. Since M is paracompact (we'll take the definition of manifold to include paracompactness, as is often done) we can find a locally finite refinement of this cover. So we have a locally finite cover of A by simple neighbourhoods U_λ. Suppose A is not compact, then this cover does not have a finite subcover. In particular, this means we can find an infinite non-repeating sequence of points $q_k \in U_k \cap A$.

If the sequence q_k has an accumulation point q, then every neighbourhood of q would contain an infinite number of the q_k, and hence every neighbourhood of q would intersect non-trivially with infinitely many of the U_k, contradicting the locally finite requirement. Thus q_k has no accumulation points. Since the closure of each simple neighbourhood is compact (being diffeomorphic to a closed ball in $\mathbb{R}^n$), it follows that each of the simple neighbourhoods can contain at most finitely many of the q_k (an infinite sequence in a compact set always has an accumulation point).

We will now use this sequence of points together with the properties of the cover to construct an inextendible past-directed causal curve from p that does not intersect S. This will contradict the fact that S is a Cauchy surface. We will frequently take subsequences, but will keep the same index to keep the notation from becoming too bad.

Let $p_1 = p$. Since this is in A, there is some simple neighbourhood, say U_1 (relabelling if necessary), such that $p_1 \in U_1$. Since for all k, $q_k \in A \subset J^-(p)$, there is a future-directed causal curve α_k connecting each of these points to p_1. All but finitely many of the q_k lie outside U_1, so all but finitely many of the α_k must intersect the compact boundary of U_1, and hence the intersection points $r_{1,k}$ must accumulate at at least one point. Choose one of these accumulation points and call it $p_2 \in \partial U_1$.

What can we say about this point p_2? Well the points $r_{1,k}$ lie inside a normal neighbourhood of p_1 (by definition of simple neighbourhood) and along a causal curve through p_1, hence there is a unique past-directed causal geodesic segment $\gamma_{1,k}$ from p_1 to $r_{1,k}$. Using the familiar trick of applying $\exp_p^{-1}$ on the normal neighbourhood (so that we just get a sequence of line segments in T_pM converging to some other line segment–which corresponds to a causal geodesic in M), we know that these causal geodesic segments tend to a causal geodesic segment γ_1 from p_1 to p_2. So in particular $p_2 \in J^-(p)$. Moreover, since $r_{1,k} \geq q_k$ and $q_k \in D^+(S)$, we know that $f(r_{1,k}) \geq f(q_k) \geq 0 = f(S)$. Since f is smooth hence continuous, $f(p_1) > f(p_2) \geq 0$ so $p_2 \in D^+(S)$ too. Thus $p_2 \in A$.

Now we essentially iterate this construction. Since $p_2 \in A$ but not in U_1, there must be some distinct simple neighbourhood from the cover, say U_2, with $p_2 \in U_2$. Since U_2 is an open neighbourhood of p_2, and p_2 is an accumulation point for the $r_{1,k}$, infinitely many of the $r_{1,k}$ lie in U_2. Now U_2 can contain at most finitely many of the q_k, so infinitely many of the curves α_k from q_k to p_1 which pass through the points $r_{1,k} \in U_2$ must intersect the boundary of U_2 at least twice (since they cannot stay in U_2), and at least one of these intersections must take place at an earlier "time" (using the global time function f) than $r_{1,k}$ (this is because infinitely many of the $r_{1,k}$ are *inside* U_2, so the causal curves (along which f must increase) α_k must have crossed the boundary of U_2 at an earlier time). This gives a second sequence of intersection points $r_{2,k}$ with $f(r_{2,k}) < f(r_{1,k})$, which have some accumulation point p_3 in the compact set ∂U_2. Using essentially the same argument as above, since we are working in a normal neighbourhood, this gives us a sequence of past-directed causal (since α_k goes between $r_{2,k}$ and $r_{1,k}$ and is causal) geodesic segments $\gamma_{2,k}$ joining $r_{2,k}$ to $r_{1,k}$ that converge to a causal geodesic segment γ_2 from p_3 to p_2, which shows that $p_3 \in J^-(p)$ (apply proposition 3.7). As above, the facts that $r_{2,k} \geq q_k$, $q_k \in D^+(S)$, and $f(r_{1,k}) > f(r_{2,k}) \geq f(q_k) \geq 0 = f(S)$ imply that $f(p_1) > f(p_2) > f(p_3) \geq 0$, and so $p_3 \in A$.

Repeating this argument, we get an infinite sequence p_i in A which cannot converge (since they all lie in different sets from the locally finite cover), together with causal geodesic segments joining each consecutive pair of them. Using theorem 3.4, we can perturb this broken geodesic curve so that it is a smooth causal geodesic passing through all the points p_i. Moreover $f(p_i) > 0$ for all i so that this curve never intersects S. This is the desired contradiction. ■

4 Focal Points

With the notation for a globally hyperbolic manifold as above, we can apply theorem 2.7 to deduce that $\exp^{\perp}$ is a diffeomorphism $U \to V$ where U is some neighbourhood of 0 in NS and V is some neighbourhood of S in M. We can thus define a function t at least locally by

$$t(q) = \left|(\exp^{\perp})^{-1}(q))\right|.$$

$t(q)$ is really the time that elapses for an observer travelling away from S along the unique timelike geodesic from S to q, which we call the normal geodesic from S to q (since it is orthogonal to S). By Gauss' lemma from differential geometry, $X := grad(t)$ is a timelike vector field with the vectors being also the tangent vector fields of the geodesics $\exp^{\perp}(tu)$ for the unit vectors $u \in NS$. Note that the dual one form to $X := grad(t)$ is dt (definition of gradient). We define K to be the total covariant derivative of dt, i.e. $K = \nabla dt$, so that K is a covariant 2-tensor. In components

$$K_{ab} = X_{a;b}$$

Since the connection is torsion free, K is a symmetric tensor ($K_{ab} = K_{ba}$). Moreover since when we restrict X to one of the normal geodesics we get the tangent vector field of that geodesic, it follows that $\nabla_X X = 0$ and hence since dt is dual to X, $\nabla_X dt = 0$, or in components

$$X^b X_{a;b} = 0 \tag{4}$$

Definition The *expansion* θ of the family $\exp^{\perp}(tu)$, $u \in U \subset NS$ of timelike normal geodesics is the trace of K (which also equals the divergence of X).

In components,

$$\theta = Tr(K) = g^{ab} X_{a;b} = X^a_{;a}$$

We now prove the *Raychaudhuri equation* which, in this situation, looks as follows:

Theorem 4.1 *With notation as above, the following equation holds*

$$X\theta + Tr(K^2) + Ric(X, X) = 0$$

or equivalently in components

$$X^k \theta_{;k} + X^a_{;b} X^b_{;a} + R_{ab} X^a X^b = 0$$

Proof We start by computing $X\theta$

$$X^k \theta_{;k} = X^k X^a_{;ak} = X^k (X^a_{ka} - R^a_{kab} X^b) = X^k (X^a_{ka} - R_{kb} X^b) \tag{5}$$

by the Ricci identity and the definition of the Ricci curvature. Now since $\nabla_X X = 0$ we have

$$0 = (X^k X^a_{;k})_{;a} = X^k_{;a} X^a_{;k} + X^k X^a_{;ka}$$

and substituting this into (5) we get

$$X^k \theta_{;k} = X^k (X^a_{ka} - R_{kb} X^b) = -X^k_{;a} X^a_{;k} - R_{kb} X^b X^k$$

which is the required result. ■

We now prove a result that will be essential to the proof of the singularity theorem. For convenience, this theorem will be stated in terms of geodesics reaching conjugate points (or singularities) in the *future*, although they apply equally to the past since we can always reverse the time orientation on M.

Theorem 4.2 *Suppose M is globally hyperbolic and satisfies $Ric(X,X) \geq 0$ for all timelike vectors X. Suppose further that S is a time slice, and that the expansion θ_0 at $p \in S$ is negative. Consider the unique timelike geodesic $\alpha(t)$ starting at p and with initial velocity orthogonal to S (recall S is a spacelike hypersurface). If α can be extended to a proper time $\frac{3}{-\theta_0}$ then α contains a conjugate point.*

Proof With the condition on the Ricci curvature, the Raychaudhuri equation yields

$$X\theta + Tr(K^2) \leq 0 \tag{6}$$

Recall that

$$(A,B) := Tr(A^T B)$$

defines an inner product on square matrices. The Cauchy-Schwarz inequality states

$$|(A,B)|^2 \leq (A,A)(B,B)$$

If we put B equal to the identity matrix, this gives

$$(Tr(A))^2 \leq nTr(A^T A)$$

for $n \times n$ matrices. Now K is symmetric and $\theta = Tr(K)$, so if we apply this to $Tr(K^2)$ we get

$$Tr(K^2) \geq \frac{1}{4}\theta^2. \tag{7}$$

In fact, we can get a slightly stronger inequality by using equation (4). If we use a basis whose first vector is X, then by (4) we have

$$X^b X^a_{;b} = 0$$

and since K is symmetric, its matrix with respect to this basis takes the form

$$\begin{pmatrix} 0 & 0 \\ 0 & \tilde{K} \end{pmatrix}$$

where $\tilde{K}$ is some 3×3 matrix. It follows that (7) can be improved to

$$Tr(K^2) \geq \frac{1}{3}\theta^2. \tag{8}$$

We can put this into (6). Since the dual one form to X is dt, we can just write $\frac{d\theta}{dt}$ instead of $X\theta$. So (6) becomes

$$\frac{d\theta}{dt} + \frac{\theta^2}{3} \leq 0$$

so

$$\int_0^t \frac{1}{\theta^2}\frac{d\theta}{dt}dt \leq -\int_0^t \frac{1}{3}dt$$
$$\Rightarrow \qquad \frac{1}{\theta} \geq \frac{1}{\theta_0} + \frac{t}{3}.$$

And now since θ_0 is negative, as t increases from 0 the right side approaches zero. Thus the continuous function $\frac{1}{\theta}$ starts from a negative value $\frac{1}{\theta_0}$ at $t = 0$ and approaches zero. This is possible only if θ diverges in a finite time, which we can see can be no greater than $\frac{3}{-\theta_0}$. This is a contradiction since θ is a continuous function, so one of our assumptions must be false. We have made two assumptions, (1) that α can indeed be extended to a proper time of at least $\frac{3}{-\theta_0}$ and (2) that $\exp^\perp$ is non-singular, i.e. has no conjugate points, along the curve α up to $t = \frac{3}{-\theta_0}$ (which allows us to define X and write down the Raychaudhuri equation). Consequently, one of these two assumptions must be false. ■

5 The Main Theorem

We return to studying causality and build up quickly to the singularity theorem stated at the beginning of this essay.

Definition For $p, q \in M$, we define the *time separation* $\tau(p, q)$ to be

$$\tau(p, q) = \sup\{L(\alpha) | \alpha \text{ is a future-pointing causal curve from p to q}\}$$

where as before $L(\alpha)$ denotes the length (or proper time elapsed) along α.

Lemma 5.1 *1.* $\tau(p, q) > 0$ *iff* $p \ll q$

2. If $p \leq q \leq r$ *then* $\tau(p, q) + \tau(q, r) \leq \tau(p, r)$.

Proof (1) If $p \ll q$ then there is a timelike curve between them and so $\tau(p, q) > 0$. Conversely, if $\tau(p, q) > 0$ then there is a causal curve between them with positive length. So this curve cannot be a null geodesic, and hence by theorem 3.4 we can deform it into a timelike curve. Thus $p \ll q$.

(2) Fix $\epsilon > 0$ and choose causal curves α from p to q and β from q to r such that $\tau(p, q) < L(\alpha) - \epsilon/2$ and $\tau(q, r) < L(\beta) - \epsilon/2$. Then $\tau(p, q) + \tau(q, r) < L(\alpha) + L(\beta) - \epsilon$. The piecewise smooth causal curve γ which consists of following α and β in succession has length $L(\alpha) + L(\beta)$ and goes from p to r. Thus $\tau(p, q) + \tau(q, r) < \tau(p, r) - \epsilon$, and since $\epsilon > 0$ was arbitrary, we get the result. ■

Lemma 5.2 $\tau : M \times M \to [0, \infty]$ *is lower semicontinuous.*

Remark A function $f : X \to \mathbb{R}$ is *semicontinuous* if for each $\epsilon > 0$ and each $x \in X$ there is a neighbourhood N of x such that $f(N) \subset (-\infty, f(x) + \epsilon]$. Alternatively, f is continuous with respect to the topology $\{(-\infty, a) | a \in [-\infty, +\infty]\}$. For the proof, we follow [O] but I will fill in a few details.

Proof Fix $\epsilon > 0$, and suppose $q \in I^+(p)$ (the other cases are trivial and uninteresting) so we can get a timelike curve α from q to p, and in fact we can assume that $L(\alpha) > \tau(p,q) - \epsilon/3$. (This is because we can always find a causal curve β from p to q with this property, and since $q \in I^+(p)$, this curve will not be a null geodesic, hence by an arbitrarily small deformation, we get a timelike curve β_v from p to q. The function $h(v) = \int |g(\beta', \beta') - g(\beta_v', \beta_v')| du$ will be continuous (since for example the deformation is continuous and happens on a compact subset of M) and so we can make the change in time separation as small as we like by choosing v sufficiently small.)

Take a normal neighbourhood N of q and let q_1 be a point in $N \cap \alpha$ before q (which exists since N is open). Now, inside a normal neighbourhood, the distance function is well behaved and depends continuously on the endpoints (this follows e.g., from the fact that N is diffeomorphic to a neighbourhood of 0 in T_qM), and the time separation separation is just the length of the unique geodesic between each pair of timelike (or null) separated points. With this in mind, we can choose a neighbourhood $V \subset J^+(q_1)$ of q such that for all q' in V, $\tau(q_1, q')$ is within $\epsilon/3$ of $\tau(q_1, q)$, so that in particular we have $\tau(q_1, q') > \tau(q_1, q) - \epsilon/3$. Also note that $\tau(q_1, q)$ is at least as large as the length of α from q_1 to q, by definition of τ.

We perform the same construction to get a corresponding neighbourhood U of the point p and a special point p_1 on α. Then consider $p' \in U$ and $q' \in V$. We can connect them with a causal curve by first taking the geodesic segment from p' to p_1, then following α to q_1, and then taking the geodesic segment from q_1 to q'. This curve will have length at least $L(\alpha) - 2\epsilon/3 > \tau(p,q) - \epsilon$, as required. ■

Lemma 5.3 *Let $A = D^+(S) \cap J^-(p)$ be the compact set considered in theorem 3.9, and let α_k be a sequence of future-pointing causal curves starting at some point $q \in S$ and going to p. Then there is a future-pointing causal geodesic γ from q to p and a subsequence α_m such that $\lim_{m\to\infty} L(\alpha_m) \leq L(\gamma)$.*

Proof By covering A with a finite (A is compact) number of simple neighbourhoods and arguing as in the proof of theorem 3.9 (except simpler now since A is covered with only finitely many neighbourhoods; we adapt the same notation as in that proof) we get a finite set of points p_i which are accumulation points for the intersection points of the curves α_n with the boundaries of the simple neighbourhoods. Also, again following the line of reasoning in theorem 3.9, inside each simple neighbourhood, letting $r_{i,k}$ denote the intersection points in the appropriate subsequence, we get that the causal geodesic segment from $r_{i,k}$ to $r_{i+1,k}$ tends to the geodesic segment from p_i to p_{i+1}, which is therefore causal. Since time separation depends continuously on endpoints at least inside a normal neighbourhood, the proper time of the geodesic segment from $r_{i,k}$ to $r_{i+1,k}$, which is greater than or equal to the proper time of α_k between these two points (since causal geodesics maximize proper time at least inside normal neighbourhoods), tends to the time separation between p_i and p_{i+1}.

Let γ_k denote the broken causal geodesic formed by joining the consecutive intersection points $r_{i,k}$ for $i = 1, ..., n$ with causal geodesic segments (this happens in a normal neighbourhood, so these segments are unique), and let γ be likewise for the points p_i. Hence by the remarks above we have

$$L(\alpha_k) \leq L(\gamma_k)$$

and the right hand side tends to $L(\gamma)$, so that $L(\gamma)$ is an upper bound for infinitely many of the $L(\alpha_k)$. Thus choosing an appropriate subsequence α_m gives the desired result. ■

Proposition 5.4 *With S, p as in theorem 3.9, let $q \in A = D^+(S) \cap J^-(p)$, and suppose in fact $q < p$. Then there is a causal curve from q to p of length $\tau(q,p)$.*

Proof We can choose a sequence of causal curves from q to p whose lengths converge to $\tau(q,p)$ (by the definition of τ). By lemma 5.3, there is a broken causal geodesic from q to p whose proper time is greater than or equal to the proper times of some subsequence of the sequence of curves. Since the proper time of the curves in that subsequence must also tend to $\tau(q,p)$, the proper time of the broken causal geodesic must equal $\tau(q,p)$. ■

Lemma 5.5 *The compact subsets of $\mathbb{R}$ with respect to the topology $T := \{(-\infty, a) | a \in [-\infty, +\infty]\}$ are the subsets which contain their supremum (finite).*

Proof Suppose $U \subset \mathbb{R}$ contains its supremum m, then clearly any open cover of U would have to contain $(-\infty, m+\epsilon)$ for some $\epsilon > 0$. Then $(-\infty, m+\epsilon)$ is a finite subcover.

Conversely, if U is such that every open cover has a finite subcover, then in particular, U must be bounded above. Therefore it has a finite supremum. If the supremum is not contained in the set then we can take a sequence of intervals $(-\infty, a_k)$ with a_k approaching the supremum from below. This will be an open cover without a finite subcover, a contradiction. ■

Theorem 5.6 *Let M be globally hyperbolic, and $S = f^{-1}(0)$ a time slice with f a global time function. Let A be as above. Let $p \in D^+(S)$, then there exists a timelike geodesic orthogonal to S which achieves the maximal possible time separation between p and any point in S.*

Proof Since $A \cap S$ is compact (A is compact and S is closed), the function $h(x) := \tau(x,p)$, which is lower semicontinuous (i.e. T-continuous), maps $A \cap S$ onto some T-compact set in $\mathbb{R}$. In the lemma just proved, we showed that the T-compact sets in $\mathbb{R}$ are precisely those sets that contain their supremum. Thus h has a maximum at some q in $A \cap S$. Given a point $q \in A \cap S$, there is a causal curve from q to p with maximal proper time by proposition 5.4. This gives the desired result. ■

And finally,

Theorem 5.7 *Let (M, g) be a Lorentzian manifold satisfying the following:*

1. *M is globally hyperbolic.*
2. *$Ric(v,v) \geq 0$ for all timelike vectors v, where Ric is the Ricci tensor.*
3. *There is a spacelike "time slice" S such that the expansion scalar $\theta \leq \theta_0 < 0$ on S.*

Then M is singular.

Remark As stated, this theorem shows the existence of a singularity in the future. This is why the statement here bounds the expansion from above by a negative constant, i.e. we are considering a contracting universe. By reversing time orientation, it can be made to apply to the past as well (then the expansion is bounded below by positive constant as in the statement given in the introduction; this statement would apply to an expanding universe, such as our own!)

Proof Suppose there exists a timelike geodesic α parametrized by proper time starting from S and reaching a parameter value greater than $\frac{3}{-\theta_0}$, say $\frac{3}{-\theta_0} + \epsilon$ at the point $p \in M$. By theorem 5.6, there is a timelike geodesic γ from S to p which maximizes the proper time. Since it maximizes proper time, in particular its proper time must be greater than or equal to $\frac{3}{-\theta_0} + \epsilon$ (since it must beat α). But theorem 4.2 guarantees that γ must reach a conjugate point no later than when it reaches the proper time parameter value $\frac{3}{-\theta_0}$, and hence by theorem 2.11 γ must stop maximizing proper time before p. This is a contradiction, and this proves the result. ■

6 Quantum Singularities

In classical physics, it makes sense to define a singular spacetime as one that is geodesically incomplete. This is because free-falling classical particles travel along geodesics, so an inextendible geodesic that cannot be extended past a certain finite proper time parameter value represents a loss of predictive power. As the authors point out in [HM], this is no longer necessarily the case for quantum particles. The authors define a spacetime to be quantum mechanically singular if the evolution of some quantum state is not uniquely defined (for all time).

As an example, the authors consider a system whose Hamiltonian is proportional to the Laplacian operator on some manifold M (in the usual representation, this could be the Hamiltonian for a free, non-relativistic particle on M). Initially the Laplacian is defined on some subset of L^2 (for example, smooth functions having compact support). Although it is symmetric, it may not be fully self-adjoint (an operator is self-adjoint if it has the same domain as its adjoint). The usual technique employed is to extend the Laplacian to have a larger domain (for example, the Fourier transform, defined initially on the Schwarz space, has a unique unitary extension to L^2 sometimes called the Fourier-Plancherel transform; or it may be possible to extend an operator to act on certain distributions, etc.) and make it self-adjoint; and once one is in possession of a self-adjoint operator, the usual theorems apply and you can exponentiate the Hamiltonian to get the (deterministic) evolution of the quantum system.

The essential problem that can arrise then is that there could be more than one self-adjoint extension, and different extensions could give rise to different evolutions of the same state. For the Laplacian at least, this problem never arrises in a classically non-singular manifold: the authors cite a theorem from functional analysis that on a geodesically complete manifold, the Laplacian has a unique self-adjoint extension. In some cases (the authors give examples, but my knowledge of quantum mechanics/functional analysis is not adequate to comment on them) the Laplacian has a unique self-adjoint extension even though the spacetime is geodesically incomplete. These are examples of spacetimes that are classically singular but not quantum mechanically singular.

The authors go on to spell this out in a bit more detail in a static spacetime. They write the Klein-Gordon equation in the form

$$\partial_t^2 \psi = V D^a (V D_a \psi) - V^2 m^2 \psi =: -A\psi$$

where $V^2 = -\xi^b \xi_b$, ξ is the timelike Killing field (the spacetime is static), and D is the

spatial covariant derivative on the surfaces of constant time. If A has a unique self-adjoint extension A_E which is positive definite, then (using the spectral theorem) you can take its positive "square root" (I take it this is a generalization of what happens when going from the Klein-Gordon equation to the Dirac equation). The wave equation is then

$$i\frac{d\psi}{dt} = (A_E)^{\frac{1}{2}}\psi$$

and the evolution is

$$\psi(t) = \exp[-it(A_E)^{\frac{1}{2}}]\psi(0). \tag{9}$$

The equation (9) won't in general agree for different self-adjoint extensions, so it is not clear what the evolution should be if more than one self-adjoint extension is possible.

The authors provide a number of examples of classically singular spherically symmetric spacetimes:

$$ds^2 = -V(r)^2dt^2 + V(r)^{-2}dr^2 + R(r)^2d\Omega.$$

Some of these remain singular for quantum states (notably the Schwarzschild solution), but others are not quantum-mechanically singular.

7 Bibliography

[O] Barrett O'Neill (1983), *Semi-Riemannian Geometry With Applications to Relativity.* Academic Press, New York.

[N] Jose Natario (2009), *Relativity and Singularities–A Short Introduction for Mathematicians.* Available on arXiv:math/0603190v3.

[HM] Gary Horowitz and Donald Marolf (1995), *Quantum Probes of Spacetime Singularities.* Phys. Rev. D. 52, 5670.

[BG] Keith Burns and Marian Gidea (2005), *Differential Geometry and Topology.* Chapman & Hall, New York.

[HE] Stephen Hawking and G.F.R. Ellis (1973), *The Large Scale Structure of Space-time.* Cambridge University Press, Cambridge, U.K.