CHAPTER ONE

Classical Theory

S. W. Hawking

IN THESE LECTURES, Roger Penrose and I will put forward our related but rather different viewpoints on the nature of space and time. We shall speak alternately and shall give three lectures each, followed by a discussion on our different approaches. I should emphasize that these will be technical lectures. We shall assume a basic knowledge of general relativity and quantum theory.

There is a short article by Richard Feynman describing his experiences at a conference on general relativity. I think it was the Warsaw conference in 1962. It commented very unfavorably on the general competence of the people there and the relevance of what they were doing. That general relativity soon acquired a much better reputation, and more interest, is in considerable measure due to Roger's work. Up to then, general relativity had been formulated as a messy set of partial differential equations in a single coordinate system. People were so pleased when they found a solution that they didn't care that it probably had no physical significance. However, Roger brought in modern concepts like spinors and global methods. He was the first to show that one could discover general properties without solving the equations exactly. It was his first singularity theorem that introduced me to the study of causal structure and inspired my classical work on singularities and black holes.

I think Roger and I pretty much agree on the classical work. However, we differ in our approach to quantum gravity and indeed to quantum theory itself. Although I'm regarded as a dangerous radical by particle physicists for proposing that there may be loss of quantum coherence, I'm definitely a conservative compared to Roger. I take the positivist viewpoint that a physical theory is just a mathematical model and that it is meaningless to ask whether it corresponds to reality. All that one can ask is that its predictions should be in agreement with observation. I think Roger is a Platonist at heart but he must answer for himself.

Although there have been suggestions that spacetime may have a discrete structure, I see no reason to abandon the continuum theories that have been so successful. General relativity is a beautiful theory that agrees with every observation that has been made. It may require modifications on the Planck scale, but I don't think that will affect many of the predictions that can be obtained from it. It may be only a low energy approximation to some more fundamental theory, like string theory, but I think string theory has been oversold. First of all, it is not clear that general relativity, when combined with various other fields in a supergravity theory, cannot give a sensible quantum theory. Reports of the death of supergravity are exaggerations. One year everyone believed that supergravity was finite. The next year the fashion changed and everyone said that supergravity was bound to have divergences even though none had actually been found. My second reason for not discussing string theory is that it has not made any testable predictions. By contrast, the straightforward application of quantum theory to general relativity, which I will be talking about, has already made two testable predictions. One of these predictions, the development of small perturbations during inflation, seems to be confirmed by recent observations of fluctuations in the microwave background. The other prediction, that black holes should radiate thermally, is testable in principle. All we have to do is find a primordial black hole. Unfortunately, there don't seem to be many around in this neck of the woods. If there had been, we would know how to quantize gravity.

Neither of these predictions will be changed even if string theory is the ultimate theory of nature. But string theory, at least at its current state of development, is quite incapable of making these predictions except by appealing to general relativity as the low energy effective theory. I suspect this may always be the case and that there may not be any observable predictions of string theory that cannot also be predicted from general relativity or supergravity. If this is true, it raises the question of whether string theory is a genuine scientific theory. Is mathematical beauty and completeness enough in the absence of distinctive observationally tested predictions? Not that string theory in its present form is either beautiful or complete.

For these reasons, I shall talk about general relativity in these lectures. I shall concentrate on two areas where gravity seems to lead to features that are completely different from other field theories. The first is the idea that gravity should cause spacetime to have a beginning and maybe an end. The second is the discovery that there seems to be intrinsic gravitational entropy that is not the result of coarse graining. Some people have claimed that these predictions are only artifacts of the semiclassical approximation. They say that string theory, the true quantum theory of gravity, will smear out the singularities and will introduce correlations in the radiation from black holes so that it is only approximately thermal in the coarse-grained sense. It would be rather boring if this were the case. Gravity would be just like any other field. But I believe it is distinctively different, because it shapes the arena in which it acts, unlike other fields which act in a fixed spacetime background. It is this that leads to the possibility of time having a beginning. It also leads to regions of the universe that one can't observe, which in turn gives rise to the concept of gravitational entropy as a measure of what we can't know.

In this lecture I shall review the work in classical general relativity that leads to these ideas. In my second and third lectures (Chapters 3 and 5) I shall show how they are changed and extended when one goes to quantum theory. My second lecture will be about black holes, and the third will be on quantum cosmology.

The crucial technique for investigating singularities and black holes that was introduced by Roger, and which I helped develop, was the study of the global causal structure of spacetime. Define $I^+(p)$ to be the set of all points of the spacetime M that can be reached from p by future-directed timelike curves (see fig. 1.1). One can think of $I^+(p)$ as the set of all events that can be influenced by what happens at p. There are similar definitions in which plus is replaced by minus and future by past. I shall regard such definitions as self-evident.

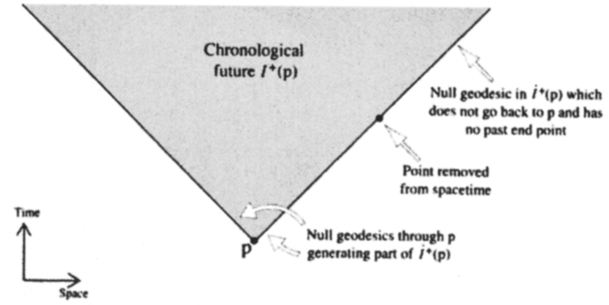
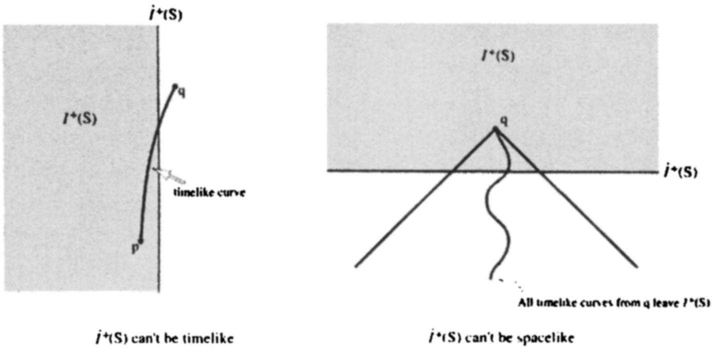

Chronological future $I^+(p)$
Null geodesic in $\dot{I}^+(p)$ which does not go back to p and has no past end point
Point removed from spacetime
Time
Space
p
Null geodesics through p generating part of $\dot{I}^+(p)$

Figure 1.1 The chronological future of a point p.



$\dot{I}^+(S)$ can't be timelike

$\dot{I}^+(S)$ can't be spacelike

Figure 1.2 The boundary of the chronological future cannot be timelike or spacelike.

One now considers the boundary $\dot{I}^+(S)$ of the future of a set S. It is fairly easy to see that this boundary cannot be timelike. For in that case, a point q just outside the boundary would be to the future of a point p just inside. Nor can the boundary of the future be spacelike, except at the set S itself. For in that case every past-directed curve from a point q, just to the future of the boundary, would cross the boundary and leave the future of S. That would be a contradiction with the fact that q is in the future of S (fig. 1.2).

One therefore concludes that the boundary of the future is null apart from the set S itself. More precisely, if q is in the boundary of the future but is not in the closure of S, there is a past-directed null

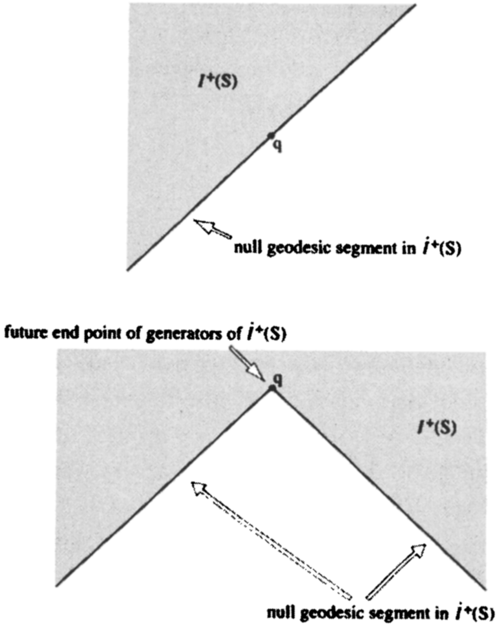

$I^+(S)$
q
null geodesic segment in $\dot{I}^+(S)$
future end point of generators of $\dot{I}^+(S)$
q
$I^+(S)$
null geodesic segment in $\dot{I}^+(S)$

Figure 1.3 *Top*: The point q lies in the boundary of the future, so there is a null geodesic segment in the boundary which passes through q. *Bottom*: If there is more than one such segment, the point q will be their future endpoint.

geodesic segment through q lying in the boundary (see fig. 1.3). There may be more than one null geodesic segment through q lying in the boundary, but in that case q will be a future endpoint of the segments. In other words, the boundary of the future of S is generated by null geodesics that have a future endpoint in the boundary and pass into the interior of the future if they intersect another generator. On the other hand, the null geodesic generators can have past endpoints only on S. It is possible, however, to have spacetimes in which there are generators of the boundary of the future of a set S that never intersect S. Such generators can have no past endpoint.

A simple example of this is Minkowski space with a horizontal line segment removed (see fig. 1.4). If the set S lies to the past of the

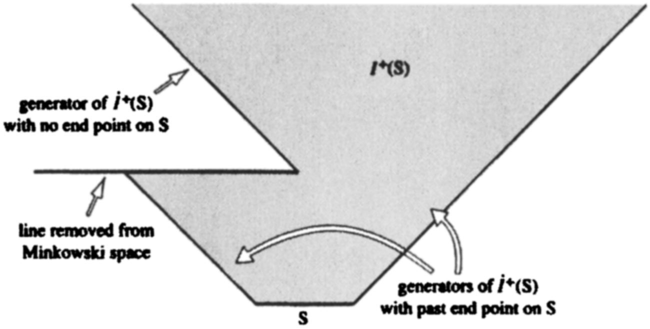


Figure 1.4 As a line has been removed from Minkowski space, the boundary of the future of the set *S* has a generator with no past endpoint.

horizontal line, the line will cast a shadow and there will be points just to the future of the line that are not in the future of *S*. There will be a generator of the boundary of the future of *S* that goes back to the end of the horizontal line. However, as the endpoint of the horizontal line has been removed from spacetime, this generator of the boundary will have no past endpoint. This spacetime is incomplete, but one can cure this by multiplying the metric by a suitable conformal factor near the end of the horizontal line. Although spaces like this are very artificial, they are important in showing how careful you have to be in the study of causal structure. In fact, Roger Penrose, who was one of my Ph.D. examiners, pointed out that a space like the one I just described was a counterexample to some of the claims I made in my thesis.

To show that each generator of the boundary of the future has a past endpoint on the set, one has to impose some global condition on the causal structure. The strongest and physically most important condition is that of global hyperbolicity. An open set *U* is said to be globally hyperbolic if

1. For every pair of points *p* and *q* in *U* the intersection of the future of *p* and the past of *q* has compact closure. In other words, it is a bounded diamond shaped region (fig. 1.5).
2. Strong causality holds on *U*. That is there are no closed or almost closed timelike curves contained in *U*.

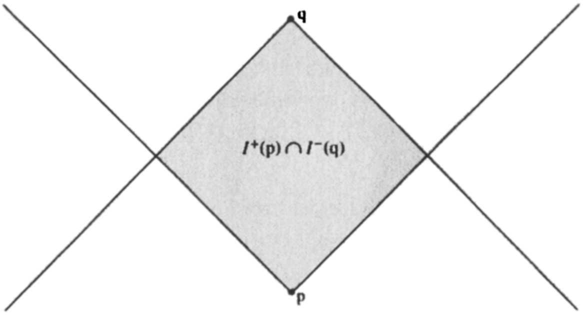


Figure 1.5 The intersection of the past of q and the future of p has compact closure.

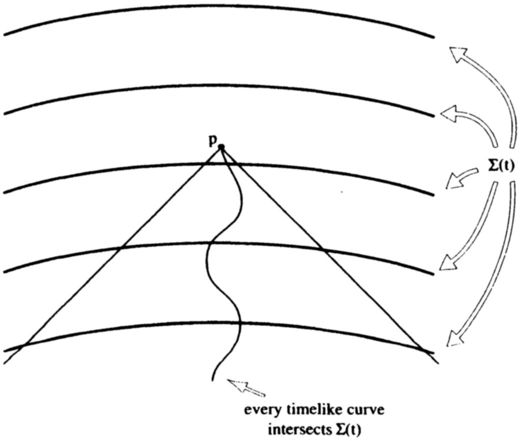


Figure 1.6 A family of Cauchy surfaces for U.

The physical significance of global hyperbolicity comes from the fact that it implies that there is a family of Cauchy surfaces $\Sigma(t)$ for U (see fig. 1.6). A Cauchy surface for U is a spacelike or null surface that intersects every timelike curve in U once and once only. One can predict what will happen in U from data on the Cauchy surface, and one can formulate a well-behaved quantum field theory on a globally hyperbolic background. Whether one can formulate a sensible

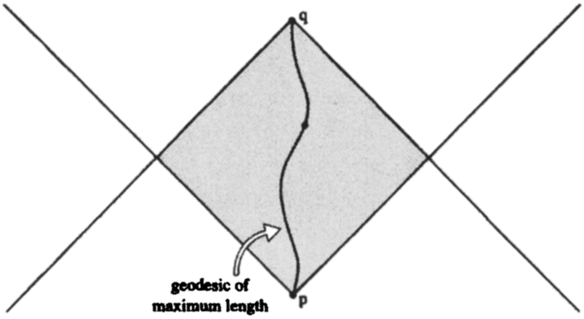


Figure 1.7 In a globally hyperbolic space, there is a geodesic of maximum length joining any pair of points that can be joined by a timelike or null curve.

quantum field theory on a nonglobally hyperbolic background is less clear. So global hyperbolicity may be a physical necessity. But my viewpoint is that one shouldn't assume it because that may be ruling out something that gravity is trying to tell us. Rather, one should deduce that certain regions of spacetime are globally hyperbolic from other physically reasonable assumptions.

The significance of global hyperbolicity for singularity theorems stems from the following. Let U be globally hyperbolic and let p and q be points of U that can be joined by a timelike or null curve. Then there is a timelike or null geodesic between p and q which maximizes the length of timelike or null curves from p to q (fig. 1.7). The method of proof is to show that the space of all timelike or null curves from p to q is compact in a certain topology. One then shows that the length of the curve is an upper semicontinuous function on this space. It must therefore attain its maximum, and the curve of maximum length will be a geodesic because otherwise a small variation will give a longer curve.

One can now consider the second variation of the length of a geodesic γ. One can show that γ can be varied to a longer curve if there is an infinitesimally neighboring geodesic from p which intersects γ again at a point r between p and q. The point r is said to be conjugate to p (fig. 1.8). One can illustrate this by considering two points p and q

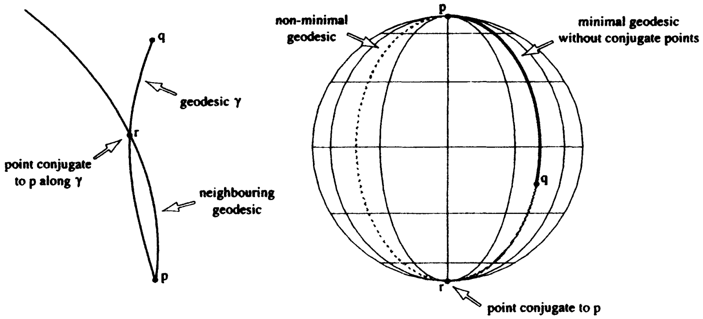


Figure 1.8 *Left*: if there is a conjugate point r between p and q on a geodesic, it is not the geodesic of minimum length. *Right*: The nonminimal geodesic from p to q has a conjugate point at the south pole.

on the surface of the Earth. Without loss of generality, one can take p to be at the north pole. Because the Earth has a positive definite metric rather than a Lorentzian one, there is a geodesic of minimal length, rather than a geodesic of maximum length. This minimal geodesic will be a line of longitude running from the north pole to the point q. But there will be another geodesic from p to q which runs down the back from the north pole to the south pole and then up to q. This geodesic contains a point conjugate to p at the south pole where all the geodesics from p intersect. Both geodesics from p to q are stationary points of the length under a small variation. But now in a positive definite metric the second variation of a geodesic containing a conjugate point can give a shorter curve from p to q. Thus, in the example of the Earth, we can deduce that the geodesic that goes down to the south pole and then comes up is not the shortest curve from p to q. This example is very obvious. However, in the case of spacetime one can show that under certain assumptions there ought to be a globally hyperbolic region in which there should be conjugate points on every geodesic between two points. This establishes a contradiction which shows that the assumption of geodesic completeness, which can be taken as a definition of a nonsingular spacetime, is false.

The reason one gets conjugate points in spacetime is that gravity is an attractive force. It therefore curves spacetime in such a way that neighboring geodesics are bent toward each other rather than away. One can see this from the Raychaudhuri or Newman-Penrose equation, which I will write in a unified form.

Raychaudhuri-Newman-Penrose Equation

$$\frac{d\rho}{dv} = \rho^2 + \sigma^{ij}\sigma_{ij} + \frac{1}{n}R_{ab}l^a l^b,$$

where $n = 2$ for null geodesics,
$n = 3$ for timelike geodesics.

Here v is an affine parameter along a congruence of geodesics with tangent vector l^a which is hypersurface orthogonal. The quantity ρ is the average rate of convergence of the geodesics, while σ measures the shear. The term $R_{ab}l^a l^b$ gives the direct gravitational effect of the matter on the convergence of the geodesics.

Einstein Equation

$$R_{ab} - \frac{1}{2}g_{ab}R = 8\pi T_{ab}.$$

Weak Energy Condition

$$T_{ab}v^a v^b \geq 0$$

for any timelike vector v^a.

By the Einstein equations, it will be nonnegative for any null vector l^a if the matter obeys the so-called weak energy condition. This says that the energy density T_{00} is nonnegative in any frame. The weak energy condition is obeyed by the classical energy momentum tensor of any reasonable matter, such as a scalar or electromagnetic field or a fluid with a reasonable equation of state. It may not, however, be satisfied locally by the quantum mechanical expectation value of the energy momentum tensor. This will be relevant in my second and third lectures (chapters 3 and 5).

Suppose the weak energy condition holds, and that the null geodesics from a point p begin to converge again and that ρ has the positive value ρ_0. Then the Newman-Penrose equation would imply that the convergence ρ would become infinite at a point q within an affine parameter distance $\frac{1}{\rho_0}$ if the null geodesic can be extended that far.

If $\rho = \rho_0$ at $v = v_0$ then $\rho \geq \frac{1}{\rho^{-1}+v_0-v}$. Thus there is a conjugate point before $v = v_0 + \rho^{-1}$.

Infinitesimally neighboring null geodesics from p will intersect at q. This means the point q will be conjugate to p along the null geodesic γ joining them. For points on γ beyond the conjugate point q there will be a variation of γ that gives a timelike curve from p. Thus γ cannot lie in the boundary of the future of p beyond the conjugate point q. So γ will have a future endpoint as a generator of the boundary of the future of p (fig. 1.9).

The situation with timelike geodesics is similar, except that the strong energy condition that is required to make $R_{ab}l^a l^b$ nonnegative for every timelike vector l^a is, as its name suggests, rather stronger. It is still, however, physically reasonable, at least in an averaged sense, in classical theory. If the strong energy condition holds, and the timelike geodesics from p begin converging again, then there will be a point q conjugate to p.

Strong Energy Condition

$$T_{ab}v^a v^b \geq \frac{1}{2}v^a v_a T$$

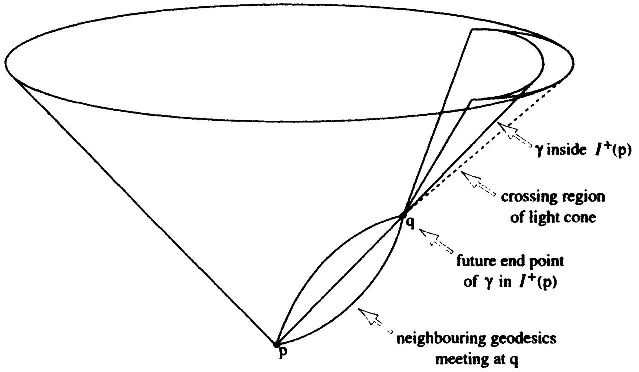

γ inside I^+(p)
crossing region of light cone
q
future end point of γ in I^+(p)
neighbouring geodesics meeting at q
p

Figure 1.9 The point q is conjugate to p along null geodesics, so a null geodesic γ that joins p to q will leave the boundary of the future of p at q.

Finally, there is the generic energy condition. This says that first the strong energy condition holds. Second, every timelike or null geodesic encounters some point where there is some curvature that is not specially aligned with the geodesic. The generic energy condition is not satisfied by a number of known exact solutions. But these are rather special. One would expect it to be satisfied by a solution that was "generic" in an appropriate sense. If the generic energy condition holds, each geodesic will encounter a region of gravitational focussing. This will imply that there are pairs of conjugate points if one can extend the geodesic far enough in each direction.

The Generic Energy Condition

1. The strong energy condition holds.
2. Every timelike or null geodesic contains a point where $l_{[a}R_{b]cd[e}l_{f]}l^c l^d \neq 0$.

One normally thinks of a spacetime singularity as a region in which the curvature becomes unboundedly large. However, the trouble with that as a definition is that one could simply leave out the singular points and say that the remaining manifold was the whole of spacetime. It is therefore better to define spacetime as the maximal manifold on which the metric is suitably smooth. One can then recognize the occurrence of singularities by the existence of incomplete geodesics that cannot be extended to infinite values of the affine parameter.

Definition of Singularity

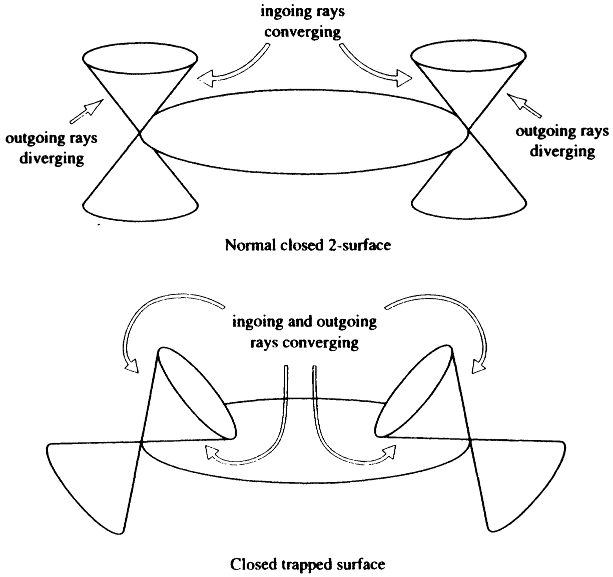
A spacetime is singular if it is timelike or null geodesically incomplete but cannot be embedded in a larger spacetime.

This definition reflects the most objectionable feature of singularities, that there can be particles whose history has a beginning or end at a finite time. There are examples in which geodesic incompleteness can occur with the curvature remaining bounded, but it is thought that generically the curvature will diverge along incomplete geodesics. This is important if one is to appeal to quantum effects to solve the problems raised by singularities in classical general relativity.

Between 1965 and 1970 Penrose and I used the techniques I have described to prove a number of singularity theorems. These theorems had three kinds of conditions. First there was an energy condition such as the weak, strong, or generic energy conditions. Then there was some global condition on the causal structure such as that there shouldn't be any closed timelike curves. And finally, there was some condition that gravity was so strong in some region that nothing could escape.

Singularity Theorems

1. Energy condition.
2. Condition on global structure.
3. Gravity strong enough to trap a region.



Normal closed 2-surface

Closed trapped surface

Figure 1.10 At a normal closed surface, the outgoing null rays from the surface diverge, while the ingoing rays converge. On a closed trapped surface, both the ingoing and outgoing null rays converge.

This third condition could be expressed in various ways. One way would be that the spatial cross section of the universe was closed, for then there would be no outside region to escape to. Another would be that there was what was called a closed trapped surface. This is a closed two-surface such that both the ingoing and outgoing null geodesics orthogonal to it were converging (fig. 1.10). Normally if you have a spherical two-surface in Minkowski space, the ingoing null geodesics are converging but the outgoing ones are diverging. But in the collapse of a star the gravitational field can be so strong that the light cones are tipped inward. This means that even the outgoing null geodesics are converging.

The various singularity theorems show that spacetime must be timelike or null geodesically incomplete if different combinations of the three kinds of conditions hold. One can weaken one condition if

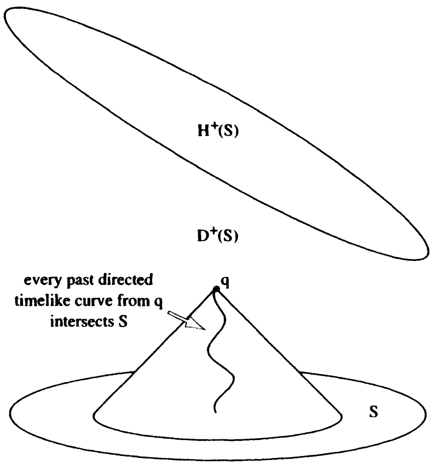


Figure 1.11 The future Cauchy development $D^+(S)$ of a set S and its future boundary, the Cauchy horizon $H^+(S)$.

one assumes stronger versions of the other two. I shall illustrate this by describing the Hawking-Penrose theorem. This has the generic energy condition, the strongest of the three energy conditions. The global condition is fairly weak, that there should be no closed timelike curves. And the no-escape condition is the most general, that there should be either a trapped surface or a closed spacelike three-surface.

For simplicity, I shall just sketch the proof for the case of a closed spacelike three-surface S. One can define the future Cauchy development $D^+(S)$ to be the region of points q from which every past-directed timelike curve intersects S (fig. 1.11). The Cauchy development is the region of spacetime that can be predicted from data on S. Now suppose that the future Cauchy development was compact. This would imply that the Cauchy development would have a future boundary called the *Cauchy horizon*, $H^+(S)$. By an argument similar to that for the boundary of the future of a point, the Cauchy horizon would be generated by null geodesic segments without past endpoints. However, since the Cauchy development is assumed to be compact, the Cauchy horizon will also be compact. This means that

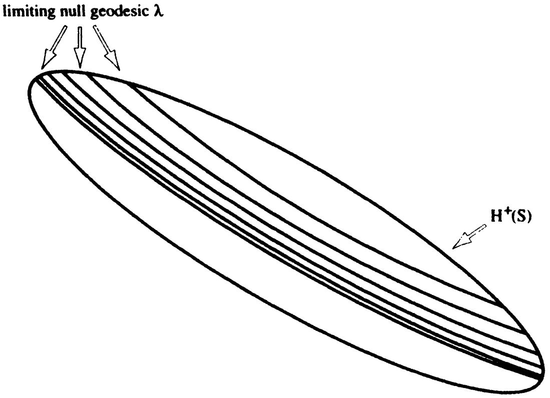


Figure 1.12 There is a limiting null geodesic λ in the Cauchy horizon which has no past or future endpoints in the Cauchy horizon.

the null geodesic generators will wind around and around inside a compact set. They will approach a limiting null geodesic λ that will have no past or future endpoints in the Cauchy horizon (fig. 1.12). But if λ were geodesically complete, the generic energy condition would imply that it would contain conjugate points p and q. Points on λ beyond p and q could be joined by a timelike curve. But this would be a contradiction because no two points of the Cauchy horizon can be timelike separated. Therefore either λ is not geodesically complete and the theorem is proved, or the future Cauchy development of S is not compact.

In the latter case one can show there is a future-directed timelike curve, γ from S, that never leaves the future Cauchy development of S. A rather similar argument shows that γ can be extended to the past to a curve that never leaves the past Cauchy development $D^-(S)$ (fig. 1.13). Now consider a sequence of points x_n on γ tending to the past and a similar sequence y_n tending to the future. For each value of n the points x_n and y_n are timelike separated and are in the globally hyperbolic Cauchy development of S. Thus, there is a timelike geodesic of maximum length λ_n from x_n to y_n. All the λ_n will cross the compact spacelike surface S. This means that there will be

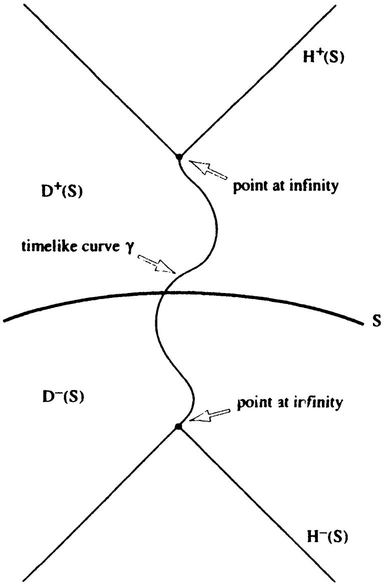


Figure 1.13 If the future (past) Cauchy development is not compact, there is a future (past) directed timelike curve from S that never leaves the future (past) Cauchy development.

a timelike geodesic λ in the Cauchy development which is a limit of the timelike geodesics λ_n (fig. 1.14). Either λ will be incomplete, in which case the theorem is proved, or it will contain conjugate points because of the generic energy condition. But in that case λ_n would contain conjugate points for n sufficiently large. This would be a contradiction because the λ_n are supposed to be curves of maximum length. One can therefore conclude that the spacetime is timelike or null geodesically incomplete. In other words, there is a singularity.

The theorems predict singularities in two situations. One is in the future in the gravitational collapse of stars and other massive bodies. Such singularities would be an end of time, at least for particles moving on the incomplete geodesics. The other situation in which singularities are predicted is in the past, at the beginning of the present expansion of the universe. This led to the abandonment

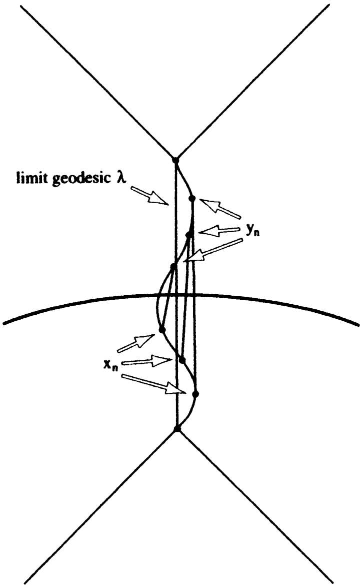


Figure 1.14 The geodesic λ, which is the limit of the γ_n, will have to be incomplete, because otherwise it would contain conjugate points.

of attempts (mainly by the Russians) to argue that there was a previous contracting phase and a nonsingular bounce into expansion. Instead, almost everyone now believes that the universe, and time itself, had a beginning at the big bang. This is a discovery far more important than a few miscellaneous unstable particles, but not one that has been so well recognized by Nobel prizes.

The prediction of singularities means that classical general relativity is not a complete theory. Because the singular points have to be cut out of the spacetime manifold, one cannot define the field equations there and cannot predict what will come out of a singularity. With the singularity in the past the only way to deal with this problem seems to be to appeal to quantum gravity. I shall return to this in my third lecture (Chapter 5). But the singularities that are predicted in the future seem to have a property that Penrose has called *cosmic censorship*. That is, they conveniently occur in places like black holes that are hidden from external observers. So any breakdown of predictability that may occur at these singularities won't affect what happens in the outside world, at least not according to classical theory.

Cosmic Censorship

Nature abhors a naked singularity.

However, as I shall show in my next lecture, there is unpredictability in the quantum theory. This is related to the fact that gravitational fields can have intrinsic entropy that is not just the result of coarse graining. Gravitational entropy, and the fact that time has a beginning and may have an end, are the two themes of my lectures because they are the ways in which gravity is distinctly different from other physical fields.

The fact that gravity has a quantity that behaves like entropy was first noticed in the purely classical theory. It depends on Penrose's *cosmic censorship conjecture*. This is unproved, but it is believed to be true for suitably general initial data and equations of state. I shall use a weak form of cosmic censorship. One makes the approximation of treating the region around a collapsing star as asymptotically flat. Then, as Penrose showed, one can conformally embed the spacetime manifold M in a manifold with boundary $\bar{M}$ (fig 1.15). The boundary ∂M will be a null surface and will consist of two components, future and past null infinity, called $\mathcal{I}^+$ and $\mathcal{I}^-$. I shall say that weak cosmic censorship holds if two conditions are satisfied. First, it is assumed that the null geodesic generators of $\mathcal{I}^+$ are complete in a certain conformal metric. This implies that observers far from the collapse live to an old age and are not wiped out by a thunderbolt singularity sent out from the collapsing star. Second, it is assumed that the past of $\mathcal{I}^+$ is globally hyperbolic. This means there are no naked singularities that can be seen from large distances. Penrose has a stronger form of cosmic censorship, which assumes that the whole spacetime is globally hyperbolic. But the weak form will suffice for my purposes.

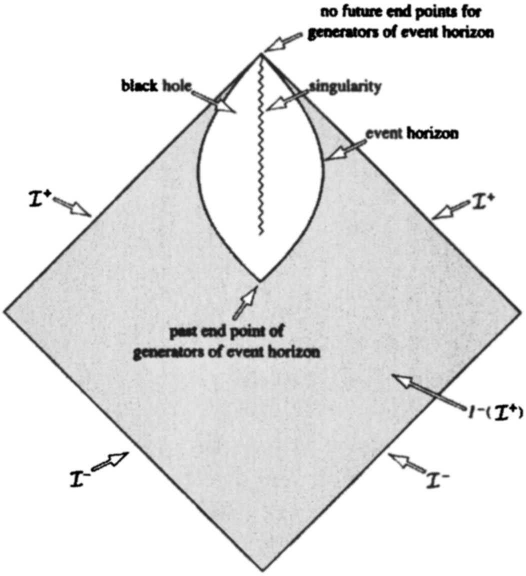


Figure 1.15 A collapsing star conformally embedded in a manifold with boundary.

Weak Cosmic Censorship

1. $\mathcal{I}^+$ and $\mathcal{I}^-$ are complete.
2. $I^-(\mathcal{I}^+)$ is globally hyperbolic.

If weak cosmic censorship holds, the singularities that are predicted to occur in gravitational collapse can't be visible from $\mathcal{I}^+$. This means that there must be a region of spacetime that is not in the past of $\mathcal{I}^+$. This region is said to be a black hole, because no light or anything else can escape from it to infinity. The boundary of the black hole region is called the *event horizon.* Because it is also the boundary of the past of $\mathcal{I}^+$, the event horizon will be generated by null geodesic segments that may have past endpoints but don't have any future endpoints. It then follows that if the weak energy condition holds,

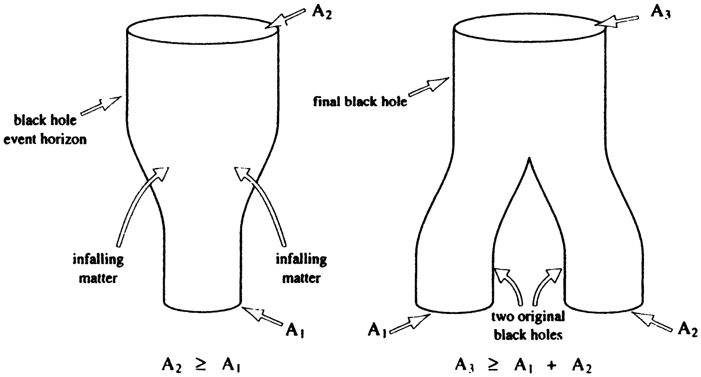


Figure 1.16 When we throw matter into a black hole, or allow two black holes to merge, the total area of the event horizons will never decrease.

the generators of the horizon can't be converging. For if they were, they would intersect each other within a finite distance.

This implies that the area of a cross section of the event horizon can never decrease with time, and in general will increase. Moreover, if two black holes collide and merge together, the area of the final black hole will be greater than the sum of the areas of the original black holes (fig. 1.16). This is very similar to the behavior of entropy according to the second law of thermodynamics. Entropy can never decrease and the entropy of a total system is greater than the sum of its constituent parts.

Second Law of Black Hole Mechanics

$$\delta A \geq 0$$

Second Law of Thermodynamics

$$\delta S \geq 0$$

First Law of Black Hole Mechanics

$$\delta E = \frac{\kappa}{8\pi}\delta A + \Omega\delta J + \Phi\delta Q$$

First Law of Thermodynamics

$$\delta E = T\delta S + P\delta V$$

The similarity with thermodynamics is increased by what is called the *first law of black hole mechanics.* This relates the change in mass of a black hole to the change in the area of the event horizon and the change in its angular momentum and electric charge. One can compare this to the first law of thermodynamics, which gives the change in internal energy in terms of the change in entropy and the external work done on the system. One sees that if the area of the event horizon is analogous to entropy, then the quantity analogous to temperature is what is called the surface gravity of the black hole κ. This is a measure of the strength of the gravitational field on the event horizon. The similarity with thermodynamics is further increased by the so-called *zeroth law of black hole mechanics*: the surface gravity is the same everywhere on the event horizon of a time-independent black hole.

Zeroth Law of Black Hole Mechanics

κ is the same everywhere on the horizon of a time-independent black hole.

Zeroth Law of Thermodynamics

T is the same everywhere for a system in thermal equilibrium.

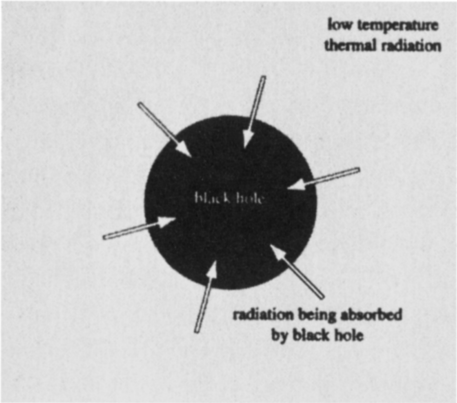


Figure 1.17 A black hole in contact with thermal radiation will absorb some of the radiation, but classically cannot send anything out.

Encouraged by these similarities, Bekenstein (1972) proposed that some multiple of the area of the event horizon actually was the entropy of a black hole. He suggested a generalized second law: the sum of this black hole entropy and the entropy of matter outside black holes would never decrease.

Generalized Second Law

$$\delta(S + cA) \geq 0$$

However, this proposal was not consistent. If black holes have an entropy proportional to horizon area, they should also have a nonzero temperature proportional to surface gravity. Consider a black hole that is in contact with thermal radiation at a temperature lower than the black hole temperature (fig. 1.17). The black hole will absorb some of the radiation but won't be able to send anything out, because

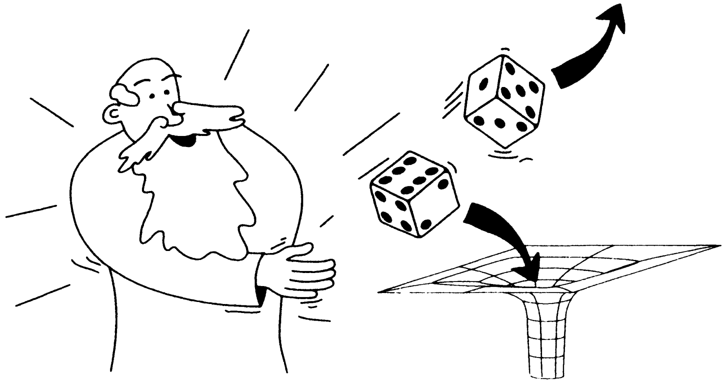


Figure 1.18

according to classical theory nothing can get out of a black hole. One thus has heat flow from the low-temperature thermal radiation to the higher-temperature black hole. This would violate the generalized second law because the loss of entropy from the thermal radiation would be greater than the increase in black hole entropy. However, as we shall see in my next lecture, consistency was restored when it was discovered that black holes are sending out radiation that was exactly thermal. This is too beautiful a result to be a coincidence or just an approximation. So it seems that black holes really do have intrinsic gravitational entropy. As I shall show, this is related to the nontrivial topology of a black hole. The intrinsic entropy means that gravity introduces an extra level of unpredictability over and above the uncertainty usually associated with quantum theory. So Einstein was wrong when he said, "God does not play dice." Consideration of black holes suggests, not only that God does play dice, but that he sometimes confuses us by throwing them where they can't be seen (fig. 1.18).