4
The physical significance of curvature

In this chapter we consider the effect of space–time curvature on families of timelike and null curves. These could represent flow lines of fluids or the histories of photons. In § 4.1 and § 4.2 we derive the formulae for the rate of change of vorticity, shear and expansion of such families of curves; the equation for the rate of change of expansion (Raychaudhuri's equation) plays a central role in the proofs of the singularity theorems of chapter 8. In § 4.3 we discuss the general inequalities on the energy–momentum tensor which imply that the gravitational effect of matter is always to tend to cause convergence of timelike and of null curves. A consequence of these energy conditions is, as is seen in § 4.4, that conjugate or focal points will occur in families of non-rotating timelike or null geodesics in general space–times. In § 4.5 it is shown that the existence of conjugate points implies the existence of variations of curves between two points which take a null geodesic into a timelike curve, or a timelike geodesic into a longer timelike curve.

4.1 Timelike curves

In chapter 3 we saw that if the metric was static there was a relation between the magnitude of the timelike Killing vector and the Newtonian potential. One was able to tell whether a body was in a gravitational field by whether, if released from rest, it would accelerate with respect to the static frame defined by the Killing vector. However, in general, space–time will not have any Killing vectors. Thus one will not have any special frame against which to measure acceleration; the best one can do is to take two bodies close together and measure their relative acceleration. This will enable one to measure the gradient of the gravitational field. If one thinks of the metric as being analogous to the Newtonian potential, the gradient of the Newtonian field would correspond to the second derivatives of the metric. These are described by the Riemann tensor. Thus one would expect that the relative

acceleration of two neighbouring bodies would be related to some components of the Riemann tensor.

In order to investigate this relation more precisely we shall examine the behaviour of a congruence of timelike curves with timelike unit tangent vector $\mathbf{V}$ ($g(\mathbf{V}, \mathbf{V}) = -1$). These curves could represent the histories of small test particles, in which case they would be geodesics, or they might represent the flow lines of a fluid. If this were a perfect fluid, then by (3.10)

$$(\mu + p)\, \dot{V}^a = -p_{;b}\, h^{ab}, \tag{4.1}$$

where $\dot{V}^a = V^a{}_{;b}\, V^b$ is the acceleration of the flow lines and $h^a{}_b = \delta^a{}_b + V^a V_b$ is the tensor which projects a vector $\mathbf{X} \in T_q$ into its component in the subspace H_q of T_q orthogonal to V. One may also think of h_{ab} as the metric in H_q (cf. §2.7).

Suppose $\lambda(t)$ is a curve with tangent vector $\mathbf{Z} = (\partial/\partial t)_\lambda$. Then one may construct a family $\lambda(t, s)$ of curves by moving each point of the curve $\lambda(t)$ a distance s along the integral curves of $\mathbf{V}$. If one now defines $\mathbf{Z}$ as $(\partial/\partial t)_{\lambda(t,s)}$ it follows from the definition of the Lie derivative (see §2.4) that $L_{\mathbf{V}}\mathbf{Z} = 0$ or in other words that

$$\frac{\mathrm{D}}{\partial s} Z^a = V^a{}_{;b} Z^b. \tag{4.2}$$

One may interpret $\mathbf{Z}$ as representing the separation of points equal distances from some arbitrary initial points along two neighbouring curves. If one adds a multiple of $\mathbf{V}$ to $\mathbf{Z}$ then this vector will represent the separation of points on the same two curves but at different distances along the curves. It is really only the separation of neighbouring curves that one is interested in, not the separation of particular points on these curves. One is thus concerned only with $\mathbf{Z}$ modulo a component parallel to $\mathbf{V}$, i.e. only with the projection of $\mathbf{Z}$ at each point q into the space Q_q consisting of equivalence classes of vectors which differ only by addition of a multiple of $\mathbf{V}$. This space can be represented as the subspace H_q of T_q consisting of vectors orthogonal to $\mathbf{V}$. The projection of $\mathbf{Z}$ into H_q will be denoted by ${}_\perp Z^a = h^a{}_b Z^b$. In the case of a fluid one can regard ${}_\perp\mathbf{Z}$ as the distance between two neighbouring particles of the fluid as measured in their rest frame.

From (4.2) it follows that

$${}_\perp\frac{\mathrm{D}}{\partial s}({}_\perp Z^a) = V^a{}_{;b}\, {}_\perp Z^b. \tag{4.3}$$

This gives the rate of change of the separation of two infinitesimally

neighbouring curves as measured in H_q. Operating again with $\mathrm{D}/\partial s$ and projecting into H_q, one finds

$$h^a{}_b \frac{\mathrm{D}}{\partial s}\left(h^b{}_c \frac{\mathrm{D}}{\partial s}{}_\perp Z^c\right) = h^a{}_b (V^b{}_{;cd}{}_\perp Z^c V^d + V^b{}_{;c} V^c{}_{;d} V_e Z^e V^d + V^b{}_{;c} V^c V^e{}_{;d} Z_e V^d + V^b{}_{;c} h^c{}_e Z^e{}_{;d} V^d).$$

Changing the order of the derivatives in the first term and using (4.2), this reduces to

$$h^a{}_b \frac{\mathrm{D}}{\partial s}\left(h^b{}_c \frac{\mathrm{D}}{\partial s}{}_\perp Z^c\right) = -R^a{}_{bcd}{}_\perp Z^c V^b V^d + h^a{}_b \dot{V}^b{}_{;c}{}_\perp Z^c + \dot{V}^a \dot{V}_b{}_\perp Z^b. \quad (4.4)$$

This equation, known as the deviation or Jacobi equation, gives the relative acceleration, i.e. the second time derivative of the separation, of two infinitesimally neighbouring curves as measured in H_q. We see that this depends only on the Riemann tensor if the curves are geodesics.

In Newtonian theory, the acceleration of each particle is given by the gradient of the potential Φ and therefore the relative acceleration of two particles with separation Z^a is $\Phi_{;ab} Z^b$. Thus the Riemann tensor term $R_{abcd} V^b V^d$ is analogous to the Newtonian $\Phi_{;ac}$. The effect of this 'tidal force' term can be seen, for example, by considering a sphere of particles freely falling towards the earth. Each particle moves on a straight line through the centre of the earth but those nearer the earth fall faster than those further away. This means that the sphere does not remain a sphere but is distorted into an ellipsoid with the same volume.

In order to investigate the deviation equation further we shall introduce dual orthonormal bases $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{E}_4$ and $\mathbf{E}^1$, $\mathbf{E}^2$, $\mathbf{E}^3$, $\mathbf{E}^4$ of T_q and T^*_q at some point q on an integral curve $\gamma(s)$ of $\mathbf{V}$, with $\mathbf{E}^4 = \mathbf{V}$. One would like to propagate them along $\gamma(s)$ to obtain similar such bases at each point of $\gamma(s)$. However, if one parallelly propagates them along $\gamma(s)$ (i.e. so that $\mathrm{D}/\partial s$ of each vector is zero) $\mathbf{E}_4$ will not remain equal to $\mathbf{V}$, and $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ will not remain orthogonal to $\mathbf{V}$, unless $\gamma(s)$ is a geodesic. We therefore introduce a new derivative along $\gamma(s)$ called the *Fermi derivative* $\mathrm{D}_\mathrm{F}/\partial s$. This is defined for a vector field $\mathbf{X}$ along $\gamma(s)$ by:

$$\frac{\mathrm{D}_\mathrm{F}\mathbf{X}}{\partial s} = \frac{\mathrm{D}\mathbf{X}}{\partial s} - g\left(\mathbf{X}, \frac{\mathrm{D}\mathbf{V}}{\partial s}\right)\mathbf{V} + g(\mathbf{X}, \mathbf{V})\frac{\mathrm{D}\mathbf{V}}{\partial s}.$$

It has the properties:

(i) $\dfrac{\mathrm{D_F}}{\partial s} = \dfrac{\mathrm{D}}{\partial s}$ if $\gamma(s)$ is a geodesic;

(ii) $\dfrac{\mathrm{D_F V}}{\partial s} = 0$;

(iii) if $\mathbf{X}$ and $\mathbf{Y}$ are vector fields along $\gamma(s)$ such that

$$\frac{\mathrm{D_F X}}{\partial s} = 0 = \frac{\mathrm{D_F Y}}{\partial s},$$

then $g(\mathbf{X}, \mathbf{Y})$ is constant along $\gamma(s)$;

(iv) if $\mathbf{X}$ is a vector field along $\gamma(s)$ orthogonal to $\mathbf{V}$ then

$$\frac{\mathrm{D_F X}}{\partial s} = {}_{\perp}\left(\frac{\mathrm{D X}}{\partial s}\right).$$

(This last property shows that the Fermi derivative is a natural generalization of the derivative $\mathrm{D}/\partial s$.)

Thus, if one propagates an orthonormal basis of T_q along $\gamma(s)$ so that the Fermi derivative of each basis vector is zero, one obtains an orthonormal basis at each point of $\gamma(s)$, with $\mathbf{E}_4 = \mathbf{V}$. The vectors $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ may be interpreted as giving a non-rotating set of axes along $\gamma(s)$. These could be realized physically by small gyroscopes pointing in the direction of each vector.

The definition of the Fermi derivative along $\gamma(s)$ can be extended from vector fields to arbitrary tensor fields by the usual rules:

(i) $\mathrm{D_F}/\partial s$ is a linear mapping of tensor fields of type (r, s) along $\gamma(s)$ to tensor fields of type (r, s), which commutes with contractions;

(ii) $\dfrac{\mathrm{D_F}}{\partial s}(\mathbf{K} \otimes \mathbf{L}) = \dfrac{\mathrm{D_F K}}{\partial s} \otimes \mathbf{L} + \mathbf{K} \otimes \dfrac{\mathrm{D_F L}}{\partial s}$;

(iii) $\dfrac{\mathrm{D_F} f}{\partial s} = \dfrac{\mathrm{d}f}{\mathrm{d}s}$, where f is a function.

From these rules it follows that the dual basis $\mathbf{E}^1$, $\mathbf{E}^2$, $\mathbf{E}^3$, $\mathbf{E}^4$ of ${T^*}_q$ is also Fermi-propagated along $\gamma(s)$. Using Fermi derivatives, (4.3) and (4.4) may be written as:

$$\frac{\mathrm{D_F}}{\partial s} {}_{\perp}Z^a = V^a{}_{;b} {}_{\perp}Z^b, \tag{4.5}$$

$$\frac{\mathrm{D^2_F}}{\partial s^2} {}_{\perp}Z^a = -R^a{}_{bcd} {}_{\perp}Z^c V^b V^d + h^a{}_b \dot{V}^b{}_{;c} {}_{\perp}Z^c + \dot{V}^a \dot{V}_b {}_{\perp}Z^b. \tag{4.6}$$

One may express these equations in terms of the Fermi-propagated

dual bases. As $\perp\mathbf{Z}$ is orthogonal to $\mathbf{V}$ it will have components with respect to $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ only. Thus it may be expressed as $Z^\alpha \mathbf{E}_\alpha$ where we adopt the convention that Greek indices take the values 1, 2, 3 only. Then (4.5) and (4.6) can be written in terms of ordinary derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}s} Z^\alpha = V^\alpha{}_{;\beta} Z^\beta, \tag{4.7}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} Z^\alpha = (-R^\alpha{}_{4\beta 4} + \dot{V}^\alpha{}_{;\beta} + \dot{V}^\alpha \dot{V}_\beta) Z^\beta \tag{4.8}$$

where $V^\alpha{}_{;\beta}$ are the components of $V^a{}_{;b}$ for which $a = \alpha$ and $b = \beta$. As the components Z^α obey the first order linear ordinary differential equation (4.7), they can be expressed in terms of their values at some point q by:

$$Z^\alpha(s) = A_{\alpha\beta}(s) Z^\beta|_q, \tag{4.9}$$

where $A_{\alpha\beta}(s)$ is a 3×3 matrix which is the unit matrix at q and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}s} A_{\alpha\beta}(s) = V_{\alpha;\gamma} A_{\gamma\beta}(s). \tag{4.10}$$

In the case of a fluid the matrix $A_{\alpha\beta}$ can be regarded as representing the shape and orientation of a small element of fluid which is spherical at q. This matrix can be written as

$$A_{\alpha\beta} = O_{\alpha\delta} S_{\delta\beta} \tag{4.11}$$

where $O_{\alpha\beta}$ is an orthogonal matrix with positive determinant and $S_{\alpha\beta}$ is a symmetric matrix. These will both be chosen to be the unit matrix at q. The matrix $O_{\alpha\beta}$ may be thought of as representing the rotation that neighbouring curves have undergone with respect to the Fermi-propagated basis while $S_{\alpha\beta}$ represents the separation of these curves from $\gamma(s)$. The determinant of $S_{\alpha\beta}$, which equals the determinant of $A_{\alpha\beta}$, may be thought of as representing the three-volume of the element of the surface orthogonal to $\gamma(s)$ marked out by the neighbouring curves.

At q where $A_{\alpha\beta}$ is the unit matrix, $\mathrm{d}O_{\alpha\beta}/\mathrm{d}s$ is antisymmetric and $\mathrm{d}S_{\alpha\beta}/\mathrm{d}s$ is symmetric. Thus the rate of rotation of neighbouring curves at q is given by the antisymmetric part of $V_{\alpha;\beta}$ while the rate of change of their separation from $\gamma(s)$ is given by the symmetric part of $V_{\alpha;\beta}$ and the rate of change of volume is given by the trace of $V_{\alpha;\beta}$. We therefore define the vorticity tensor as

$$\omega_{ab} = h_a{}^c h_b{}^d V_{[c;d]}, \tag{4.12}$$

the expansion tensor as

$$\theta_{ab} = h_a{}^c h_b{}^d V_{(c;d)}, \tag{4.13}$$

and the volume expansion as

$$\theta = \theta_{ab} h^{ab} = V_{a;\,b}\, h^{ab} = V^{a}{}_{;\,a}. \tag{4.14}$$

We further define the shear tensor as the trace free part of θ_{ab},

$$\sigma_{ab} = \theta_{ab} - \tfrac{1}{3} h_{ab} \theta, \tag{4.15}$$

and the vorticity vector as

$$\omega^{a} = \tfrac{1}{2} \eta^{abcd} V_{b} \omega_{cd} = \tfrac{1}{2} \eta^{abcd} V_{b} V_{c;\,d}. \tag{4.16}$$

The covariant derivative of the vector $\mathbf{V}$ may be expressed in terms of these quantities;

$$V_{a;\,b} = \omega_{ab} + \sigma_{ab} + \tfrac{1}{3} \theta h_{ab} - \dot{V}_{a} V_{b}. \tag{4.17}$$

This decomposition of the gradient of the fluid velocity vector is directly analogous to that in Newtonian hydrodynamics.

In the Fermi-propagated orthonormal basis the vorticity and expansion can be expressed in terms of the matrix $A_{\alpha\beta}$ and its inverse $A^{-1}{}_{\alpha\beta}$:

$$\omega_{\alpha\beta} = - A^{-1}{}_{\gamma[\alpha} \frac{\mathrm{d}}{\mathrm{d}s} A_{\beta]\gamma}, \tag{4.18}$$

$$\theta_{\alpha\beta} = A^{-1}{}_{\gamma(\alpha} \frac{\mathrm{d}}{\mathrm{d}s} A_{\beta)\gamma}, \tag{4.19}$$

$$\theta = (\det \mathbf{A})^{-1} \frac{\mathrm{d}}{\mathrm{d}s} (\det \mathbf{A}). \tag{4.20}$$

From the deviation equation (4.8) it follows that

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} A_{\alpha\beta} = (- R_{\alpha 4 \gamma 4} + \dot{V}_{\alpha;\,\gamma} + \dot{V}_{\alpha} \dot{V}_{\gamma}) A_{\gamma\beta}. \tag{4.21}$$

This equation enables one to calculate the propagation of the vorticity, shear and expansion along the integral curves of $\mathbf{V}$ if one knows the Riemann tensor.

Multiplying by $A^{-1}{}_{\beta\gamma}$ and taking the antisymmetric part, one obtains

$$\frac{\mathrm{d}}{\mathrm{d}s} \omega_{\alpha\beta} = 2 \omega_{\gamma[\alpha} \theta_{\beta]\gamma} + \dot{V}_{[\alpha;\,\beta]}. \tag{4.22}$$

Thus the propagation of vorticity depends on the antisymmetric gradient of the acceleration but not the 'tidal force'. Another form of the above equation is

$$\frac{\mathrm{d}}{\mathrm{d}s} (A_{\gamma\alpha} \omega_{\gamma\delta} A_{\delta\beta}) = A_{\gamma\alpha} \dot{V}_{[\gamma;\,\delta]} A_{\delta\beta}. \tag{4.23}$$

Therefore $A_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta}$ is a constant matrix if the curves are geodesics; in particular, if the curves are geodesics and the vorticity vanishes at one point on a curve, it will vanish at all points on the curve. If the curves are the flow lines of a perfect fluid it follows from (4.1) that

$$\dot{V}_{[\alpha;\beta]} = -\frac{1}{\mu+p}\omega_{\alpha\beta}\frac{\mathrm{d}p}{\mathrm{d}s}.$$

If the fluid is isentropic, this implies the conservation law:

$$WA_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta} = \text{constant}, \tag{4.24}$$

where

$$\log W = \int \frac{\mathrm{d}p}{\mu+p}.$$

This conservation law is the relativistic form of the Newtonian vorticity conservation law. In the geodesic or pressure-free case, this takes the usual form that the magnitude of the vorticity vector is inversely proportional to the area of a cross-section orthogonal to the vorticity vector of an element of the fluid. When the pressure is non-zero, there is an extra relativistic effect arising from the fact that compression of the fluid does work on the fluid and therefore increases the mass and so the inertia of an element of the fluid (cf. (3.9)). This means that the vorticity of a fluid increases less under compression than would otherwise be expected.

Multiplying (4.21) by $A^{-1}{}_{\beta\gamma}$ and taking the symmetric part, one finds

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta_{\alpha\beta} = -R_{\alpha4\beta4} - \omega_{\alpha\gamma}\omega_{\gamma\beta} - \theta_{\alpha\gamma}\theta_{\gamma\beta} + \dot{V}_{(\alpha;\beta)} + \dot{V}_\alpha\dot{V}_\beta. \tag{4.25}$$

(This equation and (4.23) can be expressed in terms of a general, non-orthonormal, non-Fermi-propagated basis by replacing the ordinary derivatives with Fermi derivatives and projecting everything into the subspace orthogonal to **V**.)

The trace of (4.25) is

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -R_{ab}V^aV^b + 2\omega^2 - 2\sigma^2 - \tfrac{1}{3}\theta^2 + \dot{V}^a{}_{;a}, \tag{4.26}$$

where

$$2\omega^2 = \omega_{ab}\omega^{ab} \geqslant 0,$$

$$2\sigma^2 = \sigma_{ab}\sigma^{ab} \geqslant 0.$$

This equation, which was discovered by Landau and independently by Raychaudhuri, will be of great importance later. From it one sees that vorticity induces expansion as might be expected by analogy with

centrifugal force while shear induces contraction. By the field equations, the term $R_{ab}V^aV^b = 4\pi(\mu + 3p)$ for a perfect fluid whose flow lines have tangent vectors V^a. Thus one would expect this term also to induce contraction. We shall give a general discussion of the sign of this term in §4.3.

The trace-free part of (4.25) is

$$\frac{\mathrm{D_F}}{\partial s}\sigma_{ab} = -C_{acbd}V^cV^d + \tfrac{1}{2}h_a{}^c h_b{}^d R_{cd} - \omega_{ac}\omega^c{}_b - \sigma_{ac}\sigma^c{}_b$$
$$-\tfrac{2}{3}\theta\sigma_{ab} + h_a{}^c h_b{}^d \dot{V}_{(c;\,d)} - \tfrac{1}{3}h_{ab}(2\omega^2 - 2\sigma^2 + \dot{V}^a{}_{;\,a} + \tfrac{1}{2}R_{cd}h^{cd}), \quad (4.27)$$

where C_{abcd} is the Weyl tensor. Since this tensor is trace-free it does not enter directly in the expansion equation (4.26). However since the term $-2\sigma^2$ occurs on the right of the expansion equation, the Weyl tensor produces convergence indirectly by inducing shear. The Riemann tensor can be expressed in terms of the Weyl tensor and the Ricci tensor:

$$R_{abcd} = C_{abcd} - g_{a[d}R_{c]b} - g_{b[c}R_{d]a} - \tfrac{1}{3}Rg_{a[c}g_{d]b}.$$

The Ricci tensor is given by the Einstein equations:

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

Thus the Weyl tensor is that part of the curvature which is not determined locally by the matter distribution. However it cannot be entirely arbitrary as the Riemann tensor must satisfy the Bianchi identities:

$$R_{ab[cd;\,e]} = 0$$

These can be rewritten as

$$C^{abcd}{}_{;\,d} = J^{abc}, \quad (4.28)$$

where

$$J^{abc} = R^{c[a;\,b]} + \tfrac{1}{6}g^{c[b}R^{;\,a]}. \quad (4.29)$$

These equations are rather similar to Maxwell's equations in electrodynamics:

$$F^{ab}{}_{;\,b} = J^a,$$

where F^{ab} is the electromagnetic field tensor and J^a is the source current. Thus in a sense one could regard the Bianchi identities (4.28) as field equations for the Weyl tensor giving that part of the curvature at a point that depends on the matter distribution at other points. (This approach has been used to analyse the behaviour of gravitational radiation in papers by Newman and Penrose (1962), Newman and Unti (1962) and Hawking (1966*a*).)

4.2 Null curves

The Riemann tensor will affect the rate of change of separation of null curves as well as that of timelike curves. For simplicity, we shall consider only null geodesics. These could represent the histories of photons; the effect of the Riemann tensor will be to distort or focus small bundles of light rays.

To investigate this, we consider the deviation equation for a congruence of null geodesics with tangent vector $\mathbf{K}$ ($g(\mathbf{K}, \mathbf{K}) = 0$). There are two important differences between this case and that of the timelike curves considered in the previous section. First, one could normalize the tangent vector $\mathbf{V}$ to the timelike curves by requiring $g(\mathbf{V}, \mathbf{V}) = -1$. In effect this means that one parametrized the curves by the arc-length s. However this is clearly impossible with null curves as they have zero arc-lengths. The best one can do is to choose an affine parameter v; then the tangent vector $\mathbf{K}$ will obey

$$\frac{\mathrm{D}}{\mathrm{d}v} K^a = K^a{}_{;b} K^b = 0.$$

However one could multiply v by a function f which was constant along each curve. Then fv would be another affine parameter and the corresponding tangent vector would be $f^{-1}\mathbf{K}$. Thus, given the curves as point sets in the manifold, the tangent vector is only really unique up to a constant factor along each curve. The second difference is that Q_q, the quotient of T_q by $\mathbf{K}$, is not now isomorphic to H_q, the subspace of T_q orthogonal to $\mathbf{K}$, since H_q includes the vector $\mathbf{K}$ itself as $g(\mathbf{K}, \mathbf{K}) = 0$. In fact as will be shown below, one is not really interested in the whole of Q_q but only in the subspace S_q consisting of equivalence classes of vectors in H_q which differ only by a multiple of $\mathbf{K}$. In the case of light rays, one can regard an element of S_q as representing the separation between two neighbouring light rays which were emitted at the same time by a source.

As before we introduce dual bases $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{E}_4$, and $\mathbf{E}^1$, $\mathbf{E}^2$, $\mathbf{E}^3$, $\mathbf{E}^4$ of T_q and T_q^* at some point q on a curve $\gamma(v)$. However we will not choose them to be orthonormal. We take $\mathbf{E}_4$ equal to $\mathbf{K}$, $\mathbf{E}_3$ to be some other null vector $\mathbf{L}$ having unit negative scalar product with $\mathbf{E}_4$ ($g(\mathbf{E}_3, \mathbf{E}_3,) = 0$, $g(\mathbf{E}_3, \mathbf{E}_4) = -1$) and $\mathbf{E}_1$ and $\mathbf{E}_2$ to be unit spacelike vectors, orthogonal to each other and to $\mathbf{E}_3$ and $\mathbf{E}_4$

$$(g(\mathbf{E}_1, \mathbf{E}_1) = g(\mathbf{E}_2, \mathbf{E}_2) = 1, \quad g(\mathbf{E}_1, \mathbf{E}_2) = g(\mathbf{E}_1, \mathbf{E}_3) = g(\mathbf{E}_1, \mathbf{E}_4) = 0, \text{etc.}).$$

Note that because of the non-orthonormal character of the basis, the form $\mathbf{E}^3$ is in fact equal to the form $-K^a g_{ab}$ and $\mathbf{E}^4$ is $-L^a g_{ab}$. It can be seen that $\mathbf{E}_1$, $\mathbf{E}_2$ and $\mathbf{E}_4$ constitute a basis for H_q while the projections into Q_q of $\mathbf{E}_1$, $\mathbf{E}_2$ and $\mathbf{E}_3$ form a basis of Q_q, and the projections of $\mathbf{E}_1$ and $\mathbf{E}_2$ form a basis of S_q. We shall normally not distinguish between a vector $\mathbf{Z}$ and its projection into Q_q or S_q. We shall call a basis having the properties of $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{E}_4$, above, *pseudo-orthonormal*. By parallelly transporting them along the geodesic $\gamma(v)$ one obtains a pseudo-orthonormal basis at each point of $\gamma(v)$.

We use this basis to analyse the deviation equation for null geodesics. If $\mathbf{Z}$ is the vector representing the separation of corresponding points on neighbouring curves, one has, as before:

$$L_{\mathbf{K}}\mathbf{Z} = 0,$$

so
$$\frac{\mathrm{D}}{\mathrm{d}v} Z^a = K^a{}_{;b} Z^b \tag{4.30}$$

and
$$\frac{\mathrm{D}^2}{\mathrm{d}v^2} Z^a = -R^a{}_{bcd} Z^c K^b K^d. \tag{4.31}$$

In the pseudo-orthonormal basis $K^a{}_{;4}$ will be zero as $\mathbf{K}$ is geodesic. Therefore one can express the 1, 2 and 3 components of (4.30) as a system of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}v} Z^\alpha = K^\alpha{}_{;\beta} Z^\beta,$$

where as before Greek indices take the values 1, 2, 3. This shows that the projection of $\mathbf{Z}$ into the space Q_q obeys a propagation equation which involves only this projection, and not the component of $\mathbf{Z}$ parallel to $\mathbf{K}$. Further $K^3{}_{;c} = 0$ since $(K^a g_{ab} K^b)_{;c} = 0$. This implies that $Z^3 = -Z^a K_a$ is constant along the geodesic $\gamma(v)$. This can be interpreted as saying that light rays emitted from the same source at different times maintain a constant separation in time. As this is the case, one is more interested in the behaviour of neighbouring null geodesics which have purely spatial separations, i.e. one is interested in vectors $\mathbf{Z}$ for which $Z^3 = 0$. The projections of such vectors will then lie in the subspace S_q and will obey the equation

$$\frac{\mathrm{d}}{\mathrm{d}v} Z^m = K^m{}_{;n} Z^n,$$

where m, n take the values 1, 2 only. This is similar to (4.7) for the timelike case, except that now one is dealing only with a two-dimensional space of connecting vectors $\mathbf{Z}$.

As in the previous section, one can express Z^m in terms of their values at some point q:

$$Z^m(v) = \hat{A}_{mn}(v)\, Z^n|_q,$$

where $\hat{A}_{mn}(v)$ is a 2×2 matrix which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}v}\hat{A}_{mn}(v) = K_{m;\,p}\,\hat{A}_{pn}(v), \tag{4.32}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}v^2}\hat{A}_{mn}(v) = -R_{m4p4}\,\hat{A}_{pn}(v). \tag{4.33}$$

As before we call the antisymmetric part of $K_{m;\,n}$ the vorticity $\hat{\omega}_{mn}$, the symmetric part the rate of separation $\hat{\theta}_{mn}$ and the trace the expansion $\hat{\theta}$. We also define the shear $\hat{\sigma}_{mn}$ as the trace-free part of $\hat{\theta}_{mn}$. They obey similar equations to the analogous quantities in the previous section:

$$\frac{\mathrm{d}}{\mathrm{d}v}\hat{\omega}_{mn} = -\hat{\theta}\hat{\omega}_{mn} + 2\hat{\omega}_{p[m}\hat{\sigma}_{n]p}, \tag{4.34}$$

$$\frac{\mathrm{d}}{\mathrm{d}v}\hat{\theta} = -R_{ab}K^aK^b + 2\hat{\omega}^2 - 2\hat{\sigma}^2 - \tfrac{1}{2}\hat{\theta}^2, \tag{4.35}$$

$$\frac{\mathrm{d}}{\mathrm{d}v}\hat{\sigma}_{mn} = -C_{m4n4} - \hat{\theta}\hat{\sigma}_{mn} - \hat{\sigma}_{mp}\hat{\sigma}_{pn} - \hat{\omega}_{mp}\hat{\omega}_{pn} + \delta_{mn}(\hat{\sigma}^2 - \hat{\omega}^2). \tag{4.36}$$

Equation (4.35) is the analogue of the Raychaudhuri equation for timelike geodesics. One sees again that vorticity causes expansion while shear causes contraction. We shall show in the next section that the Ricci tensor term $-R_{ab}K^aK^b$ will normally be negative, and so cause focussing. As before the Weyl tensor does not affect the expansion directly but causes distortion which in turn causes contraction (cf. Penrose (1966)).

4.3 Energy conditions

In the actual universe the energy–momentum tensor will be made up of contributions from a large number of different matter fields. It would therefore be impossibly complicated to describe the exact energy–momentum tensor even if one knew the precise form of the contribution of each field and the equations of motion governing it. In fact, one has little idea of the behaviour of matter under extreme conditions of density and pressure. Thus it might seem that one has little hope of predicting the occurrence of singularities in the universe from the Einstein equations as one does not know the right-hand side

of the equations. However there are certain inequalities which it is physically reasonable to assume for the energy–momentum tensor. These will be discussed in this section. It turns out that in many circumstances these are sufficient to prove the occurrence of singularities, independent of the exact form of the energy–momentum tensor.

The first of these inequalities is:

The weak energy condition

The energy–momentum tensor at each $p \in \mathscr{M}$ obeys the inequality $T_{ab} W^a W^b \geqslant 0$ for any timelike vector $\mathbf{W} \in T_p$. By continuity this will then also be true for any null vector $\mathbf{W} \in T_p$.

To an observer whose world-line at p has unit tangent vector $\mathbf{V}$, the local energy density appears to be $T_{ab} V^a V^b$. Thus this assumption is equivalent to saying that the energy density as measured by any observer is non-negative. This would seem very reasonable physically. To investigate further the significance of this assumption we use the fact that one may express the components T^{ab} of the energy–momentum tensor at p with respect to an orthonormal basis $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{E}_4$, ($\mathbf{E}_4$ timelike) in one of four canonical forms.

Type I.

$$T^{ab} = \begin{pmatrix} p_1 & & & \\ & p_2 & & 0 \\ & & p_3 & \\ & 0 & & \\ & & & \mu \end{pmatrix}.$$

This is the general case in which the energy–momentum tensor has a timelike eigenvector $\mathbf{E}_4$. This eigenvector is unique unless $\mu = -p_\alpha$ ($\alpha = 1, 2, 3$). The eigenvalue μ represents the energy–density as measured by an observer whose world-line at p has unit tangent vector $\mathbf{E}_4$ and the eigenvalues p_α ($\alpha = 1, 2, 3$) represent the principal pressures in the three spacelike directions $\mathbf{E}_\alpha$. This is the form of the energy–momentum for all observed fields with non-zero rest mass and also for all zero rest mass fields except in special cases when it is type II.

Type II.

$$T^{ab} = \begin{pmatrix} p_1 & 0 & & \\ 0 & p_2 & & 0 \\ & & \nu - \kappa & \nu \\ & 0 & & \\ & & \nu & \nu + \kappa \end{pmatrix}, \quad \nu = \pm 1.$$

This is the special case in which the energy–momentum tensor has a double null eigenvector $(\mathbf{E}_3+\mathbf{E}_4)$. The only observed occurrence of this form is for zero rest-mass fields when they represent radiation all of which is travelling in the direction $\mathbf{E}_3+\mathbf{E}_4$. In this case p_1, p_2 and κ are zero.

Type III.

$$T^{ab} = \begin{pmatrix} p & 0 & 0 & 0 \\ 0 & -\nu & 1 & 1 \\ 0 & 1 & -\nu & 0 \\ 0 & 1 & 0 & \nu \end{pmatrix}.$$

This is the special case in which the energy–momentum tensor has a triple null eigenvector $(\mathbf{E}_3+\mathbf{E}_4)$. There are no observed fields which have energy–momentum tensors of this form.

Type IV.

$$T^{ab} = \begin{pmatrix} p_1 & 0 & & \\ 0 & p_2 & & \\ & & -\kappa & \nu \\ & & \nu & 0 \end{pmatrix}, \quad \kappa^2 < 4\nu^2.$$

This is the general case in which the energy–momentum tensor has no timelike or null eigenvector. There are no observed fields which have energy–momentum tensors of this form.

For type I, the weak energy condition will hold if $\mu \geqslant 0$, $\mu+p_\alpha \geqslant 0$ $(\alpha = 1, 2, 3)$. For type II it will hold if $p_1 \geqslant 0$, $p_2 \geqslant 0$, $\kappa \geqslant 0$, $\nu = +1$. These inequalities are very reasonable requirements and are satisfied by all experimentally detected fields. The condition will not hold for the physically unrealized types III and IV.

The condition will also hold for the scalar field ϕ postulated by Brans and Dicke and by Dicke (see Dicke (1964)). This field is required to be positive everywhere. It has an energy–momentum tensor of the form (3.6) where now $m = 0$. The energy-tensor of the other fields is ϕ times what it would have been had the scalar field not existed.

The condition will not hold for the 'C'-field proposed by Hoyle and Narlikar (1963). This again is a scalar field with m zero, only this time the energy–momentum tensor has the opposite sign and so the energy density is negative. This allows the simultaneous creation of quanta of positive energy fields and of the negative energy C-field. This process occurs in the steady-state model of the universe suggested by Hoyle

and Narlikar in which, as particles move apart due to the general expansion of the universe, new matter is continually being created to keep the average density constant. There is, however, a quantum mechanical difficulty associated with such a process. For even if the cross-section for the process were very small, the infinite phase space available to the positive and negative energy quanta would seem to result in an infinite number of such pairs being produced in a finite region of space–time.

Such a catastrophe could not occur if the weak energy condition held. If a slightly stronger condition holds then creation is impossible in the sense that space–time must remain empty if it is empty at one time and no matter comes in from infinity. Conversely, matter present at one time cannot disappear and so must be present at another time. The condition is

The dominant energy condition

For every timelike W_a, $T^{ab}W_aW_b \geqslant 0$, and $T^{ab}W_a$ is a non-spacelike vector.

This may be interpreted as saying that to any observer the local energy density appears non-negative and the local energy flow vector is non-spacelike. An equivalent statement is that in any orthonormal basis the energy dominates the other components of T_{ab}, i.e.

$$T^{00} \geqslant |T^{ab}| \quad \text{for each } a, b.$$

This holds for type I if $\mu \geqslant 0$, $-\mu \leqslant p_\alpha \leqslant \mu$ $(\alpha = 1, 2, 3)$ and for type II if $\nu = +1$, $\kappa \geqslant 0$, $0 \leqslant p_i \leqslant \kappa$ $(i = 1, 2)$. In other words, the dominant energy condition is the weak energy condition with the additional requirement that the pressure should not exceed the energy density. This holds for all known forms of matter and there is in fact good reason for believing that this should be the case in all situations. For the speed of sound waves travelling in the $\mathbf{E}_\alpha$ direction is $\mathrm{d}p_\alpha/\mathrm{d}\mu$ (adiabatic) times the speed of light. Thus $\mathrm{d}p_\alpha/\mathrm{d}\mu$ must be less than or equal to one, as by postulate (a) in § 3.2 no signal can propagate faster than light. It follows that $p_\alpha \leqslant \mu$, since, for every known form of matter, the pressures are small when the density is small. (Bludman and Ruderman (1968, 1970) have shown that there might be fields for which mass renormalization could lead to pressure being greater than the density. We feel, however, that this probably indicates a failure of renormalization theory rather than that such a situation would occur.) Now consider the situation depicted in figure 9 in which there is a C^2

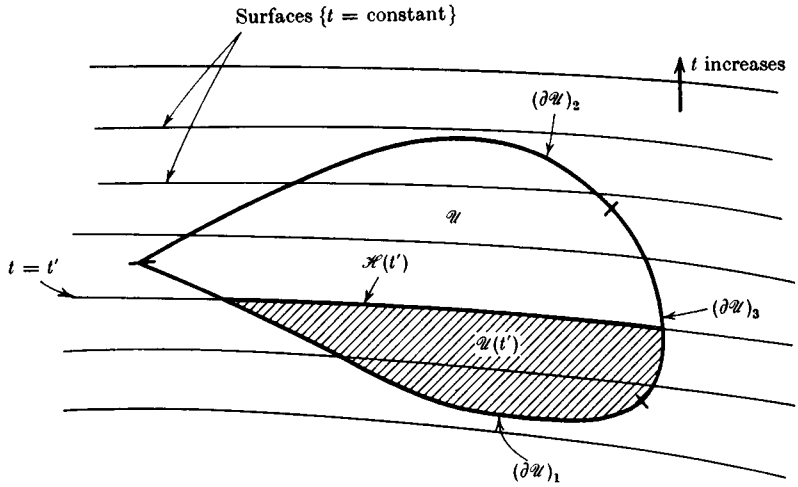

FIGURE 9. A compact region $\mathscr{U}$ of space–time with past and future non-timelike boundaries $(\partial\mathscr{U})_1$, $(\partial\mathscr{U})_2$ and timelike boundary $(\partial\mathscr{U})_3$. The part of $\mathscr{U}$ lying to the past of the surface $\mathscr{H}(t')$ (defined by $t = t'$) is $\mathscr{U}(t')$.

function t whose gradient is everywhere timelike. (It will be shown in § 6.4 that such a function will exist provided space–time is not on the verge of violating causality.) The boundary $\partial\mathscr{U}$ of the compact region $\mathscr{U}$ consists of a part $(\partial\mathscr{U})_1$, whose normal form $\mathbf{n}$ is non-spacelike and such that $n_a t_{;b} g^{ab}$ is positive, a part $(\partial\mathscr{U})_2$ whose normal form $\mathbf{n}$ is non-spacelike and such that $n_a t_{;b} g^{ab}$ is negative, and a remaining part $(\partial\mathscr{U})_3$ (which may be empty). The sign of the normal form $\mathbf{n}$ is given by the requirement that $\langle \mathbf{n}, \mathbf{X} \rangle$ be positive for all vectors $\mathbf{X}$ which point out of $\mathscr{U}$ (cf. § 2.8), $\mathscr{H}(t')$ denotes the surface $t = t'$ and $\mathscr{U}(t')$ denotes the region of $\mathscr{U}$ for which $t < t'$. For later use in § 7.4 we shall establish an inequality which holds not only for the energy–momentum tensor T^{ab} but also for any symmetric tensor S^{ab} which satisfies the dominant energy condition. Applied to the energy–momentum tensor this inequality will show that T^{ab} vanishes everywhere on $\mathscr{U}$ if it vanishes on $(\partial\mathscr{U})_3$ and on the initial surface $(\partial\mathscr{U})_1$.

Lemma 4.3.1

There is some positive constant P such that for any tensor S^{ab} which satisfies the dominant energy condition and vanishes on $(\partial\mathscr{U})_3$,

$$\int_{\mathscr{H}(t)\cap\mathscr{U}} S^{ab} t_{;a}\, d\sigma_b \leqslant -\int_{(\partial\mathscr{U})_1} S^{ab} t_{;a}\, d\sigma_b + P\int^t \left(\int_{\mathscr{H}(t')\cap\mathscr{U}} S^{ab} t_{;a}\, d\sigma_b\right) dt' + \int^t \left(\int_{\mathscr{H}(t')\cap\mathscr{U}} S^{ab}{}_{;a}\, d\sigma_b\right) dt'.$$

Consider the volume integral

$$I(t) = \int_{\mathscr{U}(t)} (S^{ab}t_{;a})_{;b}\,\mathrm{d}v = \int_{\mathscr{U}(t)} S^{ab}t_{;ab}\,\mathrm{d}v + \int_{\mathscr{U}(t)} S^{ab}{}_{;b}t_{;a}\,\mathrm{d}v.$$

By Gauss' theorem this can be transformed into an integral over the boundary of $\mathscr{U}(t)$:

$$I(t) = \int_{\partial\mathscr{U}(t)} S^{ab}t_{;a}\,\mathrm{d}\sigma_b.$$

The boundary of $\mathscr{U}(t)$ will consist of $\mathscr{U}(t) \cap \partial\mathscr{U}$ and $\mathscr{U} \cap \mathscr{H}(t)$. Since S^{ab} is zero on $(\partial\mathscr{U})_3$,

$$I(t) = \int_{\mathscr{U}(t) \cap (\partial\mathscr{U})_1} + \int_{\mathscr{U}(t) \cap (\partial\mathscr{U})_2} + \int_{\mathscr{U} \cap \mathscr{H}(t)}.$$

By the dominant energy condition, $S^{ab}t_{;a}$ is a non-spacelike vector such that $S^{ab}t_{;a}t_{;b} \geqslant 0$. As the normal form to $(\partial\mathscr{U})_2$ is non-spacelike and such that $n_a t_{;b} g^{ab} < 0$, the second term on the right will be non-negative. Thus

$$\int_{\mathscr{U} \cap \mathscr{H}(t)} S^{ab}t_{;a}\,\mathrm{d}\sigma_b \leqslant -\int_{\mathscr{U}(t) \cap (\partial\mathscr{U})_1} S^{ab}t_{;a}\,\mathrm{d}\sigma_b + \int_{\mathscr{U}(t)} (S^{ab}t_{;ab} + S^{ab}{}_{;b}t_{;a})\,\mathrm{d}v.$$

Since $\mathscr{U}$ is compact there will be some upper bound to the components of $t_{;ab}$ in any orthonormal basis whose timelike vector is in the direction of $t_{;a}$. Thus there will be some $P > 0$ such that on $\mathscr{U}$,

$$S^{ab}t_{;ab} \leqslant PS^{ab}t_{;a}t_{;b}$$

for any S^{ab} which obeys the dominant energy condition. The volume integral over $\mathscr{U}(t)$ can be decomposed into a surface integral over $\mathscr{H}(t') \cap \mathscr{U}$ followed by an integral with respect to t':

$$\int_{\mathscr{U}(t)} (PS^{ab}t_{;a}t_{;b} + S^{ab}{}_{;b}t_{;a})\,\mathrm{d}v = \int^t \left\{ \int_{\mathscr{H}(t') \cap \mathscr{U}} (PS^{ab}t_{;b} + S^{ab}{}_{;b})\,\mathrm{d}\sigma_a \right\} \mathrm{d}t',$$

where $\mathrm{d}\sigma_a$ is the surface element of $\mathscr{H}(t')$. Thus

$$\int_{\mathscr{H}(t) \cap \mathscr{U}} S^{ab}t_{;a}\,\mathrm{d}\sigma_b \leqslant -\int_{\mathscr{U}(t) \cap (\partial\mathscr{U})_1} S^{ab}t_{;a}\,\mathrm{d}\sigma_b + P\int^t \left(\int_{\mathscr{H}(t') \cap \mathscr{U}} S^{ab}t_{;a}\,\mathrm{d}\sigma_b \right) \mathrm{d}t' + \int^t \left(\int_{\mathscr{H}(t') \cap \mathscr{U}} S^{ab}{}_{;a}\,\mathrm{d}\sigma_b \right) \mathrm{d}t'. \quad \square$$

As an immediate consequence of this result one has:

The conservation theorem

If the energy–momentum tensor obeys the dominant energy condition and is zero on $(\partial\mathscr{U})_3$ and on the initial surface $(\partial\mathscr{U})_1$, then it is zero everywhere on $\mathscr{U}$.

Let
$$x(t) = \int_{\mathscr{U}(t)} T^{ab} t_{;a} t_{;b} \, dv$$
$$= \int^t \left(\int_{\mathscr{H}(t') \cap \mathscr{U}} T^{ab} t_{;a} \, d\sigma_b \right) dt' \geqslant 0.$$

Then the above lemma gives $dx/dt \leqslant Px$. But for sufficiently early values of t, $\mathscr{H}(t)$ will not intersect $\mathscr{U}$ and so x will vanish. Thus x will vanish for all t which implies that T^{ab} is zero on $\mathscr{U}$. □

From the conservation theorem it follows that if the energy–momentum tensor vanishes on a set $\mathscr{S}$, then it also vanishes on the

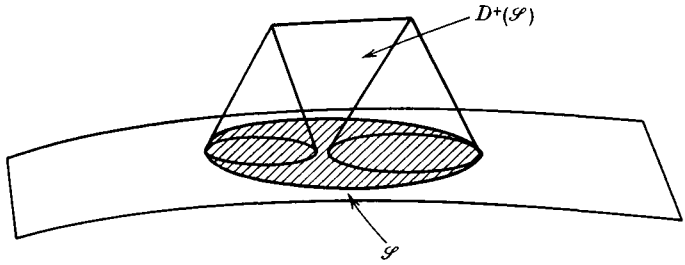

FIGURE 10. The future Cauchy development $D^+(\mathscr{S})$ of a spacelike set $\mathscr{S}$.

future Cauchy development $D^+(\mathscr{S})$, which is defined as the set of all points through which every past-directed non-spacelike curve intersects $\mathscr{S}$ (figure 10) (cf. § 6.5). For if q is any point of $D^+(\mathscr{S})$, the region of $D^+(\mathscr{S})$ to the past of q is compact (proposition 6.6.6) and may be taken as $\mathscr{U}$. This result may be interpreted as saying that the dominant energy condition implies that matter cannot travel faster than light.

For our consideration of singularities, the importance of the weak energy condition is that it implies that matter always has a converging (or more strictly nondiverging) effect on congruences of null geodesics. If the vorticity vanishes, the expansion $\hat{\theta}$ obeys the equation:

$$\frac{d}{dv}\hat{\theta} = -R_{ab} K^a K^b - 2\hat{\sigma}^2 - \tfrac{1}{2}\hat{\theta}^2.$$

Thus in this case θ will monotonically decrease along the null geodesic if $R_{ab}W^aW^b \geqslant 0$ for any null vector **W**. We shall call this the *null convergence condition*. From the Einstein equations,

$$R_{ab} - \tfrac{1}{2}g_{ab}R + \Lambda g_{ab} = 8\pi T_{ab},$$

it follows that this condition is implied by the weak energy condition, independent of the value of Λ.

From (4.26) it can be seen that the expansion θ of a timelike geodesic congruence with zero vorticity will monotonically decrease along a geodesic if $R_{ab}W^aW^b \geqslant 0$ for any timelike vector **W**. We shall call this the *timelike convergence condition*. By the Einstein equation, this condition will be satisfied if the energy–momentum tensor obeys the inequality,

$$T_{ab}W^aW^b \geqslant W^aW_a\left(\tfrac{1}{2}T - \frac{1}{8\pi}\Lambda\right).$$

This will hold for type I if

$$\mu + p_\alpha \geqslant 0, \quad \mu + \Sigma p_\alpha - \frac{1}{4\pi}\Lambda \geqslant 0,$$

and for type II if

$$\nu = +1, \quad \kappa \geqslant 0, \quad p_1 \geqslant 0, \quad p_2 \geqslant 0 \quad \text{and} \quad p_1 + p_2 - \frac{1}{4\pi}\Lambda \geqslant 0.$$

We shall say that the energy–momentum tensor satisfies the *strong energy condition* if it obeys the above inequality for $\Lambda = 0$. This is a stricter requirement than the weak energy condition but it is still physically reasonable for the total energy–momentum tensor. For the general case, type I, it would be violated only by a negative energy density or a large negative pressure (e.g. for a perfect fluid with density $1\,\mathrm{gm\,cm^{-3}}$ it can only be violated if $p < -10^{15}$ atmospheres). It holds for the electromagnetic field and for the scalar field with m zero (in particular, it holds for the scalar field of Brans and Dicke). For m non-zero, the energy–momentum tensor of a scalar field has the form (§3.3):

$$T_{ab} = \phi_{;a}\phi_{;b} - \tfrac{1}{2}g_{ab}(\phi_{;c}\phi_{;d}g^{cd} + m^2\phi^2).$$

Thus if W^a is a unit timelike vector

$$T_{ab}W^aW^b - \tfrac{1}{2}W_aW^aT = (\phi_{;a}W^a)^2 - \frac{1}{2}\frac{m^2}{\hbar^2}\phi^2 \tag{4.37}$$

which may be negative. However by the equation of the scalar field

$$\frac{1}{2}\frac{m^2}{\hbar^2}\phi^2 = \tfrac{1}{2}\phi\phi_{;ab}g^{ab}.$$

Inserting this in (4.37) and integrating over a region $\mathscr{U}$, one obtains

$$\frac{1}{2}\int_{\mathscr{U}} (g^{ab} + 2W^a W^b)\,\phi_{;a}\phi_{;b}\,\mathrm{d}\sigma - \frac{1}{2}\int_{\partial\mathscr{U}} \phi\phi_{;a}g^{ab}\,\mathrm{d}\sigma_b.$$

The first term will be non-negative since $g^{ab} + 2W^a W^b$ is a positive definite metric and the second term will be small compared to the first if the region $\mathscr{U}$ is large compared to the wavelength $\hbar/m$. For π mesons, which may be described classically by a scalar field with $m = 6 \times 10^{-25}$ gm, this wavelength is 3×10^{-13} cm. Thus although the energy–momentum tensor of π mesons may not satisfy the strong energy condition at every point, this should not affect the convergence of timelike geodesics over distances greater than 10^{-12} cm. This might possibly lead to a breakdown of the singularity theorems in chapter 8 when the radius of curvature of space–time becomes less than 10^{-12} cm but such a curvature would be so extreme that it might well count as a singularity (§10.2).

4.4 Conjugate points

In §4.1 we saw that the components of the vector which represented the separation between a curve $\gamma(s)$ and a neighbouring curve in a congruence of timelike geodesics, satisfied the Jacobi equation:

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}Z^\alpha = -R_{\alpha 4 \beta 4}Z^\beta \quad (\alpha, \beta = 1, 2, 3). \tag{4.38}$$

A solution of this equation will be called a *Jacobi field* along $\gamma(s)$. Since a solution may be specified by giving the values of Z^α and $\mathrm{d}Z^\alpha/\mathrm{d}s$ at some point on $\gamma(s)$ there will be six independent Jacobi fields along $\gamma(s)$. There will be three independent Jacobi fields which vanish at some point q of $\gamma(s)$. They may be expressed as:

$$Z^\alpha(s) = A_{\alpha\beta}(s)\frac{\mathrm{d}}{\mathrm{d}s}Z^\beta\big|_q,$$

where

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}A_{\alpha\beta}(s) = -R_{\alpha 4 \gamma 4}A_{\gamma\beta}(s), \tag{4.39}$$

and $A_{\alpha\beta}(s)$ is a 3×3 matrix which vanishes at q. These Jacobi fields may be thought of as representing the separation of neighbouring geodesics through q. As before one may define the vorticity, shear and

expansion of the Jacobi fields along $\gamma(s)$ which vanish at q:

$$\omega_{\alpha\beta} = A^{-1}{}_{\gamma[\beta}\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha]\gamma}, \tag{4.40}$$

$$\sigma_{\alpha\beta} = A^{-1}{}_{\gamma(\beta}\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha)\gamma} - \tfrac{1}{3}\delta_{\alpha\beta}\theta, \tag{4.41}$$

$$\theta = (\det \mathbf{A})^{-1}\frac{\mathrm{d}}{\mathrm{d}s}(\det \mathbf{A}). \tag{4.42}$$

These will obey the equations derived in §4.1, with $\dot{V}_\alpha = 0$. In particular

$$A_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta} = \frac{1}{2}\left(A_{\gamma\alpha}\frac{\mathrm{d}}{\mathrm{d}s}A_{\gamma\beta} - A_{\gamma\beta}\frac{\mathrm{d}}{\mathrm{d}s}A_{\gamma\alpha}\right)$$

will be constant along $\gamma(s)$. But it vanishes at q where $A_{\alpha\beta}$ is zero. Thus $\omega_{\alpha\beta}$ will be zero wherever $A_{\alpha\beta}$ is non-singular.

We shall say that a point p on $\gamma(s)$ *is conjugate to* q *along* $\gamma(s)$ if there is a Jacobi field along $\gamma(s)$, not identically zero, which vanishes at q and p. One may think of p as a point where infinitesimally neighbouring geodesics through q intersect. (Note, however, that it may be only *infinitesimally* neighbouring geodesics which intersect at p; there need not be two distinct geodesics from q passing through p.) The Jacobi fields along $\gamma(s)$ which vanish at q are described by the matrix $A_{\alpha\beta}$. Thus a point p is conjugate to q along $\gamma(s)$ if and only if $A_{\alpha\beta}$ is singular at p. The expansion θ is defined as $(\det \mathbf{A})^{-1}\,\mathrm{d}(\det \mathbf{A})/\mathrm{d}s$. Since $A_{\alpha\beta}$ obeys (4.39) where $R_{\alpha 4\gamma 4}$ is finite, $\mathrm{d}(\det \mathbf{A})/\mathrm{d}s$ will be finite. Thus a point p will be conjugate to q along $\gamma(s)$ if θ becomes infinite there. The converse will also be true since $\theta = \mathrm{d}\log(\det \mathbf{A})/\mathrm{d}s$ and $A_{\alpha\beta}$ can be singular only at isolated points or else it would be singular everywhere.

Proposition 4.4.1

If at some point $\gamma(s_1)$ $(s_1 > 0)$, the expansion θ has a negative value $\theta_1 < 0$ and if $R_{ab}V^aV^b \geqslant 0$ everywhere then there will be a point conjugate to q along $\gamma(s)$ between $\gamma(s_1)$ and $\gamma(s_1 + (3/-\theta_1))$, provided that $\gamma(s)$ can be extended to this parameter value. (This may not be possible if space–time is geodesically incomplete. In chapter 8 we shall interpret such incompleteness as evidence of the existence of a singularity.)

The expansion θ of the matrix $A_{\alpha\beta}$ obeys the Raychaudhuri equation (4.26):

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -R_{ab}V^aV^b - 2\sigma^2 - \tfrac{1}{3}\theta^2$$

where we have used the fact that the vorticity is zero. All the terms on the right-hand side are negative. Thus for $s > s_1$

$$\theta \leqslant \frac{3}{s-(s_1+(3/-\theta_1))}.$$

So θ will become infinite and there will be a point conjugate to q for some value of s between s_1 and $s_1+(3/-\theta_1)$. □

In other words, if the timelike convergence condition holds and if the neighbouring geodesics from q start converging on $\gamma(s)$, then some infinitesimally neighbouring geodesic will intersect $\gamma(s)$ providing that $\gamma(s)$ can be extended to large enough values of the parameter s.

Proposition 4.4.2

If $R_{ab}V^aV^b \geqslant 0$ and if at some point $p = \gamma(s_1)$ the tidal force $R_{abcd}V^bV^d$ is non zero, there will be values s_0 and s_2 such that $q = \gamma(s_0)$ and $r = \gamma(s_2)$ will be conjugate along $\gamma(s)$, providing that $\gamma(s)$ can be extended to these values.

A solution of (4.39) along $\gamma(s)$ is uniquely determined by the values of $A_{\alpha\beta}$ and $\mathrm{d}A_{\alpha\beta}/\mathrm{d}s$ at p. Consider the set P consisting of all such solutions for which $A_{\alpha\beta}|_p = \delta_{\alpha\beta}$, $(\mathrm{d}A_{\alpha\beta}/\mathrm{d}s)|_p$ is symmetric with trace $\theta|_p \leqslant 0$. For each solution in P there will be some $s_3 > s_1$ for which $A_{\alpha\beta}(s_3)$ is singular, since either $\theta|_p < 0$, in which case this follows from the previous result, or $\theta|_p = 0$, in which case $(\mathrm{d}\sigma_{\alpha\beta}/\mathrm{d}s)|_p$ is non-zero which will then cause σ^2 to be positive and so cause θ to become negative for $s > s_1$. The members of the set P are in one–one correspondence with the space S of all symmetric 3×3 matrices with non-positive trace (i.e. with the values of $\mathrm{d}A_{\alpha\beta}/\mathrm{d}s)|_p$). There is thus a map η from S to $\gamma(s)$ which assigns to each initial value $(\mathrm{d}A_{\alpha\beta}/\mathrm{d}s)|_p$ the point on $\gamma(s)$ where $A_{\alpha\beta}$ first becomes singular. The map η is continuous. Further if any component of $(\mathrm{d}A_{\alpha\beta}/\mathrm{d}s)|_p$ is very large, the corresponding point on $\gamma(s)$ will lie near p, since in the limit the term $R_{\alpha4\gamma4}$ in (4.39) becomes irrelevant and the solution resembles the flat space case. Thus there is some $C > 0$ and some $s_4 > s_1$ such that if any component of $(\mathrm{d}A_{\alpha\beta}/\mathrm{d}s)|_p$ is greater than C, the corresponding point on $\gamma(s)$ will be before $\gamma(s_4)$. However the subspace of S consisting of all matrices all of whose components are less than or equal to C, is compact. This shows that there is some $s_5 > s_1$ such that $\eta(S)$ is contained in the segment from $\gamma(s_1)$ to $\gamma(s_5)$. Consider now a point $r = \gamma(s_2)$ where $s_2 > s_5$. If there is no point conjugate to r between r and p, the Jacobi fields which are zero at r

must have an expansion θ which is positive at p (otherwise they would be in the set P which represents all families of Jacobi fields with zero vorticity which have non-positive expansion at p). It follows from the previous result that there is then a point $q = \gamma(s_0)$ $(s_0 < s_1)$ which is conjugate to r along $\gamma(s)$. □

In a physically realistic solution (though not necessarily in an exact one with a high degree of symmetry), one would expect every timelike geodesic to encounter some matter or some gravitational radiation and so to contain some point where $R_{abcd} V^b V^d$ was non-zero. Thus it would be reasonable to assume that in such a solution every timelike geodesic would contain pairs of conjugate points, provided that it could be extended sufficiently far in both directions.

We shall also consider the congruence of timelike geodesics normal to a spacelike three-surface, $\mathscr{H}$. By a *spacelike three-surface*, $\mathscr{H}$, we mean an imbedded three-dimensional submanifold defined locally by $f = 0$ where f is a C^2 function and $g^{ab} f_{;a} f_{;b} < 0$ when $f = 0$. We define $\mathbf{N}$, the unit normal vector to $\mathscr{H}$, by $N^a = (-g^{bc} f_{;b} f_{;c})^{-\frac{1}{2}} g^{ad} f_{;d}$ and the second fundamental tensor $\boldsymbol{\chi}$ of $\mathscr{H}$ by $\chi_{ab} = h_a{}^c h_b{}^d N_{c;d}$, where $h_{ab} = g_{ab} + N_a N_b$ is called the first fundamental tensor (or induced metric tensor) of $\mathscr{H}$ (cf. §2.7). It follows from the definition that $\boldsymbol{\chi}$ is symmetric. The congruence of timelike geodesics orthogonal to $\mathscr{H}$ will consist of the timelike geodesics whose unit tangent vector $\mathbf{V}$ equals the unit normal $\mathbf{N}$ at $\mathscr{H}$. Then one has:

$$V_{a;b} = \chi_{ab} \quad \text{at} \quad \mathscr{H}. \tag{4.43}$$

The vector $\mathbf{Z}$ which represents the separation of a neighbouring geodesic normal to $\mathscr{H}$ from a geodesic $\gamma(s)$ normal to $\mathscr{H}$, will obey the Jacobi equation (4.38). At a point q on $\gamma(s)$ at $\mathscr{H}$ it will satisfy the initial condition:

$$\frac{\mathrm{d}}{\mathrm{d}s} Z^\alpha = \chi_{\alpha\beta} Z^\beta. \tag{4.44}$$

We shall express the Jacobi fields along $\gamma(s)$ which satisfy the above condition as

$$Z^\alpha(s) = A_{\alpha\beta}(s)\, Z^\beta|_q,$$

where

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} A_{\alpha\beta} = -R_{\alpha 4 \gamma 4} A_{\gamma\beta} \tag{4.45}$$

and at q, $A_{\alpha\beta}$ is the unit matrix and

$$\frac{\mathrm{d}}{\mathrm{d}s} A_{\alpha\beta} = \chi_{\alpha\gamma} A_{\gamma\beta}. \tag{4.46}$$

We shall say that a point p on $\gamma(s)$ *is conjugate to* $\mathscr{H}$ *along* $\gamma(s)$ if there is a Jacobi field along $\gamma(s)$ not identically zero, which satisfies the initial conditions (4.44) at q and vanishes at p. In other words, p is conjugate to $\mathscr{H}$ along $\gamma(s)$ if and only if $A_{\alpha\beta}$ is singular at p. One may think of p as being a point where neighbouring geodesics normal to $\mathscr{H}$ intersect. As before $A_{\alpha\beta}$ will be singular where and only where the expansion θ becomes infinite. At q, the initial value of $A_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta}$ will be zero, therefore $\omega_{\alpha\beta}$ will be zero on $\gamma(s)$. The initial value of θ will be $\chi_{ab}g^{ab}$.

Proposition 4.4.3

If $R_{ab}V^aV^b \geqslant 0$ and $\chi_{ab}g^{ab} < 0$, there will be a point conjugate to $\mathscr{H}$ along $\gamma(s)$ within a distance $3/(-\chi_{ab}g^{ab})$ from $\mathscr{H}$, provided that $\gamma(s)$ can be extended that far.

This may be proved using the Raychaudhuri equation (4.26) as in proposition 4.4.1. □

We shall call a solution of the equation:

$$\frac{\mathrm{d}^2}{\mathrm{d}v^2}Z^m = -R_{m4n4}Z^n \quad (m,n = 1,2)$$

along a null geodesic $\gamma(v)$, *a Jacobi field along* $\gamma(v)$. The components Z^m could be thought of as the components, with respect to the basis $\mathbf{E}_1$ and $\mathbf{E}_2$, of a vector in the space S_q at each point q. We shall say that p is conjugate to q along the null geodesic $\gamma(v)$ if there is a Jacobi field along $\gamma(v)$, not identically zero, which vanishes at q and p. If $\mathbf{Z}$ is a vector connecting neighbouring null geodesics which pass through q, the component Z^3 will be zero everywhere. Thus p can be thought of as a point where infinitesimally neighbouring geodesics through q intersect. Representing the Jacobi fields along $\gamma(v)$ which vanish at q by the 2×2 matrix $\hat{A}_{mn}$,

$$Z^m(v) = \hat{A}_{mn}\frac{\mathrm{d}}{\mathrm{d}v}Z^n|_q.$$

One has as before: $\hat{A}_{lm}\hat{\omega}_{lk}\hat{A}_{kn} = 0$, so the vorticity of the Jacobi fields which are zero at p vanishes. Also p will be conjugate to q along $\gamma(v)$ if and only if

$$\hat{\theta} = (\det\hat{\mathbf{A}})^{-1}\frac{\mathrm{d}}{\mathrm{d}v}(\det\hat{\mathbf{A}})$$

becomes infinite at p. Analogous to proposition 4.4.1, we have:

Proposition 4.4.4

If $R_{ab}K^aK^b \geqslant 0$ everywhere and if at some point $\gamma(v_1)$ the expansion $\hat{\theta}$ has the negative value $\hat{\theta}_1 < 0$, then there will be a point conjugate to q along $\gamma(v)$ between $\gamma(v_1)$ and $\gamma(v_1 + (2/-\hat{\theta}_1))$ provided that $\gamma(v)$ can be extended that far.

The expansion $\hat{\theta}$ of the matrix $\hat{A}_{mn}$ obeys (4.35):

$$\frac{\mathrm{d}}{\mathrm{d}v}\hat{\theta} = -R_{ab}K^aK^b - 2\hat{\sigma}^2 - \tfrac{1}{2}\hat{\theta}^2,$$

and so the proof proceeds as before. □

Proposition 4.4.5

If $R_{ab}K^aK^b \geqslant 0$ everywhere and if at $p = \gamma(v_1)$, $K^cK^dK_{[a}R_{b]cd[e}K_{f]}$ is non-zero, there will be v_0 and v_2 such that $q = \gamma(v_0)$ and $r = \gamma(v_2)$ will be conjugate along $\gamma(v)$ provided $\gamma(v)$ can be extended to these values.

If $K^cK^dK_{[a}R_{b]cd[e}K_{f]}$ is non zero then so is R_{m4n4}. The proof is then similar to that of proposition 4.4.2. □

As in the timelike case, this condition will be satisfied for a null geodesic which passes through some matter provided that the matter is not pure radiation (energy–momentum tensor type II of §4.3) and moving in the direction of the geodesic tangent vector **K**. It will be satisfied in empty space if the null geodesic contains some point where the Weyl tensor is non-zero and where **K** does not lie in one of the directions (there are at most four such directions) at that point for which $K^cK^dK_{[a}C_{b]cd[e}K_{f]} = 0$. It therefore seems reasonable to assume that in a physically realistic solution every timelike or null geodesic will contain a point at which $K^aK^bK_{[c}R_{d]ab[e}K_{f]}$ is not zero. We shall say that a space–time satisfying this condition satisfies the *generic condition*.

Similarly we may also consider the null geodesics orthogonal to a spacelike two-surface $\mathscr{S}$. By a *spacelike two-surface* $\mathscr{S}$, we mean an imbedded two-dimensional submanifold defined locally by $f_1 = 0$, $f_2 = 0$ where f_1 and f_2 are C^2 functions such that when $f_1 = 0$, $f_2 = 0$ then $f_{1;a}$ and $f_{2;a}$ are non-vanishing and not parallel and

$$(f_{1;a} + \mu f_{2;a})(f_{1;b} + \mu f_{2;b})g^{ab} = 0$$

for two distinct real values μ_1 and μ_2 of μ. Then any vector lying in the two-surface is necessarily spacelike. We shall define $N_1{}^a$ and $N_2{}^a$, the

two null vectors normal to $\mathscr{S}$, as proportional to $g^{ab}(f_{1;b}+\mu_1 f_{2;b})$ and $g^{ab}(f_{1;b}+\mu_2 f_{2;b})$ respectively, and normalize them so that

$$N_1{}^a N_2{}^b g_{ab} = -1.$$

One can complete the pseudo-orthonormal basis by introducing two spacelike unit vectors $Y_1{}^a$ and $Y_2{}^a$ orthogonal to each other and to $N_1{}^a$ and $N_2{}^a$. We define the two null second fundamental tensors of $\mathscr{S}$ as:

$$_n\chi_{ab} = -N_{nc;d}(Y_1{}^c Y_{1a}+Y_2{}^c Y_{2a})(Y_1{}^d Y_{1b}+Y_2{}^d Y_{2b}),$$

where n takes the values 1, 2. The tensors $_1\chi_{ab}$ and $_2\chi_{ab}$ are symmetric.

There will be two families of null geodesics normal to $\mathscr{S}$ corresponding to the two null normals $N_1{}^a$ and $N_2{}^a$. Consider the family whose tangent vector $\mathbf{K}$ equals $\mathbf{N}_2$ at $\mathscr{S}$. We may fix our pseudo-orthogonal basis $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, $\mathbf{E}_4$ by taking $\mathbf{E}_1 = \mathbf{Y}_1$, $\mathbf{E}_2 = \mathbf{Y}_2$, $\mathbf{E}_3 = \mathbf{N}_1$, $\mathbf{E}_4 = \mathbf{N}_2$ at $\mathscr{S}$ and parallelly propagating along the null geodesics. The projection into the space S_q of the vector $\mathbf{Z}$ representing the separation of neighbouring null geodesics from the null geodesic $\gamma(v)$ will satisfy (4.30) and the initial conditions

$$\frac{\mathrm{d}}{\mathrm{d}v} Z^m = {}_2\chi_{mn} Z^n \tag{4.47}$$

at q on $\gamma(v)$ at $\mathscr{S}$. As before the vorticity of these fields will be zero. The initial value of the expansion θ will be $_2\chi_{ab}g^{ab}$. Analogous to proposition 4.4.3 we have:

Proposition 4.4.6

If $R_{ab}K^aK^b \geqslant 0$ everywhere and $_2\chi_{ab}g^{ab}$ is negative there will be a point conjugate to $\mathscr{S}$ along $\gamma(v)$ within an affine distance $2/(-{}_2\chi_{ab}g^{ab})$ from $\mathscr{S}$. □

From their definition, the existence of conjugate points implies the existence of self-intersections or caustics in families of geodesics. A further significance of conjugate points will be discussed in the next section.

4.5 Variation of arc-length

In this section we consider timelike and non-spacelike curves which are piecewise C^3 but which may have points at which their tangent

vector is discontinuous. We shall require that at such points the two tangent vectors

$$\left.\frac{\partial}{\partial t}\right|_- \quad \text{and} \quad \left.\frac{\partial}{\partial t}\right|_+ \quad \text{satisfy} \quad g\left(\left.\frac{\partial}{\partial t}\right|_-, \left.\frac{\partial}{\partial t}\right|_+\right) = -1,$$

that is, they point into the same half of the null cone.

Proposition 4.5.1

Let $\mathscr{U}$ be a convex normal coordinate neighbourhood about q. Then the points which can be reached from q by timelike (respectively non-spacelike) curves in $\mathscr{U}$ are those of the form $\exp_q(\mathbf{X})$, $\mathbf{X} \in T_q$ where $g(\mathbf{X}, \mathbf{X}) < 0$ (respectively $\leqslant 0$). (Here, and for the rest of this section, we consider the map exp to be restricted to the neighbourhood of the origin in T_q which is diffeomorphic to $\mathscr{U}$ under $\exp_q$.)

In other words, the null geodesics from q form the boundary of the region in $\mathscr{U}$ which can be reached from q by timelike or non-spacelike curves in $\mathscr{U}$. This is fairly obvious intuitively but because it is fundamental to the concept of causality we shall prove it rigorously. We first establish the following lemma:

Lemma 4.5.2

In $\mathscr{U}$ the timelike geodesics through q are orthogonal to the three-surfaces of constant σ ($\sigma < 0$) where the value of σ at $p \in \mathscr{U}$ is defined to be $g(\exp_q^{-1} p, \exp_q^{-1} p)$.

The proof is based on the fact that the vector representing the separation of points equal distances along neighbouring geodesics remains orthogonal to the geodesics if it is so initially. More precisely, let $\mathbf{X}(t)$ denote a curve in T_q, where $g(\mathbf{X}(t), \mathbf{X}(t)) = -1$. One must show that the corresponding curves $\lambda(t) = \exp_q(s_0 \mathbf{X}(t))$ (s_0 constant) in $\mathscr{U}$, where defined, are orthogonal to the timelike geodesics $\gamma(s) = \exp_q(s\mathbf{X}(t_0))$ (t_0 constant). Thus in terms of the two-surface α defined by $x(s, t) = \exp_q(s\mathbf{X}(t))$, one must prove that

$$g\left(\left(\frac{\partial}{\partial s}\right)_\alpha, \left(\frac{\partial}{\partial t}\right)_\alpha\right) = 0$$

(see figure 11). Now

$$\frac{\partial}{\partial s} g\left(\frac{\partial}{\partial s}, \frac{\partial}{\partial t}\right) = g\left(\frac{\mathrm{D}}{\partial s}\frac{\partial}{\partial s}, \frac{\partial}{\partial t}\right) + g\left(\frac{\partial}{\partial s}, \frac{\mathrm{D}}{\partial s}\frac{\partial}{\partial t}\right).$$

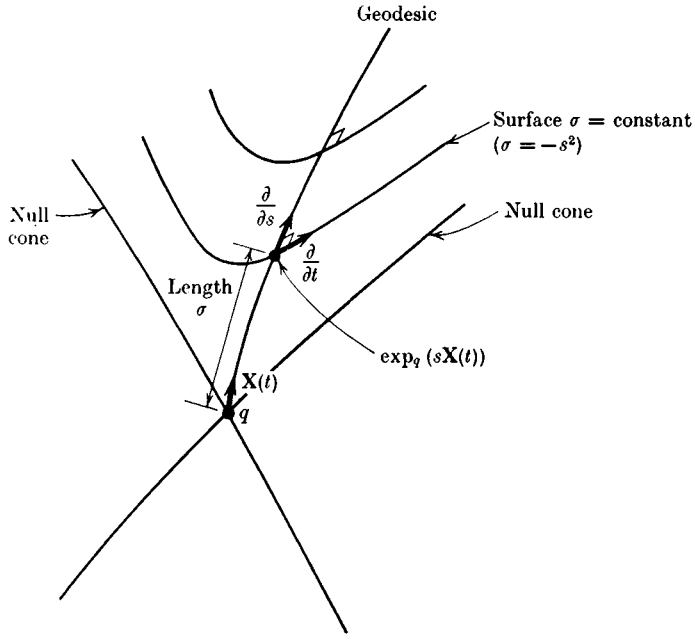

FIGURE 11. In a normal neighbourhood, surfaces at constant distance from q are orthogonal to the geodesics through q.

The first term on the right is zero as $\partial/\partial s$ is the unit tangent vector to the timelike geodesics from q. In the second term one has from the definition of the Lie derivative that

$$\frac{\mathrm{D}}{\partial s}\frac{\partial}{\partial t} = \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial s}.$$

Thus
$$\frac{\partial}{\partial s} g\left(\frac{\partial}{\partial s}, \frac{\partial}{\partial t}\right) = g\left(\frac{\partial}{\partial s}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial s}\right) = \frac{1}{2}\frac{\partial}{\partial t} g\left(\frac{\partial}{\partial s}, \frac{\partial}{\partial s}\right) = 0.$$

Therefore $g(\partial/\partial s, \partial/\partial t)$ is independent of s. But at $s = 0$, $(\partial/\partial t)_\alpha = 0$. Thus $g(\partial/\partial s, \partial/\partial t)$ is identically zero. □

Proof of proposition 4.5.1. Let C_q denote the set of all timelike vectors at q. These constitute the interior of a solid cone in T_q with vertex at the origin. Let $\gamma(t)$ be a timelike curve in $\mathscr{U}$ from q to p and let $\bar{\gamma}(t)$ be the piecewise C^2 curve in T_q defined by $\bar{\gamma}(t) = \exp_q^{-1}(\gamma(t))$. Then identifying the tangent space to T_q with T_q itself, one has

$$(\partial/\partial t)_\gamma|_q = (\partial/\partial t)_{\bar{\gamma}}|_q.$$

Therefore at q, $(\partial/\partial t)_{\bar{\gamma}}$ will be timelike. This shows that the curve $\bar{\gamma}(t)$ will enter the region C_q. But $\exp_q(C_q)$ is the region of $\mathscr{U}$ on which σ is negative and in which by the previous lemma the surfaces of constant σ are spacelike. Thus σ must monotonically decrease along $\gamma(t)$ since $(\partial/\partial t)_\gamma$ being timelike can never be tangent to the surfaces of constant σ and since at any non-differentiable point of $\gamma(t)$ the two tangent vectors point into the same half of the null cone. Therefore $p \in \exp_q(C_q)$ which completes the proof for timelike curves. To prove that a non-spacelike curve $\gamma(t)$ remains in $\exp_q(\bar{C}_q)$, one performs a small variation of $\gamma(t)$ which makes it into a timelike curve. Let $\mathbf{Y}$ be a vector field on T_q such that in $\mathscr{U}$ the induced vector field $\exp_{q*}(\mathbf{Y})$ is everywhere timelike and such that $g(\mathbf{Y}, (\partial/\partial t)_\gamma|_q) < 0$. For each $\epsilon \geqslant 0$ let $\beta(r, \epsilon)$ be the curve T_q starting at the origin such that the tangent vector $(\partial/\partial r)_\beta$ equals $(\partial/\partial t)_{\bar{\gamma}}|_{t=r} + \epsilon \mathbf{Y}|_{\beta(r,\epsilon)}$. Then $\beta(r, \epsilon)$ depends differentiably on r and ϵ. For each $\epsilon > 0$, $\exp_q(\beta(r, \epsilon))$ is a timelike curve in $\mathscr{U}$ and so is contained in $\exp_q(C_q)$. Thus the non-spacelike curve $\exp_q(\beta(r, 0)) = \gamma(r)$ is contained in $\overline{\exp_q(C_q)} = \exp_q(\bar{C}_q)$. □

Corollary

If $p \in \mathscr{U}$ can be reached from q by a non-spacelike curve but not by a timelike curve, then p lies on a null geodesic from q. □

The length of a non-spacelike curve $\gamma(t)$ from q to p is

$$L(\gamma, q, p) = \int_q^p \left[-g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right]^{\frac{1}{2}} \mathrm{d}t,$$

where the integral is taken over the differentiable sections of the curve.

In a positive definite metric one may seek the shortest curve between two points but in a Lorentz metric there will not be any shortest curve as any curve can be deformed into a null curve which has zero length. However, in certain cases there will be a longest non-spacelike curve between two points or between a point and a spacelike three-surface. We deal first with the situation when the two points are close together. We shall then derive necessary conditions in the general case when the two points are not close. The sufficient condition in this case will be dealt with in § 6.7.

Proposition 4.5.3

Let q and p lie in a convex normal neighbourhood $\mathscr{U}$. Then, if q and p can be joined by a non-spacelike curve in $\mathscr{U}$, the longest such curve is the unique non-spacelike geodesic curve in $\mathscr{U}$ from q to p. Moreover,

defining $\rho(q,p)$ as the length of this curve if it exists, and as zero otherwise, $\rho(q,p)$ is a continuous function on $\mathscr{U}\times\mathscr{U}$.

By the definition of convex normal neighbourhoods (§2.5), there is a unique geodesic $\gamma(t)$ in $\mathscr{U}$ with $\gamma(0)=q, \gamma(1)=p$. Since this geodesic depends differentiably on its endpoints, the function

$$\sigma(q,p)=\int_0^1 g\left(\left(\frac{\partial}{\partial t}\right)_\gamma,\left(\frac{\partial}{\partial t}\right)_\gamma\right)\mathrm{d}t$$

will be differentiable on $\mathscr{U}\times\mathscr{U}$. (This function σ is the same as that in lemma 4.5.2.) Thus $\rho(q,p)$ will be continuous on $\mathscr{U}\times\mathscr{U}$ since it equals $[-\sigma(q,p)]^{\frac{1}{2}}$ if $\sigma<0$ and is zero otherwise. It now remains to show that if q and p can be joined by a timelike curve in $\mathscr{U}$ then the timelike geodesic γ between them is the longest such curve. Let $\alpha(s,t)$ be $\exp_q(s\mathbf{X}(t))$ as before where $g(\mathbf{X}(t),\mathbf{X}(t))=-1$. If $\lambda(t)$ is a timelike curve in $\mathscr{U}$ from q to p, it can be represented as $\lambda(t)=\alpha(f(t),t)$. Then

$$\left(\frac{\partial}{\partial t}\right)_\lambda=f'(t)\left(\frac{\partial}{\partial s}\right)_\alpha+\left(\frac{\partial}{\partial t}\right)_\alpha.$$

Since the two vectors on the right are mutually orthogonal by lemma 4.5.2. and since $g((\partial/\partial s)_\alpha,(\partial/\partial s)_\alpha)=-1$, this gives

$$g\left(\left(\frac{\partial}{\partial t}\right)_\lambda,\left(\frac{\partial}{\partial t}\right)_\lambda\right)=-(f'(t))^2+g\left(\left(\frac{\partial}{\partial t}\right)_\alpha,\left(\frac{\partial}{\partial t}\right)_\alpha\right)\geqslant-(f'(t))^2,$$

the equality holding if and only if $(\partial/\partial t)_\alpha=0$ and hence if and only if λ is a geodesic curve. Thus

$$L(\lambda,q,p)\leqslant\int_q^p f'(t)\,\mathrm{d}t=\rho(q,p),$$

the equality holding if and only if λ is the unique geodesic curve in $\mathscr{U}$ from q to p. □

We shall now consider the case where q and p are not necessarily contained in a convex normal neighbourhood $\mathscr{U}$. By considering small variations we shall derive necessary conditions for a timelike curve $\gamma(t)$ from q to p to be the longest such curve from q to p. A *variation* α *of* $\gamma(t)$ is a C^{1-} map $\alpha\colon(-\epsilon,\epsilon)\times[0,t_p]\to\mathscr{M}$ such that

(1) $\alpha(0,t)=\gamma(t)$;

(2) there is a subdivision $0=t_1<t_2\ldots<t_n=t_p$ of $[0,t_p]$ such that α is C^3 on each $(-\epsilon,\epsilon)\times[t_i,t_{i+1}]$;

(3) $\alpha(u,0)=q$, $\alpha(u,t_p)=p$;

(4) for each constant u, $\alpha(u,t)$ is a timelike curve.

The vector $(\partial/\partial u)_\alpha|_{u=0}$ will be called the *variation vector* $\mathbf{Z}$. Conversely, given a continuous, piecewise C^2 vector field $\mathbf{Z}$ along $\gamma(t)$ vanishing at q and p, we may define a variation α for which $\mathbf{Z}$ will be the variation vector by:

$$\alpha(u,t) = \exp_r(u\mathbf{Z}|_r),$$

where $u \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$ and $r = \gamma(t)$.

Lemma 4.5.4

The variation of the length from q to p under α is

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = \sum_{i=1}^{n-1}\int_{t_i}^{t_{i+1}} g\left(\frac{\partial}{\partial u}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\left(\frac{\partial f}{\partial t}\right)\frac{\partial}{\partial t}\right\}\right)\mathrm{d}t + \sum_{i=2}^{n-1} g\left(\frac{\partial}{\partial u}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right),$$

where $f^2 = g(\partial/\partial t, \partial/\partial t)$ is the magnitude of the tangent vector and $[f^{-1}\partial/\partial t]$ is the discontinuity at one of the singular points of $\gamma(t)$.

We have:

$$\begin{aligned}\left.\frac{\partial L}{\partial u}\right|_{u=0} &= \Sigma\frac{\partial}{\partial u}\int\left(-g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right)^{\frac{1}{2}}\mathrm{d}t \\ &= -\Sigma\int g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)f^{-1}\,\mathrm{d}t \\ &= -\Sigma\int g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)f^{-1}\,\mathrm{d}t \\ &= -\Sigma\int\left\{\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\right)f^{-1} - g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)f^{-1}\right\}\mathrm{d}t.\end{aligned}$$

Integrating the first term by parts one has the required formula. □

One may simplify the formula by choosing the parameter t to be the arc-length s. Then $g(\partial/\partial t, \partial/\partial t) = -1$. We shall denote by $\mathbf{V}$ the unit tangent vector $\partial/\partial s$. One has:

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = \sum_{i=1}^{n-1}\int_{t_i}^{t_{i+1}} g(\mathbf{Z}, \dot{\mathbf{V}})\,\mathrm{d}s + \sum_{i=2}^{n-1} g(\mathbf{Z}, [\mathbf{V}])$$

where $\dot{\mathbf{V}} = \mathrm{D}\mathbf{V}/\partial s$ is the acceleration. From this one sees again that a necessary condition for $\gamma(t)$ to be the longest curve from q to p is that it *should be an unbroken geodesic curve* as otherwise one could choose a variation which would yield a longer curve.

One may also consider a timelike curve $\gamma(t)$ from a spacelike three-surface $\mathscr{H}$ to a point p. A variation α of this curve is defined as before except that condition (3) is replaced by:

(3) $\alpha(u, 0)$ lies on $\mathscr{H}$, $\alpha(u, t_p) = p$.

Thus at $\mathscr{H}$ the variation vector $\mathbf{Z} = \partial/\partial u$ lies in $\mathscr{H}$.

Lemma 4.5.5

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} g(\dot{\mathbf{V}}, \mathbf{Z})\, ds + \sum_{i=2}^{n-1} g(\mathbf{Z}, [\mathbf{V}]) + g(\mathbf{Z}, \mathbf{V})|_{s=0}.$$

The proof is as for lemma 4.5.4. □

From this one sees that a necessary condition for $\gamma(t)$ to be the longest curve from $\mathscr{H}$ to p is that it is an *unbroken geodesic curve orthogonal to* $\mathscr{H}$.

We have seen that, under a variation α, the first derivative of the length of a timelike geodesic curve is zero. To proceed further we shall calculate the second derivative. We define a two-parameter variation α of a geodesic curve $\gamma(t)$ from q to p as a C^1 map:

$$\alpha\colon (-\epsilon_1, \epsilon_1) \times (-\epsilon_2, \epsilon_2) \times [0, t_p] \to \mathscr{M}$$

such that

(1) $\alpha(0, 0, t) = \gamma(t)$;

(2) there is a subdivision $0 = t_1 < t_2 < \ldots < t_n = t_p$ of $[0, t_p]$ such that α is C^3 on each

$$(-\epsilon_1, \epsilon_1) \times (-\epsilon_2, \epsilon_2) \times [t_i, t_{i+1}];$$

(3) $\alpha(u_1, u_2, 0) = q, \quad \alpha(u_1, u_2, t_p) = p$;

(4) for all constant u_1, u_2, $\alpha(u_1, u_2, t)$ is a timelike curve.

We define

$$\mathbf{Z}_1 = \left(\frac{\partial}{\partial u_1}\right)_\alpha \Big|_{\substack{u_1=0 \\ u_2=0}},$$

$$\mathbf{Z}_2 = \left(\frac{\partial}{\partial u_2}\right)_\alpha \Big|_{\substack{u_1=0 \\ u_2=0}},$$

as the two variation vectors. Conversely given two continuous, piecewise C^2 vector fields $\mathbf{Z}_1$ and $\mathbf{Z}_2$ along $\gamma(t)$ one may define a variation for which they will be the variation vectors, by:

$$\alpha(u_1, u_2, t) = \exp_r (u_1 \mathbf{Z}_1 + u_2 \mathbf{Z}_2),$$
$$r = \gamma(t).$$

Lemma 4.5.6

Under the two-parameter variation of the geodesic curve $\gamma(t)$, the second derivative of the length will be:

$$\left.\frac{\partial^2 L}{\partial u_2\, \partial u_1}\right|_{\substack{u_1=0 \\ u_2=0}} = \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} g\left(\mathbf{Z}_1, \left\{\frac{\mathbf{D}^2}{\partial s^2}(\mathbf{Z}_2 + g(\mathbf{V}, \mathbf{Z}_2)\,\mathbf{V}) - \mathbf{R}(\mathbf{V}, \mathbf{Z}_2)\,\mathbf{V}\right\}\right) ds$$
$$+ \sum_{i=2}^{n-1} g\left(\mathbf{Z}_1, \left[\frac{\mathbf{D}}{\partial s}(\mathbf{Z}_2 + g(\mathbf{V}, \mathbf{Z}_2)\,\mathbf{V})\right]\right).$$

By lemma 4.5.4, one has:

$$\frac{\partial L}{\partial u_1}\bigg|_{\substack{u_1=0\\u_2=0}} = \Sigma\int g\left(\frac{\partial}{\partial u_1}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\left(\frac{\partial f}{\partial t}\right)\frac{\partial}{\partial t}\right\}\right)\mathrm{d}t + \Sigma g\left(\frac{\partial}{\partial u_1}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right).$$

Therefore

$$\begin{aligned}\frac{\partial^2 L}{\partial u_2\,\partial u_1}\bigg|_{\substack{u_1=0\\u_2=0}} = \Sigma\int g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\left(\frac{\partial f}{\partial t}\right)\frac{\partial}{\partial t}\right\}\right)\mathrm{d}t\\ -\Sigma\int g\left(\frac{\partial}{\partial u_1}, \left\{f^{-2}\left(\frac{\partial f}{\partial u_2}\right)\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-1}\frac{\mathrm{D}}{\partial u_2}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right.\right.\\ \left.\left. -2f^{-3}\left(\frac{\partial f}{\partial u_2}\right)\left(\frac{\partial f}{\partial t}\right)\frac{\partial}{\partial t} + f^{-2}\left(\frac{\partial^2 f}{\partial u_2\,\partial t}\right)\frac{\partial}{\partial t} + f^{-2}\left(\frac{\partial f}{\partial t}\right)\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}\right\}\right)\mathrm{d}t\\ +\Sigma g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right) + \Sigma g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial u_2}\left[f^{-1}\frac{\partial}{\partial t}\right]\right).\end{aligned}$$

The first and third terms vanish as $\gamma(t)$ is an unbroken geodesic curve. In the second term one can write:

$$\begin{aligned}\frac{\mathrm{D}}{\partial u_2}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} &= -\mathbf{R}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t} + \frac{\mathrm{D}}{\partial t}\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}\\ &= -\mathbf{R}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t} + \frac{\mathrm{D}^2}{\partial t^2}\frac{\partial}{\partial u_2}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 f}{\partial u_2\,\partial t} &= -\frac{\partial}{\partial t}\left(f^{-1}g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right)\\ &= -\frac{\partial}{\partial t}\left\{f^{-1}\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u_2}, \frac{\partial}{\partial t}\right)\right) - f^{-1}g\left(\frac{\partial}{\partial u_2}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)\right\}.\end{aligned}$$

In the fourth term:

$$\frac{\mathrm{D}}{\partial u_2}\left[f^{-1}\frac{\partial}{\partial t}\right] = \left[f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2} + f^{-3}g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}, \frac{\partial}{\partial t}\right)\frac{\partial}{\partial t}\right].$$

Then taking t to be the arc-length s, one obtains the required result. □

Although it is not immediately obvious from the appearance of the expression, one knows from its definition that it is symmetric in the two variation vector fields $\mathbf{Z}_1$ and $\mathbf{Z}_2$. One sees that it only depends on the projections of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ into the space orthogonal to $\mathbf{V}$. Thus we can confine our attention to variations α whose variation vectors are orthogonal to $\mathbf{V}$. We shall define T_γ to be the (infinite-dimensional) vector space consisting of all continuous, piecewise C^2 vector fields along $\gamma(t)$ orthogonal to $\mathbf{V}$ and vanishing at q and p. Then $\partial^2 L/\partial u_2\,\partial u_1$

will be a symmetric map of $T_\gamma \times T_\gamma$ to R^1. One may think of it as a symmetric tensor on T_γ and write it as:

$$L(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{\partial^2 L}{\partial u_2 \partial u_1}\bigg|_{\substack{u_1=0\\u_2=0}}, \quad \mathbf{Z}_1, \mathbf{Z}_2 \in T_\gamma.$$

One may also calculate the second derivative of the length from $\mathscr{H}$ to p of a geodesic curve $\gamma(t)$ normal to $\mathscr{H}$. One proceeds as before except that one endpoint of $\gamma(t)$ is allowed to vary over $\mathscr{H}$ instead of being fixed.

Lemma 4.5.7

The second derivative of the length of $\gamma(t)$ from $\mathscr{H}$ to p is:

$$\frac{\partial^2 L}{\partial u_2 \partial u_1}\bigg|_{\substack{u_1=0\\u_2=0}} = \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} g\left(\mathbf{Z}_1, \left\{\frac{\mathrm{D}^2}{\partial s^2}\mathbf{Z}_2 - \mathbf{R}(\mathbf{V}, \mathbf{Z}_2)\,\mathbf{V}\right\}\right) \mathrm{d}s$$
$$+ \sum_{i=2}^{n-1} g\left(\mathbf{Z}_1, \left[\frac{\mathrm{D}}{\partial s}\mathbf{Z}_2\right]\right) + g\left(\mathbf{Z}_1, \frac{\mathrm{D}}{\partial s}\mathbf{Z}_2\right)\bigg|_{\mathscr{H}} - \chi(\mathbf{Z}_1, \mathbf{Z}_2)\bigg|_{\mathscr{H}},$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ have been taken orthogonal to $\mathbf{V}$ and $\chi(\mathbf{Z}_1, \mathbf{Z}_2)$ is the second fundamental tensor of $\mathscr{H}$.

The first two terms are as for lemma 4.5.6. The extra terms are:

$$\frac{\mathrm{D}}{\partial u_2} g\left(\frac{\partial}{\partial u_1}, f^{-1}\frac{\partial}{\partial t}\right)\bigg|_{\mathscr{H}} = f^{-1} g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\bigg|_{\mathscr{H}}$$
$$+ f^{-3} g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\bigg|_{\mathscr{H}} + f^{-1} g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}\right)\bigg|_{\mathscr{H}}.$$

The second term vanishes as $\partial/\partial u_1$ is orthogonal to $\partial/\partial t$. If one takes t to be the arc-length s, then $\partial/\partial t$ will be equal to the unit normal $\mathbf{N}$ at $\mathscr{H}$. Since the endpoint of $\gamma(t)$ is restricted to varying over $\mathscr{H}$, $\partial/\partial u_1$ will always be orthogonal to $\mathbf{N}$. Thus

$$g\left(D\frac{\partial}{\partial u_2}\frac{\partial}{\partial u_1}, \mathbf{N}\right) = \frac{\partial}{\partial u_2} g\left(\frac{\partial}{\partial u_1}, \mathbf{N}\right) - g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial u_2}\mathbf{N}\right) = -\chi\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial u_2}\right). \quad \square$$

We shall say that a timelike geodesic curve $\gamma(t)$ from q to p is *maximal* if $L(\mathbf{Z}_1, \mathbf{Z}_2)$ is negative semi-definite. In other words, if $\gamma(t)$ is not maximal there is a small variation α which yields a longer curve from p to q. Similarly we shall say that a timelike geodesic curve from $\mathscr{H}$ to p normal to $\mathscr{H}$ is *maximal* if $L(\mathbf{Z}_1, \mathbf{Z}_2)$ is negative semi-definite, so if $\gamma(t)$ is not maximal there is a small variation which yields a longer curve from $\mathscr{H}$ to p.

Proposition 4.5.8

A timelike geodesic curve $\gamma(t)$ from q to p is maximal if and only if there is no point conjugate to q along $\gamma(t)$ in (q, p).

Suppose there is no conjugate point in (q,p). Then introduce a Fermi-propagated orthonormal basis along $\gamma(t)$. The Jacobi fields along $\gamma(t)$ which vanish at q will be represented by a matrix $A_{\alpha\beta}(t)$ which will be non-singular in (q,p), but which will be singular at q and possibly at p. Since conjugate points are isolated, $\mathrm{d}(\log\det \mathbf{A})/\mathrm{d}s$ will be infinite where $A_{\alpha\beta}$ is singular. Thus a C^0, piecewise C^2 vector field $\mathbf{Z}\in T_\gamma$ can be expressed in $[q,p]$ as

$$Z^\alpha = A_{\alpha\beta}\, W^\beta,$$

where W^β is C^0, piecewise C^1 on $[q,p]$. Then,

$$\begin{aligned} L(\mathbf{Z},\mathbf{Z}) &= \Sigma\int_0^{s_p} A_{\alpha\beta}\, W^\beta\left\{\frac{\mathrm{d}^2}{\mathrm{d}s^2}(A_{\alpha\delta}\, W^\delta) + R_{\alpha 4\gamma 4} A_{\gamma\delta}\, W^\delta\right\}\mathrm{d}s \\ &\quad + \Sigma A_{\alpha\beta}\, W^\beta\left[\frac{\mathrm{d}}{\mathrm{d}s}(A_{\alpha\delta}\, W^\delta)\right] \\ &= \lim_{\epsilon\to 0+}\Sigma\int_\epsilon^{s_p} A_{\alpha\beta}\, W^\beta\left\{2\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha\delta}\frac{\mathrm{d}}{\mathrm{d}s}W^\delta + A_{\alpha\delta}\frac{\mathrm{d}^2}{\mathrm{d}s^2}W^\delta\right\}\mathrm{d}s \\ &\quad + \Sigma A_{\alpha\beta}\, W^\beta A_{\alpha\delta}\left[\frac{\mathrm{d}}{\mathrm{d}s}W^\delta\right] \\ &= -\Sigma\int_0^{s_p}\left\{A_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}s}W^\beta A_{\alpha\delta}\frac{\mathrm{d}}{\mathrm{d}s}W^\delta + W^\beta\left(\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha\beta}A_{\alpha\delta}\right.\right. \\ &\qquad \left.\left. - A_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha\delta}\right)\frac{\mathrm{d}}{\mathrm{d}s}W^\delta\right\}\mathrm{d}s. \end{aligned}$$

(We take the limit because the second derivative of W^δ may not be defined at q.) But

$$\left(\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha\beta}\,A_{\alpha\delta} - A_{\alpha\beta}\frac{\mathrm{d}}{\mathrm{d}s}A_{\alpha\delta}\right) = -2A_{\alpha\beta}\omega_{\alpha\gamma}A_{\gamma\delta} = 0.$$

Therefore $L(\mathbf{Z},\mathbf{Z}) \leqslant 0$.

Conversely, suppose there is a point $r\in(q,p)$ conjugate to q along $\gamma(t)$. Let $\mathbf{W}$ be the Jacobi field along γ which vanishes at q and r. Let $\mathbf{K}\in T_\gamma$ be such that

$$K^a g_{ab}\frac{\mathrm{D}}{\partial s}W^b = -1 \quad \text{at} \quad r.$$

Extend $\mathbf{W}$ to p by putting it zero in $[r, p]$. Let $\mathbf{Z}$ be $\epsilon\mathbf{K} + \epsilon^{-1}\mathbf{W}$, where ϵ is some constant. Then

$$L(\mathbf{Z}, \mathbf{Z}) = \epsilon^2 L(\mathbf{K}, \mathbf{K}) + 2L(\mathbf{K}, \mathbf{W}) + 2\epsilon^{-2} L(\mathbf{W}, \mathbf{W}) = \epsilon^2 L(\mathbf{K}, \mathbf{K}) + 2.$$

Thus by taking ϵ small enough, $L(\mathbf{Z}, \mathbf{Z})$ may be made positive. □

One may obtain similar results for the case of a timelike geodesic curve $\gamma(t)$ orthogonal to $\mathscr{H}$, from $\mathscr{H}$ to p.

Proposition 4.5.9

A timelike geodesic curve $\gamma(t)$ from $\mathscr{H}$ to p is maximal if and only if there is no point in $(\mathscr{H}, q)$ conjugate to $\mathscr{H}$ along γ. □

We shall also consider variations of a non-spacelike curve $\gamma(t)$ from q to p. We shall be interested in the circumstances under which it is possible to find a variation α of $\gamma(t)$ which makes $g(\partial/\partial t, \partial/\partial t)$ negative everywhere, or in other words, yields a timelike curve from q to p. Under a variation α:

$$\begin{aligned}\frac{\partial}{\partial u}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right) &= 2g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = 2g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right) \\ &= 2\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\right) - 2g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right). \qquad (4.48)\end{aligned}$$

In order to obtain a timelike curve from q to p, one requires this to be less than or equal to zero everywhere on $\gamma(t)$.

Proposition 4.5.10

If p and q are joined by a non-spacelike curve $\gamma(t)$ which is not a null geodesic they can also be joined by a timelike curve.

If $\gamma(t)$ is not a null geodesic curve from p to q, there must be some point at which the tangent vector is discontinuous, or there must be some open interval on which the acceleration vector $(\mathrm{D}/\partial t)(\partial/\partial t)$ is non-zero and not parallel to $\partial/\partial t$. Consider first the case where there are no discontinuities. One has

$$g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = \frac{1}{2}\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right) = 0.$$

This shows that $(\mathrm{D}/\partial t)(\partial/\partial t)$ is a spacelike vector where it is non-zero and not parallel to $\partial/\partial t$. Let $\mathbf{W}$ be a C^2 timelike vector field along $\gamma(t)$

such that $g(\mathbf{W}, \partial/\partial t) < 0$. Then one will obtain a timelike curve from p to q under the variation whose variation vector is

$$\mathbf{Z} = x\mathbf{W} + y\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}$$

with
$$x = c^{-1}\mathrm{e}^{b}\int_{t_q}^{t}\mathrm{e}^{-b}(1-\tfrac{1}{2}ya^2)\,\mathrm{d}t,$$

where
$$a^2 = g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right),$$

$$c = -g\left(\mathbf{W}, \frac{\partial}{\partial t}\right),$$

$$b = -\int_{t_q}^{t} c^{-1}g\left(W, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)\mathrm{d}t,$$

and y is a C^2 non-negative function on $[p, q]$ such that $y_p = y_q = 0$ and

$$\int_{t_q}^{t_p}\mathrm{e}^{-b}(1-\tfrac{1}{2}ya^2)\,\mathrm{d}t = 0.$$

Suppose now there is some subdivision $t_q < t_1 < t_2 < \ldots < t_p$ such that the tangent vector $\partial/\partial t$ is continuous on each segment $[t_i, t_{i+1}]$. If a segment $[t_i, t_{i+1}]$ is not a null geodesic curve, it can be varied to give a timelike curve between its endpoints. Thus one has only to show that one can obtain a timelike curve from a non-spacelike curve $\gamma(t)$ made up of null geodesic segments whose tangent vectors are not parallel at points of discontinuity $\gamma(t_i)$. The parameter t can be taken to be an affine parameter on each segment $[t_i, t_{i+1}]$. The discontinuity $[\partial/\partial t]|_{t_i}$ will be a spacelike vector, as it is the difference between two non-parallel null vectors in the same half of the null cone. Thus one can find a C^2 vector field $\mathbf{W}$ along $[t_{i-1}, t_{i+1}]$ such that $g(\mathbf{W}, \partial/\partial t) < 0$ on $[t_{i-1}, t_i]$ and $g(\mathbf{W}, \partial/\partial t) > 0$ on $[t_i, t_{i+1}]$. Then a timelike curve between $\gamma(t_{i-1})$ and $\gamma(t_{i+1})$ will be obtained from the variation with variation vector field $\mathbf{Z} = x\mathbf{W}$, where $x = c^{-1}(t_{i+1}-t_i)(t-t_{i-1})$ for $t_{i-1} \leqslant t \leqslant t_i$, and $x = c^{-1}(t_i - t_{i-1})(t_{i+1}-t)$ for $t_i \leqslant t \leqslant t_{i+1}$, where $c = -g(\mathbf{W}, \partial/\partial t)$. □

Thus if $\gamma(t)$ is not a geodesic curve, it can be varied to give a timelike curve. If it is a geodesic curve, the parameter t may be taken to be an affine parameter. One then sees that a necessary, but not sufficient, condition for a variation to yield a timelike curve is that the variation vector $\partial/\partial u$ should be orthogonal to the tangent vector $\partial/\partial t$ everywhere on $\gamma(t)$, since otherwise $(\partial/\partial t)\,g(\partial/\partial u, \partial/\partial t)$ would be positive somewhere on $\gamma(t)$. For such a variation the first derivative $(\partial/\partial u)\,g(\partial/\partial t, \partial/\partial t)$ will be zero and so one will have to examine the second derivative.

We shall therefore consider a two-parameter variation α of a null geodesic $\gamma(t)$ from q to p. The variation α will be defined as before except that, for the reason given above, we shall restrict ourselves to variations whose variation vectors

$$\left.\frac{\partial}{\partial u_1}\right|_{\substack{u_1=0\\u_2=0}} \quad \text{and} \quad \left.\frac{\partial}{\partial u_2}\right|_{\substack{u_1=0\\u_2=0}}$$

are orthogonal to the tangent vector $\partial/\partial t$ on $\gamma(t)$.

It is not convenient to study the behaviour of L under such a variation since $(-g(\partial/\partial t, \partial/\partial t))^{\frac{1}{2}}$ is not differentiable when $g(\partial/\partial t, \partial/\partial t) = 0$. Instead we shall consider the variation in:

$$\Lambda \equiv -\sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) \mathrm{d}t.$$

Clearly a necessary but not sufficient condition that a variation α of $\gamma(t)$ should yield a timelike curve from q to p is that Λ should become positive.

One has

$$\begin{aligned}\frac{1}{2}\frac{\partial^2}{\partial u_2 \partial u_1}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right) &= \frac{\partial^2}{\partial u_2 \partial t}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\right) - \frac{\partial}{\partial u_2}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)\right) \\ &= \frac{\partial^2}{\partial u_2 \partial t}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\right) - g\left(\frac{\partial}{\partial u_1}, \left\{\frac{\mathrm{D}^2}{\partial t^2}\frac{\partial}{\partial u_2} - \mathbf{R}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t}\right\}\right)\end{aligned}$$

and so

$$\begin{aligned}\frac{1}{2}\left.\frac{\partial^2 \Lambda}{\partial u_2 \partial u_1}\right|_{\substack{u_1=0\\u_2=0}} = \Sigma \int g\left(\frac{\partial}{\partial u_1}, \left\{\frac{\mathrm{D}^2}{\partial t^2}\frac{\partial}{\partial u_2} - \mathbf{R}\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t}\right\}\right) \mathrm{d}t \\ + \Sigma g\left(\frac{\partial}{\partial u}, \left[\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}\right]\right), \quad (4.49)\end{aligned}$$

This formula is very similar to that for the variation of the length of a timelike curve. It can be seen that the variation of Λ is zero for a variation vector proportional to the tangent vector $\partial/\partial t$ since $\partial/\partial t$ is null and $\mathbf{R}(\partial/\partial t, \partial/\partial t)(\partial/\partial t) = 0$ as the Riemann tensor is antisymmetric. Such a variation would be equivalent to simply reparametrizing $\gamma(t)$. Thus if one wants a variation which will give a timelike curve one need consider only the projection of the variation vector into the space S_q at each point q of $\gamma(t)$. In other words, introducing a pseudo-orthonormal basis $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \mathbf{E}_4$ along $\gamma(t)$ with $\mathbf{E}_4 = \partial/\partial t$, the variation of Λ will depend only on the components Z^m of the variation vector ($m = 1, 2$).

Proposition 4.5.11

If there is no point in $[q, p]$ conjugate to q along $\gamma(t)$ then $\mathrm{d}^2\Lambda/\mathrm{d}u^2|_{u=0}$ will be negative for any variation α of $\gamma(t)$ whose variation vector $\partial/\partial u|_{u=0}$ is orthogonal to the tangent vector $\partial/\partial t$ on $\gamma(t)$ and is not everywhere zero or proportional to $\partial/\partial t$. In other words, if there is no point in $[q, p]$ conjugate to q then there is no small variation of $\gamma(t)$ which gives a timelike curve from q to p.

The proof is similar to that for proposition 4.5.8, using instead the 2×2 matrix $\hat{A}_{mn}$ of §4.2. □

Proposition 4.5.12

If there is a point r in (q, p) conjugate to q along $\gamma(t)$ then there will be a variation of $\gamma(t)$ which will give a timelike curve from q to p.

The proof is a bit finicky since one has to show that the tangent vector becomes timelike everywhere. Let W^m be the components in the space S (see §4.2) of the Jacobi field which vanishes at q and r. It obeys

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} W^m = -R_{m4n4} W^n,$$

where for convenience t has been taken to be an affine parameter. Since W^m will be at least C^3 and since $\mathrm{d}W^m/\mathrm{d}t$ is not zero at q and r, one can write $W^m = f\hat{W}^m$ where $\hat{W}^m$ is a unit vector and f and $\hat{\mathbf{W}}$ are C^2. Then

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f + hf = 0,$$

where

$$h = \hat{W}^m \frac{\mathrm{d}^2}{\mathrm{d}t^2} \hat{W}^m + R_{m4n4} \hat{W}^m \hat{W}^n.$$

Let $x \in [r, p]$ be such that W^m is not zero in $[r, x]$. Let h_1 be the minimum value of h in $[r, x]$. Let $a > 0$ be such that $a^2 + h_1 > 0$ and let $b = \{-f(\mathrm{e}^{at} - 1)^{-1}\}|_x$. Then the field

$$Z^m = \{b(\mathrm{e}^{at} - 1) + f\} \hat{W}^m$$

will vanish at q and x and will satisfy

$$Z^m \left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} Z^m + R_{m4n4} Z^n \right) > 0 \quad \text{in} \quad (q, x).$$

We shall choose a variation $\alpha(u, t)$ of $\gamma(t)$ from q to x such that the

components in S of its variation vector $\partial/\partial u|_{u=0}$ equals Z^m and such that

$$g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\Big|_{u=0}$$

satisfies

$$g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\Big|_{u=0} + g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}\right)\Big|_{u=0} = \begin{cases} -\epsilon t & \text{for} \quad 0 \leqslant t \leqslant \frac{1}{4}t_x, \\ \epsilon(t - \frac{1}{2}t_x) & \text{for} \quad \frac{1}{4}t_x \leqslant t \leqslant \frac{3}{4}t_x, \\ \epsilon(t_x - t) & \text{for} \quad \frac{3}{4}t_x \leqslant t \leqslant t_x, \end{cases}$$

where t_x is the value of t at x, and $\epsilon > 0$ but less than the least value of Z^m $(\mathrm{d}^2Z^m/\mathrm{d}t^2 + R_{m4n4}Z^n)$ in the range $\frac{1}{4}t_x \leqslant t \leqslant \frac{3}{4}t_x$. Then by (4.49) $(\partial^2/\partial u^2)\, g(\partial/\partial t, \partial/\partial t)$ will be negative everywhere in $[q, x]$ and so for sufficiently small u, α will give a timelike curve from q to x. If one joins this curve to the section of γ from x to p, one will obtain a non-spacelike curve from q to p which is not a null geodesic curve. Thus there will be a variation of this curve which gives a timelike curve from q to p. □

By similar methods one can prove:

Proposition 4.5.13

If $\gamma(t)$ is a null geodesic curve orthogonal to a spacelike two-surface $\mathscr{S}$ from $\mathscr{S}$ to p and if there is no point in $[\mathscr{S}, p]$ conjugate to $\mathscr{S}$ along γ, then no small variation of γ can give a timelike curve from $\mathscr{S}$ to p . □

Proposition 4.5.14

If there is a point in $(\mathscr{S}, p)$ conjugate to $\mathscr{S}$ along p, then there is a variation of γ which gives a timelike curve from $\mathscr{S}$ to p. □

These results on variations of timelike and non-spacelike curves will be used in chapter 8 to show the non-existence of longest geodesics.