Foundations of General Relativity

Klaas Landsman

May 24, 2018

Contents

1 General differential geometry

Readers are supposed to be roughly familiar with this material and should look for proofs or examples elsewhere (see e.g. the Literature at the end). General relativity (GR) requires a certain way of presenting it, however, including an emphasis on coordinates and indices. This is needed both for the PDE part of the course as well as for an understanding of the physics literature.

1.1 Manifolds

1. A ***space*** always means a topological space. The topology of a space X (i.e. the set of its open sets) is denoted by $\mathscr{O}(X)$, so that $U \in \mathscr{O}(X)$ means that $U \subseteq X$ and U is open.

2. A ***(topological) manifold*** of dimension n is a paracompact Hausdorff space M such that any $x \in M$ has a nbhd $U \in \mathscr{O}(M)$ homeomorphic to some $U \in \mathscr{O}(\mathbb{R}^n)$ (equivalently, any $x \in M$ has a nbhd $U' \in \mathscr{O}(M)$ homeomorphic to $\mathbb{R}^n$ itself, or to some open ball in $\mathbb{R}^n$).[1]

3. A ***chart*** on M is a pair (U, φ) where $U \in \mathscr{O}(M)$ and $\varphi : U \to \mathbb{R}^n$ is an injective open map. We write $V = \varphi(U)$. Physicists think of a chart (U, φ) as a ***coordinate system*** on U, in that one writes $\varphi : U \to \mathbb{R}^n$ as $(\varphi^1, \ldots, \varphi^n)$, where $\varphi^i : U \to \mathbb{R}$ in terms of the standard basis of $\mathbb{R}^n$ ($i = 1, \ldots, n$), and the ***coordinates*** $(x^1, \ldots, x^n)$ of $x \in U$ of x are $x^i = \varphi^i(x)$.

4. A C^k***-atlas*** on M (where $k \in \mathbb{N} \cup \{\infty\}$) is a collection of charts $(U_\alpha, \varphi_\alpha)$, where $M = \cup_\alpha U_\alpha$ (i.e. the U_α form an open cover of M), and, whenever $U_{\alpha\beta} = U_\alpha \cap U_\beta$ is not empty, writing $V_{\alpha\beta} = \varphi_\alpha(U_{\alpha\beta}) \subset \mathbb{R}^n$, the map $\varphi_\beta \circ \varphi_\alpha^{-1} : V_{\alpha\beta} \to \mathbb{R}^n$ is C^k.

5. Two C^k-atlases $(U_\alpha, \varphi_\alpha)$ and $(U'_{\alpha'}, \varphi'_{\alpha'})$ on a topological manifold M are ***equivalent*** if their union is a C^k-atlas, i.e., if all transition functions $\varphi'_{\beta'} \circ \varphi_\alpha^{-1}$ and $\varphi_\beta \circ (\varphi'_{\alpha'})^{-1}$ (if defined) are C^k (this is indeed an equivalence relation). A C^k***-structure*** on M is an equivalence class of C^k atlases on M. A ***smooth manifold*** is a manifold with C^∞ structure.

6. Until further notice we henceforth assume that M is a smooth manifold equipped with some C^∞ atlas $(U_\alpha, \varphi_\alpha)$. A ***smooth function*** $f \in C^\infty(M)$ is a map $f : M \to \mathbb{R}$ such that for each α, the map $f \circ \varphi_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is smooth.

7. Similarly, for two smooth manifolds M, N we say that a map $\psi : M \to N$ is ***smooth*** provided one and hence each of the following equivalent conditions are satisfied:

   (a) For each $f \in C^\infty(N)$ the pullback $\psi^* f \equiv f \circ \psi$ is smooth, i.e., in $C^\infty(M)$;

   (b) For any chart (U, φ) on M and chart $(\tilde{U}, \tilde{\varphi})$ on N such that $U' = \psi(U) \cap \tilde{U} \neq \emptyset$, the function $\tilde{\varphi} \circ \psi \circ \varphi^{-1} : V' \to \tilde{V}$ is smooth, where $V' = \varphi(\psi^{-1}(U')) \subset V$.

   If $N = M$, an invertible smooth map $\psi : M \to M$ with smooth inverse is called a ***diffeomorphism***. Such maps form a group $\mathrm{Diff}(M)$ called the ***diffeomorphism group*** of M.

In the absence of contrary statements, all maps between smooth things will be smooth.

[1] It follows that M is locally compact. If M is connected, then in the above definition 'paracompact' is equivalent to 'second countable'. If M is not connected, then second countability is a stronger assumption, which is equivalent to M being paracompact with at most countably many connected components. See e.g. `http://math.harvard.edu/~hirolee/pdfs/2014-fall-230a-lecture-02-addendum.pdf`. For our application of manifolds to GR the assumption that M be second countable will do.

1.2 Tangent bundle

1. A ***derivation*** of an algebra A (over $\mathbb{R}$) is a linear map $\delta : A \to A$ satisfying

$$\delta(ab) = \delta(a)b + a\delta(b). \tag{1.1}$$

We write $\mathrm{Der}(C^\infty(M))$ for the set of all derivations on $C^\infty(M)$, seen as a (commutative) algebra with respect to pointwise operations. This is a $C^\infty(M)$-module, where the appropriate map $C^\infty(M) \times \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(M))$ is the obvious one, $(f\delta)(g) = f\delta(g)$. In addition, $\mathrm{Der}(C^\infty(M))$ is a Lie algebra,[2] under the bracket

$$[\delta_1, \delta_2] = \delta_1 \circ \delta_2 - \delta_2 \circ \delta_1. \tag{1.2}$$

2. For $M = \mathbb{R}^n$, it can be shown that each derivation of $C^\infty(\mathbb{R}^n)$ takes the form

$$\delta f(x) = \sum_{j=1}^{n} X^j(x) \frac{\partial f(x)}{\partial x^j} \equiv \delta_X(f)(x) \equiv Xf(x) \equiv X_x(f), \tag{1.3}$$

where $X \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ is an (old-fashioned) vector field on $\mathbb{R}^n$. Conversely, (1.3) defines a derivation δ_X for each vector field X, and this gives a bijection $X \leftrightarrow \delta_X$ between the set $\mathfrak{X}(\mathbb{R}^n)$ of all vector fields on $\mathbb{R}^n$ and the set $\mathrm{Der}(C^\infty(\mathbb{R}^n))$ of all derivations on $C^\infty(\mathbb{R}^n)$. In fact, this bijection is an isomorphism of $C^\infty(\mathbb{R}^n)$ modules, where $\mathfrak{X}(\mathbb{R}^n)$ carries the obvious $C^\infty(\mathbb{R}^n)$ action given by $(fX)^j(x) = f(x)X^j(x)$. Thus we may, and often will, identify $\mathrm{Der}(C^\infty(\mathbb{R}^n))$ with $\mathfrak{X}(\mathbb{R}^n)$ by looking at a vector field X as the corresponding derivation δ_X. Since a vector field $X : \mathbb{R}^n \to \mathbb{R}^n$ is given by its components $X^k : \mathbb{R}^n \to \mathbb{R}$, with $X^k \in C^\infty(\mathbb{R}^n)$, we have $\mathfrak{X}(\mathbb{R}^n) \cong \oplus^n C^\infty(\mathbb{R}^n)$ as a $C^\infty(\mathbb{R}^n)$ module, and hence also

$$\mathrm{Der}(C^\infty(\mathbb{R}^n)) \cong \mathfrak{X}(\mathbb{R}^n) \cong \oplus^n C^\infty(\mathbb{R}^n) \tag{1.4}$$

is a free $C^\infty(\mathbb{R}^n)$ module (namely the n-fold direct sum of $C^\infty(\mathbb{R}^n)$ with itself).

3. If we now *define* the vector fields $\mathfrak{X}(M)$ as $\mathrm{Der}(C^\infty(M))$ we are ready, but there is a more geometric way to define vector fields on manifolds à la $C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, namely as sections of the ***tangent bundle*** TM to M. First, a (real, locally trivial) k-dimensional ***vector bundle*** over M is an open surjective map $\pi : E \to M$, where E is a manifold, such that:

   (a) For each $x \in M$, the ***fiber*** $E_x = \pi^{-1}(x)$ is a k-dimensional (real) vector space, i.e. $E_x \cong \mathbb{R}^k$ (where k is independent of x).

   (b) M has an open cover (U_i) with diffeomorphisms $\Phi_i : \pi^{-1}(U_i) \to U_i \times \mathbb{R}^k$ such that:

      i. Each restriction $\Phi_i : E_x \to \{x\} \times \mathbb{R}^k$ is an isomorphism of vector spaces ($x \in U_i$);

      ii. If $U_{ij} \equiv U_i \cap U_j \neq \emptyset$, then $\Phi_{ij} \equiv \Phi_i \circ \Phi_j^{-1} : U_{ij} \times \mathbb{R}^k \to U_{ij} \times \mathbb{R}^k$ is the identity on the first coordinate and a vector space isomorphism on the second one.

[2] A ***Lie algebra*** (over $\mathbb{R}$) is a (real) vector space over $\mathbb{K}$ equipped with a bilinear map $[\cdot,\cdot] : A \times A \to A$ that satisfies $[a,b] = -[b,a]$ (and hence $[a,a] = 0$) as well as $[a,[b,c]] + [c,[a,b]] + [b,[c,a]] = 0$ for all $a,b,c \in A$. In finite dimension every Lie algebra comes from a Lie group (Lie's Third Theorem), but even in the case at hand one may regard $\mathrm{Der}(C^\infty(M))$ as the Lie algebra of $\mathrm{Diff}(M)$, seen as a Lie group in an appropriate (difficult) way.

A ***vector bundle map*** from $\pi_1 : E \to M$ to $\pi_2 : F \to N$ is a pair $(\varphi_f : E \to F, \varphi_b : M \to N)$ such that $\pi_2 \circ \varphi_f = \varphi_b \circ \pi_1$, and each ensuing map $\varphi_f : E_x \to F_{\varphi_b(x)}$ is linear.

The simplest k-dimensional vector bundle over M is $E = M \times \mathbb{R}^k$ with π given by projection on the first coordinate (this is called a ***trivial bundle***), but it turns out that there are many other examples (unless M is simply connected). A ***section*** (or ***cross-section***) of E is a map $s : M \to E$ such that $\pi \circ s = \mathrm{id}_M$ (i.e., $\pi(s(x)) = x$ for each $x \in M$). Cross-sections of $E = M \times \mathbb{R}^k$ are simply given by maps $\tilde{s} : M \to \mathbb{R}^k$, so that $s(x) = (x, \tilde{s}(x))$, whence

$$\Gamma(M \times \mathbb{R}^k) \cong C^\infty(M, \mathbb{R}^k), \tag{1.5}$$

where $\Gamma(E)$ is the set of smooth sections of E. Under the action $C^\infty(M) \times \Gamma(E) \to \Gamma(E)$ given by $(fs)(x) = f(x)s(x)$, $\Gamma(E)$ is a finitely generated projective module over $C^\infty(M)$.[3] The *Serre–Swan Theorem* provides an isomorphism between finitely generated projective modules $\mathscr{E}$ over $C^\infty(M)$ and vector bundles $E \to M$ over M, in such a way that $\mathscr{E} \cong \Gamma(E)$. A key step in the construction of $E = \cup_{x \in M} E_x$ (disjoint union) from $\mathscr{E}$ is the identification

$$E_x = \mathscr{E}/(C^\infty(M;x) \cdot \mathscr{E}) = \mathscr{E}/\sim_x, \tag{1.6}$$

where $C^\infty(M;x) = \{f \in C^\infty(M) \mid f(x) = 0\}$ and $C^\infty(M;x) \cdot \mathscr{E}$ is the linear span of all fs, $f \in C^\infty(M;x)$, $s \in \mathscr{E}$, so that $s_1 \sim_x s_2$ iff $s_1 - s_2 \in C^\infty(M;x) \cdot \mathscr{E}$. Then E_x is a vector space under the linear structure inherited from $\mathscr{E}$ (e.g. $[s_1]_x + [s_2]_x = [s_1 + s_2]_x$, $0 = [0]_x$ etc., where $[s]_x$ is the equivalence class of s with respect to $\sim_x$). Subsequently, the smooth structure of E may be (re)constructed from $\mathscr{E}$ by reinterpreting $\hat{s} \in \mathscr{E}$ as a map $s : M \to E$ through $s(x) = [s]_x \in E_x$, and requiring $\hat{s} \to s$ to be an isomorphism $\mathscr{E} \xrightarrow{\cong} \Gamma(E)$.[4]

4. The tangent bundle $\pi : TM \to M$ is the vector bundle constructed from $\mathscr{E} = \mathrm{Der}(C^\infty(M))$ according to the above procedure.[5] In this case, we have a (linear) isomorphism

$$\mathrm{Der}(C^\infty(M))/\sim_x \cong \mathrm{Der}_x(C^\infty(M)), \tag{1.7}$$

where the the right-hand side is the (vector) space of ***point derivations*** at x, defined as linear maps $\delta_x : C^\infty(M) \to \mathbb{R}$ that satisfy

$$\delta_x(fg) = \delta_x(f)g(x) + f(x)\delta_x(g). \tag{1.8}$$

Each derivation $\delta \in \mathrm{Der}(C^\infty(M))$ defines a point derivation $\delta_x \in \mathrm{Der}_x(C^\infty(M))$ by

$$\delta_x(f) = \delta(f)(x), \tag{1.9}$$

and the isomorphism (1.7) is given by $[\delta]_x \mapsto \delta_x$. The fibers $TM_x \equiv T_xM$ of the bundle

$$TM = \cup_{x \in M} T_xM, \tag{1.10}$$

which by definition is the tangent bundle, may therefore be written as

$$T_xM = \mathrm{Der}_x(C^\infty(M)). \tag{1.11}$$

[3] A $C^\infty(M)$-module $\mathscr{E}$ is called ***finitely generated projective*** if there exists a $C^\infty(M)$-module $\mathscr{F}$ such that $\mathscr{E} \oplus \mathscr{F}$ is free, i.e. isomorphic to a direct sum of copies of $C^\infty(M)$.

[4] This isomorphism sends $C^\infty(M;x) \cdot \mathscr{E}$ to $\Gamma(E;x) = \{s \in \Gamma(E) \mid s(x) = 0\}$, so that $\Gamma(E)/\Gamma(E;x) \cong E_x$.

[5] $\mathrm{Der}(C^\infty(M))$ may no longer be free over $C^\infty(M)$, as in the case $M = \mathbb{R}^n$, but using charts one can show that it is at least finitely generated projective.

As can be seen in local charts (where the situation is the same as for $M = \mathbb{R}^n$), the point derivations at x form an n-dimensional vector space with basis $(\partial_1, \ldots, \partial_n)$, where $\partial_i = \partial/\partial x^i$, seen as an element of T_xM, maps $f \in C^\infty(M)$ to $\partial_i f(x)$. Thus TM is an n-dimensional vector bundle over M, whose smooth structure is defined such that each derivation δ of $C^\infty(M)$ is given by a cross-section $x \mapsto \delta_x$ of TM, where $\delta_x \in T_xM$. Thus

$$\mathrm{Der}(C^\infty(M)) \cong \Gamma(TM) \equiv \mathfrak{X}(M). \tag{1.12}$$

Consequently, a ***vector field*** X on M, written $X \in \mathfrak{X}(M)$, is a map $x \mapsto X_x$ (or $x \mapsto X(x)$), where $x \in M$ and $X_x \in T_xM$, closely related to (but to be distinguished from) the corresponding derivation $\delta_X \in \mathrm{Der}(C^\infty(M))$; the connections is

$$X_x(f) = \delta_X(f)(x). \tag{1.13}$$

Hence we think of a *vector field* $X \in \mathfrak{X}(M)$ as the collection of all vectors $X_x \in T_xM$, whereas we think of the corresponding *derivation* as a single global operation on $C^\infty(M)$.

5. *Point* derivations ***push forward*** under maps $\psi : M \to N$: for $x \in M$ we have linear maps

$$\psi'_x : T_xM \to T_{\psi(x)}N; \tag{1.14}$$

$$(\psi'_x \delta_x)(g) = \delta_x(\psi^* g) \ \ (g \in C^\infty(N)). \tag{1.15}$$

Collecting these maps gives a vector bundle map $\psi' : TM \to TN$ (also called ψ_* or $T\psi$).

However, *derivations* (or vector fields) push forward only if $\psi : M \to N$ is a *diffeomorphism*: the map $\psi_* : \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(N))$, or $\psi_* : \mathfrak{X}(M) \to \mathfrak{X}(N)$, is given by[6]

$$\psi_*(\delta) = (\psi^{-1})^* \circ \delta \circ \psi^*. \tag{1.16}$$

6. One may study ***tangent vectors*** $X_x \in T_xM$ in their own right (i.e., not necessarily as the values of some vector field X at x). Each tangent vector is (*nomen est omen!*) tangent to some ***curve*** γ through x, i.e. a map $\gamma : I \to M$ where $I \subset \mathbb{R}$ is some open or closed interval we often (as in: now) assume to contain 0, such that $\gamma(0) = x$. In other words,

$$X_x(f) = \frac{d}{dt} f(\gamma(t))_{|t=0}, \tag{1.17}$$

which *symbolically* may be written as $X_x = \dot\gamma \equiv d\gamma/dt$, or even as $X_x = d/dt$, with γ understood. This description gives a geometric perspective on the push-forward of T_xM just described: if $X = d\gamma/dt$ is tangent to γ, then $\psi' X = d(\psi \circ \gamma)/dt$ is tangent to $\psi(\gamma)$.

In a chart $\varphi : U \to \mathbb{R}^n$ with $x \in U$, the components X^i_φ of X_x are defined by

$$X^i_\varphi = X\varphi^i(x) = \frac{d}{dt}\varphi^i(\gamma(t))_{|t=0} = \frac{d}{dt}\gamma^i(t)_{|t=0}, \tag{1.18}$$

where $\gamma^i(t) = \varphi^i(\gamma(t))$. Strictly speaking, we have $\varphi_* X_x = \sum_{i=1}^n X^i_\varphi \partial_i \in T_{\varphi(x)}\mathbb{R}^n$; in practice, this is often written as $X = \sum_i X^i \partial_i \in T_xM$, leaving the role of the chart φ implicit.

[6] One needs $(\psi^{-1})^*$ even if $N = M$, since $\delta \circ \psi^*$ fails to be a derivation of $C^\infty(M)$. Please check!

However, the precise version (1.18) gives the transformation rule for vectors under a change of charts (i.e. of coordinates): if $x \in U_\alpha \cap U_\beta$, then (1.17) and (1.18) imply

$$X^i_\beta = \sum_j \frac{\partial x^i_\beta}{\partial x^j_\alpha} X^j_\alpha, \tag{1.19}$$

where $X^i_\beta \equiv X^i_{\varphi_\beta}$ etc., and the coordinates $x^i_\beta = \varphi^i_\beta(x)$ of x with respect to φ_β are seen as functions of the coordinates $x^i_\alpha = \varphi_\alpha(x)$ of x with respect to φ_α, namely by putting

$$x^i_\beta(x_\alpha) = \varphi^i_\beta \circ \varphi_\alpha^{-1}(x_\alpha), \tag{1.20}$$

which is really a restatement of the tautology $\varphi^i_\beta = \varphi^i_\beta \circ \varphi_\alpha^{-1} \circ \varphi_\alpha$ (on $U_\alpha \cap U_\beta$).

In both differential geometry and GR it is important to distinguish (1.19), which is a *change of coordinates formula for a given tangent vector*, from a similar formula that expresses in coordinates the *push-forward of a tangent vector* under a map $\psi : M \to M$. Suppose for simplicity that $x \in U$ and also $\psi(x) \in U$. Then, writing $X^i_\varphi \equiv X^i$ as above, as well as $\psi^i = \varphi^i \circ \psi \circ \varphi^{-1}$ (which near x is a function from V to $\mathbb{R}$), we have

$$(\psi' X)^i = \sum_j \frac{\partial \psi^i}{\partial x^j} X^j. \tag{1.21}$$

Increasing potential confusion, although (1.19) gives different coordinate descriptions of the same vector X in TM, it may also be seen as the formula for the push-forward of the vector $\varphi'_\alpha X$ in $T\mathbb{R}^n$ under the map $\varphi_\beta \circ \varphi_\alpha^{-1}$ from V_α to V_β within $\mathbb{R}^n$.

7. Vector fields X (or, equivalently, derivations) may be 'integrated', at least *locally*, in the following sense. We say that a curve $\gamma : I \to M$ ***integrates*** X if $X_{\gamma(t)} = d\gamma(t)/dt$, or

$$X_{\gamma(t)}(f) = \frac{d}{dt} f(\gamma(t)), \ \ (t \in I), \tag{1.22}$$

for each f defined in a nbhd of $\gamma(I)$. Describing γ and X by their coordinate functions $\gamma^j : I \to \mathbb{R}$ and $X^j : V \to \mathbb{R}$ relative to some chart $\varphi : U \to V$, eq. (1.22) becomes

$$\frac{d\gamma^j(t)}{dt} = X^j(\gamma^1(t), \ldots, \gamma^n(t)), \ (j = 1, \ldots, n). \tag{1.23}$$

For given X, an integrating curve γ is therefore found by solving a system of n coupled ODE, subject to some initial condition. The theory of ODE shows that for smooth X (as we assume), this can always be done locally: for each $x_0 \in M$ there exists an open interval $I \subset \mathbb{R}$ (with $0 \in I$) and a curve $\gamma : I \to M$ on which (1.22) holds with $\gamma(0) = x_0$. This solution is unique in the sense that if two curves $\gamma_1 : I_1 \to M$ and $\gamma_2 : I_2 \to M$ both satisfy (1.22) with $\gamma_1(0) = \gamma_2(0) = x_0$, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$. Taking unions, it follows that there exists a maximal interval I on which γ is defined. However, curves that integrate X may not be defined for all t, i.e., for $I = \mathbb{R}$. This complicates the important concept of a ***flow*** of a vector field X, which is meant to encapsulate all integral curves of X.

In the simplest case where for any $x \in M$ there is a curve $\gamma : \mathbb{R} \to M$ satisfying (1.22) with $\gamma(0) = x$, we say that $X \in \mathfrak{X}(M)$ is ***complete***.[7] In that case, the ***flow*** of X is a smooth map $\psi : \mathbb{R} \times M \to M$, written $\psi_t(x) \equiv \psi(t,x)$, that satisfies

$$\psi_0(x) = x; \tag{1.24}$$

$$X_{\psi_t(x)} f = \frac{d}{dt} f(\psi_t(x)) \tag{1.25}$$

for all $x \in M$, $t \in \mathbb{R}$, and $f \in C^\infty(M)$. Thus the flow ψ of X gives "the" integral curve γ of X through x_0 by $\gamma(t) = \psi_t(x_0)$. Any complete vector field has a unique flow. Uniqueness implies both that M is a disjoint union of the integral curves of X (which can never cross each other because of the uniqueness of the solution), and the composition rule

$$\psi_s \circ \psi_t = \psi_{s+t}. \tag{1.26}$$

From a group-theoretic point of view, a flow is therefore an action of $\mathbb{R}$ (as an additive group) on M *that in addition integrates* X in the sense of (1.25). In particular, (1.26) implies $\psi_{-t} = \psi_t^{-1}$, so that each $\psi_t : M \to M$ is automatically a diffeomorphism of M.

If X is not complete (a case that will be of great interest to GR!), we first define the ***domain*** $D_X \subset \mathbb{R} \times M$ of ψ as the set of all $(t,x) \in \mathbb{R} \times M$ for which there exists an open interval $I \subset \mathbb{R}$ containing 0 and t as well as a (necessarily unique) curve $\gamma : I \to M$ that satisfies (1.22) with initial condition $\gamma(0) = x$. Obviously $\{0\} \times M \subset D_X$, and (less trivially) it turns out that D_X is open. Then a flow of X is a map $\psi : D_X \to M$ that satisfies (1.24) for all x and (1.25) for all $(t,x) \in D_X$. Eq. (1.26) then holds whenever defined.

8. As a first application of flows, let us define the ***Lie derivative*** $\mathscr{L}_X Y$ of some vector field $Y \in \mathfrak{X}(M)$ with respect to another vector field $X \in \mathfrak{X}(M)$ by

$$\mathscr{L}_X Y(x) = \lim_{t \to 0} \frac{Y_{\psi_t(x)} - \psi_t'(Y_x)}{t} = \lim_{t \to 0} \frac{\psi_{-t}'(Y_{\psi_t(x)}) - Y_x}{t} \tag{1.27}$$

where ψ is the flow of X. Note that $Y_{\psi_t(x)} - Y_x$ would be undefined, since $Y_{\psi_t(x)} \in T_{\psi_t(x)}M$ whilst $Y_x \in T_x M$ and these are different vector spaces; the push-forward ψ_t' serves to move Y_x to $T_{\psi_t(x)}M$. A simple computation (Frankel, §4.1) then yields the well-known result

$$\mathscr{L}_X Y = [X,Y], \tag{1.28}$$

where the ***commutator*** is defined by $[X,Y]f = X(Y(f)) - Y(X(f))$. Note that neither XY nor YX is a vector field, yet $[X,Y] \in \mathfrak{X}(M)$ is, as may be checked by seeing vector fields as derivations; see the comments after (1.1). Thus $\mathfrak{X}(M)$ is a Lie algebra.

In coordinates, where $X = \sum_i X^i \partial_i$ and $Y = \sum_j Y^j \partial_j$, we have $[X,Y] = \sum_i [X,Y]^i \partial_i$, with

$$[X,Y]^i = \sum_j (X^j \partial_j Y^i - Y^j \partial_j X^i). \tag{1.29}$$

[7] A *sufficient* condition for X to be complete is that it has *compact support* (so if M is compact, then every vector field is complete).

1.3 Cotangent bundle and other tensor bundles

Now that we have the tangent bundle TM, all other vector bundles relevant to GR follow. First, the ***cotangent bundle*** T^*M is defined as $T^*M = \cup_{x\in M} T_x^*M$, where the fibers

$$T_x^*M = (T_xM)^* \equiv \mathrm{Hom}(T_xM, \mathbb{R}) \tag{1.30}$$

consist of all linear maps $\theta_x : T_xM \to \mathbb{R}$, i.e., are the dual vector spaces to T_xM, and the smooth structure of T^*M stipulates that elements $\theta \in \Gamma(T^*M) \equiv \Omega^1(M) \equiv \Omega(M)$, called ***covectors*** (or ***1-forms***), consist of those maps $x \mapsto \theta_x$ for which the function $x \mapsto \theta_x(X_x)$ from M to $\mathbb{R}$ is smooth for each $X \in \mathfrak{X}(M)$. Since $T_xM \cong \mathbb{R}^n$ we also have $T_x^*M \cong \mathbb{R}^n$, so that, like TM, also T^*M is an n-dimensional vector bundle over M. In a coordinate systems (x^i) defined by some chart ,T_x^*M has basis $(dx^1, \ldots, dx^n)$ defined by $dx^i(\partial_j) = \delta^i_j$; this is the ***dual basis*** to the standard basis $(\partial_1, \ldots, \partial_n)$ of T_xM defined earlier. Writing $\theta = \sum_i \theta_i dx^i$, the components θ_i are given by

$$\theta_i = \theta(\partial_i). \tag{1.31}$$

In particular, any $f \in C^\infty(M)$ defined a cross-section $df \in \Omega(M)$ by

$$df_x = \sum_i \left(\frac{\partial f}{\partial x^i}\right)(x) dx^i, \tag{1.32}$$

or, free of coordinates, by

$$df(X) = X(f). \tag{1.33}$$

1. More generally, let (e_a) be a basis of T_xM, with dual basis (ω^a) of T_x^*M (i.e. $\omega^a(e_b) = \delta^a_b$). Once again, if we expand $\theta = \sum_a \theta_a \omega^a$, we have $\theta_a = \theta(e_a)$. This may be done at a single point, but bases like $(\partial_1, \ldots, \partial_n)$ and $(dx^1, \ldots, dx^n)$ are defined at each $x \in U$ on which the coordinates $x^i = \varphi^i(x)$ are defined. Similarly, some basis (e_a) may be defined at each $x \in U$, where $U \in \mathscr{O}(M)$ is not even necessarily the domain of a chart. In that case (e_a) is called a (moving) ***frame*** or an ***n-bein***. Abstractly, if $E \to M$ is a k-dimensional vector bundle, one may locally find k linearly independent cross-sections $(u_1, \ldots, u_k)$ of E and expand any $s \in \Gamma(E)$ by $s(x) = \sum_j s_j(x) u_j(x)$, where $s_j \in C^\infty(M)$ and $u_j \in \Gamma(E)$.

2. Whereas tangent vectors *push forward* from M to N under maps $\psi : M \to N$, covectors ***pull back*** from N to M, like functions: besides the pull-back $\psi^* : C^\infty(N) \to C^\infty(M)$ on functions, any (smooth) ψ map induces a pullback $\psi^* : \Omega(N) \to \Omega(M)$ on 1-forms by

$$(\psi^*\theta)_x(X_x) = \theta_{\psi(x)}(\psi'_x X_x), \tag{1.34}$$

   where $\theta \in \Omega(N)$ and $X_x \in T_xM$. For any $f \in C^\infty(N)$ with $df \in \Omega(N)$, this yields

$$\psi^*(df) = d(\psi^* f). \tag{1.35}$$

   However, a decent vector bundle map $\psi^* : T^*N \to T^*M$ is defined only if ψ is a diffeomorphism: with $\theta_y \in T_yN$, $y \in N$, and $x = \psi^{-1}(y) \in M$, $\psi_y^*(\theta_y) \in T_x^*M$ is defined by

$$(\psi_y^*\theta_y)(X_x) = \theta_y(\psi'_x X_x). \tag{1.36}$$

   If ψ is merely injective, then we still obtain a map $\psi^* : T^*(\psi(M)) \to T^*M$ in this way.

3. Using the canonical isomorphism $V^{**} \cong V$ for any *finite-dimensional* vector space V, given by the map $v \mapsto \hat{v}$ from V to V^{**}, where $\hat{v}(\theta) = \theta(v)$, we reinterpret TM as $T^{**}M$, in that we now look at T_xM as $(T_x^*M)^*$. To blur the distinction between V and V^{**} one may write $\langle \theta, v \rangle$ for $\theta(v)$, and $\langle v, \theta \rangle$ for $\hat{v}(\theta)$, and simply declare that $\langle \theta, v \rangle = \langle v, \theta \rangle$. In this spirit, for any $(k,l) \in \mathbb{N} \times \mathbb{N}$ we define a vector bundle $T^{(k,l)}M$ over M via its fibers

$$T_x^{(k,l)}M = \mathrm{Hom}((T_xM)^k \times (T_x^*M)^l, \mathbb{R}), \tag{1.37}$$

i.e. the vector space of $k+l$-fold multilinear maps from $(T_xM)^k \times (T_x^*M)^l$ to $\mathbb{R}$, with total space $T^{(k,l)}M = \cup_{x \in M} T_x^{(k,l)}M$. We then define $\Gamma(T^{(k,l)}M)$ as the set of cross-sections $x \mapsto \tau_x$ (where $\tau_x \in T_x^{(k,l)}M$) for which the map $x \mapsto \tau_x(X_1(x), \ldots, X_k(x); \theta^1(x), \ldots, \theta^l(x))$ from M to $\mathbb{R}$ is smooth for each $(X_1, \ldots, X_k; \theta^1, \ldots, \theta^l)$ with $X_i \in \mathfrak{X}(M)$ and $\theta^j \in \Omega(M)$. As before, this equips $T^{(k,l)}M$ with a manifold structure (in that we declare $\Gamma(T^{(k,l)}M)$ to be the space of smooth cross-sections of $T^{(k,l)}M$). Equivalently, we may define $T_x^{(k,l)}M$ as the tensor product of k copies of T_x^*M and l copies of T_xM, making $T^{(k,l)}M$ the (vector bundle) tensor product of k copies of T^*M and l copies of TM. We then have

$$T^{(0,0)}M = M \times \mathbb{R}; \tag{1.38}$$

$$T^{(1,0)}M = T^*M; \tag{1.39}$$

$$T^{(0,1)}M = TM. \tag{1.40}$$

In GR, $T^{(2,0)}M$ (carrying the metric) and $T^{(3,1)}M$ (where curvature lives) will also be important. Elements of $\Gamma(T^{(k,l)}M)$ are called ***tensors*** (or ***tensor fields***, in which case each τ_x is regarded as a tensor). If $\alpha_i \in T_x^*M$ $(i = 1, \ldots, k)$ and $v_j \in T_xM$ $(j = 1, \ldots, l)$, then

$$\alpha_1 \otimes \cdots \otimes \alpha_k \otimes v_1 \otimes \cdots \otimes v_l \in T_x^{(k,l)}M,$$

having to be a multilinear map from $(T_xM)^k \times (T_x^*M)^l$ to $\mathbb{R}$, is naturally defined by

$$\alpha_1 \otimes \cdots \otimes \alpha_k \otimes v_1 \otimes \cdots \otimes v_l(X_1, \ldots, X_k; \theta^1, \ldots, \theta^l) = \alpha_1(X_1) \cdots \alpha_k(X_k) v_1(\theta^1) \cdots v_l(\theta^l).$$

All this can be rewritten in terms of *indices*. In terms of the (coordinate) basis $(\partial_1, \ldots, \partial_n)$ of T_xM with dual basis $(dx^1, \ldots, dx^n)$ of T_x^*M, the fiber $T_x^{(k,l)}M$ then has a basis

$$(dx^{i_1} \otimes \cdots \otimes dx^{i_k} \otimes \partial_{j_1} \otimes \cdots \otimes \partial_{j_l}), \tag{1.41}$$

where all indices run from 1 to n. *Thus $T^{(k,l)}M$ is an n^{k+l}-dimensional vector bundle.* Like vectors, tensors at x may be specified by their components with respect to some basis of T_xM and associated dual basis of T_x^*M, In the usual coordinate basis (∂_i) we have

$$\tau_x = \tau^{j_1 \cdots j_l}_{i_1 \cdots i_k}(x)\, dx^{i_1} \otimes \cdots dx^{i_k} \otimes \partial_{j_1} \otimes \cdots \otimes \partial_{j_l}; \tag{1.42}$$

$$\tau^{j_1 \cdots j_l}_{i_1 \cdots i_k}(x) = \tau_x(\partial_{i_1}, \ldots, \partial_{i_k}; dx^{j_1}, \ldots, dx^{j_l}), \tag{1.43}$$

where we use the ***Einstein summation convention: repeated indices are summed over***. Thus the right-hand side of (1.42) should really be preceded by $\sum_{i_1,\ldots,i_k,j_1,\ldots,j_l=1}^{n}$. Similarly, in an arbitrary basis (e_a) of T_xM with dual basis (θ^a) of T_x^*M one has

$$\tau_x = \tau^{b_1 \cdots b_l}_{a_1 \cdots a_k}(x)\, \theta^{a_1} \otimes \cdots \theta^{a_k} \otimes e_{b_1} \otimes \cdots \otimes e_{b_l}; \tag{1.44}$$

$$\tau^{b_1 \cdots b_l}_{a_1 \cdots a_k}(x) = \tau_x(e_{a_1}, \ldots, e_{a_k}; \theta^{b_1}, \ldots, \theta^{b_l}). \tag{1.45}$$

4. We write $\mathfrak{X}^{(k,l)}(M)$ for $\Gamma(T^{(k,l)}M)$, so that $\mathfrak{X}^{(0,0)}(M) = C^\infty(M)$, $\mathfrak{X}^{(0,1)}(M) = \mathfrak{X}(M)$, and $\mathfrak{X}^{(1,0)}(M) = \Omega(M)$. A tensor $\tau \in \mathfrak{X}^{(k,l)}(M)$ of type (k,l) maps k vector fields $(X_1,\ldots,X_k)$ and l covector fields $(\theta^1,\ldots,\theta^l)$ to a smooth function on M by pointwise evaluation, i.e.

$$\tau : \mathfrak{X}(M)^k \times \Omega(M)^l \to C^\infty(M); \tag{1.46}$$

$$\tau(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) : x \mapsto \tau_x(X_1(x),\ldots X_k(x);\theta^1(x),\ldots,\theta^l(x)). \tag{1.47}$$

This map is evidently $k+l$- multilinear linear over $C^\infty(M)$, in that

$$\tau(f_1X_1,\ldots,f_kX_k,g_1\theta^1,\ldots,g_l\theta^l) = f_1\cdots f_k\cdot g_1\cdots g_l\cdot\tau(X_1,\ldots,X_k;\theta^1,\ldots,\theta^l), \tag{1.48}$$

for all $f_i, g_j \in C^\infty(M)$; here we use the fact that $\mathfrak{X}(M)$ and $\Omega(M)$ are $C^\infty(M)$ modules.

Conversely, a map $\tau : \mathfrak{X}(M)^k \times \Omega(M)^l \to C^\infty(M)$ satisfying (1.48) is given by a tensor $\tau \in \mathfrak{X}^{(k,l)}(M)$ through (1.47). The proof is easy in local coordinates, where (1.48) yields

$$\begin{aligned}\tau(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) &= \tau(X_1^{i_1}\partial_{i_1},\ldots,X_k^{i_k}\partial_{i_k};\theta^1_{j_1}dx^{j_1},\ldots\theta^l_{j_l}dx^{j_l})\\ &= X_1^{i_1}\cdots X_k^{i_k}\cdot\theta^1_{j_1}\cdots\theta^l_{j_l}\tau(\partial_{i_1},\ldots,\partial_{i_k};dx^{j_1},\ldots dx^{j_l}),\end{aligned} \tag{1.49}$$

so if we define the components $\tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)$ of τ_x by (1.43) and subsequently define τ_x itself by (1.42), we have found the desired tensor that reproduces the given map τ via (1.47).[8]

5. Eqs. (1.42) - (1.43) imply the transformation properties of tensors under changes of coordinates (i.e. charts), which physicists even use to *define* tensors: in the situation of (1.19),

$$(\tau_\beta)^{j_1\cdots j_l}_{i_1\cdots i_k}(x_\beta) = \frac{\partial x^{j_1}_\beta}{\partial x^{j'_1}_\alpha}\cdots\frac{\partial x^{j_l}_\beta}{\partial x^{j'_l}_\alpha}\cdot\frac{\partial x^{i'_1}_\alpha}{\partial x^{i_1}_\beta}\cdots\frac{\partial x^{i'_k}_\alpha}{\partial x^{i_k}_\beta}\cdot(\tau_\alpha)^{j'_1\cdots j'_l}_{i'_1\cdots i'_k}(x_\alpha), \tag{1.50}$$

where the 'new' coordinates $(x_\beta) = (x^1_\beta,\ldots,x^n_\beta)$ are functions of the 'old' coordinates $(x_\alpha) = (x^1_\alpha,\ldots,x^n_\alpha)$, cf. (1.20), and hence the matrix $(\partial x^{i'_1}_\alpha/\partial x^{i_1}_\beta)$ is defined as the inverse of the matrix $(\partial x^{i_1}_\beta/\partial x^{i'_1}_\alpha)$, *both* seen as functions of the (x^i_α). Note that the argument x_β in (1.50) refers to the *same point* $x \in M$ as the argument x_α (but in *different coordinates*).

6. Let $\psi : M \to N$ be smooth. Through its coordinate expression, we may then define

$$\psi^{(0,l)}_x : T^{(0,l)}_x M \to T^{(0,l)}_{\psi(x)}N; \tag{1.51}$$

$$\psi^{(0,l)}_x\tau_x = \tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)\,\psi'_x(\partial_{j_1})\otimes\cdots\otimes\psi'_x(\partial_{j_l}), \tag{1.52}$$

which combine into a single (vector bundle) map $\psi^{(0,l)} : T^{(0,l)}M \to T^{(0,l)}N$. However, as in the special case $\psi^{(0,1)} = \psi'$ (from TM to TN), we are generally unable to define maps $\mathfrak{X}^{(0,l)}(M) \to \mathfrak{X}^{(0,l)}(N)$. Similarly, ψ induces maps $\psi^{(k,0)} : \mathfrak{X}^{(k,0)}(N) \to \mathfrak{X}^{(k,0)}(M)$ by the obvious generalization of (1.34), but in general we cannot define maps $T^{(k,0)}N \to T^{(k,0)}M$.

[8] Similarly for vector bundles $E \to M$: a map $\tau : \Gamma(E) \to \Gamma(E)$ is induced by a cross-section of the vector bundle $\mathrm{End}(E)$ iff it is $C^\infty(M)$-linear. Here $\mathrm{End}(E) = \cup_{x\in M}\mathrm{Hom}(E_x,E_x)$, topologized (as usual) by asking precisely those maps $x \mapsto L_x$, where $L_x \in \mathrm{Hom}(E_x,E_x)$, to be smooth for which all maps $x \mapsto L_x s(x)$ are smooth, $s \in \Gamma(E)$.

These defects can be overcome if ψ is *invertible* (with smooth inverse), in which case we may as well take $N = M$ and assume that $\psi : M \to M$ is a diffeomorphism. Then ψ acts on vectors in TM, whereas ψ^{-1} acts on covectors (1-forms) in T^*M. We may then define

$$\psi_x^{(k,l)}\tau_x = \tau^{j_1\cdots j_l}_{i_1\cdots i_k}(x)\cdot(\psi^{-1})^*_x(dx^{i_1})\otimes\cdots\otimes(\psi^{-1})^*_x(dx^{i_k})\otimes\psi'_x(\partial_{j_1})\otimes\cdots\otimes\psi'_x(\partial_{j_l}), \quad (1.53)$$

as an element of $T^{(k,l)}_{\psi(x)}M$; note that $(\psi^{-1})^*_x$ maps T^*_xM to $T^*_{\psi(x)}M$ whilst ψ'_x maps T_xM to $T_{\psi(x)}M$. This gives corresponding formulae for cross-sections. Thus a *diffeomorphism* $\psi : M \to M$ induces all maps $\psi^{(k,l)}_* : T^{(k,l)}M \to T^{(k,l)}M$ as well as (abuse of notation) $\psi^{(k,l)}_* : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k,l)}(M)$, recovering $\psi^{(0,1)}_* = \psi' = \psi_*$. We may also replace ψ in (1.53) by ψ^{-1}. This gives similar maps we denote by $\psi^*_{(k,l)}$, recovering $\psi^*_{(1,0)} = \psi^*$.

7. A natural operation on tensors, which is often used in GR, is ***tensoring***: if $\tau_1 \in \mathfrak{X}^{(k_1,l_1)}(M)$ and $\tau_2 \in \mathfrak{X}^{(k_2,l_2)}(M)$, then $\tau_1\otimes\tau_2 \in \mathfrak{X}^{(k_1+k_2,l_1+l_2)}(M)$ is defined by *concatenation*, i.e.

$$\tau_1\otimes\tau_2(X_1,\ldots,X_{k_1},Y_1,\ldots Y_{k_2};\theta^1,\ldots,\theta^{l_1},\rho^1,\ldots,\rho^{l_2}) = \quad (1.54)$$

$$\tau_1(X_1,\ldots,X_{k_1};\theta^1,\ldots,\theta^{l_1})\cdot\tau_2(Y_1,\ldots Y_{k_2};\rho^1,\ldots,\rho^{l_2}). \quad (1.55)$$

Indeed, $\mathfrak{X}^{(k,l)}(M)$ itself arose in this way by tensoring copies of $\mathfrak{X}^{(1,0)}(M)$ and $\mathfrak{X}^{(0,1)}(M)$.

8. Another important operation for GR is ***(index) contraction***: If $k > 0$ and $l > 0$, then a tensor $\tau \in \mathfrak{X}^{(k,l)}(M)$ may be contracted along one fixed upper and one lower index, say i and j (the result depends on this choice) to a tensor $\sigma \in \mathfrak{X}^{(k-1,l-1)}(M)$ with two fewer indices. Let (e_a) be a basis of T_xM, with dual basis (ω^a) of T^*_xM (i.e. $\omega^a(e_b) = \delta^a_b$); in local coordinates one could take the (∂_i) basis, with dual (dx^i). Then

$$\sigma^{b_1,\ldots,\hat{b}_i,\ldots,b_l}_{a_1,\ldots,\hat{a}_j,\ldots,a_k}(x) = \tau^{b_1,\ldots,a,\ldots,b_l}_{a_1,\ldots,a,\ldots,a_k}(x), \quad (1.56)$$

where, according to our standing Einstein summation convention, a is summed over, and (as usual) a hat means that the given index is omitted. This is easily seen to be independent of the basis. In GR (and also in Riemannian geometry), an important application will be to the *Riemann tensor* $R \in \mathfrak{X}^{(3,1)}(M)$, which is contracted to the *Ricci tensor* $R_{ab} = R^c_{acb}$.

9. The ***Lie derivative*** $\mathscr{L}_X$ may be extended to a map $\mathscr{L}^{(k,l)}_X \equiv \mathscr{L}_X : \mathfrak{X}^{(k,l)}M \to \mathfrak{X}^{(k,l)}M$ by

$$\mathscr{L}_X\tau = \lim_{t\to 0} t^{-1}(\psi_t^*(\tau) - \tau) \qquad (\tau \in \mathfrak{X}^{(k,l)}M), \quad (1.57)$$

cf. (1.27). In local coordinates, this gives the following explicit formula:

$$\begin{aligned}(\mathscr{L}_X\tau)^{j_1\cdots j_l}_{i_1\cdots i_k} &= X^i\partial_i\tau^{j_1\cdots j_l}_{i_1\cdots i_k} + (\partial_{i_1}X^i)\tau^{j_1\cdots j_l}_{i\cdots i_k} + \cdots + (\partial_{i_n}X^i)\tau^{j_1\cdots j_l}_{i_1\cdots i} \\ &\quad - (\partial_j X^{j_1})\tau^{j\cdots j_l}_{i_1\cdots i_k} - \cdots - (\partial_j X^{j_l})\tau^{j_1\cdots j}_{i_1\cdots i_k}, \end{aligned} \quad (1.58)$$

of which (1.29) is clearly a special case. It follows from either (a)–(d) or (1.58) that

$$[\mathscr{L}_X,\mathscr{L}_Y] = \mathscr{L}_{[X,Y]}. \quad (1.59)$$

One may equivalently define the $\mathscr{L}_X$ as the unique linear maps satisfying the rules:

(a) $\mathscr{L}_X f = Xf$ for functions $f \in C^\infty(M) \equiv \mathfrak{X}^{(0,0)}M$;
(b) $\mathscr{L}_X Y = [X,Y]$ for vector fields $Y \in \mathfrak{X}(M) \equiv \mathfrak{X}^{(0,1)}M$;
(c) $\mathscr{L}_X(\theta(Y)) = (\mathscr{L}_X\theta)(Y) + \theta(\mathscr{L}_X Y)$ for covector fields $\theta \in \Omega(M) \equiv \mathfrak{X}^{(1,0)}M$;
(d) $\mathscr{L}_X(\sigma\otimes\tau) = (\mathscr{L}_X\sigma)\otimes\tau + \sigma\otimes\mathscr{L}_X\tau$ (***Leibniz rule***) for all higher-order tensors.

2 Metric differential geometry

The material in his chapter may no longer be familiar to all readers, and so it will be developed in some more detail compared to the previous chapter, but since this is not primarily a course in (semi) Riemannian geometry but a course in GR, proofs and examples will remain terse.

2.1 (Semi) Riemannian metrics

The main tensor in this course will be the ***metric tensor*** $g \in \mathfrak{X}^{(2,0)}M$, for which each bilinear map $g_x : T_xM \times T_xM \to \mathbb{R}$ is *symmetric* (i.e. $g_x(X_x,Y_x) = g_x(Y_x,X_x)$) and *nondegenerate* (in that $g_x(X_x,Y_x) = 0$ all $Y_x \in T_xM$ iff $X_x = 0$). It follows from elementary linear algebra that each g_x can be diagonalized, in that T_xM has a basis (e_a) for which $g_x(e_a,e_b) = \varepsilon_a\delta_{ab}$, where $\varepsilon_a = \pm 1$. Furthermore, the number of positive and negative ε_a is independent of the basis and is called the ***signature*** of g_x. If M is connected, then the signature is independent of x, and even if M is not, we assume this. Thus the signature is a property of g, usually denoted by

$$(+\cdots+-\cdots-) \text{ or } (-\cdots-+\cdots+).$$

1. The metric is called ***Riemannian*** if the signature is $(+\cdots+)$, i.e., if each g_x is positive definite (which, given the assumption of symmetry, implies that it is nondegenerate, so a Riemannian metric is one for which each g_x is symmetric and positive definite).

2. The metric is called ***semi-Riemannian*** in all other cases (except $(-\cdots-)$, which by a trivial change of sign in g may be turned into the Riemannian case).

3. The metric is called ***Lorentzian*** if $\dim(M) = 4$ and the signature is $(-+++)$. Hence

$$g_x = \eta = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.1}$$

   with respect to a basis (e_a) of the above kind, which is duly called ***orthonormal***.

The Lorentzian case is the one of interest to GR, but we will often invoke examples from Riemannian geometry in order to explain some contrast with the Lorentzian case. Later on, the 3+1 split of M will be such that we look for Riemannian submanifolds of M (to be defined later).

The simplest example of a ***Lorentzian manifold*** (i.e., a manifold with Lorentzian metric) is $\mathbb{R}^4$ with the standard basis and g_x defined by (2.1) for all x. More precisely, we relabel the usual coordinates of $\mathbb{R}^4$ as (x^0,x^1,x^2,x^3), so that $T_x\mathbb{R}^4 \cong \mathbb{R}^4$ has the canonical basis $(\partial_0,\partial_1,\partial_2,\partial_3)$, with respect to which $g_{00} = g(\partial_0,\partial_0) = -1$, $g_{ii} = g(\partial_i,\partial_i) = 1$ for $i = 1,2,3$, and $g_{\mu\nu} = 0$ whenever $\mu \neq \nu$. Here we have introduced a convention often used in the (physics) literature: Greek indices μ,ν etc. run from 0 to 3, whereas Latin indices i,j etc. run from 1 to 3. Both Greek and Latin indices midway in the alphabet refer to the canonical coordinate basis $\partial_\mu = \partial/\partial x^\mu$ or $\partial_i = \partial/\partial x^i$, whereas indices a,b etc. typically refer to arbitrary bases (e_a). The above example $(\mathbb{R}^4,\eta)$ is called ***Minkowski space-time***, equipped with ***Minkowski metric*** η. It is the basis of Einstein's *special* theory of relativity, of which the *general* theory of relativity is some kind of a generalization. *What* kind exactly remains a source of (largely philosophical) debate: certainly, Einstein did not succeed in making all motion 'relative', as he originally intended.

2.2 Lowering and raising indices

Let (M,g) be a (semi) Riemannian manifold. Since g is nondegenerate, the distinction between vectors and covectors is blurred, because we now have canonical ('musical') isomorphisms

$$\flat_x : T_xM \to T_x^*M,\ \ \flat_x(X) \equiv X^\flat;\ \ X^\flat(Y) = g_x(X,Y); \tag{2.2}$$

$$\sharp_x : T_x^*M \to T_xM,\ \ \sharp_x(\theta) \equiv \theta_\sharp;\ \ g(\theta_\sharp, X) = \theta(X), \tag{2.3}$$

which maps are obviously each other's inverse, and induce mutually inverse maps

$$\flat : \mathfrak{X}(M) \to \Omega(M); \tag{2.4}$$

$$\sharp : \Omega(M) \to \mathfrak{X}(M) \tag{2.5}$$

by pointwise application. This leads to the ***lowering and raising of indices***, which is crucial to almost any computation in GR. At any point x (which we omit) we define (g^{ab}) as the inverse (matrix) to (g_{ab}), where $g_{ab} = g(e_a, e_b)$ in some basis e_a (so that $g^{ab}g_{bc} = \delta^a_c$). Obviously,

$$X^\flat_a = g_{ab}X^b; \tag{2.6}$$

$$\theta^a_\sharp = g^{ab}\theta_b, \tag{2.7}$$

which notation may then be extended to any tensor, *where the 'sharp' and 'flat' signs are usually omitted*. For example, (2.6) - (2.7) are simply written as $X_a = g_{ab}X^b$ and $\theta^a = g^{ab}\theta_b$, and for say the Riemann tensor $R \in \mathfrak{X}^{(3,1)}(M)$ (with abuse of notation) we may define $R \in \mathfrak{X}^{(4,0)}(M)$ by

$$R_{abcd} = g_{ae}R^e_{bcd}. \tag{2.8}$$

The contraction process explained at the end of the previous chapter, which in principle has nothing to do with the metric, may now elegantly be rewritten in terms of the metric by, e.g.,

$$R_{ab} = R^c_{acb} = g^{cd}R_{dacb}, \tag{2.9}$$

end hence may be repeated even in case where the original version doesn't apply, as in

$$R = g^{ab}R_{ab}. \tag{2.10}$$

If $R \in \mathfrak{X}^{(3,1)}(M)$ is the Riemann tensor, so that its first contraction $R \in \mathfrak{X}^{(2,0)}(M)$ is the Ricci tensor, this second contraction yields the ***Ricci scalar***, which again plays a central role in GR.[9] Indeed, as we shall see, GR revolves around the ***Einstein tensor*** $G \in \mathfrak{X}^{(2,0)}(M)$, defined by

$$G_{ab} = R_{ab} - \tfrac{1}{2}g_{ab}R. \tag{2.11}$$

Abstractly, *lowering* an index is a map $\flat : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k+1,l-1)}(M)$ (provided $l > 0$ of course), whose definition depends on the index. Taking the first (upper) index for simplicity, we have

$$T^\flat(X_1,\ldots,X_{k+1};\theta^1,\ldots,\theta^{l-1}) = T(X_2,\ldots,X_{k+1};X_1^\flat,\theta^1,\ldots,\theta^{l-1}). \tag{2.12}$$

Similarly, *raising* an index is a map $\sharp : \mathfrak{X}^{(k,l)}(M) \to \mathfrak{X}^{(k-1,l+1)}(M)$ ($k > 0$), which is defined, for example once again on the first (lower) index, by

$$T_\sharp(X_1,\ldots,X_{k-1};\theta^1,\ldots,\theta^{l+1}) = T(\theta^1_\sharp,X_1,\ldots,X_{k-1};\theta^2,\ldots,\theta^{l+1}). \tag{2.13}$$

[9] Readers who don't like the use of the same symbol for (in this case) four different things may *either* want to introduce different notations for each different object (such as 'Riemann', 'Ricci', and 'R'), which still doesn't solve the notation problem for raising and lowering indices except by reinstalling the 'sharp' and 'flat' symbols each time, *or* use Penrose's ***abstract index notation***, where for example R^a_{bcd} does *not* refer to the components of R in some basis, as in our notation, but simply indicates that $R \in \mathfrak{X}^{(3,1)}$. Indices defining the components of some tensor should then be added, which often leads to typographically horrible expressions (see e.g. Malament (2012).

2.3 Geodesics

Intuitively, geodesics are paths of shortest lengths between two given points. This idea only makes sense in the Riemannian case (as opposed to the semi-Riemannian case), with which we therefore start: we will then find a redefinition of a geodesic that does make sense also on semi-Riemannian manifolds. So, at least initially, let (M,g) be a Riemannian manifold. It will be convenient to use closed intervals $I=[a,b]$ as the domains of curves $\gamma: I\to M$.[10]

1. The ***length*** of a curve $\gamma:[a,b]\to M$ is defined as
$$L(\gamma)=\int_a^b dt\,\sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))}\equiv\int_a^b dt\,\|\dot\gamma(t)\|, \tag{2.14}$$
where $\dot\gamma(t)\in T_{\gamma(t)}M$ is the tangent vector to the curve, i.e. $\dot\gamma(t)f=df(\gamma(t))/dt$, cf. (1.17). So in coordinates one has $\gamma(t)=(\gamma^1(t),\dots,\gamma^n(t))$, where $\gamma^i:[a,b]\to\mathbb{R}$, and hence
$$g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))=g_{ij}(\gamma(t))\frac{d\gamma^i(t)}{dt}\frac{d\gamma^j(t)}{dt}\equiv g_{ij}(\gamma(t))\dot\gamma^i(t)\dot\gamma^j(t). \tag{2.15}$$
The length of γ is independent of its parametrization (i.e. it only depends on the image $\gamma([a,b])$ in M),[11] as opposed to its (kinetic) ***energy***, which is defined as
$$E(\gamma)=\int_a^b dt\,g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))=\int_a^b dt\,\|\dot\gamma(t)\|^2. \tag{2.16}$$
Both functionals extend to piecewise smooth curves, simply by splitting the integrals.

2. If M is connected, any two points can be connected by a smooth curve, and hence we can define the ***distance*** $d(x,y)$ between $x,y\in M$ as the infimum of $L(\gamma)$ over all smooth curves $\gamma:[0,1]\to M$ with $\gamma(0)=x$ and $\gamma(1)=y$ (one may equivalently use piecewise smooth curves here, which can alway be smoothened near their bends). This is a metric on M, whose metric topology coincides with the original topology of M.[12]

3. A ***geodesic*** (between two given points) is a curve of extremal length. We will not precisely explain what this problem in the calculus of variations means, since our goal is merely to derive the alternative (re)definition below that is valid also for the semi-Riemannian case, and so we just explain how this extremal problem is solved. In general, a functional
$$S(\gamma)=\int_a^b dt\,\mathscr{L}(\gamma(t),\dot\gamma(t)) \tag{2.17}$$
is minimized or maximized by some curve γ iff the ***Euler–Lagrange equations*** hold:
$$\frac{d}{dt}\frac{\partial\mathscr{L}}{\partial\dot\gamma^i}-\frac{\partial\mathscr{L}}{\partial\gamma^i}=0. \tag{2.18}$$
Short of giving an introduction to the calculus of variations,[13] here is a heuristic derivation of (2.18). Let $\gamma_s(t)$ a family of curves indexed by s, such that endpoints are fixed, that is,
$$\gamma_s(a)=\gamma(a);\qquad\qquad\gamma_s(b)=\gamma(b). \tag{2.19}$$

[10] Recall our standing assumption that maps, including curves, are smooth.

[11] This is an easy calculation, see e.g. Jost, Lemma 1.4.3.

[12] See e.g. Jost, pp. 14–15.

[13] We will do so later on, when discussing the derivation of the Einstein equations from the Hilbert action. In some sense that is easier, since one can work with Banach spaces. Here, the appropriate space of curves in M does not even have a linear structure and has be treated as an infinite-dimensional manifold modeled on a Banach space.

Then the extremality condition

$$\frac{dS(\gamma_s)}{ds} = 0 \tag{2.20}$$

gives, on repeatedly using the chain rule and a partial integration,

$$\begin{aligned}\frac{dS(\gamma_s)}{ds} &= \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i} \frac{\partial \gamma_s^i}{\partial s} + \frac{\partial \mathscr{L}}{\partial \dot{\gamma}_s^i} \frac{\partial \dot{\gamma}_s^i}{\partial s} \right) = \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i} \frac{\partial \gamma_s^i}{\partial s} + \frac{\partial \mathscr{L}}{\partial \dot{\gamma}_s^i} \frac{\partial}{\partial t} \frac{\partial \gamma_s^i}{\partial s} \right) \\ &= \int_a^b dt \left(\frac{\partial \mathscr{L}}{\partial \gamma_s^i} - \frac{\partial}{\partial t} \frac{\partial \mathscr{L}}{\partial \dot{\gamma}_s^i} \right) \frac{\partial \gamma_s^i}{\partial s} + \left|_b^a \frac{\partial \mathscr{L}}{\partial \dot{\gamma}_s^i} \frac{d \gamma_s^i}{ds} \right.. \end{aligned} \tag{2.21}$$

Then (2.19) gives $d\gamma_s(a)/ds = d\gamma_s(b)/ds = 0$, so that, for arbitrary γ_s and hence $\partial\gamma_s/\partial s$, eq. (2.20) implies (2.18), in which s is dropped and hence $\partial/\partial t$ becomes d/dt.

For the energy functional (2.16) the Euler–Lagrange equations (2.18) are

$$\ddot{\gamma}^i(t) + \Gamma^i_{jk}(\gamma(t)) \dot{\gamma}^j(t) \dot{\gamma}^k(t) = 0, \tag{2.22}$$

or briefly $\ddot{\gamma}^i + \Gamma^i_{jk} \dot{\gamma}^j \dot{\gamma}^k = 0$, where $\ddot{\gamma} = d^2\gamma/dt^2$, and the ***Christoffel symbols*** are given by

$$\Gamma^i_{jk} = \tfrac{1}{2} g^{im} (g_{mj,k} + g_{mk,j} - g_{jk,m}), \tag{2.23}$$

where we have introduced another useful notational convention from GR:

$$\tau^{j_1 \cdots j_l}_{i_1 \cdots i_k, j} = \partial_j \tau^{j_1 \cdots j_l}_{i_1 \cdots i_k}. \tag{2.24}$$

Warning: the Christoffel symbols do *not* form the components of a would-be tensor "$\Gamma \in \mathfrak{X}^{(2,1)}(M)$": physicists see this from their incorrect behaviour under coordinate transformations, whereas mathematicians note that Γ fails the 'tensoriality test' stated before (1.49). We will see, however, that the Γ-*symbols* do combine into the Riemann *tensor*!

To derive (2.22) for (2.16), i.e., for $\mathscr{L}(\gamma(t), \dot{\gamma}(t)) = g_{ij}(\gamma(t)) \dot{\gamma}^i(t) \dot{\gamma}^j(t)$, one uses

$$\frac{\partial \mathscr{L}}{\partial \gamma^i} = g_{jk,i} \dot{\gamma}^j \dot{\gamma}^k; \tag{2.25}$$

$$\frac{d}{dt} \frac{\partial \mathscr{L}}{\partial \dot{\gamma}^i} = 2 \frac{d}{dt} g_{ij} \dot{\gamma}^j = 2 (g_{ij,k} \dot{\gamma}^k \dot{\gamma}^j + g_{ij} \ddot{\gamma}^j) = (g_{ij,k} + g_{ik,j}) \dot{\gamma}^k \dot{\gamma}^j + 2 g_{ij} \ddot{\gamma}^j. \tag{2.26}$$

Whereas solutions of (2.22) extremize the energy for *any* parametrization), for the length functional (2.14), the Euler–Lagrange equations only take the form (2.22) iff $\|\dot{\gamma}(t)\|$ is constant; in paticular, if $\|\dot{\gamma}(t)\| = 1$ for all $t \in I$ we say that γ is parametrized by ***arc length***.[14] *We define a geodesic as a curve γ that satisfies* (2.22). This in turn *implies* that $\|\dot{\gamma}(t)\|$ is constant, as can be shown by simply computing $d(\|\dot{\gamma}(t)\|^2)/dt$ from (2.15). This time-derivative equals $g_{ij,k} \dot{\gamma}^i \dot{\gamma}^j \dot{\gamma}^k + 2 g_{ij} \ddot{\gamma}^i \dot{\gamma}^j$. Eliminating $\ddot{\gamma}^i$ via (2.22) then leads to complete cancellation to zero (for a neater calculation see footnote 22). The definition of a geodesic therefore depends on the parametrization of γ: a reparametrized geodesic may no longer satisfy (2.22), except when the reparametrization is *affine*, i.e. $t' = at + b$.[15]

[14] One has $L(\gamma)^2 \leq 2(b-a)E(\gamma)$ for any $\gamma : [a,b] \to M$, with equality iff $\|\dot{\gamma}(t)\|$ is constant (Jost, Lemma 1.4.2).

[15] Nonetheless, it is easy to show that some curve γ can be reparametrized so as to *become* a geodesic iff the right-hand side of (2.22) equals $f \cdot \dot{\gamma}^i$ for some $f \in C^\infty(M)$; cf. Malament, Prop. 1.7.9.

We started with the intuitive idea of geodesics as shortest paths between given endpoints. We now have to add some nuances. First, in $M = \mathbb{R}^n$ with flat metric (i.e. $g_{ij} = \delta_{ij}$ in the usual coordinates) geodesics are straight lines and indeed *always* form *shortest* paths between two given points.[16] But this is exceptional. For example, on the sphere, geodesics are great circles, and hence one has two geodesics between two generic points: one of minimal length and one of maximal length. These lengths coincide iff the two points are polar opposites, in which case one has infinitely many geodesics between them.

Second, in the intuitive idea of geodesics the focus was on the endpoints, whereas in defining geodesics as solutions to the ODE (2.22), the focus is rather on the initial point $\gamma(0)$ and the initial velocity $\dot{\gamma}(0)$; indeed, the solution to (2.22) is uniquely defined by these data, expect for the interval I. Like any solution to an ODE, γ has some maximal domain $I \subseteq \mathbb{R}$ on which it is defined, and this domain may not equal $\mathbb{R}$. If all geodesics γ with given $\gamma(0)$ and $\dot{\gamma}(0)$ can be defined on $I = \mathbb{R}$ we say that (M, g) is ***geodesically complete***. The ***Hopf-Rinow Theorem*** states that a Riemannian manifold (M, g) is geodesically complete iff it is complete in the metric d derived from g.[17] In particular, any compact Riemannian manifold is complete. Trivial examples of incomplete Riemannian manifolds are provided by open bounded sets $\Omega \subset \mathbb{R}^n$ with flat metric inherited from $\mathbb{R}^n$. Many Lorentzian manifolds of interest to GR (undoubtedly including our universe) are geodesically *in*complete, the proof of which (by means of the famous *singularity theorems* of Hawking and Penrose from the 1960s) forms one of the highlights of GR.

2.4 Linear connections

The definition of a geodesic as a curve γ whose tangent vector $\dot{\gamma}$ satisfies (2.22) along the entire curve (i.e. for each t where $\gamma(t)$ is defined) was inspired by the Riemannian case, but it clearly makes sense for semi-Riemannian manifolds, too. We now move on to give a geometric perspective on the Christoffel symbols Γ^i_{jk} and hence on the curious ***geodesic equation*** (2.22).

1. A ***linear connection*** on M (which is the same thing as a connection on the tangent bundle TM, see below), or, equivalently, a ***covariant derivative*** on $\mathfrak{X}(M)$, associates to each vector field $X \in \mathfrak{X}(M)$ a linear map $\nabla_X : \mathfrak{X}(M) \to \mathfrak{X}(M)$, such that:

   (a) The map $X \mapsto \nabla_X$ is $\mathbb{R}$-linear as well as $C^\infty(M)$-linear, i.e.

   $$\nabla_{fX} Y = f \nabla_X Y \quad (f \in C^\infty(M)); \tag{2.27}$$

   (b) The map $Y \mapsto \nabla_X Y$ is $\mathbb{R}$-linear but not $C^\infty(M)$-linear: it satisfies the ***Leibniz rule***

   $$\nabla_X (fY) = (Xf)Y + f \nabla_X Y \quad (f \in C^\infty(M)). \tag{2.28}$$

   This definition also makes sense on any open $U \in \mathcal{O}(M)$, and in fact if $x \in U$, then $\nabla_X Y(x)$ only depends on the value of X at x and the restriction of Y to U (this follows from (2.27) - (2.28) and the definition of a manifold). Hence we may compute covariant derivatives locally: take a (local) frame (e_a) for $\mathfrak{X}(M)$ (recall that this consists of n maps $e_a : U \to TM$ such that at each $x \in U$ the vectors $e_a(x) \in T_x M$ form a basis of $T_x M$, $a = 1, \ldots, n$), with

[16] According to Newtonian mechanics, in the absence of forces particles move on geodesics (a property that will reappear through the back door in GR).

[17] See e.g. Jost, Theorem 1.4.8.

dual basis (ω^a) (i.e. the $\omega^a(x) \in T_x^*M$ satisfy $\omega^a(e_b) = \delta_b^a$). The given connection ∇ is then completely characterized by its ***connection coefficients*** ω_{ab}^c, defined (at each x) by

$$\nabla_{e_a} e_b = \omega_{ab}^c e_c. \tag{2.29}$$

Indeed, from (1.44) - (1.45) we may write $X = X^a e_a$, where $X^a = \omega^a(X) \in C^\infty(U)$, so

$$\begin{aligned} \nabla_X Y &= \nabla_{X^a e_a}(Y^b e_b) = X^a \nabla_{e_a}(Y^b e_b) = X^a (e_a(Y^b) \cdot e_b + Y^b \nabla_{e_a} e_b) \\ &= X^a (e_a(Y^c) + Y^b \omega_{ab}^c) e_c. \end{aligned} \tag{2.30}$$

We write $\nabla_X Y^a$ for $(\nabla_X Y)^a$, so that $\nabla_X Y = (\nabla_X Y^a) e_a$. We therefore have

$$\nabla_X Y^a = XY^a + \omega_{bc}^a X^b Y^c, \tag{2.31}$$

where XY^a is the (defining) action of the vector field X on the function $Y^a \in C^\infty(U)$. In terms of the canonical coordinate basis $(e_\mu = \partial_\mu)$, $(\omega^\nu = dx^\nu)$, we therefore have

$$\omega_{\mu\nu}^\rho = dx^\rho(\nabla_\mu \partial_\nu); \tag{2.32}$$

$$\nabla_X Y^\rho = X^\mu (\partial_\mu Y^\rho + \omega_{\mu\nu}^\rho Y^\nu); \tag{2.33}$$

$$\nabla_\mu Y^\rho = \partial_\mu Y^\rho + \omega_{\mu\nu}^\rho Y^\nu, \tag{2.34}$$

where $\nabla_\mu = \nabla_{\partial_\mu}$; we could have written $Y^\rho_{,\mu}$ for $\partial_\mu Y^\rho$, and even the ***semicolon notation*** $Y^\rho_{;\mu}$ for $\nabla_\mu Y^\rho$ is *en vogue* among physicists (we give the general form later on).

2. Linear connections formalize Levi-Civita's notion of ***parallel transport***.[18] It follows from (2.31) or (2.33) that $\nabla_X Y$ only depends on the values of Y along the flow lines of X, for

$$\nabla_X Y^a(x) = \frac{d}{dt} Y^a(\psi_t(x))_{|t=0} + \omega_{bc}^a(x) X^b(x) Y^c(x), \tag{2.35}$$

where ψ is the flow of X. Conversely, given some curve $\gamma : I \to M$ with tangent vectors $\dot\gamma$ *defined along* γ *only*, the covariant derivative $\nabla_{\dot\gamma} Y$ of Y along γ is well defined for any vector field Y defined near $\gamma(I)$ or even on $\gamma(I)$ alone,[19] for in (local) coordinates we have

$$\begin{aligned} \nabla_{\dot\gamma} Y^\rho_{\gamma(t)} &= \dot\gamma^\mu(t)(\partial_\mu Y^\rho_{\gamma(t)} + \omega_{\mu\nu}^\rho(\gamma(t)) Y^\nu_{\gamma(t))} \\ &= \frac{d}{dt} Y^\rho_{\gamma(t)} + \omega_{\mu\nu}^\rho(\gamma(t)) \frac{d\gamma^\mu(t)}{dt} Y^\nu_{\gamma(t)}, \end{aligned} \tag{2.36}$$

where $\gamma^\mu : I \to \mathbb{R}$ are the coordinates of the curve (in some given chart), as before. We then say that some vector $Y \in T_x M$ is *parallel-transported along* γ (with $\gamma(0) = x$) by a vector field $t \mapsto Y_{\gamma(t)}$ defined along γ (i.e. $Y_{\gamma(t)} \in T_{\gamma(t)} M$) if $Y_{\gamma(t)}$ satisfies

$$\nabla_{\dot\gamma} Y = 0. \tag{2.37}$$

This generalizes the idea of freely moving vectors in $\mathbb{R}^n$ from place to place (which one does without any thought) to arbitrary (semi) Riemannian manifolds; the price one pays is that such motions can only be carried out once a linear connection has been defined. Of course, the ***flat connection*** on $\mathbb{R}^n$ (with flat metric $g = \delta$), defined in the standard coordinates by $\omega_{\mu\nu}^\rho = 0$ and hence $\nabla_\mu = \partial_\mu$, reproduces the naive and original case.

[18] See G. Iurato, On the history of Levi-Civitas parallel transport, `arXiv:1608.04986`.

[19] Abstractly, one equips the pullback $\gamma^* TM = \{(X,t) \in TM \times I \mid \pi(X) = \gamma(t)\}$ of $\pi : TM \to M$ with respect to $\gamma : I \to M$, seen as a vector bundle over I, with a connection $\gamma^*\nabla$, defined by $(\gamma^*\nabla)_X Y = \nabla_{f'X} Y$. See Jost, §4.1.

3. Like the Christoffel symbols, the connection coefficients do not form the components of a tensor (the relation between the two will shortly be clarified). However, various tensors may be defined in terms of the connection. For now, we just define the ***torsion*** $\tau_\nabla \in \mathfrak{X}^{(2,1)}(M)$ of a given linear connection ∇ by

$$\tau_\nabla(X,Y,\theta) = \theta(\nabla_X Y - \nabla_Y X - [X,Y]); \tag{2.38}$$

a simple computation shows that this expression is $C^\infty(M)$-linear in each entry, so our 'tensoriality test' shows τ is indeed a tensor of the said kind. The commutator vanishes in the coordinate basis (∂_μ), so that

$$\tau^\rho_{\mu\nu} = \omega^\rho_{\mu\nu} - \omega^\rho_{\nu\mu}, \tag{2.39}$$

and hence the connection ∇ is ***torsion-free*** iff $\omega^\rho_{\mu\nu} = \omega^\rho_{\nu\mu}$, i.e., iff $\nabla_\mu \partial_\nu = \nabla_\nu \partial_\mu$.

4. We now define a ***geodesic*** *with respect to a linear connection* ∇ as a curve γ for which

$$\nabla_{\dot\gamma}\dot\gamma = 0, \tag{2.40}$$

i.e., the tangent vector $\dot\gamma$ to γ is parallel transported along γ.[20] Using (local) coordinates, (2.40) may be brought into a form that is strikingly similar to (2.22): since according to (2.36) with $Y \rightsquigarrow \dot\gamma$ the expression $\dot\gamma^\mu \partial_\mu \dot\gamma^\rho$ is just $d^2\dot\gamma^\rho/dt^2 \equiv \ddot\gamma^\rho$, we obtain

$$\ddot\gamma^\rho + \omega^\rho_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu = 0. \tag{2.41}$$

Thus it is obvious that geodesics are insensitive to the torsion (2.39) of the connection.

Eq. (2.41) looks like the geodesic equation (2.22), with the difference that in (2.41) the coefficients $\omega^\rho_{\mu\nu}$ are defined by (2.32) in terms of an arbitrary linear connection ∇, whereas those in (2.22) are the Christoffel symbols (2.23) defined by the metric. Their relationship is:

Theorem 1 (Levi-Civita) *Any (semi) Riemannian manifold (M,g) admits a unique linear connection ∇ (called the* **Levi-Civita connection***) that satisfies:*

1. *The torsion τ_∇ associated to ∇ vanishes;*
2. *The connection ∇ and the metric g are related by the following property:*

$$X(g(Y,Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z) \;\; (X,Y,Z \in \mathfrak{X}(M)). \tag{2.42}$$

This means that the connection coefficients of ∇ are the Christoffel symbols, i.e.,

$$\omega^\rho_{\mu\nu} = \Gamma^\rho_{\mu\nu}. \tag{2.43}$$

Proof: using torsion-freeness in the form $\nabla_X Y - \nabla_Y X = [X,Y]$, eq. (2.42) may be rewritten as

$$g(\nabla_X Y, Z) = \tfrac{1}{2}(Xg(Y,Z) + Yg(Z,X) - Zg(X,Y) - g(X,[Y,Z]) + g([X,Y],Z) + g(Y,[Z,X])), \tag{2.44}$$

which shows both existence and uniqueness of ∇. In a coordinate basis, where once again all commutators vanish, eq. (2.44) immediately gives (2.43) with (2.23).

[20] As before, this definition depends on the parametrization of γ, which ambiguity may be resolved similarly.

2.5 General connections on vector bundles

For a more general understanding of the above constructions, as well as for a clean extension of linear connections from vector fields to arbitrary tensors (which one often needs in GR), it is useful to briefly discuss connections on arbitrary vector bundles.

1. A ***connection on a vector bundle*** $E \to M$ associates to each vector field $X \in \mathfrak{X}(M)$ a linear map $\nabla_X : \Gamma(E) \to \Gamma(E)$, such that:

   (a) The map $X \mapsto \nabla_X$ is $\mathbb{R}$-linear as well as $C^\infty(M)$-linear in X, cf. (2.27);

   (b) The map $s \mapsto \nabla_X s$ is $\mathbb{R}$-linear but not $C^\infty(M)$-linear: it satisfies the ***Leibniz rule***

   $$\nabla_X(fs) = (Xf)s + f\nabla_X s \ \ (f \in C^\infty(M)). \tag{2.45}$$

   A linear connection is just a connection on the tangent bundle; the general story is almost the same, including the localization of $\nabla_X s(x)$ to the flow lines of X arbitrarily close to x, and hence to any $U \in \mathscr{O}(M)$, $x \in U$. In particular, define a local frame (u_a), where $\mathsf{a} = 1, \ldots, k = \dim(E_x)$, i.e. the rank of E) by the properties that $u_\mathsf{a} \in \Gamma(U, E)$ (i.e. the restriction of $\Gamma(E) \equiv \Gamma(M, E)$ to some $U \in \mathscr{O}(M)$) and $(u_\mathsf{a}(x))$ forms a basis of E_x for all $x \in U$. This once again yields ***connection coefficients*** defined by

   $$\nabla_\mu u_\mathsf{b} = C^\mathsf{c}_{\mu\mathsf{b}} u_\mathsf{c}; \tag{2.46}$$

   the difference with the tangent bundle is that the three indices carried by C are no longer of the same type: b and c label basis vectors in E_x, whereas μ refers to the canonical coordinate base of $T_x M$ (recall that $\nabla_\mu = \nabla_{\partial_\mu}$). Writing $s(x) = s^\mathsf{a}(x) u_\mathsf{a}(x)$, we now have

   $$\nabla_\mu s^\mathsf{a} = \partial_\mu s^\mathsf{a} + C^\mathsf{a}_{\mu\mathsf{b}} s^\mathsf{b}, \tag{2.47}$$

   cf. (2.34). This is often written as

   $$\nabla_\mu s = \partial_\mu s + \omega_\mu s, \tag{2.48}$$

   in which either s is seen as a vector with components s^a relative to the given basis (u_a) and hence ω_μ is a matrix with components $C^\mathsf{a}_{\mu\mathsf{b}}$, or $s \in \Gamma(E)$ and $\omega_\mu(x) \in \mathrm{Hom}(E_x, E_x)$.[21]

   A vector bundle E may be equipped with a ***metric***, i.e. nondegenerate symmetric bilinear form $g_x : E_x \times E_x \to \mathbb{R}$ defined for each $x \in M$, smooth in x in the sense that for any $s, t \in \Gamma(E)$ the function $g(s,t) : M \to \mathbb{R}$ defined by $x \mapsto g_x(s(x), t(x))$ is smooth. For example, a (semi) Riemannian metric on M is a metric on $E = TM$ in precisely this sense. A connection ∇ on E is then called ***metric*** if for all $s, t \in \Gamma(E)$ we have

   $$X(g(s,t)) = g(\nabla_X s, t) + g(s, \nabla_X t). \tag{2.49}$$

   For example, the Levi-Civita connection on TM is obviously metric in this sense.

[21] Even more abstractly, connections may be regarded as maps $\nabla : \Gamma(E) \to \Gamma(T^*M \otimes E) \equiv \Omega^1(E)$, i.e. the space of E-valued 1-forms, that satisfy $\nabla(fs) = df \otimes s + f\nabla s$; the connection with the main text is $\nabla_X s = \nabla s(X)$. In that case we may write $\nabla = d + \omega$, where $\omega \in \Omega^1(\mathrm{Hom}(E,E))$, i.e. ω is a 1-form taking values in the vector bundle $\mathrm{Hom}(E,E)$. Even more generally (for those familiar with the de Rham complex $\Omega^\bullet(M)$ and its relative $\Omega^\bullet(E)$), we may define $\nabla : \Omega^p(E) \to \Omega^{p+1}(E)$, where $p = 0, \ldots, k$ with $\Omega^0(E) \equiv \Gamma(E)$, as the unique extension of the above map $\nabla : \Omega^0(E) \to \Omega^1(E)$ that satisfies $\nabla(\alpha \otimes s) = d\alpha \otimes s + (-1)^p \alpha \wedge \nabla s$, where $\alpha \in \Omega^p(M)$ and $s \in \Gamma(E)$.

2. Furthermore, take $E = T^*M$, and define ∇^* in coordinates through its components by

$$\nabla^*_\mu \theta_\nu = \partial_\mu \theta_\nu - \Gamma^\rho_{\mu\nu}\theta_\rho, \tag{2.50}$$

where the $\Gamma^\rho_{\mu\nu}$ are the Christoffel symbols defined by some (semi) Riemannian metric on M, cf. (2.23). This turns out to be a connection indeed (check the axioms), whose rationale (notably of the minus sign!) is the Leibniz-type property (or product rule)

$$X(\theta(Y)) = (\nabla^*_X\theta)(Y) + \theta(\nabla_X Y), \tag{2.51}$$

which may look even more elegant in the form

$$\nabla_X\langle\theta, Y\rangle = \langle\nabla^*_X\theta, Y\rangle + \langle\theta, \nabla_X Y\rangle, \tag{2.52}$$

where by *fiat* we have declared that on functions (such as $\langle\theta, Y\rangle \equiv \theta(Y)$) the covariant derivative ∇_X is simply X, i.e. $\nabla_X f \equiv Xf$, $f \in C^\infty(M)$. Eq. (2.51) or (2.52) might, of course, have been used to define $\nabla^* : \Omega(M) \to \Omega(M)$ in the first place, yielding (2.50). In fact, any linear connection defines a dual connection ∇^* on T^*M by (2.51).

3. Combining (2.34) and (2.50), we define a covariant derivative $\nabla^{(k,l)} : \mathfrak{X}^{(k,l)} \to \mathfrak{X}^{(k,l)}$ by

$$\begin{aligned}\nabla^{(k,l)}_\mu \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} = \partial_\mu \tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} + \Gamma^{\rho_1}_{\mu\sigma}\tau^{\sigma\cdots\rho_l}_{\nu_1\cdots\nu_k} + \cdots + \Gamma^{\rho_l}_{\mu\sigma}\tau^{\rho_1\cdots\sigma}_{\nu_1\cdots\nu_k} \\ - \Gamma^\sigma_{\mu\nu_1}\tau^{\rho_1\cdots\rho_l}_{\sigma\cdots\nu_k} - \cdots - \Gamma^\sigma_{\mu\nu_k}\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\sigma},\end{aligned} \tag{2.53}$$

where the left-hand side, which may also be written $\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k;\mu}$, really means $(\nabla^{(k,l)}_\mu\tau)^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k}$.

For those who don't like such 'definitions by formula', we note that $\nabla^{(k,l)}$ is the unique connection on $T^{(k,l)}M$ that, similarly to (2.52), satisfies the Leibniz rule

$$\begin{aligned}X(\tau(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l)) = (\nabla^{(k,l)}_X\tau)(X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) \\ + \tau(\nabla_X X_1,\ldots,X_k,\theta^1,\ldots,\theta^l) + \cdots \\ + \tau(X_1,\ldots,X_k,\theta^1,\ldots,\nabla^*_X\theta^l),\end{aligned} \tag{2.54}$$

where the case $k = l = 0$ is taken to mean $\nabla^{(0,0)}_X = X$ on $\mathfrak{X}^{(0,0)}(M) = C^\infty(M)$. Eq. (2.54) recovers $\nabla^{(0,1)} = \nabla$ on $\mathfrak{X}^{(0,1)}(M) = \mathfrak{X}(M)$ as well as $\nabla^{(1,0)} = \nabla^*$ on $\mathfrak{X}^{(1,0)}(M) = \Omega(M)$.

This construction of $\nabla^{(k,l)}$ works for any linear connection ∇, but if the latter is the Levi-Civita connection, then (2.54) implies that its defining property (2.42) comes down to[22]

$$\nabla^{(2,0)}g \equiv \nabla g = 0; \tag{2.55}$$

also in general, one usually writes ∇ for any $\nabla^{(k,l)}$. Physicists write (2.55) as

$$g_{\mu\nu;\sigma} = 0. \tag{2.56}$$

Alternatively, one may recall the description of $T^{(k,l)}M$ as the (vector bundle) tensor product of k copies of T^*M and l copies of TM, and introduce the ***tensor product connection***.

[22] As an application, let us show once again that $d(\|\dot\gamma(t)\|)/dt = 0$ for geodesics γ: using (2.54), (2.55), and (2.40), we obtain $d(\|\dot\gamma(t)\|^2)/dt = dg(\dot\gamma,\dot\gamma)/dt = \dot\gamma(g(\dot\gamma,\dot\gamma)) = (\nabla_{\dot\gamma}g)(\dot\gamma,\dot\gamma) + g(\nabla_{\dot\gamma}\dot\gamma,\dot\gamma) + g(\dot\gamma,\nabla_{\dot\gamma}\dot\gamma) = 0+0+0 = 0$.

Given two vector bundles $E^{(1)} \to M$ and $E^{(2)} \to M$, with connections $\nabla^{(1)}$ and $\nabla^{(2)}$, there is a unique connection $\nabla^{(1\otimes 2)}$ on $E^{(1)} \otimes E^{(2)}$ that satisfies the product rule[23]

$$\nabla^{(1\otimes 2)}(s^{(1)} \otimes s^{(2)}) = \nabla^{(1)}(s^{(1)}) \otimes s^{(2)} + s^{(1)} \otimes \nabla^{(2)}(s^{(2)}). \tag{2.57}$$

This may be iterated to the tensor product of finitely many vector bundles, and hence (for any linear connection ∇) the connection $\nabla^{(k,l)}$ defined by (2.53) or (2.54) is just the tensor product of the individual connections on each copy of TM or T^*M present in $T^{(k,l)}M$.

It follows from (2.51) that (for any ∇) the connection $\nabla^{(k,l)}$ *commutes with contraction*. Contracting the first upper and lower indices and writing $\sigma^{\rho_2\cdots\rho_l}_{\nu_2\cdots\nu_k} = \tau^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k}$, one has

$$(\nabla^{(k,l)}_{\nu_1}\tau)^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k} = (\nabla^{(k,l)}_{\nu_1}\sigma)^{\rho_2\cdots\rho_l}_{\nu_2\cdots\nu_k}, \tag{2.58}$$

and similarly for any other pair of upper and lower indices. In particular, this makes the physicists' notation $\tau^{\nu_1\rho_2\cdots\rho_l}_{\nu_1\nu_2\cdots\nu_k;\mu}$ unambiguous. For example, for the Ricci tensor we have

$$R_{\mu\nu;\sigma} = R^{\rho}_{\mu\rho\nu;\sigma}. \tag{2.59}$$

If ∇ satisfies (2.55), then $\nabla^{(k,l)}$ in addition commutes with contraction in the metric sense explained before (2.10), so that e.g., using (2.56), for the Ricci scalar we have

$$R_{,\sigma} = R_{;\sigma} = (g^{\mu\nu}R_{\mu\nu})_{;\sigma} = g^{\mu\nu}{}_{;\sigma}R_{\mu\nu} + g^{\mu\nu}R_{\mu\nu;\sigma} = g^{\mu\nu}R_{\mu\nu;\sigma}. \tag{2.60}$$

Finally, $\nabla^{(k,l)}$ may be used to rewrite the formula (1.58) for the Lie derivative as

$$\begin{aligned}\mathscr{L}_X\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} &= \nabla_X\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu_k} + (\nabla_{\nu_1}X^{\nu})\tau^{\rho_1\cdots\rho_l}_{\nu\cdots\nu_k} + \cdots + (\nabla_{\nu_n}X^{\nu})\tau^{\rho_1\cdots\rho_l}_{\nu_1\cdots\nu} \\ &\quad - (\nabla_{\rho}X^{\rho_1})\tau^{\rho\cdots\rho_l}_{\nu_1\cdots\nu_k} - \cdots - (\nabla_{\rho}X^{\rho_l})\tau^{\rho_1\cdots\rho}_{\nu_1\cdots\nu_k},\end{aligned} \tag{2.61}$$

since all Christoffel symbols cancel out (check!).[24] For example, using (2.55) we obtain

$$\mathscr{L}_X g_{\mu\nu} = (\nabla_\mu X^\rho)g_{\rho\nu} + (\nabla_\nu X^\rho)g_{\mu\rho} = X_{\nu;\mu} + X_{\mu;\nu}. \tag{2.62}$$

A vector field X for which $\mathscr{L}_X g = 0$, and hence $X_{\nu;\mu} + X_{\mu;\nu} = 0$, is called a ***Killing field***.[25] Flows of Killing fields are ***isometries***, that is, diffeomorphisms preserving the metric. In the notation of (1.53), this means that $\psi_t^{(2,0)} g = g$, which is usually written as $\psi_t^* g = g$. Since $[\mathscr{L}_X, \mathscr{L}_Y] = \mathscr{L}_{[X,Y]}$ Killing fields always form a Lie algebra, whose associated Lie group (up to global issues) is the subgroup of $\mathrm{Diff}(M)$ consisting of isometries.

In Minkowski space-time $(\mathbb{R}^4, \eta)$ the Christoffels symbols vanish (at least in the usual coordinates), so that $\nabla_\mu = \partial_\mu$. Hence Killing fields satisfy the equation

$$X_{\nu,\mu} + X_{\mu,\nu} = 0. \tag{2.63}$$

The general solution is a 10-dimensional vector space (within $\mathfrak{X}(\mathbb{R}^4)$) with basis

$$X^{\mu}_{(\nu)}(x) = \delta^{\mu}_{\nu} \ \ (\nu = 0,1,2,3); \tag{2.64}$$

$$X^{\mu}_{(\rho\sigma)}(x) = x_\rho\delta^{\mu}_{\sigma} - x_\sigma\delta^{\mu}_{\rho}, \ \ (\rho,\sigma = 0,1,2,3), \tag{2.65}$$

or $X_{(\nu)} = \partial_\nu$ and $X_{(\rho\sigma)} = x_\rho\partial_\sigma - x_\sigma\partial_\rho$, where $x_\rho = \eta_{\rho\sigma}x^\sigma$. This is the Lie algebra of the Poincaré-group (which is the subgroup of $GL_4(\mathbb{R})$ preserving the Minkowski metric η).

[23] If we realize $V \otimes W$ as $\mathrm{Hom}(V^* \times W^*)$, i.e. the vector space of bilinear maps from $V^* \times W^*$ to $\mathbb{R}$, then, for $v \in V$ and $w \in W$, the element $v \otimes w \in V \otimes W$ is defined by $v \otimes w(\alpha,\beta) = \alpha(v)\beta(w)$, where $\alpha \in V^*$ and $\beta \in W^*$.

[24] $\mathscr{L}_X$ is not a connection (as it fails to be $C^\infty(M)$-linear in X), but $\mathscr{L}_X$ and ∇_X both satisfy the Leibniz rule.

[25] Named after the German mathematician Wilhelm Killing (1847–1923), not the movie about Cambodja.

3 Curvature

The notion of curvature was originally introduces by Gauß in the context of lines in $\mathbb{R}^2$ and $\mathbb{R}^3$ and surfaces in $\mathbb{R}^3$. The modern approach via connections is highly abstract (and hence very powerful), but we shall recover at least some of the original ideas of Gauß c.s. later on.

3.1 Curvature tensor

For any connection ∇ on a vector bundle $E \to M$, the following map, indexed by $X, Y \in \mathfrak{X}(M)$,

$$\Omega(X,Y) : \Gamma(E) \to \Gamma(E); \tag{3.1}$$

$$\Omega(X,Y) = [\nabla_X, \nabla_Y] - \nabla_{[X,Y]} \tag{3.2}$$

is easily checked to be $C^\infty(M)$-linear in its argument $s \in \Gamma(E)$, so that $\Omega(X,Y)$ defines a cross-section of $\Gamma(\mathrm{Hom}(E,E))$.[26] In addition, $\Omega(X,Y)$ is $C^\infty(M)$-linear in X and Y, so that in the usual basis (∂_μ) associated to a chart defining coordinates (x^μ) we may write

$$[\nabla_\mu, \nabla_\nu] s(x) = \Omega_{\mu\nu}(x) s(x), \tag{3.3}$$

where $\Omega_{\mu\nu} = \Omega(\partial_\mu, \partial_\nu)$ is a linear map $E_x \to E_x$. Relative to a local frame (u_a) for $\Gamma(E)$ in which $s(x) = s^\mathsf{a}(x) u_\mathsf{a}(x)$, with $s^\mathsf{a} \in C^\infty(U)$ (see text after (2.45)), we may therefore write

$$[\nabla_\mu, \nabla_\nu] s^\mathsf{a}(x) = \Omega^\mathsf{a}_{\mathsf{b}\mu\nu}(x) s^\mathsf{b}(x), \tag{3.4}$$

from which it should be clear that the ***curvature tensor*** Ω has four indices: the first two (i.e. a and b) refer to a basis of E_x, whereas the last two (viz. μ and ν) refer to a basis of $T_x M$.

3.2 Riemann tensor

In the case $E = TM$ we now turn to this distinction is blurred, but even there it is good to keep it in mind. So we now take the Levi-Civita connection ∇ on TM, and hence have

$$\Omega(X,Y) : \mathfrak{X}(M) \to \mathfrak{X}(M); \tag{3.5}$$

$$\Omega(X,Y)Z = ([\nabla_X, \nabla_Y] - \nabla_{[X,Y]})Z, \tag{3.6}$$

where $X, Y, Z \in \mathfrak{X}(M)$, and (3.6) is $C^\infty(M)$-linear in each of the three separately. Hence

$$R(\theta, Z, X, Y) = \theta(\Omega(X,Y)Z) \tag{3.7}$$

defines a tensor $R \in \mathfrak{X}^{(3,1)}(M)$ called the ***Riemann tensor***.[27] Or, if we lower the first index,

$$R(W,Z,X,Y) = g(W, (\Omega(X,Y)Z)). \tag{3.8}$$

Its components are

$$R^\rho_{\sigma\mu\nu} = R(\omega^\rho, \partial_\sigma, \partial_\mu, \partial_\nu), \tag{3.9}$$

[26] See footnote 8. This argument is not necessary for what follows, but it does give additional insight.

[27] Bernhard Georg Friedrich Riemann (1826–1866) was one of the greatest and most influential mathematicians in recent history. His *Habilitationsschrift* from 1854 entitled *Über die Hypothesen, welche der Geometrie zu Grunde liegen* is a blueprint for modern differential geometry, especially from a metric point of view. You can find it for example on `https://www.maths.tcd.ie/pub/HistMath/People/Riemann/Geom/Geom.pdf`. Riemann also anticipated applications to physics, though not in the specific way Einstein eventually used his work.

and similarly to (3.4) we equivalently have, for any vector $Z = Z^\rho \partial_\rho$,

$$[\nabla_\mu, \nabla_\nu] Z^\rho = R^\rho_{\sigma\mu\nu} Z^\sigma. \tag{3.10}$$

Either way, using (2.34) one easily obtains the expression

$$R^\rho_{\sigma\mu\nu} = \Gamma^\rho_{\sigma\nu,\mu} - \Gamma^\rho_{\sigma\mu,\nu} + \Gamma^\rho_{\mu\tau}\Gamma^\tau_{\nu\sigma} - \Gamma^\rho_{\nu\tau}\Gamma^\tau_{\mu\sigma}, \tag{3.11}$$

where the Christoffel symbols are defined by (2.23), i.e.,

$$\Gamma^\rho_{\mu\nu} = \tfrac{1}{2} g^{\rho\sigma}(g_{\sigma\mu,\nu} + g_{\sigma\nu,\mu} - g_{\mu\nu,\sigma}). \tag{3.12}$$

Regarding R and Γ as matrices and hence omitting their first two indices, (3.11) reads[28]

$$\Omega_{\mu\nu} = \partial_\mu \Gamma_\nu - \partial_\nu \Gamma_\mu + [\Gamma_\mu, \Gamma_\nu]. \tag{3.13}$$

1. It is a nontrivial exercise to prove the ***Bianchi identities***

$$\Omega(X,Y)Z + \Omega(Y,Z)X + \Omega(Z,X)Y = 0; \tag{3.14}$$

$$(\nabla_X R)(Y,Z) + (\nabla_Y R)(Z,X) + (\nabla_Z R)(X,Y) = 0, \tag{3.15}$$

which in coordinates read

$$R^\rho_{\sigma\mu\nu} + R^\rho_{\mu\nu\sigma} + R^\rho_{\nu\sigma\mu} = 0; \tag{3.16}$$

$$R^\rho_{\sigma\mu\nu;\tau} + R^\rho_{\sigma\tau\mu;\nu} + R^\rho_{\sigma\nu\tau;\mu} = 0. \tag{3.17}$$

Defining the ***Ricci tensor*** as before by

$$R_{\mu\nu} = R^\sigma_{\mu\sigma\nu}, \tag{3.18}$$

and the ***Ricci scalar*** by

$$R = g^{\mu\nu} R_{\mu\nu}, \tag{3.19}$$

and finally the ***Einstein tensor*** by

$$G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R, \tag{3.20}$$

the second Bianchi identity (3.17) implies what is often called *the* ***Bianchi identity*** of GR:

$$\nabla_\mu G^{\mu\nu} = 0. \tag{3.21}$$

Using the metric, once more, we may lower the upper index on R by defining

$$R_{\rho\sigma\mu\nu} = g_{\rho\tau} R^\tau_{\sigma\mu\nu}, \tag{3.22}$$

which leads to some more identities satisfied by R:

$$R_{\rho\sigma\nu\mu} = -R_{\rho\sigma\mu\nu}; \tag{3.23}$$

$$R_{\sigma\rho\mu\nu} = -R_{\rho\sigma\mu\nu}; \tag{3.24}$$

$$R_{\mu\nu\rho\sigma} = R_{\rho\sigma\mu\nu}, \tag{3.25}$$

of which the first is trivial from (3.10) and hence did not require lowering indices, the second states that each map $\Omega(X,Y)$ is an isometry of T_xM, and the third is conceptually bizarre, since, as we explained, the first pair of indices plays a completely different role from the second (and yet one might apparently interchange them).

[28] Regarding a connection as a map $\nabla : \Omega^p(E) \to \Omega^{p+1}(E)$, as in footnote 21, the corresponding curvature is simply defined as $\nabla^2 : \Omega^p(E) \to \Omega^{p+2}(E)$, so that $\nabla^2 u = R \wedge u$ for some $R \in \Omega^2(E)$. The Bianchi identity (3.15) below then simply reads $\nabla R = 0$. The simplest way to prove (3.15) is to use geodesic normal coordinates, cf. §3.3.

2. The symmetries (3.23) - (3.25) enable one to count the number of independent components of the Riemann tensor in various dimensions n, namely $n^2(n^2-1)/12$. Therefore:

   (a) In $n=2$ one has just R_{1212} which (in the Riemannian case) equals

   $$R_{1212} = \det(g)\cdot K, \tag{3.26}$$

   where K is the scalar (Gaussian) curvature; or, more directly, $K = R^{12}{}_{12}$. A simple computation using (3.18), (3.19), and (3.23) - (3.25) then shows that

   $$R = 2K. \tag{3.27}$$

   (b) In $n=3$ the Riemann tensor has 6 independent components, as does the Ricci tensor, so these two carry the same geometric information.

   (c) In $n=4$ (the case of interest to physics) the Riemann tensor has 20 independent components, whereas the Ricci tensor only has 10 (as does the Einstein tensor). The geometric information in the Riemann tensor that is not passed in to the Ricci tensor is contained in the ***Weyl tensor***, which in any dimension $n>2$ is defined by

   $$C_{\rho\sigma\mu\nu} = R_{\rho\sigma\mu\nu} + \frac{2}{n-2}(g_{\rho[\nu}R_{\mu]\sigma} + g_{\sigma[\mu}R_{\nu]\rho}) + \frac{2}{(n-1)(n-2)}(R\cdot g_{\rho[\mu}g_{\nu]\sigma}), \tag{3.28}$$

   where $[\cdots]$ denotes *antisymmetrization* in the enclosed indices, much as $(\cdots)$ denotes *symmetrization* in the enclosed indices. In the case at hand, we therefore have

   $$g_{\rho[\nu}g_{\mu]\sigma} = g_{\rho\nu}g_{\mu\sigma} - g_{\rho\mu}g_{\nu\sigma}.$$

   Using this notation, we may write (3.23) - (3.24) as $R_{(\rho\sigma)\mu\nu} = R_{\rho\sigma(\mu\nu)} = 0$, and similarly we give the symmetries Weyl's tensor inherits from Riemann's:

   $$C_{(\rho\sigma)\nu\mu} = 0; \tag{3.29}$$

   $$C_{\sigma\rho(\mu\nu)} = 0; \tag{3.30}$$

   $$C_{\mu\nu\rho\sigma} = C_{\rho\sigma\mu\nu}, \tag{3.31}$$

   The fact that the Weyl tensor is 'complementary' to the Ricci tensor comes from

   $$C_{\mu\nu} \equiv C^{\sigma}_{\mu\sigma\nu} = 0. \tag{3.32}$$

   The Weyl tensor is also called the ***conformal tensor***, since it has he following property (cf. Hawking & Ellis, p. 42): a conformal scaling of the metric $g \mapsto \hat{g} = c\cdot g$, where $c \in C^\infty(M)$ is strictly positive, does not change $C^{\rho}_{\sigma\mu\nu}$ (exercise!).

3.3 Curvature and geodesics

We now give an interpretation of curvature through ***geodesic deviation***, which physicists like. Let $U \in \mathscr{O}(\mathbb{R}^2)$ be connected and let $\gamma : U \to M$ be a family of curves: with $(s,t) \in U$ we write $\gamma_s(t) \equiv \gamma(s,t)$, regarding t as the 'time' parameter on each curve γ_s, and s as a parameter labeling the curves. Apart from the usual vector field tangent to $\gamma_s(t)$ along the t-flow, i.e.,

$$\dot{\gamma}_s \equiv \gamma_*(\partial/\partial t) = \frac{\partial \gamma_s}{\partial t}, \tag{3.33}$$

on $\gamma(U)$, which gives the tangent vectors to each γ_s for fixed s as t 'runs',[29] we now also have a second vector field tangent to $\gamma_s(t)$ along the s-flow, i.e.,

$$\gamma_s{}' \equiv \gamma_*(\partial/\partial s) = \frac{\partial \gamma_s}{\partial s}. \tag{3.34}$$

Let ∇ be the Levi-Civita connection on TM. For any vector field Z defined on $\gamma(U)$, abbreviate

$$\nabla_s Z \equiv \nabla_{\gamma_s{}'} Z; \tag{3.35}$$

$$\nabla_t Z \equiv \nabla_{\dot\gamma_s} Z. \tag{3.36}$$

Since $[\partial/\partial s, \partial/\partial t] = 0$ on $U \subset \mathbb{R}^2$, on $\gamma(U)$ we have

$$[\gamma_s{}', \dot\gamma_s] = 0. \tag{3.37}$$

Therefore, because ∇ is torsion-free we have the important identity

$$\nabla_t \gamma_s{}' = \nabla_s \dot\gamma_s. \tag{3.38}$$

Another application of (3.37), with (3.6), is that for any $Z \in \mathfrak{X}(\gamma(U))$ we have

$$[\nabla_t, \nabla_s] Z = \Omega(\dot\gamma_s, \gamma_s{}') Z. \tag{3.39}$$

Now assume that each curve $t \mapsto \gamma_s(t)$ is a geodesic, so that $\nabla_t \dot\gamma_s = 0$, and take $Z = \dot\gamma_s$. Using also (3.38), eq. (3.39) becomes ***Jacobi's equation of geodesic deviation***

$$\nabla_t^2 \gamma_s{}' = \Omega(\dot\gamma_s, \gamma_s{}') \dot\gamma_s; \tag{3.40}$$

$$\nabla_t^2 \left(\frac{\partial \gamma_s^\rho}{\partial s} \right) = R^\rho_{\sigma\mu\nu} \frac{\partial \gamma_s^\sigma}{\partial t} \frac{\partial \gamma_s^\mu}{\partial t} \frac{\partial \gamma_s^\nu}{\partial s}. \tag{3.41}$$

We now change perspective and start from a *single* geodesic γ. We then define a ***Jabobi field*** along γ as any vector field J, defined along γ, that satisfies Jacobi's equation

$$\nabla_t^2 J = \Omega(\dot\gamma, J) \dot\gamma; \tag{3.42}$$

$$\nabla_t^2 J^\rho = R^\rho_{\sigma\mu\nu} \frac{d\gamma^\mu}{dt} \frac{d\gamma^\sigma}{dt} J^\nu. \tag{3.43}$$

Clearly, any one-parameter family of geodesics produces a Jacobi field along any fixed one of them by the above procedure, and conversely:

Proposition 2 . *Any solution J of* (3.42) *or* (3.43) *along γ enables one to extend γ to a one-parameter family (γ_s) for which $\gamma = \gamma_0$ and*

$$J = \frac{d\gamma_s}{ds}(s = 0). \tag{3.44}$$

This will be proved in the next subsection, since we need the exponential map for the proof. Since (3.42) or (3.43) is linear in J, we have a vector space J_γ of Jacobi fields along $\gamma : [a,b] \to M$. Since any such J solves a second-order ODE, it is determined by $J(0)$ and $\nabla_t J(0)$, so that

$$\dim(\mathrm{J}_\gamma) = 2n. \tag{3.45}$$

If the initial conditions are $J(a) = c_1 \dot\gamma(a)$ and $\nabla_t J(a) = c_2 \dot\gamma(a)$, then, since the curvature term in (3.42) drops out (why?), the solution is simply $J(t) = (c_1 + (t-a)c_2)\dot\gamma(t)$. Hence the component of J along $\dot\gamma$ is uninteresting and one usually studies Jacobi fields *orthogonal* to the given γ: a similar computation shows that if $J(0) \perp \dot\gamma(0)$ and $\nabla_t J(0) \perp \dot\gamma(0)$, then $J(t) \perp \dot\gamma(t)$ for all t.

[29] See (1.14) for the notation γ'. However, note that in $\gamma_s{}'$ the prime denotes the s-derivative.

3.4 The exponential map

In both cases (i.e. Riemannian or Lorentzian), take $x_0 \in M$ and define $\mathscr{V}_{x_0} \subseteq T_{x_0}M$ as the set of vectors $X \in T_{x_0}M$ for which the geodesic γ_X emanating at x with initial velocity X (i.e., $\gamma_X(0) = x$ and $\dot{\gamma}_X(0) = X$) is defined at least on the entire interval $[0,1]$; if (M,g) is complete, then $\mathscr{V}_{x_0} = T_{x_0}M$ for all x. The ***exponential map*** $\exp_{x_0} : \mathscr{V}_{x_0} \to M$ is defined by

$$\exp_{x_0}(X) = \gamma_X(1). \tag{3.46}$$

1. Each $x \in M$ has a ***normal neighbourhood*** U_{x_0} for which there exists a *star-shaped* open subset $\mathscr{U}_{x_0} \subset \mathscr{V}_{x_0}$ such that $\exp_{x_0} : \mathscr{U}_{x_0} \to U_{x_0}$ is a diffeomorphism.[30] Hence any point

$$x = \exp_{x_0}(X) \in U_{x_0} \tag{3.47}$$

is connected to x_0 by a geodesic *within* U_{x_0}, viz. γ_X, where $X = \exp_{x_0}^{-1}(x)$. If $t \mapsto \gamma_X(t)$ solves (2.22), then so does $t \mapsto \gamma_{\rho X}(t/\rho)$ for any $\rho > 0$, and since their two initial conditions are the same we have $\gamma_X(t) = \gamma_{\rho X}(t/\rho)$ for $\rho > 0$. Consequently, we have

$$\gamma_X(t) = \exp_{x_0}(tX) = \gamma_{tX}(1), \tag{3.48}$$

This curve γ_X is the *unique* geodesic from x_0 to x (up to an affine reparametrization) *within* U_{x_0} (that is, there may be other geodesics from x_0 to x, but these leave U_{x_0}). To see this (cf. O'Neill, Prop. 3.31), consider an arbitrary geodesic $c : [0,1] \to M$ with $c(0) = x$ and $c(1) = x$, and take $Y = \dot{c}(0)$. Uniqueness of geodesics c with given $c(0)$ and $\dot{c}(0)$, yields $c(t) = \gamma_Y(t)$. Then $c([0,1]) \subset U_{x_0}$ implies $Y \in \mathscr{U}_{x_0}$, and the endpoint matching condition $\gamma_Y(1) = x = \gamma_X(1)$ then enforces $Y = X$, which of course implies $c = \gamma_X$.

2. *Jacobi fields give the push-forward of the exponential map.* For each $X \in \mathscr{V}_{x_0}$ we have $(\exp_{x_0})'_X : T_X(T_{x_0}M) \to T_xM$. Identifying $T_X(T_{x_0}M) \cong T_{x_0}M$ (i.e. $Z \in T_{x_0}M$ is identified with $d/dt(X+tZ)_{|t=0} \in T_X(T_{x_0}M)$), this becomes a linear map $(\exp_{x_0})'_X : T_{x_0}M \to T_xM$. Take $Z \in T_{x_0}M$ (not necessarily orthogonal to $X = \dot{\gamma}_X(0)$) and let $J_Z(t)$ be the Jacobi field along γ_X with boundary conditions $J(0) = 0$ and $\nabla_t J_Z(0) = Z$. Then for each $t \in [0,1]$),[31]

$$(\exp_{x_0})'_X(Z) = J_Z(1). \tag{3.49}$$

3. The exponential map leads to the idea of ***(geodesic) normal coordinates*** (GNC) relative to both some $x_0 \in M$ and a choice of an orthonormal basis of $T_{x_0}M$, defined (at least) on the chart U_{x_0}: the normal coordinates of $x \in U_{x_0}$ are the coordinates of $\exp_{x_0}^{-1}(x) \in T_{x_0}M$ with respect to the given basis of $T_{x_0}M$. It is a simple exercise to show that in these coordinates

$$x_0^\mu = 0; \tag{3.50}$$
$$g_{\mu\nu}(0) = \delta_{\mu\nu} \qquad \text{(Riemannian case)}; \tag{3.51}$$
$$g_{\mu\nu}(0) = \eta_{\mu\nu} \qquad \text{(Lorentzian case)}; \tag{3.52}$$
$$g_{\mu\nu,\rho} = (0) \;\Rightarrow\; \Gamma^\rho_{\mu\nu}(0) = 0. \tag{3.53}$$

[30] A subset $V \subset W$ of a vector space is ***star-shaped*** if $v \in V$ implies $tv \in V$ for all $t \in [0,1]$, see O'Neill, §3.30.

[31] The idea of the proof is to construct J_Z à la (3.44) from the family $\gamma_s(t) = \exp_{x_0}(tX + stZ)$ of geodesics. Then $\gamma_0 = \gamma_X$ and $J = (d\gamma_s/ds)_{|s=0}$ coincides with the right-hand side of (3.49). Furthermore, $J(0) = 0$ and $\nabla_t J(0) = Z$. Hence $J = J_Z$, as explained after (3.44), so that $J_Z(t) = (\exp_{x_0})'_{tX}(tZ)$. Then take $t = 1$.

Furthermore, by (3.48), in GNC geodesics γ_X emanating from x_0 are simply given by

$$\gamma_X^\mu(t) = X^\mu t, \tag{3.54}$$

and hence at $t = 1$ we have $\gamma_X^\mu(1) = x^\mu$. Eq. (2.22) then implies that in GNC,

$$\Gamma^\rho_{\mu\nu}(x)x^\mu x^\nu = 0. \tag{3.55}$$

Since the velocity $\|\dot\gamma_X(t)\| = \sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))}$ is constant along γ_X, in GNC we have

$$g_{\mu\nu}(x)x^\mu x^\nu = g_{\mu\nu}(0)x^\mu x^\nu, \tag{3.56}$$

since the left-hand side equals $\|\dot\gamma_X(1)\|^2$ and the right-hand side is $\|\dot\gamma_X(0)\|^2$.

4. We now prove Proposition 2. Given $\gamma(t)$ and $J(t)$, let $c(s)$ be the unique geodesic with

$$c(0) = \gamma(0); \tag{3.57}$$
$$c'(0) = J(0), \tag{3.58}$$

where $s \in (-\delta,\delta)$ for some $\delta > 0$, and $c'(s) = \partial c(s)/\partial s$ as usual. Then define vector fields $V(s)$ and $W(s)$ along $c(s)$ as the unique solutions of

$$\nabla_{c'}V(s) = 0; \tag{3.59}$$
$$V(0) = \dot\gamma(0); \tag{3.60}$$
$$\nabla_{c'}W = 0; \tag{3.61}$$
$$W(0) = \nabla_t J(0). \tag{3.62}$$

Then the following family does the job:

$$\gamma_s(t) = \exp_{c(s)}(tV(s) + stW(s)). \tag{3.63}$$

- For fixed s, this is $\gamma_s : t \mapsto \exp_{x_s}(tX_s)$, with $x_s = c(s)$ and $X_s = V(s) + sW(s)$. Now

$$\exp_{x_s}(tX_s) = \gamma_{tX_s}(1) = \gamma_{X_s}(t) \tag{3.64}$$

  by (3.48), so $\gamma_s = \gamma_{X_s}$, emanating from $\gamma_s(0) = x_s$. This is surely a geodesic!

- To prove (3.44), we initially put

$$\tilde J(t) = \frac{\partial\gamma_s(t)}{\partial s}(s=0). \tag{3.65}$$

  Then, using (3.57) - (3.64), we compute

$$\tilde J(0) = \frac{\partial \exp_{c(s)}(0)}{\partial s}(s=0) = \frac{dc(s)}{ds}(s=0) = c'(0) = J(0); \tag{3.66}$$

$$\begin{aligned}\nabla_t\tilde J(0) &= \nabla_t\frac{\partial}{\partial s}\gamma_s(t)_{|s=t=0} = \nabla_s\frac{\partial}{\partial t}\gamma_s(t)_{|s=t=0}\\ &= \nabla_{c'}(V(s)+sW(s))_{|s=0} = W(0) = \nabla_t J(0).\end{aligned} \tag{3.67}$$

  Since J and $\tilde J$ solve the same Jacobi equation along γ, this implies $\tilde J = J$.

5. The notable ***Gauß Lemma*** sharpens this to

$$g_{\mu\nu}(x)x^\mu = g_{\mu\nu}(0)x^\mu, \tag{3.68}$$

or, in coordinate-free form, for arbitrary $X \in \mathscr{V}_{x_0}$ and $Y \in T_{x_0}M$),

$$g_x((\exp_{x_0})'_X(X), (\exp_{x_0})'_X(Y)) = g_{x_0}(X,Y). \tag{3.69}$$

This states that the radial component of any vector along a geodesic preserves its length under the exponential map; the presence of the curvature in the right-hand side of (3.42) prevents $\exp_{x_0}$ from being an isometry (which it is in flat space). To see that (3.69) is equivalent to (3.68), note that according to (3.54), in GNC we have

$$((\exp_{x_0})'_X(X))^\mu = X^\mu, \tag{3.70}$$

so if we write $Y \in T_X(T_{x_0}M) \cong T_{x_0}M$ as $Y = d(X+sY)_{|s=0}$, by definition of the push-forward $(\exp_{x_0})'_X$ we obtain $(\exp_{x_0})'_X(Y) = d(\exp_{x_0}(X+sY))_{|s=0}$, which in GNC gives

$$((\exp_{x_0})'_X(Y))^\mu = Y^\mu. \tag{3.71}$$

Hence the left-hand side of (3.69) is $g_{\mu\nu}(x)X^\mu Y^\nu$, and since the right-hand side is obviously $g_{\mu\nu}(0)X^\mu Y^\nu$, we have proven the said equivalence.

To prove (3.68) and hence (3.69),[32] we note that (3.55) with (3.12) implies

$$(2g_{\mu\rho,\nu} - g_{\mu\nu,\rho})x^\mu x^\nu = 0. \tag{3.72}$$

Furthermore, taking (3.56) at arbitrary t, we have

$$g_{\mu\nu}(tx)x^\mu x^\nu = g_{\mu\nu}(0)x^\mu x^\nu, \tag{3.73}$$

whence

$$tg_{\mu\nu,\rho}(tx)x^\mu x^\nu + 2g_{\mu\rho}(tx)x^\mu = 2g_{\mu\rho}(0)x^\mu, \tag{3.74}$$

by taking ∂_ρ of both sides. Combining (3.72) and (3.74) yields

$$\frac{d}{dt}(tg_{\mu\rho}(tx)x^\mu - tg_{\mu\rho}(0)x^\mu) = 0. \tag{3.75}$$

Hence we may evaluate the expression between brackets at $t = 1$, which gives (3.68).

Combing (3.48), (3.49), and (3.69) then gives, along the geodesic γ_X (at least for $t \in [0,1]$),

$$g_{\gamma_X(t)}(J_X(t), J_Y(t)) = t^2 g_{x_0}(X,Y). \tag{3.76}$$

For example, on $M = \mathbb{R}^n$ with Euclidean metric (i.e. $g_{ij} = \delta_{ij}$) one simply has $J_Z(t) = tZ$.

[32] Eq. (3.69) may also be proved directly from (3.49). See O'Neill, Lemma 5.1 or Jost, Corollary 4.2.2.

3.5 Riemannian versus Lorentzian geodesics

Though formally defined in the same way, there are huge differences between geodesics in Riemannian manifolds and those in Lorentzian manifolds. First, a vector $X \in T_xM$ is called:

- ***timelike*** if $g_x(X,X) < 0$;
- ***null*** if $g_x(X,X) = 0$ and $X \neq 0$;
- ***spacelike*** if $g_x(X,X) > 0$;
- ***causal*** if $g_x(X,X) \leq 0$ and $X \neq 0$ (i.e. X is either timelike or null).

Let us denote the set of these vectors at T_xM by $\mathscr{I}_x$, $\mathscr{N}_x$, $\mathscr{S}_x$, and $\mathscr{J}_x$, respectively.

Similarly, a curve γ is called timelike (etc.) if all its tangent vectors $\dot\gamma$ are timelike (etc.). In physics, timelike curves are potential trajectories of massive particles, whereas massless particles move on null curves. More generally, physical information is supposed to spread only along causal curves (this will be one of the theorems of hyperbolic PDE theory).

In the standard basis of $\mathbb{R}^4$ with Minkowski metric, a vector like $(1,0,0,0)$ is timelike, $(1,1,0,0)$ is null, and $(0,1,0,0)$ is spacelike. Thus $\gamma(t) = (t,0,0,0)$ is a timelike curve, (even a geodesic), $\gamma(t) = (t,t,0,0)$ is a null geodesic, etc.

1. We call a Lorentzian manifold (M,g) ***time orientable*** if there exists a timelike vector field $T \in \mathfrak{X}(M)$.[33] In Minkowski space-time, just think of $T_x = (1,0,0,0)$ for all $x \in \mathbb{R}^4$. With

$$\mathscr{D}_x^+ = \{X_x \in T_xM \mid \ g_x(T_x,X_x) < 0\}, \tag{3.77}$$

and $\mathscr{D}_x^- = -\mathscr{D}_x^+$, we say that a vector X_x is ***future-directed*** (fd) if $X_x \in \mathscr{D}_x^+$, and ***past-directed*** if $X_x \in \mathscr{D}_x^-$. For example, T_x itself is future-directed.[34] We denote the set of future-directed timelike vectors by $\mathscr{I}_x^+ = \mathscr{I}_x \cap \mathscr{D}_x^+$, etc. Thus $\mathscr{I}_x^+$ is the interior of $\mathscr{J}_x^+$ and $\mathscr{N}_x^+ = \partial\mathscr{I}_x^+$ is the (topological) boundary of $\mathscr{I}_x^+$; obviously, $\mathscr{N}_x^\pm = \mathscr{I}_x^\pm \cup \mathscr{D}_x^\pm$. The set $\mathscr{N}_x = \mathscr{N}_x^+ \cup \mathscr{N}_x^-$ is called the ***light cone*** at x; it is of supreme interest to GR.

Similar terminology applies to curves, e.g. γ is *future-directed* iff $\dot\gamma(t) \in \mathscr{D}^+_{\gamma(t)}$ for all t.

   (a) For $x,y \in M$ we say that $x \ll y$ (or: x ***precedes*** y) if there exists a future-directed timelike curve starting at x and ending at y. Maximizing the length of such curves, one could replace 'curve' by 'geodesic' in this definition, and either way, one could equivalently state the definition in terms of *piecewise* smooth curves or geodesics: this is because the concatenation of two curves (or geodesic) $x \to y$ and $y \to z$, which is merely piecewise smooth, can be "smoothened" so as to become a smooth curve (or geodesic) $x \to z$. In particular, the relation $\ll$ is transitive.[35] This defines sets

$$I^+(x) = \{y \in M \mid x \ll y\}; \tag{3.78}$$

$$I^-(x) = \{y \in M \mid y \ll x\}, \tag{3.79}$$

   as well as sets $J^\pm(x)$ defined like $I^\pm(x)$ with the relation $\ll$ replaced by $<$, where $x < y$ iff there exists a future-directed *causal* curve (or geodesic) starting at x and ending at y.

[33] We will later introduce the separate and more subtle concept of a *time-function*.

[34] Although defined also for spacelike vectors, the concept of future-directed is only used in practice for causal vectors. Two timelike vectors X and Y both lie in either $\mathscr{D}_x^+$ or $\mathscr{D}_x^-$ iff $g_x(X,Y) < 0$. See O'Neill, Lemma 5.29.

[35] See e.g. R. Penrose, *Techniques of Differential Topology in Relativity* (SIAM, 1972), Prop. 2.23.

2. We may define ***length*** in the Lorentzian (and generally in the semi-Riemannian) case by

$$\|X\| = \sqrt{|g(X,X)|}, \tag{3.80}$$

and although for spacelike vectors X this behaves as expected (triangle equality etc.), for causal vectors one finds the usual inequalities in the *opposite* direction, namely

$$\|X+Y\| \geq \|X\| + ||Y\|; \tag{3.81}$$
$$|g(X,Y)| \geq \|X\| \cdot \|Y\|. \tag{3.82}$$

Nonetheless, one may also define the length of a curve $\gamma : [a,b] \to M$ almost as in the Riemannian case, cf. (2.14), namely by the parametrization-independent expression

$$L(\gamma) = \int_a^b dt\, \|\dot{\gamma}(t)\|. \tag{3.83}$$

3. One then has the following contrast between the Riemannian and the Lorentzian cases:

   (a) *Riemannian case* (R): any two 'nearby' points x, y (in that $y \in U_x$) are connected by a unique curve γ of *minimal* length (necessarily a geodesic) compared to all other curves c from x to y *within* U_x. In this case, (3.83) is of course given by (2.14).

   (b) *Lorentzian case* (L): any two points x and $y \in Ux$ with $x \ll y$ are connected by a unique timelike fd curve γ of *maximal* length (which is necessarily a geodesic) compared to all other fd timelike curves c from x to y *within* U_x. In that case,

$$L(\gamma) = \int_a^b dt\, \sqrt{-g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}. \tag{3.84}$$

The proof is as follows. Define the (unit) ***radial vector field*** R on U_x at $z = \exp_x(Z)$ by

$$R_z = \frac{(\exp_x)'_Z(Z)}{\|(\exp_x)'_Z(Z)\|}, \tag{3.85}$$

so that $g(R_z, R_z) = +1$ (R) and -1 (L). Taking a curve c as defined above, decompose

$$\dot{c} = \pm g(\dot{c}, R)R + N, \tag{3.86}$$

where $g(N,R) = 0$, again with the + sign for R and the - sign for L. It follows that

$$\|\dot{c}\|^2 = g(\dot{c}, R)^2 \pm g(N,N), \tag{3.87}$$

with $g(N,N) \geq 0$ also for L (as the vector R is timelike, and hence N is spacelike). Hence

$$\|\dot{c}\| \geq g(\dot{c}, R) \ \ (R); \tag{3.88}$$
$$\|\dot{c}\| \leq -g(\dot{c}, R) \ \ (L), \tag{3.89}$$

since for L we have $g(\dot{c}, R) < 0$. We define the ***radius function*** $r : U_x \to \mathbb{R}^+$ by

$$r(\exp_x(Z)) = \|Z\|. \tag{3.90}$$

For any curve $t \mapsto c(t) = \exp_x(C(t))$ in U_x (assumed timelike in the L-case) with $c(0) = x$,

$$\frac{d}{dt} r \circ c(t) = \frac{d}{dt}\|C(t)\| = \frac{d}{dt}\sqrt{\pm g_x(C(t),C(t))} = \pm \frac{g_x(\dot{C}(t),C(t))}{\sqrt{\pm g_x(C(t),C(t))}}$$
$$= \pm \frac{g_{c(t)}((\exp_x)'_{C(t)}(\dot{C}(t)),(\exp_x)'_{C(t)}(C(t)))}{\sqrt{\pm g_{c(t)}((\exp_x)'_{C(t)}(C(t)),(\exp_x)'_{C(t)}(C(t)))}} = \pm g_{c(t)}(\dot{c},R), \qquad (3.91)$$

where we used Gauß's lemma in both the denominator and the numerator. Therefore,

$$L(c) = \int_0^1 dt\, \|\dot{c}(t)\| \geq \int_0^1 dt\, g(\dot{c}(t),R)_{c(t)} = \Big|_0^1 r \circ c = r(y) \qquad \text{(R)};$$
$$L(c) = \int_0^1 dt\, \|\dot{c}(t)\| \leq -\int_0^1 dt\, g(\dot{c}(t),R)_{c(t)} = \Big|_0^1 r \circ c = r(y) \qquad \text{(L)}.$$

On the other hand, "the" geodesic within U_x from x to $y = \exp_x(Y)$ is given by γ_Y, where

$$L(\gamma_Y) = \int_0^1 dt\, \|\dot{\gamma}_Y(t)\| = \|\dot{\gamma}_Y(0)\| = \|Y\| = r(y), \qquad (3.92)$$

since for geodesics $\gamma = \gamma_Y$ the velocity $\|\dot{\gamma}(t)\|$ is t-independent. Thus we conclude that:[36]

$$L(c) \geq L(\gamma_Y) \qquad \text{(R)}; \qquad (3.93)$$
$$L(c) \leq L(\gamma_Y) \qquad \text{(L)}. \qquad (3.94)$$

We finally prove uniqueness of γ, in that we have strict inequalities in (3.93) - (3.94) except when c is a (necessarily affine) reparametrization of γ_Y. This goes back to (3.87), which yields equalities in (3.88) - (3.89) iff $g(N,N) = 0$, which is the case iff $\dot{(c)}$ is proportional to the radial vector field R. This yields the claim.

4. Similarly but with (even) more effort, one can prove that *the causal structure of a Lorentzian manifold 'near' $x \in M$ is determined by its linearized structure in T_xM*, in the sense that

$$I^{\pm}(x) \cap U_x = \exp_x(\mathscr{I}_x^{\pm} \cap \mathscr{U}_x); \qquad (3.95)$$
$$N^{\pm}(x) \cap U_x = \exp_x(\mathscr{N}_x^{\pm} \cap \mathscr{U}_x); \qquad (3.96)$$
$$J^{\pm}(x) \cap U_x = \exp_x(\mathscr{J}_x^{\pm} \cap \mathscr{U}_x), \qquad (3.97)$$

where $N^{\pm}(x) = I^{\pm}(x) \cap J^{\pm}(x)$. In (other) words, timelike/null/causal curves (or geodesics) emanating from any point $x \in M$ are precisely the images of their linearized counterparts in T_xM under the exponential map $\exp_x$, at least in the neighbourhood $U_x \subset M$ of x where this map is a diffeomorphism.

Although this may sound obvious, the proof is quite nontrivial.[37] The main point is that according to Gauß's Lemma (3.69), the two relevant notions of being timelike (etc.) and future-directed are preserved by the exponential map.

[36] In the timelike (L) case one can decrease the length of a timelike geodesic γ by pushing it towards null curves: *If $x, y \in M$ can be connected by a timelike curve, then for any $\varepsilon > 0$ there is a timelike curve c (far from a geodesic!) with length $L(c) < \varepsilon$* (although there is no such curve with length zero). See Malament, Prop. 2.3.2.

[37] See also Proposition 4.5.1 in Hawking & Ellis, whose (rather vague) proof is based on (our) Lemma 3 below, which implies that for a timelike geodesic, each level set $\mathscr{S}_\rho$ must be spacelike. Consequently, for a timelike geodesic γ the function $t \mapsto g_{\gamma(t)}(\dot{\gamma}(t),\dot{\gamma}(t))$ is monotonically decreasing (since it can never pick up a positive contribution from a spacelike component), so that if γ starts out as a timelike geodesic (as is determined by its tangent vector and hence by the state of affairs in T_xM), it must always remain timelike. A sharper form with a complete proof is Proposition 2.1 in Senovilla.

5. We close this section with a very neat and intuitive application of Gauß's Lemma. Let $\rho > 0$ be such that $B_\rho = \{Y \in T_{x_0}M \mid \|Y\| \leq \rho\}$ lies in $\mathscr{V}_{x_0}$, and define the level sets

$$S_\rho = \{Y \in T_{x_0}M \mid \|Y\| = \rho\}; \tag{3.98}$$
$$\mathscr{S}_\rho = \exp_{x_0}(S_\rho). \tag{3.99}$$

In the Riemannian case $\mathscr{S}_\rho$ is connected, whereas in the Lorentzian case it is disconnected: in Minkowski space-time it has one sheet at spacelike distance from x_0 and two sheets–one in the fuure and one in the past–at timelike from distance from x_0. In what follows one should avoid null geodesics (since these do not cross $\mathscr{S}_\rho$). First, each straight curve $t \mapsto tX$ in $T_{x_0}M$ crosses S_ρ orthogonally: if $Z \in T_{tX}S_\rho$, where $\rho = t\|X\|$, then $Z = dc(\lambda)/d\lambda_{|\lambda=0}$ for some curve $c(\lambda)$ in S_ρ with $c(0) = tX$, i.e., $|g(c(\lambda), c(\lambda))| = \rho^2$, whence $dg(c(\lambda), c(\lambda)) = /d\lambda_{|\lambda=0}$, whence $g(X,Z) = 0$ (as in $\mathbb{R}^n$ with flat metric).

The point is that (3.48) and (3.69) imply the same for the images under $\exp_{x_0}$: for any $W \in T_w\mathscr{S}_\rho \subset T_wM$ with $w = \gamma_X(t)$ and hence $\rho = t\|X\|$, we have $g_w(W, \dot\gamma_X(t)) = 0$, hence:

Lemma 3 *Each (timelike) geodesic γ_X from x_0 crosses the level set $\mathscr{S}_\rho$ orthogonally.*

3.6 Conjugate points: definition

What happens for 'far away' (instead of 'nearby') points in regard to the extremizing properties of geodesics? This question is answered through the notion of *conjugate points*, which also play an important role in the singularity theorems of GR. To motivate their definition, we compute the second variation of the length functional (3.83). Let us do the Riemannian case and insert the appropriate sign(s) for the Lorentzian case at the end. First, we recompute the first variation, using the powerful notion of the covariant derivative that was not yet available to us in §2.3. Note that, in contrast to our discussion of Jacobi fields, here we neither assume that each γ_s is a geodsic, nor (for later use in computing the second derivative) that it is parametrized by arc length (i.e. has constant speed). Using (2.54) and (2.55), (3.35) - (3.36), and (3.38), we obtain

$$\begin{aligned}\frac{dL(\gamma_s)}{ds} &= \int_a^b dt\, \frac{\partial}{\partial s}\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t), \dot\gamma_s(t))} \\ &= \int_a^b dt\, \frac{g_{\gamma_s(t)}(\nabla_s\dot\gamma_s(t), \dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t), \dot\gamma_s(t))}} = \int_a^b dt\, \frac{g_{\gamma_s(t)}(\nabla_t {\gamma_s}'(t), \dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t), \dot\gamma_s(t))}}.\end{aligned} \tag{3.100}$$

If we now do put $s = 0$ (with $\gamma_0 = \gamma$) and do assume constant speed, say $\|\dot\gamma(t)\| = v$, we continue:

$$\begin{aligned}\int_a^b dt\, \frac{g_{\gamma(t)}\nabla_t\gamma'(t), \dot\gamma(t))}{\sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}} &= \frac{1}{v}\int_a^b dt\, [\partial_t(g_{\gamma(t)}(\gamma'(t), \dot\gamma(t))) - g_{\gamma(t)}(\gamma'(t), \nabla_t\dot\gamma(t))] \\ &= \frac{1}{v}\left(\Big|_b^a g(\gamma', \dot\gamma) - \int_a^b dt\; g_{\gamma(t)}(\gamma'(t), \nabla_{\dot\gamma}\dot\gamma(t))\right),\end{aligned} \tag{3.101}$$

since $\nabla_t = \nabla_{\dot\gamma}$. For fixed-endpoint variations, where $\gamma'(a) = \gamma'(b) = 0$, we therefore obtain

$$L'(\gamma) \equiv \frac{dL(\gamma_s)}{ds}(s=0) = -\frac{1}{v}\int_a^b dt\; g_{\gamma(t)}(\gamma'(t), \nabla_{\dot\gamma}\dot\gamma(t)), \tag{3.102}$$

since the boundary term in (3.101) vanishes. Thus we see that the extremality condition $L'(\gamma) = 0$ enforces the geodesic equation (2.40), since γ' in (3.102) is arbitrary and g is nondegenerate.

We now compute the second derivative of $L(\gamma_s)$ from (3.100):[38]

$$\begin{aligned}\frac{d^2L(\gamma_s)}{ds^2}(s=0) &= \int_a^b dt\,\frac{\partial}{\partial s}\left(\frac{g_{\gamma_s(t)}(\nabla_t{\gamma_s}'(t),\dot\gamma_s(t))}{\sqrt{g_{\gamma_s(t)}(\dot\gamma_s(t),\dot\gamma_s(t))}}\right)(s=0)\\ &= \frac{1}{v}\int_a^b dt\,[g_{\gamma_s(t)}(\nabla_s\nabla_t{\gamma_s}'(t),\dot\gamma_s(t)) + g_{\gamma_s(t)}(\nabla_t{\gamma_s}'(t),\nabla_s\dot\gamma_s(t))](s=0)\\ &\quad - \frac{1}{v^3}\int_a^b dt\,[g_{\gamma(t)}(\nabla_t\gamma'(t),\dot\gamma(t))]^2, \end{aligned} \tag{3.103}$$

where we used (3.38) to obtain the last term. We rewrite the first term using (3.39), which gives

$$\begin{aligned} g_{\gamma_s}(\nabla_s\nabla_t{\gamma_s}',\dot\gamma_s)_{|s=0} &= g_\gamma([\nabla_s,\nabla_t]\gamma',\dot\gamma) + g_\gamma(\nabla_t\nabla_s\gamma',\dot\gamma)\\ &= -g_\gamma(\Omega(\dot\gamma,\gamma')\gamma',\dot\gamma) - g_\gamma(\nabla_s\gamma',\nabla_t\dot\gamma) + \frac{d}{dt}g_\gamma(\nabla_s\gamma',\dot\gamma). \end{aligned} \tag{3.104}$$

In the last line, the first term equals $-R_\gamma(\dot\gamma,\gamma',\dot\gamma,\gamma')$, the second term vanishes for geodesics, and for fixed-endpoint variations the third term as usual vanishes upon integration $\int_a^b dt$. Furthermore, we use (3.38), so that $g_\gamma(\nabla_t{\gamma_s}',\nabla_s\dot\gamma_s) = g_\gamma(\nabla_t{\gamma_s}',\nabla_t{\gamma_s}')$. Introducing the component

$$\gamma'_\perp = \gamma' - v^{-2}g(\gamma',\dot\gamma)\dot\gamma \tag{3.105}$$

of γ' that is perpendicular to $\dot\gamma$, we have, omitting terms containing $\nabla_t\dot\gamma \equiv \nabla_{\dot\gamma}\dot\gamma = 0$,

$$g_\gamma(\nabla_t\gamma',\nabla_t\gamma') - \frac{1}{v^2}[g_\gamma(\nabla_t\gamma',\dot\gamma)]^2 = g_\gamma(\nabla_t\gamma'_\perp,\nabla_t\gamma'_\perp). \tag{3.106}$$

Up to a boundary term vanishing upon integration for fixed-endpoint variations, we may replace the right-hand side by $-g_\gamma(\gamma'_\perp,\nabla_t^2\gamma'_\perp)$. By the symmetries of the Riemann tensor, we have

$$-R_\gamma(\dot\gamma,\gamma',\dot\gamma,\gamma') = -R_\gamma(\dot\gamma,\gamma'_\perp,\dot\gamma,\gamma'_\perp) = R_\gamma(\gamma'_\perp,\dot\gamma,\dot\gamma,\gamma'_\perp) = g(\gamma'_\perp,\Omega(\dot\gamma,\gamma'_\perp)\dot\gamma), \tag{3.107}$$

so that we finally obtain ***Synge's formula*** for the second variational derivative of $L(\gamma)$:[39]

$$L''(\gamma) \equiv \frac{d^2L(\gamma_s)}{ds^2}(s=0) = -\frac{1}{v}\int_a^b dt\,g_{\gamma(t)}(\gamma'_\perp(t),\nabla_t^2\gamma'_\perp(t) - \Omega(\dot\gamma(t),\gamma'_\perp(t))\dot\gamma(t)). \tag{3.108}$$

Note that we did not assume that the curves γ_s were geodesics, except $\gamma_0 \equiv \gamma$. In the Lorentzian case, for *timelike* curves, one obtains exactly the same formula *without the minus sign*, which goes back to the one in (3.84); we invite the reader to redo the calculation for this case.[40]

As in calculus, $L(\gamma)$ is a local minimum iff $L''(\gamma) > 0$, whereas it is a local maximum iff $L''(\gamma) < 0$. It is clear from (3.108) and (3.42) that the critical case $L''(\gamma) = 0$ appears precisely when $\gamma'_\perp$ is a Jacobi field. This motivates the following definition:

[38] We elaborate on Joos, Proof of Theorem 4.1.1, p. 169.

[39] It is quite remarkable that not just in the first variation (3.102), where it is expected, but also in the second variation (3.108), only the *first* s-derivative of the family γ_s appears.

[40] See also O'Neill, Theorem 10.4.

Definition 4 *A* **conjugate point** *along a geodesic $\gamma : [a,b] \to M$ relative to $\gamma(a)$ is a point $\gamma(c)$, $c \in [a,b]$ for which there exists a nonzero Jacobi field J along $\gamma([a,c])$ that vanishes at a and c.*

If J arises from a variation of γ as in (3.44), then the boundary condition $J(a) = J(c) = 0$ means that the variation γ_s fixes the endpoints of γ.[41]

Theorem 5 *1.* Riemannian case: *A geodesic $\gamma : [a,b] \to M$ locally* **minimizes** *the length of curves from $\gamma(a)$ to $\gamma(b)$ iff there is* no *conjugate point on γ that lies between x and y.*

2. Lorentzian case*: A* timelike *geodesic $\gamma : [a,b] \to M$ locally* **maximizes** *the length of curves from $\gamma(a)$ to $\gamma(b)$ iff there is* no *conjugate point on γ that lies between x and y.*

The "$\Leftarrow$" part may be proved by remarking that, as we saw in §3.5.3, in the Lorentzian case timelike geodesics start out maximizing length, so that $L''(\gamma) < 0$. According to (3.108), this remains the case until a conjugate point is encountered, so if this is never the case, one will have $L''(\gamma) < 0$ forever (or at least as long as the geodesic is defined). Likewise in the R case.

For the '$\Rightarrow$" part, we show that the sign of $L''(\gamma)$ may indeed change once a conjugate point (at which its value is zero) has been crossed; in the L case, $L''(\gamma)$ then becomes positive, and a timelike geodesic can be constructed that is longer than the given one, whereas in the R case the opposite sign change leads to new and *shorter* geodesics between the given endpoints).[42]

Indeed, let $c \in (a,b)$, with associated Jacobi field J along $\gamma([a,c])$ for which $J(a) = 0$ and $J(c) = 0$. Then $\nabla_t J(c) \neq 0$ (since otherwise $J \equiv 0$), and by Proposition 2 there exists a one-parameter family of geodesics (γ_s) for which $J = \gamma'_{|s=0}$; since only the component of J that is orthogonal to $\dot\gamma$ is relevant, we can make J orthogonal to $\dot\gamma$ altogether, cf. the discussion after the statement of Proposition 2. Furthermore, we extend J from $\gamma([a,c])$ to $\gamma([a,b])$ by making it zero on $(c,b]$. Now find any vector field K along $\gamma : [a,b] \to M$ that is also orthogonal to $\dot\gamma$ and in addition satisfies the boundary conditions

$$K(a) = K(b) = 0; \tag{3.109}$$

$$g_{\gamma(a)}(\nabla_t J, K) = 0; \tag{3.110}$$

$$g_{\gamma(c)}(\nabla_t J, K) = -v. \tag{3.111}$$

This is possible, since unlike the Jacobi field J, the vector field K is not meant to satisfy any particular equation. We now take $\varepsilon > 0$ and consider the vector field $M = \varepsilon K + \varepsilon^{-1} J$. For any family of curves for which $\gamma'_{|s=0} = M$, we then compute the second variation (3.108), in which by construction $\gamma'_\perp$ is replaced by M. Since J satisfies the Jacobi equation, the term proportional to ε^{-2}, which only involves J, vanishes. The term proportional to ε^2, which only involves K, stands, call it $C\varepsilon^2$ (where C may have either sign). One of the cross terms proportional to $\varepsilon \cdot \varepsilon^{-1} = 1$, involving each of J and K linearly, vanishes by the Jacobi equation for J. In the L case to be specific (where the - sign in (3.108) has to be deleted), the other cross term contributes

$$L''(\gamma) = C\varepsilon^2 + \frac{1}{v}\int_a^c dt\, g_{\gamma(t)}(J(t), \nabla_t^2 K(t) - \Omega(\dot\gamma(t), K(t))\dot\gamma(t)). \tag{3.112}$$

[41]The idea should be clear from the two-sphere, where a continuous family of geodesics emanates from (say) the South Pole, meeting again at the north pole (which, then, is conjugate to the South Pole relative to any of these geodesics). The converse is not true, however: the existence of a nonzero Jacobi field J along $\gamma([a,c])$ that vanishes at both a and c does not guarantee the existence of even two geodesics from $\gamma(a)$ to $\gamma(c)$.

[42]Our proof is based on the final part of the proof of Hawking & Ellis, Prop. 4.5.8. For alternative proofs see Jost, Theorem 4.3.1, for R and O'Neill, Proposition 10.10 and Theorem 10.17, or Wald, Theorem 9.5.3, for L.

Here, using (2.54) and (2.55), we have

$$g_{\gamma(t)}(J(t),\nabla_t^2 K(t)) = \frac{d}{dt}(g_{\gamma(t)}(J(t),\nabla_t K(t))) - g_{\gamma(t)}(\nabla_t J(t),\nabla_t K(t)), \tag{3.113}$$

of which the first term vanishes upon integration, as $J(a) = J(c) = 0$. The second term gives

$$-g_{\gamma(t)}(\nabla_t J(t),\nabla_t K(t)) = -\frac{d}{dt}(g_{\gamma(t)}(\nabla_t J(t),K(t))) + (g_{\gamma(t)}(\nabla_t^2 J(t),K(t)), \tag{3.114}$$

whose last term combines with the curvature term in (3.112) to contribute

$$g_{\gamma(t)}(K(t),\nabla_t^2 J(t) - \Omega(\dot\gamma(t),J(t))\dot\gamma(t)),$$

which vanishes by the Jacobi equation for J (using the symmetries of the Riemann tensor R). Finally, the first term in (3.114) gives, upon integration, +1, so that overall we obtain $L''(\gamma) = C\varepsilon^2 + 1$. Whatever the sign of C, for ε small enough we can arrange $L''(\gamma) > 0$, and so, since it started out negative, the sign of $L''(\gamma)$ has changed across a conjugate point, as claimed.[43]

3.7 Conjugate points: existence

In GR (especially in the context of the singularity theorems) the existence of conjugate points is proved in a very specific way, which we now explain. The following constructions on a Lorentzian manifold may be performed in either the timelike or the null case, and since it is enough to make our point we take the simpler former case. We start from a fd timelike vector field $u \in \mathfrak{X}(U)$ defined locally on some open $U \subset M$, normalized such that, at each $x \in U$,

$$g_x(u_x,u_x) = u_\mu(x)u^\mu(x) = -1. \tag{3.115}$$

Integrating this vector field, one obtains a ***congruence of timelike curves*** in U, i.e. a foliation of U by timelike curves c; *vice versa*, such a congruence yields $u = \dot c$ as its tangent. Two examples:

1. The field u could be the 4-velocity of some (relativistic) fluid moving in the cosmos.

2. In the 3+1 split of M considered later, we will assume the existence of a ***time-function*** $\mathsf{t} : M \to \mathbb{R}$ with (nowhere vanishing) *timelike* gradient vector field $\nabla\mathsf{t}$, defined by

$$\nabla\mathsf{t} = \sharp(d\mathsf{t}); \tag{3.116}$$
$$(\nabla\mathsf{t})^\mu = g^{\mu\nu}\partial_\nu\mathsf{t}. \tag{3.117}$$

   One then takes $u = n$ to be unit vector field proportional to $\nabla\mathsf{t}$, in other words, one defines

$$n = -L\nabla\mathsf{t}; \tag{3.118}$$
$$L = 1/\sqrt{-g(\nabla\mathsf{t},\nabla\mathsf{t})}. \tag{3.119}$$

[43] It is by no means excluded that there may be other variations for which $L''(\gamma)$ remains negative (for example, by picking some K for which the sign in (3.111) is positive). All that has been proved is the existence of a family of variations for which the sign does change, which is enough to prove the theorem. A more precise way to handle this situation is to introduce the *index* form for the second variation of L, which, across a conjugate point, loses its property of being negative definite (L) or positive definite (R). See Jost, O'Neill, etc.

The function L introduced here is called the ***lapse***.[44] It follows that the 3-d hypersurfaces

$$\Sigma_t = \{x \in M \mid \mathrm{t}(x) = t\} \tag{3.120}$$

are orthogonal to $\nabla \mathrm{t}$. Conversely, one could start by assuming a foliation $M = \cup_t \Sigma_t$ of M by spacelike hypersurfaces Σ_t, and define $u = n$ as a unit normal vector field to the Σ_t.

Such a timelike vector field u defines a fair amount of derived tensors, each of some importance:

$$a^\mu = u^\nu \nabla_\nu u^\mu \qquad \textbf{\textit{(acceleration)}}; \tag{3.121}$$

$$h^\mu_\nu = \delta^\mu_\nu + u^\mu u_\nu \qquad \textbf{\textit{(spatial projection)}}; \tag{3.122}$$

$$k_{\mu\nu} = h^\rho_\mu h^\sigma_\nu \nabla_\rho u_\sigma \qquad \textbf{\textit{(- extrinsic curvature)}}; \tag{3.123}$$

$$\omega_{\mu\nu} = k_{[\mu\nu]} \qquad \textbf{\textit{(vorticity)}}; \tag{3.124}$$

$$\sigma_{\mu\nu} = k_{(\mu\nu)} - \tfrac{1}{3}\theta h_{\mu\nu} \qquad \textbf{\textit{(shear)}}; \tag{3.125}$$

$$\theta = \nabla_\mu u^\mu \qquad \textbf{\textit{(expansion)}} \tag{3.126}$$

where (as agreed earlier) $k_{(\mu\nu)} = \frac{1}{2}(k_{\mu\nu} + k_{\nu\mu})$ and $k_{[\mu\nu]} = \frac{1}{2}(k_{\mu\nu} - k_{\nu\mu})$. It follows that

$$k_{\mu\nu} = \tfrac{1}{3}\theta h_{\mu\nu} + \sigma_{\mu\nu} + \omega_{\mu\nu}; \tag{3.127}$$

$$\nabla_\mu u_\nu = -u_\mu a_\nu + k_{\mu\nu}. \tag{3.128}$$

Eq. (3.127) is trivial. The second can be checked by contracting both sides first with u^μ, then with u^ν, and finally with vectors orthogonal to u. The first contraction merely reproduces the definition (3.121). For the second we use (3.115), (2.54), and (2.55) to compute

$$0 = \partial_\mu g(u,u) = (\nabla_\mu g)(u,u) + g(\nabla_\mu u, u) + g(u, \nabla_\mu u) = 0 + 2g(u, \nabla_\mu u), \tag{3.129}$$

whence $u^\nu \nabla_\mu u_\nu = g(u, \nabla_\mu u) = 0$. Hence the second contraction gives $0 = 0 + 0$. Finally, the third contracting reproduces the definition (3.123). What is the meaning of (3.121) - (3.126)? The interpretation of $a = \nabla_u u$ should be clear; it vanishes for congruences of geodesics, for which

$$k_{\mu\nu} = \nabla_\mu u_\nu. \tag{3.130}$$

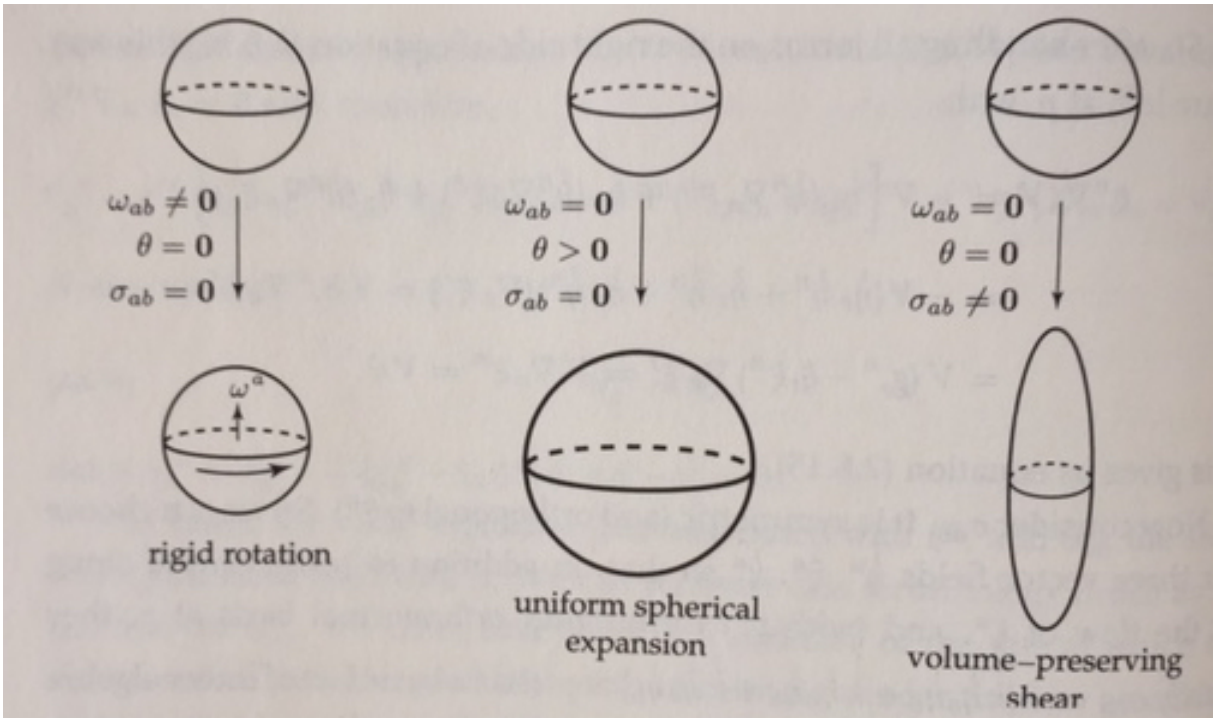
Furthermore, eq. (3.115) implies

$$h^\mu_\nu u^\nu = 0, \tag{3.131}$$

and if $g(u,v) = 0$, then $h^\mu_\nu v^\nu = v^\mu$, so that h_x projects onto the orthogonal complement of u_x, In the second example, this is $T_x\Sigma_t$, in which case the tensor $h_{\mu\nu}$ is a four-dimensional version of the three-dimensional induced metric in Σ_t, in that $h_{\mu\nu} = g(h^\rho_\mu \partial_\rho, h^\sigma_\nu \partial_\sigma)$, as is easily checked.

We return to the extrinsic curvature in Chapter 6; this geometric term only makes sense in the second example above. The three remaining terms, on the other hand, refer to the first example of fluids: the vorticity tensor (which vanishes in the second example) describes the rotation of the fluid, the shear (which is traceless) describes the directed volume-preserving expansion (or, if negative, the contraction), and θ gives the rate of total volume increase (or, if negative, the decrease) under the flow. See picture (Malament, p. 174, without permission):

[44] The existence of a time-function makes M time-orientable, with $T = -\nabla \mathrm{t}$; in Minkowski space-time, with $\mathrm{t} = x^0$, this would be $T = \partial_t$, whence the minus sign in T. The minus sign in (3.118) then makes n future-directed.



We now derive the fundamental ***Raychaudhuri equation*** for θ. Using (3.10), we compute

$$\begin{aligned} u^\sigma \nabla_\sigma (\nabla_\mu u_\nu) &= u^\sigma (\nabla_\mu \nabla_\sigma + [\nabla_\sigma, \nabla_\mu]) u_\nu \\ &= \nabla_\mu (u^\sigma \nabla_\sigma u_\nu) - (\nabla_\mu u^\sigma) \nabla_\sigma u_\nu + R_{\nu\rho\sigma\mu} u^\sigma u^\rho . \end{aligned} \tag{3.132}$$

For geodesics the first term vanishes. Eqs. (3.126), (3.130), and (3.127) then yield, along u,

$$\nabla_u \theta \equiv \dot{\theta} = -\tfrac{1}{3}\theta^2 - \sigma_{\mu\nu}\sigma^{\mu\nu} + \omega_{\mu\nu}\omega^{\mu\nu} - R_{\mu\nu} u^\mu u^\nu . \tag{3.133}$$

Since $\sigma_{\mu\nu}$ is symmetric, we have $\sigma_{\mu\nu}\sigma^{\mu\nu} = \mathrm{Tr}\,(\sigma^2) \geq 0$, where σ is the matrix with components $\sigma^\mu_\nu = h^{\mu\rho}\sigma_{\rho\nu}$. In the coming proofs of the singularity theorems we will apply this equation in the context of example 2 above, where $\omega = 0$. Furthermore, natural positive energy conditions on the matter content of the universe in combination with the Einstein equations give

$$R_{\mu\nu} u^\mu u^\nu \geq 0. \tag{3.134}$$

Therefore, the Raychaudhuri equation (3.133) gives $\dot{\theta} + \frac{1}{3}\theta^2 \leq 0$. If we now assume that

$$\theta_0 \equiv \theta(t_0) < 0 \tag{3.135}$$

at some time $t = t_0 \in [a,b]$, then $\theta(t) \neq 0$ near t_0 and hence $d(\theta^{-1})/dt \geq \frac{1}{3}$ near t_0, or

$$\theta^{-1}(t) \geq \theta_0^{-1} + \tfrac{1}{3}t. \tag{3.136}$$

This implies that θ^{-1} reaches the value zero, or $\theta \to -\infty$, at some $t_s \in (t_0, t_0 + 3/|\theta_0|]$, provided, of course, that the geodesic in question can indeed be extended to t_s.

We now transform this into a conclusion about conjugate points. Intuitively, gravity is attractive and leads to positive curvature as in (3.134), making geodesics converge, as on the sphere. Mathematically, we return to Jacobi's equation (3.42) - (3.43). We take some fixed

geodesic $\gamma : [a,b] \to M$ with $\dot\gamma = u$, and consider the three-dimensional vector space of Jacobi fields along γ with initial conditions

$$J(a) = 0; \tag{3.137}$$

$$\dot J(a) \perp \dot\gamma(a). \tag{3.138}$$

It is convenient to introduce a three-dimensional frame $(e_1(t), e_2(t), e_3(t))$ along $\gamma(t)$ that is an orthonormal basis of $T^{\perp}_{\gamma(t_0)}\dot\gamma$ and satisfies $\nabla_{\dot\gamma} e_i = 0$; this guarantees that the frame remains orthonormal as well as orthogonal to $\dot\gamma$.[45] Then

$$J = J^i e_i \equiv \sum_{i=1}^{3} J^i e_i, \tag{3.139}$$

with $J^i = g(J, e_i)$. Therefore, since $\dot\gamma = u$ and ∇ is torsion-free (which gives $\nabla_{\dot\gamma} J = \nabla_J \dot\gamma$), we may compute, using $\omega_{\mu\nu} = 0$ and hence $k_{ij} = k_{ji}$ (this is not really necessary, but convenient):

$$\dot J^i \equiv \frac{dJ^i}{dt} = \nabla_{\dot\gamma} J^i = \nabla_{\dot\gamma} g(J, e_i) = g(\nabla_{\dot\gamma} J, e_i) = g(\nabla_J \dot\gamma, e_i) = J^j g(\nabla_j u, e_i) = k_{ij} J^j. \tag{3.140}$$

Linearity of Jacobi's equation–in $J(t)$) and hence also in the initial data $\dot J(a)$, cf. (3.137)–gives

$$J^i(t) = A_{ij}(t) \dot J^j(t_0) \tag{3.141}$$

for some 3×3 matrix $A(t)$, so we have

$$\dot J^i(t) = \dot A_{ij}(t) \dot J^j(t_0) = k_{ij}(t) J^j(t) = k_{ij}(t) A_{jk}(t) \dot J^k(t_0), \tag{3.142}$$

so that $\dot A_{ik} = k_{ij} A_{jk}$, or $k = \dot A A^{-1}$, and hence, since $\theta = \mathrm{tr}(k) = \mathrm{tr}(k)$, we finally obtain

$$\theta = \mathrm{tr}(\dot A A^{-1}). \tag{3.143}$$

Now A itself is finite along γ, and so is $\dot A$, so if, in the scenario just considered, θ starts out with some negative value at t_0, it can only blow up at t_s if $A(t_s)^{-1}$ does, i.e., if $A(t_s)$, which equals the identity at $t = t_0$, has an eigenvalue zero. But this implies that there exists some initial value $J(t_0)$ for which $J(t_s) = 0$, which by definition means that $\gamma(t_s)$ is a conjugate point with respect to $\gamma(t_0)$. In summary: if $\theta(\gamma(t_0)) < 0$ somewhere along γ, then $\gamma(t_s)$ is a conjugate point with respect to $\gamma(t_0)$ iff $\lim_{t \to t_s} \theta(t) = -\infty$, and hence we have proved an important result:

Proposition 6 *Let γ be an element of a congruence of timelike geodesics (or, equivalently, a timelike vector field u such that $\nabla_u u = 0$, with $\dot\gamma = u$) with vanishing vorticity (which is the case iff the congruence is orthogonal to some foliation of M or $U \in \mathscr{O}(M)$ by spacelike hypersurfaces). If the positive curvature condition* (3.134) *holds along the congruence, and if in addition $\theta(\gamma(t_0)) < 0$ somewhere along γ, then γ has a (later) conjugate point relative to $\gamma(t_0)$, provided that the geodesic in question can indeed be extended from t_0 all the way to t_s.*

[45] This simple construction works because γ is a geodesic. Along more general curves one needs the so-called ***Fermi derivative*** $\nabla^F_{\dot\gamma} e_i$ instead of the covariant derivative $\nabla_{\dot\gamma} e_i$, see Hawking & Ellis, §4.1.

4 Singularity Theorems

Proposition 6 is a key to the singularity theorems of Hawking and Penrose, which were suggested by the earliest exact solutions to Einstein's equations.[46]

$$ds^2 = -dt^2 + a(t)^2(d\chi^2 + f(\chi)^2(d\theta^2 + \sin^2\theta d\varphi^2)), \tag{4.1}$$

where ds^2 is just the physicists' notation for the metric, the space-time is $M = (0,\infty)\times\Sigma$, where $\Sigma = S^3$ (the 3-sphere) and $f(\chi) = \sin\chi$ for $k = 1$ (positive curvature), $\Sigma = \mathbb{R}^3$ and $f(\chi) = \chi$ for $k = 0$ (no curvature), and $\Sigma = H^3$ (the 3-dimensional hyperboloid) and $f(\chi) = \sinh\chi$ for $k = -1$ (negative curvature). Finally, the function $a(t)$ depends on the precise matter content of the Universe. These are the three versions of the ***Friedman (–Robertson–Walker) Universe***, and the point of interest here is that although the metric looks reasonably well behaved as $t \to 0$ (e.g. $a(t) \sim t^{2/3}$ for a dust-filled spatially flat universe, which is admittedly not smooth at $t = 0$), the Ricci scalar (and hence the geometry) blows up. The precise form of $R(t)$ again depends on the matter, but in the same case one finds $R(t) \sim t^{-2}$. Note that *the point $t = 0$ is not included in the space-time M* (where we assumed smoothness of all things!).

Potential or actual singularities are even more glaring in the ***Schwarzschild solution***

$$ds^2 = \left(1 - \frac{2m}{r}\right)dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \tag{4.2}$$

where $m > 0$ is the mass of some gravitating object and $M = \mathbb{R}\times\Sigma$, where at least initially, in polar coordinates (r,θ,φ), the spatial part $\Sigma \subset \mathbb{R}^3$ is restricted to $r > 2m$. Here the value $r = 2m$ looks threatening, as does $r = 0$ (although the latter is not in the domain of the solution).

Even Hilbert and Einstein himself were confused about the meaning of these apparent or real singularities,[47] but today it is clear that $r = 2m$ is just a singularity of the coordinate system in which the Schwarzschild solution is expressed,[48] whereas $r = 0$ would be a real singularity, where, as in the Friedman solution, the curvature blows up.[49] Nonetheless, the precise definition of a "real" singularity remained unclear until the 1960s. One almost paradoxical feature of the problem is that singularities exclude smoothness (of any relevant geometric object, like geodesics or curvature), whereas (M,g) is smooth by definition, so that potential points where g is zero or some curvature invariant is infinite are excluded from space-time! This marks a decisive difference with say singularities in the electro-magnetic field (or any other field except gravity), which are definable on a given space-time background. Furthermore, there may be singular situations with regular curvature (as in the case of gravitational shock waves).

Whatever the precise definition of a singularity, until the 1960s it was also quite unclear whether "real" singularities were generic or exceptional (in the sense of only occurring in very special solutions with a high degree of symmetry, and hence being absent in "realistic" solutions); for example, Einstein maintained the latter. This was all sorted out by Hawking and Penrose (and a few others) in the period 1965–1970; the subject was essentially closed with the appearance of the book by Hawking & Ellis in 1973. The upshot is that, roughly speaking, *a space-time is deemed* **singular** *iff it contains an incomplete causal geodesic.*

[46] This motivation is merely heuristic and hence we run ahead of our later rigorous discussion of these solutions.

[47] See J. Earman, *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Space-times* (OUP, 1995).

[48] This is not to say that nothing interesting happens at $r = 2m$. Indeed, in this region t becomes spacelike and r becomes timelike, and the hypersurface $r = 2m$ is an *event horizon*, a concept to be defined later.

[49] This time the singularity is detected by the strange scalar $R^{\rho\sigma\mu\nu}R_{\rho\sigma\mu\nu}$, which goes like r^{-6} as $r \to 0$.

Definition 7 *A (smooth or continuous) curve $c : [a,b) \to M$ (where $b \in \mathbb{R}$, $b > a$) is* **(future) extendible** *if it has a (smooth or continuous) extension $c : [a,b] \to M$, and is* **(future) inextendible** *otherwise.*[50] *Equivalently, defining an* **endpoint** *of $c : I \to M$ as a point $z \in M$ such that for any nbhd U of z there is $t \in I$ such that $c(s) \in U$ for all $s \geq t$, c is future (in)extendible iff it has (no) endpoint.*[51] **Past (in)extendibility** *for curves $c : (a,b] \to M$ is defined analogously. A curve is* **incomplete** *if it is inextendible and has finite arc length, and* **complete** *otherwise.*[52]

In the Riemannian case, the ***Hopf-Rinow Theorem*** (already mentioned just before §2.4) states that a Riemannian manifold (M,g) is geodesically complete iff it is complete in the metric d derived from g. In this theorem, using the terminology of the above definition we may obviously replace geodesic completeness by the condition that every geodesic with finite arc length has an endpoint (and hence can be extended). Contrapositively, (M,d) is (metrically) incomplete iff there is at least one geodesic with finite arc length but no endpoint. For example, for $M = \mathbb{R}$ (with flat metric) the geodesic $c : [0,1) \to M$ defined by $c(t) = t$ is extendible and has endpoint $z = 1$, but the same curve is inextendible in $M = \mathbb{R}\backslash\{1\}$, where indeed it has no endpoint. Thus $\mathbb{R}$ is (metrically = geodesically) complete, whereas $\mathbb{R}\backslash\{1\}$ is not. Of course, though illustrative, this is a somewhat trivial case, since we may simply add the point $z = 1$ to the latter space.

To exclude such trivial cases, we extend the previous definition as follows.

Definition 8 *A Lorentzian manifold (M,g) is* **extendible** *if there exist a Lorentzian manifold (M',g') and an isometric embedding $i : M \hookrightarrow M'$ (so that $i^*g' = g$), and* **inextendible** *if this is not the case. It is* **incomplete** *if it contains an incomplete geodesic, and* **singular** *if it is incomplete and either inextendible, or has no extension in which all its incomplete geodesics extend to complete ones (i.e. in any extension at least one geodesic remains incomplete).*[53]

> 'Timelike geodesic completeness has an immediate physical significance in that it present the possibility that there could be freely moving observers or particles whose histories did not exist after (or before) a finite interval of proper time. This would appear to be an even more objectionable feature than infinite curvature and so it seems appropriate to regard such a space as singular. (...) We shall therefore adopt the view that *timelike and null geodesic completeness are minmum conditions for space-time to be considered singularity-free.* Therefore, if a space-time is timelike or null geodesically incomplete, we shall say that it has a singularity.' (Hawking & Ellis, p. 258).

So the example just given of a one-dimensional manifold (with flat metric) with a point removed is incomplete but non-singular. The FRW universe, on the other hand, is really singular, since it has past-directed timelike geodesics ending at $t = 0$ (we cannot prove this rigorously now, but it turns out that this space-time is inextendible because the curvature blows up as $t \to 0$).

In line with this quotation (which departs from the more accurate Definition 8), the Hawking–Penrose singularity theorems and related results "merely" prove the existence of incomplete (timelike or null) geodesics; inextendibility of space-time has to be established separately.

[50] A curve $c : [a,b] \to M$ is always extendible to $c : [a,b+\varepsilon)$, for some $\varepsilon > 0$

[51] Ir is easy to show that c has an endpoint iff $c(I)$ lies in a compact subset of M (O'Neill, Lemma 1.56).

[52] So for $b = \infty$ a geodesic $c : [a,\infty) \to M$ is always complete.

[53] If one replaces "Lorentzian manifold" in these definitions by "space-time", one assumes that the data satisfy the Einstein equations. We will return to this matter.

4.1 Global hyperbolicity and existence of geodesics of maximal length

The singularity theorems à la Hawking and Penrose are proved by the following strategy:

1. Physical arguments (from either the expansion of the universe or the clustering of matter) justify the curvature condition (3.134) and the condition (3.135) on the expansion θ.
2. These imply the existence of conjugate points on timelike geodesics (cf. Proposition 6).
3. According to Theorem 5, geodesics with conjugate points cannot maximize length.
4. Global assumptions imply that such geodesics *do* maximize length (cf. Proposition 12).
5. This apparent contradiction is resolved by realizing that the existence of conjugate points is based on the assumptions that all geodesics in question can be extended at least to the first conjugate point, so that the real conclusion is geodesic incompleteness of space-time.

In particular, against the expectations of Einstein himself singularities turn out to be *generic*. The simplest result in this direction is Hawking's singularity theorem from his 1966 PhD Thesis.

Clause 4 (to be addressed in the next section) and clause 1 are both based on the physically reasonable assumption of *global hyperbolicty*, which will also play a key role in the discussion of the Cauchy problem for the Einstein equations. For convenience, from now on we say:

Definition 9 *A* **space-time** *is an oriented and time-oriented connected Lorentzian manifold.*

Let (M,g) be a space-time. We return to the relation $\ll$, which was defined by $x \ll y$ if there exists a future-directed timelike curve (or geodesic) starting at x and ending at y. Further to the sets $I^{\pm}(x)$ as defined in (3.78) - (3.79), we more generally define

$$I^+(A) = \cup_{x\in A} I^+(x) = \{y \in M \mid \exists x \in A : x \ll y\} \qquad (A \subset M); \tag{4.3}$$

$$I^-(A) = \cup_{x\in A} I^-(x) = \{y \in M \mid \exists x \in A : y \ll x\} \qquad (A \subset M). \tag{4.4}$$

Here the first equality signs are definitions, and establishing the second ones are exercises. Similarly, one defines $J^{\pm}(A)$, where $\ll$ is replaced by $<$, where $x < y$ iff there exists a future-directed *causal* curve (or geodesic) starting at x and ending at y. Transitivity of $\ll$ then gives

$$I^+(I^+(A)) = I^+(A), \tag{4.5}$$

and it easy to show that for any $A \subset M$ the sets $I^{\pm}(A)$ are open (see O'Neill, Lemma 14.3).

In Minkowski space-time, $I^+(x)$ is the open set enclosed by the future light-cone emanating from x, $J^+(x)$ is its closure, and $J^+(x) - I^+(x)$ is its boundary, that is,

$$I^+(x) = \{y \in \mathbb{R}^4 \mid (y^0 - x^0)^2 - \sum_{i=1}^{3} (y^i - x^i)^2 > 0, y^0 > x^0\}; \tag{4.6}$$

$$J^+(x) = \{y \in \mathbb{R}^4 \mid (y^0 - x^0)^2 - \sum_{i=1}^{3} (y^i - x^i)^2 \geq 0, y^0 \geq x^0\}; \tag{4.7}$$

$$\partial I^+(x) = \{y \in \mathbb{R}^4 \mid (y^0 - x^0)^2 - \sum_{i=1}^{3} (y^i - x^i)^2 = 0, y^0 \geq x^0\}. \tag{4.8}$$

Furthermore, for $x \ll y$ the so-called ***double cone*** or ***diamond***

$$J(x,y) \equiv J^+(x) \cap J^-(y) \tag{4.9}$$

is compact in Minkowski space-time.[54] However, these properties need not be true in arbitrary space-times: simply removing a point shows that neither $J^{\pm}(x)$ nor $J(x,y)$ need to be closed (let alone compact). Properties that do prevail in general include:[55]

- If $x \ll y$ and $c : [a,b] \to M$ is a causal curve from x to y such that $c([a,b]) \cap I^+(x) = \emptyset$, then c is a null curve.
- $I^+(x)$ is the interior of $J^+(x)$. Equivalently: the boundary $\partial I^+(x)$ is a null surface.[56]
- $J^+(x) \subseteq \overline{I^+(x)}$, with equality iff $J^+(x)$ is closed.

We saw that the relation $\ll$ is transitive. Is it also anti-symmetric? By convention, we do *not* allow curves of zero length, that is, one has $x \ll x$ iff there is a genuine closed timelike curve from x to x (as in Gödel's solution to the Einstein equations, or in the Taub–NUT solution).[57] Hence $\ll$ is an order relation iff (M,g) contains no closed *timelike* curves, in which case we say that the space-time satisfies the ***chronology condition***. However, both the singularity theorems and well-posedness of the Cauchy problem for the Einstein equations require stronger conditions.

Definition 10 *A space-time (M,g) is called:*

1. **causal** *if it contains no closed causal curves.*
2. **strongly causal** *if any nbhd U_x of any $x \in M$ contains $V_x \in \mathscr{O}(M)$ such that any timelike (or, equivalently, causal) curve with endpoints in V_x entirely lies in V_x.*[58]
3. **globally hyperbolic** *if it is (strongly) causal and all sets $J(x,y)$ are compact.*[59]

The idea of strong causality is that there aren't even any timelike curves that start and end arbitrarily closely *near* x, which is something like a chronology condition stabilized against perturbations. If $J(x,y)$ fails to be compact, there is an incomplete causal curve emanating at x that disappears into some singularity. This curve lies in the past of y and hence is "visible" from y (which is deemed undesirable). We will need two implications (which are even equivalent definitions) of the notion of global hyperbolicity, namely, with details in the next two sections:

- Compactness of the space $C(x,y)$ of continuous fd causal curves from x to y ($x \ll y$);
- Existence of a Cauchy surface in M.

[54] $J(x,y)$ is the smallest subset of M that contains all fd causal curves from x to y ($J(x,y) = \emptyset$ unless $x \leq y$).

[55] See R. Geroch & G. Horowitz, Global structure of spacetimes, *General Relatvity: An Einstein Centenary Survey*, eds. S.W. Hawking & W. Israel, pp. 212–293 (CUP, 1979), for the simple proofs. This article is a very useful introduction to the ideas discussed in this chapter, to which Geroch himself made decisive contributions.

[56] This means that for any $y \in \partial I^+(x)$ there is a null geodesic γ emanating from y that lies in $\partial I^+(x)$.

[57] Or, for a very simple example: take the ***Minkowski hypercylinder*** $M = \{(x^0,\vec{x}) \in \mathbb{R}^4 \mid 0 \leq x^0 \leq 1\}/\sim$, where $(0,\vec{x}) \sim (1,\vec{x})$, with induced Minkowski metric. Then $I^+(x) = I^-(x) = M$ for all $x \in M$.

[58] Equivalently, for any compact $K \subset M$ any causal curve $c : (a,b) \to K$ can be extended to $c : [a,b] \to K$, where the case $a = -\infty$ and/or $b = \infty$ is included by asking that $\lim_{t \to \pm\infty} c(t) \in K$ also. See e.g. Wald, Lemma 8.2.1.

[59] Here it does not matter if we impose the condition for all $x,y \in M$, all $x \ll y$, or all $x < y$. It can be shown that if all sets $J(x,y)$ are compact, then causality and strong causality are equivalent, see Minguzzi & Sanchez, `arxiv:gr-qc/0609119` or Bernal & Sanchez, `arxiv:gr-qc/0611138`.

4.2 Existence of geodesics of maximal length

We now complete our journey from local to global properties of geodesics by proving the existence of fd timelike geodesics of maximal length. For $x \ll y$, let $C(x,y)$ be the space of *continuous* fd causal curves c from x to y, up to reparametrization (i.e. one uses the *image* $c([a,b])$ in M rather than the *function* $c : [a,b] \to M$),[60] topologized by letting any open nbhd of $c \in C(x,y)$ consist of all fd causal curves γ whose image lies in some open nbhd of $c([a,b])$ in M.[61]

Proposition 11 *A time-oriented Lorentzian manifold (M,g) is globally hyperbolic iff the space $C(x,y)$ of continuous fd causal curves from x to y is compact for all $x \ll y$.*[62]

This is a pretty difficult result, so we just sketch the outline of the proof.[63] If $C(x,y)$ is compact, then so is $J(x,y)$; this follows from the continuity of the evaluation map

$$\mathrm{ev} : C(x,y) \times [0,1] \to M; \tag{4.10}$$

$$\mathrm{ev}(c,t) = c(t). \tag{4.11}$$

The argument below will show that (M,g) is strongly causal. For the converse implication, we need to turn M into a metric space inducing the manifold topology, which cannot come from the Lorentzian metric g, but, in a singularly ugly move, comes from an ancillary Riemannian metric h, see §2.3.2. The Arzelà–Ascoli Theorem then states that $C(x,y)$ is compact iff:[64]

1. Each set $\{c(t) \mid c \in C(x,y)\} \subset M$, $t \in (0,1)$, is bounded (and hence has compact closure);
2. The family $C(x,y)$ is equicontinuous, i.e., for each $t \in [0,1]$ and each $\varepsilon > 0$ there is $\delta > 0$ such that if $|s-t| < \delta$, then $d(c(s),c(t)) < \varepsilon$ for all $c \in C(x,y)$.

It is easy to show that both conditions are satisfied iff there is some $0 < K_{x,y} < \infty$ such that

$$L_h(c) < K_{x,y} \text{ for all } c \in C(x,y), \tag{4.12}$$

where L_h is the length computed from h (and similarly, in the next step d_h will be the distance computed from h). Indeed, if this is the case, then

$$d_h(x,c(t)) \leq L_h(c) < K_{x,y}, \tag{4.13}$$

which makes the set $\{c(t) \mid c \in C(x,y)\}$ in clause 1 of the Arzelà–Ascoli Theorem bounded. Assuming c is parametrized by arc length, we have

$$L_h(c(s,t)) = L_h(c)|s-t|, \tag{4.14}$$

and hence

$$d_h(c(s),c(t)) \leq L_h(c)|s-t| < K_{x,y}. \tag{4.15}$$

This proves equicontinuity. We now finish the proof with a few observations:

[60] Since the previous definition of fd causality of a *smooth* (or C^1) curve relies on its tangent vectors, one calls a *continuous* curve $c : [a,b] \to M$ fd causal if for any $t \in (a,b)$ with normal nbhd $U_{c(t)}$ (cf. §3.4), and any $t' < t''$ such that $c([t',t'']) \subset U_x$ one has $c(t') < c(t'')$, i.e., there exists a smooth fd causal curve from $c(t')$ to $c(t'')$. This in fact implies that c is locally Lipschitz and hence C^1 almost everywhere (of course with an fd cause tangent vector).

[61] This is the quotient of the compact-open topology on $C([0,1],M)$ by the equivalence relation given by reparametrization, restricted to those timelike curves c in $C([0,1],M)$ that satisfy $c(0) = x$ and $c(1) = y$, see C.J.S Clarke, *The Analysis of Space-Time Singularities* (CUP, 1993), §6.2.2.

[62] This was Leray's original definition of global hyperbolicity in 1952, cf. Hawking & Ellis, §6.6).

[63] See Hawking & Ellis, P rop. 6.6.2, or Choquet-Bruhat, Theorem XII.10.2, fo complete proofs.

[64] Since the parametrization of $c \in C(x,y)$ does not matter, we put $c : [0,1] \to M$.

1. Contrapositively, the non-existence of a uniform bound $K_{x,y}$ for $L_h(c)$ blasts both clauses 1 and 2.

2. Eq. (4.12), in turn, is guaranteed when $J(x,y)$ is compact (in which case it may be covered with finitely many open sets of the kind U_x) and M is strongly causal (which prevents causal curves from re-entering U_x arbitrarily often, a possibility that would indefinitely increase the length of a curves and hence drive $K_{x,y}$ to infinity).

3. The last comment also proves the necessity of strong causality for compactness of $C(x,y)$.

This completes the sketch of the proof of Proposition 11.

The following result is crucial for the Hawking–Penrose singularity theorems:

Proposition 12 *If (M,g) is globally hyperbolic, then any $x \in M$ and $y \in I^+(x)$ are connected by a smooth fd timelike geodesic of maximal length (among all curves from x to y).*

The full proof takes pages to develop in detail,[65] so we just give an outline.[66] Through approximation of continuous curves by smooth ones, the length functional $c \mapsto L(c)$ defined by (3.84) makes sense on $C(x,y)$, and is *upper* semicontinuous.[67] Via compactness of $C(x,y)$, global hyperbolicity then implies that L assumes its maximum

$$\ell(x,y) = \sup\{L(\gamma) \mid \gamma\colon [0,1] \to M, \gamma(0) = x, \gamma(1) = y, \gamma \text{ fd timelike curve}\}, \tag{4.16}$$

at some curve $\gamma \in C(x,y)$. This *a priori* merely continuous curve is in fact a (smooth) geodesic.[68] Moreover, the maximum geodesic may be found as follows: if (c_n) is a sequence of curves in $C(x,y)$ for which $L(c_n) \to \ell(x,y)$, then $c_n \to \gamma$.[69]

[65] See Penrose, *passim*, Hawking & Ellis, §6.7, O'Neill, Chapter 14, Senovilla, Ch. 2.

[66] There is also a second proof, which starts from the pointwise length function $\ell : M \times M \to [0,\infty]$ defined by (4.16), where we put $\ell(x,y) = 0$ except for $x \ll y$. If ℓ is finite, then it is *lower* semicontinuous. Indeed, by definition of the supremum, for any $\varepsilon > 0$ there exists some $c_{xy} : x \to y$ for which $L(c_{xy}) = \ell(x,y) - \frac{1}{3}\varepsilon$. Furthermore, by §3.5.3 we can find a nbhd U_y of (say) y such that for any $z \in U$ there exists a curve $c_{yz} : y \to z$ with length $L(c_{yz}) < \frac{1}{3}\varepsilon$. Violation of lower semicontinuity of ℓ at y would mean that $L(c_{xz}) < \ell(x,y) - \varepsilon$ for all $z \in U$ and all curves $c_{xz} : x \to z$, but in fact the concatenation of c_{xy} and c_{yz}, which has length $L(c_{xy}) + L(c_{yz})$, can be smoothened so as to have length within $\frac{1}{3}\varepsilon$ of the latter, which leads to a contradiction. If M is globally hyperbolic, then ℓ is also finite and continuous: since $J(x,y)$ is compact, the cover (V_z), where $z \in J(x,y)$, has a finite subcover (V_{z_i}). This makes $\ell(x,y)$ finite, since each fd timelike curve from x to y can enter each (V_{z_i}) at most once, and its segment within (V_{z_i}) has finite length. Compactness of $J(x,y)$ also yields upper semicontinuity of $\ell(x,y)$ (see Hawking & Ellis, p. 216). For each $z \in J(x,y)$, define $W_z = (U_z \cap J(x,y))^-$, which is a closed and hence compact subset of $J(x,y)$, and consider the function $z \mapsto \ell(x,z) + \ell(z,y)$ on W_z. This function is continuous and hence takes a maximum at say z_1, to which (by §3.5) there is a unique fd timelike geodesic γ_{xz_1} from x to z_1. Restarting this construction from z_1 and repeating the process extends this geodesic, and a proof by contradiction to compactness (cf. Hawking & Ellis, p. 216–217) shows that the ensuing geodesic eventually reaches y and maximizes $\ell(x,y)$.

[67] That is, for each $c \in C(x,y)$ and each $\varepsilon > 0$ there is a nbhd Γ of c such that $L(\gamma) \leq L(c) + \varepsilon$ for all $\gamma \in \Gamma$. See Hawking & Ellis, Lemma 6.7.2 or Wald, Prop. 9.4.1. *Increasing* the length of a fd timelike curve c can only be done by adding fd timelike pieces, which can be done only in a limited way in a small nbhd V of c. *Decreasing* its length, on the other hand, can be done at will even within V by moving c close to a chain of almost null directions (see footnote 36). Hence lower semicontinuity (i.e., $L(\gamma) \geq L(c) - \varepsilon$ for all $\gamma \in \Gamma$) typically fails.

[68] For the proof that a continuous timelike curve of maximal length must be a geodesic see O'Neill, Proposition 14.19 (note that Hawking & Ellis, p. 215, only arrive at C^1 geodesics, as does Choquet-Bruhat, Theorem XII.9.5). Roughly, if between any two of its points it would not be a geodesic, then, if necessary chopping the non-geodesic part up into smaller pieces, we could construct a nearby geodesic whose length would be longer, see §3.5.3.

[69] This follows from an extension of Weierstrass's theorem from topology: if K is compact, then an upper semicontinuous function $f : K \to \mathbb{R}$ has a maximum (and a lower semicontinuous function $g : K \to \mathbb{R}$ has a minimum). In the present context see also O'Neill, Lemma 14.14 and Proposition14.19.

4.3 Global hyperbolicity and Cauchy surfaces

Both the singularity theorems and the solution of the Cauchy problem for the Einstein equations rely on the fundamental notion of a *Cauchy surface*, defined as follows.

Definition 13 *A* **Cauchy surface** *in a space-time (M,g) s a subset $\Sigma \subset M$ with the property that each inextendible timelike curve intersects Σ in exactly one point.*

There are various easy consequences of this definition, which we will not need and hence will not prove, but they do clarify the idea and hence we state some of them for completeness.[70]

Proposition 14 *Let (M,g) be a space-time with Cauchy surface $\Sigma \subset M$. Then:*

1. *Any other possible Cauchy surface in M is diffeomorphic to Σ;*
2. *Σ is a three-dimensional embedded spacelike submanifold of M;*[71]
3. *Σ is* **achronal** *in the sense that for all $x,y \in \Sigma$ it cannot be that $x \ll y$;*
4. *Every causal curve meets Σ (though not necessarily in one point);*
5. *Defining the* **domain of dependence** *(or* **Cauchy development***) $D^+(S)$ of a subset $S \subset M$ as the set of all points $y \in M$ for which every past-directed timelike curve (or geodesic) emanating from y intersects S, and similarly, the* **domain of influence** *$D^-(S)$ by changing past-directed to future-directed in the previous definition, one has*

$$D(\Sigma) \equiv D^+(\Sigma) \cup D^-(\Sigma) = M, \tag{4.17}$$

   which is necessary and sufficient for a closed achronal set $\Sigma \subset M$ to be Cauchy surface.

6. *Defining the* **future/past Cauchy horizon** *$H^+(S)/H^-(S)$ of any subset $S \subset M$ by*

$$H^+(S) = D^+(S) - I^-(D^+(S)); \tag{4.18}$$
$$H^-(S) = D^-(S) - I^+(D^-(S)), \tag{4.19}$$

   so that e.g. $H^+(S)$ consist of all $x \in D^+(S)$ that precede no point in $D^+(S)$,[72] *we have*

$$H^+(\Sigma) = H^-(\Sigma) = \emptyset, \tag{4.20}$$

   and this condition equally well holds iff a closed achronal set $\Sigma \subset M$ is a Cauchy surface.

[70] See e.g. Hawking & Eliis, Ch. 6, O'Neill, Ch. 14, Wald, Ch. 8, or Minguzzi & Sanchez. The entire theory was initiated by Geroch and others in the late 1960s in the topological case, and was extended to the smooth case by Bernal and Sanchez between 2003–2005. For the smooth results see the review *Recent progress on the notion of global hyperbolicity*, `arXiv:gr-qc/0712.1933`, based on the following three papers by Bernal and Sanchez:
On smooth Cauchy hypersurfaces and Geroch's splitting theorem, `arXiv:gr-qc/0306108`,
Smoothness of time-functions and the metric splitting of globally hyperbolic spacetimes, `arXiv:gr-qc/0401112`,
Further results on the smoothability of Cauchy hypersurfaces and Cauchy time functions, `arXiv:gr-qc/0512095`.

[71] The notion of an embedded submanifold will be given in §6.

[72] A point lying beyond the future Cauchy horizon of S will be influenced by events outside S, and so $H^+(S)$ measures the failure of S to be a Cauchy surface.

The idea is that everything happening at $y \in D^+(S)$ is determined by the state of affairs at S. For example, in $2d$ Minkowski space-time $\mathbb{R}^2$, for $S = \mathbb{R}$, the x-axis, $D^+(S)$ is the upper half plane and $D^-(S)$ is the lower half-plane, so that S is a Cauchy surface. Removing just $(0,0)$ from the x-axis removes the interior of the forward light cone emanating from $(0,0)$ from the earlier $D^+(S)$. Taking $S_1 = \{0\} \times [-1,1]$, it follows that $D^+(S)$ consists of the triangle with vertices $(-1,0)$, $(1,0)$, and $(0,1)$, the associated Cauchy horizon $H^+(S)$ consists of the two upper sides of this triangle. Removing $(0,0)$ from S removes the double cone with vertices $(0,0)$, $(-\frac{1}{2},\frac{1}{2})$, $(0,1)$, and $(\frac{1}{2},\frac{1}{2})$ from $D^+(S)$, whereas $H^+(S)$ suddenly consists of two zig-zag teeth (draw!).

The following result (often used as the definition of global hyperbolicity!) is very deep:

Theorem 15 *A space-time (M,g) is globally hyperbolic iff it has a Cauchy surface Σ.*

The proof is based on the construction of a time-function $\mathtt{t} : M \to \mathbb{R}$, see §3.7.2. If c is any timelike curve, then $g(\nabla \mathtt{t}, \dot{c}) = \dot{c}(\mathtt{t})$, whose left-hand side is non-zero. Hence $\mathtt{t}$ either increases or decreases along fd timelike curves, and (if necessary changing its sign) we assume $\mathtt{t}$ *increases*.

To construct $\mathtt{t}$, we once again take some auxiliary Riemannian metric h on M, as well as some at most countable open cover (V_n) with precompact elements (i.e. V_n^- is compact for each n), so that $M = \cup_n V_n$, with some associated partition of unity (ϕ_n) subordinate to the cover.[73]

We then turn the standard Riemannian measure μ_h induced by h, defined in coordinates by

$$d\mu_h(x) = \sqrt{h(x)}dx^0 \cdots dx^3, \tag{4.21}$$

see also the next section, into a probability measure $\nu_h = \chi\mu_h$, where $\chi : M \to \mathbb{R}^+$ is defined by

$$\chi = \sum_n 2^{-n} \frac{\phi_n}{\int_{V_n} d\mu_n \phi_n}. \tag{4.22}$$

The define functions $\omega^\pm : M \to \mathbb{R}^+$ by $\omega^\pm(x) = \nu_h(J^\pm(x))$, in terms of which

$$\mathtt{t}(x) = \ln\left(\frac{\omega^-(x)}{\omega^+(x)}\right). \tag{4.23}$$

Fairly technical arguments then show that:

1. $\mathtt{t}$ is continuous because $J^\pm$ are closed (which follows from global hyperbolicity);
2. $\mathtt{t}$ is strictly increasing along fd timelike curves (*idem*);
3. Each level set
$$\Sigma_t = \{x \in M \mid \mathtt{t}(x) = t\} \tag{4.24}$$
is a Cauchy surface.

Corollary 16 *For a globally hyperbolic space-time (M,g) with Cauchy surface Σ we have*

$$M \cong \mathbb{R} \times \Sigma; \tag{4.25}$$

$$M = \cup_{t\in\mathbb{R}} \Sigma_t. \tag{4.26}$$

More specifically: M is diffeomorphic to $\mathbb{R} \times \Sigma$ in such a way that under the pertinent diffeomorphism each subset $\{t\} \times \Sigma \subset \mathbb{R} \times \Sigma$ corresponds to a Cauchy surface $\Sigma_t \subset M$.

Note that (4.25) is quite intuitive: since each inextendible fd timelike curve hits Σ exactly once, we may take such a curve c, normalize the time-function $\mathtt{t}(x)$ such that $\Sigma = \Sigma_0$, and define a map $M \to \mathbb{R} \times \Sigma$ by $x \mapsto (t,\sigma)$, where $t = \mathtt{t}(x)$ and $\sigma \in \Sigma$ is the point where c hits Σ.

Given the definition (4.24) of Σ_t, eq. (4.26) is almost a triviality.

[73] This means that $\varphi_n \in C_c^\infty(V_n)$ and $\sum_n \varphi_n(x) = 1$ for all $x \in M$.

4.4 Hawking's singularity theorem

In his honor, we now discuss Hawking's singularity theorem from 1965/66, which regarding both its assumptions and the design of its proof remains a model for all subsequent results. Its underlying intuition comes from the FRW cosmologies, and so it describes Big Bang type singularities (black hole type singularities are covered by Penrose's singularity theorems, which unfortunately we cannot discuss at this stage because of their heavy reliance on null geodesics).

Let (M,g) be a globally hyperbolic space-time with Cauchy surface Σ, and recall the situation described in §3.7. We consider a congruence of timelike geodesics (γ) emanating from Σ,[74] with initial velocities (i.e. tangent vectors) $\dot\gamma = u$ orthogonal to Σ; this is called the ***normal geodesic congruence*** emanating from Σ. Cosmological applications require these to be past directed (pd), but if our universe is ever going to approach a big crunch,[75] the same construction works for future-directed geodesics.[76] Thus we also obtain the quantities defined in (3.121) - (3.126), especially the latter (i.e. the expansion $\theta = \nabla_\mu u^\mu = \mathrm{tr}(k)$ of the congruence).

To understand what follows, one more geometric construction is needed, which will be studied in great detail in our discussion of the Cauchy problem to come, namely the ***extrinsic curvature*** of $\Sigma \subset M$. This is a tensor field $K \in \mathfrak{X}^{(2,0)}(\Sigma)$ initially defined merely on Σ by

$$K(X,Y) = -g(\nabla_X N, Y), \tag{4.27}$$

where $X,Y \in \mathfrak{X}(\Sigma)$ and N is the normal vector field on Σ (whose sign is a matter of convention). This definition is predicated on the fact that $\nabla_X N$ is tangent to Σ, which is an easy consequence of the property $g(N,N) = -1$. Similarly, it is easy to show that K is symmetric, namely:

$$k(X,Y) = -g(\nabla_X N, Y) = g(N, \nabla_X Y) = g(N, \nabla_Y X) = k(Y,X). \tag{4.28}$$

From K, we define the ***mean (extrinsic) curvature*** $H : \Sigma \to \mathbb{R}$ of Σ as

$$H(x) = \mathrm{tr}(K_x) = \sum_{i=1}^{2} K_x(e_i(x), e_i(x)), \tag{4.29}$$

where $(e_i(x))$ is any orthogonal basis of $T_x\Sigma$, $x \in \Sigma$. Since k in (3.123) is spatial (because of the projections h in its definition), because of the (conventional) minus sign in (4.27), on Σ we have

$$\theta = -H. \tag{4.30}$$

Let us give some Riemannian examples, which can be found in almost all pertinent textbooks:

- For the sphere of radius a, i.e., $S^2_a \subset \mathbb{R}^3$ (with flat metric), we have $H = -2/a$.
- For the cylinder of radius a, i.e., $C^2_a \subset \mathbb{R}^3$ (with flat metric), we have $H = -1/a$.
- For any plane in $\mathbb{R}^3$ (with flat metric) we have $H = 0$.

Here the normal vectors used in the definition (4.27) are *outward*, and we see from these examples that negative K, and hence *positive* θ, gives diverging geodesics normally emanating from Σ. By the same token, *negative* θ gives *converging* normal geodesics; as stated, we assume this in the *past* direction. After this preparation we are in a position to state Hawking's theorem.

[74] Or some relatively open subset thereof–the congruence may be defined *locally*, but even so *global* hyperbolicity is needed to prove existence of geodesics with maximum length.

[75] Current observations show that this will not happen, since at the moment the expansion is even accelerating.

[76] Regarding (4.26), note that a normal congruence at $\Sigma = \Sigma_0$ may no longer be orthogonal to other level sets Σ_t.

Theorem 17 *Let a space-time (M,g) be globally hyperbolic with Cauchy surface Σ. Assume:*

1. *The curvature satisfies $R_{\mu\nu}\dot\gamma^\mu\dot\gamma^\nu \geq 0$ along all timelike geodesics γ;*

2. *The mean extrinsic curvature of Σ is (uniformly) positive in the* past *direction.*

If $H > H_0 > 0$ in clause 2, then no past directed timelike geodesic emanating from Σ can have (arc) length greater than $3/K_0$, and hence (M,g) has incomplete geodesics.

It is sufficient to prove this for timelike geodesic *normally* emanating from Σ, since other timelike geodesics are even shorter (exercise). The proof is by contradiction:

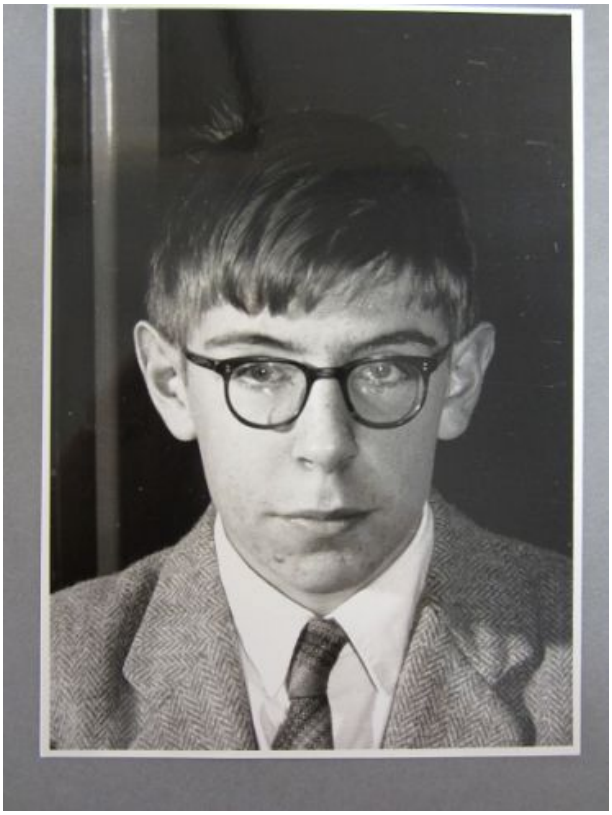
1. If there is such a geodesic, say from $x \in \Sigma$ to $y \ll x$, then by Proposition 12 (with past and future exchanged) there is one of maximal length, call it γ (this step uses global hyperbolicity).

2. By Theorem 5, γ cannot have conjugate points.

3. By Proposition 6, however, γ does have conjugate points (this step uses the assumptions on the curvature $R_{\mu\nu}$ of (M,g) and on the extrinsic curvature of $\Sigma \subset M$), provided γ can be extended far enough.

4. Hence the geodesic in question cannot exist, and the conclusion follows.

Note that the time to reach the singularity *increases* as the mean extrinsic curvature Σ decreases, in accordance with intuition: less curvature means less focusing.



Stephen Hawking in 1960

5 The Einstein equations

On Thursday November 25, 1915, Albert Einstein wrote down the immortal equations[77]

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi T_{\mu\nu}, \tag{5.1}$$

whose left-hand side we have already seen, and whose right-hand side will be explained in §5.2. The ***Einstein equations*** (5.1) are widely considered to be the most beautiful equations in all of physics (or perhaps even all of science), and they are certainly the most accurately tested ones. The relate the geometry of space-time, construed as a Lorentzian manifold (M,g), to its matter content, given by $T_{\mu\nu}$; as Misner, Thorne, and Wheeler (1973) put it: "matter tells space how to curve" (followed by: "and space tells matter how to move," namely on geodesics).[78] In this sense, (5.1) are 10 coupled second-order partial differential equations for the components $g_{\mu\nu}$ of the metric *given* $T_{\mu\nu}$, but in fact there will be additional equations for the matter fields contained in $T_{\mu\nu}$, which also depend on the metric, and one should really consider the total system.

5.1 The Hilbert action

As noticed independently by Hilbert and Einstein in 1916, the Einstein equations (5.1) can be derived from a variational principle. The geometrical quantity to be extremized in order to obtain the left-hand side is the ***(Einstein-) Hilbert action*** for the gravitational field, defined by

$$S_G(g) = \int_M d^4x \sqrt{-g(x)} R(x), \tag{5.2}$$

where $g \equiv \det(g)$ is the determinant of the matrix $g_{\mu\nu}$ (in any basis), and R is the Ricci scalar

$$R = g^{\mu\nu} R_{\mu\nu}, \tag{5.3}$$

cf. (2.10) and (2.9). More precisely, we assume M is orientable, and (5.2) should either be written as a sum over various coordinate patches using a partition of unity, or else in a geometric form (for which we have hardly developed the machinery).[79] As in the geodesic case, we now consider a family of metrics g_s, and compute $dS_G(g_s)/ds$. This requires some preparation.

[77] In fact, Einstein used a somewhat different notation; what he literally wrote was $G_{im} = -\kappa(T_{im} - \frac{1}{2} g_{im} T)$. For the moment we omit the infamous ***cosmological constant*** Λ, which Einstein added in 1917 on the left-hand side through a term $\Lambda g_{\mu\nu}$ in order to stabilize the Universe, but after he recognized the expansion of the Universe he withdrew it and called his introduction "the biggest blunder of his life" (his real blunder, though, was missing the theoretical derivation of an expanding Universe from (5.1), which was left to Friedman and Lemaître.). The cosmological constant made a spectacular come-back at the end of the 20th Century, when it was discovered that the Universe expands more rapidly than could be explained by (5.1) with known forms of matter. It has become customary to move the term $\Lambda g_{\mu\nu}$ to the right-hand side and regard it as an unknown contribution to $T_{\mu\nu}$, called ***dark energy***, which is estimated to comprise as much as 70% of the total energy of the Universe! See R.P. Kirshner, *The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos* (Princeton University Press, 2002) for the inside story of this discovery and its history.

[78] The second part is a consequence of (5.1), but the proof is tricky and we will show this only for a fluid.

[79] A manifold is called ***orientable*** if there is an atlas (within the equivalence class of atlases defining the manifold, cf. §1.1) for which all transition functions $\varphi_\beta \circ \varphi_\alpha^{-1}$ have positive Jacobian. An ***orientation*** of an orientable manifold is an atlas satisfying this condition. It can be shown that M is orientable iff it admits a nowhere vanishing n-form $\omega \in \Omega^n(M)$; one then only accepts charts φ whose coordinates $(x^1,\ldots,x^n)$ satisfy $\omega(\partial_1,\ldots,\partial_n) > 0$. In the presence of a metric we then normalize ω such that in all coordinates $\omega(\partial_1,\ldots,\partial_n) = \sqrt{|g|}$, where again $g \equiv \det(g)$, i.e., $\omega_x = \sqrt{|g(x)|}\, dx^1 \wedge \cdots \wedge dx^n$. This condition is well defined, since ω keeps this form under coordinate transformations (exercise: one has $\sqrt{|g(x_\beta)|} = J_{\alpha\beta}^{-1} \sqrt{|g(x_\alpha)|}$, where $J_{\alpha\beta} = \det|\partial x_\beta^\mu / \partial x_\alpha^\nu|$ is the Jacobian of the coordinate transformation from x_α to x_β). For any $f \in C_c^\infty(M)$ one then has $\int_M f\omega = \int_M d^n x \sqrt{|g(x)|} f(x)$. We also assume sufficient decay of the integrand in (5.2) for the integral to make sense (though not necessarily $R \in C_c^\infty(M)$).

1. In any coordinate system we have (for Lorentzian siganture, as we assume throughout)

$$\partial_\mu \sqrt{-g} = \sqrt{-g}\,\Gamma^\rho_{\mu\rho}. \tag{5.4}$$

Since the first term in (3.12) cancels the last if $\nu = \rho$, we have $\Gamma^\rho_{\mu\rho} = \frac{1}{2} g^{\rho\sigma} g_{\rho\sigma,\mu}$. Diagonalizing the symmetric invertible matrix $(g_{\rho\sigma})$, yielding nonzero eigenvalues $(\lambda_0, \ldots, \lambda_3)$ and realizing that $(g^{\rho\sigma})$ is its inverse gives

$$g^{\rho\sigma} g_{\rho\sigma,\mu} = \frac{\partial_\mu \lambda_0}{\lambda_0} + \cdots + \frac{\partial_\mu \lambda_3}{\lambda_3}. \tag{5.5}$$

But also

$$2\frac{\partial_\mu \sqrt{-g}}{\sqrt{-g}} = g^{-1}\partial_\mu g = \frac{\partial_\mu(\lambda_0 \cdots \lambda_3)}{\lambda_0 \cdots \lambda_3} = \frac{\partial_\mu \lambda_0}{\lambda_0} + \cdots + \frac{\partial_\mu \lambda_3}{\lambda_3}. \tag{5.6}$$

2. For any vector field X we define its ***divergence*** as

$$\nabla \cdot X = \nabla_\mu X^\mu. \tag{5.7}$$

Eq. (5.4) then implies

$$\sqrt{-g}\,\nabla \cdot X = \partial_\mu(\sqrt{-g} X^\mu), \tag{5.8}$$

and hence, by Stokes's Theorem (= Divergence Theorem = Gauß's Theorem),[80]

$$\int_M d^4x \sqrt{-g(x)}\,\nabla \cdot X(x) = \int_{\partial M} d^3\vec{\sigma} \cdot X, \tag{5.9}$$

where ∂M is the boundary of M (if $M = \emptyset$, then the right-hand side vanishes assuming sufficient decay of X at infinity), and $d^3\vec{\sigma}$ is the (outward) normal volume element of ∂M.

3. Each of the three terms in the integrand $\sqrt{-g}\, g^{\mu\nu} R_{\mu\nu}$ in (5.2) depends on the metric $g_{\mu\nu}$ and hence has to be varied. The variation of the Ricci tensor seems the most complicated case, but surprisingly it contributes a divergence term and hence makes no contribution to the Einstein equations (5.2). This is surprising, because definitions (3.11) and (3.18) give

$$R_{\mu\nu} = \Gamma^\rho_{\mu\nu,\rho} - \Gamma^\rho_{\mu\rho,\nu} + \Gamma^\rho_{\rho\sigma}\Gamma^\sigma_{\nu\mu} - \Gamma^\rho_{\nu\sigma}\Gamma^\sigma_{\rho\mu}, \tag{5.10}$$

whose first two terms contain second-order derivaties of $g_{\mu\nu}$. Their variation would therefore in principle be expected to give a fourth-order PDE, but this does not happen.[81]

[80] Continuing the previous footnote: eq. (5.8) takes the abstract form $\mathscr{L}_X \omega = \omega \nabla \cdot X$. ***Cartan's formula*** for the Lie derivative of exterior forms states that $\mathscr{L}_X = d i_X + i_X d$, where $X \in \mathfrak{X}(M)$, i.e., for any p-form $\alpha \in \Omega^p(M)$, $p > 0$, we have $\mathscr{L}_X \alpha = d(i_X \alpha) + i(d\alpha)$, where $d : \Omega^p(M) \to \Omega^{p+1}(M)$ is the ***exterior derivative*** (defined in coordinates by $(d\alpha)_{\mu_1 \cdots \mu_{p+1}} = \partial_{\mu_1} \alpha_{\mu_2 \cdots \mu_{p+1}}$) and $i_X : \Omega^p(M) \to \Omega^{p-1}(M)$ is the ***insertion operation***, defined in coordinates by $(i_X \alpha)_{\mu_2 \cdots \mu_p} = X^{\mu_1} \alpha_{\mu_1 \mu_2 \cdots \mu_p}$. Since $\omega \in \Omega^n(M)$ we must have $d\omega = 0$, so that Cartan's formula gives $\mathscr{L}_X \omega = d(i_X \omega)$, and hence, with the first equation in this footnote, $\omega \nabla \cdot X = d(i_X \omega)$. The abstract version of Stokes's Theorem states that $\int_M d\alpha = \int_{\partial M} \alpha$, for any $\alpha \in \Omega^n(M)$, so that $\int_M \omega \nabla \cdot X = \int_{\partial M} i_X \omega$, which is (5.9).

[81] ***Lovelock's Theorem*** states that in $d = 4$ the Einstein–Hilbert action (5.2) is the *only* possible geometric quantity giving rise to second-order PDE in the components of the metric, expect for adding a constant Λ to the Ricci scalar R, which would lead to a cosmological constant in (5.1). The proof is a rather dull kind of bookkeeping; see A. Navarro & J. Navarro, Lovelock's Theorem revisited, `https://arxiv.org/pdf/1005.2386.pdf`.

Einstein temporarily used ***unimodular coordinates***, in which $g \equiv \det(g) = -1$. In such coordinates he wrote down the Lagrangian $\mathscr{L}_E = -g^{\mu\nu}\Gamma^\rho_{\nu\sigma}\Gamma^\sigma_{\rho\mu}$, partly inspired by the Lagrangian for the free electromagnetic field $-\frac{1}{4} g^{\mu\nu} F_{\nu\sigma} F_\mu{}^\sigma$, and partly by the fact that (this times almost trivially) it gives second-order PDE. This corresponds to the fourth term in (5.10); the third vanishes if $g = 1$, cf. (5.4), and the first two terms merely bring a divergence. So Einstein had essentially the right Lagrangian already in 1913, of which $\sqrt{-g}R$ is the correct geometric form. See H.R. Brown, *Physical Relativity: Space-time structure from a dynamical perspective* (OUP, 2005), §9.2.

4. Indeed, writing $\delta F(g) = dF(g_s)/ds_{|s=0}$ and $d(g_s)_{\mu\nu}/ds_{|s=0} = \delta g_{\mu\nu}$, we claim that

$$g^{\mu\nu}\delta R_{\mu\nu} = \nabla \cdot X; \tag{5.11}$$

$$X^\mu = \nabla_\nu \delta g^{\mu\nu} - \nabla^\mu \delta g^\nu_\nu, \tag{5.12}$$

where indices are always raised and lowered with the metric $g = g_{s=0}$. However, this leads to the ambiguous notation $\delta g^{\mu\nu}$, which could mean either $(\delta g)^{\mu\nu} = g^{\mu\rho}g^{\nu\sigma}\delta g_{\rho\sigma}$, or $\delta(g^{\mu\nu}) = -g^{\mu\rho}g^{\nu\sigma}\delta g_{\rho\sigma}$, see below. To avoid this we will henceforth write $d_{\mu\nu}$ for $\delta g_{\mu\nu}$, so that the first equation above becomes $d^{\mu\nu} = g^{\mu\rho}g^{\nu\sigma}d_{\rho\sigma}$, and the second is

$$\delta g^{\mu\nu} = -g^{\mu\rho}g^{\nu\sigma}d_{\rho\sigma}, \tag{5.13}$$

which follows from $g^{\mu\nu}g_{\nu\rho} = \delta^\mu_\rho$, and hence $0 = \delta(g^{\mu\nu}g_{\nu\rho}) = (\delta g^{\mu\nu})g_{\nu\rho} + g^{\mu\nu}d_{\nu\rho}$.

The key step in the proof of (5.11) -(5.12) is the relation

$$\delta\Gamma^\rho_{\mu\nu} = \tfrac{1}{2}(\nabla_\mu d^\rho_\nu + \nabla_\nu d^\rho_\mu - \nabla^\rho d_{\mu\nu}), \tag{5.14}$$

as can be showed by a lengthy computation, but also by the following instructive trick:

(a) First note that although the coefficients $\Gamma^\rho_{\mu\nu}$ do not form the components of a tensor, their variation $\delta\Gamma^\rho_{\mu\nu}$ does. This is true far more generally: if ∇ and $\tilde\nabla$ are connections on a vector bundle E, then $(\nabla_X - \tilde\nabla_X)s$ is $C^\infty(M)$-linear in $s \in \Gamma(E)$ (unlike $\nabla_X s$ and $\tilde\nabla_X s$), since the spoiler $(Xf)s$ in the Leibniz rule (2.45) drops out of the difference. As a case in point, let ∇ be the Levi-Civita connection for a given metric g and let $\tilde\nabla$ be the one for some other metric $\tilde g$. We then have a tensor $\hat\Gamma \in \mathfrak{X}^{(2,1)}(M)$, defined by $\hat\Gamma(X,Y,\theta) = \theta(\nabla_X Y - \tilde\nabla_X Y)$, whose connection coefficients are $\Gamma^\rho_{\mu\nu} - \tilde\Gamma^\rho_{\mu\nu}$, cf. (2.32). In particular, we make take $\tilde g = g_s$, and since $\delta\Gamma^\rho_{\mu\nu}(g) = \lim_{s\to 0}(\Gamma^\rho_{\mu\nu}(g_s) - \Gamma^\rho_{\mu\nu}(g))/s$, we may conclude that the coefficients $\delta\Gamma^\rho_{\mu\nu}$ form the components of a tensor $\delta\Gamma$.

(b) Let σ and τ be tensors of the same type, say $(1,1)$ Then $\sigma = \tau$ is true iff for each $x \in M$ one has $\sigma^\nu_\mu(x) = \tau^\nu_\mu(x)$ in just *one* specific coordinate system (x^μ) defined on some nbhd U of x, which system may even depend on x. For in that case we have $\sigma_x(\partial_\mu, dx^\nu) = \tau_x(\partial_\mu, dx^\nu)$, and so, by $C^\infty(M)$-linearity of σ and τ in its arguments, $\sigma(X,\theta) = \tau(X,\theta)$, where we write $X = X^\mu\partial_\mu$ and $\theta = \theta_\nu dx^\nu$ as usual, for some $X^\mu \in C^\infty(U)$ and $\theta_\nu \in C^\infty(U)$. And similarly for tensors of any type (k,l).

(c) It therefore suffices to verify (5.14) in geodesic normal coordinates, where at $x = x_0$ we have $\nabla = \partial$, cf. (3.53). In GNC one does not even need (5.13), since $\delta g^{\rho\sigma}$ in (3.12) multiplies terms that vanish at x_0, and hence (5.14) is almost trivial.

Similarly, noting that in GNC the variation $\delta R_{\mu\nu}$ only employs the first two terms in (5.10), in which $\delta(\Gamma^\rho_{\mu\nu,\rho}) = \partial_\rho\delta\Gamma^\rho_{\mu\nu}$ (etc.) can be computed from (5.14), one obtains

$$\delta R_{\mu\nu} = \tfrac{1}{2}(\nabla_\rho\nabla_\mu d^\rho_\nu + \nabla_\rho\nabla_\nu d^\rho_\mu - \nabla_\mu\nabla_\nu d^\rho_\rho - \nabla^\rho\nabla_\rho d_{\mu\nu}), \tag{5.15}$$

where we note that the third term is symmetric in μ and ν because of (3.10) and (3.24). Contraction with $g^{\mu\nu}$ then makes the first two terms identical to each other, and similarly, the last two, and immediately leads to (5.11) - (5.12).

5. The computation of $\delta\sqrt{-g}$ is based on the relation $\partial g/\partial g_{\mu\nu} = g^{\mu\nu}g$,[82] which implies

$$\delta\sqrt{-g} = \frac{\partial\sqrt{-g}}{\partial g_{\mu\nu}}d_{\mu\nu} = -\frac{1}{2\sqrt{-g}}\frac{\partial g}{\partial g_{\mu\nu}}d_{\mu\nu} = \tfrac{1}{2}\sqrt{-g}\,g^{\mu\nu}d_{\mu\nu}. \tag{5.16}$$

6. Since we already know $\delta g_{\mu\nu}$ from (5.13), we are finally in a position to compute:

$$\begin{aligned} S'_G(g) &= \frac{dS_G(g_s)}{ds}(s=0) = \int_M d^4x\,\delta(\sqrt{-g}g^{\mu\nu}R_{\mu\nu}) \\ &= \int_M d^4x\,[(\delta\sqrt{-g})g^{\mu\nu}R_{\mu\nu} + \sqrt{-g}\,(\delta g^{\mu\nu})R_{\mu\nu} + \sqrt{-g}\,g^{\mu\nu}\delta R_{\mu\nu}] \\ &= \int_M d^4x\,\sqrt{-g}\,(\tfrac{1}{2}g^{\mu\nu}R - R^{\mu\nu})d_{\mu\nu} + \int_{\partial M} d^3\vec{\sigma}^{\mu}(\nabla^{\nu}d_{\mu\nu} - \nabla_{\mu}d^{\nu}_{\nu}) \\ &= \int_M d^4x\,\sqrt{-g}\,(R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R)\delta g^{\mu\nu}, \end{aligned} \tag{5.17}$$

   where we used (5.13) to obtain the last term, and assume hat $d_{\mu\nu}$ has compact support: if $\partial M = \emptyset$ this immediately gives the last line, and if not, $d_{\mu\nu}$ should vanish on ∂M.[83]

   If there were no matter in the Universe, requiring $S'_G(g) = 0$ for arbitrary variations $d_{\mu\nu}$ (or, equivalently, $\delta g^{\mu\nu}$) therefore already gives us the ***vacuum Einstein equations***

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = 0. \tag{5.18}$$

7. It was a fact of great importance to Einstein that the gravitational action (5.2) is, as he called it, *generally covariant*, i.e., invariant under arbitrary coordinate transformations. We would now rather say that $S_G(g)$ is invariant under (orientation-preserving) diffeomorphisms, so if we consider special variations for which $g_s = \varphi_s^* g$, where φ_s is a one-parameter group of diffeomorphisms of M arising as the flow of a vector field $X \in \mathfrak{X}(M)$ having compact support (in which case it is complete), we have $S_G(\varphi^* g) = S_G(g)$,[84] and hence $S'_G(g) = 0$ for *any* metric g (i.e., whether or not $S'_G(g) = 0$ for *arbitrary* variations). On the other hand, as a special case of (1.57), for the above variations $g_s = \varphi_s^* g$ we have

$$\frac{dg_s}{ds}(s=0) = \mathscr{L}_X g, \tag{5.19}$$

   and for these specific variations we therefore have $d_{\mu\nu} = \nabla_\mu X_\nu + \nabla_\nu X_\mu$, where we used (2.62). Using the notation (3.20), as well as the symmetry of $G_{\mu\nu}$, we therefore have

$$\begin{aligned} 0 = S'_G(g) &= -\int_M d^4x\,\sqrt{-g}\,G^{\mu\nu}(\nabla_\mu X_\nu + \nabla_\nu X_\mu) \\ &= 2\int_M d^4x\,\sqrt{-g}\,(\nabla_\mu G^{\mu\nu})X_\nu - 2\int_M d^4x\,\sqrt{-g}\,\nabla_\mu(G^{\mu\nu}X_\nu). \end{aligned} \tag{5.20}$$

   As in (5.17) the last term is a boundary integral, which vanishes if X has compact support. The first term must vanish for arbitrary X, which recovers the Bianchi identity (3.21).

[82] This follows from linear algebra: $\partial g/\partial g_{\mu\nu} = m^{\mu\nu}$, i.e. the minor = cofactor of $g_{\mu\nu}$, and $g^{\mu\nu} = m^{\nu\mu}/g$.

[83] To be honest, we have not even defined manifolds with boundary …

[84] In the notation of the previous footnotes, we have $S_G(g) = \int_M \omega_g R_g$, where we have now explicitly indicated the g-dependence of ω and R. Then $\varphi^*\omega_g = \omega_{\varphi^* g}$ and $\varphi^* R_g = R_{\varphi^* g}$, so that $\omega_{\varphi^* g}R_{\varphi^* g} = \varphi^*\omega_g\varphi^* R_g = \varphi_*(\omega_g R_g)$. For any top-dimensional form $\alpha \in \Omega^n(M)$ (with compact support) one has $\int_M \varphi^*\alpha = \int_M \alpha$, so we may compute

$$S_G(\varphi^* g) = \int_M \omega_{\varphi^* g}R_{\varphi^* g} = \int_M \varphi^*(\omega_g R_g) = \int_M \omega_g R_g = S_G(g).$$

5.2 The energy-momentum tensor

The left-hand side of the Einstein equation (5.1) describes the geometry of space-time. The right-hand side $T_{\mu\nu}$ (times 8π), called the ***energy-momentum tensor***, describes the matter content of the universe. The first thing one infers from (5.1) is that $T \in \mathfrak{X}^{(2,0)}(M)$ has to satisfy

$$T_{\mu\nu} = T_{\nu\mu}, \tag{5.21}$$

or $T(X,Y) = T(Y,X)$. This makes index raising unambiguous, so that we may freely write T^{μ}_{ν} for either $g^{\mu\rho}T_{\rho\nu}$ or $g^{\mu\rho}T_{\nu\rho}$. As a case in point, the physical interpretation of $T_{\mu\nu}$ is that $T^{\mu}_{\nu}\dot{c}^{\nu}$ is the energy-momentum four-vector of matter, relative to an object (sometimes mistakenly described as an "observer", as if there were such things throughout the universe!) moving along a timelike (or even null) curve c. We will usually work in the setting of §3.7, so that $u = \dot{c}$ is a timelike unit vector normalized by (3.115), interpreted as the four-velocity of an "observer" (*sic*) moving along with whatever matter is described by T. In that case,

$$E = T(u,u) = T_{\mu\nu}u^{\mu}u^{\nu} \tag{5.22}$$

is the (relative) ***energy density*** of the matter. Similarly, one has a (covariant) ***momentum density***

$$P^{\mu} = -h^{\mu}_{\nu}T^{\nu}_{\rho}u^{\rho}, \tag{5.23}$$

cf. (3.122), which is orthogonal to u, i.e., $g(P,u) = 0$. The fully orthogonal projection of T, viz.

$$S_{\mu\nu} = h^{\rho}_{\mu}h^{\sigma}_{\nu}T_{\rho\sigma}, \tag{5.24}$$

is the ***stress tensor*** (of the given matter): if X and Y are spacelike unit vectors orthogonal to u, then $S(X,Y)$ is the force exerted by the matter in the direction X on the spacelike unit surface element normal to Y (and *vice versa*, since $S(X,Y) = S(Y,X)$). This gives the decomposition

$$T_{\mu\nu} = S_{\mu\nu} + P_{\mu}u_{\nu} + P_{\nu}u_{\mu} + Eu_{\mu}u_{\nu}. \tag{5.25}$$

Since the Einstein equations may be rewritten as

$$R_{\mu\nu} = 8\pi(T_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}T), \tag{5.26}$$

where $T = T^{\mu}_{\mu} = g^{\mu\nu}T_{\mu\nu}$ is the trace of T, it is often useful to know that, as implied by (5.25),

$$T = S - E, \tag{5.27}$$

where $S = g^{\mu\nu}S_{\mu\nu}$ is purely spatial, i.e. $S = \sum_{i=1}^{3} S(e_i,e_i)$ for some o.n.b. (e_i) orthogonal to u. For example, we may now rewrite the curvature condition (3.134) in Hawking's Theorem 17 as

$$E \geq -S. \tag{5.28}$$

More generally, the ***strong energy condition (SEC)*** requires for *any* timelike vector field ξ that

$$T(\xi,\xi) = T_{\mu\nu}\xi^{\mu}\xi^{\nu} \geq \tfrac{1}{2}g_{\mu\nu}\xi^{\mu}\xi^{\nu}T = \tfrac{1}{2}g(\xi,\xi)T. \tag{5.29}$$

Since the trace T may well be negative, this strengthens the ***weak energy condition (WEC)***

$$T(\xi,\xi) \geq 0. \tag{5.30}$$

For $\xi = u$, this just means $E \geq 0$. To complete this list of energy conditions, we also mention the ***(strengthened) dominant energy condition ((S)DEC)***, which requires (5.30) *and* $T^\mu_\nu \xi^\nu$ to be causal (timelike, provided $T_{\mu\nu} \neq 0$). We also have the fundamental ***conservation law***

$$\nabla_\mu T^\mu_\nu = 0, \tag{5.31}$$

which follows either from the Bianchi identity (3.21) and Einstein's equation (5.1), or from an argument like the one in §5.1.7, provided T^μ_ν can be derived from an action principle, see below.

Proposition 18 *Suppose a symmetric tensor $T_{\mu\nu}$ satisfies DEC and and* (5.31)*. If $S \subset M$ is an achronal set on which $T_{\mu\nu} = 0$, then $T_{\mu\nu}$ also vanishes on $D(S)$, cf.* (4.17)*.*

This is in fact a very hard result (see Hawking & Ellis, §4.3 for an equivalent claim).[85] To see SDEC in action, we mention another difficult result, making an insight of Einstein's rigorous:[86]

Proposition 19 *Suppose a symmetric tensor $T_{\mu\nu}$ satisfies SDEC and and* (5.31)*. Let $c : I \to M$ be a curve such that $T_{\mu\nu} = 0$ outside any nbhd of $c(I)$ but $T_{\mu\nu}(c(t)) \neq 0$ for some $t \in I$. Then c can be reparametrized (if necessary) so as to become a timelike geodesic, cf.* (2.40)

The idea is that $T_{\mu\nu}$ describes a point-like "test-particle", which moves under the influence of gravity but does not act as a source. Note that the Einstein equations (5.1) are not even assumed!

A much simpler result can be derived for so-called ***dust***, with energy-momentum tensor

$$T_{\mu\nu} = \rho u_\mu u_\nu, \tag{5.32}$$

where $\rho \in C^\infty(M)$ is the mass density and u is as above, including (3.115). Eq. (5.31) gives

$$\nabla_\mu(\rho u^\mu) \cdot u + \rho \nabla_u u = 0. \tag{5.33}$$

Since $g(u, \nabla_u u) = 0$ because of (5.31), contraction with u yields two independent conditions

$$\nabla_\mu(\rho u^\mu) = 0; \tag{5.34}$$

$$\nabla_u u = 0, \tag{5.35}$$

of which the first is a conservation law and the second is just the geodesic equation for u. Eq. (5.32) is a special case of the energy-momentum tensor of a ***perfect fluid***, which is given by

$$T_{\mu\nu} = (\varepsilon + p) u_\mu u_\nu + p g_{\mu\nu} = \varepsilon u_\mu u_\nu + p h_{\mu\nu}, \tag{5.36}$$

where the energy density ε is related by the pressure density p through some equation of state, such as $p = 0$ (dust, as above) or $p = \frac{1}{3}\varepsilon$ (ultrarelativistic fluid). Eq. (5.31) now gives

$$(\varepsilon + p)\nabla_\mu u^\mu + u(\varepsilon) = 0; \tag{5.37}$$

$$(\varepsilon + p)\nabla_u u^\mu + h^{\mu\nu}\partial_\nu p = 0, \tag{5.38}$$

called the (relativistic) ***Euler equations***. The quantities (5.22) - (5.24) are obviously given by

$$E = \varepsilon; \tag{5.39}$$

$$P = 0; \tag{5.40}$$

$$S_{\mu\nu} = p h_{\mu\nu}, \tag{5.41}$$

so that $S = 3p$ and $T = 3p - \varepsilon$. The energy conditions then come down to (nontrivial exercise!):

[85] Our formulation of Proposition 18 follows Malament, Prop. 2.5.1.

[86] See Geroch & Jang, Motion of a body in general relativity, *J. Math. Phys.* 16, 65–67 (1975).

- SEC holds iff $\varepsilon + p \geq 0$ and $\varepsilon + 3p \geq 0$;
- WEC holds iff $\varepsilon + p \geq 0$ and $\varepsilon \geq 0$;
- DEC and SDEC coincide in the case of (5.36) and both hold iff $\varepsilon \geq |p|$.

Except for fluids,[87] most energy-momentum tensors of interest to GR are derived from an action principle, like the Einstein equations in which they appear. The idea is that the "coupling" of gravity to matter is described by a functional $S_M(g,\varphi)$, where φ generically stands for all matter fields, so that, analogously to (5.17), one has

$$S'_M(g,\varphi) = -\tfrac{1}{2}\int_M d^4x\sqrt{-g}\,T_{\mu\nu}\delta g^{\mu\nu}, \tag{5.42}$$

where the prime has the same meaning as in §5.1 (varying the metric), or, as physicists write,[88]

$$T_{\mu\nu} = -2\frac{\delta S_M(g,\varphi)}{\delta g^{\mu\nu}}. \tag{5.43}$$

In this notation, the Einstein equation (5.1) then simply states that

$$\frac{\delta}{\delta g^{\mu\nu}}(S_G(g) + 16\pi S_M(g,\varphi)) = 0. \tag{5.44}$$

This equation for the metric $g_{\mu\nu}$ is to be supplemented with equations for the field(s), viz.[89]

$$\frac{\delta S_M(g,\varphi)}{\delta\varphi} = 0. \tag{5.45}$$

The simplest example is a ***scalar field*** $\varphi \in C^\infty(M)$, whose action functional is

$$S_M(g,\varphi) = -\tfrac{1}{2}\int_M d^4x\sqrt{-g}\,(g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi + V(\varphi)) \equiv -\tfrac{1}{2}\int_M (g(\nabla\varphi,\nabla\varphi) + V(\varphi)), \tag{5.46}$$

where $V : \mathbb{R} \to \mathbb{R}$ is a "potential" (which for a free field equals $V(\varphi) = \frac{1}{2}m^2\varphi^2$). The computation (5.17), with $R_{\mu\nu}$ replaced by $\partial_\mu\varphi\partial_\nu\varphi$ (so that there isn't even a boundary term) gives

$$T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \tfrac{1}{2}g_{\mu\nu}(g(\nabla\varphi,\nabla\varphi) + V(\varphi)). \tag{5.47}$$

Another case of interest is the ***electromagnetic field*** $A \in \Omega^1(M)$, with $F = dA \in \Omega^2(M)$, or

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = \nabla_\mu A_\nu - \nabla_\nu A_\mu, \tag{5.48}$$

where the last equality follows because ∇ is torsion-free. The (free) action is

$$S_M(g,A) = -\tfrac{1}{4}\int_M d^4x\sqrt{-g}\,g^{\mu\rho}g^{\nu\sigma}F_{\mu\nu}F_{\rho\sigma} \equiv -\tfrac{1}{4}\int_M F^2, \tag{5.49}$$

with $F^2 = F_{\mu\nu}F^{\mu\nu}$, from which a brief computation yields the energy-momentum tensor

$$T_{\mu\nu} = g^{\rho\sigma}F_{\mu\rho}F_{\nu\sigma} - \tfrac{1}{4}g_{\mu\nu}F^2, \tag{5.50}$$

where the last term comes from the variation of $\sqrt{-g}$ and the first one comes from $\delta(g^{\mu\rho}g^{\nu\sigma})$.

[87] Even for ideal fluids one has a (constrained) action principle due to A.H. Taub, but it is extremely contrived.

[88] In order to obtain the correct Einstein equations one is, of course, free to vary prefactors and even signs in (5.43) and (5.44), but our choice matches the convention for $T_{\mu\nu}$ in quantum field theory (with respect to which one should actually multiply Newton's constant G with the factor 16π in (5.44) and with 8π in (5.1).

[89] We might as well write these as $\delta(S_G(g) + S_M(g,\varphi))/\delta\varphi = 0$, since $S_G(g)$ is independent of φ.

5.3 Electromagnetism: gauge invariance and constraints

We elaborate on electromagnetism, since it allows us to make an important conceptual point with regard to the Einstein equations. First, the equation of motion for A_μ, obtained by varying A_μ (but not $g_{\mu\nu}$) in (5.49), or indeed in any action $S_M(g,A) = \int_M d^4x\sqrt{-g}\,\mathscr{L}(A,dA)$, is

$$\frac{\delta S_M(g,A)}{\delta A_\mu} = \frac{\partial \mathscr{L}}{\partial A_\mu} - \nabla_\nu \frac{\partial \mathscr{L}}{\partial(\partial_\nu A_\mu)} = 0, \tag{5.51}$$

where (compared to the usual Euler–Lagrange equation in flat space) the covariant derivative ∇_μ appears because of (5.8). For the specific action (5.49) this immediately yields

$$\nabla_\nu F^{\nu\mu} = 0, \tag{5.52}$$

which may, more intrinsically,[90] be written in terms of the Hodge dual as $d*F = 0$ (similarly, the other half of the Maxwell equations is $dF = 0$, which is automatic given $F = dA$), or as

$$\Box A_\mu - \nabla_\mu(\nabla_\nu A^\nu) = 0, \tag{5.53}$$

where $\Box = g^{\rho\sigma}\nabla_\rho\nabla_\sigma$ is the covariant d'Alembertian. To make our point it is enough to work in Minkowski space-time, in which $\nabla_\mu = \partial_\mu$, $A^0 = -A_0$, $A^i = A_i$ $(i = 1,2,3)$, and

$$\Box = -\partial_t^2 + \Delta. \tag{5.54}$$

In parallel with the discussion in §5.1.7, the action (5.49) is ***gauge invariant***, in that we have $S_M(A + d\lambda) = S_M(A)$, say for all $\lambda \in C_c^\infty(\mathbb{R}^4)$. This invariance under $\delta A_\mu = \partial_\mu\lambda$ yields

$$0 = \int_{\mathbb{R}^4} d^4x\,\partial_\nu F^{\nu\mu}\partial_\mu\lambda = -\int_{\mathbb{R}^4} d^4x\,\lambda\,\partial_\mu\partial_\nu F^{\nu\mu} \tag{5.55}$$

for all $\lambda \in C_c^\infty(\mathbb{R}^4)$, which gives the ***Bianchi identity for electromagnetism***, i.e.

$$\partial_\mu\partial_\nu F^{\nu\mu} = 0. \tag{5.56}$$

This is so obvious (in view of the antisymmetry of F) as to be disappointing, but it must be stressed that (5.56) is similar to (3.21) in being an *identity*, which holds irrespective of the equations of motion. See below for its thrust! Another consequence of gauge invariance is that

the equations of motion (5.53) *are simultaneously underdetermined and overdetermined*:

- They are *underdetermined* in that: if A solves (5.53), then so does $A + d\lambda$, $\lambda \in C_c^\infty(\mathbb{R}^4)$;
- They are *overdetermined* in that the initial values are constrained (i.e. cannot be arbitrary).

The first point is immediate from (5.53). For the second, since (5.53) looks hyperbolic we set up a Cauchy problem and give initial data $A_\mu(\vec{x})$ and $\dot{A}_\mu(\vec{x})$ at $x^0 \equiv t = 0$, where $\vec{x} = (x^1,x^2,x^3)$. However, defining the ***electric field*** in covariant form by $E_\mu = F_{\mu\nu}n^\nu$, or with respect to the 4-velocity $u = (1,0,0,0)$, by $E_i = F_{i0} = \partial_i A_0 - \partial_0 A_i$ $(i = 1,2,3)$, eq. (5.52) for $\mu = 0$ reads

$$C \equiv \partial^\nu F_{\nu 0} = \partial_i F_{i0} = \nabla\cdot\vec{E} = \Box A_0 - \partial_0(\partial_\nu A^\nu) = \Delta A_0 - \partial_0(\nabla\cdot\vec{A}) = 0. \tag{5.57}$$

[90] In coordinates the Hodge dual of F is $*F_{\mu\nu} = \frac{1}{2}g^{\alpha\rho}g^{\beta\sigma}\varepsilon_{\rho\sigma\mu\nu}F_{\alpha\beta}$, where ε is the Levi-Civita tensor.

This is not an evolution equation but a ***constraint*** on the initial data, called the ***Gauß law***. To address the first problem we pick a ***gauge condition***, which we take to be the ***Lorentz gauge***.

$$G \equiv \partial_\nu A^\nu = 0. \tag{5.58}$$

Preparing for GR, we introduce the notation $R_\mu = \partial^\nu F_{\nu\mu}$, so that (5.52) is $R_\mu = 0$, and also

$$R^L_\mu \equiv R_\mu + \partial_\mu G = \Box A_\mu, \tag{5.59}$$

so that instead of the awkward equations of motion (5.52) or (5.53) we would like to put

$$R^L_\mu = 0. \tag{5.60}$$

In order to solve (5.52) or (5.53) via (5.60), we now proceed as follows:

1. Solve the wave equation (5.60) for each $\mu = 0,1,2,3$ (in fact the case $\mu = 0$ will be trivial, see below), subject to initial data $A_\mu(\vec{x})$ and $\dot{A}_\mu(\vec{x})$ at $t = 0$ *that respect both the constraint*

   $$C(0,\vec{x}) = \Delta A_0(\vec{x}) - \partial_i \dot{A}_i(\vec{x}) = 0, \tag{5.61}$$

   and the gauge condition

   $$G(0,\vec{x}) \equiv \partial_i A_i(\vec{x}) - \dot{A}_0(\vec{x}) = 0. \tag{5.62}$$

   To show that this can indeed be done, first take $A_0(\vec{x}) = \dot{A}_0(\vec{x}) = 0$ (which, incidentally, solves (5.60) by $A_0(x) = 0$), so that (5.61) and (5.62) become $\partial_i \dot{A}_i = 0$ and $\partial_i A_i = 0$, respectively. For example, take $\dot{A}_i(\vec{x}) = 0$ but $A_i(\vec{x}) \neq 0$ arbitrary, and solve the elliptic equation $\Delta\lambda = -\partial_i A_i$ for λ. Replacing A_i by $A_i + \partial_i\lambda$ then satisfies (5.62).

2. From the definitions (5.61) and (5.62) of C and G, respectively, we immediately obtain

   $$\dot{G} = -C + R^L_0. \tag{5.63}$$

   From the Bianchi identity (5.56), i.e. $\partial_\mu\partial_\nu\partial^\mu A^\nu = \partial_\mu\partial_\nu\partial^\nu A^\mu$ (overkill!) we find

   $$\Box G = \partial^\mu R^L_\mu. \tag{5.64}$$

   Eqs. (5.60), (5.61), and (5.63) imply $\dot{G}(t = 0,\vec{x}) = 0$. Eqs. (5.64) and (5.60) also imply

   $$\Box G(x) = 0. \tag{5.65}$$

   With the initial conditions $G(t = 0,\vec{x}) = 0$, this implies $G(x) = 0$ for all $x \in \mathbb{R}^4$ by the standard theory of the wave equation. This is called the *propagation of the gauge*.

3. Similarly for the constraint (5.57). Using the Bianchi identity (5.56), we obtain

   $$\dot{C} = -\partial_0\partial_\nu F^{\nu 0} = -\partial_\mu\partial_\nu F^{\nu\mu} + \partial_i\partial_\nu F^{\nu i} = \partial_i\partial_\nu F^{\nu i} = \partial_i R_i = \partial_i(R^L_i - \partial_i G). \tag{5.66}$$

   Assuming the initial value (5.61) as well as *either* the 'gauged' equations of motions (5.60) for $\mu = i$ and the gauge condition and (5.62), which implied $G = 0$, *or* the dynamical Maxwell equations (5.52) for $\mu = i$, eq. (5.66) implies $\dot{C}(x) = 0$ for all x, so that, once again given (5.61), we obtain $C(x) = 0$ altogether, i.e. *propagation of the constraint*.

In conclusion, Maxwell's equations (5.52) or (5.53) may be solved by solving the 'gauged' Maxwell equations (5.60) subject to initial data $A_\mu(t = 0,\vec{x})$ and $\dot{A}_\mu(t = 0,\vec{x})$ that respect both the (initial data) constraint (5.61) and the (initial data) gauge condition (5.62). Indeed, as we have seen, together with (5.60) these two conditions on the initial data guarantee that both the constraint (5.57) and the gauge condition (5.58) hold everywhere, and the latter implies that the 'gauged' equations (5.60) actually coincide with the original ones, i.e. (5.52) or (5.53). In particular, since (5.60) is hyperbolic, the usual theory of the wave equation shows that the solution is unique (given the initial conditions). The case of GR will be quite similar!

5.4 General relativity: diffeomorphism invariance and constraints

To start, Einstein's equations (5.1) have two key features analogous to Maxwell's equations:

- They are *underdetermined* in that: if a metric g solves (5.1), then so does $\psi^* g$, for any diffeomorphism ψ of M (not necessarily an isometry, for which case the claim is trivial!).
- They are *overdetermined* in that the initial values are constrained (i.e. cannot be arbitrary).

The first point was already made in connection with the action principle, cf. §5.1.7, but of course it also follows from Einstein equations (5.1) themselves, which free of coordinates read

$$G(g) = 8\pi T(g, \varphi), \tag{5.67}$$

where G is the Einstein tensor (3.20). From (1.53) with $\psi \rightsquigarrow \psi^{-1}$, (2.44), (3.6) and (3.8) we obtain $R(\psi^* g) = \psi^* R(g)$ (where we explicitly denote the dependence of the Riemann tensor R on the metric g), and similarly for the Ricci tensor, the Ricci scalar, and the Einstein tensor, i.e.

$$G(\psi^* g) = \psi^* G(g). \tag{5.68}$$

Similarly, the energy-momentum tensor $T(g, \varphi)$ should be constructed in such a way that

$$T(\psi^* g, \psi^* \varphi) = \psi^* T(g, \varphi), \tag{5.69}$$

and hence Einstein's equation for g, i.e., $G(g) - 8\pi T(g, \varphi) = 0$ implies

$$G(\psi^* g) - 8\pi T(\psi^* g, \psi^* \varphi) = \psi^*(G(g) - 8\pi T(g, \varphi)) = \psi^* 0 = 0. \tag{5.70}$$

In what follows we just discuss the vacuum case ($T = 0$), since the general case is similar.[91] Our discussion takes place in coordinates (which is typical for PDE aspects of the Einstein equations), but in the next chapter we will also develop a purely geometric view of the situation.

From (3.11), (3.12), and (3.18) we easily obtain, in any coordinate system,

$$R_{\mu\nu} = -\tfrac{1}{2} g^{\rho\sigma} g_{\mu\nu,\rho\sigma} - \tfrac{1}{2} g^{\rho\sigma}(g_{\rho\sigma,\mu\nu} - g_{\sigma\nu,\mu\rho} - g_{\mu\rho,\sigma\nu}) + F(g, \partial g), \tag{5.71}$$

where $F(g, \partial g)$ contains only first derivatives of the metric.[92] For the Einstein tensor this gives

$$G_{\mu\nu} = -\tfrac{1}{2} g^{\rho\sigma}(g_{\mu\nu,\rho\sigma} + g_{\rho\sigma,\mu\nu} - g_{\sigma\nu,\mu\rho} - g_{\mu\rho,\sigma\nu} - g_{\mu\nu} g^{\alpha\beta}(g_{\alpha\beta,\rho\sigma} - g_{\sigma\alpha,\rho\beta})) + \tilde{F}(g, \partial g). \tag{5.72}$$

Although this point will be studied in great detail in the next two chapters, we now point out (though in a somewhat superficial and coordinate-dependent way) that the ten (vacuum) Einstein equations $G_{\mu\nu} = 0$ (and more generally the full equations $G_{\mu\nu} = 8\pi T_{\mu\nu}$) come in two groups:

- The six ***dynamical equations*** $G_{ij} = 0$, where $i, j = 1, 2, 3$ as usual;
- The four ***constraints*** $C_\mu \equiv G_{\mu 0} = 0$, where $\mu = 0, 1, 2, 3$.

[91] The discussion revolves around second derivatives of $g_{\mu\nu}$ in the Einstein equation, which are absent in $T_{\mu\nu}$.

[92] We will later see that in the relevant PDE theory only the highest derivatives of the unknown functions count.

This can be seen from (5.72), using a coordinate system where $g_{\mu 0} = 0$ (so that $g_{00}g^{00} = 1$):[93] none of the constraints contain second-order derivatives of the components of the metric.

As in (5.53), the first term in (5.71), which is essentially $\Box g_{\mu\nu}$, is favourable towards a good PDE theory (as we will see, it makes the spatial components of g satisfy a hyperbolic evolution equation), but the other three terms, which are the anlogue of the second term in (5.53), ruin this and hence should be removed by a clever choice of coordinates (which makes them disappear).

The simplest way to do this (introduced by Choquet-Bruhat) is to use the ***wave gauge***[94]

$$W^\mu \equiv \Box_g x^\mu = 0, \tag{5.73}$$

where the ***covariant D'Alembertian*** $\Box_g$ is defined, on any tensor, by

$$\Box_g = g^{\rho\sigma}\nabla_\rho\nabla_\sigma. \tag{5.74}$$

In (5.73) the coordinate functions x^μ are *scalar functions*,[95], so that, $\partial_\sigma x^\mu$ being a 1-form,

$$\begin{aligned} W^\mu = \Box_g x^\mu = g^{\rho\sigma}\nabla_\rho\partial_\sigma x^\mu = g^{\rho\sigma}(\partial_\rho\partial_\sigma - \Gamma^\nu_{\rho\sigma}\partial_\nu)x^\mu = g^{\rho\sigma}(\partial_\rho\delta^\mu_\sigma - \Gamma^\nu_{\rho\sigma}\delta^\mu_\nu) \\ = -g^{\rho\sigma}\Gamma^\mu_{\rho\sigma}. \end{aligned} \tag{5.75}$$

Using (3.12), this yields a key result, where $H(g,\partial g)$ has a similar meaning as $F(g,\partial g)$:

$$g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho = g^{\rho\sigma}(g_{\rho\sigma,\mu\nu} - g_{\sigma\nu,\mu\rho} - g_{\mu\rho,\sigma\nu}) + H(g,\partial g). \tag{5.76}$$

Therefore, analogously to (5.60), the ***wave-gauged*** (or ***reduced***) ***vacuum Einstein equations***

$$R^W_{\mu\nu} \equiv R_{\mu\nu} + \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho) = 0, \tag{5.77}$$

take the desirable (quasi-linear hyperbolic) form (starting with the D'Alembertian):

$$R^W_{\mu\nu} = -\tfrac{1}{2}g^{\rho\sigma}g_{\mu\nu,\rho\sigma} + I(g,\partial g) = 0. \tag{5.78}$$

From (5.77) we also define the ***reduced Einstein tensor***

$$G^W_{\mu\nu} = R^W_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R^W = G_{\mu\nu} + \tfrac{1}{2}(g_{\mu\rho}\partial_\nu W^\rho + g_{\nu\rho}\partial_\mu W^\rho - g_{\mu\nu}\partial_\rho W^\rho), \tag{5.79}$$

so that, *provided the metric solves* (5.77), the Einstein tensor is related to the wave gauge by

$$G_{\mu\nu} = \tfrac{1}{2}(g_{\mu\nu}\partial_\rho W^\rho - g_{\mu\rho}\partial_\nu W^\rho - g_{\nu\rho}\partial_\mu W^\rho). \tag{5.80}$$

In particular, the four constraints C_μ are linear combinations of the four time-derivatives $\dot{W}^\mu$ and of the W^μ themselves and their spatial derivatives. Conversely, the $\dot{W}^\mu$ are linear combinations of the constraints C_μ and the W^μ and their spatial derivatives. For example, in coordinates where (at $t = 0$) one has $g_{00} = -1$ and $g_{0i} = 0$, these linear relations are simply

$$C_0 = \tfrac{1}{2}(\dot{W}^0 - \partial_j W^j); \tag{5.81}$$

$$C_i = \tfrac{1}{2}(\partial_i W^0 - g_{ij}\dot{W}^j). \tag{5.82}$$

[93] This restriction is only necessary to see that $G_{00} = 0$ is a constraint.

[94] Coordinates satisfying (5.73) are often called ***harmonic*** or ***wave coordinates***. See Choquet-Bruhat), §VI.7.

[95] As opposed to components of a 4-vector. Choquet-Bruhat even writes $x^{(\mu)}$ as a warning.

So if we impose both the constraints and the gauge conditions at $t = 0$, i.e.,

$$C_\mu(t = 0, \vec{x}) = 0; \tag{5.83}$$
$$W^\mu(t = 0, \vec{x}) = 0 \tag{5.84}$$

and also assume the reduced Einstein equations (5.77), then automatically,

$$\dot{W}^\mu(t = 0, \vec{x}) = 0. \tag{5.85}$$

Parallel to step 2 for electromagnetism, we apply the Bianchi identities $\nabla^\mu G_{\mu\nu} = 0$, i.e.

$$g^{\mu\rho}(\partial_\rho G_{\mu\nu} - \Gamma^\sigma_{\rho\mu} G_{\sigma\nu} - \Gamma^\sigma_{\rho\nu} G_{\mu\sigma}) = 0, \tag{5.86}$$

to (5.80). This gives a hyperbolic quasi-linear PDE for W^ρ whose principal term is $\partial^\mu \partial_\mu W^\rho$, since to leading order the first two terms on the right-hand side of (5.80) cancel out. Given the initial conditions (5.84) and (5.85), by quasi-linear hyperbolic PDE theory this implies

$$W^\mu(x) = 0 \tag{5.87}$$

altogether; the underlying assumptions were the reduced Einstein equations (5.77) or (5.79) and the initial value conditions (5.83) - (5.84).

Step 3 for electromagnetism also applies here: once again, the Bianchi identities, in full:

show that the constraints C_μ satisfy a linear homogeneous first-order symmetric system of of PDE's, provided we assume *either* the reduced Einstein equations $G^W_{ij} = 0$ and the gauge condition as above, *or*, equivalently, the original Einstein equations $G_{ij} = 0$ (which would spoil homogeneity). If we assume $C_\mu = 0$ at $t = 0$, the unique solution of this system is $C_\mu(x) = 0$ at all x. So just as for the gauge condition, assuming the constraints at $t = 0$ and the reduced Einstein equations (of which this time only the spatial and hence dynamical part is needed) guarantees that the constraints are always satisfied. In sum, if we assume:

1. The constraints $C_\mu = 0$ at $t = 0$;
2. The wave gauge condition $W^\mu = 0$ at $t = 0$;
3. The (quasi-linear hyperbolic) reduced Einstein equations $R^W_{\mu\nu} = 0$ or $G^W_{\mu\nu} = 0$,

then the full (vacuum) Einstein equations $G_{\mu\nu} = 0$ or $R_{\mu\nu} = 0$ hold (at least locally).

We will make this more precise in the final chapter on quasi-linear hyperbolic PDE's.

6 Submanifolds

Differential geometry started with the study of two-dimensional submanifolds S of $\mathbb{R}^3$ (i.e. surfaces) by Gauß, and in GR a crucial role will be played by (spacelike) three-dimensional submanifolds S of a four-dimensional Lorentzian manifold M. This leads to an interplay between the intrinsic geometric properties of S and the additional ('extrinsic') geometry obtained from its embedding $S \hookrightarrow \mathbb{R}^3$ or $S \hookrightarrow M$. This interplay was already analyzed by Gauß himself.

6.1 Basic definitions

One may define a submanifold S of M (where M is any manifold) in two equivalent ways: either as a subset $S \subset M$ of M with certain (good) properties, or as a manifold in its own right (a concept already defined, of course) plus an explicit map $F : S \to M$ with certain properties. The former leads to the latter by considering the inclusion map $S \hookrightarrow M$, whereas the latter leads to the former by identifying S with its image $F(S) \subset M$ (which may lead to some confusion!).

1. Let S be a manifold. A map $F : S \to M$ defines a ***submanifold*** $F(S) \subset M$ provided:[96]

   (a) F is a homeomorphism onto its image $F(S)$ (in particular, F is injective);

   (b) $F'_u : T_uS \to T_{F(u)}M$ is injective for all $u \in S$ (equivalently, has rank equal to $\dim(S)$).

2. A subset $S \subset \mathbb{R}^n$ is a ***k-dimensional submanifold of*** $\mathbb{R}^n$ iff each $u \in S$ has a nbhd $U \in \mathscr{O}(\mathbb{R}^n)$ in $\mathbb{R}^n$ such that $S \cap U$ is homeomorphic to some $W \in \mathscr{O}(\mathbb{R}^k)$ via $\xi : W \to \mathbb{R}^n$ (i.e. $S \cap U \cong \xi(W)$), for which in addition each map $\xi'_y : T_y\mathbb{R}^k \to T_{\xi(y)}\mathbb{R}^n$ is injective ($y \in W$).

   Subsequently, we say that $S \subset M$ is a ***k-dimensional submanifold of a manifold*** M (where $\dim(M) = n$) iff for each $u \in S$ and each chart (for M) $\varphi : U \to V \subset \mathbb{R}^n$ with $x \in U$, the image $\varphi(S \cap U)$ is a k-dimensional submanifold of $\mathbb{R}^n$ (in the sense just defined).

3. These definitions are equivalent in the way just mentioned (see e.g. Andrews, Prop. 3.2.1).

The main point is to exclude unwanted things like $\alpha \subset \mathbb{R}^2$ (intersection) or $\angle \subset \mathbb{R}^2$ (corner).

6.2 Classical theory of surfaces

The classical theory of surfaces $\Sigma \subset \mathbb{R}^3$ was largely based on local constructions. Let $U \subset \mathbb{R}^2$ be open and let $F : U \to \mathbb{R}^3$ satisfy the two conditions above, with image $\Sigma = F(U)$. The standard coordinates $u = (u^1, u^2)$ on U induce the same coordinates on Σ (i.e. the point $F(u^1, u^2) \in \Sigma \subset \mathbb{R}^3$ is said to have coordinates (u^2, u^2), too) and come with three vector fields on Σ, defined by

$$\vec{x}_1 = F'(\partial/\partial u^1); \tag{6.1}$$

$$\vec{x}_2 = F'(\partial/\partial u^2); \tag{6.2}$$

$$\vec{N} = \frac{\vec{x}_1 \times \vec{x}_2}{\|\vec{x}_1 \times \vec{x}_2\|}, \tag{6.3}$$

where injectivity of F' implies that the denominator in (6.3) is nonzero. For the same reason, the triple $(\vec{x}_1, \vec{x}_2, \vec{N})$ forms a basis of $T_u\mathbb{R}^3 \cong \mathbb{R}^3$, whilst $(\vec{x}_1, \vec{x}_2)$ is a basis of $T_{F(u)}\Sigma$, $u \in U$.

[96] Recall that F is smooth. Technically, we define an ***embedded submanifold***. A weaker notion, called an ***immersed submanifold***, in which the first condition is dropped, makes sense but wil not be used in these notes.

We let early Greek alphabet indices α, β etc. run through 1, 2, and also $i = 1,2,3$, so that $F^i : U \to \mathbb{R}$ are the coordinates of F, regarded as functions of (u^1, u^2). Then (6.1) - (6.2) is just

$$x^i_\alpha = \partial F^i / \partial u^\alpha \equiv \partial_\alpha F^i \qquad (\alpha = 1,2; i = 1,\ldots,3), \tag{6.4}$$

In the 19th century two tensors on Σ were identified (to be used in a very similar way in GR):

1. The ***first fundamental form*** $\tilde{g}$ is the metric induced by the Euclidean metric δ on $\mathbb{R}^3$, i.e.

$$\tilde{g} = F^* \delta, \tag{6.5}$$

where $F^* \equiv F^{(2,0)} : \mathfrak{X}^{(2,0)}(\mathbb{R}^3) \to \mathfrak{X}^{(2,0)}(\Sigma)$ is the pullback of $F : U \to \mathbb{R}^3$ defined after (1.52). This simply means that $\tilde{g}$ is the restriction of δ to $\Sigma \subset \mathbb{R}^3$, that is,

$$\tilde{g}(X,Y) = \delta(X,Y) = \langle X, Y \rangle, \tag{6.6}$$

where $\langle \cdot, \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^3$, and $X, Y \in \mathfrak{X}(\Sigma)$. In the (u^1, u^2) coordinates on Σ, the components of $\tilde{g}$ are given by

$$\tilde{g}_{\alpha\beta} = \tilde{g}(\partial_\alpha, \partial_\beta) = \langle \vec{x}_\alpha, \vec{x}_\beta \rangle, \tag{6.7}$$

where we collectively write α and β for u and v. By construction,

$$\tilde{g}_{\alpha\beta} = \sum_{i=1}^{3} \frac{\partial F^i}{\partial u^\alpha} \cdot \frac{\partial F^i}{\partial u^\beta}, \tag{6.8}$$

from which we see that although the (∂_1, ∂_2) basis is orthonormal in $U \subset \mathbb{R}^2$, its push-forward to Σ may no longer be orthonormal in $\mathbb{R}^3$ (this depends on F).

2. To define the ***second fundamental form*** $\tilde{k}$, we first observe that for vector field $X \in \mathfrak{X}(\Sigma)$ is also a vector field on $\mathbb{R}^3$ (restricted to Σ), so that along Σ we may define the 3-vectors

$$\nabla_X \vec{N} = X^\alpha \frac{\partial \vec{N}}{\partial u^\alpha}, \tag{6.9}$$

where $X = X^\alpha \vec{x}_\alpha$; if $X_u \equiv X_{F(u)}$ is tangent to a curve $F(\gamma^1(t), \gamma^2(t))$, then $X^\alpha = d\gamma^\alpha / dt_{|t=0}$; we may then also write $\nabla_X \vec{N}(u,v) = d\vec{N}(\gamma^1(t), \gamma^2(t))/dt_{|t=0}$ (the notation ∇_X is used because from a higher perspective one uses covariant differentiation with respect to the Levi-Civita connection defined by the flat metric δ on $\mathbb{R}^3$). One could also simply say

$$\nabla_X N^i = XN^i = X^\alpha \partial_\alpha N^i \qquad (i = 1,2,3), \tag{6.10}$$

which is (2.31) with vanishing Christoffel symbols (in $\mathbb{R}^3$). Since $\langle \vec{N}, \vec{N} \rangle = 1$, we have

$$0 = X(1_\Sigma) = X(\langle \vec{N}, \vec{N} \rangle) = \langle \nabla_X \vec{N}, \vec{N} \rangle + \langle \vec{N}, \nabla_X \vec{N} \rangle = 2\langle \nabla_X \vec{N}, \vec{N} \rangle, \tag{6.11}$$

so that $\nabla_X \vec{N}$ is orthogonal to $\vec{N}$ (in $\mathbb{R}^3$), and hence it must be tangent to Σ. In other words, we have the ***Weingarten map*** (with a minus sign for historical reasons)

$$W : T\Sigma \to T\Sigma; \tag{6.12}$$

$$X \mapsto -\nabla_X \vec{N}. \tag{6.13}$$

Strictly speaking we should write $W_{(u,v)} : T_{(u,v)}\Sigma \to T_{(u,v)}\Sigma$, etc. Finally, we define $\tilde{k}$ by

$$\tilde{k}(X,Y) = \tilde{g}(W(X),Y) = -\tilde{g}(\nabla_X \vec{N}, Y) = -\langle \nabla_X \vec{N}, Y\rangle. \tag{6.14}$$

It is easy to show that the second fundamental form thus defined is *symmetric*, i.e.,

$$\tilde{k}(X,Y) = \tilde{k}(Y,X), \tag{6.15}$$

or $\langle \nabla_Y \vec{N}, X\rangle = \langle \nabla_X \vec{N}\rangle$. To see this, note that $\langle \vec{N}, X\rangle = 0$ (since X and Y are tangent to Σ and hence orthogonal to $\vec{N}$), hence $0 = Y(\langle \vec{N}, X\rangle) = \langle \nabla_Y \vec{N}, X\rangle + \langle \vec{N}, \nabla_Y X\rangle$. Since ∇ (as the flat Levi-Civita connection on $\mathbb{R}^3$) is torsion-free, we have $\nabla_Y X = \nabla_X Y - [X,Y]$, so

$$\langle \nabla_Y \vec{N}, X\rangle = -\langle \vec{N}, \nabla_Y X\rangle = -\langle \vec{N}, \nabla_X Y\rangle + \langle \vec{N}, [X,Y]\rangle = -\langle \vec{N}, \nabla_X Y\rangle = \langle \nabla_X \vec{N}, Y\rangle, \tag{6.16}$$

because $\langle \vec{N}, [X,Y]\rangle = 0$ (because $[X,Y]$ is tangent to Σ whenever X and Y are). This computation also yields an alterative expression for $\tilde{k}$, which is manifestly symmetric:

$$\tilde{k}_{\alpha\beta} = \langle \vec{x}_{\alpha\beta}, \vec{N}\rangle; \tag{6.17}$$

$$\vec{x}_{\alpha\beta} \equiv \partial_\beta \vec{x}_\alpha; \tag{6.18}$$

in terms of $F : U \to \mathbb{R}^3$, the components of the vector $\vec{x}_{\alpha\beta}$ are $x^i_{\alpha\beta} = \partial^2 F^i / \partial u^\alpha \partial u^\beta$.

3. Many computations in the theory of surfaces use the ***Gauß–Weingarten equations***

$$\vec{x}_{\alpha\beta} = \tilde{\Gamma}^\gamma_{\alpha\beta} \vec{x}_\gamma + \tilde{k}_{\alpha\beta} \vec{N}; \tag{6.19}$$

$$\partial_\alpha \vec{N} = -\tilde{k}^\beta_\alpha \vec{x}_\beta, \tag{6.20}$$

where the $\tilde{\Gamma}^\gamma_{\alpha\beta}$ are the Christoffel symbols (as originally introduced!) associated to the metric $\tilde{g}$ on Σ, and $\tilde{k}^\beta_\alpha = \tilde{g}^{\beta\gamma}\tilde{k}_{\alpha\gamma}$, where $(\tilde{g}^{\beta\gamma})$ is the inverse matrix to $(\tilde{g}_{\beta\gamma})$, as usual. Weingarten's eq. (6.20) is just a restatement of (6.14). Gauß's eq. (6.19) is simply the expansion of the 3-vectors $\vec{x}_{\alpha\beta}$ in terms of the basis $(\vec{x}_u, \vec{x}_v, \vec{N})$. The specific form $\tilde{k}_{\alpha\beta}$ of the coefficient of $\vec{N}$ follows from (6.17). To derive the coefficient of $\vec{x}_\gamma$, let us assume (6.19) for initially unknown coefficients $\tilde{\Gamma}^\gamma_{\alpha\beta}$. We then obtain

$$\langle \vec{x}_\gamma, \vec{x}_{\alpha\beta}\rangle = \tilde{\Gamma}^\delta_{\alpha\beta}\langle \vec{x}_\gamma, \vec{x}_\delta\rangle = \tilde{g}_{\gamma\delta}\tilde{\Gamma}^\delta_{\alpha\beta}, \tag{6.21}$$

so that $\tilde{\Gamma}^\gamma_{\alpha\beta} = g^{\gamma\delta}\langle \vec{x}_\delta, \vec{x}_{\alpha\beta}\rangle$. The relation (2.23) then follows from (6.18), which yields

$$2\langle \vec{x}_\delta, \vec{x}_{\alpha\beta}\rangle = \partial_\beta\langle \vec{x}_\delta, \vec{x}_\alpha\rangle + \partial_\alpha\langle \vec{x}_\delta, \vec{x}_\beta\rangle - \partial_\delta\langle \vec{x}_\alpha, \vec{x}_\beta\rangle. \tag{6.22}$$

4. The classical theory also heavily relies on the ***Gauß–Codazzi equations***

$$\tilde{R}^\delta_{\alpha\gamma\beta} = \tilde{k}^\delta_\gamma \tilde{k}_{\alpha\beta} - \tilde{k}^\delta_\beta \tilde{k}_{\alpha\gamma}; \tag{6.23}$$

$$\tilde{k}_{\alpha\beta,\gamma} + \tilde{\Gamma}^\delta_{\alpha\beta}\tilde{k}_{\gamma\delta} = \tilde{k}_{\alpha\gamma,\beta} + \tilde{\Gamma}^\delta_{\alpha\gamma}\tilde{k}_{\beta\delta}, \tag{6.24}$$

where $\tilde{R}^\delta_{\alpha\gamma\beta}$ is the Riemann tensor as defined (in terms of the metric $\tilde{g}$ on Σ) in (3.11), indices are raised with the aid of $\tilde{g}$ as usual (e.g. $\tilde{k}^\delta_\gamma = \tilde{g}^{\delta\beta}\tilde{k}_{\beta\gamma}$), and $\tilde{k}_{\alpha\gamma,\beta} = \partial_\beta \tilde{k}_{\alpha\gamma}$.

The Gauß–Codazzi equations (for $\tilde{g}$ and $\tilde{k}$) follow from the identity

$$\partial_\gamma \partial_\beta \vec{x}_\alpha = \partial_\beta \partial_\gamma \vec{x}_\alpha, \tag{6.25}$$

i.e., $\partial_\gamma \vec{x}_{\alpha\beta} = \partial_\beta \vec{x}_{\alpha\gamma}$. Indeed, from the Gauß–Weingarten equations one easily finds

$$\begin{aligned} \vec{x}_{\alpha\beta\gamma} - \vec{x}_{\alpha\gamma\beta} &= (\tilde{R}^\delta_{\alpha\gamma\beta} - \tilde{k}^\delta_\gamma \tilde{k}_{\alpha\beta} + \tilde{k}^\delta_\beta \tilde{k}_{\alpha\gamma})\vec{x}_\delta \\ &\quad + (\tilde{k}_{\alpha\beta,\gamma} + \tilde{\Gamma}^\delta_{\alpha\beta}\tilde{k}_{\gamma\delta} - \tilde{k}_{\alpha\gamma,\beta} + \tilde{\Gamma}^\delta_{\alpha\gamma}\tilde{k}_{\beta\delta})\vec{N}, \end{aligned} \tag{6.26}$$

so that Gauß's equation (6.23) is the tangential (to Σ) component of (6.25), whilst Codazzi's equation (6.24) is its normal component. Eq. (6.23) is especially interesting, since it relates the intrinsic geometry of Σ (represented by its Riemann curvature tensor) to its extrinsic geometry (represented by the second fundamental form). In fact, Gauß's famous *Theorema Egregium* easily follows from (6.23). Following Gauß, we define

$$K = \det(W) = \det(\tilde{k})/\det(\tilde{g}) \ \textbf{\textit{(Gaußcurvature)}}; \tag{6.27}$$

$$H = \mathrm{tr}(W) = \mathrm{tr}(\tilde{g}^{-1}\tilde{k}) = \ \textbf{\textit{(mean curvature)}}, \tag{6.28}$$

where W is the Weingarten map (6.12) - (6.13), and $\tilde{k}$ and $\tilde{g}$ are the matrices defined by (6.7) and (6.17). In terms of the eigenvalues κ_1 and κ_2 of W we therefore have

$$K = \kappa_1 \kappa_2; \tag{6.29}$$

$$H = \kappa_1 + \kappa_2. \tag{6.30}$$

Then the *Theorema Egregium* is nothing but the relation (3.27) we already saw.

6.3 Hypersurfaces in arbitrary (semi) Riemannian manifolds

For applications to GR we need a similar theory, in which $\Sigma \subset M$ is a submanifold of codimension one of a Lorentzian manifold M, i.e. $\dim(\Sigma) = m$ and $\dim(M) = m+1 \equiv n$ (where $m = 3$ for GR); such a submanifold is often called a ***hypersurface***. Without much extra effort, we will develop the theory in both the Riemannian and the Lorentzian case (the general semi-Riemannian case requires too many adaptations). Thus we assume that M carries either a Riemannian or a Lorentzian metric g, with associated Levi-Civita connection ∇ on TM, and that Σ carries the induced metric $\tilde{g} \in \mathfrak{X}^{(2,0)}(\tilde{M})$ defined by the inclusion $\imath : \Sigma \hookrightarrow M$, i.e.

$$\tilde{g} = \imath^* g, \tag{6.31}$$

which simply means that $\tilde{g}_x(X_x, Y_x) = g_x(X_x, Y_x)$ for any $X_x, Y_x \in T_x\Sigma \subset T_xM$, with $x \in \Sigma$. In both cases, we assume $(\Sigma, \tilde{g})$ to be a Riemannian manifold in its own right (which in the Lorentzian case is not automatic and forces Σ to be spacelike). The induced metric $\tilde{g}$ induces an associated Levi-Civita connection $\tilde{\nabla}$ on $T\Sigma$, whose relationship with ∇ we will now unearth. The ensuing Gauß–Weingarten equations require a choice of normal unit vectors $N_x \in T_xM$ to $T_x\tilde{M}$ (i.e. $g_x(N_x, X_x) = 0$ for all $X_x \in T_x\tilde{M}$, where $x \in \tilde{M}$), generalizing (6.3). In general, there is no canonical choice of N_x, but any two choices differ by a sign and we assume that we can make a smooth choice $x \mapsto N_x$ throughout $\tilde{M}$.[97] The normalization of N_x carries a "timelike" subtlety:

$$g_x(N_x, N_x) = 1 \ \ (\text{Riemannian case}); \tag{6.32}$$

$$g_x(N_x, N_x) = -1, (\text{Lorentzian case}). \tag{6.33}$$

[97] A sufficient condition is that $\tilde{M}$ be connected and simply connected (cf. Kobayashi & Nomizu, Vol. 2, p. 5). In GR the presence of a time orientation will fix N, which we may require to be future directed.

Then the orthogonal projection (which is independent of the choice of N_x) onto $T_x\tilde{M}$ is

$$\pi_x : T_xM \to T_x\tilde{M} \subset T_xM \tag{6.34}$$

$$\pi_x(X_x) = X_x - g_x(X_x, N_x)N_x; \quad \text{(Riemannian case)}; \tag{6.35}$$

$$\pi_x(X_x) = X_x + g_x(X_x, N_x)N_x; \quad \text{(Lorentzian case)}, \tag{6.36}$$

since one requires $\pi_x(N_x) = 0$ (and $\pi_x(X_x) = X_x$ if $X_x \in T_x\tilde{M}$).

1. The key to the entire (metric) theory of hyersurfaces is the result

$$\pi(\nabla_X Y) = \tilde{\nabla}_X Y \ \ (X, Y \in \mathfrak{X}(\tilde{M})), \tag{6.37}$$

where the covariant derivative $\tilde{\nabla}_X Y$ on the right-hand side is clearly defined (as an element of $\mathfrak{X}(\tilde{M})$), but also the covariant derivative $\nabla_X Y$ in M on the left-hand side is well defined, even though Y is merely a vector field on $\tilde{M}$ rather than on all of M: as in the comment preceding (2.35), if $X \in \mathfrak{X}(\tilde{M})$ and $Y \in \mathfrak{X}(M)$, then the value of $\nabla_X Y$ only depends on the restriction of Y to $\tilde{M}$ (indeed, it only depends on the values of Y along the flow lines on X, which lie in $\tilde{M}$), and so $\nabla_X Y$ is defined (as a vector field on $\tilde{M}$) even when $Y \in \mathfrak{X}(\tilde{M})$.[98]

To prove (6.37), we write $\nabla'_X Y$ for $\pi(\nabla_X Y)$, so that (in the Lorentzian case for simplicity)

$$\nabla'_X Y = \nabla_X Y + g(\nabla_X Y, N)N. \tag{6.38}$$

We first check that ∇' is a covariant derivative on $\mathfrak{X}(\tilde{M})$. Linearity in Y is obvious (since both g and ∇_X are linear), as is rule (2.27). Rule (2.28) follows from the corresponding rule for ∇ and the property $g((Xf)Y, N) = (Xf)g(Y,N) = 0$ (since $Y \in \mathfrak{X}(\tilde{M})$). To make the identification $\nabla' = \tilde{\nabla}$ we next need to check that ∇' is torsion-free, which is the case:

$$\begin{aligned} \nabla'_X Y - \nabla'_Y X &= \nabla_X Y - \nabla_Y X + g(\nabla_X Y - \nabla_Y X, N)N \\ &= [X,Y] + g([X,Y],N)N \\ &= [X,Y], \end{aligned} \tag{6.39}$$

since ∇ (being the Levi-Civita connection on TM) is torsion-free, and $[X,Y] \in \mathfrak{X}(\tilde{M})$, assuming $X, Y \in \mathfrak{X}(\tilde{M})$, so that $g([X,Y],N) = 0$. Finally, ∇' should satisfy (2.42), i.e.

$$X(\tilde{g}(Y,Z)) = \tilde{g}(\nabla'_X Y, Z) + \tilde{g}(Y, \nabla'_X Z) \ \ (X,Y,Z \in \mathfrak{X}(\tilde{M})). \tag{6.40}$$

This is quite obvious, since for $X, Y, Z \in \mathfrak{X}(\tilde{M})$ we have

$$\tilde{g}(\nabla'_X Y, Z) = g(\nabla'_X Y, Z) = g(\nabla_X Y + g(\nabla_X Y, N)N, Z) = g(\nabla_X Y, Z), \tag{6.41}$$

since $g(N,Z) = 0$, and so the right-hand side of (6.40) equals $g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$. By (2.42) for ∇ and g, this in turn equals $X(g(Y,Z)) = X(\tilde{g}(Y,Z))$, and we are done.

2. Eq. (6.37) implies the general ***Gauß–Weingarten equations***, where still $X, Y \in \mathfrak{X}(\tilde{M})$:

$$\nabla_X Y = \tilde{\nabla}_X Y + \tilde{k}(X,Y)N \ \ \text{(Riemann)}; \tag{6.42}$$

$$\nabla_X Y = \tilde{\nabla}_X Y - \tilde{k}(X,Y)N \ \ \text{(Lorentz)}; \tag{6.43}$$

$$\nabla_X N = -W(X). \tag{6.44}$$

[98] In other words, if one insists that $\nabla_X : \mathfrak{X}(M) \to \mathfrak{X}(M)$, one may extend $Y \in \mathfrak{X}(M)$ to any vector field on M and if $X \in \mathfrak{X}(\tilde{M})$, then $\nabla_X Y$ is independent of the extension.

Here (6.44) is the *definition* of the (generalized) Weingarten map $W_x : T_x\tilde{M} \to T_x\tilde{M}$ (as before, since $g(N,N) = \pm 1$, we have $g(\nabla_X N, N) = 0$ and hence $\nabla_X N \in TM$). Furthermore, taking the (metric) inner product of (6.42) - (6.43) with N, and using (6.32) - (6.33) as well as the relation

$$g(\nabla_X Y, N) = -g(Y, \nabla_X N), \tag{6.45}$$

which is proved in the same way as in the text between (6.15) and (6.16), we obtain

$$\tilde{k}(X,Y) = g(W(X),Y), \tag{6.46}$$

This once again defines the (generalized) ***second fundamental form*** $\tilde{k} \in \mathfrak{X}^{(2,0)}(\tilde{M})$. The same calculation (6.16) as before shows that $\tilde{k}$ is symmetric, as in (6.15), viz.

$$\tilde{k}(X,Y) = -g(\nabla_X N, Y) = g(N, \nabla_X Y) = g(N, \nabla_Y X) = \tilde{k}(Y,X). \tag{6.47}$$

3. We now derive the general ***Gauß–Codazzi equations***, which, for $W,X,Y,Z \in \mathfrak{X}(\tilde{M})$, are:

$$R(W,Z,X,Y) = \tilde{R}(W,Z,X,Y) + \tilde{k}(W,Y)\tilde{k}(X,Z) - \tilde{k}(W,X)\tilde{k}(Y,Z) \quad \text{(Riemann)}; \tag{6.48}$$

$$R(W,Z,X,Y) = \tilde{R}(W,Z,X,Y) + \tilde{k}(W,X)\tilde{k}(Y,Z) - \tilde{k}(W,Y)\tilde{k}(X,Z) \quad \text{(Lorentz)}; \tag{6.49}$$

$$R(N,Z,X,Y) = (\tilde{\nabla}_X \tilde{k})(Y,Z) - (\tilde{\nabla}_Y \tilde{k})(X,Z), \tag{6.50}$$

where $R \in \mathfrak{X}^{(3,1)}(M)$ and $\tilde{R} \in \mathfrak{X}^{(3,1)}(\tilde{M})$ are the Riemann curvature tensor for the Levi-Civita connection ∇ on TM and $\tilde{\nabla}$ on $T\tilde{M}$, respectively. The Codazzi relation (6.50) is the same for the Riemannian and the Lorentzian cases. These equations follow from two computations, which we perform for the Lorentzian case,[99] i.e. using (6.43). The first is:

$$\begin{aligned}
\nabla_X \nabla_Y Z &= \nabla_X (\tilde{\nabla}_Y Z - \tilde{k}(Y,Z)N) \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z - \tilde{k}(X, \tilde{\nabla}_Y Z)N - X(\tilde{k}(Y,Z)) \cdot N - \tilde{k}(Y,Z)\nabla_X N \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z + W(X)\tilde{k}(Y,Z) - (\tilde{k}(X, \tilde{\nabla}_Y Z) + X(\tilde{k}(Y,Z)))N.
\end{aligned} \tag{6.51}$$

The second computation, which uses torsion-freeness of $\tilde{\nabla}$, i.e.

$$\tilde{\nabla}_X Y - \tilde{\nabla}_Y X = [X,Y], \tag{6.52}$$

is

$$\begin{aligned}
\nabla_{[X,Y]} Z &= \tilde{\nabla}_{[X,Y]} Z - \tilde{k}([X,Y],Z)N \\
&= \tilde{\nabla}_{[X,Y]} Z - (\tilde{k}(\tilde{\nabla}_X Y, Z) - \tilde{k}(\tilde{\nabla}_Y X, Z))N.
\end{aligned} \tag{6.53}$$

The definition (3.6) of curvature, combined with the 'covariant Leibniz rule'

$$X(\tilde{k}(Y,Z)) = (\tilde{\nabla}_X \tilde{k})(Y,Z) + \tilde{k}(\tilde{\nabla}_X Y, Z) + \tilde{k}(Y, \tilde{\nabla}_X Y), \tag{6.54}$$

which is a special case of (2.54),[100] then yields, after some neat cancellations,

$$\begin{aligned}
\Omega(X,Y)Z &= (\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})Z \\
&= \tilde{\Omega}(X,Y)Z + W(X)\tilde{k}(Y,Z) - W(Y)\tilde{k}(X,Z) \\
&\quad + ((\tilde{\nabla}_Y \tilde{k})(X,Z) - (\tilde{\nabla}_X \tilde{k})(Y,Z))N.
\end{aligned} \tag{6.55}$$

Taking the (metric) inner product with W and using (6.46) yields Gauß's equation (6.49), whereas the inner product with N and using (6.33) yields Codazzi's equation (6.50).

[99] The reader is invited to prove the Riemannian case (6.48) with (6.50) him/herself.

[100] Recall that unlike $\tilde{k}$, the metric is covariantly constant, i.e. $\tilde{\nabla}_X \tilde{g} = 0$ for all $X \in \mathfrak{X}(\tilde{M})$, cf. (2.56).

6.4 Fundamental theorem for hypersurfaces

The preceding material comes to a head in the *fundamental theorem for hypersurfaces*, which was proved (by different means) in the 19th century. We discuss the proof in some detail, since it will turn out to be a good preparation for the 3+1 split of the Einstein equations later on.[101]

Theorem 20 *Let $(\tilde{M},\tilde{g})$ be a connected and simply connected m-dimensional Riemann manifold equipped with a second tensor $\tilde{k}\in\mathfrak{X}^{(2,0)}(\tilde{M})$ satisfying the Gauß–Codazzi equations*

$$\tilde{R}(W,Z,X,Y)+\tilde{k}(W,Y)\tilde{k}(X,Z)-\tilde{k}(W,X)\tilde{k}(Y,Z)=0; \tag{6.56}$$

$$(\tilde{\nabla}_X\tilde{k})(Y,Z)-(\tilde{\nabla}_Y\tilde{k})(X,Z)=0. \tag{6.57}$$

Then there exists an isometric embedding $F:\tilde{M}\to\mathbb{R}^{m+1}$ for which the second fundamental form is the given tensor $\tilde{k}$, and such an embedding is unique up to Euclidean motions (i.e. up to isometries, which are combinations of translations and rotations).

For general $\tilde{M}$ the above theorem holds at least locally, in that any $u_0\in\tilde{M}$ has a connected and simply connected neighbourhood $U\in\mathscr{O}(\tilde{M})$ for which the above claims hold.

Note that (6.56) - (6.57) arise from (6.48) - (6.50) by putting $R=0$ (because $\mathbb{R}^{m+1}$ is equipped with the flat Euclidean metric), and have (6.23) - (6.24) as their coordinate version. The latter were admittedly written down and derived for $m=2$, but simply letting the indices α,β etc. run from 1 to m rather than from 1 tot 2 immediately generalizes our treatment of the classical theory of surfaces to any dimension (alas with some loss to visualisability).

We just prove the local version of Theorem 20 by PDE methods, which is enough to make our point, namely showing the role of the Gauß–Codazzi equations as integrability conditions.

Let us initially assume we found an $F:U\to\mathbb{R}^{m+1}$ satisfying the conditions in the theorem. Its uniqueness may be reformulated as the conjunction of the following local conditions:

1. For arbitrary $x_0\in\mathbb{R}^{m+1}$, the map F satisfies $F(u_0)=x_0$;
2. For some fixed orthonormal basis $(e_1,\ldots,e_m)$ of $T_{u_0}\tilde{M}$ and an arbitrary orthonormal basis $(f_1,\ldots,f_{m+1})$ of $T_{x_0}\mathbb{R}^{m+1}\cong\mathbb{R}^{m+1}$, its derivative satisfies $F'_{u_0}(e_\alpha)=f_\alpha$ $(\alpha=1,\ldots,m)$.

Without loss of generality we may choose geodesic normal coordinates on U relative to u_0, cf. (3.52) - (3.53), so that $e_\alpha=\partial_\alpha\equiv\partial/\partial u^\alpha$ is indeed orthonormal at least at u_0. Furthermore, we may pick coordinates (x^i) on $\mathbb{R}^{m+1}$ $(i=1,\ldots,m+1)$ such that $f_i=\partial/\partial x^i$ for $i=1,\ldots,m$. The components $F^i(u^\alpha)$ of $F:U\to\mathbb{R}^{m+1}$ then satisfy the (initial) condition

$$\frac{\partial F^i}{\partial u^\alpha}(u_0)=\delta^i_\alpha\ \ (\alpha=1,\ldots,m,\,i=1,\ldots,m); \tag{6.58}$$

$$\frac{\partial F^{m+1}}{\partial u^\alpha}(u_0)=0\ \ (\alpha=1,\ldots,m). \tag{6.59}$$

In addition to F, we have to define a normal field $\vec{N}$ on U, whose components N^i satsify

$$N^i(u_0)=0\ \ (i=1,\ldots,m); \tag{6.60}$$

$$N^{m+1}(u_0)=1. \tag{6.61}$$

[101] Our proof is based on Kobayashi & Nomizu, Vol. 2, §VII.7.

If we recall (6.4) as well as (2.50), whose asterisk we omit, for each $i = 1,\dots,m+1$ we have

$$(\tilde{\nabla}_{\partial/\partial u^\alpha}\, dF^i)_\beta = x^i_{\alpha\beta} - \tilde{\Gamma}^\gamma_{\alpha\beta} x^i_\gamma, \tag{6.62}$$

so that, introducing 1-forms $\theta^i \in \Omega(U)$ for each $i = 1,\dots,m+1$ via

$$\theta^i = dF^i, \tag{6.63}$$

Using the notation $\tilde{\nabla}_\alpha = \tilde{\nabla}_{\partial/\partial u^\alpha}$, Gauß's equation (6.19) for $(\vec{x}_\alpha)$ is then equivalent to

$$(\tilde{\nabla}_\alpha \theta^i)_\beta = \tilde{k}_{\alpha\beta} N^i \ \ (\alpha,\beta = 1,\dots,m). \tag{6.64}$$

Conversely, if $\theta^i \in \Omega(U)$ satisfies (6.64), then there exists $F^i \in C^\infty(U)$ such that (6.63) holds. We start with a computation which is valid for any $\theta^i \in \Omega(U)$ and uses the Leibniz rule (2.54):[102]

$$\begin{aligned} d\theta^i(X,Y) &= X(\theta^i(Y)) - Y(\theta^i(X)) - \theta^i([X,Y]) \\ &= (\tilde{\nabla}_X \theta^i)(Y) + \theta^i(\tilde{\nabla}_X Y) - (\tilde{\nabla}_Y \theta^i)(X) - \theta^i(\tilde{\nabla}_Y X) - \theta^i([X,Y]) \\ &= (\tilde{\nabla}_X \theta^i)(Y) - (\tilde{\nabla}_Y \theta^i)(X) + \theta^i(\tau(X,Y)) \\ &= (\tilde{\nabla}_X \theta^i)(Y) - (\tilde{\nabla}_Y \theta^i)(X), \end{aligned} \tag{6.65}$$

since the Levi-Civita connection $\tilde{\nabla}$ is torsion-free, cf. (2.38). Eq. (6.64), then gives

$$d\theta^i(\partial_\alpha,\partial_\beta) = (\tilde{\nabla}_\alpha \theta^i)(\partial_\beta) - (\tilde{\nabla}_\beta \theta^i)(\partial_\alpha) = N^i(\tilde{k}_{\alpha\beta} - \tilde{k}_{\beta\alpha}) = 0 \tag{6.66}$$

by symmetry of the second fundamental form $\tilde{k}$. The Poincaré Lemma then gives (6.63).

It is convenient to replace the 1-forms θ^i by the corresponding vector fields $Z^i = \sharp(\theta^i)$ on U $(i = 1,\dots,m+1)$, in terms of which (6.64) becomes (writing Z^β_i for $(Z^i)^\beta$)

$$\frac{\partial Z^\beta_i}{\partial u^\alpha} + \tilde{\Gamma}^\beta_{\alpha\gamma} Z^\gamma_i = N^i \tilde{k}^\beta_\alpha, \tag{6.67}$$

Similarly, in terms of Z_i Weingarten's equation (6.20) becomes

$$\frac{\partial N^i}{\partial u^\alpha} = -\tilde{k}_{\alpha\beta} Z^\beta_i, \tag{6.68}$$

We may rewrite the coupled PDE's (6.67) and (6.68) on U, $i = 1,\dots,m+1$, more elegantly as

$$\tilde{\nabla}_X Z^i = N^i W(X); \tag{6.69}$$

$$XN^i = -\tilde{k}(X,Z^i), \tag{6.70}$$

for $X \in \mathfrak{X}(U)$ and $N^i \in C^\infty(U)$, subject to the initial conditions (6.60) - (6.61) for N^i, as well as

$$Z^\alpha_i(u_0) = \delta^\alpha_i \ \ (\alpha = 1,\dots,m,\, i = 1,\dots,m); \tag{6.71}$$

$$Z^\alpha_{m+1}(u_0) = 0 \ \ (\alpha = 1,\dots,m). \tag{6.72}$$

[102] In the first line we use the identity $d\omega(X,Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X,Y])$, valid for any $\omega \in \Omega(U)$.

We derived (6.69) - (6.70) with (6.71) - (6.72) from the existence of $F : U \to \mathbb{R}^{m+1}$ with the desired properties (as stated in the theorem). Conversely, if we can solve these equations for Z^i (and N^i), we may construct F via $\theta^i = \flat(Z^i)$ and (6.63), having the right properties.

We now show that this can be done. To begin with, we show that the integrability conditions for (6.69) - (6.70) are the Gauß–Codazzi equations (which should come as no surprise, since (6.69) - (6.70) are a version of the Gauß–Weingarten equations). From (6.70) we derive both

$$[X,Y]N^i = -X\tilde{k}(Y,Z^i) + Y\tilde{k}(X,Z^i); \tag{6.73}$$

$$[X,Y]N^i = -\tilde{k}([X,Y],Z^i). \tag{6.74}$$

so that $X\tilde{k}(Y,Z^i) - Y\tilde{k}(X,Z^i) = \tilde{k}([X,Y],Z^i)$; a computation very similar to (6.65) then rewrites this as Codazzi's eq. (6.57). Similarly, practically the same computation as (6.51) - (6.55), using (6.57), shows that (6.69) implies Gauß's eq. (6.56). Thus the Gauß–Codazzi equations are *necessary* for the solvability of (6.69) - (6.70), which explains their role in Theorem 20.

To show that they are also *sufficient*, we have to make our hands dirty (as usual in PDE theory). We take geodesic normal coordinates (u^α) relative to $u_0 \in U$ (it may be necessary to shrink U in order to make it a normal nbhd) and some fixed orthonormal basis $(e_1,\dots,e_m)$ of $T_{u_0}\tilde{M}$, so that the coordinates $(u^1,\dots,u^m)$ specify the point $u = \gamma_{\vec{u}}(1)$, where $\gamma_{\vec{u}}$ is the (unique) geodesic having $\gamma_{\vec{u}}(0) = u_0$ and $\dot{\gamma}_{\vec{u}}(0) = u^\alpha e_\alpha$ (summation convention!), as usual.

For *fixed* $u \in U$, define a vector field Z^i and a function N^i along this geodesic $\gamma_{\vec{u}}$ by solving

$$\tilde{\nabla}_{\dot{\gamma}_{\vec{u}}} Z^i = N^i W(\dot{\gamma}_{\vec{u}}); \tag{6.75}$$

$$\dot{\gamma}_{\vec{u}} N^i = -\tilde{k}(\dot{\gamma}_{\vec{u}}, Z^i), \tag{6.76}$$

at least for $t \in [0,1]$, or, in coordinates, where $Z^i = (Z^1_i,\dots,Z^m_i)$ as above, and $tu = \gamma_{\vec{u}}(t)$,

$$\frac{dZ^\beta_i(t)}{dt} + u^\gamma \tilde{\Gamma}^\beta_{\gamma\alpha}(tu) Z^\alpha_i(t) = N^i(t)\tilde{k}^\beta_\alpha(tu)u^\alpha; \tag{6.77}$$

$$\frac{dN_i(t)}{dt} = -\tilde{k}_{\alpha\beta}(tu)u^\alpha Z^\beta_i, \tag{6.78}$$

with initial conditions $Z^\alpha_i(0) = \delta^\alpha_i$ $(i \le m)$, $Z^\alpha_{m+1}(0) = 0$, $N^i(0) = 0$ for $i \le m$, and $N_{m+1}(0) = 1$, cf. (6.71) - (6.72) and (6.60) - (6.61). Here we identified $Z^i(t)$ with $Z^i(tu)$, etc. These solutions exist and are unique by standard ODE theory. Finally, define $Z^i \in \mathfrak{X}(U)$ and $N^i \in C^\infty(U)$ by

$$Z^i(u) = Z^i(1); \tag{6.79}$$

$$N^i(u) = N^i(1), \tag{6.80}$$

where of course the Z^i and N^i on the right-hand side depend on u by construction. We claim that this pair (Z^i,N^i) solves (6.69) - (6.70) with the right initial conditions (6.71) - (6.72) and (6.60) - (6.61). To prove this, it is convenient to introduce two *constant* vector fields on U by

$$X = \partial_\alpha \ \ (\alpha = 1,\dots,m); \tag{6.81}$$

$$Y = a^\alpha \partial_\alpha, \tag{6.82}$$

where $(a^1,\dots,a^n)$ are the normal coordinates of some *fixed* $a \in U$. The equations

$$\tilde{\nabla}_Y Z^i = N^i W(Y); \tag{6.83}$$

$$YN^i = -\tilde{k}(Y,Z^i) \tag{6.84}$$

then hold along the geodesic $\gamma_{\vec{a}}(t)$ for $t \in [0,1]$, since there they coincide with (6.75) - (6.76).

We claim that along $\gamma_{\vec{a}}(t)$ the functions (Z^i, N^i) defined by (6.79) - (6.80) also satisfy

$$\tilde{\nabla}_Y(\tilde{\nabla}_X Z^i - N^i W(X)) = (XN^i + \tilde{k}(X,Z^i))W(Y); \tag{6.85}$$

$$Y(XN^i + \tilde{k}(X,Z^i)) = -\tilde{k}(Y, \tilde{\nabla}_X Z^i - N^i W(X)), \tag{6.86}$$

which equations are none other than (6.83) - (6.84) with the substitutions

$$Z^i \rightsquigarrow \tilde{\nabla}_X Z^i - N^i W(X); \tag{6.87}$$

$$N^i \rightsquigarrow XN^i + \tilde{k}(X,Z^i). \tag{6.88}$$

Now note that the initial conditions to (6.85) - (6.86) follow from those to (6.83) - (6.84), viz.

$$\tilde{\nabla}_X Z^i(u_0) - N(u_0)^i W_{u_0}(X) = 0; \tag{6.89}$$

$$XN^i(u_0) + \tilde{k}_{u_0}(X,Z^i) = 0. \tag{6.90}$$

Indeed, by the construction of geodesic normal coordinates, *at the point* u_0, the pair (Z^i, N^i) satisfies (6.83) - (6.84) for any Y, and so in particular it does so for X. The point now is that, (6.85) - (6.86) being a first-order system, its unique solution with initial conditions zero is zero, which by (6.87) - (6.88) shows that (Z^i, N^i) solves (6.69) - (6.70), with given initial conditions.

It remains to derive (6.85) - (6.86) from (6.83) - (6.84) and the Gauß-Codazzi equations. The argument should be familiar by now, but here we go! To derive (6.85), we compute

$$\begin{aligned}
\tilde{\nabla}_Y(\tilde{\nabla}_X Z^i - N^i W(X)) &= \tilde{\nabla}_Y \tilde{\nabla}_X Z^i - (YN^i)W(X) - N^i \tilde{\nabla}_Y(W(X)) \\
&= \tilde{\nabla}_X \tilde{\nabla}_Y Z^i - \Omega(X,Y)Z^i - (YN^i)W(X) - N^i((\tilde{\nabla}_Y W)(X) + W(\tilde{\nabla}_Y X)) \\
&= \tilde{\nabla}_X(N^i W(Y)) + \tilde{k}(X,Z^i)W(Y) - \tilde{k}(Y,Z^i)W(X) \\
&\quad - (YN^i)W(X) - N^i((\tilde{\nabla}_Y W)(X) + W(\tilde{\nabla}_Y X)) \\
&= (XN^i + \tilde{k}(X,Z^i))W(Y) + N^i(\tilde{\nabla}_X(W(Y)) - (\tilde{\nabla}_Y W)(X) - W(\tilde{\nabla}_Y X)) \\
&= (XN^i + \tilde{k}(X,Z^i))W(Y),
\end{aligned} \tag{6.91}$$

where we use Gauß in the form (6.55) to pass to the second line, we use (6.84) to cancel the term $\tilde{k}(Y,Z^i)W(X)$ on the previous line, and finally the coefficient of N^i in the penultimate line is zero by Codazzi's equation (6.57), which emerges after using (2.54) to write $\tilde{\nabla}_X(W(Y) = (\tilde{\nabla}_X W)(Y) + W(\tilde{\nabla}_X Y)$, and noting that $W(\tilde{\nabla}_X Y) - W(\tilde{\nabla}_Y X) = W(\tilde{\nabla}_X Y - \tilde{\nabla}_Y X) = 0$ because $\tilde{\nabla}_X Y = \tilde{\nabla}_Y X$, since $\tilde{\nabla}$ is torsion-free and $[X,Y] = 0$ for the constant vector fields (6.81) - (6.82).

Similarly, to derive (6.86), using (6.84), (2.54), Codazzi's (6.57), and (6.83), we compute

$$\begin{aligned}
Y(XN^i + \tilde{k}(X,Z^i)) &= XYN^i + Y\tilde{k}(X,Z^i) = -X\tilde{k}(Y,Z^i) + Y\tilde{k}(X,Z^i) \\
&= (\tilde{\nabla}_Y \tilde{k})(X,Z^i) - (\tilde{\nabla}_X \tilde{k})(Y,Z^i) + \tilde{k}(\tilde{\nabla}_Y X, Z^i) - \tilde{k}(\tilde{\nabla}_X Y, Z^i) \\
&\quad - \tilde{k}(Y, \tilde{\nabla}_X Z^i) + \tilde{k}(X, \tilde{\nabla}_Y Z^i) \\
&= -\tilde{k}(Y, \tilde{\nabla}_X Z^i) + \tilde{k}(X, N^i W(Y)) \\
&= -\tilde{k}(Y, \tilde{\nabla}_X Z^i - N^i W(X)),
\end{aligned} \tag{6.92}$$

since $\tilde{k}(X, W(Y)) = \tilde{k}(Y, W(X))$; in coordinates this is the identity $\tilde{k}_{\alpha\gamma} g^{\gamma\delta} \tilde{k}_{\delta\beta} = \tilde{k}_{\beta\gamma} g^{\gamma\delta} \tilde{k}_{\delta\alpha}$.

7 The Einstein equations as PDE's

In this chapter we transform the Einstein equations (5.1) into a system of six hyperbolic evolution equations en four elliptic constraint equations. Some first steps in this direction have already been taken in §5.4, especially the introduction of a suitable gauge condition, but to really get the analysis going we need a so-called 3 +1 split of space-time. This split is far from unique and understanding this non-uniqueness is an essential part of the analysis.[103]

We initially assume we already *have* a space-time (M,g), where g solves the Einstein equations (for some energy-momentum tensor). We then choose a spacelike Cauchy (hyper)surface $\Sigma \subset M$, and apply the theory of the previous chapter, obtaining initial data $(\tilde{g},\tilde{k})$ on Σ, constrained by the Gauß-Codazzi equations. In this chapter we investigate the implications of the assumption that g solves the Einstein equations; this amounts to rewriting these abstract generally covariant equations in $d = 4$ in terms of concrete non-covariant $3+1$-dimensional data.

7.1 Lapse and shift

> 'First we must step back and note that the problem of picking an appropriate coordinate system typically is split into two parts: choosing a time slicing (i.e., a time coordinate), and picking a spatial gauge (i.e., spatial coordinates). The time slicing determines what shape the spatial slices Σ_t take in the enveloping spacetime. The lapse L determines how the shape of the slices Σ_t changes in time, since it relates the advance of proper time to coordinate time along the normal vector N^μ connecting one spatial slice to the next. Picking a time slicing or a time coordinate therefore amounts to making a choice for the lapse function. Letting the lapse vary with position across the spatial slice takes advantage of the freedom that proper time can advance at different rates at different points on a given slice. The shift S^i, on the other hand, determines how spatial points at rest with respect to a normal observer N^μ are relabeled on neighboring slices. The spatial gauge or spatial coordinates is therefore imposed by a choice for the shift vector.'
> (Baumgarte & Shapiro, *Numerical Relativity*, p. 88).

We assume (M,g) is globally hyperbolic and hence has a Cauchy surface (see Definition 13) as well as a time function $\mathsf{t} : M \to \mathbb{R}$ for which $g(\nabla\mathsf{t},\nabla\mathsf{t}) < 0$; see §3.7.2 and the proof of Theorem 15. We already introduced the ***lapse function*** $L = 1/\sqrt{-g(\nabla\mathsf{t},\nabla\mathsf{t})}$ and the associated normalized timelike vector field $N = -L\nabla\mathsf{t}$, cf. (3.118) - (3.119), so that $g(N,N) = -1$, and N is normal to any (necessarily spacelike) hypersurface Σ_t, see (4.24): indeed, of $X \in T_x\Sigma_t$, then

$$g(X,\nabla\mathsf{t}) = (d\mathsf{t})(X) = X\mathsf{t} = 0, \tag{7.1}$$

since t is constant along Σ_t. In what follows, we take $\Sigma = \Sigma_0$ (or any other fixed value of t).

We now choose coordinates (x^0,x^1,x^2,x^3) adapted to the foliation $M = \cup_t\Sigma_t$, in that $x^0 = t$ (more precisely, $x^0(x) = t$ provided $x \in \Sigma_t$), and the x^i are (typically local) coordinates initially on Σ ($i = 1,2,3$), but subsequently on any slice Σ_t, since if $y \in \Sigma_t$, the flow line of the vector field $\nabla\mathsf{t}$ (or N) hits Σ in exactly one point $x_0 \in \Sigma$; if the latter has coordinates $x_0 = (0,x^1,x^2,x^3)$, the former is assigned coordinates $y = (t,x^1,x^2,x^3)$. This construction also gives a diffeomorphism $M \cong \mathbb{R}\times\Sigma$, which maps $y \in \Sigma_t$ to $(t,\vec{x})$, where $(0,\vec{x}) \in \Sigma$ is related to y as just explained.

[103] A good reference for this chapter is Gourgoulhon (whose $\vec{m}$ is our e_0), though we rewrote all his calculations.

Given (local) spatial coordinates (x^1,x^2,x^3) on Σ any $x \in \Sigma$ one then has natural (local) tangent vectors $e_i = \partial_i$ to Σ, as well as a natural one-form $\theta^0 = d\mathsf{t}$. The completion of (e_i) to a basis by adding (∂_0) is somewhat defective, in that the latter may not be orthogonal to Σ. To correct for this one introduces a ***shift vector*** $S = S^i\partial_i$ (sums over i,j are of course from 1 to 3), where the S^i are called the ***shift functions***, such that $e_0 = \partial_0 - S$ *is* orthogonal to Σ. We then have a frame (e_a) with dual coframe (θ^b), that is, $\theta^a(e_b) = \delta^a_b$ for $a,b = 0,1,2,3$, defined by

$$e_0 = \partial_t - S^i\partial_i; \qquad e_i = \partial_i; \tag{7.2}$$

$$\theta^0 = dt; \qquad \theta^i = dx^i + S^i dt. \tag{7.3}$$

. By definition of the lapse and the shift, we then have the useful relations

$$g = -L^2(\theta^0)^2 + \tilde{g}_{ij}\theta^i\theta^j; \tag{7.4}$$

$$e_0 = LN = -L^2\nabla\mathsf{t}; \tag{7.5}$$

$$d\mathsf{t} = dt; \tag{7.6}$$

$$\nabla\mathsf{t} = g^{\mu 0}\partial_\mu; \tag{7.7}$$

$$L = 1/\sqrt{-g^{00}}; \tag{7.8}$$

$$S^i = -g^{i0}/g^{00}; \tag{7.9}$$

$$N_\mu = (-L,0,0,0); \tag{7.10}$$

$$N^\mu = (1/L, -S/L). \tag{7.11}$$

Consequently, the metric and its inverse take the form

$$g_{\mu\nu} = \begin{pmatrix} -L^2 + S_jS^j & S_1 & S_2 & S_3 \\ S_1 & \tilde{g}_{11} & \tilde{g}_{12} & \tilde{g}_{13} \\ S_2 & \tilde{g}_{21} & \tilde{g}_{22} & \tilde{g}_{23} \\ S_3 & \tilde{g}_{31} & \tilde{g}_{32} & \tilde{g}_{33} \end{pmatrix} \equiv \begin{pmatrix} -L^2 + S_jS^j & S_i \\ S_i & \tilde{g}_{ij} \end{pmatrix}; \tag{7.12}$$

$$g^{\mu\nu} = \begin{pmatrix} -1/L^2 & S^i/L^2 \\ S^i/L^2 & \tilde{g}^{ij} - S^iS^j/L^2 \end{pmatrix}, \tag{7.13}$$

where $\tilde{g}^{ij}$ is the matrix inverse to $\tilde{g}_{ij}$ and spatial indices are raised and lowered with these spatial metric (so that e.g. $S_jS^j = \tilde{g}_{ij}S^iS^j$). Thus L and S^i may also simply be seen as parametrizations of the non-spatial components of the metric. In particular, $S^i = 0$, which is possible even globally, as shown only relatively recently,[104] corresponds to $g^{i0} = g_{0i} = 0$, so that the metric g assumes a block diagonal form. If, in addition, $L = 1$, then $g^{00} = g_{00} = -1$ but this choice is generally *not* globally possible; we will see shortly that the flow lines of the vector field would be geodesics in that case, whose focusing and hence crossing (in the presence of positive curvature) obviously invalidates the underlying coordinate system. See also the end of this section.

In switching between four-dimensional and $3+1$-dimensional arguments and computations, it turns out to be convenient to have a $4d$-version of the $3d$-objects $\tilde{g}$ and $\tilde{k}$ defined on Σ (and indeed on each hypersurface Σ_t). These are given in any coordinates by, cf. (3.122) - (3.122),[105]

$$\tilde{g}_{\mu\nu} = g_{\mu\nu} + N_\mu N_\nu; \tag{7.14}$$

$$\tilde{k}_{\mu\nu} = -\tilde{g}^\rho_\mu\tilde{g}^\sigma_\nu\nabla_\rho N_\sigma. \tag{7.15}$$

[104] See the second paper by Bernal and Sanchez cited in footnote 70.

[105] Note the minus sign in (7.15) compared to (3.123), which is a consequence of different conventions in fluid mechanics and differential geometry.

Note that indices are raised and lowered with g, so that $\tilde{g}^\nu_\mu = \delta^\nu_\mu + N_\mu N^\nu$ (also called h^ν_μ), taken at $x \in M$, is the matrix of the orthogonal projection operator $\pi_x : T_xM \to T_x\Sigma$ defined by g, cf. (6.36). Unlike the original $\tilde{g} \in \mathfrak{X}^{(2,0)}(\Sigma)$, the new $\tilde{g} \in \mathfrak{X}^{(2,0)}(M)$ is defined on any pair of vectors $X, Y \in T_xM$ ($x \in \Sigma$), though the extension is somewhat trivial in that $\tilde{g}(X,N) = 0$ for any $X, Y \in T_xM$, whilst $\tilde{g}(X,Y)$ defined from (7.14) equals the original $\tilde{g}(X,Y)$ defined from (6.31). Hence the ambiguous notation is admissible and it is always clear which $\tilde{g}$ is meant. Likewise for $\tilde{k}$ in (7.15). In terms of the projection π_x, for all $x \in \Sigma$ and $X, Y \in \mathfrak{X}(M)$ we have (check!)

$$\tilde{g}_x(X,Y) = g(\pi_x(X), \pi_x(Y)); \tag{7.16}$$

$$\tilde{k}_x(X,Y) = k(\pi_x(X), \pi_x(Y)), \tag{7.17}$$

where $k \in \mathfrak{X}^{(2,0)}(M)$ is defined by $k(X,Y) = -g(\nabla_X N, Y)$, or $k_{\mu\nu} = -\nabla_\mu N_\nu$. This also yields

$$\nabla_\mu N_\nu = -\tilde{k}_{\mu\nu} - N_\mu A_\nu, \tag{7.18}$$

where the 'acceleration' A of the vector field N is defined by $A = \nabla_N N$, so that the flow of N is geodesic iff $A = 0$. To prove (7.18), one may separately check the $N-N$, the $N-\Sigma$, the $\Sigma-N$, and the $\Sigma-\Sigma$ contractions. For example, $N^\mu N^\nu \nabla_\mu N_\nu = N^\mu A_\mu$, which equals the right-hand side $-N^\mu N^\nu(-\tilde{k}_{\mu\nu} + N_\mu A_\nu)$, since $\tilde{k}_{\mu\nu} N^\mu N^\nu = 0$ by (7.17) and $N^\mu N_\mu = g(N,N) = -1$. Etc.

We now shed completely new light on the extrinsic curvature $\tilde{k}$ of $\Sigma \subset M$ by showing that

$$\tilde{k} = -\tfrac{1}{2}\mathscr{L}_N \tilde{g} \tag{7.19}$$

$$= -\tfrac{1}{2} L^{-1} \mathscr{L}_{e_0} \tilde{g}, \tag{7.20}$$

seen equalities between symmetric tensors in either $\mathfrak{X}^{(2,0)}(\Sigma)$ or $\mathfrak{X}^{(2,0)}(M)$; in the former case the proof of (7.19) in fact implies that $\mathscr{L}_N \tilde{g} \in \mathfrak{X}^{(2,0)}(\Sigma)$. In arbitrary coordinates, we have

$$\tilde{k}_{\mu\nu} = -\tfrac{1}{2}\mathscr{L}_N \tilde{g}_{\mu\nu}, \tag{7.21}$$

$$= -\tfrac{1}{2} L^{-1} \mathscr{L}_{e_0} \tilde{g}_{\mu\nu}. \tag{7.22}$$

In Σ-adapted coordinates we may restrict to spatial indices: used (7.2) and (1.58), eq. (7.22) is

$$(\partial_t - \mathscr{L}_S)\tilde{g}_{ij} = -2L\tilde{k}_{ij}, \tag{7.23}$$

which is an important step towards the $3+1$ decomposition of the Einstein equations.

To derive (7.19) we first use the $(1,0)$ case of (2.61) with $X = N$ to compute

$$\mathscr{L}_N N_\mu = N^\nu \nabla_\nu N_\mu + (\nabla_\mu N^\nu) N_\nu = N^\nu \nabla_\nu N_\mu \equiv \nabla_N N_\mu, \tag{7.24}$$

since the in second term, $(\nabla_\mu N^\nu) N_\nu$ vanishes because $g(N,N) = N^\nu N_\nu = -1$, for

$$N^\nu \nabla_\mu N_\nu = g(N, \nabla_\mu N) = \tfrac{1}{2}\partial_\mu g(N,N) = \tfrac{1}{2}\partial_\mu(-1) = 0. \tag{7.25}$$

Using this as well as (7.18), the $(2,0)$ case of (2.61) with $X = N$ then gives

$$\mathscr{L}_N(N_\mu N_\nu) = N_\mu \nabla_N N_\nu + N_\nu \nabla_N N_\mu = N_\mu A_\nu + N_\nu A_\mu. \tag{7.26}$$

From (7.14), (2.62), (7.18), and (7.26) we then obtain, at last,

$$\mathscr{L}_N \tilde{g}_{\mu\nu} = \mathscr{L}_N(g_{\mu\nu} + N_\mu N_\nu) = -2\tilde{k}_{\mu\nu} - N_\mu A_\nu - N_\nu A_\mu + N_\mu A_\nu + N_\nu A_\mu = -2\tilde{k}_{\mu\nu}. \tag{7.27}$$

We derive (7.20) from (7.19) using a general fact, namely, writing $A_\mu = N^\nu \nabla_\nu N_\mu$ as before,

$$A_\mu = \tilde{\partial}_\mu(\ln L) = L^{-1}\tilde{g}^\nu_\mu \partial_\nu L, \tag{7.28}$$

where we use the notation $\tilde{\partial}_\mu = \tilde{g}^\nu_\mu \partial_\nu$ for the derivative along Σ.[106] Note that the projection $\tilde{g}^\nu_\mu$ reconfirms that A is tangent to Σ (i.e., orthogonal to N), which we already knew because of $g(N, \nabla_N N) = 0$. Using (7.18) and (7.14), eq. (7.28) is equivalent with

$$\nabla_N N_\nu \equiv N^\mu \nabla_\mu N_\nu = L^{-1}(N^\mu N_\nu \partial_\mu + \partial_\nu)L, \tag{7.29}$$

which we will now prove, using torsion-freeness of ∇, which implies $\nabla_\mu \partial_\nu f = \nabla_\nu \partial_\mu f$ for any $f \in C^\infty(M)$. We write (7.10) as $N_\mu = -L\partial_\mu t$ and compute

$$\begin{aligned} N^\mu \nabla_\mu N_\nu &= -N^\mu \nabla_\mu (L\partial_\nu t) = -N^\mu(\partial_\mu L \partial_\nu t + L\nabla_\nu \partial_\mu t) = L^{-1}N^\mu N_\nu \partial_\mu L - LN^\mu \nabla_\nu (L^{-1}N_\mu) \\ &= L^{-1}N^\mu N_\nu \partial_\mu L - N^\mu N_\mu \partial_\nu L^{-1} - LN^\mu \nabla_\nu N_\mu = L^{-1}(N^\mu N_\nu \partial_\mu + \partial_\nu)L, \end{aligned} \tag{7.30}$$

where we used (7.25). Using (7.5), (1.58), and (7.19), we then compute

$$\begin{aligned} \mathscr{L}_{e_0}\tilde{g}_{\mu\nu} &= \mathscr{L}_{LN}\tilde{g}_{\mu\nu} = L\mathscr{L}_N \tilde{g}_{\mu\nu} + N_\mu \partial_\nu L + N_\nu \partial_\mu L + (\partial_\mu L)N^\rho N_\rho N_\nu + (\partial_\nu L)N^\rho N_\rho N_\mu \\ &= L\mathscr{L}_N \tilde{g}_{\mu\nu} + N_\mu \partial_\nu L + N_\nu \partial_\mu L - N_\nu \partial_\mu L - N_\mu \partial_\nu L \\ &= -2L\tilde{k}_{\mu\nu}. \end{aligned} \tag{7.31}$$

This exemplifies a general phenomenon concerning $\mathscr{L}_{e_0}$: if any tensor $\tau \in \mathfrak{X}^{(k,0)}(M)$ satisfies

$$\tau(X_1, \ldots, X_k) = \tau(\pi(X_1), \ldots, \pi(X_k)), \tag{7.32}$$

i.e., τ is purely spatial, or, equivalently $\tau(X_1, \ldots, X_k) = 0$ if $X_i = N$ for at least one i, then also

$$\mathscr{L}_{e_0}\tau(X_1, \ldots, X_k) = \mathscr{L}_{e_0}\tau(\pi(X_1), \ldots, \pi(X_k)), \tag{7.33}$$

that is, also $\mathscr{L}_{e_0}\tau$ is purely spatial. This most easily follows from the Leibniz rule for $\mathscr{L}$ and hence the case $k = 1$. Since $e_0 = LN$ we may as well derived $(\mathscr{L}_{e_0}\tau)(e_0) = 0$ from the assumption $\tau_{e_0}(e_0) = 0$: using (1.58) and $\mathscr{L}_{e_0}e_0 = [e_0, e_0] = 0$, we obtain

$$(\mathscr{L}_{e_0}\tau)(e_0) = e_0(\tau(e_0)) + \tau(\mathscr{L}_{e_0}e_0) = 0 + 0 = 0.$$

7.2 Beyond Gauß-Codazzi

Further steps towards the $3+1$ decomposition of the Einstein equations involve the Gauß-Codazzi identities (6.49) - (6.50), we which for future use we write in general coordinates as

$$\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta \tilde{g}^\rho_\gamma \tilde{g}^\sigma_\delta R_{\rho\sigma\mu\nu} = \tilde{R}_{\gamma\delta\alpha\beta} + \tilde{k}_{\gamma\alpha}\tilde{k}_{\delta\beta} - \tilde{k}_{\gamma\beta}\tilde{k}_{\alpha\delta}; \tag{7.34}$$

$$\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta \tilde{g}^\rho_\gamma N^\sigma R_{\rho\sigma\mu\nu} = \tilde{\nabla}_\beta \tilde{k}_{\alpha\gamma} - \tilde{\nabla}_\alpha \tilde{k}_{\beta\gamma}, \tag{7.35}$$

where we recall that $(\tilde{g}^\mu_\nu)$ is the matrix of the orthogonal projection of T_xM onto $T_x\Sigma$, so that in Σ-adapted coordinates we may rewrite these expressions as

$$R_{ijkl} = \tilde{R}_{ijkl} + \tilde{k}_{ik}\tilde{k}_{jl} - \tilde{k}_{il}\tilde{k}_{jk}; \tag{7.36}$$

$$R_{i0kl} = \tilde{\nabla}_l \tilde{k}_{ik} - \tilde{\nabla}_k \tilde{k}_{il}, \tag{7.37}$$

[106] This is consistent with notation $\tilde{\nabla}$ for the covariant derivative within Σ defined with respect to $\tilde{g}$ because of (6.37), which in coordinates reads $\tilde{g}^\nu_\mu \nabla_\nu Y^\rho = \tilde{\nabla}_\mu Y^\rho$.

where the left-hand side of (7.37) is only valid for zero shift; more generally, it would be $N^\sigma R_{i\sigma kl}$. The other cases follow from the symmetries (3.23) - (3.25) of R. Thus the Gauß relation describes the value of the Riemann tensor at four orthogonal vectors, whereas the Codazzi relation gives its value at three spatial and one orthogonal direction. In GR these will lead to a geometric formulation of the constraints inherent in the Einstein equations. For the dynamical (evolution) equations we will need the case of two spatial and two orthogonal vectors; unlike the previous two cases, which just rely on the embedding $\Sigma \subset M$, this new case will contain expressions like $\mathscr{L}_{e_0}\tilde{k}$, which unlike the above $\tilde{\nabla}_l \tilde{k}_{ik}$ involves derivatives in the orthogonal direction. Thus the case of two orthogonal vectors relies on the time function, or, equivalently, on the foliation $M = \cup_t \Sigma_t$ (at least near $\Sigma \equiv \Sigma_0$). The ensuing ***Ricci identity*** we need reads[107]

$$R(W,N,X,N) = L^{-1}(\mathscr{L}_{e_0}\tilde{k}(X,W) + \tilde{\nabla}_W \tilde{\nabla}_X L) + \tilde{k}^2(X,W), \tag{7.38}$$

where $X, W \in Tx\Sigma$. In general coordinates this expression reads

$$\tilde{g}^\rho_\alpha \tilde{g}^\mu_\beta N^\sigma N^\nu R_{\rho\sigma\mu\nu} = L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_\alpha \tilde{\nabla}_\beta L) + \tilde{k}^2_{\alpha\beta}, \tag{7.39}$$

where $\tilde{k}^2_{\alpha\beta} \equiv \tilde{k}_{\alpha\rho}\tilde{k}^\rho_\beta$, in which the indices on $\tilde{k}$ are raised and lowered with either $\tilde{g}$ or g (this does not matter because any action of the terms $N_\mu N_\nu$ in (7.14) contracts to zero on $\tilde{k}$), and $\tilde{\nabla}_\beta L = \tilde{\partial}_\beta L$. In adapted coordinates (with in addition a zero shift vector, as before), this is

$$R_{i0j0} = L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{ij} + \tilde{\nabla}_i \tilde{\nabla}_j L) + \tilde{k}_{il}\tilde{k}^l_j. \tag{7.40}$$

To derive (7.39), we first note that (7.18) and (7.28) give

$$\nabla_\mu N_\nu = -\tilde{k}_{\mu\nu} - N_\mu \tilde{\partial}_\mu(\ln L). \tag{7.41}$$

As in the derivation of the Gauß–Codazzi equations, we start from (3.10), this time with $Z = N$:

$$\begin{aligned} R^\rho_{\sigma\mu\nu}N^\sigma &= (\nabla_\mu\nabla_\nu - \nabla_\nu\nabla_\mu)N^\sigma = -\nabla_\mu(\tilde{k}^\rho_\nu + N_\nu\tilde{\partial}^\rho L) + \nabla_\nu(\tilde{k}^\rho_\mu + N_\mu\tilde{\partial}^\rho L) \\ &= \nabla_\nu\tilde{k}^\rho_\mu - \nabla_\mu\tilde{k}^\rho_\nu + (\nabla_\nu N_\mu - \nabla_\mu N_\nu)\tilde{\partial}^\rho L + (N_\mu\nabla_\nu - N_\nu\nabla_\mu)\tilde{\partial}^\rho L. \end{aligned} \tag{7.42}$$

This gives

$$N^\sigma N^\nu R_{\rho\sigma\mu\nu} = \nabla_N \tilde{k}_{\rho\mu} - N^\nu \nabla_\mu \tilde{k}_{\rho\nu} + \tilde{\partial}_\mu(\ln L)\tilde{\partial}_\rho(\ln L) + \nabla_\mu \tilde{\partial}_\rho L + N_\mu \nabla_N \tilde{\partial}_\rho L, \tag{7.43}$$

whose last term will vanish upon contraction with $\tilde{g}^\mu_\beta$ in (7.39). We rewrite the second term $N^\nu\nabla_\mu\tilde{k}_{\rho\nu}$ using the fact that $N^\nu\tilde{k}_{\rho\nu} = 0$ and hence also $\nabla_\mu(N^\nu\tilde{k}_{\rho\nu}) = 0$. This gives

$$-N^\nu\nabla_\mu\tilde{k}^\rho_\nu = \tilde{k}^\rho_\nu\nabla_\mu N^\nu = -\tilde{k}^\rho_\nu\tilde{k}^\nu_\mu - \tilde{k}^\rho_\nu N_\mu\tilde{\partial}^\nu(\ln L), \tag{7.44}$$

whose last term will disappear upon contraction with $\tilde{g}^\mu_\beta$ in (7.39). We now replace the covariant derivative in the first term $\nabla_N\tilde{k}_{\rho\mu}$ by a Lie derivative. Our favorite rule (2.61) gives

$$\mathscr{L}_{e_0}\tilde{k}_{\rho\mu} = \nabla_{e_0}\tilde{k}_{\rho\mu} + (\nabla_\mu e^\nu_0)\tilde{k}_{\rho\nu} + (\nabla_\rho e^\nu_0)\tilde{k}_{\mu\nu}, \tag{7.45}$$

in which on the right-hand side we substitute $e_0 = LN$ and hence $\nabla_{e_0} = L\nabla_N$ (recall that unlike the Lie derivative $\mathscr{L}_X$, the covariant derivative ∇_X is $C^\infty(M)$-linear in X). In the remaining

[107] This relation is better named after Darmois (1927) and ADM (i.e. Arnowitt, Deser, and Misner, 1962).

terms we use (7.41). Many of the ensuing terms drop out after contraction with $\tilde{g}^{\rho}_{\alpha}\tilde{g}^{\mu}_{\beta}$, and after a lengthy but straightforward computation we obtain

$$\tilde{g}^{\rho}_{\alpha}\tilde{g}^{\mu}_{\beta}\nabla_N\tilde{k}_{\rho\mu} = L^{-1}\nabla_{e_0}\tilde{k}_{\alpha\beta} + 2\tilde{k}^2_{\alpha\beta}. \tag{7.46}$$

Using (7.45) and (7.46) in (7.43) finally gives (7.38), as follows:

$$\begin{aligned}\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}N^{\sigma}N^{\nu}R_{\rho\sigma\mu\nu} &= L^{-1}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + 2\tilde{k}^2_{\alpha\beta} - \tilde{k}^2_{\alpha\beta} + \tilde{\partial}_{\alpha}(\ln L)\tilde{\partial}_{\beta}(\ln L) + \tilde{\nabla}_{\alpha}\tilde{\partial}_{\beta}L \\ &= L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}L) + \tilde{k}^2_{\alpha\beta}. \end{aligned}\tag{7.47}$$

Subsequently, for the Einstein equations we do not need the full Riemann tensor $R_{\rho\sigma\mu\nu}$ but its contractions $R_{\mu\nu} = R^{\rho}_{\mu\rho\nu} = g^{\rho\sigma}R_{\rho\mu\sigma\nu}$ (the Ricci tensor) and $R = g^{\mu\nu}R_{\mu\nu}$ (the Ricci scalar). The corresponding contracted Gauß relations easily follow from (7.34), and are given by

$$\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}R_{\mu\nu} + \tilde{g}^{\sigma}_{\alpha}\tilde{g}^{\nu}_{\beta}N^{\mu}N^{\rho}R_{\rho\sigma\mu\nu} = \tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - \tilde{k}^2_{\alpha\beta}; \tag{7.48}$$

$$R + 2N^{\mu}N^{\nu}R_{\mu\nu} = \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - \mathrm{Tr}\,(\tilde{k}^2), \tag{7.49}$$

where we used the identity $\tilde{g}^{\alpha\gamma}\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\rho}_{\gamma} = \tilde{g}^{\rho\mu} = g^{\rho\mu} + N^{\rho}N^{\mu}$, and wrote

$$\mathrm{Tr}\,(\tilde{k}) = \tilde{k}^{\mu}_{\mu} = g^{\mu\nu}\tilde{k}_{\mu\nu} = \tilde{g}^{\mu\nu}\tilde{k}_{\mu\nu}; \tag{7.50}$$

$$\mathrm{Tr}\,(\tilde{k}^2) = \tilde{g}^{\mu\nu}\tilde{k}^2_{\mu\nu} = \tilde{g}^{\mu\nu}\tilde{k}_{\mu\rho}\tilde{k}^{\rho}_{\nu} = \tilde{g}^{\mu\nu}\tilde{g}^{\rho\sigma}\tilde{k}_{\mu\rho}\tilde{k}_{\nu\sigma}. \tag{7.51}$$

In adapted coordinates with vanishing shift, where $\mathrm{Tr}\,(\tilde{k}) = \tilde{g}^{ij}\tilde{k}_{ij}$ etc., these relations would be

$$R_{ij} + R_{0i0j} = \tilde{R}_{ij} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{ij} - \tilde{k}^2_{ij}; \tag{7.52}$$

$$R + 2R_{00} = \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - \mathrm{Tr}\,(\tilde{k}^2). \tag{7.53}$$

Similarly, the contracted Codazzi relations (which stop at one stage) follow from (7.35) as

$$N^{\mu}\tilde{g}^{\nu}_{\alpha}R_{\mu\nu} = \tilde{\partial}_{\alpha}\mathrm{Tr}\,(\tilde{k}) - \tilde{\nabla}_{\mu}\tilde{k}^{\mu}_{\alpha}, \tag{7.54}$$

which in the same notation as (7.52) comes down to

$$R_{oi} = \partial_i\mathrm{Tr}\,(\tilde{k}) - \tilde{\nabla}_j\tilde{k}^j_i. \tag{7.55}$$

The contractions of (7.39) are slightly more involved. First, (7.48) and (7.39) give

$$\tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - \tilde{k}^2_{\alpha\beta} - \tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}R_{\mu\nu} = L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}L) + \tilde{k}^2_{\alpha\beta}, \tag{7.56}$$

from which we obtain

$$\tilde{g}^{\mu}_{\alpha}\tilde{g}^{\nu}_{\beta}R_{\mu\nu} = -L^{-1}(\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}L) + \tilde{R}_{\alpha\beta} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\alpha\beta} - 2\tilde{k}^2_{\alpha\beta}. \tag{7.57}$$

Contracting both sides with $\tilde{g}^{\alpha\beta}$, and defining $\tilde{\Delta} = \tilde{g}^{\alpha\beta}\tilde{\nabla}_{\alpha}\tilde{\nabla}_{\beta}$, gives

$$R + N^{\mu}N^{\nu}R_{\mu\nu} = -L^{-1}(\tilde{g}^{\alpha\beta}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} + \tilde{\Delta}L) + \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - 2\mathrm{Tr}\,(\tilde{k}^2), \tag{7.58}$$

Since $\mathscr{L}_{e_0}\tilde{g}_{\alpha\beta} = -2L\tilde{k}_{\alpha\beta}$ by (7.22), we have $\mathscr{L}_{e_0}\tilde{g}^{\alpha\beta} = 2L\tilde{k}^{\alpha\beta}$, cf. (5.13), and hence

$$\tilde{g}^{\alpha\beta}\mathscr{L}_{e_0}\tilde{k}_{\alpha\beta} = \mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) - \tilde{k}_{\alpha\beta}\mathscr{L}_{e_0}\tilde{g}^{\alpha\beta} = \mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) - 2L\mathrm{Tr}\,(\tilde{k}^2), \tag{7.59}$$

where of course $\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) = e_0(\mathrm{Tr}\,(\tilde{k}))$. Hence (7.58) may be rewritten as

$$R + N^{\mu}N^{\nu}R_{\mu\nu} = -L^{-1}(\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) + \tilde{\Delta}L) + \tilde{R} + \mathrm{Tr}\,(\tilde{k})^2. \tag{7.60}$$

Using (7.49), we finally obtain the twice contracted version of (7.39), namely

$$R = \tilde{R} - 2L^{-1}(\mathscr{L}_{e_0}\mathrm{Tr}\,(\tilde{k}) + \tilde{\Delta}L) + \mathrm{Tr}\,(\tilde{k})^2 + \mathrm{Tr}\,(\tilde{k}^2). \tag{7.61}$$

7.3 The 3+1 decomposition of the Einstein equations

We now have all information for projecting the Einstein equations (5.1), with $T_{\mu\nu}$ decomposed according to (5.25), in three different directions, namely, contracting with:[108]

- The *spatial* part $\tilde{g}^\mu_\alpha \tilde{g}^\nu_\beta$, which gives the ***dynamical equations***

$$\mathscr{L}_{e_0}\tilde{k}_{\mu\nu} = -\tilde{\nabla}_\mu\tilde{\nabla}_\nu L + L(\tilde{R}_{\mu\nu} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{\mu\nu} - 2\tilde{k}^2_{\mu\nu} + 4\pi((S-E)\tilde{g}_{\mu\nu} - 2S_{\mu\nu})); \quad (7.62)$$

$$\mathscr{L}_{e_0}\tilde{g}_{\mu\nu} = -2L\tilde{k}_{\mu\nu}. \quad (7.63)$$

  These follow from (5.26), (7.57), (5.27), and (7.22). As already noted, in Σ-adapted coordinates eq. (7.63) becomes (7.23), and with (7.62), one may write the system as

$$(\partial_t - \mathscr{L}_S)\tilde{k}_{ij} = -\tilde{\nabla}_i\tilde{\nabla}_j L + L(\tilde{R}_{ij} + \mathrm{Tr}\,(\tilde{k})\tilde{k}_{ij} - 2\tilde{k}^2_{ij} + 4\pi((S-E)\tilde{g}_{ij} - 2S_{ij})); \quad (7.64)$$

$$(\partial_t - \mathscr{L}_S)\tilde{g}_{ij} = -2L\tilde{k}_{ij}, \quad (7.65)$$

  where, using (1.58) and (2.62), respectively, the two Lie derivatives may be written as

$$\mathscr{L}_S\tilde{k}_{ij} = S^l\partial_l\tilde{k}_{ij} + \tilde{k}_{jl}\partial_i S^l + \tilde{k}_{il}\partial_j S^l; \quad (7.66)$$

$$\mathscr{L}_S\tilde{g}_{ij} = \tilde{\nabla}_i S_j + \tilde{\nabla}_j S_i. \quad (7.67)$$

- The *timelike* part $N^\mu N^\mu$, which gives the so-called ***Hamiltonian constraint***

$$\tilde{R} + \mathrm{Tr}\,(\tilde{k})^2 - \mathrm{Tr}\,(\tilde{k}^2) = 16\pi E, \quad (7.68)$$

  which follows from (5.1) and (7.49); it plays a key role in (canonical) quantum gravity.

- The *mixed* part $\tilde{g}^\mu_\alpha N^\nu$ or $\tilde{g}^\nu_\beta N^\mu$, producing the ***momentum constraint***

$$\tilde{\nabla}_\mu\tilde{k}^\mu_\nu - \tilde{\nabla}_\nu\mathrm{Tr}\,(\tilde{k}) = 8\pi P_\nu, \quad (7.69)$$

  which follows from (5.1), whose $g_{\mu\nu}R$ term contracts to zero, and (7.54). Equivalently,

$$\tilde{g}^{jl}\tilde{\nabla}_l\tilde{k}_{ij} - \tilde{\nabla}_i\mathrm{Tr}\,(\tilde{k}) = 8\pi P_i. \quad (7.70)$$

Altogether, in adapted coordinates, eqs. (7.64), (7.64), (7.68), and (7.70) form a coupled system of 16 PDE's for 16 unknown functions $(\tilde{g}_{ij}, \tilde{k}_{ij}, L, S^i)$ defined on the Cauchy hypersurface Σ, where the $\tilde{k}_{ij}$ may be exchanged for the time-derivaties $\partial_t\tilde{g}_{ij}$ through (7.65), leaving 10 coupled PDE's for 10 unknowns $(\tilde{g}_{ij}, L, S^i)$, similar to the original covariant Einstein equations (which are 10 coupled PDE's for the 10 components $g_{\mu\nu}$ of the four-dimensional metric). In the latter case, the spatial part consists of *six evolution equations*, whereas the other two parts contain only first time derivatives of the spatial metric and no time derivates of the lapse and shift functions at all; hence these act as *four constraints* on the initial data $(\tilde{g}_{ij}, \partial_t\tilde{g}_{ij})$, or, in general, on $(\tilde{g}_{ij}, \tilde{k}_{ij})$. Also cf. §5.4. The lapse and shift functions are not determined by the equations at all and hence can be (more or less) freely chosen; doing so amounts to fixing a (local) gauge,

[108] The letters S and $S_{\mu\nu}$ on the right-hand sides below refer to the energy-momentum tensor, whereas the S in $\mathscr{L}_S$ on the left and the S^i on the right refer to the shift vector, sorry!

which preferably makes the evolution equations hyperbolic, and in addition is favorable for existence and uniqueness results and/or numerical computations.

The most primitive way of doing this is to try $S^i = 1$ $L = 1$, or, equivalently, $g^{0i} = g_{0i} = 0$ and $g^{00} = g_{00} = -1$, which is possible at least locally. This implies that the flow lines to the normal vector field N are geodesics, which, recalling that $A = \nabla_N N$, follows from (7.28). In such coordinates the evolution equations (7.64) - (7.65), in vacuum for simplicity, become

$$\partial_t^2 \tilde{g}_{ij} = \tilde{g}^{kl}(\tilde{g}_{ij,kl} + \tilde{g}_{kl,ij} - \tilde{g}_{lj,ik} - \tilde{g}_{il,jk}) + \cdots, \tag{7.71}$$

where the dots stand for terms with first or no derivatives. Though hyperbolic, this is a poor system because of geodesic focusing, which may lead to coordinate singularities at times where regular solutions might exist (the advantage though is that time equals proper time along N).

A more sophisticated gauge, in the spirit of the analysis in §5.4, is the (covariant) wave gauge $\Box_g x^\mu = 0$, which in the (noncovariant) 3 +1 split yields the lapse and shift functions as

$$(\partial_t - S^j\partial_j)L = -\mathrm{Tr}\,(\tilde{k})L^2; \tag{7.72}$$

$$(\partial_t - S^j\partial_j)S^i = -(\tilde{g}^{ij}\partial_j \ln L + \tilde{g}^{jk}\tilde{\Gamma}^i_{jk})L^2. \tag{7.73}$$

Both analytic and numerical goals (notably stability) are even better served with a combination of the $\mu = 0$ component of this gauge, i.e. $\Box_g x^0 = 0$, and zero shift $S^i = 0$ (called the *harmonic gauge*). Since (7.73) is empty in that case, we are just left with

$$\partial_t L = -\mathrm{Tr}\,(\tilde{k})L^2. \tag{7.74}$$

Another popular choice (especially for stability results) is *maximal slicing*, defined by

$$\mathrm{Tr}\,(\tilde{k}) = 0, \tag{7.75}$$

i.e. each Σ_t has mean zero curvature, which implies that, when Σ_t is compact, it has maximal volume (compared with other slicings). Since $\mathrm{Tr}\,(\tilde{k})$ is just (minus) the expansion θ in the Raychaudhuri equation (3.133) and the singularity theorems, putting it equal to zero clearly prevents at least the focusing mechanism studied in that context and hence avoids coordinate singularities arising through focusing. Eq. (7.72) fixes L through

$$\tilde{\Delta} L = -\mathrm{Tr}\,(\tilde{k}^2)L. \tag{7.76}$$

See the book by Baumgarte & Shapiro cited at the beginning of this chapter for more information on optimal slicings, and Choquet-Bruhat & Ruggeri, Hyperbolicity of the 3 + 1 system of Einstein equations, *Communications in Mathematical Physics* 89, 269–275 (1983) for a proof of hyperbolicity of the Einstein equations in the harmonic gauge.

7.4 Existence and maximality of solutions

In this section we restrict ourselves to the vacuum case $T_{\mu\nu} = 0$; most conclusions survive in the presence of matter, though often at the cost of additional assumptions and complications.

The analysis in §5.4 shows that the Einstein equations $G_{\mu\nu} = 0$ or $R_{\mu\nu} = 0$, seen as PDE's for the components $g_{\mu\nu}$ of the metric, are both overdetermined because of the presence of constraints on the initial data, and underdetermined because of diffeomorphism invariance. The first point was clarified in §7.3 and means that the initial data $(\Sigma, \tilde{g}_{ij}, \tilde{k}_{ij})$, where $(\Sigma, \tilde{g}_{ij})$ is some $3d$ Riemannian manifold and $\tilde{k}_{ij}$ is an additional symmetric tensor on Σ of type $(2,0)$, satisfy the Hamiltonian constraint (7.68), with $E = 0$ in the vacuum case, as well as the momentum constraint (7.70), with $P_i = 0$ in *vacuo*. These constraints will be studied in more detail in §7.5

The traditional strategy to address these problems, due to Choquet-Bruhat, was explained in §5.4: one solves the *reduced* Einstein equations (5.77), which are quasi-linear second-order hyperbolic PDE's, and imposes both the wave gauge condition (5.73) and the (ungauged) constraints at $t = 0$, cf. (5.83) - (5.84). This guarantees that the wave gauge (5.73) as well as the constraints hold at all times–at least for which solutions to (5.77) exist–and hence, by (5.77), also the original Einstein equations (5.1) are solved. This procedure by no means restores uniqueness: any diffeomorphism ψ of M that is the identity before and at $\Sigma \subset M$ but is nontrivial at later times does not change the initial conditions yet $\psi^* g$ solves the Einstein equations, and generally $\psi^* g \neq g$ at later times (of course, $\psi^* g$ will not satisfy the wave gauge).

In any case, this procedure leads to solutions that are *local in space and local in time*:

- Locality in *space* follows from the use of specific coordinates (i.e. those satisfying (5.73).
- Locality in *time* is all the existence theorems for quasi-linear second-order hyperbolic PDE's provide (we discuss the function spaces for solutions in the next chapter).

In what follows, we improve this situation twofold in a way specific to the Einstein equations.

Local existence in *space* turns into global existence in space by globalizing the gauge. This is done as follows. First, a well-known concept in Riemannian geometry is that of a *harmonic map* $h : M \to \hat{M}$ between Riemannian manifolds (M, g) and $(\hat{M}, \hat{g})$, where h is assumed smooth or at least C^2. These maps can be described abstractly, but is is easier (and sufficient for our purposes) to use local coordinates (x^μ) on M and likewise $(\hat{x}^i)$ on $\hat{M}$. Any map $h : M \to \hat{M}$ has an associated *energy* functional, defined by

$$E(h) = \int_M d^3x \sqrt{g(x)} e_x(h); \tag{7.77}$$

$$e_x(h) = \tfrac{1}{2} g^{\mu\nu}(x) \hat{g}_{ij}(h(x)) \frac{\partial h^i(x)}{\partial x^\mu} \frac{\partial h^j(x)}{\partial x^\nu}, \tag{7.78}$$

where h^i are the components of h relative to the $(\hat{x}^i)$. This expression turns out to be independent of the coordinates.[109] For example, if $M = [a, b]$ with flat metric, then $E(f)$ is the energy (2.16) of a curve in N. Another example is $N = \mathbb{R}$ with flat metric, in which case $E(h) = \int_M \nabla h \cdot \nabla h$ is the *Dirichlet integral* of h (which plays a fundamental role in the theory of the Laplace equation $\Delta h = 0$ on M). It can be shown that h extremizes $E(h)$ iff it solves

$$g^{\mu\nu} \left(\frac{\partial^2 h^i(x)}{\partial x^\mu \partial x^\nu} - \Gamma^\rho_{\mu\nu}(x) \frac{\partial h^i(x)}{\partial x^\rho} + \hat{\Gamma}^i_{jk}(h(x)) \frac{\partial h^j(x)}{\partial x^\mu} \frac{\partial h^k(x)}{\partial x^\nu} \right) = 0, \tag{7.79}$$

[109] See e.g. Jost, §8.1.

where and $\Gamma^{\rho}_{\mu\nu}$ and $\hat{\Gamma}^{i}_{jk}$ are the Christoffel symbols for g and $\hat{g}$, respectively. Thus h is called *harmonic* if it solves (7.79). Exactly the same constructions work in Lorentzian geometry, in which case a solution of (7.79) is called a *wave map*. In that case, standard hyperbolic PDE theory yields existence and uniqueness of solutions $h_{|\Sigma}$ and $\dot{h}_{|\Sigma}$ subject to initial conditions on a Cauchy surface in M (where $\dot{h} = \partial_t h$ as defined earlier for $g_{\mu\nu}$).

In order to provide the right version of the wave gauge enabling global solutions in space, we pick some fiducial metric γ on Σ and put the metric $\hat{g} = -dt^2 + \gamma$ on $\hat{M} = \mathbb{R} \times \Sigma$. Anticipating that (M, g) will be globally hyperbolic, we write $M = \mathbb{R} \times \Sigma$, and declare that g satisfies the $\hat{g}$-wave gauge iff the identity map $\mathrm{id} : M \to \hat{M}$ is a wave map (i.e. with respect to g and $\hat{g}$). It follows from the coordinate-independent nature of (7.79) that this condition is coordinate-independent also; one can also see this explicitly by noting that g satisfies the $\hat{g}$-wave gauge iff $\hat{W}^{\mu} = 0$ for each $\mu = 0, 1, 2, 3$, where, cf. (5.73) and (5.75),

$$\hat{W}^{\mu} = g^{\rho\nu}(\hat{\Gamma}^{\mu}_{\rho\nu} - \Gamma^{\mu}_{\rho\nu}). \tag{7.80}$$

Since the difference between two connections (metric or otherwise) is a tensor, this confirms the purely geometric and hence global nature of the $\hat{g}$-wave gauge. In particular, unlike the case of the original wave gauge, the index μ is now a true vector index in that $\hat{W}^{\mu}$ are the components of a vector. Thus the coordinate-dependence of the original wave gauge has been traded for $\hat{g}$-dependence.[110] We now follow the same steps as for the wave gauge, replacing W by $\hat{W}$ from (5.76) till the end of §5.4, with the same concusions: the reduced Einstein equations are quasi-linear and hyperbolic, the gauge and the constraints propagate, etc., with the difference that none of the arguments now depend on the choice of local coordinates on Σ and hence any solution is globally defined (in space). Existence of g (local in time, so on $I \times \Sigma$ for some open interval $I \subset \mathbb{R}$) solving the Einstein equations once again follows from PDE theory (see Theorem 30 in the next chapter), and the appropriate statement of uniqueness is as follows.

To be precise, we write a space-time solving Einstein's equations for given initial data $(\Sigma, \tilde{g}, \tilde{k})$ as $(M, g, \imath)$, where $\imath : \Sigma \hookrightarrow M$ injects the given manifold Σ into M; in particular, $\tilde{g} = \imath^* g$ and $\tilde{k}$ are the first and second fundamental forms of the embedding, respectively.[111] If (M, g) is globally hyperbolic with Cauchy surface $\imath(\Sigma)$, then the triple $(M, g, \imath)$ is called a ***Cauchy development*** (or ***globally hyperbolic development = GHD***) of the initial data $(\Sigma, \tilde{g}, \tilde{k})$, which will be fixed throughout the following discussion (and are always assumed to satisfy the constraints). Note that, as stated at the end of Theorem 30, the 'global in space, local in time' space-times arising from the solution of the Einstein equations in a $\hat{g}$-wave gauge are in fact globally hyperbolic and hence provide Cauchy developments or GHD of the initial data. These solutions arose from a very special procedure, but the general situation is as follows. [112]

Proposition 21 (Geometric uniqueness of solutions of Einstein's equations) *Any two Cauchy developments $(M_1, g_1, \imath_1)$ and $(M_2, g_2, \imath_2)$ of the same (smooth) initial data are locally isometric, in that $\imath_1(\Sigma)$ and $\imath_2(\Sigma)$ have open neighbourhoods U_1 and U_2 in M_1 and M_2, respectively, such that (U_1, g_1) and (U_2, g_2) are isometric through a diffeomorphism $\psi : U_1 \to U_2$ satisfying*

$$\psi^* g_2 = g_1; \tag{7.81}$$

$$\psi \circ \imath_1 = \imath_2. \tag{7.82}$$

[110] If $\Sigma = \mathbb{R}^3$ and $\gamma = \delta$, one recovers the original wave gauge, but only in Euclidean coordinates!

[111] Recall that a ***space-time*** is an oriented and time-oriented connected Lorentzian manifold, cf. Definition 9.

[112] Corollary 16 states: (M, g) *globally hyperbolic* $\Rightarrow M \cong \mathbb{R} \times \Sigma$, but the converse is true as well.

The proof is rather technical (cf. Choquet-Bruhat, Theorem 8.4, p. 168), but the idea is to construct wave maps $h_i : \hat{M} \to M_i$ ($i = 1,2$), suitably shrunk to as to become diffeomorphisms, and define $g_i' = h_i^* g_i$ on $\hat{M}$; this step brings both g_1 and g_2 into the $\hat{g}$ wave gauge. These new metrics then solve the same equations (namely the reduced Einstein equations and the $\hat{g}$ wave gauge condition) with the same initial conditions, and hence they must coincide by local uniqueness result from hyperbolic PDE's. From $g_1' = g_2'$ we obtain $g_2 = (h_1^{-1} \circ h_2)^* g_1 = \psi^* g_1$.

We now come to the (formal) culmination of the 'local in time' approach to the Einstein equations. A *maximal Cauchy development* $(M_{\max}, g_{\max}, \imath_{\max})$ of given (smooth) initial data $(\Sigma, \tilde{g}, \tilde{k})$ is a Cauchy development with the property that for any other Cauchy development $(M, g, \imath)$ of these data there exists an embedding $\psi : M \to M_{\max}$ that preserves time orientation, metric, and Cauchy surface, i.e., one has $\psi^* g_{\max} = g$ and $\psi \circ \imath = \imath_{\max}$, cf. (7.81) - (7.82).

Theorem 22 (Choquet-Bruhat and Geroch) *Each smooth initial data $(\Sigma, \tilde{g}, \tilde{k})$ set satisfying the constraints has a maximal Cauchy development $(M_{\max}, g_{\max}, \imath_{\max})$, which is unique up to time-orientation-preserving isometries fixing the Cauchy surface $\imath(\Sigma) \subset M_{\max}$, as in* (7.82)*.*

Both for understanding the claim and outlining its proof it is useful to rephrase Theorem 22 in terms of partially ordered sets (posets). As mentioned above, Cauchy developments even of fixed initial data are far from unique due to diffeomorphism invariance of the Einstein equations, but we consider two solutions equivalent if they can be transformed onto each other by a diffeomorphism respecting $\imath$ as well as time orientation: thus we say that

$$(M_1, g_1, \imath_1) \cong (M_2, g_2, \imath_2) \tag{7.83}$$

iff there is a time-orientation preserving diffeomorphism $\psi : M_1 \to M_2$ satisfying (7.81) - (7.82). This is an equivalence relation on the set $\mathrm{GHD}(\Sigma, \tilde{g}, \tilde{k})$ of all globally hyperbolic (i.e. Cauchy) developments of the data $(\Sigma, \tilde{g}, \tilde{k})$, and we denote the (quotient) set of its equivalence classes by $[\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k})$. As usual, we write $[M, g, \imath]$ for the equivalence class of $(M, g, \imath)$. Initially, put

$$(M_1, g_1, \imath_1) \leq (M_2, g_2, \imath_2) \tag{7.84}$$

iff there is a *embedding* $\psi : M_1 \to M_2$ such that (7.81) - (7.82) hold (and, being an embedding, ψ must satisfy conditions 1 and 2 in §6.6.1). This fails to be a partial ordering on $\mathrm{GHD}(\Sigma, \tilde{g}, \tilde{k})$ (it fails the antisymmetry axiom), but it does descend to a partial ordering on $[\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k})$, i.e., by abuse of notation we have $[M_1, g_1, \imath_1] \leq [M_2, g_2, \imath_2]$ provided (7.84) holds. This makes $([\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k}) \leq)$ a poset. Recall that a ***top element*** $\top \in P$ of a poset $(P, \leq)$ is an element for which $x \leq \top$ for all $x \in P$; a top is unique if it exists.[113] Theorem 22 then comes down to:

Theorem 23 *The poset $([\mathrm{GHD}](\Sigma, \tilde{g}, \tilde{k}) \leq)$ has a top element (which is necessarily unique).*

In their article in 1969, Choquet-Bruhat and Geroch, sketched a proof based on Zorn's lemma, which they even had to use twice. Since Zorn's Lemma has no place in mathematical physics, we now outline a recent constructive proof due to Sbierski.[114] Note, in particular, that Theorem 23 *implies* that the maximal Cauchy development $(M_{\max}, g_{\max}, \imath_{\max})$ is unique up to isometry.

To get a glimpse of the proof, we first rephrase Proposition 31 in terms of the above poset:

[113] This is not to be confused with a *maximal element* $m \in P$, where for all $x \in P$ one has $m \leq x$ iff $x = m$. Maximal elements are typically non-unique if they exist, and even if they are unique they may not be top elements.

[114] See J. Sbierski, On the existence of a Maximal Cauchy Development for the Einstein Equations - a dezornification, `https://arxiv.org/pdf/1309.7591.pdf`, or *Ann. Henri Poincaré* 17, 3-1–329 (2016). It must be admitted that the set-theoretic complications in this proof, which our outline omits, including weak versions of the Axiom of Choice, make this proof hardly more attractive than the original one by Choquet-Bruhat and Geroch.

Any two Cauchy developments $(M_1, g_1, \imath_1)$ *and* $(M_2, g_2, \imath_2)$ *of given initial data have a common Cauchy development* $(M, g, \imath)$*, in that* $(M, g, \imath) \leq (M_1, g_1, \imath_1)$ *as well as* $(M, g, \imath) \leq (M_2, g_2, \imath_2)$.

Indeed, take $M = U_1$, with $\psi_1 : M \to M_1$ given by the embedding $i : U_1 \subset M_1$, and $\psi_2 : M \to M_2$ defined by $\psi_2 = \psi \circ i$, where ψ is the map from Proposition 31. More strongly, we even have:

Lemma 24 *Any two Cauchy developments* $(M_1, g_1, \imath_1)$ *and* $(M_2, g_2, \imath_2)$ *have a* maximal *common Cauchy development* $(M', g', \imath')$*, in that any other common Cauchy development satisfies*

$$(M, g, \imath) \leq (M', g', \imath'). \tag{7.85}$$

Indeed, if $\{U_\alpha\}$ is the set of all U_1's appearing in Proposition 31, i.e. $U_\alpha \subset M_1$ and given maps $\psi_\alpha : U_\alpha \to M_2$, etc., then one may simply take the union $M' = \cup_\alpha U_\alpha$, with the obvious embedding $M' \subset M_1$, and map $\psi : M' \to M_2$ given by $\psi(x) = \psi_\alpha(x)$ if $x \in U_\alpha$. Conversely:

Lemma 25 *Any two Cauchy developments* $(M_1, g_1, \imath_1)$ *and* $(M_2, g_2, \imath_2)$ *have a common* extension $(M_{12}, g_{12}, \imath_{12})$*, in that* $(M_1, g_1, \imath_1) \leq (M_{12}, g_{12}, \imath_{12})$ *as well as* $(M_2, g_2, \imath_2) \leq (M_{12}, g_{12}, \imath_{12})$.

From this step onwards, the constructions become a bit ugly. Define

$$M_{12} = (M_1 \bigsqcup M_2) / \sim, \tag{7.86}$$

where $M_1 \bigsqcup M_2$ is the disjoint union of M_1 and M_2, and we say that $x \sim y$ if either $x = y$ or $x \in M' \subset M_1$ and $y = \psi(x)$, where $\psi : M' \to M_2$ has just been defined. This space naturally inherits a metric g_{12} from (M_1, g_1) and (M_2, g_2), since for $x \in M_1 \backslash M'$ we may put $g_{12}([x]) = g_1(x)$, for $y \in M_2 \backslash \psi(M')$ we have $g_{12}([y]) = g_2(y)$, noting that $[x] = x$ and $[y] = y$ in those cases, whereas for $x \in M_1$ and $y = \psi(x)$, so that $[x] = [y]$, we put $g_{12}([x]) = g_1(x)$ or $g_2(y)$; these coincide since ψ is an isometry. The obvious maps $M_1 \hookrightarrow M_{12}$ and $M_2 \hookrightarrow M_{12}$ are isometries for g_{12} by constructions. Similarly, we obtain embeddings $\Sigma \hookrightarrow M_{12}$ and $\Sigma \hookrightarrow M_{12}$ from the given ones $\Sigma \hookrightarrow M_1$ and $\Sigma \hookrightarrow M_2$ (the main difficulty in the proof is to show that M_{12} is a Hausdorff space).

The construction of the maximal space-time $M_{\max}$ is an extension of (7.86): one defines

$$M_{\max} = \left(\bigsqcup_\lambda M_\lambda \right) / \sim, \tag{7.87}$$

where $\{M_\lambda\}$ is the set of all Cauchy developments (of the given initial data), and we identify $x \in M_1$ and $y \in M_2$ (where 1 and 2 are generic values of λ) iff $x \sim y$ as defined after (7.86). Also, the constructions of the metric $g_{\max}$, the embedding $\imath_{\max}$, and the (isometric) embeddings $\psi_\lambda : M_\lambda \to M_{\max}$ are entirely similar to the case (7.86) just explained; maximality is obvious.

However, it by no means follows that $(M_{\max}, g_{\max}, \imath_{\max})$ is maximal *as a solution to the vacuum Einstein equations with given initial data*, for this is a weaker notion than a Cauchy development of these data: the difference lies in Σ being a Cauchy surface for (M, g) in the latter case, but not in the former. All that follows is that for any such space-time (M', g') in which $(M_{\max}, g_{\max}, \imath_{\max})$ can be (properly) isometrically embedded, the ensuing copy of Σ in M' (arising from $\Sigma \hookrightarrow M \hookrightarrow M'$) cannot be a Cauchy surface. In particular, $\Sigma \subset M'$ has a nonempty Cauchy horizon, which indicates an end to predictability (at least from the point of view of Σ). The *strong cosmic censorship* hypothesis (or conjecture) excludes this possibility, at least for 'generic' initial data (it cannot always be true). This is a very active area of research.[115]

[115] See Ringström's book for an introduction and Mihalis Dafermos, *The cosmic censorship conjectures in classical general relativity*, `https://www.youtube.com/watch?v=ZBYAbejIvB4`, for a more recent overview (2017).

7.5 Conformal analysis of the constraints: Lichnerowicz equation

The initial value constraints (7.68) - (7.70) may also be analyzed from a PDE point of view. In the simplest case the metric is *static*, which means that (M,g) has a timelike Killing vector field u^μ *and* has a foliation $M = \cup_t \Sigma_t$ whose leaves Σ_t are orthogonal to u^μ (equivalently, $\omega_{\mu\nu} = 0$, see (3.124)). In that case, in the right (i.e. adapted) coordinates the $g_{\mu\nu}$ are time-independent, as for the Minkowski metric or the Schwarzschild solution.[116]. Hence $\tilde{k} = 0$, and if we also assume vacuum for simplicity, then the only constraint on the ensuing initial data $(\Sigma, \tilde{g}_{ij})$ is

$$\tilde{R} = 0. \tag{7.88}$$

This is a vastly underdetermined system, since the 6 independent components of the metric $\tilde{g}_{ij}$ are subject to just one equation, but this doesn't mean that the solution is trivial. This is a problem in pure Riemannian geometry, which was first addressed by Yamabe in 1960. Yamabe argued–a complete proof was only given in 1984 by Schoen–that any Riemannian metric γ on a compact manifold Σ (without boundary) admits a ***conformal rescaling***

$$\tilde{g} = \Omega^4 \gamma, \tag{7.89}$$

where the ***conformal factor*** $\Omega \in C^\infty(\Sigma)$ is strictly positive (so that $\tilde{g}$ is a Riemannian metric on Σ), such that the Ricci scalar $\tilde{R} = \tilde{R}_{\tilde{g}}$ of $\tilde{g}$ is constant.[117] Straightforward computations give

$$\tilde{R} = -8\Omega^{-5} L_\gamma \Omega, \tag{7.90}$$

where the linear differential operator L_γ is given by

$$L_\gamma = \Delta_\gamma - \tfrac{1}{8} R_\gamma, \tag{7.91}$$

in which $\Delta_\gamma = \gamma^{ij}\nabla_i\nabla_j$ is the Laplacian on Σ defined by γ, and R_γ is the Ricci scalar defined by γ (though three-dimensional, we omit tildes on geometric quantities defined by γ; those with a tilde are defined by $\tilde{g}$). Given γ, eq. (7.88) then becomes an equation for the scalar Ω, namely

$$L_\gamma \Omega = 0. \tag{7.92}$$

This is a linear elliptic PDE, which can indeed be solved if Σ is compact. In GR, this argument applies more generally (e.g. assuming $\Omega \to 0$ at infinity in the non-compact case).

Ellipticity is here to stay, but linearity is typical of the assumption $\tilde{k} = 0$, and will be replaced by gruesome nonlinearities in general. Indeed, already the next case, where $\tilde{k}_{ij} \neq 0$ but $\mathrm{Tr}\,(\tilde{k}) = 0$, is highly nonlinear.[118] The constraints (7.68) - (7.70), again in the vacuum case, simplify to

$$\tilde{R} - \mathrm{Tr}\,(\tilde{k}^2) = 0; \tag{7.93}$$

$$\tilde{g}^{jl}\tilde{\nabla}_l \tilde{k}_{ij} = 0. \tag{7.94}$$

[116] *Birkhoff's Theorem* implies that rotational symmetry implies that the metric is static, see e.g. Hawking & Ellis.

[117] In the context of GR, adding a cosmological constant Λ modifies (7.88) to $\tilde{R} = 2\Lambda$.The possible signs of $\tilde{R}$, i.e. $\tilde{R} = 0, \pm 1$ up to rescaling, are restricted by the topology of Σ and define the so-called *Yamabe class* of Σ.

[118] Foliations with $\mathrm{Tr}\,(\tilde{k}) = 0$ are called ***maximal slicings***. This is related to the *Plateau Problem*: if $\Sigma \subset M$ has $\mathrm{Tr}\,(\tilde{k}) = 0$, and $\mathscr{S} \subset \Sigma$ is a two-dimensional surface, then the volume of any three-dimensional $S \subset \Sigma$ with $\partial S = \mathscr{S}$ is maximal compared to the volume of competing $S \subset M$ subject to $\partial S = \mathscr{S} \subset \Sigma$. In the purely Riemannian Plateau Problem the volume (or, as in the original problem in one dimension lower, the surface area of the enclosed region) would be minimal, but in the Lorentzian case it is rather maximal, for similar reasons why the length of timelike geodesics is maximal rather than minimal: excursions of S outside Σ are in the timelike direction and the signature of the Lorentzian metric then *reduces* the volume (rather than increasing it as in the Plateau Problem).

We now also choose some symmetric tensor k_{ij} on Σ, such that

$$\gamma^{jl}\nabla_l k_{ij} = 0, \tag{7.95}$$

but freely otherwise. It is easy to show that if we relate $\tilde{k}$ to k via

$$\tilde{k}_{ij} = \Omega^{-2} k_{ij}, \tag{7.96}$$

then (7.95) implies (7.94) and hence only (7.93) remains, which turns out to be equivalent to

$$L_\gamma \Omega + \tfrac{1}{8}\mathrm{Tr}\,(k^2)\Omega^{-7} = 0. \tag{7.97}$$

This equation can be analyzed by traditional methods from nonlinear elliptic PDE's (notably by constructing both sub- and super-solutions, i.e. replacing "$= 0$" by "≤ 0" and "≥ 0").

We move to the general case. Here is it customary and physically relevant to move to a ***transverse traceless*** version of k and $\tilde{k}$, where the traceless part is easy to define, namely

$$\tilde{\sigma}_{ij} = \tilde{k}_{ij} - \tfrac{1}{3}\mathrm{Tr}\,(\tilde{k})\tilde{g}_{ij}; \tag{7.98}$$

$$\sigma_{ij} = k_{ij} - \tfrac{1}{3}\mathrm{Tr}\,(k)\gamma_{ij}. \tag{7.99}$$

Adding energy-momentum and using the scaling (7.96), this reformulates the constraints as

$$L_\gamma \Omega + \tfrac{1}{8}\mathrm{Tr}\,(\sigma^2)\Omega^{-7} - \tfrac{1}{12}\mathrm{Tr}\,(k)^2\Omega^5 = -2\pi E\Omega^5; \tag{7.100}$$

$$\nabla_j \sigma_{ij} - \tfrac{2}{3}(\nabla_i \mathrm{Tr}\,(k))\Omega^6 = 8\pi P_i \Omega^{10}, \tag{7.101}$$

the first of which (i.e. the Hamiltonian constraint) is called the ***Lichnerowicz equation***. Defining the transverse part of σ and $\tilde{\sigma}$ is less straightforward: there exists a decomposition

$$\sigma_{ij} = \sigma_{ij}^{\mathrm{TT}} + (\hat{K}_\gamma X)_{ij}, \tag{7.102}$$

where $\sigma_{ij}^{\mathrm{TT}}$ in traceless and transverse in the sense that

$$\mathrm{Tr}\,(\sigma) \equiv \gamma^{ij}\sigma_{ij} = 0; \tag{7.103}$$

$$\nabla^i \sigma_{ij}^{\mathrm{TT}} = 0, \tag{7.104}$$

and X is some vector field, on which the ***conformal Killing operator*** $\hat{K}_\gamma$ acts by

$$(\hat{K}_\gamma X)_{ij} = \nabla_i X_j + \nabla_j X_i - \tfrac{2}{3}\gamma_{ij}\nabla_k X^k. \tag{7.105}$$

This generalizes the usual Killing operator $K_\gamma X = \nabla_i X_j + \nabla_j X_i$, whose solutions $K_\gamma X = 0$ are vector fields whose flow φ_t consists of isometries, i.e., $\varphi_t^* \gamma = \gamma$; vector fields solving $\hat{K}_\gamma X = 0$ are vector fields whose flow φ_t consists of *conformal isometries*, in that $\varphi_t^* \gamma = \Omega\gamma$ for some $\Omega > 0$, as above. The difficult part is the reconstruction of σ_{ij} from its transverse traceless part $\sigma_{ij}^{\mathrm{TT}}$ and X, which may be done by solving a conformal version of the Laplace equation, viz.

$$\hat{\Delta}_\gamma X^i = \nabla_j (\hat{K}_\gamma X)^{ij} = \Delta X^i + \tfrac{1}{3}\nabla^i \nabla_j X^j + R^i_j X^j. \tag{7.106}$$

Note that the kernel of $\hat{\Delta}_\gamma$ consists of conformal Killing vectors. Likewise for $\tilde{g}$ and $\tilde{\sigma}_{ij}$. In terms of the *free data* γ_{ij}, $\sigma_{ij}^{\mathrm{TT}}$, and $\tau \equiv \mathrm{Tr}\,(k)$, the *determined data* Ω and X are found by solving the final (conformal) version of the constraints, namely

$$L_\gamma \Omega + \tfrac{1}{8}\mathrm{Tr}\,(\sigma_{\mathrm{TT}}^2)\Omega^{-7} - \tfrac{1}{12}\tau^2\Omega^5 = -2\pi E\Omega^5; \tag{7.107}$$

$$\hat{\Delta}_\gamma X^i - \tfrac{2}{3}(\nabla_i \tau)\Omega^6 = 8\pi P_i \Omega^{10}. \tag{7.108}$$

Once this has been done, $\tilde{g}_{ij}$ and $\tilde{k}_{ij}$ can be (re)constructed via (7.89) and

$$\tilde{k}_{ij} = (\hat{K}_\gamma X_{ij} + \sigma_{ij}^{\mathrm{TT}})\Omega^{-10} + \tfrac{1}{3}\tau\Omega^{-4}\gamma_{ij}, \tag{7.109}$$

and these solve the original constraints (7.68) - (7.70) in terms of the above free data.

8 Quasi-linear hyperbolic PDE's

8.1 Background

1. ***Multi-indices.*** Let $n > 0$ and $x \in \mathbb{R}^n$. It will convenient to write $x = (x_1, \ldots, x_n)$ rather than our usual $(x^1, \ldots, x^n)$. Let $\alpha = (\alpha_1, \ldots \alpha_n)$, with $\alpha_i \in \mathbb{N}$ (where $0 \in \mathbb{N}$), and put

$$|\alpha| = \sum_{i=1}^{n} \alpha_i; \tag{8.1}$$

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \equiv \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} \equiv \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}; \tag{8.2}$$

$$x^\alpha = x_1^{\alpha^1} \cdots x_n^{\alpha^n}. \tag{8.3}$$

2. ***Distributions.*** For each measurable (usually open) subset $\Omega \subseteq \mathbb{R}^n$, let $\mathscr{D}(\Omega)$ be $C_c^\infty(\Omega)$ as a set, equipped with the topology in which $\varphi_\lambda \to \varphi$ iff there is a compact $K \subset \Omega$ such that $\mathrm{supp}(\varphi_\lambda) \subseteq K$ for all λ, and for all multi-indices α one has (implying $\mathrm{supp}(\varphi) \subseteq K$):

$$\|D^\alpha(\varphi_\lambda - \varphi)\|_\infty \to 0. \tag{8.4}$$

Elements of $\mathscr{D}(\Omega)$ are called ***test functions***. A linear map $u : \mathscr{D}(\Omega) \to \mathbb{C}$ is continuous iff for each compact $K \subset \Omega$ there is $m \in \mathbb{N}$ and $C > 0$ such that for all α with $|\alpha| \leq m$,

$$|\langle u, \varphi \rangle| \equiv |u(\varphi)| \leq C\|D^\alpha \varphi\|_\infty. \tag{8.5}$$

Distributions are elements of the space $\mathscr{D}'(\Omega)$ of all continuous maps $u : \mathscr{D}(\Omega) \to \mathbb{C}$.[119] This space carries the weak topology, in which $u_\lambda \to u$ iff $\langle u_\lambda, \varphi \rangle \to \langle u, \varphi \rangle$ for each $\varphi \in \mathscr{D}(\Omega)$. In this topology, $\mathscr{D}(\Omega)$ is dense in $\mathscr{D}'(\Omega)$, where $u \in \mathscr{D}(\Omega)$ defines $u \in \mathscr{D}'(\Omega)$ through the L^2 inner product, i.e., $\langle u, \varphi \rangle = \langle \overline{u}, \varphi \rangle_{L^2(\Omega)}$. Adding a middle man gives

$$\mathscr{D}(\Omega) \subset L^2(\Omega) \subset \mathscr{D}'(\Omega), \tag{8.6}$$

in which each embedding is continuous and dense. This is an example of a ***Gelfand triple***.

Let Ω be open in $\mathbb{R}^n$. For each α, the ***weak derivative*** $D^\alpha u$ of $u \in \mathscr{D}'(\Omega)$ is defined by

$$\langle D^\alpha u, \varphi \rangle = (-1)^{|\alpha|} \langle u, D^\alpha \varphi \rangle. \tag{8.7}$$

This definition may be motivated by faking the formula $\langle u, \varphi \rangle = \int_\Omega u(x)\varphi(x)$, which on repeated partial integration gives (8.7). Any linear partial differential operator may therefore be regarded as a map $L : \mathscr{D}'(\Omega) \to \mathscr{D}'(\Omega)$, with adjoint $L^* : \mathscr{D}(\Omega) \to \mathscr{D}(\Omega)$, i.e.,

$$\langle Lu, \varphi \rangle = \langle u, L^* \varphi \rangle. \tag{8.8}$$

For example, if $L = D^\alpha$, then $L^* = (-1)^{|\alpha|} D^\alpha$. The derivatives in Lu are called ***weak***, those in $L^*\varphi$ being ***classical***. Similarly, a solution $u \in \mathscr{D}'(\Omega)$ of a linear PDE $Lu = F$ (with initial conditions), i.e. $\langle Lu, \varphi \rangle = \langle u, L^* \varphi \rangle$ for all $\varphi \in \mathscr{D}(\Omega)$, is called ***weak***.

[119] Here and in what follows, if $\Omega = \cup_i K_i$ for compact i, "for each compact K" may be replaced by "for each K_i."

One has to be careful with (8.8) if Ω is not open. For example, if $\Omega = [0,\infty)\times\mathbb{R}^n$ and $L = -\Box = \partial_t^2 - \Delta$, then (due to boundary terms in partial integration) the inhomogeneous wave equation $Lu = F$ with initial conditions $u(0,x) = f$ and $\dot{u}(0,x) = g(x)$ becomes

$$-\int_0^\infty dt \int_{\mathbb{R}^n} d^n x\, u \Box\varphi = \int_0^\infty dt \int_{\mathbb{R}^n} d^n x F\varphi + \int_{\mathbb{R}^n} d^n x\, g(x)\varphi(0,x) - f(x)\dot{\varphi}(0,x). \quad (8.9)$$

For $\Omega = \mathbb{R}^n$, another widely used space of distributions is based on the space of ***rapidly decreasing (test) functions*** $\mathscr{S}(\mathbb{R}^n)$, which consists of those $f \in C^\infty(\mathbb{R}^n)$ for which the function $x \mapsto x^\alpha D^\beta f$ is bounded for all multi-indices α and β. One often writes

$$\langle x\rangle = (1 + \|x\|^2)^{1/2}, \quad (8.10)$$

and uses $x \mapsto \langle x\rangle^\alpha D^\beta f$, which of course gives the same space. The topology on $\mathscr{S}(\mathbb{R}^n)$ is such that $\varphi_\lambda \to \varphi$ iff for all $l,m \in \mathbb{N}$ and multi-indices α and β with $|\alpha| \le l$ and $|\beta| \le m$,

$$\|x^\alpha D^\beta(\varphi_\lambda - \varphi)\|_\infty \to 0. \quad (8.11)$$

The (weak) topology on the space $\mathscr{S}'(\mathbb{R}^n)$ of ***tempered distributions*** has $u_\lambda \to u$ iff there are $l,m \in \mathbb{N}$ and $C > 0$ such that for all α with $|\alpha| \le l$ and β with $|\beta| \le m$ one has

$$|\langle u, \varphi\rangle| \le C\|x^\alpha D^\beta \varphi\|_\infty. \quad (8.12)$$

Similarly to (8.6), one has a Gelfand triple (i.e. the embeddings are continuous and dense)

$$\mathscr{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n), \quad (8.13)$$

and since $\mathscr{D}(\mathbb{R}^n) \subset \mathscr{S}(\mathbb{R}^n)$ continuously, and hence $\mathscr{S}'(\mathbb{R}^n) \subset \mathscr{D}'(\mathbb{R}^n)$, this extends to

$$\mathscr{D}(\mathbb{R}^n) \subset \mathscr{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n) \subset \mathscr{D}'(\mathbb{R}^n). \quad (8.14)$$

3. ***Sobolev spaces.*** For any $s \in \mathbb{N}$, based on (8.6), define the ***Sobolev space***

$$H^s(\Omega) = \{u \in L^2(\Omega) \mid D^\alpha u \in L^2(\Omega)\, \forall \alpha : |\alpha| \le s\}, \quad (8.15)$$

where accordingly the derivatives inherent in D^α are weak. Clearly, $H^0(\Omega) = L^2(\Omega)$, but it can be shown that all $H^s(\Omega)$ are Hilbert spaces with respect to the inner product

$$\langle u, v\rangle_s = \sum_{|\alpha|\le s} \langle D^\alpha u, D^\alpha v\rangle, \quad (8.16)$$

where $\sum_{|\alpha|\le s}$ means $\sum_{\alpha:|\alpha|\le s}$, and $\langle\cdot,\cdot\rangle$ is the inner product in $L^2(\Omega)$ (note the danger of ambiguous notation here: $\langle\cdot,\cdot\rangle_p$ often denotes the inner product in L^p, but here $\langle\cdot,\cdot\rangle_s$ stands for the inner product in H^s; in our notation the inner product in L^2 would be $\langle\cdot,\cdot\rangle_0$).

For $\Omega = \mathbb{R}^n$ a different perspective on Sobolev spaces comes from the ***Fourier transform***

$$\hat{f}(\xi) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} d^n x\, f(x) e^{-i\xi x}; \quad (8.17)$$

$$\check{f}(x) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} d^n \xi\, f(\xi) e^{i\xi x}, \quad (8.18)$$

which make sense as Lebesgue integrals for $f \in L^1(\mathbb{R}^n)$. If one also has $\hat{f} \in L^1(\mathbb{R}^n)$, then

$$\check{\hat{f}} = f. \tag{8.19}$$

The scope of these formulae may be extended in at least three different ways:[120]

(a) Eq. (8.17) yields a unitary isomorphism $L^2(\mathbb{R}^n) \xrightarrow{\cong} L^2(\mathbb{R}^n)$ of Hilbert spaces.

(b) The Fourier transform also defines a linear homeomorphism $\mathscr{S}(\mathbb{R}^n) \xrightarrow{\cong} \mathscr{S}(\mathbb{R}^n)$.

(c) Defining $\hat{f}$ for $f \in \mathscr{S}'(\mathbb{R}^n)$ by $\langle \hat{f}, \varphi \rangle = \langle f, \check{\varphi} \rangle$, the Fourier transform (8.17) even defines a linear homeomorphism $\mathscr{S}'(\mathbb{R}^n) \xrightarrow{\cong} \mathscr{S}'(\mathbb{R}^n)$ of tempered distributions.

Returning to Sobolev spaces, for $\Omega = \mathbb{R}^n$ may now (re)define, *for any* $s \in \mathbb{R}$,

$$H^s(\mathbb{R}^n) = \{u \in \mathscr{S}'(\mathbb{R}^n) \mid \xi \mapsto \langle \xi \rangle^s \hat{u}(\xi) \in L^2(\mathbb{R}^n)\}, \tag{8.20}$$

with inner product

$$\langle u, v \rangle_s = \int_{\mathbb{R}^n} d^n\xi \, \langle \xi \rangle^{2s} \overline{\hat{u}}(\xi) \hat{v}(\xi) = \int_{\mathbb{R}^n} d^n\xi \, (1 + \|\xi\|^2)^s \overline{\hat{u}}(\xi) \hat{v}(\xi) \tag{8.21}$$

For $s \in \mathbb{N}$ this reproduces (8.15) as a vector space (a fact that is not obvious), but of course the inner products (8.16) and (8.21) are different (yet they induce equivalent norms), and so for $s \in \mathbb{N}$ one has to specify which one is used. This makes a difference neither for the ***Sobolev embedding theorem***, which states that for $m \geq 0$ and $s > m + \frac{1}{2}n$, one has

$$H^s(\mathbb{R}^n) \subset C_b^m(\mathbb{R}^n), \tag{8.22}$$

where the embedding is continuous with respect to the norm $\|u\|_{m,\infty} = \sum_{|\alpha| \leq m} \|D^\alpha u\|_\infty$, nor for the fundamental ***Sobolev duality theorem***, which states that for any $s \in \mathbb{R}$ one has

$$H^s(\mathbb{R}^n)^* \cong H^{-s}(\mathbb{R}^n), \tag{8.23}$$

i.e. $\Lambda \in H^s(\mathbb{R}^n)^*$ linearly, bijectively, and isometrically corresponds to $f \in H^{-s}(\mathbb{R}^n)$ via

$$\Lambda(u) = \int_{\mathbb{R}^n} d^n x \, f(x) u(x) \equiv \langle f, u \rangle. \tag{8.24}$$

Finally, for $s > 0$ we have our third Gelfand triple

$$H^s(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset H^{-s}(\mathbb{R}^n), \tag{8.25}$$

which analogously to (8.14) may be extended to a 'Gelfand quintuple'

$$\mathscr{S}(\mathbb{R}^n) \subset H^s(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset H^{-s}(\mathbb{R}^n) \subset \mathscr{S}'(\mathbb{R}^n). \tag{8.26}$$

[120] Less well known, if one equips $C_c^\infty(\mathbb{R}^n)$ with the unusual norm $\|f\|_0 = \max\{\|f\|_\infty, \|\hat{f}\|_\infty\}$, with associated completion denoted by $C_0^*(\mathbb{R}^n)$, then (8.17) yields an isometric isomorphism $C_0^*(\mathbb{R}^n) \xrightarrow{\cong} C_0^*(\mathbb{R}^n)$ as Banach spaces. For C*-algebra experts we note that the Fourier transform also yields an isomorphism $C^*(\mathbb{R}^n) \xrightarrow{\cong} C_0(\mathbb{R}^n)$ of commutative C*-algebras (here $C^*(\mathbb{R}^n)$ is the completion of $C_c^\infty(\mathbb{R}^n)$ in the operator norm obtained by letting $f \in C_c^\infty(\mathbb{R}^n)$ act on $L^2(\mathbb{R}^n)$ by convolution, whereas $C_0(\mathbb{R}^n)$ carries the supremum-norm). In this case (which follows from the Riemann–Lebesgue lemma) the Fourier transform is a special case of the Gelfand transform.

For our kind of PDE's, $\mathbb{R}^n$ will be space, and time needs to be treated separately. Typically, for fixed $T > 0$ one considers Banach spaces like $C([0,T],H^s(\mathbb{R}^n))$, with norm

$$\|u\|_\infty = \sup_{t\in[0,T]} \|u(t)\|_s, \tag{8.27}$$

or $C^1([0,T],H^s(\mathbb{R}^n))$ with analogous norm, or $L^p([0,T],H^s(\mathbb{R}^n))$, $1 \leq p < \infty$, normed by

$$\|u\|_p = \left(\int_0^T dt\, (\|u(t)\|_s)^p\right)^{1/p}, \tag{8.28}$$

or $L^\infty([0,T],H^s(\mathbb{R}^n))$, with norm

$$\|u\|_\infty = \operatorname{ess\,sup}_{t\in[0,T]} \|u(t)\|_s. \tag{8.29}$$

Here we define $L^p([0,T],H^s(\mathbb{R}^n))$, $1 \leq p < \infty$, as the completion of $C([0,T],H^s(\mathbb{R}^n))$ in the norm (8.28), and also (avoiding Banach space-valued measurable functions), *define* $L^\infty([0,T],H^s(\mathbb{R}^n))$ as the (Banach) dual of $L^1([0,T],H^{-s}(\mathbb{R}^n))$, in that we identify $f \in L^\infty([0,T],H^s(\mathbb{R}^n))$ with the functional $\Lambda_f \in (L^1([0,T],H^{-s}(\mathbb{R}^n)))^*$ given by, cf. (8.24),

$$\Lambda_f(g) = \int_0^T dt\, \langle f(t), g(t)\rangle. \tag{8.30}$$

8.2 Linear wave equations

To see such spaces in action, as before we consider the free wave equation on $\mathbb{R}^{n+1}$, i.e.

$$(-\partial_t^2 + \Delta)u = F; \tag{8.31}$$

$$u(0,x) = f; \;\; \dot{u}(0,x) = g(x), \tag{8.32}$$

For $F = 0$ and $n = 1, 3$, the (unique) solution (known since the 18th century) is

$$u(t,x) = \tfrac{1}{2}\left(f(x+t) - f(x-t) + \int_{x-t}^{x+t} dy\, g(y)\right); \qquad (n=1); \tag{8.33}$$

$$u(t,x) = \frac{1}{4\pi t^2}\int_{|y-x|=t} d\sigma^2(y)\left(t g(y) + f(y) - \sum_{i=1}^3 \partial_i f(y)(x_i - y_i)\right); \qquad (n=3). \tag{8.34}$$

From this, we see that in $n = 1$ the solution at (t,x) only depends on initial data within its causal past $J^-(x,t)$, intersected with the Cauchy surface $\Sigma = \{(x^0 = 0, x), x \in \mathbb{R}^n\}$. Indeed, recall the causal past $J^-(t,x)$,emanating from (t,x), and its boundary $L^-(t,x)$, i.e. the past light cone,

$$J^-(t,x) = \{(y^0,y) \in \mathbb{R}^{n+1}, |y^0 - x^0| \geq |y-x|, y^0 \leq x^0\}; \tag{8.35}$$

$$L^-(t,x) = \{(y^0,y) \in \mathbb{R}^{n+1}, |y^0 - x^0| = |y-x|, y^0 \leq x^0\}, \tag{8.36}$$

cf. (4.7) - (4.8) with $y^0 \geq x^0$ replaced by $y^0 \leq x^0$ (as well as x by (t,x), etc.). In $n = 1$, we have

$$\Sigma \cap J^-(x,t) = \{(y^0 = 0, y), y \in [x-t, x+t]\}, \tag{8.37}$$

whereas in $n = 3$ the solution $u(t,x)$ even depends on the initial data at $\Sigma \cap L^-(x,t)$ only, since

$$\Sigma \cap L^-(t,x) = \{(y^0 = 0, y), |y-x| = t\}. \tag{8.38}$$

An analogous phenomenon holds in the inhomogeneous case $F \neq 0$, in which case the solution

$$u(t,x) = \frac{1}{4\pi} \int_{L^-(t,x)} d\sigma^3(s,y) \, \frac{F(s,y)}{|(s-t,y-x)|}, \tag{8.39}$$

for zero initial data for simplicity, clearly depends on the values of F at the past light cone $L^-(t,x)$; in other words, $F(s,y)$ only influences u along the forward light cone emanating from (s,y). The situation in $n = 3$ (and also in all higher *odd* spatial dimensions), in which both initial data f, g and the inhomogeneous term F affect the solution only along future light rays is called the ***strong Huygens principle***. The (ordinary) ***Huygens principle***, then, formalizes the situation in $n = 1, 2$, and all higher *even* dimensions, in which the entire causal future of (s,y) affects the solution, or, equivalently, $u(t,x)$ only depends on data within its causal past.

An explicit solution for any F, f, and g may be written down using the Fourier transform:

$$\hat{u}(t,\xi) = \cos(t|\xi|)\hat{f}(\xi) + \frac{\sin(t|\xi|)}{|\xi|}\hat{g}(\xi) + \int_0^t ds \, \frac{\sin((t-s)|\xi|)}{|\xi|}\hat{F}(s,\xi); \tag{8.40}$$

as the notation indicates, the formula (8.17) is only applied to the x-variable, and, within the function classes to be discussed, the actual solution $u(t,x)$ may be (re)constructed from (8.18). Although the space-time and causal structure of the solution is not at all obvious from this formula, the advantage is that (8.40) easily implies an ***energy inequality***: for any $s \in \mathbb{Z}$,

$$\|u(t,\cdot)\|_{s+1} + \|\dot{u}(t,\cdot)\|_s \leq C_{s,T} \left((\|f\|_{s+1} + \|g\|_s) + \int_0^T d\tau \, \|F(\tau,\cdot)\|_s \right), \tag{8.41}$$

where $0 < T < \infty$, provided that $F \in L^1([0,T], H^s(\mathbb{R}^n))$, $f \in H^{s+1}(\mathbb{R}^n)$, and $g \in H^s(\mathbb{R}^n)$, so that the right-hand side makes sense. The proof is an exercise, using the fact that (8.21) implies

$$\|u(t)\|_s^2 = \int_{\mathbb{R}^n} d^n\xi \, (1 + \|\xi\|^2)^s \, |\hat{u}(t,\xi)|^2. \tag{8.42}$$

Corollary 26 *For any $T > 0$ and $s \in \mathbb{Z}$, the free wave equation* (8.31) - (8.32) *with initial conditions $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, and $F \in L^1([0,T], H^s(\mathbb{R}^n))$, has a unique solution*

$$u(t,x) \in C([0,T], H^{s+1}(\mathbb{R}^n)) \cap C^1([0,T], H^s(\mathbb{R}^n)), \tag{8.43}$$

Uniqueness follows either from the derivation of the explicit solution (8.40) from the initial data, or from (8.41): if u_1 and u_2 both solve (8.31) - (8.32), then $u = u_1 - u_2$ solves (8.31) for $F = f = g = 0$, so that the right-hand side and hence the left-hand side of (8.41) vanishes, etc.

We now turn to linear wave equations of the form $Lu = F$ with initial data (8.32), and

$$L = g^{\rho\sigma}(t,x)\partial_\rho\partial_\sigma + b^\rho(t,x)p_\rho + a(t,x). \tag{8.44}$$

Since we don't have an explicit solution, the derivation of a suitable energy inequality (to be used as a lemma for proving existence, uniqueness, and analytic properties of solutions) will have to be *a priori*.[121] A particularly useful energy inequality for the operator (8.44) is

$$\sum_{|\alpha|\leq 1} \|D^\alpha u(t,\cdot)\|_s \leq C'_{s,T} \left(\sum_{|\alpha|\leq 1} \|D^\alpha u(0,\cdot)\|_s + \int_0^t d\tau \, \|Lu(\tau,\cdot)\|_s \right), \tag{8.45}$$

[121] These *a priori* derivations are straightforward but very lengthy, and therefore we simply state the results without derivation; for (8.45) see Sogge, §I.3 and Luk, §4. See also Ringström for similar estimates.

valid for any $0 < t < T < \infty$, $s \in \mathbb{Z}$, and u such that (8.43) holds, as well as $Lu \in L^1([0,T], H^s)$.[122] This inequality immediately gives uniqueness by the same argument as for the free wave equation, but existence and regularity require a more advanced, functional-analytic argument. In order to explain the reasoning, let us first take a simpler situation.[123] For $\Omega \subseteq \mathbb{R}^n$, let

$$L : \mathscr{D}'(\Omega) \to \mathscr{D}'(\Omega) \tag{8.46}$$

be a linear operator, e.g. as in (8.44), with adjoint $L^* : \mathscr{D}(\Omega) \to \mathscr{D}(\Omega)$ defined by (8.8). As already mentioned, the PDE $Lu = F$ (with zero initial conditions for simplicity) then means

$$\langle u, L^*\varphi \rangle = \langle F, \varphi \rangle \tag{8.47}$$

for all $\varphi \in \mathscr{D}(\Omega)$. Throughout the argument, we must assume that, for any net (φ_λ) in $\mathscr{D}(\Omega)$,

$$L^*\varphi_\lambda \to L^*\varphi \;\Rightarrow\; \varphi_\lambda \to \varphi. \tag{8.48}$$

If L^* is a bijection, and $F \in \mathscr{D}'(\Omega)$, which is the very least regularity to impose, then we are done at the coarsest level of proving existence and uniqueness of a solution $u \in \mathscr{D}'(\Omega)$, since its value at $\psi \in \mathscr{D}'(\Omega)$ is given by finding the unique $\varphi \in \mathscr{D}(\Omega)$ for which $\psi = L^*\varphi$ and putting

$$\langle u, \psi \rangle = \langle F, \varphi \rangle \;\; (\psi = L^*\varphi). \tag{8.49}$$

The assumption (8.48) then implies that if $\psi_\lambda \to \psi$, i.e., $L^*\varphi_\lambda \to L^*\varphi$, then $\varphi_\lambda \to \varphi$, and hence $\langle F, \varphi_\lambda \rangle \to F, \varphi \rangle$ since $F \in \mathscr{D}'(\Omega)$ by assumption, and hence $\langle u, \psi_\lambda \rangle \to \langle u, \psi \rangle$, since $\langle u, \psi_\lambda \rangle = \langle F, \varphi_\lambda \rangle$. Thus u is a continuous linear functional on $\mathscr{D}(\Omega)$ and hence $u \in \mathscr{D}'(\Omega)$.

If L^*, still assumed to be injective, merely has dense range $\mathrm{ran}(L^*) \subset \mathscr{D}(\Omega)$, then one still has existence and uniqueness of u, since for $\psi \in \mathrm{ran}(L^*)$ eq. (8.49) still works, whereas for ψ outside the range of L^* we may write $\psi = \lim_\lambda L^*\varphi_\lambda$ and then $\langle u, \psi \rangle = \lim_\lambda \langle F, \varphi_\lambda \rangle$.

Finally, if L^*, still injective, does not have dense range, the Hahn–Banach Theorem (for locally convex vector spaces) yields existence of u by extending the solution $u : \mathrm{ran}(L^*) \to \mathbb{C}$ constructed above to a continuous linear map $u : \mathscr{D}'(\Omega) \to \mathbb{C}$, but one loses uniqueness. Fortunately, in many applications to PDE's uniqueness still follows from energy inequalities.

Such inequalities also play a central role in refining the above argument. Suppose one has two Gelfand(ish) triples $\mathscr{D}(\Omega) \subset W \subset \mathscr{D}'(\Omega)$ and $\mathscr{D}(\Omega) \subset Z \subset \mathscr{D}'(\Omega)$, where W and Z are Banach spaces and all inclusion maps are continuous with dense image, and suppose that

$$\|\varphi\|_Z \leq C\|L^*\varphi\|_W \;\; (\forall\, \varphi \in \mathscr{D}(\Omega)). \tag{8.50}$$

This 'energy condition' supersedes the continuity assumption (8.48) within $\mathscr{D}(\Omega)$, and is also more powerful in that it clearly implies that L is injective, which is an essential condition for the whole analysis to apply in the first place. Furthermore, the inequality (8.50) implies:

Provided L^* is injective, for any $F \in Z^*$ there is a solution $u \in W^*$ to $Lu = F$.

Noting that $\mathscr{D}(\Omega) \subset Z$ implies $Z^* \subset \mathscr{D}'(\Omega)$, and similarly $\mathscr{D}(\Omega) \subset W$ implies $W^* \subset \mathscr{D}'(\Omega)$, compared with the earlier argument where the assumption $F \in \mathscr{D}'(\Omega)$ gave a solution $u \in$

[122] Moreover, the derivation requires that $g^{\mu\nu}(t,x)$, $b^\mu(t,x)$, and $a(t,x)$ be C^∞ with uniform bounds on all derivatives, where $(t,x) \in [0,T] \times \mathbb{R}^n$, as well as $\sum_{\mu,nu} |g^{\mu\nu}(t,x) - \eta^{\mu\nu}| \leq \frac{1}{2}$, where η is the Minkowski metric.

[123] See e.g. A. Vasy, *Partial Differential Equations* (AMS, 2015), Chapter 17.

$\mathscr{D}'(\Omega)$, we have now strengthened the assumption to $F \in Z^* \subset \mathscr{D}'(\Omega)$, and, given (8.50), accordingly strengthened the conclusion $u \in \mathscr{D}'(\Omega)$ to $u \in W^* \subset \mathscr{D}'(\Omega)$. Indeed, noting that

$$\mathrm{ran}(L^*) \subset \mathscr{D}(\Omega) \subset W, \tag{8.51}$$

let $\psi \in \mathrm{ran}(L^*)$, so $\psi = L\varphi$, and define a linear map $u : W \to \mathbb{C}$ initially on $\mathrm{ran}(L^*) \subset W$ by

$$\langle u, L^*\varphi \rangle_{W^*-W} = \langle F, \varphi \rangle_{Z^*-Z}. \tag{8.52}$$

Because of (8.50), if $L^*\varphi_\lambda \to L^*\varphi$ in W, then $\varphi_\lambda \to \varphi$ in Z, and hence on the assumption $F \in Z^*$, the functional u defined by (8.52) is continuous on $\mathrm{ran}(L^*)$ in the (norm) topology of W. Once again, the Hahn–Banach Extension Theorem (but this time simply for Banach spaces) gives a continuous extension $u : W \to \mathbb{C}$, i.e. $u \in W^*$, as claimed.

We now show how the energy estimate (8.45) implies an estimate à la (8.50). For any $T > 0$, we replace u in (8.45) by $\varphi \in C_c^\infty((0,T) \times \mathbb{R}^n)$, which certainly satisfies the assumptions validating (8.45), and replace L by L^*. Then $D^\alpha u(0,\cdot)$ is replaced by $D^\alpha \varphi(0,\cdot) = 0$. Furthermore, for any multi-index α, $s \in \mathbb{R}$, $k \in \mathbb{N}$, and $\varphi \in H^s$, by definition of the Sobolev spaces we have

$$\|\varphi\|_{-s} \leq C' \sum_{|\alpha| \leq k} \|D^\alpha \varphi\|_{-s-k}. \tag{8.53}$$

With $k = 1$, also using the trivial $\int_0^t d\tau\, g(\tau) \leq \int_0^T d\tau\, g(\tau)$ whenever $0 < t < T$ and $g(\tau) \geq 0$, in this case with $g(\tau) = \|L^*\varphi(\tau,\cdot)\|_{-s-1}$, we find, for any $s \in \mathbb{Z}$ and $\varphi \in C_c^\infty((0,\infty) \times \mathbb{R}^n)$,

$$\|\varphi(t,\cdot)\|_{-s} \leq C \int_0^T d\tau\, \|L^*\varphi(\tau,\cdot)\|_{-s-1}. \tag{8.54}$$

This is a special case of (8.50), with

$$W = L^1([0,T], H^{-s-1}(\mathbb{R}^n)); \tag{8.55}$$

$$Z = C([0,T], H^{-s}(\mathbb{R}^n)); \tag{8.56}$$

$$W^* = L^\infty([0,T], H^{s+1}(\mathbb{R}^n)); \tag{8.57}$$

$$Z^* \supset L^1([0,T], H^s(\mathbb{R}^n)); \tag{8.58}$$

the precise form of Z^* (which is the space of bounded measures on $[0,T]$ taking values in H^s is not needed here). Assuming zero initial conditions for the moment, the abstract argument above therefore gives a solution $u \in L^\infty([0,T], H^{s+1}(\mathbb{R}^n))$ for $F \in L^1([0,T], H^s(\mathbb{R}^n))$, which, by the original energy inequality (8.45) is also unique. More advanced arguments involving elliptic regularity further push to solution into (8.43).[124] Finally, the case of nonzero initial data f, g can be reduced to the case $f = g = 0$ by a standard trick: for given F, let v solve $Lv = F$ for zero initial data, define $w(t,x) = f(x) + tg(x)$, and $u = v + w$ solves $Lu = F$ for given f, g. Thus:

Theorem 27 *For any $T > 0$, let L be defined by* (8.44)*, including all assumption stated afterwards. For any $s \in \mathbb{Z}$, the linear wave equation $Lu = F$, with $F \in L^1([0,T], H^s(\mathbb{R}^n))$ and initial conditions $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, see* (8.32)*, has a unique solution*

$$u(t,x) \in C([0,T], H^{s+1}(\mathbb{R}^n)) \cap C^1([0,T], H^s(\mathbb{R}^n)). \tag{8.59}$$

Corollary 28 *In the setting of the previous theorem, if F, f, and g are smooth, then so is u.*

This follows from the Sobolev embedding theorem (8.22). With further effort, one can also show that the causal properties of the solution relative to F and the initial data f, g are the same as for the free wave equation, except that the *strong* Huygens principle need not apply (but the 'oridinary' one, implying causal propagation of inital data and F, always does).

[124] See Sogge, p. 20.

8.3 Quasi-linear wave equations

In either the (naive) wave gauge or its refinement the $\hat{g}$-wave gauge, the Einstein equations (5.79) - (5.80) take the abstract form form $Lu = F$, where $u = g_{\mu\nu}$ and L is like (8.44), with the difference that in $L = g^{\rho\sigma}(u)\partial_\rho\partial_\sigma$ the coefficient of the highest- (i.e. second) order derivative now depends on u, and furthermore $F = F(u,\partial u)$ depends on u and ∂u. Such equations (in the more general case that g and F may depend on u, ∂u, and even (t,x)) are called ***quasi-linear***,[125] and if the signature of g is Lorentzian, as we of course assume, the PDE is ***hyperbolic***.

We assume for the moment that u takes values in $\mathbb{R}$; the generalization to $u = (g_{\mu\nu})$ taking values in $\mathbb{R}^{10}$, is straightforward and will be outlined later. It is also sufficient for later applications to GR to assume that $g^{\rho\sigma} : \mathbb{R} \to \mathbb{R}$ is smooth, as is $F : \mathbb{R}\times\mathbb{R}^{n+1} \to \mathbb{R}$. So we study

$$g^{\rho\sigma}(u)\partial_\rho\partial_\sigma u = F(u,\partial u). \tag{8.60}$$

As opposed to truly nonlinear hyperbolic PDE's, the quasi-linear case is relatively easy because it can be solved by reduction to the linear case, and one can only feel fortunate that the Einstein equations (at least in a suitable gauge) fall into this category. The solution method is a generalization of the Picard iteration procedure from ODE's:[126] start from the initial data

$$u_0(x) = f(x) = u(0,x), \tag{8.61}$$

and iteratively define u_{k+1} as the solution to the inhomogeneous linear PDE

$$g^{\rho\sigma}(u_k)\partial_\rho\partial_\sigma u_{k+1} = F(u_k,\partial u_k), \tag{8.62}$$

subject to the initial conditions $u_{k+1}(0,x) = f(x)$ and $\dot{u}_{k+1}(0,x) = g(x)$, as for u itself.[127] For some given function $u_k(t,x)$, eq. (8.62) is the type of PDE studied in the previous section. Hence Theorem 27 guarantees a solution for any $T > 0$, but convergence of the iteration and uniformity of the energy inequality (8.41) in k gives a weaker conclusion compared to the linear case:[128]

Theorem 29 *For smooth $F(u,\partial u)$ and smooth $g^{\rho\sigma}(u)$ sufficiently close to the Minkowski metric,*[129] *eq.* (8.60)*, with initial conditions $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, has a unique solution*

$$u \in L^\infty([0,T],H^{s+1}(\mathbb{R}^n)); \tag{8.63}$$

$$\dot{u} \in L^\infty([0,T],H^s(\mathbb{R}^n)), \tag{8.64}$$

provided $s > \frac{1}{2}n$ (i.e. $s > 3/2$ for $n = 3$). Here T is either arbitrary (as in the linear case), or there exists $T_* = T_*(\|f\|_{s+1},\|g\|_s)$ such that $\|D^\alpha u\|_\infty = \infty$ on $[0,T_*)\times\mathbb{R}^n$, for some $|\alpha| \le 2$.

This solution depends continuously on the initial data (so that the Cauchy problem for (8.60) *is well posed in the sense of Hadamard) in the obvious way, i.e., if $f_k \to f$ in $H^{s+1}(\mathbb{R}^n)$ and $g_k \to g$ in $H^s(\mathbb{R}^n)$, then $u_k \to u$ in $L^\infty([0,T],H^{s+1}(\mathbb{R}^n))$ with $\dot{u}_k \to \dot{u}$ in $L^\infty([0,T],H^s(\mathbb{R}^n))$.*

Finally, if $f \in C_c^\infty(\mathbb{R}^n)$ and $g \in C_c^\infty(\mathbb{R}^n)$, then $u \in C^\infty([0,T]\times\mathbb{R}^n)$, cf. Corollary 28.

Of course, $T < \min\{T_*\}$. For a trivial example with $T_* < \infty$, take $(\partial_t^2 - \Delta)u = u^3$ with $u(0,x) = \dot{u}(0,x) = 1$ (times a cutoff function), so that $u(t,x) = 1/(1-t)$ (for small x), and hence $T^* = 1$.

[125] In fluid mechanics all these dependencies occur, see e.g. Taylor, Chapter 16, but the abstract theory is similar.

[126] Recall that an ODE $u'(t) = f(t,u(t))$ with initial condition $u(0) = u_0$, which is equivalent to the integral equation $u(t) = u_0 + \int_0^t ds\, f(s,u(s))$, may be solved by iteration from $u_0(t) = u_0$ and $u_{k+1}(t) = u_0 + \int_0^t ds\, f(s,u_k(s))$. For suitably regular f, this sequence (u_k) uniformly converges to a solution u on some interval $[0,T]$.

[127] This works if $f,g \in C_c^\infty(\mathbb{R}^n)$. For initial data $f \in H^s(\mathbb{R}^n)$ and $g \in H^{s+1}(\mathbb{R}^n)$ one needs to approximate f and g within the spaces mentioned by sequences (f_k) and (g_k) in $fC_c^\infty(\mathbb{R}^n)$, respectively, upon which the initial conditions for (8.62) change into $u_{k+1}(0,x) = f_{k+1}(x)$ and $\dot{u}_{k+1}(0,x) = g_{k+1}(x)$.

[128] See Sogge, §I.4, Luk, §6, Choquet-Bruhat, Appendix III, or Ringström, Chapter 9.

[129] Think of $\sum_{\rho,\sigma}\|g^{\rho\sigma} - \eta^{\rho\sigma}\|_\infty \le \frac{1}{2}$, as in Sogge. Even for initial data $f \in H^{s+1}(\mathbb{R}^n)$ and $g \in H^s(\mathbb{R}^n)$, one can make further (rather contrived) regularity assumptions on $g^{\rho\sigma}$ and F that push u into (8.59). See Ringström, Ch. 9.

8.4 Application to GR

Theorem 29 applies to the Einstein equations in the ($\hat{g}$-) wave gauge, with the following changes:

- Instead of a single unknown u we now have 10 unknowns $g_{\mu\nu}$, with one equation for each (but the ensuing system is coupled, since $g^{\rho\sigma}$ is a function of all $g_{\mu\nu}$ and so is $F(g,\partial g)$).
- The Cauchy surface $\{t=0\}\subset\mathbb{R}^{n+1}$ is replaced by a $3d$ (Riemannian) manifold Σ.
- The initial data $u(0,\cdot)=f$ and $\dot{u}(0,\cdot)=g$ are replaced by the Cauchy data $(\tilde{g}_{ij},\tilde{k}_{ij})$ on Σ.
- Using either local coordinate patches and a partition of unity, or a background metric $\hat{e}$ on Σ making the construction coordinate-independent (like the $\hat{g}$-wave gauge), one can define Sobolev spaces $H^s(\Sigma)$ for any $s\in\mathbb{R}$ (in view of $s<\frac{1}{2}n+1$ in Theorem 29, $s\in\mathbb{N}$ is enough).[130] This construction may be extended from functions on Σ to arbitrary tensors $\tau\in\mathfrak{X}^{(k,l)}(\Sigma)$, yielding Sobolev spaces $H^s_{(k,l)}(\Sigma)$. Thus one may say, e.g., $\tilde{k}\in H^s_{(2,0)}(\Sigma)$.
- The PDE (8.62) is replaced by the reduced (vacuum) Einstein equations (5.79) or (5.80).

Theorem 30 *Let $s>3/2$.*[131] *For initial data $(\Sigma,\tilde{g}_{ij},\tilde{k}_{ij})$ where $\tilde{g}$ is sufficiently close to $\hat{e}$ and*

$$\tilde{g}\in H^{s+1}_{(2,0)}(\Sigma); \tag{8.65}$$

$$\tilde{k}\in H^{s}_{(2,0)}(\Sigma), \tag{8.66}$$

there is $T>0$ such that the reduced vacuum Einstein equations (5.79) - (5.80) *or their counterparts in a $\hat{g}$-wave gauge, have a unique solution g on $M=[0,T]\times\Sigma$, where*

$$g_{\mu\nu}\in C([0,T],H^{s+1}(\Sigma))\cap C^1([0,T],H^s(\Sigma)); \tag{8.67}$$

$$\partial_\rho g_{\mu\nu}\in C([0,T],H^s(\Sigma)). \tag{8.68}$$

Note that (8.22) *and* (8.65) - (8.66) *imply that for $s>3/2$ one has $\tilde{g}\in C^1(\Sigma)$ and $\tilde{k}\in C(\Sigma)$, whilst* (8.67) - (8.68) *then imply $g\in C^1(M)$ and hence $\partial g\in C(M)$. This solution continuously depends on the initial data on the initial data in that $\tilde{g}_l\to\tilde{g}$ in $H^{s+1}_{(2,0)}(\Sigma)$ and $\tilde{k}_l\to\tilde{k}$ in $H^s_{(2,0)}(\Sigma)$ imply $g_l\to g$ in $L^\infty([0,T],H^{s+1}(\Sigma))$ as well as $\partial_\rho g_l\to g$ in $L^\infty([0,T],H^s(\Sigma))$. Finally, if the initial data $(\tilde{g},\tilde{k})$ on Σ are smooth, then so is g, in which case (M,g) is globally hyperbolic.*[132]

This theorem concerns the Einstein equations in specific gauges in which they are hyperbolic. We have already seen that the Einstein equations as such ar not hyperbolic at all and fail to have unique solutions. The general situations was already described in Proposition 31, which as stated was valid for smooth initial data, and as such has now been justified by Theorem 30. It may now be sharpened, since it also holds for 'rough' initial data of the kind described in the above theorem; the proof (by reduction to the wave gauge) goes through virtually unchanged.

Proposition 31 may also be localized, in which case it is best seen as a *causality* result:

[130] See Taylor, *Partial Differential Equations*, Vol. I, Ch. 4, Ringström, Ch. 15, or Choquet-Bruhat, Appendix I.

[131] Choquet-Bruhat's original existence proof had $s>3/2$ but (geometric) uniqueness required $s>5/2$, see Choquet-Bruhat, Theorem 8.4, p. 168 (note that her s is our $s=1$ so that our $s>\frac{1}{2}n$ is her $s>\frac{1}{2}n+1$, etc.). For the sharper $s>3/2$ for existence and uniqueness, also in Proposition 31 and Theorem 22, see P.T. Chruściel, On maximally globally hyperbolic vacuum space-times, *J. Fixed Point Theory Appl.* 14, 325–353 (2014), Thm. 1.1.

[132] See Theorem 11.10 in Chapter XII of Choquet-Bruhat, but the claim is natural given $M=[0,T]\times\Sigma$.

Proposition 31 *Let $(\tilde{g}_{ij},\tilde{k}_{ij})$ and $(\tilde{g}'_{ij},\tilde{k}'_{ij})$ be (smooth) initial data on Σ that coincide on some submanifold $\Sigma_0 \subset \Sigma$. Then any two Cauchy developments $([0,T]\times\Sigma,g)$ and $([0,T']\times\Sigma,g')$ of these data are isometric when restricted to $D^+(\Sigma_0) \subset [0,T'']\times\Sigma_0$, where $T'' = \min\{T,T'\}$.*

This does not follow from (the proof of) Proposition 31 alone (i.e. by reduction to a wave gauge); in addition, one needs a uniqueness (or causality) result for nonlinear wave equations. For (8.60) and analogous equations, this states that if $\Sigma_0 \subset \Sigma$ is a submanifold, then $u = u'$ and $\dot{u} = \dot{u}'$ on Σ_0 implies $u = u'$ on $D^+(\Sigma_0)$. Equivalently, if the initial data f and g vanish on Σ_0, then the solution u vanishes on $D^+(\Sigma)$. This is a localized uniqueness result (for $\Sigma_0 = \Sigma$ it is simply the uniqueness claim for solutions of (8.60)), but it is at the same a causality result, stating that information (i.e. initial data) propagates causally, i.e. within the forward light-cone.

Much as uniqueness is proved from an energy inequality, the localized uniqueness of the above kind is proved from a localized energy inequality. We merely explain this for the free wave equation $\Box u = 0$ in $\mathbb{R}^{n+1}$, but the principle is the same also in Lorentzian geometry.[133]

For any $0 \leq t \leq R$, $(t,x) \in \mathbb{R}^{n+1}$, and (reasonable) function $u(t,x)$, define

$$E(t,x,R) = \tfrac{1}{2}\int_{|y-x|\leq R-t} d^n y\,[\dot{u}(t,y)^2 + \nabla u(t,y)\cdot\nabla u(t,y)]. \tag{8.69}$$

This is just the energy of u, restricted to the ball $B(x;R-t) \subset \mathbb{R}^n$. If $\Box u = 0$, then,

$$0 \leq s \leq t \;\;\Rightarrow\;\; 0 \leq E(t,x,R) \leq E(s,x,R). \tag{8.70}$$

That is, $t \mapsto E(t,x,R)$ is monotonically non-increasing. Fix $R > 0$, and note that

$$E(0,x,R) = \tfrac{1}{2}\int_{B(x,R)} d^n y(g(y)^2 + \nabla f(y)\cdot\nabla f(y)). \tag{8.71}$$

Eq. (8.71) implies that if $f(y) = g(y) = 0$ for all y such that $|y-x| \leq R$, then $E(0,x,R) = 0$, and hence $E(t,x,R) = 0$ for all $0 \leq t \leq R$ by (8.70), and hence $u(t,x) = 0$ by (8.71). Taking $R = t$ shows that if $f(y) = g(y) = 0$ for all y such that $|y-x| \leq t$, then $u(t,x) = 0$. In other words, if $f = g = 0$ within $\Sigma_0 \subset \Sigma$ (defined as the $t = 0$ hyperplane $\mathbb{R}^n_0$ in $\mathbb{R}^{n+1}$), then $u = 0$ within $D^+(\Sigma)$. Equivalently, if $u_1 = u_2$ and $\dot{u}_1 = \dot{u}_2$ at Σ_0, then $u_1 = u_2$ in $D^+(\Sigma_0)$. In case of the Einstein equations, $u_1 = u_2$ becomes $g_1 \cong g_2$ (isometrically), as we have seen, but otherwise the reasoning is the same, ultimately based on the property $g_1 = g_2$ if both metrics are brought into the same gauge. In sum, the solutions of the Einstein equations satisfy all desirable properties:

1. *Existence* global in space and local in time (with satisfactory regularity dictated by regularity of the initial data $(\Sigma,\tilde{g}_{ij},\tilde{k}_{ij})$, including smoothness for smooth initial data;

2. *Uniqueness* up to diffeomorphism;

3. *Causal propagation*, in that initial data at $\Sigma_0 \subset \Sigma$ determine the solution within $D^+(\Sigma_0)$, or, equivalently, $g_{\mu\nu}(t,x)$ is determined by initial data within the causal past $J^-(t,x)$ of (t,x), both statements again up to diffeomorphism;

4. *Cauchy stability*, in that the $4d$ metric g continuously depends on the initial data $(\Sigma,\tilde{g}_{ij},\tilde{k}_{ij})$, as formalized in Theorem 30.

Nothing is implied about global existence in time. Thus the next step is a proof of at least timelike geodesic *completeness* (cf. Definition 8), which is a very active area of research.[134]

[133] See Choquet-Bruhat, Appendix III, Theorem 2.15.

[134] Largely initiated by Christodoulou & Klainerman, *The Global Nonlinear Stability of the Minkowski Space* (Princeton University Press, 1994); see also Christodoulou (2008).

Literature

Books

J.K. Beem, P.E. Ehrlich, K. Easley, *Global Lorentzian Geometry, 2nd Edition* (M. Dekker, 1996).

Y. Choquet-Bruhat, *General Relativity and the Einstein Equations* (OUP, 2009).

Y. Choquet-Bruhat, C. DeWitt-Morette, *Analysis, Manifolds and Physics, Revised Edition* (Elsevier, 1982).

D. Christodoulou, *Mathematical Problems of General Relativity I* (EMS, 2008).

T. Frankel, *The Geometry of Physics, 2nd Edition* (CUP, 2004).

S.W. Hawking & G.F.R. Ellis, *The Large Scale Structure of Space-Time* (CUP, 1973).

J. Jost, *Riemannian Geometry and Geometric Analysis* (Springer, 2002).

S. Kichenassamy, *Nonlinear Wave Equations* (M. Dekker, 1996).

S. Klainerman & F. Nicolò, *The Evolution Problem in General Relativity* (Birkhäuser, 2003).

S. Kobayashi & K. Nomizu, *Foundations of Differential Geometry, Vols. 1, 2* (Wiley, 1963, 1969).

D. Malament, *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory* (UCP, 2012).

C.W. Misner, K.S. Thorne, J.A. Wheeler, *Gravitation* (Freeman, 1973).

B. O'Neill, *Semi-Riemannian Geometry* (Academic Press, 1983).

R. Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe* (Knopf, 2004).

C.D. Sogge, *Lectures on Non-Linear Wave Equations, 2nd Edition* (International Press, 2008).

A.D. Rendall, *Partial Differential Equations in General Relativity* (OUP, 2008).

H. Ringström, *The Cauchy Problem in General Relativity* (EMS, 2009).

M.E. Taylor, *Partial Differential Equations III: Nonlinear Equations* (Springer, 1996).

R.M. Wald, *General Relativity* (University of Chicago Press, 1984).

Online resources

S. Aretakis, *Lecture Notes on General Relativity* (2013). `https://web.math.princeton.edu/~aretakis/columbiaGR.pdf`.

B. Andrews, *Lectures on Differential Geometry*. `http://maths-people.anu.edu.au/~andrews/DG/`.

P.T. Chruściel, *An introduction to the Cauchy problem for the Einstein equations* (2010). `https://homepage.univie.ac.at/piotr.chrusciel/teaching/Cauchy/Roscoff.pdf`.

H. Friedrich & A.D. Rendall, *The Cauchy Problem for the Einstein Equations* (2000). `arXiv:gr-qc/0002074`.

E. Gourgoulhon, *3+1 Formalism in General Relativity - Bases of Numerical Relativity*. `arXiv:gr-qc/0703035`.

G. Heckman, *Introduction to Riemannian Geometry*. `https://www.math.ru.nl/~heckman/DiffGeom.pdf`.

R. Hocking, T. Talbot, M. Webb, *Existence and Uniqueness for the Vacuum Einstein Equations* (2013). `https://cmouhot.files.wordpress.com/1900/10/project-vacuumeinstein.pdf`.

J. Luk, *Introduction to Nonlinear Wave Equations*. `https://www.dpmms.cam.ac.uk/~jl845/NWnotes.pdf`.

I. Marcut, *Manifolds*. `http://www.math.ru.nl/~imarcut/index_files/lectures_2016.pdf`.

E. Minguzzi, M. Sanchez, The causal hierarchy of spacetimes, `https://arxiv.org/abs/gr-qc/0609119`.

A.D. Rendall, *Theorems on Existence and Global Dynamics for the Einstein Equations.* Living Reviews in Relativity. `https://link.springer.com/article/10.12942/lrr-2005-6`.

R. Schoen, *Topics in Differential Geometry* (2009). `http://math.stanford.edu/~schoen/trieste2012/`.

J.M.M. Senovilla, Singularity theorems and their consequences. `arXiv:1801.04912`. *GRG* 29, 701–848 (1997).

J. Smulevici, *Lectures on Lorentzian geometry and hyperbolic pdes* (2017). `https://www.math.u-psud.fr/~smulevic/lgpdes.pdf`.