

Gravitational Lensing from a Geometric Viewpoint

Volker Perlick

TU Berlin, Institute of Theoretical Physics, Sekr. PN 7-1
10623 Berlin, Germany. email: vper0433@w421zrz.physik.tu-berlin.de

Abstract. The theory of gravitational lensing is discussed in a Lorentzian manifold setting. To that end we fix a point p (observer at a particular instant) and a time-like curve γ (worldline of a light source) in a 4-dimensional Lorentzian manifold (spacetime) and we investigate how many past-pointing lightlike geodesics (light rays) go from p to γ . If there is more than one such geodesic, then we are in a gravitational lensing situation. Among other things, we study the geometry of light cones and we use the theory of conjugate points and cut points to find necessary and sufficient criteria for gravitational lensing; we discuss a Morse theory, based on a general relativistic version of Fermat's principle, to characterize the number of images for gravitational lensing situations in globally hyperbolic spacetimes; and we discuss gravitational lensing in asymptotically simple and empty spacetimes, giving an elementary proof for an odd number theorem in this situation.

1 Introduction

According to general relativity the path of a light ray is influenced by the gravitational field of massive objects. The verification of this effect during a total Sun eclipse in the year 1919 made Einstein's theory famous all over the world. It was soon realized by Eddington [12] and Chwolson [10] that, in principle, this deflection of light by massive objects might lead to the effect that an observer sees two or more distinct images of one and the same light source. Also, Chwolson [10] mentioned the possibility that, in cases of axial symmetry, the light source might appear as a ring around the deflecting mass. Those effects are usually summarized under the name *gravitational lensing*. For many decades it was not clear if gravitational lensing is, indeed, realized in Nature. It was not before 1979 that a promising candidate for gravitational lensing was found. In this year Walsh, Carlswell and Weyman [81] published their results on the double quasar 0957 +561 and suggested that in this case we see two images of one and the same quasar, produced by the gravitational field of an intervening galaxy. Since then, a great number of further gravitational lens candidates have been found, including multiple quasars, radio rings and luminous arcs. This has led to the effect that gravitational lensing is one of the most rapidly developing field in astronomy, in particular from an observational but also from a theoretical point of view. There is a comprehensive monograph on the subject by Schneider, Ehlers and Falco [70]

and there is a great number of review articles, including a regularly updated electronic review by Wambsganss [82] from which literature, in particular on the present status of observations, can be traced.

In this article we want to approach the theory of gravitational lensing from the viewpoint of Lorentzian geometry. This is somewhat unusual insofar as the majority of theoretical work on gravitational lensing is done in a quasi-Newtonian approximation formalism which was developed, in essence, by Sjur Refsdal in the 1960s and which is discussed in full detail, e. g., in Schneider, Ehlers and Falco [70]. In the standard version of this approximation formalism one restricts to a purely spatial description, as opposed to a spacetime description, and light rays are represented by straight lines in Euclidean 3-space, with the only exception that they may have a sharp bend when passing through a particular plane which is known as the “deflector plane”. (There is also a variant with several deflector planes.) This formalism has proven very powerful for calculating particular models. On the other hand, one should keep in mind that it is only an approximation. Thus, it is perfectly fine if it is used for quantitative calculations where the approximative assumptions are satisfied, but it is not the complete story as far as qualitative aspects of the theory are concerned. Gravitational lensing, by its very nature, is a general relativistic effect and it can be understood only on the basis of a 4-dimensional spacetime description, i. e., in terms of Lorentzian geometry.

Therefore, the following strategy seems to be appropriate for studying the theory of gravitational lensing. In the beginning one should concentrate on getting an understanding of gravitational lensing in terms of spacetime diagrams and becoming familiar with the 4-dimensional geometry involved. The present article tries to serve this purpose. After that, one should study the passage to the quasi-Newtonian formalism which involves several approximative assumptions. Some of these assumptions are easily understood, such as that the gravitational field should be weak and that the deflection angles should be small. However, in addition one needs some assumptions whose interpretation is less obvious. So the passage to the quasi-Newtonian approximation is, in fact, a rather subtle issue. These problems are carefully discussed by Seitz, Schneider and Ehlers [73], related material can also be found in Schneider, Ehlers and Falco [70] and in Sasaki [68]. In the third step one is then ready to study how the quasi-Newtonian formalism is operating. This is what is done in the majority of the theoretical literature on gravitational lensing and what is reviewed in full detail, e. g., in Schneider, Ehlers and Falco [70]. For mathematical aspects of gravitational lensing in the quasi-Newtonian approximation formalism, in particular for the theory of caustics, we also refer to a forthcoming book by Petters, Levine and Wambsganss [65].

For our plan to study gravitational lensing in a Lorentzian geometry setting we assume that light propagation can be described in terms of rays and we restrict to light rays in vacuo, i. e., we exclude the case that the light rays

are influenced by a medium, e. g., in terms of diffraction, on their way from the light source to the observer. Moreover, we shall restrict to the case that both the observer and the light source may be considered as pointlike. We are then naturally led to studying past-pointing lightlike geodesics (light rays) from a point (observer at a particular instant) to a timelike curve (worldline of the light source) in a Lorentzian manifold (spacetime). If there are two or more such geodesics, then we are in a gravitational lensing situation. In essence, our analysis will be kinematical throughout, although Einstein's field equation will be mentioned occasionally.

The article is organized as follows. In Section 2 we recall some basic notions from Lorentzian geometry and fix some conventions as to terminology and notation. These conventions will be essential for understanding the following, so the reader is kindly requested to read this section carefully. In Section 3 we study gravitational lensing situations in arbitrary spacetimes. In Section 4 we specialize to the case of globally hyperbolic spacetimes and in Section 5 we further specialize to asymptotically simple and empty spacetimes.

Many mathematical results will be simple corollaries of standard theorems from Lorentzian geometry which can be found in the books by Hawking and Ellis [29], Wald [80], O'Neill [50], or Beem, Ehrlich and Easley [4]. For theorems proven in one of those books the proof is not repeated here unless in cases where this seemed instructive. Also, the material on Morse theory, i. e., Subsections 3.5 and 4.2, turned out to be so technical that for proves the reader must be referred to the quoted original papers. In all other cases proves are given in (hopefully) sufficient detail.

2 Some Basic Notions of Spacetime Geometry

According to general relativity, a spacetime is a 4-dimensional Lorentzian manifold. For convenience we shall consider only Lorentzian manifolds that are *time-orientable*, i. e., we assume that it is possible to distinguish between future and past in a globally consistent way. More precisely, we use the following definition.

Definition 1. A *spacetime* is a triple $(\mathcal{M}, g, \mathcal{T}^+)$ where

- (a) \mathcal{M} is a connected 4-dimensional real C^∞ manifold whose topology satisfies the axiom of second countability and the Hausdorff axiom and is, thus, paracompact;
- (b) g is a Lorentzian metric on \mathcal{M} , i. e., a symmetric covariant second rank C^∞ tensor field which has signature $(+, +, +, -)$ at each point;
- (c) \mathcal{T}^+ is a *time orientation* for (\mathcal{M}, g) , i. e., the set of timelike tangent vectors $\{X \in T\mathcal{M} \mid g(X, X) < 0\}$ consists of exactly two connected components and \mathcal{T}^+ is one of those components.

For a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, we denote the *Levi-Civita connection* of g by ∇ and we denote the *Riemannian curvature tensor* of ∇ by R .

A linear subspace N of the tangent space $T_p\mathcal{M}$ is called (i) *spacelike* if g is positive definite on N , (ii) *lightlike* if g is positive semi-definite but not positive definite on N , and (iii) *timelike* otherwise. The property of being spacelike, lightlike or timelike is assigned to a vector in $T_p\mathcal{M}$ if the linear space generated by this vector has the respective property, to a differentiable curve if its tangent vector has the respective property everywhere and to a submanifold of \mathcal{M} if its tangent space has the respective property everywhere. Moreover, a differentiable curve is called *causal* if its tangent vector is either timelike or lightlike at each point.

Here and in the following, by a *curve* in \mathcal{M} we mean a map from a real interval I into \mathcal{M} . (A “real interval” is a connected subset of \mathbb{R} that contains more than one point. I may be open, half-open or closed.) Quite generally, we shall use a lower case greek letter for a curve, e. g., $\gamma : I \rightarrow \mathcal{M}$, and we shall use the corresponding boldface letter for the image set of this curve, e. g., $\gamma = \{\gamma(s) \mid s \in I\}$.

By a *geodesic* we always mean what is more fully called an “affinely parametrized geodesic”, i. e., a C^∞ curve $\lambda : I \rightarrow \mathcal{M}$ such that $\nabla_{\lambda'}\lambda' = 0$. This leaves the freedom of changing the parameter affinely, $I \rightarrow \tilde{I}$, $s \mapsto as + b$ with $a, b \in \mathbb{R}$, $a \neq 0$. However, when counting geodesics we shall tacitly identify two geodesics if one is a reparametrization of the other. That is to say, in a sentence such as “There are two geodesics λ_1 and $\lambda_2 \dots$ ” it goes without saying that λ_2 is not just a reparametrization of λ_1 . Please note that, according to this rule, a periodic geodesic gives rise to infinitely many geodesics between any two points on this geodesic.

Our study will be concentrating upon lightlike geodesics, which are to be interpreted as light rays. It is well-known and easily verified that under a conformal transformation $g \mapsto e^{2f}g$ of the metric g , with an arbitrary C^∞ function $f : \mathcal{M} \rightarrow \mathbb{R}$, the lightlike geodesics undergo a reparametrization but are unchanged otherwise. Since we are interested only in the paths of lightlike geodesics and not in their particular parametrizations, we could therefore allow for arbitrary conformal transformations of the metric, i. e., we could prescribe a conformal equivalence class rather than a particular metric. However, we shall not do so because we want to occasionally discuss additional assumptions on the spacetime which are not conformally invariant, such as, e. g., conditions on the Ricci tensor.

The totality of all geodesics issuing from a point p give rise to the *exponential map*

$$\exp_p : \mathcal{W}_p \rightarrow \mathcal{M}. \quad (1)$$

This map is defined on a subset \mathcal{W}_p of the tangent space $T_p\mathcal{M}$ by setting $\exp_p(X) = \lambda(1)$ where $\lambda : [0, 1] \rightarrow \mathcal{M}$ is the geodesic with $\lambda'(0) = X$. It is well known that the maximal domain \mathcal{W}_p on which this map is well-defined is an open subset of $T_p\mathcal{M}$ that contains the origin. In general, \mathcal{W}_p is not all of $T_p\mathcal{M}$, thereby reflecting the fact that a geodesic may arrive at the “boundary” of \mathcal{M} (using the word “boundary” in a colloquial manner) before its affine parameter has reached the value 1. In general, the map \exp_p

need not be injective on its maximal domain \mathcal{W}_p , thereby indicating that the geodesics issuing from p may reconverge and, eventually, intersect each other. However, it is well known that there is an open neighborhood \mathcal{W}_p^o of the origin in $T_p\mathcal{M}$ such that the restriction of \exp_p to \mathcal{W}_p^o is a diffeomorphism onto its image. Thus, sufficiently short pieces of geodesics issuing from p do not intersect. A neighborhood of p that is contained in the image of \mathcal{W}_p^o under \exp_p is called a *normal neighborhood*.

This notion can be used to assign the property of being timelike or causal to continuous curves which need not be differentiable. A curve $\gamma : I \rightarrow \mathcal{M}$ is called *timelike* (or *causal*, respectively) if it is continuous and if each $s \in I$ has a neighborhood \tilde{I} in I such that for any two parameter values s_1 and s_2 in \tilde{I} there is a timelike (or causal, respectively) geodesic from $\gamma(s_1)$ to $\gamma(s_2)$ which is completely contained in a normal neighborhood of $\gamma(s)$. If this geodesic is future-pointing whenever $s_1 < s_2$, γ is called *future-pointing*; otherwise γ is called *past-pointing*.

Moreover, we shall frequently use the following standard definition.

Definition 2. For a point p in a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, we define the *chronological future* $\mathcal{I}^+(p)$ (and the *chronological past* $\mathcal{I}^-(p)$, respectively) of p as the set of all points $q \in \mathcal{M}$ that can be reached from p along a future-pointing (or past-pointing, respectively) timelike curve.

$\mathcal{I}^+(p)$ and $\mathcal{I}^-(p)$ are obviously open subsets of \mathcal{M} for every $p \in \mathcal{M}$. However, as long as the causal structure of spacetime has not been restricted this is more or less the only statement that can be made about these two sets. The following causality notions will be of relevance for us.

Definition 3. (a) A spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ is called *causal* at a point $p \in \mathcal{M}$ if there is no closed causal curve through p .

(b) A spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ is called *future-distinguishing* (or *past-distinguishing*, respectively) at p if every neighborhood \mathcal{U} of p contains a neighborhood \mathcal{V} of p such that a future-pointing (or past-pointing, respectively) causal curve from p that has left \mathcal{U} cannot reenter \mathcal{V} . An equivalent condition is that the equation $\mathcal{I}^+(p) = \mathcal{I}^+(q)$ (or the equation $\mathcal{I}^-(p) = \mathcal{I}^-(q)$, respectively) implies the equation $p = q$.

(c) A spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ is called *strongly causal* at p if every neighborhood \mathcal{U} of p contains a neighborhood \mathcal{V} of p such that no causal curve intersects \mathcal{V} more than once.

It is easy to check that the strong causality condition implies both the future-distinguishing and the past-distinguishing condition and that either distinguishing condition implies the causality condition. For a rather detailed discussion of these well known notions we refer to Hawking and Ellis [29], to O'Neill [50] and to Beem, Ehrlich and Easley [4]. In particular, illustrative examples of spacetimes satisfying some causality assumptions but violating others are given in Figures 37 and 38 of Hawking and Ellis [29].

3 Gravitational Lensing in Arbitrary Spacetimes

In this section we discuss the geometry of gravitational lensing situations in arbitrary spacetimes $(\mathcal{M}, g, \mathcal{I}^+)$. To that end we fix a point $p \in \mathcal{M}$ and a timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$. We interpret p as an event where an observation takes place (i. e. “an astronomer here and now in his observatory”) and we interpret γ as the worldline of a light source, e. g., a distant quasar. This interpretation is, of course, based on the assumption that the spatial extension of the light source and of the observer can be neglected, i. e., that they can be considered as pointlike. Our assumption of γ being timelike means that the light source moves at a subluminal velocity. There is no need to specify the parametrization of γ (e. g., to proper time parametrization $g(\gamma', \gamma') = -1$) since we are primarily interested in the set γ and not in a particular parametrization. – The question we want to discuss is the following (see Figures 1 and 2).

How many past-pointing lightlike geodesics are there that start at the point p and terminate on the worldline γ ?

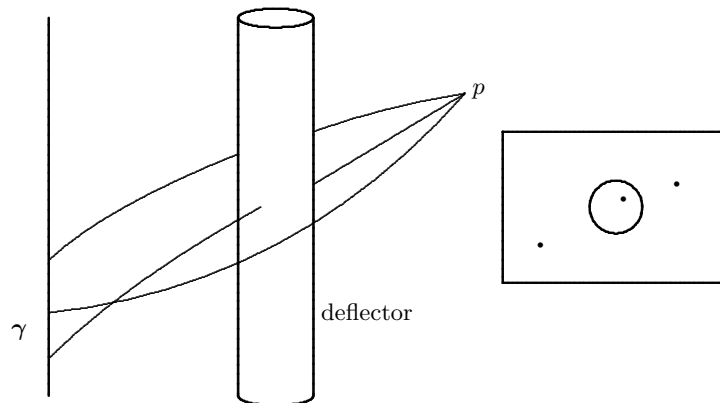


Fig. 1. This is a typical spacetime diagram of a multiple imaging situation. In correspondence with the three past-pointing lightlike geodesics from p to γ , an observer at p would see three images of a light source with worldline γ , as indicated in the insert. Actually, one of the three images is hidden behind the deflector if the latter is non-transparent.

According to the rules of general relativity, every lightlike geodesic can be interpreted as a light ray traveling under the influence of gravity alone. (As

outlined in the Introduction, we shall not be concerned with light rays influenced by a medium throughout this text.) Thus, each past-pointing lightlike geodesic from p to γ gives rise to an image of the light source γ at the celestial sphere of the observer p . To justify this interpretation one may view each lightlike geodesic as representing a thin bundle of almost parallel light rays which is focused by the observer's eye lense onto his or her retina. – The following cases are to be distinguished.

Case A: *There is no past-pointing lightlike geodesic from p to γ .* Then the observer at p does not see any image of the light source γ . Situations of this kind are far from being unusual. They may occur even for an inextendible worldline γ in Minkowski space, viz., if γ asymptotically approaches the past light cone of p . Please note that, in general, the non-existence of a past-pointing lightlike geodesic from p to γ does not imply the non-existence of a past-pointing causal curve from p to γ . In other words, even if p cannot receive a freely traveling light ray from γ it is very well possible that p can be causally influenced by γ . The Gödel cosmos provides an interesting example of this kind, see, e. g. Hawking and Ellis [29], Section 5.7. In this spacetime any two points can be joined by a past-pointing causal curve; however, the lightlike geodesics issuing from some point are restricted to a cylindrical region.

Case B: *There is exactly one past-pointing lightlike geodesic from p to γ .* Then the observer at p sees exactly one image of the light source γ . This is the situation naively taken for granted in pre-relativistic astronomy.

Case C: *There are at least two but not more than denumerably many past-pointing lightlike geodesics from p to γ .* Then the observer at p sees finitely or infinitely many distinct images of γ at his or her celestial sphere. In view of Einstein's field equation one may think of a heavy mass ("deflector"), occupying a worldtube between γ and p , whose gravitational field causes a bending of light rays. Figure 1 shows a typical three-image-configuration. Please note that, generically, different lightlike geodesics from p to γ intersect the worldline γ at different points. In other words, the various images seen at p show the light source at different ages. Astronomers use the term *time delay* for this phenomenon.

Case D: *There are more than denumerably many past-pointing lightlike geodesics from p to γ .* E. g., there may be a continuous one-parameter family of lightlike geodesics from p to γ such that the light source γ appears at the celestial sphere of the observer p as an arc or, in situations of axially symmetry, as a ring, see Figure 2. Such rings are often called *Einstein rings* although Chwolson [10] and not Einstein was the first to mention this phenomenon. We shall prove later in Proposition 12 that all members of a continuous one-parameter family of light rays from p to γ necessarily meet γ at the same point. In other words, all parts of such an extended image show the light source at the same age.

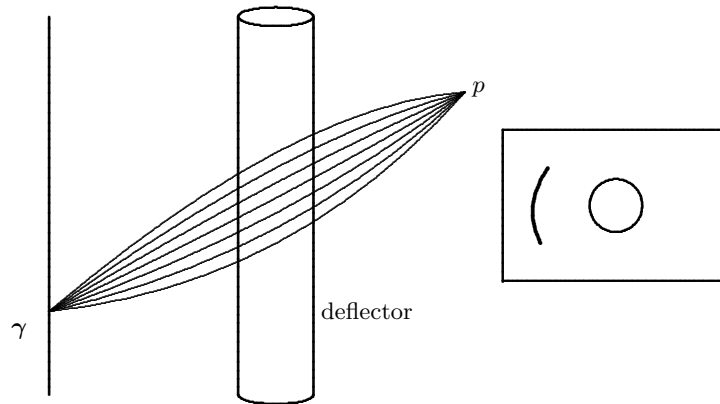


Fig. 2. In the situation depicted in this spacetime diagram there is a one-parameter family of past-pointing lightlike geodesics from p to γ . Correspondingly, an observer at p would see an extended image, such as an arc, of a light source with worldline γ . The insert indicates what is seen by the observer at p .

Whenever Case C or Case D occurs astronomers speak of *multiple imaging* by the *gravitational lens effect*. In Proposition 10 below we shall prove that Case D is “exceptional” in the sense that, under a small perturbation of the point p (keeping the worldline γ fixed), Case D always disappears. In this sense, Case C is the “generic” multiple imaging case. However, this does not mean that Case D is of no interest from a physical point of view. First, a systematic investigation of Case D situations in a spacetime will give interesting information on Case C situations. Second, the “non-genericity” of Case D situations is based on our idealization to view the light source and the observer as spatially non-extended, i. e., as pointlike. For extended light sources it is not true that Case D situations can be destroyed by an arbitrarily small perturbation. That is the reason why, actually, astronomers do observe arcs and rings produced by gravitational lensing.

It is now our goal to characterize spacetime geometries that lead to multiple imaging, with special emphasis on the difference between Case C and Case D situations. This will be done in the next three subsections. We begin with a review of the notions of conjugate points and of cut points in Subsection 3.1. These notions will be instrumental to study the local and global geometry of light cones in Subsection 3.2 and, thereupon, to give criteria for multiple imaging in Subsection 3.3. Subsections 3.4 and 3.5 are devoted to a version of Fermat’s principle on arbitrary spacetimes that is very useful in studying gravitational lensing situations.

3.1 Conjugate Points and Cut Points

The notion of conjugate points is used to characterize the situation that neighboring geodesics undergo a (partial) focusing effect. We are interested in lightlike geodesics in a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ that pass through a particular point p . (In applications to gravitational lensing this will be the point where the observation takes place and we shall be interested in lightlike geodesics that issue from p into the past.) In other words, we are interested in C^∞ curves $\lambda : I \rightarrow \mathcal{M}$ that satisfy the conditions

$$\begin{aligned} \nabla_{\lambda'} \lambda' &= 0, & (2) \\ g(\lambda', \lambda') &= 0, & (3) \\ \lambda'(s_o) &= p, & (4) \end{aligned}$$

where s_o denotes a fixed parameter value $s_o \in I$.

To investigate the behavior of geodesics which are close to λ we consider a one-parameter variation of λ , i. e., a C^∞ map $\eta :]-\varepsilon_o, \varepsilon_o[\times I \rightarrow \mathcal{M}$ with $\eta(0, \cdot) = \lambda$ where ε_o is some positive real number. We assume that not only λ but also all the varied curves $\eta(\varepsilon, \cdot)$, for $0 < |\varepsilon| < |\varepsilon_o|$, satisfy the three conditions (2), (3), (4). By differentiation with respect to the variational parameter ε this implies that the variational vector field $J : I \rightarrow T\mathcal{M}$, which is defined by $J(s) = \eta(\cdot, s)'(0)$, satisfies the three conditions

$$\begin{aligned} \nabla_{\lambda'} \nabla_{\lambda'} J - R(\lambda', J, \lambda') &= 0, & (5) \\ g(\lambda', \nabla_{\lambda'} J) &= 0, & (6) \\ J(s_o) &= 0, & (7) \end{aligned}$$

see Figure 3. For any vector field J along λ that satisfies these three conditions, one may think of the “arrow-head” of J as tracing a neighboring lightlike geodesic through p in linear approximation. (5) is called the *equation of geodesic deviation* or the *Jacobi equation* and any solution J of this equation is called a *Jacobi field* along λ .

The equations (5), (6) and (7) are obviously satisfied by any multiple of the tangent field, $J(s) = f(s) \lambda'(s)$, with $f(s_o) = 0$. Such a solution of (5), (6) and (7) is called *trivial* since it represents an infinitesimally neighboring geodesic which is just a reparametrization of λ . We are now ready to define the notion of conjugate points. For $s_1 \in I \setminus \{s_o\}$, one says that the point $q = \lambda(s_1)$ is *conjugate* to $p = \lambda(s_o)$ along λ if there is a non-trivial solution J of (5), (6) and (7) such that $J(s_1)$ is parallel to $\lambda'(s_1)$. It is easy to check that the conjugacy of q to p along λ is independent of which affine parametrization has been chosen for λ . Also, it is known that in a compact section of a lightlike geodesic there are at most finitely many points conjugate to a given point p , see, e. g., Beem, Ehrlich and Easley [4], Theorem 10.77. (The same result is true for timelike geodesics as well, but not for spacelike ones. An example where a whole interval is conjugate to a point along a spacelike geodesic was contrived by Helfer [30]. The reader is cautioned against a proof that conjugate points are always isolated, along any geodesic in a semi-Riemannian manifold of any signature, suggested by O’Neill [50], Exercise 8, p. 299. This

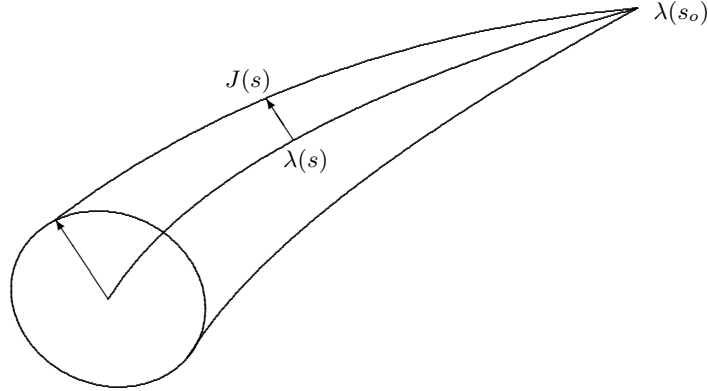


Fig. 3. The solutions J of (5), (6) and (7) form an “infinitesimal bundle of light rays” around λ . The timelike dimension is suppressed in the picture.

is one of the very few mistakes in this otherwise excellent text-book; a basis of Jacobi fields with the desired properties need not exist.)

It follows directly from the definitions that a conjugate point indicates a partial focusing effect in the following sense. If, for a lightlike geodesic λ , the point $\lambda(s_1)$ is conjugate to the point $\lambda(s_o)$, then a one-parameter family of lightlike geodesics issuing from $\lambda(s_o)$ is being refocused into the point $\lambda(s_1)$ to within linear approximation.

In addition, conjugate points indicate that a geodesic loses its extremizing property, according to the following well-known proposition.

Proposition 1. *Let $\lambda : I \rightarrow \mathcal{M}$ be a lightlike geodesic in a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ and consider two different parameter values s_o and s_1 in I . If, for some $s \in]s_o, s_1[$, the point $\lambda(s)$ is conjugate to $\lambda(s_o)$ along λ , then there is a C^∞ variation of $\lambda|_{[s_o, s_1]}$ such that all varied curves are timelike curves from $\lambda(s_o)$ to $\lambda(s_1)$. Conversely, the existence of such a variation implies that there must be a point $\lambda(s)$ conjugate to $\lambda(s_o)$ in the half-open interval $s \in]s_o, s_1]$.*

For a proof we refer to Hawking and Ellis [29], Proposition 4.5.11 and Proposition 4.5.12. It might be helpful to consult O’Neill [50], Chapter 10, Proposition 48, in addition.

Proposition 1 says that an observer moving at subluminal velocity may catch up with a light ray λ after the latter has passed through a conjugate point and that, for observers staying close to λ , this is impossible otherwise. Here the restriction to observers staying close to λ is essential. Observers “taking a short cut” may very well catch up with a lightlike geodesic even if the latter is free of conjugate points. For investigating the extremizing property of a lightlike geodesic from a global point of view, not just with respect

to neighboring curves, the notion of conjugate points is not appropriate. Instead, one has to consider the notion of “cut points” which was introduced in Riemannian geometry (i. e., for positive definite metrics) by Poincaré [66] for a special situation and by Whitehead [85] in generality. For lightlike geodesics in a Lorentzian manifold, this notion can be introduced in the following way, cf. Beem, Ehrlich and Easley [4].

For any two points p and q in \mathcal{M} , we denote by $d(p, q)$ the Lorentzian pseudo-distance between p and q , i. e.

$$d(p, q) = \sup_{\beta} \int_0^1 \sqrt{|g(\beta'(s), \beta'(s))|} ds, \quad (8)$$

where the supremum is to be taken over all causal curves $\beta : [0, 1] \rightarrow \mathcal{M}$ with $\beta(0) = p$ and $\beta(1) = q$. Owing to the well-known fact (cf. Beem, Ehrlich and Easley [4], p. 75) that any causal curve is differentiable almost everywhere, it is not necessary to restrict to differentiable causal curves to make sure that the integral in (8) does exist. Whenever p and q are causally related, such that the supremum is to be taken over a non-empty set, the existence of the supremum is guaranteed but it may be infinite.

Now let $\lambda : I \rightarrow \mathcal{M}$ be a past-pointing lightlike geodesic. For s_o and s_1 in I with $s_o < s_1$, $\lambda(s_1)$ is called the *past cut point* of $\lambda(s_o)$ along λ if $d(\lambda(s_o), \lambda(s)) = 0$ for $s \in]s_o, s_1]$ and $d(\lambda(s_o), \lambda(s)) > 0$ for all $s \in I$ with $s > s_1$. Thus, the past cut point occurs where λ loses its extremizing property among causal curves with respect to the pseudo-distance. In this situation, for $s > s_1$ the point $\lambda(s)$ can be reached from $\lambda(s_o)$ along a past-pointing causal curve which is not a lightlike geodesic. By a well-known theorem (see, e. g., Hawking and Ellis [29], Proposition 4.5.10) this implies that $\lambda(s)$ can be reached from $\lambda(s_o)$ along a past-pointing timelike curve. Thus, beyond the past cut point λ intersects some past-pointing timelike curve which started together with λ at $\lambda(s_o)$.

It is not difficult to check that the past cut point of p along λ is independent of which past-pointing affine parametrization has been chosen for λ . Also, it follows directly from the definition that the past cut point is unique if it exists. The non-existence of the past cut point may have two quite different reasons. Either $d(\lambda(s_o), \lambda(s)) = 0$ for all $s \in I$ with $s > s_o$, i. e., the extremizing property is always preserved; or $d(\lambda(s_o), \lambda(s)) > 0$ for all $s \in I$ with $s > s_o$, i. e., the extremizing property never holds. Clearly, the latter case is possible only if the past distinguishing condition of Definition 3 (b) is violated. In other words, the past-distinguishing property guarantees that a sufficiently short past-pointing lightlike geodesic is extremizing.

There is, of course, a completely analogous definition of the *future cut point* along a lightlike geodesic. However, we shall concentrate upon the past cut point because this will be the relevant notion in view of gravitational lensing situations.

In Riemannian geometry the notions of cut points and conjugate points are related by the easily remembered rule: “The cut point comes first”, see, e. g., Klingenberg [35], Proposition 2.1.7. The proof of this result is based on the well-known fact that a sufficiently short Riemannian geodesic always extremizes the Riemannian distance. We have just seen that for lightlike geodesics in Lorentzian manifolds the analogous fact need not be true unless the distinguishing property is satisfied. Therefore, the rule “The cut point comes first” can be proven for distinguishing spacetimes only. The precise statement reads as follows.

Proposition 2. *Let $\lambda : I \rightarrow \mathcal{M}$ be a past-pointing lightlike geodesic in a spacetime that satisfies the past-distinguishing property at the point $p = \lambda(s_o)$. Assume that, for some parameter value $s_1 > s_o$, the point $\lambda(s_1)$ is conjugate to $\lambda(s_o)$ along λ . Then the past cut point $\lambda(s)$ of $\lambda(s_o)$ along λ exists and it is $s_o < s \leq s_1$.*

Proof. Since the past-distinguishing property is satisfied at p , the equation $d(p, \lambda(s)) = 0$ holds for $s \in [s_o, s_o + \varepsilon[$ whenever ε is a sufficiently small positive number. By Proposition 1, $d(p, \lambda(s)) > 0$ for $s \in]s_1, s_1 + \delta[$ for arbitrarily small positive δ . Thus, the past cut point must lie in the parameter interval $]s_o, s_1]$. \square

We end this subsection with a proposition saying that, under the past-distinguishing assumption, any intersection of two past-pointing lightlike geodesics starting from a point p is indicated by the occurrence of a past cut point on each of those geodesics, please cf. Beem, Ehrlich and Easley [4], Lemma 9.13.

Proposition 3. *Let $(\mathcal{M}, g, \mathcal{T}^+)$ be a spacetime that satisfies the past-distinguishing condition at a point $p \in \mathcal{M}$. Assume that there is a point $q \in \mathcal{M}$ that can be reached from p along two past-pointing lightlike geodesics. Then the past cut point of p exists on each of those geodesics (and it comes on or before q).*

Proof. We may parametrize the two past-pointing lightlike geodesics such that $\lambda_1(0) = \lambda_2(0) = p$ and $\lambda_1(1) = \lambda_2(1) = q$. Then $\lambda_1'(1)$ and $\lambda_2'(1)$ are linearly independent since otherwise λ_2 would be a reparametrization of λ_1 . This follows from the uniqueness theorem for the geodesic equation and from the fact that, owing to our past-distinguishing condition, a closed lightlike geodesic through p cannot exist. Then, for small positive ε , the curve $\lambda_1|_{[0,1]}$ joined to the curve $\lambda_2|_{[1,1+\varepsilon]}$ gives a causal curve which is not an (unbroken) lightlike geodesic. By a well-known theorem (see Hawking and Ellis [29], Proposition 4.5.10) this implies that the point $\lambda_2(1 + \varepsilon)$ can be reached from $p = \lambda_1(0)$ by a timelike curve, thus $d(p, \lambda_2(1 + \varepsilon)) > 0$. On the other hand, the past-distinguishing condition at p guarantees that $d(p, \lambda_2(s_o + \delta)) = 0$ for small positive δ . Hence, the past cut point of p along λ_2 comes on or before $q = \lambda_2(1)$. \square

Moreover, one might ask if the past cut point itself can be reached from p along a second past-pointing lightlike geodesic. A theorem to that effect can be proven only under the assumption of global hyperbolicity and will be postponed until Section 4, see Proposition 14 below. In the Riemannian case, an analogous result holds on *complete* Riemannian manifolds and is the content of the celebrated *Poincaré Theorem*, proven by Poincaré [66] for a special case and by Whitehead [85] in its generality. It is this property that gave rise to the name “cut point”.

3.2 The Geometry of Light Cones

In a spacetime, the lightlike geodesics issuing from a point p into the past make up the so-called *past light cone* of p . In general, the past light cone need not be an immersed (let alone embedded) submanifold of \mathcal{M} , i. e., even its local structure may be very complicated. The failure of past light cones to be submanifolds is crucial for gravitational lensing. In this subsection we use the notions of conjugate points and cut points to investigate whether the past light cone is an immersed or embedded submanifold.

Please recall that the totality of all geodesics issuing from a point p are given in terms of the exponential map (1). For our study of gravitational lensing we are interested in lightlike geodesics issuing from p into the past. Therefore, we restrict the exponential map to the 3-dimensional submanifold

$$\mathcal{C}_p^- = \{X \in \mathcal{W}_p \setminus \{0\} \mid X \text{ is lightlike and past-pointing}\} \quad (9)$$

of $T_p\mathcal{M}$. In (9) $\mathcal{W}_p \subseteq T_p\mathcal{M}$ denotes the maximal domain of \exp_p . For the sake of convenience, we introduce the abbreviation

$$e_p^- = \exp_p|_{\mathcal{C}_p^-} : \mathcal{C}_p^- \longrightarrow \mathcal{M} \quad (10)$$

for the restriction of the exponential map to \mathcal{C}_p^- . The image of this map e_p^- is the *past light cone* of p , i. e., the set of all events $q \in \mathcal{M}$ that can be reached from p along a past-pointing lightlike geodesic. This set determines what is visible for an observer at p . In particular, it determines whether p observes a gravitational lensing situation.

e_p^- is a C^∞ map from a 3-dimensional manifold into a 4-dimensional manifold. Thus, at any $X \in \mathcal{C}_p^-$ the rank of the differential $T_X e_p^-$ cannot be bigger than 3. If the rank is equal to 3, e_p^- is an immersion at X , i. e., the past light cone with vertex p is an immersed submanifold near the point $e_p^-(X)$. This is, of course, necessarily the case for vectors $X \in \mathcal{C}_p^- \cap \mathcal{W}_p^o$, where \mathcal{W}_p^o denotes the domain on which the exponential map is a diffeomorphism. This reflects the well-known fact that the past light cone with vertex p is a 3-dimensional manifold if we restrict to sufficiently short lightlike geodesics issuing from p . For $X \in \mathcal{C}_p^- \setminus \mathcal{W}_p^o$, however, the number $m = 3 - \text{rank}(T_X e_p^-)$ may be bigger than 0. Comparison with the preceding subsection shows that

this is the case if and only if the point $e_p^-(X)$ is conjugate to p along the geodesic generated by X . The number m is called the *multiplicity* of this conjugate point. Since, obviously, $(T_X e_p^-)(X) \neq 0$, the multiplicity m may be either 1 or 2. It is important to realize that, in any case, the image of the differential $T_X e_p^-$ is a lightlike subspace. For a proof it suffices to realize that (2), (6) and (7) imply $g(\lambda', J) = 0$, so the tangent vector of λ and all “connecting vectors” with infinitesimally neighboring lightlike geodesics starting from the same point span a lightlike subspace. (Here we make use of the well-known fact that a timelike vector cannot be orthogonal to a lightlike vector, i. e., that the equations $g(\lambda', \lambda') = 0$ and $g(\lambda', J) = 0$ imply the equation $g(J, J) \geq 0$.) In particular, this proves the well-known fact that the light cone is a 3-dimensional *lightlike* submanifold at each point where $T_X e_p^-$ has maximal rank.

The union of all points which are conjugate to p , along any past-pointing lightlike geodesic issuing from p , is called the *past lightlike conjugate locus* of p or the *caustic* of the past light cone of p . In other words, the caustic is the set of all points where the past light cone fails to be an immersed submanifold of \mathcal{M} . At caustic points, the light cone typically forms edges or vertices whose geometry might be arbitrarily complicated. If one restricts to caustics which are *stable* against perturbations in a certain sense, then a local classification of caustics is possible with the help of Arnold’s singularity theory of Lagrangian or Legendrian maps, see Arnold, Gusein-Zade and Varchenko [3] or Arnold [2]. This formalism has been applied to *wavefronts* in general relativity, a notion which includes light cones as special cases, by Friedrich and Stewart [18], by Hasse, Kriele and Perlick [28] and, in a particularly elegant way, by Low [43]. (In [28] the proof of Theorem 4.4 is incorrect. A corrected version is going to appear.) In the case of globally hyperbolic spacetimes the formalism of Low even allows to tackle the problem of *globally* classifying the caustics of light cones, although this has not been carried through until now. For the sake of comparison the reader should also consult Petters’ work [62] [64] [65] on caustics in the quasi-Newtonian approximation formalism of gravitational lensing. Unfortunately, the subject of classifying stable caustics is so technical that we cannot go into this matter here for lack of space.

It is important to realize that a light cone may fail to be an embedded submanifold of \mathcal{M} even if its caustic is empty. At the end of this subsection we shall illustrate this claim by an example where a light cone develops transverse self-intersections without ever failing to be an immersed submanifold of \mathcal{M} , see Figure 4 below. The relevant notion for finding out whether a light cone is an embedded submanifold is the notion of cut points, and not the notion of conjugate points. To work this out, we have to take a closer look at the chronological past $\mathcal{I}^-(p)$ of a point p , please recall Definition 2.

In Minkowski space, the lightlike geodesics issuing from p into the past make up the boundary $\partial \mathcal{I}^-(p)$ of $\mathcal{I}^-(p)$. In spacetimes with a complicated causal structure, however, those lightlike geodesics may penetrate into the

open set $\mathcal{I}^-(p)$, i. e., the past light cone of p may have a non-void intersection with $\mathcal{I}^-(p)$. In spacetimes with drastic causality violations (such as, e. g., the Gödel cosmos, see Hawking and Ellis [29], Section 5.7) $\mathcal{I}^-(p)$ may even be all of \mathcal{M} such that $\partial\mathcal{I}^-(p)$ is empty and the past light cone of p is completely contained in $\mathcal{I}^-(p)$. Quite generally, $\partial\mathcal{I}^-(p)$ can be characterized in the following way.

Proposition 4. *For any point p in a spacetime, the set $\partial\mathcal{I}^-(p)$ is either empty or a 3-dimensional achronal closed embedded C^{1-} submanifold of \mathcal{M} . (A subset of a spacetime is achronal if it is impossible to connect any two of its points by a timelike curve. A C^{1-} manifold is a topological manifold whose transition maps satisfy a Lipschitz condition.)*

For a proof we refer to Hawking and Ellis [29], Proposition 6.3.1. With the Lorentzian distance function d defined by (8), we get the following result for a past-pointing lightlike geodesic λ with $\lambda(s_o) = p$. A point $\lambda(s)$ with $s > s_o$ is in the boundary of $\mathcal{I}^-(p)$ if $d(p, \lambda(s)) = 0$ and it is in the open set $\mathcal{I}^-(p)$ if $d(p, \lambda(s)) > 0$. Thus, the past cut point of p along a lightlike geodesic can be characterized as the point where this geodesic leaves the boundary of $\mathcal{I}^-(p)$ and penetrates into the open set $\mathcal{I}^-(p)$. The set of all past cut points of p along lightlike geodesics through p is called the *past lightlike cut locus* of p . We shall now prove that in past-distinguishing spacetimes the failure of a past light cone to be an embedded submanifold is indicated by a non-empty past lightlike cut locus.

Proposition 5. *Assume that the spacetime $(\mathcal{M}, g, \mathcal{I}^+)$ satisfies the past-distinguishing property at a point p . If the past lightlike cut locus of p is empty, then the map e_p^- defined in (10) is a C^∞ embedding, i. e., the past light cone of p is an embedded C^∞ submanifold of \mathcal{M} .*

Proof. If the past lightlike cut locus is empty, Proposition 2 implies that no past-pointing lightlike geodesic starting at p can have a point conjugate to p . Hence, the map e_p^- is a C^∞ immersion. Together with the past-distinguishing condition, the same assumption implies that such a geodesic must stay on $\partial\mathcal{I}^-(p)$ forever, i. e., the past light cone of p must be completely contained in $\partial\mathcal{I}^-(p)$. But then Proposition 4 guarantees that the past light cone has no self-intersection and no almost self-intersection. Hence, e_p^- must even be an embedding. \square

In Section 4 below we shall prove that the converse of this proposition is true in globally hyperbolic spacetimes, see Proposition 15. – We now illustrate the properties of conjugate loci and cut loci with three examples.

Example 1:

Figure 4 shows the past light cone of a point p in a spacetime with a non-transparent deflector. To have a concrete example, the reader may consider the spacetime metric

$$g = -dt^2 + dz^2 + dr^2 + k^2 r^2 d\varphi^2, \quad (11)$$

with some constant $0 < k < 1$, on $\mathcal{M} = \mathbb{R}^2 \times (\mathbb{R}^2 \setminus \{0\})$. Here (t, z) denote Cartesian coordinates on \mathbb{R}^2 and (r, φ) denote polar coordinates on $\mathbb{R}^2 \setminus \{0\}$. This can be interpreted as the spacetime around a static non-transparent string, see Vilenkin [79], Hiscock [32] and Gott [24]. (Vilenkin in his pioneering paper discussed this metric in connection with the linearized Einstein field equation; it was then realized independently by Hiscock and Gott that Vilenkin's results remain true even if the full Einstein equation is used.) One should think of the string as being situated at the z -axis. Since the latter is not part of the spacetime, it is indeed justified to speak of a *non-transparent* string. It is easy to see that the metric (11) induces on each plane $t = \text{const.}$, $z = \text{const.}$ the geometry of a cone; i. e., this metric has a “conic singularity” along the z -axis.

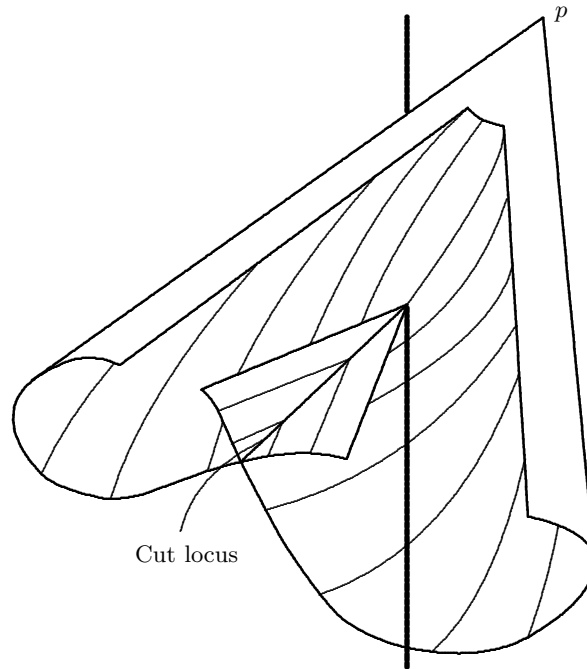


Fig. 4. In a spacetime with non-transparent deflector, e. g., a non-transparent string, light cones may develop self-intersections without failing to be immersed submanifolds. It is geometrically evident that the situation depicted here gives rise to double imaging for an observer at p .

It is an instructive exercise to verify that in this string spacetime each past light cone qualitatively looks like the one depicted in Figure 4. (Clearly, in Figure 4 one spatial dimension is suppressed. This missing dimension corresponds to the z direction in the string example. The fat vertical line in Fig-

ure 4 actually indicates that a two-dimensional world sheet has been excised from spacetime, viz., the total history of the z -axis in the string example.) The caustic of this light cone is empty, i. e., there are no points conjugate to p along any past-pointing lightlike geodesic from p . The past lightlike cut locus of p , however, is not empty, thereby illustrating our earlier claim that a light cone may fail to be an embedded submanifold without failing to be an immersed submanifold. Moreover, Figure 4 nicely exemplifies our general result that each past-pointing lightlike geodesic from p enters $\mathcal{I}^-(p)$ exactly when passing through the cut locus.

Example 2:

We now modify Example 1 by switching to a transparent deflector. In the case of the string metric (11) this can be done by changing the metric in the neighborhood of the z -axis in such a way that there is no longer a singularity, i. e., by “rounding off the tip of the cone” which represents each plane $t = \text{const.}$, $z = \text{const.}$ The region around the z -axis in which the metric has been changed can then be interpreted as the interior region of a transparent string. The resulting light cone looks like the one depicted in Figure 5. The fact that now there are new lightlike geodesics (in comparison to Figure 4) that pass through the interior region of the deflector gives rise to the formation of conjugate points, i. e., the light cone is no longer everywhere an immersed submanifold of \mathcal{M} . More precisely, the light cone develops two cuspidal edges that meet in a so-called *swallow-tail* at the point denoted by q in the figure. These two cuspidal edges (including the point q) make up the caustic of the light cone. The point q also belongs to the cut locus which looks quite similar to the one in Figure 4. A special role is played by the past-pointing lightlike geodesic λ starting from p that passes through the point q . (In Figure 4 this geodesic was blocked by the deflector, so it did not reach a point analogous to q .) q is conjugate to p and, at the same time, the past cut point of p along λ . All neighboring geodesics emanating from p pass through the cut point first (where they enter into $\mathcal{I}^-(p)$) and reach their first conjugate point afterwards (where they smoothly slip over the cuspidal edge). Again we emphasize that one spatial dimension is suppressed in Figure 5.

Example 3:

By excising a neighborhood of the point q from the spacetime in Figure 5 we are led back to a light cone of the kind considered in Figure 4. It is more interesting to leave a neighborhood of q untouched and instead to remove a worldline from spacetime that intersects the lightlike geodesic λ between p and q . (This may then be reinterpreted as the worldline of a non-transparent deflector, although some adjustments are necessary to reconcile this new interpretation with Einstein’s field equation.) The interesting new feature of this modified example is that the geodesic λ is blocked before it reaches the point q (contrary to Figure 5) but that the point q is still part of the spacetime (contrary to Figure 4). It is obvious that in this new situation

the past light cone of p united with $\{p\}$ is no longer a closed subset of \mathcal{M} since q is in the closure of the light cone but does not belong to it. By the same token, the past lightlike conjugate locus and the past lightlike cut locus of p are no longer closed subsets of \mathcal{M} . In Section 4 we shall see that this is possible only in spacetimes that are not globally hyperbolic.

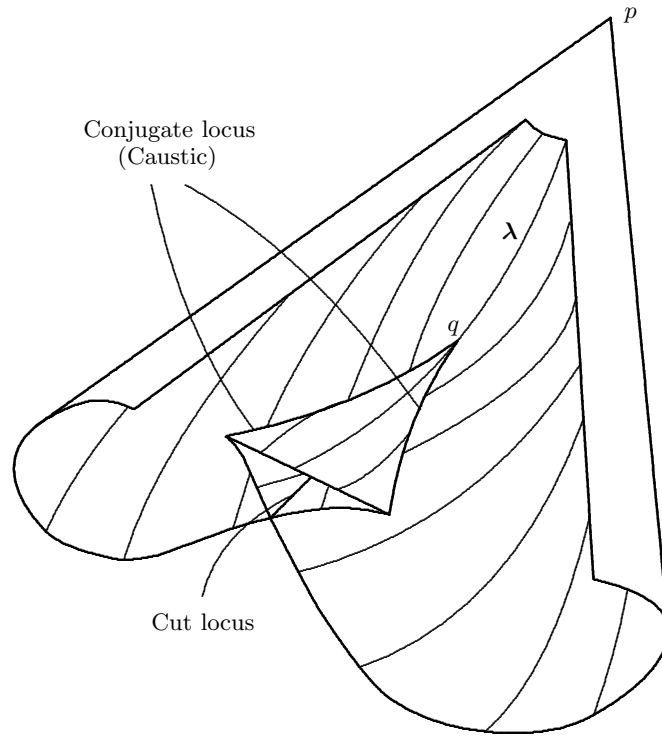


Fig. 5. In a spacetime with a transparent deflector, e. g., a transparent string, light cones typically fail to be immersed submanifolds. It is geometrically evident that the situation depicted here gives rise to triple imaging for an observer at p , cf. Figure 1.

These three examples are quite instructive but they are, of course, not universal in view of gravitational lensing situations. Completely new features may occur if the missing spatial dimension is taken into account, in particular in spacetimes without symmetry. Also, causality violations or a non-trivial topological structure of spacetime (apart from a “hole” that is meant to model a non-transparent deflector) may change the global features of light cones dramatically.

3.3 Criteria for Multiple Imaging

We are now ready to turn to the discussion of multiple imaging situations. To find out how many images an observer at p would see of a light source with worldline γ we have to determine the intersection of γ with the past light cone of p . The following proposition shows that in past-distinguishing spacetimes the occurrence of cut points is necessary for multiple imaging.

Proposition 6. *Assume that the spacetime $(\mathcal{M}, g, \mathcal{I}^+)$ is past-distinguishing at a point p and that the past lightlike cut locus of p is empty. Then for every timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$ there is at most one past-pointing lightlike geodesic that starts at p and terminates on γ .*

Proof. By contradiction, assume that we have two past-pointing lightlike geodesics λ_1 and λ_2 from p to γ . If they meet γ at the same point, Proposition 3 shows that the past lightlike cut locus of p cannot be empty. So let us assume that the two geodesics meet γ in two different points $\lambda_1(s_1) = q_1 \neq \lambda_2(s_2) = q_2$, with q_1 in the past of q_2 (say). Then we can join the section of γ between q_1 and q_2 to λ_2 to get a past-pointing causal curve from p to q_1 that is not a lightlike geodesic. By a well-known theorem (see Hawking and Ellis [29], Proposition 4.5.10) this implies that q_1 is in $\mathcal{I}^-(p)$. Together with the past-distinguishing assumption this makes sure that the past cut point of p along λ_1 must exist which gives the desired contradiction. \square

A slightly weaker version of this result, assuming the strong causality condition rather than the past-distinguishing condition, was given in Perlick [58].

Next we give a sufficient criterion for multiple imaging. The examples studied at the end of the preceding subsection suggest that a past light cone forms several sheets after past-pointing light rays have passed through cut points or conjugate points, and that this gives rise to multiple imaging situations. The following proposition puts this general idea into precise form.

Proposition 7. *Fix, in an arbitrary spacetime $(\mathcal{M}, g, \mathcal{I}^+)$, a point p and a past-pointing lightlike geodesic $\lambda : I \rightarrow \mathcal{M}$ with $\lambda(s_o) = p$. Assume that, for some parameter $s_1 > s_o$ in I , $\lambda(s_1)$ is a conjugate point or the past cut point (or both) of p along λ . Then, for every parameter value $s \in I$ with $s > s_1$, there is a timelike curve γ through $\lambda(s)$ that can be reached from p along at least two past-pointing lightlike geodesics.*

Proof. We first show that, for $s > s_1$, the point $\lambda(s)$ can be reached from p along a past-pointing timelike curve. If $\lambda(s_1)$ is the past cut point of p along λ , this follows directly from the definition of cut points. If it is a conjugate point, it follows from Proposition 1. Now we take such a timelike curve and perturb it slightly near p . In this way we get a timelike curve that intersects the past light cone of p in $\lambda(s)$ and in another point close to p , so it can be reached from p along two different past-pointing lightlike geodesics. \square

Together with Proposition 5 this result implies that multiple imaging takes place whenever a past light cone fails to be an embedded submanifold. This is true, in particular, whenever a past light cone forms a caustic.

Proposition 7 can be used to prove that multiple imaging occurs in large classes of spacetimes. E. g., it is well known that, under conditions which are to be considered as fairly general from a physical point of view, a lightlike geodesic must be either incomplete or contain a pair of conjugate points. Those “fairly general conditions” are, e. g., the weak energy condition and the so-called generic condition. We do not want to go into this matter here. We just mention that results of this kind have played a crucial part in the development of the Penrose-Hawking singularity theorems and we refer to the detailed discussion in Hawking and Ellis [29], in particular to Proposition 4.4.5. We also mention that the weak energy condition need not hold pointwise but that some integrated version of the weak energy condition would do, see Tipler [77], Borde [6], Roman [67] and Kánnár [34]. In view of these results it seems justified to say that the occurrence of conjugate points along lightlike geodesics is the rule rather than the exception. But then Proposition 7 implies that the occurrence of multiple imaging is the rule rather than the exception.

One has to keep in mind that the worldline γ in Proposition 7 must be constructed in a particular way. There is no guarantee that the real universe for which (\mathcal{M}, g) is a mathematical model contains a real light source (i. e., a galaxy or a quasar) that travels on this worldline γ . Therefore it would be nice to have an analogous proposition in which both the point p and the worldline γ are to be prescribed. Such a proposition holds in globally hyperbolic spacetimes and will be proven in Section 4, see Proposition 17 below.

Proposition 7, which is a fairly simple corollary of standard theorems, was given in Perlick [58]. Already earlier, Padmanabhan and Subramanian [51] had shown that the existence of conjugate points along a lightlike geodesic is sufficient for multiple imaging. However, their proof is completely different from ours and it uses a lot of additional assumptions on the topological and causal structure of spacetime most of which slip in surreptitiously. On the basis of these additional assumptions, Padmanabhan and Subramanian [51] were also able to show that the existence of conjugate points along lightlike geodesics is necessary for multiple imaging. We have already emphasized that this is not true in arbitrary spacetimes, please recall Example 1 at the end of the preceding subsection. We are now going to investigate topological and causal conditions on the spacetime that allow to prove such a result.

Example 1 might suggest that in simply connected spacetimes multiple imaging situations without conjugate points cannot occur. A more careful analysis shows that it is not the topology of (4-dimensional) spacetime but rather the topology of (3-dimensional) space that matters. To make this notion precise we have to consider a timelike C^∞ vector field V (*observer field*) on \mathcal{M} . The existence of such an observer field is guaranteed on all of \mathcal{M} owing to the time-orientability assumption (c) of Definition 1, cf., e. g., Wald

[80], Lemma 8.1.1. With such a V chosen, we may call any two points of \mathcal{M} *equivalent* if they lie on a common integral curve of V . The corresponding quotient space, equipped with the quotient topology, will be denoted by \mathcal{S}_V and can be interpreted as the *space* with respect to the observer field V . Please note that \mathcal{S}_V need not satisfy the Hausdorff axiom; as a counter-example one may consider any timelike vector field on Minkowski space with one point removed. Also, there is no guarantee that \mathcal{S}_V admits a smooth manifold structure such that the natural projection $\pi_V : \mathcal{M} \rightarrow \mathcal{S}_V$ becomes a submersion; as a counter-example one may consider a timelike vector field V with an integral curve that is almost periodic. For later purpose we state the following result.

Proposition 8. *Let $(\mathcal{M}, g, \mathcal{T}^+)$ be a spacetime that does not contain a closed timelike curve and let V be a timelike C^∞ vector field on \mathcal{M} . If the quotient space \mathcal{S}_V satisfies the Hausdorff axiom, \mathcal{S}_V admits a C^∞ structure such that the natural projection $\pi_V : \mathcal{M} \rightarrow \mathcal{S}_V$ makes \mathcal{M} into a fiber bundle over \mathcal{S}_V with typical fiber diffeomorphic to \mathbb{R} .*

For a proof we refer to Harris [27], Theorem 2. Note that Harris' assumption of V being complete is unnecessary since every nowhere vanishing vector field on \mathcal{M} can be made into a complete vector field by multiplication with an appropriate positive function. To prove this, one puts a complete Riemannian metric h on \mathcal{M} . This is possible since, by a famous theorem of Whitney [86] (see also, e. g., Hirsch [31], p. 55) every n -dimensional paracompact manifold can be smoothly embedded as a closed submanifold into \mathbb{R}^{2n+1} ; pulling back the Euclidean metric gives the desired complete Riemannian metric. It is then easy to check that the vector field $h(V, V)^{-1/2} V$ is complete, cf., e. g., Abraham and Marsden [1], Proposition 2.1.21.

For the following consideration we only need the topological structure on \mathcal{S}_V . We define for any point $p \in \mathcal{M}$ the set \mathcal{S}_V^p , called the *space visible to p* with respect to V , in the following way. We say that a point in \mathcal{S}_V is in \mathcal{S}_V^p if and only if the integral curve of V which is represented by that point either passes through p or can be reached from p along a past-pointing lightlike geodesic in \mathcal{M} . We are now ready to formulate the desired proposition.

Proposition 9. *Choose a timelike C^∞ vector field V on an arbitrary spacetime $(\mathcal{M}, g, \mathcal{T}^+)$. Fix a point $p \in \mathcal{M}$ and assume that \mathcal{S}_V^p , the space visible to p with respect to V , is simply connected. If the past lightlike conjugate locus of p is empty, any integral curve of V can be reached from p along at most one past-pointing lightlike geodesic.*

Proof. As the past lightlike conjugate locus of p is empty, the map e_p^- of (10) is an immersion. Then its image, the past light cone of p , is an immersed lightlike submanifold of \mathcal{M} . Since a timelike vector cannot be tangent to a lightlike submanifold, each integral curve of V intersects the image of e_p^- transversely. Hence, the combination of e_p^- with the projection $\pi_V : \mathcal{M} \rightarrow \mathcal{S}_V$ gives

a homeomorphism locally around each point, i. e., it gives a covering map from \mathcal{C}_p^- onto \mathcal{S}_V^p . As a covering map onto a simply connected space must be a (global) homeomorphism, no integral curve of V can intersect the past light cone of p more than once and the past light cone of p cannot have self-intersections or almost self-intersections. The latter implies that e_p^- is even an embedding, i. e., it is impossible that a point can be reached from p along two different past-pointing lightlike geodesics. \square

The condition of \mathcal{S}_V^p being simply connected prohibits, in particular, situations such as in Example 1 where a non-transparent deflector is modeled by a hole. However, it may also be violated in situations with transparent deflectors, viz., if the visible universe has a non-trivial spatial topology.

For some situations of interest, at least, Proposition 9 says that multiple imaging requires the occurrence of conjugate points. This is a valuable result since the existence of conjugate points along a lightlike geodesic allows to estimate the Ricci tensor along that geodesic. If we take Einstein's field equation into account, this estimate of the Ricci tensor can be rewritten as an estimate on the energy density, see Padmanabhan and Subramanian [51]. It is this observation that makes it physically interesting to investigate whether in a multiple imaging situation conjugate points must occur.

We summarize the results found so far in the following way. The occurrence of conjugate points or of cut points along a past-pointing lightlike geodesic is always sufficient for multiple imaging. If the past-distinguishing condition is satisfied at the observer's position, the occurrence of cut points is necessary as well. If the space visible to p (with respect to an observer field) has a simply connected topology, then the occurrence of conjugate points is also necessary.

We end this subsection with a group of propositions characterizing the special situation that a worldline γ meets the caustic of the past light cone of p . We first show that this is an exceptional situation.

Proposition 10. *Let $\gamma : I \rightarrow \mathcal{M}$ be a timelike C^∞ curve in an arbitrary spacetime $(\mathcal{M}, g, \mathcal{T}^+)$. Then the set of all points $p \in \mathcal{M}$ such that γ does not meet the caustic of the past light cone of p is dense in \mathcal{M} .*

Proof. For each p in some open subset $\mathcal{U} \subseteq \mathcal{M}$, we consider the map e_p^- of (10) and identify its domain \mathcal{C}_p^- with $\mathbb{R}^3 \setminus \{0\}$. This can be done with the help of local coordinates in the tangent bundle. Then the assignment $p \mapsto e_p^-$ gives a continuous embedding from \mathcal{U} into the space $C^1(\mathbb{R}^3 \setminus \{0\}, \mathcal{M})$ of C^1 maps from $\mathbb{R}^3 \setminus \{0\}$ into \mathcal{M} , equipped with the weak (or compact-open) topology. For the definition of this topology we refer to Hirsch [32], p. 34. By the transversality theorem (see, e. g., Hirsch [32], Theorem 2.1), the maps which are transverse to γ form a dense subset of $C^1(\mathbb{R}^3 \setminus \{0\}, \mathcal{M})$. Thus, the points p for which e_p^- is transverse to γ are dense in \mathcal{U} . Please recall that, by definition, e_p^- is transverse to γ at a point $q = e_p^-(X)$ if either $q \notin \gamma$ or the image of $T_X e_p^-$ and the tangent space of γ span all of $T_q \mathcal{M}$. Clearly, if γ

meets the caustic of the past light cone of p at $e_p^-(X)$, transversality cannot be satisfied since the image of $T_X e_p^-$ is at most two-dimensional. \square

In other words, by a “small perturbation” of the point p we can always achieve that a given worldline γ stays away from the caustic. But then multiple imaging situations are restricted by the following result.

Proposition 11. *If, in an arbitrary spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, a timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$ does not meet the caustic of the past light cone of a point p , then there are at most denumerably many past-pointing lightlike geodesics that start at p and terminate on γ .*

Proof. Consider the pre-image $\mathcal{A}_{p,\gamma} = (e_p^-)^{-1}(\gamma)$ of γ under the map (10). If γ does not meet the caustic, e_p^- is an immersion at each point $X \in \mathcal{A}_{p,\gamma}$, i. e., it maps a neighborhood of X in C_p^- onto a 3-dimensional submanifold which is lightlike and, thus, transverse to γ . This proves that the points in $\mathcal{A}_{p,\gamma}$ are isolated, i. e., that there are only finitely many in each compact subset of C_p^- . Since $C_p^- \simeq \mathbb{R}^3 \setminus \{0\}$ can be covered with denumerably many compact sets this completes the proof. \square

These two propositions justify our earlier claim that for multiple imaging situations Case C is generic and Case D is exceptional. However, again we emphasize that these results crucially depend on our idealization of assuming a pointlike source. They are, of course, no longer true if the worldline γ is replaced with a worldsheet or a worldtube.

Proposition 11 implies that a Case D situation is possible only if γ meets the caustic of the past light cone of p . In other words, if this light cone does not develop a caustic, then it is impossible for the observer at p to see an extended image such as an arc or a ring. We end this section by proving our earlier claim that, in a Case D situation, all parts of an extended (and connected) image show the light source at the same age.

Proposition 12. *Let p be a point and $\gamma : I \rightarrow \mathcal{M}$ a timelike C^∞ curve in a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$. Let $\mathcal{A}_{p,\gamma}$ denote the pre-image of γ under the map e_p^- which was introduced in (10). Then each connected component of $\mathcal{A}_{p,\gamma}$ is mapped by e_p^- onto a single point.*

Proof. Assume the image is not a single point. Then, by continuity of the exponential map, it is a one-dimensional timelike submanifold, viz., a portion of γ . This contradicts the fact that the image of $T_X e_p^-$ is always lightlike, so the light cone cannot contain a timelike curve. \square

3.4 Fermat’s Principle

For many applications it is useful to characterize the lightlike geodesics between a point and a timelike curve in a spacetime as the solutions of a variational problem. There are several versions of such a variational problem which

may be viewed as general-relativistic generalizations of the traditional *Fermat principle*. The oldest versions, which hold on static or stationary spacetimes only, date back to Weyl [84] and Levi-Civita [41]. They are also discussed in several modern text-books and review articles, see, e. g., Frankel [17] or Straumann [76] for the static case and Landau and Lifschitz [40] or Brill [8] for the stationary case. For a discussion from a mathematical point of view we refer to Masiello [44]. Here we want to present a more general version of Fermat's principle which holds on *arbitrary* spacetimes. Its formulation is due to Kovner [36] and the proof that the solution curves of this variational problem are, indeed, the lightlike geodesics was given by Perlick [55]. The same version of Fermat's principle is also discussed in Schneider, Ehlers and Falco [70]. As an aside, we mention that this version of Fermat's principle may be generalized to the case of light rays in media, see Perlick [59] for a detailed exposition, and to the case of extended (i. e., non-pointlike) observers and light sources, see Perlick and Piccione [60]. According to the framework of this article we shall not discuss these generalizations here.

It is our goal to characterize, in an arbitrary spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, the lightlike geodesics from a point p to a timelike worldline γ by a variational principle. To that end we have to specify (i) the set of *trial curves*, i. e., the set of curves among which the solutions to the variational problem are to be sought, and (ii) the functional that is to be extremized. The set of trial curves, denoted by $C_{p,\gamma}^\infty$ henceforth, is defined in the following way.

Definition 4. Fix, in an arbitrary spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, a point $p \in \mathcal{M}$ and an embedded timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$. Then $C_{p,\gamma}^\infty$ is, by definition, the set of all C^∞ immersions $\lambda : [0, 1] \rightarrow \mathcal{M}$ with the following properties.

- (a) λ is lightlike, i. e., $g(\lambda'(s), \lambda'(s)) = 0$ for all $s \in [0, 1]$.
- (b) λ starts at p and terminates on γ , i. e., $\lambda(0) = p$ and there is a $\tau(\lambda) \in I$ such that $\lambda(1) = \gamma(\tau(\lambda))$. Since γ is assumed to be an embedding, this defines a unique assignment $\lambda \mapsto \tau(\lambda)$.
- (c) $g(\lambda'(1), \gamma'(\tau(\lambda))) < 0$, where $\tau(\lambda) \in I$ is defined through (b).

Roughly speaking, the space of trial curves can be characterized as the set of all ways to go from p to γ at the speed of light. Our decision to define all trial curves on the interval $[0, 1]$ is a matter of convenience only. Condition (c) of Definition 4 restricts to future-oriented or past-oriented curves, depending on whether γ is future-oriented or past-oriented. For applications to gravitational lensing we are interested in the case that γ is a *past-pointing* parametrization of the worldline of a light source, see Figure 1.

Condition (b) of Definition 4 defines a map $\tau : C_{p,\gamma}^\infty \rightarrow \mathbb{R}$. We refer to τ as to the *arrival time functional* henceforth. This will be the functional to be extremized. – Finally, we need the following definition.

Definition 5. For a curve $\lambda \in C_{p,\gamma}^\infty$, as defined in Definition 4, a C^∞ *variation* of λ in $C_{p,\gamma}^\infty$ is a C^∞ map $\eta :]-\varepsilon_o, \varepsilon_o[\times [0, 1] \rightarrow \mathcal{M}$, for some $\varepsilon_o > 0$, such that $\eta(0, \cdot) = \lambda$ and $\eta(\varepsilon, \cdot) \in C_{p,\gamma}^\infty$ for all $\varepsilon \in]-\varepsilon_o, \varepsilon_o[$. The vector

field $X : [0, 1] \rightarrow T\mathcal{M}$, $s \mapsto X(s) = (\eta(\cdot, s))'(0)$ is called the *variational vector field* of η . X is called *non-trivial* if $X(s)$ and $\lambda'(s)$ are non-collinear for some $s \in [0, 1]$.

We want to prove that among all trial curves the lightlike geodesics are the stationary points of the arrival time. To that end we shall need the following characterization of variational vector fields.

Lemma 1. *For a C^∞ vector field $X : [0, 1] \rightarrow T\mathcal{M}$ along $\lambda \in C_{p,\gamma}^\infty$, the following two properties are equivalent.*

- (a) X is the variational vector field of a C^∞ variation η of λ in $C_{p,\gamma}^\infty$.
- (b) $g(\nabla_{\lambda'} X, \lambda') = 0$, $X(0) = 0$, and $X(1) \parallel \gamma'(\tau(\lambda))$.

The implication “(a) \Rightarrow (b)” is obvious since the desired properties of X follow just by differentiating the defining properties of trial curves. For a proof of the converse implication, which is more cumbersome, the reader is referred to Perlick [55].

We are now ready to formulate and prove the general-relativistic Fermat principle.

Theorem 1. (Fermat’s principle) *Let $(\mathcal{M}, g, \mathcal{T}^+)$ be an arbitrary spacetime, fix a point $p \in \mathcal{M}$ and an embedded timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$. Then for a trial curve $\lambda \in C_{p,\gamma}^\infty$ the following two properties are equivalent.*

- (a) λ is a geodesic or a reparametrization thereof.
- (b) For all C^∞ variations η of λ in $C_{p,\gamma}^\infty$, the equation $\left. \frac{d}{d\varepsilon} \tau(\eta(\varepsilon, \cdot)) \right|_{\varepsilon=0} = 0$ holds true.

Proof. We first observe that the definition of the arrival time τ implies the equation $\eta(\varepsilon, 1) = \gamma(\tau(\eta(\varepsilon, \cdot)))$ for each C^∞ variation η of λ in $C_{p,\gamma}^\infty$. Differentiating with respect to ε and setting $\varepsilon = 0$ yields

$$X(1) = \gamma'(\tau(\lambda)) \left. \frac{d}{d\varepsilon} \tau(\eta(\varepsilon, \cdot)) \right|_{\varepsilon=0}. \quad (12)$$

We now prove the implication “(a) \Rightarrow (b)”. By assumption, there is a C^∞ function $w : [0, 1] \rightarrow \mathbb{R}$ such that $\nabla_{\lambda'} \lambda' = w \lambda'$. Now let X be the variational vector field of a C^∞ variation η of λ in $C_{p,\gamma}^\infty$. Then we find $w g(X, \lambda') = g(X, \nabla_{\lambda'} \lambda') = g(X, \lambda')' - g(\nabla_{\lambda'} X, \lambda')$. The last term vanishes by Lemma 1. Upon integration, we find

$$g(X(1), \lambda'(1)) = g(X(0), \lambda'(0)) \exp\left(\int_0^1 w(s) ds\right). \quad (13)$$

Since $X(0) = 0$, this implies $g(X(1), \lambda'(1)) = 0$. But then the desired result can be read from (12) because the timelike tangent vector to γ cannot be orthogonal to the lightlike tangent vector to λ . – To prove the converse implication “(b) \Rightarrow (a)” we define a vector field $U_\lambda : [0, 1] \rightarrow T\mathcal{M}$ by parallel transporting the vector $\gamma'(\tau(\lambda))$ along λ , i. e., $\nabla_{\lambda'} U_\lambda = 0$ and

$U_\lambda(1) = \gamma'(\tau(\lambda))$. Clearly, U_λ is everywhere timelike, so $g(\lambda', U_\lambda)$ has no zeros. Now let $Z : [0, 1] \rightarrow T\mathcal{M}$ be any C^∞ vector field along λ with $Z(0) = 0$ and $Z(1) = 0$. We use this Z to define a new vector field X along λ by

$$X(s) = Z(s) - \left(\int_0^s \frac{g(\nabla_{\lambda'} Z, \lambda')}{g(U_\lambda, \lambda')} \Big|_{\bar{s}} d\bar{s} \right) U_\lambda(s). \quad (14)$$

With the help of Lemma 1 it is easy to verify that X is the variational vector field of a C^∞ variation η of λ in $C_{p,\gamma}^\infty$. Hence, we can read from (12) that our hypothesis implies $X(1) = 0$. As $Z(1) = 0$ and $g(U_\lambda(1), \lambda'(1)) \neq 0$, the integral in (14) must vanish for $s = 1$. Upon integration by parts, this results in

$$\int_0^1 g \left(Z, \nabla_{\lambda'} \frac{\lambda'}{g(U_\lambda, \lambda')} \right) \Big|_s ds = 0. \quad (15)$$

Since Z was an arbitrary C^∞ vector field along λ with $Z(0) = 0$ and $Z(1) = 0$, the fundamental lemma of variational calculus implies that

$$\nabla_{\lambda'} \frac{\lambda'}{g(U_\lambda, \lambda')} = 0, \quad (16)$$

i. e., that $\nabla_{\lambda'} \lambda'$ is a multiple of λ' . \square

Theorem 1 can be phrased as saying that among all ways to go from p to γ at the speed of light, the light rays are characterized as the stationary points of the arrival time τ . The analogy to the traditional Fermat principle is obvious. For some applications it might be convenient to choose γ as parametrized by proper time, $g(\gamma', \gamma') = -1$. However, Theorem 1 is true for any other smooth parametrization as well. The arrival time functional $\tau : C_{p,\gamma}^\infty \rightarrow \mathbb{R}$ changes, of course, if the parametrization is changed, but the new arrival time functional has the same stationary points as the old one.

Please note that, by Theorem 1, a light ray may be a local minimum, a local maximum or a saddle of τ . We shall see in the next subsection that, actually, local maxima do not occur.

From Theorem 1 we can easily rederive the more special versions of Fermat's principle which are given in many text-books on general relativity. To illustrate this claim we consider the special case of a *conformally static spacetime*. More precisely, we need the following assumptions.

- (a) \mathcal{M} is diffeomorphic to $\mathcal{S} \times \mathbb{R}$, with some 3-dimensional manifold \mathcal{S} .
- (b) The metric takes the form

$$g = e^{2f(x,t)} \left(h_{\mu\nu}(x) dx^\mu \otimes dx^\nu - dt \otimes dt \right) \quad (17)$$

where t denotes the projection from $\mathcal{M} \simeq \mathcal{S} \times \mathbb{R}$ onto the second factor, $x = (x^1, x^2, x^3)$ are coordinates on \mathcal{S} and the Einstein summation convention is used for greek indices running from 1 to 3. (Coordinates on \mathcal{S} are used for

notational convenience only. It will not be necessary to assume that \mathcal{S} can be covered by a single coordinate system.)

(c) $t \circ \gamma$ is constant, i. e., γ is vertical with respect to the product structure of $\mathcal{M} \simeq \mathcal{S} \times \mathbb{R}$.

In this situation the spacetime geometry is time independent up to an overall factor $e^{2f(x,t)}$ and the worldline γ is at rest in the “space” \mathcal{S} . Fermat’s principle takes its simplest form if we use for γ a parametrization adapted to t , i. e., $dt(\gamma') = 1$. Since all trial curves are lightlike, we can then read from (17) that the arrival time functional is given by

$$\tau(\lambda) = \int_0^1 \sqrt{h_{\mu\nu}(x(s)) \frac{dx^\mu(s)}{ds} \frac{dx^\nu(s)}{ds}} ds. \quad (18)$$

Here $s \mapsto x(s)$ denotes the projection onto the first factor of $s \mapsto \lambda(s) \in \mathcal{M} \simeq \mathcal{S} \times \mathbb{R}$. By (18), $\tau(\lambda)$ is exactly the *length* of the projected curve $s \mapsto x(s)$, measured with the spatial metric $h = h_{\mu\nu}(x) dx^\mu \otimes dx^\nu$. Hence, in this special case Theorem 1 says that a trial curve is a light ray if and only if its projection to \mathcal{S} traces out an h -geodesic. This result is due to Weyl [84], apart from the fact that Weyl restricted to the static case, i. e., he did not allow the function f in (17) to depend on t . There is a straightforward generalization from the (conformally) static to the (conformally) stationary case which is essentially due to Levi-Civita [41]. The solution curves are then no longer h -geodesics but modified by a kind of Coriolis force. For a detailed discussion, including several examples, the reader is referred to Perlick [56].

3.5 Morse Index Theory for Fermat’s Principle

Fermat’s principle admits several interesting applications to gravitational lensing. E. g., Schneider [69] has shown that Fermat’s principle can be used in the derivation of the so-called lens equation of the quasi-Newtonian approximation formalism, see also Schneider, Ehlers and Falco [70]. Again in the quasi-Newtonian approximation, Blandford and Narayan [5] have used Fermat’s principle to give a topological classification of images. Owing to the approximation assumptions, in these situations it suffices to consider Fermat’s principle on conformally static spacetimes. An application to gravitational lensing of Theorem 1 where the conformally static or conformally stationary version would not do was worked out by Kovner [36]. He considered a gravitational wave sweeping over a gravitational lensing situation and calculated, to within certain approximations, the influence of the wave on the arrival times and on the positions of the images at the observer’s sky. This line of thought was further developed by Faraoni [14] who assumed the spacetime to be a first order (but non-stationary) perturbation of Minkowski space and used a coordinate version of Fermat’s principle, given in Perlick [55], to calculate integral formulae for the arrival time and for the deflection angle.

Here we want to discuss a different application of Fermat's principle to gravitational lensing. The basic idea is to formulate a Morse theory for Fermat's principle, in analogy to the classical Morse theory of Riemannian geometry, and to investigate the significance of the Morse relations in view of gravitational lensing. As a first step towards this goal, we establish a Morse index theorem for Fermat's principle, thereby investigating whether a solution curve yields a local minimum, a local maximum or a saddle of the arrival time functional. A full Morse theory has to presuppose a globally hyperbolic spacetime and will be the subject of Subsection 4.2 below.

As a preparation, it is certainly useful to recall the classical Morse index theory of Riemannian geometry which was developed by Morse [48] in the 1930s. Let p and q be two points in a Riemannian manifold (\mathcal{N}, h) , i. e., in a manifold with a positive definite metric. Then, among all sufficiently regular curves $\alpha : [0, 1] \rightarrow \mathcal{N}$ with $\alpha(0) = p$ and $\alpha(1) = q$, the geodesics are characterized as the stationary points of the *energy functional* $E(\alpha) = \int_0^1 h(\alpha'(s), \alpha'(s)) ds$. To find out whether a geodesic α is a local minimum, a local maximum or a saddle of E one has to calculate the Hessian $\text{Hess}_\alpha(E)$ of E at the point α (i. e., the "second variation") and to determine the index and the extended index of $\text{Hess}_\alpha(E)$. Please recall that the *index* (or the *extended index*, respectively) of a bilinear form is the maximal dimension of a subspace on which this bilinear form is negative definite (or negative semi-definite, respectively). The classical Morse index theorem says that $\text{Hess}_\alpha(E)$ is non-degenerate if and only if the endpoint $\alpha(1)$ is not conjugate to the initial point $\alpha(0)$ along the geodesic α and that the extended index of $\text{Hess}_\alpha(E)$ is equal to the number of points $\alpha(s)$, $s \in]0, 1]$, that are conjugate to $\alpha(0)$ along α . Here each conjugate point is to be counted with its multiplicity. Based on the Morse index theorem, Morse was able to establish a number of theorems, now summarized under the name of Morse theory, to the effect that the number of geodesics joining two points p and q in a *complete* Riemannian manifold is related to the topology of the space of sufficiently regular curves joining these two points. Morse proved these results by considering the energy functional on the finite-dimensional space of broken geodesics with N breakpoints between p_1 and p_2 , and then letting $N \rightarrow \infty$. For a detailed review of Morse's work we refer to Milnor [46]. Later Palais and Smale [52] [53] brought forward a fresh approach to Morse theory by considering functionals on infinite-dimensional Hilbert manifolds. It was then no longer necessary to approximate the space of trial curves by N -dimensional spaces and to consider the limit $N \rightarrow \infty$ afterwards. It is this Palais-Smale version of Morse theory we want to apply to Fermat's principle.

It should be mentioned that a Morse index theory for the geodesic variational problem (i. e., extremizing the energy functional between two points) exists not only for geodesics in Riemannian manifolds but also for timelike and lightlike geodesics in Lorentzian manifolds, see Beem, Ehrlich and Easley [4] for a detailed exposition. However, this is not what we are interested in.

We want to characterize lightlike geodesics between a point and a timelike curve (not between two points), and the functional we are going to extremize is the arrival time (not the Lorentzian analogue of the energy functional).

The following exposition closely follows Perlick [57]. It is our goal to establish a Morse index theorem for Fermat's principle in an infinite-dimensional Hilbert manifold setting à la Palais-Smale. To that end we have to modify Fermat's principle, as it was given in Theorem 1, a little bit. First, we observe that, according to Definition 4, all trial curves $\lambda \in C_{p,\gamma}^\infty$ are of class C^∞ . It is well known that C^∞ maps from one manifold into another do not form a Hilbert manifold but, at the very best, a Fréchet manifold. Since this is too weak for applying Morse theory, we shall replace the C^∞ condition on the trial curves by a Sobolev H^r condition in order to get a Hilbert manifold. Second, we observe that the arrival time functional τ is invariant under reparametrization. As a consequence, its Hessian at a solution curve λ is always degenerate because it vanishes on the infinite dimensional vector space of trivial variational vector fields (please recall Definition 5). We shall solve this problem by imposing a parametrization fixing condition upon the trial curves.

To work this out we have to introduce the Hilbert manifold of Sobolev H^r curves. For background material on this subject the reader is referred to Schwartz [71]. The same book also contains a review of the Palais-Smale version of Morse theory. For $f_1, f_2 \in C^\infty([0, 1], \mathbb{R}^n)$, we define

$$\langle f_1 | f_2 \rangle_r = \sum_{i=0}^r \int_0^1 f_1^{(i)}(s) \cdot f_2^{(i)}(s) ds \quad (19)$$

where $f_1^{(i)}$ denotes the i -th derivative of f_1 and the dot denotes the standard scalar product in \mathbb{R}^n . It is easy to check that this scalar product makes $C^\infty([0, 1], \mathbb{R}^n)$ into a real pre-Hilbert space. The completion of this pre-Hilbert space is, by definition, the *Sobolev space* $H^r([0, 1], \mathbb{R}^n)$. For $r = 0$ this gives the real Lebesgue space $L^2([0, 1], \mathbb{R}^n)$ whose complex version is known to every physicist from quantum mechanics. For integers $r \geq 1$, $H^r([0, 1], \mathbb{R}^n)$ can be (and will be henceforth) identified with the space of all C^{r-1} maps from $[0, 1]$ into \mathbb{R}^n whose r -th derivatives exist almost everywhere and are locally square integrable.

Now we introduce the notion of H^r curves in a manifold. Let \mathcal{M} be a real, finite-dimensional C^∞ manifold whose topology satisfies the Hausdorff axiom and the second countability axiom. Then we define

$$H^r([0, 1], \mathcal{M}) = \left\{ \lambda : [0, 1] \longrightarrow \mathcal{M} \mid j \circ \lambda \in H^r([0, 1], \mathbb{R}^n) \right\} \quad (20)$$

where $j : \mathcal{M} \longrightarrow \mathbb{R}^n$ is a C^∞ embedding. A well-known theorem of Whitney [86] (see also, e. g., Hirsch [31], p. 55) guarantees the existence of such an embedding for $n \geq 2 \dim(\mathcal{M}) + 1$. It is easy to show that the set $H^r([0, 1], \mathcal{M})$ is

independent of which j has been chosen. Moreover, it is a well-known result of Palais and Smale [53] that the inclusion map $H^r([0, 1], \mathcal{M}) \rightarrow H^r([0, 1], \mathbb{R}^n)$ induced by j makes $H^r([0, 1], \mathcal{M})$ into a C^∞ submanifold of the Hilbert space $H^r([0, 1], \mathbb{R}^n)$ and that the manifold structure thereby established on $H^r([0, 1], \mathcal{M})$ is, again, independent of j . Thus, we may view $H^r([0, 1], \mathcal{M})$ as an infinite dimensional real C^∞ Hilbert manifold in its own right.

To define the modified space of trial curves we restrict the original space $C_{p,\gamma}^\infty$ of Definition 4 by a parametrization fixing condition. For $\lambda \in C_{p,\gamma}^\infty$ we define a vector field $U_\lambda : [0, 1] \rightarrow T\mathcal{M}$ by parallel transporting the vector $\gamma'(\tau(\lambda))$ along λ , as in the proof of Theorem 1. Then the condition $g(U_\lambda, \lambda') = \text{const.}$ singles out exactly one parametrization for each trial curve. Please note that this condition singles out an affine parametrization along each geodesic. Now we define the modified space of trial curves in the following way.

Definition 6. Fix a point p and a timelike embedded C^∞ curve $\gamma : I \rightarrow \mathcal{M}$. Then the space $H_{p,\gamma}^2$ is, by definition, the set of all $\lambda \in H^2([0, 1], \mathcal{M})$ with the following properties.

- (a) $g(\lambda', \lambda') = 0$.
- (b) $\lambda(0) = p$ and there is a $\tau(\lambda) \in I$ such that $\lambda(1) = \gamma(\tau(\lambda))$.
- (c) $g(U_\lambda, \lambda') = \text{const.} < 0$.

It can be shown that $H_{p,\gamma}^2$ is, indeed, an infinite dimensional C^∞ Hilbert submanifold of $H^2([0, 1], \mathcal{M})$ and that the arrival time functional $\tau : H_{p,\gamma}^2 \rightarrow \mathbb{R}$ defined by (b) is a C^∞ map. For a proof of these facts we refer to Perlick [57]. This result remains true if the H^2 condition in Definition 6 is replaced with an H^r condition for $r > 2$; it is not true, however, for $r = 1$. Now Fermat's principle, i. e., Theorem 1, can be reformulated in the following way.

Theorem 2. *A curve $\lambda \in H_{p,\gamma}^2$ is a geodesic if and only if the differential of the arrival time functional $\tau : H_{p,\gamma}^2 \rightarrow \mathbb{R}$ has a zero at λ .*

The proof, which is worked out in Perlick [57], is a straightforward translation of the proof of Theorem 1 into an H^2 setting.

For the Morse index theorem we have to calculate the Hessian $\text{Hess}_\lambda(\tau)$ of τ at a geodesic λ . We find the following result.

Theorem 3. (Morse index theorem) *Let $\lambda \in H_{p,\gamma}^2$ be a geodesic. The index of $\text{Hess}_\lambda(\tau)$ is equal to the number of points $\lambda(s)$, $s \in]0, 1[$, that are conjugate to $\lambda(0)$ along λ . The extended index of $\text{Hess}_\lambda(\tau)$ is equal to the number of points $\lambda(s)$, $s \in]0, 1]$, that are conjugate to $\lambda(0)$ along λ . In both cases conjugate points are to be counted with their multiplicities.*

The proof of this theorem is given in Perlick [57]. The strategy of this proof is to relate the second variational formula for Fermat's principle to the second variational formula for the geodesic variational problem. For lightlike

geodesics, the latter is worked out in Beem, Ehrlich and Easley [4], Chapter 10. We have already emphasized the differences between these two variational problems. Nonetheless, their second variational formulae turn out to be essentially the same. Thereby, the proof of Theorem 3 comes as a corollary of the Morse index theorem proven in Beem, Ehrlich and Easley.

Theorem 3 has the following immediate consequences.

(a) $\text{Hess}_\lambda(\tau)$ is non-degenerate if and only if $\lambda(1)$ is not conjugate to $\lambda(0)$ along λ .

(b) The index and the extended index of $\text{Hess}_\lambda(\tau)$ are finite for all geodesics $\lambda \in H_{p,\gamma}^2$. Hence, λ cannot be a local maximum of τ .

(c) A geodesic $\lambda \in H_{p,\gamma}^2$ is a strict local minimum of τ if and only if λ does not contain a point conjugate to $\lambda(0)$

(d) A geodesic $\lambda \in H_{p,\gamma}^2$ is a saddle of τ if it contains a point $\lambda(s)$ which is conjugate to $\lambda(0)$ for some $s \in]0, 1[$.

In view of gravitational lensing, the index has the following interpretation. At each conjugate point, infinitesimally neighboring light rays “cross over” from one side of λ to the other. This is associated with a side-reversion of the image. Thus, light rays with an even index yield a mirror image of those with an odd image. This is observable if our pointlike light source (e. g., the core of a galaxy) is surrounded by some non-symmetrical structure (e. g., irregular lobes or jets).

In Subsection 4.2 below we use the Morse index theorem to develop a full Morse theory for light rays joining a point and a timelike curve in a globally hyperbolic spacetime and we discuss applications to gravitational lensing.

4 Gravitational Lensing in Globally Hyperbolic Spacetimes

We have seen in the preceding section that the geometry of multiple imaging situations is strongly influenced by the topological and causal structure of spacetime. Such global effects are usually ignored in the astronomical literature on gravitational lensing where, typically more implicitly than explicitly, the deflector is assumed to be embedded in a universe without topological or causal pathologies.

In this section we get somewhat closer to the standard astronomer’s point of view by restricting to spacetimes without causal pathologies. More precisely, we are going to consider spacetimes that are globally hyperbolic according to the following definition.

Definition 7. For a spacetime $(\mathcal{M}, g, \mathcal{T}^+)$, a subset \mathcal{S} of \mathcal{M} is called a *Cauchy surface* if each inextendible causal curve in \mathcal{M} intersects \mathcal{S} in exactly one point. A spacetime is called *globally hyperbolic* if it admits a Cauchy surface.

The name “globally hyperbolic” was introduced by Leray [39] in 1952. It refers to the fact that a global existence and uniqueness theorem for hyperbolic partial differential equations can be established only on spacetimes with this property. Actually, our Definition 7 of global hyperbolicity does not coincide with Leray’s original definition, but it is well known that the two definitions are equivalent, see, e. g., Wald [80], p. 209. Basic properties of globally hyperbolic spacetimes are also reviewed in Hawking and Ellis [29], in O’Neill [50] and in Beem, Ehrlich and Easley [4].

One can show that a Cauchy surface \mathcal{S} is a 3-dimensional topological submanifold of \mathcal{M} , see, e. g., Theorem 8.1.3 and Theorem 8.3.1 in Wald [80]. Mimicking the proof of Proposition 6.3.1 in Hawking and Ellis [29], one may even show that \mathcal{S} is a C^{1-} (i. e., Lipschitz) submanifold of \mathcal{M} . In general, a Cauchy surface will not be a C^1 submanifold of \mathcal{M} .

In Subsection 3.3 we have introduced, for each timelike C^∞ vector field V on \mathcal{M} , the quotient space $\mathcal{S}_V = \mathcal{M}/\sim$, where two points in \mathcal{M} are considered equivalent if they can be connected by an integral curve of V . We shall now use Proposition 8 to show that in a globally hyperbolic spacetime \mathcal{S}_V comes not only with a topological but even with a differentiable structure. To that end, we first observe that the existence of a Cauchy surface makes sure that there are no closed timelike curves in \mathcal{M} . Since every integral curve of V intersects a Cauchy surface \mathcal{S} exactly once, the restriction of the natural projection

$$\pi_V : \mathcal{M} \longrightarrow \mathcal{S}_V \tag{21}$$

to \mathcal{S} gives a homeomorphism from \mathcal{S} onto \mathcal{S}_V . Since \mathcal{S} , being a topological submanifold of a Hausdorff space, must satisfy the Hausdorff axiom, this proves that \mathcal{S}_V satisfies the Hausdorff axiom. By Proposition 8, there is a C^∞ structure on \mathcal{S}_V such that $\pi_V : \mathcal{M} \longrightarrow \mathcal{S}_V$ makes \mathcal{M} into a fiber bundle over \mathcal{S}_V with fiber diffeomorphic to \mathbb{R} . This argument implies that any two Cauchy surfaces in \mathcal{M} must be homeomorphic and that, for any two timelike C^∞ vector fields V and V' on \mathcal{M} the quotient manifolds \mathcal{S}_V and $\mathcal{S}_{V'}$ must be homeomorphic. According to a famous theorem of Moise [47] any 3-dimensional topological manifold admits exactly one differentiable structure. Hence, \mathcal{S}_V and $\mathcal{S}_{V'}$ must even be diffeomorphic.

Geroch [20] has established the important fact that every globally hyperbolic spacetime admits a continuous function $t : \mathcal{M} \longrightarrow \mathbb{R}$ such that the set $t^{-1}(t_o)$ is a Cauchy surface for each $t_o \in \mathbb{R}$. It is widely believed that such a *Cauchy time function* t can be chosen differentiable, employing a smoothing argument of Seifert [72]. However, the details of the proof have never been worked out completely, and several dedicated experts who tried to do so failed. In any case, every globally hyperbolic spacetime admits a continuous Cauchy time function t which can be combined with the C^∞ projection π_V determined by a timelike C^∞ vector field V to give a homeomorphism $(\pi, t) : \mathcal{M} \longrightarrow \mathcal{S}_V \times \mathbb{R}$. Hence, the topology of a globally hyperbolic spacetime is determined by the topology of any of its Cauchy surfaces. Note, however,

that $\mathcal{S}_1 \times \mathbb{R}$ may be homeomorphic to $\mathcal{S}_2 \times \mathbb{R}$ without \mathcal{S}_1 being homeomorphic to \mathcal{S}_2 . Therefore, the topology of a globally hyperbolic spacetime does *not* determine the topology of its Cauchy surfaces. E. g., Newman and Clarke [49] contrived a spacetime with topology \mathbb{R}^4 that admits a Cauchy surface which is not homeomorphic to \mathbb{R}^3 .

It is a matter of debate whether the assumption of global hyperbolicity is to be considered as “reasonable” from a physical point of view. It is certainly true that most physicists consider the validity of a global existence and uniqueness theorem for wave equations as a requirement any “reasonable” spacetime should satisfy. Moreover, only globally hyperbolic spacetimes arise from globally solving the initial value problem of Einstein’s (vacuum) field equation. From the viewpoint of global Lorentzian geometry, however, global hyperbolicity is a very strong assumption. In particular, any globally hyperbolic spacetime has to satisfy, at each point p , the strong causality condition and, thus, the distinguishing conditions and the causality condition, recall Definition 3. Also, it is easy to verify that removing a point, a worldline or a worldtube from any spacetime necessarily results in a spacetime that is *not* globally hyperbolic. This remark is important for gravitational lensing where non-transparent deflectors are modeled by excising worldlines or worldtubes from spacetime. – We summarize these observations in the following way.

In view of gravitational lensing situations, restricting to globally hyperbolic spacetimes means restricting to the case of a transparent deflector in a spacetime without causality violation whose topology is a product of space and time.

Since a Cauchy surface may have a complicated topology, the restriction to globally hyperbolic spacetimes does not exclude universes with “handles” etc.

4.1 Criteria for Multiple Imaging in Globally Hyperbolic Spacetimes

In Subsection 3.3 we have formulated several criteria for multiple imaging in arbitrary spacetimes. These results can be considerably strengthened if we restrict to globally hyperbolic spacetimes. The reason is that light cones in globally hyperbolic spacetimes cannot be too pathological. In particular, the following technically important proposition holds true.

Proposition 13. *For a point p in a globally hyperbolic spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ the past light cone of p united with $\{p\}$ gives a closed subset of \mathcal{M} .*

Proof. Let q_n be a sequence in the past light cone of p that converges towards a point $q \neq p$ in \mathcal{M} . We want to show that q is, again, in the past light cone of p . To that end we choose a Cauchy surface \mathcal{S} through q . This is possible since, by the above-mentioned result of Geroch, every globally hyperbolic spacetime can be foliated into (continuously embedded) Cauchy surfaces. Convergence

of the q_n implies that their pre-image under e_p^- , using the notation of (10), is contained in a compact subset of $T_p\mathcal{M}$. So, by passing to a subsequence we find a sequence X_n in \mathcal{C}_p^- with $e_p^-(X_n) = q_n$ that converges towards a vector X in the closure of \mathcal{C}_p^- . Since $q \neq p$, X is different from zero and, thus, lightlike. We do not know yet if X is in the domain of the exponential map. Let λ_n denote the geodesic with $\lambda_n'(0) = X_n$ and λ the geodesic with $\lambda'(0) = X$. Since \mathcal{S} is a Cauchy surface, λ_n must intersect \mathcal{S} in a point \tilde{q}_n and λ must intersect \mathcal{S} in a point \tilde{q} . Since geodesics depend continuously on their initial conditions, the \tilde{q}_n converge towards \tilde{q} . On the other hand, the q_n converge towards q . As each geodesic λ_n intersects \mathcal{S} exactly once, this is possible only if $q = \tilde{q}$ which implies that q is in the past light cone of p . \square

In addition, it can be shown that in the globally hyperbolic case the past lightlike conjugate locus and the past lightlike cut locus of p are closed subsets of \mathcal{M} , see Beem, Ehrlich and Easley [4], Propositions 9.27 and 9.29 in combination with our Proposition 13. We have already seen in Example 3 at the end of Subsection 3.2 that this is not true without the assumption of global hyperbolicity. – The following proposition says that for globally hyperbolic spacetimes the name “cut point” is, indeed, justified because such a point indicates an intersection of geodesics. It was already mentioned that the analogous statement for complete Riemannian manifolds is known as *Poincaré Theorem* and dates back to Poincaré [66] and Whitehead [85].

Proposition 14. (Poincaré Theorem for lightlike geodesics) *Let p and q be two points in a globally hyperbolic spacetime $(\mathcal{M}, g, \mathcal{T}^+)$. Assume that q is in the past lightlike cut locus but not in the past lightlike conjugate locus of p . Then there are at least two past-pointing lightlike geodesics from p to q . The past light cone of p has a transverse self-intersection at q .*

A proof can be found in Beem, Ehrlich and Easley [4], Theorem 9.15. The light cone must have a transverse self-intersection at q since otherwise the two lightlike geodesics would arrive with collinear tangent vectors at q . As the assumption of global hyperbolicity excludes the possibility of having a closed lightlike geodesic through p , this collinearity would imply that the two geodesics are the same.

If q is in the cut locus and in the conjugate locus, then it may be impossible to reach it from p along a second geodesic, even in a globally hyperbolic spacetime. This is exemplified by the point q in Figure 5. However, since a conjugate point indicates an intersection with an “infinitesimally neighboring geodesic” the name “cut point” might be viewed as justified in this case as well.

Proposition 14 has the consequence that in globally hyperbolic spacetimes Proposition 5 admits the following converse.

Proposition 15. *Fix a point p in a globally hyperbolic spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ and assume that the map e_p^- of (10) is a C^∞ embedding, i. e., that the past*

light cone of p is an embedded submanifold of \mathcal{M} . Then the past lightlike cut locus of p is empty.

Proof. By contradiction, assume that q is the past cut point of p along some lightlike geodesic. If q is also conjugate to p we are done since this proves that e_p^- is not even an immersion. If q is not conjugate to p , Proposition 14 implies that the past light cone of p has a self-intersection at q , so e_p^- cannot be an embedding. \square

Together with Propositions 5, 6 and 7 this result implies that in a globally hyperbolic spacetime there is a multiple imaging situation for an observer at p if and only if the past light cone of p fails to be an embedded submanifold of \mathcal{M} .

Propositions 14 and 15 are not true without the assumption of global hyperbolicity. To see this consider the light cone of Figure 4. Divide a spherical section of this cone which is close to p into two hemispheres and excise one of them, together with its boundary, from spacetime. Then, if the division has been chosen appropriately, the light cone becomes an embedded submanifold since one half of the light rays is cut off before the light cone forms a self-intersection. However, the cut locus remains unchanged since the remaining light rays can be reached by the same timelike curves from p as before.

Another important feature of globally hyperbolic spacetimes is the following existence result for lightlike geodesics.

Proposition 16. *Let p be a point and γ an inextendible timelike curve in a globally hyperbolic spacetime $(\mathcal{M}, g, \mathcal{I}^+)$ such that $\gamma \cap \mathcal{I}^-(p) \neq \emptyset$. Then there is a past-pointing lightlike geodesic λ from p to γ that is completely contained in the boundary of $\mathcal{I}^-(p)$. This geodesic does not pass through the past cut point of p or through a point conjugate to p before it reaches γ .*

Proof. Global hyperbolicity implies that γ cannot be completely contained in $\mathcal{I}^-(p)$ because it must reach every Cauchy surface in the future of p . Therefore our assumptions imply that γ intersects $\partial\mathcal{I}^-(p)$ in some point q . It is well known (see, e. g., Wald [80], Theorem 8.1.6) that every point in $\partial\mathcal{I}^-(p)$ can be reached from p along a past-pointing lightlike geodesic. This geodesic cannot pass through the past cut point of p before it reaches q since after passing through the cut point a lightlike geodesic stays inside the open set $\mathcal{I}^-(p)$. By Proposition 2 this implies that the geodesic cannot pass through a point conjugate to p before reaching q . \square

This proposition gives, in particular, sufficient conditions for the existence of a past-pointing lightlike geodesic from p to γ that does not contain a point conjugate to p . A similar result was proven by Uhlenbeck [78], Corollary 4.8, with the help of Morse theory. In Subsection 4.2 we shall comment on her work in more detail.

As a corollary of Proposition 16 we immediately get the following sufficiency criterion for multiple imaging. As an illustration the reader may use Figure 5 with an appropriately placed worldline γ .

Proposition 17. *Let p be a point and $\gamma : I \rightarrow \mathcal{M}$ an inextendible timelike curve in a globally hyperbolic spacetime $(\mathcal{M}, g, \mathcal{I}^+)$. Assume that there is a past-pointing lightlike geodesic λ from p to γ that passes through a point conjugate to p or through the past cut point of p (or both) before it reaches γ . Then there are at least two past-pointing lightlike geodesics from p to γ .*

Proof. The existence of the lightlike geodesic λ implies that γ intersects the closure of $\mathcal{I}^-(p)$. Since γ is inextendible, this means that it must intersect $\mathcal{I}^-(p)$. But then Proposition 16 gives a past-pointing lightlike geodesic from p to γ which is different from λ because the latter contains a conjugate point or the past cut point of p . \square

This proposition says that, under certain assumptions, the existence of conjugate points or cut points along a lightlike geodesic is sufficient for multiple imaging. Such a sufficiency criterion was already proven in Proposition 7 for arbitrary spacetimes. The new feature of Proposition 17 is that in a globally hyperbolic spacetime the worldline γ can be freely prescribed (except for the condition of being inextendible, i. e., “sufficiently long”). In view of applications to gravitational lensing, this is a great advantage since a real light source such as a galaxy or a quasar cannot be expected to travel along a worldline that is constructed as in the proof of Proposition 7.

4.2 Morse Theory in Globally Hyperbolic Spacetimes

In Subsection 3.5 we have established a Morse index theory for Fermat’s principle. Now we want to discuss the possibility of developing a full-fledged Morse theory, relating the number of solution curves to the topology of the space of trial curves. Whereas the Morse index theory works perfectly well on an arbitrary spacetime, the full Morse theory requires a globally hyperbolic spacetime. This is in analogy to the case of Riemannian geodesics where the Morse index theory works on arbitrary Riemannian manifolds but the full Morse theory has to presuppose a complete Riemannian manifold.

For the geodesic problem on complete Riemannian manifolds, Morse theory exists in two versions. The first version, invented by Morse [48] in the 1930s and nicely reviewed by Milnor [46], considers for the space of trial curves the finite dimensional manifold of broken geodesics between two points with N breakpoints. For the final results one has to consider the limit $N \rightarrow \infty$. The second version, brought forward by Palais and Smale [52] [53] in the 1960s, considers for the space of trial curves the infinite dimensional Hilbert manifold of H^1 curves between two points. (For the notion of H^r curves please recall Subsection 3.5.) Both versions have been carried over to general relativity. For Lorentzian manifolds, the most natural analogue of the geodesic problem in complete Riemannian manifolds is the timelike geodesic problem in globally hyperbolic spacetimes. A Morse theory for this situation was developed independently by Uhlenbeck [78] and Woodhouse [87] who both used

finite dimensional approximation techniques in the spirit of Morse and Milnor. An attempt to find an infinite dimensional Hilbert manifold version, in the spirit of Palais and Smale, was brought forward by Everson and Talbot [13] but, unfortunately, turned out to be fatally flawed, see Erratum to [13]. In any case, this timelike geodesic problem has no relevance to gravitational lensing where we are interested in lightlike geodesics between a point and a timelike curve, not in timelike geodesics between two points.

For our purpose, the relevant variational problem is Fermat's principle. A Morse theory based on a version of Fermat's principle in globally hyperbolic spacetimes was invented by Uhlenbeck [78] who, in analogy to her treatment of the timelike geodesic problem in the same paper, used finite dimensional approximation techniques. Her results were used by McKenzie [45] to formulate conditions under which in a gravitational lensing situation the number of images must be odd. As to an infinite dimensional version of Morse theory for Fermat's principle, the most natural starting point seems to be the formalism established in Subsection 3.5. Unfortunately, the parametrization fixing condition on the trial curves, i. e., condition (c) of Definition 6, together with the fact that the trial curves are of type H^2 rather than of type H^1 , leads to many technical problems. For that reason Giannoni, Masiello and Piccione [22] [23] used a slightly different Hilbert manifold setting as the starting point. This approach led, indeed, to a full Morse theory for lightlike geodesics between a point and a timelike curve in a globally hyperbolic spacetime. In the rest of this subsection we shall review their main result and discuss some of its implications. Since the mathematical details are highly technical we have to refer to the original articles [22] and [23] for the proofs.

It was already mentioned that the general setting for treating variational problems in terms of infinite dimensional Hilbert manifolds is due to Palais and Smale [52] [53]. In this setting one considers differentiable functions $F : \mathcal{X} \rightarrow \mathbb{R}$ on a real Hilbert manifold \mathcal{X} . In applications to variational problems, \mathcal{X} is the space of trial curves (or, more generally, trial maps) which is typically infinite dimensional, F is the functional to be extremized, and the critical points of F (i. e., the points where the differential of F has a zero) are the solutions of the variational problem. It is the goal to relate the number of critical points of F to the topology of \mathcal{X} . More precisely, one wants to relate the number N_k of critical points where the Hessian of F has index k to the k -th Betti number B_k of \mathcal{X} . Formally, B_k is defined for each topological space \mathcal{X} in terms of the k -th singular homology space $H_k(\mathcal{X})$ with coefficients in a field \mathbb{F} (The results of Morse theory are true for any choice of \mathbb{F}). For the definition of singular homology spaces the reader is referred, e. g., to Dold [11], p. 32, or to Spanier [75], p. 173. $H_k(\mathcal{X})$ is a vector space over \mathbb{F} and B_k is, by definition, the dimension of this vector space. Geometrically, B_0 is the number of connected components of \mathcal{X} and, for $k \geq 1$, B_k can be interpreted as the number of those "holes" in \mathcal{X} that prevent a k -cycle with coefficients in \mathbb{F} from being a boundary. In particular, if \mathcal{M} is contractible to a point, then

$B_k = 0$ for all $k \geq 1$. Palais and Smale were able to establish the following result, see Corollary (3), p. 338, in Palais [52].

Assume that $F : \mathcal{X} \rightarrow \mathbb{R}$ is of class C^3 at least and satisfies the following conditions.

- (1) F is a *Morse function*, i. e., at each critical point of F the Hessian of F is non-degenerate.
- (2) F is bounded from below.
- (3) F satisfies the so-called *Condition C*, also known as *Palais-Smale condition*: There is a complete Riemannian metric h on \mathcal{X} such that the following holds. If \mathcal{S} is any subset of \mathcal{X} on which F is bounded and $\|dF\|$ is not bounded away from zero, then there is a critical point of F adherent to \mathcal{S} . Here $\|\cdot\|$ denotes the norm induced by the metric h .

Then the Morse inequalities

$$N_k \geq B_k, \quad k \geq 0 \quad (22)$$

and the relation

$$\sum_{k=0}^{\infty} (-1)^k N_k = \sum_{k=0}^{\infty} (-1)^k B_k \quad (23)$$

hold true. The right-hand side of (23) is, by definition, the *Euler characteristic* χ of \mathcal{X} . If one introduces the notation $N_+ = \sum_{i=0}^{\infty} N_{2i}$ and $N_- = \sum_{i=0}^{\infty} N_{2i+1}$, then (23) takes the form

$$N_+ - N_- = \chi. \quad (24)$$

Please note that the N_k and B_k need not be finite.

The geometric idea behind this result is the following. It turns out that the topology of the sublevel set $\mathcal{X}_t = \{x \in \mathcal{X} \mid F(x) \leq t\}$ remains unchanged if t varies over an interval which does not contain a critical value of F (i. e., a value taken by F at some critical point). On intervals containing a critical value, the topology of the sublevel set changes by “attaching a handle” for each critical point, with the special type of the handle determined by the index of the Hessian of F at the critical point. This result was first proven by Morse for functions on compact (and thus finite dimensional) manifolds, see Milnor [46]. In that case Condition C is automatically satisfied. As a matter of fact, Condition C was introduced as a sufficient condition for proving the same “handle-body theorem” without the compactness assumption.

If we want to apply this general result to our variational problem, we have to check if the assumptions on F are satisfied by the arrival time functional $\tau : H_{p,\gamma}^2 \rightarrow \mathbb{R}$ discussed in Subsection 3.5. The Morse index theorem tells us that τ is a Morse function if and only if γ does not intersect the caustic of the past light cone of p . By Proposition 10, this is the case for almost all p once γ has been chosen. Moreover, the Morse index theorem tells us that N_k is the number of past-ponting lightlike geodesics from p to γ that pass through

k conjugate points before arriving at γ , counting each conjugate point with multiplicity. The second condition of F being bounded from below is easily checked to be true on globally hyperbolic spacetimes. The main problem comes with the third condition, i. e., with Condition C. Giannoni, Masiello and Piccione [22] [23] found it necessary to modify the whole setting a little bit before they were able to verify Condition C. First they considered H^1 trial curves, rather than H^2 trial curves, i. e., curves which are differentiable almost everywhere and whose derivative is locally square integrable. Unfortunately, the equation $g(\lambda', \lambda') = 0$ (almost everywhere) does not define a submanifold of $H^1([0, 1], \mathcal{M})$, contrary to the H^2 case. Therefore Giannoni, Masiello and Piccione replaced this with the equation $g(\lambda', \lambda') = -\varepsilon^2$ which, for fixed $\varepsilon > 0$, defines a submanifold and considered the limit $\varepsilon \rightarrow 0$ afterwards. Second, they dropped the parametrization condition (c) of Definition 6. This has the effect that every critical point of the arrival time functional now comes together with all its (H^1) reparametrizations, i. e., τ cannot be a Morse function on this modified space of trial curves. Therefore Giannoni, Masiello and Piccione switched to a new functional Q that is related to the arrival time functional in a similar fashion as the energy functional $E(\alpha) = \int_0^1 h(\alpha'(s), \alpha'(s)) ds$ to the length functional $\ell(\alpha) = \int_0^1 \sqrt{h(\alpha'(s), \alpha'(s))} ds$ in Riemannian geometry. In this modified setting Giannoni, Masiello and Piccione were, indeed, able to verify Condition C, thereby establishing a full Morse theory for the variational problem at hand. Their main result can be phrased in the following way.

Theorem 4. (Morse relations for lightlike geodesics) *Let $(\mathcal{M}, g, \mathcal{T}^+)$ be a globally hyperbolic spacetime, fix a point $p \in \mathcal{M}$ and a past-pointing timelike C^∞ curve $\gamma : I \rightarrow \mathcal{M}$ from an open interval I into \mathcal{M} such that $p \notin \gamma$. Assume that γ is closed in \mathcal{M} and does not intersect the caustic of the past light cone of p . Let $H_{p,\gamma}^1$ denote the topological subspace of $H^1([0, 1], \mathcal{M})$ consisting of all $\lambda \in H^1([0, 1], \mathcal{M})$ with $\lambda(0) = p$, $\lambda(1) \in \gamma$, $g(\lambda', \lambda') = 0$ and λ' past-pointing almost everywhere. Let B_k be the k -th Betti number of $H_{p,\gamma}^1$ and N_k the number of past-pointing lightlike geodesics from p to γ that pass through k conjugate points before reaching γ , counting each conjugate point with multiplicity. Then the Morse relations (22) and (23) hold true.*

This result is implied by Theorem 1.7 of Giannoni, Masiello and Piccione [23]. Actually, they prove a slightly more general result since they consider curves confined to a subset A of \mathcal{M} with certain properties. Our Theorem 4 is the version for $A = \mathcal{M}$ in which case the assumptions placed by Giannoni, Masiello and Piccione upon the functional are satisfied, for each pair (p, γ) , if and only if the spacetime is globally hyperbolic.

The Morse relations have several interesting implications. Information on the N_k , i. e., on the number of images in gravitational lensing situations, place restrictions upon the Betti numbers and the other way round. If we want to make full use of such results we need, of course, some methods of

determining the topology of the curve space $H_{p,\gamma}^1$ which is a difficult task in general. Before commenting on this problem we list some consequences of the Morse relations. Under the assumptions of Theorem 4, the following is true.

(a) The Morse inequality for $k = 0$, i. e. $N_o \geq B_o$, implies that there are at least B_o past-pointing lightlike geodesics from p to γ which are free of conjugate points, where B_o is the number of connected components of $H_{p,\gamma}^1$. This strengthens Proposition 16.

(b) If $H_{p,\gamma}^1$ is non-empty (i. e., $\gamma \cap \mathcal{I}^-(p) \neq \emptyset$) and not contractible, we have $B_o \geq 1$ and $B_k \geq 1$ for some $k \geq 1$. Then there is a multiple imaging situation, $N_o + N_k \geq 2$, and at least one past-pointing lightlike geodesic from p to γ must contain a conjugate point.

(c) If the number of past-pointing lightlike geodesics from p to γ is finite, then all the Betti numbers B_k must be finite.

(d) If we write the Morse relation (23) in the form of (24), we find $N_+ + N_- = 2N_- + \chi$. Thus, in a gravitational lensing situation with finitely many images the total number of images is odd if and only if the Euler characteristic χ is odd.

It is an interesting problem to determine all globally hyperbolic spacetimes in which gravitational lensing always leads to an odd number of images. (Please recall that the assumption of global hyperbolicity implicitly restricts to transparent deflectors.) With the help of Morse theory we have reduced this to the problem of determining the Euler characteristic of the curve space $H_{p,\gamma}^1$, for each pair (p, γ) that satisfies the assumptions of Theorem 4. In some special cases, this can be achieved in the following way, please cf. McKenzie [45] for a similar investigation.

Let us assume that the assumptions of Theorem 4 are satisfied and choose a timelike C^∞ vector field V on \mathcal{M} such that V is tangent to γ . This is possible, see Proposition 5.1 in Giannoni, Masiello and Piccione [22]. Then the projection (21) defines a map $\lambda \mapsto \hat{\lambda} = \pi_V \circ \lambda$ from $H_{p,\gamma}^1$ to the space $\hat{H}_{p,\gamma}^1 = \{\hat{\lambda} \in H^1([0, 1], \mathcal{S}_V) \mid \hat{\lambda}(0) = \pi_V(p), \hat{\lambda}(1) = \pi_V(\gamma)\}$. This map is obviously continuous. Moreover, it is injective since a past-pointing lightlike curve is uniquely determined by its initial point and by its spatial projection. In general, however, it need not be surjective because for some curves in the target space the lightlike lift may terminate (at the ‘‘boundary’’ of \mathcal{M}) before γ has been reached. This certainly happens whenever there is a *particle horizon* in the sense that for some point $q \in \mathcal{I}^-(p)$ there is no past-pointing causal curve from q to γ . As a simple example where particle horizons occur one may consider Minkowski space restricted to the region $t > 0$. Let us say that the *lightlike lifting property* is satisfied for the pair (p, γ) if the map $\lambda \mapsto \hat{\lambda} = \pi_V \circ \lambda$ gives a homeomorphism from $H_{p,\gamma}^1$ onto $\hat{H}_{p,\gamma}^1$. As C^∞ curves are dense in the set of H^1 curves and the lifting procedure is obviously H^1 -continuous, the lightlike lifting property is satisfied if every C^∞ curve in $\hat{H}_{p,\gamma}^1$ is the projection of a curve λ in $H_{p,\gamma}^1$. An analytical condition that guarantees the lightlike lifting property is the so-called *metric growth*

condition of Uhlenbeck [78] which was employed by McKenzie [45]. We can now prove the following result.

Theorem 5. (Odd number theorem) *Assume that all the assumptions of Theorem 4 are satisfied and let V be a timelike C^∞ vector field on \mathcal{M} such that V is tangent to γ . Moreover, assume that the lightlike lifting property is satisfied for (p, γ) and that the space \mathcal{S}_V is contractible. Then the number of past-pointing lightlike geodesics from p to γ is (infinite or) odd.*

Proof. If \mathcal{S}_V is contractible, the curve space $\hat{H}_{p,\gamma}^1$ is contractible. To prove this one fixes a particular C^∞ curve $\hat{\lambda}_o \in \hat{H}_{p,\gamma}^1$ and considers a differentiable map $\phi : [0, 1] \times [0, 1] \times \mathcal{S}_V \rightarrow \mathcal{S}_V$ such that $\phi(s, 0, x) = x$ and $\phi(s, 1, x) = \hat{\lambda}_o(s)$ for all $s \in [0, 1]$ and $x \in \mathcal{S}_V$. The existence of such a map is guaranteed since \mathcal{S}_V is contractible. (It is true that contractibility is defined in terms of homotopies of *continuous* maps. However, according to a well-known theorem every continuous map between two manifolds is homotopic to a C^∞ map, see, e. g., Bott and Tu [7], Proposition 17.8, p. 213.) Now the desired contraction $\Phi : [0, 1] \times \hat{H}_{p,\gamma}^1 \rightarrow \hat{H}_{p,\gamma}^1$ is defined by $\Phi(t, \hat{\lambda})(s) = \phi(s, t, \hat{\lambda}(s))$. Since the lightlike lifting property is satisfied, this implies that $H_{p,\gamma}^1$ is contractible, i. e., the Morse relations hold with $B_o = 1$ and $B_k = 0$ for $k > 0$ which implies $\chi = 1$. If we write the Morse relation (23) in the form of (24), we find $N_+ + N_- = 2N_- + 1$, i. e., $N_+ + N_-$ is (infinite or) odd. \square

This theorem can be phrased as saying that a transparent deflector produces an odd number of images provided that there are no particle horizons and the spatial topology is trivial. It is, of course, true that particle horizons do occur in many cosmological models which are of physical interest. So, in a sense, the lightlike lifting property may be considered as a reasonable assumption only in gravitational lensing situations where cosmological aspects can be ignored.

An obvious example where the lightlike lifting property is satisfied is the case that V is a complete and hypersurface-orthogonal conformal Killing vector field. In that case Fermat's principle reduces to the geodesic problem for a Riemannian metric h on \mathcal{S}_V , as outlined at the end of Subsection 3.4, and global hyperbolicity is easily checked to be equivalent to completeness of the Riemannian manifold (\mathcal{S}_V, h) . Hence, the Morse theory for Fermat's principle reduces to the standard Morse theory for Riemannian geodesics.

It is an open problem to determine the Euler characteristic of the curve space $H_{p,\gamma}^1$ in cases where particle horizons do occur, i. e., where the lightlike lifting property is violated. Apparently no results in this direction exist so far. The so-called "chronological homotopy theory" employed by Woodhouse [87] might be of help in this connection. The latter is closely related to the Lorentzian fundamental groups of Smith [74] and to the notion of "future one-connectedness" of Flaherty [15] [16] which is also discussed in Beem, Ehrlich and Easley [4].

The formalism presented in this subsection has a (much simpler) analogue in the quasi-Newtonian approximation of gravitational lensing. A Morse the-

ory for the latter situation was developed by Petters [61] [63]. In that case the space of trial curves is genuinely finite dimensional, so one can apply the techniques of Morse without having to consider a limit $N \rightarrow \infty$. In particular, Petters [61] used this formalism to prove an odd number theorem. Already earlier, it was shown by Burke [9] that in the quasi-Newtonian approximation every transparent deflector produces an odd number of images. Burke used a fairly simple argument from differential topology, rather than Morse theory, cf. Schneider, Ehlers and Falco [70], p. 172, and Lombardi [42]. A similar argument will be used in the next section to prove an odd number theorem, without invoking Morse theory, for asymptotically simple and empty spacetimes, see Theorem 6 below.

It should be mentioned that, actually, there are several gravitational lens candidates where an even number of images is observed. Usually astronomers are not troubled by this fact because they found good reasons to assume that in those cases one of the images is too faint to be seen. Also, it might be possible that one image is hidden behind the deflector, or that two images are so close together that they are mistaken for being just one image.

5 Gravitational Lensing in Asymptotically Simple and Empty Spacetimes

In elementary optics one often considers “light sources at infinity” which are characterized by the fact that all light rays emitted from such a source are parallel to each other. In this section we want to introduce the notion of “light sources at infinity” for general-relativistic spacetimes. To that end we have to restrict to a special class of spacetimes called “asymptotically simple and empty”. Roughly speaking, an asymptotically simple spacetime is a spacetime for which the notion of “(future- or past-pointing) light rays going out to infinity” makes sense. The following definition, which is essentially due to Penrose [54], puts this vague idea into precise form, cf., e. g. Hawking and Ellis [29], p. 222.

Definition 8. A spacetime $(\mathcal{M}, g, \mathcal{T}^+)$ is called *asymptotically simple* if there is a strongly causal spacetime $(\tilde{\mathcal{M}}, \tilde{g}, \tilde{\mathcal{T}}^+)$ with the following properties.

- (a) \mathcal{M} is an open submanifold of $\tilde{\mathcal{M}}$ with a non-empty boundary $\partial\mathcal{M}$.
- (b) There is a C^∞ function $\Omega : \tilde{\mathcal{M}} \rightarrow \mathbb{R}$ such that $\mathcal{M} = \{p \in \tilde{\mathcal{M}} | \Omega(p) > 0\}$, $\partial\mathcal{M} = \{p \in \tilde{\mathcal{M}} | \Omega(p) = 0\}$, and the equation $\tilde{g} = \Omega^2 g$ holds on \mathcal{M} .
- (c) Every inextendible lightlike geodesic in \mathcal{M} has two endpoints on $\partial\mathcal{M}$.

$(\mathcal{M}, g, \mathcal{T}^+)$ is called *asymptotically simple and empty* if, in addition,

- (d) there is a neighborhood \mathcal{U} of $\partial\mathcal{M}$ in $\tilde{\mathcal{M}}$ such that the Ricci tensor of g vanishes on $\mathcal{U} \cap \mathcal{M}$.

Asymptotically simple and empty spacetimes are good models for isolated gravitating bodies. Condition (d) of Definition 8 is a way of saying that, sufficiently far away from the gravitating body under consideration, Einstein’s

vacuum field equation is satisfied. This is a reasonable model for a deflector producing gravitational lensing as long as cosmological aspects can be ignored.

Conditions (b) and (c) of Definition 8 imply that in an asymptotically simple spacetime all lightlike geodesics are complete. Indeed, since on \mathcal{M} the equation $\tilde{g} = \Omega^2 g$ is supposed to hold, every lightlike g -geodesic becomes a lightlike \tilde{g} -geodesic by changing the affine parameter according to $ds/d\tilde{s} = \Omega^{-2}$. As Ω is zero on $\partial\mathcal{M}$, this implies that $s \rightarrow \pm\infty$ if the geodesic approaches $\partial\mathcal{M}$. Hence, it is justified to interpret the elements of $\partial\mathcal{M}$ as points at infinity or, more precisely, as those points at infinity which can be reached along light rays. Thus, our plan to consider “light sources at infinity” naturally leads to considering “worldlines” contained in $\partial\mathcal{M}$.

In view of gravitational lensing, the observation that in an asymptotically simple and empty spacetime all lightlike geodesics are complete has the following interesting consequence. By a well-known theorem (see Hawking and Ellis [29], Proposition 4.4.5) a complete lightlike geodesic must contain a pair of conjugate points if the weak energy condition and the so-called “generic condition” are satisfied along this geodesic. By Proposition 7, the occurrence of conjugate points gives rise to multiple imaging. This result may be interpreted as saying that in almost all physically reasonable spacetimes which are asymptotically simple and empty multiple imaging takes place.

Before turning our attention to “light sources at infinity” we have to recall some basic facts about asymptotically simple and empty spacetimes. First we use part (b) of Definition 8 to define a vector field Z on $\tilde{\mathcal{M}}$ by the equation $d\Omega = \tilde{g}(Z, \cdot)$. It is well-known that condition (d) of Definition 8 implies that Z is non-vanishing and \tilde{g} -lightlike at each point of $\partial\mathcal{M}$. For a proof we refer to Hawking and Ellis [29], p. 222. (Please note that Hawking and Ellis include the assumption of $d\Omega$ having no zeros on $\partial\mathcal{M}$ into the definition of asymptotically simple spacetimes. However, it is a well-known result of Penrose [54] that, with the additional assumption (d) of asymptotical emptiness, this property must be automatically satisfied.) As a consequence, $\partial\mathcal{M}$ is a \tilde{g} -lightlike hypersurface of $\tilde{\mathcal{M}}$, ruled by the integral curves of Z which are \tilde{g} -lightlike geodesics (up to parametrization). Those lightlike geodesics are called the *generators* of $\partial\mathcal{M}$.

In combination with assumption (c) of Definition 8, the property of $\partial\mathcal{M}$ being a \tilde{g} -lightlike hypersurface implies that $\partial\mathcal{M}$ has two connected components: \mathcal{I}^+ (pronounced “scri plus”) where future-pointing g -lightlike geodesics terminate and \mathcal{I}^- (pronounced “scri minus”) where past-pointing g -lightlike geodesics terminate. – We now state an important proposition, essentially due to Geroch [21], which determines the global structure of asymptotically simple and empty spacetimes.

Proposition 18. *Let $(\mathcal{M}, g, \mathcal{I}^+)$ be an asymptotically simple and empty spacetime. Then $(\mathcal{M}, g, \mathcal{I}^+)$ is globally hyperbolic and every Cauchy surface is homeomorphic to \mathbb{R}^3 . Either component \mathcal{I}^\pm of $\partial\mathcal{M}$ can be diffeomorphi-*

cally mapped onto $S^2 \times \mathbb{R}$ in such a way that each generator of \mathcal{J}^\pm is mapped onto an \mathbb{R} -line. Here S^2 denotes the 2-dimensional sphere.

If the reader wants to verify the proof of this proposition he or she should consult Newman and Clarke [49] who clarified a subtlety overlooked in the original work of Geroch [21] and in Hawking and Ellis [29], Proposition 6.4.9..

After these preparations we are now ready to discuss gravitational lensing situations with light sources at infinity. To that end we consider, in an asymptotically simple and empty spacetime, a sequence of timelike C^∞ curves $\gamma_n : I \rightarrow \mathcal{M}$ that approach, for $n \rightarrow \infty$, a curve $\gamma : I \rightarrow \mathcal{J}^-$. We want to assume that γ is an immersed curve of class C^1 at least, and that the limit is in the C^1 sense, i. e., that not only $\lim_{n \rightarrow \infty} \gamma_n(s) = \gamma(s)$ in $\tilde{\mathcal{M}}$ but also $\lim_{n \rightarrow \infty} \gamma'_n(s) = \gamma'(s)$ in $T\tilde{\mathcal{M}}$. Since $\gamma'_n(s)$ is g -timelike and thus \tilde{g} -timelike, $\gamma'(s)$ is either \tilde{g} -timelike or \tilde{g} -lightlike. The first case is impossible, since \mathcal{J}^- is a lightlike hypersurface with respect to \tilde{g} , and the second case is possible only if $\gamma'(s)$ is tangent to a generator of \mathcal{J}^- . We are thus led to the following conclusion. In an asymptotically simple and empty spacetime, the worldline of a light source at infinity is to be identified with (a section of) a generator of \mathcal{J}^- .

Please note that this does *not* mean that light sources at infinity move at the speed of light. The (physical) metric g is not defined on $\partial\mathcal{M}$, i. e., it does not make sense to speak of the causal character of a curve $\gamma : I \rightarrow \mathcal{J}^-$ with respect to g . The (unphysical) metric \tilde{g} is but a formal device to introduce a geometric structure on the set of points at infinity. The causal character of curves in $\partial\mathcal{M}$ with respect to \tilde{g} has no direct physical interpretation.

Henceforth we restrict to light sources at infinity with inextendible worldlines, i. e., to (maximal) generators of \mathcal{J}^- . From Proposition 18 we know that the set of generators of \mathcal{J}^- is a manifold diffeomorphic to the 2-sphere S^2 . Hence, the set of all light sources at infinity is in one-to-one correspondence with the points of S^2 . On the other hand, we can consider for any $p \in \mathcal{M}$ the set of all one-dimensional g -lightlike subspaces of $T_p\mathcal{M}$. This, again, gives a manifold diffeomorphic to S^2 which may be called the *sky* at p . Clearly, each point of this manifold determines a g -lightlike past-pointing geodesic through p uniquely up to parametrization (i. e., it determines a light ray arriving at p), and vice versa. Hence, the points of this manifold can, indeed, be identified with the points at the celestial sphere of an observer at p . This construction defines for each $p \in \mathcal{M}$ a C^∞ map

$$f_p : S^2 \rightarrow S^2 \tag{25}$$

by assigning to each point x of the sky at p a light source $f_p(x)$ at infinity by extending the lightlike geodesic tangent to x until it reaches $\mathcal{J}^- \simeq S^2 \times \mathbb{R}$ and projecting onto the first factor afterwards. Henceforth we refer to this map f_p as to the *lens map* at p for light sources at infinity. The lens map can be written in an obvious way with the help of the exponential map. Based

on the same idea, one may try to establish a similar lens map in arbitrary spacetimes. The problem is that in the general situation there is no natural “source sphere”, i. e., no analogue of the sphere at infinity. Nonetheless, a kind of general lens map can be established, as was recently demonstrated by Frittelli and Newman [19]. Their formalism, which is based on the Hamilton-Jacobi equation for families of light rays, will not be used in this article.

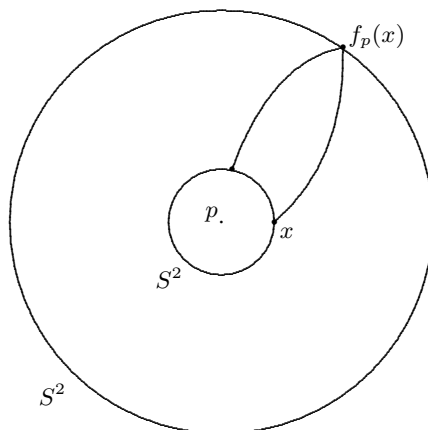


Fig. 6. In this picture the smaller sphere is meant to represent the sky at p and the bigger sphere is meant to represent light sources at infinity. The lens map (25) assigns to each point x of the sky at p a light source $f_p(x)$ at infinity. If f_p is not one-to-one, then there is multiple imaging for light sources at infinity.

The lens map (25) obviously gives all informations on how light sources at infinity are seen by an observer at p . For each light source at infinity, represented by a point $y \in S^2$, the set $f_p^{-1}(y)$ gives all points at the sky of p where this light source is seen, see Figure 6. If $f_p^{-1}(y)$ consists of more than one point, then there is multiple imaging. Please recall that $y \in S^2$ is called a *regular value* of f_p if for all $y \in S^2$ with $f_p(x) = y$ the differential $T_x f_p : T_x S^2 \rightarrow T_y S^2$ has maximal rank (i. e., is surjective). It is easy to check that y is a regular value of the lens map if and only if the generator represented by y does not intersect the caustic of the past light cone of p . Here and in the rest of this subsection the term “light cone” always refers to the light cone in $\hat{\mathcal{M}}$ with respect to the metric \tilde{g} since we need the extension of the “physical” light cone to \mathcal{J}^- . Owing to the well-known Theorem of Sard (see, e. g., Guillemin and Pollack [26], p. 39, or Hirsch [31], p. 69) almost all points $y \in S^2$ are regular values of the lens map. This is in agreement with Proposition 10 according to which the situation that a light source passes through a caustic point is to be viewed as “exceptional”. By compactness of S^2 , $f_p^{-1}(y)$ is finite for any regular value y . Hence, the observer at p sees

finitely many images of each light source at infinity that does not pass through the caustic of the past light cone of p .

We shall now establish the remarkable result that, for each light source at infinity which does not pass through the caustic of the observer's past light cone, the number of images is odd. Based on this observation we shall then prove that the same result is true for a light source moving inside \mathcal{M} provided that its worldline is inextendible and does not approach \mathcal{J}^- . This *odd number theorem* for light sources in asymptotically simple and empty spacetimes will emerge as an application of elementary differential topology; in particular, it will not be necessary to invoke the considerable technical apparatus of Morse theory that was discussed in Subsection 4.2 above.

The proof of our odd number theorem is based on a simple idea. However, the details of the proof look a little bit involved since it is necessary to introduce some cumbersome notation. For that reason we first give the general idea upon which the proof is based. This general idea of an odd-number argument is popular with astronomers who usually present it on the understanding that space and time can be described in a Newtonian fashion. Then the argument goes like this (cf. Schneider, Ehlers and Falco [70], p. 176). Consider all light rays that reach at a particular instant an observer at the point p . Parametrize these light rays with time, in a past-pointing way such that they have $t = 0$ on their arrival at p . Then for all $t > 0$ we can consider the *wavefront* \mathcal{W}_t , defined as the set of all points in space that are crossed by at least one of the considered light rays at the time t . For small t , \mathcal{W}_t is a sphere around p . For larger t , \mathcal{W}_t will develop self-intersections because the light rays are influenced by gravitating masses in the universe. Whenever the wavefront crosses a light source, given by a point moving in space in dependence of time, this gives rise to an image of the light source seen at p . As the wavefronts develop as continuous deformations of a sphere, the following definition makes sense. We say that the light source is *outside* of the wavefront \mathcal{W}_t if the position of the light source at t can be connected to p by a curve that intersects \mathcal{W}_t an odd number of times and *inside* otherwise. Here we have to restrict to curves that intersect the wavefront transversely and only at points where the wavefront is an immersed submanifold, i. e., not at caustic points. Now we assume that the light source itself never touches a wavefront with a tangent velocity vector and that it stays away from caustic points. Then, whenever the light source meets a wavefront, it changes from outside to inside or vice versa. For small t the light source is outside. For large t it is inside, provided that all light rays go out to infinity and the light source does not go out to infinity. This implies that, in total, the wavefront crosses the light source an odd number of times, i. e., that there is an odd number of images.

Instead of the somewhat vague notion of being inside or outside one may use the so-called "mapping degree" from differential topology. Using this terminology the above argument was written down, apparently for the first time,

in the introduction of McKenzie [45]. Gottlieb [25] tried to translate this general idea into a Lorentzian manifold setting and came to the conclusion that the argument does not work without quite special and unrealistic assumptions. However, this conclusion is largely based on the fact that Gottlieb implicitly restricts to multiple imaging situations without time delay. What we want to show in the following is that, with the help of the mapping degree, the above-mentioned odd number argument can be made into a precise theorem in asymptotically simple and empty spacetimes. The crucial point is, of course, that in this case it is guaranteed that all light rays go out to infinity. As an aside, we mention that a slightly different mapping degree argument was used by Lombardi [42] to prove an odd number theorem in the quasi-Newtonian approximation formalism. This is closely related to Burke's [9] original proof of an odd number theorem, again in the quasi-Newtonian approximation, using the index of vector fields. The latter is also discussed in Schneider Ehlers and Falco [70].

The essential tool for our proof is the *mapping degree* of a C^1 map $F : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ between oriented manifolds of the same dimension. In the following we briefly summarize the basic facts about this notion. For more background material the reader is referred to Westenholz [83], Guillemin and Pollack [26] and Dold [11] who, in this order, treat the subject with an increasing amount of abstract mathematics. First we recall that, by the Sard theorem already mentioned above, almost all points $y \in \mathcal{M}_2$ are regular values of F . If, for such a regular value y , the pre-image $F^{-1}(y)$ is contained in a compact set and thus finite, the *local degree of F at y* can be defined by the equation

$$\deg_y(F) = \sum_{x \in F^{-1}(y)} \text{sgn}(x). \quad (26)$$

where $\text{sgn}(x)$ is, by definition, equal to $+1$ if the tangent map $T_x F$ preserves orientation and equal to -1 if $T_x F$ reverses orientation. One can then establish the following facts.

(a) If \mathcal{M}_1 and \mathcal{M}_2 are both compact without boundary, $\deg_y(F)$ is the same for all regular values y . In this case one calls $\deg(F) = \deg_y(F)$ simply the *degree of F* . The degree is a homotopic invariant in the sense that a second C^1 map $\tilde{F} : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ has the same degree as F if and only if F can be continuously deformed into \tilde{F} .

(b) If \mathcal{M}_1 and \mathcal{M}_2 are compact with smooth boundaries, and if F restricts to a map $\partial F : \partial\mathcal{M}_1 \rightarrow \partial\mathcal{M}_2$, then

$$\deg_y(F) = \deg(\partial F) \quad (27)$$

for all regular values $y \in \mathcal{M}_2 \setminus \partial\mathcal{M}_2$.

We are now ready to state and prove the desired odd number theorem.

Theorem 6. *For any point p in an asymptotically simple and empty spacetime (\mathcal{M}, g, T^+) , the following holds true.*

(a) *The lens map (25) has degree one, $\deg(f_p) = 1$. As a consequence, each generator of \mathcal{J}^- that does not pass through the caustic of the past light cone of p can be reached from p along an odd number of lightlike geodesics. (We have already seen above that this number is finite.)*

(b) *Let I be an open interval and $\gamma : I \rightarrow \mathcal{M}$ a timelike embedded C^∞ curve such that γ is closed in \mathcal{M} and has no endpoint on \mathcal{J}^- . (This is a way of saying that γ is inextendible in \mathcal{M} and, in the past direction, does not go out to infinity approaching the velocity of light.) If γ does not pass through the caustic of the past light cone of p , then the number of past-pointing lightlike geodesics from p to γ is finite and odd.*

Proof. Fix a curve γ that satisfies the assumptions of (b). We want to construct a timelike C^∞ vector field V on \mathcal{M} that is tangent to γ and smoothly extends to the vector field Z on \mathcal{J}^- given by $\tilde{g}(Z, \cdot) = d\Omega$. First we choose a future-pointing timelike C^∞ vector field V_1 on some neighborhood \mathcal{U}_1 of γ in \mathcal{M} such that V_1 is tangent to γ . This is possible since γ is an embedding and γ is closed in \mathcal{M} ; the proof can be patterned after the proof of Proposition 5.1 in Giannoni, Masiello and Piccione [22]. Then we choose a future-pointing timelike C^∞ vector field V_2 on an open subset \mathcal{U}_2 of \mathcal{M} whose closure in $\tilde{\mathcal{M}}$ covers \mathcal{J}^- such that V_2 smoothly extends to Z on \mathcal{J}^- . The existence of such a vector field V_2 follows from the fact that \mathcal{J}^- is a closed embedded \tilde{g} -lightlike submanifold of $\tilde{\mathcal{M}}$. Since γ does not approach \mathcal{J}^- , the domains of V_1 and V_2 can be chosen disjoint. Finally, we choose a future-pointing timelike C^∞ vector field V_3 on an open subset \mathcal{U}_3 of \mathcal{M} such that \mathcal{U}_1 , \mathcal{U}_2 and \mathcal{U}_3 cover \mathcal{M} and the closure of \mathcal{U}_3 in $\tilde{\mathcal{M}}$ has void intersection with γ and with \mathcal{J}^- . Then we get the desired vector field V by combining V_1 , V_2 and V_3 with a partition of unity. For this vector field V , we consider the projection (21) onto the 3-manifold of integral curves of V . By Proposition 18, \mathcal{S}_V is homeomorphic to \mathbb{R}^3 . Since, by the theorem of Moise [47] already mentioned above, any 3-dimensional topological manifold admits a unique differentiable structure, \mathcal{S}_V must even be diffeomorphic to \mathbb{R}^3 or, what is the same, to the open unit ball $B = \{x \in \mathbb{R}^3 \mid |x| < 1\}$. Since V extends to the vector field Z on \mathcal{J}^- that is tangent to the generators, π_V extends to a C^∞ map

$$\bar{\pi}_V : \mathcal{M} \cup \mathcal{J}^- \rightarrow \bar{B} \quad (28)$$

between manifolds with boundaries. – Now the vector field V defines a timelike vector V_p at the point p . We choose three spacelike tangent vectors E_1, E_2, E_3 at p with $\tilde{g}(E_\mu, V_p) = 0$ and $\tilde{g}(E_\mu, E_\nu) = -\tilde{g}(V_p, V_p) \delta_{\mu\nu}$. Then each $x \in \mathbb{R}^3$ defines a past-pointing lightlike \tilde{g} -geodesic λ_x by initial conditions $\lambda_x(0) = p$ and $\lambda'_x(0) = x^1 E_1 + x^2 E_2 + x^3 E_3 - |x| V_p$. Since we use the \tilde{g} -affine parametrization, rather than the g -affine parametrization, λ_x arrives at \mathcal{J}^- at a finite parameter value $v(x) \in \mathbb{R}$. Having established the projection (28), the map $x \mapsto \lambda_x$ and the map $x \mapsto v(x)$, we are now ready to prove part (a) of the proposition. For each $x \in S^2$, the curve $\bar{\pi}_V \circ \lambda_x$ in \bar{B} starts at the origin and reaches the boundary at some parameter value $v(x)$.

Hence, for each $r \in]0, 1]$, there is a unique parameter value $u(r, x) \leq v(x)$ such that $|\bar{\pi}_V(\lambda_x(s))|$ is smaller than r for $0 < s < u(r, x)$ and equal to r for $s = u(r, x)$. Then the assignment $x \mapsto \bar{\Phi}_r(x) = \frac{1}{r}\bar{\pi}_V(\lambda_x(u(r, x)))$ gives a C^∞ map $\bar{\Phi}_r : S^2 \rightarrow S^2$. For $r = 1$ we get the lens map (25), $\bar{\Phi}_1 = f_p$. For r sufficiently small, $\bar{\Phi}_r$ is an orientation preserving diffeomorphism, hence $\deg(\bar{\Phi}_r) = 1$. This follows from the fact that the light cone looks like the Minkowski light cone if we restrict to sufficiently short light rays. Since the degree is a homotopic invariant, letting r vary from some small value up to the value 1 shows that the lens map has degree one. Now each generator of \mathcal{J}^- is the pre-image of a point $y \in S^2 = \partial\bar{B}$ under the map $\bar{\pi}_V$. This point y is a regular value of the lens map if and only if the generator does not meet the caustic of the past light cone of p . Under this condition $f_p^{-1}(y)$ is finite, by compactness of S^2 . Let us denote by n_\pm the number of points $x \in f_p^{-1}(y)$ such that $\text{sgn}(x) = \pm 1$. Then the definition of the degree implies $n_+ - n_- = \deg(f_p)$. Since $\deg(f_p) = 1$, this gives $n_+ + n_- = 2n_- + 1$, i. e., the number of points in $f_p^{-1}(y)$ is odd. – Now we prove part (b). To that end we consider the map $F : \bar{B} \rightarrow \bar{B}$ defined by $F(x) = \bar{\pi}_V(\lambda_x(v(x)|x|))$. Clearly, the restriction of this map to the boundary gives the lens map $\partial F = f_p : \partial\bar{B} = S^2 \rightarrow \partial\bar{B} = S^2$. If our curve γ does not meet the caustic of the past light cone of p , the point $y \in B$ with $\bar{\pi}_V^{-1}(y) = \gamma$ is a regular value of F . Hence, by (27), $\deg_y(F) = \deg(f_p) = 1$. By compactness of \bar{B} , the set $F^{-1}(y)$ is finite. As in the proof of part (a), we get $n_+ + n_- = 2n_- + 1$, where n_\pm denotes the number of elements $x \in F^{-1}(y)$ with $\text{sgn}(x) = \pm 1$. Thus, the number of elements in $F^{-1}(y)$ is odd. \square

The fact that the lens map has degree one implies, in particular, that the lens map is surjective which was not obvious from the start. The reader should also consult Kozameh, Lamberti and Reula [38] who state in Lemma 1 a result which is essentially equivalent to the fact that, in our terminology, the lens map has degree one. This paper [38] belongs to a long series of articles by Ted Newman, Carlos Kozameh and various coauthors on studying general relativity in terms of the Hamilton-Jacobi equation for families of lightlike geodesics. For reviews on this topic we refer to Chapter 7 of Joshi [33] and to Kozameh [37]. In particular, these authors have found interesting results on the geometry of “light cone cuts at infinity”, i. e., of intersections of light cones with \mathcal{J}^+ or \mathcal{J}^- in an asymptotically simple and empty spacetime, which are of relevance in view of gravitational lensing.

As stressed already at the beginning of this subsection, an asymptotically simple and empty spacetime is a good model for an *isolated* gravitating body, i. e., if cosmological aspects are ignored. Rudiments of cosmology can be introduced by modifying condition (d) of Definition 8. E. g., one could require $\text{Ric} = \Lambda g$ near $\partial\mathcal{M}$ with a positive or negative cosmological constant Λ , rather than the vacuum field equation $\text{Ric} = 0$. The resulting spacetimes are called *asymptotically deSitter* for $\Lambda > 0$ and *asymptotically anti-deSitter* for $\Lambda < 0$. It was verified already by Penrose [54] that then $\partial\mathcal{M}$ is no longer

\tilde{g} -lightlike but rather \tilde{g} -spacelike for $\Lambda > 0$ and \tilde{g} -timelike for $\Lambda < 0$. In the latter case we can consider immersed worldlines in $\partial\mathcal{M}$ which are \tilde{g} -timelike. For such “light sources at infinity” in an asymptotically anti-deSitter spacetime we have Fermat’s principle in the version of Theorem 1, viewed in the spacetime $(\tilde{\mathcal{M}}, \tilde{g}, \tilde{T}^+)$, at our disposal. This observation was used by Woolgar [88] to prove a positive energy theorem for asymptotically anti-deSitter spacetimes.

The class of asymptotically simple spacetimes seems particularly appropriate for discussing gravitational lensing in a Lorentzian geometry setting. As only a few results in this direction have been worked out so far, all dedicated experts are invited to join the work in this interesting field.

References

1. Abraham, R., Marsden, J. (1978) Foundations of mechanics. Benjamin-Cummings, Reading, Massachusetts
2. Arnold, V. (1990) Singularities of Caustics and Wave Fronts. Kluwer, Dordrecht
3. Arnold, V., Gusein-Zade, S., Varchenko, A. (1985) Singularities of Differentiable Maps . I . Birkhäuser, Boston
4. Beem, J., Ehrlich, P., Easley, K. (1996) Global Lorentzian Geometry. Dekker, New York
5. Blandford, R., Narayan, R. (1986) Fermat’s principle, caustics, and the classification of gravitational lens images. *Astrophys. J.* **310**, 568–582
6. Borde, A. (1987) Geodesic focusing, energy conditions and singularities. *Class. Quantum Grav.* **4**, 343–356
7. Bott, R., Tu, L. W. (1982) Differential forms in algebraic topology. Springer, New York
8. Brill, D. (1973) Observational contacts of general relativity. In Israel, W. (Ed.) *Relativity, Astrophysics and Cosmology, Proceedings of Banff Summer School 1972*. Reidel, Dordrecht, 127–152
9. Burke, W. L. (1981) Multiple gravitational imaging by distributed masses. *Astrophys. J.* **244**, L1
10. Chwolson, O. (1924) Über eine mögliche Form fiktiver Doppelsterne. *Astronomische Nachrichten* **221**, 329
11. Dold, A. (1980) Lectures on algebraic topology. Springer, Berlin
12. Eddington, A. S. (1920) Space, time, and gravitation. Cambridge Univ. Press, Cambridge
13. Everson, J., Talbot, C. (1976) Morse theory on timelike and causal curves. *Gen. Rel. Grav.* **7**, 609–622. Erratum (1978) **9**, 1047
14. Faraoni, V. (1992) Nonstationary gravitational lenses and the Fermat principle. *Astrophys. J.* **398**, 425–428
15. Flaherty, F. (1975) Lorentzian manifolds of non-positive curvature. I. *Proc. Symp. Pure Math.* **27**, No. 2, 395–399
16. Flaherty, F. (1975) Lorentzian manifolds of non-positive curvature. II. *Proc. Amer. Math. Soc.* **48**, 199–202
17. Frankel, T. (1979) Gravitational curvature. Freeman, San Francisco

18. Friedrich, H., Stewart, J. (1983) Characteristic initial data and wavefront singularities in general relativity. *Proc. Roy. Soc. London A* **385**, 345–371
19. Frittelli, S., Newman, Ezra T. (1998) An exact universal gravitational lensing equation. preprint gr-qc/9810017
20. Geroch, R. (1970) Domain of dependence. *J. Math. Phys.* **11**, 417–449
21. Geroch, R. (1971) Space-time structure from a global viewpoint. In Sachs, R. K. (Ed.) *General relativity and cosmology*, Enrico Fermi School, Course XLVII. Academic Press, New York, 71–103
22. Giannoni, F., Masiello, A., Piccione, P. (1997) A variational theory for light rays in stably causal Lorentzian manifolds: Regularity and multiplicity results. *Commun. Math. Phys.* **187**, 375–415
23. Giannoni, F., Masiello, A., Piccione, P. (1998) A Morse theory for light rays on stably causal Lorentzian manifolds. *Ann. Inst. H. Poincaré, Physique Théorique* **69**, 359–412
24. Gott, J. R. (1985) Gravitational lensing effects of a vacuum string: exact solutions. *Astrophys. J.* **288**, 422–427
25. Gottlieb, D. (1994) A gravitational lens need not produce an odd number of images. *J. Math. Phys.* **35**, 5507–5510
26. Guillemin, V., Pollack, A. (1974) *Differential topology*. Prentice-Hall, Englewood Cliffs, NJ
27. Harris, S. (1992) Conformally stationary spacetimes. *Class. Quantum Grav.* **9**, 1823–1827
28. Hasse, W., Kriele, M., Perlick, V. (1996) Caustics of wavefronts in general relativity. *Class. Quantum Grav.* **13**, 1161–1182
29. Hawking, S. W., Ellis, G. F. R. (1973) *The large scale structure of space-time*. Cambridge Univ. Press, Cambridge
30. Helfer, A. D. (1994) Conjugate points on spacelike geodesics or pseudo-self-adjoint Morse-Sturm-Liouville systems. *Pacific J. Math.* **164**, 321–340
31. Hirsch, M. W. (1976) *Differential topology*. Springer, New York
32. Hiscock, W. (1985) Exact gravitational field of a string. *Phys. Rev. D* **31**, 3288–3290
33. Joshi, P. (1993) *Global aspects in gravitation and cosmology*. Clarendon Press, Oxford
34. Kánnár, J. (1991) A note on the existence of conjugate points. *Class. Quantum Grav.* **8**, L179–L184
35. Klingenberg, W. (1982) *Riemannian geometry*. De Gruyter, Berlin
36. Kovner, I. (1990) Fermat principle in gravitational fields. *Astrophys. J.* **351**, 114–120
37. Kozameh, C. (1998) Dynamics of null surfaces in general relativity. In Dadhich, N., Narlikar, J. (Eds.) *Gravitation and relativity: At the turn of the millenium*, Proceedings of GR 15 Conference, 1997. IUCAA, Pune, 139–152
38. Kozameh, C., Lamberti, P. W., Reula, O. (1991) Global aspects of light cone cuts. *J. Math. Phys.* **32**, 3423–3426
39. Leray, J. (1952) *Hyperbolic differential equations*. Institute for Advanced Study, Princeton
40. Landau, L.D., Lifshits, E.M. (1959) *Course of theoretical physics. II: Theory of fields*. Addison–Wesley, Reading, Massachusetts and Pergamon, London
41. Levi-Civita, T. (1917) *Statica Einsteiniana*. *Atti della reale accademia dei lincei, Seria quinta, Rendiconti, Classe di scienze fisiche, matematiche e naturali* **26**, 458–470

42. Lombardi, M. (1998) An application of the topological degree to gravitational lenses. *Modern Phys. Lett. A* **13**, 83–86
43. Low, R. (1998) Stable singularities of wave-fronts in general relativity. *J. Math. Phys.* **39**, 3332–3335
44. Masiello, A. (1994) Variational methods in Lorentzian geometry. Pitman Research Notes in Mathematics Series 309. Longman Scientific & Technical, Essex
45. McKenzie, R. H. (1985) A gravitational lens produces an odd number of images. *J. Math. Phys.* **26**, 1592–1596
46. Milnor, J. (1963) Morse theory. Ann. Math. Studies No. 51, Princeton
47. Moise, E. (1956) Affine structures in 3-manifolds. V. *Ann. Math.* **56**, 96–114
48. Morse, M. (1934) The calculus of variations in the large. Am. Math. Soc. Colloquium Publications XVIII, Providence, Rhode Island
49. Newman, R. P. C., Clarke, C. J. S. (1987) An \mathbb{R}^4 spacetime with a Cauchy surface which is not \mathbb{R}^3 . *Class. Quantum Grav.* **4**, 53–60
50. O’Neill, B. (1983) Semi-Riemannian geometry. Academic Press, New York
51. Padmanabhan, T., Subramanian, K. (1988) The focusing equation, caustics and the condition of multiple imaging by thick gravitational lenses. *Mon. Not. Roy. Astron. Soc.* **233**, 265–284
52. Palais, R. (1963) Morse theory on Hilbert manifolds. *Topology* **2**, 299–340
53. Palais, R., Smale, S. (1964) A generalized Morse theory. *Bull. Amer. Math. Soc.* **70**, 165–172
54. Penrose, R. (1964) Conformal treatment of infinity. In deWitt, C. M., deWitt, B. (Eds.) *Relativity, groups and topology*, Les Houches Summer School 1963. Gordon and Breach, New York, 565–587
55. Perlick, V. (1990) On Fermat’s principle in general relativity. I. The general case. *Class. Quantum Grav.* **7**, 1319–1331
56. Perlick, V. (1990) On Fermat’s principle in general relativity. II. The conformally stationary case. *Class. Quantum Grav.* **7**, 1849–1867
57. Perlick, V. (1995) Infinite dimensional Morse theory and Fermat’s principle in general relativity. I. *J. Math. Phys.* **36**, 6915–6928
58. Perlick, V. (1996) Criteria for multiple imaging in Lorentzian manifolds. *Class. Quantum Grav.* **13**, 529–537
59. Perlick, V. (1999) Ray optics, Fermat’s principle and applications to general relativity. To appear in *Lecture Notes of Physics*, Series m, Springer, Heidelberg
60. Perlick, V., Piccione, P. (1998) A general-relativistic Fermat principle for extended light sources and extended receivers. *Gen. Rel. Grav.* **30**, 1461–1476
61. Petters, A. (1992) Morse theory and gravitational microlensing. *J. Math. Phys.* **33**, 1915–1931
62. Petters, A. (1993) Arnold’s singularity theory and gravitational lensing. *J. Math. Phys.* **34**, 3555–3581
63. Petters, A. (1995) Multiplane gravitational lensing. I. Morse theory of image counting. *J. Math. Phys.* **36**, 4263–4275
64. Petters, A. (1995) Multiplane gravitational lensing. II. Global geometry of caustics. *J. Math. Phys.* **36**, 4276–4295
65. Petters, A., Levine, H., Wambsganss, J. (1999) Singularity theory and gravitational lensing. To appear with Birkhäuser, Boston
66. Poincaré, H. (1905) Sur les lignes géodésiques des surfaces convexes. *Trans. Amer. Math. Soc.* **6**, 237–274

67. Roman, T. A. (1988) On the “averaged weak energy condition” and Penrose’s singularity theorem. *Phys. Rev. D* **37**, 546–548
68. Sasaki, M. (1993) Cosmological gravitational lens equation. Its validity and limitation. *Prog. Theor. Phys.* **83**, 467–491
69. Schneider, P. (1984) A new formulation of gravitational lens theory, time-delay and Fermat’s principle. *Astron. Astrophys.* **143**, 413–420
70. Schneider, P., Ehlers, J., Falco, E. (1992) *Gravitational lenses*. Springer, Heidelberg
71. Schwartz, J. T. (1969) *Nonlinear functional analysis*. Gordon and Breach, New York
72. Seifert, H.-J. (1967) *Kausale Lorentzräume*. Doctoral Thesis, Hamburg University
73. Seitz, S., Schneider, P., Ehlers, J. (1994) Light propagation in arbitrary spacetimes and the gravitational lens approximation. *Class. Quantum Grav.* **11**, 2345–2373
74. Smith, J. W. (1960) Fundamental groups on a Lorentz manifold. *Amer. J. Math.* **82**, 873–890
75. Spanier, E. (1966) *Algebraic topology*. McGraw-Hill, New York
76. Straumann, N. (1984) *General relativity and relativistic astrophysics*. Springer, Berlin
77. Tipler, F. J. (1978) General relativity and conjugate ordinary differential equations. *J. Diff. Equat.* **30**, 165–174
78. Uhlenbeck, K. (1975) A Morse theory for geodesics on a Lorentz manifold. *Topology* **14**, 69–90
79. Vilenkin, A. (1981) Gravitational field of vacuum domain walls and strings. *Phys. Rev. D* **23**, 852–857
80. Wald, R. (1984) *General relativity*. University of Chicago Press, Chicago
81. Walsh, D., Carlswell, R., Weyman, R. (1979) 0957 +561 A,B: twin quasistellar objects or gravitational lens? *Nature* **279**, 381–384
82. Wambsganss, J. (1998) *Gravitational lensing in astronomy*.
<http://www.livingreviews.org/Articles/Volume1/1998-12wamb/>
83. Westenholtz, C. v. (1978) *Differential forms in mathematical physics*. North-Holland, Amsterdam
84. Weyl, H. (1917) Zur Gravitationstheorie. *Annalen der Physik* **54**, 117–145
85. Whitehead, J. H. C. (1935) On the covering of a complete space by the geodesics through a point. *Ann. Math.* **36**, 679–704
86. Whitney, H. (1936) Differentiable manifolds. *Ann. Math.* **37**, 645–680
87. Woodhouse, N. (1976) An application of Morse theory to space-time geometry. *Commun. Math. Phys.* **46**, 135–152
88. Woolgar, E. (1994) The positivity of energy for asymptotically anti-deSitter spacetimes. *Class. Quantum. Grav.* **11**, 1881–1900