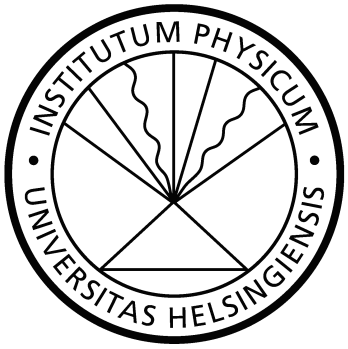

Master's Thesis
Theoretical Physics

Ray-Tracing Based Polarized Radiative Transfer in General Spacetimes

Matias Mannerkoski
2018

Supervisors: Dr. Pauli Pihajoki
Prof. Peter Johansson

Examiners: Prof. Peter Johansson
Dr. Hannu Kurki-Suonio

University of Helsinki
Department of Physics

P.O.Box 64 (Gustaf Hällströmin katu 2)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Physics

Tekijä — Författare — Author
Matias Mannerkoski

Työn nimi — Arbetets titel — Title

Ray-Tracing Based Polarized Radiative Transfer in General Spacetimes

Oppiaine — Läroämne — Subject
Theoretical Physics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	January 2018	124

Tiivistelmä — Referat — Abstract

This thesis presents a ray-tracing based method for performing polarized radiative transfer in arbitrary spacetimes and a numerical implementation of said method. This method correctly accounts for general relativistic effects on the propagation of radiation, and the polarized images and spectra it produces can be directly compared with observations. Thus it is well suited for studying systems where relativistic effects are significant, such as compact astrophysical objects.

The ray-tracing method is based on several approximations, which are discussed in depth. The most important one of these is the geometric optics approximation, which is derived starting from Maxwell's equations. In the geometric optics approximation, high frequency radiation is described as amplitudes or intensities which are propagated along geodesic rays. Additional assumptions about the properties of the radiation field allow describing it and its interaction with matter using the formalism of kinetic theory, which leads to a simple transfer equation along rays. This transfer equation is valid in arbitrary spacetimes, and forms the basis for the ray-tracing method.

The ray-tracing method presented in this work and various similar methods described in the literature are not suited for analytic computations using realistic models. Instead numerical methods are needed. Such numerical methods are implemented in a general fashion in the Arcmancer library (paper in preparation), of which large parts were implemented as a part of this work. The implementation details of Arcmancer are described and its features are compared to those available in other similar codes. Tests of the accuracy of the numerical methods as well as example applications are also presented, including a novel computation of a gravitational lensing event in a binary black hole system. The implementation is found to be correct and easily applicable to a variety of problems.

Avainsanat — Nyckelord — Keywords
General relativity, ray-tracing, polarization, radiative transfer

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1. Introduction

1.1. Background

Compact astrophysical objects, such as neutron stars and black holes, are unlike anything found within the confines of the Solar System. Their extreme properties, such as their strong gravitational fields and high densities, probe the limits of our current knowledge of physics. This makes them extremely interesting objects to study. Electromagnetic radiation, which includes visible light, radio waves and X-rays, has historically been the most important way of observing the Universe outside of the Solar System. It is also the most important way of observing compact objects, although other methods, such as gravitational wave detectors are also significant. To make use of such observations, they must be compared to theoretical models describing the object under study. This is done by modelling the behaviour of the radiation on its way to the observational equipment. One method of performing this modelling of radiative transfer is ray-tracing, which follows the propagation of radiation through space along individual lines or rays.

The main observable properties of electromagnetic radiation are its intensity and polarization at different frequencies. Intensity is simply the brightness of the radiation, while polarization is related to the direction of oscillation of the electromagnetic field in the wave. As radiation propagates through matter, these properties change due to interactions between the matter and the radiation field. These interactions, which typically depend on the frequency of radiation as well as the properties of matter, allow recovering a variety of information about the matter from the observed radiation. For example, in thermal equilibrium matter emits and absorbs radiation in a way that is often well approximated by an ideal black-body spectrum. As a result, the frequency dependence of intensity allows determining the temperature of the matter, with deviations from the ideal black-body spectrum revealing additional information. Another important process is the rotation of the direction of polarization due to magnetic fields, which is known as Faraday rotation.

Often it is possible to model radiative transfer using fairly simple, essentially New-

tonian physics, with some relevant parts of quantum theory also taken into account. However, it is well known that a more correct description of macroscopic physics is given by the framework of Einstein's general relativity, which includes both gravitational effects as well as the effects predicted by the simpler theory of special relativity. Relativistic effects on the propagation of radiation include the apparent bending of the propagation path, known as gravitational lensing due to the similarity of the effect to that of optical lenses, as well as changes in the observed frequency, known as red- and blueshift, both due to the gravitational field and the relative motion of the source. The rotation of massive objects also causes an additional gravitational effect, known as frame dragging, which causes the apparent rotation of the direction of polarization.

General relativity reveals many intuitive notions about time and space to be incorrect by introducing the concept of a curved spacetime whose geometry is controlled by its matter content. As a result correctly accounting for general relativistic effects requires significantly more complicated mathematics than the Newtonian case. Thus it is well justified to use simpler models when possible: the effects of special relativity only become important when relative velocities are large, usually at least 10% of the speed of light, while gravitational effects on radiation become significant usually only near very compact and massive objects, such as black holes and neutron stars. Consequently, the main motivation for a general relativistic formulation of radiative transfer lies in modelling compact objects. However, the limits on the applicability of the Newtonian models depend also on the accuracy of observations. For example, the weak deflection of light by the Sun's gravitational field is one of the famous early pieces of observational evidence in favour of general relativity. General relativistic effects also become important on cosmological scales, in the form of cosmological redshift and gravitational lensing by galaxies and galaxy clusters.

Modelling general relativistic radiative transfer is currently very important, as existing and upcoming instruments have sufficient resolving power to compare observations to general relativistic predictions at high precision. One such instrument is the Event Horizon Telescope, which combines several radio telescopes around the world into a single instrument using very-long-baseline interferometry (VLBI) and finished its first observations in April 2017 (Doeleman, 2017). Using VLBI, the effective resolving power of the system becomes comparable to a single telescope thousand of kilometres in diameter, which is expected to be enough to resolve for instance the shadow the central black hole of the Milky Way casts on the radiation emitted by the matter in its surroundings. This will allow, for example, testing a prediction of general relativity that black holes

are parametrized only by their mass, spin and electric charge, commonly known as the no-hair theorem (Johannsen and Psaltis, 2010; Broderick et al., 2014; Johannsen et al., 2016). Observations of radiation from accretion flows around black holes at high resolution allow also determining parameters of the black hole, such as its mass and spin, as well as the properties of the accreting matter (Huang et al., 2009; Dexter et al., 2010).

Other upcoming instruments of interest are various X-ray detectors such as the Neutron star Interior Composition Explorer (NICER), which has been operational since June 2017 (Gendreau et al., 2012; Gendreau and Arzoumanian, 2017), and the X-ray Imaging Polarimetry Explorer (XIPE) (Soffitta et al., 2013), which is still in a planning phase. These can be used to measure polarized radiation pulse profiles from neutron stars, which allows determining properties of the stars, such as their masses and radii (Lo et al., 2013; Miller and Lamb, 2015). The relation between the masses and radii of neutron stars will constrain models describing the behaviour of high density matter in addition to general relativity. Magnetic fields near neutron stars are also extremely strong, of the order of 10^4 – 10^{11}T. Thus neutron stars also provide environments which allow testing the predictions of quantum electrodynamics in extremely strong magnetic fields, which include strong effects on the polarization of radiation (Novick et al., 1977).

The main method of modelling general relativistic radiative transfer is ray-tracing, which models the propagation of radiation along rays of light, whose paths are governed by geometric optics. This method is reasonably straightforward to apply, and as a result ray-tracing based radiative transfer has been extensively applied to modelling compact objects, often in the form of specialized codes tailored to the specific problem under consideration. Apart from the most trivial cases, the implementation of radiative transfer using ray-tracing requires numerical methods that are quite similar between different applications. This leads to a large amount of duplicated effort. Thus it would be preferable to have a general purpose implementation of general relativistic radiative transfer that could be easily applied to a variety of problems. Obvious requirements to achieve this purpose are support for using different spacetime geometries, which describe the gravitational properties of the system, as well as algorithms and tools which make as few assumptions as possible about the underlying spacetime. However, it appears that so far such an implementation has not been made publicly available, or even created.

1.2. Aims of this Thesis

This thesis presents a ray-tracing method for performing polarized radiative transfer in arbitrary spacetimes and a numerical implementation of said method. Chapters 2

– 4 present the theoretical justifications for the method, paying close attention to the approximations under which the method is valid. This theoretical background of the method is available in the literature, but there does not appear to be one single source that would discuss the general theory of radiative transfer in a self-contained manner even at the level of detail presented here. Rather, the material is scattered between various research articles, with many authors making heavy use of heuristic arguments and implicit assumptions. Thus the aim here is to make these assumptions explicit and identify which parts are supported by rigorous derivations. The main focus in the theoretical treatment lies with the general theory which is based on treating interactions between radiation and matter as small corrections to the vacuum case, and consequently specific processes of matter-radiation interaction are not discussed in depth. The discussion is also based entirely on the classical theory of electromagnetism, and any corrections caused by quantum electrodynamics are assumed to be unimportant or to be similar to the effects of interaction with matter.

While the theoretical discussion of radiative transfer using ray-tracing is mainly based on existing literature, this thesis also presents new work, in the form of a new public[1] C++ library Arcmancer (Pihajoki, Mannerkoski et al. in prep.), which allows easily applying a numerical implementation of the ray-tracing method to various problems in general user-specified spacetimes. It aims to fill the need for a general purpose implementation of general relativistic radiative transfer. Compared to other public and non-public ray-tracing codes described in the literature, Arcmancer has several novel features in addition to the support for general spacetimes, such as support for using multiple coordinate systems at the same time. Large portions of this library were implemented as a part of this work, and the details of this implementation are discussed in chapter 5. Chapter 5 also includes a short comparison to the features available in other ray-tracing codes.

In chapter 6 the accuracy of the numerical procedures implemented in Arcmancer is investigated, and some example applications are presented. Of particular interest are an accuracy comparison between computations performed using different coordinate systems and a computation of images and light curves from an approximation of a binary black hole system, as it does not appear that such computations exist in the literature. Chapter 7 contains a summary of this work and concluding remarks.

[1]The Arcmancer library will be made available together with the accompanying publication.

1.3. Conventions

Units

In this work, the system of units is chosen so that $G = c = 4\pi\epsilon_0 = 1$, where G is the gravitational constant, c the speed of light and ϵ_0 the vacuum permittivity. This system of units agrees for example with the one used by Misner et al. (1973), and allows for example distances to be conveniently measured in units of mass. Additionally, the reduced Planck constant $\hbar$ and the Boltzmann constant k_B are also set to unity, although this choice does not affect the majority of the equations in this work. In this unit system, which is commonly known as Planck units, physical quantities can be represented simply as bare numbers. However, it is possible to use for example SI-units to express quantities where convenient, as the units can be interpreted also as numerical factors.

Sign Conventions

The spacetime metric tensor is taken to have the signature (+ - - -), which differs from the (- + + +) convention that is used for example by Misner et al. (1973). This choice is mainly a matter of personal preference, as both of these conventions are used in the literature, and converting between them is usually a simple matter of inserting factors of -1 in the correct places. There are also other arbitrary choices of sign, for example when relating the electric and magnetic fields to electromagnetic field tensors, which cause some equations to have differing signs when different sources are compared.

Notation

The main notational conventions concern the notation of tensors, which use both index and index-free notation, depending on which is more convenient. In index-free notation, tangent vectors and general tensors are written as u, T, while differential forms are written as $\boldsymbol{\omega}, \boldsymbol{F}$. Three-vectors, i.e. vectors in ordinary three dimensional space, are written as $\vec{v}$. In index notation, all tensors including differential forms are denoted as $u^\mu, F_{\mu\nu}$, with Greek indices running over the values 0–3 when discussing spacetimes, corresponding to (t, x, y, z) in a local Lorentz frame. Indices written using Latin letters $i, j, k \ldots$ run only over 1–3 corresponding to the spatial components (x, y, z). When the mathematics is discussed on a more general level, the indices run over all the dimensions of the manifold. The Einstein summation convention is used, with sums over repeated indices being implicit, e.g. $a^\alpha b_\alpha = \sum_{\alpha=0}^{3} a^\alpha b_\alpha$.

2. Geometry and General Relativity

This chapter begins with an overview of the principles of general relativity, which motivate the methods and approach of this thesis. Differential geometry is the main mathematical tool of relativistic physics, and this chapter also includes the basics of differential forms and the tangent bundle, which are part of the differential geometric tools needed for the theoretical discussion of radiation. The reader is assumed to be familiar with the basic notions of differential geometry, such as tensors, manifolds and curvature, but for reference some of these basic concepts are also discussed in Appendix A.

2.1. Principles of General Relativity

This section provides a brief description of general relativity and discusses its core principles that are central to this work. It also serves to set up some terminology. For detailed treatments of general relativity, see e.g. Misner et al. (1973) or Carroll (2004).

2.1.1. Spacetime and Special Relativity

General relativity builds on special relativity, which explains the observation that all observers measure the speed of light in vacuum c to be the same, regardless of the observers motion relative to each other. The main feature of special relativity is that the natural time and position coordinates of different observers mix with each other depending on the relative motion of the observers. This mixing of time and space leads to the concept of *spacetime*, which is conveniently modelled as a flat, pseudo-Riemannian manifold. This manifold admits global Cartesian coordinate systems or charts (t, x, y, z), in which the metric is given by

$$\left(g_{\mu\nu}\right) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{2.1}$$

This metric is known as the Minkowski metric. For each observer in constant non-accelerated motion there exist charts where the observers spatial (x, y, z)-coordinates are constant, known as the *inertial coordinates* of the observer.

Even though spacetime is a flat manifold in special relativity, it is nevertheless non-Euclidean, with the square norm of tangent vectors, also known as *four-vectors* in the context of relativity, taking also negative values. The sign of the square norm $g_{\mu\nu}V^\mu V^\nu = V^\mu V_\mu$ of a vector V allows classifying vectors into *timelike*, *spacelike* and *null* or *lightlike* vectors, based on whether the square norm is positive, negative or zero, respectively. Timelike vectors can be associated with the direction of the t-axis or the direction of time of some observer, with the unit vector along that direction given by the observers four-velocity u, $u^\mu u_\mu = 1$. Motion with a spatial velocity of c is associated with a null vector, which ensures that the spatial velocity is the same for all observers. Finally, spacelike vectors correspond to directions that appear purely spatial to some observers.

2.1.2. Gravitation

The Newtonian model of gravitation is incompatible with special relativity, as the instantaneous propagation of gravitational signals causes violations of causality. General relativity gives a description of gravitation that does not have this problem by replacing the flat spacetime of special relativity with a spacetime whose geometry is dynamic.

A central feature of gravitation is the familiar observation that all objects fall with the same acceleration, regardless of their mass. This property, known as the universality of gravitation, led to the formulation of the *equivalence principle*, which states that locally the effects of gravitation are indistinguishable from those of an accelerating or non-inertial reference frame. The equivalence principle means that locally the laws of physics are the same as in special relativity, regardless of any gravitational fields.

The equivalence principle also implies that gravitation should be regarded as an effect of the curvature of spacetime, as the observable effects of gravitation are non-local, just as those of curvature. Exactly this is done in general relativity, where the gravitational effects of matter are caused by its coupling to the spacetime curvature through the Einstein field equation

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = T_{\mu\nu}\,, \tag{2.2}$$

where $R_{\mu\nu}$ is the Ricci tensor and $R = R^\mu{}_\mu$ is the Ricci scalar. The matter content is described by the stress-energy tensor $T_{\mu\nu}$ which contains contributions both from the

mass and the pressure of the matter. However, this equation will not be used in this work, apart from the fact that in vacuum it implies that $R_{\mu\nu} = 0$.

2.1.3. Physics in Curved Spacetime

As the spacetime of general relativity is dynamic and curved as opposed to the static and flat spacetime of special relativity, the physical laws used in the context of special relativity may need to be modified to be compatible with it. However, the equivalence principle means that such modifications should be in a sense trivial, as local physical laws cannot couple to curvature. This gives the *minimal coupling principle* for generalizing special relativistic equations to curved spacetimes: a physical law written in coordinate independent form is equally valid regardless of the geometry of spacetime. Here coordinate independent form means an equation that is written in terms of intrinsically geometric quantities, such as tensors and covariant derivatives, so that the system of coordinates used does not enter the equation.

Usually the only complication in the above procedure is that in the context of special relativity it is possible to work with a global preferred system of inertial coordinates, and thus use equations that are not coordinate independent. Here the concept of *local Lorentz frames* or local inertial frames becomes useful. A local Lorentz frame is a reference frame where the spacetime appears locally like the flat spacetime of special relativity: the metric takes the form of the Minkowski metric and the first partial derivatives of the metric components vanish, which causes the Levi-Civita connection coefficients to also vanish. It is always possible to construct such a reference frame using Riemann normal coordinates, which are described in Appendix A.5. Using a local Lorentz frame it is easy to check if a coordinate invariant equation reduces to the known special relativistic form, or even to a non-relativistic equation. For example, it is easy to check that the correct equation of motion of a test particle in a gravitational field is the geodesic equation

$$u^\nu \nabla_\nu u^\mu = \frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu{}_{\alpha\beta} \frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau} \frac{\mathrm{d}x^\beta}{\mathrm{d}\tau} = 0, \tag{2.3}$$

as in a local Lorentz frame it reduces to

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} = 0, \tag{2.4}$$

which just states that the particle undergoes no acceleration, as there are no forces. Here x^μ are the particles coordinates, τ is the proper time, which coincides with the t

coordinate of the particle's local Lorentz frame, $u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}$ is the particle's four-velocity and $\Gamma^\mu{}_{\alpha\beta}$ are the connection coefficients, given by equation (A.18).

2.2. Differential Forms

This section gives a brief introduction to differential forms and exterior calculus, based on Nakahara (2003). In this work differential forms are used for two purposes. In chapter 3, they allow re-expressing electrodynamics of flat spacetime in an elegant form that is simple to generalize to arbitrary spacetimes, while in chapter 4 they are used in the formalism of relativistic kinetic theory to define integration measures in a coordinate free manner.

2.2.1. Differential Forms and the Exterior Derivative

Differential forms of order r, r-forms, are completely antisymmetric tensors of type $(0, r)$. Given a basis of 1-forms $\{\mathrm{d}x^\mu\}$, a general r-form $\boldsymbol{\omega}$ can be expressed in terms of its components as

$$\boldsymbol{\omega} = \frac{1}{r!}\omega_{\mu_1\ldots\mu_r}\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}. \tag{2.5}$$

Here the multiple wedge product $\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}$ is an antisymmetric tensor product of the basis forms, defined as

$$\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r} = \sum_{P\in S_r} \mathrm{sgn}(P)\mathrm{d}x^{\mu_{P(1)}} \otimes \cdots \otimes \mathrm{d}x^{\mu_{P(r)}}, \tag{2.6}$$

where P is a permutation, an element of the permutation group S_r of r elements, and $\mathrm{sgn}(P)$ is the sign of the permutation, 1 for even permutations and -1 for odd permutations. The wedge product extends to the product of an r-form $\boldsymbol{\omega}$ and a q-form $\boldsymbol{\sigma}$ in a straightforward manner, so that $\boldsymbol{\omega} \wedge \boldsymbol{\sigma}$ is an $(r+q)$-form.

While in general differential forms are only defined at a single point on the underlying manifold M, here the components $\omega_{\mu_1\ldots\mu_r}$ of an r-form $\boldsymbol{\omega}$ are assumed to be smooth functions on M, making $\boldsymbol{\omega}$ a smooth r-form field. The antisymmetry of the wedge product implies that $\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}$ vanishes if the same index appears more than once in the set $\{\mu_i\}$. From this follows that Ω^r, the space of smooth r-forms on M, has dimension $\binom{m}{r}$, where $m = \dim M$, with forms with $r > m$ vanishing identically.

The *exterior derivative* d is a linear differential operator that maps r-forms to $(r+1)$-forms. Its action on an r-form $\boldsymbol{\omega}$ is given by

$$\mathrm{d}\boldsymbol{\omega} = \frac{1}{r!}\left(\frac{\partial}{\partial x^\nu}\omega_{\mu_1\ldots\mu_r}\right)\mathrm{d}x^\nu \wedge \mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}. \tag{2.7}$$

The exterior derivative satisfies

$$\mathrm{d}(\boldsymbol{\omega} \wedge \boldsymbol{\xi}) = \mathrm{d}\boldsymbol{\omega} \wedge \boldsymbol{\xi} + (-1)^r \boldsymbol{\omega} \wedge \mathrm{d}\boldsymbol{\xi} \tag{2.8}$$

for an r-form $\boldsymbol{\omega}$ and an arbitrary form $\boldsymbol{\xi}$, and for smooth functions, which can be regarded as 0-forms, it satisfies

$$\mathrm{d}f(X) = \frac{\partial f}{\partial x^\mu}\mathrm{d}x^\mu(X) = X^\mu \frac{\partial f}{\partial x^\mu} = X(f), \tag{2.9}$$

where f is a function and X a tangent vector. If the exterior derivative of a form $\boldsymbol{\omega}$ vanishes, i.e. $\mathrm{d}\boldsymbol{\omega} = 0$, the form $\boldsymbol{\omega}$ is said to be *closed*. From the antisymmetry of the wedge product follows that $\mathrm{dd}\boldsymbol{\sigma} = 0$ for all $\boldsymbol{\sigma}$, alternatively stated as $\mathrm{d}^2 = 0$. Consequently all forms $\boldsymbol{\omega}$ satisfying $\boldsymbol{\omega} = \mathrm{d}\boldsymbol{\sigma}$ for some $\boldsymbol{\sigma}$, which are said to be *exact*, are also closed. However, all closed forms are not in general exact, with non-exact closed forms being related to the topological structure of M. *Poincaré's lemma* states that all closed forms are exact in regions of M which are contractible, i.e. which can be smoothly deformed into a single point.

Differential forms can be understood intuitively in terms of a geometric interpretation, which is discussed at length by Misner et al. (1973). In this interpretation, m-forms on an m-dimensional manifold can associated with a honeycomb-like structure of oriented boxes, or alternatively a density of points with orientation. Lower order forms are then associated with shapes of higher dimensionality, with for example $(m-1)$ forms corresponding to tubes or oriented lines and 1-forms to $(m-1)$-dimensional surfaces. A changing form corresponds to a changing density of these geometrical structures, which causes the structures to have edges. The exterior derivative can then be interpreted as constructing new forms from these edges, with closed forms naturally having no edges. This interpretation and Stokes' theorem, which is discussed in the next section, are illustrated in figure 2.1.

2.2.2. Integration and Stokes' Theorem

Differential forms allow the definition of integration on general orientable manifolds. A smooth non-vanishing m-form $\boldsymbol{\omega}$ on an m-dimensional manifold can be identified as an oriented volume element, a *volume form*, which defines a volume measure on M. A volume form can be written as $\boldsymbol{\omega} = h(p)\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m$, where the orientation of the basis is defined by the order of the basis forms. There are two different equivalence classes of orientations corresponding to even and odd permutations of the basis, with the two different classes related by a change of sign. The integral of a function f over a subset U of the manifold with respect to the volume form $\boldsymbol{\omega}$ is defined as

$$\int_U f\boldsymbol{\omega} = \int_{\phi(U)} f(\phi^{-1}(x))h(\phi^{-1}(x))\,\mathrm{d}x^1 \ldots \mathrm{d}x^m\,, \tag{2.10}$$

where ϕ is the coordinate function corresponding to the coordinates x, and the right hand side is a standard integral in $\mathbb{R}^m$.

While the choice of a volume form used for integrating functions is in general arbitrary, on pseudo-Riemannian manifolds there exists a *natural volume form* related to the metric. It is given by

$$\boldsymbol{\epsilon} = \sqrt{|g|}\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m, \tag{2.11}$$

where $g = \det\left(g_{\mu\nu}\right)$ is the determinant of the matrix of metric components in the basis $\{\mathrm{d}x^\mu\}$. This form is the natural volume form for two-reasons: it has the same expression regardless of coordinate system; and in an orthonormal frame it is just $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m$, which coincides with the standard volume measure $\mathrm{d}^m x$. Carroll (2004) calls $\boldsymbol{\epsilon}$ the Levi-Civita tensor, as in a locally flat frame the components $\epsilon_{\mu_1 \ldots \mu_m}$ coincide with the antisymmetric Levi-Civita symbol.

Integration using differential forms allows also for a generalization of the familiar identities from vector calculus, the generalized Stokes' theorem

$$\int_U \mathrm{d}\boldsymbol{\alpha} = \int_{\partial U} \boldsymbol{\alpha}, \tag{2.12}$$

where $\boldsymbol{\alpha}$ is an $(m-1)$-form, and integration of $\boldsymbol{\alpha}$ over the boundary ∂U of U is defined as

$$\int_{\partial U} \boldsymbol{\alpha} = \int_{\partial U} \iota^*\boldsymbol{\alpha}. \tag{2.13}$$

Here ι is an inclusion map that embeds the $(m-1)$-dimensional manifold ∂U as a submanifold of M, and ι^* is the *pullback* along that map, which takes forms on M to forms on ∂U. Assuming that the coordinates on M are x^μ and those on ∂U are y^μ, the pullback satisfies

$$\iota^*(\boldsymbol{\alpha} \wedge \boldsymbol{\beta}) = (\iota^*\boldsymbol{\alpha}) \wedge (\iota^*\boldsymbol{\beta}) \tag{2.14}$$

$$\iota^* \mathrm{d}\boldsymbol{\alpha} = \mathrm{d}\iota^*\boldsymbol{\alpha} \tag{2.15}$$

$$\iota^* \mathrm{d}x^\mu = \frac{\partial x^\mu}{\partial y^\nu} \mathrm{d}y^\nu. \tag{2.16}$$

In the geometric interpretation of forms, the pullback to a submanifold corresponds simply to taking the intersection of a form with the submanifold. Integrals over submanifolds that are not boundaries are defined similarly. A rigorous definition of integration over subsets and boundaries makes use of chains of simplexes, but this extra technical overhead is not particularly relevant here.

2.2.3. Operations on Forms

A useful operation on differential forms is the *interior product* i_X with a tangent vector X. It is defined by

$$\mathrm{i}_X \boldsymbol{\omega} = \frac{1}{(r-1)!} X^\nu \omega_{\nu\mu_2\ldots\mu_r} \mathrm{d}x^{\mu_2} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}, \tag{2.17}$$

where $\boldsymbol{\omega}$ is an r-form and $\mathrm{i}_X\boldsymbol{\omega}$ an $(r-1)$-form. The interior product has properties analogous to the exterior derivative, satisfying

$$\mathrm{i}_X^2 = 0 \tag{2.18}$$

$$\mathrm{i}_X(\boldsymbol{\omega} \wedge \boldsymbol{\xi}) = \mathrm{i}_X\boldsymbol{\omega} \wedge \boldsymbol{\xi} + (-1)^r \boldsymbol{\omega} \wedge \mathrm{i}_X\boldsymbol{\xi}, \tag{2.19}$$

where again $\boldsymbol{\omega}$ is an r-form and $\boldsymbol{\xi}$ is arbitrary. The Lie derivative of a form is also conveniently expressed using the interior product

$$\mathcal{L}_X\boldsymbol{\omega} = \mathrm{d}\mathrm{i}_X\boldsymbol{\omega} + \mathrm{i}_X\mathrm{d}\boldsymbol{\omega}. \tag{2.20}$$

Another useful identity is

$$\mathrm{d}f \wedge \mathrm{i}_X\boldsymbol{\omega} = X(f)\boldsymbol{\omega} \tag{2.21}$$

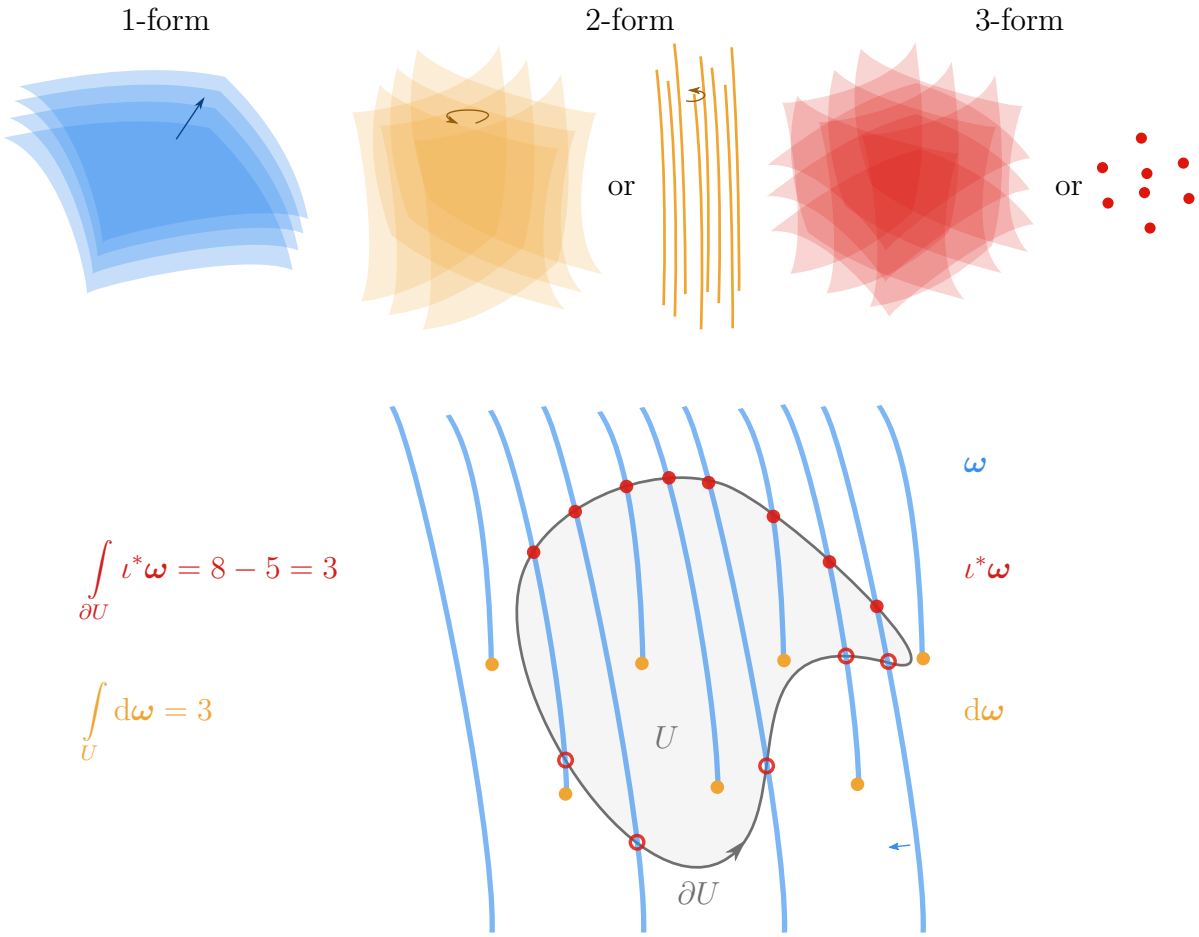

Figure 2.1.: The geometric interpretation of forms. The geometric structures corresponding to different forms in three-dimensional space are depicted at the top, with orientations indicated for the 1- and 2-forms. The density of the structures depends on the scalar multipliers of the form, similarly to how the length of an arrow is used to represent the magnitude of a vector. The lower part illustrates Stokes' theorem and the exterior derivative in two dimensions. A changing 1-form $\boldsymbol{\omega}$ is described by an increasing density of lines, and as a result some lines end or begin. The exterior derivative takes the endings of the lines, turning them into a 2-form which can be integrated over the region U. On the other hand, the pullback $\iota^*\boldsymbol{\omega}$ onto the boundary ∂U is a 1-form that can be integrated over the one dimensional boundary. The open and closed circles correspond to negative and positive values, which depend on the orientations indicated by arrows. Integration corresponds to counting the circles, so Stokes' theorem corresponds to the fact that the total sums of the circles corresponding to $\mathrm{d}\boldsymbol{\omega}$ and $\iota^*\boldsymbol{\omega}$ in U and ∂U are the same.

for a function f and an m-form $\boldsymbol{\omega}$ on an m-dimensional manifold.

Since $\binom{m}{r} = \binom{m}{m-r}$, the spaces of r-forms and $(m-r)$-forms on an m-dimensional manifold have the same dimension and are isomorphic. On pseudo-Riemannian manifolds this isomorphism is given explicitly by the *Hodge dual* $\star$, which is defined by

$$\star(\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_r}) = \frac{1}{(m-r)!}\epsilon^{\mu_1 \ldots \mu_r}{}_{\nu_{r+1}\ldots\nu_m}\mathrm{d}x^{\nu_{r+1}} \wedge \cdots \wedge \mathrm{d}x^{\nu_m}, \tag{2.22}$$

where $\epsilon^{\mu_1 \ldots \mu_r}{}_{\nu_{r+1}\ldots\nu_m}$ are the components of the Levi-Civita tensor (2.11), with indices raised using the metric.

2.3. The Tangent Bundle

The separate tangent spaces T_xM associated with points x of the manifold M can be joined together to form a new manifold, the *tangent bundle*

$$TM = \bigcup_{x \in M} T_xM, \tag{2.23}$$

which in the case of a four-dimensional spacetime is 8-dimensional. The tangent bundle forms a natural setting for relativistic kinetic theory and radiative transfer, discussed in chapter 4.

2.3.1. Basic Concepts

The tangent bundle is an example of the more general concept of a *fibre bundle*. Detailed discussion of the tangent bundle and fibre bundles in general can be found in various textbooks on differential geometry (e.g. Nakahara, 2003; Lee, 1997). For the purposes of this work it is enough to know few of the main properties of the tangent bundle and fibre bundles in general.

As a fibre bundle, the tangent bundle has a map $\pi : TM \to M$ called the *projection*, with its inverse image $\pi^{-1}(x) = T_xM$ called the *fibre* at x. The manifold M is called the *base space*. Additionally, each point in TM is a pair (x, p), with $x \in M$ and $p \in T_xM$, and can consequently be assigned the coordinates (x^μ, p^μ), where x^μ are the coordinates of x in some coordinate chart, and $p = p^\mu \frac{\partial}{\partial x^\mu}$. This gives an explicit isomorphism between $\pi^{-1}(U_i)$ and $U_i \times \mathbb{R}^m$, with U_i a coordinate neighbourhood. Note that as m-dimensional vector spaces the tangent spaces T_xM of an m-dimensional manifold are all isomorphic to $\mathbb{R}^m$. This isomorphism, known as the *local trivialization*, can be characterized as the

tangent bundle, and fibre bundles in general, looking locally like the direct product of two manifolds. In the case of the tangent bundle, other properties of fibre bundles such as coordinate transformation properties under change of chart follow from the properties of tangent vectors. Under the coordinate transformation $x^\mu \mapsto y^\mu$, the p^μ coordinates, which are just components of a tangent vector, must transform as $p^\mu \mapsto \frac{\partial y^\mu}{\partial x^\mu} p^\mu$. Figure 2.2 illustrates the basic concept of the tangent bundle.

2.3.2. Metric

For the purposes of kinetic theory, the tangent bundle needs to be given some additional structure, namely a metric making it a Riemannian manifold. This allows defining natural volume forms, which are necessary for integrating functions on the tangent bundle. A natural metric on the tangent bundle should of course be defined in terms of the base space metric and behave correctly under coordinate transformations, so that the definition of the metric has the same form regardless of coordinates. It should also respect the intuitive notion of directions along the fibres and the base space being orthogonal to each other. The following material is discussed in more detail by Sarbach and Zannias (2014), Lindquist (1966, appendix) and Sasaki (1958).

First the notion of directions along the fibre and the base space should be made precise. The connection on the base space allows for a natural splitting of the tangent space of the tangent bundle $T_{(x,p)}TM$ into *horizontal* and *vertical* components, vertical being the direction along a fibre and horizontal the direction along the base space. The horizontal component is defined using geodesics of the base space, by demanding that the tangent vector of the geodesic lifted to the tangent bundle is horizontal. The *horizontal lift* of a geodesic $\gamma : \lambda \mapsto x^\mu$ with $\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = p^\mu$ is given simply by $\tilde{\gamma} : \lambda \mapsto (x^\mu, p^\mu)$, so its tangent vector L is

$$L = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}\frac{\partial}{\partial x^\mu} + \frac{\mathrm{d}p^\mu}{\mathrm{d}\lambda}\frac{\partial}{\partial p^\mu} = p^\mu\left(\frac{\partial}{\partial x^\mu} - \Gamma^\alpha{}_{\mu\nu}p^\nu\frac{\partial}{\partial p^\alpha}\right) = p^\mu e_\mu, \tag{2.24}$$

where the last equality is taken as the definition of the horizontal basis vectors e_μ. Equivalently, all lifts of geodesics of the base space are integral curves of the vector field $L = p^\mu e_\mu$. Horizontal lifts of tangent vectors are defined similarly by simply replacing $\frac{\partial}{\partial x^\mu}$ with e_μ. The basis vectors of the vertical component are simply $\frac{\partial}{\partial p^\mu}$, which are clearly along the fibre. The basis 1-forms corresponding to e_μ are $\mathrm{d}x^\mu$, and the forms corresponding to $\frac{\partial}{\partial p^\mu}$ are

$$\boldsymbol{\theta}^\mu = \mathrm{d}p^\mu + \Gamma^\mu{}_{\alpha\nu}p^\nu \mathrm{d}x^\alpha. \tag{2.25}$$

Lindquist (1966) calls this basis the *connection basis*, as it is defined by the connection. It is simple to check that with these definitions, horizontal and vertical vectors do not mix under coordinate transformations, as expected. However, this would not be true if instead one attempted to interpret the vectors $\frac{\partial}{\partial x^\mu}$ as being horizontal. This split into horizontal and vertical components is illustrated in figure 2.2.

The conditions on the natural metric stated above give an almost unique choice of metric, which is defined by the conditions

$$e_\mu \cdot e_\nu = g_{\mu\nu} \tag{2.26}$$

$$e_\mu \cdot \frac{\partial}{\partial p^\nu} = 0 \tag{2.27}$$

$$\frac{\partial}{\partial p^\mu} \cdot \frac{\partial}{\partial p^\nu} = g_{\mu\nu}, \tag{2.28}$$

where $g_{\mu\nu}$ is the metric of the base space in the coordinate basis. The metric $\hat{g}_{AB}$ in the connection basis is therefore

$$(\hat{g}_{AB}) = \begin{pmatrix} \left(g_{\mu\nu}\right) & 0 \\ 0 & \left(g_{\mu\nu}\right) \end{pmatrix}, \tag{2.29}$$

where A, B run over all the indices of the basis. In addition to the naturalness conditions, with this metric the dot products between lifted tangent vectors agree with the dot products of the base space tangent vectors. Further, it can be shown that the horizontal lifts of geodesics of the base space $p^\mu e_\mu$ are also geodesics with respect to the Levi-Civita connection of this metric, and that the fibre at any point x is flat (Sasaki, 1958).

It is now possible to construct a natural volume form on the tangent bundle. In the case of a four-dimensional spacetime it is given by

$$\begin{aligned} \boldsymbol{\epsilon}_{TM} &= \sqrt{|\hat{g}|}\mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \boldsymbol{\theta}^0 \wedge \boldsymbol{\theta}^1 \wedge \boldsymbol{\theta}^2 \wedge \boldsymbol{\theta}^3 \\ &= |g|\mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}p^0 \wedge \mathrm{d}p^1 \wedge \mathrm{d}p^2 \wedge \mathrm{d}p^3 \\ &= \boldsymbol{\epsilon} \wedge \boldsymbol{\pi}, \end{aligned} \tag{2.30}$$

where

$$\boldsymbol{\pi} = \sqrt{|g|}\mathrm{d}p^0 \wedge \mathrm{d}p^1 \wedge \mathrm{d}p^2 \wedge \mathrm{d}p^3 \tag{2.31}$$

is the natural volume form on a fibre T_xM with the metric induced from the bundle metric.

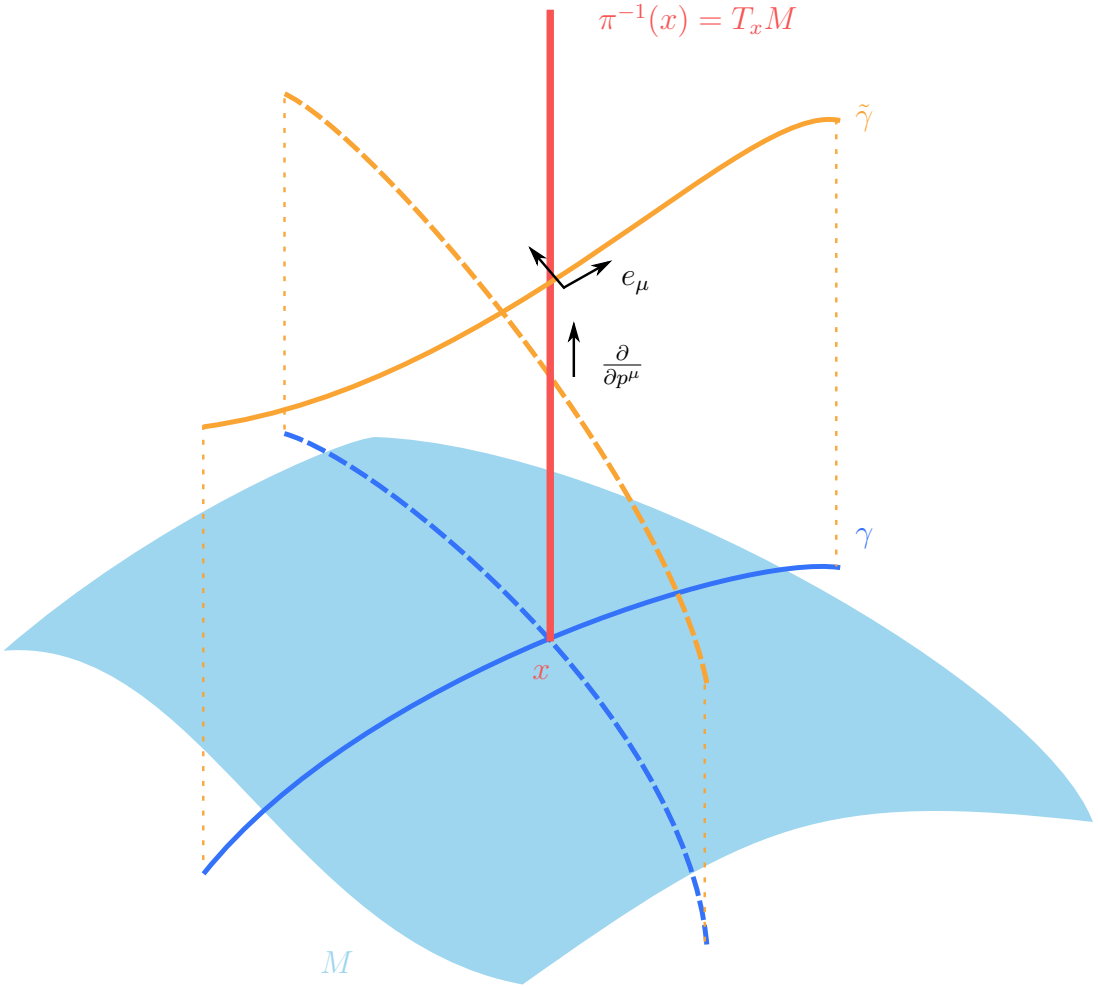

Figure 2.2.: A sketch of the tangent bundle. Each fibre $\pi^{-1}(x)$ or tangent space T_xM at some point x on the base space M can be thought of as lying above the base space. The horizontal lift $\tilde{\gamma}$ of a geodesic lies likewise above the base space geodesic γ. Another geodesic, indicated with dashed lines, crosses γ at x, but the horizontal lifts of the geodesics are separated in the vertical direction, as they have different tangent vectors.

3. Electrodynamics

This chapter focuses on the theory of electromagnetic radiation in terms of the electromagnetic field. The first section contains a short review of the basic description of the electromagnetic field in terms of the electric and magnetic fields in flat spacetime. Plane electromagnetic waves and the quantities describing their polarization are also discussed. After this, in section 3.2, the equations governing the electromagnetic field are generalized to arbitrary spacetimes, and the wave equation governing the propagation of radiation is derived. Finally, in section 3.3, the solution to the wave equation is discussed in the high frequency limit using the geometric optics approximation. The properties of the geometric optics solution will be used in the theory of radiative transfer, which is discussed in chapter 4.

3.1. Review of Basic Electrodynamics

This section contains a review of the basic description of electromagnetism in terms of electric and magnetic fields. Also, the quantities used for the description polarized electromagnetic radiation are introduced. Most of the material in this section is covered in e.g. Jackson (1998).

3.1.1. Maxwell's equations

The fundamental equations of classical electrodynamics are Maxwell's equations. In a non-relativistic setting they are usually given as

$$\boldsymbol{\nabla} \cdot \vec{E} = 4\pi\rho \tag{3.1}$$

$$\boldsymbol{\nabla} \cdot \vec{B} = 0 \tag{3.2}$$

$$\boldsymbol{\nabla} \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0 \tag{3.3}$$

$$\boldsymbol{\nabla} \times \vec{B} - \frac{\partial \vec{E}}{\partial t} = 4\pi\vec{J}, \tag{3.4}$$

where $\vec{E}$ and $\vec{B}$ are the electric and magnetic field 3-vectors, while ρ and $\vec{J}$ are the charge and current densities, respectively. Maxwell's equations imply also the conservation of charge

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \vec{J} = 0. \tag{3.5}$$

The charge and current densities are coupled to the fields through the Lorentz force, which makes solving the equations difficult in general. For a particle with charge q the Lorentz force is given by

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = q\Big(\vec{E} + \vec{v} \times \vec{B}\Big), \tag{3.6}$$

where $\vec{p}$ and $\vec{v}$ are the particle's momentum and velocity, respectively. This can be extended to the charge and current densities by replacing point particles with continuous particle densities.

With certain restrictions, which will be commented on in section 3.2.1, the homogeneous equations (3.2) and (3.3) imply that the electric and magnetic field can be written in terms of a scalar potential ϕ and a vector potential $\vec{A}$ as

$$\vec{E} = -\boldsymbol{\nabla}\phi - \frac{\partial \vec{A}}{\partial t} \tag{3.7}$$

$$\vec{B} = \boldsymbol{\nabla} \times \vec{A}. \tag{3.8}$$

This description in terms of potentials is not unique, as the field quantities are invariant under transformations of the form

$$\vec{A}' = \vec{A} + \boldsymbol{\nabla} f \tag{3.9}$$

$$\phi' = \phi - \frac{\partial f}{\partial t}, \tag{3.10}$$

where f is some suitably smooth function. These transformations are known as *gauge* transformations, with a particular choice of potentials called a choice of gauge.

3.1.2. Waves and Polarization

Maxwell's equations lead to inhomogeneous wave equations for the potentials

$$\left(\frac{\partial^2}{\partial t^2} - \nabla^2\right)\phi = 4\pi\rho \tag{3.11}$$

$$\left(\frac{\partial^2}{\partial t^2} - \nabla^2\right)\vec{A} = 4\pi\vec{J}, \tag{3.12}$$

where the Lorenz gauge condition

$$\frac{\partial \phi}{\partial t} + \boldsymbol{\nabla} \cdot \vec{A} = 0 \tag{3.13}$$

has been used. In vacuum these give homogeneous wave equations also for the electric field

$$\left(\frac{\partial^2}{\partial t^2} - \nabla^2 \right) \vec{E} = 0 \tag{3.14}$$

and similarly for the magnetic field. The solutions to these equations can be decomposed into plane waves of the form

$$\vec{E} = \vec{\mathcal{E}} e^{i\left(2\pi\nu t - \vec{k}\cdot\vec{r}\right)}, \tag{3.15}$$

where ν is the wave's frequency, $\vec{k}$ the wave vector and $\vec{\mathcal{E}}$ is a constant amplitude. Here only the real part of $\vec{E}$ is considered physical, and complex numbers are used only for mathematical convenience. For a wave travelling in the z direction this can be written as

$$\vec{E} = (\mathcal{E}^x \vec{e}_x + \mathcal{E}^y \vec{e}_y) e^{2\pi\nu i(t-z)}, \tag{3.16}$$

where $\vec{e}_i$ are unit basis vectors. The magnetic field $\vec{B}$ is given by a similar expression, but for an electromagnetic wave in vacuum it can be directly determined from the electric field, and does not need to be considered separately.

The direction and behaviour of the electric field of an electromagnetic wave defines the wave's polarization. The polarization state of a plane wave can be described by four real parameters, as the wave (3.16) has two complex components. In observations these parameters are often chosen to be the *Stokes parameters* or intensities, which can be determined by intensity measurements through suitable optical components. For a wave travelling in the z-direction the Stokes parameters are given by

$$I = \left\langle \overline{E}^x E^x + \overline{E}^y E^y \right\rangle \tag{3.17}$$

$$Q = \left\langle \overline{E}^x E^x - \overline{E}^y E^y \right\rangle \tag{3.18}$$

$$U = \left\langle \overline{E}^x E^y + \overline{E}^y E^x \right\rangle \tag{3.19}$$

$$V = -i \left\langle \overline{E}^x E^y - \overline{E}^y E^x \right\rangle, \tag{3.20}$$

where $\overline{E}^x$ denotes the complex conjugate of E^x, and the time average $\langle\rangle$ allows extending the definition to a superposition of multiple plane waves. The parameter I describes the total intensity of the radiation, and is related to the amount of energy carried by the wave.

The parameters Q and U describe the amount and direction of linear polarization, while V describes the amount of circular polarization, with positive values corresponding to a right-handed rotation of the electric field vector as seen by a stationary observer. There are two different conventions used for the positive direction of the V parameter, here the IAU/IEEE definition is used (Shcherbakov and Huang, 2011; Hamaker and Bregman, 1996).

For monochromatic light with a definite polarization state the Stokes parameters can be related to angles describing the polarization ellipse, which is the ellipse swept out by the electric field vector as seen by a stationary observer. These angles are given by

$$\psi = \frac{1}{2} \arctan \frac{U}{Q} \tag{3.21}$$

$$\beta = \frac{1}{2} \arctan \frac{V}{\sqrt{Q^2 + U^2}}, \tag{3.22}$$

and their relationship to the polarization ellipse is shown in figure 3.1. It is also useful to define the degree of polarization

$$P = \frac{\sqrt{Q^2 + U^2 + V^2}}{I}, \tag{3.23}$$

which satisfies $0 \leq P \leq 1$, and describes the total amount of polarized light. For a simple plane wave $P = 1$, while other values are possible for more complicated statistical mixtures of waves.

In a rotation about the z-axis by an angle χ, the Stokes parameters transform as

$$\mathbf{I}' = \begin{pmatrix} I' \\ Q' \\ U' \\ V' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\chi & -\sin 2\chi & 0 \\ 0 & \sin 2\chi & \cos 2\chi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix} = \mathbf{R}(\chi)\mathbf{I}, \tag{3.24}$$

where $\mathbf{I}$ is a vector of the Stokes parameters, the *Stokes vector*. Here it is interesting to note that the parameters remain invariant in rotations by an angle of π. This fact can also be readily inferred from the description in terms of the polarization ellipse, which is likewise invariant under such rotations.

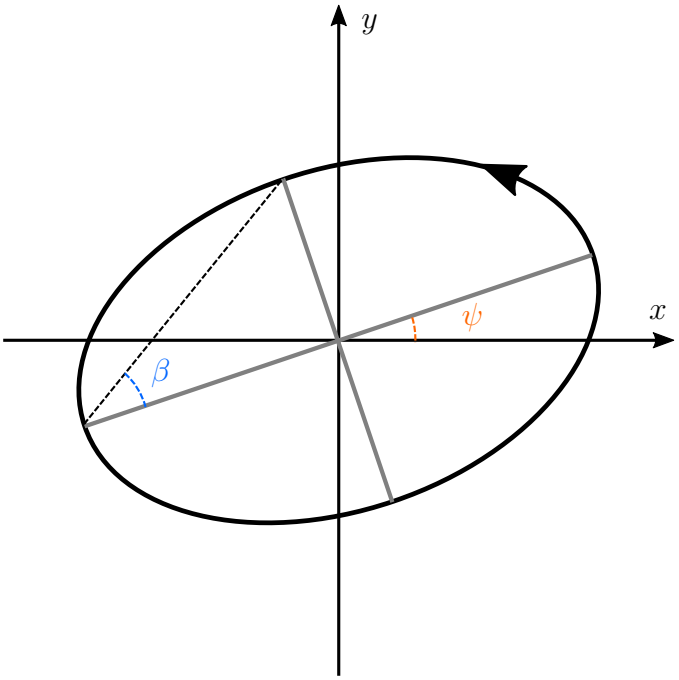

Figure 3.1.: The relation of the angles β, ψ (equations (3.21), (3.22)) to the polarization ellipse. The angles shown correspond to $V > 0$, $Q > U > 0$, with the positive direction of rotation indicated by an arrow. The wave propagates in the z direction, out from the page.

3.2. Electrodynamics in Curved Spacetime

Maxwell's equations are invariant under Lorentz transformations, even if it is impossible to see that at a glance from the form given in section 3.1.1. To make this invariance obvious, and to enable generalization to arbitrary curved spacetimes, the equations can be rewritten in terms of differential forms. The following presentation is based on Misner et al. (1973), with various signs adjusted due to the differing signature conventions.

3.2.1. The Field Tensors and Equations

To find a coordinate invariant formalism for Maxwell's equations, the various three-vector quantities need to be recast in terms of tensors. It turns out that all the relevant quantities can be described using differential forms, allowing the tools of exterior calculus to be used.

First, define the *Faraday tensor* or 2-form

$$\boldsymbol{F} = \frac{1}{2} F_{\mu\nu} \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu, \tag{3.25}$$

with components given in a local Lorentz frame as

$$\left(F_{\mu\nu}\right) = \begin{pmatrix} 0 & E^x & E^y & E^z \\ -E^x & 0 & -B^z & B^y \\ -E^y & B^z & 0 & -B^x \\ -E^z & -B^y & B^x & 0 \end{pmatrix}, \tag{3.26}$$

where E^i, B^i are the components of the electric and magnetic fields, respectively. Another important quantity is the dual of the Faraday tensor, the *Maxwell tensor* $\star\boldsymbol{F}$, which has the components

$$\left(\star F_{\mu\nu}\right) = \begin{pmatrix} 0 & -B^x & -B^y & -B^z \\ B^x & 0 & -E^z & E^y \\ B^y & E^z & 0 & -E^x \\ B^z & -E^y & E^x & 0 \end{pmatrix}, \tag{3.27}$$

again given in a local Lorentz frame. These tensors unify the electric and magnetic fields into a single, geometric object. The charge and current densities are also unified into a single object, the *current 3-form* $\boldsymbol{J}$, which is given by

$$\boldsymbol{J} = -j_\alpha \star \mathrm{d}x^\alpha \tag{3.28}$$

$$(j_\alpha) = \left(\rho, -\vec{J}\right). \tag{3.29}$$

The vector j^α is also a useful quantity, the *current four-vector*.

With these definitions, Maxwell's equations can be written as

$$\mathrm{d}\boldsymbol{F} = 0 \tag{3.30}$$

$$\mathrm{d}\star\boldsymbol{F} = 4\pi\boldsymbol{J}. \tag{3.31}$$

Here equation (3.30) corresponds to equations (3.2) and (3.3), and equation (3.31) to (3.1) and (3.4). The homogeneous equation (3.30) allows defining a *potential 1-form*

$$\boldsymbol{F} = \mathrm{d}\boldsymbol{A} \tag{3.32}$$

$$\boldsymbol{A} = A_\alpha \mathrm{d}x^\alpha \tag{3.33}$$

$$(A_\alpha) = \left(\phi, -\vec{A}\right), \tag{3.34}$$

in any contractible region of spacetime. The restriction to contractible regions is a consequence of Poincaré's lemma, which guarantees that closed forms are exact only in

that case. This same restriction applies naturally also to the formalism where ϕ and $\vec{A}$ are regarded as unrelated objects, but is made evident here. While there are cases where this topological restriction does have important consequences, for the applications considered in this work it can be ignored, as the regions where the potential needs to be defined can always be restricted to contractible ones.

The identity $\mathrm{d}^2 = 0$ allows easy determination of several properties of Maxwell's equations. For example, conservation of charge

$$\mathrm{d}\boldsymbol{J} = -\partial_\alpha j^\alpha \boldsymbol{\epsilon} = -\left(\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \vec{J}\right)\boldsymbol{\epsilon} = 0 \tag{3.35}$$

follows directly from (3.31). Another important case are the gauge transformations of the potential 1-form, which are are just shifts by an exact form,

$$\boldsymbol{A'} = \boldsymbol{A} - \mathrm{d}f. \tag{3.36}$$

These clearly leave $\boldsymbol{F}$ invariant, as

$$\boldsymbol{F'} = \mathrm{d}\boldsymbol{A'} = \mathrm{d}(\boldsymbol{A} - \mathrm{d}f) = \mathrm{d}\boldsymbol{A} - \underbrace{\mathrm{dd}f}_{=0} = \boldsymbol{F}. \tag{3.37}$$

More generally the potential could be shifted by a closed form, but here the Poincaré's lemma again guarantees that such forms are exact, and can be written as $\mathrm{d}f$.

As the operations used in equations (3.30) and (3.31) are fundamentally geometrical, they can be used as is in any spacetime with arbitrary curvature, reducing always to the Maxwell's equations of flat spacetime in a local Lorentz frame. Thus they are the correct generalization of Maxwell's equations to a curved spacetime, assuming that the minimal coupling principle is valid.

3.2.2. The Field Four-Vectors and Lorentz Force

The electric field $\vec{E}$ and the magnetic field $\vec{B}$, as seen by some observer with four-velocity u, can be associated with four-vectors E^μ and B^μ, such that in the local Lorentz frame of the observer

$$(E^\mu) = (0, \vec{E}), \tag{3.38}$$

$$(B^\mu) = (0, \vec{B}). \tag{3.39}$$

From the local component forms (3.26) and (3.27) of the Faraday and Maxwell tensors $\boldsymbol{F}$, $\star\boldsymbol{F}$, and the definition of the Hodge dual (2.22), it can be seen that

$$E_\mu = u^\nu F_{\mu\nu}\,, \tag{3.40}$$

$$B_\mu = u^\nu \star F_{\mu\nu} = \frac{1}{2} u^\nu \epsilon^{\alpha\beta}{}_{\mu\nu} F_{\alpha\beta} \tag{3.41}$$

In a Lorentz frame instantaneously comoving with a charged particle, only the electric field enters the Lorentz force, so

$$\frac{\mathrm{D}p_\alpha}{\mathrm{d}\tau} = u^\nu \nabla_\nu p_\alpha = qE_\alpha = qF_{\alpha\beta}\, u^\beta. \tag{3.42}$$

where u^α is the 4-velocity of the particle and $p_\alpha = mu_\alpha$ is the four-momentum of the particle. The Lorentz force law is therefore given by

$$\frac{\mathrm{D}p_\alpha}{\mathrm{d}\tau} = qF_{\alpha\beta}\, u^\beta, \tag{3.43}$$

which in non-comoving frames includes also the magnetic field term, just as expected.

3.2.3. Wave Equation of the Potential

Inserting the definition of the potential into equation (3.31) gives

$$\mathrm{d}\star\mathrm{d}\boldsymbol{A} = 4\pi \boldsymbol{J}. \tag{3.44}$$

This corresponds to the wave equations (3.11) and (3.12), but this correspondence is not at all obvious. To make this correspondence easier to see, and to simplify further manipulations, it is useful to express equation (3.44) in terms of covariant derivatives and make a transition to index notation.

Since the connection coefficients of the Levi-Civita connection are symmetric in the lower indices, it is possible to replace the partial derivatives in the definition of the exterior derivative (2.7) with corresponding covariant derivatives. As an explicit example of this, consider

$$\begin{aligned} \mathrm{d}\boldsymbol{A} &= (\partial_\alpha A_\beta)\mathrm{d}x^\alpha \wedge \mathrm{d}x^\beta \\ &= (\nabla_\alpha A_\beta)\mathrm{d}x^\alpha \wedge \mathrm{d}x^\beta + \underbrace{\left(A_\nu \Gamma^\nu{}_{\alpha\beta}\right)\mathrm{d}x^\alpha \wedge \mathrm{d}x^\beta}_{=0} \\ &= (\nabla_\alpha A_\beta)\mathrm{d}x^\alpha \wedge \mathrm{d}x^\beta. \end{aligned} \tag{3.45}$$

The left hand side of equation (3.44) can now be written out as

$$\begin{aligned} \mathrm{d}\star\mathrm{d}\boldsymbol{A} &= \Big(\nabla_\alpha\Big[\frac{1}{2}\epsilon^{\mu\nu}{}_{\beta\gamma}\nabla_\mu A_\nu\Big]\Big)\mathrm{d}x^\alpha\wedge\mathrm{d}x^\beta\wedge\mathrm{d}x^\gamma \\ &= \frac{1}{2}\big(\epsilon_{\mu\nu\beta\gamma}\nabla_\alpha\nabla^\mu A^\nu\big)\mathrm{d}x^\alpha\wedge\mathrm{d}x^\beta\wedge\mathrm{d}x^\gamma, \end{aligned} \tag{3.46}$$

where the fact that $\nabla_\alpha\epsilon_{\mu\nu\beta\gamma}=0$ was used. The right hand side reads likewise

$$\begin{aligned} 4\pi\boldsymbol{J} &= -\frac{4\pi}{6}j_\mu\epsilon^\mu{}_{\alpha\beta\gamma}\mathrm{d}x^\alpha\wedge\mathrm{d}x^\beta\wedge\mathrm{d}x^\gamma \\ &= -\frac{4\pi}{6}j^\mu\epsilon_{\mu\alpha\beta\gamma}\mathrm{d}x^\alpha\wedge\mathrm{d}x^\beta\wedge\mathrm{d}x^\gamma. \end{aligned} \tag{3.47}$$

From here, it is possible to read out and equate the components of the forms, finally resulting in

$$\nabla_\alpha\nabla^\alpha A^\beta - \nabla_\alpha\nabla^\beta A^\alpha = 4\pi j^\beta. \tag{3.48}$$

Imposing the Lorenz gauge condition $\nabla_\alpha A^\alpha = 0$ and using the result

$$\nabla_\alpha\nabla_\beta A^\alpha = \nabla_\beta\nabla_\alpha A^\alpha + R^\beta{}_\alpha A^\alpha, \tag{3.49}$$

which follows directly from the definition of the Riemann tensor (A.23), equation (3.48) can be written in the form

$$\nabla_\alpha\nabla^\alpha A^\beta - R^\beta{}_\alpha A^\alpha = 4\pi j^\beta. \tag{3.50}$$

To see that this does indeed correspond to the wave equations (3.11) and (3.12) in flat spacetime, it is enough to note that there $R^\beta{}_\alpha = 0$, $(A^\alpha) = (\phi, \vec{A})$ and $\nabla_\alpha\nabla^\alpha = \frac{\partial^2}{\partial t^2} - \nabla^2$.

3.3. The Geometric Optics Approximation

The behaviour of high frequency radiation can be approximately described using the laws of geometric optics, which in flat spacetime describe the radiation as propagating along straight lines, rays of light, that also refract and reflect at material interfaces. This description generalizes to curved spacetimes, and in this section the general laws of geometric optics describing the propagation of high frequency radiation in vacuum are derived from the wave equation (3.50). The derivation presented here is based on Misner et al. (1973) section 22.5. The results derived here will be later used as a basis for the theory of radiative transfer, which naturally is not very interesting in a vacuum. The

applicability of the idealized vacuum solution to realistic systems will be discussed in section 3.3.4.

3.3.1. The Wave Equation at the Geometric Optics Limit

In the geometric optics approximation, the potential A is assumed to decompose into an amplitude that varies slowly in spacetime, and a rapidly varying phase. Explicitly, the potential is written as

$$A^\mu = \mathcal{A}^\mu e^{i\theta}, \tag{3.51}$$

where the real phase θ is assumed to vary much more rapidly than the complex amplitude $\mathcal{A}^\mu$ in almost all directions in spacetime. It is also assumed that the amplitude of the wave is small enough, so that the effects of the wave on the spacetime do not need to be taken into account. From (3.51) it is already possible to infer that locally the potential appears to be a plane wave, but this will be made more explicit in section 3.3.2. Note that here the use of complex variables is done only for mathematical convenience, with only the real parts of measurable quantities considered physical, just as in section 3.1.2.

Next, the potential will be expanded in a series in terms of a small, constant parameter ε, and the equations corresponding to the wave equation (3.50) will be found in the limit $\varepsilon \to 0$, known as the *geometric optics limit*. This method of approximately solving similar differential equations is also known as the Wentzel-Kramers-Brillouin (WKB) or the eikonal approximation. A suitable expansion parameter ε can be found by considering the characteristic length scales of the system, as measured in the local Lorentz frames of some family of observers. An obvious small length scale of the system is the scale where the phase θ changes appreciably, which corresponds to the wavelength Λ of the radiation. There are several large length scales in the system, the most important ones being the scale where the curvature of the background spacetime becomes significant and the scale where the amplitude $\mathcal{A}^\mu$ of the wave changes considerably. The smaller of these length scales is chosen to be the large length scale L of the system, and the expansion parameter is chosen to be $\varepsilon = \Lambda/L$.

Now the series expansion of the potential can be written down. The phase θ is inversely proportional to the wavelength, so it can be written as $\theta = \theta_0 \varepsilon^{-1}$, with θ_0 not depending on ε. The amplitude $\mathcal{A}^\mu$ can be simply expanded in powers of ε,

$$\mathcal{A}^\mu = a^\mu + \varepsilon b^\mu + \varepsilon^2 c^\mu + \mathcal{O}\left(\varepsilon^3\right), \tag{3.52}$$

so

$$A^\mu = \Big(a^\mu + \varepsilon b^\mu + \varepsilon^2 c^\mu + \mathcal{O}\Big(\varepsilon^3\Big)\Big) e^{i\theta_0/\varepsilon}. \tag{3.53}$$

Assuming a vacuum background, the wave equation (3.50) simplifies to

$$\nabla_\alpha \nabla^\alpha A^\mu = 0, \tag{3.54}$$

since in vacuum $j^\mu = 0$ and $R^\mu{}_\nu = 0$. The Ricci tensor $R^\mu{}_\nu$ could be ignored also without the assumption of a vacuum background, as $R^\mu{}_\nu \propto L^{-2} \propto \varepsilon^2$ if L is the curvature length scale. This can be seen for example by noting that like $\nabla_\alpha \nabla^\alpha$, the Ricci tensor has units of $[\text{Length}]^{-2}$, and the only relevant length scale is the curvature length scale. The series expansion (3.53) can now be inserted into the wave equation (3.54) and the Lorenz gauge condition $\nabla_\mu A^\mu = 0$. For the gauge condition this yields

$$\begin{aligned} 0 &= \nabla_\mu \Big((a^\mu + \varepsilon b^\mu + ...) e^{i\theta_0/\varepsilon}\Big) \\ &= \Big((a^\mu + \varepsilon b^\mu + ...)\frac{i}{\varepsilon}\nabla_\mu \theta_0 + \nabla_\mu (a^\mu + \varepsilon b^\mu + ...)\Big) e^{i\theta_0/\varepsilon}, \end{aligned} \tag{3.55}$$

and similarly for the wave equation

$$\begin{aligned} 0 = \Big(&- \frac{1}{\varepsilon^2}\nabla_\nu \theta_0 \nabla^\nu \theta_0 \Big(a^\mu + \varepsilon b^\mu + \varepsilon^2 c^\mu ...\Big) \\ &+ \frac{i}{\varepsilon}[(a^\mu + \varepsilon b^\mu + ...)\nabla_\nu \nabla^\nu \theta_0 + 2\nabla^\nu \theta_0 \nabla_\nu (a^\mu + \varepsilon b^\mu + ...)] \\ &+ \nabla_\nu \nabla^\nu (a^\mu + \varepsilon b^\mu + ...)\Big) e^{i\theta_0/\varepsilon}. \end{aligned} \tag{3.56}$$

At this point, it is useful to define the *wave vector* $k^\mu = \nabla^\mu \theta = \varepsilon^{-1}\nabla^\mu \theta_0$. To keep the dependence on ε explicit, define also $\kappa^\mu = \varepsilon k^\mu$. Now, as in the end the limit $\varepsilon \to 0$ will be taken, it is enough to only consider the terms of order 0 or lower in ε. The terms of different orders must also satisfy the equations independently, so that taking the limit is sensible. The gauge condition (3.55) gives the equations

$$0 = \frac{i}{\varepsilon}\kappa_\mu a^\mu \tag{3.57}$$

$$0 = \nabla_\mu a^\mu + i\kappa_\mu b^\mu, \tag{3.58}$$

while the wave equation (3.56) gives

$$0 = \varepsilon^{-2}\kappa^\nu\kappa_\nu a^\mu \tag{3.59}$$

$$0 = \frac{i}{\varepsilon}(i\kappa^\nu\kappa_\nu b^\mu + a^\mu\nabla_\nu\kappa^\nu + 2\kappa^\nu\nabla_\nu a^\mu) \tag{3.60}$$

$$0 = -\kappa^\nu\kappa_\nu c^\mu + ib^\mu\nabla_\nu\kappa^\nu + 2i\kappa^\nu\nabla_\nu b^\mu + \nabla_\nu\nabla^\nu a^\mu. \tag{3.61}$$

Misner et al. (1973) actually ignore the 0th order equations (3.58) and (3.61) completely, simply stating that they control corrections to the geometric optics limit. This can be seen to be due to the b terms which do not contribute to the potential A when $\varepsilon \to 0$, and can be shown by a simple degrees-of-freedom argument: the 0th order equations contain five constraints on the four components b^μ and the four directional derivatives $\kappa^\nu\nabla_\nu b^\mu$, so there is clearly enough freedom for the 0th order equations to be always satisfied at the geometric optics limit. Equation (3.59) gives $\kappa^\nu\kappa_\nu = 0$ for a non-trivial solution, so the term containing b in (3.60) vanishes. Therefore in the geometric optics limit, the vacuum wave equation (3.54) reduces to the equations

$$k^\mu k_\mu = 0 \tag{3.62}$$

$$k_\mu a^\mu = 0 \tag{3.63}$$

$$a^\mu\nabla_\nu k^\nu + 2k^\nu\nabla_\nu a^\mu = 0, \tag{3.64}$$

where the wave vector k has been restored, as the dependence on ε is no longer needed.

3.3.2. Rays and Propagation Equations

The equations (3.62)–(3.64) allow deriving a set of equations describing rays of light and the propagation of quantities along those rays. The equations governing the rays follow from equation (3.62), which directly states that the wave vector k is a null vector. Taking a covariant derivative of this equation yields

$$0 = 2k_\nu\nabla^\mu k^\nu = 2k^\nu\nabla^\mu\nabla_\nu\theta = 2k^\nu\nabla_\nu\nabla^\mu\theta = 2k^\nu\nabla_\nu k^\mu, \tag{3.65}$$

as the covariant derivatives of the scalar θ commute. This is the geodesic equation for an affinely parametrized geodesic with tangent vector k. Equation (3.62) gives $k^\mu\partial_\mu\theta = \frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = 0$, so the phase is constant along the geodesic and the wave propagates along it. Thus, the null geodesics with the wave vector k as their tangent can be identified as the rays of light of geometric optics. The fact that light propagates along null geodesics is

familiar from any introductory discussion of general relativity, where it is usually stated without derivation as an obvious generalization of the straight null rays of flat spacetime.

In a local Lorentz frame a series expansion of the phase gives in the vicinity of the observer

$$\theta \approx \theta(0) + k_\mu x^\mu = \theta(0) + k^t t - \vec{k} \cdot \vec{x}. \tag{3.66}$$

Comparing this with equation (3.15), it can be seen that this is just the plane wave solution with $k^t = 2\pi\nu = \omega$, where ν is the frequency of the wave and ω the corresponding angular frequency. As $k^t = u_\mu k^\mu$, with u the four-velocity of the observer, this gives an useful way of finding the frequency of the wave as seen by any given observer. Since $k \propto \varepsilon^{-1}$, here it can also be seen that the geometric optics limit $\varepsilon \to 0$ corresponds to the limit of infinite frequency $\nu \to \infty$.

The propagation equations for the amplitude of the wave follow from the remaining equations. Here it is useful to consider separately a real scalar amplitude a and a complex polarization vector f, which are related to the amplitude a by

$$a^\mu = \mathsf{a} f^\mu, \tag{3.67}$$

$$f^\mu \overline{f}_\mu = -1, \tag{3.68}$$

$$\mathsf{a}^2 = -a^\mu \overline{a}_\mu. \tag{3.69}$$

The propagation equation for the scalar amplitude follows from equation (3.64) and its complex conjugate by calculating

$$\begin{aligned} 2\mathsf{a} k^\mu \nabla_\mu \mathsf{a} &= k^\mu \nabla_\mu \mathsf{a}^2 \\ &= -k^\mu \Big(a^\nu \nabla_\mu \overline{a}_\nu + \overline{a}_\nu \nabla_\mu a^\nu \Big) \\ &= \frac{1}{2} (a^\nu \overline{a}_\nu + a^\nu \overline{a}_\nu) \nabla_\mu k^\mu \\ &= -\mathsf{a}^2 \nabla_\mu k^\mu, \end{aligned} \tag{3.70}$$

which gives the propagation equation

$$k^\mu \nabla_\mu \mathsf{a} = -\frac{1}{2} \mathsf{a} \nabla_\mu k^\mu. \tag{3.71}$$

This equation can also be written in the form

$$\nabla_\mu \Big(\mathsf{a}^2 k^\mu \Big) = 0, \tag{3.72}$$

which is a conservation law. Misner et al. (1973) interpret this as the conservation law for a Newtonian photon number, with the electromagnetic wave conceived as being associated with a group of well localized particles propagating along the rays of light. In flat spacetime this corresponds to the conservation of energy.

With equation (3.71), the behaviour of the polarization vector can now be extracted from equation (3.64):

$$\begin{aligned} 0 &= \mathsf{a} f^\mu \nabla_\nu k^\nu + 2k^\nu \nabla_\nu (\mathsf{a} f^\mu) \\ &= 2\mathsf{a} k^\nu \nabla_\nu f^\mu + f^\mu \underbrace{\left(\mathsf{a}\nabla_\nu k^\nu + 2k^\nu \nabla_\nu \mathsf{a}\right)}_{=0}. \end{aligned} \tag{3.73}$$

This implies that the polarization vector is parallel transported along the rays:

$$k^\nu \nabla_\nu f^\mu = 0. \tag{3.74}$$

The parallel transport of the polarization vector is a natural generalization of the constancy of polarization of a plane wave in flat spacetime. The polarization vector is also restricted by the gauge condition, equation (3.63), which gives the condition $k_\mu f^\mu = 0$, i.e. polarization vector is orthogonal to the wave vector. The Lorenz gauge condition does not fix the polarization vector completely, as there still remains some freedom to specify the gauge. Most notably, it is possible to add an arbitrary multiple of k to the polarization vector.

3.3.3. The Field Tensor

For later use, expressions for the Faraday tensor and the electric field four vector are needed in this approximation. The Faraday tensor is given by

$$\begin{aligned} F_{\mu\nu} &= \nabla_\mu A_\nu - \nabla_\nu A_\mu \\ &= \left[\frac{i}{\varepsilon}(\kappa_\mu a_\nu - \kappa_\nu a_\mu) + \nabla_\mu a_\nu - \nabla_\nu a_\mu\right] e^{i\theta} \\ &= i(k_\mu a_\nu - a_\mu k_\nu) e^{i\theta}, \end{aligned} \tag{3.75}$$

as the ε^{-1} term dominates in the geometric optics limit. The electric field four-vector is then given by

$$\begin{aligned} E^\mu &= F^\mu{}_\nu \, u^\nu \\ &= i(k^\mu a_\nu u^\nu - \omega a^\mu) e^{i\theta}. \end{aligned} \tag{3.76}$$

It is always possible to choose the gauge so that $a_\nu u^\nu = 0$, simply by taking $a^\mu \mapsto a^\mu - \omega^{-1} a^\nu u_\nu k^\mu$. In this case the expression for the electric field simplifies to

$$E^\mu = -i\omega a^\mu e^{i\theta}, \tag{3.77}$$

which is again evidently locally a plane wave.

3.3.4. Validity of the Approximation and Interaction with Plasma

The equations governing the geometric optics approximation only hold exactly in vacuum and in the infinite frequency limit, whereas any physically interesting application requires at least a finite frequency, and often some interaction with matter. Therefore, to justify using the equations derived above also for these cases, the errors in the approximation should be quantified.

The problem of infinite frequency does not appear to be too serious at first glance, as finite frequencies can be described simply by not taking the $\varepsilon \to 0$ limit. Now all the higher order terms in (3.53) contribute to the potential, but as they are of order $\mathcal{O}(\varepsilon)$, it is safe to ignore them as long as ε is small enough. However, the definition of the expansion parameter ε in terms of length scales as measured by some set of observers causes some further problems for applying this reasoning, as the set of observers used for the definition is not specified in any way. In particular, Mashhoon (1987) argues that the dependence on curvature length scales allows in most cases the construction of a set of observers for which $\varepsilon > 1$ for all finite frequencies. Anyhow, this appears to be more of a problem with the approximation method used rather than with the physical results, and should not cause any actual problems for applications.

A more concrete physical issue for applications is the assumption of a vacuum background. In general any matter content will alter the propagation of radiation, as the radiation will generate a current j through the Lorentz force (equation 3.43), and this current also enters the wave equation (3.50). These effects are analysed by Breuer and Ehlers (1980, 1981) in the case of a cold plasma, which is highly relevant for astrophysical applications.

In the case of an unmagnetized plasma the modifications are simple: the rays are now timelike curves instead of null geodesics and the amplitude transport equation contains terms related the motion of the plasma, in addition to the terms present in equation (3.64). The property of the plasma that determines the paths of the rays is the *plasma frequency* ω_p, which is the characteristic frequency of oscillations in the plasma.

The plasma frequency determines the index of refraction in the plasma rest frame, i.e. the speed at which the radiation propagates relative to the plasma. As happens when dealing with other materials with varying index of refraction, a changing plasma frequency will effectively act as a force on the ray, causing it to bend.

In the case of a magnetized plasma, additional effects appear due to the anisotropy introduced by the magnetic field. Now there will be two different polarization eigenmodes, which will be in general two different elliptically polarized modes. These modes will in general follow different trajectories, but in the high frequency limit the spatial paths coincide in the plasma rest frame, with only the propagation velocities being different. This difference in propagation velocities leads to the phenomenon of *Faraday rotation*, which causes the polarization direction of a superposition of the eigenmodes to rotate. The strength of the effects related to the magnetic field are proportional to the Larmor frequency ω_L, which is the frequency of electron gyration in the magnetic field.

In both cases the corrections to the vacuum solution are proportional to the ratios ω_p/ω and ω_L/ω, with ω the angular frequency of the radiation in the plasma rest frame. Therefore the vacuum solution for the ray path should be a usable approximation if $\omega \gg \omega_p, \omega_L$. When this can be considered to hold depends of course on the application, as over sufficiently large distances even small differences in propagation velocity can cause observable differences. On the other hand, the propagation equation for the amplitude clearly requires modifications to correctly capture the physically interesting behaviour. These modifications will be included in a more general fashion in chapter 4.

4. Radiative Transfer

This chapter discusses the theory of radiative transfer, which can be roughly summarized as being a macrophysical description of radiation as compared to the microphysical description in terms of the electromagnetic field, similarly to how the continuum description of fluids is a macrophysical counterpart to the microphysical description in terms of molecules. The first section of this chapter gives an overview of the non-relativistic phenomenological theory of radiative transfer, while the rest of this chapter is focused on generalizing the theory to general spacetimes. This generalization proceeds by combining the geometric optics approximation with ideas from the statistical description of matter using kinetic theory. The final result is a relativistic radiative transfer equation, which can be solved using the ray-tracing algorithm, described at the end of this chapter.

4.1. Overview of Non-Relativistic Transfer Theory

This section gives a brief overview of the non-relativistic theory of radiative transfer. The non-relativistic description is not compatible with gravitational effects or large velocities, but the emission and absorption coefficients it employs can be related to the quantities used in the relativistic description. These coefficients are available in the literature for a variety of different systems, so connecting the relativistic description to the non-relativistic one is very useful.

The non-relativistic theory of radiative transfer is usually formulated as a phenomenological theory by assuming that radiation propagates along straight rays, i.e. by working in the geometric optics limit in flat spacetime (Rybicki and Lightman, 2008). The central quantity describing the radiation field is the *specific intensity*, which is defined as

$$I_\nu = \frac{\delta E}{\delta A \delta t \delta \nu \delta \Omega}, \tag{4.1}$$

that is, the infinitesimal amount of energy δE flowing in the direction of the normal of a surface with area δA within a solid angle $\delta\Omega$ in a frequency interval $\delta\nu$ around the

frequency ν in a time interval δt. Often the specific intensity, having units of

$$\frac{[\text{Power}]}{[\text{Area}][\text{Frequency}][\text{Solid angle}]},$$

is called simply the intensity, which may lead to confusion with the plane wave Stokes intensity parameter I, which has units of [Power]/[Area]. Quantities related to the Stokes intensity I can be obtained from the specific intensity by integrating it over solid angle and frequency. The integral over solid angle is known as the *specific flux* $F_\nu = \int \mathrm{d}\Omega\, I_\nu$, and integrating it over frequency yields the total or *bolometric flux* $F = \int \mathrm{d}\nu\, F_\nu$, which has the same units as the plane wave Stokes intensity I.

From geometry, it follows that specific intensity is conserved in empty space. When the radiation interacts with matter, which is assumed to behave as a continuum, conservation of energy allows deriving the transfer equation

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = j_\nu - \alpha_\nu I_\nu, \tag{4.2}$$

where s is the distance along a ray, j_ν the emission coefficient describing emission of radiation and α_ν the absorption coefficient describing the absorption of radiation. The effects of scattering of radiation can be included as term of the form

$$\int \mathrm{d}\Omega\, S I_\nu, \tag{4.3}$$

where S describes how strongly radiation is scattered from one propagation direction to another, but for convenience this term can be taken to be a part of j_ν. When there is no emission, equation (4.2) has the solution

$$I_\nu(s) = I_\nu(0) \exp\Big(-\int_0^s \alpha_\nu \,\mathrm{d}s\Big). \tag{4.4}$$

This makes it convenient to define the *optical depth* τ as

$$\tau = \int_0^s \alpha_\nu \,\mathrm{d}s\,. \tag{4.5}$$

When the optical depth through a region is large, i.e. $\tau \gtrsim 1$, the intensity of the radiation is strongly attenuated and the region is said to be optically thick. Conversely, regions with low optical depth ($\tau \ll 1$) are said to be optically thin.

Polarization of the radiation can be included in the description by generalizing

the Stokes parameters $\mathbf{I}$ of a plane electromagnetic wave to the *specific intensity Stokes parameters* $\mathbf{I}_\nu$, which are usually called simply the Stokes parameters, causing a possibility for confusion. Degl'Innocenti and Landolfi (2006) define these by the energy measured by an ideal detector through polarizing filters. This process corresponds to applying the formulas (3.17) – (3.20) to a field $\vec{E}$ which consists of a superposition of plane waves with wave-vectors $\vec{k}$ within the solid angle $\delta\Omega$ and frequencies ν within an interval $\delta\nu$, so that the intensity parameter I is replaced with the specific intensity I_ν. Explicitly, for instance the Q_ν parameter is given by

$$Q_\nu = \kappa \left\langle \overline{E}^x E^x - \overline{E}^y E^y \right\rangle, \tag{4.6}$$

where κ is an unspecified proportionality factor with dimension

$$[\kappa] = \frac{1}{[\text{Frequency}][\text{Solid angle}]}$$

to account for the different dimensions of the specific intensity and plane wave Stokes parameters.

This definition, like other similar definitions found in the literature, is not quite general or rigorous enough for the purposes of section 4.3.3, where the specific intensity Stokes parameters are related to a relativistic distribution tensor. To simplify these later arguments, here the defining feature of the specific intensity Stokes parameters $\mathbf{I}_\nu$ is taken to be the relation

$$\int \mathrm{d}\nu \, \mathrm{d}\Omega \, \mathbf{I}_\nu = \frac{1}{8\pi}\mathbf{I} \tag{4.7}$$

to the plane wave parameters $\mathbf{I}$ for a system with a single, monochromatic plane wave. Essentially this means that the specific intensity parameters should be given by $\mathbf{I}_{\nu'}(\vec{n}') = \frac{1}{8\pi}\mathbf{I}\delta(\nu - \nu')\delta(\vec{n} - \vec{n}')$, where $\vec{n}$ is the direction of the wave's propagation and ν its frequency. The factor of $1/8\pi$ relates the energy flux of a plane wave to the Stokes I of equation (3.17). The delta functions have the correct units and appear to correspond to the unspecified constant factor κ of the definition given by Degl'Innocenti and Landolfi (2006) in the limit $\delta\nu, \delta\Omega \to 0$.

It is possible to define fluxes of the Stokes parameters in the same way they are defined for the specific intensity. For example the specific flux of Q_ν is simply $F_{\nu,Q} = \int \mathrm{d}\Omega \, Q_\nu$. The specific intensity Stokes parameters also satisfy a generalized version of the transfer equation (4.2)

$$\frac{\mathrm{d}\mathbf{I}_\nu}{\mathrm{d}s} = \mathbf{J}_\nu - \mathbf{M}_\nu \mathbf{I}_\nu. \tag{4.8}$$

Here the emissivity coefficient j_ν is replaced by a vector of emissivity coefficients $\mathbf{J}_\nu$, which gives the emissivities for each of the Stokes parameters. The absorption coefficient α_ν is replaced by the Mueller matrix $\mathbf{M}_\nu$, which has the general form

$$\mathbf{M} = \begin{pmatrix} \alpha_I & \alpha_Q & \alpha_U & \alpha_V \\ \alpha_Q & \alpha_I & \rho_V & -\rho_U \\ \alpha_U & -\rho_V & \alpha_I & \rho_Q \\ \alpha_V & \rho_U & -\rho_Q & \alpha_I \end{pmatrix}. \tag{4.9}$$

The α terms describe the absorption of each of the Stokes parameters, while the ρ terms describe conversion between different polarization states. This conversion between polarization states is a result of differing propagation velocities of the polarization eigenmodes, as was mentioned in section 3.3.4.

Expressions for the emissivity vector $\mathbf{J}_\nu$ and Mueller matrix $\mathbf{M}_\nu$ can be derived from electrodynamics, typically by assuming that the background matter is homogeneous over large enough distances so that Fourier analysis can be applied (Melrose and McPhedran, 2005). For the emissivity, it is often necessary to also apply quantum theory. When the anisotropic effects causing the difference in propagation velocities of the eigenmodes are sufficiently weak, the components of the Mueller matrix are linear combinations of the components of the polarization tensor $\alpha^i{}_j$, which relates the currents induced in a material to the vector potential of the electromagnetic wave, $J^i = \alpha^i{}_j A^j$. Strong anisotropic effects cause the different polarization eigenmodes to propagate along different trajectories, as discussed in section 3.3.4, in which case the description in terms of Stokes parameters breaks down.

4.2. Kinetic Theory

The theory of radiative transfer can be generalized to general background spacetimes using a formalism based on kinetic theory. This section gives an overview of the relevant parts of relativistic kinetic theory, based on Ehlers (1971). The focus is mostly on the core concepts and an intuitive understanding of the geometry, as thoroughly treating the various technical details is beyond the scope of this work. Other discussions of relativistic kinetic theory include Lindquist (1966) and Sarbach and Zannias (2014).

For simplicity, it is assumed that the system under consideration is a gas of classical particles with well defined positions that move along geodesics. Ehlers (1971) includes also electromagnetic interactions between particles, and generalizations to general Hamil-

tonian systems should also be possible, but these additional complications are not useful for this work. It is not obvious that a description derived for particles applies also to the electromagnetic field, and in general it does not. For example, interference effects of light cannot be explained using classical particles. However, as will be discussed in section 4.3, it does turn out that in a suitable limit electromagnetic radiation behaves like localized massless particles that follow null geodesics, at least where the core quantities of kinetic theory are concerned.

4.2.1. The Phase Space

In non-relativistic kinetic theory the statistical state of a system is described by a distribution function which gives the average particle density at an instant of time in a position-momentum phase space with coordinates $(\vec{x}, \vec{p})$. For a Hamiltonian system the time evolution of a particle's position in this phase space is completely determined from its coordinates, so that knowing the distribution function at one time allows finding the state of the system at all other times, which is a powerful tool.

To generalize this description to be compatible with relativity, the dependence on a preferred time coordinate needs to be removed. This can be done by replacing the description in terms of particle positions at a particular time by their *world lines*, which give the whole history of a particle. A suitable setting for this is a subset of the tangent bundle TM with a metric defined as in section 2.3. The local trivialization is given by (x^μ, p^μ), where x is the spacetime position of the particle and p its four-momentum, which for massive particles is proportional to the four-velocity u as $p = mu$. To be precise, the four-momentum considered here is not in general directly related to the canonical momentum cotangent vector of Hamiltonian mechanics, but this distinction is not important for the special case needed for this work.

Assuming that the particles are not affected by any long range interactions, the world lines on the tangent bundle are horizontal lifts of geodesics, integral curves of the vector field $L = p^\mu e_\mu$. Naturally, the mass m of a particle remains constant. This is also expressed by $L(m^2) = L(g_{\mu\nu}p^\mu p^\nu) = 0$, which can be verified to hold with a direct calculation. As a consequence, particles are confined to the *mass shell*

$$\mathcal{P}_{m,x}M = \Big\{p \in T_xM \mid g_{\mu\nu}p^\mu p^\nu = m^2,\ p^t > 0 \text{ in any local Lorentz frame}\Big\}, \tag{4.10}$$

with the restriction $p^t > 0$ ensuring that particles do not travel backwards in time. The relativistic phase space of a single type of particles is then the bundle of mass shells over

the spacetime

$$\mathcal{P}_m M = \Big\{(x,p) \in TM \mid g_{\mu\nu} p^\mu p^\nu = m^2,\ p^t > 0 \text{ in any local Lorentz frame}\Big\}, \tag{4.11}$$

which is a submanifold of the tangent bundle. These concepts are illustrated in figure 4.1.

This phase space is 7-dimensional, with one of the momentum coordinates being determined by the mass constraint. The vector field L is tangent to the phase space, so it is useful to also define its restriction to the phase space

$$L_m = p^\alpha \left(\frac{\partial}{\partial x^\alpha} - \Gamma^i{}_{\alpha\gamma} p^\gamma \frac{\partial}{\partial p^i} \right), \tag{4.12}$$

with i going over the three independent momentum coordinates. Often it is convenient to choose the p^t coordinate in a Lorentz frame to be determined by the mass shell constraint, so that i goes over the spatial momentum coordinates, but any choice for the set of three independent coordinates works as well.

4.2.2. Volume Forms

On the tangent bundle the natural volume form was given by equation (2.30) as $\boldsymbol{\epsilon} \wedge \boldsymbol{\pi}$, but for defining the distribution function, it is necessary to define volume forms to act as integration measures also on the mass shell $\mathcal{P}_{m,x} M$, phase space $\mathcal{P}_m M$ and most importantly on hypersurfaces pierced by particle worldlines. While the choice of volume forms to use for integration is essentially arbitrary, to be physically meaningful they should only depend on the physical properties of the system, i.e. the particle world line tangent L_m and the momentum p in addition to the tangent bundle volume form. The volume forms should also be conserved along the particle flow, just as is guaranteed by Liouville's theorem in non-relativistic kinetic theory. This property corresponds to a vanishing Lie derivative along L_m.

The volume form $\boldsymbol{\pi}_m$ of the mass shell should obviously be related to the volume form $\boldsymbol{\pi}$ of the tangent space, but making sure that the definition is valid even for $m = 0$ requires some care. A suitable definition turns out to be

$$p_\mu \boldsymbol{\pi}_m = \iota^* \left(\mathrm{i}_{\frac{\partial}{\partial p^\mu}} \boldsymbol{\pi} \right), \tag{4.13}$$

where ι is a map embedding the mass shell in the tangent space. To simplify notation,

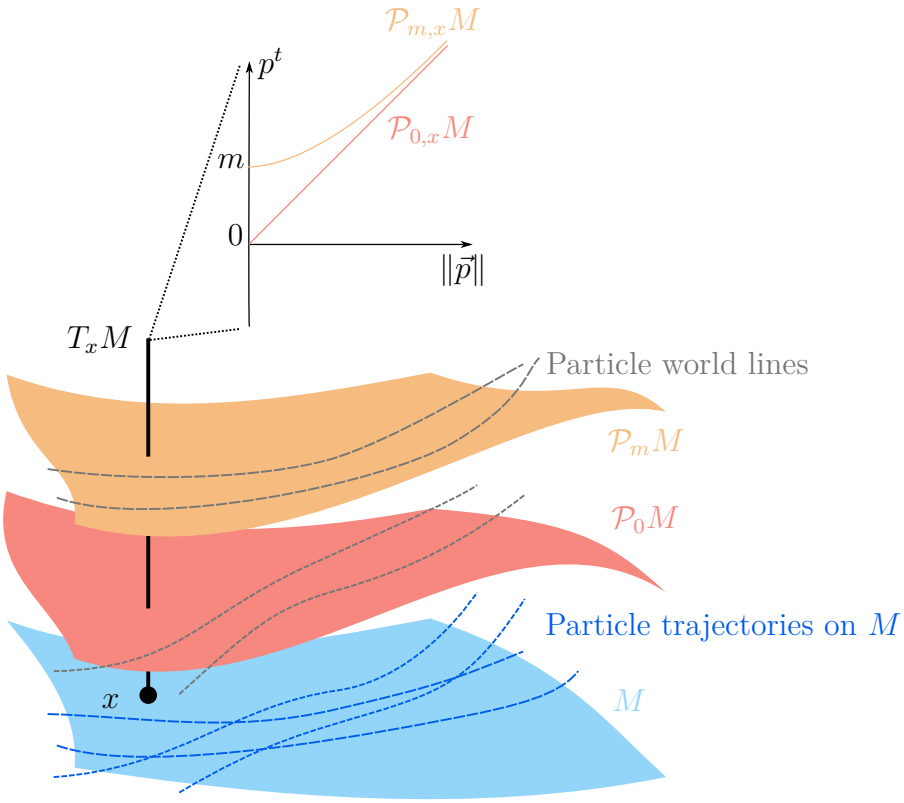

Figure 4.1.: Sketch of the phase spaces and mass shells. While the geodesic trajectories of particles on the spacetime manifold M may cross each other, the corresponding world lines on the tangent bundle TM do not. The phase spaces $\mathcal{P}_mM$ of particles of different mass are separate regions of TM, with the mass shell $\mathcal{P}_{0,x}M$ of massless particles being qualitatively different from the mass shell $\mathcal{P}_{m,x}M$ of massive particles.

explicit pullbacks will be omitted from now on, with the submanifold the forms are defined on being understood from context. In a local Lorentz frame the volume form is given explicitly by

$$\boldsymbol{\pi}_m = \frac{1}{p_t}\mathrm{d}p^x \wedge \mathrm{d}p^y \wedge \mathrm{d}p^z = \frac{\mathrm{d}^3p}{p^t}, \tag{4.14}$$

since $p_t = p^t$. This is easily seen from the definition when p^t is chosen as the coordinate determined by the mass shell condition, but holds of course also for other choices, since $\mathrm{d}p^t = (p^t)^{-1}p_i\mathrm{d}p^i$. The phase space volume form is now given by $\boldsymbol{\Omega}_m = \boldsymbol{\epsilon}\wedge\boldsymbol{\pi}_m$, analogously to the tangent bundle volume form.

The Lie derivative along L_m of the phase space volume form defined this way is

$$\mathcal{L}_{L_m}\boldsymbol{\Omega}_m = \mathrm{d}(\mathrm{i}_{L_m}\boldsymbol{\Omega}_m) + \mathrm{i}_{L_m}\underbrace{\mathrm{d}\boldsymbol{\Omega}_m}_{=0} = \mathrm{d}(\mathrm{i}_{L_m}\boldsymbol{\Omega}_m). \tag{4.15}$$

The last term can be evaluated in a local Lorentz frame, where it can be seen that

$$\mathrm{d}(\mathrm{i}_{L_m}\boldsymbol{\Omega}_m) = \mathrm{d}\left[\left(p^\alpha \mathrm{i}_{\frac{\partial}{\partial x^\alpha}}\boldsymbol{\epsilon}\right)\wedge\boldsymbol{\pi}_m - \boldsymbol{\epsilon}\wedge\left(\Gamma^j{}_{\alpha\beta}p^\alpha p^\beta \mathrm{i}_{\frac{\partial}{\partial p^j}}\boldsymbol{\pi}_m\right)\right] = 0, \tag{4.16}$$

as the derivatives of the metric and the connection coefficients vanish, the coordinates x^α and p^j are independent coordinates of the phase space and any non-zero derivatives cannot contribute due to wedge products with $\boldsymbol{\epsilon}$ and $\boldsymbol{\pi}_m$.

The form $\boldsymbol{\omega}_m = \mathrm{i}_{L_m}\boldsymbol{\Omega}_m$ can be used as a volume form on hypersurfaces pierced by the wordlines of particles. It is also conserved along the particle flow,

$$\mathcal{L}_{L_m}\boldsymbol{\omega}_m = \mathrm{d}\underbrace{(\mathrm{i}_{L_m}\mathrm{i}_{L_m}\boldsymbol{\Omega}_m)}_{=0} + \mathrm{i}_{L_m}\underbrace{\mathrm{d}\boldsymbol{\omega}_m}_{=0} = 0. \tag{4.17}$$

This form plays a central role in defining the distribution function, so it is useful to note that when restricted to a spacelike surface in a local Lorentz frame it reduces to

$$\boldsymbol{\omega}_m = \frac{p^t}{p^t}\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z\wedge\mathrm{d}p^x\wedge\mathrm{d}p^y\wedge\mathrm{d}p^z = \mathrm{d}^3x\,\mathrm{d}^3p\,, \tag{4.18}$$

which is the volume form of the phase space in non-relativistic kinetic theory.

4.2.3. The Distribution Function and the Boltzmann Equation

Now it is possible to define the distribution function. A microstate corresponding to a definite configuration of particle world lines can be completely described by a functional $\Sigma \to N_m(\Sigma)$ counting the number of world lines piercing any oriented 6-dimensional hypersurface Σ in the phase space. The orientation of the surface determines whether a world line is counted positively or negatively, essentially based on whether the particle passes the surface in the inward or outward direction. The statistical state of the system is then described by $\langle N_m\rangle$, the ensemble average of N_m. With some technical but reasonable assumptions, it can be shown that there exists an unique, non-negative function $\mathcal{N}$ on the phase space, which satisfies

$$\langle N_m(\Sigma)\rangle = \int_\Sigma \mathcal{N}\boldsymbol{\omega}_m, \tag{4.19}$$

for any surface Σ. A slight technical problem with this definition is that $\boldsymbol{\omega}_m$ vanishes on 6-dimensional surfaces in the phase space which L_m is tangent to, so that the integral is not properly defined on those surfaces. However, this causes no problems to applying

the definition, as clearly no particle world lines can pierce those surfaces either.

Short range interactions between particles can now be included in the form of collisions. Mathematically these are conveniently modelled in terms of annihilations and creations of particles, with annihilations corresponding to terminating world lines and creations to world lines beginning at some point. Normal collisions where the total particle number is conserved are then modelled as a pair of annihilation and creation events, with the events located in different parts of the mass shell. The total number of newly created particles, i.e. the difference between creations and annihilations in a region D of the phase space, is given given by $\langle N(\partial D)\rangle$. Using Stokes' theorem and other identities from section 2.2

$$\begin{aligned}
\langle N_m(\partial D)\rangle &= \int_{\partial D} \mathcal{N}\boldsymbol{\omega}_m \\
&= \int_D \mathrm{d}(\mathcal{N}\boldsymbol{\omega}_m) \\
&= \int_D \left(\mathrm{d}\mathcal{N}\wedge\boldsymbol{\omega}_m + \mathcal{N}\underbrace{\mathrm{d}\boldsymbol{\omega}_m}_{=0}\right) \\
&= \int_D \mathrm{d}\mathcal{N}\wedge(\mathrm{i}_{L_m}\boldsymbol{\Omega}_m) \\
&= \int_D L_m(\mathcal{N})\boldsymbol{\Omega}_m.
\end{aligned}$$

This can be intuitively understood with the geometrical interpretation of differential forms, illustrated in figure 4.2. The form $\mathcal{N}\boldsymbol{\omega}_m$ corresponds to a density of lines along the vector field L_m, which is just the average density of world lines. Its exterior derivative $\mathrm{d}(\mathcal{N}\boldsymbol{\omega}_m)$ is then interpreted as the density of world lines beginning at a point, with terminating lines counted negatively. These naturally correspond to creation and annihilation events, so $\mathrm{d}(\mathcal{N}\boldsymbol{\omega}_m) = L_m(\mathcal{N})\boldsymbol{\Omega}_m$ measures the total density of collision events. As $L_m(\mathcal{N}) = \frac{\mathrm{d}\mathcal{N}}{\mathrm{d}\lambda}$, where λ is the affine parameter of geodesics with p as their tangent, this can formally summarized in the equation

$$L_m(\mathcal{N}) = \frac{\mathrm{d}\mathcal{N}}{\mathrm{d}\lambda} = \left(\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}\lambda}\right)_{\mathrm{coll}}, \tag{4.20}$$

the relativistic form of the Boltzmann equation, where $\left(\frac{\mathrm{d}\mathcal{N}}{\mathrm{d}\lambda}\right)_{\mathrm{coll}}$ is the collision density per unit affine parameter.

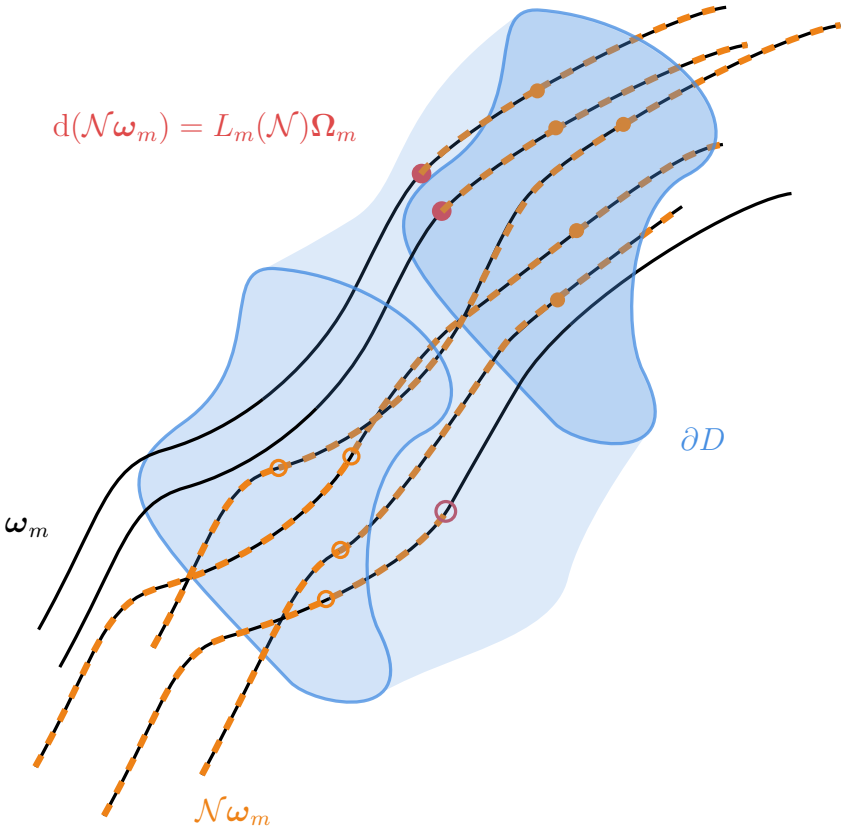

Figure 4.2.: Sketch of the correspondence between particle world lines and differential forms. The form $\boldsymbol{\omega}_m$ is identified with a density of tubes or lines along the particle flow, indicated as black lines. Since $\mathrm{d}\boldsymbol{\omega}_m = 0$, the lines do not end. The form $\mathcal{N}\boldsymbol{\omega}_m$ describes the average density of particle world lines, indicated with orange dashed lines. The pull-back of $\mathcal{N}\boldsymbol{\omega}_m$ onto ∂D is indicated with orange circles, with filled circles counted positively and open circles negatively. Collision events can cause particle trajectories to end, and as result $\mathrm{d}(\mathcal{N}\boldsymbol{\omega}_m) \neq 0$. The points corresponding to non-zero $\mathrm{d}(\mathcal{N}\boldsymbol{\omega}_m)$ are indicated with red circles. The total sums of the orange and red circles are equal, corresponding to Stokes' theorem. In this particular example $\langle N_m(\partial D)\rangle = \int_{\partial D} \mathcal{N}\boldsymbol{\omega}_m = 5 - 4 = 1$, so the total number of newly created particles is 1, as can be verified directly counting the creations and annihilations $\int_D L_m(\mathcal{N})\boldsymbol{\Omega}_m = 2 - 1 = 1$.

4.3. Relativistic Radiative Transfer

The specific intensity I_ν is an observer dependent quantity, just as energy, so it cannot be used as a basis for the relativistic description of radiative transfer. Instead, with some assumptions the radiation field can be described using kinetic theory. In this kinetic description the specific intensity is replaced with an observer independent distribution function on the tangent bundle, which is generalized to a distribution tensor to account for polarization. The relativistic generalization of the transfer equation (4.8) is then obtained from the Boltzmann equation with a suitable collision term.

4.3.1. The Distribution Tensor

Kinetic theory is an attractive formalism for the description of radiation, due to the similarities between the radiative transfer equation (4.8) and the Boltzmann equation (4.20), as well as between particle world lines and rays of light in the geometric optics limit. However, the fundamentally different natures of idealized point particles and the electromagnetic field poses a problem when applying kinetic theory to electromagnetic radiation, namely the existence and interpretation of the distribution function.

In the case of massive particles, the distribution function has a simple interpretation as a number density of particles on surfaces in the phase space. However, electromagnetic radiation does not easily admit such an interpretation, as it does not consist of localizable particles. Often it is nevertheless assumed that the distribution function $\mathcal{N}$ gives the number density of localizable, Newtonian photons that follow the rays of light in the geometric optics limit (Lindquist, 1966; Misner et al., 1973). The amount of these photons can even be related to the squared amplitude a^2, as it is proportional to the energy density of the wave, in which case the conservation equation (3.72) is just the equation for conservation of the particle number. This approach appears intuitive and gives the correct result in the limit where the kinetic description is valid, but it is not obvious what that limit is.

Instead, it is possible to define a distribution function in terms of the electromagnetic field variables, and to show that under suitable conditions it behaves like the distribution function of localizable particles. This definition also naturally extends the scalar distribution function to a distribution tensor, which includes the polarization of the radiation. Doing this in general spacetimes is quite complicated, so here it is taken as given that such a construction is possible, with the full details being discussed by Bildhauer (1989a,b). However, to make the basic idea behind the distribution tensor

clearer, it is illustrated here in the special case of flat spacetime and plane waves. This is also useful for connecting the general formalism to the one used in section 4.1.

The most important property of the distribution function is that it gives a statistical description of state of the system, which allows ignoring details like the exact positions of particles. In terms of the field quantities, a similar statistical description is given by the two-point correlation tensor

$$B_{\mu\nu}(x_1, x_2) = \left\langle \overline{A}_\mu(x_1) A_\nu(x_2) \right\rangle, \tag{4.21}$$

where $\langle\rangle$ is an ensemble average over the state of the field. Now it is possible to define a distribution tensor on the whole tangent bundle by using a Wigner transform

$$N_{\mu\nu}(x, k) = -\frac{1}{4(2\pi)^5} \int \mathrm{d}^4 y \, e^{ik^\alpha y_\alpha} B_{\mu\nu}\left(x + \frac{1}{2}y, x - \frac{1}{2}y\right), \tag{4.22}$$

where k is a tangent vector. When this is generalized to curved spacetimes, a necessary assumption is that the correlation length of the system is much shorter than the curvature length scales of the system. This assumption, similar to the $\varepsilon \to 0$ limit of geometric optics, allows defining the correlation tensor B and its integral transform in an unambiguous manner using a suitable parallel transport operator, as the correlation tensor is non-zero only over a small region.

Applying the definition (4.22) to a plane wave

$$A^\mu(x) = a^\mu e^{ik^\mu x_\mu}, \tag{4.23}$$

yields the distribution tensor

$$\begin{aligned} N_{\mu\nu}(x, q) &= -\frac{1}{4(2\pi)^5} \int \mathrm{d}^4 y \, e^{iq^\alpha y_\alpha} \left\langle \overline{a}_\mu a_\nu \right\rangle e^{ik^\alpha y_\alpha} \\ &= -\frac{1}{8\pi} \left\langle \overline{a}_\mu a_\nu \right\rangle \delta^{(4)}(k - q), \end{aligned} \tag{4.24}$$

where δ is the Dirac delta function. This can be easily generalized to an arbitrary superposition of uncorrelated plane waves, as the uncorrelated cross terms vanish when taking the ensemble average.

The distribution tensor can be reduced to a scalar distribution function $N = N^\mu{}_\mu$. In the case of a single plane wave this has the expression

$$N(x, q) = N_\mu{}^\mu(x, q) = -\frac{1}{8\pi} \left\langle \mathsf{a}^2 \overline{f}_\mu f^\mu \right\rangle \delta^{(4)}(k - q) = \frac{1}{8\pi} \mathsf{a}^2 \delta^{(4)}(k - q), \tag{4.25}$$

where the decomposition into a polarization vector and a scalar amplitude, given by equation (3.67), was used. Here the scalar function N can be interpreted as describing the total amplitude of the wave, while the distribution tensor simply decomposes this amplitude into the different polarization states. When there is just a single wave with definite polarization, the distribution tensor can be written in terms of the scalar distribution and the polarization vectors as $N_{\mu\nu} = -N\overline{f}_\mu f_\nu$. However, in general the ensemble average does not decompose like this, allowing for partially polarized radiation.

In the geometric optics limit, the distribution tensor is non-zero only when $k^\mu k_\mu = 0$. To better match the description used in kinetic theory, the distribution tensor can therefore be restricted to the phase space $\mathcal{P}_0 M$, the future light cone. Since k is a null vector, this can be formally done by setting $N_{\mu\nu}(x,k) = 2\delta(k^\mu k_\mu)\mathcal{N}_{\mu\nu}(x,k)$, where $\mathcal{N}_{\mu\nu}(x,k)$ is defined only on $\mathcal{P}_0 M$. This can be seen by calculating

$$\begin{aligned}\int\limits_{T_xM,k^t>0} 2h\delta(k^\mu k_\mu)\boldsymbol{\pi} &= \int\limits_{T_xM,k^t>0} h\frac{2}{2k^t}\Big(\delta(k^t - \left|\vec{k}\right|) - \delta(k^t + \left|\vec{k}\right|)\Big)\,\mathrm{d}^4k \\ &= \int\limits_{\mathcal{P}_{0,x}M} h\frac{\mathrm{d}^3k}{k^t} = \int\limits_{\mathcal{P}_{0,x}M} h\boldsymbol{\pi}_m,\end{aligned} \tag{4.26}$$

where a local Lorentz frame is assumed for the middle equalities, and h is an arbitrary function. The scalar function N corresponds to a scalar function $\mathcal{N}$ on $\mathcal{P}_0 M$ in a similar manner.

4.3.2. Transfer Equation for the Distribution Tensor

By applying the geometric optics expansion (3.53) to the definition of the scalar distribution function of the electromagnetic field, Bildhauer (1989a) shows that in the vacuum geometric optics limit with vanishing correlation length it satisfies the equation

$$L_0(\mathcal{N}) = k^\mu\left(\frac{\partial}{\partial x^\mu} - \Gamma^i{}_{\mu\nu}k^\nu\frac{\partial}{\partial k^i}\right)\mathcal{N} = 0, \tag{4.27}$$

which is just the collisionless Boltzmann equation (4.20). As the scalar function is related to the distribution tensor as $\mathcal{N} = \mathcal{N}^\mu{}_\mu$, and since the polarization vector is parallel transported along rays, it is natural to expect the generalization to the case of the distribution tensor to be

$$k^\mu \mathrm{D}_\mu \mathcal{N}^{\rho\sigma} = k^\mu\left(\nabla_\mu - \Gamma^i{}_{\mu\nu}k^\nu\frac{\partial}{\partial k^i}\right)\mathcal{N}^{\rho\sigma} = 0, \tag{4.28}$$

where the operator horizontal covariant derivative D_μ has been introduced for convenience. Again, Bildhauer (1989a) shows that this equation is valid in the appropriate limit, with small corrections proportional to the Riemann tensor appearing when the geometric optics expansion parameter $\varepsilon \neq 0$.

An intuitive reason for why the Boltzmann equation, which assumes localizable particles, holds also for electromagnetic radiation is that the assumption of vanishing correlation length effectively forces the field to be describable in terms of well localized wave packets that cannot interfere with each other. In the geometric optics limit these wave packets follow the null geodesic rays, effectively behaving like massless particles and making the kinetic description in terms of the distribution function and Boltzmann equation valid. This is similar to how classical particles emerge from a quantum mechanical description in terms of a wave function in the appropriate limit. The assumption of vanishing correlation length should hold quite generally when the geometric optics limit is valid, as most sources of radiation are highly random at small scales. However, there appear to be cases where it is at least in principle possible for the geometric optics approximation to remain valid while the kinetic description fails due to long range correlations. One such contrived example is interference in an asymptotically planar and coherent wave which is bent onto itself due to gravitational lensing.

For the purposes of radiative transfer, interaction with matter needs to be included. Motivated by the correspondence with kinetic theory, the processes of absorption, emission and scattering of radiation can be interpreted as annihilation, creation and collisions of particles, so adding a general collision term to describe these processes seems natural. These effects must be small for the underlying approximations to remain valid, so the collision term has to be linear in $\mathcal{N}^{\alpha\beta}$, giving an equation analogous to the non-relativistic transfer equation (4.8)

$$k^\mu \mathrm{D}_\mu \mathcal{N}^{\alpha\beta} = J^{\alpha\beta} + H^{\alpha\beta\gamma\delta}\mathcal{N}_{\gamma\delta}\,. \tag{4.29}$$

Here $J^{\alpha\beta}$ describes emissive processes that do not depend on the distribution tensor, in analogy with the emissivity vector $\mathbf{J}_\nu$ of equation (4.8), while $H^{\alpha\beta\gamma\delta}$ corresponds to the processes described by the Mueller matrix $\mathbf{M}_\nu$. They are assumed to depend only on the local properties of the background matter, and they generalize the modifications to the propagation equation for the wave amplitude discussed in section 3.3.4. As was the case for the non-relativistic equation, here it would also be possible to include a generic scattering term of the form

$$\int S^{\alpha\beta\gamma\delta}\mathcal{N}_{\gamma\delta}\,\boldsymbol{\pi}_0, \tag{4.30}$$

where S gives the scattering amplitude between different parts of the fibre or different wave vectors, but for simplicity it can again be considered to be included in $J^{\alpha\beta}$.

In principle it is possible to derive expressions for the tensors J and H from electrodynamics by assuming suitable properties for the background matter. Analogously to the non-relativistic case mentioned at the end of section 4.1, Gammie and Leung (2012) find an expression for H in terms of the polarization tensor α by adding a current term of the form $j^\mu = \alpha^\mu{}_\nu a^\nu$ to the scalar amplitude propagation equation (3.71) which is of order ε^{-1} in the geometric optics expansion. However, it is not clear under which conditions this can be done, as the interaction between matter and electromagnetic radiation decreases for high frequencies, which would suggest that the polarization tensor is actually of the order $\varepsilon \propto \nu^{-1}$. This means that some additional assumptions are needed compared to the case of flat spacetime, but Gammie and Leung (2012) do not make clear what these assumptions are. This problem can be avoided to some extent with an argument based on the equivalence principle. When the kinetic description is valid, i.e. when the corrections to the geometric optics approximation can be ignored, space appears flat and matter homogeneous over any correlated regions of the radiation field. This means that the interaction of radiation and matter can be locally described using the non-relativistic results, and the equivalence principle then allows taking these to hold in general. In section 4.3.4 this will be used to allow basing computations on known expressions for the emissivity vector $\mathbf{J}_\nu$ and Mueller matrix $\mathbf{M}_\nu$.

A general derivation would nevertheless be useful in quantifying the error caused by all the necessary approximations, as well as making sure that they are internally consistent. For example, the effects contributing to J and H are related to the modifications to ray propagation discussed in section 3.3.4, and as a result the conditions under which it is reasonable to include the effects on the propagation of the distribution tensor while ignoring the modifications to the ray path are not completely clear. Including refractive effects in the formalism would allow quantifying this error as well as correctly handling regions where refractive effects cannot be ignored. Broderick and Blandford (2003, 2004) add refractive effects to the formalism discussed in the following sections, but while their approach appears plausible, their treatment uses several approximations and does not seem to be completely general or rigorous. In general, correctly handling the differing ray paths of the polarization eigenmodes as well as conversion between changing modes along the rays seems to be a difficult problem.

Another issue that would be clarified by a general derivation are the exact approximations required to arrive at the linear and local collision term used here, although

assuming that the material properties that determine J and H are constant over distances much larger than the correlation length is likely to be enough. A rigorous derivation of the source terms and the inclusion of refractive effects would require extending the derivation of the transfer equation by Bildhauer to include the response of the background matter, but it appears that this has not been done. However, the non-relativistic case of waves in an anisotropic, refractive medium has been treated with some generality using similar methods (Lau and Watson, 1970; Ryzhik et al., 1996), and a relativistic generalization of these treatments should be possible.

4.3.3. The Invariant Stokes Parameters

The complex distribution tensor $\mathcal{N}^{\gamma\delta}(x,k)$ is Hermitian by construction, corresponding to 16 degrees of freedom. However, only four of these degrees of freedom are physical, corresponding to the four Stokes parameters. The other degrees of freedom correspond to the gauge freedom of the potential, which is not physically observable. Extracting the physical degrees of freedom allows reducing the computational load of numerically solving the transfer equation, as well as making the connection to the non-relativistic transfer equation clear.

While there are various ways to argue that $I_\nu \propto \nu^3\mathcal{N}$ (Misner et al., 1973, section 22.6 contains one based on counting Newtonian photons), these do not seem to naturally generalize to rest of the specific Stokes parameters. Instead, it can be shown that when a ray propagates along the z-axis in a local Lorentz frame the specific intensity Stokes parameters can be defined to be

$$I_\nu = -16\pi^4\nu^3(\mathcal{N}^{xx} + \mathcal{N}^{yy}) \tag{4.31}$$

$$Q_\nu = -16\pi^4\nu^3(\mathcal{N}^{xx} - \mathcal{N}^{yy}) \tag{4.32}$$

$$U_\nu = -16\pi^4\nu^3(\mathcal{N}^{xy} + \mathcal{N}^{yx}) \tag{4.33}$$

$$V_\nu = i16\pi^4\nu^3(\mathcal{N}^{xy} - \mathcal{N}^{yx}). \tag{4.34}$$

This can be done by showing that they satisfy the condition (4.7) when applied to the distribution tensor of a plane wave (4.24). From the expression for the electric field (3.76)

or (3.77) it is easy to see that the Stokes parameters (3.17)–(3.20) of a plane wave are

$$I = 4\pi^2\nu^2 \left\langle \overline{a}^x a^x + \overline{a}^y a^y \right\rangle \tag{4.35}$$

$$Q = 4\pi^2\nu^2 \left\langle \overline{a}^x a^x - \overline{a}^y a^y \right\rangle \tag{4.36}$$

$$U = 4\pi^2\nu^2 \left\langle \overline{a}^x a^y + \overline{a}^y a^x \right\rangle \tag{4.37}$$

$$V = -i4\pi^2\nu^2 \left\langle \overline{a}^x a^y - \overline{a}^y a^x \right\rangle . \tag{4.38}$$

Noting that $\mathrm{d}\nu\,\mathrm{d}\Omega = \nu^{-1}(2\pi)^{-2}\,\mathrm{d}^3k\,/k^t$, a direct computation gives

$$\begin{aligned}
\int \mathrm{d}\nu'\,\mathrm{d}\Omega\, I_{\nu'} &= -4\pi^2 \int \frac{\mathrm{d}^3k'}{k'^t} \nu'^2 (\mathcal{N}^{xx} + \mathcal{N}^{yy}) \\
&= -4\pi^2 \int \mathrm{d}^4k'\, \nu'^2 (N^{xx} + N^{yy}) \\
&= \frac{1}{8\pi} \int \mathrm{d}^4k\, (k'^t)^2 \left\langle \overline{a}^x a^x + \overline{a}^y a^y \right\rangle \delta^{(4)}(k - k') \\
&= \frac{1}{8\pi} 4\pi^2\nu^2 \left\langle \overline{a}^x a^x + \overline{a}^y a^y \right\rangle \\
&= \frac{1}{8\pi} I.
\end{aligned} \tag{4.39}$$

This holds with trivial changes for the other parameters, so equations (4.31) – (4.34) satisfy the requirements set on the specific intensity Stokes parameters by equation (4.7).

The relation between the specific intensity Stokes parameters and the distribution tensor depends on an arbitrary choice of local Lorentz frame, which is quite inconvenient for practical applications. A set of parameters that depends only on the choice of orientation of the x, y-axes is given by

$$\mathcal{I} = \frac{\mathbf{I}_\nu}{\nu^3} = \nu^{-3} \begin{pmatrix} I_\nu \\ Q_\nu \\ U_\nu \\ V_\nu \end{pmatrix}. \tag{4.40}$$

These parameters are the *invariant Stokes parameters*, and up to a constant factor they are given by the x, y components of the distribution tensor. It should be kept in mind that the values of $\mathcal{I}$ depend on the position on the tangent bundle, given by the spatial point x and the wave vector k, as well as the choice of basis vectors e_x, e_y defining the x, y-axes. These explicit dependencies are omitted from the notation in order to simplify it.

For showing that the invariant Stokes parameters do not depend on the choice of observer, it is useful to define a projection operator

$$P^{\mu}{}_{\nu}(u,k) = \delta^{\mu}{}_{\nu} - u^{\mu}u_{\nu} + w^{\mu}w_{\nu}, \tag{4.41}$$

where u is the four-velocity of an observer, and

$$w^{\mu} = \frac{k^{\mu}}{k^{\nu}u_{\nu}} - u^{\mu} \tag{4.42}$$

is the spatial part of k as seen by the observer. This operator projects vectors orthogonal to both the observers four-velocity and the wave vector, and is known as the *screen projection* operator (Gammie and Leung, 2012). Let $e_i{}^{\mu}$, $i = x, y$, be the x, y basis vectors of some observer, satisfying $e_i{}^{\mu}u_{\mu} = e_i{}^{\mu}k_{\mu} = 0$. The x, y components of the distribution tensor as seen by this observer are

$$\mathcal{N}_{ij} = \mathcal{N}_{\mu\nu}\, e_i{}^{\mu} e_j{}^{\nu}. \tag{4.43}$$

Note that the i, j-indices can be raised with the Minkowski metric, so that $\mathcal{N}_{ij} = \mathcal{N}^{ij}$.

Another observer with four-velocity u' can define their basis vectors by

$$e'_i{}^{\mu} = P^{\mu}{}_{\nu}(u', k) e_i{}^{\nu}, \tag{4.44}$$

as a straightforward computation shows that the basis vector defined this way are still orthonormal. In this case the projection operator acts more like a rotation. Now the components of the distribution tensor are

$$\begin{aligned} \mathcal{N}'_{ij} &= \mathcal{N}_{\mu\nu}\, e'_i{}^{\mu} e'_j{}^{\nu} \\ &= \mathcal{N}_{\mu\nu}\, P^{\mu}{}_{\alpha} P^{\nu}{}_{\beta}\, e_i{}^{\alpha} e_j{}^{\beta}. \end{aligned} \tag{4.45}$$

The gauge condition $a^{\mu}k_{\mu} = 0$ leads to $\mathcal{N}_{\mu\nu}\, k^{\mu} = 0$, which in turn gives

$$\mathcal{N}_{\mu\nu}\, P^{\mu}{}_{\alpha}(u,k) = \mathcal{N}_{\mu\nu}\left(\delta^{\mu}{}_{\alpha} + \frac{u^{\mu}k_{\alpha}}{k^{\beta}u_{\beta}}\right), \tag{4.46}$$

so the final result is

$$\begin{aligned} \mathcal{N}'_{ij} &= \mathcal{N}_{\mu\nu}\left(\delta^{\mu}{}_{\alpha} + \frac{u^{\mu}k_{\alpha}}{k^{\gamma}u_{\gamma}}\right)\left(\delta^{\nu}{}_{\beta} + \frac{u^{\nu}k_{\beta}}{k^{\gamma}u_{\gamma}}\right) e_i{}^{\alpha} e_j{}^{\beta} \\ &= \mathcal{N}_{\alpha\beta}\, e_i{}^{\alpha} e_j{}^{\beta} = \mathcal{N}_{ij}\,. \end{aligned} \tag{4.47}$$

This means that the x, y components of the distribution tensor, and therefore the invariant Stokes parameters, agree for all observers who define their basis vectors according to equation (4.44).

4.3.4. Transfer Equation for the Invariant Stokes Parameters

It is possible to derive a transport equation for $\mathcal{I}$ by defining the e_x, e_y basis vectors along the ray. The simplest equation is given when the basis vectors ${e_i}^\mu$ are parallel transported along the ray, $k^\mu \nabla_\mu {e_i}^\nu = 0$. The components $\mathcal{N}_{ij}$ given by (4.43) in a frame defined this way are scalars, and using equation (4.29) their transfer equation is

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{N}_{ij}}{\mathrm{d}\lambda} &= \frac{\mathrm{D}\mathcal{N}_{ij}}{\mathrm{d}\lambda} \\
&= k^\alpha \mathrm{D}_\alpha \left(\mathcal{N}_{\mu\nu} {e_i}^\mu {e_j}^\nu \right) \\
&= {e_i}^\mu {e_j}^\nu k^\alpha \mathrm{D}_\alpha \mathcal{N}_{\mu\nu} \\
&= {e_i}^\mu {e_j}^\nu \left(J_{\mu\nu} + {H_{\mu\nu}}^{\alpha\beta} \mathcal{N}_{\alpha\beta} \right) \\
&= J_{ij} + {H_{ij}}^{\alpha\beta} \mathcal{N}_{\alpha\beta} \,.
\end{aligned} \tag{4.48}$$

Inserting the definition of the invariant Stokes parameters $\mathcal{I}$, it follows that they are transported along a ray according to

$$\frac{\mathrm{d}\mathcal{I}}{\mathrm{d}\lambda} = \mathcal{J} - \mathcal{M}\mathcal{I}, \tag{4.49}$$

where the vector $\mathcal{J}$ and matrix $\mathcal{M}$ are derived from the components of J and H in the parallel transported basis. The explicit relation between the matrices and tensors is relatively straightforward to work out (Gammie and Leung, 2012), but it is not needed here. Both $\mathcal{J}$ and $\mathcal{M}$ are invariant between different observers with aligned basis vectors, just as $\mathcal{I}$ is.

Equation (4.49) is the relativistic generalization of the transfer equation (4.8). While there are obviously multiple alternate but equivalent formulations of the relativistic transfer equation in terms of the Stokes parameters or some other parametrization of polarization state, equation (4.49) appears to be a fairly convenient formulation, and is used by multiple authors (e.g. Shcherbakov and Huang, 2011; Dexter, 2016). The relation between the relativistic and non-relativistic emission and absorption can be found by considering the equation in flat spacetime with a homogeneously moving background fluid. The dependence between the affine parameter λ and spatial distance s can be easily

found in the fluid rest frame with the z-axis aligned with the ray, where

$$\frac{\mathrm{d}s}{\mathrm{d}\lambda} = k^z = C\nu. \tag{4.50}$$

Here the affine parameter λ is scaled by an arbitrary factor C for generality, with k now a scaled version of the wave vector. The arbitrary factor C is convenient for practical applications, as it can be used to handle for instance different systems of units.

The relation (4.50) also holds in general spacetimes, as it uses only local quantities in a Lorentz frame. In flat spacetime the components of k are constant, so the frequency ν is constant along the ray and equation (4.49) can be written as

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{I}}{\mathrm{d}\lambda} &= \mathcal{J} - \mathcal{M}\mathcal{I} \\
\nu^{-3}\frac{\mathrm{d}s}{\mathrm{d}\lambda}\frac{\mathrm{d}\mathbf{I}_\nu}{\mathrm{d}s} &= \mathcal{J} - \nu^{-3}\mathcal{M}\mathbf{I}_\nu \\
\frac{\mathrm{d}\mathbf{I}_\nu}{\mathrm{d}s} &= C^{-1}\left(\nu^2\mathcal{J} - \nu^{-1}\mathcal{M}\mathbf{I}_\nu\right) \\
&= \mathbf{J}_\nu - \mathbf{M}_\nu\mathbf{I}_\nu.
\end{aligned}$$

Therefore the invariant versions of the emissivity vector and the Mueller matrix are given as

$$\mathcal{J} = C\nu^{-2}\mathbf{J}_\nu \tag{4.51}$$

$$\mathcal{M} = C\nu\mathbf{M}_\nu. \tag{4.52}$$

These are local quantities, so by the equivalence principle these equalities will continue to hold in arbitrary spacetimes and with arbitrary fluid motion, allowing the invariant quantities to be found from values computed in the fluid rest frame.

4.3.5. Rotation of the Emissivity and Mueller Matrix

A technical problem in computing the invariant $\mathcal{J}$ and $\mathcal{M}$ from known expressions for $\mathbf{J}_\nu$ and $\mathbf{M}_\nu$ is that they are usually given in some preferred basis, which in general does not coincide with the parallel transported basis. The following procedure for performing the transition from the preferred basis to the parallel transported basis is adapted from the one given by Shcherbakov and Huang (2011).

Let the preferred basis be a local Lorentz frame with unit basis vectors a_μ, $\mu = t, x, y, z$, where the timelike vector a_t is typically given by the fluid's four-velocity u,

i.e. $a_t = u$, so that the frame is the rest frame of the fluid. These and the various other basis vectors discussed in the following are illustrated in figure 4.3. A standard way to align this frame with respect to a preferred direction, which is typically the direction of an external magnetic field $\vec{B}_0$ as seen in the frame, is to set $\vec{a}_z$ along $\vec{B}_0$ and then set $\vec{a}_x = \vec{a}_z \times \vec{k}$, $\vec{a}_y = \vec{k} \times \vec{a}_x$. Here the three-vectors are the spatial parts of the corresponding four-vectors, i.e. in this frame $(k^\mu) = \left(k^t, \vec{k}\right)$. Now the natural basis vectors for the polarization quantities are $\vec{b}_x = \vec{a}_x$ and $\vec{b}_y = \vec{a}_y \cos\theta - \vec{a}_z \sin\theta$, where θ is the angle between $\vec{a}_z$ and $\vec{k}$. In the b basis, the equations for $\mathbf{J}_\nu$ and $\mathbf{M}_\nu$ depend on the ray tangent k only through the frequency ν and the angle θ, simplifying them considerably.

The parallel transported basis vectors e_x and e_y can in general have components along a_t, but as was shown previously the relevant quantities are invariant when the basis vectors are replaced with $P^\mu{}_\nu(e_t, k)e_i^\nu$, where P is the projection operator defined in equation (4.41). The projected parallel transported basis vectors $\vec{e}_x, \vec{e}_y$, which are now purely spatial, and the polarization basis of the frame are then related by a rotation by an angle χ. From the geometry of the situation it can be seen that

$$\sin\chi = -\vec{b}_y \cdot \vec{e}_x = \vec{e}_x \cdot \vec{a}_{z\perp}/\|\vec{a}_{z\perp}\| \tag{4.53}$$

$$\cos\chi = \vec{b}_y \cdot \vec{e}_y = -\vec{e}_y \cdot \vec{a}_{z\perp}/\|\vec{a}_{z\perp}\| \tag{4.54}$$

$$\vec{a}_{z\perp} = \vec{a}_z - \left(\vec{k} \cdot \vec{a}_z\right)\vec{k}. \tag{4.55}$$

In terms of four-vectors these can be conveniently expressed using

$$\vec{e}_i \cdot \vec{a}_{z\perp} = -g_{\mu\alpha}P^\mu{}_\nu(a_t, k)e_i^\nu P^\alpha{}_\beta(a_t, k)a_z^\beta = -e_i^\mu a_{z\mu} - \left(a_t{}^\mu e_{i\mu}\right)\frac{k_\mu a_z{}^\mu}{k_\mu a_t{}^\mu}. \tag{4.56}$$

Now $\mathcal{J}$ and $\mathcal{M}$ can be calculated from the known $\mathbf{J}_\nu$ and $\mathbf{M}_\nu$ as

$$\mathcal{J} = C\nu^{-2}\mathbf{R}(\chi)\mathbf{J}_\nu \tag{4.57}$$

$$\mathcal{M} = C\nu\mathbf{R}(\chi)\mathbf{M}_\nu\mathbf{R}(-\chi), \tag{4.58}$$

where C is a constant fixed by the scaling of k with $C\nu = a_t{}^\mu k_\mu$. The angle θ that is needed for computing $\mathbf{J}_\nu$ and $\mathbf{M}_\nu$ can likewise be found from

$$\cos\theta = \vec{k} \cdot \vec{a}_z = -k_\mu a_z{}^\mu. \tag{4.59}$$

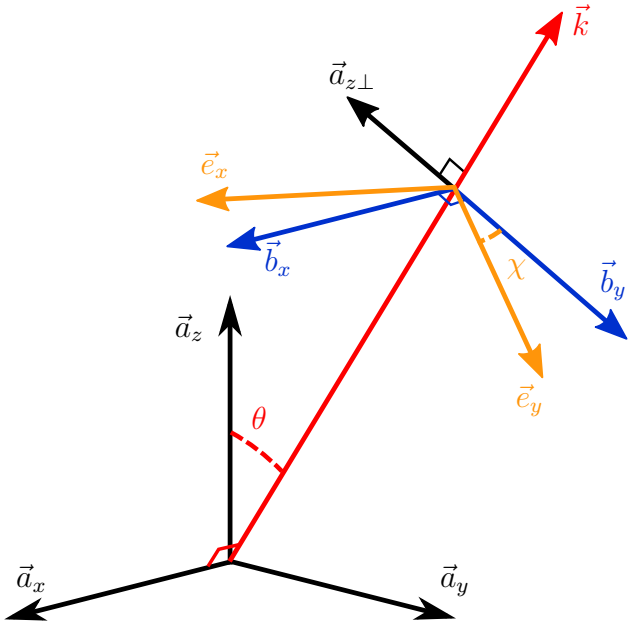

Figure 4.3.: The different basis vectors used for defining the polarization axes.

4.4. Ray-Tracing

In the most general case with scattering, the radiative transfer equation (4.29) or (4.49) can only be solved by solving the equation simultaneously in the whole phase space, as scattering couples together rays with different wave vectors. Often the effects of scattering can be ignored, in which case the radiative transfer equation is a linear differential equation along a ray. In this case it is possible to compute the observed image of a radiation source by using ray-tracing.

There are various slightly different approaches to applying ray-tracing, but the core principle behind them all is intuitive: solve the paths of the rays arriving at the observer and solve the radiative transfer equation along them. The following procedure will be applied in the subsequent chapters of this work, and is suited to solving the transfer equation for the invariant Stokes parameters (4.49) without any assumptions about the underlying spacetime. The main features of this procedure are illustrated in figure 4.4.

The ray-tracing procedure begins by constructing an image plane, which models an observational instrument, and determining the directions of the rays arriving at this image plane. Next, the image plane is divided into pixels, and for the ray arriving at each pixel the following steps are performed:

1. Given the initial tangent vector of the arriving ray, solve the geodesic equation backwards in time along the ray to determine the ray's path until it is terminated by hitting some object or is determined to escape to infinity.

2. Fix the basis vectors e_x, e_y used for defining the Stokes parameters in the image plane and parallel transport them along the ray to define the basis used for the radiative transfer calculation.

3. Set the initial values of the invariant Stokes parameters in the parallel transported basis depending on whether the ray terminates at a radiating surface, such as a star, or in vacuum.

4. Solve the radiative transfer equation (4.49) forwards along the ray to the observer, using the parallel transported basis to find the invariant emissivity and Mueller matrix as described in section 4.3.5.

After these steps, the invariant Stokes parameters arriving at each pixel of the image plane are known, and other quantities of interest may be computed from them. The details of implementing these steps numerically are described in the next chapter.

Alternative procedures include fixing the basis vectors e_i globally based on a preferred direction such as the magnetic field (Broderick and Blandford, 2003, 2004) or solving the transfer equation of the distribution tensor (suggested by Gammie and Leung, 2012). It is also possible to approximate scattering effects with ray-tracing by using Monte Carlo methods (e.g. Connors et al., 1980), and depending on the system under consideration this may be crucial for getting correct results (e.g. Schnittman and Krolik, 2009).

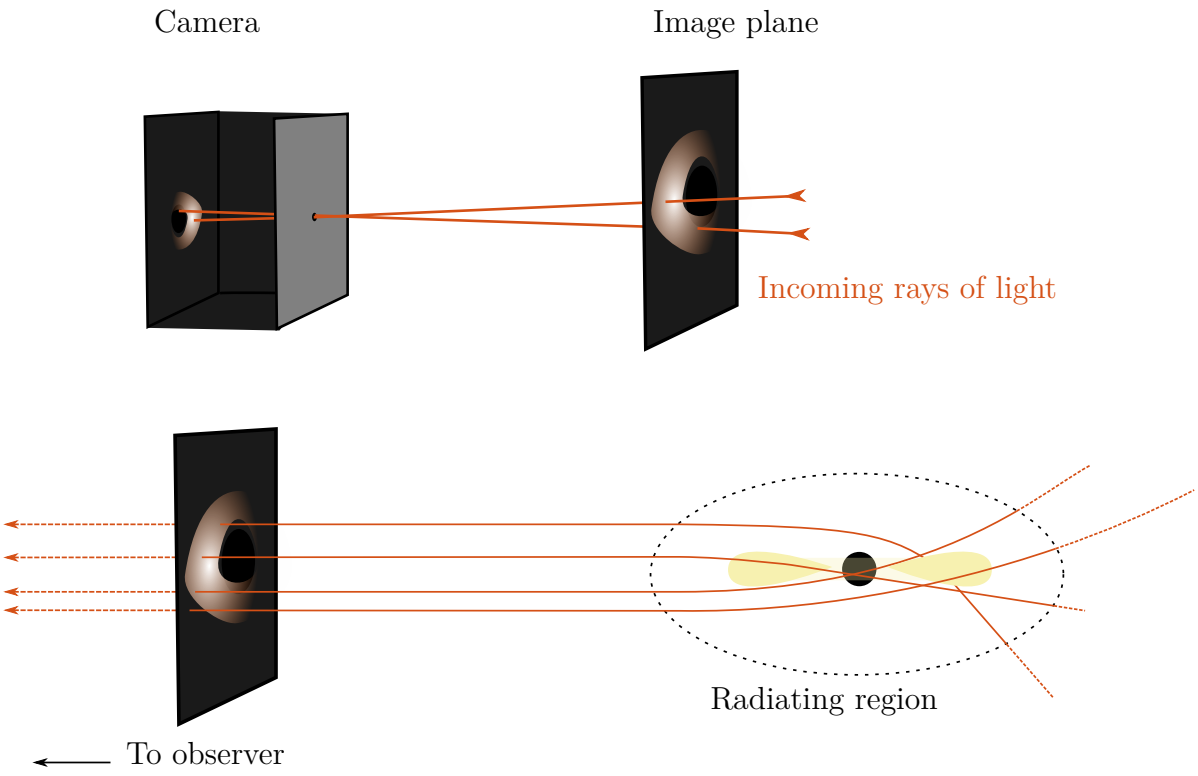

Figure 4.4.: A simple model of an observational instrument is the pinhole camera, pictured at the top. Although the actual image is formed inside the camera, the same image is also formed on an imaginary plane in front of the camera. When the observer is far from the object, the rays pass through this image plane nearly perpendicularly, and the plane can be taken to be comparatively close to the observed object. The rays need to be solved where they are indicated in solid lines, and the radiative transfer equation needs to be solved only inside the radiating region, as in vacuum the equation is trivial.

5. Numerical Methods: the ARCMANCER Library

The ray-tracing procedure described in section 4.4 is impossible to perform analytically in all but the most trivial cases, and thus the use of numerical methods is required. This chapter describes the implementation of ARCMANCER, a C++14 header-only library which provides tools for applying the ray tracing procedure in arbitrary user specified spacetimes. These tools are implemented in a fairly general manner, and can also be applied to other problems which involve computing curves or tensor arithmetic on Riemannian and pseudo-Riemannian manifolds. This generality also facilitates extending the ray-tracing procedure used here, for example to include the various possible approaches to approximating scattering effects. Unlike many other scientific software projects, as a library ARCMANCER is not an executable program that performs some predefined computation. Instead the classes and functions defined in ARCMANCER are used by other codes to define and perform the computations needed for their specific applications. In addition to the C++ library, ARCMANCER also includes Python bindings for the core radiative transfer features, which simplifies performing small scale computations and allows easy prototyping.

The ARCMANCER project was started by Pauli Pihajoki after he noted the lack of a public, general purpose ray-tracing library. This resulted in a code that was fairly general and complete enough to be applicable to some problems (Pihajoki et al., 2016; Nättilä and Pihajoki, 2017), but was not yet deemed ready to be made publicly available. At this point the features of ARCMANCER included the core necessities for performing polarized radiative transfer: computing geodesics of a user defined metric in a single coordinate system with two parallel transported basis vectors, finding the intersections of these geodesics with user defined surfaces and solving the radiative transfer equation along them. After I joined the project, the first version of the code was rewritten almost completely to increase its generality even further, with the implementation of basic differential geometric objects seeing the most major changes. As a result, the code

gained several new features which increase its capabilities, such as transport of general objects including arbitrary tensors along general curves. Other improvements also make it easier to avoid programming mistakes and simplify implementing radiative transfer computations considerably compared to the previous version of Arcmancer. This rewrite and redesign of Arcmancer forms a major part of the work associated with this thesis. In particular, the implementations of generic tensors and transport of objects along curves are mainly my work, as is the generation of image plane initial conditions, but I have also worked on most of the other currently implemented features.

The following sections describe the structure and main features of Arcmancer, as well as the numerical methods used in the implementation of these features. Compared to other codes aimed at relativistic ray-tracing, Arcmancer includes several novel features, such as support for using multiple coordinate charts on the same manifold, which simplifies for example combining results from multiple sources using different coordinate systems as well as avoiding coordinate singularities. Some of the other codes described in the literature are discussed and compared to Arcmancer in section 5.8. Appendix B includes some simple code examples which illustrate using the main features of Arcmancer through the Python bindings.

5.1. Overview of Arcmancer

Before a detailed discussion of the features and implementation of Arcmancer, an overview of its structure and components is presented. The Arcmancer library can be divided into three basic components: utilities, differential geometry and radiative transfer. The utilities component includes various tools used in the implementation of the other components, such as cubic spline interpolation (section 5.5), and is intended mainly for internal use by the library. The geometry and radiative transfer components are the main parts of the library, and contain almost all of the functionality used by applications.

Differential Geometry

The differential geometry component forms the core of Arcmancer, and is applicable to a wide variety of problems. The main classes forming the differential geometry implementation are

- `MetricSpace`, `Chart` and `ManifoldPoint`,

  which describe a manifold and points on it;

- `Tensor`,

  which is a generic implementation of tensors;

- `LorentzFrame`,

  which implements local inertial frames;

- `ParametrizedPoint` and `ParametrizedCurve`,

  which implement generic parametrized curves on manifolds; and

- `Surface`,

  which is a base class for surfaces on manifolds.

In order to support general manifolds, all of the above classes are templatized at least based on the manifold they exist on, but this information on template parameters is omitted for brevity. For example, in Minkowski spacetime the actual class representing a point would be `ManifoldPoint<MinkowskiSpacetime>`. Here the template parameter `MinkowskiSpacetime` is a class derived from `MetricSpace` implementing the Minkowski spacetime. Arcmancer includes several such implementations of common spacetimes. Additionally, the differential geometry component includes various convenience functions, such as projections of vectors. Sections 5.2 – 5.5 discuss the differential geometry component in further detail.

Radiative Transfer

The radiative transfer component is built on top of the capabilities provided by the differential geometry component. Its main classes are

- `PolarizationFrame`,

  which manages the polarization basis vectors;

- `StokesVector` and `PolarizationSpectrum`,

  which describe invariant Stokes vectors and their frequency spectrum; and

- `ImagePlane`,

  which handles ray initial conditions on the image plane.

The actual radiative transfer computation is performed by the `radiation_transfer` function. Sections 5.6 and 5.7 discuss the details of these classes and functions.

5.2. Representation of Geometric Objects

The core functionality of ARCMANCER consists of the general and flexible implementation of basic geometric objects, such as manifolds, tensors and curves, which are described in this section. The approach taken in the implementation is to closely imitate the abstract mathematical formalism of differential geometry, which ensures generality and simplifies converting mathematics to code. On the practical side, much of the implementation's flexibility is provided by C++ template metaprogramming, and as a result the additional runtime overhead from this flexibility is small. In general the C++ language allows using fairly high levels of abstraction with zero runtime overhead.

5.2.1. Manifolds

The central object on which everything else depends on is of course the manifold with a metric tensor field. In addition to the metric tensor, important structures are the actual points on the manifold, as well as the coordinate charts which give the coordinate representations of those points. ARCMANCER supports defining manifolds of in principle arbitrary dimension and metric signature, but naturally the most complete set of functionality is provided for spacetimes, i.e. four-dimensional Lorentzian manifolds with metric signature $(+---)$ or $(-+++)$.

A manifold is represented by an instance of a class which extends a base `MetricSpace` class providing functionality common to all manifolds with a metric tensor. The class representing the manifold defines a set of charts that can be used on the manifold. These charts are represented by `Chart` objects, which act as identifying tags for the charts and contain basic information on the chart, such as a descriptive name. In the current implementation, the domain of the chart cannot be explicitly specified, but this will likely be added in the future, as charts that do not cover the entire manifold are very common. For each chart on the manifold, functions returning the components of the metric tensor $g_{\mu\nu}$ and its partial derivatives $\frac{\partial g_{\mu\nu}}{\partial x^\alpha}$ need to be defined. These are the only functions that need to be defined for a manifold when only a single chart is used, as the inverse metric tensor $g^{\mu\nu}$ can be computed numerically by inverting the component matrix, and the Levi-Civita connection coefficients can also be numerically computed from these quantities.

For manifolds with multiple charts, some additional function definitions are necessary. These functions give the transition functions, which map coordinates from one chart to another, and the Jacobian matrices associated with those transitions. The number of

required functions grows quickly with the number of charts, as for each pair of charts the transition function and Jacobian are needed in both directions. To ease this process, transition functions and Jacobians can be automatically generated by chaining together already defined functions and numerically inverting the Jacobian matrices. This feature is implemented by using standard graph algorithms provided by the Boost Graph library (Siek et al., 2001), with which it is possible to find the shortest paths between charts through the directed graph of transition functions.

Points on manifolds are represented by `ManifoldPoint` objects. Internally they store the coordinates of the point in some definite chart, but this is abstracted away by providing a method that returns the coordinates corresponding to the point in any of the charts defined on the manifold. As a result, from the perspective of the code dealing with points the `ManifoldPoint` objects behave just like the abstract mathematical points, which are mapped to coordinates by the coordinate function ϕ of a chart. This approach prevents programming errors related to the use of multiple coordinate systems, as the coordinate system always needs to be explicitly specified before getting access to the coordinates. In comparison, when points are represented as bare arrays of numbers, as is commonly done in codes dealing with points in some space, the coordinate system used needs to be known from the context. This easily results in hard to find programming mistakes and incorrect computational results when dealing with multiple coordinate systems.

5.2.2. Tensors

Tensors are another core component necessary for performing relativistic and differential geometric computations. The tensor implementation in Arcmancer supports tensors of in principle arbitrary rank, with practicality and memory consumption setting some limits on the usability of high rank tensors. Key features of the tensor implementation are automatic computation of components in different charts, generation of the parallel transport derivatives and type checked operations between tensors. All tensors, including tangent and cotangent vectors, use the same generic implementation.

An object of the `Tensor` class stores a `ManifoldPoint` corresponding to the base point of the tensor space on the manifold, and the components of the tensor in some definite chart. Similarly to `ManifoldPoint`, `Tensor` abstracts away the internal component representation, so that to other parts of the code the `Tensor` objects appear to be an abstract mathematical entity. In addition to the coordinate bases of the charts defined on the manifold, it is also possible to work with tensor components in a given Lorentz

frame (see section 5.2.3), which greatly simplifies computation of observed quantities. The component storage, and most of the functionality relating to operations on tensor components, is implemented using the tensor module of the Eigen numeric linear algebra library (Guennebaud, Jacob, et al., 2010), which provides access to fast operations on multidimensional arrays. The tensor module is still listed as an incomplete and unsupported part of the Eigen library, but the feature set and performance was found to already be good enough for use in Arcmancer, with certain performance critical sections of code requiring some additional adjustments.

The positions of the tensor indices are represented at the type level, which allows the compiler to type-check tensor operations, forbidding operations that would not produce a tensor, such as contractions of two contravariant indices. Most of the common tensor operations, such as raising and lowering indices, tensor products and arithmetic operations are implemented. The binary operations also check that both tensors are located at the same point on the manifold. In addition to the generic tensor operation, some commonly used operations such as vector dot products and various projections are implemented as separate optimized functions. These optimized functions use the standard linear algebra facilities of Eigen, which have better performance than the corresponding tensor operations.

In addition to type-checking tensor operations, the index position information is also needed for computing the components of a tensor in different charts, as the transformation rules for co- and contravariant indices are different. Similarly, the implementation of parallel transport needs this information to compute the correct contractions with the connection coefficients. These computations are generated using a compile time loop over the indices to generate a sequence of calls to a function performing the computation for a single index. This approach allows a capable compiler to eliminate the unused branches and loops in the computation, causing no additional performance penalty over explicitly writing implementations for low rank tensors such as the vector types.

5.2.3. Lorentz Frames

Local Lorentz frames are a powerful tool for dealing with observers and the quantities they observe in a manner that is independent of the coordinate systems used. In Arcmancer, a local Lorentz frame is represented by an object of the `LorentzFrame` class. A `LorentzFrame` object can be constructed by specifying three non-degenerate directions. The first of these directions is the observer's four-velocity u, which becomes the time-like basis vector e_t of the frame. The remaining two directions are orthonormalized to give

the e_z and e_x basis vectors, and the e_y basis vector is computed using the cross product in the observer's space-like subspace, given by

$$e_y^\mu = -e_t^\alpha \epsilon_\alpha{}^\mu{}_{\rho\sigma} e_z^\rho e_x^\sigma. \tag{5.1}$$

The change of basis from the e_α basis vectors to a coordinate basis $\frac{\partial}{\partial x^\mu}$ is given by the component matrix $e_\alpha{}^\mu$ with $e_\alpha = e_\alpha{}^\mu \frac{\partial}{\partial x^\mu}$. These matrices work just like the Jacobian matrices between different charts, allowing the `Tensor` class to also work with components expressed in the basis of the Lorentz frame.

Other functionality of the `LorentzFrame` class includes generating null vectors with a given spatial component, which is useful for setting ray-tracing initial conditions, and the ability to be parallel transported along curves. These properties are essential in the procedure used in generating the image plane, as described in section 5.7.

5.2.4. Curves

ARCMANCER supports computing parametrized curves that satisfy equations of motion of the form

$$\frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda} = k^\mu \tag{5.2}$$

$$\frac{\mathrm{d}k^\mu}{\mathrm{d}\lambda} = f^\mu(x, k). \tag{5.3}$$

It is also possible to transport arbitrary objects along these curves, as long as they can be represented as an array of real numbers and satisfy a similar first order differential equation. The most important application for relativistic computations are naturally curves that are geodesics, and the parallel transport of tensors along those geodesics, in which case $f^\mu = -\Gamma^\mu{}_{\alpha\beta} k^\alpha k^\beta$ for the tangent vector of the geodesic, and similarly for the transported tensor. Other applications that should also be easy to implement using this system include any systems which resemble a test particle under an external force. Some examples of such systems are rays of light in a refractive medium and extreme mass-ratio inspirals, where the dynamics can be modelled as a gravitational self-force (Barack, 2009). Additionally, the curve implementation supports finding intersections with surfaces, which are described in section 5.2.5.

A curve is represented as a set of discrete points, which are stored as instances of the `ParametrizedPoint` class at different values of the curve parameter λ. Each `ParametrizedPoint` contains the point x on the manifold and the tangent vector k

of the curve, as well as any transported objects. All curves are therefore computed as curves in the tangent bundle or some other, more complicated bundle if there are transported objects. In addition to containing the different objects at a point on the curve, `ParametrizedPoint` also handles conversions between the separate component representations of the objects and their equations of motion, and the single unstructured array required as an input for the numerical integration procedure described in section 5.4. The storage of the points and solving the equations of motion is handled by the `ParametrizedCurve` class. It also contains the functionality to approximate a continuous curve using the discrete set of solution points through interpolation, which is discussed in section 5.5.

5.2.5. Surfaces

Arcmancer supports surfaces which can be defined by an equation of the form

$$S(x) = 0, \tag{5.4}$$

where S is some scalar function on the manifold. Surfaces defined this way are therefore $(m-1)$-dimensional submanifolds of an m-dimensional manifold, but in a spacetime spacelike slices of such surfaces usually coincide with the intuitive notion of a surface in 3-dimensional space. In addition to the function S, implementations of surfaces need to also provide the gradient $\mathrm{d}S$ and the velocity vector of an observer comoving with the surface, which allows defining impact angles of curves with the surface. It is also possible to define regions that act as holes in the surface.

Some uses for these surfaces are representing surfaces of physical objects such as neutron stars or extremely optically thick regions and placing computational domain boundaries on the manifold, for example to limit the rays used in radiative transfer only to the radiating volume or to prevent curves from hitting singularities.

5.3. Automatic Chart Selection

To allow operations between objects that use different charts internally, some way of choosing the chart to use for the computations is needed. In principle the choice could be arbitrary, but due to the limitations of floating point arithmetic on computers, a poor choice of chart can cause rounding-off errors in the result. It is also desirable to avoid the coordinate singularities often present in the charts, such as the polar singularity of

spherical coordinates. Another consideration that can be applied when computing curves is that the choice of chart affects how complicated the curve appears to be. A simple example of this would be circles and straight lines on a flat plane, where the correct choice of polar or Cartesian coordinates greatly simplifies the coordinate expression for the curve.

A source for floating point errors are the differences in the magnitudes of the coordinate basis vectors $\frac{\partial}{\partial x^\mu}$. If the magnitudes have large differences, the numerical values of the tensor components can likewise have large differences, even if the only physical difference is in orientation. This motivates the main method of chart selection in ARCMANCER, which uses the singular value decomposition (e.g. Press et al., 1992) of the matrix of metric components $(g_{\mu\nu})$ to find the chart where the basis vectors have the least difference in magnitude. A singular value decomposition of a real rectangular matrix $\mathbf{M}$ consists of orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix $\mathbf{\Sigma}$, which satisfy $\mathbf{M} = \mathbf{U\Sigma V}$. The diagonal entries Σ_{ii} are the singular values of $\mathbf{M}$ and describe how $\mathbf{M}$ scales different directions, giving the lengths of the axes of an ellipsoid that results from the application of the transformation $\mathbf{M}$ to a unit sphere. The matrices $\mathbf{U}$ and $\mathbf{V}$ correspond to rotations, which determine the final orientation of the ellipsoid. Therefore the ratio of the minimum and maximum singular value, or the inverse of the condition number, of $(g_{\mu\nu})$ quantifies how different the magnitudes of the basis vectors are. The chart where this ratio is the largest is then chosen as the chart to use for calculations, as in that chart the difference in the magnitudes of the basis vectors is the smallest.

Another method of chart selection is used for choosing the chart to use for integration of the curve equations of motion. First, the charts are compared using singular values as before, but in cases where different charts have similar condition numbers this may not be enough to choose the best chart as the behaviour of the curve depends also on its direction. In this case, which is detected by the ratio of singular values being under some configurable limit, the derivatives of the curve's tangent vector are compared, and the chart with the smaller derivatives is chosen. This leads to the selection of the chart where the coordinate representation of the curve is the straightest in the Euclidean sense, which should be the easiest to integrate numerically. This chart selection procedure is used at the beginning of each curve computation, and may also be triggered automatically during integration if configurable limits on the number of rejected integration steps are exceeded.

It does not appear that much work has been done relating to the automatic numerical selection of coordinate systems, likely because most numerical computations only

employ a single chart. As a result, the methods described above do not have a strong theoretical foundation, and should at this point be treated with some caution. However, these methods do seem to produce reasonable results. Performing the careful analysis necessary to show their correctness and optimality, or finding alternative methods, would however allow greatly improving the support for using multiple charts. This together with proper handling of partially overlapping coordinate charts suggests an interesting topic for future work.

Having automatic chart selection enabled is not always desirable, for example due to the non-trivial computational overhead caused by the selection methods or for testing purposes. Hence it is also possible to manually fix the chart that is used for all calculations. In this case computations behave similarly to systems where there is only one chart available, but it is still possible to input points and tensors using all the available coordinate charts.

5.4. Solving the Equation of Motion of Curves

The equation of motion for the combined system of the point on a manifold, the tangent vector of the curve and the object transported along the curve is of the general form

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\lambda} = \mathbf{f}(\mathbf{y}), \tag{5.5}$$

where $\mathbf{y}$ is a vector, i.e. a simple array, containing the components of the system. A general class of methods that can be used for solving systems of this form are the Runge-Kutta methods.

5.4.1. Runge-Kutta Methods

Runge-Kutta methods (e.g. Hairer et al., 2008) are a class of numerical methods that find an approximate value for $\mathbf{y}_1 \approx \mathbf{y}(t_0 + h)$ given $\mathbf{y}_0 = \mathbf{y}(t_0)$, with $\mathbf{y}(t)$ satisfying a general first order ordinary differential equation of the form

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}). \tag{5.6}$$

The parameter h is called the *stepsize*. An s-stage explicit Runge-Kutta method has the form

$$\mathbf{y}_1 = \mathbf{y}_0 + h \sum_{i=1}^{s} b_i \mathbf{k}_i, \tag{5.7}$$

$$\mathbf{k}_i = \mathbf{f}\left(t_0 + c_i h, \mathbf{y}_0 + h \sum_{j=1}^{i-1} a_{ij} \mathbf{k}_j \right), \tag{5.8}$$

with a_{ij}, b_i and c_i being coefficients specific to the method. This corresponds to computing a set of intermediate estimates for the function y between t_0 and $t_0 + h$. Explicit methods are easy to implement as all the terms can be computed directly. There are also implicit Runge-Kutta methods, where the sum in (5.8) extends to $j = s$, leading to a set of linear equations that need to be solved for the $\mathbf{k}_i$. A Runge-Kutta method is of order p if

$$\|\mathbf{y}(t_0 + h) - \mathbf{y}_1\| \leq K h^{p+1} \tag{5.9}$$

for some constant K and a suitable norm $\|\cdot\|$, i.e. the error for a single step is of order $\mathcal{O}(h^{p+1})$. To find the solution over some large interval, the Runge-Kutta method is applied iteratively to the result of the last step.

In ARCMANCER, the Dormand-Prince 5(4) (`dopri5`) method (Dormand and Prince, 1980) provided by the Boost Odeint library (Ahnert and Mulansky, 2011) is used for solving the equations of motion. It is a fifth order explicit Runge-Kutta method with error estimation provided by an embedded fourth order method, which allows automatically changing the stepsize to avoid excess work while maintaining the desired local accuracy. The numerical tolerances for a single step can be configured by the user to achieve the desired balance between accuracy and performance. The configurable parameters are the absolute tolerance ϵ_{abs} and the relative tolerance ϵ_{rel}, which Odeint uses to determine an error parameter

$$\mu = \max_i \left(\frac{|e_i|}{\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \left(|y_i| + \left| \frac{\mathrm{d}y_i}{\mathrm{d}t} \right| \right)} \right), \tag{5.10}$$

where y_i are the components of the vector $\mathbf{y}$ and e_i the corresponding error estimates provided by the embedded method. This error parameter is used to determine the new stepsize as

$$h_{\text{new}} = \begin{cases} h_{\text{current}} \max\left(0.9\mu^{-1/(O_E-1)}, 0.2\right) & \text{if} \quad \mu > 1, \\ h_{\text{current}} \max\left(0.9\mu^{-1/O_S}, 0.2\right) & \text{if} \quad \mu < 0.5, \\ h_{\text{current}} & \text{otherwise} \quad . \end{cases} \tag{5.11}$$

For the `dopri5` method the order parameters O are $O_S = 5$ and $O_E = 4$. The stepsize has also configurable upper and lower bounds that override the automatically selected size.

5.4.2. Surface Intersections

To support computing intersection points with surfaces, the integration system has two modes of operation. In normal operation the integrated equation is supplied directly by `ParametrizedPoint`. In addition to the automatic stepsize control provided by the `dopri5` integrator, the stepsize is also reduced when surfaces are near the current position to avoid stepping too far through a surface. If the curve passes through a surface, which is detected by a change of sign of the function S between successive points, the integration system changes the mode to solve the exact intersection point. This mode uses Henon's trick (Henon, 1982), which performs a change of integration variable from the curve parameter λ to the value of the function S, by using the chain rule and the result

$$\frac{\mathrm{d}S(x(\lambda))}{\mathrm{d}\lambda} = k^\mu \partial_\mu S = \mathrm{i}_k \mathrm{d}S. \tag{5.12}$$

The transformed equation is then

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}S} = (k^\mu \partial_\mu S)^{-1} \mathbf{f}(\mathbf{y}), \tag{5.13}$$

which is integrated to the value $S = 0$ using the same `dopri5` method. This way the intersection point is found quickly and accurately in a single iteration.

5.5. Curve Interpolation

The adaptive integration method used for computing curve points results in a curve represented by unevenly spaced points, with large gaps between adjacent points where the curve is comparatively straight. However, radiative transfer computations and various other uses require points on the curve at arbitrary values of curve parameter. It would be possible to compute these points using the same integration routine by specifying in advance the required curve parameter values, but this would be inefficient and would lose the advantages offered by the adaptive solver. For efficiently computing curve points at arbitrary values of curve parameter within the computed curve span it is instead possible to use interpolation of the integrated curve points.

The method used for curve interpolation in ARCMANCER is cubic spline interpolation (e.g. Press et al., 1992), which is efficient and has good accuracy. Given the values $\mathbf{y}_i$ of a one parameter vector valued function $\mathbf{y}(\lambda)$ at the points λ_i, $i = 1 \dots N$, the values $\mathbf{y}(\lambda)$ are approximated by a *spline*, which is defined piecewise by cubic polynomials $\tilde{\mathbf{y}}_i$ on each interval $\lambda \in [\lambda_i, \lambda_{i+1}]$. These polynomials are constructed so that the values at the known points are correct, i.e. $\tilde{\mathbf{y}}_i(\lambda_i) = \mathbf{y}_i$ and $\tilde{\mathbf{y}}_i(\lambda_{i+1}) = \mathbf{y}_{i+1}$, with the first and second derivatives of the polynomials matching at the edges of intervals

$$\left.\frac{\mathrm{d}\tilde{\mathbf{y}}_i}{\mathrm{d}\lambda}\right|_{\lambda=\lambda_{i+1}} = \left.\frac{\mathrm{d}\tilde{\mathbf{y}}_{i+1}}{\mathrm{d}\lambda}\right|_{\lambda=\lambda_{i+1}} \tag{5.14}$$

$$\left.\frac{\mathrm{d}^2\tilde{\mathbf{y}}_i}{\mathrm{d}\lambda^2}\right|_{\lambda=\lambda_{i+1}} = \left.\frac{\mathrm{d}^2\tilde{\mathbf{y}}_{i+1}}{\mathrm{d}\lambda^2}\right|_{\lambda=\lambda_{i+1}}. \tag{5.15}$$

The polynomials satisfying these conditions are given by

$$\tilde{\mathbf{y}}_i(\lambda) = A_i\mathbf{y}_i + B_i\mathbf{y}_{i+1} + C_i\mathbf{y}''_i + D_i\mathbf{y}''_{i+1}, \tag{5.16}$$

where A_i, B_i, C_i and D_i are the polynomials

$$A_i = \frac{\lambda_{i+1} - \lambda}{\lambda_{i+1} - \lambda_i} \tag{5.17}$$

$$B_i = \frac{\lambda - \lambda_i}{\lambda_{i+1} - \lambda_i} \tag{5.18}$$

$$C_i = \frac{1}{6}(A_i^3 - A_i)(\lambda_{i+1} - \lambda_i)^2 \tag{5.19}$$

$$D_i = \frac{1}{6}(B_i^3 - B_i)(\lambda_{i+1} - \lambda_i)^2, \tag{5.20}$$

and $\mathbf{y}''_i = \left.\frac{\mathrm{d}^2\tilde{\mathbf{y}}_i}{\mathrm{d}\lambda^2}\right|_{\lambda=\lambda_i}$ are unknown constants.

The unknown $\mathbf{y}''_i$ are determined uniquely by the continuity of the derivatives, given the values $\mathbf{y}''_1$ and $\mathbf{y}''_N$. A natural choice is to set $\mathbf{y}''_1 = \mathbf{y}''_N = 0$, but as the derivatives of the curve coordinates are known, an even better choice is to use the values solved from

$$\frac{\mathrm{d}\tilde{\mathbf{y}}}{\mathrm{d}\lambda} = \frac{\mathbf{y}_{i+1} - \mathbf{y}_i}{\lambda_{i+1} - \lambda_i} - \frac{3A^2 - 1}{6}(\lambda_{i+1} - \lambda_i)\mathbf{y}''_i + \frac{3B^2 - 1}{6}(\lambda_{i+1} - \lambda_i)\mathbf{y}''_{i+1} \tag{5.21}$$

with $\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\lambda} = \frac{\mathrm{d}\tilde{\mathbf{y}}}{\mathrm{d}\lambda}$ at $\lambda = \lambda_1, \lambda_N$. The resulting equations form a linear tridiagonal system, which can be efficiently solved in $\mathcal{O}(N)$ operations. The solution to the system of equations is stored, after which the evaluation of the value of the interpolation spline at

arbitrary λ in the interval $[\lambda_1, \lambda_N]$ is fast, as it only requires finding the correct interval $[\lambda_i, \lambda_{i+1}]$ and evaluating a polynomial with known coefficients.

In ARCMANCER the interpolation spline is constructed the first time a point on a curve is interpolated after the curve has been extended by solving the equation of motion. The simplest method constructs the interpolation spline over the whole span of the curve, representing all points in a single chart, but it is also possible to construct splines over only a part of the curve. It should be noted that the whole section of the curve which is interpolated needs to be able to be represented in a single chart, which needs to be taken into account if only partially overlapping charts are used.

5.6. Solving the Radiative Transfer Equation

The first steps of the ray-tracing procedure described in section 4.4 were solving the path of the ray and parallel transporting the polarization basis vectors along the ray. This is simple to implement using the previously described curve computation capabilities together with `PolarizationFrame` objects, which can be parallel transported along curves like a `LorentzFrame`, but only contain the necessary two basis vectors. Here it is also possible to use surfaces to restrict the curve to the radiating region to optimize the computation.

After the ray has been computed, the radiative transfer equation (4.49) needs to be solved along it. This is done using the `radiation_transfer` function. The radiative transfer equation is of the form

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\lambda} = \mathbf{f}(x(\lambda), \mathbf{y}) \tag{5.22}$$

and is a special case of equation (5.6), which can be solved using Runge-Kutta methods as described in section 5.4.1. Here $\mathbf{y}$ and $\mathbf{f}$ are vector valued functions, and $x(\lambda)$ is the point on a curve at the parameter value λ, including the tangent vector and other objects transported along the curve. ARCMANCER provides a general `line_integral` function for solving these types of equations along curves using similar adaptive integrators as for computing curves, with the point $x(\lambda)$ evaluated using interpolation. Thus the only remaining technical question in the implementation of the `radiation_transfer` function is defining the function $\mathbf{f}$.

Obviously, the function $\mathbf{f}$ is simply the right hand side of the transfer equation (4.49)

$$\mathbf{f}(x, \mathcal{I}) = \mathcal{J}(x) - \mathcal{M}(x)\mathcal{I}. \tag{5.23}$$

The matrices $\mathcal{M}$ and $\mathcal{J}$ are computed based on two user provided functions which represent a fluid model and a radiation model. The fluid model returns the properties of the radiating fluid at a given point, including data such as the fluid four-velocity, the magnetic field four-vector and the density of the fluid. These details are needed both for computing the Mueller matrix $\mathbf{M}_\nu$ and emissivity $\mathbf{J}_\nu$, which is the task of the user provided radiation model, as well computing $\mathcal{M}$ and $\mathcal{J}$ from these, which is done following the procedure described in section 4.3.5.

This separation of the radiation and fluid models makes the interface for the radiative transfer computations simple and flexible. It also allows for some optimizations, avoiding unnecessary repeated computation of the same values. In the approximation used in this work the ray path does not depend on the frequency, and thus the radiation transfer equation can be simultaneously solved for multiple frequencies by combining the equations for each frequency into a single system. This significantly improves performance as the same interpolated curve points and computed fluid variables can be used for all frequencies, and only the radiation model output needs to be computed separately for each frequency.

A problem with the current implementation is handling of systems which have high absorption or Faraday rotation coefficients, as in these cases the Runge-Kutta integrator often needs to take excessively small steps due to the rapid decay or oscillation of the Stokes intensities. In these cases the radiative transfer equation is said to become stiff, and would require different methods for an efficient solution. However, it should be relatively straightforward to add such methods in the future. Implementing radiative transfer which accounts for refractive effects should also be simple due to the support for general forces acting on the curves. The necessary modifications to the transport equation would of course first need to be derived from theory, with one possible approach given by Broderick and Blandford (2003, 2004), as was mentioned in section 4.3.2.

5.7. Image Generation

The typical use case for radiative transfer calculations is the production of images from some model in order to compare them with observations. To do this, the radiative transfer calculation is performed for a ray passing through each of the image pixels. As astrophysical objects are very far away, it is unfeasible to perform the computation over the whole path of the ray from the observed object to the observer. Usually over a large part of the rays path the space is approximately flat and empty, in which case

the rays forming the image are approximately perpendicular to some plane relatively near the observed object, as was illustrated in figure 4.4. Naturally this plane needs to be far enough from any massive objects so that the spacetime is flat also at the plane's position. In any given coordinate system it is possible to find analytic expressions for these image planes and the directions of the rays, but they can also be constructed without explicit reference to the coordinates. ARCMANCER implements such a construction in the `ImagePlane` class, which allows the same image plane generation code to work in all asymptotically flat spacetimes.

The procedure starts by defining a Lorentz frame at the centre of the image plane. This frame defines the directions along the plane and normal to it, with the z-axis corresponding to the plane normal and the incoming ray direction. The image plane is defined by specifying the physical distances d_x, d_y to the edges of the plane and the number of pixels in each direction. The coordinate system on the image plane is defined so that at the corresponding edges $x = \pm d_x$ and $y = \pm d_y$. From this the x and y coordinates of each pixel are computed by requiring that the pixels' coordinates are evenly spaced, and the location of the pixel is found by propagating a spacelike geodesic with a tangent vector $v = xe_x + ye_y$ for an affine parameter distance of 1. The basis vectors are parallel transported along the geodesic to define the directions at each pixel, as illustrated in figure 5.1. This is essentially a construction of the Riemann normal coordinates corresponding to the observer at the centre of the image plane.

After the initial conditions are generated on the image plane, the image is computed using a user provided function that performs the desired calculation for each image pixel. This allows easily performing various different computations to form images of the desired quantities. In addition to computing radiative transfer over a single ray for each pixel, it is also easy to perform computations using multiple curves, or to compute images of quantities such as optical thickness or ray endpoint coordinates. Since the computations for each pixel are independent, the computation of the image is naively parallelized, which allows fully utilizing modern multi-core CPUs.

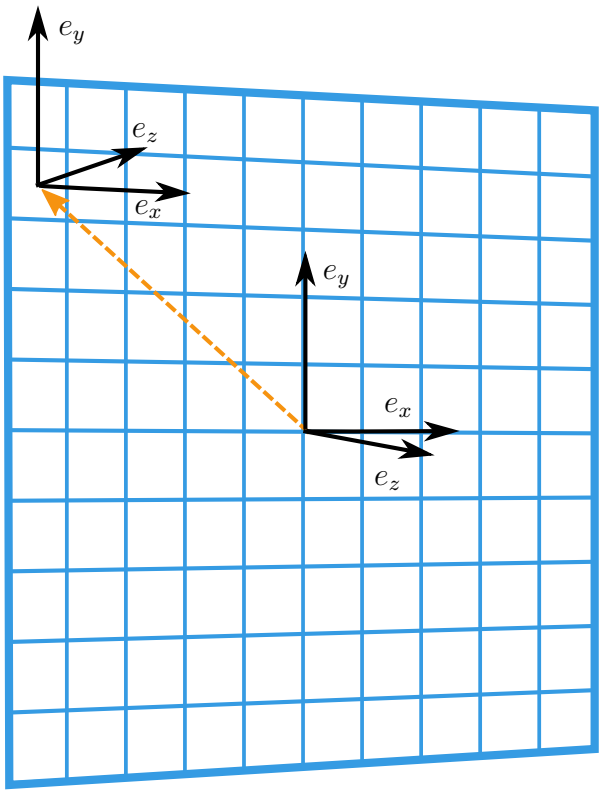

Figure 5.1.: A sketch of the image plane construction, where the frame defined at the image plane origin is parallel transported to the centre of each pixel along a spatial geodesic, indicated with an orange dashed line. The frames also naturally include a e_t basis vector, but this is not shown here.

5.8. Comparison to Other Codes

As ray-tracing is such a conceptually simple procedure, it is natural that there are already several other codes which implement similar features as ARCMANCER. An important application of ray-tracing based radiative transfer is modelling accretion flows around spinning black holes described by the Kerr spacetime, which is discussed further in section 6.2.1. The importance and relative simplicity of this special case has led to the development of several codes, although most of them are not public. Publicly available codes which can perform polarized radiative transfer in the Kerr spacetime include ASTRORAY (Shcherbakov and McKinney, 2013) and GRTRANS (Dexter, 2016). GRTRANS applies semianalytic formulas to compute the ray geodesics and parallel transported basis vectors, which reduces the computational load of computing the rays, but also prevents generalizing the code to other spacetimes. GRTRANS also includes methods specially adapted to solving the radiative transfer equation which should handle

optically thick cases better than the method currently used by ARCMANCER.

Other public codes which deal with polarization or radiative transfer in Kerr spacetime include KERTAP (Chen et al., 2015), `GRay` (Chan et al., 2013, 2017) and `ipole` (Mościbrodzka and Gammie, 2017). KERTAP can compute rays and parallel transport polarization vectors, but does not implement radiative transfer. `GRay` implements non-polarized radiative transfer, but does so using highly parallel GPU processing, giving it high performance. However, its integrator uses a predetermined stepsize, so its accuracy is not completely guaranteed. The main feature of interest in `ipole` is that it uses a formulation of the radiative transfer equation which is directly based on the distribution tensor. There are also several private codes, such as the one used by Broderick and Blandford (2003, 2004), which includes refractive plasma effects in addition to polarized radiative transfer. While such effects are not currently included in the radiative transfer method used by ARCMANCER, it should be relatively simple to implement them using the generic facilities described in the previous sections.

The specialized codes trade generality to potentially improved performance provided by analytic simplification of the computations, but this means that their applicability is restricted, and modifying them to perform computations in other spacetime geometries would require a significant amount of work. There are also codes which support arbitrary spacetimes similarly to ARCMANCER, such as GYOTO (Vincent et al., 2011) and Motion4D (Müller and Grave, 2009). However, these do not support polarized radiative transfer, with GYOTO only including scalar transfer and no parallel transport of vectors. While Motion4D appears to support parallel transport of vectors, it is unclear if it could be used for polarized radiative transfer without heavy modification, as its main purpose appears to be simple visualization of objects in curved spacetimes instead of producing results comparable to observations. The support for different spacetimes in these codes is also much more rudimentary than in ARCMANCER, as they only allow specifying a single metric with an implicit coordinate system, compared to the support for multiple explicit coordinate systems described here.

Overall, the features of ARCMANCER allow reproducing most of the capabilities of previously available codes with more generality. Combined with the features of ARCMANCER that are not available or even implementable in other codes without a complete rewrite, it seems that the only significant advantage other, more specialized codes have over ARCMANCER is their potential for better performance. This suggests that ARCMANCER can be used for most applications where relativistic ray-tracing is needed.

6. Validation and Applications

In this chapter the numerical procedures described in chapter 5 are shown to be implemented correctly in the ARCMANCER library, and their accuracy is evaluated. The effects of using different coordinate charts on the accuracy of numerically solving the geodesic equation are also investigated. After this, representative example applications to radiation from black hole accretion disks are shown to illustrate the capabilities of the library. These include post-processing the output of a general relativistic magnetohydrodynamic simulation and the computation of light curves from a lensing event in a binary black hole system.

6.1. Accuracy of Radiative Transfer in Flat Spacetime

The first and simplest test that can be carried out is checking that the radiative transfer solver gives the correct results. In flat spacetime the rays are straight, so the accuracy of the curve computation and interpolation is not a concern, and the radiative transfer implementation can be evaluated separately.

The radiative transfer equation (4.49) does not have many simple solutions, but the case where $\mathcal{J}$ and $\mathcal{M}$ are constant has the solution

$$\mathcal{I}(\lambda) = \exp(-\lambda\mathcal{M})(\mathcal{I}(0) - \mathcal{M}^{-1}\mathcal{J}) + \mathcal{M}^{-1}\mathcal{J}, \tag{6.1}$$

where the exponential of a matrix is defined in the standard way through the Taylor series expansion. Figure 6.1 compares this analytical solution to the numerical solution with randomly generated numerical values for $\mathcal{M}$ and $\mathcal{J}$. As is evident, the numerical solver behaves as expected, with the error staying within the set tolerances.

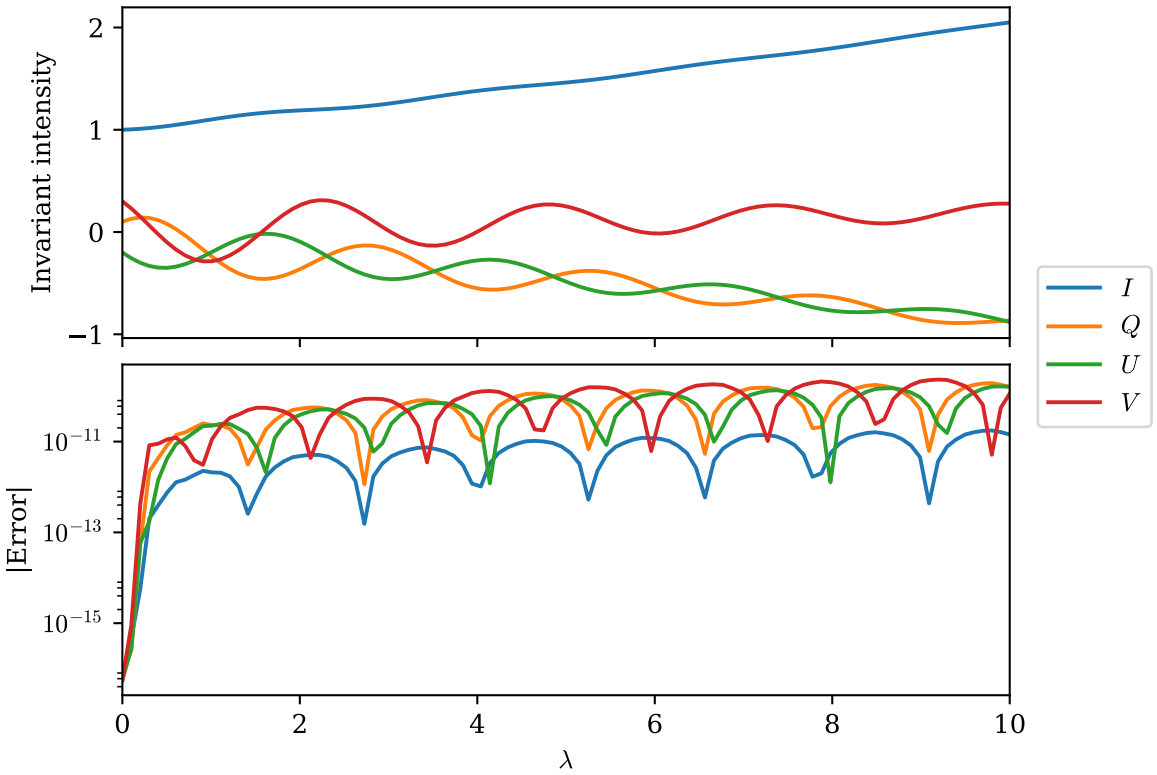

Figure 6.1.: A solution to the radiative transfer equation with constant coefficients. The top panel shows the values of the different components of the invariant Stokes vector $\mathcal{I}$, and the bottom panel shows the error between the analytic solution and the numerical solution computed with ARCMANCER. The integration tolerances were set to $\epsilon_{\text{abs}} = \epsilon_{\text{rel}} = 10^{-10}$.

6.2. Applications in the Kerr Spacetime

An important application of relativistic radiative transfer codes is modelling the emission from accretion around black holes. Astrophysical black holes can be modelled to a good approximation by the Kerr spacetime, which contains a stationary rotating black hole in an asymptotically flat background.

6.2.1. Properties of the Kerr Spacetime

To understand the applications presented here, some background theory on the Kerr solution is necessary, although the discussion here is by necessity of limited scope. For more details, see e.g. Visser (2007).

In analytic work a common choice of coordinates are the spherical Boyer-Lindquist coordinates (t, r, θ, ϕ) (Boyer and Lindquist, 1967), in which the Kerr metric is given as

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{2Mr}{\rho^2}\Big(a\sin^2\theta\,\mathrm{d}\phi + \mathrm{d}t\Big)^2 - \rho^2\left(\frac{\mathrm{d}r^2}{\Delta} + \mathrm{d}\theta^2\right) - \Big(r^2 + a^2\Big)\sin^2\theta\,\mathrm{d}\phi^2\,, \quad (6.2)$$

where M is the mass of the black hole, a is related to the angular momentum J of the black hole as $a = J/M$, and

$$\rho^2 = r^2 + a^2\cos^2\theta$$
$$\Delta = r^2 - 2Mr + a^2.$$

It is also useful to define the dimensionless spin parameter $\chi = a/M$. In the $M = 0$ limit the metric reduces to the Minkowski metric, with the relation between the Cartesian and Boyer-Lindquist coordinates given by

$$\begin{aligned} x &= \sqrt{r^2 - a^2}\sin\theta\cos\phi \\ y &= \sqrt{r^2 - a^2}\sin\theta\sin\phi \\ z &= r^2\cos\theta. \end{aligned} \quad (6.3)$$

The metric expressed in the Boyer-Lindquist coordinates becomes singular at multiple different locations. Of these only the singularity at $r = 0$ is a physical curvature singularity which cannot be removed by coordinate transformations. The other singularities are the standard singularity of spherical coordinates at $\theta = 0, \pi$, and the singularities occurring when $\Delta = 0$. The equation $\Delta = 0$ has two solutions given by

$$r_\pm = M\Big(1 \pm \sqrt{1 - \chi^2}\Big), \quad (6.4)$$

corresponding to the outer and inner event horizon. The outer event horizon can be described as a surface beyond which no signals can escape to the outside, while the inner horizon is physically irrelevant as it is unobservable and unstable. For $\chi^2 > 1$, the event horizons vanish, leaving the curvature singularity exposed. This is usually considered unphysical.

While the relatively simple expression for the metric in Boyer-Lindquist coordinates is preferred for analytic computations, the coordinate singularities are problematic for numerical work. A pair of coordinate systems that are more well behaved are the outgoing and ingoing Cartesian Kerr-Schild coordinates $(\tilde{t}, x, y, z)$, which are related to the Boyer-

Lindquist coordinates by the transformations (using the sign convention of Carter (1968))

$$\begin{aligned} \mathrm{d}t &= \mathrm{d}\tilde{t} \pm \frac{2Mr}{\Delta}\,\mathrm{d}r \\ \mathrm{d}\phi &= \mathrm{d}\varphi \pm \frac{a}{\Delta}\,\mathrm{d}r \\ x + iy &= (r \mp ia)e^{i\varphi}\sin\theta \\ z &= r\cos\theta \end{aligned}$$

with the upper sign corresponding to the outgoing and lower sign to the ingoing coordinates. In these coordinates the metric takes the form

$$g_{\mu\nu} = \eta_{\mu\nu} - \frac{2Mr}{\rho^2} l_\mu l_\nu, \tag{6.5}$$

where $\eta_{\mu\nu}$ is the Minkowski metric, r is regarded as a function of the coordinates, and

$$l_\mu \mathrm{d}x^\mu = \frac{r(x\mathrm{d}x + y\mathrm{d}y) - a(x\mathrm{d}y - y\mathrm{d}x)}{r^2 + a^2} + \frac{z\mathrm{d}z}{r} + \mathrm{d}\tilde{t} \tag{6.6}$$

for the ingoing coordinates and

$$l_\mu \mathrm{d}x^\mu = \sqrt{\frac{r^2 - z^2}{r^2(x^2 + y^2)(a^2 + r^2)}}(r(x\mathrm{d}x + y\mathrm{d}y) + a(x\mathrm{d}y - y\mathrm{d}x)) + \frac{z\mathrm{d}z}{r} - \mathrm{d}\tilde{t} \tag{6.7}$$

for the outgoing coordinates. The vectors l_μ satisfy $g^{\mu\nu}l_\mu l_\nu = \eta^{\mu\nu}l_\mu l_\nu = 0$, and l^μ are tangent to null geodesics that are respectively ingoing or outgoing, with the ingoing geodesics satisfying $r \to \infty$ as $\tilde{t} \to -\infty$ and the outgoing geodesics satisfying instead $r \to \infty$ as $\tilde{t} \to \infty$.

Both the outgoing and ingoing coordinates remain well behaved at the radius of the event horizon, but a more detailed analysis shows that the regions beyond the outer event horizon in the two coordinate systems are actually different regions of the manifold, and that there are in fact two different outer event horizons. The ingoing coordinates are regular over the future event horizon, through which future directed timelike and null geodesics can only pass in the inward direction, while the outgoing coordinates go through the past event horizon, through which timelike and null geodesics can only pass in the outward direction. The different regions covered by the charts are illustrated in figure 6.2.

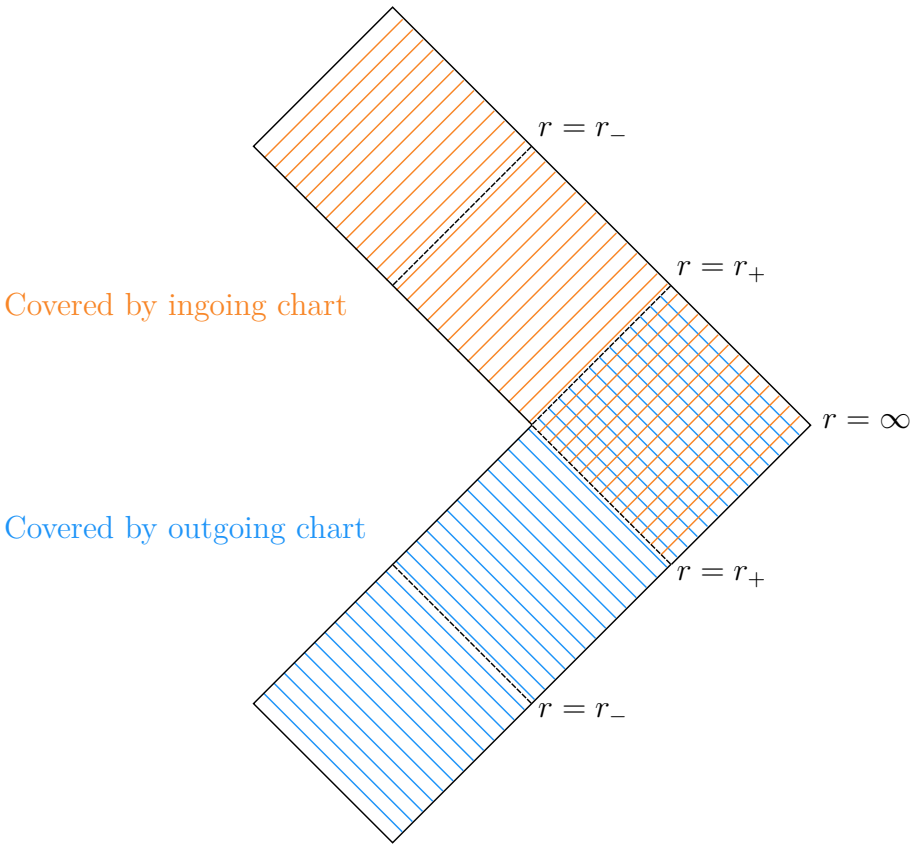

Figure 6.2.: A schematic conformal diagram showing the regions covered by the out- and ingoing Kerr-Schild coordinate charts, indicated by coloured lines. The Boyer-Lindquist chart covers the intersection of the Kerr-Schild charts, which is also the exterior of the hole. Based on Boyer and Lindquist (1967).

6.2.2. Accuracy of Geodesics and Parallel Transport

This section focuses on the accuracy and correctness of the numerical solutions to the geodesic and parallel transport equations implemented in Arcmancer. The Kerr spacetime is well suited for this purpose, as it has several symmetries corresponding to conserved quantities. These conserved quantities are important, as they can serve as a proxy for estimating the accuracy of the other computed quantities, such as the coordinates of the curve points. Directly establishing the accuracy of the curve coordinates is in general difficult, as analytical solutions of the geodesic equation are typically very complicated and often only given in terms of various special functions, if they exist at all. On the other hand, this difficulty of handling analytic solutions is a sign that the Kerr spacetime is sufficiently non-trivial, so that the correctness and accuracy of the numerical solutions can be safely assumed to generalize to other spacetimes as well.

There are four constants of motion which characterize a geodesic in the Kerr spacetime. The first one is the square norm of the tangent vector $k^\mu k_\mu$, which is also conserved along geodesics on all manifolds. Since the metric in Boyer-Lindquist coordinates (6.2) does not depend on the coordinates t and ϕ, the basis vectors ∂_t and ∂_ϕ are also Killing vectors, corresponding to the conservation of the k_t and k_ϕ components of the tangent k of a geodesic. The fourth constant of motion is the Carter's constant C (Carter, 1968), which can be written as

$$C = k_\theta^2 + k_\phi^2 \cot^2\theta - a^2 \cos^2\theta (k_t^2 - k^\mu k_\mu). \tag{6.8}$$

Comparison to a Semi-Analytic Solution

It is useful to first consider a special case where the analytical expressions can be handled relatively easily. This allows directly validating the computed coordinates and tangent vector components, as well as establishing a correspondence between the errors in the conserved quantities and the coordinates. When $a = 0$, the Kerr metric reduces to the Schwarzschild solution, which is spherically symmetric. As a result all geodesics can be reduced to lie on the equatorial plane $\theta = \pi/2$. From the normalization of the four-velocity of a massive particle $u^\mu u_\mu = 1$ it follows that

$$\left(\frac{\mathrm{d}r}{\mathrm{d}\tau}\right)^2 = u_t^2 - \left(1 - \frac{2M}{r}\right)\left(\frac{u_\phi^2}{r^2} + 1\right), \tag{6.9}$$

for further details see e.g. Carroll (2004). This equation does not lead to a simple expression for the coordinates $x^\mu(\tau)$, but it and similar equations for the other coordinates in terms of r can be integrated numerically independently of ARCMANCER's implementation. The result for $\tau(r)$ can likewise be numerically inverted to yield $r(\tau)$, which allows converting the results for $t(r)$ and $\phi(r)$ to functions of τ.

Figure 6.3 shows the result of comparing ARCMANCER's output to the results of a numerical solution of equation (6.9), which is computed with a tolerance close to the limit of floating point precision. The results from ARCMANCER were computed at various integrator tolerances, with all calculations using the ingoing Kerr-Schild coordinates, which was automatically selected by the chart selection heuristic. Clearly the solution computed by ARCMANCER is correct to the accuracy expected at the set tolerances, with the total error mostly increasing steadily due to the accumulation of error at each step. As expected, the magnitude of the errors scales proportionally to the tolerance, which can be seen from the fairly equal spacing of the different curves. The error in

most quantities is also far lower than the set tolerances. This is also expected, as the stepsize selection is based on the maximum error in any one of the propagated quantities. The error in the conserved u_t and u_ϕ is of the same order of magnitude as the error in the coordinates, so the conserved quantities appear to be good proxies for the overall accuracy of the solution.

Image Computation and the Effect of the Choice of Coordinate Chart

It is now possible to evaluate the accuracy of the integration in a more complicated situation using the conserved quantities. As the main application of ARCMANCER is radiative transfer, it is of course interesting to look at the accuracy at different parts of an image of the vicinity of the black hole. It is also of interest to compare the accuracy of the computations performed in different coordinate charts to see what kind of an effect it has. It seems that such comparisons have not been done to any great extent, most likely due to the fact that using different coordinate systems has been fairly difficult in other codes.

Figures 6.4 and 6.5 show errors in various conserved quantities and the total number of integration steps taken in the case of null geodesics or light rays propagated backwards in time from an image plane situated far away from the black hole. The basis vectors needed for polarised radiative transfer are also parallel transported along the rays. The computations are performed in all the coordinate charts currently implemented for the Kerr spacetime. Overall the errors are fairly small, showing that the solutions to the geodesic and parallel transport equations are computed correctly and accurately in all charts, but there are nevertheless large differences between the different charts.

For most quantities the outgoing Kerr-Schild coordinates shows the best behaviour, with the errors being small and nearly constant over the whole image. The outgoing chart is also the one chosen by the automatic chart selection, so this gives evidence for the effectiveness of the selection method. In the case of the outgoing chart it appears that the changes in the error over the image can be fairly well explained simply by the differences in the step count N. As each step adds additional error to the final result, areas where more steps are needed to maintain the desired accuracy of a single step will have a higher total error. This correspondence between the total step count and the errors appears to hold fairly well also in the case of the ingoing Kerr-Schild coordinates, where the step count increases considerably near the event horizon in the centre of the image. This is related to the metric becoming singular at the past horizon.

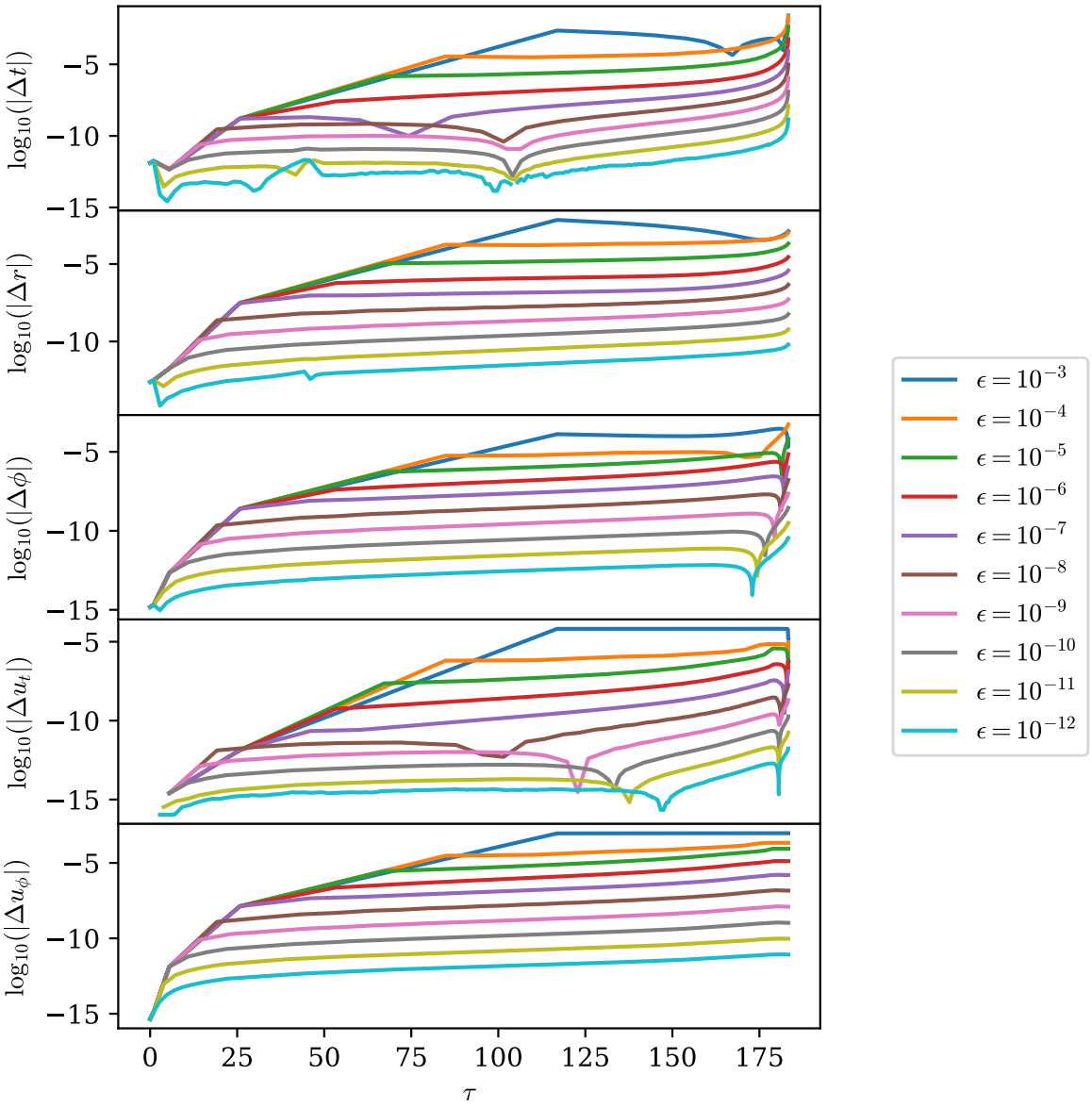

Figure 6.3.: Error in the coordinates t, r, ϕ and the conserved quantities u_t, u_ϕ as a function of proper time τ of an infalling timelike equatorial geodesic in Schwarzschild spacetime. The geodesic starts at $r = 50M$ with $u_t = 1$, $u_\phi = 3$ and terminates at $r = 2.1M$. The error is estimated by comparing ARCMANCER's results to an independent semi-analytic solution, which limits the reliability of very small errors, as is apparent from the initial errors in the coordinates. The integration tolerances used by ARCMANCER were set to $\epsilon_{\text{abs}} = \epsilon_{\text{rel}} = \epsilon$, with the values of ϵ shown in the legend.

The metric becomes singular at the horizon also in the Boyer-Lindquist coordinates, which is reflected in the larger errors in the conserved square norm of the tangent $k^\mu k_\mu$ as well as the dot product with the parallel transported basis vector $e_x^\mu k_\mu$. However, the integration step count is fairly constant and large over the whole image, with the main feature being the further increased number of steps near the polar coordinate singularities at $\theta = 0, \pi$. This suggests that the Boyer-Lindquist chart has poor numerical behaviour also far from the black hole. The Boyer-Lindquist chart also has by far the largest errors in the conserved k_ϕ and Carter's constant C. This is likely at least partly due to the magnification of smaller errors in the actually computed tangent vector components k^μ by the angular parts of the metric which are proportional to r^2 and thus very large far from the hole. The small error in k_t supports this interpretation. Regardless of the exact cause, the large errors are undesirable and suggest that the Boyer-Lindquist chart and other spherical charts should be avoided in numerical computations.

The differences in the near horizon behaviour of the different charts are further illustrated by figure 6.6, which shows the same outgoing null geodesic near the event horizon in different charts. The geodesic is almost straight in the outgoing Kerr-Schild coordinates, but coils tightly around the black hole in the other two charts. As a result the geodesic is far easier to handle numerically in the outgoing chart, and in fact impossible to represent completely in the others. Outgoing null geodesics are quite common in radiative transfer calculations in the Kerr spacetime, so using the outgoing chart there is preferable. Other charts would require some workarounds, such as stopping the geodesics some distance outside the horizon. On the other hand, for an ingoing curve the behaviour of the charts would be reversed, so that the ingoing Kerr-Schild chart would be the best choice.

Interpolation

Finally, the accuracy of the curve interpolation method, used for example during the radiative transfer computations, should be established. Figure 6.7 shows the error in the interpolated coordinates and tangent vector components compared to a solution with a much smaller stepsize. The interpolated results are fairly accurate, but the error is up to a few orders of magnitude larger than for the directly integrated points. Using tighter tolerances leads to smaller stepsizes which reduces also the interpolation error, but at when very high accuracy is required the benefit over directly integrating each point diminishes. However, typical applications do not require such extreme accuracy, and interpolation gives a considerable boost to performance compared to direct integration.

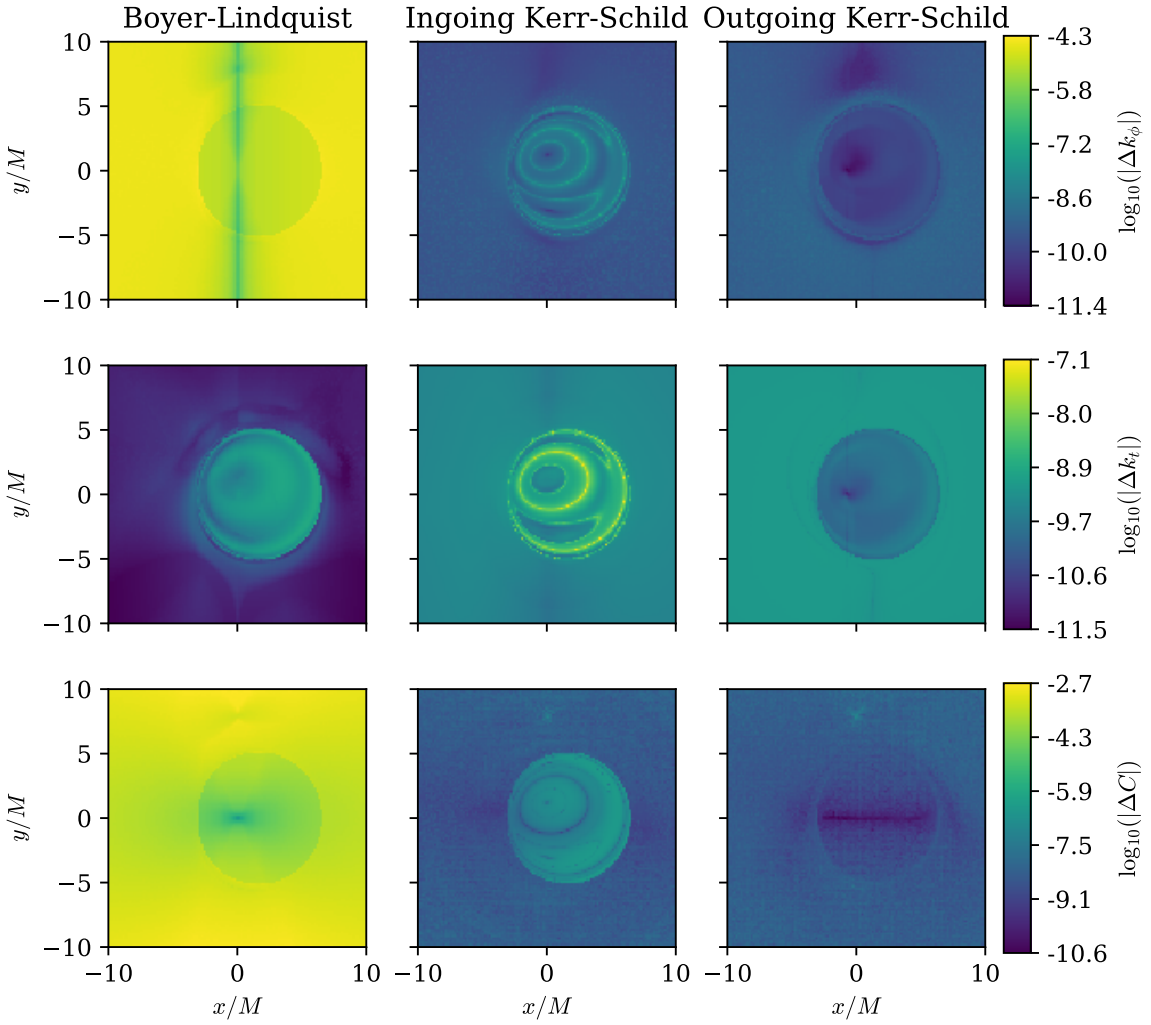

Figure 6.4.: Maximum errors in conserved quantities along rays propagated backwards from an image plane located at $r_0 = 10^6 M, \theta_0 = 50°$ computed in different coordinate charts in Kerr spacetime with spin $\chi = 0.95$. The rays were propagated until they intersected a spherical surface at $1.01 r_+$ or to an affine parameter value of $\lambda = -2r_0$, with $\lambda = 0$ initially. All computations were performed at tolerances of $\epsilon_{\text{abs}} = \epsilon_{\text{rel}} = 1 \times 10^{-10}$. The quantities shown here are the two conserved components of the tangent covector k_t and k_ϕ, as well as the Carter's constant C. Each row shows the same quantity with the same scaling, with the scales and quantities indicated at the right. Figure 6.5 shows further quantities from the same computations. The coordinates x, y are the normal coordinates of the image plane.

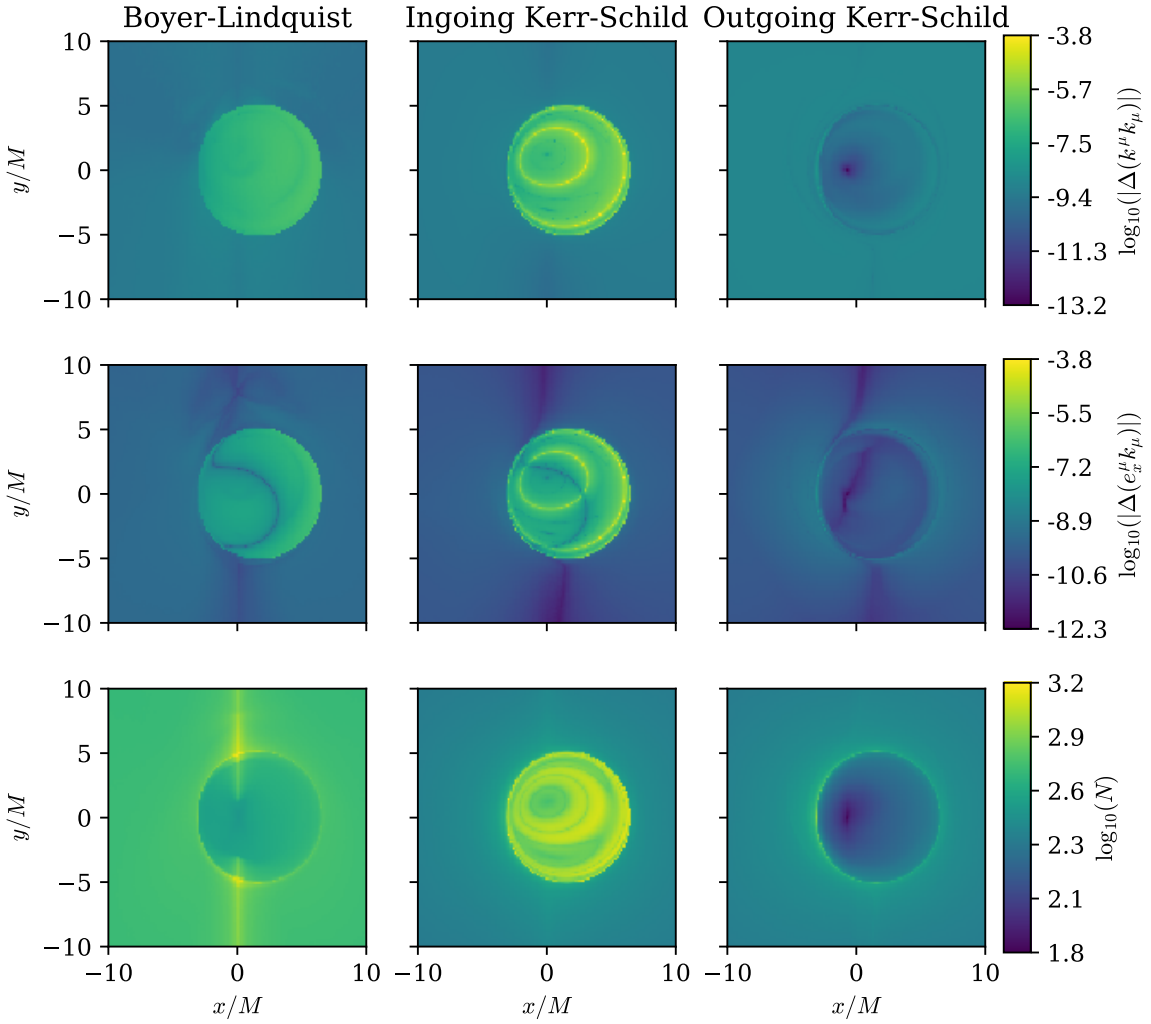

Figure 6.5.: Similar to figure 6.4, but here the error in the conserved dot products $k^\mu k_\mu$ and $e_x^\mu k_\mu$ is shown, with the last row showing the total number of steps taken N. The vector e_x is one of the basis vectors used in polarized radiative transfer computations, and is parallel transported along the ray. Like the other conserved quantities, the error in $e_x^\mu k_\mu$ is indicative of the error in the parallel transport of the vector e_x. The number of steps N is indicative of the computational load of each ray, with a lower number of steps requiring less time to compute. The error in the final result also increases with the number of steps, as only the error of a single step is controlled.

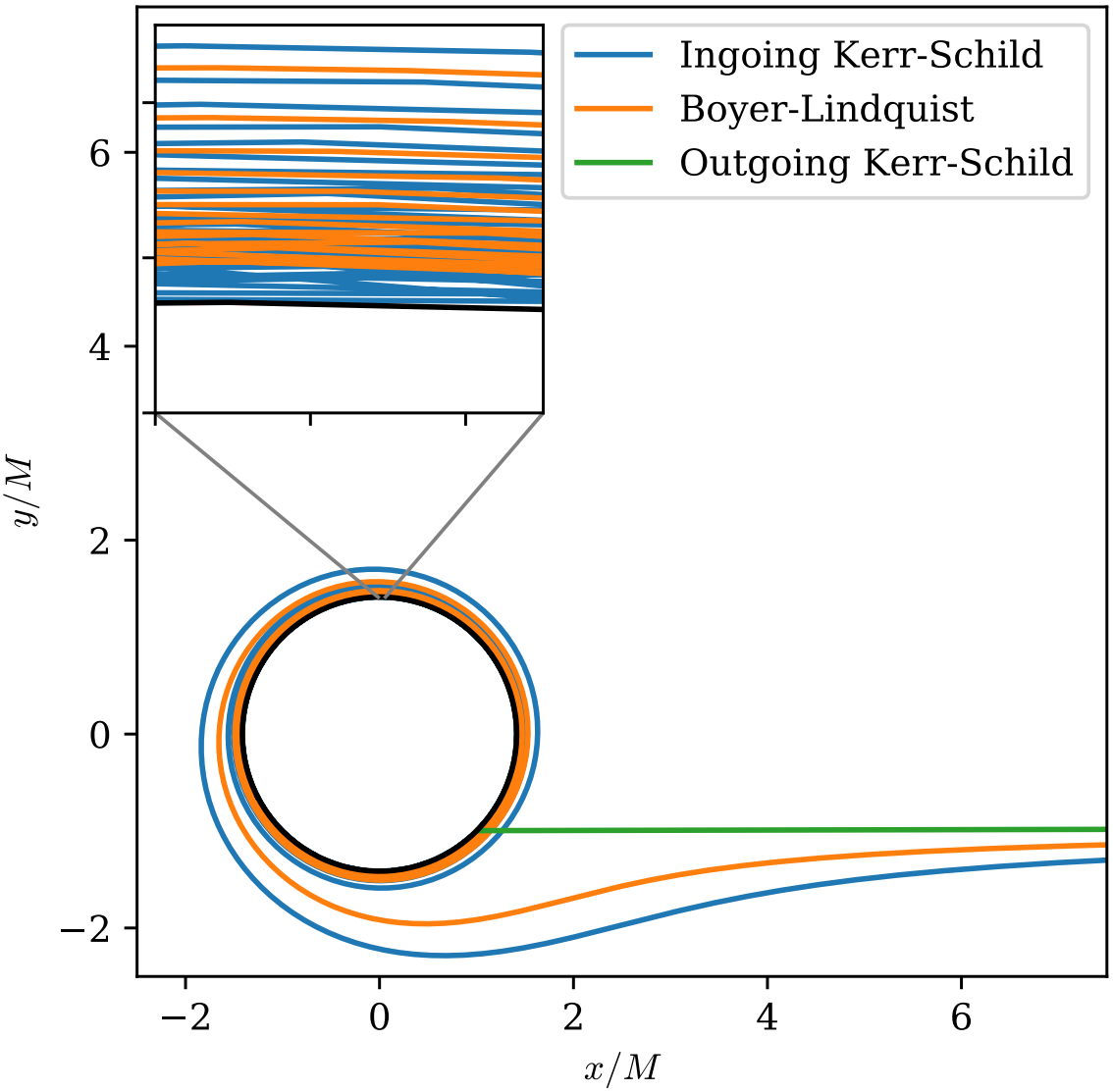

Figure 6.6.: The same outgoing null geodesic represented in different charts near a $\chi = 1$ black hole. The geodesic originates at the outer past event horizon (black), but the plotted part is limited to start just outside it. The curves are plotted in the x, y plane by simply ignoring the t coordinates, with the spherical Boyer-Lindquist coordinates converted to Cartesian coordinates using the transformation (6.3). The outgoing coordinates are adapted to radial outgoing null geodesics, so the corresponding curve (green) is almost straight, while in the two other coordinate systems the curve winds around the black hole an infinite number of times. The inset shows a zoomed in view of the curves in the ingoing Kerr-Schild (blue) and Boyer-Lindquist (orange) charts near the horizon.

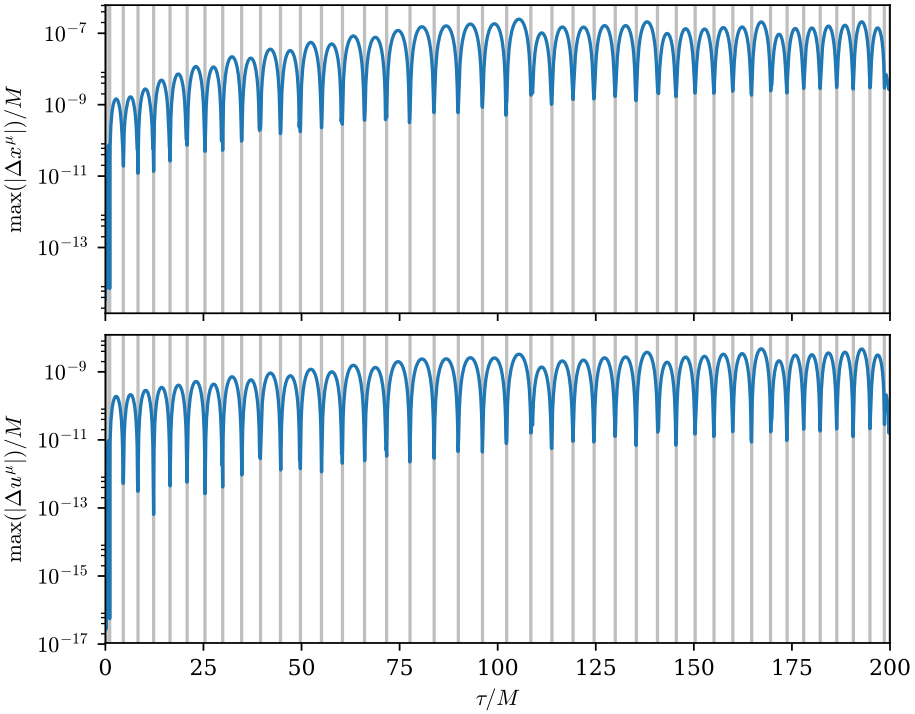

Figure 6.7.: The error in interpolated coordinates x^μ and tangent vector components u^μ along the orbit of a massive particle around a black hole with $\chi = 0.95$. The particle is at a distance of about $r = 30M$ from the hole, and the shown proper time duration τ covers roughly a third of the period of the orbit. The errors shown are the maximum errors among the components, and the interpolated curve was computed with the tolerances set to $\epsilon = 10^{-10}$. The grey vertical lines show the positions of the integrated points.

6.2.3. Accretion Disks

Having confirmed in the previous sections that the numerical solutions of the ray and radiative transfer equations are implemented correctly and produce accurate results, it is now possible to apply them to physically interesting systems. The most important radiating systems in the Kerr spacetime are accretion disks around black holes. They typically consist of heated plasma with potentially strong magnetic fields, making their dynamics extremely complicated. While there are several analytic solutions describing stationary accretion disks (see e.g. Abramowicz and Fragile, 2013, for a review), only numerical simulations can capture the full dynamics. Ray-tracing has been widely applied to post-process such simulations (e.g. Schnittman et al., 2006; Shcherbakov et al., 2012), and here an example application of Arcmancer to this task is given. However, to make interpreting the results easier, it is first useful to look at the purely geometric effects of the curved spacetime on the image of a flat disk.

Geometric Effects on the Image of a Flat Disk

The main geometric effects on the image of an accretion disk are the distortion of the image due to gravitational lensing, red- and blueshift of the emitted light and the rotation of the polarization basis vectors. The redshift of light is described by the ratio of observed and emitted frequencies $\mathcal{G} = \nu_0/\nu$, with $\mathcal{G} < 1$ corresponding to a shift to lower frequencies, i.e. redshift, and $\mathcal{G} > 1$ to blueshift. Figures 6.8 and 6.9 show the images of a rotating flat disk around a Schwarzschild black hole (spin $\chi = 0$) and a spinning Kerr black hole with spin $\chi = 0.9$, respectively, viewed from slightly above the equatorial plane at an inclination of $\theta = 80°$. The disk is implemented as a surface on the $\theta = \pi/2$ or $z = 0$ plane, with the velocity of the rotation of the surface determined by the condition for a stable circular orbit. Although here no radiation model is used to focus on the geometric effects, it is also possible to use similar surface configurations paired with a suitable radiation model to approximate geometrically thin but optically thick accretion disks. This method is employed in the example of section 6.3

In images of black holes a distinctive feature is the dark *shadow* of the black hole, which corresponds to the rays originating from within the past horizon. In figures 6.8 and 6.9 this shadow is shown as black since no observable radiation can originate from the shadow, even though here the images do not represent actual radiation. The darkness of the shadow is explained by considering actual astrophysical black holes, which are well approximated by the Kerr spacetime only long after their formation. In spacetimes corresponding to astrophysical black holes there is no past horizon, as it would be located in the far past before the formation of the hole. In such a spacetime, the rays that cross the past horizon in the Kerr spacetime would instead encounter the matter that is collapsing to form the hole. However, the radiation from this matter is redshifted to become unobservable, resulting in the black shadow.

The main geometric effects to the images of the disks are readily apparent in the two figures. Note that the whole upper surface of the disk is visible due to lensing, with the lower surface being also partially visible in the lensed image under the hole. Additionally, the shadow of the black hole appears much larger than the horizon radius of $r_+ \leq 2M$. In the case of the Schwarzschild hole there is also a thin ring of lensed images around the shadow, but for the spinning hole this feature is obscured by the disk. The image of the disk around the spinning hole is asymmetric due to the spin, with the shadow of the hole shifted to the right and the images of constant Boyer-Lindquist ϕ coordinate lines curling around the hole's shadow. However, the interpretation of the effect on the coordinate lines is somewhat ambiguous, as the coordinate system itself

is different for holes of different spin. The redshift and polarization features are quite similar for both cases, being mainly caused by the motion of the disk and the simple lensing effects, but there is a clear additional rotation of the polarization basis near the shadow of the rotating hole.

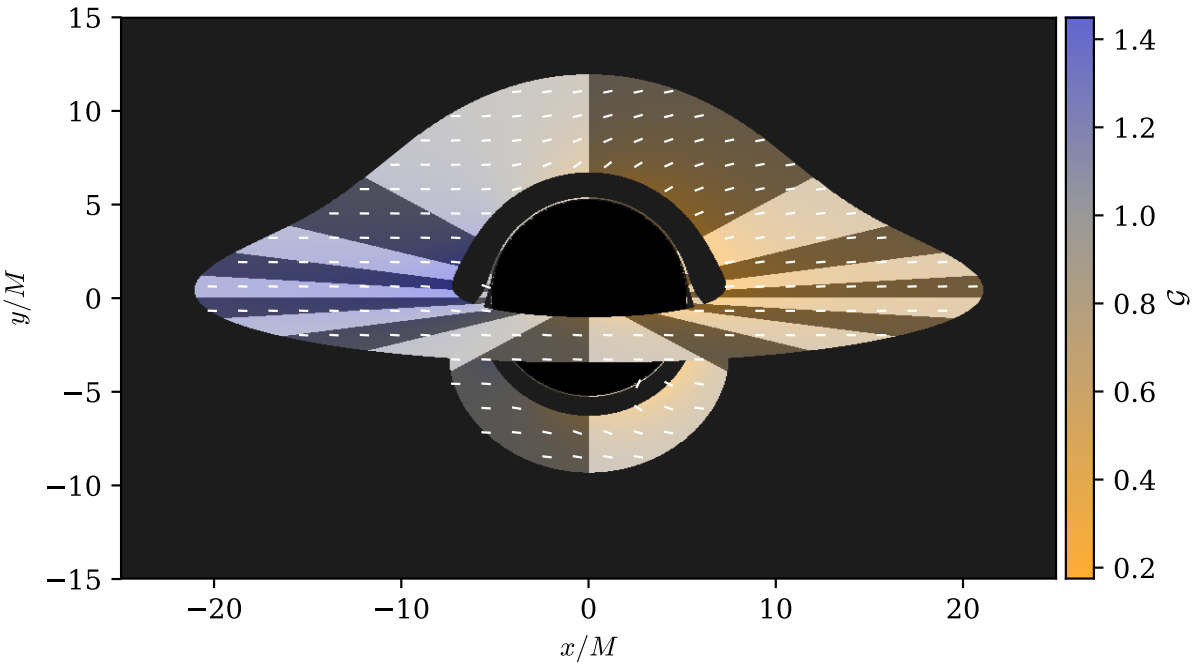

Figure 6.8.: A flat disk on the equatorial plane of a Schwarzschild ($\chi = 0$) black hole, restricted between the innermost stable circular orbit $r_{\mathrm{ISCO}} = 6M$ and $r_{\mathrm{max}} = 20M$, with a velocity corresponding to a stable circular orbit. The grey stripes are evenly spaced in the Boyer-Lindquist ϕ coordinate, while the colour shading shows the redshift $\mathcal{G}$. The scale of the redshift colouring is shown on the right. The white bars show the rotation of the standard basis defined with respect to the surface normal compared to the parallel transported basis, with horizontal bars corresponding to no rotation that would affect the polarization. The shadow of the black hole is shown in black, while the background sky is dark grey. The inclination of the observer is $\theta = 80°$.

Magnetohydrodynamic Disk

Accretion disks around black holes consists of plasma, which can be approximately modelled using the ideal magnetohydrodynamics (MHD) approximation. In this approximation the plasma is modelled as a system consisting of a continuous fluid and an electromagnetic field, which satisfy the ideal MHD condition of vanishing Lorentz force in the fluid rest frame, i.e. $u_\mu F^{\mu\nu} = 0$, where u is the fluid four-velocity. The system

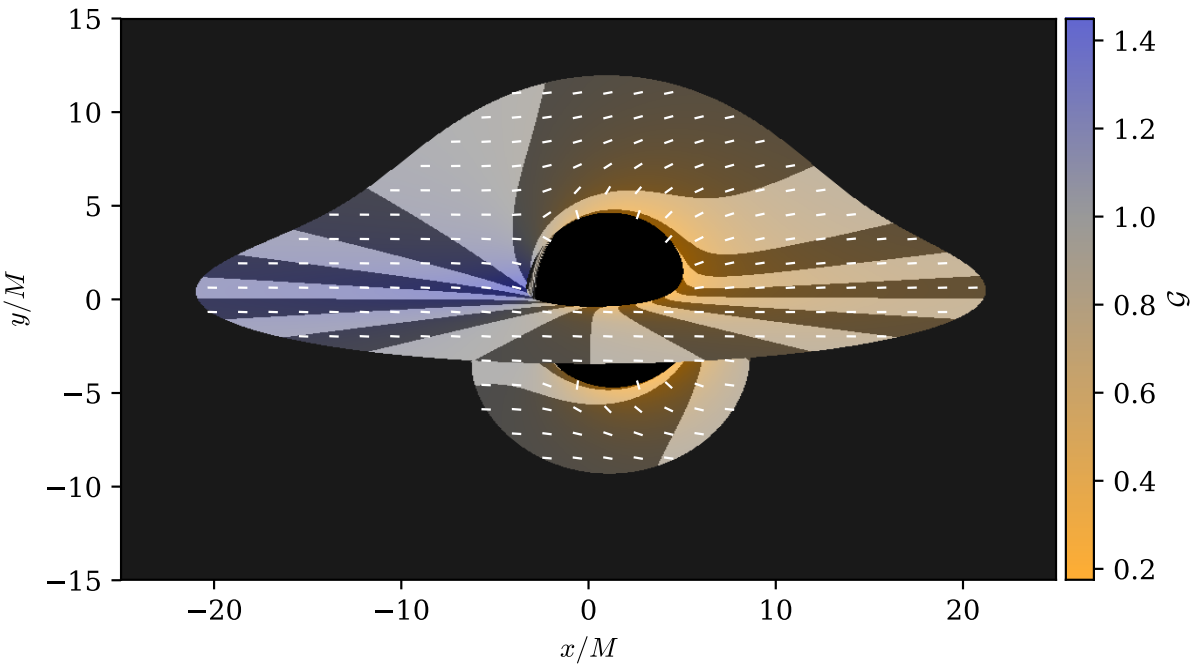

Figure 6.9.: Similar to figure 6.8, but now the black hole is spinning ($\chi = 0.9$), which causes the inner boundary of the disk to move inwards to $r_{\mathrm{ISCO}} \approx 2.32M$. The redshift colour scale is the same as in figure 6.8.

is then governed by Maxwell's equations (3.30) and (3.31), the conservation of particle number

$$\nabla_\mu(\rho u^\mu) = 0 \tag{6.10}$$

as well as the conservation law for stress-energy

$$\nabla_\mu T^{\mu\nu} = \nabla_\mu\left[\left(\rho + P + U + \frac{1}{4\pi}B^2\right)u^\mu u^\nu + \left(P + \frac{1}{8\pi}B^2\right)g^{\mu\nu} - \frac{1}{4\pi}B^\mu B^\nu\right] = 0, \tag{6.11}$$

where ρ, P, U are the density, pressure and internal energy density of the fluid, respectively, B^μ is the magnetic field four-vector of the fluid rest frame and $B^2 = -B^\mu B_\mu$. The pressure and the internal energy are additionally related by the equation of state $P = (\Gamma - 1)U$, where Γ is a constant. These general relativistic MHD (GRMHD) equations can be solved numerically, with one numerical solution algorithm implemented by the code known as HARM (Gammie et al., 2003; Noble et al., 2006).

In its default configuration, HARM solves the evolution of an accretion disk around a Kerr black hole assuming axisymmetry, which allows describing each timestep of the computation by a set of values on a two-dimensional grid. The output of HARM for each stored timestep includes for instance the pressure and internal energy of the fluid

at each point on the grid, as well as the components of the fluid four-velocity and the magnetic field four-vector. These grids of values are interpolated using linear interpolation to produce a continuous description of a fluid which can be used in radiative transfer computations with ARCMANCER. For this example only a single snapshot of HARM output is used, namely the one used by Dexter (2016) for their test of the GRTRANS code, as this data is publicly available with GRTRANS. As a consequence the evolution of the plasma during the propagation of the light is ignored. This is known as the fast-light approximation, and is valid when the propagation of the light through the plasma is much faster than the evolution of the plasma. This approximation simplifies handling the data somewhat, but the main reason for using a dataset consisting of a single snapshot is that such a dataset was easily available.

The dataset used in this example describes a turbulent torus of plasma with a mostly toroidal magnetic field. The densest area of the plasma torus is located between $r = 5M$ and $r = 20M$, and is about $10M$ thick. The data is scaled to have a black hole mass of $M = 4 \times 10^6 M_\odot$, corresponding to the central black hole Sgr A* of the Milky Way, and an accretion rate of $\dot{M} \approx 2.4 \times 10^{-11} M_\odot/\text{year}$, which are also the values used by Dexter (2016). Here $M_\odot \approx 2 \times 10^{30}\,\text{kg}$ is the solar mass. This scaling fixes the units used, but other parameters used in this example, such as the observer inclination, are chosen arbitrarily without reference to any particular existing object. With this scaling the plasma has a hydrogen number density of the order of $n \sim 10^7 \text{cm}^{-3}$ and a temperature of $T \sim 10^{11}\text{K}$, so it is fairly thin and very hot. Magnetic fields in this region have a strength of about 1 mT. With these parameters the electron plasma frequency is $\omega_p \approx 2 \times 10^2\,\text{Hz}$ and the Larmor frequency is $\omega_L \approx 2 \times 10^8\,\text{Hz}$, so based on the discussion in section 3.3.4 the approximations used in this work should hold well at least for frequencies above 10^{11}Hz. This frequency range includes for example the frequencies used by millimetre-wave radio telescopes and is thus highly relevant observationally.

In this example the emissivity and Mueller matrix coefficients are computed from the formulas for thermal synchrotron radiation given in Dexter (2016). Synchrotron radiation is emitted by relativistic electrons gyrating in a magnetic field, and is the main source of radiation in a hot plasma. In a magnetic field the motion of electrons can be approximately described as circular motion in the plane perpendicular to the magnetic field, which causes the emission of radiation, as well as a linear motion along the magnetic field direction. Relativistic effects cause the radiation to be emitted in a narrow beam around the electron's direction of motion, which leads to the emitted intensity having a dependence on the angle to the magnetic field. For high frequency radiation the emitted

intensity falls off exponentially with frequency. The emitted radiation is mostly linearly polarized in the plane of gyration, i.e. perpendicular to the magnetic field, although there is also some circular polarization.

The electrons can also absorb radiation in a magnetic field, and for a thermal distribution the strength of this absorption is given by Kirchhoff's law

$$\alpha_\nu = \frac{j_\nu}{B_\nu(T)}, \tag{6.12}$$

where α_ν and j_ν are the absorption and emission coefficients corresponding to some Stokes parameter, and B_ν is the black-body function (Rybicki and Lightman, 2008). Radiation from the much more massive ions is significantly weaker and can be safely ignored, but other processes which are ignored here, such as Compton scattering, may have an effect which would modify the results. For simplicity, the temperature of the electrons is taken to be the same as the ion temperature which is given by the HARM output. In reality the temperature of the electrons would likely differ from that of the ions, or the electrons distribution may not be thermal at all. More realistic models attempt to take this into account by, for example, computing the electron temperature as some function of the ion temperature (e.g. Mościbrodzka et al., 2016).

Figures 6.10 and 6.11 show the intensity and polarization features of the observed image at frequencies $\nu = 1.5 \times 10^{11}$ Hz and $\nu = 4 \times 10^{11}$ Hz, which correspond to radio waves with a wavelength of the order of a millimetre and are within the operational range of current radio telescopes. The observer inclination is the same 80° as in figure 6.9, and the black hole spin is likewise nearly identical at $\chi = 0.9375$, allowing easily identifying the same main geometrical features of the image. However, now the effect of redshift is even more obvious, as the intensity is strongly affected by the frequency dependent interaction between the radiation and the plasma, as well as the ν^3 dependence between the observed specific intensity and the invariant Stokes parameters. The distorted shape of the hole's shadow is now also more readily visible, due to the transparency of the optically thin plasma. The arc-like features in the strongly lensed portion of the image are caused by the axisymmetry assumption, and would likely look significantly different in a simulation of the plasma without any symmetry assumptions.

The only difference between figures 6.10 and 6.11 is their observed frequency, so the strong frequency dependence of the radiation model is clearly apparent. In the intensity images it is evident that the plasma becomes more optically thick in the lower frequencies. This is especially visible in the left side of figure 6.10 which is significantly blueshifted

and therefore samples even lower frequencies than what are observed. In these lower frequencies the depolarizing effects of the plasma are also stronger, as can be seen by looking at the areas where the polarized intensity vanishes. The direction of polarization is mostly vertical, which matches the plane of rotation of electrons in the mainly toroidal magnetic field in the plasma, but the lower frequency image also shows areas where the direction is almost horizontal, possibly due to Faraday rotation. The lensed ring around the black hole's shadow also shows strong rotation of the polarization due to geometric effects.

The frequency dependence of the observed intensity is better quantified by computing a spectrum of fluxes integrated over the whole image, which also correspond to observations where the observed object is not resolved, appearing instead point-like. Figure 6.12 shows such spectra corresponding to the different Stokes parameters, as well as the corresponding total degree of polarization and polarization angle. The fluxes are given in arbitrarily normalized units, as their numerical values depend on the angular size of the image pixels. This in turn depends on the distance between the observer and the image plane, which is completely arbitrary in this example, so choosing some particular value for it adds no new information. In the figure the intensity can be seen to reach a maximum between the frequencies of figures 6.10 and 6.11. The fact that the intensity increases even though the emitted intensity decreases in this frequency range with increasing frequency can be explained to some extent by also taking into account absorption: in this frequency range the absorption coefficients also decrease rapidly, counteracting the effect of decreased emission. The observer inclination and other properties of the system also contribute to this, as in the lower frequencies the most blueshifted regions whose intensity is the highest are partially obscured.

Another obvious feature in the spectrum is that the direction of linear polarization rotates towards purely vertical and degree of polarization increases with increasing frequency. This is as expected for this configuration, since the toroidal field causes emitted radiation to be mainly vertically polarized, and the depolarizing effects and the Faraday rotation decrease with increasing frequency. In general the different parts of the spectrum also correspond to different parts of the accretion disk. In this example this effect is not particularly obvious, but can be seen to some extent from the images at different frequencies.

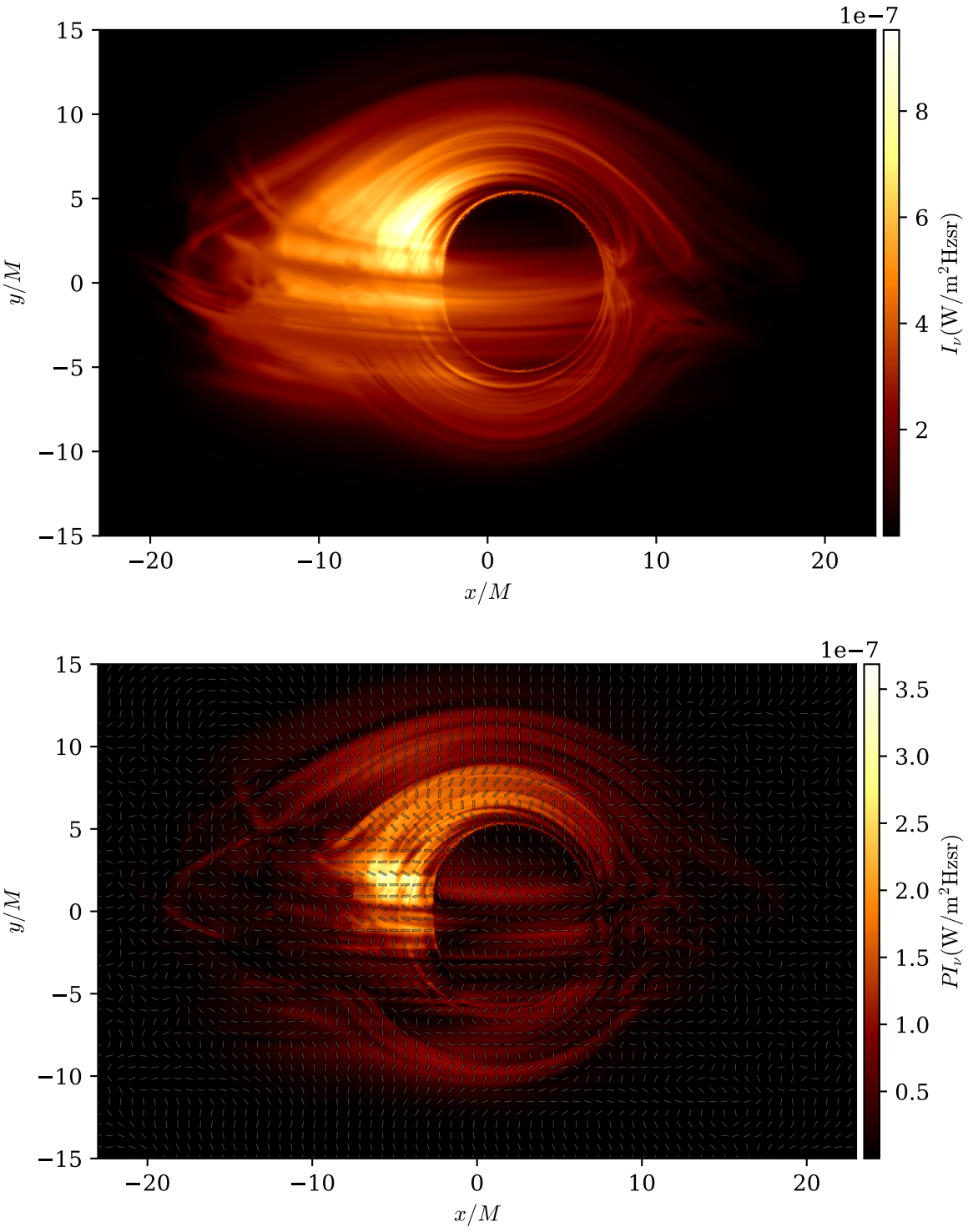

Figure 6.10.: Specific intensity I_ν at the frequency $\nu = 1.5 \times 10^{11}$ Hz (top) and polarized intensity $PI_\nu = \sqrt{Q_\nu^2 + U_\nu^2 + V_\nu^2}$ (bottom) computed from a single snapshot of a HARM simulation around a Kerr black hole with spin $\chi = 0.9375$ with an observer inclination of $\theta = 80°$. The bars in the bottom image show the direction of the principal axis of the polarization ellipse.

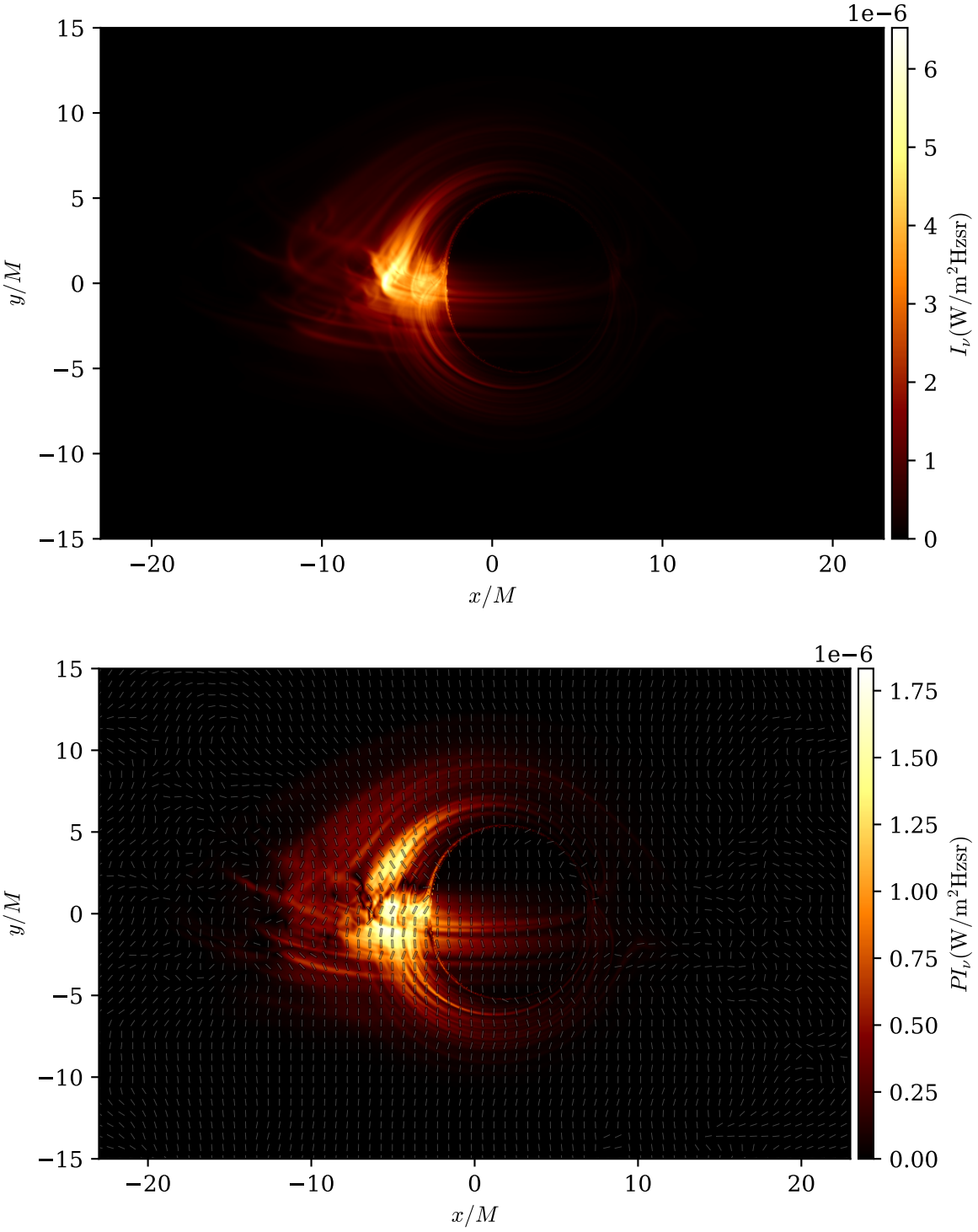

Figure 6.11.: Like figure 6.10, but now the frequency is $\nu = 4 \times 10^{11}$ Hz. Note how most of the radiation now comes from the heavily boosted inner part of the disk, which is now more clearly visible due to reduced absorption.

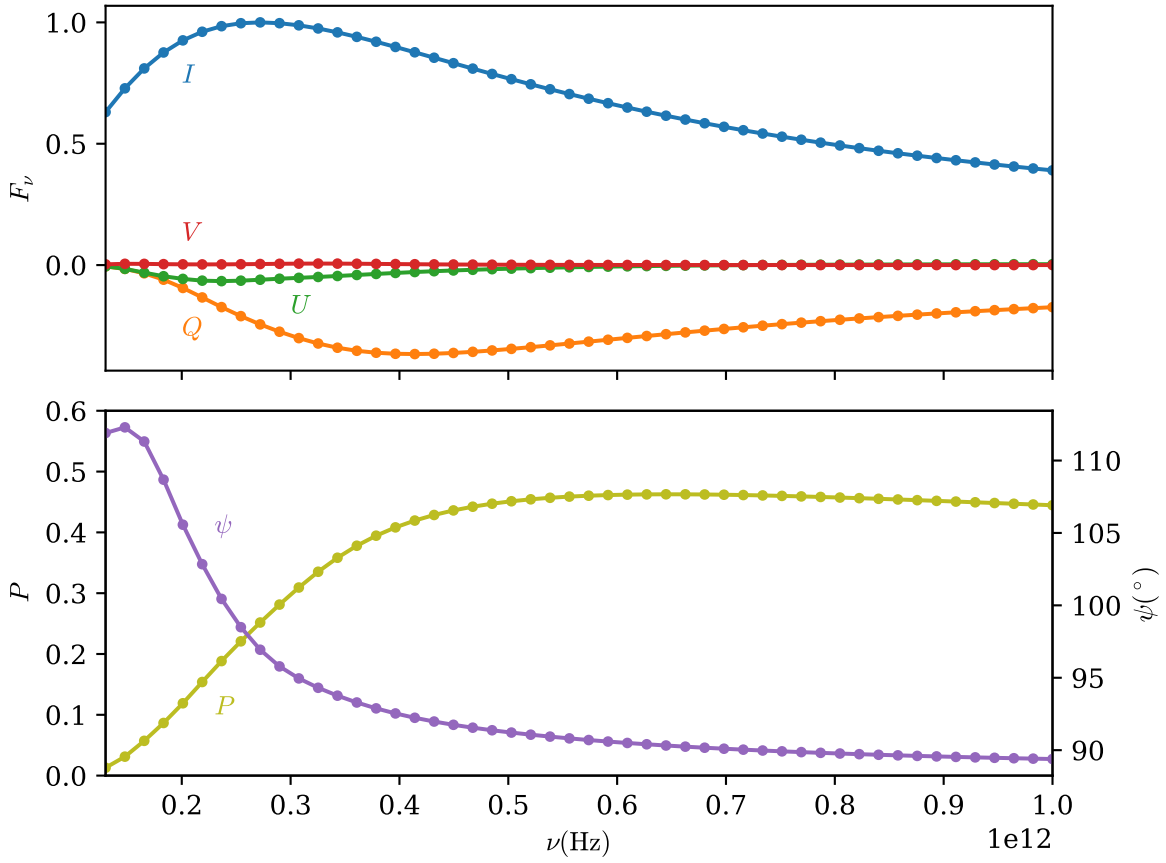

Figure 6.12.: Specific fluxes $F_{\nu,i}$, $i = I, Q, U, V$, of the Stokes parameters integrated over the image plane at different frequencies in arbitrary units (top) and the corresponding polarization degree P and the polarization angle ψ (bottom) as a function of observed frequency ν. The values are computed using the same configuration as figures 6.10 and 6.11.

6.3. Binary Black Hole Lensing

An interesting class of systems with complex gravitational effects are systems of two black holes. While completely self-consistent models of such systems would require solving the Einstein field equations as well as the magnetohydrodynamic equations governing the accreting matter, it is possible to model systems where the holes are far from each other using simple approximations. Here such an approximation is used to model the gravitational lensing effects in a binary system of non-rotating black holes. While the physical accuracy of the results is limited due to the approximations used, this example nevertheless serves well to illustrate the flexibility of ARCMANCER. The computation performed here is likely to be the first of its kind, as the various other available ray-tracing codes discussed in section 5.8 do not support it. However, the somewhat related problem of microlensing of distant active galactic nuclei, i.e. supermassive black holes, by foreground stars has been studied using other methods (e.g. Rauch and Blandford, 1991). In those studies the distances between the foreground lens and the background object are far greater than in the case considered here.

There is no known analytical solution to the two-body problem in general relativity, which has led to significant effort being put into numerical simulations of binary systems (e.g. Pretorius, 2005). While it should be possible to use the results from such a simulation with ARCMANCER, here for simplicity the metric is approximated as

$$g_{\mu\nu} = \eta_{\mu\nu} - \frac{2M}{r} l_\mu l_\nu - \frac{2m}{r'} l'_\mu l'_\nu, \tag{6.13}$$

where M, m are the masses of the primary and secondary holes, l_μ is given by (6.7) with $a = 0$ and the primed quantities are evaluated with the origin shifted by a time dependent offset giving the position of the secondary hole. This metric is simply a superposition of the non-flat parts of the Kerr-Schild forms of the Schwarzschild metric, so provided that the two holes are far enough from each other it reduces to the Schwarzschild metric near the holes. However, this does not mean that the metric solves the non-linear Einstein field equation (2.2), but the difference from an exact solution is likely to be negligible for sufficiently distant holes, due to the r^{-1} dependence of the non-flat parts. Additional effects arising from the motion of the secondary hole are ignored, apart from a trivial dependence on the $\tilde{t}$ coordinate, so in addition to the requirement that the holes be far from each other, the coordinate velocity of the secondary is required to be small for this approximation to hold. One of these motion related effects that this approximation ignores are the gravitational waves emitted by the system. More accurate analytic approximations

of the metric of a binary system are also available in the literature (Alvi, 2000), but they are considerably more complicated.

In addition to the metric, the radiation source needs also to be modelled using some approximation. In this example the source of radiation is a physically thin but optically thick accretion disk around the primary hole, which is modelled as an infinitesimally thin surface like the disk shown in figure 6.8. For this to be realistic, the gravitational effects of the secondary hole on the disk need to be negligible, which should also be the case for holes that are far from each other. The radiation is modelled as black-body radiation with the temperature computed from the Novikov-Thorne thin disk solution (e.g. Abramowicz and Fragile, 2013), which is a thin disk rotating with Keplerian velocity, i.e. the velocity of a circular orbit of a point particle. In addition to the mass of the central hole, which is set to $M = 5 \times 10^6 M_\odot$ for this example, the properties of the Novikov-Thorne disk depend on two additional parameters: the mass accretion rate of the central black hole $\dot{M}$ and a viscosity parameter α. The accretion rate is set to $\dot{M} = 0.3 L_{\text{Edd}}$, where L_{Edd} is the Eddington luminosity or limit, which is the maximum luminosity of a system where radiation pressure is balanced by gravitation. The viscosity parameter is set to $\alpha = 0.1$, which should be reasonably realistic (King et al., 2007). With these parameters the disk temperature is about $8 \times 10^5\,\text{K}$. Polarization features and angle dependent darkening of the radiation are computed assuming that the disk acts like a semi-infinite electron scattering atmosphere, which emits light partially linearly polarized parallel to the disk surface (Chandrasekhar, 1960).

Figures 6.13 and 6.14 show individual images as well as light curves from a lensing event in this binary black hole approximation. The mass ratio of the holes is $M/m = 20$, and the secondary hole passes on a linear path with respect to the coordinates nearly collinearly between the observer and the primary hole, with the distance between the two holes being $1000M$ at closest approach. The coordinate velocity of the secondary hole is $v \approx 0.03$, corresponding to the velocity of a circular orbit. This small velocity together with the fairly large distance between the holes should make the approximate metric and accretion disk used fairly accurate. As previously, the specific intensities are given with exact units, but the flux light curves are scaled arbitrarily.

The exact features of the observed images and light curves depend strongly on the exact configuration of the system, so a thorough analysis of them is not attempted here. The main notable feature of the light curves in figure 6.14 is the smooth asymmetric peak, which is caused by the gravitational lensing of the secondary hole magnifying the brightest portion of the disk, as can be seen by comparing the light curve with the

images in figure 6.13. Associated with this peak the polarization degree also increases and the polarization angle rotates, although these effects are fairly weak. The change in polarization is likely mostly due to differently polarized portions of the disk being amplified and suppressed due to the lensing, but the motion of the secondary hole may also have an effect. It is also interesting to note how far the lensing effects of the secondary hole extend, having a notable effect on the image at distances of about $100m$.

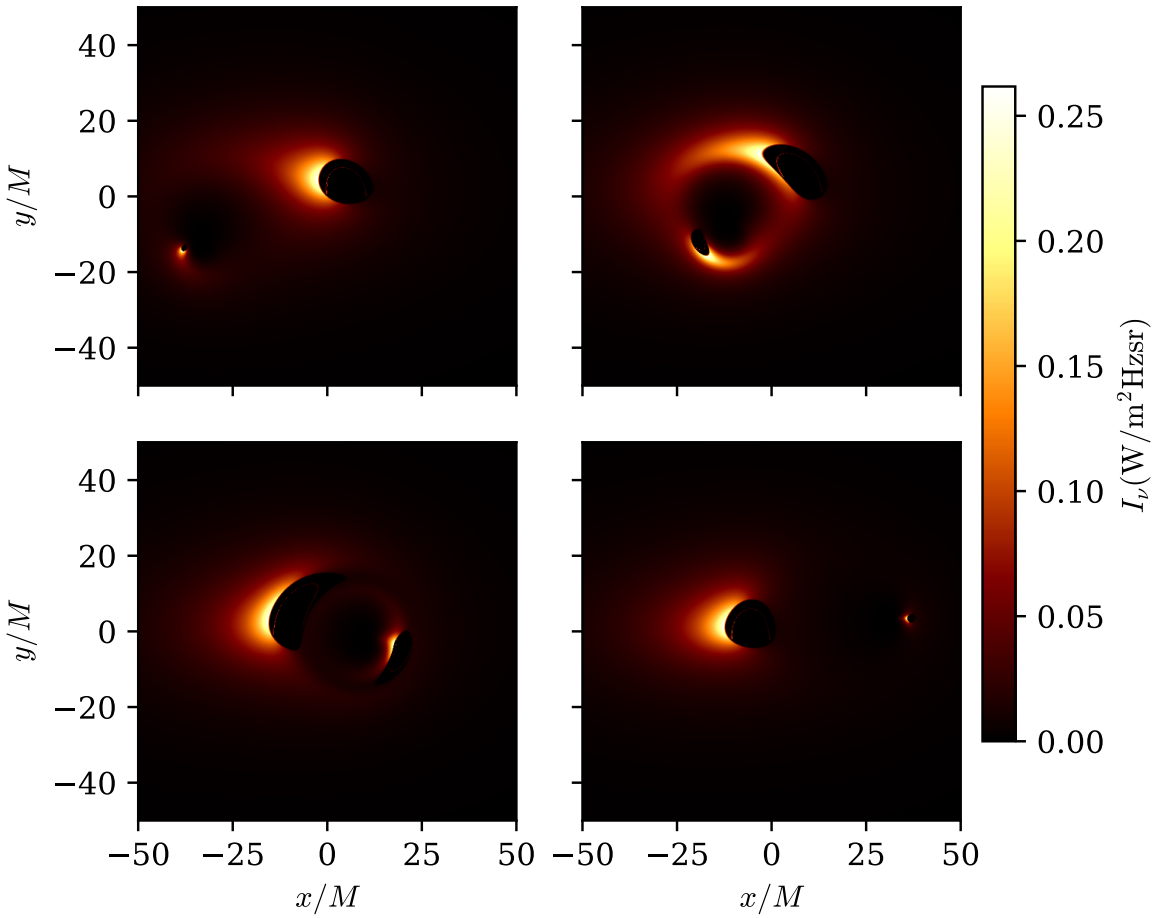

Figure 6.13.: Sequence of images from a lensing event caused by a $m = 0.05M$ black hole moving between the observer and the primary black hole with mass $M = 5 \times 10^6 M_\odot$ at a minimum distance of $r_{\min} = 1000M$. The observer inclination is $\theta = 60°$, and the images show the specific intensity observed at the frequency $\nu = 5 \times 10^{16}$ Hz, which is within the range of modern X-ray detectors. The times of the images are evenly spaced, and are indicated in the light curve of figure 6.14.

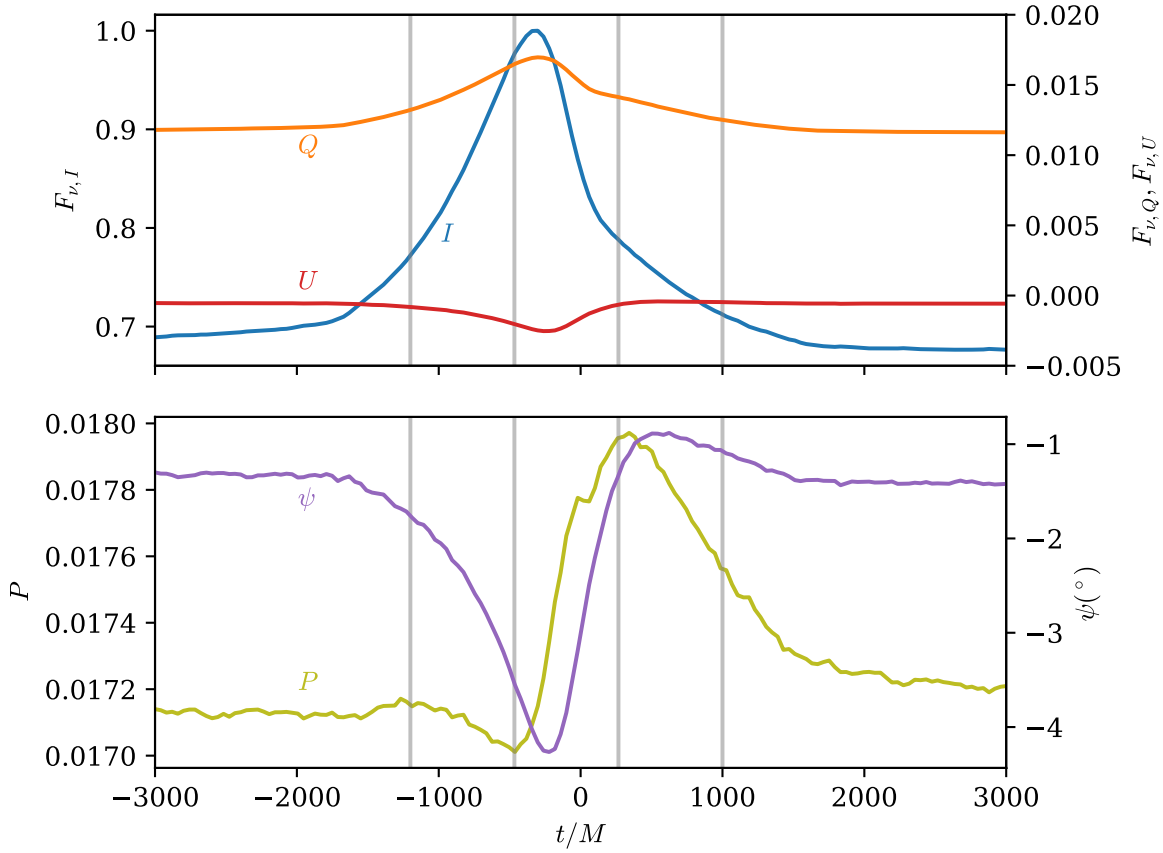

Figure 6.14.: Light curves of the fluxes corresponding to different Stokes parameters integrated over the image plane in arbitrary units (top) and the corresponding polarization degree P and polarization angle ψ (bottom) for the lensing event shown in figure 6.13. The grey vertical lines mark the times of the images shown in figure 6.13.

7. Conclusions

This thesis presents a ray-tracing based method for performing polarized radiative transfer in general spacetimes, and a code library which implements said method, Arcmancer. The ray-tracing method is justified by the geometric optics approximation, which shows that at the limit of infinite frequency radiation propagates along null geodesics in vacuum. With some additional assumptions, the radiation field in this limit can be described in terms of a kinetic distribution tensor, which describes the average intensity and polarization state of the field and satisfies a simple transport equation along the rays, or alternatively in terms of the invariant Stokes parameters which satisfy a similarly simple equation with fewer degrees of freedom. In the case where the scattering of radiation can be neglected, the transport equation can be straightforwardly solved along each ray, leading to the ray-tracing prescription which is implemented in Arcmancer.

However, even though similar ray-tracing approaches to radiative transfer have been used for a long time, it appears that the theoretical foundation of the method has not yet been made completely rigorous. While the finite frequency corrections to the idealized propagation equations have been derived in the vacuum case and can be neglected in most applications, it does not appear that similar analysis has been carried out in the presence of matter where refractive effects may become important, at least in the context of general relativity. For the same reason, in the treatment given here, the terms describing interaction with matter were added based on heuristic arguments, instead of deriving them from Maxwell's equations. As a result, there is some uncertainty to the amount of error caused by assuming that the radiation propagates as if in a vacuum, although it seems likely that the error is small in most applications where relativistic effects are important. Rigorously derived corrections to the vacuum geometric optics limit would allow correctly handling cases where such corrections cannot be ignored, or at least allow estimating the validity of the approximate computations.

In general, application of the ray-tracing method to realistic problems requires numerical computations. The Arcmancer library contains the necessary generic numerical tools for applying the ray-tracing method for radiative transfer in arbitrary

spacetimes, and is the main new contribution in this work. This library allows studying radiation in different spacetime geometries with considerably less effort than was previously possible, as other available ray-tracing implementations have been either heavily specialized to a single spacetime geometry, or lack the necessary capabilities for polarized radiative transfer. Key components of Arcmancer were implemented as a part of this work, including large parts of the differential geometry component. This new differential geometry code includes several novel features, such as support for multiple coordinate systems and tensors of arbitrary rank. Another notable feature is the programming interface, which is more closely modelled after the mathematical abstractions than is typically done in similar codes.

The key features of Arcmancer were also tested to show their correctness. The implementation of the solution of geodesics and parallel transport of vectors along them was shown to match analytical results to the configured tolerances in the case of the Kerr spacetime, which describes a spinning black hole. These results should generalize to arbitrary spacetimes, as the tests were computed employing multiple different coordinate systems, and the implementation contains no preferential treatment of any particular spacetime. It was also shown that the choice of coordinate system may have a significant effect on the computational performance and accuracy of the solved geodesics. Thus the simple interoperation between different coordinate systems which is implemented in Arcmancer may give important benefits, as the most suitable coordinate system can be easily selected depending on the problem, either manually or automatically. The support for multiple coordinate charts can also be further improved in the future, for example by improving the support for only partially overlapping charts.

The solution of the radiative transfer equation along the rays was also found to be accurate. This can also be improved on in the future, as different radiative transfer solvers that may provide even better accuracy and computational performance can be easily implemented on top of the generic capabilities for solving curves and parallel transport. These general capabilities can also be used to implement methods for dealing with scattering effects or to include the refractive effects of plasmas, provided a suitable theoretical basis for modelling those effects. Nevertheless, the current capabilities of Arcmancer already allow applying the code to a wide range of problems.

Some of these possible applications were illustrated by the example applications, which include post-processing a numerical accretion disk simulation and the computation of lensing light curves in a binary black hole system. These examples show several characteristic features of such systems, such as strong lensing of the disk image and

strong asymmetry of the disk brightness due to relativistic boosting. The results from post-processing the accretion disk simulation are comparable to those published in the past few years, while the binary black hole computation is likely to be the first of its kind. The lensing event in the binary black hole system displayed complicated behaviour in its light curve, which might allow identifying such an event from observations. However, the properties of the light curve depend strongly on the configuration of the system, and further study is required to determine if there are any consistent and unique characteristics which would aid in detection. A testament to the capabilities of Arcmancer is also the speed at which these applications were developed: for example the binary black hole computation was implemented in only a few days.

Other physical systems to which Arcmancer can be applied in the future include neutron stars, which were not discussed in this work. Applications to neutron stars benefit greatly from the generality of Arcmancer, as neutron stars are very complex systems. This complexity has for instance led to the development of several different spacetimes to approximately describe them. Some of these spacetimes have already been implemented in the library, laying the groundwork necessary for such applications. Here of great interest are also numerical solutions of neutron star spacetimes, which provide even greater accuracy, and which are expected to be usable with Arcmancer without any great complications. The comparison of observational signatures of these different spacetimes should be fairly simple to perform with the general methods described in this work. In conclusion, Arcmancer can be considered to be a significant addition to the tools available for modelling various systems where general relativistic effects are important.

Bibliography

M. A. Abramowicz and P. C. Fragile. Foundations of black hole accretion disk theory. *Living Reviews in Relativity*, 16(1):1, 2013.

K. Ahnert and M. Mulansky. Odeint – solving ordinary differential equations in C++. *AIP Conference Proceedings*, 1389(1):1586–1589, 2011.

K. Alvi. Approximate binary-black-hole metric. *Physical Review D*, 61(12):124013, 2000.

L. Barack. Gravitational self-force in extreme mass-ratio inspirals. *Classical and Quantum Gravity*, 26(21):213001, 2009.

S. Bildhauer. Transport equations for freely propagating photons in curved spacetimes: a derivation by Wigner transformation. *Classical and Quantum Gravity*, 6(8):1171, 1989a.

S. Bildhauer. Freely propagating polarized radiation in curved space-times. *Astronomy and Astrophysics*, 219:25–28, 1989b.

R. H. Boyer and R. W. Lindquist. Maximal analytic extension of the Kerr metric. *Journal of mathematical physics*, 8(2):265–281, 1967.

R. A. Breuer and J. Ehlers. Propagation of high-frequency electromagnetic waves through a magnetized plasma in curved space-time. I. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 370, pages 389–406. The Royal Society, 1980.

R. A. Breuer and J. Ehlers. Propagation of high-frequency electromagnetic waves through a magnetized plasma in curved spaces-time. II. Application of the asymptotic approximation. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 374, pages 65–86. The Royal Society, 1981.

A. Broderick and R. Blandford. Covariant magnetoionic theory - I. Ray propagation. *Monthly Notices of the Royal Astronomical Society*, 342:1280–1290, July 2003.

A. Broderick and R. Blandford. Covariant magnetoionic theory - II. Radiative transfer. *Monthly Notices of the Royal Astronomical Society*, 349:994–1008, Apr. 2004.

A. E. Broderick, T. Johannsen, A. Loeb, and D. Psaltis. Testing the no-hair theorem with Event Horizon Telescope observations of Sagittarius A. *The Astrophysical Journal*, 784 (1):7, 2014.

S. Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Addison Wesley, 2004.

B. Carter. Global Structure of the Kerr Family of Gravitational Fields. *Physical Review*, 174:1559–1571, Oct. 1968.

C. Chan, D. Psaltis, and F. Özel. GRay: A massively parallel GPU-based code for ray tracing in relativistic spacetimes. *The Astrophysical Journal*, 777(1):13, 2013.

C. Chan, L. Medeiros, F. Özel, and D. Psaltis. GRay2: A general purpose geodesic integrator for Kerr spacetimes. *arXiv preprint arXiv:1706.07062*, 2017.

S. Chandrasekhar. *Radiative Transfer*. Courier Corporation, 1960.

B. Chen, R. Kantowski, X. Dai, E. Baron, and P. Maddumage. Algorithms and programs for strong gravitational lensing in Kerr space-time including polarization. *The Astrophysical Journal Supplement Series*, 218(1):4, 2015.

P. Connors, R. Stark, and T. Piran. Polarization features of X-ray radiation emitted near black holes. *The Astrophysical Journal*, 235:224–244, 1980.

M. Degl'Innocenti and M. Landolfi. *Polarization in Spectral Lines*. Astrophysics and Space Science Library. Springer Netherlands, 2006.

J. Dexter. A public code for general relativistic, polarised radiative transfer around spinning black holes. *ArXiv e-prints*, Feb. 2016.

J. Dexter, E. Agol, P. C. Fragile, and J. C. McKinney. The submillimeter bump in Sgr A* from relativistic MHD simulations. *The Astrophysical Journal*, 717(2):1092, 2010.

S. S. Doeleman. Seeing the unseeable. *Nature Astronomy*, 1(10):646, 2017.

J. Dormand and P. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19 – 26, 1980. ISSN 0377-0427.

J. Ehlers. General-relativistic kinetic theory of gases. In *Relativistic Fluid Dynamics*, pages 301–388. Centro Internazionale Matematico Estivo, 1971.

C. F. Gammie and P. K. Leung. A formalism for covariant polarized radiative transport by ray tracing. *The Astrophysical Journal*, 752:123, June 2012.

C. F. Gammie, J. C. McKinney, and G. Tóth. HARM: a numerical scheme for general relativistic magnetohydrodynamics. *The Astrophysical Journal*, 589(1):444, 2003.

K. Gendreau and Z. Arzoumanian. Searching for a pulse. *Nature Astronomy*, 1(12):895, 2017.

K. C. Gendreau, Z. Arzoumanian, and T. Okajima. The Neutron star Interior Composition ExploreR (NICER): an explorer mission of opportunity for soft X-ray timing spectroscopy. In *Proc. SPIE*, volume 8443, page 844313, 2012.

G. Guennebaud, B. Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics. Springer-Verlag Berlin Heidelberg, 2008.

J. Hamaker and J. Bregman. Understanding radio polarimetry. III. Interpreting the IAU/IEEE definitions of the Stokes parameters. *Astronomy and Astrophysics Supplement Series*, 117(1):161–165, 1996.

M. Henon. On the numerical computation of Poincaré maps. *Physica D Nonlinear Phenomena*, 5:412–414, Sept. 1982.

L. Huang, S. Liu, Z.-Q. Shen, Y.-F. Yuan, M. J. Cai, H. Li, and C. L. Fryer. Polarized emission of Sagittarius A. *The Astrophysical Journal*, 703(1):557, 2009.

J. D. Jackson. *Classical Electrodynamics, 3rd Edition*. Wiley, 1998.

T. Johannsen and D. Psaltis. Testing the no-hair theorem with observations in the electromagnetic spectrum. II. Black hole images. *The Astrophysical Journal*, 718(1): 446, 2010.

T. Johannsen, C. Wang, A. E. Broderick, S. S. Doeleman, V. L. Fish, A. Loeb, and D. Psaltis. Testing general relativity with accretion-flow imaging of Sgr A. *Physical review letters*, 117(9):091101, 2016.

A. R. King, J. Pringle, and M. Livio. Accretion disc viscosity: how big is alpha? *Monthly Notices of the Royal Astronomical Society*, 376(4):1740–1746, 2007.

C.-W. Lau and K. M. Watson. Radiation transport along curved ray paths. *Journal of Mathematical Physics*, 11(11):3125–3137, 1970.

J. M. Lee. *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media, 1997.

R. W. Lindquist. Relativistic transport theory. *Annals of Physics*, 37:487–518, May 1966.

K. H. Lo, M. C. Miller, S. Bhattacharyya, and F. K. Lamb. Determining neutron star masses and radii using energy-resolved waveforms of X-ray burst oscillations. *The Astrophysical Journal*, 776(1):19, 2013.

B. Mashhoon. Wave propagation in a gravitational field. *Physics Letters A*, 122(6-7): 299–304, 1987.

D. B. Melrose and R. C. McPhedran. *Electromagnetic processes in dispersive media*. Cambridge University Press, 2005.

M. C. Miller and F. K. Lamb. Determining neutron star properties by fitting oblate-star waveform models to X-ray burst oscillations. *The Astrophysical Journal*, 808(1):31, 2015.

C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. W.H. Freeman and Co., San Francisco, 1973.

M. Mościbrodzka and C. Gammie. ipole-semianalytic scheme for relativistic polarized radiative transport. *Monthly Notices of the Royal Astronomical Society*, 2017.

M. Mościbrodzka, H. Falcke, and H. Shiokawa. General relativistic magnetohydrodynamical simulations of the jet in M 87. *Astronomy & Astrophysics*, 586:A38, 2016.

T. Müller and F. Grave. Motion4D–a library for lightrays and timelike worldlines in the theory of relativity. *Computer Physics Communications*, 180(11):2355–2360, 2009.

M. Nakahara. *Geometry, topology and physics*. CRC Press, 2003.

J. Nättilä and P. Pihajoki. Radiation from rapidly rotating oblate neutron stars. *arXiv preprint arXiv:1709.07292*, 2017.

S. C. Noble, C. F. Gammie, J. C. McKinney, and L. Del Zanna. Primitive variable solvers for conservative general relativistic magnetohydrodynamics. *The Astrophysical Journal*, 641(1):626, 2006.

R. Novick, M. Weisskopf, J. Angel, and P. Sutherland. The effect of vacuum birefringence on the polarization of X-ray binaries and pulsars. *The Astrophysical Journal*, 215:L117–L120, 1977.

P. Pihajoki, A. Rantala, and P. H. Johansson. Ray-tracing and polarized radiative transfer in general relativity. *Proceedings of the International Astronomical Union*, 12(S324): 347–350, 2016.

P. Pihajoki, M. Mannerkoski, et al. ARCMANCER code paper. In preparation, 2018.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C (2nd Ed.): The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1992.

F. Pretorius. Evolution of binary black-hole spacetimes. *Physical review letters*, 95(12): 121101, 2005.

K. P. Rauch and R. D. Blandford. Microlensing and the structure of active galactic nucleus accretion disks. *The Astrophysical Journal*, 381:L39–L42, 1991.

G. B. Rybicki and A. P. Lightman. *Radiative processes in astrophysics*. John Wiley & Sons, 2008.

L. Ryzhik, G. Papanicolaou, and J. B. Keller. Transport equations for elastic and other waves in random media. *Wave motion*, 24(4):327–370, 1996.

O. Sarbach and T. Zannias. The geometry of the tangent bundle and the relativistic kinetic theory of gases. *Classical and Quantum Gravity*, 31(8):085013, 2014.

S. Sasaki. On the differential geometry of tangent bundles of Riemannian manifolds. *Tohoku Mathematical Journal, Second Series*, 10(3):338–354, 1958.

J. D. Schnittman and J. H. Krolik. X-ray polarization from accreting black holes: the thermal state. *The Astrophysical Journal*, 701(2):1175, 2009.

J. D. Schnittman, J. H. Krolik, and J. F. Hawley. Light curves from an MHD simulation of a black hole accretion disk. *The Astrophysical Journal*, 651(2):1031, 2006.

R. V. Shcherbakov and L. Huang. General relativistic polarized radiative transfer: building a dynamics-observations interface. *Monthly Notices of the Royal Astronomical Society*, 410:1052–1063, Jan. 2011.

R. V. Shcherbakov and J. C. McKinney. Submillimeter quasi-periodic oscillations in magnetically choked accretion flow models of SgrA. *The Astrophysical Journal Letters*, 774(2):L22, 2013.

R. V. Shcherbakov, R. F. Penna, and J. C. McKinney. Sagittarius A* accretion flow and black hole parameters from general relativistic dynamical and polarized radiative modeling. *The Astrophysical Journal*, 755(2):133, 2012.

J. Siek, L.-Q. Lee, and A. Lumsdaine. Boost Graph Library. http://www.boost.org/doc/libs/1_58_0/libs/graph/doc/index.html, 2001.

P. Soffitta, X. Barcons, R. Bellazzini, J. Braga, E. Costa, G. W. Fraser, S. Gburek, J. Huovelin, G. Matt, M. Pearce, et al. XIPE: the X-ray imaging polarimetry explorer. *Experimental Astronomy*, 36(3):523–567, 2013.

F. H. Vincent, T. Paumard, E. Gourgoulhon, and G. Perrin. GYOTO: a new general relativistic ray-tracing code. *Classical and Quantum Gravity*, 28(22):225011, 2011.

M. Visser. The Kerr spacetime: A brief introduction. *arXiv preprint arXiv:0706.0622*, 2007.

A. Basic Concepts of Differential Geometry

This Appendix contains a brief overview of the basic concepts of differential geometry used in this work. It serves as a quick reference for some of the main definitions, but is not intended to be completely comprehensive. Textbooks dealing with the subject include Misner et al. (1973); Carroll (2004) and Nakahara (2003).

A.1. Manifolds

Differentiable manifolds are the central objects in differential geometry, and can be thought of as a generalization of geometrical objects such as surfaces embedded in euclidean space. They also serve as the foundation for the formulation of general relativity. Nakahara (2003) defines an m-dimensional differentiable manifold M as having the following properties:

1. M is a topological space.
2. M is provided with a family of pairs $\{(U_i, \phi_i)\}$.
3. $\{U_i\}$ is a family of open sets which covers M, $\cup_i U_i = M$. ϕ_i is a homeomorphism from U_i onto an open subset $U_i' \subset \mathbb{R}^m$.
4. Given U_i, U_j such that $U_i \cap U_j \neq \emptyset$, the map $\psi_{ij} = \phi_i \circ \phi_j^{-1}$ from $\phi_j(U_i \cap U_j)$ to $\phi_i(U_i \cap U_j)$ is infinitely differentiable.

The exact properties of a topological space are not important here. It suffices to say that it has enough structure to tell which points are adjacent to each other, allowing for the definition of continuous functions. A homeomorphism is simply a continuous bijection between topological spaces, whose inverse is also continuous. The third property is usually informally summarized as M being a space that looks locally like $\mathbb{R}^m$, while the fourth property corresponds to the intuitive notion of a space being smooth.

The pairs (U_i, ϕ_i) are called *charts*, with U_i being the *coordinate neighbourhood* and ϕ_i the *coordinate function*. The coordinate function can be represented as a set $\{x^a(p)\}$ of m functions $M \ni p \mapsto x^a(p) \in \mathbb{R}$, the coordinates of the point p in the chart. The set of charts $\{(U_i, \phi_i)\}$ is an *atlas*, and the *transition functions* ψ_{ij} relate the coordinates of a point in one chart to the coordinates in another chart. While computations are usually most practical to perform using some concrete chart or coordinate system, it is important to note that the points of the manifold, as well as other geometrical constructs, exist independently of the coordinates.

A.2. Vectors and Tensors

Curves

One of the simplest geometrical constructs on a manifold are curves. An open curve in a manifold M is a map $c : (a, b) \to M$, with (a, b) an open subset of $\mathbb{R}$. It is assumed that the curve does not intersect with itself, so the map c is an injection. For convenience it is also assumed that $0 \in (a, b)$, which is just a matter of shifting the parametrization of the map c suitably.

Tangent Vectors

The curves on a manifold allow for the definition of the *tangent vectors* of these curves. The directional derivative of a function $f : M \to \mathbb{R}$ along a curve c at a point $p = c(0)$ is given by

$$\left.\frac{\mathrm{d}f(c(\lambda))}{\mathrm{d}\lambda}\right|_{\lambda=0} = \frac{\partial f \circ \phi^{-1}}{\partial x^a} \left.\frac{\mathrm{d}x^a(c(\lambda))}{\mathrm{d}\lambda}\right|_{\lambda=0} = X^a \frac{\partial f}{\partial x^a} = X(f), \tag{A.1}$$

which defines the tangent vector X at the point p. More precisely X is defined using the equivalence class of curves with the same directional derivative along them at p. This definition identifies tangent vectors as linear maps from the set of smooth functions $\mathcal{F}$ to the reals $\mathbb{R}$, which give the rate of change of a function $f \in \mathcal{F}$ in the direction of the vector. These tangent vectors are elements of a vector space T_pM, the *tangent space* of the manifold at p. The partial derivatives $\frac{\partial}{\partial x^a}$ form a natural basis for this vector space, the *coordinate basis* of the chart.

Cotangent Vectors

Another useful vector space is the *cotangent space* T_p^*M, whose elements are *cotangent vectors* or one-forms. One-forms can be extended to higher order differential forms as is discussed in section 2.2. The cotangent space T_p^*M is the dual space of T_pM, meaning that the one-forms $\boldsymbol{\omega} \in T_p^*M$ are linear maps $\boldsymbol{\omega} : T_pM \to \mathbb{R}$. A one-form can be expanded in the coordinate basis as $\boldsymbol{\omega} = \omega_a \mathrm{d}x^a$, with the basis one-forms being defined by the relation

$$\mathrm{d}x^a\left(\frac{\partial}{\partial x^b}\right) = \frac{\partial x^a}{\partial x^b} = \delta^a_b, \tag{A.2}$$

where δ^a_b is the Kronecker delta, which is defined by $\delta^a_b = 1$ if $a = b$, otherwise $\delta^a_b = 0$. In addition to being maps from smooth functions to the reals, tangent vectors can also be considered to be dual vectors of one-forms, that is linear maps defined by $X(\boldsymbol{\omega}) = \boldsymbol{\omega}(X)$.

Tensors

Both tangent and cotangent vectors are special cases of more general multilinear objects, *tensors*. A tensor T of type (q, r) is a linear map from q elements of T_p^*M and r elements of T_pM to the reals. In the coordinate basis it can be expanded as

$$T = T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r} \frac{\partial}{\partial x^{a_1}} \otimes \cdots \otimes \frac{\partial}{\partial x^{a_q}} \otimes \mathrm{d}x^{b_1} \otimes \cdots \otimes \mathrm{d}x^{b_r}. \tag{A.3}$$

From this expression it is clear that the components can be extracted as

$$T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r} = T\left(\mathrm{d}x^{a_1}, \ldots, \frac{\partial}{\partial x^{b_1}}, \ldots\right), \tag{A.4}$$

so in the basis of some other chart with coordinates x'^a, the components can be expressed as

$$T'^{a_1 \ldots a_q}{}_{b_1 \ldots b_r} = T^{c_1 \ldots c_q}{}_{d_1 \ldots d_r} \frac{\partial x'^{a_1}}{\partial x^{c_1}} \cdots \frac{\partial x'^{a_q}}{\partial x^{c_q}} \frac{\partial x^{d_1}}{\partial x'^{b_1}} \cdots \frac{\partial x^{d_r}}{\partial x'^{b_r}}, \tag{A.5}$$

as $\mathrm{d}x'^a\left(\frac{\partial}{\partial x^b}\right) = \frac{\partial x'^a}{\partial x^b}$. The quantity $\frac{\partial x'^a}{\partial x^b}$ is simply the Jacobian matrix of the coordinate transformation between the two charts. This extends to non-coordinate bases by replacing the coordinate basis vectors with general basis vectors e_a and one-forms $\boldsymbol{\theta}^a$.

Writing out the basis vectors or denoting partial application of tensors quickly becomes unwieldy as the rank $q + r$ of the grows, so for most tensor computations it is convenient to use the so called abstract index notation, where $T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r}$ is taken to denote the actual tensor T and not just its components in some particular basis. In this

notation the expressions $T^a_{\ b}\, A^a$, $T(\,,A)$ and $T^a_{\ b}\, \frac{\partial}{\partial x^a} \otimes \mathrm{d}x^b \left(A^c \frac{\partial}{\partial x^c} \right)$ are therefore considered equivalent. While tensors are defined only at a single point on the manifold, they can be extended to tensor fields by assigning a tensor to every point on the manifold. In the context of general relativity tensors are generally assumed to be smooth tensor *fields*.

The Metric Tensor

A manifold M can be given additional structure by introducing a symmetric type $(0,2)$-tensor field g_{ab}, the *metric tensor*. The metric defines the inner product of vectors A and B as

$$A \cdot B = g_{ab} A^a B^b. \tag{A.6}$$

The components of the metric can be seen to be just the inner products between the basis vectors e_a, $g_{ab} = e_a \cdot e_b$. Manifolds where the inner product is positive definite are known as Riemannian manifolds, otherwise they are pseudo-Riemannian. Pseudo-Riemannian manifolds are of particular importance to general relativity, as the spacetime is modelled as one. The metric also gives the manifold a measure of distance, at least in the case of a Riemannian manifold, with an infinitesimal arc length given as

$$\mathrm{d}s^2 = g_{ab}\, \mathrm{d}x^a\, \mathrm{d}x^b\,, \tag{A.7}$$

where the $\mathrm{d}x^a$ are now interpreted as infinitesimal displacements along the curve. This allows giving a Riemannian manifold also the structure of a metric space, for which a well defined distance between any two points exists. For the pseudo-Riemannian case relevant to relativity this does not work perfectly, as for example multiple points are at a distance of zero from each other, which contradicts the intuitive notion of a distance measure. Nevertheless, the $\mathrm{d}s^2$ notation is commonly used as a shorthand when giving an explicit expression for the components of the metric.

The metric gives a natural isomorphism between the spaces T_pM and T_p^*M, by allowing the identification of a vector A and a one-form $\boldsymbol{A}$ by demanding that

$$A \cdot B = \boldsymbol{A}(B), \tag{A.8}$$

for all vectors B, which gives $A_a = g_{ab} A^b$, referred to as lowering an index. An index can be raised with the inverse metric, i.e. $A^\nu = g^{\mu\nu} A_\mu$, which leads to the relation $g^{\mu\lambda} g_{\lambda\nu} = \delta^\mu_\nu$, so the component matrix of the inverse metric $(g^{\mu\nu})$ is just the matrix inverse of $(g_{\mu\nu})$. Raising and lowering indices extends to tensors of higher rank in an obvious way.

A.3. The Lie Derivative

Vector fields can be used to define a derivative operation on general tensors that does not depend on any additional structure on the manifold. This operation is the *Lie derivative* $\mathcal{L}_X$, which takes the derivative of a tensor field along the flow of the vector field X.

The *flow* generated by a vector field X is a map $\sigma : \mathbb{R} \times M \to M$, satisfying

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^a(t, x_0) = X^a(\sigma(t, x_0)), \tag{A.9}$$

$$\sigma^a(0, x_0) = x_0^a. \tag{A.10}$$

For a fixed t, the flow gives a map $\sigma_t(x) = \sigma(t, x)$. The Lie derivative of a vector field Y along a vector field X is now defined as

$$\mathcal{L}_X Y = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left((\sigma_{-\varepsilon})_* \, Y\Big|_{\sigma_\epsilon(x)} - Y\Big|_x \right) = \left(X^a \frac{\partial Y^b}{\partial x^a} - Y^a \frac{\partial X^b}{\partial x^a} \right) \frac{\partial}{\partial x^b}. \tag{A.11}$$

Here f_* signifies the pushforward along a map f, which is given by

$$f_* \frac{\partial}{\partial y^a} = \frac{\partial x^b}{\partial y^a} \frac{\partial}{\partial x^b} \tag{A.12}$$

for a general map $f : y^a \mapsto x^a$.

The definition of the Lie derivative can be extended to general tensor fields by defining $\mathcal{L}_X g = X(g)$ for scalar functions g, and by demanding that it commutes with contraction of indices and obeys the Leibniz rule with respect to the tensor product

$$\mathcal{L}_X(T_1 \otimes T_2) = (\mathcal{L}_X T_1) \otimes T_2 + T_1 \otimes (\mathcal{L}_X T_2), \tag{A.13}$$

for arbitrary tensor fields T_1 and T_2. This leads for example to the formula (2.20) for general differential forms. It should be noted that the Lie derivative is not a directional derivative along a vector field apart from scalar functions, which can be seen from the dependence on the derivative of X in equation (A.11).

A.4. Connection and Curvature

The Covariant Derivative

To define the correct directional derivative of a tensor along a vector, the manifold needs some additional structure, a *connection*. It defines how vectors are parallel transported to a nearby point, or equivalently which elements of the vector spaces T_xM and $T_{x+\Delta x}M$ are considered to be equal as $\Delta x \to 0$. An affine connection ∇ gives the *covariant derivative* of a tangent vector Y in the direction of X as

$$\nabla_X Y = X^a \nabla_a \Big(Y^b e_b \Big) = X^a \left(\frac{\partial Y^b}{\partial x^a} + Y^c \Gamma^b{}_{ca} \right) e_b, \tag{A.14}$$

where $e_b = \frac{\partial}{\partial x^b}$ are the coordinate basis vectors, and the connection coefficients Γ are defined by

$$\nabla_{e_a} e_b = \nabla_a e_b = e_c \Gamma^c{}_{ba}. \tag{A.15}$$

The definition of the covariant derivative extends to general tensors by setting $\nabla_a f = \frac{\partial f}{\partial x^a}$ for scalar functions, and by demanding that the Leibniz rule is satisfied with respect to the tensor product, just as in the case of the Lie derivative. This gives for a general tensor T the formula

$$\begin{aligned} \nabla_c T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r} =& \frac{\partial}{\partial x^c} T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r} + \Gamma^{a_1}{}_{dc} T^{d a_2 \ldots a_q}{}_{b_1 \ldots b_r} \\ &+ \cdots + \Gamma^{a_i}{}_{dc} T^{a_1 \ldots a_{i-1} d \ldots a_q}{}_{b_1 \ldots b_r} \\ &+ \cdots - \Gamma^{d}{}_{b_i c} T^{a_1 \ldots a_q}{}_{b_1 \ldots b_{i-1} d \ldots b_r} \\ &- \cdots - \Gamma^{d}{}_{b_r c} T^{a_1 \ldots a_q}{}_{b_1 \ldots b_{r-1} d} \,. \end{aligned} \tag{A.16}$$

The object $\nabla_c T^{a_1 \ldots a_q}{}_{b_1 \ldots b_r}$ is a type $(q, r+1)$ tensor.

The Levi-Civita Connection

The connection coefficients $\Gamma^a{}_{bc}$ specify the connection completely, but their values are in principle arbitrary. However, there exists a unique connection that is metric compatible, satisfying

$$\nabla_a g_{bc} = 0, \tag{A.17}$$

and that is also torsion-free, $\Gamma^a{}_{bc} = \Gamma^a{}_{cb}$. This connection is the *Levi-Civita connection*, and its connection coefficients are given by

$$\Gamma^a{}_{bc} = \frac{1}{2} g^{ad} (\partial_b g_{dc} + \partial_c g_{bd} - \partial_d g_{bc}). \tag{A.18}$$

General relativity assumes that the connection is the Levi-Civita connection, so in other parts of this work the connection is always assumed to be the Levi-Civita connection.

Parallel Transport

The covariant derivative allows defining *parallel transport* of tensors along curves. A tensor T is said to be parallel transported along a curve with a tangent vector u if it satisfies

$$\nabla_u T = \frac{\mathrm{D}T}{\mathrm{d}\lambda} = 0, \tag{A.19}$$

where λ is the curve parameter, and the notation $\frac{\mathrm{D}}{\mathrm{d}\lambda} = u^a \nabla_a$ is introduced in analogy to $\frac{\mathrm{d}}{\mathrm{d}\lambda} = u^a \frac{\partial}{\partial x^a}$.

Geodesics

A special class of curves are affinely parametrized *geodesics*, whose tangent vectors are parallel transported along the curve

$$u^a \nabla_a u^b = 0. \tag{A.20}$$

This corresponds to the second order equation

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}\lambda^2} + \Gamma^a{}_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \frac{\mathrm{d}x^c}{\mathrm{d}\lambda} = 0, \tag{A.21}$$

which can be also obtained by extremizing the curve length

$$s = \int \sqrt{\left| g_{ab} \frac{\mathrm{d}x^a}{\mathrm{d}\lambda} \frac{\mathrm{d}x^b}{\mathrm{d}\lambda} \right|} \, \mathrm{d}\lambda \,. \tag{A.22}$$

Therefore geodesics can be considered to be the generalization of straight lines in euclidean space to general manifolds. Equation (A.21) is invariant under affine reparametrizations $\lambda' = a\lambda + b$, corresponding to a rescaling of the tangent vector $u' = au$.

Curvature Tensors

The connection is related to the curvature of the manifold, characterised by the Riemann curvature tensor $R^a{}_{bcd}$. It can be defined by

$$(\nabla_c \nabla_d - \nabla_d \nabla_c) V^a = R^a{}_{bcd} V^b \tag{A.23}$$

for arbitrary vectors V, from which it is possible to extract an explicit expression for its components. The Riemann tensor quantifies the non-commutativity of the covariant derivative, or equivalently the change in a vector after parallel transport around a small loop, which is caused by the curvature of the manifold. Other quantities related to the curvature are the Ricci tensor $R_{ab} = R_{ba} = R^c{}_{acb}$ and the Ricci scalar $R = R^a{}_a$.

A.5. Riemann Normal Coordinates

Geodesics allow explicitly constructing a local coordinate system at any point p on the manifold, in which the connection coefficients vanish at p. These coordinate systems are known as Riemann normal coordinates at p, and are an important concept in the mathematical formulation of general relativity. The construction begins by choosing the basis vectors ∂_a to give a desired form for the metric tensor g_{ab} at p. This corresponds to the process of diagonalizing the matrix of metric components. Points x in the vicinity of p have a unique geodesic connecting them and p, and thus these points can be given coordinates $x^a = k^a$, where $k^a \partial_a = \frac{\mathrm{d}}{\mathrm{d}\lambda}$ is the tangent vector at p of a geodesic γ connecting p and x, parametrized so that $\gamma(\lambda = 0) = p$ and $\gamma(\lambda = 1) = x$. In this coordinate system, curves of the form $x^a(\lambda) = \lambda k^a$ are by construction geodesics, with the tangent vector having constant components $\frac{\mathrm{d}x^a}{\mathrm{d}\lambda} = k^a$, giving $\frac{\mathrm{d}^2 x^a}{\mathrm{d}\lambda^2} = 0$. Combining this with the geodesic equation gives

$$\Gamma^a{}_{bc} k^b k^c = 0 \tag{A.24}$$

for arbitrary k at p, so the connection coefficients vanish at p. Metric compatibility of the connection also implies that the partial derivatives of the metric tensor vanish at p.

B. Code Examples

This Appendix contains some simple code examples of using ARCMANCER's Python interface to perform ray tracing and radiative transfer. These examples are kept fairly short to highlight the power of the generic tools implemented in ARCMANCER. Performing similar computations using the C++ library directly is done in a similar way and is not much more complicated. Additional, more complex examples are included with the ARCMANCER library.

B.1. Radiative Transfer

This example shows how radiative transfer computations are performed along a single ray. Here the background spacetime is flat, but the computation works in the same way regardless of the background. To compute images, this computation would need to be performed in the `compute_func` of the example in section B.2. For simplicity, here the fluid is taken to have a constant temperature and a constant velocity along the ray direction, with other fluid properties ignored. The radiation model is simply black-body emission with constant absorption. This example outputs the spectrum in figure B.1, and the details of the code are explained in the comments displayed in light blue.

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3
4  # Import the relevant parts of the Arcmancer Python module
5  from pyarcmancer import StokesVector, PolarizationSpectrum, RadiationData
6  # Geometric objects are templated based on the spacetime, so each spacetime
7  # has its own Python submodule.
8  from pyarcmancer.minkowski import *
9
10 # Setup the spacetime.
11 M = MinkowskiSpacetime()
12
13 # Point at the end of the ray.
14 x0 = ManifoldPoint(M.cartesian, # Which coordinate system is used.
15                    [0, 0, 0, 0] # The coordinates of the point.
16                   )
17 # Ray tangent. The ray comes in along the positive x-axis.
```

```
k = TangentVector(x0, # The point where the vector is defined.
                  M.cartesian, # Coordinate system used for the components.
                  [1,-1,0,0] # Components of the vector.
                 )
# The polarization basis.
pol_frame = PolarizationFrame(
                TangentVector(x0, M.cartesian, [0,0,0,1]), # vertical
                TangentVector(x0, M.cartesian, [0,0,-1,0]) # horizontal
            )
# Construct the ray.
# Geodesic is a version of ParametrizedCurve with external forces disallowed.
# The name of the type mirrors the C++ template syntax
# Geodesic<MinkowskiSpacetime, PolarizationFrame>,
# which defines the transported object type.
ray = Geodesic_PolarizationFrame(k, pol_frame)
# Extend the ray backwards in time.
ray.compute(-100)

# Define the fluid and radiation models.
# Constructs the fluid model for given temperature and velocity.
def fluid(T, v):
# Returns a function that satisfies the fluid model interface,
# i.e. computes the properties of the fluid at a point X.
    return lambda X: \
        FluidData(1, # Number density, ignored in this example.
                  T,
                  0, 0, # Pressure and magnetic field strength, also ignored.
# The direction of the magnetic field / polarization reference direction.
                  TangentVector(X, M.cartesian, [0,0,0,1]),
# The four-velocity of the fluid.
                  normalized(TangentVector(X, M.cartesian, [1,v,0,0]))
                 )

# Radiation model.
# Arguments: fluid data, rest frame frequency of radiation nu,
# angle between ray and the reference direction th.
def radiation(fluid_data, nu, th):
# Black-body emission
    B_nu = 4 * np.pi * nu**3 / np.expm1(2 * np.pi * nu/fluid_data.temperature)
    j_I = B_nu
    j_Q = .5*j_I

# Constant absorption coefficients
    a_I = 1
    a_Q = .1 # Does not follow Kirchoff's law to get net polarization.

# Other emissivities and Mueller matrix components are set to vanish.
    j_U = j_V = a_U = a_V = r_Q = r_U = r_V = 0
    return RadiationData(j_I, j_Q, j_U, j_V,
                         a_I, a_Q, a_U, a_V,
                         r_Q, r_U, r_V,
                         1 # Unit system scaling factor.
                        )
```

```
71
72  # Initial spectrum of invariant Stokes vectors, all vanish.
73  # This determines also the observed frequencies.
74  frequencies = np.linspace(.1, 20, 100)
75  initial_vals = PolarizationSpectrum(
76                      [StokesVector(nu, [0,0,0,0]) for nu in frequencies])
77
78  # Compute radiation transfer for two fluid velocities.
79  res1 = radiation_transfer(ray, # Compute transfer over this ray.
80                            k.components(M.cartesian)[0], # Normalization factor.
81                            initial_vals,
82                            fluid(10, 0),
83                            radiation)
84
85  res2 = radiation_transfer(ray, k.components(M.cartesian)[0], initial_vals,
86                            fluid(10, .5), radiation)
87
88  # Extract the observed Stokes vectors.
89  I1 = np.array([sv.rest_frame_IQUV_vector for sv in res1.stokes_vectors])
90  I2 = np.array([sv.rest_frame_IQUV_vector for sv in res2.stokes_vectors])
91
92  # Plot the results.
93  plt.plot(frequencies, I1[:,0], label='$I_1$')
94  plt.plot(frequencies, I1[:,1], label='$Q_1$')
95  plt.plot(frequencies, I2[:,0], label='$I_2$')
96  plt.plot(frequencies, I2[:,1], label='$Q_2$')
97
98  plt.xlim(.1,20)
99  plt.xlabel(r'$\nu$')
100 plt.ylabel(r'$I,Q$')
101 plt.legend()
102 plt.savefig('rad_transfer.pdf', bbox_inches='tight')
```

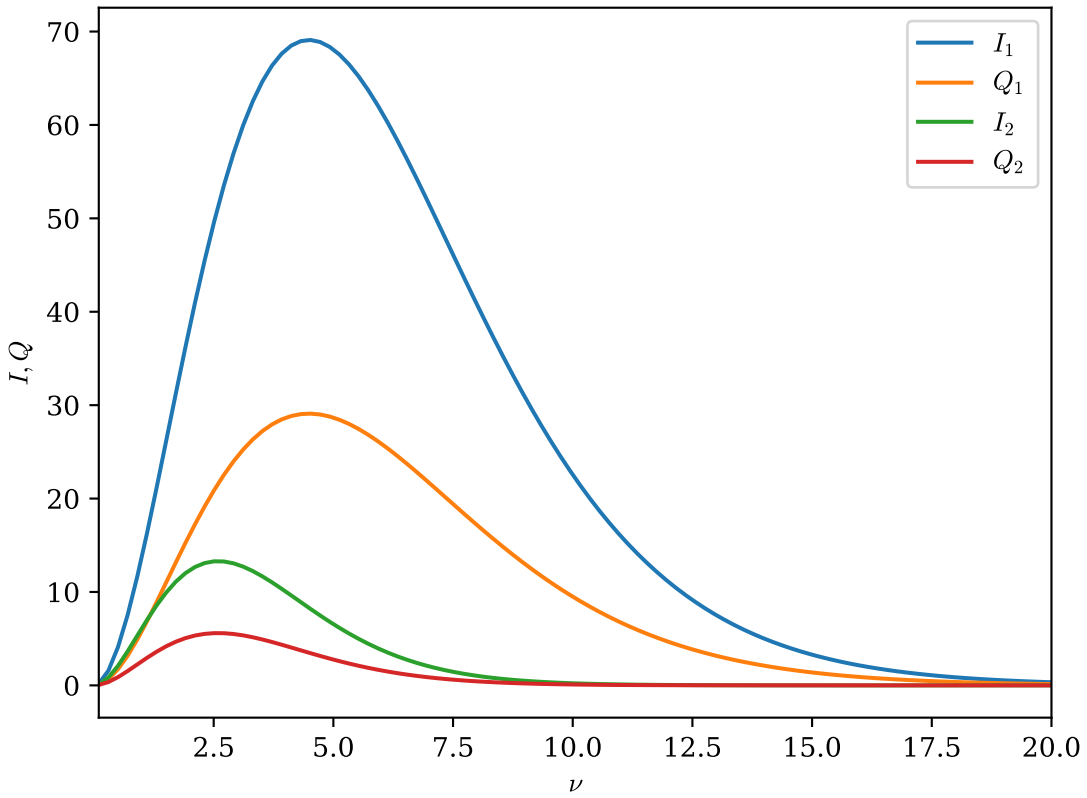

Figure B.1.: The spectrum produced by this example. The intensities labelled I_1, Q_1 are computed for stationary fluid, while I_2, Q_2 are computed for a fluid receding at a velocity of $v = .5$, resulting in redshift of the spectrum.

B.2. Image of the Kerr Black Hole Shadow

This simple example computes the image of the shadow of a Kerr black hole, which corresponds to the rays of light which would need to originate inside the past event horizon. Additionally, a grid based on the Boyer-Lindquist θ, ϕ coordinates is drawn on the background sky. The script outputs the image in figure B.2, and runs in approximately 4 minutes on a basic laptop computer in a single thread for an image of 400×400 pixels. Due to the limitations of Python, parallel computation of images is not possible through the Python interface. Note that the runtime scales linearly with the total number of pixels.

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 # Import the Arcmancer python module.
5 import pyarcmancer
6 from pyarcmancer.kerr import *
```

```
7
8  # Silence some logging we're not interested in.
9  pyarcmancer.Log.set_level(pyarcmancer.LogLevel.notice)
10
11 # Setup the spacetime.
12 chi = .99
13 M = KerrSpacetime(1, chi)
14
15 # Use lower tolerances and disable automatic chart switching for a faster
16 # calculation.
17 M.configuration.absolute_tolerance = 1e-6
18 M.configuration.relative_tolerance = 1e-6
19 M.use_fixed_chart(M.kerr_schild_out)
20
21 # Setup the observer and image plane
22 r0 = 5000 # image plane distance
23 inc = 90  # inclination in degrees
24
25 # Point at the centre of the image plane
26 x0 = ManifoldPoint(M.boyer_lindquist, [0, r0, np.radians(inc), 0])
27
28 # The local frame defining the image plane directions.
29 obs_frame = LorentzFrame(x0, # Location of the frame.
30                          M.boyer_lindquist,
31                          [1,0,0,0], # t basis vector components
32                          [0,1,0,0], # z basis vector components
33                          [0,0,0,1]  # x basis vector components
34                         )
35 # Number of pixels
36 xbins= 400
37 ybins = xbins
38 # Half-widths of the image plane
39 xspan = 10
40 yspan = ybins/xbins*xspan # Ensure square pixels
41
42 # Construct the image plane
43 ip_shape = ImagePlane.PlaneShape(xbins, ybins, xspan, yspan)
44 image_plane = ImagePlane(ip_shape, obs_frame)
45
46 # Spherical surfaces representing the event horizon and the far away sky.
47 sky_sphere = OutgoingKerrSphere(M, 5*r0)
48 eh_sphere = OutgoingKerrSphere(M, M.horizon_r)
49
50 # The computation to perform for a single pixel.
51 # The function receives an object containing e.g. ray initial conditions as its
52 # only argument, and can perform arbitrary computations,
53 # for example radiative transfer.
54 def compute_func(ip):
55 # Construct the ray and extend it backwards in time for a long distance,
56 # or until it hits either of the surfaces.
57     ray = Geodesic(ip.ray_tangent)
58     ray.compute(-30*r0, [sky_sphere, eh_sphere])
59
```

```
# Return hit point coordinates and whether the ray escaped, i.e. hit the sky.
    th, phi = ray.front.point.coordinates(M.boyer_lindquist)[2:]
    ret = {'hit_sky': ray.front_termination.hit_surface
                      and ray.front_termination.surface_index == 0,
           'th' : th,
           'phi': phi}
    return ret

# Perform the computation.
image_plane.compute(compute_func)

# Extract the image data into arrays for plotting.
th      = np.zeros((ybins,xbins))
phi     = np.zeros((ybins,xbins))
hit_sky = np.zeros((ybins,xbins), dtype=bool)

for row in image_plane.data_points:
    for val in row:
        th[val.iy,val.ix] = val.data['th']
        phi[val.iy,val.ix] = val.data['phi']
        hit_sky[val.iy, val.ix] = val.data['hit_sky']

# Show a simple coordinate grid on the sky.
grid = .5 + .3 * np.round((1 + np.cos(5 * th) * np.sin(5 * phi)) / 2)
img = np.zeros((ybins, xbins))
img[hit_sky] = grid[hit_sky]

# Draw image and save.
plt.imshow(img, origin='lower', cmap='gray', interpolation='nearest',
           extent=(-xspan, xspan, -yspan, yspan))
plt.xlabel("$x/M$")
plt.ylabel("$y/M$")
plt.savefig("kerr_shadow.pdf", bbox_inches='tight')
```

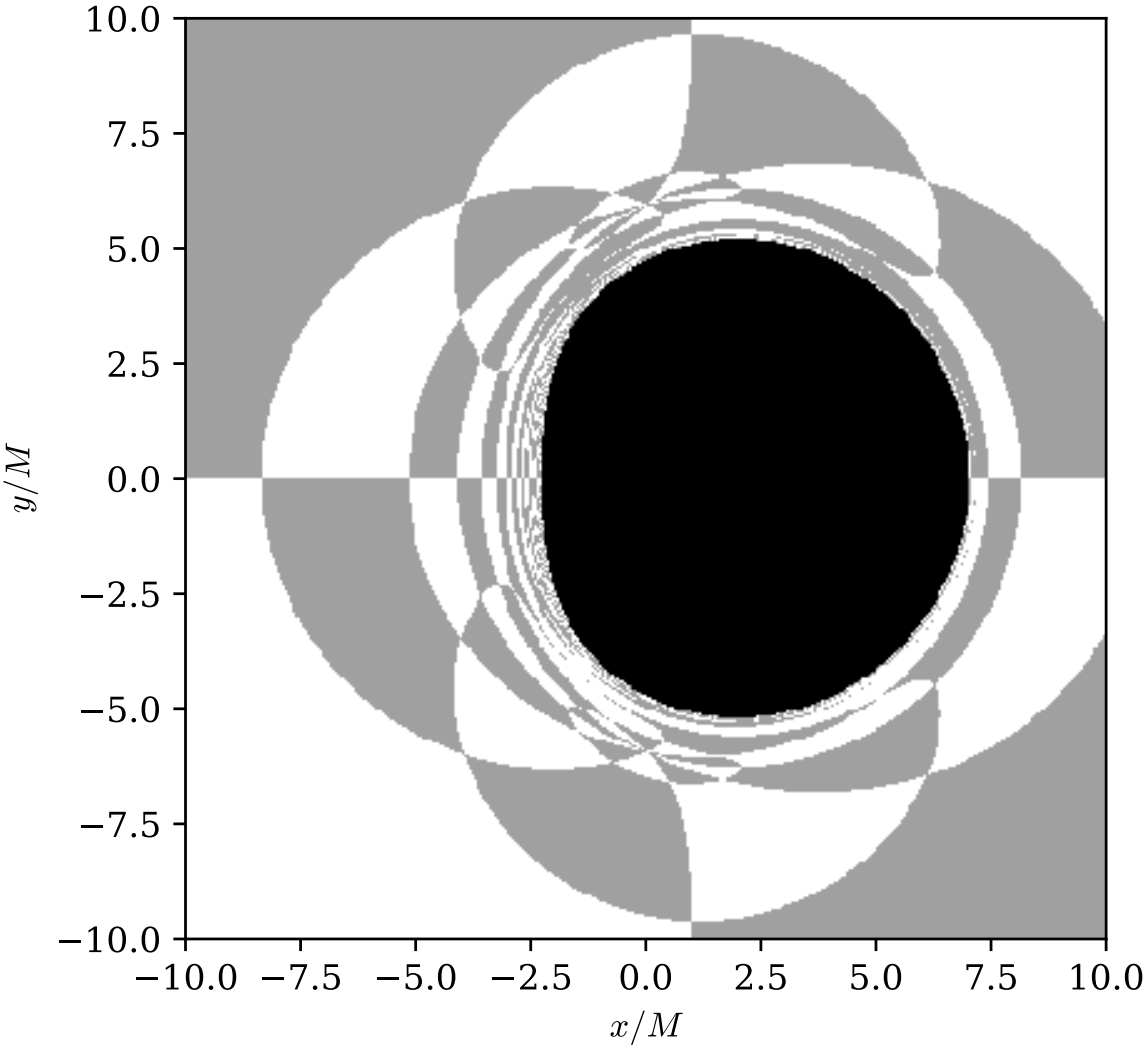

Figure B.2.: The image produced by this example. The coordinate grid placed on the sky is strongly distorted by gravitational lensing.