Self-force and radiation reaction in general relativity

Leor Barack and Adam Pound

Mathematical Sciences, University of Southampton, Southampton SO17 1BJ,
United Kingdom

Abstract. The detection of gravitational waves from binary black-hole mergers by the LIGO-Virgo Collaboration marks the dawn of an era when general-relativistic dynamics in its most extreme manifestation is directly accessible to observation. In the future, planned (space-based) observatories operating in the millihertz band will detect the intricate gravitational-wave signals from the inspiral of compact objects into massive black holes residing in galactic centers. Such inspiral events are extremely effective probes of black-hole geometries, offering unparalleled precision tests of General Relativity in its most extreme regime. This prospect has in the past two decades motivated a programme to obtain an accurate theoretical model of the strong-field radiative dynamics in a two-body system with a small mass ratio. The problem naturally lends itself to a perturbative treatment based on a systematic expansion of the field equations in the small mass ratio. At leading order one has a pointlike particle moving in a geodesic orbit around the large black hole. At subsequent orders, interaction of the particle with its own gravitational perturbation gives rise to an effective "self-force", which drives the radiative evolution of the orbit, and whose effects can be accounted for order by order in the mass ratio.

This review surveys the theory of gravitational self-force in curved spacetime and its application to the astrophysical inspiral problem. We first lay the relevant formal foundation, describing the rigorous derivation of the equation of self-forced motion using matched asymptotic expansions and other ideas. We then review the progress that has been achieved in numerically calculating the self-force and its physical effects in astrophysically realistic inspiral scenarios. We highlight the way in which, nowadays, self-force calculations make a fruitful contact with other approaches to the two-body problem and help inform an accurate universal model of binary black hole inspirals, valid across all mass ratios. We conclude with a summary of the state of the art, open problems and prospects.

Our review is aimed at non-specialist readers and is for the most part self-contained and non-technical; only elementary-level acquaintance with General Relativity is assumed. Where useful, we draw on analogies with familiar concepts from Newtonian gravity or classical electrodynamics.

Contents

1. Introduction

Black holes are "the simplest macroscopic objects in the universe", goes the famous quote from S. Chandrasekhar [1]. Indeed, we expect astrophysical black holes, when in isolation, to be described in *exact* form in terms of the 2-parameter family of

Kerr solutions to the field equations of General Relativity (GR). But put two such black holes in orbit around each other, and they form a strikingly complex dynamical system. No closed-form solutions are known, and even numerical solutions have been forbiddingly hard to obtain until well into the 21st century. The complexity of the gravitational two-body problem in GR stands in stark contrast to its elementary nature in the context of point-particle Newtonian gravity, where all possible orbital configurations are simple conical sections. Point-particle idealizations are problematic in GR, but even in scenarios where they can make sense in some effective way, the orbital dynamics remains very complicated. In classical GR, a gravitationally bound system of two masses (subject only to gravitational forces) admits no stationary configurations: gravitational waves constantly carry orbital energy away from the system, and back-reaction from that radiation gradually drives the two objects closer together. Given enough time, the two bodies eventually merge. If the two bodies are initially Kerr black holes, a single, larger Kerr black hole eventually forms. The detailed description of this inspiral and merger process has been a major computational challenge and a key theme in gravitational research for almost 50 years. A central motivation has been the desire to predict the exact pattern of gravitational waves through which such systems can manifest themselves observationally.

The landmark observation of merging black holes by the LIGO-Virgo Collaboration in 2015 [2] has conclusively established the existence in nature of black holes, inspiralling black-hole binaries, and gravitational waves, all consistent with GR. Once mathematical curiosities, black holes and gravitational waves are now firmly in the realm of observational astronomy. For the first time, we are given direct observational access to a natural process where general-relativistic dynamics plays out at its most extreme. The continued advance of *gravitational-wave astronomy* will bring unprecedented opportunities to probe relativistic theory in its most dynamical regime. The theoretical modelling of gravitational-wave sources is an integral part of that programme, as the realization of the exciting science prospects relies crucially on the availability of accurate source models.

The LIGO-Virgo discoveries can serve as a case in point: without an accurate model of the inspiral and merger it would have not been possible to extract the physical parameters of GW150914 (the first event detected and the brightest so far) at the precision with which they were reported [2], and some of the other events would have likely been missed altogether [3]. Analysis by the LIGO-Virgo Collaboration [4] concluded that the quality of science extractable from future observations may well be limited not by experimental precision but by the accuracy of available theoretical models. (This would be the case if detected mergers were to involve more rapidly spinning black holes than in the mergers already observed, or larger mass disparities between the two merging black holes.) The desire to maximize the science return from gravitational-wave experiments thus continues to drive the theory programme to improve waveform models across the full parameter space relevant to observation.

Of particular interest is the inspiral scenario where one of the black holes is much lighter than the other: the so-called "extreme-mass-ratio inspiral", or EMRI. Nature abounds with EMRIs. They come in the form of stellar-mass black holes (or neutron stars) that are captured into inspiral orbits around massive black holes—the kind of behemoth black holes, of masses of order 10^6 to 10^9 solar masses, that reside in the cores of many galaxies, including our own Milky Way. Astrophysical EMRIs emit gravitational waves in millihertz frequencies, which cannot be detected by existing detectors (seismic gravity-gradient perturbations restrict the operation

of ground-based detectors to frequencies well above 1 Hz). But they will be prime targets for the planned space-based detector LISA (the Laser Interferometer Space Antenna [5, 6]), whose peak sensitivity will be in the millihertz band. EMRIs are extraordinary natural laboratories for strong-gravity physics. In EMRIs, back-reaction from emitted gravitational waves modifies the orbit on a timescale much larger than the orbital period, so the inspiral is slow and gradual—"adiabatic". In a typical LISA EMRI, the captured object spends the last few years of inspiral in a tight orbit around the massive black hole (with an orbital revolution period of order an hour), moving at a significant fraction of the speed of light and emitting some 10^5–10^6 detectable gravitational-wave cycles. The intricate gravitational-wave signal cleanly encodes within it an extremely detailed map of the spacetime geometry around the black hole: nature's "ideal experiment" in strong gravity! Studies have shown how information from EMRI signals could be used to measure the central object's mass and spin with exquisite accuracies, confirm whether it is a Kerr black hole as GR predicts, and eliminate or tightly constrain a host of proposed alternatives to GR. LISA could observe up to hundreds of such EMRIs per year [7].

However, even the most luminous EMRI signals are expected to be weaker than the instrumental noise of LISA in its current design. They could be extracted only by filtering the signal against accurate theoretical waveform templates. A detailed model of the radiative evolution and emitted radiation in EMRI systems is thus a prerequisite to being able to tap into the rich science encapsulated in their signals. Gravitational-wave detectors measure *phase* modulations, so the model has to capture correctly the coherent phase evolution of the emitted EMRI signal over the entire inspiral; and it must ideally do so over the entire relevant parameter space of astrophysical EMRIs, allowing for arbitrary orbital configurations, masses, and any other relevant physical attributes. The exciting prospects of observing EMRIs with LISA have, over the past 20 years or so, driven a concerted effort by theorists to develop a faithful model of EMRIs within GR.

The EMRI domain of the relativistic two-body problem presents unique challenges. The disparate lengthscales and long inspiral time make it hard to tackle using direct Numerical Relativity (NR) methods of the kind that inform the LIGO-Virgo searches (i.e., ones based on numerically solving the full nonlinear set of Einstein's equations). These methods are extremely computationally expensive already in the comparable-masses regime relevant to LIGO-Virgo searches (indeed, these searches rely on phenomenological approximants that interpolate the very sparse database of NR waveforms available), and they become completely intractable for EMRIs. Also inadequate are methods based on weak-gravity or slow-motion approximations, such as the post-Newtonian (PN) approach, in which GR is treated perturbatively as an expansion about Newtonian gravity. PN methods have been greatly successful in predicting and explaining a variety of observed GR phenomena, from Mercury's perihelion advance to the orbital decay in the famous Hulse–Taylor binary pulsar [8]. They also play an important part in modelling the early stage of inspiral for LIGO-Virgo merger searches. But PN methods are wholly inappropriate in the EMRI case, where the entire interesting part of the inspiral occurs low inside the deep gravitational well of the central black hole, where gravity is extreme. One must not rely on any weak-field approximation when modelling EMRIs.

Fortunately, the EMRI problem is naturally amenable to a different type of perturbative treatment: one based on an expansion in the small mass ratio (the small body's mass m divided by the massive black hole's mass M), which, for

4

EMRIs detectable by LISA, is as small as 10^{-5}–10^{-7}. At "zeroth order", the small object is a pointlike test particle, whose motion is unaffected by its own gravitational field or internal structure. The particle traces a *geodesic* orbit—the curved-space equivalent of a straight line—in the Kerr spacetime associated with the large black hole. Geodesic orbits around a Kerr black hole can be highly complex (in particular, they are generically *ergodic*—i.e., space-filling; see Sec. 5.1 and Fig. 4 below), but they are well understood and can be described, essentially, in closed form. Such geodesic orbits approximate true EMRI orbits well only over periods much shorter than the timescale of gravitational radiation-reaction. Radiation-reaction effects come into play at the next order of the expansion: one now treats the gravitational field of the small object as a small perturbation of the background Kerr geometry, satisfying a linearized version of Einstein's field equations, and one then considers the back-reaction from that perturbation on the particle's orbit. Still viewed as a trajectory in the fixed Kerr background, the particle's orbit now experiences a small acceleration with respect to the original geodesic. This acceleration is interpreted as being caused by an effective "gravitational self-force" (GSF) attributed to the particle's interaction with its own gravitational field. In this picture, it is the GSF which drives the slow radiative inspiral, and whose work upon the particle converts orbital energy into gravitational-wave energy. Higher orders in the perturbative expansion account for nonlinear interactions of the gravitational field with the particle and with itself, as well as for dynamical effects coming from the small object's internal structure. (The leading-order effect of the particle's spin enters the equation of motion already at linear order.)

In this manner, one obtains successive approximations for the EMRI orbit and emitted radiation through a systematic expansion in the mass ratio. At each order, one derives an effective equation of motion that describes the EMRI orbit as a trajectory on the fixed Kerr geometry, coupled to a set of field equations that govern the gravitational field, including observable radiation. Analysis suggests [9] that to construct a sufficiently accurate EMRI model for future LISA searches, one must derive and solve the equations of motion through second order in the perturbative expansion (i.e., accounting for self-acceleration terms that scale with the square of m/M). This is a formidable theoretical and computational challenge. To tackle it with sufficient rigour one must take a step back and revisit a fundamental question about the nature of motion in GR: how does a "small" object move in a curved spacetime, which it itself influences?

Historically, many attempts to address this question have been in the PN context of weak fields, going at least as far back as the seminal work of Einstein, Infeld, and Hoffmann [10]. In the fully relativistic arena, significant effort has gone into simply establishing that the equivalence principle (or geodesic principle) is actually a derivable result of the Einstein field equations, not an independent postulate: in the limit of zero mass and size, the Einstein equations alone dictate that all objects, no matter their internal composition, move on geodesics of the external spacetime. Ref. [11] provides an early review of work along these lines, and Ref. [12] a very recent one.

A major step beyond this was taken in the 1970s by Dixon [13], who showed precisely how the geodesic approximation is altered by the finite size of a body (building on earlier work by Mathisson [14] and Papapetrou [15], among others). He derived an exact equation of motion for an arbitrary material body, in which the body's multipole moments couple to the curvature of the external spacetime to induce corrections to geodesic motion. However, his method fails if the body's

own gravity is strong, and he hence restricted his result to *test* bodies, which are of finite size but do not affect the spacetime geometry. He reasoned that extending his method to gravitating bodies would require identifying and somehow subtracting the body's own (highly nonlinear) "self-field", which dominates the metric in the body's neighbourhood but, in analogy with Newtonian gravity, might be expected to have no direct effect on its motion.

In the 1980s, another major step was taken by Thorne and Hartle [16], this time building on work by D'Eath [17] and Kates [18], among others. Using a perturbative expansion in the limit of small mass and size, they established a sort of *generalized* equivalence principle: at least at low orders in perturbation theory, a gravitating object, be it a material body or a black hole, moves as a test body, governed by Dixon's laws, in what locally appears to be the external spacetime. If finite-size effects are neglected, the object moves on a geodesic of that external metric. However, unlike the metric in the usual equivalence principle, this metric is only *effectively* external. It is influenced by the object itself, in a way that Thorne and Hartle did not determine. Hence, like Dixon, they left open the difficult task of finding an appropriate division of the physical metric into a self-field and an "external" remainder.

A decade later, motivated by the emerging need for EMRI models, Mino, Sasaki, and Tanaka [19] and Quinn and Wald [20] overcame that hurdle, deriving an equation of motion, now known as the MiSaTaQuWa equation, that included the effect of the object's own field at first order in perturbation theory. Detweiler and Whiting [21, 22] then showed that the MiSaTaQuWa equation is equivalent to the geodesic equation in a certain linearized vacuum metric, thereby identifying (through first perturbative order) the effective external metric required to complete Thorne and Hartle's results.

This article reviews the significant progress that has been made on the EMRI problem over the two decades since the derivation of the MiSaTaQuWa equation. Our intention is to provide readers with a non-specialist introduction to the subject (the first of this kind, we believe), covering both foundational and computational aspects. We begin with an elementary-level introduction to self-force theory in curved spacetime: Section 2 reviews the foundations of electromagnetic self-force theory in flat and curved spacetimes, and Sec. 3 then covers the essentials of gravitational self-force theory. In Sec. 4 we survey the computational techniques that have been developed in order to enable the application of GSF theory to the astrophysical EMRI problem. Then, in Sec. 5, we describe a perturbative approach to the problem of the long-term radiative evolution of the orbit (given the GSF), based on a systematic two-timescale expansion. At this point we turn to discuss actual calculations of the GSF and its effects in EMRI systems. Section 6 covers calculations of dissipative effects and the long-term orbital evolution in EMRIs, while in Sec. 7 we focus on non-dissipative physical effects, such as the GSF-induced modification of the rates of periastron advance and spin precession. Then, in Sec. 8, we review work comparing the predictions of GSF calculations with those of full NR simulations and of PN calculations. Such synergistic studies can inform the development of a universal model of binary inspirals, valid across all mass ratios; we discuss this idea, and the "effective one body" (EOB) framework that enables it. We conclude, in Sec. 9, with a summery of progress and a discussion of open problems and prospects.

There already exist several, more expert-oriented review texts on EMRI physics and the self-force. The most comprehensive review of self-force foundations, including self-contained derivations, is the Living Review article by Poisson, Pound and Vega [23] (last updated in 2011). There is a more recent, pedagogical review of foundations

6

by A. Pound [24], which covers also more recent work of importance. A. Harte's [25] reviews the non-perturbative approach to the problem of motion in GR, carrying on Dixon's programme. Computational methods for EMRIs have been reviewed by Barack in [26] and, more recently, by Wardell in [27]. For reviews of EMRI science with LISA, see, for example, [28, 29, 7].

2. Electromagnetic self-force in flat and curved spacetimes

In order to understand GSF physics, it is worthwhile to examine the ways in which it differs from Newtonian physics—and the perhaps surprising ways in which it remains the same. Consider the Newtonian analog of an EMRI: the Kepler problem, specifically the case in which a smaller body, such as a planet, orbits a much larger one, such as the Sun. We learn early in our physics education that if the smaller body is perfectly spherical and of uniform density, then it can be treated as a point mass m. It creates a gravitational potential[2] $\Phi(\vec{x}) = -\frac{m}{|\vec{x}-\vec{z}(t)|}$, which diverges at its location, $\vec{x} = \vec{z}(t)$. But we also learn that it does not "feel" the field created by this potential; instead, it only responds to the *external* field of the larger body, obeying the equation of motion

$$m\frac{d^2\vec{z}}{dt^2} = -m\vec{\nabla}\Phi_{\text{ext}}(\vec{z}),\tag{1}$$

where $\Phi_{\text{ext}}(\vec{x}) = -\frac{M}{|\vec{x}-\vec{Z}(t)|}$ and $\vec{Z}(t)$ is the trajectory of the larger body (of mass M). Finally, when we solve the equation of motion (1), we learn that the (bound) solutions are eternal, periodic ellipses.

To a large extent, GSF physics describes the breakdown of each of these results, and in this review we will detail how that breakdown occurs. However, we will also frequently emphasise the countervailing fact: that so long as its various elements are appropriately generalized, much of the Newtonian picture remains remarkably valid.

2.1. Nonrelativistic self-force

Perhaps some years after learning about the Kepler problem, we first encounter a self-force, through which an object does "feel its own field", in the case of a nonrelativistic accelerating charge in flat spacetime. At first, the facts in electromagnetism appear very much the same as in Newtonian gravity: if a point charge q is static, it produces a Coulomb potential $\phi = \frac{q}{|\vec{x}-\vec{z}|}$, just like the gravitational potential; and if we can treat it as a test charge, then it feels only the external fields, just like the point mass does, obeying the Lorentz-force law

$$m\frac{d^2\vec{z}}{dt^2} = \vec{F}_{\text{ext}} := q(\vec{E}_{\text{ext}} + \vec{v}\times\vec{B}_{\text{ext}}),\tag{2}$$

where $\vec{v} = d\vec{z}/dt$. However, this is no longer the case if we stop treating it as a test charge. If we actually take into account the change in the field due to the particle's acceleration, we find the motion obeys the Abraham-Lorentz equation:[3]

$$m\frac{d^2\vec{z}}{dt^2} = \vec{F}_{\text{ext}} + \frac{2}{3}\frac{q^2}{m}\frac{d\vec{F}_{\text{ext}}}{dt}.\tag{3}$$

---

[2] To keep expressions simple, throughout this review we use geometric units in which $G = c = 1$.

[3] The force $\frac{2}{3}\frac{q^2}{m}\frac{d\vec{F}_{\text{ext}}}{dt}$ is often written as $\frac{2}{3}q^2\frac{d^3\vec{z}}{dt^3}$, making the equation of motion third order in time and leading to physically pathological solutions. We instead write it in the "order-reduced" form [30] (sometimes called the Landau-Lifshitz equation), in which $\frac{d^3\vec{z}}{dt^3}$ is replaced by its leading

The second term is a self-force. More specifically, it is a radiation-reaction force. Unlike the Coulomb potential of the static charge, the Liénard-Wiechert potentials of the accelerating charge contain an unbound piece, which carries energy-momentum out to infinity in the form of radiation. That emission causes a recoil, pushing the particle in the opposite direction. Because of this effect, Eq. (3) differs from Newtonian gravity not only in the presence of a self-force, but in the fact that the self-force is dissipative. For that reason, the classical hydrogen atom, consisting of an electron in orbit around a proton, is famously unstable: due to its emission of radiation, the orbiting electron would lose energy, causing it to spiral inward until the atom collapsed. Hence, it is impossible to construct an electromagnetic analog of the Kepler problem.

This connection between local and global effects is an important theme in self-force theory. Locally, the particle's energy can only be changed by the self-force. But the self-force removes an amount of energy from the particle precisely equal to that carried off to infinity in the form of radiation. In fact, the most straightforward way of arriving at Eq. (3) is by finding the simplest force that ensures this energy balance.

A more rigorous way of deriving Eq. (3), and one which introduces a second central theme, is by considering a small charge distribution and then taking the limit as it shrinks to zero size in a self-similar way, such that its charge, mass, and size all go to zero [34]. The notion of a point charge, and in particular a test charge, arises as the leading nontrivial approximation in this limit; the Abraham-Lorentz force (along with, in general, some finite-size effects) appears at the first subleading order. This type of limiting procedure plays a crucial role in self-force theory.

2.2. Relativistic self-force in flat spacetime

Despite our description of it, the Abraham-Lorentz equation in the form (3) does not evince a particularly direct relationship between the self-force and the particle's field. To obtain a more direct, physically compelling picture, and to begin to recover some of the Newtonian description, it is convenient to consider the relativistic generalization of Eq. (3), known as the Abraham-Lorentz-Dirac equation:

$$m\frac{D^2 z^\mu}{d\tau^2} = F_{\text{ext}}^\mu + \frac{2}{3}\frac{q^2}{m}(\delta_\nu^\mu + u^\mu u_\nu)\frac{DF_{\text{ext}}^\nu}{d\tau}. \tag{4}$$

Here $F_\mu^{\text{ext}} := qF_{\mu\nu}^{\text{ext}}u^\nu$ is the covariant form of the Lorentz force, τ is the particle's proper time, $u^\mu := dz^\mu/d\tau$ its four-velocity, and $D/d\tau := u^\mu\nabla_\mu$ the covariant derivative along its worldline.

Like the Abraham-Lorentz force, the self-force in Eq. (4) is dissipative. And like Eq. (3), Eq. (4) can be surmised from a simple conservation law, as Dirac did [35]— specifically, from the conservation of stress-energy inside a small tube around the particle's worldline. Alternatively, it can be rigorously derived from conservation of stress-energy of a small, extended charge distribution, using the limiting procedure mentioned above [34]; a second-order extension of the equation has also been recently

approximant $\frac{d}{dt}(\vec{F}_{\text{ext}}/m)$. Solutions to the order-reduced equation approximate solutions to the third-order equation in a meaningful sense, but they are physically well behaved [31, 32, 33]. Moreover, the order-reduced form is in fact the correct one from a more fundamental perspective. The third-order form is derived from treating the charge as an exact point particle. The order-reduced form, on the other hand, automatically follows from considering an asymptotically small but extended charge distribution [34]. As we discuss below, here we take the position that in a classical theory, point particles are only ever approximations to extended objects, and so we favour the order-reduced form as a matter of principle.

derived using the same method [36]. This, besides being more physically compelling, bypasses the need for an infinite mass renormalization, which is essential in Dirac's derivation. However, we are primarily interested in the form the equation takes, not in how it is derived. Specifically, we wish to establish how the radiation-reaction force in it relates to the particle's own field.

We first examine the form of the field. If the Lorenz gauge condition $\nabla_\mu A^\mu = 0$ is satisfied, then the potential A^μ sourced by the particle satisfies the wave equation

$$\Box A^\mu = -4\pi j^\mu, \tag{5}$$

where $\Box := \eta^{\mu\nu}\nabla_\mu\nabla_\nu$ is the flat-space d'Alembertian, $\eta_{\mu\nu}$ is the metric of flat spacetime, and

$$j^\mu = q u^\mu \frac{\delta^3(\vec{x} - \vec{z})}{u^t} \tag{6}$$

is the particle's charge-current density. The standard retarded and advanced solutions to this equation are

$$A_\pm^\mu(x) = \int G_{\pm\,\mu'}^\mu(x, x') j^{\mu'}(x') d^4x', \tag{7}$$

where, in Cartesian coordinates $(t, \vec{x})$,

$$G_{\pm\,\mu'}^\mu = \delta_{\mu'}^\mu \frac{\delta(t - t' \mp |\vec{x} - \vec{x}'|)}{|\vec{x} - \vec{x}'|} \tag{8}$$

are the retarded (upper sign) and advanced (lower sign) Green's functions for $\Box$, and $d^4x' = dt' d^3x'$ is the spacetime volume element. Primed indices correspond to tensors at $x'^\mu = (t', \vec{x}')$. Due to the delta function in Eq. (8), the retarded and advanced solutions $A_\pm^\mu$ are entirely determined by the state of the particle at the retarded and advanced time, respectively; see Fig. 1.

Of course we are primarily interested in the physical, retarded solution. It contains both time-symmetric and time-antisymmetric pieces, which we can obtain by splitting the retarded Green's function into corresponding pieces:

$$G_{+\,\mu'}^\mu = G_{S\,\mu'}^\mu + G_{R\,\mu'}^\mu, \tag{9}$$

where

$$G_{S\,\mu'}^\mu = \frac{1}{2}(G_{+\,\mu'}^\mu + G_{-\,\mu'}^\mu) \tag{10}$$

and

$$G_{R\,\mu'}^\mu = \frac{1}{2}(G_{+\,\mu'}^\mu - G_{-\,\mu'}^\mu). \tag{11}$$

$G_{\mu\mu'}^S$ is a symmetric Green's function, satisfying $G_{\mu\mu'}^S(x, x') = G_{\mu'\mu}^S(x', x)$ and $\Box G_{S\,\mu'}^\mu = -4\pi\delta_{\mu'}^\mu\delta(x - x')$. $G_{\mu\mu'}^R$, on the other hand, is an antisymmetric homogeneous solution, satisfying $G_{\mu\mu'}^R(x, x') = -G_{\mu'\mu}^R(x', x)$ and $\Box G_\mu^{R\,\mu'} = 0$. (Note that we use the symbols x and x^μ interchangeably to label a point.) Substituting this split into Eq. (7) gives us the corresponding split of the retarded field,

$$A_+^\mu = A_S^\mu + A_R^\mu. \tag{12}$$

The *singular field* $A_S^\mu = \int G_{S\,\mu'}^\mu j^{\mu'} dV'$ is a relativistic generalization of the Coulomb field. It satisfies Eq. (5), it is time-symmetric, and locally, near the particle, it behaves
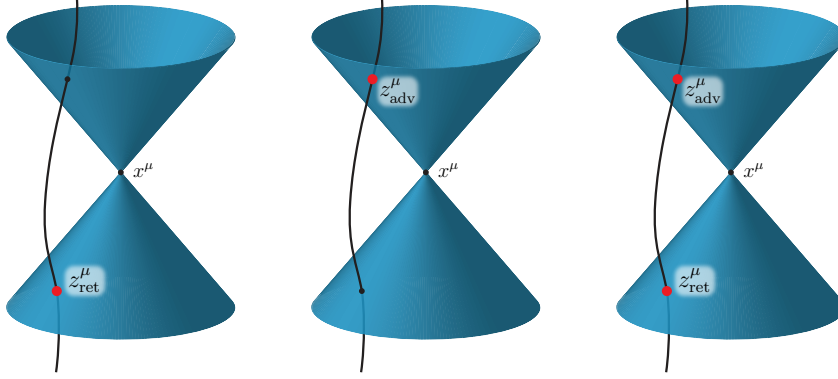
9

Figure 1: Relevant points for the retarded field A_+^μ (left), advanced field A_-^μ (middle), and singular and regular fields A_S^μ and A_R^μ (right) in flat spacetime. A_+^μ at the point x^μ depends on the state of the particle at the retarded point $z_{\mathrm{ret}}^\mu = z^\mu(\tau_{\mathrm{ret}})$, where the particle's worldline intersects x^μ's past light cone. A_-^μ at x^μ depends on the state of the particle at the advanced point $z_{\mathrm{adv}}^\mu = z^\mu(\tau_{\mathrm{adv}})$, where the particle's worldline intersects x^μ's future light cone. A_S^μ and A_R^μ each depend on the state of the particle at both z_{ret}^μ and z_{adv}^μ.

as $A_S^\mu \sim q u^\mu / |\vec{x} - \vec{z}|$, becoming singular at the particle's location. The *regular field* $A_R^\mu = \int G_{R\,\mu'}^\mu j^{\mu'} dV'$, on the other hand, is an unbound field. It satisfies the homogeneous equation

$$\Box A_R^\mu = 0, \tag{13}$$

it is time-antisymmetric, and it is regular (actually, smooth) at the particle's location.

Now return to the equation of motion (4). By explicitly evaluating the Faraday tensor associated with the regular field, $F_{\mu\nu}^R = A_{\mu;\nu}^R - A_{\nu;\mu}^R$, on the particle, and comparing the result to the right-hand side of Eq. (4) (before the order reduction described in footnote 3), one finds that the equation of motion can also be written as

$$m \frac{D^2 z^\mu}{d\tau^2} = F_{\mathrm{ext}}^\mu + q F_{R\,\nu}^\mu u^\nu. \tag{14}$$

In line with its interpretation as a generalization of the Coulomb field, A_S^μ does not appear in the equation of motion. But A_R^μ exerts an ordinary Lorentz force on the particle. Combined with the fact that A_R^μ is a homogeneous field, this suggests that from the particle's perspective, A_R^μ is indistinguishable from an *external* field. If we define the *effective* external field $\tilde{A}_{\mathrm{ext}}^\mu = A_{\mathrm{ext}}^\mu + A_R^\mu$ (and associated Faraday tensor $\tilde{F}_{\mu\nu}^{\mathrm{ext}}$), then the equation of motion is simply the Lorentz-force law

$$m \frac{D^2 z^\mu}{d\tau^2} = q \tilde{F}_{\mathrm{ext}\,\nu}^\mu u^\nu. \tag{15}$$

Equation (15) provides an alternative description of the self-force, one that is not tied to dissipation or radiation-reaction, and one much closer to the Newtonian picture: whatever its field does, a particle is always governed by the Lorentz force exerted by what it perceives to be the "external" field. However, we stress that this is only an effective external field. Away from the particle, A_R^μ *is not physical*. It

10

depends not only on the retarded point $z^\mu(\tau_{\text{ret}})$ on the particle's worldline, but also on the advanced point $z^\mu(\tau_{\text{adv}})$ (see Fig. 1). Hence, it is not causal. Only in the limit to the particle, where the retarded and advanced points merge, does it become physically meaningful.

After the inception of the EMRI modelling programme in 1996, this idea of a particle (or small object) behaving as a test particle in an effective external field was most famously advocated by Detweiler. It occupies a central place in self-force theory, and we will return to it at every stage of this review.

2.3. Electromagnetic self-force in curved spacetime

The move to curved spacetime brings a major change to the physics of the problem. In flat spacetime, waves propagate at the speed of light, along null rays: but in curved spacetime, waves scatter off the spacetime curvature, causing solutions to propagate not just *on* lightcones, but also *within* them. Because of this, the retarded potential A^μ_+ depends not only on the state of the particle at the retarded point $z^\mu(\tau_{\text{ret}})$, but on its state at all prior points $z^\mu(\tau < \tau_{\text{ret}})$, as illustrated in Fig. 2. This causes an important change to the equation of motion (4), which becomes

$$m\frac{D^2 z^\mu}{d\tau^2} = F^\mu_{\text{ext}} + q^2(\delta^\mu_\nu + u^\mu u_\nu)\left(\frac{2}{3m}\frac{DF^\nu_{\text{ext}}}{d\tau} + \frac{1}{3}R^\nu{}_\rho u^\rho\right) + 2q^2 u_\nu \int_{-\infty}^{\tau^-} \nabla^{[\mu}G^{\nu]}_{+\,\mu'} u^{\mu'} d\tau,$$
(16)

where $R^\nu{}_\rho$ is the Ricci tensor of the spacetime, and $G^\mu_{+\,\mu'}$ is the retarded Green's function for the curved-space wave equation [Eq. (17), below]. The final term in this equation is a "tail". It is an integral over the entire past history of the particle, up to $\tau^- = \tau - 0^+$, accounting for all the waves that have scattered back to the particle after having been created by it in its past.

Equation (16) was first derived by DeWitt and Brehme [37] (as corrected by Hobbs [38]) using the same approach as Dirac, considering conservation of stress-energy within a small tube around the particle. Like in the case of flat spacetime, the most rigorous derivation follows from considering the point-particle limit of an extended charge distribution; this has been done by Harte [39, 25], who derived the exact equation of motion of an arbitrary charge distribution and then took the point-particle limit. But also like in the flat-space case, for the moment we are more interested in the form of the equation than its derivation.

Despite the changes in the physics of the solution, the fundamental picture from the preceding section remains valid: the particle feels a Lorentz force due to an effective external field $\tilde{A}^{\text{ext}}_\mu = A^{\text{ext}}_\mu + A^{\text{R}}_\mu$, and the equation of motion (16) can be rewritten in the form (15).

To motivate the form of the regular field A^{R}_μ, we begin with the field equation that the particle's potential A^μ satisfies. In the Lorenz gauge, it reads

$$\Box A^\mu - R^\mu{}_\nu A^\nu = -4\pi j^\mu,$$
(17)

where $\Box := g^{\mu\nu}\nabla_\mu\nabla_\nu$, and $g_{\mu\nu}$ is the metric of the spacetime. The retarded solution is given by $A^\mu_+ = \int G^\mu_{+\,\mu'} j^{\mu'} dV'$, where $dV' = \sqrt{-g'}d^4x'$ is a covariant volume element, with g' being the determinant of $g_{\mu\nu}$ at the integration point x'^μ. We wish to split this solution into appropriate singular and regular pieces in analogy with Eq. (12). We first note that in curved spacetime, the self-force can plainly not be described as the Lorentz force exerted by the potential $\frac{1}{2}(A^\mu_+ - A^\mu_-)$: just as the retarded solution A^μ_+
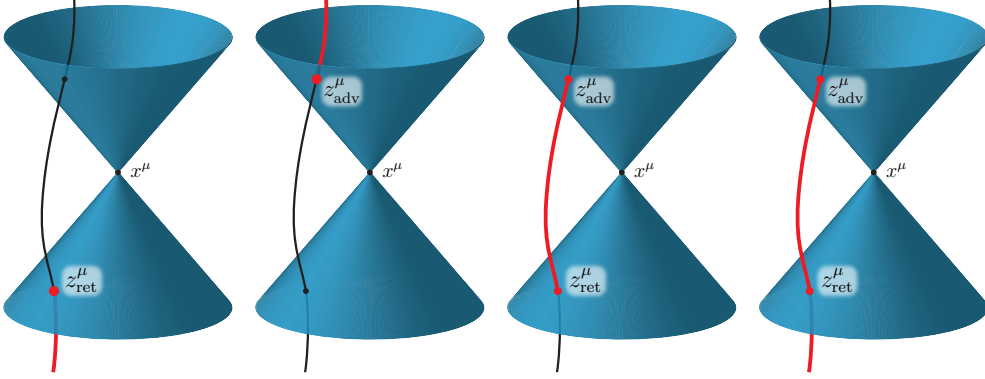
Figure 2: Relevant points for (left to right) the retarded field A_+^μ, advanced field A_-^μ, singular field $A_{\rm S}^\mu$, and regular field $A_{\rm R}^\mu$ in curved spacetime. A_+^μ at the point x^μ depends not just on the state of the particle at the retarded point on x^μ's past light cone, but also on the particle's state at all points *within* the past light cone. Analogously, A_-^μ depends on the state of the particle at all points on and within x^μ's future light cone. $A_{\rm S}^\mu$ depends on the state of the particle at all points on *and outside* x^μ's past and future light cones. $A_{\rm R}^\mu$ depends on the state of the particle at the advanced point $z_{\rm adv}^\mu = z^\mu(\tau_{\rm adv})$ and at all prior points $z^\mu(\tau < \tau_{\rm adv})$.

depends on the entire past history of the particle, the advanced solution A_-^μ depends on its entire *future* history, as shown in Fig. 2. This acausality persists even in the limit to the particle, unlike that of the regular field in flat space, and so it cannot give rise to the correct, physical self-force. An appropriate alternative was found by Detweiler and Whiting [22], who defined the modified two-point functions

$$G_{{\rm S}\,\mu'}^{\mu} = \frac{1}{2}(G_{+\,\mu'}^{\mu} + G_{-\,\mu'}^{\mu} - H^{\mu}{}_{\mu'}) \tag{18}$$

and

$$G_{{\rm R}\mu'}^{\mu} = \frac{1}{2}(G_{+\,\mu'}^{\mu} - G_{-\,\mu'}^{\mu} + H^{\mu}{}_{\mu'}). \tag{19}$$

Here $H^{\mu}{}_{\mu'}$ is a symmetric homogeneous solution, satisfying $H_{\mu\mu'}(x, x') = H_{\mu'\mu}(x', x)$ and $\Box H^{\mu}{}_{\mu'} - R^{\mu}{}_\nu H^{\nu}{}_{\mu'} = 0$. It is chosen such that $G_{{\rm R}\mu'}^{\mu}(x, x')$ has support at all points x' except those in the chronological past of x. This ensures that the corresponding field has no dependence on points $z^\mu(\tau > \tau_{\rm adv})$ in the chronological future; once again, see Fig. 2.

With this choice of $H^{\mu}{}_{\mu'}$, the singular and regular fields $A_{\rm S}^\mu = \int G_{{\rm S}\mu'}^{\mu} j^{\mu'} dV'$ and $A_{\rm R}^\mu = \int G_{{\rm R}\mu'}^{\mu} j^{\mu'} dV'$ possess all of the same key properties as in flat spacetime. $A_{\rm S}^\mu$ satisfies the inhomogeneous Eq. (17), and near the particle it behaves as a Coulomb field. $A_{\rm R}^\mu$ satisfies the homogeneous wave equation and is smooth at the particle. Off the particle, it is acausal, depending on all points on the worldline prior to the advanced point $z^\mu(\tau_{\rm adv})$. But like in flat spacetime, when evaluated on the particle, it becomes causal, depending only on points in the past. Most importantly, evaluating $A_\mu^{\rm R}$ and its derivatives on the particle reveals that the DeWitt-Brehme equation, (16), is equivalent to Eq. (14) and therefore to Eq. (15).

12

However, there is one significant change from flat spacetime. While the Green's function $G^{\mathrm{S}}_{\mu\mu'}$ remains symmetric in its arguments and indices, $G^{\mathrm{R}}_{\mu\mu'}$ does not remain antisymmetric, due to the presence of $H_{\mu\mu'}$ in Eq. (19). Because of this, unlike the purely dissipative self-force in flat spacetime, the self-force in curved spacetime has a conservative piece; it is no longer simply a radiation-reaction force. But despite this change, and despite the more complicated physics of wave propagation in curved spacetime, the essential picture remains unchanged: the particle obeys the Lorentz-force law (15) in what it perceives to be the external field.

3. Gravitational self-force and the generalized equivalence principle

3.1. Perturbation theory in GR and the failure of the point particle description

In our discussion of the electromagnetic self-force, we said that the results are rigorously justified by considering the limit of an asymptotically small charge distribution, with the point particle and its field emerging from that limit. However, the resulting equations, and the discussion of the physics, could be couched almost entirely in terms of the point particle. In the case of a gravitating source in GR, this is no longer true: the point-particle approximation fails; the field of the small object cannot, in general, be expressed as that of a point particle.

This failure stems from the nonlinearity of the Einstein field equations. From a physical perspective, we know the Einstein equations imply that a sufficiently dense mass distribution will collapse to form a black hole, not a point particle. From a mathematical perspective, we know that the Einstein equations with a point-particle source do not have a well-defined solution within any suitable class of functions [40, 41].

Let us examine how this failure manifests itself in our problem. We consider an object of mass m moving in a spacetime with a much larger external length scale $\mathcal{L} \gg m$; in an EMRI, $\mathcal{L}$ can be the mass M of the large black hole, for example (in geometrical units where mass has dimensions of length). Now we wish to take advantage of the separation of scales by expanding the exact metric of our system, $\mathbf{g}_{\mu\nu}$, in the limit $m/\mathcal{L} \to 0$. The metric reads

$$\mathbf{g}_{\mu\nu} = g_{\mu\nu} + \epsilon h^{(1)}_{\mu\nu} + \epsilon^2 h^{(2)}_{\mu\nu} + O(\epsilon^3), \tag{20}$$

where we have introduced ϵ as a formal expansion parameter to count powers of $m/\mathcal{L}$; it will be set to unity at the end of a calculation. The zeroth-order term in Eq. (20), $g_{\mu\nu}$, is referred to as the background metric. In the case of an EMRI, it is the metric of the large black hole. The corrections $h^{(n)}_{\mu\nu}$ describe the gravitational perturbations created by the small object.

The metric (20) must satisfy the Einstein equation $G_{\mu\nu}[\mathbf{g}] = 8\pi T_{\mu\nu}$, where $G_{\mu\nu}[\mathbf{g}]$ is the Einstein tensor of the spacetime and $T_{\mu\nu}$ is the stress-energy tensor of the system's matter content. For simplicity, suppose that the small object represents the only matter, such that $T_{\mu\nu}$ is the stress-energy tensor of the small object itself. If we substitute (20) into the Einstein equation, then the left-hand side becomes

$$G_{\mu\nu}[\mathbf{g}] = G_{\mu\nu}[g] + \epsilon\delta G_{\mu\nu}[h^{(1)}] + \epsilon^2\left(\delta G_{\mu\nu}[h^{(2)}] + \delta^2 G_{\mu\nu}[h^{(1)}]\right) + O(\epsilon^3), \tag{21}$$

where $\delta G_{\mu\nu}[h^{(n)}]$ is linear in $h^{(n)}_{\mu\nu}$ and $\delta^2 G_{\mu\nu}[h^{(1)}]$ has the schematic form $\partial h^{(1)}_{\mu\nu}\partial h^{(1)}_{\alpha\beta} + h^{(1)}_{\mu\nu}\partial^2 h^{(1)}_{\alpha\beta}$. Let us also suppose that in this limit, $T_{\mu\nu}$ is approximately that of a point

particle, such that $T_{\mu\nu} = \epsilon T_{\mu\nu}^{(1)} + \epsilon^2 T_{\mu\nu}^{(2)} + O(\epsilon^3)$, where $T_{\mu\nu}^{(1)}$ is the stress-energy of a point mass moving in the background $g_{\mu\nu}$. (We will momentarily delay the question of whether this makes sense in the case of a black hole, for which $T_{\mu\nu}$ identically vanishes at all points in the spacetime manifold.)

Through first order in ϵ, no fundamental problem arises. Keeping only the first-order terms in the Einstein equation, we arrive at the *linearized* Einstein equation with a point-particle source:

$$\delta G_{\mu\nu}[h^{(1)}] = 8\pi T_{\mu\nu}^{(1)}. \tag{22}$$

This equation is analogous to Eq. (17) for the electromagnetic potential. Its solutions can be expressed in terms of Green's functions, just as in the preceding sections. Like in the electromagnetic case, the retarded field splits into Detweiler-Whiting singular and regular fields. The singular field behaves as $h_{\mu\nu}^{\mathrm{S}(1)} \sim \frac{m}{r}$ near the particle, where r is a measure of distance to the particle's worldline. The regular field $h_{\mu\nu}^{\mathrm{R}(1)}$ is again a smooth vacuum solution that contains the backscattered waves that arise from propagation within, not just on, the light cones of the background spacetime.

But now suppose we read off the *second*-order term in the Einstein equation. It is

$$\delta G_{\mu\nu}[h^{(2)}] = 8\pi T_{\mu\nu}^{(2)} - \delta^2 G_{\mu\nu}[h^{(1)}]. \tag{23}$$

The second-order perturbation $h_{\mu\nu}^{(2)}$ is sourced by the quadratic combinations of $h_{\mu\nu}^{(1)}$ in $\delta^2 G_{\mu\nu}[h^{(1)}]$, which generically behave like $\sim m^2/r^4$ near the particle.[4] This singularity is too strong to be integrated, and because it is constructed from a quadratic operation on an integrable function (as opposed to a linear one), it is not even well defined as a distribution. The other source in Eq. (23), $8\pi T_{\mu\nu}^{(2)}$, if it is well defined at all, must be a distribution solely supported on the particle's worldline. Hence, it cannot cure $\delta^2 G_{\mu\nu}$'s nondistributional divergence, and the field equation (23) is itself ill defined.

Because of this failure of the point-particle treatment, in gravity we must face the small object's extended size head on. However, since the object is small, we still wish to avoid directly including its potentially complicated internal dynamics in the Einstein equations. A principal goal of self-force theory is therefore to generalize the point-particle approximation: to reduce the object to a few "bulk" properties (such as mass and spin) supported on a worldline, without representing it as a delta function stress-energy tensor. Like the point-particle approximation in electromagnetism, this reduction is achieved by considering an extended object in the limit of zero mass and size. Of course, a critical ingredient in the reduction is an equation of motion for the representative worldline, and it is there that the GSF appears.

Before proceeding to describe the limiting procedure and its results, we first note one of its crucial outcomes: at linear order, it establishes that the point-particle approximation *is* valid. That is, Eq. (22) is correct even though Eq. (23) is not. It is also correct even if $T_{\mu\nu}^{(1)}$ is not the first-order approximation to $T_{\mu\nu}$; it remains valid even if the small object is a black hole and $T_{\mu\nu}$ identically vanishes, for example. Given these facts, much of the GSF literature takes Eq. (22) as its starting point, and from that point it describes the GSF in a manner precisely analogous to the electromagnetic case. However, rather than presenting that description immediately, we will instead examine how it emerges from the more fundamental picture of an asymptotically small object.

---

[4] We will, however, mention a fine-tuned way of skirting this generic behavior in Sec. 3.6.

3.2. Point particle limits and multipole moments

The key idea in generalizing the notion of a point particle is to focus not on the small object itself, but on the gravitational field in its immediate neighbourhood. Rather than thinking of the point-particle approximation as a statement about the object's stress-energy tensor, we can take it to be a statement about the object's field.

To illustrate this idea, we return to the simpler context of Newtonian gravity. Consider an isolated material body described by a mass distribution ρ. It sources a Newtonian potential Φ satisfying

$$\partial^i \partial_i \Phi = 4\pi\rho, \tag{24}$$

where we have adopted Cartesian coordinates $x^i = \vec{x} = (x, y, z)$. Written in terms of the Green's function $G(\vec{x}, \vec{x}') = -\frac{1}{|\vec{x}-\vec{x}'|}$, the potential reads

$$\Phi(\vec{x}) = \int G(\vec{x}, \vec{x}')\rho(\vec{x}')d^3x'. \tag{25}$$

Now, suppose we are only interested in the potential outside the body. Further suppose that the body is small, with a characteristic size $\ell \ll |\vec{x}|$, and choose any representative worldline $z^\mu = (t, z^i(t))$ in its interior. We can then expand $G(\vec{x}, \vec{x}')$ in the integrand of Eq. (25) as

$$G(\vec{x}, \vec{x}') = -\frac{1}{r} - \frac{\delta x'^i n_i}{r^2} - \frac{(3\delta x'^i \delta x'^j - \delta r'^2 \delta^{ij})n_i n_j}{2r^3} + O(\ell^3/r^4), \tag{26}$$

where $r := |\vec{x} - \vec{z}|$, $\delta\vec{x}' := \vec{x}' - \vec{z}$, $\delta r' := |\delta\vec{x}'|$, and $\vec{n} := \frac{\vec{x}-\vec{z}}{r}$. Equation (25) then becomes an expansion in terms of the body's multipole moments:

$$\Phi = -\frac{m}{r} - \frac{m^i n_i}{r^2} - \frac{m^{ij} n_i n_j}{2r^3} + O(m\ell^3/r^4), \tag{27}$$

where $m := \int \rho(x')d^3x'$ is the body's mass, $m^i := \int \rho(x')\delta x'^i d^3x'$ its dipole moment, and $m^{ij} := \int \rho(x')(3\delta x'^i \delta x'^j - \delta r'^2 \delta^{ij})d^3x'$ its quadrupole moment. The dipole moment m^i measures the distance between z^i and the body's center of mass. Hence, by choosing the representative worldline z^μ to be that of the body's center of mass, we can set $m^i = 0$.

Equation (27) is the potential *outside* a small but extended body. But suppose we took it to be the potential at all points off the body's worldline [i.e., $x^i \neq z^i(t)$]. It then corresponds precisely to the potential sourced by a "structured" point particle with a mass distribution

$$\rho = m\delta^3(x^i - z^i) - m^i\partial_i\delta^3(x^i - z^i) + \frac{1}{6}m^{ij}\partial_i\partial_j\delta^3(x^i - z^i) + \ldots \tag{28}$$

One can easily check that this source, when substituted into Eq. (25), yields precisely the field (27). If we include only the first term in the source (28), it represents the standard, structureless point mass. The potential it sources in that case is identical to one sourced by a perfectly spherical body, with z^i at the body's center.

The above expansion procedure gives us a precise way of thinking of point particles as approximations to extended objects. But it also demonstrates that, in a meaningful sense, the delta function source (28) is not fundamental to the approximation. Rather

than thinking of the mass distribution $\rho = m\delta^3(\vec{x} - \vec{z})$ as the defining characteristic of the approximation, we can instead think of the singularity in the potential, $\Phi \sim \frac{m}{|\vec{x}-\vec{z}|}$, as primitive; the delta distribution is simply an intermediary that encodes this behavior of the field. This way of thinking, far from being novel, is how point particles were thought of prior to Dirac's invention of his delta function, and it played a significant role in early derivations of equations of motion in GR [42, 10, 43, 11]. Unlike a delta function, it survives in the nonlinear arena of Einstein's theory.

It is worth stressing that here we take the view that a point particle is purely an approximation to an extended object, with no fundamental status on its own. It emerges from an asymptotic expansion in the limit of small size, $\ell \ll r$—or equivalently, large relative distance, $r \gg \ell$. As we get closer to the object, we "see" more of its structure, in the form of sensitivity to its higher multipole moments. If we are very close, at distances comparable to the object's size, $r \sim \ell$, then the approximation breaks down entirely.

3.3. Matched asymptotic expansions and the local form of the metric

Even in GR, any sufficiently well-behaved stress-energy tensor can be reduced to an infinite set of multipole moments, as Dixon showed [13]. But unlike in Newtonian gravity, there is no simple way of translating these moments into an expression for the gravitational field outside the object. Fortunately, there is actually no need to do so: we can obtain the field in a small region outside the object, expressed in terms of a discrete set of multipole moments, directly from the field equations. Instead of specifying a stress-energy tensor, we need only specify the object's moments.[5] We achieve this using the method of matched asymptotic expansions [44, 45]. This method, which derives from singular perturbation theory, has become a standard means of obtaining equations of motion of small objects; see Refs. [17, 46, 18, 47, 16, 48, 49, 19, 21, 50, 51, 52, 53, 54, 55, 56, 57] for a sample. In particular, this method was used in the first derivation of the MiSaTaQuWa equation [19], and it has been the basis of most later foundational work. Our presentation of the method follows Refs. [55, 58].

Matched expansions are used when an ordinary expansion breaks down in a small region. The key idea of the method is to introduce a second limit, one which magnifies this problematic region, and to perform a second expansion in that new limit. In our case, this region is a neighbourhood of the small object itself. As we mentioned in the previous section, the point-particle approximation breaks down at a distance $r \sim \ell$ from the object. The expansion (20) of the metric fails at that same distance. This is intuitively sensible: Eq. (20) treats the small object's field as a small perturbation of the external universe, but sufficiently near the object, its gravity will dominate over that of external sources, and it cannot be treated as a small perturbation.

Let us assume the object is compact, such that $m \sim \ell$, and ϵ now counts powers of $m/\mathcal{L}$ or $\ell/\mathcal{L}$. If we think of $\mathcal{L}$ as being of order 1, then the region $r \sim \ell$, which we call the "body zone", is equivalent to $r \sim \epsilon$. To zoom in on this region, we introduce a scaled distance $\tilde{r} = r/\epsilon$. In terms of this scaled distance, the body zone corresponds to $\tilde{r} \sim 1$. Hence, we can zoom in on the body zone by taking the limit $\epsilon \to 0$ at fixed $\tilde{r}$, and we perform our new expansion in this limit. This contrasts with the original, ordinary expansion (20), which is performed in the limit $\epsilon \to 0$ at fixed r. That limit

---

[5] Of course, if one wishes to model a particular type of body with a specific $T^{\mu\nu}$, then one must infer its moments from that $T^{\mu\nu}$. However, that becomes, in some sense, an independent problem.

shrinks the object to zero size while holding external lengths fixed. The limit at fixed $\tilde{r}$ instead keeps lengths of size ϵ fixed—in particular, the size of the small object—and blows up external lengths out to infinity.

We shall illustrate the method in the Newtonian case. Suppose we have a small body of mass density ρ in an external gravitational field. Again assume the object is compact (but otherwise arbitrarily structured), and take $\mathcal{L}$ to be the typical scale over which the external field varies. In a region outside the external sources, the total field satisfies $\nabla^a \nabla_a \Phi = 4\pi\rho$. Outside all sources, including the body, it satisfies the homogeneous equation $\nabla^a \nabla_a \Phi = 0$. If we expand in powers of ϵ at fixed r, we have $\Phi(r,\epsilon) = \Phi_0(r) + \epsilon\Phi_1(r) + \epsilon^2\Phi_2(r) + O(\epsilon^3)$, analogous to Eq. (20). If we expand in powers of ϵ at fixed $\tilde{r}$, we have instead $\Phi(r,\epsilon) = \hat{\Phi}_0(\tilde{r}) + \epsilon\hat{\Phi}_1(\tilde{r}) + \epsilon^2\hat{\Phi}_2(\tilde{r}) + O(\epsilon^3)$. The background field in the first expansion, Φ_0, is the field due to external sources in the absence of the body; the background in the second expansion, $\hat{\Phi}_0(\tilde{r})$, is the field of the small body itself in the absence of external sources. Let us call the first expansion an "outer expansion" and denote it $E_{\text{out}}\Phi = \Phi_{\text{out}}(r,\epsilon)$, and the second expansion an "inner expansion" and denote it $E_{\text{in}}\Phi = \Phi_{\text{in}}(\tilde{r},\epsilon)$, where $E_{\text{in/out}}$ represent the expansion operations.

Since Φ_{out} and Φ_{in} are both expansions of the same function, they must agree with (or "match") each other when suitably compared. To make the comparison, we perform an inner expansion of Φ_{out}, giving us $E_{\text{in}}\Phi_{\text{out}} = E_{\text{in}}E_{\text{out}}\Phi$, and we perform an outer expansion of Φ_{in}, giving us $E_{\text{out}}\Phi_{\text{in}} = E_{\text{out}}E_{\text{in}}\Phi$. When written as functions of ϵ and r, each of these operations yields a double expansion in the limit $r \to 0$ and $\epsilon \to 0$, with the forms

$$E_{\text{in}}\Phi_{\text{out}} = \sum_{n=0}^{\infty} \sum_{p=-\infty}^{\infty} r^p \epsilon^n \Phi_{n,p}, \tag{29}$$

$$E_{\text{out}}\Phi_{\text{in}} = \sum_{n=0}^{\infty} \sum_{q=-\infty}^{\infty} \left(\frac{\epsilon}{r}\right)^q \epsilon^n \hat{\Phi}_{n,q}, \tag{30}$$

where the coefficients are independent of r and ϵ. For a sufficiently well-behaved Φ, we will have $E_{\text{in}}E_{\text{out}}\Phi = E_{\text{out}}E_{\text{in}}\Phi$,[6] implying the matching condition

$$\Phi_{n,p} = \hat{\Phi}_{n+p,-p}. \tag{31}$$

Equation (29) represents the behaviour of the outer expansion in the limit $r \to 0$, very near the worldline relative to the external length scale $\mathcal{L}$, while Eq. (30) represents the behaviour of the inner expansion in the limit $\tilde{r} \to \infty$, very far from the body relative to the internal length scale ℓ. We can expect the resulting double expansions to be accurate when ϵ/r and $r/\mathcal{L}$ are both small. As illustrated in Fig. 3, this range, $\ell \ll r \ll \mathcal{L}$, describes a "buffer region" between the body zone and the external universe. It can be thought of as a "local far-field" region, being simultaneously in the small body's local neighbourhood ($r \ll \mathcal{L}$) and in its far field ($r \gg \ell$).

Without performing any calculations, we can use the matching condition (31) to constrain the forms of Φ_{in} and Φ_{out}. Since there are no negative powers of ϵ in the outer expansion (i.e., no $\Phi_{n,p}$ for $n < 0$), Eq. (31) implies that $\hat{\Phi}_{n,q}$ vanishes for $q < -n$; likewise, since there are no negative powers of ϵ in Φ_{in} (i.e., no $\hat{\Phi}_{n,q}$ for

---

[6] For discussions of the criteria that guarantee this condition is satisfied, see Ref. [44]. Reference [53] uses a particularly clear, though unnecessarily strong set of criteria.
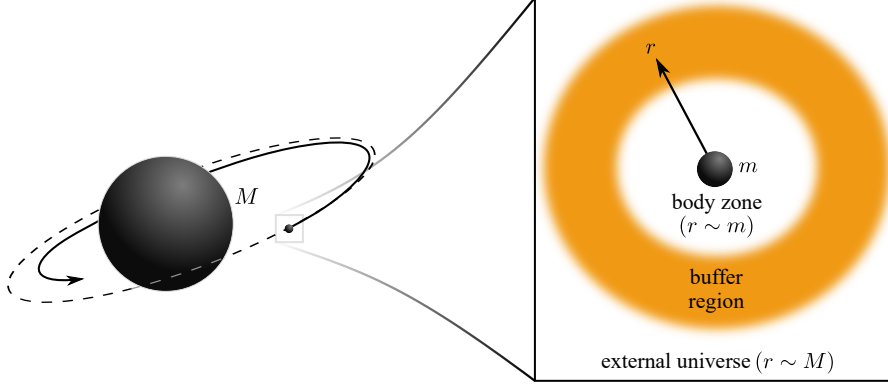
Figure 3: Regions involved in matched asymptotic expansions, specialized to the case of an EMRI. The body zone corresponds to distances $r \sim m$ from the small object; the inner expansion, performed in the limit $\epsilon \to 0$ at fixed r/ϵ, is presumed to be accurate there. The external universe corresponds to distances $r \sim M$; the outer expansion, performed in the limit $\epsilon \to 0$ at fixed r, is presumed to be accurate there. The buffer region corresponds to $m \ll r \ll M$, lying between the other two; the double expansions in the limits $\epsilon \to 0$ and $r \to 0$ are expected to be accurate there.

$n < 0$), Eq. (31) dictates that $\Phi_{n,p}$ vanishes for $p < -n$. Therefore we have

$$\hat{\Phi}_0 = \hat{\Phi}_{0,0} \quad + \frac{\epsilon}{r}\hat{\Phi}_{0,1} \quad + \frac{\epsilon^2}{r^2}\hat{\Phi}_{0,2} + O(\epsilon^3/r^3), \tag{32a}$$

$$\epsilon\hat{\Phi}_1 = r\hat{\Phi}_{1,-1} \quad + \epsilon\hat{\Phi}_{1,0} \quad + \frac{\epsilon^2}{r}\hat{\Phi}_{1,1} + O(\epsilon^3/r^2), \tag{32b}$$

$$\epsilon^2\hat{\Phi}_2 = r^2\hat{\Phi}_{2,-2} + \epsilon r\hat{\Phi}_{2,-1} + \epsilon^2\hat{\Phi}_{2,0} \quad + O(\epsilon^3/r), \tag{32c}$$

and

$$\Phi_0 = \Phi_{0,0} \quad + r\Phi_{0,1} \quad + r^2\Phi_{0,2} \quad + O(r^3), \tag{33a}$$

$$\epsilon\Phi_1 = \frac{\epsilon}{r}\Phi_{1,-1} \quad + \epsilon r^0\Phi_{1,0} \quad + \epsilon r\Phi_{1,1} \quad + O(\epsilon r^2), \tag{33b}$$

$$\epsilon^2\Phi_2 = \frac{\epsilon^2}{r^2}\Phi_{2,-2} + \frac{\epsilon^2}{r}\Phi_{2,-1} + \epsilon^2 r^0\Phi_{2,0} + O(\epsilon^2 r). \tag{33c}$$

The matching condition further dictates that the coefficients in the nth row of Eq. (32) match, term by term, those in the nth column of Eq. (33). However, we often only care about the particular case of the first row and first column.

In our context, we are primarily interested in the outer expansion; we only use the inner expansion to inform the outer. The form (33) of the outer expansion near the worldline is valid regardless of what field equation Φ satisfies. We can now further constrain it using the field equation. Substituting Eq. (33) into $\partial^i\partial_i\Phi_n = 0$, one finds the familiar result that the terms of the form $\frac{\Phi_{n,-p}}{r^p}$ (with $p > 0$) must be linear combinations of spherical harmonics $Y^{p-1,m}$, and those of the form $r^p\Phi_{n,p}$ (with $p \geq 0$) must be linear combinations of harmonics $Y^{p,m}$. We can also write this as $\Phi_{n,-p} = -m^{(n)}_{i_1\cdots i_{p-1}}n^{i_1}\cdots n^{i_{p-1}}$ and $\Phi_{n,p} = \phi^{(n)}_{i_1\cdots i_p}n^{i_1}\cdots n^{i_p}$ for some symmetric and trace-free tensors $m^{(n)}_{i_1\cdots i_{p-1}}(t)$ and $\phi^{(n)}_{i_1\cdots i_p}(t)$.

18

Now let us interpret the terms in the expansion. With suggestive renamings of the tensors $m^{(n)}_{i_1 \cdots i_p}$, we can conveniently sort the terms into two groups,

$$\Phi^{\mathrm{S}}_1 = -\frac{m(t)}{r}, \tag{34a}$$

$$\Phi^{\mathrm{S}}_2 = -\frac{m_i(t)n^i}{r^2} - \frac{\delta m(t)}{r}, \tag{34b}$$

$$\vdots$$

and

$$\Phi^{\mathrm{R}}_n = \phi^{(n)}(t) + \phi^{(n)}_i(t)x^i + \phi^{(n)}_{ij}(t)x^i x^j + \ldots, \tag{35}$$

with the total field given by their sum, $\Phi_n = \Phi^{\mathrm{R}}_n + \Phi^{\mathrm{S}}_n$. Although derived in a very different way, these two groups represent (Newtonian versions of) Detweiler-Whiting singular and regular fields. We can think of $\Phi^{\mathrm{S}} = \sum_{n\geq 1}\epsilon^n\Phi^{\mathrm{S}}_n$ as the body's self-field. It is characterized by a set of multipole moments $m_{i_1 \cdots i_p}$ and corrections $\delta^n m_{i_1 \cdots i_p}$ to them, with Φ^{S}_n containing moments up to $m_{i_1 \cdots i_n}$. We can identify these moments as those of the body itself, *without* having to integrate over the body's interior. To understand this, note that the most singular term at a given order ϵ^n, $\Phi_{n,-n}/r^n$, corresponds to a term in the expansion (32a). That expansion is identical to Eq. (27), the far-field expansion of the field of an isolated body, and instead of defining the body's moments as integrals over its interior, we can define them directly as the coefficients in that far-field expansion. On the other hand, the R field $\Phi^{\mathrm{R}} = \sum_n \epsilon^n\Phi^{\mathrm{R}}_n$ defined from Eq. (35) has the form of a Taylor expansion of a smooth external field, with no direct dependence on the body's moments. Hence, we can think of $\tilde{\Phi} = \Phi_0 + \Phi^{\mathrm{R}}$ as an effective external field. In Sec. 3.6 we will present an example of how the fields Φ^{R}_n actually arise in a Newtonian context.

From the above analysis, we see that solely from the vacuum field equation and the matching condition, we can tightly constrain the local form of the field outside the body: it is given by Eqs. (34) and (35), expressed in terms of the body's multipole moments and some smooth fields Φ^{R}_n. We can freely specify the multipole moments; this is equivalent to specifying the body's material composition. We can also freely specify the fields Φ^{R}_n; this is equivalent to specifying the external environment. However, it is more practical to leave the Φ^{R}_n free at this stage. They will be determined by solving the global problem with whatever external sources are present.

With the local form of the field known, we can now forget about the matching procedure and extend our solution down to $r = 0$ as we did when discussing the point-particle approximation. This effectively replaces the physical field inside the body with a fictitious one. But crucially, it does *not* alter the physical field outside the body. At $r = 0$, the S field becomes singular, while the R field remains a smooth solution to the vacuum equation $\partial^i\partial_i\Phi^{\mathrm{R}}_n = 0$, in precise analogy with the Detweiler-Whiting fields.

Although the calculations become far more involved in GR, all of the essential ideas remain the same. We assume an outer expansion of the form (20) along with a complementary inner expansion at fixed $\tilde{r}$. The matching condition dictates that the fields $h^{(n)}_{\mu\nu}$ have forms precisely analogous to those in Eq. (33). After further constraining the form of the perturbations using the vacuum Einstein equations, one finds that they can again be conveniently written in the form $h^{(n)}_{\mu\nu} = h^{\mathrm{S}(n)}_{\mu\nu} + h^{\mathrm{R}(n)}_{\mu\nu}$,

where in analogy with Eq. (34), the S field has the schematic form [55, 58]

$$h_{\mu\nu}^{\text{S}(1)} \sim \frac{m}{r} + O(r^0), \tag{36}$$

$$h_{tt}^{\text{S}(2)} \sim h_{ij}^{\text{S}(2)} \sim \frac{m^2 + m_i n^i}{r^2} + O(1/r), \tag{37a}$$

$$h_{ti}^{\text{S}(2)} \sim \frac{\epsilon_{ijk} s^j n^k}{r^2} + O(1/r), \tag{37b}$$

and so on at higher orders. [Here, in the curved-space case, (t, x^i) are specialized to be comoving coordinates centred on $\vec{z}$.] In analogy with Eq. (35), the R field has the form

$$h_{\mu\nu}^{\text{R}(n)} = \phi_{\mu\nu}^{(n)}(t) + \phi_{\mu\nu i}^{(n)}(t)x^i + \phi_{\mu\nu ij}^{(n)}(t)x^i x^j + \dots \tag{38}$$

As is clear from the presence of the object's spin s^i in $h_{\mu\nu}^{\text{S}(2)}$, the forms of these fields do differ somewhat from the Newtonian ones. In the Newtonian case, the S field was written entirely in terms of "mass moments" $m_{i_1 \cdots i_\ell}$; in GR there are an additional set of moments $s_{i_1 \cdots i_\ell}$, called "current moments", of which s_i is the first. In $h_{\mu\nu}^{\text{S}(3)}$, both the mass and current quadrupole moments, m_{ij} and s_{ij}, would appear; in $h_{\mu\nu}^{\text{S}(4)}$, the octupole moments; and so forth. Similarly, the R field is more complex than its Newtonian analogue. In Newtonian gravity each of the coefficients in the expansion (35) was symmetric and trace-free; in GR the tensors $\phi_{\mu\nu i_1 \cdots i_\ell}^{(n)}$ are not all symmetric and trace-free, though they are uniquely defined in terms of a certain set of symmetric-trace-free "seed" tensors, in a procedure best explained in Refs. [58, 59].

These differences aside, the S and R fields' properties closely parallel those of the Newtonian fields. In particular, the singular field $h_{\mu\nu}^{\text{S}} = \sum_n \epsilon^n h_{\mu\nu}^{\text{S}(n)}$ carries local information about the object; every term in its expansion, at all orders in r and ϵ, is proportional to one of the object's multipole moments (or to some nonlinear combination of them). The regular field $h_{\mu\nu}^{\text{R}} = \sum_n \epsilon^n h_{\mu\nu}^{\text{R}(n)}$ is a smooth vacuum perturbation for all $r \geq 0$, locally independent of the object; prior to imposing boundary conditions, no terms in the series (38) depend on any of the object's moments. One can also show that, like the Detweiler-Whiting regular field in electromagnetism, $h_{\mu\nu}^{\text{R}}$ is causal on the worldline [59]. We can combine it with the external background $g_{\mu\nu}$ to form a metric $\tilde{g}_{\mu\nu} = g_{\mu\nu} + h_{\mu\nu}^{\text{R}}$ that we can interpret as the effective external geometry. This interpretation will be further bolstered when we consider the equation of motion.

Nevertheless, though the S and R fields have desirable properties, one should keep in mind that only their sum $h_{\mu\nu}^{(n)} = h_{\mu\nu}^{\text{S}(n)} + h_{\mu\nu}^{\text{R}(n)}$ represents a truly physical field. We could have split that field into alternative choices of S and R fields, or foregone the split altogether. For concreteness we have adopted the definitions in Refs. [56, 58], but there are many (in fact, infinitely many) other possible choices. Unlike in the Newtonian case, the singular fields in GR, under most sensible definitions, will contain nondivergent terms proportional to r^n with $n \geq 0$, and because of this, there is no obviously preferred singular-regular split. Ultimately, one can make any convenient choice and derive useful equations in terms of that choice. In Sec. 3.5 we will comment on some alternative choices that have been made in the literature.

Before proceeding, we recapitulate the essential point of this section: Eqs. (36) and (37), along with the Taylor series for $\tilde{g}_{\mu\nu}$, represent the form of the metric in the

buffer region outside any compact object. This form is valid regardless of whether the small object is a material body or a black hole. If it is a material body, then r is some measure of distance to a representative worldline z^μ in the body's interior, as in the Newtonian case. If the object is instead a black hole, then clearly there is no representative worldline in its physical interior. However, we can *associate* a worldline with it based on the behaviour of the field outside of it, with $r = 0$ serving to define that worldline. This associated worldline emerges from the limiting procedure, existing not in the physical spacetime but in the background spacetime with metric $g_{\mu\nu}$ (or equivalently, in the effective spacetime with metric $\tilde{g}_{\mu\nu}$).

3.4. Point particles, punctures, and effective sources

As we described in Sec. 3.1, a traditional point-particle description fails at nonlinear orders. However, after the local analysis of the last section, we are now equipped to adopt the more general viewpoint of Sec. 3.2: rather than approximating the object as a delta function stress-energy tensor, we replace it with a local singularity in the metric. We call this singularity a *puncture* in the spacetime.

To understand how we use this idea, first begin with the vacuum Einstein equations outside the object. At first and second order, they have the form of Eqs. (22) and (23) with the stress-energy terms set to zero,

$$\delta G_{\mu\nu}[h^{(1)}] = 0, \tag{39}$$

$$\delta G_{\mu\nu}[h^{(2)}] = -\delta^2 G_{\mu\nu}[h^{(1)}]. \tag{40}$$

The physically correct solutions to these equations satisfy a free boundary-value problem: in a small region around $r = 0$, they must satisfy Eqs. (36)–(37), which we can think of as boundary conditions; and the location of that boundary region (or equivalently, of $r = 0$) is free to move in response to the solution.

We solve this free boundary value problem by taking the following steps. We first extend the locally derived fields $h_{\mu\nu}^{(n)} = h_{\mu\nu}^{\mathrm{S}(n)} + h_{\mu\nu}^{\mathrm{R}(n)}$ down to all $r > 0$, replacing the physical field in the object's interior but leaving intact the physical field in its exterior. The fields $h_{\mu\nu}^{(n)}$ then satisfy Eqs. (39)–(40) for all $r > 0$. But they do not satisfy these equations on the domain $r \geq 0$; and for $n > 1$, they cannot (in general) be made to satisfy distributional equations on that domain. To obtain equations that can be solved on that domain, we next change our field variable from $h_{\mu\nu}^{(n)}$ to, essentially, $h_{\mu\nu}^{\mathrm{R}(n)}$. More precisely, we introduce a "puncture field" $h_{\mu\nu}^{\mathcal{P}(n)} \approx h_{\mu\nu}^{\mathrm{S}(n)}$ by truncating the expansions (36) and (37) at some finite order, and we then solve for the "residual field" $h_{\mu\nu}^{\mathcal{R}(n)} := h_{\mu\nu}^{(n)} - h_{\mu\nu}^{\mathcal{P}(n)} \approx h_{\mu\nu}^{\mathrm{R}(n)}$. This field satisfies the equations

$$\delta G_{\mu\nu}[h^{\mathcal{R}(1)}] = -\delta G_{\mu\nu}[h^{\mathcal{P}(1)}], \tag{41}$$

$$\delta G_{\mu\nu}[h^{\mathcal{R}(2)}] = -\delta^2 G_{\mu\nu}[h^{(1)}] - \delta G_{\mu\nu}[h^{\mathcal{P}(2)}] \tag{42}$$

for all $r > 0$. If $h_{\mu\nu}^{\mathcal{P}(n)}$ approximates $h_{\mu\nu}^{\mathrm{S}(n)}$ sufficiently well, then the sources in these equations (referred to as *effective* sources) are integrable, and we can define the equations for all $r \geq 0$; note that here the sources are defined at $r = 0$ nondistributionally, by taking the limit from $r > 0$. We can also make $h_{\mu\nu}^{\mathcal{P}(n)}$ go to zero outside some region containing $r = 0$, such that outside that region, $h_{\mu\nu}^{\mathcal{R}(n)}$ reduces to the physical field $h_{\mu\nu}^{(n)}$.

With this procedure, we recover the physical field $h_{\mu\nu}^{(n)} = h_{\mu\nu}^{\mathcal{P}(n)} + h_{\mu\nu}^{\mathcal{R}(n)}$ outside the object, along with the effective field $h_{\mu\nu}^{\mathcal{R}(n)}$ in a region containing $r = 0$. We can also say that we are recovering a local approximation to $h_{\mu\nu}^{\mathrm{R}(n)}$: if we make $h_{\mu\nu}^{\mathcal{P}(n)}$ approximate $h_{\mu\nu}^{\mathrm{S}(n)}$ to sufficiently high order in r, then we can make the Taylor expansions of $h_{\mu\nu}^{\mathcal{R}(n)}$ and $h_{\mu\nu}^{\mathrm{R}(n)}$ identical to any finite order in r. This allows us to *exactly* obtain $h_{\mu\nu}^{\mathrm{R}(n)}$ at $r = 0$, along with any number of its derivatives there.

However, recall that in Sec. 3.1, we mentioned that at first order, the point-particle approximation *is* valid. To see how the approximation is justified, note that $\delta G_{\mu\nu}[h^{(1)}]$ is well defined as a distribution on $r \geq 0$, since it is a linear operator acting on an integrable function. We can hence *define* the stress-energy $T_{\mu\nu}^{(1)} := \frac{1}{8\pi}\delta G_{\mu\nu}[h^{(1)}]$, and a short calculation, beginning from Eq. (36), establishes that this is precisely the stress-energy of a point mass. In other words, a vacuum perturbation $h_{\mu\nu}^{(1)}$ with the local form (36) is *identical* to the field sourced by a point mass sitting at $r = 0$, a result first shown (to our knowledge) in Ref. [17] (see also [53, 55]). Hence, at first order we can rightly think of the object as a point particle. One can then solve the equation $\delta G_{\mu\nu}[h^{(1)}] = 8\pi T_{\mu\nu}^{(1)}$ directly for $h_{\mu\nu}^{(1)}$, and if necessary, afterward subtract $h_{\mu\nu}^{\mathrm{S}(1)}$ to recover $h_{\mu\nu}^{\mathrm{R}(1)}$. From the point-particle description, one can then also return to the puncture scheme, writing $\delta G_{\mu\nu}[h^{\mathcal{R}(1)}] = 8\pi T_{\mu\nu}^{(1)} - \delta G_{\mu\nu}[h^{\mathcal{P}(1)}]$. With $\delta G_{\mu\nu}[h^{\mathcal{P}(1)}]$ interpreted distributionally, unlike in Eq. (41), it cancels the delta function in $8\pi T_{\mu\nu}^{(1)}$, leaving the same nondistributional remainder as if we start with Eq. (41).

Section 4 discusses more technical issues related to both the puncture scheme and the point-particle description. For the moment, we stress the physical picture: we have replaced the object with a singular puncture in spacetime. At linear order, but *not* at nonlinear orders, this is precisely equivalent to approximating the object as a classical point particle (i.e., as a delta function material source). The puncture, or the particle, moves on the curve z^μ. We shall next consider its equation of motion.

3.5. Generalized equivalence principle

Unlike the derivation of the local field outside the object, the derivation of the equation of motion is essentially different in GR than in Newtonian gravity. In the Newtonian case, the gravitational force is an independent postulate of the theory, separate from the field equation. We can use the field equation to obtain the field as a functional of, say, the body's center of mass $\vec{z}$ and its multipole moments, but this leaves both $\vec{z}$ and all the moments freely specifiable. To obtain an equation of motion for $\vec{z}$, we must then appeal to Newton's law of gravitation. To do this for an asymptotically small body, we must sum the forces on each mass element, from which we find that the center of mass satisfies $m\frac{d^2\vec{z}}{dt^2} = -\int \rho\vec{\nabla}\Phi dV$. Due to Newton's third law, all the internal forces cancel one another, leaving us with $m\frac{d^2\vec{z}}{dt^2} = -\int \rho\vec{\nabla}\Phi_{\mathrm{ext}}dV$. When expanded in powers of ϵ, at leading order this reads $m\frac{d^2\vec{z}}{dt^2} = -m\vec{\nabla}\Phi_0 + O(\epsilon^2)$, recovering the equation of motion of the point particle in Eq. (1).[7]

In GR, on the other hand, the equation of motion and field equation are inextricably linked. This complicates some derivations, because it means we cannot solve the field equation without simultaneously finding the equation of motion. However, it also enables a different method of deriving the equation of motion.

---

[7] For a more sophisticated treatment of the Newtonian problem, see the review in Ref. [25].

22

Consider a generic material object with stress-energy $T^{\mu\nu}$. The object's motion is determined by the conservation law $\nabla_\nu T^{\mu\nu} = 0$. By virtue of the contracted Bianchi identity $\nabla_\nu G^{\mu\nu} = 0$, the motion is also equally determined by the Einstein field equation, $G^{\mu\nu} = 8\pi T^{\mu\nu}$. In fact, we can go a step further: as first pointed out by Weyl and Einstein [60, 42, 10, 43], because the motion is encoded in the field even *outside* the object, it can also be determined from the *vacuum* Einstein equations in the object's exterior. This means that we can obtain the equation from the same local analysis described in Sec. 3.3, without ever having to consider the object's interior.

Concretely, let us choose some representative worldline z^μ. This can be *any* worldline near the object's center of mass, in the sense that it lies in the body zone (or equivalently, in the sense that the mass dipole moment m_i is of order ϵ^2). The linearized Einstein equation (39) applied to the field (36) dictates that at zeroth order, for an arbitrarily structured compact object, the worldline must be a geodesic of the background spacetime $g_{\mu\nu}$ [17, 53]:

$$\frac{D^2 z^\mu}{d\tau^2} = O(\epsilon). \tag{43}$$

This is the standard equivalence principle, which states that any sufficiently small object moves on a geodesic of the external spacetime (i.e., it is in free fall in that spacetime).

Similarly, Eq. (40) applied to the fields (36) and (37) determines that at first order, the center of mass (defined by the condition $m_i \equiv 0$) satisfies [53, 55][8]

$$\frac{D^2 z^\mu}{d\tau^2} = -\frac{1}{2}\epsilon g^{\mu\nu}\left(2h^{\mathrm{R}(1)}_{\nu\rho;\sigma} - h^{\mathrm{R}(1)}_{\rho\sigma;\nu}\right) u^\rho u^\sigma - \frac{\epsilon}{2m} R^\mu{}_{\alpha\beta\gamma} u^\alpha s^{\beta\gamma} + O(\epsilon^2), \tag{44}$$

where the spin tensor $s^{\beta\gamma}$ has components $s^{ij} = \epsilon^{ijk}s_k$ (in the comoving coordinate system introduced above; other components vanish). The first term on the right-hand side is the first-order GSF; the second is the Mathisson-Papapetrou spin force. One can rearrange this equation to equivalently write it in the more compelling form

$$\frac{\tilde{D}^2 z^\mu}{d\tilde\tau^2} = -\frac{\epsilon}{2m}\tilde{R}^\mu{}_{\alpha\beta\gamma}\tilde{u}^\alpha s^{\beta\gamma} + O(\epsilon^2), \tag{45}$$

where now the proper time $\tilde\tau$, four-velocity $\tilde{u}^\mu = dz^\mu/d\tilde\tau$, covariant derivative $\tilde{D}/d\tilde\tau = \tilde{u}^\mu\tilde\nabla_\mu$, and Riemann tensor are all defined with respect to the effective metric $\tilde{g}_{\mu\nu} = g_{\mu\nu} + \epsilon h^{\mathrm{R}(1)}_{\mu\nu} + O(\epsilon^2)$. Equation (45) is the equation of motion of a (spinning) *test body* in the effective metric. This result is directly analogous to the electromagnetic case, in which the particle moved as a test particle in the effective field $\tilde{A}^\mu$.

In principle, the next-order equation of motion could be obtained from the third-order vacuum Einstein equation, in the same manner as the zeroth- and first-order equations of motion. In practice it has instead been derived [56, 54] by solving the vacuum field equations in the *inner* expansion to sufficiently high order and then appropriately transforming to the outer expansion. As of the writing of this review, such a derivation has only been performed in the restricted case of an object with vanishing moments s_i, δs_i, m_{ij}, and s_{ij}. The equation of motion for such an object, as derived by one of us in Ref. [56] (see [57] for more details) is found to be

$$\frac{D^2 z^\mu}{d\tau^2} = -\frac{1}{2}\left(g^{\mu\nu} - h^{\mu\nu}_{\mathrm{R}}\right)\left(2h^{\mathrm{R}}_{\nu\rho;\sigma} - h^{\mathrm{R}}_{\rho\sigma;\nu}\right)u^\rho u^\sigma + O(\epsilon^3), \tag{46}$$

---

[8] This result for a spinning object was first derived, in the case of a BH, in Ref. [61] following Thorne and Hartle's approach to matched expansions [16].

where now $h^{\mathrm{R}}_{\mu\nu} = \epsilon h^{\mathrm{R}(1)}_{\mu\nu} + \epsilon^2 h^{\mathrm{R}(2)}_{\mu\nu}$. This can be rearranged to obtain the geodesic equation in the effective metric $\tilde{g}_{\mu\nu} = g_{\mu\nu} + h^{\mathrm{R}}_{\mu\nu}$ [56, 59],

$$\frac{\tilde{D}^2 z^\mu}{d\tilde{\tau}^2} = O(\epsilon^3). \tag{47}$$

Once again, we find that the object moves as a test body in the effective metric. This is a generalized equivalence principle: if we neglect finite-size effects (i.e., spin and higher moments), then any object, no matter its internal structure, falls freely with respect to what it perceives as the external metric. In the context of the puncture scheme described in the preceding section, it is the puncture that moves on this worldline, and in Eq. (46) we replace $h^{\mathrm{R}}_{\mu\nu}$ with $h^{\mathcal{R}}_{\mu\nu}$.

Here we have written all these results in terms of the particular, locally defined regular field introduced in Sec. 3.3. At first glance, it may seem that this regular field is quite remote from the Green's-function-based Detweiler-Whiting field introduced in the electromagnetic case, despite sharing the same key properties. However, the two definitions are intimately related. At linear order, we could have adopted Detweiler-Whiting definitions of $h^{\mathrm{S}(1)}_{\mu\nu}$ and $h^{\mathrm{R}(1)}_{\mu\nu}$ by beginning with the point-particle field equation (22), writing the physical solution in terms of a retarded Green's function, and then adopting a split into S and R fields precisely analogous to Eqs. (18) and (19) [22]. Despite their seemingly very different definitions, these S and R fields precisely agree with the locally defined ones of Sec. 3.3 at least through order r^2 [23, 62].

Nevertheless, one should keep in mind the warning we gave in Sec. 3.3: the split of the physical field into an R field and an S field is not unique. With the particular definitions used here, taken from Ref. [56], the effective external metric $\tilde{g}_{\mu\nu}$ at nonlinear orders has all the desirable, "physical" properties of the Detweiler-Whiting R field: it is a smooth vacuum metric, it is causal when evaluated on the worldline [59], and at least through second order, the object moves as a test body in it. But other choices can be made that preserve all these properties [62], and one should be wary of ascribing too much physical significance to $\tilde{g}_{\mu\nu}$. Furthermore, one can make choices that do not possess all the desirable properties of the Detweiler-Whiting split, but which are nonetheless perfectly practical. For example, in Ref. [54] Gralla defines a singular-regular split that can be used in a puncture scheme at second order; but the regular field in his case is not a vacuum solution, and the equation of motion is not a geodesic in his effective metric. Similarly, in Ref. [63] Harte defines an effective metric in which the body moves as a test body even in the fully nonlinear regime (in the restricted case of a material body); but his effective metric is not a vacuum metric, and it is not suitable for use in a puncture scheme, as it would actually become singular at the worldline in the point-particle limit.

We also note that, aside from the choice of singular-regular split, there are several other differences between various equations of motion that have been derived. At first order, the relation between them is thoroughly established, but at second order, few comparisons have been made. The results summarized in this and the preceding section were derived by one of us in Refs. [55, 56, 58, 57]. Using a very similar method, an alternative form of the second-order equation was derived by Gralla [54]. Though the relation between these results is well understood in broad strokes [59, 57], they have not yet been compared in detail. Section 5 below comments on a principal difference between them. Another result was also derived much earlier, using a more

axiomatic approach, by Rosenthal [52], and Ref. [64] contains yet another perspective. A main difference between the Pound, Gralla, and Rosenthal results is the *gauge* (or class of gauges) in which they are derived. For example, Rosenthal's result was derived in a physically counterintuitive gauge in which the first-order GSF actually vanishes. We turn to this issue of gauge in the next section.

3.6. Gauge freedom

In the context of perturbation theory, GR's gauge freedom corresponds to our ability to make an infinitesimal coordinate transformation, $x^\mu \to x^\mu + \epsilon \xi^\mu + O(\epsilon^2)$. If we expand such a transformation's effect on the metric $g_{\mu\nu} + \epsilon h^{(1)}_{\mu\nu} + O(\epsilon^2)$, we find that it induces a change $h^{(1)}_{\mu\nu} \to h^{(1)}_{\mu\nu} - \nabla_\mu \xi_\nu - \nabla_\nu \xi_\mu$; transformation laws at higher orders in ϵ are easily derived.

To see how this freedom plays into the discussion in the preceding sections, we once again refer to the Newtonian case. More specifically, we return to the Kepler problem recalled at the beginning of Sec. 2, following an example due to Detweiler and Poisson [65, 66]. Here we can clearly see the singular-regular split: the total gravitational potential of the system, $\Phi(\vec{x}) = -\frac{m}{|\vec{x}-\vec{z}|} - \frac{M}{|\vec{x}-\vec{Z}|}$, naturally divides into the S potential sourced by the small mass,

$$\Phi^{\mathrm{S}}(\vec{x}) = -\frac{m}{|\vec{x}-\vec{z}|}, \tag{48}$$

and the regular, external potential sourced by the large mass,

$$\Phi_{\mathrm{ext}}(\vec{x}) = -\frac{M}{|\vec{x}-\vec{Z}|}. \tag{49}$$

As described by Eq. (1), the motion of m is entirely governed by Φ_{ext}. Nevertheless, even in this scenario we can find what might be called GSF effects—if by that we simply mean a force on m that scales with m^2. Let us place the origin of our coordinates at the system's center of mass. This implies $\vec{x}_{\mathrm{CM}} = \frac{M\vec{Z}+m\vec{z}}{M+m} = 0$, meaning that the position of M is given by $\vec{Z} = -\frac{m}{M}\vec{z}$. If we substitute this into the potential in Eq. (49), we find

$$\Phi_{\mathrm{ext}} = -\frac{M}{r} + \frac{mz^i n_i}{r^2} + O\left(\frac{M}{r}\frac{m^2}{M^2}\right), \tag{50}$$

where $r := |\vec{x}|$ and $n^i := x^i/r$. (Note that r and n^i here are not to be confused with the r and n^i of previous sections, which referred to a radial distance and radial unit vector centered on z^i.) In this form, the "external potential" that m feels has contributions proportional to m. In analogy with the definitions in Sec. 3.3, we can idenfity $\Phi_0 := -\frac{M}{r}$ as the "background" field, $\Phi^{\mathrm{R}}_1 := \frac{z^i n_i}{r^2}$ as the "first-order regular potential", and

$$F^i_1 = -m\partial^i \Phi^{\mathrm{R}}_1 \tag{51}$$

as the corresponding "first-order GSF".

However, this split of the external field is coordinate dependent. If we had instead chosen the origin of our coordinates to lie at the large mass's position $\vec{Z}$, then we would have had $\vec{Z} = 0$. The external potential would then have been simply $\Phi_{\mathrm{ext}} = -\frac{M}{r}$. The "higher-order" pieces Φ^{R}_n would all have been identically zero, and we would have had $F^i_1 = 0$. (Though, because the coordinate system in this case is accelerated,

pseudo-forces do arise.)[9] We can understand this as an example of the gauge freedom described above: if we start in the center-of-mass coordinates, we can transform to M-centered coordinates with the small translation $x^i \rightarrow x^i + \frac{m}{M} z^i$. This coordinate transformation entirely "gauges away" the regular potential $\Phi_1^{\rm R}$.

From this example we see that what appears as a perturbative piece of the external field entirely depends on our choice of gauge. The same applies in GR: with an appropriate choice of gauge, the regular field $h_{\mu\nu}^{\rm R}$ and its first derivative can be eliminated at all points along the worldline z^μ, and the GSF along with them. Unlike in the Newtonian case, $h_{\mu\nu}^{\rm R}$ cannot be *entirely* eliminated; it does contain gauge-invariant information. Yet the freedom to alter (or even completely eliminate) the GSF may seem to complicate its interpretation. However, this does not pose a fundamental problem or invalidate the equation of motion (46). Consider again the Kepler problem. The Newtonian equation of motion (1) is valid in any inertial Cartesian coordinate system; it is invariant under Galilean transformations $x^i \rightarrow x'^i = A^i{}_j x^j + v^i t + \zeta^i$, where $A^i{}_j$ is a constant rotation matrix and v^i and ζ^i are constants. Likewise, if we expand it in powers of m/M as described above, the equation of motion

$$ m\frac{d^2 z^i}{dt^2} = -m\partial^i \left[\Phi_0 + \frac{m}{M}\Phi_1^{\rm R} + O\left(\frac{m^2}{M^2}\right) \right] \tag{52} $$

is invariant under infinitesimal Galilean transformations, $x^i \rightarrow x^i + (m/M)\xi^i$, where $\xi^i = \epsilon^i{}_{jk}\phi^k x^j + v^i t + \zeta^i$ and ϕ^k is the axis of rotation. Under such a transformation, the external field is altered, with $\Phi_{\rm ext} = \Phi_0 + (m/M)\Phi_1^{\rm R} + O(m^2/M^2)$ becoming $\Phi_0 + (m/M)\Phi_1'^{\rm R} + O(m^2/M^2)$, where $\Phi_1'^{\rm R} = \Phi_1^{\rm R} - (m/M)\xi^i\partial_i\Phi_0$ is the new regular field. But the left- and right-hand sides of Eq. (52) both transform in the same way, such that in the new frame, the equation takes the identical form in terms of $\Phi_1'^{\rm R}$. In the same way, the geodesic equation in the effective external metric, $\frac{\tilde{D}^2 z^\mu}{d\tilde{\tau}^2} = 0$, is invariant under any smooth coordinate transformation, and its expansion (46) is correspondingly valid in *all* gauges, at least if the transformation between them is smooth. For such transformations, $h_{\mu\nu}^{\rm S(1)}$ is invariant, while $h_{\mu\nu}^{\rm R(1)}$ transforms to $h_{\mu\nu}^{\rm R(1)} - \nabla_\mu\xi_\nu - \nabla_\nu\xi_\mu$, just as if the effective external metric were the complete physical metric. Extensions of these transformation laws to second order are discussed in Ref. [59].

Many of the foundational results in GSF theory were derived in a specific gauge called the Lorenz gauge, in which a "trace-reversed" variable $\bar{h}_{\mu\nu} := h_{\mu\nu} - \frac{1}{2}g_{\mu\nu}g^{\alpha\beta}h_{\alpha\beta}$ satisfies the gauge condition $\nabla^\mu\bar{h}_{\mu\nu} = 0$, in analogy with $\nabla^\mu A_\mu = 0$. In this gauge, the linearized Einstein tensor takes a simple, hyperbolic form,

$$ \delta G_{\mu\nu}[h] = -\frac{1}{2}\Box\bar{h}_{\mu\nu} - R_\mu{}^\alpha{}_\nu{}^\beta\bar{h}_{\alpha\beta}, \tag{53} $$

where $R_\mu{}^\alpha{}_\nu{}^\beta$ is the Riemann tensor of the background metric. This is the gauge in which the MiSaTaQuWa equation, the first-order Detweiler-Whiting decomposition, the local results in Sec. 3.3, and the second-order equation of motion (46) were first derived. Among other advantages, it allows one to work with 4D Green's functions in a generic vacuum background spacetime, just as in the electromagnetic case. In terms of

---

[9] The equation of motion in this case is generally written in the effective-one-body form $\mu\frac{d^2 z^i}{dt^2} = -\mu(M+m)n^i/r^2$, where $\mu = mM/(M+m)$ is the "reduced mass" of the "effective body" in orbit around an effective central mass $M+m$ at $r=0$. In this form, the pseudo-forces have been grouped with the acceleration on the left-hand side.

the retarded Green's function, the first-order equation of motion (44) (in the absence of spin) takes a form analogous to the electromagnetic equation of motion (16),

$$\frac{D^2 z^\mu}{d\tau^2} = -\frac{\epsilon}{2} g^{\mu\nu} (2h_{\nu\rho\sigma}^{\text{tail}} - h_{\rho\sigma\nu}^{\text{tail}}) u^\rho u^\sigma + O(\epsilon^2),\tag{54}$$

where the tail field is given by

$$h_{\alpha\beta\gamma}^{\text{tail}} = m \int_{-\infty}^{\tau^-} \nabla_\gamma G_{\mu\nu\mu'\nu'}^+ u^{\mu'} u^{\nu'} d\tau.\tag{55}$$

This is the original form of the MiSaTaQuWa equation, prior to its reformulation by Detweiler and Whiting.

However, the Lorenz gauge is not always optimal for calculations in particular background spacetimes, such as the black-hole spacetime of an EMRI. Considerable work has gone into understanding the role of gauge in GSF theory [67, 53, 68, 69, 70, 59, 57], and the local field and equation of motion have now been derived in a variety of gauges. Of particular interest are gauges that are either more or less singular than the Lorenz gauge; this takes us outside the class of smoothly related gauges discussed above, demanding fresh consideration of the singular-regular split with each new choice of gauge. Section 4.3 below discusses how more singular gauges become useful in practical computations. More recently, Ref. [57] has constructed a class of "highly regular" gauges, which eliminate the most singular, $\sim 1/|x^a - z^a|^2$ term in $h_{\mu\nu}^{(2)}$. We briefly describe the implications of such a gauge choice in the conclusion.

4. Survey of computational methods

While the first-order equation of motion (54) had been known since 1997, recasting it in a form amenable to numerical computation, concretely for EMRI systems, took some more years of development. In this section we describe the principles of the two main computational frameworks that have been devised and implemented for actual (numerical) computation of the GSF and its effects in EMRI systems: the *mode-sum* and *puncture* methods. For simplicity we restrict the discussion here to the *first-order* GSF; we shall briefly cover recent progress on the second-order problem in the concluding section of this review. Our description here avoids technical detail as much as possible and does not assume familiarity with specific techniques in black-hole perturbation theory. For an expert-level review of GSF computation methods we refer readers to [26] or the more recent [27].

The starting point for any (first-order) GSF computation is the equation of motion (44) [or the original MiSaTaQuWa form (54)]. Omitting the spin and $O(\epsilon^2)$ terms in that equation, we write it here in the more compact form

$$m\frac{D^2 z^\alpha}{d\tau^2} = \lim_{x \to z} m\nabla^{\alpha\beta\gamma} h_{\beta\gamma}^R(x) =: F^\alpha(z),\tag{56}$$

where $\nabla^{\alpha\beta\gamma}$ represents the differential operator appearing on the right-hand side of (44), and we have omitted the superscript '(1)' in $h_{\beta\gamma}^{R(1)}$ for brevity. We further write $h_{\beta\gamma}^R$ as the difference between the full, physical (retarded) metric perturbation sourced by the particle, and the S field:

$$F^\alpha(z) = \lim_{x \to z} m\nabla^{\alpha\beta\gamma} h_{\beta\gamma}^R(x) = \lim_{x \to z} m\nabla^{\alpha\beta\gamma} \left[h_{\beta\gamma}(x) - h_{\beta\gamma}^S(x) \right].\tag{57}$$

The argument 'x' represents a field point in a neighbourhood of the particle's worldline, and 'z' is the worldline point where the GSF is calculated. We suppose for the moment that the full perturbation $h_{\alpha\beta}$ satisfies the Lorenz-gauge form of the linearized Einstein field equations with a point-particle source, which we write here schematically as

$$\tilde{\Box} h_{\alpha\beta} = S_{\alpha\beta}. \tag{58}$$

The explicit form of the wave-like operator $\tilde{\Box}$ can be found in Eq. (53) above, and $S_{\alpha\beta}$ is the linearized energy-momentum tensor associated with the particle, with support (a delta function) confined to the particle's worldline. Equation (58) has to be solved with physical, "retarded" boundary conditions, which are most conveniently imposed at infinity and on the black hole's event horizon. The conditions are that no radiation should be coming in from (past null) infinity, and that no radiation should be coming out from inside the black hole.

From a computational point of view, solving (58) and evaluating (57) pose two main technical difficulties. The first difficulty is that the Lorenz-gauge field equations (58) constitute a complicated set of coupled partial differential equations (PDEs). Even though these equations are linear and manifestly hyperbolic, solving them numerically is computationally expensive and technically challenging (due, in particular, to the need to resolve the diverging field near the particle with sufficient accuracy, and due also to the occurrence of certain mode instabilities—see below). This problem has often been referred to in GSF literature as the "gauge problem": when it was first derived, the correct GSF was only known in the Lorenz gauge, meaning one had to calculate the metric perturbation in a gauge which, despite being convenient for describing the local singularity near the particle, does not sit very well with the global symmetries of the black hole background. The second technical difficulty is the "subtraction problem": to implement Eq. (57) and obtain the GSF, one has to subtract one divergent quantity from another (which is usually given only numerically), before taking the regular limit to the particle. This is obviously problematic in practice.

In the next two subsections we will describe two methods for tackling the subtraction problem. Then, in subsection 4.3, we will see how the gauge problem has been overcome.

4.1. Mode-sum method

The mode-sum method [71, 72] is a general procedure addressing the subtraction problem. The basic idea is very simple. Instead of directly subtracting the divergent field $\nabla^{\alpha\beta\gamma} h^S_{\beta\gamma}$ from the divergent field $\nabla^{\alpha\beta\gamma} h_{\beta\gamma}$, first decompose each of these fields into multipolar-mode components (using a basis of angular harmonics defined on spheres around the *large* black hole), and then perform the subtraction at the level of individual modes. Finally, add up all "regularized" modal contributions. The benefit of such an approach is twofold: First, due to the particular Coulomb-like form of the singularity in the field near the particle, the individual multipole modes of $\nabla^{\alpha\beta\gamma} h^S_{\beta\gamma}$ (and of $\nabla^{\alpha\beta\gamma} h_{\beta\gamma}$) have *finite* values even at the location of the particle, so one only ever subtracts *finite* quantities in one's calculation. Second, the perturbation field $h_{\beta\gamma}$ is typically solved for mode-by-mode anyway, so the necessary input for the mode-sum procedure is readily available without any extra work. The mode-sum formula simply prescribes the "correct" quantity that has to be subtracted from each modal

contribution, so that the resulting regularized modal contributions add up to give the correct, physical GSF.

To describe this more precisely, let us define the fields $F_{\text{full}}^\alpha(x) := m\nabla^{\alpha\beta\gamma}h_{\beta\gamma}$ and $F_{\text{S}}^\alpha(x) := m\nabla^{\alpha\beta\gamma}h_{\beta\gamma}^S$ in the (Kerr) black-hole geometry, and introduce the standard Boyer-Lindquist coordinates (t, r, θ, φ) covering the exterior of the black hole.[10] Then let us consider the decomposition of F_{full}^α and F_{S}^α into spherical-harmonic modes defined on spheres of constant t and r around the black hole: $F_{\text{full}}^\alpha = \sum_{\ell=0}^\infty F_{\text{full}}^{\alpha\ell}$, where $F_{\text{full}}^{\alpha\ell} = \sum_{m=-\ell}^\ell F_{\ell m}(t, r)Y_{\ell m}(\theta, \varphi)$, with $Y_{\ell m}$ being the usual spherical harmonics, and similarly for F_{S}^α. Equation (57) thus becomes

$$F^\alpha(z) = \lim_{x \to z} \sum_{\ell=0}^\infty \left[F_{\text{full}}^{\alpha\ell}(x) - F_{\text{S}}^{\alpha\ell}(x) \right]. \tag{59}$$

The individual mode sums of $F_{\text{full}}^{\alpha\ell}$ and $F_{\text{S}}^{\alpha\ell}$ both diverge at the particle: in the multipolar space, the Coulomb-like particle singularity has turned into a large-ℓ ("ultraviolet") divergence. However, the mode sum of the difference $F_{\text{full}}^{\alpha\ell}(x) - F_{\text{S}}^{\alpha\ell}$ converges faster than any power of $1/\ell$ everywhere, even at the particle, since $h_{\beta\gamma}^R(x)$ in Eq. (57) is a smooth field. This implies that $F_{\text{full}}^{\alpha\ell}$ and $F_{\text{S}}^{\alpha\ell}$ share the same ultraviolet singularity structure. In fact, based on the detailed form of the singular field, one can show [71, 72] $F_{\text{S}}^{\alpha\ell}(z) \sim A^\alpha\ell + B^\alpha + C^\alpha/\ell + \cdots$ at large ℓ, where A^α, B^α and C^α are ℓ-independent expansion coefficients that encode the ultraviolet structure. We can thus write

$$F^\alpha(z) = \sum_{\ell=0}^\infty \left[F_{\text{full}}^{\alpha\ell}(z) - A^\alpha\ell + B^\alpha + C^\alpha/\ell \right] - D^\alpha,$$

$$D^\alpha := \sum_{\ell=0}^\infty \left[F_S^{\alpha\ell}(z) - A^\alpha\ell + B^\alpha + C^\alpha/\ell \right], \tag{60}$$

where the two individual sums are convergent (at least as $\sim \ell^{-1}$). For reasons that will become clear in Sec. 5, the GSF has historically been calculated by approximating the source orbit as a geodesic, neglecting the acceleration caused by the force. To that end, the explicit values of the so-called "regularization parameters", A^α, B^α, C^α and D^α, have been derived analytically for arbitrary geodesic orbits in Kerr spacetime [73, 26]. In particular, it has been shown that when the acceleration is neglected, C^α and D^α always vanish identically (even in a broad class of non-Lorenz gauges [69]). One thus arrives at the final, working form of the mode-sum formula:

$$F^\alpha(z) = \sum_{\ell=0}^\infty \left[F_{\text{full}}^{\alpha\ell}(z) - A^\alpha\ell + B^\alpha \right]. \tag{61}$$

This provides a practical means of evaluating the GSF at any point along a given (geodesic) orbit: First obtain the multipole modes of the physical field $h_{\alpha\beta}$ by solving a suitable version of the linearized Einstein equations mode by mode (this is usually done numerically). From each mode then subtract the analytically given quantity $A^\alpha\ell + B^\alpha$, and finally add up all the modal contributions.

---

[10] The Boyer-Lindquist coordinates (t, r, θ, φ) have the standard intuitive meaning of spherical-polar coordinates (+time) in the weak-gravity region far away from the black hole. They coincide with the similarly labelled Schwarzschild coordinates in the limit of a vanishing black-hole spin.

The above schematic description suppresses some important detail. For instance, the values $F_{\text{full}}^{\alpha\ell}(z)$ and $F_S^{\alpha\ell}(z)$ usually depend on the direction from which the limit $x \to z$ is taken (and so does the value of the parameter A^α), so in the above procedure one is required to specify a particular directional limit. (Of course, the value of the GSF itself is independent of the choice of direction.) Also, the multipole modes into which the field $h_{\alpha\beta}$ is decomposed are not the usual spherical-harmonic modes, but (for instance, in the case of a Schwarzschild background) a suitable tensorial generalization thereof. So, obtaining the modal contributions $F_{\text{full}}^{\alpha\ell}$ involves the additional step of re-expanding the tensorial modes in a basis of spherical harmonics. There is, in fact, a variant of the mode-sum procedure that employs a fully tensorial mode decomposition (with suitable regularization parameters), avoiding that extra complication [74]. There is also a generalization of the mode-sum formula to accelerated (non-geodesic) orbits, allowing one to account for the particle's self-acceleration [75]. Finally, we mention that the mode sum in (61) converges only slowly: the summand typically falls off only as $\sim \ell^{-2}$. This means one normally has to compute a large number of ℓ-modes ($\ell_{\text{max}} \sim 50$ is typical), which can become computationally expensive. This problem can be mitigated through the use of "high-order regularization parameters", which analytically capture some of the higher-order terms in the $1/\ell$ expansion of $F_S^{\alpha\ell}(z)$ [76, 77]; these higher-order terms can be constructed such that they sum to zero, but including them accelerates the rate of convergence. For more details, refer to [26, 27].

The mode-sum scheme, in its many variants, has been the primary framework for GSF calculations, and it is responsible for many of the results to be presented in later sections. In particular, it was employed in the recent first calculation of the GSF for generic bound geodesic orbits around a Kerr black hole [78, 79].

4.2. Puncture (or "effective source") method

In Sec. 3.4 we introduced the puncture method as a means of circumventing the nonintegrability of the second-order source, and a minimal version of the method was proposed for that purpose at least as far back as Ref. [80]. However, the puncture method was first fully developed and implemented [81, 82, 83] at first order, with a different motivation: tackling problems where a multipole decomposition, and hence mode-sum regularization, is awkward. This is the case with the Lorenz-gauge metric perturbation on a Kerr background, for which no natural basis of harmonics is known.

As described in Sec. 3.4, the puncture method addresses the subtraction problem differently from mode-sum. Here, the "regularization" is performed already at the level of the field equation (58). Instead of solving for the physical field $h_{\alpha\beta}$ and then subtracting the singular field, one solves directly for a local approximation to the regular field $h_{\alpha\beta}^{\text{R}}$. Specifically, we design an analytic function $h_{\alpha\beta}^{\mathcal{P}}(x)$ that approximates the singular field $h_{\alpha\beta}^{\text{S}}(x)$ near the particle sufficiently well that

$$\lim_{x \to z}(h_{\alpha\beta}^{\mathcal{P}} - h_{\alpha\beta}^{\text{S}}) = 0 \quad \text{and also} \quad \lim_{x \to z}(\nabla^{\alpha\beta\gamma}h_{\beta\gamma}^{\mathcal{P}} - \nabla^{\alpha\beta\gamma}h_{\beta\gamma}^{\text{S}}) = 0. \tag{62}$$

Then it is perfectly allowable to replace the true singular field in Eq. (57) with its puncture-field approximant:

$$F^\alpha(z) = \lim_{x \to z} m\nabla^{\alpha\beta\gamma}h_{\beta\gamma}^{\mathcal{R}}(x), \tag{63}$$

where we have introduced the residual field $h_{\beta\gamma}^{\mathcal{R}} := h_{\beta\gamma} - h_{\beta\gamma}^{\mathcal{P}}$. We then make $h_{\beta\gamma}^{\mathcal{R}}$ the

subject of the field equation (58):

$$\tilde{\Box}h^{\mathcal{R}}_{\alpha\beta} = S_{\alpha\beta} - \tilde{\Box}h^{\mathcal{P}}_{\alpha\beta} =: S^{\mathrm{eff}}_{\alpha\beta}. \tag{64}$$

The effective source $S^{\mathrm{eff}}_{\alpha\beta}$ contains no delta function on the particle's worldline; its residual non-smoothness there is determined by how well our puncture field approximates the singular field. The field $h^{\mathcal{R}}_{\alpha\beta}$ is at least once differentiable (unlike the physical field $h_{\alpha\beta}$, which is divergent), and it directly yields the GSF, via Eq. (63).

In practice, it is convenient to restrict the support of $S^{\mathrm{eff}}_{\alpha\beta}$ in Eq. (64) to within a small region around the particle's worldline, so as to avoid having to control the behavior of the puncture field and effective source far from the particle. This can be achieved with a suitable window function, or by introducing a "worldtube" around the worldline, such that one solves for $h^{\mathcal{R}}_{\alpha\beta}$ inside the worldtube and for the original perturbation $h_{\alpha\beta}$ outside it, with the analytically known value of the puncture field used to communicate between the two variables across the boundary of the tube. This scheme can be implemented numerically without any multipole decomposition, directly evolving the hyperbolic PDE (64) from initial conditions in 3+1 dimensions (space+time) [84]. However, the method has also proved useful when applied in conjunction with a mode decomposition. One can separate the field into azimuthal modes, $\sim e^{im\varphi}$, and evolve each of the m-modes separately in 2+1 dimensions [85, 86, 87, 88]; this is possible and useful even on a Kerr background, thanks to its axial symmetry. When a full multipole expansion is possible, such as on a Schwarzschild background, the puncture scheme can be implemented in 1+1 dimensions, multipole mode by multipole mode [83]. It can even be applied in the frequency domain [89, 74].

The utility and significance of the puncture idea becomes fully manifest when coming to solve the second-order field equation (23). Here, applying the mode-sum method becomes impossible in general. Recall that the source term appearing in the second-order equation (23) is sufficiently singular that the equation does not actually admit a globally valid solution. But even restricting to $x^{\mu} \neq z^{\mu}$, the singularity in the second-order solution, described by Eq. (37), is strong enough (a consequence of the distributionally ill-defined source) that its individual ℓ modes diverge at the particle. This means that even if one were given the modes of the retarded field, one could not apply mode-sum regularization to extract the regular field. For these reasons, the puncture idea takes on a more fundamental status in second-order GSF calculations. The basic idea of the puncture can also be applied to control the behavior of the second-order solution near the horizon and at large distances, in circumstances where one would otherwise encounter infrared-type divergences [90].

4.3. Alternative choices of gauge

The application of both the mode-sum and puncture methods involves, in some form, solving linear field equations for the metric perturbation (or for its multipole modes), which must usually be done numerically. In the above discussion we have referred specifically to the Lorenz-gauge form of these equations, (58) or (64). This form is convenient for a number of reasons: First and foremost, it yields the perturbation field in a gauge consistent with that assumed in the original GSF formulation, and so ready to be used directly in calculations. Related to this, the singularity of $h^{\mathrm{S}(1)}_{\mu\nu}$ in the Lorenz gauge has an intuitive, Coulomb-like form. Finally, the field equations themselves are hyperbolic and form part of a mathematically well-posed initial-value problem. Indeed, much of the initial progress in GSF calculations came from direct

31

Lorenz-gauge implementations [91, 92, 93], and the approach is still a popular one, in conjunction with either mode-sum regularization [94] or a puncture [87, 74].

However, the direct Lorenz-gauge approach has several serious weaknesses. Prime among these is the fact that the Lorenz-gauge field equations (58) cannot be decomposed into individual, decoupled multipole modes on a Kerr background, in any known form. This restricts one to time-domain numerical evolutions in 3+1 or 2+1 dimensions, which are computationally expensive and cumbersome.[11] Second, even on a Schwarzschild background where the equations are separable into (tensorial-type) spherical harmonics, they still constitute a complicated set of 10 equations that couple between the various tensorial components. Third, time-domain evolutions of the Lorenz-gauge equations appear to suffer from linear instabilities associated with certain nonphysical gauge modes [87], whose removal still constitutes an open problem.

The above complications have motivated the development of methods of calculating the GSF in alternative gauges, facilitated by theoretical work to extend the GSF formalism beyond the Lorenz gauge [67, 53, 68, 69, 54, 70, 59, 57]. These developments have focused on the most traditionally useful gauges in black hole perturbation theory: the Regge-Wheeler-Zerilli gauge [95, 96] (or closely related "EZ" gauge [97]) in the Schwarzschild case; and the so-called "radiation gauge" in the Kerr case [98]. In these gauges, the first-order metric perturbation can be obtained (or "reconstructed") from one or more scalar quantities that satisfy fully separable field equations, reducing the numerical calculation of $h_{\alpha\beta}$ to solving a set of ordinary differential equations. For example, in the case of the radiation gauge, the scalar quantity is the linear perturbation of one of the Newman-Penrose curvature scalars—either Ψ_0 or Ψ_4—constructed from the Riemann tensor; the separable field equation is then the well-known Teukolsky equation.

Unfortunately, as alluded to in Sec. 3.6, these alternative gauges become poorly behaved in the presence of a point-particle source, introducing pathological singularities into the metric perturbation. These singularities violate the generic form of the local field described in Sec. 3.3. However, by analyzing the local form of the transformation to the Lorenz gauge, Ref. [70] showed that nevertheless, the GSF and related quantities can be rigorously calculated from a mode-sum formula in the radiation gauge. Along with subsequent followup work [99, 100, 101], this result has effectively resolved the "gauge problem".[12]

Combined with significant development of practical computational methods, these results have enabled vital progress over the past decade, particularly in the Kerr case, where Lorenz-gauge calculations are most cumbersome [103, 104, 105, 106, 107, 108, 109]. This programme has recently culminated in a calculation of the GSF for generic (eccentric and inclined) bound geodesic orbits, by van de Meent [78, 79]. Figures 7 and 8 in Sec. 6 below show sample outputs from this calculation, which is an implementation of the radiation-gauge approach.

In both the Regge-Wheeler-Zerilli and radiation gauges, such calculations have been further facilitated by the semi-analytical method of Mano, Suzuki, and Takasugi [110], in which solutions to the relevant scalar-like field equations are expressed as infinite sums of hypergeometric functions, involving certain coefficients that

---

[11] A Fourier expansion in both t and φ would reduce the problem to that of solving an elliptic set in two spatial dimensions, but this approach has not been explored yet.

[12] Although no comparable result has been derived in the Regge-Wheeler-Zerilli gauge, current work is expected to soon rectify that [102]. The usability of this gauge is, anyhow, restricted to the Schwarzschild case.

are determined numerically. This approach is now enabling very high-precision calculations of the GSF in Kerr spacetime [111, 112]. It even makes entirely analytical treatments possible within a post-Newtonian framework, where the equations are solved, and the GSF computed, order by order in a small-frequency expansion. Such analytical calculations have been performed to very high post-Newtonian order [113, 114, 115, 116, 117, 118, 119, 120].

5. Perturbative approaches to orbital evolution in EMRIs

A calculation of the local GSF acting on the small object in an EMRI system is only a first step (crucial as it is) in the programme to model the long-term orbital dynamics and emitted gravitational waves. One must also devise a method that uses the GSF information to construct a sufficiently accurate description of the evolving orbit and emitted radiation. In this section we review the formulation of such a method, based on a systematic perturbative expansion that exploits the adiabatic nature of the inspiral process.

But before we do so, we must review some of the pertinent properties of bound *geodesic* orbits in Kerr geometry. The next subsection summarizes the essentials of "orbital mechanics" of test particles in Kerr, and in Sec. 5.2 we will move on to discuss the problem of evolution under the effect of the GSF.

5.1. Geodesic orbits around a Kerr black hole

In Keplerian mechanics, the motion of a test mass in a spherically symmetric stationary gravitational potential is very simple. The orbit has a conserved total energy E (kinetic+potential) associated with the time-translation invariance of the background potential, and a conserved orbital angular momentum $\mathbf{L}$ associated with its rotational invariance. The conservation of $\mathbf{L}$ leads to a planar motion, and the particular inverse-distance form of the potential guarantees that bound orbits are exactly periodic. The situation is quite similar in GR when one considers the motion of a test (pointlike, non-spinning) mass in the spherically symmetric and stationary gravitational potential of a Schwarzschild black hole: again, the symmetries of the background potential enable one to identify a conserved total energy E[13] and a conserved angular momentum $\mathbf{L}$, and the latter's existence means orbits are planar. However, in GR, bound orbits are no longer exactly periodic: the deviation of the potential away from the inverse-distance law produces the well-known relativistic precession effect, whose observational imprint on Mercury's orbit (the "anomalous" perihelion advance) was famously used in validating the infant GR theory. Instead, bound geodesic orbits in a Schwarzschild potential exhibit a combination of two periodic motions. Each orbit has an epicyclic period T_r, equal to the time between two successive periapsis passages, and a "rotational" period T_φ associated with the mean azimuthal motion. While in the Keplerian theory $T_r = T_\varphi$, in GR it is always the case that $T_r > T_\varphi$, corresponding to an advance of the periapsis.

The situation is more involved when the central black hole is spinning. The gravitational potential of a Kerr black hole is still stationary, so orbital energy is still conserved. However, the potential is no longer spherically symmetric, so the total orbital angular momentum $\mathbf{L}$ is not conserved and orbits are generally non-planar.

---

[13] In the relativistic treatment, it is customary to include the particle's "rest mass", mc^2, in E, so that E now represents the sum of kinetic, potential, and rest-mass energies.

Nonetheless, the potential remains *axially* symmetric, meaning the projection L_z of **L** along the symmetry axis (the direction of the black hole's spin) *is* conserved. As a result, the orbital plane performs a simple-precession motion about the direction of the black hole's spin. This is an example of the Lense-Thirring effect associated with coupling between spin and orbital angular momentum in GR. Bound geodesic orbits around a Kerr black hole are thus (generically) *tri*-periodic: in addition to T_r and T_φ, they posses a third, "longitudinal" period T_z, equal to the time interval between two successive minima of the object's longitudinal angle. The combination of two precessional (libration-type) motions traces out a complicated trajectory, which is generically *ergodic* (space-filling): loosely speaking, each point within a certain torus-shaped volume around the black hole eventually gets visited by the particle. Figure 4 displays a sample geodesic orbit, illustrating this behavior (left orbit in the figure; the special behavior of the orbit on the right will be discussed further below).
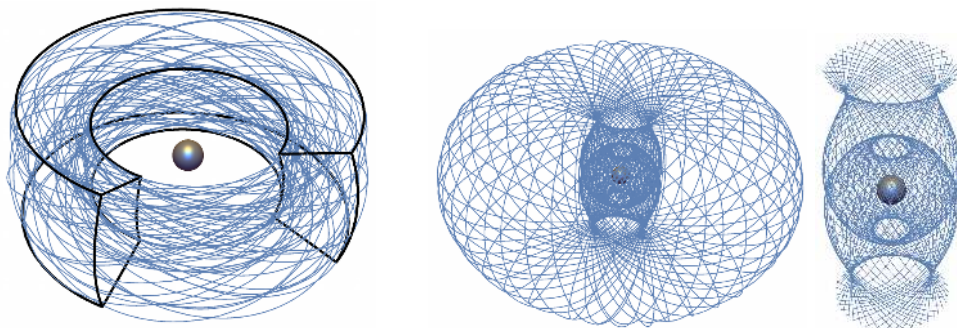

Figure 4: *Left:* Typical geodesic orbit around a Kerr black hole. The orbit is 3-periodic, and it ergodically fills the interior of the outlined torus-shaped region. *Right:* The special case of a resonant orbit. Here the radial and longitudinal periods are in a $3:2$ ratio, and the motion is no longer ergodic. Note how the orbit is instead confined to a certain 2-dimensional surface (topologically, a self-intersecting cylinder). The rightmost figure expands the central region of the resonant orbit, for clarity.

The conserved quantities E and L_z constitute first integrals of the geodesic equation of motion. The particle's conserved mass, m, is a third such integral. Not long after the discovery of the Kerr metric, B. Carter identified a *fourth* integral, Q, now known as the Carter constant [121]. Q is associated with a more subtle symmetry of the Kerr geometry. It does not have a simple physical interpretation or a Newtonian analogue, except in the weak-field or Schwarzschild limits, where it roughly corresponds to the square of the "remaining" component of angular momentum, $L_x^2 + L_y^2$. Orbits that are initially equatorial remain equatorial (due to the symmetry of the Kerr geometry under reflection across the "equatorial" plane, i.e. the plane orthogonal to the black hole's spin direction) and have $Q = 0$. The trio of constants $\{E, L_z, Q\}$ completely and uniquely parametrizes all geodesic orbits in Kerr spacetime, up to initial phases.[14]

The existence of four first integrals of motion—the trio $\{E, L_z, Q\}$, in addition

---

[14] However, curiously, it turns out that the trio of periods $\{T_r, T_\varphi, T_\theta\}$ is *not* a good parametrization: there are (infinitely many) pairs of physically distinct orbits that exhibit the same three periods [122].

to m—allows us to write the geodesic equations of motion in a convenient first-order form. Moreover, as first noted by Carter [121], one can choose the "time" parameter along the orbit so that the radial and longitudinal libration motions become manifestly decoupled from one another. Let $x_{\rm p}^\alpha(\lambda) = \{t_{\rm p}(\lambda), r_{\rm p}(\lambda), \theta_{\rm p}(\lambda), \varphi_{\rm p}(\lambda)\}$ represent the particle's geodesic trajectory in Boyer-Lindquist coordinates, where λ is Carter's time parameter.[15] The equations of geodesic motion in Kerr geometry take the remarkably simple form

$$
\begin{aligned}
\frac{dr_{\rm p}}{d\lambda} &= \pm\sqrt{R(r_{\rm p})}, \\
\frac{d\theta_{\rm p}}{d\lambda} &= \pm\sqrt{\Theta(\cos\theta_{\rm p})}, \\
\frac{d\varphi_{\rm p}}{d\lambda} &= \Phi_r(r_{\rm p}) + \Phi_\theta(\theta_{\rm p}), \\
\frac{dt_{\rm p}}{d\lambda} &= \mathcal{T}_r(r_{\rm p}) + \mathcal{T}_\theta(\theta_{\rm p}),
\end{aligned}
\tag{65}
$$

where each of the right-hand-side functions depends only on its indicated argument (as well as on the constant parameters E/m, L/m and Q/m, and on the black hole's mass and spin); the explicit form of these functions can be found in Carter's paper, Ref. [121]. Given initial conditions [say, $x_{\rm p}^\alpha(0)$], the radial and longitudinal motions can be immediately (and independently) determined from the first two of these equations. Then, supplied with $r_{\rm p}(\lambda)$ and $\theta_{\rm p}(\lambda)$, one can solve for the azimuthal motion using the third equation. The fourth of the equations in (65) relates the parameter λ to the standard time coordinate t.

Geodesic motion in Kerr is thus manifestly *integrable*, to use the language of dynamical systems. As such, it can be conveniently formulated in terms of action-angle variables. This will prove useful when we later discuss the effect of the self-force. We let $J_\alpha = \{E, L, Q, m\}$ be our action variables,[16] and $q^\alpha = \{q^t, q^r, q^\theta, q^\varphi\}$ be our generalized angle variables, associated with the t, r, θ and φ motions. The equations of geodesic motion in Kerr then take the form

$$
\dot{J}_\alpha = 0, \qquad \dot{q}^\alpha = \omega^\alpha(J_\mu),
\tag{66}
$$

The four quantities ω^α are generalized "frequencies" associated with q^α; their relationships with the physical (Boyer-Lindquist) frequencies, for bound orbits, were derived by Schmidt [124]. The parameters J_α are the orbit's *principal* elements. They describe the "shape" of the orbit and determine physical attributes such as the orbital eccentricity and semimajor axis. The parameters q^α are the orbit's *positional* elements. They contain the orbit's phase information and determine physical attributes such as the (time-dependent) direction of the periapsis and orientation of the orbital plane. In Eq. (66) we allow the overdots to denote differentiation with respect to any suitable parameter along the orbit—like λ, t, or proper time, with suitable redefinitions of ω^α and q^α; we will make a particular choice of parameter later in our discussion.

In general, the radial and longitudinal libration motions are incommensurate, leading to ergodic behavior, as mentioned above. However, there is a special class

---

[15] λ is often referred to as "Mino time" in more recent literature, after Y. Mino, who emphasized its utility in the context of EMRI physics [123].

[16] Strictly speaking, the action variables are certain integrals over the phase-space torus. But those integrals are (invertible) functions of $\{E, L, Q, m\}$, and so we allow ourselves to refer to $\{E, L, Q, m\}$ themselves as the action variables.

of geodesic orbits for which ω^r/ω^θ is a rational number. Such "resonant" orbits are non-ergodic and precisely periodic: the orbit completes a certain integer number of radial cycles at the same time it completes a certain integer number of longitudinal cycles. Figure 4 above shows an example. The unusual periodic nature of resonant orbits manifests itself more profoundly the smaller those integers are; resonances with small integers are sometimes called "strong". During the gradual radiative inspiral of an EMRI system, the orbit will become momentarily tangent to numerous such resonances, including, generically, ones that are strong [125]. Such resonant crossings have interesting dynamical consequences, mentioned further below.

A distinctive property of orbits around black holes, of no analogue in Keplerian mechanics, is the presence of an *innermost stable orbit*. In the Keplerian problem, circular orbits (for example) exist at arbitrarily small radii around the center of attraction: the gravitational pull can always be countered with a sufficiently strong centrifugal force by endowing the particle with a sufficiently large tangential velocity. Moreover, all such orbits are dynamically *stable*, in the sense that a small perturbation applied to the orbit (e.g., one that takes it away from being circular) remains small over time. Contrast this with the situation in GR for, e.g., a Schwarzschild black hole. There, no physically allowable amount of angular momentum can support circular geodesic motion below $r = 3M$ (where M is the black hole's mass, with its event horizon at $r = 2M$), and the innermost *stable* circular orbit (ISCO) is as far out as $r = 6M$. In fact, all stable bound geodesic orbits around a Schwarzschild black hole, even highly eccentric ones, reach their perapsis above $r = 4M$. For rotating black holes, the ISCO location depends on the amount of spin. A higher spin rate gives rise to a smaller ISCO radius for objects that co-rotate with the black hole, and to a larger ISCO radius for counter-rotation motion.

5.2. Self-consistent and two-timescale descriptions of the long-term orbital evolution

Though the geodesic dynamics is complex, on the face of it, the long-term inspiral seems fairly simple. We know that the GSF has a dissipative effect, causing the small object to spiral into the central black hole. Because the GSF is small, we also know that this process must be slow; over a single radial period, for example, the worldline traced out by the object must be very nearly a geodesic of the background spacetime. Hence, we can say that the inspiral is approximately *adiabatic*, slowly evolving through a sequence of background geodesics. Occasionally this evolution will pass through one of the resonances described in the previous section, but for the most part, the evolution is apparently innocuous.

However, despite the simplicity of this physical description, there is surprising difficulty in obtaining an accurate mathematical one. The challenge arises precisely from what makes the physics deceptively simple: the slowness of the inspiral. If we continue to think of the inspiral as an evolution through a sequence of geodesics, we can describe it as a slow change of the "constants" of motion, E, L_z, and Q. They change at the rate $\dot{E}/E \propto m/M^2$, introducing the large time scale $\sim M^2/m$ into the system. This time scale is traditionally called the *radiation-reaction time*, denoted $t_{\rm rr}$. Just as the object's small size $\sim m \ll M$ led to a failure of ordinary perturbation theory (motivating the use of matched expansions), the presence of the large scale $t_{\rm rr}$ does likewise. To elucidate this, let us write Eq. (20) in the more explicit form of an ordinary Taylor series,

$$\mathsf{g}_{\mu\nu}(x,\eta) = g_{\mu\nu}(x) + \eta h^{(1)}_{\mu\nu}(x) + \eta^2 h^{(2)}_{\mu\nu}(x) + O(\eta^3), \tag{67}$$

where x^μ is any suitable set of background coordinates, such as the Boyer-Lindquist coordinates associated with the central black hole, and $\eta := m/M$ is the small mass ratio. (Here we have set $\epsilon = 1$ and for clarity redefined each of the perturbations $h_{\mu\nu}^{(n)}$ with a factor η^n pulled out.) Now, it should be clear that the metric perturbation produced by the object will depend on the object's worldline z^μ, which may lead us to expect that each of the perturbations $h_{\mu\nu}^{(n)}$ depend on z^μ. But the worldline satisfying the equation of motion (46) plainly depends on η, while the coefficients $h_{\mu\nu}^{(n)}$ in the expansion (67) are independent of η. It follows that $h_{\mu\nu}^{(n)}$ cannot depend on the whole of z^μ. Instead, we can only utilize an expansion of the form (67) if we also expand z^μ in the same way:

$$z^\mu(\tau, \eta) = z_0^\mu(\tau) + \eta z_1^\mu(\tau) + \eta^2 z_2^\mu(\tau) + O(\eta^3). \tag{68}$$

In this treatment of the worldline, the zeroth-order worldline z_0^μ is a geodesic of the background spacetime $g_{\mu\nu}$, and the GSF introduces small corrections to that worldline. Instead of equations of the form (46), one obtains evolution equations for the individual terms z_n^μ. $h_{\mu\nu}^{(1)}$ depends only on z_0^μ and creates the GSF that drives z_1^μ; $h_{\mu\nu}^{(2)}$ depends on z_1^μ in addition to z_0^μ, and it (together with $h_{\mu\nu}^{(1)}$) drives z_2^μ; and so on. This approach, first used systematically by Gralla and Wald [53] and carried to second order by Gralla [54], is the only consistent way to apply ordinary perturbation theory to the problem. Unfortunately, it is not suitable for treating long-term effects. Suppose the small object initially moves tangentially to a geodesic z_0^μ. As the inspiral progresses, the object moves further and further from z_0^μ, until the expansion (68) breaks down.[17]

So ordinary perturbation theory fails. To account for the long-term changes in the worldline, we must avoid the expansion (68) of z^μ, and hence the expansion (67). Instead, we naturally seek an asymptotic expansion that allows the metric perturbation to depend on the full, unexpanded z^μ. We may write this as

$$\mathsf{g}_{\mu\nu}(x, \eta) = g_{\mu\nu}(x) + \eta h_{\mu\nu}^{(1)}(x; z) + \eta^2 h_{\mu\nu}^{(2)}(x; z) + O(\eta^3), \tag{69}$$

with each coefficient containing an implicit functional dependence on z^μ. This is the type of expansion implicitly used in the preceding sections. It leads to the coupled system of equations (46), (41), and (42). It is called the *self-consistent* approximation, a name first applied by Gralla and Wald [53], indicating that in it, the trajectory z^μ is obtained by solving the coupled field equations and equation of motion together, self-consistently. This self-consistent formulation was used in the original derivations of the MiSaTaQuWa equation, and it was put on a systematic basis, extendable to any order in perturbation theory, in Ref. [55], which first explicitly introduced the expansion (69).

The self-consistent approximation successfully eliminates the growing errors associated with the expansion (68), and it has the advantage of being formulated in a generic (vacuum) background spacetime. However, it is not quite ideal for the EMRI problem. It does not capitalize on the particular, adiabatic character of the orbital

---

[17] This breakdown generally occurs not on the radiation-reaction time, but on the much shorter *dephasing time*, $t_{dph} \sim M/\sqrt{\eta}$. As we can guess from the form of the equation (46), the most rapid growth in z_1^μ is quadratic in time, such that ηz_1^μ grows to order η^0 after t_{dph}; this growth occurs specifically in "phase" variables such as the azimuthal phase φ and the angle variables q^α. The dephasing time hence corresponds to the time over which the gravitational wave sourced by a geodesic particle grows significantly out of phase with the wave sourced by a self-accelerated one.

inspiral, which is only slowly evolving and hence very nearly triperiodic. Furthermore, it is not designed to accurately incorporate a second type of slow change in the system: the slow evolution of the large black hole. Over time, the black hole absorbs energy and angular momentum in the form of gravitational waves, causing its mass and spin to slowly change. Because the expansion (69) does not have a built-in way to accurately track this long-term effect, a naive implementation of the self-consistent approximation would exhibit growing errors, just as the expansion (67) exhibited growing errors due to the long-term changes in the orbit. While Eq. (69) can be tweaked to eliminate this inaccuracy [126], an approximation more specifically tailored to EMRIs is preferable. Such an approximation is offered by a *multiscale expansion* (also known as a "two-timescale" expansion), a method of singular perturbation theory [45] that, in the case of an EMRI, expresses the evolving worldline as a function of both "slow time" and "fast time" variables. By likewise writing the metric perturbation in terms of these variables, we may split the equation of motion and field equations into corresponding slow- and fast-time equations. At fixed values of slow time, the fast-time equations have the same triperiodicity as a background geodesic, providing "snapshots" of the inspiral on the orbital time scale. The slow-time equations then govern the smooth evolution from one snapshot to the next.

This approximation scheme is most easily described in terms of the action-angle variables (J_α, q^α) introduced in the previous section. As a slow time, we may use $\tilde{t} = \eta t$;[18] when Boyer-Lindquist time t is comparable to the radiation-reaction time (i.e., $t \sim M/\eta$), the slow time is of order 1 (i.e., $\tilde{t} \sim M$). As fast-time variables, we can use the angle variables q^α. In place of Eq. (67) or (69), the expansion of the metric then becomes

$$\mathrm{g}_{\mu\nu}(t, x^a, \eta) = g_{\mu\nu}(x^a) + \eta h^{(1)}_{\mu\nu}(x^a, \tilde{t}, q^\alpha) + \eta^2 h^{(2)}_{\mu\nu}(x^a, \tilde{t}, q^\alpha) + O(\eta^3), \qquad (70)$$

where x^a can be any set of coordinates on spacetime slices of constant t, such as the Boyer-Lindquist coordinates $x^a = (r, \theta, \varphi)$. The coefficients in the expansion are required to be bounded functions of $\tilde{t}$, such that they do not grow large with time, and to be periodic functions of q^α, with Fourier expansions

$$h^{(n)}_{\mu\nu} = \sum_{k_\alpha} h^{(n,k_\alpha)}_{\mu\nu}(x^a, \tilde{t}) e^{-ik_\alpha q^\alpha}, \qquad (71)$$

where the sum runs over sets of integer constants k_α. From the coefficients $h^{(n,k_\alpha)}_{\mu\nu}$, one can obtain the slow evolution of the "constants" J_α, which instead of Eq. (66) now satisfy an equation of the form

$$\frac{dJ_\alpha}{dt} = \sum_{k_A} \left[\eta G^{(1,k_A)}_\alpha(\tilde{t}) + \eta^2 G^{(2,k_A)}_\alpha(\tilde{t}) + O(\eta^3) \right] e^{-ik_A q^A}. \qquad (72)$$

The Fourier coefficients $G^{(n,k_A)}_\alpha$ are constructed from those of the GSF (and thence, from the coefficients $h^{(n,k_\alpha)}_{\mu\nu}$). Here, unlike in Eq. (71), the sum is over pairs of integers $k_A = (k_r, k_\theta)$; because the background is stationary and axially symmetric, the "forces" in Eq. (72) are independent of q^t and q^ϕ. Once the evolution of the action variables is determined, one can obtain the evolution of the angle variables simply by

---

[18] Refs. [127, 90] discuss more refined choices of slow time.

integrating [19]

$$\frac{dq^\alpha}{dt} = \omega^\alpha(J_A) = \omega^\alpha_{(0)}(\tilde{t}) + \eta\omega^\alpha_{(1)}(\tilde{t}) + O(\eta^2);\tag{73}$$

that is,

$$q^\alpha = \frac{1}{\eta}\int \omega^\alpha d\tilde{t} = \frac{1}{\eta}\left[q^\alpha_{(0)}(\tilde{t}) + \eta q^\alpha_{(1)}(\tilde{t}) + O(\eta^2)\right].\tag{74}$$

Finally, with $J_\alpha(t,\eta)$ and $q^\alpha(t,\eta)$ known, one obtains the metric perturbations (71) as ordinary functions of t, x^a, and η.

A formulation of this sort was first put forward by Hinderer and Flanagan [9], who developed the mulstiscale expansion of the equation of motion. Besides working out the basic formalism, they also used it to establish which ingredients are needed to compute each of the terms in Eqs. (72) and (74). This is particularly important in the case of the phase evolution, since accurately tracking the phase of the waveform is the critical requirement for matched filtering. Hinderer and Flanagan found that in order to obtain the leading term in the phase evolution (74), one requires only the time-averaged dissipative piece of the first-order force; they refer to this as an "adiabatic approximation". In order to obtain both the leading and subleading term in (74), one requires the entire first-order force, along with the time-averaged dissipative piece of the second-order force; Hinderer and Flanagan refer to this as a "post-adiabatic approximation". Because the subleading term in (74) is of order 1, a post-adiabatic approximation is necessary for accurate modelling, and we can conclude that EMRI science requires at least part of the second-order metric perturbation. This has been the primary motivation for developing second-order GSF theory, and our primary reason for including a description of it in this review.

Given its clear advantages, the multiscale expansion is the most promising method of tackling long-term evolution, and following Hinderer and Flanagan's expansion of the equation of motion, work has recently begun on the corresponding expansion of the field equations [90, 126]. Yet the multiscale method has failings of its own. One limitation is that it breaks down on large distances, where length scales become comparable to the radiation-reaction time scale [90]. This failure manifests itself as an infrared divergence in the retarded solution. A similar failure can also occur near the large black hole's event horizon, which (in the context of wave propagation) plays a role similar to that of infinity. Overcoming these failures calls for the introduction of additional, complementary expansions near infinity and the horizon, which can be combined with the multiscale expansion by once again appealing to the method of matched asymptotic expansions.

5.3. Transient resonances

Even putting aside the effects on large distances, the multiscale expansion encounters an entirely different, equally problematic phenomenon: transient orbital resonances. Generically, away from a resonance, all the $k_A \neq 0$ modes in Eq. (72) are oscillatory, averaging out to zero on the radiation-reaction time. The long-term, average evolution is then driven by the approximately constant, $k_A = 0$ modes in Eq. (72), giving rise

---

[19] In general, the right-hand side of Eq. (73) will contain oscillatory dependence on q^α, but such oscillations can be removed with a near-identity transformation [125, 128]. (We return to this type of transformation in Sec. 6.2 below.)

to an equation of the form

$$\frac{d\langle J_\alpha \rangle}{dt} = \eta G_\alpha^{(1,0)}(\tilde{t}) + \eta^2 G_\alpha^{(2,0)}(\tilde{t}) + O(\eta^3), \tag{75}$$

where $\langle \cdot \rangle$ denotes the average over the (q^r, q^θ) torus. However, this situation changes near a resonance, where for some period of time, one of the $k_A \neq 0$ modes becomes approximately stationary. We can see how this occurs, and its dynamical consequences, by examining how the phase $\psi = k_A q^A$ evolves near a resonance. Suppose a resonance occurs at a time $t_{\rm res}$, meaning that the ratio $\omega^r(t_{\rm res})/\omega^\theta(t_{\rm res})$ is rational. For some integers $k_A = k_A^{\rm res}$, then, the combination $k_A^{\rm res}\omega^A(t_{\rm res}) = k_r^{\rm res}\omega^r + k_\theta^{\rm res}\omega^\theta$ vanishes. Now consider the *resonant phase* $\psi^{\rm res} := k_A^{\rm res} q^A$ near $t_{\rm res}$. Expanding $q^A(t)$ around $q^A(t_{\rm res})$, recalling $dq^A/dt = \omega^A(\tilde{t})$, we obtain

$$\psi^{\rm res}(t) = \psi^{\rm res}(\tilde{t}_{\rm res}) + k_A^{\rm res}\omega^A(\tilde{t}_{\rm res})(t-t_{\rm res}) + \frac{\eta}{2}k_A^{\rm res}\left.(d\omega^A/d\tilde{t})\right|_{\tilde{t}_{\rm res}}(t-t_{\rm res})^2 + O[(t-t_{\rm res})^3], \tag{76}$$

In usual circumstances, away from a resonance, the second term on the right dominates the evolution, and the phase varies on the orbital timescale $\sim 1/(k_A\omega^A) \sim M$, causing $e^{-ik_A q^A}$ to oscillate and average to zero. But because $k_A^{\rm res}\omega^A(\tilde{t}_{\rm res}) = 0$, near the resonance the third term dominates the evolution. The phase then varies slowly, on the long timescale $\sim 1/\sqrt{\eta k_A^{\rm res}(d\omega^A/d\tilde{t})} \sim M/\sqrt{\eta}$. Over periods shorter than this, the terms $G_A^{(n,k_A^{\rm res})}e^{-i\psi^{\rm res}}$ in Eq. (72) become approximately stationary. They then appear as additional terms on the right-hand side of Eq. (75), altering the average rate of change. After the time $M/\sqrt{\eta}$, when the orbit has completed its passage through resonance, the additional driving term in Eq. (75) will have shifted the action variables J_α by an amount $\sim (dJ_\alpha/dt)/\sqrt{\eta} \sim \sqrt{\eta}$. This induces a corresponding order-$\sqrt{\eta}$ shift to the orbital frequencies, inducing an order-$\sqrt{\eta}$ term in Eq. (73). Since the remainder of the inspiral, after the resonance, lasts a time of order M/η, the shift in the frequencies leads to a dramatic, $\sim 1/\sqrt{\eta}$ cumulative shift in the orbital phases.

The emergence of the dynamical time scale $M/\sqrt{\eta}$ near resonances violates the assumed form (70) of the metric, causing the multiscale approximation to break down. Like the failures at large distances, this can presumably be overcome by introducing another approximation scheme near resonance, one that accurately accommodates the new time scale there. However, even if the passage through resonance is accurately modelled, resonances nevertheless lead to an overall loss of accuracy—a subtle consequence of the sensitive dependence of the dynamics on the resonant phase $\psi^{\rm res}$. As illustrated in Fig. 5, the "shape" of the resonant orbit is strongly dependent upon the value of $\psi^{\rm res}$. Orbits with the same principal parameters (e.g., the same $\{E, L_z, Q\}$) can look very different depending on that value. Unsurprisingly, this leads to a situation where the details of the radiative dynamics across the resonance depend sensitively on $\psi^{\rm res}$. Indeed, it has been demonstrated with explicit calculations of radiative fluxes from resonant orbits [129] that $\langle dE/dt \rangle$, $\langle dL_z/dt \rangle$ and (especially) $\langle dQ/dt \rangle$ vary significantly as functions of $\psi^{\rm res}$. Thus, to model the radiative transition across a resonance one must know the exact phase of the orbit as it enters the resonance.

To appreciate the consequences of such phase dependence, suppose we include the first- and second-order forces prior to the resonance, expecting this to suffice for an accurate model. The error in $\psi^{\rm res}$ upon entering the resonance will then be small, $\sim \epsilon$, but this will generate an order-ϵ^2 error in the driving force, leading to an order-$\epsilon^{3/2}$
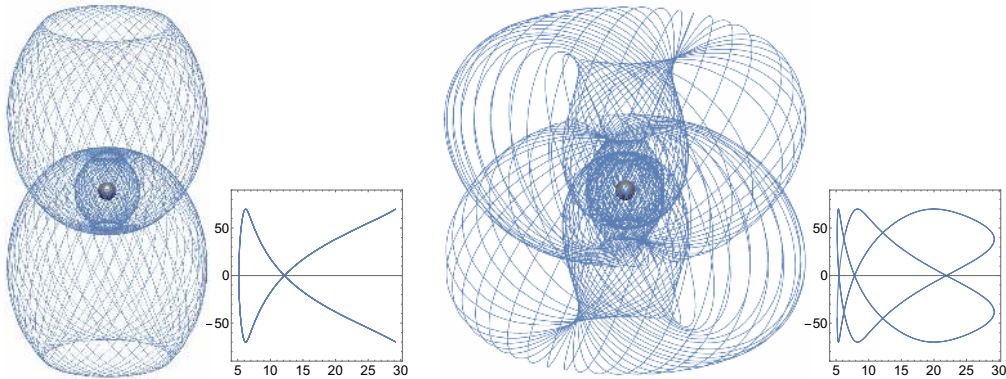
40

Figure 5: The shape—and radiative dynamics—of a resonant orbit depend sensitively on the resonant phase ψ^{res}, here defined as the value of the longitudinal angle θ at a periapsis of the resonant orbit. The two 3:2 resonant orbits viewed here, and the orbit shown on the right in Fig. 4, all have the same parameters $\{E, L_z, Q\}$ but different resonant phases ($\psi^{\mathrm{res}} = 20°$ for the orbit in Fig. 4, $\psi^{\mathrm{res}} = 90°$ for the orbit on the left here, and $\psi^{\mathrm{res}} = 45°$ for the orbit on the right here). The two orbits are viewed from just above the equatorial plane. Next to each orbit we display the corresponding Lissajous curve, showing the orbit in the plane of radius r (horizontal axis, in units of M) and inclination angle (vertical axis, measured from the equator in degrees).

error in the frequencies, and from there an order-$\sqrt{\epsilon}$ error in the phases. While that error remains small, it is larger than the order-ϵ error we may have expected from including the second-order force. Similarly, if the orbit then goes on to encounter another resonance, this order-$\sqrt{\epsilon}$ phase error will be boosted to an order-1 error. With each successive passage through resonance, we lose half an order of phase accuracy.

However, we should note that a strong resonant effect of this sort only arises when k_A^{res} are *small* integers. It is then that the effect of the otherwise-oscillatory forcing terms in Eq. (72) can build up coherently over a long time. In contrast, high-order resonances with large values of k_A^{res} are efficient in "averaging away" such coherent buildup, so their resonant effect is suppressed and generally negligible. Studies [130, 131] show that essentially all astrophysically relevant systems encounter at least one strong resonance while in the LISA band. A recent study [132] suggested, however, that LISA-relevant EMRIs are likely to pass through only one such resonance. It estimated that, as a result, resonances are expected to reduce the number of EMRI detections with LISA by no more than $\sim 4\%$.

The basic dynamical features of transient resonances in EMRIs were first described by Flanagan and Hinderer in [133], in the framework of the two-timescale expansion. (An alternative, but equivalent, mathematical description was later given by Gair and collaborators [134, 132].) Ref. [133] demonstrated with an explicit calculation, using a simple post-Newtonian model for the GSF, that a low-order resonant crossing of a typical LISA EMRI can last a few hundred orbital cycles. This was later confirmed in [130] using a more accurate model of the radiation reaction. The impact of resonances on EMRI modelling has been examined in several subsequent works [129, 135, 125, 136, 137, 131]. A complete treatment, including the expansion of the Einstein equations and matching to the multiscale expansion, is still lacking.

5.4. Spin and finite-size effects

So far in our discussion of orbital evolution, we have not specified the form of the force. In Sec. 3.5, we indicated that in addition to the GSF, a small compact object is also subject to forces arising from its finite size. While the GSF arises from a somewhat indirect interaction of a small object with its own field, the finite-size forces arise from direct couplings between the object's multipole moments and the external curvature. This is first seen in the Papapetrou spin term in the first-order equation of motion (44). At each higher order, one additional pair of multipole moments appear (one mass moment and one current moment). Corrections to the lower-order forces may (potentially) also arise. So, for example, in the second-order equation of motion, we expect the object's quadrupole moments to appear, presumably in precisely the form they do for a test body [13] (though Ref. [36] has indicated some subtleties in this expectation). We may also expect a correction to the Papapetrou term, coming from coupling of the object's spin to the curvature of $h_{\mu\nu}^{\mathrm{R}(1)}$. Finally, the spin will act as a source in the second-order field equation (or equivalently, appear as a term in the second-order puncture), thereby altering $h_{\mu\nu}^{\mathrm{R}(2)}$ and the GSF that arises from it.

At present, the precise form of these finite-size effects remains to be confirmed; as mentioned in Sec. 3.5, all derivations of the second-order equation of motion have specialized to approximately spherical, nonspinning objects. However, based on the two-timescale analysis, we can assess, in advance, which of these terms are required to accurately model an EMRI. Specifically, we require only the time-averaged, dissipative piece of the second-order force. For a test body, the force arising from the quadrupole moments [138] seems purely conservative, suggesting that this term is not needed for EMRI modelling. On the other hand, there are tentative hints that the correction to the Papapetrou force will contain dissipative effects [139]. (Note also that spin introduces a new frequency, necessitating the introduction of a new phase variable in the multiscale expansion.) In any case, second-order finite-size forces certainly warrant further investigation.

As part of our discussion of actual EMRI calculations in Sec. 6 below, we will review work to incorporate the first-order, Papapetrou force into EMRI models.

5.5. Gauge dependence

One crucial issue we have not yet broached is that of the gauge dependence of the orbital evolution. Because the GSF is entirely gauge dependent, the orbital evolution may appear to be pure gauge. In a superficial sense, this is true: as an extreme example, we could choose a gauge in which the GSF identically vanishes, in which case no inspiral would seem to occur at all. However, this choice would lead to secular growth of the metric perturbation, causing perturbation theory to eventually break down. To avoid this and remain within the realm of perturbation theory, we must restrict the class of allowed gauges. For example, we may confine our attention to gauges compatible with the multiscale expansion (70). Physically, this corresponds to adopting a coordinate system that respects the orbit's physical triperiodicity on short time scales. Such restrictions also play an important role in the computation of various physical quantities, as described in Sec. 7.

6. Actual calculations for EMRIs: dissipative effects and orbital evolution

We now begin our survey of concrete results from numerical GSF calculations in EMRI systems. We have compiled two sets of such results. The first, to be covered in the present section, explores the radiative effects of the GSF and their impact on the orbital evolution. We focus particularly on two approximation schemes that have been employed to model the evolution; these methods are distinct from the two-timescale expansion described above, but we can assess their accuracy using the two-timescale framework. The second set of results, to be covered in Sec. 7, comprises calculations of various specific conservative physical effects of the GSF (along with other quantities constructed from $h_{\mu\nu}^{\mathrm{R}}$). Section 8 then presents tests of GSF results against other approaches to the two-body problem, and describes the productive synergy being pursued between GSF physics and these approaches.

6.1. Balance laws and adiabatic evolution

In the previous section, we noted that to accurately model an EMRI waveform, we must work to at least second order in perturbation theory. However, we also described the inspiral as approximately adiabatic, slowly evolving from one geodesic to the next. For the purposes of simply detecting a sufficiently bright EMRI, without accurately identifying the source orbit's parameters, this leading-order, adiabatic approximation may well suffice.

In its most primitive form, such an approximation corresponds to simply combining the slowness of the inspiral with the conservation of energy and angular momentum. To motivate this, we return to what is perhaps the most intuitive aspect of GSF physics: back-reaction from emission of gravitational waves. The dissipative piece of the GSF "does work" on the particle, dissipating its orbital energy and angular momentum and thereby driving its gradual inspiral deeper and deeper into the potential well of the central black hole. The particle's lost energy and angular momentum is transferred into the gravitational field, which then carries them away to infinity in the form of gravitational waves. In this picture, the loss of orbital energy and angular momentum is "balanced" by the emitted radiation. This intuitive picture is basically a valid one (at least at first order in η), except that—unlike in the analogous flat-space relativistic electrodynamics problem that inspires it—such a balance cannot usually be established in a momentary sense: there is no meaningful way of relating the momentary rate of (say) orbital-energy dissipation to the momentary flux of energy in the gravitational waves. (This impossibility traces back to the fundamental absence of a notion of local energy in GR.)

However, even in GR, in certain circumstances, it is possible to formulate balance relations that hold in a certain "time-averaged" sense. For example, it was shown in [140] that, for a freely falling point mass whose orbit begins and ends at infinity, the total work done by the first-order GSF exactly balances the total energy carried away in (linearized) gravitational waves. More pertinently for our context, it has been shown for bound geodesic orbits in Kerr that balance relations hold, within an adiabatic approximation (and at first-order in η), when a suitable orbital averaging is applied [141, 142, 129] (see also [143] for early work to this end). These "balance

laws" for energy and angular momentum take the simple form

$$\langle F_t^{\mathrm{diss}}/u^t \rangle = \langle \dot{\mathcal{E}}_\infty \rangle + \langle \dot{\mathcal{E}}_{\mathcal{H}} \rangle,$$
$$-\langle F_\varphi^{\mathrm{diss}}/u^t \rangle = \langle \dot{\mathcal{L}}_\infty \rangle + \langle \dot{\mathcal{L}}_{\mathcal{H}} \rangle, \tag{77}$$

where on the right-hand side are time-averaged asymptotic fluxes of energy ($\dot{\mathcal{E}}$) and angular momentum ($\dot{\mathcal{L}}$) out to infinity ('∞') and down the event horizon of the black hole ('$\mathcal{H}$'), and throughout the rest of this section an overdot denotes d/dt. Following the logic of the adiabatic approximation, quantities on both sides are calculated while approximating the orbit as a fixed geodesic: for example, F^μ is the GSF generated by the field $h_{\mu\nu}^{\mathrm{R}(1)}$ of a point mass moving on a geodesic. We have used $\langle \cdot \rangle$ to denote time averaging (with respect to time t); for intrinsically periodic geodesic orbits (any orbit in Schwarzschild, or circular, equatorial or resonant orbits in Kerr) it suffices to average over one period of the motion, but in general (for ergodic geodesics such as the one on the left in Fig. 4) one must average over an infinite amount of time (while still treating the orbit as a fixed geodesic), or, equivalently, over the orbital (q^r, q^θ) torus. On the left-hand side of the balance equations (77) are the averaged t and φ components of the local GSF, normalized by the time component of the local four-velocity, u^t (which simply translates from the local proper time used in the definition of the GSF to the usual coordinate time used in defining the asymptotic fluxes). The left-hand sides here are always positive, and so are the fluxes $\langle \dot{\mathcal{E}}_\infty \rangle$ and $\langle \dot{\mathcal{L}}_\infty \rangle$. However, the fluxes $\langle \dot{\mathcal{E}}_{\mathcal{H}} \rangle$ and $\langle \dot{\mathcal{L}}_{\mathcal{H}} \rangle$ of radiation absorbed by the black hole can be either positive or—for certain orbits in Kerr—negative. Negative horizon fluxes mark *superradiant* behavior, in which some of the black hole's rotational energy and angular momentum are, in effect, transferred *to* the orbit [144, 145, 146].

We may also write the balance laws (77) in the more suggestive form

$$\langle \dot{E} \rangle = -\langle \dot{\mathcal{E}}_\infty \rangle - \langle \dot{\mathcal{E}}_{\mathcal{H}} \rangle,$$
$$\langle \dot{L}_z \rangle = -\langle \dot{\mathcal{L}}_\infty \rangle - \langle \dot{\mathcal{L}}_{\mathcal{H}} \rangle. \tag{78}$$

If we compute the fluxes on the right-hand side for a given geodesic with parameters E and L_z, these laws allow us to evolve to a new geodesic with parameters $E + \langle \dot{E} \rangle \delta t$ and $L + \langle \dot{L}_z \rangle \delta t$. This gives us a way of deducing the dominant, adiabatic evolution without actually resorting to a calculation of the local GSF. Computational methods for asymptotic fluxes in black-hole perturbation theory have been well developed since the early 1970s, and have been performed many times using frequency-domain methods (e.g, [144, 147, 145, 146, 148]) as well as time-domain ones [149, 150, 91, 92, 151]. Such calculations are much less computationally expensive than local GSF calculations, and can be done in the convenient framework of Teukolsky's perturbation formalism, working with Ψ_0 or Ψ_4 instead of the full metric perturbation. This approach has been taken by Hughes and collaborators [152, 153] in order to compute the adiabatic evolution of (certain types of) EMRIs, circumventing the need for an explicit GSF calculation.

Recall, however, that generic geodesic orbits in Kerr are characterized by *three*—not two—constants of motion: to get a full description of the adiabatic evolution for a generic orbit one must also be able to calculate the evolution of the Carter constant Q, in addition to the evolution of E and L_z. It is not difficult to write down $\dot{Q}$ directly in terms of the local GSF [154], but there is no known way to relate $\langle \dot{Q} \rangle$ to some asymptotic fluxes of radiation. This problem was solved, at the practical level,

thanks to a breakthrough idea by Mino [123, 155] and follow-up work by Sago and collaborators [141, 142], who derived a simple, practical formula for $\langle \dot{Q} \rangle$. The formula involves both quantities that are encoded in the asymptotic radiation (and are readily calculable within the Tuekolsky formalism), and quantities that are locally defined as integrals along the orbit (but are nonetheless also easily calculated). With this, it is now possible to calculate the evolution of generic EMRI orbits in Kerr, at leading, adiabatic order, without performing an actual GSF calculation. Although originally formulated outside the context of a two-timescale expansion, it can be shown that this evolution scheme is precisely equivalent to solving the leading-order equations in the two-timescale approximation (which Hinderer and Flanagan referred to as "adiabatic order" for that reason). There is an ongoing programme to numerically implement this approach [156, 157, 151, 153].

In addition to enabling adiabatic evolution, the balance law (77) provides an important benchmark for GSF calculations. GSF codes all involve both a calculation of the local GSF and a calculation of the global perturbation field, from which the fluxes $\langle \dot{\mathcal{E}}_{\infty/\mathcal{H}} \rangle$ and $\langle \dot{\mathcal{L}}_{\infty/\mathcal{H}} \rangle$ may readily be extracted. Thus, GSF codes offer the opportunity to test the balance relations (77). [Or, conversely, depending on one's point of view, the balance relations (77) offer the opportunity to test GSF codes.] Explicit numerical calculations demonstrating the validity of (77) have been carried out for a variety of cases, including circular [92] and eccentric [93] orbits in Schwarzschild, eccentric orbits in the equatorial plane of a Kerr black hole [78], and, most recently, generic (eccentric and inclined) bound orbits in Kerr [79].

Finally, the balance laws have also provided an important test of *second*-order calculations. In a recent series of papers [62, 158, 90, 159], practical methods and computational tools have been developed for calculations of the second-order perturbation produced by a particle on a quasicircular inspiral orbit in Schwarzschild geometry. The monopolar (spherically symmetric) piece of the second-order field equation (42) contains several dissipative terms, including a term equal to the flux of energy at infinity, a term equal to the flux of energy through the horizon, and a term describing the rate of orbital-energy dissipation (note that all these quantities are, in fact, second order in η). It can be analytically shown that this portion of the field equation precisely yields the balance law (77). This result, yet to be published, was derived by one of us (AP) recently. It can be said to be a first concrete physical outcome from a second-order GSF calculation. Because it involves nontrivial properties of the puncture $h_{\mu\nu}^{\mathcal{P}(2)}$ appearing in Eq. (42), it stands as a major test of the second-order formalism.

6.2. Computation of inspiral orbits with the full (first-order) GSF

The leading-order, adiabatic, flux-based calculations described above capture the main, average dissipative effect of the (first-order) GSF, but they completely neglect the conservative piece of the force, as well as subleading effects of the dissipative first-order force that average out at leading order in the two-timescale approximation. As discussed in Sec. 5.2, both types of neglected effects of the first-order GSF contribute at the first post-adiabatic order, together with that of the averaged dissipative second-order GSF. These post-adiabatic terms, including the conservative piece of the GSF, do have an important secular effect on the phase evolution in EMRI systems [160], and as already stressed, it is important to include them in EMRI models.

The most direct way of incorporating these effects would be to directly integrate

45

the equation of motion in a self-consistent manner, by solving the coupled pair of equations (44) and (41). This is a computationally challenging task. In a time-domain implementation, one would need to compute the GSF time-step by time-step, and at each step accelerate the orbit that sources the field by a suitable amount. This direct calculation has only been attempted so far using a scalar-field toy model [161] with short orbital sequences; computing entire EMRI inspirals using this method does not seem to be a realistic prospect without a major improvement in numerical methodology. Frequency-domain implementations seem to be even less suitable for such direct integrations of the equation of motion, as, at least in their current form, they all assume a fixed geodesic source, with a given, unevolving frequency spectrum.

In Sec. 5.2, we have advocated for the two-timescale approach as an alternative (one which would be amenable to frequency-domain methods). However, to date there has been no direct implementation of that approach. Instead, orbital-evolution calculations so far have been based on the idea of *osculating geodesics* (named after the method of osculating orbits in Newtonian celestial mechanics). In this approach, the inspiral orbit is reconstructed as a smooth sequence of geodesics, each lying tangent to (or "osculating") the true orbit at a particular moment. This amounts to modelling the true orbit as an evolving geodesic with dynamical orbital elements. At each instant t_0, one then approximates the GSF by computing it as if, for all $t < t_0$, the particle had been moving on the geodesic that is instantaneously tangential to the evolving orbit at that instant t_0. This GSF is then used to calculate the momentary rate of change in each of the orbital elements—principal as well as, importantly, positional; this is a key difference from the adiabatic approximation, which likewise models the orbit as a smooth sequence of geodesics, but does not account for the GSF's conservative effect on the evolution of the positional elements. Equations governing the forced evolution of all orbital elements in the Schwarzschild case were obtained in Ref. [162], and Ref. [163] later generalized the formalism to Kerr.

It is important to point out that the osculating-geodesics equations are precisely equivalent to the original equations of motion; they themselves involve no approximation, as long as the forcing term is known exactly for the evolving orbit. The only non-numerical source of error introduced in this procedure is in the use of "geodesic" GSF information (i.e., the GSF calculated while treating the orbit as a fixed geodesic rather than as the true, evolving orbit). The resulting error in the momentary self-acceleration formally scales with the square of the small mass ratio, and is *a priori* comparable to the error introduced by neglecting the second-order GSF. Assessment of the actual magnitude of error from using the "geodesic" GSF would require comparison with calculations based on a direct time-domain evolution.

To implement the above idea in practice, rather than calculating the "geodesic" GSF on the fly, one prepares a large database of GSF values covering the region of interest in the parameter space of geodesic orbits (this is a 2D space in the case of a nonspinning particle in Schwarzschild, and a 3D space if either the particle or central black hole is spinning). Each entry in that database (labelled by, e.g., $\{E, L_z, Q\}$) contains numerical GSF values calculated as a function of the phase variables (e.g., $\{q^r, q^\theta\}$) along a fixed geodesic orbit. This can be done relatively quickly using frequency-domain GSF codes. The data is then interpolated across the space, so that one now has analytic interpolation formulas for the GSF components at each point in the (4D or 6D) phase space of the problem. The preparation of this database is computationally heavy, but done once and for all. One can then use the interpolated GSF information to evolve any orbit, with any given initial conditions.

The first implementations of the osculating-geodesics method, with the full first-order GSF in Schwarzschild and using the "geodesic" GSF approximation, were presented by Warburton and collaborators in [164, 108]. Figure 6, reproduced from Ref. [108], shows sample results from this work. Here, the eccentricity, e, and semilatus rectum, p, are used as convenient, geometrically intuitive parameters for bound geodesic orbits in Schwarzschild. The plot shows the computed evolution of several EMRI orbits with mass ratios $\eta = 10^{-5}$ in the plane of e and p, for a range of initial conditions; each solid (black) curve represents a single EMRI orbit, starting at the blue curve on the right, and evolving leftward (i.e., inward) until the last stable orbit, represented by the near-vertical (red) curve on the left. We can see how radiation reaction drives the orbit to be more and more circular (until very near the final plunge where the eccentricity can briefly increase). Important information is provided in this plot by the dashed curves: these are contours that mark the number of radians by which the periastron position will rotate under the effect of the conservative GSF, from a given point until plunge. Stated differently, they indicate the total phase error that one would be making by neglecting the effect of the conservative GSF. (The negative values of the indicated phase implies that the conservative GSF works against the usual periastron advance.) These results illustrate the importance of accounting for conservative effects in EMRI models.
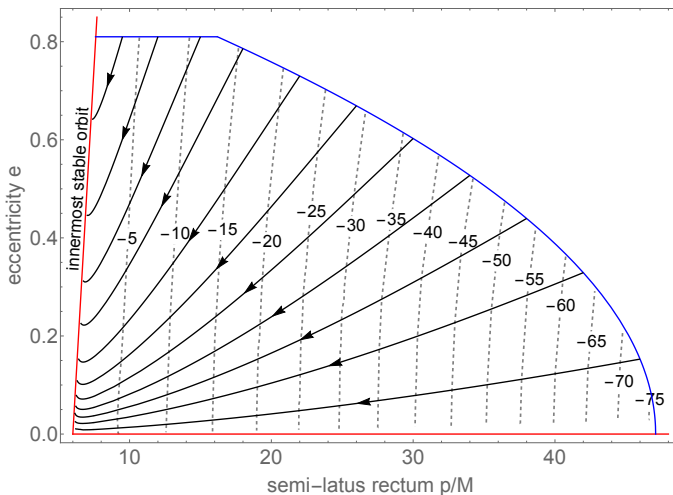

Figure 6: Sample inspiral orbits around a Schwarzschild black hole, as calculated by Osburn *et al.* in [108] for a mass ratio $\eta = 10^{-5}$. Each solid black curve tracks the evolution of a particular EMRI orbit in the e–p plane (eccentricity vs. semilatus rectum), from entering the LISA band (blue curve, assuming a black-hole mass of $M = 10^6 M_\odot$), until reaching the innermost stable orbit (red line on the left). During the evolution, the conservative piece of the GSF acts to decrease the rate of periastron advance; the dashed contour lines (with associated numerical values) indicate the total amount of periastron phase (in radians) accumulated due to this effect, from a given moment until plunge.

The full inspiral trajectories of the osculating-geodesics method are still rather slow to evaluate in practice (Ref. [108] reports a computation time of minutes to hours

for each such trajectory, depending on mass ratio). That is because the forcing terms in the evolution equations depend explicitly on the orbital phases, so one must resolve the inspiral trajectory over the short orbital timescale. Van de Meent and Warburton [128] showed how this can be avoided using a simple trick, by means of what is known in the theory of dynamical systems as a *near-identity transformation*. The idea is to apply a "small" transformation to the phase-space variables of the problem, such that the resulting forcing terms in the evolution equations no longer depend on the orbital phase, while the solution to the modified problem remains uniformly close to the original solution. In effect, this procedure "pushes" the phase dependence one order higher in the small parameter of the problem (mass ratio, in our case), effectively averaging away all orbital-scale oscillations at leading order. Ref. [128] applied this method to inspiral orbits in Schwarzschild spacetime, demonstrating the substantial gain in computational efficiency that it offers.

While the osculating-geodesics evolution model of Refs. [164, 108, 128] represents significant progress (compared to the leading-order adiabatic model), it still misses several important pieces, all formally second order in the GSF approximation, but all expected to contribute at the first post-adiabatic order just like the first-order conservative GSF. Missing is the dissipative piece of the second-order GSF, and missing too are the aforementioned corrections due to the use of the "geodesic" GSF approximation. If the small particle is spinning, the Papapetrou force must also be included, along with second-order dissipative forcing terms associated with the spin. The first step in this direction was taken in recent work [165], which explores the impact of the Papapetrou term on the orbital evolution. Ref. [24] has also suggested how the osculating geodesics method might be extended to second order, potentially providing an alternative to the two-timescale expansion.

Work on orbital evolution has so far concentrated on the Schwarzschild problem. However, "geodesic" GSF calculations are now available for generic bound geodesic orbits in Kerr [78, 79]—Figs. 7 and 8 display sample outputs of those calculations. With these results, we expect the osculating-geodesics programme to now extend its reach to the Kerr case. The main hurdle here is computational: one must populate the (now 3D) space of Kerr geodesics sufficiently densely with GSF data, and devise an efficient interpolation method over that space. One would need to extend the application of the near-identity-transformation method [128] to the Kerr case. There is, additionally, ongoing work to develop more efficient ways to compute the orbital evolution in Kerr based on a two-timescale expansion with a judicious choice of orbital elements and gauge [166]. Inspirals in Kerr also generically undergo resonant crossings, which, as discussed before, demand a special treatment.

6.3. Sustained resonances?

An intriguing question, explored in another work by van de Meent [125], is that of the possibility of *sustained* resonances in EMRIs. In the general theory of resonant dynamical systems, a sustained resonance is one in which the driving force does not take the system across the resonance and away from it (as in the usual, transient-type resonances discussed in Sec. 5.3), but rather it keeps the system oscillating around the resonance for a prolonged amount of time—typically $\Delta t \propto \eta^{-1}$ rather than $\eta^{-1/2}$, in the EMRI case. The conditions for a sustained resonance can be deduced from an effective-potential equation satisfied by the resonant phase, of the form $\dot{\psi}^2 = K - V(\psi)$, where $K := V(\psi^{\mathrm{res}})$, and in the EMRI case, the potential V depends on the self-forcing
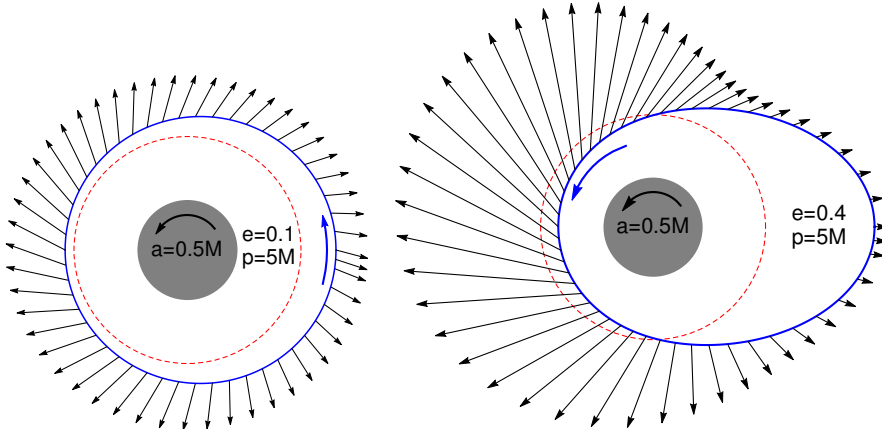
Figure 7: Sample output from van de Meent's code [78], showing the first-order GSF along eccentric geodesic orbits in the equatorial plane of a Kerr black hole. We show here two orbits, of the same semilatus rectum but differing eccentricities: $e = 0.1$ on the left and $e = 0.4$ on the right. In both cases the orbital angular velocity is aligned with the black hole's spin, which is set at half the maximal allowable rate. For clarity, the orbits are plotted in the r–q^r plane, where q^r is the phase of the radial epicyclic motion (not the usual azimuthal phase φ; this is why the actual precessional motion of the orbit is not visible here). The two orbits are shown on the same scale, and the central holes are also to scale. Dashed (red) circles mark the innermost circular stable orbit (ISCO), for reference. Along each orbit we indicate with arrows the spatial projection of the GSF onto the equatorial plane. The normalization of the GSF vector projection was chosen arbitrarily, but it is fixed along each orbit and between the two orbits; the actual magnitude of the GSF projection at the pericenter of the orbit on the left is $\sim 0.0237\eta^2$. The code presented in [78] is a frequency-domain implementation of the radiation-gauge approach with mode-sum regularization, and in its latest version [79] it is capable of computing the GSF along (essentially) any bound geodesic orbit around any Kerr black hole.

terms [125]. A sustained resonance requires both that $V(\psi)$ has a minimum in which ψ can be "trapped", and importantly, that such trapping can actually occur during the normal evolution of the EMRI. The latter condition requires that the system enters the resonance with an appropriate value of K. Ref. [125] established that the "window" of suitable K values, as expressed, e.g., in terms of an interval of orbital phase, is extremely narrow, being proportional to $\eta^{1/2}$. Thus, even if sustained resonances are dynamically possible in EMRIs (i.e., there exist configurations for which V has a minimum), which is not yet clear, the proportion of systems likely to be captured into one will be very small. It appears unlikely that LISA will observe a sustained resonance in an EMRI system.

6.4. Resonant kicks

There is also a different type of resonance that occurs in EMRI systems. Eccentric orbits, even around a Schwarzschild black hole, can experience resonances between
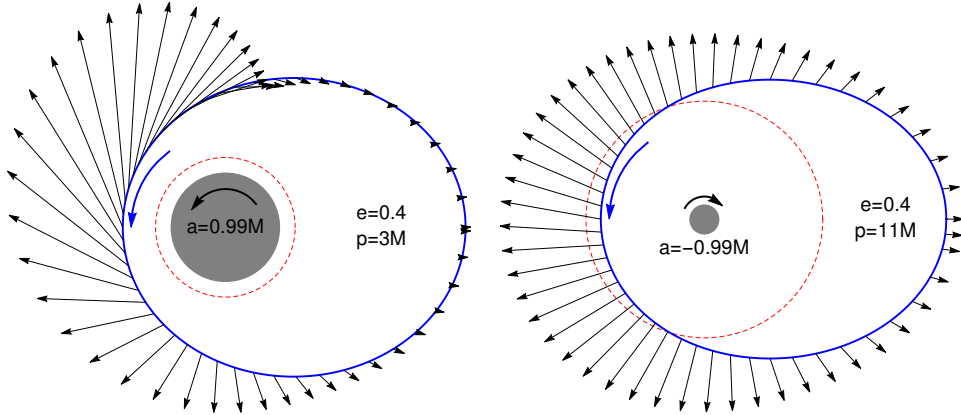
Figure 8: Same as in Fig. 7, this time with a central hole spinning at 99% the maximal allowable rate. The orbit on the left co-rotates with the black hole, while that on the right is counter-rotating. (Counter-rotating orbits encounter their ISCOs relatively far out, so their adiabatic radiative evolution terminates while still relatively far from the event horizon.) Note the two orbits are *not* shown on the same scale, and use a different normalization of the GSF vector. The actual magnitude of the GSF vector projection at pericenter is $\sim 0.3856\eta^2$ for the left orbit and (a mere) $\sim 0.00768\eta^2$ for the orbit on the right.

the epicyclic r motion and the rotational φ motion: instances when $n_r T_r = n_\phi T_\phi$ with n_r, n_φ some (small) integers. In such instances, the precession of the periastron momentarily halts. The intrinsic dynamics is unaffected, but the usual azimuthal symmetry (on average) of the emitted gravitational waves is broken, and the effect of asymmetric emission can build up coherently over a large number of orbits—the episode lasts an amount of time $\propto \eta^{-1/2}$, just as an r–θ resonance. During such an r–φ resonance, gravitational waves carry to infinity a net amount of linear momentum in a particular direction, resulting in a net "kick" to the system's center of mass in the opposite direction. The magnitude of the kick velocity is expected to scale as $\eta^{3/2}$— a factor η^2 for the momentary magnitude of linear momentum in the gravitational waves, times a factor $\eta^{-1/2}$ for the buildup duration.

In Ref. [167], van de Meent performed a detailed analysis of r–φ resonances for equatorial EMRIs in Kerr, numerically computing the magnitudes and directions of the resulting kicks. (This followed a similar analysis by Hirata for θ–φ circular-orbit resonances [168], where, however, the effect appeared to be much less pronounced.) His main results are summarized in Fig. 9. An intriguing feature is the misalignment of the kick direction with the direction of periastron: this is due to the "light-bending" effect of the strong gravitational field around the black hole. The maximal kick velocities attainable are around $v \sim 30,000 \times \eta^{3/2}$ km/sec, for a 1:2 r–φ resonance at a very high eccentricity outside a very rapidly spinning Kerr black hole. Even in such extreme (and astrophysically less likely) configurations, the effect can become observationally important only with η in the not-so-extreme regime. For example, $\eta = 10^{-2}$ yields $v = 30$ km/sec, only just comparable to the escape velocity of a typical globular cluster (the likely nesting place for putative intermediate-mass black holes). It appears very
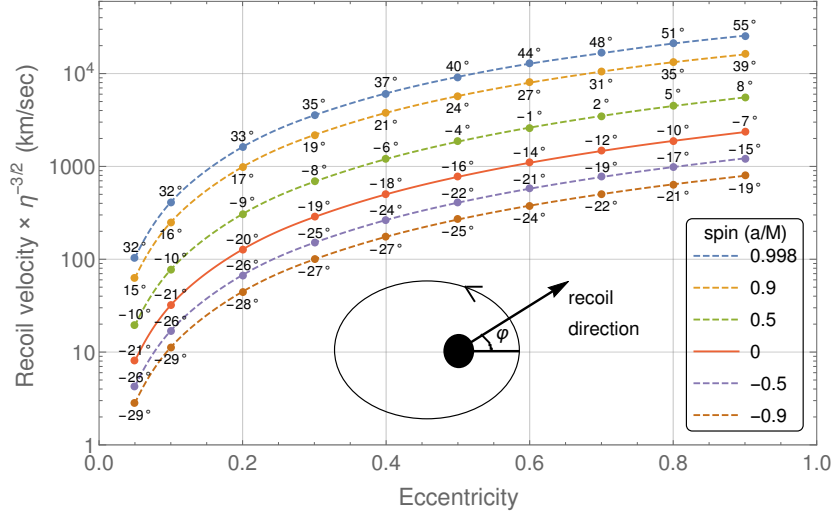
50

Figure 9: Resonant kicks imparted to eccentric equatorial orbits around a Kerr black hole as they encounter a 1:2 r–φ resonance (based on data from Ref. [167]). Plotted is the magnitude of recoil velocity (divided by $\eta^{3/2}$), as a function of orbital eccentricity, for a range of black-hole spins [from top to bottom: $a/M = 0.998, 0.9, 0.5, 0, -0.5$ and -0.9, with positive (negative) values corresponding to prograde (retrograde) orbits]. Indicated next to each data point is the direction of recoil in the equatorial plane, measured as an azimuthal angle φ relative to the on-resonance direction of the periastron (with φ increasing in the direction of motion).

unlikely that resonant kicks will have observational implications for LISA-type EMRIs.

7. Actual calculations for EMRIs: Conservative effects

In our discussion so far we have encountered several times the fact that the GSF in EMRIs has, in addition to the obvious dissipative effect responsible for the emission of gravitational waves and for the orbital decay, also a range of non-dissipative, or *conservative* effects. This was manifest, for example, in Fig. 6, which illustrated how the GSF acts to decrease the rate of periastron advance in what leads to a significant cumulative effect on the long-term phase evolution. This effect comes from a local order-η correction to a *positional* element of the orbit (i.e., a phase), which (at first order) has no impact on the rate of energy or angular momentum dissipation. Heuristically, the presence of the small mass "deforms" the background gravitational potential in a way that also affects the details of the dynamics of positional elements.

At least at first order, this conservative effect can be cleanly disentangled from that of dissipation by writing the GSF as a sum of a time-symmetric piece $F_\alpha^{\rm cons}$ ("conservative force") and a time-antisymmetric piece $F_\alpha^{\rm diss}$ ("dissipative force"),[20] and then considering the equation of motion with the full GSF replaced with either

---

[20] The precise definition of $F_\alpha^{\rm cons}$ and $F_\alpha^{\rm diss}$ alludes to the retarded and advanced metric perturbations (the gravitational analogues of the potentials A_-^μ and A_+^μ of Sec. 2): $F_\alpha^{\rm cons}$ is the force exerted by the perturbation $\frac{1}{2}(h_{\alpha\beta+} + h_{\alpha\beta-}) - h_{\alpha\beta}^{\rm S}$, while $F_\alpha^{\rm diss}$ is exerted by $\frac{1}{2}(h_{\alpha\beta+} - h_{\alpha\beta-}) - h_{\alpha\beta}^{\rm S}$.

$F_\alpha^{\rm cons}$ or $F_\alpha^{\rm diss}$. What we may call the "conservative dynamics" is thus described by the solution of the equation of motion

$$m\frac{D^2 z^\mu}{d\tau^2} = F_{\rm cons}^\alpha(z), \tag{79}$$

obtained from (56) by setting $F_{\rm diss}^\alpha$ to zero.

As a simple example, consider the circular orbit of a non-spinning particle around a Schwarzschild black hole. Working in Schwarzschild coordinates (t, r, θ, φ) and setting (without loss of generality) $\theta = \pi/2$ and $\dot\theta = 0$ at some initial moment, the full (first-order) equations of motion (56) explicitly read, for *any* orbit in Schwarzschild,

$$\dot E \; = \; - F_t, \tag{80}$$

$$m\ddot r = \; -\frac{m}{2}\frac{dV(r; L_z)}{dr} + F^r, \tag{81}$$

$$m\ddot\theta = F^\theta, \tag{82}$$

$$\dot L_z \; = F_\varphi, \tag{83}$$

where $E = m\left(1 - \frac{2M}{r}\right)\dot t$, $L_z = mr^2\dot\varphi$, an overdot denotes (here and throughout this section) differentiation with respect to proper time, and $V(r; L_z)$ is a certain effective potential [see Eq. (85) and Fig. 10 below]. Since, from symmetry, we have $F^\theta = 0$ on the equatorial plane, Eq. (82) tells us that the orbit remains confined to that plane, as expected. In the absence of a GSF, Eqs. (80) and (83) turn into conservation equations for the geodesic energy E and angular momentum L_z. Now, in the specific case of a *circular* orbit, the particular symmetries of the setup imply $F_\alpha^{\rm diss} = \{F_t, 0, 0, F_\varphi\}$ and $F_\alpha^{\rm cons} = \{0, F_r, 0, 0\}$, so that the t and φ component of the GSF are purely dissipative while its r component is purely conservative. To consider the purely conservative dynamics in this example thus amounts to setting $F_t = 0 = F_\varphi$, turning Eqs. (80) and (83) into conservation equations. However, a conservative component of the GSF, F_r, still features in Eq. (81). Let us explore the effect of this conservative forcing term, highlighting the role of gauge freedom in the problem.

To begin, we notice that, in the absence of the GSF, Eq. (81) describes motion in a one-dimensional effective potential. Circular orbits are defined via the conditions $\dot r = 0 = \ddot r$, which determine the values of E and L_z as functions of the orbital radius, r_0. They also determine the value of the orbital frequency[21] $\Omega := \frac{d\varphi}{dt} = \sqrt{M/r_0^3}$ (the same frequency as a Keplerian orbit of the same radius). The conservative forcing term F^r in Eq. (81) modifies slightly the coordinate location of the circular-orbit equilibrium, and with it also the functional relation between Ω and r_0: a short calculation gives

$$\Omega = \sqrt{\frac{M}{r_0^3}}\left(1 - \frac{r_0^2(r_0 - 3M)}{2mM(r_0 - 2M)}F^r\right). \tag{84}$$

We see that the conservative force F^r appears to slightly "shift" (by an amount of order η) the value of the orbital frequency at a given coordinate radius r_0. But there is an important subtlety here. We recall that both the radius and the GSF are gauge dependent. Expressing the perturbed metric in different gauges changes the identification of the value r_0 with a particular physical radius, and it also changes the value of F^r. In fact, a short calculation [92] reveals that a small coordinate

transformation $r \to r - \xi^r$ (where ξ^r is order-η) shifts the radius r_0 and the value of F^r in precisely the right way so as to cancel each other's order-η contribution on the right-hand side of Eq. (84), leaving Ω *invariant* under any such transformation. The orbital frequency, unlike the coordinate radius, is a genuine, physical, "gauge-invariant" attribute of the orbit.[22] In this example, the conservative GSF does not have a clear physical ("observable") significance. Rather, it simply tells us how the invariant frequency is related to radius in any given gauge.

Our first example serves to highlight the kind of subtleties one faces in trying to interpret the physical effect of the conservative GSF, given the gauge ambiguity of the underlying perturbation theory. However, one should by no means conclude that the conservative effect of the GSF is a pure gauge artifact. Indeed, in the rest of this section we will explore a handful of genuine physical effects of the conservative force, which influence the conservative sector of the post-geodesic dynamics in an (essentially) gauge-invariant way. We will also describe invariant quantities constructed directly from the effective metric $\tilde{g}_{\mu\nu}$, which have no direct reference to the GSF. Over the past decade a significant proportion of work in the field has been dedicated to obtaining a quantitative description of these effects in Schwarzschild or Kerr spacetimes. The main motivation for such work is to enable comparisons and synergies with other approaches to the two-body problem, as we shall describe later, in Sec. 8.

Like in the preceding section, we focus here on first-order effects, and we omit the superscript "(1)" from quantities such as $h_{\mu\nu}^{\mathrm{R}(1)}$.

7.1. Frequency of the innermost stable circular orbit

In Sec. 5.1 we mentioned the existence of innermost stable orbits as a distinct characteristic of the strong-field dynamics around black holes in GR. Here we will consider the deformation of that innermost orbit due to the conservative piece of the GSF. To this end, we start with a slightly more detailed discussion of the situation in the geodesic case, restricting first to circular orbits around a Schwarzschild black hole. In this case, the existence and location of an ISCO can be readily inferred from the radial equation of motion (81), where for now we omit the GSF term. In that equation, the effective potential is given by

$$V(r; L_z) = \left(1 - \frac{2M}{r}\right)\left(\frac{L_z^2}{m^2 r^2} + 1\right). \tag{85}$$

Figure 10 shows $V(r)$ for a number of L_z values. It starts off resembling the familiar Keplerian potential at large distances, and like it (for sufficiently large L_z) it reaches a minimum at some radius depending on L_z. But whereas the Keplerian potential goes off to diverge at small separations, the Schwarzschild potential further develops a peak before dropping (to zero) towards the horizon. The minimum of the effective potential marks the location of a circular orbit, where the effective radial force vanishes; it is a *stable* circular orbit since the effective radial force acts as a restoring force in the vicinity of the minimum [note the sign of $\ddot{r}$ in Eq. (81)]. The radius of the stable circular orbit decreases with decreasing angular momentum L_z. In the Keplerian case one can support a stable circular orbit at arbitrarily small separations by selecting a

---

[22] The sense in which Ω is said to be gauge invariant will be clarified further below, in Sec. 7.6. Ω is not actually invariant under arbitrary gauge transformations.

suitable value of the angular momentum,[23] but not so in the Schwarzschild case: here, as L_z decreases, the maximum of $V(r)$ migrates outward to larger radii, until, at some critical value $L_z = L_c$, it merges with the minimum, at which point both extrema disappear. No minimum, or maximum, occur for $L_z < L_c$. The point at which the maximum and minimum merge, and the effective radial force vanishes, is the ISCO.
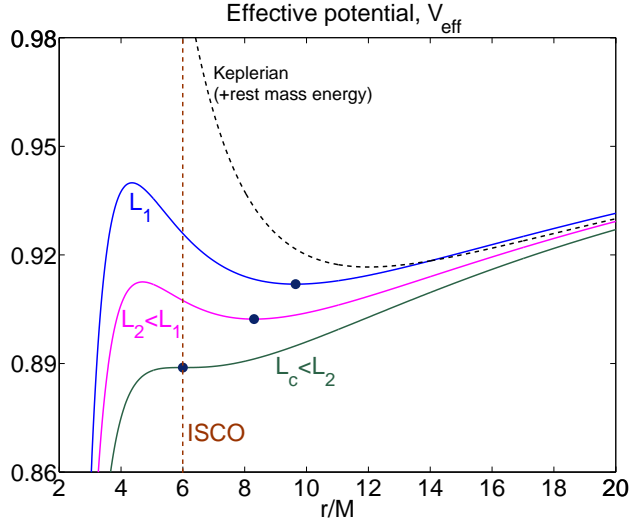

Figure 10: Effective radial potential for geodesic orbits around a Schwarzschild black hole. The radial velocity is governed by $\dot{r}^2 = (E/m)^2 - V(r; L_z)$, with the effective potential $V(r; L_z)$ [Eq. (85)] shown here for three values of the angular momentum L_z. For comparison we also show (dashed line) the familiar Keplerian effective potential (with rest mass energy m added, to match the convention for V in the relativistic case). Stable circular orbits correspond to minima of the effective potential, and no such orbits exist for $L < L_c$.

The effective-potential picture can also be invoked to describe the situation when the orbit evolves under the effect of the full first-order GSF (including dissipation). Again, for simplicity, consider the adiabatic inspiral of (quasi-)circular orbits in a Schwarzschild geometry. As angular momentum is radiated away, the effective potential $V(r; L_z)$ evolves through a sequence similar to that shown in Fig. 10, with the minimum migrating leftward. It has been shown [169] that, in such a situation, orbital energy dissipates at just the right rate so as to keep an initially circular orbit circular. Thus, the orbit evolves through a sequence of circular geodesics, sitting (roughly) at the bottom of the potential well as it gradually shifts leftward in Fig. 10. This adiabatic process proceeds so long as L_z is sufficiently above L_c. As the angular momentum approaches its critical value, the effective radial restoring force weakens, and at some point can no longer counter the object's effective radial inertia inwards. At that point, the gradual inspiral transitions into a direct *plunge* into the black hole, which proceeds approximately along a fixed geodesic, with (almost) fixed E and $L_z \sim L_c$. The details of the transition from adiabatic inspiral to plunge across

[23] In the limit $r \to 0$ one would have to take L_z to zero as $\sim r^{1/2}$, corresponding to a diverging tangential velocity $v \sim r^{-1/2}$.

the ISCO were first worked out (for a Kerr black hole) in Ref. [170] (see also [171]), which, in particular, concluded that the radial thickness of the transition regime scales as $\Delta r \propto \eta^{2/5}$. Thus, for a true inspiral process under the effect of the full GSF, the notion of a well localized ISCO is replaced with that of a "smeared out" transition regime.

However, a precisely localized notion of an ISCO is recovered when considering the conservative dynamics alone, "switching off" the dissipative piece of the GSF in the equations of motion (80)–(83). The remaining, conservative piece of the GSF perturbs the effective restoring force and thereby perturbs the value of r at which it vanishes (i.e., the ISCO's coordinate location). The coordinate radius $r_{\rm isco}$ of the GSF-shifted ISCO is, of course, gauge dependent, but the order-η frequency shift $\Delta\Omega_{\rm isco} := \Omega(r_{\rm isco}) - \Omega_0(6M)$ [where $\Omega(r_0)$ is given in Eq. (84) and $\Omega_0(r_0) = \sqrt{M/r_0^3}$] provides a genuine, physically meaningful measure of the conservative GSF effect. To compute $\Delta\Omega_{\rm isco}$ in practice one has to calculate the contribution of the conservative GSF to the effective restoring force acting on slightly eccentric geodesic orbits, in the limit of vanishing eccentricity, near the ISCO. This calculation was first performed, numerically, for a Schwarzschild black hole, in Ref. [172].[24] The result is

$$\frac{\Delta\Omega_{\rm isco}}{\Omega_0(6M)} = 0.25101546\eta, \tag{86}$$

with an error bar $\pm 5 \cdot 10^{-8}\eta$.[25] This order-η frequency shift is hardly an "observable" effect: after all, for true EMRIs it is typically much smaller than the radial extent of the transition regime, which is proportional to $\eta^{2/5}$. Nonetheless, the value provides an invaluable handle on the strong-field conservative dynamics and a benchmark against which calculations based on other approaches can be tested. In Sec. 8 we will describe the important role that the GSF ISCO-shift result (86) has played in advancing studies of the two-body problem in GR.

So far we have identified the ISCO via the condition of a vanishing effective restoring force. It is natural to ask whether one could instead locate the ISCO by minimizing a suitable conserved energy. Indeed, for circular *geodesic* orbits, there is a well-defined conserved energy, given as a function of radius by $E = m(1 - 2M/r_0)(1 - 3M/r_0)^{-1/2}$, whose minimization correctly gives $r_0 = 6M$. One does not *a priori* expect there to exist a similar notion of conversed energy in the *perturbed* EMRI spacetime, which is lacking a time-translation symmetry. However, in 2011 Le Tiec and collaborators [176] have proposed precisely such an energy function, defined on the conservative sector of the perturbed geometry for circular orbits, whose existence and form came as direct consequences of the so-called "first law of binary black-hole mechanics", which had been formulated at around the same time [177, 178]. Le Tiec *et al.* showed, as an important test, that minimization of their energy recovers the direct GSF result (86), providing a first strong vindication of the first law in the strong-field regime. The same notion of energy, generalized to the Kerr geometry, came out later from a systematic Hamiltonian formulation of the conservative dynamics [166], and in Ref. [179] it was used to calculate the ISCO frequency shift, through

---

[24] A similar calculation, of the ISCO shift in the toy problem of the scalar-field self-force acting on a scalar charge, was performed earlier in Ref. [173].

[25] We quote here the more numerically accurate value obtained later in [174]. Ref. [172] used a definition of frequency based on the Lorenz-gauge time coordinate, which suffers from a certain pathology. The value given here corrects for that pathology, following [175]. (More on that in Sec. 7.6 below.)

order η, for orbits in the equatorial plane of a Kerr black hole (as a function of the black hole's spin). This calculation was more recently confirmed with a direct GSF restoring-force analysis in Ref. [78].

In Sec. 7.5 we shall return to the first law of binary black holes in our discussion of redshift and its relation to binding energy.

7.2. Periastron advance of slightly eccentric orbits

Consider a circular geodesic orbit around a Schwarzschild black hole, with radius $r_0 > 6M$, like the one represented by the black dot sitting at the minimum of the upper curve in Fig. 10. Now consider a small perturbation of that orbit away from circularity (e.g., by endowing it with a small additional amount of energy at a fixed angular momentum). At first order, the radial motion will then exhibit simple harmonic oscillations around the base circular orbit: expanding the right-hand side of Eq. (81) (without the GSF term) about $r = r_0$, recalling $dV/dr = 0$ at $r = r_0$, gives $\ddot{r} \simeq -\omega^2(r - r_0)^2$, with $\omega^2 = \frac{M(r_0 - 6M)}{r_0^3(r_0 - 3M)}$. This frequency corresponds to proper time τ. To convert it to a frequency with respect to the usual coordinate time t—call it Ω_r—we write $\Omega_r = (d\tau/dt)\omega$, where the "redshift" factor is $d\tau/dt = (1 - 3M/r_0)$. Thus we have

$$\Omega_r = \Omega \left(1 - \frac{6M}{r_0}\right)^{1/2}, \tag{87}$$

where $\Omega = (M/r_0^3)^{1/2}$, recall, is the azimuthal frequency of the base circular geodesic. Note how Ω_r vanishes at the ISCO, as expected, and how this radial oscillation frequency is always smaller than Ω: it takes the object longer to complete one radial oscillation than it does to complete one φ rotation. This is the origin of the celebrated effect of relativistic periastron advance. The frequency of periastron precession is $\Omega - \Omega_r$, and over a full radial period, $\Delta t = 2\pi/\Omega_r$, it gains a φ-phase equal to $(\Omega - \Omega_r)\Delta t = 2\pi(\Omega/\Omega_r - 1)$. Thus, recalling Eq. (87), we obtain that the amount of periastron advance per radial cycle is given by

$$\delta = 2\pi \left[\left(1 - \frac{6M}{r_0}\right)^{-1/2} - 1\right]. \tag{88}$$

This is a textbook result, and the reason we have elaborated on its derivation here is that it should now be straightforward to understand how it is modified by the action of the conservative GSF. The GSF modifies (at order η) the effective restoring force near the minimum of the potential, and with it it modifies (at order η) the value of Ω_r for a given circular orbit. At this point we must be mindful of gauge ambiguity, and we therefore adopt the frequency Ω, and not the coordinate radius r_0, as a label of the base circular orbits. We then say that the conservative GSF modifies the *functional relation* between the two invariant frequencies Ω_r and Ω. We express this in the form

$$\Omega_r(\Omega) = \Omega_r^{(0)}(\Omega) + \Delta\Omega_r(\Omega), \tag{89}$$

where $\Omega_r^{(0)}(\Omega)$ is the background geodesic relation, and $\Delta\Omega_r(\Omega)(\sim \eta)$ is the GSF correction to it. The background relation itself is defined by

$$\Omega_r^{(0)}(\Omega) := \Omega \left(1 - \frac{6M}{r_\Omega(\Omega)}\right)^{1/2}, \tag{90}$$

56

obtained from (87) by replacing r_0 with its geodesic expression in terms of Ω, $r_\Omega(\Omega) = (M/\Omega^2)^{1/3}$—but note this replacement is not unique, and one sometimes finds it more natural to define $r_\Omega(\Omega) = [(M+m)/\Omega^2]^{1/3}$ (this recalls the actual relation in the analogue Kepler problem). Note that how we choose to define the background function $\Omega_r^{(0)}(\Omega)$ affects what we mean by the "GSF correction" $\Delta\Omega_r$, and one should be aware of this ambiguity; what remains unambiguous is the *sum* of the two terms on the right-hand side of Eq. (89). Finally, now that we have $\Omega_r(\Omega)$ through order η, the derivation of the periastron advance $\delta(\Omega)$ proceeds just as in the geodesic case, leading to an expression of the form

$$\delta(\Omega) = \delta^{(0)}(\Omega) + \Delta\delta(\Omega), \tag{91}$$

where the background relation $\delta^{(0)}(\Omega)$ is the one obtained from (88) by replacing $r_0 \to r_\Omega(\Omega)$.

A calculation of $\Delta\delta(\Omega)$ requires knowledge of the conservative piece of the GSF on slightly eccentric orbits, in the limit of vanishing eccentricity. This calculation was first performed (numerically) in Ref. [180] for a Schwarzschild black hole, and later for equatorial orbits in Kerr in Ref. [181]. The GSF results were shown to agree nicely with post-Newtonian calculations in the weak-field (large r_0) regime.

The function $\delta(\Omega)$ provides a powerful diagnostic of the conservative dynamics in EMRIs, and it has been utilized in that role on many occasions. The comparison with PN calculations, besides providing much reassurance, also tells us about the convergence properties of the PN expansion (at least through order η) and benchmarks its performance in the strong field. As we shall describe in Sec. 8, it was even used to decide between two initially conflicting recent results for the fourth-order PN equation of motion. Also in Sec. 8 we will describe how $\delta(\Omega)$ has been used to facilitate rewarding 3-way comparisons between GSF, PN and full NR calculations, and to help calibrate the potentials of EOB theory.

7.3. Spin precession and self-torque

Another familiar effect of GR dynamics is the one referred to (somewhat nondescriptively) as the "geodetic effect", or "geodesic precession": the behavior associated with the failure of a spin vector to return to itself after being parallelly transported along a closed curve in curved spacetime. This is also known as the *de Sitter precession*, after W. de Sitter, who, in as early as 1916, predicted that the rotation axis of the Earth–Moon system as it moves around the Sun should experience a precession of ~ 1.9 arcseconds per century due to GR corrections to the Newtonian dynamics [182] (a prediction since confirmed using Lunar laser ranging experiments). A much larger de Sitter effect of ~ 6.6 arcseconds per *year* was measured using a gyroscope in a polar Earth orbit as part of the Gravity Probe B experiment [183]. Larger still is the effect observed in the (only known) double-pulsar system, PSR J0737-3039, where the spin of one of the pulsars appears to be precessing at a rate of ~ 4.8 *degrees* per year [184]. Given the binary's 2.45-hour orbital period, this translates to $\sim 3.7 \times 10^{-6}$ radians of precession angle per radian of orbital revolution—still a rather meagre effect when measured this way.

A much more extreme manifestation of spin precession can be found in the late inspiral of compact-object binaries, including EMRIs, where its rate can reach $O(1)$ radians per radian of orbital revolution. We can quantify this precisely in the case of a

spinning test particle in a circular orbit around a Schwarzschild black hole, in the small-spin limit where the particle can be assumed to follow a geodesic of the background geometry. Then the angle of spin precession per radian of orbital revolution is given in exact form by the simple expression

$$\psi = 1 - \left(1 - \frac{3M}{r_0}\right)^{1/2}, \tag{92}$$

and we see that ψ can be as large as ~ 0.3 (at the ISCO, $r_0 = 6M$) in this scenario: that is some $105°$ rotation of the spin axis over a single orbital period! This primary geodesic effect is perturbed when the particle's finite (nonzero) spin magnitude and/or mass are considered. A finite spin magnitude affects the motion in two ways: by giving rise to a Mathisson-Papapetrou force term in the equation of motion [recall Eq. (44)], and by perturbing the metric of spacetime. Ignoring both these effects (by taking the limit of vanishing spin magnitude), we can ask how the finite mass of the particle modifies the precession rate ψ in the conservative dynamics (i.e., ignoring the dissipative piece of the GSF as elsewhere in this section).

At linear order in m, the answer can be easily found thanks to a result by Harte [63], who showed that, as the particle moves on a geodesic of the effective metric $\tilde{g}_{\alpha\beta}$, its spin vector gets parallelly transported in that same effective metric. Just as geodesic motion in $\tilde{g}_{\alpha\beta}$ corresponds to self-acceleration in $g_{\alpha\beta}$, parallel transport in $\tilde{g}_{\alpha\beta}$ corresponds to a *self-torque* in $g_{\alpha\beta}$: an effective torque exerted on the spinning particle by its own perturbation, which, in particular, modifies the rate of precession ψ away from its geodesic value given in (92). Again using the orbital frequency Ω as an "invariant" label of the circular orbit, we write

$$\psi(\Omega) = \psi^{(0)}(\Omega) + \Delta\psi(\Omega), \tag{93}$$

where the background relation $\psi^{(0)}(\Omega)$ is the one obtained from (92) by replacing $r_0 \to r_\Omega(\Omega)$. An explicit formula for the $O(\eta)$ self-torque correction $\Delta\psi(\Omega)$, which involves $h^{\rm R}_{\alpha\beta}$ and its first derivatives on the particle, was derived in Ref. [185]. This work also performed a first numerical calculation of $\Delta\psi(\Omega)$, showing agreement with the predictions of PN theory at large r_0 (small Ω) and quantifying its gradual breakdown in the strong field. As we describe in the next section, follow-up work extended this comparison to very high order in the PN expansion, and also used the GSF results for $\Delta\psi(\Omega)$ to provide further calibration of the EOB model. In this fashion, the self-torque effect has been utilized as an additional diagnostic of the strong-field dynamics, complimentary to that provided by $\delta(\Omega)$.

In Ref. [186], Akcay *et al.* introduced a generalized notion of ψ applicable to any bound (eccentric) geodesic orbit in Schwarzschild geometry. Here, the self-torque effect is computed over a full radial cycle, and expressed in terms of the two frequencies Ω and Ω_r. Again, a successful comparison was made with equivalent expressions calculated analytically in a PN framework. This programme of comparison was further pursued in [187]. Similar calculations are underway in Kerr [188, 189].

7.4. Self-tides

Spin-precession information, just like the GSF, is encoded in the *first* derivatives of the effective metric $\tilde{g}_{\alpha\beta}$ (evaluated on the orbit). Other invariant quantities of interest may be constructed from higher-order derivatives, as first suggested by Bini and Damour

in [190]. In Ref. [139], Dolan *et al.* considered a set of *tidal* invariants constructed from second derivatives of $\tilde{g}_{\alpha\beta}$. To motivate these quantities, we refer back to the Newtonian scenario of a small body immersed in an external gravitational potential Φ_{ext}. The body's center of mass z^i is subject to an acceleration $a_i = -\partial_i \Phi_{\text{ext}}(z)$. Relative to the center of mass, a mass element inside the body, at $z^i + \delta x^i$, experiences a *tidal* acceleration $a_i^{\text{tidal}} = -[\partial_i \Phi_{\text{ext}}(z + \delta x) - \partial_i \Phi_{\text{ext}}(z)] = -\Phi_{,ij}^{\text{ext}}(z)\delta x^j + O[(\delta x)^2]$. The quantity $\mathcal{E}_{ij} = \Phi_{,ij}^{\text{ext}}(z)$, referred to as the quadrupole tidal moment, characterizes the (leading-order) tidal forces on the body.

In GR, tidal forces are described by the Riemann curvature tensor. Thinking of $\tilde{g}_{\alpha\beta}$ as the external metric, we can construct a relativistic tidal moment $\mathcal{E}_{\alpha\beta} := \tilde{R}_{\alpha\beta\gamma\delta}\tilde{u}^\beta \tilde{u}^\delta$, where $\tilde{R}_{\mu\nu\beta\delta}$ is the Riemann tensor associated with $\tilde{g}_{\alpha\beta}$, and $\tilde{u}^\alpha$ is the particle's four-velocity normalized in $\tilde{g}_{\alpha\beta}$. Unlike in Newtonian gravity, in GR there is an additional tidal moment, $\mathcal{B}_{\alpha\beta} := \frac{1}{2}\epsilon_{\alpha\gamma}{}^{\mu\nu}\tilde{R}_{\mu\nu\beta\delta}\tilde{u}^\gamma \tilde{u}^\delta$, where $\epsilon_{\alpha\beta\mu\nu}$ is the fully antisymmetric Levi-Civita tensor. While $\mathcal{E}_{\alpha\beta}$ characterizes stretching, $\mathcal{B}_{\alpha\beta}$ characterizes twisting; they are sometimes referred to as electric- and magnetic-type tidal moments, respectively, or as the tidal field and the "frame-drag" field. Focusing on circular orbits in the equatorial plane of a Kerr black hole, Dolan *et al.* showed that the eigenvalues of these tensors are invariant quantities. They identified four nontrivial such eigenvalues, three associated with $\mathcal{E}_{\alpha\beta}$ and one with $\mathcal{B}_{\alpha\beta}$. Letting $\lambda \in \{\lambda_i^E, \lambda^B\}$ ($i = 1, 2, 3$), they then split each λ into "background" and "particle" contributions in the usual way,

$$\lambda(\Omega) = \lambda^{(0)}(\Omega) + \Delta\lambda(\Omega), \tag{94}$$

giving explicit expressions for the "self-tides" contribution $\Delta\lambda(\Omega)$ in terms of $h_{\alpha\beta}^{\text{R}}$ and its first and second derivatives on the particle. Ref. [139] then performed a numerical calculation of $\Delta\lambda(\Omega)$ in the case of a Schwarzschild background, and demonstrated how the results agree at small Ω with the analytical predictions of an independent post-Newtonian calculation. The latter was truly independent in that it had no reference to a point particle or self-force: it was based on [191] and [192], where the approximate form of the metric near a small black hole immersed in a weakly-curved tidal environment was derived. The agreement is thus nontrivial and highly informative.

However, there is good reason to question whether λ genuinely describes the physical tidal environment of the small object. After all, $\tilde{g}_{\alpha\beta}$ is *not* the true metric; it is only an effective one. We know that the particle follows a geodesic in this effective metric and that its spin gets parallelly transported in it. But what is the physical significance of the tidal field associated with it? Given the degree of ambiguity involved in the very definition of the field $h_{\alpha\beta}^{\text{R}}$ (recall our discussion in Sec. 3.3), should we expect it to have a clear physical meaning at all? The agreement with the post-Newtonian analysis suggests that, in fact, it has. Additional light was shed on this question in Ref. [57] using a systematic matched-asymptotic-expansions analysis. This work demonstrated explicitly that $\mathcal{E}_{\alpha\beta}$ and $\mathcal{B}_{\alpha\beta}$ are *not* equal to at least one particular pair of tidal moments that have been used to describe the local spacetime geometry near a tidally perturbed small object. But it also showed that due to an inherent ambiguity in the definition of tidal moments in GR, one *can* write the metric of the local spacetime in such a way as to interpret $\mathcal{E}_{\alpha\beta}$ and $\mathcal{B}_{\alpha\beta}$ as the physical tidal moments. Hence, while there is some danger of over-interpreting $h_{\alpha\beta}^{\text{R}}$, the tidal moments constructed from it *can* be thought of as one *particular* characterization of the object's tidal environment.

In [193], Nolan *et al.* have advanced this programme one step further by defining and calculating *octupolar* tidal invariants associated with $\tilde{g}_{\alpha\beta}$, which involve *third* derivatives of the perturbation $h^{\mathrm{R}}_{\alpha\beta}$. Here too, focusing on circular orbits in Schwarzschild, they were able to obtain accurate numerical results and demonstrate their consistency with analytical PN calculations.

7.5. Redshift and binding energy

We have kept for last a conservative GSF effect that, in fact, had been the first to be identified and analyzed (by Detweiler in 2008), the first to enable a direct comparison between GSF and PN results (again by Detweiler), and one whose fundamental role in perturbation theory has since been gradually revealed through the work of many. If we have kept it for last it is only because its physical interpretation is somewhat more subtle. In [66], Detweiler introduced the *redshift variable $z(\Omega)$* as a diagnostic of the conservative dynamics of circular orbits in Schwarzschild geometry.[26] Now often referred to as "Detweiler's redshift", it is defined as follows. Consider the particle's four-velocity $\tilde{u}^{\alpha}$, as normalized in the effective metric $\tilde{g}_{\alpha\beta}$ [i.e., $\tilde{g}_{\alpha\beta}\tilde{u}^{\alpha}\tilde{u}^{\beta} = -1$ through $O(\eta)$]. Detweiler's redshift is then

$$z := (\tilde{u}^t)^{-1} = \frac{d\tilde{\tau}}{dt}, \tag{95}$$

where, like in Sec. 3.5, $\tilde{\tau}$ is proper time as measured in $\tilde{g}_{\mu\nu}$. Labelling the circular orbit by its frequency Ω, as usual, we then extract the $O(\eta)$ effect by writing

$$z(\Omega) = z^{(0)}(\Omega) + \Delta z(\Omega). \tag{96}$$

The value of z on the zeroth-order geodesic in the Schwarzschild background is $z^{(0)}(\Omega) = 1 - \frac{3M}{r_\Omega(\Omega)}$. Detweiler showed that $\Delta z(\Omega)$ can be extracted from the value of the field $h^{\mathrm{R}}_{\alpha\beta}$ on the particle alone (the derivatives of $h^{\mathrm{R}}_{\alpha\beta}$ are not needed), and that z depends on $h^{\mathrm{R}}_{\alpha\beta}$ only through the simple scalar contraction $h^{\mathrm{R}}_{\alpha\beta}u^{\alpha}u^{\beta}$. He also showed that $\Delta z(\Omega)$ is an invariant quantity, in a certain useful sense (to be clarified in the next subsection).

What is the physical interpretation of z? It is tempting to suggest that z measures gravitational redshift: the ratio between proper-time lapse along the orbit and proper-time lapse of inertial observers far away. However, we must recall z is calculated in the effective geometry $g_{\alpha\beta} + h^{\mathrm{R}}_{\alpha\beta}$, not in the true, physical geometry (the actual redshift in the latter is predominantly due to the singular piece of the physical metric perturbation, which diverges on the particle). In the case that the small object is a black hole, a more strictly physical interpretation has emerged in recent years: z is (up to a proportionality factor) equal to the *surface gravity*, κ, of the small black-hole's horizon. This relationship was first surmised based on the fact that the two quantities play analogous roles in the first law of binary black hole mechanics (discussed further below) [177]. Using a matched-asymptotic-expansions analysis, one of us [194] later showed that this relationship is *exact* through linear order in the mass ratio. The surface-gravity interpretation was impressively confirmed by Zimmerman *et al.* even in fully nonlinear GR, using full NR simulations of black-hole binaries [195].[27] Thus,

---

[26] To be pedantic, we note that Ref. [66] actually considered the inverse redshift, $z^{-1} = \tilde{u}^t$. The use of z became prevalent in more recent literature.

[27] More precisely stated, the surface-gravity interpretation of z was confirmed in Ref. [195] assuming the validity of the first law. Conversely, if one assumes that interpretation, the results of [195] can be considered a test of the first law.

Δz describes the $O(\eta)$ contribution to the surface gravity of the small hole's horizon due to back-reaction from its own field.

The redshift variable $z(\Omega)$ has played a central role in the development of both GSF and PN methods over the past decade. Soon after its introduction, it was used as a benchmark in first comparisons of numerical GSF results obtained in different coordinate gauges [196, 105], in facilitating a first contact with PN theory [66], and in benchmarking first results in Kerr [106]. It later enabled high-precision comparisons with PN theory and the accurate calibration of the EOB potentials, both activities to be described in Sec. 8 below. Because $z(\Omega)$ requires only the perturbation field and not its derivatives, it is relatively easy to extract from numerical solutions of the field equations, and thus provides a first testing point in developing new GSF codes.

But besides possessing practical value, z also turned out to be of some fundamental significance. In 2011, Le Tiec and collaborators [177] formulated the "first law of binary mechanics": a conjectured[28] variational formula, which, remarkably, when applied to the circular-orbit EMRI setup, relates the global properties of the system—binding energy and total angular momentum—to the local redshift z as measured on the particle. Using this formula it is possible to write down explicit expressions for the binary's binding energy and angular momentum through $O(\eta^2)$ (i.e., including the leading-order self-gravity term) in terms of $\Delta z(\Omega)$ alone [176]. This is remarkable, because a similar calculation of the global energy and angular momentum using systematic perturbation theory alone would require knowledge of the second-order metric perturbation. The first law thus appears to provide a strikingly "easy" access to second-order information! At the practical level, one then gains access to an additional, important gauge-invariant property of the binary, namely its binding energy, with an ability to compute the GSF correction to it without having to go to second order. This, indeed, has been exploited in many synergistic studies with PN theory and EOB. Knowledge of the binding energy also allowed, as we have seen before, an "easy" route to the determination of the ISCO frequency shift in the Kerr case [179] (through minimization of that energy—a calculation later confirmed with a direct GSF calculation). The relation between Detweiler's redshift and binding energy manifests itself clearly in the recently developed Hamiltonian formulation of EMRI dynamics [166], in which the two-body interaction piece of the Hamiltonian is expressed as a simple function of Δz.

The original (Detweiler's) definition of z as an invariant variable relies crucially on the existence of helical symmetry in the problem, as in the circular-orbit setup (ignoring dissipation). Later on, in [199], a generalized notion of redshift was introduced, applicable to any bound periodic orbit in Schwarzschild. The Barack–Sago generalized redshift is given by

$$z(\Omega, \Omega_r) = \langle \tilde{u}^t \rangle^{-1} = \frac{\tilde{\tau} \text{ period}}{t \text{ period}}, \tag{97}$$

where $\langle \cdot \rangle$ denotes an average with respect to proper time over a complete radial period of the motion, and on the right-hand side we have, one finds, simply the ratio between the radial period measured in proper time $\tilde{\tau}$ and the radial period measured in t. This, it was shown, is invariant within a class of gauges consistent with the periodicity of the setup (i.e., ones in which the metric perturbation has a bi-periodic spectrum

---

[28] The validity of the first law was rigorously established in a PN context and for a range of certain circumstances—cf. [197, 198].

with fundamental frequencies $\{\Omega, \Omega_r\}$). These two frequencies are also used as an invariant label of the bound geodesics,[29] and in splitting z into "background" and "self-force" contributions in the usual way. Calculations of the generalized $\Delta z(\Omega, \Omega_r)$, with detailed comparisons to PN predictions, were carried out in, e.g., [200, 201] for Schwarzschild and [109, 120] for Kerr. The introduction of a generalized redshift has motivated the formulation of a similarly generalized first-law of binary mechanics—by Le Tiec in 2015 [197]—which divorced itself from its own reliance on helical symmetry. This, in turn, provided an important link between $\Delta z(\Omega, \Omega_r)$ and some of the EOB potential terms associated with radial momentum, giving a new handle on the latter [202, 203, 120].

7.6. Comment on invariants and quasi-invariants

In closing this section, it is appropriate that we make more precise the notion of gauge invariance implied in our discussion. We have announced that the frequency Ω was a gauge-invariant parametrization of circular orbits, and that $\Delta\delta(\Omega)$, $\Delta\phi(\Omega)$, $\Delta\lambda(\Omega)$ and $\Delta z(\Omega)$ were gauge-invariant functional relations. But in truth, none of these quantities are gauge invariant in the usual mathematical sense. In perturbation theory, a linearized quantity (e.g., a field) is said to be gauge invariant if its value is unaffected by the action of any linear diffeomorphism of the form $x^\alpha \to x^\alpha - \xi^\alpha$, where ξ^α is a small gauge displacement (of order η, in our case). Examples are the components of the Riemann tensor linearized about flat spacetime, or the components of the Ricci tensor linearized about any vacuum spacetime.[30] But consider, for example, the frequency $\Omega = d\varphi/dt$ of a circular orbit. It is easy to find gauge displacements ξ^α that *do* change Ω: take, for instance, $\xi^\alpha = \varpi\, t\, \delta_\varphi^\alpha$ (where ϖ is some order-η constant), under which $\varphi \to \varphi - \varpi t$ and thus $\Omega \to \Omega - \varpi$. This gauge transformation, which is perfectly legitimate mathematically, has clearly modified the "gauge-invariant" frequency. What happened here, of course, is that ξ^α generated rotation with respect to the original system. It is not surprising to find that the frequency defined with respect to the rotating system differs from that defined in the original system.

In light of this example, we must refine our language. When we say Ω is "gauge invariant", what we really mean is that it is invariant under a certain, suitably restricted class of "physically reasonable" gauge transformations, which must exclude, in our case, ones that artificially generate rotation. But our example points to a true practical dilemma often encountered in GSF calculations. In perturbation theory gauge conditions are rarely given as explicit coordinate conditions and more typically formulated in terms of conditions on the metric perturbation itself. Given a metric perturbation in a particular gauge, how do we tell if the gauge is "physically reasonable"? Or, equivalently, given two perturbations that, we are told, are related via a "nonphysical" gauge transformation (as in the above example), how can we tell which of the two gauges (if any) is the "good" one?

---

[29] As mentioned in an earlier footnote, the two frequencies are not actually good global parameters, due to the isofrequency phenomena studied in [122]. However, one can still use them as parameters with due caution in local regions of the parameter space.

[30] For a Kerr background, Aksteiner and Bäckdahl [204] have identified a set of 18 independent gauge-invariant linearized perturbations (including the 10 components of the Ricci perturbation), defined from linear combinations of second and third derivatives of the metric perturbation, which constitutes a minimal "generating" set in the sense that all other geometrical invariants can be derived from it by taking derivatives and linear combinations.

These are important questions, because they impact on our ability to assign a proper physical interpretation to our GSF results and, as a result, on our ability to compare our results with those obtained in other gauges or using other methods. To address these questions, we must accompany the definition of each of our "invariant" quantities [Ω, $\Delta z(\Omega)$, etc.] with a precise statement about the class of "physically reasonable" gauges within which it remains invariant. That statement should be in the form of conditions on the metric perturbation, and these conditions must refer to some genuine invariant attributes of the spacetime in question. In the case of circular orbits, one such attribute is *helical symmetry*, and another is *asymptotic flatness*. One demands that the metric perturbation respects the helical symmetry, meaning $(\partial_t + \Omega\partial_\varphi)h_{\alpha\beta} = 0$, and that it has the right fall-off at infinity. (In our above example of a "rotating" gauge it would be the form of the metric perturbation at infinity that would identify it as a "noninertial" one and thus exclude it from consideration.) Alternative conditions must be imposed in setups that are not helically symmetric, such as eccentric-orbit configurations or inclined orbits in Kerr. These can rely on an "averaged" notion of symmetry or on the periodic properties of the problem, in combination with asymptotic flatness [199, 78]. The formulation of conditions to define a class of suitable gauges is a subtle matter, which depends on the invariant quantity being calculated and on what the calculation aims to achieve. For example, when the goal is to enable comparison with a PN calculation, one must ensure that both calculations are done within the same class of gauges.

In practice, however, one often finds in GSF calculations that the gauges most convenient to work with (e.g., the Lorenz or radiation gauges) fall outside the class of gauges deemed suitable for a particular calculation. One then has to introduce, *a posteriori*, a correction to the quantity being calculated, which accounts for the effect of a gauge transformation from the working gauge (say, Lorenz or radiation) onto the class of physically suitable gauges. As an example, consider the calculation of the ISCO frequency shift described above, in Sec. 7.1. The Lorenz-gauge metric perturbation used for that calculation in [172] turns out to suffer from a minor gauge pathology: the monopole part of the perturbation does not decay at infinity but rather approaches a constant value. This is easily cured with a simple gauge transformation (a small shift of the time coordinate) away from Lorenz—but that gauge transformation changes the value of $\Delta\Omega_{\mathrm{isco}}$. The value quoted in Eq. (86) is the gauge-corrected one, and it is that value one can meaningfully compare to, say, PN results. Similar gauge corrections were necessary, and have been introduced, in other calculations of "invariant" quantities, such as in the radiation-gauge calculation of the periastron advance in Kerr [181].

In summary, the various physical quantities whose GSF corrections we have considered in this section—Ω_{isco}, δ, ψ, etc.—are not gauge invariant in the strict mathematical sense, but we may refer to them as *quasi-invariants*. They are invariant within a certain class of gauges that must be defined *ad hoc*, depending on the purpose of our calculation. The process of identifying a suitable class of gauges, and of performing any necessary gauge adjustments, is an important prerequisite in the programme of comparison and synergy with other methods. We turn to discuss this programme next.

8. Comparisons and synergies

As we have seen, comparison of results from GSF calculations with the analytical predictions of the PN theory provides a powerful overall check of both the numerical

calculation itself and the GSF theory underpinning it. In fact, this is a mutual test of both GSF and PN methods. Each of these frameworks relies on subtle regularization procedures and involves a chain of subtle computational manipulations. Thus, finding an agreement on the "final value" of a concrete physical quantity is greatly nontrivial and highly gratifying.

But such comparisons can achieve more than just mutual tests. The GSF and PN methods are both systematic approximation approaches to the two-body problem in GR, each based on a perturbative expansion in a different limiting domain of the problem. The PN method expands about the limit of large separation (weak, Newtonian gravity), keeping the mass ratio η arbitrary, while the GSF method expands about the limit $\eta \to 0$ (geodesic motion), keeping the separation arbitrary. Combining these two complementary approximation methods can be useful in at least three different ways. First, using GSF results as a benchmark, one can directly study the convergence properties of the PN expansion in the strong field, at least through $O(\eta)$. Similarly, analysis of how PN terms depend on η may provide hints about the convergence properties of the GSF expansion. Second, one can use GSF information to determine high-order terms in the PN expansion that are hard to obtain with a direct PN calculation (hence pushing the validity of the PN expansion further into the strong-field regime). Similarly, one can use PN information to "predict" certain high-order GSF terms. Finally, one can hope to inform a model of the two-body problem in an intermediate domain that is perhaps not accessible to either of the approximation methods separately.

This synergistic programme can be expanded to include NR, which directly solves the fully nonlinear Einstein field equations that describe the inspiral and merger dynamics in black-hole binaries. The idea of such a three-fold synergy is illustrated in Fig. 11, showing the respective applicability domain of each of the three methods, NR, PN and GSF, in the essential two-parameter space of the two-body problem: mass ratio vs. separation (or strength of gravitational interaction). NR simulations can model the two-body dynamics of two comparable-mass black holes close to their merger, but become computationally prohibitive when the separation is large or the mass ratio too small ($\lesssim 1:10$, in practice). In each of these situations, a "small parameter" presents itself in the problem, making it amenable to a perturbative description: PN for large separations, and GSF for small mass ratios. The synergistic studies to be reviewed in the rest of this section seek to explore the interfaces between the three approaches and reach across them. This is often done in the framework of the EOB approach, which aims to provide a universal, semi-analytical model of the two-body dynamics across its entire parameter space; this, too, will be reviewed below.

For a fuller review of work on the overlap between the various approaches to the binary black-hole problem, see [205].

8.1. Interface with post-Newtonian theory

As mentioned in Sec. 7, a first direct contact with PN theory was made by Detweiler in a 2008 paper [66], utilizing the relation $z(\Omega)$ for circular orbits in Schwarzschild as a benchmark for the comparison. Using large-r_0 fits to increasingly more accurate numerical GSF results as they became available, later calculations have extracted the $O(\eta)$ piece of $z(\Omega)$ to a very high PN order [206, 207, 111, 208, 112]. Such high-order PN terms of $z(\Omega)$ have also been obtained via a direct analytical calculation
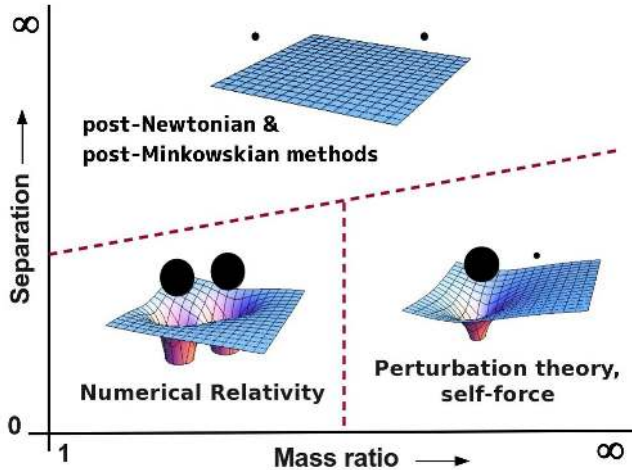
Figure 11: Domains of the two-body problem in GR. The dashed lines schematically delineate the applicability domains of the three main approximation methods. Cartoons illustrate the principle behind each approach: PN expands about flat space, GSF expands about the exact fixed geometry of a central black hole, and NR tackles the full nonlinear dynamics. Synergistic work seeks to interface between results from the three types of calculations.

[115, 117], using the semi-analytical approach based on the Mano–Suzuki–Takasugi method, which we have mentioned in Sec. 4.3. Conversely, $O(\eta^2)$ terms of $z(\Omega)$ have been used to "predict" certain components of the second-order GSF, which are yet to be directly calculated [209]. There is a large body of work on the PN expansion of other physical GSF quantities—see, e.g., [118, 120, 187, 210] and references therein.

The ISCO-shift calculation described in Sec. 7 typifies the way in which a GSF result can serve as a strong-field benchmark against which the performance of the PN approximation may be assessed. Long before the advent of GSF methods, the value of the ISCO shift had been providing a comparison point for various resummation techniques that have been suggested over the years for improving the convergence of the PN expansion in the strong field (see, e.g., [211, 212]). The GSF calculation of the ISCO shift in 2009 supplied a much-sought-for "exact" value that could reliably, for the first time, discriminate between the various PN models (and between the various ways of usefully defining the ISCO within a PN framework). Work by Favata [213] examined the performance of about two dozen PN schemes against the GSF value of the ISCO shift, ruling most of them out.

Another opportunity for GSF calculations to inform the development of PN theory was provided more recently, with the direct derivation of the fourth-PN equation of motion, independently (and using different methods) by Damour *et al.* [214, 215] and by Bernard *et al.* [216]. Both derivations used GSF results to determine the values of certain parameters associated with ambiguities in their respective regularization methods. Furthermore, accurate GSF results for the periastron precession [181] were later utilized to resolve an initial discrepancy between the results of the two derivations [217, 218]. (A complete, ambiguity-free, fully PN derivation has very recently been performed, arriving at the same result as the earlier, GSF-informed derivations [219].)

8.2. Interface with Numerical Relativity

The regime of the binary black-hole problem where the two masses are comparable and strongly interacting can only be tackled reliably using NR tools. This is the regime most relevant to LIGO's black-hole merger sources, and indeed, the analysis of signals from these sources has been relying on theoretical waveforms produced using NR methods. It is desirable to extend the reach of NR methods to systems of more disparate masses: this is important for future LIGO observations [4], and crucial for potential detections of intermediate-mass-ratio inspirals (IMRIs) with space-based or third-generation ground-based detectors [220]. The problem is that NR simulations become fundamentally intractable for masses that are too disparate, because of the extra computational burden that comes with the need to resolve the two lengthscales and to track the evolution over a larger number of orbital cycles. This motivates a synergistic approach that meshes together results from NR and GSF calculations. The ultimate goal is to construct a model that is valid across all mass ratios, but the synergy can pay significant dividends along the way, as illustrated by work already done, which we now review.

The potential of a GSF-NR synergy was first demonstrated in Ref. [221]. That work exploited two then-recent developments: on one hand, the calculation of a concrete physical effect of the GSF—its contribution to the periastron advance of orbits around a Schwarzschild black hole [93, 180]; and, on the other hand, the development of a highly accurate spectral NR code, enabling, for the first time, the extraction of local orbital phase information [222]. In Ref. [221], the periastron advance $\delta(\Omega)$ for a range of mass ratios between $\eta = 1\!:\!1$ and $\eta = 1\!:\!8$ was calculated in NR and compared with the "predictions" of GSF and PN calculations (as well as with those of the EOB model; see below). The analysis allowed, for the first time, researchers to quantitatively assess the performance of the first-order GSF approximation in the comparable-masses regime, using the NR result as a benchmark. It nicely demonstrated how that approximation improves with decreasing η, and gave a first handle on the magnitude of second-order (and higher) GSF effects. Most remarkably, it illustrated how the first-order GSF approximation can prove applicable far beyond its natural $\eta \ll 1$ domain: with the simple replacement of η with the symmetric mass ratio $mM/(M + m)^2$ $[= \eta + O(\eta^2)]$, it was shown that the first-order GSF reproduces the NR result for $\delta(\Omega)$ with a good accuracy for *any* mass ratio—even for $\eta = 1$! This curious result has since been demonstrated also with physical quantities other than $\delta(\Omega)$, but it is yet to be explained on theoretical grounds.

In Ref. [223], Le Tiec and collaborators used NR results for spinning black holes to predict (in effect, "calculate") the GSF correction to $\delta(\Omega)$ for orbits around a Kerr black hole, before direct GSF calculations in Kerr were available. Their results were confirmed (and improved upon) only very recently, with the direct GSF calculation by van de Meent [181]. Further work (mentioned above), by Zimmerman and collaborators [195], has shown how the redshift function $z(\Omega)$, including GSF corrections, can be extracted from full NR simulations. This programme has motivated the development of precision tools for extraction of orbital frequencies from NR simulations [224]. Here, comparison with existing GSF results provides a benchmark for testing the extraction procedure. Conversely, NR results for quantities for which direct GSF results are not yet available, such as the orbital-plane precession, may in the future provide a benchmark for testing the eventual GSF results. Ref. [224] also studied the impact of resonance transitions in NR evolutions. No clear effect has been

observed for the mass ratios studied ($\eta \geq 1:7$), but future work with smaller η may allow for an interesting comparison with corresponding GSF calculations.

The GSF-NR synergy has a large potential scope that is only now beginning to be realized. Besides testing each other, GSF and NR techniques may be directly combined to improve the performance of either. GSF codes can greatly benefit from the incorporation of advanced numerical methods that are used routinely in NR (like mesh refinement, or the use of highly accurate spectral methods for solving PDEs). On the other hand, GSF information may be directly incorporated in NR codes in order to ease the computational burden when the mass ratio is small. According to one such proposal [225], one could "excise" a region around the smaller black holes out of the computational domain, using GSF information to set boundary conditions on the excision boundary. It is hoped that such, or similar ideas could enable one to tackle the IMRI regime of the two-body problem, which remains seemingly inaccessible to any one of the individual approaches on its own.

8.3. Calibration of the Effective-One-Body model

The EOB formalism [226, 171, 227] (see [228] for a review) is an analytical framework that aims to provide a universal description of the dynamics of coalescing binaries of black holes and other compact objects, valid for any mass ratio and over the entire merger process. To achieve this, the model seeks to generalize the effective-one-body reduction of the Kepler problem: just as the Newtonian two-body problem can be reduced to the motion of an effective body in a central potential, the EOB model posits that the two-body problem in GR (at least in its conservative sector) can be reduced to geodesic motion of an effective body in a deformed Schwarzschild geometry. By design, the model automatically reproduces the true dynamics in the test-particle ($\eta \to 0$) limit and to all known PN orders. Dissipative terms and spin effects are incorporated into the model in a modular way, based on resummed PN expressions. Beyond that, the analytical EOB model involves certain *a priori* unspecified "potentials", and the idea is that these EOB potentials may then be "calibrated" using whatever information becomes available from other, more systematic calculations of the two-body dynamics. There is an ongoing, intensive programme to calibrate the EOB potentials using numerical data from NR simulations; see [229, 230, 231] and references therein. This activity has proven hugely successful, and NR-calibrated EOB waveform templates now play a central role in LIGO data analysis. The EOB idea has also been applied to tidally interacting neutron-star systems (e.g., [232, 233]) and to the problem of hyperbolic scattering of black holes [234, 210].

In a 2010 paper [175], Damour first highlighted the potential of (the then-just-emerging) GSF results to provide a new, powerful type of calibration for the EOB model. Information from GSF calculations is highly numerically accurate, conveniently comes split into conservative and dissipative effects, and most importantly, gives a handle on the small-mass-ratio portion of the problem, which is inaccessible to NR. Damour demonstrated in [175] how then-available GSF results usefully constrain the $O(\eta)$ piece of the EOB potentials, and proposed a list of physical GSF quantities whose transcription into EOB language could provide further calibration.

Thus began a fruitful exchange between the EOB and GSF programmes. The initial transcription between the GSF and EOB languages was done using the periastron advance $\delta(\Omega)$ [175, 180] as a physically meaningful invariant quantity. Later

work utilized Detweiler's redshift for that purpose, for circular orbits in Schwarzschild [174, 113, 114, 115] and Kerr [235], then for eccentric orbits in Schwarzschild [203, 201], and finally for eccentric orbits in Kerr [120]. Further work utilized the self-torque on circular [190] or eccentric [187] orbits, and also the self-tides [236]. Ref. [209] explored the prospect of calibrating the EOB potentials at $O(\eta^2)$ using a future GSF calculation of the redshift through second order. The calibration of the dissipative sector of the EOB model has also been improved, in the EMRI domain, using perturbative energy-flux calculations [237, 238, 239, 240].

There are many prospects for future GSF calculations to further inform the EOB model (and the implications to EOB of some recent GSF results, such as the ISCO shift in Kerr [179], have yet to be considered). Referring back to Fig. 11, it can be said that the EOB approach provides a convenient framework for fusing together results from the very distinct methods of NR, PN and GSF, with the aim of covering the entire plane shown in the figure. GSF information is particularly valuable, as it gives a handle on a "remote", hitherto inaccessible corner of the parameter space. The hope is that such an EOB-facilitated universal model could one day provide an accurate, reliable, and computationally efficient description of all binary sources relevant to gravitational-wave astronomy, from LIGO's merging comparable-mass black holes and neutron stars to LISA's EMRIs, and everything in between.

9. Frontiers

The self-force programme provides a fine example of experiment-driven progress in theoretical physics. The renewed interest in this old problem of classical relativity, the volume of activity over the past two decades, and the significant progress that has been achieved, all no doubt owe themselves to the exciting prospects of observing gravitational waves from EMRIs in the not-so-distant future. It remains a central goal of the programme to provide an accurate theoretical model of astrophysical EMRIs to enable the identification and interpretation of EMRI signals in the datastream of LISA. There is still much to be done to achieve that goal, and in this concluding section we will review the main outstanding issues, both foundational and computational.

9.1. Open problems and prospects: foundational issues

The foundations of self-force theory are by now fairly well developed. Given an arbitrarily structured, small, compact object, there is a concrete algorithm, involving matched asymptotic expansions, for obtaining the metric (up to the smooth vacuum perturbation $h^{\mathrm{R}}_{\mu\nu}$) in the object's local neighbourhood to any order in perturbation theory. This local metric can be converted into a singular puncture in the spacetime, effectively reducing the object to a point-particle-like structure characterized by the object's multipole moments. The position of the puncture then acts as a representation of the "position" of the object, and the equation of motion governing it follows from two things: the vacuum Einstein field equations in the puncture's local neighbourhood, and a condition that identifies it with the object's center of mass. Once the puncture and its equation of motion are in hand, the physical field everywhere outside the object, an effective field inside it, and its mean motion can all be computed (through second perturbative order) from the coupled set of equations (41), (42), and (46). At first perturbative order, this description can be reduced to that of a point particle interacting with the Detweiler-Whiting free radiation field $h^{\mathrm{R}(1)}_{\mu\nu}$. At both first and

68

second order, the physical picture that emerges is that of a pointlike object that moves as a test body in an effective external spacetime.

However, work remains to be done even at this general level. As stressed in Sec. 5.2, for optimal EMRI signal analysis we require the equation of motion through second perturbative order. Although it has been derived explicitly through that order, the derivations have so far been restricted to the special case of a nonspinning, spherical object. For a generic object, the second-order equation of motion will involve the object's spin and quadrupole moments. At least hypothetically, those forces can be extracted from the fully nonlinear results of Harte [25]. However, those results are, at least in principle, not applicable to black holes. One would also need to be sure of how Harte's multipole moments (defined by integration over a material stress-energy tensor) relate to those that would appear in the metric outside the body. Hence, a first-principles derivation based on matched asymptotic expansions would be preferable. We can expect two new subtleties to arise in such a derivation. First, there should be multiple possible "good" definitions of center of mass, corresponding to the multiple "spin supplementary" conditions studied in PN, test-body, and fully nonlinear contexts [241]. Second, the evolution of the object's quadrupole moments cannot be determined from the vacuum EFE alone; instead, it will require knowledge of the object's internal composition. If the object is a black hole, then the no-hair theorem dictates that at leading order its quadrupole moments are those of the Kerr metric, uniquely determined by its mass and spin. But if it is a star, the quadrupole moments must be determined from the star's equation of state.

Beyond this extension to nonspherical, spinning objects, there are several other foundational issues to explore. For example, the second-order punctures in Refs. [58] and [54], which are in different gauges, should be directly compared. The class of "highly regular gauges" in Ref. [57] should also be examined; this is a class of gauges in which the singularity in the second-order Einstein tensor $\delta^2 G_{\mu\nu}$ is significantly ameliorated. Counter to the general discussion in Sec. 3.1, in this special class one *can* define a distributional stress-energy $T^{(2)}_{\mu\nu}$ and a distributionally well-defined field equation of the form (23). With a well-defined distributional source, one could then define useful quasilocal, Green's-function-based definitions of the second-order singular and regular fields, analogous to the Detweiler-Whiting definitions at first order. This would complement the purely local definitions described in Sec. 3.3. It would also potentially enable an alternative path to calculations of second-order effects [242]. However, $T^{(2)}_{\mu\nu}$ itself has not yet been derived, and many practical details of numerical schemes in a highly regular gauge have yet to be worked out.

Another area that should be explored is the effective-field-theory approach to self-force theory [243]. This approach, based on a presumed point-particle effective action together with some regularization prescription, has only been applied at first order. But if extended to second order, it would provide an alternative path to a Green's-function formulation at second order [244, 245]. Given that effective-field-theory calculations are generally quite efficient, it may also enable simple derivations of finite-size effects and even of the third-order self-force [244, 245].

There are also more abstruse open problems. Self-force theory could be extended to the case of a photon, for example, perhaps building on work on the ultrarelativistic limit [246]. And although there have been several studies in the case of an object that is both massive and charged [247, 248, 249], and of objects in alternative theories of gravity [250], there is considerable room to develop these cases from first principles.

Extensions to higher dimensions could also be pursued [251].

But the greatest remaining challenges do not lie in purely foundational issues. Instead they sit at the level of practical formulations in the specific case of an EMRI. We enumerate some of those challenges below.

9.2. Open problems and prospects: computational issues

There are three main (interrelated) open problems facing concrete calculations of EMRI spacetimes: improving upon and extending the reach of current first-order GSF calculations; developing and implementing an evolution scheme that remains accurate over a complete inspiral; and finding practical ways of calculating second-order GSF effects.

At first order, as we have seen, state-of-the-art numerical codes are now capable of outputting the value of the first-order GSF along (essentially) any given, fixed, bound geodesic orbit around a Kerr black hole. The most numerically efficient approach is based on a semi-analytical treatment of the frequency-domain perturbation equations in the Teukolsky formalism, followed by metric reconstruction and mode-sum regularization in the radiation gauge [78]. But even this method remains, in general, very computationally expensive, and suffers from several weaknesses. In particular, it does not perform well at large orbital eccentricities—not only because of the slow convergence of the frequency-mode sum, but also because of what appears to be delicate cancellations between different mode contributions [78]. Continuous-spectrum methods could help tackle high-eccentricity scenarios, as well as unbound (parabolic- or hyperbolic-type) orbits; preliminary work by Hopper and collaborators [252, 253] nicely illustrates the applicability of such an approach.

However, the problems of highly eccentric and unbound orbits are perhaps more naturally tackled using a time-domain approach. Time-domain methods and working codes have greatly improved over the past decade, but many challenges remain. In the most direct time-domain approach, based on the Lorenz-gauge formulation, there remains an unsolved problem of linearly growing gauge modes that contaminate numerical evolutions [87]. And there remains a problem of slowly damped junk radiation due to the imprecise determination of (the unknown) initial conditions for the evolution. There has been much recent work to improve the performance of time-domain codes, with a variety of techniques being introduced. These included the use of adaptive mesh refinement [254, 255, 88], hyperboloidal and compactified coordinates [256], gauge-damping drivers [87], and pseudospectral methods [257]. There is an ongoing programme of development, and one should expect to see a gradual improvement in code performance over time. We expect to see a step-function improvement from moving to work with curvature scalars (instead of metric components) as numerical variables, using the method proposed in Ref. [258]. This approach, which also automatically resolves the problem of linearly growing gauge modes, requires further development; in particular, the nontrivial extension from Schwarzschild to Kerr must be carried out. Further work is also required on eliminating initial junk radiation, perhaps through the formulation of more precise initial data for numerical evolutions.

All of the above applies to calculations for fixed geodesics. There are significantly greater challenges in the way of accurately simulating complete inspirals. The multiscale treatment of the system appears to be the optimal method for tackling this problem, as it naturally accounts for the system's two time scales and would

70

enable efficient frequency-domain calculations. But it has not yet been fully worked out. In Sec. 5.2, we noted that a naive implementation of the multiscale expansion leads to infrared divergences, which must be resolved by again using the method of matched asymptotic expansions, matching the two-timescale expansion to alternative expansions in the far field and near the horizon. Furthermore, the multiscale approach is complicated by the presence of transient resonances. These resonances introduce a third, transient time scale into the system, leading to a breakdown of the two-timescale expansion. The obvious way of handling this time scale is to utilize yet another expansion, this one tailored to the near-resonance regime. A complete description of the system would then be formed from the combination of the two-timescale, near-resonance, near-horizon, and near-infinity expansions.

Several alternatives to the two-timescale expansion could be pursued. Most obviously, one could directly tackle the coupled system of equations (41), (42), and (46), self-consistently evolving the gravitational field and the orbit that sources it. This may have substantial advantages in the less adiabatic regions of the parameter space where a two-timescale expansion would struggle (e.g., the transition from inspiral to plunge, highly eccentric orbits, and perhaps even the transition across resonances). But such an approach faces numerous obstacles. As described in Sec. 5.2, a naive implementation would incur secularly growing errors due to the long-term changes in the central black hole. Although this can likely be overcome [126], self-consistent evolution would prohibit many of the methods that have succeeded for fixed geodesic sources. Because the orbital frequencies change with time, discrete-spectrum frequency-domain methods would not apply. This suggests a time-domain approach. Currently, there is yet to be developed a time-domain code that computes the self-consistent evolution of the orbit under the GSF effect—such a calculation has only been performed so far for a scalar-charge toy model [161]. A code of that sort could come with a prohibitive computational cost: time-domain implementations of the perturbation equations are extremely computationally expensive, and with current technology it cannot be hoped that one would be able to self-consistently evolve an EMRI orbit for more than $O(10)$ orbits. However, as described above, time-domain methods are making advances, and they could ultimately be capable of performing a full self-consistent evolution.

Some more radical ideas for self-consistent evolutions include (1) the use of a fully spectral algorithm akin to SpEC [259] for solving the self-consistent perturbation equations; or (2) the use of a fully nonlinear NR code to solve for the full, nonperturbative metric of the EMRI system, with an excision boundary in an intermediate-scale buffer zone around the black hole [225]. A third idea would be to attempt self-consistent evolution without resorting to the time domain. The time-evolving frequency spectrum may be addressed using a *wavelet* decomposition instead of the standard Fourier decompositions: unlike Fourier harmonics, wavelets are localized in time (as well as in frequency), and in principle, a basis of such functions could naturally describe the slowly evolving perturbation in the EMRI problem. The details of all three proposals are yet to be worked out and their applicability and advantage assessed.

Another obvious, less radical alternative to the two-timescale method would be an extension of the method of osculating geodesics. To date, this is the only method that has been used to evolve orbits in practice (aside from leading-order, adiabatic evolutions), but currently it can only accommodate the first-order GSF. Accurate EMRI modelling will require some second-order effects, and the method will have to

be extended to account for them. Reference [24] provided an initial sketch of how that might be done.

Making that leap to second order, regardless of which evolution scheme is used, is now the most significant challenge within the programme to obtain accurate EMRI waveforms for LISA. So far, work to that end has been restricted to the simple test case of quasicircular orbits in Schwarzschild geometry, and even in that case only partial initial results have been obtained. We expect this to remain at the forefront of activity in the field over the coming years. The challenges here are significant. Even in the Schwarzschild case, extending the current frequency-domain (two-timescale), Lorenz-gauge-based method [74, 159] to noncircular orbits is quite nontrivial. The principal technical issue here is how to construct the second-order perturbation from a sum over inhomogeneous frequency modes while avoiding complications associated with the non-smoothness of the field at the particle. We envisage using a variant of the method of "extended particular solutions" introduced in [260] (itself a generalization of the method of "extended homogeneous solutions" [261]), which elegantly circumvents the problem. But the details are yet to be worked out.

A yet more challenging task is the extension of second-order calculations to Kerr: here, the existing frequency-domain Lorenz-gauge-based approach is invalid in its current form, due to the lack of separability of the field equations in Kerr. One could possibly work with coupled field equations, but we suspect the resulting scheme would be cumbersome to the point of intractability. It might be possible to tackle the second-order Kerr problem in the time domain, using self-consistent evolution in the Lorenz gauge; but in addition to the problems described above, this would involve an evolution of the field equations with a non-compact source. That presents its own potential difficulties, and it is yet to be attempted. It would be highly desirable to instead extend the radiation-gauge method that has facilitated calculations at first order. This would require a formalism of second-order metric reconstruction from curvature scalars satisfying separable equations on Kerr. The prospects for obtaining such a formalism in the near future remain unclear. For the purpose of EMRI modelling, it may also be possible to work entirely with curvature scalars. Sufficiently accurate EMRI models do not require *all* the information from second order, but only certain time-averaged dissipative effects. At first order, these effects can be computed entirely from curvature scalars; if the same is true at second order, one could bypass the need for a second-order metric-reconstruction formalism.

Besides ultimately providing accurate waveform templates for analysis of LISA data, self-force calculations can, on a shorter timescale, inform the gradual development of approximate EMRI waveforms that can be used in data-analysis studies. Such approximate models—known as "kludges"—may even prove useful as crude search templates in a first stage of a hierarchical, multi-step search algorithm. The idea of kludge templates has a long history in EMRI studies. The first kludge models [262, 263] incorporated back-reaction effects in an approximate manner, using post-Newtonian formulas, but there is a long-running programme [264, 265, 266, 267] aimed to improve the accuracy of the models using numerical data from actual self-force calculations. Efforts so far have concentrated on the simple Schwarzschild problem, but as self-force calculations in Kerr are now becoming available there is much scope for further development of kludges of increasing precision across the full parameter space of astrophysical EMRIs. We envisage using a simple, perhaps semi-analytical kludge model as an adjustable platform, with parameters that can be calibrated by fitting to numerical self-force data (in much the same way that the

EOB model of merging black hole binaries is calibrated using NR data). It is not unreasonable to expect that some form of an advanced self-force-calibrated kludge model could ultimately suffice for both detection and parameter extraction of LISA EMRIs.

Acknowledgments

[1] Chandrasekhar S 1985 The mathematical theory of black holes *Oxford, UK: Clarendon (1992) 646 p., OXFORD, UK: CLARENDON (1985) 646 P.*

[2] Abbott B P *et al.* (Virgo, LIGO Scientific) 2016 *Phys. Rev. Lett.* **116** 061102 (*Preprint* 1602.03837)

[3] Abbott B P *et al.* (Virgo, LIGO Scientific) 2016 *Phys. Rev. Lett.* **116** 241103 (*Preprint* 1606.04855)

[4] Abbott B P *et al.* (Virgo, LIGO Scientific) 2017 *Class. Quant. Grav.* **34** 104002 (*Preprint* 1611.07531)

[5] The *laser interferometer space antenna* (ESA website), https://www.elisascience.org/

[6] The *laser interferometer space antenna* (NASA website), https://lisa.nasa.gov/

[7] Babak S, Gair J, Sesana A, Barausse E, Sopuerta C F, Berry C P L, Berti E, Amaro-Seoane P, Petiteau A and Klein A 2017 *Phys. Rev.* **D95** 103012 (*Preprint* 1703.09722)

[8] Hulse R A and Taylor J H 1975 *Astrophys. J.* **195** L51

[9] Hinderer T and Flanagan E E 2008 *Phys. Rev.* **D78** 064028 (*Preprint* 0805.3337)

[10] Einstein A, Infeld L and Hoffmann B 1938 *Annals of Mathematics* **39** 65–100

[11] Infeld L and Schild A 1949 *Rev. Mod. Phys.* **21**

[12] Geroch R and Weatherall J O 2017 (*Preprint* 1707.04222)

[13] Dixon W G 1974 *Phil. Trans. Roy. Soc. Lond. A* **277** 59

[14] Mathisson M 1937 *Acta Phys. Pol.* **6** 163

[15] Papapetrou A 1952 *Proc. R. Soc. A* **64** 248–258

[16] Thorne K S and Hartle J B 1985 *Phys. Rev. D* **31** 1815

[17] D'Eath P 1975 *Phys. Rev. D* **11** 1387

[18] Kates R 1980 *Phys. Rev. D* **22** 1853

[19] Mino Y, Sasaki M and Tanaka T 1997 *Phys. Rev. D* **55** 3457–3476

[20] Quinn T C and Wald R M 1997 *Phys. Rev. D* **56** 3381–3394

[21] Detweiler S L 2001 *Phys. Rev. Lett.* **86** 1931–1934 (*Preprint* gr-qc/0011039)

[22] Detweiler S L and Whiting B F 2003 *Phys. Rev. D* **67** 024025

[23] Poisson E, Pound A and Vega I 2011 *Living Rev. Rel.* **14** 7 (*Preprint* 1102.0529)

[24] Pound A 2015 *Fund. Theor. Phys.* **179** 399–486 (*Preprint* 1506.06245)

[25] Harte A I 2015 *Fund. Theor. Phys.* **179** 327–398 (*Preprint* 1405.5077)

[26] Barack L 2009 *Class. Quant. Grav.* **26** 213001 (*Preprint* 0908.1664)

[27] Wardell B 2015 *Fund. Theor. Phys.* **179** 487–522 (*Preprint* 1501.07322)

[28] Amaro-Seoane P, Schutz B F and Thornburg J 2011 (*Preprint* 1102.3647)

[29] Amaro-Seoane P, Gair J R, Pound A, Hughes S A and Sopuerta C F 2015 *J. Phys. Conf. Ser.* **610** 012002 (*Preprint* 1410.0958)

[30] Landau L and Lifshitz E 1975 *The Classical Theory of Fields* 4th ed (New York: Pergamon Press, Oxford)

[31] Flanagan E and Wald R 1996 *Phys. Rev. D* **54** 6233–6283 (*Preprint* gr-qc/9602052)

[32] Spohn H 2000 *EPL* **50** 287

[33] Rohrlich F 2008 *Phys. Rev. E* **77** 046609

[34] Gralla S E, Harte A I and Wald R M 2009 *Phys. Rev. D* **80** 024031 (*Preprint* 0905.2391)

[35] Dirac P 1938 *Proc. R. Soc. London A* **167** 148

[36] Moxon J and Flanagan a 2017 (*Preprint* 1711.05212)

[37] DeWitt B and Brehme R 1960 *Ann. Phys. (N.Y.)* **9** 220–259

[38] Hobbs J 1968 *Ann. Phys. (N.Y.)* **47** 141–165
[39] Harte A I 2009 *Class.Quant.Grav.* **26** 155015 (*Preprint* 0903.0167)
[40] Geroch R and Traschen J 1987 *Phys. Rev. D* **36**(4) 1017–1031 URL
    http://link.aps.org/doi/10.1103/PhysRevD.36.1017
[41] Steinbauer R and Vickers J A 2009 *Class.Quant.Grav.* **26** 065001 (*Preprint* 0811.1376)
[42] Einstein A and Grommer J 1927 *Sitzungsber. Preuss. Akad. Wiss. Phys. Math. Kl.* **2**
[43] Einstein A and Infeld L 1949 *Can. J. Math.* **1** 209
[44] Eckhaus W 1979 *Asymptotic Analysis of Singular Perturbations* (New York: Elsevier
    North-Holland)
[45] Kevorkian J and Cole J D 1996 *Multiple scale and singular perturbation methods* (New York:
    Springer)
[46] D'Eath P D 1996 *Black holes: gravitational interactions* (New York: Oxford University Press)
[47] Kates R E 1980 *Phys. Rev. D* **22** 1879
[48] Futamase T 1987 *Phys. Rev. D* **36** 321
[49] Futamase T and Itoh Y 2007 *Living Reviews in Relativity* **10** URL
    http://www.livingreviews.org/lrr-2007-2
[50] Poisson E 2004 *Phys. Rev. D* **69** 084007 (*Preprint* gr-qc/0311026)
[51] Rosenthal E 2006 *Phys. Rev. D* **73** 044034
[52] Rosenthal E 2006 *Phys. Rev. D* **74** 084018
[53] Gralla S E and Wald R M 2008 *Class. Quant. Grav.* **25** 205009 [Erratum: Class. Quant.
    Grav.28,159501(2011)] (*Preprint* 0806.3293)
[54] Gralla S E 2012 *Phys. Rev. D* **85** 124011 (*Preprint* 1203.3189)
[55] Pound A 2010 *Phys. Rev. D* **81** 024023
[56] Pound A 2012 *Phys. Rev. Lett.* **109** 051101 (*Preprint* 1201.5089)
[57] Pound A 2017 *Phys. Rev.* **D95** 104056 (*Preprint* 1703.02836)
[58] Pound A 2012 *Phys. Rev. D* **86** 084019 (*Preprint* 1206.6538)
[59] Pound A 2015 *Phys. Rev. D* **92** 044021 (*Preprint* 1506.02894)
[60] Weyl H 1921 *Raum, Zeit, Materie, 4th ed.* (Berlin: Springer) chap 36
[61] Mino Y, Sasaki M and Tanaka T 1997 (*Preprint* gr-qc/9705073)
[62] Pound A and Miller J 2014 *Phys. Rev. D* **89** 104020 (*Preprint* 1403.1843)
[63] Harte A I 2012 *Class. Quant. Grav.* **29** 055012 (*Preprint* 1103.0543)
[64] Detweiler S 2012 *Phys. Rev. D* **85** 044048 (*Preprint* 1107.2098)
[65] Detweiler S L and Poisson E 2004 *Phys. Rev. D* **69** 084019 (*Preprint* gr-qc/0312010)
[66] Detweiler S L 2008 *Phys. Rev.* **D77** 124026 (*Preprint* 0804.3529)
[67] Barack L and Ori A 2001 *Phys. Rev.* **D64** 124003 (*Preprint* gr-qc/0107056)
[68] Gralla S E and Wald R M 2011 *Class. Quant. Grav.* **28** 177001 (*Preprint* 1104.5205)
[69] Gralla S E 2011 *Phys. Rev.* **D84** 084050 (*Preprint* 1104.5635)
[70] Pound A, Merlin C and Barack L 2014 *Phys. Rev.* **D89** 024009 (*Preprint* 1310.1513)
[71] Barack L and Ori A 2000 *Phys. Rev.* **D61** 061502 (*Preprint* gr-qc/9912010)
[72] Barack L 2001 *Phys. Rev.* **D64** 084021 (*Preprint* gr-qc/0105040)
[73] Barack L and Ori A 2003 *Phys. Rev. Lett.* **90** 111101 (*Preprint* gr-qc/0212103)
[74] Wardell B and Warburton N 2015 *Phys. Rev.* **D92** 084019 (*Preprint* 1505.07841)
[75] Heffernan A, Ottewill A C, Warburton N, Wardell B and Diener P 2017 (*Preprint*
    1712.01098)
[76] Heffernan A, Ottewill A and Wardell B 2012 *Phys. Rev.* **D86** 104023 (*Preprint* 1204.0794)
[77] Heffernan A, Ottewill A and Wardell B 2014 *Phys. Rev.* **D89** 024030 (*Preprint* 1211.6446)
[78] van de Meent M 2016 *Phys. Rev.* **D94** 044034 (*Preprint* 1606.06297)
[79] van de Meent M 2017 (*Preprint* 1711.09607)
[80] Rosenthal E 2005 *Phys. Rev. D* **72** 121503 (*Preprint* gr-qc/0508050)
[81] Barack L and Golbourn D A 2007 *Phys. Rev.* **D76** 044020 (*Preprint* 0705.3620)
[82] Barack L, Golbourn D A and Sago N 2007 *Phys. Rev.* **D76** 124036 (*Preprint* 0709.4588)
[83] Vega I and Detweiler S L 2008 *Phys. Rev.* **D77** 084008 (*Preprint* 0712.4405)
[84] Vega I, Diener P, Tichy W and Detweiler S L 2009 *Phys. Rev.* **D80** 084021 (*Preprint*
    0908.2138)
[85] Dolan S R and Barack L 2011 *Phys. Rev.* **D83** 024019 (*Preprint* 1010.5255)
[86] Dolan S R, Barack L and Wardell B 2011 *Phys. Rev.* **D84** 084001 (*Preprint* 1107.0012)
[87] Dolan S R and Barack L 2013 *Phys. Rev.* **D87** 084066 (*Preprint* 1211.4586)
[88] Thornburg J and Wardell B 2017 *Phys. Rev.* **D95** 084043 (*Preprint* 1610.09319)
[89] Warburton N and Wardell B 2014 *Phys. Rev.* **D89** 044046 (*Preprint* 1311.3104)
[90] Pound A 2015 *Phys. Rev.* **D92** 104047 (*Preprint* 1510.05172)
[91] Barack L and Lousto C O 2005 *Phys. Rev.* **D72** 104026 (*Preprint* gr-qc/0510019)

[92] Barack L and Sago N 2007 *Phys. Rev.* **D75** 064021 (*Preprint* `gr-qc/0701069`)

[93] Barack L and Sago N 2010 *Phys. Rev.* **D81** 084021 (*Preprint* `1002.2386`)

[94] Osburn T, Forseth E, Evans C R and Hopper S 2014 *Phys. Rev.* **D90** 104031 (*Preprint* `1409.4419`)

[95] Regge T and Wheeler J A 1957 *Phys. Rev.* **108** 1063

[96] Zerilli F J 1970 *Phys. Rev. D* **2** 2141

[97] Thompson J E, Whiting B F and Chen H 2017 *Class. Quant. Grav.* **34** 174001 (*Preprint* `1611.06214`)

[98] Chrzanowski P L 1975 *Phys. Rev.* **D11** 2042–2062

[99] Shah A G and Pound A 2015 *Phys. Rev. D* **91** 124022 (*Preprint* `1503.02414`)

[100] Merlin C, Ori A, Barack L, Pound A and van de Meent M 2016 *Phys. Rev.* **D94** 104066 (*Preprint* `1609.01227`)

[101] van De Meent M 2017 *Class. Quant. Grav.* **34** 124003 (*Preprint* `1702.00969`)

[102] Thompson J Private communication

[103] Keidl T S, Friedman J L and Wiseman A G 2007 *Phys. Rev.* **D75** 124009 (*Preprint* `gr-qc/0611072`)

[104] Keidl T S, Shah A G, Friedman J L, Kim D H and Price L R 2010 *Phys. Rev.* **D82** 124012 [Erratum: Phys. Rev.D90,no.10,109902(2014)] (*Preprint* `1004.2276`)

[105] Shah A G, Keidl T S, Friedman J L, Kim D H and Price L R 2011 *Phys. Rev.* **D83** 064018 (*Preprint* `1009.4876`)

[106] Shah A G, Friedman J L and Keidl T S 2012 *Phys. Rev.* **D86** 084059 (*Preprint* `1207.5595`)

[107] Merlin C and Shah A G 2015 *Phys. Rev.* **D91** 024005 (*Preprint* `1410.2998`)

[108] Osburn T, Warburton N and Evans C R 2016 *Phys. Rev.* **D93** 064024 (*Preprint* `1511.01498`)

[109] van de Meent M and Shah A G 2015 *Phys. Rev.* **D92** 064025 (*Preprint* `1506.04755`)

[110] Mano S, Suzuki H and Takasugi E 1996 *Prog. Theor. Phys.* **95** 1079–1096 (*Preprint* `gr-qc/9603020`)

[111] Shah A G, Friedman J L and Whiting B F 2014 *Phys. Rev.* **D89** 064042 (*Preprint* `1312.1952`)

[112] Johnson-McDaniel N K, Shah A G and Whiting B F 2015 *Phys. Rev.* **D92** 044007 (*Preprint* `1503.02638`)

[113] Bini D and Damour T 2014 *Phys. Rev.* **D89** 064063 (*Preprint* `1312.2503`)

[114] Bini D and Damour T 2014 *Phys. Rev.* **D89** 104047 (*Preprint* `1403.2366`)

[115] Bini D and Damour T 2015 *Phys. Rev.* **D91** 064050 (*Preprint* `1502.02450`)

[116] Bini D and Damour T 2015 *Phys. Rev.* **D91** 064064 (*Preprint* `1503.01272`)

[117] Kavanagh C, Ottewill A C and Wardell B 2015 *Phys. Rev.* **D92** 084025 (*Preprint* `1503.02334`)

[118] Hopper S, Kavanagh C and Ottewill A C 2016 *Phys. Rev.* **D93** 044010 (*Preprint* `1512.01556`)

[119] Kavanagh C, Ottewill A C and Wardell B 2016 *Phys. Rev.* **D93** 124038 (*Preprint* `1601.03394`)

[120] Bini D, Damour T and Geralico A 2016 *Phys. Rev.* **D93** 124058 (*Preprint* `1602.08282`)

[121] Carter B 1968 *Phys. Rev.* **174** 1559–1571

[122] Warburton N, Barack L and Sago N 2013 *Phys. Rev.* **D87** 084012 (*Preprint* `1301.3918`)

[123] Mino Y 2003 *Phys. Rev.* **D67** 084027 (*Preprint* `gr-qc/0302075`)

[124] Schmidt W 2002 *Class. Quant. Grav.* **19** 2743 (*Preprint* `gr-qc/0202090`)

[125] van de Meent M 2014 *Phys. Rev.* **D89** 084033 (*Preprint* `1311.4457`)

[126] Flanagan E, Hinderer T, Moxon J and Pound A The two-body problem in general relativity in the extreme-mass-ratio limit via multiscale expansions. i. foundations in preparation

[127] Mino Y and Price R 2008 *Phys. Rev. D* **77** 064001 (*Preprint* `0801.0179`)

[128] Van De Meent M and Warburton N 2018 (*Preprint* `1802.05281`)

[129] Flanagan a E, Hughes S A and Ruangsri U 2014 *Phys. Rev.* **D89** 084028 (*Preprint* `1208.3906`)

[130] Ruangsri U and Hughes S A 2014 *Phys. Rev.* **D89** 084036 (*Preprint* `1307.6483`)

[131] Berry C P L, Cole R H, Cañizares P and Gair J R 2016 *Phys. Rev.* **D94** 124042 (*Preprint* `1608.08951`)

[132] Mihaylov D P and Gair J R 2017 (*Preprint* `1706.06639`)

[133] Flanagan E E and Hinderer T 2012 *Phys. Rev. Lett.* **109** 071102 (*Preprint* `1009.4923`)

[134] Gair J, Yunes N and Bender C M 2012 *J. Math. Phys.* **53** 032503 (*Preprint* `1111.3605`)

[135] Isoyama S, Fujita R, Nakano H, Sago N and Tanaka T 2013 *PTEP* **2013** 063E01 (*Preprint* `1302.4035`)

[136] Brink J, Geyer M and Hinderer T 2015 *Phys. Rev. Lett.* **114** 081102 (*Preprint* `1304.0330`)

[137] Brink J, Geyer M and Hinderer T 2015 *Phys. Rev.* **D91** 083001 (*Preprint* `1501.07728`)

[138] Steinhoff J and Puetzfeld D 2012 *Phys. Rev.* **D86** 044033 (*Preprint* `1205.3926`)

[139] Dolan S R, Nolan P, Ottewill A C, Warburton N and Wardell B 2015 *Phys. Rev.* **D91** 023009 (*Preprint* `1406.4890`)
[140] Quinn T C and Wald R M 1999 *Phys. Rev.* **D60** 064009 (*Preprint* `gr-qc/9903014`)
[141] Sago N, Tanaka T, Hikida W, Ganz K and Nakano H 2006 *Prog. Theor. Phys.* **115** 873–907 (*Preprint* `gr-qc/0511151`)
[142] Ganz K, Hikida W, Nakano H, Sago N and Tanaka T 2007 *Prog. Theor. Phys.* **117** 1041–1066 (*Preprint* `gr-qc/0702054`)
[143] Gal'tsov V 1982 *J. Phys. A* **15** 3737
[144] Teukolsky S A and Press W H 1974 *Astrophys. J.* **193** 443–461
[145] Hughes S A 2000 *Phys. Rev.* **D61** 084004 [Erratum: Phys. Rev.D90,no.10,109904(2014)] (*Preprint* `gr-qc/9910091`)
[146] Glampedakis K and Kennefick D 2002 *Phys. Rev.* **D66** 044002 (*Preprint* `gr-qc/0203086`)
[147] Poisson E 1995 *Phys. Rev.* **D52** 5719–5723 [Addendum: Phys. Rev.D55,7980(1997)] (*Preprint* `gr-qc/9505030`)
[148] Drasco S and Hughes S A 2006 *Phys. Rev.* **D73** 024027 [Erratum: Phys. Rev.D90,no.10,109905(2014)] (*Preprint* `gr-qc/0509101`)
[149] Martel K 2004 *Phys. Rev.* **D69** 044025 (*Preprint* `gr-qc/0311017`)
[150] Poisson E 2004 *Phys. Rev.* **D70** 084044 (*Preprint* `gr-qc/0407050`)
[151] Sundararajan P A, Khanna G, Hughes S A and Drasco S 2008 *Phys. Rev.* **D78** 024022 (*Preprint* `0803.0317`)
[152] Hughes S A, Drasco S, Flanagan E E and Franklin J 2005 *Phys. Rev. Lett.* **94** 221101 (*Preprint* `gr-qc/0504015`)
[153] Hughes S A 2017 Adiabatic and post-adiabatic approaches to extreme mass ratio inspiral *Proceedings, 14th Marcel Grossmann Meeting on Recent Developments in Theoretical and Experimental General Relativity, Astrophysics, and Relativistic Field Theories (MG14): Rome, Italy, July 12-18, 2015* vol 2 pp 1953–1959 (*Preprint* `1601.02042`) URL `https://inspirehep.net/record/1414738/files/arXiv:1601.02042.pdf`
[154] Ori A 1997 *Phys. Rev.* **D55** 3444–3456
[155] Mino Y 2005 *Class. Quant. Grav.* **22** S717–S738 (*Preprint* `gr-qc/0506002`)
[156] Drasco S, Flanagan E E and Hughes S A 2005 *Class. Quant. Grav.* **22** S801–846 (*Preprint* `gr-qc/0505075`)
[157] Sundararajan P A, Khanna G and Hughes S A 2007 *Phys. Rev.* **D76** 104005 (*Preprint* `gr-qc/0703028`)
[158] Pound A 2014 *Phys. Rev.* **D90** 084039 (*Preprint* `1404.1543`)
[159] Miller J, Wardell B and Pound A 2016 *Phys. Rev.* **D94** 104018 (*Preprint* `1608.06783`)
[160] Pound A, Poisson E and Nickel B G 2005 *Phys. Rev.* **D72** 124001 (*Preprint* `gr-qc/0509122`)
[161] Diener P, Vega I, Wardell B and Detweiler S 2012 *Phys. Rev. Lett.* **108** 191102 (*Preprint* `1112.4821`)
[162] Pound A and Poisson E 2008 *Phys. Rev.* **D77** 044013 (*Preprint* `0708.3033`)
[163] Gair J R, Flanagan E E, Drasco S, Hinderer T and Babak S 2011 *Phys. Rev.* **D83** 044037 (*Preprint* `1012.5111`)
[164] Warburton N, Akcay S, Barack L, Gair J R and Sago N 2012 *Phys. Rev.* **D85** 061501 (*Preprint* `1111.6908`)
[165] Warburton N, Osburn T and Evans C R 2017 (*Preprint* `1708.03720`)
[166] Fujita R, Isoyama S, Le Tiec A, Nakano H, Sago N and Tanaka T 2017 *Class. Quant. Grav.* **34** 134001 (*Preprint* `1612.02504`)
[167] van de Meent M 2014 *Phys. Rev.* **D90** 044027 (*Preprint* `1406.2594`)
[168] Hirata C M 2011 *Phys. Rev.* **D83** 104024 (*Preprint* `1011.4987`)
[169] Kennefick D and Ori A 1996 *Phys. Rev.* **D53** 4319–4326 (*Preprint* `gr-qc/9512018`)
[170] Ori A and Thorne K S 2000 *Phys. Rev.* **D62** 124022 (*Preprint* `gr-qc/0003032`)
[171] Buonanno A and Damour T 2000 *Phys. Rev.* **D62** 064015 (*Preprint* `gr-qc/0001013`)
[172] Barack L and Sago N 2009 *Phys. Rev. Lett.* **102** 191101 (*Preprint* `0902.0573`)
[173] Diaz-Rivera L M, Messaritaki E, Whiting B F and Detweiler S L 2004 *Phys. Rev.* **D70** 124018 (*Preprint* `gr-qc/0410011`)
[174] Akcay S, Barack L, Damour T and Sago N 2012 *Phys. Rev.* **D86** 104041 (*Preprint* `1209.0964`)
[175] Damour T 2010 *Phys. Rev.* **D81** 024017 (*Preprint* `0910.5533`)
[176] Le Tiec A, Barausse E and Buonanno A 2012 *Phys. Rev. Lett.* **108** 131103 (*Preprint* `1111.5609`)
[177] Le Tiec A, Blanchet L and Whiting B F 2012 *Phys. Rev.* **D85** 064039 (*Preprint* `1111.5378`)
[178] Blanchet L, Buonanno A and Le Tiec A 2013 *Phys. Rev.* **D87** 024030 (*Preprint* `1211.1060`)
[179] Isoyama S, Barack L, Dolan S R, Le Tiec A, Nakano H, Shah A G, Tanaka T and Warburton

        N 2014 *Phys. Rev. Lett.* **113** 161101 (*Preprint* `1404.6133`)

[180] Barack L, Damour T and Sago N 2010 *Phys. Rev.* **D82** 084036 (*Preprint* `1008.0935`)

[181] van de Meent M 2017 *Phys. Rev. Lett.* **118** 011101 (*Preprint* `1610.03497`)

[182] de Sitter W 1916 *Mon. Not. Roy. Astron. Soc.* **76** 699–728

[183] Everitt C W F, DeBra D B, Parkinson B W, Turneaure J P, Conklin J W, Heifetz M I, Keiser
        G M, Silbergleit A S, Holmes T, Kolodziejczak J, Al-Meshari M, Mester J C, Muhlfelder B,
        Solomonik V G, Stahl K, Worden P W, Bencze W, Buchman S, Clarke B, Al-Jadaan A,
        Al-Jibreen H, Li J, Lipa J A, Lockhart J M, Al-Suwaidan B, Taber M and Wang S 2011
        Phys. Rev. Lett. **106**(22) 221101

[184] Breton R P, Kaspi V M, Kramer M, McLaughlin M A, Lyutikov M, Ransom S M, Stairs I H,
        Ferdman R D, Camilo F and Possenti A 2008 *Science* **321** 104–107 (*Preprint* `0807.2644`)

[185] Dolan S R, Warburton N, Harte A I, Le Tiec A, Wardell B and Barack L 2014 *Phys. Rev.*
        D89 064011 (*Preprint* `1312.0775`)

[186] Akcay S, Dempsey D and Dolan S R 2017 *Class. Quant. Grav.* **34** 084001 (*Preprint*
        `1608.04811`)

[187] Kavanagh C, Bini D, Damour T, Hopper S, Ottewill A C and Wardell B 2017 (*Preprint*
        `1706.00459`)

[188] Akcay S 2017 *Phys. Rev.* **D96** 044024 (*Preprint* `1705.03282`)

[189] van de Meent M (private communication)

[190] Bini D and Damour T 2014 *Phys. Rev.* **D90** 024039 (*Preprint* `1404.2747`)

[191] Taylor S and Poisson E 2008 *Phys. Rev.* **D78** 084016 (*Preprint* `0806.3052`)

[192] Johnson-McDaniel N K, Yunes N, Tichy W and Owen B J 2009 *Phys. Rev.* **D80** 124039
        (*Preprint* `0907.0891`)

[193] Nolan P, Kavanagh C, Dolan S R, Ottewill A C, Warburton N and Wardell B 2015 *Phys.
        Rev.* **D92** 123008 (*Preprint* `1505.04447`)

[194] Pound A 2015 Self-force effects on the horizon geometry of a small black hole (unpublished
        notes)

[195] Zimmerman A, Lewis A G M and Pfeiffer H P 2016 *Phys. Rev. Lett.* **117** 191101 (*Preprint*
        `1606.08056`)

[196] Sago N, Barack L and Detweiler S L 2008 *Phys. Rev.* **D78** 124024 (*Preprint* `0810.2530`)

[197] Le Tiec A 2015 *Phys. Rev.* **D92** 084021 (*Preprint* `1506.05648`)

[198] Blanchet L and Le Tiec A 2017 *Class. Quant. Grav.* **34** 164001 (*Preprint* `1702.06839`)

[199] Barack L and Sago N 2011 *Phys. Rev.* **D83** 084023 (*Preprint* `1101.3331`)

[200] Akcay S, Le Tiec A, Barack L, Sago N and Warburton N 2015 *Phys. Rev.* **D91** 124014
        (*Preprint* `1503.01374`)

[201] Bini D, Damour T and Geralico a 2016 *Phys. Rev.* **D93** 104017 (*Preprint* `1601.02988`)

[202] Akcay S and van de Meent M 2016 *Phys. Rev.* **D93** 064063 (*Preprint* `1512.03392`)

[203] Bini D, Damour T and Geralico A 2016 *Phys. Rev.* **D93** 064023 (*Preprint* `1511.04533`)

[204] Aksteiner S and Bäckdahl T 2018 (*Preprint* `1803.05341`)

[205] Le Tiec A 2014 *Int. J. Mod. Phys.* **D23** 1430022 (*Preprint* `1408.5505`)

[206] Blanchet L, Detweiler S L, Le Tiec A and Whiting B F 2010 *Phys. Rev.* **D81** 064004
        (*Preprint* `0910.0207`)

[207] Blanchet L, Detweiler S L, Le Tiec A and Whiting B F 2010 *Phys. Rev.* **D81** 084033
        (*Preprint* `1002.0726`)

[208] Blanchet L, Faye G and Whiting B F 2014 *Phys. Rev.* **D90** 044017 (*Preprint* `1405.5151`)

[209] Bini D and Damour T 2016 *Phys. Rev.* **D93** 104040 (*Preprint* `1603.09175`)

[210] Bini D and Damour T 2017 (*Preprint* `1706.06877`)

[211] Damour T, Jaranowski P and Schaefer G 2000 *Phys. Rev.* **D62** 084011 (*Preprint*
        `gr-qc/0005034`)

[212] Blanchet L and Iyer B R 2003 *Class. Quant. Grav.* **20** 755 (*Preprint* `gr-qc/0209089`)

[213] Favata M 2011 *Phys. Rev.* **D83** 024027 (*Preprint* `1008.4622`)

[214] Damour T, Jaranowski P and Schäfer G 2014 *Phys. Rev.* **D89** 064058 (*Preprint* `1401.4548`)

[215] Damour T, Jaranowski P and Schäfer G 2015 *Phys. Rev.* **D91** 084024 (*Preprint* `1502.07245`)

[216] Bernard L, Blanchet L, Bohé A, Faye G and Marsat S 2016 *Phys. Rev.* **D93** 084037 (*Preprint*
        `1512.02876`)

[217] Damour T, Jaranowski P and Schäfer G 2016 *Phys. Rev.* **D93** 084014 (*Preprint* `1601.01283`)

[218] Bernard L, Blanchet L, Bohé A, Faye G and Marsat S 2017 *Phys. Rev.* **D95** 044026 (*Preprint*
        `1610.07934`)

[219] Marchand T, Bernard L, Blanchet L and Faye G 2017 (*Preprint* `1707.09289`)

[220] Veitch J, Pürrer M and Mandel I 2015 *Phys. Rev. Lett.* **115** 141101 (*Preprint* `1503.05953`)

[221] Le Tiec A, Mroue A H, Barack L, Buonanno A, Pfeiffer H P, Sago N and Taracchini A 2011

Phys. Rev. Lett. **107** 141101 (*Preprint* 1106.3278)

[222] Boyle M, Brown D A, Kidder L E, Mroue A H, Pfeiffer H P, Scheel M A, Cook G B and Teukolsky S A 2007 *Phys. Rev.* **D76** 124038 (*Preprint* 0710.0158)

[223] Le Tiec A *et al.* 2013 *Phys. Rev.* **D88** 124027 (*Preprint* 1309.0541)

[224] Lewis A G M, Zimmerman A and Pfeiffer H P 2017 *Class. Quant. Grav.* **34** 124001 (*Preprint* 1611.03418)

[225] Schutz B F *Presentation at the 20th Capra meeting on radiation reaction in genaral relativity, Chapel Hill (2017)*

[226] Buonanno A and Damour T 1999 *Phys. Rev.* **D59** 084006 (*Preprint* gr-qc/9811091)

[227] Damour T 2001 *Phys. Rev.* **D64** 124013 (*Preprint* gr-qc/0103018)

[228] Damour T and Nagar A 2011 *Fundam. Theor. Phys.* **162** 211–252 [,211(2009)] (*Preprint* 0906.1769)

[229] Buonanno A, Pan Y, Pfeiffer H P, Scheel M A, Buchman L T and Kidder L E 2009 *Phys. Rev.* **D79** 124028 (*Preprint* 0902.0790)

[230] Taracchini A *et al.* 2014 *Phys. Rev.* **D89** 061502 (*Preprint* 1311.2544)

[231] Bohé A *et al.* 2017 *Phys. Rev.* **D95** 044028 (*Preprint* 1611.03703)

[232] Damour T and Nagar A 2010 *Phys. Rev.* **D81** 084016 (*Preprint* 0911.5041)

[233] Bernuzzi S, Nagar A, Dietrich T and Damour T 2015 *Phys. Rev. Lett.* **114** 161103 (*Preprint* 1412.4553)

[234] Damour T, Guercilena F, Hinder I, Hopper S, Nagar A and Rezzolla L 2014 *Phys. Rev.* **D89** 081503 (*Preprint* 1402.7307)

[235] Bini D, Damour T and Geralico A 2015 *Phys. Rev.* **D92** 124058 [Erratum: Phys. Rev.D93,no.10,109902(2016)] (*Preprint* 1510.06230)

[236] Bini D and Damour T 2014 *Phys. Rev.* **D90** 124037 (*Preprint* 1409.6933)

[237] Nagar A and Akcay S 2012 *Phys. Rev.* **D85** 044025 (*Preprint* 1112.2840)

[238] Nagar A and Shah A 2016 *Phys. Rev.* **D94** 104017 (*Preprint* 1606.00207)

[239] Harms E, Lukes-Gerakopoulos G, Bernuzzi S and Nagar A 2016 *Phys. Rev.* **D94** 104010 (*Preprint* 1609.00356)

[240] Lukes-Gerakopoulos G, Harms E, Bernuzzi S and Nagar A 2017 (*Preprint* 1707.07537)

[241] Costa L F O and Natário J 2015 *Fund. Theor. Phys.* **179** 215–258 (*Preprint* 1410.6443)

[242] Wardell B, Galley C R, Zenginoğlu A, Casals M, Dolan S R and Ottewill A C 2014 *Phys. Rev. D* **89** 084021 (*Preprint* 1401.1506)

[243] Galley C R and Hu B L 2009 *Phys. Rev.* **D79** 064002 (*Preprint* 0801.0900)

[244] Galley C R 2012 *Class. Quant. Grav.* **29** 015010 (*Preprint* 1012.4488)

[245] Galley C R 2012 *Class. Quant. Grav.* **29** 015011 (*Preprint* 1107.0766)

[246] Galley C R and Porto R A 2013 *JHEP* **11** 096 (*Preprint* 1302.4486)

[247] Zimmerman P and Poisson E 2014 *Phys. Rev.* **D90** 084030 (*Preprint* 1406.5111)

[248] Linz T M, Friedman J L and Wiseman A G 2014 *Phys. Rev.* **D90** 084031 (*Preprint* 1406.5112)

[249] Zimmerman P 2015 *Phys. Rev.* **D92** 064040 (*Preprint* 1505.03915)

[250] Zimmerman P 2015 *Phys. Rev.* **D92** 064051 (*Preprint* 1507.04076)

[251] Harte A I, Taylor P and Flanagan a 2018 (*Preprint* 1804.03702)

[252] Hopper S and Cardoso V 2017 (*Preprint* 1706.02791)

[253] Hopper S 2017 (*Preprint* 1706.05455)

[254] Thornburg J 2011 *Gen. Rel. Grav.* **43** 1211–1251 (*Preprint* 0909.0036)

[255] Thornburg J 2010 (*Preprint* 1006.3788)

[256] Harms E, Bernuzzi S, Nagar A and Zenginoglu A 2014 *Class. Quant. Grav.* **31** 245004 (*Preprint* 1406.5983)

[257] Canizares P, Sopuerta C F and Jaramillo J L 2010 *Phys. Rev.* **D82** 044023 (*Preprint* 1006.3201)

[258] Barack L and Giudice P 2017 *Phys. Rev.* **D95** 104033 (*Preprint* 1702.04204)

[259] The *spectral einstein code*, https://www.black-holes.org/code/SpEC.html

[260] Hopper S and Evans C R 2013 *Phys. Rev.* **D87** 064008 (*Preprint* 1210.7969)

[261] Barack L, Ori A and Sago N 2008 *Phys. Rev.* **D78** 084021 (*Preprint* 0808.2315)

[262] Barack L and Cutler C 2004 *Phys. Rev.* **D69** 082005 (*Preprint* gr-qc/0310125)

[263] Babak S, Fang H, Gair J R, Glampedakis K and Hughes S A 2007 *Phys. Rev.* **D75** 024005 [Erratum: Phys. Rev.D77,04990(2008)] (*Preprint* gr-qc/0607007)

[264] Huerta E A and Gair J R 2009 *Phys. Rev.* **D79** 084021 [Erratum: Phys. Rev.D84,049903(2011)] (*Preprint* 0812.4208)

[265] Huerta E A, Kumar P, Gair J R and McWilliams S T 2014 *Phys. Rev.* **D90** 024024 (*Preprint* 1403.0561)

[266] Chua A J K and Gair J R 2015 *Class. Quant. Grav.* **32** 232002 (*Preprint* 1510.06245)

[267] Chua A J K, Moore C J and Gair J R 2017 *Phys. Rev.* **D96** 044005 (*Preprint* 1705.04259)