Topics in Gravitational-Wave Astronomy

Theoretical studies, source modelling & statistical methods

ALVIN J. K. CHUA

CLARE HALL / INSTITUTE OF ASTRONOMY / UNIVERSITY OF CAMBRIDGE

A dissertation submitted for the degree of Doctor of Philosophy

Supervisor: Dr Jonathan R. Gair

1 MARCH 2017

Abstract

Astronomy with gravitational-wave observations is now a reality. Much of the theoretical research in this field falls under three broad themes: the mathematical description and physical understanding of gravitational radiation and its effects; the construction of accurate and computationally efficient waveform models for astrophysical sources; and the improved statistical analysis of noisy data from interferometric detectors, so as to extract and characterise source signals. The doctoral thesis presented in this dissertation is an investigation of various topics across these themes.

Under the first theme, we examine the direct interaction between gravitational waves and electromagnetic fields in a self-contained theoretical study; this is done with a view to understanding the observational implications for highly energetic astrophysical events that radiate in both the gravitational and electromagnetic sectors. We then delve into the second theme of source modelling by developing and implementing an improved waveform model for the extreme-mass-ratio inspirals of stellar-mass compact objects into supermassive black holes, which are an important class of source for future space-based detectors such as the Laser Interferometer Space Antenna.

Two separate topics are explored under the third theme of data analysis. We begin with the procedure of searching for gravitational-wave signals in detector data, and propose several combinatorial compression schemes for the large banks of waveform templates that are matched against putative signals, before studying the usefulness of these schemes for accelerating searches. After a gravitational-wave source is detected, the follow-up process is to measure its parameters in detail from the data; this is addressed as we apply the machine-learning technique of Gaussian process regression to gravitational-wave data analysis, and in particular to the formidable problem of parameter estimation for extreme-mass-ratio inspirals.

Acknowledgements

The completion of a doctoral thesis is an intrinsically individual and (some might say) thankless experience. It is however by no means isolated, and should not pass without a word of thanks to those who have helped.

I am indebted to my supervisor Jon Gair for the opportunity of working with him, the overarching guidance he has provided, and indeed just for being so good at what he does. His broad and creative approach to research is inspiring, while his experience and stature in the field have been an invaluable conduit for my exposure to gravitational-wave astronomy.

Further gratitude goes to Chris Moore, who is always ready to talk science, and equally adept at bouncing ideas around or combing through the details. Daily life in the office has been an enriching—and often exhausting—learning experience thanks to both him and Rob Cole, who has provided much help as well over the years. I have likewise enjoyed my interactions with other past members of the IoA gravity group, some of whom I have had the additional pleasure of collaborating with: Christopher Berry, Tom Callister, Priscilla Cañizares, Sonke Hee, Eliu Huerta, and Steve Taylor.

For graciously serving as both my academic referees and thesis examiners, as well as for advice freely proffered and often heeded, I thank Leor Barack and Anthony Lasenby. I am also grateful for science- and job-related discussions with various researchers at the IoA and around the globe: Mustafa Amin, Stas Babak, John Baker, Anthony Challinor, Scott Field, Ik Siong Heng, Scott Hughes, Donald Lynden-Bell, Nikku Madhusudhan, Michele Vallisneri, Maarten van de Meent, and Yan Wang (both of them).

Finally, I greatly appreciate the assorted assistance afforded by Debbie Peterson and Margaret Harding at the IoA, Irene Hills and Magda Bergman at Clare Hall, and the good people at the Cambridge Trust.

This thesis is dedicated to my family—for everything, and then some.

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the preface and specified in the text.

It is not substantially the same as any that I have submitted, or that is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution, except as declared in the preface and specified in the text.

I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other university or similar institution, except as declared in the preface and specified in the text.

All concepts, results and data obtained by others are properly referenced. The use of the collaborative "we" in the text is a stylistic choice.

Some of the material in this dissertation has been adapted from the following journal publications:

- A. J. K. Chua, P. Cañizares & J. R. Gair, *Electromagnetic signatures of far-field gravitational radiation in the 1+3 approach*, Class. Quantum Grav. **32**, 015011 (2015).

- A. J. K. Chua & J. R. Gair, *Improved analytic extreme-mass-ratio inspiral model for scoping out eLISA data analysis*, Fast Track Communication, Class. Quantum Grav. **32**, 232002 (2015).

- A. J. K. Chua, *Augmented kludge waveforms and Gaussian process regression for EMRI data analysis*, J. Phys.: Conf. Ser. **716**, 012028 (2016).

- A. J. K. Chua & J. R. Gair, *Tunable compression of template banks for fast gravitational-wave detection and localisation*, Phys. Rev. D **93**, 122001 (2016).

- C. J. Moore, C. P. L. Berry, A. J. K. Chua & J. R. Gair, *Improving gravitational-wave parameter estimation using Gaussian process regression*, Phys. Rev. D **93**, 064001 (2016).

- C. J. Moore, A. J. K. Chua, C. P. L. Berry & J. R. Gair, *Fast methods for training Gaussian processes on large datasets*, R. Soc. Open Sci. **3**, 160125 (2016).

Explicit citation of these papers indicates results that have since been updated (as in Chapter 4), or work done mainly in collaboration (as in Chapter 6).

This dissertation is shorter than 60,000 words in length, including the abstract, tables, footnotes and appendices. The word count is 52,000 words (evaluated using Kile).

— Alvin J. K. Chua
1 March 2017

Contents

List of Figures and Tables

Chapter 1

Introduction

Gravitational waves (GWs) are radiative perturbations in the curvature of spacetime. They are a confirmed prediction of Einstein's general theory of relativity [1], and hence an essential feature of any other viable theory of gravitation. The measurement of GW effects is on the long list of tests that general relativity has undergone (and passed with flying colours), but the intrinsic weakness of these effects means they can only be detected for astrophysical sources in much stronger gravitational fields than those probed by Solar System experiments. This allows GWs to place exclusive constraints on gravity in the strong-field regime, as well as to reveal and describe highly energetic sources that might be inaccessible to traditional electromagnetic astronomy.

Radio observations of the Hulse–Taylor binary pulsar PSR B1913+16 [2] in the 1970s provided the first compelling evidence for the existence of GWs. This pulsar and its neutron star companion are losing orbital energy through the emission of GWs, which causes the orbit of the binary to shrink; the rate of decrease in orbital period that is predicted by general relativity agrees exquisitely with the measured value. A number of similar binaries were later found and observed with even greater precision [3, 4], lending further credence to GWs and their description in the underlying theory.

The detection of GWs in a more direct sense (i.e. measuring their effects as they pass through a detector) is notoriously difficult, due to the weak coupling of gravity to matter. Efforts to directly detect astrophysical GWs began with the resonant detectors of the 1960s, which were designed to pick up resonantly amplified antenna oscillations caused by a passing wave [5, 6]. Although unsuccessful, these attempts paved the way for the adoption of laser interferometry in GW detection, and the eventual construction of large-scale interferometers such as the Laser Interferometer Gravitational-wave Observatory (LIGO) [7]. These highly sensitive instruments operate by measuring the phase difference of light along their angled arms, which changes in response to the miniscule distortions of spacetime as a GW passes through.

In 2015, the twin Advanced LIGO detectors situated at both ends of the continental United States simultaneously recorded a transient GW signal from the merger of two stellar-mass black holes [8]. This discovery was essentially a pair of first detections: not just that of GWs, but also the energy radiated directly by a black-hole system (which is dark in the electromagnetic sector by definition). It was shortly followed by another confirmed black-hole binary merger in the same observational run [9, 10]. More detections of stellar-mass GW sources are anticipated during the presently ongoing run and into the next decade, as Advanced LIGO is brought up to full design sensitivity.

While the LIGO discoveries are the culmination of a generation-long research and development programme targeting the direct detection of GWs, they are also the herald of a brand new era in astronomy—one that extends its reach beyond the electromagnetic spectrum for the first time, and into the radiation spectrum of gravity. GW sources with characteristic masses of less than 10^3 Solar masses radiate most strongly in the kilohertz (10^1–10^4 Hz) band of the gravitational spectrum. These sources will be found and observed by a global detector network that comprises Advanced LIGO and other ground-based interferometers such as Advanced Virgo [11] and KAGRA [12].

Another exciting prospect in GW astronomy is the detection of longer-lived signals from sources exceeding 10^5 Solar masses; these will involve the supermassive black holes (SMBHs) that reside at the centres of galaxies, and thus revolutionise the study of formation rates and evolution scenarios for such objects. There will be an abundance of these signals in the sensitivity band of millihertz (10^{-4}–10^0 Hz) space-based detectors such as the proposed Laser Interferometer Space Antenna (LISA) [13][1], or more ambitious missions such as DECIGO [15] and TianQin [16]. Although LISA is still over a decade away

[1] This is the recently submitted L3 mission proposal for ESA's Cosmic Vision programme, in which the design of LISA has been restored from the down-scoped eLISA instrument [14] to an earlier configuration with higher sensitivity.

from coming to fruition, recent results from the precursor LISA Pathfinder satellite [17] have already successfully demonstrated the technological viability of observing GW sources from space [18].

At even lower frequencies, precise timing observations of Galactic millisecond pulsars by radio telescopes allow arrays of such pulsars to be used as natural detectors in the nanohertz (10^{-9}–10^{-6} Hz) gravitational band. An international consortium of three pulsar timing array collaborations [19] has been formed to detect and study GW sources in this band. The most promising source for these arrays is the stochastic background signal from the cosmological population of inspiralling SMBH binaries, for which detection is expected within the next ten years [20]. With the advent of pulsar timing arrays and space-based detectors, a near-complete observational coverage of the gravitational spectrum can be obtained—this will in turn unlock the full scientific potential of multiband GW astronomy [21], thereby enhancing the synergy of the nascent field with its electromagnetic counterpart.

1.1 Chapter synopses

We provide an introductory overview of GW astronomy in Chapter 2, with a strong focus on its theoretical aspects: the mathematical/physical framework of GWs in general relativity, the astrophysical modelling of source signals, and the statistical analysis of detector data.

In Chapter 3, the interactions between far-field GWs and electromagnetic fields are studied using the 1+3 approach to general relativity. Many known effects are rederived in this framework, while new results are also found. These interactions are typically extremely weak even under astrophysical conditions, but can give rise to potentially observable electromagnetic signatures.

Chapter 4 describes the development and implementation of an improved approximate waveform model for the extreme-mass-ratio inspirals of stellar-mass compact objects into supermassive black holes, which will be observable by LISA and other space-based detectors. The computationally efficient new model uses the three frequencies of an orbit around a spinning Kerr black hole to better match the accuracy of more expensive models.

In Chapter 5, a general method of tunable compression is proposed for the large banks of waveform templates that are used in GW detection searches. Various combinatorial schemes under this method are investigated with toy waveform models. The best one is applied to a simple example featuring actual waveforms, and yields results that are promising for practical implementation.

Chapter 6 explores the viability of using the machine-learning technique known as Gaussian process regression in GW parameter estimation. The method employs a Bayesian approach to account for the systematic bias that might occur due to theoretical error in waveform models. Through low-dimensional studies, it is demonstrated to be potentially useful for measuring the astrophysical parameters of extreme-mass-ratio inspirals.

Finally, concluding remarks and additional discussion of possible future research directions are given in Chapter 7.

1.2 Preliminaries

Readers are assumed to have at least a rudimentary knowledge of differential geometry, general relativity, Bayesian statistics and signal analysis.

1.2.1 Conventions

Throughout this dissertation, we use the common set of conventions laid out in this section, unless otherwise stated here (and in the respective chapters). Notation in each chapter has been chosen for intuitiveness, concordance with standard notation, and to avoid symbolic overload. Due to the variety of topics covered in this dissertation, adherence to the first two constraints may give rise to symbols that are multiply defined across chapters (or even across sections in the same chapter), but we trust context to prevent confusion.

- Latin (spacetime) indices run from 0 to 3, while Greek (space) indices run from 1 to 3; exceptions to this rule will be explicitly flagged. The Einstein summation convention applies only to lower–upper pairs of indices.

- The metric signature is $(-,+,+,+)$, and the Riemann tensor sign convention is $R_{ab} = R^{c}{}_{acb}$.

- The partial-derivative operator ∂_a is denoted by an index comma:

$$T_{,a} := \partial_a T. \tag{1.1}$$

  The covariant-derivative operator ∇_a is denoted by an index semicolon:

$$T_{;a} := \nabla_a T. \tag{1.2}$$

- Symmetrisation is denoted by round brackets around indices, e.g.

$$T_{(ab)} := \frac{1}{2}(T_{ab} + T_{ba}). \tag{1.3}$$

  Skew-symmetrisation is denoted by square brackets around indices, e.g.

$$T_{[abc]} := \frac{1}{6}(T_{abc} + T_{bca} + T_{cab} - T_{acb} - T_{bac} - T_{cba}). \tag{1.4}$$

- Geometrised units such that $c = G = 1$ are adopted, where c is the speed of light in vacuum and G is the gravitational constant. Exceptions occur where units are explicitly restored, and also throughout Chapter 3 where the cosmological convention $c = 8\pi G = \mu_0 = 1$ is used (with μ_0 being the magnetic permeability in vacuum).

- Cartesian coordinates are implied unless otherwise stated:

$$[x^a] := (t, x, y, z). \tag{1.5}$$

- In the context of time-varying quantities $q(t)$, q_0 denotes $q(0)$. Overdots are reserved for time derivatives with respect to coordinate time; any other time derivatives will use Leibniz's notation.

- Vectors, matrices and tensors are typically denoted by boldface letters or symbols. Their individual components are typically denoted by indices on the corresponding italicised letters or symbols. Overhats might be used to emphasise that a vector has unit norm, but are often omitted to ease notation when emphasis is unnecessary.

- Complex conjugation is denoted by a superscript $*$, and Fourier transforms are denoted by an overtilde.
- The common (base-10) logarithm is denoted by $\lg$, while the natural logarithm is denoted by $\ln$.
- The expected value of a random variable/field is denoted by $\mathbb{E}$.

1.2.2 Abbreviations

AAK	Augmented analytic kludge
AK	Analytic kludge
EMRI	Extreme-mass-ratio inspiral
EMW	Electromagnetic wave
GPR	Gaussian process regression
GW	Gravitational wave
LIGO	Laser Interferometer Gravitational-wave Observatory
LISA	Laser Interferometer Space Antenna
NK	Numerical kludge
NR	Numerical relativity
PN	Post-Newtonian
ROC	Receiver operating characteristic
SMBH	Supermassive black hole
SNR	Signal-to-noise ratio

Table 1.1: Glossary of abbreviations.

Chapter 2

Gravitational-wave astronomy

Theoretical research in the field of GW astronomy falls generally under three broad themes (although this taxonomy is neither standard nor sharply demarcated). The first and most fundamental is the study of mathematical frameworks in which the physics of GWs may be described and understood. Another theme is the usage of such frameworks in constructing waveform models for astrophysical GW sources, allowing general relativity and other theories of gravitation to be tested against observations. Finally, advanced statistical and computational methods are required for the extraction of GW signals from the noisy detector data, and the characterisation of their source properties; these two procedures are known respectively as detection and parameter estimation.

An overview of the basic approaches that underpin these three themes is provided in this chapter. It focuses on introducing context and relevant concepts for the rest of the dissertation, and is not intended to be comprehensive.

2.1 Gravitational waves in general relativity

GWs are most simply described using the linearised form of general relativity, which is a sound approximation to the full theory in the weak-field region of GW sources, i.e. where the background is near-Minkowski (flat or slightly curved) and the spacetime perturbations are small. In this section, we outline the canonical description of GWs in linearised general relativity, following the influential treatment of Misner, Thorne & Wheeler [22].

2.1.1 Linearised formalism

The metric g_{ab} of a GW-perturbed spacetime may be written as

$$g_{ab} = \eta_{ab} + h_{ab}, \tag{2.1}$$

where $[\eta_{ab}] = \mathrm{diag}(-1, 1, 1, 1)$ represents the Minkowski metric in Cartesian coordinates, and the components h_{ab} of the metric perturbation are small ($|h_{ab}| \ll 1$). We substitute (2.1) into the Einstein field equations

$$R_{ab} - \frac{1}{2} g_{ab} R = 8\pi T_{ab}, \tag{2.2}$$

and linearise in h_{ab} by neglecting terms that are $O(|h_{ab}|^2)$. In this approximation, all indices may be raised and lowered using η_{ab} instead of g_{ab}; expressions for the Ricci tensor R_{ab} and scalar curvature R (in terms of the metric components) then depend only on second partial derivatives of h_{ab}.

By further writing h_{ab} in trace-reversed form

$$h_{ab} = \bar{h}_{ab} - \frac{1}{2} \mathrm{tr}(\bar{h}) \eta_{ab}, \tag{2.3}$$

where $\mathrm{tr}(\bar{h}) := \bar{h}_a{}^a = -h_a{}^a$, and imposing the Lorenz gauge condition

$$\bar{h}^{ab}{}_{,b} = 0, \tag{2.4}$$

Equation (2.2) reduces to a Minkowski-space wave equation for $\bar{h}_{ab}$ that is sourced by the stress–energy tensor T_{ab}, i.e.

$$\Box \bar{h}_{ab} = -16\pi T_{ab}, \tag{2.5}$$

where $\Box \bar{h}_{ab} := \bar{h}_{ab,c}{}^c$ defines the flat-space d'Alembert operator $\Box$.

The wave equation (2.5) is valid as long as the gravitational field is weak, e.g. far from any GW source, or close to a nearly Newtonian source (characterised by $T_{00} \gg |T_{ab}|$ for all nonzero a or b). A similar equation applies for strong fields in full general relativity, but with a curved-space d'Alembertian and an additional effective source term $-16\pi \Delta T_{ab}$ that contains all quadratic-

and higher-order terms in h_{ab}. The general solution to (2.5) is given by the retarded integral

$$\bar{h}_{ab}(t, \mathbf{x}) = 4 \int d^3\mathbf{x}' \, \frac{T_{ab}(t - r, \mathbf{x}')}{r}, \tag{2.6}$$

where $\mathbf{x} := [x^\mu]$, and $r := |\mathbf{r}|$ with $\mathbf{r} = \mathbf{x} - \mathbf{x}'$.

Far from any GW source ($T_{ab} = 0$), the simplest vacuum solution to (2.5) is the monochromatic plane wave

$$\bar{h}_{ab} = \Re[A_{ab} \exp{(i k_c x^c)}], \tag{2.7}$$

where the four-wavevector k_a is null ($k_a k^a = 0$) and the symmetric amplitude tensor A_{ab} is transverse to k_a ($A_{ab}k^b = 0$). The four constraints $A_{ab}k^b = 0$ on the ten independent components A_{ab} arise from the Lorenz condition (2.4), which is a partial gauge fixing; the remaining gauge freedom may be used to further constrain the plane-wave solution. We choose the standard transverse–traceless gauge by imposing the three spatial conditions

$$A_{0\mu} = 0, \tag{2.8}$$

and the trace-free condition

$$\mathrm{tr}(A) = 0, \tag{2.9}$$

which leaves two dynamical degrees of freedom in A_{ab}. With $\mathrm{tr}(\bar{h}) = 0$ in (2.3) as a result of (2.9), the metric perturbation in this gauge reduces to its transverse and traceless part

$$h^{\mathrm{TT}}_{ab} := h_{ab} = \bar{h}_{ab}. \tag{2.10}$$

2.1.2 Gravitational-wave generation

For an isolated and nearly Newtonian system, the mass monopole moment is precisely the total mass–energy $M = \int d^3\mathbf{x}' \, T_{00}$, which is a conserved quantity and contributes no radiation ($\dot{M} = 0$). Likewise, the first time derivative of the mass dipole moment $m_\mu = \int d^3\mathbf{x}' \, T_{00} x'_\mu$ corresponds to the total linear momentum, which is also conserved with no associated radiation ($\ddot{m}_\mu = 0$). Hence the lowest-order radiation emitted by such a system is quadrupolar in nature, i.e. generated by a time-varying mass quadrupole moment.[2]

Any astrophysical system that spins or orbits with a degree of rotational asymmetry will emit GWs, due to the oscillation of its mass quadrupole moment. The far-field radiation from most GW sources is described to high accuracy by the well-known quadrupole formula, which is valid in the slow-motion approximation (i.e. as long as the characteristic source size is small compared to the wavelength of the GWs emitted). This formula may in principle be used to compute gravitational waveforms for any such source, provided the dynamics of its mass distribution are known.

The local mass quadrupole moment for a nearly Newtonian, slow-motion source is given by

$$I_{\mu\nu}(t') = \int d^3\mathbf{x}'\, T_{00}(t', \mathbf{x}') x'_\mu x'_\nu, \tag{2.11}$$

where the mass–energy density T_{00} may be evaluated at time t' since retardation over the extent of the source will be negligible for a distant observer, i.e. $|\mathbf{x}'| \ll |\mathbf{x}|$. It is convenient to consider the traceless part of (2.11):

$$\mathcal{I}_{\mu\nu} = I_{\mu\nu} - \frac{1}{3}\mathrm{tr}(I)\eta_{\mu\nu}, \tag{2.12}$$

which simplifies various expressions (e.g. (2.22) and (2.23)) and is known as the reduced mass quadrupole moment. In the far field of the source, we only require the traceless part of $\mathcal{I}_{\mu\nu}$ (equivalently, of $I_{\mu\nu}$) after projecting transverse to the direction of wave propagation. For a plane wave propagating along the z-axis, the nonzero components of $\mathcal{I}^{\mathrm{TT}}_{\mu\nu}$ reduce to

$$\mathcal{I}^{\mathrm{TT}}_{11} = -\mathcal{I}^{\mathrm{TT}}_{22} = \frac{1}{2}(\mathcal{I}_{11} - \mathcal{I}_{22}), \tag{2.13}$$

$$\mathcal{I}^{\mathrm{TT}}_{12} = \mathcal{I}^{\mathrm{TT}}_{21} = \mathcal{I}_{12}. \tag{2.14}$$

Finally, to obtain the quadrupole formula, we use the result

$$\int d^3\mathbf{x}'\, T_{\mu\nu} = \frac{1}{2}\partial_0^2 \int d^3\mathbf{x}'\, T_{00} x'_\mu x'_\nu, \tag{2.15}$$

[2] This argument is largely heuristic; a more rigorous derivation of the result is given in [23].

which follows from the flat-space conservation equations

$$T^{ab}{}_{,b} = 0. \tag{2.16}$$

Via (2.6), (2.12), (2.15) and the far-field approximation $r = |\mathbf{x}|$, (the spatial part of) the metric perturbation in transverse–traceless gauge is then related to the second time derivative of $\mathcal{I}^{\mathrm{TT}}_{\mu\nu}$ by

$$h^{\mathrm{TT}}_{\mu\nu}(t, \mathbf{x}) = \frac{2}{r}\ddot{\mathcal{I}}^{\mathrm{TT}}_{\mu\nu}(t - r). \tag{2.17}$$

2.1.3 Gravitational-wave propagation

Energy and momentum are carried by propagating GWs, much as they are by electromagnetic radiation. The stress–energy tensor for a GW—or indeed any gravitational field—cannot be defined locally, since the equivalence principle implies the existence of a local reference frame in which the field vanishes.

However, it is valid to define an effective GW stress–energy tensor that is smeared out over several gravitational wavelengths; this is given in linearised theory by the Isaacson tensor

$$T^{\mathrm{GW}}_{ab} = \frac{1}{32\pi}\langle h^{\mathrm{TT}}_{cd,a} h^{\mathrm{TT}}_{ef,b}\rangle \eta^{ce}\eta^{df}, \tag{2.18}$$

where the angled brackets denote an average over many wave cycles. Equation (2.18) is subject to the usual conservation of energy and momentum via (2.16). Although its reinsertion into (2.2) is uninformative in linearised theory (since $T^{\mathrm{GW}}_{ab} = O(|h_{ab}|^2) = 0$), the energy–momentum associated with the GW field in full general relativity can contribute significantly to the background curvature and result in non-negligible backscattering or tail effects.

The linearised description of GW propagation on Minkowski space is extensible to slightly curved spacetimes, where the weak curvature might be due to the GWs themselves or any nearby matter/gravitational fields. This is because the plane-wave solution (2.7) and the broader approach of geometrical "optics" remain valid in the short-wave approximation (i.e. as long as the gravitational wavelength is small compared to the background radius of curvature). On a flat or slightly curved spacetime, the stress–energy tensor (2.18) for a plane and monochromatic GW simplifies to

$$T_{ab}^{\mathrm{GW}} = \frac{1}{32\pi} A^2 k_a k_b, \tag{2.19}$$

where $A^2 := A_{ab}A^{ab}/2$. In addition, the geometrical-optics approach allows ray effects such as redshift, lensing, and Faraday rotation of the polarisation plane to be characterised for GWs in the same way as for electromagnetic radiation.

2.1.4 Gravitational-wave effects

As seen in Section 2.1.1, the field of a GW in general relativity has two dynamical degrees of freedom. These are known as the plus (h_+) and cross ($h_\times$) linear-polarisation modes; for a plane wave propagating along the z-axis, they are defined with respect to the chosen (x, y)-coordinates as

$$h_+ := h_{11}^{\mathrm{TT}} = -h_{22}^{\mathrm{TT}}, \tag{2.20}$$

$$h_\times := h_{12}^{\mathrm{TT}} = h_{21}^{\mathrm{TT}}. \tag{2.21}$$

In the field of a passing plus-polarised GW ($h_\times = 0$), the proper distance between two points with initial separation ℓ in the x-direction oscillates with h_+ as $(g_{11})^{1/2}\ell \approx (1 + h_+/2)\ell$. For two points with initial separation ℓ in the y-direction, it oscillates as $(g_{22})^{1/2}\ell \approx (1 - h_+/2)\ell$. This is an effective stretching and squeezing of space in two orthogonal directions. A cross-polarised GW ($h_+ = 0$) has a similar effect, since its metric is mapped to the plus-polarised one with $h_+ \equiv h_\times$ by a $\pi/4$ rotation in the polarisation plane.

Although a GW is technically a distortion of both space and time, the metric perturbation in the far field equals its transverse and traceless part, and so in transverse–traceless gauge the far-field effects are purely spatial as described above. The strength of a far-field GW may thus be interpreted as a dimensionless strain of size $O(|h|)$, where $h := h_+ + ih_\times$. This strain is driven by the evolution of a distant GW source via (2.17); it allows the detection of such sources through interferometry, where the two arms of an interferometer are alternately lengthened and shortened by the passage of a GW.

Another key effect of GWs in astrophysical settings is gravitational radiation reaction: energy E and angular momentum L^μ are carried away from GW sources, leading to phenomena such as spin-down in neutron stars or inspiralling in binaries. We obtain the rates of energy and angular-momentum loss (averaged over time and solid angle) for a source by integrating the respective fluxes of T^{GW}_{ab} through a surrounding sphere of radius r. Via (2.17) and the general version of (2.18) for an arbitrary gauge, these rates are related to the time derivatives of $\mathcal{I}_{\mu\nu}$ by

$$\langle\dot{E}\rangle = -\int d\Omega\, T^{\mathrm{GW}}_{01} r^2 = -\frac{1}{5}\langle\dddot{\mathcal{I}}_{\mu\nu}\dddot{\mathcal{I}}^{\mu\nu}\rangle, \tag{2.22}$$

$$\langle\dot{L}^\mu\rangle = -\int d\Omega\, \epsilon^{\mu\nu\rho} x_\nu T^{\mathrm{GW}}_{\rho 1} r^2 = -\frac{2}{5}\epsilon^{\mu\nu\rho}\langle\ddot{\mathcal{I}}_{\nu\sigma}\dddot{\mathcal{I}}_{\rho\tau}\rangle\eta^{\sigma\tau}, \tag{2.23}$$

where we have converted to spherical polar coordinates $[x^\mu] = (r, \theta, \phi)$, and $\epsilon^{\mu\nu\rho}$ is the three-dimensional Levi-Civita symbol.

2.2 Sources and waveform modelling

In this section, we provide a brief overview of GW sources (adapted from reviews such as [24–26]), with particular emphasis on the coalescing binary systems that are relevant to much of the material in this dissertation. Different strategies are used to model such binaries in the comparable-mass and extreme-mass-ratio regimes; we summarise these methods and the present status of waveform development for both cases.

2.2.1 Source overview

Gravity is many orders of magnitude weaker than the other fundamental interactions. The matter terms on the right-hand sides of (2.2) and (2.17) contain a suppressed multiplicative constant G/c^4 that is actually $\sim 10^{-45}$ in SI units. Inserting realistic scales of mass, length and time for any conceivable man-made source (e.g. the mass and orbital speed of the International Space Station) into

(2.17), we find $|h| \lesssim 10^{-32}$; such strains are impossible to detect with modern instruments, which are around 10 orders of magnitude less sensitive. Hence detectable GW sources are necessarily of astrophysical or cosmological origin, and must comprise extremely massive objects moving at near-relativistic speeds to make up for their vast distances from local detectors.

The best-understood types of GW sources are "clean" astrophysical systems that are relatively straightforward to model and identify. Chief among these are binary coalescences, which are some of the most ubiquitous sources in the sensitivity bands of ground- and space-based interferometers [27, 28], and indeed the first to be detected. These coalescences refer to binary systems in the final stage of their evolution, spanning from the late inspiral (starting at orbital separations small enough that the source is in band) to the post-merger ringdown (ending as the oscillations of the final object decay to zero). Other types of well-modelled GW sources include the early-inspiral stage of lower-mass binaries that do not merge in band, but are close enough for their weaker signals to be detected, and also "continuous" sources such as individual neutron stars that are asymmetric and rapidly spinning.

Transient GW signals that are less cleanly modelled may be classified as burst sources. These are generated by processes such as core-collapse supernovae [29], glitches in the internal dynamics of neutron stars [30], and the formation of cusps or kinks in the one-dimensional topological spacetime defects known as cosmic strings [31]. Short-duration GW signals from binary systems in particular scenarios are also referred to as bursts, e.g. the radiation emitted close to periapsis by long-period, nearly parabolic binaries in the extreme-mass-ratio regime [32], or that from late-stage binary coalescences with little or no inspiral evolution (such that most of the signal strength is attributed to the highly nonlinear merger phase).

The descriptions above refer to GW sources that are strong or close enough for their signals to be individually identified. However, the Universe is also filled with countless similar systems that are unresolvable; the superposition of signals from such sources forms a stochastic background of GWs, which is modelled as a correlated random field and searched for in the responses of independent detectors over long observing times. Like any random field, a stochastic background may be characterised by its power spectral density $S_h(f)$, which is integrated over the redshift and mass distributions of the sources under consideration. In addition to astrophysical backgrounds, stochastic backgrounds of primordial GWs can also arise through early-Universe processes such as the amplification of quantum fluctuations by inflation [33], or first-order cosmological phase transitions [34].

2.2.2 Binary-coalescence modelling

We now summarise the more general methods used to model the orbital evolution and GW emission of a coalescing binary system with component masses $(m_1, m_2 \geq m_1)$. It is useful to introduce the standard definitions of total mass M, reduced mass μ, (small) mass ratio q and symmetric mass ratio η, i.e.

$$M := m_1 + m_2, \tag{2.24}$$

$$\mu := \frac{m_1 m_2}{m_1 + m_2}, \tag{2.25}$$

$$q := \frac{m_1}{m_2}, \tag{2.26}$$

$$\eta := \frac{m_1 m_2}{(m_1 + m_2)^2} = \frac{\mu}{M} = \frac{q}{(1+q)^2}. \tag{2.27}$$

The inspiral stage of a binary coalescence is extremely well modelled at large orbital separations, and its GW signal is essentially a "chirp" that increases in frequency and amplitude. For illustrative purposes, we describe here the radiation emitted by a binary of non-spinning point masses. Via (2.17), (2.22) and (2.23), it is possible to derive expressions for the binary's GW amplitude $|h|$, as well as the evolution of its orbital frequency f and eccentricity e due to radiation reaction. At leading order in f, these are given by

$$|h| \propto \frac{\mathcal{M}^{5/3} f^{2/3}}{r}(1 + O(e^2)), \tag{2.28}$$

$$\dot{f} \propto \mathcal{M}^{5/3} f^{11/3}(1 + O(e^2)), \tag{2.29}$$

$$\dot{e} \propto -\mathcal{M}^{5/3} f^{8/3} e(1 + O(e^2)), \tag{2.30}$$

where the chirp mass $\mathcal{M}$ is defined as

$$\mathcal{M} := \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}} = M^{2/5}\mu^{3/5} = M\eta^{3/5}. \tag{2.31}$$

For a Newtonian chirp (i.e. for a binary in an instantaneous Keplerian orbit), there are no higher-order terms in f, and the right-hand sides of (2.28)–(2.30) truncate after terms that are $O(e^4)$. A closed-form expression for $e(f)$ may be obtained by solving the system of equations (2.29) and (2.30); at small eccentricities, this expression reduces to $e \propto f^{-19/18}$ [35, 36]. Equations (2.28) and (2.29) thus decouple from (2.30) for a Newtonian binary. They show that the chirp mass and distance of such a source may be estimated from its inspiral alone (since $(|h|, f, \dot{f})$ are measurable from the GW signal).

Equations (2.28)–(2.30) for a relativistic binary include higher-order terms in f, which are known as post-Newtonian (PN) corrections [37, 38]. The corresponding expansions are done in a PN parameter that is associated with a typical internal speed of the system, e.g. the average squared orbital speed v^2. Corrections that are $O(v^{2n})$ relative to leading order are referred to as nPN terms. The PN formalism is well suited to systems in which internal speeds are small, in the sense that expansions are not required to high order. In general, any GW observable may be written as a PN expansion with respect to its expression as evaluated in the Newtonian potential. For example, expansions of the GW phase include terms with different dependencies on η, which allows both masses of the binary to be estimated (rather than just the chirp mass).

For a binary that is nearly circular, Kepler's third law gives $v^2 \propto (fM)^{2/3}$. By transforming to dimensionless coordinates x^a/M, we can put (2.28)–(2.30) into a form that is explicitly dependent on v^2 and independent of M, i.e.

$$|h| \propto \frac{\eta v^2}{\bar{r}}(1 + O(e^2)), \tag{2.32}$$

$$\frac{d\bar{f}}{d\bar{t}} = M^2\dot{f} \propto \eta v^{11}(1 + O(e^2)), \tag{2.33}$$

$$\frac{de}{d\bar{t}} = M\dot{e} \propto -\eta v^8 e(1 + O(e^2)), \tag{2.34}$$

where the overbars denote dimensionless quantities. These equations illustrate the total-mass invariance of waveforms in dimensionless coordinates, which is often exploited in the modelling of GW sources.

The merger stage of a binary coalescence occurs deep within the nonlinear dynamical strong field, at which point PN methods become both impractical and inaccurate. Numerical relativity (NR) methods [39] (i.e. direct numerical integration of the Einstein field equations) are typically used to model this stage. The radiative degrees of freedom in a gravitational field are encoded in the Weyl tensor C_{abcd} (the fully traceless part of the Riemann tensor), which is reduced in the Newman–Penrose formalism [40] to five scalar contractions Ψ_n of C_{abcd} with a complex null tetrad. For waveform modelling, Ψ_4 is the most relevant as it describes outgoing radiation at null infinity; it is related to the metric perturbation by

$$\Psi_4 = -\ddot{h}_+ + i\ddot{h}_\times \tag{2.35}$$

in the limit as $r \to \infty$.

In NR simulations, Ψ_4 is expanded in terms of the spin-weighted spherical harmonics ${}_sY_{lm}$ [41] with $s = -2$, i.e.

$$\Psi_4(t, r, \theta, \phi) = \sum_{l=2}^{\infty} \sum_{m=-l}^{l} \Psi_{4,lm}(t, r)\, {}_{-2}Y_{lm}(\theta, \phi). \tag{2.36}$$

This facilitates computation (to some desired l) of the modes $\Psi_{4,lm}$, which are then extracted on a sphere of finite radius. Breakthroughs in NR during the mid-2000s [42–44] have enabled the highly accurate evolution of binary coalescences from several orbits before plunge (the start of the merger stage) right through to ringdown. However, NR waveforms are expensive to compute, and are limited at present to $\lesssim 100$ cycles—typically with small initial orbital separations ($\lesssim 20M$) and mass ratios close to unity ($\gtrsim 1/20$) [45].

The end product of a binary coalescence is a vibrating black hole or relativistic star that gradually settles into a stable configuration. Although this ringdown stage may be modelled within NR, its waveforms are described more simply through black-hole perturbation theory as a superposition of quasinormal modes [46], i.e. exponentially decaying sinusoids. The framework for studying black-hole perturbations is built upon early calculations by Regge & Wheeler [47] and Zerilli [48]; for a rotating Kerr black hole, it is centred around the Teukolsky master equation [49] describing the linear perturbations of various scalar fields ψ on the spacetime.

In the spherical-like Boyer–Lindquist coordinates, the Kerr line element for the spacetime around a mass M with angular momentum Ma is written as

$$ds^2 = -\frac{\Delta}{\Sigma}(dt - a\sin^2\theta\, d\phi)^2 + \frac{\sin^2\theta}{\Sigma}((r^2+a^2)d\phi - a\, dt)^2 + \frac{\Sigma}{\Delta}dr^2 + \Sigma\, d\theta^2, \quad (2.37)$$

where $\Delta = r^2 - 2Mr + a^2$ and $\Sigma = r^2 + a^2\cos^2\theta$. The Teukolsky equation takes the general form

$$ {}_s\Theta\psi = 4\pi\Sigma T, \quad (2.38)$$

where T is the source term (written in Newman–Penrose form), and ${}_s\Theta$ is the Teukolsky operator for a field of spin weight s:

$$\begin{aligned} {}_s\Theta = &-\left(\frac{(r^2+a^2)^2}{\Delta} - a^2\sin^2\theta\right)\partial_t^2 - \frac{4Mar}{\Delta}\partial_t\partial_\phi - \left(\frac{a^2}{\Delta} - \frac{1}{\sin^2\theta}\right)\partial_\phi^2 \\ &+\Delta^{-s}\partial_r\Delta^{s+1}\partial_r + \frac{1}{\sin\theta}\partial_\theta\sin\theta\,\partial_\theta + 2s\left(\frac{a(r-M)}{\Delta} + \frac{i\cos\theta}{\sin^2\theta}\right)\partial_\phi \\ &+2s\left(\frac{M(r^2-a^2)}{\Delta} - r - ia\cos\theta\right)\partial_t - s^2\cot^2\theta + s. \end{aligned} \quad (2.39)$$

For outgoing gravitational radiation, we set $s = -2$ and $\psi = \rho^{-4}\Psi_4$ with $\rho = 1/(r - ia\cos\theta)$; the Teukolsky equation is then separable into its radial and angular components via the expansion (2.36), such that the Kerr quasinormal modes (i.e. the eigenfunctions of (2.38) with $T = 0$) and in particular their complex frequencies may be obtained analytically (e.g. [50]).

2.2.3 Comparable-mass mergers

Although the three stages of a binary coalescence have very different practical descriptions (the two-body problem at large and small separations, then one-body perturbations), there are regimes of overlap in which the computed waveforms can be cross-validated and stitched together smoothly. Approaches such as those described above may be used to construct inspiral–merger–ringdown waveform models for comparable-mass binaries with $q \gtrsim 10^{-1}$. These models are applicable to binary systems of stellar-mass compact objects—white dwarfs, neutron stars or stellar-origin ($\lesssim 10^2\, M_\odot$) black holes—as well as SMBH binaries, through a trivial rescaling by total mass.

The best-developed comparable-mass models incorporate at some level the spins of both objects in the binary, whether these are assumed to be aligned with the orbital angular momentum, or generically aligned and hence precessing. One powerful framework for achieving this is the effective-one-body formalism [51], in which the relative motion of two spinning masses $((m_1, \mathbf{S}_1), (m_2, \mathbf{S}_2))$ is mapped to the geodesic motion of an effective particle $(\mu, \mathbf{S}_\mu)$ on a deformed Kerr spacetime $(M, \mathbf{S}_M)$. The effective-one-body approach has been used to construct a succession of comparable-mass models, the latest of which includes precessing spins [52]. These models typically produce inspiral–merger waveforms that are smoothly connected to quasinormal modes, with NR simulations employed to calibrate unknown coefficients and free parameters in the merger–ringdown.

While most of the methods and models described so far provide time-domain waveforms $h(t)$, it is often desirable to directly compute $\tilde{h}(f)$ for ease of use in data analysis algorithms. This is done in the other well-known family of models, dubbed Phenom [53], which use NR-fitted phenomenological parameters to write their waveforms as closed-form analytic expressions in the frequency domain. Phenom waveforms rival those from the effective-one-body approach in accuracy, and have also been extended to account for the effects of spin precession [54].

2.2.4 Extreme-mass-ratio inspirals

The standard strategies adopted in the modelling of comparable-mass binary coalescences are not well suited to the regime where $q \ll 1$, with the possible exception of the effective-one-body approach (although work in this area is presently limited to circular and equatorial orbits, and requires calibration against more accurate models [55]). This is because the inspiral of the smaller object includes many orbits in the strong field of the larger object, and so the PN treatment is inadequate on its own, while NR simulations are hampered by the greatly differing scales of mass, length and time in the system.

Extreme-mass-ratio inspirals (EMRIs) involve a compact object in orbit around a rotating SMBH, with $q \lesssim 10^{-4}$ such that $(m_1, m_2, \eta) \approx (\mu, M, q)$. They are expected to be an important and detectable source of GWs at millihertz frequencies [56]. The most readily available EMRI waveforms are generated using mixed-formalism models known as "kludges" [57–59], which are characterised by their computational efficiency and modular construction. In these models, evolution of the orbit over long timescales is based on PN expressions in the limit of (non-spinning) test-particle motion on Kerr geometry; the modular setup then allows the introduction of relativistic orbital corrections on both short and long timescales. Flat-space multipole formulae (e.g. (2.17)) sourced by the compact object are used for fast waveform generation to a reasonable approximation, even though the background spacetime is strongly curved.

Fuller EMRI treatments require black-hole perturbation theory and the computation of the gravitational self-force on the compact object [60, 61]. At first order in q, the gravitational field of a test particle on Kerr spacetime is a linear perturbation of the background metric, i.e. $g_{ab} \to g_{ab} + qh_{ab}$. Motion on the perturbed spacetime depends only on some regular part h^{R}_{ab} of the metric perturbation, and may be interpreted either as geodesic motion on $g_{ab} + qh^{\mathrm{R}}_{ab}$, or as Kerr geodesic motion that is altered by an effective external force. This self-force is written as

$$F^a = \mu \nabla^{abc} h^{\mathrm{R}}_{bc}, \tag{2.40}$$

with the force operator ∇^{abc} given in terms of the Kerr covariant derivative by

$$\nabla^{abc} = \frac{1}{2}(g^{ad}u^b - 2g^{ab}u^d - u^a u^b u^d)u^c \nabla_d, \tag{2.41}$$

where u^a is the test particle's four-velocity. The adiabatic (time-averaged), dissipative effects of the self-force at first order in q are encoded in solutions of the Teukolsky equation (2.38) with compact-object source; while these are expensive to compute, they are more accurate than kludges and have been used to generate full EMRI waveforms for eccentric and inclined Kerr orbits [62, 63].

Although Teukolsky-based waveforms are potentially sufficient for detection purposes, they suffer from a build-up of error in the orbital phase over the long inspiral lifetime. To accurately recover the orbital parameters of EMRIs, waveform models must include the next dominant self-force phenomenon of resonances, where the concurrence of the various orbital frequencies causes transient jumps in the evolution of the compact object's orbit [64]. They also need to account for higher-order phase effects due to the conservative first-order self-force and the dissipative second-order self-force [65]; these calculations are the aim of the ongoing self-force programme. Full self-force models might be too computationally intensive to directly provide waveforms for data analysis, and hence will likely be used to inform kludges instead [66–68].

The actual merger and ringdown of an extreme-mass-ratio coalescence are not as important from a data analysis standpoint, since most of the source information will typically be contained in the extensive inspiral stage. Regardless, the merger is now a relatively straightforward transition between the last stable orbit of the compact object and its geodesic plunge [69], while the ringdown is again described by the quasinormal modes of the final Kerr black hole.

Finally, coalescing binary systems with $10^{-3} \lesssim q \lesssim 10^{-2}$ are known as intermediate-mass-ratio inspirals. Less work has been done on constructing waveform models for such sources; this is due to the scarcity of direct observational evidence for intermediate-mass black holes, as well as the limitations of NR and perturbation theory at intermediate mass ratios. However, several hybrid methods have been applied to the modelling of these systems [70, 71], and the implied existence of $\sim 10^2\,M_\odot$ black holes in light of the first LIGO detection might presage their eventual discovery.

2.3 Detectors and data analysis

We now describe some of the existing and proposed GW detectors introduced in Chapter 1, before outlining the signal processing framework used to analyse their data (with particular emphasis on the standard approaches to detection and parameter estimation for laser interferometers). The material in this section has been adapted from reviews such as [25, 26, 72, 73].

2.3.1 Detector overview

The radiation spectrum of likely sources in GW astronomy will be spanned by an envisioned network of man-made and naturally occurring interferometric detectors with different arm lengths and noise backgrounds. These fall into three main categories: ground-based detectors (which are sensitive at kilohertz or "high" frequencies), space-based detectors (at millihertz or "low" frequencies), and pulsar timing arrays (at nanohertz or "very low" frequencies).

Ground-based detectors will observe stellar-mass binary coalescences, asymmetric neutron stars, supernovae and other low-mass GW sources. Modern Michelson interferometers such as Advanced LIGO [7] and Advanced Virgo [11] have arms that are $\sim 10^3\,\mathrm{m}$ long, and employ Fabry–Pérot cavities to increase the effective arm length to $\sim 10^5\,\mathrm{m}$, hence improving sensitivity at sub-kilohertz frequencies. They also use vibration isolation systems and materials with low mechanical loss to screen out instrumental noise, but environmental changes in the local gravitational field (e.g. seismic vibrations, or atmospheric pressure gradients) are an irreducible source of gravity gradient noise at $\lesssim 10^1\,\mathrm{Hz}$. Future variants such as KAGRA [12] and the Einstein Telescope [74] will explore the same frequency band with additional noise reduction strategies, including underground construction and cryogenic cooling.

Gravity gradient noise falls off rapidly away from the Earth, and so GWs at lower frequencies can be detected by an interferometer in space. Space-based detectors will observe Galactic white-dwarf binaries in the early-inspiral stage, and more distant sources such as EMRIs and the binary coalescences of smaller ($\lesssim 10^7\,M_\odot$) SMBHs. The archetypal design for a space interferometer is that of LISA [13]: three freely flying test masses housed in drag-free spacecraft, separated by $\sim 10^9\,\mathrm{m}$, and placed in an Earth-trailing orbit around the Sun. LISA is most sensitive in the millihertz band, below which it is limited by noise from long-timescale drifts in the relative acceleration of the test masses. At decihertz frequencies, LISA's arm length exceeds multiple wavelengths and its sensitivity is reduced due to partial signal cancellation; however, future detectors with shorter test-mass separations (e.g. DECIGO [15], the geocentric TianQin [16], and other proposed missions [75]) will extend the range of space-based GW observations up to the lower frequency limits of ground interferometers.

Arrays of stable millisecond pulsars in the Milky Way emit radio pulses that arrive at Earth with exquisite regularity in time, thus effectively acting as a network of natural "interferometers" with Galactic-scale ($\sim 10^{20}\,\mathrm{m}$) arm lengths. A GW that passes through the Galaxy will perturb the paths taken by the pulses, and hence their times of arrival, in a correlated way; this effect is then distinguishable from the noise in pulse arrival times due to intrinsic pulsar processes, which are uncorrelated. The peak sensitivity of a pulsar timing array occurs near its lower frequency limit, i.e. the cut-off determined by the total observation time. For a five-year baseline, this corresponds to the nanohertz frequency band. Pulsar timing arrays such as the IPTA [19] and a future programme supported by the Square Kilometre Array [76] will observe inspiralling SMBH binaries at the higher end of their mass range, as well as the stochastic GW background of signals from such sources.

Apart from interferometric detectors, there is another broad class of detector that works by amplifying the vibrations induced in a material object due to the passage of a GW. Such instruments are known as resonant detectors [77, 78]. Their sensitivities have gradually been surpassed by those of laser interferometers; furthermore, they typically operate in a narrow high-frequency band ($\sim 10^3\,\mathrm{Hz}$) where there are fewer anticipated astrophysical or cosmological sources, and so are less relevant to the content of this dissertation.

2.3.2 Signal processing

When analysing data from laser interferometers, the two polarisation modes $h_{+,\times}$ are not measured directly. The intermediate step involves at least two independent detector-response functions $h_{I,II}$, which are orthogonal in the sense that the noise in both channels is uncorrelated. These two data streams are obtained from either separate interferometers or noise-correlated ones (e.g. in shared-arm configurations such as LISA); as long as the noise is not perfectly correlated, $h_{I,II}$ can always be orthogonalised in the usual way.

Working in arbitrary coordinates, we denote the orthonormal coordinate frame adopted in (2.20) and (2.21) as $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \equiv (\mathbf{p}, \mathbf{q}, \mathbf{r})$ (with $\mathbf{r}$ pointing from the source to the detector), and define the polarisation basis tensors

$$H^{+}_{\mu\nu} := p_\mu p_\nu - q_\mu q_\nu, \tag{2.42}$$

$$H^{\times}_{\mu\nu} := p_\mu q_\nu + q_\mu p_\nu, \tag{2.43}$$

such that the far-field metric perturbation may be written as

$$h^{\mathrm{TT}}_{\mu\nu} = h_+ H^{+}_{\mu\nu} + h_\times H^{\times}_{\mu\nu}. \tag{2.44}$$

For a single detector whose two arms are aligned with the unit vectors $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$, the strain signal caused by the passage of a GW is given by

$$h_I = \frac{1}{2} h^{\mathrm{TT}}_{\mu\nu} (\ell_1^\mu \ell_1^\nu - \ell_2^\mu \ell_2^\nu). \tag{2.45}$$

This response function is proportional to the sine of the angle α between the arms [79], and is explicitly related to $h_{+,\times}$ via (2.44):

$$h_I = (h_+ F_I^+ + h_\times F_I^\times) \sin\alpha, \tag{2.46}$$

$$F_I^{+,\times} \sin\alpha = \frac{1}{2} H^{+,\times}_{\mu\nu} (l_1^\mu l_1^\nu - l_2^\mu l_2^\nu), \tag{2.47}$$

where the antenna pattern functions $F_I^{+,\times}(\theta_S, \phi_S, \psi_S)$ depend on the sky location (θ_S, ϕ_S) and polarisation angle ψ_S of the source [80]. A second channel h_{II} is required for the recovery of both modes $h_{+,\times}$, and is similarly defined as (the orthogonal part of) the strain signal in another linearly independent detector.

GW strain data from laser interferometers is often characterised by a low instantaneous signal-to-noise ratio (SNR), due to the high levels of instrumental and environmental noise. However, the data may be processed with template filters generated from waveform models, which allows the build-up of SNR over time. In the standard signal-processing framework for GW astronomy, data from a pair of orthogonal detectors may be written as the complex time series[3] $x(t) = h(t) + n(t)$, where the source signal $h = h_I + ih_{II}$ is a de-

[3] Another common and more general approach is to consider the real time-series data from individual detectors; the SNRs for multiple orthogonal detectors may then be summed in quadrature to give a combined SNR for the detector network.

terministic function of time (and some unknown source parameters), and the detector noise $n = n_I + in_{II}$ in both channels is assumed to be a stationary and zero-mean stochastic process.

The process known as matched filtering involves the cross-correlation of x with some optimal template filter $F(t)$ that maximises the output SNR

$$\rho := \frac{\mathbb{E}[(x|F)]}{\sqrt{\mathbb{E}[(n|F)(n|F)]}}, \tag{2.48}$$

where the (zero-lag) cross-correlation inner product $(\cdot|\cdot)$ on the function space of finite-length time series is defined as

$$(a|b) := \int dt\, a(t)b^*(t). \tag{2.49}$$

Since n is stationary, its autocorrelation function $R_n(t,t') := \mathbb{E}[n(t)n^*(t')]$ may be written as an even function of one variable, i.e.

$$R_n(\tau) = \mathbb{E}[n(t)n^*(t-\tau)]. \tag{2.50}$$

By the Wiener–Khinchin theorem, the (two-sided) noise power spectral density $\mathcal{S}_n(f)$ is simply the Fourier transform of $R_n(\tau)$:

$$\mathcal{S}_n(f) := \int d\tau\, R_n(\tau)\exp\left(-2\pi i f\tau\right). \tag{2.51}$$

The solution to the variational problem $\delta\rho/\delta F = 0$ is then $h(t) \propto (R_n * F)(t)$, which by the convolution theorem gives

$$\tilde{F}(f) \propto \frac{\tilde{h}(f)}{\mathcal{S}_n(f)}, \tag{2.52}$$

and so the matched filter is proportional to the noise-weighted signal itself.

It is convenient to shift the noise weighting into the inner product (2.49); the resultant noise-weighted inner product $\langle\cdot|\cdot\rangle$ is real and defined as

$$\langle a|b\rangle := (a|F_b) = \int dt\, a(t)F_b^*(t) = \int df\, \frac{\tilde{a}(f)\tilde{b}^*(f)}{\mathcal{S}_n(f)} \tag{2.53}$$

via the Plancherel theorem, with F_b denoting the matched filter corresponding to the waveform template b (i.e. $\tilde{F}_b := \tilde{b}/\mathcal{S}_n$).

With a filter that is matched to the source signal h, the output SNR (2.48) simplifies to

$$\rho = \frac{\mathbb{E}[\langle x|h\rangle]}{\sqrt{\mathbb{E}[\langle n|h\rangle\langle n|h\rangle]}} = \sqrt{\langle h|h\rangle}, \tag{2.54}$$

where we have used the noise identities

$$\mathbb{E}[\langle n|h\rangle] = 0, \tag{2.55}$$

$$\mathbb{E}[\langle n|h\rangle\langle n|h'\rangle] = \langle h|h'\rangle. \tag{2.56}$$

Equation (2.54) may be interpreted as the "true" SNR of the source. If the filter is matched not to h but some other waveform h', the output SNR is instead

$$\rho' = \frac{\langle h|h'\rangle}{\sqrt{\langle h'|h'\rangle}}, \tag{2.57}$$

which is $\leq \rho$ by the Cauchy–Schwarz inequality.

The common principle of comparing detector data against many templates and identifying the best-matching one is used in both detection and parameter estimation. Since the SNR might be reduced significantly if the identified template is not close to optimal, the generating waveform model should be as accurate as possible. At the same time, it must also be computationally efficient due to the large number of templates required in both procedures. Different strategies apply for detection, which can be viewed as a global search on the SNR surface over the model parameter space, and parameter estimation, which is an exploration (over a more localised region) of the probability that the template-subtracted data is consistent with detector noise.

2.3.3 Detection

To discern the presence of a signal buried in detector noise, one standard approach is to perform a targeted search by filtering the data with a template bank, i.e. a discrete set of precomputed (or rapidly computed) unit-SNR templates that densely span an extended region of the model parameter space. This may be done while the detector is online, in order to identify high-SNR candidates for prompt follow-up. A detection is then claimed if the output SNR for any template in the bank surpasses a specified significance threshold that is determined by the properties of the concurrent noise background.

Various prescriptions for template placement may be used to ensure proper coverage of the search region. The most straightforward is to grid the templates on a lattice that is uniform with respect to some parameter-space metric. Such a metric is often defined by way of the overlap $\mathcal{O}(\cdot|\cdot)$ between two waveforms, which is given by

$$\mathcal{O}(a|b) := \frac{\langle a|b\rangle}{\sqrt{\langle a|a\rangle\langle b|b\rangle}}. \tag{2.58}$$

This bilinear form takes the value of ± 1 for linearly dependent waveforms and 0 for orthogonal waveforms, and hence provides a measure of the accuracy with which one waveform represents another.

The overlap is related to several other commonly used quantities in the GW literature. For example, the maximal overlap between the source signal and a (discrete or continuous) set of templates from a waveform model is known as the “fitting factor” of that set [81]. The “match” between two waveforms refers to their overlap maximised over some chosen subset of parameters; these are typically extrinsic to the source and more rapidly searched over [82]. We also have the “mismatch” between neighbouring templates in a bank, which is just their match subtracted from unity, as well as the “minimal match” for a bank with uniform mismatch, which is the fraction of optimal SNR it retains when the signal falls maximally far from the nearest templates [83].

Template banks are well suited to detection searches for simply modelled signals across a low-dimensional parameter space, but their computational efficiency scales poorly with waveform cost and search volume. Banks of $\sim 10^5$ approximate templates for stellar-mass binary coalescences are employed in low-latency LIGO pipelines to search a reduced four-dimensional parameter space (the component masses and aligned-spin magnitudes) for coincident signals in the two detectors. The significance of a candidate signal is assessed against the SNR distribution of false coincident signals due to noise in either data stream. A frequentist approach is used to estimate this distribution: the timestamps of one data stream are artificially shifted many times, with the offsets being large enough that all coincident signals in the time-shifted data sets are necessarily uncorrelated between detectors (and interpreted as noise).

Apart from targeted searches, the LIGO pipelines also incorporate generic searches that are performed without the use of waveform models. These generally involve the identification of coincident excess power in spectrograms of the data, i.e. time–frequency plots of the short-time power spectral density; they provide a complementary search for (stronger) signals from well-modelled sources, and are also suitable for detecting the transient signals from burst sources. As the field of GW astronomy matures, the standard detection algorithms based on template banks and spectrograms will likely be improved or supplanted by modern statistical techniques such as deep neural networks [84] for targeted searches, or compressed sensing [85] for generic ones.

2.3.4 Parameter estimation

Detection searches are performed by maximising the SNR-related quantity $\langle x|h\rangle$ over the parameter space of the waveform model h; in parameter estimation, it is the noise-related quantity $\langle x-h|x-h\rangle$ that is minimised instead. The probability that $x-h$ is consistent with Gaussian detector noise supplies a likelihood function for the model parameters $\boldsymbol{\lambda}$ given the data, i.e.

$$L(\boldsymbol{\lambda}|x) \equiv p(x|\boldsymbol{\lambda}) \propto \exp\left(-\frac{1}{2}\langle x-h(\boldsymbol{\lambda})|x-h(\boldsymbol{\lambda})\rangle\right). \tag{2.59}$$

This allows the optimisation problem to be cast in the framework of Bayesian statistics, such that confidence regions around the optimal parameter point may be obtained from the posterior probability density

$$p(\boldsymbol{\lambda}|x) = \frac{L(\boldsymbol{\lambda}|x)\pi(\boldsymbol{\lambda})}{Z(x)}. \tag{2.60}$$

The parameter prior π is specified as appropriate or taken to be uninformative, while the model evidence $Z = \int d\boldsymbol{\lambda}\, L\pi$ is a normalising factor that is unused in parameter estimation. Although Z is expensive to evaluate explicitly, it can be used for the Bayes-factor comparison of different waveform models, and also provides a Bayesian approach to the detection of weaker signals via a comparison with the null-model evidence $Z_0 \propto \exp\left(-1/2\langle x|x\rangle\right)$.

Grid-based searches of the posterior surface with template banks are too computationally inefficient to estimate parameters at the precision admitted by most waveform models. The measurement precision for a model $h(\boldsymbol{\lambda})$ at some point in parameter space is determined by its waveform derivatives $\partial h/\partial\boldsymbol{\lambda}$ at that point. In the case of high SNR, the parameter estimation errors $\Delta\boldsymbol{\lambda}$ due to Gaussian noise have the normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}^{-1})$, where the Fisher information matrix $\boldsymbol{\Gamma}$ is given by [86, 87]

$$\Gamma_{ij} = \left\langle \frac{\partial h}{\partial\lambda_i} \Big| \frac{\partial h}{\partial\lambda_j} \right\rangle. \tag{2.61}$$

The root-mean-square errors in the general case can then be approximated as

$$\Delta\lambda_i \approx \sqrt{(\Gamma^{-1})_{ii}}, \tag{2.62}$$

which places stringent and typically impractical upper limits on the grid spacing of a template bank that is put to this purpose.

Parameter estimation algorithms thus make use of stochastic search methods instead, in order to map out (relevant parts of) the posterior surface at higher resolution. The unnormalised posterior density $L\pi$ in (2.60) is sampled over a region of parameter space containing the detected signal; this is usually localised beforehand by the initial detection algorithms, although stochastic methods can in principle search extended regions and adjust their sampling resolution adaptively. Estimates of the source parameters and their associated errors are obtained directly from the distribution of samples, after it has converged to the underlying posterior density.

The most successful class of stochastic sampling methods for GW parameter estimation are the Markov-chain-Monte-Carlo algorithms, which have seen extensive use due to their simplicity and versatility. These typically involve sampling with a random walk across parameter space, but directed by a probability density that proposes points at each step, as well as some criterion for accepting or rejecting points. The speed and reliability of these algorithms can also be improved through techniques such as simulated annealing [88] and parallel tempering [89]. Other well-known methods that have been applied to GW parameter estimation include nested sampling [90], which explores nested contours of increasing probability with a number of live points, and evolutionary algorithms [91], which use genetic concepts such as breeding and fitness to evolve a population of points across parameter space.

Chapter 3

Einstein–Maxwell interactions

GWs from astrophysical sources can interact with background electromagnetic fields, giving rise to distinctive and potentially detectable electromagnetic signatures. In this chapter, we study such interactions for far-field gravitational radiation using the 1+3 approach to relativity. Linearised equations for the Maxwell field on perturbed Minkowski space are derived and solved analytically. The inverse Gertsenshteĭn conversion of GWs in a static electromagnetic field is rederived, and the resultant electromagnetic radiation is shown to be significant for highly magnetised neutron stars in compact binary systems.

We also obtain a variety of nonlinear interference effects for interacting gravitational and electromagnetic plane waves, although wave–wave resonances previously described in the literature are absent when the electric–magnetic self-interaction is taken into account. The fluctuation and amplification of electromagnetic energy flux as the strength of the GW increases towards the gravitational–electromagnetic frequency ratio is a possible signature of gravitational radiation from extended astrophysical sources.

The material in this chapter has been adapted from [92].

3.1 Background

The strongest GW sources for present and future detectors are highly energetic astrophysical events, many of which will be accompanied by distinctive and detectable electromagnetic signals. Such electromagnetic counterparts can aid the detection of GW sources through improved event rate prediction, the provision of search triggers, the confirmation of individual detections, and enhanced parameter estimation. Once an operational network of GW observatories is fully realised, the synergy of complementary information from gravitational and electromagnetic observations will establish GWs as an important component of multimessenger astronomy [93–95].

As observations of GW sources and their electromagnetic counterparts improve in precision, so too must models of such dual sources, to account for any correlations between the two types of signal. A nascent line of research towards this end is the direct coupling between gravitational and electromagnetic fields in the strong-field regime. Recent work in this area has focused on the electromagnetic signatures of gravitational perturbations on various curved spacetimes; the perturbed Einstein–Maxwell equations have been solved for Schwarzschild [96], slowly rotating Kerr–Newman [97] and equal-mass binary Kerr [98], with the numerical involvement increasing as per the complexity of the spacetime.

The problem of Einstein–Maxwell coupling for gravitational radiation in flat space is older and more analytically tractable than that in curved space, leading to a better characterisation of the (albeit weaker) interactions between far-field GWs and electromagnetic fields. One such effect is the resonant conversion of a GW into an electromagnetic wave (EMW)—and vice versa—in the presence of a static electromagnetic field [99–101]. The direct signatures of GWs on EMWs have also been studied; these include frequency splitting [102], intensity fluctuations [102, 103], deflection of rays [103–105] and gravitationally induced Faraday rotation of the EMW polarisation [105–110]. Indirect GW detection schemes using microlensing [111] and phase modulation [112] effects on light have been proposed as well.

Among various frameworks suited to the study of interacting GWs and electromagnetic fields is the 1+3 covariant approach to general relativity, in which spacetime is locally split into time and space via the introduction of a fundamental timelike congruence [113–115]. This approach is most commonly employed in the cosmological setting, and in particular has been used to describe electromagnetic signatures of the large-scale tensor perturbations associated with primordial GWs [116–120]. It may also be applied to gravitational–electromagnetic interactions in a general spacetime [121], although any inhomogeneity in the spacetime typically renders the governing equations intractable due to tensor–vector and tensor–tensor coupling [122].

Such difficulties with the 1+3 formalism may be partially overcome by extending the spacetime splitting to a 1+1+2 decomposition in the case of locally rotationally symmetric spacetimes, which have a preferred spatial direction [122]; this method has been used to semi-analytically model the electromagnetic signature of a Schwarzschild ringdown [123]. The 1+3 approach has also been applied to the interaction of far-field GWs and electromagnetic fields in the presence of a magnetised plasma [124, 125]. Finally, recent work on Minkowski-space GWs and EMWs within the 1+3 framework has uncovered resonant interactions between the two under specific conditions [120, 126].

As any resonant amplification of electromagnetic fields by gravitational radiation might be important for GW detection, we take a more detailed look at flat-space interactions between GWs and electromagnetic fields using the 1+3 approach to relativity. We provide in Section 3.2 a brief primer to the 1+3 formalism, which is seldom encountered in the context of GW astronomy. In Section 3.3, we derive linearised evolution and constraint equations for the electromagnetic field on GW-perturbed flat space, and approximate these on exact Minkowski space. This framework is applied to simple models of static and radiative electromagnetic fields in Section 3.4, where we consider the resultant effects in astrophysical settings and discuss their implications for dual observations of GW sources.

We rederive in Section 3.4.1 the resonant induction of an EMW by a GW in a static electromagnetic field, and estimate that for highly magnetised neutron stars in compact binary systems, the energy radiated through this process might be non-negligible with respect to the magnetic dipole radiation. In Section 3.4.2, we find no resonant interaction between plane GWs and EMWs after considering electric–magnetic self-interaction contributions that have been omitted in previous work [120, 126]. However, nonlinear interference effects are shown to be significant in a regime where the GW strength approaches the GW–EMW frequency ratio from below; the resultant fluctuation and amplification of electromagnetic energy flux is a potentially stronger signature of gravitational radiation than other geometrical-optics effects in the literature.

3.2 The 1+3 approach to relativity

The 1+3 covariant formalism [113–115] has been applied to many problems in general relativity. In the context of gravitational perturbations, the approach has limited benefits due to the generally intractable evolution equations that arise, although it is fairly well-suited to the analytical study of GWs in spatially homogeneous spacetimes [122].

For a general spacetime, we define a local 1+3 threading into time and space by introducing a fundamental timelike congruence of worldlines on the manifold. The timelike vector field tangential to this congruence is given by

$$u^a := \frac{dx^a}{d\tau}, \tag{3.1}$$

where τ is proper time along the worldlines and $u_a u^a = -1$. In the absence of vorticity, such a threading foliates the spacetime into spacelike hypersurfaces orthogonal to u^a. Every tensor field $\mathcal{T}$ on the spacetime may then be decomposed into its timelike and spacelike parts via contraction with, respectively, (3.1) and the spatial projection tensor[4]

$$P_{ab} := g_{ab} + u_a u_b. \tag{3.2}$$

Decomposing the covariant derivative $\mathcal{T}_{;a}$ in the above way defines the covariant time ($\dot{\mathcal{T}}$) and space ($\mathcal{T}_{:a}$) derivatives of $\mathcal{T}$. If $\mathcal{T}$ is a spatially projected tensor, we may write

$$\dot{\mathcal{T}} = \mathcal{T}_{;a} u^a, \tag{3.3}$$

$$\mathcal{T}_{:a} = \mathcal{T}_{;a} + \dot{\mathcal{T}} u_a. \tag{3.4}$$

Following the conventions of [114], we also define the spacetime volume form $\epsilon_{abcd} = \epsilon_{[abcd]}$ as the fully antisymmetric pseudotensor with $\epsilon_{0123} := |\det[g_{ab}]|^{1/2}$, and its projection onto each instantaneous rest space as $\epsilon_{abc} := u^d \epsilon_{dabc}$ (such that the spatial curl is right-handed).

For a general fluid solution to the Einstein field equations (2.2), the stress–

[4] The usual 1+3 notation for the projection tensor is h_{ab}, by unfortunate coincidence.

energy tensor decomposition with respect to (3.1) is given by

$$T_{ab} = \mu u_a u_b + pP_{ab} + 2q_{(a}u_{b)} + \pi_{ab}, \tag{3.5}$$

where μ is the mass–energy density, p is the isotropic pressure, q_a is the energy flux vector and π_{ab} is the symmetric and traceless anisotropic stress tensor. The matter field fully determines the local Ricci curvature, but not the nonlocal curvature encoded in the Weyl tensor C_{abcd}, which is obtained from the Riemann tensor R_{abcd} by subtracting its various trace terms:

$$C_{abcd} = R_{abcd} - (g_{a[c}R_{d]b} - g_{b[c}R_{d]a} - \frac{1}{3}g_{a[c}g_{d]b}R). \tag{3.6}$$

By analogy with electromagnetism, the Weyl tensor splits into its "electric" and "magnetic" parts; these are given respectively by

$$E_{ab} = C_{acbd}u^c u^d, \tag{3.7}$$

$$H_{ab} = \frac{1}{2}\epsilon_a{}^{cd}C_{cdbe}u^e. \tag{3.8}$$

Both matter and geometry affect motion along the worldlines, which is characterised by the kinematical quantities in the decomposition

$$u_{a;b} = \frac{1}{3}\vartheta P_{ab} + \sigma_{ab} + \omega_{ab} - \dot{u}_a u_b. \tag{3.9}$$

The vorticity tensor $\omega_{ab} = u_{[a:b]}$ and the (non-gravitational) acceleration vector $\dot{u}_a$ vanish in GW-perturbed Minkowski space, and are henceforth taken to be zero. With this simplification, the evolution of the expansion scalar $\vartheta = u_a{}^{:a}$ is governed by the Raychaudhuri equation

$$\dot{\vartheta} = -\frac{1}{3}\vartheta^2 - 2\sigma^2 - \frac{1}{2}(\mu + 3p), \tag{3.10}$$

where $\sigma^2 := \sigma_{ab}\sigma^{ab}/2$. The shear tensor σ_{ab} is the symmetric and traceless part of $u_{(a:b)}$, and evolves (keeping all quantities symmetric and traceless) as

$$\dot{\sigma}_{ab} = -\frac{2}{3}\vartheta\sigma_{ab} - \sigma_{ac}\sigma_b{}^c - E_{ab} + \frac{1}{2}\pi_{ab}. \tag{3.11}$$

Equations (3.10) and (3.11) are accompanied by the two spatial constraints

$$\mathrm{div}\sigma_{ab} := \sigma_{ab}{}^{:b} = \frac{2}{3}\vartheta_{:a} - q_a, \tag{3.12}$$

$$\mathrm{curl}\sigma_{ab} := \epsilon_{acd}\sigma_b{}^{d:c} = H_{ab}, \tag{3.13}$$

where all quantities in (3.13) are again kept symmetric and traceless.

An electromagnetic field F_{ab} on the spacetime splits into its electric and magnetic parts in the usual way:

$$E_a = F_{ab}u^b, \tag{3.14}$$

$$B_a = \frac{1}{2}\epsilon_{abc}F^{bc}. \tag{3.15}$$

It evolves in accordance with Maxwell's equations

$$F_{ab}{}^{;b} = J_a, \tag{3.16}$$

$$F_{[ab;c]} = 0, \tag{3.17}$$

where $J_a = \rho_{\mathrm{E}}u_a + \mathcal{J}_a$ is the four-current source with charge density ρ_{E} and spatially projected current $\mathcal{J}_a$. Equations (3.16) and (3.17) may also be decomposed with respect to (3.1). With $\mathrm{div}V_a := V_a{}^{:a}$ and $\mathrm{curl}V_a := \epsilon_{abc}V^{c:b}$ for a spatially projected vector V_a, we write the first-order electromagnetic evolution equations (keeping all quantities spatially projected) as

$$\dot{E}_a = -\frac{2}{3}\vartheta E_a + \sigma_{ab}E^b + \mathrm{curl}B_a - \mathcal{J}_a, \tag{3.18}$$

$$\dot{B}_a = -\frac{2}{3}\vartheta B_a + \sigma_{ab}B^b - \mathrm{curl}E_a, \tag{3.19}$$

along with the two spatial constraints

$$\mathrm{div}E_a = \rho_{\mathrm{E}}, \tag{3.20}$$

$$\mathrm{div}B_a = 0. \tag{3.21}$$

The electromagnetic field in turn curves the background spacetime via its stress–energy tensor T_{ab}^{EM}. This may be written in the fluid form (3.5) with

$$\mu = \frac{1}{2}(E^2 + B^2), \tag{3.22}$$

$$p = \frac{1}{6}(E^2 + B^2), \tag{3.23}$$

$$q_a = \mathcal{S}_a, \tag{3.24}$$

$$\pi_{ab} = \frac{1}{3}(E^2 + B^2)P_{ab} - E_a E_b - B_a B_b, \tag{3.25}$$

where $E^2 := E_a E^a$ and $B^2 := B_a B^a$ are the squared field magnitudes and $\mathcal{S}_a := \epsilon_{abc} E^b B^c$ is the covariant Poynting vector. Coupling of the Einstein and Maxwell fields is manifest in (3.10)–(3.13) and (3.18)–(3.25); we explore this in the following sections for the case of a radiative gravitational field.

3.3 Linearised far-field equations

We consider the interactions between gravitational radiation and electromagnetic fields in perturbed Minkowski space with the metric $\tilde{\eta}_{ab} = \eta_{ab} + h_{ab}$. In the transverse–traceless gauge, $\tilde{\eta}_{00} = -1$ and we may choose $u^a = \delta^a_0$, where δ^a_b is the Kronecker delta. Gravitational radiation is covariantly described by the electric and magnetic parts of the Weyl tensor, and more simply in flat space by the shear tensor, which is related to the metric perturbation by

$$\sigma_{ab} = \frac{1}{2}\dot{h}_{ab}^{\mathrm{TT}}. \tag{3.26}$$

Apart from the vanishing vorticity and acceleration, other kinematical and geometrical quantities are greatly simplified by the flatness and symmetry of the spacetime; for example, the expansion scalar and electric Weyl tensor reduce at linear perturbative order to $\vartheta = 0$ and (via (3.11)) $E_{ab} = -\dot{\sigma}_{ab}$.

Second-order evolution equations for an electromagnetic field $\{E_a, B_a\}$ on the spacetime may be derived from (3.18) and (3.19) by spatially projecting their covariant time derivatives, making use of (3.10)–(3.13) to simplify the expressions. These wave-like equations are sourced by the kinematical quantities ϑ and σ_{ab}, with additional Ricci and Weyl curvature terms arising from the non-commutativity of derivatives. Similar equations have been derived for a fully general spacetime, where the only assumption is a single perfect-fluid matter field with a barotropic equation of state [121].

For most astrophysical GW sources we expect to observe, the electromagnetic luminosity ($\sim 10^{37}$ W for a typical galaxy) is dwarfed by the gravitational luminosity (some significant fraction of $c^5/G \sim 10^{52}$ W) [25], and so the energy carried by gravitational radiation is generally much greater than that stored in the electromagnetic field. This translates to $E^2 \sim B^2 \ll \sigma^2 \ll 1$ in geometric units. The wave-like equations for E_a and B_a contain source terms of three sizes: $\sim \sigma E$, $\sim \sigma^2 E$ and $O(E^3)$, with the last arising from the back-reaction of the electromagnetic field on the background spacetime via (3.22)–(3.25). Considering only the leading (in σ) terms at linear order in E, we find

$$\tilde{\Box} E_a = \sigma_{ab}\dot{E}^b + 2\dot{\sigma}_{ab}E^b + \epsilon_{abc}\sigma^{cd}B_d{}^{:b} + \epsilon_{abc}\sigma^{cd:b}B_d + (\tilde{\mathrm{curl}}\,\sigma_{ab})B^b, \tag{3.27}$$

$$\tilde{\Box} B_a = \sigma_{ab}\dot{B}^b + 2\dot{\sigma}_{ab}B^b - \epsilon_{abc}\sigma^{cd}E_d{}^{:b} - \epsilon_{abc}\sigma^{cd:b}E_d - (\tilde{\mathrm{curl}}\,\sigma_{ab})E^b, \tag{3.28}$$

where $\tilde{\Box}\mathcal{T} := \ddot{\mathcal{T}} - \mathcal{T}_{:a}{}^{:a}$ for any spatially projected tensor $\mathcal{T}$. Here and henceforth, a tilde over an operator explicitly indicates that we are raising and lowering indices with the perturbed metric $\tilde{\eta}$.

Equations (3.27) and (3.28), along with the divergence constraints $\tilde{\mathrm{div}} E_a = \tilde{\mathrm{div}} B_a = 0$ from (3.20) and (3.21), govern the evolution of electromagnetic fields in the presence of weak-field gravitational radiation. They are coupled to the first-order evolution and constraint equations (3.11)–(3.13) for σ_{ab}, which may be cast as a constrained wave-like equation in similar fashion to the derivation of (3.27) and (3.28). The shear equations contain terms that are $\sim \sigma^2$, $\sim \sigma^3$ and $O(E^2\sigma)$; at linear order in σ, however, we have

$$\tilde{\Box}\sigma_{ab} = 0 \tag{3.29}$$

with $\tilde{\mathrm{div}}\,\sigma_{ab} = 0$ (which is the transversality condition implied by (3.26)). Hence it is reasonable to treat σ_{ab} as a fixed background of gravitational radiation that drives oscillations in the electromagnetic field via (3.27) and (3.28).

For weak-field calculations, it is convenient to replace the perturbed Minkowski metric $\tilde{\eta}_{ab}$ with an exact one, which simplifies index manipulation and any harmonic expansion of tensor fields. This approximation is trivially valid for (3.29) and its divergence constraint (where replacing the perturbed operators with their Minkowski counterparts only introduces terms that are quadratic- or higher-order in σ), but not so for (3.27) and (3.28). Using a perturbative approach, we consider the gravitationally coupled electromagnetic field as the sum of a free field and an induced first-order perturbation, i.e.

$$E_a = E_a^{(0)} + E_a^{(1)}, \tag{3.30}$$

$$B_a = B_a^{(0)} + B_a^{(1)}, \tag{3.31}$$

where $\{E_a^{(0)}, B_a^{(0)}\}$ is a vacuum Maxwell solution, $E^{(1)} \ll E^{(0)}$ and $B^{(1)} \ll B^{(0)}$. Denoting operators on Minkowski space with overbars, we have

$$\bar{\Box} E_a^{(0)} = \bar{\Box} B_a^{(0)} = 0, \tag{3.32}$$

$$\bar{\mathrm{div}} E_a^{(0)} = \bar{\mathrm{div}} B_a^{(0)} = 0. \tag{3.33}$$

Substituting (3.30)–(3.33) into the equations for $\{E_a, B_a\}$ yields wave-like equations for the induced field that are essentially (3.27) and (3.28) with linear corrections. These corrections are due to the difference operators $(\tilde{\Box} - \bar{\Box})$ and $(\tilde{\mathrm{div}} - \bar{\mathrm{div}})$ giving rise to terms that are $\sim \sigma E$ and non-negligible with respect to (3.27) and (3.28). The induced field equations read

$$\bar{\Box} E_a^{(1)} = F[E_a^{(0)}] + G[B_a^{(0)}] + \{\text{linear corrections}\}[E_a^{(0)}], \tag{3.34}$$

$$\bar{\Box} B_a^{(1)} = F[B_a^{(0)}] - G[E_a^{(0)}] + \{\text{linear corrections}\}[B_a^{(0)}], \tag{3.35}$$

$$\bar{\mathrm{div}} E_a^{(1)} = \{\text{linear corrections}\}[E_a^{(0)}], \tag{3.36}$$

$$\bar{\mathrm{div}} B_a^{(1)} = \{\text{linear corrections}\}[B_a^{(0)}]. \tag{3.37}$$

Here F and G are linear maps defined by (3.27) and (3.28) as

$$F[V_a] := \sigma_{ab} \dot{V}^b + 2\dot{\sigma}_{ab} V^b, \tag{3.38}$$

$$G[V_a] := \epsilon_{abc}\sigma^{cd}V_d^{\ :b} + \epsilon_{abc}\sigma^{cd:b}V_d + (\bar{\mathrm{curl}}\sigma_{ab})V^b, \tag{3.39}$$

with the covariant time and space derivatives equal to their partial counterparts at linear perturbative order.

The divergences of the induced electromagnetic field contain terms that are generally nonzero, even in the absence of sources. Equation (3.36) in particular has been interpreted as an effective four-current generator for the induced field [102], although there is no similar analogy for its magnetic counterpart (3.37). A more suitable comparison might be to think of the corrections in (3.36) and (3.37) as "polarisation" and "magnetisation" effects generated by the spacetime perturbations, with E_a and $B_a^{(0)}$ playing the respective roles of the electric displacement and auxiliary magnetic fields [127].

In this work, we consider a far-field background GW that is plane, monochromatic and linearly polarised with constant amplitude. The geometrical-optics approximation is valid whenever the gravitational wavelength is much smaller than the background radius of curvature, i.e. across the distant wave zone of a typical astrophysical source and well into its local wave zone [23]. More realistic (multimodal) inspiral-type waveforms may be built from superpositions of our simplified model, with the resultant imprint on the electromagnetic field bearing the characteristics of the source waveform.

The background GW is a solution to (3.29) (with $\tilde{\bar{\Box}} = \bar{\Box}$) and $\bar{\mathrm{div}}\sigma_{ab} = 0$. Choosing coordinates x^a such that it propagates in the positive z-direction with zero initial phase, the wave is described by the real part of

$$\sigma_{ab} = \sigma \exp\left(-ik(t-z)\right)p_{ab}, \tag{3.40}$$

where σ_{ab} has been promoted to a complex tensor. The unit polarisation tensor p_{ab} has nonzero components $p_{11} = -p_{22} = \cos 2\alpha$ and $p_{12} = p_{21} = \sin 2\alpha$ for some wave polarisation angle α. Any linear corrections in (3.34)–(3.37) are obtained in the usual way with the metric perturbation, which is given by

$$h_{ab}^{\mathrm{TT}} = \Re\left[\frac{2i}{k}\sigma_{ab}\right] \tag{3.41}$$

in accordance with (3.26).

3.4 Solutions and observational implications

We now consider two simple models for the free field $\{E_a^{(0)}, B_a^{(0)}\}$ in (3.34)–(3.37), and discuss their astrophysical implications. Section 3.4.1 deals with the effects of gravitational radiation on a static electromagnetic field, while GW–EMW interactions are examined in Section 3.4.2.

3.4.1 Static electromagnetic field

When an EMW propagates through a static electromagnetic field, it is resonantly converted to a GW of the same frequency and wavevector; the GW is sourced by a stress–energy tensor proportional to both the radiative and static electromagnetic fields [99]. Astrophysical GWs generated through this "Gertsenshteĭn process" are generally too weak to be of practical interest [6]. The Gertsenshteĭn effect and its inverse process, where a GW in a static electromagnetic field induces an EMW proportional to both fields, might nevertheless be relevant for detecting individual gravitons [128] or high-frequency GWs [129].

The inverse Gertsenshteĭn process is as inefficient as its counterpart, and the fraction of gravitational energy converted is small ($\lesssim 10^{-10}$) even under typical conditions in pulsar environments [100]. However, the energy in the induced EMW might be comparable to that radiated conventionally by astrophysical systems where both the gravitational radiation and magnetic field are strong (but still in the far-field regime of Section 3.3). Hence it is worthwhile to derive the inverse Gertsenshteĭn effect within our framework, and to revisit the feasibility of detecting it in observations.

For a plane GW propagating in a uniform magnetic field, the field component in the direction of the wavevector does not affect the induced EMW. Considering only the projection of the magnetic field onto the (x, y)-plane, the free field may be written as

$$E_a^{(0)} = 0, \tag{3.42}$$

$$B_a^{(0)} = B^{(0)} p_a^{(0)}, \tag{3.43}$$

with the unit polarisation vector $p_a^{(0)} = (0, \cos\beta, \sin\beta, 0)$ for some field polarisation angle β. All linear corrections in (3.34)–(3.37) vanish for static and uniform electromagnetic fields, and we expect separable solutions to the system. Isolating the spatial dependence in our ansatz as a scalar harmonic, we promote (3.30) and (3.31) to complex vectors and write

$$E_a^{(1)} = \mathcal{E}_a \exp{(ikz)}, \tag{3.44}$$

$$B_a^{(1)} = \mathcal{B}_a \exp{(ikz)}, \tag{3.45}$$

where $\{\mathcal{E}_a, \mathcal{B}_a\}$ depends only on time.

Equations (3.34)–(3.37) now simplify to an ordinary differential equation in time for the sole independent component of the induced field. Solving this with homogeneous initial conditions, we arrive at

$$\mathcal{E}_a = \epsilon_a{}^{bc}\mathcal{B}_b\delta_c^3, \tag{3.46}$$

$$\mathcal{B}_a = \frac{1}{2}hB^{(0)}(kt\exp{(-ikt)} - \sin kt)p_a^{(1)}, \tag{3.47}$$

where $h = 2\sigma/k$ and $p_a^{(1)} = p_a{}^b p_b^{(0)} = (0, \cos{(2\alpha - \beta)}, \sin{(2\alpha - \beta)}, 0)$. Equations (3.46) and (3.47) describe a plane, monochromatic and linearly polarised EMW; its amplitude is given by

$$E^{(1)} = B^{(1)} = \frac{1}{2}hB^{(0)}(k^2t^2 - kt\sin{(2kt)} + \sin^2 kt)^{\frac{1}{2}}, \tag{3.48}$$

which is proportional to time for large t.

The period $T_{\mathrm{GW}} = 2\pi/k$ and strength h of the sinusoidal background GW determine a natural timescale $T_{\mathrm{GW}}/(2\pi h)$, at which $B^{(1)} \sim B^{(0)}$ and higher-order perturbations to the electromagnetic field become significant. In reality, the linear growth in (3.48) is contingent on a steady build-up of oscillations over time, and is more of an upper bound for EMWs induced by chirp- or ringdown-type GWs with evolving frequency and/or amplitude. We incorporate such waveforms with the generalised model

$$\sigma_{ab} = \frac{1}{2}(k + \dot{k}t)h\exp{(-i(k + \frac{1}{2}\dot{k}t)t + ikz - \lambda t)}p_{ab}, \tag{3.49}$$

where the spatial dependence has been left unchanged from (3.40) to main-

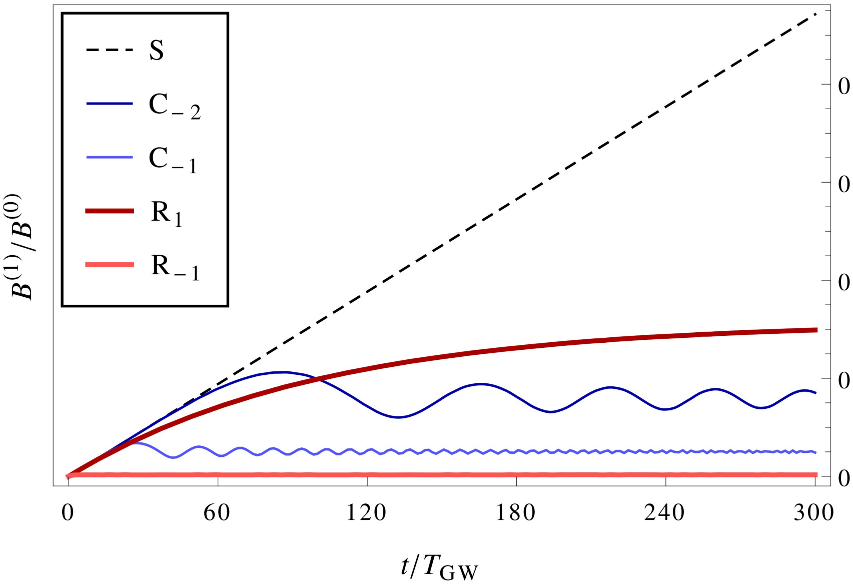

Figure 3.1: Induced EMW amplitude relative to background magnetic field strength for different gravitational waveforms with $f_{\rm GW} = 10\,{\rm Hz}$ and $h_I = 10^{-3}$: a sinusoid (S), a chirp with $\dot{f}_{\rm GW} = 10^n\,{\rm Hz/s}$ (C_n), and a ringdown with $\tau_{\rm GW} = 10^n\,{\rm s}$ (R_n).

tain separability. When $\lambda = 0$, (3.49) describes a linear chirp with constant chirp rate $\dot{f}_{\rm GW} = \dot{k}/(2\pi)$, while for $\dot{k} = 0$ it gives a ringdown with damping timescale $\tau_{\rm GW} = 1/\lambda$. Equations (3.34)–(3.37) may then be solved analytically to yield Fresnel-like integrals in the chirp case, and solutions with bounded exponential growth in the ringdown case.

The inverse Gertsenshteĭn effect is potentially significant in the context of neutron star binaries, since the induced EMW is proportional in strength to both h and $B^{(0)}$. While the stable GWs from the early inspirals of such systems might be conducive to resonant growth, any associated magnetic fields will have fallen off drastically where the wave zone for gravitational radiation begins. Hence we consider a neutron star binary coalescence in an interaction region I with the strongest possible GW strain h_I and magnetic field strength B_I, i.e. at the inner edge $R_I := c/(2\pi f_{\rm GW})$ of the local wave zone [23].

Figure 3.1 shows the ratio $B^{(1)}/B^{(0)}$ for various gravitational waveforms, using typical values of (initial) frequency $f_{\rm GW} = 10\,{\rm Hz}$ and measured strain $h_\oplus = 10^{-21}$ ($h_I := h_\oplus R_\oplus / R_I = 10^{-3}$) that correspond to a neutron star binary coalescence at $R_\oplus = 10^2\,\mathrm{Mpc}$. With such a large interaction strain, we have $B^{(1)} \nearrow B^{(0)}$ in just 300 GW periods for the sinusoidally driven EMW, which is well within the $\sim 10^4$ waveform cycles observable by LIGO. In general, however, the induced EMW amplitude is reduced with increasing variability in the gravitational waveform. The inverse Gertsenshteĭn effect is insignificant for the R_{-1} waveform, and hence completely negligible for actual stellar-mass ringdowns with their damping timescales of $\sim 10^{-5}\,\mathrm{s}$.

By Poynting's theorem [127], the spacetime-averaged power density transferred from a sinusoidal GW to its induced EMW is (at leading order in time)

$$-\langle \dot{u}_{\mathrm{GW}} \rangle = \langle \mathrm{d\bar{i}v}\mathcal{S}_a \rangle = \frac{1}{8\mu_0} h_I^2 B_I^2 \omega_{\mathrm{GW}}^2 t, \tag{3.50}$$

where units have been restored and $\omega_{\mathrm{GW}} = 2\pi f_{\mathrm{GW}}$.[5] The GW energy density is given via (2.18) by

$$u_{\mathrm{GW}} = \frac{c^2}{32\pi G} h_I^2 \omega_{\mathrm{GW}}^2. \tag{3.51}$$

Hence the fraction of gravitational energy converted in the region I is

$$\Upsilon = \frac{2\pi G}{\mu_0 c^2} B_I^2 t^2, \tag{3.52}$$

in accordance with the original Gertsenshteĭn result [99, 100]. Even for a neutron star binary coalescence containing a magnetar[6] with radius $R_S = 10^4\,\mathrm{m}$ and surface field strength $B_S = 10^{11}\,\mathrm{T}$ ($B_I := B_S R_S^3 / R_I^3 = 10^3\,\mathrm{T}$), Υ over 10^4 GW periods is small ($\sim 10^{-9}$).

At leading order in time, the time-averaged Poynting flux of the induced EMW at the interaction distance R_I is given by

$$\langle \mathcal{S} \rangle := \sqrt{\langle \mathcal{S}_a \rangle \langle \mathcal{S}^a \rangle} = \frac{c}{24\mu_0} h_I^2 B_I^2 \omega_{\mathrm{GW}}^2 t^2. \tag{3.53}$$

Like the magnetic dipole radiation emitted by a neutron star, the Gertsenshteĭn radiation typically dwarfs the beamed radiation arising from synchrotron

[5] Note that the covariant Poynting vector is now $\mathcal{S}_a = \epsilon_a{}^{bc}\Re[E_b]\Re[B_c]$.

[6] Magnetars are highly magnetised neutron stars with typical periods of 1 to $10\,\mathrm{s}$ and surface fields ranging from 10^9 to 10^{11} T [130].

emission in the magnetosphere, but can neither propagate through the ionised interstellar medium nor be detected by existing radio telescopes due to its low frequency ($\lesssim 10^3\,\mathrm{Hz}$). It is more instructive to compare (3.53) with the angle-averaged flux density of the maximal magnetic dipole radiation at R_I, which is given by [131]

$$\langle S_{\mathrm{dip}}\rangle = \frac{1}{6\mu_0 c^3} B_S^2 \omega_{\mathrm{dip}}^4 R_S^6 R_I^{-2}, \tag{3.54}$$

where ω_{dip} is the neutron star's angular velocity.

For a neutron star binary coalescence containing a millisecond pulsar with radius $R_S = 10^4\,\mathrm{m}$ and surface field $B_S = 10^6\,\mathrm{T}$, we have $\langle \mathcal{S}\rangle \sim 10^9\,\mathrm{W/m^2}$ after 300 GW periods and $\langle S_{\mathrm{dip}}\rangle \sim 10^{17}\,\mathrm{W/m^2}$. If the pulsar is replaced by a similarly sized magnetar with a $1\,\mathrm{s}$ period and $10^{11}\,\mathrm{T}$ surface field, the average flux generated through the Gertsenshteĭn process after 300 GW periods is $\sim 10^{19}\,\mathrm{W/m^2}$—a good 10^4 times larger than that due to the magnetar's dipole radiation. Although this excess flux cannot be detected directly, it should in principle contribute significantly to the heating of any bipolar outflows or nearby interstellar clouds. The resultant secondary emission of pulsed electromagnetic radiation (with pulse frequency f_{GW}) might then be observable by conventional telescopes across a range of bands, depending on the composition of the surrounding nebula.

3.4.2 Electromagnetic radiation

Interactions between gravitational and electromagnetic radiation in the far field are most prominently characterised by a variety of interference-like (but fully nonlinear) effects on the latter. For our framework, we consider a free EMW that is plane, monochromatic and linearly polarised; since electromagnetic wavelengths are typically much shorter than gravitational and astrophysical length scales, our choice is motivated by the validity of geometrical optics as much as the suitability of plane harmonics to the tensor–vector contractions in (3.38) and (3.39). The EMW is described by the real part of

$$E_a^{(0)} = E^{(0)} \exp\left(i(n_b x^b + \psi)\right) p_a^{(0)}, \tag{3.55}$$

$$B_a^{(0)} = \frac{1}{n} \epsilon_a{}^{bc} n_b E_c^{(0)}, \tag{3.56}$$

where the four-wavevector $n_a = n(-1, \sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ has the usual polar and azimuthal angles (with respect to the z-axis), and ψ is the initial phase relative to (3.40). The unit polarisation vector now lies in the plane orthogonal to the spatial wavevector n_μ, and is defined such that $p_3^{(0)} = \sin\theta\sin\gamma$ for some wave polarisation angle γ.

For separable solutions, the tensor–vector contractions in (3.38) and (3.39) motivate the ansatz

$$E_a^{(1)} = \frac{1}{2}(\mathcal{E}_a^{(+)} \exp{(i m_\mu^{(+)} x^\mu)} + \mathcal{E}_a^{(-)} \exp{(i m_\mu^{(-)} x^\mu)}), \tag{3.57}$$

$$B_a^{(1)} = \frac{1}{2}(\mathcal{B}_a^{(+)} \exp{(i m_\mu^{(+)} x^\mu)} + \mathcal{B}_a^{(-)} \exp{(i m_\mu^{(-)} x^\mu)}), \tag{3.58}$$

where $m_\mu^{(\pm)} := n_\mu \pm k_\mu$ (with $k_\mu = k\delta_\mu^3$) are spatial wavevectors associated with the first-order perturbation, and we have used the phasor multiplication rule

$$\Re[\exp{(i\Phi)}]\Re[\exp{(i\Psi)}] = \frac{1}{2}\Re[\exp{(i|\Phi+\Psi|)} + \exp{(i|\Phi-\Psi|)}]. \tag{3.59}$$

The scalar Helmholtz harmonics $\exp{(i m_\mu^{(\pm)} x^\mu)}$ decouple from (3.34) and (3.35), leaving a system of ordinary differential equations in time for $\{\mathcal{E}_a^{(\pm)}, \mathcal{B}_a^{(\pm)}\}$. Although the choice of ansatz is amenable to plane-wave solutions, the divergences (3.36) and (3.37) depend on the angular configuration $\{\theta, \phi, \alpha, \gamma\}$ of the waves, and are in general nonzero. As it turns out, the full system (3.34)–(3.37) of evolution and constraint equations is inconsistent with (3.57) and (3.58) for all but two wave configurations: parallel ($\theta = 0$) and antiparallel ($\theta = \pi$), both of which yield plane-wave perturbations.

GW–EMW interactions have previously been studied in the 1+3 formalism, but neglecting the electric–magnetic self-interaction terms in the evolution equations (3.27) and (3.28) (i.e. setting $G = 0$ in (3.39)). This decouples the spatial dependence without explicit knowledge of the covariant Helmholtz harmonics, and for parallel waves the resultant ordinary differential equation describes a resonantly driven oscillator with natural and driving frequency $m = n + k$ [120, 126]. By considering the full equations, however, we find that the effect of G is to cancel the terms due to F when $\theta = 0$, such that (3.34) and (3.35) become homogeneous wave equations. In other words, parallel waves do not interact at all. Such cancellation does not occur for antiparallel waves, although we find no resonant interaction either. The lack of interaction between parallel GWs and EMWs is a known result that has been obtained via other approaches in the literature [104, 106, 107].

For a general interaction angle θ, (3.34)–(3.37) do not admit two-mode solutions of the form (3.57) and (3.58). Nevertheless, the $m^{(\pm)}$-modes are dominant when the frequency ratio $\rho := k/n$ is small, in the sense that the evolution and constraint equations are consistent at leading order when $\rho \sec\theta \ll 1$. Since $\rho \lesssim 10^{-4}$ in most astrophysical scenarios (the highest-frequency GW sources have $f_{\rm GW} \sim 10^3\,{\rm Hz}$ [25], while $f_{\rm EM} \sim 10^7\,{\rm Hz}$ is the lowest frequency that modern radio telescopes are sensitive to [132, 133]), the two-mode ansatz is justifiable and valid for all angular configurations except orthogonal waves.

Solving the ordinary differential equations for $\{\mathcal{E}_a^{(\pm)}, \mathcal{B}_a^{(\pm)}\}$ with homogeneous initial conditions, we obtain a wave perturbation with a complicated dependence on $\{k, n, \theta, \phi, \alpha, \gamma\}$ (see Appendix A). The solution (A.1)–(A.10) remains linearly polarised, however, as its components have a common phase offset ψ and time dependence

$$\begin{aligned}\mathcal{E}_a^{(\pm)}, \mathcal{B}_a^{(\pm)} \quad &\propto \quad m^{(\pm)} \exp\left(-i(n \pm k)t\right) \\ &\qquad -m^{(\pm)} \cos\left(m^{(\pm)}t\right) + i(n \pm k) \sin\left(m^{(\pm)}t\right), \end{aligned} \tag{3.60}$$

where $m^{(\pm)} = (k^2 + n^2 \pm 2kn\cos\theta)^{1/2}$. This represents an effective splitting of the EMW frequency into four perturbation frequencies $m^{(\pm)}$ and $n \pm k$, along with the original free frequency n. The amplitude of the wave perturbation vanishes in the limit for parallel waves, and is $\sim hE^{(0)}$ for antiparallel waves; its characteristic size for general θ is

$$E^{(1)}, B^{(1)} = O(hE^{(0)}/\rho), \tag{3.61}$$

which indicates that nonlinear interference effects between GWs and EMWs might become significant as $h \nearrow \rho$. When $h > \rho$, higher-order perturbations come into play and the validity of the perturbative approach is limited.

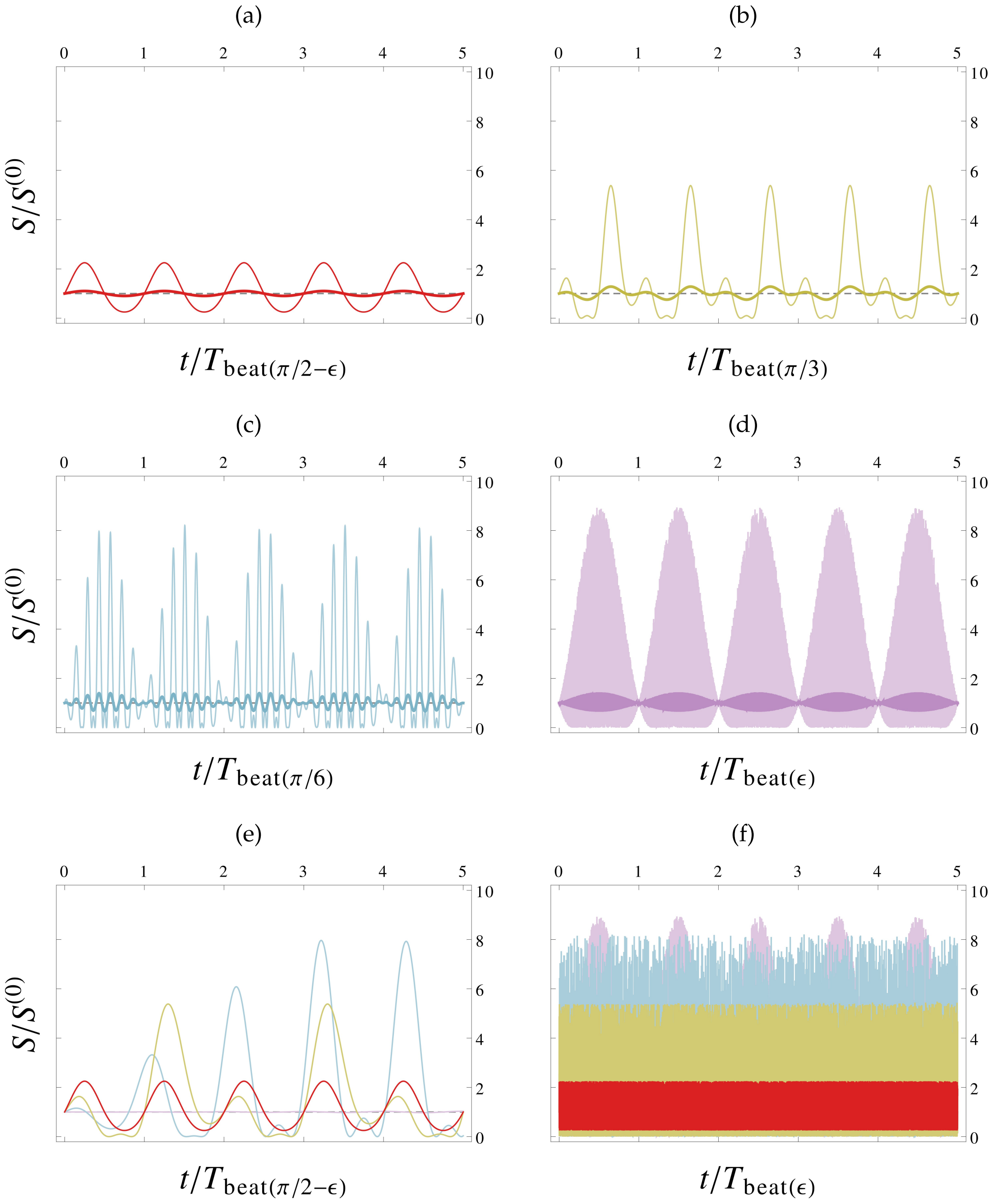

Figure 3.2: Perturbed Poynting flux envelope for $h = \rho$ at different interaction angles between $\theta = \pi/2$ and $\theta = 0$ ((a)–(d)), along with comparisons of the four configurations over different timescales ((e) and (f)). The bolded curves in (a)–(d) are for $h = 10^{-1}\rho$.

To illustrate the behaviour in the $h \sim \rho$ regime, we define here a complex covariant Poynting vector[7]

$$S_a := \epsilon_a{}^{bc} E_b B_c^*, \tag{3.62}$$

which gives the envelope $S := (S_a^* S^a)^{1/2}$ of the usual Poynting vector magnitude for the full field $\{E_a, B_a\}$. Due to the presence of cross terms in (3.62), the Poynting flux envelope is spatially periodic on the gravitational length scale $2\pi/k$. Figure 3.2 shows the relative flux envelope $S/S^{(0)}$ for a plus-polarised GW ($\alpha = 0$) and an EMW in the (y, z)-plane ($(\phi, \gamma) = (\pi/2, 0)$), where S is evaluated at $x^\mu = 0$ and $S^{(0)}$ is the constant flux envelope of the free EMW.

In accordance with previous results [126], there is an emergence of θ-dependent beats in the perturbed EMW, with frequency given by the greatest common divisor of the spectrum $\{n, m^{(\pm)}, n \pm k\}$. It is useful to define an approximate beat period $T_{\mathrm{beat}(\theta)} := 2\pi/(k - (m^{(+)} - m^{(-)})/2)$, which describes much of the beat structure for most values of θ. As the interaction angle decreases from $\pi/2 - \epsilon$ to ϵ (where $\epsilon = 10^{-3}$), the peaks for the extremal case $h = \rho$ increase from around $S/S^{(0)} = 2$ to a limiting value of $S/S^{(0)} = 9$. Additionally, we find significant nonlinear amplification of the beats as h is raised from $10^{-1}\rho$ to ρ. Beating effects are essentially negligible for $h \lesssim 10^{-3}\rho$.

There is an overall flux increase apparent in Figure 3.2, attributable to the transfer of energy from the GW to the electromagnetic field as in Section 3.4.1. For a clearer picture of this flux amplification and its dependence on interaction angle, we require the Poynting flux averaged over finite time intervals T, which we denote explicitly as $\langle \mathcal{S} \rangle_T = (\langle \mathcal{S}_a \rangle_T \langle \mathcal{S}^a \rangle_T)^{1/2}$ with

$$\langle \mathcal{S}_a \rangle_T = \frac{1}{T} \int_0^T dt \, \mathcal{S}_a. \tag{3.63}$$

Considering the same angular configuration as before, Figure 3.3 shows a sequence of polar plots (with respect to interaction angle) for $\langle \mathcal{S} \rangle_T / \langle \mathcal{S}^{(0)} \rangle$ with increasing T, where $\langle \mathcal{S} \rangle_T$ is evaluated at $x^\mu = 0$ and $\langle \mathcal{S}^{(0)} \rangle$ for the free EMW is effectively constant over gravitational timescales. When $h \sim \rho$, the overall flux in the forward sector $|\theta| < \pi/2$ is approximately doubled for small $|\theta|$ after

[7] Our definition differs by a factor of 1/2 from the conventional complex Poynting vector, which is used to calculate time-averaged flux for sinusoidal plane waves.

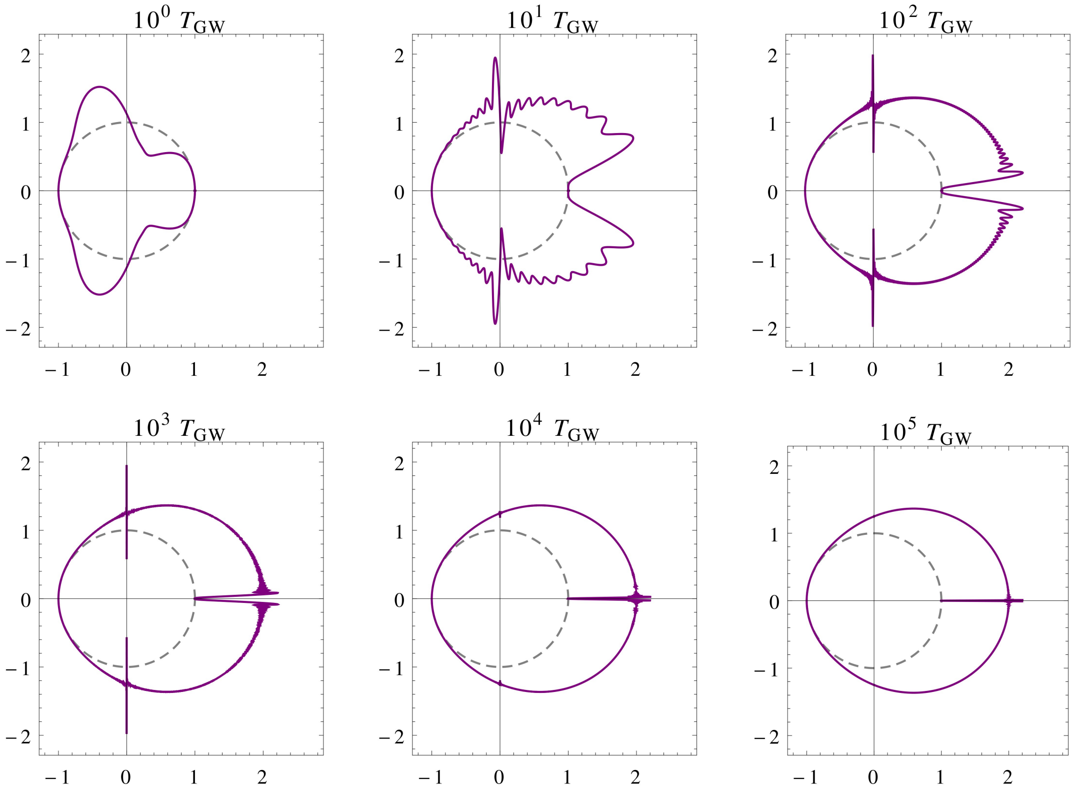

Figure 3.3: Perturbed time-averaged Poynting flux $\langle \mathcal{S} \rangle_T / \langle \mathcal{S}^{(0)} \rangle$ as a radial function of interaction angle, for different time intervals between $T = 10^0 T_{\rm GW}$ and $T = 10^5 T_{\rm GW}$. Each plot is for $h = \rho$, with $\theta = 0$ on the positive horizontal axis such that the GW propagates to the right.

just 10^2 GW periods. There is little to no flux amplification in the backward sector $|\theta| > \pi/2$. We note that the interaction between parallel waves vanishes as expected, with the beat frequency and the induced field itself going to zero smoothly as $|\theta| \to 0$; the seemingly pathological behaviour of $\langle \mathcal{S} \rangle_T / \langle \mathcal{S}^{(0)} \rangle$ at $\theta = 0$ is attributable to the non-smoothness of the time-averaging operation (3.63) in the limit as $T \to \infty$.

The nonlinear interference depicted in Figures 3.2 and 3.3 is potentially relevant for GW sources with electromagnetic counterparts that are long-lived (lasting at least several GW periods), and preferably high-frequency ($\rho \lesssim 10^{-10}$) for effects to be significant at low interaction strains. Possible counterparts for a compact binary system are a pulsar component as in Section 3.4.1 or, more promisingly, an extended electromagnetic source such as a bipolar outflow or interstellar cloud around the binary. If an extended source emits radiation in the band $f_{\mathrm{EM}} \sim f_{\mathrm{GW}}/h_I$, its radiation profile might be characterised by intensity fluctuations and overall flux amplification at small angular distances from the binary's sky location; the fluctuations should increase in frequency to $f_{\mathrm{beat}(\pi/2)} = f_{\mathrm{GW}}$ as the interaction angle widens, then diminish rapidly at larger angular distances as h_I falls below $10^{-1}\rho$.

Figure 3.3 effectively describes the flux amplification at different interaction angles, but there is actually a tiny deflection of the perturbed time-averaged Poynting vector $\langle \mathcal{S}_a \rangle$ in the direction of the GW. The original Poynting vector, averaged over all time, is given simply by

$$\left\langle \mathcal{S}_a^{(0)} \right\rangle = \frac{1}{2} \Re \big[S_a^{(0)} \big]. \tag{3.64}$$

Its perturbed counterpart reduces to

$$\langle \mathcal{S}_a \rangle = \left\langle \mathcal{S}_a^{(0)} \right\rangle + \frac{1}{2} \left\langle \Re \big[S_a^{(1)} \big] \right\rangle + \frac{1}{2} \left\langle \Re \Big[\epsilon_a{}^{bc} E_b^{(1)} B_c^{(1)} \Big] \right\rangle, \tag{3.65}$$

since both cross terms average to zero over all time. The first two terms in (3.65) depend only on the angular configuration of the waves and have no spatial dependence, while the spatial dependence in the third term is negligible for $\rho \ll 1$. We consider the deflection angle Θ_{def} between (3.64) and (3.65) with the same angular configuration as before; expanding the angle in powers of both h and ρ, we find

$$\Theta_{\mathrm{def}} = O(\min\{h^2/\rho, \rho\}), \tag{3.66}$$

which is valid in the forward sector but away from $\theta = 0$, where Θ_{def} goes sharply to $\pi/2$ due to the time-averaging operation.

Equation (3.66) becomes $\Theta_{\mathrm{def}} = O(h^2/\rho)$ for $h < \rho$, such that the deflection of the time-averaged Poynting vector varies with both f_{GW} and f_{EM}. This is a new result, although a frequency-dependent deflection of time-averaged flux does not necessarily imply the dispersion of light by GWs. The maximal angle $\Theta_{\mathrm{def}} \sim h$ agrees with previous results for the ray deflection angle in other approaches [104, 111]. A hydrogen-line radio wave passing the neutron star binary coalescence of Section 3.4.1 with an impact parameter corresponding to $h_I \sim \rho \sim 10^{-8}$ will have its Poynting vector deflected by $\sim 10^{-3}\,\mathrm{arcsec}$; this is comparable to the deflection due to conventional gravitational lensing by the same system ($\sim 10^{-2}\,\mathrm{arcsec}$). Such angular deviations are too small to be observed directly, but might be amenable to microlensing techniques.

Another well-documented GW–EMW interaction is the gravitational analogue of Faraday rotation experienced by an EMW in the field of a passing GW; if the projection of the EMW polarisation vector onto the GW polarisation plane is aligned with the plus mode, it will undergo a slight (oscillatory) rotation as long as the cross mode is nonzero [106, 107, 109]. In our framework, there is indeed no rotation for a plus-polarised GW ($\alpha = 0$) and an aligned EMW ($(\phi, \gamma) = (\pi/2, 0)$), since $\Re[E_a^{(0)}]$ and $\Re[E_a]$ are parallel. We consider the rotation angle[8] Θ_{rot} between $E_a^{(0)}$ and E_a for a cross-polarised GW ($\alpha = \pi/4$) and the same EMW at $x^\mu = 0$; expanding the angle in powers of h, we find

$$|\Theta_{\mathrm{rot}}| = O(h), \tag{3.67}$$

in accordance with previous results [106, 107]. The rotation angle also oscillates at $\sim f_{\mathrm{GW}}$ as expected [106], with beat frequency $f_{\mathrm{beat}(\theta)}$. Again, since we have $h \lesssim \rho \lesssim 10^{-4}$ in most astrophysical scenarios, the gravitationally in-

[8] We use the complex fields to smooth out oscillations on the electromagnetic timescale; the angle Θ between two complex vectors V_a and W_a is given by $\cos\Theta = \Re[V_a^* W^a]/(VW)$.

duced Faraday rotation due to the passage of a far-field GW will typically be $\lesssim 10\,\mathrm{arcsec}$ and difficult to detect using modern techniques.

3.5 Discussion

In this chapter, we have studied far-field interactions between gravitational radiation and electromagnetic fields in the 1+3 approach to relativity, with a view to characterising observable signatures on any electromagnetic radiation emitted by astrophysical GW sources. Linearised evolution and constraint equations for the electromagnetic field on a GW-perturbed spacetime have been approximated and solved perturbatively on Minkowski space, where the relevant harmonic expansions are explicitly known and analytically tractable.

We have rederived the inverse Gertsenshteĭn effect by applying this framework to the interaction of a plane GW with a static electromagnetic field, and have also considered the resonantly induced electromagnetic radiation in an astrophysical setting. Order-of-magnitude calculations have shown that the Gertsenshteĭn radiation is comparable to the magnetic dipole radiation for highly magnetised neutron stars in compact binary systems; in the presence of a surrounding nebula, this might lead to a secondary emission of electromagnetic radiation pulsed at the GW frequency.

Several geometrical-optics effects have been found in the case of interacting GWs and EMWs. There is no resonant growth of the electromagnetic field as found in other work, due to the additional consideration of electric–magnetic self-interaction contributions here. We have also demonstrated that the nonlinear fluctuation and amplification of electromagnetic energy flux becomes significant as the GW strain approaches the GW–EMW frequency ratio from below, and might serve as a distinctive astrophysical signature of gravitational radiation emitted near or within an extended electromagnetic source.

From the various assumptions and approximations employed in this work, it is evident that the analytical advantages of the 1+3 approach are limited even for our simple model. A calculation of second-order perturbations induced by the first-order fields via (3.34)–(3.37) might provide a clearer picture of interacting waves in the $h \sim \rho$ regime, although a rapid blow-up of the full field (signalling the breakdown of the perturbative approach) is more likely. Numerical solutions of (3.27) and (3.28) or their unlinearised versions might be worth pursuing in this case, both to verify results from the perturbative approach and to facilitate more accurate models by extending the framework into the $h > \rho$ regime.

Our results have observational implications for two types of astrophysical source: compact sources with large values of h and $B^{(0)}$ (for the inverse Gertsenshteĭn effect to be relevant), and extended ones with a wide range of interaction angle (for more prominent nonlinear interference effects). They are not restricted to any specific example suggested here, however; neither have we considered scenarios where the gravitational and electromagnetic sources are separate. Detailed source models that incorporate far-field gravitational–electromagnetic interactions—or any Einstein–Maxwell coupling in general—will be an asset to the detection and observation of GW sources at present, and indeed the larger realm of multimessenger astronomy in the future.

Chapter 4

Augmented kludge waveforms

To prepare for the formulation of the LISA mission over the next few years, several outstanding and urgent questions in data analysis must be addressed using mock data challenges. These data challenges will require accurate and computationally affordable waveform models for anticipated sources such as EMRIs. Previous data challenges have made use of the well-known analytic kludge waveforms of Barack & Cutler, which are extremely quick to generate but dephase relative to more accurate waveforms within hours, due to their mismatched radial, polar and azimuthal frequencies.

In this chapter, we describe an augmented Barack–Cutler waveform model that uses a frequency map to the correct Kerr frequencies, along with updated PN evolution equations and a local trajectory fit to a more accurate model. The augmented kludge waveforms stay in phase for months and may be generated with virtually no additional computational cost.

The material in this chapter is adapted from [59, 134], but has been updated to include new unpublished results that were obtained with the latest version of the waveform model [135].

4.1 Background

With the recent success of the LISA Pathfinder mission [18], space-based GW astronomy is now one step closer to becoming a reality. The launch and operation of Pathfinder provides an early demonstration of the technology necessary for proposed detectors such as LISA to probe the source-rich and scientifically rewarding millihertz GW sky. Results from the mission also offer vital insights into the noise properties of the LISA instrument; these will be used over the next few years to address several important open questions in data analysis, which must be dealt with prior to mission adoption at the end of this decade. Accurate and computationally affordable models of likely sources will allow such issues to be investigated in forthcoming mock data challenges, and are therefore urgently needed.

EMRIs are an important type of source for LISA and other space-based detectors. Radiation reaction from the emission of GWs causes the orbit of the compact object around the central black hole to shrink and circularise adiabatically. During the final years of inspiral, the orbital dynamics display extreme relativistic effects that arise because the compact object is deep within the strong-field region of the black-hole spacetime. These effects are imprinted on the GW signal from the source; measuring them will allow us to map out the multipole structure of the spacetime, and hence to test the strong-field validity of black-hole solutions in general relativity [136].

The optimal detection and identification of EMRI signals depends on the availability of a waveform model that is as accurate as possible. While the extreme mass ratio prohibits the use of NR simulations, it does allow the source to be modelled faithfully within the framework of black-hole perturbation theory. At first order in mass ratio, such models are based on solving for the outgoing radiation field Ψ_4 with the Teukolsky equation (2.38) (sourced by a compact object moving along a Kerr geodesic). Orbital evolution may be introduced in Teukolsky-based models by balancing the change in orbital energy and angular momentum against the radiation fluxes of "snapshot" geodesic waveforms. The ongoing development of self-force calculations will eventually provide a more accurate evolution that accounts for both dissipative and conservative self-interaction effects at higher orders in mass ratio.

For scoping out data analysis, the large parameter space of EMRI models and the complexity of their waveforms necessitates the use of templates that can be generated as quickly as possible. As Teukolsky-based waveforms are computationally expensive, they have been supplemented by approximate waveforms designed for robust use in data analysis. The waveform model used for previous mock LISA data challenges [137] is the analytic kludge (AK)

[57], in which the orbit is built from Keplerian ellipses, with relativistic inspiral, periapsis precession and Lense–Thirring precession imposed using analytic PN evolution equations. The AK model is extremely quick to compute, but is less accurate than the numerical kludge (NK) model [58], which combines Kerr geodesics with Teukolsky-fitted PN orbital evolution for greater accuracy. In both kludge models, generation of the waveform from the orbit is sped up from Teukolsky-based models by using a flat-space approximation.

In this chapter, we describe an augmented analytic kludge (AAK) waveform model based on a frequency mapping method. The orbital frequency and two precession rates in the AK model are matched to appropriate combinations of the fundamental frequencies describing the radial, polar and azimuthal components of geodesic orbits in Kerr spacetime [138]. We also update the model with suitable PN evolution equations and include an additional fit to the NK model, which itself shows excellent agreement with the Teukolsky-based geodesic and inspiral waveforms in [62, 63]. The length of time over which the AAK waveform stays in phase with the NK waveform is typically increased by a few orders of magnitude, while the added computational cost is insignificant as the map-and-fit is only performed locally.

A brief overview of the AK and NK models is given in Section 4.2, while the AAK model is presented in Section 4.3. In Section 4.3.1, we first introduce the Kerr fundamental frequencies and the parameter-space map they induce in the AK model, before describing the technical details of the AAK implementation in Section 4.3.2. The performance of the augmented model is then compared to that of the original AK model in Section 4.3.3, with the more accurate but slower NK model used as the benchmark for both.

4.2 Kludge waveform models

A kludge in the context of EMRI modelling is any approximate model that uses a combination of formalisms to generate waveforms quickly and extensively for sampling purposes. Kludge waveforms capture many qualitative features of more accurate EMRI waveforms, and (owing to their modular construction) can be modified to incorporate self-force information as it becomes available.

Two widely used kludges are introduced briefly in this section: the AK waveform of Barack & Cutler [57], which is very fast to compute and provides the basis for our new model, and the NK waveform of Babak et al. [58], which we take as a fiducial model for calibration and benchmarking purposes. Other approximate EMRI models exist but at varying levels of implementation (e.g. [139, 140]), and we do not consider them in this work (apart from the PN fluxes of Sago & Fujita [141], which are used as part of the AAK model).

Assuming the spin of the compact object is negligible, an EMRI can be described by 14 parameters: the two masses $(\mu, M \gg \mu)$ of the system, the three components of the central black hole's spin vector $\mathbf{S}$, three constants $\mathbf{E}$ describing the compact object's (instantaneous) orbit, the three components of the compact object's position vector $\mathbf{X}$ with respect to the black hole, and the three components of the system's position vector $\mathbf{R}$ with respect to the Solar System. Of these 14 degrees of freedom, seven are extrinsic to the source: two in $\mathbf{S}$ and one in $\mathbf{X}$ (corresponding to spatial rotation of the source), three in $\mathbf{R}$ (corresponding to spatial translation), and one in $\mathbf{E}$ (corresponding to temporal translation). The parameters of an EMRI model are often chosen to decouple the intrinsic degrees of freedom from the extrinsic ones, which are generally cheaper to search over during data analysis [142].

Schematically, the main ingredients of a kludge waveform model are then (i) the evolution of the orbital constants along the inspiral (i.e. the "phase-space" trajectory), using PN or fitted fluxes $\mathbf{F}$:

$$\dot{\mathbf{E}} = \mathbf{F}(\mu, M, \mathbf{S}, \mathbf{E}); \tag{4.1}$$

(ii) the construction of the compact object's worldline (i.e. the "configuration-space" trajectory), using geodesic or flux-derived expressions $\mathbf{G}$:

$$\dot{\mathbf{X}} = \mathbf{G}(\mu, M, \mathbf{S}, \mathbf{E}); \tag{4.2}$$

and (iii) the generation of the waveform field h at the detector, using some weak-field multipole formula H:

$$h(t) = H(\mathbf{X}, \mathbf{R}). \tag{4.3}$$

4.2.1 Analytic kludge

In the AK model [57], both the orbital trajectory and the waveform are computed in a flat-space approximation, with relativistic effects such as inspiralling and precession added separately. The trajectory is built out of rotating Keplerian ellipses. Radiation reaction is introduced in phase space, where the orbital constants describing a Keplerian ellipse are evolved with PN equations. In configuration space, the orientation of this ellipse is also evolved with PN equations to simulate relativistic precession. The waveform is then generated using the Peters–Mathews mode-sum approximation for Keplerian orbits [35], in which the mass quadrupole moment is decomposed into harmonics of the Keplerian orbital frequency.

Since a Keplerian orbit is confined within the plane normal to its angular momentum vector $\mathbf{L}$, the AK waveform is constructed in an $\mathbf{L}$-based coordinate frame

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})_{\hat{\mathbf{L}}} := \left(\frac{(\hat{\mathbf{R}} \cdot \hat{\mathbf{L}})\hat{\mathbf{L}} - \hat{\mathbf{R}}}{\mathcal{S}_{\mathbf{L},\mathbf{R}}}, \frac{\hat{\mathbf{R}} \times \hat{\mathbf{L}}}{\mathcal{S}_{\mathbf{L},\mathbf{R}}}, \hat{\mathbf{L}} \right), \tag{4.4}$$

and projected transverse to the wave frame

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})_{\mathrm{AK}} := \left(\frac{\hat{\mathbf{R}} \times \hat{\mathbf{L}}}{\mathcal{S}_{\mathbf{L},\mathbf{R}}}, \frac{\hat{\mathbf{L}} - (\hat{\mathbf{L}} \cdot \hat{\mathbf{R}})\hat{\mathbf{R}}}{\mathcal{S}_{\mathbf{L},\mathbf{R}}}, -\hat{\mathbf{R}} \right), \tag{4.5}$$

where the normalisation factor $\mathcal{S}_{\mathbf{L},\mathbf{R}} := (1 - (\hat{\mathbf{L}} \cdot \hat{\mathbf{R}})^2)^{1/2}$. These two frames are made time-varying (with respect to a fixed heliocentric and ecliptic-based frame [79]) through the forced precession of $\mathbf{L}$.

The two waveform polarisations in the transverse–traceless gauge (with respect to $(\hat{\mathbf{x}}, \hat{\mathbf{y}})_{\mathrm{AK}}$) are given by the n-mode sums

$$h_+ = \sum_{n=1}^{\infty} h_n^+, \tag{4.6}$$

$$h_\times = \sum_{n=1}^{\infty} h_n^\times, \tag{4.7}$$

with

$$h_n^+ = (1 + (\hat{\mathbf{R}} \cdot \hat{\mathbf{L}})^2)(b_n \sin 2\tilde{\gamma} - a_n \cos 2\tilde{\gamma}) + (1 - (\hat{\mathbf{R}} \cdot \hat{\mathbf{L}})^2)c_n, \tag{4.8}$$

$$h_n^\times = 2(\hat{\mathbf{R}} \cdot \hat{\mathbf{L}})(b_n \cos 2\tilde{\gamma} + a_n \sin 2\tilde{\gamma}), \tag{4.9}$$

where $\tilde{\gamma}$ is an azimuthal angle in the orbital plane measuring the direction of periapsis with respect to $(\hat{\mathbf{R}}\cdot\hat{\mathbf{L}})\hat{\mathbf{L}}-\hat{\mathbf{R}}$ (i.e. the orthogonal projection of $\hat{\mathbf{z}}_{\mathrm{AK}}$ onto the plane normal to $\hat{\mathbf{L}}$).[9] The functions (a_n, b_n, c_n) describe the changing mass quadrupole moment of a Keplerian orbit with mean anomaly $\Phi(t)$, eccentricity e and orbital angular frequency $\dot{\Phi}$, and are given by [35]

$$\begin{aligned} a_n &= -n\mathcal{A}(J_{n-2}(ne) - 2eJ_{n-1}(ne) + (2/n)J_n(ne) \\ &\quad +2eJ_{n+1}(ne) - J_{n+2}(ne)) \cos n\Phi, \end{aligned} \tag{4.10}$$

$$b_n = -n\mathcal{A}(1 - e^2)^{1/2}(J_{n-2}(ne) - 2J_n(ne) + J_{n+2}(ne)) \sin n\Phi, \tag{4.11}$$

$$c_n = 2\mathcal{A}J_n(ne) \cos n\Phi, \tag{4.12}$$

where the J_n are Bessel functions of the first kind, and $\mathcal{A} = (\dot{\Phi}M)^{2/3}\mu/|\mathbf{R}|$ in the extreme-mass-ratio limit.

In the ecliptic-based coordinate system, the sky position $\hat{\mathbf{R}} \equiv (\theta_S, \phi_S)$ of the source and the black-hole spin orientation $\hat{\mathbf{S}} \equiv (\theta_K, \phi_K)$ are effectively constant. It is convenient to represent $\hat{\mathbf{L}}$ in ecliptic coordinates with respect to $\hat{\mathbf{S}}$ (where both vectors are moved by parallel transport to the Solar System barycentre). We have

$$\hat{\mathbf{L}} = \hat{\mathbf{S}} \cos\iota + \left(\frac{\hat{\mathbf{z}} - (\hat{\mathbf{z}}\cdot\hat{\mathbf{S}})\hat{\mathbf{S}}}{|\hat{\mathbf{z}} - (\hat{\mathbf{z}}\cdot\hat{\mathbf{S}})\hat{\mathbf{S}}|} \cos\alpha + \frac{\hat{\mathbf{S}}\times\hat{\mathbf{z}}}{|\hat{\mathbf{S}}\times\hat{\mathbf{z}}|} \sin\alpha \right) \sin\iota, \tag{4.13}$$

where $\hat{\mathbf{z}} = [0, 0, 1]^T$ is normal to the ecliptic plane, ι is the inclination angle between $\hat{\mathbf{L}}$ and $\hat{\mathbf{S}}$, and α is an azimuthal angle in the spin-equatorial plane measuring the direction of $\hat{\mathbf{L}} - (\hat{\mathbf{L}} \cdot \hat{\mathbf{S}})\hat{\mathbf{S}}$ with respect to $\hat{\mathbf{z}} - (\hat{\mathbf{z}} \cdot \hat{\mathbf{S}})\hat{\mathbf{S}}$ (i.e. the

[9] We have changed some of the notation in [57] for consistency, since we are constructing a hybrid model using different formalisms. For example: the notation for the angles γ and $\tilde{\gamma}$ has been swapped; the notation ν for the orbital frequency $\dot{\Phi}/(2\pi)$ is unused; the notation for the inclination λ is now ι.

angle between the orthogonal projections of $\hat{\mathbf{L}}$ and $\hat{\mathbf{z}}$ onto the plane normal to $\hat{\mathbf{S}}$). Furthermore, since $\tilde{\gamma}$ is neither intrinsic nor extrinsic, it is useful to define the purely intrinsic parameter $\gamma := \tilde{\gamma} - \beta$, where $\beta = \beta(\hat{\mathbf{R}}, \hat{\mathbf{S}}, \hat{\mathbf{L}}) = \beta(\theta_S, \phi_S, \theta_K, \phi_K, \iota, \alpha)$ is an azimuthal angle in the orbital plane measuring the direction of $\hat{\mathbf{L}} \times \hat{\mathbf{S}}$ with respect to $(\hat{\mathbf{R}} \cdot \hat{\mathbf{L}})\hat{\mathbf{L}} - \hat{\mathbf{R}}$.

Only one of the six parameters comprising $\mathbf{E} = (e, \iota, \dot{\Phi})$ and $\mathbf{X} = (\Phi(t), \gamma, \alpha)$ in the above Keplerian setup changes with time. In the AK model, $(e, \dot{\Phi}, \gamma, \alpha)$ are promoted to functions of time and evolved with mixed-order PN expressions that depend on $(\mu, M, a = |\mathbf{S}|/M, \mathbf{E})$ [143–146], while ι is approximated as constant (since the inclination angle of a typical EMRI varies extremely slowly [147]). The Keplerian orbit shrinks and circularises as $\dot{\Phi}(t)$ and $e(t)$ increase and decrease respectively. From (4.13), the time dependence of the orbital orientation $\hat{\mathbf{L}}(t)$ is confined to $\alpha(t)$, where $\dot{\alpha}$ is precisely the angular rate of Lense–Thirring precession. Finally, the angular rate of periapsis precession is given by $\dot{\gamma} + \dot{\alpha}$ since $\gamma(t)$ is measured with respect to $\hat{\mathbf{L}}(t) \times \hat{\mathbf{S}}$.

While the waveform field $h_{+,\times}$ is effectively planar at the Solar System and may be calculated in the fixed heliocentric frame (as opposed to a detector-centric one), the rotational and orbital motion of LISA in the ecliptic plane must be factored into the detector's response to $h_{+,\times}$. In the standard LISA framework, the field is transformed into the response functions $h_{I,II}$ via (2.46) and (2.47), with $F_I^{+,\times}(\theta_S, \phi_S, \theta_K, \phi_K)$ given explicitly in [79]; these rotate with respect to $\hat{\mathbf{R}}$ as the plane of the detector along its orbit precesses around the ecliptic plane. Doppler modulation of the waveform phase (through $\Phi(t)$) is also included to correct for the orbital motion of the detector itself.

The AK model was the first waveform model used to investigate the precision of LISA parameter estimation over the full (modulo compact-object spin) EMRI parameter space [57]. Due to its computational efficiency, the model has also been employed in past mock LISA data challenges to generate injected signals in simulated data and parametrised templates for search algorithms [137]. However, the approximate waveforms it produces are demonstrably inaccurate, and will result in reduced detection and parameter estimation performance if used to analyse data sets containing realistic EMRI signals.

4.2.2 Numerical kludge

In the NK model [58], the orbital trajectory is computed in curved space with a treatment that is fully relativistic up to the evolution of orbital constants [148, 149], i.e. it is built out of Kerr geodesics. The three constants of motion for a geodesic are evolved with Teukolsky-fitted PN equations, which introduces radiation reaction. In configuration space, precession effects are obtained for free by integrating the geodesic equations along the phase-space trajectory. The curved-space coordinates of the compact object's worldline are then associated artificially with coordinates in flat space, and the waveform is generated using the standard quadrupole formula (or variants that include additional contributions from higher-order moments of mass [150, 151]).

The NK waveform is constructed in an $\mathbf{S}$-based coordinate frame

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})_{\mathbf{S}} := \left(\frac{\hat{\mathbf{R}} \times \hat{\mathbf{S}}}{\mathcal{S}_{\mathbf{S},\mathbf{R}}}, \frac{\hat{\mathbf{R}} - (\hat{\mathbf{R}} \cdot \hat{\mathbf{S}})\hat{\mathbf{S}}}{\mathcal{S}_{\mathbf{S},\mathbf{R}}}, \hat{\mathbf{S}} \right), \tag{4.14}$$

and projected transverse to the wave frame

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})_{\mathrm{NK}} := \left(\frac{\hat{\mathbf{R}} \times \hat{\mathbf{S}}}{\mathcal{S}_{\mathbf{S},\mathbf{R}}}, \frac{\hat{\mathbf{S}} - (\hat{\mathbf{S}} \cdot \hat{\mathbf{R}})\hat{\mathbf{R}}}{\mathcal{S}_{\mathbf{S},\mathbf{R}}}, -\hat{\mathbf{R}} \right), \tag{4.15}$$

where the normalisation factor $\mathcal{S}_{\mathbf{S},\mathbf{R}} := (1 - (\hat{\mathbf{S}} \cdot \hat{\mathbf{R}})^2)^{1/2}$. Aligning the z-axis with $\mathbf{S}$ is a more natural choice for the NK model, since the compact object's worldline is computed in Boyer–Lindquist coordinates. The two wave frames (4.5) and (4.15) are related by a rotation about $\mathbf{R}$.

Using the decomposition (2.44), the two waveform polarisations in the transverse–traceless gauge (with respect to $(\hat{\mathbf{x}}, \hat{\mathbf{y}})_{\mathrm{NK}}$) are given by

$$h_+ = \frac{1}{2} h_{ij}^{\mathrm{TT}} H_{kl}^{+} \delta^{ik} \delta^{jl}, \tag{4.16}$$

$$h_\times = \frac{1}{2} h_{ij}^{\mathrm{TT}} H_{kl}^{\times} \delta^{ik} \delta^{jl}, \tag{4.17}$$

where we have switched to Latin spatial indices and δ^{ij} is the Kronecker delta. Via (2.12)–(2.14) and (2.17), the far-field metric perturbation in the quadrupole approximation is given by

$$h_{ij}^{\mathrm{TT}} = \frac{2}{|\mathbf{R}|}\left(P_i^k P_j^l - \frac{1}{2}P_{ij}P^{kl}\right)\ddot{I}_{kl}, \tag{4.18}$$

where contraction with $P_{ij} := (\delta_{ij} - \hat{z}_i\hat{z}_j)_{\mathrm{NK}}$ projects transverse to $\mathbf{R}$.

In the extreme-mass-ratio limit, the mass quadrupole moment is simply

$$I_{ij} = \mu x_i x_j, \tag{4.19}$$

where the $x_i(t)$ are Cartesian components of the compact object's position vector $\mathbf{X}$ with respect to the frame (4.14) centred on the black hole. Although (4.18) (with (4.19)) is a weak-field equation in flat-space coordinates, the NK model specifies and calculates $(x_1, x_2, x_3) = (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta)$ in Boyer–Lindquist coordinates. The self-consistency of this approach clearly degrades further into the strong field, but does not severely impact the effectiveness of the NK waveforms as an approximation to Teukolsky-based ones [58].

A timelike Kerr geodesic is described fully by three first integrals of motion: the orbital energy E, the projection L_z of the orbital angular momentum $\mathbf{L}$ onto $\mathbf{S}$, and the quadratic Carter constant Q. Along such an orbit, $(r(t), \theta(t), \phi(t))$ are obtained by integrating the geodesic equations for a test particle in the Kerr spacetime (2.37); these are written in canonical form as [22]

$$\Sigma\frac{dr}{d\tau} = \pm\sqrt{V_r}, \tag{4.20}$$

$$\Sigma\frac{d\theta}{d\tau} = \pm\sqrt{V_\theta}, \tag{4.21}$$

$$\Sigma\frac{d\phi}{d\tau} = V_\phi, \tag{4.22}$$

$$\Sigma\frac{dt}{d\tau} = V_t, \tag{4.23}$$

where τ is proper time along the worldline and $\Sigma = r^2 + a^2\cos^2\theta$. The potential functions $V_{r,\theta,\phi,t}$ are given by

$$V_r(r) = P^2 - (r^2 + (L_z - aE)^2 + Q)\Delta, \tag{4.24}$$

$$V_\theta(\theta) = Q - \cos^2\theta \left(a^2(1-E^2) + \frac{L_z^2}{\sin^2\theta} \right), \tag{4.25}$$

$$V_\phi(r,\theta) = \frac{L_z}{\sin^2\theta} - aE + \frac{aP}{\Delta}, \tag{4.26}$$

$$V_t(r,\theta) = aL_z - a^2E\sin^2\theta + \frac{(r^2+a^2)P}{\Delta}, \tag{4.27}$$

with $P = E(r^2+a^2) - aL_z$ and $\Delta = r^2 - 2Mr + a^2$.

In practice, it is convenient to work with alternative parametrisations of (E, L_z, Q). For a bound orbit, the geodesic may be specified by the parameters $(r_p, r_a, \theta_{\min})$ (the values of r at periapsis and apoapsis, and the minimal value of θ respectively), which fully describe the range of motion in the radial and polar coordinates. The roots of V_r determine r_p and r_a, while the roots of V_θ determine $\cos\theta_{\min}$ (the maximal value of $\cos\theta$). Another parametrisation is (e, ι, p) (the quasi-Keplerian eccentricity, inclination and semi-latus rectum); these are defined in terms of $(r_p, r_a, \theta_{\min})$ as

$$(e, \iota, p) := \left(\frac{r_a - r_p}{r_a + r_p}, \frac{\pi}{2} - \theta_{\min}, \frac{2r_ar_p}{r_a + r_p} \right). \tag{4.28}$$

Finally, since the configuration-space parameters (r, θ) oscillate between the bounds $r_p \leq r \leq r_a$ and $\theta_{\min} \leq \theta \leq \pi - \theta_{\min}$, it is useful to define

$$(\psi, \chi) := \left(\cos^{-1}\left(\frac{p-r}{er} \right), \cos^{-1}\left(\frac{\cos\theta}{\cos\theta_{\min}} \right) \right), \tag{4.29}$$

such that ψ (the quasi-Keplerian true anomaly) and χ are the phases of radial and polar motion respectively.

The orbital constants $\mathbf{E} = (E, L_z, Q)$ in the above geodesic setup do not vary with time. Radiation reaction is added to the NK model by evolving $\mathbf{E}$ with fluxes that depend on $(\mu, M, a, \mathbf{E})$ (note that the inclination $\iota(t) = \tan^{-1}(\sqrt{Q}/L_z)$ is correctly time-dependent in this model). These fluxes are mixed-order PN expressions that have been fitted to the results of Teukolsky-based computations for circular inclined orbits [149]. Integrating the geodesic equations along the phase-space trajectory then gives $\mathbf{X} = (\psi(t), \chi(t), \phi(t))$, complete with relativistic precession. Once the waveform field $h_{+,\times}$ has been calculated via (4.16)–(4.19), the LISA response functions $h_{I,II}$ may be obtained through the method outlined in Section 4.2.1.

Waveforms from the NK model display excellent agreement with Teukolsky-based waveforms in the strong-field regime; they are reliable up to a closest approach of $r_p \approx 5M$, with typical matches of over 0.95 [58]. NK waveforms might even be accurate enough to serve as templates in actual LISA detection algorithms. However, they are still slightly expensive to generate in large numbers due to the relatively elaborate construction of the phase- and configuration-space trajectories, while added computational cost also arises in the parameter conversion $(E, L_z, Q) \leftrightarrow (e, \iota, p)$, the handling of plunge, etc.

4.3 Augmented analytic kludge

The AK model is 5–15 times faster than the NK model at generating year-long waveforms sampled at 0.2 Hz for a generic $(10^1, 10^6)M_\odot$ EMRI with low initial eccentricity ($e_0 \lesssim 0.3$); this speed-up is enhanced for longer waveform durations, but diminished for higher initial eccentricity (since more modes must be summed in the Peters–Mathews approximation).[10] However, AK waveforms suffer from severe dephasing with respect to NK waveforms, even at the early-inspiral stage. In Figure 4.1, the AK waveform for a $(10^1, 10^6)M_\odot$ EMRI with initial semi-latus rectum $p_0 = 15M$ matches the qualitative features of the corresponding NK waveform, but is a full cycle out of phase within three hours. This is due to the mismatched frequencies in the two models.

In Sections 4.3.1 and 4.3.2, we describe the construction of a hybrid model that capitalises on the benefits of both kludges. The AK model is augmented with an initial map to the fundamental frequencies of Kerr geodesic motion, which corrects the instantaneous phasing as shown in Figure 4.1. Over longer timescales, the mapped orbital trajectory is further improved through self-consistent PN evolution and a local polynomial fit to the phase-space trajectory

[10] The sums in (4.6) and (4.7) must be truncated at some arbitrary number of modes N, which directly affects both the speed and accuracy of the AK model. This number may be specified by setting a threshold for the relative power radiated into the N-th harmonic, and has been experimentally determined to scale linearly with eccentricity [57]. We use $N = \lfloor 30e_0 \rfloor$ as the default value for both the AK and AAK models.

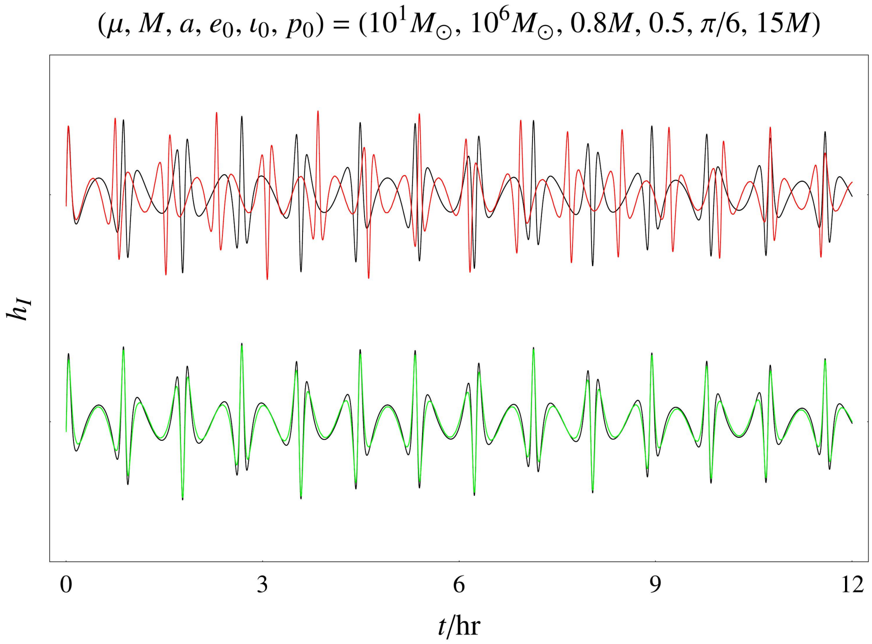

Figure 4.1: First 12 hours of AK (red) and AAK (green) waveforms overlaid on NK waveform (black), for the early inspiral of a $(10^1, 10^6)M_\odot$ EMRI with initial semi-latus rectum $p_0 = 15M$.

of the NK model. Fast algorithms for higher-order fits and plunge handling have been incorporated in the latest AAK implementation, which has been made publicly available at `github.com/alvincjk/EMRI_Kludge_Suite` as part of a software suite for generating kludges.

The initial version of the AAK model yields waveforms that can remain phase-coherent with NK waveforms for over two months, but with overlap values lower than 0.97 [59]. This is the commonly chosen minimal match for a waveform template bank that corresponds to a 90%-ideal observed event rate [83], and thus ensures the equivalent localisation of any signal detected with such banks of AAK and NK templates. In Section 4.3.3, we report further improved results for the present AAK implementation [135]. Two-month overlaps higher than 0.97 are found for EMRIs with varying spin and eccentricity; however, the overlaps still degrade with proximity to plunge, due to the divergence of the AAK and NK trajectories deep within the strong field.

4.3.1 The fundamental-frequency map

The geodesic equations (4.20)–(4.23) take a simple form with the choice of a timelike parameter $\lambda = \int d\tau/\Sigma$ [152, 153]; this decouples (4.20) and (4.21), and for a bound orbit makes the radial and polar components of motion manifestly periodic with respect to λ. For the azimuthal and temporal components (whose potentials depend only on (r,θ)), overall rates of evolution may be obtained by averaging (4.26) and (4.27) over many periods of (r,θ) motion.

From the radial and polar periods $\Lambda_{r,\theta}$, the average azimuthal rate $\langle d\phi/d\lambda\rangle$ and the average temporal rate $\langle dt/d\lambda\rangle$ (denoted Γ by analogy with the Lorentz factor), we may define three angular and dimensionless fundamental frequencies $\Omega_{r,\theta,\phi}$ for the test particle's motion with respect to coordinate time. In terms of $(r_p, r_a, \theta_{\min})$, these frequencies are written as [141, 154]

$$\Omega_r = \frac{2\pi}{\Lambda_r \Gamma}, \tag{4.30}$$

$$\Omega_\theta = \frac{2\pi}{\Lambda_\theta \Gamma}, \tag{4.31}$$

$$\Omega_\phi = \lim_{N\to\infty} \frac{1}{N^2 \Lambda_r \Lambda_\theta \Gamma} \int_0^{N\Lambda_r} d\lambda_r \int_0^{N\Lambda_\theta} d\lambda_\theta \, V_\phi(r(\lambda_r), \theta(\lambda_\theta)), \tag{4.32}$$

where $\Lambda_{r,\theta}$ and Γ are given by

$$\Lambda_r = 2\int_{r_p}^{r_a} \frac{dr}{\sqrt{V_r}}, \tag{4.33}$$

$$\Lambda_\theta = 4\int_{\theta_{\min}}^{\pi/2} \frac{d\theta}{\sqrt{V_\theta}}, \tag{4.34}$$

$$\Gamma = \lim_{N\to\infty} \frac{1}{N^2 \Lambda_r \Lambda_\theta} \int_0^{N\Lambda_r} d\lambda_r \int_0^{N\Lambda_\theta} d\lambda_\theta \, V_t(r(\lambda_r), \theta(\lambda_\theta)). \tag{4.35}$$

Expressions for $\Omega_{r,\theta,\phi}$ in terms of (e, ι, p) have been derived by Schmidt [138]; these are less compact, but have more utility in practical implementations.

In terms of the fundamental frequencies, the periapsis and Lense–Thirring precession rates are given by $\Omega_\phi - \Omega_r$ and $\Omega_\phi - \Omega_\theta$ respectively. These vanish in the Newtonian limit, where $\Omega_{r,\theta,\phi}$ approach a single orbital frequency Ω from below, i.e. $\Omega_r \nearrow \Omega_\theta \nearrow \Omega_\phi \nearrow \Omega$. The frequency Ω is then related to (e,p) by Kepler's third law:

$$\Omega = \left(\frac{1-e^2}{p}\right)^{3/2}. \tag{4.36}$$

In the AK model, Ω is associated with the quantity $\dot{\Phi}M$; however, periapsis and Lense–Thirring precession are added on top of $\dot{\Phi}$ via $\dot{\gamma}+\dot{\alpha}$ and $\dot{\alpha}$ respectively, and so Ω is the lowest frequency by construction. This inconsistency with the relativistic case leads to mismatched frequencies when supplying identical parameters to the two models, since the same value of p in (4.36) specifies the radial AK frequency while approximating the azimuthal NK frequency. In other words, the frequencies in the AK model are generally too high.

A three-dimensional endomorphism over the AK space of orbits is induced by requiring that the radial, polar and azimuthal frequencies $(\dot{\Phi}, \dot{\Phi}+\dot{\gamma}, \dot{\Phi}+\dot{\gamma}+\dot{\alpha})$ for any (e,ι,p) have the same values as the (dimensionful) relativistic frequencies $\omega_{r,\theta,\phi} := \Omega_{r,\theta,\phi}/M$. We map the parameters (M,a,p) rather than (e,ι,p) to unphysical values; this gives better results since periapsis and Lense–Thirring precession are more directly determined by the central mass and its rotation respectively. The map $(M,a,p) \mapsto (\tilde{M},\tilde{a},\tilde{p})$ is given implicitly by solving the algebraic system of equations

$$\dot{\Phi}(\tilde{M},\tilde{a},\tilde{p}) = \omega_r(M,a,p), \tag{4.37}$$

$$\dot{\gamma}(\tilde{M},\tilde{a},\tilde{p}) = \omega_\theta(M,a,p) - \omega_r(M,a,p), \tag{4.38}$$

$$\dot{\alpha}(\tilde{M},\tilde{a},\tilde{p}) = \omega_\phi(M,a,p) - \omega_\theta(M,a,p) \tag{4.39}$$

for the unphysical set $(\tilde{M},\tilde{a},\tilde{p})$, which is defined as the root closest to the physical set (M,a,p) with a Euclidean metric on parameter space.

Substituting $(\tilde{M},\tilde{a},\tilde{p})$ for (M,a,p) in the AK model provides an instantaneous correction of its frequencies at any point along the inspiral trajectory $(e(t),\iota(t),p(t))$. In principle, applying the map along the entirety of a fiducial inspiral will keep the AK waveform phase-coherent with relativistic waveforms (generated from that trajectory) until plunge. However, such an inspiral is usually more expensive to compute (as in the case of the NK model), and the additional cost from evaluating the map itself also scales linearly with the number of points sampled along the trajectory. Complications also arise as the compact object approaches the point of plunge, where the fundamental frequencies diverge and the map (4.37)–(4.39) is no longer well-defined.

4.3.2 Implementation

In order to retain the main advantage of the AK model, computational costs are kept as low as possible by evaluating the map at a small number of points and relying on independent evolution of the orbital constants over long timescales. Firstly, the NK trajectory is generated at and around the specified initial point $(M, a, e_0, \iota_0, p_0)$ over a user-defined timescale T_{fit}, which depends on the radiation-reaction timescale $T_{\text{RR}} := M^2/\mu$ and specifies the duration over which the AAK model is calibrated. The timescale T_{fit} and number of sample points N_{fit} may be adjusted adaptively based on the proximity of (e_0, ι_0, p_0) to plunge; they typically satisfy $0.1T_{\text{RR}} \lesssim T_{\text{fit}} \lesssim 10T_{\text{RR}}$ and $N_{\text{fit}} \lesssim 10$ (the latter to ensure an added computational cost of $\lesssim 1\%$). Evaluation of the map at each of the N_{fit} points gives a local "best-fit" trajectory $(\tilde{M}(t), \tilde{a}(t), e(t), \iota, \tilde{p}(t))_{\text{fit}}$ for the AAK model, where the unphysical $(\tilde{M}(t), \tilde{a}(t))_{\text{fit}}$ now change with time and $\iota_{\text{fit}} = \iota_0$ is approximately constant over the duration T_{fit}.

The actual trajectory $(\tilde{M}, \tilde{a}, e(t), \iota, \tilde{p}(t))$ in the AAK model is generated independently from the NK model, and hence more rapidly. From the mapped initial point $(\tilde{M}, \tilde{a}, e_0, \iota_0, \tilde{p}_0)$ (which also lies on the best-fit trajectory by construction), $(e(t), \tilde{p}(t))$ are evolved with 3PN $O(e^6)$ expressions given by Sago & Fujita [141], while $(\tilde{M}, \tilde{a}, \iota)$ are left constant. The configuration-space evolution of (Φ, γ, α) is performed with the appropriate combinations (4.37)–(4.39) of the fundamental frequencies, given by Sago–Fujita expressions that are consistent at 3PN $O(e^6)$ with the phase-space evolution. Higher-order 4PN $O(e^6)$ expressions [141] have also been tested, but these seem to result in poorer agreement with NK waveforms (which use fluxes of up to 3PN), possibly due to the known divergence of certain expansions beyond 3PN order [155].

With inclination constant along both the best-fit and 3PN $O(e^6)$ trajectories, sampling $(\tilde{M}, \tilde{a}, e(t), \tilde{p}(t))$ at each of the N_{fit} points allows the calculation of

$(\tilde{M}(t), \tilde{a}(t), e(t), \tilde{p}(t))_{\mathrm{fit}} - (\tilde{M}, \tilde{a}, e(t), \tilde{p}(t))$ over the duration T_{fit}. This difference trajectory is then fitted to polynomials in time, extrapolated over the lifetime of the inspiral, and added to the 3PN $O(e^6)$ evolution. In the initial AAK implementation, the coefficients of the quadratic fit are given by second-order finite-difference quotients (i.e. $N_{\mathrm{fit}} = 3$), which works well but only for small values of T_{fit}. The present version uses a quartic least-squares fit, which allows the choice of longer T_{fit} and consequently gives better long-term phase agreement with NK waveforms.

The augmentations to the AK framework are focused on improving the phase information of its waveforms, since the amplitude of a GW signal is measured far less precisely than its phase. However, the calculation of amplitude in the original AK model ($\mathcal{A}$ in (4.10)–(4.12)) is a decent approximation since it is based on $\dot{\Phi}M$, which is assigned a value $\approx \Omega_\phi$ that turns out to be correct for this purpose (see discussion around (4.36)). Hence it is the AAK amplitude that is shifted away from the fiducial NK value through the mapping of frequencies and the unphysical evolution of $\tilde{M}$. A simple adjustment is made to reverse this shift; in (4.10)–(4.12) for the AAK model, the amplitude is now given by

$$\mathcal{A} = \frac{(\omega_\phi M)^{2/3}\mu}{|\mathbf{R}|}, \tag{4.40}$$

where the azimuthal frequency ω_ϕ and the physical black-hole mass M are used in place of $\dot{\Phi} = \omega_r$ and $\tilde{M}$ (i.e. M in the original AK model) respectively.

Finally, a fast method of plunge handling has been added to the present AAK implementation; this feature is useful in general, but especially when generating large numbers of AAK waveform templates for search algorithms. The compact object plunges when its instantaneous orbit along the phase-space trajectory $\mathbf{E}(t)$ becomes unstable, i.e.

$$\frac{\partial^2 V_r(r, a, \mathbf{E})}{\partial r^2} \leq \frac{\partial V_r(r, a, \mathbf{E})}{\partial r} = V_r(r, a, \mathbf{E}) = 0, \tag{4.41}$$

where V_r is given in (4.24) with $\mathbf{E} = (E, L_z, Q)$. This point is termed the last stable orbit $\mathbf{E}_{\mathrm{LSO}}$, and is precisely the point at which the discriminant $\mathcal{D}(a, \mathbf{E})$ of the quartic polynomial $V_r(r)$ changes sign from positive (four real roots) to negative (two real roots) [156]. Since $\mathcal{D}$ is a simple analytic function of the quartic coefficients, it is computationally trivial to check for stability at every integration step for the phase-space trajectory, provided the evolution is done in terms of (E, L_z, Q).[11]

Plunge detection in the AAK model is far less straightforward than in the other two kludges, partly because the evolution is performed in terms of the quasi-Keplerian orbital parameters, and the computational benefits of the discriminant method are nullified by having to convert $(e, \iota, \tilde{p}) \to (E, L_z, Q)$. Furthermore, $(\tilde{M}, \tilde{a}, \tilde{p})$ are unphysical; the inverse of the map (4.37)–(4.39) is computationally expensive and (more crucially) ill-defined at plunge, and so cannot be used to obtain the physical parameters for stability calculations.

To circumvent these issues, the AAK model uses (4.36) with $\Omega \approx \Omega_\phi$ to obtain an approximation for the physical parameter p. While generating the phase-space trajectory, it checks (between the least- and most-bound orbits [62]) the stability of (e, ι, p) at every radiation-reaction interval T_{RR}. Once the stability changes across an interval, it then bisects that interval to find p_{LSO}, and smoothly zeroes the waveform over 10 additional orbits with a one-sided Planck-taper window [157]. The added computational cost associated with this algorithm is $\lesssim 1\%$. Although the approximation for p is crude, the phase-space trajectories in the AAK and NK models are generally divergent to begin with, and the plunge points for both models may differ significantly even if a more accurate expression is used.

4.3.3 Benchmarking

The specified initial state of an EMRI in the AAK model is described by the intrinsic parameters $(\mu, M, a, e_0, \iota_0, \gamma_0, \psi_0)$ and the extrinsic parameters $(p_0, \theta_S, \phi_S, \theta_K, \phi_K, \alpha_0, D)$, where $D := |\mathbf{R}|$.[12] In configuration space, transfor-

[11] The discriminant method is applicable to the NK model, and may speed it up slightly. Currently, the NK implementation precomputes (for the specified value of a) an interpolated p_{LSO} surface over the relevant region of $(e, \iota)_{\mathrm{LSO}}$ space, by finding and examining the roots of $V_r(r)$ numerically. It then checks for $p < p_{\mathrm{LSO}}(e, \iota)$ when generating the phase-space trajectory.

[12] For a source at cosmological redshift z, the values of D and (μ, M) are replaced with the luminosity distance $D(1+z)$ and the redshifted masses $(\mu(1+z), M(1+z))$ respectively.

mations from $\mathbf{X}_{\mathrm{AAK}} = (\psi, \gamma, \alpha)$ to $\mathbf{X}_{\mathrm{AK}} = (\Phi, \gamma, \alpha)$ and $\mathbf{X}_{\mathrm{NK}} = (\psi, \chi, \phi)$ are required for a comparison of the three waveform models. We have chosen to specify the quasi-Keplerian true anomaly ψ in the (shared) AAK parameter space, since there is no closed-form expression for ψ in terms of the mean anomaly Φ. The conversion $\psi \rightarrow \Phi$ is given by the Keplerian expressions [158]

$$\Phi = E - e \sin E, \tag{4.42}$$

$$E = \tan^{-1}\left(\frac{\sqrt{1-e^2}\sin\psi}{e+\cos\psi}\right), \tag{4.43}$$

where E is known as the eccentric anomaly.

On the other hand, the AAK model retains the AK parameters (γ, α); these have explicit meanings in the (intrinsic) $\mathbf{L}$-based coordinate frame

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})_{\mathbf{L}} := \left(\frac{\hat{\mathbf{L}} \times \hat{\mathbf{S}}}{\mathcal{S}_{\mathbf{L},\mathbf{S}}}, \frac{(\hat{\mathbf{S}} \cdot \hat{\mathbf{L}})\hat{\mathbf{L}} - \hat{\mathbf{S}}}{\mathcal{S}_{\mathbf{L},\mathbf{S}}}, \hat{\mathbf{L}}\right), \tag{4.44}$$

where the normalisation factor $\mathcal{S}_{\mathbf{L},\mathbf{S}} := (1 - (\hat{\mathbf{L}} \cdot \hat{\mathbf{S}})^2)^{1/2}$ and $\hat{\mathbf{L}}(\alpha)$ is given in ecliptic coordinates by (4.13). The unit position vector of the compact object with respect to this frame is $\hat{\mathbf{r}}_{\mathbf{L}} = [\cos(\psi + \gamma), \sin(\psi + \gamma), 0]^T$, and a change of basis to the S-based coordinate frame (4.14) gives

$$\hat{\mathbf{r}}_{\mathbf{S}} = \mathbf{Q}_{\mathbf{S}}^T \mathbf{Q}_{\mathbf{L}} \hat{\mathbf{r}}_{\mathbf{L}}, \tag{4.45}$$

where the orthogonal matrices $\mathbf{Q}_{\mathbf{S}} := [\hat{\mathbf{x}}|\hat{\mathbf{y}}|\hat{\mathbf{z}}]_{\mathbf{S}}$ and $\mathbf{Q}_{\mathbf{L}} := [\hat{\mathbf{x}}|\hat{\mathbf{y}}|\hat{\mathbf{z}}]_{\mathbf{L}}$ are formed from the triads in (4.14) and (4.44) respectively. It is then straightforward to obtain (χ, ϕ) from $\hat{\mathbf{r}}_{\mathbf{S}} = [\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta]^T$, via (4.29).

As a generic example, we consider a prograde EMRI with redshifted component masses $(\mu, M) = (10^1, 10^6)M_\odot$, spin $a = 0.5M$, and initial orbital parameters $(e_0, \iota_0, p_0) = (0.1, \pi/6, 8.25M)$. The initial semi-latus rectum is chosen such that the compact object plunges approximately one year after entering the LISA band at a representative frequency $f_{\mathrm{GW}} = 2.7\,\mathrm{mHz}$, where f_{GW} is defined as twice the azimuthal orbital frequency (i.e. the dominant GW harmonic at low eccentricity). In the AAK model, the fitting timescale and number of sample points are set to $T_{\text{fit}} \leq 10T_{\text{RR}}$ and $N_{\text{fit}} \leq 10$ respectively, with inequality in the case of adaptive adjustments.

One important result from our comparison studies is that the AK model can lead to an overestimation of SNR if used without modification. This is due to the fact that the frequencies in the AK model are generally too high for any given (e, ι, p), as mentioned in Section 4.2.1. The SNR of a signal $h = h_I + ih_{II}$ is given by (2.54), with the noise-weighted inner product (2.53) written in terms of the one-sided noise power spectral density $S_n(f > 0) := 2\mathcal{S}_n(f)$ as [87]

$$\langle a|b\rangle = 2\int_0^\infty df\,\frac{\tilde{a}(f)\tilde{b}^*(f) + \tilde{a}^*(f)\tilde{b}(f)}{S_n(f)} = 4\Re\left[\int_0^\infty df\,\frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\right]. \tag{4.46}$$

In this chapter, $S_n(f)$ is taken to be a LISA noise model for the L6A5 (six links, five-million-kilometre arms) configuration known as classic LISA, assuming the original mission requirements and including confusion noise from the foreground of Galactic white-dwarf binaries [159].

At a luminosity distance of 5 Gpc, an NK signal (sampled at 0.2 Hz) from the example EMRI described above has a one-year SNR of $\rho = 30.8$. When using the AAK model to generate the signal, a comparable value of $\rho = 32.1$ is obtained. However, an AK signal from the same EMRI has $\rho = 57.8$. To illustrate why this is the case, we consider the characteristic strain h_c of a signal and the noise amplitude h_n; these are given respectively by [26]

$$h_c(f) = 2f|\tilde{h}(f)|^2, \tag{4.47}$$

$$h_n(f) = \sqrt{fS_n(f)}, \tag{4.48}$$

such that

$$\rho^2 = \int_{-\infty}^\infty d(\ln f)\,\left(\frac{h_c(f)}{h_n(f)}\right)^2. \tag{4.49}$$

With these definitions, the area between h_c and h_f on a log–log plot gives an indication (but not an approximation) of SNR, and allows the relative detectability of signals to be estimated. The characteristic strain for the three signals and the LISA noise amplitude are shown in Figure 4.2, where the excess power in the AK signal at higher frequencies is evident, along with the consequent

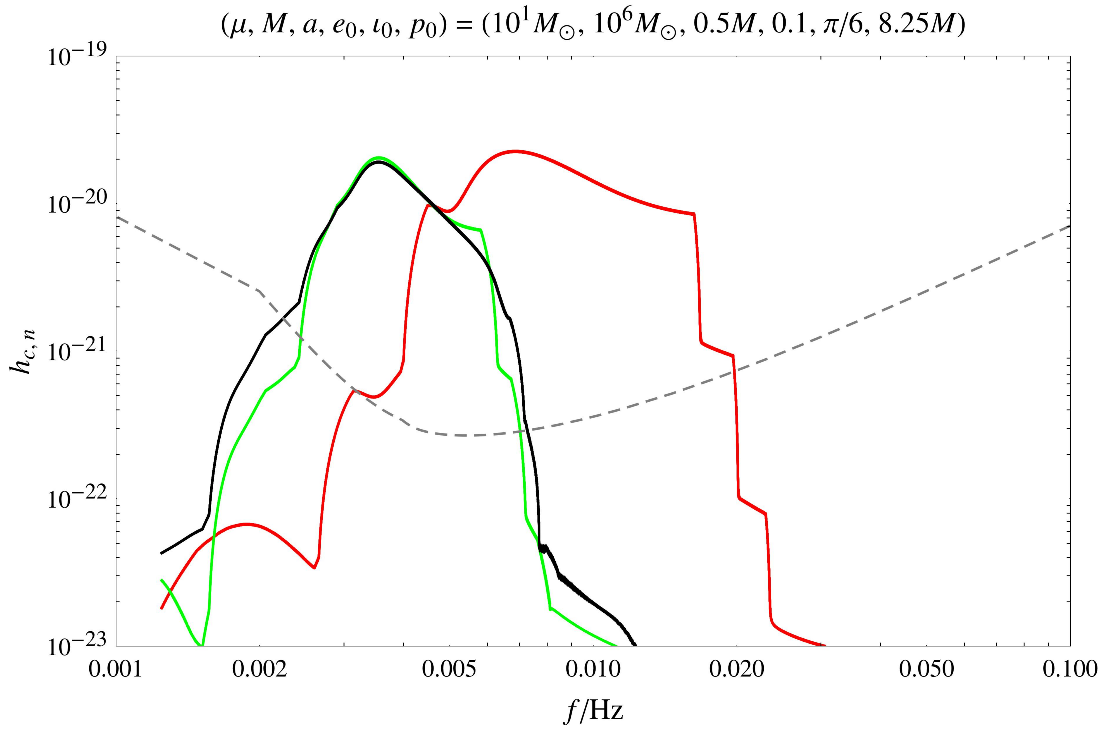

Figure 4.2: Characteristic strain for year-long AK (red), AAK (green) and NK (black) signals from the example EMRI, along with the noise amplitude (dashed) for the LISA configuration L6A5. A moving-average filter has been applied to the h_c curves, such that oscillations are smoothed out for ease of visualisation while the overall spectral profile is preserved.

$(\mu/M_\odot, a/M, e_0)$	$\mathcal{O}(\cdot\|h_{\rm NK})_{2\,\rm mth}$		$\mathcal{O}(\cdot\|h_{\rm NK})_{6\,\rm mth}$		σ
	$h_{\rm AK}$	$h_{\rm AAK}$	$h_{\rm AK}$	$h_{\rm AAK}$	
$(10^1, 0.5, 0.1)$	2.0×10^{-3}	9.5×10^{-1}	6.4×10^{-4}	1.5×10^{-1}	0.89
$(10^0, 0.5, 0.1)$	4.4×10^{-5}	5.9×10^{-1}	2.7×10^{-6}	1.3×10^{-1}	0.92
$(10^1, 0.8, 0.1)$	-1.6×10^{-3}	9.3×10^{-1}	-2.7×10^{-4}	1.3×10^{-1}	0.91
$(10^1, 0.5, 0.5)$	8.6×10^{-3}	2.3×10^{-1}	-2.0×10^{-3}	4.3×10^{-2}	0.38

Table 4.3: Comparison of the AK and initial AAK models via two- and six-month waveform overlaps $\mathcal{O}(\cdot|h_{\rm NK})$ for generic EMRIs with different compact-object mass, black-hole spin and initial eccentricity. The computational speed-up σ of both models over the NK model is given as well.

boost to SNR. This error is likely to persist for $M \gtrsim 10^6 M_\odot$, but may be mitigated for less massive central black holes as the maxima of the three h_c curves are blueshifted past the minimum of h_n.

For the purposes of this work (where the NK model is taken as fiducial), the phase accuracy of the AK and AAK models is assessed by how well their waveforms overlap with NK waveforms. In [59], the overlaps $\mathcal{O}(h_{\rm AK}|h_{\rm NK})$ and $\mathcal{O}(h_{\rm AAK}|h_{\rm NK})$ over two and six months are computed for the example EMRI, as well as for the same source with (i) $\mu = 10^0 M_\odot$, (ii) $a = 0.8M$ and (iii) $e_0 = 0.5$. The AK and initial AAK models have virtually identical computation times τ, and are both quicker than the NK model with typical speed-up factors of $\sigma := 1 - \tau/\tau_{\rm NK} \approx 0.9$ (except in the case of $e_0 = 0.5$, where $\sigma \approx 0.4$). However, the AAK model yields overlaps that are consistently higher, and by 2–3 orders of magnitude in most cases (see Table 4.3).

The speed-up factors for the AAK model are preserved by the present implementation, while its overlaps are increased across the board due to the enhanced fitting algorithm. Most notably, there is substantial improvement for EMRIs with higher initial eccentricities, although this is partly attributable to the use of a higher sample rate than that in [59] (around 0.03 Hz). The overlap and timing performance of the AAK model across 2–6 months with varying compact-object mass, black-hole spin and initial eccentricity is summarised by the plots in Figures 4.4–4.6.

In Figure 4.4, the overlap $\mathcal{O}(h_{\rm AAK}|h_{\rm NK})$ is computed for the example EMRI, as well as for the same source with $0.5 \leq \lg{(\mu/M_\odot)} \leq 2$. The minimum value of

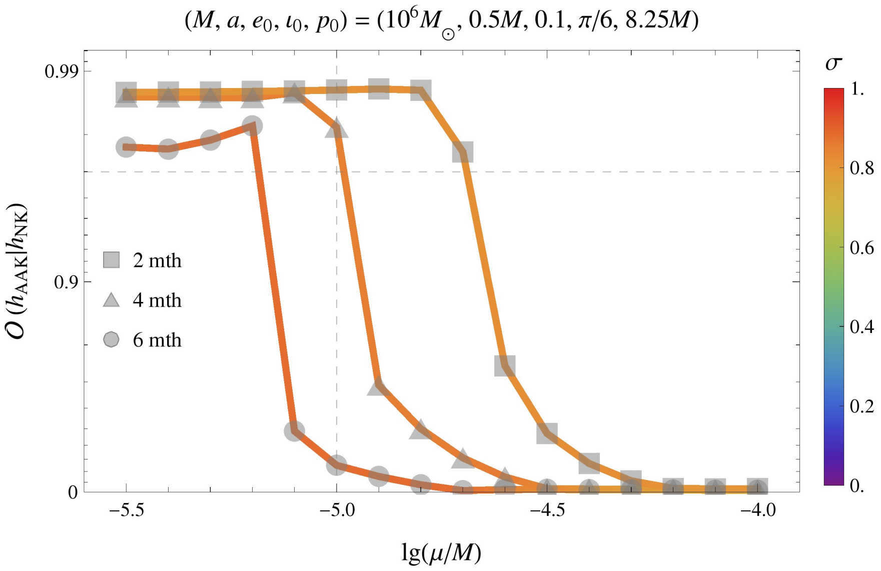

Figure 4.4: Two- to six-month overlaps between AAK and NK waveforms for generic EMRIs with varying compact-object mass. The computational speed-up σ of the AAK over the NK is coded by colour. Vertical dashed lines correspond to the example EMRI, while horizontal ones indicate the standard minimal-match value of 0.97 for template banks.

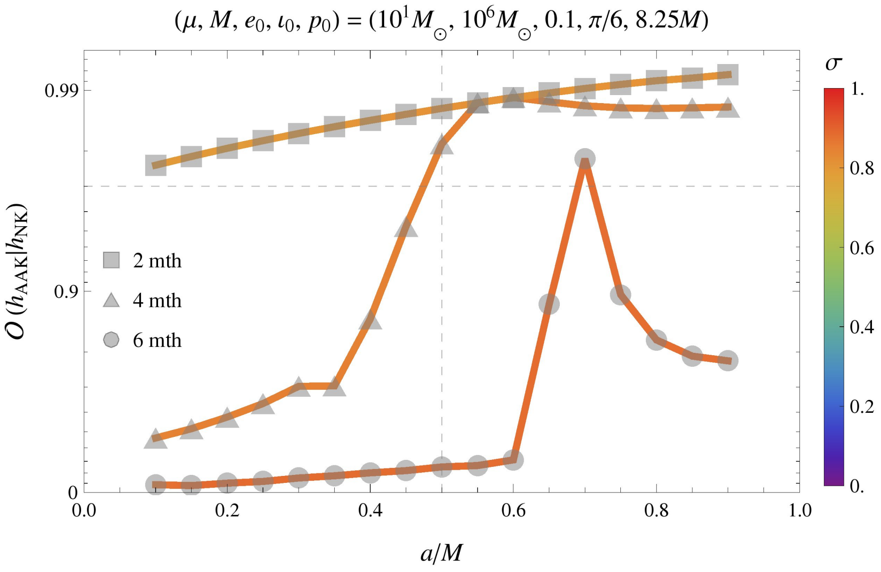

Figure 4.5: Overlaps and computational speed-up as in Figure 4.4, but for generic EMRIs with varying black-hole spin.

$\mu \approx 3M_\odot$ is chosen since the upper bound of $10T_{\mathrm{RR}}$ for T_{fit} exceeds six months for a less massive compact object, and so the overlaps in that regime will not show significant improvement (even the six-month overlap is already > 0.97). We also do not consider intermediate-mass black holes with $\mu > 100M_\odot$. Instead of choosing p_0 such that the EMRI plunges after one year (as done in [59]), we consider fixed $p_0 = 8.25M$ in this analysis; this leads the overlaps to degrade at larger rather than smaller mass ratios μ/M. The two-month overlaps are > 0.97 up to $\mu \approx 20M_\odot$, for which plunge occurs at around 5.6 months. There is greater speed-up over the NK model for longer waveform durations as expected, but all values of σ are $\gtrsim 0.8$.

Overlaps for the example EMRI with varying spin $0.1 \leq a/M \leq 0.9$ are shown in Figure 4.5. Again, all values of σ are $\gtrsim 0.8$, with greater speed-up for longer waveform durations. For fixed p_0, prograde EMRIs with lower spin start closer to plunge, and so the overlap values generally increase along with a. There appears to be an opposing effect at higher spin (possibly due to the additional degrees of freedom for error from fitting $\tilde{M}$ and $\tilde{a}$ in the AAK model) that causes a fall-off in the four- and six-month overlaps for $a \gtrsim 0.6$. The two-month overlaps are > 0.97 across the full range of considered spins, which is perhaps unsurprising since the upper bound of $10T_{\mathrm{RR}}$ for T_{fit} is also two months for a $(10^1, 10^6)M_\odot$ EMRI. However, we note here that while the phase accuracy of the AAK model may be arbitrarily increased in principle by taking $T_{\mathrm{fit}} > 10T_{\mathrm{RR}}$, the computational requirement that $N_{\mathrm{fit}} \lesssim 10$ will likely reduce the quality of the trajectory fit at early times.

The effect of varying eccentricity for the example EMRI is illustrated in Figure 4.6. We consider initial eccentricities $0.05 \leq e_0 \leq 0.5$, since there turns out to be no speed-up over the NK model ($\sigma \approx 0$) when generating a two-month AAK waveform with $e_0 = 0.5$. This is an important limitation of the Peters–Mathews approximation for waveform generation (see discussion in footnote 10), and will have to be addressed if the AAK model is to be useful in searches for high-eccentricity EMRIs. Nevertheless, the two- and even four-month overlaps are > 0.97 for $e_0 \lesssim 0.3$, with $\sigma \gtrsim 0.5$. As in the case of Figure 4.5, there is a peak in the six-month overlap; this is probably due to variance in the fit for e, and is unlikely to carry any fundamental significance.

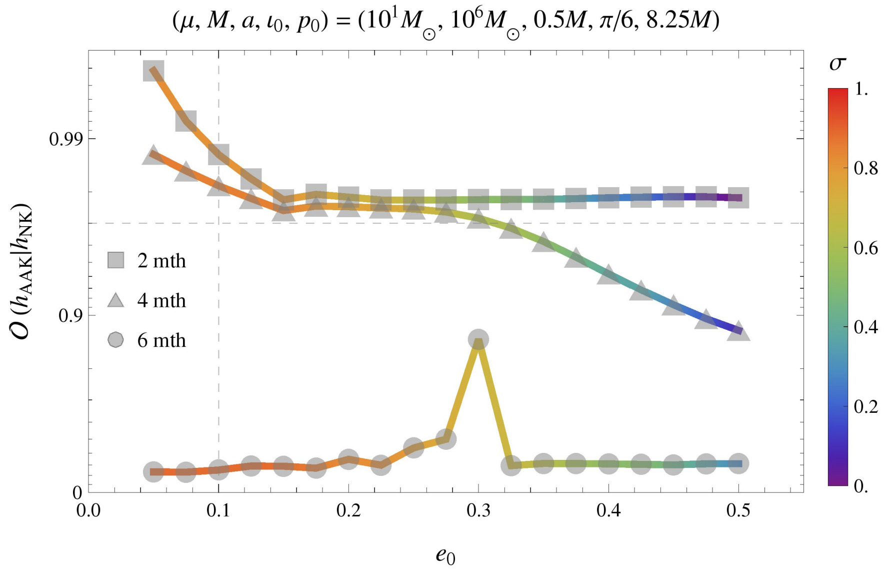

Figure 4.6: Overlaps and computational speed-up as in Figure 4.4, but for generic EMRIs with varying initial eccentricity.

Finally, we summarise here the results of a Fisher matrix calculation for a year-long AAK signal from the example EMRI considered throughout this section. The Fisher information matrix Γ for a GW signal h parametrised by $\boldsymbol{\lambda}$ is given by (2.61), where $\boldsymbol{\lambda} = (\mu, M, \mathbf{S}, \mathbf{E}, \mathbf{X}, \mathbf{R})$ in the case of EMRIs. For our AAK signal normalised to an SNR of $\rho = 30$, we find that the masses and spin can be measured to within the fractional root-mean-square errors

$$\Delta(\ln \mu) \approx 4 \times 10^{-5} \tag{4.50}$$

$$\Delta(\ln M) \approx 2 \times 10^{-5}, \tag{4.51}$$

$$\Delta(\ln (a/M)) \approx 4 \times 10^{-5}. \tag{4.52}$$

These errors are an order of magnitude better than the corresponding values for the AK model [57], and are more comparable to those cited for the NK model [66] (although the latter consider a circular, equatorial EMRI with $a = 0.9M$). They are also consistent with (the lower end of) the values reported in the L2/L3 mission selection proposal [14] for ESA's Cosmic Vision programme, where the AK model was used but with a modified plunge criterion.

4.4 Discussion

The widely used AK waveform model of Barack & Cutler is extremely quick to generate, but typically dephases within hours with respect to more accurate EMRI waveforms due to a mismatch of its radial, polar and azimuthal frequencies with the actual Kerr frequencies along its orbit. Hence using AK waveforms as signal templates for EMRIs will result in reduced SNRs for detection, as well as an inaccurate estimation of astrophysical parameters. This might limit their utility in scoping out data analysis issues for space-based GW detectors—an area where there are several important open questions that must be addressed within the next few years, in preparation for LISA mission adoption at the end of this decade.

We have proposed in this chapter an augmented AK waveform model that features a frequency map to the Kerr fundamental frequencies, self-consistent 3PN $O(e^6)$ evolution equations from Sago & Fujita, and calibration to the NK model of Babak et al. via a local phase-space trajectory fit that is extrapolated over the waveform duration. The present implementation of the model also includes various improvements such as amplitude corrections and fast plunge handling. AAK waveforms are virtually as quick to compute as their predecessors, but stay phase-coherent with NK waveforms for months and yield high overlap values (> 0.97) across extended regions of parameter space.

The most pressing deficiency in the AAK model is the ill-defined nature of (4.37)–(4.39) at the last stable orbit due to the divergent Kerr frequencies; this complicates both plunge detection and the specification of orbital parameters at plunge. Another limitation is that the mode-sum approximation (4.6) and (4.7) becomes more expensive than the quadrupole formula itself at high eccentricities ($e_0 \gtrsim 0.5$). If the model can be made robust through the resolution of these issues, it may also be upgraded with fits to improved NK models when they become available (e.g. with Kerr self-force evolution included through the framework in [67, 68]). With the era of space-based detectors drawing closer, the AAK waveform model will be an essential tool for near-future work on LISA data analysis and mock data challenges, as well as a useful addition to the inventory of EMRI waveforms for millihertz GW astronomy.

Chapter 5

Template bank compression

One strategy for reducing the online computational cost of matched-filter detection searches for GWs is to introduce a compressed basis for the waveform template bank. In this chapter, we propose and investigate several combinatorial compression schemes for a general template bank. Through offline compression, these tunable schemes are shown to yield faster detection and localisation of signals, along with moderately improved sensitivity and accuracy over coarsened banks at the same level of computational cost.

The general method is inspired by a recently proposed template bank compression scheme that potentially yields significant cost savings, but may be demonstrated to suffer from a number of fundamental problems. Our compression schemes address and rectify these problems; they might be useful for any search involving template banks, and especially in the analysis of data from space-based detectors, where online searches will be difficult due to the long-duration waveforms and large parameter spaces of some sources.

The material in this chapter has been adapted from [160].

5.1 Background

Various strategies exist to reduce the online cost of evaluating matched-filter inner products in GW detection algorithms, typically by shifting the computational burden to the preparatory offline stage. Some methods focus on making individual inner products computationally cheaper: this may be achieved across regions of parameter space through direct template interpolation [161, 162], or more generally by using a reduced-order quadrature [163].

Other methods seek to reduce the number of inner products in a grid-based search through template bank compression, i.e. the reduced-basis representation of a large template bank by a smaller set of templates [164–166].

In a recently proposed method of template bank compression [167], binary labelling is used to define a non-orthogonal basis that maximises compression losslessly (in the sense of perfect signal recovery without noise). This idea is fully general and admits a much higher compression rate than existing methods based on the eigenvalue structure of the template bank, but comes with significant penalties to detection sensitivity and localisation accuracy in the presence of detector noise. The method as originally described also suffers from an arbitrarily asymmetric treatment of templates, as well as a restrictive level of compression that limits its practicality to high-SNR signals. While the binary labelling method might be useful in the context of LISA (where source SNRs are potentially higher than for ground-based detectors), its practical applicability to GW data analysis remains undeveloped and hence unclear.

In this chapter, we introduce and develop the related method of conic compression (i.e. defining a compressed basis through conic combinations of templates) by characterising its performance under various simplifying assumptions, before investigating its viability for present and future GW detectors with a more realistic example. We propose several compression schemes, one of which subsumes a symmetric-treatment version of the binary labelling method as a particular case. These tunable schemes feature discrete transitions between zero and maximal compression, and offer fast detection and localisation of GW signals in the search space with a controlled loss (if at all) in sensitivity/accuracy. Their generality and straightforward implementation also allow them to supplement existing grid-search methods, or to rapidly identify seed points for stochastic searches in parameter estimation.

The general method of conic compression is set out in Section 5.2. Three families of conic compression schemes are then proposed in Sections 5.2.1–5.2.3: a lossy scheme based on partitions of the template bank, and two lossless schemes whose conic combinations are determined permutatively or by base representations of template labels. We calculate the optimal detection statistics for these schemes, and find that the standard maximum-overlap statistic is significantly suboptimal for detection in the lossless case. Section 5.2.4 compares the three schemes under simplified conditions, i.e. assuming the GW signal is proportional to a single template in an orthogonal template bank. The lossy partition scheme is shown to have slightly better detection sensitivity than its lossless counterparts at the same level of compression. Furthermore, while the lossless schemes provide automatic identification (i.e. localisation to a single template) of the signal upon detection, the identification accuracy falls off more rapidly with compression in the presence of noise.

We focus exclusively on the partition scheme in Section 5.3, where the orthogonality and single-template assumptions are lifted separately. As shown in Section 5.3.1, the overall performance of the scheme is partition-dependent in the case of a correlated (non-orthogonal) template bank, and must be pre-optimised by grouping highly correlated templates together. The optimised partition scheme retains the benefits of a correlated template bank up to high levels of compression, and is superior to a simple coarsening of the template bank (obtained by increasing the maximal mismatch between neighbouring templates). Section 5.3.2 discusses the case of a GW signal lying in a low-dimensional subspace of an orthogonal template bank, for which the detection sensitivity of the scheme is not significantly reduced.

In Section 5.4, we implement the optimised partition scheme for a highly correlated (maximal mismatch ≈ 0.01) template bank of $\sim 10^4$ PN waveforms, which describe only the inspiral phase of a comparable-mass binary coalescence. The scheme is shown to be viable for practical applications, as it performs well on this example up to high levels of compression and at all considered values of SNR. Its detection rate for a signal injected centrally is superior to that of the coarsening approach (especially at $\gtrsim 80\%$ compression), and this improvement is even more marked for a signal injected at the boundary of the bank. In addition, the accuracy rate for localisation of the injection to a $\lesssim 0.1\%$ region of the search space is undiminished up to $\sim 90\%$ compression, and is again higher than that of the coarsening approach.

The speed-up and enhanced accuracy in localising the GW signal with conic compression is promising for LISA data analysis, where the online use of template banks will be challenging for many sources (e.g. EMRIs with their large parameter space [168]). While the long duration of LISA signals is computationally prohibitive to fully coherent detection searches even with compression, our method is suitable for the shorter semi-coherent searches that are required for rapid electromagnetic follow-up. Conic compression might also provide an alternative to the singular-value-decomposition method used in LIGO detection pipelines for stellar-mass binary coalescences [165]: it scales well with parameter-space dimensionality, and easily matches or surpasses the order-of-magnitude computational savings from that method.

5.2 Conic compression

We consider a generic bank of N waveform templates h_n, where the template labels n are drawn from the collection $\mathbf{N} := \{n \in \mathbb{Z}^+ \,|\, n \leq N\}$, and the templates have been normalised with respect to the noise-weighted inner product (2.53) such that $\langle h_n | h_n \rangle = 1$ for all $n \in \mathbf{N}$. The inner products of the data x and the templates define N associated statistics

$$x_n := \langle x | h_n \rangle, \tag{5.1}$$

which may be used for detection and localisation in a simple grid search.

Our general method of compression is to reduce the number of statistic evaluations from N to M by considering conic (i.e. positive-coefficient) combinations of the original templates. The template labels are grouped into M sets $\mathbf{U}_m$, where the set labels m are drawn from the collection $\mathbf{M} := \{m \in \mathbb{Z}^+ \,|\, m \leq M\}$, and the sets satisfy $\bigcup_{m\in\mathbf{M}} \mathbf{U}_m = \mathbf{N}$. These sets define M conic templates

$$H_m := \sum_{n\in\mathbf{U}_m} h_n, \tag{5.2}$$

which are prepared offline (like the template bank itself), along with M associated statistics

$$X_m := \langle x | H_m \rangle = \sum_{n\in\mathbf{U}_m} x_n, \tag{5.3}$$

which are evaluated at the online stage.

Without any prior assumptions on the template bank, each template must be treated equally. This is done by ensuring that:

(a) each combination is weighted equally;

(b) each combination includes the same number of templates;

(c) each template is included in the same number of combinations.

Definition (5.2) has been chosen to satisfy Condition (a), while Condition (b) is imposed by further requiring $\operatorname{card}(\mathbf{U}_m) = \operatorname{card}(\mathbf{U}_{m'})$ for all $m, m' \in \mathbf{M}$ (where the set cardinality $\operatorname{card}(\mathbf{S})$ is the number of elements in the set $\mathbf{S}$). Condition (c) must be enforced separately in the construction of the sets. The second equality in Definition (5.3) relates the conic statistic evaluations to the original statistics (5.1), which are no longer evaluated at the online stage.

To simplify analysis, we first assume the template bank is an orthogonal set such that

$$\langle h_n | h_{n'} \rangle = \delta_{nn'}, \tag{5.4}$$

where $\delta_{nn'}$ is the Kronecker delta. We further assume the GW signal (if present) lies in the one-dimensional subspace spanned by a single template in Hilbert space, i.e.

$$x = Ah_1 + \{\text{noise}\}, \tag{5.5}$$

where $A > 0$ and the templates have been relabelled without loss of generality. It follows from (2.54) and (5.5) that $A = \rho$ is the source SNR. These orthogonal and 1-D restrictions are neither realistic nor optimal, but facilitate the analytic assessment and comparison of various compression schemes in this section. The overall performance of conic compression is generally improved by the lifting of these assumptions, which we consider in Sections 5.3 and 5.4.

In the presence of a GW signal, the expectation values and covariances of the normally distributed original statistics (5.1) are now given by

$$\mathbb{E}[x_n] = A\langle h_1 | h_n \rangle = A\delta_{1n}, \tag{5.6}$$

$$\operatorname{cov}(x_n, x_{n'}) = \langle h_n | h_{n'} \rangle = \delta_{nn'}. \tag{5.7}$$

As the labelling of templates is itself a probabilistic process with discrete uniform distribution, the original statistic vector x has the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ (with $\mu_n^{(i)} = \mathbb{E}[x_n]$ and $\Sigma_{nn'} = \mathrm{cov}(x_n, x_{n'})$), but summed over the N possible assignments i of $1 \in \mathbf{N}$ and renormalised accordingly. If the signal is absent, the distribution of x is simply $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Hence we have

$$p_1(\mathbf{x}) \propto \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_{(i)}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{(i)}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(i)}\right), \tag{5.8}$$

$$p_0(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right), \tag{5.9}$$

where p_1 and p_0 are the probability density functions of x in the respective presence or absence of a GW signal.

An optimal detection region $\mathcal{R}$ in Hilbert space maximises the detection rate $P_D = \int_{\mathcal{R}} p_1$ subject to a given false alarm rate $P_F = \int_{\mathcal{R}} p_0$; hence $p_1 = \lambda p_0$ on its boundary $\partial\mathcal{R}$ for some Lagrange multiplier λ. Using (5.6)–(5.9), we define the optimal detection statistic

$$x_{\mathrm{opt}} := \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \frac{1}{N} \exp\left(-\frac{A^2}{2}\right) \sum_{n\in\mathbf{N}} \exp\left(Ax_n\right), \tag{5.10}$$

such that the optimal detection surfaces $\partial\mathcal{R}$ are precisely the level sets of x_{opt} parametrised by λ, and a detection is claimed if x_{opt} exceeds the threshold λ_T corresponding to some fixed value of P_F.

In deriving (5.10), we have implicitly assumed a population of GW sources with equal likelihood and known signal amplitude. Equation (5.10) therefore defines the optimal statistic for detecting events drawn from such a population. For a population of sources that are not equally likely, we need to replace the sum in (5.10) with a suitably weighted sum. Similarly, for a population with a distribution of amplitudes, we need to marginalise (5.10) over A; in the case of an (improper) uniform prior over the interval $(-\infty, \infty)$, this would give a detection statistic proportional to $\sum_{n\in\mathbf{N}} \exp(x_n^2/2)$.

Any choice of population makes assumptions about the astrophysical distribution of GW sources that might not be justified. In this work, the focus is

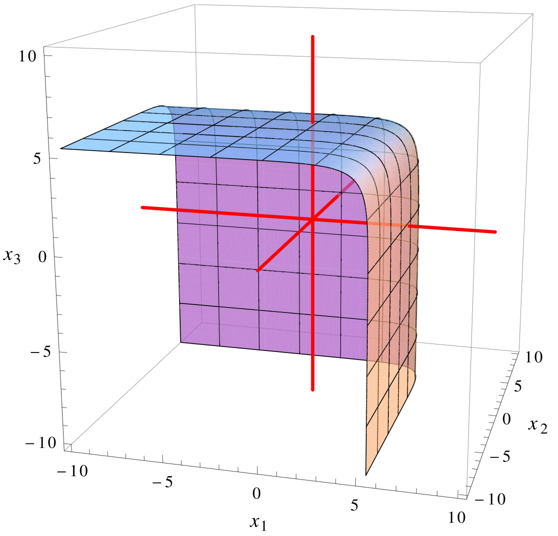

Figure 5.1: Three-dimensional projection of optimal detection surface for uncorrelated statistics x_n, at true SNR of $\rho = 2$.

on the investigation and comparison of template bank compression schemes, and so we only consider the equal-likelihood and known-amplitude population assumed in the derivation of (5.10). While the treatment of amplitude in particular is artificial, a search that is optimised for sensitivity to signal amplitudes around the detection threshold will likely be near-optimal for any given astrophysical population (and closer to optimality than a search tuned for the wrong astrophysical population). Finally, we note that although (5.10) has been derived as a frequentist optimal statistic, the same equation also arises as the Bayes factor for the presence (versus absence) of a signal, assuming flat model priors and the source population assumptions outlined above.

For sufficiently high SNR (large A), the optimal surfaces $x_{\text{opt}} = \lambda$ defined by (5.10) are well approximated by semi-infinite hypercubes (see Figure 5.1), i.e. the level sets of the standard maximum-overlap detection statistic [169]

$$x_{\max} := \max_{n \in \mathbf{N}} \{x_n\}. \tag{5.11}$$

Since the original statistics (5.1) are uncorrelated, the probability density functions of $x_{\max}$ in the presence or absence of a GW signal are obtainable explicitly.

These are given respectively by

$$q_1(x_{\max}) = F_0(x_{\max})^{N-1} f_1(x_{\max}) + (N-1) F_0(x_{\max})^{N-2} F_1(x_{\max}) f_0(x_{\max}), \quad (5.12)$$

$$q_0(x_{\max}) = N F_0(x_{\max})^{N-1} f_0(x_{\max}), \quad (5.13)$$

where $f_s(x_{\max})$ is the probability density function for the Gaussian distribution $\mathcal{N}(sA, 1)$, and $F_s(x_{\max})$ is the cumulative distribution function

$$F_s(x_{\max}) = \int_{-\infty}^{x_{\max}} du\, f_s(u). \quad (5.14)$$

For our analysis of conic compression schemes, we also require the expectation values and covariances of the normally distributed conic statistics (5.3). From (5.3), (5.6) and (5.7), it follows in the presence of a GW signal that

$$\mathbb{E}[X_m] = \sum_{n \in \mathbf{U}_m} \mathbb{E}[x_n] = A \operatorname{card}(\{1\} \cap \mathbf{U}_m), \quad (5.15)$$

$$\operatorname{cov}(X_m, X_{m'}) = \sum_{n \in \mathbf{U}_m} \sum_{n' \in \mathbf{U}_{m'}} \operatorname{cov}(x_n, x_{n'}) = \operatorname{card}(\mathbf{U}_m \cap \mathbf{U}_{m'}), \quad (5.16)$$

where the cardinalities are determined by the choice of compression scheme.

As before, the conic statistic vector $\mathbf{X}$ has the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ (now with $\mu_m^{(i)} = \mathbb{E}[X_m]$ and $\Sigma_{mm'} = \operatorname{cov}(X_m, X_{m'})$), but summed over the N possible assignments of $1 \in \mathbf{N}$ and renormalised accordingly. If the signal is absent, the distribution of $\mathbf{X}$ is again $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The probability density functions of $\mathbf{X}$ in the presence or absence of a GW signal are then given respectively by (5.8) and (5.9) with $\mathbf{x} \equiv \mathbf{X}$.

We now propose and investigate three general conic compression schemes in Sections 5.2.1–5.2.3, before comparing their performance and potential applicability in Section 5.2.4. The orthogonal and 1-D restrictions (5.4) and (5.5) are assumed throughout Section 5.2.

5.2.1 Partition scheme

The simplest method of grouping the template labels n is to take the family of sets $\mathbf{U}_m$ as a partition of $\mathbf{N}$, i.e. $\mathbf{U}_m \cap \mathbf{U}_{m'} = \emptyset$ for all distinct $m, m' \in \mathbf{M}$. Condition (c) is then automatically satisfied, while Condition (b) defines the set cardinality $P = \mathrm{card}(\mathbf{U}_m)$ for all $m \in \mathbf{M}$. It follows that $M = N/P$.

For the comparison of schemes in Section 5.2.4, it is useful to introduce a compression parameter $K \in \mathbb{Z}^+$ for each scheme, which determines the compression rate

$$\kappa := 1 - \frac{N_{\mathrm{eval}}}{N}, \tag{5.17}$$

where $N_{\mathrm{eval}} = M$ is the required number of statistic evaluations (for detection or localisation purposes). This generates a sliding scale of groupings that ranges from no compression at $K = 1$ to maximal compression at some scheme-dependent value of K. We may clearly choose $K = P$ for the partition scheme, such that maximal compression is given by $K = N$. The minimal nontrivial compression is 50% at $K = 2$, while there are diminishing returns at large K since $\kappa(K)$ is concave-down.

From (5.15) and (5.16), we now have

$$\mathbb{E}[X_m] = A\delta_{1m}, \tag{5.18}$$

$$\mathrm{cov}(X_m, X_{m'}) = P\delta_{mm'}, \tag{5.19}$$

where the sets have been relabelled such that $1 \in \mathbf{U}_1$ without loss of generality. Again considering the N possible assignments of $1 \in \mathbf{N}$, the optimal detection statistic $X_{\mathrm{opt}} := p_1(\mathbf{X})/p_0(\mathbf{X})$ follows from (5.8) and (5.9) (with $\mathbf{x} \equiv \mathbf{X}$) as

$$X_{\mathrm{opt}} = \frac{1}{M} \exp\left(-\frac{A^2}{2P}\right) \sum_{m \in \mathbf{M}} \exp\left(\frac{A}{P} X_m\right). \tag{5.20}$$

Since the conic statistics for the partition scheme remain uncorrelated, the optimal surfaces $X_{\mathrm{opt}} = \lambda$ resemble that in Figure 5.1, and in lieu of (5.20) it is valid to consider the maximum-overlap detection statistic

$$X_{\max} := \max_{m \in \mathbf{M}} \{X_m\}. \tag{5.21}$$

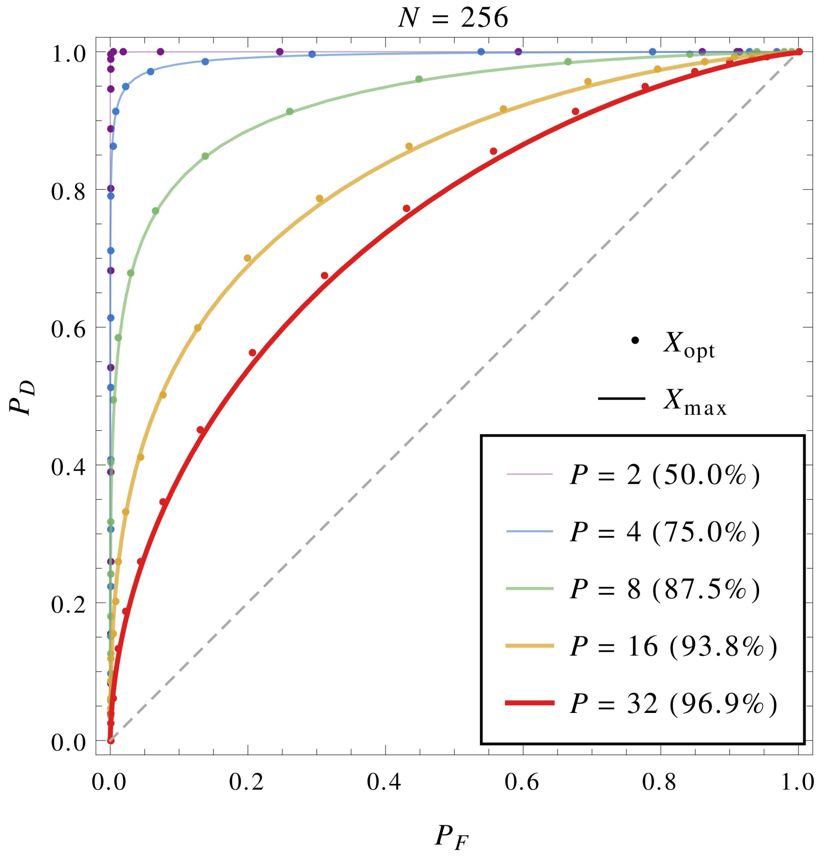

Figure 5.2: ROC curves for the partition scheme's optimal and maximum-overlap detection statistics, at different values of set cardinality P (with compression rate κ in parentheses) for $N = 256$ and a true SNR of $\rho = 10$. The dashed diagonal line indicates the worst possible performance, i.e. a random search for which the detection and false alarm rates are equal.

Plots of detection rate P_D against false alarm rate P_F for a detection statistic are known as receiver operating characteristic (ROC) curves [161]. ROC curves for both the optimal and maximum-overlap statistics are compared in Figure 5.2.[13] With increased compression, the performance of the maximum-overlap statistic falls away slightly from that of the optimal statistic, due to the lowering of effective SNR $A/\sqrt{P}$ in (5.20); nevertheless, (5.21) is a sound approximation as both sets of curves show good overall agreement.

For the partition scheme to admit a useful (i.e. populated) sliding scale of compression rates, the template bank might need to be trimmed or padded such that N has as many divisors as possible. Fixing the false alarm rate and choosing either a desired detection rate or a compression rate then allows advance determination of the conic templates (5.2) and the threshold λ_T, which is the value of λ corresponding to the fixed false alarm rate. The algorithm for detection follows as: (i) evaluate the conic statistics (5.3); (ii) claim a detection if $X_{\max} > \lambda_T$. Threshold and detection SNRs for the maximum-overlap statistic may be defined respectively as

$$\rho_T := \frac{\lambda_T}{\sqrt{\mathrm{var}(X_{\max})}}, \tag{5.22}$$

$$\rho_D := \frac{X_{\max}}{\sqrt{\mathrm{var}(X_{\max})}}. \tag{5.23}$$

An extension of the detection algorithm is required for signal identification (i.e. localisation to a single template), since the simple coarse-graining of partition compression does not distinguish between template labels in the same set. The signal is most likely to be associated with the largest conic statistic evaluation $X_{(1)}$, so the best candidate may be obtained by further evaluating all of the original statistics x_n contributing to $X_{(1)}$ and identifying the largest. This finer level of evaluations increases the computational cost by P to $N_{\mathrm{eval}} = M + P$.

For better identification accuracy at lower SNRs, we may widen our search to the i largest X_m instead, at an added computational cost of iP. The standard identification algorithms I_i for the partition scheme follow (after detection) as:

[13] The curves for (5.20) were obtained via 10^5-trial Monte-Carlo simulations, while numerical integration of (5.12) and (5.13) was used to generate the curves for (5.21).

(iii) evaluate the original statistics (5.1) for all $n \in \mathbf{V}_i$, where

$$\mathbf{V}_i := \bigcup_{j=1}^{i} \mathbf{U}_{(j)}, \tag{5.24}$$

with $\mathbf{U}_{(j)}$ corresponding to the j-th largest conic statistic evaluation $X_{(j)}$; (iv) identify $\max_{n\in\mathbf{V}_i}\{x_n\}$.

Other identification algorithms may also be considered. One such alternative is obtained by defining a further partition of $\mathbf{V}_i$ into two sets and evaluating the associated conic statistics, then identifying the set $\mathbf{V}_i'$ corresponding to the larger statistic evaluation and repeating the process with $\mathbf{V}_i \equiv \mathbf{V}_i'$ until $\mathrm{card}(\mathbf{V}_i') = 1$. This method might be useful for large values of P; it yields a smaller added computational cost of $2\log_2 iP$, but incurs a penalty to identification accuracy since the early iterations still involve coarse-grained searches.

5.2.2 Symmetric base scheme

Without an additional fine-grained search, partition compression is lossy in the sense that the GW signal is not automatically identified in the limit of zero noise. A lossless conic compression scheme proposed by Wang [167] represents each template label n in binary and assigns it to the set $\mathbf{U}_m$ if its m-th digit is 1. This binary scheme features the largest possible lossless compression ($M = \log_2 N$) and an automatic identification of the GW signal; however, it suffers from an unequal treatment of templates (i.e. it violates Conditions (b) and (c)) and hence it yields an arbitrary level of performance that depends on the initial assignment of template labels. Furthermore, the restriction to maximum compression limits its usefulness in practical applications.

We propose a compression scheme modelled on Wang's binary scheme, but symmetrised (for equal treatment of templates) and generalised to a sliding scale of base representations (for tunable compression). The template labels n are represented modulo N in base B, and each set $\mathbf{U}_m \equiv \mathbf{U}_{k,b}$ is constructed by collecting all of the labels whose k-th digit is b (this includes $b = 0$, and gives a symmetric version of the binary scheme when $B = 2$). For Conditions (b) and (c) to be satisfied, we require $\log_B N \in \mathbb{Z}^+$; it follows that $M = B\log_B N$.

The compression parameter is chosen as $K = \log_B N$, such that maximal compression is given by $K = \log_3 N \approx \ln N$ (base-2 compression is slightly suboptimal with symmetrisation). In contrast to the partition scheme, compression for the symmetric base scheme is dependent on the size of the template bank; the minimal nontrivial compression for $N = 10^2$ is nearly 80% at $K = 2$ (base-$\sqrt{N}$ compression), and over 95% for $N = 10^4$.

From (5.15) and (5.16), we have[14]

$$\mathbb{E}[X_m] = A\delta_{0b}, \tag{5.25}$$

$$\operatorname{cov}(X_m, X_{m'}) = B^{K-2}(B\delta_{kk'}\delta_{bb'} - \delta_{kk'} + 1), \tag{5.26}$$

where $m = B(k-1) + b + 1$, and the templates have been relabelled such that (5.5) becomes $x = Ah_N + \{\text{noise}\}$ without loss of generality. Considering the N possible assignments of $N \in \mathbf{N}$, the optimal detection statistic follows from (5.8) and (5.9) as

$$X_{\text{opt}} = \frac{1}{N} \exp\left(-\frac{\beta K A^2}{2} + (\beta - \alpha)A\,\operatorname{tr}(\mathbf{X})\right) \prod_{k=1}^{K} \sum_{b=0}^{B-1} \exp\left(\alpha A X_{k,b}\right) \tag{5.27}$$

with

$$\alpha := \frac{B}{N}, \tag{5.28}$$

$$\beta := \frac{M - K + 1}{NK}, \tag{5.29}$$

where $\operatorname{tr}(\mathbf{X}) := \sum_{m\in\mathbf{M}} X_m$.

The higher compression rates provided by the symmetric base scheme result from the non-empty intersections among the sets $\mathbf{U}_{k,b}$ with different values of k. As seen in (5.26), these also lead to correlations among the conic statistics $X_{k,b}$. The optimal detection surfaces given by $X_{\text{opt}} = \lambda$ differ significantly from that depicted in Figure 5.1; their projections onto the correlated subspaces are now compact hyperboloids, and no longer approach the semi-infinite hypercubes of the maximum-overlap detection statistic at high SNR (see Figure 5.3).

[14] The covariance matrix defined by (5.26) is rank-deficient, but we may take the Moore–Penrose pseudoinverse $\mathbf{\Sigma}^{+}$ as a perturbative approximation to $\mathbf{\Sigma}^{-1}$ in (5.8) and (5.9).

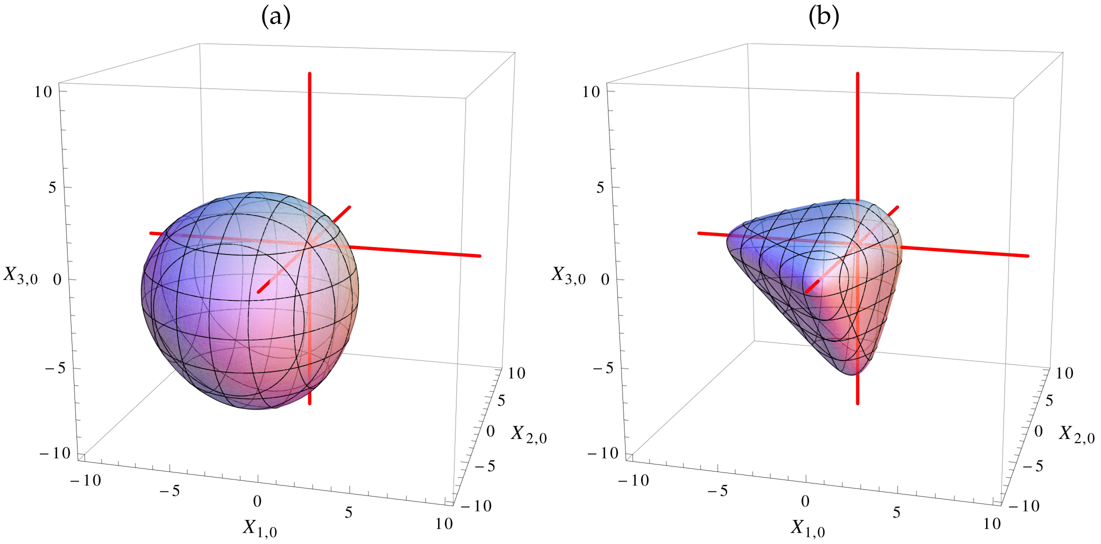

Figure 5.3: Three-dimensional projection of optimal detection surface for correlated statistics $X_{k,b}$ of symmetric base scheme with $N = 256$ and $B = 4$, at true SNRs of (a) $\rho = 10$ and (b) $\rho = 100$.

Without a simple approximation for the optimal detection statistic, the most feasible option is to use (5.27) itself with an estimate A of the true SNR. ROC curves for the estimated statistic $X_{\text{opt}}^{A=\epsilon\rho}$ with $\epsilon \in \{1/2, 2\}$ are compared against those for the optimal and maximum-overlap statistics in Figure 5.4. Not much detection sensitivity (for a fixed false alarm rate) is lost if the true SNR can be estimated to within a factor of two, while usage of the maximum-overlap statistic now incurs a more noticeable drop in performance as expected.

The restriction of N, B and K to integer values also results in more sparsely populated sliding scales than those admitted by the partition scheme. There are two possible compression rates for $N = 256$ (base-2 compression is suboptimal compared to $B = 4$), and three for $N = 6561 = 81^2 = 9^4 = 3^8$; most other values of N will admit only one or none. Notwithstanding the lack of tunability, a feasible strategy is to trim or pad the template bank such that N is a perfect square or cube, since the smallest values of K already yield high compression rates. The detection algorithm then follows as given in Section 5.2.1, with some estimated detection statistic $X_{\text{opt}}^{A=\epsilon\rho}$ in place of $X_{\max}$.

One key feature of the symmetric base scheme and other lossless methods

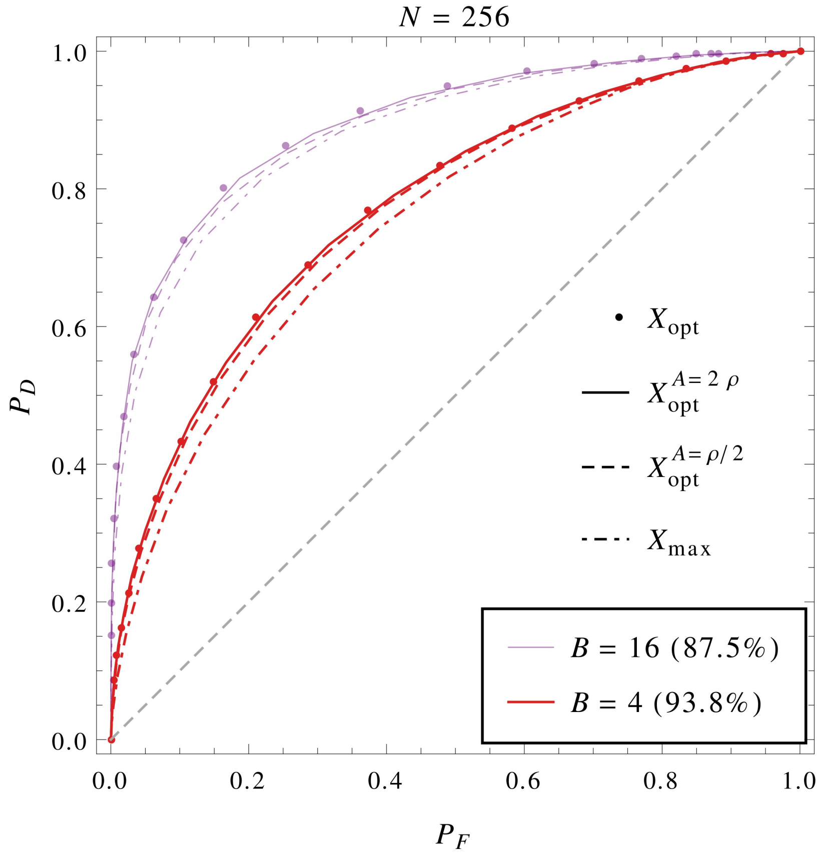

Figure 5.4: ROC curves for the symmetric base scheme's optimal, maximum-overlap and estimated detection statistics, at different values of base B (with compression rate κ in parentheses) for $N = 256$ and a true SNR of $\rho = 10$.

of compression is the automatic identification of the GW signal (upon detection). In the symmetric base scheme, the label of the identified template in base-B representation is given digit-wise by the largest conic statistic evaluation $X_{k,(1)} := \max_b\{X_{k,b}\}$ for each value of k. However, as each digit k is identified individually, the overall identification accuracy falls off severely with increasing K (i.e. the total number of digits).

A possible modification for better accuracy is to consider the $i + 1$ largest $X_{k,b}$ for each k and perform an additional fine-grained search over the $(i + 1)^K$ templates, which increases the computational cost accordingly. The standard identification algorithms I_i for the symmetric base scheme follow (after detection) as: (iii) evaluate the original statistics (5.1) for all $n \in \mathbf{V}_i$, where

$$\mathbf{V}_i := \bigcap_{k=1}^{K} \bigcup_{j=1}^{i+1} \mathbf{U}_{k,(j)}, \tag{5.30}$$

with $\mathbf{U}_{k,(j)}$ corresponding to the j-th largest conic statistic evaluation $X_{k,(j)}$ for each k; (iv) identify $\max_{n\in\mathbf{V}_i}\{x_n\}$. Automatic identification is recovered for $i = 0$, where steps (iii) and (iv) become unnecessary as $\mathrm{card}(\mathbf{V}_0) = 1$.

For large values of K (small values of B), these identification algorithms might still suffer from poor accuracy. One alternative algorithm is obtained by defining some threshold X_T and considering all conic statistic evaluations $X_{k,b} > X_T$, then performing the additional fine-grained search over all of the corresponding templates. Such a threshold may be set prior to data-taking; if $X_{k,(1)} < X_T$ for some value of k, the k-th digit of the number is unconstrained and templates corresponding to all possible choices of that digit are considered. Alternatively, X_T may be based on the data by setting $X_T = f \min_k\{X_{k,(1)}\}$ for some fixed fraction f, which ensures that at least one possible value is identified for each digit. Both approaches will in general yield increased accuracy, but they offer less control over the number of conic statistic evaluations considered and hence the overall computational cost.

5.2.3 Binomial coefficient scheme

The symmetric base labelling method is not the only construction of the sets $\mathbf{U}_m$ that preserves both lossless compression (automatic identification) and equal treatment of templates (Conditions (b) and (c)). In general, we may represent any assignment of N templates to M sets with a collection of N M-digit binary labels, where the m-th digit of each label is 1 if it appears in $\mathbf{U}_m$ and 0 otherwise. Condition (c) implies that each label must appear in exactly R sets, and hence contain exactly R 1's. In addition, Condition (b) defines the set cardinality $C = \operatorname{card}(\mathbf{U}_m)$ for all $m \in \mathbf{M}$, which yields the constraint $NR = MC$ (each of the N labels appears exactly R times across all sets, while each of the M sets contains exactly C labels). For some given integers $N \geq M \geq R$, this constraint is equivalent to the existence of

$$C = \frac{NR}{M} \in \mathbb{Z}^+, \tag{5.31}$$

which is both a necessary and sufficient condition for such a set construction to be possible [170].

We now require that the conic statistics (5.3) are correlated symmetrically, as seen in the partition scheme (but not the symmetric base scheme). This additional condition implies that the intersection of each pair of sets has fixed cardinality I, i.e. $\operatorname{card}(\mathbf{U}_m \cap \mathbf{U}_{m'}) = I$ for all distinct $m, m' \in \mathbf{M}$. Considering the family of all such intersections then yields the constraint $NR(R-1) = M(M-1)I$ (each of the N labels appears exactly ${}^{R}\mathrm{C}_2$ times across all intersections, while each of the ${}^{M}\mathrm{C}_2$ intersections contains exactly I labels). For some given integers $N \geq M \geq R$ and C satisfying (5.31), this constraint is equivalent to the existence of

$$I = \frac{NR(R-1)}{M(M-1)} = \frac{C(R-1)}{M-1} \in \mathbb{Z}^+, \tag{5.32}$$

which is a necessary (but not in general sufficient) condition for such a set construction to be possible.

The general construction of a family of sets under the constraints (5.31) and (5.32) is an open problem in combinatorial design theory (see Apppendix B). In this chapter, we restrict our focus to a special case that may be treated in greater detail. Every M-digit binary number with exactly R 1's is taken to represent a distinct template label; the set cardinality then equals the number of $(M-1)$-digit binary numbers with exactly $(R-1)$ 1's, while the intersection cardinality of each pair of sets equals the number of $(M-2)$-digit binary numbers with exactly $(R-2)$ 1's. Hence for all distinct $m, m' \in \mathbf{M}$, we have

$$N = {}^{M}\mathrm{C}_{R}, \tag{5.33}$$

$$C = {}^{M-1}\mathrm{C}_{R-1}, \tag{5.34}$$

$$I = {}^{M-2}\mathrm{C}_{R-2}, \tag{5.35}$$

such that (5.31) and (5.32) are satisfied. We refer to this as the binomial coefficient scheme. The usual ordering of the binary numbers gives a natural map onto the original label collection $\mathbf{N} = \{n \in \mathbb{Z}^{+} \,|\, n \leq N\}$, although the inverse map is analytically nontrivial (but straightforward in practice).

As the binomial coefficient scheme shares many similarities with the symmetric base scheme, we only highlight its key features in this section. The compression parameter is chosen as $K = R$, such that maximal compression is given by $K = \mathrm{cbc}^{-1}(N)/2$ (where $\mathrm{cbc}(M) := \Gamma(M+1)/\Gamma(M/2+1)^2$ is the continuous extension of the central binomial coefficient ${}^{M}\mathrm{C}_{M/2}$). Compression rates again depend on the size of the template bank; at small values of K, they are only slightly higher than those of the symmetric base scheme.

From (5.15) and (5.16), we have

$$\mathbb{E}[X_m] = A \sum_{r=1}^{R} \delta_{rm}, \tag{5.36}$$

$$\mathrm{cov}(X_m, X_{m'}) = {}^{M-2}\mathrm{C}_{R-2} \left(\frac{M-R}{R-1} \delta_{mm'} + 1 \right), \tag{5.37}$$

where the sets have been relabelled such that $1 \in \mathbf{U}_r$ for $1 \leq r \leq R$ without loss of generality. Considering the N possible assignments of $1 \in \mathbf{N}$, the optimal detection statistic follows from (5.8) and (5.9) as

$$X_{\text{opt}} = \frac{1}{N} \exp\left(-\frac{\beta K A^2}{2} + (\beta - \alpha) A \operatorname{tr}(\mathbf{X})\right) \sum_{i=1}^{N} \exp\left(\alpha A \sum_{r \in \mathbf{R}_i} X_r\right) \tag{5.38}$$

with

$$\alpha := \frac{1}{{}^{M-1}\mathrm{C}_{R-1}}, \tag{5.39}$$

$$\beta := \frac{1}{{}^{M-2}\mathrm{C}_{R-1}}, \tag{5.40}$$

where the sets $\mathbf{R}_i$ are the N distinct R-combinations of the collection $\mathbf{M}$.

All of the conic statistics X_m are correlated symmetrically, as seen in (5.37). Upon projection onto any three-dimensional subspace, the optimal detection surfaces given by $X_{\text{opt}} = \lambda$ resemble those in Figure 5.3 at both low and high SNRs. It follows that the maximum-overlap detection statistic is again an inadequate approximation to the optimal statistic, and we are compelled to use (5.38) itself (assuming an accurate estimate of true SNR). We do not include ROC curves for the binomial coefficient scheme here, as they are very similar to those in Figure 5.4.

A direct comparison of the base and binomial schemes is difficult, since there are few suitable values of N that are exactly valid for both schemes. Lack of tunability is also more of an issue for the binomial scheme: the only values of N that admit more than one nontrivial compression rate might be the Singmaster numbers [171] (which admit two as they appear six times in Pascal's triangle), and it is not known whether any number admits more than two (apart from $N = 3003 = {}^{78}\mathrm{C}_2 = {}^{15}\mathrm{C}_5 = {}^{14}\mathrm{C}_6$ [172]). The problem may be overcome by considering a more general compression scheme satisfying the conditions (5.31) and (5.32). This is beyond the scope of the current work due to the complexity of set construction (see Appendix B), but might be investigated for specific template banks in the future.

The detection algorithm for the binomial coefficient scheme is as given in Section 5.2.1, with some estimated detection statistic $X_{\text{opt}}^{A=\epsilon\rho}$ in place of $X_{\max}$. Automatic identification is available as well, with the label of the identified template given uniquely by the R largest conic statistic evaluations. For higher accurate-identification rates, a possible alternative is to consider the $R+i$ largest X_m and perform an additional fine-grained search over the ${}^{R+i}\mathrm{C}_R$ templates. The standard identification algorithms I_i for the binomial coefficient scheme follow (after detection) as: (iii) evaluate the original statistics (5.1) for all $n \in \mathbf{V}_i$, where

$$\mathbf{V}_i := \bigcup_{k=1}^{{}^{R+i}\mathrm{C}_R} \bigcap_{j\in\mathbf{J}_k} \mathbf{U}_{(j)}, \tag{5.41}$$

with $\mathbf{U}_{(j)}$ corresponding to the j-th largest conic statistic evaluation $X_{(j)}$ and the sets $\mathbf{J}_k$ given by the ${}^{R+i}\mathrm{C}_R$ distinct R-combinations of the set $\{j \in \mathbb{Z}^+ \,|\, j \leq R+i\}$; (iv) identify $\max_{n\in\mathbf{V}_i}\{x_n\}$. Automatic identification is recovered for $i=0$, where steps (iii) and (iv) become unnecessary as $\mathrm{card}(\mathbf{V}_0) = 1$.

We note that another possible scheme would be a "direct sum" of the partition scheme and either the symmetric base or binomial coefficient scheme. The collection of template labels is first partitioned into subcollections, each of which is further decomposed into smaller sets via one of the correlated schemes; these sets may also be recombined across the initial partition for increased compression. We do not consider this further here, but such an approach would overcome some of the difficulties associated with the restricted values of N for the base and binomial schemes.

5.2.4 Scheme comparison

In this section, we compare the performance of the uncorrelated partition scheme and its two correlated alternatives across three areas: compression, detection and identification (i.e. localisation to a single template). The detection and identification plots here (and throughout the rest of the chapter) were obtained using 10^5-trial Monte-Carlo simulations, and so the errors on each plot point are $\sim 10^{-3}$ for a 1-sigma binomial confidence interval.

Log–log plots of M against N for various conic compression schemes are shown in Figure 5.5, where the maximum lossless compression provided by the binary scheme [167] has also been included for reference. As alluded to in Sections 5.2.1–5.2.3, the partition scheme has the largest range of compression rates, both in terms of compression bounds (plot area) and admitted rates

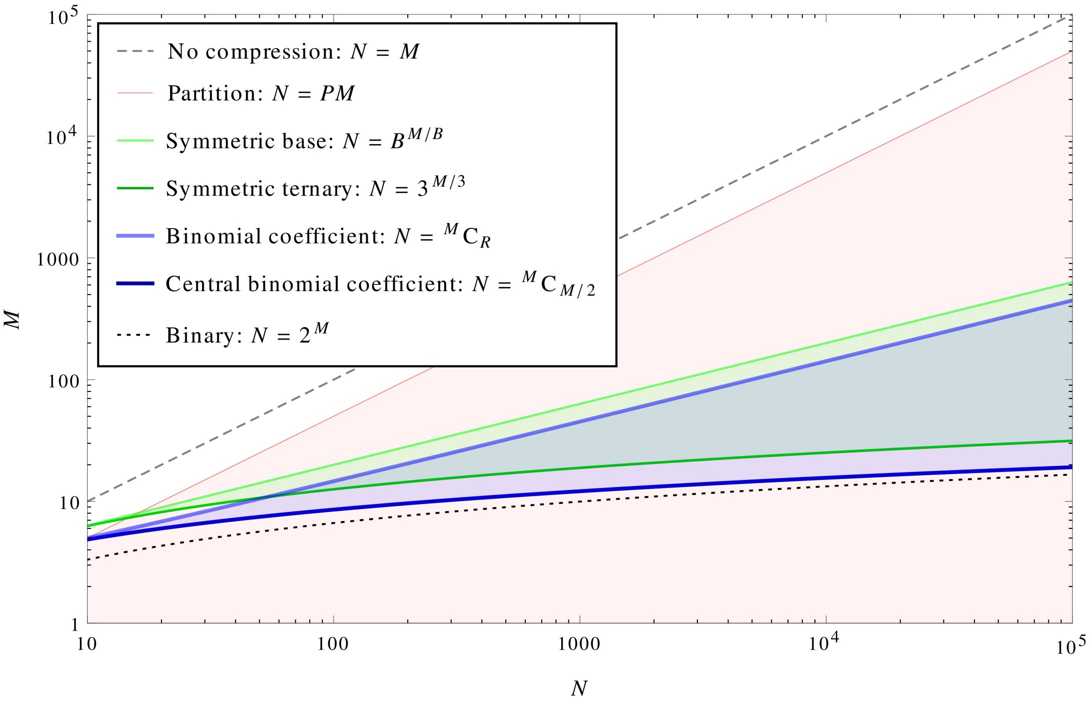

Figure 5.5: Log–log plots of M against N for various tunable/fixed conic compression schemes. For each tunable scheme, the corresponding shaded region indicates the range of possible compression rates (with the trivial compression setting $K = 1$ excluded). Not every point in this region is realisable in practice, as discussed in the text.

(discrete density, not shown). The two correlated schemes cover similar areas at lower densities in Figure 5.5, with the binomial coefficient scheme offering slightly greater compression.

Detection performance for each compression setting of a given scheme may be measured by detection sensitivity at a fixed false alarm rate (which is simply read off the corresponding ROC curve), or by a summary statistic that captures most of the information contained in an ROC curve (e.g. the area A_{ROC} under the curve). Since an ROC curve always lies above the no-discrimination line $P_D = P_F$, we define the discrimination

$$D := 2A_{\mathrm{ROC}} - 1, \tag{5.42}$$

which serves as a measure of how well the detection statistic discriminates between true and false positives. Figure 5.6(a) shows plots of discrimination against compression for the three proposed schemes at different values of true SNR, with $N \approx 256$. We use the maximum-overlap detection statistic in lieu of the optimal statistic for the partition scheme, and are compelled to choose $N = 210$ for the binomial coefficient scheme. The three schemes have comparable performance at lower SNRs, but the partition scheme begins to outpace its correlated alternatives as SNR increases.

To compare identification performance (after a true detection), we consider plots of accurate-identification rate P_I against compression, but only for the fastest standard algorithms of each scheme (i.e. I_1 for the partition scheme, and automatic identification I_0 for the correlated schemes). The rate P_I for each plot point is calculated using all and only the trials with the injected signal present, and therefore assumes perfect detection throughout ($P_D = 1$ and $P_F = 0$). This decouples identification from detection: it allows standardised comparison of the schemes at a fixed false alarm rate, and does not penalise the identification performance of any method for having inferior detection performance.

As seen in Figure 5.6(b), the usefulness of lossless compression and automatic identification is limited in the presence of noise; the addition of a simple fine-grained search to the partition scheme is enough to yield significantly higher identification accuracy at the cost of marginally lower compres-

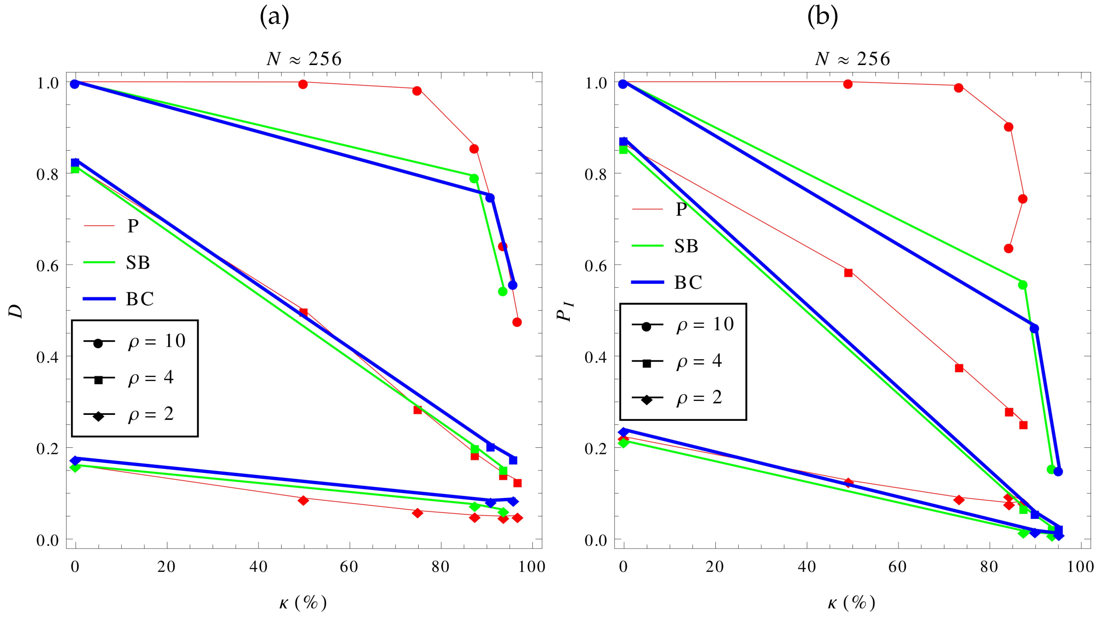

Figure 5.6: Plots of (a) discrimination D and (b) accurate-identification rate P_I against compression rate κ for the partition, symmetric base and binomial coefficient schemes, at different values of true SNR ρ for $N \approx 256$.

sion. The turnaround in accurate-identification rates for the partition scheme at larger values of P is due to the additional statistic evaluations used in the fine-grained search, which for I_1 gives $N_{\text{eval}} = M + P$ in (5.17). Since $M = N/P$, $\kappa(P)$ has one turning point. For this example, $P = 8$ and $P = 64$ provide the same level of compression; identification accuracy is higher for the former at $\rho = 10$, similar for both at $\rho = 4$, and higher for the latter at $\rho = 2$.

In summary, the partition scheme offers better overall performance than its correlated alternatives at the same level of compression. For detection, the introduced correlations among the conic statistics lead to slightly reduced detection sensitivity and discriminatory power at high SNR; furthermore, the potential benefits of lossless compression for identification turn out to be nullified by the effects of noise. Hence there appears to be little reason for using correlated schemes over the partition scheme, which is more promising as it is easy to implement and admits a relatively populated sliding scale of compression rates. We further investigate and implement the partition scheme as the representative conic compression scheme in Sections 5.3 and 5.4.

5.3 Orthogonality and subspaces

The conic compression schemes proposed in Section 5.2 are fully general, in the sense that no prior assumptions about the template bank are made apart from (5.4) and (5.5). These orthogonal and 1-D restrictions are neither realistic nor optimal, as template banks typically feature highly correlated neighbouring templates and are unlikely to contain a template exactly proportional to the GW signal itself. In this section, we discuss the (separate) lifting of each assumption for the partition scheme, and the resultant effects on detection sensitivity and localisation accuracy. Each approach may be viewed as a simplified limiting case of an actual template bank, which can always be made dense enough to include a signal-proportional template (assuming model accuracy), or orthogonally decomposed. A more realistic example with both assumptions lifted is considered in Section 5.4.

5.3.1 Non-orthogonal templates

We first consider a sufficiently dense bank of correlated (non-orthogonal) templates, such that the GW signal still lies in the 1-D subspace spanned by a single template in Hilbert space. From the first equalities in (5.6), (5.7), (5.15) and (5.16), it follows in the presence of a GW signal that

$$\mathbb{E}[X_m] = A \sum_{n \in \mathbf{U}_m} \langle h_1 | h_n \rangle, \tag{5.43}$$

$$\operatorname{cov}(X_m, X_{m'}) = \sum_{n \in \mathbf{U}_m} \sum_{n' \in \mathbf{U}_{m'}} \langle h_n | h_{n'} \rangle. \tag{5.44}$$

Any partition of **N** as in Section 5.2.1 defines a splitting of the (sorted) original mean vector and covariance matrix into $P \times 1$ blocks and $P \times P$ blocks respectively; each entry in the conic mean vector and covariance matrix is then simply the sum of entries in the corresponding block, which reflects the coarse-graining of the compression.

As a toy model for investigating non-orthogonal templates, we use a frequency-parametrised bank of sinusoidal waveforms $h = \sin(2\pi f t)$ with finite observation time T. Assuming white noise for simplicity, the inner product (2.53) may be written for real time series as

$$\langle a|b\rangle \propto \int_0^T dt\, a(t)b(t). \tag{5.45}$$

For an N-template bank with $f_{\min} \leq f \leq f_{\max}$ and $\delta f := (f_{\max} - f_{\min})/(N-1) \ll f_{\min}$, the overlaps are given by

$$\langle h_n|h_{n+\Delta n}\rangle \approx 2\int_0^1 dt\ \sin{(2\pi f_{\min} t)} \sin{(2\pi (f_{\min} + |\Delta n|\delta f)t)}, \tag{5.46}$$

where we have normalised to $T = 1$ such that f is given in waveform cycles per observation time. This sinc-like function of $\Delta n \in \mathbb{Z}$ yields a band covariance matrix for x_n; we set $N = 256$, and choose the frequency bounds such that $\mathrm{cov}(x_n, x_{n\pm1}) \approx 0.97$ (i.e. a maximal mismatch of around 0.03).

In contrast to the orthogonal case, the choice of partition generally affects the performance of the partition scheme for non-orthogonal templates. For the one-parameter template bank with overlaps given by (5.46), we consider both a randomised partition and a more optimised (but not necessarily optimal) partition with $\mathbf{U}_m = \{n \in \mathbb{Z}^+ \,|\, (m-1)P < n \leq mP\}$. We also include for comparison a uniformly spaced M-template subset of the original bank; equivalently, $\mathbf{U}_m = \{n \in \mathbb{Z}^+ \,|\, n = mP\}$ where $\bigcup_{m\in\mathbf{M}} \mathbf{U}_m \neq \mathbf{N}$. This "coarsened" template bank is obtained not through compression, but by simply reducing the correlation (increasing the maximal mismatch) between neighbouring templates. The standard detection algorithm outlined in Section 5.2.1 is then applied for the two partition schemes and the coarsening method.

Figure 5.7(a) shows plots of discrimination (using $X_{\max}$) against compression for both choices of partition and the coarsened template bank, with performance in the presence of a GW signal averaged over the N possible locations of the corresponding template in the bank. The optimised partition (with highly correlated templates grouped together) outperforms its randomised counterpart at all considered values of true SNR. It also shows improvement over the coarsening method at higher compression, which is expected as it uses information from the full N-template bank rather than just an M-template subset.

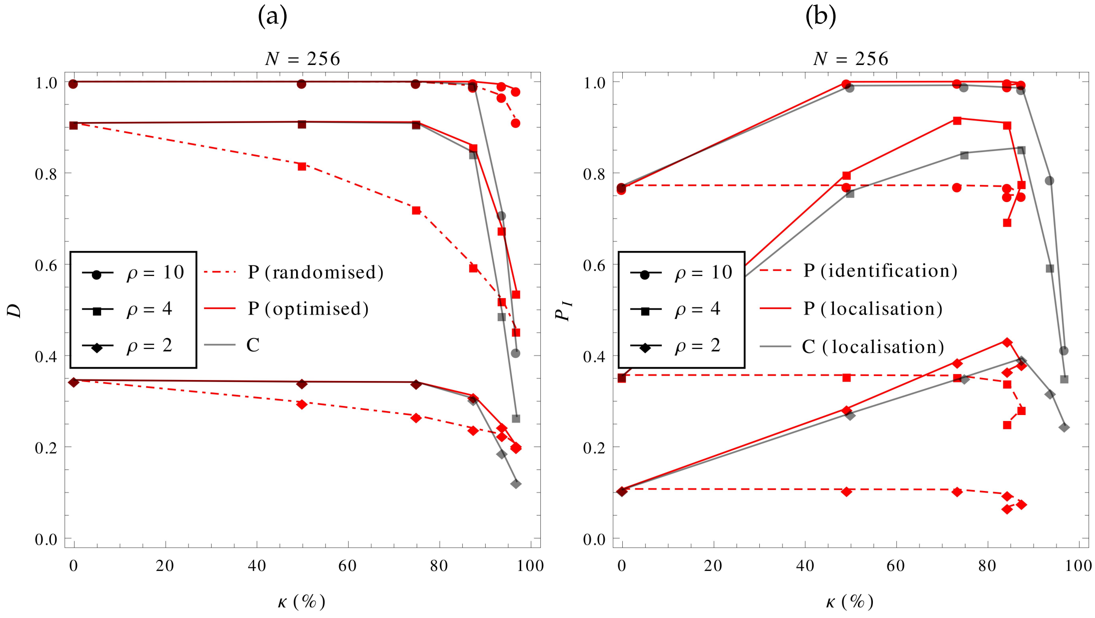

Figure 5.7: Plots of (a) discrimination D against compression rate κ for the randomised/optimised partition scheme and the coarsening method, and (b) accurate-identification/localisation rate P_I against compression rate κ for the optimised partition scheme and the coarsening method, at different values of true SNR ρ for a non-orthogonal bank. Accurate localisation here is defined as the identification of the template h_1 or one of the nearest P templates.

The largest statistic evaluation for the coarsened template bank identifies a best guess for the GW signal, but the accuracy of this identification is zero if the signal does not correspond to a template in the coarsened bank. Since the spacing of the coarsened bank is P, we may consider the best-guess template as representative of the P templates nearest to it (or $P - 1$ if P is odd), and say that the largest statistic evaluation localises a best guess for the signal. We then define the localisation to be accurate if the correct template h_1 is one of those templates (equivalently, if the identified best-guess template is h_1 or one of the nearest P templates). The identification algorithms in Section 5.2.1 also identify a single best-guess template for the partition scheme, which allows us to consider both accurate identification (to a precision of 1) and accurate localisation (to a precision of P) in the same way. Figure 5.7(b) shows plots of these accuracy rates (using I_1, which gives $N_{\mathrm{eval}} = M + P$ in (5.17)) against compression for the optimised partition scheme and the coarsening method.

As in Section 5.2.4, the turnaround in accurate-identification and localisation rates for the partition scheme is due to the additional statistic evaluations of the fine-grained search. The localisation rates increase up to some level of compression, which is mainly because "accurate" localisation is defined up to a degree of precision that degrades with compression; this effect is seen for the coarsening method as well. Localisation to within the spacing of the original template bank (i.e. identification) decreases monotonically in accuracy for the partition scheme, and will not be achievable for the majority of signals with the coarsening method. The partition scheme localises the GW signal with slightly greater accuracy than the coarsening method, and in fact identifies it with virtually no fall-off in accuracy at significant compression levels.

Increasing the correlation between neighbouring templates is known to improve the detection and localisation performance of a general template bank [173]. Results in this section illustrate that the partition scheme retains these benefits up to high levels of compression, and provides a superior alternative to simply coarsening the template bank for computational savings. The viability of conic compression becomes even more evident in Section 5.4, where we apply the partition scheme to a larger and more broadly correlated two-parameter template bank.

5.3.2 Two-dimensional subspace

Throughout Sections 5.2 and 5.3.1, we have assumed that the GW signal is exactly proportional to a template in the bank. To understand the impact on compression performance when this is not the case, we consider a bank of N uncorrelated templates obtained through some orthogonalisation procedure (e.g. as in [164–166]) on a general template bank, and a signal lying in the N-dimensional Hilbert space spanned by the orthogonal set.

If N is large, the signal is typically restricted to a low-dimensional subspace (this follows from the volume of an N-sphere). For simplicity, we assume it lies exactly between two templates in a 2-D subspace, i.e.

$$x = A(h_1 + h_2) + \{\text{noise}\}, \tag{5.47}$$

where the templates have been relabelled without loss of generality and $A = \rho/\sqrt{2}$ from (2.54). Hence the expectation values of the original and conic statistics become

$$\mathbb{E}[x_n] = A(\delta_{1n} + \delta_{2n}), \tag{5.48}$$

$$\mathbb{E}[X_m] = A\,\mathrm{card}(\{1,2\} \cap \mathbf{U}_m), \tag{5.49}$$

while their covariances remain as (5.7) and (5.16) respectively. The assumption (5.47) is the worst-case scenario for a 2-D subspace, since the signal is maximally far from both templates in the subspace.

Although it is not possible to pre-optimise the choice of partition for orthogonal templates, the performance of the partition scheme in the presence of a 2-D GW signal falls into two partition-dependent cases. At small values of P, it is more likely that the labels $1 \in \mathbf{U}_m$ and $2 \in \mathbf{U}_{m'}$ are assigned to different sets ($m \neq m'$); as P increases, so does the probability that they are assigned to the same set ($m = m'$), which improves performance (e.g. the effective SNR for detection purposes is raised by a factor of $\sqrt{2}$).

The standard detection algorithm in Section 5.2.1 is applicable for a 2-D signal, while the standard identification algorithms may be generalised at step (iv) by considering the two largest original statistic evaluations instead. Figure 5.8 shows plots of discrimination and accurate-identification rate against

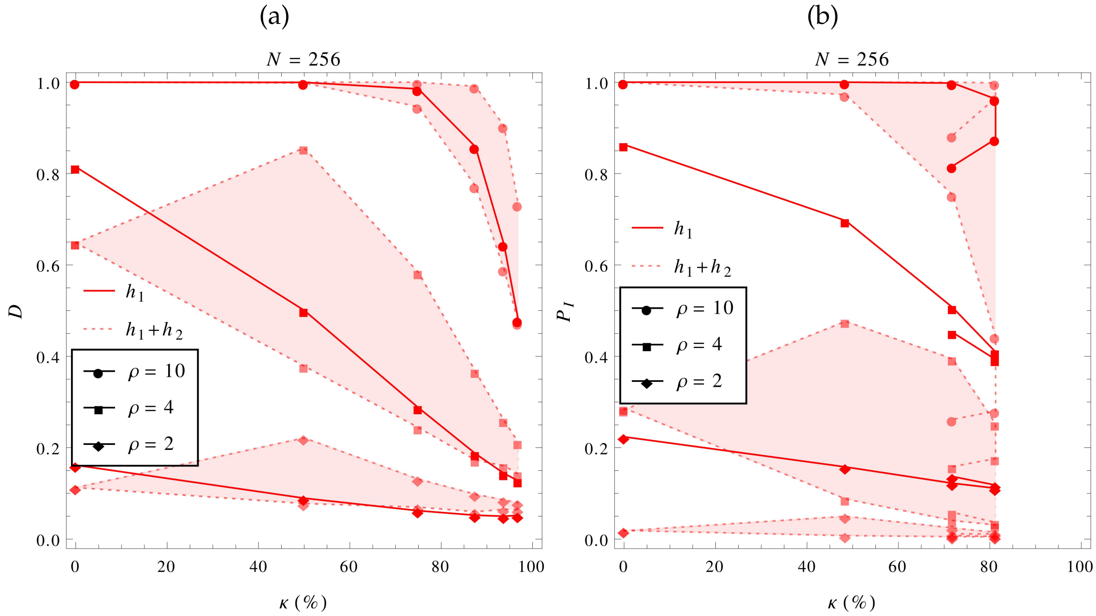

Figure 5.8: Plots of (a) discrimination D and (b) accurate-identification rate P_I against compression rate κ for the partition scheme with 1-D and 2-D signals, at different values of true SNR ρ. The higher dotted curves for each value of ρ correspond to the scenario where the template labels 1 and 2 are assigned to the same set, while the lower curves correspond to them being assigned to different sets. Accurate identification for the 2-D case is defined as the identification of both templates h_1 and h_2.

compression for a 2-D signal $\propto (h_1 + h_2)$, compared against a 1-D signal $\propto h_1$ with the same true SNR ρ. The identification algorithm I_2 is used, since the accuracy rate of I_1 falls to zero if $m \neq m'$. This gives $N_{\text{eval}} = M + 2P$ in (5.17).

For detection of a 2-D GW signal, the effectiveness of the partition scheme is reduced slightly at lower SNRs, but mitigated by the case where $m = m'$ (i.e. the higher dotted curves in Figure 5.8). Detection performance for this special case actually improves up to some level of compression, which can occur as the symmetry among all possible signals is broken (by the partitioning process). A similar effect is seen for the example in Section 5.4. The discrimination for a 1-D signal generally lies within the 2-D discrimination bounds; at higher compression rates, there is little to no detection performance lost if the signal is not confined to a 1-D subspace.

Accurate identification of a 2-D GW signal (i.e. the identification of both h_1 and h_2 in this toy model) is more problematic than in the 1-D case, since accuracy rates are reduced to begin with and fall off rapidly even at high SNR. Nevertheless, options such as lowering compression or switching to $\mathrm{I}_{i>2}$ are available for the partition scheme, which should at least allow the template with maximal signal overlap to be identified at acceptable accuracy rates.

If the true SNR is sufficiently high, the algorithms $\mathrm{I}_{i>j}$ may also be used to identify a j-dimensional GW signal described by an arbitrary linear combination of templates, i.e.

$$x = \sum_{k=1}^{j} A_k h_k + \{\text{noise}\}, \tag{5.50}$$

where $A_k > A_{k+1}$ and the templates have been relabelled without loss of generality. At step (iv) of the algorithms, each ordered weight A_k may be approximated by the k-th largest original statistic evaluation $x_{(k)}$, with the detection SNR given by

$$\rho_D = \sqrt{\sum_{k=1}^{j} x_{(k)}^2}. \tag{5.51}$$

While this method fully recovers the (relative) weights of a GW signal's j largest modes in the limit of zero noise, its accuracy might be significantly diminished for lower-SNR signals or at higher levels of compression.

5.4 Taylor-T2 example

In this section, we implement the (optimised) partition scheme described in Sections 5.2.1 and 5.3.1 for a larger and more realistic example: a two-parameter template bank of mixed-order PN waveforms, which describe the gravitational radiation emitted during the inspiral part of a comparable-mass binary coalescence. An optimised partition here (and in general) refers to a partition of the template bank for which the convex hulls of the sets are non-intersecting, such that highly correlated templates are grouped together.

The waveform family we use is the Taylor-T2 approximant [38, 174] for a circular and non-inclined binary with comparable component masses $(m_1, m_2 \geq m_1)$. These waveforms are parametrised by chirp mass (2.31) and symmetric mass ratio (2.27), and are written as PN expansions in the frequency-related variable $\xi := (G\mathcal{M}\eta^{-3/5}\dot{\phi}/c^3)^{2/3}$, where units have been restored and $\dot{\phi}$ is the orbital phase frequency. We truncate the PN expansions at finite order, specifying the phase and amplitude to 3.5PN and 2PN respectively; the resultant mixed-order waveform may be written compactly as [175]

$$h_{\mathcal{M},\eta}(t) = \frac{2G\mathcal{M}\eta^{2/5}}{c^2 R}\mathcal{A}(t)\exp\left(2i\psi(t)\right), \tag{5.52}$$

where R is the source distance (which the true SNR ρ is inversely proportional to), and expressions for the amplitude function $\mathcal{A}$ and tail-distorted orbital phase ψ are given in Appendix C.

Template bank compression is potentially more important for analysing data from space-based detectors, since the long duration and potential complexity of GW signals in their low-frequency sensitivity band greatly increases the cost of a grid-based detection search. As SMBH binary coalescences are an anticipated source for LISA, we consider a Taylor-T2 signal with the parameters $\boldsymbol{\theta}_C = (1, 0.15)$, where $\boldsymbol{\theta} := (\mathcal{M}/(10^6 M_\odot), \eta)$; this corresponds to a black-hole binary inspiral with component masses $(m_1, m_2) = (0.6, 2.5) \times 10^6 M_\odot$. The duration of the signal is set to $t_c = 1\,\mathrm{yr}$.

We also generate a bank of Taylor-T2 templates with the same duration, each normalised with respect to the inner product (2.53), where $\mathcal{S}_n(f)$ is given by an analytic approximation to the (down-scoped) LISA noise power spectral density [176]. These templates are gridded uniformly in the transformed parameters $\boldsymbol{\theta}' := \boldsymbol{\theta}_C + L(\boldsymbol{\theta} - \boldsymbol{\theta}_C)$ (128 points in each parameter), with the signal lying in the middle of the four central templates and the linear transformation L chosen such that the template overlaps are isotropic with respect to the grid (at least for the central region). The maximal mismatch of each template with its four nearest neighbours is around 0.01.

Since the $N = 16384$ templates are pre-sorted by the (skewed) square grid, an optimised (but not necessarily optimal) partition is obtained by the obvious grouping into M blocks of $\sqrt{P} \times \sqrt{P}$ templates. This particular template bank admits six nontrivial square partitions with $P \in \{4, 16, 64, 256, 1024, 4096\}$; we do not consider the case $P = 4096$, as $P = 1024$ already yields a compression rate of 99.9%. A large number of rectangular partitions (where $P = 2^i$ with $0 < i < 14$) are also possible, but we omit these here for simplicity as they are degenerate with the square partitions and among themselves. Square partitions are straightforward to generalise for various lattice choices [177–179], and will be fairly optimal as long as the templates are gridded uniformly in the parameter-space metric.

The expectation values of the original and conic statistics (given by the first equality in (5.6) and (5.43) respectively) are visualised in Figure 5.9, where the coarse-graining of the compression is evident. Overlaps for the Taylor-T2 template bank are much less localised than the toy model overlaps in Section 5.3.1; this is due to their wider cycle widths in both $\mathcal{M}$ and η, as well as a slight degeneracy in the two parameters (overlaps at the boundary of the first plot in Figure 5.9 can be as high as 0.4ρ). As the templates are so broadly correlated and the GW signal is injected right in the centre of the bank, the partition scheme is expected to perform well up to a high level of compression.

For comparison purposes, we again consider the simple coarsening method discussed in Section 5.3.1. The smaller coarsened banks are formed by selecting individual templates near the centre of each square block in the original bank, rather than by summing the templates as in the partition scheme. Detection and localisation performance for both the partition scheme and the coarsening method on the Taylor-T2 template bank with central injection is sum-

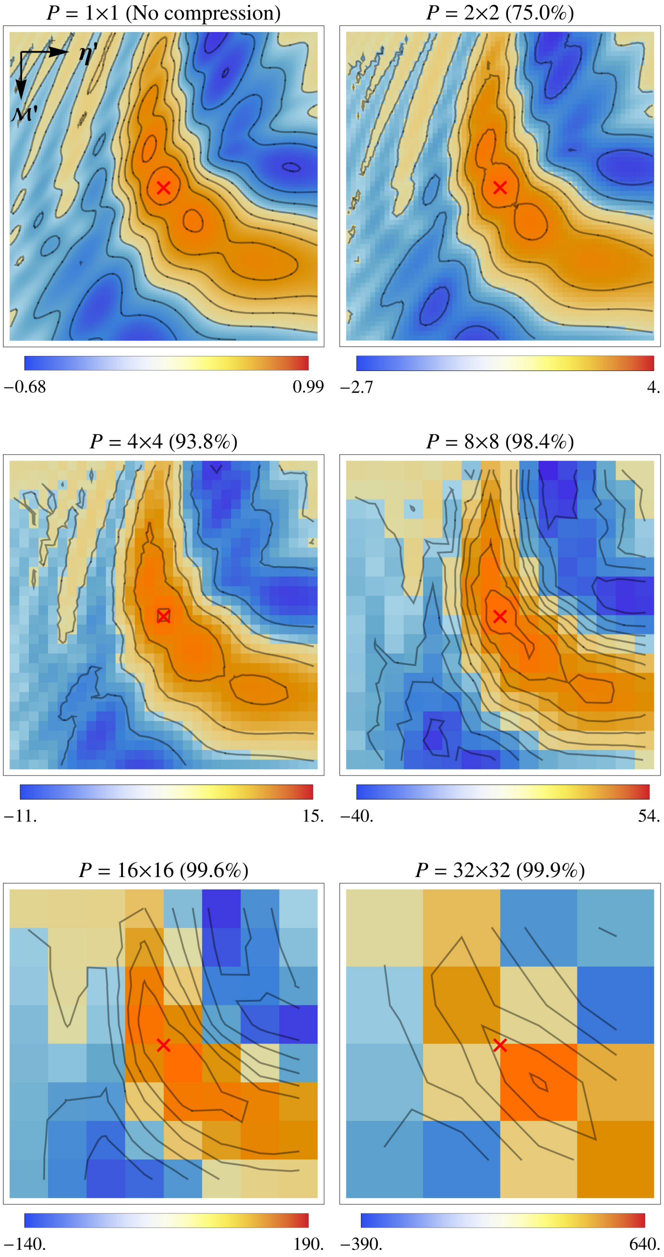

Figure 5.9: Matrix/contour plots of the expectation values $\mathbb{E}[x_n]$ and $\mathbb{E}[X_m]$ for the partition scheme, at different values of set cardinality P (with compression rate κ in parentheses) for a Taylor-T2 GW signal (red cross) injected between the four central templates of a (128×128)-template Taylor-T2 bank. The bank is gridded uniformly in linearly transformed parameters $\mathcal{M}'$ (increasing from top to bottom) and η' (increasing from left to right) with maximal mismatch ≈ 0.01. Overlap values depend on the true SNR ρ (set to 1 in these plots), and range from positive (orange) to negative (blue) in some subinterval of $(-P\rho, P\rho)$.

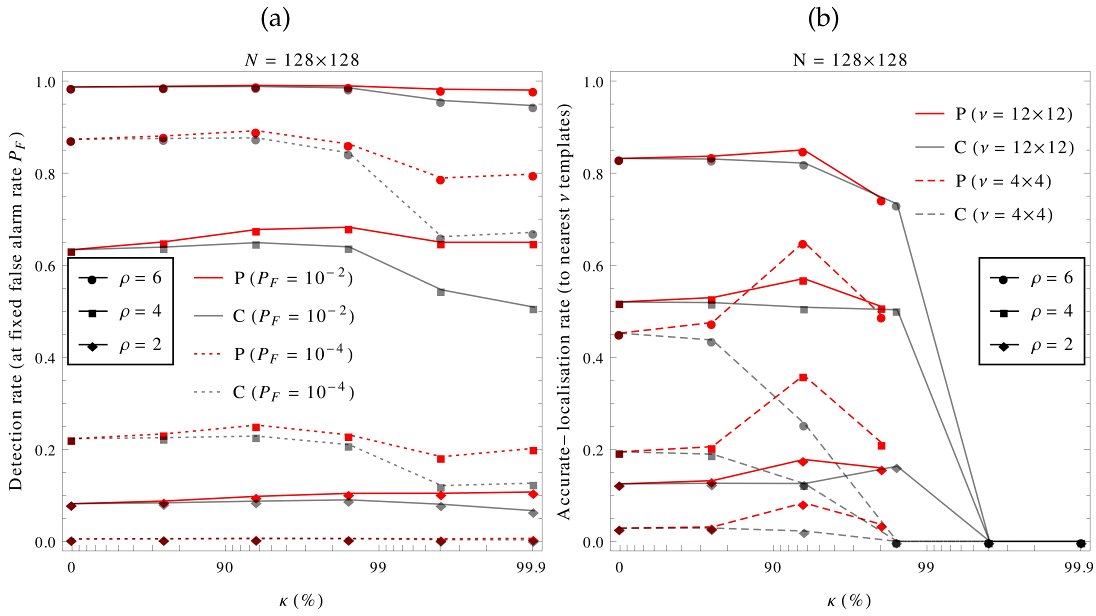

Figure 5.10: Plots of (a) detection rate P_D (at fixed false alarm rate P_F) and (b) accurate-localisation rate P_I (to nearest ν templates) against compression rate κ for the optimised partition scheme and the coarsening method, at different values of true SNR ρ for a GW signal injected with central parameters $\boldsymbol{\theta}_C$. Accurate localisation here is defined as the identification of a template within the central squares of ν templates.

marised in Figure 5.10. The semi-log plots in this section use an abscissa of $-\lg(1-\kappa)/3$, as most of the considered compression rates are $> 90\%$.

Instead of the discrimination (5.42), we quantify detection performance using the detection rates at two fixed false alarm rates $P_F = 10^{-2}$ and $P_F = 10^{-4}$ (the number of Monte-Carlo trials performed for each plot point is $\sim 10^5$, and so the errors are $\sim 10^{-3}$ for a 1-sigma binomial confidence interval). At all considered values of SNR and fixed false alarm rate, there is no fall-off in the partition scheme's detection performance up to $\kappa = 93.8\%$ (and even a slight increase, due to the special choice of central injection). While this is also the case for the coarsening method, detection rates for the partition scheme are distinctly higher at compression rates of $> 90\%$, with improvements of over 0.1 at $\kappa = 99.9\%$.

The identification algorithm I_1 is used to localise the GW signal, which gives $N_{\mathrm{eval}} = M + P$ in (5.17). Rates for accurate localisation to within two central squares of 12×12 templates (corresponding to $< 1\%$ of the entire bank) and 4×4 templates ($< 0.1\%$ of the bank) are considered. Localisation is typically improved by compression up to $\kappa = 93.7\%$, which is provided by two different values of P (see discussion in Section 5.2.4). The two values are $P = 16$, beyond which the matrix/contour plot of $\mathbb{E}[X_m]$ in Figure 5.9 loses scale-similarity to that of $\mathbb{E}[x_n]$, and $P = 1024$, for which performance is regained as each conic template incorporates more of the original templates and accuracy is added by the fine-grained search. Localisation is poorer at $\kappa = 98.0\%$, which corresponds to both $P = 64$ and $P = 256$. To reduce clutter in Figure 5.10(b), only the higher localisation rates for $\kappa = 93.7\%$ and $\kappa = 98.0\%$ are plotted. The partition scheme outperforms the coarsening method at most levels of compression, especially in the case of accurate localisation to within the smaller square of 4×4 templates.

For the special case of a centrally injected GW signal, the detection and localisation performance of the partition scheme is non-decreasing up to high levels of compression and can even rise above that of the original template bank; however, this may also be said for the coarsening method. To illustrate that the improvement of the partition scheme over the coarsening method is not simply due to the special choice of injection, we consider two other cases:

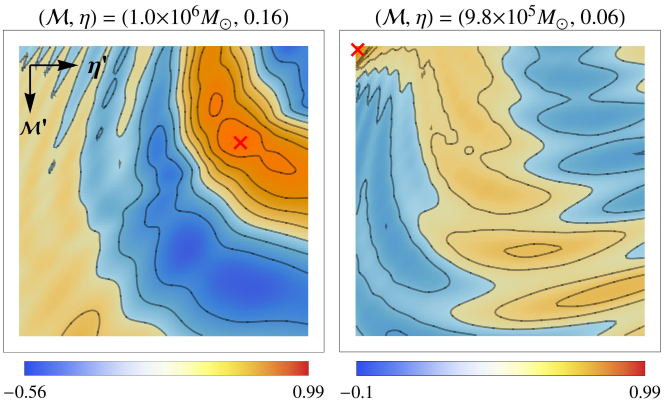

Figure 5.11: Matrix/contour plots of the expectation values $\mathbb{E}[x_n]$ for Taylor-T2 GW signals (red crosses) injected in a (128×128)-template Taylor-T2 bank at random (left) and near the boundary (right).

a Taylor-T2 signal injected with randomly drawn parameters $\boldsymbol{\theta}_R = (1.0, 0.16)$, and another injected near the boundary of the bank with the parameters $\boldsymbol{\theta}_B = (0.98, 0.06)$ (i.e. in the middle of the four corner templates with low chirp mass and symmetric mass ratio). The expectation values of the original statistics for these two injections are visualised in Figure 5.11.

Figure 5.12 shows detection and accurate-localisation rates for both the partition scheme and the coarsening method on the random and boundary injections. The random injection is actually recovered with slightly better rates than the central injection rates in Figure 5.10, but with a similar improvement of the partition scheme over the coarsening method. A more marked difference between the two methods is obtained for the boundary injection. Detection rates for both methods are now non-increasing, with the partition scheme showing greater improvement over the coarsening method; for the $\rho = 6$ case, the improvement is around 0.3 at $\kappa = 93.8\%$. Rates for accurate localisation of the boundary injection to within the corner square of 12×12 templates follow a similar trend, with a largest improvement of around 0.5 (again for the $\rho = 6$ case at $\kappa = 93.7\%$).

Detection and localisation performance for this Taylor-T2 example is injection-dependent, as it is for any realistic template bank: there is clearly no symmetry among all possible GW signals, since the templates are asymmetri-

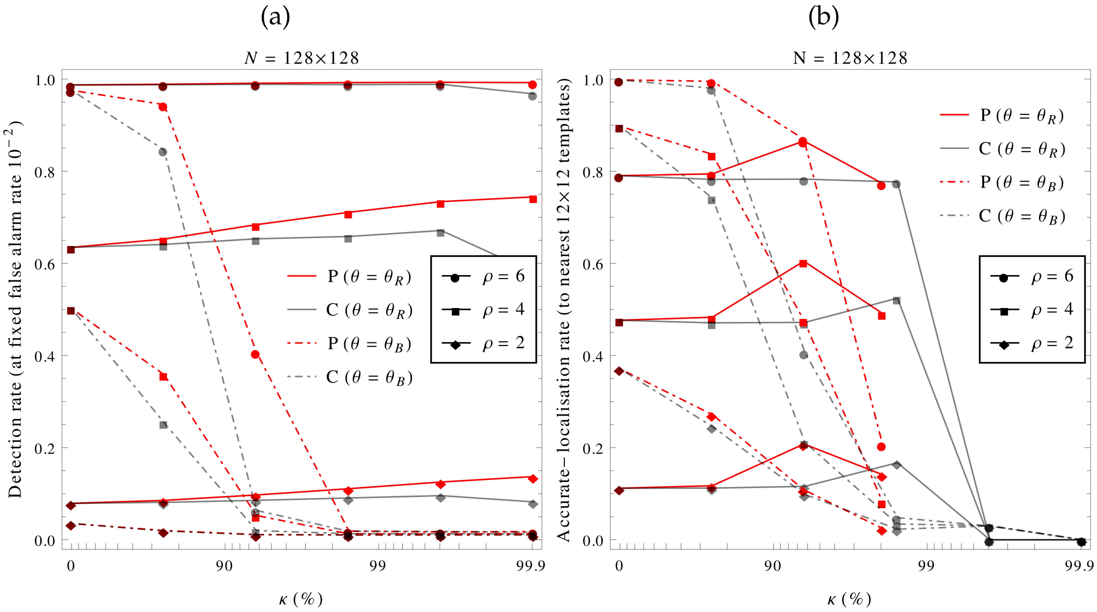

Figure 5.12: Plots of (a) detection rate P_D (at fixed false alarm rate $P_F = 10^{-2}$) and (b) accurate-localisation rate P_I (to nearest 12×12 templates) against compression rate κ for the optimised partition scheme and the coarsening method, at different values of true SNR ρ for a GW signal injected with random parameters $\boldsymbol{\theta}_R$ and boundary parameters $\boldsymbol{\theta}_B$. Accurate localisation here is defined as the identification of a template within the square of 12×12 templates nearest to each injection.

cally correlated and the signals may lie between templates. We have not undertaken a full injection-averaged analysis (similar to that in Section 5.3.1) due to the size of the template bank, but overall detection rates for such an analysis should decrease monotonically with compression as per intuition, with the partition scheme outperforming the coarsening method (as it does for the three injections presented here, as well as several others we have examined).

The partition scheme is expected to remain robust for searches in a $(d > 2)$-dimensional parameter space. As the number of templates that are highly correlated with the GW signal increases exponentially with d, enlarging the span P of each conic template at the same rate should maintain detection sensitivity while increasing the relative computational savings (which scale as $1 - 1/P$). Good scaling with parameter-space dimensionality allows conic compression to be competitive with other search techniques that reduce computational cost. For example, the method of searching over the signal time-of-arrival using fast Fourier transforms yields a logarithmic reduction in the number of search points for that parameter [177], but for multidimensional searches an overall logarithmic reduction is easily attained by the partition scheme with little impact on performance. The two methods might even be combined for greater savings, by constructing conic sums of templates aligned at a fixed reference time and using their Fourier transforms to search over time-of-arrival.

5.5 Discussion

In this chapter, we have presented and compared three tunable conic compression schemes (partition, symmetric base and binomial coefficient) for a GW template bank in a grid-based detection search. The bank is compressed in the preparatory offline stage, which yields faster detection and localisation of signals by reducing the number of inner product evaluations performed online.

A recently proposed binary labelling method, modified to ensure the equal treatment of templates, is subsumed as a particular case of the symmetric base scheme. Optimal detection statistics have been calculated for all three schemes under simplified conditions, and the standard maximum-overlap detection statistic is shown to be significantly suboptimal for the base and binomial schemes. While these two lossless schemes provide automatic identification of the GW signal upon detection, the benefits of this are negated in the presence of noise. Furthermore, the lossy partition scheme offers better overall performance than its counterparts at the same level of compression.

We have applied the partition scheme to toy models of (i) a correlated template bank with a signal-proportional template and (ii) a signal lying in the span of orthogonal templates, to show that it remains feasible under such conditions. These toy models are instructive as they represent the two limiting cases of a general template bank. Correlations among the original templates result in partition-dependent performance, but this may be optimised beforehand by grouping highly correlated templates together; the optimised partition scheme is then superior to a simple coarsening of the template bank. If the signal is proportional to a linear combination of templates in an orthogonal bank, the detection performance of the scheme is not significantly reduced.

Conic compression performs well if the original template bank is sufficiently correlated, as demonstrated by our example implementation of the optimised partition scheme for a bank of $\sim 10^4$ PN waveforms. We consider a centrally injected GW signal, a randomly injected one, and one at the boundary of the bank; again, the scheme is superior to the coarsening method across the board. The partition scheme is shown to be viable for practical applications, as it maintains good detection sensitivity and localisation accuracy up to high levels of compression and at all considered values of SNR for this more realistic template bank.

In summary, our tunable conic compression schemes—specifically the optimised partition scheme—might provide an effective method of improving the speed, detection sensitivity and localisation accuracy of GW template banks. The schemes are potentially useful for any search involving template banks, as they are fully general and may easily be adapted to supplement existing algorithms in GW data analysis pipelines. Conic compression is also particularly promising in the context of LISA data analysis, where online grid searches will be difficult as computational costs are more prohibitive; for example, the method could be used as an online tool to rapidly identify nearby sources before merger and generate alerts for electromagnetic telescopes.

Chapter 6

Gaussian process regression

The theoretical uncertainty in waveform models must be accounted for when performing Bayesian parameter estimation on data from GW detectors, especially as it could be the dominant source of error at high SNRs. A recently introduced method deals with model error by marginalising over it, using a prior probability density that is constructed through the machine-learning technique of Gaussian process regression.

In this chapter, we outline the framework of the method and summarise the results of its application to LIGO parameter estimation, before performing a preliminary investigation of its viability for EMRI sources in LISA data analysis. Through low-dimensional parameter estimation studies, we demonstrate that the method remains computationally practical in the light of various difficulties associated with searching the large EMRI parameter space. We also discuss possible strategies for its generalisation to higher-dimensional searches.

The material in this chapter is adapted from selected parts of [134, 180, 181], but the main results in Section 6.3 have yet to be published.

6.1 Background

A typical EMRI will be observed by LISA for $\sim 10^5$ orbits over the mission lifetime. Although this allows its properties to be measured very precisely, the results are consequently susceptible to any inaccuracy in the waveform models that are used in the Bayesian parameter estimation algorithms. This "theoretical error" will likely dominate over statistical error at high SNRs [182]. At the same time, EMRIs are notoriously difficult to model accurately, since the extreme mass ratio prohibits the use of NR simulations. Waveform models that employ self-force calculations are still under development; they will also be computationally expensive, and are unlikely to be used directly for parameter estimation after they become available.

Current data analysis studies for LISA are therefore heavily reliant on the kludges described in Chapter 4, which are designed for robust use in search algorithms and can be modified to include self-force information. However, it will still be challenging to do a rough detection search of the full EMRI parameter space with kludge waveforms [168], much less explore it with the precision required for accurate parameter estimates. EMRI waveforms are extremely sensitive to small changes in their parameters, and so the global peak in the vast and multimodal posterior surface is akin to the proverbial needle in a haystack. This fact, coupled with the difficulty in modelling the complex waveforms and the systematic bias due to theoretical error, makes EMRI parameter estimation the most formidable problem in LISA data analysis.

In this chapter, we investigate the machine-learning technique of Gaussian process regression (GPR) [183, 184] as a possible strategy for mitigating theoretical error in EMRI parameter estimation. Specifically, GPR is used to interpolate a small set of precomputed waveform differences between an accurate model and an approximate one; the GPR interpolant then provides a prior probability density for the waveform difference, which allows theoretical error to be marginalised over in the standard Bayesian likelihood with the approximate model [185]. The benefits of this method for GW parameter estimation are twofold: it includes information from computationally expensive waveforms while searching with faster but less accurate ones, and accounts for any residual model inaccuracy with more conservative error estimates.

An overview of the new marginalised-likelihood method is given in Section 6.2. The technique of GPR (in the context of waveform interpolation) and the training of the Gaussian process model from the precomputed set of waveform differences are introduced in Sections 6.2.1 and 6.2.2 respectively; Section 6.2.3 then briefly summarises results from a LIGO-oriented proof-of-concept study, where the method is applied to parameter estimation for comparable-mass black-hole binary mergers.

In Section 6.3, the viability of the GPR likelihood for EMRI parameter estimation is investigated, and a simple argument is provided to show that the minimal allowed density of the training set (over parameter space) is typically much lower than the sampling density of a parameter estimation search with the accurate waveforms. This is verified by one- and two-dimensional examples in Sections 6.3.1 and 6.3.2 respectively, where the relationship between the waveform-difference Fisher information matrix and the optimality of the GPR model and training set is also made clear. Finally, possible computational strategies for generalising the method to higher-dimensional searches are discussed in Section 6.4.

6.2 The marginalised likelihood

In this chapter, we consider waveform models $h(\boldsymbol{\lambda} \in \Lambda)$ over a common ℓ-dimensional parameter space Λ, which we take to be isomorphic to $\mathbb{R}^\ell$ for simplicity. The standard Bayesian likelihood $L(\boldsymbol{\lambda}|x)$ for the model parameters (given the data x) is defined in (2.59), with the usual noise-weighted inner product (2.53) on the function space W of finite-length complex time series. A maximum-likelihood estimation $\boldsymbol{\lambda}_{\mathrm{ML}}$ of the parameters may then be obtained by maximising (2.59) over Λ, such that

$$\left\langle \frac{\partial h}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}_{\mathrm{ML}}) \middle| x - h(\boldsymbol{\lambda}_{\mathrm{ML}}) \right\rangle = \mathbf{0}. \tag{6.1}$$

For a waveform model h_{acc} that provides an accurate description of the source, the waveform template at the true parameter values $\boldsymbol{\lambda}_{\mathrm{true}}$ is consistent with the source signal, i.e. $x = h_{\mathrm{acc}}(\boldsymbol{\lambda}_{\mathrm{true}}) + n$. At leading order in n, any error $\boldsymbol{\lambda}_\epsilon := \boldsymbol{\lambda}_{\mathrm{ML}} - \boldsymbol{\lambda}_{\mathrm{true}}$ in the measured parameter values is then purely statistical, in that it depends linearly on n. Using Einstein notation, we may write (6.1) as

$$\langle (\partial h_{\mathrm{acc}})_b | n - (\partial h_{\mathrm{acc}})_a (\lambda_\epsilon)^a \rangle = 0 \implies (\lambda_\epsilon)^a = (\Gamma_{\mathrm{acc}}^{-1})^{ab} \langle n | (\partial h_{\mathrm{acc}})_b \rangle, \tag{6.2}$$

where $\partial h \equiv \partial h / \partial \boldsymbol{\lambda}$ is the waveform derivative vector and $\boldsymbol{\Gamma}$ is the Fisher information matrix (2.61), with both evaluated at $\boldsymbol{\lambda}_{\mathrm{ML}}$.

In general, a waveform model h_{app} that is used for parameter estimation may only be approximate, such that $h_{\mathrm{app}}(\boldsymbol{\lambda}_{\mathrm{true}}) \neq h_{\mathrm{acc}}(\boldsymbol{\lambda}_{\mathrm{true}})$. At leading order in n, any error in the measured parameter values may be written as

$$(\lambda_\epsilon)^a = (\Gamma^{-1}_{\mathrm{app}})^{ab} \langle n | (\partial h_{\mathrm{app}})_b \rangle - (\Gamma^{-1}_{\mathrm{app}})^{ab} \langle h_\epsilon | (\partial h_{\mathrm{app}})_b \rangle, \tag{6.3}$$

where the first term is statistical in the sense of (6.2), and the second term corresponds to the theoretical error that arises from the difference h_ϵ between the approximate and accurate waveforms, i.e.

$$h_\epsilon := h_{\mathrm{app}} - h_{\mathrm{acc}}. \tag{6.4}$$

Again, all derivatives in (6.3) are evaluated at $\boldsymbol{\lambda}_{\mathrm{ML}}$, while the waveform difference h_ϵ is evaluated at $\boldsymbol{\lambda}_{\mathrm{true}}$.[15]

The statistical-error terms in (6.2) and (6.3) are inversely proportional to the waveform SNR $\rho = \langle h|h \rangle^{1/2}$, since $\partial h \propto \rho$ and $\boldsymbol{\Gamma} \propto \rho^2$. On the other hand, the theoretical-error term in (6.3) is independent of ρ. Hence the systematic bias incurred by using approximate waveforms h_{app} in (2.59) may dominate the noise uncertainty for high-SNR sources, and is likely to be the limiting factor in extracting parameter information from such signals [182].

One approach to account for this bias is to marginalise over the error of h_{app} (with respect to h_{acc}) in (2.59), by specifying a suitable prior probability density $p(h_\epsilon)$ for the waveform difference [185]. The "marginalised likelihood" is given by the (functional) integral

$$\mathcal{L} \propto \int Dh_\epsilon \, L_{\mathrm{app}} p(h_\epsilon), \tag{6.5}$$

which can be evaluated analytically if $p(h_\epsilon)$ is Gaussian (since L_{app} is also Gaussian). Such a prior may be obtained through the technique of GPR, which provides an interpolant for h_ϵ with an associated normal distribution at each point in parameter space.

[15] However, $h_\epsilon(\boldsymbol{\lambda}_{\mathrm{true}}) = h_\epsilon(\boldsymbol{\lambda}_{\mathrm{ML}})$ at leading order, which allows the theoretical error in (6.3) (and hence $\boldsymbol{\lambda}_{\mathrm{true}}$ itself, at high SNR) to be estimated for a given measurement $\boldsymbol{\lambda}_{\mathrm{ML}}$ [182].

6.2.1 The Gaussian process prior

In the GPR approach, the waveform difference $h_\epsilon \in W$ is modelled as a zero-mean Gaussian process over Λ, i.e.

$$h_\epsilon \sim \mathcal{GP}(0, k), \tag{6.6}$$

where the mean is chosen (uninformatively) as the time series $0 \in W$, and the covariance function $k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ is some symmetric and positive-definite bilinear form on Λ. For any finite set of parameter points $\{\boldsymbol{\lambda}_i \in \Lambda \,|\, i = 1, 2, ..., N\}$, the corresponding set of waveform differences $\{h_\epsilon(\boldsymbol{\lambda}_i) \in W \,|\, i = 1, 2, ..., N\}$ has a Gaussian probability distribution $\mathcal{N}(\mathbf{0}, \mathbf{K})$ on W^N, i.e.

$$p(\mathbf{v}) \propto \frac{1}{\det \mathbf{K}} \exp\left(-\frac{1}{2}\mathbf{v}^T\mathbf{K}^{-1}\mathbf{v}\right), \tag{6.7}$$

where the waveform difference vector $\mathbf{v}$ and covariance matrix $\mathbf{K}$ are given respectively by

$$\mathbf{v} = [h_\epsilon(\boldsymbol{\lambda}_i)]. \tag{6.8}$$

$$\mathbf{K} = [k(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j)], \tag{6.9}$$

It is convenient to write the quadratic form in (6.7) as

$$\mathbf{v}^T\mathbf{K}^{-1}\mathbf{v} = \mathrm{tr}(\mathbf{K}^{-1}\mathbf{M}), \tag{6.10}$$

where

$$\mathbf{M} := \mathbf{v}\mathbf{v}^T = [\langle h_\epsilon(\boldsymbol{\lambda}_i) | h_\epsilon(\boldsymbol{\lambda}_j)\rangle]. \tag{6.11}$$

In choosing a frequency-independent form $k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ for the covariance function, we have assumed that the correlations among the waveform differences across parameter space do not depend on frequency. The particular normalising factor in (6.7) is only correct for a single frequency component (i.e. if the waveform difference at each parameter point is perfectly correlated across all frequency bins), but is retained for practical reasons.[16] Finally, the inner product for waveform differences in (6.11) is chosen to be the same noise-weighted

[16] Also, the "usual" exponent of $1/2$ is absent due to integration over the complex plane.

inner product (2.53) as for the waveforms. These assumptions simplify the GPR calculations, but are also conservative in the sense that they yield less informative likelihoods; a more detailed justification is provided in [180].

From the definition of a Gaussian process, the enlarged set $\{h_\epsilon(\boldsymbol{\lambda}_i), h_\epsilon(\boldsymbol{\lambda})\}$ is again normally distributed with zero mean and the covariance matrix

$$\mathbf{K}_* := \begin{bmatrix} \mathbf{K} & \mathbf{k}_* \\ \mathbf{k}_*^T & k_{**} \end{bmatrix}, \tag{6.12}$$

where

$$\mathbf{k}_* := [k(\boldsymbol{\lambda}_i, \boldsymbol{\lambda})], \tag{6.13}$$

$$k_{**} := k(\boldsymbol{\lambda}, \boldsymbol{\lambda}). \tag{6.14}$$

If $\{h_\epsilon(\boldsymbol{\lambda}_i)\}$ is known, then the conditional probability density of $h_\epsilon(\boldsymbol{\lambda})$ given $\{h_\epsilon(\boldsymbol{\lambda}_i)\}$ is also Gaussian, i.e.

$$p(h_\epsilon(\boldsymbol{\lambda})|\{h_\epsilon(\boldsymbol{\lambda}_i)\}) \propto \frac{1}{\sigma^2} \exp\left(-\frac{1}{2}\frac{\langle h_\epsilon(\boldsymbol{\lambda}) - \mu | h_\epsilon(\boldsymbol{\lambda}) - \mu\rangle}{\sigma^2}\right), \tag{6.15}$$

where $\mu(\boldsymbol{\lambda})$ and $\sigma^2(\boldsymbol{\lambda})$ are given respectively by

$$\mu = \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{v}, \tag{6.16}$$

$$\sigma^2 = k_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*. \tag{6.17}$$

The conditional probability (6.15) forms the basis of GPR, and yields an interpolation of $h_\epsilon(\boldsymbol{\lambda})$ from a small precomputed "training" set

$$\mathcal{D} := \{(\boldsymbol{\lambda}_i, h_\epsilon(\boldsymbol{\lambda}_i)) \,|\, i = 1, 2, ..., N\}. \tag{6.18}$$

This interpolant is given by the waveform difference mean $\mu(\boldsymbol{\lambda})$, with associated variance $\sigma^2(\boldsymbol{\lambda})$; it provides a new GPR-informed waveform model

$$h_{\mathrm{GPR}} := h_{\mathrm{app}} - \mu, \tag{6.19}$$

which approximates h_{acc} via (6.4). Equation (6.15) also supplies the prior density for h_ϵ in (6.5), which evaluates to

$$\mathcal{L} \propto \frac{1}{1+\sigma^2} \exp\left(-\frac{1}{2}\frac{\langle x - h_{\mathrm{GPR}} | x - h_{\mathrm{GPR}}\rangle}{1+\sigma^2}\right). \tag{6.20}$$

The GPR marginalised likelihood has several desirable features for parameter estimation. A maximum-likelihood estimation of $\boldsymbol{\lambda}$ with (6.20) gives

$$\langle (\partial h_{\mathrm{GPR}})_b | n - h_\epsilon + \mu - (\partial h_{\mathrm{GPR}})_a (\lambda_\epsilon)^a \rangle = 0 \tag{6.21}$$

from (6.1) and (6.19) with $x = h_{\mathrm{acc}}(\boldsymbol{\lambda}_{\mathrm{true}}) + n$ and $\boldsymbol{\lambda}_{\mathrm{ML}} = \boldsymbol{\lambda}_{\mathrm{true}} + \boldsymbol{\lambda}_\epsilon$, and so

$$\begin{aligned}(\lambda_\epsilon)^a &= (\Gamma^{-1}_{\mathrm{GPR}})^{ab}\langle n | (\partial h_{\mathrm{GPR}})_b\rangle - (\Gamma^{-1}_{\mathrm{GPR}})^{ab}\langle h_\epsilon | (\partial h_{\mathrm{GPR}})_b\rangle \\ &\quad + (\Gamma^{-1}_{\mathrm{GPR}})^{ab}\langle \mu | (\partial h_{\mathrm{GPR}})_b\rangle,\end{aligned} \tag{6.22}$$

where the third term is proportional to the GPR interpolant μ and acts to cancel the second term by design. This correction greatly reduces the systematic bias due to theoretical error, provided the interpolant is performing optimally (i.e. $\mu \approx h_\epsilon$) near $\boldsymbol{\lambda}_{\mathrm{true}}$.

Another safeguard against theoretical error is the presence of the GPR variance σ^2 in (6.20). This variance is $\ll 1$ when $\mu \approx h_\epsilon$, but may become ~ 1 far from all training-set points, or in the case of a suboptimally chosen training set or covariance function. The consequent broadening of the Gaussian in (6.20) is then conservative in nature, as it usually prevents the true parameter values from being excluded at high significance.

Lastly, the premise of the GPR approach is based on the availability of accurate waveforms h_{acc} that are extremely expensive to compute, and hence unsuitable for use in Monte Carlo search algorithms with the standard likelihood. The marginalised likelihood remains computationally tractable while including information from h_{acc}, since it only uses the approximate waveforms h_{app} and adds to them some linear combination of precomputed waveform differences via (6.16). Any extra online computational cost from using the marginalised likelihood thus scales linearly with the size N of the training set. The scaling coefficient (relative to the cost of (2.59) with h_{app}) is typically small; for the analyses in Section 6.3, it is $\sim 10^{-3}$.

6.2.2 Training the Gaussian process

With the zero-mean assumption in (6.6), the waveform difference model is fully specified by the covariance function k. The standard approach is to define a functional form for k that depends on a number of hyperparameters $\boldsymbol{\theta}$, and to select values for $\boldsymbol{\theta}$ by training the Gaussian process with information from the set $\mathcal{D}$. A covariance function $k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ is stationary if it depends only on the relative position $\boldsymbol{\lambda} - \boldsymbol{\lambda}'$ of the two parameter points; it is furthermore isotropic if it depends only on

$$\tau^2 := g_{ab}(\lambda - \lambda')^a(\lambda - \lambda')^b, \tag{6.23}$$

where the g_{ab} are the $\ell(\ell + 1)/2$ independent components of some constant parameter-space metric g on Λ.

An investigation of various common isotropic (hence stationary) covariance functions in the GW context finds the performance of the GPR interpolant and the marginalised likelihood to be fairly robust against changes in the functional form for k [180]. Hence we consider a single fixed form in this chapter: the squared-exponential covariance function

$$k_{\mathrm{SE}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^2\right), \tag{6.24}$$

which is the smooth limiting case for several different families of covariance functions. The hyperparameters for the model $\mathcal{GP}(0, k_{\mathrm{SE}})$ then comprise only the metric components g_{ab} and some overall scale factor σ_f^2.

As the size of the training set $\mathcal{D}$ increases, the covariance matrix $\mathbf{K}$ rapidly becomes ill-conditioned, even for a modestly sized set with $N \gtrsim 10$. This is partly mitigated by the addition of noise to $\mathcal{D}$, such that the GPR interpolant need only pass close to (rather than through) each training-set point. Keeping σ_f^2 as the overall scale factor, we transform

$$\mathbf{K} \to \mathbf{K} + \sigma_f^2\sigma_n^2\mathbf{I}_N, \tag{6.25}$$

where $\mathbf{I}_N$ is the identity matrix, and the fractional noise variance σ_n^2 of each training-set point is taken to be uniform and fixed (i.e. not treated as a hyperparameter). In practical terms, the transformation (6.25) effectively reduces the condition number of $\mathbf{K}$, thereby facilitating its numerical inversion. We use $\sigma_n^2 = 10^{-4}$ throughout this work, which is the smallest value compatible with all of the $N \lesssim 100$ training sets considered in Section 6.3.

The most straightforward method of selecting the Gaussian process hyperparameters $\boldsymbol{\theta}$ is through maximum-likelihood estimation with the hyperlikelihood $Z(\boldsymbol{\theta}|\mathcal{D}) = p(\{h_\epsilon(\boldsymbol{\lambda}_i)\})$ in (6.7), i.e. the likelihood for the hyperparameters given the training set. In other words, an optimal set of hyperparameters $\boldsymbol{\theta}_{\mathrm{ML}}$ is obtained by maximising the log-hyperlikelihood

$$\ln Z = -\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{M}) - \ln\det\mathbf{K} + \mathrm{const} \tag{6.26}$$

over the hyperparameter space Θ.

Part of this maximisation may be done analytically, since the overall scale σ_f^2 factors out of the matrix expressions in $\ln Z$ [181]. In the case of (6.26), $\ln Z$ with $\mathbf{K} = \sigma_f^2\hat{\mathbf{K}}$ achieves a maximum in σ_f^2 at

$$\sigma_f^2 = \frac{1}{2N}\mathrm{tr}(\hat{\mathbf{K}}^{-1}\mathbf{M}). \tag{6.27}$$

Substituting (6.27) back into (6.26) then gives a scale-invariant form for $\ln Z$:

$$\ln Z = -N\ln\mathrm{tr}(\hat{\mathbf{K}}^{-1}\mathbf{M}) - \ln\det\hat{\mathbf{K}} + \mathrm{const}. \tag{6.28}$$

Equations (6.27) and (6.28) effectively reduce the dimensionality of the hyperparameter space by one (to $\dim\Theta = \ell(\ell+1)/2$ for the model $\mathcal{GP}(0, k_{\mathrm{SE}})$), which is useful for the low-dimensional searches conducted in Section 6.3.

6.2.3 Application to black-hole binary mergers

In [180], the viability of the GPR marginalised likelihood for improving GW parameter estimation is demonstrated through a one-dimensional ($\ell = 1$) study, using waveforms for merging black-hole binary systems with comparable component masses $(m_1, m_2 \geq m_1)$. Two waveform approximants implemented in the LIGO Scientific Collaboration Algorithm Library [186] are considered: the phenomenologically fitted IMRPhenomC model [187] and the analytic Taylor-F2 model [188], which are taken as accurate and approximate respectively. Even though these two waveforms are qualitatively different (IMRPhenomC describes the full inspiral–merger–ringdown while Taylor-F2 is inspiral-only), the marginalised likelihood functions as expected in reducing systematic bias.

Equation (6.20) is used to estimate the chirp mass (2.31) from synthetic data $x = h_{\mathrm{acc}}(\boldsymbol{\lambda}_{\mathrm{true}})$ (no detector noise), where $h_{\mathrm{acc}}(\boldsymbol{\lambda}_{\mathrm{true}})$ is an injected IMRPhenomC signal with $\mathcal{M}_{\mathrm{true}} = 5.045M_{\odot}$ and fixed mass ratio $q = 0.75$. As the density of the training set (with respect to some metric on Λ) is expected to be the strongest determinant of interpolation performance, two different grid sizes in $\mathcal{M}$ are considered: $\Delta\mathcal{M} = 10^{-2}M_{\odot}$ and $\Delta\mathcal{M} = 5 \times 10^{-3}M_{\odot}$. The corresponding points are gridded uniformly around $\mathcal{M}_{\mathrm{true}}$ across the range $5 \leq \mathcal{M}/M_{\odot} \leq 5.6$, such that the density of the training set is varied by fixing its span and changing its cardinality.

A GPR model with the squared-exponential covariance function (6.24) is then trained on both sets by optimising the single independent metric hyperparameter, which is chosen more intuitively in the $\ell = 1$ case to be the covariance length $\delta\mathcal{M} := (g_{\mathcal{M}\mathcal{M}})^{-1/2}$. In general (and for the considered sets), the optimal value of $\delta\mathcal{M}$ shifts as the density and cardinality of the training set are varied, but typically by less than a factor of two. The performance of the marginalised likelihood is found to be similarly robust against the choice of $\delta\mathcal{M}$ for a given training set.

Unsurprisingly, the performance of the marginalised likelihood is improved for the denser training set. Higher fidelity between the GPR waveform (6.19) and the accurate waveform is obtained across the span of the set; this is quantified by overlaps $\mathcal{O}(h_{\mathrm{GPR}}|h_{\mathrm{acc}})$ that are $\gtrsim 0.999$. The variance σ^2 associated with the waveform difference interpolant is smaller for the denser training set as well, with values that are $\lesssim 10^{-3}$ relative to σ_f^2 (which is the limiting value of σ^2 outside the span of the set). A maximum-likelihood estimation of $\mathcal{M}$ with the corresponding marginalised likelihood is therefore closer to the true value, and better constrained. The sparser training set in this study gives $\mathcal{O}(h_{\mathrm{GPR}}|h_{\mathrm{acc}}) \gtrsim 0.985$ and $\sigma^2/\sigma_f^2 \lesssim 10^{-2}$, with a marginalised likelihood that is shifted away and broader but still functional. As will be discussed in Section 6.3, this is because its density is close to some threshold determined by the optimal value of $\delta\mathcal{M}$ (which is largely independent of training-set density).

Different source SNRs in the range $8 \leq \rho \leq 64$ are also considered in this study. The marginalised likelihood for the sparser training set reduces the systematic error in the maximum-likelihood estimation of $\mathcal{M}$ from $\mathcal{M}_\epsilon = 5 \times 10^{-3} M_\odot$ (around 10-sigma for a typical LIGO source with $\rho = 16$) to $\mathcal{M}_\epsilon = 9 \times 10^{-4} M_\odot$; it also broadens to remain consistent with $\mathcal{M}_{\mathrm{true}}$ at 2-sigma, even at high SNR. These results are obtained in the regime where the overlaps between the accurate and approximate waveforms are ≈ 0.35 across the span of the training set. Although theoretical error is reduced if the approximate model is improved, results in Section 6.3 show that the marginalised likelihood remains needed for overlaps as high as ≈ 0.97, which may still lead to significant systematic bias for a typical EMRI with $\rho = 30$.

6.3 Application to extreme-mass-ratio inspirals

The detection and characterisation of EMRIs is a formidable challenge in GW data analysis, especially in the broader context of resolving these sources from a LISA data set that is likely to contain an (over)abundance of long-lived and overlapping signals [189]. Even as a standalone problem, searches of the EMRI parameter space are greatly hindered by its large volume (with respect to the Fisher information metric), which requires $\sim 10^{30}$ waveforms for full coverage in a template bank approach [168]. This is exacerbated by the long and unwieldy waveforms themselves; a sampling rate of 0.2 Hz (the characteristic frequency for an EMRI with a $10^6 M_\odot$ central black hole) yields $\sim 10^7$ samples for both channels of a year-long waveform.

Due to the $O(N^3)$ cost of computing the Cholesky decomposition for $\mathbf{K}$ in (6.16) and (6.17), it is clearly impractical—if not impossible—to cover any significant fraction of parameter space with a single training set. The present purpose of the GPR marginalised likelihood is thus restricted to precise parameter estimation in highly localised regions of parameter space. Furthermore, if the GPR approach is to be useful for EMRIs at all, the typical separation of points in the training set must be greater than the Fisher metric lengths of the accurate waveform model, which determine the characteristic resolution of grid or stochastic searches with that model [134].

A simple argument shows that this is normally the case for EMRI waveforms with $\rho > 1$. We consider a small neighbourhood of some point $\boldsymbol{\lambda}_0 \in \Lambda$, along with a covariance metric g for a Gaussian process that optimally describes the distribution of h_ϵ in that neighbourhood. The metric defines the short covariance lengths $\delta\lambda^a := (g_{aa})^{-1/2}$, i.e. the half widths of the associated hyperellipsoid in each one-dimensional parameter subspace through $\boldsymbol{\lambda}_0$. These lengths place upper bounds on the characteristic grid sizes of the training set and lower bounds on its span; an incompatible training set typically yields no peak in the log-hyperlikelihood surface (6.26), and so the regression becomes suboptimal.

From the optimality of the Gaussian process, the covariance lengths associated with g approximate the correlation lengths of h_ϵ itself. Hence we have $\delta\lambda^a \sim (\delta\lambda_{\mathrm{over}})^a$, where the vector of overlap lengths $\delta\boldsymbol{\lambda}_{\mathrm{over}}$ is defined to satisfy

$$\langle h_\epsilon(\boldsymbol{\lambda}_0)|h_\epsilon(\boldsymbol{\lambda}_0 + \mathbf{P}_a\delta\boldsymbol{\lambda}_{\mathrm{over}})\rangle = 0 \tag{6.29}$$

for each a, with $\mathbf{P}_a$ projecting $\delta\boldsymbol{\lambda}_{\mathrm{over}}$ onto the subspace corresponding to λ^a. At leading order, we then have

$$\delta\lambda^a \sim \frac{\langle h_\epsilon|h_\epsilon\rangle}{|\langle h_\epsilon|(\partial h_\epsilon)_a\rangle|} = \frac{\rho_\epsilon \sec\phi^a}{\sqrt{(\Gamma_\epsilon)_{aa}}}, \tag{6.30}$$

where ∂h_ϵ and $\boldsymbol{\Gamma}_\epsilon$ respectively denote the derivative vector and Fisher information matrix of the waveform difference, and ϕ^a is the principal inner-product angle between h_ϵ and $(\partial h_\epsilon)_a$. The covariance lengths do not scale with the SNR ρ_ϵ of the waveform difference, since the final denominator in (6.30) is $\propto \rho_\epsilon$.

We now consider the Fisher metric lengths of the waveform difference. As pointed out in [180], the Fisher lengths of the difference between two waveform models are generally larger than those of the individual models, especially if both models generate waveforms with high overlap. This might be due to the waveform difference having a lower SNR, or being more broadly correlated across parameter space. Since the first of these two factors is known, we may separate it from the analysis and consider the Fisher lengths of the unit-SNR waveform difference instead. For the EMRI examples that follow in this section, these are found to be comparable to the Fisher lengths of the unit-SNR waveforms, which is simply the statement that the waveform difference varies across parameter space in similar fashion to the waveforms themselves.

Since the short (i.e. defined analogously to $\delta\lambda^a$) Fisher metric lengths of the unit-SNR waveform difference are given by

$$(\delta\lambda_{\mathrm{Fish}})^a := \frac{\rho_\epsilon}{\sqrt{(\Gamma_\epsilon)_{aa}}}, \tag{6.31}$$

it follows that $\delta\lambda^a \gtrsim (\delta\lambda_{\mathrm{Fish}})^a$. In general, any waveform derivative with respect to a parameter that only affects amplitude will have $\phi^a \approx 0$, such that $\delta\lambda^a \sim (\delta\lambda_{\mathrm{Fish}})^a$. Nevertheless, the SNR-normalised Fisher lengths for the waveform difference (or the waveforms themselves in the EMRI case) give rough order-of-magnitude estimates for the optimal covariance lengths, and are more straightforward to obtain. These estimates may be used to provide initial guesses for the diagonal metric hyperparameters when maximising the log-hyperlikelihood with standard optimisation routines.

Finally, we compare the optimal covariance lengths $\delta\lambda^a$ to the unnormalised Fisher metric lengths of the accurate waveform model, which for EMRIs are given approximately by $(\delta\lambda_{\mathrm{Fish}})^a/\rho$ through the above argument. The former lengths determine the maximal separation of points in the training set, while the latter describe the support of the accurate posterior probability density. Since $\delta\lambda^a \gtrsim (\delta\lambda_{\mathrm{Fish}})^a$, it follows for $\rho > 1$ that the lowest training-set density allowed (while ensuring interpolation optimality) will typically be far below the sampling density required to resolve the posterior surface. Hence the GPR approach should be viable for EMRIs, and its computational benefits are enhanced for sources with higher SNR.

The validity of this simple argument is illustrated through one- and two-dimensional analyses of the marginalised likelihood in Sections 6.3.1 and 6.3.2 respectively. Kludge waveform models are used in this study, due to their availability and computational practicality. Waveforms generated from the NK model in Section 4.2.2 are taken to be accurate, since they have high fidelity with Teukolsky-based waveforms up to an orbital separation of $\approx 5M$. The latest version of the AAK in Section 4.3 is used as the approximate model; it is faster than the NK and matches its early phase evolution with high waveform overlap, but dephases gradually as the compact object approaches plunge.

In both sections, the synthetic data x is an injected NK signal from a $10^1 M_\odot$ stellar-mass black hole orbiting a $10^6 M_\odot$ SMBH, with no detector noise added. The signal is two months long and sampled at 0.2 Hz, while the source distance is adjusted such that the SNR is $\rho = 30$ with respect to an analytic approximation for the (down-scoped) LISA noise power spectral density [176]. Other orbital parameters are chosen to ensure that the NK and AAK waveforms have an overlap of ≈ 0.97, so as to investigate the scenario in which the approximate model is fairly accurate to begin with. As will be shown, EMRI parameter estimation is susceptible to significant systematic bias even at this level of theoretical error.

For the given source and waveform parameters, a single evaluation of the NK likelihood takes 29.4 s on average, as compared to an average of 5.66 s per evaluation for the AAK likelihood. This disparity in computational cost is likely to be even greater in more realistic scenarios where, for example, the NK is used as the approximate model and the accurate waveforms are provided by a self-force model. In Section 6.3.1, we verify that the added cost of using the GPR marginalised likelihood over the standard likelihood with approximate waveforms is indeed proportional to the training-set size N, and with a small scaling coefficient (as mentioned in Section 6.2.1).

6.3.1 One-dimensional example

In this section, the GPR marginalised likelihood (6.20) is used to estimate the compact-object mass of an EMRI with $\mu_{\text{true}} = 10^1 M_\odot$, assuming all other parameters are known and fixed at their true values. The covariance metric on the corresponding parameter subspace has a single component $g_{\mu\mu}$, which is optimised through the maximisation of the log-hyperlikelihood (6.26). This one-dimensional example provides a simple illustration of the relationships

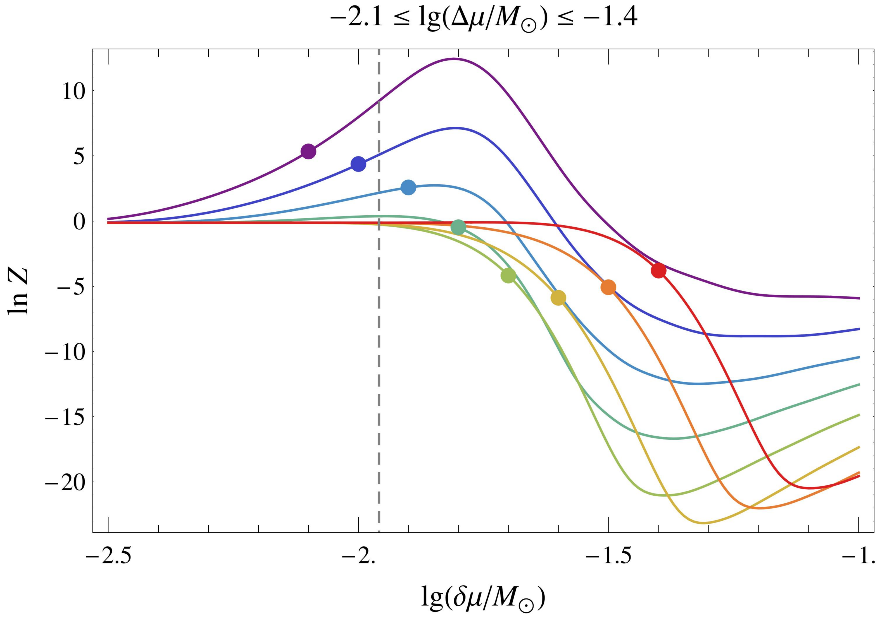

Figure 6.1: Plots of $\ln Z$ against $\lg\left(\delta\mu/M_\odot\right)$ for eight 10-point training sets with grid sizes $-2.1 \leq \lg\left(\Delta\mu/M_\odot\right) \leq -1.4$ (indicated by the abscissae of the solid circles). The vertical dashed line corresponds to the Fisher metric length $\delta\mu_{\mathrm{Fish}}$.

between the optimal covariance length $\delta\mu := (g_{\mu\mu})^{-1/2}$, the Fisher metric length $\delta\mu_{\mathrm{Fish}} := ((\Gamma_\epsilon)_{\mu\mu})^{-1/2}$, and the training-set grid size $\Delta\mu$.

The GPR model is trained on eight 10-point training sets with uniform grids, where the grid sizes are distributed in the range $-2.1 \leq \lg\left(\Delta\mu/M_\odot\right) \leq -1.4$. This range is chosen to encompass the Fisher length of the unit-SNR waveform difference[17], which is approximately constant across the spans of the considered training sets and evaluated at μ_{true} as $\lg\left(\delta\mu_{\mathrm{Fish}}/M_\odot\right) = -1.96$. Each training set is placed such that μ_{true} lies at the geometric centre of its span (and thus maximally far from the nearest points in the set).

Figure 6.1 shows plots of the log-hyperlikelihood for the eight training sets, with the optimal covariance length for each set given by the abscissa of the peak (where it exists). The optimal value $\delta\mu$ is approximately constant for all valid training sets, and falls in the narrow range $\lg\left(\delta\mu_{\mathrm{Fish}}/M_\odot\right) \leq \lg\left(\delta\mu/M_\odot\right) \leq -1.8$. In comparison to the approach of [180] described in Section 6.2.3, vary-

[17] For comparison, the Fisher lengths of the unit-SNR NK and AAK waveforms at μ_{true} are $\lg\left(\delta\mu^{\mathrm{NK}}_{\mathrm{Fish}}/M_\odot\right) = -1.86$ and $\lg\left(\delta\mu^{\mathrm{AAK}}_{\mathrm{Fish}}/M_\odot\right) = -1.85$ respectively.

ing the density of the training set by fixing its cardinality and changing its span also shifts $\delta\mu$ by less than a factor of two, which demonstrates that both span and cardinality have less impact than density on a training set's performance. Hyperlikelihood peaks emerge only for grid sizes $\lg{(\Delta\mu/M_\odot)} \leq -1.8$, indicating that $1/\delta\mu$ determines a minimum threshold for the density of an optimal training set. Finally, $\delta\mu_{\mathrm{Fish}}$ appears to set a lower bound on $\delta\mu$, which is consistent with the discussion around (6.30) and (6.31).

We now consider the marginalised likelihood itself for three other 10-point training sets. Firstly, a set $\mathcal{D}_{\mathrm{Fish}}$ is placed around μ_{true} with grid size $\Delta\mu_{\mathrm{Fish}} = \delta\mu_{\mathrm{Fish}}$; training the GPR model on this set yields an optimal covariance length $\lg{(\delta\mu/M_\odot)} = -1.82$. Two more sets $\mathcal{D}_{\mathrm{cov}}$ and $\mathcal{D}_{\mathrm{2cov}}$ are constructed in the same way, with the grid sizes $\Delta\mu_{\mathrm{cov}} = \delta\mu$ and $\Delta\mu_{\mathrm{2cov}} = 2\delta\mu$ respectively. A different optimal covariance length is found for $\mathcal{D}_{\mathrm{cov}}$, while there is no hyperlikelihood peak for $\mathcal{D}_{\mathrm{2cov}}$. Nevertheless, the performance of the marginalised likelihood is found to be practically constant across the range $\lg{(\delta\mu_{\mathrm{Fish}}/M_\odot)} \leq \lg{(\delta\mu/M_\odot)} \leq -1.8$, and so the Gaussian process is not retrained for $\mathcal{D}_{\mathrm{cov}}$ and $\mathcal{D}_{\mathrm{2cov}}$ (i.e. the same value of $\delta\mu$ is used for all three training sets).

At a source SNR of $\rho = 30$, a high overlap of ≈ 0.97 between the accurate and approximate waveforms still results in a 5-sigma bias due to theoretical error; as seen in Figure 6.2, the approximate likelihood L_{app} is peaked away from μ_{true} with $\mu_\epsilon \approx 3 \times 10^{-3} M_\odot$, while the 1-sigma length for L_{app} (and the accurate likelihood L_{acc}) is $\approx 5 \times 10^{-4} M_\odot$. The marginalised likelihood with the training set $\mathcal{D}_{\mathrm{Fish}}$ is virtually identical to L_{acc}, and is slightly broader with $\mathcal{D}_{\mathrm{cov}}$ but remains peaked near the true value. For the sparsest training set $\mathcal{D}_{\mathrm{2cov}}$, the peak of the marginalised likelihood has an error μ_ϵ similar to that of L_{app}, although it is sufficiently broadened to ensure that it is still consistent with μ_{true} at 2-sigma significance.

The one-dimensional example in this section indicates that the GPR approach might be viable for EMRIs, since even the densest considered training set $\mathcal{D}_{\mathrm{Fish}}$ has a grid size that is significantly larger than the width of the accurate likelihood. By constructing a few additional training sets with different sizes N, the marginalised likelihood is found to take an average of $5.66 + 0.013N\,\mathrm{s}$ per evaluation, i.e. $\approx 2\%$ longer than the approximate likelihood for a 10-point

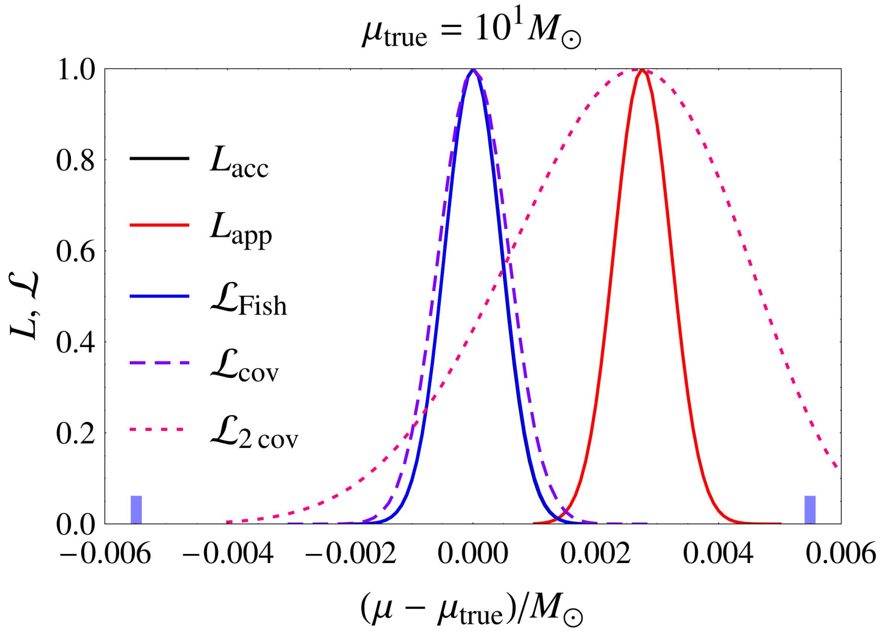

Figure 6.2: One-dimensional likelihood plots for the standard likelihood with accurate and approximate waveforms, and the marginalised likelihood with the training sets $\mathcal{D}_{\text{Fish}}$, $\mathcal{D}_{\text{cov}}$ and $\mathcal{D}_{2\text{cov}}$. The only training-set points within the horizontal plot range belong to the densest set $\mathcal{D}_{\text{Fish}}$, and are indicated by thick marks on the horizontal axis.

training set. However, it provides no computational savings over the accurate likelihood for large training sets (with $\gtrsim 2000$ points in this particular case), which may restrict its usefulness for higher-dimensional problems. This limitation, as well as the generally poor scaling of GPR with dimensionality, is discussed further in Sections 6.3.2 and 6.4.

6.3.2 Two-dimensional example

In this section, the GPR marginalised likelihood (6.20) is used to estimate the component masses of an EMRI with $(\mu, M)_{\text{true}} = (10^1, 10^6)M_\odot$, again assuming all other parameters are known and fixed at their true values. Maximisation of the log-hyperlikelihood (6.26) is now over the three independent components $(g_{\mu\mu}, g_{MM}, g_{\mu M})$ of the covariance metric on the two-dimensional parameter subspace. The eigensystem $\{(\lambda_i, \hat{\mathbf{v}}_i)\,|\,i = 1, 2\}$ of g defines a covariance ellipse with semi-principal axes $\{\lambda_i^{-1/2}\hat{\mathbf{v}}_i\}$ in the usual way.

Although the component-mass subspace is chosen for the two-dimensional example here, a straightforward search in (μ, M) is not necessarily optimal in the context of higher-dimensional parameter estimation. For example, since the central mass $M \approx \mu + M$ strongly determines the characteristic frequency of an EMRI waveform, variations in the waveform difference with respect to M may be reduced by rescaling the time coordinate to dimensionless time t/M. This approach has been investigated, and yields longer (by an order of magnitude or so) covariance lengths as expected. However, it also results in less stable derivatives and poorer interpolation; this is likely because the AAK model maps M to an unphysical mass $\tilde{M}$ that is then evolved through fits (see Section 4.3.2), and so its waveforms vary differently from the NK waveforms with respect to M. If more accurate models are used, the waveform difference will have an infinite covariance length in total mass, such that rescaling the time coordinate by $\mu + M$ reduces the component-mass subspace to a single degree of freedom (e.g. the mass ratio μ/M).

Three different training sets are considered for the (μ, M) example in this section. The first is a (6×6)-point set $\mathcal{D}_{\text{Fish}}$ with $(\mu, M)_{\text{true}}$ lying at the geometric centre of its span; its points are placed uniformly on a grid defined by the semi-principal axes $\{\lambda_i^{-1/2}\hat{\mathbf{v}}_i\}_{\mathrm{Fish}}$ of the Fisher metric ellipse (where $\{(\lambda_i, \hat{\mathbf{v}}_i)\}_{\mathrm{Fish}}$ is the eigensystem of $\boldsymbol{\Gamma}_\epsilon$ for the unit-SNR waveform difference). Two more (10×10)-point training sets $\mathcal{D}_{\mathrm{dense}}$ and $\mathcal{D}_{\mathrm{sparse}}$ are constructed on rectangular grids, with the grid sizes given by the short and long Fisher lengths respectively, i.e.

$$(\Delta\mu, \Delta M)_{\mathrm{dense}} := \left(\frac{1}{\sqrt{(\Gamma_\epsilon)_{\mu\mu}}}, \frac{1}{\sqrt{(\Gamma_\epsilon)_{MM}}}\right), \tag{6.32}$$

$$(\Delta\mu, \Delta M)_{\mathrm{sparse}} := \left(\sqrt{(\Gamma_\epsilon^{-1})_{\mu\mu}}, \sqrt{(\Gamma_\epsilon^{-1})_{MM}}\right). \tag{6.33}$$

As justified in Section 6.3.1, the GPR model is trained on a single training set ($\mathcal{D}_{\mathrm{dense}}$ in this case), and the same optimal covariance ellipse is subsequently used for all three sets. The relative placement of points in the three training sets is shown in Figure 6.3, along with the covariance and Fisher ellipses. Both ellipses are aligned and the Fisher ellipse is slightly smaller, which is consistent with the discussion around (6.30) and (6.31).

From the contour plots in Figure 6.4, the measurement of (μ, M) with the approximate likelihood L_{app} has a theoretical error of $(\mu_\epsilon, M_\epsilon) \approx (2 \times 10^{-3}, 6 \times 10^0)M_\odot$, and excludes $(\mu, M)_{\mathrm{true}}$ at beyond 2-sigma significance. The marginalised likelihood with the training set $\mathcal{D}_{\mathrm{dense}}$ is virtually identical to the accurate likelihood L_{acc}; so too is the likelihood for $\mathcal{D}_{\mathrm{Fish}}$, which is sparser and contains fewer points. More surprisingly, the training set $\mathcal{D}_{\mathrm{sparse}}$ also yields a likelihood that is very similar to L_{acc}, which indicates that a training-set density no lower than that corresponding to the long Fisher metric lengths (i.e. the half extents of the unit-SNR Fisher ellipse in each parameter) will still be optimal on the level of the marginalised likelihood. However, it may be difficult to learn the optimal covariance metric from such a training set if it is too sparse or contains too few points.

It is clear that a simple rectangular grid approach to the placement of training set points will not scale well with the dimensionality ℓ of the parameter space, but uniform placement on a grid defined by the Fisher metric eigensystem is also limited at moderately large ℓ. From our one- and two-dimensional studies, as well as preliminary investigations of a three-dimensional problem,

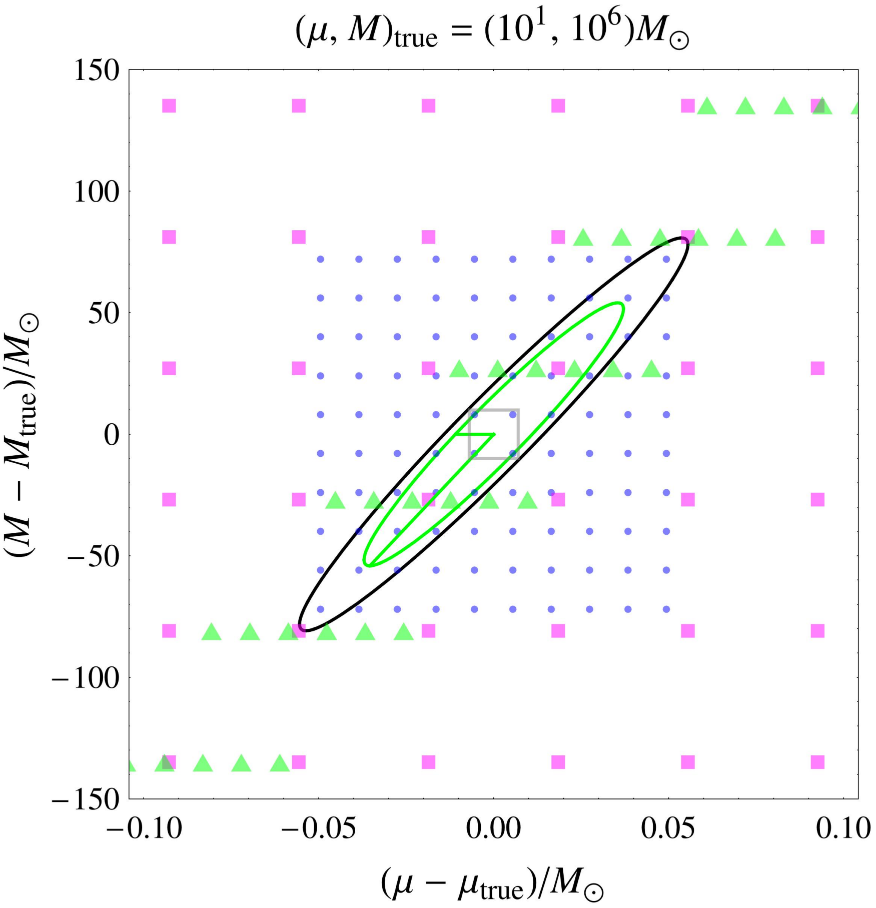

Figure 6.3: Training-set point placement around $(\mu, M)_{\text{true}}$ for $\mathcal{D}_{\text{dense}}$ (dots), $\mathcal{D}_{\text{Fish}}$ (triangles) and $\mathcal{D}_{\text{sparse}}$ (squares). The grid for $\mathcal{D}_{\text{Fish}}$ is defined by the semi-principal axes of the Fisher metric ellipse (green), and is aligned with the optimal covariance ellipse (black) learnt from $\mathcal{D}_{\text{dense}}$. The central grey square corresponds to the plot range of Figure 6.4.

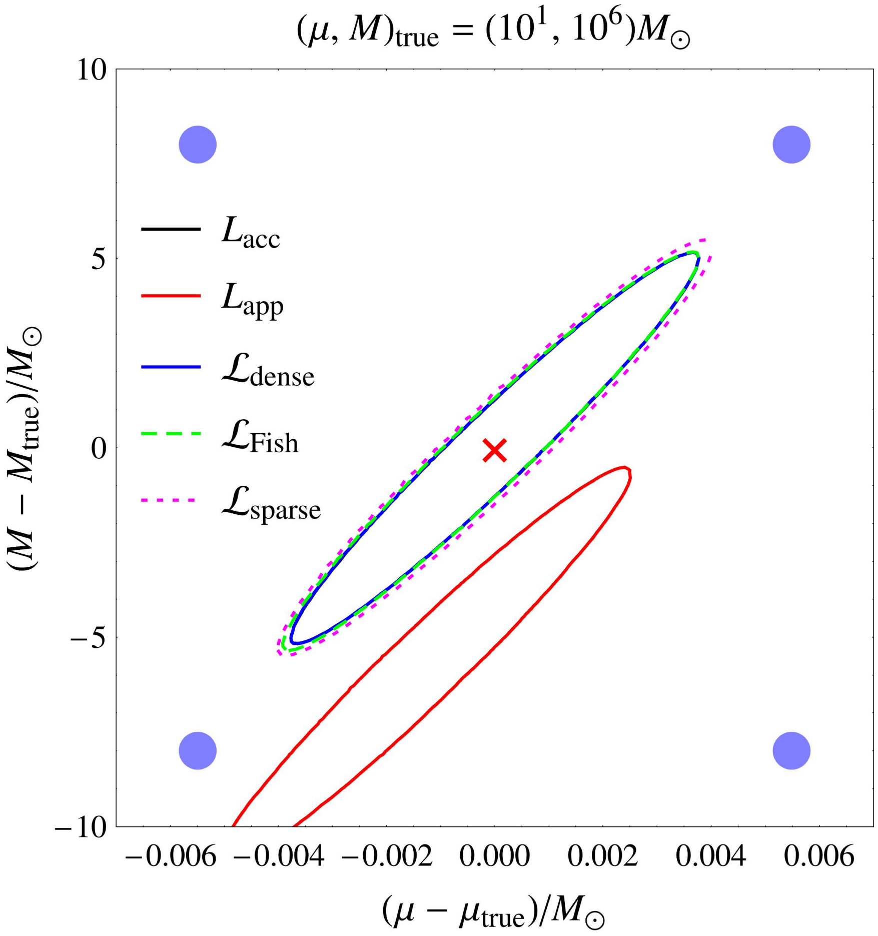

Figure 6.4: Two-dimensional likelihood contour plots for the standard likelihood with accurate and approximate waveforms, and the marginalised likelihood with the training sets $\mathcal{D}_{\rm dense}$, $\mathcal{D}_{\rm Fish}$ and $\mathcal{D}_{\rm sparse}$. All contours are 2-sigma. The only training-set points within the plot range belong to $\mathcal{D}_{\rm dense}$, and are indicated by solid circles.

six points along each eigendirection appears to be the bare minimum for learning a covariance metric that well describes the waveform difference locally. This necessitates $O(N^3)$ operations on a $6^\ell \times 6^\ell$ covariance matrix in the training stage, which is computationally challenging for $\ell > 5$.

If a suitable covariance metric can be obtained, the actual set of waveform differences used in the interpolation stage does not have to be quite as large or dense as the set used to train the GPR model. A set with just three points along each eigendirection becomes unfeasible at $\ell > 8$, since the marginalised likelihood is likely to be more expensive than the accurate likelihood for $N \sim 10^4$; still, this is enough to estimate the seven intrinsic EMRI parameters (see Section 4.2). In general, the dimensionality problem for the GPR approach might be tackled by reducing the computational cost of operations on the $N \times N$ covariance matrix, or by reducing N itself while ensuring the training set is functional. Possible methods for achieving these are beyond the scope of the present work, but are discussed in the concluding remarks of Section 6.4.

6.4 Discussion

In this chapter, we have introduced the GPR marginalised likelihood and its applications in GW data analysis, with a focus on investigating its viability for EMRI parameter estimation through low-dimensional studies. Even in the considered scenario where the waveform model used for parameter estimation has a ≈ 0.97 match with the source signal at the true parameter values, significant systematic bias from theoretical error will still arise for high-SNR ($\rho \gtrsim 30$) sources. The GPR approach is shown to mitigate this bias, and hence to be suitable for improving the accuracy of EMRI parameter estimation (albeit in highly localised regions of parameter space).

The performance of the marginalised likelihood is strongly dependent on its precomputed training set of waveform differences, which relies on the existence of a more accurate waveform model that reproduces the source signal with high fidelity. For the method to be practical, the density of the training set must be significantly lower than the resolution of a posterior search with the accurate waveforms. This is shown to be the case for EMRIs through a simple argument, and is verified with one- and two-dimensional examples. Another key result in these sections is a demonstration of how the Fisher information matrix of the waveform difference (normalised to unit SNR) may be used to inform the placement of training-set points, as well as to estimate a covariance metric that describes the waveform difference locally.

While the marginalised likelihood shows early promise for EMRI parameter estimation, it is akin to other applications of GPR in being subject to the curse of dimensionality. The number of training-set points required to search an ℓ-dimensional parameter subspace grows exponentially with ℓ, which hinders not just the offline training stage (since the covariance matrix is larger and more ill-conditioned), but also the online interpolation stage (where a new linear combination of points is computed for each likelihood evaluation).

One possible approach to these computational problems is to replace the squared-exponential covariance function (6.24) in the GPR model with a covariance function that has compact support on parameter space (e.g. the Wendland polynomials [190]), such that the covariance matrix becomes sparse. Iterative methods [191] may then be used to accelerate the Cholesky decomposition of $\mathbf{K}$ in (6.16), (6.17) and (6.26). A compact-support covariance function also reduces the number of training-set points summed in (6.16), which directly determines the evaluation speed of the marginalised likelihood.

Another strategy is to minimise the size of the training set itself. As seen throughout Section 6.3, the covariance metric that is learnt during the training stage also determines a fixed threshold for the density of a training set that functions optimally at the interpolation stage. However, it may be possible to lower this threshold density through reparametrisation or dimensionality reduction methods, e.g. the component-mass example discussed in Section 6.3.2. In the case of parameters for which this is not feasible, the number of training-set points used to cover the region of relevance may still be reduced through a non-uniform placement of points, or non-geometric prescriptions such as stochastic placement algorithms [192–194]. Full coverage of the search region with a precomputed training set might not even be necessary; one possibility is to use a smaller "moving" set that is updated adaptively as the marginalised likelihood is sampled with Markov-chain-Monte-Carlo methods.

Chapter 7

Conclusion

The historic first direct detection of GWs by Advanced LIGO at the end of 2015 [8] was but the opening event in the quick succession of exciting developments that have since unfolded in the field. Within the short span of 18 months, we have witnessed a second confirmed black-hole binary merger [9] and the unequivocal success of the LISA Pathfinder proof-of-technology mission [18], while the outlook for an impending detection with pulsar timing arrays is also much improved [20]. Against the backdrop of this rapidly evolving landscape, it is hoped that the work presented in this dissertation constitutes a modest but timely contribution to the burgeoning field of GW astronomy.

In Chapter 3, we have found that Einstein–Maxwell interactions between far-field GWs and electromagnetic fields are likely too weak for the practical purposes of GW detection and parameter estimation, except under conditions that might be contrived or extreme by astrophysical standards (e.g. those considered in Section 3.4). A characterisation of the electromagnetic perturbations driven by strong-field gravitational radiation should yield effects that are more pronounced, and will allow the modelling of GW sources in greater detail. However, EMWs induced by GWs in the strong-field regime are restricted to gravitational frequency bands ($\lesssim 10^3\,\mathrm{Hz}$) [96, 123], which presents the same difficulties for direct detection as mentioned in Section 3.4.1.

The framework in Section 3.3 was originally intended as an intermediate step towards adapting the 1+3 approach for strong-field calculations, but turned out to be analytically intractable even for the far-field problem in spherical coordinates, due to the mixing of spherical harmonic modes by the tensor–vector contractions in (3.38) and (3.39). An alternative research direction in the area of Einstein–Maxwell interactions would be to unify or extend existing semi-analytical methods [96, 97, 123] for Schwarzschild/Kerr spacetimes. Unification should be achievable in principle, as a common (and perhaps inevitable) trait of such methods is a decomposition of the coupled fields into secondary variables distributed among the axial [47] and polar [48] sectors.

The work presented in Chapter 4 is of more pressing interest to the GW community, due the scarcity of accurate and computationally efficient models that describe the information-rich waveforms from EMRIs. We have developed an augmented variant of a well-known kludge waveform model; the new AAK model retains the speed of its predecessor, while matching the phase evolution of more accurate kludges over a significant fraction of the inspiral. With a (meta)stable version of the model released online as part of a kludge software suite [135], AAK waveforms will hopefully see widespread use in the next round of mock LISA data challenges.

EMRIs are an attractive research topic due to the inherent challenges they pose for source modelling and data analysis, and there is much that remains to be done in terms of populating the EMRI waveform inventory before the LISA mission design is finalised at the end of this decade. The immediate improvements that must be made to the AAK model are workaround solutions to the problems mentioned in Section 4.4. For example, the ill-defined nature of (4.37)–(4.39) at plunge may be resolved (inelegantly) by simply starting or stopping the trajectory integration as close to plunge as numerically allowed. Other modifications will involve streamlining the model to make it as fast and robust as possible for the mock data challenges, e.g. the development of an optimised method that detects and handles plunge with a minimal amount of computational overhead, or the improvement of the fitting algorithms so as to reduce the amount of tuning presently required.

Another important research direction in the area of kludge waveform models is to bridge the present disconnect between data analysis work and the gravitational self-force programme. The most accurate kludges currently available are fitted to Teukolsky-based models, and hence account only for the adiabatic, dissipative first-order self-force; a logical next step is to incorporate the subleading “post-adiabatic” self-force phenomenon of resonances. Efforts have been made to explore the prevalence of resonance jumps [195] and their impact on EMRI detection rates [196], but the effect of resonances on the waveforms themselves has yet to be studied.

It is also crucial to set up a channel enabling kludges to be informed by higher-order self-force results once they become available. Early steps towards this goal have been made for an EMRI with a non-spinning central black hole [68]; in this work, the compact object's orbit is described as an evolving geodesic whose parameters are governed by forced equations of motion, and it is explicitly computed by supplying a model for the self-force on Schwarzschild orbits. A similar result may now be achieved for spinning black holes, by using pieces that have recently become available: the forced equations of motion for a Kerr black hole [67] and the self-force on an eccentric, equatorial Kerr orbit [197].

The third major theme of this doctoral thesis is the development of improved techniques for GW data analysis, with the two overlapping but qualitatively different procedures of detection and parameter estimation covered in Chapters 5 and 6 respectively. Future research directions under this theme are the most open-ended, as they typically involve the application of new statistical or computational methods to data analysis, as well as the study of their efficacy using realistic source signals and detector noise.

Various families of tunable compression schemes for GW template banks have been proposed in Chapter 5. In particular, the optimised partition of a template bank into non-overlapping sets performs better than other considered groupings with intersection (which are still interesting from a purely combinatorial perspective), and also yields moderate gains over a straightforward coarsening of the bank. By investigating the features of the method on a bank of $\sim 10^4$ PN waveforms, the first step towards advancing its usage in GW detection has been taken; however, the development of a practical implementation and a more extensive comparison to existing orthogonal-decomposition methods [164–166] is necessary for its adoption. Nevertheless, the generality of these compression schemes means that the method need not stand on its own, and may instead be integrated with existing algorithms for further controlled compression (if the benefits are found to be worthwhile).

In Chapter 6, we have found through low-dimensional parameter estimation studies that the GPR marginalised likelihood might be viable for improving the accuracy of searches in highly localised regions of EMRI parameter space. Possible strategies for the generalisation of the method to higher dimensions are discussed in Section 6.4. Apart from working towards realistic parameter estimation in all dimensions of the full (or at least intrinsic) parameter space, another target is to ensure that the highly localised prior support required for the likelihood to be used in sampling algorithms is compatible with the rough localisation provided by detection-stage searches. Methods that address the curse of dimensionality might simultaneously achieve this, but the problem can also be tackled from the detection end, e.g. by employing a semi-coherent search with short segments of AAK templates that are highly accurate over their duration, such that better localisation accuracy is obtained.

There are other interesting research directions associated with developing the GPR marginalised likelihood: for example, extending it in a physically motivated way to account for the varying degree of error that is expected across different frequency bins. The more general method of GPR waveform interpolation also has possible applications that are unrelated to parameter estimation, such as the smooth combination of waveform models incorporating different features (e.g. a non-spinning eccentric model for comparable-mass mergers and a circular one with spin).

Appendix A

Induced Maxwell field solution

For $\rho \sec\theta \ll 1$ and homogeneous initial conditions, the solution to (3.34)–(3.37) with the background GW in (3.40), the free EMW in (3.55) and (3.56), and the induced electromagnetic field ansatz in (3.57) and (3.58) is given by

$$\mathcal{E}_a^{(\pm)} = hE^{(0)}\xi^{(\pm)}(t)e^{i\psi}P_a^{(\pm)}, \tag{A.1}$$

$$\mathcal{B}_a^{(\pm)} = hE^{(0)}\xi^{(\pm)}(t)e^{i\psi}Q_a^{(\pm)}, \tag{A.2}$$

where

$$\begin{aligned} \xi^{(\pm)}(t) &= m^{(\pm)}\exp\left(-i(n \pm k)t\right) \\ &\quad -m^{(\pm)}\cos\left(m^{(\pm)}t\right) + i(n \pm k)\sin\left(m^{(\pm)}t\right), \end{aligned} \tag{A.3}$$

$$P_0^{(\pm)} = Q_0^{(\pm)} = 0, \tag{A.4}$$

$$
\begin{aligned}
P_1^{(\pm)} = {} & \frac{i\sin^2(\theta/2)}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
& \times \big[2n^2\sin(2\alpha-\gamma-3\phi)+6n^2\sin(2\alpha+\gamma-3\phi) \\
& -n^2\sin(2\alpha-\gamma-2\theta-3\phi)+n^2\sin(2\alpha+\gamma-2\theta-3\phi) \\
& +4n^2\sin(2\alpha+\gamma-\theta-3\phi)+4n^2\sin(2\alpha+\gamma+\theta-3\phi) \\
& -n^2\sin(2\alpha-\gamma+2\theta-3\phi)+n^2\sin(2\alpha+\gamma+2\theta-3\phi) \\
& -2(8k^2+10kn+3n^2)\sin(2\alpha-\gamma-\phi) \\
& -2n(2k+n)\sin(2\alpha+\gamma-\phi)-n^2\sin(2\alpha-\gamma-2\theta-\phi) \\
& +n^2\sin(2\alpha+\gamma-2\theta-\phi)-2n(k+2n)\sin(2\alpha-\gamma-\theta-\phi) \\
& -2kn\sin(2\alpha+\gamma-\theta-\phi)-2n(k+2n)\sin(2\alpha-\gamma+\theta-\phi) \\
& -2kn\sin(2\alpha+\gamma+\theta-\phi)-n^2\sin(2\alpha-\gamma+2\theta-\phi) \\
& +n^2\sin(2\alpha+\gamma+2\theta-\phi)\big],
\end{aligned}
\tag{A.5}
$$

$$
\begin{aligned}
P_2^{(\pm)} = {} & \frac{i\sin^2(\theta/2)}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
& \times \big[2n^2\cos(2\alpha-\gamma-3\phi)+6n^2\cos(2\alpha+\gamma-3\phi) \\
& -n^2\cos(2\alpha-\gamma-2\theta-3\phi)+n^2\cos(2\alpha+\gamma-2\theta-3\phi) \\
& +4n^2\cos(2\alpha+\gamma-\theta-3\phi)+4n^2\cos(2\alpha+\gamma+\theta-3\phi) \\
& -n^2\cos(2\alpha-\gamma+2\theta-3\phi)+n^2\cos(2\alpha+\gamma+2\theta-3\phi) \\
& +2(8k^2+10kn+3n^2)\cos(2\alpha-\gamma-\phi) \\
& +2n(2k+n)\cos(2\alpha+\gamma-\phi)+n^2\cos(2\alpha-\gamma-2\theta-\phi) \\
& -n^2\cos(2\alpha+\gamma-2\theta-\phi)+2n(k+2n)\cos(2\alpha-\gamma-\theta-\phi) \\
& +2kn\cos(2\alpha+\gamma-\theta-\phi)+2n(k+2n)\cos(2\alpha-\gamma+\theta-\phi) \\
& +2kn\cos(2\alpha+\gamma+\theta-\phi)+n^2\cos(2\alpha-\gamma+2\theta-\phi) \\
& -n^2\cos(2\alpha+\gamma+2\theta-\phi)\big],
\end{aligned}
\tag{A.6}
$$

$$
\begin{aligned}
P_3^{(\pm)} &= \frac{in\sin\theta}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
&\times\left[2(3k+n)\sin\left(2\alpha-\gamma-2\phi\right)+2(k-n)\sin\left(2\alpha+\gamma-2\phi\right)\right. \\
&-3k\sin\left(2\alpha-\gamma-\theta-2\phi\right)+k\sin\left(2\alpha+\gamma-\theta-2\phi\right) \\
&-3k\sin\left(2\alpha-\gamma+\theta-2\phi\right)+k\sin\left(2\alpha+\gamma+\theta-2\phi\right) \\
&-n\sin\left(2\alpha-\gamma+2\theta-2\phi\right)+n\sin\left(2\alpha+\gamma+2\theta-2\phi\right) \\
&\left.-n\sin\left(2\alpha-\gamma-2\theta-2\phi\right)+n\sin\left(2\alpha+\gamma-2\theta-2\phi\right)\right], \qquad \text{(A.7)}
\end{aligned}
$$

$$
\begin{aligned}
Q_1^{(\pm)} &= \frac{i\sin^2\left(\theta/2\right)}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
&\times\Big[2n^2\cos\left(2\alpha-\gamma-3\phi\right)-6n^2\cos\left(2\alpha+\gamma-3\phi\right) \\
&-n^2\cos\left(2\alpha-\gamma-2\theta-3\phi\right)-n^2\cos\left(2\alpha+\gamma-2\theta-3\phi\right) \\
&-4n^2\cos\left(2\alpha+\gamma-\theta-3\phi\right)-4n^2\cos\left(2\alpha+\gamma+\theta-3\phi\right) \\
&-n^2\cos\left(2\alpha-\gamma+2\theta-3\phi\right)-n^2\cos\left(2\alpha+\gamma+2\theta-3\phi\right) \\
&-2(8k^2+10kn+3n^2)\cos\left(2\alpha-\gamma-\phi\right) \\
&+2n(2k+n)\cos\left(2\alpha+\gamma-\phi\right)-n^2\cos\left(2\alpha-\gamma-2\theta-\phi\right) \\
&-n^2\cos\left(2\alpha+\gamma-2\theta-\phi\right)-2n(k+2n)\cos\left(2\alpha-\gamma-\theta-\phi\right) \\
&+2kn\cos\left(2\alpha+\gamma-\theta-\phi\right)-2n(k+2n)\cos\left(2\alpha-\gamma+\theta-\phi\right) \\
&+2kn\cos\left(2\alpha+\gamma+\theta-\phi\right)-n^2\cos\left(2\alpha-\gamma+2\theta-\phi\right) \\
&-n^2\cos\left(2\alpha+\gamma+2\theta-\phi\right)\Big], \qquad \text{(A.8)}
\end{aligned}
$$

$$
\begin{aligned}
Q_2^{(\pm)} &= \frac{i\sin^2(\theta/2)}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
&\times \big[-2n^2\sin(2\alpha-\gamma-3\phi)+6n^2\sin(2\alpha+\gamma-3\phi) \\
&+n^2\sin(2\alpha-\gamma-2\theta-3\phi)+n^2\sin(2\alpha+\gamma-2\theta-3\phi) \\
&+4n^2\sin(2\alpha+\gamma-\theta-3\phi)+4n^2\sin(2\alpha+\gamma+\theta-3\phi) \\
&+n^2\sin(2\alpha-\gamma+2\theta-3\phi)+n^2\sin(2\alpha+\gamma+2\theta-3\phi) \\
&-2(8k^2+10kn+3n^2)\sin(2\alpha-\gamma-\phi) \\
&+2n(2k+n)\sin(2\alpha+\gamma-\phi)-n^2\sin(2\alpha-\gamma-2\theta-\phi) \\
&-n^2\sin(2\alpha+\gamma-2\theta-\phi)-2n(k+2n)\sin(2\alpha-\gamma-\theta-\phi) \\
&+2kn\sin(2\alpha+\gamma-\theta-\phi)-2n(k+2n)\sin(2\alpha-\gamma+\theta-\phi) \\
&+2kn\sin(2\alpha+\gamma+\theta-\phi)-n^2\sin(2\alpha-\gamma+2\theta-\phi) \\
&-n^2\sin(2\alpha+\gamma+2\theta-\phi)\big], \qquad \text{(A.9)}
\end{aligned}
$$

$$
\begin{aligned}
Q_3^{(\pm)} &= \frac{in\sin\theta}{8m^{(\pm)}(k\pm n-m^{(\pm)})(k\pm n+m^{(\pm)})} \\
&\times [2(3k+n)\cos(2\alpha-\gamma-2\phi)-2(k-n)\cos(2\alpha+\gamma-2\phi) \\
&-3k\cos(2\alpha-\gamma-\theta-2\phi)-k\cos(2\alpha+\gamma-\theta-2\phi) \\
&-3k\cos(2\alpha-\gamma+\theta-2\phi)-k\cos(2\alpha+\gamma+\theta-2\phi) \\
&-n\cos(2\alpha-\gamma+2\theta-2\phi)-n\cos(2\alpha+\gamma+2\theta-2\phi) \\
&-n\cos(2\alpha-\gamma-2\theta-2\phi)-n\cos(2\alpha+\gamma-2\theta-2\phi)], \qquad \text{(A.10)}
\end{aligned}
$$

with $m^{(\pm)} = (k^2+n^2\pm 2kn\cos\theta)^{1/2}$.

Appendix B

Combinatorial design theory

The problem of constructing a family of sets $\mathbf{U}_m$ under the cardinality constraints (5.31) and (5.32) in Section 5.2.3 may be regarded geometrically as the problem of constructing a collection of N distinct points (representing template labels) and M distinct lines (representing sets) with the following properties:

(i) each point lies on exactly R lines;

(ii) each line passes through exactly C points;

(iii) any two lines intersect at exactly I points;

(iv) any two points lie on at most $R - 1$ lines.

The final property is the automatic identification condition, i.e. no two labels are assigned to exactly the same subfamily of sets.

The feasibility of carrying out such a construction (or finding additional conditions on N, M and R that ensure it is possible) is a difficult and unsolved problem in combinatorics. One special case that has been studied in detail is $R = C$ and $I = 1$. This implies that $N = M = R^2 - R + 1$, and that any two points must lie on exactly one line. Under these circumstances, the four geometrical properties define a finite projective plane of order $R - 1$ [170]. It is known that finite projective planes exist with prime orders [170], but there is no finite projective plane of order 6 [198] or 10 [199], while the existence (or otherwise) of an order-12 finite projective plane remains an open question.

The special case of finite projective planes is uninteresting from a compression-scheme point of view, as it has $N = M$ and hence achieves no compression. However, it strongly indicates that the conditions (5.31) and (5.32) are not sufficient to ensure the existence of a set construction with the four required properties. Nonetheless, valid set constructions have been found for small values of N, M and R; for example, $(N, M, R) = (10, 6, 3)$ yields $C = 5$, $I = 2$, and the set construction

$$\begin{aligned}
\mathbf{U}_1 &= \{1, 2, 3, 4, 5\}\,, \\
\mathbf{U}_2 &= \{1, 2, 6, 7, 8\}\,, \\
\mathbf{U}_3 &= \{1, 3, 6, 9, 10\}\,, \\
\mathbf{U}_4 &= \{2, 5, 8, 9, 10\}\,, \\
\mathbf{U}_5 &= \{3, 4, 7, 8, 10\}\,, \\
\mathbf{U}_6 &= \{4, 5, 6, 7, 9\}\,. \qquad \text{(B.1)}
\end{aligned}$$

Additional solutions for $(N, M, R) = (12, 9, 3)$ and $(N, M, R) = (14, 7, 3)$ also exist. No counterexamples (i.e. values of (N, M, R) satisfying (5.31) and (5.32) but admitting no set construction) have been found for $N > M$, although we have not conducted an exhaustive search.

A general compression scheme satisfying the conditions (5.31) and (5.32) might potentially admit more compression rates than the symmetric base scheme for each value of N. Given the difficulties in actually constructing the sets, however, we focus instead on the special case of “maximal representation” for fixed M and R (i.e. every M-digit binary number with exactly R 1’s represents a distinct template label); this gives the binomial coefficient scheme described in Section 5.2.3.

Appendix C

Taylor-T2 waveform expansions

The Taylor-T2 PN waveform (5.52) used in Section 5.4 describes the inspiral part of a circular and non-inclined comparable-mass binary coalescence [38, 174, 175]. With units restored, its amplitude and phase are written as expansions in the frequency-related variable

$$\xi := \left(\frac{G\mathcal{M}}{c^3 \eta^{3/5}} \dot{\phi} \right)^{2/3}, \tag{C.1}$$

with the orbital phase ϕ given to 3.5PN accuracy by

$$\begin{aligned} \phi \;=\; & -\frac{1}{\eta} \Bigg\{ \tau^{5/8} + \left(\frac{3715}{8064} + \frac{55}{96}\eta \right) \tau^{3/8} - \frac{3}{4}\pi \tau^{1/4} \\ & + \left(\frac{9275495}{14450688} + \frac{284875}{258048}\eta + \frac{1855}{2048}\eta^2 \right) \tau^{1/8} \\ & + \left(-\frac{38645}{172032} + \frac{65}{2048}\eta \right) \pi \ln\left(\frac{\tau}{\tau_0} \right) \\ & + \left[\frac{831032450749357}{57682522275840} - \frac{53}{40}\pi^2 - \frac{107}{56}\gamma + \frac{107}{448} \ln\left(\frac{\tau}{256} \right) \right. \\ & \left. + \left(-\frac{126510089885}{4161798144} + \frac{2255}{2048}\pi^2 \right) \eta + \frac{154565}{1835008}\eta^2 - \frac{1179625}{1769472}\eta^3 \right] \tau^{-1/8} \\ & + \left(\frac{188516689}{173408256} + \frac{488825}{516096}\eta - \frac{141769}{516096}\eta^2 \right) \pi \tau^{-1/4} \Bigg\}, \end{aligned} \tag{C.2}$$

where γ is the Euler–Mascheroni constant. Here τ is a time-related variable, written in terms of the binary coalescence time t_c as

$$\tau = \frac{c^3\eta^{8/5}}{5G\mathcal{M}}(t_c - t), \tag{C.3}$$

where we set $t_c = 1\,\mathrm{yr}$ for a $\sim 10^6 M_\odot$ black-hole binary inspiral.

The GW amplitude is then proportional to the 2PN amplitude function

$$\mathcal{A} = \xi\left(2 + \frac{1}{3}(-13+\eta)\xi + 4\pi\xi^{3/2} \ + \frac{1}{180}(-837 - 635\eta + 15\eta^2)\xi^2\right), \tag{C.4}$$

while the GW phase is twice the tail-distorted orbital phase

$$\psi = \phi - 3\xi^{3/2}\left(1 - \frac{\eta}{2}\xi\right)\ln\left(\frac{\xi}{\xi_0}\right), \tag{C.5}$$

with the 1PN factor of $1 - (\eta/2)\xi$ included to account for the nonlinear interaction between the gravitational field of the source and its emitted gravitational radiation [200]. Finally, the constant frequency in ξ_0 is set to $\dot{\phi}_0 = 10^{-4}\pi$, which corresponds to an approximate entry frequency of $10^{-4}\,\mathrm{Hz}$ for LISA.

Bibliography

[1] A. Einstein. On gravitational waves. *Proceedings of the Prussian Academy of Sciences*, 1918:154, 1918.

[2] R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. *The Astrophysical Journal*, 195:L51, 1975.

[3] I. H. Stairs. Testing general relativity with pulsar timing. *Living Reviews in Relativity*, 6:5, 2003.

[4] A. G. Lyne et al. A double-pulsar system: A rare laboratory for relativistic gravity and plasma physics. *Science*, 303:1153, 2004.

[5] J. Weber. Detection and generation of gravitational waves. *Physical Review*, 117:306, 1960.

[6] W. H. Press and K. S. Thorne. Gravitational-wave astronomy. *Annual Review of Astronomy and Astrophysics*, 10:335, 1972.

[7] G. M. Harry (for the LIGO Scientific Collaboration). Advanced LIGO: The next generation of gravitational wave detectors. *Classical and Quantum Gravity*, 27:084006, 2010.

[8] B. P. Abbott et al. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116:061102, 2016.

[9] B. P. Abbott et al. GW151226: Observation of gravitational waves from a 22-Solar-mass binary black hole coalescence. *Physical Review Letters*, 116:241103, 2016.

[10] B. P. Abbott et al. Binary black hole mergers in the first Advanced LIGO observing run. *Physical Review X*, 6:041015, 2016.

[11] F. Acernese et al. Advanced Virgo: A second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32:024001, 2015.

[12] K. Somiya (for the KAGRA Collaboration). Detector configuration of KAGRA—the Japanese cryogenic gravitational-wave detector. *Classical and Quantum Gravity*, 29:124007, 2012.

[13] K. Danzmann et al. LISA: Laser Interferometer Space Antenna. 2017. www.elisascience.org/files/publications/LISA_L3_20170120.pdf.

[14] P. Amaro-Seoane et al. The gravitational universe. 2013. arXiv:1305.5720 [astro-ph.CO].

[15] S. Kawamura et al. The Japanese space gravitational wave antenna: DECIGO. *Classical and Quantum Gravity*, 28:094011, 2011.

[16] J. Luo et al. TianQin: A space-borne gravitational wave detector. *Classical and Quantum Gravity*, 33:035010, 2016.

[17] K. Danzmann (for the LISA Pathfinder Team and the eLISA Consortium). LISA and its pathfinder. *Nature Physics*, 11:613, 2015.

[18] M. Armano et al. Sub-femto-g free fall for space-based gravitational wave observatories: LISA Pathfinder results. *Physical Review Letters*, 116:231101, 2016.

[19] G. Hobbs et al. The International Pulsar Timing Array project: Using pulsars as a gravitational wave detector. *Classical and Quantum Gravity*, 27:084013, 2010.

[20] S. R. Taylor et al. Are we there yet? Time to detection of nanohertz gravitational waves based on pulsar-timing array limits. *The Astrophysical Journal Letters*, 819:L6, 2016.

[21] A. Sesana. Prospects for multiband gravitational-wave astronomy after GW150914. *Physical Review Letters*, 116:231102, 2016.

[22] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. W. H. Freeman, 1973.

[23] K. S. Thorne. Multipole expansions of gravitational radiation. *Reviews of Modern Physics*, 52:299, 1980.

[24] C. Cutler and K. S. Thorne. An overview of gravitational-wave sources. In N. Bishop and S. D. Maharaj, editors, *Proceedings of the 16th International Conference on General Relativity and Gravitation*. World Scientific, 2002.

[25] B. S. Sathyaprakash and B. F. Schutz. Physics, astrophysics and cosmology with gravitational waves. *Living Reviews in Relativity*, 12:2, 2009.

[26] C. J. Moore, R. H. Cole, and C. P. L. Berry. Gravitational-wave sensitivity curves. *Classical and Quantum Gravity*, 32:015014, 2015.

[27] T. Creighton. Advanced LIGO: Sources and astrophysics. *Classical and Quantum Gravity*, 20:S853, 2003.

[28] P. Amaro-Seoane et al. Low-frequency gravitational-wave science with eLISA/NGO. *Classical and Quantum Gravity*, 29:124016, 2012.

[29] C. D. Ott. The gravitational-wave signature of core-collapse supernovae. *Classical and Quantum Gravity*, 26:063001, 2009.

[30] P. D. Lasky. Gravitational waves from neutron stars: A review. *Publications of the Astronomical Society of Australia*, 32:e034, 2015.

[31] T. Damour and A. Vilenkin. Gravitational wave bursts from cosmic strings. *Physical Review Letters*, 85:3761, 2000.

[32] L. J. Rubbo, K. Holley-Bockelmann, and L. S. Finn. Event rate for extreme mass ratio burst signals in the Laser Interferometer Space Antenna band. *The Astrophysical Journal Letters*, 649:L25, 2006.

[33] P. D. Lasky et al. Gravitational-wave cosmology across 29 decades in frequency. *Physical Review X*, 6:011035, 2016.

[34] C. Caprini et al. Science with the space-based interferometer eLISA, II: Gravitational waves from cosmological phase transitions. *Journal of Cosmology and Astroparticle Physics*, 04(2016):001, 2016.

[35] P. C. Peters and J. Mathews. Gravitational radiation from point masses in a Keplerian orbit. *Physical Review*, 131:435, 1963.

[36] P. C. Peters. Gravitational radiation and the motion of two point masses. *Physical Review*, 136:B1224, 1964.

[37] L. Blanchet and T. Damour. Post-Newtonian generation of gravitational waves. *Annales de l'Institut Henri Poincaré*, 50:377, 1989.

[38] L. Blanchet. Gravitational radiation from post-Newtonian sources and inspiralling compact binaries. *Living Reviews in Relativity*, 17:2, 2014.

[39] J. Centrella, J. G. Baker, B. J. Kelly, and J. R. van Meter. Black-hole binaries, gravitational waves, and numerical relativity. *Reviews of Modern Physics*, 82:3069, 2010.

[40] E. T. Newman and R. Penrose. An approach to gravitational radiation by a method of spin coefficients. *Journal of Mathematical Physics*, 3:566, 1962.

[41] E. T. Newman and R. Penrose. Note on the Bondi–Metzner–Sachs group. *Journal of Mathematical Physics*, 7:863, 1966.

[42] F. Pretorius. Evolution of binary black-hole spacetimes. *Physical Review Letters*, 95:121101, 2005.

[43] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower. Accurate evolutions of orbiting black-hole binaries without excision. *Physical Review Letters*, 96:111101, 2006.

[44] J. G. Baker et al. Gravitational-wave extraction from an inspiralling configuration of merging black holes. *Physical Review Letters*, 96:111102, 2006.

[45] The SXS Collaboration. SXS Gravitational Waveform Database, 2016. www.black-holes.org/waveforms.

[46] K. D. Kokkotas and B. G. Schmidt. Quasi-normal modes of stars and black holes. *Living Reviews in Relativity*, 2:2, 1999.

[47] T. Regge and J. A. Wheeler. Stability of a Schwarzschild singularity. *Physical Review*, 108:1063, 1957.

[48] F. J. Zerilli. Effective potential for even-parity Regge–Wheeler gravitational perturbation equations. *Physical Review Letters*, 24:737, 1970.

[49] S. A. Teukolsky. Rotating black holes: Separable wave equations for gravitational and electromagnetic perturbations. *Physical Review Letters*, 29:1114, 1972.

[50] E. W. Leaver. An analytic representation for the quasi-normal modes of Kerr black holes. *Proceedings of the Royal Society A*, 402:285, 1985.

[51] A. Buonanno and T. Damour. Effective one-body approach to general relativistic two-body dynamics. *Physical Review D*, 59:084006, 1999.

[52] Y. Pan et al. Inspiral–merger–ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism. *Physical Review D*, 89:084006, 2014.

[53] P. Ajith et al. Inspiral–merger–ringdown waveforms for black-hole binaries with nonprecessing spins. *Physical Review Letters*, 106:241101, 2011.

[54] M. Hannam et al. Simple model of complete precessing black-hole-binary gravitational waveforms. *Physical Review Letters*, 113:151101, 2014.

[55] N. Yunes et al. Extreme mass-ratio inspirals in the effective-one-body approach: Quasicircular, equatorial orbits around a spinning black hole. *Physical Review D*, 83:044044, 2011.

[56] P. Amaro-Seoane et al. Intermediate and extreme mass-ratio inspirals: Astrophysics, science applications and detection using LISA. *Classical and Quantum Gravity*, 24:R113, 2007.

[57] L. Barack and C. Cutler. LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy. *Physical Review D*, 69:082005, 2004.

[58] S. Babak et al. "Kludge" gravitational waveforms for a test-body orbiting a Kerr black hole. *Physical Review D*, 75:024005, 2007.

[59] A. J. K. Chua and J. R. Gair. Improved analytic extreme-mass-ratio inspiral model for scoping out eLISA data analysis. *Classical and Quantum Gravity*, 32:232002, 2015.

[60] L. Barack. Gravitational self-force in extreme mass-ratio inspirals. *Classical and Quantum Gravity*, 26:213001, 2009.

[61] E. Poisson, A. Pound, and I. Vega. The motion of point particles in curved spacetime. *Living Reviews in Relativity*, 14:7, 2011.

[62] S. A. Hughes. Evolution of circular, nonequatorial orbits of Kerr black holes due to gravitational-wave emission, II: Inspiral trajectories and gravitational waveforms. *Physical Review D*, 64:064004, 2001.

[63] S. Drasco and S. A. Hughes. Gravitational wave snapshots of generic extreme mass ratio inspirals. *Physical Review D*, 73:024027, 2006.

[64] É. É. Flanagan and T. Hinderer. Transient resonances in the inspirals of point particles into black holes. *Physical Review Letters*, 109:071102, 2012.

[65] T. Hinderer and É. É. Flanagan. Two-timescale analysis of extreme mass ratio inspirals in Kerr spacetime: Orbital motion. *Physical Review D*, 78:064028, 2008.

[66] E. A. Huerta and J. R. Gair. Influence of conservative corrections on parameter estimation for extreme-mass-ratio inspirals. *Physical Review D*, 79:084021, 2009.

[67] J. R. Gair et al. Forced motion near black holes. *Physical Review D*, 83:044037, 2011.

[68] N. Warburton et al. Evolution of inspiral orbits around a Schwarzschild black hole. *Physical Review D*, 85:061501(R), 2012.

[69] A. Ori and K. S. Thorne. Transition from inspiral to plunge for a compact body in a circular equatorial orbit around a massive, spinning black hole. *Physical Review D*, 62:124022, 2000.

[70] C. O. Lousto, H. Nakano, Y. Zlochower, and M. Campanelli. Intermediate-mass-ratio black-hole binaries: Numerical relativity meets perturbation theory. *Physical Review Letters*, 104:211101, 2010.

[71] E. A. Huerta and J. R. Gair. Intermediate-mass-ratio inspirals in the Einstein Telescope, I: Signal-to-noise ratio calculations. *Physical Review D*, 83:044020, 2011.

[72] P. Jaranowski and A. Królak. *Analysis of gravitational-wave data*. Cambridge University Press, 2009.

[73] P. Jaranowski and A. Królak. Gravitational-wave data analysis: Formalism and sample applications. *Living Reviews in Relativity*, 15:4, 2012.

[74] M. Punturo et al. The Einstein Telescope: A third-generation gravitational wave observatory. *Classical and Quantum Gravity*, 27:194002, 2010.

[75] J. Crowder and N. J. Cornish. Beyond LISA: Exploring future gravitational wave missions. *Physical Review D*, 72:083005, 2005.

[76] T. J. W. Lazio. The Square Kilometre Array pulsar timing array. *Classical and Quantum Gravity*, 30:224011, 2013.

[77] M. Maggiore. *Gravitational waves, Volume 1: Theory and experiments*. Oxford University Press, 2007.

[78] O. D. Aguiar. Past, present and future of the resonant-mass gravitational wave detectors. *Research in Astronomy and Astrophysics*, 11:1, 2010.

[79] C. Cutler. Angular resolution of the LISA gravitational wave detector. *Physical Review D*, 57:7089, 1998.

[80] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne. Spin-induced orbital precession and its modulation of the gravitational waveforms from merging binaries. *Physical Review D*, 49:6274, 1994.

[81] T. A. Apostolatos. Search templates for gravitational waves from precessing, inspiraling binaries. *Physical Review D*, 52:605, 1995.

[82] B. F. Schutz. Data processing, analysis, and storage for interferometric antennas. In D. G. Blair, editor, *The detection of gravitational waves*. Cambridge University Press, 1989.

[83] B. J. Owen. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Physical Review D*, 53:6749, 1996.

[84] D. George and E. A. Huerta. Deep neural networks to enable real-time multimessenger astrophysics. 2017. arXiv:1701.00008[astro-ph.IM].

[85] P. Addesso et al. Compressed sensing for time-frequency gravitational wave data analysis. 2016. arXiv:1605.03496[astro-ph.IM].

[86] L. S. Finn. Detection, measurement, and gravitational radiation. *Physical Review D*, 46:5236, 1992.

[87] C. Cutler and É. É. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform? *Physical Review D*, 49:2658, 1994.

[88] N. Christensen and R. Meyer. Using Markov chain Monte Carlo methods for estimating parameters with gravitational radiation data. *Physical Review D*, 64:022001, 2001.

[89] C. Röver, R. Meyer, and N. Christensen. Coherent Bayesian inference on compact binary inspirals using a network of interferometric gravitational wave detectors. *Physical Review D*, 75:062004, 2007.

[90] F. Feroz, J. R. Gair, M. P. Hobson, and E. K. Porter. Use of the MULTINEST algorithm for gravitational wave data analysis. *Classical and Quantum Gravity*, 26:215003, 2009.

[91] J. Crowder, N. J. Cornish, and J. L. Reddinger. LISA data analysis using genetic algorithms. *Physical Review D*, 73:063011, 2006.

[92] A. J. K. Chua, P. Cañizares, and J. R. Gair. Electromagnetic signatures of far-field gravitational radiation in the 1+3 approach. *Classical and Quantum Gravity*, 32:015011, 2015.

[93] E. S. Phinney. Finding and using electromagnetic counterparts of gravitational wave sources: A white paper for the Astro2010 Decadal Review. 2009. arXiv:0903.0098[astro-ph.CO].

[94] I. Mandel and R. O'Shaughnessy. Compact binary coalescences in the band of ground-based gravitational-wave detectors. *Classical and Quantum Gravity*, 27:114007, 2010.

[95] S. Ghosh and G. Nelemans. Localizing gravitational wave sources with optical telescopes and combining electromagnetic and gravitational wave data. In C. F. Sopuerta, editor, *Gravitational wave astrophysics*. Springer, 2015.

[96] H. Sotani, K. D. Kokkotas, P. Laguna, and C. F. Sopuerta. Gravitationally driven electromagnetic perturbations of neutron stars and black holes. *Physical Review D*, 87:084018, 2013.

[97] P. Pani, E. Berti, and L. Gualtieri. Scalar, electromagnetic, and gravitational perturbations of Kerr-Newman black holes in the slow-rotation limit. *Physical Review D*, 88:064048, 2013.

[98] C. Palenzuela et al. Binary black holes' effects on electromagnetic fields. *Physical Review Letters*, 103:081101, 2009.

[99] M. E. Gertsenshteĭn. Wave resonance of light and gravitational waves. *Soviet Physics JETP*, 14:84, 1962.

[100] Y. B. Zel'dovich. Electromagnetic and gravitational waves in a stationary magnetic field. *Soviet Physics JETP*, 38:652, 1974.

[101] C. Barrabès and P. A. Hogan. Interaction of gravitational waves with magnetic and electric fields. *Physical Review D*, 81:064024, 2010.

[102] F. I. Cooperstock. The interaction between electromagnetic and gravitational waves. *Annals of Physics*, 47:173, 1968.

[103] D. M. Zipoy. Light fluctuations due to an intergalactic flux of gravitational waves. *Physical Review*, 142:825, 1966.

[104] R. Fakir. Gravitational wave detection: A nonmechanical effect. *The Astrophysical Journal*, 418:202, 1993.

[105] S. Kopeikin, P. Korobkov, and A. Polnarev. Propagation of light in the field of stationary and radiative gravitational multipoles. *Classical and Quantum Gravity*, 23:4299, 2006.

[106] A. M. Cruise. An interaction between gravitational and electromagnetic waves. *Monthly Notices of the Royal Astronomical Society*, 204:485, 1983.

[107] E. Montanari. On the propagation of electromagnetic radiation in the field of a plane gravitational wave. *Classical and Quantum Gravity*, 15:2493, 1998.

[108] A. R. Prasanna and S. Mohanty. Gravitational wave-induced rotation of the plane of polarization of pulsar signals. *Europhysics Letters*, 60:651, 2002.

[109] M. Halilsoy and O. Gurtug. Search for gravitational waves through the electromagnetic Faraday rotation. *Physical Review D*, 75:124021, 2007.

[110] V. Faraoni. The rotation of polarization by gravitational waves. *New Astronomy*, 13:178, 2008.

[111] R. Ragazzoni, G. Valente, and E. Marchetti. Gravitational wave detection through microlensing? *Monthly Notices of the Royal Astronomical Society*, 345:100, 2003.

[112] G. B. Lesovik, A. V. Lebedev, V. Mounutcharyan, and T. Martin. Detection of gravity waves by phase modulation of the light from a distant star. *Physical Review D*, 71:122001, 2005.

[113] G. F. R. Ellis. Relativistic cosmology. In R. K. Sachs, editor, *Proceedings of the International School of Physics "Enrico Fermi", Course 47, General relativity and cosmology*. Academic Press, 1971.

[114] G. F. R. Ellis and H. van Elst. Cosmological models: Cargèse Lectures 1998. In M. Lachièze-Rey, editor, *Proceedings of the NATO Advanced Study Institute on Theoretical and Observational Cosmology*. Kluwer Academic Publishers, 1999.

[115] C. G. Tsagas, A. Challinor, and R. Maartens. Relativistic cosmology and large-scale structure. *Physics Reports*, 465:61, 2008.

[116] P. K. S. Dunsby, B. A. C. C. Bassett, and G. F. R. Ellis. Covariant analysis of gravitational waves in a cosmological context. *Classical and Quantum Gravity*, 14:1215, 1997.

[117] M. Marklund, P. K. S. Dunsby, and G. Brodin. Cosmological electromagnetic fields due to gravitational wave perturbations. *Physical Review D*, 62:101501(R), 2000.

[118] C. G. Tsagas, P. K. S. Dunsby, and M. Marklund. Gravitational wave amplification of seed magnetic fields. *Physics Letters B*, 561:17, 2003.

[119] C. G. Tsagas. Gravitomagnetic amplification in cosmology. *Physical Review D*, 81:043501, 2010.

[120] C. G. Tsagas. Gravitoelectromagnetic resonances. *Physical Review D*, 84:043524, 2011.

[121] C. G. Tsagas. Electromagnetic fields in curved spacetimes. *Classical and Quantum Gravity*, 22:393, 2005.

[122] C. A. Clarkson and R. K. Barrett. Covariant perturbations of Schwarzschild black holes. *Classical and Quantum Gravity*, 20:3855, 2003.

[123] C. A. Clarkson, M. Marklund, G. Betschart, and P. K. S. Dunsby. The electromagnetic signature of black hole ring-down. *The Astrophysical Journal*, 613:492, 2004.

[124] M. Marklund, G. Brodin, and P. K. S. Dunsby. Radio wave emissions due to gravitational radiation. *The Astrophysical Journal*, 536:875, 2000.

[125] M. Servin and G. Brodin. Resonant interaction between gravitational waves, electromagnetic waves, and plasma flows. *Physical Review D*, 68:044017, 2003.

[126] A. P. Kouretsis and C. G. Tsagas. Gravito-electromagnetic resonances in Minkowski space. *Physical Review D*, 88:044006, 2013.

[127] J. D. Jackson. *Classical electrodynamics*. Wiley, 1998.

[128] F. J. Dyson. Is a graviton detectable? Henri Poincaré Prize Lecture 2012. 2013. publications.ias.edu/sites/default/files/poincare2012.pdf.

[129] F-Y. Li, H. Wen, and Z-Y. Fang. High-frequency gravitational waves having large spectral densities and their electromagnetic response. *Chinese Physics B*, 22:120402, 2013.

[130] S. A. Olausen and V. M. Kaspi. The McGill magnetar catalog. *The Astrophysical Journal Supplement Series*, 212:6, 2014.

[131] F. Pacini. Rotating neutron stars, pulsars and supernova remnants. *Nature*, 219:145, 1968.

[132] H. D. Falcke et al. A very brief description of LOFAR: The Low Frequency Array. *Highlights of Astronomy*, 14:386, 2006.

[133] A. R. Taylor. The Square Kilometre Array. *Proceedings of the International Astronomical Union Symposia and Colloquia*, 291:339, 2012.

[134] A. J. K. Chua. Augmented kludge waveforms and Gaussian process regression for EMRI data analysis. *Journal of Physics: Conference Series*, 716:012028, 2016.

[135] A. J. K. Chua and J. R. Gair. EMRI Kludge Suite, 2016. github.com/alvincjk/EMRI_Kludge_Suite.

[136] J. R. Gair, M. Vallisneri, S. L. Larson, and J. G. Baker. Testing general relativity with low-frequency, space-based gravitational-wave detectors. *Living Reviews in Relativity*, 16:7, 2013.

[137] K. A. Arnaud et al. An overview of the Mock LISA Data Challenges. *AIP Conference Proceedings*, 873:619, 2006.

[138] W. Schmidt. Celestial mechanics in Kerr spacetime. *Classical and Quantum Gravity*, 19:2743, 2002.

[139] M. Sasaki and H. Tagoshi. Analytic black hole perturbation approach to gravitational radiation. *Living Reviews in Relativity*, 6:6, 2003.

[140] E. Forseth, C. R. Evans, and S. Hopper. Eccentric-orbit extreme-mass-ratio inspiral gravitational wave energy fluxes to 7PN order. *Physical Review D*, 93:064058, 2016.

[141] N. Sago and R. Fujita. Calculation of radiation reaction effect on orbital parameters in Kerr spacetime. *Progress of Theoretical and Experimental Physics*, 2015:073E03, 2015.

[142] A. Buonanno, Y. Chen, and M. Vallisneri. Detection template families for gravitational waves from the final stages of binary-black-hole inspirals: Nonspinning case. *Physical Review D*, 67:024016, 2003.

[143] B. M. Barker and R. F. O'Connell. Gravitational two-body problem with arbitrary masses, spins, and quadrupole moments. *Physical Review D*, 12:329, 1975.

[144] V. A. Brumberg. *Essential relativistic celestial mechanics*. CRC Press, 1991.

[145] W. Junker and G. Schäfer. Binary systems: Higher order gravitational radiation damping and wave emission. *Monthly Notices of the Royal Astronomical Society*, 254:146, 1992.

[146] F. D. Ryan. Effect of gravitational radiation reaction on nonequatorial orbits around a Kerr black hole. *Physical Review D*, 53:3064, 1996.

[147] S. A. Hughes. Evolution of circular, nonequatorial orbits of Kerr black holes due to gravitational-wave emission. *Physical Review D*, 61:084004, 2000.

[148] K. Glampedakis, S. A. Hughes, and D. Kennefick. Approximating the inspiral of test bodies into Kerr black holes. *Physical Review D*, 66:064005, 2002.

[149] J. R. Gair and K. Glampedakis. Improved approximate inspirals of test bodies into Kerr black holes. *Physical Review D*, 73:064037, 2006.

[150] J. D. Bekenstein. Gravitational-radiation recoil and runaway black holes. *The Astrophysical Journal*, 183:657, 1973.

[151] W. H. Press. Gravitational radiation from sources which extend into their own wave zone. *Physical Review D*, 15:965, 1977.

[152] S. Chandrasekhar. *The mathematical theory of black holes*. Clarendon Press, 1983.

[153] Y. Mino. Perturbative approach to an orbital evolution around a supermassive black hole. *Physical Review D*, 67:084027, 2003.

[154] S. Drasco and S. A. Hughes. Rotating black hole orbit functionals in the frequency domain. *Physical Review D*, 69:044015, 2004.

[155] N. Yunes and E. Berti. Accuracy of the post-Newtonian approximation: Optimal asymptotic expansion for quasicircular, extreme-mass ratio inspirals. *Physical Review D*, 77:124006, 2008.

[156] E. L. Rees. Graphical discussion of the roots of a quartic equation. *American Mathematical Monthly*, 29:51, 1922.

[157] D. J. A. McKechan, C. Robinson, and B. S. Sathyaprakash. A tapering window for time-domain templates and simulated signals in the detection of gravitational waves from coalescing compact binaries. *Classical and Quantum Gravity*, 27:084020, 2010.

[158] M. Capderou. *Satellites: Orbits and missions*. Springer, 2005.

[159] A. Klein et al. Science with the space-based interferometer eLISA: Supermassive black hole binaries. *Physical Review D*, 93:024003, 2016.

[160] A. J. K. Chua and J. R. Gair. Tunable compression of template banks for fast gravitational-wave detection and localization. *Physical Review D*, 93:122001, 2016.

[161] S. Mitra, S. V. Dhurandhar, and L. S. Finn. Improving the efficiency of the detection of gravitational wave signals from inspiraling compact binaries: Chebyshev interpolation. *Physical Review D*, 72:102001, 2005.

[162] R. J. E. Smith et al. Towards rapid parameter estimation on gravitational waves from compact binaries using interpolated waveforms. *Physical Review D*, 87:122002, 2013.

[163] P. Cañizares, S. E. Field, J. R. Gair, and M. Tiglio. Gravitational wave parameter estimation with compressed likelihood evaluations. *Physical Review D*, 87:124005, 2013.

[164] I. S. Heng. Rotating stellar core-collapse waveform decomposition: A principal component analysis approach. *Classical and Quantum Gravity*, 26:105005, 2009.

[165] K. Cannon et al. Singular value decomposition applied to compact binary coalescence gravitational-wave signals. *Physical Review D*, 82:044025, 2010.

[166] S. E. Field et al. Reduced basis catalogs for gravitational wave templates. *Physical Review Letters*, 106:221102, 2011.

[167] Y. Wang. Fast detection and automatic parameter estimation of a gravitational wave signal with a novel method. *General Relativity and Gravitation*, 47:142, 2015.

[168] J. R. Gair et al. Event rate estimates for LISA extreme mass ratio capture sources. *Classical and Quantum Gravity*, 21:S1595, 2004.

[169] S. V. Dhurandhar and B. S. Sathyaprakash. Choice of filters for the detection of gravitational waves from coalescing binaries, II: Detection in colored noise. *Physical Review D*, 49:1707, 1994.

[170] N. L. Biggs. *Discrete mathematics*. Oxford University Press, 2002.

[171] D. Singmaster. Repeated binomial coefficients and Fibonacci numbers. *Fibonacci Quarterly*, 13:295, 1975.

[172] D. Singmaster. How often does an integer occur as a binomial coefficient? *American Mathematical Monthly*, 78:385, 1971.

[173] B. S. Sathyaprakash and S. V. Dhurandhar. Choice of filters for the detection of gravitational waves from coalescing binaries. *Physical Review D*, 44:3819, 1991.

[174] L. Blanchet, B. R. Iyer, C. M. Will, and A. G. Wiseman. Gravitational waveforms from inspiralling compact binaries to second-post-Newtonian order. *Classical and Quantum Gravity*, 13:575, 1996.

[175] R. H. Cole and J. R. Gair. Likelihood smoothing using gravitational wave surrogate models. *Physical Review D*, 90:124043, 2014.

[176] P. Amaro-Seoane et al. eLISA/NGO: Astrophysics and cosmology in the gravitational-wave millihertz regime. *GW Notes*, 6:4, 2013.

[177] P. R. Brady, T. Creighton, C. Cutler, and B. F. Schutz. Searching for periodic sources with LIGO. *Physical Review D*, 57:2101, 1998.

[178] B. J. Owen and B. S. Sathyaprakash. Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement. *Physical Review D*, 60:022002, 1999.

[179] R. Prix. Template-based searches for gravitational waves: Efficient lattice covering of flat parameter spaces. *Classical and Quantum Gravity*, 24:S481, 2007.

[180] C. J. Moore, C. P. L. Berry, A. J. K. Chua, and J. R. Gair. Improving gravitational-wave parameter estimation using Gaussian process regression. *Physical Review D*, 93:064001, 2016.

[181] C. J. Moore, A. J. K. Chua, C. P. L. Berry, and J. R. Gair. Fast methods for training Gaussian processes on large datasets. *Royal Society Open Science*, 3:160125, 2016.

[182] C. Cutler and M. Vallisneri. LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms. *Physical Review D*, 76:104018, 2007.

[183] D. J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[184] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[185] C. J. Moore and J. R. Gair. Novel method for incorporating model uncertainties into gravitational wave parameter estimates. *Physical Review Letters*, 113:251101, 2014.

[186] LIGO Scientific Collaboration. LSC Algorithm Library Suite, 2016. wiki.ligo.org/DASWG/LALSuite.

[187] L. Santamaría et al. Matching post-Newtonian and numerical relativity waveforms: Systematic errors and a new phenomenological model for nonprecessing black hole binaries. *Physical Review D*, 82:064016, 2010.

[188] T. Damour, B. R. Iyer, and B. S. Sathyaprakash. Comparison of search templates for gravitational waves from binary inspiral. *Physical Review D*, 63:044023, 2001.

[189] S. Drasco. Strategies for observing extreme mass ratio inspirals. *Classical and Quantum Gravity*, 23:S769, 2006.

[190] H. Wendland. *Scattered data approximation*. Cambridge University Press, 2004.

[191] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, 2003.

[192] S. Babak. Building a stochastic template bank for detecting massive black hole binaries. *Classical and Quantum Gravity*, 25:195011, 2008.

[193] C. Messenger, R. Prix, and M. A. Papa. Random template banks and relaxed lattice coverings. *Physical Review D*, 79:104017, 2009.

[194] I. W. Harry, B. Allen, and B. S. Sathyaprakash. Stochastic template placement algorithm for gravitational wave data analysis. *Physical Review D*, 80:104014, 2009.

[195] U. Ruangsri and S. A. Hughes. Census of transient orbital resonances encountered during binary inspiral. *Physical Review D*, 89:084036, 2014.

[196] C. P. L. Berry, R. H. Cole, P. Cañizares, and J. R. Gair. Importance of transient resonances in extreme-mass-ratio inspirals. *Physical Review D*, 94:124042, 2016.

[197] M. van de Meent. Gravitational self-force on eccentric equatorial orbits around a Kerr black hole. *Physical Review D*, 94:044034, 2016.

[198] R. C. Bose. On the application of the properties of Galois fields to the problem of construction of hyper-Graeco-Latin squares. *Sankhyā: The Indian Journal of Statistics*, 3:323, 1938.

[199] C. W. H. Lam, L. Thiel, and S. Swiercz. The non-existence of finite projective planes of order 10. *Canadian Journal of Mathematics*, 41:1117, 1989.

[200] L. Blanchet and G. Schäfer. Gravitational wave tails and binary star systems. *Classical and Quantum Gravity*, 10:2699, 1993.