CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Apollo

MNRAS **478,** 28–40 (2018)
Advance Access publication 2018 April 28
doi:10.1093/mnras/sty1079

Towards a framework for testing general relativity with extreme-mass-ratio-inspiral observations

A. J. K. Chua,[1,2]★ S. Hee,[3,4] W. J. Handley,[3,4,5] E. Higson,[3,4] C. J. Moore,[6,7] J. R. Gair,[8] M. P. Hobson[3] and A. N. Lasenby[3,4]

[1]*Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA*
[2]*Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK*
[3]*Astrophysics Group, Battcock Centre, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK*
[4]*Kavli Institute for Cosmology Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[5]*Gonville and Caius College, Trinity Street, Cambridge CB2 1TA, UK*
[6]*Centro de Astrofísica e Gravitação – CENTRA, Departamento de Física, Instituto Superior Técnico – IST, Universidade de Lisboa – UL, Av. Rovisco Pais 1, P-1049-001 Lisboa, Portugal*
[7]*Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK*
[8]*School of Mathematics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, UK*

Accepted 2018 April 21. Received 2018 April 19; in original form 2018 March 26

ABSTRACT

Extreme-mass-ratio-inspiral observations from future space-based gravitational-wave detectors such as LISA will enable strong-field tests of general relativity with unprecedented precision, but at prohibitive computational cost if existing statistical techniques are used. In one such test that is currently employed for LIGO black hole binary mergers, generic deviations from relativity are represented by N deformation parameters in a generalized waveform model; the Bayesian evidence for each of its 2^N combinatorial submodels is then combined into a posterior odds ratio for modified gravity over relativity in a null-hypothesis test. We adapt and apply this test to a generalized model for extreme-mass-ratio inspirals constructed on deformed black hole spacetimes, and focus our investigation on how computational efficiency can be increased through an evidence-free method of model selection. This method is akin to the algorithm known as product-space Markov chain Monte Carlo, but uses nested sampling and improved error estimates from a rethreading technique. We perform benchmarking and robustness checks for the method, and find order-of-magnitude computational gains over regular nested sampling in the case of synthetic data generated from the null model.

Key words: gravitational waves – methods: data analysis – methods: statistical.

1 INTRODUCTION

The spate of gravitational-wave (GW) sources found by Advanced LIGO in its first two observing runs (Abbott et al. 2016a, 2017a,b,c,d) has opened up a new branch of multimessenger astronomy – one that extends its reach beyond electromagnetic radiation for the first time, and into the gravitational sector. Traditional electromagnetic telescopes will work together with a ground-based GW detector network (Dooley et al. 2015), pulsar timing arrays (Lommen 2015), and future space-based GW detectors (Amaro-Seoane et al. 2017) to discover and study a broad spectrum of sources that are astrophysical or cosmological in origin.

Astronomy with GW observations will also improve our understanding of gravitation and fundamental physics, by granting access to unprecedented tests of general relativity (GR) and its alternatives in the dynamical strong-field regime (Gair et al. 2013; Yunes & Siemens 2013). Several such tests have been performed on data from Advanced LIGO's first observing run, with no evidence for deviation from GR to date (Abbott et al. 2016a,b; Yunes, Yagi & Pretorius 2016; Abbott et al. 2018a,b); these include various GR consistency checks for measured signals, as well as the placing of constraints on non-tensorial GW polarizations in generic metric theories, and on the graviton Compton wavelength in massive-gravity theories.

One particular test focuses on the late-stage phase evolution of GW signals from black hole binary mergers, which encode several strong-field effects that are not observable for binary pulsars. This approach is a particular implementation of the parametrized post-Einsteinian framework (Yunes & Pretorius 2009), in which model-independent deviations from GR are parametrized as deformations to the post-Newtonian (PN) and phenomenological phase param-

★ E-mail: alvin.j.chua@jpl.nasa.gov

© 2018 The Author(s)
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

eters in a GR-based inspiral–merger–ringdown waveform model (Khan et al. 2016). Any phase deformations for a given signal are then constrained through comparison with the generalized model, and a Bayesian model selection framework (Li et al. 2012; Agathos et al. 2014) is used to perform a null-hypothesis test of GR.

Merging binary systems have provided the only GW signals detected so far, and are expected to be ubiquitous across the bandwidths of current and future interferometric detectors. They will be an important type of source for ESA's space-based detector LISA (Amaro-Seoane et al. 2017), in the form of massive-black hole (MBH) binary mergers and the extreme-mass-ratio inspirals (EMRIs) of stellar-origin compact objects into MBHs within galactic nuclei (Amaro-Seoane et al. 2012). EMRIs in particular have the potential to facilitate stringent tests of GR (Barack & Cutler 2007; Gair et al. 2013); this is because the phase of the GW signal can be precisely tracked over the large number of observable orbits spent by the compact object in the strong field of the central black hole.

Model-independent parametric tests of GR with EMRIs may be performed using an EMRI analogue of the generalized waveform model for comparable-mass black hole binary mergers. One such model (Gair & Yunes 2011) introduces deformations at the level of the background metric around the central MBH, by constructing analytic EMRI waveforms (Barack & Cutler 2004) on a bumpy black hole spacetime that retains an approximately conserved energy, angular momentum, and Carter constant along each geodesic orbit (Vigeland, Yunes & Stein 2011). This model shows the potential of EMRIs for testing the Kerr solution in GR, as it can place much tighter constraints on the metric deformations than current X-ray observations (Moore & Gair 2015; Moore, Chua & Gair 2017).

However, the prospect of doing precision science with EMRIs is accompanied by a high degree of technical difficulty. The length and complexity of EMRI signals lead to computationally expensive models and a highly multimodal likelihood surface in Bayesian inference problems, which exacerbates the already significant challenge of evaluating the evidence for model selection. Modern algorithms such as nested sampling can explore comparable-mass merger likelihoods efficiently (Feroz et al. 2009a; Veitch & Vecchio 2010) and are used to compute Bayes factors in the LIGO framework (Abbott et al. 2016b), but improved techniques are required to adapt these algorithms for tests of GR with EMRIs.

In this paper, we take a first step towards developing a framework for testing GR with EMRI observations. The generalized EMRI waveform model described above is trialled in a Bayesian model selection framework based on the LIGO tests, using a nested-sampling algorithm that is tailored for high-dimensional and multimodal likelihoods (Handley, Hobson & Lasenby 2015a,b). We assess the viability of a product-space method (Hee et al. 2015) in accelerating the convergence of nested sampling on EMRI likelihoods, and apply a rethreading technique (Higson et al. 2017) that provides error estimates on the Bayes factors obtained through this method.

As the generalized EMRI model is too computationally unwieldy for method development purposes, we first present a proof of principle on a toy waveform model that mirrors several of its key qualitative features. The rethreading technique is empirically validated, and is used to compare the errors attained by the product-space method and regular nested sampling, as a function of total likelihood evaluations. We find that for the toy model likelihood, the product-space method reduces by an order of magnitude the number of likelihood calls taken to reach a standard deviation of 5 per cent on the final GR/non-GR Bayes factor.

A similar but scaled-down analysis is then performed for the generalized model, with sampling restricted to a subset of EMRI parameters. Due to the greater complexity of the likelihood in this case, sampling bias is more likely to occur in shorter nested-sampling runs; we briefly demonstrate how a population of such runs may be combined to reduce this bias and ensure faster convergence in practice. Results for this more realistic EMRI likelihood show that the product-space method still attains better precision than regular nested sampling at the same level of computational cost, and indicate an even greater improvement in terms of cost to reach 5 per cent error (potentially reducing the required number of likelihood calls by two orders of magnitude).

In Section 2, we give an overview of the generalized EMRI waveform model and define a toy surrogate that shares some of its relevant features. The statistical framework for testing GR with this model is set out in Section 3; we briefly summarize the key components of the LIGO test infrastructure, and describe the product-space method and rethreading technique that are used to adapt it for EMRI tests in this work. We then investigate in Section 4 the viability of our proposed framework through a mock test of GR with the toy model, before extending our analysis to the generalized model in Section 5. Throughout this paper, we adopt geometrized units such that $c = G = 1$. Greek (spacetime) indices run from 0 to 3, while Latin (space) indices run from 1 to 3. The base-10 logarithm is denoted by lg, while the natural logarithm is denoted by ln.

2 WAVEFORM MODELS

The generalized model considered in this work is based on the analytic kludge (AK) formalism of Barack & Cutler (2004), which we summarize in Section 2.1. This GR-based EMRI model is the most computationally efficient one available, and as such has been widely used in scoping out data analysis for space-based GW detectors; it is qualitatively resemblant to more accurate EMRI models, and its quantitative fidelity may also be improved through frequency corrections (Chua & Gair 2015; Chua, Moore & Gair 2017). In Section 2.2, we describe the construction by Gair & Yunes (2011) of AK waveforms on a family of generic modified-gravity black hole spacetimes. These spacetimes are parametrized by metric deformations (or 'bumps') of different sizes, which show up in the resultant 'bumpy AK' (bAK) model as perturbations to the phase evolution at different orders. The sinusoidal toy model we introduce in Section 2.3 is designed to mimic the parameter dependence of the bAK phase evolution, at a fraction of the computational cost.

2.1 Analytic kludge model

In the AK model, the instantaneous orbit of the compact object around the central black hole is approximated as Newtonian; the orbital parameters of this Keplerian ellipse are evolved over time with mixed-order PN expressions to simulate relativistic effects such as radiation reaction (inspiralling and circularizing) and orbital precession (apsidal and Lense–Thirring). The waveform is then generated along the orbital trajectory using the Peters & Mathews (1963) mode-sum approximation.

The inertia tensor for an EMRI with masses $(\mu, M \gg \mu)$ is given by $I^{ij}(t) = \mu x^i(t) x^j(t)$, where $\boldsymbol{x}$ is the position of the compact object relative to the central black hole. For an instantaneous Newtonian orbit with orbital frequency ν, the second time derivative of I^{ij} may be decomposed into n-harmonics of ν as $\ddot{I}^{ij} = \sum_n \ddot{I}^{ij}_n$. The three independent components of $\ddot{I}^{ij}_n$ are given by

$$\ddot{I}^{11}_n = a_n + c_n, \quad \ddot{I}^{12}_n = b_n, \quad \ddot{I}^{22}_n = c_n - a_n, \tag{1}$$

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

with (Peters & Mathews 1963)

$$a_n = -n\mathcal{A}\Big[J_{n-2}(ne) - 2eJ_{n-1}(ne) + \frac{2}{n}J_n(ne) + 2eJ_{n+1}(ne) - J_{n+2}(ne)\Big]\cos(n\Phi), \tag{2}$$

$$b_n = -n\mathcal{A}\left(1-e^2\right)^{1/2}[J_{n-2}(ne) - 2J_n(ne) + J_{n+2}(ne)] \times \sin(n\Phi), \tag{3}$$

$$c_n = 2\mathcal{A}J_n(ne)\cos(n\Phi), \tag{4}$$

$$\mathcal{A} = \mu\tilde{v}^{2/3}, \tag{5}$$

where the J_n are Bessel functions of the first kind, e is the orbital eccentricity, $\Phi(t)$ is the mean anomaly (i.e. $\dot{\Phi} = 2\pi\nu$), and $\tilde{v} := 2\pi\nu M$ is the dimensionless orbital angular frequency.

At the detector location, it is convenient to work in an orthonormal coordinate frame $\{\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{r}}\}$ with

$$\hat{\boldsymbol{p}} = \frac{\hat{\boldsymbol{r}} \times \hat{\boldsymbol{L}}}{|\hat{\boldsymbol{r}} \times \hat{\boldsymbol{L}}|}, \quad \hat{\boldsymbol{q}} = \hat{\boldsymbol{p}} \times \hat{\boldsymbol{r}}, \tag{6}$$

where $\hat{\boldsymbol{r}}$ points from detector to source and $\boldsymbol{L}$ is the orbital angular momentum of the binary. In the transverse–traceless gauge, the leading-order gravitational radiation at the detector due to a source at luminosity distance D is given by (Misner, Thorne & Wheeler 1973)

$$h_{ij} = \frac{2}{D}\left(P_{ik}P_{jl} - \frac{1}{2}P_{ij}P_{kl}\right)\ddot{I}^{kl}, \quad h^{+,\times} = \frac{1}{2}h^{ij}H^{+,\times}_{ij}, \tag{7}$$

with the polarization and projection tensors

$$H^{+}_{ij} = \hat{p}_i\hat{p}_j - \hat{q}_i\hat{q}_j, \quad H^{\times}_{ij} = \hat{p}_i\hat{q}_j + \hat{q}_i\hat{p}_j, \quad P_{ij} = \delta_{ij} - \hat{r}_i\hat{r}_j, \tag{8}$$

where δ_{ij} is the Kronecker delta.

From (1)–(8), the two GW polarization amplitudes $h^{+,\times}$ for an EMRI may be written in terms of the Peters–Mathews harmonic decomposition as (Barack & Cutler 2004)

$$h^{+,\times} = \frac{1}{D}\sum_n A^{+,\times}_n, \tag{9}$$

$$A^{+}_n = C^{+}_a a_n + C^{+}_b b_n + C^{+}_c c_n, \tag{10}$$

$$A^{\times}_n = C^{\times}_a a_n + C^{\times}_b b_n, \tag{11}$$

where the coefficients $C^{+,\times}_{a,b,c}$ depend on the angular configuration $(\lambda, \tilde{\gamma}, \alpha, \theta_K, \phi_K, \theta_S, \phi_S)$ of the EMRI: the inclination λ between $\boldsymbol{L}$ and the black hole spin $\mathbf{S}$; the azimuth $\tilde{\gamma}$ of periapsis in the orbital plane (relative to $\boldsymbol{L} \times \boldsymbol{S}$); the azimuth α of $\boldsymbol{L}$ projected on to the spin-equatorial plane; the orientation (θ_K, ϕ_K) of $\boldsymbol{S}$; and the sky location (θ_S, ϕ_S). The first two angles are intrinsic to the source, while the rest are defined relative to a fixed ecliptic-based coordinate system (Cutler 1998). Explicit expressions for the coefficients are given by equations (10) and (18)–(25) in Barack & Cutler (2004).

In a relativistic EMRI, the frequency ν, eccentricity e, and inclination λ change over time due to radiation reaction, while the angles $(\tilde{\gamma}, \alpha)$ precess as well. The AK model evolves (ν, e) with 3.5PN expressions, while approximating λ as constant due to its slow evolution over the full inspiral (Hughes 2000). The fluxes $(\dot{\nu}, \dot{e})$ are given by (Junker & Schaefer 1992)

$$\dot{\nu} = \frac{96}{10\pi}\frac{\eta\tilde{v}^{11/3}}{M^2(1-e^2)^{7/2}}\left(1 + \frac{73e^2}{24} + \frac{37e^4}{96}\right) + \mathcal{O}(\tilde{v}^{13/3}), \tag{12}$$

$$\dot{e} = -\frac{304}{15}\frac{\eta\tilde{v}^{8/3}e}{M(1-e^2)^{5/2}}\left(1 + \frac{121e^2}{304}\right) + \mathcal{O}(\tilde{v}^{10/3}), \tag{13}$$

where $\eta := \mu/M$ is the mass ratio, and only the leading 2.5PN terms are explicitly presented here. [See equations 28 and 30 in Barack & Cutler (2004) for the full expressions.]

The angular rates $(\dot{\tilde{\gamma}}, \dot{\alpha})$ determine the precession of periapsis in the orbital plane, $\dot{\tilde{\gamma}} + \dot{\alpha}$, and the Lense–Thirring precession of the orbital plane around the black hole spin axis, $\dot{\alpha}$. They are given by (Junker & Schaefer 1992; Barker & O'Connell 1975)

$$\dot{\tilde{\gamma}} = 3\frac{\tilde{v}^{5/3}}{M(1-e^2)} + \mathcal{O}(\tilde{v}^2), \tag{14}$$

$$\dot{\alpha} = 2\frac{\tilde{a}\tilde{v}^2}{M(1-e^2)^{3/2}}, \tag{15}$$

where $\tilde{a} := |\boldsymbol{S}|/M^2$ is the dimensionless spin angular momentum. [See equation (29) in Barack & Cutler (2004) for the full version of (14).]

In the original AK model, the two precession rates and the orbital frequency can differ significantly from their actual values in a relativistic EMRI. Following Chua & Gair (2015), we correct the starting angular rates $(\dot{\Phi}_0, \dot{\tilde{\gamma}}_0, \dot{\alpha}_0)$ through a parameter-space map $(M, \tilde{a}, \nu) \mapsto (M', \tilde{a}', \nu')$ such that

$$\dot{\Phi}_0(M', \tilde{a}', \nu') = \omega_r(M, \tilde{a}, \nu), \tag{16}$$

$$\dot{\tilde{\gamma}}_0(M', \tilde{a}', \nu') = \omega_\theta(M, \tilde{a}, \nu) - \omega_r(M, \tilde{a}, \nu), \tag{17}$$

$$\dot{\alpha}_0(M', \tilde{a}', \nu') = \omega_\phi(M, \tilde{a}, \nu) - \omega_\theta(M, \tilde{a}, \nu), \tag{18}$$

where $\omega_{r,\theta,\phi}$ are the fundamental frequencies of radial, polar, and azimuthal motion (Schmidt 2002) on the starting Kerr geodesic. This map does not eradicate the accumulated phase error of AK waveforms over the full inspiral, but has negligible computational cost and greatly improves the quantitative accuracy over short times.

An EMRI has 14 degrees of freedom in the AK model, which neglects the spin of the compact object; the parameters of the model are chosen to decouple the seven source-intrinsic degrees of freedom from the observer-dependent extrinsic ones. We define the set of dimensionless AK parameters as

$$\Theta_{\rm AK} = \Theta_{\rm int} \cup \Theta_{\rm ext}, \tag{19}$$

$$\Theta_{\rm int} = \left\{\lg\left(\frac{\mu}{M_\odot}\right), \lg\left(\frac{M}{M_\odot}\right), \tilde{a}, e_0, \cos\lambda, \Phi_0, \tilde{\gamma}_0\right\}, \tag{20}$$

$$\Theta_{\rm ext} = \left\{\tilde{v}_0, \alpha_0, \cos\theta_K, \phi_K, \cos\theta_S, \phi_S, \lg\left(\frac{D}{\rm Gpc}\right)\right\}, \tag{21}$$

where the subscript zero denotes the value taken by a quantity at the arbitrary starting time $t = 0$.

2.2 Bumpy analytic kludge model

The family of generically deformed black hole spacetimes derived separately by Benenti & Francaviglia (1979) and Vigeland et al. (2011) provides a model-independent setting in which to construct modified-gravity EMRI waveforms for testing the Kerr metric solution in GR. These bumpy black holes are not required to satisfy the Einstein equations (as done in Collins & Hughes 2004; Vigeland & Hughes 2010), but are Kerr-like through their stationarity, axisymmetry, and admission of an approximate second-rank Killing tensor.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

They are also more general than the modified Kerr spacetimes considered in other proposed EMRI tests (Glampedakis & Babak 2006; Barack & Cutler 2007), as they allow for mass-moment deformations beyond quadrupole order.

In Boyer–Lindquist coordinates, the components of the Kerr metric around a black hole with mass M and spin angular momentum $a = \tilde{a}M$ are given by (Misner et al. 1973)

$$g^{\rm K}_{tt} = -\left(1 - \frac{2Mr}{\Sigma}\right), \quad g^{\rm K}_{t\phi} = -\frac{2Mar\sin^2\theta}{\Sigma}, \quad g^{\rm K}_{rr} = \frac{\Sigma}{\Delta},$$
$$g^{\rm K}_{\theta\theta} = \Sigma, \quad g^{\rm K}_{\phi\phi} = \left(r^2 + a^2 + \frac{2Ma^2r\sin^2\theta}{\Sigma}\right)\sin^2\theta, \tag{22}$$

where $\Sigma := r^2 + a^2\cos^2\theta$ and $\Delta := r^2 - 2Mr + a^2$. The Kerr metric is stationary and axisymmetric, and hence may be cast in Lewis–Papapetrou form via the partial coordinate transformation

$$(\rho, z) = \left(\sqrt{\Delta}\sin\theta, (r - M)\sin\theta\right), \tag{23}$$

with the temporal and azimuthal coordinates (t, ϕ) left unchanged.

Vigeland et al. (2011) consider a linear deformation of the Kerr metric in Lewis–Papapetrou form, and apply the inverse of the transformation (23) to obtain its components in Boyer–Lindquist-like coordinates (r, θ); this ensures that the deformed metric remains stationary and axisymmetric, i.e. it admits temporal and azimuthal Killing vectors t^μ and l^μ such that

$$\nabla_{(\mu}t_{\nu)} = \nabla_{(\mu}l_{\nu)} = 0. \tag{24}$$

The deformed metric components may then be written as

$$g_{\mu\nu} = g^{\rm K}_{\mu\nu} + \epsilon h_{\mu\nu}, \tag{25}$$

where $\epsilon \ll 1$ is a bookkeeping parameter for the metric deformation $h_{\mu\nu}$. Other Kerr-like properties are included by requiring the second-rank tensor $\xi_{\mu\nu} = \Delta t_{(\mu}l_{\nu)} + r^2 g_{\mu\nu}$ to approximately satisfy the Killing tensor equation, i.e.

$$\nabla_{(\lambda}\xi_{\mu\nu)} = \mathcal{O}(\epsilon^2), \tag{26}$$

and by requiring $|h_{\mu\nu}| = \mathcal{O}(1/r^2)$ as $M/r \to 0$ (such that the deformed metric retains asymptotic flatness, along with its original mass and spin angular momentum).

With the above constraints, the only nonzero components of $h_{\mu\nu}$ are h_{tt}, $h_{t\phi}$, h_{rr}, and $h_{\phi\phi}$, which depend on the black hole parameters $(M, \tilde{a})$ and three arbitrary radial functions γ_i, $i \in \{1, 3, 4\}$. By representing the nonzero components and radial functions as power series in M/r, Gair & Yunes (2011) obtain expressions for $h_{\mu\nu,n}$, $2 \leq n \leq 5$ in terms of $\gamma_{i,n}$, where these quantities are the coefficients of the $(M/r)^n$ terms in the corresponding series. If the inclination angle of a geodesic orbit is approximated as constant (as done in the AK model), a metric deformation that is purely[1] nth order in M/r turns out to be fully specified by a set of three coefficients, $\mathcal{B}_n := \{\gamma_{1,n}, \gamma_{4,n}, \gamma_{3,n+1}\}$; we refer to such a deformation as a $\mathcal{B}_n$ bump.

The AK formalism is then used to construct EMRI waveforms on the bumpy black hole spacetime, which is parametrized by the set of coefficients $\bigcup_n \mathcal{B}_n$, $2 \leq n \leq 5$ in addition to $(M, \tilde{a})$. Both the long-time-scale radiation reaction fluxes and the short-time-scale precession rates of the EMRI are altered by the $\mathcal{B}_n$ coefficients, since they perturb the three first integrals of motion along a timelike geodesic with four-velocity u^μ: the energy $E = t_\mu u^\mu$ and orbital angular momentum $L_z = l_\mu u^\mu$, which remain conserved, and the analogue of the Carter constant $Q = \xi_{\mu\nu}u^\mu u^\nu$, which is conserved at linear order in the metric deformation.

By matching the turning points of motion for an instantaneous geodesic orbit (E, L_z, Q) with those of a precessing Keplerian orbit (υ, e, λ) in the AK model, Gair & Yunes (2011) compute the leading-order corrections caused by the $\mathcal{B}_n$ deformations to the fluxes and angular rates (12)–(15). Each set of corrections at nth order in M/r depends only on a single linear combination $\epsilon_n := \gamma_{1,n} + 2\gamma_{4,n}$ of the $\mathcal{B}_n$ coefficients ($\gamma_{3,n+1}$ is at sub-leading order), which effectively reduces the additional degrees of freedom in the extended model to four. The corrections are given by

$$\delta\dot{\upsilon}_n = \frac{8}{5\pi}\frac{\eta\tilde{\upsilon}^{(2n+9)/3}}{M^2(1-e^2)^{n+5/2}}g_{\upsilon,n}\epsilon_n, \tag{27}$$

$$\delta\dot{e}_n = -\frac{16}{5}\frac{\eta\tilde{\upsilon}^{(2n+6)/3}}{M(1-e^2)^{n+3/2}}g_{e,n}\epsilon_n, \tag{28}$$

$$\delta\dot{\bar{\gamma}}_n = \frac{\tilde{\upsilon}^{(2n+1)/3}}{M(1-e^2)^{n-1}}g_{\bar{\gamma},n}\epsilon_n, \tag{29}$$

$$\delta\dot{\alpha}_n = -\frac{\tilde{\alpha}\tilde{\upsilon}^{(2n+2)/3}}{M(1-e^2)^{n-1/2}}g_{\alpha,n}\epsilon_n, \tag{30}$$

where explicit expressions for the eccentricity-dependent factors $g_{\cdot,n}$ are given by equations 301, 302, and 335–346 in Gair & Yunes (2011) (with $g_{\bar{\gamma},2} = 1/2$ and $g_{\alpha,2} = 1$).

In summary, the bAK model comprises (i) adding the corrections (27)–(30) to the corresponding evolution equations (12)–(15) in the AK model, and (ii) extending the set of model parameters to

$$\Theta_{\rm bAK} = \Theta_{\rm AK} \cup \Theta_{\mathcal{B}}, \tag{31}$$

$$\Theta_{\mathcal{B}} = \{\lg\epsilon_n \,|\, 2 \leq n \leq 5\}, \tag{32}$$

where we assume $\epsilon_n > 0$ for simplicity. The deformation parameters ϵ_n then determine the magnitudes of the $\mathcal{B}_n$ bumps, which manifest as phase drifts in the corresponding bAK waveforms. Fig. 1 illustrates how these waveforms dephase over time relative to the AK waveform (where $\epsilon_n = 0$ for all n). Full EMRI waveforms can have up to $\sim 10^5$ observable cycles, and hence might be able to constrain the leading-order bumps down to $\epsilon_n \sim 10^{-7}$ (Moore et al. 2017). For the sake of comparison, combined results for the LIGO events GW150914 and GW151226 in a less conservative single-parameter analysis placed no upper bound below $\sim 10^{-1}$ on the dimensionless deformation parameters of the generalized black hole binary merger model (Abbott et al. 2016a).

2.3 Sinusoidal toy model

Motivated by the phase behaviour of the bAK waveforms, we define a deformed sinusoidal model in which the phase evolution has similar dependence on a set of deformation parameters analogous to $\Theta_{\mathcal{B}}$. Waveforms from this toy model are several orders of magnitude faster to compute than bAK waveforms, and as such are useful for building intuition during method development. They are qualitatively equivalent to the special case of bAK waveforms from circular, equatorial EMRIs over short time-scales (or alternatively, in the geodesic limit $\eta = 0$).

[1]For simplicity, the allowed deformations in the bAK model are restricted to 'pure' bumps and their linear combinations, e.g. a $\mathcal{B}_4$ bump is not the most general fourth-order deformation, but defined as one for which the $\mathcal{B}_{2,3,5}$ coefficients are all zero. While the model might not fully represent deviations from GR at sub-leading order, it is an adequate surrogate in this work.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

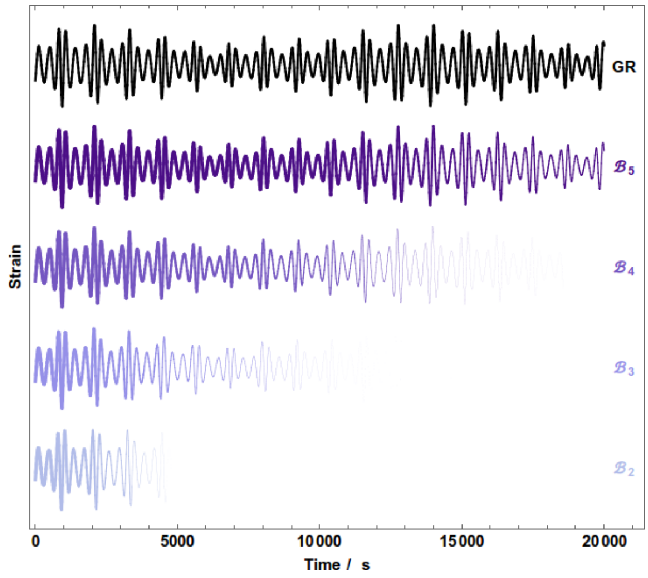

Figure 1. Comparison of bAK waveforms with common source parameters and different $\mathcal{B}_n$ bumps of the same magnitude ($\epsilon_n = 0.4$). These waveforms start in phase with the GR-based AK waveform by construction, but lose phase coherence (indicated by opacity) over time. Higher order $\mathcal{B}_n$ waveforms correspond to smaller metric deformations, and thus dephase more slowly. Figure has been reproduced from Moore et al. (2017).

Our toy model consists simply of a sine wave parametrized by amplitude A and angular frequency Ω, along with four parameters that deform the phase at different orders in time; these deformation parameters are denoted ϵ_n by way of analogy to the bAK model. The toy waveform is given by[2]

$$h = A \sin\left[\Omega\tilde{t}\left(1 + \sum_{n=2}^{5} \epsilon_n \left(\frac{\tilde{t}}{\tau}\right)^{n-1}\right)\right], \tag{33}$$

where $\Omega, \tilde{t}$, and τ are all dimensionless. The time-scale-like quantity τ has been introduced to control the overall strength of the phase drift, and is not treated as a parameter in the model. Henceforth we fix $\tau = 2T_{\rm obs}$, where $T_{\rm obs}$ is some specified waveform duration, and define the set of toy model parameters as

$$\Theta_{\rm toy} = \{A, \Omega\} \cup \Theta_{\mathcal{B}}, \tag{34}$$

where $\Theta_{\mathcal{B}}$ is defined as in (32).

3 STATISTICAL FRAMEWORK

The framework presented in this paper is based on the model-independent and parametric test infrastructure introduced by Li et al. (2012), in which the combinatorial submodels of a generalized waveform model are used in a null-hypothesis test of GR, and a nested-sampling algorithm is employed to compute the Bayesian evidence for each submodel. An overview of the pertinent concepts is given in Sections 3.1–3.3. In Section 3.4, we describe the product-space nested-sampling method considered by Hee et al. (2015); it provides an evidence-free computation of the Bayes factor for each submodel (with respect to the null submodel), thus mitigating the prohibitive computational difficulty of performing the test with a generalized EMRI waveform model. The estimation of Bayes factor errors in the method also requires a new rethreading technique (Higson et al. 2017), which we outline in Section 3.5.

3.1 Gravitational-wave likelihood

In the basic matched-filtering framework for GW data analysis, data from a single interferometric detector is written as the time series $x = h + n$, where h is the detector response to a passing GW and n is the detector noise (typically approximated as a Gaussian and stationary random process). For a LISA-like detector with three arms, two independent signals $h_{I,II}$ may be obtained; these are related on average to the two GW polarizations $h^{+,\times}$ by

$$h_{I,II} = \frac{\sqrt{3}}{2}\left(F^{+}_{I,II}h^{+} + F^{\times}_{I,II}h^{\times}\right), \tag{35}$$

where the antenna pattern functions $F^{+,\times}_{I,II}$ (Apostolatos et al. 1994) depend on the sky location and polarization angle of the source in a detector-based coordinate system. [See equations 15–17 in Barack & Cutler (2004) for the explicit expressions in ecliptic coordinates.] Doppler modulation of the waveform phase may also be included in $h^{+,\times}$ to account for the orbital motion of LISA.

For compactness, we write the GW signal as a complex time series $h = h_I + \mathrm{i}h_{II}$.[3] The signal-to-noise ratio (SNR) of h is given by $\rho := \sqrt{\langle h|h\rangle} = \sqrt{\langle h_I|h_I\rangle + \langle h_{II}|h_{II}\rangle}$, with the noise-weighted inner product $\langle\cdot|\cdot\rangle$ defined as (Cutler & Flanagan 1994)

$$\langle a|b\rangle = 2\int_0^\infty \mathrm{d}f\, \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}, \tag{36}$$

where S_n is the power spectral density of n. If the detector noise is assumed to be white (as done in Section 4), (36) simplifies to

$$\langle a|b\rangle = \frac{1}{S_n}\int_0^{T_{\rm obs}} \mathrm{d}t\, a^*(t)b(t), \tag{37}$$

which may be computed directly from the time-domain waveforms. In Section 5, S_n is given instead by an analytic approximation (Amaro-Seoane et al. 2013) to the LISA noise power spectral density.[4] Throughout this paper, we work with waveforms that are renormalized with respect to some reference waveform h' and specified amplitude ρ', i.e. $h \to \rho' h/\sqrt{\langle h'|h'\rangle}$; hence it is more useful to think of 'SNR' here as an amplitude relative to the defined norm in (36), and not to the actual noise in the data.

Given GW data x that contains a signal $h(\theta_*)$ corresponding to the model parameter values θ_*, the Bayesian likelihood $\mathcal{L} = \Pr(x|\theta)$ is defined as (Cutler & Flanagan 1994)

$$\mathcal{L}(\theta) \propto \exp\left(-\frac{1}{2}\langle x - h(\theta)|x - h(\theta)\rangle\right). \tag{38}$$

We consider only EMRI waveforms with $\rho \gtrsim 10$ throughout this paper, which is a conservative choice that is consistent with typical values of the threshold SNR for reliable detection (Babak et al. 2017; Chua et al. 2017). In this regime, the noise term in the log-likelihood is suppressed by a factor of $\gtrsim$100 relative to the leading term, and may be neglected for the purposes of model selection. Hence, to simplify analysis in Sections 4–5, we assume a particular noise realization of $n = 0$ in the synthetic data (i.e. $x = h(\theta_*)$).

[2]Although full GW models require two independent waveform modes for parameter estimation, the likelihood obtained with the single toy model mode is qualitatively sufficient for the purposes of this work.

[3]This notation is also compatible with the single real waveform mode h of the sinusoidal toy model.

[4]The noise model used corresponds to a down-scoped version of LISA, and thus gives more conservative results; the mission design has now been restored to an earlier configuration with higher sensitivity.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

Table 1. Scale for interpreting posterior odds ratios. Since $P^i_j = -P^j_i$, negative values of P^i_j correspond to reversed model odds.

P^i_j	Odds for $\mathcal{M}_i$ over $\mathcal{M}_j$
$0 \lesssim P^i_j \lesssim 1$	None
$1 \lesssim P^i_j \lesssim 3$	Slight
$3 \lesssim P^i_j \lesssim 5$	Significant
$5 \lesssim P^i_j$	Decisive

3.2 Null-hypothesis test

From Bayes' theorem, the posterior probability of a model hypothesis $\mathcal{M}$ given the data x may be written as

$$\Pr(\mathcal{M}\,|x) = \frac{\mathcal{Z}\Pi}{\Pr(x)}, \tag{39}$$

where $\mathcal{Z} = \Pr(x|\,\mathcal{M})$ and $\Pi = \Pr(\mathcal{M})$ are, respectively, the evidence and prior probability for the model. The model evidence is typically obtained by marginalizing the likelihood $\mathcal{L} = \Pr(x|\theta, \mathcal{M})$ over the model parameters, i.e.

$$\mathcal{Z} = \int \mathrm{d}\theta\ \mathcal{L}(\theta)\pi(\theta), \tag{40}$$

where $\pi = \Pr(\theta|\,\mathcal{M})$ is the parameter prior.

Two model hypotheses $\mathcal{M}_{i,j}$ may be compared through the ratio of their posterior probabilities, which quantifies the degree of belief in one model over the other. The logarithm of this posterior odds ratio is defined as

$$P^i_j := \ln\left[\frac{\Pr(\mathcal{M}_i\,|x)}{\Pr(\mathcal{M}_j\,|x)}\right] = B^i_j + \ln\left(\frac{\Pi_i}{\Pi_j}\right), \tag{41}$$

where $B^i_j := \ln(\mathcal{Z}_i\,/\,\mathcal{Z}_j)$. The quantity B^i_j is the logarithm of the Bayes factor, which is commonly used as an equivalent substitute for the posterior odds ratio through the implicit assumption $\Pi_i = \Pi_j$. A general scale for the interpretation of posterior odds ratios (Jeffreys 1961; Kass & Raftery 1995) is given in Table 1.

For a generalized waveform model that extends a GR-based model $\mathcal{M}_{\rm GR}$ through a set $\Theta_{\mathcal{B}}$ of N deformation parameters, we may define 2^N submodels whose deformation parameter sets are given by the 2^N subsets of $\Theta_{\mathcal{B}}$. The submodel corresponding to the null set is $\mathcal{M}_{\rm GR}$ itself; the remaining $2^N - 1$ submodels are modified-GR models, which we denote collectively by the hypothesis $\mathcal{M}_{\rm modGR} = \bigvee_{m\neq{\rm GR}} \mathcal{M}_m$. Even though $\mathcal{M}_{\rm modGR}$ is a collection of nested submodels, Li et al. (2012) observe that the individual submodel pieces of evidence are logically disjoint, due to the distinct integration measure on each submodel parameter space. Hence the posterior odds ratio for $\mathcal{M}_{\rm modGR}$ over $\mathcal{M}_{\rm GR}$ simplifies to

$$\begin{aligned}\frac{\Pr(\mathcal{M}_{\rm modGR}\,|x)}{\Pr(\mathcal{M}_{\rm GR}\,|x)} &= \frac{\sum_{m\neq{\rm GR}}\Pr(\mathcal{M}_m\,|x)}{\Pr(\mathcal{M}_{\rm GR}\,|x)} \\ &= \frac{1}{2^N-1}\sum_{m\neq{\rm GR}}\frac{\mathcal{Z}_m}{\mathcal{Z}_{\rm GR}},\end{aligned} \tag{42}$$

where the second equality follows from the assumptions that $\Pi_{\rm modGR} = \Pi_{\rm GR}$ and $\Pi_m = \Pi_{m'}$ for all $m, m' \neq {\rm GR}$.

This approach provides the basic framework for a test of GR with the bAK model, where the modified-GR hypothesis is compared against the null hypothesis (the AK model) through evaluation of the submodel Bayes factors in (42). A single EMRI source is considered for the assessment of our methods, but it is straightforward to generalize (42) for a population of sources (Li et al. 2012). In the bAK model (where $N = 4$), the 16 submodels may be indexed by $0 \leq m \leq 15$, whose nonzero digits in binary representation specify the deformation parameters included in each submodel. Hence the AK model is denoted $\mathcal{M}_0 \equiv \mathcal{M}_{0000}$, while the full bAK model is $\mathcal{M}_{15} \equiv \mathcal{M}_{1111}$. We use the convention that the binary digits of m from right to left correspond to ϵ_n with n= 2, 3, 4, 5, respectively; for example, $\mathcal{M}_{1010}$ has only two deformation parameters $\{\epsilon_3, \epsilon_5\}$, and is equivalent to the $\epsilon_2 = \epsilon_4 = 0$ slice of the full model.

Although the Bayesian evidence (40) has a built-in Occam penalty on model complexity, it does not fully account for the multiplicity effect (Jeffreys 1961; Scott & Berger 2010). As a larger number of deformation parameters is considered, it becomes more likely that one particular parameter will cause the inclusive submodels to fit the data well by chance. In other words, a uniform submodel prior (i.e. $\Pi_m = \Pi_{m'}$ for all $m, m' \neq 0$) might not be appropriate for the nested submodels in this framework, and more thorough prescriptions for assigning model prior probability (George & Foster 2000; Consonni, Forster & La Rocca 2013; Villa & Walker 2015) should be investigated. However, the number of deformation parameters in the bAK model is small; furthermore, the focus of this work is the evaluation of (42) with improved precision, and not the validation of its accuracy. Hence we retain a uniform submodel prior for simplicity, and it follows from (41) and (42) that

$$P^{\rm modGR}_{\rm GR} = B^{\rm modGR}_{\rm GR} = \ln\left[\sum_{m=1}^{15}\exp\left(B^m_0\right)\right] - \ln 15. \tag{43}$$

3.3 Nested sampling

For most statistical problems, (40) admits no analytic solution and is impractical to evaluate through direct numerical integration, even over a parameter space of modest dimensionality. A wealth of alternative techniques for computing the evidence has thus been developed; these range from simple estimates that use Laplace's approximation (Tierney & Kadane 1986) or Chib's method (Chib 1995), to more sophisticated sampling strategies based on concepts such as thermodynamic integration and simulated annealing (Meng & Wong 1996; Gelman & Meng 1998; Neal 2001). We employ in this work the nested sampling strategy introduced by Skilling (2004), which provides an accurate and computationally efficient means of simultaneously exploring the posterior surface and evaluating the evidence. Nested sampling has been shown to be suitable for GW likelihoods (Feroz et al. 2009a; Veitch & Vecchio 2010), and is one of the two algorithms used to compute model pieces of evidence in LIGO tests of GR (Abbott et al. 2016b).

In nested sampling, the multidimensional integral (40) is written in the one-dimensional form (Skilling 2006)

$$\mathcal{Z} = \int_0^1 \mathrm{d}X\ \mathcal{L}(X), \tag{44}$$

where the prior mass X is given by

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} \mathrm{d}\theta\,\pi(\theta), \tag{45}$$

i.e. the integral of the parameter prior over the interior of the likelihood contour $\mathcal{L} = \lambda$. A set of $N_{\rm live}$ initial points is first sampled from the prior; this set of 'live' points is then evolved across parameter space by discarding at the i-th iteration the point θ_i with the lowest likelihood value λ_i, and replacing it with one drawn from the prior but within the contour $\mathcal{L}_i = \lambda_i$.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

Skilling (2006) shows that the prior mass corresponding to each λ_i may be approximated as

$$X_i \approx \exp\left(-\frac{i}{N_{\rm live}}\right), \tag{46}$$

which shrinks exponentially from $X_0 := 1$ to zero as the live points converge on and navigate the bulk of the posterior surface. Upon the satisfaction of suitable convergence criteria, the algorithm is truncated and the evidence (44) may be approximated as

$$\mathcal{Z} \approx \sum_{i>0} w_i \, \mathcal{L}_i, \tag{47}$$

where the weights w_i are given by a Riemann sum rule, e.g. $w_i = X_{i-1} - X_i$. The set of discarded live points θ_i (often termed dead points) also serves as a set of posterior samples, through the assignment of posterior probability $p_i = w_i \, \mathcal{L}_i \, / \, \mathcal{Z}$ to each point.

The prior-mass approximation (46) introduces statistical error into the posterior probabilities p_i and the evidence estimate (47), via the weights w_i. The uncertainty in the evidence estimate depends on the absolute error of each w_i, and is dominated by the Poisson variability in the number of iterations taken to reach the posterior bulk. Hence (47) is log-normally distributed, and the standard deviation of $\ln \mathcal{Z}$ in nested sampling scales with the number of live points as (Skilling 2006)

$$\sigma_{\ln \mathcal{Z}} \propto \frac{1}{\sqrt{N_{\rm live}}}. \tag{48}$$

It is often difficult to sample from the prior under the constraint $\mathcal{L} > \lambda$, since the likelihood contours $\mathcal{L} = \lambda$ might in general be multimodal or degenerate. In this work, we make use of the nested-sampling implementation POLYCHORD (Handley et al. 2015a,b), which mitigates these difficulties through the incorporation of clustering and slice sampling algorithms. It also exhibits good scaling with the dimensionality of the parameter space, and in that sense is an improved successor to the widely adopted nested-sampling tool MULTINEST (Feroz, Hobson & Bridges 2009b; Feroz et al. 2013). The two main runtime parameters in POLYCHORD that determine sampling resolution and reliability are, respectively, $N_{\rm live}$ and $N_{\rm rep}$; the latter is the number of randomly oriented one-dimensional slices sampled at each iteration in order to decorrelate the new live point from the discarded point.

3.4 Product-space nested sampling

Methods exist for obtaining Bayes factors without explicitly evaluating pieces of evidence, and these are especially useful when the number of competing models is large. The most well known is the Savage–Dickey density ratio for nested models [and generalizations thereof (Verdinelli & Wasserman 1995; Marin & Robert 2010; Wetzels, Grasman & Wagenmakers 2010)], where the Bayes factor between a null model and an encompassing one is given by the ratio between the posterior and prior probabilities at the null point in the larger model space. Another class of methods uses the product-space representation (Carlin & Chib 1995) for a prespecified and indexed collection of competing models; this approach involves exploring the set of model indices and each model space simultaneously, and has been investigated in the context of Markov chain Monte Carlo (MCMC) algorithms (Carlin & Chib 1995; Godsill 2001; Lodewyckx et al. 2011) and nested sampling (Hee et al. 2015).

Product-space sampling and reversible-jump MCMC (Green 1995) are modern examples of transdimensional frameworks for Bayesian model selection (Sisson 2005). One disadvantage of product-space methods (as compared to reversible-jump MCMC) is the required prior specification of all the competing models. This is not an issue in our framework since the bAK submodels are fully defined by the power set of $\Theta_{\mathcal{B}}$, and the $\mathcal{B}_n$ bumps are truncated at $n = 5$ (their effects are exponentially suppressed as n increases). In the product-space representation, the 16 submodels $\mathcal{M}_m$ in Section 3.2 may be thought of as distinct 'slices' of some hypermodel $\mathcal{M}$; the parameter space of $\mathcal{M}$ is the combination[5] of all the submodel spaces, and is parametrized by

$$\Theta = \Theta_{\rm gen} \cup \{m\}, \tag{49}$$

where $\Theta_{\rm gen}$ is the parameter set for the generalized waveform model (i.e. 31 or 34) and m is the submodel index.

The posterior probability for m is given by (Hee et al. 2015)

$$\begin{aligned}
\Pr(m|x, \mathcal{M}) &= \int \mathrm{d}\theta \, \Pr(\theta, m|x, \mathcal{M}) \\
&= \frac{1}{\mathcal{Z}_{\mathcal{M}}} \int \mathrm{d}\theta \, \mathcal{L}(\theta, m)\pi(\theta, m) \\
&= \frac{\pi(m)}{\mathcal{Z}_{\mathcal{M}}} \int \mathrm{d}\theta_m \, \mathcal{L}(\theta_m)\pi(\theta_m|m) \\
&= \frac{\pi(m)\,\mathcal{Z}_m}{\mathcal{Z}_{\mathcal{M}}},
\end{aligned} \tag{50}$$

where $\mathcal{Z}_{\mathcal{M}}$ and $\mathcal{Z}_m$ are the pieces of evidence for the hypermodel and mth submodel, respectively. To obtain the third equality, we have decomposed θ into the parameters θ_m that are included in the submodel $\mathcal{M}_m$, and the parameters ϕ_m that are excluded; the integral $\int d\phi_m \, \pi(\phi_m) = 1$ then factors out of the expression.

When sampling in the hypermodel space, it is convenient to assign a uniform prior $\pi(m) = 1/16$ on the submodel index, since it is straightforward to restore the assumption $\sum_{m \neq 0} \Pi_m = \Pi_0$ in post-processing [i.e. through the final term in (43)]. We then have

$$B_0^{m'} = \ln\left(\frac{\mathcal{Z}_{m'}}{\mathcal{Z}_0}\right) = \ln\left[\frac{\Pr(m = m'|x, \mathcal{M})}{\Pr(m = 0|x, \mathcal{M})}\right], \tag{51}$$

such that the Bayes factors in (43) may be obtained directly from the sampled posterior distribution for m. A nested-sampling implementation of the product-space method was found to be viable by Hee et al. (2015) through application to a cosmological problem. In this paper, we use a similar implementation to compare a larger number of models in our GW test of GR, and perform an improved investigation of the errors on the obtained Bayes factors.

3.5 Error estimation from rethreading

If the Bayes factors in (43) are obtained by evaluating the individual submodel pieces of evidence with regular nested sampling, the standard deviation of $P_{\rm GR}^{\rm modGR}$ scales as $1/\sqrt{N_{\rm live}}$ (from propagation of the log-evidence error (48)). In the product-space method, however, the Bayes factors are ratios of posterior probabilities $p_m := \Pr(m|x, \mathcal{M})$; computing them is then a parameter estimation problem, with a different associated uncertainty that depends on the posterior errors σ_{p_m}. These errors arise from the prior-mass approximation (46), as in the case of $\sigma_{\ln \mathcal{Z}}$, but also from the dimensional reduction in the likelihood reparametrization $\mathcal{L}(\theta) \to \mathcal{L}(X)$ (Higson et al. 2017). Nevertheless, Higson et al. (2017) observe that σ_{p_m} are determined by the relative errors of the nested-sampling weights,

[5] More precisely, the hypermodel parameter space is the vector bundle of submodel spaces over the discrete set of model indices.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

and hence are typically $< \sigma_{\ln \mathcal{Z}}$. This is the key factor that underpins the improved efficiency of product-space nested sampling.

An algorithm for estimating the errors of p_m (or of any quantity derived from p_m) over a single nested-sampling run has been proposed by Higson et al. (2017). The technique is motivated by the fact that any number of runs r with N_r live points may be merged into a single run with $N_{\rm live} = \sum_r N_r$ (Skilling 2006), by combining and reordering all of their dead points [and adjusting $N_{\rm live}$ in (46) accordingly]. Conversely, it is also possible to unravel a single nested-sampling run into its constituent 'threads', i.e. a set of $N_{\rm live}$ independent runs r with $N_r = 1$. This is achieved by tracking the birth and death order of the sampling points; each thread is then formed from the sequence of replacements for an original live point.

A distribution of runs with the original number of live points may be obtained rapidly from the set of threads, by using bootstrap resampling on the set and recombining (or 'rethreading') the sample threads. These new runs contain only points that are present in the original run, but yield statistical variance in the estimated weights and p_m. In this work, we use the rethreading technique to generate 10^3 realizations of $P_{\rm GR}^{\rm modGR}$ from a single nested-sampling run, and thus to evaluate its mean and standard deviation directly. The rethreading results are also validated against those from repeated runs for the toy model in Section 4.

4 RESULTS: SINUSOIDAL TOY MODEL

To demonstrate that product-space nested sampling with rethreading can explore qualitative EMRI likelihoods with improved efficiency, we first apply it to synthetic data generated from the toy EMRI model defined in Section 2.3. Two data sets x are considered: one from the 'GR' submodel $\mathcal{M}_{0000}$, and one from the deformed submodel $\mathcal{M}_{0010}$ with a '$\mathcal{B}_3$ bump'. The signal parameters for x_{0000} and x_{0010} are $(A, \Omega) = (1, 1)$ and $(A, \Omega, \lg \epsilon_3) = (1, 1, -1.9)$, respectively. Both waveforms are generated with duration $T_{\rm obs} = 10^4$ (such that they contain $\sim 10^3$ cycles) and a dimensionless sampling rate of unity, then renormalized to SNR $\rho = 10$.

As the toy waveforms are near-sinusoidal, the GW likelihood (38) contains damped oscillations in Ω; to a lesser extent, such oscillations are also present for parameters that are highly correlated with frequency (e.g. M) in more realistic EMRI models. Fig. 2(a) shows the $\lg \epsilon_3$–Ω slice of $\ln \mathcal{L}$ for the submodel $\mathcal{M}_{0010}$ with the data x_{0000}, where Ω is seen to be relatively well localized in the moderate-SNR regime. However, the likelihood is highly degenerate in the deformation parameters. In the case of Fig. 2(a), any value for $\lg \epsilon_3$ that falls below some threshold ≈ -2 ceases to deform the waveform significantly, and hence has negligible effect on the value of $\ln \mathcal{L}$. This results in an extended region of high likelihood along the $\lg \epsilon_3$ axis, even though the true GR signal lies on the $\epsilon_3 = 0$ boundary of the $\mathcal{M}_{0010}$ space.

The deformation parameters are also degenerate among themselves, which further complicates the likelihood surface if the true signal is deformed. For the data x_{0010}, the signal is not even contained in submodels without ϵ_3; however, Fig. 2(b) shows that regions of high likelihood exist in the $\lg \epsilon_4$–$\lg \epsilon_5$ slice of $\ln \mathcal{L}$ for $\mathcal{M}_{1100}$, and that a $\lg \epsilon_3 = -1.9$ deformation can be well approximated by a larger value of ϵ_4 or ϵ_5 (or a combination of the two). Although such degeneracies are partially broken by the additional degrees of freedom in more realistic generalized GW models, they are still expected to hamper parameter estimation for the individual deformations (Li et al. 2012; Moore et al. 2017). Nevertheless, they have a less severe impact on the null-hypothesis test in Section 3.2, where the aim is not to distinguish among the deformed submodels.

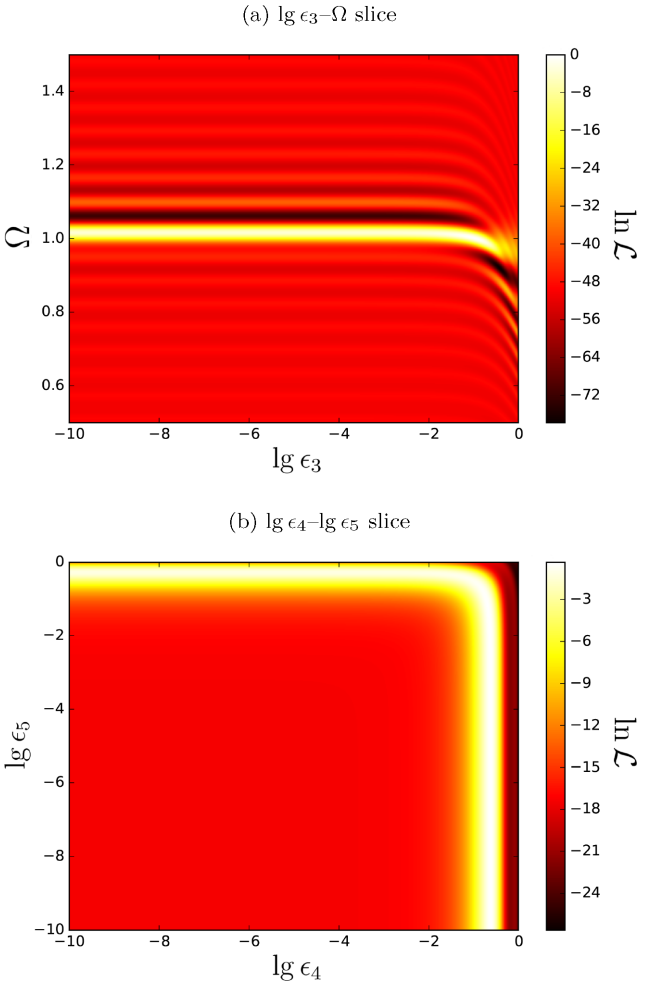

Figure 2. Log-likelihood slices for sinusoidal toy model: (a) submodel $\mathcal{M}_{0010}$ with data x_{0000}; (b) submodel $\mathcal{M}_{1100}$ with data x_{0010}.

For the two data sets, both regular and product-space nested sampling are used to obtain $P_{\rm GR}^{\rm modGR}$ with an associated standard deviation σ_P. In the regular case, (43) is evaluated piecewise by computing the individual submodel pieces of evidence, and the standard nested-sampling estimates for each log-evidence error (48) are propagated to yield σ_P. For the product-space method, $P_{\rm GR}^{\rm modGR}$ is obtained through a single nested-sampling run in the hypermodel space, and its approximate error is calculated in two ways: using the single-run rethreading technique, and from 50 repetitions of the same product-space run. The latter procedure is only to demonstrate that the methods in Sections 3.4 and 3.5 are working as expected, and is not performed in Section 5.

The product-space method manifestly provides computational savings over regular nested sampling in the null-hypothesis test, as it samples in just one parameter space instead of 16. However, the hypermodel space is more complex than a single submodel space and requires additional runtime to explore effectively, such that differences in efficiency cannot simply be estimated from the number of spaces sampled. To assess the gains when using the product-space method, we vary the POLYCHORD runtime parameter $N_{\rm live}$ to obtain different degrees of precision on each corresponding evaluation of $P_{\rm GR}^{\rm modGR}$, and compare the number of likelihood calls taken by the two methods to achieve the same σ_P. The other POLYCHORD runtime

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

parameter is set as $N_{\rm rep} = 30$, which has been chosen empirically to ensure the convergence of regular evidence estimates.

It is also instructive to study how regular and product-space nested sampling perform on the penultimate step in the evaluation of $P_{\rm GR}^{\rm modGR}$, i.e. the individual submodel Bayes factors B_0^m. Fig. 3 shows the B_0^m and associated errors that are obtained from the two methods with $N_{\rm live} = 2500$, for the GR data x_{0000}. The Occam penalty on model complexity is clearly observed in both sets of results, as the relative evidence for each submodel decreases with the number of parameters it contains. However, the product-space method appears to systematically give Bayes factors that are more pronounced (negative), and in tension with the regular results. This is likely because the entire hypermodel space is explored with the same number of live points allocated to each submodel space in the regular method, leading to a slight degree of sampling bias. The Bayes factor errors for both methods are nevertheless comparable, since the errors on a posterior over m are smaller than those on submodel evidence evaluations (as discussed in Section 3.5).

In Fig. 4, the error σ_P of $P_{\rm GR}^{\rm modGR}$ for the two data sets x_{0000} and x_{0010} is plotted against the number of likelihood calls for a sequence of regular and product-space nested-sampling runs with $100 \leq N_{\rm live} \leq 2500$. Both the rethreading and repetition error estimates for the product-space method are included; they are seen to agree well, with the latter showing more scatter (since they are computed from only 50 evaluations of $P_{\rm GR}^{\rm modGR}$, as opposed to 10^3 realizations in the rethreading technique). The rethreading error estimates are further validated through a reduced chi-squared test against the sample mean μ of $P_{\rm GR}^{\rm modGR}$ in Fig. 5, where $\chi^2 = \sum_i (P_i - \mu)^2/(12\sigma_P^2) \approx 1$ for both data sets. We also find $\mu_{0000} \sim -1$ in the GR case and $\mu_{0010} \sim 1$ in the $\mathcal{B}_3$ case, which is by design from our choices of SNR and ϵ_3 for the synthetic data.

For both data sets, it is clear that product-space nested sampling is effective at reducing the computational cost required to reach a given level of precision (or alternatively, at providing greater precision with a given number of likelihood calls). In the GR case, the average gain in efficiency (i.e. the mean horizontal distance between the blue and green curves in Fig. 4) is a factor of around 24. Furthermore, as the likelihood surface over the hypermodel space is less complex for x_{0000}, nested sampling explores it nearly as efficiently as each of the 16 submodel spaces. This is seen by comparing regular and product-space runs of equal $N_{\rm live}$, where the number of likelihood calls taken by the latter is almost exactly 16 times smaller, and its associated error is slightly lower as well.

In the $\mathcal{B}_3$ case, larger overall errors σ_P are obtained for both methods, and the average reduction in computational cost with product-space nested sampling is reduced to a factor of around 9. As expected, this is largely caused by the increased complexity of the likelihood surface over the hypermodel space for x_{0010}: additional modes are present in the parameter spaces of the eight submodels containing ϵ_3, as well as in other submodel spaces due to deformation parameter degeneracies such as that from Fig. 2(b). When comparing regular and product-space runs of equal $N_{\rm live}$, the number of likelihood calls taken by the latter is only 13 times smaller, and its associated error is now slightly higher. It follows from these results that computational savings will likely be smallest for data generated from $\mathcal{M}_{1111}$; we discuss this in Section 5.

Another contribution to the overall difference in σ_P between the two data sets arises from the construction of $P_{\rm GR}^{\rm modGR}$ itself. For the $\mathcal{B}_3$ data, the parameter space of the null submodel $\mathcal{M}_{0000}$ is a region of lower posterior probability in the product-space approach, and hence the relative posterior error on $\Pr(m = 0)$ is higher. This error propagates into every submodel Bayes factor B_0^m via (51), which increases the final error σ_P. The effect is also present for regular nested sampling, since $\mathcal{M}_{0000}$ has a higher relative evidence error as well. However, the increase in σ_P only becomes significant in the strong-deformation regime, which is unlikely for actual tests of GR; furthermore, it can also be mitigated in practice (e.g. by using the actual submodel index prior $\sum_{m\neq 0}\Pi_m = \Pi_0$ when sampling).

5 RESULTS: BUMPY ANALYTIC KLUDGE MODEL

For a more realistic generalized EMRI waveform model in the product-space approach, the dimensionality of the hypermodel parameter space is considerably higher. In the case of the bAK model, $\Theta = \Theta_{\rm AK} \cup \Theta_{\mathcal{B}} \cup \{m\}$ is 19-dimensional and parametrized by 14 GR parameters, four deformation parameters, and the submodel index. The likelihood surface over the GR parameter space is known to be highly multimodal with a large information content (Gair et al. 2004). A full exploration of this space is hampered by the significant computational cost of waveform generation (even though the AK formalism already provides the cheapest EMRI waveforms available), and is beyond the scope of this work. We instead fix all but seven of the hypermodel parameters, allowing only the component masses, deformation parameters and submodel index to vary. For our synthetic data, the intrinsic GR parameters (20) of the signal are chosen as $\Theta_{\rm int}$= (1, 6, 0.1, 0.1, 0.9, 0, 0); the waveform is 2 months long with an initial frequency of 2 mHz (such that it contains $\sim 10^4$ cycles), and is sampled at 0.2 Hz.

Instead of considering synthetic data from a deformed submodel as in Section 4, we restrict analysis here to the more realistic case of a GR signal, and investigate the effect of SNR on sampling performance. The two data sets studied in this section are both generated from $\mathcal{M}_{0000}$ with the same GR parameters, but are renormalized to SNRs $\rho = 10$ and $\rho = 100$, respectively. Fig. 6 shows the B_0^m and associated errors that are obtained from regular and product-space nested sampling with $N_{\rm live} = 500$, for the moderate-SNR case $\rho = 10$. The Occam penalty is again apparent, but the submodel Bayes factors are generally lower than in Fig. 3 even though the SNR is the same; this is likely due to the EMRI's orbital evolution reducing degeneracy in the deformed parameters. As in Fig. 3, there also appears to be a slight systematic difference between the Bayes factors from the two methods. Both the regular and product-space methods correctly favour the null submodel $\mathcal{M}_{0000}$, although the latter does so with smaller errors and fewer likelihood calls for the same number of live points.

In Fig. 7, the error σ_P of $P_{\rm GR}^{\rm modGR}$ for the $\rho = 10$ and $\rho = 100$ data sets is plotted against the number of likelihood calls for several product-space nested-sampling runs with varying POLYCHORD runtime parameters $N_{\rm live}$ and $N_{\rm rep}$. These are compared against a single regular nested-sampling run for each SNR value (due to the considerably higher computational cost of the regular method). It is clear that the efficiency gains obtained for the sinusoidal toy model are still present for the bAK model, in that every product-space run shown has both better precision and lower computational cost than the two regular runs. As indicated by the results in Section 4, these gains might be diminished in the case of data that is generated from more complex submodels such as $\mathcal{M}_{1111}$. However, since there is strong prior expectation for an actual EMRI signal to be well described by GR, the order-of-magnitude savings observed here will likely be close to what is obtained in practice.

Furthermore, since the log-evidence error (48) scales as $1/\sqrt{N_{\rm live}}$ and the number of likelihood calls increases approximately linearly with $N_{\rm live}$, an extrapolation of both regular nested-sampling runs

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

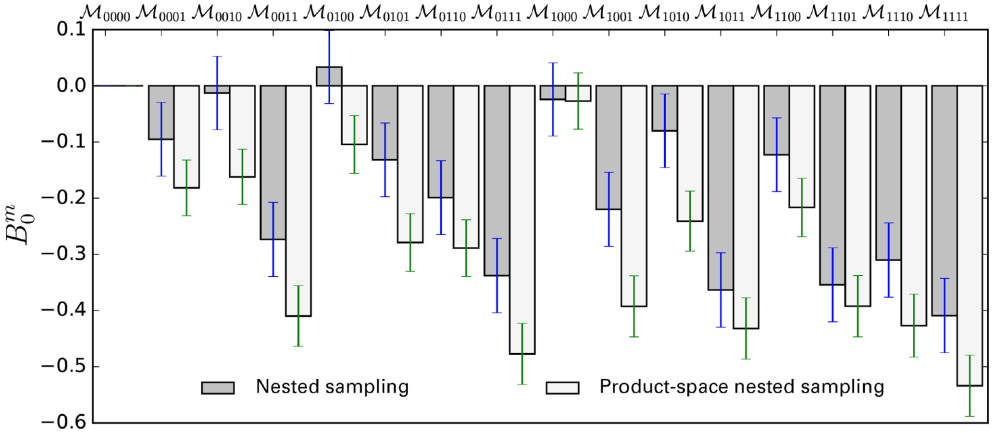

Figure 3. Submodel Bayes factors B_0^m ($B_0^0 = 0$ by definition) for regular and product-space nested sampling with sinusoidal toy data x_{0000} (SNR $\rho = 10$). Both methods have $N_{\rm live} = 2500$; regular nested sampling takes 4.2×10^7 likelihood calls in total, while product-space nested sampling takes 2.7×10^6 calls.

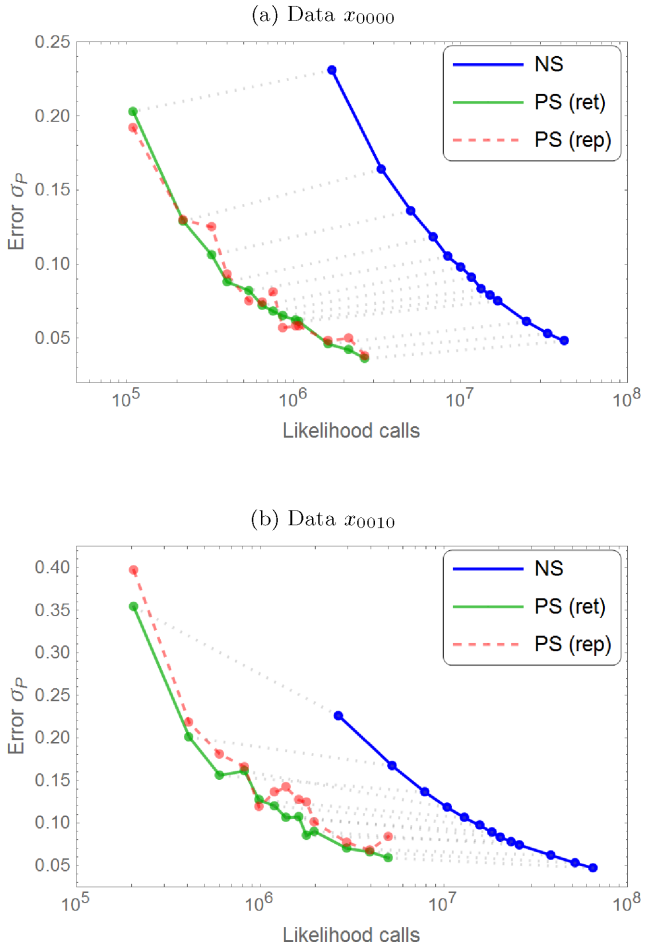

Figure 4. Error σ_P of $P_{\rm GR}^{\rm modGR}$ for (a) GR data x_{0000} and (b) $\mathcal{B}_3$ data x_{0010}. Regular nested sampling (blue) is compared to product-space nested sampling with error estimates from single-run rethreading (green) and 50 repeated runs (red, dashed). Grey dotted lines indicate regular and product-space (rethreading) runs of equal $N_{\rm live}$, ranging from 100 to 2500.

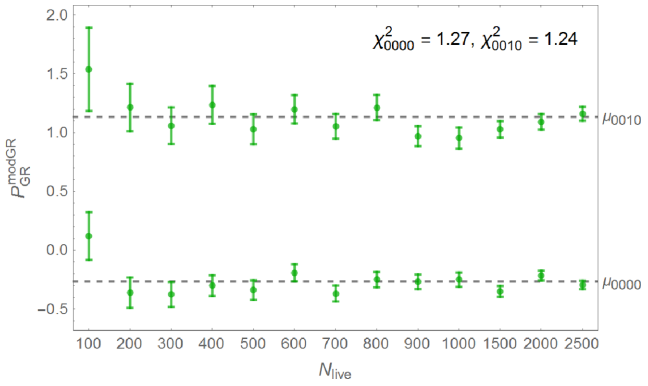

Figure 5. Reduced chi-squared test of rethreading errors against sample mean μ of $P_{\rm GR}^{\rm modGR}$, for GR data x_{0000} (bottom) and $\mathcal{B}_3$ data x_{0010} (top). Each sequence of 13 product-space runs with varying $N_{\rm live}$ corresponds to a green curve in Fig. 4. The number of degrees of freedom is 12.

to 5 per cent error ($\sim 10^8$ calls) indicates that for the bAK likelihood, the reduction of computational cost from the product-space method is boosted to around two orders of magnitude. A direct verification of this statement is impractical for this work, since the cost of each $\sim 10^7$-call run in Fig. 7 is $\approx$4000 core hours (a single call to the bAK likelihood takes $\approx$1.5 s). Nevertheless, the scaling of errors on nested-sampling pieces of evidence is both well understood and reliable (as indicated by the smoothness of the blue curves in Fig. 4), which lends credence to the validity of such an extrapolation.

Significant sampling bias is introduced into the results for the high-SNR case $\rho = 100$; this is because optima in the likelihood surface become more localized, and hence require higher sampling resolution/reliability to map out accurately. The values of $P_{\rm GR}^{\rm modGR}$ and their associated errors from the 10 high-SNR product-space runs (indicated by the light green points in Fig. 7) are plotted in Fig. 8. Computing the reduced chi-squared statistic for these runs gives $\chi^2 \approx 5$, such that σ_P for any given run appears to underrepresent the observed scatter within the set of runs. Since the rethreading technique has been validated in Section 4, the large χ^2 suggests that the hypermodel space is not being sampled consistently between runs. If any particular run misses a region of high posterior probability, then its threads will not contain enough information to provide a rethreading estimate that represents its true error; this could occur if $N_{\rm live}$ is not large enough to find sharply defined modes, or if replacement live points are correlated with discarded ones due to inadequate $N_{\rm rep}$. The sampling bias observed here (and to a lesser

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

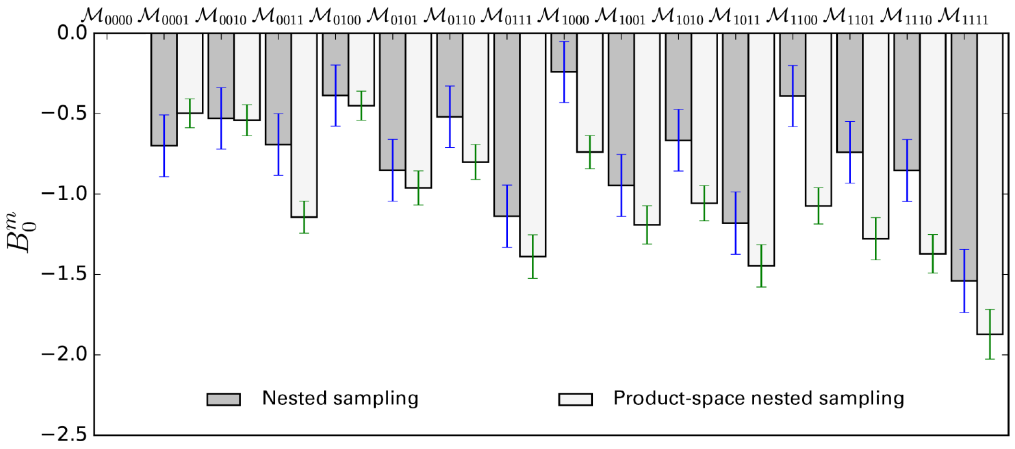

Figure 6. Submodel Bayes factors B_0^m ($B_0^0 = 0$ by definition) for regular and product-space nested sampling with bAK data x_{0000} (SNR $\rho = 10$). Both methods have $N_{\rm live} = 500$; regular nested sampling takes 1.2×10^6 likelihood calls in total, while product-space nested sampling takes 7.8×10^5 calls.

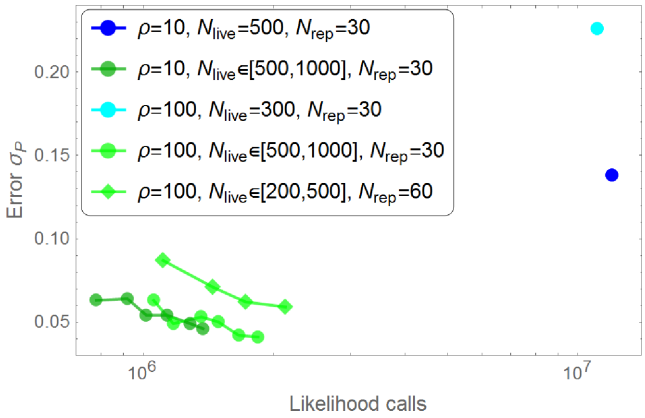

Figure 7. Error σ_P of $P_{\rm GR}^{\rm modGR}$ for GR data x_{0000} with varying SNR $\rho \in \{10, 100\}$, sampling resolution $N_{\rm live} \in [200, 1000]$, and sampling reliability $N_{\rm rep} \in \{30, 60\}$. Regular nested sampling (blue) is compared to product-space nested sampling with rethreading (green).

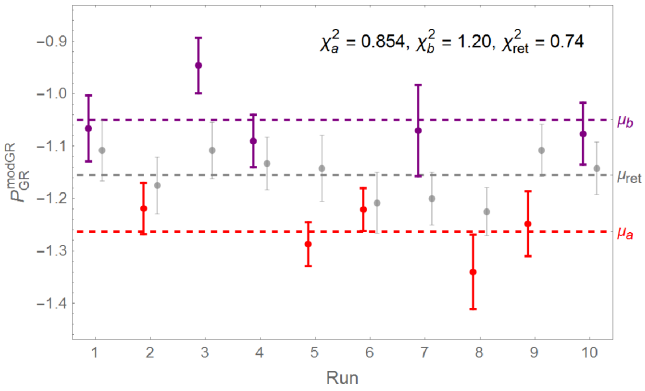

Figure 8. Reduced chi-squared test of product-space errors, for GR data x_{0000} with SNR $\rho = 100$. As the 10 original runs have inadequate sampling accuracy, they appear to be clustered around two distinct values μ_a (red) and μ_b (purple). Unravelling and rethreading them into 10 runs of equal $N_{\rm live}$ reduces the bias, and yields a single value $\mu_{\rm ret}$ (grey) as expected.

extent in Figs 3 and 6) might be better characterized with new diagnostic tests (Higson et al. 2018a) in future work.

In Fig. 8, we illustrate through a clustered chi-squared test that the obtained $P_{\rm GR}^{\rm modGR}$ and σ_P actually show a closer fit to two distinct sampling distributions with differing mean (since $\chi^2 \approx 1$ for each cluster). This analysis is not to be taken at face value, since $P_{\rm GR}^{\rm modGR}$ clearly has a unique underlying value. It does however highlight the possibility of drawing an erroneous conclusion from the null-hypothesis test due to sampling bias, especially if the runtime parameters $N_{\rm live}$ and $N_{\rm rep}$ are set too low to adequately explore high-SNR likelihoods. Furthermore, with the added complexity of the EMRI likelihood surface, it is computationally unfeasible to systematically increase these parameters without bound. We thus propose a strategy for making the product-space method robust to sampling bias; the idea is simple, and made possible by the same premise that facilitates the rethreading technique.

As discussed in Section 3.5, nested sampling permits both the unravelling and the interweaving of independent runs. Although threads from a single run might have correlations that reflect the sampling bias present in that run, such correlations are more diffuse in a set of threads from a large population of runs, and sampling bias will be reduced in runs that are randomly reconstructed from these threads. Hence the best way to utilize all of the information in the 10 high-SNR product-space runs is to unravel them into their constituent threads (5900 of them in this case) and rethread these into realizations of a run with $N_{\rm live} = 5900$, which produces the value $P_{\rm GR}^{\rm modGR} = -1.16 \pm 0.02$. The error is far smaller than that of any run in Fig. 7, which is perhaps unsurprising given the combination of information from all the runs. To demonstrate that this procedure does actually serve to decorrelate the individual threads, 10 new runs with $N_{\rm live} = 590$ are constructed through rethreading; their results (the grey points in Fig. 8) are seen to exhibit none of the sampling bias present in the original runs.

6 CONCLUSION

In this paper, we have adapted, combined, and assessed a variety of recent modelling/statistical techniques to devise a preliminary framework for testing GR with EMRI observations from future space-based GW detectors. A generalized EMRI waveform model (Gair & Yunes 2011) and its toy surrogate are trialled in a null-hypothesis test developed for LIGO sources (Li et al. 2012); the method of product-space nested sampling (Hee et al. 2015) with rethreading error estimates (Higson et al. 2017) is shown to system-

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

atically increase computational efficiency by an order of magnitude over regular evidence-based sampling.

The results and observations presented here are quite general; they are relevant not just for the outlined EMRI test of GR, but indeed any similar parametrized test that uses a generalized waveform model to describe GW sources in modified gravity (although the need to reduce computational cost is most strongly motivated for EMRIs). Product-space nested sampling with rethreading is furthermore shown to be efficient, robust and hence potentially useful for a broad range of model selection problems beyond the null-hypothesis test in this work. While much of the present analysis is exclusive to nested-sampling theory, some results (e.g. the characterization of posterior versus evidence errors) might also be applicable to other algorithms such as product-space MCMC.

Although the computational savings afforded by our proposed methods are promising, this work is only the first step in developing a practical infrastructure for testing GR with future EMRI observations. The framework should eventually incorporate other techniques for increased efficiency, such as reduced-order quadratures (Canizares et al. 2013) to accelerate individual likelihood evaluations, or dynamic nested sampling (Higson et al. 2018b) to improve sampling convergence. Finally, the actual accuracy and instructiveness of the null-hypothesis test must also be validated on data sets containing realistic source signals and detector noise, e.g. as performed by Meidam et al. (2018) for the constraints on deformation parameters in the LIGO test, but with additional focus on the final posterior odds ratio.

ACKNOWLEDGEMENTS

We thank Nicolas Yunes, Michele Vallisneri, and Stephen Taylor for useful comments on the manuscript. AJKC acknowledges support from the Jet Propulsion Laboratory (JPL) Research and Technology Development programme. SH thanks the Science and Technology Facilities Council (STFC) for financial support. CJM acknowledges financial support provided under the European Union's H2020 ERC Consolidator Grant 'Matter and strong-field gravity: New frontiers in Einstein's theory' grant agreement no. MaGRaTh646597, and networking support by the COST Action CA16104. Parts of this work were performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and funding from STFC. Parts of this work were also undertaken on the COSMOS Shared Memory system at DAMTP, University of Cambridge operated on behalf of the STFC DiRAC HPC Facility; this equipment is funded by BIS National E-infrastructure capital grant ST/J005673/1 and STFC grants ST/H008586/1, ST/K00333X/1. Parts of this work were also carried out at JPL, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

Abbott B. P. et al., 2016a, Phys. Rev. X, 6, 041015
Abbott B. P. et al., 2016b, Phys. Rev. Lett., 116, 221101
Abbott B. P. et al., 2017a, Phys. Rev. Lett., 118, 221101
Abbott B. P. et al., 2017b, Phys. Rev. Lett., 119, 141101
Abbott B. P. et al., 2017c, Phys. Rev. Lett., 119, 161101
Abbott B. P. et al., 2017d, ApJ, 851, L35
Abbott B. P. et al., 2018a, Phys. Rev. Lett. , 120, 201102
Abbott B. P. et al., 2018b, Phys. Rev. Lett., 120, 031104
Agathos M., Del Pozzo W., Li T. G. F., Van Den Broeck C., Veitch J., Vitale S., 2014, Phys. Rev. D, 89, 082001
Amaro-Seoane P. et al., 2012, Class. Quantum Gravity, 29, 124016
Amaro-Seoane P. et al., 2013, GW Notes, 6, 4
Amaro-Seoane P. et al., 2017, preprint (arXiv:1702.00786)
Apostolatos T. A., Cutler C., Sussman G. J., Thorne K. S., 1994, Phys. Rev. D, 49, 6274
Babak S. et al., 2017, Phys. Rev. D, 95, 103012
Barack L., Cutler C., 2004, Phys. Rev. D, 69, 082005
Barack L., Cutler C., 2007, Phys. Rev. D, 75, 042003
Barker B. M., O'Connell R. F., 1975, Phys. Rev. D, 12, 329
Benenti S., Francaviglia M., 1979, Gen. Relativ. Gravit., 10, 79
Canizares P., Field S. E., Gair J. R., Tiglio M., 2013, Phys. Rev. D, 87, 124005
Carlin B. P., Chib S., 1995, J. R. Stat. Soc. B, 57, 473
Chib S., 1995, J. Am. Stat. Assoc., 90, 1313
Chua A. J. K., Gair J. R., 2015, Class. Quantum Gravity, 32, 232002
Chua A. J. K., Moore C. J., Gair J. R., 2017, Phys. Rev. D, 96, 044005
Collins N. A., Hughes S. A., 2004, Phys. Rev. D, 69, 124022
Consonni G., Forster J. J., La Rocca L., 2013, Stat. Sci., 28, 398
Cutler C., 1998, Phys. Rev. D, 57, 7089
Cutler C., Flanagan E. E., 1994, Phys. Rev. D, 49, 2658
Dooley K. L., Akutsu T., Dwyer S., Puppo P., 2015, J. Phys. Conf. Ser., 610, 012012
Feroz F., Gair J. R., Hobson M. P., Porter E. K., 2009a, Class. Quantum Gravity, 26, 215003
Feroz F., Hobson M. P., Bridges M., 2009b, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2013, preprint (arXiv:1306.2144)
Gair J., Yunes N., 2011, Phys. Rev. D, 84, 064016
Gair J. R., Barack L., Creighton T., Cutler C., Larson S. L., Phinney E. S., Vallisneri M., 2004, Class. Quantum Gravity, 21, S1595
Gair J. R., Vallisneri M., Larson S. L., Baker J. G., 2013, Liv. Rev. Relativ., 16, 7
Gelman A., Meng X.-L., 1998, Stat. Sci., 13, 163
George E., Foster D. P., 2000, Biometrika, 87, 731
Glampedakis K., Babak S., 2006, Class. Quantum Gravity, 23, 4167
Godsill S. J., 2001, J. Comput. Graph. Stat. 10, 230
Green P. J., 1995, Biometrika, 82, 711
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, MNRAS, 450, L61
Handley W. J., Hobson M. P., Lasenby A. N., 2015b, MNRAS, 453, 4384
Hee S., Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 455, 2461
Higson E., Handley W., Hobson M., Lasenby A., 2017, Bayesian Anal.
Higson E., Handley W., Hobson M., Lasenby A., 2018a, preprint (arXiv:1804.06406)
Higson E., Handley W., Hobson M., Lasenby A., 2018b, preprint (arXiv:1704.03459)
Hughes S. A., 2000, Phys. Rev. D, 61, 084004
Jeffreys H., 1961, The Theory of Probability. Oxford Univ. Press, Oxford
Junker W., Schaefer G., 1992, MNRAS, 254, 146
Kass R. E., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 773
Khan S., Husa S., Hannam M., Ohme F., Pürrer M., Forteza X. J., Bohé A., 2016, Phys. Rev. D, 93, 044007
Li T. G. F. et al., 2012, Phys. Rev. D, 85, 082003
Lodewyckx T., Kim W., Lee M. D., Tuerlinckx F., Kuppens P., Wagenmakers E.-J., 2011, J. Math. Psychol., 55, 331
Lommen A. N., 2015, Rep. Prog. Phys., 78, 124901
Marin J.-M., Robert C. P., 2010, Electron. J. Stat., 4, 643
Meidam J. et al., 2018, Phys. Rev. D, 97, 044033
Meng X.-L., Wong W. H., 1996, Stat. Sin., 6, 831
Misner C. W., Thorne K. S., Wheeler J. A., 1973, Gravitation. Freeman & Co., San Francisco
Moore C. J., Gair J. R., 2015, Phys. Rev. D, 92, 024039
Moore C. J., Chua A. J. K., Gair J. R., 2017, Class. Quantum Gravity, 34, 195009
Neal R. M., 2001, Stat. Comput., 11, 125
Peters P. C., Mathews J., 1963, Phys. Rev., 131, 435

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019

Schmidt W., 2002, Class. Quantum Gravity, 19, 2743
Scott J. G., Berger J. O., 2010, Ann. Stat., 38, 2587
Sisson S. A., 2005, J. Am. Stat. Assoc., 100, 1077
Skilling J., 2004, Am. Inst. Phys. Conf. Ser., 119, 1211
Skilling J., 2006, Bayesian Anal., 1, 833
Tierney L., Kadane J. B., 1986, J. Am .Stat. Assoc., 81, 82
Veitch J., Vecchio A., 2010, Phys. Rev. D, 81, 062003
Verdinelli I., Wasserman L., 1995, J. Am. Stat. Assoc., 90, 614
Vigeland S. J., Hughes S. A., 2010, Phys. Rev. D, 81, 024030
Vigeland S., Yunes N., Stein L. C., 2011, Phys. Rev. D, 83, 104027
Villa C., Walker S., 2015, Scand. J. Stat., 42, 947
Wetzels R., Grasman R. P., Wagenmakers E.-J., 2010, Comput. Stat. Data Anal., 54, 2094
Yunes N., Pretorius F., 2009, Phys. Rev. D, 80, 122003
Yunes N., Siemens X., 2013, Liv. Rev. Relativ., 16, 9
Yunes N., Yagi K., Pretorius F., 2016, Phys. Rev. D, 94, 084002

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article-abstract/478/1/28/4989936 by University of Cambridge user on 02 January 2019