An architecture for efficient gravitational wave parameter estimation with multimodal linear surrogate models

Richard O'Shaughnessy*

Center for Computational Relativity and Gravitation,
Rochester Institute of Technology, Rochester, NY 14623, USA

Jonathan Blackman

TAPIR, Walter Burke Institute for Theoretical Physics, Mailcode 350-17,
California Institute of Technology, Pasadena, CA 91125, USA

Scott E. Field

Mathematics Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA and
Cornell Center for Astrophysics and Planetary Science,
Cornell University, Ithaca, New York 14853, USA
(Dated: June 23, 2017)

The recent direct observation of gravitational waves has further emphasized the desire for fast, low-cost, and accurate methods to infer the parameters of gravitational wave sources. Due to expense in waveform generation and data handling, the cost of evaluating the likelihood function limits the computational performance of these calculations. Building on recently developed surrogate models and a novel parameter estimation pipeline, we show how to quickly generate the likelihood function as an analytic, closed-form expression. Using a straightforward variant of a production-scale parameter estimation code, we demonstrate our method using surrogate models of effective-one-body and numerical relativity waveforms. Our study is the first time these models have been used for parameter estimation and one of the first ever parameter estimation calculations with multi-modal numerical relativity waveforms, which include all $\ell \leq 4$ modes. Our grid-free method enables rapid parameter estimation for any waveform with a suitable reduced-order model. The methods described in this paper may also find use in other data analysis studies, such as vetting coincident events or the computation of the coalescing-compact-binary detection statistic.

PACS numbers:

I. INTRODUCTION

On September 14, 2015, at 09:50:45 UTC, the Laser Interferometer Gravitational-Wave Observatory (LIGO) made the first direct observation of a gravitational-wave signal from two coalescing black hole binaries [1]. By systematically comparing the signal against approximations to the solutions of Einstein's equations, the properties of the coalescing black hole binary were inferred [2]. Similar binary black hole systems will be detected in coming years [3, 4], at a rate of up to one per day by advanced LIGO and Virgo. These discoveries will revolutionize our understanding of astrophysics (see, e.g., Refs. [5–9] and references therein) and provide tests of gravitational theory to unprecedented accuracies in the regime of strong-field dynamics with relativistic velocities [4, 10].

The vigorous pace of discovery, combined with the tantalizing opportunities afforded by low-latency and coordinated multimessenger observations [11], demand equally rapid inference: LIGO and its electromagnetic partners should prepare to reliably reconstruct the source parameters of coalescing binaries as fast as possible [12]. Es-

pecially when using the best-available waveform models, these calculations can be very costly; see, e.g., [13].

Several strategies have been developed to reduce the computational cost of parameter estimation [14–18]. Approaches that have appeared in the literature include generating the approximate solutions more quickly [19–24]; interpolating some combination of the waveform or likelihood [15, 21, 22, 25–29]; or adopting a sparse representation to reduce the computational cost of data handling [15–17, 19, 30]. Some methods, however, achieve rapid turnaround through simplifying approximations.

Two rapid strategies eschew significant approximation: reduced-order models (ROMs), a term we shall use interchangeably with surrogate models, and refactored likelihoods.

A surrogate model provides an efficient and highly accurate representation for the gravitational wave strain. Surrogate models have been applied to reproduce the radiation from complicated sources, including long duration signals [20, 26], arbitrarily many harmonic modes [20, 25], spinning binary systems [26, 31], precessing binary systems [17, 32], and neutron star models with tidal effects [33]. Moreover, as we describe in this paper, calculations that arise naturally in parameter estimation studies can be expressed in terms of simple, precomputed quantities constructed from the reduced-order representation. The result is a dramatic reduction in the

*Electronic address: oshaughn@mail.rit.edu

arXiv:1701.01137v2 [gr-qc] 22 Jun 2017

number and complexity of operations needed to carry out gravitational-wave inference.

Similarly, Pankow et al. [15] (henceforth ILE, a shorthand for "integrate likelihood over the extrinsic parameters") expressed the gravitational-wave strain using the natural basis provided by a spin-weighted spherical harmonic decomposition of the waves. By almost eliminating overhead from data handling (e.g., the cost of performing Fourier transforms and inner products needed to evaluate $\mathcal{L}$), this representation allows for rapid likelihood evaluations, enabling direct Monte Carlo integration over all "fast" variables (e.g., extrinsic parameters corresponding to the spacetime location and orientation of the binary, which leave the binary's intrinsic dynamics unchanged). The ILE framework was recently applied in [34], to directly compare GW150914 against numerical simulations of Einstein's equations, without any intermediate approximation.

In this work, we demonstrate that these two approaches can and should be naturally unified, dramatically enhancing overall performance. This combination increases the performance of ILE by removing the need for a brute-forced grid-based exploration of the intrinsic parameter space, which can be a source of error. Additionally, we present the first parameter estimation results using multimodal numerical relativity surrogates.

This paper is organized as follows. In Section II we introduce the likelihood calculation (Section II A), describe how to refactor the log-likelihood for efficient use of reduced-order gravitational-wave models (Section II B), and implement our procedure as a simple extension of an existing, production scale parameter estimation pipeline by interfacing this pipeline with low-level surrogate data-access tools (Section II C). Section III demonstrates the method using end-to-end comparisons with the traditional ILE framework. For simplicity and to facilitate illustrations and comparisons, we emphasize examples using a previously-reported and widely available nonspinning, comparable-mass effective-one-body (EOB) surrogate [20][1] provides the exact code snapshot of LAL at the time the surrogate was built. and a nonspinning numerical relativity surrogate model [25] including 77 harmonic modes up to $\ell = 8$ and trained on a mass ratio interval of $q = m_1/m_2 \in [1, 10]$, where m_1 and m_2 are the binary's component masses.

---

[1] This surrogate model, which is distributed with the `gwsurrogate` package [35], was built for the EOB model described in Ref. [36] and implemented in the routine EOBNRv2 as part of the publicly available LIGO Analysis Library (LAL) Suite. The git hash 59c12886b026c863397f191e6c2ca69ef3498616 (available, e.g., at https://github.com/lscsoft/lalsuite)

II. METHODS

A. Inference by (Monte Carlo) integration

1. Preliminaries

Given a value of the intrinsic parameters λ (eight parameters characterizing the two masses and spin vectors) and extrinsic parameters θ (four spacetime coordinates for the coalescence event; three Euler angles for the binary's orientation relative to the Earth), we can predict the response h_k of LIGO's two operational instruments, denoted as $k = \{1, 2\}$, to an impinging gravitational wave signal. Assuming a Gaussian, stationary noise model, we can evaluate the log-likelihood

$$
\begin{aligned}
\ln\mathcal{L}(\lambda, \theta) = \\
-\frac{1}{2} \sum_k \langle h_k(\lambda, \theta) - d_k | h_k(\lambda, \theta) - d_k \rangle_k - \langle d_k | d_k \rangle_k \quad (1)
\end{aligned}
$$

of LIGO's network of observatories having recorded a gravitational wave signal. Except for the overall normalization constant, and omitting calibration uncertainty, our expression (1) agrees with Eq. (1) in [2]. Here d_k is the detector data in instrument k,

$$
\langle a|b \rangle_k \equiv \int_{-\infty}^{\infty} 2df \, \frac{\tilde{a}(f)^* \tilde{b}(f)}{S_{n,k}(|f|)} \,,
$$

is a noise-weighted inner product implied by the kth-detector's noise power spectrum $S_{n,k}(f)$, $\tilde{a}(f)$ is the Fourier transform of $a(t)$, $\tilde{a}(f)^*$ denotes complex conjugation of $\tilde{a}(f)$, and f is frequency; see, e.g., [15] for more details. In practice, and as discussed in the next section, we adopt a low-frequency cutoff f_{low} such that all inner products are modified to

$$
\langle a|b \rangle_k \equiv 2 \int_{|f| > f_{\text{low}}} df \, \frac{\tilde{a}(f)^* \tilde{b}(f)}{S_{n,k}(|f|)} \,. \quad (2)
$$

A key task of any parameter estimation study is to compute the joint posterior probability of λ, θ

$$
p_{\text{post}}(\lambda, \theta) = \frac{\mathcal{L}(\lambda, \theta)p(\theta)p(\lambda)}{\int d\lambda d\theta \mathcal{L}(\lambda, \theta)p(\lambda)p(\theta)} \,, \quad (3)
$$

which follows from Bayes' theorem. Here $p(\theta)$ and $p(\lambda)$ are priors on the (independent) variables θ, λ.[2]

---

[2] For simplicity, we assume all binary black hole systems are equally likely anywhere in the universe, at any orientation relative to the detector. Future direct observations may favor a correlated distribution, including the formation of more massive black holes at larger redshift [6].

2. Fast and slow intrinsic parameters

Following [15], we partition the intrinsic parameter λ into "fast" and "slow" parameters denoted by λ_f and λ_s, respectively. In principle, this division depends entirely on the computational cost of waveform generation. "Fast" parameters are those for which new waveform evaluations can be quickly generated from existing ones as the value of λ_f changes. Typically, this is accomplished by an explicit, closed-form expression. In the original ILE study [15], the fast (slow) parameters were the extrinsic (intrinsic) parameters. By contrast, for this paper, and as we show for any other ILE-based investigation using linear surrogates that represent the scale-free general relativity solution, the only "slow" parameter is the binary system's total mass $M = m_1 + m_2$.

Having split the intrinsic parameters into a fast and slow set, we shall now view the likelihood function as $\mathcal{L}(\lambda_f, \lambda_s, \theta)$. For a fixed value of λ_s, integration over all fast parameters leads to an intermediate result[3]:

$$\mathcal{L}_{\mathrm{marg,s}}(\lambda_s) \equiv \int \mathcal{L}(\lambda, \theta)p(\theta)p(\lambda_s, \lambda_f)d\theta d\lambda_f . \quad (4)$$

Note that unlike in the original ILE framework, we have explicitly retained the prior $p(\lambda_s, \lambda_f)$ in this expression. For the ILE study [15], λ_s included all intrinsic parameters, with λ_f being empty. In that work, a function proportional to $\mathcal{L}_{\mathrm{marg,s}}(\lambda = \lambda_s)$ was evaluated on a grid; interpolated, fitted, or otherwise approximated; and hence used to generate the posterior as a function of λ_s

$$p_{\mathrm{post}}(\lambda_s) = \frac{\mathcal{L}_{\mathrm{marg,s}}(\lambda_s)}{\int d\lambda_s \mathcal{L}_{\mathrm{marg,s}}(\lambda_s)} , \quad (5)$$

which follows by integrating Eq. (3) over θ and λ_f.[4] The denominator of this quantity, the (Bayesian) evidence for our model, can be used to assess how well our model fits the data. The ILE grid-based design was intended to minimize the severe computational cost of evaluating waveforms at different values of λ_s. In Sec. II B, we show how linear surrogate models remove any need for expensive, high-dimensional grids.

If the integral appearing in Eq. (4) is performed by direct Monte Carlo integration, and this computation is repeated on a dense and uniform grid in λ_s (here, $\lambda_s = M$, the total binary mass), the posterior may be estimated using a fit-free method. ILE used this same method – henceforth denoted ILEMC – to infer posterior distributions in θ. Assuming we have N random

---

[3] In general the prior in λ_s, λ_f will not be separable: the range of allowed mass ratios will depend on total mass, for example.

[4] Note that in this expression (and in contrast to the notation in [15]), our expression for $\mathcal{L}_{\mathrm{marg}}$ includes the prior over λ, allowing us to employ the same expression for the posterior to describe the method used in this work and in [15].

samples $\{\theta_q, \lambda_{f,q}\}_{q=1}^N$ drawn from a sampling distribution $p_s(\lambda_f, \theta; \lambda_s)$ at fixed value of λ_s, then the numerical approximation $\hat{\mathcal{L}}_{\mathrm{marg,s}}(\lambda_s)$ to the true marginalized likelihood $\mathcal{L}_{\mathrm{marg,s}}(\lambda_s)$ computed by Monte Carlo integration is:

$$\hat{\mathcal{L}}_{\mathrm{marg,s}}(\lambda_s) = \frac{1}{N}\sum_{q=1}^N w_q(\lambda_s), \quad (6a)$$

$$w_q(\lambda_s) = \frac{\mathcal{L}(\lambda_s, \lambda_{f,q}, \theta_q)p(\theta_q)p(\lambda_s, \lambda_{f,q})}{p_s(\lambda_{f,q}, \theta_q; \lambda_s)} . \quad (6b)$$

We repeat this process for a uniform grid in λ_s, using the same number of samples, N, each time. As a result and in particular, we can estimate the true one-dimensional cumulative distribution $P(<x)$ by the numerical approximation $\hat{P}(<x)$:

$$P(<x) \approx \hat{P}(<x) = \frac{\sum_g \Theta(x - x_g)w_g}{\sum_g w_g} , \quad (7)$$

i.e., by a Monte Carlo integral over the interval $< x$. Here x can be any parameter in λ or θ, g indexes the Monte Carlo samples over the union of all of the values of λ_s, x_g refers to the value of parameter x for the gth sample, and $\Theta(x)$ is the Heaviside step function. Using higher-dimensional weighted density estimates (e.g., kernel density estimators or weighted histograms), we can likewise estimate the joint posterior distribution in any set of dimensions, with sufficiently dense sampling.

B. Refactored likelihood for linear surrogates

1. Surrogate-enabled ILE

A complex gravitational-wave strain

$$h(t, \vartheta, \phi; \lambda) = h_+(t, \vartheta, \phi; \lambda) - ih_\times(t, \vartheta, \phi; \lambda), \quad (8)$$

can be expressed in terms of its two fundamental polarizations h_+ and $h_\times$. Here, t denotes time, ϑ and ϕ are the polar and azimuthal angles for the direction of gravitational wave propagation away from the source. The complex gravitational-wave strain can be written in terms of spin-weighted spherical harmonics $Y_{\ell m}^{(-2)}(\vartheta, \phi)$ as

$$h(t, \vartheta, \phi; \lambda) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} \frac{D_{\mathrm{ref}}}{D}h^{\ell m}(t; \lambda)Y_{\ell m}^{(-2)}(\vartheta, \phi) , \quad (9)$$

where the sum includes all harmonic modes $h^{\ell m}(t; \boldsymbol{\lambda})$ made available by the model; where D_{ref} is a fiducial reference distance; and where D, the luminosity distance to the source, is one of the extrinsic parameters.

Following a standard ROM prescription, we assume access to a *linear* surrogate model for each harmonic mode

$$h_{\ell m}(t; \lambda) = \frac{M}{D_{\mathrm{ref}}}c_{\ell m, \alpha}(\lambda_f)W_\alpha(t/M), \quad (10)$$

associated with some fiducial distance $D_{\rm ref}$, which can be expressed as a *linear* expansion in a set of reduced basis functions W_α. Different surrogate modeling techniques prescribe different approaches for the coefficients $c_{\ell m,\alpha}$ – these details need not concern us here. Crucially, this decomposition naturally identifies the "fast" and "slow" intrinsic parameters. Since the time and total mass are coupled through the basis functions, $W_\alpha(t/M)$, we are unable to "pull" M out of the inner products (cf. Eq. (13)); computing the likelihood for new values of M will require the computation of many slow overlap integrals. Conversely, as the remaining intrinsic parameters only enter through the expansion coefficients, $c_{\ell m,\alpha}(\lambda_f)$, computing the likelihood for new values of λ_f is accomplished with fast evaluations of these coefficients (cf. Eq. (13)).

This approach requires that the coefficients are given by a known, closed-form expression and the model's temporal and parametric dependence has an affine factoriza-

tion of Eq. (10). In particular, this restriction precludes surrogate models whose temporal dependence has a *non-linear* relationship to the basis (say, by an amplitude and phase decomposition). As described later on, we do not believe this to be any real restriction in practice since all surrogates can be brought into the form (10). Indeed, the numerical relativity surrogate model used in this paper [25], expressed as two independent linear expansions of the amplitude and phase of each mode $h_{\ell m}(t,\lambda)$, was originally in a format incompatible with the representation (10).

Following Pankow et al. [15], we substitute expression (9) for $h_{\ell m}$ into the expression $h_k(t - t_k) = F_{+,k}h_+(t-t_k)+F_{\times,k}h_\times(t-t_k)$ for the detector response h_k, where $t_k = t_c - \vec{x}_k \cdot \hat{n}$ is the arrival time at the kth detector (at position $\vec{x}_k$) for a plane wave propagating along $\hat{n}$ and t_c is the time of coalescence [15]. We then substitute these expressions for h_k into the likelihood function (1) thereby generating [15]

$$
\begin{aligned}
\ln\mathcal{L}(\lambda,\theta) = &(D_{\rm ref}/D)\mathrm{Re}\sum_k\sum_{\ell m}(F_k Y_{\ell m}^{(-2)})^* Q_{k,lm}(\lambda,t_k) \\
&- \frac{(D_{\rm ref}/D)^2}{4}\sum_k\sum_{\ell m\ell'm'}\left[|F_k|^2[Y_{\ell m}^{(-2)}]^* Y_{\ell'm'}^{(-2)} U_{k,\ell m,\ell'm'}(\lambda)+\mathrm{Re}\left(F_k^2 Y_{\ell m}^{(-2)} Y_{\ell'm'}^{(-2)} V_{k,\ell m,\ell'm'}\right)\right]
\end{aligned}
\tag{11}
$$

where where $F_k = F_{+,k} - iF_{\times,k}$ are the complex-valued detector response functions of the kth detector [15] and the quantities Q, U, V depend on h and the data as

$$
Q_{k,\ell m}(\lambda,t_k) = 2\int_{|f|>f_{\rm low}}\frac{df}{S_{n,k}(|f|)}e^{2\pi i f t_k}\tilde{h}_{\ell m}^*(\lambda;f)\tilde{d}(f) ,
\tag{12a}
$$

$$
U_{k,\ell m,\ell'm'}(\lambda) = \langle h_{\ell m}|h_{\ell'm'}\rangle_k ,
\tag{12b}
$$

$$
V_{k,\ell m,\ell'm'}(\lambda) = \langle h_{\ell m}^*|h_{\ell'm'}\rangle_k .
\tag{12c}
$$

Finally, substituting (10) into (12) while fixing the value of M, we find that all three parameter-dependent functions can be expressed in terms of the surrogate interpolating functions, $c_{\ell m,\alpha}(\lambda)$, and correlations between the reduced basis functions and data:

$$
Q_{k,\ell m}(\lambda,t_k) = c_{\ell m,\alpha}(\lambda)\langle T_{t_k}W_\alpha|d_k\rangle_k ,
\tag{13a}
$$

$$
U_{k,\ell m,\ell m'}(\lambda) = c_{\ell m,\alpha}^*(\lambda)\langle W_\alpha|W_\beta\rangle_k c_{\ell'm',\beta}(\lambda),
\tag{13b}
$$

$$
V_{k,\ell m,\ell m'}(\lambda) = c_{\ell m,\alpha}(\lambda)\langle W_\alpha^*|W_\beta\rangle_k c_{\ell'm',\beta}(\lambda).
\tag{13c}
$$

Here T_τ is a time-translation operator $(T_\tau f)(t) = f(t-\tau)$. In other words, after a handful of inner product evaluations that can be computed once and for all, the likelihood can be subsequently evaluated for all intrinsic parameters λ *analytically*, as $c(\lambda)$ are known analytic expressions (made directly available by the surrogate model) and Q, U, V are easily tabulated arrays.

2. Discussion and further compression

Equation (13) needs to be evaluated for each basis function in the surrogate model. Current multimodal surrogates (including the surrogates employed here) treat each mode independently, so the total number of basis functions grows with the number of modes as well as the number of basis functions per mode. For example, in some of our parameter estimation studies we use a non-spinning NR surrogate that contains 77 modes and about 25 basis per mode, implying an enormous number of inner products (13).

The basis size can be reduced three ways. Our procedure follows a combination of the first two observations described below; we hope to explore the final approach (which requires building a new surrogate model) in future work.

First, we can eliminate superfluous modes from our expansion; for example, the $(l,m) = (8,8)$ mode is rarely practically relevant (i.e., $\langle h_{88}|h_{88}\rangle \ll \langle h|h\rangle$; cf. the caption of Fig. 1). Our ILE implementation automatically eliminates modes which are unlikely to be relevant, based on a reference set of parameters provided by the gravitational wave search.

Second, we can easily reduce the number of basis elements needed per mode. As shown in Fig. 1, surrogate errors typically converge exponentially and employ an ex-

cessively accurate basis (often with overlap errors around 10^{-10}) for parameter estimation purposes. So the basis size can be reduced by as much as a factor of 2 in our case without a significant loss in accuracy.[5] To be concrete, Eq. (13) can be expressed in terms of orthogonal basis functions, $e_\alpha(t)$, which are related to the basis functions, $W_\alpha(t)$, by a linear transformation, $W_\alpha(t) = \mathcal{V}_{\alpha\beta}e_\beta(t)$. The transformation matrix $\mathcal{V}_{\alpha\beta}$ is a necessary part of the surrogate building process [20] and is readily available for use by the ILE codes. In the orthogonal representation, the basis elements are ordered by significance; we can therefore dramatically reduce the number of basis elements needed, by adopting a basis size suitable to the comparatively lower accuracy needed for our calculations. For simplicity and modularity, however, our current implementation uses Eq. (13) directly, without additional refactoring. Instead, our current implementation can drop higher-order (orthogonal) basis functions before the computation of W_α and hence Eq. (13).

Finally, we can reduce the overall set of W_α by using another surrogate construction procedure, employing the same basis set for all the modes. This straightforward second-generation surrogate should require fewer basis coefficients, particularly since modes with the same harmonic index m share similar frequency content at early times.

3. Implementation considerations

Our approach requires the surrogate model to be expressed as a linear combination of basis elements. However, not all surrogates have this form. For example, some models are built to separately reconstruct the amplitude and phase of $h_{\ell m}$ see, e.g., [25, 26]. That said, any surrogate can be used to train a secondary "surrogate-of-a-surrogate" that has the necessary form given by Eq. (10). We have found that building a secondary surrogate is significantly easier than the original surrogate since (i) the waveform training data is already aligned [i.e., the surrogate-builder does not need to duplicate the effort needed to establish a consistent definition of the event time][6] and (ii) arbitrarily many waveform evaluations can be supplied by the primary surrogate. Moreover, several modern surrogates are already expressed in as a linear combination of basis to enable their use in
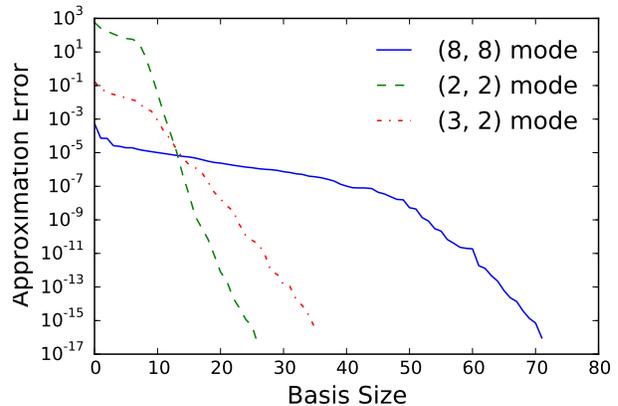
---

[5] To assess just how excessively accurate these surrogate model constructions can be, we compare the typical accuracy shown in Fig. (1) to the largest mismatch error that could introduce a significant deviation into a parameter estimation posterior: an overlap error of order $0.1/\rho^2 \simeq 10^{-5}(\rho/100)^2$. This nominal overlap error is also often significantly smaller than other systematic effects associated with intrinsic NR error due to finite resolution and extraction, as well as to effects associated with the neglect of higher-order modes.

[6] In our case, the initial surrogate was aligned so that the maximum of $\sum_{\ell m}|h_{\ell m}(t)|^2$ occurs at $t = 0$; see Eq. (2) in Ref. [25].

FIG. 1: **Convergence of the NR surrogate basis**: Convergence of the reduced basis approximation (i.e. for an optimal set of expansion coefficients $c_{\ell m,\alpha}(\lambda)$ found through orthogonal projection) of the linear NR surrogate model, $h_{S,1}^{\ell m}(t, \lambda)$, for three representative harmonic modes. The L^2-type errors, computed from the formula $\max_\lambda \|h_{S,1}^{\ell m}(\cdot;\lambda) - h_{S,2}^{\ell m}(\cdot;\lambda)\|$, are measured with respect to the original *nonlinear* surrogate, $h_{S,2}^{\ell m}(t;\lambda)$, built in Ref. [25]. These *unnormalized* errors should be compared on a relative scale. In particular, we see that, as expected, the (2,2)-mode's accuracy dominates the overall error budget. By contrast to the quadrupole mode, the (8,8) mode is relatively unimportant as evidenced by six orders of magnitude difference with the $(2,2)$ mode. (Using zero basis functions is equivalent to neglecting the mode altogether.) Our linear surrogate truncates the basis size at 22 regardless of harmonic mode index, and the ILE pipeline makes additional mode-specific basis reductions based on the truncation's impact to the overall surrogate model accuracy (see Sec. II B 2).

reduced-order-quadrature methods [17]. Our method can therefore be applied to all available surrogates without loss of generality.

As described in [20], the `gwsurrogate` package [35] provides an interface to generic surrogates for gravitational wave radiation from coalescing binaries. As part of this paper, we have extended gwsurrogate's API to allow for a convenient interface with low level surrogate waveform data as needed by the surrogate-enabled ILE pipeline. In this work we will demonstrate our method using one of the surrogates provided with it: a zero-spin, equal-mass surrogate tuned to the nonspinning effective-one-body model [36]. For each total mass M, we can extract the basis functions W_α and construct the inner products appearing in Eq. (13), once and for all. To evaluate the likelihood any mass ratio q and extrinsic parameters θ, we use the ILE likelihood [Eq. (11)], where Q, U, V are evaluated using Eq. (13) and the coefficients $c_{\ell m,\alpha}(q)$ are provided by our surrogate model.

C. Two methods to infer parameters

In Sec. III we will directly compare two ILE-type approaches to infer parameters: traditional ILE [15] and its extension developed here. While both methods have been presented for generic binary black hole systems, the parameter estimation results of Sec. III are for non-spinning binaries. And so, for concreteness, we briefly summarize these methods when specialized to such systems.

Traditional ILE. We use ILE to (i) evaluate $\mathcal{L}_{\mathrm{marg,alt}}(m_1, m_2) = \int d\theta p(\theta)\mathcal{L}(\lambda, \theta)$ by direct Monte Carlo integration, (ii) fit this function, as $\hat{\mathcal{L}}_{\mathrm{marg,alt}}$ [7], and then (iii) integrate $p_{\mathrm{post}}(\lambda) = \hat{\mathcal{L}}_{\mathrm{marg,alt}} p(\lambda)/\int d\lambda \hat{\mathcal{L}}_{\mathrm{marg,alt}} p(\lambda)$ to evaluate the posterior.

In contrast to $\mathcal{L}_{\mathrm{marg,s}}(\lambda_s)$ given by Eq. (4), which is a function only of the total binary mass $\lambda_s = M$ and which in this work is used only for normalization, the function $\mathcal{L}_{\mathrm{marg,alt}}(\lambda)$ depends on all intrinsic binary parameters λ.

We perform the integration carried out in the third step via Monte Carlo, using a uniform prior density in (m_1, m_2) such that each component mass is greater than $1M_\odot$ and the total mass is less than $200M_\odot$. The prior's boundary is defined by a right triangle with verticies at $(1M_\odot, 1M_\odot)$, $(199M_\odot, 1M_\odot)$, and $(1M_\odot, 199M_\odot)$], so in this region $p(m_1, m_2) = \frac{2}{(198M_\odot)^2}$. In an M, q coordinate system, in the region consistent with our constraints this prior has the form $p(M, q) = \frac{4}{(198M_\odot)^2}\frac{M}{(1+q)^2}$. The extra factor of 2 arises by compressing the two regions $m_1 > m_2$ and $m_1 < m_2$ into a single region in the M, q plane (i.e., by requiring $q < 1$ or $m_1 > m_2$.)

ILEMC. In the new approach described in this paper, we perform the Monte Carlo procedure described in Eq. (6) for a dense and uniform grid in total mass ($M = \lambda_s$). One-dimensional posterior distributions are found via Eq.(7).

To be concrete, in Sec. III our numerical experiment will be to (i) generate a specific list of candidate signals, (ii) prepare mock data for the expected LIGO response, and finally (iii) apply the traditional ILE and ILEMC procedures to these synthetic datasets. We assume both instruments operate at the LIGO O1 sensitivity [37]. We analyze data segments of 32s in duration sampled at a rate of $16,384$Hz.

Our candidate signals are nonspinning black hole binaries, with an inclination of 0 (in the first case) or $\pi/4$ (in the remaining cases) relative to the line of sight, and with the distance D scaled so the network signal-to-noise ratio is 25 or 20. For simplicity, all candidate signals have been

---

ID	$m_1(M_\odot)$	$m_2(M_\odot)$	ι	ρ	Model
0	35	35	0	25	EOBHM
1	100	30	$\pi/4$	20	EOBHM
2	100	50	$\pi/4$	20	EOBHM

TABLE I: **Source parameters**: This table provides the source parameters for each candidate event. All synthetic events were created as distinct realizations of Hanford and Livingston data at a GPS time 10^9s and a sky location of RA=DEC=0. Candidate signals include the $(\ell, m) = (2, \pm 1)$, $(2, \pm 2)$, $(3, \pm 3)$, $(4, \pm 4)$ and $(5, \pm 5)$ modes made available by the EOBHM model.

generated with a multimodal effective-one-body model for nonspinning binary black holes [36], henceforth denoted EOBHM. The EOB surrogate model used in this paper has been trained on EOBHM [20], but for signal injection we continue to use the original EOBHM model. The original EOBHM model is also used for parameter estimation with the traditional ILE method while the ILEMC will always use the surrogate model in its analysis. The likelihood calculation uses frequencies between $f_{\mathrm{low}} = 20\,\mathrm{Hz}$ and $2000\,\mathrm{Hz}$. A template's duration depends on the model. For EOBHM, the (2,2) mode starts at $10\,\mathrm{Hz}$, to insure $(4, 4)$ mode starts before 20 Hz. For the ROM, the entire dimensionless surrogate model is used when computing Eq. (13) so the starting frequency depends on the binary mass; for a sense of scale, at $M = 150M_\odot$ and $q = 9$, the $(4, 4)$ mode has a starting frequency of roughly $22\,\mathrm{Hz}$.

III. DEMONSTRATIONS

Figures 2 and 3 compare results obtained by traditional ILE and ILEMC applied to identical sources, summarized in Table I, using a range of candidate models that may include or omit higher order modes. Section II C provides a complete description of the demonstration's setup.

A. Single mode, non-spinning EOB surrogate model

Figure 2 shows two cumulative posterior distributions from a parameter estimation study performed on a synthetic dataset with source parameters (entry ID 0) summarized in Table I – a configuration motivated by the parameters of GW150914 [1]. The solid red curve shows the posterior distribution recovered with traditional ILE and the EOBHM model. The green curve shows the results derived from our new ILEMC approach and a surrogate trained to reproduce the $(2, 2)$ mode of EOBHM. Despite the surrogate model ignoring higher harmonic mode content we expect these two methods to produce nearly-identical posteriors. Indeed, this choice of inclination angle and mass ratio is well known to minimize the importance of higher modes. For nearly equal-mass

binaries at this relatively low total mass, previous studies have shown higher harmonics have negligible impact on parameter estimation; see, e.g., [34, 38–40].

As expected, the red (traditional ILE) and green (ILEMC) curves nearly agree. The discrete steps apparent in the ILEMC posterior (green) arise because of the Monte Carlo procedure used to estimate the cumulative; by contrast, the smooth results produced by traditional ILE follow from applying a simple approximation to the function $\mathcal{L}_{\mathrm{marg,alt}}$. In [34], this approximation was a second-order Taylor series approximation near the peak; in [15], this approximation was an interpolating function based on a discrete sampling grid.

Here, we have compared parameter estimation (PE) with and without surrogates, using a surrogate tuned to the same model. Models like EOB have parameters which have been calibrated against numerical relativity in a certain region and with a certain accuracy. They have systematic errors, which grow increasingly significant when moving away from the region where they have been best-calibrated. For example, even for these binary parameters – deep inside a thoroughly-explored region of parameter space – we do not see this level of agreement when carrying out a similar analysis using other EOB models [41].

B. Multi-mode, non-spinning EOB and NR

In Figure 3 we show results for parameter estimation via ILE with EOBHM (solid red); via ILE with only the leading-order mode from EOBHM (dotted red); via ILEMC, with a surrogate which reproduces the $(2,2)$ mode of numerical relativity simulations [25] (dotted green); via ILE, using the same surrogate model (dotted black); and with the same surrogate but including higher-order modes, via ILE (solid black). The top panels show a result with $q = 0.3$; the bottom panels show results for a binary with $q = 0.5$. By design, both scenarios are comparable to cases examined in [38] with a different parameter estimation engine (e.g., their Fig 9) using a similar model, albeit at higher SNR.

First and foremost, the bottom two panels show that, as with Figure 2, ILEMC and ILE agree when the models do. We compare calculations performed using the same NR surrogate model, using ILE (black) and ILEMC (green) for parameter inference parameter inference. For example, in every panel the dotted black and green lines – two independent parameter inference methods using the $(2,\pm2)$ modes of the NR surrogate – are in good agreement with one another.

Second, the bottom right panel shows by example that at this binary mass higher harmonics matter, since the dotted and solid lines do not agree. As illustrated clearly by directly comparable prior work like [38], higher harmonics break degeneracies and improve parameter estimation accuracy. The good agreement in the bottom panels of Figure 3 persists despite employing a com-
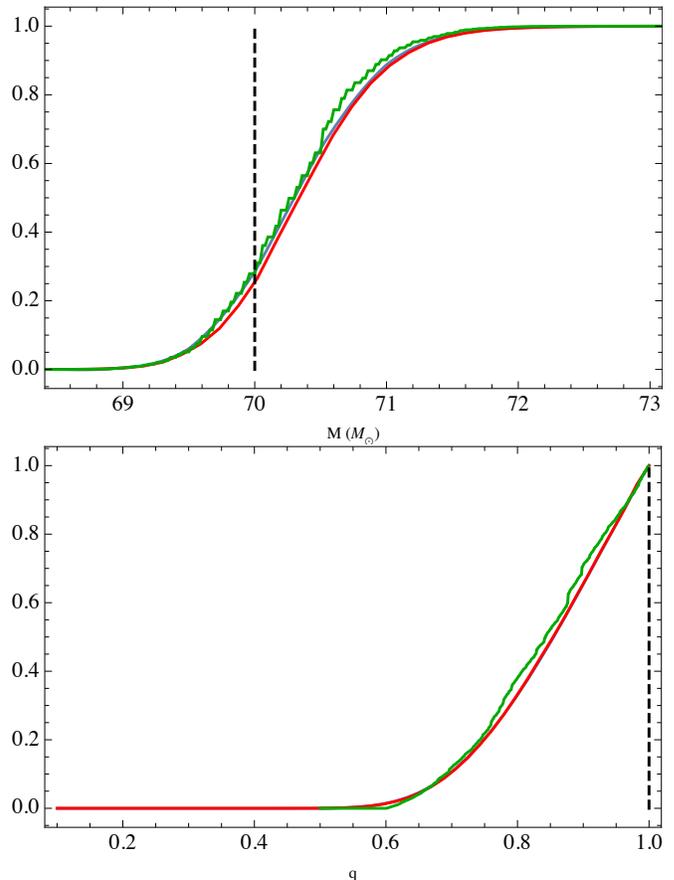

FIG. 2: **Comparison of traditional ILE and ILEMC pipelines using a GW150914-like signal.** The one-dimensional cumulative posterior distributions for total mass M (top) and mass ratio $q = m_2/m_1$ (bottom) have been derived using both the traditional ILE (thin blue or red curve) and ILEMC (green curve) frameworks. The binary parameters for this single analysis, summarized in Table I (entry "ID 0"), are depicted as a dashed vertical line. Parameter estimation with traditional ILE was performed using the EOBHM model (red and blue) while the ILEMC study uses an EOBHM-calibrated surrogate (green). To illustrate the consistency between different fitting methods used for the traditional ILE results, we have considered evaluation of the posterior with a quadratic fit (blue) and a Gaussian process (red).

pletely different model for the source and template.

Finally, in both the top and bottom panels of Figure 3, we see generally good agreement between the red curves (PE with ILE, carried out using EOBHM) and the black curves (PE with ILE, carried out using our NR surrogate). Despite systematic differences between the NR surrogate and EOBHM, in most cases this agreement persists even when higher modes are omitted (dotted lines) or used (solid lines).
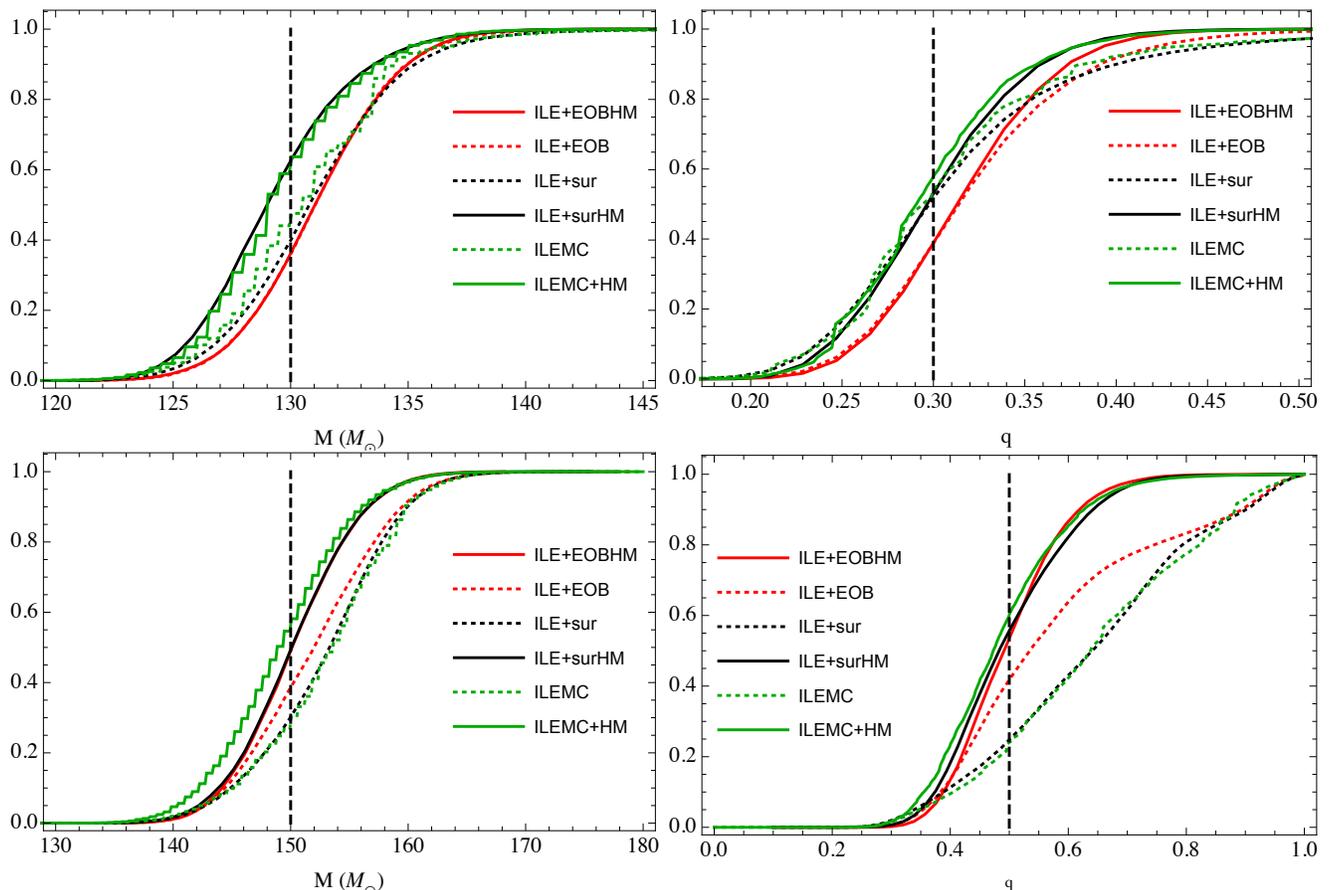
FIG. 3: **Comparison of traditional ILE and ILEMC pipelines using unequal masses and higher harmonics**: We consider gravitational wave signals for a source binary with $m_1 = 100M_\odot$, $m_2 = 30M_\odot$ (top panels) or $m_1 = 100M_\odot$ and $m_2 = 50M_\odot$ (bottom panels). A dashed vertical line depicts these injection parameter values which are also listed in Table I. Dashed lines indicate results derived using only the $(2,2)$ mode; solid lines indicate results using modes with $\ell \le 4$. The red lines indicate PE performed with a nonspinning EOB model and ILE; the green (ILEMC) and black (ILE) lines indicate PE performed with the same surrogate tuned to nonspinning NR simulations [25] but using two different methods.

IV.   CONCLUSIONS

In this paper, we unified two strategies to accelerate parameter estimation – surrogate models and factored likelihoods – and implemented the result in a production-scale environment, ready-to-use on real LIGO data. Our code can be used with any (time-domain) linear surrogate model, leveraging parallel efforts to better model the multimodal gravitational wave signal from coalescing binaries [15, 19, 20, 25].

With our existing implementation, we can rapidly reconstruct parameters of arbitrary sources whose runtime is mostly limited by the cost of a low-dimensional adaptive Monte Carlo integral. Based on operation counts, we project parameter estimates could be carried out in seconds to minutes. When operating at its theoretical limit, this approach can conceivably provide real-time parameter estimation and Bayesian evidence factors [42–44]. These rapid calculations will be helpful to improve current detection procedures or to supplement investiga-

tions into the impact of non-Gaussian noise ("glitches") on GW parameter estimation.

Finally, as an illustration of our method's broad utility, we have demonstrated how waveform modeling errors can, for sufficiently massive systems ($f_{low}M > 0.5 \times 10^{-2}$ in dimensionless units), cause us to draw less sharp and reliable conclusions about the progenitor's properties. Our method can make direct use of high-fidelity surrogate models trained on numerical relativity waveform data without any approximations to general relativity and including all harmonic modes resolved by the simulation. For heavy black holes in particular, where systematic biases are expected to be the most extreme [45–47], we demonstrate by example that modeling error such as neglecting higher harmonic modes can impact our interpretation of candidate events.

V. ACKNOWLEDGMENTS

[1] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Phys. Rev. Lett. **16**, 061102 (2016).

[2] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Phys. Rev. Lett. **116**, 241102 (2016), URL http://link.aps.org/doi/10.1103/PhysRevLett.116.241102.

[3] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), ApJL **833**, 1 (2016), 1602.03842, URL https://dcc.ligo.org/LIGO-P1500217/public/main.

[4] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), PRX **6**, 041015 (2016).

[5] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), ApJL **818**, L22 (2016), 1602.03846, URL https://dcc.ligo.org/LIGO-P1500262/public/main.

[6] K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy, Nature (London) **534**, 512 (2016), 1602.04531.

[7] C. L. Rodriguez, C.-J. Haster, S. Chatterjee, V. Kalogera, and F. A. Rasio, ApJL **824**, L8 (2016), 1604.04254.

[8] S. Bird, I. Cholis, J. B. Muñoz, Y. Ali-Haïmoud, M. Kamionkowski, E. D. Kovetz, A. Raccanelli, and A. G. Riess, Physical Review Letters **116**, 201301 (2016), 1603.00464.

[9] S. E. de Mink and I. Mandel, MNRAS **460**, 3545 (2016), 1603.02291.

[10] B. P. Abbott et al. (LIGO Scientific Collaboration, Virgo Collaboration), Phys. Rev. Lett. **116**, 221101 (2016), 1602.03841.

[11] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, et al., Living Reviews in Relativity **19** (2016), 1304.0670.

[12] The LIGO Scientific Collaboration and the Virgo Collaboration (2015), URL https://dcc.ligo.org/LIGO-T1500055/public.

[13] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), PRX **6**, 041014 (2016), URL http://adsabs.harvard.edu/abs/2016arXiv160601210T.

[14] B. Miller, R. O'Shaughnessy, B. Farr, and T. Littenberg, Phys. Rev. D **92**, 4056 (2015), URL http://xxx.lanl.gov/abs/arXiv:1506.06032.

[15] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy, Phys. Rev. D **92**, 023002 (2015), URL http://adsabs.harvard.edu/abs/2015PhRvD.92b3002P.

[16] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015).

[17] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016), 1604.08253.

[18] S. Vinciguerra, J. Veitch, and I. Mandel, Classical and Quantum Gravity **34**, 115006 (2017), URL http://stacks.iop.org/0264-9381/34/i=11/a=115006.

[19] P. Canizares, S. E. Field, J. R. Gair, and M. Tiglio, Phys. Rev. D **87**, 124005 (2013), URL http://xxx.lanl.gov/abs/arXiv:1304.0462.

[20] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Physical Review X **4**, 031006 (2014).

[21] R. J. E. Smith, K. Cannon, C. Hanna, D. Keppel, and I. Mandel, Phys. Rev. D **87**, 122002 (2013).

[22] K. Cannon, J. D. Emberson, C. Hanna, D. Keppel, and H. P. Pfeiffer, Phys. Rev. D **87**, 044008 (2013).

[23] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).

[24] A. Lundgren and R. O'Shaughnessy, Phys. Rev. D **89**, 044021 (2014), URL http://link.aps.org/doi/10.1103/PhysRevD.89.044021.

[25] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Phys. Rev. Lett. **115**, 121102 (2015), URL http://link.aps.org/doi/10.1103/PhysRevLett.115.121102.

[26] M. Pürrer, Classical and Quantum Gravity **31**, 195010 (2014).

[27] R. J. E. Smith, C. Hanna, I. Mandel, and A. Vecchio, Phys. Rev. D **90**, 044074 (2014).

[28] R. H. Cole and J. R. Gair, Physical Review D **90**, 124043 (2014).

[29] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby, MNRAS **421**, 169 (2012), 1110.2997.

[30] H. Antil, S. E. Field, F. Herrmann, R. H. Nochetto, and M. Tiglio, Journal of Scientific Computing **57**, 604 (2013).

[31] M. Pürrer, Physical Review D **93**, 064041 (2016).

[32] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, D. A. Hemberger, P. Schmidt, and R. Smith, Phys. Rev. D **95**, 104023 (2017), URL https://link.aps.org/doi/10.1103/PhysRevD.95.104023.

[33] B. D. Lackey, S. Bernuzzi, C. R. Galley, J. Meidam, and C. V. D. Broeck, arXiv preprint arXiv:1610.04742 (2016).

[34] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Phys. Rev. D **94**, 064035 (2016), URL http://link.aps.org/doi/10.1103/PhysRevD.94.064035.

[35] *Gwsurrogate*, https://pypi.python.org/pypi/gwsurrogate/.

[36] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, Physical Review D **84**, 124052 (2011).

[37] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Phys. Rev. D **93**, 112004

(2016), 1604.00439, URL http://link.aps.org/doi/10.1103/PhysRevD.93.112004.

[38] P. B. Graff, A. Buonanno, and B. S. Sathyaprakash, Phys. Rev. D **92**, 022002 (2015), 1504.04766.

[39] B. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Available as arxiv:1611.07531 (2016), URL https://arxiv.org/abs/1611.07531.

[40] J. Lange (2016), master's thesis submitted to the Rochester Institute of Technology, available as LIGO-P1600281 at https://dcc.ligo.org/LIGO-P1600281.

[41] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **86**, 024011 (2012).

[42] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **80**, 063007 (2009), 0902.0368.

[43] R. Lynch, S. Vitale, R. Essick, E. Katsavounidis, and F. Robinet, ArXiv e-prints (2015), 1511.05955.

[44] J. B. Kanner, T. B. Littenberg, N. Cornish, M. Millhouse, E. Xhakaj, F. Salemi, M. Drago, G. Vedovato, and S. Klimenko, Phys. Rev. D **93**, 022002 (2016), 1509.06423.

[45] V. Varma, P. Ajith, S. Husa, J. C. Bustillo, M. Hannam, and M. Pürrer, Phys. Rev. D **90**, 124004 (2014), 1409.2349.

[46] V. Varma and A. Parameswaran, In preparation (LIGO P1600332) (2016).

[47] J. Calderón Bustillo, S. Husa, A. M. Sintes, and M. Pürrer, Phys. Rev. D **93**, 084019 (2016), 1511.02060.