Tempered Multifidelity Importance Sampling for Gravitational Wave Parameter Estimation

Bassel Saleh,[1,*] Aaron Zimmerman,[2,†] Peng Chen,[3,‡] and Omar Ghattas[1,§]

[1] *Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712, USA*
[2] *Center for Gravitational Physics, The University of Texas at Austin, Austin, TX 78712, USA*
[3] *College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA*

(Dated: April 30, 2025)

Estimating the parameters of compact binaries which coalesce and produce gravitational waves is a challenging Bayesian inverse problem. Gravitational-wave parameter estimation lies within the class of multifidelity problems, where a variety of models with differing assumptions, levels of fidelity, and computational cost are available for use in inference. In an effort to accelerate the solution of a Bayesian inverse problem, cheaper surrogates for the best models may be used to reduce the cost of likelihood evaluations when sampling the posterior. Importance sampling can then be used to reweight these samples to represent the true target posterior, incurring a reduction in the effective sample size. In cases when the problem is high dimensional, or when the surrogate model produces a poor approximation of the true posterior, this reduction in effective samples can be dramatic and render multifidelity importance sampling ineffective. We propose a novel method of tempered multifidelity importance sampling in order to remedy this issue. With this method the biasing distribution produced by the low-fidelity model is tempered, changing the chi-squared divergence between the two distributions and thereby affecting the efficiency of importance sampling. There is an optimal temperature which maximizes the efficiency in this setting, and we propose a low-cost strategy for approximating this optimal temperature using samples from the untempered distribution. In this paper, we motivate this method by applying it to Gaussian target and biasing distributions. Finally, we apply it to a series of problems in gravitational wave parameter estimation and demonstrate improved efficiencies when applying the method to real gravitational wave detections.

I. INTRODUCTION

The direct detection of gravitational waves (GWs) [1–28] provides an unprecedented viewpoint on the most compact objects in the Universe. Observations by the Advanced LIGO [29], Advanced Virgo [30], and KAGRA [31] detectors reveal the properties of the black holes and neutron stars which emit GWs as they inspiral and coalesce. Following detection of a GW event, the next step in applying GW data to a host of problems in fundamental physics and astrophysics is to measure the properties of the binary system that produced the signal from the noisy data. The standard method for solving this inverse problem is through Bayesian inference, see e.g. [32–35].

In GW data analysis, parameter estimation presents a number of challenges. The parameters of the binary system such as the masses and spins of components and their location in the sky are highly correlated, and must be inferred with a high degree of accuracy for precision applications. Further, the GW models needed for inference can be computationally expensive [36]. Full numerical simulations of the binary evolution provide the highest fidelity predictions, but are intractably expensive for use in most algorithms. Large catalogs of simulations have been used together with likelihood interpolation and marginalization to infer the properties of GW sources, see e.g. [37]. More commonly, numerical simulations are used together with analytical approximates to build a wide hierarchy of surrogate models, with different underlying approximations, fidelity, and speed. They include phenomenological models [38–45], Effective One Body models [46–53], and surrogate models built directly on numerical simulations [36, 54–58] Cutting-edge models require fractions to tens of seconds to evaluate, see e.g. [42, 53] despite attention paid to computational efficiency. See also Ref. [59] for a recent overview of GW models as well as methods for accelerating their evaluation. The challenges of GW parameter estimation are compounded by the need of a large number of model evaluations due to the relatively low efficiency of sampling methods [33]. As the rate of GW detections rapidly increases with increasing detector sensitivity, it is important to explore methods of accelerating GW inference.

In this work we tackle the problem of accelerating GW inference from the standpoint of multifidelity methods. The goal of using a multifidelity framework is to exploit the computational speed of low-fidelity models while retaining the accuracy of a high-fidelity model, in order to achieve fast, accurate solutions to many-query problems such as Bayesian inference [60]. GW models provide a variety of options for multifidelty approaches, since there are nested models in the sense that one model may be the limit of another in the case where some physical ef-

* bassel@utexas.edu
† aaron.zimmerman@utexas.edu
‡ pchen402@gatech.edu
§ omar@oden.utexas.edu

fect is neglected. In this work we focus on the inclusion of higher modes (higher than quadrupolar radiative multipole moments) giving rise to our multifidelity hierarchy. Accounting for these higher modes of emission can break degeneracies and improve the measurement of several parameters, especially the ratio of the masses and the inclination of the orbital plane to the line of sight [10, 11, 61–64]. Meanwhile, the evaluation of the higher mode contributions increases the computational cost of the forward model roughly in proportion to the number of modes used. The expected differences between inferences with and without higher modes and their difference in computational expense make them a promising target for tempered importance sampling.

Specifically, we use the IMRPhenomXPHM [42] family of waveform models for both the low and high fidelity signal model. We treat the model as high fidelity when all of the available higher order modes are turned on, whereas for low fidelity runs, only the leading order mode is active.

One strategy to benefit from a multifidelity paradigm is through multifidelity importance sampling (MFIS) [65], in which one samples from a low-fidelity posterior and reweights the samples using high-fidelity evaluations to obtain representative samples of the high-fidelity posterior, which can be used to compute Monte Carlo integrals. Importance sampling has been explored GW inference [66], and has been applied successfully as one method to marginalize over detector calibration uncertainties [15, 16, 67], to search for signatures of binary eccentricity [35], and to improve the quality of samples from machine-learning based inference [68]. However, these applications are often limited in practice by the efficiency of importance sampling. If the two posteriors are too different from each other, the low efficiency of importance sampling means you need a huge number of low-fidelity samples to obtain desired accuracies [69]. More precisely, the effective sample size N_{eff} scales inversely with $\chi^2(p||q)$, the chi-squared divergence between the target and biasing probability distributions p and q [70].

In this work we propose a method, tempered multifidelity importance sampling, to improve the efficiency of importance sampling and broaden its domain of application. A typical problem faced when deploying importance sampling is that the chi-squared divergence is very sensitive to the support of the two distributions in the tails. If the support of the biasing distribution does not cover the target well enough in the tails, $\chi^2(p||q)$ will be large and the effective sample size will thus be small. By tempering the biasing distribution, i.e. raising the density function to some power smaller than one, the coverage expands and thus the overlap in the tails might improve as a result. The change in overall coverage may be modest, but even a small change in the tails can lead to dramatic improvements in the efficiency.

We note that an arbitrary temperature is not expected to improve the efficiency, since over-broadening the biasing distribution can have a negative effect. Further, the relationship between coverage and improved efficiency is complicated. We emphasize that while the notion of improving the coverage of the target distribution is a motivation for our exploration of tempering in the MFIS setting, ultimately our metric of success is the improvement of the efficiency. In the end, applying tempering adds an additional degree of freedom, the temperature T, which can be tuned to increase N_{eff}. In this work we develop a principled procedure for selecting a good temperature, and we find that in several of our experiments this temperature results in appreciable improvements to the efficiency.

Although we motivate the idea of tempered multifidelity importance sampling with the challenge of GW parameter estimation, our results have broad applicability. MFIS can be useful in any problem for which many model evaluations are required and a hierarchy of varying fidelities exists. These are referred to as *outer-loop applications* and include optimization, uncertainty propagation, data assimilation, control, and sensitivity analysis [60].

In the following sections, we present theoretical results demonstrating the impact of tempering on importance sampling efficiency and by extension on the Monte Carlo error in estimators using tempered samples. We argue that for arbitrary Gaussian probability densities p and q, there exists a unique temperature which minimizes $\chi^2(p||q)$, and we provide a closed form expression for this optimal temperature. We derive a practical approximation for the optimal temperature for arbitrary distributions p and q, under modest assumptions, and we validate this approximation in the Gaussian setting where it can be compared to the true solution. Finally, we apply our tempered multifidelity importance sampling algorithm to the Bayesian inference of both simulated GW observations and real GW observations. We discuss the unique challenges of this problem and propose some directions for future work.

II. METHODOLOGY

A. Bayesian Inference

In the context of an inverse problem, we start by considering the problem of finding some parameters of interest θ, given some noisy measurements of an observable quantity related to those parameters through a forward model $h(\theta)$. Here h is a vector of model predictions, i.e. a time series of observed strain values for the GW case. We assume that our observed data vector d is related to θ through

$$d = h(\theta) + n \tag{1}$$

where n is a noise term which we assume to be additive in this manner.

In many settings, including standard GW analysis, it is reasonable to assume the noise is stationary and Gaus-

sian, with zero mean and a known covariance Σ_n. In this case, $n \sim \mathcal{N}(0, \Sigma_n)$ and has a probability density p_n given by

$$p_n(n) = \frac{1}{\sqrt{|2\pi\Sigma_n|}} \exp\left(-\frac{1}{2} n^T \Sigma_n^{-1} n\right). \tag{2}$$

Substituting $n = d - h(\theta)$ yields the *likelihood function* $\mathcal{L}(d|\theta) = p_n(d - h(\theta))$, representing the probability of observing the data y given the parameter θ. We further assume a *prior probability density* $\pi(\theta)$ on the parameters.

Bayes' rule allows us to relate the likelihood and prior probability densities to the *posterior* density, which represents our knowledge of the parameter conditioned on a particular observation y. The posterior density is given by

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{Z}. \tag{3}$$

The normalization constant Z is called the *evidence*, given by

$$Z = \int \mathcal{L}(d|\theta)\pi(\theta) d\theta\,. \tag{4}$$

With the zero-mean Gaussian noise assumption and our expression for the likelihood, we have

$$\begin{aligned} p(\theta|d) &\propto \mathcal{L}(d|\theta)\pi(\theta) \\ &\propto \exp\left[-\frac{1}{2}(d - h(\theta))^T \Sigma_n^{-1} (d - h(\theta))\right] \pi(\theta)\,. \end{aligned} \tag{5}$$

In GW data analysis the evaluation of the likelihood is typically carried out in the frequency domain, where the covariance matrix of stationary Gaussian noise is diagonal. In terms of the one-sided power spectral density (PSD) $S_n(f)$, the Fourier transform of the model $\tilde{h}(\theta)$, and the frequency-domain data $\tilde{d}$, the standard GW likelihood for data taken by a single detector is (e.g. [32, 33])

$$\mathcal{L}(d|\theta) \propto \exp\left[-2\,\mathrm{Re}\sum_k \frac{|\tilde{d}_k - \tilde{h}(\theta; f_k)|^2}{S_n(f_k)} \Delta f\right], \tag{6}$$

where the sum is carried out at the discrete positive frequencies f_k and Δf is the frequency spacing. In the case of multiple detectors observing the same GW event, the noise in each detector is assumed to be independent and so the individual detector likelihoods multiply, using the same model parameters θ for each detector while accounting for the light-travel time between the detectors, as well as the different signal response functions and noise spectra.

It is important to recognize that the posterior density depends both on the observed data *and* the model $h(\theta)$, precisely through the dependence of the likelihood on $h(\theta)$. If one were to use a different model, the solution of the Bayesian inverse problem would be a different posterior probability density.

Thus, with data and model in hand, the task of Bayesian inference is to characterize the posterior given by Eq. (5). In many settings, the most useful thing to seek is a set of independent, identically distributed (i.i.d.) samples distributed according to the posterior. Once obtained, these samples can be used to compute Monte Carlo estimations of quantities that depend on the uncertain parameters, i.e.

$$\mathbb{E}_p[f(\theta)] = \int f(\theta)\, p(\theta|d)\, d\theta \;\approx\; \frac{1}{N}\sum_{i=1}^{N} f(\theta_i), \tag{7}$$

$$\{\theta_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} p(\theta|d)\,. \tag{8}$$

The principal computational challenge of Bayesian inference is the need to evaluate the potentially expensive model $h(\theta)$ a large number of times in the process of generating samples.

B. Multifidelity Importance Sampling

Sampling from a high-dimensional posterior distribution is, in general, a challenging task. The challenge of sampling from a posterior $p(\theta|d)$ is made worse when the model h is computationally expensive to evaluate. Importance sampling is one approach to exploit the speedup afforded by approximate models without sacrificing the accuracy of more sophisticated and expensive models. At its core, importance sampling involves estimating statistics of one distribution using samples drawn from another. To simplify notation, we let p and q refer to the high and low fidelity posterior distributions, respectively, as well as their corresponding probability density functions. Consider the mean of a function $f(\theta)$ with respect to the high fidelity posterior,

$$\mu := \mathbb{E}_p[f(\theta)] = \int f(\theta)\, p(\theta)\, d\theta. \tag{9}$$

We can introduce the low fidelity posterior by multiplying the integrand of Eq. (9) by unity to obtain

$$\mu = \int f(\theta)\, \frac{p(\theta)}{q(\theta)}\, q(\theta)\, d\theta = \mathbb{E}_q\left[f(\theta)\, w(\theta)\right], \tag{10}$$

$$w(\theta) := \frac{p(\theta)}{q(\theta)}\,. \tag{11}$$

In general, we call q the *biasing distribution*. Note that we have rewritten the mean μ as an expectation with respect to the biasing probability density q. This allows us to define the importance sampling Monte Carlo estimator,

$$\hat{\mu} := \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\, w(\theta_i), \qquad \{\theta_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} q, \tag{12}$$

where now the N i.i.d. samples θ_i are drawn from the biasing distribution as opposed to the original posterior distribution.

In practice, the evidence Z that appears in Eq. (3) is often ignored, and we only have evaluations of the posterior densities up to a constant. Thus, we cannot compute $w(\theta)$ explicitly. Instead we use *self-normalized* importance sampling [71], where we define our estimate of μ as

$$\tilde{\mu} := \frac{\sum_{i=1}^{N} f(\theta_i)\, w(\theta_i)}{\sum_{i=1}^{N} w(\theta_i)}. \tag{13}$$

Here, since the normalizing constants implicit in the weights cancel out, we can safely ignore them when computing $w(\theta)$.

Following the analysis in [69, 70], we see that the expected error of this self-normalized estimator is bounded by the chi-squared divergence between p and q, $\chi^2(p||q)$, given by

$$\chi^2(p||q) = \int \frac{p(\theta)^2}{q(\theta)}\, d\theta - 1\,. \tag{14}$$

Specifically, for a bounded measurable function f, the mean-squared error of the estimate in Eq. (13) is bounded by

$$\mathbb{E}\left[(\tilde{\mu} - \mathbb{E}_p[f(\theta)])^2\right] \leq 4||f||^2_{L^\infty} \frac{\chi^2(p||q)+1}{N}\,, \tag{15}$$

where $||f||_{L^\infty}$ is the L^∞ norm of f. This motivates us to search for biasing distributions that are as similar as possible to the target distribution, in the χ^2 sense.

To assess the efficiency of importance sampling, it is not usually practical to compute or approximate the χ^2 divergence directly. Instead we can compute the *effective sample size* N_{eff},

$$N_{\text{eff}} = \frac{\left[\sum_{i=1}^{N} w(\theta_i)\right]^2}{\sum_{i=1}^{N} w(\theta_i)^2}\,. \tag{16}$$

In the limit of a large number of samples θ_i drawn from the biasing distribution q, we can see that

$$\frac{N_{\text{eff}}}{N} \to \varepsilon := \frac{1}{\chi^2(p||q)+1}\,, \tag{17}$$

so that minimizing χ^2 has the effect of maximizing N_{eff}. Here we have defined the idealized efficiency ε given χ^2.

In practice we have access to only a limited number of choices for a reasonably accurate biasing distribution q. For example, in GW data analysis q may be defined by a signal model which is computationally cheaper to evaluate than the model implicit in the desired posterior p. The development of accurate and computationally efficient signal models is itself a major challenge requiring significant effort. To improve the efficiency of importance sampling without the flexibility to tune the biasing distribution through modeling, we turn to the idea of tempering q.

C. Tempering

In the setting of multifidelity importance sampling as outlined above, we propose introducing the *tempered* biasing distribution, such that its density is

$$q_T(\theta) := \frac{q(\theta)^{1/T}}{Z_T}\,. \tag{18}$$

Tempering is commonly used in Markov Chain Monte Carlo approaches to sample distributions through parallel tempering, e.g. [32, 72–75]. Here our aim is to apply the concept of tempering to improve importance sampling.

By raising the density to the power $1/T$, T acts as a "temperature" controlling the width of the distribution. This additional free parameter allows us to improve the efficiency of importance sampling without changing the underlying density q. A new constant Z_T is needed to ensure the tempered density is normalized and is therefore

$$Z_T = \int q(\theta)^{1/T}\, d\theta\,. \tag{19}$$

To gain an intuition into how tempering improves the efficiency of importance sampling, we turn to a simple toy model. Let p and q be one-dimensional Gaussians, and without loss of generality let p be zero mean. With $p = \mathcal{N}(0, \sigma_p)$ and $q = \mathcal{N}(\mu, \sigma_q)$, the chi-squared divergence is

$$\chi^2(p||q_T) = \frac{T\sigma_q^2}{\sigma_p\sqrt{2T\sigma_q^2 - \sigma_p^2}} \exp\left(\frac{\mu^2}{2T\sigma_q^2 - \sigma_p^2}\right) - 1\,. \tag{20}$$

In this case, we can maximize the efficiency ε by minimizing χ^2 analytically. The temperature that minimizes this expression, which we call T_*, is

$$T_* = \frac{3\sigma_p^2 + 2\mu^2 + \sqrt{\sigma_p^4 + 12\sigma_p^2\mu^2 + 4\mu^4}}{4\sigma_q^2}\,. \tag{21}$$

We note that for this example χ^2 diverges when $\sigma_q\sqrt{T}/\sigma_p < 1/\sqrt{2}$, and we expect in these cases that q_T is too narrow to ever captures the tails of p; one benefit of tempering is that this can be evaded by increasing T. Numerical experiments show that in practice when q is too narrow, N_{eff}/N decreases towards zero with increasing N, although this appears to be a slow and highly stochastic process.

We illustrate this simple example by plotting the idealized efficiency ε in Fig. 1 for the one-dimensional Gaussians, where by (17) and (20),

$$\varepsilon = \frac{\sqrt{2T(\sigma_q^2/\sigma_p^2) - 1}}{T(\sigma_q^2/\sigma_p^2)} \exp\left(-\frac{\mu^2/\sigma_p^2}{2T(\sigma_q^2/\sigma_p^2) - 1}\right)\,. \tag{22}$$

The efficiency depends only on two parameters, the normalized bias μ/σ_p between the distributions and the ratio

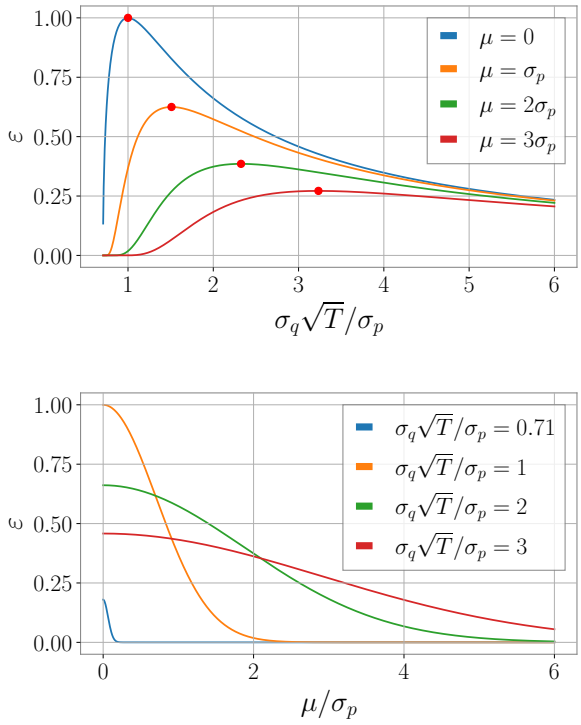

FIG. 1. Example case of the idealized efficiency in terms of the χ^2-divergence between target p and tempered q_T in the one-dimensional Gaussian case. *Top:* The efficiency plotted against the ratio of widths rescaled by temperature, for various fixed relative biases. There is a unique maximum efficiency, given by using the optimal temperature of Eq. (21). *Bottom:* The efficiency plotted against the relative bias between the Gaussians, for various fixed relative widths and temperatures. The results are proportional to a Gaussian in μ/σ_p in each case.

of the width of p to the tempered width of q_T, specifically $\sigma_q\sqrt{T}/\sigma_p$. The dependence on the normalized bias is straightforward to understand: ε is a Gaussian in μ/σ_p with variance $T(\sigma_q^2/\sigma_p^2) - 1/2$. Meanwhile, we see that for $T = 1$ and $\mu = 1$, the efficiency becomes very poor for $\sigma_q/\sigma_p < 1$, and also becomes poor as σ_q/σ_p becomes large. As the normalized bias increases, a wider biasing distribution gives better efficiency, as expected since otherwise the tails of q cannot cover the bulk of the target distribution p. We see that for each value of the bias, tempering allows us to tune the width of q_T to attain an optimal efficiency.

To map this example onto inference for GW data analysis, consider the case where p and q result from two different GW signal models. In the limit of high signal-to-noise (SNR), the widths of the posteriors scale inversely with SNR. Thus while the bias μ and the ratio of the widths σ_q/σ_p to be fixed by the differences in the signal models approximately independently of SNR, as the SNR increases the normaized bias μ/σ_p grows large. This situation is one where the efficiency is expected to be to the left of the peak of each curve in the top panel of Fig. 1, and as SNR increases we traverse these curves towards larger μ/σ_p moving vertically down, with a severe loss of efficiency. We illustrate this in the bottom panel of Fig. 1, where we plot the efficiency for fixed $\sigma_q\sqrt{T}/\sigma_p$ versus μ/σ_p. If we imagine fixing $T = 1$, then the efficiency decays monotonically with increasing μ/σ_p. Using tempering we can tune the value of $\sigma_q\sqrt{T}/\sigma_p$ to improve the efficiency, moving onto a different curve.

From this example we can take away another lesson, namely that some amount of tempering is expected to improve the efficiency in many situations, even if the optimal temperature is not known. The danger is in moving to the right of the optimal efficiency, where in any case the decay in efficiency is less severe with increasing temperature.

This toy model can be readily extended to multi-dimensional Gaussians, $p = \mathcal{N}(0, \Sigma_p)$ and $q = \mathcal{N}(\mu, \Sigma_q)$. The efficiency of importance sampling and effect of tempering depends on the details of the shapes of the covariance matrices and the direction of the bias μ, but for isotropic Gaussians and a μ which is of order unity in all dimensions, the effect of increasing the dimension is that the efficiency is roughly that of the one-dimensional efficiency raised to the number of dimensions n, $\varepsilon \sim (\epsilon_{1\mathrm{D}})^n$. Thus we expect that the efficiency of importance sampling can be quite poor in a high number of dimensions. Meanwhile, the analytic form of the optimal temperature is nearly the same as in the one-dimensional case. Further details are given in Appendix A.

D. Approximate optimal temperature

In order to maximize the efficiency of importance sampling for a given p and q, we would seek a temperature which minimizes $\chi^2(p||q_T)$. In practice we cannot access the ideal optimal temperature T_*. Performing a numerical search for T_* is expected to be impractical, since each evaluation of a candidate temperature requires sampling from q_T. Another challenge is that the efficiency must be estimated from samples, and so is inherently stochastic, as discussed in Sec. III B, making numerical searches for T_* unreliable. However, we can work out an approximation to the optimal temperature under the condition that q is sufficiently close to p and assuming that T is sufficiently close to 1. In this case we find an estimate for the optimal temperature $T_\delta \approx T_*$ which can be computed in practice using samples from q, provided that q is normalized (and hence its evidence is known). We find that

$$T_\delta = 1 - \frac{\mathbb{E}_q[(p/q - 1)\log q]}{\mathrm{Var}_q[\log q]}\,. \tag{23}$$

The density p does not to be normalized for this calculation, since the p/q term can be estimated using self-normalized weights. See Appendix B for a detailed

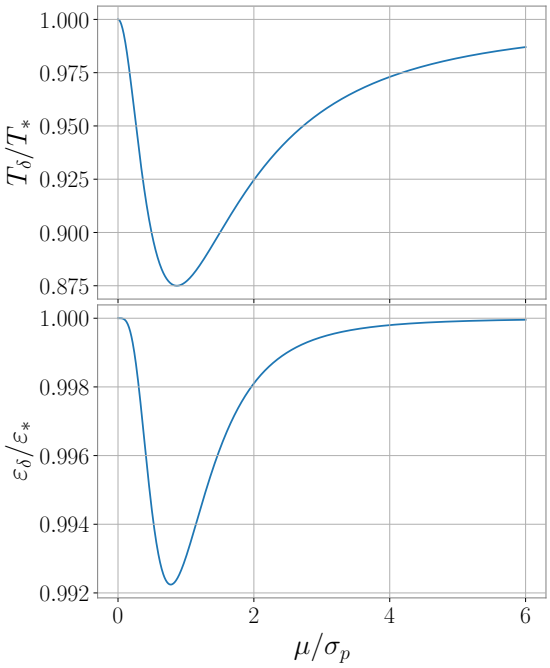

FIG. 2. Comparisons between the optimal temperature T_* from and the approximated temperature T_δ from for the one-dimensional Gaussian example, using Eqs. (21) and (24). *Top:* Ratio of the temperatures T_δ/T_*, which is a function of μ/σ_p only. *Bottom:* Ratio of the idealized efficiencies $\varepsilon_\delta/\varepsilon_*$ achieved for the optimal temperature T_* and the approximated temperature T_δ, which is a function of μ/σ_p only.

derivation and more specific treatment of the assumptions.

Equation (23) provides a practical path to estimating a good temperature for tempering. The idea would be to first sample from q to attain i.i.d. samples, use these to estimate T_δ, and perform a second round of sampling from the tempered q_T. As long as the computational expense of sampling twice using the model implicit in q is less than that of sampling from p, tempering can lead to accurate inferences with less cost.

We can examine the accuracy of the approximation (23) in the Gaussian case and compare it to the exact solution in Eq. (21). Note that we do this analysis in the Gaussian case purely for building intuition and gaining some analytical insight into the this temperature estimate. In the general case, no assumptions about the Gaussianity of the distributions is necessary. As before, when $p = \mathcal{N}(0, \sigma_p)$ and $q = \mathcal{N}(\mu, \sigma_q)$ we obtain

$$T_\delta = \frac{\sigma_p^2 + \mu^2}{\sigma_q^2} \,. \tag{24}$$

It is perhaps remarkable that the integrals involved give such a simple expression for T_δ in this case, and we give the full derivation in Appendix C. We see that in the limit $\mu = 0$, our estimate T_δ agrees with the optimal temperature T_*. Their ratio depends only on the normalized bias μ/σ_p, and is given in the top panel Fig. 2. Of greater interest is the effect of the approximation on the efficiency ε. The lower panel of Fig. 2 gives the ratio of the efficiencies $\varepsilon_\delta/\varepsilon_*$ for the optimal and approximated temperatures. This ratio also only depends on the normalized bias, and we see that is is very close to unity for all values.

The fact that the idealized efficiencies depend only on the rescaled bias μ/σ_p for both temperature choices may initially be surprising, but this can be understood as follows. The efficiency before tempering only depends on the normalized bias and the ratio of widths, and tempering only allows us to adjust the ratio of the widths without impacting the bias. Our choice of temperature fixes the ratio of the widths, in a manner that depends on the value of the normalized bias. The result is a tempered efficiency that depends only on the bias.

III. NUMERICAL RESULTS

With the notion of tempered importance sampling defined, we turn to applications of this method to multi-fidelity inference. We carry out a sequence of computational experiments to test the effectiveness of tempered importance sampling in GW parameter estimation. We use pairs of high- and low-fidelity GW models to recover the parameters of both simulated and real GW data. Following the approach of [66], our high-fidelity models incorporate higher radiative multipole moments, while our low fidelity models include only the dominant quadrupolar emission. After reviewing our analysis setup, we provide examples of lower-dimensional GW inference that demonstrate the effects of tempering on importance sampling in controlled cases. We then present results from simulated GW signals from aligned-spin systems (injections), as well as results from two real events from the third observing campaign of the LIGO, Virgo, KAGRA Collaborations.

A. Analysis details

We carry out two kinds of numerical experiments: injections of high-fidelity models into simulated data followed by Bayesian parameter estimation, and inference of real GW data. In the injection-recovery experiments we used IMRPhenomXHM [41], a model which assumes that the spin components are aligned with the orbital angular momentum of the binary, and hence neglects the effects of orbital precession. For these the waveforms were injected into zero noise using the high fidelity model and the posteriors sampled using the low fidelity model. We carry out a number of such experiments, in both restricted lower-dimensional cases as well as over the full 11

parameters. The parameter choices for these injections are shown in Table I. We also examine two events from the third gravitational wave transient catalog (GWTC-3) [16], using open data from the Gravitational Wave Open Science Center [76, 77]. For these runs, we used IMRPhenomXPHM [42], which allows for generic spins and models orbital precession, resulting in a total of 15 parameters. Tempering is carried out by scaling the power spectral density values by the appropriate temperature before sampling.

We use the following software tools for our computational experiments. The `bilby` [34, 35] Python package was employed to set up the inference problems in all the GW experiments we conducted. We use the dynamic nested sampling algorithm [78] as implemented in the `dynesty` package [79] as it is used in `bilby` to sample the posterior distributions. The use of nested sampling [80, 81] is important for our chosen approach, since we need the evidence to normalize our biasing density q in order to compute our temperature estimates T_δ. For processing the samples and visualizing the posteriors, we use `pesummary` [82].

For our injections, our priors are standard agnostic choices: uniform in detector-frame component masses, localization uniform in Euclidean volume (neglecting cosmological effects at the relatively low distances used in this study), inclination angle uniform in $\cos\iota$, polarization angle and coalescence phase uniform within their allowed ranges, and time of coalescence uniform in a window of 0.2 s centered on the injection time. We denote the dimensionless aligned-spin components of the spins as S_{1z} and S_{2z} in this work. Our priors in these components are the projection onto the orbital angular momentum of dimensionless spin vectors isotropic in orientation and with a magnitude uniform in $[0, 0.99]$. The result is a prior peaked around $S_iz = 0$ for each component, see e.g. [83]. For our GW likelihood we assume a two detector network composed of LIGO Hanford and LIGO Livingston. We use the design noise curve `aLIGO_ZERO_DET_high_P_psd` [84] as our baseline PSD in both detectors. We integrate the noise-weighted inner product from $f_{\rm low} = 20$ Hz to $f_{\rm high} = 1024$ Hz except where noted.

For our analysis of real GW events, our priors and analysis settings mirror those used in GWTC-3 [16, 85], with the following exceptions: we did not marginalize over calibration uncertainty [86, 87], and we use the Euclidean distance prior $p(d_L) \propto d_L^2$ rather than accounting for cosmological expansion. We also did not marginalize our likelihood over coalescence time or luminosity distance during sampling.

B. Error estimation

An important point to keep in mind when carrying out parameter estimation in practice is that we cannot access idealized quantities like the efficiency ε, the optimal T to minimize χ^2, or approximate temperatures defined using the distributions p and q. In all cases we instead must estimate these quantities through the samples we gather when carrying out parameter estimation. This means that reported results, including our computed efficiency of importance sampling $N_{\rm eff}/N$ and our temperature estimate T_δ are Monte Carlo estimates and carry some uncertainty.

$\mathcal{M}[M_\odot]$	q	S_{1z}	S_{2z}	α	δ	ι	ψ	ϕ_c	$t_c[s]$
30	0.5	0.4	0.3	1.3	-1.21	1	2.6	2.3	1126259642.413

TABLE I. Parameter values used in our injection studies. The component masses are m_1 and m_2, here expressed in terms of the chirp mass $\mathcal{M} = (m_1m_2)^{3/5}/(m_1+m_2)^{1/5}$ and mass ratio $q = m_2/m_1$. Also listed are the components of the dimensionless spins aligned with the orbital angular momentum S_{1z} and S_{2z}, the right ascension α, the declination δ, the inclination of the orbital angular momentum to the line of sight ι, the polarization angle ψ, the phase of coalescence ϕ_c, and the time of coalescence t_c. All angles are expressed in radians. We vary d_L between 200 Mpc and 1200 Mpc in order to scan over SNR values between 108 and 18, respectively.

Usually parameter estimation routines generate a sufficient number of samples N that these Monte Carlo uncertainties are small, and if this is not the case more samples can be gathered. However we find that in practice importance sampling for high dimensional distributions can have poor efficiencies, resulting in a small number of effective samples. The Monte Carlo error associated with these samples can be large, much larger than expected for a given N, and in some cases this prevents us from usefully estimating T_δ, as discussed below in Sec. III D.

It is thus important to have a method for quantifying the uncertainties of out estimators. Resampling methods provide simple and practical approaches for assessing the uncertainties and even biases associated with a set of samples. In this study we use a bootstrap analysis [88, 89], drawing a set of N samples with replacement from the N samples representing our distribution to get a new bootstrapped estimator. We repeat this 1000 times and compute the variance of our estimators. This allows us to estimate the uncertainties in $N_{\rm eff}/N$, and display them as 1-σ error bars on our plots. Wherever practical we ensured that we had enough samples from q and q_T to reliably estimate the efficiency of our MFIS and tempered MFIS approaches.

C. 2D and 4D CBC parameter estimation

We begin by studying lower dimensional GW inference problems in order to understand the effects of tempering. For these 2D and 4D parameter estimation experiments, we fix all but a few of the parameters at their injection values. Figures 3-7 show a comparison between the importance sampling efficiency for an untempered low-fidelity posterior and the efficiency for a posterior tem-

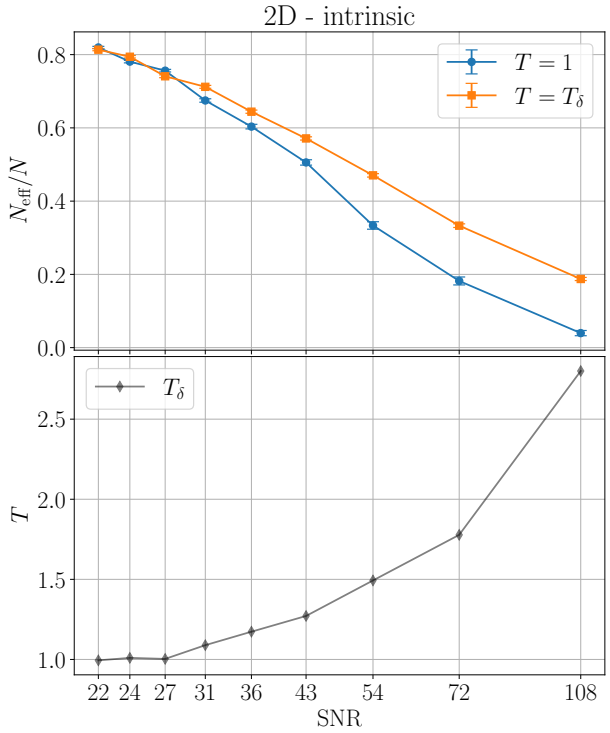

FIG. 3. Result of tempering for the 2D intrinsic problem, where the inferred parameters are the mass ratio and chirp mass.

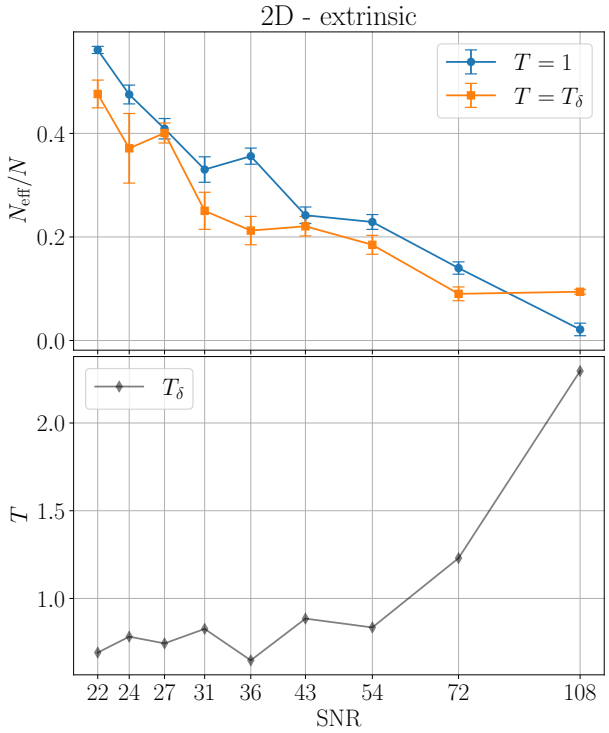

FIG. 4. Result of tempering for the 2D extrinsic problem, where the inferred parameters are the luminosity distance and inclination angle.

pered at our approximate temperature T_δ. This comparison is plotted as a function of SNR, and the corresponding values of T_δ are also provided.

Figure 3 shows the impact of tempering for a 2-dimensional inference problem in which only the chirp mass $\mathcal{M}$ and the mass ratio q are sampled over. The posteriors for this simple problem are in the Gaussian regime for the range of SNRs we explore, SNR > 20. As such the results are well-modeled by our Gaussian expectations. In fact the efficiency $N_{\rm eff}/N$ of importance sampling the low-fidelity posteriors with the high-fidelity model is fit well by a Gaussian as a function of SNR. This is the expected behavior when q and p are both Gaussian, and a fit to Eq. (22) reveals $(\sigma_q^2 - \sigma_p^2/2)^{1/2}/\mu \approx 38/\text{SNR}$ in this case. We also see from the top panel of Fig. 3 that there is a range of SNRs for which tempering at T_δ produces marked improvement in the efficiency.

Figure 4 shows the analogous result for a case in which we sample over the extrinsic parameters luminosity distance d_L and the inclination ι. In this case the expectations from our Gaussian model do not hold at these SNR values. The approximate temperatures recovered from our method are below 1, and the resulting tempered distributions mostly hurt the efficiency until the highest SNR case.

We turn next to analogous experiments for 4D. Figure 5 shows corner plots for the recovery on an SNR 72 signal for two cases. These corner plots compare the high- and low-fidelity posteriors to the tempered low-fidelity posteriors. In the first (left panel) we sample over the four intrinsic parameters of the aligned-spin models, the chirp mass $\mathcal{M}$, the mass ratio q, and the dimensionless aligned-spin components S_{1z} and S_{2z}. The posteriors appear to be fairly Gaussian, indicating that we expect improvement in $N_{\rm eff}/N$ with tempering. In the second (right panel) we sample over four extrinsic parameters, specifically the luminosity distance d_L, the inclination ι, and the sky position given by right ascension α and declination δ. The posteriors in this case are more complicated than for the intrinsic parameters, with non-Gaussian features and correlations that vary across parameter space. In this case our intuition from the Gaussian examples may not apply directly. For some further context on how these posteriors visually change for different SNRs, see Appendix D.

Figure 6 illustrates the effect of tempering the 4D inference over the intrinsic parameters. As in the 2D case, the efficiency is fit well by a Gaussian, with $(\sigma_q^2 - \sigma_p^2/2)^{1/2}/\mu \approx 34/\text{SNR}$. The 4D intrinsic case shows systematic improvement in the efficiency, but it is more modest than the improvement seen in 2D. Fig. 7 illustrates the effect when sampling over the intrinsic parameters. This case does not show consistent improvement with tempering when using our estimate T_δ for the op-

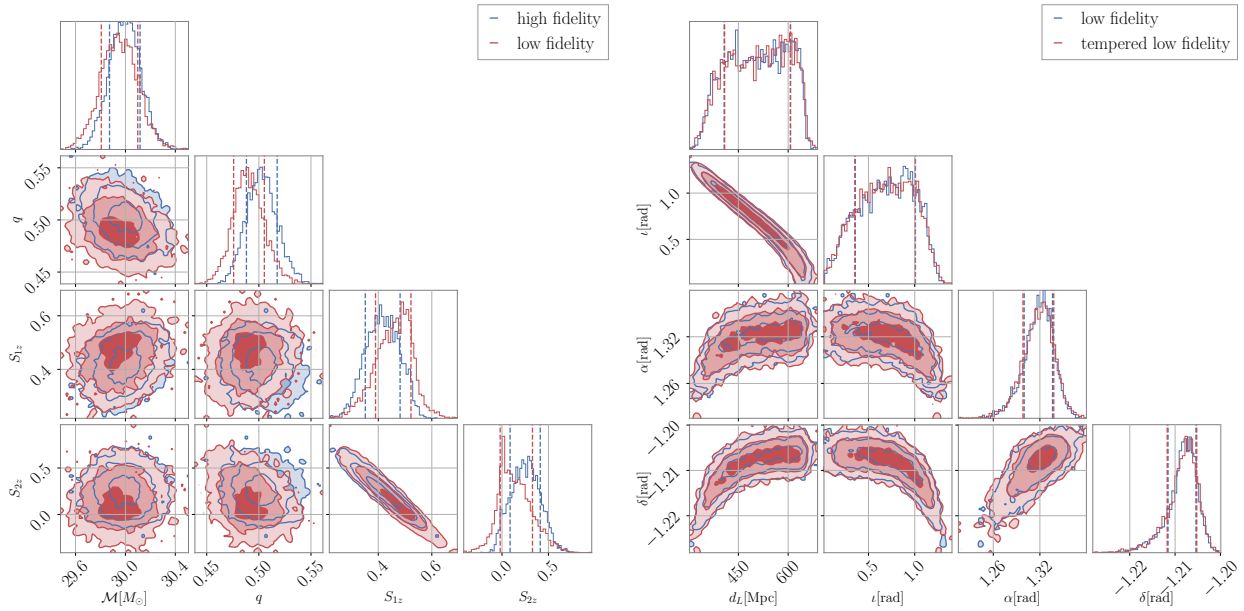

FIG. 5. 4D corner plots for $d_L = 300$ (SNR 72). We compare recovery of the high-fidelity injection with high- and low-fidelity models, along with the tempered low-fidelity recovery. For visual clarity, we show the 90% credible isocontours for each distribution. *Left:* Recovery of the four intrinsic parameters, with the remaining parameters fixed to their injected values. *Right:* Recovery of four extrinsic parameters, with the remaining parameters fixed to their injected values

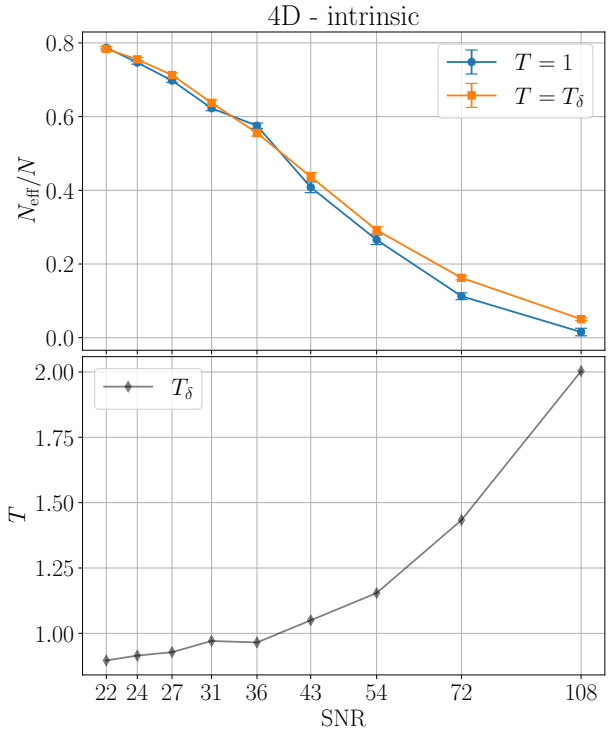

FIG. 6. Result of tempering for the 4D intrinsic problem, where the inferred parameters are the mass ratio, chirp mass, and spin magnitudes.

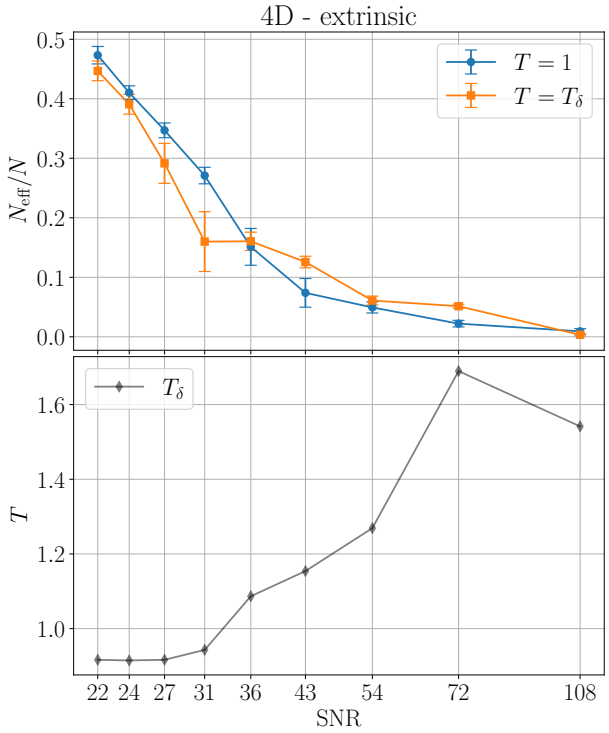

FIG. 7. Result of tempering for the 4D extrinsic problem, where the inferred parameters are the luminosity distance, inclination angle, and sky position.

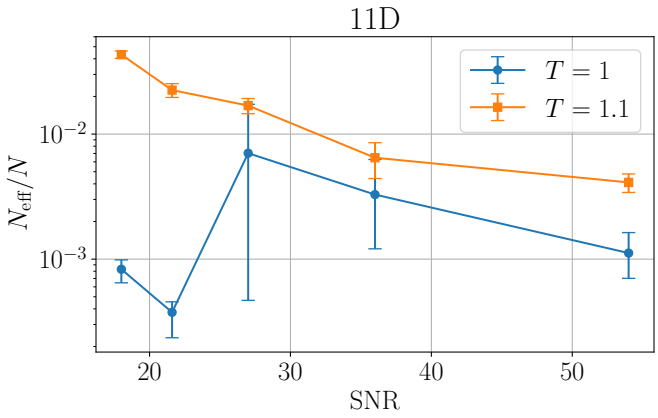

FIG. 8. Result of tempering for the 11-dimensional aligned-spin injection. For this experiment, we tempered all the runs at $T = 1.1$ instead of using the approximation T_δ.

timal temperature, though some improvement is seen at higher SNRs. Similar to the 2D case, the initial temperature estimates T_δ are below 1, but generally climb with SNR.

These low-dimensional experiments are useful for guiding our expectations and intuition for higher-dimensional problems of practical interest. We see that for well-behaved cases, such as in the intrinsic parameter inferences, our approximation for the optimal temperature leads to improvements in efficiencies, and the behavior of both the efficiency and the impact of tempering follows our expectations from the Gaussian models. However, our intuition from the simplest models does not appear to apply to the extrinsic parameters, and these show a case where tempering can hurt the efficiency of MFIS.

D. Full aligned-spin CBC parameter estimation

Having seen mixed results for tempering in our lower-dimensional tests, we turn to a more realistic case of the 11D recovery of an injected GW signal from an aligned-spin binary black hole coalescence. As with our 2D and 4D tests, our injected signal has no added noise (we assume a zero-noise realization of the random detector noise). We find a very poor efficiency for our lower-fidelity recover of the high-fidelity injection. This is seen in Fig. 8 across a range of SNR values, where the $T = 1$ recovery commonly has efficiencies of $\sim 1\%$. The efficiencies are roughly flat across the SNRs tested. Our low efficiencies are consistent with the poor efficiency of the zero-noise injection and recovery presented in [66] using similar analysis choices.[1]

[1] See Table 1 of [66], and note that the differences in their waveform models are expected to be even larger than ours, likely accounting for another factor of ~ 10 decrease in the efficiency.

Initially, we were concerned that these low efficiencies would mean that our estimates for T_δ would be unreliable, given the small $N_{\rm eff}$ for the samples we drew from q. Therefore, inspired by the observation that in many cases any amount of tempering improves the efficiency of importance sampling in our Gaussian examples, for this experiment we opted for a different prescription and simply tempered each case by the same temperature $T = 1.1$, drawing a similar number $N \approx 2 \times 10^4$ samples. The resulting efficiencies are seen in Fig. 8. While these efficiencies are still poor overall, we see a several times improvement at moderate SNRs with no particular loss of efficiency at higher SNRs as compared to the $T = 1$ case.

The success of uniform tempering in 11D points towards another potential way to benefit from tempered MFIS. Rather our proposed two-step process, first performing standard inference with $T = 1$ to get samples from q and then estimating T_δ with these, prior experience or theoretical analysis can provide a proposed temperature for a single step of tempered MFIS.

E. Parameter Estimation for GWTC-3 Events

In order to test our method in a fully realistic situation, we apply it to two events from GWTC-3. Our goal is to probe the regime in which the two posteriors are very similar as well as the regime in which they substantially differ. For the former scenario, we choose GW191222_033537 (hereafter GW191222), a fairly typical binary black hole signal which favors equal masses, small effective spin parameter $\chi_{\rm eff}$, and no strong signs of orbital precession. This is an example of an event for which the higher mode content of the gravitational wave is expected to be small, meaning our low and high fidelity models produce similar results. This means that the overall efficiency of multifidelity importance sampling is higher, since q is very similar to p, but also that the margin of improvement is smaller since there is little extra information coming from the high fidelity model.

GW200129_065458 (hereafter GW200129) was selected for the opposite reason; it is a high SNR event where standard analysis with IMPhenomXPHM infers large component spins, clear orbital precession [16, 90] and some posterior weight towards unequal masses $m_2/m_1 \sim 1/2$ where higher modes make greater contributions to the signal. Parameter estimation for this event is systematically different across signal models, e.g. [16, 90–92], and it is also complicated by non-Gaussian noise in the raw data, which must be mitigated [16, 93].

For these reasons we expect the choice of signal model to have greater impact for GW200129, and therefore the distributions p and q inferred with and without higher modes to be more different. We see evidence for this in the corner plots of Fig. 9, which shows marginalized 2D and 1D posteriors for GW200129 for selected intrinsic and extrinsic parameters. We show both the samples pre-

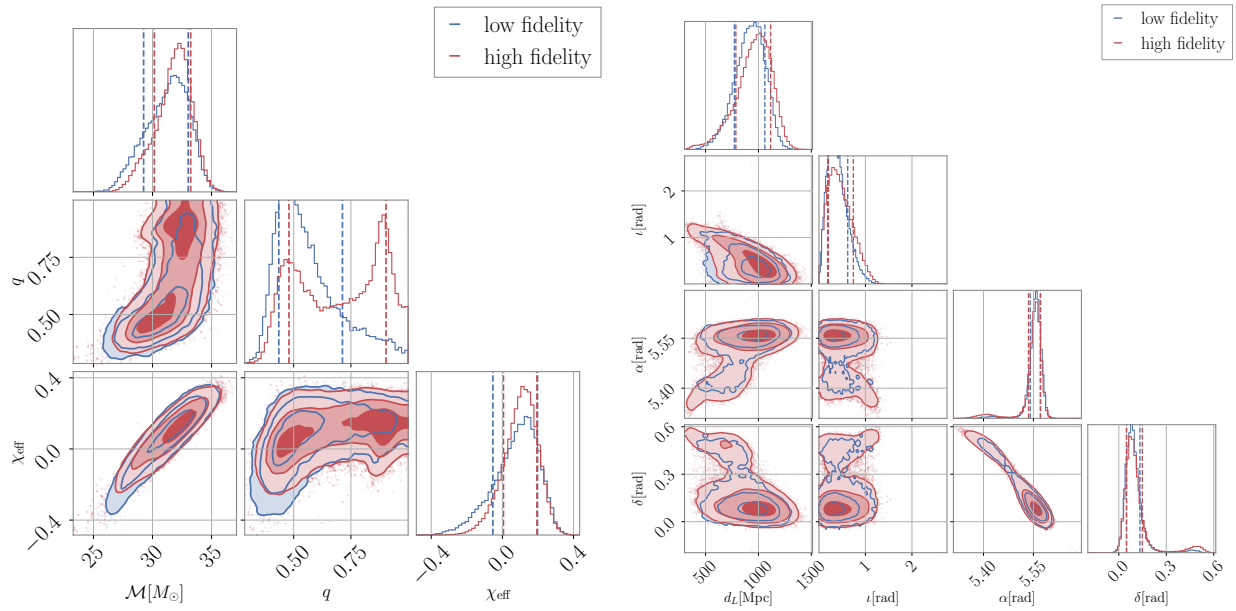

FIG. 9. Corner plots for GW200129, illustrating recovery with the high-fidelity model, which includes precession and higher mulitipolar emission, and the low-fidelity model which includes only precession and the dominant quadrupolar emission. Also shown is the result of tempering the low-fidelity distribution at $T = T_\delta$. Unlike in Fig. 5, these plots show different marginals of the full 15D inference, and the effect of tempering on the marginals of this high-dimensional posterior is nontrivial. However, in Fig. 10 we see the improvement in efficiency as a result of tempering quite clearly, despite it being harder to visually interpret the corner plots.

sented in GWTC-3 using the IMRPhenomXPHM model with a complete set of modes (high-fidelity), and our own low-fidelity recovery of this event. The primary difference in the instrinsic parameters appears to be the absence in the low-fidelity recovery of a second mode at larger q values, which impacts both the effective spin $\chi_{\rm eff}$ and chirp mass $\mathcal{M}$ inferences. While it is natural to attribute this to the difference in the models (and the presence or absence of multiple posterior modes is model-dependent for GW200129, see e.g. [16, 90, 92]), we cannot rule out that our four independent low-fidelity runs failed to recover this distinct region of probability. The extrinsic parameters also show differences in recovery, notably poorer coverage of the tails of the high-fidelity posteriors by the low-fidelity in several regions. The clear difference between these posteriors should result in lower overall efficiency, but greater potential for improvement with tempering.

Figure 10 shows the observed improvement in efficiency for both events, in this case using violin plots to visualize the uncertainty on the efficiency as estimated from bootstrap resampling. Notably importance sampling with real events provides generally better efficiencies than our 11D, zero-noise case even though the dimensionality is higher here. This was observed also in the analysis of [66]. These higher efficiencies, together with analysis settings that draw a larger number of samples for the $T = 1$ inference for GW200129, allow us to reliably estimate T_δ for these real events and temper using it.

As expected, GW191222 has a much higher efficiency than GW200129. For GW191222 the estimated optimal temperature T_δ is close to unity, and tempering improves the efficiency of MFIS, with the median of the tempered efficiency estimate well into the tail of the uncertainty of the standard analysis. For GW200129, tempering provides a clear improvement, more than doubling the efficiency of importance sampling.

IV. DISCUSSION

MFIS is a promising framework for GW inference because in this domain, there exists a rich hierarchy of waveform models, with a range of computational costs and including a variety of physical effects. The high computational cost of some of the best models means that exploiting samples from cheaper models via importance sampling has the potential to greatly reduce computational costs and accelerate inference. In practice, direct application of importance sampling can lead to very low efficiencies, as seen in the existing literature, e.g. [66, 68].

Our goal in this work is to present a method to improve the efficiency of importance sampling generally, and explore its application to GW inference. We introduce the idea of tempered MFIS, where the biasing distribution is tempered in order to reduce the chi-squared divergence and increase the efficiency. Further, we derive a practical estimate for the temperature needed to achieve this improvement, at the cost of generating an initial set of

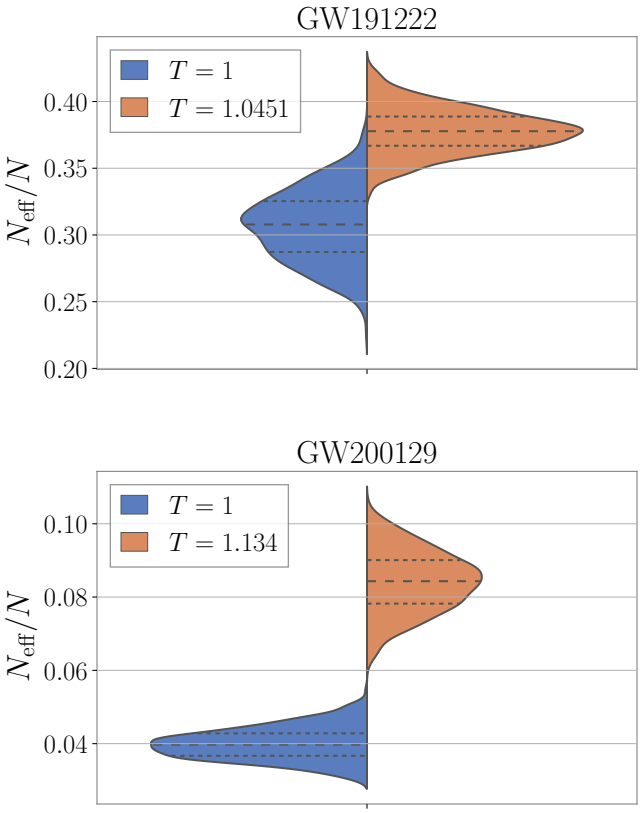

FIG. 10. Result of tempering for the events GW191222 and GW200129. Note the improvement in efficiency, subject to uncertainty estimated through bootstrap resampling.

samples from the untempered biasing distribution.

By carefully investigating this idea of tempering in controlled settings (such as our Gaussian experiments), we arrive at a principled methodology for improving importance sampling efficiency via the relatively cheap calculation of our approximately optimal temperature. In doing so, we extract several other general insights: low efficiency is often caused by samples drawn from poorly overlapped tails, importance sampling is highly sensitive to the shapes of the distributions, and the overall efficiency scales poorly with dimension.

Furthermore, the mixed success of applying our tempering method in the GW experiments suggests that this is not a one-size-fits-all solution, and that care must be taken in understanding and modifying these distributions. However, the cases in which the efficiency did improve in our experiments motivates the future pursuit of other ways to modify biasing distributions, the goal being a more robust and reliable method for battling low efficiency. Two of our results are especially promising: first that our estimate for the optimal temperature improves the efficiency of importance sampling for real GW data, and secondly that an principled guess for a temperature uniformly improves the efficiency for simulated signals injected into zero noise across a range of signal strengths, in one case by more than an order of magnitude.

Finally, we use bootstrap resampling techniques to estimate uncertainties in quantities estimated from samples such as the efficiency and temperature. The application of resampling methods to generate frequentist-based uncertainties on quantities like the effective number of samples is of independent interest in GW data analysis. There are a variety of such resampling methods, and some such as jackknife resampling allow one to not only estimate these uncertainties but also correct for biases present in the estimators [88]. Future work may pursue these methods to better capture and account for the uncertainties inherent in our sample-based efficiency and temperature estimates.

Our method may be useful in various applications of importance sampling to GWs. It would be especially interesting to understand if tempering can be applied to methods such as those used in [68], where importance sampling is applied to correct samples drawn using a learned normalizing flow. The success of tempering may lead to simple ideas to improve the construction of these machine-learning based methods.

ACKNOWLEDGMENTS

We thank Ethan Payne and Dingcheng Luo for useful discussions. We especially thank Colm Talbot for useful discussions and technical advice throughout this project. We also thank Carl-Johan Haster for a careful reading of this manuscript and for helpful comments. B.S. and O.G. were supported by DOE grant DE-SC0019303 during a portion of this work, and B.S. by a CNS Catalyst Grant at UT Austin during this work. A.Z. was supported by NSF Grants PHY-2207594 and PHY-2308833 while carrying out this work. P.C. was partially supported by NSF Grants #2245674, #2245111, and #2325631. This paper has preprint numbers UT-WI-19-2024 and LIGO-P2400206.

This research has made use of data or software obtained from the Gravitational Wave Open Science Center [77], a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch

Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.

This work makes use of the `lalsuite` [94], `gwpy` [95], `bilby` [34, 96, 97] `dynesty` [79, 98], and `pesummary` [82, 99] software packages.

Appendix A: Multi-dimensional Gaussian example

In this appendix we collect results on the efficiency of importance sampling in the context of multi-dimensional Gaussians. Without loss of generality, place the mean of p at the origin, letting $p = \mathcal{N}(0, \Sigma_p)$ and $q = \mathcal{N}(\mu, \Sigma_q)$. We could further rotate and rescale our coordinates to make p a unit Gaussian, but for clarity we retain Σ_p explicitly. Provided that

$$\Gamma := 2\Sigma_p^{-1} - \Sigma_q^{-1}\,. \tag{A1}$$

is positive definite, these densities can be inserted into Eq. (14) and the Gaussian integral resolved. Let $\Sigma := \Gamma^{-1}$, then

$$\begin{aligned}\chi^2(p||q) + 1 &= \frac{|\Sigma_q|}{|\Sigma_p|^{1/2}} \frac{\exp\left[\frac{1}{2}\mu^\top(\Sigma_q^{-1} + \Sigma_q^{-1}\Sigma\Sigma_q^{-1})\mu\right]}{|2\Sigma_q - \Sigma_p|^{1/2}} \\ &= \frac{|\Sigma_q|}{|\Sigma_p|^{1/2}} \frac{\exp\left[\mu^\top(2\Sigma_q - \Sigma_p)^{-1}\mu\right]}{|2\Sigma_q - \Sigma_p|^{1/2}}\,. \end{aligned} \tag{A2}$$

In the second line we have used a variant of the Woodbury identity [100], specifically

$$(A + B)^{-1} = A^{-1} - A^{-1}B(\mathbb{1} + A^{-1}B)^{-1}A^{-1}\,. \tag{A3}$$

Tempering then makes the replacement $\Sigma_q \to T\Sigma_q$ in Eq. (A2).

In the case of isotropic n-dimensional Gaussians, the expressions simplify further. With $\Sigma_p = \sigma_p^2\mathbb{1}$, $\Sigma_q = \sigma_q^2\mathbb{1}$ we also have $\Sigma = \sigma_p^2\sigma_q^2/(2\sigma_q^2 - \sigma_p^2)\mathbb{1}$ and so

$$\chi^2(p||q) + 1 = \left(\frac{T\sigma_q^2}{\sigma_p\sqrt{2T\sigma_q^2 - \sigma_p^2}}\right)^n \exp\left[\frac{\mu^\top\mu}{2T\sigma_q^2 - \sigma_p^2}\right]\,. \tag{A4}$$

This result agrees with the $n = 1$ result from Eq. (20). Further, if the bias μ as $O(1)$ entries in all n dimensions, then $\mu^\top\mu \sim O(n)$ and hence the result Eq. (A4) is just Eq. (20) raised to the power n. If on the other hand there are some unbiased directions, the efficiency is penalized by an effective dimension that is less than n.

Continuing consideration of the isotropic Gaussian case, we can solve for the optimal temperature

$$T_* = \frac{3n\sigma_p^2 + 2\mu^\top\mu + \sqrt{n^2\sigma_p^4 + 12n\sigma_p^2\mu^\top\mu + 4(\mu^\top\mu)^2}}{4n\sigma_q^2}\,. \tag{A5}$$

Again under the assumption that $\mu^\top\mu \sim O(n)$, we see that the factors of n cancel out of every term, resulting in essentially the same optimal temperature as for the $n = 1$ case.

From this we conclude that as the dimension of the densities under consideration increases, we expect the efficiency of importance sampling to decrease rapidly, scaling as ϵ_{1D}^n for n relevant dimensions. Meanwhile, the best temperatures for tempering will remain similar to the estimates for $n = 1$.

Appendix B: Approximate optimal temperature derivation

Let p and q be probability densities over a space X, such that the measure corresponding to p is absolutely continuous with respect to the analogous measure for q. The χ^2-divergence between p and q is

$$\chi^2(p||q) = \int \frac{p^2}{q}dx - 1\,, \tag{B1}$$

Recall that p and q are densities and therefore functions of $x \in X$. We write these functions without this explicit argument for visual clarity.

It is more convenient for our derivation to work with the inverse temperature $\beta = 1/T$ in this Appendix, rather than T. The *tempered density* q_β is therefore defined as

$$q_\beta := \frac{q^\beta}{Z_\beta}\,, \qquad Z_\beta = \int q^\beta dx\,. \tag{B2}$$

Our goal is to find β_* such that

$$\beta_* = \arg\min_\beta \chi^2(p||q_\beta)\,, \tag{B3}$$

The value of β that minimizes the χ^2-divergence maximizes the efficiency of using q_β as a biasing density for importance sampling.

Since we cannot find β_* in general, we derive an approximate expression β_δ by assuming that p and q are approximately the same and that $\beta_* \approx \beta_\delta$ is close to unity, in the following sense. We define a bookkeeping parameter ϵ that tracks small quantities, and write

$$p = q(1 + \epsilon r)\,, \tag{B4}$$

$$\beta = 1 + \epsilon\,\delta\beta\,. \tag{B5}$$

We track quantities at leading order in ϵ to solve for the $\delta\beta$ that minimizes χ^2 at this order. This bookkeeping parameter falls out of the final solution, which is linear in the small r deviation between p and q.

We start by taking the derivative of $\chi^2(p||q_\beta)$ with respect to β. We have

$$\chi^2(p||q_\beta) = \int \frac{p^2}{q_\beta} dx - 1 = Z_\beta \int \frac{p^2}{q^\beta} dx - 1\,, \tag{B6}$$

so that

$$\frac{d}{d\beta}\chi^2(p||q_\beta) = Z_\beta \int \frac{d}{d\beta}\frac{p^2}{q^\beta} dx + \int \frac{p^2}{q^\beta} dx \frac{dZ_\beta}{d\beta} \tag{B7}$$

$$= -\int q^\beta dx \int \frac{p^2}{q^\beta} \log q\, dx + \int \frac{p^2}{q^\beta} dx \int q^\beta \log q\, dx\,. \tag{B8}$$

Using the substitution $\beta = 1 + \epsilon\,\delta\beta$ and Taylor expanding q^β and p^2/q^β around $\epsilon\,\delta\beta = 0$, we find

$$q^\beta = q^{1+\epsilon\delta\beta} = q(1 + \epsilon\delta\beta \log q + \mathcal{O}(\epsilon^2)) \tag{B9}$$

and additionally, using $p = q(1+\epsilon r)$,

$$\frac{p^2}{q^\beta} = \frac{p^2}{q(1+\epsilon\delta\beta\log q + \mathcal{O}(\epsilon^2)} \tag{B10}$$

$$= \frac{p^2}{q}(1 - \epsilon\delta\beta\log q + \mathcal{O}(\epsilon^2)) \tag{B11}$$

$$= \frac{q^2(1+\epsilon r)^2}{q}(1 - \epsilon\delta\beta\log q + \mathcal{O}(\epsilon^2)) \tag{B12}$$

$$= q(1 - \epsilon\delta\beta\log q + 2\epsilon r + \mathcal{O}(\epsilon^2))\,. \tag{B13}$$

Thus, keeping terms only to first order in ϵ, we have

$$q^\beta \approx q(1 + \epsilon\delta\beta\log q)\,, \tag{B14}$$

$$\frac{p^2}{q^\beta} \approx q(1 - \epsilon\delta\beta\log q + 2\epsilon r)\,. \tag{B15}$$

Plugging these approximations back into Eq. (B8), we find that up to first order in ϵ,

$$\begin{aligned}
\frac{d}{d\beta}\chi^2(p||q_\beta) &\approx -\left(1 + \epsilon\delta\beta\int \log q\, dx\right)\int q(1 - \epsilon\delta\beta\log q + 2\epsilon r)\log q\, dx \\
&\quad + \int q(1 - \epsilon\delta\beta\log q + 2\epsilon r)dx \int q(1+\epsilon\delta\beta\log q)\log q\, dx \\
&= -(1 + \epsilon\delta\beta\mathbb{E}_q[\log q])(\mathbb{E}_q[\log q] - \epsilon\delta\beta\mathbb{E}_q[(\log q)^2] + 2\epsilon\mathbb{E}_q[r\log q]) + (1 - \epsilon\delta\beta\mathbb{E}_q[\log q])(\mathbb{E}_q[\log q] + \epsilon\delta\beta\mathbb{E}_q[(\log q)^2]) \\
&\approx -(\mathbb{E}_q[\log q] - \epsilon\delta\beta\mathbb{E}_q[(\log q)^2] + 2\epsilon\mathbb{E}_q[r\log q] + \epsilon\delta\beta\,(\mathbb{E}_q[\log q])^2) + \mathbb{E}_q[\log q] + \epsilon\delta\beta\mathbb{E}_q[(\log q)^2] - \epsilon\delta\beta(\mathbb{E}_q[\log q])^2 \\
&= 2\epsilon\delta\beta\mathbb{E}_q[(\log q)^2] - 2\epsilon\delta\beta(\mathbb{E}_q[\log q])^2 - 2\epsilon\mathbb{E}_q[r\log q]\,.
\end{aligned} \tag{B16}$$

The result is

$$\frac{d}{d\beta}\chi^2(p||q_\beta) \approx 2\epsilon(\delta\beta\,\mathrm{Var}_q[\log q] - \mathbb{E}_q[r\log q])\,. \tag{B17}$$

Next, since our goal is to find the value of β that minimizes $\chi^2(p||q_\beta)$, we set this expression equal to zero and solve for $\delta\beta$. We find

$$\beta_\delta = 1 + \frac{\mathbb{E}_q[r\log q]}{\mathrm{Var}_q[\log q]}\,, \tag{B18}$$

where we denote the inverse temperature β_δ to distinguish this approximation from the true minimizer β_*.

Finally, we recall our goal is to compute an approximation T_δ for the optimal temperature T_*. At the order we have worked, there are two natural choices for representing T_δ in terms of $\delta\beta$. We can re-expand in small ϵ, $T_\delta \approx 1 - \epsilon\,\delta\beta$, or use a resummed version $T_\delta \approx 1/(1+\epsilon\,\delta\beta)$. These agree to leading order in ϵ but can differ appreciably for moderate temperatures. By applying both to our one-dimensional Gaussian example from Sec. II C, we find that the former choice (re-expanding in small quantities) dramatically outperforms the latter resummed estimate, which suffers from divergences at moderate biases. Remembering then that $r = (1/\epsilon)(p/q - 1)$, we obtain our result Eq. (23).

This approach to estimating T_* boils down to seeking the minimum of $\chi^2(p||q)$ using a single Newton-Raphson step starting from $\beta = 1$. If β_* is sufficiently close to $\beta = 1$ the method is guaranteed to give a temperature β_δ closer to β_*, but as is standard for applications of Newton's method, if $\beta = 1$ is too far from β_* this method can fail catastrophically. Note that while further iterations would improve the temperature estimate in the convergent case, each iterate requires sampling from q_T and so may be prohibitively expensive.

Appendix C: Approximate optimal temperature in the Gaussian case

In the case that p and q are both Gaussian, we can find the optimal temperature T_* explicitly, as given in Eq. (21). We now aim to use our approximation as given in Eq. (24) to obtain T_δ in terms of μ, σ_p, and σ_q.

In this case,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_p^2}\right), \tag{C1}$$

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma_q} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right). \tag{C2}$$

We start by noting that

$$\mathbb{E}_q[(p/q-1)\log q] = \int p \log q \, dx - \int q \log q \, dx\,, \tag{C3}$$

$$\mathrm{Var}_q[\log q] = \int q(\log q)^2 \, dx - \left(\int q \log q \, dx\right)^2. \tag{C4}$$

Next,

$$\begin{aligned}
\int q \log q \, dx &= \frac{1}{\sqrt{2\pi}\sigma_q} \int \left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2} - \log\sqrt{2\pi}\sigma_q\right) \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx \\
&= -\frac{1}{2\sqrt{2\pi}\sigma_q^3} \int (x-\mu)^2 \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx - \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_q} \int \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx \\
&= -\frac{1}{2\sqrt{2\pi}\sigma_q^3}(\sqrt{2\pi}\sigma_q^3) - \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_q}(\sqrt{2\pi}\sigma_q) \\
&= -\frac{1}{2} - \log\sqrt{2\pi}\sigma_q\,.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\int q(\log q)^2 dx &= \frac{1}{\sqrt{2\pi}\sigma_q} \int \left(\frac{(x-\mu)^4}{4\sigma_q^4} + \frac{(x-\mu)^2}{\sigma_q^2}\log\sqrt{2\pi}\sigma_q + (\log\sqrt{2\pi}\sigma_q)^2\right) \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx \\
&= \frac{1}{4\sqrt{2\pi}\sigma_q^5} \int (x-\mu)^4 \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx + \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_q^3} \int (x-\mu)^2 \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx \\
&\quad + \frac{(\log\sqrt{2\pi}\sigma_q)^2}{\sqrt{2\pi}\sigma_q} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2}\right) dx \\
&= \frac{1}{4\sqrt{2\pi}\sigma_q^5}(3\sqrt{2\pi}\sigma_q^5) + \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_q^3}(\sqrt{2\pi}\sigma_q^3) + \frac{(\log\sqrt{2\pi}\sigma_q)^2}{\sqrt{2\pi}\sigma_q}(\sqrt{2\pi}\sigma_q) \\
&= \frac{3}{4} + \log\sqrt{2\pi}\sigma_q + (\log\sqrt{2\pi}\sigma_q)^2\,.
\end{aligned}$$

Finally,

$$\begin{aligned}
\int p \log q \, dx &= \frac{1}{\sqrt{2\pi}\sigma_p} \int \left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_q^2} - \log\sqrt{2\pi}\sigma_q\right) \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_p^2}\right) dx \\
&= -\frac{1}{2\sqrt{2\pi}\sigma_p\sigma_q^2} \int (x-\mu)^2 \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_p^2}\right) dx - \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_p} \int \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_p^2}\right) dx \\
&= -\frac{1}{2\sqrt{2\pi}\sigma_p\sigma_q^2}(\sqrt{2\pi}\sigma_p(\sigma_p^2+\mu^2)) - \frac{\log\sqrt{2\pi}\sigma_q}{\sqrt{2\pi}\sigma_p}(\sqrt{2\pi}\sigma_p) \\
&= -\frac{\sigma_p^2+\mu^2}{2\sigma_q^2} - \log\sqrt{2\pi}\sigma_q\,.
\end{aligned}$$

Putting it all together, several cancellations yield the simplified results

$$\mathbb{E}_q[(p/q-1)\log q] = -\frac{\sigma_p^2+\mu^2}{2\sigma_q^2} - \log\sqrt{2\pi}\sigma_q + \frac{1}{2} + \log\sqrt{2\pi}\sigma_q = \frac{1}{2}\left(1-\frac{\sigma_p^2+\mu^2}{\sigma_q^2}\right), \tag{C5}$$

$$\mathrm{Var}_q[\log q] = \frac{3}{4} + \log\sqrt{2\pi}\sigma_q + (\log\sqrt{2\pi}\sigma_q)^2 - (\frac{1}{4} + \log\sqrt{2\pi}\sigma_q + (\log\sqrt{2\pi}\sigma_q)^2) = \frac{1}{2}. \tag{C6}$$

Thus,

$$\delta\beta \approx 1 - \frac{\sigma_p^2+\mu^2}{\sigma_q^2}. \tag{C7}$$

By re-expanding in the expected smallness of $\delta\beta$, we have finally

$$T_\delta = \frac{\sigma_p^2+\mu^2}{\sigma_q^2}, \tag{C8}$$

which is used in Sec. II D.

These steps can be generalized with some effort for generic multi-dimensional Gaussians, giving at the same level of approximation

$$T_\delta = \frac{1}{n}\left(\mathrm{Tr}[\Sigma_q^{-1}\Sigma_p] + \mu^\top\Sigma_q^{-1}\mu\right), \tag{C9}$$

which demonstrates that this temperature estimate tends to remain $O(1)$ even as n grows, as required for our approximation.

Appendix D: Additional corner plots for 4D experiments

In this appendix, we provide additional corner plots for the 4D injection-recovery experiments outlined in section III C. Figure 5 in that section shows the corner plots for intrinsic and extrinsic parameter recoveries for an SNR of 72. Here we provide the analogous plots for two additional SNRs, 108 and 54.

These two cases bookend the case shown in Fig. 5, exhibiting higher and lower SNRs. We focus first on the case where we sample over the intrinsic parameters (left hand panels of Fig. 11). In the higher SNR case, it is visually clear that the coverage of the high-fidelity intrinsic posteriors are improved when tempering the low-fidelity posteriors. The lower SNR case shows a smaller and more subtle change, but both cases exhibit improved efficiency, with better relative improvement for the higher SNR case.

The high SNR case helps to illustrate that when tempering improves the coverage of the target distribution, the efficiency of importance sampling is improved and the reason is visually clear. Indeed one can see from the expression for $N_{\rm eff}$, Eq. (16), why the coverage of the tails can have outsized impact. When the biasing distribution does not cover the tail of the target, a rare sample drawn from the edge of the biasing distribution ends up with a large weight, which can then overwhelm the weights of the other samples and strongly reduce $N_{\rm eff}$. Meanwhile, it is less clear visually for the lower SNR case that the coverage is significantly improved, but we see that tempering still improves the efficiency of important sample, illustrating that tempering can help even when the changes to the tempered distribution are subtle.

The extrinsic parameter inferences (right panels of Fig. 11) are harder to interpret. In these case the posteriors show that the low-fidelity model already mostly covers the tails of the high-fidelity distribution, even before tempering, with the exception seeming to be the right ascension parameter. Our estimates for the optimal temperature also prove to be less reliable in these cases, since in several cases tempering with T_δ does not yield improved efficiency. This is the case for the highest SNR example.

The SNR 72 case shown in Fig. 5 falls between these, in terms of the visual impact of tempering. This is much clearer in the intrinsic case, where the estimate T_δ appears to be a decent approximation of the optimal temperature, and the amount of tempering increases predictably with increased bias. Again, the extrinsic inferences resist such a simple interpretation, especially since the untempered low-fidelity posteriors are not as dramatically biased relative to their high-fidelity counterparts.

We further note that these corner plots only show probability isocontours of marginal distributions, comparing only one or two parameters at a time. They therefore do not display the full information content of the posteriors. This further justifies these corner plots as qualitative guides for our intuition, rather than as precise measures of the effect of tempering in higher dimensions.

[1] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "Observation of Gravitational Waves from a Binary Black Hole Merger," Phys. Rev. Lett. **116**, 061102 (2016),

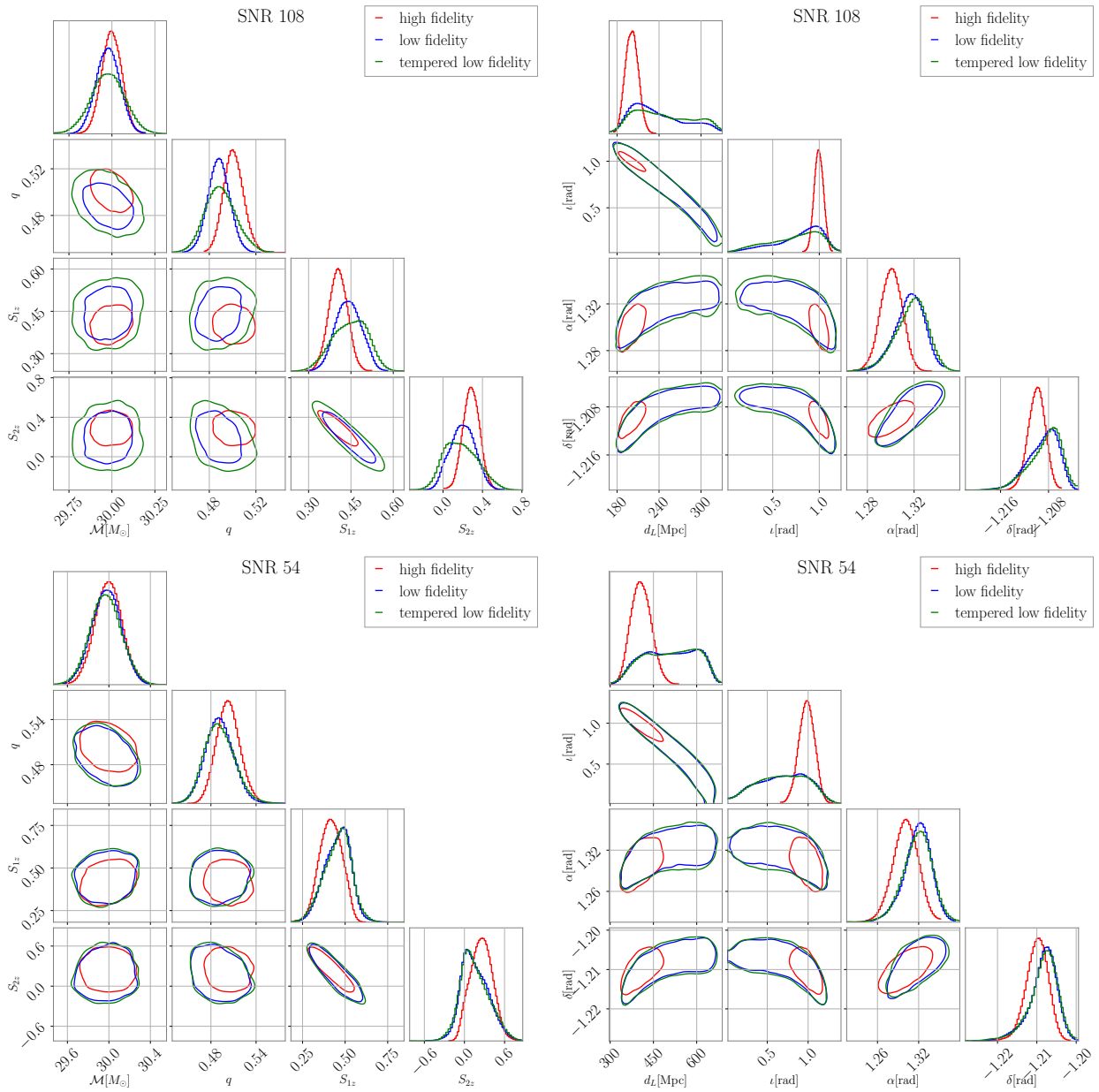

FIG. 11. 4D corner plots for the intrinsic and extrinsic cases, for two additional SNRs, 108 and 54. These bookend the SNR 72 case presented in figure 5 and place that result in context. We see that for the highest SNR, especially in the intrinsic case, the change in overlap in the tails of the distribution is more dramatic for the tempered distribution versus the untempered one, while for the lower SNR case the change is significantly more subtle.

arXiv:1602.03837 [gr-qc].

[2] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence," Phys. Rev. Lett. **116**, 241103 (2016), arXiv:1606.04855 [gr-qc].

[3] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "Binary Black Hole Mergers in the first Advanced LIGO Observing Run," Phys. Rev. X **6**, 041015 (2016), [Erratum: Phys.Rev.X 8, 039903 (2018)], arXiv:1606.04856 [gr-qc].

[4] Benjamin P. Abbott *et al.* (LIGO Scientific, VIRGO), "GW170104: Observation of a 50-Solar-Mass Binary Black Hole Coalescence at Redshift 0.2," Phys. Rev. Lett. **118**, 221101 (2017), [Erratum: Phys.Rev.Lett. 121, 129901 (2018)], arXiv:1706.01812 [gr-qc].

[5] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence," Phys. Rev. Lett. **119**, 141101 (2017), arXiv:1709.09660 [gr-qc].

[6] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral," Phys. Rev. Lett. **119**, 161101 (2017), arXiv:1710.05832 [gr-qc].

[7] B. . P. . Abbott *et al.* (LIGO Scientific, Virgo),

"GW170608: Observation of a 19-solar-mass Binary Black Hole Coalescence," Astrophys. J. Lett. **851**, L35 (2017), arXiv:1711.05578 [astro-ph.HE].

[8] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs," Phys. Rev. X **9**, 031040 (2019), arXiv:1811.12907 [astro-ph.HE].

[9] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "GW190425: Observation of a Compact Binary Coalescence with Total Mass $\sim 3.4 M_\odot$," Astrophys. J. Lett. **892**, L3 (2020), arXiv:2001.01761 [astro-ph.HE].

[10] R. Abbott *et al.* (LIGO Scientific, Virgo), "GW190412: Observation of a Binary-Black-Hole Coalescence with Asymmetric Masses," Phys. Rev. D **102**, 043015 (2020), arXiv:2004.08342 [astro-ph.HE].

[11] R. Abbott *et al.* (LIGO Scientific, Virgo), "GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object," Astrophys. J. Lett. **896**, L44 (2020), arXiv:2006.12611 [astro-ph.HE].

[12] R. Abbott *et al.* (LIGO Scientific, Virgo), "GW190521: A Binary Black Hole Merger with a Total Mass of $150 M_\odot$," Phys. Rev. Lett. **125**, 101102 (2020), arXiv:2009.01075 [gr-qc].

[13] R. Abbott *et al.* (LIGO Scientific, Virgo), "GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run," Phys. Rev. X **11**, 021053 (2021), arXiv:2010.14527 [gr-qc].

[14] R. Abbott *et al.* (LIGO Scientific, KAGRA, VIRGO), "Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences," Astrophys. J. Lett. **915**, L5 (2021), arXiv:2106.15163 [astro-ph.HE].

[15] R. Abbott *et al.* (LIGO Scientific, VIRGO), "GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run," Phys. Rev. D **109**, 022001 (2024), arXiv:2108.01045 [gr-qc].

[16] R. Abbott *et al.* (LIGO Scientific, Virgo, KAGRA), "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run," Phys. Rev. X **13**, 041039 (2023), arXiv:2111.03606 [gr-qc].

[17] "Observation of Gravitational Waves from the Coalescence of a $2.5-4.5\ M_\odot$ Compact Object and a Neutron Star," (2024), arXiv:2404.04248 [astro-ph.HE].

[18] Alexander H. Nitz, Collin Capano, Alex B. Nielsen, Steven Reyes, Rebecca White, Duncan A. Brown, and Badri Krishnan, "1-OGC: The first open gravitational-wave catalog of binary mergers from analysis of public Advanced LIGO data," Astrophys. J. **872**, 195 (2019), arXiv:1811.01921 [gr-qc].

[19] Alexander H. Nitz, Thomas Dent, Gareth S. Davies, Sumit Kumar, Collin D. Capano, Ian Harry, Simone Mozzon, Laura Nuttall, Andrew Lundgren, and Márton Tápai, "2-OGC: Open Gravitational-wave Catalog of binary mergers from analysis of public Advanced LIGO and Virgo data," Astrophys. J. **891**, 123 (2020), arXiv:1910.05331 [astro-ph.HE].

[20] Alexander H. Nitz, Collin D. Capano, Sumit Kumar, Yi-Fan Wang, Shilpa Kastha, Marlin Schäfer, Rahul Dhurkunde, and Miriam Cabero, "3-OGC: Catalog of Gravitational Waves from Compact-binary Mergers," Astrophys. J. **922**, 76 (2021), arXiv:2105.09151 [astro-ph.HE].

[21] Alexander H. Nitz, Sumit Kumar, Yi-Fan Wang, Shilpa Kastha, Shichao Wu, Marlin Schäfer, Rahul Dhurkunde, and Collin D. Capano, "4-OGC: Catalog of Gravitational Waves from Compact Binary Mergers," Astrophys. J. **946**, 59 (2023), arXiv:2112.06878 [astro-ph.HE].

[22] Barak Zackay, Tejaswi Venumadhav, Liang Dai, Javier Roulet, and Matias Zaldarriaga, "Highly spinning and aligned binary black hole merger in the Advanced LIGO first observing run," Phys. Rev. D **100**, 023007 (2019), arXiv:1902.10331 [astro-ph.HE].

[23] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga, "New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO," Phys. Rev. D **100**, 023011 (2019), arXiv:1902.10341 [astro-ph.IM].

[24] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga, "New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo," Phys. Rev. D **101**, 083030 (2020), arXiv:1904.07214 [astro-ph.HE].

[25] Barak Zackay, Liang Dai, Tejaswi Venumadhav, Javier Roulet, and Matias Zaldarriaga, "Detecting gravitational waves with disparate detector responses: Two new binary black hole mergers," Phys. Rev. D **104**, 063030 (2021), arXiv:1910.09528 [astro-ph.HE].

[26] Seth Olsen, Tejaswi Venumadhav, Jonathan Mushkin, Javier Roulet, Barak Zackay, and Matias Zaldarriaga, "New binary black hole mergers in the LIGO-Virgo O3a data," Phys. Rev. D **106**, 043009 (2022), arXiv:2201.02252 [astro-ph.HE].

[27] Ajit Kumar Mehta, Seth Olsen, Digvijay Wadekar, Javier Roulet, Tejaswi Venumadhav, Jonathan Mushkin, Barak Zackay, and Matias Zaldarriaga, "New binary black hole mergers in the LIGO-Virgo O3b data," (2023), arXiv:2311.06061 [gr-qc].

[28] Digvijay Wadekar, Javier Roulet, Tejaswi Venumadhav, Ajit Kumar Mehta, Barak Zackay, Jonathan Mushkin, Seth Olsen, and Matias Zaldarriaga, "New black hole mergers in the LIGO-Virgo O3 data from a gravitational wave search including higher-order harmonics," (2023), arXiv:2312.06631 [gr-qc].

[29] J. Aasi *et al.* (LIGO Scientific), "Advanced LIGO," Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[30] F. Acernese *et al.* (VIRGO), "Advanced Virgo: a second-generation interferometric gravitational wave detector," Class. Quant. Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[31] T. Akutsu *et al.* (KAGRA), "Overview of KAGRA: Detector design and construction history," PTEP **2021**, 05A101 (2021), arXiv:2005.05574 [physics.ins-det].

[32] J. Veitch *et al.*, "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library," Phys. Rev. D **91**, 042003 (2015), arXiv:1409.7215 [gr-qc].

[33] Eric Thrane and Colm Talbot, "An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models," Publ. Astron. Soc. Austral. **36**, e010 (2019), [Erratum: Publ.Astron.Soc.Austral. 37, e036 (2020)],

arXiv:1809.02293 [astro-ph.IM].
[34] Gregory Ashton *et al.*, "BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy," Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].
[35] I. M. Romero-Shaw *et al.*, "Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue," Mon. Not. Roy. Astron. Soc. **499**, 3295–3319 (2020), arXiv:2006.00714 [astro-ph.IM].
[36] Scott E. Field, Chad R. Galley, Jan S. Hesthaven, Jason Kaye, and Manuel Tiglio, "Fast prediction and evaluation of gravitational waveforms using surrogate models," Phys. Rev. X **4**, 031006 (2014), arXiv:1308.3565 [gr-qc].
[37] James Healy, Carlos O. Lousto, Jacob Lange, and Richard O'Shaughnessy, "Application of the third RIT binary black hole simulations catalog to parameter estimation of gravitational waves signals from the LIGO-Virgo O1/O2 observational runs," Phys. Rev. D **102**, 124053 (2020), arXiv:2010.00108 [gr-qc].
[38] P. Ajith *et al.*, "Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins," Phys. Rev. Lett. **106**, 241101 (2011), arXiv:0909.2867 [gr-qc].
[39] Mark Hannam, Patricia Schmidt, Alejandro Bohé, Leïla Haegel, Sascha Husa, Frank Ohme, Geraint Pratten, and Michael Pürrer, "Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms," Phys. Rev. Lett. **113**, 151101 (2014), arXiv:1308.3271 [gr-qc].
[40] Geraint Pratten, Sascha Husa, Cecilio Garcia-Quiros, Marta Colleoni, Antoni Ramos-Buades, Hector Estelles, and Rafel Jaume, "Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes," Phys. Rev. D **102**, 064001 (2020), arXiv:2001.11412 [gr-qc].
[41] Cecilio García-Quirós, Marta Colleoni, Sascha Husa, Héctor Estellés, Geraint Pratten, Antoni Ramos-Buades, Maite Mateu-Lucena, and Rafel Jaume, "Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries," Phys. Rev. D **102**, 064002 (2020), arXiv:2001.10914 [gr-qc].
[42] Geraint Pratten *et al.*, "Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes," Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].
[43] Héctor Estellés, Marta Colleoni, Cecilio García-Quirós, Sascha Husa, David Keitel, Maite Mateu-Lucena, Maria de Lluc Planas, and Antoni Ramos-Buades, "New twists in compact binary waveform modeling: A fast time-domain model for precession," Phys. Rev. D **105**, 084040 (2022), arXiv:2105.05872 [gr-qc].
[44] Hang Yu, Javier Roulet, Tejaswi Venumadhav, Barak Zackay, and Matias Zaldarriaga, "Accurate and efficient waveform model for precessing binary black holes," Phys. Rev. D **108**, 064059 (2023), arXiv:2306.08774 [gr-qc].
[45] Jonathan E. Thompson, Eleanor Hamilton, Lionel London, Shrobana Ghosh, Panagiota Kolitsidou, Charlie Hoy, and Mark Hannam, "PhenomXO4a: a phenomenological gravitational-wave model for precessing black-hole binaries with higher multipoles and asymmetries," Phys. Rev. D **109**, 063012 (2024), arXiv:2312.10025 [gr-qc].
[46] A. Buonanno and T. Damour, "Effective one-body approach to general relativistic two-body dynamics," Phys. Rev. D **59**, 084006 (1999), arXiv:gr-qc/9811091.
[47] Alessandra Buonanno and Thibault Damour, "Transition from inspiral to plunge in binary black hole coalescences," Phys. Rev. D **62**, 064015 (2000), arXiv:gr-qc/0001013.
[48] Thibault Damour, "Coalescence of two spinning black holes: an effective one-body approach," Phys. Rev. D **64**, 124013 (2001), arXiv:gr-qc/0103018.
[49] Andrea Taracchini *et al.*, "Effective-one-body model for black-hole binaries with generic mass ratios and spins," Phys. Rev. D **89**, 061502 (2014), arXiv:1311.2544 [gr-qc].
[50] Alessandro Nagar, Alice Bonino, and Piero Rettegno, "Effective one-body multipolar waveform model for spin-aligned, quasicircular, eccentric, hyperbolic black hole binaries," Phys. Rev. D **103**, 104021 (2021), arXiv:2101.08624 [gr-qc].
[51] Alessandro Nagar, Piero Rettegno, Rossella Gamba, Simone Albanesi, Angelica Albertini, and Sebastiano Bernuzzi, "Analytic systematics in next generation of effective-one-body gravitational waveform models for future observations," Phys. Rev. D **108**, 124018 (2023), arXiv:2304.09662 [gr-qc].
[52] Lorenzo Pompili *et al.*, "Laying the foundation of the effective-one-body waveform models SEOBNRv5: Improved accuracy and efficiency for spinning nonprecessing binary black holes," Phys. Rev. D **108**, 124035 (2023), arXiv:2303.18039 [gr-qc].
[53] Antoni Ramos-Buades, Alessandra Buonanno, Héctor Estellés, Mohammed Khalil, Deyan P. Mihaylov, Serguei Ossokine, Lorenzo Pompili, and Mahlet Shiferaw, "Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes," Phys. Rev. D **108**, 124037 (2023), arXiv:2303.18046 [gr-qc].
[54] Jonathan Blackman, Scott E. Field, Chad R. Galley, Béla Szilágyi, Mark A. Scheel, Manuel Tiglio, and Daniel A. Hemberger, "Fast and Accurate Prediction of Numerical Relativity Waveforms from Binary Black Hole Coalescences Using Surrogate Models," Phys. Rev. Lett. **115**, 121102 (2015), arXiv:1502.07758 [gr-qc].
[55] Jonathan Blackman, Scott E. Field, Mark A. Scheel, Chad R. Galley, Daniel A. Hemberger, Patricia Schmidt, and Rory Smith, "A Surrogate Model of Gravitational Waveforms from Numerical Relativity Simulations of Precessing Binary Black Hole Mergers," Phys. Rev. D **95**, 104023 (2017), arXiv:1701.00550 [gr-qc].
[56] Vijay Varma, Scott E. Field, Mark A. Scheel, Jonathan Blackman, Lawrence E. Kidder, and Harald P. Pfeiffer, "Surrogate model of hybridized numerical relativity binary black hole waveforms," Phys. Rev. D **99**, 064045 (2019), arXiv:1812.07865 [gr-qc].
[57] Vijay Varma, Scott E. Field, Mark A. Scheel, Jonathan Blackman, Davide Gerosa, Leo C. Stein, Lawrence E. Kidder, and Harald P. Pfeiffer, "Surrogate models for precessing binary black hole simulations with unequal masses," Phys. Rev. Research. **1**, 033015 (2019), arXiv:1905.09300 [gr-qc].
[58] Lalit Pathak, Amit Reza, and Anand S. Sengupta, "Fast and faithful interpolation of numerical relativ-

ity surrogate waveforms using meshfree approximation," (2024), arXiv:2403.19162 [gr-qc].
[59] Niayesh Afshordi *et al.* (LISA Consortium Waveform Working Group), "Waveform Modelling for the Laser Interferometer Space Antenna," (2023), arXiv:2311.01300 [gr-qc].
[60] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger, "Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization," SIAM Review **60**, 550–591 (2018), publisher: Society for Industrial and Applied Mathematics.
[61] Juan Calderón Bustillo, Pablo Laguna, and Deirdre Shoemaker, "Detectability of gravitational waves from binary black holes: Impact of precession and higher modes," Phys. Rev. D **95**, 104038 (2017), arXiv:1612.02340 [gr-qc].
[62] Jacob Lange, Richard O'Shaughnessy, and Monica Rizzo, "Rapid and accurate parameter inference for coalescing, precessing compact binaries," (2018), arXiv:1805.10457 [gr-qc].
[63] Prayush Kumar, Jonathan Blackman, Scott E. Field, Mark Scheel, Chad R. Galley, Michael Boyle, Lawrence E. Kidder, Harald P. Pfeiffer, Bela Szilagyi, and Saul A. Teukolsky, "Constraining the parameters of GW150914 and GW170104 with numerical relativity surrogates," Phys. Rev. D **99**, 124005 (2019), arXiv:1808.08004 [gr-qc].
[64] Yiwen Huang, Carl-Johan Haster, Salvatore Vitale, Vijay Varma, Francois Foucart, and Sylvia Biscoveanu, "Statistical and systematic uncertainties in extracting the source properties of neutron star - black hole binaries with gravitational waves," Phys. Rev. D **103**, 083001 (2021), arXiv:2005.11850 [gr-qc].
[65] Benjamin Peherstorfer, Tiangang Cui, Youssef Marzouk, and Karen Willcox, "Multifidelity importance sampling," Computer Methods in Applied Mechanics and Engineering **300**, 490–509 (2016).
[66] Ethan Payne, Colm Talbot, and Eric Thrane, "Higher order gravitational-wave modes with likelihood reweighting," Phys. Rev. D **100**, 123017 (2019), arXiv:1905.05477 [astro-ph.IM].
[67] Ethan Payne, Colm Talbot, Paul D. Lasky, Eric Thrane, and Jeffrey S. Kissel, "Gravitational-wave astronomy with a physical calibration model," Phys. Rev. D **102**, 122004 (2020), arXiv:2009.10193 [astro-ph.IM].
[68] Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf, "Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference," Phys. Rev. Lett. **130**, 171403 (2023), arXiv:2210.05686 [gr-qc].
[69] Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart, "Importance sampling: Intrinsic dimension and computational cost," Statistical Science , 405–431 (2017).
[70] Terrence Alsup and Benjamin Peherstorfer, "Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and Bayesian inverse problems," arXiv:2010.11708 [cs, math, stat] (2021), arXiv: 2010.11708.
[71] Art B. Owen, *Monte Carlo theory, methods and examples* (`https://artowen.su.domains/mc/`, 2013).
[72] Robert H. Swendsen and Jian-Sheng Wang, "Replica Monte Carlo Simulation of Spin-Glasses," Phys. Rev. Lett. **57**, 2607 (1986).
[73] Charles J. Geyer, "Markov chain monte carlo maximum likelihood," in *Proc. 23rd Symp. Interface, Computing Science and Statistics*, edited by MK Elaine and MK Selma (Interface Foundation of North America, New York, 1991).
[74] David J. Earl and Michael W. Deem, "Parallel tempering: Theory, applications, and new perspectives," Phys. Chem. Chem. Phys. **7**, 3910–3916 (2005).
[75] W. D. Vousden, W. M. Farr, and I. Mandel, "Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations," Monthly Notices of the Royal Astronomical Society **455**, 1919–1937 (2015), https://academic.oup.com/mnras/article-pdf/455/2/1919/18514064/stv2422.pdf.
[76] R. Abbott *et al.* (KAGRA, VIRGO, LIGO Scientific), "Open Data from the Third Observing Run of LIGO, Virgo, KAGRA, and GEO," Astrophys. J. Suppl. **267**, 29 (2023), arXiv:2302.03676 [gr-qc].
[77] LIGO, Virgo, and KAGRA Scientific Collaborations, "Gravitational Wave Open Science Center," `https://www.gw-openscience.org/`.
[78] Edward Higson, Will Handley, Mike Hobson, and Anthony Lasenby, "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation," Statistics and Computing **29**, 891–913 (2019), arXiv:1704.03459 [stat.CO].
[79] Joshua S. Speagle, "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences," Mon. Not. Roy. Astron. Soc. **493**, 3132–3158 (2020), arXiv:1904.02180 [astro-ph.IM].
[80] John Skilling, "Nested Sampling," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Series, Vol. 735, edited by Rainer Fischer, Roland Preuss, and Udo Von Toussaint (AIP, 2004) pp. 395–405.
[81] John Skilling, "Nested sampling for general Bayesian computation," Bayesian Analysis **1**, 833 – 859 (2006).
[82] Charlie Hoy and Vivien Raymond, "PESummary: the code agnostic Parameter Estimation Summary page builder," SoftwareX **15**, 100765 (2021), arXiv:2006.06639 [astro-ph.IM].
[83] Ken K. Y. Ng, Salvatore Vitale, Aaron Zimmerman, Katerina Chatziioannou, Davide Gerosa, and Carl-Johan Haster, "Gravitational-wave astrophysics with effective-spin measurements: asymmetries and selection biases," Phys. Rev. D **98**, 083007 (2018), arXiv:1805.03046 [gr-qc].
[84] "Advanced ligo anticipated sensitivity curves," https://dcc.ligo.org/LIGO-T0900288/public (2015).
[85] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration, "GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run — Parameter estimation data release," (2021).
[86] B. P. Abbott *et al.* (LIGO Scientific, Virgo), "Properties of the Binary Black Hole Merger GW150914," Phys. Rev. Lett. **116**, 241102 (2016), arXiv:1602.03840 [gr-qc].
[87] Will M. Farr, Benjamin Farr, and Tyson Littenberg, "Modelling calibration errors in cbc waveforms," LIGO-

T1400682 (2014).
[88] Bradley Efron, *The jackknife, the Bootstrap and Other Resampling Plans* (SIAM, 1982) CBMS-NSF Regional Conference Series in Applied Mathematics.
[89] David W. Hogg, Jo Bovy, and Dustin Lang, "Data analysis recipes: Fitting a model to data," (2010), arXiv:1008.4686 [astro-ph.IM].
[90] Mark Hannam *et al.*, "General-relativistic precession in a black-hole binary," Nature **610**, 652–655 (2022), arXiv:2112.11300 [gr-qc].
[91] Vijay Varma, Sylvia Biscoveanu, Tousif Islam, Feroz H. Shaik, Carl-Johan Haster, Maximiliano Isi, Will M. Farr, Scott E. Field, and Salvatore Vitale, "Evidence of Large Recoil Velocity from a Black Hole Merger Signal," Phys. Rev. Lett. **128**, 191102 (2022), arXiv:2201.01302 [astro-ph.HE].
[92] Tousif Islam, Avi Vajpeyi, Feroz H. Shaik, Carl-Johan Haster, Vijay Varma, Scott E. Field, Jacob Lange, Richard O'Shaughnessy, and Rory Smith, "Analysis of GWTC-3 with fully precessing numerical relativity surrogate models," (2023), arXiv:2309.14473 [gr-qc].
[93] Ethan Payne, Sophie Hourihane, Jacob Golomb, Rhiannon Udall, Richard Udall, Derek Davis, and Katerina Chatziioannou, "Curious case of GW200129: Interplay between spin-precession inference and data-quality issues," Phys. Rev. D **106**, 104017 (2022), arXiv:2206.11932 [gr-qc].
[94] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration, "LVK Algorithm Library - LALSuite," Free software (GPL) (2018).
[95] Duncan Macleod, Scott Coughlin, Alex Southgate, Derek Davis, Matt Pitkin, rngeorge, paulaltin, Joseph Areeda, Patrick Godwin, Leo Singer, Vivien Raymond, Eric Quintero, aromerorodriguez, Thomas Massinger, Pierre Chanial, François Rozet, Evan Goetz, David Keitel, Ethan Marx, Katrin Leinweber, Martin Beroiz, and The Gitter Badger, "gwpy/gwpy: Gwpy 3.0.8," (2024).
[96] Isobel M. Romero-Shaw, Paul D. Lasky, and Eric Thrane, "Searching for Eccentricity: Signatures of Dynamical Formation in the First Gravitational-Wave Transient Catalogue of LIGO and Virgo," Mon. Not. Roy. Astron. Soc. **490**, 5210–5216 (2019), arXiv:1909.05466 [astro-ph.HE].
[97] Greg Ashton, Moritz Hübner, Paul Lasky, and Colm Talbot, "Bilby: A user-friendly bayesian inference library," (2019).
[98] Sergey Koposov, Josh Speagle, Kyle Barbary, Gregory Ashton, Ed Bennett, Johannes Buchner, Carl Scheffler, Ben Cook, Colm Talbot, James Guillochon, Patricio Cubillos, Andrés Asensio Ramos, Ben Johnson, Dustin Lang, Ilya, Matthieu Dartiailh, Alex Nitz, Andrew McCluskey, and Anne Archibald, "joshspeagle/dynesty: v2.1.3," (2023).
[99] Charlie Hoy, Vivien Raymond, Aditya Vijaykumar, Duncan Macleod, Colm Talbot, Matt Pitkin, Edward Fauchon, Gregory Ashton, Nikhil Sarin, Ian Harry, John Veitch, Nils Leif Fischer, and Sebastian Khan, "pesummary/pesummary: 0.13.0 release," (2021).
[100] Max A. Woodbury, *Inverting modified matrices* (Princeton University, Princeton, NJ, 1950) p. 4, statistical Research Group, Memo. Rep. no. 42,.