A more efficient approach to parallel-tempered Markov-chain Monte Carlo for the highly structured posteriors of gravitational-wave signals

Benjamin Farr and Vicky Kalogera

Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) & Department of Physics and Astronomy, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA

Erik Luijten

Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) & Department of Materials Science and Engineering & Department of Engineering Sciences and Applied Mathematics, Northwestern University, 2220 Campus Drive, Evanston, Illinois 60208, USA

(Received 29 September 2013; published 3 July 2014)

We introduce a new Markov-chain Monte Carlo (MCMC) approach designed for the efficient sampling of highly correlated and multimodal posteriors. Parallel tempering, though effective, is a costly technique for sampling such posteriors. Our approach minimizes the use of parallel tempering, only applying it for a short time to build a proposal distribution that is based upon estimation of the kernel density and tuned to the target posterior. This proposal makes subsequent use of parallel tempering unnecessary, allowing all chains to be cooled to sample the target distribution. Gains in efficiency are found to increase with increasing posterior complexity, ranging from tens of percent in the simplest cases to over a factor of 10 for the more complex cases. Our approach is particularly useful in the context of parameter estimation of gravitational-wave signals measured by ground-based detectors, which is currently done through Bayesian inference with MCMC, one of the leading sampling methods. Posteriors for these signals are typically multimodal with strong nonlinear correlations, making sampling difficult. As we enter the advanced-detector era, improved sensitivities and wider bandwidths will drastically increase the computational cost of analyses, demanding more efficient search algorithms to meet these challenges.

DOI: 10.1103/PhysRevD.90.024014 PACS numbers: 07.05.Kf, 04.80.Nn, 04.30.Db, 95.85.Sz

I. INTRODUCTION

In the coming years, the detectors of the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo Collaboration (LVC) will come online following a multiyear endeavor to upgrade the instruments. This so-called "advanced-detector era" will ultimately bring a projected factor of 10 increase in range and a broadened band of sensitivity reaching down to 10 Hz from the previous era's lower limit of 40 Hz [1,2]. This additional sensitivity at lower frequencies makes the detectors sensitive to gravitational waves (GWs) from compact binary mergers even earlier in their inspiral phase. This has a tremendous impact on the computational cost of analyses, as waveform models become up to a factor of ∼40 longer than in previous analyses.

To estimate the parameters of a GW source, the LVC parameter estimation (PE) algorithms (found in the LALINFERENCE software package [3,4]) compute $\sim 10^7$–10^8 model waveforms that are compared to the interferometric data. Because the generation of these waveforms constitutes the computational bottleneck of the analysis, the longer waveforms required for advanced LVC parameter estimation will increase analysis times by up to a factor of ∼50. PE analyses required several hours to several days to analyze a GW candidate that entered the band of sensitivity at 40 Hz [3]. Without further optimization the analysis of individual GW candidates in the advanced detector era will become prohibitively long, requiring improvements to both model waveforms and PE algorithms. The present work addresses inefficiencies of the PE methods currently employed, particularly focusing on Markov-chain Monte Carlo (MCMC) methods, and offers an approach that can significantly reduce the total number of waveforms generated during a given analysis.

We propose a new analysis method that adopts a longer burn-in phase than standard MCMC, relying on parallel tempering only to produce a rough estimate of the target posterior through an approach based upon estimation of the kernel density. The resulting proposal more efficiently generates uncorrelated samples from multimodal and correlated posteriors than parallel tempering, eliminating the need for the latter and allowing all chains to sample from the target posterior. We emphasize that although this algorithm was developed to aid in the parameter estimation of GW sources, the techniques are not problem-specific, and can potentially be applied to other MCMC algorithms to increase the efficiency of estimating highly structured posteriors.

In Sec. II we give a brief introduction to the noise and the signal models for GWs from binary inspirals. Section III outlines the MCMC methods employed by

LALINFERENCE in the LVC's last science run. Section IV describes the new MCMC strategy that we have developed to increase sampling efficiency.

II. SIGNAL AND NOISE MODELS

The Bayesian PE algorithms used to analyze LVC data depend on models for both the noise and the signal. To provide context, here we briefly discuss those models. The most accurate models for GWs produced by compact binary systems are those generated by simulations that numerically solve the full nonlinear differential equations of general relativity. However, this approach is computationally far too expensive to be used for PE analyses. Therefore, in lieu of numerical waveforms, PE algorithms rely on approximate methods such as post-Newtonian expansion [5] or the effective-one-body formalism [6] to generate model waveforms for a given set of physical parameters $\boldsymbol{\theta}$.

The GWs produced by a quasicircular compact binary system of masses m_1 and m_2 are parametrized by fifteen parameters [5],

$$\boldsymbol{\theta} = \{\mathcal{M}_c, q, \boldsymbol{S}_1, \boldsymbol{S}_2, \iota, D_L, \psi, \alpha, \delta, t_c, \phi_c\}, \tag{1}$$

where $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ is the chirp mass, $q = m_2/m_1$ the asymmetric mass ratio defined such that $0 < q \leq 1$, $\boldsymbol{S}_i$ the spin vector of the ith binary component, ι the inclination of the orbital plane relative to the observer's line of sight, D_L the luminosity distance, ψ the polarization angle, α the right ascension, δ the declination, t_c the time of coalescence, and ϕ_c the phase at coalescence. For the purposes of this work, we will focus only on mergers of nonspinning compact objects where $|\boldsymbol{S}_1| = |\boldsymbol{S}_2| = 0$, reducing the parameter space to nine dimensions.

The current noise model used for PE analyses assumes the noise to be stationary and Gaussian, with a power spectral density that is estimated via Welch's method [7] near the time of interest (i.e., trigger time) [3]. However, real detector noise is often nonstationary and non-Gaussian, with occasional glitches and nonstationarities not currently accounted for in the noise model [8] that can potentially bias parameter estimates. More sophisticated noise modeling is outside the scope of this work, but remains an area of active research [9,10].

III. MCMC TECHNIQUES AND PARAMETER ESTIMATION

The posterior probability $p(\boldsymbol{\theta}|d)$ of the parameter set $\boldsymbol{\theta}$ given the data d is calculated according to Bayes's theorem, a framework for updating prior information $\pi(\boldsymbol{\theta})$ based on newly measured data,

$$p(\boldsymbol{\theta}|d) = \frac{\pi(\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})}{p(d)}, \tag{2}$$

where the likelihood $\mathcal{L}(\boldsymbol{\theta}) = p(d|\boldsymbol{\theta})$ is the probability of measuring the data d given the parameter set $\boldsymbol{\theta}$, and $p(d)$ is the marginal likelihood. The likelihood function for a detector network is given by the product of individual detector likelihoods [11],

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_i \exp\left[-2\int_0^\infty \frac{|\tilde{d}_i(f) - \tilde{h}_i(f,\boldsymbol{\theta})|^2}{S_{n,i}(f)} df\right], \tag{3}$$

where $\tilde{d}_i(f)$, $\tilde{h}_i(f,\boldsymbol{\theta})$, and $S_{n,i}(f)$ are the ith detector's data, modeled signal, and one-sided noise power spectral density, respectively, in the frequency domain.

To define our formalism and notation we briefly summarize the basics of Bayesian analysis and MCMC methodology. The posterior distributions (2) of compact binary GW signals in the LVC are typically estimated using multiple sampling algorithms. Nested sampling [12], MultiNest [13], and MCMC [14,15] have all proven to be effective sampling techniques. Here we introduce several improvements aimed at the MCMC approach, but some of the proposed techniques (in particular the tuned jump proposal outlined in Sec. IV A) may improve the efficiency of other sampling techniques as well.

MCMC methods produce samples at a density proportional to that of the target posterior distribution by constructing a Markov chain whose equilibrium distribution is proportional to the posterior distribution. Our MCMC implementation uses the Metropolis-Hastings algorithm [16,17], which requires a proposal density $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ to generate a new sample $\boldsymbol{\theta}'$ given the current sample $\boldsymbol{\theta}$. Such a proposal is accepted with a probability $r_s = \min(1, \alpha)$, where

$$\alpha = \frac{Q(\boldsymbol{\theta}|\boldsymbol{\theta}')p(\boldsymbol{\theta}'|d)}{Q(\boldsymbol{\theta}'|\boldsymbol{\theta})p(\boldsymbol{\theta}|d)}. \tag{4}$$

If accepted, $\boldsymbol{\theta}'$ is added to the chain; otherwise, $\boldsymbol{\theta}$ is repeated.

Chains are typically started at a random location in parameter space, requiring some number of iterations before dependence on this location is lost. The samples collected during this burn-in period are necessary, but do not provide useful data, as they are typically discarded when the posterior is estimated. Furthermore, adjacent samples in the chain are usually correlated, requiring the chain to be "thinned" by its integrated autocorrelation time (ACT). We refer to the samples remaining after discarding burn-in and thinning by the ACT as the effective samples.

The efficiency of the Metropolis-Hastings algorithm is largely dependent on the choice of proposal density, since that is what governs the acceptance rates and ACTs. The most commonly used proposal density is a Gaussian centered on $\boldsymbol{\theta}$. The width of this Gaussian for each parameter will affect the acceptance rate of the proposal. Widths that are too large will cause low acceptance rates,

whereas widths that are too small will lead to strongly correlated samples and large ACTs. For the idealized case of a posterior on $\mathbb{R}^d$ composed of independent and identically distributed components such that $p(\theta_1, \theta_2, \ldots, \theta_d) = f(\theta_1)f(\theta_2)\ldots f(\theta_d)$, where f is a one-dimensional smooth density, it can be shown that the optimal acceptance rate is approximately 23.4% [18,19]. This value, applicable only to the local Gaussian jump proposal, provides the optimum balance between acceptance rate and ACT. In principle, proposals can achieve arbitrarily high acceptance rates and yet produce uncorrelated samples—as shown in the context of MCMC schemes in statistical mechanics [20–22]. Nevertheless, we find that for typical situations in GW data analysis, targeting an acceptance rate of 23.4% allows for relatively consistent ACTs for all posteriors. Therefore, during the burn-in period we scale the one-dimensional Gaussian widths of all proposal densities to approximately achieve this acceptance rate. This adaptation is limited to the first $\sim 10^5$ likelihood evaluations, and is removed after the burn-in.

Gaussian jump proposals are typically sufficient for unimodal posteriors and spaces without strong correlations between parameters. However, there are many situations where strong parameter correlations exist and/or multiple isolated modes appear spread across the multidimensional parameter space. When parameters are strongly correlated, the ideal jumps would be along these correlations. This makes the one-dimensional jumps in the model parameters very inefficient. Furthermore, to sample between isolated modes, a chain must make a large number of jumps through regions of low probability. To properly weigh these modes, a Markov chain must alternate between them frequently. Two commonly used techniques to achieve this are parallel tempering (PT) and differential evolution.

A. Parallel tempering

Tempering introduces a "temperature" T into the likelihood function, resulting in a modified posterior,

$$p_T(\theta|d) \propto \pi(\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})^{\frac{1}{T}}. \tag{5}$$

Increasing temperatures above $T = 1$ reduces the contrast of the likelihood surface, shortening and broadening the peaks in the distribution and making them easier to sample. PT originates from Monte Carlo simulations in condensed-matter physics, starting from replica-exchange Monte Carlo [23] and then generalized to the full exchange of "configurations" [24] (cf., Ref. [25] for a review). It exploits the "flattening" of the distributions with increasing temperature to construct an ensemble of tempered chains with temperatures spanning $T = 1$ to some maximum temperature $T_{\max}$. Chains at higher temperatures are more likely to accept jumps to lower posterior values and hence more likely to explore parameter space and move between isolated modes. Regions of higher posterior value found by the high-temperature chains are then passed down through the temperature ensemble via swaps between chains at adjacent temperatures. Such swaps are proposed periodically and accepted at a rate $r_s = \min(1, \omega_{ij})$, where

$$\omega_{ij} = \left(\frac{\mathcal{L}(\boldsymbol{\theta}_j)}{\mathcal{L}(\boldsymbol{\theta}_i)}\right)^{\frac{1}{T_i}-\frac{1}{T_j}}, \tag{6}$$

with $T_i < T_j$. This technique greatly increases the probability of the $T = 1$ chain sampling between modes, but does so by creating many additional chains whose samples are ultimately discarded, since they are not drawn from the target posterior. In our calculations, the temperatures T_i are distributed logarithmically. Every 100 iterations, swaps are proposed sequentially between adjacent chains starting from the highest-temperature pair. All runs using the standard PT approach are done using eight chains, consistent with the analyses conducted during the last LVC science run [3]. It should be noted that this approach subtly violates detailed balance, as the location of a hot chain can be passed to the $T = 1$ chain in a single sequence of swaps, but the reverse is not possible. Tests carried out as part of the work in [3] showed this to have no measurable effect on posterior estimates, thus we use the same parallel swapping method here for consistency.

B. Differential evolution

Differential evolution attempts to solve the multimodal sampling problem by leveraging information gained previously in the run [26]. It does so by drawing two previous samples $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ from the chain and proposing a new sample $\boldsymbol{\theta}'$ according to

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \gamma(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1), \tag{7}$$

where γ is a free coefficient. Fifty percent of the time we use this as a mode-hopping proposal, with $\gamma = 1$. In the case where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}$ are in the same mode, this proposes a sample from the mode containing $\boldsymbol{\theta}_2$. The other 50% of the time we choose γ uniformly between 0 and 1 to sample along correlations. This proves useful when linear correlations are encountered, but performs poorly on nonlinear correlations.

C. Previous implementation

The MCMC implementation used during the last LVC science run by LALINFERENCE employed a combination of PT and differential evolution [4]. Eight tempered chains were typically employed, with computation time per chain ranging from several hours to 1–2 weeks, depending on the waveform model used. Although this approach proved effective at sampling multimodal distributions, it required up to several thousand CPU hours for a single run due to the number of samples collected at $T > 1$ that did not contribute to the estimation of the posterior.

IV. PARALLEL-TEMPERED TUNING

We propose a pragmatic approach to address the high computational cost associated with the conventional PT implementation. PT is effective at proposing jumps between isolated modes, but requires $n_{\text{temps}} - 1$ additional likelihood evaluations for each sample in the $T = 1$ chain, where n_{temps} is the number of parallel chains. Differential evolution is a computationally less expensive method to propose intermodal jumps, but the differential evolution buffer (i.e., sampling history) must first be filled with samples across the posterior. Even if the history of a chain represents a perfect sampling of the posterior, there is only a probability $(n-1)/n^2$ of drawing an intermodal jump vector originating from the mode the chain is currently in, for the case of a posterior with n modes of equal weight.

To remedy this situation, we use parallel tempering only during the burn-in phase. The purpose of this short PT phase is to allow the $T = 1$ chain to collect samples from each of the isolated modes of the posterior. Once collected, these samples are used to produce a specialized jump proposal that is tuned to the target posterior. This new proposal eliminates the need for PT chains, thus the $T > 1$ chains can be cooled to $T = 1$, where they sample independently using the tuned proposal. Figure 1 shows a rough schematic of this PT-tuned approach.

A. PT-tuned jump proposal

Central to this approach is a method for producing a proposal distribution from the samples collected during the PT burn-in phase. A kernel-density estimator (KDE) is an obvious choice, as it produces a continuous distribution from a sample set, and is trivial to draw samples from. Given a set of samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ drawn from the target distribution f, its Gaussian KDE is given by

$$\hat{f}_h(\boldsymbol{x}) \propto \sum_{i=1}^{n} \exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{x}_i)^{\mathrm{T}} \cdot \Sigma^{-1} \cdot (\boldsymbol{x}-\boldsymbol{x}_i)}{2h^2}\right), \tag{8}$$

where Σ is the covariance of the sample and h is the bandwidth, which we have defined using Scott's Rule [27] to be

$$h = n^{-1/(d+4)}, \tag{9}$$

with d the number of dimensions. A sample is drawn from the estimated distribution by first drawing a point $\boldsymbol{x}_i$ from the sample, then drawing a point from a Gaussian centered on that point with covariance Σ.

However, KDEs tend to artificially broaden the modes of multimodal distributions, which results in a poor estimate of the posterior and in low proposal acceptance rates. To avoid such "over-smoothing" we first cluster the collected samples, identifying isolated areas of high posterior density and effectively partitioning the parameter space into subspaces. With each partition containing a single mode of the posterior, an individual KDE can be used to estimate the posterior in each partition with little over-smoothing. These individual KDEs are weighted by the fraction of total samples contained within the partition, then combined to produce a single estimate of the posterior distribution across the full parameter space.

For the clustering step we have elected to partition the PT samples using OPTICS ("Ordering Points To Identify the Clustering Structure") [28], a density-based algorithm designed to order a set of samples based on their density in parameter space. A tree-based method [29] is used to extract the clustering structure from this ordering. Before clustering, the data is scaled in each dimension by the standard deviation of the data in that dimension, to remove the variation in scales between different dimensions. More sophisticated distance measures such as the Mahalanobis distance [30] were also tested, but did not yield significant improvements compared to the normalized Euclidean distance method. This approach does not require the number of clusters to be known *a priori*, nor does it expect clusters to follow a particular distribution. The only input parameters required are the maximum distance ε to determine nearest neighbors, and the minimum number of points N_{min} defining a cluster. Once the clustering tree is determined, each "leaf" is treated as a partition for which a KDE is calculated.

A sample is generated from this proposal density by first drawing a leaf from the tree, where leaf c is drawn with a probability

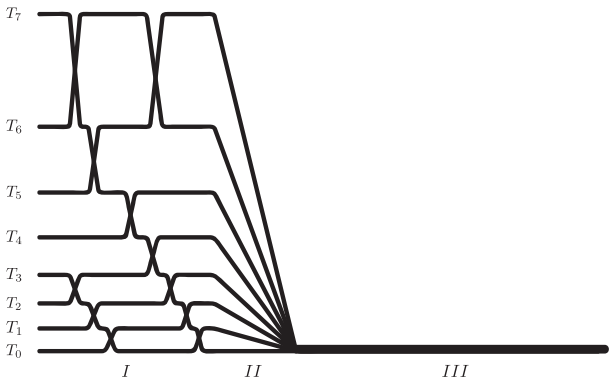

FIG. 1. Schematic of the parallel-tempering tuned approach. Each line represents a chain with temperature increasing vertically. Swaps in chain locations show when PT is in effect. Phase I is the parallel-tempered burn-in which ends after ~500 effective samples, at which point the $T = 1$ chain shares its differential evolution buffer with the other chains and the specialized proposal is tuned using its samples. During phase II the $T > 1$ chains are linearly annealed to $T = 1$ over the course of ~10 ACTs. Phase III produces all samples used to estimate the posterior, where chains sample independently using a jump proposal optimized to the target posterior.

$$\gamma_c = \frac{N_c}{\sum_{\ell \in \mathcal{C}} N_\ell}, \tag{10}$$

where N_ℓ is the number of samples in leaf ℓ and $\mathcal{C}$ the set of all leaves in the tree. A sample is then drawn from the estimate of the posterior in the leaf's subspace $p_c(\boldsymbol{\theta})$ as estimated by the kernel-density estimator.

To illustrate the faithfulness of the clustered KDE of a distribution compared to the simple KDE, we combined several two-dimensional (2D) Gaussians with random sizes and orientations. Samples were drawn from this "true" distribution in Fig. 2, which were then used to estimate the underlying distribution using both methods. Subsequently, samples were drawn from these estimates and compared to the true distribution. Even in this simple 2D case the clustered KDE can be seen to be remarkably more faithful to the target distribution (Fig. 2b), resulting in much higher acceptance rates for the proposal.

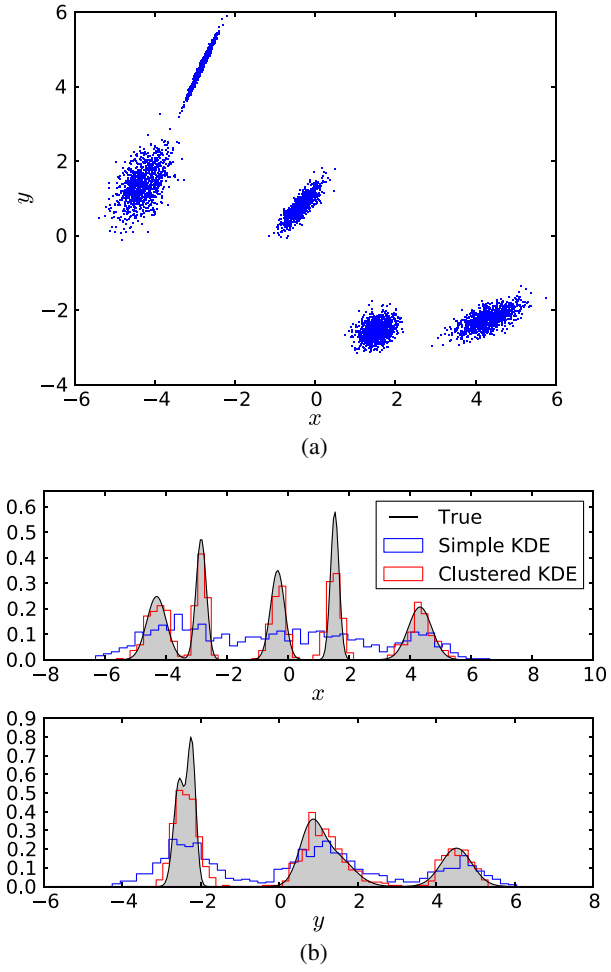

FIG. 2 (color online). An illustration of the clustered-KDE approach to estimating a distribution from a set of samples. (a) Samples drawn from a distribution composed of several 2D Gaussians with random sizes and orientations. The distribution was then estimated from these samples using both the simple KDE and the clustered-KDE methods. (b) Comparison of samples drawn from these estimates and the original set drawn from the true distribution.

To ensure that detailed balance is maintained, the forward and backward jump probabilities must be computed to determine the acceptance probability (4). In this case $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, the probability of proposing a jump to $\boldsymbol{\theta}'$ from $\boldsymbol{\theta}$, is given by

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}') = \sum_{\ell \in \mathcal{C}} \gamma_\ell p_\ell(\boldsymbol{\theta}'). \tag{11}$$

Since the jump proposal is independent of the chain's current location, proposals are never correlated. This reduces ACTs and thereby increases the effective sample size for a chain of a given length.

Lastly, the KDE is able to accurately estimate distributions with modes of arbitrary shape. This makes the proposal efficient for proposing jumps along nonlinear correlations as well, addressing a shortcoming of differential evolution.

B. Annealing for efficient use of chains

During the parallel-tempering phase, typically several hundred effective samples are collected by the $T = 1$ chain. A PT phase of ~500 effective samples has proven sufficient for GW analysis, but will likely require tuning when applied to problems with longer burn-in periods. Once all modes in the posterior have been sampled to some extent, the PT-tuned jump proposal and differential evolution buffer will propose frequent intermodal jumps. This eliminates the need for PT, allowing us to anneal all chains to $T = 1$ where they independently sample the target posterior distribution. Here we have chosen the annealing function of chain i to decrease linearly with iteration number from its original temperature T_i to $T = 1$ over the course of $100\ell_{\mathrm{PT}}$ iterations, where ℓ_{PT} is the ACT of the $T = 1$ chain during the parallel-tempering phase. The cooling rate and precise mathematical form of the cooling function were found to have no strong effect on the sampling efficiency of this approach.

Once all chains have reached $T = 1$, no further exchanges between chains are proposed. From this point onwards all chains draw samples from the target distribution, and owing to the PT-tuned jump proposal are able to do so with shorter ACTs. The set of jump proposals used for the final phase consists of 20% PT-tuned proposals, 50% differential-evolution draws, 25% Gaussian proposals, and 5% proposals that account for an exact degeneracy between ϕ_c and ψ. Testing showed this set of proposals to be effective for simulated GW data sets; however, extensive testing to find the optimal proposal set was outside the scope of this work.

V. EFFICIENCY TESTS

To achieve the most efficient analysis we must minimize the number of likelihood computations that are ultimately discarded. This means minimizing the length of chains with $T > 1$, and minimizing the ACTs of chains sampling the target posterior.

To compare the efficiency of the algorithms we define an effective sampling rate

$$r_{\text{eff}} = \frac{\sum_{i=1}^{n_{\text{chains}}} N_{\text{eff},i}}{\sum_{i=1}^{n_{\text{chains}}} N_{\text{iter},i}}, \tag{12}$$

where $N_{\text{eff},i}$ is the number of effective samples collected by chain i, and $N_{\text{iter},i}$ the total number of likelihood calculations performed for chain i. If we consider the entire ensemble of chains in a run, the effective sampling rate is the number of uncorrelated samples (at $T = 1$) divided by the total number of likelihood computations. For a run using only parallel tempering, $N_{\text{eff},i} = 0$ for $i > 1$, since only the $T = 1$ chain samples from the target distribution. For all analyses to follow, 12 chains were run in parallel. For the GW analyses, a maximum temperature T_{max} was chosen for each simulation such that $\mathcal{L}_{\text{max}}$ followed

$$\frac{1}{T_{\text{max}}} \log(\mathcal{L}_{\text{max}}) \sim 10, \tag{13}$$

where $\mathcal{L}_{\text{null}}$ is the likelihood of measuring the observed data with no signal present, and 10 was chosen to ensure that the chain with the highest temperature would be effectively sampling the prior distribution.

A. Analytical likelihoods

To test the ability of our approach to sample posteriors with correlated parameters and multiple modes we tested it on three different distributions. We chose these distributions in a 15-dimensional space to emulate the dimensionality of the parameter space that this proposal will ultimately need to handle (i.e., spinning compact binaries). The distributions used for testing consisted of (i) a multivariate Gaussian (200:1 ratio between largest and smallest widths), (ii) a bimodal distribution with two isolated multivariate Gaussians of the same shape and orientation, separated by eight standard deviations, and (iii) a Rosenbrock function [31,32],

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{d-1} [(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2]. \tag{14}$$

The performance of the proposed method is compared to that of the previous implementation (Sec. III C), which used only parallel tempering and differential evolution. In all cases the chains ran until $\sim 10^3$ effective samples were collected, not including the burn-in. With standard PT this amounts to running all chains for $\sim$1000 effective samples after burn-in, whereas for the new method each chain is run for about $1000/n_{\text{chains}}$ effective samples after annealing.

The one-dimensional marginalized posteriors recovered by both methods pass one-sample Kolmogorov-Smirnov (K-S) tests against the analytical one-dimensional functions for the unimodal and bimodal multivariate Gaussian likelihoods, and two-sample K-S tests against each other for the Rosenbrock likelihood. Acceptance rates for the PT-tuned proposal were in the range 6%–20%. This demonstrates that the new proposals are able to generate successful jumps around nontrivial likelihood surfaces without the use of parallel tempering.

Table I compares the ACTs and effective sampling rates of the two approaches for each of the tested analytical likelihoods. In all cases the PT-tuned proposal produces chains with shorter ACTs, although the improvement is minimal for the unimodal likelihood. The simple structure of a single Gaussian makes the PT burn-in criterion of $\sim$500 effective samples unnecessarily long, since no alternative modes need to be found or weighed. This minimizes the gain in efficiency possible, since the majority of the chain's total length (85% in this case) is still done in the PT phase. In practice even standard PT is not needed to sample such a simple distribution, and we merely include it here as a benchmark. Furthermore, these comparisons are sensitive to both the length of the burn-in and the total number of effective samples collected. Since the burn-in procedure of the new method contains both a parallel-tempering phase of hundreds of effective samples and the annealing phase, it is typically much more expensive. Thus, the reduction in ACT from the PT-tuned proposal must be substantial enough to warrant the burn-in, making situations in which the posterior is highly structured the best candidates for improvement. In addition, for long runs (i.e., large numbers of effective samples), the cost of the burn-in period becomes less important and even small improvements in the ACT can result in far fewer likelihood computations over the course of the run. We note that this approach provides more efficient sampling both due to a decrease in

TABLE I. Efficiency comparison between standard parallel tempering and the PT-tuned proposal for various test functions. The autocorrelation times (ACTs) and effective sampling rates r_{eff} reported for standard PT are the median values from ten runs with different random seeds. The PT-tuned ACTs are the median values from the 12 chains after burn-in.

	Standard PT		PT-tuned		
Distribution	ACT	r_{eff}	ACT	r_{eff}	$\frac{r_{\text{eff,new}}}{r_{\text{eff,old}}}$
Unimodal	280	3.3×10^{-4}	200	4.2×10^{-4}	1.26
Bimodal	850	1.3×10^{-4}	120	1.2×10^{-3}	9.02
Rosenbrock	3280	3.5×10^{-5}	470	3.4×10^{-4}	9.71

TABLE II. Efficiency comparison between standard parallel tempering and the PT-tuned approach for ten randomly selected simulated nonspinning gravitational-wave signals. The values reported for standard parallel tempering are collected from the $T = 1$ chain of an analysis with eight parallel chains. The autocorrelation times for the new method are the median values from the 12 chains after burn-in.

	Standard PT		PT-tuned		
Event	ACT	$r_{\rm eff}$	ACT	$r_{\rm eff}$	$\frac{r_{\rm eff,new}}{r_{\rm eff,old}}$
1	1300	8.5×10^{-5}	1040	1.2×10^{-4}	1.4
2	2700	4.6×10^{-5}	190	5.8×10^{-4}	13
3	2160	5.2×10^{-5}	340	3.1×10^{-4}	5.9
4	1440	7.6×10^{-5}	430	3.2×10^{-4}	4.2
5	4220	2.8×10^{-5}	5500	9.4×10^{-5}	3.3
6	840	1.3×10^{-4}	270	2.4×10^{-4}	1.9
7	1540	7.8×10^{-5}	2030	9.8×10^{-5}	1.3
8	560	1.8×10^{-4}	200	2.7×10^{-4}	1.5
9	1460	7.9×10^{-5}	300	3.6×10^{-4}	4.6
10	960	7.9×10^{-5}	310	3.0×10^{-4}	3.9

ACT and due to the fact that all chains contribute to the posterior.

B. Simulated GW data

To ensure that our findings are relevant to the analysis of GW signals, we performed additional comparisons using simulated Gaussian detector noise containing simulated gravitational-wave signals. GW signals were generated using the TaylorF2 template family [33]. For this work we only included nonspinning compact binary mergers, restricting the parameter space to nine dimensions. Ten different signals were tested with total masses ranging from 1.4 to 12.8 $M_\odot$, and network signal-to-noise ratios between 12.5 and 63.4. Again, we found that comparisons of the estimated posteriors pass the two-sided K-S test and are consistent with the injected signal. The clustering phase identified a varying number of clusters when estimating the posterior, ranging from 2 to 12 clusters. Acceptance rates for the subsequent PT-tuned proposal fell in the range of ~0.2%–18%, compared to the ~0.2%–10% achieved from differential evolution, and ~22%–25% from the local one-dimensional Gaussian proposals. The 0.2% acceptance rate was an outlier, with typical acceptance rates on the upper end of this range. However, even with such a low acceptance rate the run still outperformed the standard PT approach.

Table II compares the ACTs and effective sampling rates of the two methods for ten randomly selected nonspinning simulated gravitational-wave signals. The PT-tuned approach shows improved efficiencies over standard PT for all "events," but unlike our findings for the analytical likelihoods, here the ACTs of the chains are not always smaller for the PT-tuned proposal. Nevertheless, in these cases the cost due to longer ACTs is still compensated by the fact that all chains contribute to the effective sampling. The wide range in efficiency improvement, from factors of 1.3 to 13 in these tests, can be attributed to several factors. The complexity of the posterior depends strongly on the injection parameters, ranging from unimodal with little correlation to multimodal and highly correlated. Moreover, for some runs the posterior was poorly estimated when constructing the jump proposal, a natural consequence of estimating the posterior before it has been exhaustively sampled. This is the cause for the small efficiency improvement seen for event 7, which had proposal acceptance rates of ~0.2%–0.3%. Yet, despite these low acceptance rates, the chains were still able to sample the whole posterior, and did so more efficiently than standard PT.

VI. SUMMARY

We have presented an alternative implementation to standard parallel tempering, designed to minimize the time spent on parallel tempering, thus avoiding likelihood computations not contributing to the estimation of the posterior. By parallel tempering only long enough for the $T = 1$ chain to identify the modes of the posterior, a jump proposal can be tuned to the target posterior. These tuned jump proposals allow sampling of multiple modes and/or along nonlinear correlations without the continued use of parallel tempering, while also reducing ACTs. These benefits come with the trade-off of a more expensive burn-in process than traditional parallel tempering. However we find that for highly structured (i.e., non-Gaussian) likelihood surfaces the proposed approach proves to be worth this cost. The gains in efficiency increase with the complexity of the posterior, making us optimistic that the more complex posteriors encountered in the parameter space of spinning compact binary mergers will see even larger increases in sampling efficiency. However, due to the computational cost of spinning analyses we have not included them in this study, leaving them as the subject of future work.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grants No. PHY-0969820 (V. K. and E. L.) and No. DGE-0824162 (B. F.). V. K. also thanks the Aspen Center for Physics for its hospitality while she worked on this project.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), arXiv:1304.0670.
[2] G. M. Harry and LIGO Scientific Collaboration, Classical Quantum Gravity **27**, 084006 (2010).
[3] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese *et al.*, Phys. Rev. D **88**, 062001 (2013).
[4] LIGO Scientific Collaboration and Virgo Collaboration Parameter Estimation Group (to be published).
[5] L. Blanchet, Living Rev. Relativity **9**, 4 (2006).
[6] A. Buonanno and T. Damour, Phys. Rev. D **59**, 084006 (1999).
[7] P. D. Welch, IEEE Trans. Audio Electroacoust. **15**, 70 (1967).
[8] L. Blackburn, L. Cadonati, S. Caride, S. Caudill, S. Chatterji *et al.*., Classical Quantum Gravity **25**, 184004 (2008).
[9] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **82**, 103007 (2010).
[10] T. B. Littenberg, M. Coughlin, B. Farr, and W. M. Farr, Phys. Rev. D **88**, 084044 (2013).
[11] L. S. Finn and D. F. Chernoff, Phys. Rev. D **47**, 2198 (1993).
[12] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010).
[13] F. Feroz, M. P. Hobson, and M. Bridges, Mon. Not. R. Astron. Soc. **398**, 1601 (2009).
[14] M. van der Sluys, V. Raymond, I. Mandel, C. Röver, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, Classical Quantum Gravity **25**, 184011 (2008).
[15] M. V. van der Sluys, C. Röver, A. Stroeer, V. Raymond, I. Mandel, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, Astrophys. J. Lett. **688**, L61 (2008).
[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[17] W. K. Hastings, Biometrika **57**, 97 (1970).
[18] A. Gelman, G. O. Roberts, and W. R. Gilks, in *Bayesian Statistics, 5 (Alicante, 1994)*, Oxford Science Publications (Oxford University Press, New York, 1996), p. 599.
[19] G. O. Roberts and J. S. Rosenthal, Can. J. Stat. **26**, 5 (1998).
[20] J. Liu and E. Luijten, Phys. Rev. Lett. **92**, 035504 (2004).
[21] J. Liu and E. Luijten, Phys. Rev. E **71**, 066701 (2005).
[22] D. W. Sinkovits, S. A. Barr, and E. Luijten, J. Chem. Phys. **136**, 144111 (2012).
[23] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[24] C. Geyer, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (Interface Foundation, Fairfax, VA, 1991), p. 156.
[25] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).
[26] C. J. F. Ter Braak, Stat. Comput. **16**, 239 (2006).
[27] D. W. Scott, in *Multivariate Density Estimation* (Wiley, New York, 2008), p. 125.
[28] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99 (ACM, New York, 1999) p. 49.
[29] J. Sander, X. Qin, Z. Lu, N. Niu, and A. Kovarsky, in *Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD'03 (Springer, Berlin, 2003), p. 75.
[30] P. C. Mahalanobis, in *Proceedings of the National Institute of Sciences of India*, Vol. 2 (Indian National Science Academy, New Delhi, 1936), p. 49.
[31] H. H. Rosenbrock, Computer Journal (UK) **3**, 175 (1960).
[32] L. C. W. Dixon and D. J. Mills, J. Optim. Theory Appl. **80**, 175 (1994).
[33] L. Blanchet, T. Damour, G. Esposito-Farèse, and B. R. Iyer, Phys. Rev. Lett. **93**, 091101 (2004).