Adaptive Parallel Tempering Markov chain Monte Carlo algorithm for compact binary gravitational wave parameter estimation

2316440[1]

[1] *School of Physics and Astronomy*
University of Glasgow

ABSTRACT

This work develops a simple adaptive (AP) Parallel Tempering Markov chain Monte Carlo (PTMCMC) algorithm for the purposes of sampling compact binary gravitational wave parameter posterior probability distributions, responding to the problems of their potential multimodality, existence of local extrema and the a priori unknown correlation structure. The algorithm is tested on 3 multivariate analytical distributions (unimodal correlated Gaussian, bimodal uncorrelated Gaussian, Rosenbrock's banana function) and subsequently applied to a range of simulated non-eccentric spinless binary black hole signals of different signal-to-noise ratio, injected into a two-detector interferometer network with simulated frequency domain stationary Gaussian noise. Samples from the analytical distributions show no statistically significant deviations from the targets under the Kolmogorov-Smirnov test at 5% significance level. The algorithm recovers the compact binary injection parameters within the 90% credible intervals of their approximated marginal posterior distributions for all simulated signals. For the injections with network signal-to-noise ratio above 5, the bulk of the probability mass of the sampled marginal posteriors is enclosed within a well-localised region of high probability density.

Keywords: Markov chain Monte Carlo — gravitational wave astronomy

1. INTRODUCTION

Since the first detection of a gravitational wave (GW) signal from a binary black hole (BBH) merger in September 2015 [1] the LIGO Scientific Collaboration and Virgo Collaboration (LVC) have documented further ~ 50 compact binary coalescence signals during its first three observing runs (O1, O2, O3a) [2; 3]. This number is predicted to grow [4] as the Advanced LIGO and Virgo instruments are undergoing sensitivity improvements and the KAGRA detector in Japan is expected to join the search in O4 [3], while planning continues for LIGO-India [5] and the space-based interferometer LISA [6].

Detection of the transient GW signals has opened several new domains of astrophysical inference. Population analysis of the compact binary sources allowed for the study of BBH merger rates and parameter distributions [7], while analysis of the detected waveforms permitted for strong-field regime tests of Einstein's theory of General Relativity (GR) by means such as quantifying deviations from the post-Newtonian coefficients of the waveform models predicted by GR [8; 9]. The GW170817 [10] binary neutron star (BNS) merger event with its detected electromagnetic counterpart together with well-localised BBH signals were used as standard sirens in cosmological inference of the Hubble constant [11]; measurements similar to this might have a future potential of resolving the current Hubble tension [12]. More cutting edge research awaits as the search continues for a GW background and continuous GW signals from pulsars [13].

All of the aforementioned studies performed with the use of GW transients ultimately rely on interferometer data analysis and source parameter estimation. In the event of a gravitational wave passing through the interferometer network, the recorded data undergoes a twofold process of analysis. Firstly, the GW signal has to be identified within the noisy detector output - this currently involves cross-correlating the data with members of a large dataset of model waveform templates in a technique called matched filtering [14]. Once a data segment containing a candidate signal is identified, the focus of the analysis moves to model selection and parameter estimation [15; 16], which ultimately aims to recover the marginal probability density functions (PDFs) of the a priori unknown source parameters describing the GW signal buried under persistent interferometer noise. Relying on the ability to describe the interferometer noise and equipped with numerical approximations of the model waveforms predicted by GR, the so called posterior PDF of the source parameters is recovered within a Bayesian framework.

In the case of a non-eccentric binary, full characterisation of the source involves as many as 15 free parameters [16]. The process of marginalising the fully dimensional joint posterior PDF in order to retrieve its 1-dimensional kernels is notorious for involving computationally expensive multivariate integrals, impossible to be solved analytically or even accurately approximated with a brute force numerical approach of summing over a discrete grid within any conceivable time[1]. Although there is active research aimed at overcoming this problem by performing GW parameter inference using deep learning [17; 18], to date GW parameter estimation is carried out using stochastic sampling of the posterior PDF with methods such as Markov chain Monte Carlo (MCMC) [16; 19] or Nested Sampling [16]. Generally speaking, stochastic sampling is aimed at obtaining a finite sample of points belonging to the parameter space, drawn from a target distribution (in the GW case the target is the joint parameter posterior). The problem of finding the marginal distributions is then reduced to simply projecting the histogram of the samples onto each of the posterior dimensions, while marginal expectation values of the parameters (or functions thereof) are approximated by arithmetic averages over the sampled points.

In the case of MCMC algorithms, the draw from the target distribution is obtained by performing a biased random walk within the parameter space [20]. The algorithm begins by initialising a 'walker' at a random starting location within the parameter space. From then on, state updates ('jumps' to different locations within the domain of the PDF) are proposed and subsequently accepted or rejected depending on the shape of the PDF. Provided the jump proposal distribution and the transition probability meet sufficient requirements, the chain of accepted states is a sample drawn from the target distribution. The simplest MCMC formulation, the Metropolis-Hastings algorithm [21], requires as little as the ability to compute ratios of the target PDF at pairs of locations within the parameter space. This proves particularly useful in many Bayesian scenarios, where the target posterior is known only up to a difficult to compute normalising factor (the so called *evidence*), as is the case in the GW inference problem.

There are two main issues that arise from implementing the standard Metropolis-Hastings algorithm for the purposes of parameter estimation. Firstly, in order to ensure efficient sampling, the jump proposal distribution has to be scaled adequately to the target PDF, which has an a priori unknown scale and correlation structure. Secondly, the algorithm performs poorly on multimodal distributions (and in fact on any distributions with pronounced local extrema) as traversing the low probabilty regions between seperated modes requires a sequence of unlikely transitions.

Many advanced sampling algorithms addressing these problems have been developed [19]. Scaling the proposal distribution is usually solved by introducing adaptive algorithms [22] which can adjust the proposal as more is learnt about the target from the chain of accepted samples. A particularly interesting solution to the issue of multimodality comes from approaches which view the target PDF as an energy distribution, including ingenious methods such as Hamiltonian Monte Carlo (HMC) [23] or Parallel Tempering (PTMCMC) [24]. The HMC sampler moves according to simulated Hamiltonian dynamics - provided the walker is equipped with enough momentum a local minimum can be escaped. Parallel Tempering solves the problem of local minima by exchanging states between multiple chains run in parallel, each of which samples the PDF at a different temperature; low probability (high energy) states become more accessible as the temperature of the system increases. Both these methods with various adaptations have been implemented for GW posterior sampling.

This report is devoted to developing and testing a simple Python-based adaptive PTMCMC algorithm targeted at sampling posterior PDFs of compact binary transient gravitational wave parameters. This is done largely for didactic purposes (for the author's own education), given that more advanced PTMCMC based techniques are already in use within the *LALInference* software package [16] developed for this exact purpose. However, the algorithm which is put forward takes a fundamentally different approach to proposal scaling, by implementing the Adaptive Proposal [25] method, which is not switched off after the initial sampling period (burn-in). Even though such adaptation might violate the ergodicity of the process (the sampled distribution might differ from the target), with the use of methods such as mode tracing [26] it could be relatively simply transformed into an efficient vanishing adaptation which could be of interest in both GW parameter estimation and other sampling problems.

Section 2 outlines the principles of GW parameter estimation necessary to formulate the posterior PDF of the compact binary source parameters, which is the intended target distribution of the devised sampling algorithm. Section 3 introduces the PTMCMC algorithm developed for this problem. Section 4 is devoted to validating the algorithm on analytical distributions. Section 5 presents the parameter estimation results obtained by testing the algorithm on

[1] This is a simple consequence of the high dimensionality of the posterior - finite grid size scales like N^D, where N is the number of gridpoints along a dimension and D is the number of dimensions. Even a very coarse grid of ~ 100 points per dimension would require 10^{30} posterior evaluations, each of which involves generating a computationally costly waveform model for a different set of source parameters.

a number of BBH signals of different signal-to-noise ratio (SNR), simulated in frequency domain and injected into a two-detector network with simulated stationary Gaussian noise. The work concludes with a short discussion (Section 6) suggesting possible routes of improving the algorithm for future use.

2. GW PARAMETER ESTIMATION

This section outlines the principles of Bayesian parameter estimation for compact binary GW sources. This includes a description the basic data model (2.1), introduction to the fundamental Bayesian framework required to recover the posterior PDF from the interferometer data (2.2) and formulation of the parameter likelihood function for a noisy data segment containing a GW signal (2.3-2.6), coherently extended to a multiple detector network. The content of this chapter is largely a synthesis of the descriptions presented in [15; 16] and in the Statistical Astronomy (STA1) lecture notes [27] prepared by Prof. Graham Woan at the University of Glasgow.

2.1. *Data model & GW parameters*

A gravitational wave passing through a laser interferometer is recorded as the fractional change in the length of the interferometer arms as a function of time t, referred to as the strain [1]. In the presence of a gravitational wave signal $h(t)$, the measured strain data $d(t)$ is a sum of the signal and the interferometer noise, $n(t)$, i.e.

$$d(t) = h(t) + n(t). \tag{2.1}$$

The functional form of the signal h is determined by parameters $\boldsymbol{\theta}$ of the coalescing compact binary source. In the case of a non-eccentric binary, the parameters [16] required to fully characterise $h(t)$ are

$$\boldsymbol{\theta} = (m_1, m_2, D_{\rm L}, \alpha, \delta, \iota, \psi, \phi_{\rm c}, t_{\rm c}, \boldsymbol{S}_1, \boldsymbol{S}_2), \tag{2.2}$$

the primary m_1 and secondary m_2 mass of the binary, luminosity distance to the source $D_{\rm L}$, source position in the sky determined by its right ascention α and declination δ, the inclination angle ι between the normal to the orbital plane of the binary and the line of sight, polarisation angle ψ, orbital phase $\phi_{\rm c}$ at a reference time $t_{\rm c}$ (eg. the time of coalescence) and two 3-dimensional spin vectors $\boldsymbol{S}_1, \boldsymbol{S}_2$ (one per each mass in the binary). This gives a total of 15 unknown parameters, which are to be estimated.

2.2. *Bayesian Inference*

Given the detected strain data containing a GW signal, the interest lies in estimating the unknown source parameters $\boldsymbol{\theta}$ or, more precisely, in determining the PDF of $\boldsymbol{\theta}$ over an assumed prior parameter range. This probability density function can be viewed [16; 15] in the Bayesian framework [27] as the *posterior probability* $p(\boldsymbol{\theta}|\boldsymbol{d}, I)$ of the hypothesis that the GW signal buried under the noise is described by the particular parameter vector $\boldsymbol{\theta}$, given the detected data $\boldsymbol{d}$.

Baye's Theorem allows us to express this posterior probability as

$$p(\boldsymbol{\theta}|\boldsymbol{d}, I) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{d}|I)}, \tag{2.3}$$

where I represents the set of background assumptions[2] (subsequently not mentioned explicitly) and

- $p(\boldsymbol{d}|\boldsymbol{\theta}, I)$ is the *likelihood* of observing the observed data, given that the signal is described by parameters $\boldsymbol{\theta}$,
- $p(\boldsymbol{\theta}|I)$ is the *prior* probability of the parameters $\boldsymbol{\theta}$, reflecting the initial state of belief in the hypothesis that the source can be parametrised by the particular vector $\boldsymbol{\theta}$, and
- $p(\boldsymbol{d}|I)$ is the *evidence*. It is the probability of detecting the data independent of $\boldsymbol{\theta}$. In the process of posterior sampling it acts effectively as a normalisation constant and hence can be omitted.

In the light of Equation 2.3, we can recover the kernel of the posterior probability density of $\boldsymbol{\theta}$, given that we can model the likelihood function. This in turn, as described in the next section, rests upon the ability to model both the detector noise and, given $\boldsymbol{\theta}$, the gravitational wave signal h.

Having obtained the likelihood function (and having specified the priors), one can in principle evaluate (up to a constant factor) the posterior probability for any vector $\boldsymbol{\theta}$ and therefore obtain a numerical value proportional to the

[2] These include for example the assumption that there exists a GW signal in the data.

probability that the source has these particular parameters. This however is not yet the desired result - one would like to know the probability densities corresponding separately to each of the parameters included in $\boldsymbol{\theta}$. In other words, we are after the *marginalised* posterior probabilities of the form

$$p(\theta_i|\boldsymbol{d}) \propto \int \cdots \int_U p(\boldsymbol{\theta}|\boldsymbol{d})\, d\theta_1 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_n, \tag{2.4}$$

where $U \subseteq \mathbb{R}^n$ corresponds to the domain of definition of parameters $\{\theta_j\}\backslash\theta_i$ and n is the dimensionality of the problem.

These (Eq. 2.4) are exactly the integrals which will be approximated by stochastic sampling of the posterior with the use of the PTMCMC algorithm described in Section 3. Given a sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{d})$, the marginal posterior probability density function for parameter θ_i evaluated at $\theta_i = b$ over a sufficiently small interval $[a, c]$ containing b will be approximately proportional to the number of samples for which θ_i falls within the given interval, i.e.

$$p(\theta_i|\boldsymbol{d})|_{\theta_i=b} \propto |\{\boldsymbol{\theta}\,|\, a \leq \theta_i \leq c \ \wedge\ \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{d})\}|, \tag{2.5}$$

which is equivalent to saying that the marginal posterior for θ_i corresponds to the projection of the sample histogram onto parameter θ_i, provided sufficiently small binning.

2.3. *Likelihood of a GW signal in a stationary Gaussian noise*

As explained in the previous subsection, in order to obtain the posterior PDF of the GW parameters (Eq. 2.3) for stochastic sampling and parameter inference, we first need to formulate the likelihood function describing the probability of observing the detected data, given a vector of source parameters $\boldsymbol{\theta}$.

Consider a vector $\boldsymbol{d} = (d_0, d_1, \ldots, d_{N-1})$ of N discrete strain measurements equally distributed over the data segment duration T measured in seconds, i.e. $d_n = d(n\Delta t)$, where $\Delta t = T/N$ [s] is the sample spacing[3]. Then the (dimensional) discrete Fourier transform of $\boldsymbol{d}$ is the complex frequency series $\tilde{\boldsymbol{d}} = (\tilde{d}_{-L}, \tilde{d}_{-L+1}, \ldots, \tilde{d}_0, \ldots, \tilde{d}_{L-1})$, where $L = N/2$ and

$$\tilde{d}_k = \Delta t \sum_{n=0}^{N-1} d_n e^{-2\pi i f_k n \Delta t}\ [\mathrm{Hz}^{-1}], \quad f_k = \frac{k}{T}\ [\mathrm{Hz}], \tag{2.6}$$

describes the contribution of the frequency f_k to the time domain data, and where for simplicity we have assumed that N is even. Analogically we can define the time domain vectors $\boldsymbol{h}, \boldsymbol{n}$, and their frequency domain counterparts $\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{n}}$, corresponding to the signal and the noise respectively.

Due to linearity of the Fourier transform and recalling the assumed data model (Eq. 2.1), it is straightforward to conclude that in each frequency bin the strain transform is equal to the sum of the transforms of the signal and the noise, i.e.

$$\tilde{d}_k = \tilde{h}_k + \tilde{n}_k. \tag{2.7}$$

By simple rearrangement, the noise in the k^{th} frequency bin is equal to the residual between the data and the signal,

$$\tilde{n}_k = \tilde{d}_k - \tilde{h}_k. \tag{2.8}$$

Under the assumption [15] that $\tilde{n}_k$ can be modelled as stationary Gaussian noise with zero mean and variance σ_k^2, given a parameter vector $\boldsymbol{\theta}$ and a corresponding model signal h, the probability of observing the data $\tilde{d}_k$ is equal to the probability that the residual $r = \tilde{d}_k - \tilde{h}_k$ can be attributed to the Gaussian noise $\tilde{n}_k$, i.e.

$$p(\tilde{d}_k|\boldsymbol{\theta}, \sigma_k^2) \propto \exp\left[-\frac{(\tilde{d}_k - \tilde{h}_k)^2}{2\sigma_k^2}\right]. \tag{2.9}$$

Further, assuming [16] that the measurements in each frequency bin are *independent*, the total probability of observing data $\boldsymbol{d}$ given parameters $\boldsymbol{\theta}$ is the product of the probabilities over all frequency bins, i.e.

$$p(\boldsymbol{d}|\boldsymbol{\theta}, \{\sigma_k^2\}) = \prod_k p(\tilde{d}_k|\boldsymbol{\theta}, \sigma_k^2) \propto \prod_k \exp\left[-\frac{(\tilde{d}_k - \tilde{h}_k)^2}{2\sigma_k^2}\right] = \exp\left[\sum_k -\frac{(\tilde{d}_k - \tilde{h}_k)^2}{2\sigma_k^2}\right]. \tag{2.10}$$

[3] Equivalently, the sampling frequency $F_s = 1/\Delta t = N/T$ [Hz].

The above is exactly the Bayesian likelihood from Eq 2.3.

It can be shown (for derivation see: Appendix in [15]) that the variance of the noise, σ_k^2 will be given by

$$\sigma_k^2 = \langle |\tilde{n}_k|^2 \rangle = \frac{T}{2} S(f_k) \ [\mathrm{Hz}^{-2}] \tag{2.11}$$

where $S(f_k)$ is the one-sided power spectral density (PSD) of the interferometer noise evaluated at the frequency of the k^{th} frequency bin.

Using the above expression for the noise variance and noting the symmetry of the positive and negative frequency components of the Fourier transform, we can rewrite the likelihood given in Eq. 2.10 as

$$p(\boldsymbol{d}|\boldsymbol{\theta}, S(f)) \propto \exp\left[-\frac{2}{T}\sum_{k>0}\frac{(\tilde{d}_k - \tilde{h}_k)^2}{S(f_k)}\right], \tag{2.12}$$

which depends only on quantities that are known (i.e. data) or can be approximated[4] (noise PSD) or modelled (GW signal). This much detail is sufficient for the purposes of the parameter estimation simulation performed as part of this work, however the reader should keep in mind that both the PSD estimation and especially the model signal generation are nontrivial problems.

2.4. *Coherent analysis in a detector network*

So far the previous chapter explained how to obtain a likelihood function for the strain measured within a single detector. This section shows how the analysis is coherently extended to a network consisting of multiple detectors.

The frequency domain strain within a detector D can be [15; 16] decomposed as

$$\tilde{\boldsymbol{h}} = F_+(\alpha, \delta, \psi, t_{\mathrm{c}})\tilde{\boldsymbol{h}}_+ + F_\times(\alpha, \delta, \psi, t_{\mathrm{c}})\tilde{\boldsymbol{h}}_\times \tag{2.13}$$

where $\tilde{\boldsymbol{h}}_+$, $\tilde{\boldsymbol{h}}_\times$ are the plus and cross polarisations of the GW signal, the same in every detector for a given set of source parameters $\boldsymbol{\theta}$, and F_+, $F_\times$ are detector specific antenna response functions.

On top of the variation introduced by the detector dependent antenna response pattern (Eq. 2.13), the signals in different detectors are time shifted with respect to one another due to variations in signal travel time from the source. The standard way of accounting for this [16] is to see the coalescence time t_{c} as the time of coalescence that would be measured by a detector located at the centre of the Earth. Then, for each detector we can calculate the respective time delay τ, related to the difference between the light[5] travel time to the centre of the Earth and to the detector. In the frequency domain this time delay corresponds to a frequency modulation, i.e.

$$\tilde{h}_\tau(f) = \tilde{h}_0(f)e^{-2\pi i f \tau} \tag{2.14}$$

where $\tilde{h}_0(f)$ is the frequency domain signal in the reference frame of the centre of the Earth (Eq. 2.13) and $\tilde{h}_\tau(f)$ is the signal in the detector with time delay τ.

Defining then $\tilde{\boldsymbol{h}}^D$ to be the frequency domain signal vector in a detector D, with components given by $\tilde{h}_k^D = \tilde{h}_\tau(f_k)$, we can treat the network coherently by extending the single detector likelihood given by Eq. 2.12 to

$$p(\{\boldsymbol{d}^D\}|\boldsymbol{\theta}, \{S^D(f)\}) \propto \prod_D p(\boldsymbol{d}^D|\boldsymbol{\theta}, S^D(f)) = \prod_D \exp\left[-\frac{2}{T}\sum_{k>0}\frac{(\tilde{d}_k^D - \tilde{h}_k^D)^2}{S^D(f_k)}\right], \tag{2.15}$$

where $\{\boldsymbol{d}^D\}$ denotes the set of data vectors detected in a detector network $\{D\}$ and $\{S^D(f)\}$ is the corresponding set of one-sided interferometer PSDs.

2.5. *Phase marginalised likelihood*

Coalescence phase, even though necessary to fully describe the signal, is a nuisance parameter of no astrophysical interest as it is a property of the model and not the event itself. It is possible to analytically marginalise the likelihood

[4] In practice the noise PSD is not known exactly and has to be estimated over a neighbouring data segment which does not contain the signal [16].

[5] Gravitational waves travel at the speed of light.

over the phase parameter prior to sampling, which is often performed in order to reduce the dimensionality and correlations within the posterior [16]. In order for this simulation to closely resemble the data analysis process at LIGO, such phase marginalisation is performed. Without going into the technical details, for which I refer the reader to [28], starting at Eq. 2.15, the phase marginalised likelihood for a detector network $\{D\}$ can be written as

$$p(\{\boldsymbol{d}^D\}|\boldsymbol{\theta}', \{S^D(f)\}) \propto \int p(\{\boldsymbol{d}^D\}|\boldsymbol{\theta}, \{S^D(f)\}) d\phi_c \propto \exp\left[-\frac{2}{T}\sum_{k>0,D}\frac{|\tilde{d}_k^D|^2 + |\tilde{h}_k^D|^2}{S^D(f_k)}\right] \mathrm{I}_0\left[\frac{4}{T}\left|\sum_{k>0,D}\frac{\tilde{d}_k^{*D}\tilde{h}_k^D}{S^D(f_k)}\right|\right] \tag{2.16}$$

where $\boldsymbol{\theta}'$ is the vector of remaining signal parameters (i.e. all but ϕ_c), $\tilde{h}$ is evaluated at $\phi_c = 0$, '$*$' denotes the complex conjugate and I_0 is the modified Bessel function of the first kind.

2.6. *Loglikelihood*

For computational convenience posterior sampling is performed in logspace. Taking logarithms in Eq. 2.3 we can write the logarithm of the phase marginalised detector network posterior as

$$\ln p(\boldsymbol{\theta}'|\{\boldsymbol{d}^D\}) \propto L + \ln p(\boldsymbol{\theta}'), \tag{2.17}$$

where $\ln p(\boldsymbol{\theta}')$ is the *logprior* and L is the phase marginalised detector network *loglikelihood*, given by the natural logarithm of Eq. 2.16. That is,

$$L \coloneqq \underbrace{-\frac{2}{T}\sum_{k>0,D}\frac{|\tilde{d}_k^D|^2 + |\tilde{h}_k^D|^2}{S^D(f_k)}}_{=A} + \ln \mathrm{I}_0\left[\underbrace{\frac{4}{T}\left|\sum_{k>0,D}\frac{\tilde{d}_k^{*D}\tilde{h}_k^D}{S^D(f_k)}\right|}_{=B}\right] = A + \ln \mathrm{I}_{0e}(B) + B, \tag{2.18}$$

where $\mathrm{I}_{0e}(x) = \exp(-|x|)\mathrm{I}_0(x)$ is the exponentially scaled modified Bessel function of the first kind, implemented[6] in order to prevent numerical overflows within the regular I_0.

Since the priors are more context specific than the general form of the likelihood, the prior choices are reviewed in Section 5, when introducing the parameters of the BBH injection simulations performed to test the sampling algorithm.

3. ADAPTIVE PTMCMC ALGORITHM

This section describes the adaptive PTMCMC algorithm developed for sampling compact binary GW posteriors. I begin with introducing the Metropolis-Hastings algorithm (3.1), which is followed by its extension to Parallel Tempering (3.2) and a description of the implemented adaptation (3.3).

3.1. *Metropolis-Hastings*

The Metropolis-Hastings algorithm, presented by W. K. Hastings in 1970 as a generalisation of a sampling algorithm proposed by N. Metropolis in 1953 [21], is a method of using Markov chains to obtain a stochastic sample from a probability distribution of interest (target PDF), developed for the purposes of efficient approximation of high dimensional integrals. It is the simplest MCMC algorithm and it is the foundation of many of the more advanced sampling methods, including Parallel Tempering.

Assume we want to obtain a set of N samples $\{\boldsymbol{\theta}_i\}_{i=1}^N$ drawn from a target distribution $\pi(\boldsymbol{\theta})$. In the case of GW parameter estimation problem, $\pi(\boldsymbol{\theta})$ is the likelihood multiplied by the prior, proportional to the posterior probability $p(\boldsymbol{\theta}|\boldsymbol{d})$ (Eq. 2.3). The algorithm begins by initialising a random walker at a location $\boldsymbol{\theta}_1$ belonging to the domain of definition of π (parameter space) and such that $\pi(\boldsymbol{\theta}_1) \neq 0$. Then, the algorithm suggests a sequence of transitions (jumps) to new locations within the parameter space drawn from the so called *proposal distribution*. Under a symmetric proposal distribution, and given the current state (most recently accepted location) of the walker is $\boldsymbol{\theta}_i$, a jump to a new location $\boldsymbol{\theta}_{i+1}$ is accepted with probability

$$p = \min\left(1, \frac{\pi(\boldsymbol{\theta}_{i+1})}{\pi(\boldsymbol{\theta}_i)}\right), \tag{3.1}$$

[6] The exponentially scaled modified Bessel function was calculated using the `i0e` function from the *Scipy.special* Python function library.

i.e. with probability 1 if the proposed location corresponds to a higher probability density, or with probability equal to the ratio of the probability densities of the two states otherwise. This transition probability is referred to as the *Metropolis criterion* and it ensures that the algorithm generates a discrete-time Markov chain whose stationary distribution is the target π, or simply that the chain of accepted samples $\{\boldsymbol{\theta}_i\}_{i=1}^N$ is a random draw form π. In practice it takes time for the chain to lose dependence on the initial position before convergence to the target is reached. This time is referred to as the *burn-in*, and the samples corresponding to this period are discarded.

Algorithm 1 shows the pseudo-code of the M-H algorithm with a multivariate Gaussian proposal distribution and logspace rejection sampling based transition, as implemented for the purposes of this work, following [29].

Algorithm 1: Metropolis-Hastings MCMC

Input: Function π which calculates the value of the target distribution $\pi(\boldsymbol{\theta})$ given $\boldsymbol{\theta}$.
Output: set of N samples $\{\boldsymbol{\theta}_i\}_{i=1}^N \sim \pi$.
Initialize $\boldsymbol{\theta}_1$ s.t. $\pi(\boldsymbol{\theta}_1) \neq 0$.
Initialize Σ, covariance matrix of the jump proposal distribution.
Set counter $i = 1$.
while $i \leq N$ **do**
  draw jump proposal $\boldsymbol{\theta}_{i+1} \sim \mathcal{N}(\boldsymbol{\theta}_i, \Sigma)$
  draw random number $r \sim \mathcal{U}(0,1)$
  if $\ln \pi(\boldsymbol{\theta}_{i+1}) - \ln \pi(\boldsymbol{\theta}_i) > \ln r$ // this is equivalent to the Metropolis criterion (Eq. 3.1)
  then
    accept proposal: $\boldsymbol{\theta}_{i+1} \leftarrow \boldsymbol{\theta}_{\rm new}$
  else
    reject proposal: $\boldsymbol{\theta}_{i+1} \leftarrow \boldsymbol{\theta}_i$
  end
  $i \leftarrow i + 1$
end

3.2. *Parallel Tempering*

Parallel Tempering [24] is an extension of the Metropolis-Hastings algorithm which encourages a fuller posterior PDF sampling by exchanging states between multiple Markov chains run in parallel, each of which is sampling a differently *tempered* posterior. *Tempering* refers to introducing a temperature T s.t.

$$p_T(\boldsymbol{\theta}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{\theta})^{\frac{1}{T}} p(\boldsymbol{d}), \tag{3.2}$$

where p_T is the tempered posterior. It is clear that if the temperature $T = 1$, p_T is the target posterior, however as $T \to \infty$, the tempered version of the posterior approaches the prior. An example of how a bimodal Gaussian distribution scales with temperature is shown in Fig. 1.

While several tempered Metropolis-Hastings chains are run in parallel, every n samples the algorithm proposes a swap between the states $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$ of adjacent chains i, j s.t. $T_i < T_j$. This swap is accepted with probability

$$p_s = \min\left(1, \frac{p(\boldsymbol{d}|\boldsymbol{\theta}_j)^{\frac{1}{T_i} - \frac{1}{T_j}}}{p(\boldsymbol{d}|\boldsymbol{\theta}_i)}\right), \tag{3.3}$$

which is an extension of the Metropolis criterion. (Appendix A explains how this transition probability relates to the concepts of statistical mechanics).

MCMC walkers initialized at higher temperatures will sample from a distribution that is closer in shape to the prior, thus making lower likelihood states more available, and hence allowing for a fuller exploration of the parameter space. Feeding higher temperature states down the temperature ladder increases the odds of escaping local maxima in the $T = 1$ chain (the chain that samples the 'true' target posterior) and promotes configurations otherwise unavailable to this chain, such as intermodal jumps. This feature makes PTMCMC algorithms particularly efficient at sampling not only multimodal distributions, but also any distributions with complicated local features (such as local extrema or inteference patterns, both of which are seen in GW posteriors), allowing for fast convergence to the regions in the probability space where the bulk of the probability mass is contained.

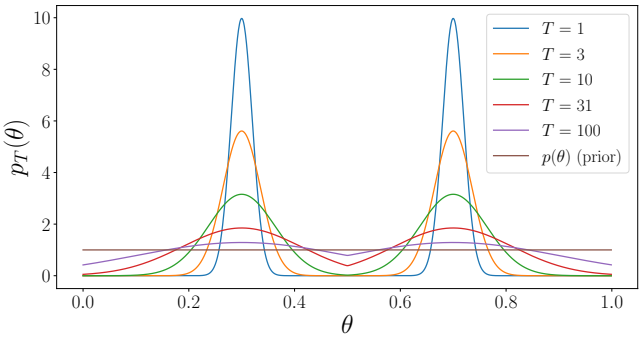


Figure 1. Tempered versions of a bimodal Gaussian distribution with two seperated modes. The tempered posteriors shown in the figure correspond to 5 temperatures distributed logarithmically in range $T \in [1, 100]$. When $T = 1$, $p_T(\theta)$ corresponds to the target distribution, while for $T > 1$, $p_T(\theta)$ is given by the LHS of Equation 3.2. As $T \to \infty$, $p_T(\theta) \to p(\theta)$.

Algorithm 2: Parallel Tempering MCMC

Input: Function p_T and L which, given $\boldsymbol{\theta}$, evaluate the value of the target distribution tempered by T, and the loglikelihood at $\boldsymbol{\theta}$ respectively.
Output: set of N samples from the target posterior $\{\boldsymbol{\theta}_i\}_{i=1}^{N} \sim p_{T=1}$.
Initialize $\boldsymbol{\theta}_1$ s.t. $p_{T=1}(\boldsymbol{\theta}_1) \neq 0$.
Initialize temperatures $T_1 = 1 < \ldots < T_k$.
Initialize $\{\Sigma_{T_i}\}_{i=1}^{k}$, covariance matrices of jump proposal distributions.
Set n - number of Metropolis-Hastings updates before state swap between chains is proposed.
Set counter $i = 1$.
while $i \leq \frac{N}{n}$ **do**
  for T **in** *temperatures* **do**
    draw $n-1$ samples $\{\boldsymbol{\theta}_{i,T}\} \sim p_T$ using the M-H algorithm with proposal covariance Σ_T `// see: Algorithm 1`
    adapt Σ_T `// see: Algorithm 3`
    record current state of the walker $\boldsymbol{\phi}_T \leftarrow \boldsymbol{\theta}_{n-1,T}$
  end
  for T **in** *decreasing order of temperatures* **do**
    draw random number $r \sim \mathcal{U}(0, 1)$
    if $(\frac{1}{T_j} - \frac{1}{T_{j+1}})(\mathrm{L}(\boldsymbol{\phi}_{T_{j+1}}) - \mathrm{L}(\boldsymbol{\phi}_{T_j})) > \ln r$ `// this is equivalent to swap acceptance probability in Eq. 3.3`
     then
      accept swap: $\boldsymbol{\phi}_{T_j} \leftrightarrow \boldsymbol{\phi}_{T_{j+1}}$
    end
  end
  for T **in** *temperatures* **do**
    update chain states after all swaps have been proposed: $\boldsymbol{\theta}_{n,T} \leftarrow \boldsymbol{\phi}_T$
  end
  $i \leftarrow i + 1$
end

The pseudo-code formulation of the PTMCMC algorithm written for the purposes of this work is presented in Algorithm 2. Fig 2 shows a comparison between the performance of the standard Metropolis-Hastings and the PTMCMC on a simple 1D bimodal Gaussian distribution.

3.3. *Adaptive Proposal*

In order to ensure fast convergence of the chains to the target distribution, the jump proposal distribution has to be scaled according to the target. Jumps which are too large relative to the features of the target will tend to propose locations of low probability (i.e. will miss the modes of the distribution), which will result in low acceptance rates and

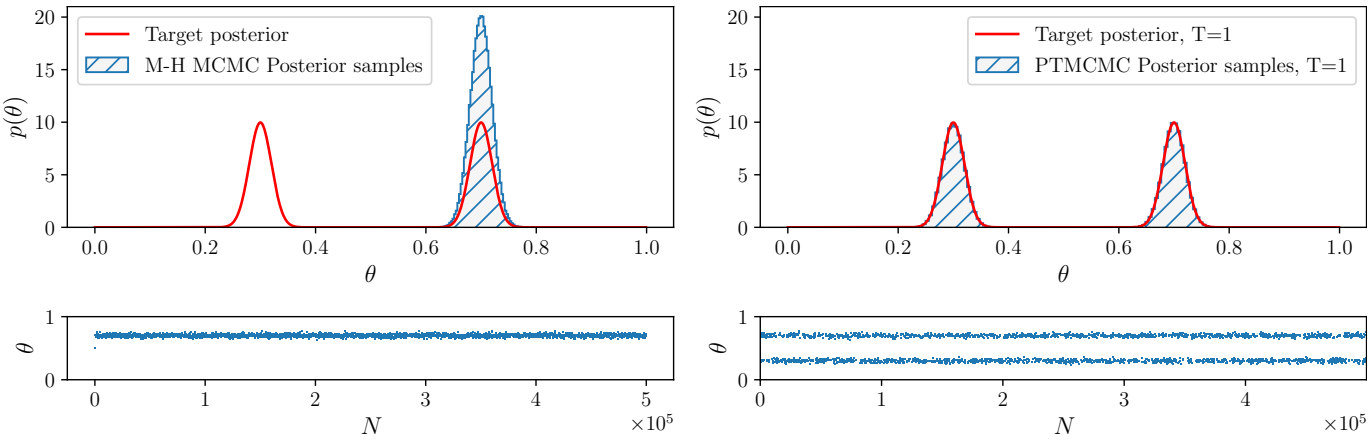


Figure 2. Comparison between a standard Metropolis-Hastings (M-H) and a Parallel Tempered (PTMCMC) sampling runs of 5×10^5 samples from a bimodal Gaussian distribution with modes separated by 20σ. Both chains are initiated at $\theta = 0.5$, exactly between the modes, with fixed Gaussian jump proposal distribution with $\sigma_p = 2.38\sigma$, tuned to a single mode. The top panel shows the sample histograms plotted against the target posterior while the bottom panel shows the corresponding MCMC chains (location of the walker within the probability space as a function of sample number N). The M-H algorithm (left) gets locked at the first found isolated mode and fails to fully explore the posterior. This is remedied by the PTMCMC algorithm (right) - state swaps between tempered chains allow the $T = 1$ chain to frequently traverse the low probability region between the modes, ensuring even sampling of the posterior. (For plotting purposes the chains in the bottom panel were thinned by a factor of 200).

failing to adequately explore the probability space. At the other extreme, jumps which are too small will lead to very high acceptance rates, at the cost of strongly correlated samples and slow convergence.

Given an N-dimensional target distribution and a multivariate Gaussian jump proposal, manually scaling it to achieve the best autocorrelation structure would require specifying $N(N+1)/2$ covariance matrix entries[7] describing not only the jump variance along every individual dimension but also the covariance of the 2D correlation structure for every pair of parameters. This is not an easy task, in particular when we have little or no prior information about the scales and correlations of the target, which is usually the case.

A way of overcoming this problem is adapting the proposal distribution as we learn more about the target from the already accepted samples of the ongoing sampling run, in order to achieve desired acceptance rates. The adaptation implemented in this work is the simple Adaptive Proposal (AP) devised by Haario *et al.* in [25], based on recalculating the covariance matrix of the proposal distribution using a fixed number of past samples. Algorithm 3 provides the corresponding pseudo-code. Following the implementation in [25], the covariance matrix is scaled by a dimensionality dependent constant, which ensures an optimal acceptance rate of 20-50% across all tempered chains. Additionally, the matrix is altered by adding the identity matrix times a small constant (in the current implementation fixed to 10^{-12}) in order to stop the covariance matrix from becoming singular, should that occur otherwise. The adaptation is turned on after initial 1000 samples are obtained.

Algorithm 3: Adaptive Proposal (AP)

Assume PTMCMC state swaps are proposed every n samples, N is the dimensionality of the target posterior and $\{\boldsymbol{\theta}_T\}$ is the set of n most recently accepted samples from the posterior at temperature T.

Before a swap is proposed:

for T *in temperatures* **do**

  $\Sigma_T \leftarrow \frac{2.38^2}{N}\mathrm{cov}(\{\boldsymbol{\theta}_T\}) + \epsilon I_N$; // $\epsilon = 10^{-12}$

end

[7] This is a diagonal matrix, so one only needs to choose the diagonal and half of the off-diagonal entries. Hence the number of choices is less than N^2.

As discussed in [30], the AP method in principle violates the ergodicity of the chain and hence the limiting distribution of the algorithm can differ slightly[8] from the target. A possible alternative is the Adaptive Metropolis (AM) algorithm [30] which recursively performs the adaptation using the whole history of past samples, as opposed to only a fixed number used by the AP adaptation. The AM algorithm, even though the adaptation still violates the Markovian property[9] of the chain, preserves the required ergodic properties [30]. In fact any *vanishing* adaptation (i.e. any adaptation that becomes less significant as the number of samples grows) will result in an algorithm which preserves asymptotical convergence to the target, as discussed extensively in [22].

Regardless of the loss of formal ergodicity, I have chosen to implement the non-vanishing adaptation of the AP algorithm, as it better complements Parallel Tempering, particularly in the case of potentially multimodal distributions. The reasoning behind it is as follows - given a distribution with distant isolated modes enclosing comparable masses of probability (and after both modes had been found by the chain), the AM adaptation (which recalculates the distribution covariance based on the whole sampling history) will favour jumps too large to adequately explore a single mode, often landing in the low probability region between the modes[10]. The 'forgetful' AP algorithm however, can be tuned to remember only samples from the most recently visited distribution mode and hence promotes jumps scaled adequately to that mode. In practice this is done by recalculating the jump proposal covariance every time the algorithm proposes a state swap between adjacent chains. If a swap occurs every 100 samples, at the time of the swap the distribution covariance gets recalculated based on the 100 most recently visited points (Algorithm 3). Since in the case of isolated modes spontaneous intermodal jumps will be very unlikely, the covariance will correspond to the scale of a single mode. As a result, the jump proposal will be well-tuned to the local shape of the distribution, whereas the occurrence of intermodal jumps will be ensured by Parallel Tempering. This approach will be most efficient for distributions where different modes exhibit a similar scale and correlation structure. However, given that only a fraction of the intermodal jumps proposed between the tempered chains is accepted, this is still a justified solution.

Fig. 3 shows how the implemented adaptation reduces sample correlations and shortens the burn-in period of the chain in the case of an example BBH posterior.

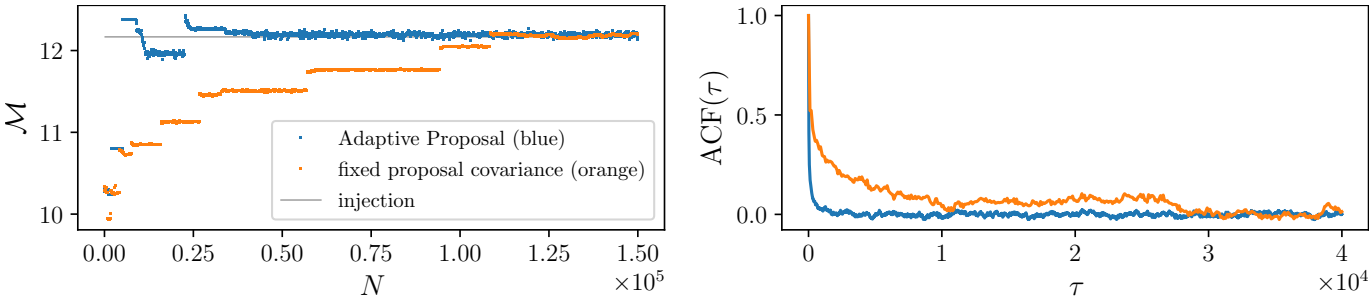


Figure 3. Comparison between example PTMCMC chains ($T = 1$) sampling the same GW posterior with the AP adaptation (blue) and without it (orange) and their respective normalised autocorrelation functions after burn-in subtraction. The injected chirp mass is $\mathcal{M} \approx 12.17$. Both chains are initiated at the same location. Adaptation reduces the burn-in period from approximately 120,000 to 40,000 samples (left) and reduces the sample correlation at most lags τ (right), improving the quality of the sample.

4. ALGORITHM VALIDATION

In order to validate the sampling algorithm, its performance was tested on three standard distributions - a unimodal multivariate correlated Gaussian, a bimodal multivariate uncorrelated[11] Gaussian, both in 9 dimensions, and a 3-dimensional Rosenbrock's banana function. Dimensionality of the Gaussians is equal[12] to the dimensionality of a simple

[8] According to [25] this is in most cases insignificant.

[9] Markovian property refers to future states depending only on the present state of the process, and not the past.

[10] A possible solution to this adaptation problem is discussed in Section 6.

[11] Initially I have planned to run this test on a correlated bimodal Gaussian, but the correlation decreases the frequency of accepted intermodal jumps and increases the integrated autocorrelation times of the chains. This results in having to perform long sampling runs (at least few million samples) to achieve a decent number of uncorrelated samples to perform the K-S test - such long runs were not really feasible due to technical limitations of remote access to the school machines.

[12] The choice of test distribution dimensionality was made prior to deciding on using phase marginalised posteriors, which have only 8 free parameters.

spinless GW posterior. The Rosenbrock function is supposed to resemble the GW posterior correlations between source sky position and coalescence time, which tend to form a 3-dimensional elongated banana shape in the probability space (Fig. 5).

4.1. *Test distributions*

Following the approach in [16] the parameters of the Gaussian test distributions were chosen so that they closely mimic the scales and correlations present in a typical GW parameter posterior. For the unimodal Gaussian this is accomplished by specifying the distribution covariance matrix to be equal to the covariance matrix of the posterior samples of one of the simulated GW signals. The priors were set to uniform within a prior range of $\pm 5\sigma$, centred on the distribution mean in each of the 9 dimensions.

The bimodal uncorrelated Gaussian was set to be a sum of two unimodal Gaussians, each specified as above, with the difference that the off-diagonal covariance matrix entries were set to 0. The distribution modes were separated by 10σ in each dimension; the prior range was scaled adequately, to extend a further $\pm 5\sigma$ beyond the modes.

The 3D Rosenbrock function loglikelihood was formulated as

$$f(\boldsymbol{\theta}) = -\frac{1}{20}\sum_{i=1}^{2}\left[100(\theta_{i+1}-\theta_i^2)^2+(1-\theta_i)^2\right] \quad \text{where} \quad \boldsymbol{\theta} = (\theta_1,\theta_2,\theta_3)^{\mathbf{T}} \in \mathbb{R}^3 \tag{4.1}$$

where the constant rescaling of 1/20 preserves the narrow ridge-like structure of the distribution, as implemented in [31]. The prior was set to uniform within the relevant bounds, so that bulk of the probability mass was enclosed within.

4.2. *Sampling parameters*

Each sampling run was performed by the adaptive PTMCMC algorithm and consisted of 10^6 posterior samples. Sampling was performed in logspace. The distributions were sampled by 8 (unimodal Gaussian) and 10 (bimodal Gaussian, Rosenbrock) parallel chains with temperatures distributed logarithmically in the range $T \in [1, 1000]$. The temperatures and the number of chains were chosen to ensure efficient mixing between chains and suppression of local features in the highest temperature chain. Swaps between the chains were proposed every 100 samples. The walkers were initialized randomly within the prior range, ensuring the value of the distribution at the initial point is non-zero. The initial covariance matrix of the jump proposal distribution was chosen to be a diagonal matrix with each variance entry randomly scaled to between 0.1 and 10 times the true variance of the distribution along the corresponding dimension. This is supposed to simulate a ± 1 order of magnitude error margin on the initial estimation of the parameter variances. After the sampling was completed, each set of samples was visually examined to estimate the burn-in period. Samples corresponding to the burn-in were subsequently discarded from the analysis. The corner plots of the samples are presented in Appendix C.

4.3. *Sample thinning*

In order to obtain an uncorrelated MCMC sample, the raw samples are thinned by the longest integrated autocorrelation time, τ_{ac}, of their 1-dimensional chains. The integrated autocorrelation time approximates the average number of Markov chain iterations (lag) between two uncorrelated samples. Following the approach presented in [32] an estimate of τ_{ac} is found by summing the values of the normalised autocorrelation function of the chain up to a suitably chosen lag M, i.e.

$$\tau_{\text{ac}} \approx \tau_{\text{ac}}(M) = 1 + 2\sum_{\tau=1}^{M}\rho(\tau), \quad \text{where} \quad M = \min(\{L | L \leq 5\tau_{\text{ac}}(L)\}) \ll N \tag{4.2}$$

where $\rho(\tau) = \text{ACF}(\tau)/\text{ACF}(0)$ is the normalised autocorrelation function of the chain at lag τ.[13]

This autocorrelation time estimate is obtained separately for each chain and the maximal obtained value is used to thin the samples, so that only every $\max(\tau_{\text{ac}})$ sample is used.

[13] Note - this discards the autocorrelation contributions at lags greater than M - this is done to reduce the variance in the estimator.

4.4. *Kolmogorov-Smirnov test*

Statistical comparison of the obtained MCMC samples against the target distributions is performed by the Kolmogorov-Smirnov (K-S) test [33] on the uncorrelated samples corresponding to the first distribution dimension. The quantity analysed by the K-S test is the supremum of the absolute value of the difference between the sample empirical cumulative distribution function (ECDF) and the continuous target distribution cumulative distribution function (CDF). This supremum, referred to as the $D_{K\text{-}S}$ statistic, will follow the Kolmogorov-Smirnov distribution, independent of the target and determined solely by the sample size. Given a $D_{K\text{-}S}$ statistic we can then associate with it a p-value corresponding to the probability of obtaining a $D_{K\text{-}S}$ statistic equally or more extreme under the assumption that the null-hypothesis, that the sample follows the target distribution, is true. Using the standard 5% significance level, if the p-value falls below 0.05 the null-hypothesis will be rejected. Otherwise, for $p > 0.05$ I will claim that the the sample shows no statistically significant deviation from the target distribution.

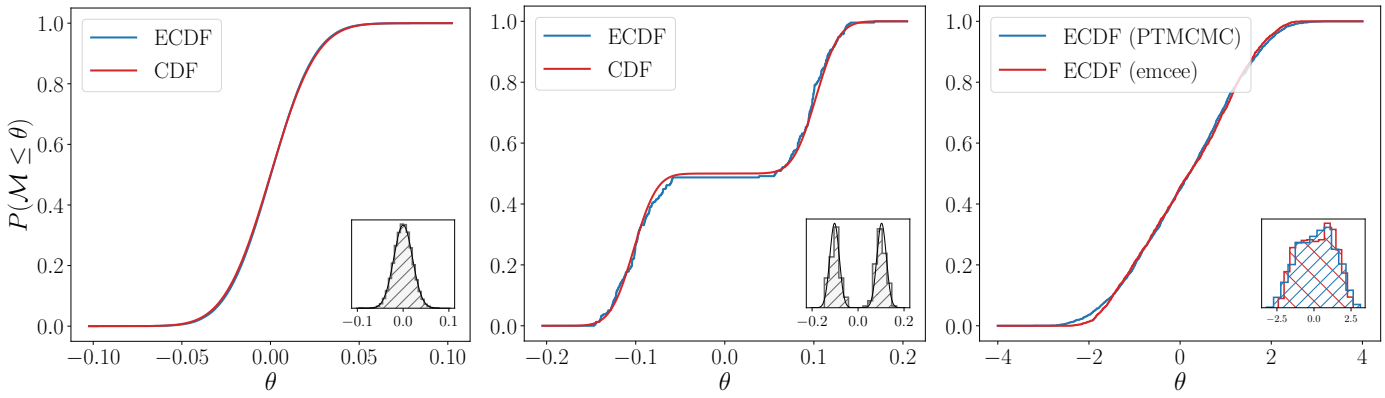
Table 1 presents the results of the K-S test and Fig. 4 shows the corresponding cumulative distributions and sample histograms. All p-values are well above the 0.05 margin, so there is no statistically sufficient evidence in the samples to contradict the null-hypothesis that the samples are drawn from their respective targets. Note that due to lack of analytical formulation of the CDF of the Rosenbrock distribution, the 2-sample K-S test was performed instead, where ECDF of the PTMCMC samples was compared against the ECDF of samples obtained by a popular ensemble MCMC sampler *emcee* [34].

Table 1. Kolmogorov-Smirnov test results.

Target Distribution	sample size, N [a]	τ_{ac} [b]	effective sample size, N/τ_{ac}	$D_{K\text{-}S}$	p-value
Unimodal Gaussian	9.90×10^5	420	2358	0.0114	0.91
Bimodal Gaussian	9.90×10^5	4212	236	0.0563	0.43
Rosenbrock function	9.99×10^5	486	2056	0.0277	0.41

[a] Sample size corrected for the burn-in period.

[b] Estimated autocorrelation time for the worst chain sampling from the target ($T = 1$) posterior.



(Unimodal Gaussian) (Bimodal Gaussian) (Rosenbrock function)

Figure 4. Comparison between the ECDF of the uncorrelated PTMCMC samples and the CDF of the target marginal along the 1st dimension. The supremum of the absolute value of the difference between the two gives the $D_{K\text{-}S}$ statistic. The histograms in the corner show the corresponding samples plotted against the target distribution. For the Rosenbrock function, the comparison is performed between two samples: one obtained by my adaptive PTMCMC, one by the *emcee* sampler.

5. PARAMETER ESTIMATION OF SIMULATED BBH INJECTIONS

The developed sampling algorithm was applied to simulated binary black hole (BBH) signals of different amplitudes injected into a two detector network (L1, H1) with simulated frequency domain noise following O3 PSDs.

I discuss the choice of the waveform model and injection parameters (5.1), generation of frequency domain noise (5.2), the choice of priors (5.3) and present the parameter estimation results (5.4).

5.1. *Waveform model and simulation parameters*

The injected signals are inspiral-merger-ringdown waveforms simulated using the non-precessing IMRPhenomD approximant, generated in frequency domain using `SimInspiralChooseFDWaveform` from the *LALsimulation* Python package. Following the O3 data analysis [3], the lower frequency cut-off[14] for the signals was set to $f_{\rm min} = 20\,{\rm Hz}$, while the reference frequency was set to $f_{\rm ref} = 40\,{\rm Hz}$. For each detector in the network, each of the simulated signals was first modified by the antenna response pattern of the detector and then frequency modulated according to the respective time delay[15]. Thus processed signal was then injected into fake intereferometer noise, generated as outlined in Section 5.2.

All signal parameters, but the source luminosity distance, were kept constant across all injections, fixed at values chosen to approximate a signal representative of the population of binaries detected in O3a, motivated by the population parameter distributions presented in [3]. The luminosity distance to the source was varied to obtain a linear distribution of injections in optimal network SNR, between SNR of 5 and 30, with a spacing of 5. The optimal SNR for a detector network $\{D\}$ is calculated by adding the individual detector SNRs in quadrature, i.e.

$$\mathrm{SNR}(\{D\}) = \sqrt{\sum_D \mathrm{SNR}^2(D)} \approx \sqrt{\frac{4}{T}\sum_{k>0,\,D}\frac{|\tilde{h}_k^D|^2}{S^D(f_k)}}. \tag{5.1}$$

The suitable luminosity distances were chosen after interpolating the distance as a function of network SNR, keeping the remaining parameters fixed. The parameters of all the simulated injections are presented in Table 2.

Even though the analysis does not enter the time domain, in order to generate the frequency domain signals we still need to specify the signal sampling frequency $F_{\rm s}$ and the signal duration T. These determine the maximal (Nyquist) frequency ($f_{\rm max} = F_{\rm s}/2$) and the frequency spacing ($\Delta f = 1/T$) of the complex frequency series. These were chosen to be $F_{\rm s} = 2048\,{\rm Hz}$ and $T = 8\,{\rm s}$ respectively, after ensuring the lowest chirp mass signal allowed by the prior fully fits[16] into the analysed data segment.

5.2. *Simulated frequency domain noise*

For the purposes of simulating the parameter estimation process, the simulated signals were injected into fake interferometer noise. The noise was modelled to follow the O3 design PSDs as given in [39; 40] and, for simplicity, it was generated in frequency domain. Recalling that we have adopted a stationary Gaussian noise model with zero mean and noise variance given by Eq. 2.11, following [15] we can write

$$\frac{T}{2}S(f_k) = \sigma_k^2 = \langle|\tilde{n}_k|^2\rangle = \langle x_k^2 + y_k^2\rangle = \langle x_k^2\rangle + \langle y_k^2\rangle \tag{5.2}$$

where x_k and y_k are respectively the real and imaginary parts of the noise in the $k^{\rm th}$ frequency bin, i.e. $\tilde{n}_k = x_k + iy_k$, and where we have used linearity of the expectation operator.

Assuming that the power is evenly split between the real and imaginary part of the noise, it follows that x_k and y_k have equal variance γ_k^2. Modelling both x_k and y_k as having zero mean, this variance is simply equal to the expectation of their squared values, i.e.

$$\gamma_k^2 = \langle x_k^2\rangle = \langle y_k^2\rangle \tag{5.3}$$

and hence, using Eq. 5.2,

$$\gamma_k^2 = \frac{T}{4}S(f_k). \tag{5.4}$$

Therefore, given a model noise power spectral density PSD $S(f)$, a realisation of this noise is obtained in the frequency domain by randomly drawing the complex and imaginary parts of the noise component in each frequency

[14] At lower frequencies detector strain is dominated by the noise.

[15] The routines used for this were respectively `ComputeDetAMResponse` and `TimeDelayFromEarthCenter` from *LAL* Python library.

[16] The final frequency and time duration of the longest (minimal chirp mass) signal allowed by the prior were estimated using `SimInspiralGetFinalFreq` and `SimInspiralChirpTimeBound` functions from the *LALSimulation* library.

bin from a Gaussian distribution with zero mean and variance given by Eq. 5.4, i.e.

$$\tilde{n}_k = x_k + iy_k, \quad x_k \sim \mathcal{N}\left(0, \frac{T}{4}S(f_k)\right), \quad y_k \sim \mathcal{N}\left(0, \frac{T}{4}S(f_k)\right). \tag{5.5}$$

In order to allow for better comparison of the results, the same noise realisations were used for all of the analysed events. In practice this is done by assigning a fixed integer seed to each of the detectors in the network and using this seed to set the state of the pseudo random number generator before generating the detector noise.

5.3. *Priors*

The prior probability density functions used in the calculation of the posterior ought to reflect the initial state of belief in the distributions of the parameters, independent of the detection. These, as implemented in this work, naturally follow from the astrophysical assumptions made about the underlying population of the compact binary sources. One might argue however, that using non-uniform (informative) priors on the parameters of *simulated* signals is a fallacy unless the choice of the parameters is determined by a random draw from the underlying compact binary merger population. I have not formally performed such a random draw, yet the choice of parameters was motivated by the merger population observed in the O3a detection run. For that reason, and in order for the posterior to closely resemble the posteriors present while analysing real detections, the astrophysically motivated priors were used.

For sampling purposes, the problem was reparametrised in terms of the chirp mass $\mathcal{M}$ and mass ratio[17] q, defined [35] as

$$\mathcal{M} \coloneqq \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}; \quad q \coloneqq \frac{m_1}{m_2}, \quad m_2 \leq m_1. \tag{5.6}$$

The resultant joint prior on $\mathcal{M}$ and q has to be modified by the absolute value of the Jacobian determinant associated with this transformation. Given a uniform mass prior this gives

$$p_{\mathcal{M},q}(\mathcal{M}, q) = p_{m_1,m_2}(m_1(\mathcal{M},q), m_2(\mathcal{M},q)) \left|\frac{\partial(m_1,m_2)}{\partial(\mathcal{M},q)}\right| \propto \mathcal{M}(1+q)^{2/5} q^{-6/5} = \mathcal{M}^{-1} m_2^2. \tag{5.7}$$

The luminosity distance prior reflects the belief that compact binary systems are distributed uniformly in the comoving volume of the universe[18] and that mergers occur at a uniform rate in the source frame[19]. This results in a prior uniform in the comoving volume-time, given by

$$p(D_\mathrm{L}) \propto \frac{1}{(1+z)} \frac{\mathrm{d}V_\mathrm{c}}{\mathrm{d}D_\mathrm{L}} \propto \frac{D_\mathrm{L}^2}{(1+z)^2} \left(\frac{D_\mathrm{L} E(z)}{D_H} + (1+z)^2\right)^{-1} \tag{5.8}$$

[36; 3], where $\mathrm{d}V_\mathrm{c}/\mathrm{d}D_\mathrm{L}$ is the differential comoving volume element, z is the redshift corresponding to the luminosity distance D_L, $E(z) = H_0/H(z)$ and $D_H = c/H_0$ is the Hubble distance - all under the assumed flat ($\Omega_k = 0$) ΛCDM cosmology with Planck 2018 parameters $H_0 = 67.4$ and $\Omega_m = 0.315$ [37]. The first factor of $(1+z)^{-1}$ accounts for time dilation end effectively converts the merger rate in the source frame to the frame of the detector.

The priors on the sky position $\{\alpha, \delta\}$ and the inclination angle ι follow from requiring that both the spatial distribution and the orientation of the binaries are isotropic (i.e. have no preferred direction), i.e.

$$p(\delta) \propto \cos(\delta), \quad p(\iota) \propto \sin(\iota). \tag{5.9}$$

All the remaining priors are uniform within suitable intervals $[\theta_\mathrm{min}, \theta_\mathrm{max}]$, as presented in Appendix B. The overall prior is the product of individual priors, which allows us to recover the overall logprior, i.e.

$$p(\boldsymbol{\theta}) = \prod_i p(\theta_i) \quad \Longrightarrow \quad \ln p(\boldsymbol{\theta}) = \ln \mathcal{U}(\boldsymbol{\theta}_\mathrm{min}, \boldsymbol{\theta}_\mathrm{max}) + \ln \prod_{\theta \in X} p(\theta), \tag{5.10}$$

[17] Note that this is not the most standard definition of the mass ratio - usually in literature we see $q \coloneqq m_2/m_1$ where m_1 is the primary mass, so that $0 < q \leq 1$. The definition assumed in this work follows the definitions introduced in [35]. Theoretically this alternative definition would be problematic as q becomes unbounded for $m_2 \to 0$, which would result in an infinite prior domain. However, within the range of BBH masses involved in this simulation, such exotic mass ratios will lie beyond the prior range, i.e. necessarily $1 \leq q \leq \max(m_1)/\min(m_2) \ll \infty$.

[18] i.e. uniform in the universe volume taking into account the expansion of the universe.

[19] For relatively nearby sources a $p(D_\mathrm{L}) \propto \mathrm{d}V/\mathrm{d}D_\mathrm{L} \propto D_\mathrm{L}^2$ prior is often used, where D_L is treated simply as the radial coordinate of a spherical universe of volume V. This approximation however is only reasonable up to a $D_\mathrm{L} \sim$ few 10^2 Mpc, which is below the prior range of the simulated signals.

where $X = \{(\mathcal{M}, q), D_{\rm L}, \delta, \iota\}$ and

$$\ln \mathcal{U}(\boldsymbol{\theta}_{\rm min}, \boldsymbol{\theta}_{\rm max}) = \begin{cases} 0, & \text{for } \theta_{i,\rm min} \leq \theta_i \leq \theta_{i,\rm max}\, \forall i, \\ -\infty, & \text{otherwise.} \end{cases} \tag{5.11}$$

The above is the logprior[20] used in posterior sampling.

5.4. *Parameter estimation results*

Table 2 presents parameter estimation results for the simulated BBH injections, obtained by logspace posterior sampling with the developed adaptive PTMCMC algorithm. The logarithm of the sampled distribution is the sum of the phase marginalised loglikelihood (Eq. 2.18) and the logprior (Eq. 5.10). Sampling was performed by 8 parallel chains with temperatures distributed logarithmically in the range $T \in [1, 1000]$ and with state swaps proposed every 100 samples. The quoted values correspond to the median and the 90% credible interval centred at the median, both estimated from the empirical cumulative distribution function of the samples. The samples used for obtaining the presented estimates are all 10^6 samples ($T = 1$), short of the initial burn-in, which was estimated individually for each sample run by visual inspection of the chains and subsequently discarded. Corner plots displaying 2D and 1D histograms of the PTMCMC samples are attached in Appendix C. An extract from one of these is presented in Fig. 5.

Table 2. Adaptive PTMCMC parameter estimation results of simulated BBH injections.

		$\mathcal{M}/\mathrm{M}_\odot$	q	$D_{\rm L}$/Mpc	ι/rad	ψ/rad	α/rad	δ/rad	$t_{\rm c}$/s
-	**Injected**	**12.167**	**2.00**	–	**0.78** $(\pi/4)$	**0.60**	**4.50**	**−0.50**	**1126259446**
optimal SNR	$\boldsymbol{D}_{\rm L}$/Mpc					Estimated			
30	**346**	$12.175^{+0.047}_{-0.046}$	$2.01^{+0.14}_{-0.15}$	391^{+90}_{-153}	$0.62^{+0.57}_{-0.43}$	$1.60^{+1.41}_{-1.46}$	$4.53^{+0.26}_{-0.24}$	$-0.51^{+0.15}_{-0.08}$	$1126259446.0001^{+0.0023}_{-0.0011}$
25	**416**	$12.176^{+0.056}_{-0.054}$	$2.01^{+0.17}_{-0.18}$	462^{+123}_{-183}	$0.65^{+0.60}_{-0.47}$	$1.50^{+1.48}_{-1.36}$	$4.51^{+0.33}_{-0.27}$	$-0.50^{+0.18}_{-0.10}$	$1126259446.0000^{+0.0029}_{-0.0011}$
20	**519**	$12.176^{+0.072}_{-0.067}$	$2.02^{+0.21}_{-0.23}$	560^{+173}_{-237}	$0.66^{+0.63}_{-0.47}$	$1.58^{+1.43}_{-1.43}$	$4.56^{+0.45}_{-0.41}$	$-0.52^{+0.29}_{-0.09}$	$1126259446.0004^{+0.0046}_{-0.0015}$
15	**693**	$12.177^{+0.095}_{-0.089}$	$2.03^{+0.28}_{-0.31}$	748^{+249}_{-331}	$0.67^{+0.66}_{-0.47}$	$1.58^{+1.41}_{-1.44}$	$4.52^{+0.54}_{-0.43}$	$-0.51^{+0.35}_{-0.11}$	$1126259446.0001^{+0.0059}_{-0.0014}$
10	**1039**	$12.177^{-0.154}_{-0.131}$	$2.06^{+0.42}_{-0.54}$	1106^{+500}_{-503}	$0.70^{+0.66}_{-0.51}$	$1.49^{+1.48}_{-1.35}$	$4.43^{+0.77}_{-0.48}$	$-0.46^{+0.52}_{-0.16}$	$1126259446.0000^{+0.0083}_{-0.0016}$
5	**2078**	$12.380^{+2.556}_{-2.731}$	$1.87^{+1.16}_{-0.69}$	3020^{+1783}_{-1612}	$0.95^{+0.56}_{-0.65}$	$1.61^{+1.37}_{-1.44}$	$4.33^{+1.81}_{-4.10}$	$-0.00^{+1.18}_{-0.87}$	$1126259446.0024^{+0.0398}_{-0.0439}$

NOTE—The results should be interpreted as follows: $\gamma^{+b}_{-a} \implies P(\theta \leq \gamma) \approx 0.5$ and $P(\gamma - a \leq \theta \leq \gamma + b) \approx 0.9$.

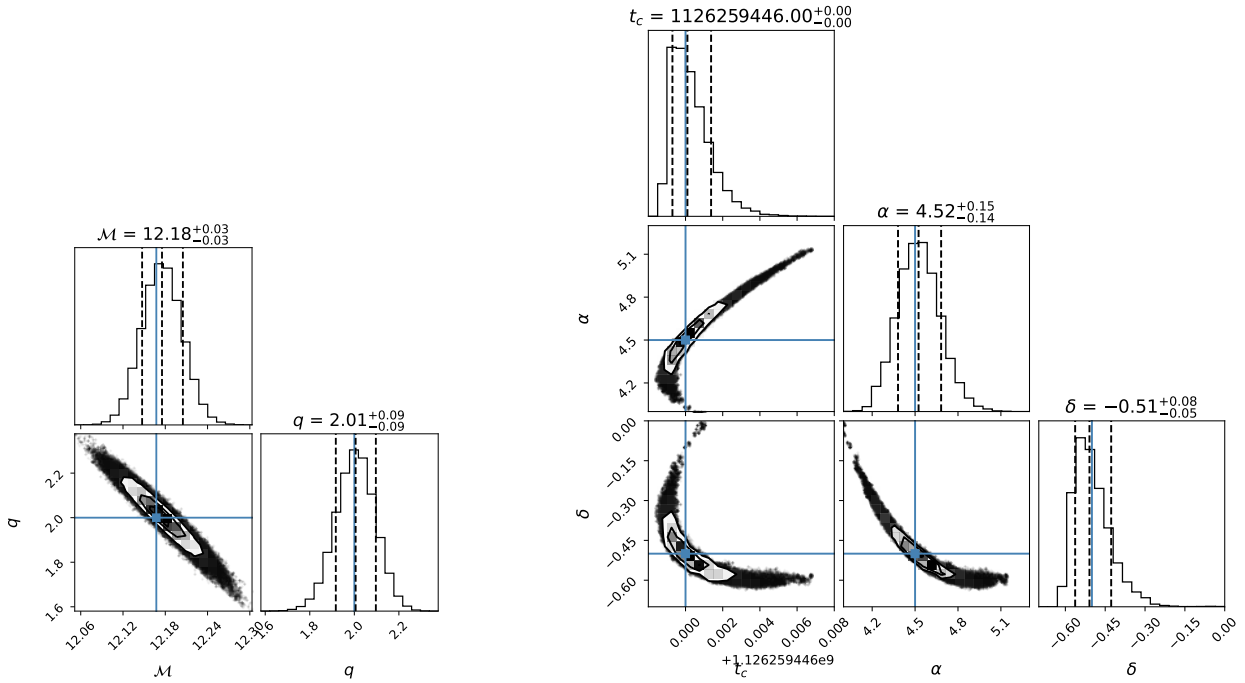
All injected parameters fall within the 90% credible intervals of the sampled marginal distributions for all analysed injections. For each of the parameters of the injections with SNR > 5, but the polarisation angle ψ, which is generally poorly recovered, the bulk of the probability mass of the 1-dimensional marginal distributions is contained within a well localised and clearly defined mode. As expected, as the signal SNR is decreased, the distribution tails become more elongated and the modes are less pronounced. This results in the distribution becoming gradually harder to sample. For the lowest SNR signal, situated below the network detectibility threshold of SNR ~ 12 [38], the noise level is high enough that the 90% credible regions encompass all or nearly all of the prior range and the results are no longer informative. This is an expected behaviour.

The burn-in periods were $\sim 5 \times 10^4$, with fastest converging chains requiring as little as $\sim 10^4$ samples to find the main distribution modes. Proposal adaptation resulted in efficient and nearly constant acceptance rates of $41 - 43\%$ across all tempered chains for all injections. The algorithm does not seem to run into any problems when sampling complicated posterior structures, such as narrow elongated ridges. More tests (ex. involving Fisher information) are needed to conclude whether the recovered posterior scales and correlations match the true shape of the posterior.

6. CONCLUSIONS

Bayesian Inference of GW source parameters involves computing high dimensional integrals and hence requires stochastic sampling techniques, such as MCMC methods. In this work I have developed a simple adaptive Parallel

[20] Note - with this definition some of the parameters have a double prior, i.e. a uniform prior within the range $[\theta_{\rm min}, \theta_{\rm min}]$ multiplied by the informative prior. This is equivalent to the informative prior defined on the domain of the uniform prior, so exactly what was to be obtained.



(chirp mass - mass ratio) (time - right ascention - declination)

Figure 5. Posterior samples for chosen BBH injection parameters of the SNR = 30 signal. Blue solid lines correspond to the true parameter values. Dashed lines mark the 68% credible interval centred at the sample median.

Tempering (PTMCMC) sampler responding to this problem, formulated for the purposes of sampling highly correlated and potentially multimodal GW parameter posterior probability distributions. The algorithm has been validated on 3 analytical distributions and tested on a simulated BBH signal with varying SNR, buried under simulated interferometer noise. The results of these tests were promising - K-S test showed no statistically significant deviations from the respective targets, while for the BBH simulation the injection parameters were well recovered by the algorithm for all events above the detectibility threshold.

There is a number of improvements that the algorithm could benefit from. Firstly, Parallel Tempering should be re-written to make use of parallel processing, so that adding the tempered chains does not increase the computation time. This is a technical improvement which would not influence the quality of the samples, but would largely speed up the sampling process.

Another, and arguably more pressing issue is to consider replacing the non-vanishing AP adaptation with a vanishing adaptation which would ensure asymptotical convergence to the correct target. A relatively simple alternative would be the AM adaptation which uses the whole history of past samples to recalculate the proposal covariance. However, for this to be efficient on multimodal distributions, the AM adaptation should be performed independently within each mode. This could be achieved with the use of a method of mode tracing proposed in [26]. An alternative would be to formally derive the limiting distribution of the Markov chains under the current AP adaptation and show that the deviation is not significant in the case of the typical GW posteriors, or otherwise considering switching the adaptation off at a suitable moment in the sampling process.

Lastly, the most obvious forwards direction is testing the algorithm on a whole range of GW posteriors for different compact sources, parameter choices, waveform models and network configurations. This should include both cross-tests with the already established algorithms and real data parameter estimation, since it should be kept in mind that recovering injection parameters of simulated signals is simpler than a real event analysis in which the signal and the noise model are only approximations and not exact data descriptions.

The Python code written for the purposes of this project, including the adaptive PTMCMC sampler and the BBH injection simulation routines is available at github.com/2316440/AdaptivePTMCMC.

REFERENCES

[1] Abbott, B. P., *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Observation of Gravitational Waves from a Binary Black Hole Merger*. Phys. Rev. Lett., 116(6):061102, (2016).

[2] Abbott, B. P., *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *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).

[3] Abbott, R., *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), *GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run*. arXiv:2010.14527 [gr-qc] (2020).

[4] Abbott, B. P., *et al.* (KAGRA Collaboration, LIGO Scientific Collaboration and Virgo Collaboration), *Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA*. Living Reviews in Relativity, 21(1):3 (2020).

[5] Iyer, B., *LIGO-India, Proposal of the Consortium for Indian Initiative in Gravitational-wave Observations (IndIGO)*. LIGO Document M1100296-v2 (2011).

[6] Danzmann, K., *et al.*, *LISA Laser Interferometer Space Antenna.* A proposal in response to the ESA call for L3 mission concepts. (2017).

[7] Abbott, B. P., *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo*. Astrophys. J. 882, L24 (2019).

[8] Abbott, B.P., *et al.* (LIGO Scientific and Virgo Collaboration), *Tests of General Relativity with GW150914*. Phys. Rev. Lett. 116, 221101 (2016).

[9] Abbott, B. P., *et al.* (LIGO Scientific and Virgo Collaboration), *Tests of General Relativity with the Binary Black Hole Signals from the LIGO-Virgo Catalog GWTC-1*. Phys. Rev. D 100, 104036 (2019).

[10] Abbott, B. P., *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral*. Phys. Rev. Lett. 119, 161101 (2017).

[11] Abbott, B. P., *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *A gravitational-wave standard siren measurement of the Hubble constant*. Nature 551, 85–88 (2017).

[12] Beenakker, W., Venhoek, D., *A structured analysis of Hubble tension*. arXiv:2101.01372 [astro-ph.CO] (2021).

[13] Shoemaker, D., *et al.* (LIGO Scientific Collaboration), *Gravitational wave astronomy with LIGO and similar detectors in the next decade*. eprint arXiv:1904.03187 (2019).

[14] Bustillo, J. C., *et al.*, *Sensitivity of gravitational wave searches to the full signal of intermediate-mass black hole binaries during the first observing run of Advanced LIGO*. Phys. Rev. D 97, 024016 (2018).

[15] Veitch, J., Vecchio, A., *Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network*. Phys. Rev. D 81, 062003. (2010).

[16] Veitch, J., *et al.*, *Parameter Estimation for Compact Binaries with Ground-Based Gravitational-Wave Observations Using the LALInference Software Library*. Phys. Rev. D 91, 042003 (2015).

[17] Cuoco, E., *et al.*, *Enhancing gravitational-wave science with machine learning*. Mach. Learn.: Sci. Technol. **2** 011002 (2021).

[18] Krastev, P. G., *et al.*, *Detection and parameter estimation of gravitational waves from binary neutron-star mergers in real LIGO data using deep learning*. Physics Letters B, Vol. 815, (2021).

[19] Luengo, D., *et al.*, *A survey of Monte Carlo methods for parameter estimation*. EURASIP J. Adv. Signal Process. 2020, 25 (2020).

[20] Hogg, D. W., Foreman-Mackey, D., *Data analysis recipes: Using Markov Chain Monte Carlo*. arXiv:1710.06068 [astro-ph.IM] (2017).

[21] Hastings, W. K., *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. Biometrika, vol. 57, no. 1, 1970, pp. 97–109. (1970).

[22] Andrieu, C., Thoms, J., *A tutorial on adaptive MCMC*. Stat Comput 18, 343–373 (2008).

[23] Radford, N., *MCMC using Hamiltonian dynamics*. Handbook of Markov Chain Monte Carlo. 10.1201/b10905-6 (2012).

[24] Earl, D. J., Deem, M. W., *Parallel Tempering: Theory, Applications, and New Perspectives*. Physical Chemistry Chemical Physics 7.23, 3910 (2005).

[25] Haario, H., Saksman, E., Tamminen, J., *Adaptive proposal distribution for random walk Metropolis algorithm.* Computational Statistics 14, 375–395 (1999).

[26] Pompe, E., Holmes, Ch., Łatuszyński, K., *A framework for adaptive MCMC targeting multimodal distributions.* Ann. Statist. 48, no. 5, 2930–2952. (2020).

[27] Woan, G., *AA12M Statistical Astronomy (STA1).* https://radio.astro.gla.ac.uk/numerical/index.html [accessed: 11 November 2020] (2019).

[28] Veitch, J., Del Pozzo, W., *Analytic marginalisation of phase parameter.* LIGO Document T1300326-v1 (2013).

[29] Messenger, Ch., *STA2 - lecture 6.* School of Physics & Astronomy, University of Glasgow (2019).

[30] Haario, H., Saksman, E., Tamminen, J., *An Adaptive Metropolis Algorithm.* Bernoulli, 7(2), 223-242. (2001).

[31] Pagani, F., Wiegand, M., Nadarajah, S., *An n-dimensional Rosenbrock Distribution for MCMC Testing.* Preprint (2019).

[32] Sokal A., *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms.* In: DeWitt-Morette C., Cartier P., Folacci A. (eds) *Functional Integration.* NATO ASI Series (Series B: Physics), vol 361. Springer, Boston, MA. (1997).

[33] Wayne, D. W., *Kolmogorov–Smirnov one-sample test.* Applied Nonparametric Statistics (2nd ed.). Boston: PWS-Kent. pp. 319–330. ISBN 978-0-534-91976-4 (1990).

[34] Foreman-Mackey, D. *et al.*, *emcee v3: A Python ensemble sampling toolkit for affine-invariant MCMC.* Journal of Open Source Software, 2 4(43), 1864 (2019).

[35] Abbott, B. P. *et al.* (LIGO Scientific Collaboration, Virgo Collaboration). *The basic physics of the binary black hole merger GW150914.* Ann. Phys., 529, No. 1-2. 1600209 (2017).

[36] Hogg, D. W., *Distance measures in cosmology.* arXiv:astro-ph/9905116 (1999).

[37] Aghanim, N. *et al.*, *Planck 2018 results. VI. Cosmological parameters.* arXiv:1807.06209 [astro-ph.CO] (2018).

[38] Abbott, B. P. *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), *Search for intermediate mass black hole binaries in the first observing run of Advanced LIGO.* Phys. Rev. D 96, 022001 (2017).

[39] Soni, S. (LIGO Scientific Collaboration), LIGO-G1900992 document (2019).

[40] Soni, S. (LIGO Scientific Collaboration), LIGO-G1900993 document (2019).

STATEMENT OF WORK

Due to timetabling clashes with the Mathematics part of my degree and upon agreement of Prof. Woan and Dr. Veitch I have worked on this project during the first semester. I had weekly 1 hour supervision meetings with Dr. Veitch (a total of 9), whereas my individual project work was spread over the weeks according to my availability. The supervision time in the meetings was divided between me and another Astronomy & Mathematics student (Student X) who found herself in a similar situation.

I have chosen to explore the topic of MCMC sampling algorithms out of my own interest, however it was Dr. Veitch who suggested I should read into samplers addressing the issue of multimodality such as Hamiltonian Monte Carlo and Parallel Tempering. From then on I have worked on the entirety of the project single-handedly and hence everything presented in this report is my own independent work.

Having said this, however, I cannot fail to acknowledge the advantages of working alongside Student X. Even though our projects had little in common, we supported each other throughout the process, reported our findings to one another and discussed whatever doubts we had. Even though we could be of little meritorical help to one another, often casting our doubts into words was enough to help each other move forwards. On top of that, since Student X's work involved using MCMC methods, she was kind enough to apply my sampler to her problem and test its performance against *emcee*. The sampled posteriors showed no statistically significant disparities.

ACKNOWLEDGEMENTS

I would like to list several opportunities and resources provided to me in the past years of studying at the School of Physics & Astronomy, which made it possible for me to have a particularly smooth start at this project. I was lucky to have worked with Dr. Veitch & Dr. Messenger as a research student in the summer of 2018, where I was exposed to the basics of Bayesian statistics, GW parameter estimation, operation of the ground-based interferometer network and the *LALSimulation* package for simulating GW injections. I have also chosen to attend the Statistical Astronomy lectures in year 3. Prof. Woan's lectures on Bayesian statistics is where my intuitive understanding of the likelihood of a signal in noisy data comes from, whereas Dr. Messenger has introduced me to the basics of the Metropolis-Hastings algorithm and to important statistical concepts such as p-values.

Lastly I would like to thank Dr. Veitch for all his useful advice and support during the course of this project - without it I would not have nearly enough confidence to do any of this.

APPENDIX

A. PARALLEL TEMPERING - STATISTICAL MECHANICS ANALOGY

The use of 'temperature' in Parallel Tempering is not coincidental, as the algorithm can be thought of in terms of statistical mechanics. If we associate the energy of a state $\boldsymbol{\theta}$ with the negative loglikelihood, i.e.

$$E(\boldsymbol{\theta}) = -\ln p(\mathbf{d}|\boldsymbol{\theta}), \tag{A1}$$

then the probability of the chain occupying the state $\boldsymbol{\theta}$ in the system at thermodynamic temperature T is described by the Boltzmann distribution, i.e.

$$p(\mathbf{d}|\boldsymbol{\theta})_T \propto \exp\left[-\frac{E(\boldsymbol{\theta})}{T}\right] = p(\mathbf{d}|\boldsymbol{\theta})^{\frac{1}{T}} \tag{A2}$$

in units where $k_{\rm B} = 1$. This is exactly the tempered version of the likelihood in Eq. 3.2.

If a swap between two chains at temperatures T_i and T_j and states $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ is proposed, we can understand the acceptance probability of this swap (Eq 3.3) as the probability that a transition between the corresponding energy levels occurs at temperature T_i and that the opposite transition occurs in the system at temperature T_j. By expressing this joint probability in terms of the corresponding Boltzmann factors we arrive at

$$p_s \propto \exp\left[\frac{E(\boldsymbol{\theta}_i) - E(\boldsymbol{\theta}_j)}{T_i}\right] \times \exp\left[\frac{E(\boldsymbol{\theta}_j) - E(\boldsymbol{\theta}_i)}{T_j}\right] = \exp\left[(E(\boldsymbol{\theta}_i) - E(\boldsymbol{\theta}_j))\left(\frac{1}{T_i} - \frac{1}{T_j}\right)\right] = \frac{p(\mathbf{d}|\boldsymbol{\theta}_j)^{\frac{1}{T_i}-\frac{1}{T_j}}}{p(\mathbf{d}|\boldsymbol{\theta}_i)}, \tag{A3}$$

which is exactly the statement of Eq. 3.3.

In the higher temperature system the probability of transitions between states separated by a larger energy difference increases and hence analogically the walker sampling from a tempered posterior can visit regions of low likelihood more often, effectively smoothing the distribution.

B. PRIOR BOUNDS

Table B.1. Prior bounds

θ	$\theta_{\rm min}$	$\theta_{\rm max}$	units
m_1	15	25	$M_\odot$
m_2	6	15	$M_\odot$
$D_{\rm L}$	0	5000	Mpc
α	0	2π	rad
δ	$-\frac{\pi}{2}$	$\frac{\pi}{2}$	rad
ι	0	$\frac{\pi}{2}$	rad
ϕ	0	π	rad
$t_{\rm c}$	$t_{\rm c} - 0.05$	$t_{\rm c} + 0.05$	s

NOTE—In principle the inclination angle ι can take values in the interval $[0, \pi]$. However, for a two-detector network the posterior is symmetric w.r.t. $\iota = \pi/2$, and hence a narrower prior was used, corresponding to only half of the full prior range. It is a somewhat questionable choice, as it leaves out a whole other ι mode outside the prior range.

C. CORNER PLOTS

Solid lines indicate the true modes (or injection parameters in GW case) of the marginal distributions, whereas the dashed lines correspond to the sample median and 68% credible interval.

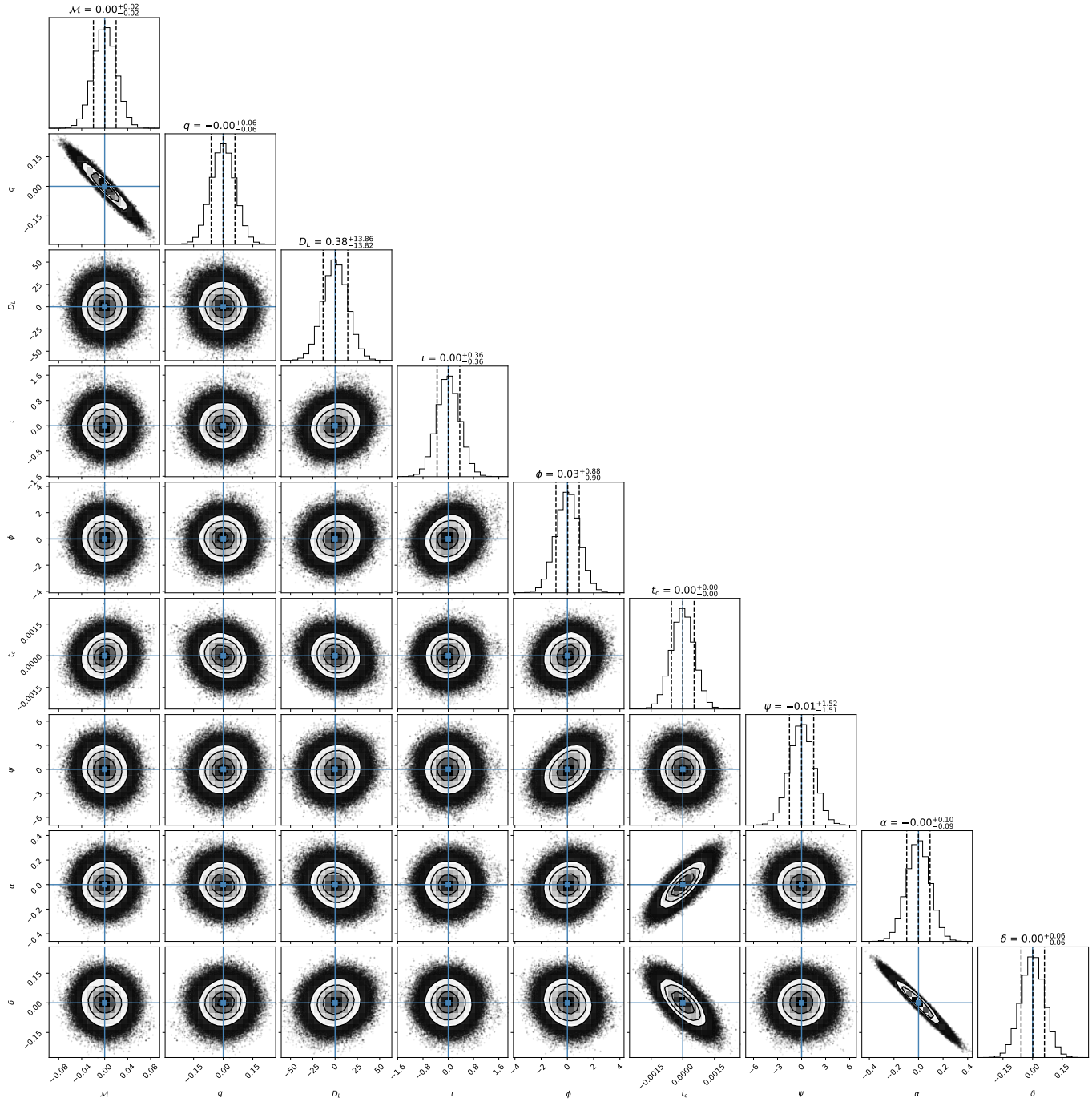


Figure 6. Adaptive PTMCMC posterior samples for Unimodal 9D Correlated Gaussian target. $N = 9.9 \times 10^5$ posterior samples ($10^6 -$ burn-in), sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

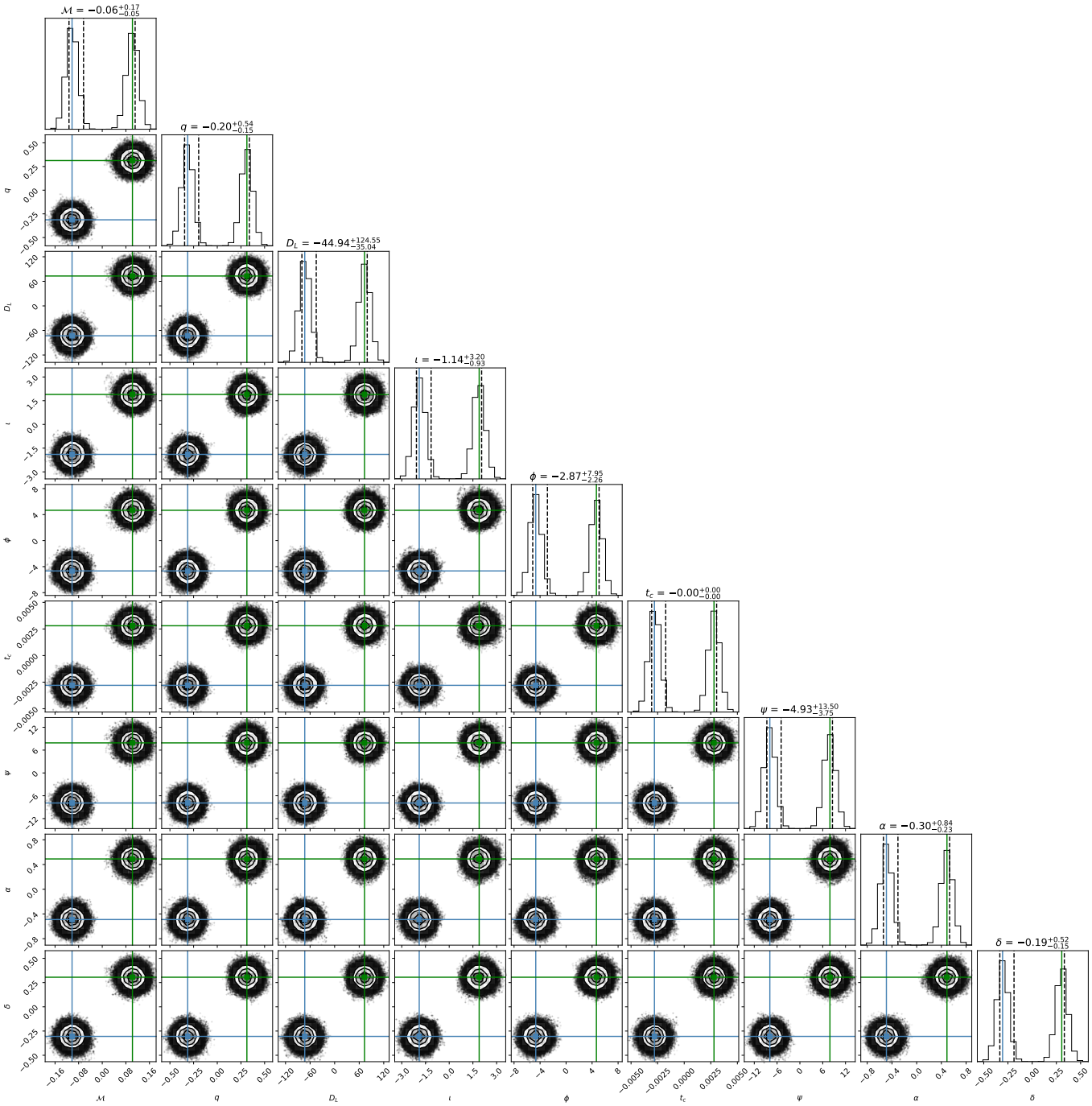
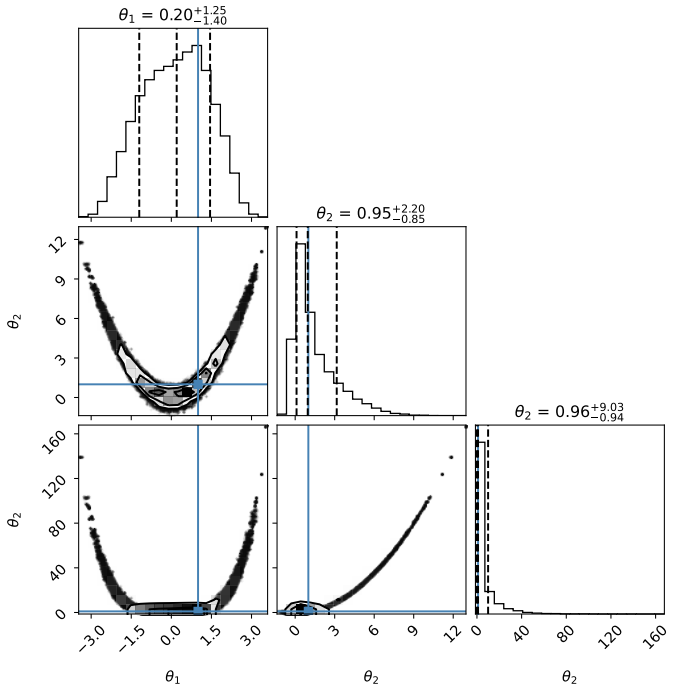
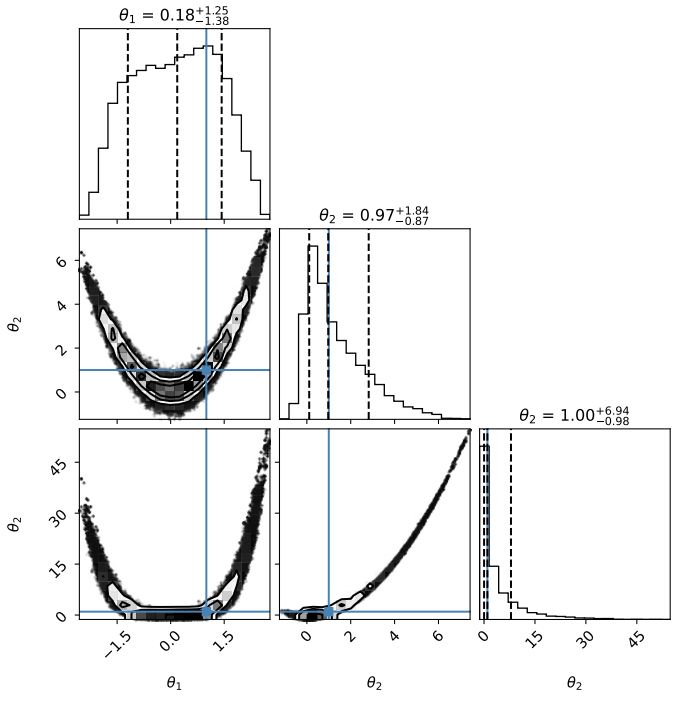


Figure 7. Adaptive PTMCMC posterior samples for Bimodal 9D uncorrelated Gaussian target. $N = 9.9 \times 10^5$ ($10^6 -$ burn-in), sampling in 10 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples. The algorithm successfully finds and explores both modes, however it fails to sample both modes evenly, resulting in shifting of the distribution median (middle dashed line in the diagonal histograms) towards the left mode. This is accounted for by thinning the samples by the integrated autocorrelation time (which picks up the time the walker spends on average within a single mode). However, this effect should be easily eliminated by performing longer sampling runs, so that the time spent within a single mode becomes an insignificant fraction of the sampling duration.



(Adaptive PTMCMC)



(emcee)

Figure 8. Adaptive PTMCMC (top) and *emcee* (bottom) posterior samples from the 3D Rosenbrock's banana function. $N = 9.99 \times 10^5$ (10^6 − burn-in) samples, sampling in 10 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples (PTMCMC); 10 walkers distributed uniformly across the prior space (*emcee*). My adaptive PTMCMC sees longer distribution tails - it is unclear whether this is a sign that my sampler fails to converge to the correct target, or that it has outperformed the *emcee* sampler. More test are required to resolve this.

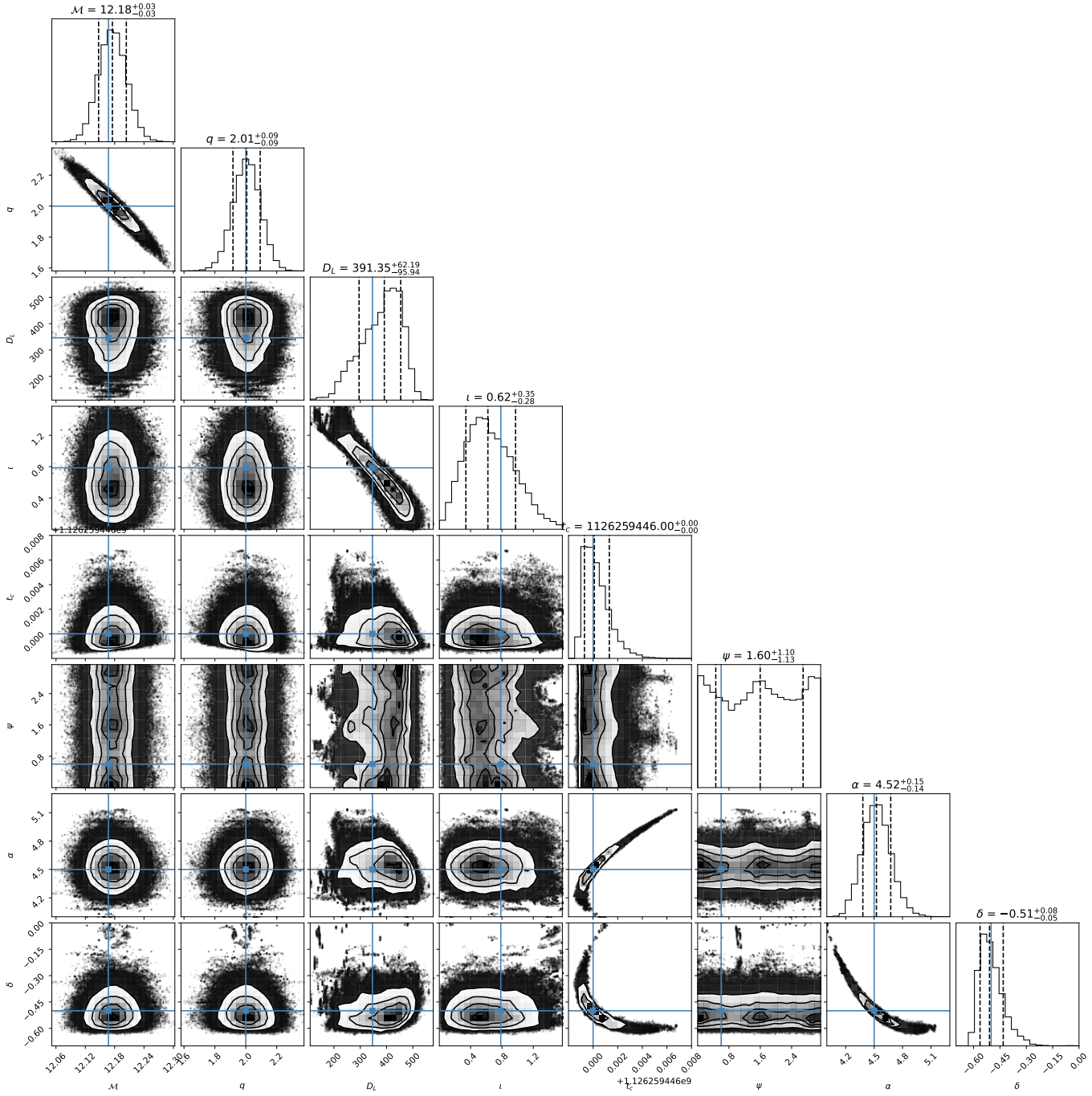


Figure 9. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=30. Sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

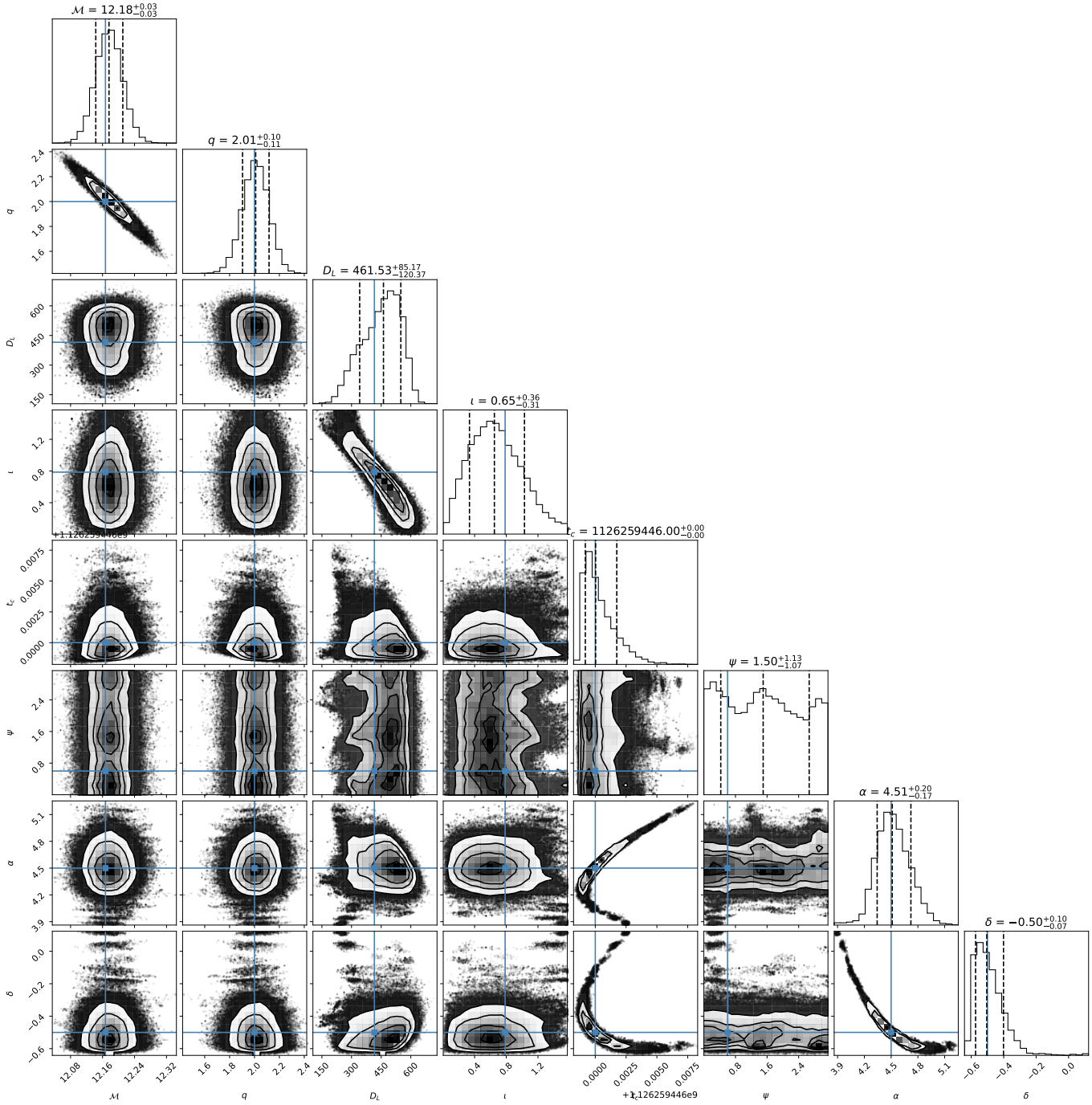


Figure 10. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=25. Sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

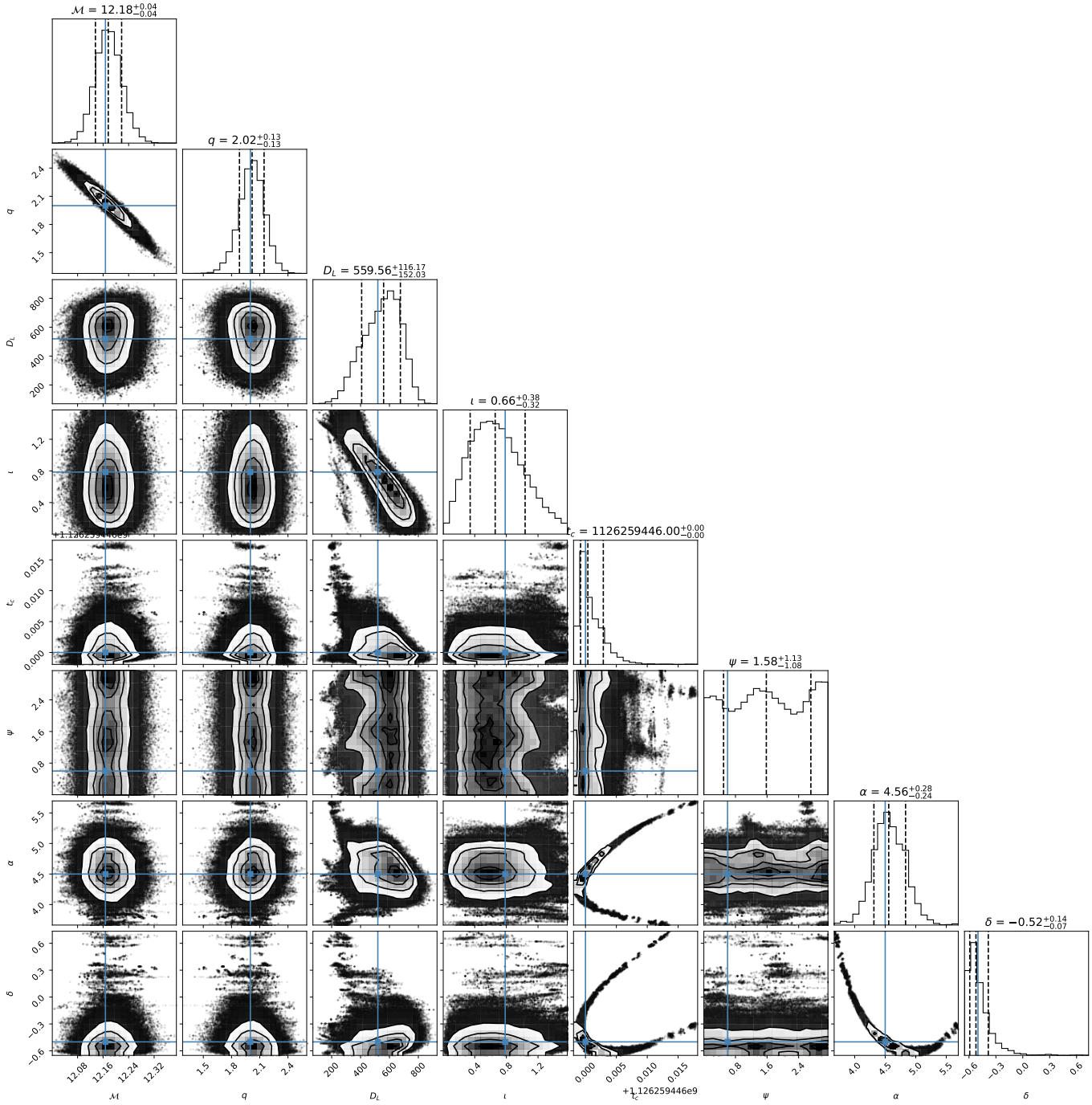


Figure 11. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=20. Sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

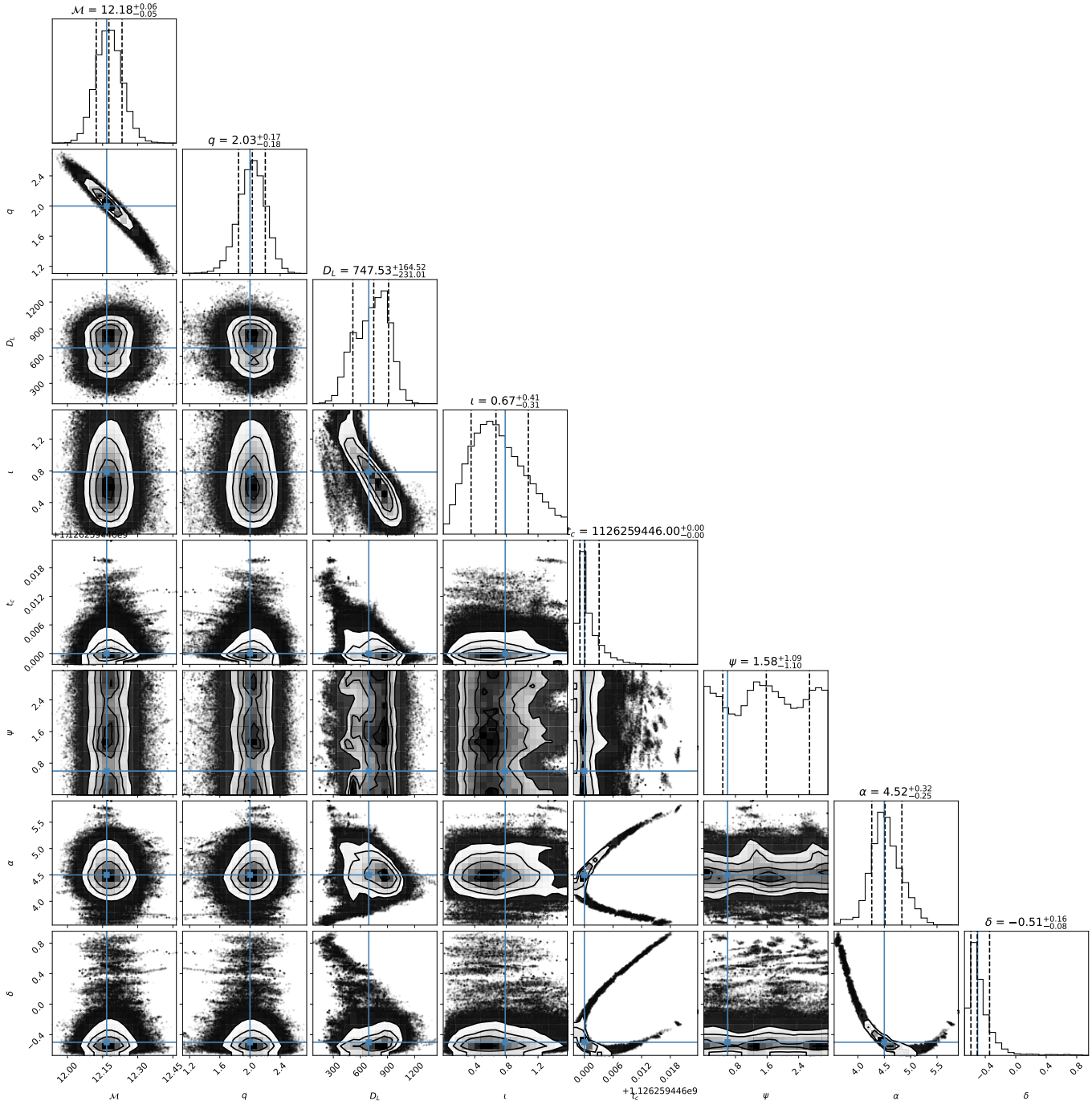


Figure 12. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=15. Sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

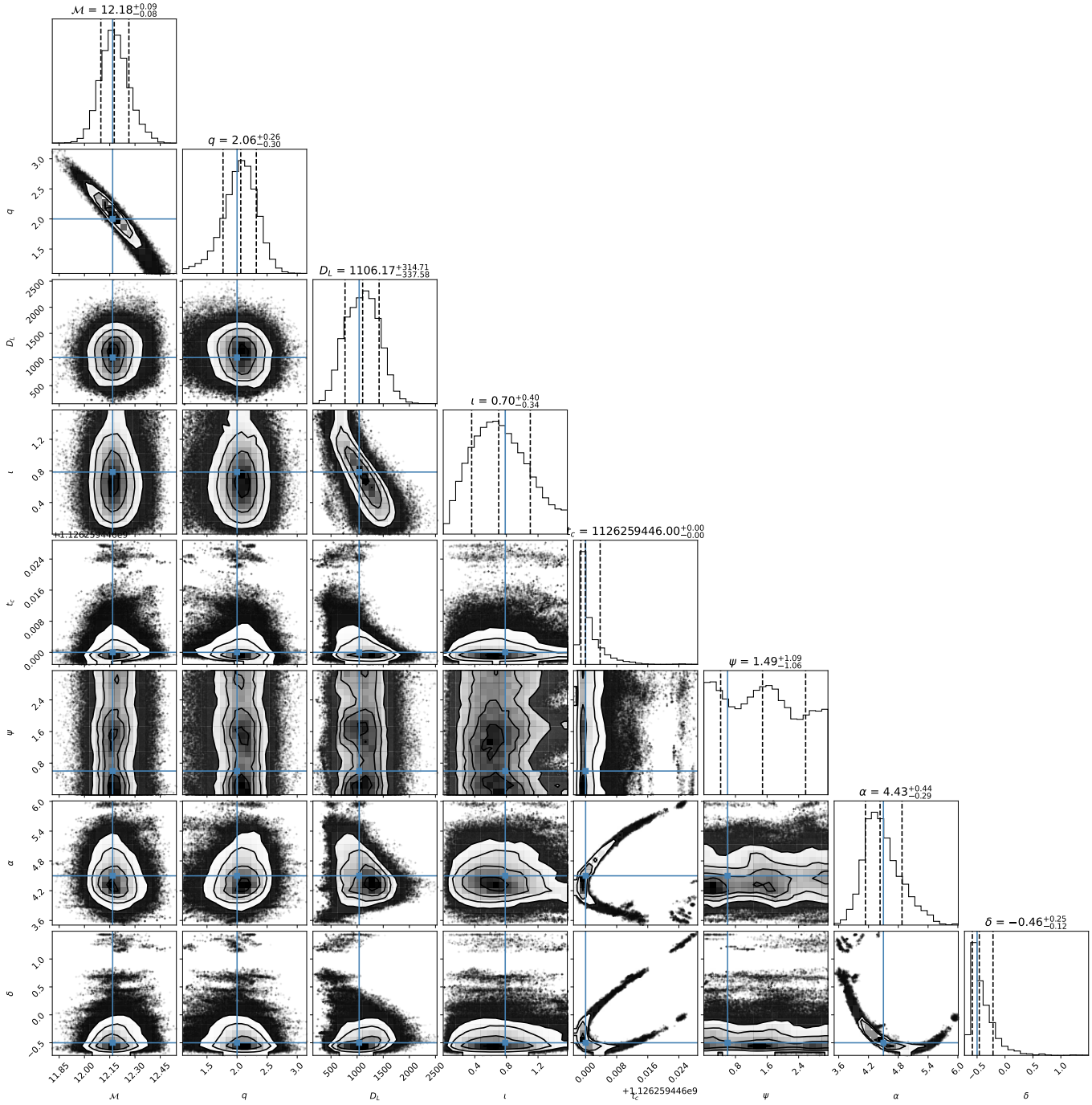


Figure 13. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=10. Sampling in 8 temperature chains distributed logarithmically between $T = 1$ and $T = 10^3$, state swaps proposed every 100 samples.

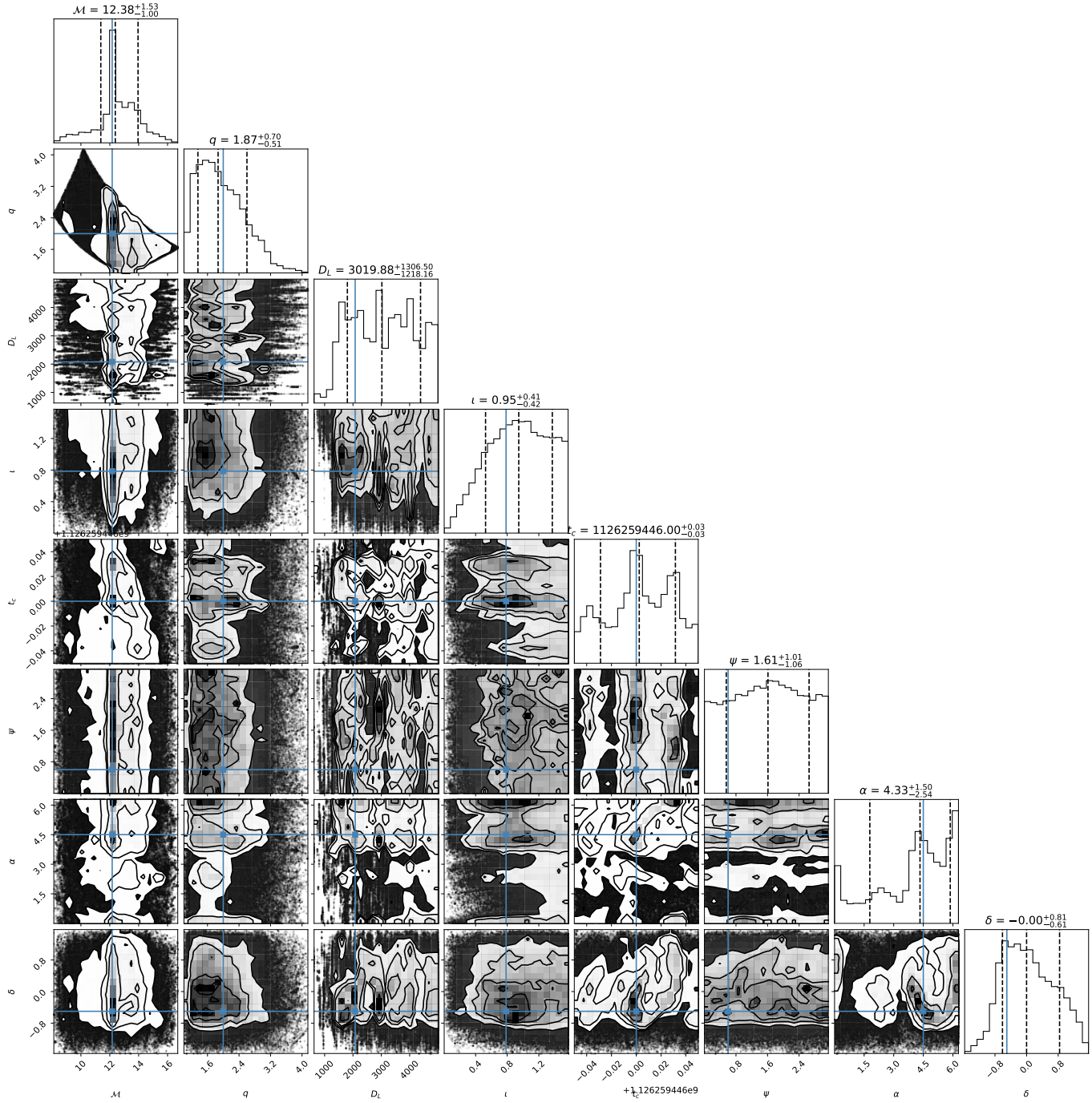


Figure 14. Adaptive PTMCMC posterior samples ($\sim 10^6$) for the BBH injection with optimal network SNR=5. SNR is low enough that the features of the distribution get lost in the noise and the sampler frequently traverses the entire prior range.