

# Accurate characterization of the stochastic gravitational-wave background with pulsar timing arrays by likelihood reweighting

Sophie Hourihane<sup>1,2,\*</sup> Patrick Meyers<sup>1,†</sup> Aaron Johnson<sup>1,‡</sup> Katerina Chatzioannou<sup>1,2,§</sup> and Michele Vallisneri<sup>3,||</sup>

<sup>1</sup>*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*

<sup>2</sup>*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

<sup>3</sup>*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*



(Received 12 December 2022; accepted 31 March 2023; published 26 April 2023)

An isotropic stochastic background of nanohertz gravitational waves creates excess residual power in pulsar-timing-array datasets, with characteristic interpulsar correlations described by the Hellings-Downs function. These correlations appear as nondiagonal terms in the noise covariance matrix, which must be inverted to obtain the pulsar-timing-array likelihood. Searches for the stochastic background, which require many likelihood evaluations, are therefore quite computationally expensive. We propose a more efficient method: we first compute approximate posteriors by ignoring cross correlations and then reweight them to exact posteriors via importance sampling. We show that this technique results in accurate posteriors and marginal likelihood ratios, because the approximate and exact posteriors are similar, which makes reweighting especially accurate. The Bayes ratio between the marginal likelihoods of the exact and approximate models, commonly used as a detection statistic, is also estimated reliably by our method, up to ratios of at least  $10^6$ .

DOI: 10.1103/PhysRevD.107.084045

## I. INTRODUCTION

The nanohertz stochastic gravitational-wave (GW) background can be detected through the induced delay on the times of arrival of pulses from millisecond pulsars [1–3]. Recent evidence that the datasets collected by the three major pulsar-timing-array (PTA) consortia all include excess timing noise of common amplitude and spectral shape [4–7] suggests that we might be getting closer to detection [8,9]. However, since such common-spectrum noise may arise from a non-GW astrophysical or terrestrial source [5,10] (even if this seems unlikely in current data [4,11]), a GW detection claim needs to wait for the finding that the excess noise is correlated across pulsars with the characteristic angular pattern known as the Hellings-Downs curve [2].

In PTA data analysis, timing noise is represented as a Gaussian process with covariance matrix  $C_{aibj}$ , where  $a, b$  range over pulsars and  $i, j$  over timing measurements (or equivalently frequency components). For common-spectrum uncorrelated noise, the matrix factorizes as  $C_{ij}\delta_{ab}$ ; for an isotropic GW background, it is given by  $C_{ij}\Gamma_{ab}$ , with  $\Gamma_{ab} = \Gamma(\theta_{ab})$  the Hellings-Downs correlation coefficient, a function of the angular separation  $\theta_{ab}$  between pairs of pulsars.

The PTA data model includes several other stochastic components, but GW detection is usually formulated by comparing a common process (CP) model that includes common-spectrum uncorrelated noise and an “HD” model that includes common-spectrum Hellings-Downs–correlated noise.<sup>1</sup> By contrast, information about the GW amplitude and spectral shape is carried primarily by the autocorrelation terms (the  $C_{aibj}$  elements with  $a = b$ ).

Parameter estimation and model selection for the CP and HD models are both typically handled through stochastic sampling, which requires repeated evaluations of the data likelihood. Since the CP excess-noise covariance matrix factorizes across pulsars but the HD matrix does not [12], the likelihood is significantly slower to compute for the latter model (e.g., a factor of  $\sim 25$  for the NANOGrav 12.5 yr dataset, which will only grow larger as more pulsars are observed). The number of likelihood evaluations is magnified by the thinning of sample chains [typically by  $N_t \sim \mathcal{O}(10^3)$ ] and by the use of parallel tempering schemes [typically by  $N_c \sim \mathcal{O}(10)$  temperatures] which require many likelihood evaluations per CP posterior sample. The overall cost can be prohibitive for the HD model, particularly when multiple background analyses (e.g., “sky scrambles” [13,14] and “phase shifts” [14]) are required to estimate the significance of a result.

\*sohour@caltech.edu

†pmeyers@caltech.edu

‡aaronj@caltech.edu

§kchatzioannou@caltech.edu

||vallis@caltech.edu

<sup>1</sup>In the NANOGrav 12.5 yr stochastic background analysis that initially reported the evidence for a common process [4], the CP and HD models are labeled model 2A and 3A, respectively.

Methods to optimize PTA search strategies in both data acquisition and modeling have been studied extensively. On the data acquisition side, studies found the most impactful observing cadences and radio frequency bands for detecting a GW background (GWB) [15–17]. On the modeling side, improvements in computational efficiency have been made by using Fourier basis methods [18,19] to characterize red-noise processes, as opposed to dense covariance matrix approaches [20]. More recently, the factorized likelihood approach reduces the wall clock time needed to evaluate a *CP*-only model by a factor proportional to the number of pulsars [12], and Hamiltonian Monte Carlo methods have been implemented to improve sampling efficiency [21].

In this study we propose an approach that further mitigates computational cost of producing posterior samples for the HD model in terms of both CPU and wall clock time. Rather than exploring the HD model stochastically, we reuse parameter-estimation results for the inexpensive *CP* model and “reweight” them to obtain posteriors and marginal likelihoods under the HD model. Specifically, a thinned set of *CP*-model samples yields a set of weighted HD-model samples, with weights equal to the ratios of the HD and *CP* likelihoods. The computational gains are realized by performing only one HD likelihood evaluation per HD posterior sample and parallelizing the calculation of weights.

The general reweighting formalism can be applied to any combination of models, though convergence and low sampling error depend on stochastic chains for the original posterior having a sufficient number of samples in the support of the target posterior. This is the case for the HD and *CP* posteriors, since both are dominated by single-pulsar autocorrelation terms. Additionally, the two models share the same parameters and corresponding priors. In this paper we apply the reweighting formalism to simulated PTA data and compare posteriors and marginal likelihoods obtained by reweighting and by brute-force sampling. We find that (i) the posteriors recovered through reweighting are statistically unbiased; and that (ii) the HD vs *CP* Bayes factors (the ratios of marginal likelihoods) agree with the “hypermodel” method typically used in PTA analyses [22] to within 10% uncertainty for Bayes factors  $\in [10^{-3}, 10^7]$ .

The rest of the paper is organized as follows. In Sec. II we introduce the general reweighting formalism following [23]. In Sec. III we describe the HD and *CP* models in more detail. In Sec. IV we present results from simulated data that validate the reweighting approach. In Sec. V we conclude by discussing the application of our method and its computational gains.

## II. POSTERIOR REWEIGHTING

Samples distributed according to one posterior distribution can, under some circumstances, be reweighted to estimate a second posterior distribution; this is a form of importance sampling. In this section, we present the general

methodology behind this posterior reweighting following Ref. [23] and describe how it can be used to also estimate the marginal likelihood of a model and the Bayes factor between models.

The posterior distribution,  $p(\theta|d, T)$  for a target model *T* with parameters  $\theta$  given data *d* can be written explicitly in terms of the Bayes theorem,

$$p(\theta|d, T) = \frac{\mathcal{L}(d|\theta, T)\pi(\theta|T)}{\mathcal{Z}_T}, \quad (1)$$

where  $\mathcal{L}(d|\theta, T)$  is the likelihood,  $\pi(\theta|T)$  is the prior, and  $\mathcal{Z}_T$  is the marginal likelihood (also known as evidence, though we do not use this term here). We rewrite this target posterior distribution in terms of the likelihood and prior for another “approximate” model *A*,

$$p(\theta|d, T) = \frac{\mathcal{L}(d|\theta, A)\frac{\mathcal{L}(d|\theta, T)}{\mathcal{L}(d|\theta, A)}\pi(\theta|A)\frac{\pi(\theta|T)}{\pi(\theta|A)}}{\mathcal{Z}_T} \quad (2)$$

$$= w_{\mathcal{L}}(d|\theta)\pi(\theta)\frac{\mathcal{L}(d|\theta, A)\pi(\theta|A)}{\mathcal{Z}_T}. \quad (3)$$

In the last line we have introduced weights given by the ratio of the likelihoods and priors of the two models

$$w_{\mathcal{L}}(d|\theta) = \frac{\mathcal{L}(d|\theta, T)}{\mathcal{L}(d|\theta, A)}, \quad (4)$$

$$\pi(\theta) = \frac{\pi(\theta|T)}{\pi(\theta|A)}; \quad (5)$$

we can also combine the weights to get

$$w(d|\theta) = w_{\mathcal{L}}(d|\theta)\pi(\theta). \quad (6)$$

Given  $N_s$  posterior samples  $\theta_s \sim p(\theta|d, A)$  for model *A*, we can resample them with weights  $w(d|\theta_s)$  to obtain a posterior sampling of model *T*; the marginal likelihood  $\mathcal{Z}_T$  can also be estimated as

$$\mathcal{Z}_T = \int d\theta \mathcal{L}(d|\theta, T)\pi(\theta|T) \quad (7)$$

$$= \mathcal{Z}_A \int d\theta w_{\mathcal{L}}(d|\theta)\pi(\theta)p(\theta|d, A). \quad (8)$$

The integral in Eq. (8) can be approximated with Monte Carlo integration

$$\mathcal{Z}_T \approx \frac{\mathcal{Z}_A}{N_s} \sum_{s=1}^{N_s} w_{\mathcal{L}}(d|\theta_s)\pi(\theta_s) = \mathcal{Z}_A \bar{w}, \quad (9)$$

where  $\bar{w}$  is the mean of the weights,  $w(d|\theta)$ . If we are interested in model selection between the approximate

and target models, the Bayes factor between them is then simply

$$\mathcal{B}_A^T = \bar{w}. \quad (10)$$

Though the reweighting procedure is mathematically exact, it is subject to sampling errors, especially if the approximate and target posteriors are too disjoint. We quantify sampling error with the “effective number of samples”  $n_{\text{eff}}$ —the approximate number of samples drawn independently from the target posterior that would approximate  $\mathcal{Z}_T$  as accurately as the reweighting estimate (9). Reference [24] estimates  $n_{\text{eff}}$  as

$$n_{\text{eff}} \approx \frac{\sum_s w_{\mathcal{L}}(d|\theta_s) w_{\pi}(\theta_s)]^2}{\sum_s [w_{\mathcal{L}}(d|\theta_s) w_{\pi}(\theta_s)]^2} = \frac{N_s}{1 + (\frac{\sigma_w}{\bar{w}})^2}, \quad (11)$$

where  $\sigma_w$  is the standard deviation of the weights. We also define the efficiency

$$\mathcal{E} \equiv \frac{n_{\text{eff}}}{N_s}. \quad (12)$$

It follows from Eq. (10) that the error  $\sigma_B$  on the mean  $\mathcal{B}_A^T$  is

$$\sigma_B = \frac{\sigma_w}{\sqrt{n_{\text{eff}}}} = \frac{\sigma_w}{\sqrt{\mathcal{E} N_s}}. \quad (13)$$

If we represent the target posterior by a set of equal-weight samples by performing a weighted redraw from the approximate distribution, then Eq. (11) makes intuitive sense. It implies that a few samples with high weights (relative to  $\bar{w}$ ), will result in the same sample being drawn many times and lead to comparatively lower  $n_{\text{eff}}$ . Equivalently, such high individual weights (relative to  $\bar{w}$ ) increase  $\sigma_w$  and thus decrease  $n_{\text{eff}}$ . In the limit of vanishing variance,  $n_{\text{eff}} \rightarrow N_s$ , while as variance grows  $n_{\text{eff}} \rightarrow 0$ .

We can also use the weights to estimate the statistical distance between the approximate and target posteriors in the form of the Kullback-Leibler (KL) divergence. That is,

$$\mathcal{D}_{KL}(A||T) \equiv \int d\theta p_T(\theta) \ln \frac{p_A(\theta)}{p_T(\theta)}, \quad (14)$$

$$\approx \sum_{\theta \sim p_A} \frac{1}{N_s} \ln \frac{p_A(\theta)}{p_T(\theta)}, \quad (15)$$

$$= \ln(\bar{w}) - \overline{\ln(w)}, \quad (16)$$

where we have written the posteriors  $p_K(\theta) \equiv p(\theta|d, K)$  for  $K \in \{A, T\}$  and  $\overline{\ln(w)}$  is the average of the log of the weights. This equation can be used in combination with Eq. (11) to provide guidance when reweighting leads to low efficiency.

The main reason for low efficiency is that a region of low posterior for the approximate model overlaps with a region of high posterior for the target model. Samples in this region get very high weights and can lead to poor reconstruction of the target posterior. Understanding when and where this can occur is an area of active research and has led to other forms of importance sampling [25].

### III. THE PULSAR-TIMING-ARRAY MODELS FOR STOCHASTIC GRAVITATIONAL WAVES

In this section we discuss the statistical framework used to detect a stochastic GW background with an array of regularly timed millisecond pulsars. We first introduce a Gaussian likelihood that includes the full interpulsar correlations induced by GWs (the Hellings-Downs model). We then introduce a secondary model that ignores interpulsar correlations and includes GWs as a common (but uncorrelated) power-law spectrum in each pulsar’s residuals (a common process model). We claim evidence for GWs when a dataset significantly favors HD over CP.

These two models contain the same parameters and priors; furthermore, the posterior distributions of model parameters are not affected strongly by the inclusion of interpulsar correlations. This makes the CP likelihood a good approximate distribution for the HD likelihood. As we discuss below, the CP model is significantly faster to evaluate, and so we will use CP as our approximate likelihood  $\mathcal{L}_A$ , while the HD model will be the target  $\mathcal{L}_T$ . An added bonus of our choice of these models is that, in implementing the reweighting scheme discussed above to speed up computation of the HD posterior, we naturally also calculate Bayes factors that can be used as a GW detection statistic.

#### A. Pulsar-timing-array likelihood

A detailed presentation of the PTA likelihood derivation can be found in Refs. [3,14,18,19,26,27]; in this section we describe only the relevant details. A reader familiar with PTA analyses can skip to Sec. III C.

Pulse arrival times [time(s) of arrival (TOA)] are affected by both deterministic and stochastic processes. The deterministic contribution (described more fully in [14]) contains terms relating to the motion of the pulsar, such as sky location, rotation period, etc., as well as individually resolvable GW sources such as continuous waves. An initial solution for the timing model is subtracted from the measured TOA, leaving behind the fit residuals  $\delta t$ . The uncertainties on this timing model are described by a Taylor expansion in timing parameters  $\epsilon$  with partial derivatives (comprising the design matrix)  $\mathbf{M}$  evaluated at the initial timing solution.

The stochastic contribution to the TOA is due to a combination of the intrinsic low-frequency spin noise (or “red noise”) of individual pulsars (IRN) and a common,

stochastic process induced by a GW background. We model both as stationary zero-mean Gaussian random processes with Fourier vector bases. It follows that the Fourier coefficients  $\mathbf{a}$  (the basis weights) are described entirely by their covariance  $\phi|_{\eta} = \langle \mathbf{a}_{ai} \mathbf{a}_{bj} \rangle$ . Here indices  $a, b$  index pulsars,  $i, j$  index frequencies, brackets indicate the ensemble average, and the  $\eta$  are the “hyperparameters” associated with the distribution of  $\mathbf{a}$ .

With both the deterministic and stochastic contributions to the noise modeled, the timing residuals  $\mathbf{r}$  are

$$\mathbf{r} = \delta\mathbf{t} - \mathbf{M}\boldsymbol{\epsilon} - \mathbf{F}\mathbf{a} = \delta\mathbf{t} - \mathbf{Tb}, \quad (17)$$

where the matrix  $\mathbf{F}$  collects the Fourier basis vectors evaluated at the TOA and where we have introduced

$$\mathbf{T} \equiv [\mathbf{MF}], \quad \mathbf{b} \equiv \begin{bmatrix} \boldsymbol{\epsilon} \\ \mathbf{a} \end{bmatrix}, \quad (18)$$

for ease of notation. The residuals  $\mathbf{r}$  should now be white and Gaussian, with a covariance matrix  $\mathbf{N}$  that describes the uncertainty associated with each TOA observation. The likelihood is then

$$\mathcal{L}(\delta\mathbf{t}|\mathbf{b}) = \frac{\exp(-\frac{1}{2}\mathbf{r}^T \mathbf{N}^{-1} \mathbf{r})}{\sqrt{\det 2\pi\mathbf{N}}}. \quad (19)$$

We complement the likelihood with the Gaussian-process prior for the Fourier components,

$$\pi(\mathbf{a}|\eta) = \frac{\exp(-\frac{1}{2}\mathbf{a}^T \phi|_{\eta}^{-1} \mathbf{a})}{\sqrt{\det 2\pi\phi|_{\eta}}}. \quad (20)$$

The Gaussian form of the likelihood and prior means that we can marginalize analytically over the  $\mathbf{a}$ , leaving only the hyperparameters  $\eta$ . A similar choice is made for the timing model correction prior  $\pi(\boldsymbol{\epsilon})$  [14]. The marginalized likelihood is then

$$\mathcal{L}(\delta\mathbf{t}|\eta) = \int d\mathbf{b} \mathcal{L}(\delta\mathbf{t}|\mathbf{b}) \pi(\mathbf{a}|\eta) \pi(\boldsymbol{\epsilon}), \quad (21)$$

$$\propto \frac{\exp(-\frac{1}{2}\delta\mathbf{t}^T \mathbf{C}^{-1} \delta\mathbf{t})}{\sqrt{\det(2\pi\mathbf{C})}}, \quad (22)$$

where  $\mathbf{C} = \mathbf{N} + \mathbf{T}\mathbf{B}\mathbf{T}^T$  is the covariance kernel, and

$$\mathbf{B} = \begin{bmatrix} \infty & 0 \\ 0 & \phi \end{bmatrix}. \quad (23)$$

Here  $\infty$  represents a formal limit of covariance for a uniform unbounded prior on  $\boldsymbol{\epsilon}$ .

## B. Pulsar-timing-array stochastic models

We model both the IRN and the GWB as power laws in the frequency domain.<sup>2</sup> The model hyperparameters are then  $\eta = (A^a, \gamma^a, A_{\text{GW}}, \gamma_{\text{GW}})$  where  $\gamma^a$ ,  $\gamma_{\text{GW}}$  and  $A^a$ ,  $A_{\text{GW}}$  are the negative spectral indices and amplitudes of the IRN and GW power laws, respectively. We split  $\phi$  into its two contributions; one from the IRN and the other the common GWB

$$\phi = \phi^{\text{IRN}} + \phi^{\text{GW}}. \quad (24)$$

By stationarity, both the IRN and the GWB are uncorrelated between frequencies. Therefore  $\phi$  will contain no cross-frequency terms, and  $\phi_{ai,bj} \propto \delta_{ij}$ .

By definition, the IRN is an independent process in each pulsar,

$$\phi|_{\eta(ai),(bj)}^{\text{IRN}} = \kappa_i(\eta_a) \delta_{ab} \delta_{ij}. \quad (25)$$

The IRN power  $\kappa_i(\eta_a)$  in frequency bin  $i$  for pulsar  $a$  is modeled as the power law

$$\kappa_i(\eta_a) = \kappa_i(A_a, \gamma_a) = \frac{A_a^2}{12\pi^2 T} \frac{1}{\text{yr}^{-1}} \left( \frac{\nu_i}{\text{yr}^{-1}} \right)^{-\gamma_a} \text{yr}^2, \quad (26)$$

where  $T$  is the total observation time and  $\nu_i$  is the frequency associated with bin  $i$ .

In contrast to the IRN, the GW background is correlated between pulsars,

$$\phi|_{\eta(ai),(bj)}^{\text{GW}} = \Gamma_{ab} \kappa_i(\eta_{\text{GW}}) \delta_{ij}. \quad (27)$$

Here  $\kappa_i$  is again given by Eq. (26), except that every pulsar has the same amplitude  $A_{\text{GW}}$  and spectral index  $\gamma_{\text{GW}}$ . The function  $\Gamma_{ab}$  describes GW correlations between pulsars  $a$  and  $b$  and is known as the Hellings-Downs curve [12], Eq. (5)].

The  $a = b$  components of Eq. (27) represent the excess-noise power induced by the GWB in each pulsar. Half of this power is caused by the “Earth term” in the pulsar GW response and contributes to interpulsar correlations; the other half is caused by the “pulsar term” and is statistically uncorrelated among pulsars. The first indications of a GWB in PTA data will appear through these diagonal self-correlations [8,30], so they could be detected using the *CP* model as well as the *HD* model. However, evidence for *CP* could also be caused by physical effects such as the solar wind [31] or by model misspecification, such as

<sup>2</sup>Power laws are not the only choice for the distribution of Fourier coefficients. Other choices include (but are not limited to) a free spectral model with independent densities for each Fourier frequency and a broken power law [4,28,29].

incorrect priors [10] or poor IRN models [11]. Other mechanisms can induce interpulsar correlations that are inconsistent with the Hellings-Downs curve, such as clock errors with monopolar correlations [31,32] or Solar System ephemeris errors with dipolar correlations [33–36]. Thus, the detection of HD correlations through the off-diagonal terms of  $\phi_{\text{GW}}$  is considered the decisive factor in claiming a GWB detection, and the *CP* vs HD Bayes factor is used as a GWB detection statistic [4,6,7,11].

### C. Implementation and computational considerations

The standard PTA likelihood (22) requires the inverse noise covariance matrix  $\mathbf{C}^{-1}$  and therefore the inverse of  $\phi|_{\eta}$ . Although the PTA analysis software, such as ENTERPRISE [37], is optimized to speed up the likelihood evaluation, inversion becomes the most expensive computation when  $\phi|_{\eta}$  is not pulsar diagonal. For instance, for the NANOGrav 12.5 yr dataset each HD-model likelihood is  $\sim 25$  slower than the corresponding *CP* likelihood. This factor applies to 45 pulsars over a 12.9 yr dataset and will increase with the number of pulsars.

The current workhorse method to compute  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  is a hypermodel Markov chain Monte Carlo sampler [22,38,39]. In such an analysis, a discrete metaparameter tracks the current model (HD or *CP*) while the sampler jumps between them. The final Bayes factor is the number of samples in the HD model divided by the number of samples in the *CP* model. The two posteriors are also selected by the value of the metaparameter. As the evidence for a GWB becomes stronger, the HD model will be sampled more often than the *CP* model, slowing down calculations further.<sup>3</sup>

Despite the difference in likelihood functions and computation time, the posteriors for *CP* and HD are generally quite similar. Figure 1 displays the marginalized one- and two-dimensional  $\gamma_{\text{GW}}$  and  $\log_{10} A_{\text{GW}}$  posteriors and the  $\ln \mathcal{L}$  distributions for *CP* (blue) and HD (green), as recovered by hypermodel sampling. The similarity between the posteriors and the  $\sim 25\times$  likelihood speedup suggest that this problem is well suited for the reweighting method introduced in Sec. II.<sup>4</sup> The HD posterior created by reweighting the *CP* posterior is plotted in orange and is almost identical to the direct-sampling HD posterior. The efficiency of the reweighting method as posterior differences is discussed in the next section.

<sup>3</sup>A constant added to the *CP* log-likelihood can mitigate this particular issue and result in a comparable number of samples in each model. That constant should be close to the Bayes factor, which is unknown *a priori* in real data. In our study we estimated this constant by using the likelihood ratio between the *CP* and HD models evaluated at the injected parameters. This ensured that both models contained enough samples particularly in high Bayes factor regimes.

<sup>4</sup>Here we use identical priors between the target and the original distribution, meaning the prior weights  $w_{\pi}(\theta) = 1$ .

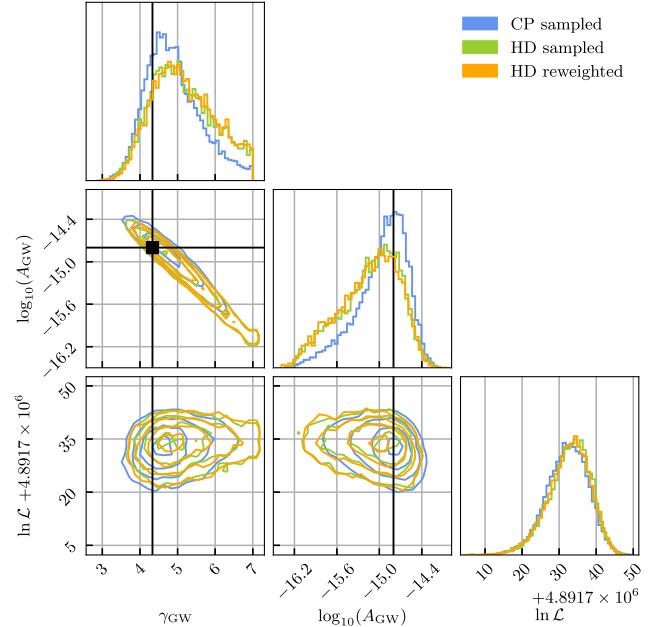


FIG. 1. Posteriors for  $\gamma_{\text{GW}}$  and  $\log_{10} A_{\text{GW}}$  and  $\ln \mathcal{L}$  distribution for simulated PTA data with a  $\log_{10}(A_{\text{GW}}) = -14.8$  GWB. We show histograms for direct sampling of *CP* (blue), HD (green), and for *CP*-to-HD reweighting (orange). Black lines indicate the injected values. For this plot we selected one of our simulations with the most visually different *CP* and HD GW posteriors. Even so, the direct-sampling and reweighted HD posteriors are almost identical. The reweighted posterior is well sampled, with 51% efficiency.

## IV. DEMONSTRATION OF THE METHOD

To show that we can safely reweight *CP* to HD, we simulate PTA datasets containing GWBs with different amplitudes and demonstrate that reweighting yields unbiased Bayes factors and posteriors.

### A. Bayes factors

To test  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  recovery, we simulate 100 datasets for 45 pulsars over 12.9 years, using maximum-likelihood red-noise hyperparameters from the 12.5 yr NANOGrav dataset [8,40].<sup>5</sup> Each simulation includes a power-law GWB with  $\log_{10} A_{\text{GW}}$  varying uniformly between  $-15$  and  $-14$ . We set  $\gamma_{\text{GW}}$  to  $13/3$ , the theoretical value for a GW background from supermassive black-hole binaries [41]. For each simulated dataset, we obtain a thinned set of *CP* posterior samples using PTMCMCSampler [38]. We reweight the *CP* posterior sample to the HD model and calculate  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  following Eq. (10). To verify the accuracy of these reweighted  $\mathcal{B}_{\text{CP}}^{\text{HD}}$ 's, we obtain an independent estimate from hypermodel runs on the same simulations. We compare the reweighted and hypermodel Bayes factors in Fig. 2, finding

<sup>5</sup>The NANOGrav 12.5 yr dataset is actually 12.9 years in length.

them in excellent agreement. The top panel shows the  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  estimates plotted against  $\log_{10} A_{\text{GW}}$ ; the bottom panel shows the relative difference of the  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  estimates (reweighted minus hypermodel, divided by their average). Marker colors encode reweighting efficiency. The mean relative difference is  $-0.5 \pm 4\%$ , so we observe no systematic effect. The maximum relative difference is 10%, small enough that it could not affect a GWB detection claim. Error bars are computed by combining (in quadrature) reweighted Bayes factor errors from Eq. (13) and hypermodel Bayes factor errors from the bootstrap method of [42]. Bayes factor differences are not strongly correlated with the injected GW amplitude or the Bayes factor, although the difference uncertainties are inversely correlated with efficiency [see Eq. (13)].

Figure 2 shows also that as we increase the simulated amplitude, the sampling efficiency tends to decrease. This

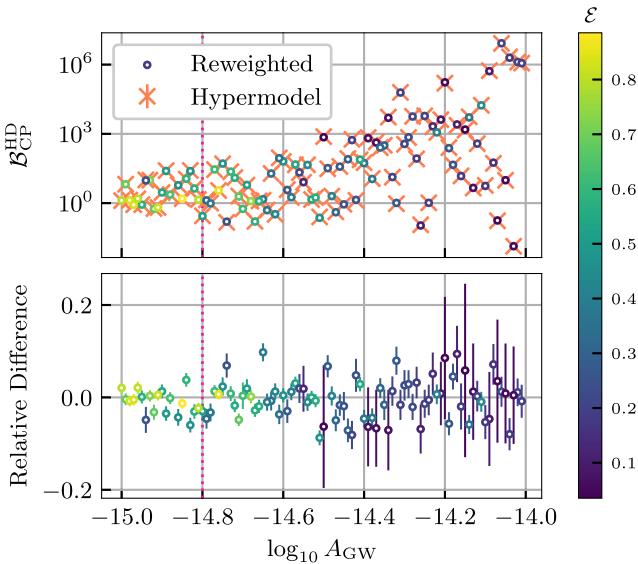


FIG. 2. Top:  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  vs simulated GWB amplitude. Bayes factors recovered via reweighting, Eq. (10), are colored by their efficiency  $\mathcal{E}$ , Eq. (12). Bayes factors recovered via the hypermodel are plotted as coral X's. The hypermodel error is calculated with a bootstrap method described in [42] whereas the reweighting error is estimated with Eq. (13), although both errors are too small to see. Bottom: relative difference in the hypermodel and recovered Bayes factors, again colored by efficiency. The error bars are propagated from the hypermodel and reweighting errors above. As the GWB amplitude increases, the efficiency decreases due to the distribution of the weights broadening as in Eq. (11). The relative difference between these Bayes factors is usually small, typically  $-0.5 \pm 4\%$ , but can be as large as 10%. A 10% difference in  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  is not large enough to change a detection conclusion to a nondetection conclusion or vice versa and therefore we can consider the difference small. For instance, a Bayes factor of 100 would lead to the same qualitative conclusion as a Bayes factor of 110. The pink vertical line in both plots is  $\log_{10} A_{\text{GW}} = -14.8$ , the posterior plotted in Fig. 1 to demonstrate that this posterior is typical.

is expected; as the amplitude of the GWB increases, the off-diagonal terms in Eq. (27) become more significant. The likelihood can then change between the two models significantly, which affects  $\bar{w}$ , and can even be maximized in different parts of parameter space. Such conditions can lead to a large spread in the weights as some points get heavily upweighted and others get downweighted. From Eq. (11), a large spread in the weights means that  $n_{\text{eff}}$  will decrease, and more samples from the *CP* distribution will be needed in order to faithfully represent the HD posterior and calculate  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  accurately: see Eq. (13). In our simulated datasets, however, the recovered Bayes factor remains within 10% of that calculated with the hypermodel even in regions where  $\mathcal{B}_{\text{CP}}^{\text{HD}} > 10^6$ .

In order to study the relation between the model posterior similarity and the efficiency of the reweighting procedure, we compute the Kullback-Leibler divergence [43], which quantifies the difference between two distributions. We plot the relationship between the KL divergence and the efficiency in Fig. 3. The upper plot shows total KL divergence [Eq. (16)] vs efficiency [Eq. (12)] for the *CP* and HD posteriors. As the KL divergence increases, the posteriors become more distinct and the sampling efficiency decreases. The bottom plot shows the fractional contributions of different model parameters to the total KL divergence.<sup>6</sup> We split the 92 parameters into four sets: the IRN amplitudes and spectral indices (pink and red, respectively) and the GW background amplitude and spectral index (blue and gold, respectively). We compute the partial KL divergence of the *CP* and HD marginalized posteriors for each parameter and sum those of the IRN parameters. The fractional contribution is then obtained by dividing those partial KL divergences by the total. The set of all red-noise parameters contributes more to the total divergence than do the GWB parameters individually. The set of all IRN amplitude posteriors is the major contributor to the divergence ( $55 \pm 11\%$ ), followed by the set of all IRN spectral indices ( $27 \pm 8\%$ ); the contribution from  $A_{\text{GW}}$  and  $\gamma_{\text{GW}}$  are roughly equivalent at percent level,  $9 \pm 8\%$  each.

## B. Posterior recovery

Figure 1 offered visual confirmation that the GWB parameter posteriors under the HD model are recovered without bias via reweighting. In this section we confirm these initial findings through a more extensive percentile-percentile (P-P) test [44]. We generate 100 simulations similar to those described in Sec. IV A, except that each

<sup>6</sup>The total KL divergences are not directly comparable to the KL divergences of the various marginalized posteriors; the total KL divergence is equal to the sum of the marginalized KL divergences only when parameters are uncorrelated. Although this is not the case in our analysis, the normalized marginal KL divergences still inform us of parameters that most greatly influence the total KL divergence.

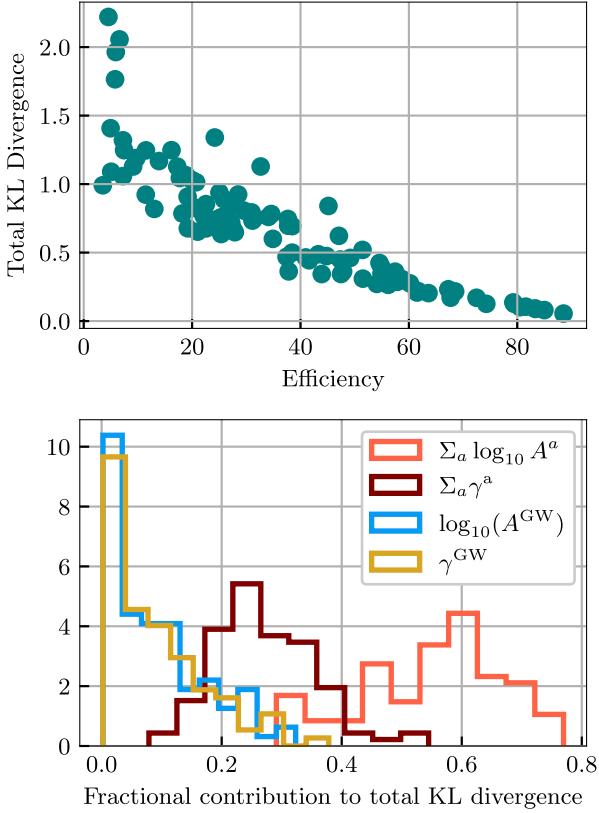


FIG. 3. Top: total KL divergence, Eq. (16), vs efficiency, Eq. (12), between the  $CP$  and the reweighted HD posterior. As the KL divergence increases, the posteriors become more distinct, and the sampling efficiency decreases. Bottom: fractional contributions to total KL divergence from sets of parameters including all the IRN amplitudes and spectral indices (pink and red, respectively) and the GWB amplitude and spectral index (blue and gold, respectively). The set of all IRN parameters contribute more to the total divergence than the GWB parameters individually.

simulated parameter is drawn from its analysis prior, as required to achieve Bayesian coverage. The priors for the spectral indices are  $\gamma_{GW}, \gamma^a \in U[2, 6]$ , and the priors for the amplitudes are  $\log_{10} A_{GW} \in U[-15, -12]$  and  $\log_{10} A^a \in U[-16, -14]$ . We recover  $CP$  posteriors from these simulations with direct sampling, and then reweight and resample those posteriors to the HD model.

The P-P test is a standard measure of bias in recovered posteriors. Datasets are first simulated by drawing parameters from their priors and adding Gaussian noise. The posterior of each dataset is then sampled. The percentile of each of the “true” or injected values is calculated in the marginalized, one-dimensional posterior of each parameter. For a set of unbiased posteriors, the injected value will be distributed according to each posterior. That is, the percentile of where each injected value lands in a 1D marginalized posterior will be distributed uniformly between the 0th and 100th percentile, the x-axis of Fig. 4. This test of uniformity in posterior space is represented with the cumulative

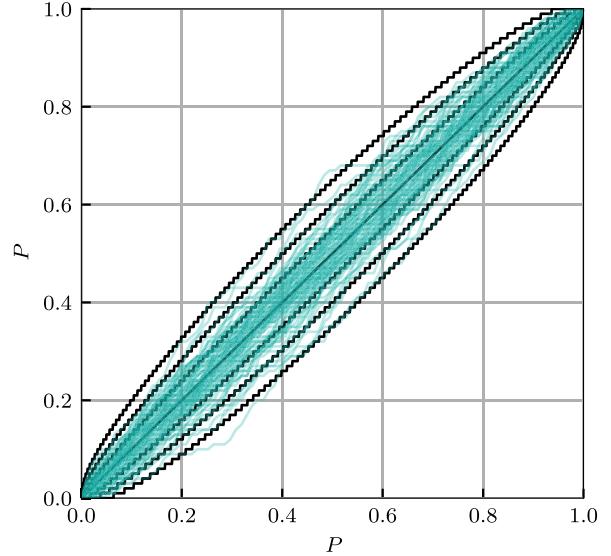


FIG. 4. P-P plots for all 92 reweighted, HD-model parameters (teal) with the 1-, 2-, and 3- $\sigma$  standard deviations (black). The y-axis is the percentile of each parameter’s injected value in its marginalized posterior. The x-axis is the percentile of the sorted y-axis values. The recovered posteriors are consistent with expectations, suggesting that the posterior recovery is unbiased.

distribution function (CDF) of the posterior’s percentile. Since the CDF of a uniform distribution between 0 and 1 (0th and 100th percentile) is a line of slope 1, the P-P plot is usually represented this way. A P-P plot showing a line consistently below (above) the line  $x = y$  is indicative of parameter bias of overestimating (underestimating) the parameter value. An S-curve going above (below) then below (above) the diagonal is indicative of an overestimate (underestimate) of the posterior’s standard deviation.

Figure 4 shows the corresponding P-P plots. The 92 different parameters ( $\gamma^a, A^a$  for 45 pulsars as well as  $\gamma_{GW}^a, A_{GW}^a$  for the GWB) are plotted in teal. The expected 1-, 2-, and 3- $\sigma$  confidence intervals are plotted in black. The recovered posteriors agree with expectations; only two lines briefly leave the three-sigma error bars. This suggests that the reweighting method neither over- nor underestimates parameters systematically, as would be the case if some parameters were always above or below the diagonal; nor does it recover incorrect variance, as would be indicated by S-curves around the diagonal.

### C. Bayes factor recovery on extended dataset

To this point, we have demonstrated that likelihood reweighting is a promising tool for recovering accurate Bayes factors and unbiased posteriors in a simulated dataset with 12.9 years of timing data, 45 pulsars, and a range of injected GWB amplitudes. As PTAs continue to collect more data, it becomes natural to ask at what point the reweighting scheme could fail, either by misestimating Bayes factors or by exhibiting low efficiencies. We examine

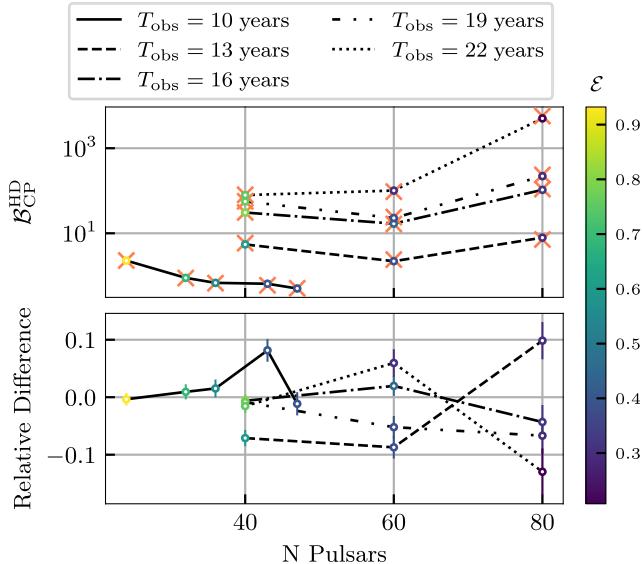


FIG. 5. Top:  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  vs number of pulsars for sets of fixed observation times between 10 and 22 years,  $T_{\text{obs}}$ . Bayes factors recovered via reweighting, Eq. (10), are colored by their efficiency  $\mathcal{E}$ , Eq. (12). Bayes factors recovered via the hypermodel are plotted as coral X's. The error estimate of each point is described in the caption of Fig. 2. For a fixed number of pulsars, an increase in the observation time leads to a higher Bayes factor. In each case, as  $N$  increases, the efficiency decreases. Additionally, as the observation time increases, efficiency tends to decrease, albeit less distinctly. The relative difference between the direct sampling and reweighting Bayes factors remains quite small and is independent of both  $T_{\text{obs}}$  and  $N$ . The small errors and high efficiencies (each greater than 20%) imply that likelihood reweighting remains reliable when additional time and pulsars are added.

the performance of likelihood reweighting after the addition of additional pulsars and additional observation time to the dataset. We find that, even for 80 pulsars and 22 years of data, the efficiency remains above 20% and the errors between the Bayes factor calculated with direct sampling and reweighting are comparable to those in Sec. IV A.

To create this extended dataset, we simulate realistic pulsars and add additional observing time to each pulsar. To create new pulsars, we sample sky locations by fitting existing pulsar locations with a kernel density estimate and sample from it. Each new pulsar is assigned white-noise parameters and observing epochs (plus Gaussian scatter) from an existing pulsar. To simulate additional years of data, to each pulsar we add TOA with Gaussian scatter. The red-noise parameters for the new pulsars are drawn from the IRN prior. For existing pulsars, the red-noise amplitudes were set from the maximum-likelihood draw as in Sec. IV A. The GWB was injected with  $A_{\text{GW}} = 1.92 \times 10^{-15}$ , the median posterior amplitude of the NANOGrav 12.5 yr analysis [4]. In total, we simulated 22 years of data in 80 pulsars; below we present results based subsets of that data.

In Fig. 5 we plot  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  as a function of the number of pulsars  $N$  and for different observation durations  $T_{\text{obs}}$ . We find that the relative difference between the  $\mathcal{B}_{\text{CP}}^{\text{HD}}$  recovered by direct sampling and by reweighting remain within 10% of each other, suggesting that the reweighting scheme remains valid for these extended datasets. Moreover, we find that while the ratio between the HD and CP likelihood computation times is approximately constant across extended observation time, the ratio scales with the number of pulsars (ranging between 10 and 40). Thus, as more pulsars are added to the dataset, reweighting becomes more important.

## V. DISCUSSION AND CONCLUSIONS

We have introduced a reweighting method to efficiently and reliably obtain GW posterior and marginal likelihood for a GWB model in PTA data analysis. We first compute an inexpensive approximate posterior (CP) that omits pulsar-pulsar correlations, then reweight it to a full posterior (HD) that includes them. We have validated this method by comparing reweighted posteriors and Bayes factors with distributions and factors obtained with direct sampling. Reweighting appears to be reliable and unbiased. Even in cases with low reweighting efficiency (as defined by the reduction in the number of effective samples), the reweighted Bayes factor estimate remained robust up to  $\mathcal{B}_{\text{CP}}^{\text{HD}} > 10^6$ , far larger than required for a confident GWB detection.

Even though our method requires evaluating the computationally expensive HD likelihood, it is still much more efficient than direct stochastic sampling. This is due to the additional evaluations required for the latter, which do *not* need to be repeated when reweighting. Direct sampling results in very autocorrelated sample chains, which are thinned [by factors  $N_t \sim \mathcal{O}(10^3)$ , on the order of the chain autocorrelation length] to obtain quasi-independent samples. By contrast, reweighting is applied *after* thinning, reducing the number of HD likelihood evaluations by  $N_t$ . In addition, the weights of Eq. (4) can be computed in parallel on multiple cores, allowing a further wall clock speedup (by the number  $N_p$  of parallel processes). Finally, if parallel tempering was used to sample the approximate model, only samples from the coldest chain should need be reweighted, decreasing the necessary number of computations by a factor of the number of chains  $N_c \sim \mathcal{O}(10)$ .

While the reweighting procedure is mathematically exact, the method is subject to sampling error; reweighted posteriors could have too few effective samples to accurately reflect the true distribution. Constructing generic diagnostic tools for such situations can be challenging, as the effective number of independent samples  $n_{\text{eff}}$  can vary between applications. In such cases, estimating the Bayes factor sampling error or inspecting posteriors visually can help identify undersampling. If  $n_{\text{eff}}$  is low, a few strategies are available. The simplest is to increase the number of samples

for the approximate model. A more sophisticated option involves the importance sampling of the approximate model by concentrating on the region of parameter space that the target seems to prefer. In the most extreme case, so many approximate-model samples are needed that the method becomes less efficient than direct sampling. This happens when the efficiency drops to the ratio of likelihood computation times (e.g., to 1/25 for the NANOGrav 12.5 yr dataset). If parallel tempering is used, then “hot” chains, with a correspondingly broader posterior, could be used in situations where efficiency is low due to a lack of samples from the approximate distribution available to estimate tails in the target distribution.

The reweighting formalism is generic and can be applied to any pair of approximate and target distributions. For example, one could model a clock error by including a process with monopolar correlations in addition to the HD correlations. In this situation, extra parameters are added to the target model, which requires drawing samples from some proposal distribution for the new parameters (see [45,46] for examples of reweighting between models with varying numbers of parameters). In practice, sampling error (efficiency) increases (decreases) if the approximate and target posteriors do not overlap, as quantified in Fig. 3 using the Kullback-Leibler divergence.

Throughout this work we have presented examples that are based on the NANOGrav 12.5 yr analysis. Although our simulations are consistent with the NANOGrav dataset and our understanding of the stochastic GWB, we have not simulated realistic radio frequency noise such as dispersion measure variations or solar wind fluctuations. More “advanced” noise modeling adds numerous extra parameters to each pulsar to measure chromatic effects [31,33,47–49] increasing the complexity of the analysis. Given that most of these additional parameters impact only individual pulsar measurements, a factorized-likelihood approach to estimate the *CP* model, followed by this reweighting scheme could significantly reduce the wall clock time of an analysis that uses more advanced noise models.

In the context of PTA searches for GWBs, the reweighting formalism introduced in this paper offers an accurate and computationally efficient shortcut to GW posteriors and HD vs *CP* Bayes factors. In this paper we tested the method on simulated datasets with increasing GWB amplitudes, which served a proxy for increased observing time and number of pulsars. Our results suggest that reweighting remains robust for PTA datasets with Bayes factors of at least  $10^6$ , orders of magnitude larger than current results. Thus, our method can reliably characterize the GWB from PTA datasets for the foreseeable future and into the detection regime.

## ACKNOWLEDGMENTS

We thank Ethan Payne for useful discussions about reweighting and Ken Olum, Steve Taylor, and Paul Baker for comments on the paper draft. Numerical investigations were performed using services provided by the OSG Consortium [50,51], which is supported by the National Science Foundation Grants No. 2030508 and No. 1836650. Additional computing resources were provided by Caltech’s Theoretical AstroPhysics Including Relativity and Cosmology (TAPIR) group. In addition to ENTERPRISE [37,52], our software stack included SCIPY [53], MATPLOTLIB [54], NUMPY [55], PANDAS [56], and CORNER [57]. S. H. and P. M. M. acknowledge the VIPER PTA Summer School at Vanderbilt University, which was funded under NSF CAREER-2146016. A. D. J. and K. C. acknowledge support from the Caltech and Jet Propulsion Laboratory President’s and Director’s Fund and the Sloan Foundation. S. H. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1745301. P. M. M. and M. V. were supported by the NANOGrav Physics Frontiers Center, National Science Foundation (NSF), Grant No. 2020265. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

---

- [1] S. Detweiler, *Astrophys. J.* **234**, 1100 (1979).
- [2] R. W. Hellings and G. S. Downs, *265*, L39 (1983).
- [3] S. R. Taylor, [arXiv:2105.13270](https://arxiv.org/abs/2105.13270).
- [4] Z. Arzoumanian *et al.* (NANOGrav), *Astrophys. J. Lett.* **905**, L34 (2020).
- [5] B. Goncharov *et al.*, *Astrophys. J. Lett.* **917**, L19 (2021).
- [6] S. Chen *et al.*, *Mon. Not. R. Astron. Soc.* **508**, 4970 (2021).
- [7] J. Antoniadis *et al.*, *Mon. Not. R. Astron. Soc.* **510**, 4873 (2022).
- [8] Z. Arzoumanian *et al.* (NANOGrav Collaboration), *Astrophys. J.* **859**, 47 (2018).
- [9] N. Pol, S. R. Taylor, and J. D. Romano, *Astrophys. J.* **940**, 173 (2022).
- [10] A. Zic, G. Hobbs, R. M. Shannon, D. Reardon, B. Goncharov, N. D. R. Bhat, A. Cameron, S. Dai, J. R. Dawson, M. Kerr *et al.*, *Mon. Not. R. Astron. Soc.* **516**, 410 (2022).
- [11] B. Goncharov *et al.*, *Astrophys. J.* **932**, L22 (2022).

[12] S. R. Taylor, J. Simon, L. Schult, N. Pol, and W. G. Lamb, *Phys. Rev. D* **105**, 084049 (2022).

[13] N. J. Cornish and L. Sampson, *Phys. Rev. D* **93**, 104047 (2016).

[14] S. R. Taylor, L. Lentati, S. Babak, P. Brem, J. R. Gair, A. Sesana, and A. Vecchio, *Phys. Rev. D* **95**, 042002 (2017).

[15] K. J. Lee, C. G. Bassa, R. Karuppusamy, M. Kramer, R. Smits, and B. W. Stappers, *Mon. Not. R. Astron. Soc.* **423**, 2642 (2012).

[16] M. T. Lam, M. A. McLaughlin, J. M. Cordes, S. Chatterjee, and T. J. W. Lazio, *Astrophys. J.* **861**, 12 (2018).

[17] M. T. Lam, *Astrophys. J.* **868**, 33 (2018).

[18] L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren, *Phys. Rev. D* **87**, 104021 (2013).

[19] R. van Haasteren and M. Vallisneri, *Phys. Rev. D* **90**, 104012 (2014).

[20] R. van Haasteren, Y. Levin, P. McDonald, and T. Lu, *Mon. Not. R. Astron. Soc.* **395**, 1005 (2009).

[21] G. E. Freedman, A. D. Johnson, R. van Haasteren, and S. J. Vigeland, *Phys. Rev. D* **107**, 043013 (2023).

[22] S. Hee, W. J. Handley, M. P. Hobson, and A. N. Lasenby, *455*, 2461, (2016).

[23] E. Payne, C. Talbot, and E. Thrane, *Phys. Rev. D* **100**, 123017 (2019).

[24] V. Elvira, L. Martino, and C. P. Robert, [arXiv:1809.04129](https://arxiv.org/abs/1809.04129).

[25] A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry, Pareto smoothed importance sampling (2015), [arXiv:1507.02646](https://arxiv.org/abs/1507.02646).

[26] R. van Haasteren and Y. Levin, *Mon. Not. R. Astron. Soc.* **428**, 1147 (2013).

[27] Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. J. Chamberlin, S. Chatterjee, B. Christy, J. M. Cordes, N. J. Cornish, K. Crowter, P. B. Demorest *et al.*, *Astrophys. J.* **821**, 13 (2016).

[28] S. R. Taylor, J. Simon, and L. Sampson, *Phys. Rev. Lett.* **118**, 181102 (2017).

[29] L. Sampson, N. J. Cornish, and S. T. McWilliams, *Phys. Rev. D* **91**, 084055 (2015).

[30] J. D. Romano, J. S. Hazboun, X. Siemens, and A. M. Archibald, *Phys. Rev. D* **103**, 063027 (2021).

[31] C. Tiburzi, G. Hobbs, M. Kerr, W. Coles, S. Dai, R. Manchester, A. Possenti, R. Shannon, and X. You, *Mon. Not. R. Astron. Soc.* **455**, 4339 (2016).

[32] M. Tinto, *Phys. Rev. D* **97**, 084047 (2018).

[33] B. Goncharov *et al.*, *Mon. Not. R. Astron. Soc.* **502**, 478 (2021).

[34] M. Vallisneri *et al.* (NANOGrav Collaboration), [arXiv:2001.00595](https://arxiv.org/abs/2001.00595).

[35] R. N. Caballero *et al.*, *Mon. Not. R. Astron. Soc.* **481**, 5501 (2018).

[36] E. Roeber, *Astrophys. J.* **876**, 55 (2019).

[37] J. A. Ellis, M. Vallisneri, S. R. Taylor, and P. T. Baker, ENTERPRISE: Enhanced numerical toolbox enabling a robust pulsar inference suite, Zenodo, [10.5281/zenodo.4059815](https://doi.org/10.5281/zenodo.4059815) (2020).

[38] J. Ellis and R. van Haasteren, *jellis18/ptmcmc sampler*: Official release, [10.5281/zenodo.1037579](https://doi.org/10.5281/zenodo.1037579) (2017).

[39] S. R. Taylor, R. van Haasteren, and A. Sesana, *Phys. Rev. D* **102**, 084039 (2020).

[40] NANOGrav Scientific Collaboration, Nanograv data, <https://data.nanograv.org/>.

[41] E. S. Phinney, [arXiv:astro-ph/0108028](https://arxiv.org/abs/astro-ph/0108028).

[42] D. W. Heck, A. M. Overstall, Q. F. Gronau, and E.-J. Wagenmakers, *Stat. Comput.* **29**, 631 (2018).

[43] S. Kullback and R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).

[44] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, Fourth Edition (Marcel Dekker, Inc., New York, NY, 2003), p. 145.

[45] I. M. Romero-Shaw, N. Farrow, S. Stevenson, E. Thrane, and X.-J. Zhu, *Mon. Not. R. Astron. Soc.* **496**, L64 (2020).

[46] I. M. Romero-Shaw, P. D. Lasky, E. Thrane, and J. C. Bustillo, *Astrophys. J. Lett.* **903**, L5 (2020).

[47] C. Tiburzi *et al.*, *Mon. Not. R. Astron. Soc.* **487**, 394 (2019).

[48] C. Tiburzi *et al.*, *Astron. Astrophys.* **647**, A84 (2021).

[49] J. S. Hazboun *et al.* (NANOGrav Collaboration), *Astrophys. J.* **929**, 39 (2022).

[50] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick, in *J. Phys. Conf. Ser.* **78**, 012057 (2007).

[51] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein, in *2009 WRI World Congress on Computer Science and Information Engineering* (IEEE Computer Society, Washington, DC, 2009), Vol. 2, pp. 428–432.

[52] S. R. Taylor, P. T. Baker, J. S. Hazboun, J. Simon, and S. J. Vigeland, Enterprise extensions (2021), v2.3.3, [https://github.com/nanograv/enterprise\\_extensions](https://github.com/nanograv/enterprise_extensions).

[53] P. Virtanen *et al.* (SciPy 1.0 Contributors), *Nat. Methods* **17**, 261 (2020).

[54] J. D. Hunter, *Comput. Sci. Eng.* **9**, 90 (2007).

[55] C. R. Harris *et al.*, *Nature (London)* **585**, 357 (2020).

[56] The Pandas Development Team, *pandas-dev/pandas*: PAN-DAS, [10.5281/zenodo.7794821](https://doi.org/10.5281/zenodo.7794821) (2020).

[57] D. Foreman-Mackey, *J. Open Source Software* **1**, 24 (2016).