The Astrophysical Journal, 932:105 (8pp), 2022 June 20 https://doi.org/10.3847/1538-4357/ac6f5e

Gravitational-wave Statistics for Pulsar Timing Arrays: Examining Bias from Using a Finite Number of Pulsars

Aaron D. Johnson[1], Sarah J. Vigeland[1], Xavier Siemens[1,2], and Stephen R. Taylor[3]

[1] Center for Gravitation, Cosmology and Astrophysics, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, WI 53201, USA; johnsoad@uwm.edu

[2] Department of Physics, Oregon State University, Corvallis, OR 97331, USA

[3] Department of Physics and Astronomy, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37235, USA

Received 2022 January 25; revised 2022 May 10; accepted 2022 May 11; published 2022 June 21

Abstract

Recently, many different pulsar timing array (PTA) collaborations have reported strong evidence for a common stochastic process in their data sets. The reported amplitudes are in tension with previously computed upper limits. In this paper, we investigate how using a subset of a set of pulsars biases Bayesian upper limit recovery. We generate 500 simulated PTA data sets, based on the NANOGrav 11 yr data set with an injected stochastic gravitational-wave background (GWB). We then compute the upper limits by sampling the individual pulsar likelihoods, and combine them through a factorized version of the PTA likelihood to obtain upper limits on the GWB amplitude, using different numbers of pulsars. We find that it is possible to recover an upper limit (95% credible interval) below the injected value, and that it is significantly more likely for this to occur when using a subset of pulsars to compute the upper limit. When picking pulsars to induce the maximum possible bias, we find that the 95% Bayesian upper limit recovered is below the injected value in 10.6% of the realizations (53 of 500). Further, we find that if we choose a subset of pulsars in order to obtain a lower upper limit than when using the full set of pulsars, the distribution of the upper limits obtained from these 500 realizations is shifted to lower-amplitude values.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858)

1. Introduction

Pulsar timing arrays (PTAs) aim to detect gravitational waves in the nanohertz frequency regime by looking for correlations between times of arrival of radio signals from millisecond pulsars (MSPs; Taylor 2021). Pulsar timing models predict the pulse times of arrival from a pulsar based on that pulsar's astrophysical properties. The differences between the predicted and measured times of arrival are the timing residuals (Verbiest et al. 2021). By modeling these residuals, we attempt to reveal gravitational-wave signals hidden in our data.

The first such gravitational-wave signal detected by PTAs is expected to come from a stochastic gravitational-wave background (GWB) made up of gravitational waves emitted by a cosmological population of supermassive binary black holes (SMBBHs; Rosado et al. 2015). Assuming that these SMBBHs are circular and only evolve due to gravitational-wave emission, the characteristic strain spectrum is given by Phinney (2001):

$$h_c(f) = A_{\rm GWB}\left(\frac{f}{f_{\rm yr}}\right)^{-2/3}, \tag{1}$$

where the amplitude $A_{\rm GWB}$ depends on the SMBBH population and galaxy merger rate, and $f_{\rm yr}$ is the reference frequency corresponding to $1\ {\rm yr}^{-1}$. Based on models of the SMBBH population, we expect to detect the GWB with PTAs within the next 5 yr (Taylor et al. 2016; Pol et al. 2021).

Recently, multiple PTAs have found evidence of a common spectrum stochastic process. The NANOGrav collaboration reported a common spectrum process with a median strain amplitude of 1.92×10^{-15} in an analysis of their 12.5 yr data set (Arzoumanian et al. 2020); the PPTA reported a median amplitude of 2.2×10^{-15} (Goncharov et al. 2021); the EPTA reported a median amplitude of 2.95×10^{-15} (Chen et al. 2021); and the IPTA, which used combined data from its constituent collaborations' older data sets, reported a median amplitude of 2.8×10^{-15} (Antoniadis et al. 2022). The amplitude of this process is in tension with some previously published upper limits on the amplitude of the GWB.

The NANOGrav collaboration placed an upper limit of $A < 1.45 \times 10^{-15}$, based on an analysis of 34 pulsars timed for up to 11 yr (Arzoumanian et al. 2018). These pulsars consist of the ones from the 11 yr data set that have been timed for more than 3 yr. The PPTA placed an upper limit of $A < 10^{-15}$, based on an analysis of the four pulsars that had the highest timing precision, which were timed for up to 11 yr (Shannon et al. 2015). The EPTA placed an upper limit of $A < 3 \times 10^{-15}$, based on an analysis of six pulsars timed for up to 18 yr (Lentati et al. 2015). These six pulsars were chosen to minimize the dimensionality required to search over. Additionally, the least sensitive pulsar of the six affected the result at the 2% level.

There are several possible explanations for this apparent discrepancy between the earlier results and the most recent ones. Early work did not model the uncertainty in the position of the solar system barycenter, and, as shown in Vallisneri et al. (2020), the choice of solar system ephemeris can significantly affect detection statistics. The choice of the prior on the pulsar's intrinsic red noise also has a significant effect on the upper limit on a common stochastic process, as shown in Hazboun et al. (2020), due to the covariance between the two.

Here, we investigate how Bayesian GWB upper limits are affected by the use of a finite number of pulsars. We generate simulated PTA data with an injected GWB, and compare the Bayesian upper limits computed by analyzing the entire PTA versus using only a subset of the pulsars. We show that a wide range of possible upper limits can be computed when only a small number of pulsars are used to compute the upper limit. Furthermore, it is possible to find an upper limit that is lower than the injected value of the GWB, and this occurs more often when using a subset of pulsars.

This paper is organized as follows. In Section 2, we discuss the procedure by which we simulate the pulsars and compute the upper limits. Section 3 details how the upper limits change with different combinations of pulsars. Finally, in Section 4, we discuss our results and make concluding remarks about what this means for the number of pulsars that are used in PTA data sets.

2. Methods

We use methods that are, to a large extent, the same as those in previous papers that set Bayesian upper limits (Lentati et al. 2015; Shannon et al. 2015; Arzoumanian et al. 2016, 2018). The significant differences include using a factorized PTA likelihood and grid-approximating the posterior for each individual pulsar, instead of sampling using a Metropolis–Hastings Markov Chain Monte Carlo algorithm. All models here, as in previous papers setting Bayesian upper limits, use a 30-frequency power-law pulsar intrinsic red noise and GWB given by

$$\rho(f) = \frac{A^2}{12\pi^2}\frac{1}{T}\left(\frac{f}{\mathrm{yr}^{-1}}\right)^{-\gamma}\mathrm{yr}^2, \tag{2}$$

where $\rho(f) = S(f)\Delta f$, where $S(f)$ is the power spectral density and $\Delta f = 1/T$.

All of the results in this paper use Bayesian methods. The frequentist methods that have been used previously to set upper limits via the optimal statistic (Anholm et al. 2009; Demorest et al. 2012; Chamberlin et al. 2015) are not considered here. Importantly, the Bayesian and frequentist upper limits have different interpretations and do not coincide in general (Röver et al. 2011). Furthermore, the optimal statistic, which was used to set upper limits in Arzoumanian et al. (2016), only looks at the cross-correlations between different pulsars, while the Bayesian methods used to set upper limits in previous papers look only at autocorrelations, so the two are fundamentally different and it is difficult to compare them.

2.1. Factorized Likelihood

When interpulsar correlations are not included, the PTA likelihood can be factored into a product of individual pulsars (Arzoumanian et al. 2020; Taylor et al. 2022):

$$p(\{d_j\}_N|\{\boldsymbol{\theta}_j\}_N, A_{\mathrm{GWB}}) = \prod_{j=1}^{N} p(d_j|\boldsymbol{\theta}_j, A_{\mathrm{GWB}}), \tag{3}$$

where d_j are the data, θ_j are the intrinsic noise parameters, and A_{GWB} is the common red process amplitude for the jth pulsar. Using the factorized likelihood allows for the rapid computation of upper limits with more than one pulsar. We use a grid approximation on the individual pulsar models, as described in Section 2.3, then multiply the marginalized common red process amplitude posteriors for each pulsar that we want in the combined upper limit. These new posteriors are then reweighted from a log-uniform to a uniform prior on A_{GWB}, by multiplying by

$$f(x) = \frac{10^x}{10^{x_{\max}} - 10^{x_{\min}}}, \tag{4}$$

where $x = \log_{10} A_{\mathrm{GWB}}$, and $x_{\max}$, $x_{\min}$ are the maximum and minimum values of the uniform prior for the log amplitude, respectively. From this reweighted marginalized posterior, we can take the 95% Bayesian upper limit easily, by interpolating the posterior and using a cumulative sum until we reach 0.95,

$$\sum_{x_{\min}}^{x_{95\%}} p(A_{\mathrm{GWB}}|d) = 0.95, \tag{5}$$

and then finding the $\log_{10} A_{\mathrm{GWB}}$ value corresponding to $x_{95\%}$ where the sum was truncated. All the following discussions use the 95% upper limit as computed here.

2.2. Simulations

We simulate 500 sets of pulsars using `TEMPO2` (Edwards et al. 2006; Hobbs et al. 2006) and `libstempo` (Vallisneri 2020), with the observation baselines, observing cadences, and noise properties based on the 11 yr NANOGrav data set (Arzoumanian et al. 2018). The full 11 yr NANOGrav data set contains 45 pulsars. Due to the large number of upper limits that need to be computed for the following sections, we only simulate 22 of the 45 pulsars that have been timed for more than 6 yr. Pulsars with shorter timing baselines contribute less to the upper limit than ones that have been observed for many years. Because of this, we do not expect that removing these pulsars will significantly affect the results here.

Each pulsar contains white noise related to the uncertainty in the pulsar times of arrival, intrinsic red noise similar to that in the 11 yr data set, and an injected GWB with an amplitude $A_{\mathrm{GWB}} = 10^{-15}$ and a spectral index $\gamma_{\mathrm{GWB}} = 13/3$. The GWB injection includes Hellings and Downs cross-correlations (Hellings & Downs 1983), but we only use the autocorrelations to set the upper limits, as was done in many previous PTA papers (Shannon et al. 2015; Lentati et al. 2016; Arzoumanian et al. 2016, 2018). Lentati et al. (2015) did use cross-correlations, but found that the upper limits were consistent with their autocorrelation-only analysis. Similarly, we also find that including cross-correlations does not change the upper limit in the cases where the upper limit falls below the injected amplitude. The autocorrelations dominate the recovery of a GWB when the number of pulsars N_p is relatively small, since all of the cross-correlation coefficients are less than 1; however, for large numbers of pulsars, the cross-correlations become more significant, since the number of cross-correlation terms increases as $\mathcal{O}(N_p^2)$. In this particular set of simulated pulsars, the cross-correlations are too weak compared to the autocorrelations to affect the upper limits.

2.3. Software and Implementation

We use `enterprise` (Ellis et al. 2020) to set up a model for our simulated pulsar sets, with priors as in Table 1. Here, we use a grid approximation to obtain each pulsar's posterior, which is rendered effective by the low dimensionality of the parameter space. We use a power-law model for both the red

Table 1
Priors for the Model Used to Analyze Each Individual Simulated Pulsar

Parameter	Description	Interval
$A_{\rm RN}$	log-uniform	[−20, −11]
$\gamma_{\rm RN}$	uniform	[0, 7]
$A_{\rm GWB}$	log-uniform	[−20, −11]
$\gamma_{\rm GWB}$	constant	13/3
EFAC	constant	1

Note. The intrinsic red noise parameters have been labeled with "RN," and the common red process amplitude and spectral index have been labeled with "GWB." The spectral index for the common red process has been fixed in each model to 13/3. EFAC is a multiplicative factor on the time of arrival uncertainties.

noise intrinsic to each pulsar and the red noise common among all the pulsars. Therefore, we set up our grid for each individual pulsar over (1) the intrinsic red noise amplitude, (2) the intrinsic red noise spectral index, and (3) the common red process amplitude. Care must be taken, since models from `enterprise` return a log-likelihood. To facilitate the use of the grid approximation, we subtract the maximum log-likelihood evaluated on the grid from all points before exponentiation to return the posterior that could then be marginalized. We use a Nelder–Mead algorithm (Gao & Han 2012; Virtanen et al. 2020) to find this maximum, starting from several random locations in the parameter space. Once the maximum is found, we evaluate 300 points of the $\log_{10} A_{\rm GWB}$ marginalized posterior. To evaluate each point, we marginalize over the intrinsic red noise parameters simultaneously using `scipy.integrate.dblquad` (Virtanen et al. 2020). This reduces the number of evaluations by allowing the adaptive integration routine to decide how many points are required, instead of using a uniform grid. We model each pulsar individually and then postprocess using the factorized likelihood, as discussed in Section 2.1.

3. Results

Using the above methods and software, we investigate how bias may appear when using a subset of pulsars. We start this section with three specific combinations and discuss how their cumulative upper limits change as we add more pulsars. Next, we generalize to all possible combinations and average over all realizations to discover trends in how adding more pulsars affects the distribution of the upper limits that are possible. We then investigate a single combination for every realization to see how the upper limits change as we add more pulsars. Some of these pulsars hold more influence over the upper limits than others. By using the Kullback–Leibler divergence (KL; Kullback & Leibler 1951), we enumerate and examine these pulsars. Finally, we investigate bias by comparing the distributions of the upper limits obtained when computing the minimum cumulative upper limit given by a sequence of 22 pulsars for all 500 realizations.

3.1. Combinations and Upper Limits

One of the goals of this work is to investigate how the choice of which pulsars to use affects the computation of the upper limits. Initially, we consider three different combinations: time span, single-pulsar upper limit, and a combination that uses a greedy algorithm to get a low upper limit.

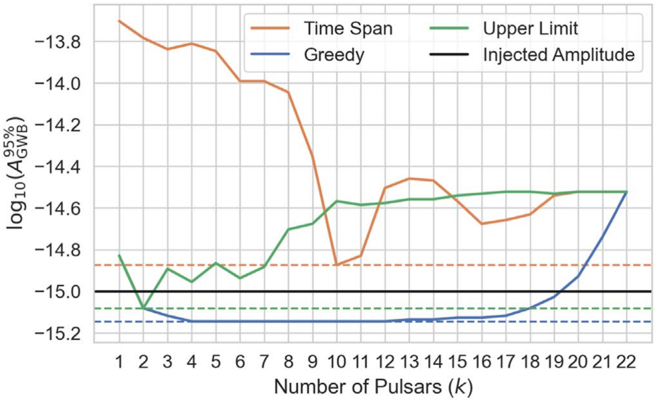

Figure 1. Cumulative upper limits computed using three different upper limit combinations. While the different combinations vary significantly, they agree once the last pulsar has been added. The dashed lines are the minimum values that are achieved by each combination. Both the single-pulsar upper limit and the greedy upper limit combinations drop below the injected value when a subset of the pulsars is used.

1. The time span combination is, as its name suggests, a combination that sorts pulsars by their observation time spans.
2. After taking each pulsar and individually computing an upper limit, we can order these upper limits from lowest to highest. We call this the single-pulsar upper limit combination. This was the combination used in the NANOGrav 9 yr stochastic GWB search (Arzoumanian et al. 2016). When using this method, the upper limit dropped to a minimum, then increased with each added pulsar, until it eventually saturated.
3. The last combination is a greedy algorithm in which we build up the upper limit pulsar by pulsar. The lowest individual pulsar upper limit takes the first slot in the combination. Next, the upper limit is computed for the first slot and each of the remaining 21 pulsars. The pulsar from the remaining 21 that gives the lowest two-pulsar upper limit is put into the second slot. By continuing in this fashion until all the pulsars have been used, we find that the combination attains a minimum that is much lower than the time span combination.

Because there is only one combination when using all 22 pulsars, the combinations' upper limits converge as we use more pulsars. However, these three combinations clearly show that there can be a large variance in the upper limits when we use fewer pulsars. One (particularly bad) realization using the three combinations discussed here is shown in Figure 1. Stopping with too few pulsars in either the single-pulsar upper limit or the greedy upper limit combination returns a value that is below the injected amplitude. In the greedy upper limit combination, this is especially pronounced: the upper limits calculated with between 2 and 19 pulsars yield a value below the injected amplitude.

3.2. All Combinations

Other than the combinations listed here, there are many other upper limit combinations that are possible. The factorized PTA likelihood allowed us to compute the upper limits for all possible combinations of 22 pulsars—a feat that would not otherwise be possible with current computers. In Figure 2, we show the distributions of these upper limits for all $\binom{22}{k}$

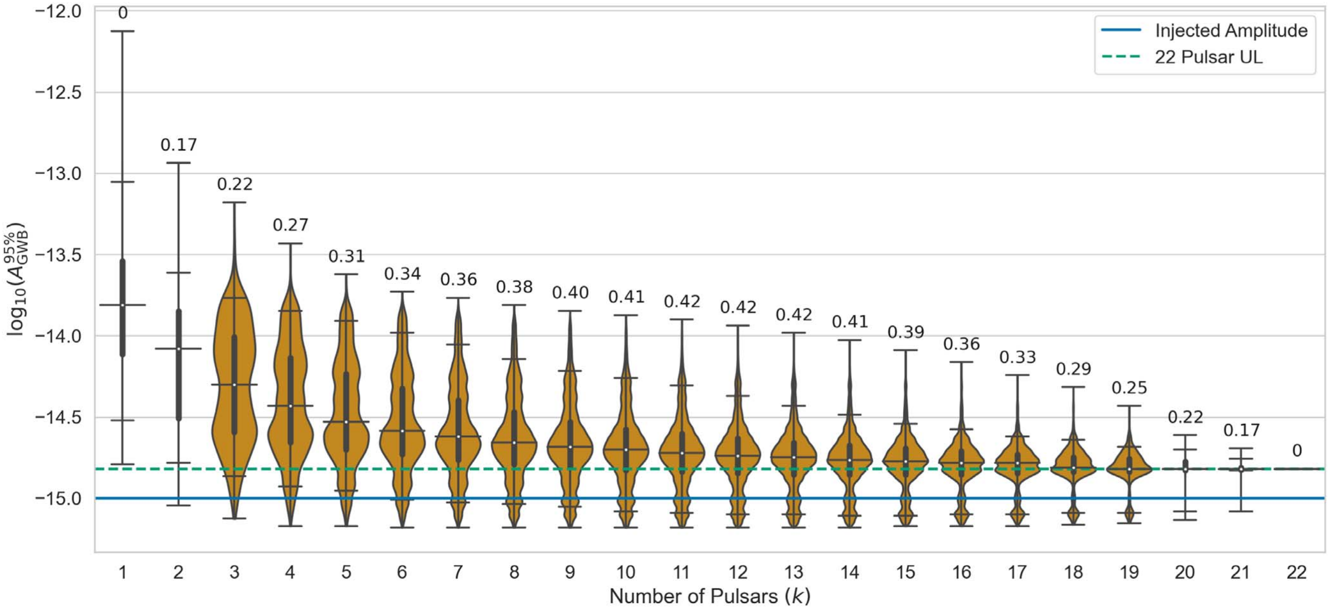

Figure 2. Violin plot showing the distribution of upper limits given by combinations of k pulsars for a single realization of the GWB. The number of pulsars k is given on the horizontal axis. The minimum, 5%, median, 95%, and maximum values are given by the horizontal lines on each violin. The bold bar around the white dot showing the median value gives the 25%–75% values. Violins have been removed for values of k that have fewer than 300 combinations. This realization has a clear multimodal structure, in which one of the modes goes below the injected amplitude for subsets consisting of 15 or more pulsars. The number above each violin shows the number of combinations below the injected amplitude (the solid blue line) divided by the number of combinations below the 22-pulsar upper limit (the green dashed line) when using k pulsars. When choosing pulsar combinations that give a lower upper limit than when using the full set, this gives the probability of randomly selecting a combination that results in an upper limit below the injected amplitude.

combinations with a given k on each violin. This realization is worrisome, because an entire mode of the multimodal structure ends up below the injected amplitude. The median upper limit decreases monotonically as we increase the number of pulsars, and it reaches its minimum value when using the full pulsar set. As we increase the number of pulsars used, the spread of the distribution of possible upper limits decreases, until we are left with only a single point when using the entire set of pulsars. The number above each violin in Figure 2 shows the number of combinations below the injected amplitude (the blue solid line) divided by the number of combinations below the 22-pulsar upper limit (the green dashed line) when using k pulsars. When choosing pulsar combinations that give a lower upper limit than when using the full set, this gives the probability of randomly selecting a combination that results in an upper limit below the injected amplitude. If we try to find an upper limit lower than when using the full data set in this realization, we risk ending up with an upper limit below the injected amplitude.

After investigating the combinations that end up below the injected amplitude, we find that there are some commonalities among these combinations. Pulsars J1640+2224 and J1909–3744 are used in nearly every combination, while J1713+0747 is left out until all the pulsars have been added. However, we also find that the pulsars that are included and left out is realization-dependent: another realization's combinations that fall below the injected amplitude do not necessarily include or exclude these particular pulsars.

Apart from being particularly influential on the upper limits (see Section 3.5), these pulsars behave the same as all the other pulsars in our data set. The recovered values of their intrinsic red noise amplitude and spectral indices are consistent with the injected values in every realization. Further, we used simulations that do not include any extra astrophysical effects that may be mismodeled. While removing J1640+2224 and J1909–3744 "fixes" this particular realization, in the sense that the amplitude upper limit is no longer below the injected value, seven other realizations remain with upper limits below the injected amplitude. Additionally, we cannot know beforehand that these pulsars cause problems without knowing the true value of the GWB amplitude. Including the rest of the pulsars in the data set similarly pulls the upper limit back up to reasonable values in all but two realizations (0.4%).

3.3. All Combinations of All Realizations

Figure 3 shows the distributions of the upper limits averaged across all 500 realizations of the GWB. The median upper limit again decreases monotonically as the number of pulsars increases and the range of the upper limits decreases. However, there are realizations that have upper limits that drop below the injected value, with as few as two pulsars and as many as 22 pulsars. When using the subset of pulsars that yields the lowest upper limit in all the realizations, we find the $\log_{10} A_{\mathrm{GWB}}$ upper limit below the injected value in 53 (10.6%) of 500 realizations.

Two realizations (0.4%) remain below the injected value even when using the entire pulsar set. In both cases, a single pulsar that strongly disfavors the GWB at and above its injected value dominates the upper limit, with a marginalized $\log_{10} A_{\mathrm{GWB}}$ posterior localized to values below the injected value. Upon multiplying this pulsar's $\log_{10} A_{\mathrm{GWB}}$ posterior by others, the other posteriors are forced to zero above the injected amplitude, resulting in the overall upper limit falling below the injected amplitude.

3.4. Single Combination of All Realizations

Following the previous sections, we consider a single sequence of pulsars: from the shortest observation time span to the longest. This allows us to look at the trends that exist

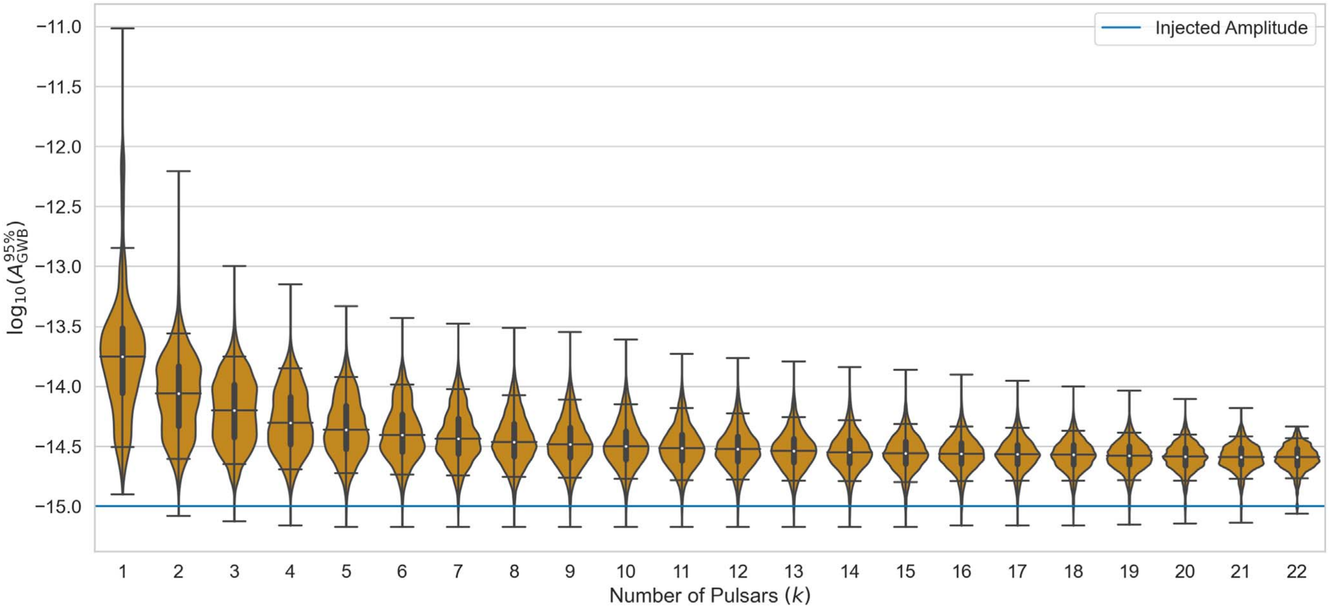

Figure 3. Violin plot showing the distribution of upper limits given by combinations of k pulsars averaged across 500 realizations. The number of pulsars k is given on the horizontal axis. The minimum, 5%, median, 95%, and maximum values are given by the horizontal lines on each violin. The bold vertical bars around the median value give the 25%–75% values. The $\log_{10}A_{\mathrm{GWB}}$ 95% upper limit falls below the injected amplitude value in two to 53 (0.4%–10.6%) of 500 realizations, depending on the subset of pulsars used.

across the realizations as a result of increasing the number of pulsars used in computing the cumulative upper limits. In Figure 4, we show the cumulative upper limits obtained when adding pulsars in this order. Each pulsar added decreases the upper limit, on average, until about 13 pulsars. At this point, the upper limit saturates in this specific combination. However, as shown in Figure 1, this saturation does not happen for some pulsar combinations.

As can be seen in Figure 4, the pulsars affect the upper limits differently between realizations. For example, assuming that the pulsars are added in the same order, as they are here, J1713+0747 may lower the upper limit in one realization, then increase the upper limit in the next. However, some pulsars influence the upper limits more than others.

3.5. Influential Pulsars

In order to work out which pulsars are most influential for the cumulative upper limits, we use the KL divergence,

$$D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right), \tag{6}$$

as a measure of the difference between P and Q, where we take $P(x)$ as the $\log_{10}A_{\mathrm{GWB}}$ posterior, computed with all 22 pulsars in the data set, and we take $Q(x)$ as the $\log_{10}A_{\mathrm{GWB}}$ posterior, computed while "dropping out" the pulsar whose influence we would like to check. Importantly, P remains the same for every pulsar that we are dropping out (although it will be slightly different between realizations). Figure 5 shows three columns for the pulsars that have $D_{\mathrm{KL}} > 0.5$ for any realization. In the left column, it shows the KL divergence, as described above. In the middle column, we plot the $\log_{10}A_{\mathrm{GWB}}$ posterior averaged over all the realizations for the full data set, the full data set without one pulsar, and the pulsar that was dropped. The D_{KL} value between the 22-pulsar mean posterior and the 21-pulsar mean posterior (both shown in the middle column) appears as a vertical line in the plots in the left column. In the right column, we have histograms of the upper limits for these same combinations for all 500 realizations.

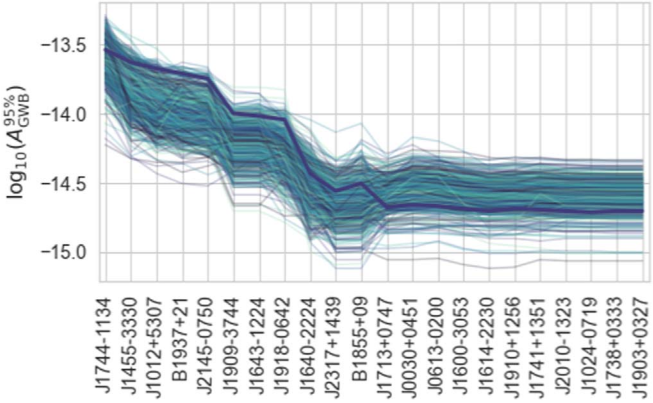

Figure 4. Cumulative upper limits for a sequence of pulsars sorted by observation time from shortest to longest. Each line corresponds to an individual realization (out of 500). The bold line represents one of these realizations. Most realizations in this sequence do not drop below the injected $\log_{10}A_{\mathrm{GWB}}$. The pulsars added in this combination have varied behavior between realizations: in some realizations, the pulsars increase the upper limit, while in others they decrease the upper limit.

As shown in Figure 5, J1640+2224 and J2317+1439 appear to slightly lower the mean posteriors and the upper limits once they are added to the set of pulsars used to compute the upper limits. J1713+0747 tends to increase the mean posterior and upper limits, while J1909–3744 does not affect the mean posterior or the upper limits significantly in either direction. For every pulsar on these plots, there are realizations where adding the pulsars does not significantly change the posterior, and we see this manifest as a $D_{\mathrm{KL}} \approx 0$. The pulsars not shown in these plots have smaller KL divergences between the full pulsar set and with one pulsar removed. This does not mean that we

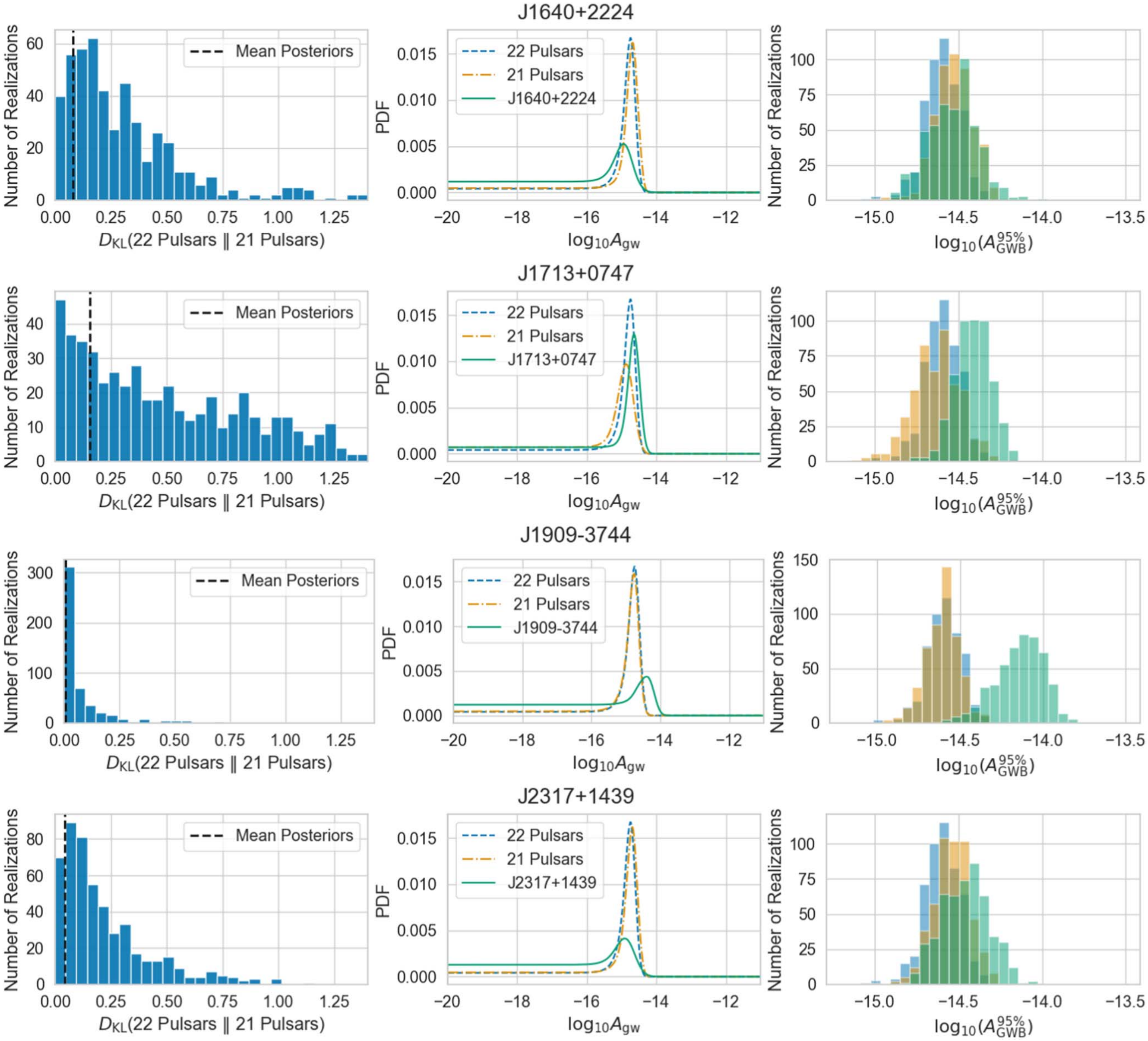

Figure 5. The plots in the left column contain the KL divergence computed using the $\log_{10}A_{\mathrm{GWB}}$ posterior, using all pulsars as the first argument, and the $\log_{10}A_{\mathrm{GWB}}$ posterior, using all but one pulsar as the second argument. The title of each subplot shows the pulsar that is removed from the second argument of the KL divergence. We cut out any pulsars that do not have $D_{\mathrm{KL}} > 0.5$. In the middle column, we plot the $\log_{10}A_{\mathrm{GWB}}$ posterior averaged over all realizations for the full 22-pulsar set, the 21-pulsar set, and the single pulsar that has been dropped. A vertical dashed line shows the D_{KL} between the mean posteriors of the 22 pulsars and 21 pulsars in the plots in the left column. In the right column, we plot the upper limits associated with each of the 500 realizations, with colors that correspond to the legend in the middle column. Each divergence computed shows that even though these pulsars are often influential on the upper limits, there are realizations where the posteriors with and without the pulsars are close.

should drop these pulsars when computing the upper limits, but that they show less influence on the overall upper limit once 21 pulsars have been included.

This analysis explains the results of the realization in Section 3.2. The two pulsars that were included in most of the combinations below the injected amplitude, J1640+2224 and J1909–3744, appear in this list of our most influential pulsars in these simulated data sets. Further, J1640+2224 tends to move the upper limit in a downward direction. J1713+0747, in contrast, tends to move the upper limit in an upward direction, and therefore it is left out until the last few pulsars.

3.6. Bias

In computing all possible combinations of the upper limits, we find that many realizations have combinations that result in upper limits below the injected amplitude value of the GWB. Out of 500 realizations, 53 (10.6%) realizations have at least one combination that gives a 95% upper limit below the injected value. This number drops to two (0.4%) realizations when using the full data set for every realization.

Figure 6 shows the upper limits obtained for all 500 realizations, where the upper limits have been found using three different possible methods. In one method, we computed the upper limit using all 22 pulsars (the green histogram on the right). In another method, we ranked the pulsars based on their

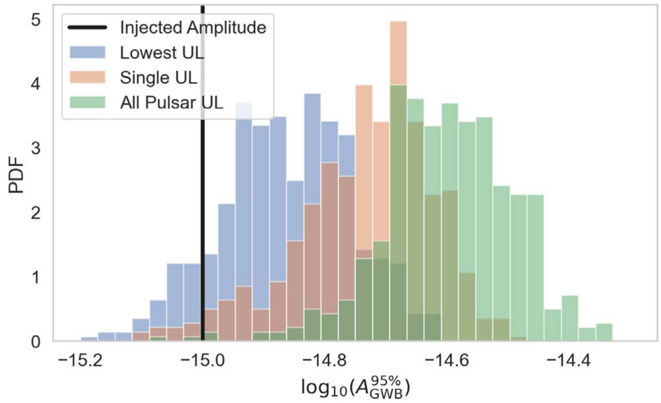

Figure 6. The upper limits obtained for all 500 realizations, where the upper limits have been found using three different methods. In one method, we computed the upper limit using all 22 pulsars (the green histogram on the right). In another method, we ranked the pulsars based on their single-pulsar upper limits on the GWB, and then combined those pulsars one by one until adding another pulsar caused the upper limit to increase. We took this minimum upper limit as the true upper limit (the orange histogram in the middle). In the final method, we looked at all possible combinations of pulsars and chose the lowest possible upper limit that could be obtained (the blue histogram on the left). The upper limits obtained using either the second or third method tend to be lower than those obtained using all 22 pulsars, resulting in a systematic shift toward lower amplitudes for these histograms (the blue on the left and the orange in the middle). The blue histogram on the left has 53 (10.6%) realizations below the injected amplitude (the vertical black line), the orange histogram in the middle has 12 (2.4%) realizations below the injected amplitude, and the green histogram on the right has two (0.4%) realizations below the injected amplitude.

single-pulsar upper limits on the GWB, and then combined those pulsars one by one until adding another pulsar caused the upper limit to increase. We took this minimum upper limit as the true upper limit (the orange histogram in the middle). In the final method, we looked at all possible combinations of pulsars and chose the lowest possible upper limit that could be obtained (the blue histogram on the left).

Note that when using the second or third method, we necessarily end up either using all 22 pulsars or using some subset of them, and the computed upper limits must either be equivalent to the upper limit obtained using all 22 or must be smaller, which is why the blue (left) and orange (middle) histograms are shifted to lower values relative to the green (right) histogram. As shown in Figure 6, the upper limits obtained using either the second or third method tend to be lower than those obtained using all 22 pulsars; furthermore, we find that 12 (2.4%) realizations of the orange (middle) histogram and 53 (10.6%) realizations of the blue (left) histogram out of 500 have an upper limit that falls below the injected amplitude. These results demonstrate that choosing a subset of pulsars in order to obtain the lowest possible upper limit results in a biased measurement.

4. Discussion and Conclusion

In this paper, we use simulated PTA data to study how the choice of which pulsars to include in GWB analyses can bias the upper limits on the GWB. By factorizing the PTA likelihood (Taylor et al. 2022), we are able to compute all possible combinations of upper limits for each realization of the GWB. This method limits us to an autocorrelation-only analysis. However, we find that including cross-correlations does not change the upper limits in the cases that the upper limits fall below the injected value. In every realization of 500, we find that the median and spread of the distribution of the upper limits decrease monotonically as the number of pulsars used increases. In some realizations, the probability of finding a value below the injected amplitude is significant when picking combinations that give upper limits below those returned by using all 22 pulsars. When using all pulsars to set the upper limit, we find that the upper limit is below the injected value in just two of the 500 realizations.

By investigating the sequences of upper limits resulting from different combinations of pulsars, we have shown that we can bias our upper limit to lower $\log_{10} A_{\rm GWB}$ values by using a subset of pulsars. In 53 (10.6%) of 500 realizations, the upper limit falls below the injected amplitude when choosing the minimum value in the lowest upper limit combination sequence. While this is the maximum bias to lower values of $\log_{10} A_{\rm GWB}$ that we can find, it is far from the only set of combinations that is biased toward lower values. Picking the lowest upper limit of any sequence of upper limits given by a combination of pulsars will always result in either the full set of pulsars being used or a distribution of upper limits from the 500 realizations shifted toward lower $\log_{10} A_{\rm GWB}$ values.

Multiple PTA experiments have recently published results reporting the detection of a common stochastic process whose amplitude is in tension with previously published upper limits on the amplitude of such a process. This work helps to explain one possible reason for this discrepancy. The earlier published work used significantly fewer pulsars compared to the number being used in the most recent papers, and, as shown in this paper, using a small number of pulsars to set the upper limits can lead to bias and can even result in upper limits that are lower than the true amplitude. The range of possible upper limits decreases as we increase the number of pulsars, and therefore using as many pulsars as we can reduces the probability that the upper limit that we obtain is below the actual value of the GWB. Furthermore, using more pulsars has the added benefit of producing a finer grid of angular separations of pulsar pairs, increasing our sensitivity to the cross-correlations that are characteristic of the GWB. In order to avoid bias and improve detection prospects, the best strategy for PTAs is to include as many MSPs as possible.

We thank Jeff Hazboun for helpful suggestions early on in this work. Joe Simon provided code to combine the posteriors from individual pulsar analyses into the posteriors from a joint analysis. We also thank Michele Vallisneri for valuable comments. Finally, we thank the anonymous reviewer, whose suggestions led to new interesting plots and discussion.

This work was supported by NSF Physics Frontier Center grant 1430284. ADJ and SJV were supported by UWM Discovery and Innovation grant 101X410. SJV was supported by NSF PHY-2011772. SRT acknowledges support from NSF AST-200793, PHY-2020265, and PHY-2146016. SRT also acknowledges support from Vanderbilt University College of Arts & Science Dean's Faculty Fellowship program.

ORCID iDs

Aaron D. Johnson https://orcid.org/0000-0002-7445-8423
Sarah J. Vigeland https://orcid.org/0000-0003-4700-9072
Xavier Siemens https://orcid.org/0000-0002-7778-2990
Stephen R. Taylor https://orcid.org/0000-0003-0264-1453

References

Anholm, M., Ballmer, S., Creighton, J. D., Price, L. R., & Siemens, X. 2009, PhRvD, 79, 084030

Antoniadis, J., Arzoumanian, Z., Babak, S., et al. 2022, MNRAS, 510, 4873

Arzoumanian, Z., Baker, P. T., Blumer, H., et al. 2020, ApJL, 905, L34

Arzoumanian, Z., Baker, P. T., Brazier, A., et al. 2018, ApJ, 859, 47

Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2016, ApJ, 821, 13

Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2018, ApJS, 235, 37

Chamberlin, S. J., Creighton, J. D., Siemens, X., et al. 2015, PhRvD, 91, 044048

Chen, S., Caballero, R. N., Guo, Y. J., et al. 2021, MNRAS, 508, 4970

Demorest, P. B., Ferdman, R. D., Gonzalez, M., et al. 2012, ApJ, 762, 94

Edwards, R. T., Hobbs, G. B., & Manchester, R. N. 2006, MNRAS, 372, 1549

Ellis, J. A., Vallisneri, M., Taylor, S. R., & Baker, P. T. 2020, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE v3.0.0, Zenodo, doi:10.5281/ZENODO.4059815

Gao, F., & Han, L. 2012, Comput. Optim. Appl., 51, 259

Goncharov, B., Shannon, R. M., Reardon, D. J., et al. 2021, ApJL, 917, L19

Hazboun, J. S., Simon, J., Siemens, X., & Romano, J. D. 2020, ApJL, 905, L6

Hellings, R. W., & Downs, G. S. 1983, ApJ, 265, L39

Hobbs, G. B., Edwards, R. T., & Manchester, R. N. 2006, MNRAS, 369, 655

Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79

Lentati, L., Shannon, R. M., Coles, W. A., et al. 2016, MNRAS, 458, 2161

Lentati, L., Taylor, S. R., Mingarelli, C. M. F., et al. 2015, MNRAS, 453, 2577

Phinney, E. S. 2001, arXiv:astro-ph/0108028

Pol, N. S., Taylor, S. R., Zoltan Kelley, L., et al. 2021, ApJL, 911, L34

Rosado, P. A., Sesana, A., & Gair, J. 2015, MNRAS, 451, 2417

Röver, C., Messenger, C., & Prix, R. 2011, arXiv:1103.2987

Shannon, R. M., Ravi, V., Lentati, L. T., et al. 2015, Sci, 349, 1522

Taylor, S. R. 2021, arXiv:2105.13270

Taylor, S. R., Simon, J., Schult, L., Pol, N., & Lamb, W. G. 2022, PhRvD, 105, 084049

Taylor, S. R., Vallisneri, M., Ellis, J. A., et al. 2016, ApJ, 819, L6

Vallisneri, M. 2020, libstempo: Python wrapper for Tempo2, Astrophysics Source Code Library, ascl:2002.017

Vallisneri, M., Taylor, S. R., Simon, J., et al. 2020, ApJ, 893, 112

Verbiest, J., Oslowski, S., & Burke-Spolaor, S. 2021, arXiv:2101.10081

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261