Expediting Astrophysical Discovery with Gravitational-Wave Transients Through Massively Parallel Nested Sampling

Rory J. E. Smith$^{1,2,}$ and Gregory Ashton$^{1,2,}$

$^1$School of Physics and Astronomy, Monash University, Vic 3800, Australia
$^2$OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Clayton VIC 3800, Australia

(Dated: September 27, 2019)

Understanding the properties of transient gravitational waves and their sources is of broad interest in physics and astronomy. Extracting information from gravitational waves involves comparing the signals and noise in experimental data to models of signals and noise. The most physically accurate models typically come with a large computational overhead which can render data analysis extremely time consuming, or possibly even prohibitive. In some cases highly specialized optimizations can mitigate these issues, though they can be difficult to implement, as well as to generalize to arbitrary models of the data. Here, we propose a general solution to the large run time of astrophysical inference using a parallelized nested sampling algorithm. The reduction in the run-time of inference scales almost linearly with the number of parallel processes running on a high-performance computing cluster. By utilizing a pool of several hundred CPUs in a high-performance cluster, the large analysis times of many astrophysical inferences can be alleviated while simultaneously ensuring that any gravitational-wave signal model can be used “out of the box”, i.e., without additional optimization or approximation. Our method will be useful to both the LIGO-Virgo-KAGRA collaborations and the wider scientific community performing astrophysical analyses on gravitational-wave transients.

Introduction—Gravitational-wave transients from merging binary black holes and binary neutron stars are now being routinely detected by the Advanced LIGO and Virgo detector network [1]. Such compact-binary systems offer unprecedented means to study strong-field gravity [2,3], matter at supra-nuclear densities [4], and stellar astrophysics [5]. With data in the public domain [7], methods to infer the properties of gravitational waves are of broad interest to various communities in physics and astronomy. An outstanding problem in astrophysical inference is how to expedite measurements of the properties of gravitational-waves and their sources, mitigating analysis times of months to years in some important cases [8,9]. In this Letter we show that wall-time of astrophysical inference can be significantly reduced using a highly flexible, massively-parallel nested sampling algorithm [10] deployed at scale on a high-performance CPU cluster. This work is not the first to use parallelism in inference. Algorithms such as RIFT [11] achieve rapid parameter estimation by precomputing aspects of the inference problem in parallel and approximating expensive functions using cheaper interpolation methods. Also, a class of approximations collectively known as “reduced order methods” [8,9,12,15] can facilitate low-latency parameter estimation. Both these methods utilize an “offline/online” decomposition in which expensive computations are first performed offline – possibly in parallel – in order to facilitate fast online analyses, using problem-specific approximations.

The benefit of our method can be summarized as follows: While nested sampling can efficiently make inferences on multi-dimensional astrophysical parameter spaces, the models that describe the data – and which facilitate inference – are expensive. Approximate methods can reduce the cost of models but frequently face the “curse of dimensionality”, i.e., they become exponentially more difficult to construct for realistic models of increasingly higher dimensionality [15]. Parallelism closes the loop on these two problems by bypassing the need for approximations whilst simultaneously reducing the wall-time cost of expensive analyses. The reduction in wall time scales almost-linearly with the number of CPUs in the cluster.

Our method affords a high degree of flexibility and scalability, and can facilitate inference at greatly reduced wall time when deployed on several hundred CPUs. Importantly, our method can be used to reduce the wall time of inference using any astrophysical gravitational-wave signal model without additional approximation or optimization. This represents a major advancement to techniques to mitigate the wall time of inference in gravitational-wave astronomy.

Inference—Astrophysical inference generally consists of two parts: Parameter estimation and hypothesis testing [17,21]. Parameter estimation entails computing the posterior probability density of the source parameters given the experimental data, e.g., the masses and spins of binary black holes. Hypothesis testing entails computing the Bayesian “evidence”: The probability of the data given an hypotheses. With the evidence, one can quantify the relative probability of the data under competing hypotheses, e.g., How much more probable is it that the data contain a signal described by General Relativity than by an alternative theory of gravity?.

Inferences made about the (astro)physics of gravitational-wave transients generally rely on (i) a model for the underlying gravitational-wave signal present in the data, (ii) a statistical description of the
detector-noise properties, and possibly \( (iii) \), a model for deterministic features of the noise, e.g., its power spectral density \([8, 21]\). Stochastic sampling algorithms—such as nested sampling \([19, 20, 22]\) or Markov Chain Monte Carlo \([19, 21]\)—are employed to search the parameter space of the models and estimate posterior densities and evidences.

Bayesian inference relates the probability of astrophysical parameters \( \theta \) to experimental data \( d \), and an hypothesis for the data \( H \), via Bayes theorem \([19]\):

\[
p(\theta|d, H) = \frac{p(d|\theta, H) L(d|\theta, H)}{\mathcal{Z}(d|H)},
\]

(1)

Here, \( p(d|\theta, H) \) is the posterior probability density of the astrophysical parameters \( \theta \) given \( d \) and \( H \); \( L \) is the likelihood of \( d \) given \( \theta \) and \( H \); \( p(\theta|H) \) is the prior probability of \( \theta \); and \( \mathcal{Z}(d|H) \) is the evidence of \( d \) given \( H \).

The posterior density is the target for parameter estimation, while the evidence is the target for hypothesis testing. Both the posterior density and evidence can be estimated to high accuracy using nested sampling \([19, 20]\). Assuming the priors can be defined, the primary input to inference algorithms is the likelihood function. We will consider a likelihood function which describes the probability of LIGO-Hanford, LIGO-Livingston and Virgo data given \( (i) \) an hypothesis that the data contain a signal plus Gaussian noise \( (H_S) \), and \( (ii) \) parameters \( \theta \) which describe the signal. This likelihood is the basis for most inferences in GW astronomy \([19, 20, 23]\), and constitutes the dominant cost of inference \([8, 21]\). The log likelihood is:

\[
\ln L(d|\theta, H_S) \propto \sum_{k=\{H, L, V\}} -\frac{1}{2} \langle d_k - h(\theta), d_k - h(\theta) \rangle,
\]

(2)

where we have used the usual “noise-weighted inner product”:

\[
\langle d_k, h(\theta) \rangle = 4 \Re f \sum_i \tilde{d}_k(f_i) \tilde{h}(\theta; f_i) S_n(f_i).
\]

(3)

Here \( \tilde{d}_k \) and \( \tilde{h} \) are the Fourier transforms of the time-domain data and signal, \( S_n \) is the noise power spectral density. It is important to stress that the choice of signal model \( h(\theta) \) defines a particular signal hypothesis, and in practice many different signal models are often used to analyze a given data set \([2]\). The particular choice of signal model is generally determined by the requirements of astrophysical data analysis. The framework which we have outlined is flexible enough to accommodate a large variety of astrophysical signal models, and can be extended to accommodate models for the noise \([24]\). For a given model, the CPU-time of inference scales with the signal’s bandwidth multiplied by its duration \([15]\). The overall cost is set by the intrinsic complexity of the model \([8, 9]\). For signal models defined in the time domain, \( h \) is computed by first evaluating the model in the time domain and subsequently taking the discrete Fourier transform. It is important to note that time-domain signal models can be significantly more computationally expensive than those defined directly in the frequency domain. Many time-domain models require solving costly coupled ODEs to evaluate the signal at discrete times, e.g., \([25]\).

Together with the additional cost of the Fourier transform, the relative cost of using time domain models in inference can be between one to two orders of magnitude more expensive than frequency-domain models \([8, 9]\).

Signal models—Many of the most accurate models of gravitational-wave signals are computationally expensive and constitute the dominant computational cost of the likelihood, and hence the entire inference process. Used “out of the box” in serial sampling algorithms, the wall time of analyses can be between several days to several months, or even years \([8, 9]\). As such, models are employed based on trade offs between accuracy and speed, often in conjunction with highly tailored approximations or optimizations \([8, 9, 11, 12, 21]\).

For analyses of binary black hole systems we consider two waveform models: a full-precession model (SEOBNRv3) \([25, 27]\) which includes generic two-spin inspiral precession dynamics, and a numerical relativity “surrogate model” \((NRSur7dq2)\) \([13]\). NRSur7dq2 is built from the results of around 700 numerical relativity simulations, covers spin magnitudes up to 0.8 and mass ratios up to 2, and includes all \( l \leq 4 \) modes. For binary neutron stars, we use an effective-precession model (IMRPhenomPv2NRT) \([25, 32]\) which includes the effect of precessing spin on the heavier of the two bodies, and models the tidal deformability of neutron stars through two tidal deformability parameters. These models are employed routinely in LIGO/Virgo analyses, and numerous other studies \([2]\).

Our choice of models is motivated by three considerations: \( (i) \) accuracy; \( (ii) \) flexibility; and \( (iii) \) extensibility/augmentibility of the models. Both SEOBNRv3 and IMRPhenomPvNRT have been used in published analyses on events which appeared in the first gravitational-wave transient catalogue \((GWTC-1)\) \([2]\), and can model a wide range of binary black hole and binary neutron star systems observable by LIGO/Virgo \([25, 27]\). NRSur7dq2 has been used in a variety of studies, such as when the effects of higher-order modes become important, see, e.g., \([33, 35]\), and to model the effects of gravitational-wave memory, e.g., \([39]\). Furthermore, the models are built on a underlying framework which is extendable, meaning that future signal models may share many of the features of the current generation of these models. Thus, any performance gains that can be achieved on this generation of models should also be achievable with future generations of waveform models.
Parallel Nested Sampling—Nested sampling is a stochastic-sampling method designed primarily to estimate the evidence $Z(d|\mathcal{H})$ [22] which appears in the denominator of Eq. [1], and which is the primary ingredient in Bayesian hypothesis testing. As a byproduct, nested sampling also produces the posterior density $p(\theta|d, \mathcal{H})$. Importantly, nested sampling is scalable to high-dimensional and irregularly shaped parameter spaces [10, 20]. This affords a large degree of flexibility and ensures that nested sampling is well suited to extensions of the likelihood function in Eq. (2) e.g., by increasing the dimensionality of the parameter space to include parameters that model features of the noise, or parameters that describe signals in alternative theories of gravity.

Nested sampling computes the evidence by estimating the “marginalized likelihood function”:

$$Z(d|\mathcal{H}) = \int_\theta d\theta \pi(\theta) \mathcal{L}(d|\theta, \mathcal{H})$$

$$= \int_{X=0}^1 dX \mathcal{L}(d|X, \mathcal{H}).$$

(4)

(5)

The second line transforms the integral over the multidimensional parameter space $\theta$ into a one-dimensional integral over the prior mass $dX = d\theta \pi(\theta)$.

Qualitatively, the integral is estimated by drawing successive “live points” from a bounded prior – together with a sample-acceptance/rejection rule – to estimate the change in prior mass $dX_i$ and evidence $Z_i(d|\mathcal{H})$ at each iteration $i$. The boundaries of the prior are updated after a user-specified number of iterations [10]. The algorithm terminates upon a stopping condition, e.g., when the change in the evidence between successive iterations $\Delta Z$ falls below a user-defined threshold. Assuming that sampling from the prior incurs a negligible cost, the dominant cost of the algorithm comes from evaluating the likelihood function in Eq. (2). Throughout, we will use an implementation of nested sampling from the 	exttt{dynesty} package [10], which is freely available, easy to use, and easily parallelizable. We note that other parallel nested sampling tools exist, e.g., 	exttt{cpnest} [40], though we have found that 	exttt{dynesty} is more straightforward to deploy at scale on a high-performance computer cluster.

We can speedup the overall run time by parallelizing the sampling step. The key observation is to note that because the prior is by definition a priori known, we can draw an arbitrary number of prior samples in parallel at each iteration. Thus, by utilizing a pool of $n_{\text{cores}}$ CPUs in a cluster to draw $n_{\text{live}}$ samples in parallel within the current prior bound. The scaling relation for the speedup

$$S(n_{\text{cores}}, n_{\text{live}}) = n_{\text{live}} \ln(1 + n_{\text{cores}}/n_{\text{live}})$$

(6)

The overall decrease in the run time will, in practice, not scale exactly because of factors such as intra-CPU communication and other overheads. However, as we show in the following sections, the reduction in wall-time is consistent with the above scaling relation. Using 	exttt{dynesty}, the parallelization can be accomplished using the Message Passing Interface (MPI) [42] pool which facilitates efficient communication between CPUs in a cluster. We use an implementation of the gravitational-wave likelihood function and priors based on the gravitational-wave analysis library 	exttt{bilby} [20], and modified to facilitate efficient MPI communication [43].

We note that there are other aspects of the inference framework which could be parallelized in some cases. Graphical processor units (GPUs) have been demonstrated to improve the efficiency of inference by an order of magnitude by parallelizing the computation of signal models which admit a closed form expression [41], i.e., where all components of the time/frequency series can be evaluated simultaneously. As we have noted, this is not applicable to models such as SEOBNRv3, and so we do not consider GPU parallelization here.

Example wall times of end-to-end analyses—In Table I we provide estimates of realistic speedups using parallel nested sampling as a function of the number of CPUs $n_{\text{cores}} = (1, 415, 830)$. In all examples, we use $n_{\text{live}} = 830$ live points$^1$. The results for $n_{\text{cores}} = (415, 830)$ are measured, while the results for $n_{\text{cores}} = 1$ are estimated using the scaling relation Eq. (6). The maximum of $n_{\text{cores}} = 830$ is set by the total available cores on the cluster SSTAR$^2$. $^{45}$ used for the analysis. We consider four archetypal cases based off of observed signals described in GWTC-1 [2]: (i) a heavy binary black hole system, in this case with chirp mass $M_c \approx 25 M_\odot$, (ii) a “moderate mass” binary black hole system, with chirp mass similar to GW151012 ($M_c \approx 15 M_\odot$), (iii) a “low mass” binary black hole system with chirp mass similar to GW151226 ($M_c \approx 9 M_\odot$), and (iv) a binary neutron star system with chirp mass similar to GW170817 ($M_c \approx 1.2 M_\odot$). These four cases represent analyses of increasing wall time$^3$ and therefore represent typical analyses of increasing computational complexity. For each analysis, we “inject” a simulated signal into simulated LIGO and Virgo detector noise. We assume design-sensitivity Advanced

---

$^1$ We match the number of cores to a multiple of the live points, as well as ensuring that the prior bounds are updated every $2n_{\text{live}}$ iterations. This ensures that no samples are discarded when the bounds are updated.

$^2$ All of the runs were performed on Intel Xeon E5-2660 (Sandybridge) CPUs with a 2.2GHz clock rate. Nodes are networked via non-blocking QDR infiniband.

$^3$ From a given reference frequency, the signal duration (to leading order) is proportional to $M_c^{-5/3}$. 

3
LIGO/Virgo noise power spectral densities [46]. For binary black hole analyses, we sample over a fourteen dimensional space of astrophysical parameters characterizing sources in General Relativity: two masses, six spin components, luminosity distance, inclination, right ascension, declination, and polarization phase, and phase at coalescence. We numerically marginalize over the “nuisance parameter” time at coalescence [47]. For the binary neutron star analysis, we sample over the same fourteen parameters, plus two parameters which quantify the stars’ tidal deformability [32].

The binary black hole analyses use a low(high) frequency of 20Hz(1024Hz). The simulated signals have signal-to-noise ratios \((\rho_{150914}, \rho_{151012}, \rho_{151226}) = (15,12,12)\). Our analysis on the GW170817-like system uses a low(high) frequency of 15Hz(2048Hz) and uses data with duration \(T = 512s\) (c.f. \(T = 128s\) from 20Hz [28]). The simulated signal has signal-to-noise ratio \(\rho_{170817} = 50\). While Advanced LIGO has little sensitivity below 20Hz at design sensitivity, it is nonetheless interesting to demonstrate that such analyses can be performed in principle. Such an analysis could be useful for very loud signals which may have appreciable signal-to-noise ratio between 15Hz to 20Hz.

We find wall times consistent with the speedup scaling relation given by Eq. (4). Importantly, we find that for inference on GW170817-like binary neutron stars using IMRPhenomPv2NRT, analysis times can be reduced to around 2.5 days using 830 CPUs (c.f., 3.8 years using 1 CPU). For the lowest-mass binary black hole system (GW151226-like), we find that analyses can be performed in around half a day using 830 CPUs with SEOBNRv3. For the heaviest binary black hole system in our table (GW170809-like) we find that with 830 CPUs, analyses can be performed in around 1.6 hours.

Implications — We have focused on “vanilla” inference problems where the only free parameters are those of signals described by (approximations to) General Relativity. Nested sampling methods have been shown to be robust for estimating evidences and posteriors in model spaces that have many tens to hundreds of parameters [10, 39]. Thus, our results demonstrate that provided the inference problem is dominated by the cost of the likelihood function, then parallelized nested sampling will offer comparable speedups for inferences in which the models and model spaces are significantly larger and more complex than those which we have considered. For example, it is increasingly common for analyses to estimate not just signal parameters but also those of the noise model [24]. Additionally, signal models in alternative theories of gravity are often parameterized by many more than the 15-17 parameters which describe binary black hole and binary neutron star signals in General Relativity [50]. Thus, our method is highly extensible to a wide class of important data analysis problems.

Parallelized nested sampling may also serve as a useful tool in tackling inference on signals as seen by third-generation detectors, e.g., Einstein Telescope and Cosmic Explorer. The analysis challenge for these instruments will be significantly more complex: many signals will be in-band simultaneously, and signals may be in band for up to several tens of minutes [51, 53]. Thus methods which can alleviate aspects of the wall-time of inference will be valuable as the demands and complexity of data analysis increase.

Comparison to alternative methods — Several alternative strategies exist to attempt to mitigate the cost of inference. As we will argue, our method is more flexible that currently-existing techniques, and can be applied more broadly. A class of techniques known collectively as “reduced order methods” have been successful at reducing the cost of inference using certain signal models by up to a factor of around 300, and are employed by the LIGO Scientific Collaboration in production-level analyses [8, 9, 12]. However, they are difficult to apply broadly to all classes of signal models, in particular to fully-precessing time-domain signal models due to the curse of dimensionality, see, e.g., [13, 14, 54]. Moreover, they typically require highly specialized knowledge to construct and as such may not be readily utilized by the larger physics and astronomy communities. Other methods offer various degrees of parallelism: Monte Carlo methods [19] have been employed to facilitate using expensive models such as SEOBNRv3 [2]. However, the efficiency of the algorithm is typically poor and the model is expensive. These two issues compound in such a way as to make the wall time and CPU time are large. Moreover, MCMC techniques are far less accurate at estimating evidences [50]. Significantly, the convergence of nested sampling scales relatively weakly with the number of dimensions of the model space [55], making it an ideal candidate for computing evidences and posteriors for models with many tens to hundreds of parameters [10, 39] such as would be the case when noise models are included in the likelihood function [24] or when signal models describe physics from alternative theories of gravity [2]. This is in contrast to parallel “grid-based” methods [11, 21] in which the complexity scales as a power of the dimensionality of the model which rapidly becomes prohibitive as the space increases significantly.

An embarrassingly parallel method known as likelihood reweighting [37, 37] aims to obtain posterior samples and evidences from a target likelihood function, i.e., the one of interest, by leveraging samples and evidences obtained using a reference likelihood function that is computationally cheaper to evaluate [37].

4 For example, the analyses on the GWTC-1 event GW170608 used 120 parallel MCMC chains running continuously for around two months [39].
Reweighting is extremely efficient when the target and reference posteriors are similar. For instance, analyses using higher-order mode likelihoods on the GWTC-1 events can generate posterior samples with between 7% – 60% efficiency [37]. However, there are two drawbacks to reweighting. First, when the target and reference likelihoods differ significantly, the overall efficiency can be poor. For very loud events, e.g., SNR $\sim 50$, as seen by aLIGO/Virgo the efficiency can be around 0.1% [37], meaning that many thousands of reference analyses have to be performed in order to generate a satisfactory number of effective samples through reweighting. In practice this can lead to a very high overall CPU time, though due to the embarrassingly parallel nature of the problem, a low wall time. Secondly, the choice of a good reference likelihood is not always obvious. Here, a trade off between accuracy and speed has to be made, and several of the fastest waveform models have restrictions in, e.g., mass ratio and spin [8], which could make the target and reference likelihoods diverge in regions of parameter space, thus introducing a potential source of inefficiency in the reweighting procedure.

**Acknowledgements** — This work is supported through Australian Research Council (ARC) Centre of Excellence CE170100004. The analyses presented in this paper were performed using the supercomputer cluster at the Swinburne University of Technology (SSTAR). This document has LIGO Document number P1900255-v1. We are grateful to SSTAR system admins for their support with all things MPI and for their patience. We would like to thank Eve Chase for providing information about the GWTC-1 analyses performed using SEOBNRv3. We are grateful for insightful comments from Vivien Raymond, Eve Chase, Richard O'Shaughnessy, Moritz Hubner, and Michele Vallisneri. Additional thanks to Joshua Speagle for pointing out the scaling relation for parallel nested sampling.

**Outlook** — Parallelized nested sampling, deployed on a high-performance CPU cluster, reduces the wall-time of inference almost linearly with the number of parallel CPUs in the cluster. Because it does not approximate either gravitational-wave signal models, or the statistical properties of the data, it is highly flexible, extendable, easy to implement, and can be used in a broad variety of analyses. While the method is computationally expensive, it nonetheless affords greatly expedited inferences on gravitational waves provided one has access to a high-performance computer cluster. Given the availability of clusters in large collaborations and research institutes, together with cloud-computing resources, parallelized nested sampling should be a useful tool to both the LIGO Scientific Collaboration as well as to independent research groups. By expediting inference, we enable astrophysical discoveries to be made on much shorter time scales, in some cases from years to hours. Faster algorithms allow deeper exploration into the data, enabling the use of more complex models of the data which otherwise may not be practical to use.

### Table I: Example wall times for analyses using SEOBNRv3, IMRPhenomPv2NRT and NRsurr7d2q using $n_{\text{cores}} = (1, 415, 830)$ CPUs. The results for $n_{\text{cores}} = 1$ are measured, whereas the results for $n_{\text{cores}} = 1$ are estimated using the scaling relation Eq. (6). Each of the four cases correspond to analyses on simulated signals injected into Advanced LIGO and Advanced Virgo-like noise, and are chosen to reflect "typical" LIGO/Virgo sources from GWTC-1 [2]: three binary black holes (GW170809-like, GW151012-like, GW151226-like), and a binary neutron star (GW170817-like).

| Number of CPUs | SEOBNRv3 | IMRPhenomPv2NRT | NRsurr7d2q |
|----------------|----------|-----------------|------------|
| 1              | 38.3 d   | 1.6 hr          | 26.4 d     |
| 415            | 2.7 hr   | 1.1 d           | 1.5 hr     |
| 830            | 1.6 hr   | 15.5 hr         | 1.1 hr     |
| 2              | 2.4 d    | 27.9 hr         | 4.5 d      |
| 415            | 2.4 d    | 57.5 d          | 3.6 hr     |
| 830            | 2.4 d    | 2.6 hr          | 2.4 d      |
| 3.8 yr         | 2.7 hr   | 1.1 d           | 3.6 hr     |
| 4.5 d          | 2.6 hr   | 3.3 hr          | 2.4 d      |
| 2.4 d          | 2.6 hr   | 62.3 d          | 3.6 hr     |
| 62.3 d         | 3.6 hr   | 3.6 hr          | 2.4 d      |

[1] GraceDB – Gravitational-Wave Candidate Event Database (2019), URL https://gracedb.ligo.org/superevents/public/05/.
[2] B. P. Abbott et al. (LIGO Scientific, Virgo) (2019), 1903.04467.
[3] B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. 123, 011102 (2019), URL https://link.aps.org/doi/10.1103/PhysRevLett.123.011102.
[4] B. P. Abbott et al. (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. 116, 221101 (2016), URL https://link.aps.org/doi/10.1103/PhysRevLett.116.221101.
[5] B. P. Abbott et al. (The LIGO Scientific Collaboration and the Virgo Collaboration), Phys. Rev. Lett. 121, 161101 (2018), URL https://link.aps.org/doi/10.1103/PhysRevLett.121.161101.
[6] B. P. Abbott et al. (LIGO Scientific, Virgo) (2018), 1811.12940.
[7] M. Purrer, Class. Quant. Grav. 31, 195010 (2014).
