Accelerated Bayesian parameter estimation and model selection for gravitational waves with normalizing flows

Alicja Polanska
Mullard Space Science Laboratory
University College London
alicja.polanska.22@ucl.ac.uk

Thibeau Wouters
Utrecht University
Nikhef
t.r.i.wouters@uu.nl

Peter T. H. Pang
Nikhef
Utrecht University
thopang@nikhef.nl

Kaze K. W. Wong
Center for Computational Astrophysics
Flatiron Institute
Department of Applied Mathematics and Statistics
Johns Hopkins University
wwong24@jhu.edu

Jason D. McEwen
Mullard Space Science Laboratory
University College London
Alan Turing Institute
jason.mcewen@ucl.ac.uk

Abstract

We present an accelerated pipeline, based on high-performance computing techniques and normalizing flows, for joint Bayesian parameter estimation and model selection and demonstrate its efficiency in gravitational wave astrophysics. We integrate the JIM inference toolkit, a normalizing flow-enhanced Markov chain Monte Carlo (MCMC) sampler, with the learned harmonic mean estimator. Our Bayesian evidence estimates run on 1 GPU are consistent with traditional nested sampling techniques run on 16 CPU cores, while reducing the computation time by factors of $5\times$ and $15\times$ for 4-dimensional and 11-dimensional gravitational wave inference problems, respectively. Our code is available in well-tested and thoroughly documented open-source packages, ensuring accessibility and reproducibility for the wider research community.

1 Introduction

In many scientific fields Bayesian inference is an indispensable tool for extracting new knowledge from observations, providing a principled statistical framework for parameter estimation and model selection. In Bayesian statistics, the posterior probability distribution $p(\theta \,|\, d, M)$ encodes information about the parameters θ, of model M, given the observed data d. By Bayes' theorem the posterior is given by

$$p(\theta \,|\, d, M) = \frac{p(d \,|\, \theta, M) p(\theta \,|\, M)}{p(d \,|\, M)}, \tag{1}$$

where $p(d \,|\, \theta, M)$ is the likelihood, $p(\theta \,|\, M)$ the prior and $p(d \,|\, M) \equiv z$ the Bayesian evidence. The likelihood quantifies how well a model and a given set of parameters θ describe the data. The prior

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

reflects our existing beliefs about the parameters. Typically, Markov chain Monte Carlo (MCMC) sampling techniques are used to explore the posterior distribution for parameter estimation, from which parameter estimates and their uncertainties can be computed. The Bayesian evidence z, also called the marginal likelihood, is a normalization factor and is computed as

$$z = p(d \mid M) = \int d\theta \, p(d \mid \theta, M) p(\theta \mid M). \quad (2)$$

The evidence is a crucial quantity for comparing competing models, allowing us to provide a statistically principled preference for one model over another [1]; although, it is computationally difficult to calculate.

One particular scientific field that heavily relies on Bayesian inference is gravitational wave (GW) astrophysics. Since 2015, Advanced LIGO [2] and Advanced Virgo [3] have detected $\mathcal{O}(100)$ GW signals originating from mergers of black holes and neutron stars [4–7], revealing a novel and reliable way of observing and studying the Universe. Evidence estimates have allowed, for instance, one to identify the nature of sources of GWs [8] and discern between various waveform models encoding different underlying physics [9], thereby advancing our understanding of GW sources.

Nested sampling algorithms [10, 11] are widely used to compute the Bayesian evidence. However, this method of estimation is tightly coupled to the sampling strategy, inhibiting the adoption of accelerated sampling techniques. The sampling must be performed in a nested manner, which means these methods can be computationally expensive, especially for high-dimensional parameter spaces and multimodal posteriors. A fast and scalable alternative is therefore of paramount importance for various scientific disciplines. In GW astrophysics, for instance, telescope operators require information regarding the nature of the source in low latency to identify potential electromagnetic counterparts of GW events. Moreover, next-generation GW detectors, such as the Einstein Telescope [12] and the Cosmic Explorer [13], will have increased sensitivities which result in longer signal durations and more events to analyze, enhancing the demand for efficient inference methods [14]. Previous attempts have accelerated nested sampling algorithms using machine learning [15, 16] or make use of simulation-based inference [17–25].

Recently, the JIM[1] inference toolkit [26] was introduced, which accelerates parameter estimation by using normalizing flow-enhanced MCMC sampling as well as hardware accelerators such as graphical processing units (GPUs) and tensor processing units (TPUs). However, MCMC methods like JIM do not provide the Bayesian evidence, which is necessary for Bayesian model selection. In this work, we augment JIM with a scalable evidence estimator decoupled from the sampling method – the learned harmonic mean estimator with normalizing flows [27, 28], implemented in the `harmonic` Python package[2]. Since, unlike nested sampling, the learned harmonic mean is agnostic to the sampling strategy, it is possible to realise the acceleration provided by JIM and still perform accurate evidence estimation. Other methods of evidence estimation decoupled from the sampling strategy have been recently proposed [29–31], but we choose the learned harmonic mean due to several advantages discussed in depth in Refs. [28, 32]. We demonstrate, using an example from the field of GW astrophysics, that our pipeline provides accurate evidence estimates while only requiring a fraction of the computational cost required by the traditional methods.

2 Methodology

We construct an accelerated pipeline to, first, sample the posterior distribution and, second, compute the Bayesian evidence. We leverage normalizing flows, at both the sampling and evidence estimation stages. Moreover, we use a sampler that leverages the high-performance computing techniques of JAX [33].

2.1 Normalizing flows

Normalizing flows are generative models that transform a simple base distribution into a complex one through a series of invertible, differentiable mappings with learned parameters. The flow can be trained on samples from the distribution of interest by minimising the forward Kullback-Leibler (KL)

[1] `https://github.com/kazewong/jim`

[2] `https://github.com/astro-informatics/harmonic`

divergence. For a more extensive review of normalizing flows we refer the reader to references [34, 35]. Both JIM and harmonic use rational-quadratic spline flows [36], where piecewise rational-quadratic functions are used in the transformations. They are able to encode nonlinear and local relationships, allowing for a more expressive and powerful architecture than affine transformations [37].

2.2 JIM inference toolkit

In Ref. [26], the authors introduced JIM, an inference toolkit implemented in JAX [33], and applied it to GW astrophysics as an example. JIM supports GPU-accelerated differentiable gravitational waveform models [38] and can therefore make use of efficient gradient-based samplers such as the Metropolis-adjusted Langevin algorithm [39] or Hamiltonian Monte Carlo [40]. In order to further accelerate the parameter estimation, JIM makes use of FLOWMC[3] , a normalizing flow-enhanced MCMC sampler implemented in JAX [41, 42]. It accelerates traditional MCMC by adapting a global proposal density distribution to the target distribution with normalizing flow on the fly. It has been shown that JIM can accurately infer the parameters of GW signals originating from merging black holes [26] and neutron stars [43]. Moreover, JIM achieves this at a fraction of the computational cost of conventional methods that nested sampling, e.g. BILBY [44–46]. However, previously JIM could not provide evidence estimates for model comparison.

2.3 Learned harmonic mean estimator

To compute the Bayesian evidence from posterior samples, we consider the recently proposed learned harmonic mean estimator [27, 47, 28, 32]. The learned harmonic mean is a scalable estimator of the evidence based on posterior samples, which is therefore agnostic to the sampler used and can be integrated with the FLOWMC sampler used in JIM. While the original harmonic mean estimator [48] suffered from instability [49], the learned harmonic mean solves this issue by leveraging machine learning techniques [27]. The reciprocal evidence $\rho = z^{-1}$ is estimated as

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{p(d \,|\, \theta_i, M) p(\theta_i \,|\, M)}, \quad \theta_i \sim p(\theta | d, M), \tag{3}$$

where N is the number of samples and $\varphi(\theta)$ is a learned normalized target distribution that must be concentrated within the posterior. Recently, the authors of Ref. [28] integrated normalizing flows into the learned harmonic mean estimator, which provide a robust approach to ensure the learned target distribution is indeed concentrated within the posterior. Specifically, a temperature parameter is introduced to scale the variance of the base distribution of the flow by a factor $0 < T < 1$. The concentrated flow is then used as the target $\varphi(\theta)$. The authors show that the estimates are robust to different values of T. The method is implemented in the harmonic package written in JAX.

The posterior distributions encountered in GW physics are often multimodal, which can prove challenging. In particular, due to the topology-preserving nature of their transformations, normalizing flows tend to struggle with multimodality [50], and are prone to mode-covering behaviour when trained using forward KL divergence [51]. These problems are mitigated by the fact that for the learned harmonic mean estimator to be accurate, it is not necessary to achieve a very close approximation of the posterior [27]. However, a poor approximation can potentially be problematic if it leads to regions of high flow density in regions where posterior density is low. To facilitate the learning of the multimodal distributions, we consider a multimodal base distribution (a sum of normal distributions with an identity covariance matrix), which leads to an improvement in harmonic diagnostics. In future work we plan to investigate in detail our method's robustness to multimodality, including for GW events with a low signal to noise ratio, by numerically studying the influence of the base distribution choice, as well as considering other flow approaches designed to deal with multimodality [e.g. 52, 53, 50, 54].

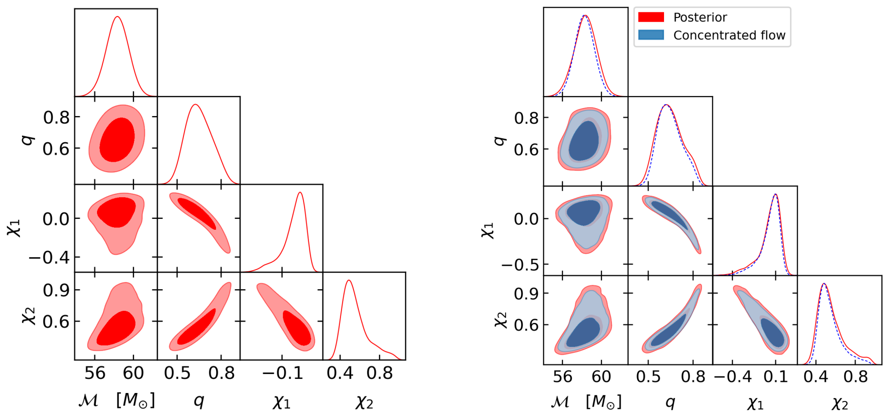
3 Results

We validate and benchmark our pipeline, combining JIM with harmonic, by computing the Bayesian evidence of a simulated GW signal from a binary black hole merger as an example. The signal is

[3] https://github.com/kazewong/flowMC

Table 1: Total wall times to compute the evidence estimates for the examples discussed in the main text. We run BILBY on 16 CPU cores and JIM + `harmonic` on 1 GPU.

Example	Method	$\log(z)$	Sampling time	Evidence estimation time
4D	BILBY	390.33 ± 0.11	31.3 min	–
	JIM + `harmonic`	$390.360^{+0.006}_{-0.006}$	3.4 min	1.9 min
11D	BILBY	378.29 ± 0.15	3.5 h	–
	JIM + `harmonic`	$378.420^{+0.09}_{-0.08}$	11.8 min	2.4 min



(a) BILBY (b) JIM + `harmonic`

Figure 1: Corner plots for the 4-dimensional posterior samples from (a) BILBY and (b) JIM used for inference (solid red) alongside the concentrated flow at $T = 0.8$ used in the learned harmonic mean (dashed blue).

injected into a realization of a Gaussian noise time series coming from a network of the two Advanced LIGO [2] and the Advanced Virgo detector [3] at their design sensitivities. The recovery of the injected signal is performed in two examples. In the first example, we only recover the 4-dimensional intrinsic parameter space, comprising of the masses and the aligned spins of the systems, while fixing other parameters to their injected values. In the second example, we recover the full 11-dimensional parameter space, which includes the extrinsic parameters. The setup of these inferences in described in detail in Appendix A.

For both examples, we compute the Bayesian evidence with nested sampling, and with JIM combined with `harmonic`. For the former, we use BILBY [44], employing DYNESTY [55] as the nested sampling library. More specifically, we use PARALLEL-BILBY [46] and parallelize the computation with 1000 live points over 16 cores on a single Intel Xeon Silver 4310 Processor central processing unit (CPU). For JIM and `harmonic`, we use a single NVIDIA A100-40GB GPU to perform the inference.

When estimating the evidence with `harmonic` from the JIM posterior samples, we divide the samples into two sets with an equal number of chains, using one to train the flow and the other to estimate the evidence. The samples are thinned, keeping only every tenth sample in each chain, which achieved an accurate estimate at reduced time and memory demands.

As an additional check we also estimate the evidence on posterior samples obtained from nested samples via rejection sampling with BILBY. We randomly shuffle these posterior samples, as the output contains more samples from higher density regions towards the end, introducing bias into the train–inference split. Then we divide them into 20 chains before using `harmonic`. We obtain results consistent with nested sampling evidence estimates.

The evidence estimates for the 4-dimensional example along with their computation times are shown in the first two rows of Table 1. We use a rational-quadratic spline flow with 6 layers and 8 spline bins and a unimodal base in the learned harmonic mean estimator, and set the temperature parameter to $T = 0.8$. We find that the evidence estimates obtained using BILBY are in close agreement with our estimates obtained from JIM samples with `harmonic`. However, our pipeline achieves a speedup factor of $5.4\times$ relative to BILBY at performing this calculation even for this relatively low dimensional example. Figure 1 shows corner plots of BILBY samples as as well as the half of JIM samples used for inference, alongside the concentrated flow of `harmonic`. The plot shows visual agreement between JIM and BILBY, and demonstrates that the flow of `harmonic` is concentrated in the posterior of JIM. We perform additional sanity checks, described in [27] to further validate the results of the learned harmonic mean. In particular we inspect the estimates of error, kurtosis and the ratio between the square root variance of variance and variance estimates.

We repeat the same procedure for the 11-dimensional example. We use the same thinning procedure and employ 5 layers with 64 bins for the rational-quadratic spline flow. Because the multimodal features are more pronounced in this posterior, we use a multimodal base consisting of three normal distributions with an identity covariance matrix, one centered at 0 in all dimensions, one centered at 0 except for dimensions ϕ_c, ψ centered at 1, and finally one centered at 0 except for ϕ_c, ψ, α, δ centered at 2. This heuristic choice introduces the underlying multimodality into the flow at the start of the training and results in improved diagnostics. The results of this analysis are shown in the last two rows of Table 1, with the corner plots shown in Appendix B. Evidence estimates are again in close agreement. However, our pipeline is $14.8\times$ faster than BILBY.

4 Conclusions

In this work we have constructed an end-to-end pipeline that accelerates Bayesian inference, including both parameter estimation and also model selection. Our pipeline combines the efficient MCMC sampling of JIM with the learned harmonic mean estimator implemented in `harmonic` to compute the evidence. To demonstrate the effectiveness of our pipeline, we applied it to a simulated GW event and inferred its 4-dimensional intrinsic and complete 11-dimensional parameter spaces. We have shown that our pipeline provides accurate evidence estimates at only a fraction of the time required by traditional methods, with a speedup of $5.4\times$ and $14.8\times$ for the 4- and 11-dimensional examples respectively. In future work, we aim to further investigate the optimal treatment of multimodal distributions, exploring the various approaches proposed in literature [e.g. 52, 53, 50, 54], and apply the methods presented here to real data of observed GW events. Both JIM and `harmonic` are open-source, well-documented and tested. Therefore, the pipeline introduced in this work can directly be applied in other scientific fields relying on Bayesian inference.

Acknowledgments and Disclosure of Funding

AP is supported by the UCL Centre for Doctoral Training in Data Intensive Science (STFC grant number ST/W00674X/1). TW and PTHP are supported by the research program of the Netherlands Organization for Scientific Research (NWO). JDM is supported by EPSRC (grant number EP/W007673/1) and STFC (grant number ST/W001136/1). This work was supported by collaborative visits funded by the Cosmology and Astroparticle Student and Postdoc Exchange Network (CASPEN), as well as a G-Research grant. The authors acknowledge the computational resources provided by the LIGO Laboratory's CIT cluster, which is supported by National Science Foundation Grants PHY-0757058 and PHY0823459.

References

[1] Roberto Trotta. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378(1):72–82, 05 2007.

[2] J. Aasi et al. Advanced LIGO. *Class. Quant. Grav.*, 32:074001, 2015.

[3] F. Acernese et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Class. Quant. Grav.*, 32(2):024001, 2015.

[4] B. P. Abbott et al. GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs. *Phys. Rev. X*, 9(3):031040, 2019.

[5] R. Abbott et al. GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *Phys. Rev. X*, 11:021053, 2021.

[6] R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Phys. Rev. X*, 13(4):041039, 2023.

[7] R. Abbott et al. GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run. *Phys. Rev. D*, 109(2):022001, 2024.

[8] A. G. Abac et al. Observation of Gravitational Waves from the Coalescence of a 2.5–4.5 M $_\odot$ Compact Object and a Neutron Star. *Astrophys. J. Lett.*, 970(2):L34, 2024.

[9] Benjamin P Abbott et al. Model comparison from LIGO–Virgo data on GW170817's binary components and consequences for the merger remnant. *Class. Quant. Grav.*, 37(4):045006, 2020.

[10] John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859, 2006.

[11] Greg Ashton, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, Matthew Griffiths, Will Handley, Michael Habeck, et al. Nested sampling for physical scientists. *Nature Reviews Methods Primers*, 2(1):39, 2022.

[12] M. Punturo et al. The third generation of gravitational wave observatories and their science reach. *Class. Quant. Grav.*, 27:084007, 2010.

[13] Matthew Evans et al. A Horizon Study for Cosmic Explorer: Science, Observatories, and Community. 9 2021.

[14] Marica Branchesi et al. Science with the Einstein Telescope: a comparison of different designs. *JCAP*, 07:068, 2023.

[15] Michael J. Williams, John Veitch, and Chris Messenger. Nested sampling with normalizing flows for gravitational-wave inference. *Phys. Rev. D*, 103(10):103006, 2021.

[16] Michael J. Williams, John Veitch, and Chris Messenger. Importance nested sampling with normalising flows. *Mach. Learn. Sci. Tech.*, 4(3):035011, 2023.

[17] Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy. *Nature Phys.*, 18(1):112–117, 2022.

[18] Alex Kolmus, Grégory Baltus, Justin Janquart, Twan van Laarhoven, Sarah Caudill, and Tom Heskes. Fast sky localization of gravitational waves using deep learning seeded importance sampling. *Phys. Rev. D*, 106(2):023032, 2022.

[19] Alex Kolmus, Justin Janquart, Tomasz Baka, Twan van Laarhoven, Chris Van Den Broeck, and Tom Heskes. Tuning neural posterior estimation for gravitational wave inference. 3 2024.

[20] Alvin J. K. Chua and Michele Vallisneri. Learning Bayesian posteriors with neural networks for gravitational-wave inference. *Phys. Rev. Lett.*, 124(4):041102, 2020.

[21] Stephen R. Green, Christine Simpson, and Jonathan Gair. Gravitational-wave parameter estimation with autoregressive neural network flows. *Phys. Rev. D*, 102(10):104057, 2020.

[22] Stephen R. Green and Jonathan Gair. Complete parameter inference for GW150914 using deep learning. *Mach. Learn. Sci. Tech.*, 2(3):03LT01, 2021.

[23] Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-Time Gravitational Wave Science with Neural Posterior Estimation. *Phys. Rev. Lett.*, 127(24):241103, 2021. DINGO paper.

[24] Maximilian Dax, Stephen R. Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference. *Phys. Rev. Lett.*, 130(17):171403, 2023.

[25] Uddipta Bhardwaj, James Alvey, Benjamin Kurt Miller, Samaya Nissanke, and Christoph Weniger. Sequential simulation-based inference for gravitational wave signals. *Phys. Rev. D*, 108(4):042004, 2023.

[26] Kaze W. K. Wong, Maximiliano Isi, and Thomas D. P. Edwards. Fast Gravitational-wave Parameter Estimation without Compromises. *Astrophys. J.*, 958(2):129, 2023.

[27] Jason D McEwen, Christopher GR Wallis, Matthew A Price, and Alessio Spurio Mancini. Machine learning assisted Bayesian model comparison: learnt harmonic mean estimator. *arXiv preprint arXiv:2111.12720*, 2021.

[28] Alicja Polanska, Matthew A Price, Davide Piras, Alessio Spurio Mancini, and Jason D McEwen. Learned harmonic mean estimation of the Bayesian evidence with normalizing flows. *arXiv preprint arXiv:2405.05969*, 2024.

[29] Alan Heavens, Yabebal Fantaye, Arrykrishna Mootoovaloo, Hans Eggers, Zafiirah Hosenie, Steve Kroon, and Elena Sellentin. Marginal Likelihoods from Monte Carlo Markov Chains. *arXiv preprint arXiv:1704.03472*, 2017.

[30] He Jia and Uros Seljak. Normalizing constant estimation with gaussianized bridge sampling. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–14. PMLR, 2020.

[31] Rahul Srinivasan, Marco Crisostomi, Roberto Trotta, Enrico Barausse, and Matteo Breschi. floZ: Evidence estimation from posterior samples with normalizing flows. *arXiv preprint arXiv:2404.12294*, 2024.

[32] Davide Piras, Alicja Polanska, Alessio Spurio Mancini, Matthew A Price, and Jason D McEwen. The future of cosmological likelihood-based inference: accelerated high-dimensional parameter estimation and model comparison. *arXiv preprint arXiv:2405.12965*, 2024.

[33] Roy Frostig, Matthew James Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, 4(9), 2018.

[34] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[35] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, November 2021.

[36] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

[37] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

[38] Thomas D. P. Edwards, Kaze W. K. Wong, Kelvin K. H. Lam, Adam Coogan, Daniel Foreman-Mackey, Maximiliano Isi, and Aaron Zimmerman. ripple: Differentiable and Hardware-Accelerated Waveforms for Gravitational Wave Data Analysis. 2 2023.

[39] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

[40] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. 1 2017.

[41] Kaze W. K. Wong, Marylou Gabrié, and Daniel Foreman-Mackey. flowMC: Normalizing flow enhanced sampling package for probabilistic inference in JAX. *J. Open Source Softw.*, 8(83):5021, 2023.

[42] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proc. Nat. Acad. Sci.*, 119(10):e2109420119, 2022.

[43] Thibeau Wouters, Peter T. H. Pang, Tim Dietrich, and Chris Van Den Broeck. Robust parameter estimation within minutes on gravitational wave signals from binary neutron star inspirals. 4 2024.

[44] Gregory Ashton et al. BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy. *Astrophys. J. Suppl.*, 241(2):27, 2019.

[45] I. M. Romero-Shaw et al. Bayesian inference for compact binary coalescences with bilby: validation and application to the first LIGO–Virgo gravitational-wave transient catalogue. *Mon. Not. Roy. Astron. Soc.*, 499(3):3295–3319, 2020.

[46] Rory J. E. Smith, Gregory Ashton, Avi Vajpeyi, and Colm Talbot. Massively parallel Bayesian inference for transient gravitational-wave astronomy. *Mon. Not. Roy. Astron. Soc.*, 498(3):4492–4502, 2020.

[47] A Spurio Mancini, MM Docherty, MA Price, and JD McEwen. Bayesian model comparison for simulation-based inference. *RAS Techniques and Instruments*, 2(1):710–722, 2023.

[48] Michael A. Newton and Adrian E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994.

[49] Radford M. Neal. Contribution to the discussion of “Approximate Bayesian inference with the weighted likelihood bootstrap” by Newton MA, Raftery AE. *JR Stat Soc Ser A (Methodological)*, 56:41–42, 1994.

[50] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pages 2133–2143. PMLR, 2020.

[51] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[52] Zachary Ziegler and Alexander Rush. Latent normalizing flows for discrete sequences. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2019.

[53] Laurent Dinh, Jascha Sohl-Dickstein, Hugo Larochelle, and Razvan Pascanu. A RAD approach to deep mixture models. *arXiv preprint arXiv:1903.07714*, 2019.

[54] Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4936. PMLR, 2022.

[55] Joshua S. Speagle. dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Mon. Not. Roy. Astron. Soc.*, 493(3):3132–3158, 2020.

[56] Sascha Husa, Sebastian Khan, Mark Hannam, Michael Pürrer, Frank Ohme, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. *Phys. Rev. D*, 93(4):044006, 2016.

[57] Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *Phys. Rev. D*, 93(4):044007, 2016.

Parameter	Description	Injected value	Prior
$\mathcal{M}$	detector-frame chirp mass $[M_\odot]$	60	$[25, 100]$
q	mass ratio m_2/m_1	0.65	$[0.125, 1]$
χ_1	first component aligned spins	0.12	$[-0.99, 0.99]$
χ_1	second component aligned spins	0.53	$[-0.99, 0.99]$
d_L	luminosity distance $[\mathrm{Mpc}]$	2500	$[500, 4000]$
t_c	coalescence time $[\mathrm{s}]$	0	$[-0.01, 0.01]$
ϕ_c	coalescence phase	0.4	$[0, 2\pi]$
ι	inclination angle	2.5	$[0, 2\pi]$
ψ	polarization angle	0.4	$[0, \pi]$
α	right ascension	2.5	$[0, 2\pi]$
δ	declination	2.5	$[0, 2\pi]$

Table 2: Description of the parameters used in the GW simulations, their injected value and uniform prior ranges.

A Setup of simulated gravitational wave signal

The source parameters $\boldsymbol{\theta}$ of a GW signal are inferred from the data by computing their posterior distributions. We set the duration of the simulated signal considered in this work to 4 seconds and analyze the signal in the frequency domain, setting the frequency range to $[20, 2048]$ Hz. We use the GW approximant `IMRPhenomD` [56, 57] for injecting and analysing the signal.

In Tab. 2 below, we provide the parameters used in the GW simulations and the values chosen for the simulated signal. All priors used in the analyses are uniform priors in the ranges shown in Tab. 2. We adopt the standard convention that m_1 refers to the heavier black hole and m_2 to the lighter black hole of the binary system such that the mass ratio $q = m_2/m_1$ is bounded above by 1. When given a GW, represented as time series $d(t)$ and a gravitational waveform model $h(t; \boldsymbol{\theta})$, with θ the parameters of Tab. 2, the likelihood function is given by

$$\begin{aligned}\log p(d \,|\, \theta, M) &= -\frac{1}{2} \langle d - h(\boldsymbol{\theta}), d - h(\boldsymbol{\theta}) \rangle + \text{normalization constant} \\ &= \langle d, h(\boldsymbol{\theta}) \rangle - \frac{1}{2} \langle h(\boldsymbol{\theta}), h(\boldsymbol{\theta}) \rangle - \frac{1}{2} \langle d, d \rangle + \text{normalization constant}.\end{aligned} \tag{4}$$

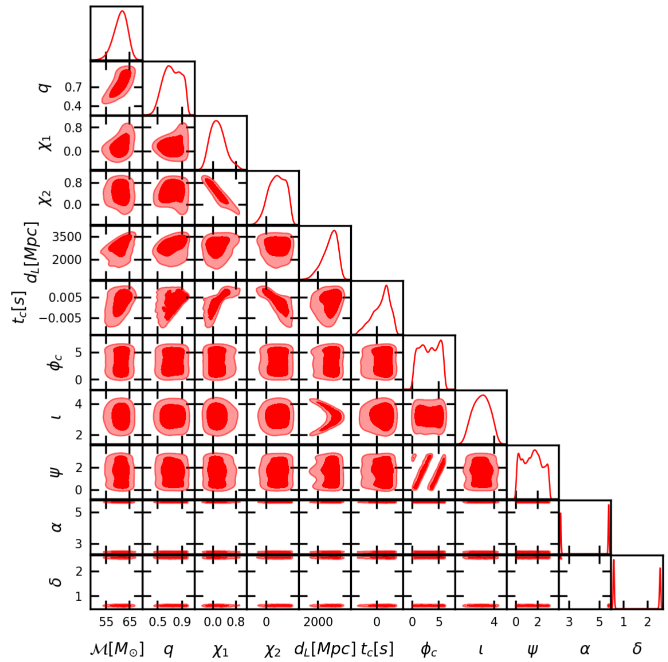
Here, the noise-weighted inner product $\langle a, b \rangle$ is defined as

$$\langle a, b \rangle = 4\mathrm{Re} \int_{f_{\mathrm{low}}}^{f_{\mathrm{high}}} \mathrm{d}f \, \frac{\tilde{a}(f) \tilde{b}^*(f)}{S_n(f)} \,, \tag{5}$$

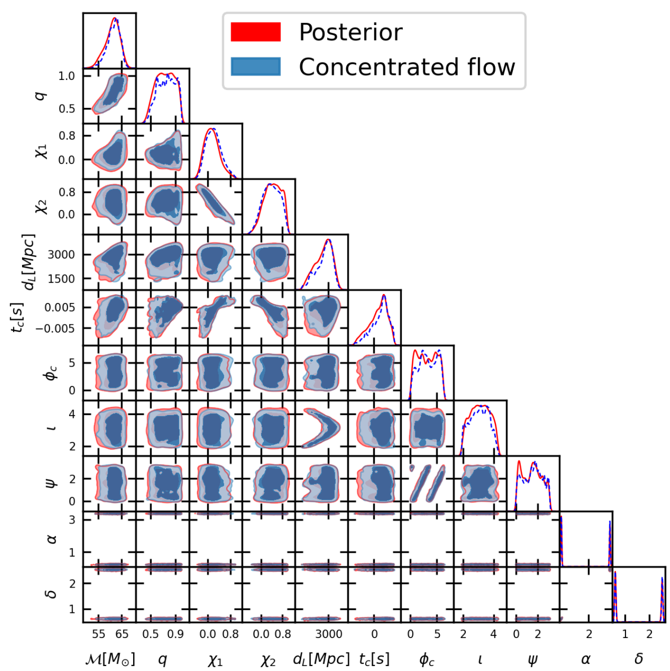
with $S_n(f)$ representing the one-sided power spectral density (PSD), the asterisk symbol denoting complex conjugation and $\tilde{x}(f)$ representing the Fourier transform of a time series $x(t)$.

As shown in Eq. 5, the $\langle d, d \rangle$ term and the normalization constant do not depend on the source parameters $\boldsymbol{\theta}$, which is often neglected in analyses, e.g., in JIM. The Bayesian evidence obtained with such convention is then equivalent to the Bayes factor of the signal hypothesis against the noise hypothesis, thus the Bayes factor quoted by BILBY.

B Corner plot for the 11D example



(a) BILBY



(b) JIM + `harmonic`

Figure 2: Corner plots for the 11-dimensional posterior samples from (a) BILBY and (b) JIM used for inference (solid red) alongside the concentrated flow at $T = 0.8$ used in the learned harmonic mean (dashed blue).