Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks

Niall Jeffrey [1,2]
[1] Laboratoire de Physique de l'École Normale Supérieure,
ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France
[2] Department of Physics & Astronomy, University College London, Gower Street, London, UK
niall.jeffrey@phys.ens.fr

Benjamin D. Wandelt [3,4]
[3]Institut d'Astrophysique de Paris (IAP), UMR 7095, CNRS, Sorbonne Université, France
[4]Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, USA
bwandelt@iap.fr

Abstract

High-dimensional probability density estimation for inference suffers from the "curse of dimensionality". For many physical inference problems, the full posterior distribution is unwieldy and seldom used in practice. Instead, we propose direct estimation of lower-dimensional marginal distributions, bypassing high-dimensional density estimation or high-dimensional Markov chain Monte Carlo (MCMC) sampling. By evaluating the two-dimensional marginal posteriors we can unveil the full-dimensional parameter covariance structure. We additionally propose constructing a simple hierarchy of fast neural regression models, called Moment Networks, that compute increasing moments of any desired lower-dimensional marginal posterior density; these reproduce exact results from analytic posteriors and those obtained from Masked Autoregressive Flows. We demonstrate marginal posterior density estimation using high-dimensional LIGO-like gravitational wave time series and describe applications for problems of fundamental cosmology.

1 Introduction

Estimating the posterior probability density $p(\boldsymbol{\theta}|\boldsymbol{x})$ of a set of parameters $\boldsymbol{\theta}$ given some observed data $\boldsymbol{x}$ is often the primary objective of problems of inference, prediction, or generation. The object $p(\boldsymbol{\theta}|\boldsymbol{x})$ encapsulates all belief and uncertainties about the unknown quantities $\boldsymbol{\theta}$. With this aim in mind, recent advances in neural density estimation have improved our ability to estimate the density $p(\boldsymbol{\theta}|\boldsymbol{x})$ from a set of training examples $\{\boldsymbol{x}_i, \boldsymbol{\theta}_i\}$.

Estimating such probability densities with neural density methods, such as Mixture Density Networks [5], or recent state-of-the-art normalizing flow methods, such as Masked Autogregressive Flows (MAF [26]), provide an excellent way to quantify uncertainty for predicted or inferred parameters and signals $\boldsymbol{\theta}$. Used for likelihood-free inference (also known as simulation-based inference [7, 11]) these density estimation methods can estimate conditional probability densities for parameters and data, either the posterior or the likelihood [3, 25].

For high-dimensional signals, estimation of the full joint density is often not useful and, instead, summaries of lower-dimensional marginal densities are the final goal. For example, the marginal posterior density per pixel, or subsets of pixels, could serve to quantify uncertainty in a reconstructed image.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

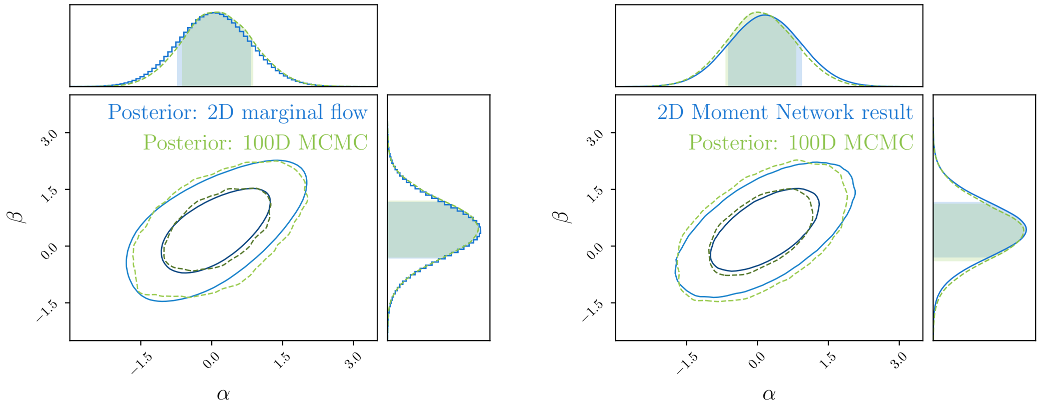

Figure 1: 100-dimensional data model with known reference distribution evaluated with 10^7 MCMC samples. Direct 2D marginal posterior estimation using a MAF ensemble (*left panel*) and representation of 2D Moment Network result (*right panel*) both trained with 8×10^4 simulations.

In this example, the joint posterior marginal for pairs of pixel parameters $\boldsymbol{\theta} = [\alpha, \beta]$ given some observed data $\boldsymbol{x}_{obs}$

$$p(\alpha, \beta | \boldsymbol{x}_{obs}) = \int p(\alpha, \beta, \boldsymbol{\theta}' | \boldsymbol{x}_{obs}) \, \mathrm{d}\boldsymbol{\theta}' \ , \tag{1}$$

would marginalize over all possible values of all other parameters (i.e. the other pixels and latent parameters) $\boldsymbol{\theta}'$. If this were evaluated for all pairs of parameters, all 2D marginal moments of the high-dimensional posterior distribution would be characterized.

In this contribution we present two complementary approaches to evaluate the two-dimensional marginal posterior distributions, marginal flows and Moment Networks (Sec. 2). In Sec. 3 we demonstrate the two methods in comparison to a known underlying posterior density (sampled with MCMC), and show a simulated gravitational wave data model, where the underlying time-ordered signal values form the high-dimensional parameter space to be inferred. In Sec. 4 we describe seemingly intractable problems in cosmological inference that can be solved using marginal posterior density estimation.

2 Marginal posterior density estimation

Motivation In practice, for many physical inference problems with high-dimensional parameter spaces, the full high-dimensional posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}_{obs})$ is not necessary or even interpretable. Instead, the inference goal are the marginal one- and two-dimensional posterior distributions of the parameters [e.g. 1, 20, 29].

Even if full posterior sampling is possible through sophisticated MCMC techniques, e.g. in hierarchical models with known distributions (rather than in a likelihood-free framework), the number of posterior samples needed to compute a d-dimensional marginal grows exponentially with d. For high-dimensional problems, the limited number of independent samples that result in practice only allow for the computation of low-dimensional marginals of the posterior density.

We therefore make the marginal densities the target of our inference problem. We take inspiration from the simplicity of integration when a Monte Carlo sampled representation of the posterior distribution is available. In this case, marginalization is trivial: it amounts simply to ignoring the parameter dimensions to be marginalized over. In this work we show that we can directly bring this powerful notion to simulation-based inference; it allows us to estimate the marginal posterior density (or its moments) directly. This is a powerful approach whenever we are dealing with a large number parameters and effectively removes the practical limitation of simulation-based inference techniques to applications with a low-dimensional parameter space.

Marginal flows Many popular and powerful density estimation methods can be categorized as *normalizing flows*. These use a series of bijective functions to transform from simple known densities (e.g. unit normal) to the target density [19, 21]. MAFs represent the estimated density q as a transformation of a unit normal through a series of autoregressive functions [26, 27]. The networks are trained to give an estimate q of the target distribution p by minimizing a Monte Carlo estimate of the Kullback-Leibler divergence [22]. For a sampling distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ this would be

$$U(\boldsymbol{\varphi}) = -\sum_{i=1}^{N} \log q(\boldsymbol{x}_i|\boldsymbol{\theta}_i;\boldsymbol{\varphi}) \tag{2}$$

with varying network parameters $\boldsymbol{\varphi}$ over the forward-modelled mock data $\boldsymbol{x}_i$. In this same likelihood-free framework, one can directly estimate the posterior distributions for subsets of the large parameter set. For any two parameters α and β of the full $\boldsymbol{\theta}$, one can directly estimate $p(\alpha,\beta|\boldsymbol{x})$ by minimizing

$$U(\boldsymbol{\varphi}) = -\sum_{i=1}^{N} \log q(\alpha_i,\beta_i|\boldsymbol{x}_i;\boldsymbol{\varphi}) \ . \tag{3}$$

The resulting density will indeed be an estimate of the marginal posterior for the chosen parameter pair (eq. 1) if all parameters of $\boldsymbol{\theta}$ (not just the chosen pair) are drawn from the prior $p(\boldsymbol{\theta})$ to generate the training data. This procedure also avoids the need for data compression steps [4, 10], as we condition on high-dimensional data $\boldsymbol{x}$ rather than estimating its density, and any nuisance parameters are automatically marginalized away, provided they have been sufficiently sampled in the training data.

Moment Networks In practice, posterior estimates often serve principally to compute posterior moments. Moment Networks allow us to side-step the problem of estimating the posterior density and directly skip to estimation of location, scale, and covariance of the parameters (and possibly higher-order moments). When this is sufficient, Moment Networks allow the use of far simpler neural network architectures, which reduces risk of training failure, and boosts inference speed.

We begin by noting that if we find some function of our data $\mathcal{F}(\boldsymbol{x})$ that minimizes an L_2 loss over the distribution of possible training examples $\{\boldsymbol{x}_i,\boldsymbol{\theta}_i\}$,

$$J_0 = \int ||\boldsymbol{\theta} - \mathcal{F}(\boldsymbol{x})||^2 p(\boldsymbol{x},\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{\theta} \ , \tag{4}$$

then $\mathcal{F}$, which we represent as a neural network, evaluated for the observed data is the mean of the posterior distribution $\mathcal{F}(\boldsymbol{x}_{obs}) = \langle\boldsymbol{\theta}\rangle_{\boldsymbol{\theta}|\boldsymbol{x}_{obs}}$. It is therefore possible to create a hierarchy of networks to generate further moments of the posterior distribution. For example, the function $\mathcal{G}$ that minimizes

$$J_1 = \int ||(\boldsymbol{\theta} - \mathcal{F}_{\text{fixed}}(\boldsymbol{x}))^2 - \mathcal{G}(\boldsymbol{x})||^2 p(\boldsymbol{x},\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{\theta} \ , \tag{5}$$

for fixed, already trained $\mathcal{F}$, is such that $\mathcal{G}(\boldsymbol{x}_{obs})$, is the set of posterior variances [2, 16]. The objective functions for marginal posterior parameter covariances can be similarly constructed.

By sampling the full parameter space from the prior distributions $p(\boldsymbol{\theta})$, the functions $\mathcal{F}$ and $\mathcal{G}$ can be combined to output the posterior means, variances, and covariances for subsets of the full set of parameters; the marginalization over other parameters is implicitly done during training. This result is exact and independent of the true underlying posterior or prior distributions.

The Moment Network solves for the marginal posterior moments by construction, and therefore does not suffer the problem of mode collapse in variational inference with multi-modal posteriors, which can lead to underpredicted uncertainty. Outside the likelihood-free framework, if one does have information about the functional form of the posterior, one can fit the posterior parameters to the marginal moments (see Sec. 4).

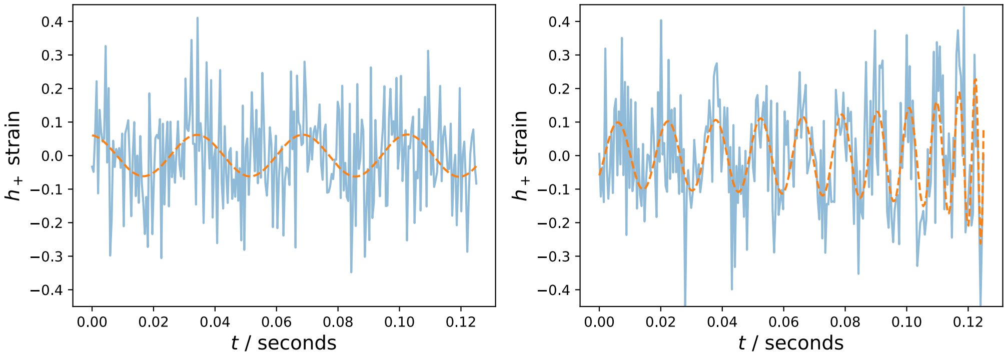

Figure 2: Two example simulated gravitation wave time series signals for the strain "+" polarization h_+ with realistic LIGO-like noise. The dashed line shows the true strain values over time.

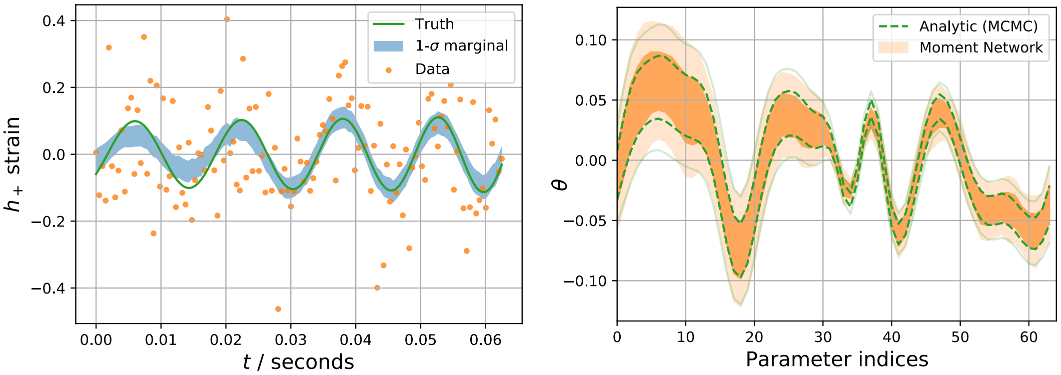

Figure 3: *Left panel*: Moment Network (MN) estimate of the 1-σ standard deviation per strain $h_+(\Delta t)$ parameter given 62.5ms of data (Fig. 2). *Right panel*: For each of 64 parameters (c.f. time step), contours are the marginal posterior σ from MN (*shaded orange*) and MCMC (*dashed green*).

3 Experiments

High-dimensional inference: We use a 100-dimensional parameter inference toy model to demonstrate marginal posterior estimation for pairs of parameters. The model consists of 100-element data vectors with non-stationary Gaussian noise and a Gaussian prior distribution with non-trivial covariance introducing parameter correlation. We estimate the marginal posterior density for parameter pairs using MAF (using the pyDELFI package [3]) and estimate the 2D marginal mean, variance and covariance using Moment Networks. The results are represented in Fig. 1.

As a reference, we can directly sample the posterior distribution using high-dimensional MCMC, which took 10^7 draws from the likelihood. The normalizing flow result would have been intractable if the density estimation target was with respect to the full parameter space or to the data space. With a marginal flow, changing the target to the pairs of parameters, it is simple (with a basic 2-GPU:12GB set-up) to evaluate all marginal posterior pairs (often represented as a so-called "corner plot").

With the same set-up, the Moment Network hierarchy was able to accurately evaluate the means, variances and covariances of the marginals (see Fig. 1) with a few seconds of training and evaluation in $\mathcal{O}(10^{-2})$s without requiring any sampling or grid evaluation. For many practical applications of inference in the physical sciences, these marginal joint moments would be the final goal.

Gravitational wave signal demonstration: Fig. 2 shows two example simulated gravitational wave time series. The two signals (*dashed orange*) are ~ 0.12s intervals from the 1s before a binary black hole merger using the SEOBNRv4 model [6].

We have simplified the problem for this demonstration by removing all "geometric" effects (black hole spin, inclination, detector geometry) and use only the $+$ polarization for the detector strain $h_{+,\times}$. We do, however, sample the merger events with an independent prior distribution for each mass $p(M) = \mathcal{U}(10, 30)M_{\rm sol}$ and distance $p(\chi) = \mathcal{U}(500, 1500)\rm{Mpc}$. The noise is LIGO-like noise, which, along with the signal, was generated using the pyCBC package for 35000 simulations.

With the time series elements forming a high-dimensional parameter space, the *left panel* of Fig. 3 shows a representation of the marginal posterior standard deviation for each of 128 parameters for a simulated data set. This result was evaluated using the trained Moment Network. The *right panel* shows a validation case of similar complexity to the gravitational wave model, but with known likelihood. The Moment Network trained on simulations matches accurately with a long-run MCMC chain. This validates our approach.

4 Discussion & cosmological applications

Though the direct density estimation of marginal posteriors is much more robust than the estimation of the full posterior, it may still suffer from well-known issues of density estimation. Moment Networks optimize a completely different set of objective functions to return estimates of the posterior moments. This affords an opportunity to cross-validate, as moments of the estimated marginal posteriors should match those from the Moment Network. If the results are inconsistent for an initial set of simulations, then there may be insufficient network complexity (the network complexity for both methods scales similarly) or insufficient number of simulations.

Thus far, density estimation likelihood-free inference in cosmology has generally been limited to a few parameters [e.g. 3, 8, 17, 23, 30, 32]. Though simulation-based inference of cosmological fields (including dark matter) can be integral to many analyses [e.g. 9, 18, 28, 31], it can be intractable to estimate the full posterior due to high-dimensionality. With the approach we propose, joint marginal posteriors (and associated moments) for reconstructed cosmological fields can be directly evaluated.

For cases where it is possible to sample, marginal flows and Moment Networks still provide advantages. One particularly ambitious cosmological sampler BORG [14, 15] samples the $\sim 10^7-$dimensional posterior density of the initial conditions of the Universe for given galaxy data, using a non-linear forward model including the physics and data effects. Though the full posterior is sampled, the complexity of the sampler and the inherently sequential nature of MCMC limits the number of independent samples to $\mathcal{O}(10^3)$; sufficient only to estimate low-dimensional marginal posteriors and their moments. The approach proposed in this work could use similar computational resources to generate simulations to train marginal flows and Moment Networks in parallel (rather than sequentially) and efficiently output low-dimensional marginal posteriors and moments.

Beyond general high-dimensional inference, the principal motivation for this work (Sec. 2), we plan to explore a wide range of further applications of marginal flows and Moment Networks to probe the fundamental physics of the Universe in future studies.

Demonstration code can be found at: github.com/NiallJeffrey/MomentNetworks

Broader Impact

This work provides a robust approach to quantify uncertainty from high-dimensional parameter spaces by estimating marginal posterior distributions or their associated moments directly. This has immediate application for parameter and model inference in astrophysics and cosmology, and the physical sciences more generally. We note that if the method is misunderstood or misapplied, incorrect uncertainty quantification or risk analysis would follow. To mitigate this, diagnostic and validation methods can be applied (e.g. ensembles of neural density estimators or quantile tests) or, as proposed in this work, by comparing results between likelihood-free methods (e.g. marginal flows and Moment Networks). The approach in this work can be applied to signal inference and prediction more generally (including fast image analysis, time series prediction, forecasting, and quantifying uncertainty for decision making).

Acknowledgements

Software used pyDELFI (github.com/justinalsing/pydelfi) for MAF density estimation implementation [3]; chainconsumer (samreay.github.io) for Fig. 1 [13]; emcee (emcee.readthedocs.io) for MCMC sampling [12]; pyCBC (pycbc.org) for gravitational wave data simulation [24].

The authors thank Tom Charnock for useful discussions. NJ acknowledges funding from the École Normale Supérieure (ENS). BDW acknowledges support by the ANR BIG4 project, grant ANR-16-CE23-0002 of the French Agence Nationale de la Recherche; and the Labex ILP (reference ANR-10-LABX-63) part of the Idex SUPER, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme Investissements d'avenir under the reference ANR-11-IDEX-0004-02. The Flatiron Institute is supported by the Simons Foundation.

References

[1] T. M. C. Abbott, F. B. Abdalla, A. Alarcon, J. Aleksić, S. Allam, S. Allen, A. Amara, J. Annis, J. Asorey, S. Avila, and et al. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. *PRD*, 98(4):043526, Aug 2018.

[2] J. Adler and O. Öktem. Deep Bayesian Inversion. *arXiv e-prints*, page arXiv:1811.05910, Nov. 2018.

[3] J. Alsing, T. Charnock, S. Feeney, and B. Wand elt. Fast likelihood-free cosmology with neural density estimators and active learning. *MNRAS*, 488(3):4440–4458, Sept. 2019.

[4] J. Alsing and B. Wandelt. Generalized massive optimal data compression. *MNRAS*, 476(1):L60–L64, May 2018.

[5] C. Bishop. Mixture density networks. Working paper, Aston University, 1994.

[6] A. Bohé, L. Shao, A. Taracchini, A. Buonanno, S. Babak, I. W. Harry, I. Hinder, S. Ossokine, M. Pürrer, V. Raymond, and et al. Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors. *Physical Review D*, 95(4), Feb 2017.

[7] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020.

[8] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe, and K. Cranmer. Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning. *The Astrophysical Journal*, 886(1):49, Nov 2019.

[9] J. Caldeira, W. Wu, B. Nord, C. Avestruz, S. Trivedi, and K. Story. Deepcmb: Lensing reconstruction of the cosmic microwave background with deep neural networks. *Astronomy and Computing*, 28:100307, Jul 2019.

[10] T. Charnock, G. Lavaux, and B. D. Wandelt. Automatic physical inference with information maximizing neural networks. *PRD*, 97(8):083004, Apr. 2018.

[11] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

[12] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC Hammer. *PASP*, 125(925):306, Mar. 2013.

[13] S. R. Hinton. ChainConsumer. *The Journal of Open Source Software*, 1:00045, Aug. 2016.

[14] J. Jasche, F. Leclercq, and B. Wandelt. Past and present cosmic structure in the sdss dr7 main sample. *Journal of Cosmology and Astroparticle Physics*, 2015(01):036–036, Jan 2015.

[15] J. Jasche and B. D. Wandelt. Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *Monthly Notices of the Royal Astronomical Society*, 432(2):894–913, Apr 2013.

[16] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge University Press, 2003.

[17] N. Jeffrey, J. Alsing, and F. Lanusse. Likelihood-free inference with neural compression of DES SV weak lensing map statistics, Nov 2020. staa3594, arXiv:2009.08459.

[18] N. Jeffrey, F. Lanusse, O. Lahav, and J.-L. Starck. Deep learning dark matter map reconstructions from DES SV weak lensing data. *MNRAS*, 492(4):5023–5029, Mar. 2020.

[19] D. Jimenez Rezende and S. Mohamed. Variational Inference with Normalizing Flows. *arXiv e-prints*, page arXiv:1505.05770, May 2015.

[20] S. Joudaki and KiDS Collaboration. KiDS-450 + 2dFLenS: Cosmological parameter constraints from weak gravitational lensing tomography and overlapping redshift-space galaxy clustering. *MNRAS*, 474(4):4894–4924, Mar. 2018.

[21] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

[22] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[23] P. Lemos, N. Jeffrey, L. Whiteway, O. Lahav, N. I. Libeskind, and Y. Hoffman. The sum of the masses of the Milky Way and M31: a likelihood-free inference approach. *arXiv e-prints*, page arXiv:2010.08537, Oct. 2020.

[24] A. Nitz, I. Harry, D. Brown, C. M. Biwer, J. Willis, T. D. Canton, C. Capano, L. Pekowsky, T. Dent, A. R. Williamson, G. S. Davies, S. De, M. Cabero, B. Machenschalk, P. Kumar, S. Reyes, D. Macleod, dfinstad, F. Pannarale, T. Massinger, M. Tápai, L. Singer, S. Kumar, S. Khan, S. Fairhurst, A. Nielsen, SSingh087, shasvath, B. U. V. Gadre, and I. Dorrington. gwastro/pycbc: Pycbc release v1.16.10. *Zenodo*, Oct. 2020.

[25] G. Papamakarios and I. Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

[26] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[27] G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 837–848. PMLR, 16–18 Apr 2019.

[28] M. A. Petroff, G. E. Addison, C. L. Bennett, and J. L. Weiland. Full-sky cosmic microwave background foreground cleaning using machine learning, 2020.

[29] PlanckCollaboration. Planck 2018 results. VI. Cosmological parameters. *AAP*, 641:A6, Sept. 2020.

[30] D. K. Ramanah, R. Wojtak, Z. Ansari, C. Gall, and J. Hjorth. Dynamical mass inference of galaxy clusters with neural flows, 2020.

[31] M. Shirasaki, N. Yoshida, and S. Ikeda. Denoising weak lensing mass maps with deep learning. *PRD*, 100(4):043527, Aug. 2019.

[32] P. L. Taylor, T. D. Kitching, J. Alsing, B. D. Wandelt, S. M. Feeney, and J. D. McEwen. Cosmic shear: Inference from forward models. *Physical Review D*, 100(2), Jul 2019.