Arbitrary Marginal Neural Ratio Estimation for Simulation-based Inference

François Rozet
University of Liège
francois.rozet@uliege.be

Gilles Louppe
University of Liège
g.louppe@uliege.be

Abstract

In many areas of science, complex phenomena are modeled by stochastic parametric simulators, often featuring high-dimensional parameter spaces and intractable likelihoods. In this context, performing Bayesian inference can be challenging. In this work, we present a novel method that enables amortized inference over arbitrary subsets of the parameters, without resorting to numerical integration, which makes interpretation of the posterior more convenient. Our method is efficient and can be implemented with arbitrary neural network architectures. We demonstrate the applicability of the method on parameter inference of binary black hole systems from gravitational waves observations.

1 Introduction

Formally, a simulator is a stochastic forward model that takes a vector of parameters $\theta \in \Theta$ as input, samples internally a series $z \in \mathcal{Z} \sim p(z|\theta)$ of latent variables and finally produces an observation $x \in \mathcal{X} \sim p(x|\theta, z)$ as output, thereby defining an implicit likelihood $p(x|\theta)$. This likelihood typically is *intractable* as it corresponds to $p(x|\theta) = \int_{\mathcal{Z}} p(x, z|\theta) \, \mathrm{d}z$, the integral of the joint likelihood $p(x, z|\theta)$ over *all* possible trajectories through the latent space $\mathcal{Z}$. In Bayesian inference, we are interested in the posterior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta')p(\theta') \, \mathrm{d}\theta'} \tag{1}$$

for some observation(s) x and a prior $p(\theta)$, which not only involves the intractable likelihood $p(x|\theta)$ but also an intractable integral over the parameter space Θ. The omnipresence of this problem gave rise to a rapidly expanding field of research [1] commonly referred to as *simulation-based* inference. Recent approaches [2–5] are to learn a surrogate model $\hat{p}(\theta|x)$ of the posterior and, then, proceed as if the latter was tractable.

However, domain scientists are not always interested in the full set of simulator parameters at once. In particular, when interpreting posterior predictions, they generally study several small parameter subsets, like singletons or pairs, while ignoring the others. This task corresponds to estimating the marginal posterior $p(\theta_a|x) = \int_{\Theta_b} p(\theta|x) \, \mathrm{d}\theta_b$ over parameter subspaces $\Theta_a \leq \Theta$ of interest, while the complement subspaces $\Theta_b : \Theta_a \times \Theta_b = \Theta$ are unobserved. To this end, most applications [4, 5] resort to numerical integration of a surrogate $\hat{p}(\theta|x)$ of the full posterior, which is computationally expensive if Θ_b is large.

A solution to get rid of numerical integration is to learn directly a surrogate $\hat{p}(\theta_a|x)$ by considering θ_b as part of the latent variables. If we are interested in a single or a few predetermined subspaces, this approach is reasonable and leads to accurate estimation of marginal posteriors [6, 7]. However, if we need to choose *arbitrarily* the subspace Θ_a at inference time, this solution is not viable anymore as there exists an exponential number $(2^{\dim(\Theta)} - 1)$ of marginal posteriors.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

Contribution We build upon neural ratio estimation (NRE) [3, 8] to enable integration-less marginal posterior estimation over arbitrary parameter subspaces. The key idea is to introduce, as input of the ratio estimator, a binary mask $a \in \{0, 1\}^{\dim(\Theta)}$ indicating the current subspace Θ_a of interest. Intuitively, this allows the network to distinguish the subspaces and, thereby, to learn a different ratio for each of them. Our method, dubbed arbitrary marginal neural ratio estimation (AMNRE), can be implemented with arbitrary neural network architectures, including multi-layer perceptrons (MLPs) [9] and residual networks [10]. AMNRE is an amortized method, meaning that inference is simulation-free and can be repeated several times with distinct observations, without retraining. The counterpart is that AMNRE could require a lot of training simulations to produce accurate predictions. The implementation is available at `https://github.com/francois-rozet/amnre`.

Related work Imputation methods [11–13] were the first to introduce a binary mask to condition networks with respect to which features are missing. This trick allowed to train a single generative network for all combinations of missing features. Our method differs in that it does not generate likely replacements for the missing features but evaluates the likeliness, conditionally to an observation, of those that are provided.

2 Arbitrary marginal neural ratio estimation

NRE The principle of NRE [3] is to train a classifier network $d_\phi : \Theta \times \mathcal{X} \mapsto [0, 1]$ to discriminate between pairs (θ, x) equally sampled from the joint distribution $p(\theta, x)$ and the product of the marginals $p(\theta)p(x)$. Formally, the optimization problem is

$$\phi^* = \arg\min_\phi \mathop{\mathbb{E}}_{p(\theta,x)p(\theta')} \Big[\mathcal{L}(d_\phi(\theta, x)) + \mathcal{L}(1 - d_\phi(\theta', x))\Big], \tag{2}$$

where $\mathcal{L}(p) = -\log p$ is the negative log-likelihood. For this task, the decision function modeling the Bayes optimal classifier [3] is

$$d(\theta, x) = \frac{p(\theta, x)}{p(\theta, x) + p(\theta)p(x)}, \tag{3}$$

thereby defining the likelihood-to-evidence (LTE) ratio

$$r(\theta, x) = \frac{d(\theta, x)}{1 - d(\theta, x)} = \frac{p(\theta, x)}{p(\theta)p(x)} = \frac{p(x|\theta)}{p(x)} = \frac{p(\theta|x)}{p(\theta)}. \tag{4}$$

Consequently, NRE gives access to an estimator $\log r_\phi(\theta, x) = \text{logit}(d_\phi(\theta, x))$ of the LTE log-ratio and a surrogate $\hat{p}(\theta|x) = r_\phi(\theta, x)p(\theta)$ for the posterior density.

AMNRE With the additional binary mask $a \in \{0, 1\}^{\dim(\Theta)}$, the classifier takes the form $d_\phi(\theta_a, x, a)$ and the optimization problem becomes

$$\phi^* = \arg\min_\phi \mathop{\mathbb{E}}_{p(\theta,x)p(\theta')} \mathop{\mathbb{E}}_{p(a)} \Big[\mathcal{L}(d_\phi(\theta_a, x, a)) + \mathcal{L}(1 - d_\phi(\theta'_a, x, a))\Big], \tag{5}$$

where $\theta_a = (\theta_i : a_i = 1)$ and $p(a)$ is a mask distribution. In this context, the Bayes optimal classifier (see Appendix A) is

$$d(\theta_a, x, a) = \frac{p(\theta_a, x)}{p(\theta_a, x) + p(\theta_a)p(x)}, \tag{6}$$

meaning that AMNRE gives access to an estimator $\log r_\phi(\theta_a, x, a) = \text{logit}(d_\phi(\theta_a, x, a))$ of all marginal LTE log-ratios and a surrogate $\hat{p}(\theta_a|x) = r_\phi(\theta_a, x, a)p(\theta_a)$ for all marginal posteriors.

AMNRE does not have any particular architectural requirements, with the notable exception of the variable input size of θ_a. To make the method more convenient, θ_a is replaced by the element-wise product $\theta \cdot a$ ($\theta_a \cdot 1$ and $\theta_b \cdot 0$), carrying the same information at fixed size. The mask a is still required as input since a zero in $\theta \cdot a$ does not unambiguously indicate a zero in a. To prevent numerical stability issues when $d_\phi(\theta_a, x, a) \to 1$, the approximate log-ratio $\log r_\phi(\theta_a, x, a)$ is extracted from the neural network and the class prediction is recovered by application of the sigmoid function.

The mask distribution is an important part of AMNRE's training. If some masks a have a small probability $p(a)$ to be selected, it is likely that the estimator will not model their respective marginal posteriors as well as other, more frequent masks. In our experiments, we adopt a uniform mask distribution $p(a) = (2^{\dim(\Theta)} - 1)^{-1}$, leaving the study of this aspect to future work.

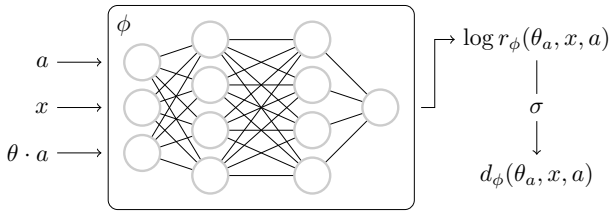

Figure 1: Illustration of AMNRE's classifier architecture.

3 Experiments and results

3.1 Simple likelihood and complex posterior

Papamakarios et al. [14] introduce a toy simulator with a 5-dimensional parameter space $\Theta \subseteq \mathbb{R}^5$, for which the likelihood is tractable. Despite its *simple likelihood*, the simulator has a *complex posterior* (SLCP) with four symmetric modes. Hence, SLCP is a non-trivial posterior estimation benchmark that allows to retrieve the ground-truth posterior through Markov chain Monte Carlo (MCMC) sampling [15, 16] of the likelihood.

We apply AMNRE on SLCP and compare the learned surrogates with the ground-truth posterior. Training details are provided in Appendix C. In Figure 2, we observe that AMNRE 1d and 2d surrogates are in close agreement with the ground-truth. The structure of the distribution, represented by the credible regions, is modeled correctly, even in low density regions. We also note that the four symmetric modes (see θ_3 and θ_4) are properly recovered, which is sometimes challenging for traditional sampling methods. Concerning the parameter θ_5, we observe that the network is slightly underconfident around the mode, which could indicate that, among the five parameters of SLCP, θ_5 is the hardest to infer. Finally, in Figure 4, we see that AMNRE is also able to recover the full 5d posterior and the predictions are very consistent with the 1d and 2d surrogates.

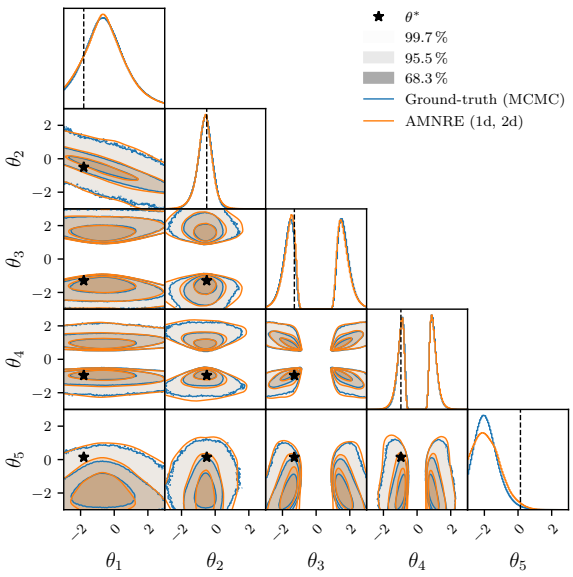

Figure 2: Ground-truth posterior against AMNRE 1d and 2d marginal surrogates, for an observation x^* of the SLCP testing set. Density is averaged over three training instances. Contours represent the 68.3 %, 95.5 % and 99.7 % highest posterior density regions. Stars represent the true parameters θ^* of the observation.

3.2 Gravitational waves

In recent years, the observations of gravitational waves (GWs) from compact binary coalescences systems have had a massive impact on our understanding of the Universe, partly thanks to inference of the systems' parameters. To obtain posterior samples, the LIGO/Virgo collaboration currently applies MCMC [15, 16] or nested sampling [17, 18] algorithms to involved physical models of the likelihood of emitted waves [19, 20]. With these approaches, posterior calculation typically takes days for binary black hole (BBH) mergers and has to be repeated from scratch for each observation.

As a proof of concept, we employ AMNRE to infer the full 15-dimensional set of precessing quasi-circular BBH parameters, given GW observations from the LIGO/Virgo detectors. The simulator details are provided in Appendix B. After training (see Appendix C for details), we evaluate the learned surrogate model on data surrounding GW150914, the first recorded GW event [21]. As reference, we use the posterior samples produced by Bilby [20] with the `dynesty` [18, 22] nested

sampler (MIT License), which leverages the true likelihood. It takes 3 days for Bilby to complete the posterior inference of GW150914, while our network builds histograms (100 bins per dimension) of all 1d and 2d marginal posteriors in about 1 second, on a single 1080Ti GPU.

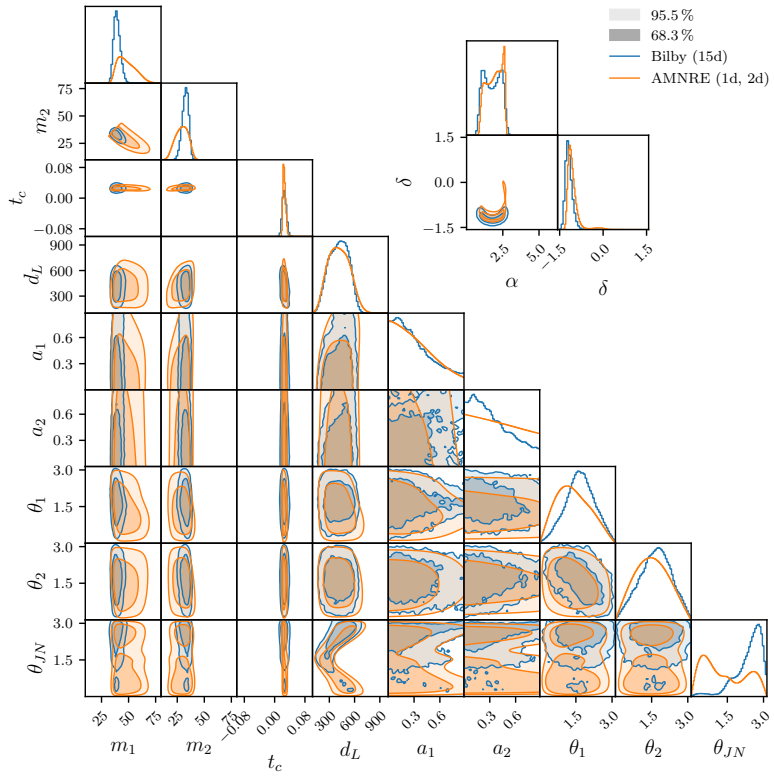

Figure 3: AMNRE 1d and 2d surrogate marginal posteriors against marginalized Bilby posterior samples over a subset of the parameters, for the GW150914 observation. Density is averaged over three training instances.

As can be seen in Figure 3, the surrogate marginal posteriors of AMNRE share the same structure as the marginalized posterior inferred by Bilby. For some parameter subsets, especially those containing the masses m_1 and m_2 and the inclination angle θ_{JN}, the predictions present significant inaccuracies. For the masses, the surrogates are underconfident but predict the correct modes. For other parameters, including the coalescence time t_c, luminosity distance d_L and sky location (α, δ), the surrogates are in close agreement with Bilby.

4 Conclusions

This work introduces AMNRE, a novel simulation-based inference method that enables integration-less marginal posterior estimation over arbitrary parameter subspaces. Through our experiment with the SLCP toy simulator, we demonstrate that the proposed algorithm is indeed able to recover the ground-truth posterior and marginalize it arbitrarily. This experiment also highlights the capacity of AMNRE to model multi-modal distributions, even using a very basic MLP architecture.

The second experiment consists in applying AMNRE to the problem of BBH parameter inference from GW observations. This proof of concept demonstrates that AMNRE is able to analyze GW events several order of magnitude faster (seconds instead of days) than traditional sampling methods. However, if most of the surrogate marginal posteriors seem accurate, some present significant inaccuracies. Possible causes are a lack of estimator expressiveness or insufficient simulation budget; aspects we do not properly study in this work. Still, we believe these results to be a promising demonstration of the applicability of AMNRE for convenient interpretation of the posterior in challenging scientific settings.

Acknowledgments

The authors would like to thank Antoine Wehenkel, Arnaud Delaunoy and Joeri Hermans for the insightful discussions and comments. Gilles Louppe is recipient of the ULiège - NRB Chair on Big Data and is thankful for the support of the NRB.

References

[1] Kyle Cranmer et al. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[2] David Greenberg et al. "Automatic posterior transformation for likelihood-free inference". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2404–2414.

[3] Joeri Hermans et al. "Likelihood-free mcmc with amortized approximate ratio estimators". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4239–4248.

[4] Pedro J Gonçalves et al. "Training deep neural density estimators to identify mechanistic models of neural dynamics". In: *Elife* 9 (2020), e56261.

[5] Stephen R Green et al. "Complete parameter inference for GW150914 using deep learning". In: *Machine Learning: Science and Technology* 2.3 (2021), 03LT01.

[6] Arnaud Delaunoy et al. "Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization". In: *arXiv preprint arXiv:2010.12931* (2020).

[7] Benjamin Kurt Miller et al. "Truncated Marginal Neural Ratio Estimation". In: *arXiv preprint arXiv:2107.01214* (2021).

[8] Kyle Cranmer et al. "Approximating likelihood ratios with calibrated discriminative classifiers". In: *arXiv preprint arXiv:1506.02169* (2015).

[9] Kurt Hornik et al. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[10] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[11] Jinsung Yoon et al. "Gain: Missing data imputation using generative adversarial nets". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5689–5698.

[12] Mohamed Ishmael Belghazi et al. "Learning about an exponential amount of conditional distributions". In: *arXiv preprint arXiv:1902.08401* (2019).

[13] Yang Li et al. "ACFlow: Flow models for arbitrary conditional likelihoods". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5831–5841.

[14] George Papamakarios et al. "Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 837–848.

[15] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970).

[16] Ming-Hui Chen et al. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

[17] John Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian analysis* 1.4 (2006), pp. 833–859.

[18] Edward Higson et al. "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation". In: *Statistics and Computing* 29.5 (2019), pp. 891–913.

[19] John Veitch et al. "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library". In: *Physical Review D* 91.4 (2015), p. 042003.

[20] Gregory Ashton et al. "BILBY: a user-friendly Bayesian inference library for gravitational-wave astronomy". In: *The Astrophysical Journal Supplement Series* 241.2 (2019), p. 27.

[21] Benjamin P Abbott et al. "Observation of gravitational waves from a binary black hole merger". In: *Physical review letters* 116.6 (2016), p. 061102.

[22] Joshua S Speagle. "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences". In: *Monthly Notices of the Royal Astronomical Society* 493.3 (2020), pp. 3132–3158.

[23] Sebastian Khan et al. “Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era”. In: *Physical Review D* 93.4 (2016), p. 044007.

[24] Alejandro Bohé et al. “PhenomPv2–technical notes for the LAL implementation”. In: *LIGO Technical Document, LIGO-T1500602-v4* (2016).

[25] Djork-Arné Clevert et al. “Fast and accurate deep network learning by exponential linear units (elus)”. In: *arXiv preprint arXiv:1511.07289* (2015).

[26] Sergey Ioffe et al. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[27] Diederik P Kingma et al. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[28] Ilya Loshchilov et al. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[29] Ilya Loshchilov et al. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Sections 3 and 4.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] Some assumptions are implicit.
   (b) Did you include complete proofs of all theoretical results? [No] An *informal* proof of the correctness of AMNRE is provided in Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The implementation is available as a public GitHub repository.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the caption of Figure 5 (a single 1080TI GPU).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Appendix B.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Correctness of AMNRE

Reformulating (5), we have

$$\begin{aligned}
L &= \iiint_{\Theta\times\mathcal{X}\times\Theta} p(\theta,x)p(\theta') \mathop{\mathbb{E}}_{p(a)}\left[\mathcal{L}(d_\phi(\theta_a,x,a)+\mathcal{L}(1-d_\phi(\theta'_a,x,a))\right] \mathrm{d}\theta\,\mathrm{d}x\,\mathrm{d}\theta' \\
&= \iint_{\Theta\times\mathcal{X}} \mathop{\mathbb{E}}_{p(a)}\left[p(\theta,x)\mathcal{L}(d_\phi(\theta_a,x,a))+p(\theta)p(x)\mathcal{L}(1-d_\phi(\theta_a,x,a))\right] \mathrm{d}\theta\,\mathrm{d}x \\
&= \mathop{\mathbb{E}}_{p(a)} \iint_{\Theta_a\times\mathcal{X}} \underbrace{\left[p(\theta_a,x)\mathcal{L}(d_\phi(\theta_a,x,a))+p(\theta_a)p(x)\mathcal{L}(1-d_\phi(\theta_a,x,a))\right]}_{\ell(d_\phi(\theta_a,x,a))} \mathrm{d}\theta_a\,\mathrm{d}x,
\end{aligned}$$

which is minimized only if each term $\ell(d_\phi(\theta_a,x,a))$ is itself minimized. Assuming $p(\theta_a)p(x)>0$, if $p(\theta_a,x)=0$, ℓ is uniquely minimized by the value 0. Otherwise, if $p(\theta_a,x)>0$, the minimum is only reached by a value q such that

$$\begin{aligned}
0 &= \frac{\mathrm{d}\ell(q)}{\mathrm{d}q} \\
&= p(\theta_a,x)\frac{\mathrm{d}\mathcal{L}(q)}{\mathrm{d}q} + p(\theta_a)p(x)\frac{\mathrm{d}\mathcal{L}(1-q)}{\mathrm{d}q} \\
&= p(\theta_a,x)\frac{-1}{q} + p(\theta_a)p(x)\frac{1}{1-q} \\
\Leftrightarrow \quad q &= \frac{p(\theta_a,x)}{p(\theta_a,x)+p(\theta_a)p(x)} = d(\theta_a,x,a).
\end{aligned}$$

Importantly, if $p(\theta_a,x)=0$, $d(\theta_a,x,a)=0$ and still minimizes ℓ. Therefore, as long as $p(\theta_a)p(x)>0$, $d(\theta_a,x,a)$ is the optimal classifier.

B Simulators

B.1 Simple likelihood and complex posterior

In this toy simulator, $\theta\in\mathbb{R}^5$ parametrizes a 2d multivariate Gaussian from which four points are independently sampled to construct an observation x. The generative process [14] is

$$\begin{aligned}
&\theta_i \sim \mathcal{U}(-3,3) \quad \text{for} \quad i=1,\dots,5 \\
&s_1=\theta_3^2, \quad s_2=\theta_4^2, \quad \rho=\tanh(\theta_5) \\
&\mu=(\theta_1,\theta_2), \quad \Sigma=\begin{pmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{pmatrix} \\
&x=(z_1,\dots,z_4) \quad \text{where} \quad z_j\sim\mathcal{N}(\mu,\Sigma),
\end{aligned}$$

for which the likelihood $p(x|\theta)=\prod_j p(z_j|\theta)$ is tractable.

B.2 Gravitational waves

As we do not have enough knowledge in the domain, we borrow the BBH simulator implemented by Green et al. [5] (MIT License). We succinctly describe the generative process in this section. For more information, please refer to the original paper or the implementation.

Prior We perform inference over the full 15-dimensional set of precessing quasi-circular BBH parameters: component masses (m_1, m_2), reference phase ϕ_c, coalescence time t_c, luminosity distance d_L, spin magnitudes (a_1, a_2), spin angles $(\theta_1, \theta_2, \phi_{12}, \phi_{JL})$, inclination angle θ_{JN}, polarization angle ψ, and sky location (α, δ). To analyze GW150914, we take a prior uniform over

$$\begin{aligned} 10\,\mathrm{M}_\odot \leq m_i &\leq 80\,\mathrm{M}_\odot \\ -0.1\,\mathrm{s} \leq t_c &\leq 0.1\,\mathrm{s} \\ 100\,\mathrm{Mpc} \leq d_L &\leq 1000\,\mathrm{Mpc} \\ 0 \leq a_i &\leq 0.88 \end{aligned}$$

and standard over the remaining quantities. We take $t_c = 0$ to be the trigger time of GW150914 and constraint $m_1 \geq m_2$.

Waveform generation The simulator generates waveforms using the `IMRphenomPv2` frequency-domain processing model [23, 24] and assumes stationary Gaussian noise with respect to the noise power spectral density (PSD) estimated from $1024\,\mathrm{s}$ of detector data prior to GW150914. The frequency ranges from 20 to $1024\,\mathrm{Hz}$ and each waveform has a duration of $8\,\mathrm{s}$. The waveforms are whitened with respect to the estimated PSD.

Waveform processing The observations are quite large (16 384 features) and, thereby, impractical to store on disk and feed to a neural network. To alleviate this problem, the waveforms are compressed to a reduced-order basis corresponding to the first 128 components of a singular value decomposition (SVD). Using more SVD components did not help producing better predictions, likely due to the higher ratio of noise in less significant components.

Since an observation corresponds to two waveforms from two geographically distant detectors (H1 and L1) and frequency-domain signals are represented by complex-valued vectors, each processed observation is a vector of $128 \times 2 \times 2 = 512$ real-valued numbers.

Noise For the training set, the noise of the detectors is not added to the stored waveforms. Instead, noise realizations are sampled with respect to the PSD in real time during training, which effectively increases the size of the training set.

C Experimental details

Datasets For each simulator, we use three fixed datasets of pairs $(\theta, x) \sim p(\theta, x)$ to train, validate and test AMNRE, respectively. The sizes of the datasets are provided in Table 1.

Table 1: Dataset sizes for each simulator.

Simulator	Training set	Validation set	Testing set
SLCP	1 048 576	131 072	131 072
GW	4 194 304	131 072	131 072

Architectures For SLCP, we use an MLP with 7 hidden layers of 256 neurons and ELU [25] activation functions. For GW, the classifier is a residual residual network [10] consisting of 17 residual blocks of 2 linear layers with 512 neurons and ELU [25] activation functions. In the blocks, we insert batch normalization layers [26] before the activation functions.

Training All networks are optimized with the AdamW [27, 28] stochastic optimization algorithm. At each epoch, the batches are built by sampling without replacement from the training set. The independent parameters θ' are obtained by shifting circularly ($i \leftarrow i + 1$ and $n \leftarrow 1$) the batch of parameters θ. Each element in the batch has a different mask, sampled from the uniform mask distribution.

For SLCP, we apply a "reduce on plateau" scheduling to the learning rate, that is, we divide the learning rate by a factor 2 each time the loss on the validation set has not decreased for 7 consecutive

epochs. The training stops when the learning rate reaches 10^{-6} or lower. For GW, we apply a learning rate cosine annealing [29] over 512 epochs. Other hyperparameters are provided in Table 2.

Table 2: Training hyperparameters.

Hyperparameter	SCLP	GW
Optimizer	AdamW	AdamW
Weight decay	10^{-4}	10^{-4}
Batch size	1024	1024
Batches per epoch	256	1024
Epochs	-	512
Scheduling	reduce on plateau	cosine annealing
Initial learning rate	10^{-3}	2×10^{-4}
Final learning rate	10^{-6}	10^{-6}

D Additional figures

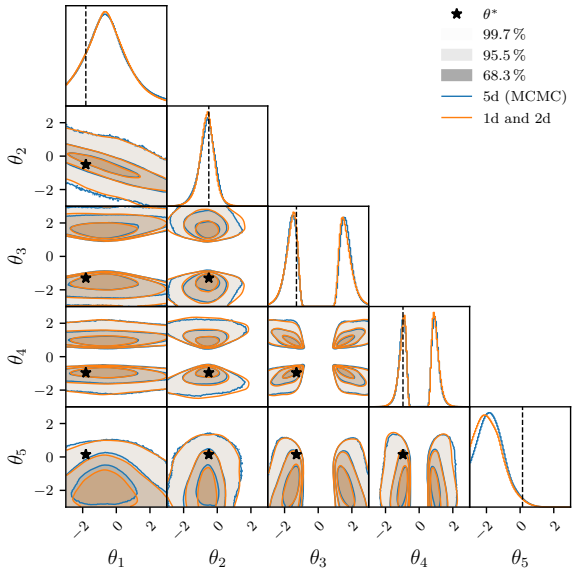

Figure 4: AMNRE full 5d surrogate posterior against 1d and 2d marginal surrogates, for an observation of the SLCP testing set. The predictions of AMNRE for the marginal posteriors are consistent with the predictions for the full posterior, marginalized onto the 1d and 2d subspaces.

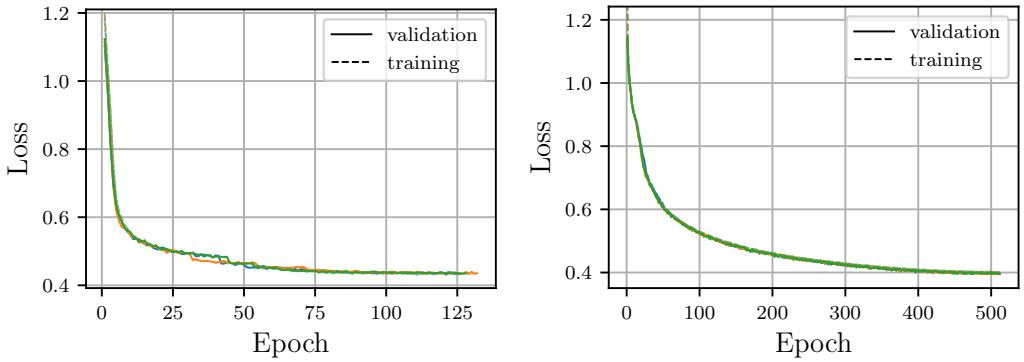

Figure 5: Mean training and validation losses of AMNRE surrogate models for SLCP (left) and GW (right) simulators. Each color corresponds to a different training instance. All instances converge without signs of overfitting. Training takes around 5 minutes for SLCP and 8 hours for GW, on a single 1080Ti GPU.

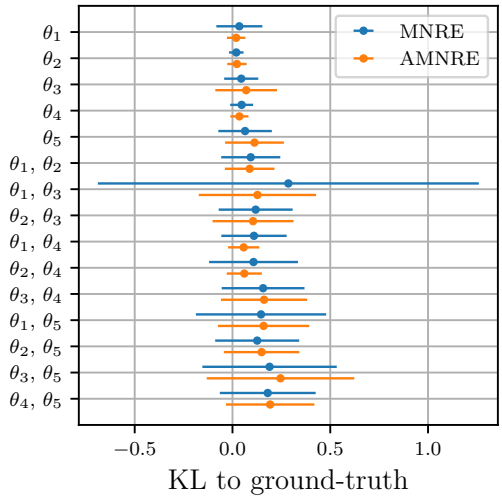

Figure 6: KL divergence to the marginalized ground-truth posterior of 1d and 2d surrogate marginal posterior histograms. The bars represent the mean and standard deviation over 64 observations from the SLCP testing set. AMNRE does not diverges more from the ground-truth than MNRE [3, 7], despite using a single network for all subspaces.

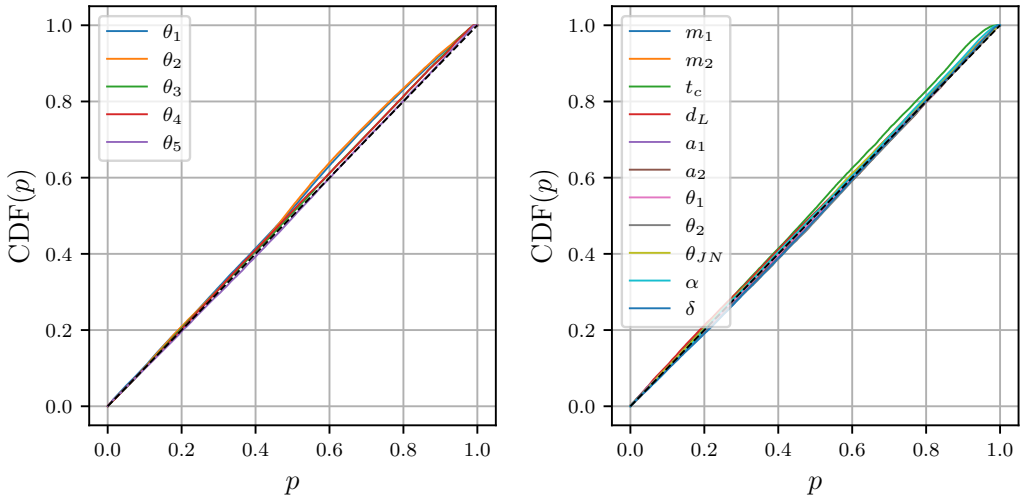

Figure 7: Cumulative distribution function (CDF) of the percentiles p of 8192 parameters θ^* in the one-dimensional surrogate marginal posteriors $\hat{p}(\theta_i|x^*)$ for pairs (θ^*, x^*) of the SLCP (left) and GW (right) testing sets. If the surrogate posterior is consistent with the prior, *i.e.* if $\mathbb{E}_{p(x)}\left[\hat{p}(\theta|x)\right] \approx p(\theta)$, the percentiles should be distributed uniformly between 0 and 1. Since the CDFs lie close to the diagonal, we conclude that the surrogates are consistent with the prior.

Importantly, one *cannot* conclude that the network models properly the posterior from this result, as *any* distribution consistent with the prior, including the prior itself, would present diagonal CDFs. Green et al. [5] inadvertently draw this erroneous conclusion.