Prior Dependence in Neural Ratio Estimation

Nabil Salama[1,2,3] **Katherine Fraser**[1,3] **Po-Wen Chang**[3] **Benjamin Nachman**[4]
[1]University of California, Berkeley
[2]University of Hamburg
[3]Lawrence Berkeley National Laboratory
[4]Stanford University
nabil.salama@studium.uni-hamburg.de

Abstract

Neural ratio estimation is an effective technique for using neural networks to learn likelihood ratios from simulated data, which can be used for applications such as parameter inference and reweighting. While it can be theoretically shown that in principle the resulting likelihood-to-evidence ratio should be independent of the prior used to select the training data, that independence relies on achieving an idealized classifier, which is not possible in practice. It is therefore unclear to what extent prior independence survives under realistic training conditions, particularly for more complicated higher-dimensional problems. In this paper, we test the prior independence of parameter estimation for a simple Gaussian toy example. We plan to extend this study to realistic particle physics datasets in the future.

1 Neural Ratio Estimation

Neural ratio estimation (NRE) [1, 2] is a technique for inference in which likelihood ratios are learned directly from simulated data without an explicit parameterized form for the likelihood. It is advantageous for problems where modeling the likelihood explicitly is intractable. One way to perform neural ratio estimation is to use a parameterized classifier.

In parameterized neural networks, a set of parameters $\boldsymbol{\theta}$ is provided as an additional input to a classifier alongside the features $\boldsymbol{x}$. As the distribution of $\boldsymbol{x}$ depends on $\boldsymbol{\theta}$, the network learns a continuous interpolation over the parameter space, enabling a single classifier to generalize its predictions for the output label y across varying physical hypotheses [3]. Parameterized classifiers are advantageous because they permit modeling the parameter dependence of the data without training separate models for each parameter value. They have been used in a wide range of particle physics applications [4, 5, 6, 7, 8], including the dependence of the Top-quark mass m_t on the Pythia tuning parameters [9].

To evaluate the likelihood-to-evidence ratio $p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{x})$ with a parameterized neural network, one trains a binary classifier that distinguishes between samples drawn from the parameter-dependent joint distribution and those drawn from a parameter-independent data distribution:

$$\begin{aligned} X_0 = (\boldsymbol{x}, \boldsymbol{\theta}) \sim p(\boldsymbol{x})p(\boldsymbol{\theta}), \quad Y_0 = (y = 0) \\ X_1 = (\boldsymbol{x}, \boldsymbol{\theta}) \sim p(\boldsymbol{x}, \boldsymbol{\theta}), \quad Y_1 = (y = 1) \end{aligned} \tag{1}$$

For balanced class priors and an ideal classifier, the "likelihood ratio trick" [10] ensures that the trained classifier output can be expressed in terms of the likelihood ratio.

$$\frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{1 - f(\boldsymbol{x}, \boldsymbol{\theta})} = r(\boldsymbol{x}, \boldsymbol{\theta}) := \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})\, p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}, \tag{2}$$

2026 Conference on Physics and AI (PAI26).

where $f \in [0, 1]$ is the trained neural network output, and we have used Bayes theorem in the last step. From this expression, we see that the classifier output can be used to perform inference by obtaining a distribution over parameters.

While the likelihood-to-evidence ratio is theoretically independent of the prior used to generate the training data, this independence relies on achieving the optimal classifier. In realistic settings, deviations from this ideal behavior can introduce a dependence on the training prior, and such effects have been observed in some cases [11, 12]. In this work, we explore how the choice of parameter prior $p(\boldsymbol{\theta})$ influences the likelihood ratio through its impact on parameter estimation. In particular, we investigate the extent to which the prior used to generate training data affects the efficacy of parameter estimation in practice using a simple Gaussian toy example, a common benchmark in simulation-based inference studies [13].

2 Analysis Setup

We use NRE to infer the true $\boldsymbol{\theta}$ values that describe the means of the distributions from the data samples. To evaluate how the inferred values of $\boldsymbol{\theta}$ (and hence the inferred likelihood-to-evidence ratio) depend on the choice of prior for $\boldsymbol{\theta}$, we generate d-dimensional data samples $\boldsymbol{x}$ (with $d \in [1, 6]$) from multidimensional Gaussians $\mathcal{N}(\boldsymbol{\theta}, \sigma_d = 2)$. The parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ are drawn from four distinct prior distributions $p_i(\boldsymbol{\theta})$ normalized over $[0, 10]^d$: a uniform prior, a normal prior $\mathcal{N}(5, 4)$, an exponential prior with $\lambda = -0.1$, and a discrete grid with spacing $h = 0.2$.

For each prior distribution, the training samples for the data-independent class Y_0 were drawn according to $X_0 = (\boldsymbol{x}_0, \boldsymbol{\theta}_a) \sim p_i(\boldsymbol{x})\, p_i(\boldsymbol{\theta})$, and similarly for the data-dependent class Y_1 with $X_1 = (\boldsymbol{x}_1, \boldsymbol{\theta}_c) \sim p_i(\boldsymbol{x}, \boldsymbol{\theta})$ from distributions

$$\boldsymbol{\theta}_a, \boldsymbol{\theta}_b, \boldsymbol{\theta}_c \overset{\text{iid}}{\sim} p_i(\boldsymbol{\theta}), \qquad \boldsymbol{x}_0 \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_b), \qquad \boldsymbol{x}_1 \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_c). \tag{3}$$

To improve robustness, we compute the likelihood-to-evidence ratio using 15 independently drawn samples for each prior, and combine them according to:

$$r_i(\boldsymbol{\theta} \mid \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}) = \frac{\prod_{k=1}^{m} r_i(\boldsymbol{x}^{(k)}, \boldsymbol{\theta})}{\int_\Omega \prod_{k=1}^{m} r_i(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}')\, d\boldsymbol{\theta}'}. \tag{4}$$

The classifier used for ratio estimation is a small multi-layer perceptron architecture with input $(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathbb{R}^{2d}$, 4 dense layers with 64 units each and ReLU activation, trained in a supervised manner for 5 epochs for every prior configuration. The output is a single sigmoid-activated unit and BCE loss. The dataset is as described around Eq. 3 with a total sample size of 10^5 and a training/test/validation split of 70% / 15% / 15%. To test the classifiers' training this way, the parameter θ is estimated using Eq. 2, by evaluating each of the trained classifiers with 15 unique samples for 25^d uniformly spaced values of $\boldsymbol{\theta}$.

3 Results

The performance of NRE in approximating the likelihood can be evaluated through its accuracy in parameter inference and reweighting. In this work, we restrict our study to parameter inference for brevity. We judge the quality of inference using the maximum likelihood ratio estimate (MLE) and highest likelihood ratio density (HLD) intervals, where the latter can be interpreted as the uncertainty on the former. Figure 1 illustrates how these quantities are obtained. Mathematically, the MLE $\hat{\boldsymbol{\theta}}_i$ for prior i is defined as the argument maximizing the likelihood ratio; and the bias $b_{i,j}$ in dimension axis $j \in [1, d]$ is defined as the distance between $\hat{\boldsymbol{\theta}}_i$ and the true parameter $\theta_{i,j}$:

$$\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta} \in \Omega} r_i(\boldsymbol{\theta} \mid \mathbf{X}_i), \qquad b_{i,j} = \hat{\theta}_{i,j} - \theta_{i,j} \tag{5}$$

with $\mathbf{X}_i = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\right)$. The α-HLD $\mathcal{H}_i^{(\alpha)}$ is defined as the interval covering the fraction α of the likelihood ratio surrounding the maximum:

$$\mathcal{H}_i^{(\alpha)} = \left\{\boldsymbol{\theta} \in \Omega : r_i(\boldsymbol{\theta} \mid \mathbf{X}_i) \geq c_i^{(\alpha)}\right\}, \qquad c_i^{(\alpha)} \text{ s.t.} \int_{\mathcal{H}_i^{(\alpha)}} r_i(\boldsymbol{\theta} \mid \mathbf{X}_i)\, d\boldsymbol{\theta} = \alpha\,. \tag{6}$$

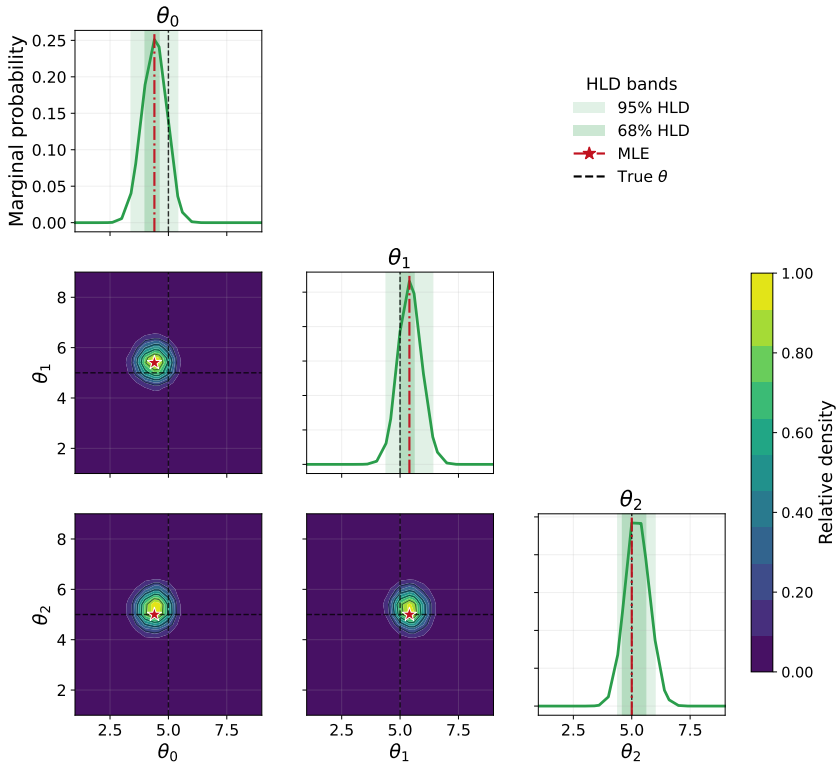

Figure 1: One and two-dimensional projections of the three-dimensional likelihood ratio, with training data sampled from an exponential prior around the true inference point $\boldsymbol{\theta} = (5.0, 5.0, 5.0)$.

From here, the HLD interval width $\Delta_{i,j}^{(\alpha)}$ along dimension axis j is defined as the distance of the interval boundaries:

$$\Delta_{i,j}^{(\alpha)} = U_{i,j}^{(\alpha)} - L_{i,j}^{(\alpha)}, \qquad L_{i,j}^{(\alpha)} = \min_{\boldsymbol{\theta} \in \mathcal{H}_i^{(\alpha)}} \theta_j, \qquad U_{i,j}^{(\alpha)} = \max_{\boldsymbol{\theta} \in \mathcal{H}_i^{(\alpha)}} \theta_j\,. \tag{7}$$

The bias and HLD intervals are shown in figure 2, which compares the inferred parameters to the true values for four different training priors. The consistent overlap of the HLD intervals with the true parameter across all four panels indicates that parameter inference—and thus the likelihood ratio—is robust to the choice of prior in this simple one-dimensional setting.

We can also define the coverage of a given HLD interval as the fraction of estimated values $\hat{\theta}_{i,j}$ that lie within that interval. This coverage, as well as width and bias, are calculated dimension-axis-wise (j) for all l^d inference cases for each prior and all $|D|$ dimensionalities. Averaging across the former two yields the measures of mean coverage, width, and bias, which we use for $d > 1$:

$$\bar{C}_i^{(\alpha)} = \frac{1}{dl^d} \sum_{k=1}^{l^d} \sum_{j=0}^{d} C_{i,j,k}^{(\alpha)} \qquad \bar{\Delta}_i^{(\alpha)} = \frac{1}{dl^d} \sum_{k=1}^{l^d} \sum_{j=0}^{d} \Delta_{i,j,k}^{(\alpha)} \qquad \bar{b}_i = \frac{1}{dl^d} \sum_{k=1}^{l^d} \sum_{j=0}^{d} b_{i,j,k}\,. \tag{8}$$

These metrics are shown in figure 3 for dimensions 1 through 6. From this figure we observe that differences in coverage, width, and bias across priors are minimal, while performance depends on dimensionality. The grid prior delivers slightly improved performance compared to the others.

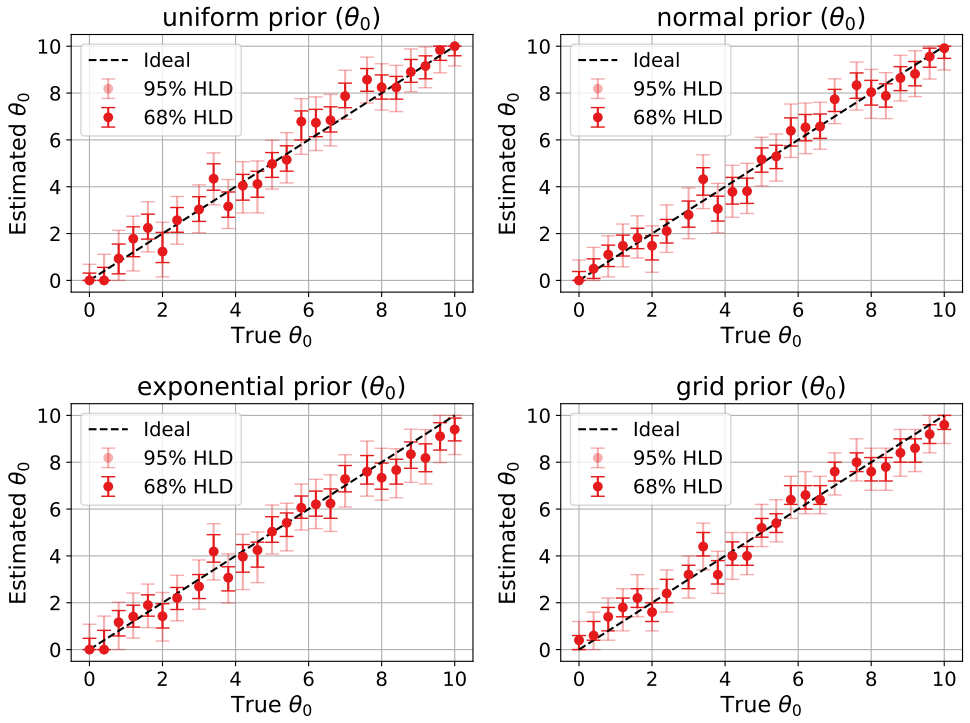

Figure 2: Parameter estimates and HLD intervals inferred from parameterized classifiers with different training data priors. Each panel compares the true parameter with the 68% and 95% HLD intervals for a given prior.

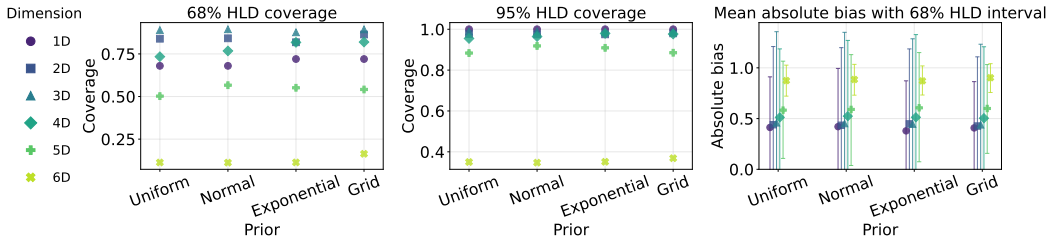

Figure 3: Summary of metrics for different prior choices and data dimensionality. The left and middle panels show the mean coverage of 68% and 95% intervals. The right panel shows the mean absolute bias and 68% HLD width. Differences between priors are insignificant compared to differences in dimensionality.

4 Conclusion

This work examines the robustness of the theoretical insight that neural ratio estimation using parameterized neural networks should be prior-independent. Specifically, we consider Gaussian datasets in different dimensions and compare the choice of various priors for each. We find only minimal dependence on the training prior in this set of toy examples.

While approximate prior-independence is observed in this simple example, several extensions merit further study. In particular, a comparison based on reweighting performance will be presented in a longer forthcoming version of this work. It is also unclear whether this behavior persists in more complex settings, motivating similar studies on more realistic datasets and with more complex NRE algorithms which have been developed to mitigate some of these challenges [14]. Should significant deviations arise, it would be worthwhile to explore the design of an optimal training prior specific to realistic particle physics problems.

References

[1] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

[2] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proc. Nat. Acad. Sci.*, 117(48):30055–30062, 2020.

[3] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5):235, 2016.

[4] Anders Andreassen and Benjamin Nachman. Neural networks for full phase-space reweighting and parameter tuning. *Physical Review D*, 101(9):091901, 2020.

[5] Benjamin Nachman and Jesse Thaler. E pluribus unum ex machina: Learning from many collider events at once. *Physical Review D*, 103(11):116013, 2021.

[6] Arindam Bhattacharya, Katherine Fraser, and Matthew D. Schwartz. Observable optimization for precision theory: machine learning energy correlators. *JHEP*, 01:151, 2026.

[7] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4239–4248. PMLR, 2020.

[8] Benjamin Kurt Miller, Christoph Weniger, and Patrick Forré. Contrastive neural ratio estimation. In Sanmi Koyejo, Shakir Mohamed, Anima Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh, editors, *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 3262–3278. Curran Associates, Inc., 2022.

[9] Forrest Flesher, Katherine Fraser, Charles Hutchison, Bryan Ostdiek, and Matthew D. Schwartz. Parameter inference from event ensembles and the top-quark mass. *Journal of High Energy Physics*, 2021(9):58, 2021.

[10] Shahzar Rizvi, Mariel Pettee, and Benjamin Nachman. Learning likelihood ratios with neural network classifiers. *Journal of High Energy Physics*, 2024(02):136, 2024.

[11] Benjamin Kurt Miller, Alex Cole, Patrick Forré, Gilles Louppe, and Christoph Weniger. Truncated Marginal Neural Ratio Estimation. In *35th Conference on Neural Information Processing Systems*, 7 2021.

[12] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful. 10 2021.

[13] Jan-Matthis Lueckmann, Jan Boelts, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. Benchmarking simulation-based inference, 2021.

[14] Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation. 8 2022.