Likelihood-free inference with an improved cross-entropy estimator

Markus Stoye,[1] **Johann Brehmer,**[2] **Gilles Louppe,**[3] **Juan Pavez,**[4] **and Kyle Cranmer**[2]
[1] Reexen [2] New York University [3] University of Liège [4] Federico Santa María Technical University
markus.stoye@reexen.com, johann.brehmer@nyu.edu, g.louppe@uliege.be,
juan.pavezs@alumnos.usm.cl, kyle.cranmer@nyu.edu

Abstract

Phenomena in many areas of science are often modeled by complex computer simulations that do not have a tractable likelihood function. Such implicit models provide a major challenge for inference. Recently, techniques for likelihood-free inference have been developed in which neural networks are trained as surrogate models using the joint likelihood ratio and joint score as training data. These quantities characterize the latent process of the simulator and can often be extracted from its runs. We extend this approach and show how this augmented training data can be used to provide a new, lower-variance cross-entropy estimator. In a real-life particle-physics example we demonstrate that this new loss function leads to an improved sample efficiency compared to previous methods.

1 Introduction

Many real-world phenomena are best described by computer simulations. Such simulators often implement a stochastic generative process, which is based on a mechanistic model and parametrized by θ. In practice, these simulators are used to generate samples of observations $x \sim p(x|\theta)$, but the density is only defined implicitly through the simulation code. Often, the generative process involves latent variables and the density

$$p(x|\theta) = \int \mathrm{d}z \; p(x, z|\theta) \tag{1}$$

is intractable because of the integral over a large (and possibly highly structured) latent space. Without a tractable likelihood, statistical inference on the parameters θ given observed data x is challenging. This problem has prompted the development of *likelihood-free inference* methods such as Approximate Bayesian Computation [1–4] and neural density or neural density ratio estimation algorithms [5–23]. Nearly all of these established methods treat the simulator as a black box and only use its capability to generate samples for a specified values of θ.

In Refs. [24–26] a new paradigm was introduced that exploits additional information that can be extracted from the simulation. In particular, within the simulation where the latent variables z are available, it is often possible to extract the *joint likelihood ratio* $r(x, z)$ and the *joint score* $t(x, z)$,

$$r(x, z|\theta_0, \theta_1) = \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)} \quad \text{and} \quad t(x, z|\theta_0) = \nabla_\theta \log p(x, z|\theta)\Big|_{\theta_0}, \tag{2}$$

which are dependent on the latent variables z corresponding to a particular sample.

It was then shown that certain loss functionals that use the joint likelihood ratio and the joint score are minimized by the likelihood ratio $r(x|\theta_0, \theta_1)$, an otherwise intractable quantity. This motivates a family of new techniques for likelihood-free inference in which the joint likelihood ratio and joint

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

score are used as training data for neural networks. These networks serve as surrogate models for the intractable likelihood or likelihood ratio. Experiments showed these new methods to be more sample-efficient than previously established neural density and neural density ratio estimation techniques. The authors of Refs. [24–26] coined the term "mining gold" for the process of extracting the joint likelihood ratio and joint score from the simulator – while the augmented data require some effort to extract, they are extremely valuable.

While the loss functionals originally proposed in Refs. [24–26] have the correct minima, they are not necessarily the most sample efficient. In particular, the proposed losses are based on the mean squared error (MSE) between the network prediction and the joint likelihood ratio or joint score, which is often dominated by few samples for which these quantities are large. Here we extend and improve that original work with two new algorithms for likelihood-free inference. The key improvement are two new loss functions, which use an improved estimator for the cross entropy based on the joint likelihood ratio and joint score. After introducing these new algorithm in Sec. 2, we show its performance in a problem from particle physics in Sec. 3, before giving our conclusions in Sec. 4.

2 Cross-entropy estimation with augmented data

Consider the problem of estimating the likelihood ratio $r(x|\theta_0,\theta_1)$ given two balanced samples: $(x_i,z_i) \sim p(x,z|\theta_0)$, labeled with $y_i = 0$, and $(x_i,z_i) \sim p(x,z|\theta_1)$, labeled $y_i = 1$. For each simulated event we assume that the joint likelihood ratio $r(x_i,z_i|\theta_0,\theta_1)$ and joint score $t(x_i,z_i|\theta_0)$ are also available.

The familiar binary cross-entropy loss functional is defined as

$$L[g(x)] = -\int \mathrm{d}x \, \Big[p(x|y=1)\,\log(g(x)) + p(x|y=0)\,\log(1-g(x))\Big] \,. \tag{3}$$

Typically and without using the joint likelihood ratio, the two terms are sampled separately, defining a high-variance estimator of the cross entropy. But in the scenario where we have access to the joint likelihood ratio, we can rewrite this as

$$\begin{aligned} L[g(x)] &= -\int \mathrm{d}x \; p(x,z)\Big[s(x_i,z_i|\theta_0,\theta_1)\log(g(x)) + (1-s(x,z|\theta_0,\theta_1))\log(1-g(x))\Big] \\ &\approx -\frac{1}{N}\sum_{(x_i,z_i)\sim p(x_i,z_i)} \Big[s(x_i,z_i|\theta_0,\theta_1)\log(g(x_i)) + (1-s(x_i,z_i|\theta_0,\theta_1))\log(1-g(x_i))\Big] \\ &=: L_{\text{ALICE}}[g(x)] \end{aligned} \tag{4}$$

where we define $p(x,z) \equiv {}^1\!/_2(p(x,z|\theta_0) + p(x,z|\theta_1))$ and $s(x,z|\theta_0,\theta_1) = [r(x,z|\theta_0,\theta_1)+1]^{-1}$.

This improved cross-entropy estimator uses the exact $s(x,z|\theta_0,\theta_1)$ in place of the class label $y_i \in \{0,1\}$, thus reducing the variance. In this way the samples drawn according to $y=0$ also provide information about the second $y=1$ term in the loss function, and vice versa. By minimizing the loss function we get an estimator $\hat{s}(x) = \arg\min_g L_{\text{ALICE}}$, which can directly be translated to an estimator for the likelihood ratio

$$\hat{r}(x|\theta_0,\theta_1) = \frac{1-\hat{s}(x)}{\hat{s}(x)} \,. \tag{5}$$

This defines the ALICE inference method[1], which consists of mining the joint likelihood ratio from the simulator, training a neural network on the improved cross-entropy estimator in Eq. (4), and using this surrogate model for statistical inference on θ.

It is straightforward to show that the minimum of L_{ALICE} in the limit of the exact integral (or infinite samples) corresponds to the true likelihood ratio function $r(x|\theta_0,\theta_1)$. The interesting question is how it performs with finite samples. It is to be expected that a likelihood ratio estimator based on the ALICE estimator for the cross-entropy should perform as least as well as an estimator based on the standard cross-entropy estimator in Eq. (3) (the "CARL" technique).

In analogy to the CASCAL and RASCAL methods of Refs. [24–26], we can define an additional inference method which uses the joint score, i. e. an additional piece of information that describes the

[1] **A**pproximate **l**ikelihood with **i**mproved **c**ross-entropy **e**stimator

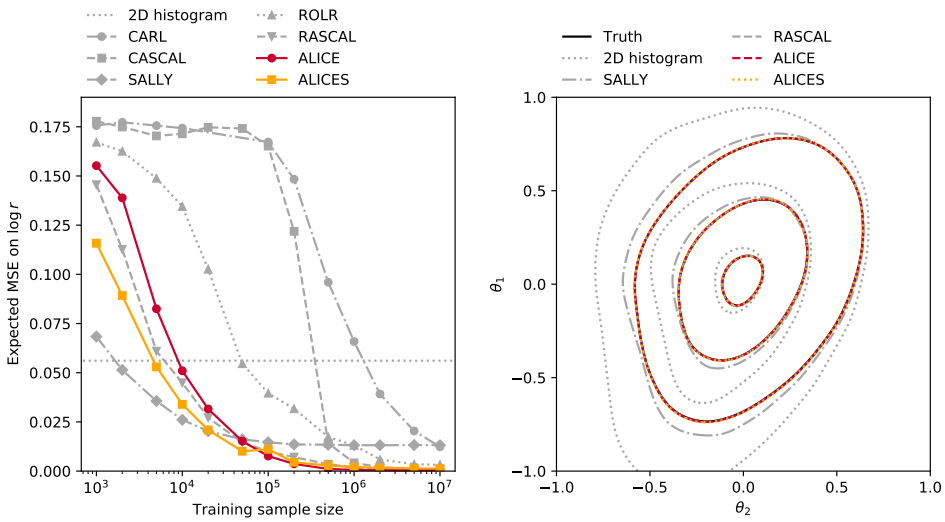

Figure 1: Left: Estimator fidelity as a function of the training sample size. As a metric we use the expected mean squared error on the log likelihood ratio, see Ref. [25]. The new methods (red, orange lines) are more sample efficient than the similar ROLR and RASCAL techniques. Right: Corresponding expected exclusion limits, assuming 36 events distributed according to $\theta = (0, 0)^T$ and based on a large training set with 10^7 samples. We find an excellent performance of the ALICE and ALICES methods, virtually indistinguishable from the RASCAL method and the true likelihood ratio.

local (tangential) behavior of the likelihood function. If a parameterized likelihood ratio estimator is implemented with a differentiable architecture such as a neural network, we can calculate the gradient of the output $\hat{s}(x|\theta_0, \theta_1)$ with respect to θ_0 and similarly calculate the corresponding score

$$\hat{t}(x|\theta_0, \theta_1) = \nabla_\theta \log \hat{r}(x|\theta_0, \theta_1) = \nabla_\theta \log \left(\frac{1 - \hat{s}(x_i|\theta, \theta_1)}{\hat{s}(x_i|\theta, \theta_1)} \right) \tag{6}$$

of the $\hat{r}$ estimator. For a perfect $\hat{r}$ (or equivalently $\hat{s}$) estimator, this corresponding score $\hat{t}$ will also minimize the squared error loss with respect to the joint score $t(x, z|\theta_0, \theta_1)$, which can be extracted from the simulator [24–26]. Turning this argument around, we can use the joint score to guide the training of the estimator. This is the idea behind the ALICES[2] technique, which is based on the loss function

$$L_{\text{ALICES}}[g] = L_{\text{ALICE}}[g] - \frac{1}{N} \sum_{(x_i,z_i)\sim p(x_i,z_i)} \left[\alpha \, (1-y_i) \left| t(x_i, z_i|\theta_0, \theta_1) - \nabla_\theta \log \left(\frac{1 - g(x_i|\theta, \theta_1)}{g(x_i|\theta, \theta_1)} \right) \right|_{\theta_0} \right|^2 \Bigg] . \tag{7}$$

The factor $(1 - y_i)$ is necessary to guarantee the correct minimum of the squared error on the score. The hyper-parameter α weights the two terms in the loss function. This loss is the natural extension of the the CASCAL loss function, but we expect it to reduce the variance compared to the CASCAL approach for finite sample size.

3 Experiments

We experiment with the new methods in the particle physics problem introduced in Refs. [24, 25]. In this real-world problem, the outcome of proton-proton collisions is characterized by 42 observables, from which likelihood ratios and confidence limits on two model parameters are derived. We first consider an idealized setting neglecting the detector response where the likelihood function is tractable, which provides us with ground truth that can be used to evaluate the performance of the algorithms. For a detailed description of the setup, see Ref. [25].

[2] **A**pproximate **l**ikelihood with **i**mproved **c**ross-entropy **e**stimator and **s**core

We compare the new ALICE and ALICES methods to the similar CARL, ROLR, CASCAL, and RASCAL techniques introduced in Refs. [24, 25] as well as to the SALLY and SALLINO methods. SALLY and SALLINO approximate a statistical model that is accurate in the neighborhood of $\theta = (0, 0)^T$. The methods are very sample efficient, but make approximations that limit their asymptotic performance.

Except for the new loss functions, we used the same architectures and hyper-parameters as in Ref. [25]. In particular, we use fully connected networks with five hidden layers, 100 units each, and $\tanh$ activation functions for both approaches. For ALICES we use $\alpha = 5$, which was found to give a good performance for the closely related CASCAL method [25].

The left panel in Fig. 1 shows the quality of the likelihood ratio estimate based on various sized training samples for the new methods and compares them to the inference techniques presented in Ref. [25]. As a performance metric we use an expected mean squared error on the log likelihood ratio, as defined in Ref. [25]. Unsurprisingly, the ALICE and ROLR methods clearly outperform CARL, which does not have access to the joint likelihood ratio. More significantly, we find that ALICE outperforms ROLR, which does have access to the joint likelihood ratio. We conjecture that this improvement can be attributed to the lower variance of the cross-entropy compared to the squared error. More surprisingly, the ALICE method also outperforms the RASCAL method for larger training sample sizes ($\geq 10^5$), even though ALICE does not have access to the joint score.

For smaller training sample sizes ($\leq 10^5$) the ALICES method outperforms the ALICE method, which is not surprising given the additional information available during training. For larger training sample sizes ($\geq 10^5$), the variance of the score actually deteriorates the performance of ALICES compared to ALICE. We did not perform hyper-parameter tuning for α as a function of the training sample size, which should ensure that ALICES performs at least as well as ALICE. We leave a systematic tuning of the α parameter and an analysis of sources of variance in this approach for future work.

The right panel in Fig. 1 shows how this performance translates to high-quality inference results that are virtually indistinguishable from the ground-truth limits and substantially better than the baseline histogram analysis.

4 Conclusions

In this work, we have extended recently developed inference techniques for the setting in which the likelihood is only implicitly defined through a stochastic generative model or simulator. By exploiting the joint likelihood ratio that can be extracted from the simulator, we introduced an improved cross-entropy estimator. This improved cross-entropy estimator is used to define two new likelihood-free inference techniques: ALICE and ALICES.

Our experiments comparing ALICE and ALICES with the other recently developed techniques indicate that they are significantly more sample efficient than ROLR, CASCAL, and RASCAL techniques. We attribute this to the lower variance of the improved cross-entropy estimator. For smaller training sample sizes, there are still advantages to the SALLY and SALLINO techniques.

We note that it is possible to use a hybrid of the traditional cross-entropy of Eq. 3 and the improved cross-entropy Eq. 4. This would be useful in situations where one may not have access to the joint ratio for practical reasons or because some training samples come from real data instead of a simulation.

The ubiquity of simulators and other implicit models indicates there is enormous potential for likelihood-free inference techniques. The use of augmented data improves the sample efficiency of these techniques significantly, and these results motivate further study of variance reduction techniques that leverage this augmented data.

Acknowledgments

JB, KC, and GL are grateful for the support of the Moore-Sloan data science environment at NYU. KC and GL were supported through the NSF grants ACI-1450310 and PHY-1505463. JP was partially supported by the Scientific and Technological Center of Valparaíso (CCTVal) under Fondecyt grant BASAL FB0821. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] D. B. Rubin: 'Bayesianly justifiable and relevant frequency calculations for the applied statistician'. Ann. Statist. 12 (4), p. 1151, 1984. URL https://doi.org/10.1214/aos/1176346785.

[2] M. A. Beaumont, W. Zhang, and D. J. Balding: 'Approximate bayesian computation in population genetics'. Genetics 162 (4), p. 2025, 2002.

[3] J. Alsing, B. Wandelt, and S. Feeney: 'Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology' , 2018. arXiv:1801.01497.

[4] T. Charnock, G. Lavaux, and B. D. Wandelt: 'Automatic physical inference with information maximizing neural networks'. Phys. Rev. D97 (8), p. 083004, 2018. arXiv:1802.03537.

[5] T. Kanamori, S. Hido, and M. Sugiyama: 'A least-squares approach to direct importance estimation'. Journal of Machine Learning Research 10 (Jul), p. 1391, 2009.

[6] Y. Fan, D. J. Nott, and S. A. Sisson: 'Approximate Bayesian Computation via Regression Density Estimation'. ArXiv e-prints , 2012. arXiv:1212.1479.

[7] L. Dinh, D. Krueger, and Y. Bengio: 'NICE: Non-linear Independent Components Estimation'. ArXiv e-prints , 2014. arXiv:1410.8516.

[8] D. Jimenez Rezende and S. Mohamed: 'Variational Inference with Normalizing Flows'. ArXiv e-prints , 2015. arXiv:1505.05770.

[9] K. Cranmer, J. Pavez, and G. Louppe: 'Approximating Likelihood Ratios with Calibrated Discriminative Classifiers' , 2015. arXiv:1506.02169.

[10] K. Cranmer and G. Louppe: 'Unifying generative models and exact likelihood-free inference with conditional bijections'. J. Brief Ideas , 2016.

[11] L. Dinh, J. Sohl-Dickstein, and S. Bengio: 'Density estimation using Real NVP'. ArXiv e-prints , 2016. arXiv:1605.08803.

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

[13] B. Paige and F. Wood: 'Inference Networks for Sequential Monte Carlo in Graphical Models'. ArXiv e-prints , 2016. arXiv:1602.06701.

[14] R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann: 'Likelihood-free inference by ratio estimation'. ArXiv e-prints , 2016. arXiv:1611.10242.

[15] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle: 'Neural Autoregressive Distribution Estimation'. ArXiv e-prints , 2016. arXiv:1605.02226.

[16] A. van den Oord, S. Dieleman, H. Zen, et al.: 'WaveNet: A Generative Model for Raw Audio'. ArXiv e-prints , 2016. arXiv:1609.03499.

[17] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu: 'Conditional Image Generation with PixelCNN Decoders'. ArXiv e-prints , 2016. arXiv:1606.05328.

[18] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu: 'Pixel Recurrent Neural Networks'. ArXiv e-prints , 2016. arXiv:1601.06759.

[19] M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander: 'Likelihood-free inference via classification'. Statistics and Computing p. 1–15, 2017.

[20] D. Tran, R. Ranganath, and D. Blei: 'Hierarchical implicit models and likelihood-free variational inference'. In I. Guyon, U. V. Luxburg, S. Bengio, et al. (eds.), 'Advances in Neural Information Processing Systems 30', p. 5523–5533, 2017.

[21] G. Louppe and K. Cranmer: 'Adversarial Variational Optimization of Non-Differentiable Simulators'. ArXiv e-prints , 2017. arXiv:1707.07113.

[22] G. Papamakarios, T. Pavlakou, and I. Murray: 'Masked Autoregressive Flow for Density Estimation'. ArXiv e-prints , 2017. arXiv:1705.07057.

[23] G. Papamakarios, D. C. Sterratt, and I. Murray: 'Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows'. ArXiv e-prints , 2018. arXiv:1805.07226.

[24] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez: 'Constraining Effective Field Theories with Machine Learning'. Phys. Rev. Lett. 121 (11), p. 111801, 2018. arXiv:1805.00013.

[25] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez: 'A Guide to Constraining Effective Field Theories with Machine Learning'. Phys. Rev. D98 (5), p. 052004, 2018. arXiv:1805.00020.

[26] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer: 'Mining gold from implicit models to improve likelihood-free inference' , 2018. arXiv:1805.12244.