Likelihood-free MCMC with Amortized Approximate Ratio Estimators

Joeri Hermans [1] **Volodimir Begy** [2] **Gilles Louppe** [1]

Abstract

Posterior inference with an intractable likelihood is becoming an increasingly common task in scientific domains which rely on sophisticated computer simulations. Typically, these forward models do not admit tractable densities forcing practitioners to make use of approximations. This work introduces a novel approach to address the intractability of the likelihood and the marginal model. We achieve this by learning a flexible amortized estimator which approximates the likelihood-to-evidence ratio. We demonstrate that the learned ratio estimator can be embedded in MCMC samplers to approximate likelihood-ratios between consecutive states in the Markov chain, allowing us to draw samples from the intractable posterior. Techniques are presented to improve the numerical stability and to measure the quality of an approximation. The accuracy of our approach is demonstrated on a variety of benchmarks against well-established techniques. Scientific applications in physics show its applicability.

1. Introduction

Domain scientists are generally interested in the posterior

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\boldsymbol{\theta})p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} \tag{1}$$

which relates the parameters $\boldsymbol{\theta}$ of a model or theory to observations $\mathbf{x}$. Although Bayesian inference is an ideal tool for such settings, the implied computation is generally not. Often the marginal model $p(\mathbf{x}) = \int p(\boldsymbol{\theta})p(\mathbf{x} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}$ is intractable, making posterior inference using Bayes' rule impractical. Methods such as Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970) bypass the dependency on the marginal model by evaluating the ratio of posterior densities between consecutive states in the Markov chain. This allows the posterior to be approximated numerically, provided that the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$ are tractable. We consider the equally common and more challenging setting, the so-called likelihood-free setup, in which the likelihood cannot be evaluated in a reasonable amount of time or has no tractable closed-form expression. However, drawing samples from the forward model is possible.

Contributions We introduce a Bayesian inference algorithm for scientific applications where (i) a forward model is available, (ii) the likelihood is intractable, and (iii) accurate approximations are important to do science. Central to this work is a novel amortized likelihood-to-evidence ratio estimator which allows for the direct estimation of the posterior density function for arbitrary model parameters $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and observations $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$. We exploit this ability to amortize the estimation of acceptance ratios in MCMC, enabling us to draw posterior samples. Finally, we develop a necessary diagnostic to probe the quality of the approximations in intractable settings.

2. Background

2.1. Markov chain Monte Carlo

MCMC methods are generally applied to sample from a posterior probability distribution with an intractable marginal model, but for which point-wise evaluations of the likelihood are possible (Metropolis et al., 1953; Hastings, 1970; MacKay, 2003). Posterior samples are drawn from the target distribution by collecting dependent states $\boldsymbol{\theta}_{0:T}$ of a Markov chain. The mechanism for transitioning from $\boldsymbol{\theta}_t$ to the next state $\boldsymbol{\theta}'$ depends on the algorithm at hand. However, the acceptance of a transition $\boldsymbol{\theta}_t \rightarrow \boldsymbol{\theta}'$, for $\boldsymbol{\theta}'$ sampled from a proposal mechanism $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t)$, is usually determined by evaluating some form of the posterior ratio

$$\frac{p(\boldsymbol{\theta}' \mid \mathbf{x})}{p(\boldsymbol{\theta}_t \mid \mathbf{x})} = \frac{p(\boldsymbol{\theta}')p(\mathbf{x} \mid \boldsymbol{\theta}') \,/\, p(\mathbf{x})}{p(\boldsymbol{\theta}_t)p(\mathbf{x} \mid \boldsymbol{\theta}_t) \,/\, p(\mathbf{x})} = \frac{p(\boldsymbol{\theta}')p(\mathbf{x} \mid \boldsymbol{\theta}')}{p(\boldsymbol{\theta}_t)p(\mathbf{x} \mid \boldsymbol{\theta}_t)}. \tag{2}$$

We observe that (i) the normalizing constant $p(\mathbf{x})$ cancels out within the ratio, thereby bypassing its intractable evaluation, and (ii) how the likelihood ratio is central in assessing the quality of a candidate state $\boldsymbol{\theta}'$ against state $\boldsymbol{\theta}_t$.

[1]University of Liège, Belgium [2]University of Vienna, Austria. Correspondence to: Joeri Hermans <joeri.hermans@doct.uliege.be>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Metropolis-Hastings Metropolis-Hastings (MH) (Metropolis et al., 1953; Hastings, 1970) is a straightforward implementation of Equation 2 in which the proposal mechanism $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t)$ is typically a tractable distribution. These components are combined to compute the acceptance probability ρ of a transition $\boldsymbol{\theta}_t \to \boldsymbol{\theta}'$:

$$\rho = \min\left(1,\ \frac{p(\boldsymbol{\theta}')p(\mathbf{x} \mid \boldsymbol{\theta}')}{p(\boldsymbol{\theta}_t)p(\mathbf{x} \mid \boldsymbol{\theta}_t)} \frac{q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}')}\right). \tag{3}$$

The choice of an appropriate transition distribution is important to maximize the effective sample size (sampling efficiency) and to reduce the autocorrelation.

Hamiltonian Monte Carlo Hamiltonian Monte Carlo (HMC) (Neal, 2011; Duane et al., 1987; Betancourt, 2017) improves upon the sampling efficiency of Metropolis-Hastings by reducing the autocorrelation of the Markov chain. This is achieved by modeling the density $p(\mathbf{x} \mid \boldsymbol{\theta})$ as a potential energy function

$$U(\boldsymbol{\theta}) \triangleq -\log p(\mathbf{x} \mid \boldsymbol{\theta}), \tag{4}$$

and attributing some kinetic energy,

$$K(\mathbf{m}) \triangleq \frac{1}{2}\mathbf{m}^2 \tag{5}$$

with momentum $\mathbf{m} \sim p(\mathbf{m})$ to the current state $\boldsymbol{\theta}_t$. A new state $\boldsymbol{\theta}'$ can be proposed by simulating the Hamiltonian dynamics of $\boldsymbol{\theta}_t$. This is achieved by leapfrog integration of $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ over a fixed number of steps with initial momentum $\mathbf{m}$. Afterwards, the acceptance ratio

$$\min\left(1,\ \exp\left(U(\boldsymbol{\theta}') - U(\boldsymbol{\theta}_t) + K(\mathbf{m}') - K(\mathbf{m})\right)\right) \tag{6}$$

is computed to assess the quality of the candidate state $\boldsymbol{\theta}'$.

2.2. Approximate likelihood ratios

The most powerful test-statistic to compare two hypotheses $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ for an observation $\mathbf{x}$ is the likelihood ratio (J. Neyman, 1933)

$$r(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \triangleq \frac{p(\mathbf{x} \mid \boldsymbol{\theta}_0)}{p(\mathbf{x} \mid \boldsymbol{\theta}_1)}. \tag{7}$$

Cranmer et al. (2015) have shown that it is possible to express the test-statistic through a change of variables $\mathbf{d}(\cdot)\colon \mathbb{R}^d \mapsto [0, 1]$. This observation can be used in a supervised learning setting to train a classifier $\mathbf{d}(\mathbf{x})$ to distinguish samples $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}_0)$ with class label $y = 1$ from $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}_1)$ with class label $y = 0$. The decision function modeled by the optimal classifier $\mathbf{d}^*(\mathbf{x})$ is in this case

$$\mathbf{d}^*(\mathbf{x}) = p(y = 1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta}_0)}{p(\mathbf{x} \mid \boldsymbol{\theta}_0) + p(\mathbf{x} \mid \boldsymbol{\theta}_1)}, \tag{8}$$

thereby obtaining the likelihood ratio as

$$r(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{\mathbf{d}^*(\mathbf{x})}{1 - \mathbf{d}^*(\mathbf{x})}. \tag{9}$$

In the literature, this is known as the likelihood ratio trick (LRT) (Cranmer et al., 2015; Mohamed & Lakshminarayanan, 2016; Gutmann et al., 2017; Dutta et al., 2016; Tran et al., 2017; Brehmer et al., 2020) and is especially prominent in the area of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Uehara et al., 2016; Turner et al., 2018; Azadi et al., 2018).

Often we are interested in computing the likelihood ratio between many arbitrary hypotheses. Training $\mathbf{d}(\mathbf{x})$ for every possible pair of hypotheses becomes impractical. A solution proposed by (Cranmer et al., 2015; Baldi et al., 2016) is to parameterize the classifier $\mathbf{d}$ with $\boldsymbol{\theta}$ (typically by injecting $\boldsymbol{\theta}$ as a feature) and train $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ to distinguish between samples from $p(\mathbf{x} \mid \boldsymbol{\theta})$ and samples from an arbitrary but fixed reference hypothesis $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$. In this setting, the decision function modeled by the optimal classifier (Cranmer et al., 2015) is

$$\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x} \mid \boldsymbol{\theta}) + p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})}, \tag{10}$$

thereby defining the likelihood-to-reference ratio as

$$r(\mathbf{x} \mid \boldsymbol{\theta}) \triangleq r(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{ref}}) = \frac{\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})}{1 - \mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})}. \tag{11}$$

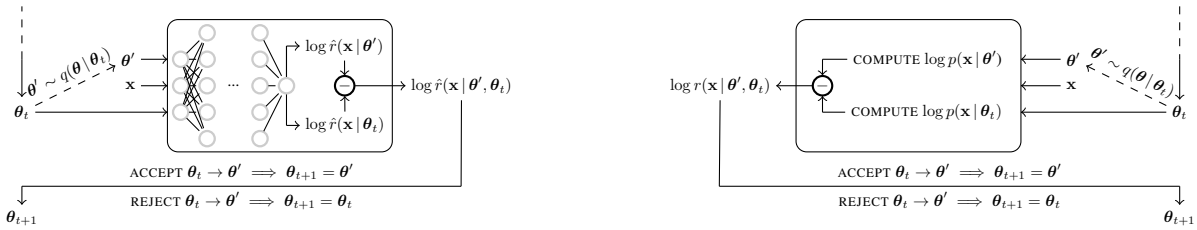
Subsequently, the likelihood ratio between arbitrary hypotheses $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ can then be expressed as

$$r(\mathbf{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{r(\mathbf{x} \mid \boldsymbol{\theta}_0)}{r(\mathbf{x} \mid \boldsymbol{\theta}_1)}. \tag{12}$$

3. Method

We propose a method to draw samples from a posterior with an intractable likelihood and marginal model. As noted earlier, MCMC samplers rely on the likelihood ratio to compute the acceptance ratio. We propose to remove the dependency on the intractable likelihoods $p(\mathbf{x} \mid \boldsymbol{\theta}')$ and $p(\mathbf{x} \mid \boldsymbol{\theta}_t)$ by directly modeling their ratio using an amortized ratio estimator $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta}', \boldsymbol{\theta}_t)$. We call this method amortized approximate likelihood ratio MCMC (AALR-MCMC). Figure 1 provides a schematic overview of the proposed method.

Likelihood-free Metropolis-Hastings Adapting MH to the likelihood-free setup is achieved by replacing the computation of the intractable likelihood ratio in Equation 3 with $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta}', \boldsymbol{\theta}_t)$. The algorithm remains otherwise unchanged. We summarize the likelihood-free Metropolis-Hastings sampler in Appendix A.



(a) AALR-MCMC does not have to evaluate the likelihood, but instead computes an approximation of the likelihood ratio.

(b) Vanilla MCMC computes the likelihood(s) whenever a transition needs to be assessed.

Figure 1. Overview showing *(a)* the proposed method AALR-MCMC and *(b)* traditional MCMC when evaluating the transition from the current state $\boldsymbol{\theta}_t$ to a candidate state $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_t)$. Both methods rely on the acceptance ratio as a test-statistic to evaluate the quality of the proposed transition $\boldsymbol{\theta}_t \to \boldsymbol{\theta}'$. AALR-MCMC does not depend on the evaluation of the (intractable) likelihood. Rather, it relies on an amortized estimator (Section 3.1) to approximate the likelihood ratio $r(\mathbf{x} \mid \boldsymbol{\theta}', \boldsymbol{\theta}_t)$.

Likelihood-free Hamiltonian Monte Carlo The first step in making HMC likelihood-free, is by showing that $U(\boldsymbol{\theta}_t) - U(\boldsymbol{\theta}')$ reduces to the log-likelihood ratio,

$$\begin{aligned} U(\boldsymbol{\theta}_t) - U(\boldsymbol{\theta}') &= \log p(\mathbf{x} \mid \boldsymbol{\theta}') - \log p(\mathbf{x} \mid \boldsymbol{\theta}_t) \\ &= \log r(\mathbf{x} \mid \boldsymbol{\theta}', \boldsymbol{\theta}_t). \end{aligned} \tag{13}$$

To simulate the Hamiltonian dynamics of $\boldsymbol{\theta}_t$, we require a likelihood-free definition of $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. Within our framework, $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ can be expressed as

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = -\frac{\nabla_{\boldsymbol{\theta}} r(\mathbf{x} \mid \boldsymbol{\theta})}{r(\mathbf{x} \mid \boldsymbol{\theta})}. \tag{14}$$

This form can be recovered by a differentiable $\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})$, as expanding $r(\mathbf{x} \mid \boldsymbol{\theta})$ in Equation 14 yields

$$-\frac{\nabla_{\boldsymbol{\theta}} r(\mathbf{x} \mid \boldsymbol{\theta})}{r(\mathbf{x} \mid \boldsymbol{\theta})} = -\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x} \mid \boldsymbol{\theta}). \tag{15}$$

Having likelihood-free alternatives for $U(\boldsymbol{\theta}) - U(\boldsymbol{\theta}')$ and $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$, we can replace these components in HMC to obtain a likelihood-free HMC sampler. This procedure is summarized in Appendix A. While likelihood-free HMC does not rely on the intractable likelihood, it still depends on the computation of $\nabla_{\boldsymbol{\theta}} \hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ to recover $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. This can be a costly operation depending on the size of the ratio estimator. Similar to HMC, the sampler requires careful tuning to maximize the sampling efficiency.

3.1. Improving the ratio estimator $\hat{r}$

Simply relying on the amortized likelihood-to-reference ratio estimator $\hat{r}$ does not yield satisfactory results, even when considering simple toy problems. Experiments indicate that the choice of the mathematically arbitrary reference hypothesis $\boldsymbol{\theta}_{\text{ref}}$ does have a significant effect on the approximated likelihood ratios in practice. Other independent studies (Dutta et al., 2016) observe similar issues and also conclude that the reference hypothesis $\boldsymbol{\theta}_{\text{ref}}$ is a sensitive hyper-parameter which requires careful tuning for the problem at hand. We find that poor inference results occur in the absence of support between $p(\mathbf{x} \mid \boldsymbol{\theta})$ and $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$, as illustrated in Figure 2. In this example, the evaluation of the approximate ratio $\hat{r}$ for an observation $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}^*)$ is undefined when the observation $\mathbf{x}$ does not have density in $p(\mathbf{x} \mid \boldsymbol{\theta})$ and $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$, or either of the densities is numerically negligible. Therefore, the decision function modeled by the optimal classifier $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ outside of the space covered by $p(\mathbf{x} \mid \boldsymbol{\theta})$ and $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$ is undefined. Practically, this implies that the ratio $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ can take on an arbitrary value which is detrimental to the inference procedure because multiple solutions for $\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})$ exist.

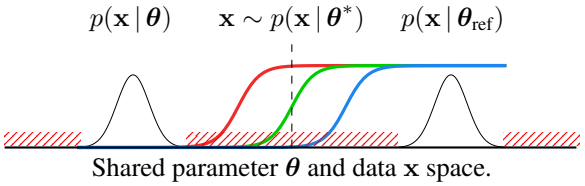


Figure 2. Consider having access to an optimal classifier $\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})$ modeling $r(\mathbf{x} \mid \boldsymbol{\theta})$ with $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta}^*)$. This ratio is undefined for $\mathbf{x}$ as neither $p(\mathbf{x} \mid \boldsymbol{\theta})$ nor $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$ puts numerically non-negligible density on $\mathbf{x}$. This implies that $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ and its decision function $\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})$ can take on arbitrary values in regions not covered by $p(\mathbf{x} \mid \boldsymbol{\theta})$ or $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$ (striped areas) because no such training data exists. The red, green and blue lines depict optimal decision functions as they all minimize the criterion which captures the ability to classify between samples from $p(\mathbf{x} \mid \boldsymbol{\theta})$ and $p(\mathbf{x} \mid \boldsymbol{\theta}_{\text{ref}})$. However, the functions have different approximations of $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$.

To overcome the issues associated with a fixed reference hypothesis, we propose to train the parameterized classifier to distinguish dependent sample-parameter pairs $(\mathbf{x}, \boldsymbol{\theta}) \sim p(\mathbf{x}, \boldsymbol{\theta})$ with class label $y = 1$ from independent sample-parameter pairs $(\mathbf{x}, \boldsymbol{\theta}) \sim p(\mathbf{x})p(\boldsymbol{\theta})$ with class label $y = 0$. This modification results in the optimal classifier

$$\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x}, \boldsymbol{\theta}) + p(\mathbf{x})p(\boldsymbol{\theta})}, \tag{16}$$

Algorithm 1 Optimization of $\mathbf{d}_\phi(\mathbf{x}, \boldsymbol{\theta})$.

Inputs:	Criterion ℓ (e.g., BCE)
	Implicit generative model $p(\mathbf{x} \mid \boldsymbol{\theta})$
	Prior $p(\boldsymbol{\theta})$
Outputs:	Parameterized classifier $\mathbf{d}_\phi(\mathbf{x}, \boldsymbol{\theta})$
Hyperparameters:	Batch-size M

while not converged do
  Sample $\boldsymbol{\theta} \leftarrow \{\boldsymbol{\theta}_m \sim p(\boldsymbol{\theta})\}_{m=1}^M$
  Sample $\boldsymbol{\theta}' \leftarrow \{\boldsymbol{\theta}'_m \sim p(\boldsymbol{\theta})\}_{m=1}^M$
  Simulate $\mathbf{x} \leftarrow \{\mathbf{x}_m \sim p(\mathbf{x} \mid \boldsymbol{\theta}_m)\}_{m=1}^M$
  $\mathcal{L} \leftarrow \ell(\mathbf{d}_\phi(\mathbf{x}, \boldsymbol{\theta}),\ 1) + \ell(\mathbf{d}_\phi(\mathbf{x}, \boldsymbol{\theta}'),\ 0)$
  $\phi \leftarrow$ **OPTIMIZER**$(\phi,\ \nabla_\phi \mathcal{L})$
end while
return $\mathbf{d}_\phi$

and thereby in the likelihood-to-evidence ratio

$$\frac{\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})}{1 - \mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})} = \frac{p(\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x})p(\boldsymbol{\theta})} = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})}{p(\mathbf{x})} = r(\mathbf{x} \mid \boldsymbol{\theta}). \quad (17)$$

This formulation ensures that the likelihood-to-evidence ratio will always be defined everywhere it needs to be evaluated, as the joint $p(\mathbf{x}, \boldsymbol{\theta})$ is consistently supported by the product of marginals $p(\mathbf{x})p(\boldsymbol{\theta})$.

We summarize the procedure for learning the classifier $\mathbf{d}^*(\mathbf{x}, \boldsymbol{\theta})$ and the corresponding ratio estimator $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ in Algorithm 1. The algorithm amounts to the minimization of the binary cross-entropy (BCE) loss of a classifier $\mathbf{d}_\phi$. We provide a proof in Appendix B that demonstrates that it results in the optimal discriminator $\mathbf{d}^*$.

Although the usage of the marginal model instead of an arbitrary reference hypothesis vastly improves the accuracy of $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$, obtaining the likelihood-to-evidence ratio $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ by transforming the output of $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ can still be susceptible to numerical errors. This may happen in the saturating regime where the classifier $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ is able to (almost) perfectly discriminate samples from $p(\mathbf{x} \mid \boldsymbol{\theta})$ and $p(\mathbf{x})$. We prevent this issue by extracting $\log \hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ from the neural network before applying the sigmoidal projection in the output layer, since $\log \hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ is the logit of $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$. This choice also mitigates a vanishing gradient when computing $\nabla_{\boldsymbol{\theta}} \log \hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ or $\nabla_{\mathbf{x}} \log \hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$.

Finally, approximating the likelihood-to-evidence ratio also enables the direct estimation of the posterior density as $\hat{p}(\boldsymbol{\theta} \mid \mathbf{x}) = p(\boldsymbol{\theta})\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$. This is useful in low-dimensional model parameter spaces, where scanning is a reasonable strategy.

3.2. Receiver operating curve diagnostic

Likelihood-free computations are challenging to verify as the likelihood is by definition intractable. A robust strategy is necessary to verify the quality of the approximation before making any scientific conclusion based on a likelihood-free approach. Inspired by Cranmer et al. (2015), we identify issues in our ratio-estimator $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ by evaluating the identity $p(\mathbf{x} \mid \boldsymbol{\theta}) = p(\mathbf{x})\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$. If $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$ is exact, then a classifier should not be able to distinguish between samples from $p(\mathbf{x} \mid \boldsymbol{\theta})$ and the reweighted marginal model $p(\mathbf{x})\hat{r}(\mathbf{x} \mid \boldsymbol{\theta})$. The discriminative performance of the classifier can be assessed by means of a ROC curve. A diagonal ROC (AUC = 0.5) curve indicates that a classifier is insensitive and $\hat{r}(\mathbf{x} \mid \boldsymbol{\theta}) = r(\mathbf{x} \mid \boldsymbol{\theta})$. This result can also be obtained if the classifier is not powerful enough to extract any predictive features. Figure 3 provides an illustration of this diagnostic.

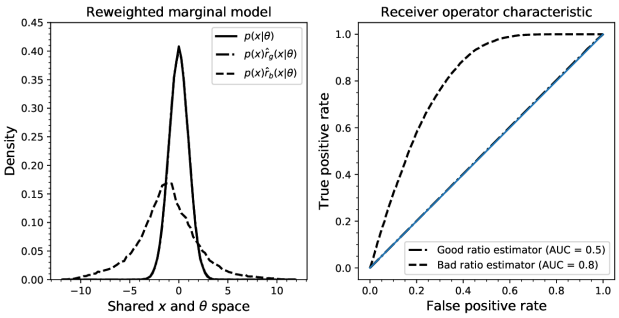


Figure 3. This figure demonstrates the diagnostic presented in Section 3.2. We train two ratio estimators. The first approximates the ratio $r(\mathbf{x} \mid \boldsymbol{\theta})$ well, while the other does not. We denote these estimators as $\hat{r}_g(\mathbf{x} \mid \boldsymbol{\theta})$ and $\hat{r}_b(\mathbf{x} \mid \boldsymbol{\theta})$ respectively. The test diagnostic is applied to a single test hypothesis $\boldsymbol{\theta} = 0$. *(Left):* Marginal model reweighted using $\hat{r}_g(\mathbf{x} \mid \boldsymbol{\theta})$ and $\hat{r}_b(\mathbf{x} \mid \boldsymbol{\theta})$. It is clear that $\hat{r}_b(\mathbf{x} \mid \boldsymbol{\theta})$ does not properly approximate $r(\mathbf{x} \mid \boldsymbol{\theta})$, as the reweighted marginal model is distinguishable from the test hypothesis $p(\mathbf{x} \mid \boldsymbol{\theta} = 0)$. *(Right):* A classifier is trained to distinguish between samples from the test hypothesis and the reweighted marginal models. The ROC curve indicates that the classifier could not extract any predictive features for samples $\mathbf{x} \sim p(\mathbf{x})$ reweighted by $\hat{r}_g(\mathbf{x} \mid \boldsymbol{\theta})$, indicating a good approximation of $r(\mathbf{x} \mid \boldsymbol{\theta})$ by $\hat{r}_g(\mathbf{x} \mid \boldsymbol{\theta})$.

4. Related work

Algorithms such as ABC (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002; Marin et al., 2011) tackle the problem of Bayesian inference by collecting proposal states $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ whenever an observation $\mathbf{x}$ produced by the forward model $\mathbf{x} \sim p(\mathbf{x} \mid \boldsymbol{\theta})$ resembles an observation $\mathbf{x}_o$. Formally, a proposal state $\boldsymbol{\theta}$ is accepted whenever a compressed observation $\sigma(\mathbf{x})$ (low-dimensional summary statistic) satisfies $d(\sigma(\mathbf{x}), \sigma(\mathbf{x}_o)) < \epsilon$ for some distance function d and acceptance threshold ϵ. *The resulting approximation of the posterior will only be exact whenever the summary statistic is sufficient and* $\epsilon \to 0$ (Beaumont et al., 2002). Several procedures have been proposed to improve

the acceptance rate by guiding simulations based on previously accepted states (Toni et al., 2008; Marjoram et al., 2003; Wegmann et al., 2009). Other works investigated learning summary statistics (Fearnhead & Prangle, 2012; Dinev & Gutmann, 2018; Wong et al., 2018). Contrary to these methods, AALR-MCMC does not actively use the simulator during inference and learns a direct mapping from data and parameter space to likelihood-to-evidence ratios.

Other approaches take the perspective to cast inference as an optimization problem (Neal & Hinton, 1998; Hoffman et al., 2013). In variational inference, a parameterized posterior over parameters of interest is optimized (Salimans et al., 2015). Amortized variational inference (Gershman & Goodman, 2014; Ritchie et al., 2016) expands on this idea by using generative models to capture inference mappings. Recent work in (Louppe et al., 2017) proposes a novel form of variational inference by introducing an adversary in combination with REINFORCE-estimates (Williams, 1992; Sutton et al., 2000) to optimize a parameterized prior. Others have investigated meta-learning to learn parameter updates (Pesah et al., 2018). However, these works only provide point-estimates.

Sequential approaches such as SNPE-A (Papamakarios & Murray, 2016), SNPE-B (Boelts et al., 2019) and APT/SNPE-C (Greenberg et al., 2019) iteratively adjust an approximate posterior parameterized as a mixture density network or a normalizing flow. Instead of learning the posterior directly, SNL (Papamakarios & Murray, 2018) makes use of autoregressive flows to model an approximate likelihood. AALR-MCMC mirrors SNL as the trained conditional density estimator is plugged into MCMC samplers to bypass the intractable marginal model. This allows SNL to approximate the posterior numerically. Contrary to our approach, SNL cannot directly provide estimates of the posterior posterior density function.

The usage of ratios is explored in several studies. CARL (Cranmer et al., 2015) models likelihood ratios for frequentist tests. As shown in Section 3.1, CARL does not produce accurate results in some cases. LFIRE (Dutta et al., 2016) models a likelihood-to-evidence ratio by logistic regression and relies on the usage of summary statistics. Unlike us, they require samples from the marginal model and a specific (reference) likelihood, while we only require samples from the joint $p(\mathbf{x}, \boldsymbol{\theta})$. Therefore, LFIRE requires retraining for every evaluation of different $\boldsymbol{\theta}$.

Finally, an important concern of likelihood-free inference is minimizing the number of simulation calls. Active simulation strategies such as BOLFI (Gutmann & Corander, 2016) and others (Ong et al., 2017; Meeds & Welling, 2014) achieve this through Bayesian optimization. Emulator networks (Lueckmann et al., 2018) exploit the uncertainty within an ensemble to guide simulations. Recent works (Brehmer et al., 2020; 2018) significantly reduce the amount of required simulations, provided joint likelihood ratios and scores can be extracted from the simulator.

5. Experiments

5.1. Setup

We compare AALR-MCMC against rejection ABC and established modern posterior approximation techniques such as SNL, SNPE-A SNPE-B and APT. We allocate a *simulation budget* of one million forward passes. Sequential approaches such as SNPE-A, SNPE-B, and APT spread this budget equally across 100 rounds. These rounds focus the simulation budget to iteratively improve the approximation of a *single* posterior. For SNL, due to the computational constraints of its inner MCMC sampling step, we limit the simulation budget to 100000 forward passes spread equally over 100 rounds. Unless stated otherwise, our evaluations assess the posterior estimate obtained in the final round. Although the ratio estimator in AALR-MCMC is trained *once* to model all posteriors (amortization), we only examine the posterior of interest $p(\boldsymbol{\theta} \,|\, \mathbf{x} = \mathbf{x}_o)$. This choice puts our method at a disadvantage since the task of amortized inference is more complex compared to fitting of a single posterior. We stress that from a scientific point of view, accuracy of the approximation is preferred over simulation cost. All experiments are repeated 25 times. AALR-MCMC makes use of the likelihood-free Metropolis-Hastings sampler. Implementation guidelines are discussed in Appendix C. Experimental details, additional results and plots demonstrating several other aspects are discussed in Appendix D. Code is available at https://github.com/montefiore-ai/hypothesis.

5.1.1. BENCHMARK PROBLEMS

Tractable problem Given a model parameter sample $\boldsymbol{\theta} \in \mathbb{R}^5$, the forward generative process is defined as:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\theta}} &= (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1), \\ s_1 = \boldsymbol{\theta}_2^2, \;\; s_2 &= \boldsymbol{\theta}_3^2, \;\; \rho = \tanh(\boldsymbol{\theta}_4), \\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}} &= \begin{bmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{bmatrix}, \end{aligned}$$

$$\text{with } \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_4) \text{ where } \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$$

The likelihood is $p(\mathbf{x} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{4} \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_\theta,\ \boldsymbol{\Sigma}_\theta)$, with a uniform prior $p(\boldsymbol{\theta})$ between $[-3, 3]$ for every $\boldsymbol{\theta}_i$. The resulting posterior is non-trivial due to squaring operations, which are responsible for the presence of multiple modes. An observation $\mathbf{x}_o$ is generated by conditioning the forward model on $\boldsymbol{\theta}^* = (0.7, -2.9, -1.0, -0.9, 0.6)$ as in Papamakarios & Murray (2018) and Greenberg et al. (2019).

Detector calibration We like to determine the offset $\boldsymbol{\theta} \in \mathbb{R}$ of a particle detector from the collision point given

Algorithm	Tractable problem	Detector calibration	Population model	M/G/1
ABC	-6.668 ± 0.000	-2.180 ± 0.000	N/A	N/A
SNPE-A	-6.141 ± 1.227	-1.775 ± 1.775	$\mathbf{7.024} \pm 0.515$	1.177 ± 0.937
SNPE-B	-5.693 ± 0.809	-1.075 ± 0.226	-0.632 ± 0.843	1.105 ± 0.384
APT	-4.441 ± 0.487	-2.004 ± 0.753	6.366 ± 0.432	-2.741 ± 3.356
SNL	$\mathbf{-4.060} \pm 0.308$	N/A	N/A	N/A
AALR-MCMC (ours)	-4.126 ± 0.004	$\mathbf{-1.005} \pm 0.074$	6.482 ± 0.214	$\mathbf{2.302} \pm 0.189$

Table 1. Posterior log probabilities $\log p(\boldsymbol{\theta} = \boldsymbol{\theta}^* \mid \mathbf{x} = \mathbf{x}_o)$ for generating parameters $\boldsymbol{\theta}^*$ and observation $\mathbf{x}_o$. For SNPE-A, SNPE-B and APT we directly extracted the posterior log probability from the mixture of Gaussians. Since the proposed ratio estimator models the log likelihood-to-evidence ratio, we compute $\log p(\boldsymbol{\theta} = \boldsymbol{\theta}^* \mid \mathbf{x} = \mathbf{x}_o)$ as $\log r(\mathbf{x} = \mathbf{x}_o \mid \boldsymbol{\theta} = \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta} = \boldsymbol{\theta}^*)$. Assessing the quality of a method exclusively based on the observed log posterior probabilities is potentially **misleading**, as the metric does not take the structure of the posterior into account. As such, we provide this table for historic reasons to comply with previous studies such as Papamakarios & Murray (2016), Boelts et al. (2019) and Greenberg et al. (2019).

a detector response $\mathbf{x}_o$. Our particle detector emulates a 32×32 spherical uniform grid such that $\mathbf{x} \in \mathbb{R}^{1024}$. Every detector pixel measures the momentum of particles passing through the detector material. The `pythia` simulator (Sjöstrand et al., 2008) generates electron-positron ($e^- e^+$) collisions and is configured according to the parameters derived by the Monash tune (Skands et al., 2014). The collision products and their momenta are processed by `pythiamill` (Crosby, 2020) to compute the response of the detector by simulating the interaction of the collision products with the detector material. We consider a prior $p(\boldsymbol{\theta}) \triangleq \mathcal{U}(-30, 30)$ with $\mathbf{x}_o$ generated at $\boldsymbol{\theta}^* = 0$.

Population model The Lotka-Volterra model (Lotka, 1920) describes the evolution of predator-prey populations. The population dynamics are driven by a set of differential equations with parameters $\boldsymbol{\theta} \in \mathbb{R}^4$. An observation describes the population counts of both groups over time. Simulations are typically compressed into a summary statistic $\bar{\mathbf{x}} \in \mathbb{R}^9$ (Papamakarios & Murray, 2018; Greenberg et al., 2019). We also follow this approach to remain consistent. The prior $p(\boldsymbol{\theta}) \triangleq \mathcal{U}(-10, 2)$ (log-scale) for every $\boldsymbol{\theta}_i$. We generate an observation from the narrow oscillating regime $\boldsymbol{\theta}^* = (-4.61, -0.69, 0, -4.61)$.

M/G/1 queuing model This model describes a queuing system of continuously arriving jobs at a single server and is described by a model parameter $\boldsymbol{\theta} \in \mathbb{R}^3$. The time it takes to process every job is uniformly distributed in the interval $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$. The arrival time between two consecutive jobs is exponentially distributed according to the rate $\boldsymbol{\theta}_3$. An observation $\mathbf{x}$ are 5 equally spaced percentiles of inter-departure times, i.e., the 0th, 25th, 50th, 75th and 100th percentiles. An observation $\mathbf{x}_o$ is generated by conditioning the forward model on $\boldsymbol{\theta}^* = (1.0, 5.0, 0.2)$ as in Papamakarios & Murray (2018). We consider a uniform prior $p(\boldsymbol{\theta}) \triangleq \mathcal{U}(0, 10) \times \mathcal{U}(0, 10) \times \mathcal{U}(0, 0.333)$.

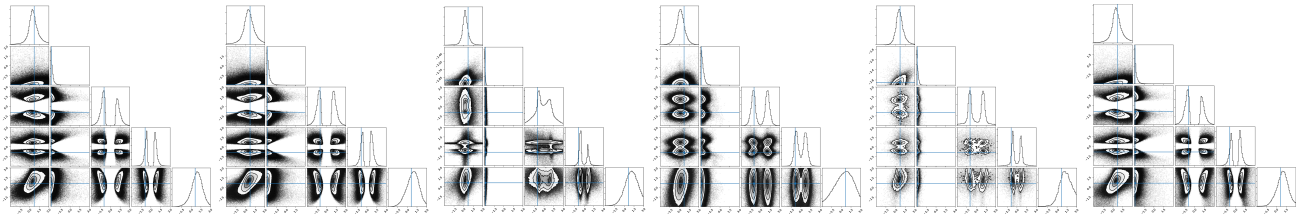
5.2. Results

Table 1 shows the posterior log probabilities of the generating parameter $\boldsymbol{\theta}^*$ for an observation $\mathbf{x}_o$. Additionally, the ROC diagnostic for our method reports AUC = 0.58 for the tractable problem, AUC = 0.5 for the detector calibration and M/G/1 benchmarks, and AUC = 0.55 for the population evolution model. These results demonstrate that the proposed ratio estimator provides accurate ratio estimates.

If we assess the quality of the methods exclusively based on the log posterior probabilities in Table 1, we could argue that all methods are close to each other in terms of approximation, with AALR-MCMC yielding the best results for the detector calibration and M/G/1 and SNL and SNPA-A respectively producing the most accurate inference for the tractable problem and the population evolution model. However, this is potentially misleading, as the metric does not take the structure of the posterior into account. Again, we stress that the ability of inference techniques to approximate the posterior accurately is critical in scientific applications which seek to, for instance, constrain the model parameter $\boldsymbol{\theta}$. To demonstrate this point, we focus on the tractable

Algorithm	MMD	ROC AUC
AALR-MCMC (ours)	$\mathbf{0.05} \pm 0.005$	$\mathbf{0.58} \pm 0.0080$
AALR-MCMC (LRT)	0.53 ± 0.004	0.99 ± 0.0001
ABC	0.29 ± 0.004	0.98 ± 0.0007
SNPE-A	0.21 ± 0.070	0.93 ± 0.0305
SNPE-B	0.20 ± 0.061	0.91 ± 0.0409
APT	0.17 ± 0.036	0.83 ± 0.0145
SNL	0.11 ± 0.091	0.63 ± 0.0564

Table 2. Results for the tractable benchmark. AALR-MCMC outperforms all other methods across in terms of accuracy and robustness (low variance). Numerical errors introduced by MCMC might have contributed to these results. The MMD scores are in agreement with Greenberg et al. (2019).



(a) MCMC groundtruth (b) AALR-MCMC (c) SNPE-A (d) SNPE-B (e) APT (f) SNL

Figure 4. Posteriors from the tractable benchmark. The experiments are repeated 25 times and the approximate posteriors are subsampled from those runs. An objective visual assessment can be made: AALR-MCMC shares the same structure with the MCMC truth, demonstrating its accuracy. Some runs of the other methods were not consistent, contributing to the variance observed in Table 2.

problem and carry out two distinct quantitative analyses between the samples of the approximate posterior and the MCMC groundtruth. The first computes the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) while the latter trains a classifier to compute the ROC AUC. Results are summarized in Table 2 while Figure 4 shows the approximations and the groundtruth. Both AALR-MCMC and SNL accurately model the true posterior, but SNPA-A, SNPA-B and APT clearly fail to do so. The observed discrepancy between the LRT and the proposed ratio estimator indicates that the improvements in Section 3.1 are *critical*.

In addition to comparing the final approximations, we evaluate the accuracy of the approximations with respect to a given simulation budget. In doing so we challenge our method even further, as sequential approaches are specifically designed to be simulation efficient. We expect sequential approaches to obtain more accurate approximations with less simulations. The results of this evaluation are shown in Figure 5. With the exception of SNL which produces results comparable to ours, we unexpectedly find that the sequential approaches were not able to outperform our method on this (toy) problem, even though AALR-MCMC and its ratio estimator tackle the harder task of amortized inference. This demonstrates the accuracy and robustness of our method.

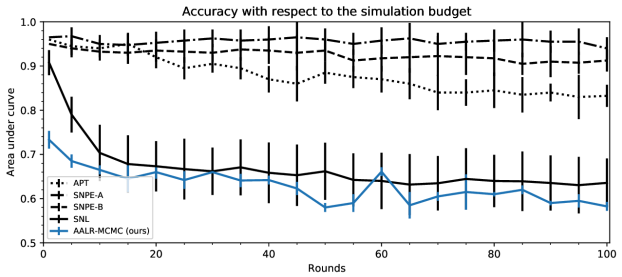


Figure 5. We evaluate the accuracy of the approximations with respect to different simulation budgets on the tractable benchmark. The accuracy is obtained by computing the ROC AUC between samples from the approximation and the MCMC groundtruth. Except for SNL which yields comparable results, sequential approaches are not able to outperform AALR-MCMC.

5.3. Demonstrations: strong gravitational lensing

The following demonstrations will showcase several aspects of our method while considering the problem of strong gravitational lensing. We use `autolens` (Nightingale et al., 2018) to simulate the telescope optics, imaging sensors and physics governing strong lensing. The simulation black-box encapsulates these components. The output of the simulation is a high-dimensional observation $\mathbf{x} \in \mathbb{R}^{128\times128}$ with uninformative data dimensions. We use a ratio estimator based on RESNET-18 (He et al., 2016) parameterized by $\boldsymbol{\theta}$ in the fully connected trunk. Appendix D.6 discusses the setups and the simulation models in detail.

5.3.1. MARGINALIZATION OF NUISANCE PARAMETERS

Often scientists are aphetic about a posterior describing all model parameters. Rather, they are interested in a posterior in which nuisance parameters have been marginalized out. This is easily achieved within our framework by including all parameters (including nuisance parameters) to the simulation model, but only presenting the parameters of interest to the ratio estimator during training. The training procedure remains otherwise unchanged. This problem focuses on recovering the Einstein radius $\boldsymbol{\theta} \in \mathbb{R}$ of a gravitational lens. We are not interested in the parameters describing the source and foreground galaxy (15 parameters). Figure 6 depicts our posterior approximation, ROC diagnostic and observation $\mathbf{x}_o$ with $\boldsymbol{\theta}^* = 1.66$ and prior $p(\boldsymbol{\theta}) \triangleq \mathcal{U}(0.5, 3.0)$.

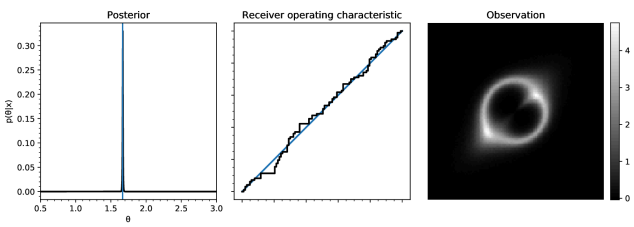


Figure 6. (Left): Approximation of the posterior. *(Middle):* Diagonal ROC diagnostic, indicating a good approximation of the posterior. *(Right):* Observation associated with the posterior.

5.3.2. AMORTIZATION ENABLES POPULATION STUDIES

Consider a set of n independent and identically distributed observations $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. The amortization of the ratio estimator allows additional observations to be included in the computation of the posterior $p(\boldsymbol{\theta}\,|\mathcal{X})$ without requiring new simulations or retraining. This allows us to efficiently undertake population studies. Bayes' rule tells us

$$\begin{aligned} p(\boldsymbol{\theta}\,|\mathcal{X}) &= \frac{p(\boldsymbol{\theta}) \prod_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}\,|\,\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \prod_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}\,|\,\boldsymbol{\theta}) d\,\boldsymbol{\theta}}, \\ &\approx \frac{p(\boldsymbol{\theta}) \prod_{\mathbf{x}\in\mathcal{X}} \hat{r}(\mathbf{x}\,|\,\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) \prod_{\mathbf{x}\in\mathcal{X}} \hat{r}(\mathbf{x}\,|\,\boldsymbol{\theta}) d\,\boldsymbol{\theta}}. \end{aligned} \tag{18}$$

The denominator can efficiently be approximated by Monte Carlo sampling using the ratio estimator $\hat{r}(\mathbf{x}\,|\,\boldsymbol{\theta})$. However, with MCMC the denominator cancels out within the ratio between consecutive states $\boldsymbol{\theta}_t \to \boldsymbol{\theta}'$. Thereby obtaining

$$\frac{\hat{p}(\boldsymbol{\theta}'|\mathcal{X})}{\hat{p}(\boldsymbol{\theta}_t|\mathcal{X})} = \frac{p(\boldsymbol{\theta}') \prod_{\mathbf{x}\in\mathcal{X}} \hat{r}(\mathbf{x}\,|\,\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_t) \prod_{\mathbf{x}\in\mathcal{X}} \hat{r}(\mathbf{x}\,|\,\boldsymbol{\theta}_t)}. \tag{19}$$

We consider the same simulation model as in Section 5.3.1, with the exception that the Einstein radius used to simulate a gravitational lens is not $\boldsymbol{\theta}$, but instead drawn from $\mathcal{N}(\boldsymbol{\theta}, 0.25)$. We reduce the uncertainty about the generating parameter $\boldsymbol{\theta}^* = 2$ by modeling the posterior $\hat{p}(\boldsymbol{\theta}\,|\mathcal{X})$. This is demonstrated in Figure 7. All individual posteriors (dotted lines) are derived using the same pretrained ratio estimator. The posterior $\hat{p}(\boldsymbol{\theta}\,|\mathcal{X})$ is approximated using the formalism described above.

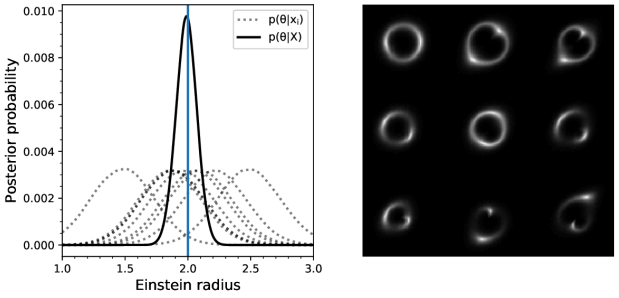


Figure 7. (Left): The dotted lines represent the posteriors $\hat{p}(\boldsymbol{\theta}\,|\,\mathbf{x} = \mathbf{x}_i)$ for every independent and identically distributed observation $\mathbf{x}_i$, while the solid line depicts the posterior $\hat{p}(\boldsymbol{\theta}\,|\mathcal{X})$. All posteriors are derived using the same pretrained ratio estimator. *(Right)*: Observations sampled from $p(\mathbf{x}\,|\,\boldsymbol{\theta} = \boldsymbol{\theta}^*)$.

5.3.3. BAYESIAN MODEL SELECTION

Until now we only considered posteriors with continuous model parameters. We turn to a setting in which scientists are interested in a discrete space of models. *In essence casting classification as Bayesian model selection, allowing us the quantify the uncertainty among models (classes) with respect to an observation.* We demonstrate the task of model selection by computing the posterior $\hat{p}(m|\,\mathbf{x})$ across a space of 10 models $\mathcal{M} = \{m_0, \dots, m_9\}$. The index i of a model m_i corresponds to the number of source galaxies present in the lensing system. The categorical prior $p(m)$ is uniform. Figure 8 shows $\hat{p}(m|\,\mathbf{x})$ and the associated diagnostic for different observations. Both posteriors were computed using the same ratio estimator.

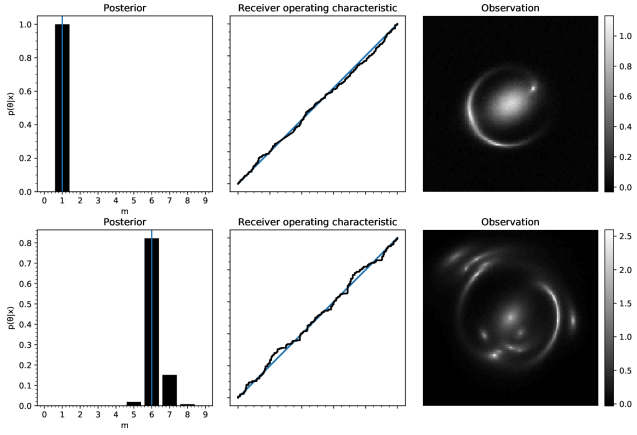


Figure 8. Posterior $\hat{p}(m|\,\mathbf{x})$ over the model space $\mathcal{M}$. Both diagnostics are diagonal. *(Top)*: Lensing system with a single source galaxy. *(Bottom):* Lensing system with 6 different source galaxies. The MAP of the posterior $\hat{p}(m|\,\mathbf{x})$ identifies the correct number of source galaxies, despite abundant lensing artifacts.

5.4. Estimator capacity and sequential ratio estimation

The amortization of our ratio estimator requires sufficient representational capacity to accurately approximate $r(\mathbf{x}\,|\,\boldsymbol{\theta})$, which of course directly depends on the complexity of the task at hand. As we explore in Appendix E, if the capacity of the ratio estimator is too low, then the quality of inference is impaired.

However, increasing the capacity of a ratio estimator to match the complexity of the inference problem is not always a viable strategy, nor easy to determine beforehand. We observe that for a trained classifier $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ with insufficient capacity (AUC > 0.5) the posterior $\hat{p}(\boldsymbol{\theta}\,|\,\mathbf{x} = \mathbf{x}_o)$ is typically larger compared to the true posterior. Since the true decision function cannot be modeled, the loss of the classifier $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ is indeed necessarily larger than the loss of the optimal classifier, which effectively means that the classifier $\mathbf{d}(\mathbf{x}, \boldsymbol{\theta})$ should not be able to exclude sample-parameter pairs $(\mathbf{x}, \boldsymbol{\theta})$. This is a desirable property because the generating parameters $\boldsymbol{\theta}^*$ should not be excluded either. From this observation, we can run a sequential ratio estimation procedure in which the posterior for $\mathbf{x} = \mathbf{x}_o$ is refined iteratively across a series of rounds. Starting with the initial prior $p_0(\boldsymbol{\theta}) := p(\boldsymbol{\theta})$, we improve the posterior by setting as prior for the next round, $p_{t+1}(\boldsymbol{\theta})$, the posterior

$\hat{p}_t(\boldsymbol{\theta} \,|\, \mathbf{x} = \mathbf{x}_o)$ obtained at the previous round. At each iteration, the training procedure is repeated and eventually terminates based on the ROC diagnostic (AUC = 0.5).

To demonstrate this sequential ratio estimation procedure, let us assume the population model setting. Our ratio estimator is a low-capacity MLP with 3 layers and 50 hidden units. In every round t, 10,000 sample-parameter pairs are drawn from the joint $p(\mathbf{x}, \boldsymbol{\theta})$ with prior $p_t(\boldsymbol{\theta})$ for training. The following AUC scores were obtained: .99, .92, .54, and finally .50, terminating the algorithm.

Let us finally note that some time after the first version of this work, Durkan et al. (2020) identified that the sequential ratio estimation procedure outlined here is strongly related to APT/SNPE-C, in the sense that both approaches can actually be viewed as instances of a more general and unified constrastive learning scheme.

6. Summary and discussion

This work introduces a novel approach for Bayesian inference.We achieve this by replacing the intractable evaluation of the likelihood ratio in MCMC with an amortized likelihood ratio estimator. We demonstrate that a straightforward application of the likelihood ratio trick to MCMC is insufficient. We solve this by modeling the likelihood-to-evidence ratio for arbitrary observations $\mathbf{x}$ and model parameters $\boldsymbol{\theta}$. This implies that a pretrained ratio estimator can be used to infer the posterior density function of arbitrary observations. A theoretical argument demonstrates that the training procedure yields the optimal ratio estimator. The accuracy of an approximation can easily be verified by the proposed diagnostic. No summary statistics are required, as the technique directly learns mappings from observations and model parameters to likelihood-to-evidence ratios. Our framework allows for the usage of off-the-shelf neural architectures such as RESNET (He et al., 2016). Experiments highlight the accuracy and robustness of our method.

Simulation efficiency We take the point of view that accuracy of the approximation is preferred over simulation cost. This is the case in many scientific disciplines which seek to reduce the uncertainty over a parameter of interest. Despite the experimental handicap, we have shown that existing simulation efficient approaches are not able to outperform our method in terms of accuracy with respect to a certain (and small) simulation budget.

ACKNOWLEDGMENTS

The authors would like to thank Antoine Wehenkel and Matthia Sabatelli for the insightful discussions and comments. Joeri Hermans would like to thank the National Fund for Scientific Research for his FRIA scholarship. Gilles Louppe is recipient of the ULiège - NRB Chair on Big data and is thankful for the support of NRB.

References

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., and Odena, A. Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*, 2018.

Baldi, P., Cranmer, K., Faucett, T., Sadowski, P., and Whiteson, D. Parameterized neural networks for high-energy physics. *Eur. Phys. J. C*, 76(5):235, April 2016. ISSN 1434-6044, 1434-6052. URL https://doi.org/10.1140/epjc/s10052-016-4099-4.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002. URL http://www.genetics.org/content/162/4/2025.

Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017. URL https://arxiv.org/abs/1701.02434.

Boelts, J., Lueckmann, J.-M., Goncalves, P. J., Sprekeler, H., and Macke, J. H. Comparing neural simulations by neural density estimation. In *2019 Conference on Cognitive Computational Neuroscience*, pp. 1289–1299. Cognitive Computational Neuroscience, 2019. URL https://doi.org/10.32470/ccn.2019.1291-0.

Brehmer, J., Cranmer, K., Louppe, G., and Pavez, J. A guide to constraining effective field theories with machine learning. *Phys. Rev. D*, 98(5):052004, September 2018. ISSN 2470-0010, 2470-0029. URL https://doi.org/10.1103/physrevd.98.052004.

Brehmer, J., Louppe, G., Pavez, J., and Cranmer, K. Mining gold from implicit models to improve likelihood-free inference. *Proc Natl Acad Sci USA*, 117(10):5242–5249, February 2020. ISSN 0027-8424, 1091-6490. URL https://doi.org/10.1073/pnas.1915980117.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015. URL https://arxiv.org/abs/1506.02169.

Crosby, O. PYTHIA, pythia, April 2020. URL https://doi.org/10.1002/9783527809080.cataz13886.

Dinev, T. and Gutmann, M. U. Dynamic likelihood-free inference via ratio estimation (dire). *arXiv preprint arXiv:1810.09899*, 2018.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, September 1987. ISSN 0370-2693. URL https://doi.org/10.1016/0370-2693(87)91197-x.

Durkan, C., Murray, I., and Papamakarios, G. On Contrastive Learning for Likelihood-free Inference. *arXiv e-prints*, art. arXiv:2002.03712, February 2020.

Dutta, R., Corander, J., Kaski, S., and Gutmann, M. U. Likelihood-free inference by ratio estimation. *arXiv preprint arXiv:1611.10242*, 2016.

Fearnhead, P. and Prangle, D. Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, May 2012. ISSN 1369-7412. URL https://doi.org/10.1111/j.1467-9868.2011.01010.x.

Gershman, S. and Goodman, N. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2404–2414, Long Beach, California, USA, 2019. PMLR. URL http://proceedings.mlr.press/v97/greenberg19a.html.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Gutmann, M. U. and Corander, J. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research*, 17 (1):4256–4302, 2016. URL http://jmlr.org/papers/v17/15-017.html.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. Likelihood-free inference via classification. *Stat Comput*, 28(2):411–425, March 2017. ISSN 0960-3174, 1573-1375. URL https://doi.org/10.1007/s11222-017-9738-6.

Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970. ISSN 1464-3510, 0006-3444. URL https://doi.org/10.1093/biomet/57.1.97.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, June 2016. URL https://doi.org/10.1109/cvpr.2016.90.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

J. Neyman, E. P. IX. on the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706):289–337, February 1933. ISSN 0264-3952, 2053-9258. URL https://doi.org/10.1098/rsta.1933.0009.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pp. 971–980, 2017.

Kormann, R., Schneider, P., and Bartelmann, M. Isothermal elliptical gravitational lens models. *Astron. Astrophys.*, 284:285–299, 1994.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. ISSN 0018-9219. URL https://doi.org/10.1109/5.726791.

Lotka, A. J. Analytical note on certain rhythmic relations in organic systems. *Proc Natl Acad Sci USA*, 6(7):410–415, June 1920. ISSN 0027-8424, 1091-6490. URL https://doi.org/10.1073/pnas.6.7.410.

Louppe, G., Hermans, J., and Cranmer, K. Adversarial variational optimization of non-differentiable simulators. *arXiv preprint arXiv:1707.07113*, 2017. URL https://arxiv.org/abs/1707.07113.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. *arXiv preprint arXiv:1805.09294*, 2018. URL https://arxiv.org/abs/1805.09294.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. Approximate Bayesian computational methods. *Stat Comput*, 22(6):1167–1180, October 2011. ISSN 0960-3174, 1573-1375. URL https://doi.org/10.1007/s11222-011-9288-2.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, December 2003. ISSN 0027-8424, 1091-6490. URL https://doi.org/10.1073/pnas.0306899100.

Meeds, E. and Welling, M. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint arXiv:1401.2838*, 2014. URL https://arxiv.org/abs/1401.2838.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606, 1089-7690. URL https://doi.org/10.1063/1.1699114.

Mohamed, S. and Lakshminarayanan, B. Learning in Implicit Generative Models. *ArXiv e-prints*, October 2016.

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011. URL https://arxiv.org/abs/1206.1901.

Neal, R. M. and Hinton, G. E. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pp. 355–368. Springer Netherlands, 1998. URL https://doi.org/10.1007/978-94-011-5014-9_12.

Nightingale, J. W., Dye, S., and Massey, R. J. AutoLens: Automated modeling of a strong lens's light, mass, and source. *Mon. Not. R. Astron. Soc.*, 478(4):4738–4784, May 2018. ISSN 0035-8711, 1365-2966. URL https://doi.org/10.1093/mnras/sty1264.

Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. Variational Bayes with synthetic likelihood. *Stat Comput*, 28(4):971–988, August 2017. ISSN 0960-3174, 1573-1375. URL https://doi.org/10.1007/s11222-017-9773-3.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pp. 1028–1036, 2016.

Papamakarios, G. and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. *arXiv preprint arXiv:1805.07226*, 2018. URL https://arxiv.org/abs/1805.07226.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pesah, A., Wehenkel, A., and Louppe, G. Recurrent machines for likelihood-free inference. *arXiv preprint arXiv:1811.12932*, 2018.

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. Population growth of human y chromosomes: A study of y chromosome microsatellites. *Mol. Biol. Evol.*, 16(12):1791–1798, December 1999. ISSN 0737-4038, 1537-1719. URL https://doi.org/10.1093/oxfordjournals.molbev.a026091.

Ritchie, D., Horsfall, P., and Goodman, N. D. Deep amortized inference for probabilistic programs. *arXiv preprint arXiv:1610.05735*, 2016.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Sjöstrand, T., Mrenna, S., and Skands, P. A brief introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178(11):852–867, June 2008. ISSN 0010-4655. URL https://doi.org/10.1016/j.cpc.2008.01.036.

Skands, P., Carrazza, S., and Rojo, J. Tuning PYTHIA 8.1: The monash 2013 tune. *Eur. Phys. J. C*, 74(8):3024, August 2014. ISSN 1434-6044, 1434-6052. URL https://doi.org/10.1140/epjc/s10052-014-3024-y.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997. URL http://www.genetics.org/content/genetics/145/2/505.full.pdf.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface.*, 6(31):187–202, July 2008. ISSN 1742-5689, 1742-5662. URL https://doi.org/10.1098/rsif.2008.0172.

Tran, D., Ranganath, R., and Blei, D. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pp. 5523–5533, 2017.

Turner, R., Hung, J., Saatci, Y., and Yosinski, J. Metropolis-Hastings generative adversarial networks. *arXiv preprint arXiv:1811.11357*, 2018.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

Wegmann, D., Leuenberger, C., and Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182 (4):1207–1218, June 2009. ISSN 0016-6731, 1943-2631. URL https://doi.org/10.1534/genetics.109.102509.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn*, 8(3-4):229–256, May 1992. ISSN 0885-6125, 1573-0565. URL https://doi.org/10.1007/bf00992696.

Wong, W., Jiang, B., Wu, T.-y., and Zheng, C. Learning summary statistic for approximate Bayesian computation via deep neural network. *STAT SINICA*, pp. 1595–1618, 2018. ISSN 1017-0405. URL https://doi.org/10.5705/ss.202015.0340.