

# Direct Amortized Likelihood Ratio Estimation

Adam D. Cobb, Brian Matejek, Daniel Elenius, Anirban Roy, Susmit Jha

Neuro-Symbolic Computing and Intelligence (NuSCI) Research Group  
Computer Science Laboratory, SRI International  
adam.cobb@sri.com

## Abstract

We introduce a new amortized likelihood ratio estimator for likelihood-free simulation-based inference (SBI). Our estimator is simple to train and estimates the likelihood ratio using a single forward pass of the neural estimator. Our approach directly computes the likelihood ratio between two competing parameter sets, which differs from the previous approach of comparing two neural network output values. We refer to our model as the direct neural ratio estimator (DNRE). As part of introducing the DNRE, we derive a corresponding Monte Carlo estimate of the posterior. We benchmark our new ratio estimator and compare it to ratio estimators in the current literature. We show that our new ratio estimator often outperforms these previous approaches. As a further contribution, we introduce a new derivative estimator for likelihood ratio estimators that enables us to compare likelihood-free Hamiltonian Monte Carlo (HMC) with random-walk Metropolis-Hastings (MH). We show that HMC is equally competitive, which has not been previously shown. Finally, we include a novel real-world application of SBI using our neural ratio estimator to design a quadcopter. Code is available at <https://github.com/SRI-CSL/dnre>.

## 1 Introduction

In many scientific applications, we rely on complex simulators to provide us with a set of observations,  $\mathbf{x}$ , for a corresponding set of parameters,  $\theta$ . Common examples range from simulating Computational Fluid Dynamics (or CFD), to flight simulators and computational biology. Significant domain expertise goes into building such simulators to determine the potential functional relationship between  $\theta$  and  $\mathbf{x}$ . However, we often do not know the likelihood function,  $p(\mathbf{x}|\theta)$ , which is the probability of the output given the input. Without the analytical form of the likelihood, we cannot easily rely on the machinery of Bayesian inference to perform model inversions. For example, given a certain desired output or observation, we might want to know which distribution of parameters could give us this output. In particular, if one were designing a cyber-physical system such as a quadcopter, then a reasonable question to ask could be: “What size should I make the fuselage such that I achieve a hover time of 20 seconds?” If we only have access to a

simulator that goes from fuselage parameters to hover time, then this inversion can be challenging. Likelihood-free inference aims to solve this problem of performing inference over simulators (or black-box models). This paradigm is often referred to as simulation-based inference (SBI) (Cranmer, Brehmer, and Louppe 2020).

Since one of the main objectives of SBI is to perform Bayesian inference to learn the posterior over the parameters of a simulator, all solutions require accounting for a lack of an analytical likelihood. They revolve around Bayes’ rule and the corresponding Bayesian inference approaches that currently exist in the literature. One approach is to directly estimate the posterior using a density estimator, which is referred to as Neural Posterior Estimation (Papamakarios and Murray 2016). NPE learns the posterior distribution using a normalizing flow (Rezende and Mohamed 2015) in a way that allows one to both evaluate and sample from  $p(\theta|\mathbf{x})$ . Unlike NPE, an alternative is Neural Likelihood Estimation (NLE) (Papamakarios, Sterratt, and Murray 2019), which trains a neural density estimator to learn the likelihood directly. The result is the estimated  $p(\mathbf{x}|\theta)$  which can be used with the prior,  $p(\theta)$  to sample from the posterior through Markov chain Monte Carlo (MCMC). Neural Ratio Estimation (NRE) is the third neural-based approach and is the one explored in this paper. Unlike the other two estimation approaches, NRE does not use a density estimator and relies on a standard neural network that aims to estimate the ratio of two likelihoods (Cranmer, Pavez, and Louppe 2015; Thomas et al. 2016; Hermans, Begy, and Louppe 2020). NRE is closely tied to MCMC in that it is used to estimate the ratio in the MH step. While there have been many papers that have focused on comparison between all aforementioned approaches, including their sequential and amortized variants (e.g. Lueckmann et al. (2021)), the focus of this paper is to explore within the existing range of available ratio estimators.

Our paper is structured as follows. In the next section we highlight related work. In Section 3 we provide preliminary theory on likelihood ratio estimation. We then introduce DNRE along with likelihood-free HMC in Section 4. Finally we perform an empirical evaluation of the performance of DNRE compared to previous ratio estimators in Section 5, leading up to a novel real-world example of designing a quadcopter. We conclude in Section 6.

**Contributions.** In this paper, we build on the work of Hermans, Begy, and Louppe (2020) and Thomas et al. (2016, 2022) by introducing a new amortized neural likelihood ratio estimator that *directly* computes the likelihood ratio between two sets of parameters and only requires a single pass through the network to achieve this estimation. We derive a corresponding new Monte Carlo estimate of the posterior distribution when using DNRE. We also introduce a new gradient estimator that can be applied to both our approach and previous approaches. This gradient estimator is more numerically stable than the previous one. We benchmark DNRE along with the baselines both with and without HMC on standard SBI tasks showing that DNRE can often outperform the baselines. These experiments also show likelihood-free HMC to be competitive, which has not been previously shown. Our final contribution is the introduction of a novel design example for quadcopters.

## 2 Related Work

Neural ratio estimation has become one of the main neural network based approaches to performing SBI, alongside NPE and NLE. One of the early works in bringing neural ratio estimation into the area of SBI was by Cranmer, Pavez, and Louppe (2015). This work showed that likelihood ratios are invariant under a specific class of dimensionality reduction maps. This led to the use of classifiers for approximating the likelihood ratio statistic. The result of this work was to perform likelihood-free inference for ratios between a freely varying parameter and a reference parameter and has since been demonstrated across multiple domains such as in high-energy physics (Baldi et al. 2016). See the corresponding code library, *carl*, for further insights (Louppe, Cranmer, and Pavez 2016).

In follow-up work (Thomas et al. 2016, 2022; Hermans, Begy, and Louppe 2020) it was highlighted that relying on a ratio estimation that requires a specific reference parameter,  $\theta_{\text{ref}}$ , can lead to scenarios where the denominator of the ratio,  $p(\mathbf{x}|\theta)/p(\mathbf{x}|\theta_{\text{ref}})$ , may not provide sufficient coverage for the numerator while exploring  $\theta$ . In particular, in the works by Thomas et al., they focus on a logistic regression model and directly show how the approach of requiring a reference parameter can cause large variances in the ratio estimation. Hermans, Begy, and Louppe (2020) also highlight this issue with results on their tractable benchmark problem in their paper. Both these works derive a likelihood-to-evidence ratio estimator as their solution to this problem. The idea being that the support of the evidence,  $p(\mathbf{x})$ , will have better coverage of a likelihood while varying the parameter,  $\theta$ . This likelihood-to-evidence ratio approach is now the one adopted within the SBI community and will be described in Section 3.

In the original work by Cranmer, Pavez, and Louppe (2015), they highlight that it might be possible to learn a classifier which is a function of both  $\theta$  and  $\theta_{\text{ref}}$ , as well as  $\mathbf{x}$ . However, in the description of this approach they write “it is not clear whether the optimal decision function can be expected for data generated from [two sets of parameters] never jointly encountered during learning” and therefore do not go with this approach. Hermans, Begy, and

Louppe (2020) also paraphrase this predicted likely outcome of jointly training pairs of parameters as being “impractical”. While these works suggest a difficulty in jointly learn pairs, neither of these papers directly explore whether this challenge exists in practice. Our work directly explores this research question and shows that it is feasible to build a ratio estimator parameterized by parameter pairs and often can outperform the likelihood-to-evidence ratio approach.

## 3 Preliminaries

The likelihood ratio between two hypotheses  $\theta$  and  $\theta_{\text{ref}}$  for an observation  $\mathbf{x}$  is given by:

$$r(\mathbf{x}|\theta, \theta_{\text{ref}}) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta_{\text{ref}})}. \quad (1)$$

Cranmer, Pavez, and Louppe (2015) showed that a change of variables allows one to learn a function,  $d(\mathbf{x})$ , to classify between samples,  $\mathbf{x} \sim p(\mathbf{x}|\theta)$ , and samples,  $\mathbf{x} \sim p(\mathbf{x}|\theta_{\text{ref}})$ , by providing the classifier with corresponding 1 and 0 labels respectively. We note that this formulation means the classifier is indirectly dependent on  $\theta$  and  $\theta_{\text{ref}}$  through the data generation process. The optimal decision function is therefore given by:

$$d^*(\mathbf{x}) = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta) + p(\mathbf{x}|\theta_{\text{ref}})}. \quad (2)$$

The likelihood ratio is then defined using the classifier,

$$r(\mathbf{x}|\theta, \theta_{\text{ref}}) = \frac{d(\mathbf{x})}{1 - d(\mathbf{x})}. \quad (3)$$

Instead of the classifier implicitly depending on  $\theta$ , an extension is to directly pass the numerator’s parameters to the classifier, while holding  $\theta_{\text{ref}}$  as a reference parameter (Cranmer, Pavez, and Louppe 2015; Baldi et al. 2016). The resulting classifier is now written as  $d(\mathbf{x}, \theta)$ , but is still indirectly dependent on the reference  $\theta_{\text{ref}}$ .

To overcome the challenge of selecting an appropriate reference hypothesis, both Thomas et al. (2016) and Hermans, Begy, and Louppe (2020) derive the likelihood-to-evidence ratio,

$$r(\mathbf{x}|\theta) = \frac{d^*(\mathbf{x}, \theta)}{1 - d^*(\mathbf{x}, \theta)} = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})p(\theta)} = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x})}, \quad (4)$$

where  $d^*(\mathbf{x}, \theta)$  still directly depends on  $\theta$ , but the denominator is now the evidence  $p(\mathbf{x})$  and therefore avoids the implicit dependence of the reference parameter. Training this classifier is where Thomas et al. (2016) and Hermans, Begy, and Louppe (2020) differ. The former requires sampling from the marginal, whereas the latter only uses samples from the joint. With the latter approach, the learned classifier is trained to distinguish between sample-parameter pairs  $(\mathbf{x}, \theta) \sim p(\mathbf{x}, \theta)$  and independent sample pairs  $(\mathbf{x}, \theta) \sim p(\mathbf{x})p(\theta)$ . As part of this approach Hermans, Begy, and Louppe (2020) sample two independent  $\theta$ ’s from the prior,  $\{\theta, \theta'\}$ , and simulate  $\mathbf{x}$  according to  $p(\mathbf{x}|\theta)$ . This gives the sample from the joint,  $(\mathbf{x}, \theta)$ , and the independent pair,  $(\mathbf{x}, \theta')$ . To avoid the additional cost of drawing a new independent sample  $\theta'$  the paper applies a shift to the existing

vector of  $\theta$ 's to take advantage of the independence between the prior samples. The parameterized classifier is trained using a binary cross entropy loss whereby pairs  $(\mathbf{x}, \theta)$  are given the label 1 and  $(\mathbf{x}, \theta')$  are given the label 0. The output of the network is an estimate of the log ratio,  $\log \hat{r}$ .

Once learned, the likelihood-to-evidence ratio estimator can also be used to estimate the posterior directly by multiplying the ratio by the prior,  $p(\theta|\mathbf{x}) \approx \hat{r}(\mathbf{x}|\theta)p(\theta)$ . However, to estimate the likelihood ratio as in Equation (1), one needs to apply two forward passes through the network with two sets of parameters to get  $r(\mathbf{x}|\theta, \theta') = r(\mathbf{x}|\theta)/r(\mathbf{x}|\theta')$ . This form can then be used to perform likelihood-free MCMC by replacing the likelihood ratio inside the MH acceptance step.

#### 4 DNRE: Direct Amortized Neural Likelihood Ratio Estimation

In this section we present our approach of Direct Amortized Neural Likelihood Ratio Estimation (DNRE), which directly parameterizes the classifier with the two parameter pairs,  $\theta$  and  $\theta'$ , as  $d_\phi(\mathbf{x}, \theta, \theta')$ . As a result we introduce the new optimal classifier,  $d^*$  as follows:

$$\begin{aligned} d^*(\mathbf{x}, \theta, \theta') &= \frac{p(\mathbf{x}, \theta)p(\theta')}{p(\mathbf{x}, \theta)p(\theta') + p(\mathbf{x}, \theta')p(\theta)} \\ &= \frac{p(\mathbf{x}|\theta)p(\theta)p(\theta')}{p(\mathbf{x}|\theta)p(\theta)p(\theta') + p(\mathbf{x}|\theta')p(\theta')p(\theta)} \\ &= \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta) + p(\mathbf{x}|\theta')}, \end{aligned} \quad (5)$$

which leads to the new direct amortized likelihood estimator:

$$r(\mathbf{x}|\theta, \theta') = \frac{d^*(\mathbf{x}, \theta, \theta')}{1 - d^*(\mathbf{x}, \theta, \theta')} = \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta')}. \quad (6)$$

A potential advantage of this approach is the additional information provided to the classifier for label  $y = 0$ . This is when the denominator,  $p(\mathbf{x}|\theta')$ , must be greater than the numerator,  $p(\mathbf{x}|\theta)$ . We force this relationship to be the case by swapping  $\theta$  and  $\theta'$  such that the parameter in the denominator becomes the one that generated the observation  $\mathbf{x}$ . This is compared to learning to distinguish between the joint distribution and independent samples which will result in a softer decision boundary for the classifier to learn.

The key difference when training the new Direct Neural Ratio Estimator (DNRE), compared to the original NRE is that during the training, we pass the set of three inputs  $(\mathbf{x}, \theta, \theta')$ , instead of the two. We ensure that the order is consistent while training using the binary cross entropy loss. In our case we assign the ordered triplet  $(\mathbf{x}, \theta, \theta')$  a label 1, and swap  $\theta$  and  $\theta'$  for the label 0. We highlight that by explicitly incorporating both sets of parameters we aim to learn the optimal direct likelihood ratio estimator  $r(\mathbf{x}|\theta, \theta')$  in Equation (6). Algorithm 1 displays the training loop for the new estimator. The algorithm follows that of (Hermans, Begy, and Louppe 2020), except for two key differences: (1) including both  $\theta$  and  $\theta'$  as estimator inputs and (2) swapping the inputs for the zero label in the binary cross entropy loss.

---

#### Algorithm 1: Optimization of $d_\phi(\mathbf{x}, \theta, \theta')$

---

$\ell$ : Criterion (BCE)  
 $p(\mathbf{x}|\theta)$ : Implicit generative model  
**Inputs:**  $p(\theta)$ : Prior  
 $N$ : Number of steps  
 $M$ : Batch-size  
**Output:**  $d_\phi(\mathbf{x}, \theta, \theta')$ : Parameterized classifier

```

1: for  $i = 1$  to  $N$  do
2:    $\theta \leftarrow \{\theta_m \sim p(\theta)\}_{m=1}^M$ 
3:    $\theta' \leftarrow \{\theta'_m \sim p(\theta)\}_{m=1}^M$ 
4:    $\mathbf{x} \leftarrow \{\mathbf{x}_m \sim p(\mathbf{x}|\theta)\}_{m=1}^M$ 
5:    $\mathcal{L} \leftarrow \ell(d_\phi(\mathbf{x}, \theta, \theta'), 1) + \ell(d_\phi(\mathbf{x}, \theta', \theta), 0)$ 
6:    $\phi \leftarrow \text{OPTIMIZER}(\phi, \nabla_\phi \mathcal{L})$ 
7: end for
8: return  $d_\phi$ 

```

---

#### Monte Carlo Posterior Approximation

Unlike the likelihood-to-evidence ratio proposed in Equation (4), the DNRE only requires one pass through the neural classifier to estimate the likelihood ratio. This means when performing MCMC, one only needs to call the classifier once per evaluation of the MH acceptance step. This direct estimation is therefore more convenient. However, a significant advantage that the original estimator has over our proposed estimator, is that the posterior distribution can be explicitly approximated by adding the log prior to the predicted log likelihood-to-evidence ratio.

To overcome the challenge of estimating the posterior using DNRE, we must integrate out the  $\theta'$  in the denominator numerically, using  $M$  samples. This requires an inverse trick that we derive in the log-space:

$$\log p(\theta|\mathbf{x}) \approx -\log \text{SumExp}\{-\log \hat{r}(\mathbf{x}|\theta, \theta'_i)\} + \log M + \log p(\theta). \quad (7)$$

The logSumExp is applied to the negated output,  $\log \hat{r}$ , of the neural network to estimate the log of the inverse of the likelihood-to-evidence ratio ( $p(\mathbf{x})/p(\mathbf{x}|\theta)$ ). The summation adds a  $\log M$  term, which is removed in the log-space. A final negation reverses the initial negation, and gives the estimated likelihood-to-evidence ratio. Then, adding the log-prior to gives the log posterior. We show this derivation in full starting from the exponentiated output of the predicted DNRE,  $r(\mathbf{x}|\theta, \theta'_i)$ :

$$\begin{aligned} \frac{1}{p(\theta|\mathbf{x})} &\approx \frac{1}{p(\theta)} \frac{1}{M} \sum_i \frac{1}{r(\mathbf{x}|\theta, \theta'_i)} \\ &= \frac{1}{p(\theta)} \frac{1}{M} \sum_i \frac{p(\mathbf{x}|\theta'_i)}{p(\mathbf{x}|\theta)} \\ &= \frac{1}{p(\theta)} \frac{p(\mathbf{x})}{p(\mathbf{x}|\theta)}. \end{aligned}$$

For higher-dimensional  $\theta'$  this is computationally intractable, but for smaller dimensional problems we can use

this form to compare posteriors in a manner consistent with previous approaches within the literature. In particular, this form allows us to analyze how the expected coverage (Hermans et al. 2021) varies with the sample size  $M$ . We will explore this in the experiments section.

### Likelihood-free Hamiltonian Monte Carlo

Although not empirically explored in Hermans, Begy, and Louppe (2020), they make a recommendation as to how one could implement Hamiltonian Monte Carlo (HMC) using a neural ratio estimator. HMC is a gradient-based Markov chain Monte Carlo sampling scheme that augments the original parameter space with additional momentum parameters,  $\mathbf{m}$ , in order to sample using Hamiltonian dynamics (Duane et al. 1987; Neal et al. 2011). HMC defines the potential energy function as  $U(\theta) = -\log[p(\mathbf{x}|\theta)p(\theta)]$  and the kinetic energy function as  $K(\mathbf{m}) = \mathbf{m}^\top \mathbf{m}/2$ . The two requirements of running an HMC algorithms are access to  $\nabla U(\theta)$  and access to the MH acceptance step of  $\rho = \min(0, -U(\theta^*) + U(\theta) - K(\mathbf{m}^*) + K(\mathbf{m}))$ , where  $\theta^*$  and  $\mathbf{m}^*$  are the proposed parameter and momenta pair. For NRE and BNRE,  $\rho$  is derived from two passes through the estimator such that  $-U(\theta^*) + U(\theta) = \log r(\mathbf{x}|\theta^*) - \log r(\mathbf{x}|\theta) + \log p(\theta^*) - \log p(\theta)$ . For our new estimator DNRE, we only require one pass such that  $-U(\theta^*) + U(\theta) = \log r(\mathbf{x}|\theta^*, \theta) + \log p(\theta^*) - \log p(\theta)$ . To estimate  $\nabla U(\theta)$ , Hermans, Begy, and Louppe (2020) use the chain rule to derive:

$$\nabla_\theta U(\theta) = -\frac{\nabla_\theta r(\mathbf{x}|\theta)}{r(\mathbf{x}|\theta)}. \quad (8)$$

This estimate of the derivative can be numerically unstable in practice as it requires exponentiating the output of the estimator for the denominator, which we have observed can often be a small number. A more simple approach is to estimate  $\nabla U(\theta)$  by treating the classifier as the approximate log ratio which allows us to separate the two terms. Therefore in the case of the NRE we can achieve the approximated derivative of the log likelihood as:

$$\begin{aligned} \nabla_\theta \log r(\mathbf{x}|\theta) &\approx \nabla_\theta \log p(\mathbf{x}|\theta) - \nabla_\theta \log p(\mathbf{x}) \\ &= \nabla_\theta \log p(\mathbf{x}|\theta). \end{aligned} \quad (9)$$

We can also follow the same approximation for our DNRE approach to get the derivative of the log likelihood:

$$\begin{aligned} \nabla_\theta \log r(\mathbf{x}|\theta, \theta') &\approx \nabla_\theta \log p(\mathbf{x}|\theta) - \nabla_\theta \log p(\mathbf{x}|\theta') \\ &= \nabla_\theta \log p(\mathbf{x}|\theta). \end{aligned} \quad (10)$$

We have found these estimators to be more stable as it avoids the exponential and division operations. For DNRE, we need to account for the  $\theta'$  when estimating the derivative. Therefore we simply choose to sample  $\theta'$  from its prior during our implementation. For estimators that closely match the true ratio, the contribution from  $\theta'$  to the derivative should tend to zero, following the logic of Equation (10).

## 5 Experiments

In this section we compare three amortized estimators: Neural Ratio Estimator (NRE) (Hermans, Begy, and Louppe

2020), Balanced Neural Ratio Estimator (BNRE) (Delaunoy et al. 2022), and our new approach of Direct Neural Ratio Estimator (DNRE). The baseline approaches are implemented using the *LAMPE* Python package (Rozet 2022). BNRE uses the same loss function as NRE, but with an additional regularization term to encourage a more balanced and conservative overall estimation of the posterior. We start with a toy example to assess the ability of modeling a simple ground truth likelihood ratio. We then explore how the Monte Carlo posterior approximation performs in practice using the two moons dataset example. Finally, we compare the performance of all three ratio estimators using a common SBI benchmark. In our final experiment we demonstrate a full scale example for the design of a quadcopter.

### Illustrative Ratio Approximation Example

Our initial comparison investigates the ability of all three approaches to fit to a simple one-dimensional hierarchical Gaussian model,

$$x \sim \mathcal{N}(\theta, \sigma^2), \quad \theta \sim \mathcal{N}(0, \sigma^2).$$

All models were trained over 1000 epochs with neural network architectures of three layers of 64 units. We use a training set of size 10,000 and a validation set of size 5,000. For comparison to the ground truth, we set the numerator to  $p(x|\theta = 0.0)$  and set the  $\theta'$  in the denominator to range from the minimum to the maximum of the training set. Figure 1 displays the results for three different standard deviations,  $[0.1, 0.3, 0.5]$ . We see that the DNRE slightly outperforms NRE, which is supported by the results of Table 4 in Appendix A, and BNRE seems to struggle when in regions where the difference between the two log likelihoods is large (greater than around two, which corresponds to a probability ratio of just over seven).

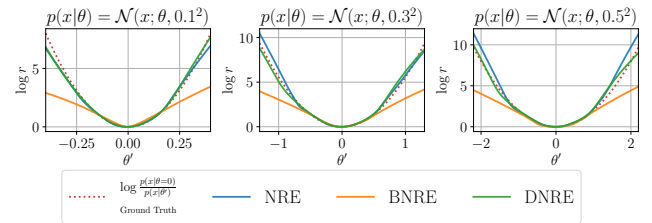


Figure 1: Comparison to ground truth log likelihood ratio for the Gaussian model while varying the denominator parameter  $\theta'$  and keeping the numerator parameter constant at  $\theta = 0.0$ . DNRE outperforms NRE, followed by BNRE.

### Posterior Approximation Performance

For the next experiment, we work with the two moons dataset, which is a typical low-dimensional benchmark within the SBI literature (Greenberg, Nonnenmacher, and Macke 2019; Lueckmann et al. 2021). We explore the quality of the estimator by analyzing our new posterior approximation from Equation (7). Figure 2 displays the expected coverage plot, whereby perfectly calibrated posteriors follow the dashed line, with more conservative posteriors

falling above this line and overconfident ones falling below (Hermans et al. 2021). Conservative estimators are generally more desirable for scientific applications. This figure highlights the importance of the number of samples needed to realize the estimated calibration of the DNRE approximated posterior. Increasing the number of these samples leads to the best performing (slightly conservative) expected coverage performance.

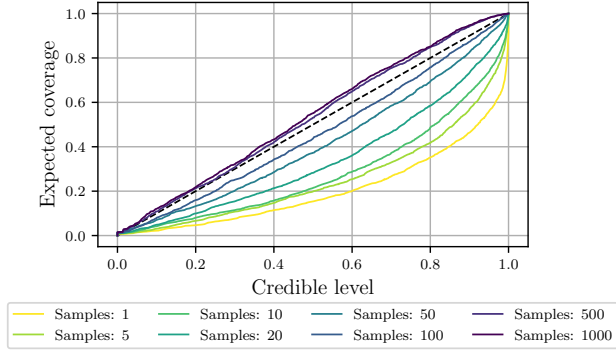


Figure 2: Expected coverage with varying sample size  $M$  for the two moons dataset. The larger the sample size, the more balanced the posterior estimate becomes.

In Figure 3 we compare the inferred posteriors of NRE, BNRE, and DNRE. We also overlay the reference samples from the ground truth conditional distribution. The main illustration of this figure is to show that our new approach is correctly approximating the posterior distribution.

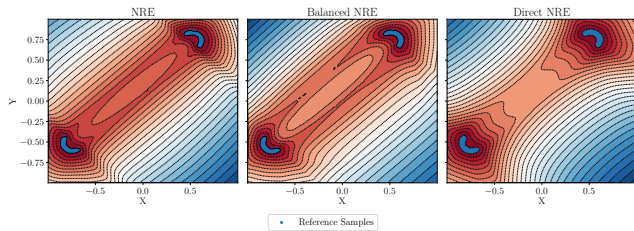


Figure 3: Contour plot comparing the three amortized estimators for the two moons dataset: Neural Ratio Estimator (NRE), Balanced Neural Ratio Estimator (BNRE), and Direct Neural Ratio Estimator (DNRE). For the DNRE, we used  $M = 20,000$ .

## Benchmark results

In this section we perform a quantitative comparison between the three ratio estimators, where we use the SBI Benchmark examples of Lueckmann et al. (2021) (see Appendix B). All estimators have an architecture of five layers of 64 units, using the Exponential Linear Unit non-linearity between layers. For all approaches we applied the same grid search over both the learning rate and standard deviation of the proposal distribution for the random-walk MH sampling scheme. The best model for each task was selected

according to the C2ST score applied for 1,000 samples using the first observation of the benchmark. The C2ST score is a classifier-based test trained to distinguish between reference posterior samples and samples from the likelihood-free inference (see Appendix B). The results in Table 1 are averaged across the 10 available observations in the benchmark. A score of 0.5 is optimal, whereas 1.0 would be the worst score. While the C2ST metric cannot capture all the key indicators of a good estimator, these results do indicate the promise of DNRE since it outperforms the baselines. We see that DNRE achieves the best score in five out of the ten benchmarks, followed by NRE with four and BNRE with one.

## Comparison to HMC

In addition to comparing the results for all ratio estimators across the benchmark examples using random-walk MH, we also run HMC using the gradient estimators shown in Equations (9) and (10). We display these results in Table 2. As with the random-walk MH sampler, we perform grid search using the same setup, except we now additionally vary the trajectory length instead of the variance of the proposal distribution. Since HMC can be notoriously difficult to tune, we resort to extending the grid search to include desired acceptance rates ranging between 0.5 to 0.8. This uses the dual averaging scheme as introduced in Hoffman, Gelman et al. (2014) and as implemented in Cobb et al. (2019).<sup>1</sup> We use the same number of parallel chains; same number of collected samples; and the same thinning of each chain as for the random-walk MH sampler. Again, we see that DNRE achieves the lowest score in five of the ten experiments, however these experiments are interestingly not for the same experiments as in Table 1.

We can also compare inference through HMC with random-walk MH, since it is useful to establish whether either sampling approach is in general more useful on this benchmark. Table 3 displays the mean C2ST of the MH results minus the mean C2ST of the HMC results. When HMC has a lower (better) C2ST, we highlight this in orange, and we do the same for MH by highlighting when it is better using blue. A potentially useful result of this table is that there are clear winners for certain tasks. For the two moons and the SLCP Distractors task, random-walk MH outperforms across all estimators, whereas for Gaussian linear and Bernoulli GLM, HMC outperforms across all estimators. This observation could imply that certain SBI tasks are suited to a particular sampling scheme. Additionally, there are thirteen instances of each sampling scheme performing best, further suggesting that both HMC and MH samplers should be considered for all future SBI approaches since we cannot conclude from these results that either sampling scheme is better.

## Estimated Posterior Under Ground Truth $\theta$

The benchmark provided by Lueckmann et al. (2021) also provides the ground truth,  $\theta^*$ , for each observation. As a re-

<sup>1</sup>We additionally average across all chains at each step, which results in reliable desired acceptance rates.

Approach	TM	GL	LV	SIR	SLCP
NRE	$0.559 \pm 0.026$	<b><math>0.533 \pm 0.015</math></b>	<b><math>0.995 \pm 0.004</math></b>	<b><math>0.803 \pm 0.083</math></b>	$0.925 \pm 0.039$
BNRE	<b><math>0.544 \pm 0.033</math></b>	$0.59 \pm 0.019$	$0.997 \pm 0.002$	$0.944 \pm 0.029$	$0.893 \pm 0.038$
DNRE	$0.587 \pm 0.027$	$0.576 \pm 0.034$	$0.998 \pm 0.002$	$0.871 \pm 0.064$	<b><math>0.826 \pm 0.078</math></b>
	GLU	SLCP D	B GLM	GM	B GLM R
NRE	$0.618 \pm 0.025$	$0.982 \pm 0.008$	<b><math>0.781 \pm 0.046</math></b>	$0.751 \pm 0.015$	$0.819 \pm 0.035$
BNRE	$0.607 \pm 0.021$	$0.984 \pm 0.007$	$0.807 \pm 0.039$	$0.755 \pm 0.016$	$0.862 \pm 0.047$
DNRE	<b><math>0.597 \pm 0.025</math></b>	<b><math>0.98 \pm 0.011</math></b>	$0.813 \pm 0.069$	<b><math>0.747 \pm 0.015</math></b>	<b><math>0.777 \pm 0.062</math></b>

Table 1: C2ST SBI Metropolis-Hastings Benchmark Results.

Approach	TM	GL	LV	SIR	SLCP
NRE	$0.657 \pm 0.026$	$0.531 \pm 0.015$	<b><math>0.996 \pm 0.005</math></b>	<b><math>0.803 \pm 0.082</math></b>	$0.913 \pm 0.042$
BNRE	<b><math>0.562 \pm 0.07</math></b>	$0.543 \pm 0.019$	$0.997 \pm 0.003$	$0.945 \pm 0.029$	<b><math>0.891 \pm 0.039</math></b>
DNRE	$0.753 \pm 0.14$	<b><math>0.519 \pm 0.006</math></b>	$0.998 \pm 0.001$	$0.891 \pm 0.043$	$0.892 \pm 0.098$
	GLU	SLCP D	B GLM	GM	B GLM R
NRE	$0.613 \pm 0.025$	$0.995 \pm 0.003$	$0.776 \pm 0.044$	$0.752 \pm 0.016$	$0.83 \pm 0.044$
BNRE	<b><math>0.598 \pm 0.025</math></b>	$0.991 \pm 0.009$	$0.788 \pm 0.036$	$0.751 \pm 0.015$	$0.86 \pm 0.047$
DNRE	$0.665 \pm 0.162$	<b><math>0.989 \pm 0.008</math></b>	<b><math>0.738 \pm 0.038</math></b>	<b><math>0.747 \pm 0.013</math></b>	<b><math>0.77 \pm 0.109</math></b>

Table 2: C2ST SBI HMC Benchmark Results.

Approach	TM	GL	LV	SIR	SLCP	GLU	SLCP D	B GLM	GM	B GLM R
NRE	-0.098	0.002	-0.001	0.000	0.012	0.005	-0.013	0.005	-0.001	-0.011
BNRE	-0.018	0.047	0.000	-0.001	0.002	0.009	-0.007	0.019	0.004	0.002
DNRE	-0.166	0.057	0.000	-0.020	-0.066	-0.068	-0.009	0.075	0.000	0.007

Table 3: C2ST SBI comparison between Table 1 and Table 2. The results show the means of the MH minus the HMC means. Better performance by HMC is given by negative values. Better performance by MH is given by positive values.

sult it is possible to report the posterior probability of  $\theta^*$  according to each estimator. Table 5 in Appendix B displays the average log posterior probabilities across all observations for each estimator. We see that DNRE tends to assign the highest posterior probability when averaging over all the ground truth observation and parameter pairs.

### Quadcopter Design

For our final experiment we perform SBI for aircraft design. In particular, we run a custom flight dynamics pipeline which combines CAD software (PTC 2022) with a computational flight simulation model (Walker et al. 2022; Bapty et al. 2022; Cobb et al. 2023). To use the pipeline, one can define an aircraft design and allow the simulator to determine performances such as its drag, lift, maximum flight time and hover time. We want to explore how we should parameterize a quadcopter such that we achieve a certain performance level. The observations,  $\mathbf{x} \in \mathbb{R}^7$ , consist of: the number of interferences; the mass; the maximum flight distance; the maximum hover time; the maximum lateral speed; the maximum control input at the maximum flight distance; and the maximum power at the maximum speed. The design parameters,  $\theta \in \mathbb{R}^{19}$ , consist of: the arm length; four fuselage shape parameters; and seven ‘x’ and ‘y’ locations

of electrical devices inside the fuselage. We highlight some of the key parameters in Figure 4.

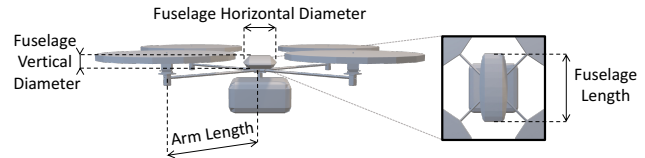


Figure 4: Schematic of the quadcopter design which highlights the key design parameters related to the fuselage and the arm length.

**Training** We train all three neural ratio estimators using a five layer fully connected architecture with 128 hidden units per layer. We use a learning rate of 0.001 and train all models for 2,000 epochs using a batch size of 512 with 3,592 training points and 898 validation points.

**Design Specification** SBI is useful tool for design when using black-box models to simulate performance. In this case the sole objective is to perform Bayesian inference over the simulation model, such that we can specify objectives



(=  $\mathbf{x}$ ) to sample designs and evaluate their suitability. Here we specify a design objective of:

Interferences	0	Flight Distance (m)	500.0
Mass (Kg)	3	Hover Time (s)	20.0
Speed (m/s)	30.0	Power at Max. Speed (W)	2000.0
Control Input	0.7		

To see how this design objective compares to existing designs in the training data, in the supplementary materials we include Figure 7 which displays the empirical distribution of the observations along with this specific design objective superimposed. Notably, this is a particularly demanding specification when looking at the need of a low mass quadcopter with a high hover time. Figure 5 displays the result of performing likelihood-free inference using random walk MH with the DNRE approach. We also highlight the highest ranked design, according to DNRE, from a randomly sampled test set of 1,059 quadcopters. Interestingly, we see that designs meeting our specification come from a distribution that require long arm lengths and a minimal sized fuselage. We hypothesize that our zero interference requirement has led longer arm lengths to prevent the propellers intersecting with other components. The choice of a small fuselage (short diameters and short length) will help reduce the mass since the average mass of 5.38 Kg is much higher than our desired design choice. Appendix C includes the comparable corner plots for NRE and BNRE.

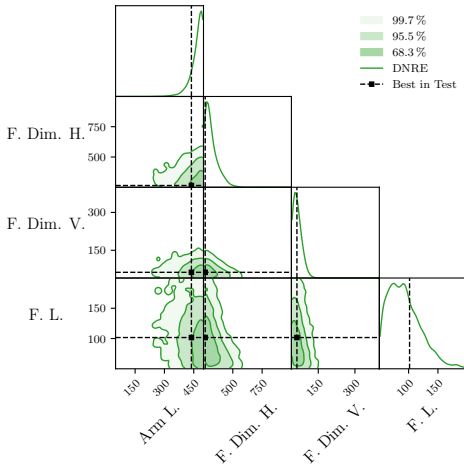


Figure 5: Corner plot displaying materialized samples from likelihood-free inference using DNRE for a subset of the parameters as highlighted in Figure 4. The black square corresponds to the highest ranked test design in the test set. The high density on the longest arm length combined with parameters leading to a small fuselage likely follow the objective of reducing interferences between the propellers and reducing the overall mass.

**Ranking** We can use the estimated likelihood ratio to perform a model comparison for different  $\theta$ 's. In particular we can use all three ratio estimators to compare all the designs in the test set. We look to find the top 100 designs according to each likelihood estimator and check for consistency

across estimators. The following ranking matrix displays the number of overlapping test samples in both the top 100 (highlighted in bold) and the top 10 out of 1,059 data points:

Overlap: <b>Top-100</b> , Top-10			
	NRE	BNRE	DNRE
NRE	-	<b>93</b>	<b>89</b>
BNRE	9	-	<b>87</b>
DNRE	8	8	-

All estimators have a high overlap or design rankings for both top 10 and top 100 suggesting that all approaches are valid. Furthermore, there is a clear 'winner' in the test set since all propose the same top performing design which is the one superimposed on Figure 5.

**Improving an Existing Design** We demonstrate the utility of performing likelihood-free HMC with DNRE by taking a promising existing design and improving it by reducing structural interferences. In Appendix C we include the full case study that demonstrates how we can use our newly derived gradient-based sampling scheme with our proposed direct neural ratio estimator to evolve a seed design from 60 structural interferences to 4. This is while successfully preserving the other original design performance metrics as evaluated through the flight simulation software. Figure 6 displays sub-sampled quadcopters taken along the Markov chain.



Figure 6: Sub-sampled quadcopter designs taken along the likelihood-free HMC chain using DNRE. The initial seed design on the far left has multiple structural interferences, including sensors that cut through the fuselage. As we move along the chain we see the design morph into our desired structure with very few interferences. This is achieved through increasing the arm length and changing the shape of the fuselage, as well as varying the placement of the interior sensing components.

## 6 Conclusion

The contribution of this paper is to demonstrate that directly learning to approximate the likelihood ratio between two pairs of parameters presents itself as a viable option for likelihood-free inference and often outperforms competing approaches on standard SBI benchmarks. As part of our contribution, we derive a new Monte Carlo estimator for the posterior distribution when using our DNRE approach. We also derive a simple likelihood gradient estimator that can be successfully used to perform HMC. We are therefore the first to compare random walk MH with HMC for likelihood ratio estimation approaches. We find that HMC is a viable MCMC approach and can outperform random-walk MH. Finally, we introduce a novel application of SBI for the design of a quadcopter and we release this data as part of the supplementary materials.

## Acknowledgements

This material is based upon work supported by the United States Air Force and DARPA under Contract No. FA8750-20-C-0002 and FA8750-23-C-0519. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.

## References

- Baldi, P.; Cranmer, K.; Faucett, T.; Sadowski, P.; and Whiteson, D. 2016. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76: 1–7.
- Bapty, T.; Whittington, S.; Walker, J. D.; Hite, J.; Swenson, B.; Owens, K.; Eisele, F.; Scott, J.; and Owens, R. 2022. Design Oracle for AI-Based CPS Design. *Design Automation for CPS and IoT (DESTION)*.
- Cobb, A. D.; Baydin, A. G.; Markham, A.; and Roberts, S. J. 2019. Introducing an Explicit Symplectic Integration Scheme for Riemannian Manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:1910.06243*.
- Cobb, A. D.; Roy, A.; Elenius, D.; Heim, F. M.; Swenson, B.; Whittington, S.; Walker, J. D.; Bapty, T.; Hite, J.; Ramani, K.; McComb, C.; and Jha, S. 2023. AircraftVerse: A Large-Scale Multimodal Dataset of Aerial Vehicle Designs. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Cranmer, K.; Brehmer, J.; and Louppe, G. 2020. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48): 30055–30062.
- Cranmer, K.; Pavez, J.; and Louppe, G. 2015. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*.
- Delaunoy, A.; Hermans, J.; Rozet, F.; Wehenkel, A.; and Louppe, G. 2022. Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation. *arXiv preprint arXiv:2208.13624*.
- Duane, S.; Kennedy, A. D.; Pendleton, B. J.; and Roweth, D. 1987. Hybrid Monte Carlo. *Physics letters B*, 195(2): 216–222.
- Greenberg, D.; Nonnenmacher, M.; and Macke, J. 2019. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2404–2414. PMLR.
- Hermans, J.; Begy, V.; and Louppe, G. 2020. Likelihood-free MCMC with amortized approximate ratio estimators. In *International conference on machine learning*, 4239–4248. PMLR.
- Hermans, J.; Delaunoy, A.; Rozet, F.; Wehenkel, A.; and Louppe, G. 2021. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*.
- Hoffman, M. D.; Gelman, A.; et al. 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623.
- Louppe, G.; Cranmer, K.; and Pavez, J. 2016. carl: a likelihood-free inference toolbox.
- Lueckmann, J.-M.; Boelts, J.; Greenberg, D.; Goncalves, P.; and Macke, J. 2021. Benchmarking Simulation-Based Inference. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 343–351. PMLR.
- Neal, R. M.; et al. 2011. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11): 2.
- Papamakarios, G.; and Murray, I. 2016. Fast  $\varepsilon$ -free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29.
- Papamakarios, G.; Sterratt, D.; and Murray, I. 2019. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 837–848. PMLR.
- PTC. 2022. Creo CAD Software. <https://www.ptc.com/en/products/creo>. Accessed: 2023-11-01.
- Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International conference on machine learning*, 1530–1538. PMLR.
- Rozet, F. 2022. LAMPE. <https://github.com/francois-rozet/lampe>. Accessed: 2024-01-08.
- Thomas, O.; Dutta, R.; Corander, J.; Kaski, S.; and Gutmann, M. U. 2016. Likelihood-free inference by ratio estimation. *arXiv preprint arXiv:1611.10242*.
- Thomas, O.; Dutta, R.; Corander, J.; Kaski, S.; and Gutmann, M. U. 2022. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1): 1–31.
- Walker, J. D.; Heim, F. M.; Surampudi, B.; Bueno, P.; Carpenter, A.; Chocron, S.; Cutshall, J.; Lammons, R.; Bapty, T.; Swenson, B.; and Whittington, S. 2022. A Flight Dynamics Model for Exploring the Distributed Electrical eVTOL Cyber Physical Design Space. *Design Automation for CPS and IoT (DESTION)*.