Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation

Arnaud Delaunoy[*]
University of Liège
a.delaunoy@uliege.be

Joeri Hermans[*]
Unaffiliated
joeri@peinser.com

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

Antoine Wehenkel
University of Liège
antoine.wehenkel@uliege.be

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

Abstract

Modern approaches for simulation-based inference rely upon deep learning surrogates to enable approximate inference with computer simulators. In practice, the estimated posteriors' computational faithfulness is, however, rarely guaranteed. For example, Hermans et al. [1] show that current simulation-based inference algorithms can produce posteriors that are overconfident, hence risking false inferences. In this work, we introduce Balanced Neural Ratio Estimation (BNRE), a variation of the NRE algorithm [2] designed to produce posterior approximations that tend to be more conservative, hence improving their reliability, while sharing the same Bayes optimal solution. We achieve this by enforcing a balancing condition that increases the quantified uncertainty in small simulation budget regimes while still converging to the exact posterior as the budget increases. We provide theoretical arguments showing that BNRE tends to produce posterior surrogates that are more conservative than NRE's. We evaluate BNRE on a wide variety of tasks and show that it produces conservative posterior surrogates on all tested benchmarks and simulation budgets. Finally, we emphasize that BNRE is straightforward to implement over NRE and does not introduce any computational overhead.

1 Introduction

Many areas of science and engineering use parametric computer simulations to describe complex stochastic generative processes. In this setting, Bayesian inference provides a principled framework to identify parameters matching empirical observations. Computer simulations, however, define the necessary likelihood function only implicitly, which prevents its evaluation and the use of classical inference algorithms. To overcome this obstacle, recent simulation-based inference (SBI) algorithms [3] build upon deep learning surrogates to approximate parts of the Bayes rule and enable approximate inference. For example, [4, 5] build a surrogate of the likelihood function while [6, 7, 2, 8, 9] approximate the likelihood-to-evidence ratio. The posterior can also be targeted directly with variational inference, as proposed by [10, 11, 5]. These algorithms are either amortized or run sequentially to drive the training towards a target observation and improve the simulation efficiency of the procedure [10, 12, 2, 11, 4, 8, 5]. However, sequential methods have the drawback of being computationally expensive to diagnose as the surrogates are only valid for the target observation [1]. Truncated marginal neural ratio estimation [9] alleviates this issue by introducing a sequential algorithm that builds a surrogate valid in a local region around the target.

[*]Equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Since modern simulation-based inference algorithms rely on deep learning surrogates, concerns naturally arise regarding their computational faithfulness and whether they are sufficiently adequate for the inference task of interest. In Bayesian inference, these concerns can be at least partially addressed with diagnostics designed to probe the correct behaviour of the inference method, such as $\hat{R}$ diagnostics for MCMC [13], or to assess the quality of posterior approximations directly. The latter include diagnostics such as simulation-based calibration (SBC) [14] or coverage-based diagnostics [15, 1]. As discussed by Hermans et al. [1], posterior approximations must be conservative to guarantee reliable inferences, even when approximations are not faithful. For example, in the physical sciences, where the goal is often to constrain parameters of interest, wrongly excluding plausible values could drive the scientific inquiry in the wrong direction, whereas failing to exclude implausible values because of (too) conservative estimations is much less detrimental. Unfortunately, the same authors also demonstrate that current simulation-based inference algorithms can lead to overconfident surrogates and therefore false inferences.

Scientific use cases requiring conservative inference include for example the study of dark matter models in particle physics and astrophysics [16], which could be cold, warm, or hot dark matter. In general, thermal dark matter models are described by a single parameter, the dark matter thermal relic mass, which can be intuitively thought of as the energy the dark matter particle had in the Early Universe. Small values correspond to warm or hot dark matter, while high values are descriptive of cold dark matter. Applying an inference algorithm without diagnosing the learned estimator could lead to constraints that are tighter than they should be. For example, whenever an overconfident estimator produces posterior estimates that favor cold dark matter models, it could simultaneously reject alternative models, such as the extensively studied Sterile Neutrino, a potential candidate for the Warm Dark Matter particle. Making a scientific statement in this direction therefore requires the uttermost care to not wrongly exclude values of the thermal relic mass that are actually plausible.

In this work, we develop a novel algorithm that not only converges to exact inference as the simulation budget increases, but which is also more likely to produce conservative surrogates in small simulation budget regimes. Towards this objective, we propose a variant of the NRE algorithm called Balanced Neural Ratio Estimation (BNRE), which enforces a balancing condition on the binary neural classifier to increase the reliability of its posterior approximations.

The structure of the manuscript is outlined as follows. Section 2 describes the formalism and the necessary background. Section 3 describes BNRE and provides theoretical arguments towards its conservativeness and reliability. Section 4 illustrates our main results and provides insights regarding the behaviour of the method. Finally, Section 5 discusses related work while Section 6 summarizes our contributions and hints at future work.

2 Background

2.1 Statistical formalism

This work is concerned with simulation-based inference algorithms that produce posterior approximations $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ under the following semantics. **Target parameters** $\boldsymbol{\vartheta}$ denote the parameters of the model and we make the reasonable assumption that the prior $p(\boldsymbol{\vartheta})$ is tractable. The model is generically expressed as a computer program, a simulator, that describes the forward dynamics of interest based on the input parameters $\boldsymbol{\vartheta}$. The simulator implicitly defines the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. While we cannot directly evaluate the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$, we can execute the computer program to generate synthetic **observables** $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. Every observable $\boldsymbol{x}_o$ is tied to **ground truth** parameters $\boldsymbol{\vartheta}^*$ whose forward evaluation within the simulator produced $\boldsymbol{x}^*$.

Of special importance to Bayesians is the notion of a **credible region**, which is a domain Θ within the target parameter space that satisfies $\int_\Theta p(\boldsymbol{\vartheta} \mid \boldsymbol{x} = \boldsymbol{x}^*)\, \mathrm{d}\boldsymbol{\vartheta} = 1 - \alpha$ for some observable $\boldsymbol{x}^*$ and confidence level $1 - \alpha$. Because many such regions exist, we target the credible region with the smallest volume, also known as the highest posterior density region [17, 18].

2.2 Neural ratio estimation

Neural Ratio Estimation (NRE) is an established approach in the simulation-based inference literature both from frequentist [6] and Bayesian [7, 2, 8, 9] perspectives. In essence, all protocols rely on

the density-ratio trick [19, 20, 6] to construct a surrogate of the likelihood ratio. In this work, we consider an amortized estimator $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ of the intractable likelihood-to-evidence ratio $r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = p(\boldsymbol{\vartheta}, \boldsymbol{x})/p(\boldsymbol{\vartheta})p(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\vartheta})/p(\boldsymbol{x})$ that can be learned by training a binary classifier $\hat{d} : \mathbf{X} \times \boldsymbol{\Theta} \mapsto [0, 1]$ to distinguish between samples of the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ with class label 1 and samples of the product of marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ with class label 0, with equal label marginal probability. For the binary cross-entropy loss, the Bayes optimal classifier is

$$d(\boldsymbol{\vartheta}, \boldsymbol{x}) = \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} = \sigma\left(\log \frac{p(\boldsymbol{\vartheta}, \boldsymbol{x})}{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. Given target parameters $\boldsymbol{\vartheta}$ and an observable $\boldsymbol{x}$ supported by $p(\boldsymbol{\vartheta})$ and $p(\boldsymbol{x})$ respectively, the learned classifier $\hat{d}$ provides an approximation for the log likelihood-to-evidence ratio $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ because $\log r(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \text{logit}(d(\boldsymbol{\vartheta}, \boldsymbol{x})) \approx \text{logit}(\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})) = \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. The log posterior density function is approximated as $\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = \log p(\boldsymbol{\vartheta}) + \log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$.

3 Balanced binary classification for neural ratio estimation

Following Hermans et al. [1], let us first define the **expected coverage probability** of the $1 - \alpha$ highest posterior density regions derived from the posterior estimator $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ as

$$\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}\left[\mathbb{1}\left(\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1 - \alpha)\right)\right], \tag{2}$$

where the function $\Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1 - \alpha)$ yields the $1 - \alpha$ highest posterior density region of $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$. This diagnostic probes the conservativeness of the posterior estimator (or the lack thereof) and can be interpreted as the expected frequentist coverage $\mathbb{E}_{p(\boldsymbol{\vartheta})}\mathbb{E}_{p(\boldsymbol{x} \mid \boldsymbol{\vartheta})}\left[\mathbb{1}\left(\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}(1 - \alpha)\right)\right]$.

In this work, a posterior estimator has coverage at the confidence level $1 - \alpha$ whenever the expected coverage probability is larger or equal to the nominal coverage probability, $1 - \alpha$. We say that a posterior estimator is **conservative** when it has coverage for all confidence levels. The expected coverage probability can be plotted for various levels α, which allows to visually identify conservative posterior estimators. The expected coverage can also be shown to be a special case of the SBC diagnostic [14] (see Appendix A), further motivating the usage of expected coverage.

Our main objective is to restrict the hypothesis space of the approximate classifiers $\hat{d}$ to those leading to conservative posterior estimators, hence solving the reliability concerns of NRE. Towards this goal, we construct a hypothesis space of **balanced classifiers** and show both theoretically and empirically that they lead to posterior estimators that tend to be more conservative.

3.1 Balanced binary classification

Definition 1. *A classifier $\hat{d}$ is balanced if* $\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}\left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right] = \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right]$, *or*

$$\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}\left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right] + \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right] = 1. \tag{3}$$

Theorem 1. *Any balanced classifier $\hat{d}$ satisfies* $\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}\left[\dfrac{d(\boldsymbol{\vartheta}, \boldsymbol{x})}{\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}\right] \geq 1.$

Proof. The integral form of the balancing condition

$$\iint \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{x} = 1 \tag{4}$$

implies that $\big(p(\boldsymbol{x}, \boldsymbol{\vartheta}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ is a valid density, both integrating to 1 and positive everywhere. Therefore, its Kullback-Leibler (KL) divergence with $p(\boldsymbol{\vartheta}, \boldsymbol{x})$ is positive. Through

Jensen's inequality, we obtain

$$
\begin{aligned}
0 &\leq \mathrm{KL}\left(p(\boldsymbol{\vartheta},\boldsymbol{x}) \,\middle|\middle|\, \big(p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})\right) \\
&\leq \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{\vartheta},\boldsymbol{x})}{\big(p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})}\right] \\
&\leq \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \frac{d(\boldsymbol{\vartheta},\boldsymbol{x})}{\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})}\right] \\
\Rightarrow \quad 1 &\leq \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\exp\left(\log \frac{d(\boldsymbol{\vartheta},\boldsymbol{x})}{\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})}\right)\right] = \mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\frac{d(\boldsymbol{\vartheta},\boldsymbol{x})}{\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})}\right]. \qquad \square
\end{aligned}
$$

Theorem 2. *Any balanced classifier $\hat{d}$ satisfies $\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\dfrac{1-d(\boldsymbol{\vartheta},\boldsymbol{x})}{1-\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})}\right] \geq 1$.*

Proof. Similar to Theorem 1, see Appendix B. $\square$

Theorem 1 shows that, in expectation over the joint distribution $p(\boldsymbol{\vartheta},\boldsymbol{x})$, a balanced classifier $\hat{d}$ tends to make predictions whose probability values $\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})$ are smaller than the exact probability values $d(\boldsymbol{\vartheta},\boldsymbol{x})$. In other words, a balanced classifier $\hat{d}$ tends to be less confident than the Bayes optimal classifier d. Similarly, Theorem 2 shows that, in expectation over the product of the marginals $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$, a balanced classifier tends to make predictions whose probability values $1-\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x})$ are smaller than the exact probability values $1-d(\boldsymbol{\vartheta},\boldsymbol{x})$, hence showing that a balanced classifier $\hat{d}$ tends to also be less confident than the Bayes optimal classifier d. We note however that these two theorems hold only in expectation, which implies that neither $\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x}) \leq d(\boldsymbol{\vartheta},\boldsymbol{x})$ for all $\boldsymbol{\vartheta},\boldsymbol{x}$ nor $1-\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x}) \leq 1-d(\boldsymbol{\vartheta},\boldsymbol{x})$ for all $\boldsymbol{\vartheta},\boldsymbol{x}$ can generally be guaranteed.

Theorem 3. *The Bayes optimal classifier $d(\boldsymbol{\vartheta},\boldsymbol{x})$ is balanced.*

Proof. Replacing the Bayes optimal classifier

$$
d(\boldsymbol{\vartheta},\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{\vartheta},\boldsymbol{x})}{p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \tag{5}
$$

in the integral form of the balancing condition, we have

$$
\begin{aligned}
&\iint \big(p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)d(\boldsymbol{\vartheta},\boldsymbol{x})\,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{x} \\
&= \iint \frac{\big(p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\,p(\boldsymbol{\vartheta},\boldsymbol{x})}{p(\boldsymbol{\vartheta},\boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{x} \\
&= \iint p(\boldsymbol{\vartheta},\boldsymbol{x})\,\mathrm{d}\boldsymbol{\vartheta}\,\mathrm{d}\boldsymbol{x} = 1. \qquad \square
\end{aligned}
$$

Theorem 3 states that the Bayes optimal classifier is balanced. Therefore, **minimizing the cross-entropy loss while restricting the model hypothesis space to balanced classifiers results in the same Bayes optimal classifier of Eqn. 1.**

3.2 Balanced neural ratio estimation

We now extend the NRE algorithm to enforce the balancing condition. The previous results show that enforcing the condition should result in more conservative classifiers $\hat{d}$ and therefore to dispersed posterior approximations. Let us first note that Theorem 1 can be expressed as $\mathbb{E}_{p(\boldsymbol{x})}[\mathbb{E}_{p(\boldsymbol{\vartheta}\,|\,\boldsymbol{x})}[d(\boldsymbol{\vartheta},\boldsymbol{x})/\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x}]] \geq 1$, which can (ideally) be achieved when the inner expectation is larger than 1 for all $\boldsymbol{x}$. In this case, the classifier $\hat{d}$ will be such that $\hat{d}(\boldsymbol{\vartheta},\boldsymbol{x}) \leq d(\boldsymbol{\vartheta},\boldsymbol{x})$ in

regions of high posterior density. Then,

$$\frac{\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})} \leq \frac{d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}, \text{ which is equivalent to } \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}) \leq r(\boldsymbol{x} \mid \boldsymbol{\vartheta}), \tag{6}$$

and $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ since $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = p(\boldsymbol{\vartheta})\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta})$. Similarly, Theorem 2 implies $1 - d(\boldsymbol{\vartheta}, \boldsymbol{x}) \geq 1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})$ in regions of high prior density, which results in $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$. Between those two opposite effects, the constraint on $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ that will dominate depends on whether $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) > p(\boldsymbol{\vartheta})$ or $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) < p(\boldsymbol{\vartheta})$. If $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) > p(\boldsymbol{\vartheta})$, then $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$, whereas if $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) < p(\boldsymbol{\vartheta})$ then $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \leq \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$. Overall, imposing the balancing condition will therefore result in approximate posteriors that lie between the prior and the exact posterior, without being more confident than they should.

Practically, the balancing condition can be targeted through a regularization penalty. For the binary cross-entropy $\mathcal{L}[\hat{d}] \triangleq -\mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})}[\log \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}[\log(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}))]$ and given that the balancing condition only depends on samples from $p(\boldsymbol{x})p(\boldsymbol{\vartheta})$ and $p(\boldsymbol{x}, \boldsymbol{\vartheta})$, the full loss functional including the balancing condition can be expressed as

$$\mathcal{L}_b \left[\hat{d}\right] \triangleq \mathcal{L} \left[\hat{d}\right] + \lambda \left(\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right] + \mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\right] - 1\right)^2, \tag{7}$$

where λ is a (scalar) hyper-parameter controlling the strength of the balancing condition's contribution. The training procedure is summarized in Algorithm 1. Since a classifier is balanced if the balancing condition cancels out, λ could, in principle, be set arbitrarily large. However, as the balancing condition is estimated via Monte Carlo sampling, setting λ to a large value could impair the classifier's learning ability. We found that $\lambda = 100$ works well across many problem domains with varying simulation budgets.

Algorithm 1 Training algorithm for Balanced Neural Ratio Estimation (BNRE).

Inputs:	Implicit generative model $p(\boldsymbol{x} \mid \boldsymbol{\vartheta})$ (simulator) and prior $p(\boldsymbol{\vartheta})$
Outputs:	Approximate classifier $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$ parameterized by ψ
hyper-parameters:	Balancing condition strength λ (default = 100) and batch-size n

repeat

Sample data from the joint $\{\boldsymbol{\vartheta}_i,\ \boldsymbol{x}_i \sim p(\boldsymbol{\vartheta}, \boldsymbol{x}),\ y_i = 1\}_{i=1}^{n/2}$

Sample data from the marginals $\{\boldsymbol{\vartheta}_i,\ \boldsymbol{x}_i \sim p(\boldsymbol{\vartheta})p(\boldsymbol{x}),\ y_i = 0\}_{i=n/2+1}^{n}$

$\mathcal{L}[\hat{d}_\psi] = -\frac{1}{n} \sum_{i=1}^{n} y_i \log \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i) + (1 - y_i) \log(1 - \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i))$

$\mathcal{B}[\hat{d}_\psi] = \frac{2}{n} \sum_{i=1}^{n/2} \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i) + \frac{2}{n} \sum_{i=n/2+1}^{n} \hat{d}_\psi(\boldsymbol{\vartheta}_i, \boldsymbol{x}_i)$

$\psi = \texttt{minimizer\_step}(\texttt{params=}\psi,\ \texttt{loss=}\mathcal{L}[\hat{d}_\psi] + \lambda(\mathcal{B}[\hat{d}_\psi] - 1)^2)$

until convergence

return $\hat{d}_\psi(\boldsymbol{\vartheta}, \boldsymbol{x})$.

4 Experiments

We start by providing an extensive validation of BNRE on a broad range of benchmarks demonstrating that the proposed method alleviates the problem. Section 4.2 follows up with an illustrative demonstration on the behaviour of BNRE and its hyper-parameters. Code is available at `https://github.com/montefiore-ai/balanced-nre`.

4.1 Extensive validation

Setup We evaluate the expected coverage of posterior estimators produced by both NRE and BNRE on various problems. Those benchmarks cover a diverse set of problems from particle physics (Weinberg), epidemiology (Spatial SIR), queueing theory (M/G/1), population dynamics (Lotka Volterra, and astronomy (Gravitational Waves). They are representative of real scientific applications of simulation-based inference. A more detailed description of the benchmarks can be found in Appendix C. The architectures and hyper-parameters used for each problem are defined in Appendix

D. Our evaluation considers simulation budgets of increasing size, ranging from $2^{10} = 1024$ to $2^{17} = 131,072$ samples, and credibility levels from 0.05 to 0.95. For every simulation budget, we train 5 posterior estimators for 500 epochs and determine the credible region by evaluating the approximated posterior density function in a discretized and empirically normalized grid of the parameter space with sufficient resolution. The subsequent credible region is the set of parameters whose estimated (and normalized) posterior density is higher or equal to an inclusion threshold fitted to obtain the desired credibility level $1 - \alpha$. Details on this procedure are described in Appendix E. The expected coverage probability is estimated on 10000 unseen samples from the joint $p(\boldsymbol{\vartheta}, \boldsymbol{x})$, for each considered credibility level.

Expected coverage The expected coverage curves and their interpretation are detailed in Figure 1. We observe that NRE often produces posterior estimators that are overconfident, especially for small simulation budgets. However, NRE's reliability increases with the availability of training data. By contrast, **BNRE produces posterior estimators that are conservative on all benchmarks for all simulation budgets**. Figure 2 explores the same phenomena through a quantity which we call the coverage AUC, highlighting the effect of the simulation budget. Coverage AUC corresponds to the integrated signed area between the expected coverage curve and the diagonal of a particular simulation. From this quantity it is evident there is a clear distinction between NRE and BNRE with respect to the available simulation budget. Both methods have the tendency to converge towards 0, indicating both methods are moving closer to the Bayes optimal classifier. However, the difference between these methods lies with how this solution is approached. While NRE can approach this limit from both sides, BNRE consistently produces coverage AUC's above 0, corresponding to conservative posterior approximations, and therefore exhibits the desired behaviour (in expectation).

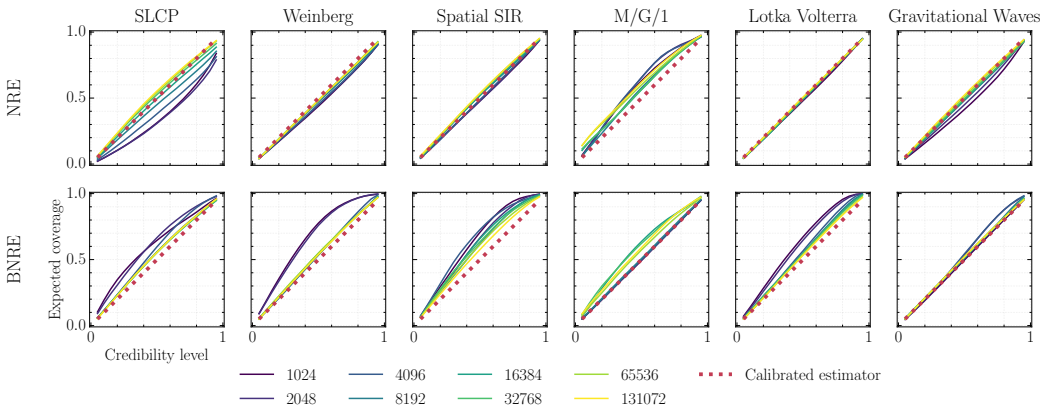

Figure 1: Expected coverage for increasing simulation budgets. A perfectly calibrated posterior has an expected coverage probability equal to the nominal coverage probability and hence produces a diagonal line. A conservative estimator has an expected coverage curve at or above the diagonal line, while an overconfident estimator produces curves below the diagonal line. The diagnostic therefore provides an immediate visual interpretation. We observe that NRE can produce overconfident estimators, while BNRE always produces coverage curves above the diagonal line and therefore the desired behaviour: conservative posterior approximations. The means over 5 runs are reported.

Statistical performance In addition to the reliability of the posteriors, we evaluate and compare the statistical performance of the posterior approximations produced by NRE and BNRE. We estimate the expected approximate log posterior density $\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$ over a large number of pairs $\boldsymbol{\vartheta}, \boldsymbol{x}$. It captures how well the posterior surrogates $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ approximate the true posteriors $p(\boldsymbol{\vartheta} \mid \boldsymbol{x})$ since $\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] = -\mathbb{E}_{p(\boldsymbol{x})}\mathrm{KL}\left[p(\boldsymbol{\vartheta} \mid \boldsymbol{x}) \,||\, \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right] + \mathbb{E}_{p(x)}\mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\log p(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$ [21].

Figure 3 shows our results. We observe that enforcing the balancing condition for $\lambda = 100$ is associated with a loss in statistical performance. However, the loss in statistical performance is eventually recovered by increasing the simulation budget. In fact, practitioners might be inclined to favor reliability over statistical performance [1], although it is always a trade-off that depends on the use case. Nevertheless, it is possible to improve the statistical performance by tuning the surrogate, or by increasing the available simulation budget as we have demonstrated.

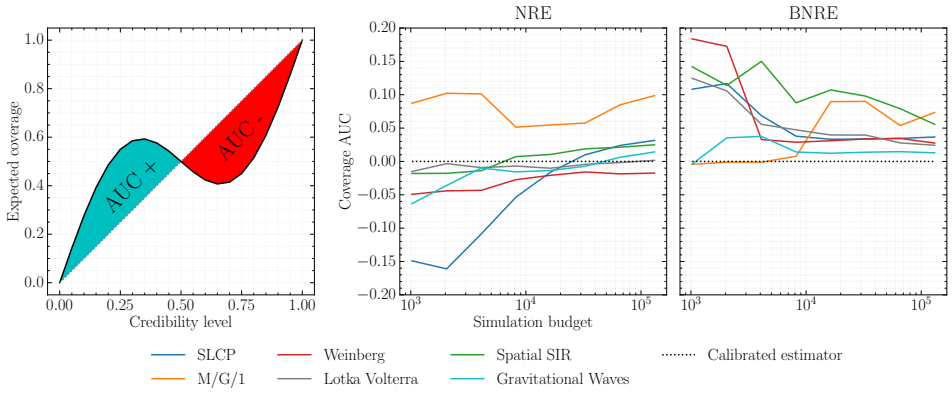

Figure 2: Coverage AUC measures the integrated signed area between the expected coverage curve and the diagonal. A perfectly calibrated posterior has an expected coverage probability equal to the nominal coverage probability, producing a diagonal line and has a coverage AUC of 0, as shown on the left subplot. A conservative estimator on the other hand has a coverage AUC larger than 0 and an overconfident estimator smaller than 0. We observe that while NRE can produce coverage AUC both below or above 0, BNRE always produces a coverage AUC larger than 0, implying that its posterior approximations are conservative on average. The means over 5 runs are reported. A complete overview, including standard deviations, are provided in Appendix F.

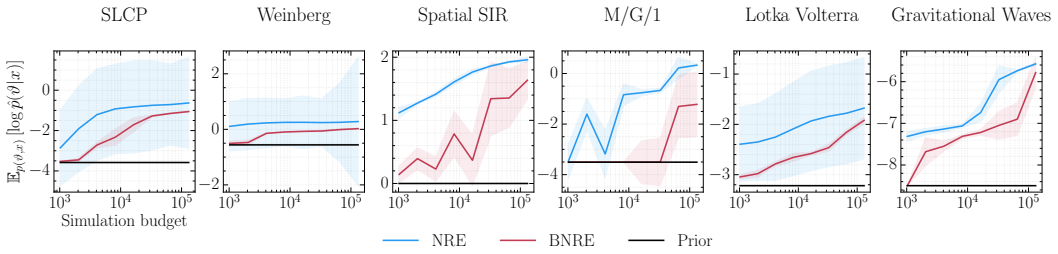

Figure 3: Expected value $\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})\right]$ of the approximate log posterior density of the nominal parameters with respect to the simulation budget. We observe that BNRE produces log posterior densities lower than NRE. This shows that enforcing the balancing condition to have more reliable posterior approximates comes at the price of a small loss in information gain. However, BNRE improves over the prior and eventually converges towards NRE as the simulation budget increases. Solid lines represent the mean over 5 runs and shaded areas represent the standard deviation.

4.2 In-depth analysis

In this section, we consider the Weinberg benchmark as described in Appendix C. The quality of the posterior approximations produced by BNRE is initially discussed with respect to the simulation budget. Afterwards, the effects of the hyper-parameter λ are studied.

Quality assessment Because the expected coverage does not capture the quality of an approximation in terms of information gain, we complement our assessment with a bias and variance analysis of the posterior approximations. Let us consider the expected squared error over the approximate posterior $\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*\right)^2\right]$, where $\boldsymbol{\vartheta}^*$ is the ground truth parameter value. With $\bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) = \mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\boldsymbol{\vartheta}\right]$, we decompose $\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*\right)^2\right]$ as

$$\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}(\boldsymbol{x})\right)^2\right] + 2\left(\bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \boldsymbol{\vartheta}^*\right)\underbrace{\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}(\boldsymbol{x})\right)\right]}_{=0} + \mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \boldsymbol{\vartheta}^*\right)^2\right]$$

$$= \mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})}\left[\left(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}(\boldsymbol{x})\right)^2\right] + \left(\bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \boldsymbol{\vartheta}^*\right)^2.$$

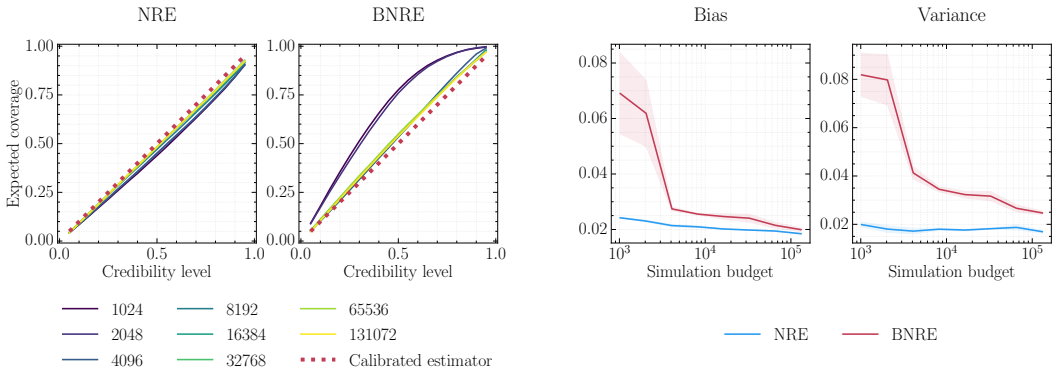

Figure 4: Comparison between NRE and BNRE in terms of expected coverage, bias and variance on the Weinberg benchmark. On the left side, the coverage is shown with respect to the simulation budget represented by the colormap. The bias and variance are represented on the right side of the plot. BNRE is run with $\lambda = 100$. Consistent with our previous observations in Figure 3, we observe that the gap in both bias and variance reduces as the simulation budget increases. Futhermore, in contrast with NRE, the posterior approximations of BNRE are tending towards being increasingly calibrated while at the same time being conservative. Solid lines represent the mean over 5 runs and shaded areas represent the standard deviation.

The expectation over the joint distribution $p(\boldsymbol{\vartheta}^*, \boldsymbol{x})$ of the expected squared error can hence be decomposed in a bias term defined as

$$\text{bias}(\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})) \triangleq \mathbb{E}_{p(\boldsymbol{\vartheta}^*, \boldsymbol{x})} \left[\left(\bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \boldsymbol{\vartheta}^* \right)^2 \right], \tag{8}$$

which can be interpreted as the expected discrepancy between the nominal value $\boldsymbol{\vartheta}^*$ and the expected posterior value $\bar{\boldsymbol{\vartheta}}$. The variance term is

$$\text{variance}(\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})) \triangleq \mathbb{E}_{p(\boldsymbol{\vartheta}^*, \boldsymbol{x})} \left[\mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x})} \left[\left(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}(\boldsymbol{x}) \right)^2 \right] \right] \tag{9}$$

and measures the dispersion of the posterior approximations. Note that these terms differ from the typical statistical bias and variance of point estimators since we are considering full posterior estimators. In particular, the bias of the Bayes optimal model does not necessarily reduce to 0.

Figure 4 shows the evolution of expected coverage, bias and variance with respect to the available simulation budget. By taking all plots into consideration with respect to the simulation budget, we can validate that – as suggested by theorems 1 and 2 – the increase in expected coverage is tied to an increase in variance. However, this increase comes at the price of a slight increase in bias. Consistent with our previous observations in Figure 3, we observe that the gap in both bias and variance reduces as the simulation budget increases. The bias gets close to 0 for high simulation budgets, showing that the bias induced by BNRE vanishes as the simulation-budget increases. A bias and variance analysis for all remaining benchmarks is discussed in Appendix G.

Effects of λ Finally, Figure 5 shows the effect the hyper-parameter λ on the posterior approximations, their expected coverage and the balancing condition. BNRE is run 5 times for λ ranging from 1 to 2^{15} and for a fixed simulation budget of 1024. Initially, the effect on the posterior approximations is limited for small values of λ. However, once λ increases, the balancing condition forces the posterior approximations to become increasingly dispersed and conservative. Eventually, at least for this specific simulation budget, the posterior approximation reduces to the prior as the balancing condition becomes dominant over the cross-entropy term. Although the global optimum remains unchanged as stated by Theorem 3, large λ values are likely to impair the training procedure. In particular, a large λ can inflate the statistical noise of the Monte Carlo estimation of the balancing condition and make the classifier $\hat{d}$ degenerate to a classifier that is trivially balanced such as the random classifier $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) = 0.5$ for all $\boldsymbol{\vartheta}, \boldsymbol{x}$. In this case, $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = 1$ for all $\boldsymbol{\vartheta}, \boldsymbol{x}$ and the approximate posterior degenerates to the prior. This effect is directly evident from Figure 5, starting from $\lambda \simeq 1000$. In practice, λ should be sufficiently large such that the approximate classifier is balanced,

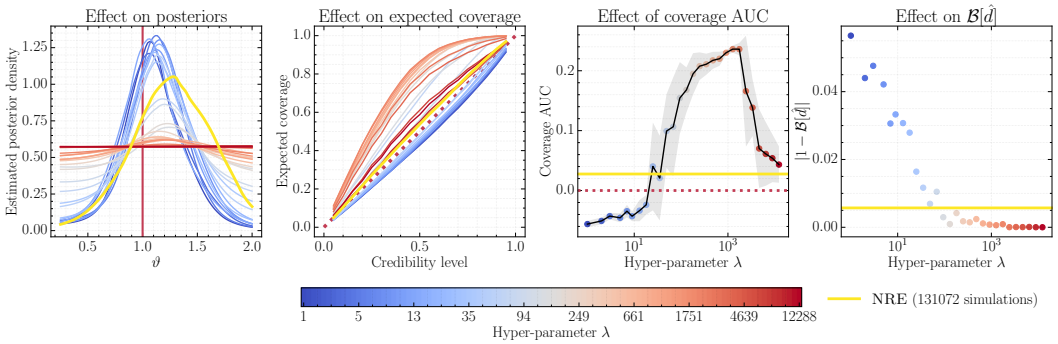

Figure 5: Effect of the hyper-parameter λ for a fixed simulation budget of 1024. The first plot from left to right shows the evolution of the approximate posterior for a given observation at a fixed $\boldsymbol{\vartheta}^*$, indicated by the red vertical line. This approximate posterior is compared to NRE trained on a large simulation budget, shown in yellow and serving as a proxy for the true posterior. The second plot illustrates the empirical expected coverage. The third plot provides a summarized view of the second plot using the coverage AUC as summary statistic. The fourth plot shows that classifiers are becoming increasingly more balanced as λ increases. In addition, the plots show that λ is directly tied to the statistical performance and reliability of the posterior approximations. Classifiers trained with small λ's are associated with (relatively) tight posteriors and overconfident approximations, while classifiers trained with larger values of λ are increasingly more dispersed and conservative until the posterior approximations reduce to the prior due to inflated statistical noise of the Monte Carlo estimation of the balancing condition. Furthermore, the expected coverage plot shows the estimator is almost perfectly calibrated and implicitly balanced. Immediately visible from the various posterior approximations in the leftmost subplot, is the fact that BNRE produces overconfident and biased approximations in the presence of a small simulation budget and a small λ, indicated by their dark blue color. However, the balancing condition can be applied to the underlying estimator to improve its reliability by increasing λ. Ideally, λ should be as small as possible to maximize predictive performance, while at the same time remain sufficiently large to guarantee coverage. From the 3$^{\text{th}}$ subplot from the left, in this particular problem setting, that happens at the point where the coverage AUC transitions from being negative to positive ($\lambda \approx 25.0$).

while maximizing the statistical performance of the posterior estimator. Therefore, we recommend to start with a small value for λ and to gradually increase λ until the posterior estimator becomes conservative. We empirically found $\lambda = 100$ to be a reasonably good default value leading to good performance across all considered benchmarks with various model architectures.

5 Related work

In the Bayesian setting, BNRE improves the reliability of NRE by constraining the classifier hypothesis space to balanced classifiers, which results in more conservative posteriors. Towards the same objective of conservative and reliable approximate posteriors, Hermans et al. [1] have shown empirically that ensembling posterior estimators increases their expected coverage. Since the two solutions are complementary, we suggest that ensembling BNRE is a safe practice to follow. To the best of our knowledge, no other related work exists to make Bayesian simulation-based inference algorithms more conservative and reliable.

In the frequentist setting, Cranmer et al. [6] make use of neural ratio estimation to learn likelihood ratio test statistics. They show that the classifier $\hat{d}$ does not need to be exact for the statistic to remain the most powerful, provided that the approximate likelihood ratio is monotonic with exact likelihood ratio. When this is not the case, robust inference remains possible by calibrating the classifier, at the price of a loss in statistical power. Similarly, for frequentist likelihood-free inference, Dalmasso et al. [22] use classifiers to estimate likelihood ratio statistics and propose a procedure for guaranteeing valid hypothesis tests and confidence sets. Finally, Dalmasso et al. [23] propose a practical procedure for the Neyman construction of confidence sets with finite-sample guarantees of nominal coverage as well as diagnostics that estimate conditional coverage over the entire parameter space.

In this work, we make the assumption that the simulator is well-specified, in the sense that it accurately models the real data generation process. However, this assumption is often violated. To overcome this issue, Generalized Bayesian inference (GBI) extends Bayesian inference by replacing the likelihood term by with arbitrary loss function [24]. Those loss functions can be designed to mitigate specific types of misspecifications and enable robust inference, even with intractable likelihoods [25–27]. Power likelihood losses have also been shown to increase robustness to model misspecification [28]. It consists in raising the likelihood to a power to control the impact it has over the prior. The lower the power of likelihood, the lower the importance given to the data and the higher the uncertainty of the posterior. It can either be set based on practitioner knowledge or derived from observed data [29]. Following the same objective, Miller and Dunson [30] introduce coarsened posteriors that condition on a neighborhood of the empirical data distribution rather than on the data itself. This neighborhood is derived from a distance function that, when set to the relative entropy, allows the approximation of coarsened posteriors by a power posterior. Recently, Dellaporta et al. [31] applied Bayesian non-parametric learning to SBI, making inference with misspecified simulator models both robust and computationally efficient.

6 Conclusions and future work

In this work, we introduced Balanced Neural Ratio Estimation (BNRE), a variation of neural ratio estimation designed to produce more conservative posterior estimators, even when the likelihood-to-evidence ratio estimator is not computationally faithful. We provide theoretical arguments suggesting that enforcing the balancing condition should lead to more conservative posteriors without sacrificing exactness in the large simulation budget regime. Our theoretical results are experimentally validated on benchmarks of varying complexity.

Nevertheless, our inference algorithm comes with limitations that practitioners should keep in mind. First, we emphasize that theorems 1 and 2 hold only in expectation, which means that we cannot provide any guarantee at the level of single inferences. Second, the balancing condition is enforced through a regularization penalty that is not estimated exactly. This implies that the classifier $\hat{d}$ is rarely strictly balanced, although close to be, in which case theorems 1 and 2 do not hold. Third, the benefits of BNRE remain to be assessed in high-dimensional parameter spaces. In particular, the posterior density must be evaluated on a discretized grid over the parameter space to compute credibility regions, which currently prohibits the accurate computation of expected coverage in the high-dimensional setting. In conclusion, **BNRE should not be viewed as a way to obtain conservative posterior estimators with 100% reliability, but rather as a way to increase the reliability of the posterior estimators with minimal effort and no computational overhead**.

Looking forward, the balancing condition could potentially be applied to other simulation-based inference algorithms. Future works could include a generalization to neural posterior estimation (NPE). In fact, the likelihood-to-evidence ratio can be extracted from an approximate posterior by removing its dependence on the prior, $\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \log \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) - \log p(\boldsymbol{\vartheta})$, which in turn can be expressed as a classifier $\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) = \sigma(\log \hat{r}(\boldsymbol{x} \mid \boldsymbol{\vartheta}))$ on which the balancing condition can be evaluated and enforced. Although our work focuses on amortized approximate inference, the balancing condition could also be applied to sequential inference algorithms to increase their reliability.

Finally, although our initial motivation is framed within the field of simulation-based inference, our theoretical results are directly applicable to **any binary classification task** by replacing the joint and marginal distributions in the balancing condition with the distributions of the two considered classes. Therefore, it provides an easy-to-implement modification for high-risk classification problems.

Acknowledgments and Disclosure of Funding

Arnaud Delaunoy, Joeri Hermans and Antoine Wehenkel would like to thank the National Fund for Scientific Research (F.R.S.-FNRS) for their scholarships. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the National Fund for Scientific Research (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

References

[1] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting A Crisis In Simulation-Based Inference. *arXiv e-prints*, art. arXiv:2110.06581, October 2021.

[2] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4239–4248. PMLR, 13–18 Jul 2020.

[3] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

[4] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[5] Manuel Glöckler, Michael Deistler, and Jakob H Macke. Variational methods for simulation-based inference. In *International Conference on Learning Representations*, 2021.

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

[7] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, Michael U Gutmann, et al. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 2016.

[8] Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pages 2771–2781. PMLR, 2020.

[9] Benjamin K Miller, Alex Cole, Patrick Forré, Gilles Louppe, and Christoph Weniger. Truncated marginal neural ratio estimation. *Advances in Neural Information Processing Systems*, 34: 129–143, 2021.

[10] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in neural information processing systems*, pages 1028–1036, 2016.

[11] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[12] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems*, 30, 2017.

[13] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[14] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

[15] David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pages 1830–1840. PMLR, 2021.

[16] Joeri Hermans, Nilanjan Banik, Christoph Weniger, Gianfranco Bertone, and Gilles Louppe. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Monthly Notices of the Royal Astronomical Society*, 507(2):1999–2011, 2021.

[17] Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

[18] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 1973.

[19] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[21] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 343–351. PMLR, 13–15 Apr 2021.

[22] Niccolò Dalmasso, Rafael Izbicki, and Ann Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In *International Conference on Machine Learning*, pages 2323–2334. PMLR, 2020.

[23] Niccolo Dalmasso, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: Bridging classical statistics and machine learning in simulation and uncertainty quantification. *arXiv preprint arXiv:2107.03920*, 2021.

[24] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[25] Sebastian M Schmon, Patrick W Cannon, and Jeremias Knoblauch. Generalized posteriors in approximate bayesian computation. *arXiv preprint arXiv:2011.08644*, 2020.

[26] Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, Chris Oates, et al. Robust generalised bayesian inference for intractable likelihoods. *arXiv preprint arXiv:2104.07359*, 2021.

[27] Lorenzo Pacchiardi and Ritabrata Dutta. Generalized bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*, 2021.

[28] Peter Grünwald and Thijs Van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

[29] Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

[30] Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

[31] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust bayesian inference for simulator-based models via the mmd posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR, 2022.

[32] Kyle Cranmer, Lukas Heinrich, Tim Head, and Gilles Louppe. "Active Sciencing" with Reusable Workflows. `https://github.com/cranmer/active_sciencing`, 2017.

[33] Alexander Y Shestopaloff and Radford M Neal. On bayesian inference for the m/g/1 queue with efficient mcmc sampling. *arXiv preprint arXiv:1401.5548*, 2014.

[34] Alfred J Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415, 1920.

[35] Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118(2972):558–560, 1926.

[36] LIGO Scientific Collaboration. LIGO Algorithm Library - LALSuite. free software (GPL), 2018.

[37] C. M. Biwer, Collin D. Capano, Soumi De, Miriam Cabero, Duncan A. Brown, Alexander H. Nitz, and V. Raymond. PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals. *Publ. Astron. Soc. Pac.*, 131(996):024503, 2019. doi: 10.1088/1538-3873/aaef0b.

A Expected coverage as a special case of simulation-based calibration

Simulation-based calibration (SBC) [14] provides a way to diagnose the faithfulness of an approximate posterior distribution $\hat{p}(\theta|x)$. Given an observation $\boldsymbol{x}^* \sim p(\boldsymbol{x})$, Talts et al. [14] prove that, for any one-dimensional statistic $f : \Theta \mapsto \mathbb{R}$, the rank statistic

$$r(\boldsymbol{\vartheta}^*) = \mathbb{E}_{p(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)}\big[\mathbb{1}[f(\boldsymbol{\vartheta}) \leq f(\boldsymbol{\vartheta}^*)]\big] \tag{10}$$

of posterior samples $\boldsymbol{\vartheta}^* \sim p(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$ is uniformly distributed over the interval $[0, 1]$. Consequently, any deviation from the uniform distribution for the approximate rank statistic

$$\hat{r}(\boldsymbol{\vartheta}^*) = \mathbb{E}_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)}\big[\mathbb{1}[f(\boldsymbol{\vartheta}) \leq f(\boldsymbol{\vartheta}^*)]\big] \tag{11}$$

indicates some error in the approximate posterior $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$. As this holds for any statistic f, it also holds for $f(\boldsymbol{\vartheta}) = \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$. In this special case, if $\hat{r}(\boldsymbol{\vartheta}^*) = \alpha$, a proportion $1 - \alpha$ of samples $\boldsymbol{\vartheta} \sim \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$ have an approximate posterior density larger than $\boldsymbol{\vartheta}^*$. In other words, it means that $\boldsymbol{\vartheta}^*$ resides within the $1 - \alpha$ highest posterior density region $\Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)}(1 - \alpha)$ of $\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$. Therefore, we have

$$P(\hat{r}(\boldsymbol{\vartheta}^*) \geq \alpha) = \mathbb{E}_{p(\boldsymbol{\vartheta}^* \mid \boldsymbol{x}^*)}\left[\mathbb{1}[\boldsymbol{\vartheta}^* \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)}(1 - \alpha)]\right] \tag{12}$$

and since $\hat{r}(\boldsymbol{\vartheta}^*)$ should be uniformly distributed, $P(\hat{r}(\boldsymbol{\vartheta}^*) \geq \alpha)$ should be equal to $1 - \alpha$. In practice, this test cannot be performed locally for a given $\boldsymbol{x}^*$ as we cannot sample from the unknown posterior distribution $p(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$. Instead, SBC checks globally that $\hat{r}(\boldsymbol{\vartheta}^*)$ is uniformly distributed over pairs $(\boldsymbol{\vartheta}^*, \boldsymbol{x}^*) \sim p(\boldsymbol{\vartheta}, \boldsymbol{x})$ sampled from the joint distribution, which, in the special case $f(\boldsymbol{\vartheta}) = \hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)$, comes down to check that

$$\mathbb{E}_{p(\boldsymbol{\vartheta}^*, \boldsymbol{x}^*)}\left[\mathbb{1}[\boldsymbol{\vartheta}^* \in \Theta_{\hat{p}(\boldsymbol{\vartheta} \mid \boldsymbol{x}^*)}(1 - \alpha)]\right] = 1 - \alpha \tag{13}$$

is satisfied for all $\alpha \in [0, 1]$. We recognize here the expected coverage diagnostic used in Hermans et al. [1] and this work.

B Proof of Theorem 2

Theorem 2. *Any balanced classifier $\hat{d}$ satisfies* $\mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\dfrac{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}\right] \geq 1.$

Proof. From the integral form of the balancing condition, we have

$$\begin{aligned}
1 &= \iint \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x} \\
&= 2 - \iint \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x} \\
&= \iint p(\boldsymbol{\vartheta}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x} + \iint p(\boldsymbol{\vartheta})p(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x} - \iint \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x} \\
&= \iint \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\big(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\big) \,\mathrm{d}\boldsymbol{\vartheta} \,\mathrm{d}\boldsymbol{x},
\end{aligned}$$

which implies that $\big(p(\boldsymbol{x}, \boldsymbol{\vartheta}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\big(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\big)$ is a valid density, integrating to 1 and positive everywhere. Therefore, its Kullback-Leibler divergence with $p(\boldsymbol{\vartheta})p(\boldsymbol{x})$ is positive and, using Jensen's inequality, we have

$$\begin{aligned}
0 &\leq \mathrm{KL}\left(p(\boldsymbol{\vartheta})p(\boldsymbol{x}) \,\middle|\middle|\, \big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\big(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\big)\right) \\
&\leq \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}{\big(p(\boldsymbol{\vartheta}, \boldsymbol{x}) + p(\boldsymbol{\vartheta})p(\boldsymbol{x})\big)\big(1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})\big)}\right] \\
&\leq \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\log \frac{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}\right] \\
\Rightarrow \quad 1 &\leq \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\exp\left(\log \frac{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}\right)\right] = \mathbb{E}_{p(\boldsymbol{\vartheta})p(\boldsymbol{x})}\left[\frac{1 - d(\boldsymbol{\vartheta}, \boldsymbol{x})}{1 - \hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x})}\right].
\end{aligned}$$

□

C Benchmarks

The *SLCP* simulator models a fictive problem with 5 parameters. The observable $\boldsymbol{x}$ is composed of 8 scalars which represent the 2D-coordinates of 4 points. The coordinate of each point is sampled from the same multivariate Gaussian whose mean and covariance matrix are parametrized by $\boldsymbol{\vartheta}$. We consider an alternative version of the original task [4] by inferring the marginal posterior density of 2 of those parameters. In contrast to its original formulation, the likelihood is not tractable due to the marginalization.

The *Weinberg* problem [32] concerns a simulation of high energy particle collisions $e^+e^- \to \mu^+\mu^-$. The angular distributions of the particles can be used to measure the Weinberg angle $\boldsymbol{x}$ in the standard model of particle physics. From the scattering angle, we are interested in inferring Fermi's constant $\boldsymbol{\vartheta}$.

The *Spatial SIR* model [1] involves a grid-world of susceptible, infected, and recovered individuals. Based on initial conditions and the infection and recovery rate $\boldsymbol{\vartheta}$, the model describes the spatial evolution of an infection. The observable $\boldsymbol{x}$ is a snapshot of the grid-world after some fixed amount of time.

M/G/1 [33] models a processing and arrival queue. The problem is described by 3 parameters $\boldsymbol{\vartheta}$ that influence the time it takes to serve a customer, and the time between their arrivals. The observable $\boldsymbol{x}$ is composed of 5 equally spaced quantiles of inter-departure times.

The *Lotka-Volterra* population model [34, 35] describes a process of interactions between a predator and a prey species. The model is conditioned on 4 parameters $\boldsymbol{\vartheta}$ which influence the reproduction and mortality rate of the predator and prey species. We infer the marginal posterior of the predator parameters from time series representing the evolution of both populations over time. The specific implementation is based on a Markov Jump Process as in Papamakarios et al. [4].

Gravitational Waves (GW) are ripples in space-time emitted during events such as the collision of two black-holes. They can be detected through interferometry measurements $\boldsymbol{x}$ and convey information about celestial bodies, unlocking new ways to study the universe. We consider inferring the masses $\boldsymbol{\vartheta}$ of two black-holes colliding through the observation of the gravitational wave as measured by LIGO's dual detectors [36, 37].

D Architectures and hyper-parameters

Table 1 summarizes the architectures and hyper-parameters used for each benchmark. The classifier architectures are separated into two parts: the embedding and the head networks. The embedding network ϕ compresses the observable into a set of features. The head network f then uses those features $\phi(\boldsymbol{x})$ concatenated with the parameters $\boldsymbol{\vartheta}$ to predict the class,

$$\hat{d}(\boldsymbol{\vartheta}, \boldsymbol{x}) = f(\boldsymbol{\vartheta}, \phi(x)).$$

The learning rate is scheduled during training. Table 1 provides the initial learning rates. Those are then divided by 10 each time no improvement was observed on the validation loss for 10 epochs. Further details can be found in the code repository attached to this manuscript.

Table 1: Architectures and training hyper-parameters

	SLCP	M/G/1	Weinberg	Lotka-V.	Spatial SIR	GW
Embedding network	None	None	None	CNN	Resnet-18	CNN
Embedding layers	/	/	/	8	/	13
Embedding channels	/	/	/	8	/	16
Convolution type	/	/	/	Conv1D	Conv2D	Dilated Conv1D
Head network	MLP	MLP	MLP	MLP	MLP	MLP
Head layers	6	6	6	3	3	3
Head hidden neurons	256	256	256	128	256	128
Learning rate	0.001	0.001	0.001	0.001	0.001	0.001
Epochs	500	500	500	500	500	500
Batch size	256	256	256	256	256	256

E Estimation of the expected coverage probability

We describe in this section the methodology used to estimate the expected coverage probability

$$\mathbb{E}_{p(\boldsymbol{\vartheta},\boldsymbol{x})}\left[\mathbb{1}\left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x})}(1-\alpha)\right]\right].$$

We consider n test simulations $(\boldsymbol{\vartheta}_i^*, \boldsymbol{x}_i) \sim p(\boldsymbol{\vartheta})p(\boldsymbol{x}\,|\,\boldsymbol{\vartheta})$ and compute their associated approximate posteriors $\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)$ in a discretized and empirically normalized grid of the parameter space. The associated credible region is the highest density credible region, i.e. a credible region of the form

$$\Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)}(1-\alpha) = \{\boldsymbol{\vartheta} : \hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i) \geq \gamma\}. \tag{14}$$

The threshold γ is computed using a dichotomic search to produce a credible region of level $1-\alpha$. We then estimate the empirical expected coverage probability by the proportion of nominal parameters $\boldsymbol{\vartheta}_i^*$ that falls in their associated credible region $\Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)}(1-\alpha)$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left[\boldsymbol{\vartheta}_i^* \in \Theta_{\hat{p}(\boldsymbol{\vartheta}\,|\,\boldsymbol{x}_i)}(1-\alpha)\right].$$

F Standard deviations of Coverage AUCs

Figure 6 shows the coverage AUC for various simulation budgets. The mean and standard deviation over 5 runs are reported.

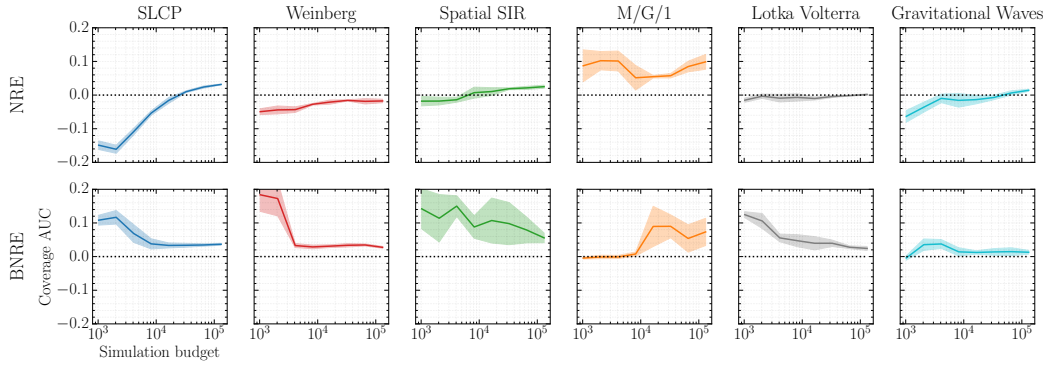

Figure 6: Coverage AUC measures the integrated signed area between the expected coverage curve and the diagonal. A perfectly calibrated posterior has an expected coverage probability equal to the nominal coverage probability, producing a diagonal line and has a coverage AUC of 0, as shown on the left subplot. A conservative estimator on the other hand has a coverage AUC larger than 0 and an overconfident estimator smaller than 0. We observe that while NRE can produce coverage AUC both below or above 0, BNRE always produces a coverage AUC larger than 0, implying that its posterior approximations are conservative. Solid lines represent the mean over 5 runs and shaded areas represent the standard deviation.

G Complete bias and variance analysis

Figure 7 shows the evolution of the bias and variance w.r.t. the simulation budget on a wide variety of benchmarks. We observe that observations made on Weinberg in Section 4 generalize to all benchmarks. The variance obtained with BNRE is always higher or equal than the one obtained with NRE as suggested by Theorems 1 and 2. In addition, as suggested by Theorem 3, the bias and variance obtained with BNRE converges, as NRE, to the Bayes optimal solution.

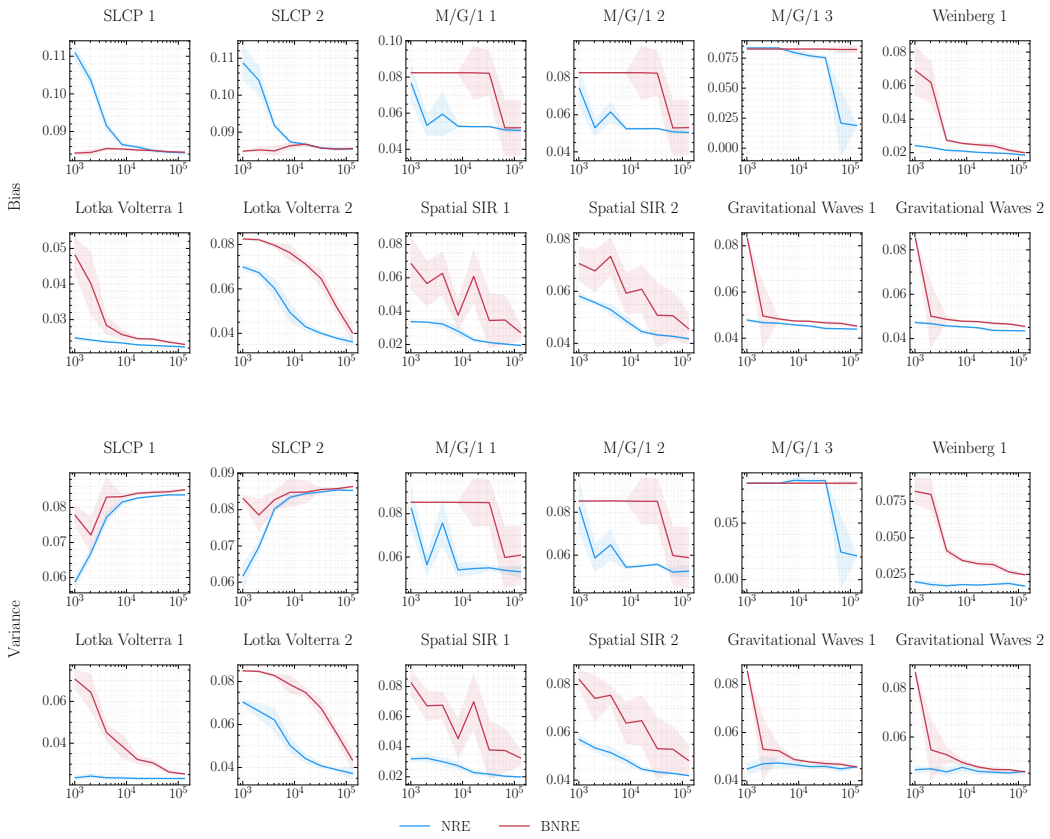

Figure 7: Evolution of the bias and variance w.r.t. the simulation budget. The bias and variance are estimated as described in Section 4 and are scaled to account for the prior's spread, permitting a direct comparison between the benchmarks. Marginals are considered when dealing with multidimensional parameter spaces. Those are denoted by an index following the benchmark name.