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.

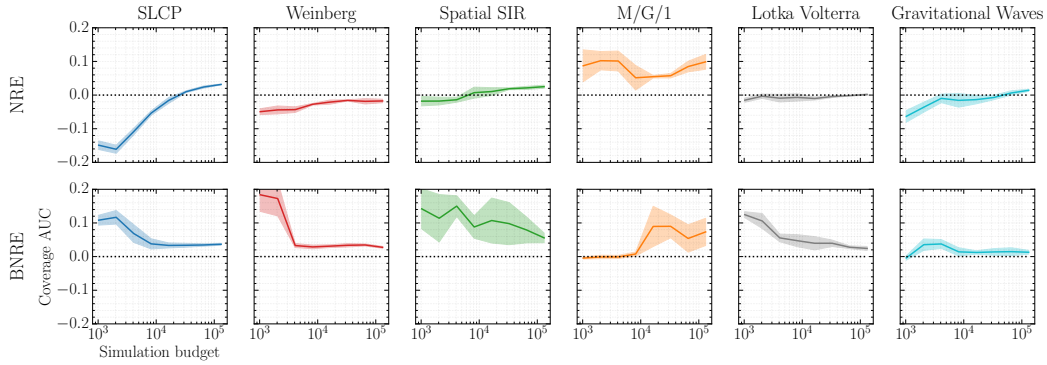

Figure 6: Coverage AUC measures the integrated signed area between the expected coverage curve and the diagonal. A perfectly calibrated posterior has an expected coverage probability equal to the nominal coverage probability, producing a diagonal line and has a coverage AUC of 0, as shown on the left 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.

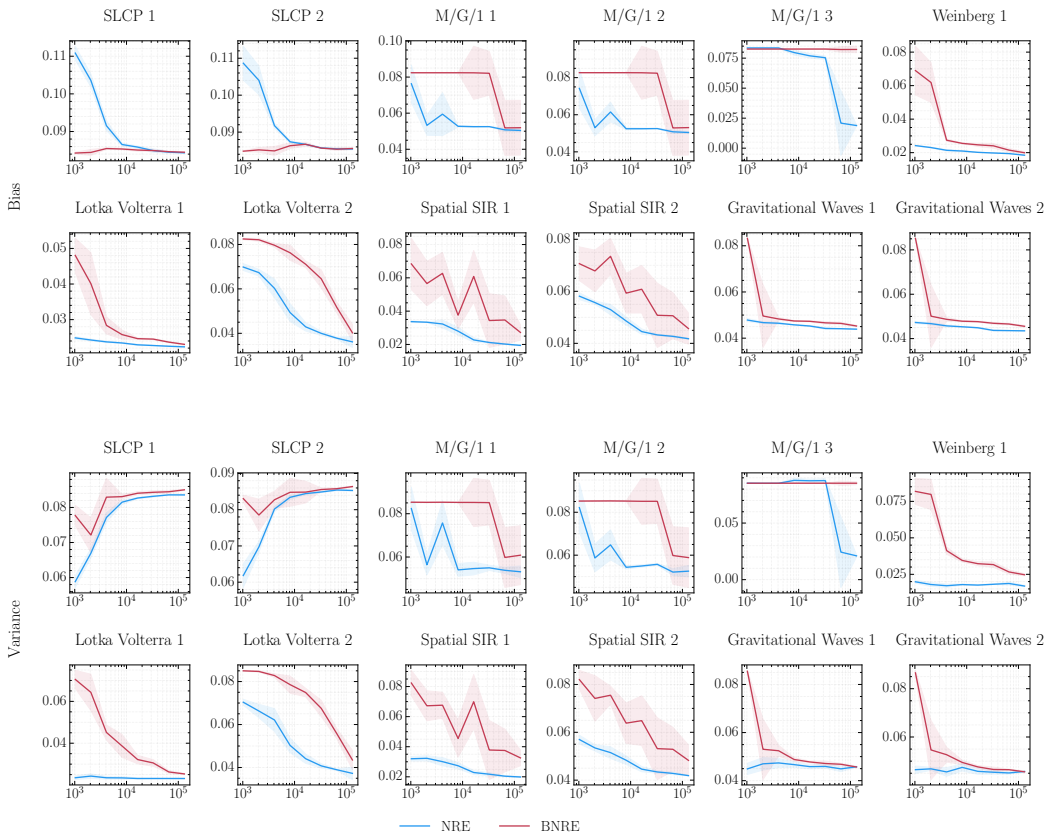

Figure 7: Evolution of the bias and variance w.r.t. the simulation budget. The bias and variance are estimated as described in Section 4 and are scaled to account for the prior's spread, permitting a direct comparison between the benchmarks. Marginals are considered when dealing with multidimensional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