Sampling-Based Accuracy Testing of Posterior Estimators for General Inference

Pablo Lemos [1 2 3 4 *] **Adam Coogan** [1 2 3 *] **Yashar Hezaveh** [1 2 3] **Laurence Perreault-Levasseur** [1 2 3]

Abstract

Parameter inference, i.e. inferring the posterior distribution of the parameters of a statistical model given some data, is a central problem to many scientific disciplines. Generative models can be used as an alternative to Markov Chain Monte Carlo methods for conducting posterior inference, both in likelihood-based and simulation-based problems. However, assessing the accuracy of posteriors encoded in generative models is not straightforward. In this paper, we introduce 'Tests of Accuracy with Random Points' (TARP) coverage testing as a method to estimate coverage probabilities of generative posterior estimators. Our method differs from previously-existing coverage-based methods, which require posterior evaluations. We prove that our approach is necessary and sufficient to show that a posterior estimator is accurate. We demonstrate the method on a variety of synthetic examples, and show that TARP can be used to test the results of posterior inference analyses in high-dimensional spaces. We also show that our method can detect inaccurate inferences in cases where existing methods fail.

1. Introduction

The task of parameter inference, i.e. determining the values of unknown parameters θ in a statistical model consistent with observed data x, is a ubiquitous task in scientific analyses. While multiple well-established approaches such as Markov-chain Monte Carlo (MCMC), variational inference (VI) and nested sampling (Skilling, 2006) already exist, there has been a recent shift towards applying machine learning for posterior inference amortized over different

observations (e.g. Zhu & Zabaras, 2018; Papernot & Mc-Daniel, 2018; Charnock et al., 2020; Wilson & Izmailov, 2020; Zuo et al., 2020). This approach involves training a model (typically a neural network) to approximate the true posterior distribution as a function of the observation. The goal is to efficiently infer the posterior for new data, eliminating the need for costly MCMC runs for each new observation.

Simulation-based inference (SBI, e.g. Cranmer et al., 2020), also known as likelihood-free inference (LFI) or implicit likelihood inference (ILI), has gained significant popularity in recent years (e.g. Ong et al., 2018; Perreault Levasseur et al., 2017; Gonçalves et al., 2020; Dax et al., 2021; Alsing et al., 2019; Wagner-Carena et al., 2021; Legin et al., 2021; Brehmer, 2021; Coogan et al., 2020; Montel et al., 2022; Coogan et al., 2022; Brehmer et al., 2019; Chen et al., 2020; Mishra-Sharma & Cranmer, 2022; Karchev et al., 2022b; Hermans et al., 2021a; Anau Montel & Weniger, 2022; de Witt et al., 2020; Marlier et al., 2021; Karchev et al., 2022a; Ramesh et al., 2022). SBI does not require an explicit expression for the likelihood, and instead merely relies on having a simulator to generate training data. The SBI framework allows handling complex, high-dimensional data and models that are difficult or intractable to analyze using traditional likelihood-based methods.

Early developments of SBI include the introduction of Rejection Approximate Bayesian Computation (ABC) (Rubin, 1984; Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003; Fearnhead & Prangle, 2012), but today SBI has evolved to encompass more powerful, neural network-powered, amortized methods, such as Neural Ratio Estimation (NRE) (Cranmer et al., 2015; Thomas et al., 2022; Hermans et al., 2020; Durkan et al., 2020; Miller et al., 2022b); Neural Posterior Estimation (NPE) (Rezende & Mohamed, 2015; Papamakarios & Murray, 2016; Lueckmann et al., 2018; Lueckmann et al., 2017; Greenberg et al., 2019) and Neural Likelihood Estimation (NLE) (Price et al., 2018; Papamakarios et al., 2019; Frazier et al., 2022). Recently there has been substantial interest in applying SBI in high-dimensional parameter spaces. Generative models, such as Generative Adversarial Networks GANs (Goodfellow et al., 2014), Normalizing Flows (Dinh et al., 2014; Rezende & Mohamed, 2015; Papamakarios et al., 2021), Variational Autoencoders (Kingma & Welling, 2013) and

---

[1]Mila – Quebec AI Institute, Montreal, Quebec, Canada [2]Université de Montréal, Montreal, Quebec, Canada [3]CIELA Institute, Montreal, Quebec, Canada [4]Flatiron Institute Center for Computational Mathematics, 162 5th Ave, 3rd floor, New York, NY 10010, USA. Correspondence to: Pablo Lemos <pablo.lemos@umontreal.ca>.

Score-Based/Diffusion Models (Song et al., 2020; Ho et al., 2020; Sohl-Dickstein et al., 2015), are powerful ways to encode approximate posteriors in such settings.

Convergence tests for MCMC methods, such as the Gelman-Rubin statistic (Gelman & Rubin, 1992), the effective sample size and the integrated autocorrelation time, are well-established. However, these assess the diversity of samples rather than directly guaranteeing that the posterior is being sampled correctly. For SBI, testing for accuracy of the estimated posterior is often performed using coverage probabilities (but see also Guo et al. (2017)), relying on the evaluation of the density of the posteriors. (Schall, 2012; Prangle et al., 2013; Cranmer et al., 2020; Hermans et al., 2021b). Coverage probabilities measure the proportion of the time that a certain interval contains the true parameter value. However, coverage probability calculations based on evaluations of the learned posterior distributions are not applicable to samples obtained from those generative models where such evaluations are not available, such as GANs and diffusion models. Furthermore, and more importantly, these coverage probability tests are a necessary but not sufficient diagnostic to assess the accuracy of the estimated posterior.

Other methods have been suggested as alternative validations for SBI (Lueckmann et al., 2021; Dalmasso et al., 2020; Deistler et al., 2022). For example, Simulation-Based Callibration (SBC) (Talts et al., 2018), proposes an interesting technique that uses only samples, but can only be used for one-dimensional posteriors and not the full-dimensional space. Another method, proposed by Linhart et al. (2022), is an efficient way to assess posterior accuracy but is designed specifically for normalizing flows and cannot be applied in other inference settings. None of these methods can be applied to assess the accuracy of inference for high-dimensional variables.

The goal is this paper is to introduce a framework for testing the accuracy of parameter inference using only samples from the true joint distribution of data x and the parameters of interest θ, $p(x, \theta)$, and samples from the estimated posterior distribution $\hat{p}(\theta|x)$.

Our novel contributions are a proof of the necessary and sufficient conditions to verify the accuracy of posterior estimators through coverage checks (Theorem 3), along with a methods that practically implements this theorem (§3.2). We begin by introducing all necessary notation in §2. We then introduce our method in §3, present our experiments in §4, and summarize our findings in §5. Our code is available at https://github.com/Ciela-Institute/tarp.

2. Formalism

In this section, we introduce some basic concepts and build up to our key theoretical result (Theorem 3). The coverage testing procedure introduced in the following section is essentially a practical implementation of this theorem.

2.1. Notation

As stated in the introduction, we are interested continuous-valued parameters $\theta \in U \subset \mathbb{R}^n$ and observations $x \in V \subset \mathbb{R}^m$ taken from (subsets of) Euclidean space, with joint density $p(\theta, x)$. We denote our posterior estimator by $\hat{p}(\theta|x)$ (which could be a neural network or MCMC sampler, for example) and assume we can also use it to generate samples of θ.

With these preliminaries, we make two basic definitions:

Definition 1. *A posterior estimator $\hat{p}(\theta|x)$ is* **accurate** *if*

$$\hat{p}(\theta|x) = p(\theta|x) \quad \forall (x, \theta) \sim p(x, \theta) \,. \quad (1)$$

Definition 2. *A **credible region generator** $\mathcal{G} : \hat{p}, \alpha, x \mapsto W \subset U$ for a given credibility level $1 - \alpha$ and observation x is a function satisfying*

$$\int_{\mathcal{G}(\hat{p}, \alpha, x)} \mathrm{d}\theta \, \hat{p}(\theta|x) = 1 - \alpha \,. \quad (2)$$

Note that there are an infinite number of such generators. A commonly-used one is the highest-posterior density region generator, defined in §3.1.

Next, we introduce two central definitions for this work, adapted from Hermans et al. (2021b) (henceforth H21).

Definition 3. *The **coverage probability** for a generator $\mathcal{G}$, credibility level $1 - \alpha$ and datum x is*

$$\mathrm{CP}(\hat{p}, \alpha, x, \mathcal{G}) = \mathbb{E}_{p(\theta|x)} \left[\mathbb{1} \left(\theta \in \mathcal{G}(\hat{p}, \alpha, x) \right) \right] \,. \quad (3)$$

Definition 4. *The **expected coverage probability** for a generator $\mathcal{G}$ and credibility level α is the coverage probability averaged over the data distribution:*

$$\mathrm{ECP}(\hat{p}, \alpha, \mathcal{G}) = \mathbb{E}_{p(x)} \left[\mathrm{CP}(\hat{p}, \alpha, x, \mathcal{G}) \right] \,. \quad (4)$$

2.2. Coverage probability

We now demonstrate some basic facts about estimators with correct coverage probabilities. We begin with a straightforward result:

Theorem 1. *The posterior has coverage probability $\mathrm{CP}(p, \alpha, x, \mathcal{G}) = 1 - \alpha$ for all values of x and any credible region generator $\mathcal{G}(p, \alpha, x)$.*

Proof Substituting $\hat{p}(\theta|x) = p(\theta|x)$, the definition of coverage probability becomes:

$$\mathrm{CP}(p, \alpha, x, \mathcal{G}) = \mathbb{E}_{p(\theta|x)} \left[\mathbb{1}(\theta \in \mathcal{G}(p, \alpha, x) \right]$$
$$= \int_{\mathcal{G}(p, \alpha, x)} \mathrm{d}\theta \, p(\theta|x) \quad (5)$$
$$= 1 - \alpha \,,$$

where the last line follows from the definition of a credible region. ∎

It follows trivially from this that the posterior has $\text{ECP}(p, \alpha, x, \mathcal{G}) = 1 - \alpha$ as well.

Next, we prove the more interesting reverse direction of this theorem, which requires introducing another type of credible region generator.

Definition 5. *A **positionable credible region generator** $\mathcal{P}_{\theta_r}(\hat{p}, \alpha, x)$ generates credible regions positioned at θ_r, a freely-chosen point in parameter space, in the sense that*

$$\lim_{\alpha \to 1} \mathcal{P}_{\theta_r}(\hat{p}, \alpha, x) = \{\theta_r\} \qquad (6)$$

for all x and θ_r. The regions' shapes are not important: they could be, for example, balls or hypercubes.

Lastly, we denote the average of a function $f(\cdot)$ over a credible region Θ positioned at θ_r as

$$\overline{f(\theta_r)}(\Theta) := \frac{1}{\text{vol}[\Theta]} \int_{\Theta} d\theta \, f(\theta) . \qquad (7)$$

When $f(\cdot)$ is a probability density function, $\overline{f(\cdot)}(\Theta)$ is as well, since it is the convolution of $f(\cdot)$ with the density $\mathbb{1}(\theta \in \Theta)$.

Theorem 2. *Suppose the coverage probability of a posterior estimator is equal to $1 - \alpha$ for a positionable credible region generator $\mathcal{P}_{\theta_r}$ for all θ_r, x and α. Further, suppose that $\hat{p}(\cdot|x)$ and $p(\cdot|x)$ are both continuous on their domains. Then $\hat{p}(\cdot|x) = p(\cdot|x)$.*

Proof Define $\Theta := \mathcal{P}_{\theta_r}(\hat{p}, \alpha, x)$ for clarity.

The integral in the definition of the coverage probability can be written as

$$\begin{aligned} \text{CP}(\hat{p}, \alpha, x, \mathcal{P}_{\theta_r}) &= 1 - \alpha \\ &= \int_{\Theta} d\theta \, p(\theta|x) \\ &= \text{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) , \end{aligned} \qquad (8)$$

where first equality follows by assumption. Since we've assumed $\hat{p}(\cdot|x)$ has support everywhere that $p(\cdot|x)$ has support, the volume of the credible region is positive. By the definition of a credible region, we also have

$$1 - \alpha = \int_{\Theta} d\theta \, \hat{p}(\theta|x) = \text{vol}[\Theta] \, \overline{\hat{p}(\cdot|x)}(\Theta) . \qquad (9)$$

Setting this equal to the previous expression yields $\overline{\hat{p}(\cdot|x)}(\Theta) = \overline{p(\cdot|x)}(\Theta)$, which holds for all θ_r and x by assumption. Taking $\alpha \to 1$ (i.e., making Θ small) gives the desired result. ∎

2.3. Expected coverage probability

The previous result is still not very useful, since it is computationally very expensive to calculate the coverage probability of a posterior estimator. Practically, doing so requires producing histograms of the samples from $p(\theta, x)$ in x, which may be high-dimensional. However, as pointed out in H21, it's much simpler to compute the *expected* coverage probability.

The next theorem is our main theoretical result: correct expected coverage is enough to verify the posterior estimator is accurate, as long as it is correct for any function $\theta_r(x)$ defining the positions of the credible regions.

Theorem 3. *Suppose the expected coverage probability of $\hat{p}$ is equal to $1 - \alpha$ for a positionable credible region generator $\mathcal{P}_{\theta_r}$ for all α, x, and $\theta_r(\cdot)$ assigning a position to the credible regions as a function of x. Further suppose that $\hat{p}(\cdot|x)$ has support everywhere that $p(\cdot|x)$ has support, and that both functions are continuous on their domains. Then $\hat{p}(\cdot|x) = p(\cdot|x)$.*

Proof Again, let $\Theta := \mathcal{P}_{\theta_r}(\hat{p}, \alpha, x)$ for clarity.

First, we leverage the definition of credible regions to find an expression for the volume of Θ:

$$1 - \alpha = \int_{\Theta} d\theta \, \hat{p}(\theta|x) = \text{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) , \qquad (10)$$

which implies

$$\text{vol}[\Theta] = \frac{1 - \alpha}{\overline{p(\cdot|x)}(\Theta)} . \qquad (11)$$

This allows us to expand and simplify the expression for the expected coverage:

$$\begin{aligned} \text{ECP}(\hat{p}, \alpha, \mathcal{P}_{\theta_r}) &= 1 - \alpha \\ &= \int dx \, p(x) \int_{\Theta} d\theta \, p(\theta|x) \\ &= \int dx \, p(x) \, \text{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) \\ &= (1 - \alpha) \int dx \, p(x) \frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} . \end{aligned} \qquad (12)$$

Canceling the factors of $1 - \alpha$ gives that the integral in the last line is equal to 1.

By assumption, this holds for *any* choice of position function $\theta_r(x)$. We can therefore take the functional derivative of the integral with respect to $\theta_r(x)$. Recalling that the averages
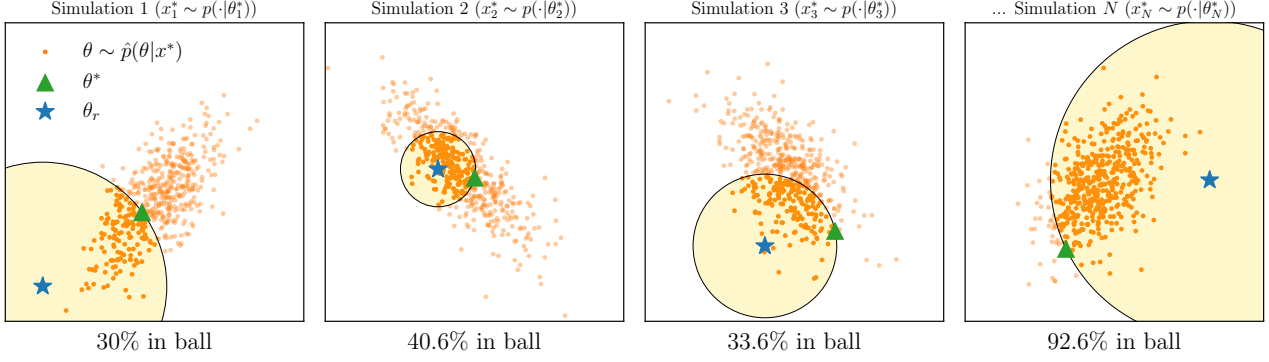
Figure 1. A graphical illustration of the proposed coverage test for assessing the quality of a posterior estimator $\hat{p}$. Given a set of simulations (panels), we draw samples from the posterior estimator (orange points). We sample a reference parameter point θ_r, and determine the fraction of points f falling within a ball centered on θ_r extending to the true parameter point θ^* used to generate the simulation (ball indicated in yellow, f indicated below each panel). Our coverage test aggregates the statistics of f, providing a necessary and sufficient way to guarantee the accuracy of $\hat{p}$.

in the integrand depend on θ_r, we obtain

$$0 = \frac{\delta}{\delta \theta_r(x)} \int dx \, p(x) \frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \qquad (13)$$

$$= \int dx \, \delta \theta_{r,i}(x) \, p(x) \frac{\partial}{\partial \theta_{r,i}} \left(\frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \right) \qquad (14)$$

$$= \int dx \, \delta \theta_{r,i}(x) \frac{\overline{p(\cdot|x)}(\Theta) \, p(x)}{\overline{\hat{p}(\cdot|x)}(\Theta)}$$
$$\times \left[\frac{\partial \log \overline{p(\cdot|x)}(\Theta)}{\partial \theta_{r,i}} - \frac{\partial \log \overline{\hat{p}(\cdot|x)}(\Theta)}{\partial \theta_{r,i}} \right], \qquad (15)$$

where the i subscript indexes the components of θ_r. Since this expression must hold for all variations $\delta \theta_{r,i}$, the integrand must evaluate to zero (i.e., the Euler-Lagrange equation must be satisfied). By assumption, the factor outside the braces in the integrand is nonzero, implying

$$\frac{\partial \log \overline{p(\cdot|x)}(\Theta)}{\partial \theta_{r,i}} = \frac{\partial \log \overline{\hat{p}(\cdot|x)}(\Theta)}{\partial \theta_{r,i}} . \qquad (16)$$

This implies $\log \overline{p(\cdot|x)}(\Theta) = \log \overline{\hat{p}(\cdot|x)}(\Theta) + c(x)$, for some x-dependent integration constant c. But since the functions inside the logarithms themselves densities, we have $c(\cdot) = 0$. Taking the limit $\alpha \to 1$ gives $\hat{p}(\theta|x) = p(\theta|x)$. ∎

The coverage testing method we will introduce in the next section is effectively a practical implementation of this theorem.

3. Our method

With our main theoretical result proven (c.f. Theorem 3), in this section we use it to first explain the blind spots of

typical coverage probability calculations and then introduce our new coverage checking procedure.

3.1. High posterior density coverage testing

Before introducing the proposed method, we first discuss HPD coverage.

Definition 6. *We define the HPD credible region generator $\mathcal{H}(\hat{p}, \alpha, x)$ as the generator that produces the region with mass $1 - \alpha$ occupying the smallest possible volume in U[1].*

Note that this is not a positionable credible region generator. This can be used combined with Def. 4 to calculate High-Posterior Density Expected Coverage Probabilities (HPD ECPs). HPD ECPs are often used to assess coverage (Hermans et al., 2021b; Rozet et al., 2021; Miller et al., 2022a; Deistler et al., 2022; Tejero-Cantero et al., 2020).

Perhaps the most intuitive way of calculating expected coverage probability using HPD regions is to compute such a region for all possible values of α,[2] then calculate the expected coverage using (3). In practice, however, there is a more efficient calculation of expected coverage probabilities, which is derived from the following result:

Remark 1. *A pair (θ^*, x^*), and a posterior estimator $\hat{p}(\theta|x)$ uniquely define a HPD confidence region as:*

$$\{\theta \in U \mid \hat{p}(\theta|x^*) \geq \hat{p}(\theta^*|x^*)\} . \qquad (17)$$

This, in turn, defines a corresponding **HPD confidence level** *$1 - \tilde{\alpha}_{\text{HPD}}(\hat{p}, \theta^*, x^*)$, as the integral of $\hat{p}(\theta|x)$ over that region.*

---

[1] Note this is ill-defined for the uniform density function.

[2] Note that previous works such as Perreault Levasseur et al. (2017) have attempted to perform accuracy testing from a handful of values of α. This test is not nearly as restrictive as scanning over all possible values of α as is typically used for coverage testing.
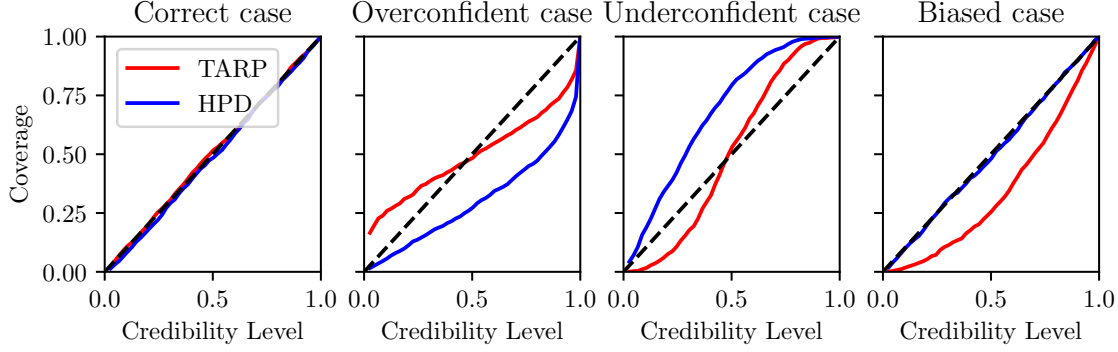
Figure 2. Results on the Gaussian toy model for all four cases described in §4.1. The red line shows the method presented in this paper, while the blue shows the HPD region.

Algorithm 1 Calculation of $\mathrm{ECP}(\hat{p}, \alpha, \mathcal{H})$ using highest posterior density regions, from a set of simulations $\{\theta_i, x_i\}$, $i \in [1, N_{\mathrm{sims}}]$

Generate n samples $\{\theta_{ij}\} \sim \hat{p}(\theta|x_i)$ for each simulation x_i.
for $i \leftarrow 1$ to N_{sims} **do**
  $f_i = (1/\mathrm{n}) \cdot \sum_{j=1}^n \mathbb{1}\left[\hat{p}(\theta_{ij}|x_i) < \hat{p}(\theta_i^*|x_i)\right]$
end for
$\mathrm{ECP}(\hat{p}, \alpha, \mathcal{H}) = (1/N_{\mathrm{sims}}) \sum_{i=1}^{N_{\mathrm{sims}}} \mathbb{1}\left(f_i < 1 - \alpha\right)$

We can then rederive an important result for this HPD confidence level:

Lemma 1. *We can calculate the ECP of the $1 - \alpha$ highest posterior density regions as:*

$$\mathrm{ECP}(\hat{p}, \alpha, \mathcal{H}) = \mathbb{E}_{p(\theta,x)}\left[\mathbb{1}\left(\tilde{\alpha}_{\mathrm{HPD}}(\hat{p}, \theta, x) \geq \alpha\right)\right]. \quad (18)$$

Proof Firstly, we notice that:

$$\theta^* \in \mathcal{H}\left(\hat{p}, \alpha, x^*\right) \Leftrightarrow \tilde{\alpha}_{\mathrm{HPD}}(\hat{p}, \theta^*, x^*) \geq \alpha. \quad (19)$$

This follows from the fact that, if $\theta^* \in \mathcal{H}\left(\hat{p}, \alpha, x^*\right)$, then the HPD confidence region defined by (θ^*, x^*) is contained in $\mathcal{H}\left(\hat{p}, \alpha, x^*\right)$.

Then, from (4), it follows that (18) is true. ∎

This result can be used in practice to calculate the HPD ECP from samples of the true joint distribution $p(\theta, x)$, as shown in Algorithm 1. As previously discussed, this algorithm requires explicit evaluations of the posterior estimator. We try to provide more intuitive connections between both definitions in §A.

As is well-known in the literature, estimating the ECP with HPD regions is not enough to demonstrate a posterior estimator is accurate. Theorem 3 reveals why: by definition, the HPD region generator is not positionable. Positionability is critical to the proof of the theorem, since it requires varying the position function $\theta_r(x)$.

To concretely demonstrate how considering only HPD coverage can fail, we consider the interesting case discussed in H21 of $\hat{p}(\theta|x) = p(\theta)$. From the definition of ECP,

$$\begin{aligned}
\mathrm{ECP}(\hat{p}, \alpha, \mathcal{H}) &= \mathbb{E}_{p(x,\theta)}[\mathbb{1}(\theta \in \mathcal{H}(\hat{p}, \alpha)] \\
&= \mathbb{E}_{p(\theta)}[\mathbb{1}(\theta \in \mathcal{H}(\hat{p}, \alpha)] \\
&= \int_{\mathcal{H}(\hat{p},\alpha)} \mathrm{d}\theta\, p(\theta) \\
&= 1 - \alpha.
\end{aligned} \quad (20)$$

In the second line, we used the fact that HPD generator is independent of x in this case $\mathcal{H}(\hat{p}, \alpha, x) = \mathcal{H}(\hat{p}, \alpha)$. We recognize the third line as the definition of a credible region for the prior, yielding the fourth line. This means that $\hat{p}(\theta|x)$ has perfect HPD ECP in this case.

We now introduce a coverage testing method that remedies such blind spots.

3.2. Distance to random point coverage testing

The method proposed here generates spherical credible regions around position θ_r:

Definition 7. *Given a distance metric $d : U \times U \rightarrow R$, We define the generator of TARP regions $\mathcal{D}_{\theta_r}(\hat{p}, \alpha, x, d)$ as the positionable generator that produces credible regions of credibility level $1 - \alpha$:*

$$\mathcal{D}_{\theta_r}(\hat{p}, \alpha, x, d) := \{\theta \in U \mid d(\theta, \theta_r) \leq R(\hat{p}, \alpha, x)\}, \quad (21)$$

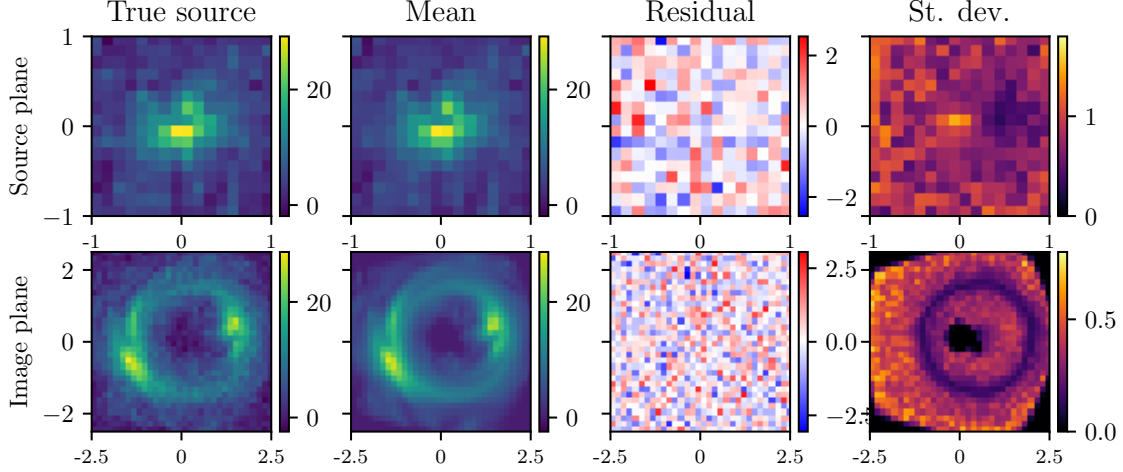*where $R(\hat{p}, \alpha, x)$ is such that (2) is satisfied.*

Figure 3. An example of one of the lensing simulations performed. The top panels show the (latent) source plane that we are trying to infer, while the bottom panels show the distorted images. From left to right, the plot shows the truth, mean, and standard deviation of the samples from the posterior estimator (in the case of this figure, the 'exact' estimator), and the residuals. The noise in the observations is set to 1 on the color scales shown here.

Algorithm 2 Calculation of $\text{ECP}(\hat{p}, \alpha, \mathcal{D}_{\theta_r})$ using the TARP method, using a set of simulations $\{\theta_i, x_i\}, i \in \{1, \ldots, N_{\text{sims}}\}$, parameter distance metric $d : U \times U \to \mathbb{R}^{\geq 0}$ and reference point sampling distribution $\tilde{p}(\cdot|x)$.

Generate n samples $\{\theta_{ij}\} \sim \hat{p}(\theta|x_i)$ for each simulation x_i, where $j = \{1, \ldots, n\}$.
for $i \leftarrow 1$ to N_{sims} **do**
$\quad \theta_r \sim \tilde{p}(\theta_r|x)$ {Generate reference point}
$\quad f_i = (1/\text{n}) \cdot \sum_{j=1}^{n} \mathbb{1}\left[d(\theta_{ij}, \theta_r) < d(\theta_i^*, \theta_r)\right]$
end for
$\text{ECP}(\hat{p}, \alpha, \mathcal{D}_{\theta_r}) = (1/N_{\text{sims}}) \sum_{i=1}^{N_{\text{sims}}} \mathbb{1}(f_i < 1 - \alpha)$

From this result, and similarly to the previous section, a key result follows:

Remark 2. *A pair (θ^*, x^*), and a posterior estimator $\hat{p}(\theta|x)$ uniquely define a **TARP**[3] credible region for a given d and θ_r:*

$$\{\theta \in U \mid d(\theta, \theta_r) \leq d(\theta^*, \theta_r)\} \quad (22)$$

*This, in turn, defines a corresponding **TARP confidence level** $1 - \tilde{\alpha}_{\text{TARP}}(\hat{p}, \theta^*, \theta_r, d)$. as the integral of $\hat{p}(\theta|x)$ over that region.*

We can calculate expected coverage similar to the HPD case:

Lemma 2. *We can calculate the ECP of the $1 - \alpha$ TARP regions as:*

$$\text{ECP}(\hat{p}, \alpha, \mathcal{D}_{\theta_r}) = \mathbb{E}_{p(\theta,x)}\left[\mathbb{1}(\tilde{\alpha}_{\text{TARP}}(\hat{p}, \theta^*, \theta_r, x^*, d) \geq \alpha)\right]. \quad (23)$$

Proof Let $\mathcal{D}_{\theta_r}(x^*, \alpha, \hat{p}, d)$ be a ball centered at θ_r with radius $R(\hat{p}, \alpha, x)$ and credibility $1 - \alpha$. Similarly, the

---

[3]TARP is short for "Test of Accuracy with Random Points. A previous version of this paper used the name DRP ("Distance to Random Point").

TARP region defined by (θ^*, x^*) has the same center, radius $d(\theta^*, \theta_r)$, and credibility $1 - \tilde{\alpha}$ for some $\tilde{\alpha}$. It then follows that:

$$\theta^* \in \mathcal{D}_{\theta_r}(x^*, \alpha, \hat{p}, d) \Leftrightarrow d(\theta^*, \theta_r) \leq R(\hat{p}, \alpha, x). \quad (24)$$

Since R is a monotonic function of α and the regions are centered on the same point, we have

$$d(\theta^*, \theta_r) \leq R(\hat{p}, \alpha, x) \Leftrightarrow \tilde{\alpha} \geq \alpha. \quad (25)$$

Then by (4) we have (23). ∎

With this, we have everything we need to formulate our algorithm, which is presented in Algorithm 2. While similar to Algorithm 1, there are three key differences to this algorithm:

- TARP implements Theorem 3's requirement that coverage holds for all possible ways of choosing the positions of the credible regions by randomly sampling θ_r from some distribution $\tilde{p}(\theta|x)$ that can depend on x.

- TARP probes credible regions of smaller size (i.e., larger α) as the number of posterior samples, simulations, and reference points tested is increased. Following the logic of the proof of Theorem 3, this means it asymptotically tests whether the averages of $\hat{p}(\theta|x)$ and $p(\theta|x)$ match on smaller and small balls.

- TARP does not require explicit evaluations of the posterior estimator $\hat{p}$: it only requires calculating distances between parameters sampled from $\hat{p}$ and θ_r.

In the following section, we test the proposed method in a series of experiments and compare its performance with that of HPD coverage probabilities.

4. Experiments

We apply our algorithm, described in Algorithm 2 to three different experiments. For all experiments, we normalize all parameters θ to the range $[0, 1]$, and unless otherwise specified, we generate reference points uniformly in the D-dimensional hypercube $x \in [0, 1]^D$ where D is the dimensionality of the parameter space. We use the Euclidean or L2 distance as a metric to calculate TARP regions. We explore the dependence on the reference point distribution and the distance metric in §4.2.

4.1. Gaussian Toy Model

As a first example, we can use a simple Gaussian toy model. In this model, we assume that all the posterior distributions are Gaussian. Therefore, we can generate samples from the posterior for a validation simulation from the estimated mean and covariance matrix. We first generate 'simulations', by uniform sampling in our parameter space, $\theta^* \sim \mathcal{U}(-5, 5)$. We also randomly generate the diagonal elements of the covariances matrices Σ of our posterior estimates by sampling from $\log \sigma \sim \mathcal{U}(-5, -1)$, and set the off-diagonal elements to 0. To validate, we also need to know the mean of the posterior distributions. We consider three cases:

- Firstly, we draw these from a normal distribution $\mathcal{N}(\theta^*, \Sigma)$. This means that the coverage probabilities should show a uniform distribution. We call this the *correct case*.

- Secondly, we draw the true values from $\mathcal{N}(\theta^*, 0.5\Sigma)$ and $(\mathcal{N}(\theta^*, 2\Sigma))$. This means that the posterior samples come from a distribution that is too narrow (wide), and are therefore overconfident (underconfident)

- Lastly, we want to build a *biased case*. For this, we pick the means to be equal to:

$$\theta^* - \text{sign}(\theta^*) \cdot Z \left(1 - \frac{|\theta^*|}{5}\right) \cdot \sigma, \quad (26)$$

where Z is the inverse survival function. The idea with this example is to create a position-dependent bias: The furthest the true value is from the origin, the more biased the posterior is. We have specifically designed this bias in a way that HPD coverage probabilities will be blind to it. However, the point of this example is to show that there are biases that HPD can be blind to, but the random nature of TARP should be able to detect. The function (26) is plotted in appendix §C

For each of these cases, we want to compare how this method compares to the HPD coverage probability test. Because in this toy model we know the correct posterior, we can easily compute both HPD and TARP coverage probabilities. To pick the TARP reference points, we use the prior $(\tilde{p}(\theta_r|x) = p(\theta_r))$.

The results for our Gaussian toy model are shown in Fig. 2. In each panel, the x-axis shows the credibility level $1 - \alpha$, while the y-axis shows the expected coverage $\text{ECP}(\hat{p}, \alpha, \mathcal{G})$. For an accurate posterior estimator, $\text{ECP}(\hat{p}, \alpha \mathcal{G}) = 1 - \alpha$, $\forall \alpha \in (0, 1)$ as described in §2, which would then lead to the diagonal black dashed diagonal line. We see in the first panel that that is indeed the case for the 'correct' case, which is accurate by construction. We found consistent results amongst all values of D we tested, going up to $D = 1000$.

The second and third panels show the over and underconfident cases, respectively. We see how these cases lead to different coverage plots than the HPD method. This is not entirely unexpected: For underconfident estimators, the TARP regions from randomly selected points are more likely to cover approximately half of the posterior estimator $\alpha \sim 0.5$, while for overconfident estimators, they are likely to cover either very little $\alpha \sim 1$ or a lot $\alpha \sim 0$. We expand this intuition, including some figures, in §B. Finally, in the fourth panel, we see how the biased case cannot be detected by the HPD region but is detectable by TARP. This shows how, as explained in §2, $\text{ECP}(\hat{p}, \alpha) = 1 - \alpha$ does not mean the posterior is accurate for HPD regions, but it does for TARP regions.

We also repeated this example for the case of Gaussian distributions with nondiagonal elements in the covariance matrix. To do this, we randomly generated arrays of size $D(D-1)/2$, we then converted them into lower triangular matrices, which we used as the Cholesky decomposition of the covariance matrix. We found that adding non-zero elements to the covariance matrix did not change our results.

4.2. Dependence on θ_r distribution and distance metric

All the results of the Gaussian Toy Model, shown in Fig. 2, rely on two choices, specified in §3.2: A distribution to draw reference points θ_r from, and a distance metric $d(\cdot, \cdot)$. Therefore, it is key to study the dependence of our method
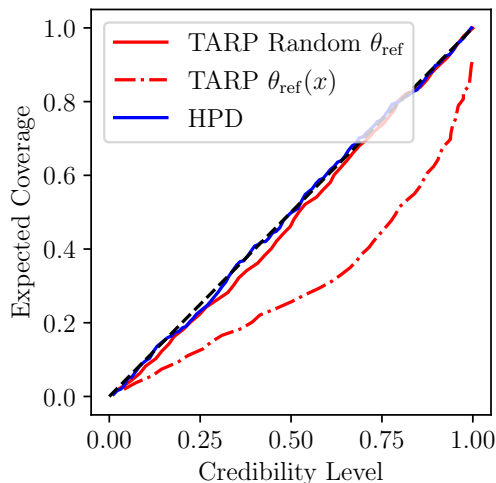
Figure 4. Expected coverage vs credibility level for the uninformative posterior estimator described in §4.3. The blue line shows the coverage calculated using HPD regions, while the red lines use TARP regions. The continuous line uses reference points that are independent of x, while the dot-dashed line uses reference points that depend on x.

on different choices of both things. Firstly, we repeated all four versions of the Gaussian Toy Model experiment, drawing θ_r from various distributions:

- A uniform distribution, both covering the wide range $\theta_r \sim \mathcal{U}(0, 1)$, and covering only part of the range $\theta_r \sim \mathcal{U}(0, 0.5)$

- A normal distribution, centered at $\theta = 0.5$, and with standard deviation varying between 0.01 and 0.1

- θ_r with a fixed value, either at the center of parameter space $\theta_r = 0.5$ or at a different location.

We also repeated our experiments using the Manhattan or L1 distance, instead of L2. We found very similar curves to those shown in Fig. 2 for the correct, overconfident, and underconfident cases. In the biased case, the different θ_r distributions led to different curves, but all of them clearly showed there was a bias. These figures are shown in §D. Therefore, we conclude that the proposed method is robust to different distributions for θ_r, and choices of distance metric.

4.3. Revealing when estimators are uninformative

As our second benchmark, we consider the case mentioned before in which the learned posterior estimator is equal to the prior $\hat{p}(\theta|x) = p(\theta)$. The reason why we are interested

in this example is that, in that case, the expected coverage probability calculated using HPD regions will be equal to $1 - \alpha$ for any value of α, as previously discussed. However, with TARP we have the ability to avoid this blindspot by sampling reference points in a manner dependent on x.

To make this concrete, we consider a one-dimensional example with a Gaussian prior $p(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$. Our 'forward model' in this case is simply generating a number n_x of data points, from $\{x_i\}_{i=1}^{n_x} \sim N(\theta, \sigma_x^2)$. In this conjugate model, we can easily derive the true posterior:

$$p\left(\theta|\, \{x_i\}_{i=1}^{n_x}\right) = \mathcal{N}(\mu|m, s), \tag{27}$$

$$s = \left(\frac{1}{\sigma_0^2} + \frac{n + x}{\sigma_x + x^2}\right)^{-1}, \tag{28}$$

$$m = s\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i x_i}{\sigma_x^2}\right) \tag{29}$$

We fix $n_x = 50$, $\mu_0 = 0$, $\sigma_0 = 1$ and $\sigma_x = 0.1$. We generate 500 samples from the forward model, and calculate expected coverage from an 'uninformative estimator' $\hat{p}(\theta|x) = p(\theta)$ in three ways: 1) using HPD regions, 2) using TARP regions where θ_r is drawn randomly from $\mathcal{U}(0, 1)$, and 3) using TARP regions where $\theta_r = x_0 + u$, where x_0 is the first observation, and $u \sim \mathcal{U}(0, 1)$. We expect the first two methods to have ECP equal to $1 - \alpha$, but not for the third.

We show the results in Fig. 4. First, we notice that when we use HPD regions, we get the correct expected coverage, even though the estimator is wrong (validating the theoretical discussion in § 2). This means that, in this case, HPD coverage could fool us into thinking our estimator is accurate when in reality it is completely uninformative. Interestingly, the same happens when we use TARP regions with reference points selected randomly from the prior (red line). This is because, as discussed in §2, Theorem 3 only holds in both directions when the choice of the region depends on x. Finally, as anticipated, the expected coverage is *not* $1 - \alpha$ when the sampling distribution for θ_r has some x-dependence. Therefore, we see how even when we introduce a small dependence on x to $\tilde{p}(\theta_r|x)$ in TARP reveals that the posterior estimator is not accurate. We further explore how the dependence of $\tilde{p}(\theta_r|x)$ on x affects our results in §D.

4.4. Gravitational Lensing

To test our algorithm in a more realistic and high-dimensional setting, we consider a simplified astrophysics problem: gravitational lensing source reconstruction. Gravitational lensing occurs in nature when light rays from a distant galaxy move along curved rather than straight paths due to the mass of another intervening galaxy (the 'lens') (Treu, 2010). The result is a highly-distorted, ring-shaped
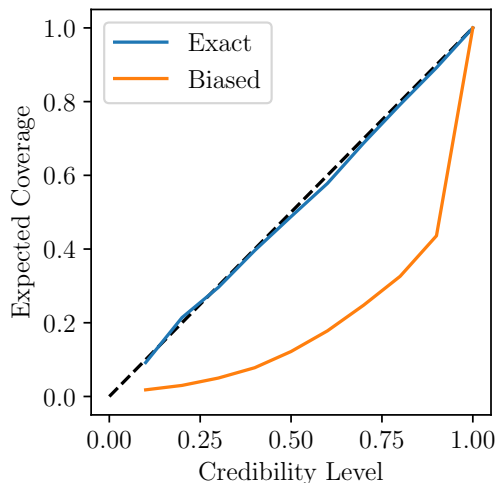
Figure 5. Expected coverage probability vs credibility level for our lensing example, for which tests based on HPD coverage are intractable. We see how, as expected, the exact posterior estimator (blue) accurately characterizes the posterior while the biased estimator (orange) does not.

image of the background galaxy. The goal of source reconstruction is to infer from a noisy image what the light from the source galaxy looks like without distortions, assuming the mathematical form of the distortions is perfectly known. In this high-dimensional setting, coverage checks based on the posterior's HPD region are intractable.

The simulator in this scenario samples the source galaxy's light θ from a multivariate-normal distribution that we fit to a dataset of galaxy images (Stone & Courteau, 2019; Stone et al., 2021). A matrix A encoding the lensing distortions are then applied, and the final observation is produced by adding Gaussian pixel noise of standard deviation σ_n, so that $x \sim \mathcal{N}(A\theta, \sigma_n^2)$. For computational convenience, we use 16×16-pixel source images and 32×32-pixel observations.

As shown in Adam et al. (2022) and reviewed in §E, posterior samples of θ can be generated using techniques from diffusion modeling. In general, this approach yields subtly biased posterior samples. However, with our multivariate-normal prior on θ, it is possible to generate unbiased posterior samples. We refer to samples from these as 'biased' and 'exact' in our results.

Fig. 5 shows the results for both the exact and the biased posterior estimators, using 500 simulations, and 1000 posterior samples per simulation, and sampling θ_r from the prior. As expected, our method gets the correct coverage for the exact estimator. It is important to stress that generative models are needed for parameter spaces of this dimensionality (256 parameters), and no previously existing methods could

calculate ECPs to assess the accuracy of such models. The biased estimator, on the other hand, produces a similar curve to that of the bottom right panel of Fig. 2, which indicates that it is indeed biased.

5. Conclusions

Testing the accuracy of estimated posteriors is a key element of parameter inference. While there exist well-establish convergence diagnostics for established sampling methods like MCMC, it is difficult to directly test the accuracy of posterior inferences, particularly those computed using deep learning methods. This is the case for both likelihood-based and simulation-based inference. In this paper, we introduced TARP coverage probabilities as a new technique to test the accuracy of estimated posteriors using posterior samples alone, when explicit posterior evaluations are not available. While our focus is testing posterior estimators based on generative machine learning models, our method could equally well be used to test the correctness of MCMC samples, although potentially at a great computational cost.

We have shown that this test is sufficient to prove that the inference is accurate, while other similar tests were necessary but not sufficient. We also tested the impact of the choice of $\hat{p}(\theta_r|x)$ and the distance metric used by the TARP method and found that they do not significantly affect our results. The exception to this is the case where the posterior estimator is equal to the prior, in which case TARP only works if $\hat{p}(\theta_r|x)$ has some dependency on x. It is left up to the user of the method to determine whether this is a risk.

We applied our test successfully to a variety of inference problems, in particular in cases where alternative methods fail, and showed that it scales well to high-dimensional posteriors. Therefore, we propose TARP coverage probabilities as a tool to test the accuracy of future posterior inference analyses from generative models.

6. Broader Impact

Our work is focused on checking the correctness of statistical inferences, which is an important open issue. We expect our work to have a positive societal impact by increasing the trustworthiness of machine learning applications to scientific problems across a wide variety of domains. As with any statistical method, however, incorrect application of our method (particularly through poor choice of the sampling distribution for θ_{ref}) could lead to invalid conclusions.

References

Adam, A., Coogan, A., Malkin, N., Legin, R., Perreault-Levasseur, L., Hezaveh, Y., and Bengio, Y. Posterior samples of source galaxies in strong gravitational lenses

with score-based priors. *arXiv preprint arXiv:2211.03812*, 2022.

Alsing, J., Charnock, T., Feeney, S., and Wandelt, B. Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notices of the Royal Astronomical Society*, 488(3):4440–4458, 2019.

Anau Montel, N. and Weniger, C. Detection is truncation: studying source populations with truncated marginal neural ratio estimation. In *36th Conference on Neural Information Processing Systems*, 11 2022.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Brehmer, J. Simulation-based inference in particle physics. *Nature Reviews Physics*, 3(5):305–305, January 2021. doi: 10.1038/s42254-021-00305-6.

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., and Cranmer, K. Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning. *Astrophys. J.*, 886(1):49, 2019. doi: 10.3847/1538-4357/ab4c41.

Charnock, T., Perreault-Levasseur, L., and Lanusse, F. Bayesian Neural Networks. *arXiv e-prints*, art. arXiv:2006.01490, June 2020. doi: 10.48550/arXiv.2006.01490.

Chen, Y., Zhang, D., Gutmann, M., Courville, A., and Zhu, Z. Neural approximate sufficient statistics for implicit models. *arXiv preprint arXiv:2010.10079*, 2020.

Coogan, A., Karchev, K., and Weniger, C. Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies. In *34th Conference on Neural Information Processing Systems*, 10 2020.

Coogan, A., Anau Montel, N., Karchev, K., Grootes, M. W., Nattino, F., and Weniger, C. One never walks alone: the effect of the perturber population on subhalo measurements in strong gravitational lenses. 9 2022.

Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., and Malz, A. I. Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362, 2020.

Dax, M., Green, S. R., Gair, J., Macke, J. H., Buonanno, A., and Schölkopf, B. Real-time gravitational wave science with neural posterior estimation. *Phys. Rev. Lett.*, 127:241103, Dec 2021. doi: 10.1103/PhysRevLett.127.241103. URL https://link.aps.org/doi/10.1103/PhysRevLett.127.241103.

de Witt, C. S., Gram-Hansen, B., Nardelli, N., Gambardella, A., Zinkov, R., Dokania, P., Siddharth, N., Espinosa-Gonzalez, A. B., Darzi, A., Torr, P., and Baydin, A. G. Simulation-based inference for global health decisions. 2020. doi: 10.48550/ARXIV.2005.07062. URL https://arxiv.org/abs/2005.07062.

Deistler, M., Goncalves, P. J., and Macke, J. H. Truncated proposals for scalable and hassle-free simulation-based inference. *arXiv preprint arXiv:2210.04815*, 2022.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Durkan, C., Murray, I., and Papamakarios, G. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2771–2781. PMLR, 2020.

Fearnhead, P. and Prangle, D. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. Bayesian inference using synthetic likelihood: asymptotics and adjustments. *Journal of the American Statistical Association*, (just-accepted):1–28, 2022.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical science*, pp. 457–472, 1992.

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., and Macke, J. H. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, sep 2020. ISSN 2050-084X. doi: 10.7554/eLife.56261. URL https://doi.org/10.7554/eLife.56261.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *stat*, 1050:10, 2014.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2404–2414. PMLR, 2019.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. *arXiv e-prints*, art. arXiv:1706.04599, June 2017. doi: 10.48550/arXiv.1706.04599.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pp. 4239–4248. PMLR, 2020.

Hermans, J., Banik, N., Weniger, C., Bertone, G., and Louppe, G. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Mon. Not. Roy. Astron. Soc.*, 507(2):1999–2011, 2021a. doi: 10.1093/mnras/stab2181.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021b.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Karchev, K., Anau Montel, N., Coogan, A., and Weniger, C. Strong-Lensing Source Reconstruction with Denoising Diffusion Restoration Models. In *36th Conference on Neural Information Processing Systems*, 11 2022a.

Karchev, K., Trotta, R., and Weniger, C. SICRET: Supernova Ia Cosmology with truncated marginal neural Ratio EsTimation. 9 2022b. doi: 10.1093/mnras/stac3785.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Legin, R., Hezaveh, Y., Levasseur, L. P., and Wandelt, B. Simulation-Based Inference of Strong Gravitational Lensing Parameters. 12 2021.

Linhart, J., Gramfort, A., and Rodrigues, P. L. Validation diagnostics for sbi algorithms based on normalizing flows. *arXiv preprint arXiv:2211.09602*, 2022.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. *arXiv e-prints*, art. arXiv:1805.09294, May 2018.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 343–351. PMLR, 2021.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

Marlier, N., Brüls, O., and Louppe, G. Simulation-based bayesian inference for multi-fingered robotic grasping, 2021. URL https://arxiv.org/abs/2109.14275.

Miller, B. K., Cole, A., Weniger, C., Nattino, F., Ku, O., and Grootes, M. W. swyft: Truncated marginal neural ratio estimation in python. *Journal of Open Source Software*, 7(75):4205, 2022a. doi: 10.21105/joss.04205. URL https://doi.org/10.21105/joss.04205.

Miller, B. K., Weniger, C., and Forré, P. Contrastive Neural Ratio Estimation. 10 2022b.

Mishra-Sharma, S. and Cranmer, K. Neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess. *Phys. Rev. D*, 105(6):063017, 2022. doi: 10.1103/PhysRevD.105.063017.

Montel, N. A., Coogan, A., Correa, C., Karchev, K., and Weniger, C. Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation. *Mon. Not. Roy. Astron. Soc.*, 518(2):2746–2760, 2022. doi: 10.1093/mnras/stac3215.

Ong, V. M.-H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. Likelihood-free inference in high dimensions with synthetic likelihood. *Computational Statistics & Data Analysis*, 128:271–291, 2018.

Papamakarios, G. and Murray, I. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848. PMLR, 2019.

Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Papernot, N. and McDaniel, P. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv e-prints*, art. arXiv:1803.04765, March 2018. doi: 10.48550/arXiv.1803.04765.

Perreault Levasseur, L., Hezaveh, Y. D., and Wechsler, R. H. Uncertainties in Parameters Estimated with Neural Networks: Application to Strong Gravitational Lensing. *Astrophys. J. Lett.*, 850(1):L7, 2017. doi: 10.3847/2041-8213/aa9704.

Prangle, D., Blum, M., Popovic, G., and Sisson, S. Diagnostic tools of approximate bayesian computation using the coverage property." arxiv preprint. *arXiv preprint arXiv:1301.3166*, 412, 2013.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

Ramesh, P., Lueckmann, J.-M., Boelts, J., Tejero-Cantero, Á., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. GATSBI: Generative Adversarial Training for Simulation-Based Inference. *arXiv e-prints*, art. arXiv:2203.06481, March 2022. doi: 10.48550/arXiv.2203.06481.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rozet, F. et al. Arbitrary marginal neural ratio estimation for likelihood-free inference. 2021.

Rubin, D. B. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151 – 1172, 1984. doi: 10.1214/aos/1176346785. URL https://doi.org/10.1214/aos/1176346785.

Schall, R. The empirical coverage of confidence intervals: Point estimates and confidence intervals for confidence levels. *Biometrical journal*, 54(4):537–551, 2012.

Skilling, J. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833 – 859, 2006. doi: 10.1214/06-BA127. URL https://doi.org/10.1214/06-BA127.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. URL http://arxiv.org/abs/1503.03585.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Stone, C. and Courteau, S. The Intrinsic Scatter of the Radial Acceleration Relation. *The Astrophysical Journal*, 882 (1):6, September 2019. doi: 10.3847/1538-4357/ab3126.

Stone, C., Courteau, S., and Arora, N. The Intrinsic Scatter of Galaxy Scaling Relations. *The Astrophysical Journal*, 912(1):41, May 2021. doi: 10.3847/1538-4357/abebe4.

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

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P., Greenberg, D., and Macke, J. sbi: A toolkit for simulation-based inference. *The Journal of Open Source Software*, 5(52):2505, August 2020. doi: 10.21105/joss.02505.

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. U. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1):1–31, 2022.

Treu, T. Strong lensing by galaxies. *Annual Review of Astronomy and Astrophysics*, 48:87–125, 2010.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wagner-Carena, S., Park, J. W., Birrer, S., Marshall, P. J., Roodman, A., and Wechsler, R. H. Hierarchical Inference with Bayesian Neural Networks: An Application to Strong Gravitational Lensing. *Astrophys. J.*, 909(2):187, 2021. doi: 10.3847/1538-4357/abdf59.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4697–4708. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf.

Zhu, Y. and Zabaras, N. Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366:415–447, August 2018. doi: 10.1016/j.jcp.2018.04.018.

Zuo, Y., Chen, C., Li, X., Deng, Z., Chen, Y., Behler, J., Csányi, G., Shapeev, A. V., Thompson, A. P., Wood, M. A., and Ong, S. P. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A*, 124(4):731–745, 2020. doi: 10.1021/acs.jpca.9b08723. URL https://doi.org/10.1021/acs.jpca.9b08723. PMID: 31916773.
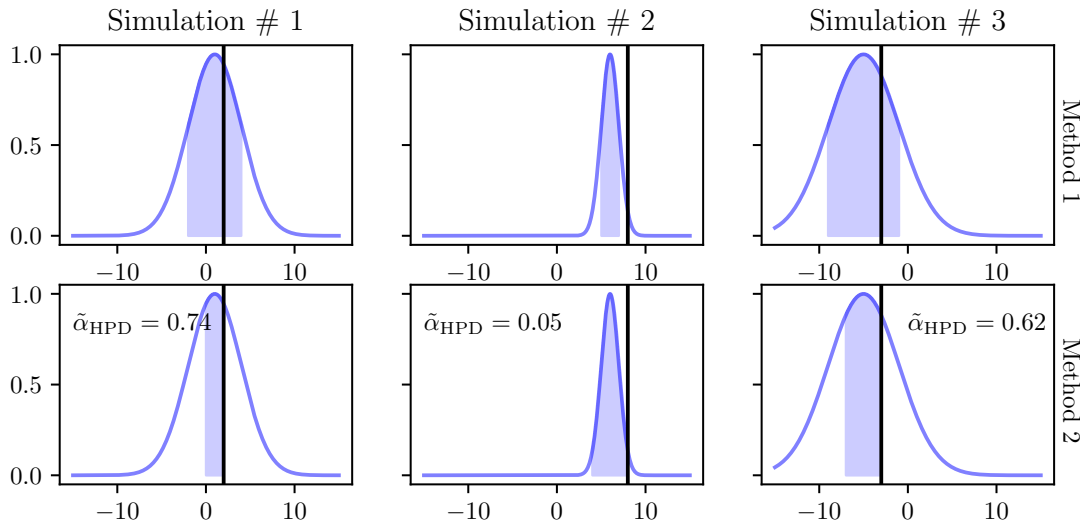
Figure 6. This figure illustrates the intuition behind the two ways of calculating high posterior density coverages. Each column shows one of three simulations in a toy example. The blue curves show the predicted posteriors, and the black vertical lines show the corresponding truths. We want to calculate the coverage for the 68% credibility level ($\alpha = 0.32$). The first approach, shown in the top row, would consist of calculating the $1 - \alpha$ credibility region, then checking how often the truth is in the said region (in this case twice, so the coverage is $2/3$). The alternative approach, described in §3.1, and shown in the bottom row, is to find the HPD region defined by the truth, then find how often $\tilde{\alpha}_{\text{HPD}} < \alpha$, which is again twice. The plot illustrates the fact that these two approaches are exactly equivalent.

A. Connection between both definitions

§3.1 discussed the differences between two possible methods for calculating coverage probabilities, both for HPD and TARP regions. We try to build more intuition behind that connection in this appendix. Focusing first on the case of HPD regions, shown in Fig. 6. The first method, perhaps more intuitive but far more inefficient, would be to calculate the $1 - \alpha$ credibility region, then check how often the truth is in the said region for each simulation, and for multiple values of alpha (notice the nested loop). The second method, a consequence of the very important Lemma 1, and already used by Algorithm 1 would be to find the HPD region defined by the truth for each simulation, and its corresponding credibility level $1 - \tilde{\alpha}_{\text{HPD}}$. We can then calculate the coverage for the $1 - \alpha$ level as $\sum_{i=1}^{N} \tilde{\alpha}_{\text{HPD}} \geq \alpha$, where N is the number of simulations.

A similar logic applies to TARP credibility regions. While we could find the radius from the reference point, such that α reaches a certain value, it is far more computationally efficient, and equivalent, to use the credibility regions defined by the true values, as shown in Fig. 7. This is the method used by Algorithm 2.

B. Intuition about over and under confident plots

Practitioners used to applying coverage probabilities to validate SBI analysis will be used to seeing over- and underconfident curves, such as those in the blue curves of Fig. 2. However, the same figure shows how the TARP method produces different curves for over- and underconfident posterior estimators. The aim of this appendix is to provide some intuition behind these differences.

Firstly, we focus on underconfident posteriors, shown in the top panel of Fig. 8. In this case, we see that the TARP coverage tends to be close to 0.5. This is because regardless of where the random reference point is, if the truth is close to the peak of the posterior, the TARP area is likely to cover approximately half of the distribution. On the other hand, for overconfident posteriors, shown in the bottom panel of Fig. 8, we see that the TARP coverage tends to be close to either 0 or 1. This is because regardless of where the random reference point is, if the truth is far from the peak of the posterior, the TARP area is likely to cover either the whole distribution, or none of it.
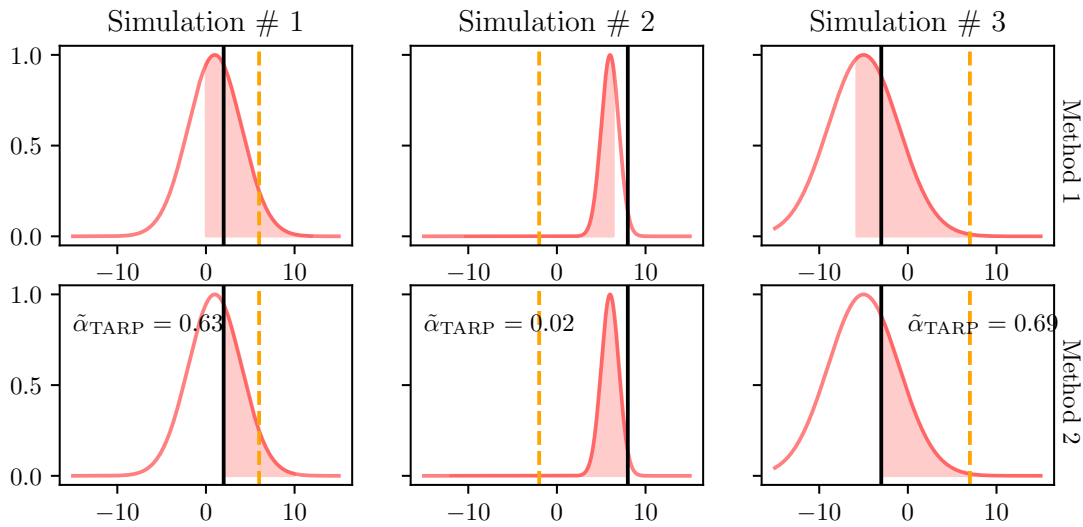
Figure 7. Similarly to Fig. 6, this figure illustrates the intuition behind the two ways of calculating 'distance to random point' coverages. Each column shows one of three simulations in a toy example. The red curves show the predicted posteriors, and the black vertical lines show the corresponding truths. The orange lines show the randomly selected reference points. We want to calculate the coverage for the 68% credibility level ($\alpha = 0.32$). The first approach, shown in the top row, would consist of finding the $1 - \alpha$ credibility region centered around the reference point, then checking how often the truth is in the said region (in this case twice, so the coverage is $2/3$). The alternative approach, shown in the bottom row, is to find the TARP region defined by the reference and the truth, then find how often $\tilde{\alpha}_{\mathrm{HPD}} < \alpha$, which is again twice. The plot illustrates the fact that these two approaches are exactly equivalent.
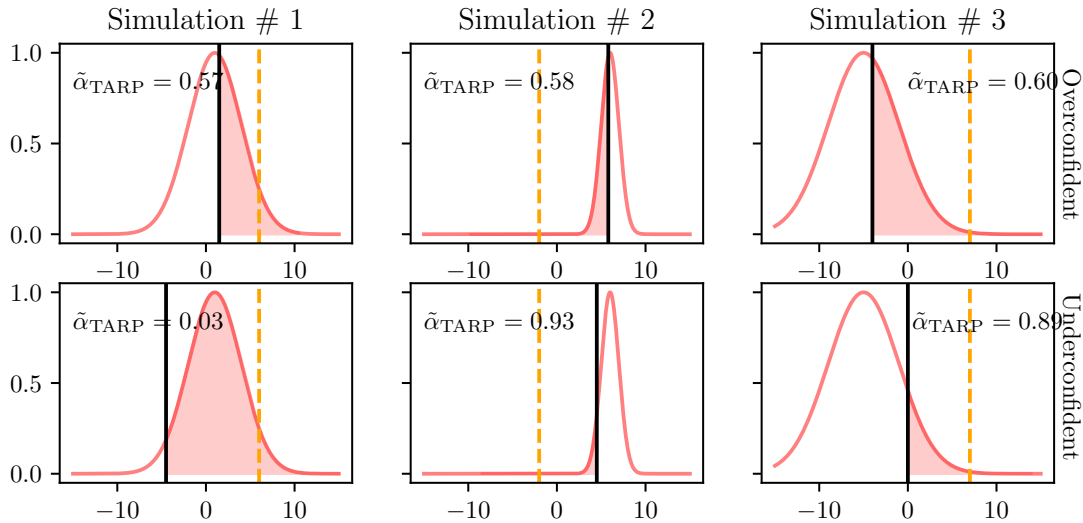

Figure 8. This figure illustrates the reason behind the shapes of over and underconfident curves obtained using TARP coverage, such as those in Fig. 2. The top shows three example simulations with overconfident predictions, while the bottom shows underconfident predictions. The figure shows that TARP coverages tend to be close to 0.5, while for overconfident regions they tend to be close to either 0 or 1.
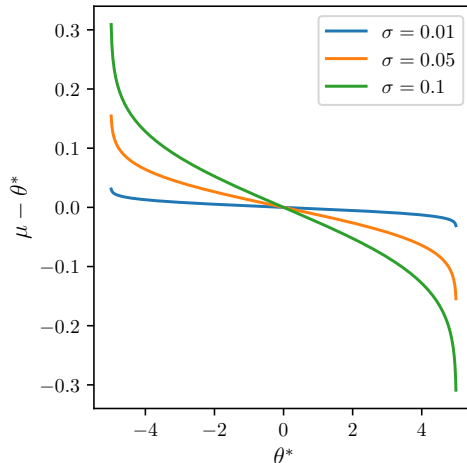
Figure 9. The position-dependent function used as the mean in the *biased case* in §4.1, for three different values of sigma.

C. Biased case experiment function

Fig. 9 shows the function (26), used in §4.1 for the one-dimensional case, as the means of the normal distributions. The function shows how, when θ^* is zero, the distributions are centered at the correct value, whereas as we move away from zero, the posterior estimator will be increasingly biased.

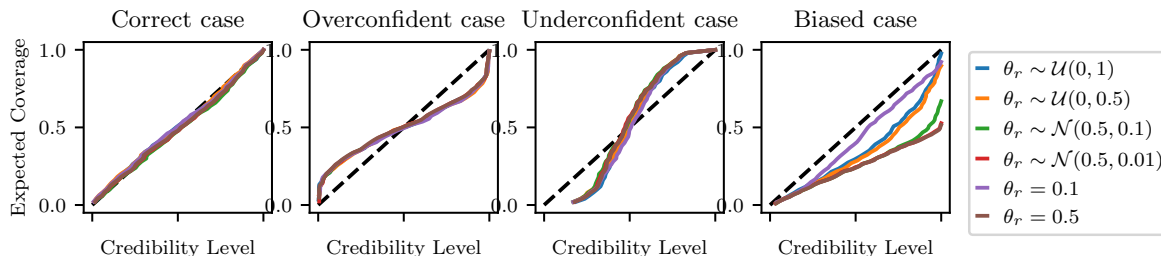## D. Dependence on θ_r distribution and distance metric

Figure 10. The Gaussian Toy Model experiment described in §4.1, drawing the reference points θ_r from different distributions, as described in §4.2.

Fig. 10 shows the same as Fig. 2, but varying the distribution used to draw θ_r. We find that this only makes a difference in the biased case, but even then there is clear evidence of bias for all distributions. Fig. 11 shows the same, comparing the use of L1 and L2 as distance metrics. We find no appreciable differences in this case. We, therefore, conclude that our method is robust to choices of θ_r distribution and distance metric.

In §4.3, we discussed how when we have a distribution $\tilde{p}(\theta_r|x)$ that has some dependency on x, the TARP method reveals an inaccurate posterior estimator, in the case when the posterior estimator is simply recovering the prior. Fig. 12 shows what happens to this experiment for different distributions. We see how the distributions that do not depend on x, shown in continuous lines, do not detect the inaccurate posterior estimator as expected. On the other hand, the distributions that depend strongly on x, shown as dotted lines, very clearly detect the inaccurate estimator. Finally, we show a distribution with a weaker dependence on x, where TARP does lie away from the diagonal line, but much closer than the other x-dependent distributions, as expected.
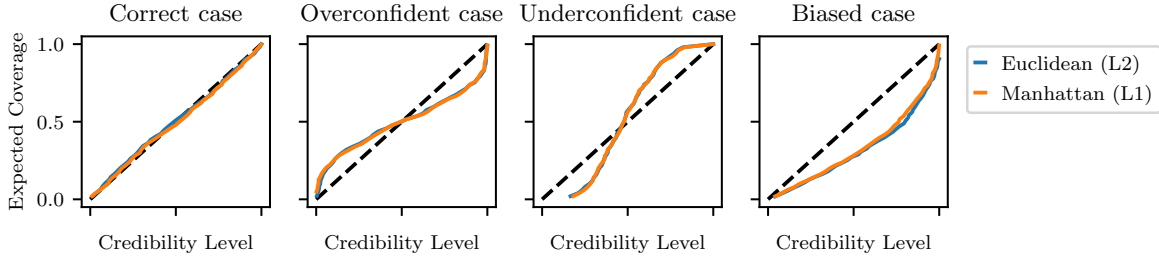
Figure 11. The Gaussian Toy Model experiment described in §4.1, using L1 or L2 distance metrics, as described in §4.2.
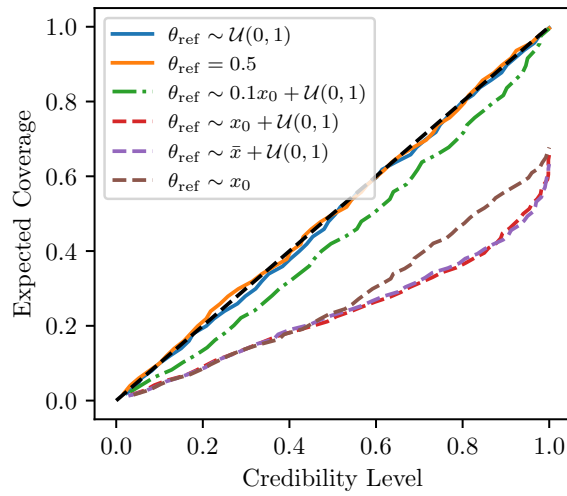

Figure 12. The same expected coverage vs credibility level plot shown in Fig. 4, for different $\tilde{p}(\theta_r|x)$ distributions. The continuous lines show distributions that do not depend on x, the dash-dotted line shows a distribution with a weak dependence on x, and the dotted lines show a stronger dependence.

E. Gravitational lensing experiment details

As shown in Adam et al. (2022), gravitational lensing source reconstruction can be performed using techniques from score-based modeling. Here we summarize the key ideas behind score-based modeling and how we generate biased and exact posterior samples.

Score-based modeling works by perturbing a training dataset sampled from a prior $p(\theta)$ with noise of increasing scales indexed by $t \in [0, T]$. Here $t = 0$ corresponds to unperturbed data $(p_0(\theta) = p(\theta))$ and $t = T$ corresponds to perturbing the data so much it is buried under noise and follows a Gaussian distribution $(p_T(\theta) = \mathcal{N}(\theta|0, \sigma_T^2))$. The noising process be described by the stochastic differential equation (SDE) (Song et al., 2020)

$$\mathrm{d}\theta_t = f(t, \theta)\,\mathrm{d}t + g(t)\,\mathrm{d}w \;, \tag{30}$$

where w is a standard Wiener process. Using denoising score-matching (Hyvärinen & Dayan, 2005; Vincent, 2011; Song et al., 2020), a neural network can be trained to approximate the time-dependent prior score $\boldsymbol{\nabla}_\theta p_t(\theta)$, where $p_t(\theta)$ is the distribution over data perturbed by the noising process up to time t. Given the prior score, samples can be generated by solving the corresponding reverse SDE (RSDE) backward in time, starting with samples from p_T:

$$\mathrm{d}\theta = \left[f(t, \theta) - g^2(t)\boldsymbol{\nabla}_\theta \log p_t(\theta)\right]\mathrm{d}t + g(t)\,\mathrm{d}w \;, \tag{31}$$

where here $\mathrm{d}t$ is a negative timestep.

For simplicity, instead of fitting a score-based model, we fit a multivariate Gaussian to the PROBES dataset of galaxy images as our prior, giving $p(\theta) = \mathcal{N}(\mu_0, \Sigma_0)$. We use the variance-exploding SDE from Song et al. (2020) as our noise process. The prior at time t is thus $p_t(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0 + \sigma_t^2\mathbb{I})$, where σ_t^2 is the variance of the noise process at time t. This expression can be used to evaluate the prior score analytically.

To modify the sampling procedure to generate samples from $p(\theta|x)$ for some observation x, we must condition the score in the RSDE, replacing the prior score with the posterior score:

$$\mathrm{d}\theta = \left[f(t, \theta) - g^2(t)\boldsymbol{\nabla}_\theta \log p_t(\theta|x)\right]\mathrm{d}t + g(t)\,\mathrm{d}w \;, \tag{32}$$

By Bayes' rule, the posterior score is

$$\boldsymbol{\nabla}_\theta \log p_t(\theta|x) = \boldsymbol{\nabla}_\theta \log p_t(x|\theta) + \boldsymbol{\nabla}_\theta \log p_t(\theta)\,, \tag{33}$$

where the first term on the RHS is the score of the likelihood. As pointed out in Adam et al. (2022), this time-dependent likelihood is in general intractable but can be approximated as

$$\hat{p}_t(x|\theta) = \mathcal{N}(x|A\theta, \sigma_n^2\mathbb{I} + \sigma_t^2 AA^T) \approx p_t(x|\theta)\,, \tag{34}$$

where the matrix A encodes the lensing distortions and σ_n is the standard deviation of the noise in the observation (see § 4.4). *However*, when $p(\theta)$ is a multivariate Gaussian, the time-dependent likelihood *is tractable*, evaluating to

$$p_t(x|\theta) = \mathcal{N}(x|A\theta_c(\theta), \sigma_n^2\mathbb{I} + A\Sigma_c A^T)\,, \tag{35}$$

where

$$\Sigma_c := \left(\Sigma_0^{-1} + \sigma_t^{-2}\mathbb{I}\right)^{-1}\,, \qquad \theta_c(\theta) := \sigma_t^{-2}\Sigma_c \theta\,. \tag{36}$$

We, therefore, have two methods for sampling the posterior for the source galaxy's light: solving the RSDE (32) using the exact time-dependent likelihood ((35)) or the approximate, biased one ((34)). We refer to these as the 'exact' and 'biased' samplers respectively.

Finally, we solve both the exact and biased RSDEs by discretizing with the Euler-Maruyama method (see e.g. Song et al. (2020)). We find 300 steps are sufficient to ensure convergence.