
Cost-aware simulation-based inference

Ayush Bharti
Aalto University

Daolang Huang
Aalto University

Samuel Kaski
Aalto University
University of Manchester

François-Xavier Briol
University College London

Abstract

Simulation-based inference (SBI) is rapidly becoming the preferred framework for estimating parameters of intractable models in science and engineering. A significant challenge in this context is the large computational cost of simulating data from complex models, and the fact that this cost often depends on parameter values. We therefore propose *cost-aware SBI methods* which can significantly reduce the cost of existing sampling-based SBI methods, such as neural SBI and approximate Bayesian computation. This is achieved through a combination of rejection and self-normalised importance sampling, which reduces the number of expensive simulations needed. Our approach is studied extensively on models from epidemiology to telecommunications engineering, where we obtain significant reductions in the overall cost of inference.

1 INTRODUCTION

Many scientific disciplines use computational and mechanistic models to study complex physical, biological, or sociological phenomena. Such models, or *simulators*, are often poorly suited for standard Bayesian inference methods due to their intractable likelihood functions. Simulation-based inference (SBI; Cranmer et al. (2020)) addresses this issue by approximating posteriors using repeated simulations from the model instead of evaluations of the likelihood. These simulations are typically used to measure similarity with the observed data (Lintusaari et al., 2017; Sisson, 2018; Briol et al., 2019), or for training a conditional density estimator (Lueckmann et al., 2021; Zammit-Mangion et al., 2024), and a large number of these is therefore

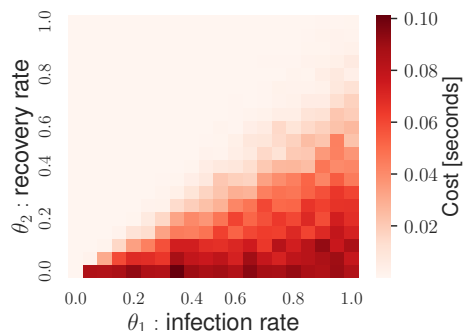


Figure 1: Estimated cost of the temporal susceptible-infected-recovered model considered in Section 4.2.

required. A significant bottleneck limiting the use of SBI in many real-world cases is therefore the cost of simulators. Even if a model takes only a minute for a single simulation, state-of-the-art SBI often requires tens of thousands of simulations which can take several days to run. Needless to say, costlier simulators taking hours or days per simulation, such as most tsunami models (Behrens and Dias, 2015), wind farm models (Kirby et al., 2023) or nuclear fusion simulators (Hoppe et al., 2021), are currently out of the reach of SBI.

Surprisingly, existing SBI methods tend to ignore the fact that the cost of a simulator will often depend on the parameter value for which we are simulating, and are therefore unable to take advantage of this property. For example, many models of disease spread in epidemiology have varying cost (Kypraios et al., 2017; McKinley et al., 2018); as the recovery rate decreases and the infection rate increases, the disease spreads faster, and the computational cost of a simulation increases. This is illustrated in Figure 1 for a model we will study in Section 4.2, in which case some parametrisations are up to 10 times more expensive than cheaper parametrisations. Varying cost is also a feature in many other real-world applications, including aircraft vehicle design simulators (Cobb et al., 2023) where simulating aircraft with large number of propellers and wings is slower, and graph-based radio propagation models where simulating graphs with more nodes takes longer (Bharti et al., 2022a).

In this paper, we tackle this issue by proposing the first family of *cost-aware* alternatives to popular SBI methods such as neural posterior estimation (NPE) (Papamakarios and Murray, 2016; Radev et al., 2022), neural likelihood estimation (NLE) (Papamakarios et al., 2019) and approximate Bayesian computation (ABC) (Sisson, 2018). Cost-aware SBI uses self-normalised importance sampling with an importance distribution constructed to encourage sampling from the cheaper parameterisations of the model. This leads to SBI methods capable of significant computational savings without compromising on accuracy. We demonstrate these advantages on real-world simulators from the fields of epidemiology and radio propagation, leading to significant reductions in the cost of inference in both cases. Interestingly, cost-aware SBI complements, rather than replaces, many of the existing approaches for sample-efficiency in the literature.

2 BACKGROUND

Consider a parametric model (i.e. a family of Borel probability distributions) on some space $\mathcal{X} \subseteq \mathbb{R}^d$ with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. We will assume this is a simulator-based model, meaning that the density $p(\cdot | \theta)$ (and therefore the likelihood) associated with this model cannot be evaluated point-wise, but it is possible to simulate independent realisations $x \sim p(x | \theta)$ ¹ for a fixed $\theta \in \Theta$ (albeit at a potentially significant computational cost). Given m_o observations $x_o = (x_{o1}, \dots, x_{om_o}) \in \mathcal{X}^{m_o}$ and a prior $p(\theta)$, we are interested in the posterior

$$p(\theta | x_o) = \frac{p(x_o | \theta)p(\theta)}{\int_{\Theta} p(x_o | \theta)p(\theta)d\theta} \propto p(x_o | \theta)p(\theta),$$

where \propto denotes proportionality up to a multiplicative constant. As the likelihood is intractable, SBI methods rely on simulations from the joint $(x, \theta) \sim p(x, \theta) = p(x | \theta)p(\theta)$ to approximate the posterior distribution. This is typically achieved by first sampling from the prior $\theta \sim p(\theta)$ and then simulating m data points from the model $(x_1, \dots, x_m) \sim p(x | \theta)$. Suppose the expected cost of simulating a single realisation from the model with parameter θ is given by $c(\theta)$, where $c : \Theta \rightarrow (0, \infty)$ is called the *cost function*. Then, the total expected cost of simulating m realisations from n parameter values $\theta_1, \dots, \theta_n$ becomes $\sum_{i=1}^n c(\theta_i)m$. In this paper, the cost function will usually refer to computational cost (measured in units of time), but it could conceptually represent other forms of cost such as financial cost or memory cost. We now introduce some popular SBI methods, emphasising the computationally costly operations in each method.

¹We abuse notation by writing $x \sim p(x | \theta)$ for independent sampling from the distribution with density $p(\cdot | \theta)$.

Neural posterior estimation (NPE). NPE (Radev et al., 2022) is a neural SBI method that uses a conditional density estimator, such as a normalising flow (Papamakarios et al., 2021), to learn an approximate mapping from data to the posterior: $x \mapsto p(\theta | x)$. Let $q_\phi(\theta | x)$ denote such a conditional normalising flow parameterised by the vector ϕ . The parameter ϕ can be estimated by minimising the empirical loss

$$\begin{aligned} \ell_{\text{NPE}}(\phi) &= -\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \log q_\phi(\theta_i | x_{ij}) \\ &\approx -\mathbb{E}_{\theta \sim p(\theta)} [\mathbb{E}_{x \sim p(x | \theta)} [\log q_\phi(\theta | x)]], \end{aligned} \quad (1)$$

using the dataset $\{(\theta_i, \{x_{ij}\}_{j=1}^m)\}_{i=1}^n$ generated from the joint. After training, inference is amortised; given the estimated flow parameters $\hat{\phi}$ and a new observed dataset x_o , an approximation of the posterior is given by simply evaluating $p_{\text{NPE}}(\theta | x_o) := q_{\hat{\phi}}(\theta | x_o)$. However, training q_ϕ requires many realisations from the simulator in order to accurately approximate the expected loss, which can be a significant bottleneck if the simulator is computationally expensive.

Neural likelihood estimation (NLE). In NLE (Papamakarios et al., 2019), a conditional density estimator is used to learn a mapping from the parameters to the likelihood: $\theta \mapsto p(x | \theta)$. Given simulations from the joint, the density estimator $q_\phi(x | \theta)$ is trained by minimising the following empirical loss

$$\begin{aligned} \ell_{\text{NLE}}(\phi) &= -\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \log q_\phi(x_{ij} | \theta_i) \\ &\approx -\mathbb{E}_{\theta \sim p(\theta)} [\mathbb{E}_{x \sim p(x | \theta)} [\log q_\phi(x | \theta)]]. \end{aligned} \quad (2)$$

The trained model is used to obtain an approximate posterior $p_{\text{NLE}}(\theta | x_o) \propto q_{\hat{\phi}}(x_o | \theta)p(\theta)$ which can typically be sampled from using Markov chain Monte Carlo (MCMC). Once again, the costly step is simulating the realisations used to estimate the expected loss.

Approximate Bayesian computation (ABC). ABC (Lintusaari et al., 2017; Sisson, 2018) is an SBI framework that relies on a distance $\varrho : \mathcal{X}^{m_o} \times \mathcal{X}^m \rightarrow [0, \infty)$ to create an approximate posterior

$$p_{\text{ABC}}(\theta | x_o) \propto \mathbb{E}_{x \sim p(x | \theta)} [1_{\{\varrho(x_o, x) \leq \epsilon\}}] p(\theta).$$

The classic accept-reject ABC method (Pritchard et al., 1999) to approximate p_{ABC} involves repeatedly sampling $\theta \sim p(\theta)$ from the prior, simulating data $x \sim p(x | \theta)$ using the simulator, and then accepting θ if the distance $\varrho(x_o, x)$ falls below a tolerance threshold $\epsilon > 0$. Even the more advanced ABC methods based on sequential Monte Carlo (Beaumont et al., 2009; Del Moral et al., 2011) and regression adjustment

(Beaumont et al., 2002) rely on this two-step process. To ensure ABC provides an accurate approximation of $p(\theta|x_o)$, we must take ϵ to be small (Frazier et al., 2018), but this comes at the cost of a high rejection rate which increases the number of simulator calls. ABC can therefore be very computationally demanding when dealing with expensive simulators.

Related work. To tackle the challenge of expensive simulations, a number of *sample-efficient* SBI methods have been proposed. A full review is beyond the scope of this paper, but these approaches can be broadly categorised as follows: (i) methods that focus on improved sampling of the posterior, hence reducing the total number of parameters for which simulations are needed e.g. through sequential sampling (Sisson et al., 2007; Beaumont et al., 2009; Del Moral et al., 2011; Greenberg et al., 2019; Papamakarios et al., 2019; Hermans et al., 2020) or Gaussian process surrogates (Meeds and Welling, 2014; Wilkinson, 2014; Gutmann and Corander, 2016; Järvenpää et al., 2019); (ii) methods that improve upon independent simulations of data given a fixed parameter value, e.g. using quasi-Monte Carlo or Bayesian quadrature (Niu et al., 2023; Bharti et al., 2023), hence reducing the number of samples needed per parameter value; and (iii) methods using side-information to avoid unnecessary simulations, e.g. multi-fidelity methods (Prescott and Baker, 2020, 2021; Warne et al., 2022; Prescott et al., 2024), methods based on partial simulations (Prangle, 2016), and expert-in-the-loop approaches (Bharti et al., 2022a). Our proposed cost-aware sampling method falls in the third category as it uses knowledge of the cost of simulations to guide sampling in the parameter space. Importantly, it is complementary to all of the approaches described above, including methods that parallelise computations (Kulkarni and Moritz, 2023). Note that importance sampling has previously been used in the context of SBI by Dax et al. (2023); Prangle and Viscardi (2023), but their focus was solely on variance reduction rather than on reducing the computational cost without compromising on the variance.

3 METHOD

We now describe cost-aware importance sampling in Section 3.1, discuss practical considerations in Sections 3.2 and 3.3, then present cost-aware SBI in Section 3.4.

3.1 Cost-aware importance sampling

Let π be a target density on $\Theta \subseteq \mathbb{R}^p$. In the context of SBI, π will either be the prior or the approximate posterior, but for now it can be considered arbitrary. We assume that sampling independent realisations $\theta \sim \pi$

is cheap, however, each sampled θ will be used for a downstream task with expected cost $c(\theta)$. In SBI, $c(\theta)$ will be the expected cost of simulating from the simulator. In scenarios where the cost varies across Θ , the standard approach of sampling n times independently from π (in a cost-agnostic manner) can lead to a large downstream cost. We therefore aim to reduce the downstream cost by using $c(\theta)$ to guide our sampling towards cheaper regions of Θ . Instead of sampling independently from the target $\pi(\theta)$, we propose to use a *cost-aware proposal* distribution with density

$$\tilde{\pi}_g(\theta) \propto \frac{\pi(\theta)}{g(c(\theta))},$$

and re-weight samples so as to approximate π . Here, $g : (0, \infty) \rightarrow (0, \infty)$ is a non-decreasing function that governs the degree to which we penalise for cost, and the composition $g \circ c$ must be strictly positive. The assumption that g is non-decreasing is needed to ensure that larger costs are penalised at least as much as smaller costs, and the assumption that $g \circ c$ is strictly positive can trivially be satisfied for any cost function which is strictly positive. The function g is a user-specified choice, and we discuss its selection in Section 3.3. Interestingly, when g is a constant function, our approach does not penalise for cost and therefore reduces to i.i.d. sampling (and hence to standard SBI methods when used in that context). More generally, $\tilde{\pi}_g$ can be interpreted as a form of exponential tilting (Siegmund, 1976) of π with the function $\exp(-\log g(c(\theta)))$, but we are not aware of any focus on computational cost of a downstream task in this literature. The closest approach is that of Neyman allocation (Neyman, 1934), which solves the stratification problem when sampling cost varies; however, the notion of cost in this setting is that of sampling from π , rather than the cost of a downstream task.

Outside of a few specific choices of π , c , and g , we will typically not have access to closed-form expressions for the normalisation constant of the proposal $\tilde{\pi}_g$ (see Table 4 in Appendix B.1 for a few exceptions where we do). However, this does not create an issue for sampling from $\tilde{\pi}_g$, which we propose to do via rejection sampling (von Neumann, 1951) with proposal π . Samples from $\tilde{\pi}_g$ are generated by first sampling candidates $\theta \sim \pi$, and then accepting those candidates with probability $A(\theta) = \tilde{\pi}_g(\theta)/M\pi(\theta)$, where $M > 0$ is a constant that satisfies the condition $\tilde{\pi}_g(\theta) \leq M\pi(\theta)$. The following proposition formalises the conditions under which the cost-aware proposal $\tilde{\pi}_g$ is a valid density, and gives an expression for $A(\theta)$. The proof is in Appendix A.1.

Proposition 1. *Let π be a density function on Θ and assume $g : (0, \infty) \rightarrow (0, \infty)$ is a non-decreasing function and $g \circ c$ is strictly positive; i.e. $g_{\min} := \inf_{\theta \in \Theta} g(c(\theta)) > 0$. Then, $\tilde{\pi}_g$ is also a density function.*

In addition, we can sample from $\tilde{\pi}_g$ using rejection sampling with proposal π and acceptance probability

$$A(\theta) = \frac{g_{\min}}{g(c(\theta))}. \quad (3)$$

This rejection sampling algorithm is very widely applicable since there are no restrictions on π , and the conditions on g and $g \circ c$ are minimal as previously discussed. In addition, the acceptance probability $A(\theta)$ does not require knowledge of the normalisation constant of $\tilde{\pi}_g$. It only depends on c , which we assume known, and g_{\min} , which is available since we can choose g . In particular, it does not rely on any costly down-stream task, making the approach computationally cheap. Alternative proposals, such as those used in adaptive rejection sampling (Gilks and Wild, 1992; Gilks et al., 1995), may provide an algorithm with slightly smaller rejection rate, but the tractability of the acceptance probability makes this approach particularly convenient.

We note that when the minima of g and c are small, then g_{\min} may also be small. This means that $A(\theta)$ will approach zero, leading to a higher rejection rate. The cost of sampling from $\tilde{\pi}_g$ will increase in such scenarios, but this is not a significant issue as, for SBI, π usually corresponds to a simple distribution such as a uniform or a Gaussian, which is extremely cheap to sample from. Furthermore, this issue can also be alleviated by choosing the penalty function to be of the form $g(c(\theta)) = \max(1, h(c(\theta)))$ for some non-decreasing function h , which ensures that $g_{\min} = 1$, and hence $A(\theta)$ is not too close to zero.

Now that we can sample from $\tilde{\pi}_g$, we can replace samples from π with those from $\tilde{\pi}_g$ to minimise the total downstream cost. The fact that we sample from $\tilde{\pi}_g$ instead of π can be accounted for by weighting the realisations using self-normalised importance sampling (Trotter and Tukey, 1956). The unnormalised weights in this case can be computed as

$$w(\theta) = \frac{\pi(\theta)}{\tilde{\pi}_g(\theta)} = \frac{B\pi(\theta)g(c(\theta))}{\pi(\theta)} \propto g(c(\theta)), \quad (4)$$

where B is the normalisation constant for $\tilde{\pi}_g$. Given realisations $\{\theta_i\}_{i=1}^n$ from $\tilde{\pi}_g$, normalised weights can be obtained from the unnormalised weights as follows:

$$w_{\text{Ca}}(\theta_i) = \frac{w(\theta_i)}{\sum_{j=1}^n w(\theta_j)} = \frac{g(c(\theta_i))}{\sum_{j=1}^n g(c(\theta_j))}. \quad (5)$$

The advantage of using self-normalisation is that the weights do not depend on the normalisation constants of either the proposal $\tilde{\pi}_g$ or the target π .

Now that we have our cost-aware importance sampling scheme for π , we may be interested in computing the expected value of an arbitrary integrand $f : \Theta \rightarrow \mathbb{R}$

with respect to π , i.e., $\mu = \int_{\Theta} f(\theta)\pi(\theta)d\theta$. This can be achieved using the estimator $\hat{\mu}_n^{\text{Ca}} = \sum_{i=1}^n w_{\text{Ca}}(\theta_i)f(\theta_i)$, which is consistent and has bounded weights and finite variance σ_{Ca}^2 under very mild conditions. In addition, we can relate the variance of this estimator (a key measure of its efficiency) to the variance σ_{MC}^2 of the Monte Carlo estimator $\hat{\mu}_n^{\text{MC}} = \frac{1}{n} \sum_{i=1}^n f(\theta_i)$; see Appendix A.2 for the proof.

Proposition 2. *Suppose g is non-decreasing and $g_{\min} > 0$. Then, $\hat{\mu}_n^{\text{Ca}} \rightarrow \mu$ with probability 1 as $n \rightarrow \infty$.*

Furthermore, assume $g_{\max} = \sup_{\theta \in \Theta} g(c(\theta)) < \infty$ and f is π -square-integrable ($\int_{\Theta} f(\theta)^2 \pi(\theta)d\theta < \infty$). Then

$$\frac{g_{\min}}{ng_{\max}} \leq w_{\text{Ca}}(\theta_i) \leq \frac{g_{\max}}{ng_{\min}} \quad \forall i \in \{1, \dots, n\},$$

and we have that

$$\frac{g_{\min}}{g_{\max}} \left(\sigma_{\text{MC}}^2 - \frac{\mu^2}{n} \right) \leq \sigma_{\text{Ca}}^2 \leq \frac{g_{\max}}{g_{\min}} \left(\sigma_{\text{MC}}^2 - \frac{\mu^2}{n} \right).$$

The assumptions are again extremely minimal. Assuming f is π -square-integrable is standard and required for Monte Carlo to have finite variance. The assumption that $g_{\max} < \infty$ holds when c is upper bounded, or can be enforced by choosing g of the form $g(c(\theta)) = \min(h(c(\theta)), C)$, where h is a strictly-positive non-decreasing function and C is some large constant.

A direct corollary of this result is that the cost-aware estimator does not have infinite variance, which is a regular issue for importance sampling (see Chapter 9 of Owen (2013)). It also implies that cost-aware sampling is at worst g_{\max}/g_{\min} less efficient than Monte Carlo, meaning we would like g_{\min}/g_{\max} to be large.

As a side note, we remark that we could be tempted to use the optimal self-normalised importance sampling proposal; i.e. the proposal which minimises the variance. However, this is proportional to $|f(\theta) - \mu|\pi(\theta)$ (Hesterberg, 1988, Ch. 2), which corresponds to having $g(c(\theta)) \propto |f(\theta) - \mu|^{-1}$ and is sadly not a non-decreasing function. This choice is therefore not a valid proposal for our framework.

Another alternative measure of efficiency is the *effective sample size (ESS)* (Owen, 2013, Section 9.3):

$$\text{ESS} = \frac{(\sum_{i=1}^n w(\theta_i))^2}{n \sum_{i=1}^n w(\theta_i)^2} = \frac{(\sum_{i=1}^n g(c(\theta_i)))^2}{n \sum_{i=1}^n g(c(\theta_i))^2}.$$

This is an attractive measure since the variance is integrand-specific but the ESS is not. Again, we can bound this quantity using our minimal assumptions; see Appendix A.3 for the proof.

Proposition 3. *Suppose g is non-decreasing and $0 <$*

$g_{\min} \leq g_{\max} < \infty$. Then:

$$\left(\frac{g_{\min}}{g_{\max}}\right)^2 \leq ESS \leq \left(\frac{g_{\max}}{g_{\min}}\right)^2.$$

Similarly to the variance result, efficiency (as measured by the ESS) can potentially be improved by taking the ratio g_{\min}/g_{\max} as large as possible.

3.2 Cost-versus-efficiency trade-off

The penalty function g can have a significant impact on cost-aware importance sampling since it leads to a cost-versus-efficiency trade-off. For efficiency, we have already seen that we should choose g so as to maximise g_{\min}/g_{\max} . Of course, this is a simplistic view as it does not take cost into consideration. We therefore also introduce the notion of *computational gain* (CG), which is the ratio of the expected cost of the downstream task using Monte Carlo to that of using cost-aware sampling:

$$CG = \frac{\int_{\Theta} c(\theta) \pi(\theta) d\theta}{\int_{\Theta} c(\theta) \tilde{\pi}_g(\theta) d\theta}.$$

The CG allows us to ascertain the reduction in cost we can expect to have given a particular choice of g ; for instance, $CG = 2$ implies a 50% reduction in cost. The CG can be straightforwardly estimated using Monte Carlo samples from π and $\tilde{\pi}_g$ prior to running any expensive downstream tasks (and in some rare cases can even be obtained in closed-form; see Appendix B.2). We now bound the CG; see Appendix A.4 for the proof.

Proposition 4. *Suppose g is non-decreasing and $0 < g_{\min} \leq g_{\max} < \infty$. Then: $1 \leq CG \leq \frac{g_{\max}}{g_{\min}}$.*

Since we want to minimise the cost, or equivalently maximise the CG, we should maximise g_{\max}/g_{\min} , or equivalently minimise g_{\min}/g_{\max} . Unfortunately, this is in direct contradiction to our findings on maximising the ESS, which demonstrates a clear trade-off. Note that having a lower bound for the CG of 1 guarantees that we will never increase our expected cost when using cost-aware importance sampling.

3.3 Choosing the penalty function

Given the cost-versus-efficiency trade-off above, the question of how to choose g remains. Since we do not require running expensive simulations from the model to compute either the CG or the ESS, we propose to use these as a criterion. More precisely, we select g such that the quantity $CG \times ESS$ is above or close to 1 (as would be the case for standard SBI methods). This ensures that the reduction in cost does not come at the expense of significant degradation in performance.

Algorithm 1 Cost-aware SBI

Input: Number of samples n , target p , penalty function g , observed data x_o .

[Optional: Cost function c . If not available, replace c by an estimate \hat{c} in the algorithm below.]

Sample parameters $\theta_1, \dots, \theta_n$ from the cost-aware proposal \tilde{p}_g through rejection sampling using c .

Simulate from the model: $x_i \sim p(x | \theta_i)$, $i = 1, \dots, n$. Perform SBI with real data x_o , simulated data $\{(\theta_i, \{x_{ij}\}_{j=1}^m)\}_{i=1}^n$ and weights $\{w_{Ca}(\theta_i)\}_{i=1}^n$.

Output: Weighted samples from the SBI posterior.

In cases where the expensive regions of Θ are important to sample from to obtain good efficiency (for instance, in SBI when the true posterior lies in the computationally costly region), choosing a single penalty function based on CG and ESS alone may lead to sub-optimal results. Therefore, we recommend an approach based on *multiple importance sampling* (i.e. combining several importance sampling estimators), with components inspired by the defensive mixture approach of Hesterberg (1995). We consider J importance distributions, which include the target $\tilde{\pi}_1 = \pi$ and $J - 1$ cost-aware proposals $\tilde{\pi}_j(\theta) \propto \pi(\theta)/g_j(c(\theta))$ for $j = 2, \dots, J - 1$. Given $\theta_{ij} \sim \tilde{\pi}_j$ for $i = 1, \dots, n_j$ and $j = 1, \dots, J$, an integral $\mu = \int_{\Theta} f(\theta) \pi(\theta) d\theta$ can be estimated using *multiple cost-aware (mCa) importance sampling* as

$$\hat{\mu}^{mCa} = \frac{1}{J} \sum_{j=1}^J \sum_{i=1}^{n_j} w_{Ca,j}(\theta_{ij}) f(\theta_{ij})$$

where $w_{Ca,j}$ corresponds to the weights in Equation (5) computed with penalisation function g_j . In practice, we found $n_1 = \dots = n_J$ and $J = 4$ to work well and propose to use penalisation functions of the form $g_j(c(\theta)) = c(\theta)^{k_j}$. A few of the components k_j can be selected based on the $CG \times ESS$ metric, while the others can be higher powers. By doing so, we get some samples from the expensive regions, whilst also getting the cost benefits from using larger values of k_j .

Note that Csonka et al. (2001) also considered cost for multiple importance sampling, but their notion of cost refers to the cost of sampling from an importance distribution rather than the cost of a downstream task. Their method is therefore not applicable to our setting.

3.4 Cost-aware simulation-based inference

We are now ready to apply our cost-aware sampling strategy to SBI. Here, π typically corresponds to the prior or posterior, and $c(\theta)$ is the expected cost of a single simulation from the model with parameter value θ , and will be measured using runtime on a given machine. Our approach encourages sampling parameter

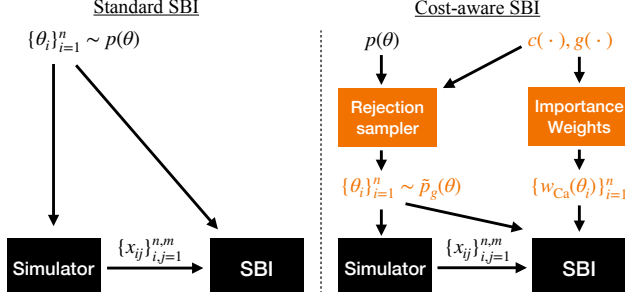


Figure 2: Flowchart of SBI and Ca-SBI. Ca-SBI utilises the cost function c and a penalty function g to (i) sample from the cost-aware proposal \tilde{p}_g , and (ii) compute the cost-aware weights. Step 1 reduces the overall cost from using the simulator, while step 2 guarantees we are sampling from the target SBI posterior.

values corresponding to cheaper parametrisations of the simulator, and hence reduces the overall cost of SBI. It is summarised in Algorithm 1 and Figure 2.

If $\theta \mapsto c(\theta)$ is not known *a priori*, we can estimate it by evaluating the simulator once for a few values of the parameters, say $(\theta_1, \dots, \theta_{\tilde{n}})$, $\tilde{n} \ll n$, and recording the computational time $\{y_i\}_{i=1}^{\tilde{n}}$. We can see these measurements as noisy evaluations of the expected cost: $y_i = c(\theta_i) + \varepsilon_i$. By fitting a simple model such as a polynomial or a Gaussian process (Rasmussen and Williams, 2006) using $\{(\theta_i, y_i)\}_{i=1}^{\tilde{n}}$, we can get an estimate $\hat{c}(\theta)$ of the cost function. Then, the cost-aware version of SBI methods can be implemented as per Algorithm 1. Note that these initial simulations are not wasted since they can be recycled for SBI.

Neural SBI. In the case of NPE/NLE, we need to estimate the losses in Equations 1 and 2 respectively. This requires approximating nested expectations with respect to the prior p , which can be expensive when p places significant mass in regions of Θ where the simulator is expensive. We therefore use a cost-aware importance sampling: given samples $(\theta_1, \dots, \theta_n) \sim \tilde{p}_g$ obtained using rejection sampling and the corresponding simulated data $x_{i1}, \dots, x_{im} \sim p(x|\theta_i)$, we can estimate the *cost-aware NPE* (Ca-NPE) and *cost-aware NLE* (Ca-NLE) losses as

$$\begin{aligned}\hat{\ell}_{\text{Ca-NPE}}(\phi) &= -\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n w_{\text{Ca}}(\theta_i) \log q_\phi(\theta_i | x_{ij}) \\ \hat{\ell}_{\text{Ca-NLE}}(\phi) &= -\frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n w_{\text{Ca}}(\theta_i) \log q_\phi(x_{ij} | \theta_i).\end{aligned}$$

Approximate Bayesian computation. The target is now the ABC approximate posterior $p_{\text{ABC}}(\theta | x_o) \propto$

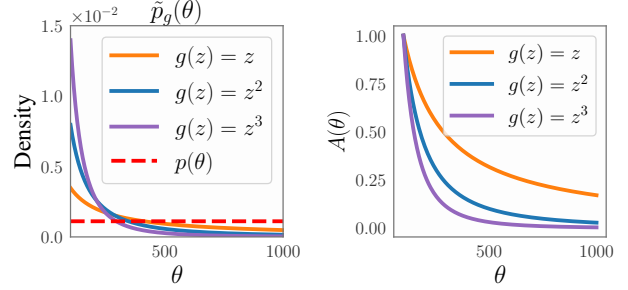


Figure 3: *Left:* Cost-aware prior $\tilde{p}_g(\theta)$ for different penalty functions $g(z) = z^k$ using prior $\mathcal{U}(10^2, 10^3)$ for the Gamma experiment. The cost increases linearly with θ , see Figure 8(a). *Right:* Acceptance probability $A(\theta)$ as a function of θ for different g functions.

$\mathbb{E}_{x \sim p(x|\theta)}[1_{\{\ell(x_o, x) \leq \epsilon\}}]p(\theta)$ and we use a proposal

$$\tilde{p}_{\text{ABC}}(\theta | x_o) \propto \mathbb{E}_{x \sim p(x|\theta)}[1_{\{\ell(x_o, x) \leq \epsilon\}}]\tilde{p}_g(\theta),$$

where $\tilde{p}_g(\theta) \propto p(\theta)/g(c(\theta))$. The *cost-aware ABC* algorithm therefore consists of sampling parameter values from \tilde{p}_g through rejection sampling, using the accept/reject mechanism of ABC, and returning all accepted samples $(\theta_1, \dots, \theta_{n_\epsilon})$ weighted by $w_{\text{Ca}}(\theta_i) = w(\theta_i)/\sum_{i=1}^{n_\epsilon} w(\theta_i)$. The target ABC posterior $p_{\text{ABC}}(\theta | x_o)$ can then be approximated as $\sum_{i=1}^{n_\epsilon} w_{\text{Ca}}(\theta_i) \delta_{\theta_i}$, where δ_{θ_i} is a Dirac delta mass at θ_i . Note that this is a consistent estimator despite the fact that the normalisation constants of $\tilde{p}_g(\theta)$, $\tilde{p}_{\text{ABC}}(\theta | x_o)$ and $p(\theta | x_o)$ are all unknown.

4 EXPERIMENTS

We will now demonstrate the performance of our cost-aware versions of ABC, NPE, and NLE against their standard counterparts. We use the `sbi` library (Tejero-Cantero et al., 2020) for NPE and NLE. We use penalties of the form $g(z) = z^k$, and present results with different choices of k . We also present results for multiple importance sampling with $J = 4$ components with equal proportions, which includes the target p . The code is available at <https://github.com/huangdaolang/cost-aware-sbi>.

4.1 Illustrative Example: Gamma simulator

We begin with an illustrative example which does not require SBI since it has a tractable likelihood, but which allows us to assess cost-aware SBI in detail. When simulating from a $\text{Gamma}(\theta, 1)$ using the popular Ahrens-Dieter acceptance-rejection method (Ahrens and Dieter, 1982), the number of uniform draws required increases with the shape parameter θ . The computational cost scales linearly with θ , and we found that

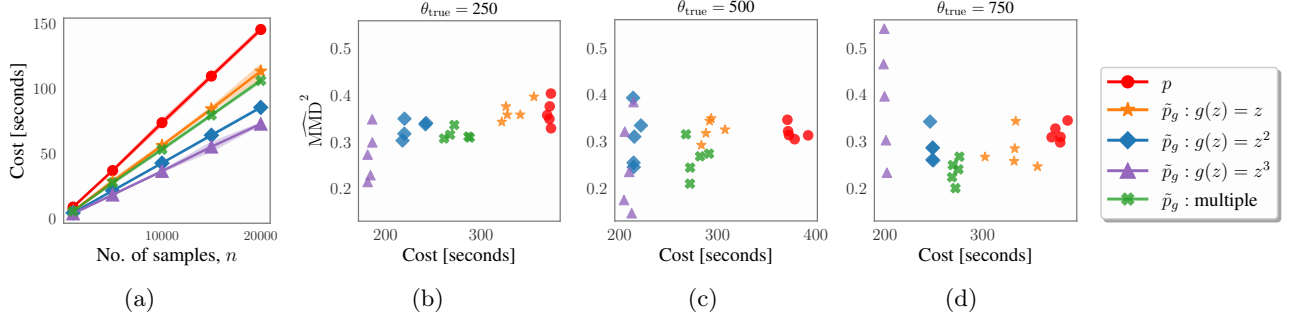


Figure 4: **The Gamma experiment.** (a) Cost of simulating n data-points, each with $m = 500$ Gamma samples, using the prior and different cost-aware proposals. (b)-(d) MMD between the ABC posteriors and the true posterior for different values of θ_{true} over five independent runs with $n = 50,000$ and $\epsilon = 0.05$. Sample mean and standard deviation of m points are taken as statistics. The corresponding NPE plots are shown in Appendix C.3.

Table 1: Cost-efficiency trade-off for Gamma($\theta, 1$).

$g(z)$		SBI	Ca-SBI			
		-	$z^{0.5}$	z	z^2	z^3
ESS	1.0	0.94	0.80	0.42	0.16	
CG	1.0	1.11	1.25	1.62	2.06	
Cost SBI/Cost Ca-SBI	1.0	1.13	1.26	1.65	2.0	
CG \times ESS	1.0	1.05	1.0	0.68	0.33	

$c(100) \approx 0.002$ seconds and $c(1000) \approx 0.02$ seconds for $m = 500$ samples on our machine. Taking p to be uniform on $\Theta = [10^2, 10^3]$, we can obtain \tilde{p}_g for $g(z) = z^k$, $k = 1, 2, 3$ in closed-form; see Figure 3 (left). As expected, the cost-aware proposals place less mass on the computationally expensive regions of Θ (i.e. larger values of θ) and the extent of the penalisation is controlled by k . Figure 3 (right) shows the acceptance probabilities $A(\theta)$ used to sample from \tilde{p}_g via rejection sampling, which naturally reduces as the cost increases. This is not a problem since the cost of simulating from p is very small (of the order of 10^{-7} s).

To choose g , we analyse the cost-efficiency trade-off by simulating $n = 20,000$ samples from the Gamma simulator in Table 1. First, we see that CG is very close to the ratio of the total cost of simulations using SBI and Ca-SBI for different g , indicating that CG is a good proxy for measuring the cost benefits. Unsurprisingly, ESS decreases as k increases whilst the cost increases when k increases. This confirms that CG \times ESS is a reasonable metric for balancing efficiency and cost when selecting g . Taking $k = 1$ yields a value of 1.0, similar to using the prior p for sampling. We therefore take the prior and cost-aware proposals with $k = \{1, 2, 3\}$ as the components for multiple importance sampling.

We plot the cost of simulating $m = 500$ data points from the Gamma simulator for different sampling distributions as a function of number of parameter samples

n in Figure 4(a). Naturally, the cost reduces as the exponent k increases, with $k = 2$ and $k = 3$ taking less than half the time compared to using the prior. In Figure 4(b)-(d), we measure the cost-versus-efficiency trade-off using the maximum mean discrepancy (MMD) (Gretton et al., 2012) between a reference posterior (obtained using the pymc3 library (Salvatier et al., 2016)) and the ABC posteriors as a function of cost for different true parameter values. The corresponding plots for NPE are shown in Appendix C.3. As the penalty on the cost increases (by increasing k), the posterior approximation improves when the true value lies in the low-cost region ($\theta_{\text{true}} = 250$) but can degrade slightly otherwise (see $\theta_{\text{true}} = 750$). The multiple importance sampling approach consistently achieves similar posterior accuracy as using the prior (i.e. classical rejection ABC), with the added benefit of around 25% reduction in total simulation cost.

4.2 Epidemiology models

We now consider variants of the SIR model in epidemiology, where SBI methods have been extensively applied (McKinley et al., 2009; Neal, 2012; McKinley et al., 2014; Kypraios et al., 2017; McKinley et al., 2018). As discussed in Section 1, the cost for these models can vary significantly depending on parameters such as infection and recovery rates. We consider three variants: the homogeneous, temporal, and Bernoulli SIR. These models have between 1 and 3 parameters impacting the cost. Detailed descriptions and their corresponding estimated cost functions are in Appendix C.4.

We compare the performance in terms of MMD between a reference NPE posterior trained on $n = 50,000$ samples and the NPE (or Ca-NPE) posteriors trained on a smaller set of $n = 5,000$ samples drawn from $p(\theta)$ or our $\tilde{p}_g(\theta)$ with different choice of g . We compute the time saved by Ca-NPE as 1 minus the ratio

Table 2: NPE and Ca-NPE on three SIR models. The mean and standard deviation from 50 runs are reported. Ca-NPE reaches comparable performance to NPE when k is small whilst taking significantly less time than NPE.

	$\widehat{\text{MMD}}^2(\downarrow)$					Time saved (\uparrow)			
	NPE	Ca-NPE $g(z) = z^{0.5}$	Ca-NPE $g(z) = z$	Ca-NPE $g(z) = z^2$	Ca-NPE multiple	Ca-NPE $g(z) = z^{0.5}$	Ca-NPE $g(z) = z$	Ca-NPE $g(z) = z^2$	Ca-NPE multiple
Homogen.	0.02(0.02)	0.02(0.01)	0.02(0.02)	0.23(0.08)	0.05(0.04)	16%(2)	38%(2)	70%(2)	30%(5)
Temporal	0.03(0.03)	0.06(0.03)	0.07(0.03)	0.07(0.03)	0.05(0.04)	36%(4)	65%(2)	85%(1)	24%(5)
Bernoulli	0.02(0.00)	0.02(0.00)	0.02(0.01)	0.04(0.01)	0.02(0.00)	23%(4)	37%(4)	47%(3)	25%(6)

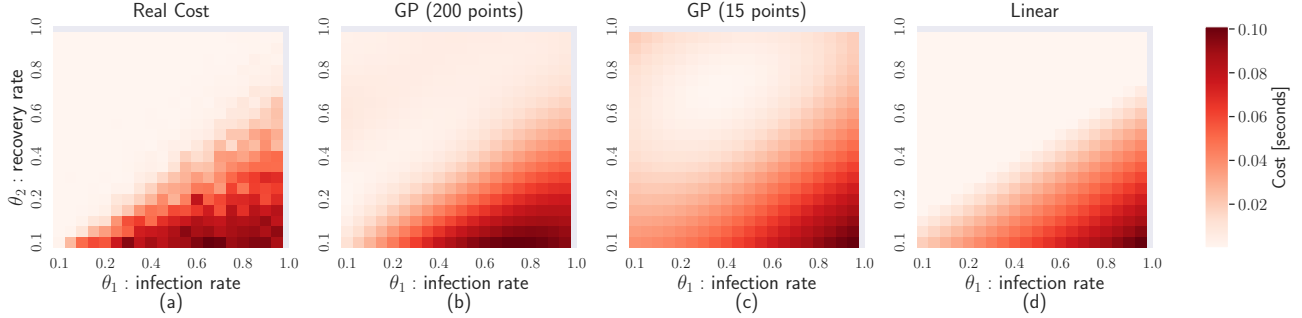


Figure 5: Cost function estimate of the temporal SIR model at varying levels of accuracy. (a) The real cost function estimated using a 20×20 grid with 50 samples for each grid (same as Figure 1). (b) Estimated cost using a GP model trained on 200 data points. (c) Estimated cost using a GP model trained on only 15 data points. (d) Estimated cost using a linear model trained on 200 data points.

of the total simulation time using \tilde{p}_g to that of using p . We report numbers for penalty function with $k = \{0.5, 1, 2\}$ and the multiple proposal in Table 2. Our Ca-NPE method reduces simulation time across different SIR models, whilst maintaining comparable accuracy to the standard NPE approach. For instance, Ca-NPE with $g(z) = z$ achieves a 37% reduction (2.4 hrs) in simulation cost without sacrificing performance in the Bernoulli SIR model. Similarly, for the temporal SIR model, $k = 2$ provides the most significant time savings of 85% (380 s), with only a slight increase in MMD. Similar results are observed when using the C2ST and marginal two-sample KS tests metrics; see Appendix C.4.

We now also study the impact of the accuracy of the cost function estimator using the temporal SIR model. Apart from using a GP to estimate the cost function with 200 data-points, we also estimate the cost using a GP trained on just 15 points, and with a linear model trained on 200 points. The corresponding cost plots are shown in Figure 5, along with the real cost function (same as Figure 1). We then use these different cost estimates to run Ca-NPE with different penalty function, and report the results in Table 3. We observe that the accuracy of the cost estimate does not impact the accuracy of the posterior estimation, as the MMD values for all the cases are similar. Moreover, even if the cost function is not accurately estimated, we are

still able to reduce the computational cost of doing SBI through a cost-aware approach; see for example the GP(15 points) and linear cost models. Hence, running our cost-aware method with rough estimates of the cost function also provides computational gain, without losing out on posterior accuracy.

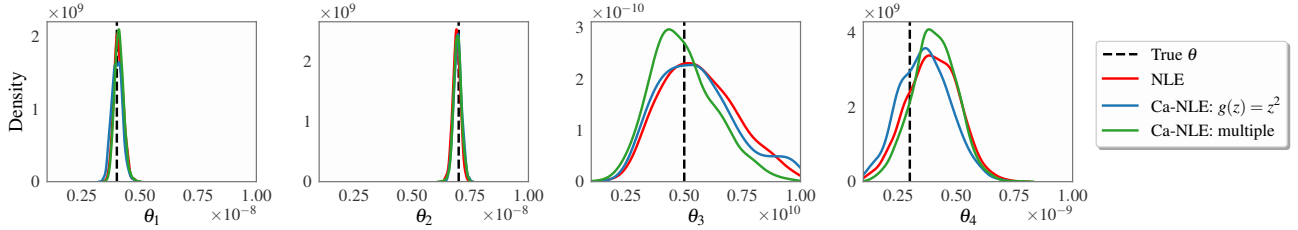
4.3 Radio propagation model

Finally, we apply our method on a computationally expensive real-world simulator with four parameters from the radio propagation field; see (Bharti et al., 2022b; Huang et al., 2023). Driven by an underlying point process, this model simulates an 801-dimensional complex-valued time series data which represents a radio signal. The cost of simulation increases linearly with the arrival rate of the point process, which is one of the four parameters. The other three parameters, governing the amplitude and noise present in the signal, do not affect the cost. For each parameter θ , we simulate $m = 50$ iid time series and use six summary statistics based on temporal moments (Bharti et al., 2022a) of the signals, as described in Appendix C.5.

We generate $n = 10,000$ realisations using a uniform prior, which took 15.6 hrs. In contrast, simulating the same number of samples from our cost-aware proposal took 8.8 hrs with $g(z) = z^2$ and 11.6 hrs with multiple importance sampling ($k = \{1, 2, 3\}$), leading to a

Table 3: Effect of inaccurate cost function estimation on Ca-SBI. MMD (\downarrow) and time saved (\uparrow) for Ca-NPE on the temporal SIR model with different cost models. The mean and standard deviation from 50 runs are reported.

Cost model	Ca-NPE ($g(z) = z^{0.5}$)		Ca-NPE ($g(z) = z$)		Ca-NPE ($g(z) = z^2$)		Ca-NPE (multiple)	
	$\widehat{\text{MMD}}^2$	Time saved	$\widehat{\text{MMD}}^2$	Time saved	$\widehat{\text{MMD}}^2$	Time saved	$\widehat{\text{MMD}}^2$	Time saved
GP (200 points)	0.06(0.03)	36%(4)	0.07(0.03)	65%(2)	0.07(0.03)	85%(1)	0.05(0.04)	24%(5)
GP (15 points)	0.06(0.03)	8%(3)	0.06(0.03)	33%(2)	0.08(0.03)	69%(1)	0.05(0.04)	15%(3)
Linear	0.07(0.02)	19%(5)	0.06(0.02)	34%(6)	0.06(0.02)	61%(2)	0.07(0.02)	27%(3)


 Figure 6: Marginals of the approximate posterior for the radio propagation model using NLE and our Ca-NLE method with $n = 10,000$. The Ca-NLE methods perform similar to NLE, whilst saving hours of computation.

cost reduction of 44% and 26%, respectively. All three posterior distributions are very similar, with multiple importance sampling yielding the most concentrated posterior around the true value for the arrival rate θ_3 . Therefore, using our approach, we save hours of computational time without any degradation in performance. We plot the corresponding NLE posteriors in Figure 6 and observe similar results.

Interestingly, Ca-SBI is embarrassingly parallelisable and can therefore lead to further computational gains if the right resources for parallelisation are available to the user. To demonstrate this, we generated $n = 10,000$ realisations again from the radio propagation model by parallelising computation with 50 cores for both standard SBI and Ca-SBI with $g(z) = z^2$. As expected, the total run time reduced by a factor of approximately 1/50, with standard SBI taking 21 mins (instead of 15.6 hrs) and Ca-SBI taking 10.6 mins (instead of 8.8 hrs), resulting in a computational speed-up of around 49.5%, similar to what we observed before parallelisation.

5 CONCLUSION

We proposed a family of cost-aware importance sampling distributions which can reduce the overall cost when downstream tasks are expensive and have a cost which varies over the sampling space. For SBI methods, this is particularly helpful when the cost of simulating from the model varies across parameters. As our method only depends on the prior, we could show that it can be applied to the most popular SBI methods including ABC, NPE, and NLE. Other methods such as regression-adjustment ABC (Beaumont et al., 2002),

generalised Bayesian inference (Pacchiardi et al., 2024), Bayesian synthetic likelihood (Price et al., 2018; Frazier et al., 2023), and neural ratio estimation (Durkan et al., 2020; Hermans et al., 2020) could also be made cost-aware in a similar manner.

Limitations and future work. One interesting extension of the work could be to use cost-aware sampling beyond SBI. As importance sampling suffers in high dimensions, our method may not work as well if there are many parameters that affect the cost function. The performance also degrades when the true value lies in the high-cost region. However, these issues could be alleviated by using adaptive importance sampling (Cornuet et al., 2012; Martino et al., 2015) where the penalty function is adapted based on an initial coarse estimate of where the posterior lies. For SBI, a limitation of our approach is that it does not apply to optimisation-based SBI methods such as minimum distance estimation (Briol et al., 2019; Key et al., 2021; Dellaporta et al., 2022), the method of simulated moments (Hall, 2003), or regression-based methods (Gutmann and Corander, 2016). Developing cost-aware versions of these methods is an interesting avenue for future work, and could be approached with cost-aware Bayesian optimisation as proposed in Lee et al. (2020).

Acknowledgements

The authors are grateful to Art Owen and Dennis Prangle for pointing out relevant related work. AB, DH and SK were supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI). AB was also supported

by the Research Council of Finland grant no. 362534. SK was also supported by the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1]. FXB was supported by the EPSRC grant [EP/Y022300/1].

References

- Ahrens, J. H. and Dieter, U. (1982). Generating Gamma variates by a modified rejection technique. *Communications of the ACM*, 25(1):47–54. 6
- Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990. 2, 3
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035. 3, 9
- Behrens, J. and Dias, F. (2015). New computational methods in tsunami science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2053):20140382. 1
- Bharti, A., Briol, F.-X., and Pedersen, T. (2022a). A general method for calibrating stochastic radio channel models with kernels. *IEEE Transactions on Antennas and Propagation*, 70(6):3986–4001. 1, 3, 8
- Bharti, A., Filstroff, L., and Kaski, S. (2022b). Approximate Bayesian computation with domain expert in the loop. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1893–1905. 8
- Bharti, A., Naslidnyk, M., Key, O., Kaski, S., and Briol, F.-X. (2023). Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2289–2312. 3
- Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019). Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*. 1, 9
- Cobb, A. D., Roy, A., Elenius, D., Heim, F. M., Swenson, B., Whittington, S., Walker, J. D., Bapty, T., Hite, J., Ramani, K., McComb, C., and Jha, S. (2023). Aircraftverse: A large-scale multimodal dataset of aerial vehicle designs. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 1
- Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812. 9
- Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062. 1
- Csonka, F., Szirmay-Kalos, L., and Antal, G. (2001). Cost-driven multiple importance sampling for monte-carlo rendering. Technical report, TR-186-2-01-19, Institute of Computer Graphics and Algorithms, Vienna. 5
- Dax, M., Green, S. R., Gair, J., Pürner, M., Wildberger, J., Macke, J. H., Buonanno, A., and Schölkopf, B. (2023). Neural importance sampling for rapid and reliable gravitational-wave inference. *Physical Review Letters*, 130(17):171403. 3
- Del Moral, P., Doucet, A., and Jasra, A. (2011). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020. 2, 3
- Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 943–970. 9
- Durkan, C., Murray, I., and Papamakarios, G. (2020). On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, volume 119, pages 2771–2781. 9
- Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607. 3
- Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. (2023). Bayesian inference using synthetic likelihood: Asymptotics and adjustments. *Journal of the American Statistical Association*, 118(544):2821–2832. 9
- Friedman, J. H. (2003). On multivariate goodness-of-fit and two-sample testing. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, 1:311. 22
- Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*. 20
- Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv:1707.07269*. 17
- Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4):455. 4
- Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337. 4
- Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). Automatic posterior transformation for

- likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2404–2414. 3, 17
- Gretton, A., Borgwardt, K., Rasch, M. J., and Scholkopf, B. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773. 7
- Gutmann, M. U. and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47. 3, 9
- Hall, A. R. (2003). Generalized method of moments. *A companion to theoretical econometrics*, pages 230–255. 9
- Hermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4239–4248. 3, 9
- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194. 5
- Hesterberg, T. C. (1988). *Advances in importance sampling*. Stanford University. 4
- Hodges Jr, J. (1958). The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486. 22
- Hoppe, M., Embreus, O., and Fülöp, T. (2021). Dream: A fluid-kinetic framework for tokamak disruption runaway electron simulations. *Computer Physics Communications*, 268:108098. 1
- Huang, D., Bharti, A., Souza, A. H., Acerbi, L., and Kaski, S. (2023). Learning robust statistics for simulation-based inference under model misspecification. In *Thirty-seventh Conference on Neural Information Processing Systems*. 8, 23
- Järvenpää, M., Gutmann, M. U., Pleska, A., Vehtari, A., and Marttinen, P. (2019). Efficient acquisition rules for model-based approximate Bayesian computation. *Bayesian Analysis*, 14(2):595 – 622. 3
- Key, O., Fernandez, T., Gretton, A., and Briol, F.-X. (2021). Composite goodness-of-fit tests with kernels. In *NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling*. 9
- Kirby, A., Briol, F.-X., Dunstan, T. D., and Nishino, T. (2023). Data-driven modelling of turbine wake interactions and flow resistance in large wind farms. *Wind Energy*, 26(9):875–1011. 1
- Kulkarni, S. and Moritz, C. A. (2023). Improving effectiveness of simulation-based inference in the massively parallel regime. *IEEE Transactions on Parallel and Distributed Systems*, 34(4):1100–1114. 3
- Kypriaios, T., Neal, P., and Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Mathematical Biosciences*, 287:42–53. 1, 7
- Lee, E. H., Perrone, V., Archambeau, C., and Seeger, M. (2020). Cost-aware Bayesian optimization. *arXiv preprint arXiv:2003.10870*. 9
- Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66:66–82. 1, 2
- Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*. 22
- Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). Benchmarking simulation-based inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 343–351. 1
- Martino, L., Elvira, V., Luengo, D., and Corander, J. (2015). An adaptive population importance sampler: Learning from uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437. 9
- McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1). 7
- McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. (2014). Simulation-based Bayesian inference for epidemic models. *Computational Statistics & Data Analysis*, 71:434–447. 7
- McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2018). Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, 33(1):4–18. 1, 7
- Meeds, E. and Welling, M. (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, page 593–602. 3
- Neal, P. (2012). Efficient likelihood-free Bayesian computation for household epidemics. *Statistics and Computing*, 22:1239–1256. 7
- Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767. 17
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–625. 3

- Niu, Z., Meier, J., and Briol, F.-X. (2023). Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. *Electronic Journal of Statistics*, 17(1):1411–1456. 3
- Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. <https://artowen.su.domains/mc/>. 4, 14
- Pacchiardi, L., Khoo, S., and Dutta, R. (2024). Generalized Bayesian likelihood-free inference. *Electronic Journal of Statistics*, 18(2):3628–3686. 9
- Papamakarios, G. and Murray, I. (2016). Fast ϵ -free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1036–1044. 2
- Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64. 2
- Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30. 17
- Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 837–848. 2, 3, 17
- Prangle, D. (2016). Lazy ABC. *Statistics and Computing*, 26:171–185. 3
- Prangle, D. and Viscardi, C. (2023). Distilling importance sampling for likelihood free inference. *Journal of Computational and Graphical Statistics*, 32(4):1461–1471. 3
- Prescott, T. P. and Baker, R. E. (2020). Multifidelity approximate Bayesian computation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):114–138. 3
- Prescott, T. P. and Baker, R. E. (2021). Multifidelity approximate Bayesian computation with sequential Monte Carlo parameter sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):788–817. 3
- Prescott, T. P., Warne, D. J., and Baker, R. E. (2024). Efficient multifidelity likelihood-free Bayesian inference with adaptive computational resource allocation. *Journal of Computational Physics*, 496:112577. 3
- Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11. 9
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798. 2
- Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2022). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466. 2
- Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press. 6
- Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55. 7
- Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, pages 673–684. 3
- Sisson, S. A. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC. 1, 2
- Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765. 3
- Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. (2020). sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505. 6, 17
- Trotter, H. and Tukey, J. (1956). Conditional Monte Carlo for normal samples. In *Proceedings of the Symposium on Monte Carlo Methods*, pages 64–79. John Wiley and Sons. 4
- von Neumann, J. (1951). Various technique used in connection with random digits. *National Bureau of Standards Applied Math Series*, 12:36–38. 3
- Warne, D. J., Prescott, T. P., Baker, R. E., and Simpson, M. J. (2022). Multifidelity multilevel Monte Carlo to accelerate approximate Bayesian parameter inference for partially observed stochastic processes. *Journal of Computational Physics*, 469:111543. 3
- Wilkinson, R. (2014). Accelerating ABC methods using Gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 1015–1023. 3
- Zammit-Mangion, A., Sainsbury-Dale, M., and Huser, R. (2024). Neural methods for amortised parameter inference. *arxiv 2404.12484*. 1

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Section 2 and Section 3]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, see Section 4]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, see zipped file attached.]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes, see Section 3]
 - (b) Complete proofs of all theoretical results. [Yes, see Appendix A]
 - (c) Clear explanations of any assumptions. [Yes, see Section 3]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. Instructions are in Appendix C.]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, see Section 4 and Appendix C.1]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, see Section 4]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, see Appendix C.1]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes, see Appendix C.1]
 - (b) The license information of the assets, if applicable. [Yes, see Appendix C.1]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

Appendix A contains the proofs of the theoretical results presented in the main text. In Appendix B, we derive closed-form expressions for the cost-aware proposal and the computational gain (CG) in certain cases. Appendix C consists of the implementation details and additional results for the experiments conducted in Section 4.

A Proof of theoretical results

A.1 Proof of Proposition 1

Proof. Consider the cost-aware proposal $\tilde{\pi}_g(\theta) \propto \pi(\theta)/g(c(\theta))$. In order to ensure that $\tilde{\pi}_g$ is indeed a probability density function, we need to show that (i) it is non-negative, and (ii) can be normalised; i.e. there exists a normalisation constant $B > 0$ which is finite. For (i), we know that π is a probability density function, and that $g_{\min} > 0$, therefore $\tilde{\pi}_g$ is a ratio of non-negative quantities and must be non-negative. For (ii), we can once again use the existence of $g_{\min} > 0$ to guarantee the normalisation constant is finite since:

$$B = \int_{\Theta} \frac{\pi(\theta)}{g(c(\theta))} d\theta \leq \frac{\int_{\Theta} \pi(\theta) d\theta}{\inf_{\theta \in \Theta} g(c(\theta))} = \frac{1}{g_{\min}}. \quad (6)$$

Here, the inequality comes from the fact that $g \circ c$ is strictly positive, so we must have that $B < 1/g_{\min} < \infty$. This concludes our proof that $\tilde{\pi}_g$ is a valid probability density function.

To sample from proposal $\tilde{\pi}_g$ using rejection sampling, we need to find a constant $M > 0$ such that the condition $\tilde{\pi}_g(\theta) \leq M\pi(\theta)$ holds. To that end,

$$\frac{\tilde{\pi}_g(\theta)}{\pi(\theta)} = \frac{\pi(\theta)}{Bg(c(\theta))\pi(\theta)} = \frac{1}{Bg(c(\theta))} \leq \frac{1}{Bg_{\min}} = M,$$

where the first equality follows from the definition of $\tilde{\pi}_g$, the second equality by cancellation of π in the numerator and denominator, and the inequality by the fact that g_{\min} is the infimum of $g \circ c$. Note that $g \circ c$ is strictly positive, and therefore $M < \infty$. The acceptance probability $A(\theta)$ for the rejection sampler can hence be written as

$$A(\theta) = \frac{\tilde{\pi}_g(\theta)}{M\pi(\theta)} = \frac{\pi(\theta)Bg_{\min}}{Bg(c(\theta))\pi(\theta)} = \frac{g_{\min}}{g(c(\theta))},$$

which completes the proof. \square

A.2 Proof of Proposition 2

Proof. We start by proving almost-sure convergence. According to Theorem 9.2 of Owen (2013, Ch. 9), the self-normalised importance sampling estimator $\hat{\mu}_n^{\text{Ca}}$ converges to μ in probability as n goes to infinity if the proposal $\tilde{\pi}_g$ is a probability density function on $\Theta \subseteq \mathbb{R}^p$ and $\tilde{\pi}_g(\theta) > 0$ whenever the target $\pi(\theta) > 0$. From Proposition 1, we know that $\tilde{\pi}_g(\theta) = \pi(\theta)/Bg(c(\theta))$ is a valid density with $0 < B < \infty$, and that $\inf_{\theta \in \Theta} g(c(\theta)) = g_{\min} > 0$. As a result, the denominator $Bg(c(\theta))$ is greater than zero for all $\theta \in \Theta$, and hence $\pi(\theta) > 0$ whenever $\tilde{\pi}_g(\theta) > 0$, which completes the proof of the first statement.

We now move on to proving the bounds on the weights. Recall that

$$w_{\text{Ca}}(\theta_i) = \frac{w(\theta_i)}{\sum_{j=1}^n w(\theta_j)} = \frac{g(c(\theta_i))}{\sum_{j=1}^n g(c(\theta_j))} \leq \frac{\sup_{\theta \in \Theta} g(c(\theta))}{n \inf_{\theta \in \Theta} g(c(\theta))} = \frac{g_{\max}}{ng_{\min}}.$$

Similarly, we also have a lower bound by taking an infimum on the numerator and supremum on the denominator:

$$w_{\text{Ca}}(\theta_i) \geq \frac{g_{\min}}{ng_{\max}}.$$

Finally, we prove the bounds on the variance σ_{Ca}^2 . Since we are considering a self-normalised importance sampling estimator, the variance of $\hat{\mu}_n^{\text{Ca}}$ can be written as (Owen, 2013, Ch. 9):

$$n\sigma_{\text{Ca}}^2 = \int_{\Theta} \frac{\pi(\theta)^2(f(\theta) - \mu)^2}{\tilde{\pi}_g(\theta)} d\theta = \int_{\Theta} \frac{Bg(c(\theta))\pi(\theta)^2(f(\theta) - \mu)^2}{\pi(\theta)} d\theta = B \int_{\Theta} g(c(\theta))\pi(\theta)(f(\theta) - \mu)^2 d\theta \quad (7)$$

Upper bounding this expression, we get:

$$\begin{aligned} n\sigma_{\text{Ca}}^2 &\leq Bg_{\max} \int_{\Theta} \pi(\theta)(f(\theta) - \mu)^2 d\theta \leq \frac{g_{\max}}{g_{\min}} \int_{\Theta} \pi(\theta)(f(\theta) - \mu)^2 d\theta \\ &= \frac{g_{\max}}{g_{\min}} \int_{\Theta} \pi(\theta)(f^2(\theta) - 2f(\theta)\mu + \mu^2) d\theta = \frac{g_{\max}}{g_{\min}} (n\sigma_{\text{MC}}^2 - \mu^2). \end{aligned}$$

To derive the above, we first used the definition of our cost-aware proposal, then used Hölder's inequality to take out the supremum of g and the bound in Equation (6) to upper bound B . We then concluded by expanding the square and simplifying using the definitions of μ and σ_{MC}^2 . Dividing by n on both sides of the inequality gives our upper bound.

Similarly, we can get the lower bound by using a lower bound on B (rather than an upper bound). Such a bound can be obtained as:

$$B = \int_{\Theta} \frac{\pi(\theta)}{g(c(\theta))} d\theta \geq \frac{\int_{\Theta} \pi(\theta) d\theta}{\sup_{\theta \in \Theta} g(c(\theta))} = \frac{1}{g_{\max}}. \quad (8)$$

Using Equation (8) to lower bound Equation (7) and taking an infimum over g , we get

$$n\sigma_{\text{Ca}}^2 \geq \frac{g_{\min}}{g_{\max}} \int_{\Theta} \pi(\theta)(f(\theta) - \mu)^2 d\theta = \frac{g_{\min}}{g_{\max}} (n\sigma_{\text{MC}}^2 - \mu^2).$$

Once again, dividing both sides by n gives the required bound. This concludes our proof. \square

A.3 Proof of Proposition 3

Proof. We start with the upper bound on the ESS, and take a similar approach to that used to bound the weights in Proposition 2:

$$\text{ESS} = \frac{(\sum_{i=1}^n g(c(\theta_i)))^2}{n \sum_{i=1}^n g(c(\theta_i))^2} \leq \frac{\sup_{\theta \in \Theta} (\sum_{i=1}^n g(c(\theta)))^2}{\inf_{\theta \in \Theta} n \sum_{i=1}^n g(c(\theta))^2} = \frac{n^2 g_{\max}^2}{n^2 g_{\min}^2} = \left(\frac{g_{\max}}{g_{\min}} \right)^2.$$

Similarly (but taking an infimum on the numerator and supremum on the denominator), we get a corresponding lower bound

$$\text{ESS} = \frac{(\sum_{i=1}^n g(c(\theta_i)))^2}{n \sum_{i=1}^n g(c(\theta_i))^2} \geq \frac{\inf_{\theta \in \Theta} (\sum_{i=1}^n g(c(\theta)))^2}{\sup_{\theta \in \Theta} n \sum_{i=1}^n g(c(\theta))^2} = \left(\frac{g_{\min}}{g_{\max}} \right)^2,$$

which concludes our proof. \square

A.4 Proof of Proposition 4

Proof. The fact that the CG is lower-bounded by 1 is trivial and simply follows from the fact that $\tilde{\pi}_g$ is explicitly putting less mass than π in regions of higher cost.

For the upper bound, we start by proving an explicit expression for the CG in terms of B and $g(c(\theta))$:

$$\text{CG} = \frac{\int_{\Theta} c(\theta) \pi(\theta) d\theta}{\int_{\Theta} c(\theta) \tilde{\pi}_g(\theta) d\theta} = \frac{\int_{\Theta} c(\theta) \pi(\theta) d\theta}{\int_{\Theta} c(\theta) \frac{\pi(\theta)}{Bg(c(\theta))} d\theta} = B \frac{\int_{\Theta} c(\theta) \pi(\theta) d\theta}{\int_{\Theta} c(\theta) \frac{\pi(\theta)}{g(c(\theta))} d\theta}$$

Using the above, we get

$$\text{CG} \leq \frac{B}{\inf_{\theta \in \Theta} \frac{1}{g(c(\theta))}} \left(\frac{\int_{\Theta} c(\theta) \pi(\theta) d\theta}{\int_{\Theta} c(\theta) \pi(\theta) d\theta} \right) = Bg_{\max} \leq \frac{g_{\max}}{g_{\min}}.$$

where the last inequality follows from the fact that $B \leq \frac{1}{g_{\min}}$, as established in Equation (6). \square

B Closed-form expressions of the cost-aware proposal and related quantities

B.1 Rejection sampling with closed-form proposal

In this section, we derive closed-form expressions for the cost-aware proposal in order to plot its density for different penalty functions in Figure 3. Instead of sampling from the prior $p(\theta)$ (taken to be a $\mathcal{U}(\theta_{\min}, \theta_{\max})$ with $0 < \theta_{\min} < \theta_{\max} < \infty$), we propose to generate parameter values from the cost-aware proposal distribution $\tilde{p}_g = \frac{p(\theta)}{g(c(\theta)) \times B}$: where $B = \int_{\Theta} \frac{p(\theta)}{g(c(\theta))} d\theta < \infty$ is the normalisation constant that ensures \tilde{p}_g is a density. Although our method does not require B to be known, it can be computed in closed-form for certain choices of the cost function $c(\theta)$, penalty function g , and the prior p as given in Table 4.

Table 4: Closed-form expressions for the normalisation constant B of cost-aware densities.

$c(\theta)$	$g(z)$	B
$\alpha\theta + \beta$	$z^k, k > 1$	$\frac{1}{(\theta_{\max} - \theta_{\min})} \left[\frac{(\alpha\theta + \beta)^{1-k}}{\alpha(1-k)} \right]_{\theta_{\min}}^{\theta_{\max}}$
$\alpha\theta + \beta$	z	$\frac{1}{(\theta_{\max} - \theta_{\min})} \left[\frac{\log(\alpha\theta + \beta)}{\alpha} \right]_{\theta_{\min}}^{\theta_{\max}}$
$\alpha\theta^2$	1	$\frac{1}{(\theta_{\max} - \theta_{\min})} \left[-\frac{1}{\alpha^2\theta} \right]_{\theta_{\min}}^{\theta_{\max}}$

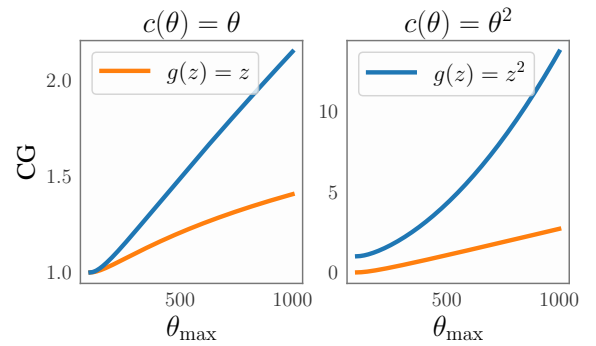
B.2 Closed-form expressions for computational gain (CG)

In case of a uniform prior distribution $\mathcal{U}(\theta_{\min}, \theta_{\max})$, the computational gain (CG) can be computed in closed-form for linear and quadratic choices of cost and penalty function. We report the CG expressions in Table 5, and plot the CG as a function of θ_{\max} for linear and quadratic cost functions in Figure 7. Naturally, the gains are more significant when the cost is quadratic and $g(z) = z^2$.

Table 5: Closed-form expressions of computational gain for uniform prior $\mathcal{U}(\theta_{\min}, \theta_{\max})$.

$c(\theta)$	$g(z)$	CG
$\alpha\theta$	z	$\frac{(\theta_{\max} + \theta_{\min}) \log(\frac{\theta_{\max}}{\theta_{\min}})}{2(\theta_{\max} - \theta_{\min})}$
$\alpha\theta^2$	z	$\frac{(\theta_{\max}^3 - \theta_{\min}^3)}{3(\theta_{\max} - \theta_{\min})\theta_{\min}\theta_{\max}}$
$\alpha\theta$	z^2	$\frac{(\theta_{\max}^2 - \theta_{\min}^2)}{2\theta_{\max}\theta_{\min} \log(\theta_{\max}/\theta_{\min})}$
$\alpha\theta^2$	z^2	$\frac{(\theta_{\max}^3 - \theta_{\min}^3)^2}{9(\theta_{\max} - \theta_{\min})^2 \theta_{\max}^2 \theta_{\min}^2}$

Figure 7: CG vs. θ_{\max} for different choices of penalty function g under uniform prior $\mathcal{U}(1, \theta_{\max})$.



We now present the derivation of the closed-form expressions of CG from Table 5:

Case 1: $c(\theta) = \alpha\theta$, $g(z) = z$. We start by computing the expected cost of the cost-aware method:

$$\begin{aligned} \int_{\Theta} c(\theta) \tilde{p}_g(\theta) d\theta &= \int_{\Theta} \frac{c(\theta)p(\theta)}{Bg(c(\theta))} d\theta = \frac{1}{B} = \left(\int_{\Theta} \frac{p(\theta)}{g(c(\theta))} d\theta \right)^{-1} = \left(\int_{\Theta} \frac{p(\theta)}{c(\theta)} d\theta \right)^{-1} \\ \text{CG} &= \left(\int_{\Theta} c(\theta)p(\theta) d\theta \right) \left(\int_{\Theta} \frac{p(\theta)}{c(\theta)} d\theta \right) = \left(\frac{\alpha}{(\theta_{\max} - \theta_{\min})} \int_{\Theta} \theta d\theta \right) \left(\frac{1}{\alpha(\theta_{\max} - \theta_{\min})} \int_{\Theta} \frac{1}{\theta} d\theta \right) \\ &= \frac{(\theta_{\max} + \theta_{\min}) \log(\frac{\theta_{\max}}{\theta_{\min}})}{2(\theta_{\max} - \theta_{\min})} \end{aligned}$$

Case 2: $c(\theta) = \alpha\theta^2$, $g(z) = z$

$$\begin{aligned} \text{CG} &= \left(\frac{\alpha}{(\theta_{\max} - \theta_{\min})} \int_{\Theta} \theta^2 d\theta \right) \left(\frac{1}{\alpha(\theta_{\max} - \theta_{\min})} \int_{\Theta} \frac{1}{\theta^2} d\theta \right) \\ &= \left(\frac{1}{(\theta_{\max} - \theta_{\min})^2} \left[\frac{\theta^3}{3} \right]_{\theta_{\min}}^{\theta_{\max}} \right) \left[-\frac{1}{\theta} \right]_{\theta_{\min}}^{\theta_{\max}} = \frac{(\theta_{\max}^3 - \theta_{\min}^3)}{3(\theta_{\max} - \theta_{\min})\theta_{\min}\theta_{\max}} \end{aligned}$$

Case 3: $c(\theta) = \alpha\theta$, $g(z) = z^2$

$$\begin{aligned} \int_{\Theta} c(\theta) \tilde{p}_g(\theta) d\theta &= \frac{1}{B} \int_{\Theta} \frac{c(\theta)p(\theta)}{c(\theta)^2} d\theta = \left(\int_{\Theta} \frac{p(\theta)}{c(\theta)} d\theta \right) \left(\int_{\Theta} \frac{p(\theta)}{c(\theta)^2} d\theta \right)^{-1} \\ \text{CG} &= \left(\frac{\int_{\Theta} c(\theta)p(\theta) d\theta}{\int_{\Theta} \frac{p(\theta)}{c(\theta)} d\theta} \right) \left(\int_{\Theta} \frac{p(\theta)}{c(\theta)^2} d\theta \right) \\ &= \left(\frac{\alpha \left[\theta^2 \right]_{\theta_{\min}}^{\theta_{\max}}}{2(\theta_{\max} - \theta_{\min})} \right) \left(\frac{1}{\alpha^2(\theta_{\max} - \theta_{\min})} \left[-\frac{1}{\theta} \right]_{\theta_{\min}}^{\theta_{\max}} \right) \left(\frac{1}{\alpha(\theta_{\max} - \theta_{\min})} \left[\log \theta \right]_{\theta_{\min}}^{\theta_{\max}} \right) \\ &= \frac{(\theta_{\max} + \theta_{\min})}{2(\log \theta_{\max} - \log \theta_{\min})} \left(\frac{1}{\theta_{\min}} - \frac{1}{\theta_{\max}} \right) = \frac{(\theta_{\max}^2 - \theta_{\min}^2)}{2\theta_{\max}\theta_{\min} \log(\theta_{\max}/\theta_{\min})} \end{aligned}$$

Case 4: $c(\theta) = \alpha\theta^2$, $g(z) = z^2$

$$\begin{aligned} \text{CG} &= \left(\frac{\alpha}{(\theta_{\max} - \theta_{\min})} \int_{\Theta} \theta^2 d\theta \right) \left(\frac{1}{\alpha^2(\theta_{\max} - \theta_{\min})} \int_{\Theta} \frac{1}{\theta^4} d\theta \right) \left(\frac{1}{\alpha(\theta_{\max} - \theta_{\min})} \int_{\Theta} \frac{1}{\theta^2} d\theta \right) \\ &= \left(\frac{\left[\theta^3 \right]_{\theta_{\min}}^{\theta_{\max}}}{3(\theta_{\max} - \theta_{\min})} \right) \left(-\frac{1}{3} \left[\frac{1}{\theta^3} \right]_{\theta_{\min}}^{\theta_{\max}} \right) \left(\frac{1}{\left[-\frac{1}{\theta} \right]_{\theta_{\min}}^{\theta_{\max}}} \right) = \frac{(\theta_{\max}^3 - \theta_{\min}^3)^2}{9(\theta_{\max} - \theta_{\min})^2 \theta_{\max}^2 \theta_{\min}^2} \end{aligned}$$

C Additional experimental details and results

C.1 Implementation details

We used the `sbi` package (Tejero-Cantero et al., 2020) (<https://sbi-dev.github.io/sbi/>, Version: 0.22.0, License: Apache 2.0) to implement NPE and Ca-NPE. Specifically, we chose the NPE-C model (Greenberg et al., 2019) with Masked Autoregressive Flow (MAF) (Papamakarios et al., 2017) as the inference network. We used the default configuration with 50 hidden units and 5 transforms for MAF, and training with a fixed learning rate 5×10^{-4} . We implemented NLE using the NLE-A model (Papamakarios et al., 2019) with the same network configuration as in NPE. For posterior estimation, we used MCMC and chose slice sampling (Neal, 2003) as the MCMC method. For the ABC implementation in the Gamma experiment, we use rejection-ABC algorithm with a tolerance threshold of $\epsilon = 0.05$ and $n = 50,000$, in contrast with $n = 5000$ for NPE, as rejection-ABC requires much more samples to achieve a reasonable level of posterior approximation. We compute the Euclidean distance between the sample mean and standard deviation of the observed and simulated data, computed from $m = 500$ iid samples. We used a Dell Precision 7550 laptop with Intel i7-10750H processor for running the Gamma experiments and generating the data for the radio propagation model. All the epidemiology experiments are carried out on a CPU cluster consisting of Xeon E5 2680 processors.

C.2 Maximum mean discrepancy hyperparameters

We use a Gaussian kernel for implementing MMD, and select the lengthscale parameter using the median heuristic (Garreau et al., 2017) computed on the observed dataset (i.e. we take it proportional to the median distance between observations). This is one of the most widely used parameter selection methods for MMD, and we provide the hyperparameter values in Table 6 and Table 7. The amplitude parameter of the kernel is set to 1 as it just acts a multiplicative constant and therefore does not matter so long as it is selected to be the same for all methods.

Table 6: Lengthscales used for the Gamma simulator.

θ_{true}	Lengthscale
250	0.48
500	0.68
750	0.81

Table 7: Lengthscales used the SIR simulators.

SIR model	Lengthscale
Homogeneous	0.10
Temporal	0.19
Bernoulli	0.32

 Table 8: Values of g_{\min} and g_{\max} for the Gamma simulator.

	g_{\min}	g_{\max}
$g(z) = z$	10^{-3}	2.8×10^{-3}
$g(z) = z^2$	10^{-6}	7.8×10^{-6}
$g(z) = z^3$	10^{-9}	2.2×10^{-8}

C.3 Additional results for the Gamma experiment

Here we present the additional results related to the Gamma experiment. We estimate the cost of sampling $m = 500$ iid points from the Gamma simulator, averaged over 50 runs, using a linear model in Figure 8(a). The parameter θ is varied in the range $[100, 1000]$. The g_{\min} and g_{\max} values for different choices of g is in Table 8.

In Figure 8(b)-(d), we plot the MMD between the NPE posteriors and the true posterior for the Gamma simulator, analogous to the ABC plots shown in Figure 4. Note that we use $n = 5000$ points in the training data for NPE, unlike the $n = 50,000$ data points we simulate for ABC, as NPE achieves superior performance than ABC with much fewer samples. Similar to the ABC results, the multiple importance sampling approach achieves similar posterior accuracy as using the prior, whilst reducing total simulation cost by roughly 25%. When the true value lies in the low-cost region ($\theta_{\text{true}} = 250$), the NPE posterior approximation improves with increasing penalty exponent k , as most of the simulated samples in the training data are from that region. Contrarily, the MMD for NPE increases with k when θ_{true} is in the high-cost regions, as the NPE is trained with much fewer samples around θ_{true} . The approximate posterior distributions obtained using NPE and ABC are shown in Figure 9 and Figure 10, respectively.

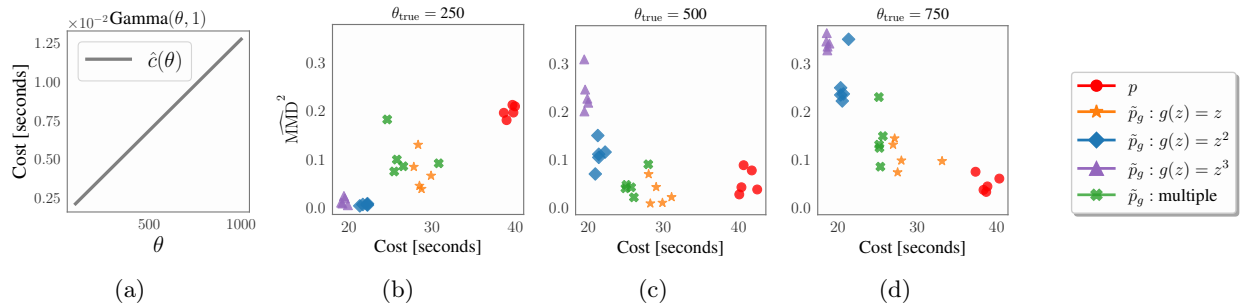


Figure 8: (a) Estimated cost of the Gamma simulator. (b)-(d) MMD between the NPE posteriors and the true posterior for different values of θ_{true} over five independent runs with $n = 5000$. Sample mean and standard deviation of m points are taken as data.

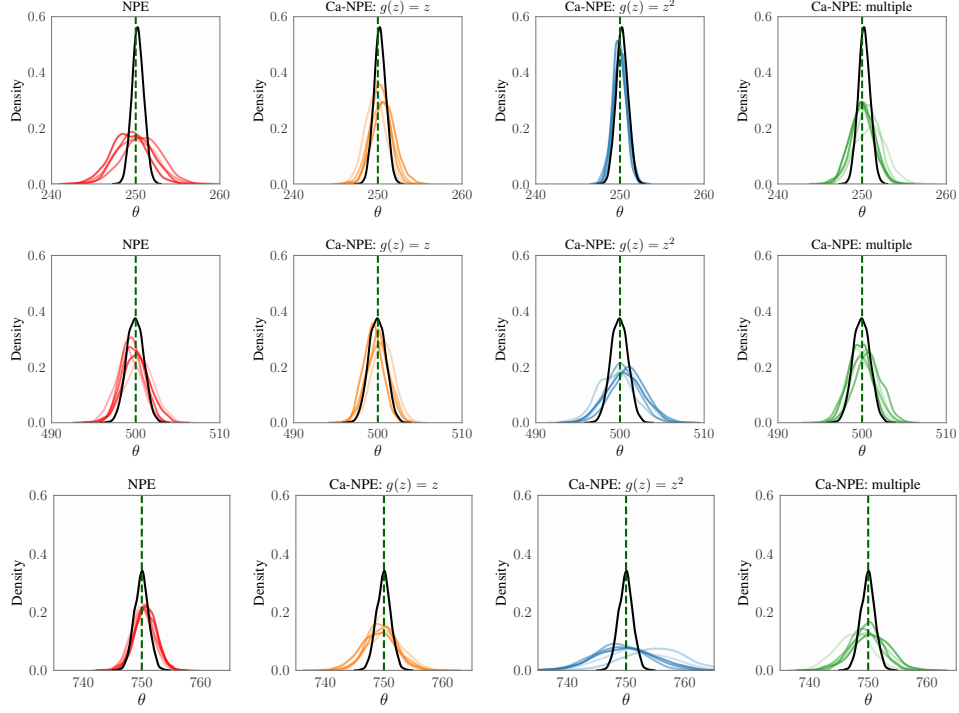


Figure 9: Kernel density estimates of NPE posteriors for the Gamma simulator using $\theta_{\text{true}} = 250$ (top row), $\theta_{\text{true}} = 500$ (middle row), and $\theta_{\text{true}} = 750$ (bottom row) with $n = 5000$ samples. The true posterior is in black.

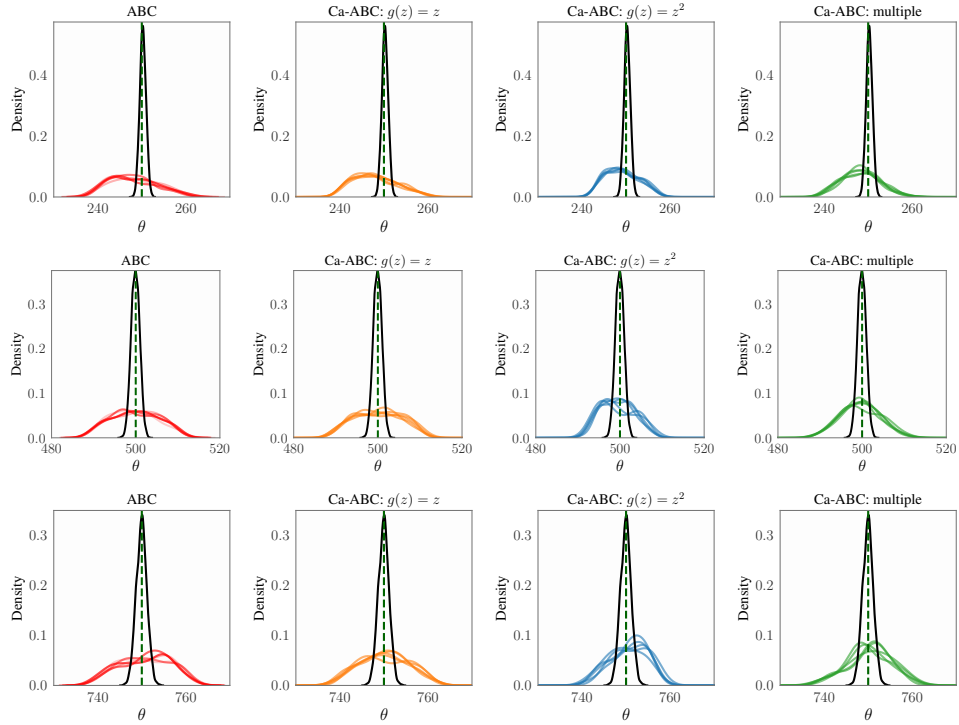


Figure 10: Kernel density estimates of ABC posteriors for the Gamma simulator using $\theta_{\text{true}} = 250$ (top row), $\theta_{\text{true}} = 500$ (middle row), and $\theta_{\text{true}} = 750$ (bottom row) with $n = 5000$ samples. The true posterior is in black.

Algorithm 2 Simulation of Homogeneous SIR model

Input: population size N , infection rate θ_1 , dispersion parameter k
 Initialize $i = 1$, $s = N - 1$
while $i > 0$ **do**
 Sample $I \sim \text{Gamma}(k, k)$
 Sample $Z \sim \text{Poisson}(\theta_1 \cdot I)$ ▷ Sample infectious period
 for $1, \dots, Z$ **do**
 Sample $u \sim \text{Uniform}(0, 1)$
 if $u < \frac{s}{N}$ **then**
 $s \leftarrow s - 1$
 $i \leftarrow i + 1$
 end if
 end for
 $i \leftarrow i - 1$ ▷ individual recovers or dies
end while
Output: Final epidemic size $N - s$.

C.4 Details and additional results for epidemiology experiments

We considered three models: the homogeneous SIR, the temporal SIR and a Bernoulli SIR. The homogeneous SIR is a simplified model with a single infection rate parameter, where simulation time increases as the rate rises. The temporal SIR model, which accounts for the progression of an epidemic outbreak, has two parameters controlling the infection and recovery rates, both of which impact simulation cost as shown in Figure 1. The Bernoulli SIR is a more complex version based on the temporal SIR, incorporating an additional parameter to control the probability of transmission between individuals. All three parameters of the Bernoulli SIR can influence the simulation cost.

Homogeneous SIR assumes a closed population where each individual has an equal probability of contacting any other individual. Initially, one individual is infected, and the epidemic spreads through infectious contacts until no infectious individuals remain. The detailed algorithm for this model is described in Algorithm 2. For simulations, the population size N is set to 10,000, and $k = 1$. We choose a uniform prior for the infection rate $\theta_1 \sim \mathcal{U}(1, 10)$, and we use $\theta_{\text{true}} = [5]^\top$ to generate the observed data. The estimated simulation costs w.r.t. parameters are shown in Figure 11(a).

Temporal SIR extends the basic SIR model by incorporating temporal dynamics, where the infection and removal events are tracked over time. This allows for the analysis of the epidemic’s progression. The detailed algorithm for this model is described in Algorithm 3. We discretise the outbreak lasted time into a number of bins and calculate the number of removals in each bin as the summary statistics. For simulations, the population size N is set to 1000, and the number of bins is set to $n_b = 10$. We choose uniform priors for the infection rate $\theta_1 \sim \mathcal{U}(0.1, 1.0)$ and the removal rate $\theta_2 \sim \mathcal{U}(0.1, 1.0)$, and set $\theta_{\text{true}} = [0.5, 0.5]^\top$ to generate the observed data.

Bernoulli SIR simulates the spread of an epidemic on a network represented by a Bernoulli random graph. Each individual in the network can infect their neighbours with a probability determined by θ_3 . The model tracks infection and removal events until the epidemic ends. The detailed algorithm for this model is described in Algorithm 4. For simulations, the population size N is set to 1000, and the number of bins is set to $n_b = 10$. We choose uniform priors for the infection rate $\theta_1 \sim \mathcal{U}(0.1, 1.0)$ and the removal rate $\theta_2 \sim \mathcal{U}(0.1, 1.0)$, and the edge probability $\theta_3 \sim \mathcal{U}(0.1, 1.0)$. We use $\theta_{\text{true}} = [0.5, 0.5, 0.5]^\top$ to generate the observed data. The estimated simulation costs w.r.t. parameters are shown in Figure 11(b).

Cost function. For epidemiology models, since $c(\theta)$ is not known in advance, we randomly sampled a small dataset $\{(\theta_i, y_i)\}_{i=1}^m$ with $m = 200$ parameter and computational time pairs, and estimated it by fitting a Gaussian process (GP) model, see Figure 11 for the plots. We used GPyTorch (Gardner et al., 2018) (<https://github.com/cornellius-gp/gpytorch>) to implement GP models.

Algorithm 3 Simulation of Temporal SIR model

Input: population size N , number of bins n_b , infection rate θ_1 , removal rate θ_2
 Initialize $i = 1$, $s = N - 1$, $t = 0$
 Initialize lists **times** = $[t]$, **types** = $[1]$ \triangleright 1 for infection, 2 for removal
while $i > 0$ **do**
 $\tau \leftarrow \text{Exp}(\frac{\theta_1}{N}is + \theta_2 I)$
 Simulate $u \sim \text{Uniform}(0, 1)$
 if $u < \frac{\theta_1 s}{\theta_1 s + N\theta_2}$ **then**
 $i \leftarrow i + 1$
 $s \leftarrow s - 1$
 Append 1 to **types**
 else
 $i \leftarrow i - 1$
 Append 2 to **types**
 end if
 $t \leftarrow t + \tau$
 Append t to **times**
end while
 $T \leftarrow$ last element of **times**
 Identify the times at which removals occurred from **times** based on **types**
 Discretise the interval $[0, T]$ into n_b bins and count the number of removals in each bin
Output: Final epidemic size $N - s$, duration of the epidemic T , removals per bin

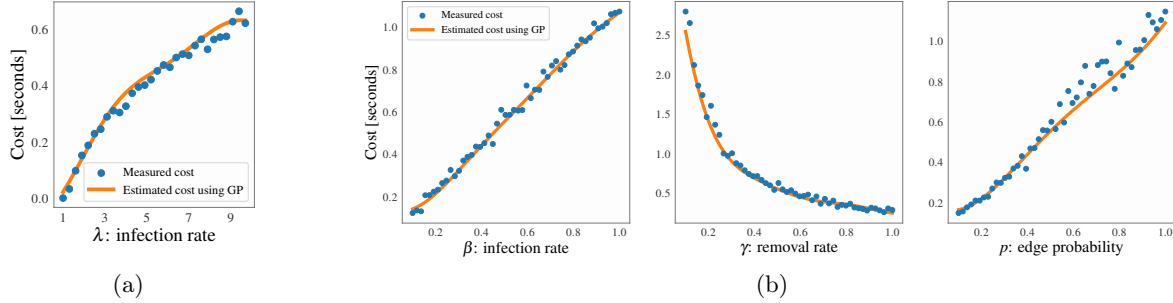


Figure 11: Costs of sampling one data point from (a) the Homogeneous SIR model, and (b) the Bernoulli SIR model, calculated as the average over 20 simulation runs. The corresponding GP fit for estimating the cost function is shown in orange. The GP is able to estimate the cost function well with limited amount of samples.

Components for multiple importance sampling. We use the $\text{CG} \times \text{ESS}$ metric to select the components for the recommended multiple importance sampling scheme, similar to Table 1. Taking the penalty function to be of the form $g(z) = z^k$, we compute $\text{CG} \times \text{ESS}$ for $k = \{0.25, 0.5, 0.75, 1, 1.5, 2\}$, see Table 9 for the values. In order to have the same components across all the epidemiology models, we select $k = \{0.5, 1, 2\}$ as the three components apart from the prior. This gives us the right balance between efficiency and cost. The corresponding values for g_{\min} and g_{\max} are reported in Table 10.

Table 9: $\text{CG} \times \text{ESS}$ for the epidemiology models.

$g(z)$	SBI	Ca-SBI					
		$z^{0.25}$	$z^{0.5}$	$z^{0.75}$	$z^{1.0}$	$z^{1.5}$	z^2
Homogen.	1.0	1.04	1.06	1.04	0.99	0.81	0.53
Temporal	1.0	0.93	0.87	0.73	0.56	0.26	0.14
Bernoulli	1.0	1.09	1.07	0.95	0.84	0.50	0.23

Table 10: g_{\min} and g_{\max} values for epidemiology models.

	$g(z) = z^{0.5}$		$g(z) = z$		$g(z) = z^2$	
	g_{\min}	g_{\max}	g_{\min}	g_{\max}	g_{\min}	g_{\max}
Homogen.	0.1498	0.7929	0.0225	0.6288	0.0005	0.3953
Temporal	0.0930	0.3250	0.0087	0.1056	0.0001	0.0112
Bernoulli	0.1197	0.9190	0.0143	0.8447	0.0002	0.7134

Algorithm 4 Simulation of Bernoulli SIR model

Input: population size N , number of bins n_b , infection rate θ_1 , removal rate θ_2 , edge probability θ_3
 Initialize $t = 0$, $r = 0$, set individual 1 infectious and all other individuals susceptibles
 Initialize list **times** = $[t]$
 Let \mathcal{I} and \mathcal{S} denote the sets of infections and susceptible, respectively.
 Generate connectivity matrix G , where for $i < j$, $G_{ij} = G_{ji} \sim \text{Bern}(\theta_3)$
while $|\mathcal{I}| > 0$ **do**
 Simulate $\tau \sim \text{Exp}(\theta_1 \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{S}} G_{ij} + \theta_2 |\mathcal{I}|)$
 $t \leftarrow t + \tau$
 Simulate $u \sim \text{Uniform}(0, 1)$
 if $u < \frac{\theta_1 \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{S}} G_{ij}}{\theta_1 \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{S}} G_{ij} + \theta_2 |\mathcal{I}|}$ **then**
 Sample J from \mathcal{S} with $P(J = j) = \sum_{i \in \mathcal{I}} G_{ij} / \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{S}} G_{ik}$
 $\mathcal{S} \leftarrow \mathcal{S} \setminus \{J\}$, $\mathcal{I} \leftarrow \mathcal{I} \cup \{J\}$
 else
 Sample K uniformly from \mathcal{I}
 Remove individual K ; $\mathcal{I} \leftarrow \mathcal{I} \setminus \{K\}$
 Append t to **times**
 $r \leftarrow r + 1$
 end if
end while
 $T \leftarrow$ last element of **times**
 Discretise the interval $[0, T]$ into n_b bins and count the number of removals in each bin
Output: Final epidemic size r , duration of the epidemic T , removals per bin

Additional performance metrics. Apart from the MMD, we also report additional performance metrics such as the classifier 2-sample test (C2ST) (Friedman, 2003; Lopez-Paz and Oquab, 2017) and the marginal two-sample Kolmogorov-Smirnov (KS) (Hodges Jr, 1958) test scores for the SIR experiments. C2ST involves training a classifier to distinguish between two distributions, which in our case is the reference posterior and the inferred posterior. The test outputs the classification accuracy, shown in Table 11, which at best is 0.5 in case the two distributions are the same. The two-sample KS test statistic, reported in is Table 12, quantifies the distance between the empirical cumulative distribution functions of two distributions. Our conclusions remain the same.

Table 11: C2ST scores (\downarrow) for NPE and Ca-NPE on the three SIR models. The mean and standard deviation from 50 runs are reported.

	NPE	Ca-NPE ($g(z) = z^{0.5}$)	Ca-NPE ($g(z) = z$)	Ca-NPE ($g(z) = z^2$)	Ca-NPE (multiple)
Homogen.	0.56(0.04)	0.55(0.02)	0.55(0.03)	0.77(0.06)	0.59(0.04)
Temporal	0.60(0.03)	0.66(0.05)	0.70(0.03)	0.75(0.04)	0.65(0.03)
Bernoulli	0.93(0.01)	0.93(0.01)	0.93(0.01)	0.92(0.02)	0.92(0.02)

Table 12: Two-sample KS test scores (\downarrow) for NPE and Ca-NPE on the three SIR models. The mean and standard deviation from 50 runs are reported.

	NPE	Ca-NPE ($g(z) = z^{0.5}$)	Ca-NPE ($g(z) = z$)	Ca-NPE ($g(z) = z^2$)	Ca-NPE (multiple)
Homogen.	0.12(0.05)	0.11(0.04)	0.11(0.05)	0.49(0.14)	0.15(0.06)
Temporal	[0.12(0.06),	[0.11(0.05),	[0.13(0.05),	[0.13(0.07),	[0.12(0.04),
	0.12(0.06)]	0.13(0.07)]	0.13(0.07)]	0.19(0.07)]	0.13(0.06)]
Bernoulli	[0.12(0.03),	[0.13(0.03),	[0.14(0.03),	[0.15(0.03),	[0.12(0.03),
	0.06(0.03),	0.05(0.02),	0.07(0.03),	0.13(0.03),	0.05(0.02),
	0.08(0.03)]	0.09(0.02)]	0.09(0.03)]	0.12(0.04)]	0.10(0.04)]

C.5 Details of the radio propagation experiment

The radio propagation model used in Section 4.3 is a stochastic of how the environment affects the transmission of radio signals between a transmitter and a receiver. By adjusting the parameters of this model, one can simulate

different communication environments, such as a small room, a large hall or a corridor, which then helps in testing new wireless communication systems. A pseudocode for simulating from this model is given in Algorithm 5, see Huang et al. (2023) for a more detailed description.

This model has four parameters, out of which, only θ_3 affects the computational cost, as shown in Figure 12. We use a linear model to estimate the cost function. For quadratic penalty function $g(z) = z^2$, $g_{\min} = 1.7$ and $g_{\max} = 139.2$. Again, we use the CG \times ESS metric to select the components for the multiple importance sampling variant of our method, see Table 13 for the values. We see the $g(z) = z$ yields the same CG \times ESS as the standard SBI method, which means that we do not lose out on efficiency with this choice of g . Hence, we take the cost-aware components to be $\{z, z^2, z^3\}$ in order to get computational advantages from the higher powers.

Figure 13 shows the NPE results for the radio propagation model, analogous to the NLE results of Figure 6 from the main text. The observations remain the same: the posteriors obtained from NPE and Ca-NPE (for both multiple cost-aware importance sampling and $g(z) = z^2$) are similar, indicating that Ca-NPE reduces cost without sacrificing posterior accuracy.

Figure 12: Cost of sampling $m = 50$ iid realisations from the radio propagation model as a function of θ_3 .

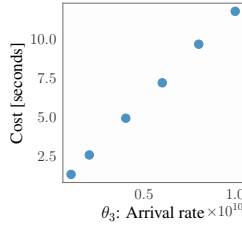


Table 13: CG \times ESS for the radio propagation model.

$g(z)$	SBI	Ca-SBI			
	-	$z^{0.5}$	z	z^2	z^3
CG \times ESS	1.0	1.06	1.0	0.60	0.22

Algorithm 5 Simulation of the radio propagation model

Input: Bandwidth $B = 4$ GHz, length of time-domain signal $N_s = 801$, $m = 50$ realisations, reverberation gain θ_1 , reverberation time θ_2 , arrival rate θ_3 , noise variance θ_4

Set $\Delta f = B/(N_s - 1)$ and $t_{\max} = 1/\Delta f$

while $i \leq m$ **do**

 Sample $N_{\text{points}} \sim \text{Poisson}(\theta_3 t_{\max})$

 Sample $\tau_j \sim \mathcal{U}(0, t_{\max})$, $j = 1, \dots, N_{\text{points}}$ and arrange them in ascending order

 Sample $\beta_j \sim \mathcal{CN}(0, \sigma_P^2(\tau_j))$, $j = 1, \dots, N_{\text{points}}$, where $\sigma_P^2(\tau) = \theta_1 \exp(-\tau/\theta_2)/\theta_3 B$

 Compute $H_l = \sum_{j=1}^{N_{\text{points}}} \beta_j \exp(-i2\pi l \Delta f \tau_j)$, $l = 1, \dots, N_s$ (here, i refers to the imaginary unit)

 Sample $W_l \sim \mathcal{CN}(0, \theta_4)$ and compute $Y_l = H_l + W_l$, $l = 1, \dots, N_s$

 Take inverse Fourier transform of Y_1, \dots, Y_{N_s} to obtain the time-domain signal y_1, \dots, y_{N_s}

 Compute the temporal moments as $M_{ik} = \int t^k y(t) dt$ for $k = 0, 1, 2$

end while

Compute the sample mean and sample variance of $\{M_{ik}\}_{i=1}^m$.

Output: Data vector $\mathbf{x} \in \mathbb{R}^6$ consisting of three means and three variance estimates of temporal moments.

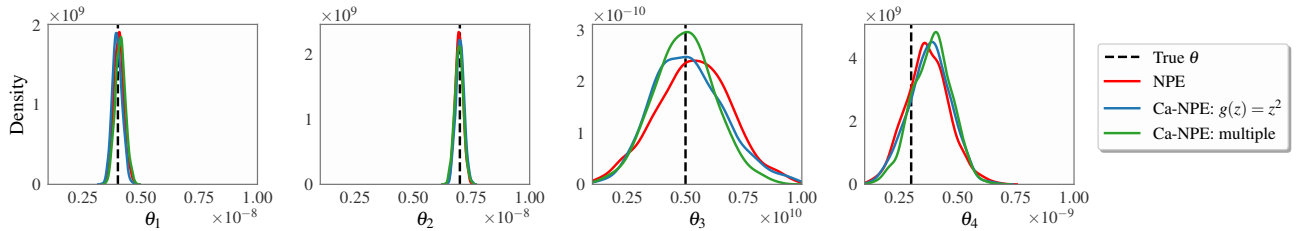


Figure 13: Marginals of the approximate posterior for the radio propagation model using NPE and our Ca-NPE method with $n = 10,000$. The Ca-NPE methods perform similar to NPE.