

---

# Amortized Monte Carlo Integration

---

Adam Goliński<sup>\*12</sup> Frank Wood<sup>3</sup> Tom Rainforth<sup>\*1</sup>

## Abstract

Current approaches to amortizing Bayesian inference focus solely on approximating the posterior distribution. Typically, this approximation is, in turn, used to calculate expectations for one or more target functions—a computational pipeline which is inefficient when the target function(s) are known upfront. In this paper, we address this inefficiency by introducing AMCI, a method for *amortizing Monte Carlo integration* directly. AMCI operates similarly to amortized inference but produces three distinct amortized proposals, each tailored to a different component of the overall expectation calculation. At runtime, samples are produced separately from each amortized proposal, before being combined to an overall estimate of the expectation. We show that while existing approaches are fundamentally limited in the level of accuracy they can achieve, AMCI can theoretically produce arbitrarily small errors for any integrable target function using only a single sample from each proposal at runtime. We further show that it is able to empirically outperform the theoretically optimal self-normalized importance sampler on a number of example problems. Furthermore, AMCI allows not only for amortizing over datasets but also amortizing over target functions.

## 1. Introduction

At its core, Bayesian modeling is rooted in the calculation of expectations: the eventual aim of modeling is typically to make a decision or to construct predictions for unseen data, both of which take the form of an expectation under the posterior (Robert, 2007). This aim can

---

<sup>\*</sup>Equal contribution <sup>1</sup>Department of Statistics, University of Oxford, United Kingdom <sup>2</sup>Department of Engineering Science, University of Oxford, United Kingdom <sup>3</sup>Department of Computer Science, University of British Columbia, Vancouver, Canada. Correspondence to: Adam Goliński <adamg@robots.ox.ac.uk>.

thus be summarized in the form of one or more expectations  $\mathbb{E}_{p(x|y)}[f(x)]$ , where  $f(x)$  is a target function and  $p(x|y)$  is the posterior distribution on  $x$  for some data  $y$ , which we typically only know up to a normalizing constant  $p(y)$ . More generally, expectations with respect to distributions with unknown normalization constant are ubiquitous throughout the sciences (Robert & Casella, 2013).

Sometimes  $f(x)$  is not known up front. Here it is typically convenient to first approximate  $p(x|y)$ , e.g. in the form of Monte Carlo (MC) samples, and then later use this approximation to calculate estimates, rather than addressing the target expectations directly.

However, it is often the case in practice that a particular target function, or class of target functions, is known a priori. For example, in decision-based settings  $f(x)$  takes the form of a loss function, while any posterior predictive distribution constitutes a set of expectations with respect to the posterior, parameterized by the new input. Though often overlooked, it is well established that in such *target-aware* settings the aforementioned pipeline of first approximating  $p(x|y)$  and then using this as a basis for calculating  $\mathbb{E}_{p(x|y)}[f(x)]$  is suboptimal as it ignores relevant information in  $f(x)$  (Hesterberg, 1988; Wolpert, 1991; Oh & Berger, 1992; Evans & Swartz, 1995; Meng & Wong, 1996; Chen & Shao, 1997; Gelman & Meng, 1998; Lacoste-Julien et al., 2011; Owen, 2013; Rainforth et al., 2018b). As we will later show, the potential gains in such scenarios can be arbitrarily large.

In this paper, we extend these ideas to *amortized* inference settings (Stuhlmüller et al., 2013; Kingma & Welling, 2014; Ritchie et al., 2016; Paige & Wood, 2016; Le et al., 2017; 2018a; Webb et al., 2018), wherein one looks to amortize the cost of inference across different possible datasets by learning an artifact that assists the inference process at runtime for a given dataset. Existing approaches do not operate in a target-aware fashion, such that even if the inference network learns proposals that perfectly match the true posterior for every possible dataset, the resulting estimator is still sub-optimal.

To address this, we introduce AMCI, a framework for performing *Amortized Monte Carlo Integration*. Though still based on learning amortized proposals distributions, AMCI varies from standard amortized inference ap-

proaches in three respects. First, it operates in a target-aware fashion, incorporating information about  $f(x)$  into the amortization artifacts. Second, rather than using self-normalization, AMCI employs three distinct proposals for separately estimating  $\mathbb{E}_{p(x)} [p(y|x) \max(f(x), 0)]$ ,  $\mathbb{E}_{p(x)} [-p(y|x) \min(f(x), 0)]$ , and  $\mathbb{E}_{p(x)} [p(y|x)]$ , before combining these into an overall estimate. This breakdown allows for arbitrary performance improvements compared to self-normalized importance sampling (SNIS). Finally, to account for cases in which multiple possible target functions may be of interest, AMCI also allows for amortization over parametrized functions  $f(x; \theta)$ .

Remarkably, AMCI is able to achieve an arbitrarily low error at run-time using only a single sample from each proposal given sufficiently powerful amortization artifacts, contrary to the fundamental limitations on the accuracy of conventional amortization approaches. This ability is based around its novel breakdown of the target expectation into separate components, the subsequent utility of which extends beyond the amortized setting we consider here.

## 2. Background

### 2.1. Importance Sampling

Importance Sampling (IS), in its most basic form, is a method for approximating an expectation  $\mathbb{E}_{\pi(x)} [f(x)]$  when it is either not possible to sample from  $\pi(x)$  directly, or when the simple MC estimate,  $\frac{1}{N} \sum_{n=1}^N f(x_n)$  where  $x_n \sim \pi(x)$ , has problematically high variance (Hesterberg, 1988; Wolpert, 1991). Given a proposal  $q(x)$  from which we can sample and for which we can evaluate the data, it forms the following estimate

$$\mu := \mathbb{E}_{\pi(x)} [f(x)] = \int f(x) \frac{\pi(x)}{q(x)} q(x) dx \quad (1)$$

$$\approx \hat{\mu} := \frac{1}{N} \sum_{n=1}^N f(x_n) w_n \quad (2)$$

where  $x_n \sim q(x)$  and  $w_n := \pi(x_n)/q(x_n)$  is known as the importance weight of sample  $x_n$ .

In practice, one often does not have access to the normalized form of  $\pi(x)$ . For example, in Bayesian inference settings, we typically have  $\pi(x) = p(x|y) \propto p(x, y)$ . Here we can use our samples to both approximate the normalization constant and the unnormalized integral. Thus if  $\pi(x) \propto \gamma(x)$ , we have

$$\mathbb{E}_{\pi(x)} [f(x)] = \frac{\int \frac{f(x)\gamma(x)}{q(x)} q(x) dx}{\int \frac{\gamma(x)}{q(x)} q(x) dx} \approx \frac{\sum_{n=1}^N f(x_n) w_n}{\sum_{n=1}^N w_n} \quad (3)$$

where  $x_n \sim q(x)$ , and  $w_n := \gamma(x_n)/q(x_n)$ . This approach is known as self-normalized importance sampling (SNIS). Conveniently, we can also construct the SNIS estimate lazily by calculating the empirical measure, i.e. stor-

ing weighted samples,

$$\pi(x) \approx \sum_{n=1}^N w_n \delta_{x_n}(x) / \sum_{n=1}^N w_n \quad (4)$$

and then using this to construct the estimate in (3) when  $f(x)$  becomes available. As such, we can also think of SNIS as a method for Bayesian inference as, informally speaking, the empirical measure produced can be thought of as an approximation of the posterior.

For a general unknown target, the optimal proposal, i.e. the proposal which results in estimator having lowest possible variance, is the target distribution  $q(x) = \pi(x)$  (see e.g. (Rainforth, 2017, 5.3.2.2)). However, this no longer holds if we have some information about  $f(x)$ . In this target-aware scenario, the optimal behavior turns out to depend on whether we are self-normalizing or not.

For the non-self-normalized case, the optimal proposal can be shown to be  $q^*(x) \propto \pi(x) |f(x)|$  (Owen, 2013). Interestingly, in the case where  $f(x) \geq 0 \forall x$ , this leads to an exact estimator, i.e.  $\hat{\mu} = \mu$  (with  $\hat{\mu}$  as per (2)). To see this, notice that the normalizing constant for  $q^*(x)$  is  $\int \pi(x) f(x) dx = \mu$  and hence  $q^*(x) = \pi(x) f(x) / \mu$ . Therefore, even when  $N = 1$ , any possible value of the resulting sample  $x_1$  yields an  $\hat{\mu}$  satisfying  $\hat{\mu} = f(x_1) \pi(x_1) / q^*(x_1) = \mu$ .

In the self-normalized case, the optimal proposal instead transpires to be  $q^*(x) \propto \pi(x) |f(x) - \mu|$  (Hesterberg, 1988). In this case, one can no longer achieve a zero variance estimator for finite  $N$  and nonconstant  $f(x)$ . Instead, the achievable error is lower bounded by (Owen, 2013)

$$\mathbb{E}[(\hat{\mu} - \mu)^2] \geq \frac{1}{N} \left( \mathbb{E}_{\pi(x)} [|f(x) - \mu|] \right)^2, \quad (5)$$

creating a fundamental limit on the performance of SNIS, even when information about  $f(x)$  is incorporated.

Given that these optimal proposals make use of the true expectation  $\mu$ , we will clearly not have access to them in practice. However, they provide a guide for the desirable properties of a proposal and can be used as targets for adaptive IS methods (see (Bugallo et al., 2017) for a recent review).

### 2.2. Inference Amortization

Inference amortization involves learning an *amortization artifact* that takes in datasets and produces proposals tailored to the corresponding inference problems. This amortization artifact typically takes the form of a parametrized proposal,  $q(x; \varphi(y; \eta))$ , which takes in data  $y$  and produces proposal parameters using an *inference network*  $\varphi(y; \eta)$ , which itself has parameters  $\eta$ . When clear from the context, we will use the shorthand  $q(x; y, \eta)$  for this proposal.

Though the exact process varies with context, the inference network is usually trained either by drawing latent-data sample pairs from the joint  $p(x, y)$  (Paige & Wood, 2016;

Le et al., 2017; 2018b), or by drawing mini-batches from a large dataset using stochastic variational inference approaches (Hoffman et al., 2013; Kingma & Welling, 2014; Rezende et al., 2014; Ritchie et al., 2016). Once trained, it provides an efficient means of approximately sampling from the posterior of a particular dataset, e.g. using SNIS.

Out of several variants, we focus on the method introduced by Paige & Wood (2016), as this is the one AMCI builds upon. In their approach,  $\eta$  is trained to minimize the expectation of  $D_{KL} [p(x|y) || q(x; y, \eta)]$  across possible datasets  $y$ , giving the objective

$$\begin{aligned} \mathcal{J}(\eta) &= \mathbb{E}_{p(y)} \left[ D_{KL} [p(x|y) || q(x; y, \eta)] \right] \\ &= \mathbb{E}_{p(x,y)} [-\log q(x; y, \eta)] + \text{const wrt } \eta \end{aligned} \quad (6)$$

We note that the distribution  $p(y)$  over which we are taking the expectation is actually chosen somewhat arbitrarily: it simply dictates how much the network prioritizes a good amortization for one dataset over another; different choices are equally valid and imply different loss functions.

This objective requires us to be able to sample from the joint distribution  $p(x, y)$  and it can be optimized using gradient methods since the gradient can be easily evaluated:

$$\nabla_{\eta} \mathcal{J}(\eta) = \mathbb{E}_{p(x,y)} [-\nabla_{\eta} \log q(x; y, \eta)]. \quad (7)$$

### 3. AMCI

Amortized Monte Carlo integration (AMCI) is a framework for amortizing the cost of calculating expectations  $\mu(y, \theta) := \mathbb{E}_{\pi(x;y)} [f(x; \theta)]$ . Here  $y$  represents changeable aspects of the reference distribution  $\pi(x; y)$  (e.g. the dataset) and  $\theta$  represents changeable parameters of the target function  $f(x; \theta)$ . The reference distribution is typically known only up to a normalization constant, i.e.  $\pi(x; y) = \gamma(x; y)/Z$  where  $\gamma(x; y)$  can be evaluated pointwise, but  $Z$  is unknown. AMCI can still be useful in settings where  $Z$  is known, but here we can simply use its known value rather than constructing a separate estimator.

Amortization can be performed across  $y$  and/or  $\theta$ . When amortizing over  $y$ , the function does not need to be explicitly parameterized; we just need to be able to evaluate it pointwise. Similarly, when amortizing over  $\theta$ , the reference distribution can be fixed. In fact, AMCI can be used for a parameterized set of conventional integration problems  $\int_{x \in \mathcal{X}} f(x; \theta) dx$  by exploiting the fact that

$$\int_{x \in \mathcal{X}} f(x; \theta) dx = \mathbb{E}_{\pi(x)} [f(x; \theta) / \pi(x)] \quad (8)$$

for any  $\pi(x)$  where  $\pi(x) \neq 0 \forall x \in \mathcal{X}$  for which  $f(x) \neq 0$ .

For consistency of notation with the amortized inference literature, we will presume a Bayesian setting in the rest of this section, i.e.  $\pi(x; y) = p(x|y)$  and  $\gamma(x; y) = p(x, y)$ .

### 3.1. Estimator

Existing amortized inference methods implicitly evaluate expectations using SNIS (or some other form of self-normalized estimator (Paige & Wood, 2016; Le et al., 2018a)), targeting the posterior as the optimal proposal  $q^*(x; y) \approx p(x|y)$ . Not only is this proposal suboptimal when information about the target function is available, there is a lower bound on the accuracy the SNIS approach itself can achieve as shown in (5).

AMCI overcomes these limitations by breaking down the overall expectation into separate components and constructing separate estimates for each. We can first break down the target expectation into the ratio of the “unnormalized expectation” and the normalization constant:

$$\begin{aligned} \mu(y, \theta) &:= \mathbb{E}_{p(x|y)} [f(x; \theta)] = \frac{\mathbb{E}_{p(x|y)} [f(x; \theta) p(y)]}{\mathbb{E}_{p(x)} [p(y|x)]} \\ &= \frac{\mathbb{E}_{q_1(x;y,\theta)} \left[ \frac{f(x;\theta)p(x,y)}{q_1(x;y,\theta)} \right]}{\mathbb{E}_{q_2(x;y)} \left[ \frac{p(x,y)}{q_2(x;y)} \right]} =: \frac{E_1}{E_2} \end{aligned} \quad (9)$$

where  $q_1(x; y, \theta)$  and  $q_2(x; y)$  are two separate proposals, used respectively for each of the two expectations  $E_1$  and  $E_2$ . We note that the proposal  $q_1(x; y, \theta)$  may depend not only on the observed dataset  $y$ , but also on the parameters of the target function  $\theta$ .

We can now generate separate MC estimates for  $E_1$  and  $E_2$ , and take their ratio to estimate the overall expectation:

$$\begin{aligned} \mu(y, \theta) &\approx \hat{\mu}(y, \theta) := \hat{E}_1 / \hat{E}_2 \quad \text{where} \\ \hat{E}_1 &:= \frac{1}{N} \sum_{n=1}^N \frac{f(x'_n; \theta) p(x'_n, y)}{q_1(x'_n; y, \theta)} \quad x'_n \sim q_1(x; y, \theta) \\ \hat{E}_2 &:= \frac{1}{M} \sum_{m=1}^M \frac{p(x_m, y)}{q_2(x_m; y)} \quad x_m \sim q_2(x; y). \end{aligned} \quad (10)$$

The key idea behind AMCI is that we can now **separately train each of these proposals to be good estimators for their respective expectation**, rather than rely on a single proposal to estimate both, as is implicitly the case for SNIS.

Consider, for example, the case where  $f(x; \theta) \geq 0$ . If  $q_1(x; y, \theta) \propto f(x; \theta) p(x|y)$  and  $q_2(x; y) \propto p(x|y)$  then both  $\hat{E}_1$  and  $\hat{E}_2$  will form exact estimators (as per Section 2.1), even if  $N = M = 1$ . Consequently, we achieve an exact estimator for  $\mu(y, \theta)$ , allowing for arbitrarily large improvements over any SNIS estimator, because SNIS forces  $q_1(x; y, \theta)$  and  $q_2(x; y)$  to be the same distribution.

More generally, the optimal proposal for  $E_1$  and  $E_2$  are  $q_1(x; y, \theta) \propto |f(x; \theta)| p(x|y)$  and  $q_2(x; y) \propto p(x|y)$  respectively, with the latter always resulting in an exact estimator for  $E_2$ . Thus the more  $|f(x; \theta)| p(x|y)$  varies from  $p(x|y)$ , the worse the conventional approach of only amortizing

the posterior will perform, while the harder it becomes to construct a reasonable SNIS estimator even when information about  $f(x; \theta)$  is incorporated. Separately learning  $q_1(x; y, \theta)$  and  $q_2(x; y)$  means that each will become a better individual proposal and the overall estimator improves.

It turns out that we do not actually require the previous assumption of  $f(x; \theta) \geq 0 \forall x, \theta$  to achieve a zero variance estimator. Specifically, if we let<sup>1</sup>

$$f^+(x; \theta) = \max(f(x; \theta), 0) \quad \text{and} \quad (11)$$

$$f^-(x; \theta) = -\min(f(x; \theta), 0) \quad (12)$$

denote truncations of the target function into its positive and negative components (as per the concept of positivisation (Owen, 2013, 9.13)), then we can break down the overall expectation as follows

$$\begin{aligned} \mu(y, \theta) &= \frac{\mathbb{E}_{q_1^+(x; y, \theta)} \left[ \frac{f^+(x; \theta) p(x, y)}{q_1^+(x; y, \theta)} \right] - \mathbb{E}_{q_1^-(x; y, \theta)} \left[ \frac{f^-(x; \theta) p(x, y)}{q_1^-(x; y, \theta)} \right]}{\mathbb{E}_{q_2(x; y)} \left[ \frac{p(x, y)}{q_2(x; y)} \right]} \\ &=: \frac{E_1^+ - E_1^-}{E_2} \end{aligned} \quad (13)$$

where we now have three expectations and three proposals. Analogously to (10), we can construct estimates for each expectation separately and then combine them:

$$\begin{aligned} \mu(y, \theta) &\approx \hat{\mu}(y, \theta) := (\hat{E}_1^+ - \hat{E}_1^-) / \hat{E}_2 \quad \text{where} \\ \hat{E}_1^+ &:= \frac{1}{N} \sum_{n=1}^N \frac{f^+(x_n^+; \theta) p(x_n^+, y)}{q_1^+(x_n^+; y, \theta)} \quad x_n^+ \sim q_1^+(x; y, \theta) \\ \hat{E}_1^- &:= \frac{1}{K} \sum_{k=1}^K \frac{f^-(x_k^-; \theta) p(x_k^-, y)}{q_1^-(x_k^-; y, \theta)} \quad x_k^- \sim q_1^-(x; y, \theta) \\ \hat{E}_2 &:= \frac{1}{M} \sum_{m=1}^M \frac{p(x_m, y)}{q_2(x_m; y)} \quad x_m \sim q_2(x; y), \end{aligned} \quad (14)$$

which forms the AMCI estimator. The theoretical capability of this estimator is summarized in the following result, the proof for which is given in Appendix B.

**Theorem 1.** *If the following hold for a given  $\theta$  and  $y$ ,*

$$\mathbb{E}_{p(x)} [f^+(x; \theta) p(y|x)] < \infty \quad (15)$$

$$\mathbb{E}_{p(x)} [f^-(x; \theta) p(y|x)] < \infty \quad (16)$$

$$\mathbb{E}_{p(x)} [p(y|x)] < \infty \quad (17)$$

*and we use the corresponding set of optimal proposals  $q_1^+(x; y, \theta) \propto f^+(x; \theta) p(x, y)$ ,  $q_1^-(x; y, \theta) \propto f^-(x; \theta) p(x, y)$ , and  $q_2(x; y) \propto p(x, y)$ , then the AMCI*

<sup>1</sup>Practically, it may sometimes be beneficial to truncate the proposal about another point,  $c$ , by instead using  $f^+(x; \theta) = \max(f(x; \theta) - c, 0)$  and  $f^-(x; \theta) = -\min(f(x; \theta) - c, 0)$ , then adding  $c$  onto our final estimate.

*estimator defined in (14) satisfies*

$$\mathbb{E} [\hat{\mu}(y, \theta)] = \mu(y, \theta), \quad \text{Var} [\hat{\mu}(y, \theta)] = 0 \quad (18)$$

*for any  $N \geq 1$ ,  $K \geq 1$ , and  $M \geq 1$ , such that it forms an exact estimator for that  $\theta, y$  pair.*

Though our primary motivation for developing the AMCI estimator is its attractive properties in an amortization setting, we note that it may still be of use in static expectation calculation settings. Namely, the fact that it can achieve an arbitrarily low mean squared error for a given number of samples means it forms an attractive alternative to SNIS more generally, particularly when we are well-placed to hand-craft highly effective proposals and in adaptive importance sampling settings.

We note that individual elements of this estimator have previously appeared in the literature. For example, the general concept of using multiple proposals has been established in the context of multiple importance sampling (Veitch & Guibas, 1995). The use of two separate proposals for the unnormalized target and the normalizing constant (i.e. (10)), on the other hand, was recently independently suggested by Lamberti et al. (2018) in a non-amortized setting. However, we believe that the complete form of the AMCI estimator in (14) has not previously been suggested, nor its theoretical benefits or amortization considered.

## 3.2. Amortization

To evaluate (14), we need to learn three amortized proposals  $q_1^+(x; y, \theta)$ ,  $q_1^-(x; y, \theta)$ , and  $q_2(x; y)$ .

Learning  $q_2(x; y)$  is equivalent to the standard inference amortization problem and so we will just use the objective given by (6), as described in section 2.2.

The approaches for learning  $q_1^+(x; y, \theta)$  and  $q_1^-(x; y, \theta)$  are equivalent, other than the function that is used in the estimators. Therefore, for simplicity, we introduce our amortization procedure in the case where  $f(x; \theta) \geq 0 \forall x, \theta$ , such that we can need only learn a single proposal,  $q_1(x; y, \theta)$ , for the numerator as per (10). This trivially extends to the full AMCI setup by separately repeating the same training procedure for  $q_1^+(x; y, \theta)$  and  $q_1^-(x; y, \theta)$ .

### 3.2.1. FIXED FUNCTION $f(x)$

We first consider the scenario where  $f(x)$  is fixed (i.e. we are not amortizing over function parameters  $\theta$ ) and hence in this section we drop the dependence of  $q_1$  on  $\theta$ .

To learn the parameters  $\eta$  for the first amortized proposal  $q_1(x; y, \eta)$ , we need to adjust the target in (6) to incorporate the effect of the target function. Let  $E_1(y) := \mathbb{E}_{p(x)} [f(x) p(y|x)]$  and  $g(x|y) := \frac{f(x) p(x, y)}{E_1(y)}$ , i.e. the normalized optimal proposal for  $q_1$ . Naively adjusting (6)

leads to a double intractable objective

$$\begin{aligned} \mathcal{J}'_1(\eta) &= \mathbb{E}_{p(y)} [D_{KL}(g(x|y) || q_1(x; y, \eta))] \\ &= \mathbb{E}_{p(y)} \left[ - \int_{\mathcal{X}} \frac{f(x) p(x, y)}{E_1(y)} \log q_1(x; y, \eta) dx \right] \\ &\quad + \text{const wrt } \eta. \end{aligned} \quad (19)$$

Here the double intractability comes from the fact we do not know  $E_1(y)$  and, at least at the beginning of the training process, we cannot estimate it efficiently either.

To address this, we use our previous observation that the expectation over  $p(y)$  in the above objective is chosen somewhat arbitrarily. Namely, it dictates the relative priority of different datasets  $y$  during training and not the optimal proposal for each individual datapoint; disregarding the finite capacity of the network, the global optimum is still always  $D_{KL}[g(x|y) || q_1(x; y, \eta)] = 0, \forall y$ . We thus maintain a well-defined objective if we choose a different reference distribution over datasets. In particular, if we take the expectation with respect to  $h(y) \propto p(y)E_1(y)$ , we get

$$\begin{aligned} \mathcal{J}_1(\eta) &= \mathbb{E}_{h(y)} [D_{KL}(g(x|y) || q_1(x; y, \eta))] \\ &= c^{-1} \mathbb{E}_{p(x, y)} [-f(x) \log q_1(x; y, \eta)] \\ &\quad + \text{const wrt } \eta \end{aligned} \quad (20)$$

where  $c = \mathbb{E}_{p(y)} [E_1(y)] > 0$  is a positive constant that does not affect the optimization—it is the normalization constant for the distribution  $h(y)$ —and can thus be ignored. Each term in this expectation can now be evaluated directly, meaning we can again run stochastic gradient descent algorithms to optimize it. Note that this does not require evaluation of the density  $p(x, y)$ , only the ability to draw samples.

Interestingly, this choice of  $h(y)$  can be interpreted as giving larger importance to the values of  $y$  which yield larger  $E_1(y)$ . Informally, we could think about this choice as attempting to minimize the L1 errors of our estimates, that is  $\mathbb{E}_{p(y)} [|E_1(y) - \hat{E}_1(y)|]$ , presuming that the error in our estimation scales as the magnitude of the true value  $E_1(y)$ .

More generally, if we choose  $h(y) \propto p(y)E_1(y)\lambda(y)$  for some positive evaluable function  $\lambda : \mathcal{Y} \rightarrow \mathbb{R}^+$ , we get a tractable objective of the form

$$\mathcal{J}_1(\eta; \lambda) = \mathbb{E}_{p(x, y)} \left[ - \frac{f(x)}{\lambda(y)} \log q_1(x; y, \eta) \right]$$

up to a constant scaling factor and offset. We can thus use this trick to adjust the relative preference given to different datasets, while ensuring the objective is tractable.

### 3.2.2. PARAMETERIZED FUNCTION $f(x; \theta)$

As previously mentioned, AMCI also allows for amortization over parametrized functions, to account for cases in which multiple possible target functions may be of interest. We can incorporate this by using *pseudo prior*  $p(\theta)$  to

generate example parameters during our training.

Analogously to  $h(y)$ , the choice of  $p(\theta)$  determines how much importance we assign to different possible functions that we would like to amortize over. Since, in practice, perfect performance is unattainable over the entire space of  $\theta$ , the choice of  $p(\theta)$  is important and it will have an important effect on the performance of the system.

Incorporating  $p(\theta)$  is straightforward: we take the expectation of the fixed target function training objective over  $\theta$ . In this setting, our inference network  $\varphi$  needs to take  $\theta$  as input when determining the parameters of  $q_1$  and hence we let  $q_1(x; y, \theta, \eta) := q_1(x; \varphi(y, \theta; \eta))$ . If  $E_1(y, \theta) := \mathbb{E}_{p(x)} [f(x; \theta)p(y|x)]$ ,  $g(x|y; \theta) := f(x; \theta)p(x, y)/E_1(y, \theta)$ , and  $h(y, \theta) \propto p(y)p(\theta)E_1(y, \theta)$ , we get an objective which is analogous to (20):

$$\begin{aligned} \mathcal{J}_1(\eta) &= \mathbb{E}_{h(y, \theta)} [D_{KL}(g(x|y; \theta) || q_1(x; y, \theta, \eta))] \\ &= c^{-1} \cdot \mathbb{E}_{p(x, y)p(\theta)} [-f(x; \theta) \log q_1(x; y, \theta, \eta)] \\ &\quad + \text{const wrt } \eta \end{aligned} \quad (21)$$

where  $c = \mathbb{E}_{p(y)p(\theta)} [E_1(y, \theta)] > 0$  is again a positive constant that does not affect the optimization.

### 3.3. Efficient Training

If  $f(x; \theta)$  and  $p(x)p(\theta)$  are mismatched, i.e.  $f(x; \theta)$  is large in regions where  $p(x)p(\theta)$  is low, training by naïvely sampling from  $p(x)p(\theta)$  can be inefficient. Instead, it is preferable to try and sample from  $g(\theta, x) \propto p(x)p(\theta)f(x; \theta)$ . Though this is itself an intractable distribution, it represents a standard, rather than an amortized, inference problem and so it is much more manageable than the overall training. Namely, as the samples do not depend on the proposal we are learning or the datasets, we can carry out this inference process as a pre-training step that is substantially less costly than the problem of training the inference networks itself.

One approach is to construct an MCMC sampler targeting  $g(\theta, x)$  to generate the samples, which can be done upfront before training. Another is to use an importance sampler

$$\begin{aligned} \mathcal{J}_1(\eta) &= \text{const wrt } \eta \\ &\quad + c^{-1} \mathbb{E}_{q'(\theta, x)p(y|x)} \left[ - \frac{p(\theta)p(x)f(x; \theta)}{q'(\theta, x)} \log q_1(x; y, \theta, \eta) \right] \end{aligned} \quad (22)$$

where  $q'(\theta, x)$  is a proposal as close to  $g(\theta, x)$  as possible.

In the case of non-parameterized functions  $f(x)$ , there is no need to take an expectation over  $p(\theta)$ , and we instead desire to sample from  $g(x) \propto p(x)f(x)$ .

## 4. Experiments

Even though AMCI is theoretically able to achieve exact estimators with a finite number of samples, this will rarely be the case for practical problems, for which learning per-

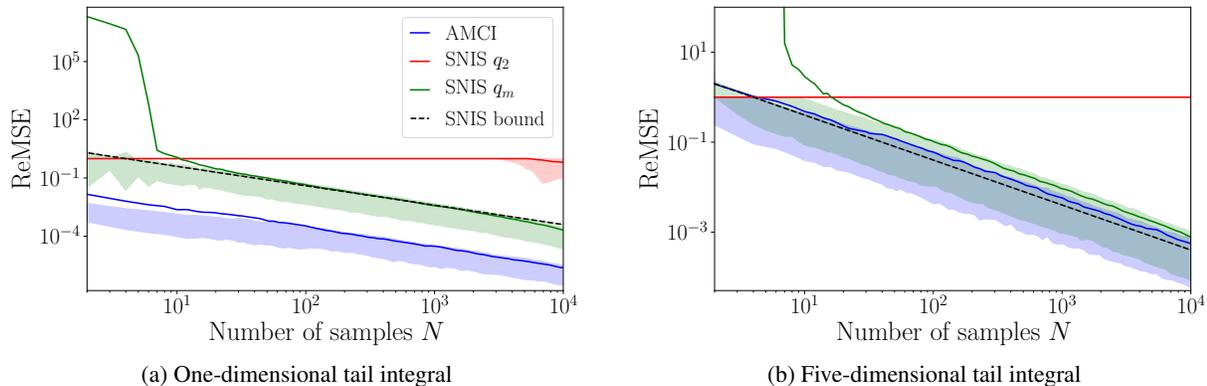


Figure 1: Relative mean squared errors (as per (25)) for [left] the one-dimensional and [right] the five-dimensional tail integral example. The solid lines for each estimator indicate the median of  $\delta(y, \theta)$  estimated using a common set of 100 samples from  $y, \theta \sim p(y)p(\theta)$ , with the corresponding  $\delta(y, \theta)$  then each separately estimated using 100 samples of the respective  $\hat{\delta}(y, \theta)$ . The shading instead shows the estimates from replacing  $\delta(y, \theta)$  with the 25% and 75% quantiles of  $\hat{\delta}(y, \theta)$  for a given  $y$  and  $\theta$ . The median of  $\delta(y, \theta)$  is at times outside of this shaded region as  $\delta(y, \theta)$  is often dominated by a few large outliers. The dashed line shows the median of  $\delta(y, \theta)$  with the  $\delta(y, \theta)$  corresponding to the ReMSE optimal SNIS estimator, namely  $(\mathbb{E}_{p(x|y)}[|f(x; \theta) - \mu(y, \theta)|])^2/N$  as per (5), which is itself estimated (with only nominal error) using  $10^6$  samples. We note that the error for SNIS with  $q_2$  proposal is to a large extent flat because there is not a single sample in the estimator for which  $f(x; \theta) > 0$ , such that they return  $\hat{\mu}(y, \theta) = 0$  and hence give  $\delta(y, \theta) = 1$ . In Figure (b) the SNIS  $q_m$  line reaches the ReMSE value of  $10^{18}$  at  $N=2$  and the y-axis limits have been readjusted to allow clear comparison at higher  $N$ . This effect is caused by the bias of SNIS: these extremely high errors for SNIS  $q_m$  arise when all  $N$  samples happen to be drawn from distribution  $q_1$ , for further explanation and the full picture see Figure 5 in Appendix A.

fect proposals is not typically realistic, particularly in amortized contexts (Cremer et al., 2018). It is therefore necessary to test its empirical performance to assert that gains are possible with inexact proposals. To this end, we investigate AMCI’s performance on two illustrative examples.

Our primary baseline is the SNIS approach implicitly used by most existing inference amortization methods, namely the SNIS estimator with proposal  $q_2(x; y)$ . Though this effectively represents the previous state-of-the-art in amortized expectation calculation, it turns out to be a very weak baseline. We, therefore, introduce another simple approach one could hypothetically consider using: training separate proposals as per AMCI, but then using this to form a mixture distribution proposal for an SNIS estimator. For example, in the scenario where  $f(x; \theta) \geq 0 \forall x, \theta$  (such that we only need to learn two proposals), we can use

$$q_m(x; y, \theta) = \frac{1}{2}q_1(x; y, \theta) + \frac{1}{2}q_2(x; y) \quad (23)$$

as an SNIS proposal that takes into account the needs of both  $E_1$  and  $E_2$ . We refer to this method as the *mixture* SNIS estimator and emphasize that it represents a novel amortization approach in its own right.

We also compare AMCI to the theoretically optimal SNIS estimator, i.e. the error bound given by (5). As we will show, AMCI is often able to empirically outperform this bound, thereby giving better performance than *any* approach based on SNIS, whether that approach is amortized

or not. This is an important result and, it particular, it highlights that the potential significance of the AMCI estimator extends beyond the amortized setting we consider here.

We further consider using SNIS with proposal  $q_1(x; y, \theta)$ . However, this transpires to perform extremely poorly throughout (far worse than  $q_2(x; y)$ ) and so we omit its results from the main paper, giving them in Appendix A.

In all experiments, we use the same number of sample from each proposal to form the estimate (i.e.  $N = M = K$ ).

An implementation for AMCI and our experiments is available at <http://github.com/talesa/amci>.

#### 4.1. Tail Integral Calculation

We start with the conceptually simple problem of calculating tail integrals for Gaussian distributions, namely

$$p(x) = \mathcal{N}(x; 0, \Sigma_1) \quad p(y|x) = \mathcal{N}(y; x, \Sigma_2) \quad (24)$$

$$f(x; \theta) = \prod_{i=1}^D \mathbb{1}_{x_i > \theta_i} \quad p(\theta) = \text{UNIFORM}(\theta; [0, u_D]^D)$$

where  $D$  is the dimensionality, we set  $\Sigma_2 = I$ , and  $\Sigma_1$  is a fixed covariance matrix (for details see Appendix C).

This problem was chosen because it permits easy calculation of the ground truth expectations by exploiting analytic simplifications, while remaining numerically challenging for values of  $\theta$  far away from the mean when we do not use these simplifications. We performed one and

five-dimensional variants of the experiment.

We use normalizing flows (Rezende & Mohamed, 2015) to construct our proposals, providing a flexible and powerful means of representing the target distributions. Details are given in Appendix C. Training was done by using importance sampling to generate the values of  $\theta$  and  $x$  as per (22) with  $q'(\theta, x) = p(\theta) \cdot \text{HALFNORMAL}(x; \theta, \text{diag}(\Sigma_2))$ .

To evaluate AMCI and our baselines we use the relative mean squared error (ReMSE)  $\delta(y, \theta) = \mathbb{E}[\hat{\delta}(y, \theta)]$ , where

$$\hat{\delta}(y, \theta) = \frac{(\mu(y, \theta) - \hat{\mu}(y, \theta))^2}{\mu(y, \theta)^2} \quad (25)$$

and  $\hat{\mu}(y, \theta)$  is our estimate for  $\mu(y, \theta)$ . We then consider summary statistics across different  $\{y, \theta\}$ , such as its median when  $y, \theta \sim p(y)p(\theta)$ .<sup>2</sup> In calculating this,  $\delta(y, \theta)$  was separately estimated for each value of  $y$  and  $\theta$  using 100 samples of  $\hat{\delta}(y, \theta)$  (i.e. 100 realizations of the estimator).

As shown in Figure 1, AMCI outperformed SNIS in both the one- and five-dimensional cases. For the one-dimensional example, AMCI significantly outperformed all of SNIS  $q_2$ , SNIS  $q_m$ , and the theoretically optimal SNIS estimator. SNIS  $q_2$ , the approach implicitly taken by existing inference amortization methods, typically failed to place even a single sample in the tail of the distribution, even for large  $N$ . Interestingly, SNIS  $q_m$  closely matched the theoretical SNIS bound, suggesting that this amortized proposal is very close to the theoretically optimal one. However, this still constituted significantly worse performance than AMCI—taking about  $10^3$  more samples to achieve the same relative error—demonstrating the ability of AMCI to outperform the best possible SNIS estimator.

For the five-dimensional example, AMCI again significantly outperformed our main baseline SNIS  $q_2$ . Though it still also outperformed SNIS  $q_m$ , its advantage was less than in one-dimensional case, and it did not outperform the SNIS theoretical bound. SNIS  $q_m$  itself did not match the bound as closely as in the one-dimensional example either, suggesting that the proposals learned were worse than in the one-dimensional case. Further comparisons based on using the mean squared error (instead of ReMSE) are given in Appendix A and show qualitatively similar behavior.

## 4.2. Planning Cancer Treatment

To demonstrate how AMCI might be used in a more real-world scenario, we now consider an illustrative example relating to cancer diagnostic decisions. Imagine that an oncologist is trying to decide whether to administer a treatment to a cancer patient. Because the treatment is highly invasive, they only want to administer it if there is a realis-

<sup>2</sup>Variability in  $\delta(y, \theta)$  between different instances of  $\{y, \theta\}$  is considered in Figures 7 and 8 in Appendix A.

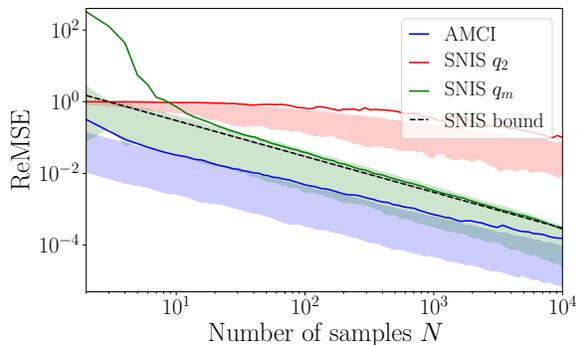


Figure 2: Relative mean squared errors for the cancer example. Conventions as per Figure 1. It is worth noting that it took about  $10^4$  more samples for the SNIS  $q_2$  estimator to achieve the same level of accuracy as the AMCI estimator.

tic chance of it being successful, i.e. that the tumor shrinks sufficiently to allow a future operation to be carried out. However, they are only able to make noisy observations about the current size of the tumor, and there are various unknown parameters pertaining to its growth, such as the patients predisposition to the treatment. To aid in the oncologists decision, the clinic provides a simulator of tumor evolution, a model of the latent factors required for this simulator, and a loss function for administering the treatment given the final tumor size. We wish to construct an amortization of this simulator, so that we can quickly and directly predict the expected loss function for administering the treatment from a pair of noisy observations of the tumor size taken at separate points in time. A detailed description of the model and proposal setup is in the Appendix C.3.

To evaluate the learned proposals we followed the same procedure as for the tail integral example. Results are presented in Figure 2. AMCI again significantly outperformed the literature baseline of SNIS  $q_2$ —it took about  $N = 10^4$  samples for SNIS  $q_2$  to achieve the level of relative error of AMCI for  $N = 2$ . AMCI further maintained an advantage over SNIS  $q_m$ , which itself again closely matched the optimal SNIS estimator. Further comparisons are given in Appendix A and show qualitatively similar behavior.

## 5. Discussion

In all experiments AMCI performed better than SNIS with either  $q_2$  or  $q_m$  for its proposal. Moreover, it is clear that AMCI is indeed able to break the theoretical bound on the achievable performance of SNIS estimators: in some cases AMCI is outperforming the best achievable error by any SNIS estimator, regardless of the proposal the latter uses. Interestingly, the mixture SNIS estimator we also introduced proved to be a strong baseline as it closely matched the theoretical baseline in both experiments. However, such an effective mixture proposal is only possible thanks to learning the multiple inference artifacts we suggest as part of the

AMCI framework, while its performance was still generally inferior to AMCI itself.

We now consider the question of when we expect AMCI to work particularly well compared to SNIS, and the scenarios where it is less beneficial, or potentially even harmful. We first note that scaling with increasing dimensionality is a challenge for both because the importance sampling upon which they rely suffers from the curse of dimensionality. However, the scaling of AMCI should be no worse than existing amortization approaches as each of the amortized proposals is trained in isolation and corresponds to a conventional inference amortization.

We can gain more insights into the relative performance of the two approaches in different settings using an informal asymptotic analysis in the limit of a large number of samples. Assuming  $f(x; \theta) \geq 0 \forall x, \theta$  for simplicity,<sup>3</sup> then both AMCI and SNIS can be expressed in the form of (10), where for SNIS we set  $q_1(x; y, \theta) = q_2(x; y)$ ,  $N = M$ , and share samples between the estimators. Separately applying the central limit theorem to  $\hat{E}_1$  and  $\hat{E}_2$  yields

$$\hat{\mu}(y, \theta) = \frac{\hat{E}_1}{\hat{E}_2} \rightarrow \frac{E_1 + \sigma_1 \xi_1}{E_2 + \sigma_2 \xi_2}, \quad \text{as } N, M \rightarrow \infty \quad (26)$$

where  $\xi_1, \xi_2 \sim \mathcal{N}(0, 1)$  and

$$\sigma_1 := \frac{1}{N} \text{Var}_{q_1(x; y, \theta)} \left[ \frac{f(x; \theta)p(x, y)}{q_1(x; y, \theta)} \right], \quad (27)$$

$$\sigma_2 := \frac{1}{M} \text{Var}_{q_2(x; y)} \left[ \frac{p(x, y)}{q_2(x; y)} \right]. \quad (28)$$

Asymptotically, the mean squared error of  $\hat{\mu}(y, \theta)$  is dominated by its variance. Thus, by taking a first order Taylor expansion of  $\text{Var}[\hat{\mu}(y, \theta)]$  about  $1/E_2$ , we get, for large  $M$ ,

$$\begin{aligned} & \mathbb{E} \left[ (\hat{\mu}(y, \theta) - \mu(y, \theta))^2 \right] \\ & \approx \frac{1}{E_2^2} \left( \sigma_1^2 + \sigma_2^2 \mu(y, \theta)^2 - 2\mu(y, \theta) \sigma_1 \sigma_2 \text{Corr}[\xi_1, \xi_2] \right) \\ & = \frac{\sigma_2^2}{E_2^2} \left( (\kappa - \text{Corr}[\xi_1, \xi_2])^2 + 1 - \text{Corr}[\xi_1, \xi_2]^2 \right) \quad (29) \end{aligned}$$

where the approximation from the Taylor expansion becomes exact in the limit  $M \rightarrow \infty$  and  $\kappa := \sigma_1 / (\mu(y, \theta) \sigma_2)$  is a measure of the relative accuracy of the two estimators. See (43) in Appendix D.1 for a more verbose derivation.

For a given value of  $\sigma_2$ , the value of  $\kappa$  for SNIS is completely dictated by the problem: in general, the larger the mismatch between  $f(x; \theta)p(x, y)$  and  $p(x, y)$ , the larger  $\kappa$  will be. This yields the expected result that the errors for SNIS become large in this setting. For AMCI, we can control  $\kappa$  through ensuring a good proposal for both  $\hat{E}_1$  and  $\hat{E}_2$ , and, if desired, by adjusting  $M$  and  $N$  (relative to a

<sup>3</sup>The results trivially generalize to general  $f(x)$  with suitable adjustment of the definition of  $\sigma_1$ .

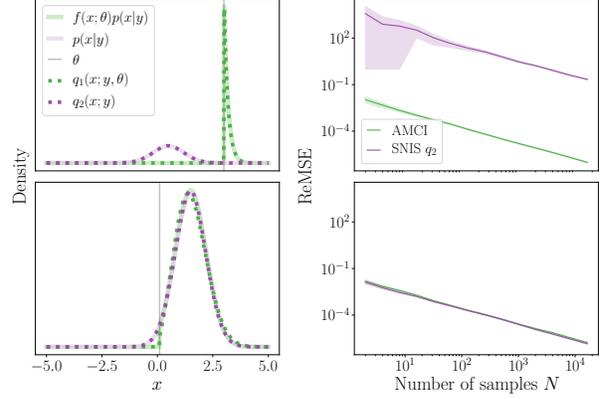


Figure 3: Results for the one-dimensional tail integral model in a setting with large mismatch [top] and low mismatch [bottom], with  $(y, \theta)$ , respectively  $(1, 3)$  and  $(3, 0.1)$ . The left column illustrates the shape of the proposal  $q_1$  and the achievable quality of fit to  $f(x; \theta)p(x|y)$ , we see that AMCI is able to learn very accurate proposals in both cases. The right column compares the performance of the AMCI and the SNIS estimators where we see that the gain for AMCI is much larger when the mismatch is large. Uncertainty bands in column two are estimated over a 1000 runs and are almost imperceptibly small.

fixed budget  $M + N$ ). Consequently, we can achieve better errors than SNIS by driving  $\kappa$  down.

On the other hand, as  $f(x; \theta)p(x, y)$  and  $p(x, y)$  become increasingly well matched, then  $\kappa \rightarrow 1$  and we find that AMCI has little to gain over SNIS. In fact, we see that AMCI can potentially be worse than SNIS in this setting: when  $f(x; \theta)p(x, y)$  and  $p(x, y)$  are closely matched, we also have  $\text{Corr}[\xi_1, \xi_2]^2 \approx 1$  for SNIS, such that we observe a canceling effect, potentially leading to very low errors. Achieving  $\text{Corr}[\xi_1, \xi_2]^2 \approx 1$  can be more difficult for AMCI, potentially giving rise to a higher error. However, it could be possible to mitigate this by correlating the estimates, e.g. through common random numbers.

To assess if this theory manifests in practice, we revisit our tail integral example, comparing large and small mismatch scenarios. The results, shown in Figure 3, agree with these theoretical findings. In Appendix D we further showing that the reusing of samples for both  $\hat{E}_1$  and  $\hat{E}_2$  in AMCI can be beneficial when the targets are well matched.

More generally, as Theorem 1 tells us that the AMCI estimator can achieve an arbitrarily low error for any given target function, while SNIS cannot, we know that its potential gains are larger the more accurate we are able to make our proposals. As such, as advances elsewhere in the field allow us to produce increasingly effective amortized proposals, e.g. through advanced normalizing flow approaches (Grathwohl et al., 2019; Kingma & Dhariwal, 2018), the larger the potential gains are from using AMCI.

## Acknowledgments

We would like to thank Yee Whye Teh for providing helpful discussions at the early stages of the project. AG is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems. FW is supported by DARPA D3M, under Cooperative Agreement FA8750-17-2-0093, Intel under its LBNL NERSC Big Data Center, and an NSERC Discovery grant. TR is supported by the European Research Council under the European Unions Seventh Framework Programme (FP7/20072013) / ERC grant agreement no. 617071. His research leading to these results also received funding from EPSRC under grant EP/P026753/1.

## References

- Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. Adaptive importance sampling: the past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.
- Chen, M.-H. and Shao, Q.-M. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 08 1997.
- Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.
- Enderling, H. and Chaplain, M. A. Mathematical modeling of tumor growth and treatment. *Current pharmaceutical design*, 20–30:4934–40, 2014.
- Evans, M. and Swartz, T. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical science*, pp. 254–272, 1995.
- Gelman, A. and Meng, X.-L. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 05 1998.
- Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. FFJORD: free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations (ICLR)*, 2019.
- Hahnfeldt, P., Panigrahy, D., Folkman, J., and Hlatky, L. Tumor development under angiogenic signaling. *Cancer Research*, 59(19):4770–4775, 1999.
- Hesterberg, T. C. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 2013.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.
- Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems (NIPS)*, 2018.
- Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.
- Lacoste-Julien, S., Huszár, F., and Ghahramani, Z. Approximate inference for the loss-calibrated Bayesian. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.
- Lamberti, R., Petetin, Y., Septier, F., and Desbouvries, F. A double proposal normalized importance sampling estimator. *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 238–242, 2018.
- Le, T. A., Baydin, A. G., and Wood, F. Inference compilation and universal probabilistic programming. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- Le, T. A., Igl, M., Jin, T., Rainforth, T., and Wood, F. Auto-encoding sequential Monte Carlo. *International Conference on Learning Representations (ICLR)*, 2018a.
- Le, T. A., Kosiorek, A. R., Siddharth, N., Teh, Y. W., and Wood, F. Revisiting reweighted wake-sleep. *arXiv:1805.10469*, 2018b.
- Meng, X.-L. and Wong, W. H. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.
- Oh, M.-S. and Berger, J. O. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.
- Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.
- Paige, B. and Wood, F. Inference networks for sequential Monte Carlo in graphical models. *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.
- Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

- Rainforth, T. *Automating inference, learning, and design using probabilistic programming*. PhD thesis, 2017.
- Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. On Nesting Monte Carlo Estimators. *Proceedings of the International Conference on Machine Learning (ICML)*, 2018a.
- Rainforth, T., Zhou, Y., Lu, X., Teh, Y. W., Wood, F., Yang, H., and van de Meent, J.-W. Inference trees: Adaptive inference with exploration. *arXiv preprint arXiv:1806.09550*, 2018b.
- Rezende, D. and Mohamed, S. Variational inference with normalizing flows. *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.
- Ritchie, D., Horsfall, P., and Goodman, N. D. Deep amortized inference for probabilistic programs. *arXiv:1610.05735*, 2016.
- Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.
- Robert, C. and Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- Stuhlmüller, A., Taylor, J., and Goodman, N. Learning stochastic inverses. *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- Veach, E. and Guibas, L. J. Optimally combining sampling techniques for Monte Carlo rendering. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 419–428, 1995.
- Webb, S., Goliński, A., Zinkov, R., Siddharth, N., Rainforth, T., Teh, Y. W., and Wood, F. Faithful inversion of generative models for effective amortized inference. *Advances in Neural Information Processing Systems (NIPS)*, 2018.
- Wolpert, R. L. Monte Carlo integration in Bayesian statistical analysis. *Contemporary Mathematics*, 115:101–116, 1991.

# Appendices for Amortized Monte Carlo Integration

Adam Goliński\*   Frank Wood   Tom Rainforth\*

## A. Additional Experimental Results

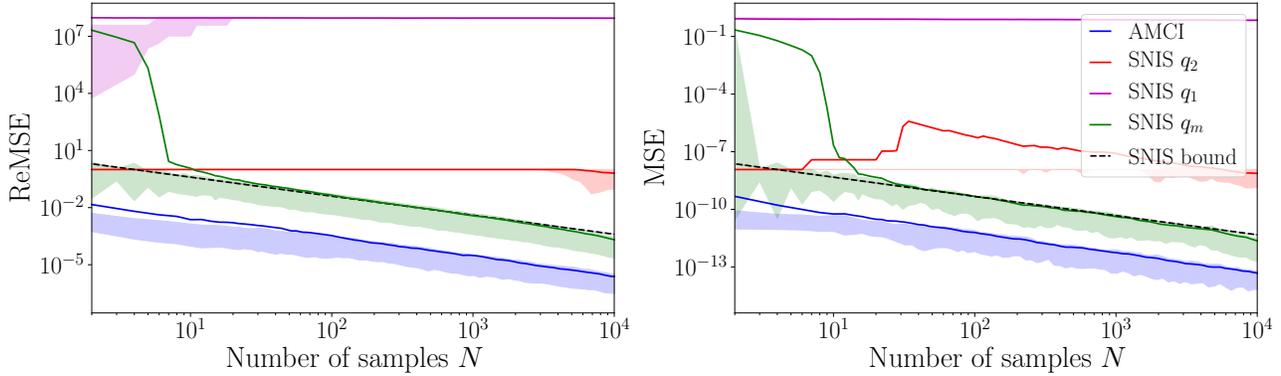


Figure 4: Additional results for one-dimensional tail integral example as per Figure 1a. [left] Relative mean squared errors (as per (25)). [right] Mean squared error  $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta))^2]$ . Conventions as per Figure 1. The results for SNIS  $q_1$  indicate that it severely underestimates  $E_2$  leading to very large errors, especially when the mismatch between  $p(x|y)$  and  $f(x; \theta)$  is as significant as in the tail integral case.

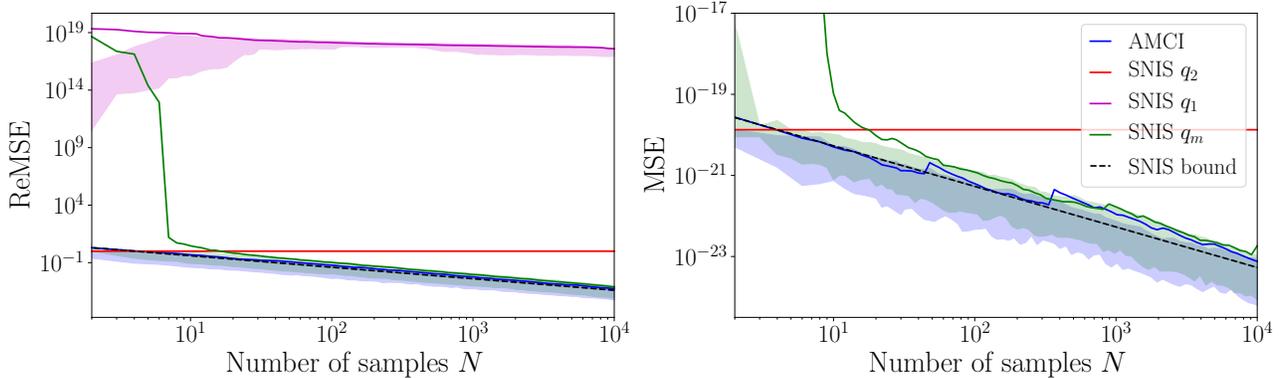


Figure 5: Additional results for five-dimensional tail integral example as per Figure 1b. [left] Relative mean squared errors (as per (25)). [right] Mean squared error  $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta))^2]$ . Conventions as per Figure 1. The y-axis limits for the MSE have been readjusted to allow clear comparison at higher  $N$ . Note that the SNIS  $q_m$  yields MSE of  $10^{-1}$  at  $N=2$ , while the SNIS  $q_1$  MSE is far away from the range of the plot for all  $N$ , giving a MSE of  $10^{-0.9}$  at  $N=2$  and  $10^{-1.2}$  at  $N=10^4$ , with a shape very similar to the ReMSE for SNIS  $q_1$  as per the left plot. The extremely high errors for SNIS  $q_m$  at low values of  $N$  arise in the situation when all  $N$  samples drawn happen to come from distribution  $q_1$ . We believe that the results presented for  $q_m$  underestimate the value of  $\delta(y, \theta)$  between around  $N=6$  and  $N=100$ , due to the fact that the estimation process for  $\delta(y, \theta)$ , though unbiased, can have a very large skew. For  $N \leq 6$  there is a good chance of at least one of the 100 trials we perform having all  $N$  samples originating from distribution  $q_1$ , such that we generate reasonable estimates for the very high errors this can induce. For  $N \geq 100$  the chances of this event occurring drop to below  $10^{-30}$ , such that it does not substantially influence the true error. For  $6 \leq N \leq 100$ , the chance the event will occur in our 100 trials is small, but the influence it has on the overall error is still significantly, meaning it is likely we will underestimate the error. This effect could be alleviated by Rao-Blackwellizing the choice of the mixture component, but this would induce a stratified sampling estimate, thereby moving beyond the SNIS framework.

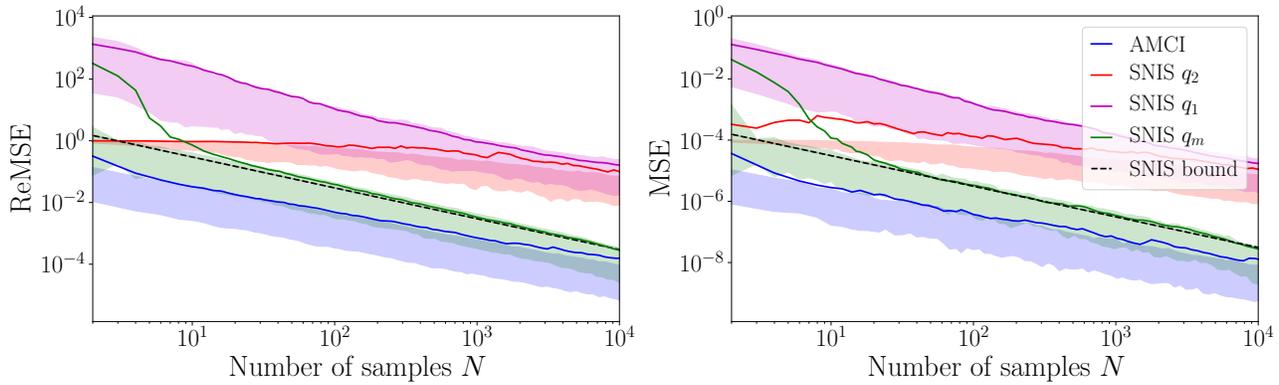


Figure 6: Additional results for cancer example as per Figure 2. [left] Relative mean squared errors (as per (25)). [right] Mean squared error  $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta))^2]$ . Conventions as per Figure 1. Here, the SNIS  $q_1$  performs much better than in the tail integral example because of smaller mismatch between  $p(x|y)$  and  $f(x; \theta)$ , meaning the estimates for  $E_2$  are more reasonable. Nonetheless, we see that SNIS  $q_1$  still performs worse than even SNIS  $q_2$ .

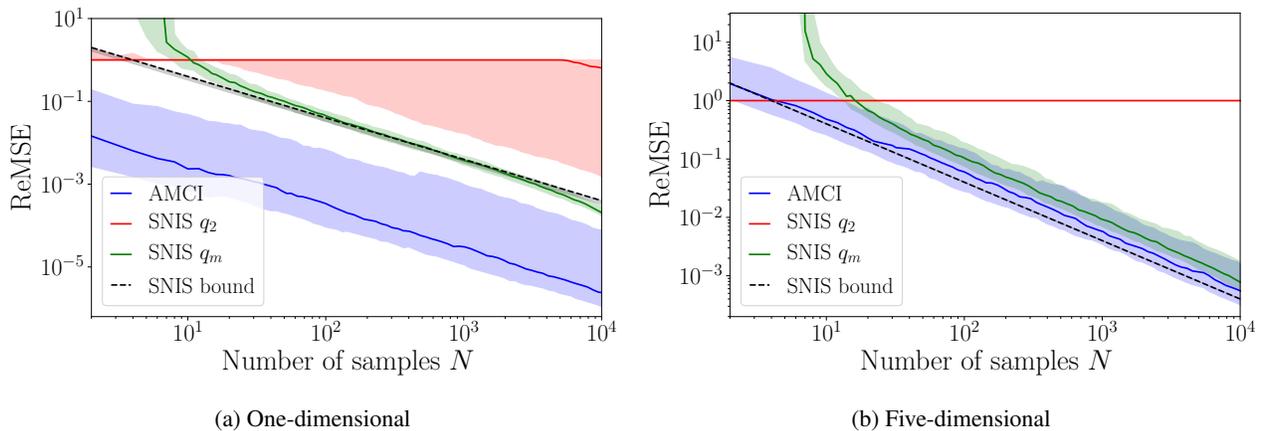


Figure 7: Investigation of the variability of the results across datapoints  $y, \theta$  for [left] the one-dimensional and [right] the five-dimensional tail integral example. Unlike previous figures, the shading shows the estimates of the 25% and 75% quantiles of  $\delta(y, \theta)$  estimated using a common set of 100 samples from  $y, \theta \sim p(y)p(\theta)$ , with the corresponding  $\delta(y, \theta)$  then each separately estimated using 100 samples of the respective  $\hat{\delta}(y, \theta)$ . The solid lines for each estimator and the dashed line remain the same as in previous figures – they indicate the median of  $\delta(y, \theta)$ . Now the dashed line also has a shaded area associated with it reflecting the variability in the SNIS bound across datapoints.

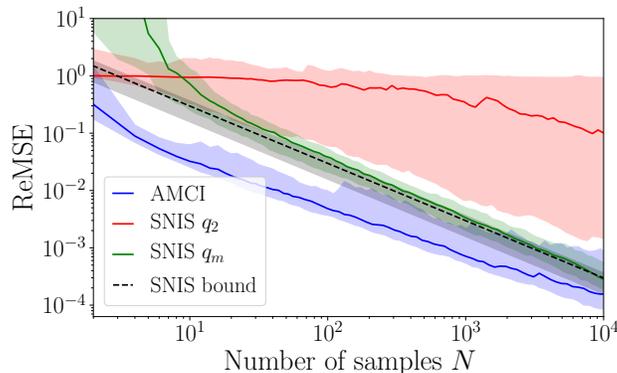


Figure 8: Investigation of the variability of the results across datapoints  $y, \theta$  for cancer example. Conventions as per Figure 7. The fact that the upper quantile of the AMCI error is larger than the upper quantile of the SNIS  $q_m$  error suggests that there are datapoints for which AMCI yields higher mean squared error than SNIS  $q_m$ . However, AMCI is still always better than the standard baseline, i.e. SNIS  $q_2$ .

## B. Proof of Theorem 1

**Theorem 1.** *If the following hold for a given  $\theta$  and  $y$ ,*

$$\mathbb{E}_{p(x)} [f^+(x; \theta)p(y|x)] < \infty \quad (15)$$

$$\mathbb{E}_{p(x)} [f^-(x; \theta)p(y|x)] < \infty \quad (16)$$

$$\mathbb{E}_{p(x)} [p(y|x)] < \infty \quad (17)$$

*and we use the corresponding set of optimal proposals  $q_1^+(x; y, \theta) \propto f^+(x; \theta)p(x, y)$ ,  $q_1^-(x; y, \theta) \propto f^-(x; \theta)p(x, y)$ , and  $q_2(x; y) \propto p(x, y)$ , then the AMCI estimator defined in (14) satisfies*

$$\mathbb{E} [\hat{\mu}(y, \theta)] = \mu(y, \theta), \quad \text{Var} [\hat{\mu}(y, \theta)] = 0 \quad (18)$$

*for any  $N \geq 1$ ,  $K \geq 1$ , and  $M \geq 1$ , such that it forms an exact estimator for that  $\theta, y$  pair.*

*Proof.* The result follows straightforwardly from considering each estimator in isolation. Note that the normalization constants for distributions  $q_1^+, q_1^-, q_2$  are  $E_1^+, E_1^-, E_2$ , respectively, e.g.  $\int f^+(x^+; \theta)p(x^+, y) dx^+ = E_1^+$ . Therefore, starting with  $\hat{E}_2$ , we have

$$\hat{E}_2 = \frac{1}{M} \sum_{m=1}^M \frac{p(x_m, y)}{q_2(x_m; y)} = \frac{1}{M} \sum_{m=1}^M \frac{p(x_m, y)}{p(x_m, y)/E_2} = E_2 \quad (30)$$

for all possible values of  $x_m$ . Similarly, for  $\hat{E}_1^+$

$$\hat{E}_1^+ = \frac{1}{N} \sum_{n=1}^N \frac{p(x_n^+, y)f^+(x_n^+; \theta)}{q_1(x_n^+; y, \theta)} = \frac{1}{N} \sum_{n=1}^N \frac{p(x_n^+, y)f^+(x_n^+; \theta)}{p(x_n^+, y)f^+(x_n^+; \theta)/E_1^+} = E_1^+ \quad (31)$$

for all possible values of  $x_n^+$ . Analogously, we have  $\hat{E}_1^- = E_1^-$  for all possible values of  $x_k^-$ . Combining all of the above, the result now follows.  $\square$

## C. Experimental details

### C.1. One-dimensional tail integral

Let us recall the model from (24),

$$p(x) = \mathcal{N}(x; 0, \Sigma_1) \quad p(y|x) = \mathcal{N}(y; x, \Sigma_2) \quad f(x; \theta) = \prod_{i=1}^D \mathbb{1}_{x_i > \theta_i} \quad p(\theta) = \text{UNIFORM}(\theta; [0, u_D]^D)$$

where for the one-dimensional example  $D = 1$  we used  $u_1 = 5$  and  $\Sigma_1 = \Sigma_2 = 1$ .

For our parameterized proposals  $q_1(x; y, \theta)$  and  $q_2(x; y)$  we used a normalizing flow consisting of 10 radial flow layers (Rezende & Mohamed, 2015) with a standard normal base distribution. The parameters of each flow were determined by a neural network taking in the values of  $y$  and  $\theta$  as input, and returning the parameters defining the flow transformations. Each network comprised of 3 fully connected layers with 1000 hidden units each layer, with relu activation functions.

Training was done by using importance sampling to generate the values of  $\theta$  and  $x$  as per (22) with

$$q'(\theta, x) = p(\theta) \cdot \text{HALFNORMAL}(x; \mu = \theta, \sigma = \Sigma_2).$$

and a learning rate of  $10^{-2}$  with the Adam optimizer Kingma & Ba (2015).

The ground truth values of  $\mu(y, \theta)$  were determined analytically using  $\mu(y, \theta) = \mathbb{E}_{p(x|y)} [f(x; \theta)] = 1 - \Phi(\theta)$ , where  $\Phi(\cdot)$  is the standard normal cumulative distribution function.

### C.2. Five-dimensional tail integral

In the context of the model definition in (24), for the five-dimensional example we used  $u_5 = 3$ ,  $\Sigma_2 = I$  and

$$\Sigma_1 = \begin{bmatrix} 1.2449 & 0.2068 & 0.1635 & 0.1148 & 0.0604 \\ 0.2068 & 1.2087 & 0.1650 & 0.1158 & 0.0609 \\ 0.1635 & 0.1650 & 1.1665 & 0.1169 & 0.0615 \\ 0.1148 & 0.1158 & 0.1169 & 1.1179 & 0.0620 \\ 0.0604 & 0.0609 & 0.0615 & 0.0620 & 1.0625 \end{bmatrix}.$$

In this case, we used a conditional masked autoregressive flow (MAF) (Papamakarios et al., 2017) with standard normal base distribution as the parameterization of our proposals  $q_1(x; y, \theta)$  and  $q_2(x; y)$ . Here the normalizing flows consisted of 16 flow layers with single 1024 hidden units layer within each flow and we used tanh rather than relu activation functions as we found this made a significant difference in terms of training stability for the distribution  $q_1$ . We did not find batch normalization to help the performance or stability significantly, and hence we have not used it. We used the conditional MAF implementation from <http://github.com/ikostrikov/pytorch-flows>.

Training was done using importance sampling to generate the values of  $\theta$  and  $x$  as per (22) with

$$q'(\theta, x) = p(\theta) \cdot \text{HALFNORMAL}(x; \mu = \theta, \sigma = \text{diag}(\Sigma_2)).$$

We used a learning rate of  $10^{-4}$  and the Adam optimizer.

The estimates of the ground truth values  $\mu(y, \theta)$  were determined numerically using an SNIS estimator with  $10^{10}$  samples and the proposal  $q(x; \theta) = \text{HALFNORMAL}(x; \mu = \theta, \sigma = \text{diag}(\Sigma_2))$ .

### C.3. Planning Cancer Treatment

As explained in the main paper, this experiment revolves around an oncologist is trying to decide whether to administer a treatment to a cancer patient. They have access to two noisy measurements of the tumor size, a simulator of tumor evolution, a model of the latent factors required for this simulator, and a loss function for administering the treatment given the final tumor size. We note that this is problem for which the target function  $f(x)$  does not have any changeable parameters (i.e.  $\theta = \emptyset$ ).

The size of the tumor is measured at the time of admission  $t=0$  and five days later ( $t=5$ ), yielding observations  $c'_0$  and  $c'_5$ . These are noisy measurements of the true sizes  $c_0$  and  $c_5$ . The loss function  $\ell(c_{100})$  is based only on the size of the tumor after  $t=100$  days of treatment. The simulator for the development of the tumor takes the form of an ordinary differential equation (ODE) and is taken from (Hahnfeldt et al., 1999; Enderling & Chaplain, 2014; Rainforth et al., 2018a).

The ODE itself is defined on two variables, the size of the tumor at time  $t$ ,  $c_t$ , and corresponding carrying capacity,  $K_t$ , where we take  $K_0=700$ . In addition to the initial tumor size  $c_0$ , the key parameter of the ODE, and the only one we model as varying across patients, is  $\epsilon \in [0, 1]$ , a coefficient determining the patient’s response to the anti-tumor treatment. The ODE now take the form

$$\frac{dc}{dt} = -\lambda c \log\left(\frac{c}{K}\right) - \epsilon c \quad \frac{dK}{dt} = \phi c - \psi K c^{2/3} \quad (32)$$

where the values of the parameters  $\phi=5.85$ ,  $\psi=0.00873$ ,  $\lambda=0.1923$  are based on those recommended in Hahnfeldt et al. (1999). We use the notation

$$c_t = \omega(K_0, c_0, \epsilon, t) \quad (33)$$

to denote the deterministic process of running an ODE solver on (32) with given inputs, up to time  $t$ , and assume the following statistical model

$$\begin{aligned} c_0 &\sim \text{GAMMA}(k = 25, \theta = 20) \\ \epsilon &\sim \text{BETA}(\alpha = 5.0, \beta = 10.0) \\ c'_t &\sim \text{GAMMA}\left(k = \frac{c_t^2}{10000}, \theta = \frac{c_t}{10000}\right). \end{aligned}$$

To summarize and relate the model to the notation from Section 3:  $x = \{c_0, \epsilon\}$ ,  $y = \{c'_0, c'_1\}$ . The function in this case is fixed to the loss function for administering the treatment given the final tumor size provided to us by the clinic

$$f(x) = \ell(\omega(700, c_0, \epsilon, t = 100)) \quad (34)$$

$$\ell(c) = \frac{1 - 2 \times 10^{-8}}{2} \left( \tanh\left(-\frac{c-300}{150}\right) + 1 \right) + 10^{-8}. \quad (35)$$

**Amortization** In this case, the amortization is performed using parametric distributions as proposals: a Gamma distribution for  $c_0$  and a Beta distribution for  $\epsilon$ , both parameterized by a multilayer perceptron with 16 layers with 5000 hidden units each. Since we do not face an overwhelming mismatch between  $f(x)$  and  $p(x)$ , unlike in the tail integral example,

the training was done by generating the values of  $x$  from the prior  $p(x)$  as per (21). We used a learning rate of  $10^{-4}$  with the Adam optimizer.

Similarly to the case of five-dimensional tail integral example, the estimates serving as ground truth values  $\mu(y)$  have been determined numerically using an SNIS estimator with  $10^9$  samples and the proposal set to the prior  $q(x) = p(x)$ .

#### C.4. Mini-batching Procedure

AMCI operates in a slightly unusual setting for neural network training because instead of having a fixed dataset, we are instead training on samples from our model  $p(x, y)$ . The typical way to perform batch stochastic gradient optimization involves many epochs over the training dataset, stopping once the error increases on the validation set. Each epoch is itself broken down into multiple iterations, wherein one takes a random mini-batch (subsample) from the dataset (without replacement) and updates the parameters based on a stochastic gradient step using these samples, with the epoch finishing once the full dataset has been used.

However, there are different ways the training can proceed when we have the ability to generate an infinite amount of data from our model  $p(x, y)$  and we now no longer have the risk of overfitting. There are two extreme approaches one could take. The first one would be sampling two large but fixed-size datasets (training and validation) before the time of training and then following the standard training procedure for the finite datasets outlined above. The other extreme would be to completely surrender the idea of dataset or epoch, and sample each batch of data presented to the optimizer directly from  $p(x, y)$ . In this case, we would not need a validation dataset as we would never be at risk of overfitting—we would finish the training once we are satisfied with the convergence of the loss value.

Paige & Wood (2016) found that the method which empirically performed best in similar amortized inference setting was one in the middle between the two extremes outlined above. They suggest a method which decides when to sample new synthetic (training and validation) datasets, based on performance on the validation data set. They draw fixed-sized training and validation datasets and optimize the model using the standard finite data procedure on the training dataset until the validation error increases. When that happens they sample new training and validation datasets and repeat the procedure. This continues until empirical convergence of the loss value. In practice, they allow a few missteps (steps of increasing value) for the validation loss before they sample new synthetic datasets, and limit the maximum number of optimization epochs performed on a single dataset.

We use the above method throughout all of our experiments. We allowed a maximum of 2 missteps w.r.t. the validation dataset and maximum of 30 epochs on a single dataset before sampling new datasets.

Note that the way training and validation datasets are generated is modified slightly when using the importance sampling approach for generating  $x$  and  $\theta$  detailed in Section 3.3. Whenever we use the objective in (22), instead of sampling the training and validation datasets from the prior  $p(x, y)$  we will sample them from the distribution  $q'(\theta, x) \cdot p(y|x)$  where  $q'$  is a proposal chosen to be as close to  $p(x)p(\theta)f(x; \theta)$  as possible.

We note that while training was robust to the number of missteps allowed, adopting the general scheme of Paige & Wood (2016) was very important in achieving effective training: we initially tried generating every batch directly from the model  $p(x, y)$  and we found that the proposals often converged to the local minimum of just sampling from the prior.

#### D. Reusing samples

The AMCI estimator in (14) requires taking  $T = N + K + M$  samples, but only  $N$ ,  $K$ , or  $M$  are used to evaluate each of the individual estimators. Given that, in practice, we do not have access to the perfectly optimal proposals, it can sometimes be more efficient to reuse samples in the calculation of multiple components of the expectation, particularly if the target function is cheap to evaluate relative to the proposal. Care is required though to ensure that this is only done when a proposal remains valid (i.e. has finite variance) for the different expectation.

To give a concrete example, in the case where  $f(x; \theta) \geq 0 \forall x, \theta$ , such that we can use a single proposal for the numerator as per (10), we could use the following estimator

$$\mu(y, \theta) \approx \frac{\alpha \hat{E}_1(q_1) + (1 - \alpha) \hat{E}_1(q_2)}{\beta \hat{E}_2(q_1) + (1 - \beta) \hat{E}_2(q_2)} \quad (36)$$

where  $\hat{E}_i(q_j)$  indicates the estimate for  $E_i$  using the samples from  $q_j$ . The level of interpolation is set by parameters  $\alpha, \beta$

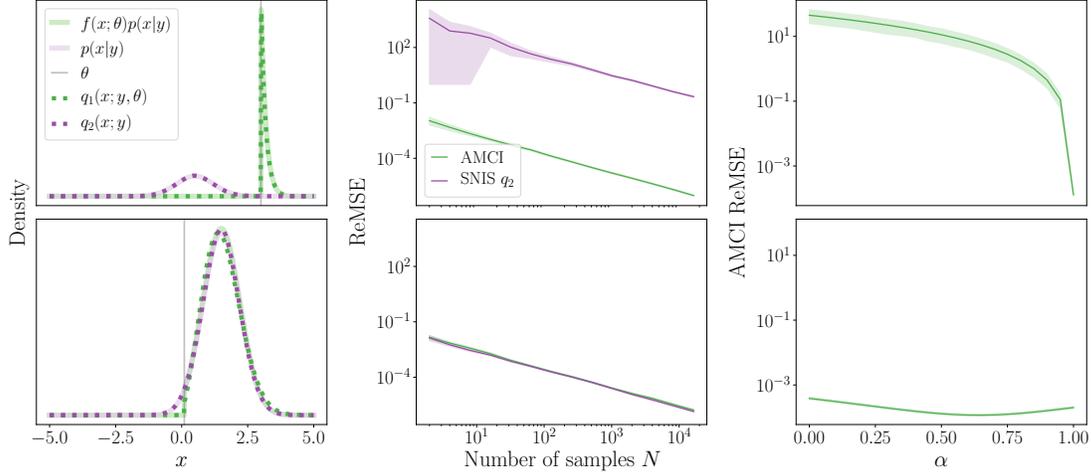


Figure 9: Extension of Figure 3. Column three presents the effects of reusing samples by varying the parameter  $\alpha$  in (36) ( $\beta = 0$ , number of samples is fixed to  $N = M = 64$ ), where we see that this sample re-usage provides small gains for the low mismatch case, but no gains in the high mismatch case. Uncertainty bands in columns two and three are estimated over a 1000 runs and are very small.

which vary between 0 and 1. If we had direct access to the optimal proposals, it would naturally be preferable to set  $\alpha = 1$  and  $\beta = 0$ , leading to a zero-variance estimator. However, for imperfect proposals, the optimal values vary slightly from this (see Appendix D.1).

In relation to our discussion in Section 5, the third column of Figure 9 shows how when  $f(x; \theta)p(x, y)$  and  $p(x, y)$  are closely matched we can decrease the error of our AMCI estimator by reusing samples through setting  $\alpha < 1$ .

Note that while it is possible to set  $\beta > 0$  for negligible extra computational cost as  $\hat{E}_2(q_1)$  depends only on weights needed for calculating  $\hat{E}_1(q_1)$ , setting  $\alpha < 1$  requires additional evaluations of the target function and so will likely only be beneficial when this is cheap relative to sampling from or evaluating the proposal.

#### D.1. Derivation of the optimal parameter values for $\alpha$ and $\beta$

In this section, we derive the optimal values of  $\alpha$  and  $\beta$  in terms of minimizing the mean squared error (MSE) of the estimator in (36). We assume that we are allocated a total sample budget of  $T$  samples, such that  $M = T - N$ .

Let the true values of the expectations in the numerator and denominator be denoted as  $E_1$  and  $E_2$ , respectively. We also define the following shorthands for the unbiased importance sampling estimators with respect to proposals  $q_1$  and  $q_2$  in (36)  $a_1 = \frac{1}{N} \sum_n \frac{f(x_n; \theta)p(x_n, y)}{q_1(x_n; y, \theta)}$ ,  $b_1 = \frac{1}{M} \sum_m \frac{f(x_m^*; \theta)p(x_m^*, y)}{q_2(x_m^*; y)}$ ,  $a_2 = \frac{1}{N} \sum_n \frac{p(x_n, y)}{q_1(x_n; y, \theta)}$ ,  $b_2 = \frac{1}{M} \sum_m \frac{p(x_m^*, y)}{q_2(x_m^*; y)}$ , where  $x_n \sim q_1(x; y, \theta)$  and  $x_m^* \sim q_2(x; y)$ .

We start by considering the estimator according to (36)

$$\mu := \frac{E_1}{E_2} \approx \hat{\mu} := \frac{\hat{E}_1}{\hat{E}_2} := \frac{\alpha a_1 + (1 - \alpha)b_1}{\beta a_2 + (1 - \beta)b_2}. \quad (37)$$

Using the central limit theorem separately for  $\hat{E}_1$  and  $\hat{E}_2$ , then we thus have, as  $N, M \rightarrow \infty$ ,

$$\hat{\mu} \rightarrow \frac{E_1 + \sigma_1 \xi_1}{E_2 + \sigma_2 \xi_2}, \quad (38)$$

where  $\xi_1, \xi_2 \sim \mathcal{N}(0, 1)$  are correlated standard normal random variables and  $\sigma_1$  and  $\sigma_2$  are the standard deviation of the estimators for the numerator and the denominator, respectively. Specifically we have

$$\begin{aligned} \sigma_1^2 &= \text{Var}[\alpha a_1 + (1 - \alpha)b_1] \\ &= \alpha^2 \text{Var}_{q_1}[a_1] + (1 - \alpha)^2 \text{Var}_{q_2}[b_1], \end{aligned}$$

which by the weak law of large numbers

$$= \frac{\alpha^2}{N} \text{Var}_{q_1}[f(x_1)w_1] + \frac{(1-\alpha)^2}{M} \text{Var}_{q_2}[f(x_1^*)w_1^*] \quad (39)$$

where  $w_1 = p(x_1, y)/q_1(x_1; y, \theta)$ ,  $w_1^* = p(x_1^*, y)/q_2(x_1^*; y)$ ,  $x_1 \sim q_1(x; y, \theta)$ , and  $x_1^* \sim q_2(x; y)$ . Analogously,

$$\sigma_2^2 = \frac{\beta^2}{N} \text{Var}_{q_1}[w_1] + \frac{(1-\beta)^2}{M} \text{Var}_{q_2}[w_1^*]. \quad (40)$$

Now going back to (38) and using Taylor's Theorem on  $1/(E_2 + \sigma_2\xi_2)$  about  $1/E_2$  gives

$$\begin{aligned} \hat{\mu} &= \frac{E_1 + \sigma_1\xi_1}{E_2} \left(1 - \frac{\sigma_2\xi_2}{E_2}\right) + O(\epsilon) \\ &= \frac{E_1}{E_2} + \frac{\sigma_1\xi_1}{E_2} - \frac{E_1\sigma_2\xi_2}{E_2^2} - \frac{\sigma_1\sigma_2\xi_1\xi_2}{E_2^2} + O(\epsilon) \end{aligned}$$

where  $O(\epsilon)$  represents asymptotically dominated terms. Note here the importance of using Taylor's theorem, instead of just a Taylor expansion, to confirm that these terms are indeed asymptotically dominated. We can further drop the  $\sigma_1\sigma_2\xi_1\xi_2/E_2^2$  term as this will be of order  $O(1/\sqrt{MN})$  and will thus be asymptotically dominated, giving

$$= \frac{E_1}{E_2} + \frac{\sigma_1\xi_1}{E_2} - \frac{E_1\sigma_2\xi_2}{E_2^2} + O(\epsilon). \quad (41)$$

To calculate the MSE of  $\hat{\mu}$ , we start with the standard bias variance decomposition

$$\mathbb{E} \left[ \left( \hat{\mu} - \frac{E_1}{E_2} \right)^2 \right] = \text{Var}[\hat{\mu}] + \left( \mathbb{E} \left[ \hat{\mu} - \frac{E_1}{E_2} \right] \right)^2. \quad (42)$$

Considering first the bias squared term, we see that this depends only on the higher order terms  $O(\epsilon)$ , while the variance does not. It straightforwardly follows that the variance term will be asymptotically dominant, so we see that optimizing for the variance is asymptotically equivalent to optimizing for the MSE.

Now using the standard relationship  $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$  yields

$$\begin{aligned} \text{Var}[\hat{\mu}] &= \text{Var} \left[ \frac{E_1}{E_2} \right] + \text{Var} \left[ \frac{\sigma_1\xi_1}{E_2} \right] + \text{Var} \left[ \frac{E_1\sigma_2\xi_2}{E_2^2} \right] + 2\text{Cov} \left[ \frac{\sigma_1\xi_1}{E_2}, -\frac{E_1\sigma_2\xi_2}{E_2^2} \right] + O(\epsilon) \\ &\approx 0 + \frac{\sigma_1^2}{E_2^2} + \frac{E_1^2\sigma_2^2}{E_2^4} - 2\frac{E_1\sigma_1\sigma_2}{E_2^3} \text{Cov}[\xi_1, \xi_2] \\ &= \frac{1}{E_2^2} \left( \sigma_1^2 + \sigma_2^2\mu^2 - 2\mu\sigma_1\sigma_2\text{Corr}[\xi_1, \xi_2] \right) \end{aligned} \quad (43)$$

since  $\text{Var}[\xi_1] = \text{Var}[\xi_2] = 1 \implies \text{Cov}[\xi_1, \xi_2] = \text{Corr}[\xi_1, \xi_2]$ ,

$$\begin{aligned} &= \frac{\alpha^2}{NE_2^2} \text{Var}_{q_1}[f(x_1)w_1] + \frac{(1-\alpha)^2}{ME_2^2} \text{Var}_{q_2}[f(x_1^*)w_1^*] + \frac{E_1^2\beta^2}{NE_2^4} \text{Var}_{q_1}[w_1] + \frac{E_1^2(1-\beta)^2}{ME_2^4} \text{Var}_{q_2}[w_1^*] \\ &\quad - 2\frac{E_1}{E_2^3} \text{Corr}[\xi_1, \xi_2] \left( \frac{\alpha^2}{N} \text{Var}_{q_1}[f(x_1)w_1] + \frac{(1-\alpha)^2}{M} \text{Var}_{q_2}[f(x_1^*)w_1^*] \right) \left( \frac{\beta^2}{N} \text{Var}_{q_1}[w_1] + \frac{(1-\beta)^2}{M} \text{Var}_{q_2}[w_1^*] \right) \end{aligned}$$

To assist in the subsequent analysis, we assume that there is no correlation,  $\text{Corr}[\xi_1, \xi_2] = 0$ . Though this assumption is unlikely to be exactly true, there are two reasons we believe it is reasonable. Firstly, because we expect to set  $\alpha \approx 1$  and  $\beta \approx 0$ , the correlation should generally be small in practice as the two estimators rely predominantly on independent sets of samples. Secondly, we believe this is generally a relatively conservative assumption: if one were to presume a particular correlation, there are adversarial cases with the opposite correlation where this assumption is damaging.

Given this assumption it is now straightforward to optimize for  $\alpha$  and  $\beta$  by finding where the gradient is zero as follows

$$\begin{aligned}\nabla_{\alpha}(\text{Var}[\hat{\mu}]E_2^2) &= \frac{2\alpha \text{Var}_{q_1}[f(x_1)w_1]}{N} - \frac{2(1-\alpha)\text{Var}_{q_2}[f(x_1^*)w_1^*]}{T-N} = 0 \\ \Rightarrow \alpha^* &= N \cdot \left( (T-N) \frac{\text{Var}_{q_1}[f(x_1)w_1]}{\text{Var}_{q_2}[f(x_1^*)w_1^*]} + N \right)^{-1}\end{aligned}\quad (44)$$

noting that

$$\nabla_{\alpha}^2(\text{Var}[\hat{\mu}]E_2^2) = \frac{\text{Var}_{q_1}[f(x_1)w_1]}{N} + \frac{\text{Var}_{q_2}[f(x_1^*)w_1^*]}{T-N} > 0$$

and hence it's a local minimum. Analogously

$$\beta^* = N \cdot \left( (T-N) \frac{\text{Var}_{q_1}[w_1]}{\text{Var}_{q_2}[w_1^*]} + N \right)^{-1}.\quad (45)$$

We note that it is possible to estimate all the required variances here using previous samples. It should therefore be possible to adaptively set  $\alpha$  and  $\beta$  by using these equations along with empirical estimates for these variances.