Appendices for Amortized Monte Carlo Integration

Adam Goliński*   Frank Wood   Tom Rainforth*

A. Additional Experimental Results

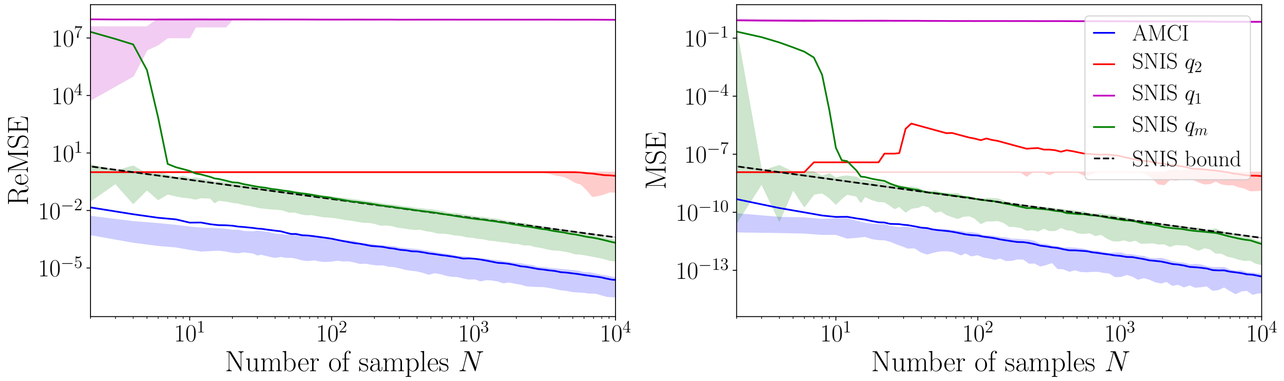


Figure 4: Additional results for one-dimensional tail integral example as per Figure 1a. [left] Relative mean squared errors (as per (25)). [right] Mean squared error $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta)^2]$. Conventions as per Figure 1. The results for SNIS q_1 indicate that it severely 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.

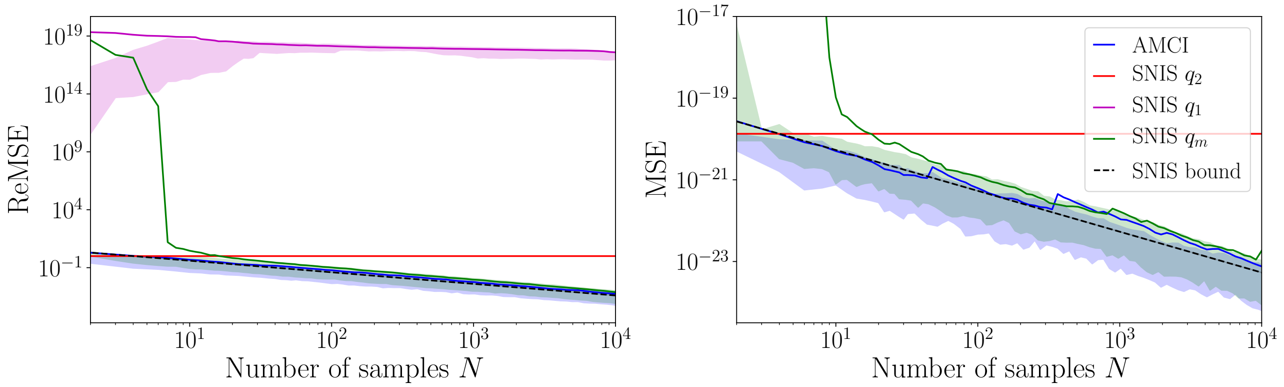


Figure 5: Additional results for five-dimensional tail integral example as per Figure 1b. [left] Relative mean squared errors (as per (25)). [right] Mean squared error $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta)^2]$. Conventions as per Figure 1. The y-axis limits for the MSE have been readjusted 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.

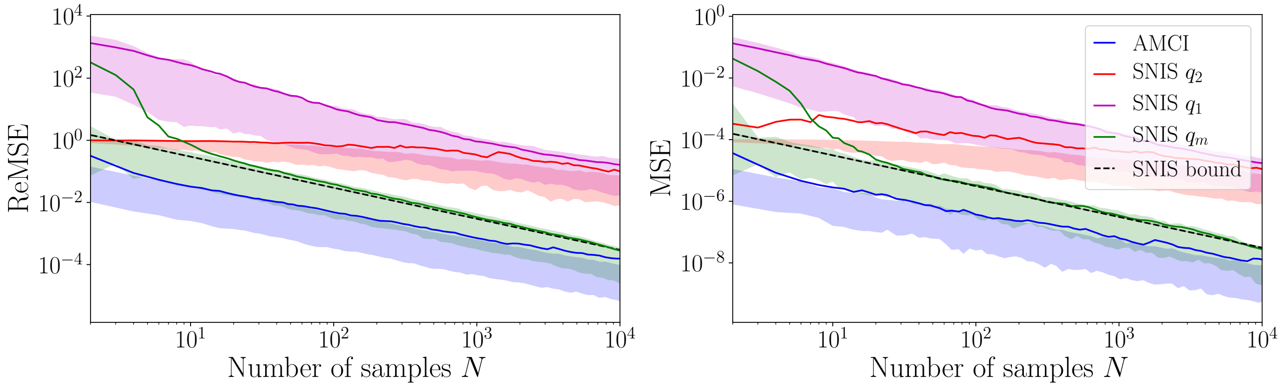
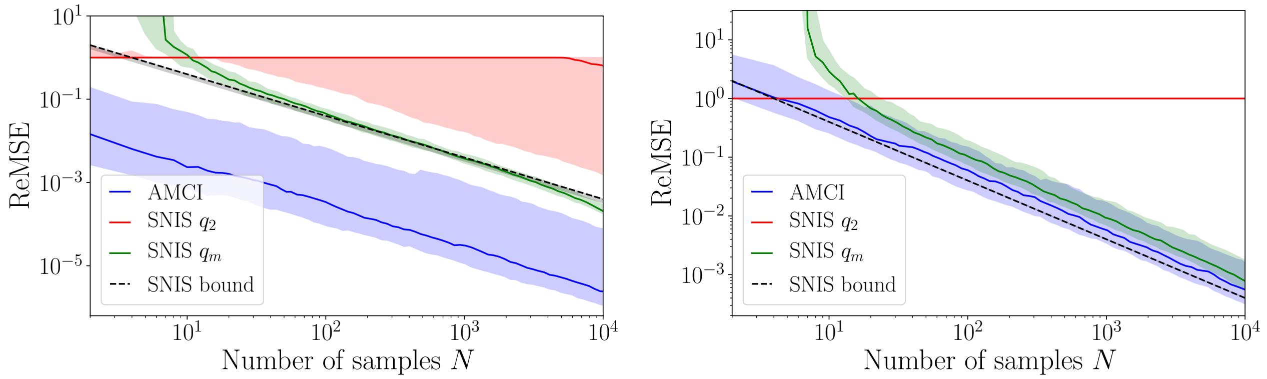


Figure 6: Additional results for cancer example as per Figure 2. [left] Relative mean squared errors (as per (25)). [right] Mean squared error $\mathbb{E}[(\mu(y, \theta) - \hat{\mu}(y, \theta)^2]$. Conventions as per Figure 1. Here, the SNIS q_1 performs much better than in the tail integral example 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 that even SNIS q_2.



(a) One-dimensional

(b) Five-dimensional

Figure 7: Investigation of the variability of the results across datapoints y, θ 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.

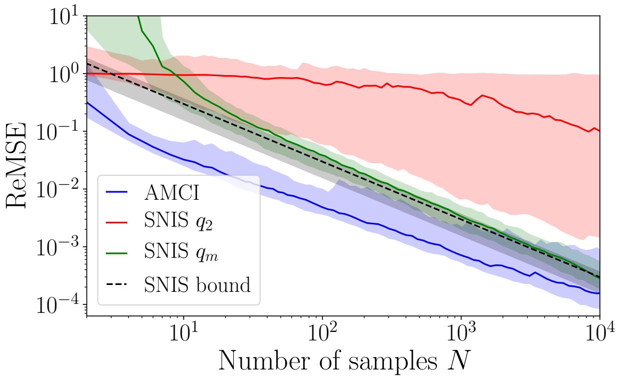


Figure 8: Investigation of the variability of the results across datapoints y, θ for cancer example. Conventions as per Figure 7. The fact that the upper quantile of the AMCI error is larger than the upper quantile of the SNIS q_m error suggests that there are datapoints for which AMCI yields higher 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 θ and y,*

$$\mathbb{E}_{p(x)}\left[f^+(x;\theta)p(y|x)\right] < \infty \tag{15}$$
$$\mathbb{E}_{p(x)}\left[f^-(x;\theta)p(y|x)\right] < \infty \tag{16}$$
$$\mathbb{E}_{p(x)}\left[p(y|x)\right] < \infty \tag{17}$$

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*

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

for any $N \geq 1$, $K \geq 1$, and $M \geq 1$, such that it forms an exact estimator for that θ, 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. $\int f^+(x^+;\theta)p(x^+,y)\,\mathrm{d}x^+ = E_1^+$. Therefore, starting with $\hat{E}_2$, we have

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

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

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

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. □

C. Experimental details

C.1. One-dimensional tail integral

Let us recall the model from (24),

$$p(x) = \mathcal{N}(x;0,\Sigma_1) \qquad p(y|x) = \mathcal{N}(y;x,\Sigma_2) \qquad f(x;\theta) = \prod\nolimits_{i=1}^{D} \mathbb{1}_{x_i>\theta_i} \qquad p(\theta) = \textsc{Uniform}(\theta;[0,u_D]^D)$$

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 θ 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 θ and x as per (22) with

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

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 $\mu(y,\theta) = \mathbb{E}_{p(x|y)}\left[f(x;\theta)\right] = 1-\Phi(\theta)$, 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 θ and x as per (22) with

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

We used a learning rate of 10^{-4} an 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 $q(x; \theta) = \textsc{HalfNormal}(x; \mu = \theta, \sigma = \text{diag}(\Sigma_2))$.

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

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -\lambda c \log\left(\frac{c}{K}\right) - \epsilon c \qquad \frac{\mathrm{d}K}{\mathrm{d}t} = \phi c - \psi K c^{2/3} \tag{32}$$

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

$$c_t = \omega(K_0, c_0, \epsilon, t) \tag{33}$$

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

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

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

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

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

Amortization In this case, the amortization is performed using parametric distributions as proposals: a Gamma distribution for c_0 and a Beta distribution for ϵ, 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 fave the risk of overfitting. There are two extremes 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 θ 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

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

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 α, β

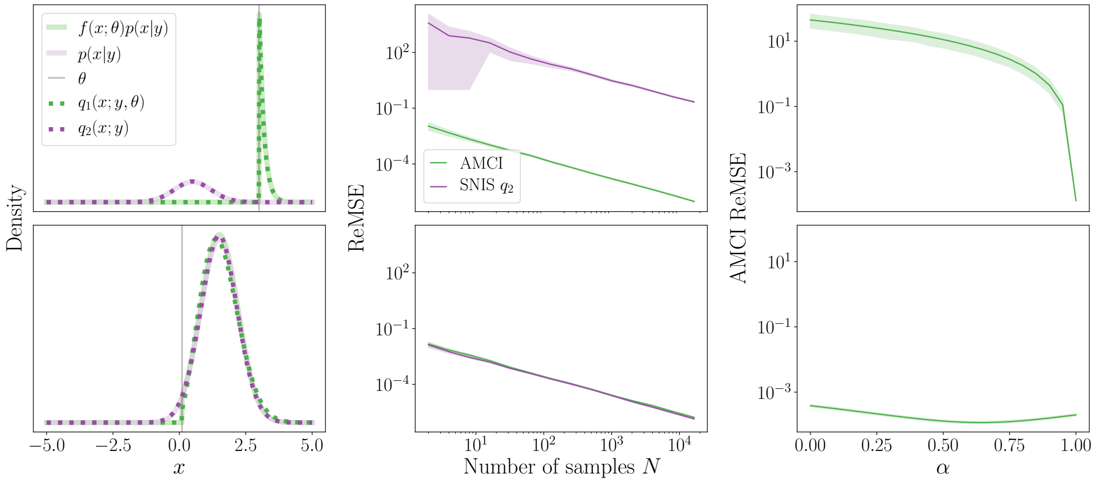


Figure 9: Extension of Figure 3. Column three presents the effects of reusing samples by varying the parameter α in (36) ($\beta = 0$, number of samples is fixed to $N = M = 64$), where we see that this sample re-usage provides small gains for the low mismatch case, but no gains in the high mismatch 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 α and β

In this section, we derive the optimal values of α and β 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) $a_1 = \frac{1}{N}\sum_n^N \frac{f(x_n;\theta)p(x_n,y)}{q_1(x_n;y,\theta)}$, $b_1 = \frac{1}{M}\sum_m^M \frac{f(x^*_m;\theta)p(x^*_m,y)}{q_2(x^*_m;y)}$, $a_2 = \frac{1}{N}\sum_n^N \frac{p(x_n,y)}{q_1(x_n;y,\theta)}$, $b_2 = \frac{1}{M}\sum_m^M \frac{p(x^*_m,y)}{q_2(x^*_m;y)}$, 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)

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

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

$$\hat{\mu} \to \frac{E_1 + \sigma_1\xi_1}{E_2 + \sigma_2\xi_2}, \tag{38}$$

where $\xi_1, \xi_2 \sim \mathcal{N}(0,1)$ are correlated standard normal random variables and σ_1 and σ_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

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

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,

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

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

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

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

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

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 $\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]+2\,\mathrm{Cov}[X,Y]$ yields

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

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

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

To assist in the subsequent analysis, we assume that there is no correlation, 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 α and β by finding where the gradient is zero as follows

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

noting that

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

and hence it's a local minimum. Analogously

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

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

References

Enderling, H. and Chaplain, M. A. Mathematical modeling of tumor growth and treatment. *Current pharmaceutical design*, 20–30:4934–40, 2014.

Hahnfeldt, P., Panigrahy, D., Folkman, J., and Hlatky, L. Tumor development under angiogenic signaling. *Cancer Research*, 59(19):4770–4775, 1999.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

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., Cornish, R., Yang, H., Warrington, A., and Wood, F. On Nesting Monte Carlo Estimators. *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.