1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–11

Amortized Monte Carlo Integration

Adam Goliński[†], **Yee Whye Teh**[†], **Frank Wood**[‡], **Tom Rainforth**[†]
[†] *University of Oxford*
[‡] *University of British Columbia*

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. In this paper, we address the inefficiency of this computational pipeline when the target function(s) are known upfront. To this end, we introduce a method for *amortizing Monte Carlo integration*. Our approach operates similarly to amortized inference but produces two amortization artifacts tailored to maximize the accuracy of the resulting expectation calculation(s). We show that while existing approaches have fundamental limitations in the level of accuracy that can be achieved for a given run time computational budget, our framework can produce arbitrary small errors for a wide range of target functions with $\mathcal{O}(1)$ computational cost at run time. Furthermore, our framework allows not only for amortizing over possible datasets but also over possible target functions.

Keywords: Monte Carlo, Amortized Inference, Integration

1. Motivation and Background

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). The eventual aim of the vast majority of Bayesian inference problems can 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 $p(y)$. Sometimes $f(x)$ is not known up front, or we care about many different $f(x)$, such that is convenient to just approximate $p(x|y)$ upfront, e.g. in the form of Monte Carlo samples, and then later use this to calculate estimates, rather than address 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. It has been well established in the literature 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 $f(x)$ (Owen, 2013; Lacoste-Julien et al., 2011). The potential gains in such situations can be substantial: the approach we introduce will be able to, in theory, produce estimates with arbitrary small mean squared errors from taking only two samples, compared with potentially arbitrarily high errors for methods which do not incorporate information about $f(x)$ (Owen, 2013).

Although it is all too often overlooked, how to adjust for target-aware settings has previously been extensively considered in the fixed-dataset context (Hesterberg, 1988; Wolpert, 1991; Oh and Berger, 1992; Evans and Swartz, 1995; Meng and Wong, 1996; Chen and Shao, 1997; Gelman and Meng, 1998; Lacoste-Julien et al., 2011). In this paper, we extend these ideas to *amortized* inference settings (Stuhlmüller et al., 2013; Kingma and Welling, 2014; Ritchie et al., 2016; Paige and Wood, 2016; Le et al., 2017, 2018; Maddison et al., 2017; Naesseth 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. Typically, this amortization artifact takes the form of a parametrized proposal, $q(x; \varphi(y))$, which takes in data y and regresses these to proposal parameters $\varphi(y)$, generally using a deep neural network. Though the exact process varies with context, the inference network is usually trained either by drawing latent-data sample pairs from the joint $p(x, y)$ (Ritchie et al., 2016; Paige and Wood, 2016; Le et al., 2017), or by drawing subdata from a large dataset using stochastic variational inference approaches (Hoffman et al., 2013; Kingma and Welling, 2014; Rezende et al., 2014). Once trained, it provides an efficient means of approximately sampling from the posterior of a particular dataset, e.g. using self-normalized importance sampling (SNIS).

2. Method

We introduce AMCI, a framework for performing *amortized Monte Carlo integration* which varies from standard amortized inference approaches in three respects. Firstly it operates in a target-aware fashion, incorporating information about $f(x)$ into the amortization artifacts, increasing the efficiency at runtime. Secondly, rather than relying purely on SNIS, AMCI amortizes and employs two separate proposals for estimating the unnormalized target integral $\mathbb{E}_{p(x)}[f(x)p(y|x)]$ and the marginal likelihood $\mathbb{E}_{p(x)}[p(y|x)]$. Such construction allows it, at least in principle, to return estimates with arbitrarily low mean squared error using just two samples when optimal proposals are used for both of the estimators. This is in contrast to standard SNIS, whose attainable variance is lower bounded by $\mathbb{E}_{p(x|y)}[|f(x) - \mathbb{E}_{p(x|y)}[f(x)]|]^2/N$ for a given $f(x)$ and number of samples N and hence so is the MSE of its estimates (Owen, 2013). Finally, to account for cases in which multiple target functions may be of interest, AMCI allows for amortization of parametrized functions $f(x; \theta)$ by extending our target distribution with a pseudo prior $p(\theta)$.

Although AMCI is strongly motivated by Bayesian settings, it can be applied in any Monte Carlo integration setting wherein we wish to calculate $\mathbb{E}_{\pi(x)}[f(x)]$ for some reference distribution $\pi(x)$, known only up to a normalizing constant. Moreover, because any integral $\int_{x\in\mathcal{X}} f(x)dx$ can be expressed as an expectation $\mathbb{E}_{q(x)}[f(x)\pi(x)/q(x)]$ through importance sampling, AMCI allows amortizing integration more generally.

Estimator AMCI seeks to overcome the limitations of SNIS by using a new estimator consisting of two separate, non-self-normalized, importance sampling estimators for $\mathbb{E}_{p(x)}[f(x)p(y|x)]$ and $\mathbb{E}_{p(x)}[p(y|x)]$. This way each estimator can use a separate, individually tuned proposal, allowing an arbitrary reduction in variance compared with SNIS.

What is more, such construction allows us to achieve a zero-variance estimate if optimal proposals are used for each of the estimators. Thus AMCI can provide superior performance, as compared to previous approaches, since in principle it can return single sample estimates

with arbitrarily low error, requiring $\mathcal{O}(1)$ computational cost at runtime. Below we present the method in the case when the function of interest $f^*(x)$ is upper or lower bounded by b and hence we can ensure $\forall x : f(x) \geq 0$ by setting $f(x) = \pm(f^*(x) - b)$. If that is not the case we can use importance sampling positivisation (Owen, 2013) to maintain the ability to obtain a zero variance estimator, see Appendix B for details.

The AMCI estimator for $\mathbb{E}_{p(x|y)}[f(x)]$ is a ratio of convex combinations of estimators w.r.t. the two different proposals:

$$\frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} = \frac{\frac{\alpha}{N}\sum_n^N \frac{f(x_n)p(x_n,y)}{q_1(x_n|y)} + \frac{1-\alpha}{M}\sum_m^M \frac{f(x_m^*)p(x_m^*,y)}{q_2(x_m^*|y)}}{\frac{\beta}{N}\sum_n^N \frac{p(x_n,y)}{q_1(x_n|y)} + \frac{1-\beta}{M}\sum_m^M \frac{p(x_m^*,y)}{q_2(x_m^*|y)}} \tag{1}$$

where different numbers of samples, N and M, are drawn from $x_n \sim q_1(x|y)$, $x_m^* \sim q_2(x|y)$, respectively. The optimal sampling proposal for the non-normalized importance sampling in the expectation in the numerator is $q_1(x|y) \propto |f(x)|p(x|y)$ while for the denominator it is $q_2(x|y) \propto p(x|y)$ (Owen, 2013).

The level of interpolation is set by parameters α, β 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, in practice, our proposals will not be perfect and so using a convex combination of the two estimators allows us to make use of all the $N + M$ samples draw at negligible extra cost. See Appendix D for the derivation of the asymptotically optimal parameter settings of α and β.

Amortization To evaluate this estimator AMCI needs to learn to amortize proposals q_1 and q_2. Following Paige and Wood (2016) our objective for amortizing q_2 is

$$\mathcal{J}_2(\eta) = \mathbb{E}_{p(y)}[D_{KL}\left[p(x|y)\,||\,q_2(x;\varphi_2(y;\eta))\right]] = \mathbb{E}_{p(x,y)}[-\log q_2(x;\varphi_2(y;\eta))] + \text{const wrt } \eta \tag{2}$$

where φ_2 is a neural network with parameters η. This objective requires the ability to sample from $p(x, y)$ and it can be optimized using gradient methods.

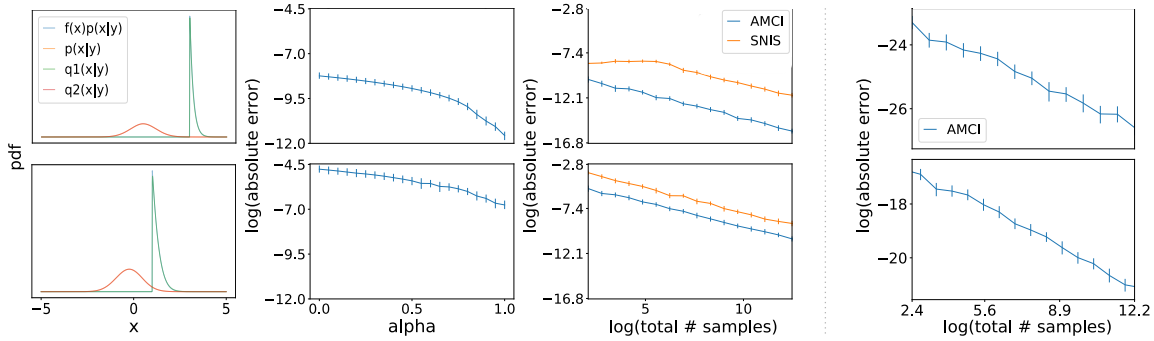
We note that the expectation over $p(y)$ in the above objective is chosen quite arbitrarily and it does not change the optimal solution to the problem (presuming an infinite capacity neural network), which is $D_{KL}\left[p(x|y)\,||\,q_2(x|y)\right] = 0\,\forall y$. Instead, it changes the relative priority of different values of different datasets y during training.

For amortizing q_1, we need to adjust the above target to incorporate the effect of the target function. Further, when the target function is parameterized, i.e. $f(x;\theta)$, we allow amortization over functions by introducing a pseudo-prior $p(\theta)$ which specifies which θ values we amortize over and their relative importance. In this case we choose to take an expectation over $h(y,\theta) \propto p(y)Z(y;\theta)p(\theta)$, where $Z(y;\theta) = \int |f(x;\theta)|\,p(x|y)\mathrm{d}x$, as this will substantially improve the tractability of the training as shown in Appendix A.

$$\mathcal{J}_1(\eta) = \mathbb{E}_{h(y,\theta)}\left[D_{KL}\left[\pi(x|y;\theta)\,||\,q_1(x;\varphi_1(y,\theta;\eta))\right]\right] \tag{3}$$

$$\propto \mathbb{E}_{p(x,y)\,p(\theta)}\left[-|f(x;\theta)|\log q_1(x;\varphi_1(y,\theta;\eta))\right] + \text{const wrt } \eta \tag{4}$$

Efficient Training If $|f(x)|$ and $p(x)$ are mismatched, i.e. $|f(x)|$ is large in regions where $p(x)$ is low, training by naïvely sampling from $p(x, y)$ can be very inefficient. Instead it is preferable to try and sample from $g(\theta, x) \propto p(\theta)p(x)\,|f(x;\theta)|$. Though this is itself an intractable distribution, it represents a standard, rather an amortized, inference problem and so it is much more manageable than the overall training problem. One simple approach is to construct an MCMC sampler targeting $g(\theta, x)$ to generate the required samples, which



(a) One dimensional example

(b) Five dimensional example

Figure 1: Results of the integration amortization experiments with a parameterized integrand $f(x;\theta)$. Rows show results for two different pairs of (y,θ). Column one in Figure (a) illustrates the shape of the proposal q_1 and the achievable quality of fit to $|f(x;\theta)|p(x|y)$. Column two presents the effects of varying the parameter α. Column three compares the performance of AMCI and SNIS estimators. In Figure (b) the line for the SNIS estimator is not presented as the proposal q_2 failed to place even one sample in the area of the tail, such all of the SNIS estimates were equal to zero. Details in Appendix C.

can be done upfront before training. Another is to construct an importance sampler, namely

$$\mathcal{J}_1'(\eta) = \mathbb{E}_{q'(\theta,x)p(y|x)}\left[-\frac{p(\theta)p(x)\,|f(x;\theta)|}{q'(\theta,x)}\log q_1(x;\varphi_1(y,\theta;\eta))\right] + \text{const wrt } \eta \tag{5}$$

where $q'(\theta,x)$ is as close to $g(\theta,x)$ as possible and could, if necessary, be pre-trained.

3. Experiments

We consider a D dimensional tail integral problem with the goal to compute $\mathbb{E}_{p(x|y)}[f(x;\theta)]$ $p(x) = \mathcal{N}(x;0,\Sigma_1)$; $p(y|x) = \mathcal{N}(y;x,\Sigma_2)$; $f(x;\theta) = \prod_{i=1}^{D} \mathbb{1}_{x_i>\theta_i}$; $p(\theta) = \textsc{Uniform}(\theta;[0,5]^D)$ and consider comparing AMCI with the SNIS estimator using the posterior approximation (i.e. q_2) as the proposal. For this problem, the posterior and true value of can be determined analytically. We are using a normalizing flow consisting of radial flow layers (Rezende and Mohamed, 2015) with standard normal base distribution as our proposal $q(x|y)$. We perform one and five-dimensional variants of the experiment. Full experimental details, including the choice of the training proposal q', are in Appendix C.

Results are presented in Figure 1. We found that fixing $\beta = 0$ universally results in the smallest error of the estimates. AMCI performed significantly better than the SNIS estimator. In the one-dimensional example, the error for AMCI is over an order of magnitude smaller than for the SNIS estimator. In the five-dimensional example, the SNIS estimator failed to place even one sample in the area of the tail, and all of the SNIS estimates were equal to zero. The initially flat line for the SNIS estimator in row one, column three in figure (a) origins from the same phenomena – often the proposal fails to place even one sample in the area yielding non-zero value for $f(x;\theta)$. In this model, the normalizing flow used for q_1 is flexible enough to match the target distribution well, and hence $\alpha = 1$ is universally optimal. However, we anticipate that in more difficult problems we will face situations when the proposal is not able to match the target distribution so well and hence $\alpha < 1$ might be optimal.

Acknowledgments

AG is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems. TR and YWT are supported in part by the European Research Council under the European Unions Seventh Framework Programme (FP7/20072013) / ERC grant agreement no. 617071. FW is supported under DARPA PPAML through the U.S. AFRL under Cooperative Agreement FA8750-14-2-0006, Sub Award number 61160290-111668.

References

Ming-Hui Chen and Qi-Man Shao. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 08 1997.

Michael Evans and Tim Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical science*, pages 254–272, 1995.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 05 1998.

Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Simon Lacoste-Julien, Ferenc Huszar, and Zoubin Ghahramani. Approximate inference for the loss-calibrated Bayesian. In Geoffrey Gordon, David Dunson, and Miroslav Dudk, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 416–424, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Tuan Anh Le, Atılım Güneş Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1338–1348, Fort Lauderdale, FL, USA, 2017. PMLR.

Tuan Anh Le, Maximilian Igl, Tom Jin, Tom Rainforth, and Frank Wood. Auto-encoding sequential Monte Carlo. In *International Conference on Learning Representations (ICLR)*, 2018.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6576–6586, 2017.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constatnts via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

Christian A Naesseth, Scott W Linderman, Rajesh Ranganath, and David M Blei. Variational sequential Monte Carlo. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Man-Suk Oh and James O Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168, 1992.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Brooks Paige and Frank Wood. Inference networks for sequential Monte Carlo in graphical models. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 2016.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, 2014.

Daniel Ritchie, Paul Horsfall, and Noah D Goodman. Deep amortized inference for probabilistic programs. *arXiv preprint arXiv:1610.05735*, 2016.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in Neural Information Processing Systems*, pages 3048–3056, 2013.

Robert L Wolpert. Monte Carlo integration in Bayesian statistical analysis. *Contemporary Mathematics*, 115:101–116, 1991.

Appendix A. Details about the amortization objective

As mentioned in the main text the choice of the distribution $p(y)$ in Equation 2 made by Paige and Wood (2016) is somewhat arbitrary. Let $\pi(x|y;\theta) = \frac{|f(x;\theta)|\,p(x|y)}{Z(y;\theta)}$, $Z(y;\theta) = \int |f(x;\theta)|\,p(x|y)\,\mathrm{d}x = \mathbb{E}_{p(x|y)}\left[|f(x;\theta)|\right]$ and $q_1(x|y;\eta) = q_1(x;\varphi_1(y,\theta;\eta))$. If we try to devise an objective for amortizing q_1 taking expectation with respect to $h(y,\theta) \propto p(y)p(\theta)$ as per the objective for amortizing q_2 we are left the intractability with respect to $Z(y;\theta)$:

$$\begin{aligned}
\mathcal{J}_1'(\eta) &= \mathbb{E}_{h(y,\theta)\propto p(y)p(\theta)}[D_{KL}(\pi(x|y;\theta)||q_1(x|y;\eta))] && (6)\\
&= \mathbb{E}_{h(y,\theta)\propto p(y)p(\theta)}\left[-\int \pi(x|y;\theta)\log q_1(x|y;\eta)\mathrm{d}x\right] + \text{const wrt } \eta && (7)\\
&= \mathbb{E}_{h(y,\theta)\propto p(y)p(\theta)}\left[-\frac{1}{Z(y;\theta)}\int |f(x;\theta)|\,p(x|y)\log q_1(x|y;\eta)\mathrm{d}x\right] + \text{const wrt } \eta && (8)\\
&\propto \mathbb{E}_{\pi(y,\theta)\propto\frac{p(y)p(\theta)}{Z(y;\theta)}}\left[-\int |f(x;\theta)|\,p(x|y)\log q_1(x|y;\eta)\mathrm{d}x\right] + \text{const wrt } \eta && (9)\\
&\propto \mathbb{E}_{\pi(x,y,\theta)\propto\frac{p(x,y)p(\theta)}{Z(y;\theta)}}\left[-|f(x;\theta)|\log q_1(x|y;\eta)\right] + \text{const wrt } \eta && (10)
\end{aligned}$$

The intractability comes from the fact we do not know $Z(y;\theta)$ and, at least at the beginning of the training process, we cannot estimate it efficiently either. Hence we cannot sample from the distribution $\pi(x,y,\theta) \propto p(x,y)p(\theta)/Z(y;\theta)$.

However, when we pick $h(y,\theta) \propto p(y)p(\theta)Z(y;\theta)$ we avoid this intractability as the terms cancel as follows

$$\begin{aligned}
\mathcal{J}_1(\eta) &= \mathbb{E}_{h(y,\theta)\propto p(y)p(\theta)Z(y;\theta)}\left[D_{KL}\left[\pi(x|y;\theta)\,||\,q_1(x|y;\eta)\right]\right] && (11)\\
&= c^{-1}\cdot\mathbb{E}_{p(x,y)p(\theta)}\left[-|f(x;\theta)|\log q_1(x|y;\eta)\right] + \text{const wrt } \eta && (12)
\end{aligned}$$

where $c = \int p(y)\,p(\theta)\,Z(y;\theta)\,\mathrm{d}y\,\mathrm{d}\theta$.

In it interesting to note that this choice of $h(y,\theta)$ can be interpreted as giving larger importance to the values of y and θ which posterior yields larger $Z(y;\theta)$. Informally, we could think about this choice as attempting to minimizing the L1 errors of our estimates, that is $\mathbb{E}_{p(y,\theta)}[||Z-\hat{Z}||_1]$.

Appendix B. Positivisation

Positivisation uses multiple importance samplers to allows one to formulate a zero variance estimator when $f(x)$ takes both positive and negative signs, and is not upper or lower bounded. Following Owen (2013), we use a standard decomposition of $f(x)$ into positive and negative parts. Define $f_+(x) = \max(f(x),0)$, and $f_-(x) = \max(-f(x),0)$. Then $f(x) = f_+(x) - f_-(x)$.

Now let $q_+(x)$ be a density function which is positive whenever $p(x)f_+(x) > 0$ and let $q_-(x)$ be a density function which is positive whenever $p(x)f_-(x) > 0$. We take $n_\pm$ samples x_a^+, x_b^- from $q_\pm(x)$ respectively. The estimator for $\mathbb{E}_{p(x)}[f(x)]$ is

$$\mathbb{E}_{p(x)}[f(x)] = \frac{1}{n_+}\sum_a^{n_+}\frac{f_+(x_a^+)p(x_a^+)}{q_+(x_a^+)} - \frac{1}{n_-}\sum_b^{n_-}\frac{f_-(x_b^-)p(x_b^-)}{q_-(x_b^-)} \tag{13}$$

The optimal sampling proposals are $q_\pm(x) \propto p(x)f_\pm(x)$, respectively. If optimal sampling proposals are used we get zero variance for sample budget $n_+ = n_- = 1$, i.e. $n = 2$.

In the AMCI setting positivisation affects the estimator of $\mathbb{E}_{p(x|y)}[f(x)]$ from Equation 1. It also implies that we need to learn three instead of two proposal distributions, which we will denote as $q_{1+}(x|y), q_{1-}(x|y)$ and $q_2(x|y)$. Instead of drawing N samples from $q_1(x|y;\eta)$, we now draw n_+, n_- samples x_a^+, x_b^- from q_1+, q_1-, respectively, such that $n_+ + n_- = N$. We also draw M samples x_m^* from q_2. The new form of the estimator is

$$\mathbb{E}_{p(x)}[f(x)p(y|x)] = \alpha \left(\frac{1}{n_+} \sum_a^{n_+} \frac{f_+(x_a^+)p(x_a^+, y)}{q_{1+}(x_a^+|y)} - \frac{1}{n_-} \sum_b^{n_-} \frac{f_-(x_b^-)p(x_b^-, y)}{q_{1-}(x_b^-|y)} \right) + \frac{1-\alpha}{M} \sum_m^M \frac{f(x_m^*)p(x_m^*, y)}{q_2(x_m^*|y)} \tag{14}$$

$$\mathbb{E}_{p(x)}[p(y|x)] = \beta \left(\frac{1}{n_+} \sum_a^{n_+} \frac{p(x_a^+, y)}{q_{1+}(x_a^+|y)} + \frac{1}{n_-} \sum_b^{n_-} \frac{p(x_b^-, y)}{q_{1-}(x_b^-|y)} \right) + \frac{1-\beta}{M} \sum_m^M \frac{p(x_m^*, y)}{q_2(x_m^*|y)} \tag{15}$$

$$\mathbb{E}_{p(x|y)}[f(x)] = \frac{\mathbb{E}_{p(x)}[f(x)p(y|x)]}{\mathbb{E}_{p(x)}[p(y|x)]} \tag{16}$$

Appendix C. Experimental details

For the one-dimensional case we used $\Sigma_1 = \Sigma_2 = 1$ while for the five-dimensional case it was

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1.2449 & 0.2068 & 0.1635 & 0.1148 & 0.0604 \\ 0.2068 & 1.2087 & 0.1650 & 0.1158 & 0.0609 \\ 0.1635 & 0.1650 & 1.1665 & 0.1169 & 0.0615 \\ 0.1148 & 0.1158 & 0.1169 & 1.1179 & 0.0620 \\ 0.0604 & 0.0609 & 0.0615 & 0.0620 & 1.0625 \end{bmatrix}$$

We used a normalizing flow consisting of radial flow layers (Rezende and Mohamed, 2015) with standard normal base distribution as our parameterized proposal $q(x|y)$. We denote the base distribution as $x_0 \sim \mathcal{N}(0, 1)$, and the radial flows transformation as $\mathcal{R}(x;\psi)$, where ψ are their parameters and they are determined by a neural network such that $x = \mathcal{R}(x_0; \varphi(y, \theta; \eta))$.

The normalizing flow for the one-dimensional example consisted of 10 radial flows, while the one for the five-dimensional example consisted of 50 flows. The neural network φ outputting the parameters of all of those flows had 3 fully connected layers with 1000 hidden units for both one and five-dimensional examples.

The $(y;\theta)$ pairs presented in Figure 1 are $(1, 3)$ in row 1, $(-0.5, 1)$ in row 2 for the one-dimensional example and $([0.9, 1.6, 1.3, -1.0, 3.5], [0, 1, 2, 3, 4])$ in row 1, $([1, 1, 4, 3, 0.5], [2, 3, 2, 3, 2])$ in row 2 for the five-dimensional example.

We use objective $\mathcal{J}_1'$ from Equation 5. The $q'(\theta, x)$ for both examples was $q'(\theta, x) = q'(\theta)q'(x|\theta)$ with $q'(\theta) = p(\theta) = \textsc{Uniform}(\theta; [0, 5]^D)$ and $q'(x|\theta) = \textsc{HalfNormal}(x; \theta, 1)$, so that every x would fall into the tail determined by θ.

Appendix D. Derivation of the optimal parameter values for the AMCI estimator

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 Equation 1. 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 Z_N and Z_D, respectively. We also define the following shorthands for the unbiased importance sampling estimators with respect to proposals q_1 and q_2 in Equation 1 $a_1 = \frac{1}{N}\sum_n^N \frac{f(x_n)p(x_n,y)}{q_1(x_n|y)}$, $b_1 = \frac{1}{M}\sum_m^M \frac{f(x_m^*)p(x_m^*,y)}{q_2(x_m^*|y)}$, $a_2 = \frac{1}{N}\sum_n^N \frac{p(x_n,y)}{q_1(x_n|y)}$, $b_2 = \frac{1}{M}\sum_m^M \frac{p(x_m^*,y)}{q_2(x_m^*|y)}$, where $x_n \sim q_1(x|y)$ and $x_m^* \sim q_2(x|y)$.

We start by considering the estimator according to Equation 1

$$\frac{Z_N}{Z_D} \approx I := \frac{\alpha a_1 + (1-\alpha)b_1}{\beta a_2 + (1-\beta)b_2}. \tag{17}$$

Using the central limit theorem, then as $N, M \to \infty$, we have

$$\to \frac{Z_N + \sigma_N \xi_N}{Z_D + \sigma_D \xi_D}, \quad \text{where} \quad \xi_N, \xi_D \sim \mathcal{N}(0,1) \tag{18}$$

are correlated standard normal random variables and σ_N and σ_D are the standard deviation of the estimators for numerator and denominator respectively. Specifically we have

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

which by the weak law of large numbers

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

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

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

Now going back to Equation 18 and using Taylor's Theorem on $1/\left(Z_D + \sigma_D \xi_D\right)$ about $1/Z_D$ gives

$$\begin{aligned} I =& \frac{Z_N + \sigma_N \xi_N}{Z_D}\left(1 - \frac{\sigma_D \xi_D}{Z_D}\right) + O(\epsilon) \\ =& \frac{Z_N}{Z_D} + \frac{\sigma_N \xi_N}{Z_D} - \frac{Z_N \sigma_D \xi_D}{Z_D^2} - \frac{\sigma_N \sigma_D \xi_N \xi_D}{Z_D^2} + O(\epsilon)\end{aligned}$$

where $O(\epsilon)$ represents asympotitically dominated terms. Note here the importance of using Taylor's theorem, rather just a Taylor expansion, to to confirm that these terms are indeed asympotitically dominated. We can further drop the $\frac{\sigma_N \sigma_D \xi_N \xi_D}{Z_D^2}$ term as this will be order $O(1/\sqrt{MN})$ and will thus be asymptotically dominated, giving

$$= \frac{Z_N}{Z_D} + \frac{\sigma_N \xi_N}{Z_D} - \frac{Z_N \sigma_D \xi_D}{Z_D^2} + O(\epsilon).$$

To calculate the MSE of I, we start with the standard bias variance decomposition

$$\mathbb{E}\left[\left(I-\frac{Z_N}{Z_D}\right)^2\right]=\mathrm{Var}\left[I\right]+\left(\mathbb{E}\left[I-\frac{Z_N}{Z_D}\right]\right)^2.$$

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

$$\begin{aligned}
\mathrm{Var}[I] &= \mathrm{Var}\left[\frac{\sigma_N\xi_N}{Z_D}\right]+\mathrm{Var}\left[\frac{Z_N\sigma_D\xi_D}{Z_D^2}\right]-2\mathrm{Cov}\left[\frac{\sigma_N\xi_N}{Z_N},\frac{Z_N\sigma_D\xi_D}{Z_D^2}\right]+O(\epsilon)\\
&= \frac{\sigma_N^2}{Z_D^2}+\frac{Z_N^2\sigma_N^2}{Z_D^4}-2\frac{\sigma_N Z_N\sigma_D}{Z_D^3}\mathrm{Cov}[\xi_N,\xi_D]+O(\epsilon)\\
&= \frac{\alpha^2}{NZ_D^2}\mathrm{Var}_{q_1}[f(x_1)w_1]+\frac{(1-\alpha)^2}{MZ_D^2}\mathrm{Var}_{q_2}[f(x_1^*)w_1^*]\\
&\quad+\frac{Z_N^2\beta^2}{NZ_D^4}\mathrm{Var}_{q_1}[w_1]+\frac{Z_N^2(1-\beta)^2}{MZ_D^4}\mathrm{Var}_{q_2}[w_1^*]\\
&\quad-2\frac{Z_N}{Z_D^3}\mathrm{Corr}[\xi_N,\xi_D]\left(\frac{\alpha^2}{N}\mathrm{Var}_{q_1}[f(x_1)w_1]+\frac{(1-\alpha)^2}{M}\mathrm{Var}_{q_2}[f(x_1^*)w_1^*]\right)\\
&\quad\times\left(\frac{\beta^2}{N}\mathrm{Var}_{q_1}[w_1]+\frac{(1-\beta)^2}{M}\mathrm{Var}_{q_2}[w_1^*]\right)
\end{aligned}$$

To assist in the subsequent analysis, we assume that there is no correlation, $\mathrm{Corr}[\xi_N,\xi_D]=0$. 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. Nonetheless, catering for non-zero correlations is something one may wish to look into in future work.

Given this assumption is now straightforward to optimize for α and β by finding where the gradient is zero as follows

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

noting that

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

and hence it's a local minimum. Analogously

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

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.