A Discussion on Amortized Monte Carlo Integration (AMCI)

April 30, 2020

- Bayesian Inference has a goal of approximating some posterior distribution.
- These approximations are inefficient in calculating expected values of functions (target functions) if the functions are known upfront.
- This is what AMCI tries to address.

Review

- Goal: Calculate $E_{p(x|y)}[f(x)]$.
  - One Approach: MC Sampling
- Problem: MANY papers in the past have shown that
  - Information of $f(x)$ known $\rightarrow$ MC Sampling is Inefficient

Solution

- Perform inference incorporating information about target function $f(x)$.
- Perform inference in amortized setting.

Importance Sampling

$$\mu := \mathbb{E}_{\pi(x)}\left[f(x)\right] = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx$$
$$\approx \hat{\mu} := \frac{1}{N}\sum_{n=1}^{N} f(x_n)w_n$$

Figure: Approximation for $E_{\pi(x)}[f(x)]$

Importance Sampling (cont.)

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

Figure: Approximation under Self-Normalized Importance Sampling (SNIS)

Importance Sampling (cont.)

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

Figure: Lower Bound of Achievable Error for the Self-Normalized Case

Inference Amortization

- Parameterized Proposal: $q(x; \varphi(y; \eta)) = q(x; y, \eta)$
- Data: y
- Inference Network: $\varphi(y; \eta)$

$$\begin{tikzcd}
\eta \arrow[r] & \varphi(y;\eta)) \arrow[d] & & y \arrow[ll] \\
 & q(x;\varphi(y;\eta)) \arrow[r] & \text{Proposal Parameters} &
\end{tikzcd}$$

AMCI

What makes AMCI unique from standard amortized inference?

ACMI ...

- operates in a target-aware fashion.
- uses 3 different proposal distributions.
- allows for amortization over parameterized target functions ($f(x; \theta)$ not just $f(x)$).

AMCI Inference Network

Method taken from Paige and Wood (2016)
Optimization Problem:

$$
\begin{aligned}
argmin_\eta \mathcal{J}(\eta) &= argmin_\eta E_{p(y)}\left[D_{KL}[p(x|y)||q(x;y,\eta)]\right] \\
&= argmin_\eta E_{p(x,y)}[-log(q(x;y,\eta))]
\end{aligned}
$$

Sampling from $p(x,y)$ can be optimized using gradient methods:

$$
\nabla_\eta \mathcal{J}(\eta) = E_{p(x,y)}[-\nabla_\eta log(q(x;y,\eta))]
$$

AMCI Introduction

- Goal of AMCI: Amortize the cost of calculating $\mu(y, \theta) := E_{\pi(x;y)}(f(x;\theta))$

$$\mu(y,\theta) := \mathbb{E}_{p(x|y)}\big[f(x;\theta)\big] = \frac{\mathbb{E}_{p(x|y)}\big[f(x;\theta)\,p(y)\big]}{\mathbb{E}_{p(x)}\big[p(y|x)\big]}$$

$$= \frac{\mathbb{E}_{q_1(x;y,\theta)}\left[\frac{f(x;\theta)p(x,y)}{q_1(x;y,\theta)}\right]}{\mathbb{E}_{q_2(x;y)}\left[\frac{p(x,y)}{q_2(x;y)}\right]} =: \frac{E_1}{E_2}$$

- Numerator: Unnormalized Expectation
- Denominator: Normalization Constant

AMCI Introduction (cont.)

Now that we have 2 expected value functions, we can take 2 MC samples.

$$\mu(y,\theta) \approx \hat{\mu}(y,\theta) := \hat{E}_1/\hat{E}_2 \quad \text{where}$$

$$\hat{E}_1 := \frac{1}{N}\sum_{n=1}^{N}\frac{f(x'_n;\theta)p(x'_n,y)}{q_1(x'_n;y,\theta)} \quad x'_n \sim q_1(x;y,\theta)$$

$$\hat{E}_2 := \frac{1}{M}\sum_{m=1}^{M}\frac{p(x_m,y)}{q_2(x_m;y)} \quad x_m \sim q_2(x;y).$$

"we can now separately train each of these proposals to be good estimators for their respective expectation"

Comparison to SNIS

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

$$\mu(y,\theta) := \mathbb{E}_{p(x|y)}\left[f(x;\theta)\right] = \frac{\mathbb{E}_{p(x|y)}\left[f(x;\theta)\,p(y)\right]}{\mathbb{E}_{p(x)}\left[p(y|x)\right]}$$

$$= \frac{\mathbb{E}_{q_1(x;y,\theta)}\left[\frac{f(x;\theta)p(x,y)}{q_1(x;y,\theta)}\right]}{\mathbb{E}_{q_2(x;y)}\left[\frac{p(x,y)}{q_2(x;y)}\right]} =: \frac{E_1}{E_2}$$

"...the more $|f(x;\theta)|p(x|y)$ varies from $p(x|y)$, the worse the conventional approach of only amortizing Amortized Monte Carlo Integration the posterior will perform, while the harder it becomes to construct a reasonable SNIS estimator even when information about $f(x;\theta)$ is incorporated."

Theoretical Zero-Variance Estimator for AMCI

We have already shown that we can achieve such an estimator for the case of importance sampling when the target function is non-negative $f(x) \geq 0$.
For our new estimator, we can relax this assumption.
Splitting our target function into its positive and negative components:

- $f^+(x;\theta) = max(f(x;\theta), 0)$
- $f^-(x;\theta) = -min(f(x;\theta), 0)$

Theoretical Zero-Variance Estimator for AMCI (cont.)

Using the target function decomposition, we have

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

From here, we again take MC samples to obtain

$$\mu(y,\theta) \approx \hat{\mu}(y,\theta) := (\hat{E}_1^+ - \hat{E}_1^-)/\hat{E}_2 \quad \textbf{where}$$

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

$$\hat{E}_1^- := \frac{1}{K}\sum_{k=1}^{K} \frac{f^-(x_k^-;\theta)p(x_k^-,y)}{q_1^-(x_k^-;y,\theta)} \quad x_k^- \sim q_1^-(x;y,\theta)$$

$$\hat{E}_2 := \frac{1}{M}\sum_{m=1}^{M} \frac{p(x_m,y)}{q_2(x_m;y)} \quad x_m \sim q_2(x;y), \quad (14)$$

Theoretical Zero-Variance Estimator for AMCI (cont.)

This leads us to our final conclusion.

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.

Existing Amortization Inference Setbacks

- Solution is suboptimal if information aboiut f(x) is available.
- There is a lower bound on the achievable error.

Putting the A in AMCI

- Benefits of Amortization: Amortizing over ...
  - y: explicit parameterization isn't needed.
  - θ: reference distribution $\pi(x; y)$ can be fixed.
- To obtain our theoretical zero-variance estimator, we need to learn 3 amortized proposals:
  - $q_1^+(x; y, \theta)$
  - $q_1^-(x; y, \theta)$
  - $q_2(x; y)$

Amortization for Fixed Function f(x)

Because we are not amortizing over θ, we drop the proposals dependence on it.
If we let $g(x|y)$ be defined as the normalized optimal proposal for q_1, we get the following objective function:

$$
\begin{aligned}
\mathcal{J}_1^{'}(\eta) &= E_{p(y)}[D_{KL}(g(x|y)||q_1(x;y,\eta))] \\
&= E_{p(y)}\left[-\int_{\mathcal{X}} \frac{f(x)p(x,y)}{E_1(y)} log(q_1(x;y,\eta))dx\right] + k
\end{aligned}
$$

where $E_1(y) = E_{p(x)}[f(x)p(y|x)]$
<u>Problem</u>: $E_1(y)$ is unknown with no good way of estimating it.
<u>Solution</u>: ...?

Making a Well-Defined Objective Function

the expectation with respect to $h(y) \propto p(y) E_1(y)$,

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

Making a Well-Defined Objective Function for $f(x;\theta)$

If $E_1(y,\theta) := \mathbb{E}_{p(x)}\left[f(x;\theta)p(y|x)\right]$, $g(x|y;\theta) := f(x;\theta)\,p(x,y)/E_1(y,\theta)$, and $h(y,\theta) \propto p(y)p(\theta)E_1(y;\theta)$, we get an objective which is analogous to (20):

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

Is AMCI practical?

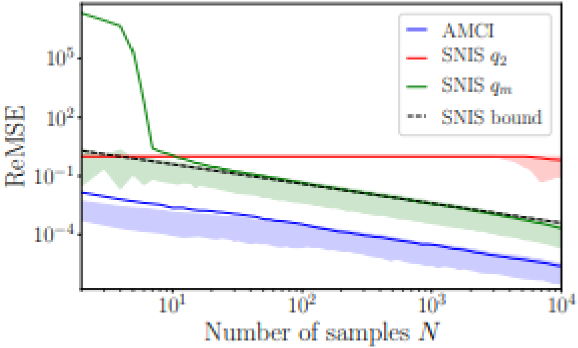
- Exact estimators are highly unlikely with low sample values under imperfect proposals.
- In order to have a proper answer to this question, an assessment of gain has to be done on imperfect proposals.

Tail Integration Experiement

This experiment is a good baseline because there is a good ground truth to compare it to (analytical methods).
Baseline estimations are evaluated by their relative mean squared error (ReMSE):

$$\delta(y,\theta) = E\left(\hat{\delta}(y,\theta)\right) = E\left(\frac{(\mu(y,\theta) - \hat{\mu}(y,\theta))^2}{\mu(y,\theta)^2}\right)$$

Result:



AMCI v. SNIS: An Asymptotic Comparison

For simplicity, assume $f(x;\theta) \geq 0, \forall x, \theta$.
If we apply the central limit theorem to the separate estimator $\hat{E}_1$ and $\hat{E}_2$, we get

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

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

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

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

Conclusion

By approximating the MSE as shown below,

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

we see that $\kappa \to 1 \implies \text{AMIC} \approx \text{SNIS}$.

Personal Discussion Questions

Questions I had that I think would be good to talk about as a group.

- What is the computation cost of amortization? If κ is at the point where AMCI is marginal to SNIS, what's the time complexity tradeoff?

> More generally, if we choose $h(y) \propto p(y)E_1(y)\lambda(y)$ for some positive evaluable function $\lambda : \mathcal{Y} \to \mathbb{R}^+$, 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.

- What is the benefit of such a function lambda? Could this adjustment be abused?