Debiasing Approximate Inference

Sebastian Nowozin

Current: Microsoft Research Cambridge

Soon: Google AI Berlin (Brain team)

Symposium on Advances in Approximate Inference – NeurIPS 2018
2\textsuperscript{nd} December 2018
The Pattern

• “Consistent” approximate inference: consistent estimator of some quantity of interest \((k \to \infty)\)
• Computationally constraints put limits on \(k\)
Example 1: Self-Normalized Importance Sampling

• Quantity of interest

\[ \mathbb{E}_{x \sim p}[f(x)] \]

• Consistent estimator, unnormalized \( \tilde{p} \)

\[ \hat{L}(\tilde{p}, q, k) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\tilde{p}(x_i) q(x_i)} \tilde{p}(x_i) f(x_i), \quad x_i \sim q \]
Example 2: AIS Evidence Estimates

• Annealed Importance Sampling: unbiased estimates of
  \[ \hat{p}_k(x) \rightarrow p(x) = \int p(x|z) p(z) \, dz \]

• However, in many applications we need to estimate (See [Salakhutdinov and Murray, 2008])
  \[ \log p(x) \]

• Naïve plug-in estimate, consistent,
  \[ \hat{L}_k = \log \hat{p}_k(x) \]

• Biased ("stochastic lower bound", sounds better)
Example 3: Importance Weighted Autoencoder (IWAE)

• Family of tighter ELBO bounds [Burda et al., 2015]
• Intractable expectation

$$\log \mathbb{E}_{z \sim q_\omega(z|x)} \left[ \frac{p_\theta(x|z) p(z)}{q_\omega(z|x)} \right]$$

• Approximate “naively” using empirical expectation

$$\hat{\mathcal{L}}_K := \log \frac{1}{K} \sum_{i=1}^{K} \frac{p_\theta(x|z) p(z)}{q_\omega(z|x)}$$

$$z_i \sim q_\omega(z|x)$$
IWAE: known results [Burda et al., 2015]

- ELBO recovery
  \[ \text{ELBO} = \hat{L}_1 \]

- Consistency

**Corollary 2** (Consistency of \( \hat{L}_K \)). For \( K \to \infty \) the estimator \( \hat{L}_K \) is consistent, that is, for all \( \epsilon > 0 \)
  \[
  \lim_{K \to \infty} P(|\hat{L}_K - \log p(x)| \geq \epsilon) = 0.
  \] (12)

- Stochastic monotonicity (== bias)
  \[ \mathbb{E}\hat{L}_1 \leq \mathbb{E}\hat{L}_2 \leq \cdots \leq \mathbb{E}\hat{L}_\infty = \log p(x) \]
Example 4: Markov Chain Monte Carlo

- Quantity of interest
  \[ \mathbb{E}_{x \sim p}[f(x)] \]

- Consistent estimator from truncated Markov chain samples
  \[ x_{t+1} \sim T_k(x_{t+1} | x_t), \quad x_0 \sim T_0(x_0) \]
  \[ \hat{L}_k = \frac{1}{k} \sum_{t=1}^{k} f(x_t) \]

- Bias due to truncation

- See [Strathmann et al., ICML 2015]
Example 5: Stochastic Metropolis-Hastings Acceptance Rates

- Stochastic Gradient MCMC method omit accept-reject step in a Metropolis-Hastings chain (e.g. SGLD [Welling and Teh, ICML 2011])
- Problem: exact (MALA) acceptance rate is intractable
  \[ \alpha_n(x \rightarrow x') = \min \left\{ 1, \frac{\tilde{p}_n(x')}{\tilde{p}_n(x)} \cdot \frac{q_n(x|x')}{q_n(x'|x)} \right\} \]
- Consistent estimator by truncation
  \[ \alpha_k(x \rightarrow x') \]
- Biased due to exponentiation and min operation
Debiasing Methods
Analytic Methods

Delta Method
Case-by-Case

Resampling Methods
Jackknife Debiasing
Bootstrap Debiasing

Stochastic Methods
Russian Roulette, Debiasing Lemma

Peter Hall’s bootstrap lecture notes, 2016

Christopher G. Small, CRC Press, 2010

[Schucany et al., JASA 1971]
and [Sharot, JASA 1976]

[Lyne et al., Statistical Science, 2015]
Importance Weighted Autoencoder (IWAE)

• Family of tighter ELBO bounds [Burda et al., 2015]
• Intractable expectation

\[
\log \mathbb{E}_{z \sim q_\omega(z|x)} \left[ \frac{p_\theta(x|z) \ p(z)}{q_\omega(z|x)} \right]
\]

• Approximate “naively”

\[
\hat{\mathcal{L}}_K := \log \frac{1}{K} \sum_{i=1}^{K} \frac{p_\theta(x|z) \ p(z)}{q_\omega(z|x)}
\]

\[z_i \sim q_\omega(z|x)\]
Delta Method for Moments

• == Taylor expansion

• Here, Taylor expand log around $\mathbb{E}[w]$, evaluate at $Y_k = \frac{1}{k} \sum w_i$

\[
\log Y_k = \log(\mathbb{E}[w] + (Y_k - \mathbb{E}[w])) \\
= \log \mathbb{E}[w] - \sum_{j=1}^{\infty} \frac{(-1)^j}{j\mathbb{E}[w]^j} \mathbb{E}[(Y_k - \mathbb{E}[w])^j]
\]
Delta Method VI

\[ \log Y_k = \log \mathbb{E}[w] - \sum_{j=1}^{\infty} \frac{(-1)^j}{j! \mathbb{E}[w]^j} \mathbb{E}[(Y_k - \mathbb{E}[w])^j] \]

- Naïve estimator
- Quantity of interest (intractable)
- Correction terms (intractable)
\[ \log \mathbb{E}[w] = \log Y_k + \sum_{j=1}^{\infty} \frac{(-1)^j}{j! \mathbb{E}[w]^j} \mathbb{E}[(Y_k - \mathbb{E}[w])^j] \]

**Quantity of interest**
(intractable)

**Naïve estimator**

**Correction terms**
(intractable)

Delta Method VI
Delta Method VI

$$\log \mathbb{E}[w] = \log Y_k + \frac{-1}{\mathbb{E}[w]} \mathbb{E}[Y_k - \mathbb{E}[w]] + \sum_{j=2}^{\infty} \frac{(-1)^j}{j \mathbb{E}[w]^j} \mathbb{E}[(Y_k - \mathbb{E}[w])^j]$$

Quantity of interest (intractable)  Naïve estimator = 0

Remaining correction terms (intractable)
Delta Method VI [Teh et al., 2007]

\[
\log \mathbb{E}[w] = \log Y_k + \frac{1}{2 \mathbb{E}[w]^2} \mathbb{E}[(Y_k - \mathbb{E}[w])^2] + \sum_{j=3}^{\infty} \frac{(-1)^j}{j \mathbb{E}[w]^j} \mathbb{E}[(Y_k - \mathbb{E}[w])^j]
\]

Quantity of interest (intractable)

Naïve estimator

Approximate using estimated moments

Remaining correction terms (intractable)

\[
\frac{\hat{\mu}_2}{2\hat{\mu}^2}
\]
Delta Method VI [Teh et al., 2007]

\[
\log \mathbb{E}[w] \approx \log Y_k + \frac{\hat{\mu}_2}{2\hat{\mu}^2}
\]

• Indeed reduces bias to $o(k^{-2})$, [Nowozin, 2018]

**Proposition 5** (Bias of $\hat{\mathcal{L}_K^D}$). We evaluate the bias of $\hat{\mathcal{L}_K^D}$ in (53) as follows.

\[
\mathbb{B}[\hat{\mathcal{L}_K^D}] = -\frac{1}{K^2} \left( \frac{\mu_3}{\mu^3} - \frac{3\mu_2^2}{2\mu^4} \right) + o(K^{-2}).
\]
[Nowozin, “Debiasing Evidence Approximations”, ICLR 2018]
Inverse sample size $\frac{1}{n}$ to $\frac{1}{n-1}$

Estimate $\hat{t}_\infty$, $\hat{t}_n$, $\hat{t}_{n-1}$
Assume we have an asymptotic expansion

$$\mathbb{E}[\hat{T}_k] = T + \frac{a_1}{k} + \frac{a_2}{k^2} + \cdots$$

Then

$$\mathbb{E}[k\hat{T}_k - (k - 1)\hat{T}_{k-1}] = k \left( T + \frac{a_1}{k} + \frac{a_2}{k^2} \right) - (k - 1) \left( T + \frac{a_1}{k-1} + \frac{a_2}{(k-1)^2} \right) + O(k^{-2})$$

$$= T + a_1 + \frac{a_2}{k} - a_1 - \frac{a_2}{k-1} + O(k^{-2})$$

$$= T - \frac{a_2}{k(k-1)} + O(k^{-2})$$

$$= T + O(k^{-2})$$
Generalized Jackknife

• Original jackknife: [Quennouille, 1949]
  • Removes first order $O(n^{-1})$ bias

• Generalization to higher-order bias removal: [Schucany et al., 1974]
  • Eliminates bias to any order
  • Variance typically increases
Sharot form of the generalized Jackknife

\[
\hat{T}_G^{(m)} = \sum_{j=0}^{m} c(n, m, j) \hat{T}_{n-j}.
\]

\[
c(n, m, j) = (-1)^j \frac{(n-j)^m}{(m-j)!j!}.
\]

\[
\hat{T}_G^{(0)} = \hat{T}_n
\]

\[
\hat{T}_G^{(1)} = n\hat{T}_n - (n-1)\hat{T}_{n-1}
\]

\[
\hat{T}_G^{(2)} = \frac{n^2}{2} \hat{T}_n - (n-1)^2\hat{T}_{n-1} + \frac{(n-2)^2}{2}\hat{T}_{n-2}
\]

- [Sharot, 1976]
- \(n\): sample size
- \(m\): order of the jackknife, \(m \geq 0\)
- \(\hat{T}_n\): original consistent estimator evaluated on \(n\) samples
Jackknife Variational Inference (JVI)

**Definition 1** (Jackknife Variational Inference (JVI)). Let $K \geq 1$ and $m < K$. The jackknife variational inference estimator of the evidence of order $m$ with $K$ samples is

$$
\hat{L}_K^{J,m} := \sum_{j=0}^{m} c(K, m, j) \bar{L}_{K-j},
$$

(20)

where $\bar{L}_{K-j}$ is the empirical average of one or more IWAE estimates obtained from a subsample of size $K-j$, and $c(K, m, j)$ are the Sharot coefficients defined in (18). In this paper we use all possible $\binom{K}{K-j}$ subsets, that is,

$$
\bar{L}_{K-j} := \frac{1}{\binom{K}{K-j}} \sum_{i=1}^{\binom{K}{K-j}} \hat{L}_{K-j}(Z_{i}^{(K-j)}),
$$

(21)

where $Z_{i}^{(K-j)}$ is the $i$'th subset of size $K-j$ among all $\binom{K}{K-j}$ subsets from the original samples $Z = (z_1, z_2, \ldots, z_K)$. We further define $L_K^{J,m} = \mathbb{E}_Z[\hat{L}_K^{J,m}]$. 
Higher-order Bias Reduction

Figure 5: Absolute bias as a function of $K$. 

Bias of log $p(x)$ evidence approximations, $P = \text{Gamma}(0.1, 1)$
Evidence Evaluations (VAE MNIST)

- Effective bias reduction
- Higher-order terms matter
Example 5: Stochastic Metropolis-Hastings Acceptance Rates

- Stochastic Gradient MCMC method omit accept-reject step in a Metropolis-Hastings chain (e.g. SGLD [Welling and Teh, ICML 2011])

- Problem: exact (MALA) acceptance rate is intractable

\[
\alpha_n(\theta \rightarrow \theta') = \min \left\{ 1, \frac{\tilde{p}_n(\theta')}{\tilde{p}_n(\theta)} \frac{q_n(\theta|\theta')}{q_n(\theta'|\theta)} \right\}
\]

- Consistent estimator by truncation

\[\alpha_k(x \rightarrow x')\]

- Biased due to exponentiation and min operation
Stochastic Acceptance Rate
(Ceperley and Dewing, “Penalty MCMC”, 1998)

\[ \log \frac{\tilde{p}_n(\theta')}{{\tilde{p}}_n(\theta)} \] is deterministic. Consider a random batch \( B \) of size \( 1 \ll m \ll n \), then approximately

\[ \log \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \sim \mathcal{N}(\mu(\theta, \theta'), \nu(\theta, \theta')) \]

Why? CLT:

\[ \log \tilde{p}_B(\theta) = \log p(\theta) + \frac{n}{|B|} \sum_{i \in B} \log p(x_i | \theta) \]
Ceperley-Dewing, Intuition, 1/2

- (Formal proof relies on relating Log-Normal distribution tail masses)
- Intuition:
  \[
  \log \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \sim \mathcal{N}(\mu(\theta, \theta'), v(\theta, \theta'))
  \]
- Then
  \[
  \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \sim \text{LogNormal}(\mu(\theta, \theta'), v(\theta, \theta'))
  \]
- And
  \[
  \mathbb{E} \left[ \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \right] = \exp \left( \mu(\theta, \theta') + \frac{1}{2} v(\theta, \theta') \right)
  \]
\[
E \left[ \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \right] = \exp \left( \mu(\theta, \theta') + \frac{1}{2} \nu(\theta, \theta') \right)
\]

\[
E \left[ \frac{\tilde{p}_B(\theta')}{\tilde{p}_B(\theta)} \exp \left( -\frac{1}{2} \nu(\theta, \theta') \right) \right] = \exp(\mu(\theta, \theta')) = \frac{\tilde{p}_n(\theta')}{\tilde{p}_n(\theta)}
\]

penalty factor, < 1

- Under assumptions of Normality: exact debiasing of stochastic MCMC
- The larger the variance, the worse the penalty
Langevin-Ceperley-Dewing (LCD)
Joint work with Alexander Gaunt (MSR Cambridge)

• Extension of Ceperley-Dewing to discretized Langevin dynamics
• Goal: Make SGLD valid for any stepsize via stochastic rejection

• Ceperley-Dewing assumes $q(\theta' | \theta) = q(\theta | \theta')$
• Simple extension using two batches $B, B'$, one for $q_B$, and one for the likelihood ratio
Normal Experiment (1D)

• Simple 1D Normal mean experiment

\[ \mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \]
\[ x_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, 1000 \]

• Infer

\[ p(\mu|x_1, \ldots, x_{1000}) \]

• Compare SGLD and LCD for different stepsizes, batchsize 64

• Initialize using true posterior, no burn-in
Normal Experiment (1D)

Kolmogorov-Smirnov statistic

- cd
- sgld
- mhb
- cd-1b

$10^{-1}$ $10^{0}$ $10^{1}$
Conclusions

• Approximate Inference is an estimation problem
• Let’s use a broader toolbox of techniques to make bespoke tradeoffs in approximate inference
Thanks!

nowozin@gmail.com

Code for JVI:
https://github.com/Microsoft/jackknife-variational-inference