A general method for debiasing a Monte Carlo estimator

Maxime Vandegar

SLAC

Context

- Let us say that we want to optimize $\log p(y)$.
- Taking a variational approach, we can introduce $q_\psi(x|y)$ and optimize the evidence lower bound:

$$ELBO(\theta, \psi) = \mathbb{E}_{q_\psi} \log p_\theta(y|x) - KL(q_\psi(y|x)||p(x))$$
$$\leq \log p(y)$$

- If the bound is not tight, one ends up optimizing for the bias without improving the marginal likelihood objective.

Context

- Two main approaches exist to tighten the bound:
  - Model $q_\psi(x|y)$ from a large distribution family so that it can closely match the posterior distribution (i.e. efficiently optimize $\mathrm{KL}(q_\psi(x|y)||p(x|y))$).
  - Tighten the bound using more work (Importance Weighted Autoencoders, IWAEs):

$$\mathcal{L}_K^{IW}(\phi,\psi) = \log \frac{1}{K}\sum_{k=1}^{K} \frac{p(x_k)p_\theta(y|x_k)}{q_\psi(x_k|y)}.$$

Variational Inference with Normalizing Flows, arXiv:1505.05770.
Importance Weighted Autoencoders , arXiv:1509.00519.

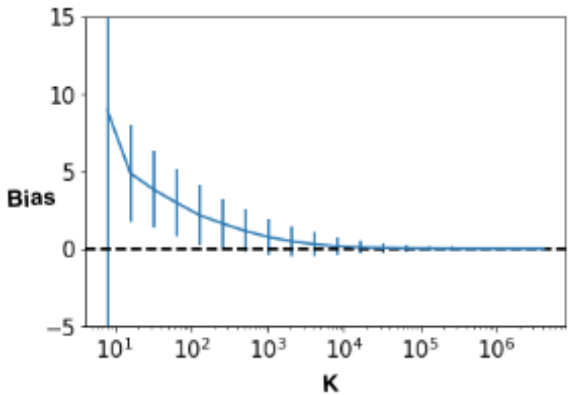
IWAEs

- The bias and variance of the estimator vanish for $K \to \infty$ at a rate $O(1/K)$.

$$\log p(y) \geq \mathcal{L}^{IW}_{K+1} \geq \mathcal{L}^{IW}_{K}$$

and

$$\lim_{K \to \infty} \mathcal{L}^{IW}_{K} = \log p(y).$$



Can we get an unbiased estimate with finite computational resources?

- The objective can be unbiased based on random truncation of an infinite convergent serie $\rightarrow$ Russian roulette, Kahn (1955).
- Rediscovered more recently and independently.
  - McLeish (2011).
  - Rhee and Glynn (2012).

Use of Different Monte Carlo Sampling Techniques.
A general method for debiasing a Monte Carlo estimator , arXiv:1005.2228.
A new approach to unbiased estimation for SDE's, arXiv:1207.2452.

A general method for debiasing a Monte Carlo estimator

McLeish, 2011

Unbiased estimator: derivation

- Assume the convergent sequence $\{X_0, X_1, ...\}$.
- Suppose $\mathcal{K}$ is a random variable taking finite non-negative integer values with

$$P(\mathcal{K} \geq k) > 0, \forall k > 0.$$

- Let us define the random variable

$$Y = X_0 + \sum_{k=1}^{K} \frac{\nabla X_k}{P(\mathcal{K} \geq k)}$$

  where $\nabla X_k = X_k - X_{k-1}$ and $K \sim P(\mathcal{K})$.

- Y can be rewritten as

$$X_0 + \sum_{k=1}^{\infty} \nabla X_k \frac{I(k \leq K)}{P(\mathcal{K} \geq k)}.$$

Unbiased estimator: derivation

- $Y = X_0 + \sum_{k=1}^{\infty} \nabla X_k \frac{I(k \leq K)}{P(\mathcal{K} \geq k)}$ is an unbiased estimate of the limit X_∞.

$$\mathbb{E}_{K \sim P(\mathcal{K})}[Y] = X_0 + \sum_{k=1}^{\infty} \nabla X_k \frac{\mathbb{E}_{K \sim P(\mathcal{K})}[I(k \leq K)]}{P(\mathcal{K} \geq k)}$$

$$= X_0 + \sum_{k=1}^{\infty} \nabla X_k = X_\infty.$$

Unbiased estimator: variance

- Any distribution satisfying $P(\mathcal{K} \geq k) > 0, \forall k > 0$ yields an unbiased estimator.
- $P(\mathcal{K} \geq k)$ should be chosen in order to provide a low-variance estimator.

Unbiased estimator: variance

- It can be shown that:

$$\sigma_Y^2 = \mathbb{E}[\sum_{k=1}^{\infty} \frac{(\nabla X_k)^2}{P(\mathcal{K} \geq k)}(1 - P(\mathcal{K} \geq k))$$

$$+2\sum_{j=1}^{\infty}\sum_{k=j+1}^{\infty} \frac{\nabla X_k \nabla X_j}{P(\mathcal{K} \geq j)}(1 - P(\mathcal{K} \geq j))].$$

- Thus, the variance can be unbiasedly estimated using:

$$\mathbb{E}[\sum_{k=1}^{N} \frac{(\nabla X_k)^2}{P(\mathcal{K} \geq k)^2}(1 - P(\mathcal{K} \geq k))$$

$$+2\sum_{j=1}^{N}\sum_{k=j+1}^{N} \frac{\nabla X_k \nabla X_j}{P(\mathcal{K} \geq j)^2 P(\mathcal{K} \geq k)}(1 - P(\mathcal{K} \geq k))].$$

Unbiased estimator: variance

- The variance can also be written as:

$$\sigma_Y^2 = \sum_{k=1}^{\infty} \frac{2(X_\infty - \xi_k)\nabla_{u_k} - \nabla_{\sigma_k^2}}{P(\mathcal{K} \geq k)} - (X_\infty - X_0)^2$$

  where $\xi_k = \frac{\mu_k + \mu_{k-1}}{2}, \nabla_{\mu_k} = \mu_k - \mu_{k-1}, \nabla_{\sigma_k} = \sigma_k - \sigma_{k-1}.$

- Suppose we wish to minimize the variance subject to a constraint on the expected value of K, i.e.

$$min_{P(\mathcal{K} \geq k)}\{\sigma_Y^2\}$$

  subject to

$$\sum_k P(\mathcal{K} \geq k) = C.$$

Unbiased estimator: variance

- Introducing the Lagrangian

$$\sigma_Y^2 - \lambda(\sum_k P(\mathcal{K} \geq k) - C$$

  and differentiating with respect to $P(\mathcal{K} \geq k)$ gives the minimum variance

$$P(\mathcal{K} \geq k) \sim c\sqrt{|2(X_\infty - \xi_k)\nabla_{\mu_k} - \nabla_{\sigma_k^2}|}$$

  where $c = \frac{1}{\lambda}$ is determined by the contraint $\sum_k P(\mathcal{K} \geq k) = C$.

- The minimum variance $P(\mathcal{K} \geq k)$ depends on X_∞ which is not practical.
- Yet, it is common to have information about the rate of convergence of the sequence that can be used to design an asymptotically appropriate sequence $P(\mathcal{K} \geq k)$.

Unbiased estimator: variance

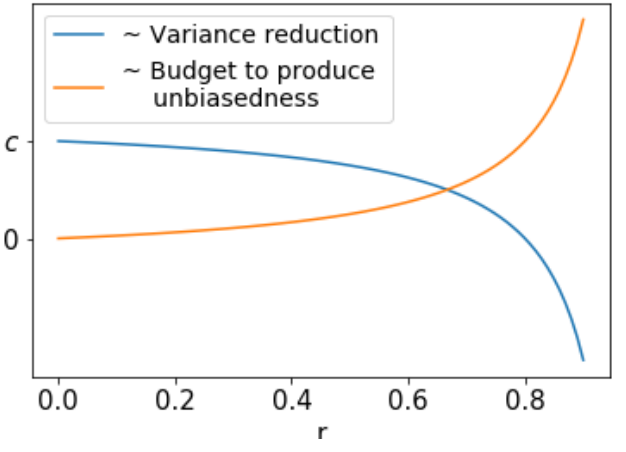
Example

- Let us assume $X_\infty - X_k \sim ar^k$, $|r| < 1$ and $X_\infty = b$.
- Let us use a shifted geometric distribution $P(\mathcal{K} \geq k) = q^{k-s}$.
- Solving the constrained problem with $\mathbb{E}_{p(\mathcal{K})}[K] = c$ gives the optimum values of $q = |r|$ and $s \sim c - \frac{|r|}{1-|r|}$.

Unbiased estimator: variance

Example

- Solving the constrained problem with $\mathbb{E}_{p(\mathcal{K})}[K] = c$ gives the optimum values of $q = |r|$ and $s \sim c - \frac{|r|}{1-|r|}$.
- Intuitively, when the rate of convergence if fast (r is small), the minimum variance is achieved by a large guarantee on the value of K (s is large).
- The residual budget $c - s = \frac{|r|}{1-|r|}$ is used to produce unbiasedness.



Unbiased estimator: variance

Example: bias-variance tradeoff

- $MSE = Variance + \cancel{Bias^2}$.
- If we were to stop after c iterations (compute X_c which is biased), the MSE would be smaller by a factor of approximately $|r|^{\frac{-2|r|}{1-|r|}}$.

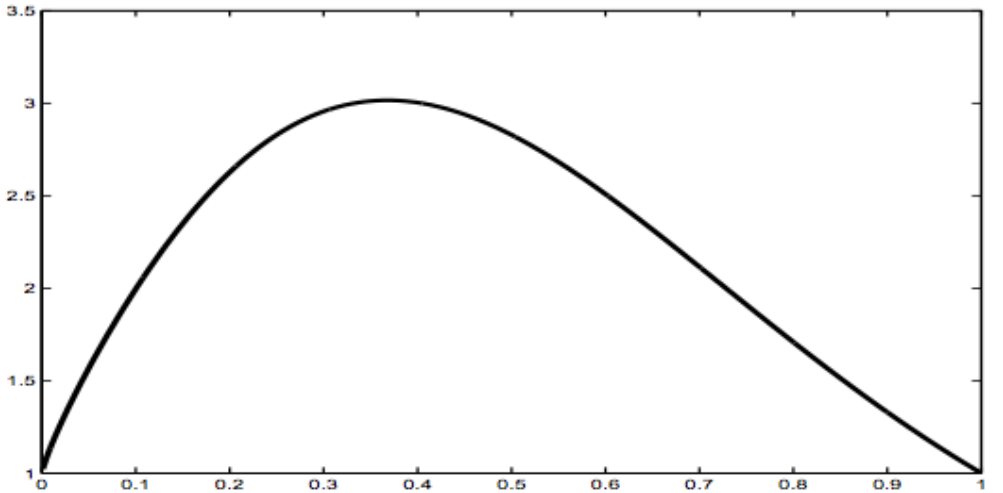


Figure 1: Relative increase in MSE due to debiasing the sequence.

Unbiased estimator: example

Simpson's rule

- Consider using the trapezoidal rule for estimating $\int_0^1 sin(\pi x)dx$.
- Here, X_k is the estimate of the integral with 2^{k+1} function evaluations on the grid $0, \Delta x, ..., 2^k \Delta x = 1$. Of course, $\lim_{k\to\infty} X_k = \int_0^1 sin(\pi x)dx$.
- $P(\mathcal{K} \geq k)$ is chosen as $p(1-p)^{k-s}$ with $p = \frac{3}{4}$ and $s = 2$ which gives an expected number of function evaluations of 7.

Unbiased estimator: example

Simpson's rule

- Variance of the Monte Carlo estimator: ≈ 0.0135.
- Variance of the unbiased estimator: $\approx 6.41 \times 10^{-6}$.
- More than a two thousand-fold gain in efficiency over crude Monte Carlo!

Let us go back to IWAEs

- An unbiased estimator of the log marginal likelihood can be computed as

$$\hat{\mathcal{L}} = \mathcal{L}_1^{IW} + \sum_{k=1}^{K} \frac{\mathcal{L}_k^{IW} - \mathcal{L}_{k-1}^{IW}}{P(\mathcal{K} \geq k)}, K \sim p(\mathcal{K}).$$

SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models , arXiv:2004.00353.

- Unbiasing the estimator may potentially introduce high variance.
  - This may not be an important issue with SGD.
  - Lower bias is prefered over lower variance.

Table 1: Test negative log-likelihood of the trained model, estimated using IWAE(k=5000). For SUMO, k refers to the expected number of computed terms.

	MNIST			OMNIGLOT		
Training Objective	k=5	k=15	k=50	k=5	k=15	k=50
ELBO (Burda et al., 2016)	86.47	—	86.35	107.62	—	107.80
IWAE (Burda et al., 2016)	85.54	—	84.78	106.12	—	104.67
ELBO (Our impl.)	85.97±0.01	85.99±0.05	85.88±0.07	106.79±0.08	106.98±0.19	106.84±0.13
IWAE (Our impl.)	85.28±0.01	84.89±0.03	84.50±0.02	104.96±0.04	104.53±0.05	**103.99±0.12**
JVI (Our impl.)	—	—	84.75±0.03	—	—	104.08±0.11
SUMO	**85.09±0.01**	**84.71±0.02**	**84.40±0.03**	**104.85±0.04**	**104.29±0.12**	**103.79±0.14**

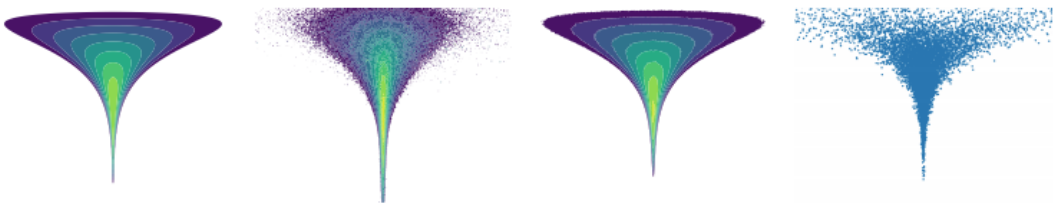


Figure 1: We trained latent variable models for posterior inference, which requires minimizing log probability under the model. Training with IWAE leads to optimizing for the bias while leaving the true model in an unstable state, whereas training with SUMO—though noisy—leads to convergence.

Credits: SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models , arXiv:2004.00353.

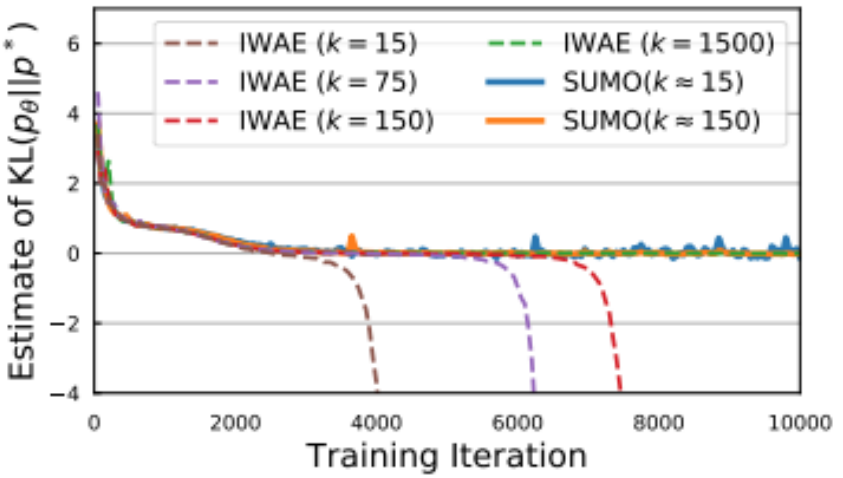


Figure 2: Training with reverse KL requires minimizing $\log p(x)$. SUMO estimates are unbiased and trains well, but minimizing the lower bound IWAE with small k leads to estimates of $-\infty$.

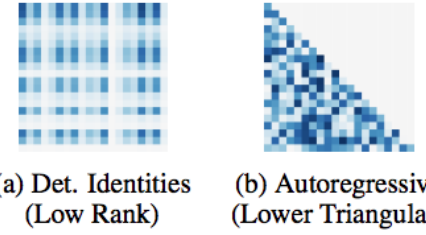
Credits: SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models , arXiv:2004.00353.

Other application

Residual Flows for Invertible Generative Modeling

Residual Flows for Invertible Generative Modeling

- Invertible residual networks (i-ResNets) are composed of simple transformations $y = x + g(x)$. These transformations are invertible (Banach fixed point theorem) if g is contractive, i.e. with Lipschitz constant strictly less than unity, which can be enforced using spectral normalization.



(a) Det. Identities (Low Rank) (b) Autoregressive (Lower Triangular)

(c) Coupling (Structured Sparsity) (d) **Unbiased Est.** (Free-form)

Figure 1: **Pathways to designing scalable normalizing flows** and their enforced Jacobian structure. Residual Flows fall under unbiased estimation with free-form Jacobian.

Residual Flows for Invertible Generative Modeling

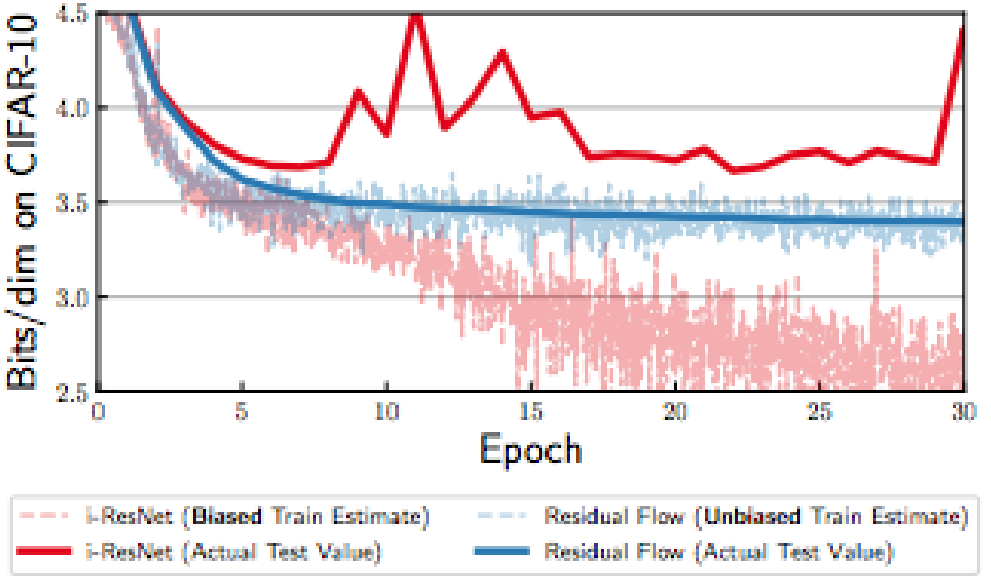
- Using the change of variables theorem allows to evaluate the log marginal likelihood:

$$\log p(y) = \log p(f(y)) + tr(\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} [J_g(y)]^k).$$

- Previous work used a fixed truncation to approximate the infinite serie.
  - This naïve approach has a bias that grows with the number of dimensions of y and the Lipschitz constant of g.
  - As such, the fixed truncation estimator requires a careful balance between bias and expressiveness, and cannot scale to higher dimensional data.

Residual Flows for Invertible Generative Modeling

- Without decoupling the objective and estimation bias, i-ResNets end up optimizing for the bias without improving the actual maximum likelihood objective.
- Residual flows use the Russian roulette estimator in order to produce an unbiased estimate of the infinite serie.
  - Competitive with state-of-the-art flow-based models.



Credits: Residual Flows for Invertible Generative Modeling, arXiv:1906.02735.

The end.