Perturbative Black Box Variational Inference

Introduction

- We establish a unified view on black box variational inference with generalized divergences as a form of biased importance sampling.
- We use these insights to construct a new variational bound with favorable properties with respect to a variance-bias trade-off.
- In our experiments, the resulting posterior covariances are closer to the true posterior, and likelihoods on heldout data are higher than with traditional black box variational inference.

Variational Inference as Biased Importance Sampling

\[ \log p(x) = \int p(x, z) dz \]

Black box variational inference (BBVI) estimates a lower bound \( \mathcal{L}(\lambda) \) on \( \log p(x) \) based on Monte Carlo samples from a variational distribution \( q_\lambda(z) \). Taking \( \mathcal{L}(\lambda) \) as a proxy for \( \log p(x) \) results in a bias and a sampling variance.

Variance-bias trade-off:

- High sampling variance, zero bias: importance sampling
  \[ p(x) = \mathbb{E}_{q_\lambda(z)} p(x, z) q_\lambda(z) \]
- Low sampling variance, large bias: black box variational inference
  \[ \log p(x) \geq \mathbb{E}_{q_\lambda(z)} \log p(x, z) q_\lambda(z) \]

A Unified View

Lower bound:

\[ p(x) \geq f(p(x)) \geq \mathbb{E}_{x \sim q_\lambda(z)} f\left( \frac{p(x, z)}{q_\lambda(z)} \right) = \mathcal{L}_f(\lambda) \]

For any concave function \( f \) with \( f(\xi) \leq \xi \):

- importance sampling: \( f = \text{id} \)
- (traditional) Kullback-Leibler BBVI: \( f = \log + \text{const.} \)
- BBVI with alpha-divergence: \( f(\alpha)(\xi) \propto \xi^{1-\alpha} \)

Observation: \( \xi = \frac{p(x|z)}{q_\lambda(z)} \) is highly peaked in \( z \)-space.

- If \( f(\xi) \) depends algebraically on \( \xi \) as in the alpha bound, then the sampling variance is high and reparameterization gradients are noisy.
- If \( f(\xi) \) depends only on \( \log \xi \), as in the KL bound, then the sampling variance is lower and reparameterization gradients are less noisy. However, the logarithm introduces a bias, i.e., the KL-bound is less tight.

Perturbative Black Box Variational Inference (PVI)

Aim: Lower bound with small bias and small gradient noise.

Taylor expansion of \( \frac{p(x|z)}{q_\lambda(z)} = \exp[\log p(x, z) - \log q_\lambda(z)] \) around a reference value of \( e^{-1} \):

\[ \mathcal{L}_{\text{PVI}}(\lambda) = e^{-V} \sum_{k=0}^{\infty} \frac{1}{k!} \mathbb{E}_{x \sim q_\lambda(z)} \left[ V_k + \log p(x, z) - \log q_\lambda(z) \right]^k \]

We show that \( p(x) \geq \mathcal{L}_{\text{PVI}}(\lambda) \) for all odd \( n \):

- \( n \rightarrow \infty \): importance sampling
- \( n = 1 \) (traditional) Kullback-Leibler VI
- \( n = 3 \) (proposed)

The proposed PVI bound is tighter than the traditional Kullback-Leibler bound. At the same time, the bound has smaller gradient noise than the alpha bound, leading to faster convergence of stochastic gradient ascent (see experiments).

Experiments

- Gaussian process regression with synthetic data
- Gaussian process classification with the UCI data sets
- Variational autoencoder (VAE)

We train a VAE on subsets of different sizes of the MNIST data set. Our proposed PVI method reaches higher predictive likelihoods when the data set is small.