Perturbative Black Box Variational Inference

Cheng Zhang*[1], Robert Bamler*[1], Manfred Opper[2], Stephan Mandt*[1]
1. Disney Research, 2. Technical University of Berlin
* Equal contribution

Disney Research

Technische Universität Berlin

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.

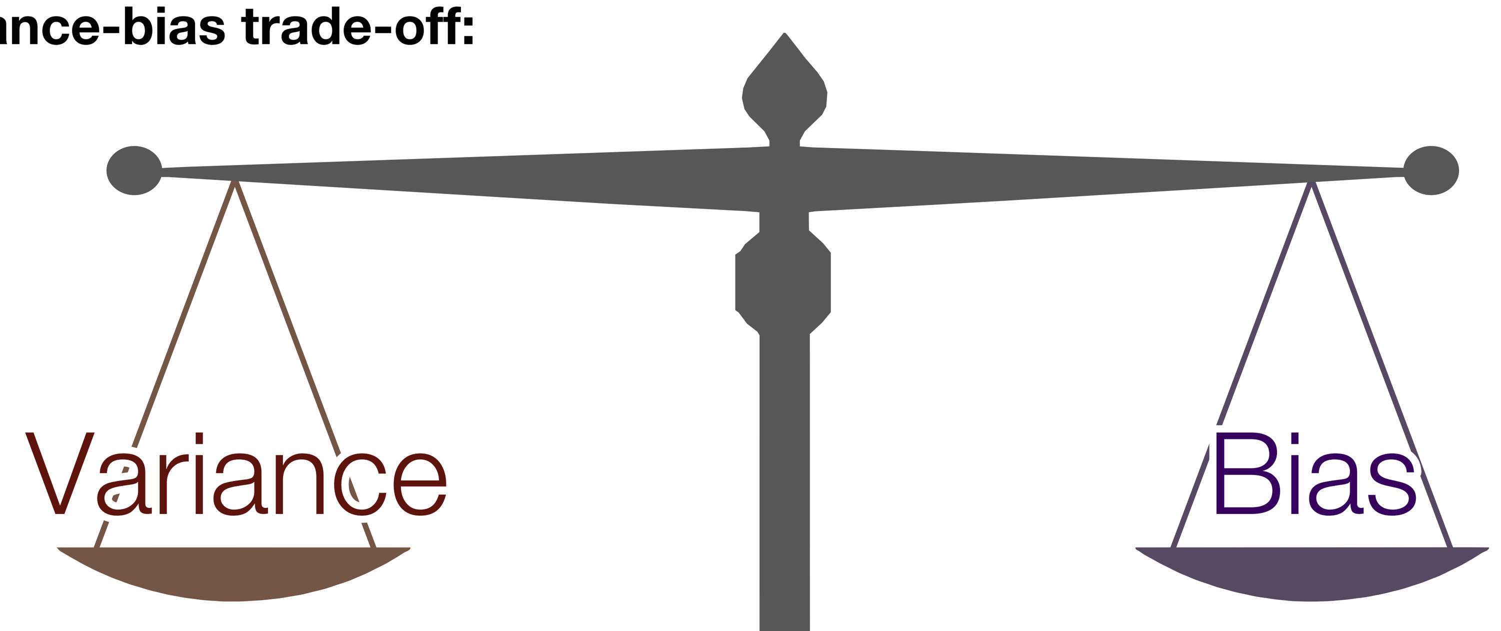
Variational Inference as Biased Importance Sampling

Model with joint p.d.f. $p(\mathbf{x}, \mathbf{z})$ (x: observed, z: latent) → **seek** posterior $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})}$

Problem: intractable denominator $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$

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

Variance-bias trade-off:



High sampling variance, zero bias:
importance sampling

$$p(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\lambda}\Bigg[\underbrace{\frac{p(\mathbf{x},\mathbf{z})}{q_\lambda(\mathbf{z})}}_{\text{importance ratio (highly peaked)}}\Bigg]$$

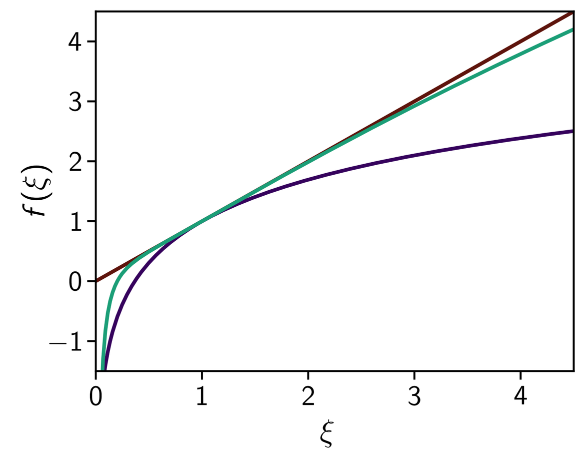
Large bias, low sampling variance:
black box variational inference

$$\log p(\mathbf{x}) \geq \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\lambda}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\lambda(\mathbf{z})}\right]}_{\text{evidence lower bound (ELBO)}}$$

A Unified View

Lower bound: $p(\mathbf{x}) \;\geq\; f(p(\mathbf{x})) \;\geq\; \mathbb{E}_{\mathbf{z}\sim q_\lambda}\left[f\left(\frac{p(\mathbf{x},\mathbf{z})}{q_\lambda(\mathbf{z})}\right)\right] \;\equiv\; \mathcal{L}_f(\lambda)$

For any concave function f with $f(\xi) \leq \xi$.
→ importance sampling: $f = \mathrm{id}$
→ (traditional) Kullback-Leibler BBVI: $f = \log + \mathrm{const.}$
→ BBVI with alpha-divergence: $f^{(\alpha)}(\xi) \propto \xi^{1-\alpha}$



Observation: $\xi \equiv \frac{p(\mathbf{x},\mathbf{z})}{q_\lambda(\mathbf{z})}$ is highly peaked in **z**-space.
⇒ If $f(\xi)$ depends algebraically on ξ, as in the alpha bound, then the sampling variance is high and reparemeterization gradients are noisy.
⇒ If $f(\xi)$ depends only on $\log \xi$, as in the KL bound, then the sampling variance is lower and reparemeterization gradients are less noisy. However, the logarithm introduces a bias, i.e., the KL-bound is less tight.

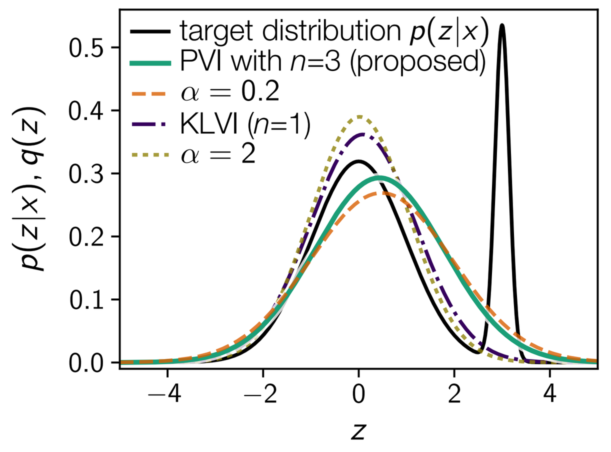
Perturbative Black Box Variational Inference (PVI)

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

Taylor expansion of $\frac{p(\mathbf{x},\mathbf{z})}{q_\lambda(\mathbf{z})} = \exp[\log p(\mathbf{x},\mathbf{z}) - \log q_\lambda(\mathbf{z})]$ around a reference value of e^{-V_0}:

$$\mathcal{L}_{f^{(n)}_{V_0}}(\lambda) \;\equiv\; e^{-V_0} \sum_{k=0}^{n} \frac{1}{k!} \mathbb{E}_{\mathbf{z}\sim q_\lambda}\left[\left(V_0 + \log p(\mathbf{x},\mathbf{z}) - \log q_\lambda(\mathbf{z})\right)^k\right]$$

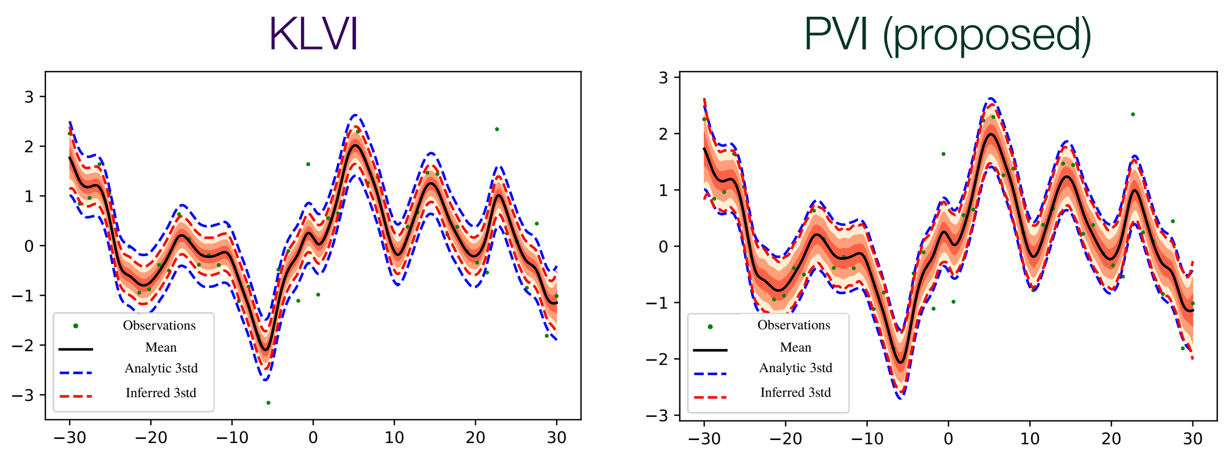
We show that $p(\mathbf{x}) \geq \mathcal{L}_{f^{(n)}_{V_0}}(\lambda)$ for all odd n.
→ $n \to \infty$: importance sampling
→ $n = 1$: (traditional) Kullback-Leibler VI
→ $n = 3$: proposed



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).

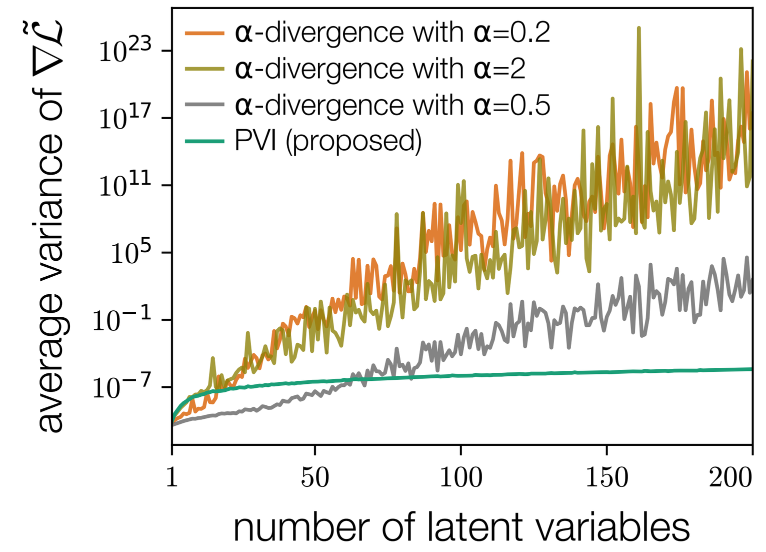
Experiments

Gaussian process regression with synthetic data



	average posterior variance
Analytic	0.0415
KLVI	0.0176
PVI (proposed)	0.0355

With increasing dimensionality of the latent space, the **gradient noise** grows exponentially for the alpha-VI bound, and only algebraically for the proposed PVI bound (green line).

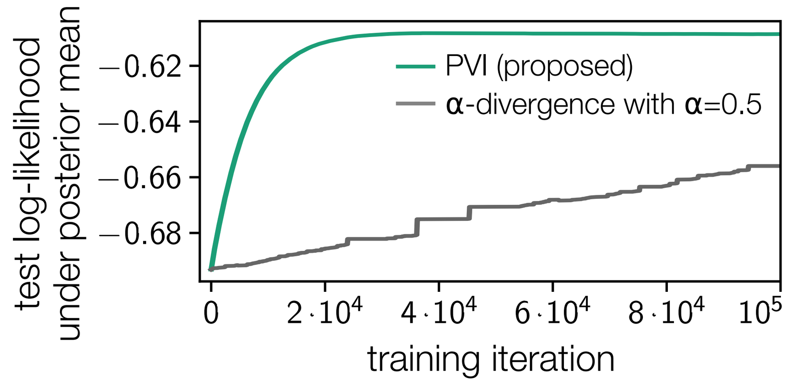


Gaussian process classification with the UCI data sets

We consistently obtain **lower error rates** on held out test sets.

Data set	Crab	Pima	Heart	Sonar
KLVI	0.22	0.245	0.148	0.212
PVI (proposed)	**0.11**	**0.240**	**0.133**	**0.173**

The proposed PVI **converges faster** than BBVI with an alpha divergence. This may be explained by the smaller gradient noise, see above.



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**.