Perturbative Black Box Variational Inference

Robert Bamler[*]
Disney Research
Pittsburgh, USA

Cheng Zhang[*]
Disney Research
Pittsburgh, USA

Manfred Opper
TU Berlin
Berlin, Germany

Stephan Mandt[*]
Disney Research
Pittsburgh, USA

firstname.lastname@{disneyresearch.com, tu-berlin.de}

Abstract

Black box variational inference (BBVI) with reparameterization gradients triggered the exploration of divergence measures other than the Kullback-Leibler (KL) divergence, such as alpha divergences. These divergences can be tuned to be more mass-covering (preventing overfitting in complex models), but are also often harder to optimize using Monte-Carlo gradients. In this paper, we view BBVI with generalized divergences as a form of biased importance sampling. The choice of divergence determines a bias-variance tradeoff between the tightness of the bound (low bias) and the variance of its gradient estimators. Drawing on variational perturbation theory of statistical physics, we use these insights to construct a new variational bound which is tighter than the KL bound and more mass covering. Compared to alpha-divergences, its reparameterization gradients have a lower variance. We show in several experiments on Gaussian Processes and Variational Autoencoders that the resulting posterior covariances are closer to the true posterior and lead to higher likelihoods on held-out data.

1 Introduction

Variational inference (VI) (Jordan et al., 1999) provides a way to convert Bayesian inference to optimization by minimizing a divergence measure. Recent advances of VI have been devoted to scalability (Hoffman et al., 2013; Ranganath et al., 2014), divergence measures (Minka, 2005; Li and Turner, 2016; Hernandez-Lobato et al., 2016), and structured variational distributions (Hoffman and Blei, 2014; Ranganath et al., 2016).

While traditional stochastic variational inference (SVI) (Hoffman et al., 2013) was limited to conditionally conjugate Bayesian models, black box variational inference (Ranganath et al., 2014) (BBVI) enables SVI on a large class of models by expressing the gradient as an expectation, and estimating it by Monte-Carlo sampling. A similar version of this method relies on reparametrized gradients and has a lower variance (Salimans and Knowles, 2013; Kingma and Welling, 2014; Rezende et al., 2014; Ruiz et al., 2016). BBVI paved the way for approximate inference in complex and deep generative models (Kingma and Welling, 2014; Rezende et al., 2014; Ranganath et al., 2015; Bamler and Mandt, 2017).

Before the advent of BBVI, divergence measures other than the KL divergence had been of limited practical use due to their complexity in both mathematical derivation and computation (Minka, 2005), but have since then been revisited. Alpha-divergences (Hernandez-Lobato et al., 2016; Dieng et al., 2017; Li and Turner, 2016) achieve a better matching of the variational distribution to different regions of the posterior and may be tuned to either fit its dominant mode or to cover its entire support. The problem with reparameterizing the gradient of the alpha-divergence is, however, that the resulting gradient estimates have large variances. It is therefore desirable to find other divergence measures with low-variance reparameterization gradients.

---

[*]Equal contribution

In this paper, we develop a new variational bound based on concepts from perturbation theory of statistical physics. Our contributions are as follows.

- We establish a view on black box variational inference with generalized divergences as a form of *biased importance sampling* (Section 3.1). The choice of divergence allows us to trade-off between a low-variance stochastic gradient and loose bound, and a tight variational bound with higher-variance Monte-Carlo gradients. As we explain below, importance sampling and point estimation are at opposite sides of this spectrum.

- We use these insights to construct a new variational bound with favorable properties. Based on perturbation theory of statistical physics, we derive a new variational bound (Section 3.2) which has a small variance and which is tighter compared to the KL-bound. The bound is easy to optimize and contains perturbative corrections around the mean-field solution.

In our experiments (Section 4), we find that the new bound is more mass-covering than the KL-bound, but its variance is much smaller than alpha divergences which have a similar mass-covering effect.

2 Related work

Our approach is related to BBVI, VI with generalized divergences, and variational perturbation theory. We thus briefly discuss related work in these three directions.

Black box variational inference (BBVI). BBVI has already been addressed in the introduction (Salimans and Knowles, 2013; Kingma and Welling, 2014; Rezende et al., 2014; Ranganath et al., 2014; Ruiz et al., 2016); it enables variational inference for many models. Our work builds upon BBVI in that it makes a large class of new divergence measures between the posterior and the approximating distribution tractable. Depending on the divergence measure, BBVI may suffer from high-variance stochastic gradients. This is a practical problem that we aim to improve in this paper.

Generalized divergences measures. Our work connects to generalized information-theoretic divergences (Amari, 2012). Minka (2005) introduced a broad class of divergences for variational inference, including alpha-divergences. Most of these divergences have been intractable in large-scale applications until the advent of BBVI. In this context, alpha-divergences were first suggested by Hernandez-Lobato et al. (2016) for local divergence minimization, and later for global minimization by Li and Turner (2016) and Dieng et al. (2017). As we show in this paper, alpha-divergences have the disadvantage of inducing high-variance gradients, since the ratio between posterior and variational distribution enters polynomially instead of logarithmically. In contrast, our approach leads to a more stable inference scheme in high dimensions.

Variational perturbation theory. Our work also relates to variational perturbation theory. Perturbation theory refers to a set of methods that aim to truncate a typically divergent power series to a convergent series. In machine learning, these approaches have been addressed from an information-theoretic perspective by Tanaka (1999, 2000). Thouless-Anderson-Palmer (TAP) equations (Thouless et al., 1977) are a special form of second-order perturbation theory and were originally developed in statistical physics. TAP equations are aimed at including perturbative corrections to the mean-field solution of Ising models. They have been adopted into Bayesian inference in (Plefka, 1982) and were advanced by many authors (Kappen and Wiegerinck, 2001; Paquet et al., 2009; Opper et al., 2013; Opper, 2015). In variational inference, perturbation theory yields extra terms to the mean-field variational objective which are difficult to calculate analytically. This may be a reason why the methods discussed are not widely adopted by practitioners. In this paper, we emphasize the ease of including perturbative corrections in a black box variational inference framework. Furthermore, in contrast to earlier formulations, our approach yields a strict lower bound to the marginal likelihood which can be conveniently optimized. Our approach is different from traditional variational perpetuation formulation, because variational perturbation theory (Kleinert, 2009) generally does not result in a bound.

3 Method

In this section, we present our main contributions. We first present our view of black box variational inference (BBVI) as a form of biased importance sampling in Section 3.1. With this view, we bridge the gap between variational inference and importance sampling. In Section 3.2, we introduce our new variational bound, and analyze its properties further in Section 3.3.
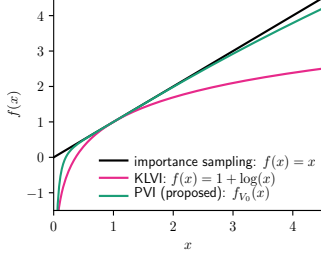
Figure 1: Different choices for f in Eq. 3. KLVI corresponds to $f(x) = \log(x) + \text{const.}$ (red), and importance sampling to $f(x) = x$ (black). Our proposed PVI bound uses f_{V_0} (green) as specified in Eq. 6, which lies between KLVI and importance sampling (we set $V_0 = 0$ for PVI here).
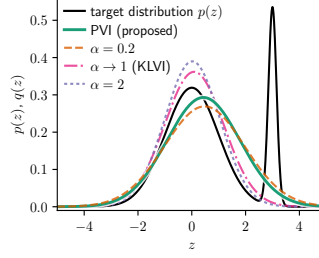
Figure 2: Behavior of different VI methods on fitting a univariate Gaussian to a bimodal target distribution (black). PVI (proposed, green) covers more of the mass of the entire distribution than the traditional KLVI (red). Alpha-VI is mode seeking for large α and mass covering for smaller α.
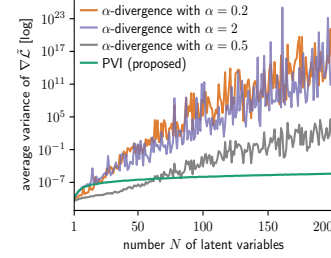
Figure 3: Sampling variance of the stochastic gradient (averaged over its components) in the optimum for alpha-divergences (orange, purple, gray) and the proposed PVI (green). The variance grows exponentially as a function of the latent dimension for α-VI, and only algebraically for PVI.

3.1 Black Box Variational Inference as Biased Importance Sampling

Consider a probabilistic model with data $\mathbf{x}$, latent variables $\mathbf{z}$, and joint distribution $p(\mathbf{x}, \mathbf{z})$. We are interested in the posterior distribution over the latent variables, $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{x}, \mathbf{z})/p(\mathbf{x})$. This involves the intractable marginal likelihood $p(\mathbf{x})$. In variational inference (Jordan et al., 1999), we instead minimize a divergence measure between a variational distribution $q(\mathbf{z}; \lambda)$ and the posterior. Here, λ are parameters of the variational distribution, and the task is to find the parameters λ^* that minimize the distance to the posterior. This is equivalent to maximizing a lower bound to the marginal likelihood.

We call the difference between the log variational distribution and the log joint distribution the *interaction energy*,

$$V(\mathbf{z}; \lambda) = \log q(\mathbf{z}; \lambda) - \log p(\mathbf{x}, \mathbf{z}). \tag{1}$$

We use V or $V(\mathbf{z})$ interchangeably to denote $V(\mathbf{z}; \lambda)$, when more convenient. Using this notation, the marginal likelihood is

$$p(\mathbf{x}) = \mathop{\mathbb{E}}_{q(\mathbf{z})} [e^{-V(\mathbf{z})}], \tag{2}$$

We call $e^{-V(\mathbf{z})} = p(\mathbf{x}, \mathbf{z})/q(\mathbf{z})$ the *importance weight*, since sampling from $q(\mathbf{z})$ to estimate the right-hand side of Eq. 2 is equivalent to importance sampling. As this is inefficient in high dimensions, we resort to variational inference. To this end, let $f(\cdot)$ be any concave function defined on the positive reals. We assume furthermore that for all $x > 0$, we have $f(x) \leq x$. Applying Jensen's inequality, we can lower-bound the marginal likelihood,

$$p(\mathbf{x}) \geq f(p(\mathbf{x})) \geq \mathop{\mathbb{E}}_{q(\mathbf{z})} [f(e^{-V(\mathbf{z}; \lambda)})] \equiv \mathcal{L}_f(\lambda). \tag{3}$$

We call this bound the f-ELBO, in comparison to the evidence lower bound (ELBO) used in Kullback-Leibler variational inference (KLVI). Figure 1 shows exemplary choices of f. We maximize $\mathcal{L}_f(\lambda)$ using reparameterization gradients, where the bound is not computed analytically, but rather its gradients are estimated by sampling from $q(\mathbf{z})$ (Kingma and Welling, 2014). This leads to a stochastic gradient descent scheme, where the noise is a result of the Monte-Carlo estimation of the gradients.

Our approach builds on the insight that black box variational inference is a type of biased importance sampling, where we estimate a lower bound of the marginal likelihood by sampling from a proposal distribution, iteratively improving this distribution. The approach is biased, since we do not estimate the exact marginal likelihood but only a lower bound to this quantity. As we argue below, the introduced bias allows us to estimate this bound more easily, because we decrease the variance of this estimator. The choice of the function f thereby trades-off between bias and variance in the following way:

- For $f = id$ being the identity, we obtain *importance sampling*. (See the black line in Figure 1). In this case, Eq. 3 does not depend on the variational parameters, so there is nothing to

optimize and we can directly sample from any proposal distribution q. Since the expectation under q of the importance weight $e^{-V(\mathbf{z})}$ gives the exact marginal likelihood, there is no bias. For a large number of latent variables, the importance weight $e^{-V(\mathbf{z})}$ becomes tightly peaked around the minimum of the interaction energy V, resulting in a very high variance of this estimator. Importance sampling is therefore on one extreme end of the bias-variance spectrum.

- For $f = \log$, we obtain the familiar *Kullback-Leibler (KL) bound*. (See the pink line in Figure 1; here we add a constant of 1 for comparison, which does not influence the optimization). Since $f(e^{-V(\mathbf{z})}) = -V(\mathbf{z})$, the bound is

$$\mathcal{L}_{KL}(\lambda) = \mathop{\mathbb{E}}_{q(\mathbf{z})}[-V(\mathbf{z})] = \mathop{\mathbb{E}}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]. \qquad (4)$$

The Monte-Carlo expectation of $\mathbb{E}_q[-V]$ has a much smaller variance than $\mathbb{E}_q[e^{-V}]$, implying efficient learning (Bottou, 2010). However, by replacing e^{-V} with $-V$ we introduce a bias. We can further trade-off less variance for more bias by dropping the entropy term. A flexible enough variational distribution will shrink to zero variance, which completely eliminates the noise. This is equivalent to point-estimation, and is at the opposite end of the bias-variance spectrum.

- Now, consider any f which is between the logarithm and the identity, for example, the green line in Figure 1 (this is the bound that we will propose in Section 3.2). The more similar f is to the identity, the less biased is our estimate of the marginal likelihood, but the larger the variance. Conversely, the more f behaves like the logarithm, the easier it is to estimate $f(e^{-V(\mathbf{z})})$ by sampling, while at the same time the bias grows.

One example of alternative divergences to the KL divergence that have been discussed in the literature are alpha-divergences (Minka, 2005; Hernandez-Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017). Up to a constant, they correspond to the following choice of f:

$$f^{(\alpha)}(e^{-V}) \propto e^{-(1-\alpha)V}. \qquad (5)$$

The parameter α determines the distance to the importance sampling case ($\alpha = 0$). As α approaches 1 from below, this bound leads to a better-behaved estimation problem of the Monte-Carlo gradient. However, unless taking the limit of $\alpha \to 1$ (where the objective becomes the KL-bound), V still enters exponentially in the bound. As we show, this leads to a high variance of the gradient estimator in high dimensions (see Figure 3 discussed below). The alpha-divergence bound is therefore similarly as hard to estimate as the marginal likelihood in importance sampling.

Our analysis relies on the observation that expectations of exponentials in V are difficult to estimate, and expectations of polynomials in V are easy to estimate. We derive a new variational bound which is a polynomial in V and at the same time results in a tighter bound than the KL-bound.

3.2 Perturbative Black Box Variational Inference

We now propose a new bound based on the considerations outlined above. This bound is tighter and more mass-covering than the KL bound, and its gradients are easy to estimate via the reparameterization approach. Since V will never be exponentiated, the reparameterization gradients have a lower variance than for unbiased alpha divergences.

We construct a function f with a free parameter V_0, where V enters only polynomially:

$$f_{V_0}(e^{-V}) = e^{-V_0}\left(1 + (V_0 - V) + \frac{1}{2}(V_0 - V)^2 + \frac{1}{6}(V_0 - V)^3\right). \qquad (6)$$

This function f is a third-order Taylor expansion of the importance weight e^{-V} in the interaction energy V around the reference energy $V_0 \in \mathbb{R}$. We introduce V_0 so that any additive constant in the log-joint distribution can be absorbed into V_0. It is easy to see that $f_{V_0}(\cdot)$ is concave for any choice of V_0, and that its graph lies below the identity function (see proof in Section 3.3). Thus, $f_{V_0}(\cdot)$ meets all conditions for Eq. 3 to hold, and we have $p(\mathbf{x}) \geq \mathcal{L}_{f_{V_0}}(\lambda)$ for all V_0 and λ. We find the optimal values for the reference energy V_0 and the variational parameters λ by simultaneously maximizing $\mathcal{L}_{f_{V_0}}(\lambda)$ over both V_0 and λ. We call $\mathcal{L}_{f_{V_0}}(\lambda)$ the perturbative variational lower bound and name the method *perturbative variational inference* (PVI). The resulting variational lower bound is

$$\mathcal{L}_{f_{V_0}}(\lambda) = e^{-V_0} \mathop{\mathbb{E}}_q \left[1 + (V_0 - V) + \frac{1}{2}(V_0 - V)^2 + \frac{1}{6}(V_0 - V)^3\right]. \qquad (7)$$

Using the reparameterization gradient representation, we can easily take gradients with respect to the variational parameters. The fact that V enters only polynomially and not exponentially leads to a low-variance stochastic gradient. This is in contrast to the alpha-divergence bound (Eq. 5), where V enters exponentially.

As a technical note, the factor e^{-V_0} is not a function of the latent variables and does not contribute to the variance, however, it may lead to numerical underflow or overflow in large models. This can be easily mitigated by considering the surrogate objective $\tilde{\mathcal{L}}_{f_{V_0}}(\lambda) \equiv e^{V_0}\mathcal{L}_{f_{V_0}}(\lambda)$. The gradients with respect to λ of $\mathcal{L}_{f_{V_0}}(\lambda)$ and $\tilde{\mathcal{L}}_{f_{V_0}}(\lambda)$ are equal up to a positive prefactor, so we can replace the former with the latter in gradient descent. The gradient with respect to V_0 is $\partial \mathcal{L}_{f_{V_0}}(\lambda)/\partial V_0 \propto \partial \tilde{\mathcal{L}}_{f_{V_0}}(\lambda)/\partial V_0 - \tilde{\mathcal{L}}_{f_{V_0}}(\lambda)$. Using the surrogate $\tilde{\mathcal{L}}_{f_{V_0}}(\lambda)$ avoids numerical underflow or overflow, as well as exponentially increasing or decreasing gradients.

Figure 1 shows several choices for f that correspond to different divergences. The red curve shows the logarithm, corresponding to the typical KL divergence bound, while the black line shows the importance sampling case. Our function f_{V_0} corresponds to the green curve. We see that it lies between the importance sampling case and the KL-divergence case and therefore has a lower bias than KLVI. Note that in this example we have set the reference energy V_0 to zero.

In Figure 2, we fit a Gaussian distribution to a one-dimensional bimodal target distribution (black line), using different divergences. Compared to KLVI (pink line), alpha-divergences are more mode-seeking (purple line) for large values of α, and more mass-covering (orange line) for small α (Li and Turner, 2016). Our PVI bound (green line) achieves a similar mass-covering effect as in alpha-divergences, but with associated low-variance reparameterization gradients. This is also seen in Figure 3 which compares the gradient variances of alpha-VI and PVI as a function of dimensions.

3.3 Theoretical Considerations

We conclude the presentation of the PVI bound by exploring several aspects theoretically. We generalize the perturbative expansion to all odd orders and recover the KL-bound as the first order. We also show that the proposed PVI method does not result in a trivial bound. In addition, we show in the supplement that PVI implicitly minimizes a valid divergence from q to the true posterior.

Generalization to all odd orders. Eq. 6 defines $f_{V_0}(e^{-V})$ as the third order Taylor expansion of e^{-V} in V around V_0. We generalize this definition to a general order n, and define for $x > 0$,

$$f_{V_0}^{(n)}(x) \equiv e^{-V_0}\sum_{k=0}^{n}\frac{(V_0 + \log x)^k}{k!} \tag{8}$$

This includes Eq. 6 in the case $n = 3$. It turns out that $\mathcal{L}_{f_{V_0}^{(n)}}(\lambda)$ is a lower bound for all odd n, because $f_{V_0}^{(n)}$ is concave and lies below the identity function for all x. To see this, note that the second derivative $\partial^2 f_{V_0}^{(n)}(x)/\partial x^2 = -e^{-V_0}(V_0 + \log x)^{n-1}/((n-1)!\,x^2)$ is non-positive everywhere for odd n. Therefore, the function is concave. Next, consider the function $g(x) = f_{V_0}^{(n)}(x) - x$, which has a stationary point at $x = x_0 \equiv e^{-V_0}$. Since g is also concave, x_0 is a global maximum, and thus $g(x) \leq g(x_0) = 0$ for all x, implying that $f_{V_0}^{(n)}(x) \leq x$. Thus, for odd n, the function $f_{V_0}^{(n)}$ satisfies all requirements for Eq. 3, and $\mathcal{L}_{f_{V_0}^{(n)}}(\lambda) \equiv \mathbb{E}_q[f_{V_0}^{(n)}(e^{-V})]$ is a lower bound on the model evidence. Note that an even order n does not lead to a valid concave function.

First order: KLVI. For $n = 1$, the lower bound is

$$\mathcal{L}_{f_{V_0}^{(1)}}(\lambda) = e^{-V_0}(1 + V_0 - \mathbb{E}_q[V]) = e^{-V_0}(1 + V_0 + \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \lambda)]) \tag{9}$$

Maximizing this bound over λ is equivalent to maximizing the evidence lower bound (ELBO) of traditional KLVI. Apart from a positive prefactor, the reference energy V_0 has no influence on the gradient of the first-order lower bound with respect to λ. This is why one can safely ignore V_0 in traditional KLVI. As a matter of fact, optimizing over V_0 results here exactly in the KL bound.

Third order: Non-triviality of the bound. When we go beyond a first-order Taylor expansion, the lower bound is no longer invariant under shifts of the interaction energy V, and we can no longer ignore the reference energy V_0. For $n = 3$, we obtain the proposed PVI lower bound, see Eq. 7. Since

| (a) KLVI | (b) MF PVI |
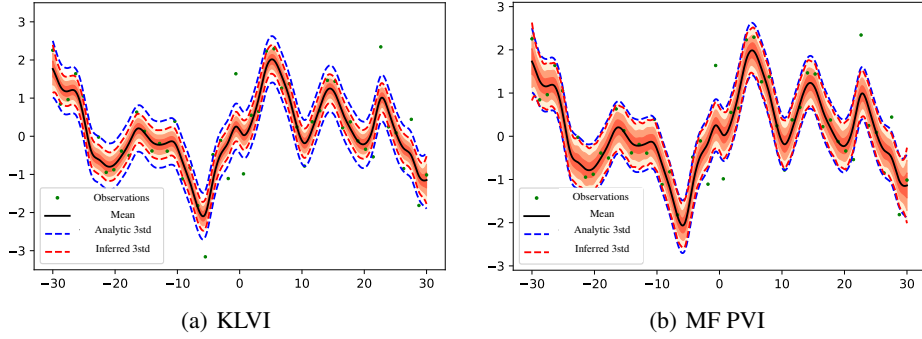
Figure 4: Gaussian process regression on synthetic data (green dots). Three standard deviations are shown in varying shades of oranges. The blue dashed lines show three standard deviations of the true posterior. The red dashed lines show the inferred three standard deviations using KLVI (a) and PVI (b). We can see that the results from our proposed PVI are close to the analytic solution while traditional KLVI underestimates the variances.

Method	Avg variances
Analytic	0.0415
KLVI	0.0176
MF PVI	0.0355

Table 1: Average variances across training examples in the synthetic data experiment. The closer to the analytic solution, the better.

Dataset	Crab	Pima	Heart	Sonar
KLVI	0.22	0.245	0.148	0.212
MF PVI	**0.11**	**0.240**	**0.1333**	**0.1731**

Table 2: The error rate of GP classification. The lower the better. Our proposed PVI consistently obtains better classification results.

the model evidence $p(\mathbf{x})$ is always positive, a lower bound would be useless if it was negative. We show that once the inference algorithm is converged, the bound at the optimum is always positive.

At the optimum, all gradients vanish. By setting the derivative with respect to V_0 of the right-hand side of Eq. 7 to zero we find that $\mathbb{E}_{q^*}[(V_0^* - V)^3] = 0$, where $q^* \equiv q(\mathbf{z}; \lambda^*)$ and V_0^* denote the variational distribution and the reference energy at the optimum, respectively. This means that the third-order term of the lower bound vanishes at the optimum. We rewrite the remaining terms as follows, which shows that the bound at the optimum is always positive:

$$\mathcal{L}_{V_0^*}(\lambda^*) = e^{-V_0^*} \mathbb{E}_{q^*}\left[1 + (V_0^* - V) + \frac{1}{2}(V_0^* - V)^2\right] = \frac{e^{-V_0^*}}{2}\left(\mathbb{E}_{q^*}\left[(1 + V_0^* - V)^2\right] + 1\right) > 0.$$

4 Experiments

We evaluate PVI with different models. First we investigate the behavior of the new bound in a controlled setup of Gaussian processes on synthetic data (Section 4.1). We then evaluate PVI based on a classification task using Gaussian processes classifiers, where we use data from the UCI machine learning repository (Section 4.2). This is a Bayesian non-conjugate setup where black box inference is required. Finally, we use an experiment with the variational autoencoder (VAE) to explore our approach on a deep generative model (Section 4.3). This experiment is carried out on MNIST data. Across all the experiments, PVI demonstrates advantages based on different metrics.

4.1 GP Regression on Synthetic Data

In this section, we inspect the inference behaviors using a synthetic dataset with Gaussian processes (GP). We generate the data according to a Gaussian noise distribution centered around a mixture of sinusoids and sample 50 data points (green dots in Figure 4). We then use a GP to model the data, thus assuming the generative process $f \sim \mathcal{GP}(0, \Lambda)$ and $y_i \sim \mathcal{N}(f_i, \epsilon)$.

We first compute an analytic solution of the posterior of the GP, (three standard deviations shown in blue dashed lines) and compare it to approximate posteriors obtained by KLVI (Figure 4 (a)) and the proposed PVI (Figure 4 (b)). The results from PVI are almost identical to the analytic solution. In contrast, KLVI underestimates the posterior variance. This is consistent with Table 1, which shows the average diagonal variances. PVI results are much closer to the exact posterior variances.

4.2 Gaussian Process Classification

We evaluate the performance of PVI and KLVI on a GP classification task. Since the model is non-conjugate, no analytical baseline is available in this case. We model the data with the following generative process:

$$f \sim \mathcal{GP}(0, K), \quad z_i = \sigma(f_i), \quad y_i \sim Bern(z_i).$$

Above, K is the GP kernel, σ indicates the sigmoid function, and $Bern$ indicates the Bernoulli distribution. We furthermore use the Matern 32 kernel,

$$K_{ij} = k(x_i, x_j) = s^2(1 + \frac{\sqrt{3}\, r_{ij}}{l}) \exp(-\frac{\sqrt{3}\, r_{ij}}{l}), \quad r_{ij} = \sqrt{(x_i - x_j)^T (x_i - x_j)}.$$

Data. We use four datasets from the UCI machine learning repository, suitable for binary classification: Crab (200 datapoints), Pima (768 datapoints), Heart (270 datapoints), and Sonar (208 datapoints). We randomly split each of the datasets into two halves. One half is used for training and the other half is used for testing. We set the hyper parameters $s = 1$ and $l = \sqrt{D}/2$ throughout all experiments, where D is the dimensionality of input x.

Table 2 shows the classification performance (error rate) for these data sets. Our proposed PVI consistently performs better than the traditional KLVI.

Convergence speed comparison. We also carry out a comparison in terms of speed of convergence, focusing on PVI and α-divergence VI. Our results indicate that the smaller variance of the reparameterization gradient leads to faster convergence of the optimization algorithm.

We train the GP classifier from Section 4.2 on the Sonar UCI data set using a constant learning rate. Figure 5 shows the test log-likelihood under the posterior mean as a function of training iterations. We split the data set into equally sized training, validation, and test sets. We then tune the learning rate and the number of Monte Carlo samples per gradient step to obtain optimal performance on the validation set after minimizing the α-divergence with a fixed budget of random samples. We use $\alpha = 0.5$ here; smaller values of α lead to even slower convergence. We optimize the PVI lower bound using the same learning rate and number of Monte Carlo samples. The final test error rate is 22% on an approximately balanced data set. PVI converges an order of magnitude faster.

Figure 3 in Section 3 provides more insight in the scaling of the gradient variance. Here, we fit GP regression models on synthetically generated data by maximizing the PVI lower bound and the α-VI lower bound with $\alpha \in \{0.2, 0.5, 2\}$. We generate a separate synthetic data set for each $N \in \{1, \ldots, 200\}$ by drawing N random data points around a sinusoidal curve. For each N, we fit a one-dimensional GP regression with PVI and α-VI, respectively, using the same data set for both methods. The variational distribution is a fully factorized Gaussian with N latent variables. After convergence, we estimate the sampling variance of the gradient of each lower bound with respect to the posterior mean. We calculate the empirical variance of the gradient based on 10^5 samples from q, and we average over the N coordinates. Figure 3 shows the average sampling variance as a function of N on a logarithmic scale. The variance of the gradient of the α-VI bound grows exponentially in the number of latent variables. By contrast, we find only algebraic growth for PVI.

4.3 Variational Autoencoder

We experiment on Variational Autoencoders (VAEs), and we compare the PVI and the KLVI bound in terms of predictive likelihoods on held-out data (Kingma and Welling, 2014). Autoencoders compress unlabeled training data into low-dimensional representations by fitting it to an encoder-decoder model that maps the data to itself. These models are prone to learning the identity function when the hyperparameters are not carefully tuned, or when the network is too expressive, especially for a moderately sized training set. VAEs are designed to partially avoid this problem by estimating the uncertainty that is associated with each data point in the latent space. It is therefore important that the inference method does not underestimate posterior variances. We show that, for small datasets, training a VAE by maximizing the PVI lower bound leads to higher predictive likelihoods than maximizing the KLVI lower bound.

We train the VAE on the MNIST dataset of handwritten digits (LeCun et al., 1998). We build on the publicly available implementation by Burda et al. (2016) and also use the same architecture and hyperparamters, with $L = 2$ stochastic layers and $K = 5$ samples from the variational distribution per gradient step. The model has 100 latent units in the first stochastic layer and 50 latent units in the second stochastic layer.
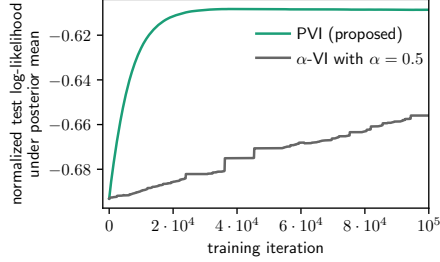
Figure 5: Test log-likelihood (normalized by the number of test points) as a function of training iterations using GP classification on the Sonar dataset. PVI converges faster than α-VI even though we tuned the number of Monte Carlo samples per training step (100) and the constant learning rate (10^{-5}) so as to maximize the performance of α-VI on a validation set.
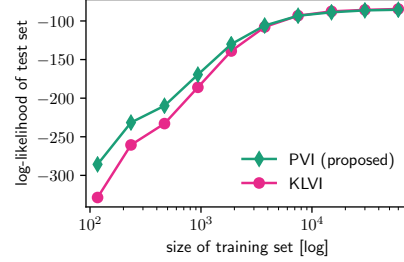
Figure 6: Predictive likelihood of a VAE trained on different sizes of the data. The training data are randomly sampled subsets of the MNIST training set. The higher value the better. Our proposed PVI method outperforms KLVI mainly when the size of the training dataset is small. The fewer the training data, the more advantage PVI obtains.

The VAE model factorizes over all data points. We train it by stochastically maximizing the sum of the PVI lower bounds for all data points using a minibatch size of 20. The VAE amortizes the gradient signal across data points by training inference networks. The inference networks express the mean and variance of the variational distribution as a function of the data point. We add an additional inference network that learns the mapping from a data point to the reference energy V_0. Here, we use a network with four fully connected hidden layers of 200, 200, 100, and 50 units, respectively.

MNIST contains 60,000 training images. To compare how much our approach leads to improvement on smaller-scale data where Bayesian uncertainty matters more, we evaluate the test likelihood after training the model on randomly sampled fractions of the training set. We use the same training schedules as in the publicly available implementation, keeping the total number of training iterations independent of the size of the training set. Different to the original implementation, we shuffle the training set before each training epoch as this turns out to increase the performance for both our method and the baseline.

Figure 6 shows the predictive log-likelihood of the whole test set, where the VAE is trained on random subsets of different sizes of the training set. We use the same subset to train with PVI and KLVI for each training set size. PVI leads to a higher predictive likelihood than traditional KLVI on subsets of the data. We explain this finding with our observation that the variational distributions obtained from PVI capture more of the posterior variance. As the size of the training set grows—and the posterior uncertainty decreases—the performance of KLVI catches up with PVI.

As a potential explanation why PVI converges to the KLVI result for large training sets, we note that $\mathbb{E}_{q^*}[(V_0^* - V)^3] = 0$ at the optimal variational distribution q^* and reference energy V_0^* (see Section 3.3). If V becomes a symmetric random variable (such as a Gaussian) in the limit of a large training set, then this condition implies that $\mathbb{E}_{q^*}[V] = V_0^*$, and PVI reduces to KLVI for large training sets.

5 Conclusion

We first presented a view on black box variational inference as a form of biased importance sampling, where we can trade-off bias versus variance by the choice of divergence. Bias refers to the deviation of the bound from the true marginal likelihood, and variance refers to its reparameterization gradient estimator. We then propose a new variational bound that connects to variational perturbation theory, and which includes corrections to the standard Kullback-Leibler bound. We showed both theoretically and experimentally that our proposed PVI bound is tighter than the KL bound, and has lower-variance reparameterization gradients compared to α-VI. In order to scale up our method to massive data sets, future work will explore stochastic versions of PVI. Since the PVI bound contains interaction terms between all data points, breaking it up into mini-batches is non-straightforward. Furthermore, the PVI and α bounds can also be combined, such that PVI further approximates α-VI. This could lead to promising results on large data sets where traditional α-VI is hard to optimize due to its variance, and traditional PVI converges to KLVI.

References

Amari, S. (2012). *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media.

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *ICML*.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. Springer.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *ICLR*.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. M. (2017). Variational inference via χ upper bound minimization. In *ICML*.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. (2016). Black-box alpha divergence minimization. In *ICML*.

Hoffman, M. and Blei, D. (2014). Structured stochastic variational inference. *CoRR abs/1404.4114*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *JMLR*, 14(1).

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2).

Kappen, H. J. and Wiegerinck, W. (2001). Second order approximations for probability models. In *NIPS*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.

Kleinert, H. (2009). *Path integrals in quantum mechanics, statistics, polymer physics, and financial markets*. World scientific.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *NIPS*.

Minka, T. (2005). Divergence measures and message passing. Technical report, Technical report, Microsoft Research.

Opper, M. (2015). Expectation propagation. In Krzakala, F., Ricci-Tersenghi, F., Zdeborova, L., Zecchina, R., Tramel, E. W., and Cugliandolo, L. F., editors, *Statistical Physics, Optimization, Inference, and Message-Passing Algorithms*, chapter 9, pages 263–292. Oxford University Press.

Opper, M., Paquet, U., and Winther, O. (2013). Perturbative corrections for approximate inference in gaussian latent variable models. *JMLR*, 14(1).

Paquet, U., Winther, O., and Opper, M. (2009). Perturbation corrections in approximate inference: Mixture modelling applications. *JMLR*, 10(Jun).

Plefka, T. (1982). Convergence condition of the TAP equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6):1971.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *AISTATS*.

Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *ICML*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Ruiz, F., Titsias, M., and Blei, D. (2016). The generalized reparameterization gradient. In *NIPS*.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic line ar regression. *Bayesian Analysis*, 8(4).

Tanaka, T. (1999). A theory of mean field approximation. In *NIPS*.

Tanaka, T. (2000). Information geometry of mean-field approximation. *Neural Computation*, 12(8).

Thouless, D., Anderson, P. W., and Palmer, R. G. (1977). Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3).

Supplementary Material to "Perturbative Black Box Variational Inference"

Robert Bamler[*]
Disney Research
Pittsburgh, USA

Cheng Zhang[*]
Disney Research
Pittsburgh, USA

Manfred Opper
TU Berlin
Berlin, Germany

Stephan Mandt[*]
Disney Research
Pittsburgh, USA

firstname.lastname@{disneyresearch.com, tu-berlin.de}

Proof that PVI minimizes a divergence

In this supplement we show that perturbative black box variational inference (PVI) minimizes a valid divergence from the variational distribution $q(\mathbf{z})$ to the true posterior distribution $p(\mathbf{z}|\mathbf{x})$. This has the important consequence that PVI converges to exact inference in the limit of an arbitrarily flexible variational family. In contrast to traditional Kullback-Leibler variational inference (KLVI) and α-VI, the divergence minimized by PVI depends on the choice of variational family.

Let f be any regularizing function that satisfies the conditions for Eq. 3 of the main text (i.e., f is concave and smaller than the identity). Let $\mathcal{L}_f(q)$ be the associated lower bound on the model evidence. We prefer this notation over the notation $\mathcal{L}_f(\lambda)$ used in the main text here, because, at this stage, we do not restrict q to a specific family of variational distributions indexed by λ. We define a divergence D_f from q to the true posterior $p(\mathbf{z}|\mathbf{x})$ by

$$D_f(p||q) \equiv f(p(\mathbf{x})) - \mathcal{L}_f(q). \tag{S1}$$

Here, the model evidence $p(\mathbf{x})$ is unknown to us, but it is a well-defined constant assuming that the model parameters are kept constant. We show that D_f is indeed a valid divergence. From Eq. 3 of the main text, we find $\mathcal{L}_f(q) \leq f(p(\mathbf{x}))$ and therefore D_f is non-negative. The lower bound $\mathcal{L}_f(q)$ is defined in Eqs. 1 and 3 of our paper as

$$\mathcal{L}_f(q) \equiv \mathbb{E}_{q(\mathbf{z})}\left[f\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right) \right]. \tag{S2}$$

Setting $q(\mathbf{z})$ to the true posterior, $p(\mathbf{x}, \mathbf{z})/p(\mathbf{x})$, yields $\mathcal{L}_f(q) = f(p(\mathbf{x}))$, and therefore sets $D_f(p||q)$ to zero. Thus, D_f is indeed a valid divergence.

Consider now the specific family of regularizing functions f_{V_0} defined in Eq. 6 of the main text. We now also restrict q to be a member of some variational family. Maximizing the corresponding lower bound $\mathcal{L}_{f_{V_0}}(q)$ simultaneously over the reference energy V_0 and over the variational parameters yields an optimal reference energy V_0^* and an optimal member q^* of the variational family. Both depend not only on the model but also on the variational family to which q is restricted. Evidently, q^* is the member of the variational family that minimizes the divergence

$$D_{f_{V_0^*}}(p||q) = f_{V_0^*}(p(\mathbf{x})) - \mathcal{L}_{f_{V_0^*}}(q). \tag{S3}$$

Here, the first term on the right-hand side is a constant (since V_0^* is). Its value is not known to us, but well defined. Thus, PVI minimizes a valid divergence to the true posterior. As a practical consequence this implies that the exact maximum of the PVI lower bound is the true posterior if the variational family is sufficiently flexible to contain it. Note that the choice of divergence that PVI minimizes depends on both the model and the variational family (via V_0^*).

---

[*]Equal contribution