Proximity Variational Inference

Jaan Altosaar
altosaar@princeton.edu
Princeton University

Rajesh Ranganath
rajeshr@cims.nyu.edu
New York University

David M. Blei
david.blei@columbia.edu
Columbia University

Abstract

Variational inference is a powerful approach for approximate posterior inference. However, it is sensitive to initialization and can be subject to poor local optima. In this paper, we develop proximity variational inference (PVI). PVI is a new method for optimizing the variational objective that constrains subsequent iterates of the variational parameters to robustify the optimization path. Consequently, PVI is less sensitive to initialization and optimization quirks and finds better local optima. We demonstrate our method on four proximity statistics. We study PVI on a Bernoulli factor model and sigmoid belief network fit to real and synthetic data and compare to deterministic annealing (Katahira et al., 2008). We highlight the flexibility of PVI by designing a proximity statistic for Bayesian deep learning models such as the variational autoencoder (Kingma and Welling, 2014; Rezende et al., 2014) and show that it gives better performance by reducing overpruning. PVI also yields improved predictions in a deep generative model of text. Empirically, we show that PVI consistently finds better local optima and gives better predictive performance.

1 Introduction

Variational inference (VI) is a powerful method for probabilistic modeling. VI uses optimization to approximate difficult-to-compute conditional distributions (Jordan et al., 1999). In its modern incarnation, it has scaled Bayesian computation to large data sets (Hoffman et al., 2013), generalized to large classes of models (Kingma and Welling, 2014; Ranganath et al., 2014; Rezende and Mohamed, 2015), and has been deployed as a computational engine in probabilistic programming systems (Mansinghka et al., 2014; Kucukelbir et al., 2015; Tran et al., 2016).

Despite these significant advances, however, VI has drawbacks. For one, it tries to iteratively solve a difficult nonconvex optimization problem and its objective contains many local optima. Consequently, VI is sensitive to initialization and easily gets stuck in a poor solution. We develop a new optimization method for VI and show that it finds better optima.

Consider a probability model $p(\mathbf{z}, \mathbf{x})$ and the goal of calculating the posterior $p(\mathbf{z} \mid \mathbf{x})$. The idea behind VI is to posit a family of distributions over the hidden variables $q(\mathbf{z}; \boldsymbol{\lambda})$ and then fit the variational parameters $\boldsymbol{\lambda}$ to minimize the Kullback-Leibler (KL) divergence between the approximating family and the exact posterior, $\text{KL}(q(\mathbf{z}; \boldsymbol{\lambda}) || p(\mathbf{z} \mid \mathbf{x}))$. The KL is not tractable so VI optimizes a proxy. That proxy is the evidence lower bound (ELBO),

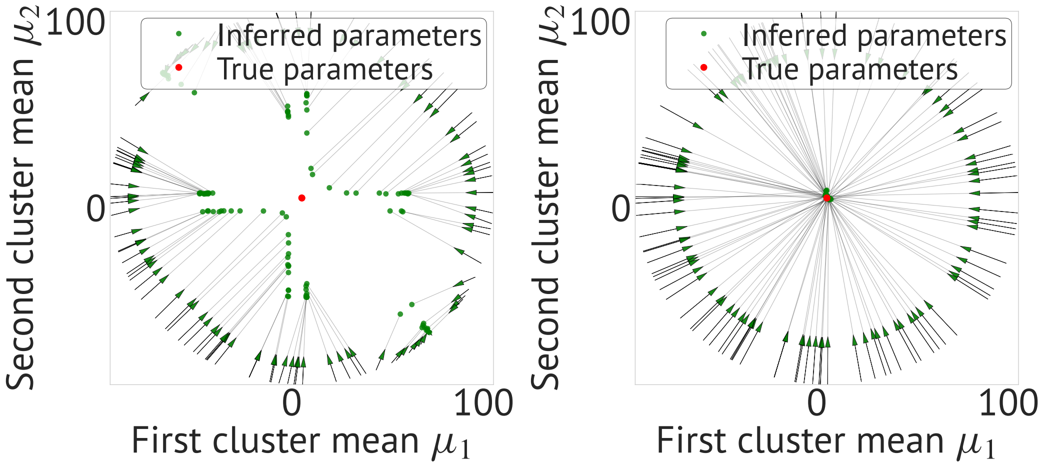
$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z}; \boldsymbol{\lambda})], \qquad (1)$$

where expectations are taken with respect to $q(\mathbf{z}; \boldsymbol{\lambda})$. Maximizing the ELBO with respect to $\boldsymbol{\lambda}$ is equivalent to minimizing the KL divergence.

The issues around VI stem from the ELBO and the iterative algorithms used to optimize it. When the algorithm zeroes (or nearly zeroes) some of the support of $q(\mathbf{z}; \boldsymbol{\lambda})$, it becomes hard to later "escape," i.e., to add support for the configurations of the latent variables that have been assigned zero probability (MacKay, 2003; Burda et al., 2015). This leads to poor local optima and to sensitivity to the starting point, where a misguided initialization will lead to such optima. These problems happen in both gradient-based and coordinate ascent methods. We address these issues with proximity variational inference (PVI), a variational inference algorithm that is specifically designed to avoid poor local optima and to be robust to different initializations.

PVI builds on the proximity perspective of gradient ascent. The proximity perspective views each step of gradient ascent as a constrained minimization of a Taylor expansion of the objective around the previous step's parameter (Spall, 2003; Boyd and Vandenberghe, 2004). The constraint, a *proximity constraint*, enforces that the next point should be inside a Euclidean ball of the previous. The step size relates to the size of that ball.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).



(a) Variational inference

(b) Proximity variational inference, Algorithm 2

Figure 1: Proximity variational inference (PVI) is robust to bad initialization. We study a Bernoulli factor model. Model parameters are randomly initialized on a ring around the known true parameters (in red) used to generate the data. The arrows start at these parameter initializations and end at the final parameter estimates (shown as green dots). **(a)** Variational inference with gradient ascent suffers from multiple local optima and cannot reliably recover the truth. **(b)** PVI with an entropy proximity statistic reliably infers the true parameters using Algorithm 2.

In VI, a constraint on the Euclidean distance means that all dimensions of the variational parameters are equally constrained. We posit that this leads to problems; some dimensions need more regularization than others. For example, consider a variational distribution that is Gaussian. A good optimization will change the variance parameter more slowly than the mean parameter to prevent rapid changes to the support. The Euclidean constraint cannot enforce this. Furthermore, the constraints enforced by gradient descent are transient; the constraints are relative to the previous iterate—one poor move during the optimization can lead to permanent optimization problems.

To this end, PVI uses proximity constraints that are more meaningful to variational inference and to optimization of probability parameters. A constraint is defined using a proximity statistic and distance function. As one example, we consider a constraint based on the entropy proximity statistic. This limits the change in entropy of the variational approximation from one step to the next. Consider again a Gaussian approximation. The entropy is a function of the variance alone and thus the entropy constraint counters the pathologies induced by the Euclidean proximity constraint. We also study constraints built from other proximity statistics, such as those that penalize the rapid changes in the mean and variance of the approximate posterior.

Figure 1 provides an illustration of the advantages of PVI. Our goal is to estimate the parameters of a factor analysis model with variational inference, i.e., using the posterior expectation under a fitted variational distribution. We run variational inference 100 times, each time initializing the estimates (the model parameters) to a different position on a ring around the truth.

In the figure, red points indicate the true value. The start locations of the green arrows indicate the initialized estimates. Green points indicate the final estimates, after optimizing from the initial points. Panel (a) shows that optimizing the standard ELBO with gradients leads to poor local optima and misplaced estimates. Panel (b) illustrates that regardless of the initialization, PVI with an entropy proximity statistic finds estimates that are close to the true value.

The rest of the paper is organized as follows. Section 2 reviews variational inference and the proximity perspective of gradient optimization. Section 3 derives PVI; we develop four proximity constraints and two algorithms for optimizing the ELBO. We study four models in Section 4: a Bernoulli factor model, a sigmoid belief network (Mnih and Rezende, 2016), a variational autoencoder (Kingma and Welling, 2014; Rezende et al., 2014), and a deep exponential family model of text (Ranganath et al., 2015). PVI outperforms classical methods for variational inference.

Related work. Recent work has proposed several related algorithms. Khan et al. (2015) and Theis and Hoffman (2015) develop a method to optimize the ELBO that imposes a soft limit on the change in KL of consecutive variational approximations. This is equivalent to PVI with identity proximity statistics and a KL distance function. Khan et al. (2016) extend both prior works to other divergence functions. Their general approach is equivalent to PVI identity proximity statistics and distance functions given by strongly-convex divergences. Compared to prior work, PVI generalizes to a broader class of proximity statistics. We develop proximity statistics based on entropy, KL, orthogonal weight matrices, and the mean and variance of the variational approximation.

The problem of model pruning in variational inference has also been studied and analytically solved in a matrix factorization model in Nakajima et al. (2013)—this method is model-specific, whereas PVI applies to a much broader class of latent variable models. Finally, deterministic annealing (Katahira et al., 2008) consists of adding a temperature parameter to the entropy term in the ELBO that is annealed to one during inference. This is similar to PVI with the entropy proximity statistic which keeps the entropy stable across iterations. Deterministic annealing enforces global penalization of low-entropy configurations of latent variables rather than the smooth constraint used in PVI, and cannot accommodate the range of proximity statistics we design in this work.

2 Variational inference

Consider a model $p(\mathbf{x}, \mathbf{z})$, where $\mathbf{x}$ is the observed data and $\mathbf{z}$ are the latent variables. As described in Section 1, VI posits an approximating family $q(\mathbf{z}; \boldsymbol{\lambda})$ and maximizes the ELBO in Equation (1). Solving this optimization is equivalent to finding the variational approximation that minimizes KL divergence to the exact posterior (Jordan et al., 1999; Wainwright and Jordan, 2008).

2.1 Gradient ascent has Euclidean proximity

Gradient ascent maximizes the ELBO by repeatedly following its gradient. One view of this algorithm is that it repeatedly maximizes the linearized ELBO subject to a proximity constraint on the current variational parameter (Spall, 2003). The name 'proximity' comes from constraining subsequent parameters to remain close in the proximity statistic. In gradient ascent, the proximity statistic for the variational parameters is the identity function $f(\boldsymbol{\lambda}) = \boldsymbol{\lambda}$, and the distance function is the square difference.

Let $\boldsymbol{\lambda}_t$ be the variational parameters at iteration t and ρ be a constant. To obtain the next iterate $\boldsymbol{\lambda}_{t+1}$, gradient ascent maximizes the linearized ELBO,

$$\begin{aligned} U(\boldsymbol{\lambda}_{t+1}) =& \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla \mathcal{L}(\boldsymbol{\lambda}_t)^\top (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) \\ & - \frac{1}{2\rho} (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t). \end{aligned} \quad (2)$$

Specifically, this is the linearized ELBO around $\boldsymbol{\lambda}_t$, subject to $\boldsymbol{\lambda}_{t+1}$ being close to $\boldsymbol{\lambda}_t$ in squared Euclidean distance.

Finding the $\boldsymbol{\lambda}_{t+1}$ which maximizes Equation (2) yields

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho \nabla \mathcal{L}(\boldsymbol{\lambda}_t). \quad (3)$$

This is the familiar gradient ascent update with a step size of ρ. The step size ρ controls the radius of the Euclidean ball which demarcates valid next steps for the parameters. Note that the Euclidean constraint between subsequent iterates is implicit in all gradient ascent algorithms.

2.2 An example where variational inference fails

We study a setting where variational inference suffers from poor local optima. Consider a factor model, with Bernoulli latent variables and Gaussian likelihood:

$$\begin{aligned} z_{ik} &\sim \text{Bernoulli}(\pi) & (4) \\ x_i &\sim \text{Gaussian}\left(\mu = \textstyle\sum_k z_{ik}\mu_k, \sigma^2 = 1\right). & (5) \end{aligned}$$

This is a "feature" model of real-valued data x; when one of the features is on (i.e., $z_{ik} = 1$), the ith mean shifts according the that feature's mean parameter (i.e., μ_k). Thus the binary latent variables z_{ik} control which cluster means μ_k contribute to the distribution of x_i.

The Bernoulli prior is parametrized by π; we choose a Bernoulli approximate posterior $q(z_k; \lambda_k) = \text{Bernoulli}(\lambda_k)$. A common approach to VI is coordinate ascent (Bishop, 2006), where we iteratively optimize each variational parameter. The optimal variational parameter for z_{ik} is

$$\lambda_{ik} \propto \exp\left\{ \mathbb{E}_{-z_{ik}} \left[-\frac{1}{2\sigma^2}(x_i - \sum_j z_{ij}\mu_j)^2 \right] \right\}. \quad (6)$$

We can use this update in a variational expectation-maximization setting. The corresponding gradient for μ_k is

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = -\frac{1}{\sigma^2} \sum_i \left(-x_i\lambda_{ik} + \lambda_{ik}\mu_k + \lambda_{ik} \sum_{j \neq k} \lambda_{ij}\mu_j \right). \quad (7)$$

Meditating on these two equations reveals a deficiency in mean-field variational inference. First, if the mean parameters μ are initialized far from the data then $q^*(z_{ik} = 1)$ will be very small. The reason is in Equation (6), where the squared difference between the data x_i and the expected cluster mean will be large and negative. Second, when the probability of cluster assignment is close to zero, λ_{ik} is small. This means that the norm of the gradient in Equation (7) will be small. Consequently, learning will be slow. We see this phenomenon in Figure 1 (a). Variational inference arrives at poor local optima and does not recover the correct cluster means.

Algorithm 1: Proximity variational inference

```
Input: Initial parameters $\boldsymbol{\lambda}_0$, proximity statistic $f(\boldsymbol{\lambda})$,
       distance function $d$
Output: Parameters $\boldsymbol{\lambda}$ of variational $q(\boldsymbol{\lambda})$ that maximize
        the ELBO objective
while $\mathcal{L}$ not converged do
    $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \text{Noise}$
    while $U$ not converged do
        Update $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_{t+1} + \rho \nabla_{\boldsymbol{\lambda}} U(\boldsymbol{\lambda}_{t+1})$
    end
    $\boldsymbol{\lambda}_t \leftarrow \boldsymbol{\lambda}_{t+1}$
end
return $\boldsymbol{\lambda}$
```

3 Proximity variational inference

We now develop proximity variational inference (PVI), a variational inference method that is robust to initialization and can consistently reach good local optima (Section 3.1). PVI alters the notion of proximity. We further restrict the iterates of the variational parameters by deforming the Euclidean ball implicit in classical gradient ascent. This is done by choosing proximity statistics that are not the identity function, and distance functions that are different than the square difference. These design choices help guide the variational parameters away from poor local optima (Section 3.2). One drawback of the proximity perspective is that it requires an inner optimization at each step of the outer optimization. We use a Taylor expansion to avoid this computational burden (Section 3.3).

3.1 Proximity constraints for variational inference

PVI enriches the proximity constraint in gradient ascent of the ELBO. We want to develop constraints on the iterates $\boldsymbol{\lambda}_t$ to counter the pathologies of standard variational inference.

Let $f(\cdot)$ be a *proximity statistic*, and let d be a differentiable distance function that measures distance between proximity statistic iterates. A *proximity constraint* is the combination of a distance function d applied to a proximity statistic f. (Recall that in classical gradient ascent, the Euclidean proximity constraint uses the identity as the proximity statistic and the square difference as the distance.) Let k be the scalar magnitude of the proximity constraint. We define the proximity update equation for the variational parameters $\boldsymbol{\lambda}_{t+1}$ to be

$$\begin{aligned} U(\boldsymbol{\lambda}_{t+1}) =& \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla \mathcal{L}(\boldsymbol{\lambda}_t)^\top (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) \\ & - \frac{1}{2\rho}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) \qquad (8) \\ & - k \cdot d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_{t+1})), \end{aligned}$$

Algorithm 2: Fast proximity variational inference

```
Input: Initial parameters $\boldsymbol{\lambda}_0$, adaptive learning rate
       optimizer, proximity statistic $f(\boldsymbol{\lambda})$, distance $d$
Output: Parameters $\boldsymbol{\lambda}$ of the variational distribution $q(\boldsymbol{\lambda})$
        that maximize the ELBO objective
while $\mathcal{L}_{proximity}$ not converged do
    $\boldsymbol{\lambda}_{t+1} =$
      $\boldsymbol{\lambda}_t + \rho(\nabla \mathcal{L}(\boldsymbol{\lambda}_t) - k \cdot (\nabla d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_t)) \nabla f(\boldsymbol{\lambda}_t))$.
    $\tilde{\boldsymbol{\lambda}} = \alpha \tilde{\boldsymbol{\lambda}} + (1 - \alpha) \boldsymbol{\lambda}_{t+1}$
end
return $\boldsymbol{\lambda}$
```

where $\tilde{\boldsymbol{\lambda}}$ is the variational parameter to which we are measuring closeness. In gradient ascent, this is the previous parameter $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_t$, but our construction can enforce proximity to more than just the previous parameters. For example, we can set $\tilde{\boldsymbol{\lambda}}$ to be an exponential moving average[1]—this adds robustness to one-update optimization missteps.

The next parameters are found by maximizing Equation (8). This enforces that the variational parameters between updates will remain close in the proximity statistic $f(\boldsymbol{\lambda})$. For example, $f(\boldsymbol{\lambda})$ might be the entropy of the variational approximation; this can avoid zeroing out some of its support. This procedure is detailed in Algorithm 1. The magnitude k of the constraint is a hyperparameter. The inner optimization loop optimizes the update equation U at each step.

3.2 Proximity statistics for variational inference

We describe four proximity statistics $f(\boldsymbol{\lambda})$ appropriate for variational inference. Together with a distance function, these proximity statistics yield proximity constraints. (We study them in Section 4.)

Entropy proximity statistic. Consider a constraint built from the entropy proximity statistic, $f(\boldsymbol{\lambda}) = \mathrm{H}(q(\mathbf{z}; \boldsymbol{\lambda}))$. Informally, the entropy measures the amount of randomness present in a distribution. High entropy distributions look more uniform across their support; low entropy distributions are peaky.

Using the entropy in Equation (8) constrains all updates to have entropy close to their previous update. When the variational distributions are initialized with large entropy, this statistic balances the "zero-forcing" issue that is intrinsic to variational inference (MacKay, 2003). Figure 1 demonstrates how PVI with an entropy constraint can correct this pathology.

KL proximity statistic. We can rewrite the ELBO to include the KL between the approximate posterior and the

[1] The exponential moving average of a variable $\boldsymbol{\lambda}$ is denoted $\tilde{\boldsymbol{\lambda}}$ and is updated according to $\tilde{\boldsymbol{\lambda}} \leftarrow \alpha \tilde{\boldsymbol{\lambda}} + (1 - \alpha)\boldsymbol{\lambda}$, where α is a decay close to one.

prior (Kingma and Welling, 2014),

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \text{KL}(q(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\lambda}) || p(\mathbf{z})).$$

Flexible models tend to minimize the KL divergence too quickly and get stuck in poor optima (Bowman et al., 2016; Higgins et al., 2016). The choice of KL as a proximity statistic prevents the KL from being optimized too quickly relative to the likelihood.

Mean/variance proximity statistic. A common theme in the problems with variational inference is that the bulk of the probability mass can quickly move to a point where that dimension will no longer be explored (Burda et al., 2015). One way to address this is to restrict the mean and variance of the variational approximation to change slowly during optimization. This constraint only allows higher order moments of the variational approximation to change rapidly. The mean $\mu = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}[\mathbf{z}]$ and variance $\text{Var}(\mathbf{z}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}[(\mathbf{z} - \mu)^2]$ are the statistics $f(\boldsymbol{\lambda})$ we constrain.

Orthogonal proximity statistic. In Bayesian deep learning models such as the variational autoencoder (Kingma and Welling, 2014; Rezende et al., 2014) it is common to parametrize the variational distribution with a neural network. Orthogonal weight matrices make optimization easier in neural networks by allowing gradients to propagate further (Saxe et al., 2013). We can exploit this fact to design an orthogonal proximity statistic for the weight matrices W of neural networks: $f(W) = WW^\top$. With an orthogonal initialization for the weights, this statistic enables efficient optimization.

We gave four examples of proximity statistics that, together with a distance function, yield proximity constraints. We emphasize that any function of the variational parameters $f(\boldsymbol{\lambda})$ can be designed to ameliorate issues with variational inference. We discuss how to select a proximity statistic in Section 5.

3.3 Linearizing the proximity constraint for fast proximity variational inference

PVI in Algorithm 1 requires optimizing the update equation, Equation (8), at each iteration. This rarely has a closed-form solution and requires a separate optimization procedure that is computationally expensive.

An alternative is to use a first-order Taylor expansion of the proximity constraint. Let ∇d be the gradient with respect to the second argument of the distance function, and $f(\tilde{\boldsymbol{\lambda}})$ be the first argument to the distance. We compute the expansion around $\boldsymbol{\lambda}_t$ (the variational parameters at step t),

Inference method	ELBO	Likelihood
Variational inference	−121.4	−113.7
Deterministic annealing	−116.8	−108.8
PVI, Entropy constraint	**−113.3**	**−106.7**
PVI, Mean/variance constraint	−114.9	−107.4

Table 1: Proximity variational inference improves on deterministic annealing (Katahira et al., 2008) and VI in a one-layer sigmoid belief network. We report the test set evidence lower bound (ELBO) and marginal likelihood on the binary MNIST dataset (Larochelle and Murray, 2011). The model has one stochastic layer of 200 latent variables. PVI outperforms deterministic annealing (Katahira et al., 2008) and the classical variational inference algorithm.

$$\begin{aligned} U(\boldsymbol{\lambda}_{t+1}) =& \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla\mathcal{L}(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) \\ & - \frac{1}{2\rho}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) \\ & - k \cdot (d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_t)) \\ & + \nabla d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_t)) \nabla f(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)). \end{aligned}$$

This linearization enjoys a closed-form solution for the variational parameters $\boldsymbol{\lambda}_{t+1}$,

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho(\nabla\mathcal{L}(\boldsymbol{\lambda}_t) - k \cdot (\nabla d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_t)) \nabla f(\boldsymbol{\lambda}_t)). \tag{9}$$

Note that setting $\tilde{\boldsymbol{\lambda}}$ to the current parameter $\boldsymbol{\lambda}_t$ removes the proximity constraint. Distance functions are minimized at zero so their derivative is zero at that point.

Fast PVI is detailed in Algorithm 2. Unlike PVI in Algorithm 1, the update in Equation (9) does not require an inner optimization. Fast PVI is tested in Section 4. The complexity of fast PVI is similar to standard VI because fast PVI optimizes the ELBO subject to the distance constraint in f. (The added complexity comes from computing the derivative of f; no inner optimization loop is required.)

Finally, note that fast PVI implies a global objective which varies over time. It is

$$\begin{aligned} \mathcal{L}_{\text{proximity}}(\boldsymbol{\lambda}_{t+1}) =& \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\boldsymbol{\lambda}_{t+1})] \\ & - k \cdot d(f(\tilde{\boldsymbol{\lambda}}), f(\boldsymbol{\lambda}_{t+1})). \end{aligned}$$

Because d is a distance, this remains a lower bound on the evidence, but where new variational approximations remain close in f to previous iterations' distributions.

4 Experiments

We developed proximity variational inference (PVI). We now empirically study PVI, variational inference, and deterministic annealing (Katahira et al., 2008).

Inference method	ELBO	Likelihood
Variational inference	−116.2	−104.9
Deterministic annealing	−102.0	−94.2
PVI, Entropy constraint	**−99.7**	**−93.2**
PVI, Mean/variance constraint	−100.7	−93.3

Table 2: Proximity variational inference improves over deterministic annealing and VI in a three-layer sigmoid belief network. The model has three layers of 200 latent variables. We report the evidence lower bound (ELBO) and marginal likelihood on the MNIST test set (Larochelle and Murray, 2011).

We first study sigmoid belief networks and find that PVI improves over deterministic annealing and VI in terms of held-out values of the ELBO and marginal likelihood. We then study a variational autoencoder model of images. Using an orthogonal proximity statistic, we show that PVI improves over classical VI by reducing overpruning. Finally, we study a deep generative model fit to a large corpus of text, where PVI yields better predictive performance with little hyperparameter tuning.[2]

Hyperparameters. For PVI, we use the inverse Huber distance for d.[3] The inverse Huber distance penalizes smaller values than the square difference. For PVI Algorithm 2, we set the exponential moving average decay constant for $\tilde{\boldsymbol{\lambda}}$ to $\alpha = 0.9999$. We set the constraint scale k (or temperature parameter in deterministic annealing) to the initial absolute value of the ELBO unless otherwise specified. We explore two annealing schedules for PVI and deterministic annealing: a linear decay and an exponential decay. For the exponential decay, the value of the magnitude at iteration t of T total iterations is set to $k \cdot \gamma^{\frac{t}{T}}$ where γ is the decay rate. We use the Adam optimizer (Kingma and Ba, 2015) unless otherwise specified.

4.1 Sigmoid belief network

The sigmoid belief network is a discrete latent variable model with layers of Bernoulli latent variables (Neal, 1992; Ranganath et al., 2015). It is used to benchmark variational inference algorithms (Mnih and Rezende, 2016). The approximate posterior is a collection of Bernoullis, parameterized by an inference network with weights and biases. We fit these variational parameters with VI, deterministic annealing (Katahira et al., 2008), or PVI, and learn the model parameters (weights and biases) using variational expectation-maximization.

We learn the weights and biases of the model with gradient ascent. We use a step size of $\rho = 10^{-3}$ and train for 4×10^6 iterations with a batch size of 20. For PVI Algorithm 2 and deterministic annealing, we grid search over exponential decays with rates $\gamma \in \{10^{-5}, 10^{-6}, ..., 10^{-10}, 10^{-20}, 10^{-30}\}$ and report the best results for each algorithm. (We also explored linear decays but they did not perform as well.) To reduce the variance of the gradients, we use the leave-one-out control variate of Mnih and Rezende (2016) with 5 samples. (This is an extension to the black box variational inference algorithm in Ranganath et al. (2014).)

Results on MNIST. We train a sigmoid belief network model on the binary MNIST dataset of handwritten digits (Larochelle and Murray, 2011). For evaluation, we compute the ELBO and held-out marginal likelihood with importance sampling on the validation set of 10^4 digits (using 5000 samples, as in Rezende et al. (2014)). In Table 1 we show the results for a model with one layer of 200 latent variables. Table 2 displays similar results for a three-layer model with 200 latent variables per layer. In both one and three-layer models the KL proximity statistic performs worse than the mean/variance and entropy statistics; it requires different decay schedules. Overall, PVI with the entropy and mean/variance proximity statistics yields improvements in the held-out marginal likelihood in comparison to deterministic annealing and VI.

4.2 Variational autoencoder

To demonstrate the value of designing proximity statistics tailored to specific models, we study the variational autoencoder (Kingma and Welling, 2014; Rezende et al., 2014). This model is difficult to optimize, and current optimization techniques yield solutions that do not use the full model capacity (Burda et al., 2015). In Section 3.2 we designed an orthogonal proximity statistic to make backpropagation in neural networks easier. We show that this statistic enables us to find a better approximate posterior in the variational autoencoder by reducing overpruning.

We fit the variational autoencoder to binary MNIST data (Larochelle and Murray, 2011) with variational expectation-maximization. The model has one layer of 100 Gaussian latent variables. The inference network and generative network are chosen to have two hidden layers of size 200 with rectified linear units (ReLUs). We use an orthogonal initialization for the inference network weights. The learning rate is set to 10^{-3} and we run VI and PVI for 5×10^4 iterations. The orthogonal proximity statistic changes rapidly during optimization, so we use constraint

[2] We also compared PVI to Khan et al. (2015). Specifically, we tested PVI on the Bayesian logistic regression model from that paper and with the same data. Because Bayesian logistic regression has a single mode, all methods performed equally well. We note that we could not apply their algorithm to the sigmoid belief network because it would require approximating difficult iterated expectations.

[3] We define the inverse Huber distance $d(x, y)$ to be $|x - y|$ if $|x - y| < 1$ and $0.5(x - y)^2 + 0.5$ otherwise. The constants ensure the function and its derivative are continuous at $|x - y| = 1$.

Inference method	ELBO	Likelihood
Variational inference	−101.0	−94.2
PVI, Orthogonal constraint	**−100.4**	**−93.9**

Table 3: Proximity variational inference with an orthogonal proximity statistic makes optimization easier in a variational autoencoder model (Kingma and Welling, 2014; Rezende et al., 2014). We report the held-out evidence lower bound (ELBO) and estimates of the marginal likelihood on the binarized MNIST (Larochelle and Murray, 2011) test set.

magnitudes $k \in \{1, 10^{-1}, 10^{-2}, ..., 10^{-5}\}$, with no decay, and report the best result.

We compute the ELBO and importance-sampled marginal likelihood estimates on the validation set. In Table 3 we show that PVI with the orthogonal proximity statistic on the weights of the inference network enables easier optimization and improves over VI.

Why does PVI improve upon VI in the variational autoencoder? The choice of rectified linear units in the inference network allows us to study overpruning of the latent code (MacKay, 2001; Burda et al., 2015). We study the fraction of 'dead ReLUs'— the fraction of rectified linear units in each layer of the inference neural network whose input is below zero. With PVI Algorithm 2 and the orthogonal proximity constraint, the inference network has 1.6% fewer dead ReLUs in the hidden layer and shows a 3.2% reduction in the output layer than in the same model learned using classic variational inference.

Once the input to a ReLU drops below zero, the unit stops receiving gradient updates. The output layer parametrizes the latent variable distribution, so this means PVI reduced the pruning of the approximate posterior and led to the utilization of 3 additional latent variables. This is the reason it outperformed a variational autoencoder fit with VI.

4.3 Deep generative model of text

Deep exponential family models, Bayesian analogues to neural networks, represent a flexible class of models (Ranganath et al., 2015). However, black box variational inference is commonly used to fit these models, which requires variance reduction (Ranganath et al., 2014). Deep exponential family models with Poisson latent variables present a challenging approximate inference problem because they are discrete and high-variance. We demonstrate that PVI with the mean/variance proximity constraint improves predictive performance in such an unsupervised model of text.

The generative process for a 1-layer deep exponential family model of text, with Poisson latent variables and Poisson likelihood, is

$$\mathbf{z} \sim \text{Poisson}(\boldsymbol{\lambda}) \tag{10}$$

$$\mathbf{x} \sim \text{Poisson}(\mathbf{z}^\top g(W)), \tag{11}$$

where W are real-valued model parameters and g is an elementwise function that maps to the positive reals (we use the softplus function). The dimension of $\mathbf{z}$ is K, so the model parameters must have shape (K, V) where V is the cardinality of the count-valued observations $\mathbf{x}$. We use this as a model of documents, so $\mathbf{x}$ is the bag-of-words representation of word counts, W represents the common factors in documents, and the per-document latent variable $\mathbf{z}$ captures which factors are prevalent in the language used in each document.

Inference method	Perplexity
Variational inference	2329
PVI, Mean/variance constraint	**2294**

Table 4: Proximity variational inference with a mean/-variance proximity statistic improves predictive performance in a deep exponential family model with Poisson latent variables. We report the held-out perplexity on the *Science* corpus of journal articles.

We study the performance of our method on a corpus of articles from the academic journal *Science*. The corpus contains 138K documents in the training set, 1K documents in the test set, and 5.9K terms. We set the latent dimension to 100, and fit the variational Poisson parameters using black box variational inference (Ranganath et al., 2014) using minibatches of size 64 and 32 samples of the latent variables to estimate the gradients.

Poisson variables have high variance, so we use the optimal control variate scaling developed in Ranganath et al. (2014) and estimate this scaling in a round-robin fashion as in Mnih and Rezende (2016) for efficiency. We use the RMSProp adaptive gradient optimizer (Tieleman and Hinton, 2012) with a step size of 0.01. For PVI Algorithm 2 with the mean/variance proximity statistic, we use an exponential decay for the constraint and test decay rates γ of 10^{-5} and 10^{-10}. We train for 10^6 iterations on the *Science* corpus, using variational expectation-maximization to learn the model parameters.

For evaluation, we keep the model parameters fixed and hold out 90% of the words in each document in the test set. Using the 10% of observed words in each document, we learn the variational parameters using PVI or variational inference with 300 iterations per document. We compute perplexity

university	fig	disease
new	dna	virus
department	protein	hiv
york	cells	aids
research	cell	human
science	gene	patients
state	binding	diseases
laboratory	two	cases
national	sequence	infection
california	proteins	infected

Table 5: The top ten words for three factors of a deep exponential family model with Poisson latent variables fit to the *Science* corpus of scientific articles. We show topics from a model fit with proximity variational inference; the topics for the same model fit with variational inference are similar.

on the held-out documents, which is given by

$$\exp\left(\frac{-\sum_{d\in\text{docs}}\sum_{w\in d}\log p(w \mid \#\text{ held-out in d})}{N_{\text{held-out words}}}\right). \tag{12}$$

Conditional on the number of held-out words in a document, the distribution over held-out words is multinomial. The mean of the conditional multinomial is the normalized Poisson rate of the document matrix-multiplied with the softplus of the weights. This is the same evaluation metric as in Ranganath et al. (2015).

The results of fitting the model to the corpus of *Science* documents are reported in Table 4 and Table 5. While the topics found by models fit with both PVI and VI are similar, PVI gives significantly better predictive performance in terms of held-out perplexity.

5 Discussion

We presented proximity variational inference, a flexible method designed to avoid bad local optima. We showed that classic variational inference gets trapped in these local optima and cannot recover. The choice of proximity statistic f and distance d enables the design of a variety of constraints that improve optimization. As examples of proximity statistics, we gave the entropy, KL divergence, orthogonal proximity statistic, and the mean and variance of the approximate posterior. We evaluated our method in four models to demonstrate that it is easy to implement, readily extensible, and leads to beneficial statistical properties of variational inference algorithms.

The empirical results also yield guidelines for choosing proximity statistics. The entropy is useful for models with discrete latent variables which are prone to quickly getting stuck in local optima or flat regions of the objective. We also saw that the KL statistic gives poor performance empirically, and that the orthogonal proximity statistic reduces pruning in deep generative models such as the variational autoencoder. In models like the deep exponential family model of text, the entropy is not tractable so the mean/variance proximity statistic is a natural choice.

Future work. Simplifying optimization is necessary for truly black-box variational inference. An adaptive magnitude decay based on the value of the constraint should further improve the technique (this could be done per-parameter). New proximity constraints are also easy to design and test. For example, the variance of the gradients of the variational parameters is a valid proximity statistic—which can be used to avoid variational approximations that have high-variance gradients. Another set of interesting proximity statistics are empirical statistics of the variational distribution, such as the mean, for when analytic forms are unavailable. We also leave the design and study of constraints that admit coordinate updates to future work.

Acknowledgments

The experiments presented in this article were performed on computational resources supported by the Princeton Institute for Computational Science and Engineering (PICSciE), the Office of Information Technology's High Performance Computing Center and Visualization Laboratory at Princeton University, Colin Raffel, and Dawen Liang. We are very grateful for their support. We also thank Emtiyaz Khan and Ben Poole for helpful discussions, and Josko Plazonic for computation support.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Józefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *International Conference on Learning Representations*.

Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., and Lerchner, A. (2016). Early Visual Concept Learning with Unsupervised Deep Learning. *ArXiv:1606.05579*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J.

(2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

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):183–233.

Katahira, K., Watanabe, K., and Okada, M. (2008). Deterministic annealing variant of variational bayes method. *Journal of Physics: Conference Series*, 95(1):012015.

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. (2016). Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Uncertainty in Artificial Intelligence*, pages 319–328.

Khan, M. E., Baqué, P., Fleuret, F., and Fua, P. (2015). Kullback-leibler proximal variational inference. In *Advances in Neural Information Processing Systems*, pages 3384–3392.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic Variational Inference in Stan. In *Neural Information Processing Systems*.

Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In *Artificial Intelligence and Statistics*, volume 15, pages 29–37.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

MacKay, D. J. (2001). Local minima, symmetry-breaking, and model pruning in variational free energy minimization. *Available online at http://www.inference.phy.cam.ac.uk/mackay/minima.pdf*.

Mansinghka, V., Selsam, D., and Perov, Y. N. (2014). Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv:1404.0099*.

Mnih, A. and Rezende, D. J. (2016). Variational inference for monte carlo objectives. In *International Conference on Machine Learning*, pages 2188–2196.

Nakajima, S., Sugiyama, M., Babacan, S. D., and Tomioka, R. (2013). Global analytic solution of fully-observed variational bayesian matrix factorization. *Journal of Machine Learning Research*.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113.

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

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Spall, J. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley.

Theis, L. and Hoffman, M. D. (2015). A trust-region method for stochastic variational inference with applications to streaming data. *Journal of Machine Learning Research*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5 - rmsprop. *COURSERA: Neural Networks for Machine Learning*.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv:1610.09787*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.