Provable Smoothness Guarantees for Black-Box Variational Inference

Justin Domke [1]

Abstract

Black-box variational inference tries to approximate a complex target distribution through a gradient-based optimization of the parameters of a simpler distribution. Provable convergence guarantees require structural properties of the objective. This paper shows that for location-scale family approximations, if the target is M-Lipschitz smooth, then so is the "energy" part of the variational objective. The key proof idea is to describe gradients in a certain inner-product space, thus permitting the use of Bessel's inequality. This result gives bounds on the location of the optimal parameters, and is a key ingredient for convergence guarantees.

1. Introduction

Variational inference (VI) approximates a complex distribution with a simpler one. Take a target distribution $p(\boldsymbol{z}, \boldsymbol{x})$ where $\boldsymbol{x}$ is observed data and $\boldsymbol{z}$ are latent variables. Let $q_{\boldsymbol{w}}(\boldsymbol{z})$ be a simpler distribution with parameters $\boldsymbol{w}$. VI algorithms minimize the (negative) "evidence lower bound"

$$-\operatorname{ELBO}(\boldsymbol{w}) = \underbrace{\mathop{\mathbb{E}}_{\mathrm{z}\sim q_{\boldsymbol{w}}}\left[-\log p(\mathrm{z}, \boldsymbol{x})\right]}_{\text{Energy term } l(\boldsymbol{w})} + \underbrace{\mathop{\mathbb{E}}_{\mathrm{z}\sim q_{\boldsymbol{w}}}\left[\log q_{\boldsymbol{w}}(\mathrm{z})\right]}_{\text{Neg-Entropy term } h(\boldsymbol{w})}, \tag{1}$$

equivalent to minimizing the KL-divergence from $q_{\boldsymbol{w}}(\boldsymbol{z})$ to $p(\boldsymbol{z}|\boldsymbol{x})$.

Traditionally, this was done with message-passing algorithms. This requires that q and p be relatively simple, essentially so that expectations of parts of $\log p$ can be computed with respect to q (Ghahramani and Beal, 2001; Winn and Bishop, 2005; Blei et al., 2017). Recent work (e.g. Salimans and Knowles, 2013; Wingate and Weber, 2013; Ranganath et al., 2014; Regier et al., 2017a; Kucukelbir et al., 2017) has focused on a "black box" model where the algorithm can only evaluate $\log p(\boldsymbol{z}, \boldsymbol{x})$ or its gradient $\nabla_{\boldsymbol{z}} \log p(\boldsymbol{z}, \boldsymbol{x})$ at chosen points $\boldsymbol{z}$. The key idea is that it is still possible to create an unbiased estimator of the gradient of ELBO, and therefore to optimize it through stochsatic gradient methods. This strategy applies to a large range of distributions, and is widely used.

It is important to know when black-box inference will work. While often empirically successful, black-box VI can and does fail to find the optimum (Yao et al., 2018; Regier et al., 2017b; Fan et al., 2015). Stochastic optimization convergence guarantees (Bottou et al., 2016) typically require two types of assumptions:

- *Gradient variance* must be controlled. The variance of VI gradient estimators has been studied (Fan et al., 2015; Xu et al., 2018; Domke, 2019), leading to the result that if $\log p$ is smooth, then the variance of reparameterization gradient estimators can be controlled. While an important step, these results alone cannot fully explain convergence behavior.
- Structural properties of the *objective itself* are needed. Fig. 1 shows an example where *exact* gradients are available. While a careful step-size and initialization appear to lead to convergence, other times there are worrying "jumps". Is any general guarantee possible?

The ultimate goal of the line of research in this paper is to obtain full convergence guarantees for practical black-box variational inference algorithms. The results in this paper are a step towards that goal.

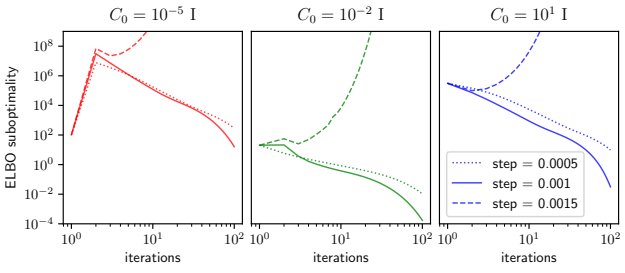

Figure 1. Gradient descent on `fires` with various step-sizes, initialized with $\boldsymbol{m} = 0$ and various C. Results are sensitive to both initialization and the stepsize. Can this be explained?

[1]College of Computing and Information Sciences, University of Massachusetts, Amherst, USA. Correspondence to: Justin Domke <domke@cs.umass.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

One very fundamental property is Lipschitz *smoothness* which means the gradient cannot change too quickly. Formally, a function f is M-smooth in the l_2 norm if $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x}-\boldsymbol{y}\|_2$. For non-convex objectives, essentially all convergence guarantees require smoothness, both for regular stochastic gradient descent (SGD) (Lee et al., 2016; Ge et al., 2015; Ghadimi and Lan, 2013), proximal SGD (Ghadimi et al., 2016), or momentum or "accelerated" SGD (Yang et al., 2016). Convergence guarantees are possible for convex objectives with or without smoothness(Bottou et al., 2016; Rakhlin et al., 2012).

Because this property is so fundamental, several works on variational inference have *assumed* that the VI objective (or part of it) is smooth. These include:

- Khan et al. (2015) Sec. 4
- Khan et al. (2016) Assumption A1
- Regier et al. (2017a) Condition 1
- Fan et al. (2016) Thm. 1
- Buchholz et al. (2018) Thm. 1
- Mohamad et al. (2018) Sec. 3.2
- Alquier and Ridgway (2017) Assumption 3.2.

Yet, to the best of our knowledge, no rigorous guarantees that this is actually true are known. The purpose of this paper is to fill that theoretical gap by providing conditions under which smoothness provably holds.

1.1. Contributions

Smoothness of the energy: (Thm. 1) Our main result is more general than variational inference: If $f(\boldsymbol{z})$ is M-smooth then $\mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z})$ is M-smooth over $\boldsymbol{w}$, when $q_{\boldsymbol{w}}$ is in the location-scale family with a "standardized" base distribution. In particular, if $\log p(\boldsymbol{z}, \boldsymbol{x})$ is M-smooth over $\boldsymbol{z}$ then the energy $l(\boldsymbol{w})$ in Eq. (1) is M-smooth. This requires no convexity assumptions.

Solution guarantees: (Thm. 7) Intuitively, structural properties of $\log p$ should imply properties of the optimal parameters $\boldsymbol{w}$. Using the above smoothness result, we show that at the optimal $\boldsymbol{w}$, all eigenvalues of the covariance of q are at least $1/M$. This is important because the neg-entropy in Eq. (1) is non-smooth when eigenvalues are very small.

Convergence considerations: Even if $\log p$ is smooth, the ELBO is *not* smooth, due to the entropy. We propose two solutions: a *projected* gradient descent scheme that leverages the above solution guarantee and a *proximal* scheme that uses the full structure of the entropy.

Understanding plain gradient descent: Given that the full ELBO is non-smooth, why does *plain* gradient descent sometimes succeed, and sometimes – even with exact gradients – produce huge "jumps" as seen in Fig. 1? We give insight into this using the smoothness result.

As a minor contribution, we extend existing work (Challis and Barber, 2013; Titsias and Lázaro-gredilla, 2014) to show that if $-\log p(\boldsymbol{z}, \boldsymbol{x})$ happens to be strongly-convex over $\boldsymbol{z}$ then so is the energy term $l(\boldsymbol{w})$ (Thm. 9). This gives another parameter-space solution guarantee where essentially the covariance of q cannot be too large (Thm. 10).

2. Preliminaries

A multivariate **location-scale family** (Geyer, 2011) is the result of drawing a sample from a "base" distribution and applying an affine transformation to it. Formally,

$$\mathsf{z} \sim \mathrm{LocScale}(\boldsymbol{m}, C, s) \iff \mathsf{z} \stackrel{d}{=} C\mathsf{u} + \boldsymbol{m},\ \mathsf{u} \sim s, \quad (2)$$

where $\stackrel{d}{=}$ indicates equality in distribution.

Black-box VI using these families was studied by Titsias and Lázaro-gredilla (2014). A simple example is the multivariate Gaussian, for which $\mathrm{LocScale}(\boldsymbol{m}, C, \mathcal{N}(0, I)) = \mathcal{N}(\boldsymbol{m}, CC^\top)$. Many families are representable, e.g. elliptical distributions such as the multivariate Student-T or Cauchy distributions. More generally, the base distribution need not be symmetric.

Notation. Let $\boldsymbol{w} = (\boldsymbol{m}, C)$ be a vector containing all components of $\boldsymbol{m}$ and C. We write $q_{\boldsymbol{w}}$ to denote $\mathrm{LocScale}(\boldsymbol{m}, C, s)$, leaving s implicit. Proofs use $\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u}) = C\boldsymbol{u} + \boldsymbol{m}$ to denote the affine mapping determined by parameters $\boldsymbol{w}$. $A \preceq B$ means that $B - A$ is positive semidefinite. We assume $\boldsymbol{z} \in \mathbb{R}^d$. Sans-serif font ($\mathsf{u}, \mathsf{z}$) distinguishes random variables.

Density. If the base distribution has a density and C is invertible, then the location-scale distribution also has a density, which is $q_{\boldsymbol{w}}(\boldsymbol{z}) = \mathrm{LocScale}(\boldsymbol{z}|\boldsymbol{m}, C, s) = \frac{1}{|C|} s\left(C^{-1}(\boldsymbol{z} - \boldsymbol{m})\right)$.

Entropy. The entropy of random variables under affine transformations is $\mathrm{Entropy}[A\mathsf{u} + \boldsymbol{b}] = \mathrm{Entropy}[\mathsf{u}] + \log|\det A|$ (Cover and Thomas, 2006, Sec. 8.6). Thus, the neg-entropy is $h(\boldsymbol{w}) = -\mathrm{Entropy}[s] - \log|\det C|$, with gradient $\nabla h(\boldsymbol{w}) = \left(0, -C^{-\top}\right)$.

Standardized Representations. We say that s is "standardized" if it has mean zero and unit variance, i.e. $\mathbb{E}_{\mathsf{u}\sim s}\,\mathsf{u} = 0$ and $\mathbb{V}_{\mathsf{u}\sim s}\,\mathsf{u} = I$. When s is standardized, the mean of the location-scale distribution is $\boldsymbol{m}$ while the variance is $CC^\top$.

A result we will use in Sec. 3.3 is that any location-scale family can be represented using a standardized base distribution, provided the variance exists: If s has mean μ and variance Σ, then $s' = \text{LocScale}(-\Sigma^{-1/2}\mu, \Sigma^{-1/2}, s)$, is standardized, yet $\text{LocScale}(\boldsymbol{m}, C, s')$ and $\text{LocScale}(\boldsymbol{m}, C, s)$ index the same set of distributions.

Bessel's inequality states that if $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k\}$ are orthonormal in some inner-product $\langle \cdot, \cdot \rangle$ with corresponding norm $\|\cdot\|$, then

$$\sum_{i=1}^{k} |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 \leq \|\boldsymbol{x}\|^2 . \tag{3}$$

This can be seen as a generalization of the Cauchy-Schwarz inequality $|\langle \boldsymbol{y}, \boldsymbol{x} \rangle|^2 \leq \|\boldsymbol{y}\|^2 \|\boldsymbol{x}\|^2$, which follows from using the singleton set $\{\boldsymbol{a}_1\} = \{\boldsymbol{y}/\|\boldsymbol{y}\|\}$ (Rooin and Bayat, 2012; Hasegawa and Karamakar).

2.1. Convergence Guarantees and Smoothness

It is impossible to review the vast optimization literature relevant to solving Eq. (1): There are many algorithms (gradient descent, stochastic gradient descent, momentum or accelerated variants, proximal or mirror descent variants) that can be analyzed with different hyper-parameters (step-sizes, iterate averaging) yielding different types of guarantees. These guarantees depend on properties of the target objective being optimized. Different distributions $p(\boldsymbol{x}, \boldsymbol{z})$ different properties for the objective (smoothness, convexity, strong convexity) or gradient estimators of it (variance bounds).

Still, at a very high level, the story is simple. To the best of our knowledge, all existing convergence guarantees require *either* smoothness (to guarantee a stationary point) *or* convexity (to guarantee a global optima), or both. For concreteness, suppose f is *only* known to be smooth (and possibly non-convex): Ghadimi and Lan (2013) analyze the the iteration $\mathsf{w}_{n+1} = \mathsf{w}_n - \gamma \mathsf{g}_n$ where $\mathsf{g}_1, \cdots \mathsf{g}_N$ are independent and $\mathbb{E}\,\mathsf{g}_n = \nabla f(\mathsf{w}_n)$. A simplified statement of their result is as follows: If (i) f is M-smooth, (ii) $\mathbb{E}\|\mathsf{g}_n - \nabla f(\mathsf{w}_n)\|_2^2 \leq \sigma^2$, and (iii) the starting point w_1 obeys $\|\mathsf{w}_1 - \boldsymbol{w}^*\|_2 \leq D$, then with a step-size of $\gamma = \min(1/M, D/(\sigma\sqrt{N}))$,

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\,\|\nabla f(\mathsf{w}_n)\|^2 \leq \frac{M^2 D^2}{N} + \frac{2DM\sigma}{\sqrt{N}}, \tag{4}$$

where the expectation is over all possible executation traces $\mathsf{w}_1, \cdots, \mathsf{w}_N$.

Both the step-size and convergence rate depend on smoothness. If there is noise ($\sigma > 0$) the convergence is $1/\sqrt{N}$. With no noise ($\sigma = 0$) convergence is $1/N$. Similar rates are known for proximal or projected stochastic gradient descent (Ghadimi et al., 2016) and stochastic gradient descent with momentum or Nesterov acceleration (Yang et al., 2016). Recent work seeks to understand when these iterations will converge to a (local) minima instead of a saddle point (Ge et al., 2015; Lee et al., 2016); here too, smoothness is a key assumption.

Similar guarantees are possible if the objective is convex or strongly convex, without requiring smoothness (Rakhlin et al., 2012; Bottou et al., 2016; Bubeck, 2015, Section 6.2). When gradients are stochastic, it may be helpful to average "minibatches" of gradient estimates, both in the convex (Bubeck, 2015, Section 6.2) and non-convex cases (Ghadimi et al., 2016).

3. Smoothness of the energy

In this section, we set $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z})$. The energy $l(\boldsymbol{w})$ in Eq. (1) is recovered when $f(\boldsymbol{z}) = -\log p(\boldsymbol{z}, \boldsymbol{x})$ (since $\boldsymbol{x}$ is constant). This is done to simplify the notation and because the results apply to general f and might be of independent interest.

3.1. Main Result

The following is the main technical result of this paper. It states that if $q_{\boldsymbol{w}}$ is a location-scale family with a zero-mean, unit variance base distribution and $f(\boldsymbol{z})$ is M-smooth, then $l(\boldsymbol{w})$ is also M-smooth. We emphasize that f is *not* assumed to be convex.

Theorem 1. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized base distribution s. If $f(\boldsymbol{z})$ is M-smooth, then $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z})$ is also M-smooth.*

Before proving this, we give four technical lemmas, all proven in Sec. 9 (in the supplement). The idea is to define a certain inner-product $\langle \cdot, \cdot \rangle_s$ over functions and a set of orthonormal functions $\{\boldsymbol{a}_i\}$ such that derivatives $l(\boldsymbol{w})$ can be written as an inner-product of $\boldsymbol{a}_i$ and $\nabla f \circ \boldsymbol{t}_w$ in $\langle \cdot, \cdot \rangle_s$, where $\circ$ indicates composition of functions.

Lemma 2. *$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_s = \mathbb{E}_{\mathsf{u}\sim s}\, \boldsymbol{a}(\mathsf{u})^\top \boldsymbol{b}(\mathsf{u})$ is a valid inner-product on squared-integrable $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^k$.*

The proof consists of verifying each of the defining properties of an inner-product.

Lemma 3. *Let $\boldsymbol{a}_i(\boldsymbol{u}) = \frac{d}{dw_i}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u})$. This is independent of $\boldsymbol{w}$ and $\frac{dl(\boldsymbol{w})}{dw_i} = \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} \rangle_s$.*

This is proven by first verifying that both $\frac{d}{dC_{ij}}\boldsymbol{t}_w(\boldsymbol{u})$ and $\frac{d}{dm_i}\boldsymbol{t}_w(\boldsymbol{u})$ are independent of $\boldsymbol{w}$, then calculating $\frac{dl}{dw_i}$ and performing some manipulations.

Lemma 4. *If s is standardized, then the functions $\{\boldsymbol{a}_i\}$ are orthonormal in $\langle \cdot, \cdot \rangle_s$.*

To prove this, note that the components $\boldsymbol{a}_i$ have two "types" namely $\frac{d}{dC_{ij}}\boldsymbol{t}_w(\boldsymbol{u})$ and $\frac{d}{dm_i}\boldsymbol{t}_w(\boldsymbol{u})$. Thus, an inner-product

$\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle_s$ reduces to the expected inner product of two such terms. It can be shown that this inner-product is one when $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ have the same type and indices, and zero otherwise.

Lemma 5. *If s is standardized, then* $\mathbb{E}_{\mathsf{u}\sim s} \left\| \boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u}) - \boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u}) \right\|_2^2 = \left\| \boldsymbol{w} - \boldsymbol{v} \right\|_2^2$.

This is shown by substituting the exact form of $\boldsymbol{t}_{\boldsymbol{w}}$ and $\boldsymbol{t}_{\boldsymbol{v}}$. Taking the expectation leads a result of $\|\Delta C\|_F^2 + \|\Delta \boldsymbol{m}\|_2^2$, where $\Delta \boldsymbol{m}$ denotes the difference of the $\boldsymbol{m}$ components of $\boldsymbol{w}$ and $\boldsymbol{v}$, and similarly for ΔC. This is equivalent to the squared Euclidean distance of $\boldsymbol{w}$ and $\boldsymbol{v}$.

Proof of Thm. 1. Take two parameter vectors, $\boldsymbol{w}$ and $\boldsymbol{v}$. Apply Lem. 3 to each component of the gradients $\nabla l(\boldsymbol{w})$ and $\nabla l(\boldsymbol{v})$ to get that

$$\begin{aligned}
&\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 \\
&= \sum_i \left(\langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} \rangle_s - \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}} \rangle_s \right)^2 \\
&= \sum_i \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} - \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}} \rangle_s^2 .
\end{aligned}$$

Lem. 4 showed that the functions $\{\boldsymbol{a}_i\}$ are orthonormal in the inner-product $\langle \cdot, \cdot \rangle_s$. Thus, by Bessel's inequality,

$$\begin{aligned}
\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 &\leq \|\nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} - \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}}\|_s^2 , \qquad (5) \\
&= \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \|\nabla f\left(\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u})\right) - \nabla f\left(\boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u})\right)\|_2^2
\end{aligned}$$

where $\|\cdot\|_s$ denotes the norm corresponding to $\langle \cdot, \cdot \rangle_s$. Now apply the smoothness of f to get that

$$\begin{aligned}
\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 &\leq M^2 \mathop{\mathbb{E}}_{\mathsf{u}\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u}) - \boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u})\|_2^2 \qquad (6) \\
&= M^2 \|\boldsymbol{w} - \boldsymbol{v}\|_2^2 , \qquad (7)
\end{aligned}$$

where the last equality follows from Lem. 5. □

Note that the *only* inequalities used in this proof are (i) Bessel's inequality and (ii) the bound on the difference of gradients of f provided by the assumption that f is M-smooth. Thus, the tightness of the final bound that $\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2 \leq M \|\boldsymbol{w} - \boldsymbol{v}\|_2$ is determined by the tightness of these two inequalities. It's natural to ask when this bound will be tight or loose. The following section will show that it is tight when f is closer to an isotropic quadratic. On the other hand, the starting assumption that f is smooth $\|\nabla f(\boldsymbol{x}) - \nabla l(\boldsymbol{y})\|_2 \leq M \|\boldsymbol{x} - \boldsymbol{y}\|_2$ might often be loose, e.g. if f is much smoother in "some directions" than others. In this case, the final bound on l will also be loose due to looseness created in moving from Eq. (5) to Eq. (6) (even M might still be the best possible *smoothness constant*).

The idea of expressing the gradient using a fixed base distribution and a transformation $\boldsymbol{t}_w(\boldsymbol{u})$ is also used in "reparameterization" type estimators (Titsias and Lázaro-gredilla, 2014; Rezende et al., 2014; Kingma and Welling, 2014). Smoothness, however, is a deterministic property of the function $l(\boldsymbol{w})$, independent of any method one might use for estimating or optimizing it.

3.2. Unimprovability

This section gives an example function $f(\boldsymbol{z})$ that is M-smooth, but leads to a function $l(\boldsymbol{w})$ that is M-smooth (but not smoother), meaning that Thm. 1 is unimprovable. Intuitively, smoothness is a quadratic upper-bound. So, it is natural to suppose that $f(\boldsymbol{z})$ is *exactly* quadratic. The following shows that in this case, $l(\boldsymbol{w})$ has a closed form.

Theorem 6. *Let $q_{\boldsymbol{w}} = \mathrm{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized base distribution s and let $f(\boldsymbol{z}) = \frac{a}{2} \|\boldsymbol{z} - \boldsymbol{z}^*\|_2^2$. Then $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z}\sim q_{\boldsymbol{w}}} f(\mathsf{z}) = \frac{a}{2}(\|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 + \|C\|_F^2)$.*

To see that Thm. 1 is unimprovable, define $\bar{\boldsymbol{w}} = (\boldsymbol{z}^*, 0_{d,d})$, where $0_{d,d}$ is a $d \times d$ matrix of zeros. Then, $l(\boldsymbol{w}) = \frac{a}{2}\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|_2^2$. This is M-smooth for $M = a$, but not for any smaller value.

3.3. Solution Guarantees

Intuitively, properties of the target distribution $p(\boldsymbol{z}|\boldsymbol{x})$ might imply properties of the variational distribution q_{w^*} at the optimal parameters $\boldsymbol{w}^*$. In particular, if $\log p(\boldsymbol{z}, \boldsymbol{x})$ is smooth over $\boldsymbol{z}$, then it is "spread out" so we might expect that $q_{\boldsymbol{w}^*}$ would also be. This section formalizes and proves a version of this intuition. This will be used in Sec. 5 to give a convergence guarantee for projected stochastic gradient descent. The core idea is that the ELBO in Eq. (1) is poorly conditioned for low-variance distributions. However, if we can guarantee that the optimum lies in a well-conditioned region, we can constrain optimization to that region.

We define $\mathcal{W}_M$ to be the set of parameters where all singular values of C are at least $1/\sqrt{M}$, i.e.

$$\mathcal{W}_M = \left\{ (\boldsymbol{m}, C) \;\middle|\; \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\} . \qquad (8)$$

We could equivalently define $\mathcal{W}_M$ to be the set of of parameters where all eigenvalues of $CC^\top$ are at least $\frac{1}{M}$, i.e. $CC^\top \succeq \frac{1}{M} I$. Recall from Sec. 2 that for standardized s, $\mathbb{V}_{\mathsf{z}\sim q_{\boldsymbol{w}}} \mathsf{z} = CC^\top$, so this is the parameters with variance at least $\frac{1}{M} I$.

The following result shows that if minimizing the ELBO with a smooth target distribution, the optimal parameters must fall in $\mathcal{W}_M$, i.e. the variance cannot be smaller than $\frac{1}{M} I$. This requires the stronger assumption that the base distribution s is spherically symmetric.

Theorem 7. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized and spherically symmetric base distribution s. Suppose $\boldsymbol{w}$ minimizes $l(\boldsymbol{w}) + h(\boldsymbol{w})$ from Eq. (1) and $\log p(\boldsymbol{z}, \boldsymbol{x})$ is M-smooth over $\boldsymbol{z}$. Then, $\boldsymbol{w} \in \mathcal{W}_M$.*

The proof of this theorem (in Sec. 11) first establishes the following Lemma.

Lemma 8. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized and spherically symmetric base distribution s. Let $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} f(\mathsf{z})$. Suppose C is diagonal and f is M-smooth. Then, $|\frac{dl(\boldsymbol{w})}{dC_{ii}}| \leq M|C_{ii}|$.*

The proof of Lem. 8 first shows that if $C_{ii} = 0$, then $\frac{dl}{dC_{ii}} = 0$, which uses that s is symmetric. Then, given an arbitrary $\boldsymbol{w}$, let $\boldsymbol{w}'$ be $\boldsymbol{w}$ with C_{ii} set to zero. Since we know from Thm. 6 that l is M-smooth, we then get that $|dl(\boldsymbol{w})/dC_{ii}| \leq \|\nabla l(\boldsymbol{w}') - \nabla l(\boldsymbol{w})\|_2 \leq M|C_{ii}|$.

Now, the proof of Thm. 7 uses the fact that if $\boldsymbol{w}$ is a minimum, then $\nabla l(\boldsymbol{w}) = -\nabla h(\boldsymbol{w})$. If C happens to be diagonal, the result is easy to show using the previous lemma along with the exact gradient of h. Given an arbitrary C, we can use the singular value decomposition of C to define another M-smooth function which must have a diagonal solution.

4. Analogous Result for Convex Functions

Smoothness and strong convexity are complementary in that they give upper and lower bounds on the eigenvalues of the Hessian. As a minor contribution, we observe that a guarantee complementary to Thm. 1 holds: if $-\log p$ is (strongly) convex, then so is $l(\boldsymbol{w})$. The example in Thm. 6 shows this result is also unimprovable.

Theorem 9. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$. If $f(\boldsymbol{z})$ is convex, then $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} f(\mathsf{z})$ is also convex. If, in addition, s is standardized and $f(\boldsymbol{z})$ is c-strongly convex, then $l(\boldsymbol{w})$ is also c-strongly convex.*

Proof. **(Convexity)** Represent l as $l(\boldsymbol{w}) = \mathbb{E}_{\mathsf{u} \sim s} f(\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u}))$ where $\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u}) = C\boldsymbol{u} + \boldsymbol{m}$. For fixed $\boldsymbol{u}$, $\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u})$ is *linear* in $\boldsymbol{w}$. Thus, given any two parameter vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ and any $\alpha, \beta \in (0, 1)$ with $\alpha + \beta = 1$, since f is convex, $l(\alpha\boldsymbol{w} + \beta\boldsymbol{v})$ is equal to

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{u} \sim s} f(\boldsymbol{t}_{\alpha\boldsymbol{w}+\beta\boldsymbol{v}}(\mathsf{u})) &= \mathop{\mathbb{E}}_{\mathsf{u} \sim s} f(\alpha\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u}) + \beta\boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u})) \\
&\leq \mathop{\mathbb{E}}_{\mathsf{u} \sim s} \alpha f(\boldsymbol{t}_{\boldsymbol{w}}(\mathsf{u})) + \beta f(\boldsymbol{t}_{\boldsymbol{v}}(\mathsf{u})) \\
&= \alpha l(\boldsymbol{w}) + \beta l(\boldsymbol{v}).
\end{aligned}$$

(Strong convexity) If f is c-strongly convex then $f(\boldsymbol{z}) = f_0(\boldsymbol{z}) + \frac{c}{2}\|\boldsymbol{z}\|_2^2$ for some convex function f_0. Thus, $l(\boldsymbol{w}) = l_0(\boldsymbol{w}) + \frac{c}{2}\mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} \|\mathsf{z}\|_2^2$, where $l_0(\boldsymbol{w}) = \mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} f_0(\mathsf{z})$ is convex by the previous reasoning. Then, it isn't too hard to show that $\mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} \|\mathsf{z}\|_2^2 = \mathbb{E}_{\mathsf{u} \sim s} \|C\mathsf{u} + \boldsymbol{m}\|_2^2 = \|C\|_F^2 + \|\boldsymbol{m}\|_2^2 = \|\boldsymbol{w}\|_2^2$. Thus, we have that $l(\boldsymbol{w}) = l_0(\boldsymbol{w}) + \frac{c}{2}\|\boldsymbol{w}\|^2$ is c-strongly convex. □

The convexity result (and proof) is essentially the same as that of Titsias and Lázaro-gredilla (2014, Appendix, Proposition 1). The strong-convexity result generalizes a previous result due to Challis and Barber (2013, Sec. 3.2) who give a strong-convexity guarantee for Gaussian variational distributions applied to targets with Gaussian priors.

The following result gives a bound on the location of the optimal parameters. The proof uses the fact that, at the optimum, $\nabla l(\boldsymbol{w}) = -\nabla h(\boldsymbol{w})$, so the exact gradient is known. However, strong convexity means that only certain gradients are possible at a given part of parameter space. (This result is complementary to Thm. 7.)

Theorem 10. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized and spherically symmetric base distribution s. Suppose $\boldsymbol{w}$ minimizes $l(\boldsymbol{w}) + h(\boldsymbol{w})$ from Eq. (1) and $-\log p(\boldsymbol{z}, \boldsymbol{x})$ is c-strongly convex over $\boldsymbol{z}$. Then, $\|C\|_F^2 + \|\boldsymbol{m} - \boldsymbol{z}^*\|_2^2 \leq \frac{d}{c}$, where $\boldsymbol{z}^* = \text{argmax}_{\boldsymbol{z}} \log(\boldsymbol{z}, \boldsymbol{x})$.*

5. Convergence Considerations

In optimizing the ELBO in Eq. (1), the negative entropy term h creates complications. The gradient is $\nabla h(\boldsymbol{w}) = (0, -C^{-\top})$ (Sec. 2), which can change arbitrarily rapidly when the singular values of C are close to zero. So $h(\boldsymbol{w})$ is *not* Lipschitz-smooth, posing a challenge for establishing convergence guarantees for pure gradient descent applied to the full ELBO. In this section, we consider two strategies for coping with this: *Projected* gradient descent, and *proximal* gradient descent. Finally, we seek to understand the performance of *regular* gradient descent seen in Fig. 1.

The following result gives one way of dealing with the fact that the negentropy is non-smooth. As in Sec. 3.3, this requires the additional assumption that the base distribution is spherically symmetric.

Theorem 11. *Let $q_{\boldsymbol{w}} = \text{LocScale}(\boldsymbol{m}, C, s)$ with parameters $\boldsymbol{w} = (\boldsymbol{m}, C)$ and a standardized and spherically symmetric base distribution s. Suppose $\log p(\boldsymbol{z}, \boldsymbol{x})$ is M-smooth. Then $\text{ELBO}(\boldsymbol{w})$ as in Eq. (1) is $2M$ smooth over $\mathcal{W}_M$ and if $\boldsymbol{w}^*$ is an optima of $\text{ELBO}(\boldsymbol{w})$, then $\boldsymbol{w}^* \in \mathcal{W}_M$.*

The proof (in Sec. 13) first shows that h is M-smooth over $\mathcal{W}_M$ by taking two arbitrary parameter vectors $\boldsymbol{w}, \boldsymbol{v} \in \mathcal{W}_M$ and using a matrix norm inequality to bound the difference of the gradients $\nabla h(\boldsymbol{w})$ and $\nabla h(\boldsymbol{v})$. Then, we combine our main result that l is smooth (Thm. 1) with the bound on the location of the optimum (Thm. 7) and the fact that h is

smooth over $\mathcal{W}_M$. (By the triangle inequality, the sum of two M-smooth functions is $2M$ smooth.)

Given this result, a natural approach to optimizing the ELBO is to use projected (stochastic) gradient descent, i.e. to iterate $\mathsf{w}' = \mathrm{proj}_{\mathcal{W}_M}(\mathsf{w} - \gamma \mathsf{g})$ where $\mathsf{g} = \nabla l(\mathsf{w}) + \nabla h(\mathsf{w})$ (or a stochastic estimator) and $\mathrm{proj}_{\mathcal{W}}$ is Euclidean projection. Thm. 13 (in Sec. 13, supplement) shows that if $\boldsymbol{w} = (\boldsymbol{m}, C)$, and C has singular value decomposition $C = USV^\top$, then

$$\mathrm{proj}_{\mathcal{W}_M}(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{W}_M} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2 = (\boldsymbol{m}, UTV^\top)$$

where T is a diagonal matrix with $T_{ii} = \max(S_{ii}, 1/\sqrt{M})$.

Another way of dealing the fact that h is non-smooth is to use *proximal* optimization (Beck and Teboulle, 2009; Parikh, 2014; Bubeck, 2015; Ghadimi et al., 2016; Ghadimi and Lan, 2012). Intuitively, the idea is as follows: With a step-size γ gradient descent on $l + h$ gives the update $\mathsf{w}' = \mathsf{w} - \gamma(\nabla l(\mathsf{w}) + \nabla h(\mathsf{w}))$, which can equivalently be seen as minimizing a linear approximation of $l + h$ at $\boldsymbol{w}$, with a quadratic penalty, i.e. setting

$$\mathsf{w}' = \operatorname*{argmin}_{\boldsymbol{v}} l(\mathsf{w}) + h(\mathsf{w}) + \langle \nabla l(\mathsf{w}) + \nabla h(\mathsf{w}), \boldsymbol{v} - \mathsf{w} \rangle + \frac{1}{2\gamma} \|\boldsymbol{v} - \mathsf{w}\|^2, \quad (9)$$

If $h(\boldsymbol{w})$ is non-smooth, even if $\boldsymbol{v}$ is close to w, $h(\mathsf{w}) + \langle \nabla h(\mathsf{w}), \boldsymbol{v} - \mathsf{w} \rangle$ can be an arbitrarily poor approximation of $h(\boldsymbol{v})$. Thus, a natural idea is to leave h *unapproximated*, i.e. to linearize l only. This would mean instead using

$$\mathsf{w}' = \operatorname*{argmin}_{\boldsymbol{v}} l(\mathsf{w}) + \langle \nabla l(\mathsf{w}), \boldsymbol{v} - \mathsf{w} \rangle + h(\boldsymbol{v}) + \frac{1}{2\gamma} \|\boldsymbol{v} - \mathsf{w}\|^2. \quad (10)$$

This is equivalent to

$$\mathsf{w}' = \operatorname*{prox}_{\gamma}[\mathsf{w} - \gamma \nabla l(\mathsf{w})],$$

where

$$\operatorname*{prox}_{\gamma}[\boldsymbol{w}] = \operatorname*{argmin}_{\boldsymbol{v}} h(\boldsymbol{v}) + \frac{1}{2\gamma} \|\boldsymbol{v} - \boldsymbol{w}\|_2^2.$$

Thm. 13 shows that if $\boldsymbol{w} = (\boldsymbol{m}, C)$ and C is triangular with a positive diagonal, then $\mathrm{prox}_\gamma(\boldsymbol{w}) = (\boldsymbol{m}, C + \Delta C)$, where ΔC is diagonal with $\Delta C_{ii} = \frac{1}{2}((C_{ii}^2 + 4\gamma)^{1/2} - C_{ii})$. Intuitively, this has the effect of keeping the diagonal entries away from 0: If C_{ii} is very small then $\Delta C_{ii} \approx \gamma$ while if C_{ii} is large, $\Delta C_{ii} \approx 0$. The proximal scheme has two advantages over projection. First, convergence rates depend on the smoothness constant of the linearized terms, which is M rather than $2M$. Second the proximal operator is faster to compute. prox takes $\Omega(d)$ time, while proj takes $\Omega(d^3)$ time, due to the need for a singular value decomposition.

6. Demonstration

To avoid complications related to stochastic gradients, we consider two settings where $l(\boldsymbol{w})$ and its gradient can be computed (nearly) exactly. Take a dataset $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$ and let X be a matrix with $\boldsymbol{x}_n$ on row n and $\boldsymbol{y}$ a vector of the values $(y_1, \cdots y_n)$. We model $p(\boldsymbol{z}, \boldsymbol{y}|X) = p(\boldsymbol{z}) \prod_{n=1}^N p(y_n|\boldsymbol{x}_n, \boldsymbol{z})$. The prior $p(\boldsymbol{z})$ is a standard Gaussian. We consider both linear regression with $p(y_n|\boldsymbol{x}_n, \boldsymbol{z}) = \mathcal{N}(y_n|\mu = \boldsymbol{z}^\top \boldsymbol{x}_n, \sigma^2 = 1)$ and binary logistic regression with $p(y_n|\boldsymbol{x}_n, \boldsymbol{z}) = \sigma(y_n \boldsymbol{x}_n^\top \boldsymbol{z})$. It can be shown that $\log p(\boldsymbol{z}, \boldsymbol{y}|X)$ is M-smooth with $M = 1 + \sigma_{\max}(XX^\top)$ for linear regression and $M = 1 + \frac{1}{4}\sigma_{\max}(XX^\top)$ for logistic regression.

For linear regression data (`boston`, `fires`), l has a closed form. For logistic regression (`australian`, `ionosphere`), we compute l via a redution to a set of pre-computed one dimensional integrals: Observe that for all $\boldsymbol{w}$, $\mathbb{E}_{\mathsf{z} \sim q_{\boldsymbol{w}}} \log p(y|\boldsymbol{x}, \mathsf{z}) = g(y\boldsymbol{x}^\top \boldsymbol{m}, \|C^\top \boldsymbol{x}\|_2)$, where $g(a, b) = \mathbb{E}_{\mathsf{t} \sim \mathcal{N}(0,1)} \log \sigma(a + b\mathsf{t})$. By pre-computing g over a grid of inputs (a, b) we can quickly evaluate $l(\boldsymbol{w})$ and its gradient via spline interpolation.

We initialize $\boldsymbol{m}$ to zero and $C = \rho I$ for a range of scaling constants ρ. Fig. 2 shows example results on two datasets. For projected or proximal gradient descent, simply initializing $C = 0$ is fine. For naive gradient descent, initialization is subtle, since too small a ρ leads to an enormous entropy gradient (and thus "jumps"), while for large ρ, all algorithms converge slowly.

Fig. 3 systematically varies ρ on various datasets. There are two seemingly strange behaviors for naive gradient descent. First, it performs very similarly to proximal gradient descent for large ρ. To understand this, note that when C is large, the entropy is locally nearly linear, and so a proximal step is similar to a naive step. Second, there is a near-symmetry between small and large ρ. Here, observe that if naive gradient descent is initialized with *small* ρ, the huge gradient of the entropy term will send the parameters to a correspondingly *large* C in the second iteration. In these examples, a carefully chosen ρ performs well, though this may be hard to find and there is no guarantee in general.

These results confirm the theory developed above. First, we see that proximal gradient descent always converges with a step-size of $\gamma = 1/M$. Thus suggests that M as derived

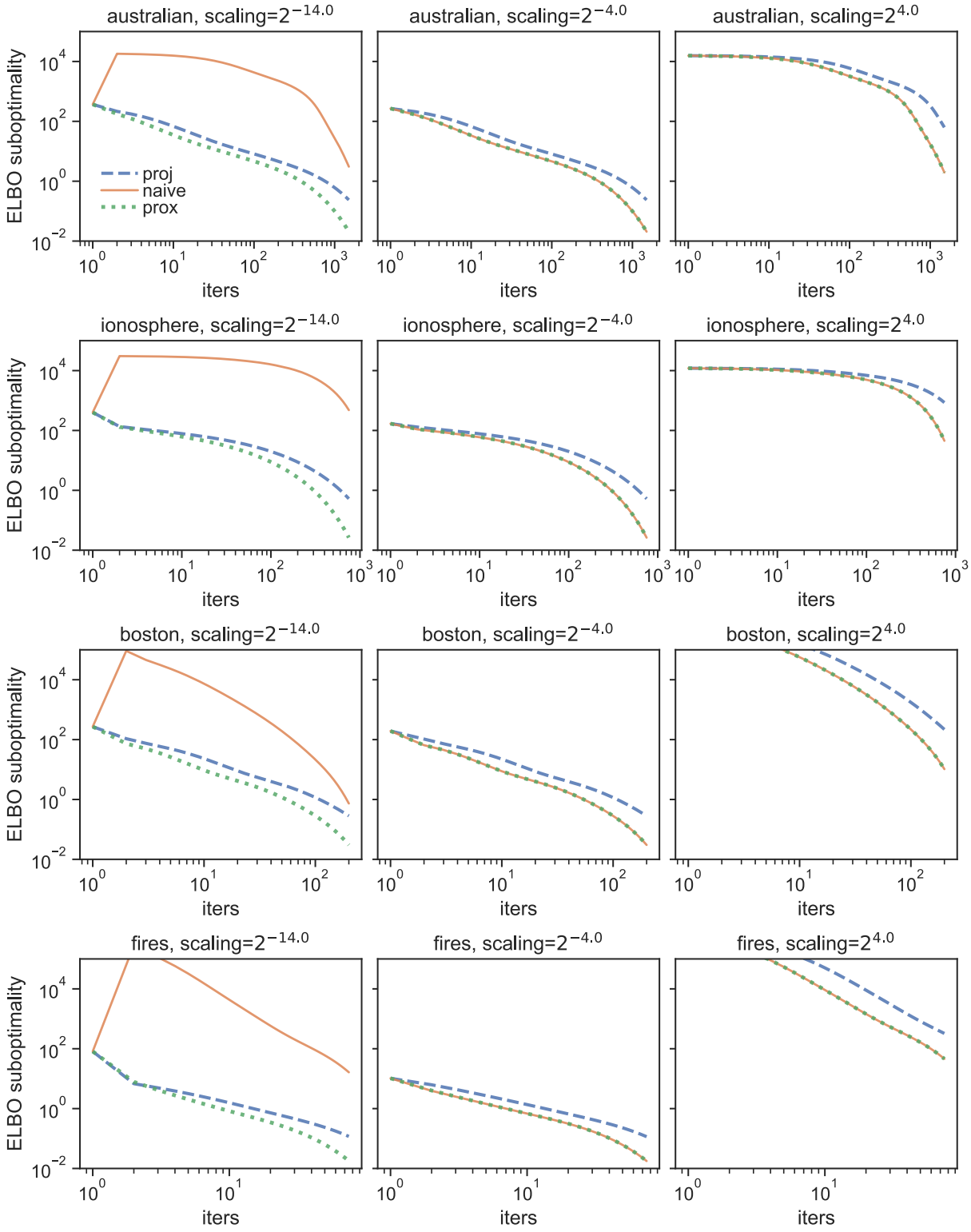

Figure 2. **Naive optimization can work well, but is sensitive to initialization.** Looseness of the objective obtained by naive gradient descent ($\gamma = 1/M$), projected gradient descent ($\gamma = 1/(2M)$) and proximal gradient descent ($\gamma = 1/M$). Optimization starts with $\boldsymbol{m} = 0$ and $C = \rho I$ where ρ is a scaling factor. Initializing $C = 0$ is fine for proximal or projected gradient descent, but naive gradient descent requires careful initialization. Results for other datasets in Sec. 8 (supplement).

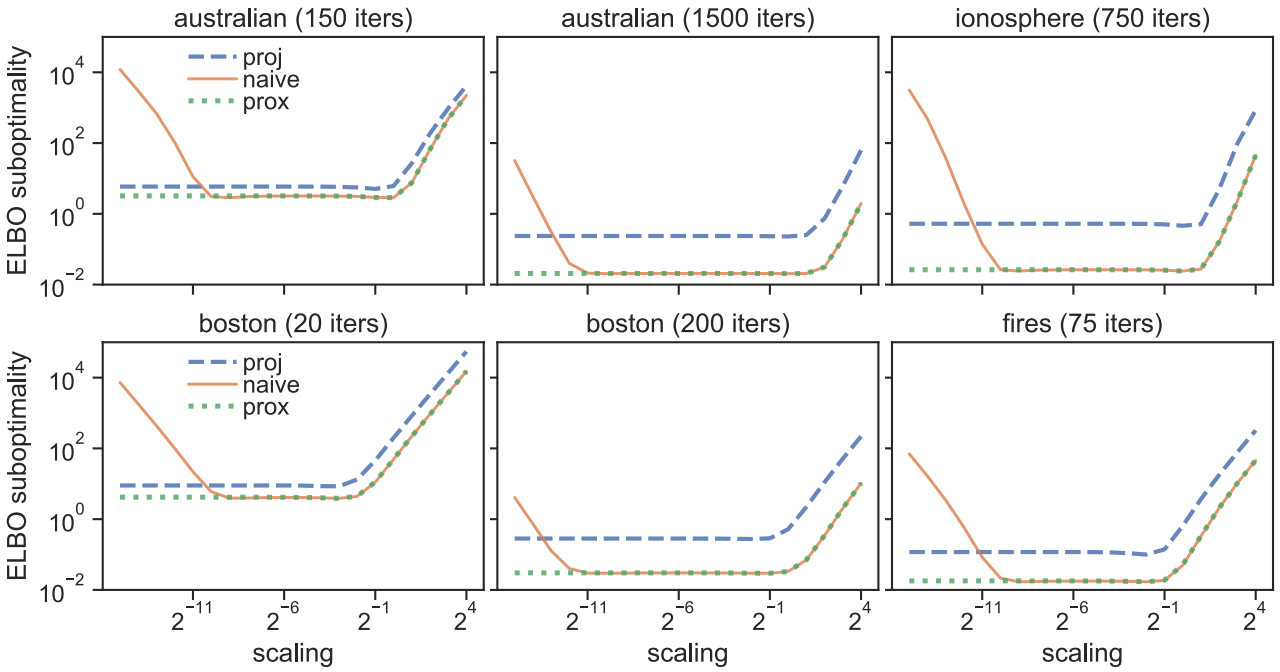

Figure 3. **Naive optimization is similar to proximal for large initial C, but worse for small C.** Results of optimizing the ELBO with different scaling factors ρ on four different datasets. The two right columns show results after enough iterations for proximal optimization to converge to less than $\approx 10^{-1}$. The left column shows results after $\frac{1}{10}$-th as many iterations. Proximal optimization starting with $C \approx 0$ always performs well. Projected gradient descent requires more iterations. Naive optimization can work well, but is not guaranteed and requires careful initialization.

in Thm. 1 is correct. Second, projected gradient descent always converges with a step-size of $1/(2M)$. This suggests that Thm. 11 is correct to assert that the optimal parameters $\boldsymbol{w}^* \in \mathcal{W}_M$ and that the ELBO is $2M$-smooth over $\mathcal{W}_M$. Finally, naive gradient "descent" truly can ascend when the parameters $\boldsymbol{w}$ start in the region where $h(\boldsymbol{w})$ is non-smooth, but behaves similarly to proximal gradient descent otherwise, confirming the discussion in Sec. 5.

7. Discussion

The primary contribution of this paper is to show that for VI with location-scale families, smoothness of $\log p(\boldsymbol{z}, \boldsymbol{x})$ implies smoothness of the free energy. This fills a theoretical gap relevant to many existing works (Khan et al., 2015; 2016; Khan and Lin, 2017; Regier et al., 2017a; Fan et al., 2016; Buchholz et al., 2018; Mohamad et al., 2018; Alquier and Ridgway, 2017). We also showed that result gives parameter-space guarantees on the location of the optimal parameters. As a minor contribution, we also give analogous guarantees for strong-convexity. Convergence guarantees for gradient-based optimization require *either* smoothness or convexity. Thus, at a very high level, this paper shows that if $\log p(\boldsymbol{z}, \boldsymbol{x})$ has the structure needed to guarantee finding $\boldsymbol{z}^* = \operatorname{argmax} \log p(\boldsymbol{z}, \boldsymbol{x})$, then it *also* has the structure to guarantee that VI with a location-scale family will converge.

While motivated by VI, the main results for smoothness (Thm. 1) and (strong) convexity (Thm. 9) are general properties of expectations parameterized by location-scale families, and so may be of independent interest.

There are several issues to consider when gauging the immediate practical impact of this work. Most importantly, the smoothness guarantee in this paper was already *true*, even if was not *known*. Thus, real-world black-box VI methods already benefit from it. Second, $\nabla l(\boldsymbol{w})$ typically must be *estimated*, and convergence guarantees need bounds on the fluctuations of the estimator. Finding better gradient estimators (and bounds) is an active research topic (Domke, 2019). Finally, the theory for projected and proximal gradient optimization is still evolving, particularly for non-convex objectives. It seems to be an open question if the "minibatches" of gradient estimates that current bounds (Ghadimi et al., 2016) use are truly required.

A result conceptually related to this paper's smoothness guarantee is used in variational boosting (Guo et al., 2016; Locatello et al., 2018): The functional gradient for non-parametric q is smooth if q is bounded below by a positive constant. While similar in spirit, this does not address traditional parametric VI.

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv:1706.09293 [cs, math, stat]*, 2017.

Amir Beck and Marc Teboulle. Gradient-based algorithms with applications to signal recovery. *Convex optimization in signal processing and communications*, pages 42–88, 2009.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *arXiv:1606.04838 [cs, math, stat]*, 2016.

Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

Alexander Buchholz, Florian Wenzel, and Stephan Mandt. Quasi-Monte Carlo Variational Inference. In *ICML*, 2018.

Edward Challis and David Barber. Gaussian Kullback-Leibler Approximate Inference. *Journal of Machine Learning Research*, 14:2239–2286, 2013.

T. M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, N.J, 2nd ed edition, 2006.

Justin Domke. Provable Gradient Variance Guarantees for Black-Box Variational Inference. In *NeurIPS*, 2019.

K. Fan, Y. Zhang, R. Henao, and K. Heller. Triply Stochastic Variational Inference for Non-linear Beta Process Factor Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 121–130, 2016.

Kai Fan, Ziteng Wang, Jeff Beck, James Kwok, and Katherine Heller. Fast Second-Order Stochastic Backpropagation for Variational Inference. In *NeurIPS*, 2015.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition. In *COLT*, page 46, 2015.

Charles J Geyer. Statistics 5101 Lecture Slides, Deck 5. http://www.stat.umn.edu/geyer/f11/5101/slides/s5.pdf, 2011.

S. Ghadimi and G. Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Guanghui Lan. Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: A Generic Algorithmic Framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Zoubin Ghahramani and Matthew Beal. Propagation Algorithms for Variational Bayesian Learning. In *NeurIPS*, 2001.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B. Dunson. Boosting Variational Inference. *arXiv:1611.05559 [cs, stat]*, 2016.

R Hasegawa and B Karamakar. Generalizations of Cauchy-Schwarz in Probability Theory.

Mohammad E. Khan, Pierre Baqué, François Fleuret, and Pascal Fua. Kullback-Leibler proximal variational inference. In *NeurIPS*, pages 3402–3410, 2015.

Mohammad Emtiyaz Khan and Wu Lin. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 878–887. PMLR, 2017.

Mohammad Emtiyaz Khan, Reza Babanezhad, Wu Lin, Mark Schmidt, and Masashi Sugiyama. Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions. In *UAI*, 2016.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient Descent Only Converges to Minimizers. In *COLT*, page 12, 2016.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Ratsch. Boosting Variational Inference: An Optimization Perspective. In *AISTATS*, pages 464–472, 2018.

Saad Mohamad, Abdelhamid Bouchachia, and Moamar Sayed-Mouchaweh. Asynchronous Stochastic Variational Inference. *arXiv:1801.04289 [cs, stat]*, 2018.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer, Place of publication not identified, 2014. OCLC: 878109549.

Neal Parikh. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, pages 449–456, 2012.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black Box Variational Inference. In *AISTATS*, 2014.

Jeffrey Regier, Michael I Jordan, and Jon McAuliffe. Fast Black-box Variational Inference through Stochastic Trust-Region Optimization. In *NeurIPS*, pages 2399–2408. Curran Associates, Inc., 2017a.

Jeffrey Regier, Michael I Jordan, and Jon McAuliffe. Fast Black-box Variational Inference through Stochastic Trust-Region Optimization. In *NeurIPS*, page 10, 2017b.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, 2014.

Jamal Rooin and Morteza Bayat. Equivalency of Cauchy-Schwarz and Bessel Inequalities. *The Mathematical Intelligencer*, 34(4):2–3, 2012.

Tim Salimans and David A. Knowles. Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. *Bayesian Anal.*, 8(4):837–882, 2013.

Michalis Titsias and Miguel Lázaro-gredilla. Doubly Stochastic Variational Bayes for non-Conjugate Inference. In *ICML*, 2014.

David Wingate and Theophane Weber. Automated Variational Inference in Probabilistic Programming. *arXiv:1301.1299 [cs, stat]*, 2013.

John Winn and Christopher M Bishop. Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

Ming Xu, Matias Quiroz, Robert Kohn, and Scott A. Sisson. On some variance reduction properties of the reparameterization trick. *arXiv:1809.10330 [cs, stat]*, 2018.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified Convergence Analysis of Stochastic Momentum Methods for Convex and Non-convex Optimization. *arXiv:1604.03257 [math, stat]*, 2016.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but Did It Work?: Evaluating Variational Inference. In *ICML*, 2018.