Benefits of Non-Linear Scale Parameterizations in Black Box Variational Inference through Smoothness Results and Gradient Variance Bounds

Alexandra Hotti
Klarna
KTH Royal Institute of Technology

Lennart Van der Goten
KTH Royal Institute of Technology
Science for Life Laboratory

Jens Lagergren
Science for Life Laboratory
KTH Royal Institute of Technology

Abstract

Black box variational inference has consistently produced impressive empirical results. Convergence guarantees require that the variational objective exhibits specific structural properties and that the noise of the gradient estimator can be controlled. In this work we study the smoothness and the variance of the gradient estimator for location-scale variational families with non-linear covariance parameterizations. Specifically, we derive novel theoretical results for the popular exponential covariance parameterization and tighter gradient variance bounds for the softplus parameterization. These results reveal the benefits of using non-linear scale parameterizations on large scale datasets. With a non-linear scale parameterization, the smoothness constant of the variational objective and the upper bound on the gradient variance decrease as the scale parameter becomes smaller. Learning posterior approximations with small scales is essential in Bayesian statistics with sufficient amount of data, since under appropriate assumptions, the posterior distribution is known to contract around the parameter of interest as the sample size increases. We validate our theoretical findings through empirical analysis on several large-scale datasets, underscoring the importance of non-linear parameterizations.

1 INTRODUCTION

Variational inference (VI; Blei et al. (2017)) is useful for approximating potentially complex posterior distributions using simpler variational posteriors. To find the variational posterior that most closely approximates the posterior, the evidence lower bound (ELBO), denoted as $\mathcal{L}(\phi)$, can be optimized using gradient descent. However, in many cases of interest, $\nabla\mathcal{L}(\phi)$ cannot be computed in closed form. As a result, black box VI (BBVI; Ranganath et al. (2014)) has become a popular alternative where Monte Carlo integration is used to form unbiased estimates of $\nabla\mathcal{L}(\phi)$.

Despite BBVI consistently producing impressive empirical results (Ranganath et al., 2014; Zhang et al., 2018; Ranganath et al., 2016; Salimans and Knowles, 2014; Tran et al., 2016; Kingma and Welling, 2022; Kviman et al., 2023, 2024), literature on convergence guarantees for BBVI remains relatively scarce. Domke et al. (2023) and Kim et al. (2023b) have as of recent made notable advancements, by establishing convergence guarantees for the location-scale family —a predominant choice in practice - for linear parameterizations of the scale matrix, wherein the scale components are optimized directly. However, in real-world applications, a non-linear transformation from $\mathbb{R}_{>0}$ to $\mathbb{R}$ is frequently used during optimization to ensure that positive-definite covariance matrices are learned.

A recent contribution in this direction comes from Kim et al. (2023b), who established convergence results for 1-Lipschitz continuous scale parameterizations, which includes the popular softplus transformation. However, they also identified certain drawbacks associated with non-linear parameterizations. Specifically, they showed that opting for a 1-Lipschitz parameterization, instead of a linear transformation, increases the smoothness constant. Additionally, non-linear scale parameterizations disrupt strong convexity, meaning that even if the posterior is strongly log-concave, the $\mathcal{L}(\phi)$ does not exhibit strong convexity.

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

In practice, the exponential transformation is another widely used parameterization (Kingma and Welling, 2019b; Kucukelbir et al., 2016). Regrettably, it is not 1-Lipschitz, rendering the aforementioned results inapplicable. In our study, we establish convergence properties for the exponential parameterization, focusing on the structural characteristics of the objective function and the inherent variance of its gradient estimator. To that end, we derive gradient variance bounds for the exponential parameterization and present tighter bounds for the softplus parameterization. These bounds reveal that we obtain a reduction in the variance of the gradient estimator for small scale parameters. Additionally, we identify conditions under which the ELBO exhibits Lipschitz smoothness and strong convexity with the exponential parameterization. For the energy term of the ELBO, we demonstrate that the magnitude of its smoothness constant depends on the number of variational parameters as well as the size of the scale components. Notably, as we learn smaller scale components, the smoothness constant decreases.

Our objective is to emphasize the advantages of non-linear parameterizations when learning small scale components. Learning posterior approximations with small standard deviations is of particular importance for sufficiently large datasets. According to the Bernstein-von Mises theorem, provided that suitable assumptions are met, the posterior distribution contracts around the parameter of interest as the sample size increases. Finally, we empirically validate our theoretical findings and demonstrate that the speed of convergence for non-linear and linear parameterizations is contingent on both the number of variational parameters and the scale of the dataset. Notably, with sufficiently large datasets, non-linear parameterizations outperform their linear counterparts.

To summarize, our contributions are as follows:

- We prove and establish the conditions under which the energy function is Lipschitz smooth with an exponential scale parameterization.
- We derive gradient variance bounds for the exponential parameterization and establish tighter bounds for the softplus parameterization.
- We specify conditions under which the energy, using a exponential parameterization, is strongly convex.
- We demonstrate that the entropy is smooth with both the exponential and softplus parameterizations.
- On a synthetic dataset, we confirm that non-linear scale parameterization's convergence speed depends on the number of data points and variational parameters. However, with a sufficient amount of data, the impact from the variational parameters diminishes.
- We further solidify these empirical findings on seven real datasets which vary in size.

2 RELATED WORK

The structural properties of $\mathcal{L}(\phi)$, namely smoothness and (strong) convexity, have previously been studied by Challis and Barber (2013); Domke (2020). In particular, Domke (2020) proved that, given a target that is M-Lipschitz smooth, the energy component of the variational objective is also M-Lipschitz smooth. Additionally, when the target is strongly convex, then the same holds true for the energy.

Furthermore, the variance of VI gradient estimators has been investigated by Kim et al. (2023c); Fan et al. (2015), Xu et al. (2019), and Domke (2019). The latter derived upper bounds on the variance of the BBVI gradient estimator in the context of linear scale parameterizations. Meanwhile, Kim et al. (2023c) established corresponding bounds for 1-Lipschitz scale parameterizations.

Recently, the results from these works were leveraged by Kim et al. (2023b) and Domke et al. (2023) to establish convergence guarantees for BBVI with linear scale parameterizations. These guarantees were derived on the assumption that the log-density of the posterior is Lipschitz-smooth and (strongly) concave. In the study by Kim et al. (2023b), additional guarantees were provided for 1-Lipschitz parameterizations, which does not include the exponential scale transformation, under the assumption that the log posterior is Lipschitz smooth and that the likelihood is $\mu-$quadratically growing. Additionally, a significant contribution was recently made by Kim et al. (2023a), who demonstrated that BBVI, when employing the sticking-the-landing gradient estimator, converges at a geometric rate.

In this paper, we build on the theoretical insights provided by Domke (2019, 2020), and Kim et al. (2023c).

3 BACKGROUND

Notation For a matrix D, let the *Frobenius* norm be $\|D\|_F := \sqrt{\mathrm{Tr}(D^T D)}$. For the sake of notational ease, we use the notation $f(\boldsymbol{\omega})$ for scalar functions $f : \mathbb{R} \rightarrow \mathbb{R}$ to indicate that f is to be applied elementwise onto $\boldsymbol{\omega}$.

Variational Inference In VI, it is common to approximate a posterior density using a simpler, often Gaussian, variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with parameters ϕ, where $\mathbf{z} \in \mathbb{R}^d$ is a latent variable and $\mathbf{x} \in \mathbb{R}^k$ is the observed data. The goal is to infer $q_\phi(\mathbf{z}|\mathbf{x})$ by minimizing the KL divergence between the true posterior and the variational posterior. However, the true posterior is typically unknown, and thus the KL divergence cannot be computed directly. VI circumvents this by instead considering an equivalent problem, which is to maximize the ELBO

$$\phi^* = \arg\max_{\phi} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right].$$

where $\log p_\theta(\mathbf{x}, \mathbf{z})$ is the (unnormalized log-density) target with parameters θ.

In VI, when using gradient descent, we iteratively optimize $\mathcal{L}(\phi)$ through the following updates:

$$\phi_{t+1} \leftarrow \phi_t + \gamma \nabla \mathcal{L}(\phi_t),$$

where γ is the step size, the gradient is computed with respect to the variational parameters ϕ_t at the current time step t, and the plus sign indicates that the ELBO should be maximized.

Location-scale family In this work, we assume that $q_\phi(\mathbf{z}|\mathbf{x})$, with parameters $\phi = (\boldsymbol{\mu}, L)$, belongs to the location-scale family with location $\boldsymbol{\mu}$ and covariance matrix $\Sigma = LL^T$, here represented through a Cholesky decomposition where L is a triangular matrix. Thus,

$$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}) \Leftrightarrow \mathbf{z} \overset{d}{=} L\boldsymbol{\epsilon} + \mu, \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}),$$

where $p(\boldsymbol{\epsilon})$ is a standardized base distribution. The random vector $\boldsymbol{\epsilon}$ is defined as $\boldsymbol{\epsilon} \triangleq (\epsilon_1, \epsilon_2, \ldots, \epsilon_d) \in \mathbb{R}^d$. The components of $\boldsymbol{\epsilon}$ are assumed to be i.i.d. with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = 1$, $\mathbb{E}[\epsilon_i^3] = 0$, and fourth moment $\mathbb{E}[\epsilon_i^4] = \kappa$.

Examples of members from the location-scale family include the Gaussian and Student-t distributions (Casella and Berger, 2021).

Black-Box Variational Inference for the Location-Scale Family Depending on the choice of the variational family and the structure of $\log p_\theta(\mathbf{x}, \mathbf{z})$, it might not always be possible to derive $\nabla \mathcal{L}(\phi)$ in closed form. BBVI tackles this by producing stochastic estimates of the gradient or parts of it. These estimates are then used to optimize the ELBO with Stochastic Gradient Descent (SGD).

For the location-scale family, a segment of the gradient can be evaluated analytically. By way of illustration, $\mathcal{L}(\phi)$ can be decomposed as

$$\mathcal{L}(\phi) := \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z})]}_{l(\phi)} \quad \underbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log q_\phi(\mathbf{z}|\mathbf{x})]}_{h(\phi)},$$

where $h(\phi)$ is the entropy and $l(\phi)$ represents the energy component. For the location-scale family, $\nabla h(\phi)$ can be derived in closed form. However, whether this is possible for the energy term hinges on the specific form of $\log p_\theta(\mathbf{x}, \mathbf{z})$.

When the target necessitates it, the reparameterization trick can be employed to obtain unbiased estimates of $\nabla l(\phi)$. Specifically, by performing a change of variable

$$\begin{aligned} \nabla l(\phi) &= \nabla \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}, \mathbf{z})] \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla \log p_\theta(x, t_\phi(\boldsymbol{\epsilon}))], \end{aligned}$$

where $t_\phi(\boldsymbol{\epsilon}) \triangleq L\boldsymbol{\epsilon} + \boldsymbol{\mu}$. Since the expectation is now taken with respect to the base distribution $p(\boldsymbol{\epsilon})$, we can use Monte Carlo integration to obtain unbiased estimates $\mathbf{g} \triangleq \nabla \log p_\theta(\mathbf{x}, t_\phi(\boldsymbol{\epsilon}))$ of $\nabla l(\phi)$. Thus, $\nabla l(\phi) = \mathbb{E}[\mathbf{g}]$.

Scale parameterizations There are numerous possible parameterization approaches for learning a scale matrix L that results in a positive definite covariance. We constrain our analysis to when L is constructed as follows (Kingma and Welling, 2019a):

$$L \triangleq L_T + D(\psi(\boldsymbol{\omega})), \tag{1}$$

where $L_T \in \mathbb{R}^{d \times d}$ is a strictly lower triangular matrix, $\psi : \mathbb{R} \to \mathbb{R}_+$ is applied element-wise to $\boldsymbol{\omega} \in \mathbb{R}^d$, and $D(\psi(\boldsymbol{\omega})) \in \mathbb{R}^{d \times d}_{>0}$ is a diagonal matrix. A special case of this parameterization occurs when it is assumed that each *off-diagonal* element of L_T is equal to zero. This is equivalent to the well-known *mean field parameterization.*

For the exponential parameterization let $\boldsymbol{\omega} \triangleq \log \boldsymbol{\sigma}$, where $\boldsymbol{\sigma} \in \mathbb{R}^d_{>0}$, and $\psi(\boldsymbol{\omega}) \triangleq e^{\boldsymbol{\omega}}$. Another choice for the scale parameterization is the softplus function where $\boldsymbol{\omega} \triangleq \log(e^{\boldsymbol{\sigma}} - 1)$ and $\psi(\boldsymbol{\omega}) \triangleq \log(1 + e^{\boldsymbol{\omega}})$. Finally, for the linear approach $\boldsymbol{\omega} \triangleq \boldsymbol{\sigma}$ (equiv. $\psi(\boldsymbol{\omega}) = \boldsymbol{\omega}$).

3.1 Convergence Properties

Lipschitz smoothness Many convergence theorems for the SGD algorithm (and augmented versions of it) require a Lipschitz smooth objective function (Reddi et al., 2016; Ghadimi and Lan, 2013; Garrigos and Gower, 2023). A function h is Lipschitz M-smooth in the ℓ_2 norm if it is differentiable and if

$$\|\nabla h(\mathbf{z}) - \nabla h(\mathbf{z}')\|_2 \leq M\|\mathbf{z} - \mathbf{z}'\|_2 \quad \forall \mathbf{z}, \mathbf{z}'$$

where $0 < M < \infty$.

Strong convexity Another structural property that can be used in conjunction with Lipschitz smoothness to prove convergence is strong convexity. Let $\lambda \geq 0$. Then, a function h is λ-strongly convex, if for every $x, y \in \mathbb{R}^d$, and $t \in [0, 1]$

$$\lambda \frac{t(1-t)}{2}\|x-y\|^2 + h(tx + (1-t)y) \leq th(x) + (1-t)h(y).$$

Variance of the gradient estimator A downside of BBVI is that the convergence rate is impacted by the variance of the stochastic gradient estimator $\mathbf{g}$. Domke (2019) studied the expected squared norm (ESN), which upper bounds the trace of the variance of the gradient estimator

$$\mathrm{Tr}(\mathrm{Var}[\mathbf{g}]) = \mathbb{E}[\|\mathbf{g}\|_2^2] - \|\mathbb{E}[\mathbf{g}]\|_2^2 \leq \mathbb{E}[\|\mathbf{g}\|_2^2].$$

Domke (2019) demonstrated that a bound directly on $\mathrm{Tr}(\mathrm{Var}[\mathbf{g}])$ is not significantly better compared to an upper bound on the ESN when $d \gg 1$. Consequently, in this paper, we also derive bounds on the ESN.

4 THEORETICAL RESULTS

For ease of notation, let $\pi(\mathbf{z}) \triangleq \log p_\theta(\mathbf{x}, \mathbf{z})$. In the case of non-linear parameterized approaches, let $\phi^\psi \triangleq (\boldsymbol{\mu}, L_T, \boldsymbol{\omega})$, and for the linear approach, let $\phi \triangleq (\boldsymbol{\mu}, L_T, \boldsymbol{\sigma})$.

4.1 Smoothness results

We now show under which assumptions $\mathcal{L}(\phi^\psi)$ is Lipschitz smooth. It is well-known that the sum of two Lipschitz smooth functions remains Lipschitz smooth, with a smoothness constant equal to the sum of their individual constants. For a derivation, refer to Lemma C.1 in the *supplementary material*. Therefore, we individually investigate the smoothness of the energy and entropy terms.

The entropy The differential entropy of a random variable with a distribution from the location-scale family has a closed form expression (Cover, 1999):

$$h(\phi) \triangleq h(\boldsymbol{\epsilon}) + \sum_{i=1}^{d} \log \sigma_i,$$

where $h(\boldsymbol{\epsilon})$ is the entropy of the standardized base distribution, which is a constant. A derivation of this result can be found in the supplementary materials.

Evidently, the partial derivatives of $h(\phi)$ w.r.t. to $\boldsymbol{\mu}$ and L_T will be equal to zero. For the linearly parameterized approach, the partial derivatives w.r.t. to a diagonal element satisfy $\frac{\partial}{\partial \sigma_i} h(\phi) = \frac{1}{\sigma_i}$, which tends to infinity as $\sigma_i \to 0$. Thus, $h(\phi)$ is not Lipschitz smooth.

To resolve this, we instead employ a non-linear approach, such that $h(\phi^\psi)$ becomes Lipschitz smooth.

Lemma 4.1. *Let $h(\cdot)$ be the entropy of a random variable from the location-scale family.*

(1) Let $\psi(\boldsymbol{\omega}) \triangleq e^{\boldsymbol{\omega}}$, then $h(\phi^\psi)$ is Lipschitz smooth with an arbitrarily small smoothness constant.

(2) Let $\psi(\boldsymbol{\omega}) \triangleq \log(1 + e^{\boldsymbol{\omega}})$, then $h(\phi^\psi)$ is Lipschitz smooth.

Proof. See the supplementary material. □

Energy Term Kim et al. (2023c) proved and established conditions for when $l(\phi^\psi)$ is smooth, with one assumption being that $\pi(\mathbf{z})$ is twice differentiable, and another that $\psi(\boldsymbol{\omega})$ is 1-Lipschitz continuous. They found that using a 1-Lipschitz parameterization increases the value of the smoothness constant compared to the linear approach, suggesting that it may be necessary to reduce the step size to ensure convergence. Nonetheless, it still remains to be demonstrated whether the exponential approach is smooth.

We now proceed to prove and identify conditions for when $l(\phi^\psi)$, with the exponential approach, is Lipschitz smooth.

Theorem 4.2. *Let $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$. Assume the following: (1) $\pi(\boldsymbol{z})$ is M-Lipschitz smooth, (2) $\|\nabla \pi(\mathbf{z})\| \leq D < \infty$, and (3) $\sigma_j \leq K < \infty, \quad \forall j$. Then*

$$\|\nabla l(\phi^\psi) - \nabla l(\phi^{\psi'})\|_2 \\ \leq K(M\sqrt{m + d(K-1)} + D)\|\phi^\psi - \phi^{\psi'}\|_2,$$

where m denotes the number of variational parameters.

Proof. See the supplementary material. □

From Theorem 4.2, we can infer that the smoothness constant of the exponential method depends on M, which is the smoothness constant of $l(\phi)$ with the linear parameterization. Whether $l(\phi^\psi)$ obtains an increased smoothness constant compared to $l(\phi)$ depends on the norm of the gradient of the target, the number of variational parameters, and the scale components' size. More specifically, the smoothness constant of $l(\phi^{\psi})$ increases as the dimensionality of the problem grows with m and d. However, the theorem also highlights that the smoothness constant of $l(\phi^{\psi})$ decreases as the diagonal scale components σ_j become smaller.

In Section 5, we empirically investigate the interplay between the dimensionality and the scale on the convergence rate.

4.2 Gradient Variance Bounds

We now derive a gradient variance bound for the exponentially parametrized approach and a tighter bound compared to the one derived in Kim et al. (2023c) for the softplus parameterization. This will be accomplished by imposing an additional constraint, namely that each $\sigma_j \leq K < \infty$.

Lemma 4.3. *Assume that each $\sigma_j \leq K < \infty$, and that L has a mean field parameterization, where each $(L_T)_{i,j} = 0$ for $i \neq j$.*

(1) When $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$, then

$$\|\nabla_{\phi^{\psi}} \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2 \leq (1 + K^2 \|\mathcal{E}\|_F) \|\nabla \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2.$$

(2) When $\psi(\boldsymbol{\omega}) = softplus(\boldsymbol{\omega})$, then

$$\|\nabla_{\phi^{\psi}} \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2 \leq (1 + (1 - e^{-K})^2 \|\mathcal{E}\|_F) \|\nabla \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2.$$

(3) (Kim et al., 2023c) When $\psi(\boldsymbol{\omega}) = \boldsymbol{\omega}$, then

$$\|\nabla_{\phi} \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2 \leq (1 + \|\mathcal{E}\|_F) \|\nabla \pi(t_{\phi}(\boldsymbol{\epsilon}))\|_2^2.$$

where $\mathcal{E}$ is a diagonal matrix where $\mathcal{E}_{ii} = \epsilon_i^2$.

Proof. See the supplementary material. □

With Lemma 4.3 we can then show the following

Lemma 4.4. *Let $\mathbf{g}$ be the gradient estimator of $l(\cdot)$, and assume that (1) each $\sigma_j \leq K < \infty$, (2) that $\mathbf{z}$ is a stationary point of π, (3) that L has a mean field parameterization, and (4) π is M-Lipschitz smooth.*

(1) When $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$, then

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&(K^2 2\sqrt{d\kappa} + 1)\|\boldsymbol{\mu} - \boldsymbol{z}\|_2^2 \\ &+ (K^2(\sqrt{d\kappa} + \sqrt{d}\kappa) + 1)\|L\|_F^2 \Big). \end{aligned} \tag{2}$$

(2) When $\psi(\boldsymbol{\omega}) = softplus(\boldsymbol{\omega})$, then

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&((1 - e^{-K})^2 2\sqrt{d\kappa} + 1)\|\boldsymbol{\mu} - \boldsymbol{z}\|_2^2 \\ &+ ((1 - e^{-K})^2(\sqrt{d\kappa} + \sqrt{d}\kappa) + 1)\|L\|_F^2 \Big). \end{aligned} \tag{3}$$

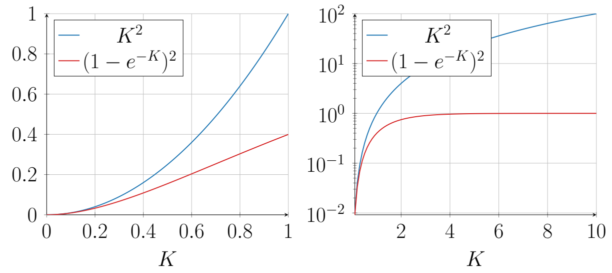

Figure 1: Comparison of the scaling factors' magnitudes in the ESN upper bounds for the exponential and softplus parameterizations for increasing values of K.

(3) (Kim et al., 2023c) When $\psi(\boldsymbol{\omega}) = \boldsymbol{\omega}$, then

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&(2\sqrt{d\kappa} + 1)\|\boldsymbol{\mu} - \boldsymbol{z}\|_2^2 \\ &+ (\sqrt{d\kappa} + \sqrt{d}\kappa + 1)\|L\|_F^2 \Big). \end{aligned} \tag{4}$$

Proof. See the supplementary material. □

Upon comparing the gradient variance bounds in Lemma 4.4, it becomes apparent that the bounds for the non-linear methods are subject to scaling factors that rely on the value of K. These scaling factors are illustrated in Figure 1. From the right-hand plot, we observe that for $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$, the upper bound on the ESN is scaled by a factor that tends to infinity as K approaches infinity. In contrast, the scaling factor for softplus is bounded from above by 1.

Nevertheless, it is more interesting, as supported by the size of the scales learned in our empirical evaluations (see Figure 4), to consider the scenario where K tends to zero. When $K < 1$, both methods yield scaling factors that reduce the magnitude of the ESN bound compared to the linear approach. The softplus parameterization achieves this reduction at a faster rate than the exponential method, as demonstrated on the left side of Figure 1. However, for sufficiently small scales, the difference in scaling factors between the two methods diminishes.

To understand how these scaling factors impact the final gradient variance bounds for the softplus parameterization, as presented in Lemma 4.4, consider Figure 2. Notice that, with our bound, the impact from increasing d is mitigated as K decreases. Conversely, as $K \to \infty$, we approach the bound derived by Kim et al. (2023c), where a significant penalty is incurred from increasing d.

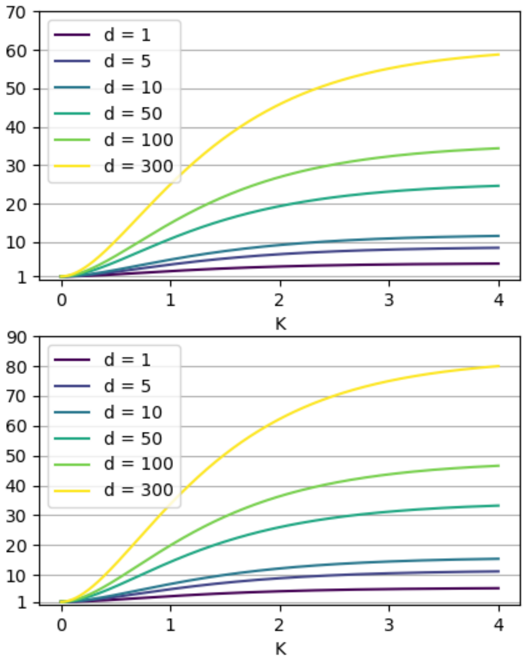

Figure 2: The scaling factors $((1-e^{-K})^2 \cdot 2\sqrt{d\kappa}+1)$ (top) and $(1-e^{-K})^2(\sqrt{d\kappa}+\sqrt{d\kappa})+1$ (bottom) found in the softplus bound on $\mathbb{E}\|\mathbf{g}\|_2^2$ (Lemma 4.4) for various numbers of variational parameters (d) and increasing scales (K).

4.3 Strong Convexity

We now prove the strong convexity of the energy under the exponential parameterization.

Theorem 4.5. *Let π be λ-strongly convex, $\psi(\omega) = e^{\omega}$, and assume that each $\sigma_j \geq \delta$, where $1 \geq \delta > 0$. Then $l(\phi^{\psi})$ is $\lambda\delta^2-$strongly convex for the mean-field parameterization.*

Proof. See the supplementary material. □

5 EXPERIMENTS

Training Infrastructure All experiments were conducted on a single NVIDIA RTX 4090 with 24 GiB of memory using the PyTorch framework (Paszke et al., 2019).

Common setup In the synthetic experiments, we used vanilla SGD, and with a smoothness constant of L, the step size was set to $\frac{1}{L}$. For the real datasets, we used the Adam optimizer and initialized the step size to $\frac{1}{L}$. We analytically estimated the smoothness constants M for each target $\pi(z)$, as described later in the experiments section. To estimate the constants K and D for the non-linear scale parameterizations (refer to Thm. 4.2), we initially generated a single optimization trace. We used the same learning rate for both the exponential and the softplus parameterizations to ensure a fair comparison and empirically found that both parameterizations exhibited similar performance. In all experiments, we use a Gaussian variational posterior with a mean-field parameterization.

Furthermore, to address the non-smoothness of the entropy term when using the linear parameterization, as described in Section 4.1, Domke (2020) used projected SGD to ensure that the diagonal scale component remained above a lower bound of $\frac{1}{\sqrt{M}}$, where M is the Lipschitz smoothness constant. In all subsequent experiments, we utilize projected SGD for the linear parameterization.

5.1 Synthetic - Increasing data

In this experiment, we examine the influence of the dataset size on the convergence speed for both linear and non-linear parameterizations. We also analyze the ESN and scale trace plots of these approaches.

Setup Consider a dataset $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ wherein each $\boldsymbol{x}_i \in \mathbb{R}^d$ is an i.i.d. sample drawn from a Gaussian distribution with a known covariance matrix $\boldsymbol{\sigma}^2 I$ and unknown mean $\boldsymbol{\mu}$. Given that $X|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 I)$ and assuming a standard normal prior for $\boldsymbol{\mu}$ as $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, I)$, the resulting posterior distribution for $\boldsymbol{\mu}$ has a closed form solution. In the univariate case, the variance of the posterior σ^* is given by $\frac{\sigma^2}{\sigma^2+N}$. Notably, the variance decreases with N. With this setup, we generated 12 synthetic datasets, where we gradually increased the number of features as $d = 20, 100, 500$ and the number of data points $N = 10, 20, 100, 1000$.

Results The trace plots over $\mathcal{L}(\phi)$ for the synthetic datasets are presented in Figure 3. For a fixed $N = 10$, we observe that with increasing d (and thus an increasing number of variational parameters), the non-linear parameterizations converge slower than their linear counterparts. However, for each value of d, the non-linear methods exhibit a faster convergence rate for increasing values of N, while the linear approach converges after roughly the same number of iterations. Furthermore, for each value of d, the non-linear methods surpass the linear ones provided that N is sufficiently large.

5.2 Real datasets

In this experiment, we compare the convergence speed of linear and non-linear approaches on both linear and logistic regression tasks across six datasets. These

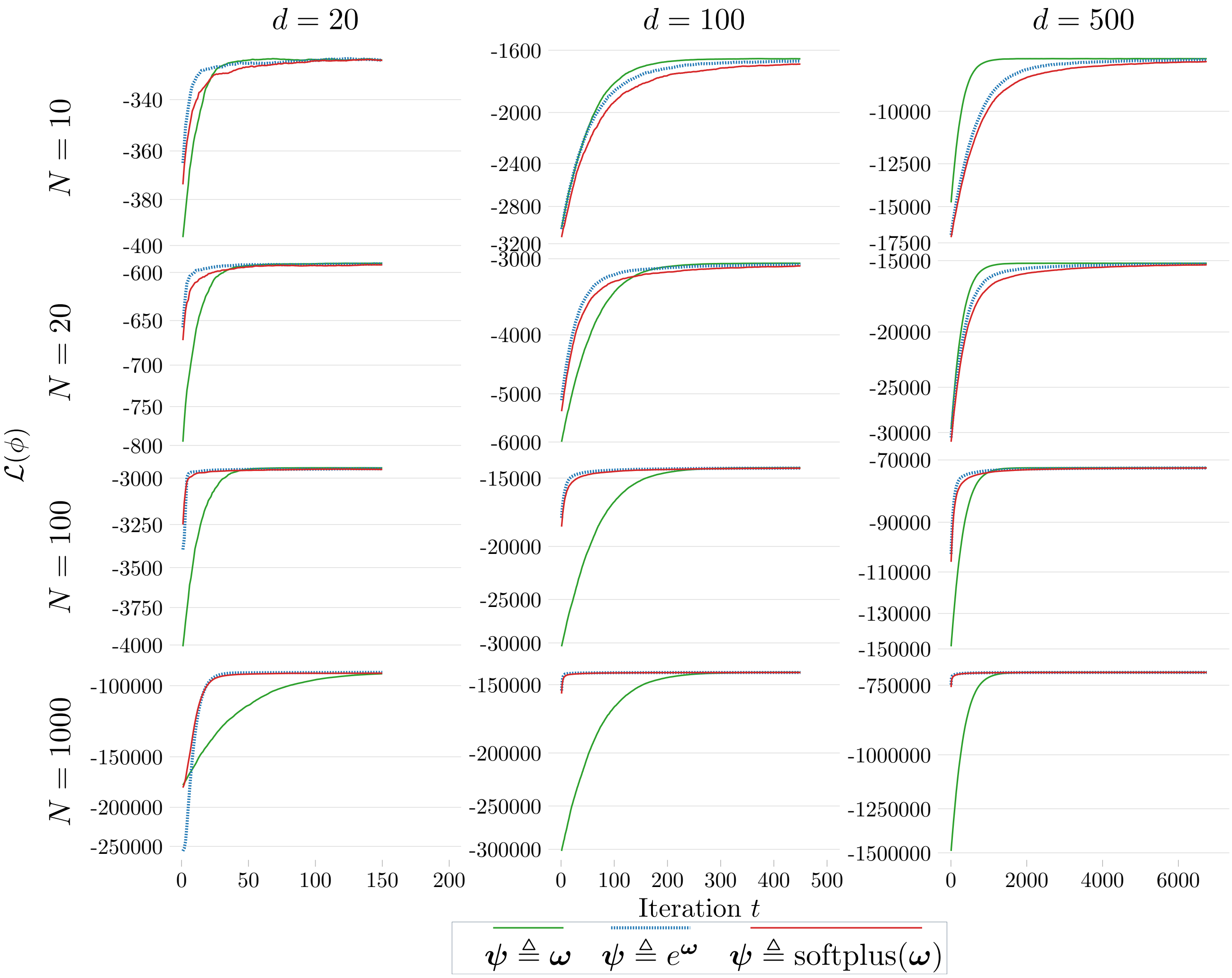

Figure 3: Comparison of convergence speed on a synthetic dataset with increasing number of features d and number of data points N.

datasets vary in both their number of examples and features, as detailed in Table 1. We opt to exclude non-numerical features and apply a *z-score* standardization on the remaining numerical features[1].

Setup Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ we posit $p(\mathbf{z}, \mathbf{y}|X) = \mathcal{N}(\mathbf{z}|0, \sigma^2 I) \prod_{n=1}^N p(y_n|\mathbf{z}, \mathbf{x}_n)$ and define $p(y_n|\mathbf{z}, \mathbf{x}_n) = \mathcal{N}(y_n|\mathbf{z}^T\mathbf{x}_n, \rho^2 I)$ for linear regression and $p(y_n|\mathbf{z}, \mathbf{x}_n) = \sigma(y_n \cdot \mathbf{z}^T\mathbf{x}_n)$ for logistic regression. We adopt the same values of $\sigma = 1$ and $\rho = 2$ as in Domke (2019). We calculated the smoothness constant M of each target $\pi(z)$, by taking the spectral norm of $\frac{1}{\sigma^2}I + cX^TX$, where $c = 0.25$ for logistic regression and $c = \frac{1}{\rho^2}$ for linear regression.

For the optimization of the fires and cpusmall datasets, the variational parameters were initialized with $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma^2 I = I$. For the other datasets, $\sigma^2 I$ was set to $0.1^2 I$.

Dataset	N	d
fires (Cortez and Morais, 2008)	517	10
cpusmall (Cheng et al., 2009)	8192	12
msd (Bertin-Mahieux, 2011)	$1,000,000$	90
ionosphere (Sigillito et al., 1989)	351	34
higgs (Whiteson, 2014)	$11,000,000$	28
buzz (Kawala et al., 2013)	$140,000$	78
cover type (Blackard, 1998)	$581,012$	10

Table 1: Overview of the considered datasets. The number of features are reported after filtering out non-numerical features.

[1] Also on the target variable in the case of linear regression

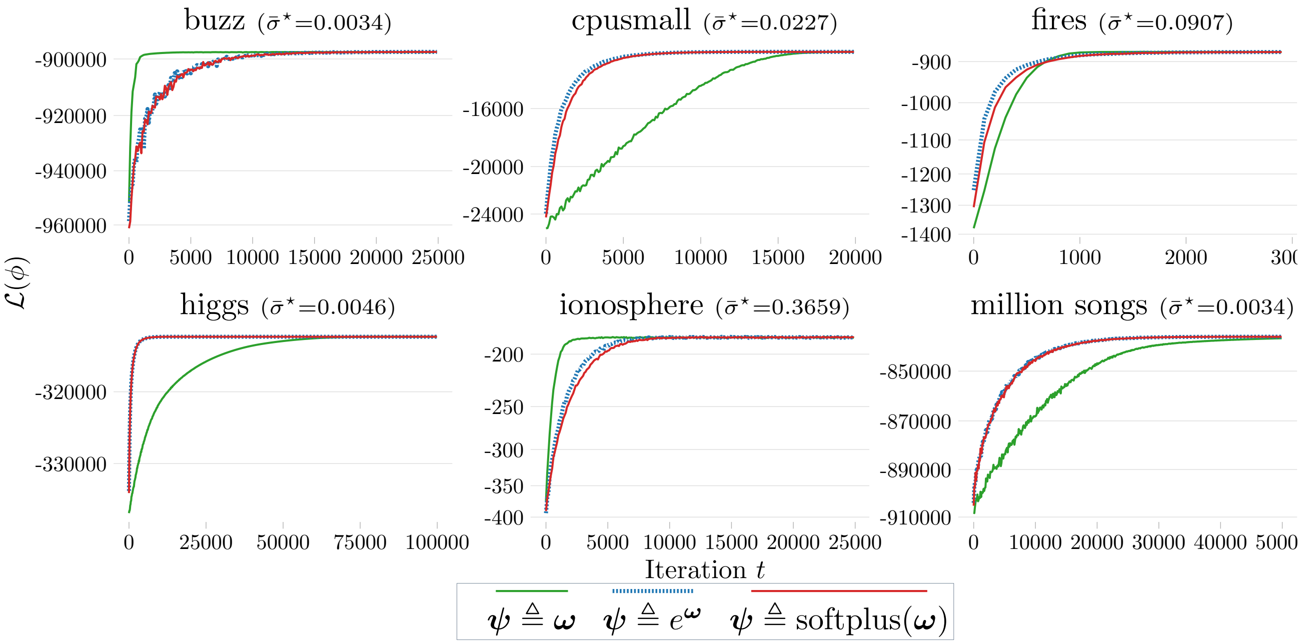

Figure 4: Comparison of convergence speed on datasets of varying sizes between linear and non-linear scale parameterizations. The values of $\mathcal{L}(\phi_t)$ are reported as averages over 10 independent training runs. For each dataset, we present the final averaged scale parameter $\bar{\sigma}^*$.

Results Figure 4 showcases the ELBO trace plots. The results highlight the influence of both the number of data points and features on the convergence of the non-linear approaches. The linear approach has an advantage on the ionosphere dataset due to its small number of data points and a relatively large number of features. On the fires dataset, which has fewer features and slightly more data points compared to ionosphere, the discrepancy in the convergence speed is reduced. Although the cpusmall dataset shares a similar feature count with the fires dataset, it comprises about 16 times more data points, leading to swifter convergence for the non-linear methods. On the other hand, despite the buzz dataset having $140,000$ data points, its higher dimensionality slows down the convergence of the nonlinear method. The msd dataset, which has marginally more features compared to buzz but seven times as many data points, exhibits faster convergence for the non-linear methods. Finally, both cover type and higgs have a large number of examples and relatively few features, which benefits the non-linear methods. Results for the cover type dataset can be found in the *supplementary material.*

It is important to note that Kim et al. (2023b) also studied large-scale problems and found that the linear parametrization often yielded superior results, which may initially seem contradictory to our findings. However, our bounds provide a coherent explanation for this seeming contradiction. Although Kim et al. (2023b) focused on problems involving small scales, their models incorporated a significantly larger number of variational parameters by employing the Cholesky parametrization, which includes non-zero off-diagonal elements. In contrast, we utilize the mean field parametrization, as this is necessitated by our gradient variance bounds. Moreover, following Domke (2020), we set our step sizes to the inverses of the estimated smoothness constants. Given that the non-linear smoothness constant decreases with K and increases with d, our learning rates varied across datasets. In contrast, Kim et al. (2023b) utilized a uniform step size of 10^{-3} across all models. We further expand on this topic in the supplementary material.

Finally, we observe that the trace plots for the exponential and softplus parameterizations become increasingly similar at smaller scales. This is anticipated, given that the distinctions between the parameterizations diminish as the scale reduces.

6 LIMITATIONS

A limitation of our work is that the smoothness results in Theorem 4.2 were only demonstrated for the exponential parameterization. Empirically, the softplus parameterization behaved similarly to the exponential parameterization, suggesting that a similar result could exist for the softplus parameterization.

7 CONCLUSION

In this study, we delved into the convergence properties of location-scale variational families featuring non-linear covariance parameterizations. In particular, we derived novel Lipschitz smoothness and strong convexity results, as well as gradient variance bounds for the widely-recognized exponential covariance parameterization. Additionally, we derived tighter gradient variance bounds for the softplus parameterization. Collectively, our theoretical and empirical results underscore the advantages of non-linear parameterizations for large-scale datasets.

Acknowledgements

First, we acknowledge the insightful comments provided by the reviewers, which have helped improve our work. This project was made possible through funding by Klarna Bank AB, the Swedish Foundation for Strategic Research under grant ID19-0052, and the Swedish Research Council (VR) 2017-04609.

References

T. Bertin-Mahieux. YearPredictionMSD. UCI Machine Learning Repository, 2011. DOI: https://doi.org/10.24432/C50K61.

J. Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877, 2017.

G. Casella and R. L. Berger. *Statistical inference.* Cengage Learning, 2021.

E. Challis and D. Barber. Gaussian kullback-leibler approximate inference. *Journal of Machine Learning Research*, 14(8), 2013.

W. Cheng, J. Hühn, and E. Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 161–168, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553395. URL `https://doi.org/10.1145/1553374.1553395`.

P. Cortez and A. Morais. Forest Fires. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C5D88D.

T. M. Cover. *Elements of information theory.* John Wiley & Sons, 1999.

J. Domke. Provable gradient variance guarantees for black-box variational inference, 2019.

J. Domke. Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning*, pages 2587–2596. PMLR, 2020.

J. Domke, G. Garrigos, and R. Gower. Provable convergence guarantees for black-box variational inference, 2023.

K. Fan, Z. Wang, J. Beck, J. Kwok, and K. A. Heller. Fast second order stochastic backpropagation for variational inference. *Advances in Neural Information Processing Systems*, 28, 2015.

G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods, 2023.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4): 2341–2368, 2013. doi: 10.1137/120880811. URL `https://doi.org/10.1137/120880811`.

F. Kawala, A. Douzal, E. Gaussier, and E. Diemert. Buzz in social media. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C56G6V.

K. Kim, Y. Ma, and J. R. Gardner. Linear convergence of black-box variational inference: Should we stick the landing?, 2023a.

K. Kim, J. Oh, K. Wu, Y.-A. Ma, and J. R. Gardner. On the convergence and scale parameterizations of black-box variational inference, 2023b.

K. Kim, K. Wu, J. Oh, and J. R. Gardner. Practical and matching gradient variance bounds for black-box variational Bayesian inference. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16853–16876. PMLR, 23–29 Jul 2023c. URL `https://proceedings.mlr.press/v202/kim23w.html`.

D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, nov 2019a. ISSN 1935-8237. doi: 10.1561/2200000056. URL `https://doi.org/10.1561/2200000056`.

D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019b.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference, 2016.

O. Kviman, R. Molén, A. Hotti, S. Kurt, V. Elvira, and J. Lagergren. Cooperation in the latent space: The benefits of adding mixture components in variational autoencoders. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18008–18022. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/kviman23a.html`.

O. Kviman, N. Branchini, V. Elvira, and J. Lagergren. Variational resampling. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2024.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL `http://arxiv.org/abs/1912.01703`.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 324–333, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/ranganath16.html`.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

T. Salimans and D. A. Knowles. On using control variates with stochastic approximation for variational bayes and its connection to stochastic linear regression, 2014.

V. Sigillito, S. Wing, L. Hutton, and K. Baker. Ionosphere. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5W01B.

D. Tran, R. Ranganath, and D. M. Blei. The variational gaussian process, 2016.

D. Whiteson. HIGGS. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5V312.

M. Xu, M. Quiroz, R. Kohn, and S. A. Sisson. Variance reduction properties of the reparameterization trick. In *The 22nd international conference on artificial intelligence and statistics*, pages 2711–2720. PMLR, 2019.

C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes.** See the Supplementary Materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.** The code can be found in the Supplementary Materials. Please refer to Section 5 for detailed information about the experiments, and to Table 1 for references to the specific datasets used.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.** Please refer to Section 5.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.** Please refer to Section 5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.** See the first paragraph in Section 5.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. **Yes.** Citations to the datasets used may be found in Table 1.

   (b) The license information of the assets, if applicable. **Not applicable**

   (c) New assets either in the supplemental material or as a URL, if applicable. **Not applicable**

   (d) Information about consent from data providers/curators. **Only public datasets are used**

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. **Not applicable**

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not applicable**

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not applicable**

Benefits of Non-Linear Scale Parameterizations in Black Box Variational Inference through Smoothness Results and Gradient Variance Bounds

A EXPERIMENTS

A.1 Additional Results

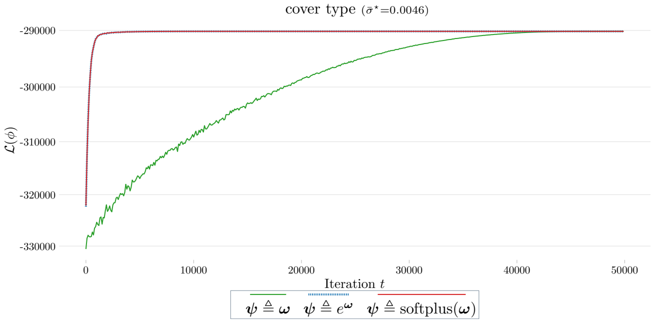

Figure 5: Comparison of convergence speed on the cover type dataset between linear and non-linear scale parameterizations. The values of $\mathcal{L}(\phi_t)$ are reported as averages over 10 independent training runs. The title indicates the final averaged scale parameter $\bar{\sigma}^*$ of the variational posterior.

A.2 Longer comparison Kim et al. (2023b)

It is important to note that Kim et al. (2023b) also studied large-scale problems and found that the linear parametrization often yielded superior results, which may initially seem contradictory to our findings. However, our bounds provide a coherent explanation for this seeming contradiction. Although Kim et al. (2023b) focused on problems involving small scales, their models incorporated a significantly larger number of variational parameters by employing the Cholesky parametrization, which includes non-zero off-diagonal elements. In contrast, we utilize the mean field parametrization, as this is necessitated by our gradient variance bounds. Moreover, following Domke (2020), we set our step sizes to the inverses of the estimated smoothness constants. Given that the non-linear smoothness constant decreases with K and increases with d, our learning rates varied across datasets. In contrast, Kim et al. (2023b) utilized a uniform step size of 10^{-3} across all models. Figure 6 illustrates the substantial advantage of the linear method, as predicted by our bounds, when increasing d, from 24 to 90, under the Cholesky parametrization. Furthermore, this effect is even more pronounced in the case of the msd dataset, as used by both us and Kim et al. (2023b), where our model employs 188 parameters compared to their 4, 559 variational parameters.

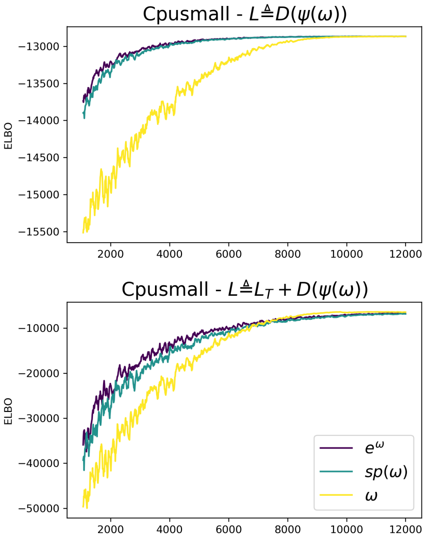

Figure 6: With the mean field parametrization (top), non-linear approaches benefit from a small d (24) and small scales, leading to faster convergence. With the full Cholesky parametrization (bottom), the parameters substantially increase (90), enhancing the relative performance of the linear method.

B EXTERNAL RESULTS

Theorem B.1 ((Cover, 1999, pp. 253-254)). *Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector. Then, for any matrix $A \in \mathbb{R}^{d\times d}$ and any vector $\boldsymbol{c} \in \mathbb{R}^d$, the differential entropy of $\boldsymbol{x}$ satisfies the following properties:*

(1) $h(\boldsymbol{x} + \boldsymbol{c}) = h(\boldsymbol{x})$,

(2) $h(A\boldsymbol{x}) = h(\boldsymbol{x}) + \log|\det(A)|$.

Lemma B.2 ((Domke, 2020)). *Let $p(\boldsymbol{\epsilon})$ be a standardized base distribution. Let $a : \mathbb{R}^d \to \mathbb{R}^k$ and $b : \mathbb{R}^d \to \mathbb{R}^k$ be squared-integrable functions. Then $\langle a, b\rangle_{p(\boldsymbol{\epsilon})} = \mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[a(\boldsymbol{\epsilon})^T b(\boldsymbol{\epsilon})]$ is a valid inner product.*

For a definition of the standardized base distribution $p(\boldsymbol{\epsilon})$, please refer to Section 3.

Lemma B.3 ((Domke, 2020)). *Let $p(\boldsymbol{\epsilon})$ be a standardized distribution and $\phi := (\boldsymbol{\mu}, L_T, \boldsymbol{\sigma})$, then*

$$\mathbb{E}_{p(\boldsymbol{\epsilon})}\|t_\phi(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2^2 = \|\phi - \phi'\|_2^2. \tag{5}$$

Lemma B.4 ((Kim et al., 2023c)). *Suppose that $t_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a location-scale family reparameterization function with $\pi : \mathbb{R}^d \to \mathbb{R}$. Then, for $\mathbf{g}_\pi \overset{d}{=} \nabla\pi(t_\phi(\boldsymbol{\epsilon}))$, and L parameterized as in Eq. 1 with the mean-field parameterization where each $L_T = \mathbf{0}$*

$$\|\nabla_\phi\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 = \|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 + \mathbf{g}_\pi^T\mathcal{E}\Psi\mathbf{g}_\pi,$$

where $\mathcal{E} \triangleq \operatorname{diag}(\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_d^2)$, and $\Psi \triangleq \operatorname{diag}(\psi'(\omega_1)^2, \psi'(\omega_2)^2, \ldots, \psi'(\omega_d)^2)$.

Lemma B.5 ((Kim et al., 2023c)). *Suppose that $t_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a location-scale family reparameterization function with $\pi : \mathbb{R}^d \to \mathbb{R}$, let ψ be a 1-Lipschitz function, and $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ where $p(\boldsymbol{\epsilon})$ is a standardized base distribution. The random vector $\boldsymbol{\epsilon}$ is defined as $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2, \ldots, \epsilon_d) \in \mathbb{R}^d$. The components of $\boldsymbol{\epsilon}$ are assumed to be i.i.d. with $\mathbb{E}[\epsilon_i] = 0$, $\operatorname{Var}[\epsilon_i] = 1$, $\mathbb{E}[\epsilon_i^3] = 0$, and $\mathbb{E}[\epsilon_i^4] = \kappa$. Then, for the mean-field parameterization*

$$\mathbb{E}\|t_\phi(\boldsymbol{\epsilon}) - \mathbf{z}\|_2^2(1 + \|\mathcal{E}\|_F) \leq (2\sqrt{\kappa d} + 1)\|\boldsymbol{\mu} - \mathbf{z}\|_2^2 + (\sqrt{d\kappa} + \kappa\sqrt{d} + 1)\|L\|_F^2. \tag{6}$$

Note that the inequality in Lemma B.5 is slightly different from Lemma 3 in Kim et al. (2023c), since there is a minor mistake in the final expression presented in Lemma 3.

Lemma B.6 ((Domke, 2019)). *Suppose that $t_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a location-scale family reparameterization function, that ψ is linear, and assume that $p(\boldsymbol{\epsilon})$ is a standardized base distribution. The random vector $\boldsymbol{\epsilon}$ is defined as $\boldsymbol{\epsilon} := (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_d) \in \mathbb{R}^d$. The components of $\boldsymbol{\epsilon}$ are assumed to be i.i.d. with $\mathbb{E}[\epsilon_i] = 0$, $\operatorname{Var}[\epsilon_i] = 1$, $\mathbb{E}[\epsilon_i^3] = 0$, and $\mathbb{E}[\epsilon_i^4] = \kappa$. Then,*

$$\mathbb{E}\|t_\phi(\boldsymbol{\epsilon}) - \mathbf{z}\|_2^2 = \|\boldsymbol{\mu} - \mathbf{z}\|_2^2 + \|L\|_F^2. \tag{7}$$

Theorem B.7 ((Domke, 2019)). *Suppose that π is M-smooth, $\boldsymbol{z}$ is a stationary point of π, and $p(\boldsymbol{\epsilon})$ is standardized with $\mathbb{E}[\epsilon_i] = 0$, $\operatorname{Var}[\epsilon_i] = 1$, $\mathbb{E}[\boldsymbol{\epsilon}_i^3] = 0$, and $\mathbb{E}[\epsilon_i^4] = \kappa$. Let $\mathbf{g} = \nabla_\phi\pi(t_\phi(\boldsymbol{\epsilon}))$ for $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ and assume that ψ is linear. Then,*

$$\mathbb{E}\|\mathbf{g}\|_2^2 \leq M^2((d+1)\|\boldsymbol{\mu} - \boldsymbol{z}\|_2^2 + (d+\kappa)\|L\|_F^2) \tag{8}$$

where $L \in \mathbb{R}^{d\times d}$.

Note that in Theorem B.7 L is not constrained to be lower triangular.

Lemma B.8 ((Kim et al., 2023b)). *Let π be convex. For a convex nonlinear ψ, the inequality*

$$\langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle \leq \mathbb{E}\langle \nabla \pi(t_{\phi}(\boldsymbol{\epsilon})), t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon}) \rangle$$

holds iff

$$\operatorname{diag}(\operatorname{Cov}(\nabla \pi(t_{\phi}(\boldsymbol{\epsilon})), \boldsymbol{\epsilon})) \succeq 0. \tag{9}$$

For the scale components parametrized as in Eq. 1, the assumption in Eq. 9 holds for the mean-field parameterization, yet may not hold otherwise.

Theorem B.9 ((Kim et al., 2023b; Domke, 2020)). *Let π be λ-strongly convex, and ψ be a scale parameterization.*

(1) If ψ is linear, the energy $l(\psi)$ is λ-strongly convex.

(2) If ψ is convex with the mean-field parameterization, the energy $l(\phi^{\psi})$ is convex.

(3) If ψ is convex with the scale parameterized as in Eq. 1, the energy $l(\phi^{\psi})$ may not be convex.

(4) If ψ is such that $\psi \in C^1(\mathbb{R}, \mathbb{R}_+)$, the energy $l(\phi^{\psi})$ is not strongly convex.

C PROOFS

Lemma C.1. *Given that $h(\phi)$ is M_h-Lipschitz smooth and $l(\phi)$ is M_l Lipschitz smooth, $\mathcal{L}(\phi)$ is $M_l + M_h$ Lipschitz smooth.*

Proof.

$$\begin{aligned}\|\nabla \mathcal{L}(\phi) - \nabla \mathcal{L}(\phi')\|_2 &= \|\nabla l(\phi) + \nabla h(\phi) - \nabla l(\phi') - \nabla h(\phi')\|_2 = \|(\nabla l(\phi) - \nabla l(\phi')) + (\nabla h(\phi) - \nabla h(\phi'))\|_2 \\ &\leq \|\nabla l(\phi) - \nabla l(\phi')\|_2 + \|\nabla h(\phi) - \nabla h(\phi')\|_2 \leq M_l\|\phi - \phi'\|_2 + M_h\|\phi - \phi'\|_2 \\ &= (M_l + M_h)\|\phi - \phi'\|_2,\end{aligned}$$

where we first apply the triangle inequality and then use the Lipschitz smoothness assumption of $h(\phi)$ and $l(\phi)$. □

Lemma 4.1. *Let $h(\cdot)$ be the entropy of a random variable from the location-scale family.*

(1) Let $\psi(\boldsymbol{\omega}) \triangleq e^{\boldsymbol{\omega}}$, then $h(\phi^{\psi})$ is Lipschitz smooth with an arbitrarily small smoothness constant.

(2) Let $\psi(\boldsymbol{\omega}) \triangleq \log(1 + e^{\boldsymbol{\omega}})$, then $h(\phi^{\psi})$ is Lipschitz smooth.

Proof. With Theorem B.1, it is straightforward to show

$$h(\phi^{\psi}) = h(\boldsymbol{\epsilon} L + \boldsymbol{\mu}) = h(\boldsymbol{\epsilon}) + \log|\det L| = h(\boldsymbol{\epsilon}) + \sum_{i=1}^{d} \log \psi(\omega_i), \tag{10}$$

and thus

$$\frac{\partial h(\phi^{\psi})}{\partial \omega_i} = \frac{1}{\psi(\omega_i)} \cdot \psi'(\omega_i). \tag{11}$$

(1) For the exponential parameterization $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$. Thus,

$$\|\nabla h(\phi^{\psi}) - \nabla h(\phi^{\psi'})\|_2 = \sqrt{\sum_{i=1}^{d} \frac{e^{\omega_i}}{e^{\omega_i}} - \frac{e^{\omega_i'}}{e^{\omega_i'}}} = 0 \leq M\|\phi^{\psi} - \phi^{\psi'}\|_2 \tag{12}$$

where $0 < M < \infty$.

(2) For the softplus parameterization we consider an equivalent definition of Lipschitz smoothness. For twice differentiable functions, $h(\phi^{\psi})$ is Lipschitz smooth if the eigenvalues of Hessian $\nabla^2 h(\phi^{\psi})$ are smaller than M_h. Any second-order partial derivatives taken w.r.t. a component of either $\boldsymbol{\mu}$ or L_T will be equal to zero. Additionally, all off-diagonal elements of $\nabla^2 h(\phi^{\psi})$ will be zero. The non-zero elements of $\nabla^2 h(\phi^{\psi})$ are the second-order partial derivatives taken twice w.r.t. the diagonal scale components σ_i. As a result the Hessian is diagonal and therefore its largest eigenvalue is the maximal value among its diagonal entries. Consequently, we consider:

$$\frac{\partial^2 h(\phi^{\psi})}{\partial \omega_i^2} = -\frac{1}{\psi(\omega_i)} \cdot \psi''(\omega_i) + \frac{1}{\psi^2(\omega_i)} \cdot {\psi'}^2(\omega_i). \tag{13}$$

For softplus $\psi(\omega) = \log(1+e^{\omega})$, $\psi'(\omega) = \frac{e^{\omega}}{1+e^{\omega}}$, and $\psi''(\omega) = \frac{e^{\omega}}{1+e^{\omega}} - \frac{(e^{\omega})^2}{(1+e^{\omega})^2}$. Thus

$$\frac{\partial^2 h(\phi^{\psi})}{\partial \omega_i^2} = -\frac{1}{\log(1+e^{\omega})} \cdot \left(\frac{e^{\omega}}{1+e^{\omega}} - \frac{(e^{\omega})^2}{(1+e^{\omega})^2}\right) + \frac{1}{(\log(1+e^{\omega}))^2} \cdot \frac{(e^{\omega})^2}{(1+e^{\omega})^2}, \tag{14}$$

and

$$|\frac{\partial^2 h(\phi^{\psi})}{\partial \omega_i^2}| = | -\frac{1}{\log(1+e^{\omega})} \cdot \left(\frac{e^{\omega}}{1+e^{\omega}} - \frac{(e^{\omega})^2}{(1+e^{\omega})^2}\right) + \frac{1}{(\log(1+e^{\omega}))^2} \cdot \frac{(e^{\omega})^2}{(1+e^{\omega})^2}| \leq 0.1671. \tag{15}$$

Therefore, $h(\phi^{\psi})$ is Lipschitz smooth with a smoothness constant M_h of 0.1671.

□

Lemma C.2. *Assume that* $0 < a, b \leq K$. *Then* $\exists$ *a* $0 < C < \infty$, *such that*

(1) $(b-a)^2 \leq C(\log(b) - \log(a))^2$.

Proof. (1) Since $\log(x)$ is continuous on the interval $[a, b]$ and differentiable on the interval (a, b), then according to the mean value theorem, there exists a point $c \in (a, b)$, such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{16}$$

Let $f(x) := \log(x)$, then by the mean value theorem

$$\frac{1}{c} = \frac{\log(b) - \log(a)}{b - a}. \tag{17}$$

We rearrange the terms and square both sides

$$(b-a)^2 = c^2(\log(b) - \log(a))^2. \tag{18}$$

Finally, the assumption that $0 < a, b \leq K$ implies that $c < K$, and thus that

$$(b-a)^2 \leq K^2(\log(b) - \log(a))^2 \tag{19}$$

□

Theorem 4.2. *Let $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$. Assume the following: (1) $\pi(\boldsymbol{z})$ is M-Lipschitz smooth, (2) $\|\nabla\pi(\mathbf{z})\| \leq D < \infty$, and (3) $\sigma_j \leq K < \infty, \quad \forall j$. Then*

$$\begin{aligned}&\|\nabla l(\phi^{\psi}) - \nabla l(\phi^{\psi'})\|_2 \\ &\leq K(M\sqrt{m + d(K-1)} + D)\|\phi^{\psi} - \phi^{\psi'}\|_2,\end{aligned}$$

where m denotes the number of variational parameters.

Proof. We want to show that $l(\phi^{\psi})$ is Lipschitz smooth. To that effect, we first compute the partial derivatives of the gradient. We first use the reparameterization trick and Lemma B.2

$$\begin{aligned}\nabla_{\phi_i^{\psi}} l(\phi^{\psi}) &= \nabla_{\phi_i^{\psi}} \mathbb{E}_{q_{\phi^{\psi}}(z|x)}\left[\pi(\boldsymbol{z})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\nabla_{\phi_i^{\psi}} \pi(t_{\phi}(\boldsymbol{\epsilon}))\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\nabla_{\phi_i^{\psi}} t_{\phi}(\boldsymbol{\epsilon})^T \nabla\pi(t_{\phi}(\boldsymbol{\epsilon}))\right] && (20)\\ &= \left(\nabla_{\phi_i^{\psi}} t_{\phi}, \nabla\pi \circ t_{\phi}\right)_{p(\boldsymbol{\epsilon})}, && (21)\end{aligned}$$

where $\circ$ is used to signify the composition of two functions, and the final equality follows from Lemma B.2.

It is straightforward to obtain the partial derivatives wrt to the location and off-diagonal scale components. This is a partial result in Lemma 4 found in Domke (2020)

$$\frac{\partial t_{\phi}(\boldsymbol{\epsilon})}{\partial \mu_i} = \boldsymbol{e}_i, \quad \frac{\partial t_{\phi}(\boldsymbol{\epsilon})}{\partial (L_T)_{i,j}} = \boldsymbol{e}_i \epsilon_j, \tag{22}$$

where $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ form a canonical basis in $\mathbb{R}^d$.

With the linear approach, used by Domke (2020), all derivatives of t_{ϕ} takes this form. Thus, the main proof idea in Domke (2020) is to show that the set $\{\frac{\partial t_{\phi}(\boldsymbol{\epsilon})}{\partial \phi_i}\}$ constitutes an orthonormal basis, which allows them to use Bessel's inequality.

Unfortunately, when ψ is non-linear, the components of the gradient where the partial derivatives are computed wrt to the diagonal scale components depend on the variational parameters ϕ^{ψ}; $\frac{\partial t_{\phi}(\boldsymbol{\epsilon})}{\partial \omega_i} = \boldsymbol{e}_i \epsilon_j \frac{\partial \psi(\omega_i)}{\partial \omega_i}$. This means that $\{\frac{\partial t_{\phi}(\boldsymbol{\epsilon})}{\partial \phi_i^{\psi}}\}$ is not an orthonormal basis and we cannot use the proof strategy employed in Domke (2020).

To unify the expression for the derivatives taken wrt to each type of variational parameter, we introduce a vector b such that

$$b := \left(1, 1, \ldots, \tfrac{\partial\psi(\omega_1)}{\partial\omega_1}, \tfrac{\partial\psi(\omega_2)}{\partial\omega_2}, \ldots, \tfrac{\partial\psi(\omega_d)}{\partial\omega_d}\right) \tag{23}$$

where the dimensionality of b corresponds to the number of variational parameters m. Then we express $\nabla_{\phi_i^{\psi}} t_{\phi}(\boldsymbol{\epsilon})$ in terms of $\nabla_{\phi_i} t_{\phi}(\boldsymbol{\epsilon})$

$$\nabla_{\phi_i^{\psi}} t_{\phi}(\boldsymbol{\epsilon}) = b_i \nabla_{\phi_i} t_{\phi}(\boldsymbol{\epsilon}) \tag{24}$$

With this expression for the partial derivatives we now form

$$\begin{aligned}\|\nabla l(\phi^{\psi}) - \nabla l(\phi^{\psi'})\|_2^2 &= \sum_{j=1}^{m}(\nabla_{\phi_j^{\psi}} l(\phi^{\psi}) - \nabla_{\phi^{\psi}{}'_j} l(\phi^{\psi'}))^2 = \sum_{j=1}^{m}((b_j \nabla_{\phi_j} t_{\phi}, \nabla\pi \circ t_{\phi})_{p(\boldsymbol{\epsilon})} - (b'_j \nabla_{\phi'_j} t_{\phi'}, \nabla\pi \circ t_{\phi'})_{p(\boldsymbol{\epsilon})})^2 \\ = \sum_{j=1}^{m}((\nabla_{\phi_j} t_{\phi}, b_j \nabla\pi \circ t_{\phi})_{p(\boldsymbol{\epsilon})} &- (\nabla_{\phi'_j} t_{\phi'}, b'_j \nabla\pi \circ t_{\phi'})_{p(\boldsymbol{\epsilon})})^2 = \sum_{j=1}^{m}(\nabla_{\phi_j} t_{\phi}, b_j \nabla\pi \circ t_{\phi} - b'_j \nabla\pi \circ t_{\phi'})_{p(\boldsymbol{\epsilon})}^2 \qquad (25)\end{aligned}$$

where we in the last three equalities exploit the linearity property of real valued inner products and the final equality follows from Eq. 22.

Now, we apply Cauchy-Schwarz inequality to each term in the sum in Eq. 25

$$
\begin{aligned}
\sum_{j=1}^{m} (\nabla_{\phi_j} t_\phi, b_j \nabla\pi \circ t_\phi - b'_j \nabla\pi \circ t_{\phi'})^2_{p(\boldsymbol{\epsilon})} &\leq \sum_{j=1}^{m} \|\nabla_{\phi_j} t_\phi(\boldsymbol{\epsilon})\|^2_{p(\boldsymbol{\epsilon})} \|b_j \nabla\pi \circ t_\phi - b'_j \nabla\pi \circ t_{\phi'}\|^2_{p(\boldsymbol{\epsilon})} \\
&= \sum_{j=1}^{m} \|b_j \nabla\pi \circ t_\phi - b'_j \nabla\pi \circ t_{\phi'}\|^2_{p(\boldsymbol{\epsilon})} \qquad (26)
\end{aligned}
$$

where $\|\cdot\|_{p(\boldsymbol{\epsilon})}$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle_{p(\boldsymbol{\epsilon})}$, defined in Lemma B.2, and $\|\nabla_{\phi_j} t_\phi(\boldsymbol{\epsilon})\|^2_{p(\boldsymbol{\epsilon})} = 1$.

We now rewrite each term in the sum in Eq. 26 and use the triangle inequality in the first inequality

$$
\begin{aligned}
\sum_{j=1}^{m} \|b_j \nabla\pi \circ t_\phi - b'_j \nabla\pi \circ t_{\phi'}\|^2_{p(\boldsymbol{\epsilon})} &= \sum_{j=1}^{m} \|b_j \nabla\pi \circ t_\phi - b_j \nabla\pi \circ t_{\phi'} + b_j \nabla\pi \circ t_{\phi'} - b'_j \nabla\pi \circ t_{\phi'}\|^2_{p(\boldsymbol{\epsilon})} \\
&\leq \sum_{j=1}^{m} (\|b_j(\nabla\pi \circ t_\phi - \nabla\pi \circ t_{\phi'})\|_{p(\boldsymbol{\epsilon})} + \|(b_j - b'_j)\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})})^2 \\
&= \sum_{j=1}^{m} (|b_j| \cdot \|\nabla\pi \circ t_\phi - \nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})} + |(b_j - b'_j)| \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})})^2.
\end{aligned}
$$

We now exploit Assumption (1), i.e. that π is $M-$Lipschitz smooth and then apply Lemma B.3

$$
\begin{aligned}
&\sum_{j=1}^{m} (|b_j| \cdot \|\nabla\pi \circ t_\phi - \nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})} + |(b_j - b'_j)| \|\nabla f \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})})^2 \\
&\leq \sum_{j=1}^{m} (|b_j| M \mathbb{E}_{p(\boldsymbol{\epsilon})} \|t_\phi(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2 + |(b_j - b'_j)| \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})})^2
\end{aligned}
\tag{27}
$$

Now we apply the triangle inequality

$$
\begin{aligned}
&\sum_{j} (|b_j| \cdot M \cdot \|\phi - \phi'\|_2 + |(b_j - b'_j)| \cdot \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})})^2 \qquad (28) \\
&\leq (\sqrt{\sum_{j} |b_j|^2 M^2 \|\phi - \phi'\|_2^2} + \sqrt{\sum_{j} |(b_j - b'_j)|^2 \cdot \|\nabla\pi \circ t_{\phi'}\|^2_{p(\boldsymbol{\epsilon})}})^2 \\
&= (M\|\phi - \phi'\|_2 \sqrt{\sum_{j} |b_j|^2} + \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})} \sqrt{\sum_{j} |(b_j - b'_j)|^2})^2 \\
&= \big(M\|\phi - \phi'\|_2 \|b\|_2 + \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})} \|b - b'\|_2\big)^2 \qquad (29)
\end{aligned}
$$

This gives us that

$$
\|\nabla l(\phi^{\psi}) - \nabla l(\phi^{\psi'})\|_2 \leq M\|\phi - \phi'\|_2 \|b\|_2 + \|\nabla\pi \circ t_{\phi'}\|_{p(\boldsymbol{\epsilon})} \|b - b'\|_2. \tag{30}
$$

We now individually investigate the terms in Eq. 30 for the log-parameterized approach. Using Assumption (3), i.e. that $\sigma_j < K < \infty$ and Lemma C.2

$$\|b-b'\|_2 = \sqrt{\sum_{j=1}^{d}(b_j - b'_j)^2} = \sqrt{\sum_{j=1}^{d}(e^{\omega_j} - e^{\omega'_j})^2} = \sqrt{\sum_{j=1}^{d}(\sigma_j - \sigma'_j)^2} \leq \sqrt{\sum_{j=1}^{d}K^2(\log\sigma_j - \log\sigma'_j)^2}$$
$$= \sqrt{\sum_{j=1}^{d}K^2(\omega-\omega')^2} \leq \sqrt{\sum_{j=1}^{d}K^2(\phi^{\psi}_j - {\phi^{\psi}}'_j)^2} = K\|\phi^{\psi} - \phi^{\psi'}\|_2. \tag{31}$$

We now consider

$$\|b\|_2 = \sqrt{\sum_{j=1}^{m-d} 1 + dK} = \sqrt{m-d+dK} = \sqrt{m+d(K-1)}. \tag{32}$$

We combine Eqs. 30-32 with Assumption (2) and get that

$$\|\nabla l(\phi^{\psi}) - \nabla l(\phi^{\psi'})\|_2^2 \leq K(M\sqrt{m+d(K-1)}+D)\|\phi^{\psi} - \phi^{\psi'}\|_2$$

□

Lemma 4.3. *Assume that each $\sigma_j \leq K < \infty$, and that L has a mean field parameterization, where each $(L_T)_{i,j} = 0$ for $i \neq j$.*

(1) When $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$, then
$$\|\nabla_{\phi^{\psi}}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+K^2\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2.$$

(2) When $\psi(\boldsymbol{\omega}) = softplus(\boldsymbol{\omega})$, then
$$\|\nabla_{\phi^{\psi}}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+(1-e^{-K})^2\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2.$$

(3) (Kim et al., 2023c) When $\psi(\boldsymbol{\omega}) = \boldsymbol{\omega}$, then
$$\|\nabla_{\phi}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2.$$

where $\mathcal{E}$ is a diagonal matrix where $\mathcal{E}_{ii} = \epsilon_i^2$.

Proof. Starting from Lemma B.4, it is straightforward to see that

$$\|\nabla_{\phi}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 = \|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 + g_\pi^T\mathcal{E}\Psi g_\pi \leq \|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 + \|\mathcal{E}\|_F\|\Psi\|_2\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \tag{33}$$

We now investigate Eq.33 for each type of parameterization. Starting with the exponential parameterization

$$\psi'(\omega_i)^2 = e^{2\omega_i} = e^{2\log\sigma_i} = \sigma_i^2 \leq K^2, \tag{34}$$

where we use the assumption that $\sigma_i \leq K < \infty$ and that $\omega_i \triangleq \log\sigma_i$.

We combine Eq.33-34 and get the following for the exponential parameterization

$$\|\nabla_{\phi}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+K^2\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2.$$

For the softplus parameterization

$$\psi'(\omega_i)^2 = (\frac{e^{\omega_i}}{1+e^{\omega_i}})^2 = (\frac{e^{\sigma_i}-1}{1+e^{\sigma_i}-1})^2 = (1-e^{-\sigma_i})^2 \leq (1-e^{-K})^2, \tag{35}$$

where we use the assumption that $\sigma_i \leq K < \infty$ and that $\omega_i \triangleq \log(e^{\sigma_i}-1)$.

We now combine Eq.33 with Eq.35 and get the following result for softplus

$$\|\nabla_{\phi^\psi}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+(1-e^{-K})^2\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2$$

Finally, for the linear parameterization

$$\psi'(\omega_i)^2 = 1^2 = 1, \tag{36}$$

as $\psi(\omega_i) = \omega_i$.

Thus, for the linear approach we get

$$\|\nabla_{\phi^\psi}\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq (1+\|\mathcal{E}\|_F)\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2.$$

□

Lemma 4.4. *Let* $\mathbf{g}$ *be the gradient estimator of* $l(\cdot)$*, and assume that (1) each* $\sigma_j \leq K < \infty$*, (2) that* $\mathbf{z}$ *is a stationary point of* π*, (3) that* L *has a mean field parameterization, and (4)* π *is M-Lipschitz smooth.*

(1) When $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}}$*, then*

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&(K^2 2\sqrt{d\kappa}+1)\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2 \\ &+(K^2(\sqrt{d\kappa}+\sqrt{d}\kappa)+1)\|L\|_F^2\Big).\end{aligned} \tag{37}$$

(2) When $\psi(\boldsymbol{\omega}) = softplus(\boldsymbol{\omega})$*, then*

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&((1-e^{-K})^2 2\sqrt{d\kappa}+1)\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2 \\ &+((1-e^{-K})^2(\sqrt{d\kappa}+\sqrt{d}\kappa)+1)\|L\|_F^2\Big).\end{aligned} \tag{38}$$

(3) (Kim et al., 2023c) When $\psi(\boldsymbol{\omega}) = \boldsymbol{\omega}$*, then*

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}\|_2^2] \leq M^2 \Big(&(2\sqrt{d\kappa}+1)\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2 \\ &+(\sqrt{d\kappa}+\sqrt{d}\kappa+1)\|L\|_F^2\Big).\end{aligned} \tag{39}$$

Proof. We aim to present a result analogous to Lemma B.5. Unlike the original lemma, our result does not require ψ to be 1-Lipschitz and instead assumes an upper bound on σ_j.

We start by taking the expectation of (1) in Lemma 4.3

$$\mathbb{E}\|\mathbf{g}\|_2^2 = \mathbb{E}\|\nabla\pi_{\phi^\psi}(t_\phi(\boldsymbol{\epsilon}))\|_2^2 \leq \mathbb{E}\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2(1+K^2\|\mathcal{E}\|_F). \tag{40}$$

Analogues to the proof of Theorem B.7, we now exploit assumption (2), which implies that $\nabla\pi_\phi(z) = 0$

$$\begin{aligned}\mathbb{E}\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))\|_2^2(1+K^2\|\mathcal{E}\|_F) &= \mathbb{E}\|\nabla\pi(t_\phi(\boldsymbol{\epsilon}))-\nabla\pi(\boldsymbol{z})\|_2^2(1+K^2\|\mathcal{E}\|_F)\\ &\le M^2\mathbb{E}\|t_\phi(\boldsymbol{\epsilon})-\boldsymbol{z}\|_2^2(1+K^2\|\mathcal{E}\|_F) = M^2(\mathbb{E}\|t_\phi(\boldsymbol{\epsilon})-\boldsymbol{z}\|_2^2+K^2\mathbb{E}\|\mathcal{E}\|_F\|t_\phi(\boldsymbol{\epsilon})-\boldsymbol{z}\|_2^2)\\ &= M^2(\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2+\|L\|_F^2+K^2\mathbb{E}\|\mathcal{E}\|_F\|t_\phi(\boldsymbol{\epsilon})-\boldsymbol{z}\|_2^2), \end{aligned} \tag{41}$$

where we use assumption (4) in the inequality, and in the last equality apply Lemma B.6.

From Lemmas B.5 and B.6 it is straightforward to see that

$$\mathbb{E}\|t_\phi(\boldsymbol{\epsilon})-\mathbf{z}\|_2^2\|\mathcal{E}\|_F \le (2\sqrt{\kappa d})\|\boldsymbol{\mu}-\mathbf{z}\|_2^2+(\sqrt{d\kappa}+\kappa\sqrt{d})\|L\|_F^2. \tag{42}$$

We now combine Eq.40-42

$$\begin{aligned}\mathbb{E}\|\mathbf{g}\|_2^2 &\le M^2\left(\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2+\|L\|_F^2+K^2\left((2\sqrt{\kappa d})\|\boldsymbol{\mu}-\mathbf{z}\|_2^2+(\sqrt{d\kappa}+\kappa\sqrt{d})\|L\|_F^2\right)\right)\\ &= M^2\left((K^2 2\sqrt{d\kappa}+1)\|\boldsymbol{\mu}-\boldsymbol{z}\|_2^2+(K^2(\sqrt{d\kappa}+\sqrt{d}\kappa)+1)\|L\|_F^2\right), \end{aligned} \tag{43}$$

which corresponds to (1) in Lemma 4.4.

The results for (2) and (3) in Lemma 4.4 can be shown through similar derivations. □

Lemma C.3. *Let $a, b \in [\delta, \infty)$ for some $\delta > 0$. Then,*

$$(a-b)^2 \ge \delta^2(\log(a)-\log(b))^2.$$

Proof. Let $f(a) = \log(a)$, then as f is continuous on $[a, b]$ and differentiable on (a, b), by the Mean Value Theorem, there exists some $c \in (a, b)$ such that:

$$\frac{1}{c} = \frac{\log(a)-\log(b)}{a-b}.$$

We square both sides and rearrange

$$\frac{1}{c^2}(a-b)^2 = (\log(a)-\log(b))^2.$$

Given that $a, b \in [\delta, \infty)$, this implies that $0 < \delta \le c$, and thus

$$\frac{1}{\delta^2}(a-b)^2 \ge (\log(a)-\log(b))^2.$$

□

Theorem 4.5. *Let π be λ-strongly convex, $\psi(\omega) = e^\omega$, and assume that each $\sigma_j \ge \delta$, where $1 \ge \delta > 0$. Then $l(\phi^\psi)$ is $\lambda\delta^2-$strongly convex for the mean-field parameterization.*

Proof. To show that the energy $l(\phi^\psi)$ is strongly convex in terms of the variational parameters ϕ^ψ, we want to show the following

$$l(\phi^\psi) \ge l(\phi^{\psi\prime}) + \langle\nabla l(\phi^{\psi\prime}), \phi^\psi - \phi^{\psi\prime}\rangle + \lambda\frac{1}{2}\|\phi^\psi - \phi^{\psi\prime}\|_2^2, \quad \forall\phi^\psi, \phi^{\psi\prime}. \tag{44}$$

To do so, we start by using the assumption that π is λ-strongly convex, to form the following inequality

$$\pi(\boldsymbol{z}) \geq \pi(\boldsymbol{z}') + \langle \nabla \pi(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \rangle + \lambda \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{z}'\|_2^2, \quad \forall \boldsymbol{z}, \boldsymbol{z}'. \tag{45}$$

We then take the expectations of both sides in the inequality in Eq. 45 and use the reparametrization trick

$$l(\phi^{\psi}) \geq l(\phi^{\psi'}) + \mathbb{E}\langle \nabla \pi(t_{\phi'}(\boldsymbol{\epsilon})), t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon}) \rangle + \lambda \frac{1}{2} \mathbb{E}\|t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2^2, \quad \forall \phi^{\psi}, \phi^{\psi'}. \tag{46}$$

From Lemma B.8, under the assumption that we use the mean field parameterization, we obtain

$$\begin{aligned} l(\phi^{\psi}) &\geq l(\phi^{\psi'}) + \mathbb{E}\langle \nabla \pi(t_{\phi'}(\boldsymbol{\epsilon})), t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon}) \rangle + \lambda \frac{1}{2} \mathbb{E}\|t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2^2 \\ &\geq l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \mathbb{E}\|t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2^2 \end{aligned} \tag{47}$$

From Lemma B.3 and the proof of (4) in Theorem B.9 we obtain

$$\begin{aligned} &l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \mathbb{E}\|t_{\phi}(\boldsymbol{\epsilon}) - t_{\phi'}(\boldsymbol{\epsilon})\|_2^2 \geq l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \left(\|L - L'\|_F^2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \right) \\ &= l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \left(\|\psi(\boldsymbol{\omega}) - \psi(\boldsymbol{\omega}')\|_2^2 + \|L_T - L_T'\|_F^2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \right) \end{aligned} \tag{48}$$

Now, let $\psi(\boldsymbol{\omega}) = e^{\boldsymbol{\omega}} = \boldsymbol{\sigma}$. Then

$$\|\psi(\boldsymbol{\omega}) - \psi(\boldsymbol{\omega}')\|_2^2 = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}'\|_2^2 = \sum_{j=1}^{d} (\sigma_j - \sigma_j')^2. \tag{49}$$

Assume that each $\sigma_j \geq \delta > 0$, and apply Lemma C.3

$$\sum_{j=1}^{d} (\sigma_j - \sigma_j')^2 \geq \sum_{j=1}^{d} \delta^2 (\log \sigma_j - \log \sigma_j')^2 = \delta^2 \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2. \tag{50}$$

We now combine this result with Eq.48 and use the assumption that $1 \geq \delta > 0$:

$$\begin{aligned} l(\phi^{\psi}) &\geq l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \left(\delta^2 \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 + \|L_T - L_T'\|_F^2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \right) \\ &\geq l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \frac{1}{2} \left(\delta^2 \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2^2 + \delta^2 \|L_T - L_T'\|_F^2 + \delta^2 \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \right) \\ &= l(\phi^{\psi'}) + \langle \nabla l(\phi^{\psi}), \phi^{\psi} - \phi^{\psi'} \rangle + \lambda \delta^2 \frac{1}{2} \|\phi^{\psi} - \phi^{\psi'}\|_2^2. \end{aligned} \tag{51}$$

Thus, we have demonstrated that the function $l(\phi^{\psi})$ is $\lambda\delta^2$-strongly convex, provided that the diagonal scale components are lower-bounded by δ. However, even though the exponential scale parameterization preserves strong convexity, it reduces the strong convexity factor of the energy compared to the linear approach, which maintains the strong convexity factor as shown in (1) in Theorem B.9.

□