Practical and Matching Gradient Variance Bounds for Black-Box Variational Bayesian Inference

Kyurae Kim [1] **Kaiwen Wu** [1] **Jisu Oh** [2] **Jacob R. Gardner** [1]

Abstract

Understanding the gradient variance of black-box variational inference (BBVI) is a crucial step for establishing its convergence and developing algorithmic improvements. However, existing studies have yet to show that the gradient variance of BBVI satisfies the conditions used to study the convergence of stochastic gradient descent (SGD), the workhorse of BBVI. In this work, we show that BBVI satisfies a *matching* bound corresponding to the *ABC* condition used in the SGD literature when applied to smooth and quadratically-growing log-likelihoods. Our results generalize to nonlinear covariance parameterizations widely used in the practice of BBVI. Furthermore, we show that the variance of the mean-field parameterization has provably superior dimensional dependence.

1. Introduction

Variational inference (VI; Jordan et al. 1999; Blei et al. 2017; Zhang et al. 2019) algorithms are fast and scalable Bayesian inference methods widely applied in fields of statistics and machine learning. In particular, black-box VI (BBVI; Ranganath et al. 2014; Titsias & Lázaro-Gredilla 2014) leverages stochastic gradient descent (SGD; Robbins & Monro 1951; Bottou 1999) for inference of non-conjugate probabilistic models. With the development of bijectors (Kucukelbir et al., 2017; Dillon et al., 2017; Fjelde et al., 2020), most of the methodological advances in BBVI have now been abstracted out through various probabilistic programming frameworks (Carpenter et al., 2017; Ge et al., 2018; Dillon et al., 2017; Bingham et al., 2019; Salvatier et al., 2016).

Despite the advances of BBVI, little is known about its theoretical properties. Even when restricted to the location-scale family (Definition 2), it is unknown whether BBVI is guaranteed to converge without having to modify the algorithms used in practice, for example, by enforcing bounded domains, bounded support, bounded gradients, and such. This theoretical insight is necessary since BBVI methods are known to be less robust (Yao et al., 2018; Dhaka et al., 2020; Welandawe et al., 2022; Dhaka et al., 2021; Domke, 2020) compared to other inference methods such as Markov chain Monte Carlo. Although progress has been made to formalize the theory of BBVI with some generality, the gap between our understanding of BBVI and the convergence guarantees of SGD remains open. For example, Domke (2019; 2020) provided smoothness and gradient variance guarantees. Still, these results do not yet yield a full convergence guarantee and do not extend to *nonlinear* covariance parameterizations used in practice.

In this work, we investigate whether recent progress in relaxing the gradient variance assumptions used in SGD (Tseng, 1998; Vaswani et al., 2019; Schmidt & Roux, 2013; Bottou et al., 2018; Gower et al., 2019; 2021b; Nguyen et al., 2018) apply to BBVI. These extensions have led to new insights that the structure of the gradient bounds can have non-trivial interactions with gradient-adaptive SGD algorithms (Zhang et al., 2022). For example, when the "interpolation assumption" (the gradient noise converges to 0; Schmidt & Roux 2013; Ma et al. 2018; Vaswani et al. 2019) does not hold, ADAM (Kingma & Ba, 2015) provably diverges with certain stepsize combinations (Zhang et al., 2022). Until BBVI can be shown to conform to the assumptions used by these recent works, it is unclear how these results relate to BBVI.

While the variance of BBVI gradient estimators has been studied before (Xu et al., 2019; Domke, 2019; Mohamed et al., 2020a; Fujisawa & Sato, 2021), the connection with the conditions used in SGD has yet to be established. As such, we answer the following question:

> *Does the gradient variance of BBVI conform to the conditions assumed in convergence guaran-*

[1]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, United States [2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States. Correspondence to: Kyurae Kim <kyrkim@seas.upenn.edu>, Jacob R. Gardner <jrgardner@seas.upenn.edu>.

Proceedings of the 40th *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

tees of SGD without modifying the implementations used in practice?

The answer is yes! Assuming the target log joint distribution is smooth and quadratically growing, we show that the gradient variance of BBVI satisfies the *ABC* condition (Assumption 2) used by Polyak & Tsypkin (1973); Khaled & Richtárik (2023); Gower et al. (2021b). Our analysis extends the previous result of Domke (2019) to covariance parameterizations involving nonlinear functions for conditioning the diagonal (see Section 2.5), as commonly done in practice. Furthermore, we prove that the gradient variance of the mean-field parameterization (Peterson & Anderson, 1987; Peterson & Hartman, 1989; Hinton & van Camp, 1993) results in better dimensional dependence compared to full-rank ones.

Overall, our results should act as a key ingredient to obtaining a full convergence guarantees of BBVI, as recently done by Kim et al. (2023).

Our contributions are summarized as follows:

❶ We provide upper bounds on the gradient variance of BBVI that matches the *ABC condition* (Assumption 2) used for analyzing SGD.

- ➤ Theorems 1 and 2 do not require any modification of the algorithms used in practice.
- ➤ Theorem 3 achieves better constants under the stronger *bounded entropy* assumption.

❷ Our analysis applies to BBVI parameterizations (Section 2.5) widely used in practice (Table 1).

- ➤ Lemma 1 enables the bounds to cover nonlinear covariance parameterizations.
- ➤ Lemma 3 and Remark 4 shows that the gradient variance of the mean-field parameterization has superior dimensional scaling.

❸ We provide a matching lower bound (Theorem 4) on the gradient variance, showing that, under the stated assumptions, the ABC condition is the weakest assumption applicable to BBVI.

2. Preliminaries

Notation Random variables are denoted in serif, while their realization is in regular font. (*i.e*, x is a realization of x, $\boldsymbol{x}$ is a realization of the vector-valued $\boldsymbol{\mathsf{x}}$.) $\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ denotes the Euclidean norm, while $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}\left(\boldsymbol{A}^\top \boldsymbol{A}\right)}$ is the Frobenius norm, where $\mathrm{tr}\left(\boldsymbol{A}\right) = \sum_{i=1}^{d} A_{ii}$ is the matrix trace.

2.1. Variational Inference

Variational inference (Peterson & Anderson, 1987; Hinton & van Camp, 1993) is a family of inference algorithms devised to solve the problem

$$\underset{\boldsymbol{\lambda} \in \mathbb{R}^p}{\text{minimize}} \;\; \mathrm{D}_{\mathrm{KL}}(q_{\psi,\boldsymbol{\lambda}}, \pi), \tag{1}$$

where $q_{\psi,\boldsymbol{\lambda}}$ is called the "variational approximation", while π is a distribution of interest, and D_{KL} is the (exclusive) Kullback-Leibler (KL) divergence.

For Bayesian inference, π is the posterior distribution

$$\pi\left(\boldsymbol{z}\right) \propto \ell\left(\boldsymbol{x} \mid \boldsymbol{z}\right) p\left(\boldsymbol{z}\right) = \ell\left(\boldsymbol{x}, \boldsymbol{z}\right),$$

where $\ell\left(\boldsymbol{x} \mid \boldsymbol{z}\right)$ is the likelihood, and $p\left(\boldsymbol{z}\right)$ is the prior. In practice, one only has access to the likelihood and the prior. Thus, Equation (1) cannot be directly solved. Instead, we can minimize the negative *evidence lower bound* (ELBO; Jordan et al. 1999) function $F\left(\boldsymbol{\lambda}\right)$.

Evidence Lower Bound More formally, we solve

$$\underset{\boldsymbol{\lambda} \in \mathbb{R}^p}{\text{minimize}} \;\; F\left(\boldsymbol{\lambda}\right),$$

where F is defined as

$$F\left(\boldsymbol{\lambda}\right) \triangleq -\mathbb{E}_{\boldsymbol{\mathsf{z}} \sim q_{\psi,\boldsymbol{\lambda}}}\left[\log \ell\left(\boldsymbol{x}, \boldsymbol{\mathsf{z}}\right)\right] - \mathrm{H}\left(q_{\psi,\boldsymbol{\lambda}}\right), \tag{2}$$

$$= -\mathbb{E}_{\boldsymbol{\mathsf{z}} \sim q_{\psi,\boldsymbol{\lambda}}}\left[\log \ell\left(\boldsymbol{x} | \boldsymbol{\mathsf{z}}\right)\right] + \mathrm{D}_{\mathrm{KL}}(q_{\psi,\boldsymbol{\lambda}}, p), \tag{3}$$

$\boldsymbol{\mathsf{z}}$	is the latent (random) variable,
$q_{\psi,\boldsymbol{\lambda}}$	is the variational distribution,
ψ	is a bijector (support transformation), and
H	is the differential entropy.

The bijector ψ (Dillon et al., 2017; Fjelde et al., 2020; Leger, 2023) is a differentiable bijective map that is used to de-constrain the support of constrained random variables. For example, when z is expected to follow a gamma distribution, using $\eta = \psi\left(z\right)$ with $\psi\left(z\right) = \log z$ lets us work with η, which can be any real number, unlike z. The use of ψ^{-1} corresponds to the automatic differentiation VI formulation (ADVI; Kucukelbir et al. 2017), which is now widespread.

2.2. Variational Family

In this work, we specifically consider the location-scale variational family with a standardized base distribution.

Definition 1 (Reparameterization Function). An affine mapping $\boldsymbol{t}_{\boldsymbol{\lambda}} : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\boldsymbol{t}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right) \triangleq \boldsymbol{C}\boldsymbol{u} + \boldsymbol{m}$$

with $\boldsymbol{\lambda}$ containing the parameters for forming the location $\boldsymbol{m} \in \mathbb{R}^d$ and scale $\boldsymbol{C} = \boldsymbol{C}\left(\boldsymbol{\lambda}\right) \in \mathbb{R}^{d \times d}$ is called the (location-scale) *reparameterization function*.

Definition 2 (Location-Scale Family). Let φ be some d-dimensional distribution. Then, $q_{\boldsymbol{\lambda}}$ such that

$$\boldsymbol{\zeta} \sim q_{\boldsymbol{\lambda}} \quad \Leftrightarrow \quad \boldsymbol{\zeta} \overset{d}{=} \boldsymbol{t}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right); \quad \boldsymbol{u} \sim \varphi$$

is said to be a member of the location-scale family indexed by the base distribution φ and parameter $\boldsymbol{\lambda}$.

This family includes commonly used variational families, such as the mean-field Gaussian, full-rank Gaussian, Student-T, and other elliptical distributions.

Remark 1 (**Entropy of Location-Scale Distributions**). The differential entropy of a location-scale family distribution (Definition 2) is

$$\mathrm{H}(q_\lambda) = \mathrm{H}(\varphi) + \log|\boldsymbol{C}|.$$

Definition 3 (**ADVI Family**; Kucukelbir et al. 2017)**.** Let q_λ be some d-dimensional distribution. Then, $q_{\psi,\lambda}$ such that

$$\boldsymbol{z} \sim q_{\psi,\lambda} \quad \Leftrightarrow \quad \boldsymbol{z} \overset{d}{=} \psi^{-1}(\boldsymbol{\zeta}); \quad \boldsymbol{\zeta} \sim q_\lambda$$

is said to be a member of the ADVI family with the base distribution q_λ parameterized with $\boldsymbol{\lambda}$.

We impose assumptions on the base distribution φ.

Assumption 1 (**Base Distribution**)**.** φ is a d-dimensional distribution such that $\boldsymbol{u} \sim \varphi$ and $\boldsymbol{u} = (u_1, \dots, u_d)$ with indepedently and identically distributed components. Furthermore, φ is **(i)** symmetric and standardized such that $\mathbb{E}u_i = 0$, $\mathbb{E}u_i^2 = 1$, $\mathbb{E}u_i^3 = 0$, and **(ii)** has finite kurtosis $\mathbb{E}u_i^4 = \kappa < \infty$.

These assumptions are already satisfied in practice by, for example, generating u_i from a univariate normal or Student-T with $\nu > 4$ degrees of freedom.

2.3. Reparameterization Trick

When restricted to location scale families (Definitions 2 and 3), we can invoke Change-of-Variable, or more commonly known as the "reparameterization trick," such that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim q_{\psi,\lambda}} \log \ell(\boldsymbol{x}, \boldsymbol{z}) &= \mathbb{E}_{\boldsymbol{\zeta}\sim q_\lambda} \log \ell\left(\boldsymbol{x}, \psi^{-1}(\boldsymbol{\zeta})\right) \\ &= \mathbb{E}_{\boldsymbol{u}\sim\varphi} \log \ell\left(\boldsymbol{x}, \psi^{-1}(\boldsymbol{t}_\lambda(\boldsymbol{u}))\right)\end{aligned}$$

through the Law of the Unconcious Statistician. Differentiating this results in the *reparameterization* or *path* gradient, which often achieves lower variance than alternatives (Xu et al., 2019; Mohamed et al., 2020b).

Objective Function For generality, we represent our objective as a composite infinite sum problem:

Definition 4 (**Composite Infinite Sum**)**.**

$$F(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{u}\sim\varphi} f(\boldsymbol{t}_\lambda(\boldsymbol{u})) + h(\boldsymbol{\lambda}),$$

where $(\boldsymbol{\lambda}, \boldsymbol{u}) \mapsto f \circ \boldsymbol{t}_\lambda : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is some bivariate stochastic function of $\boldsymbol{\lambda}$ and the "noise source" $\boldsymbol{u}$, while h is a deterministic regularization term.

By appropriately defining f and h, we retrieve the two most common formulations of the ELBO in Equation (2) and Equation (3) respectively:

Definition 5 (**ELBO Entropy-Regularized Form**)**.**

$$f_{\mathrm{H}}(\boldsymbol{\zeta}) = -\log \underbrace{\ell\left(\boldsymbol{x}, \psi^{-1}(\boldsymbol{\zeta})\right)}_{\text{Joint Likelihood}} - \log\left|\boldsymbol{J}_{\psi^{-1}}(\boldsymbol{\zeta})\right|$$

$$h_{\mathrm{H}}(\boldsymbol{\lambda}) = -\mathrm{H}(q_\lambda).$$

Definition 6 (**ELBO KL-Regularized Form**)**.**

$$f_{\mathrm{KL}}(\boldsymbol{\zeta}) = -\log \underbrace{\ell\left(\boldsymbol{x} \mid \psi^{-1}(\boldsymbol{\zeta})\right)}_{\text{Likelihood}} - \log\left|\boldsymbol{J}_{\psi^{-1}}(\boldsymbol{\zeta})\right|$$

$$h_{\mathrm{KL}}(\boldsymbol{\lambda}) = \mathrm{D}_{\mathrm{KL}}(q_\lambda, p).$$

Here, $\boldsymbol{J}_{\psi^{-1}}$ is the Jacobian of the bijector. Since $\mathrm{D}_{\mathrm{KL}}(q_\lambda, p)$ is seldomly available in tractable form, the entropy-regularized form is the most widely used, while the KL regularized is common for Gaussian processes and variational autoencoders.

Gradient Estimator We denote the M-sample estimator of the gradient of F as

$$\boldsymbol{g}_M(\boldsymbol{\lambda}) \triangleq \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}_m(\boldsymbol{\lambda}), \text{ where} \tag{4}$$

$$\boldsymbol{g}_m(\boldsymbol{\lambda}) \triangleq \nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}_m)) + \nabla h(\boldsymbol{\lambda}); \quad \boldsymbol{u}_m \sim \varphi. \tag{5}$$

We will occasionally drop $\boldsymbol{\lambda}$ for clarity.

2.4. Gradient Variance Assumptions in Stochastic Gradient Descent

Gradient Variance Assumptions in SGD For a while, most convergence proofs in SGD have relied on the "bounded variance" assumption. That is, for a gradient estimator $\boldsymbol{g}$, $\mathbb{E}\|\boldsymbol{g}\|_2^2 \leq G$ for some finite constant G. This assumption is problematic because ❶ these types of global constants result in loose bounds, ❷ and it directly contradicts the strong-convexity assumption (Nguyen et al., 2018). Thus, retrieving previously known SGD convergence rates under weaker assumptions has been an important research direction (Tseng, 1998; Vaswani et al., 2019; Schmidt & Roux, 2013; Bottou et al., 2018; Gower et al., 2019; 2021b; Nguyen et al., 2018).

ABC Condition In this work, we focus on the recently rediscovered *expected smoothness*, or *ABC*, condition (Polyak & Tsypkin, 1973; Gower et al., 2021b).

Assumption 2 (**Expected Smoothness;** *ABC*)**.** $\boldsymbol{g}$ is said to satisfy the expected smoothness condition if

$$\mathbb{E}\|\boldsymbol{g}_M(\boldsymbol{\lambda})\|_2^2 \leq 2A\left(F(\boldsymbol{\lambda}) - F^*\right) + B\|\nabla F(\boldsymbol{\lambda})\|_2^2 + C.$$

for some finite $A, B, C \geq 0$, where $F^* = \inf_{\lambda\in\mathbb{R}^p} F(\boldsymbol{\lambda})$.

As shown by Khaled & Richtárik (2023), this condition is not only strictly weaker than many of the previously used assumptions but also *generalizes* them by retrieving known convergence rates when tweaking the constants.

With the ABC condition, for non-convex L-smooth functions, under a "appropriately chosen" stepsize (otherwise the bound may blow-up as explained by Khaled & Richtárik) of $\gamma \leq 1/LB$, SGD converges to a $\mathcal{O}(LC\gamma)$ neighborhood in a $\mathcal{O}\left(\left(1+L\gamma^2 A\right)^T/(\gamma T)\right)$ rate. Minor variants of the ABC condition have also been used to prove convergence of SGD for quasar convex functions Gower et al. (2021a), stochastic heavy-ball/momentum methods Liu & Yuan (2022), and stochastic proximal methods (Li & Milzarek, 2022). Given the influx of results based on the ABC condition, connecting with it would significantly broaden our theoretical understanding of BBVI.

Table 1: Survey of Parameterizations Used in Black-Box Variational Inference

Framework	Version	Parameterizations	Conditioner	Code
TURING (Ge et al., 2018)	`v0.23.2`	Nonlinear Mean-field	softplus	link
STAN (Carpenter et al., 2017)	`v2.31.0`	Nonlinear Mean-field	exp	link
		Linear Cholesky		link
PYRO (Bingham et al., 2019)	`v0.10.1`	Nonlinear Mean-field	softplus	link
		Linear Cholesky[1]		link
PYMC3 (Salvatier et al., 2016)	`v5.0.1`	Nonlinear Mean-field	softplus	link
		Nonlinear Cholesky	softplus	link
GPYTORCH (Gardner et al., 2018)	`v1.9.0`	Linear Cholesky		link
		Linear Mean-field		link

[1] Numpyro also provides a low-rank Cholesky parameterization, which is non-linearly conditioned. But the full-rank Cholesky is linear.

* Tensorflow probability (Dillon et al., 2017) wasn't included as it doesn't provide a fully pre-configured variational family (although `tfp.experimental.vi.build_*_posterior` exists, the parameterization is user-supplied).

2.5. Covariance Parameterizations

When using the location-scale family (Definition 2), the scale matrix $\boldsymbol{C}$ can be parameterized in different ways. Any parameterization that results in a positive definite covariance $\boldsymbol{C}\boldsymbol{C}^\top \in \mathbb{S}^d_{++}$ is valid. We consider multiple parameterizations as the choice can result in different theoretical properties. A brief survey on the use of different parameterizations is shown in Table 1.

Linear Parameterization The previous results by Domke (2019) considered the matrix square root parameterization, which is linear with respect to the variational parameters.

Definition 7 (**Matrix Square Root**).

$$\boldsymbol{C}(\boldsymbol{\lambda}) = \boldsymbol{C},$$

where $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ is a matrix, $\boldsymbol{\lambda}_{\boldsymbol{C}} = \mathrm{vec}(\boldsymbol{C}) \in \mathbb{R}^{d^2}$ such that $\boldsymbol{\lambda} = (\boldsymbol{m}, \boldsymbol{\lambda}_{\boldsymbol{C}})$.

Note that $\boldsymbol{C}$ is not constrained to be symmetric so this is not a matrix square root in a narrow sense. Also, this parameterization does not guarantee $\boldsymbol{C}\boldsymbol{C}^\top$ to be positive definite (only positive semidefinite), which occasionally results in the entropy term h_{H} blowing up (Domke, 2020). Domke proposed to fix this by using proximal operators.

Nonlinear Parameterizations In practice, optimization is preferably done in unconstrained $\mathbb{R}^p$, which then positive definiteness can be ensured by explicitly mapping the diagonal elements to positive numbers. We denote this by the *diagonal conditioner* ϕ. (See Table 1 for a brief survey on their use). The following two parameterizations are commonly used, where $\boldsymbol{D} = \mathrm{diag}(\phi(\boldsymbol{s})) \in \mathbb{R}^{d\times d}$ denotes a diagonal matrix such that $D_{ii} = \phi(s_i) > 0$.

Definition 8 (**Mean-Field**).

$$\boldsymbol{C}(\boldsymbol{\lambda}, \phi) = \mathrm{diag}(\phi(\boldsymbol{s})),$$

where $\boldsymbol{s} \in \mathbb{R}^d$ and $\boldsymbol{\lambda} = (\boldsymbol{m}, \boldsymbol{s})$.

Definition 9 (**Cholesky**).

$$\boldsymbol{C}(\boldsymbol{\lambda}, \phi) = \mathrm{diag}(\phi(\boldsymbol{s})) + \boldsymbol{L},$$

where $\boldsymbol{s} \in \mathbb{R}^d$, $\boldsymbol{L} \in \mathbb{R}^{d\times d}$ is a strictly lower triangular matrix, $\boldsymbol{\lambda}_{\boldsymbol{L}} = \mathrm{vec}(\boldsymbol{L}) \in \mathbb{R}^{(d+1)d/2}$ such that $\boldsymbol{\lambda} = (\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\lambda}_{\boldsymbol{L}})$. The special case of $\phi(x) = x$ is called the "linear Cholesky" parameterization.

Diagonal conditioner For the diagonal conditioner, the softplus function $\phi(x) = \mathrm{softplus}(x) \triangleq \log(1+e^x)$ (Dugas et al., 2000) or the exponential function $\phi(x) = e^x$ is commonly used. While using these nonlinear functions significantly complicates the analysis, assuming ϕ to be 1-Lipschitz retrieves practical guarantees.

Assumption 3 (**Lipschitz Diagonal Conditioner**). The diagonal conditioner ϕ is 1-Lipschitz continuous.

Remark 2. The softplus function is 1-Lipschitz.

3. Main Results

3.1. Key Lemmas

The main challenge in studying BBVI is that the gradient of the composed function $\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))$ is different from ∇f. For the matrix square root parameterization, Domke (2019) established the connection through Lemma 1 (restated as Lemma 6 in Appendix C.1). We generalize this result to nonlinear parameterizations:

Lemma 1. *Let $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ be a location-scale reparameterization function (Definition 1) with some differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. Then, for $\boldsymbol{g}_f \triangleq \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))$,*

(i) Mean-Field

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{g}_f,$$

(ii) Cholesky

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{\Sigma}\boldsymbol{g}_f + \boldsymbol{g}_f^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I})\boldsymbol{g}_f,$$

where $\boldsymbol{U}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}$ are diagonal matrices, which the diagonals are defined as

$$U_{ii} = u_i^2, \quad \Phi_{ii} = \phi'(s_i)^2, \quad \Sigma_{ii} = \textstyle\sum_{j=1}^i u_j^2,$$

and ϕ is a diagonal conditioner for the scale matrix.

Proof. See the *full proof* in Appendix C.2.1.

Note that the relationships in this lemma are all equalities, which can be bounded with known quantities, as done in the next lemma. We note here that if any of our analyses were to be improved, this shall by done by obtaining tighter bounds on the equalities in Lemma 1.

Lemma 2. *Let $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ be a location-scale reparameterization function (Definition 1), $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and let ϕ satisfy Assumption 3.*

(i) Mean-Field

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \leq \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2,$$

where $\boldsymbol{U}$ is a diagonal matrix such that $U_{ii} = u_i^2$.

(ii) Cholesky

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \leq \left(1 + \|\boldsymbol{u}\|_2^2\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2,$$

where the equality holds for the matrix square root parameterization.

Proof. See the *full proof* in Appendix C.2.2.

Lemma 1 act as the interface between the properties of the parameterization and the likelihood f.

Remark 3 (**Variance Reduction Through ϕ**). A *nonlinear* Cholesky parameterization with a 1-Lipschitz ϕ achieves lower or equal variance compared to the matrix square root and *linear* Cholesky, where the equality is achieved with the matrix square root parameterization.

Dimension Dependence of Mean-Field The superior dimensional dependence of the mean-field parameterization is given by the following lemma:

Lemma 3. *Let the assumptions of Lemma 2 hold and $\boldsymbol{u} \sim \varphi$ satisfy Assumption 1. Then, for the mean-field parameterization,*

$$\mathbb{E}\|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2 \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right) \leq \left(\sqrt{d}\kappa + \kappa\sqrt{d} + 1\right) \|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + \left(2\kappa\sqrt{d} + 1\right) \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

Proof. See the *full proof* in Appendix C.2.3.

Remark 4 (**Superior Variance of Mean-Field**). The mean-field parameterization has $\mathcal{O}\left(\sqrt{d}\right)$ dimensional dependence compared to the $\mathcal{O}(d)$ dimensional dependence of the full-rank parameterizations in Lemma 7.

Lastly, the following lemma is the basic building block for all of our upper bounds:

Lemma 4. *Let $\boldsymbol{g}_M$ be the M-sample gradient estimator of F (Definition 4) for some function f, h and let $\boldsymbol{u}$ be some random variable. Then,*

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{1}{M}\mathbb{E}\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 + \|\nabla F(\boldsymbol{\lambda})\|_2^2.$$

Proof. See the *full proof* in Appendix C.2.4.

3.2. Upper Bounds

We restrict our analysis to the class of log-likelihoods that satisfy the following conditions:

Definition 10 (**L-smoothness**). A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if it satisfies the following for all $\boldsymbol{\zeta}, \boldsymbol{\zeta}' \in \mathbb{R}^d$:

$$\|\nabla f(\boldsymbol{\zeta}) - \nabla f(\boldsymbol{\zeta}')\|_2 \leq L\|\boldsymbol{\zeta} - \boldsymbol{\zeta}'\|_2.$$

Definition 11 (**Quadratic Functional Growth**). A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-quadratically growing if

$$\frac{\mu}{2}\|\boldsymbol{\zeta} - \bar{\boldsymbol{\zeta}}\|_2^2 \leq f(\boldsymbol{\zeta}) - f^*$$

for all $\boldsymbol{\zeta} \in \mathbb{R}^d$, where $\bar{\boldsymbol{\zeta}} = \Pi_f(\boldsymbol{\zeta})$ is a projection of $\boldsymbol{\zeta}$ onto the set of minimizers of f and $f^* = \inf_{\boldsymbol{\zeta} \in \mathbb{R}^d} f(\boldsymbol{\zeta})$.

The quadratic growth condition has first been used by (Anitescu, 2000) and is strictly weaker than the Polyak-Łojasiewicz inequality (see Karimi et al. 2016, Appendix A for the proof). Furthermore, for μ-strongly (quasar) convex functions (Hinder et al., 2020; Jin, 2020) automatically satisfy quadratic growth, but our analysis does *not* require (quasar) convexity.

Both assumptions are commonly used in SGD. For studying the gradient variance of BBVI, assuming both smoothness and quadratic growth is weaker than the assumptions of Xu et al. (2019) but stronger than those of Domke (2019), who assumed only smoothness. The additional assumption on growth is necessary to extend his results to establish the *ABC* condition.

For the variational family, we assume the followings:

Assumption 4. $q_{\psi,\lambda}$ is a member of the ADVI family (Definition 3), where the underlying q_λ is a member of the location-scale family (Definition 2) with its base distribution φ satisfying Assumption 1.

Entropy-Regularized Form First, we provide the upper bound for the ELBO in entropy-regularized form. This result does *not* require any modifications to vanilla SGD.

Theorem 1. *Let $\boldsymbol{g}_M$ be an M-sample estimate of the gradient of the ELBO in entropy regularized form (Definition 5). Also, assume that Assumption 3 and 4 hold,*

- *f_{H} is L_{H}-smooth, and*
- *f_{KL} is μ_{KL}-quadratically growing.*

Then,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq & \frac{4L_{\mathrm{H}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\ & + \frac{2L_{\mathrm{H}}^2}{M} C(d,\kappa)\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2 \\ & + \frac{4L_{\mathrm{H}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)\left(F^* - f_{\mathrm{KL}}^*\right),\end{aligned}$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*

$C(d,\kappa) = d + \kappa$ *for the Cholesky and matrix square root,*

$\bar{\zeta}_{\mathrm{KL}}$, $\bar{\zeta}_{\mathrm{H}}$ are the stationary points of f_{KL}, f_{H}, respectively, $F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, and $f_{\mathrm{KL}}^* = \inf_{\zeta\in\mathbb{R}^d} f(\zeta)$.

Proof Sketch. From Lemma 4, we can see that the key quantity of upper bounding the gradient variance is to analyze $\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|$. The bird's eye view of the proof is as follows:

❶ The relationship between $\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2$ and $\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2$ is established through Lemma 2.

❷ Then, the L_{H}-smoothness of f_{H} relates $\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2$ with $\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2$, the average squared distance from f_{H}'s stationary point.

❸ The average squared distance enables the simplification of stochastic terms through Lemmas 3 and 7. This step also introduces dimension dependence.

From here, we are now left with the $\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2$ term. One might be tempted to assume the quadratic growth assumption on f_{H} and proceed as

$$\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 \leq \frac{2}{\mu}\left(f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f_{\mathrm{H}}^*\right).$$

However, for the entropy-regularized form, this soon runs into a dead end since in

$$\begin{aligned}\mathbb{E} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f_{\mathrm{H}}^* &= F(\boldsymbol{\lambda}) - h(\boldsymbol{\lambda}) - f^* \\ &= (F(\boldsymbol{\lambda}) - F^*) + (F^* - f^*) - h_{\mathrm{H}}(\boldsymbol{\lambda}),\end{aligned}$$

the negative entropy term h_{H} is not bounded unless we rely on assumptions that need modifications to the BBVI algorithms. (*e.g.*, bounded support, bounded domain). Fortunately, the following inequality cleverly side-steps this problem:

$$\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 \leq 2\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{KL}}\|_2^2 + 2\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2, \tag{6}$$

albeit at the cost of some looseness. By converting the entropy-regularized form into the KL-regularized form, the regularizer term becomes $h_{\mathrm{KL}} = \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\lambda}}, p) \geq 0$, which is bounded below by definition, unlike the entropic-regularizer h_{H}. The proof completes by

❹ applying the quadratic growth assumption to relate the parameter distance with the function suboptimality gap, and

❺ upper bounding the KL regularizer term.

Proof. See the *full proof* in Appendix C.3.1.

Remark 5. If the bijector ψ is an identity function, ζ_{KL} and ζ_{H} are the maximum likelihood (ML) and maximum a-posteriori (MAP) estimates, respectively. Thus, with enough datapoints, the term $\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2$ will be negligible since the ML and MAP estimates will be close.

Remark 6. It is also possible the tighten the constants by a factor of two. Instead of applying Equation (6), we can use the inequality

$$(a+b)^2 \leq \left(1+\delta^2\right)a^2 + \left(1+\delta^{-2}\right)b^2,$$

for some $\delta > 0$. By setting $\delta^2 = b = \|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2$,

$$\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 \leq (1+\delta^2)\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{KL}}\|_2^2 + \delta^4 + \delta^2.$$

Since $\delta \approx 0$ as explained in Remark 5, the constant in front of the first term is tightened almost by a factor of 2. However, the stated form is more convenient for theory since the first term does not depend on $\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2$.

Remark 7. Let $\kappa_{\mathrm{cond.}} = {}^{L_{\mathrm{H}}}/_{\mu_{\mathrm{KL}}}$ be the *condition number* of the problem. For the full-rank parameterizations, the variance is bounded as $\mathcal{O}\left(L_{\mathrm{H}}\kappa_{\mathrm{cond.}}(d+\kappa)/M\right)$. The variance depends linearly on

❶ the scaling of the problem L_{H},
❷ the conditioning of the problem $\kappa_{\mathrm{cond.}}$,
❸ the dimensionality of the problem d, and
❹ the tail properties of the variational family κ,

where the number of Monte Carlo samples M linearly reduces the variance.

KL-Regularized Form We now prove an equivalent result for the KL-regularized form. Here, we do not have to rely on Equation (6) since we already start from f_{KL}, which results in better constants.

Theorem 2. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in KL-regularized form (Definition 6). Also, assume that*

- *f_{KL} is L_{KL}-smooth,*
- *f_{KL} is μ_{KL}-quadratically growing,*

and Assumption 3 and 4 hold. Then, the gradient vari-

ance is bounded above as

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)(F(\boldsymbol{\lambda}) - F^*) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)\left(F^* - f_{\mathrm{KL}}^*\right),$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*

$C(d,\kappa) = d + \kappa$ *for the Cholesky and matrix square root,*

$F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, *and* $f_{\mathrm{KL}}^* = \inf_{\boldsymbol{\zeta}\in\mathbb{R}^d} f(\boldsymbol{\zeta})$.

Proof. See the *full proof* in Appendix C.3.2.

3.3. Upper Bound Under Bounded Entropy

The bound in Theorem 1 is slightly loose due to the use of Equation (6) and Equation (29). An alternative bound with slightly tighter constants, although the gains are marignal compared to Remark 6, can be obtained by assuming the following:

Assumption 5 (Bounded Entropy). The regularization term is bounded below as $h_{\mathrm{H}}(\boldsymbol{\lambda}) \geq h_{\mathrm{H}}^*$.

For the entropy-regularized form, this corresponds to the entropy being bounded above by some constant since $h(\boldsymbol{\lambda}) = -\mathrm{H}(q_{\boldsymbol{\lambda}})$. When using the nonlinear parameterizations (Definitions 8 and 9), this assumption can be practically enforced by bounding the output of ϕ by some large S.

Proposition 1. *Let the diagonal conditioner ϕ be bounded as $\phi(x) \leq S$. Then, for any d-dimensional distribution $q_{\boldsymbol{\lambda}}$ in the location-scale family with the mean-field (Definition 8) or Cholesky (Definition 9) parameterizations,*

$$h_{\mathrm{H}}(\boldsymbol{\lambda}) = -\mathrm{H}(q_{\boldsymbol{\lambda}}) \geq -\mathrm{H}(\varphi) - \frac{d}{2}\log S.$$

Proof. From Remark 1, $\mathrm{H}(q_{\boldsymbol{\lambda}}) = \mathrm{H}(\varphi) + \log|\boldsymbol{C}|$. Since $\boldsymbol{C}$ under Definitions 8 and 9 is a diagonal or triangular matrix, the log absolute determinant is the log sum of the diagonals. The conclusion follows from the fact that the diagonals $C_{ii} = \phi(s_i)$ are bounded by S. □

This is essentially a weaker version of the bounded domain assumption, though only the diagonal elements of $\boldsymbol{C}$, $s_1, \ldots, s_d$, are bounded. While this assumption results in an admittedly less realistic algorithm, it enables a tighter bound for the entropy-regularized form ELBO.

Theorem 3. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in entropy-regularized form (Definition 5). Also, assume that*

- *f_{H} is L_{H}-smooth,*
- *f_{H} is μ_{H}-quadratically growing,*
- *h_{H} is bounded as $h_{\mathrm{H}}(\boldsymbol{\lambda}) > h_{\mathrm{H}}^*$ (Assumption 5),*

and Assumption 3 and 4 hold. Then, the gradient variance of $\boldsymbol{g}_M$ is bounded above as

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}}M} C(d,\kappa)(F(\boldsymbol{\lambda}) - F^*) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}}M} C(d,\kappa)\left(F^* - f_{\mathrm{H}}^* - h_{\mathrm{H}}^*\right),$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*

$C(d,\kappa) = d + \kappa$ *for the Cholesky parameterization,*

$F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, *and* $f_{\mathrm{H}}^* = \inf_{\boldsymbol{\zeta}\in\mathbb{R}^d} f(\boldsymbol{\zeta})$.

Proof Sketch. Instead of using Equation (6), we apply the quadratic assumption directly to f_{H}. The remaining entropic-regularizer term can now be bounded through the bounded entropy assumption.

Proof. See the *full proof* in Appendix C.3.3.

3.4. Matching Lower Bound

Finally, we present a matching lower bound on the gradient variance of BBVI. Our lower bound holds broadly for smooth and strongly convex problem instances that are well-conditioned and high-dimensional.

Theorem 4. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in either the entropy- or KL-regularized form. Also, let Assumption 4 hold where the matrix square root parameterization is used. Then, for all L-smooth and μ-strongly convex functions f such that $L/\mu < \sqrt{d+1}$, the variance of $\boldsymbol{g}_M$ is bounded below by some strictly positive constant as*

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \geq \frac{2\mu^2(d+1) - 2L^2}{ML}(F(\boldsymbol{\lambda}) - F^*) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2\mu^2(d+1) - 2L^2}{ML}\left(\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - f^*\right),$$

as long as $\boldsymbol{\lambda}$ is in a local neighborhood around the unique global optimum $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, where $F^* = F(\boldsymbol{\lambda}^*)$ and $f^* = \arg\min_{\boldsymbol{\zeta}\in\mathbb{R}^d} f(\boldsymbol{\zeta})$.

Proof Sketch. We use the fact that, with the matrix square root parameterization, if f is L-smooth, $\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))$ is also L-smooth (Domke, 2020). From this, the parameter suboptimality can be related to the function suboptimality as

$$\|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2^2 \geq (2/L)\left(\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f^*\right),$$

where $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{\zeta}}, \mathbf{0})$. For the entropy term, we circumvent the need to directly bound its value by restricting our interest to the neighborhood of the minimizer $\boldsymbol{\lambda}^*$, where the contribution of $h(\boldsymbol{\lambda}^*) - h(\boldsymbol{\lambda})$ will be marginal enough for the lower bound to hold.

Proof. See the *full proof* in Appendix C.3.4.

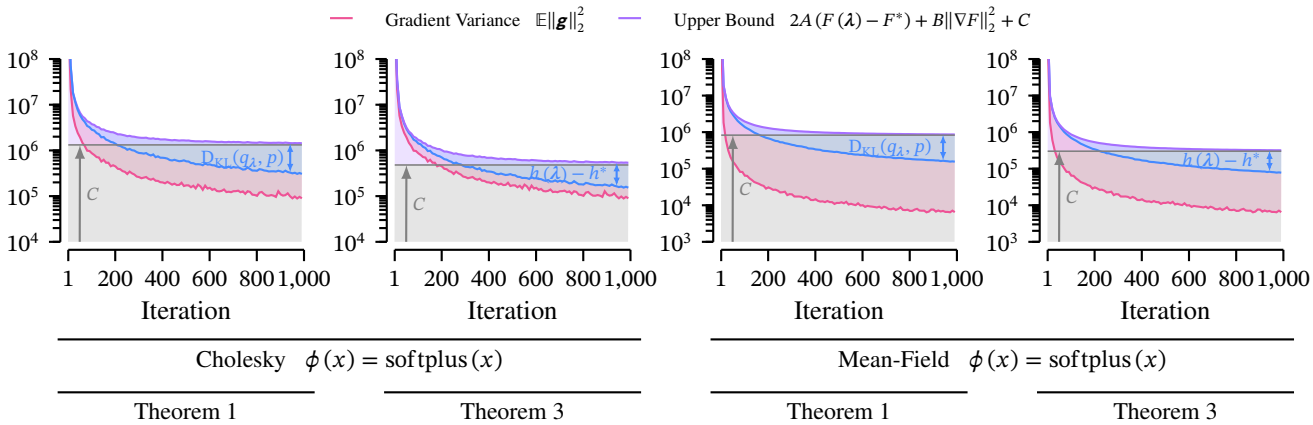

Figure 1: **Evaluation of the bounds for a perfectly conditioned quadratic target function.** The blue regions are the loosenesses resulting from either using (Theorem 1) or not using (Theorem 3) the bounded entropy assumption (Assumption 5), while the red regions are the remaining "technical loosesnesses." The gradient variance was estimated from 10^3 samples.

Remark 8 (Matching Dimensional Dependence). For well-conditioned problems such that $L/\mu < \sqrt{d+1}$, a lower bound of the same dimensional dependence with our upper bounds holds near the optimum.

Remark 9 (Unimprovability of the ABC Condition). The lower bound suggests that the *ABC* gradient variance condition is unimprovable within the class of smooth, quadratically growing functions.

4. Simulations

We now evaluate our bounds and the insights gathered during the analysis through simulations. We implemented a bare-bones implementation of BBVI in Julia (Bezanson et al., 2017) with plain SGD. The stepsize were manually tuned so that all problems converge at similar speeds. For all problems, we use a unit Gaussian base distribution such that $\varphi(u) = \mathcal{N}(u; 0, 1)$ resulting in a kurtosis of $\kappa = 3$ and use $M = 10$ Monte Carlo samples.

4.1. Synthetic Problem

To test the *ideal* tightness of the bounds, we consider quadratics achieving the tightest bound for the constants $L_{\mathrm{H}}, L_{\mathrm{KL}}, \mu_{\mathrm{H}}, \mu_{\mathrm{KL}}$ given as

$$\log \ell(\boldsymbol{x} \mid \boldsymbol{z}) = -\frac{N}{\sigma^2}\|\boldsymbol{z} - \boldsymbol{z}^*\|_2^2; \quad \log p(\boldsymbol{z}) = -\frac{1}{\lambda}\|\boldsymbol{z}\|_2^2,$$

where N simulates the effect of the number of datapoints. We set the constants as $\sigma = 0.3$, $\lambda = 8.0$, and $N = 100$, the mode $\boldsymbol{z}^*$ is randomly sampled from a Gaussian, and the dimension of the problem is $d = 20$. For the bounded entropy case, we set $S = 2.0$ (the true standard deviation is in the order of 1e-3).

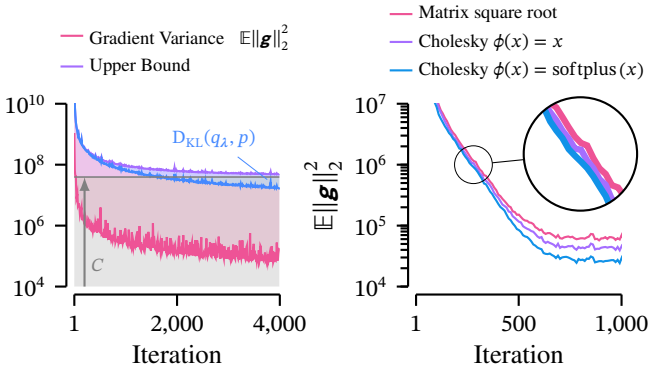

Figure 2: **Linear regression on the AIRFOIL dataset. (left) Evaluation of the upper bound (Theorem 1). (right) Comparison of the variance of different parameterizations resulting in the same $\boldsymbol{m}, \boldsymbol{C}$.**

Quality of Upper Bound The results for the Cholesky and mean-field parameterizations with a softplus bijector are shown in Figure 1. For the Cholesky parameterization, the bulk of the looseness comes from the treatment of the regularization term (blue region). The remaining "technical looseness" (red region) is relatively tight and can be shown to be tighter when using linear parameterizations ($\phi(x) = x$) and the square root parameterization, which is the tightest. However, for the mean-field parameterization, despite the superior constants (Remark 4), there is still room for improvement. Additional results for other parameterizations can be found in Appendix B.1.

4.2. Real Dataset

Model We now evaluate the theoretical results with real datasets. Given a regression dataset $(\boldsymbol{X}, \boldsymbol{y})$, we use the linear Gaussian model

$$y \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2); \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \lambda \mathbf{I}),$$

where λ and σ are hyperparameters. The smoothness and quadratic growth constants for this model are given as the max- and minimum eigenvalues of $\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{-1}\mathbf{I}$ (for f_{H}) and $\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X}$ (for f_{KL}). f^*_{KL} and f^*_{H} are given as the mode of the likelihood and the posterior, while F^* is the negative marginal log-likelihood.

Quality of Upper Bound Section 4.1 shows the result on the AIRFOIL dataset (Dua & Graff, 2017). The constants are $L_{\mathrm{H}} = 3.520 \times 10^4, \mu_{\mathrm{KL}} = 2.909 \times 10^3$. Due to poor conditioning, the bound is much looser compared to the quadratic case. We note that generalizing our bounds to utilize matrix smoothness and matrix-quadratic growth as done by (Domke, 2019) would tighten the bounds. But the theoretical gains would be marginal. Detailed information about the datasets and additional results for other parameterizations can be found in Appendix B.2.

Comparison of Parameterizations Section 4.1 compares the gradient variance resulting from the different parameterizations. For a fair comparison, the gradient is estimated on the $\boldsymbol{\lambda}$ that results in the same $\boldsymbol{m}, \boldsymbol{C}$ for all three parameterizations. This shows the gradual increase in variance by **(i)** not using a nonlinear conditioner (linear Cholesky) **(ii)** and increasing the number of variational parameters (matrix square root).

5. Related Works

Controlling Gradient Variance The main algorithmic challenge in BBVI is to control the gradient noise (Ranganath et al., 2014). This has led to various methods for reducing the variance of VI gradient estimators using control variates (Ranganath et al., 2014; Miller et al., 2017; Geffner & Domke, 2018), ensembling of estimators (Geffner & Domke, 2020), modifying the differentiation procedure (Roeder et al., 2017), quasi-Monte Carlo (Buchholz et al., 2018; Liu & Owen, 2021), and multilevel Monte Carlo (Fujisawa & Sato, 2021). Cultivating a deeper understanding of the properties of gradient variance could further extend this list.

Convergence Guarantees Obtaining full convergence guarantees has been an important task for understanding BBVI algorithms. However, most guarantees so far have relied on strong assumptions such as that the log-likelihood is Lipschitz (Chérief-Abdellatif et al., 2019; Alquier, 2021), that the gradient variance is bounded by constant (Liu & Owen, 2021; Buchholz et al., 2018; Domke, 2020; Hoffman & Ma, 2020), and that the support of q_λ is bounded (Fujisawa & Sato, 2021). Our result shows that similar results can be obtained under relaxed assumptions. Meanwhile, Bhatia et al. (2022) have recently proven a full complexity guarantee for a variant of BBVI. But similarly to Hoffman & Ma (2020), they only optimize the scale matrix $\boldsymbol{C}$, and the specifics of the algorithm diverge from the usual BBVI implementations as it uses the stochastic power iterations instead of SGD.

Gradient Variance Guarantees Studying the actual gradient variance properties of BBVI has only started to make progress recently. Fan et al. (2015) first provided bounds by assuming the log-likelihood to be Lipschitz. Under more general conditions, Domke (2019) provided tight bounds for smooth log-likelihoods, which our work builds upon. Domke's result can also be seen as a direct generalization of the results of Xu et al. (2019), which are restricted to quadratic log-likelihoods and the mean-field family. Lastly, Mohamed et al. (2020a) provides a conceptual evaluation of gradient estimators used in BBVI.

6. Discussions

Conclusions In this work, we have proven upper bounds on the gradient variance of BBVI with the location-scale family for smooth, quadratically-growing log-likelihoods. Specifically, we have provided bounds for both the ELBO in entropy-regularized and KL-regularized forms. Our guarantees work without a single modification to the algorithms used in practice, although stronger assumptions establish a tighter bound for the entropy-regularized form ELBO. Also, our bounds corresponds to the ABC condition (Section 2.4) and the *expected residual* (ER) condition, where the latter is a special case of the former with $B = 1$. The ER condition has been used by Gower et al. (2021a) for proving convergence of SGD on quasar convex functions, which generalize convex functions. The results of this paper are used by Kim et al. (2023) to establish convergence of BBVI through the results of Khaled & Richtárik (2023).

Limitations Our results have the following limitations: ❶ Our results only apply to smooth and quadratically–growing log likelihoods and ❷ the location-scale ADVI family. Also, ❸ our bounds cannot distinguish the variance of the Cholesky and matrix square root parameterizations, ❹ and empirically, the bounds for the mean-field parameterization appear loose. Furthermore, ❺ our results only work with 1-Lipschitz diagonal conditioners such as the softplus function. Unfortunately, assuming both smoothness and quadratic growth is quite restrictive, as it leaves a very small number of known distributions. Also, in practice, non-Lipschitz conditioners such as the exponential functions are widely used. While obtaining similar bounds with such conditioners would be challenging, constructing a theoretical framework that extends to such would be an important future research direction.

Acknowledgements

This work was supported by NSF award IIS-2145644.

References

Alquier, P. Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions. In *Proceedings of the International Conference on Machine Learning*, volume 193 of *PMLR*, pp. 207–218. ML Research Press, July 2021. (page 9)

Anitescu, M. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, January 2000. (page 5)

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. (page 8)

Bhatia, K., Kuang, N. L., Ma, Y.-A., and Wang, Y. Statistical and computational trade-offs in variational inference: A case study in inferential model selection. (arXiv:2207.11208), July 2022. (page 9)

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019. (pages 1, 4)

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. (page 1)

Bottou, L. On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, pp. 9–42. Cambridge University Press, first edition, January 1999. (page 1)

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, January 2018. (pages 1, 3)

Buchholz, A., Wenzel, F., and Mandt, S. Quasi-Monte Carlo variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 668–677. ML Research Press, July 2018. (page 9)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. (pages 1, 4)

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. E. A generalization bound for online variational inference. In *Proceedings of the Asian Conference on Machine Learning*, volume 101 of *PMLR*, pp. 662–677. ML Research Press, October 2019. (page 9)

Dhaka, A. K., Catalina, A., Andersen, M. R., ns Magnusson, M., Huggins, J., and Vehtari, A. Robust, accurate stochastic optimization for variational inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10961–10973. Curran Associates, Inc., 2020. (page 1)

Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J., and Vehtari, A. Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7787–7798. Curran Associates, Inc., 2021. (page 1)

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. TensorFlow distributions. (arXiv:1711.10604), November 2017. (pages 1, 2, 4)

Domke, J. Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. (pages 1, 2, 4, 5, 9, 13, 16, 17, 18, 21)

Domke, J. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 2587–2596. ML Research Press, July 2020. (pages 1, 4, 7, 9, 17, 23)

Dua, D. and Graff, C. UCI machine learning repository. 2017. (page 9)

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. (page 4)

Fan, K., Wang, Z., Beck, J., Kwok, J., and Heller, K. A. Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. (page 9)

Fjelde, T. E., Xu, K., Tarek, M., Yalburgi, S., and Ge, H. Bijectors.jl: Flexible transformations for probability distributions. In *Proceedings of The Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *PMLR*, pp. 1–17. ML Research Press, February 2020. (pages 1, 2)

Fujisawa, M. and Sato, I. Multilevel Monte Carlo variational inference. *Journal of Machine Learning Research*, 22(278):1–44, 2021. (pages 1, 9)

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. (page 4)

Ge, H., Xu, K., and Ghahramani, Z. Turing: A language for flexible probabilistic inference. In *Proceedings of the International Conference on Machine Learning*, volume 84 of *PMLR*, pp. 1682–1690. ML Research Press, 2018. (pages 1, 4)

Geffner, T. and Domke, J. Using large ensembles of control variates for variational inference. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. (page 9)

Geffner, T. and Domke, J. A rule for gradient estimator selection, with an application to variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pp. 1803–1812. ML Research Press, August 2020. (page 9)

Gower, R., Sebbouh, O., and Loizou, N. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of The International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 1315–1323. ML Research Press, March 2021a. (pages 4, 9)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5200–5209. ML Research Press, June 2019. (pages 1, 3)

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Mathematical Programming*, 188(1): 135–192, July 2021b. (pages 1, 2, 3)

Hinder, O., Sidford, A., and Sohoni, N. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of Conference on Learning Theory*, volume 125 of *PMLR*, pp. 1894–1938. ML Research Press, July 2020. (page 5)

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Annual Conference on Computational Learning Theory*, pp. 5–13, Santa Cruz, California, United States, 1993. ACM Press. (page 2)

Hoffman, M. and Ma, Y. Black-box variational inference as a parametric approximation to Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, PMLR, pp. 4324–4341. ML Research Press, November 2020. (page 9)

Jin, J. On the convergence of first order methods for quasar-convex optimization. (arXiv:2010.04937), October 2020. (page 5)

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. (pages 1, 2)

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 795–811, Cham, 2016. Springer International Publishing. (page 5)

Khaled, A. and Richtárik, P. Better theory for SGD in the nonconvex world. *Transactions of Machine Learning Research*, 2023. (pages 2, 3, 4, 9)

Kim, K., Wu, K., Oh, J., Ma, Y., and Gardner, J. R. Black-box variational inference converges. (arXiv:2305.15349), May 2023. (pages 2, 9)

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*, San Diego, California, USA, 2015. (page 1)

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14): 1–45, 2017. (pages 1, 2, 3)

Leger, J.-B. Parametrization cookbook: A set of bijective parametrizations for using machine learning methods in statistical inference. (arXiv:2301.08297), January 2023. (page 2)

Li, X. and Milzarek, A. A unified convergence theorem for stochastic optimization methods. In *Advances in Neural Information Processing Systems*, October 2022. (page 4)

Liu, J. and Yuan, Y. On almost sure convergence rates of stochastic gradient methods. In *Proceedings of the Conference on Learning Theory*, volume 178 of *PMLR*, pp. 2963–2983. ML Research Press, June 2022. (page 4)

Liu, S. and Owen, A. B. Quasi-Monte Carlo quasi-Newton in Variational Bayes. *Journal of Machine Learning Research*, 22(243):1–23, 2021. (page 9)

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern

over-parametrized learning. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3325–3334. ML Research Press, July 2018. (page 1)

Miller, A., Foti, N., D' Amour, A., and Adams, R. P. Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. (page 9)

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020a. (pages 1, 9)

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020b. (page 3)

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! Convergence without the bounded gradients assumption. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3750–3758. ML Research Press, July 2018. (pages 1, 3)

Peterson, C. and Anderson, J. R. A mean field theory learning algorithm for Neural Networks. *Complex Systems*, 1 (5):995–1019, 1987. (page 2)

Peterson, C. and Hartman, E. Explorations of the mean field theory learning algorithm. *Neural Networks*, 2(6): 475–494, January 1989. (page 2)

Polyak, B. T. and Tsypkin, Y. Z. Pseudogradient adaptation and training algorithms. *Automatic Remote Control*, 34 (3):45–68, 1973. (pages 2, 3)

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 33 of *PMLR*, pp. 814–822. ML Research Press, April 2014. (pages 1, 9)

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, September 1951. (page 1)

Roeder, G., Wu, Y., and Duvenaud, D. K. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. (page 9)

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016. (pages 1, 4)

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. (arXiv:1308.6370), August 2013. (pages 1, 3)

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pp. 1971–1979. ML Research Press, June 2014. (page 1)

Tseng, P. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, May 1998. (pages 1, 3)

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pp. 1195–1204. ML Research Press, April 2019. (pages 1, 3)

Welandawe, M., Andersen, M. R., Vehtari, A., and Huggins, J. H. Robust, automated, and accurate black-box variational inference. (arXiv:2203.15945), March 2022. (page 1)

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. Variance reduction properties of the reparameterization trick. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pp. 2711–2720. ML Research Press, April 2019. (pages 1, 3, 5, 9)

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the International Conference on Machine Learning*, PMLR, pp. 5581–5590. ML Research Press, July 2018. (page 1)

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, August 2019. (page 1)

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.-Q. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, 2022. (page 1)

TABLE OF CONTENTS

1 Introduction **1**

2 Preliminaries **2**
2.1 Variational Inference 2
2.2 Variational Family 2
2.3 Reparameterization Trick 3
2.4 Gradient Variance Assumptions in Stochastic Gradient Descent 3
2.5 Covariance Parameterizations 4

3 Main Results **4**
3.1 Key Lemmas 4
3.2 Upper Bounds 5
3.3 Upper Bound Under Bounded Entropy 7
3.4 Matching Lower Bound 7

4 Simulations **8**
4.1 Synthetic Problem 8
4.2 Real Dataset 8

5 Related Works **9**

6 Discussions **9**

A Detailed Comparison Against Domke 2019 **13**

B Additional Simulation Results **14**
B.1 Synthetic Problem 14
B.2 Real Datasets 14

C Proofs **16**
C.1 External Lemmas 16
C.2 Proof of Key Lemmas 17
C.2.1 Proof of Lemma 1 17
C.2.2 Proof of Lemma 2 18
C.2.3 Proof of Lemma 3 19
C.2.4 Proof of Lemma 4 20
C.3 Proof of Theorems 21
C.3.1 Proof of Theorem 1 21
C.3.2 Proof of Theorem 2 22
C.3.3 Proof of Theorem 3 23
C.3.4 Proof of Theorem 4 23

A. Detailed Comparison Against Domke 2019

Under the assumption that f_{H} is L_{H}-smooth and the linear full-rank Cholesky parameterization, under our notation, (Domke, 2019) prove the following bound:

$$\mathbb{E}\|\boldsymbol{g}_{M=1}\|_2^2 \leq L_{\mathrm{H}}^2 \left((d+1)\|\boldsymbol{m} - \bar{\boldsymbol{\zeta}}_{\mathrm{H}}\|_2^2 + (d+\kappa)\|\boldsymbol{C}\|_{\mathrm{F}}^2 \right).$$

We extend Domke's analysis in three original directions.

❶ Generalization to Nonlinear Parameterizations First, we generalize the bounds to support nonlinear parameterizations. In particular, Lemma 1 and Lemma 2 generalize Lemma 1 of Domke (2019) to 1-Lipschitz nonlinear conditioners. From here, the analysis becomes identical to Domke's setup, until we reach our original analysis we discuss in Item ❸.

❷ Tighter Bound for the Mean-Field Parameterization Second, for the mean-field parameterization, we prove a bound that is tighter in the large d regime,

$$\mathbb{E}\|\boldsymbol{g}_{M=1}\|_2^2 \leq L_{\mathrm{H}}^2 \left((\sqrt{d\kappa} + \kappa\sqrt{d} + 1)\|\boldsymbol{m} - \bar{\boldsymbol{\zeta}}_{\mathrm{H}}\|_2^2 + (2\kappa\sqrt{d} + 1)\|\boldsymbol{C}\|_{\mathrm{F}}^2 \right),$$

as a direct consequence of Lemma 5.

❸ Connecting with the ABC Condition Furthermore, we extend the bounds above and establish the ABC condition (Assumption 2) for the ELBO, through the quadratic function growth condition (Definition 11). Specifically, in our proof of Theorem 1, the derivation past Equation (27) is original.

B. Additional Simulation Results

B.1. Synthetic Problem

We provide additional results for the simulations with quadratics in Section 4.1.

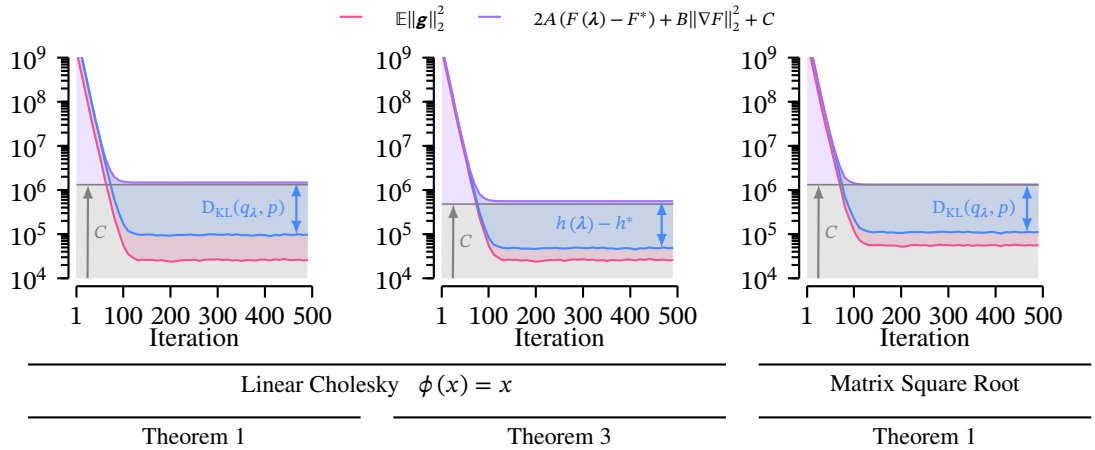

Figure 3: **Evaluation of the bounds for a perfectly conditioned quadratic target.** The blue regions are the loosenesses resulting from either using (Theorem 1) or not using (Theorem 3) the bounded entropy assumption (Assumption 5), while the red regions are the remaining "technical loosesnesses." The gradient variance was estimated from 10^3 samples.

B.2. Real Datasets

We provide detailed information and additional results for the linear regression problem in Section 4.2. The constants for the linear regression datasets are shown in Table 2, while additional results for the nonlinear Cholesky (Figure 4), linear Cholesky (Figure 5), nonlinear mean-field (Figure 6), and matrix square root (Figure 7) parameterizations are displayed.

Table 2: Properties of the Linear Regression Datasets

Dataset	d	N	L_{H}	μ_{KL}	$\kappa_{\mathrm{cond.}}$	$\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2$	**Constants for Theorem 1**	
							A	C
FERTILITY	9	100	1.840×10^3	5.017×10^2	4	5.167×10^{-9}	1.620×10^4	1.313×10^6
PENDULUM	9	630	1.525×10^4	1.897×10^3	8	1.243×10^{-10}	2.942×10^5	2.858×10^7
AIRFOIL	5	1,503	3.520×10^4	2.909×10^3	12	2.937×10^{-10}	6.815×10^5	3.936×10^7
WINE	11	1,599	5.526×10^4	1.786×10^3	31	6.628×10^{-9}	4.787×10^6	6.054×10^8

* N is the number of datapoints in the dataset, $\kappa_{\mathrm{cond.}} = L_{\mathrm{H}}/\mu_{\mathrm{KL}}$ is the condition number.

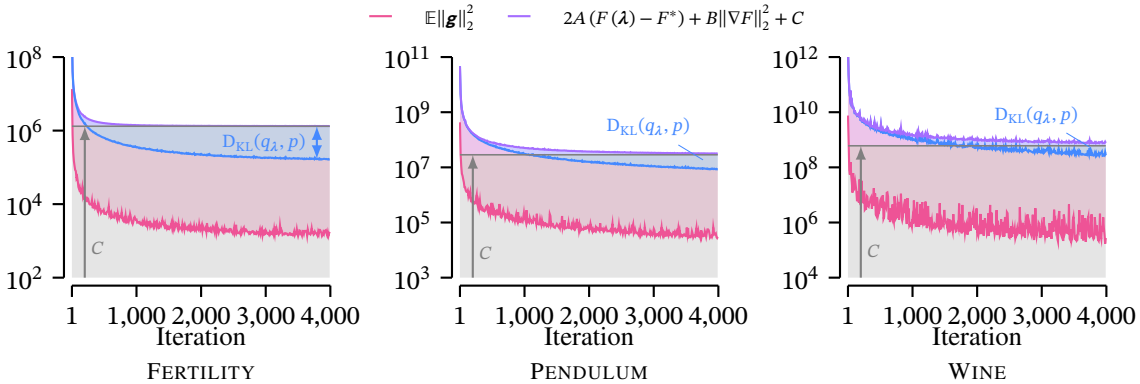

Figure 4: **Evaluation of Theorem 1 with the nonlinear Cholesky ($\phi(x) = \mathrm{softplus}(x)$) parameterization on linear regression datasets.** The gradient variance was estimated from 4×10^3 samples.

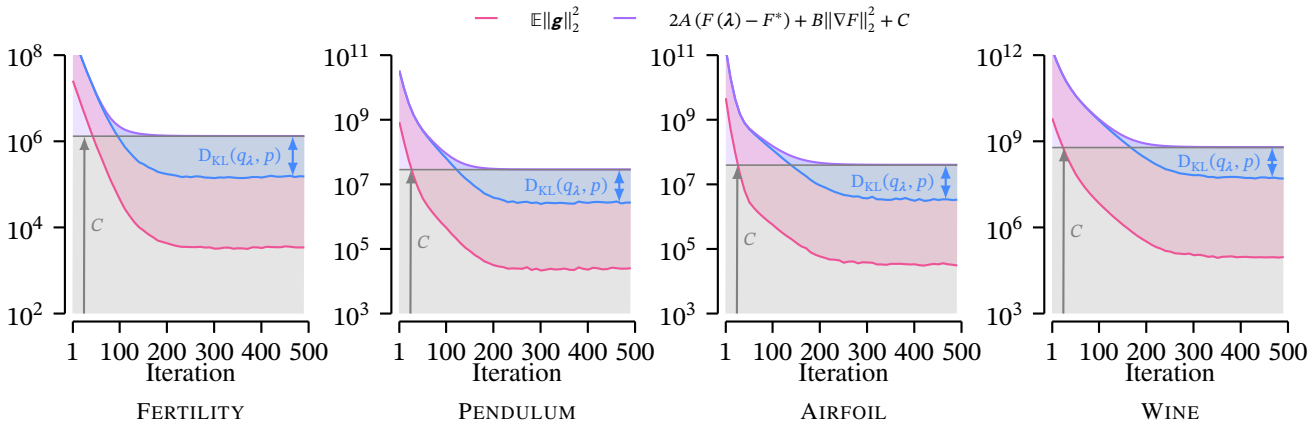

Figure 5: **Evaluation of Theorem 1 with the linear Cholesky ($\phi(x) = x$) parameterization on linear regression datasets.** The gradient variance was estimated from 4×10^3 samples.

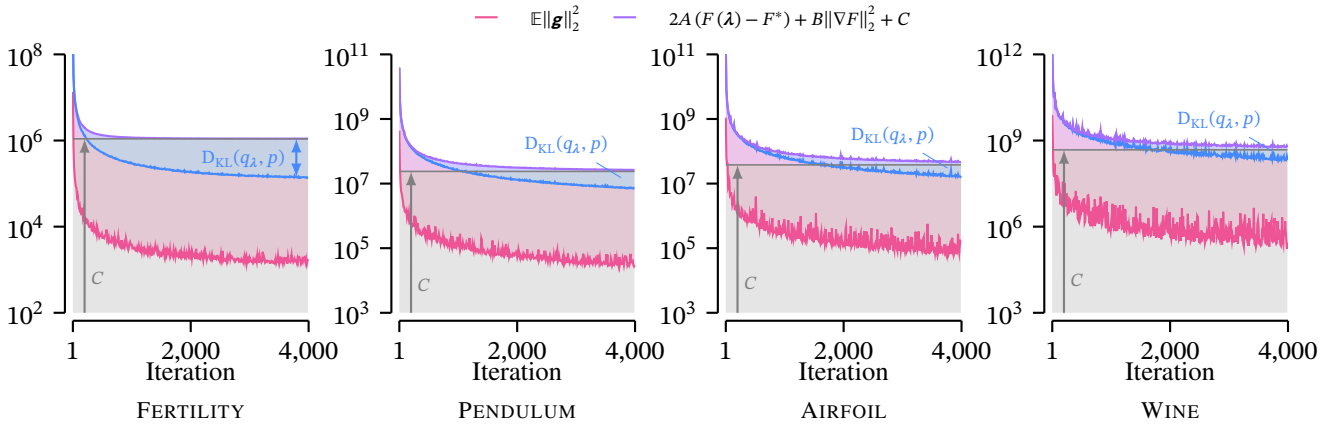

Figure 6: **Evaluation of Theorem 1 with the nonlinear mean-field ($\phi(x) = \mathrm{softplus}(x)$) parameterization on linear regression datasets.** The gradient variance was estimated from 4×10^3 samples.

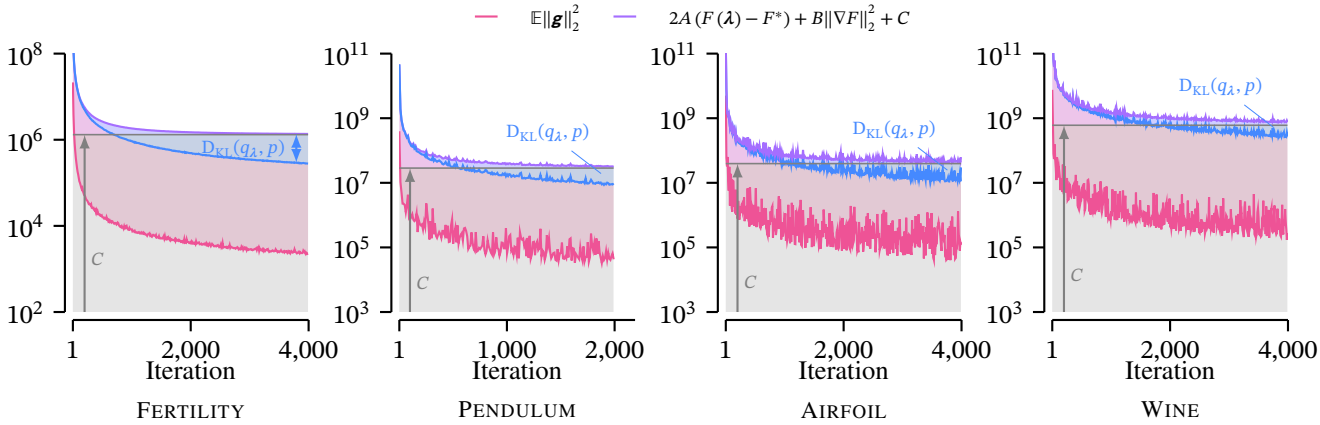

Figure 7: **Evaluation of Theorem 1 matrix square root parameterization on linear regression datasets.** The gradient variance was estimated from 4×10^3 samples.

C. Proofs

C.1. External Lemmas

Lemma 5 (Domke 2019, Lemma 9)**.** *Let* $\boldsymbol{u} = (u_1, u_2, \dots, u_d)$ *be a* d*-dimensional vector-valued random variable with zero-mean independently and identically distributed components. Then,*

$$\begin{aligned}
\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top &= \left(\mathbb{E}u_i^2\right)\mathbf{I} \\
\mathbb{E}\|\boldsymbol{u}\|_2^2 &= d\,\mathbb{E}u_i^2 \\
\mathbb{E}\boldsymbol{u}\left(1 + \|\boldsymbol{u}\|_2^2\right) &= \left(\mathbb{E}u_i^3\right)\mathbf{1} \\
\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{u}\boldsymbol{u}^\top &= \left((d-1)\left(\mathbb{E}u_i^2\right)^2 + \mathbb{E}u_i^4\right)\mathbf{I}.
\end{aligned}$$

Lemma 6 (Domke 2019, Lemma 1)**.** *Let* $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ *be a location-scale reparameterization function (Definition 1). Also, let* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *be some differentiable function. Then,*

$$\begin{aligned}
&\|\nabla_\lambda f\left(\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right)\right)\|_2^2 \\
&= \|\nabla f\left(\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right)\right)\|_2^2\left(1 + \|\boldsymbol{u}\|_2^2\right).
\end{aligned}$$

Lemma 7 (Domke 2019, Lemma 1)**.** *Let* $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ *be a location-scale reparameterizaiton function (Definition 1). Also, let* $\boldsymbol{z} \in \mathbb{R}^d$ *be some vector and* $\boldsymbol{u} \sim \varphi$ *satisfy Assumption 1. Then,*

$$\mathbb{E}\|\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right) - \boldsymbol{z}\|_2^2\left(1 + \|\boldsymbol{u}\|_2^2\right) = (d+1)\,\|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + (d+\kappa)\,\|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

Lemma 8. *Let* $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ *be a location-scale reparameterizaiton function (Definition 1). Also, let* $\boldsymbol{z} \in \mathbb{R}^d$ *be some vector, and let* $\boldsymbol{u} \sim \varphi$ *satisfy Assumption 1. Then,*

$$\mathbb{E}\|\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right) - \boldsymbol{z}\|_2^2 = \|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right) - \boldsymbol{z}\|_2^2 &= \mathbb{E}\|\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\|_2^2 \\
&= \mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{u} + 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{m} - 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{z} \\
&\quad + \boldsymbol{m}^\top\boldsymbol{m} - 2\,\boldsymbol{m}^\top\boldsymbol{z} + \boldsymbol{z}^\top\boldsymbol{z}. \qquad (7)
\end{aligned}$$

The first three terms follow as

$$\begin{aligned}
&\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{u} + 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{m} - 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{z} \\
&= \mathbb{E}\mathrm{tr}\left(\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{u}\right) + 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{m} - 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{z} \\
&= \mathrm{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top\right) + 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{m} - 2\,\mathbb{E}\boldsymbol{u}^\top\boldsymbol{C}^\top\boldsymbol{z},
\end{aligned}$$

applying Lemma 5,

$$\begin{aligned}
&= \mathrm{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\right) \\
&= \|\boldsymbol{C}\|_{\mathrm{F}}^2.
\end{aligned}$$

Applying this to Equation (7),

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{t}_\lambda\left(\boldsymbol{u}\right) - \boldsymbol{z}\|_2^2 &= \boldsymbol{m}^\top\boldsymbol{m} - 2\,\boldsymbol{m}^\top\boldsymbol{z} + \boldsymbol{z}^\top\boldsymbol{z} + \|\boldsymbol{C}\|_{\mathrm{F}}^2 \\
&= \|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

C.2. Proof of Key Lemmas

C.2.1. Proof of Lemma 1

Lemma 1. *Let $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ be a location-scale reparameterization function (Definition 1) with some differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. Then, for $\boldsymbol{g}_f \triangleq \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))$,*

(i) Mean-Field

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{g}_f,$$

(ii) Cholesky

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{\Sigma}\boldsymbol{g}_f + \boldsymbol{g}_f^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I})\boldsymbol{g}_f,$$

where $\boldsymbol{U}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}$ are diagonal matrices, which the diagonals are defined as

$$U_{ii} = u_i^2, \quad \Phi_{ii} = \phi'(s_i)^2, \quad \Sigma_{ii} = \textstyle\sum_{j=1}^i u_j^2,$$

and ϕ is a diagonal conditioner for the scale matrix.

Proof. The proof starts by applying the Chain Rule and then computing the quadratic norm of the gradient as

$$\begin{aligned}
&\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \\
&= \left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u})) \\
&= \nabla f^\top(\boldsymbol{t}_\lambda(\boldsymbol{u}))\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u})) \\
&= \boldsymbol{g}_f^\top \left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \boldsymbol{g}_f. \qquad (8)
\end{aligned}$$

Naturally, the derivative of the reparameterization function will depend on the specific parameterization used.

Proof for Cholesky Let p denote the number of scalar variational parameters such that $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)$. Then,

$$\begin{aligned}
&\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \\
&= \sum_{i=1}^d \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial m_i}\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial m_i}\right)^\top + \sum_{i=1}^d \sum_{j \le i}^d \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\right)^\top,
\end{aligned}$$

where $\lambda_{C_{ij}}$ denote the parameter responsible for the ij-th entry of $\boldsymbol{C}$, C_{ij}. Notice that, unlike for the matrix square root parameterization (Domke, 2019), the sum for C_{ij} is only over the lower triangular section.

For the derivatives with respect to m_i and C_{ij}, Domke (2020; 2019) show that

$$\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial m_i} = \mathbf{e}_i \quad \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial C_{ij}} = \mathbf{e}_i u_j, \qquad (9)$$

where $\mathbf{e}_i$ is the unit basis of the ith component.

Therefore,

$$\begin{aligned}
&\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \\
&= \sum_{i=1}^d \mathbf{e}_i \mathbf{e}_i^\top + \sum_{i=1}^d \sum_{j \le i} \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\right)^\top \\
&= \mathbf{I} + \underbrace{\sum_{i=1}^d \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ii}}}\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ii}}}\right)^\top}_{\text{diagonal of } \boldsymbol{C}} \\
&\quad + \underbrace{\sum_{i=1}^d \sum_{j<i} \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}}\right)^\top}_{\text{off-diagonal of } \boldsymbol{C}}, \qquad (10)
\end{aligned}$$

leaving us with the derivatives of the scale term.

The gradient with respect to $\lambda_{C_{ij}}$, however, depends on the parameterization. That is,

$$\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda_{C_{ij}}} = \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial C_{ij}} \frac{\partial C_{ij}}{\partial \lambda_{C_{ij}}} = \mathbf{e}_i u_j \frac{\partial C_{ij}}{\partial \lambda_{C_{ij}}}. \qquad (11)$$

For the diagonal elements, $\lambda_{C_{ii}} = s_i$. Thus,

$$\frac{\partial C_{ii}}{\partial s_i} = \frac{\partial \phi(s_i)}{\partial s_i} = \phi'(s_i). \qquad (12)$$

And for the off-diagonal elements, $\lambda_{L_{ij}} = L_{ij}$, and

$$\frac{\partial C_{ij}}{\partial L_{ij}} = 1. \qquad (13)$$

Plugging Equations (11) to (13) into Equation (10),

$$\begin{aligned}
&\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \\
&= \mathbf{I} + \underbrace{\sum_{i=1}^d (u_i \phi'(s_i))^2 \mathbf{e}_i \mathbf{e}_i^\top}_{\text{diagonal of } \boldsymbol{C}} + \underbrace{\sum_{i=1}^d \sum_{j=1, j<i} u_j^2 \mathbf{e}_i \mathbf{e}_i^\top}_{\text{off-diagonal of } \boldsymbol{C}} \\
&= \mathbf{I} + \underbrace{\sum_{i=1}^d u_i^2 (\phi'(s_i))^2 \mathbf{e}_i \mathbf{e}_i^\top}_{\text{diagonal of } \boldsymbol{C}} + \underbrace{\sum_{i=1}^d \sum_{j \le i} u_j^2 \mathbf{e}_i \mathbf{e}_i^\top - \sum_{i=1}^d u_i^2 \mathbf{e}_i \mathbf{e}_i^\top}_{\text{off-diagonal of } \boldsymbol{C}} \\
&= \mathbf{I} + \underbrace{\boldsymbol{U}\boldsymbol{\Phi}}_{\text{diagonal of } \boldsymbol{C}} + \underbrace{\boldsymbol{\Sigma} - \boldsymbol{U}}_{\text{off-diagonal of } \boldsymbol{C}} \\
&= (\mathbf{I} + \boldsymbol{\Sigma}) + \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}), \qquad (14)
\end{aligned}$$

where $\boldsymbol{U}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}$ are diagonal matrices defined as

$$\begin{aligned}\boldsymbol{\Phi} &= \operatorname{diag}\left(\left[\phi'(s_1)^2, \dots, \phi'(s_d)^2\right]\right)\\ \boldsymbol{U} &= \operatorname{diag}\left(\left[u_1^2, \dots, u_d^2\right]\right)\\ \boldsymbol{\Sigma} &= \operatorname{diag}\left(\left[u_1^2, u_1^2 + u_2^2, \dots, \textstyle\sum_{i=1}^d u_i^2\right]\right).\end{aligned}$$

The major difference with the proof of Domke (2019, Lemma 8) for the matrix square root case is that we only sum the $u_j^2 \mathbf{e}_i \mathbf{e}_i^\top$ terms over the *lower diagonal elements*. This is the variance reduction effect we get from using the Cholesky parameterization.

Coming back to Equation (8),

$$\begin{aligned}&\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2\\ &= \boldsymbol{g}_f^\top \left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda} \boldsymbol{g}\\ &= \boldsymbol{g}_f^\top \Big((\mathbf{I} + \boldsymbol{\Sigma}) + \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \Big) \boldsymbol{g}_f\\ &= \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{\Sigma} \boldsymbol{g}_f + \boldsymbol{g}_f^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \boldsymbol{g}_f. \qquad (15)\end{aligned}$$

Proof for Mean-field For the mean-field variational family, the covariance has only diagonal elements. Therefore, Equation (14) becomes

$$\left(\frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda}\right)^\top \frac{\partial \boldsymbol{t}_\lambda(\boldsymbol{u})}{\partial \lambda} = \mathbf{I} + \boldsymbol{U}\boldsymbol{\Phi},$$

and Equation (15) becomes

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \boldsymbol{g}_f^\top (\mathbf{I} + \boldsymbol{U}\boldsymbol{\Phi}) \boldsymbol{g}_f = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{g}_f.$$

□

C.2.2. Proof of Lemma 2

Lemma 2. *Let $\boldsymbol{t}_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ be a location-scale reparameterization function (Definition 1), $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and let ϕ satisfy Assumption 3.*

(i) Mean-Field

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \leq \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2,$$

where $\boldsymbol{U}$ is a diagonal matrix such that $U_{ii} = u_i^2$.

(ii) Cholesky

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \leq \left(1 + \|\boldsymbol{u}\|_2^2\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2,$$

where the equality holds for the matrix square root parameterization.

Proof. The proof continues from the result of Lemma 1.

Proof for Cholesky Lemma 1 shows that

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{\Sigma} \boldsymbol{g}_f + \boldsymbol{g}_f^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \boldsymbol{g}_f,$$

where $\boldsymbol{g}_f = \nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))$.

By the 1-Lipschitz assumption, the entries of the diagonal matrix Φ satisfy

$$\Phi_{ii} = \phi'(d_i)^2 \leq 1,$$

which means

$$\boldsymbol{\Phi} \preceq \mathbf{I} \;\Rightarrow\; \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \preceq 0 \;\Rightarrow\; \boldsymbol{g}_f{}^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \boldsymbol{g}_f \leq 0.$$

Therefore, for the full-rank Cholesky parameterization and a 1-Lipschitz conditioner ϕ,

$$\begin{aligned}&\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2\\ &= \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 + \boldsymbol{g}_f{}^\top \boldsymbol{\Sigma} \boldsymbol{g}_f + \boldsymbol{g}_f{}^\top \boldsymbol{U}(\boldsymbol{\Phi} - \mathbf{I}) \boldsymbol{g}_f\\ &\leq \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 + \boldsymbol{g}_f{}^\top \boldsymbol{\Sigma} \boldsymbol{g}_f\\ &\leq \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 + \|\boldsymbol{\Sigma}\|_{2,2} \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2\\ &= \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 + \left(\sum_{i=1}^d u_i^2\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2\\ &= \left(1 + \|\boldsymbol{u}\|_2^2\right) \|\nabla f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2,\end{aligned}$$

where $\|\boldsymbol{U}\|_{2,2}$ is the L_2 operator norm of $\boldsymbol{U}$. This upper bound coincides with that of the matrix square root parameteration. Thus, unforunately, this bound fails to acknowledge the lower variance of the Cholesky parameterization, coinciding with that of the matrix square root parameterization.

Proof for Mean-field (Definition 8) For the mean-field parameterization, Lemma 1 shows that

$$\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 = \|\boldsymbol{g}_f\|_2^2 + \boldsymbol{g}_f^\top \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{g}_f.$$

For the second term,

$$\boldsymbol{g}_f^\top \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{g}_f \leq \|\boldsymbol{U}\|_{2,2}\|\boldsymbol{\Phi}\|_{2,2}\|\boldsymbol{g}_f\|_2^2.$$

By the 1-Lipschitzness of ϕ,

$$\|\boldsymbol{\Phi}\|_{2,2} = \sigma_{\max}(\boldsymbol{\Phi}) = \max_{i=1,\dots,d} \phi'(s_i)^2 \leq 1.$$

Then,

$$\boldsymbol{g}_f^\top (\boldsymbol{U}\boldsymbol{\Phi})\, \boldsymbol{g}_f \leq \|\boldsymbol{U}\|_{2,2}\, \|\boldsymbol{g}_f\|_2^2 \tag{16}$$

$$\leq \|\boldsymbol{U}\|_{\mathrm{F}}\, \|\boldsymbol{g}_f\|_2^2, \tag{17}$$

which gives the result. Here, unlike the bounds on $\boldsymbol{\Phi}$, the bounds in Equations (16) and (17) are quite loose, and become looser as the dimensionality increases.

□

C.2.3. Proof of Lemma 3

Lemma 3. *Let the assumptions of Lemma 2 hold and $\boldsymbol{u} \sim \varphi$ satisfy Assumption 1. Then, for the mean-field parameterization,*

$$\mathbb{E}\|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2 \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right)$$
$$\leq \left(\sqrt{d\kappa} + \kappa\sqrt{d} + 1\right) \|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + \left(2\kappa\sqrt{d} + 1\right) \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

Proof. The key idea is to prove a similar result as Lemma 7, but with better constants to reflect that the mean-field parameterization has a lower variance.

First,

$$\mathbb{E}\|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2 \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right)$$
$$= \mathbb{E}\|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2 + \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2,$$

applying Lemma 8,

$$= \|\boldsymbol{m} - \boldsymbol{z}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}} + \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2. \tag{18}$$

The last term decomposes as

$$\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\|\boldsymbol{t}_\lambda(\boldsymbol{u}) - \boldsymbol{z}\|_2^2 = \underbrace{\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{u}}_{\text{Term ❶}}$$
$$+ 2\, \underbrace{\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top (\boldsymbol{m} - \boldsymbol{z})}_{\text{Term ❷}}$$
$$+ \underbrace{\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}}_{\text{Term ❸}}\, \|\boldsymbol{m} - \boldsymbol{z}\|_2^2. \tag{19}$$

We will now focus on the stochastic terms ❶-❸ one by one.

First, for Term ❶, notice that the mean-field parameterization implies that $\boldsymbol{C} = \mathrm{diag}(c_1, \dots, c_d)$. Thus,

$$\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{u} = \mathbb{E}\left(\sqrt{\sum_{i=1}^d u_i^4}\right)\left(\sum_{i=1}^d c_i^2\, u_i^2\right)$$
$$= \sum_{i=1}^d c_i^2\, \mathbb{E}\left(\sqrt{\sum_{j=1}^d u_j^4}\right) u_i^2,$$

applying Cauchy-Schwarz inequality for expectations,

$$\leq \sum_{i=1}^d c_i^2 \sqrt{\left(\mathbb{E}\sum_{j=1}^d u_j^4\right)\left(\mathbb{E}u_i^4\right)}$$

and given Assumption 1,

$$= \sum_{i=1}^d c_i^2 \sqrt{d\kappa^2}$$
$$= \kappa\sqrt{d}\, \|\boldsymbol{C}\|_{\mathrm{F}}^2. \tag{20}$$

Term ❷ can be bounded as

$$\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top (\boldsymbol{m}-\boldsymbol{z})$$

using the Cauchy-Schwarz inequality for vectors as

$$\leq \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}} \|\boldsymbol{C}\boldsymbol{u}\|_2 \|\boldsymbol{m}-\boldsymbol{z}\|_2,$$

again, applying the inequality for expectations,

$$= \sqrt{\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}^2\, \mathbb{E}\|\boldsymbol{C}\boldsymbol{u}\|_2^2}\, \|\boldsymbol{m}-\boldsymbol{z}\|_2$$

$$= \sqrt{\mathbb{E}\left(\sum_{i=1}^{d} u_i^4\right) \operatorname{tr}\left(\boldsymbol{C}^\top \boldsymbol{C}\, \mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top\right)}\, \|\boldsymbol{m}-\boldsymbol{z}\|_2,$$

from Assumption 1,

$$= \sqrt{d\kappa \operatorname{tr}\left(\boldsymbol{C}^\top \boldsymbol{C}\right)}\, \|\boldsymbol{m}-\boldsymbol{z}\|_2$$

$$= \sqrt{d\kappa}\, \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{m}-\boldsymbol{z}\|_2$$

$$= \sqrt{d\kappa} \sqrt{\|\boldsymbol{C}\|_{\mathrm{F}}^2 \|\boldsymbol{m}-\boldsymbol{z}\|_2^2},$$

and by the arithmetic mean-geometric mean inequality,

$$= \frac{\sqrt{d\kappa}}{2}\left(\|\boldsymbol{C}\|_{\mathrm{F}}^2 + \|\boldsymbol{m}-\boldsymbol{z}\|_2^2\right). \tag{21}$$

Finally, Term ❸ follows as

$$\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}} = \mathbb{E}\sqrt{\sum_{i=1}^{d} u_i^4},$$

using Jensen's inequality,

$$\leq \sqrt{\mathbb{E}\sum_{i=1}^{d} u_i^4}$$

$$= \sqrt{d\kappa}. \tag{22}$$

Combining all the results, Equation (18) becomes

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})-\boldsymbol{z}\|_2^2 \left(1+\|\boldsymbol{U}\|_{\mathrm{F}}\right) \\
&\quad \leq \|\boldsymbol{m}-\boldsymbol{z}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2 \\
&\qquad + \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{u} \\
&\qquad + 2\,\mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}\, \boldsymbol{u}^\top \boldsymbol{C}^\top \|\boldsymbol{m}-\boldsymbol{z}\|_2^2 \\
&\qquad + \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}} \|\boldsymbol{m}-\boldsymbol{z}\|_2^2
\end{aligned}$$

and applying Equations (20) to (22),

$$\begin{aligned}
&\leq \|\boldsymbol{m}-\boldsymbol{z}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2 \\
&\quad + \kappa\sqrt{d}\|\boldsymbol{C}\|_{\mathrm{F}} \\
&\quad + \kappa\sqrt{d}\left(\|\boldsymbol{C}\|_{\mathrm{F}}^2 + \|\boldsymbol{m}-\boldsymbol{z}\|_2^2\right) \\
&\quad + \sqrt{d\kappa}\, \|\boldsymbol{m}-\boldsymbol{z}\|_2^2 \\
&= \left(\sqrt{d\kappa} + \kappa\sqrt{d} + 1\right) \|\boldsymbol{m}-\boldsymbol{z}\|_2^2 + \left(2\kappa\sqrt{d}+1\right) \|\boldsymbol{C}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

C.2.4. Proof of Lemma 4

Lemma 4. *Let $\boldsymbol{g}_M$ be the M-sample gradient estimator of F (Definition 4) for some function f, h and let $\boldsymbol{u}$ be some random variable. Then,*

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{1}{M}\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\|_2^2 + \|\nabla F(\boldsymbol{\lambda})\|_2^2.$$

Proof. From the definition of variance,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \\
&\quad = \operatorname{tr}\mathbb{V}\left[\boldsymbol{g}_M\right] + \|\mathbb{E}\boldsymbol{g}_M\|_2^2,
\end{aligned}$$

following the definition in Equation (4),

$$= \operatorname{tr}\mathbb{V}\left[\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}_m\right] + \|\nabla F(\boldsymbol{\lambda})\|_2^2,$$

and then the definition in Equation (5),

$$= \operatorname{tr}\mathbb{V}\left[\frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}_m)\right) + \nabla h(\boldsymbol{\lambda})\right] + \|\nabla F(\boldsymbol{\lambda})\|_2^2,$$

by the linearity of variance,

$$\begin{aligned}
&= \frac{1}{M}\operatorname{tr}\mathbb{V}\left[\nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\right] + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\
&= \frac{1}{M}\left(\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\|_2^2 - \|\mathbb{E}\nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\|_2^2\right) \\
&\qquad + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \qquad (23) \\
&\leq \frac{1}{M}\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f\left(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\|_2^2 + \|\nabla F(\boldsymbol{\lambda})\|_2^2.
\end{aligned}$$

□

C.3. Proof of Theorems

C.3.1. Proof of Theorem 1

Theorem 1. *Let $\boldsymbol{g}_M$ be an M-sample estimate of the gradient of the ELBO in entropy regularized form (Definition 5). Also, assume that Assumption 3 and 4 hold,*

- *f_{H} is L_{H}-smooth, and*
- *f_{KL} is μ_{KL}-quadratically growing.*

Then,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq& \frac{4L_{\mathrm{H}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\ &+ \frac{2L_{\mathrm{H}}^2}{M} C(d,\kappa)\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2 \\ &+ \frac{4L_{\mathrm{H}}^2}{\mu_{\mathrm{KL}}M} C(d,\kappa)\left(F^* - f_{\mathrm{KL}}^*\right),\end{aligned}$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*
$C(d,\kappa) = d + \kappa$ *for the Cholesky and matrix square root,*

$\bar{\zeta}_{\mathrm{KL}}$, $\bar{\zeta}_{\mathrm{H}}$ are the stationary points of f_{KL}, f_{H}, respectively, $F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, and $f_{\mathrm{KL}}^* = \inf_{\zeta\in\mathbb{R}^d} f(\zeta)$.

Proof. The proof uses the L_{H}-smoothness of f_{H} such that

$$\begin{aligned}\mathbb{E}\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 &= \mathbb{E}\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla f_{\mathrm{H}}(\bar{\zeta}_{\mathrm{H}})\|_2^2 \\ &\leq L_{\mathrm{H}}^2 \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2, \end{aligned} \tag{24}$$

where $\bar{\zeta}_{\mathrm{H}}$ is a stationary point of f_{H} such that $\nabla f_{\mathrm{H}}(\bar{\zeta}_{\mathrm{H}}) = \mathbf{0}$. These steps have been previously used by Domke (2019, Theorem 3) to prove the special case for the matrix square root parameterization.

For the mean-field parameterization, we start from Lemma 2 and apply Equation (24) as

$$\begin{aligned}&\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \\ &\leq \mathbb{E}\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right) \\ &\leq L_{\mathrm{H}}^2\, \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) + \bar{\zeta}_{\mathrm{H}}\|_2^2 \left(1 + \|\boldsymbol{U}\|_{\mathrm{F}}\right),\end{aligned}$$

applying Lemma 3,

$$\begin{aligned}&\leq L_{\mathrm{H}}^2 \left(\kappa\sqrt{d} + \sqrt{\kappa d} + 1\right)\|\boldsymbol{m} - \bar{\zeta}_{\mathrm{H}}\|_2^2 \\ &\quad + L_{\mathrm{H}}^2 \left(2\kappa\sqrt{d} + 1\right)\|\boldsymbol{C}\|_{\mathrm{F}}^2,\end{aligned}$$

and since the kurtosis satisfies $\kappa \geq 1$ and thus $\kappa \geq \sqrt{\kappa}$,

$$\leq L_{\mathrm{H}}^2 \left(2\kappa\sqrt{d} + 1\right)\left(\|\boldsymbol{m} - \bar{\zeta}_{\mathrm{H}}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right). \tag{25}$$

Similarly, for the full-rank parameterizations, we start from Lemma 2 and apply Equation (24) as

$$\begin{aligned}&\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \\ &\leq \mathbb{E}\|\nabla f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \left(1 + \|\boldsymbol{u}\|_2^2\right), \\ &\leq L_{\mathrm{H}}^2\, \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 \left(1 + \|\boldsymbol{u}\|_2^2\right),\end{aligned} \tag{26}$$

applying Lemma 7,

$$= L_{\mathrm{H}}^2 \left((d+1)\|\boldsymbol{m} - \bar{\zeta}_{\mathrm{H}}\|_2^2 + (d+\kappa)\|\boldsymbol{C}\|_{\mathrm{F}}^2\right),$$

and since the kurtosis satisfies $\kappa \geq 1$,

$$\leq L_{\mathrm{H}}^2 (d+\kappa)\left(\|\boldsymbol{m} - \bar{\zeta}_{\mathrm{H}}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right). \tag{27}$$

Both Equations (25) and (27) can now be denoted as

$$\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \leq L_{\mathrm{H}}^2\, C(d,\kappa)\left(\|\boldsymbol{m} - \bar{\zeta}_{\mathrm{H}}\|_2^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right),$$

where by Lemma 8,

$$= L_{\mathrm{H}}^2\, C(d,\kappa)\, \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2, \tag{28}$$

and the constants are $C(d,\kappa) = \kappa\sqrt{d} + 1$ for mean-field and $C(d,\kappa) = d + \kappa$ for the full-rank parameterizations.

As mentioned in the sketch, it is necessary to convert the entropy-regularized form into the KL-regularized form through the following inequality:

$$\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 \leq 2\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{KL}}\|_2^2 + 2\,\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2.$$

where $\bar{\zeta}_{\mathrm{KL}} = \Pi_{f_{\mathrm{KL}}}(\bar{\zeta}_{\mathrm{H}})$ is a projection of $\bar{\zeta}_{\mathrm{H}}$ to the set of minimizers of f_{KL}. Note that the KL-regularized form does not need to be tractable; only its existence suffices. We can now apply the quadratic growth assumption as

$$\begin{aligned}\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{KL}}\|_2^2 &\leq \frac{2}{\mu_{\mathrm{KL}}}\left(\mathbb{E} f_{\mathrm{KL}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f_{\mathrm{KL}}^*\right) \\ &= \frac{2}{\mu_{\mathrm{KL}}}\left(F(\boldsymbol{\lambda}) - h_{\mathrm{KL}}(\boldsymbol{\lambda}) - f_{\mathrm{KL}}^*\right),\end{aligned}$$

and since $-h_{\mathrm{KL}}(\boldsymbol{\lambda}) = -\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\lambda}}, p) \leq 0$ by definition,

$$\leq \frac{2}{\mu_{\mathrm{KL}}}\left(F(\boldsymbol{\lambda}) - f_{\mathrm{KL}}^*\right) \tag{29}$$

$$= \frac{2}{\mu_{\mathrm{KL}}}\left(\left(F(\boldsymbol{\lambda}) - F^*\right) + \left(F^* - f_{\mathrm{KL}}^*\right)\right). \tag{30}$$

Combining Equation (28) with Equation (6),

$$\begin{aligned}&\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \\ &\leq 2\,L_{\mathrm{H}}^2\, C(d,\kappa)\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\zeta}_{\mathrm{H}}\|_2^2 + 2\,L_{\mathrm{H}}^2\, C(d,\kappa)\,\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2,\end{aligned}$$

and applying Equation (30),

$$\begin{aligned}&\leq \frac{4\,L_{\mathrm{H}}^2}{\mu_{\mathrm{KL}}} C(d,\kappa)\left(\left(F(\boldsymbol{\lambda}) - F^*\right) + \left(F^* - f_{\mathrm{KL}}^*\right)\right) \\ &\quad + 2\,L_{\mathrm{H}}^2\, C(d,\kappa)\,\|\bar{\zeta}_{\mathrm{KL}} - \bar{\zeta}_{\mathrm{H}}\|_2^2\end{aligned}$$

Plugging this into Lemma 4 yields the result. □

C.3.2. Proof of Theorem 2

Theorem 2. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in KL-regularized form (Definition 6). Also, assume that*

- *f_{KL} is L_{KL}-smooth,*
- *f_{KL} is μ_{KL}-quadratically growing,*

and Assumption 3 and 4 hold. Then, the gradient variance is bounded above as

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}} M} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}} M} C(d,\kappa)\left(F^* - f_{\mathrm{KL}}^*\right),$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*
$C(d,\kappa) = d + \kappa$ *for the Cholesky and matrix square root,*
$F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$*, and* $f_{\mathrm{KL}}^* = \inf_{\zeta\in\mathbb{R}^d} f(\zeta)$.

Proof. This proof uses the smoothness of f_{KL} instead of f_{H}. That is,

$$\mathbb{E}\|\nabla f_{\mathrm{KL}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 = \mathbb{E}\|\nabla f_{\mathrm{KL}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla f_{\mathrm{KL}}(\bar{\boldsymbol{\zeta}}_{\mathrm{KL}})\|_2^2$$

applying Equation (24),

$$\leq L_{\mathrm{KL}}^2\, \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{\zeta}}_{\mathrm{KL}}\|_2^2, \tag{31}$$

where $\bar{\boldsymbol{\zeta}}_{\mathrm{KL}}$ is a stationary point of f_{KL}.

Substituting Equation (31) in Equation (28),

$$\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{KL}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \leq L_{\mathrm{KL}}^2 C(d,\kappa)\, \mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{\zeta}}_{\mathrm{KL}}\|_2^2,$$

and by applying Equation (30) for f_{KL},

$$= \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}}} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \frac{2L_{\mathrm{KL}}^2}{\mu_{\mathrm{KL}}} C(d,\kappa)\left(F^* - f_{\mathrm{KL}}^*\right).$$

Plugging this to Lemma 4 proves the result. □

C.3.3. Proof of Theorem 3

Theorem 3. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in entropy-regularized form (Definition 5). Also, assume that*

- *f_{H} is L_{H}-smooth,*
- *f_{H} is μ_{H}-quadratically growing,*
- *h_{H} is bounded as $h_{\mathrm{H}}(\boldsymbol{\lambda}) > h_{\mathrm{H}}^*$ (Assumption 5),*

and Assumption 3 and 4 hold. Then, the gradient variance of $\boldsymbol{g}_M$ is bounded above as

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \leq \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}} M} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}} M} C(d,\kappa)\left(F^* - f_{\mathrm{H}}^* - h_{\mathrm{H}}^*\right),$$

where

$C(d,\kappa) = 2\kappa\sqrt{d} + 1$ *for mean-field,*

$C(d,\kappa) = d + \kappa$ *for the Cholesky parameterization,*

$F^* = \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, and $f_{\mathrm{H}}^* = \inf_{\boldsymbol{\zeta}\in\mathbb{R}^d} f(\boldsymbol{\zeta})$.

Proof. The proof is similar to that of Theorem 1. As mentioned in the *proof sketch*, we use the fact that the entropic regularizer is bounded such that

$$-h_{\mathrm{H}}(\boldsymbol{\lambda}) < -h_{\mathrm{H}}^*.$$

By applying the quadratic growth assumption directly to f_{H},

$$\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{\zeta}}_{\mathrm{H}}\|_2^2 \leq \frac{2}{\mu_{\mathrm{H}}}\left(\mathbb{E} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f_{\mathrm{H}}^*\right) = \frac{2}{\mu_{\mathrm{H}}}\left(F(\boldsymbol{\lambda}) - h_{\mathrm{H}}(\boldsymbol{\lambda}) - f_{\mathrm{H}}^*\right),$$

and by Assumption 5,

$$\leq \frac{2}{\mu_{\mathrm{H}}}\left(F(\boldsymbol{\lambda}) - F^*\right) + \frac{2}{\mu_{\mathrm{H}}}\left(F^* - f_{\mathrm{H}}^* - h_{\mathrm{H}}^*\right). \tag{32}$$

The proof resumes from Equation (28) as

$$\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f_{\mathrm{H}}(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \leq L_{\mathrm{H}}^2 C(d,\kappa)\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{\zeta}}_{\mathrm{H}}\|_2^2,$$

and by applying Equation (32),

$$= \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}}} C(d,\kappa)\left(F(\boldsymbol{\lambda}) - F^*\right) + \frac{2L_{\mathrm{H}}^2}{\mu_{\mathrm{H}}} C(d,\kappa)\left(F^* - f_{\mathrm{H}}^* - h_{\mathrm{H}}^*\right).$$

Plugging this to Lemma 4 proves the result. □

C.3.4. Proof of Theorem 4

Theorem 4. *Let $\boldsymbol{g}_M$ be an M-sample estimator of the gradient of the ELBO in either the entropy- or KL-regularized form. Also, let Assumption 4 hold where the matrix square root parameterization is used. Then, for all L-smooth and μ-strongly convex functions f such that $L/\mu < \sqrt{d+1}$, the variance of $\boldsymbol{g}_M$ is bounded below by some strictly positive constant as*

$$\mathbb{E}\|\boldsymbol{g}_M\|_2^2 \geq \frac{2\mu^2(d+1) - 2L^2}{ML}\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 + \frac{2\mu^2(d+1) - 2L^2}{ML}\left(\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - f^*\right),$$

as long as $\boldsymbol{\lambda}$ is in a local neighborhood around the unique global optimum $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} F(\boldsymbol{\lambda})$, where $F^* = F(\boldsymbol{\lambda}^*)$ and $f^* = \arg\min_{\boldsymbol{\zeta}\in\mathbb{R}^d} f(\boldsymbol{\zeta})$.

Proof. When using the matrix square root parameterization, Domke (2020) have shown that if f is L-smooth, $\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))$ is also L-smooth. Therefore, we have

$$\|\mathbb{E}\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \leq 2L\left(\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f^*\right). \tag{33}$$

Furthermore, let $\bar{\boldsymbol{\zeta}}$ be the minimizer of f, namely $f^* = f(\bar{\boldsymbol{\zeta}})$. From Lemma 6, we have

$$\mathbb{E}\|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 = \mathbb{E}\|\nabla f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2\left(1 + \|\boldsymbol{u}\|_2^2\right),$$

by the μ-strong convexity of f,

$$\geq 2\mu\,\mathbb{E}(f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - f^*)\left(1 + \|\boldsymbol{u}\|_2^2\right) \geq \mu^2\,\mathbb{E}\|\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{\zeta}}\|_2^2\left(1 + \|\boldsymbol{u}\|_2^2\right),$$

applying Lemma 7,

$$= \mu^2\left((d+1)\|\boldsymbol{m} - \bar{\boldsymbol{\zeta}}\|_2^2 + (d+\kappa)\|\boldsymbol{C}\|_{\mathrm{F}}^2\right),$$

and by the property of the kurtosis that $\kappa \geq 1$,

$$\geq \mu^2\,(d+1)\|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2^2,$$

where $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{\zeta}}, \mathbf{O})$.

Observe that $\bar{\boldsymbol{\lambda}}$ is the minimizer of $\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))$ such that

$$\mathbb{E} f(\boldsymbol{t}_{\bar{\boldsymbol{\lambda}}}(\boldsymbol{u})) = f(\bar{\boldsymbol{\zeta}}) = f^* \leq \mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))$$

for any $\boldsymbol{\lambda}$. Furthermore, from the L-smoothness of $\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u}))$, we have

$$\mu^2(d+1)\|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2^2 \geq \frac{2\mu^2(d+1)}{L}\left(\mathbb{E} f(\boldsymbol{t}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \mathbb{E} f(\boldsymbol{t}_{\bar{\boldsymbol{\lambda}}}(\boldsymbol{u}))\right).$$

Thus, we have

$$\mathbb{E}\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 \geq \frac{2\mu^2(d+1)}{L}\left(\mathbb{E}f(\boldsymbol{t}_\lambda(\boldsymbol{u})) - f^*\right). \tag{34}$$

Now, from Equation (23),

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_M\|_2^2 &= \frac{1}{M}\left(\mathbb{E}\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 - \|\mathbb{E}\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2\right) \\
&\quad + \|\nabla F(\boldsymbol{\lambda})\|_2^2, \\
&\text{applying Equation (33),} \\
&\geq \frac{1}{M}\left(\mathbb{E}\|\nabla_\lambda f(\boldsymbol{t}_\lambda(\boldsymbol{u}))\|_2^2 - 2L^2\left(\mathbb{E}f(\boldsymbol{t}_\lambda(\boldsymbol{u})) - f^*\right)\right) \\
&\quad + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\
&\text{applying Equation (34),} \\
&\geq \frac{2\mu^2(d+1) - 2L^2}{ML}\left(\mathbb{E}f(\boldsymbol{t}_\lambda(\boldsymbol{u})) - f^*\right) \\
&\quad + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\
&\geq \frac{2\mu^2(d+1) - 2L^2}{ML}\left(F(\boldsymbol{\lambda}) - h(\boldsymbol{\lambda}) - f^*\right) \\
&\quad + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\
&= \frac{2\mu^2(d+1) - 2L^2}{ML}\left(F(\boldsymbol{\lambda}) - F^*\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2 \\
&\quad + \frac{2\mu^2(d+1) - 2L^2}{ML}\left(F^* - f^* - h(\boldsymbol{\lambda})\right).
\end{aligned}$$

The last term

$$\frac{2\mu^2(d+1) - 2L^2}{ML}\left(F^* - f^* - h(\boldsymbol{\lambda})\right)$$

can be shown to be positive if $\boldsymbol{\lambda}$ is sufficiently close to the optimum. Let $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}} F(\boldsymbol{\lambda})$ be the minimizer of F. Then, we have

$$\begin{aligned}
F^* - f^* - h(\boldsymbol{\lambda}) &= \mathbb{E}f(\boldsymbol{t}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) + h(\boldsymbol{\lambda}^*) - f^* - h(\boldsymbol{\lambda}) \\
&= \left(\mathbb{E}f(\boldsymbol{t}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - f^*\right) + \left(h(\boldsymbol{\lambda}^*) - h(\boldsymbol{\lambda})\right),
\end{aligned}$$

where the first term is strictly positive and the second term goes to zero as $\boldsymbol{\lambda} \to \boldsymbol{\lambda}^*$. □