# Nearly Dimension-Independent Convergence of Mean-Field Black-Box Variational Inference

**Kyurae Kim**
University of Pennsylvania
kyrkim@seas.upenn.edu

**Yi-An Ma**
University of California San Diego
yianma@ucsd.edu

**Trevor Campbell**
University of British Columbia
trevor@stat.ubc.ca

**Jacob R. Gardner**
University of Pennsylvania
jacobrg@seas.upenn.edu

## Abstract

We prove that, given a mean-field location-scale variational family, black-box variational inference (BBVI) with the reparametrization gradient converges at a rate that is nearly independent of any explicit dimension dependence. Specifically, for a $d$-dimensional strongly log-concave and log-smooth target, the number of iterations for BBVI with a sub-Gaussian family to obtain a solution $\epsilon$-close to the global optimum has an explicit dimension dependence no larger than $\mathrm{O}(\log d)$. This is a significant improvement over the $\mathrm{O}(d)$ dependence of full-rank location-scale families. For heavy-tailed families, we prove a weaker $\mathrm{O}(d^{2/k})$ dependence, where $k$ is the number of finite moments of the family. Additionally, if the Hessian of the target log-density is constant, the complexity is free of any explicit dimension dependence. We also prove that our bound on the gradient variance, which is key to our result, cannot be improved using only spectral bounds on the Hessian of the target log-density.

## 1 Introduction

Variational inference (VI; Blei et al., 2017; Hinton and van Camp, 1993; Jordan et al., 1999; Peterson and Hartman, 1989) is an effective method for approximating intractable high-dimensional distributions and models with tall datasets. Among various VI algorithms, black-box VI (BBVI; Kucukelbir et al., 2017; Ranganath et al., 2014; Titsias and Lázaro-Gredilla, 2014; Wingate and Weber, 2013), which minimizes the exclusive KL divergence (Kullback and Leibler, 1951) via stochastic gradient descent (SGD; Bottou et al., 2018; Robbins and Monro, 1951) in the space of parameters, is widely used due to its flexibility to apply to a wide range of variational families with only minor modifications (Bingham et al., 2019; Carpenter et al., 2017; Fjelde et al., 2025; Ge et al., 2018; Patil et al., 2010). Specifically, location-scale variational families—in which a base distribution is mutated by an affine transformation—remain a popular choice, encompassing those with diagonal scale matrices (the "mean-field" approximation; Hinton and van Camp, 1993; Peterson and Hartman, 1989), as well as scale matrices with low rank (Ong et al., 2018; Rezende et al., 2014; Tomczak et al., 2020) and full-rank (Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014) factors.

The choice of the variational family is generally known to affect the convergence speed of BBVI, where families that are more "expressive," those that contain more complex distributions, result in slower convergence. For example, in location-scale families, it has been empirically observed that mean-field families often provide faster convergence to an accurate posterior approximation than full-rank families (Agrawal et al., 2020; Giordano et al., 2018, 2024; Ko et al., 2024; Zhang et al., 2022). This is because full-rank families often require running SGD with a smaller step size and

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

for longer; even given a large computation budget, BBVI on a full-rank family may not converge adequately (Ko et al., 2024). Therefore, choosing the expressiveness of the family corresponds to trading statistical accuracy for computational efficiency (Bhatia et al., 2022). In order to control this trade-off for our benefit, a clear theoretical understanding of the relationship between convergence speed and expressiveness is needed.

Formally, consider the setting of approximating a $\mu$-strongly log-concave and $L$-log-smooth target, where $\kappa \triangleq L/\mu$ is the condition number. For BBVI with the reparametrization gradient (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014) on a full-rank location-scale family, an $\epsilon$-close solution to the global optimum in squared distance in parameter space can be obtained after at least $\mathrm{O}(d\kappa^2\epsilon^{-1})$ iterations (Domke, 2019; Kim et al., 2023a). For mean-field location-scale families, on the other hand, the iteration complexity improves to $\mathrm{O}(\sqrt{d}\,\kappa^2\epsilon^{-1})$ (Kim et al., 2023a). While this is clearly better than the $\mathrm{O}(d)$ explicit dimension dependence of full-rank families, it has been conjectured that a better dependence is more likely (Kim et al., 2023a).

In this work, we positively resolve this conjecture by obtaining stronger convergence guarantees for BBVI on mean-field location-scale families (Section 3). In particular, under the conditions stated above, we prove that BBVI with a mean-field location-scale family with sub-Gaussian tails can obtain an $\epsilon$-accurate solution in squared distance after $\mathrm{O}((\log d)\kappa^2\epsilon^{-1})$ iterations. Heavier-tailed families achieve a weaker $\mathrm{O}(d^{2/k}\kappa^2\epsilon^{-1})$ iteration complexity guarantee, where $k$ is the number of finite moments of the variational family. For the Student-$t$ variational family with a high-enough degrees of freedom $\nu$, this corresponds to a $\mathrm{O}(d^{2/(\nu-2)})$ explicit dimension dependence. In addition, if the Hessian of the target log-density is constant, any mean-field location-scale family attains a $\mathrm{O}(\kappa^2\epsilon^{-1})$ iteration complexity without any explicit dependence on $d$.

The key element of the proof is a careful probabilistic analysis of the variance of the reparametrization gradient (Section 4): In general, the reparametrization gradient of the scale parameters contains heavy-tailed components that grow not-so-slowly in $d$. However, for mean-field families, only a *single* random coordinate turns out to be heavy-tailed. Through a probabilistic decomposition, the influence of this heavy-tailed component can be averaged out over all $d$ coordinates. Then the lighter-tailed components of the gradient dominate as $d$ increases, resulting in a benign dimension dependence (Lemma 4.1). We also provide a lower bound (Proposition 4.2) showing that our analysis cannot be improved when using only spectral bounds on the Hessian of the target log-density.

## 2 Preliminaries

**Notation** We denote random variables in sans serif (*e.g.*, $\mathsf{u}$, $\mathsf{U}$). $\mathbb{S}^d_{\succ 0} \subset \mathbb{R}^{d\times d}$ denotes the set of $d \times d$ positive definite (PD) matrices, $\mathbb{D}^d \subset \mathbb{R}^{d\times d}$ denotes the set of diagonal matrices, and $\mathbb{D}^d_{\succ 0} \subset \mathbb{D}^d \cap \mathbb{S}^{d\times d}_{\succ 0}$ is its positive definite subset. $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_2$ denote the Euclidean inner product and norm. For a matrix $A \in \mathbb{R}^{d\times d}$, $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(A^\top A)}$ is the Frobenius norm, $\|A\|_2 = \sigma_{\max}(A)$ is the $\ell_2$ operator norm, where $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ are the largest and smallest singular values.

### 2.1 Problem Setup

Our problem of interest is an optimization problem over some space $\Lambda \subseteq \mathbb{R}^p$ of the form of

$$\underset{\lambda\in\Lambda}{\text{minimize}} \left\{F(\lambda) \triangleq f(\lambda) + h\left(\lambda\right)\right\}, \quad \text{where} \quad f(\lambda) \triangleq \mathbb{E}_{\mathsf{z}\sim q_\lambda}\ell\left(\mathsf{z}\right), \tag{1}$$

$\ell : \mathbb{R}^d \to \mathbb{R}$ is a measurable function we refer to as the "target function", $h : \Lambda \to \mathbb{R}$ is a potentially non-smooth convex regularizer, and the expectation $\mathbb{E}_{\mathsf{z}\sim q_\lambda}\ell(\mathsf{z})$ is assumed to be intractable.

BBVI is a special case of Eq. (1) where $\ell = -\log\pi$ is the negative (unnormalized) log-density of some distribution $\pi$ with respect to the Lebesgue measure and $h(\lambda) = -\mathbb{H}[q_\lambda]$ is the negative differential entropy of $q_\lambda$. Then $F$ is the exclusive Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}$ (Kullback and Leibler, 1951) up to an additive constant (Jordan et al., 1999), where Eq. (1) reduces to

$$\underset{\lambda\in\Lambda}{\text{minimize}} \left\{\mathrm{D}_{\mathrm{KL}}(q_\lambda,\pi) \propto -\mathbb{E}_{\mathsf{z}\sim q_\lambda}\log\pi(\mathsf{z}) - \mathbb{H}\left(q_\lambda\right)\right\}, \tag{2}$$

We assume $\pi$ is supported on $\mathbb{R}^d$, which, unless discrete-valued variables are involved, is often valid after appropriate support transformations (Kim et al., 2023a, §2.2). Such a setup for BBVI has been proposed by Kucukelbir et al. (2017), and now encompasses most practical use of BBVI with the

reparametrization gradient as implemented in Stan (Carpenter et al., 2017), PyMC (Patil et al., 2010), Pyro (Bingham et al., 2019), and Turing (Fjelde et al., 2025; Ge et al., 2018).

For the purpose of a quantitative theoretical analysis, we will consider the following properties:

**Definition** (Smoothness)**.** *For some $\phi : \mathbb{R}^d \to \mathbb{R}$, we say $\phi$ is L-(Lipschitz )smooth if there exists some $L \in (0, +\infty)$ such that, for all $z, z' \in \mathbb{R}^d$,*

$$\|\nabla\phi(z) - \nabla\phi(z')\|_2 \leq L\|z - z'\|_2 \ .$$

**Definition** (Strong Convexity)**.** *For some $\phi : \mathbb{R}^d \to \mathbb{R}$, we say $\phi$ is $\mu$-strongly convex if there exists some constant $\mu \in (0, L]$ such that, for all $z, z' \in \mathbb{R}^d$,*

$$\langle \nabla\phi(z), z - z' \rangle \geq \phi(z) - \phi(z') + \frac{\mu}{2}\|z - z'\|_2^2 \ .$$

In the context of BBVI, assuming that $\ell = -\log\pi$ is both $\mu$-strongly convex and $L$-smooth is equivalent to assuming $\pi$ is $\mu$-strongly log-concave and $L$-log-Lipschitz smooth, respectively, which is common in the analysis of MCMC (Chewi, 2024) and VI (Arnese and Lacker, 2024; Diao et al., 2023; Domke et al., 2023; Kim et al., 2023a; Lambert et al., 2022; Lavenant and Zanella, 2024).

### 2.2 Variational Family

We consider the location-scale family (Casella and Berger, 2001, §3.5):

**Definition 2.1** (Location-Scale Variational Family)**.** *A family of distributions $\mathcal{Q}$ is referred to as a location-scale variational family if there exists some univariate distribution $\varphi$ dominated by the Lebesgue measure such that each member of $\mathcal{Q}$ indexed by $\lambda = (m, C) \in \mathbb{R}^d \times \mathcal{C}$, where $\mathcal{C} \subset \mathbb{R}^{d\times d}$ and $q_\lambda \in \mathcal{Q}$, satisfies*

$$\mathsf{z} \sim q_\lambda \qquad \Leftrightarrow \qquad \mathsf{z} \overset{\mathrm{d}}{=} \mathcal{T}_\lambda(\mathsf{u}) \ ,$$

*where*

$$\mathcal{T}_\lambda(\mathsf{u}) \triangleq C\mathsf{u} + m, \qquad \mathsf{u} \triangleq (\mathsf{u}_1, \ldots, \mathsf{u}_d), \qquad \mathsf{u}_i \overset{\text{i.i.d.}}{\sim} \varphi \ ,$$

*and $\overset{\mathrm{d}}{=}$ is equivalence in distribution. Then $\mathcal{T}_\lambda$ is referred to as the "reparametrization function," while $m$ and $C$ are referred to as the location and scale parameters, respectively.*

In addition, we impose mild regularity assumptions on the moments of the base distribution:

**Assumption 2.2.** *$\varphi$ satisfies the following: (i) It is standardized such that $\mathbb{E}\mathsf{u}_i = 0$ and $\mathbb{E}\mathsf{u}_i^2 = 1$, (ii) symmetric such that $\mathbb{E}\mathsf{u}_i^3 = 0$, and (iii) its kurtosis is finite such that $\mathbb{E}\mathsf{u}_i^4 = r_4 < \infty$.*

The location-scale family with Assumption 2.2 encompasses many variational families used in practice, such as Gaussians, Student-$t$ with a high-enough degrees of freedom $\nu$, Laplace, and so on, and enables the use of the reparametrization gradient.

While the choice of $\varphi$ gives control over the tail behavior of the family, the choice of the structure of the scale matrix $C$ gives control over how much correlation between coordinates of $\ell$ the variational approximation can represent. This ability to represent correlations is often referred to as the "expressiveness" of a variational family, where the most expressive choice is the following:

**Definition 2.3** (Full-Rank Location-Scale Family)**.** *We say $\mathcal{Q}$ is a full-rank location-scale family if it satisfies Definition 2.1 and, for any $C \in \mathcal{C}$, $C$ is invertible and the squared $C$s, $CC^\top$, span the whole space of dense $\mathbb{R}^{d\times d}$ positive definite matrices as $\{CC^\top \mid C \in \mathcal{C}\} = \mathcal{S}^d_{\succ 0}$.*

Typically, full-rank location-scale families are formed by setting $\mathcal{C}$ to be the set of invertible triangular matrices (the "Cholesky factor parametrization"; Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014) or the set of symmetric square roots (Domke, 2020; Domke et al., 2023). Adding further restrictions on $\mathcal{C}$ forms various subsets of the broader location-scale family. In this work, we focus on the case where $C \in \mathcal{C}$ is restricted to be diagonal such that $\mathcal{C} \subset \mathbb{D}^d$, which is known as the mean-field approximation (Hinton and van Camp, 1993; Peterson and Hartman, 1989):

**Definition 2.4** (Mean-Field Location-Scale Family)**.** *We say $\mathcal{Q}$ is a mean-field location-scale family if it satisfies Definition 2.1 and all $C \in \mathcal{C}$ are diagonal such that $\mathcal{C} \subset \mathbb{D}^d$ .*

### 2.3 Algorithm Setup

Recall that BBVI is essentially SGD in the space of parameters of the variational distribution. Therefore, we have to define the space of parameters. For this, we use the "linear" parametrization:

$$\Lambda = \left\{\lambda = (m, C) \mid m \in \mathbb{R}^d, C \in \mathbb{D}^d_{\succ 0}\right\} \subset \mathbb{R}^p \ . \tag{3}$$

Under this parametrization, the desirable properties of $\ell$ easily transfer to $f$. For instance, if $\ell$ is $\mu$-strongly convex and $L$-smooth, $f$ is also $\mu$-strongly convex and $L$-smooth (Domke, 2020). This contrasts with "non-linear parametrizations" commonly used in practice, such as making the diagonal positive by $C_{ii} = \exp(\lambda_{C_{ii}})$. Such practice rules out transfer of strong convexity and smoothness (Kim et al., 2023a) unless constraints such as $C_{ii} \geq \delta$ for some $\delta > 0$ are enforced (Hotti et al., 2024). (Though they can sometimes be beneficial by reducing gradient variance; Hotti et al., 2024; Kim et al., 2023b.) The flip side of using the linear parametrization is that we must now enforce the constraint $C \succ 0$. Furthermore, $h$ then becomes non-smooth with respect to $C$:

$$h(\lambda) \quad = \quad -\mathbb{H}(q_\lambda) \quad = \quad -\log|\det C| - d\,\mathbb{H}(\varphi) \quad = \quad -\textstyle\sum_{i=1}^{d} \log|C_{ii}| - d\,\mathbb{H}(\varphi)\ . \tag{4}$$

This corresponds to log-barrier functions (Parikh and Boyd, 2014, §6.7.5), which are non-smooth. Thus, the optimization algorithm must somehow deal with these difficulties (Domke, 2020).

In this work, we will rely on the proximal variant of stochastic gradient descent (SGD; Bottou, 1999; Bottou et al., 2018; Nemirovski et al., 2009; Robbins and Monro, 1951; Shalev-Shwartz et al., 2011), often referred to as stochastic proximal gradient descent (SPGD; Nemirovski et al., 2009). Proximal methods are a family of methods that rely on proximal operators (Parikh and Boyd, 2014), which are well defined as long as the following hold:

**Assumption 2.5.** *$h : \Lambda \to \mathbb{R} \cup \{+\infty\}$ is convex, bounded below, and lower semi-continuous.*

The non-smoothness of $h$ and the domain constraint are handled by the proximal operator

$$\operatorname{prox}_{\gamma h}(\lambda) \triangleq \operatorname*{arg\,min}_{\lambda' \in \Lambda} \left\{ h(\lambda') + (1/\gamma)\|\lambda - \lambda'\|_2^2 \right\} ,$$

while the intractability of $f$ is handled through stochastic estimates of $\nabla f$ (Definition 2.6). For a step size schedule $(\gamma_t)_{t\geq 0}$, $\widehat{\nabla f}$, an unbiased estimator of $\nabla f(\lambda_t) = \mathbb{E}\widehat{\nabla f}(\lambda_t; u)$, and a sequence of i.i.d. noise $(u_t)_{t\geq 0}$, for each $t \geq 0$, SPGD iterates

$$\lambda_{t+1} = \operatorname{prox}_{\gamma_t h}\big(\lambda_t - \gamma_t \widehat{\nabla f}(\lambda; u_t)\big) .$$

In the case of BBVI with a mean-field location-scale family, the proximal operator of Eq. (4) is identical to that of log-barrier functions (Parikh and Boyd, 2014, §6.7.5):

$$\operatorname{prox}_{\gamma h}(\lambda = (m, C)) = (m, C'), \quad \text{where} \quad C'_{ii} = (1/2)\left(C_{ii} + \sqrt{C_{ii}^2 + 4\gamma}\right) .$$

Instead of using SPGD, one can also use projected SGD, where $C$ is projected to a subset where $F$ is smooth (Domke, 2020) and use the "closed-form entropy" gradient $\widehat{\nabla F} \triangleq \widehat{\nabla f} + \nabla h$ (Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014). However, the resulting theoretical guarantees are indistinguishable (Domke et al., 2023), and the need for setting a closed domain of $C$ is inconvenient. Therefore, we only consider SPGD. But our results can easily be applied to projected SGD.

For $\widehat{\nabla f}$, we will use the classic *reparametrization gradient* (Ho and Cao, 1983; Rubinstein, 1992):

**Definition 2.6** (Reparametrization Gradient)**.** *For a differentiable function $\ell : \mathbb{R}^d \to \mathbb{R}$,*

$$\widehat{\nabla f}(\lambda; u) \triangleq \nabla_\lambda \ell\left(\mathcal{T}_\lambda\left(u\right)\right) = \frac{\partial \mathcal{T}_\lambda(u)}{\partial \lambda} \nabla \ell\left(\mathcal{T}_\lambda\left(u\right)\right) \ , \quad \textit{where} \quad u \sim \varphi \ ,$$

*is an unbiased estimator of $\nabla f$ such that $\nabla_\lambda \mathbb{E}_{z\sim q_\lambda} \ell(z) = \nabla f(\lambda)$.*

The reparametrization gradient, also known as the push-in gradient or pathwise gradient, was introduced to VI by Kingma and Welling (2014); Rezende et al. (2014); Titsias and Lázaro-Gredilla (2014). (See also the reviews by Glasserman 1991; Mohamed et al. 2020; Pflug 1996.) It is empirically observed to outperform alternatives (Kucukelbir et al., 2017; Mohamed et al., 2020) such as the score gradient (Glynn, 1990; Williams, 1992) and *de facto* standard whenever $\ell$ is differentiable. (Though theoretical evidence of this superiority is limited to the quadratic setting; Xu et al., 2019.)

### 2.4 General Analysis of Stochastic Proximal Gradient Descent

Analyzing the convergence of BBVI corresponds to analyzing the convergence of SPGD (or more broadly, of SGD) for the class of problems that corresponds to BBVI. For this, we will first discuss sufficient conditions for the convergence of SPGD and the resulting consequences.

**Assumption 2.7** (Lipschitz Gradients in Expectation)**.** *There exists some constant $\mathcal{L} \in [0, \infty)$ such that, for all $\lambda, \lambda' \in \Lambda$,*

$$\mathbb{E}\|\widehat{\nabla f}(\lambda; u) - \widehat{\nabla f}(\lambda'; u)\|_2^2 \leq \mathcal{L}^2 \|\lambda - \lambda'\|_2^2 .$$

**Assumption 2.8** (Bounded Variance)**.** *There exists some constant $\sigma \in [0, \infty)$ such that, for all $\lambda_* \in \arg\min_{\lambda \in \Lambda} F(\lambda)$,*

$$\mathbb{E}\|\widehat{\nabla f}(\lambda_*; \mathsf{u})\|_2^2 \leq \sigma^2 \ .$$

Both assumptions were initially used by Bach and Moulines (2011, Assumptions H2 and H4) to analyze the convergence of SGD. Here, Assumption 2.7 serves as an analog of $L$-smoothness, and thus determines the largest stepsize we can use. The strategy of combining Assumptions 2.7 and 2.8 is referred to as "variance transfer" (Garrigos and Gower, 2023, §4.3.3). Previously, for analyzing BBVI, a slightly different assumption called quadratically-bounded variance (QV)—which assumes the existence of $\alpha, \beta \in [0, +\infty)$ such that, for all $\lambda \in \Lambda$, $\mathbb{E}\|\widehat{\nabla f}(\lambda; \mathsf{u})\|_2^2 \leq \alpha\|\lambda - \lambda_*\|_2^2 + \beta$ holds—has been commonly used (Domke, 2019; Domke et al., 2023; Kim et al., 2024b). While similar, our assumptions result in a constant-factor improvement in the resulting bounds.

For the analysis, we will use a two-stage step size schedule (Gower et al., 2019, Theorem 3.2):

$$\gamma_t = \begin{cases} \gamma_0 & \text{if } t \leq t_* \\ \frac{1}{\mu}\frac{2t+1}{(t+1)^2} & \text{if } t \geq t_* + 1 \end{cases} \ , \quad \text{where} \quad 0 < \gamma_0 \leq \frac{\mu}{2\mathcal{L}^2} \tag{5}$$

This operates by first maintaining a fixed step size $\gamma_0$ until some switching time $t_* \in \{0, \ldots, T\}$, and then switches to the $1/t$ schedule of Lacoste-Julien et al. (2012).

Under Assumptions 2.7 and 2.8, we can now provide a complexity guarantee for solving Eq. (1) via SPGD. Since BBVI consists of a subset of Eq. (1), establishing Assumptions 2.7 and 2.8 and invoking the following result will constitute our complexity guarantee for BBVI.

**Proposition 2.9.** *Suppose $f$ is $\mu$-strongly convex, $h$ satisfies Assumption 2.5, and $\widehat{\nabla f}$ satisfies Assumptions 2.7 and 2.8. Then, for the global optimum $\lambda_* = \arg\min_{\lambda \in \Lambda} F(\lambda)$ and $\Delta \triangleq \|\lambda_0 - \lambda_*\|_2$, there exists some $t_*$ and $\gamma_0$ such that SPGD with the step size schedule in Eq. (5) guarantees*

$$T \geq \mathrm{O}\left\{\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + \frac{\sigma\mathcal{L}}{\mu^2}\log\left(\frac{\mathcal{L}^2}{\sigma^2}\Delta^2\right)\frac{1}{\sqrt{\epsilon}} + \frac{\mathcal{L}^2}{\mu^2}\log\left(\Delta^2\frac{1}{\epsilon}\right) + 1\right\} \quad \Rightarrow \quad \mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon \ .$$

*Proof.* See the *full proof* in Appendix B.1.1, p. 26. □

This result is a slight improvement over past analysis of SPGD with Eq. (5) (Domke et al., 2023, Theorem 7). In particular, the dependence on the initialization $\Delta$ has been improved to be logarithmic instead of polynomial. Furthermore, it encompasses the case where we have "interpolation" ($\sigma^2 = 0$; Kim et al., 2024b; Schmidt and Roux, 2013; Vaswani et al., 2019) automatically resulting in a $\mathrm{O}(\log 1/\epsilon)$ complexity. The key difference in the analysis is that we choose a different switching time $t_*$ in a way adaptive to $\sigma^2$ and $\Delta$, ensuring that the dependence on both is optimized.

For a non-strongly convex $f$, using the strategy of Domke et al. (2023, Theorem 8 and 11) should yield a corresponding $\mathrm{O}(1/\epsilon^2)$ complexity guarantee under the same set of assumptions. However, this requires fixing the horizon $T$ in advance, and it is currently unknown how to obtain an anytime $\mathrm{O}(1/\sqrt{T})$ convergence bound for SGD under Assumptions 2.7 and 2.8 or QV. If one moves away from the canonical SGD update by incorporating Halpern iterations (Halpern, 1967), it is possible to obtain any-time convergence under a QV-like assumption (Alacaoglu et al., 2025).

# 3 Main Results

## 3.1 General Result

For our results, we impose an additional assumption that is a generalization of $L$-smoothness under twice differentiability of $\ell$.

**Assumption 3.1.** *$\ell$ is twice differentiable and, for all $z \in \mathbb{R}^d$, there exist some matrix $H \in \mathbb{R}^{d \times d}$ and constant $\delta \in [0, \infty)$ satisfying*

$$\|H\|_2 < \infty \quad \textit{and} \quad \|\nabla^2 \ell(z) - H\|_2 \leq \delta \ .$$

Notably, if $\ell$ is twice differentiable, $\mu$-strongly convex, $L$-smooth, it already satisfies Assumption 3.1 with $H = \frac{L+\mu}{2}\mathrm{I}_d$ and $\delta = \frac{L-\mu}{2}$. If $\ell$ is only $L$-smooth, it satisfies it with $H = 0_{d \times d}$ and $\delta = L$. The key advantage of this assumption, however, is that it characterizes Hessians that are not necessarily well-conditioned, but almost constant. This crucially affects the dimension dependence.

Given our assumptions on the target function $\ell$, variational family $\mathcal{Q}$, and our choice of gradient estimator, we can guarantee that SPGD applied to a problem structure corresponding to BBVI (Eq. (1)) achieves a given level of accuracy $\epsilon$ after $\mathrm{O}(g(d, H, \delta, \mu, \varphi)\epsilon^{-1})$ number of iterations:

**Theorem 3.2.** *Suppose the following hold:*

1. *$\ell$ is $\mu$-strongly convex and satisfies Assumption 3.1 and $\mu \leq \sigma_{\min}(H) \leq \sigma_{\max}(H) \leq L$.*
2. *$h$ satisfies Assumption 2.5.*
3. *$\mathcal{Q}$ is a mean-field location-scale family, where Assumption 2.2 holds.*
4. *$\widehat{\nabla f}$ is the reparametrization gradient.*

*Denote the global optimum $\lambda_* = (m_*, C_*) = \arg\min_{\lambda\in\Lambda} F(\lambda)$, the irreducible gradient noise as $\sigma_*^2 \triangleq \|m_* - \bar{z}\|_2^2 + \|C_*\|_{\mathrm{F}}^2$, and the stationary point of $\ell$ as $\bar{z} \triangleq \arg\min_{z\in\mathbb{R}^d} \ell(z)$. Then there exists some $t_*$ and $\gamma_0$ such that SPGD with the step size schedule in Eq. (5) guarantees*

$$T \geq \mathrm{O}\Big\{g(d, H, \delta, \mu, \varphi)\Big(\sigma_*^2\epsilon^{-1} + \sigma_* \log\big(\|\lambda_0 - \lambda_*\|_2^2\big)\epsilon^{-1/2}\Big)\Big\} \quad \Rightarrow \quad \mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon \,,$$

*where*

$$g(d, H, \delta, \mu, \varphi) \triangleq 2\,(1 + r_4)\big(\|H\|_2^2/\mu^2\big) + 4\big(\delta^2/\mu^2\big)\Big((1/2) + r_4 + \mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2\Big) \,.$$

*Proof.* The *full proof* can be found in Appendix B.2.1, p. 32. □

Due to the identity $\|\lambda - \lambda'\|_2^2 = \mathbb{E}_{\mathsf{u}\sim\varphi^{\otimes d}}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2$ (Lemma A.3), which is the squared cost of a coupling between $q_{\lambda_T}$ and $q_{\lambda_*}$, our guarantee also translates to a guarantee in Wasserstein-2 distance: $\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon \;\Rightarrow\; \mathbb{E}\mathrm{W}_2(q_{\lambda_T}, q_{\lambda_*})^2 \leq \epsilon$. In the general case where $\delta > 0$, the dimension dependence enters through $\mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2$, which depends on the order-statistics of the base distribution $\varphi$. In case $\ell$ is a quadratic, corresponding to $\pi$ being a Gaussian target distribution in the BBVI context, there exists some $H$ such that $\nabla^2\ell(z) = H$ for all $z \in \mathbb{R}^d$. Thus, Assumption 3.1 holds with $\delta = 0$, implying a dimension-independent convergence rate. We will present additional special cases with more explicit choices of $\varphi$ in the next section.

In case we do not want to assume Assumption 3.1 and only assume that $\ell$ is $\mu$-strongly convex and $L$-smooth instead, we can replace them with the generic choices of $H = \frac{L+\mu}{2}\mathrm{I}_d$ and $\delta = \frac{L-\mu}{2}$, which hold for all $\ell$s that are $\mu$-strongly convex, $L$-smooth, and twice differentiable. This then makes the role of the condition number $\kappa \triangleq L/\mu$ more explicit.

**Corollary 3.3.** *Suppose $\ell$ is is twice differentiable, $\mu$-strongly convex, and $L$-smooth. Then, denoting the condition number as $\kappa \triangleq L/\mu$, Theorem 3.2 holds with*

$$g\Big(d, \tfrac{L+\mu}{2}\mathrm{I}_d, \tfrac{L-\mu}{2}, \mu, \varphi\Big) = (1/2)(1 + r_4)(\kappa + 1)^2 + (\kappa - 1)^2\Big((1/2) + r_4 + \mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2\Big) \,.$$

This makes the $\mathrm{O}(\kappa^2)$ condition number dependence explicit, but the downside is that we lose dimension independence in the case of ill-conditioned quadratic $\ell$s. This fact suggests that dimension dependence is more fundamentally related to how close the Hessian is to a constant rather than how well-conditioned it is.

### 3.2 Special Cases with Benign Dimension Dependence

We now present some special cases of Theorem 3.2, which has yet to exhibit an explicit dependence on dimensionality. As mentioned in the previous section, dimension dependence depends on the order statistics of $\varphi$, which is related to the tail behavior of $\varphi$.

**Variational Families with Sub-Gaussian Tails.** The most commonly used variational family in practice is the Gaussian variational family. More broadly, for sub-Gaussian variational families, $\mathsf{u}_i^2$ is sub-exponential and therefore admits a moment generating function (MGF) (Wainwright, 2019, Theorem 2.6), which leads to a $\mathrm{O}(\log d)$ explicit dimension dependence.

**Proposition 3.4.** *Suppose there exists some $t > 0$ such that the MGF of $\mathsf{u}_i^2$ satisfies $M_{\mathsf{u}_i^2}(t) < \infty$. Then*

$$\mathbb{E}\max_{i=1,\ldots,d} \mathsf{u}_i^2 \quad \leq \quad (1/t)\left(\log M_{\mathsf{u}_i^2}(t) + \log d\right) \,.$$

*For example, if $\varphi$ is a standard Gaussian, then*

$$g(d, H, \delta, \mu, \varphi) \quad \leq \quad 8\big(\|H\|_2^2/\mu^2\big) + \big(\delta^2/\mu^2\big)(22 + 16\log d) \,.$$

*Proof.* The *full proof* can be found in Appendix B.2.2, p. 33. □

**Variational Families with Finite Higher Moments.** For families with tails heavier than sub-Gaussian, however, $u_i^2$ may not have an MGF. While we then lose the $\mathrm{O}(\log d)$ dependence, we may still obtain a polynomial dependence that can be better than $\mathrm{O}(\sqrt{d})$ obtained in previous works (Kim et al., 2023b). In particular, the result that will follow states that the highest order of the available moments determines the order of dimension dependence. For Student-$t$ families, this implies that using a high-enough degree of freedom $\nu$ can make the dimension dependence benign.

**Proposition 3.5.** *Suppose, for $k \geq 2$, the kth moment of $u_i^2$ is finite as $r_{2k} = \mathbb{E}u_i^{2k} < \infty$. Then*

$$\mathbb{E} \max_{i=1,\ldots,d} u_i^2 \quad \leq \quad \sqrt{2}\, d^{1/k}\, r_{2k}^{1/k} \,.$$

*For example, if $\varphi$ is a Student-t with $\nu > 4$ degrees of freedom and unit variance, then*

$$g(d, H, \delta, \mu, \varphi) \quad \leq \quad 8(\|H\|_2^2/\mu^2) + (\delta^2/\mu^2)\left(16 + \sqrt{2}\,\nu^3 d^{\frac{2}{\nu-2}}\right) .$$

*Proof.* See the *full proof* in Appendix B.2.3, p. 35. □

# 4 Analysis of Gradient Variance

## 4.1 Overview

The key technical contribution of this work is analyzing the gradient variance and thus establishing the constants $\mathcal{L}$ (Assumption 2.7) and $\sigma^2$ (Assumption 2.8), which boils down to analyzing

$$\begin{aligned}\mathbb{E}\|\widehat{\nabla f}(\lambda; u) - \widehat{\nabla f}(\lambda'; u)\|_2^2 &= \mathbb{E}\left\|\frac{\partial \mathcal{T}_\lambda(u)}{\partial \lambda}\nabla\ell(\mathcal{T}_\lambda(u)) - \frac{\partial \mathcal{T}_{\lambda'}(u)}{\partial \lambda'}\nabla\ell(\mathcal{T}_{\lambda'}(u))\right\|_2^2 \\ &= \mathbb{E}\left\|\frac{\partial \mathcal{T}_\lambda(u)}{\partial \lambda}(\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u)))\right\|_2^2 ,\end{aligned}$$

where the equality follows from the fact that the Jacobian $\partial\mathcal{T}_\lambda(u)/\partial\lambda$ does not depend on $\lambda$. For mean-field location-scale variational families, the squared Jacobian follows as

$$\left(\frac{\partial \mathcal{T}_\lambda(u)}{\partial \lambda}\right)^\top \frac{\partial \mathcal{T}_\lambda(u)}{\partial \lambda} = \mathrm{I}_d + U^2 ,$$

where $U \triangleq \mathrm{diag}(u_1, \ldots, u_d)$ (Kim et al., 2023b). This implies that

$$\begin{aligned}&\mathbb{E}\|\widehat{\nabla f}(\lambda; u) - \widehat{\nabla f}(\lambda'; u)\|_2^2 \\ &\quad = \underbrace{\mathbb{E}\|\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u))\|_2^2}_{\triangleq V_{\text{loc}}} + \underbrace{\mathbb{E}\|U(\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u)))\|_2^2}_{\triangleq V_{\text{scale}}} . \quad (6)\end{aligned}$$

Our goal is to bound each term by $\|\lambda - \lambda'\|_2^2$.

In order to solve the expectations, we need to simplify the $\nabla\ell$ terms. For instance, for the gradient of the location $V_{\text{loc}}$, assuming that $\ell$ is $L$-smooth allows for a quadratic approximation. That is,

$$V_{\text{loc}} \quad = \quad \mathbb{E}\|\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u))\|_2^2 \quad \leq \quad L^2\mathbb{E}\|\mathcal{T}_\lambda(u) - \mathcal{T}_{\lambda'}(u)\|_2^2 \quad = \quad L^2\|\lambda - \lambda'\|_2^2 ,$$

where the last equality is by Lemma A.3.

Now, it is tempting to use the same quadratic approximation strategy for the gradient of the scale $V_{\text{scale}}$. Indeed, this strategy was used by Domke (2019) to bound the gradient variance of full-rank location-scale variational families and by Ko et al. (2024) for structured location-scale variational families. Unfortunately, this strategy does not immediately apply to mean-field families due to the matrix $U$. We somehow have to decouple $\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u))$ and $U$, but in a way that does not lose the correlation between the two; the correlation leads to cancellations critical to obtaining a tight bound. Kim et al. (2023b) used the inequality

$$V_{\text{scale}} \quad \leq \quad \mathbb{E}\|U^2\|_{\mathrm{F}}\|\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u))\|_2^2 , \quad (7)$$

which resulted in a dimension dependence of $\mathrm{O}(r_4\sqrt{d})$ after solving the expectation. The key question is whether this dimension dependence can be improved. Due to the ordering of norms $\|\cdot\|_2 \leq \|\cdot\|_{\mathrm{F}}$, it is natural to consider the tighter inequality

$$V_{\text{scale}} \quad \leq \quad \mathbb{E}\|U\|_2^2\|\nabla\ell(\mathcal{T}_\lambda(u)) - \nabla\ell(\mathcal{T}_{\lambda'}(u))\|_2^2 .$$

(This step corresponds to Eq. (8) in the proof sketch of the upcoming result.) The main challenge, however, is solving the resulting expectation in a way that is also tight with respect to $d$. We will see that this requires a careful probabilistic analysis.

### 4.2 Upper Bound on Gradient Variance

We now formally state our upper bound on the gradient variance. In the context of proving Theorem 3.2, the following lemma implies both Assumption 2.7 and Assumption 2.2. (See the proof of Theorem 3.2.) We provide a corresponding unimprovability result in Section 4.3.

**Lemma 4.1.** *Suppose Assumptions 2.2 and 3.1 hold, $\mathcal{Q}$ is a mean-field location-family, and $\widehat{\nabla f}$ is the reparametrization gradient. Then, for any $\lambda, \lambda' \in \mathbb{R}^d \times \mathbb{D}^d$.*

$$\mathbb{E}\|\widehat{\nabla f}(\lambda; \mathsf{u}) - \widehat{\nabla f}(\lambda'; \mathsf{u})\|_2^2 \leq \left\{2(1 + r_4)\|H\|_2^2 + 4\delta^2\left(1/2 + r_4 + \mathbb{E}\max_{j=1,\dots,d} \mathsf{u}_j^2\right)\right\}\|\lambda - \lambda'\|_2^2 \,.$$

*Proof Sketch.* For the proof sketch, we will assume that $\ell$ is $L$-smooth instead of taking Assumption 3.1. This will vastly simplify the analysis and let us focus on the key elements.

Recall $V_{\text{scale}}$ in Eq. (6). Applying the operator norm and the $L$-smoothness of $\ell$ yields

$$V_{\text{scale}} \quad \leq \quad \mathbb{E}\|\mathsf{U}\|_2^2\|\nabla\ell(\mathcal{T}_\lambda(\mathsf{u})) - \nabla\ell(\mathcal{T}_{\lambda'}(\mathsf{u}))\|_2^2 \quad \leq \quad L^2\mathbb{E}\|\mathsf{U}\|_2^2\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2 \,. \tag{8}$$

It remains to solve the expectation over $\mathsf{u}_1, \dots, \mathsf{u}_d$. Denote

$$\lambda = (m, C), \qquad \lambda' = (m', C'), \qquad \bar{m} \triangleq m - m', \quad \text{and} \quad \bar{C} \triangleq C - C' \,,$$

recall that $\mathcal{T}_\lambda(u) = Cu + m$ (Definition 2.1), and notice that, since $\mathsf{U}$ is a diagonal matrix, $\|\mathsf{U}\|_2^2 = \max_{i=1,\dots,d} \mathsf{u}_j^2$. Then we can rewrite Eq. (8) as

$$\begin{aligned}
V_{\text{scale}} \quad &\leq \quad L^2\mathbb{E}\Big(\max_{i=1,\dots,d} \mathsf{u}_j^2\Big)\textstyle\sum_{i=1}^d (C_{ii}\mathsf{u}_i + m_i - C'_{ii}\mathsf{u}_i - m'_i)^2 \\
&= \quad L^2\mathbb{E}\Big(\max_{i=1,\dots,d} \mathsf{u}_j^2\Big)\textstyle\sum_{i=1}^d \big(\bar{C}_{ii}\mathsf{u}_i + \bar{m}_i\big)^2 \\
&\leq \quad L^2\mathbb{E}\Big(\max_{i=1,\dots,d} \mathsf{u}_j^2\Big)\textstyle\sum_{i=1}^d \big(2\bar{C}_{ii}^2\mathsf{u}_i^2 + 2\bar{m}_i^2\big) \,. \qquad \text{(Young's inequality)}
\end{aligned}$$

The problematic term is

$$\mathbb{E}\Big(\max_{j=1,\dots,d} \mathsf{u}_j^2\Big)\textstyle\sum_{i=1}^d \bar{C}_{ii}^2\mathsf{u}_i^2 \quad = \quad \mathbb{E}\mathsf{u}_{i_*}^2\sum_{i=1}^d \bar{C}_{ii}^2\mathsf{u}_i^2 \quad = \quad \mathbb{E}\Big[\bar{C}_{i_*i_*}^2\mathsf{u}_{i_*}^4 + \textstyle\sum_{j\neq i_*}^d \bar{C}_{jj}^2\mathsf{u}_{i_*}^2\mathsf{u}_j^2\Big] \,,$$

where $i_* = \arg\max_{i=1,\dots,d} \mathsf{u}_i^2$ is the coordinate of maximum magnitude. Here, $\mathsf{u}_{i_*}^4 = \max_{i=1,\dots,d} \mathsf{u}_i^4$ is a heavy-tailed quantity that generally grows fast in $d$, unlike $\mathsf{u}_{i_*}^2$. (*e.g.*, for a Gaussian $\mathsf{u}_i$, $\mathsf{u}_{i_*}^2$ has an MGF but $\mathsf{u}_{i_*}^4$ does not.) Therefore, a benign dimension dependence might appear futile. Notice, however, that the problematic term only affects a single dimension: the maximal axis indicated by $i_*$. A probabilistic analysis reveals that as $d$ increases, the effect of $\mathsf{u}_{i_*}^4$ becomes averaged out and the effect of the remaining term involving $\mathsf{u}_i^2$ dominates. More formally,

$$\mathbb{E}\mathsf{u}_{i_*}^2\textstyle\sum_{i=1}^d \bar{C}_{ii}^2\mathsf{u}_i^2 \quad = \quad \sum_{i=1}^d \bar{C}_{ii}^2\,\mathbb{E}\big[\mathsf{u}_{i_*}^4\mathbb{1}\{i_* = i\} + \mathsf{u}_{i_*}^2\mathsf{u}_i^2\mathbb{1}\{i_* \neq i\}\big] \,,$$

where

$$\mathbb{E}\big[\mathsf{u}_{i_*}^4\mathbb{1}\{i_* = i\}\big] \quad = \quad \mathbb{E}\big[\mathsf{u}_{i_*}^4\big]\mathbb{E}[\mathbb{1}\{i_* = i\}] \quad = \quad \mathbb{E}\big[\mathsf{u}_{i_*}^4\big]\mathbb{P}[i_* = i] \quad = \quad \mathbb{E}\big[\mathsf{u}_{i_*}^4\big](1/d) \,.$$

Since the maximum of $d$ random variable is always smaller than their sum, the probability of the maximally random event, $\mathbb{P}[i_* = i] = 1/d$, kills off the dimensional growth of $\mathsf{u}_{i_*}^4$. In fact, using the crude bound $\mathbb{E}\mathsf{u}_{i_*}^4 \leq \mathbb{E}\sum_{i=1}^d \mathsf{u}_i^4 = dr_4$, where the last equality is due to Assumption 2.2, is enough to make this term independent of $d$. The remaining dimension dependence comes from $\mathsf{u}_{i_*}^2$:

$$\mathbb{E}\big[\mathsf{u}_{i_*}^2\mathsf{u}_i^2\mathbb{1}\{i_* \neq i\}\big] \;=\; \mathbb{E}\Big[\max_{j\neq i} \mathsf{u}_j^2\mathsf{u}_i^2\mathbb{1}\{i_* \neq i\}\Big] \;\leq\; \mathbb{E}\Big[\max_{j=1,\dots,d-1} \mathsf{u}_j^2\Big]\mathbb{E}\big[\mathsf{u}_i^2\big] \;=\; \mathbb{E}\max_{j=1,\dots,d-1} \mathsf{u}_j^2 \,,$$

where the last equality follows from Assumption 2.2. Therefore, we finally obtain

$$\begin{aligned}
V_{\text{scale}} &\leq 2L^2\sum_{i=1}^d\left[\Big(\mathbb{E}\max_{j=1,\dots,d-1} \mathsf{u}_j^2 + r_4\Big)\bar{C}_{ii}^2 + \mathbb{E}\max_{j=1,\dots,d} \mathsf{u}_j^2\bar{m}_i^2\right] \\
&\leq 2L^2\Big(\mathbb{E}\max_{j=1,\dots,d} \mathsf{u}_j^2 + r_4\Big)\big(\|\bar{m}\|_2^2 + \|\bar{C}\|_{\text{F}}^2\big) \\
&= 2L^2\Big(\mathbb{E}\max_{j=1,\dots,d} \mathsf{u}_j^2 + r_4\Big)\|\lambda - \lambda'\|_2^2 \,.
\end{aligned}$$

The full proof performs an analogous analysis under the more general Assumption 3.1. □

See the *full proof* in Appendix B.3.1, p. 37.

### 4.3 Unimprovability

We also demonstrate a lower bound, which implies that Lemma 4.1 cannot be improved by the spectral bounds of $\nabla^2\ell$. From Eq. (6) and the fundamental theorem of calculus,

$$\mathbb{E}\|\widehat{\nabla f}(\lambda;\mathsf{u})\|_2^2 \quad \geq \quad \mathbb{E}\|\mathsf{U}(\nabla\ell\left(\mathsf{z}\right)-\nabla\ell\left(\bar{z}\right))\|_2^2 \quad = \quad \mathbb{E}\left\|\mathsf{U}\int_0^1\nabla^2\ell(\mathsf{z}^w)(\mathsf{z}-\bar{z})\mathrm{d}w\right\|_2^2, \tag{9}$$

where $\bar{z}\in\{z\mid\nabla\ell(z)=0\}$ is a stationary point of $\ell$, $\mathsf{z}\triangleq\mathcal{T}_\lambda(\mathsf{u})$, and $\mathsf{z}^w\triangleq w\mathsf{z}+(1-w)\bar{z}$. There exists a matrix-valued function with bounded singular values that lower-bounds this quantity:

**Proposition 4.2.** *Suppose Assumption 2.2 holds and $\mathcal{Q}$ is a mean-field location-scale family. Then, for any $t>0$, $d>0$, $\mu,L\in(0,+\infty)$ satisfying $\mu\leq L$, there exists a matrix-valued function $H(z):\mathbb{R}^d\to\mathbb{S}^d_{\succ0}$ satisfying $\mu\mathrm{I}_d\preceq H\preceq L\mathrm{I}_d$ almost surely and a set of parameters $\lambda=(m,C)\in\mathbb{R}^d\times\mathbb{D}^d_{\succ0}$ such that*

$$\mathbb{E}\left\|\mathsf{U}\int_0^1 H(\mathsf{z}^w)(\mathsf{z}-\bar{z})\mathrm{d}w\right\|_2^2\geq\left\{\tfrac{(L-\mu)^2}{4}-\tfrac{L^2}{2}\tfrac{\mathbb{E}\max_{i=1,\ldots,d}\mathsf{u}_i^4}{d}\right\}c(t,\varphi)\left\{\mathbb{E}\max_{i=1,\ldots,d-1}\mathsf{u}_i^2-t\right\}\|C\|_{\mathrm{F}}^2\,.$$

*where $c(t,\varphi)>0$ is a constant only dependent on $t$ and $\varphi$.*

*Proof.* The *full proof* can be found in Appendix B.3.4, p. 43. □

For Gaussians, $\mathbb{E}\max_i\mathsf{u}_i^4$ is upper bounded as $\mathrm{O}(\sqrt{d})$ (Gumbel, 1954, Eq. 1.6), which means the negative term vanishes at a $\mathrm{O}(1/\sqrt{d})$ rate. Furthermore, $\mathbb{E}\max_i\mathsf{u}_i^2\geq\left(\mathbb{E}\max_i\mathsf{u}_i\right)^2=\Omega(\log d)$ by the well-known lower bound on the expected maximum of i.i.d. Gaussians (Wainwright, 2019, Exercise 2.11.(b)). Combining these facts with Proposition 4.2 yield a $\Omega\left(L^2\log d\right)$ bound on Eq. (9).

**Remark 4.3.** *It is not obvious that the rows of our worst-case example $H_{\mathrm{worst}}$ form conservative vector fields. This means that Proposition 4.2 does not assert the existence of a function $\ell$ that satisfies $\nabla^2\ell=H_{\mathrm{worst}}$. However, it does suggest that one cannot improve Lemma 4.1 by relying only on spectral bounds on the Hessian.*

## 5 Discussion

### 5.1 Related Works

Early results analyzing VI had to rely on assumptions that either: (i) do not hold on Gaussian targets, (ii) are difficult to verify, or (iii) require bounds on the domain (Alquier and Ridgway, 2020; Buchholz et al., 2018; Fan et al., 2015; Fujisawa and Sato, 2021; Khan et al., 2016; Liu and Owen, 2021; Nguyen et al., 2025; Regier et al., 2017). coordinate-ascent VI (CAVI), in particular, was studied on specific models (Ghorbani et al., 2019; Zhang and Zhou, 2020) only. Under general and verifiable assumptions, Xu and Campbell (2022) obtained asymptotic convergence guarantees, while partial results, such as bounds on the gradient variance (Domke, 2019; Fan et al., 2015; Kim et al., 2023b), or regularity of the ELBO (Challis and Barber, 2013; Domke, 2020; Titsias and Lázaro-Gredilla, 2014), were known.

It was only recently that non-asymptotic quantitative convergence under realizable and verifiable assumptions was established. For BBVI specifically, Hoffman and Ma (2020) first proved convergence on Gaussian targets (quadratic $\ell$), while Domke et al. (2023); Kim et al. (2023a, 2024b) proved the first results on strongly convex and smooth functions with location-scale families. Surendran et al. (2025) extended these results to non-convex smooth functions and more complex variational family parametrizations, and Cheng et al. (2024) analyzed a variant of semi-implicit VI. The results of Domke et al.; Hoffman and Ma; Kim et al., who focused on full-rank families, suggest a $\mathrm{O}(d)$ dimension dependence in the iteration complexity. On the other hand, Kim et al. (2023a) reported a $\mathrm{O}(\sqrt{d})$ dimension dependence for mean-field location-scale families, while conjecturing $\mathrm{O}(\log d)$ dependence, based on the partial result of Kim et al. (2023b). For targets with a diagonal Hessian structure, Ko et al. (2024, Corollary 1) show that mean-field families are dimension-independent.

Apart from BBVI, Wasserstein VI algorithms—which minimize the KL divergence on the Wasserstein geometry—provide non-asymptotic convergence guarantees. In particular, the algorithms by Diao et al. (2023); Lambert et al. (2022) optimize over the full-rank Gaussian family, while that of Jiang et al. (2025) optimizes over all mean-field families with bounded second moments. To guarantee $\mathbb{E}\mathrm{W}_2(q_{\lambda_T},q_{\lambda_*})^2\leq\epsilon$ on strongly log-concave and log-smooth targets, they all report an iteration complexity of $\mathrm{O}(d\epsilon^{-1}\log\epsilon^{-1})$. Meanwhile, under the same conditions, Arnese and Lacker

(2024); Lavenant and Zanella (2024) analyzed (block) CAVI, and reported an iteration complexity of $\mathrm{O}(d \log \epsilon^{-1})$. Bhattacharya et al. (2025) provides a concurrent result on CAVI, but relies on an assumption that departs from log-concavity and smoothness. Finally, Bhatia et al. (2022) analyzes a specialized algorithm optimizing over only the scale of Gaussians, which has a gradient query complexity of $\mathrm{O}(dk\epsilon^{-3})$, where $k$ is the user-chosen number of rank-1 factors in the scale matrix.

### 5.2 Conclusion

In this work, we proved that BBVI with mean-field location-scale families is able to converge with an iteration complexity with only a $\mathrm{O}(\log d)$ dimension dependence, as long as the tails of the family are sub-Gaussian. For high-dimensional targets, this suggests a substantial speed advantage over BBVI with full-rank families. In practice, the mean-field approximation can be combined with other design elements such as control variates (Boustati et al., 2020; Geffner and Domke, 2018, 2020; Miller et al., 2017; Roeder et al., 2017; Wang et al., 2024) and data-point subsampling (Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014). Our analysis strategy should easily be combined with existing analyses (Kim et al., 2024a,b) for such design elements.

For a target distribution $\pi$ with a condition number of $\kappa$ and a target accuracy level $\epsilon$, we now know how to improve the dependence on $d$ and $\epsilon$ in the iteration complexity: Using less-expressive families such as mean-field (Theorem 3.2) or structured (Ko et al., 2024) families improves the dependence on $d$, while applying control variates to gradient estimators (Kim et al., 2024b) improves the dependence on $\epsilon$. However, it is currently unclear whether the dependence on $\kappa$ is tight or improvable. If it is tight, it would be worth investigating whether this can be provably improved through algorithmic modifications, for example, via stochastic second-order optimization methods (Byrd et al., 2016; Fan et al., 2015; Liu and Owen, 2021; Meng et al., 2020; Regier et al., 2017).

Another future direction would be to develop methods that are able to adaptively adjust the computational cost between $\mathrm{O}(\log d)$ and $\mathrm{O}(d)$ by trading statistical accuracy akin to the method of Bhatia et al. (2022). Existing BBVI schemes with "low-rank(-plus-diagonal)" families (Ong et al., 2018; Rezende et al., 2014; Tomczak et al., 2020) result in a non-smooth, non-Lipschitz, and non-convex landscape. This not only rules out typical theoretical convergence guarantees but also exhibits unstable and slow convergence in practice (Modi et al., 2025). Furthermore, understanding the statistical side of this trade-off will be an important direction. As of now, our understanding is restricted to either mean-field or full-rank families (Katsevich and Rigollet, 2024; Margossian and Saul, 2023, 2025; Wang and Blei, 2019a,b; Yang et al., 2020; Zhang and Gao, 2020) with little in between except for the work of Bhatia et al. (2022).

## Acknowledgments and Disclosure of Funding

The authors thank Anton Xue for helpful discussions and the reviewers for helpful suggestions.

K. Kim, J. R. Gardner were supported through the NSF award [IIS2145644]; Y.-A. Ma was supported by the NSF Award CCF-2112665 (TILOS), the DARPA AIE program, and the CDC-RFA-FT-23-0069; T. Campbell was supported by the NSERC Discovery Grant RGPIN-2025-04208.

## References

Abhinav Agrawal, Daniel R Sheldon, and Justin Domke. Advances in black-box VI: Normalizing flows, importance weighting, and optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 17358–17369. Curran Associates, Inc., 2020. (page 1)

Ahmet Alacaoglu, Yura Malitsky, and Stephen J. Wright. Towards weaker variance assumptions for stochastic optimization. arXiv Preprint arXiv:2504.09951, 2025. (page 5)

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, 2020. (page 9)

Manuel Arnese and Daniel Lacker. Convergence of coordinate ascent variational inference for log-concave measures via optimal transport. arXiv Preprint arXiv:2404.08792, 2024. (pages 3, 9)

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459. Curran Associates, Inc., 2011. (page 5)

Kush Bhatia, Nikki Lijing Kuang, Yi-An Ma, and Yixin Wang. Statistical and computational trade-offs in variational inference: A case study in inferential model selection. arXiv Preprint arXiv:2207.11208, 2022. (pages 2, 10)

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. On the convergence of coordinate ascent variational inference. *The Annals of Statistics*, 53(3):929–962, 2025. (page 10)

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019. (pages 1, 3)

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

Léon Bottou. On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, pages 9–42. Cambridge University Press, 1 edition, 1999. (page 4)

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. (pages 1, 4)

Ayman Boustati, Sattar Vakili, James Hensman, and S. T. John. Amortized variance reduction for doubly stochastic objective. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 124 of *PMLR*, pages 61–70. JMLR, 2020. (page 10)

Alexander Buchholz, Florian Wenzel, and Stephan Mandt. Quasi-Monte Carlo variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pages 668–677. JMLR, 2018. (page 9)

R. H. Byrd, S. L. Hansen, Jorge Nocedal, and Y. Singer. A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016. (page 10)

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. (pages 1, 3)

George Casella and Roger L. Berger. *Statistical Inference*. Cengage Learning, 2 edition, 2001. (page 3)

Edward Challis and David Barber. Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, 14(68):2239–2286, 2013. (page 9)

Ziheng Cheng, Longlin Yu, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Kernel semi-implicit variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *PMLR*, pages 8248–8269. JMLR, 2024. (page 9)

Sinho Chewi. *Log-Concave Sampling*. Unpublished draft, november 3, 2024 edition, 2024. URL https://chewisinho.github.io/main.pdf. (page 3)

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *PMLR*, pages 7960–7991. JMLR, 2023. (pages 3, 9)

Justin Domke. Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 32, pages 329–338. Curran Associates, Inc., 2019. (pages 2, 5, 7, 9)

Justin Domke. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pages 2587–2596. JMLR, 2020. (pages 3, 4, 9, 32)

Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 66289–66327. Curran Associates, Inc., 2023. (pages 3, 4, 5, 9, 26)

Kai Fan, Ziteng Wang, Jeff Beck, James Kwok, and Katherine A Heller. Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, volume 28, pages 1387–1395. Curran Associates, Inc., 2015. (pages 9, 10)

Tor Erlend Fjelde, Kai Xu, David Widmann, Mohamed Tarek, Cameron Pfiffer, Martin Trapp, Seth D. Axen, Xianda Sun, Markus Hauru, Penelope Yong, Will Tebbutt, Zoubin Ghahramani, and Hong Ge. Turing.jl: A general-purpose probabilistic programming language. *ACM Transactions on Probabilistic Machine Learning*, 1(3):1–48, 2025. (pages 1, 3)

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

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. arXiv Preprint arXiv:2301.11235, 2023. (pages 5, 29)

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: A language for flexible probabilistic inference. In *Proceedings of the International Conference on Machine Learning*, volume 84 of *PMLR*, pages 1682–1690. JMLR, 2018. (pages 1, 3)

Tomas Geffner and Justin Domke. Using large ensembles of control variates for variational inference. In *Advances in Neural Information Processing Systems*, volume 31, pages 9960–9970. Curran Associates, Inc., 2018. (page 10)

Tomas Geffner and Justin Domke. Approximation based variance reduction for reparameterization gradients. In *Advances in Neural Information Processing Systems*, volume 33, pages 2397–2407. Curran Associates, Inc., 2020. (page 10)

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pages 2221–2231. JMLR, 2019. (page 9)

Ryan Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018. (page 1)

Ryan Giordano, Martin Ingram, and Tamara Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *Journal of Machine Learning Research*, 25:1–39, 2024. (page 1)

Paul Glasserman. *Gradient Estimation via Perturbation Analysis*. Number 116 in The Springer International Series in Engineering and Computer Science. Springer, New York, NY, 1991. (page 4)

Peter W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990. (page 4)

Eduard Gorbunov, Filip Hanzely, and Peter Richtarik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 680–690. JMLR, 2020. (page 29)

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pages 5200–5209. JMLR, 2019. (pages 5, 29)

E. J. Gumbel. The maxima of the mean largest value and of the range. *The Annals of Mathematical Statistics*, 25(1):76–84, 1954. (page 9)

Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967. (page 5)

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 5–13. ACM Press, 1993. (pages 1, 3)

Y. C. Ho and X. Cao. Perturbation analysis and optimization of queueing networks. *Journal of Optimization Theory and Applications*, 40(4):559–582, 1983. (page 4)

Matthew Hoffman and Yian Ma. Black-box variational inference as a parametric approximation to Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pages 4324–4341. JMLR, 2020. (page 9)

Alexandra Maria Hotti, Lennart Alexander Van der Goten, and Jens Lagergren. Benefits of non-linear scale parameterizations in black box variational inference through smoothness results and gradient variance bounds. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR*, pages 3538–3546. JMLR, 2024. (page 4)

Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Foundations of Computational Mathematics*, 2025. (page 9)

Norman L. Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. Continuous univariate distributions. volume 1 of *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York, 2 edition, 1994. (page 33)

Norman L. Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. Continuous univariate distributions. volume 2 of *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York, 2 edition, 1995. (pages 33, 35, 36)

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

Anya Katsevich and Philippe Rigollet. On the approximation accuracy of Gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409, 2024. (page 10)

Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M. Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *Journal of Optimization Theory and Applications*, 199:499–540, 2023. (page 29)

Mohammad Emtiyaz Khan, Reza Babanezhad, Wu Lin, Mark Schmidt, and Masashi Sugiyama. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 319–328, Jersey City, New Jersey, USA, 2016. AUAI Press. (page 9)

Kyurae Kim, Jisu Oh, Kaiwen Wu, Yian Ma, and Jacob R. Gardner. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 44615–44657. Curran Associates Inc., 2023a. (pages 2, 3, 4, 9, 29)

Kyurae Kim, Kaiwen Wu, Jisu Oh, and Jacob R. Gardner. Practical and matching gradient variance bounds for black-box variational Bayesian inference. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *PMLR*, pages 16853–16876. JMLR, 2023b. (pages 4, 7, 9)

Kyurae Kim, Joohwan Ko, Yi-An Ma, and Jacob R. Gardner. Demystifying SGD with doubly stochastic gradients. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *PMLR*, pages 24210–24247. JMLR, 2024a. (page 10)

Kyurae Kim, Yian Ma, and Jacob R. Gardner. Linear convergence of black-box variational inference: Should we stick the landing? In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR*, pages 235–243. JMLR, 2024b. (pages 5, 9, 10)

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, Banff, AB, Canada, 2014. (pages 2, 4)

Joohwan Ko, Kyurae Kim, Woo Chang Kim, and Jacob R. Gardner. Provably scalable black-box variational inference with structured variational families. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *PMLR*, pages 24896–24931. JMLR, 2024. (pages 1, 2, 7, 9, 10, 29)

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. (pages 1, 2, 3, 4, 10)

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. (pages 1, 2)

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $\mathrm{O}(1/t)$ convergence rate for the projected stochastic subgradient method. arXiv Preprint arXiv:1212.2002, 2012. (page 5)

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, volume 35, pages 14434–14447. Curran Associates, Inc., 2022. (pages 3, 9)

Hugo Lavenant and Giacomo Zanella. Convergence rate of random scan coordinate ascent variational inference under log-concavity. *SIAM Journal on Optimization*, 34(4):3750–3761, 2024. (pages 3, 10)

Sifan Liu and Art B. Owen. Quasi-Monte Carlo quasi-Newton in variational Bayes. *Journal of Machine Learning Research*, 22(243):1–23, 2021. (pages 9, 10)

Charles C. Margossian and Lawrence K. Saul. The shrinkage-delinkage trade-off: An analysis of factorized Gaussian approximations for variational inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 216 of *PMLR*, pages 1358–1367. JMLR, 2023. (page 10)

Charles C. Margossian and Lawrence K. Saul. Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258 of *PMLR*, pages 3466–3474. JMLR, 2025. (page 10)

Si Yi Meng, Sharan Vaswani, Issam Hadj Laradji, Mark Schmidt, and Simon Lacoste-Julien. Fast and furious convergence: Stochastic second order methods under interpolation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 1375–1386. JMLR, 2020. (page 10)

Andrew Miller, Nick Foti, Alexander D' Amour, and Ryan P Adams. Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, volume 30, pages 3708–3718. Curran Associates, Inc., 2017. (page 10)

Chirag Modi, Diana Cai, and Lawrence K. Saul. Batch, match, and patch: Low-rank approximations for score-based variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258 of *PMLR*, pages 4510–4518. JMLR, 2025. (page 10)

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020. (page 4)

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. (page 4)

Dai Hai Nguyen, Tetsuya Sakurai, and Hiroshi Mamitsuka. Wasserstein gradient flow over variational parameter space for variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistic*, volume 258 of *PMLR*, pages 1756–1764. JMLR, 2025. (page 9)

Victor M.-H. Ong, David J. Nott, and Michael S. Smith. Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27(3):465–478, 2018. (pages 1, 10)

Neal Parikh and Stephen P. Boyd. *Proximal Algorithms*, volume 1 of *Foundations and Trends® in Optimization*. Now Publishers, Norwell, MA, 2014. (page 4)

Anand Patil, David Huard, and Christopher Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4):1–81, 2010. (pages 1, 3)

Carsten Peterson and Eric Hartman. Explorations of the mean field theory learning algorithm. *Neural Networks*, 2(6):475–494, 1989. (pages 1, 3)

Georg Pflug. *Optimization of Stochastic Models: The Interface between Simulation and Optimization*. Number v.373 in The Springer International Series in Engineering and Computer Science Ser. Springer, New York, NY, 1996. (page 4)

Rana. Answer to "A bound for the expectation of the maximum independent random variables", 2017. URL https://math.stackexchange.com/q/2177201. (page 34)

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 33 of *PMLR*, pages 814–822. JMLR, 2014. (page 1)

Jeffrey Regier, Michael I Jordan, and Jon McAuliffe. Fast black-box variational inference through stochastic trust-region optimization. In *Advances in Neural Information Processing Systems*, volume 30, pages 2399–2408. Curran Associates, Inc., 2017. (pages 9, 10)

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pages 1278–1286. JMLR, 2014. (pages 1, 2, 4, 10)

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. (pages 1, 4)

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30, pages 6928–6937. Curran Associates, Inc., 2017. (page 10)

Reuven Y. Rubinstein. Sensitivity analysis of discrete event systems by the "push out" method. *Annals of Operations Research*, 39(1):229–250, 1992. (page 4)

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. arXiv Preprint arXiv:1308.6370, 2013. (page 5)

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011. (page 4)

Sobihan Surendran, Antoine Godichon-Baggioni, and Sylvain Le Corff. Theoretical convergence guarantees for variational autoencoders. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258 of *PMLR*. JMLR, 2025. (page 9)

Symbol-1. Answer to "Meaningful lower-bound of $\sqrt{a^2+b}-a$ when $a \gg b > 0$", 2022. URL https://math.stackexchange.com/q/4360503. (page 28)

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pages 1971–1979. JMLR, 2014. (pages 1, 2, 3, 4, 9, 10)

Marcin Tomczak, Siddharth Swaroop, and Richard Turner. Efficient low rank Gaussian variational inference for neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 4610–4622. Curran Associates, Inc., 2020. (pages 1, 10)

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997. (page 43)

Sharan Vaswani, Francis Bach, and Mark Schmidt. 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*, pages 1195–1204. JMLR, 2019. (page 5)

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, NY, 1st ed edition, 2019. (pages 6, 9)

Xi Wang, Tomas Geffner, and Justin Domke. Joint control variate for faster black-box variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR*, pages 1639–1647. JMLR, 2024. (page 10)

Yixin Wang and David Blei. Variational bayes under model misspecification. In *Advances in Neural Information Processing Systems*, volume 32, pages 13357–13367. Curran Associates, Inc., 2019a. (page 10)

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019b. (page 10)

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. (page 4)

David Wingate and Theophane Weber. Automated variational inference in probabilistic programming. arXiv Preprint arXiv:1301.1299, 2013. (page 1)

Ming Xu, Matias Quiroz, Robert Kohn, and Scott A. Sisson. Variance reduction properties of the reparameterization trick. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89 of *PMLR*, pages 2711–2720. JMLR, 2019. (page 4)

Zuheng Xu and Trevor Campbell. The computational asymptotics of Gaussian variational inference and the Laplace approximation. *Statistics and Computing*, 32(4), 2022. (page 9)

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. $\alpha$-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020. (page 10)

Anderson Y. Zhang and Harrison H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, 48(5):2575–2598, 2020. (page 9)

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, 2020. (page 10)

Lu Zhang, Bob Carpenter, Andrew Gelman, and Aki Vehtari. Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49, 2022. (page 1)

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we obtain a $\mathrm{O}((\log d)\kappa^2\epsilon^{-1})$ iteration complexity result for BBVI on strong log-concave and log-smooth target distributions. This corresponds to a nearly dimension-independent iteration complexity.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As stated in Remark 4.3, the main limitation of our work is that the unimprovability result Proposition 4.2 does not fully assert that Lemma 4.1 is tight for any target function $\ell$. It only shows that our specific proof strategy, which uses only spectral bounds on the Hessian of $\ell$, is unimprovable. In principle, using additional properties of the Hessian could result in a tighter bound.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are stated in either Sections 2 and 3 or the propositional statements.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not contain any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not contain any experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not contain any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not contain any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not contain any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines: The content of the paper is a theoretical study of an inference algorithm and does not involve real data.

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The content of the paper is a theoretical study of an inference algorithm and does not have direct societal consequences.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve real data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not involve real data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not involve real data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: Part of the supporting results were obtained after some minor interaction with LLMs. However, all of the proofs were written and proofread by humans. Therefore, LLMs did not play an important, original role nor did they contribute any non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

## Contents

## A Auxiliary Lemmas

**Lemma A.1.** *Suppose Assumption 2.2 holds. Then* $r_4 = \mathbb{E}u_i^4 \geq 1$.

*Proof.* By Jensen's inequality $\mathbb{E}u_i^4 \geq \left(\mathbb{E}u_i^2\right)^2$. Lastly, $\mathbb{E}u_i^4 \geq \left(\mathbb{E}u_i^2\right)^2 = 1$ by Assumption 2.2. □

**Lemma A.2.** *Suppose Assumption 2.2 holds and denote* $U = \operatorname{diag}(u_1, \ldots, u_d)$. *Then we have the following identities: (i)* $\mathbb{E}uu^\top = \mathrm{I}_d$, *(ii)* $\mathbb{E}U^2 = \mathrm{I}_d$.

*Proof.* From Assumption 2.2, we know that $\mathbb{E}u_i^2 = 1$. Then (i) follows from

$$[\mathbb{E}uu^\top]_{ij} = \mathbb{E}u_i u_j \quad = \quad \begin{cases} \mathbb{E}u_i^2 & \text{if } i = j \\ \mathbb{E}u_i \mathbb{E}u_j & \text{if } i \neq j \end{cases} \quad = \quad \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .$$

For (ii), we only need to focus on the diagonal since the off-diagonal is already zero.

$$[\mathbb{E}U^2]_{ii} \quad = \quad [\mathbb{E}\operatorname{diag}(u_1, \ldots, u_d)^2]_{ii} \quad = \quad \mathbb{E}u_i^2 \quad = \quad 1 .$$

□

**Lemma A.3.** *Suppose Assumption 2.2 holds,* $\mathcal{T}_\lambda$ *is the reparametrization function for a location-scale family, and the linear parametrization is used. Then*

$$\mathbb{E}\|\mathcal{T}_\lambda(u) - \mathcal{T}_{\lambda'}(u)\|_2^2 = \|\lambda - \lambda'\|_2^2 .$$

*Proof.* Denoting $\lambda = (m, C)$ and $\lambda' = (m', C')$,

$$\begin{aligned}
&\mathbb{E}\|\mathcal{T}_\lambda(u) - \mathcal{T}_{\lambda'}(u)\|_2^2 \\
&= \mathbb{E}\|(Cu + m) - (C'u - m')\|_2^2 \\
&= \mathbb{E}\|(C - C')u + (m - m')\|_2^2 \\
&= \mathbb{E}\|(C - C')u\|_2^2 + 2\langle (C - C')\mathbb{E}u,\, m - m'\rangle + \mathbb{E}\|m - m'\|_2^2 \\
&= \mathbb{E}\|(C - C')u\|_2^2 + \mathbb{E}\|m - m'\|_2^2 . \qquad \text{(Assumption 2.2)} \qquad (10)
\end{aligned}$$

Lastly,

$$\begin{aligned}
\mathbb{E}\|(C - C')u\|_2^2 &= \mathbb{E}u^\top (C - C')^\top (C - C')u \\
&= \mathbb{E}\operatorname{tr} u^\top (C - C')^\top (C - C')u \\
&= \operatorname{tr}(C - C')^\top (C - C')\mathbb{E}uu^\top \qquad \text{(cyclic property of trace)} \\
&= \operatorname{tr}(C - C')^\top (C - C')\mathrm{I} \qquad \text{(Lemma A.2)} \\
&= \operatorname{tr}(C - C')^\top (C - C') \\
&= \|C - C'\|_{\mathrm{F}}^2 .
\end{aligned}$$

Combining this with Eq. (10) yields the result. □

**Lemma A.4.** *Suppose* $f$ *is* $\mu$*-strongly convex and Assumption 2.7 holds. Then* $f$ *is* $L$*-Lipschitz smooth, while the constants satisfy the ordering*

$$\mu \quad \leq \quad L \quad \leq \quad \mathcal{L} .$$

*Proof.* For all $\lambda, \lambda' \in \Lambda$, the unbiasedness of $\widehat{\nabla f}$ and Jensen's inequality states that

$$\|\nabla f(\lambda) - \nabla f(\lambda')\|_2^2 \;=\; \|\mathbb{E}\widehat{\nabla f}(\lambda; u) - \mathbb{E}\widehat{\nabla f}(\lambda'; u)\|_2^2 \;\leq\; \mathbb{E}\|\widehat{\nabla f}(\lambda; u) - \widehat{\nabla f}(\lambda'; u)\|_2^2 .$$

Then the $\mu$-strong convexity of $f$ and Assumption 2.7 yields the inequality

$$\mu^2\|\lambda - \lambda'\|_2^2 \;\leq\; \|\nabla f(\lambda) - \nabla f(\lambda')\|_2^2 \;\leq\; \mathbb{E}\|\widehat{\nabla f}(\lambda; u) - \widehat{\nabla f}(\lambda'; u)\|_2^2 \;\leq\; \mathcal{L}^2\|\lambda - \lambda'\|_2^2 ,$$

from which the statement follows immediately. □

# B Proofs

## B.1 Proofs of Results in Section 2

### B.1.1 Proof of Proposition 2.9

Under the stated assumptions, we first establish a convergence bound which bounds $\mathbb{E}\|\lambda_T - \lambda_*\|_2^2$ after $T$ iterations under a given step size schedule. We will invert this convergence bound into a complexity guarantee by identifying the conditions on $T$, $t_*$, and $\gamma_0$ that guarantee $\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon$ for a given $\epsilon > 0$.

**Lemma B.1.** *Suppose $f$ is $\mu$-strongly convex, $h$ satisfies Assumption 2.5, and $\nabla f$ satisfies Assumptions 2.7 and 2.8. Then, for the global optimum $\lambda_* = \arg\min_{\lambda \in \Lambda} F(\lambda)$, any $t_*$ satisfying $4\mathcal{L}^2/\mu^2 \leq t_* \leq T$, and the step size schedule in Eq. (5), the contraction coefficient $\rho \triangleq 1 - \mu\gamma_0$ satisfies $\rho \in (0, 1)$ and the last iterate of SPGD after $T$ iterations, $\lambda_T$, satisfies*

$$\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \|\lambda_0 - \lambda_*\|_2^2 \, \rho^{t_*} \left(\frac{{t_*}^2}{T^2}\right) + 2\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + \frac{8\sigma^2}{\mu^2} \frac{T - t_*}{T^2} \, .$$

*Proof.* The *full proof* is deferred to Appendix B.1.2, p. 29. □

This is a slight generalization of a result (Domke et al., 2023, Theorem 7), where the switching time $t_*$ was fixed to some $t_* \propto \mathcal{L}/\mu$. While the choice of $t_* \propto \mathcal{L}/\mu$ results in the typical $\mathrm{O}(1/\epsilon)$ asymptotic complexity, it suffers from a suboptimal polynomial dependence on the initialization error $\Delta = \|\lambda_0 - \lambda_*\|_2$. Picking an alternative $t_*$, which is what we do in the proof, improves the iteration complexity to $\mathrm{O}\big(1/\epsilon + 1/\sqrt{\epsilon} \log \Delta^2 + \log(\Delta^2/\epsilon)\big)$.

**Proposition 2.9.** *Suppose $f$ is $\mu$-strongly convex, $h$ satisfies Assumption 2.5, and $\widehat{\nabla f}$ satisfies Assumptions 2.7 and 2.8. Then, for the global optimum $\lambda_* = \arg\min_{\lambda \in \Lambda} F(\lambda)$ and $\Delta \triangleq \|\lambda_0 - \lambda_*\|_2$, there exists some $t_*$ and $\gamma_0$ such that SPGD with the step size schedule in Eq. (5) guarantees*

$$T \geq \mathrm{O}\left\{ \frac{\sigma^2}{\mu^2} \frac{1}{\epsilon} + \frac{\sigma\mathcal{L}}{\mu^2} \log\left(\frac{\mathcal{L}^2}{\sigma^2}\Delta^2\right) \frac{1}{\sqrt{\epsilon}} + \frac{\mathcal{L}^2}{\mu^2} \log\left(\Delta^2 \frac{1}{\epsilon}\right) + 1 \right\} \quad \Rightarrow \quad \mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon \, .$$

*Proof.* Since $f$ is strongly convex and $h$ is convex, $F$ is also strongly convex. This implies that, by the property of strictly convex functions, $F$ has a unique global optimum, which we denote as $\lambda_*$.

From Lemma B.1, we have

$$\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \|\lambda_0 - \lambda_*\|_2^2 \, \rho^{t_*} \left(\frac{{t_*}^2}{T^2}\right) + 2\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + \frac{8\sigma^2}{\mu^2} \frac{T - t_*}{T^2} \, , \tag{11}$$

where $\rho = 1 - \gamma_0\mu$. We will optimize the upper bound over the parameters $t_*$, $\gamma_0$, and $T$ so that we can ensure the $\epsilon$-accuracy guarantee $\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon$.

Consider the choice

$$t_* = \min\left\{ \left\lceil \frac{1}{\log 1/\rho} \log\left(\frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) \right\rceil, T \right\} \quad \text{and} \quad \gamma_0 = \frac{\mu}{2\mathcal{L}^2} \, . \tag{12}$$

Using this, we will separately analyze the total error in Eq. (11) for the cases of $t_* = T$ and $t_* \neq T$.

The case $t_* = T$ happens only if

$$\left\lceil \frac{1}{\log 1/\rho} \log\left(\frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) \right\rceil \geq T$$

is true. Then an immediate implication is that

$$\begin{aligned} & \frac{1}{\log 1/\rho} \log\left(\frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) + 1 \geq T \\ \Leftrightarrow \quad & \log\left(\frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) \geq (\log 1/\rho)(T-1) \\ \Leftrightarrow \quad & \frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2 \geq \rho^{-(T-1)} \end{aligned}$$

$$\Leftrightarrow \qquad \|\lambda_0 - \lambda_*\|_2^2 \rho^{T-1} \geq 2\gamma_0 \frac{\sigma^2}{\mu} \; . \tag{13}$$

Considering this fact, Eq. (11) becomes

$$\begin{aligned}
\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 &\leq \|\lambda_0 - \lambda_*\|_2^2 \, \rho^T + 2\gamma_0 \frac{\sigma^2}{\mu} && (t_* = T) \\
&\leq \|\lambda_0 - \lambda_*\|_2^2 \, \rho^T + \|\lambda_0 - \lambda_*\|_2^2 \rho^{T-1} && \text{(Eq. (13))} \\
&\leq 2\|\lambda_0 - \lambda_*\|_2^2 \rho^{T-1} \; . && (\rho < 1)
\end{aligned}$$

The number of required steps for achieving the $\epsilon$-accuracy requirement follows from

$$\begin{aligned}
& 2\|\lambda_0 - \lambda_*\|_2^2 \rho^{T-1} \leq \epsilon \\
\Leftrightarrow \quad & 2\|\lambda_0 - \lambda_*\|_2^2 \frac{1}{\epsilon} \leq (1/\rho)^{T-1} \\
\Leftrightarrow \quad & \log\left(2\|\lambda_0 - \lambda_*\|_2^2 \frac{1}{\epsilon}\right) \leq (T-1)\log(1/\rho) \\
\Leftrightarrow \quad & \frac{1}{\log 1/\rho} \log\left(2\|\lambda_0 - \lambda_*\|_2^2 \frac{1}{\epsilon}\right) \leq T - 1 \\
\Leftarrow \quad & \frac{1}{1-\rho} \log\left(2\|\lambda_0 - \lambda_*\|_2^2 \frac{1}{\epsilon}\right) \leq T - 1 && (\log(1/\rho) \geq 1 - \rho) \\
\Leftrightarrow \quad & \frac{2\mathcal{L}^2}{\mu^2} \log\left(2\|\lambda_0 - \lambda_*\|_2^2 \frac{1}{\epsilon}\right) + 1 \leq T && (1 - \rho = \gamma_0 \mu = \mu^2/(2\mathcal{L}^2))
\end{aligned} \tag{14}$$

For the case $t_* \neq T$,

$$\begin{aligned}
t_* &= \left\lceil \frac{1}{\log 1/\rho} \log\left(\frac{\mu}{2\gamma_0 \sigma^2} \|\lambda_0 - \lambda_*\|_2^2\right) \right\rceil \\
&\geq \frac{1}{\log 1/\rho} \log\left(\frac{\mu}{2\gamma_0 \sigma^2} \|\lambda_0 - \lambda_*\|_2^2\right) \\
&= \frac{1}{\log \rho} \log\left(\frac{2\gamma_0 \sigma^2}{\mu} \frac{1}{\|\lambda_0 - \lambda_*\|_2^2}\right) .
\end{aligned} \tag{15}$$

This implies

$$\rho^{t_*} \leq \frac{2\gamma_0 \sigma^2}{\mu} \frac{1}{\|\lambda_0 - \lambda_*\|_2^2} \; .$$

Substituting for this in Eq. (11),

$$\begin{aligned}
\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 &\leq 2\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + 2\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + 8 \frac{\sigma^2}{\mu^2} \frac{T - t_*}{T^2} \\
&= 4\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + 8 \frac{\sigma^2}{\mu^2} \frac{T - t_*}{T^2} \\
&\leq 4\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + 8 \frac{\sigma^2}{\mu^2} \frac{1}{T} \\
&= a \frac{1}{T^2} + b \frac{1}{T} \; ,
\end{aligned}$$

which is a quadratic function of $1/T$ with the coefficients

$$a \triangleq 4\gamma_0 \frac{\sigma^2}{\mu} {t_*}^2 \qquad \text{and} \qquad b \triangleq 8 \frac{\sigma^2}{\mu^2} \; .$$

Achieving the $\epsilon$-accuracy guarantee is equivalent to finding the largest $x = 1/T$ satisfying the inequalities $x > 0$ and

$$ax^2 + bx \leq \epsilon \; .$$

By the quadratic formula, this is equivalent to finding the largest $x$ satisfying

$$0 \leq x \leq \frac{-b + \sqrt{b^2 + 4a\epsilon}}{2a} .$$

Therefore, picking any

$$T \geq \frac{2a}{-b + \sqrt{b^2 + 4a\epsilon}}$$

is sufficient to obtain an $\epsilon$-accurate solution. To make the bound more interpretable, after defining $\alpha = 4a\epsilon$ and $\beta = b$, we can use the inequality (Symbol-1, 2022)

$$\frac{\alpha}{2\sqrt{\beta^2 + \alpha}} \leq -\beta + \sqrt{\beta^2 + \alpha} .$$

Then

$$\frac{2a}{-b + \sqrt{b^2 + 4a\epsilon}} \quad \leq \quad 2a \frac{2\sqrt{b^2 + 4a\epsilon}}{4a\epsilon} \quad = \quad \sqrt{b^2 + 4a\epsilon}\, \frac{1}{\epsilon} \quad \leq \quad b\frac{1}{\epsilon} + 2\sqrt{a}\frac{1}{\sqrt{\epsilon}} ,$$

where we used the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Thus, we have

$$T \geq b\frac{1}{\epsilon} + 2\sqrt{a}\frac{1}{\sqrt{\epsilon}} \qquad \Rightarrow \qquad \mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon .$$

Substituting $t_*$ and $\gamma_0$ with the expressions in Eq. (12),

$$\begin{aligned}
& T \geq 8\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + 2\sqrt{4\gamma_0 \frac{\sigma^2}{\mu} {t_*}^2}\frac{1}{\sqrt{\epsilon}} \quad = \quad 8\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{\gamma_0}\frac{\sigma}{\mu^{1/2}} t_* \frac{1}{\sqrt{\epsilon}} \\
\Leftarrow \quad & T \geq 8\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{\gamma_0}\frac{\sigma}{\mu^{1/2}}\left(\frac{1}{\log 1/\rho}\log\left(\frac{\mu}{2\gamma_0\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) + 1\right)\frac{1}{\sqrt{\epsilon}} \qquad \text{(Eq. (15))} \\
\Leftarrow \quad & T \geq 8\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{\frac{\mu}{2\mathcal{L}^2}}\frac{\sigma}{\mu^{1/2}}\left(\frac{2\mathcal{L}^2}{\mu^2}\log\left(\frac{\mu}{2\sigma^2}\frac{2\mathcal{L}^2}{\mu}\|\lambda_0 - \lambda_*\|_2^2\right) + 1\right)\frac{1}{\sqrt{\epsilon}} \qquad (\log 1/\rho \geq 1 - \rho = \mu^2/(2\mathcal{L}^2)) \\
& \quad = 8\frac{\sigma^2}{\mu^2}\frac{1}{\epsilon} + 2\sqrt{2}\frac{\sigma}{\mathcal{L}}\left(\frac{2\mathcal{L}^2}{\mu^2}\log\left(\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) + 1\right)\frac{1}{\sqrt{\epsilon}}
\end{aligned}$$

Now, Lemma A.4 asserts that $\mathcal{L} \geq \mu$. This allows us to further simplify the term

$$\begin{aligned}
2\sqrt{2}\frac{\sigma}{\mathcal{L}}\left(\frac{2\mathcal{L}^2}{\mu^2}\log\left(\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) + 1\right) &\leq 2\sqrt{2}\frac{\sigma}{\mathcal{L}}\left(\frac{2\mathcal{L}^2}{\mu^2}\log\left(\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) + \frac{2\mathcal{L}^2}{\mu^2}\log 3\right) \\
&= 2\sqrt{2}\frac{\sigma}{\mathcal{L}}\frac{2\mathcal{L}^2}{\mu^2}\log\left(3\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) \\
&= 4\sqrt{2}\frac{\sigma\mathcal{L}}{\mu^2}\log\left(3\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right) .
\end{aligned}$$

Considering this, the sufficient condition for $\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon$ is now

$$T \geq \frac{8\sigma^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{2}\frac{\sigma\mathcal{L}}{\mu^2}\log\left(3\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right)\frac{1}{\sqrt{\epsilon}} . \tag{16}$$

Combining both cases, that is, Eqs. (14) and (16), we have

$$T \geq \max\left(\frac{8\sigma^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{2}\frac{\sigma\mathcal{L}}{\mu^2}\log\left(3\frac{\mathcal{L}^2}{\sigma^2}\|\lambda_0 - \lambda_*\|_2^2\right)\frac{1}{\sqrt{\epsilon}},\ \frac{2\mathcal{L}^2}{\mu^2}\log\left(2\|\lambda_0 - \lambda_*\|_2^2\frac{1}{\epsilon}\right) + 1\right) .$$

This implies the stated result. □

### B.1.2 Proof of Lemma B.1

The proof closely mirrors the strategy of Garrigos and Gower (2023, Theorem 12.9), which is a combination of previous analyses of SPGD (Gorbunov et al., 2020; Khaled et al., 2023) with the analysis of SGD strongly convex objectives with a decreasing step size schedule (Gower et al., 2019). The main difference is that Garrigos and Gower utilize a different condition on the gradient variance instead of Assumption 2.7. Specificially, they assume that, for all $\lambda, \lambda' \in \Lambda$, there exists some function of $L(u) : \mathrm{supp}(\boldsymbol{u}) \to [0, \infty)$ such that, for each $u \in \mathrm{supp}(\boldsymbol{u})$, the function $\widehat{\nabla f}(\lambda; u) : \Lambda \to \mathbb{R}^p$ is $L(u)$-smooth with respect to $\lambda$. This then enables the use of the "convex expected smoothness" (Gorbunov et al., 2020; Khaled et al., 2023) condition, which postulates that, for all $\lambda \in \Lambda$, there exists some $\mathcal{L} < \infty$ such that

$$\mathbb{E}\|\widehat{\nabla f}(\lambda; \boldsymbol{u}) - \widehat{\nabla f}(\lambda'; \boldsymbol{u})\|_2^2 \leq \mathcal{L}^2 \mathrm{D}_f(\lambda, \lambda') \;, \tag{17}$$

where

$$\mathrm{D}_f(\lambda, \lambda') \triangleq f(\lambda) - f(\lambda') - \langle \nabla f(\lambda'), \lambda - \lambda' \rangle \tag{18}$$

is the Bregman divergence associated with $f$. Note that Assumption 2.7 and the $\mu$-strong convexity of $f$ implies Eq. (17). Therefore, under our assumptions, one can invoke the results that assume Eq. (17), which was the strategy by some previous analyses of BBVI (Kim et al., 2023a; Ko et al., 2024). Here, we will take a more straightforward approach that uses Assumption 2.7 directly in the convergence proof, but the results are identical to the indirect approach of establishing Eq. (17).

**Lemma B.1.** *Suppose $f$ is $\mu$-strongly convex, $h$ satisfies Assumption 2.5, and $\nabla f$ satisfies Assumptions 2.7 and 2.8. Then, for the global optimum $\lambda_* = \arg\min_{\lambda \in \Lambda} F(\lambda)$, any $t_*$ satisfying $4\mathcal{L}^2/\mu^2 \leq t_* \leq T$, and the step size schedule in Eq. (5), the contraction coefficient $\rho \triangleq 1 - \mu\gamma_0$ satisfies $\rho \in (0, 1)$ and the last iterate of SPGD after $T$ iterations, $\lambda_T$, satisfies*

$$\mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \|\lambda_0 - \lambda_*\|_2^2 \, \rho^{t_*} \left(\frac{{t_*}^2}{T^2}\right) + 2\gamma_0 \frac{\sigma^2}{\mu} \frac{{t_*}^2}{T^2} + \frac{8\sigma^2}{\mu^2} \frac{T - t_*}{T^2} \;.$$

*Proof.* Since $f$ is strongly convex and $h$ is convex, $F$ is also strongly convex. This implies that $F$ has a unique global optimum, which we denote as $\lambda_*$. Furthermore, under the stated assumptions on $h$, the proximal operator $\mathrm{prox}_{\gamma h}(\cdot)$ is non-expansive for any $\gamma \in (0, \infty)$ (Garrigos and Gower, 2023, Lemma 8.17) and any $\lambda, \lambda' \in \mathbb{R}^p$ such that

$$\|\mathrm{prox}_{\gamma h}(\lambda) - \mathrm{prox}_{\gamma h}(\lambda')\|_2 \leq \|\lambda - \lambda'\|_2 \tag{19}$$

and $\lambda_*$ is the fixed-point of the deterministic proximal gradient descent step (Garrigos and Gower, 2023, Lemma 8.18) such that

$$\mathrm{prox}_{\gamma h}(\lambda_* - \gamma \nabla f(\lambda_*)) = \lambda_* \;. \tag{20}$$

Using these facts,

$$\begin{aligned}
\|\lambda_{t+1} - \lambda_*\|_2^2 &\leq \|\mathrm{prox}_{\gamma_t h}(\lambda_t - \gamma_t \widehat{\nabla f}(\lambda_t; \boldsymbol{u})) - \mathrm{prox}_{\gamma_t h}(\lambda_* - \gamma_t \nabla f(\lambda_*))\|_2^2 && \text{(Eq. (20))} \\
&\leq \|\lambda_t - \gamma_t \widehat{\nabla f}(\lambda_t; \boldsymbol{u})) - \lambda_* + \gamma_t \nabla f(\lambda_*)\|_2^2 \;. && \text{(Eq. (19))}
\end{aligned}$$

Expanding the square,

$$\|\lambda_{t+1} - \lambda_*\|_2^2 \leq \|\lambda_t - \lambda_*\|_2^2 - 2\gamma_t \langle \widehat{\nabla f}(\lambda_t; \boldsymbol{u}) - \nabla f(\lambda_*), \lambda_t - \lambda_* \rangle + \gamma_t^2 \|\widehat{\nabla f}(\lambda_t; \boldsymbol{u}) - \nabla f(\lambda_*)\|_2^2 \;.$$

Denoting the filtration of the $\sigma$-field of the iterates generated up to iteration $t$ as $\mathcal{F}_t$,

$$\begin{aligned}
&\mathbb{E}\big[\|\lambda_{t+1} - \lambda_*\|_2^2 \mid \mathcal{F}_t\big] \\
&\quad \leq \|\lambda_t - \lambda_*\|_2^2 - 2\gamma_t^2 \Big\langle \mathbb{E}\Big[\widehat{\nabla f}(\lambda_t; \boldsymbol{u}) \mid \mathcal{F}_t\Big] - \nabla f(\lambda_*), \lambda_t - \lambda_* \Big\rangle + \gamma_t^2 \mathbb{E}\Big[\|\widehat{\nabla f}(\lambda_t; \boldsymbol{u}) - \nabla f(\lambda_*)\|_2^2 \mid \mathcal{F}_t\Big] \\
&\quad = \|\lambda_t - \lambda_*\|_2^2 - 2\gamma_t \langle \nabla f(\lambda_t) - \nabla f(\lambda_*), \lambda_t - \lambda_* \rangle + \gamma_t^2 \mathbb{E}\Big[\|\widehat{\nabla f}(\lambda_t; \boldsymbol{u}) - \nabla f(\lambda_*)\|_2^2 \mid \mathcal{F}_t\Big] \;, \quad (21)
\end{aligned}$$

where the equality follows from the fact that $\widehat{\nabla f}$ is unbiased conditional on any $\lambda_t \in \Lambda$.

From the $\mu$-strong convexity of $f$,

$$-2\gamma_t \langle \nabla f(\lambda_t) - \nabla f(\lambda_*), \lambda_t - \lambda_* \rangle$$

$$
\begin{aligned}
&= -2\gamma_t\langle \nabla f(\lambda_t), \lambda_t - \lambda_* \rangle + 2\gamma_t \langle \nabla f(\lambda_*), \lambda_t - \lambda_* \rangle \\
&\leq -\gamma_t \mu \|\lambda_t - \lambda_*\|_2^2 - 2\gamma_t \big\{ f(\lambda_t) - f(\lambda_*) - \langle \nabla f(\lambda_*), \lambda_t - \lambda_* \rangle \big\} \quad (\mu\text{-strong convexity of } f) \\
&= -\gamma_t \mu \|\lambda_t - \lambda_*\|_2^2 - 2\gamma_t \mathrm{D}_f(\lambda_t, \lambda_*) \quad (\text{Eq. (18)}) \,. \qquad (22)
\end{aligned}
$$

The gradient variance at $\lambda_t$, on the other hand, can be compared against the gradient variance at $\lambda_*$ through the variance transfer strategy as

$$
\begin{aligned}
&\gamma_t^2 \mathbb{E}\Big[ \|\widehat{\nabla f}(\lambda_t; \mathit{u}) - \nabla f(\lambda_*; \mathit{u})\|_2^2 \mid \mathcal{F}_t \Big] \\
&= \gamma_t^2 \mathbb{E}\Big[ \|\widehat{\nabla f}(\lambda_t; \mathit{u}) - \widehat{\nabla f}(\lambda_*; \mathit{u}) + \widehat{\nabla f}(\lambda_*; \mathit{u}) - \nabla f(\lambda_*; \mathit{u})\|_2^2 \mid \mathcal{F}_t \Big] \\
&\leq 2\,\gamma_t^2 \mathbb{E}\Big[ \|\widehat{\nabla f}(\lambda_t; \mathit{u}) - \widehat{\nabla f}(\lambda_*; \mathit{u})\|_2^2 \mid \mathcal{F}_t \Big] + 2\gamma_t^2 \mathbb{E}\Big[ \|\widehat{\nabla f}(\lambda_*; \mathit{u}) - \nabla f(\lambda_*; \mathit{u})\|_2^2 \mid \mathcal{F}_t \Big] \quad (\text{Young's inequality}) \\
&\leq 2\gamma_t^2 \mathcal{L}^2 \|\lambda_t - \lambda_*\|_2^2 + 2\gamma_t^2 \sigma^2 \,, \quad (\text{Assumptions 2.7 and 2.8}) \\
&= 4\gamma_t^2 \frac{\mathcal{L}^2}{\mu} \Big( f(\lambda_t) - f(\lambda_*) - \langle \nabla f(\lambda_t), \lambda_t - \lambda_* \rangle \Big) + 2\gamma_t^2 \sigma^2 \quad (\mu\text{-strong convexity of } f) \\
&= 4\gamma_t^2 \frac{\mathcal{L}^2}{\mu} \mathrm{D}_f(\lambda_t, \lambda_*) + 2\gamma_t^2 \sigma^2 \,. \qquad (23)
\end{aligned}
$$

Applying Eqs. (22) and (23) to Eq. (21),

$$
\begin{aligned}
\mathbb{E}\big[ \|\lambda_{t+1} - \lambda_*\|_2^2 \mid \mathcal{F}_t \big] &\leq \|\lambda_t - \lambda_*\|_2^2 - \gamma_t \Big( \mu \|\lambda_t - \lambda_*\|_2^2 + 2\,\mathrm{D}_f(\lambda_t, \lambda_*) \Big) + 2\gamma_t^2 \bigg( 2\frac{\mathcal{L}^2}{\mu} \mathrm{D}_f(\lambda_t, \lambda_*) + \sigma^2 \bigg) \\
&= (1 - \gamma_t \mu) \|\lambda_t - \lambda_*\|_2^2 + 2\gamma_t \bigg( 2\gamma_t \frac{\mathcal{L}^2}{\mu} - 1 \bigg) \mathrm{D}_f(\lambda_t, \lambda_*) + 2\gamma_t^2 \sigma^2 \,.
\end{aligned}
$$

Taking expectation over all randomness, we obtain our general partial contraction bound

$$
\mathbb{E}\|\lambda_{t+1} - \lambda_*\|_2^2 \leq (1 - \gamma_t \mu) \mathbb{E}\|\lambda_t - \lambda_*\|_2^2 + 2\gamma_t \bigg( 2\gamma_t \frac{\mathcal{L}^2}{\mu} - 1 \bigg) \mathbb{E}[\mathrm{D}_f(\lambda_t, \lambda_*)] + 2\gamma_t^2 \sigma^2 \,. \qquad (24)
$$

Due to the form of the step size schedule, SPGD operates in two different regimes: the first stage with a fixed step size $\gamma_t = \gamma_0$ ($t \in \{0, \ldots, t_*\}$) and the second stage with a decreasing step size $\gamma_{t+1} < \gamma_t$ ($t \in \{t_* + 1, \ldots, T\}$). In the first stage, $\gamma_t = \gamma_0 \leq \frac{\mu}{2\mathcal{L}^2}$. Then the Bregman divergence term in Eq. (24) is negative such that

$$
\mathbb{E}\|\lambda_{t+1} - \lambda_*\|_2^2 \leq (1 - \gamma_t \mu) \mathbb{E}\|\lambda_t - \lambda_*\|_2^2 + 2\gamma_t^2 \sigma^2 \qquad (25)
$$

Unrolling the recursion yields

$$
\begin{aligned}
\mathbb{E}\|\lambda_{t_*} - \lambda_*\|_2^2 &\leq (1 - \gamma_0 \mu)^{t_*} \|\lambda_0 - \lambda_*\|_2^2 + 2\gamma_0^2 \sigma^2 \sum_{t=0}^{t_* - 1} (1 - \gamma_0 \mu)^t \\
&\leq (1 - \gamma_0 \mu)^{t_*} \|\lambda_0 - \lambda_*\|_2^2 + 2\gamma_0 \frac{\sigma^2}{\mu} \quad (\text{geometric series sum formula}) \\
&\leq \rho^{t_*} \|\lambda_0 - \lambda_*\|_2^2 + 2\gamma_0 \frac{\sigma^2}{\mu} \,. \qquad (26)
\end{aligned}
$$

From Lemma A.4, we deduce that $\gamma_0 \mu = \mu^2 / (2\mathcal{L}^2) \leq 1/2$, which implies $\rho \in (0, 1)$.

We now turn to the second stage, where the step size starts decreasing. Notice that Eq. (5) satisfies

$$
\gamma_t \quad = \quad \frac{1}{\mu} \frac{2t + 1}{(t + 1)^2} \quad \leq \quad \frac{1}{\mu} \frac{2t_* + 1}{(t_* + 1)^2} \quad \leq \quad \frac{1}{\mu} \frac{2}{t_*} \quad \leq \quad \frac{1}{\mu} \frac{2\mu^2}{4\mathcal{L}^2} \quad \leq \quad \frac{\mu}{2\mathcal{L}^2} \,.
$$

Therefore, $\gamma_t \leq \frac{\mu}{2\mathcal{L}^2}$ for all $t \geq 0$. Again, the Bregman term in Eq. (24) is negative such that

$$
\mathbb{E}\|\lambda_{t+1} - \lambda_*\|_2^2 \leq (1 - \gamma_t \mu) \mathbb{E}\|\lambda_t - \lambda_*\|_2^2 + 2\gamma_t^2 \sigma^2 \,.
$$

Subtituting $\gamma_t$ with the choice in Eq. (5), we obtain

$$\begin{aligned}\mathbb{E}\|\lambda_{t+1}-\lambda_*\|_2^2 &\leq \left(1-\frac{2t+1}{(t+1)^2}\right)\mathbb{E}\|\lambda_t-\lambda_*\|_2^2+2\frac{\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^4}\\ &= \frac{t^2}{(t+1)^2}\mathbb{E}\|\lambda_t-\lambda_*\|_2^2+2\frac{\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^4}\ .\end{aligned}$$

Multiplying $(t+1)^2$ to both sides,

$$(t+1)^2\mathbb{E}\|\lambda_{t+1}-\lambda_*\|_2^2 \leq t^2\mathbb{E}\|\lambda_t-\lambda_*\|_2^2+2\frac{\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^2}\ .$$

Let us choose the Lyapunov function $V_t \triangleq (t+1)^2\mathbb{E}\|\lambda_{t+1}-\lambda_*\|_2^2$ . Then the discrete derivative of the Lyapunov,

$$V_{t+1}-V_t \quad\leq\quad 2\frac{\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^2} \quad\leq\quad 8\frac{\sigma^2}{\mu^2}\ ,$$

shows that the energy is increasing only by a constant. By integrating the Lyapunov over the time interval $t=t_*,\ldots,T-1$,

$$\begin{aligned}& & V_T-V_{t_*} &\leq 8\frac{\sigma^2}{\mu^2}(T-t_*)\\ \Leftrightarrow & & V_T &\leq V_{t_*}+8\frac{\sigma^2}{\mu^2}(T-t_*)\\ \Leftrightarrow & & T^2\,\mathbb{E}\|\lambda_T-\lambda_*\|_2^2 &\leq t_*^2\mathbb{E}\|\lambda_{t_*}-\lambda_*\|_2^2+8\frac{\sigma^2}{\mu^2}(T-t_*)\\ \Leftrightarrow & & \mathbb{E}\|\lambda_T-\lambda_*\|_2^2 &\leq \frac{{t_*}^2}{T^2}\mathbb{E}\|\lambda_{t_*}-\lambda_*\|_2^2+8\frac{\sigma^2}{\mu^2}\frac{T-t_*}{T^2}\ .\end{aligned}$$

Substuting $\|\lambda_{t_*}-\lambda_*\|_2^2$ with the error in Eq. (26),

$$\begin{aligned}\mathbb{E}\|\lambda_T-\lambda_*\|_2^2 &\leq \left(\rho^{t_*}\|\lambda_0-\lambda_*\|_2^2+2\gamma_0\frac{\sigma^2}{\mu}\right)\frac{{t_*}^2}{T^2}+8\frac{\sigma^2}{\mu^2}\frac{T-t_*}{T^2}\\ &= \|\lambda_0-\lambda_*\|_2^2\,\rho^{t_*}\frac{{t_*}^2}{T^2}+2\gamma_0\frac{\sigma^2}{\mu}\frac{{t_*}^2}{T^2}+\frac{8\sigma^2}{\mu^2}\frac{T-t_*}{T^2}\ ,\end{aligned}\tag{27}$$

which is our stated result. □

### B.2 Proofs of Results in Section 3

#### B.2.1 Proof of Theorem 3.2

**Theorem 3.2.** *Suppose the following hold:*

1. *$\ell$ is $\mu$-strongly convex and satisfies Assumption 3.1 and $\mu \leq \sigma_{\min}(H) \leq \sigma_{\max}(H) \leq L$.*
2. *$h$ satisfies Assumption 2.5.*
3. *$\mathcal{Q}$ is a mean-field location-scale family, where Assumption 2.2 holds.*
4. *$\widehat{\nabla f}$ is the reparametrization gradient.*

*Denote the global optimum $\lambda_* = (m_*, C_*) = \arg\min_{\lambda\in\Lambda} F(\lambda)$, the irreducible gradient noise as $\sigma_*^2 \triangleq \|m_* - \bar{z}\|_2^2 + \|C_*\|_{\mathrm{F}}^2$, and the stationary point of $\ell$ as $\bar{z} \triangleq \arg\min_{z\in\mathbb{R}^d} \ell(z)$. Then there exists some $t_*$ and $\gamma_0$ such that SPGD with the step size schedule in Eq. (5) guarantees*

$$T \geq \mathrm{O}\Big\{ g(d, H, \delta, \mu, \varphi)\Big(\sigma_*^2\epsilon^{-1} + \sigma_* \log\big(\|\lambda_0 - \lambda_*\|_2^2\big)\epsilon^{-1/2}\Big)\Big\} \quad \Rightarrow \quad \mathbb{E}\|\lambda_T - \lambda_*\|_2^2 \leq \epsilon \,,$$

*where*

$$g(d, H, \delta, \mu, \varphi) \triangleq 2\,(1 + r_4)\big(\|H\|_2^2/\mu^2\big) + 4\big(\delta^2/\mu^2\big)\Big((1/2) + r_4 + \mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2\Big) \,.$$

*Proof.* The proof consists of establishing the sufficient conditions of Proposition 2.9 as follows:

(i) $\ell$ is $\mu$-strongly convex $\quad\Rightarrow\quad$ $f$ is $\mu$-strongly convex.

(ii) Assumption 3.1 $\quad\Rightarrow\quad$ Assumptions 2.7 and 2.8.

Under the linear parametrization, (i) was established by Domke (2020, Thm. 9). It remains to establish (ii). Therefore, the proof focuses on analyzing the variance of the gradient estimator $\widehat{\nabla f}$. Since Assumption 3.1 holds, Lemma 4.1 states that, for all $\lambda, \lambda' \in \mathbb{R}^d \times \mathbb{D}^d$, the inequality

$$\mathbb{E}\|\widehat{\nabla f}(\lambda; \mathsf{u}) - \widehat{\nabla f}(\lambda'; \mathsf{u})\|_2^2 \leq \Big\{2(1 + r_4)\|H\|_2^2 + 4\delta^2\Big(1/2 + r_4 + \mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2\Big)\Big\}\|\lambda - \lambda'\|_2^2$$

holds. Since $\Lambda \subset \mathbb{R}^d \times \mathbb{D}^d$ under the linear parametrization, this implies we satisfy Assumption 2.7 with

$$\mathcal{L}^2 = 2(1 + r_4)\|H\|_2^2 + 4\delta^2\Big(1/2 + r_4 + \mathbb{E}\max_{j=1,\ldots,d} \mathsf{u}_j^2\Big) \,. \tag{28}$$

Furthermore, For the specific choice of $\lambda_* = (m_*, C_*) = \arg\min_{\lambda\in\Lambda} F(\lambda)$ and $\bar{\lambda} = (\bar{z}, 0_{d\times d})$ (which is not part of $\Lambda$), we have the equality

$$\mathbb{E}\|\widehat{\nabla f}(\lambda_*; \mathsf{u}) - \widehat{\nabla f}\big(\bar{\lambda}; \mathsf{u}\big)\|_2^2 = \mathbb{E}\|\widehat{\nabla f}(\lambda_*; \mathsf{u}) - \widehat{\nabla f}(\bar{z}; \mathsf{u})\|_2^2 = \mathbb{E}\|\widehat{\nabla f}(\lambda_*; \mathsf{u})\|_2^2 \,.$$

This means Lemma 4.1 also implies Assumption 2.8 with the constant

$$\sigma^2 \quad = \quad \mathcal{L}^2\|\lambda_* - \bar{\lambda}\|_2^2 \quad = \quad \mathcal{L}^2\big(\|m_* - \bar{z}\| + \|C_*\|_{\mathrm{F}}^2\big) \quad = \quad \mathcal{L}^2\sigma_*^2 \,. \tag{29}$$

We are now able to invoke Proposition 2.9. Substituting $\mathcal{L}$ and $\sigma^2$ in Eq. (27) with the expressions above, we obtain the condition

$$T \geq \max\left(\frac{8\sigma_*^2\mathcal{L}^2}{\mu^2}\frac{1}{\epsilon} + 4\sqrt{2}\frac{\sigma_*\mathcal{L}^2}{\mu^2}\log\left(\frac{3}{\sigma_*^2}\|\lambda_0 - \lambda_*\|_2^2\right)\frac{1}{\sqrt{\epsilon}},\ \frac{2\mathcal{L}^2}{\mu^2}\log\left(2\|\lambda_0 - \lambda_*\|_2^2\frac{1}{\epsilon}\right) + 1\right) .$$

Using the fact $\mathcal{L} \geq \mu$ from Lemma A.4, we finally have

$$T \geq \frac{\mathcal{L}^2}{\mu^2}\max\left(8\sigma_*^2\frac{1}{\epsilon} + 4\sqrt{2}\,\sigma_*\log\left(\frac{3}{\sigma_*^2}\|\lambda_0 - \lambda_*\|_2^2\right)\frac{1}{\sqrt{\epsilon}},\ 2\log\left(2\|\lambda_0 - \lambda_*\|_2^2\frac{1}{\epsilon}\right) + 1\right) .$$

Finally, substituting for Eq. (28) yields our stated result. □

### B.2.2 Proof of Proposition 3.4

The result follows from a well-known bound on the expected maximum of sub-exponential random variables. We state the proof for completeness.

**Lemma B.2.** *Let $X_1, \dots, X_d$ be i.i.d. random variables. Suppose there exists some $t > 0$ such that their moment-generating function (MGF) satisfies $M_{X_i}(t) < \infty$. Then*

$$\mathbb{E} \max_{i=1,\dots,d} X_i \quad \leq \quad \frac{1}{t}(\log M_{X_i}(t) + \log d) \ .$$

*Proof.*

$$\begin{aligned}
\mathbb{E}\left[t \max_{i=1,\dots,d} X_i\right] &= \log \exp\left(\mathbb{E}\left[t \max_{i=1,\dots,d} X_i\right]\right) \\
&\leq \log \mathbb{E} \exp\left(t \max_{i=1,\dots,d} X_i\right) && \text{(Jensen's inequality)} \\
&= \log \mathbb{E} \max_{i=1,\dots,d} \exp(tX_i) \\
&\leq \log \mathbb{E} \sum_{i=1}^{d} \exp(tX_i) \\
&= \log \sum_{i=1}^{d} M_{X_i}(t) && \text{(Definition of MGFs)} \\
&= \log(dM_{X_i}(t)) \ . && (X_1, \dots, X_d \text{ are i.i.d.})
\end{aligned}$$

Dividing both sides by $t$ yields the statement. □

Applying Lemma B.2 to $U_i^2$ yields the result.

**Proposition 3.4.** *Suppose there exists some $t > 0$ such that the MGF of $U_i^2$ satisfies $M_{U_i^2}(t) < \infty$. Then*

$$\mathbb{E} \max_{i=1,\dots,d} U_i^2 \quad \leq \quad (1/t)\left(\log M_{U_i^2}(t) + \log d\right) \ .$$

*For example, if $\varphi$ is a standard Gaussian, then*

$$g(d, H, \delta, \mu, \varphi) \quad \leq \quad 8\big(\|H\|_2^2/\mu^2\big) + \big(\delta^2/\mu^2\big)(22 + 16 \log d) \ .$$

*Proof.* The first part of the statement is a re-statement of Lemma B.2.

For the special case of $U_i \sim \mathcal{N}(0,1)$, we know that $U_i^2 \sim \chi_1^2$ (Johnson et al., 1995, Eq. 29.1), which is the $\chi^2$ distribution with 1 degree of freedom. The MGF of $\chi_1^2$ is given as

$$M_{U_i^2}(t) = (1-2t)^{-1/2} \qquad \text{(Johnson et al., 1995, Eq. 29.6)}$$

for $t \in (0, 1/2)$. Then we can invoke Lemma B.2, which suggests

$$\mathbb{E} \max_{i=1,\dots,d} U_i^2 \leq \min_{t \in (0,1/2)} \frac{1}{t}\left(-\frac{1}{2}\log(1-2t) + \log d\right) \ .$$

Any fixed choice of $t \in (0, 1/2)$ is a valid upper bound. Picking $t = \frac{1}{2}\big(1 - \frac{1}{\mathrm{e}}\big) \geq \frac{1}{4}$ yields

$$\mathbb{E} \max_{i=1,\dots,d} U_i^2 \leq 4\left(\frac{1}{2} + \log d\right) \ . \tag{30}$$

Furthermore, the kurtosis of the standard Gaussian is $r_4 = 3$ (Johnson et al., 1994, Eq. 13.11). Plugging $r_4$ and Eq. (30) into $g$ in Theorem 3.2 yields the statement. □

### B.2.3 Proof of Proposition 3.5

The result follows from the following moment-based bound on the expected maximum of random variables, which is a non-asymptotic refinement of the proof by Rana (2017).

**Lemma B.3.** *Let $\mathsf{x}_1, \ldots, \mathsf{x}_d$ be i.i.d. non-negative random variables where, for $k \geq 2$, their $k$th moment is finite. That is, $\mathbb{E}\mathsf{x}_i^k = r_k < \infty$. Then*

$$\mathbb{E} \max_{i=1,\ldots,d} \mathsf{x}_i \quad \leq \quad d^{1/k}(k/(k-1))^{(k-1)/k} r_k^{1/k} \, .$$

*Proof.* For any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E} \max_{i=1,\ldots,d} \mathsf{x}_i &= \int_0^{\epsilon d^{1/k}} \mathbb{P}\left[\max_{i=1,\ldots,d} \mathsf{x}_i \geq t\right] \mathrm{d}t + \int_{\epsilon d^{1/k}}^{\infty} \mathbb{P}\left[\max_{i=1,\ldots,d} \mathsf{x}_i \geq t\right] \mathrm{d}t \\
&\leq \int_0^{\epsilon d^{1/k}} \mathrm{d}t + \int_{\epsilon d^{1/k}}^{\infty} d\,\mathbb{P}\left[\mathsf{x}_i \geq t\right] \mathrm{d}t && \text{(i.i.d. and } \mathbb{P}[\cdot] \leq 1) \\
&= d^{1/k}\left(\epsilon + \frac{1}{k\epsilon^{k-1}} \int_{\epsilon d^{1/k}}^{\infty} k\left(\epsilon d^{1/k}\right)^{k-1} \mathbb{P}\left[\mathsf{x}_i \geq t\right] \mathrm{d}t\right) \\
&\leq d^{1/k}\left(\epsilon + \frac{1}{k\epsilon^{k-1}} \int_{\epsilon d^{1/k}}^{\infty} k t^{k-1} \mathbb{P}\left[\mathsf{x}_i \geq t\right] \mathrm{d}t\right) && (\epsilon d^{1/k} \leq t) \\
&\leq d^{1/k}\left(\epsilon + \frac{1}{k\epsilon^{k-1}} \int_{0}^{\infty} k t^{k-1} \mathbb{P}\left[\mathsf{x}_i \geq t\right] \mathrm{d}t\right) && \text{(Decreased lower limit of integral)} \, .
\end{aligned}$$

Now, from the definition of moments, we know that

$$\begin{aligned}
\int_0^{\infty} k t^{k-1} \mathbb{P}\left[\mathsf{x}_i \geq t\right] \mathrm{d}t &= \int_0^{\infty} \int_{-\infty}^{\infty} k t^{k-1} \mathbb{1}_{x_i > t} \, \mathrm{d}\mathbb{P}[x_i] \, \mathrm{d}t \\
&= \int_{-\infty}^{\infty} \int_0^{\infty} k t^{k-1} \mathbb{1}_{x_i > t} \, \mathrm{d}t \, \mathrm{d}\mathbb{P}[x_i] && \text{(Fubini's Theorem)} \\
&= \int_{-\infty}^{\infty} \int_0^{x_i} k t^{k-1} \, \mathrm{d}t \, \mathrm{d}\mathbb{P}[x_i] \\
&= \int_{-\infty}^{\infty} x_i^k \, \mathrm{d}\mathbb{P}[x_i] \\
&= r_k \, .
\end{aligned}$$

Therefore,

$$\mathbb{E} \max_{i=1,\ldots,d} \mathsf{x}_i \leq d^{1/k}\left(\epsilon + \frac{1}{k\epsilon^{k-1}} r_k\right) .$$

The bound is minimized when setting

$$\epsilon = \left(\frac{k-1}{k} r_k\right)^{1/k} .$$

Then

$$\begin{aligned}
\mathbb{E} \max_{i=1,\ldots,d} \mathsf{x}_i &\leq d^{1/k}\left(\left(\frac{k-1}{k} r_k\right)^{1/k} + \frac{1}{k} m_k \left(\frac{k-1}{k} r_k\right)^{-(k-1)/k}\right) \\
&= d^{1/k}\left(\left(\frac{k-1}{k} r_k\right)^{1/k} + \frac{1}{k-1}\left(\frac{k-1}{k} r_k\right)^{1/k}\right) \\
&= d^{1/k}\left(1 + \frac{1}{k-1}\right)\left(\frac{k-1}{k} r_k\right)^{1/k} \\
&= d^{1/k}\left(\frac{k}{k-1}\right)^{(k-1)/k} r_k^{1/k} \, .
\end{aligned}$$

□

If the $k$th moment of $\mathsf{u}_i^2$ is finite, this then immediately implies a polynomial $\mathrm{O}(d^{1/k})$ bound on $g$.

**Proposition 3.5.** *Suppose, for $k \geq 2$, the kth moment of $\mathsf{u}_i^2$ is finite as $r_{2k} = \mathbb{E}\mathsf{u}_i^{2k} < \infty$. Then*

$$\mathbb{E} \max_{i=1,\dots,d} \mathsf{u}_i^2 \quad \leq \quad \sqrt{2}\, d^{1/k}\, r_{2k}^{1/k} \,.$$

*For example, if $\varphi$ is a Student-t with $\nu > 4$ degrees of freedom and unit variance, then*

$$g(d, H, \delta, \mu, \varphi) \quad \leq \quad 8(\|H\|_2^2/\mu^2) + (\delta^2/\mu^2)\Big(16 + \sqrt{2}\,\nu^3 d^{\frac{2}{\nu-2}}\Big) \,.$$

*Proof.* The first part of the statement directly follows from Lemma B.3, where we simplified $\left(k/k-1\right)^{(k-1)/k}$. In particular, for $k \geq 2$, $\left(k/k-1\right)^{(k-1)/k}$ is monotonically decreasing. Since an order $k \geq 2$ moment exists by the assumption on the degrees of freedom, $\left(k/k-1\right)^{(k-1)/k} \leq \sqrt{2}$.

Let's turn to the second part of the statement. We will denote a Student-$t$ distribution with $\nu$-degrees of freedom as $t_\nu$. Since $t_\nu$ does not have unit variance (Johnson et al., 1995, Eq. 28.7a), we have to set the sampling process from $\varphi$ to be

$$\mathsf{u}_i \sim \varphi \qquad \Leftrightarrow \qquad \mathsf{u}_i \stackrel{\mathrm{d}}{=} \frac{\nu-2}{\nu}\mathsf{v}_i \,, \quad \text{where} \quad \mathsf{v}_i \overset{\text{i.i.d.}}{\sim} t_\nu \,.$$

Now, it is known that $\mathsf{v}_i^2 \stackrel{\mathrm{d}}{=} \mathsf{w}_i \sim \mathrm{FDist}(1, \nu_2)$ (Johnson et al., 1995, §28.7), where $\mathrm{FDist}(\nu_1, \nu_2)$ is Fisher's $F$-distribution with $(\nu_1, \nu_2)$ degrees of freedom. The $k$th raw moment of $\mathrm{FDist}(\nu_1, \nu_2)$, denoted as $m_k \triangleq \mathbb{E}\mathsf{w}_i^k$, exists up to $2k < \nu_2 = \nu$ and is given as

$$m_k = \left(\frac{\nu_2}{\nu_1}\right)^k \frac{\Gamma(\nu_1/2+k)}{\Gamma(\nu_1/2)} \frac{\Gamma(\nu_2/2+k)}{\Gamma(\nu_2/2)} \,. \qquad \text{(Johnson et al., 1995, Eq. 27.43)}$$

This means that we can invoke Lemma B.3 as

$$\mathbb{E} \max_{i=1,\dots,d} \mathsf{u}_i^2 \quad = \quad \left(\frac{\nu-2}{\nu}\right)^2 \mathbb{E} \max_{i=1,\dots,d} \mathsf{w}_i \quad \leq \quad \sqrt{2}\left(\frac{\nu-2}{\nu}\right)^2 d^{1/k} m_k^{1/k} \,,$$

with any $k < \nu/2$.

For $m_k^{1/k}$, we can use the fact that the gamma function satisfies the recursion $\Gamma(z+1) = z\Gamma(z)$, which implies $\Gamma(a/2+k) = \Gamma(a/2)\prod_{i=0}^{k-1}(a/2+i)$ for any $a > 0$. Therefore,

$$\begin{aligned}
\left(\frac{\Gamma(a/2+k)}{\Gamma(a/2)}\right)^{1/k} &= \left(\prod_{i=0}^{k-1}\Big(\frac{a}{2}+i\Big)\right)^{1/k} \\
&\leq \frac{1}{k}\sum_{i=0}^{k-1}\Big(\frac{a}{2}+i\Big) && \text{(AM-GM inequality)} \\
&= \frac{a}{2} + \frac{1}{k}\frac{k(k-1)}{2} && \text{(geometric series sum formula)} \\
&= \frac{a+k-1}{2} \,.
\end{aligned}$$

Applying this bound to $a = \nu_2 = \nu$ and $a = \nu_1 = 1$ respectively,

$$\begin{aligned}
m_k^{1/k} &= \left(\nu^k \frac{\Gamma(1/2+k)}{\Gamma(1/2)} \frac{\Gamma(\nu/2+k)}{\Gamma(\nu/2)}\right)^{1/k} \\
&\leq \nu\, \frac{k}{2}\, \frac{\nu+k-1}{2} \\
&< \nu\, \frac{\nu}{4}\, \frac{3\nu}{4} && (k < \nu/2) \\
&< \frac{\nu^3}{4} \,.
\end{aligned}$$

Also, choosing $k = \lceil \nu/2 - 1 \rceil$, we have $d^{1/k} \leq d^{2/(\nu-2)}$. This yields

$$\mathbb{E} \max_{i=1,\dots,d} \mathsf{u}_i^2 \quad < \quad \sqrt{2}\left(\frac{\nu-2}{\nu}\right)^2 d^{\frac{2}{\nu-2}} \frac{\nu^3}{4} \quad < \quad \frac{1}{2\sqrt{2}}\nu^3 d^{\frac{2}{\nu-2}} \,. \tag{31}$$

Lastly, the kurtosis of $u_i = (\nu-2)/\nu\, v_i$ follows as (Johnson et al., 1995, Eq. 28.5)

$$r_4 \quad = \quad \left(\frac{\nu-2}{\nu}\right)^4 \mathbb{E} w_i^2 \quad = \quad \left(\frac{\nu-2}{\nu}\right)^4 \frac{3\nu^2}{(\nu-2)(\nu-4)} \quad = \quad 3\frac{(\nu-2)^3}{\nu^2(\nu-4)} \quad \leq \quad 3\,.$$

Plugging the bound in Eq. (31) and the value of $r_4$ into $g$ in Theorem 3.2 yields the statement. □

### B.3 Proofs of Results in Section 4

#### B.3.1 Proof of Lemma 4.1

Under the assumption that $\nabla^2 \ell \preceq L\mathrm{I}_d$ and twice differentiability, it is well known that $\nabla^2 \ell \preceq L\mathrm{I}_d \;\Rightarrow \ell$ is $L$-smooth. We will prove a supporting result analogous to this under Assumption 3.1, which will allow us to bound the relative growth of $\nabla \ell$.

**Lemma B.4.** *Suppose $\ell : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 3.1. Then, for any $W \in \mathbb{R}^{d\times d}$ satisfying $\|W\|_2 < \infty$,*

$$\|W(\nabla \ell(z) - \nabla \ell(z'))\|_2 \le \|WH(z - z')\|_2 + \delta \|W\|_2 \|z - z'\|_2 \,.$$

*Proof.* The *full proof* is deferred to Appendix B.3.2, p. 39. □

Using this, we can now simplify the $\nabla \ell$ terms in Eq. (6). Applying Lemma B.4 to $V_{\text{loc}}$ with $W = \mathrm{I}_d$ and Young's inequality,

$$\begin{aligned}
V_{\text{loc}} &\le \mathbb{E}(\|H(\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}))\|_2 + \delta \|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2)^2 && \text{(Lemma B.4)} \\
&\le 2\,\mathbb{E}\|H(\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}))\|_2^2 + 2\delta^2\,\mathbb{E}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2 && \text{(Young's inequality)} \\
&\le 2\,\|H\|_2^2 \mathbb{E}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2 + 2\delta^2\,\mathbb{E}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2 && \text{(Operator norm)} \\
&= 2\big(\|H\|_2^2 + \delta^2\big)\mathbb{E}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2 \\
&= 2\big(\|H\|_2^2 + \delta^2\big)\|\lambda - \lambda'\|_2^2 \,. && \text{(Lemma A.3)} \qquad (32)
\end{aligned}$$

Similarly, applying Lemma B.4 to $V_{\text{scale}}$ with $W = \mathsf{U}$ and Young's inequality,

$$\begin{aligned}
V_{\text{scale}} &\le \mathbb{E}(\|\mathsf{U}H(\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}))\|_2 + \delta\|\mathsf{U}\|_2\,\mathbb{E}\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2)^2 && \text{(Lemma B.4)} \\
&\le 2\underbrace{\mathbb{E}\|\mathsf{U}H(\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}))\|_2^2}_{V_{\text{const}}} + 2\delta^2\underbrace{\mathbb{E}\|\mathsf{U}\|_2^2\|\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u})\|_2^2}_{V_{\text{non-const}}} && \text{(Young's inequality)} \,.
\end{aligned} \qquad (33)$$

$V_{\text{const}}$ corresponds to the constant component of the Hessian $\nabla^2 \ell$, whereas $V_{\text{non-const}}$ corresponds to the non-constant residual. Denote the location and scale parameters of $\lambda$ and $\lambda'$ as

$$\lambda = (m, C) \qquad \text{and} \qquad \lambda' = (m', C') \,.$$

For $V_{\text{const}}$, we can use the following lemma:

**Lemma B.5.** *Suppose $\mathcal{T}_\lambda$ is the reparameterization operator of a mean-field location-family and Assumption 2.2 holds. Then, for any matrix $H \in \mathbb{R}^{d\times d}$ and any $\lambda, \lambda' \in \mathbb{R}^d \times \mathbb{D}^d$,*

$$\|\mathsf{U}H(\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}))\|_2^2 \le r_4 \|H\|_2^2 \|\lambda - \lambda'\|_2^2 \,.$$

See the *full proof* in Appendix B.3.3, p. 40.

The remaining part of the proof closely resembles the proof sketch of Lemma 4.1. For convenience, we first restate Lemma 4.1 and then proceed to the full proof.

**Lemma 4.1.** *Suppose Assumptions 2.2 and 3.1 hold, $\mathcal{Q}$ is a mean-field location-family, and $\widehat{\nabla f}$ is the reparametrization gradient. Then, for any $\lambda, \lambda' \in \mathbb{R}^d \times \mathbb{D}^d$.*

$$\mathbb{E}\|\widehat{\nabla} f(\lambda; \mathsf{u}) - \widehat{\nabla} f(\lambda'; \mathsf{u})\|_2^2 \le \left\{2(1 + r_4)\|H\|_2^2 + 4\delta^2\left(1/2 + r_4 + \mathbb{E}\max_{j=1,\dots,d} \mathsf{u}_j^2\right)\right\}\|\lambda - \lambda'\|_2^2 \,.$$

*Proof.* Recall Eq. (33). The proof consists of bounding the two terms $V_{\text{const}}$ and $V_{\text{non-const}}$. First, for $V_{\text{const}}$, under Assumption 3.1,

$$V_{\text{const}} \le r_4 \,\|H\|_2^2 \,\|\lambda - \lambda'\|_2^2 \,. \qquad \text{(Lemma B.5)} \qquad (34)$$

It remains to bound $V_{\text{non-const}}$, which is our main challenge.

Denote $\bar{m} \triangleq m - m'$ and $\bar{C} \triangleq C - C'$ such that

$$\begin{aligned}
\mathcal{T}_\lambda(\mathsf{u}) - \mathcal{T}_{\lambda'}(\mathsf{u}) &= (C\mathsf{u} + m) - (C'\mathsf{u} + m') \\
&= (C - C')\mathsf{u} + (m - m')
\end{aligned}$$

$$= \bar{C}\boldsymbol{u} + \bar{m} \ .$$

Then

$$\begin{aligned}
V_{\text{non-const}} &= \mathbb{E}\|\boldsymbol{U}\|_2^2 \|\mathcal{T}_\lambda(\boldsymbol{u}) - \mathcal{T}_{\lambda'}(\boldsymbol{u})\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{U}\|_2^2 \|\bar{C}\boldsymbol{u} + \bar{m}\|_2^2 \\
&\le \mathbb{E}\|\boldsymbol{U}\|_2^2 \big(2\|\bar{C}\boldsymbol{u}\|_2^2 + 2\|\bar{m}\|_2^2\big) && \text{(Young's inequality)} \\
&= \mathbb{E}\Big(\max_{j=1,\dots,d} \boldsymbol{u}_j^2\Big) \sum_{i=1}^d \big(2\bar{C}_{ii}^2 \boldsymbol{u}_i^2 + 2\bar{m}_i^2\big) \\
&= 2\mathbb{E}\sum_{i=1}^d \Big(\max_{j=1,\dots,d} \boldsymbol{u}_j^2\Big) \bar{C}_{ii}^2 \boldsymbol{u}_i^2 + 2\mathbb{E}\Big(\max_{j=1,\dots,d} \boldsymbol{u}_j^2\Big) \sum_{i=1}^d \bar{m}_i^2 \ .
\end{aligned}$$

We will focus on the first term. Denoting $i_* = \arg\max_{i=1,\dots,d} \boldsymbol{u}_i^2$, the coordinate of maximum magnitude, we can decompose the expectation by the contribution of the event $i_* = i$ and $i_* \neq i$. That is,

$$\mathbb{E}\boldsymbol{u}_{i_*}^2 \textstyle\sum_{i=1}^d \bar{C}_{ii}^2 \boldsymbol{u}_i^2 = \sum_{i=1}^d \bar{C}_{ii}\, \mathbb{E}\Big[ \underbrace{\boldsymbol{u}_{i_*}^4 \mathbb{1}\{i_* = i\}}_{V_{\max}} + \underbrace{\boldsymbol{u}_{i_*}^2 \boldsymbol{u}_i^2 \mathbb{1}\{i_* \neq i\}}_{V_{\text{non-max}}} \Big] \ .$$

The expectation of the event $i_* = i$ follows as

$$\begin{aligned}
V_{\max} &= \mathbb{E}\big[\boldsymbol{u}_{i_*}^4 \mathbb{1}\{i_* = i\}\big] \\
&= \mathbb{E}\big[\boldsymbol{u}_{i_*}^4\big] \mathbb{E}[\mathbb{1}\{i_* = i\}] && (\boldsymbol{u}_{i_*} \perp\!\!\!\perp i_*) \\
&= \mathbb{E}\big[\boldsymbol{u}_{i_*}^4\big] \mathbb{P}[i_* = i] \\
&= \mathbb{E}\big[\boldsymbol{u}_{i_*}^4\big] \frac{1}{d} \\
&\le \mathbb{E}\left[\sum_{i=1}^d \boldsymbol{u}_i^4\right] \frac{1}{d} && (\max_{j=1,\dots,d} \boldsymbol{u}_j^4 \le \textstyle\sum_{j=1}^d \boldsymbol{u}_j^4) \\
&= (dr_4) \frac{1}{d} && \text{(Assumption 2.2)} \\
&= r_4 \ . && (35)
\end{aligned}$$

On the other hand, for the event $i_* \neq i$,

$$\begin{aligned}
V_{\text{non-max}} &= \mathbb{E}\big[\boldsymbol{u}_{i_*}^2 \boldsymbol{u}_i^2 \mathbb{1}\{i_* \neq i\}\big] \\
&= \mathbb{E}\Big[\max_{j \neq i} \boldsymbol{u}_j^2 \boldsymbol{u}_i^2 \mathbb{1}\{i_* \neq i\}\Big] \\
&= \mathbb{E}\Big[\max_{j \neq i} \boldsymbol{u}_j^2 \boldsymbol{u}_i^2\Big] && (\mathbb{1} \le 1) \\
&\le \mathbb{E}\Big[\max_{j \neq i} \boldsymbol{u}_j^2\Big] \mathbb{E}\big[\boldsymbol{u}_i^2\big] && (\boldsymbol{u}_j \perp\!\!\!\perp \boldsymbol{u}_i \text{ for all } i \neq j) \\
&= \mathbb{E}\Big[\max_{j=1,\dots,d-1} \boldsymbol{u}_j^2\Big] \mathbb{E}\big[\boldsymbol{u}_i^2\big] && (\boldsymbol{u}_1, \dots, \boldsymbol{u}_d \text{ are i.i.d.}) \\
&= \mathbb{E} \max_{j=1,\dots,d-1} \boldsymbol{u}_j^2 \ . && \text{(Assumption 2.2)}
\end{aligned}$$

Therefore, we finally obtain

$$\begin{aligned}
V_{\text{non-const}} &\le 2\sum_{i=1}^d \left[\Big(\mathbb{E} \max_{j=1,\dots,d-1} \boldsymbol{u}_j^2 + r_4\Big)\bar{C}_{ii}^2 + \mathbb{E} \max_{j=1,\dots,d} \boldsymbol{u}_j^2 \bar{m}_i^2\right] \\
&\le 2\Big(\mathbb{E} \max_{j=1,\dots,d} \boldsymbol{u}_j^2 + r_4\Big)\big(\|\bar{m}\|_2^2 + \|\bar{C}\|_{\mathrm{F}}^2\big) && (\max_{j=1,\dots,d-1} \boldsymbol{u}_j^2 \le \max_{j=1,\dots,d} \boldsymbol{u}_j^2) \\
&= 2\Big(\mathbb{E} \max_{j=1,\dots,d} \boldsymbol{u}_j^2 + r_4\Big)\|\lambda - \lambda'\|_2^2 \ . && (36)
\end{aligned}$$

Combining Eqs. (6), (32) to (34) and (36) yields the statement. □

### B.3.2 Proof of Lemma B.4

**Lemma B.4.** *Suppose $\ell : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 3.1. Then, for any $W \in \mathbb{R}^{d\times d}$ satisfying $\|W\|_2 < \infty$,*

$$\|W(\nabla\ell(z) - \nabla\ell(z'))\|_2 \le \|WH(z-z')\|_2 + \delta\|W\|_2\|z-z'\|_2 \,.$$

*Proof.* From twice differentiability of $\ell$ (Assumption 3.1) and the fundamental theorem of calculus, we know that

$$\|W(\nabla\ell(z) - \nabla\ell(z'))\|_2 = \left\| W \int_0^1 \nabla^2\ell(tz + (1-t)z')(z-z')\,\mathrm{d}t \right\|_2 .$$

Denoting $z_t \triangleq tz + (1-t)z'$ for clarity,

$$\begin{aligned}
&\|W(\nabla\ell(z) - \nabla\ell(z'))\|_2 \\
&\quad= \left\| \int_0^1 W\nabla^2\ell(z_t)(z-z')\,\mathrm{d}t \right\|_2 \\
&\quad\le \int_0^1 \|W\nabla^2\ell(z_t)(z-z')\|_2\,\mathrm{d}t && \text{(Jensen's inequality)} \\
&\quad= \int_0^1 \left\|W\big(\nabla^2\ell(z_t) - H + H\big)(z-z')\right\|_2\mathrm{d}t \\
&\quad\le \int_0^1 \left\{\|WH(z-z')\|_2 + \|W\|_2\left\|\nabla^2\ell(z_t) - H\right\|_2\|z-z'\|_2\right\}\mathrm{d}t && \text{(Triangle inequality)} \\
&\quad\le \int_0^1 \left\{\|WH(z-z')\|_2 + \delta\|W\|_2\|z-z'\|_2\right\}\mathrm{d}t && \text{(Assumption 3.1)} \\
&\quad= \|WH(z-z')\|_2 + \delta\|W\|_2\|z-z'\|_2 \,.
\end{aligned}$$

□

### B.3.3 Proof of Lemma B.5

**Lemma B.5.** *Suppose $\mathcal{T}_\lambda$ is the reparameterization operator of a mean-field location-family and Assumption 2.2 holds. Then, for any matrix $H \in \mathbb{R}^{d\times d}$ and any $\lambda, \lambda' \in \mathbb{R}^d \times \mathbb{D}^d$,*

$$\|\boldsymbol{U}H(\mathcal{T}_\lambda(\boldsymbol{u}) - \mathcal{T}_{\lambda'}(\boldsymbol{u}))\|_2^2 \le r_4\|H\|_2^2\|\lambda - \lambda'\|_2^2 \,.$$

*Proof.* For clarity, let us denote $\bar{C} \triangleq C - C'$ and $\bar{m} \triangleq m - m'$ such that

$$\begin{aligned}\mathcal{T}_\lambda(\boldsymbol{u}) - \mathcal{T}_{\lambda'}(\boldsymbol{u}) &= (C\boldsymbol{u} + m) - (C'\boldsymbol{u} + m') \\ &= (C - C')\boldsymbol{u} + (m - m') \\ &= \bar{C}\boldsymbol{u} + \bar{m} \,.\end{aligned}$$

Then

$$\begin{aligned}\|\boldsymbol{U}H(\mathcal{T}_\lambda(\boldsymbol{u}) - \mathcal{T}_{\lambda'}(\boldsymbol{u}))\|_2^2 &= \|\boldsymbol{U}H\big(\bar{C}\boldsymbol{u} + \bar{m}\big)\|_2^2 \\ &= \underbrace{\mathbb{E}\|\boldsymbol{U}H\bar{C}\boldsymbol{u}\|_2^2}_{V_{\text{scale}}} + 2\underbrace{\langle \boldsymbol{U}H\bar{m}, H\bar{C}\boldsymbol{u}\rangle}_{V_{\text{cross}}} + \underbrace{\mathbb{E}\|\boldsymbol{U}H\bar{m}\|_2^2}_{V_{\text{loc}}} \,.\end{aligned}$$

$V_{\text{loc}}$ and $V_{\text{cross}}$ are straightforward. Under Assumption 2.2, it immediately follows that

$$\begin{aligned}V_{\text{loc}} &= \mathbb{E}\|\boldsymbol{U}H\bar{m}\|_2^2 \\ &= \bar{m}^\top H^\top \mathbb{E}U^2 H\bar{m} \\ &= \bar{m}^\top H^\top H\bar{m} \qquad \text{(Lemma A.2)} \\ &= \|H\bar{z}\|_2^2 \,.\end{aligned}$$

On the other hand,

$$\begin{aligned}V_{\text{cross}} &= \mathbb{E}\langle \boldsymbol{U}H\bar{m},\, \boldsymbol{U}H\bar{C}\boldsymbol{u}\rangle \\ &= \bar{m}^\top H^\top \big(\mathbb{E}\boldsymbol{U}^2 H\bar{C}\boldsymbol{u}\big) \,.\end{aligned}$$

The expectation follows as

$$\begin{aligned}\big[\mathbb{E}U^2 H\bar{C}\boldsymbol{u}\big]_i &= \mathbb{E}\boldsymbol{u}_i^2 \sum_{j=1}^d H_{ij}\bar{C}_{jj}\boldsymbol{u}_j \\ &= H_{ii}\bar{C}_{ii}\mathbb{E}\boldsymbol{u}_i^3 + \sum_{j\neq i} H_{ij}\bar{C}_{jj}\mathbb{E}\boldsymbol{u}_i^2\mathbb{E}\boldsymbol{u}_j \\ &= 0 \,. \qquad \text{(Assumption 2.2)}\end{aligned}$$

Thus, the cross term $V_{\text{cross}}$ vanishes.

$V_{\text{scale}}$ requires careful elementwise inspection in order to apply Assumption 2.2. That is,

$$\begin{aligned}V_{\text{scale}} &= \mathbb{E}\|\boldsymbol{U}H\bar{C}\boldsymbol{u}\|_2^2 \\ &= \mathbb{E}\sum_{i=1}^d \boldsymbol{u}_i^2 \left\{\sum_{j=1}^d H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right\}^2 \\ &= \mathbb{E}\sum_{i=1}^d \boldsymbol{u}_i^2 \left\{H_{ii}\bar{C}_{ii}\boldsymbol{u}_i + \sum_{j\neq i} H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right\}^2 \\ &= \mathbb{E}\sum_{i=1}^d \boldsymbol{u}_i^2 \left\{H_{ii}^2\bar{C}_{ii}^2\boldsymbol{u}_i^2 + 2H_{ii}\bar{C}_{ii}\boldsymbol{u}_i\left(\sum_{j\neq i} H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right) + \left(\sum_{j\neq i} H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right)^2\right\} \qquad \text{(expand quadratic)} \\ &= \sum_{i=1}^d \left\{H_{ii}^2\bar{C}_{ii}^2\mathbb{E}\boldsymbol{u}_i^4 + 2H_{ii}\bar{C}_{ii}\mathbb{E}\boldsymbol{u}_i^3\mathbb{E}\left(\sum_{j\neq i} H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right) + \mathbb{E}\boldsymbol{u}_i^2\mathbb{E}\left(\sum_{j\neq i} H_{ij}\bar{C}_{jj}\boldsymbol{u}_j\right)^2\right\} \qquad \text{(distribute } \boldsymbol{u}_i^2\text{)}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{d} \left\{ r_4 H_{ii}^2 \bar{C}_{ii}^2 + \mathbb{E}\left( \sum_{j \neq i} H_{ij} \bar{C}_{jj} \mathsf{u}_j \right)^2 \right\} && \text{(Assumption 2.2)} \\
&= \sum_{i=1}^{d} \left\{ r_4 H_{ii}^2 \bar{C}_{ii}^2 + \sum_{j \neq i} \left( H_{ij}^2 \bar{C}_{jj}^2 \mathbb{E}\mathsf{u}_j^2 + \sum_{k \neq j} H_{ij} \bar{C}_{jj} \mathbb{E}\mathsf{u}_j H_{ik} \bar{C}_{kk} \mathbb{E}\mathsf{u}_k \right) \right\} && \text{(expand quadratic)} \\
&= \sum_{i=1}^{d} \left\{ r_4 H_{ii}^2 \bar{C}_{ii}^2 + \sum_{j \neq i} H_{ij}^2 \bar{C}_{jj}^2 \right\} && \text{(Assumption 2.2)} \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} H_{ij}^2 \bar{C}_{jj}^2 + (r_4 - 1) \sum_{i=1}^{d} H_{ii}^2 \bar{C}_{ii}^2 \\
&= \|H\bar{C}\|_{\mathrm{F}}^2 + (r_4 - 1) \|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 \, .
\end{aligned}
$$

Combining everything,

$$
\begin{aligned}
\|H(\mathcal{T}_\lambda(\mathsf{u}) - z)\|_{U^2}^2 &= V_{\text{loc}} + 2V_{\text{cross}} + V_{\text{scale}} \\
&= \|H\bar{m}\|_2^2 + \|H\bar{C}\|_{\mathrm{F}}^2 + (r_4 - 1)\|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 \qquad (37)
\end{aligned}
$$

From the property of the Frobenius norm, for any matrix $A \in \mathbb{R}^{d \times d}$, we can decompose

$$
\|A\|_{\mathrm{F}}^2 \quad = \quad \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij}^2 \quad = \quad \sum_{i=1}^{d} A_{ii}^2 + \sum_{i=1}^{d} \sum_{i \neq j} A_{ij}^2 \quad = \quad \|\mathrm{diag}(A)\|_{\mathrm{F}}^2 + \|\mathrm{off}(A)\|_{\mathrm{F}}^2 \, ,
$$

where $\mathrm{off}(A)$ is a function that zeroes-out the diagonal of $A$. Then from Eq. (37),

$$
\begin{aligned}
\|H(\mathcal{T}_\lambda(\mathsf{u}) - z)\|_{U^2}^2 &= \|H\bar{m}\|_2^2 + \|\mathrm{off}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 + \|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 + (r_4 - 1)\|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 \\
&= \|H\bar{m}\|_2^2 + \|\mathrm{off}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 + r_4 \|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 \\
&\leq r_4 \|H\bar{m}\|_2^2 + r_4 \|\mathrm{off}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 + r_4 \|\mathrm{diag}\big(H\bar{C}\big)\|_{\mathrm{F}}^2 && \text{(Lemma A.1)} \\
&= r_4 \big( \|H\bar{m}\|_2^2 + \|H\bar{C}\|_{\mathrm{F}}^2 \big) \\
&\leq r_4 \|H\|_2^2 \big( \|\bar{m}\|_2^2 + \|\bar{C}\|_{\mathrm{F}}^2 \big) && \text{(operator norm)} \\
&= r_4 \|H\|_2^2 \|\lambda - \lambda'\|_2^2 \, ,
\end{aligned}
$$

which is the stated result. □

### B.3.4 Proof of Proposition 4.2

For any $\mu, L \in (0, \infty)$ such that $\mu \leq L$, our goal is to obtain a matrix-valued function $H_{\text{worst}} : \mathbb{R}^d \to \mathbb{S}^d_{\succ 0}$ satisfying

$$\mu \mathrm{I}_d \quad \preceq \quad H_{\text{worst}} \quad \preceq \quad L\mathrm{I}_d$$

that, under the choice $H = H_{\text{worst}}$, maximizes the quantity

$$\left\| \mathit{U} \int_0^1 H(\mathit{z}^w)(\mathit{z} - \bar{z}) \mathrm{d}w \right\|_2^2 , \tag{38}$$

where $\bar{z} \in \{z \mid \nabla \ell(z) = 0\}$ is any stationary point of $\ell$, $\mathit{z} \triangleq \mathcal{T}_\lambda(\mathit{u})$, and $\mathit{z}^w \triangleq w\mathit{z} + (1-w)\bar{z}$. Given the norm constraint, the worst-case example that maximizes Eq. (38) will be the matrix-valued function that approximately results in

$$\left\| \mathit{U} \int_0^1 H(\mathit{z}^w)(\mathit{z} - \bar{z}) \mathrm{d}w \right\|_2^2 \quad \asymp \quad L^2 \|\mathit{U}\|_2^2 \|\mathit{z} - \bar{z}\|_2^2$$

for *any* realization of $\mathit{u}$ on $\mathbb{R}^d$. For this, we will establish the relations

$$\left\| \mathit{U} \int_0^1 H(\mathit{z}^w)(\mathit{z} - \bar{z}) \mathrm{d}w \right\|_2^2 \quad = \quad \|\mathit{U} H(\mathit{z}^w)(\mathit{z} - \bar{z})\|_2^2 \quad \asymp \quad L \|\mathit{U}\|_2^2 \|\mathit{z} - \bar{z}\|_2^2 \ . \tag{39}$$

The first equality in Eq. (39) follows from identifying the conditions where $H(\mathit{z}^w)$ is independent of the value of $w$. For the specific choice of

$$m = \bar{z} = 0_d, \qquad C = \mathrm{diag}(\delta, \ldots, \delta), \qquad \text{any } \delta > 0 \ ,$$

$H(\mathit{z}^w)$ is independent of $w$ if it only depends on the quantities

$$\mathit{i}_* = \underset{i=1,\ldots,d}{\arg\max} |\mathit{z}_i^w| \qquad \text{and} \qquad \hat{\mathit{z}}^w \triangleq \frac{\mathit{z}^w}{\|\mathit{z}^w\|_2} \ . \tag{40}$$

That is, with some abuse of notation, $H(\mathit{z}^w) = H(\hat{\mathit{z}}^w, \mathit{i}_*)$.

**Lemma B.6.** *Suppose $m = \bar{z} = 0_d$, and for any $\delta > 0$, $C = \mathrm{diag}(\delta, \ldots, \delta)$. If $H(\mathit{z}^w)$ is a function of only $\mathit{i}_*$ and $\hat{\mathit{z}}^w$, then $H(\mathit{z}^w)$ is constant with respect to $w \in [0, 1]$.*

*Proof.* It suffices to show that, under the stated conditions, the values of $\mathit{i}_*$ and $\hat{\mathit{z}}^w$ are invariant to $w$. For $\hat{\mathit{z}}^w$, this trivially follows from the assumption that $\bar{z} = 0$ as

$$\hat{\mathit{z}}^w = \frac{\mathit{z}^w}{\|\mathit{z}^w\|_2} = \frac{w\mathit{z} + (1-w)\bar{z}}{\|w\mathit{z} + (1-w)\bar{z}\|_2} = \frac{w\mathit{z}}{\|w\mathit{z}\|_2} = \frac{\mathit{z}}{\|\mathit{z}\|_2} \ .$$

For $\mathit{i}_*$, we use the fact that the diagonal matrix $C$ is isotropic as

$$\underset{i=1,\ldots,d}{\arg\max} \ |\mathit{z}_i^w| \quad = \quad \underset{i=1,\ldots,d}{\arg\max} \ w\, C_{ii} |\mathit{u}_i| \quad = \quad \underset{i=1,\ldots,d}{\arg\max} \ w\delta\, |\mathit{u}_i| \quad = \quad \underset{i=1,\ldots,d}{\arg\max} \ |\mathit{u}_i| \ .$$

□

From $H(\mathit{z}^w) = H(\hat{\mathit{z}}^w, \mathit{i}_*)$, the integral in Eq. (38) can be solved as

$$\left\| \mathit{U} \int_0^1 H(\mathit{z}^w)(\mathit{z} - \bar{z}) \mathrm{d}w \right\|_2^2 \quad = \quad \|\mathit{U} H(\mathit{z}^w)(\mathit{z} - \bar{z})\|_2^2 \ .$$

It remains to construct $H$ in a way that depends only on $\hat{\mathit{z}}^w$ and $\mathit{i}_*$ such that

$$\|\mathit{U} H(\mathit{z}^w)(\mathit{z} - \bar{z})\|_2^2 \quad \asymp \quad L \|\mathit{U}\|_2^2 \|\mathit{z} - \bar{z}\|_2^2 \ .$$

Recalling the spectral constraints, this is equivalent to, for all $z \in \mathbb{R}^d$, $H$ solving the equation

$$H(\mathit{i}_*)\mathit{z} = L \, \|\mathit{z}\|_2 \, \mathrm{e}_{\mathit{i}_*} \quad \text{subject to} \quad \mu \mathrm{I}_d \leq H(\mathit{z}^w) \leq L\mathrm{I}_d \ . \tag{41}$$

Notice the equivalence

$$H(z^w)z = L\,\|z\|_2 \mathrm{e}_{i_*} \qquad \Leftrightarrow \qquad H(z^w)\,\frac{z^w}{\|z^w\|_2} = L\mathrm{e}_{i_*}\ .$$

Thus, $\hat{z}^w$ and $i_*$ contain all the information we need. The following matrix-valued function almost solves Eq. (41):

$$H_{\mathrm{worst}}(z) = \alpha \mathrm{I}_d + \frac{\beta}{2}\big(\mathrm{e}_{i_*}\hat{z}^\top + \hat{z}\,\mathrm{e}_{i_*}^\top\big), \quad \text{where} \quad \hat{z} = \frac{z}{\|z\|_2}\ . \tag{42}$$

This function is reminiscent of a householder reflector (Trefethen and Bau, 1997, Eq. 10.4) with some modifications to satisfy the eigenvalue constraint. That is, from the fact that both $\mathrm{e}_{i_*}$ and $\hat{z}$ have a unit norm, it is apparent that this matrix satisfies Assumption 3.1 with $H = \alpha \mathrm{I}_d$ and $\delta = \beta$. Furthermore, by setting the constants as

$$\alpha = \frac{L+\mu}{2} \quad \text{and} \quad \beta = \frac{L-\mu}{2}\ , \tag{43}$$

the triangle inequality asserts that the eigenvalue constraint $\mu \mathrm{I}_d \preceq H_{\mathrm{worst}} \preceq L\mathrm{I}_d$ is satisfied almost surely.

Given the specific form of $H_{\mathrm{worst}}$, we are now ready to formally prove Proposition 4.2. Let us first restate the proposition for convenience and then proceed to the proof.

**Proposition 4.2.** *Suppose Assumption 2.2 holds and $\mathcal{Q}$ is a mean-field location-scale family. Then, for any $t > 0$, $d > 0$, $\mu, L \in (0, +\infty)$ satisfying $\mu \leq L$, there exists a matrix-valued function $H(z) : \mathbb{R}^d \to \mathbb{S}^d_{\succ 0}$ satisfying $\mu \mathrm{I}_d \preceq H \preceq L\mathrm{I}_d$ almost surely and a set of parameters $\lambda = (m, C) \in \mathbb{R}^d \times \mathbb{D}^d_{\succ 0}$ such that*

$$\mathbb{E}\big\|U\textstyle\int_0^1 H(z^w)(z-\bar{z})\mathrm{d}w\big\|_2^2 \geq \left\{\frac{(L-\mu)^2}{4} - \frac{L^2}{2}\frac{\mathbb{E}\max_{i=1,\ldots,d} u_i^4}{d}\right\} c(t,\varphi)\Big\{\mathbb{E}\max_{i=1,\ldots,d-1} u_i^2 - t\Big\}\|C\|_{\mathrm{F}}^2\ .$$

*where $c(t,\varphi) > 0$ is a constant only dependent on $t$ and $\varphi$.*

*Proof.* Recall $H_{\mathrm{worst}}$ in Eq. (42). By inspection, we know that $H_{\mathrm{worst}}(z^w)$ only depends on the quantities $i_*$ and $z^w$. Then Lemma B.6 states that $w \mapsto H_{\mathrm{worst}}(z^w)$ is a constant function. Therefore,

$$\begin{aligned}
\mathbb{E}\left\|U\int_0^1 H_{\mathrm{worst}}(z^w)(z-\bar{z})\mathrm{d}w\right\|_2^2 &= \mathbb{E}\big\|UH_{\mathrm{worst}}\big(z^0\big)(z-\bar{z})\big\|_2^2 && \text{(Lemma B.6)}\\
&= \mathbb{E}\left\|U\left(\alpha\mathrm{I}_d + \frac{\beta}{2}\big(\mathrm{e}_{i_*}\hat{z}^\top + \hat{z}\,\mathrm{e}_{i_*}^\top\big)\right)z\right\|_2^2 . && \text{(Eq. (42))} \quad (44)
\end{aligned}$$

This can be decomposed as

$$\begin{aligned}
&\mathbb{E}\left\|U\left(\alpha\mathrm{I}_d + \frac{\beta}{2}\big(\mathrm{e}_{i_*}\hat{z}^\top + \hat{z}\,\mathrm{e}_{i_*}^\top\big)\right)z\right\|_2^2\\
&= \mathbb{E}\left\|\alpha Uz + \frac{\beta}{2}U\mathrm{e}_{i_*}\big(\hat{z}^\top z\big) + \frac{\beta}{2}U\hat{z}\,\big(\mathrm{e}_{i_*}^\top z\big)\right\|_2^2\\
&= \mathbb{E}\left\|\alpha Uz + \frac{\beta}{2}U\mathrm{e}_{i_*}\|z\|_2 + \frac{\beta}{2}U\hat{z}z_{i_*}\right\|_2^2\\
&= \mathbb{E}\left\|\alpha Uz + \left(\frac{\beta}{2}\|z\|_2\right)U\mathrm{e}_{i_*} + \left(\frac{\beta}{2}\hat{z}_{i_*}\right)Uz\right\|_2^2\\
&= \mathbb{E}\left\|\left(\frac{\beta}{2}\|z\|_2\right)U\mathrm{e}_{i_*} + \left(\alpha + \frac{\beta}{2}\hat{z}_{i_*}\right)Uz\right\|_2^2\\
&= \mathbb{E}\left[\frac{\beta^2}{4}\|z\|_2^2\|U\mathrm{e}_{i_*}\|_2^2 + \left(\alpha + \frac{\beta}{2}\hat{z}_{i_*}\right)^2\|Uz\|_2^2 + \beta\left(\alpha + \frac{\beta}{2}\hat{z}_{i_*}\right)\|z\|_2\big(\mathrm{e}_{i_*}^\top U^2 z\big)\right]
\end{aligned}$$

$$= \mathbb{E}\left[\frac{\beta^2}{4} u_{i_*}^2 \|z\|_2^2\right] + \underbrace{\mathbb{E}\left[\left(\alpha + \frac{\beta}{2}\hat{z}_{i_*}\right)^2 \|Uz\|_2^2\right]}_{\triangleq V_1} + \underbrace{\mathbb{E}\left[\beta\left(\alpha + \frac{\beta}{2}\hat{z}_{i_*}\right)\|z\|_2 \left(\mathrm{e}_{i_*}^\top U^2 z\right)\right]}_{\triangleq V_2} . \tag{45}$$

Here, the first term $\beta^2/4 u_{i_*}^2 \|z\|_2^2$ is the worst-case behavior we expect from solving Eq. (41). The remaining terms $V_1$ and $V_2$ are the error caused by inexactly solving Eq. (41). It suffices to show that $\beta^2/4 u_{i_*}^2 \|z\|_2^2$ dominates lower bounds on $V_1$ and $V_2$ asymptotically in $L$ and $d$.

$V_1 \geq 0$ trivially holds and can immediately be lower-bounded. $V_2$, on the other hand, is not necessarily non-negative. Therefore, we will use the bound $V_2 \geq -|\mathbb{E}V_2|$.

$$\begin{aligned}
|\mathbb{E}V_2| &\leq \beta\left(\alpha + \frac{\beta}{2}\right)\mathbb{E}\|z\|_2 \, |\mathrm{e}_{i_*}^\top U^2 z| \\
&\leq \beta\left(\alpha + \frac{\beta}{2}\right)\mathbb{E}\|z\|_2 \, u_{i_*}^2 |z_{i_*}| \\
&\leq \beta\left(\alpha + \frac{\beta}{2}\right)\left(\mathbb{E}\|z\|_2^2 u_{i_*}^2\right)^{1/2}\Big(\underbrace{\mathbb{E}u_{i_*}^2 z_{i_*}^2}_{\triangleq V_3}\Big)^{1/2} . && \text{(Cauchy-Schwarz)}
\end{aligned} \tag{46}$$

For $V_3$, we can use an argument similar to Eq. (35) where we distribute the influence of the maximum coordinate over the $d$ coordinates.

$$\begin{aligned}
V_3 = \mathbb{E}u_{i_*}^2 z_{i_*}^2 &= \mathbb{E}\sum_{i=1}^d u_{i_*}^2 z_i^2 \, \mathbb{1}\{i_* = i\} \\
&= \sum_{i=1}^d \mathbb{E}\left[u_{i_*}^2 C_{ii}^2 u_i^2 \, \mathbb{1}\{i_* = i\}\right] \\
&= \sum_{i=1}^d C_{ii}^2 \mathbb{E}\left[u_{i_*}^4\right]\mathbb{E}[\mathbb{1}\{i_* = i\}] && (u_{i_*} \perp\!\!\!\perp i_*) \\
&= \sum_{i=1}^d C_{ii}^2 \mathbb{E}\left[u_{i_*}^4\right]\mathbb{P}[i_* = i] \\
&= \frac{1}{d}\mathbb{E}\left[u_{i_*}^4\right]\|C\|_{\mathrm{F}}^2 \\
&= \frac{1}{d}\left(\mathbb{E}u_{i_*}^4\right)\mathbb{E}\|z\|_2^2 .
\end{aligned} \tag{47}$$

The last equality follows by applying Lemma A.2 to the identity $\mathbb{E}\|z\|_2^2 = \mathbb{E}u^\top C^\top C u$. By applying Eq. (47) into Eq. (46), we can now notice that $V_2$ decreases by a factor of $\mathbb{E}u_{i_*}^4/d$.

$$\begin{aligned}
\mathbb{E}V_2 &\geq -\beta\left(\alpha + \frac{\beta}{2}\right)\frac{\mathbb{E}\left[u_{i_*}^4\right]}{d}\sqrt{\mathbb{E}u_{i_*}^2\|z\|_2^2}\,\sqrt{\mathbb{E}\|z\|_2^2} \\
&\geq -\beta\left(\alpha + \frac{\beta}{2}\right)\frac{\mathbb{E}\left[u_{i_*}^4\right]}{d}\mathbb{E}\left[u_{i_*}^2\|z\|_2^2\right] . && \text{(Assumption 2.2)}
\end{aligned}$$

It is clear that $V_2$ vanishes as $d \to \infty$.

Applying the lower bound on $V_2$ into Eqs. (44) and (45), we have

$$\begin{aligned}
&\mathbb{E}\left\|U\int_0^1 H_{\mathrm{worst}}(z^w)(z - \bar{z})\mathrm{d}w\right\|_2^2 \\
&\geq \frac{\beta^2}{4}\mathbb{E}\left[u_{i_*}^2\|z\|_2^2\right] - \beta\left(\alpha + \frac{\beta}{2}\right)\frac{\mathbb{E}\left[u_{i_*}^4\right]}{d}\mathbb{E}\left[u_{i_*}^2\|z\|_2^2\right] \\
&= \left\{\frac{\beta^2}{4} - \left(\alpha\beta + \frac{\beta^2}{2}\right)\frac{\mathbb{E}\left[u_{i_*}^4\right]}{d}\right\}\mathbb{E}u_{i_*}^2\|z\|_2^2
\end{aligned}$$

$$= \left\{ \frac{(L-\mu)^2}{16} - \left( \frac{L^2-\mu^2}{4} + \frac{(L-\mu)^2}{8} \right) \frac{\mathbb{E}\big[u_{i_*}^4\big]}{d} \right\} \mathbb{E} u_{i_*}^2 \|z\|_2^2 \qquad \text{(Eq. (43))}$$

$$\geq \left\{ \frac{(L-\mu)^2}{16} - \left( \frac{L^2-\mu^2}{4} + \frac{L^2+\mu^2}{4} \right) \frac{\mathbb{E}\big[u_{i_*}^4\big]}{d} \right\} \mathbb{E} u_{i_*}^2 \|z\|_2^2 \qquad \text{(Young's inequality)}$$

$$= \left\{ \frac{(L-\mu)^2}{16} - \frac{L^2}{2} \frac{\mathbb{E} \max_{i=1,\dots,d} u_i^4}{d} \right\} \mathbb{E} u_{i_*}^2 \|Cu\|_2^2 \,. \tag{48}$$

It remains to solve the expectation.

Let us decompose the events where the $i$th coordinate attains the maximum ($i_* = i$) or not ($i_* \neq i$) as done in Lemma 4.1.

$$\begin{aligned}
\mathbb{E} u_{i_*}^2 \|Cu\|_2^2 &= \mathbb{E}\left[ \sum_{i=1}^d C_{ii} u_i^2 u_{i_*}^2 \right] \\
&= \sum_{i=1}^d C_{ii}^2 \Big\{ \mathbb{E}\big[u_i^2 u_{i_*}^2 \mathbb{1}_{i_*=i}\big] + \mathbb{E}\big[u_i^2 u_{i_*}^2 \mathbb{1}_{i_* \neq i}\big] \Big\} \\
&\geq \sum_{i=1}^d C_{ii}^2 \mathbb{E}\big[u_i^2 u_{i_*}^2 \mathbb{1}_{i_* \neq i}\big] \,.
\end{aligned}$$

We are left with the expectation over the event $i_* \neq i$. For the upper bound in Lemma 4.1, the expectation was solved by noticing that $u_i^2$ and $u_{i_*}^2$ can be made independent after upper bounding the indicator. For a lower bound, however, breaking up the expectation for $u_i^2$ and $u_{i_*}^2$ is more involved.

$$\begin{aligned}
\mathbb{E}\big[u_i^2 u_{i_*}^2 \, \mathbb{1}_{i_*=i}\big] &= \mathbb{E}\left[ u_i^2 \max_{j \neq i} u_j^2 \, \mathbb{1}_{i_*=i} \right] \\
&= \mathbb{E}\left[ u_i^2 \max_{j \neq i} u_j^2 \, \mathbb{1}\left\{ u_i^2 < \max_{j \neq i} u_j^2 \right\} \right] \,.
\end{aligned} \tag{49}$$

By introducing a free variable $t > 0$, we can break up the indicator

$$\begin{aligned}
\mathbb{1}\left\{ u_i^2 < \max_{j \neq i} u_j^2 \right\} &\geq \mathbb{1}\left\{ u_i^2 < \max_{j \neq i} u_j^2, \, t < \max_{j \neq i} u_j^2 \right\} \\
&\geq \mathbb{1}\left\{ u_i^2 < t, \, \max_{j \neq i} u_j^2 > t \right\} \\
&= \mathbb{1}\big\{ u_i^2 < t, \, \big\} \mathbb{1}\Big\{ \max_{j \neq i} u_j^2 > t \Big\} \,.
\end{aligned} \tag{50}$$

This then allows the expectation to break up between terms depending on $u_i^2$ and $\max_{j \neq i} u_j$, which is the independence that we were after. That is, applying Eq. (50) to Eq. (49),

$$\begin{aligned}
\mathbb{E}\big[u_i^2 u_{i_*}^2 \, \mathbb{1}_{i_*=i}\big] &\geq \mathbb{E}\left[ u_i^2 \max_{j \neq i} u_j^2 \, \mathbb{1}\big\{ u_i^2 < t, \, \big\} \mathbb{1}\Big\{ \max_{j \neq i} u_j^2 > t \Big\} \right] \\
&= \mathbb{E}\big[ u_i^2 \, \mathbb{1}\big\{ u_i^2 < t \big\} \big] \mathbb{E}\left[ \max_{j=1,\dots,d-1} u_j^2 \, \mathbb{1}\Big\{ \max_{j=1,\dots,d-1} u_j^2 > t \Big\} \right] \\
&= \mathbb{E}\big[ u_i^2 \, \mathbb{1}\big\{ u_i^2 < t \big\} \big] \left( \mathbb{E}\left[ \max_{j=1,\dots,d-1} u_j^2 \right] - \mathbb{E}\left[ \max_{j=1,\dots,d-1} u_j^2 \, \mathbb{1}\Big\{ \max_{j=1,\dots,d-1} u_j^2 \leq t \Big\} \right] \right) \\
&\geq \left( \int_0^t \mathbb{P}\big[u_i^2 > s\big] \mathrm{d}s \right) \left( \mathbb{E}\left[ \max_{j=1,\dots,d-1} u_j^2 \right] - t \right) \,.
\end{aligned}$$

Notice that the function $(t, \varphi) \mapsto \int_0^t \mathbb{P}\big[u_i^2 > s\big] \mathrm{d}s$ is strictly positive as long as $t > 0$ and only dependent on $t$ and the base distribution $\varphi$.

We now obtain our final result by combining the results into Eq. (48). With explicit constants,

$$\mathbb{E}\left\|\int_0^1 H_{\mathrm{worst}}(\mathsf{z}^w)(\mathsf{z}-\bar{z})\mathrm{d}w\right\|_{\mathsf{U}^2}^2 \geq \left\{\frac{(L-\mu)^2}{4} - \frac{L^2}{2}\frac{\mathbb{E}\max_{i=1,\ldots,d}\mathsf{u}_i^4}{d}\right\} \times \left(\int_0^t \mathbb{P}\big[\mathsf{u}_i^2 > s\big]\mathrm{d}s\right)\left(\mathbb{E}\left[\max_{i=1,\ldots,d-1}\mathsf{u}_i^2\right] - t\right)\|C\|_{\mathrm{F}}^2 \,.$$

Substituting $c(t,\varphi) \triangleq \int_0^t \mathbb{P}\big[\mathsf{u}_i^2 > s\big]\mathrm{d}s$ into this yields the stated result. □