Linear Convergence of Black-Box Variational Inference: Should We Stick the Landing?

Kyurae Kim
University of Pennsylvania

Yi-An Ma
University of California San Diego

Jacob R. Gardner
University of Pennsylvania

Abstract

We prove that black-box variational inference (BBVI) with control variates, particularly the sticking-the-landing (STL) estimator, converges at a geometric (traditionally called "linear") rate under perfect variational family specification. In particular, we prove a quadratic bound on the gradient variance of the STL estimator, one which encompasses misspecified variational families. Combined with previous works on the quadratic variance condition, this directly implies convergence of BBVI with the use of projected stochastic gradient descent. For the projection operator, we consider a domain with triangular scale matrices, which the projection onto is computable in $\Theta(d)$ time, where d is the dimensionality of the target posterior. We also improve existing analysis on the regular closed-form entropy gradient estimators, which enables comparison against the STL estimator, providing explicit non-asymptotic complexity guarantees for both.

1 INTRODUCTION

Despite the massive success of black-box variational inference (BBVI; Kucukelbir et al., 2017; Ranganath et al., 2014; Titsias and Lázaro-Gredilla, 2014), our understanding of its computational properties has only recently started to make progress (Domke, 2019, 2020; Domke et al., 2023a; Hoffman and Ma, 2020; Kim et al., 2023a,b). Notably, Domke et al. (2023a); Kim et al. (2023a) have independently established the convergence of "full" BBVI. This is a significant advance from the previous results where simplified versions of BBVI were analyzed (Bhatia et al., 2022; Hoffman and Ma, 2020) and results that *a-priori* assumed regularity of the ELBO (Alquier and Ridgway, 2020; Buchholz et al., 2018; Chérief-Abdellatif et al., 2019; Fujisawa and Sato, 2021; Khan et al., 2016, 2015; Liu and Owen, 2021; Regier et al., 2017). We now have rigorous convergence guarantees that, for certain well-behaved posteriors, BBVI achieves a convergence rate of $\mathcal{O}(1/T)$, corresponding to a computational complexity of $\mathcal{O}(1/\epsilon)$ (Domke et al., 2023a; Kim et al., 2023a). A remaining theoretical question is whether BBVI can achieve better rates, in particular geometric convergence rates, which is traditionally called "linear" convergence in the optimization literature (see the textbook by Nesterov 2004, §1.2.3), corresponding to a complexity of $\mathcal{O}(\log(1/\epsilon))$.

For stochastic gradient descent (SGD; Bottou, 1999; Nemirovski et al., 2009; Robbins and Monro, 1951), it is known that improving the $\mathcal{O}(1/T)$ convergence rate is challenging (Harvey et al., 2019; Rakhlin et al., 2012). This is because, once in the stationary regime, it is necessary to either decrease the stepsize or average the iterates, where the latter reduces SGD to Markov chain Monte Carlo (Dieuleveut et al., 2020). Not surprisingly, both cases result in a significant slowdown compared to deterministic gradient descent. Overall, SGD is known to achieve $\mathcal{O}\left(1/\sqrt{T}\right)$ for general convex functions and $\mathcal{O}(1/T)$ for strongly convex functions (Moulines and Bach, 2011; Nemirovski et al., 2009; Shalev-Shwartz et al., 2011; for more modern analysis techniques, see Garrigos and Gower, 2023).

Meanwhile, under a condition known as "interpolation," which assumes that the gradient variance becomes zero at the optimum, SGD is known to achieve a linear convergence rate (Schmidt and Roux, 2013). This can automatically hold for certain problems, such as empirical risk minimization (ERM) with overparameterized models, explaining the fast empirical convergence of modern machine learning models (Ma et al., 2018; Vaswani et al., 2019). Also, control variate methods such as "variance-reduced" gradients (Gower et al., 2020; Johnson and Zhang, 2013; Schmidt et al.,

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

2017) algorithmically achieve the same effect and have been successful both in theory and practice. Unfortunately, variance-reduced gradient methods are strictly restricted to the finite-sum setting, which BBVI is not part of (See §2.4 by Kim et al., 2023a). Thus, it is yet unclear how BBVI could benefit from the advances in variance-reduced gradients.

Fortunately, other types of control variates have been actively pursued in BBVI (Geffner and Domke, 2018, 2020a; Miller et al., 2017; Paisley et al., 2012; Ranganath et al., 2014; Wang et al., 2024). In particular, the *sticking-the-landing* (STL; Roeder et al., 2017) estimator satisfies the interpolation condition (e.g., achieves zero gradient variance at the optimum) when the variational family $\mathcal{Q}$ contains the true posterior π [1]. It is thus natural to ask whether existing control variate approaches such as STL are sufficient to achieve linear convergence under realizable conditions. In fact, this possibility has been mentioned by Hoffman and Ma (2020, §5).

In this work, we confirm these conjectures by establishing a linear convergence rate of BBVI with STL when the variational family contains the true posterior–*i.e.*, is perfectly specified. For a d-dimensional strongly log-concave posterior with a condition number of κ and a location-scale variational family with a full rank scale, BBVI with the STL estimator finds variational parameters ϵ-close to the global optimum at a rate of $\mathcal{O}\left(d\kappa^2 \log\left(1/\epsilon\right)\right)$. Even beyond the perfectly specified setting, our theoretical results characterize the behavior of the STL estimator in the misspecified setting, which is closer to practice. This provides some intuition as to why the comparisons between the STL and the "standard" closed-form entropy (CFE; Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014) estimators appear mixed in practice (Agrawal et al., 2020; Geffner and Domke, 2020b).

While our results are built on top of the theoretical framework of Domke et al. (2023a), a similar convergence result on the STL estimator appeared in a later, recent preprint version (Domke et al., 2023b) concurrently with this work. The details of the result differ, and we provide additional contributions specific to the STL estimator. We discuss the differences in more detail in Appendix B along with other related works.

Contributions Our contributions are summarized in the following list. An overview of the theorems is provided in Table 1 of Appendix A. We also provide an overview of previous rigorous complexity analyses on BBVI in Table 2 of Appendix B.

[1] Although the term interpolation does not literally make sense outside of the ERM context, we will stick to this term to stay in line with the SGD literature.

- **We prove that BBVI with the STL estimators can converge at a linear rate.** When the variational family is perfectly specified such that the posterior is contained in the variational family, Theorem 6 establishes this through Theorem 1. This is the first result for "full" BBVI without algorithmic simplification.
- **Our analysis encompasses variational family misspecification.** When the variational family is misspecified, the Fisher-Hyvärinen divergence between the variational posterior and the true posterior captures the behavior of STL.
- **We establish a matching lower bound on the gradient variance.** Our upper bound in Theorem 1 and the concurrent result by Domke et al. (2023b) are proven to be tight by a constant factor through Theorem 3.
- **We improve previously obtained gradient variance bounds for the CFE estimator.** In Theorem 4, we tighten the constants of the bounds previously obtained by Domke et al. (2023a). This makes the theoretical results for the CFE and STL estimators comparable.
- **We provide a parameterization with a projection operator with $\Theta(d)$ complexity.** In § 3.1, we propose a triangular scale parameterization with a corresponding projection operator that can be computed in $\Theta(d)$ time. This improves over the matrix square-root parameterization used by Domke et al. (2023a), which involved a $\mathcal{O}(d^3)$ projection operator based on the singular value decomposition.
- **We prove precise quantitative complexity guarantees for SGD with QV gradient estimators.** We obtain quantatitive non-asymptotic complexity guarantees from the "anytime convergence" results of Domke et al. (2023a).

2 PRELIMINARIES

Notation Random variables are denoted in sans-serif (*e.g.*, x, $\boldsymbol{\mathsf{x}}$), vectors are in bold (*e.g.*, $\boldsymbol{x}$, $\boldsymbol{\mathsf{x}}$), and matrices are in bold capitals (*e.g.* $\boldsymbol{A}$). For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we denote the inner product as $\boldsymbol{x}^\top \boldsymbol{x}$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle$, the ℓ_2 norm as $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}\left(\boldsymbol{A}^\top \boldsymbol{A}\right)}$ denotes the Frobenius norm, and for some matrix $\boldsymbol{B}$, $\boldsymbol{A} \succeq \boldsymbol{B}$ is the Loewner order implying that $\boldsymbol{A} - \boldsymbol{B}$ is a positive semi-definite matrix. $\mathbb{S}^d$, $\mathbb{S}^d_{++}$, $\mathbb{L}^d$, $\mathbb{L}^d_{++}$ are the set of symmetric, positive definite, triangular, and triangular matrices with strictly positive eigenvalues (Cholesky factors). $\sigma_{\min}\left(\boldsymbol{A}\right)$ is the smallest eigenvalue of $\boldsymbol{A}$.

2.1 Variational Inference

Variational inference (VI, Blei et al., 2017; Jordan et al., 1999; Zhang et al., 2019) aims to minimize the exclusive (or reverse) Kullback-Leibler (KL; Kullback and Leibler, 1951) divergence as:

$$\underset{\lambda \in \Lambda}{\text{minimize}} \; \mathrm{D}_{\mathrm{KL}}(q_\lambda, \pi) \triangleq \mathbb{E}_{z \sim q_\lambda} - \log \pi(z) - \mathbb{H}(q_\lambda),$$

where D_{KL} is the KL divergence,
$\mathbb{H}$ is the differential entropy,
π is the (target) posterior distribution,
q_λ is the variational approximation.

For Bayesian posterior inference, the KL divergence is, unfortunately, intractable. Instead, one equivalently minimizes the negative *evidence lower bound* (ELBO; Jordan et al., 1999) F such that:

$$\underset{\lambda \in \Lambda}{\text{minimize}} \; F(\lambda) \triangleq \mathbb{E}_{z \sim q_\lambda} - \log \ell(z) - \mathbb{H}(q_\lambda),$$

where $\ell(z) \propto \pi(z)$ is the unnormalized posterior proportional up to a multiplicative constant. In typical use cases of VI, we only have access to ℓ but not π, and the normalizing constant is intractable.

Black-Box Variational Inference Black-box variational inference (BBVI; Ranganath et al., 2014; Titsias and Lázaro-Gredilla, 2014) minimizes F by leveraging stochastic gradient descent (SGD; Bottou, 1999; Nemirovski et al., 2009; Robbins and Monro, 1951). By obtaining a stochastic estimate $g(\lambda)$ which is *unbiased* as $\mathbb{E}g(\lambda) = \nabla F(\lambda)$, BBVI repeats:

$$\lambda_{t+1} = \mathrm{proj}(\lambda_t - \gamma_t g),$$

where γ_t is called the stepsize. The use of the projection operator $\mathrm{proj}(\cdot)$ forms a subset of the broad SGD framework called *projected* SGD. The convergence of BBVI with projected SGD has recently been established by Domke et al. (2023a).

In addition to the KL divergence, our analysis invokes the Fisher-Hyvärinen divergence (Hyvärinen, 2005; Otto and Villani, 2000):

Definition 1 (**Fisher-Hyvärinen Divergence**). The pth order Fisher-Hyvärinen divergence between two distributions π and q is given as

$$\mathrm{D}_{\mathrm{F}^p}(q, \pi) \triangleq \mathbb{E}_{z \sim q} \|\nabla \log \pi(z) - \nabla \log q(z)\|_2^p.$$

Here, we use the pth order generalization (Huggins et al., 2018) of the original Fisher-Hyvärinen divergence. We denote the standard 2nd order Fisher-Hyvärinen divergence as $\mathrm{D}_{\mathrm{F}}(q, \pi) \triangleq \mathrm{D}_{\mathrm{F}^2}(q, \pi)$. This divergence was first defined by Otto and Villani (2000) (attributed by Zegers, 2015) as the "relative Fisher information" in the context of optimal transport. It was later introduced to the machine learning community by Hyvärinen (2005) for score matching.

2.2 Variational Family

Throughout this paper, we restrict our interest to the location-scale variational family, which has been successfully used by Domke (2019, 2020); Domke et al. (2023a); Fujisawa and Sato (2021); Kim et al. (2023a,b); Titsias and Lázaro-Gredilla (2014) for analyzing the properties of BBVI. It encompasses many practical families such as the Gaussian, Student-t, and elliptical distributions. In particular, the location-scale family is part of the broader reparameterized family:

Definition 2 (**Reparameterized Family**). Let φ be some d-variate distribution. Then, q_λ that can be equivalently represented as

$$z \sim q_\lambda \quad \Leftrightarrow \quad z \overset{d}{=} \mathcal{T}_\lambda(u); \quad u \sim \varphi,$$

where $\overset{d}{=}$ is equivalence in distribution, is said to be part of a reparameterized family generated by the base distribution φ and the reparameterization function $\mathcal{T}_\lambda$.

Naturally, this means we focus on the *reparameterization gradient estimator* (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014), often observed to achieve lower variance than alternatives (Xu et al., 2019). (See Mohamed et al. 2020 for a comprehensive overview.) From this, we obtain the location-scale family through the following reparameterization function:

Definition 3 (**Location-Scale Reparameterization Function**). A mapping $\mathcal{T}_\lambda : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$\mathcal{T}_\lambda(u) \triangleq Cu + m$$

with $\lambda \in \mathbb{R}^p$ containing the parameters for forming the location $m \in \mathbb{R}^d$ and scale $C \in \mathbb{R}^{d \times d}$ is called the location-scale reparameterization function.

For the scale matrix C, various parameterizations are used in practice, as shown by Kim et al. (2023b, Table 1). We discuss our parameterization of choice in § 2.3.

The choice for the base distribution φ completes the specifics of the variational family. For example, choosing φ to be a univariate Gaussian result in the Gaussian variational family. We impose the following general assumptions on the base distribution:

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

Overall, the assumptions on the variational family are collected as follows:

Assumption 2. The variational family is the location-scale family formed by Definitions 2 and 3 with the base distribution φ satisfying Assumption 1.

2.3 Scale Parameterization

For the scale parameterization $\boldsymbol{\lambda} \mapsto \boldsymbol{C}$, in principle, any choice that results in a positive-definite covariance matrix is valid. However, recently, Kim et al. (2023a) have shown that a seemingly innocent choice of parameterization can have a massive impact on computational performance. For example, nonlinear parameterizations can easily break the strong convexity of the ELBO (Kim et al., 2023a), which could have been otherwise obtained (Domke, 2020). Therefore, the scale parameterization is subject to the constraints:

(i) Positive Definiteness: $\boldsymbol{C}\boldsymbol{C}^\top \succ 0$.
This is needed to ensure that $\boldsymbol{C}\boldsymbol{C}^\top$ forms a valid covariance in $\mathbb{S}^d_{++}$.

(ii) Linearity: $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2 = \|\boldsymbol{m} - \boldsymbol{m}'\|_2^2 + \|\boldsymbol{C} - \boldsymbol{C}'\|_{\mathrm{F}}^2$.
As shown by Kim et al. (2023a), this constraint is sufficient to form a μ-strongly convex ELBO from a μ-strongly log-concave posterior.

(iii) Convexity: The mapping $\boldsymbol{\lambda} \mapsto \boldsymbol{C}\boldsymbol{C}^\top$ is convex on Λ_S.
This is needed to ensure that the ELBO is convex whenever the target posterior is log-concave (Domke et al., 2023a; Kim et al., 2023a).

(iv) Smooth Log-Determinant: $\boldsymbol{\lambda} \mapsto \log\det \boldsymbol{C}$ is Lipschitz smooth on Λ_S.

This condition is only required by projected SGD so that the ELBO is Lipschitz smooth on Λ_S.

Domke (2020); Domke et al. (2023a) guaranteed **(iv)** by setting the domain of $\boldsymbol{\lambda}$ to be

$$\left\{ (\boldsymbol{m}, \boldsymbol{C}) \in \mathbb{R}^d \times \mathbb{S}^d_{++} \mid \boldsymbol{C}\boldsymbol{C}^\top \succeq S^{-1}\mathbf{I} \right\},$$

with $S = L$, where L is the log-smoothness constant of the posterior. That is, $\boldsymbol{C}$ is chosen to be a proper matrix square root of the covariance $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$ such that $\boldsymbol{C} = \boldsymbol{C}^\top = \boldsymbol{\Sigma}^{1/2}$. This parameterization ensures that a proper projection operator exists onto Λ_S, where they proposed to use the singular value decomposition. This projection operator is quite costly as it imposes a $\mathcal{O}(d^3)$ complexity. We will later propose a different parameterization based on triangular matrices, where the projection operator only costs $\Theta(d)$ while obtaining the same convergence guarantees.

2.4 Gradient Estimators

The gradient estimators considered in this work are the closed-form entropy (CFE; Kucukelbir et al., 2017; Titsias and Lázaro-Gredilla, 2014) and sticking the landing (STL; Roeder et al., 2017) estimators.

Closed-From Entropy Estimator The CFE estimator is the "standard" estimator used for BBVI.

Definition 4 (Closed-Form Entropy Estimator). The closed-form entropy gradient estimator is

$$\boldsymbol{g}(\boldsymbol{\lambda}) \triangleq \nabla_{\boldsymbol{\lambda}} \log \ell\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right) + \nabla_{\boldsymbol{\lambda}} \mathbb{H}(q_{\boldsymbol{\lambda}}),$$

where the gradient of the entropy term is computed deterministically.

It can be used whenever the entropy $\mathbb{H}(q_{\boldsymbol{\lambda}})$ is available in a closed form. For location-scale families, this is always the case up to an additive constant.

Sticking-the-Landing Estimator On the other hand, the STL estimator estimates the entropy through a special Monte Carlo strategy:

Definition 5 (Sticking-the-Landing Estimator; STL). The sticking-the-landing gradient estimator

$$\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda}) \triangleq \nabla_{\boldsymbol{\lambda}} \log \ell\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right) - \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\nu}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}}$$

is given by stopping the gradient from propagating through $\log q_{\boldsymbol{\lambda}}$.

Notice that, the gradient of $\log q$ is "stopped" by the assignment $\boldsymbol{\nu} = \boldsymbol{\lambda}$. This creates a control variate effect, where the control variate cv is implicitly formed as

$$\mathsf{cv}(\boldsymbol{\lambda}; \boldsymbol{u}) = \nabla_{\boldsymbol{\lambda}} \mathbb{H}(\boldsymbol{\lambda}) + \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\nu}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}}.$$

Subtracting this to the CFE estimator leads to the STL estimator.

2.5 Quadratic Variance Condition

The convergence of BBVI has recently been established concurrently by Domke et al. (2023a); Kim et al. (2023a). However, the analysis of Domke et al. presents a broadly applicable framework based on the *quadratic variance* (QV) condition.

Definition 6 (Quadratic Variance; QV). A gradient estimator $\boldsymbol{g}$ is said to satisfy the quadratic variance condition if the following bound holds:

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{\lambda})\|_2^2 \leq \alpha\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 + \beta,$$

for any $\boldsymbol{\lambda} \in \Lambda_S$ and some $0 \leq \alpha, \beta < \infty$, where $\boldsymbol{\lambda}^*$ is a stationary point.

This basically assumes that the gradient variance grows no more than a quadratic plus a constant. For non-asymptotic analysis of SGD, this bound was first used by Moulines and Bach (2011) as an intermediate step implied by the condition:

$$\mathbb{E}\left\|\boldsymbol{g}\left(\boldsymbol{\lambda}\right)-\boldsymbol{g}\left(\boldsymbol{\lambda}'\right)\right\|_2^2 \leq \mathcal{L}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}'\|_2^2,$$

for all $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ and some $0 < \mathcal{L} < \infty$. This, combined with the assumption $\mathbb{E}\|\boldsymbol{g}\left(\boldsymbol{\lambda}^*\right)\|_2^2 < \infty$, implies the QV condition. They used this strategy to prove the convergence of SGD on strongly convex functions. Later on, Wright and Recht (2021, p. 85) directly assumed the QV condition to obtain similar results. A comprehensive convergence analysis of projected and proximal SGD with the QV condition was conducted by Domke et al. (2023a), where they also prove the convergence on general convex functions. This work will invoke the analysis of Domke et al. by establishing the QV condition of the considered gradient estimators.

2.6 Interpolation Condition

To establish the linear, or more intuitively "exponential," convergence of SGD, Schmidt and Roux (2013) have relied on the interpolation condition:

Definition 7 (**Interpolation**)**.** A gradient estimator $\boldsymbol{g}$ is said to satisfy the interpolation condition if $\mathbb{E}\|\boldsymbol{g}\left(\boldsymbol{\lambda}^*\right)\|^2 = 0$ for $\boldsymbol{\lambda}^* \in \Lambda$ such that $\|\nabla F\left(\boldsymbol{\lambda}^*\right)\| = 0$.

This assumes that the gradient variance vanishes at a stationary point, gradually retrieving the convergence behavior of deterministic gradient descent. For the QV condition, this corresponds to $\beta = 0$.

Achieving "Interpolation" Currently, there are two ways where the interpolation condition can be achieved. The first case is when interpolation is achieved *naturally*. That is, in ERM, when the model is so overparameterized that certain parameters can "interpolate" all of the data points in the train data (Ma et al., 2018; Vaswani et al., 2019), the gradient becomes 0. Otherwise, a control variate approach such as stochastic average gradient (SAG; Schmidt et al., 2017) or stochastic variance-reduced gradient (SVRG; Johnson and Zhang, 2013), and their many variants (Gower et al., 2020) can be used.

Does STL "Interpolate?" As we previously discussed, the STL estimator is essentially a control variate method. Thus, an important question is whether it can achieve the same effect, notably linear convergence, as variance-reduced SGD methods. While Roeder et al. (2017) have already shown that the STL estimator achieves interpolation when $q_{\boldsymbol{\lambda}^*} = \pi$, our research question is whether this fact can be rigorously used to establish linear convergence of SGD.

3 MAIN RESULTS

3.1 Triangular Scale Parameterization

First, we will demonstrate a parameterization that is more computationally efficient than the matrix square-root parameterization considered in § 2.3, while satisfying the constraints **(i)** to **(iv)**. We first turn our attention to the following domain for the variational parameters:

$$\Lambda_S \triangleq \left\{\left(\boldsymbol{m}, \boldsymbol{C}\right) \in \mathbb{R}^d \times \mathbb{L}_{++}^d \mid \sigma_{\min}\left(\boldsymbol{C}\right) \geq 1/\sqrt{S}\right\},$$

where $\mathbb{L}_{++}^d$ is the set of Cholesky factors. A key special case is the mean-field variational family, which is a strict subset of Λ_S, where we restrict $\boldsymbol{C}$ to be diagonal matrices. With that said, we consider the two following parameterizations:

$$\begin{aligned}
&\boldsymbol{C} = \boldsymbol{L}, && \text{(full-rank)}\\
&\boldsymbol{C} = \operatorname{diag}\left(L_{11}, \ldots, L_{dd}\right), && \text{(mean-field)}
\end{aligned}$$

where $\boldsymbol{L}$ is a Cholesky factor. In practice, the triangular matrix parameterization is most commonly used (Kucukelbir et al., 2017), and results in lower gradient variance than the square root parameterization (Kim et al., 2023b).

Projection Operator The key advantage of operating with triangular scale matrices is that the entropy is the log-sum of their eigenvalues, which turns out to be their diagonal elements. This implies that the gradient of the entropy term $\boldsymbol{\lambda} \mapsto \mathbb{H}\left(q_{\boldsymbol{\lambda}}\right)$ only resides on the diagonal subspace of $\boldsymbol{C}$. Therefore, the "smoothness" of $\boldsymbol{\lambda} \mapsto \mathbb{H}\left(q_{\boldsymbol{\lambda}}\right)$ can be achieved by only controlling the eigenvalues (or diagonal elements) of $\boldsymbol{C}$. This sharply contrasts with the square-root parameterization where the constraint is on the *singular values*, which are much harder to control. Nevertheless, the canonical Euclidean projection operator is:

Proposition 1. *The Euclidean projection operator onto* Λ_S, $\operatorname{proj}_{\Lambda_S} : \mathbb{R}^d \times \mathbb{L}^d \to \Lambda_S$, *is given as*

$$\operatorname{proj}_{\Lambda_S}\left(\boldsymbol{\lambda}\right) = \operatorname*{arg\,min}_{\boldsymbol{\lambda}' \in \Lambda_S} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2 = \left(\boldsymbol{m}, \widetilde{\boldsymbol{C}}\right),$$

where $\widetilde{\boldsymbol{C}}$ *is the projection of* $\boldsymbol{C}$ *such that*

$$\widetilde{C}_{ij} = \begin{cases} \max\left(C_{ii},\ 1/\sqrt{S}\right) & \text{for } i = j \\ C_{ij} & \text{for } i \neq j. \end{cases}$$

Proof. Since the eigenvalues of a triangular matrix are its diagonal elements, we notice that Λ_S is a constraint only on the diagonal elements of $\boldsymbol{C}$ such that $C_{ii} \geq 1/\sqrt{S}$. Conveniently, this is an element-wise half-space constraint for which the projection follows as

$$\widetilde{C}_{ii} = \operatorname*{arg\,min}_{c \geq 1/\sqrt{S}} \|C_{ii} - c\|_2^2 = \max\left(C_{ii}, 1/\sqrt{S}\right).$$

□

Theoretical Properties We will now prove that our construction is valid. It is trivial that **(i)** to **(iii)** are satisfied. We formally prove that Λ_S satisfies **(iv)**:

Proposition 2. *The entropy $\boldsymbol{\lambda} \mapsto \mathbb{H}(q_{\boldsymbol{\lambda}})$ is S-Lipschitz smooth on Λ_S.*

Proof. See Appendix C.3.

Given these results, we will hereafter assume projected SGD is run on Λ_S with the projection operator $\mathsf{proj}_{\Lambda_S}$.

3.2 Theoretical Analysis of the STL Estimator

Before presenting our analysis on BBVI gradient estimators, we will discuss a notable aspect of our strategy and the key step in our proof.

Our main assumption on the target posterior is that it is L-log(-Lipschitz) smooth:

Definition 8. π is said to be L-log-(Lipschitz) smooth if its log-density $\log \pi : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz smooth such that

$$\left\|\nabla \log \pi(\boldsymbol{z}) - \nabla \log \pi(\boldsymbol{z}')\right\|_2 \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|_2,$$

for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$ and some $0 < L < \infty$.

If this holds for π, the same bound holds for ℓ as well since they are proportional up to a constant such that $\nabla \log \ell = \nabla \log \pi$. This assumption has been used by Domke (2019) to establish similar results for the CFE estimator and is also widely used in the analysis of sampling algorithms based on log-concave analysis. (See Dwivedi et al. (2019, §2.3) for such example.) For probability measures, log-smoothness implies that the density of π can be upper bounded by some Gaussian. Naturally, this essentially corresponds to assuming π has sub-Gaussian tails.

Adaptive Bounds with the Peter-Paul Inequality Unlike the QV bounds obtained by Domke et al. (2023a), our bounds involve a free parameter $\delta \geq 0$. We call these bounds *adaptive* QV bounds.

Assumption 3 (Adaptive QV). The gradient estimator $\boldsymbol{g}$ satisfies the bound

$$\mathbb{E}\|\boldsymbol{g}(\boldsymbol{\lambda})\|_2^2 \leq (1 + C\delta)\,\widetilde{\alpha}\,\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 + (1 + C^{-1}\delta^{-1})\,\widetilde{\beta},$$

for any $\delta > 0$, any $\boldsymbol{\lambda} \in \Lambda_S$, and some $0 < \widetilde{\alpha}, \widetilde{\beta} < \infty$, where $\boldsymbol{\lambda}^*$ is a stationary point.

This is a consequence of the use of the "Peter-Paul" inequality such that

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

and can be seen as a generalization of the usual inequality $(a + b)^2 \leq 2a^2 + 2b^2$. Adjusting δ can occasionally tighten the analysis. In fact, δ can be optimized to become *adaptive* to the downstream analysis. Indeed, in our complexity analysis, δ automatically trades-off the influence of $\widetilde{\alpha}$ and $\widetilde{\beta}$ according to the accuracy budget ϵ.

Key Lemma The key first step in all of our analysis is the following decomposition:

Lemma 1. *Assume Assumption 2. The expected-squared norm of STL is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \leq (2 + \delta)\,V_1 + (2 + \delta)\,V_2 + \left(1 + 2\delta^{-1}\right)V_3,$$

where the terms are

$$V_1 = \mathbb{E}\,J_{\mathcal{T}}(\boldsymbol{u})\,\|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2$$
$$V_2 = \mathbb{E}\,J_{\mathcal{T}}(\boldsymbol{u})\,\|\nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2$$
$$V_3 = \mathbb{E}\,J_{\mathcal{T}}(\boldsymbol{u})\,\|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2,$$

for any $\delta > 0$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$. $J_{\mathcal{T}} : \mathbb{R}^d \to \mathbb{R}$ is a function depending on the variational family as

$$J_{\mathcal{T}}(\boldsymbol{u}) = 1 + \textstyle\sum_{i=1}^d u_i^2 \quad \textit{for full-rank and}$$
$$J_{\mathcal{T}}(\boldsymbol{u}) = 1 + \sqrt{\textstyle\sum_{i=1}^d u_i^4} \quad \textit{for mean-field.}$$

Proof. See Appendix C.4.1.

Here, $J_{\mathcal{T}}$ is a term that stems from the Jacobian of $\mathcal{T}$. Thus, $J_{\mathcal{T}}$ contains the properties unique to the chosen variational family. V_1 and V_2 measure how far the current variational approximation $q_{\boldsymbol{\lambda}}$ is from a stationary point $\boldsymbol{\lambda}^*$. Thus, both terms will eventually reach 0 as BBVI converges, regardless of family specification. The key is V_3, which captures the amount of mismatch between the score of the true posterior π and variational posterior $q_{\boldsymbol{\lambda}^*}$. Establishing the "interpolation condition" amounts to analyzing when V_3 becomes 0.

3.2.1 Upper Bounds

We now present our upper bound on the expected-squared norm of the STL gradient estimator.

Theorem 1. *Assume Assumption 2 and that ℓ is L-log-smooth. For the full-rank parameterization, the expected-squared norm of the STL estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \leq \alpha_{\mathrm{STL}}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 + \beta_{\mathrm{STL}}$$

where

$$\alpha_{\mathrm{STL}} = (2 + \delta)\left(L^2\left(d + k_\varphi\right) + S^2(d + 1)\right)$$
$$\beta_{\mathrm{STL}} = (1 + 2\delta^{-1})\left(2d + k_\varphi\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \ell)}$$

for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \in \Lambda_S$ and any $\delta > 0$.

Proof. See Appendix C.4.2.

Remark 1 (**Mean-Field Variational Family**). We prove an equivalent result for the mean-field variational family, Theorem 7 in Appendix C.4.3, which has an $\mathcal{O}\left(\sqrt{d}\right)$ dimensional dependence.

Remark 2 (**Interpolation Condition**). Theorem 1 encompasses both settings where the variational family is well-specified and misspecified. That is, when the variational family is well specified, *i.e.*, $\mathrm{D}_{\mathrm{F}^4}\left(q_{\lambda^*}, \pi\right) = 0$, we obtain interpolation such that $\beta_{\mathrm{STL}} = 0$.

Remark 3 (**Adaptivity of Bound**). When the variational family is well specified such that $\mathrm{D}_{\mathrm{F}^4}\left(q_{\lambda^*}, \pi\right) = 0$, we can adaptively tighten the bound by setting $\delta = 0$, where α_{STL} is reduced by a constant factor.

3.2.2 Lower Bounds

We also obtain lower bounds on the expected-squared norm of the STL estimator to analyze its best-case behavior and the tightness of the bound.

Necessary Conditions for Interpolation First, we obtain lower bounds that generally hold for all $\lambda \in \Lambda_L$ and any π. Our analysis relates the gradient variance with the Fisher-Hyvärinen divergence. This can be related back to the KL divergence through an assumption on the posterior π known as the log-Sobolev inequality. The general form of the log-Sobolev inequality was originally proposed by Gross (1975) to study diffusion processes. In this work, we use the form used by Otto and Villani (2000):

Assumption 4 (**Log-Sobolev Inequality; LSI**). π is said to satisfy the log-Sobolev inequality if, for any variational family $\mathcal{Q}$ and all $q_\lambda \in \mathcal{Q}$, the following inequality holds:

$$\mathrm{D}_{\mathrm{KL}}(q, \pi) \leq \frac{C_{\mathrm{LSI}}}{2} \mathrm{D}_{\mathrm{F}}(q, \pi).$$

Strongly log-concave distributions are known to satisfy the LSI, where the strong log-concavity constant becomes the (inverse) LSI constant (see also Villani, 2016, Theorem 9.9):

Remark 4 (Bakry and Émery, 1985). Let π be μ-strongly log-concave. Then, LSI holds with $C_{\mathrm{LSI}}^{-1} = \mu$.

We now present a lower bound which holds for all $\lambda \in \Lambda_S$ and any log-differentiable π:

Theorem 2. *Assume Assumption 2. The expected-squared norm of the STL estimator is lower bounded as*

$$\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\lambda\right)\right\|_2^2 \geq \mathrm{D}_{\mathrm{F}}(q_\lambda, \pi) \geq \frac{2}{C_{\mathrm{LSI}}} \mathrm{D}_{\mathrm{KL}}(q_\lambda, \pi),$$

for all $\lambda \in \Lambda_S$ and any $0 < S < \infty$, where the last inequality holds if π is LSI.

Proof. See Appendix C.5.1.

Corollary 1 (**Necessary Conditions for Interpolation**). *For the STL estimator, the interpolation condition does not hold if*

(i) $\mathrm{D}_{\mathrm{F}}(q_{\lambda_{\mathrm{F}}^*}, \pi) > 0$, *or,*

(ii) *when π is LSI,* $\mathrm{D}_{\mathrm{KL}}(q_{\lambda_{\mathrm{KL}}^*}, \pi) > 0$,

where $\lambda_{\mathrm{F}}^* \in \arg\min_{\lambda \in \Lambda_S} \mathrm{D}_{\mathrm{F}}(q_\lambda, \pi)$, *and* $\lambda_{\mathrm{KL}}^* \in \arg\min_{\lambda \in \Lambda_S} \mathrm{D}_{\mathrm{KL}}(q_\lambda, \pi)$,

for any $0 < S < \infty$.

Tightness Analysis The bound in Theorem 2 is unfortunately not tight regarding the constants. It, however, holds for all λ and π. Instead, we establish an alternative lower bound that holds for some λ and π but is tight regarding the dependence on d and L.

Theorem 3. *Assume Assumption 2. There exists a strongly-convex, L-log-smooth posterior and some variational parameter $\widetilde{\lambda} \in \Lambda_L$ for all $L \geq 1$ such that*

$$\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\widetilde{\lambda}\right)\right\|_2^2 \geq \left(L^2\left(d + k_\varphi\right) - 2\left(d+1\right)\right)\left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 - 2\left(k_\varphi - 1\right)\left\|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right\|_2^2,$$

where $\widetilde{\lambda} = (\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{C}})$ and $\bar{\boldsymbol{z}}$ is a stationary point of the said log posterior.

Proof. See Appendix C.5.2.

Remark 5. Theorem 3 implies that Theorem 1 with $S = L$ is tight with respect to the dimension dependence d and the log-smoothness L except for a factor of 4.

Remark 6 (**Room for Improvement**). Part of the factor of 4 looseness is due to the extreme worst case: when $\nabla \log \pi$ and $\nabla \log q_\lambda$ are anti-correlated. This worst case is unlikely to appear in practice, thus making a tighter lower bound challenging to obtain. But at the same time, we were unsuccessful at seeking a general assumption that would rule out these worst cases in the upper bound. Specifically, we tried very hard to apply coercivity/gradient monotonicity of log-concave distributions, but to no avail, leaving this to future works.

3.3 Theoretical Analysis of the CFE Estimator

We now present the analysis of the CFE estimator. While the CFE estimator has been studied in-depth by Domke (2019); Domke et al. (2023a); Kim et al. (2023b), we slightly improve the latest analysis of Domke et al. (2023a, Theorem 3). Specifically, we improve the constants and obtain an adaptive bound. This ensures that we have a fair comparison with the STL estimator.

Theorem 4. *Assume Assumption 2 and that ℓ is L-log-smooth. For the full-rank parameterization, the expected-squared norm of the CFE estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2 \leq \alpha_{\mathrm{CFE}}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 + \beta_{\mathrm{CFE}}$$

where

$$\alpha_{\mathrm{CFE}} = L^2\left(d+k_\varphi\right)(1+\delta) + (L+S)^2$$
$$\beta_{\mathrm{CFE}} = L^2\left(d+k_\varphi\right)\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2$$

for any $\boldsymbol{\lambda},\boldsymbol{\lambda}^* \in \Lambda_S$ and $\delta > 0$, where $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{z}}, \mathbf{0})$ and $\bar{\boldsymbol{z}}$ is any stationary point of f.

Proof. See Appendix C.6.1.

Remark 7 (Comparison with STL). Compared to the STL estimator, the constant α of the CFE estimator is tighter by a factor of 4. Considering Theorem 3, the constant factor difference should be marginal in practice.

Remark 8 (Intuitions on $\|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2^2$). The quantity $\|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2^2$ can be expressed in the Wasserstein-2 distance as

$$\mathrm{d}_{\mathcal{W}_2}\left(q_{\boldsymbol{\lambda}^*}, \delta_{\bar{\boldsymbol{z}}}\right) = \sqrt{\|\boldsymbol{m}^*-\bar{\boldsymbol{z}}\|_2^2 + \|\boldsymbol{C}^*\|_{\mathrm{F}}^2} = \|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2,$$

where $\delta_{\bar{\boldsymbol{z}}}$ is a delta measure centered on the posterior mode $\bar{\boldsymbol{z}}$. Also, when the variational posterior mean $\boldsymbol{m}^*$ is close to $\bar{\boldsymbol{z}}$ such that $\|\boldsymbol{m}^*-\bar{\boldsymbol{z}}\|_2^2 \approx 0$, $\|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2^2$ corresponds to the variational posterior variance as

$$\|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2^2 \approx \|\boldsymbol{C}^*\|_{\mathrm{F}}^2 = \mathrm{tr}\,\mathbb{V}_{\boldsymbol{z}\sim q_{\boldsymbol{\lambda}^*}}[\boldsymbol{z}].$$

3.4 Non-Asymptotic Complexity of Black-Box Variational Inference

Strongly Log-Concave Posteriors First, let us define the following:

Definition 9. π is said to be μ-strongly log-concave if its log-density $\log\pi : \mathbb{R}^d \to \mathbb{R}$ (equivalently $\log\ell$) satisfies the inequality

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

for all $\boldsymbol{z},\boldsymbol{z}' \in \mathbb{R}^d$ and some $\mu > 0$.

Essentially, this assumes that the log-density of π is μ-strongly convex. It also implies that the density of π is lower bounded by some Gaussian. A consequence of μ-strong log-concavity is that, combined with the constraints on the variational parameterization in § 2.3, the ELBO is also μ-strongly convex (Challis and Barber, 2013; Domke et al., 2023a; Kim et al., 2023a). μ-strongly log-concave posteriors can easily be constructed by combining a log-concave likelihood with a Gaussian prior, and are popularly used to analyze BBVI and sampling algorithms.

Theoretical Setup We now apply the general complexity results for projected SGD established in Appendix C.7 to BBVI. **(i)** strongly log-concave posteriors, **(ii)** SGD run with fixed stepsizes, and **(iii)** the full-rank variational family. This is because the convergence analyses for **(ii)** ∩ **(iii)** are the tightest. Although the bounds for the mean-field parameterization have better dependences on d, so far, it is unknown whether they are tight (Kim et al., 2023b). (See also Kim et al., 2023a, Conjecture 1.)

Complexity of BBVI on Strongly-Log-Concave π We can now plug in the constants obtained in § 3.2. This immediately establishes the iteration complexity resulting from the use of different gradient estimators.

Theorem 5 (Complexity of Fixed Stepsize BBVI with CFE). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the CFE estimator and projected SGD with a fixed stepsize applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T-\boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 2\kappa^2\left(d+k_\varphi+4\right)\left(1+2\|\bar{\boldsymbol{\lambda}}-\boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)\log\left(2\Delta^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta = \|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\|_2$, and $\kappa = L/\mu$ is the condition number.

Proof. See Appendix C.8.1.

In particular, the following theorem establishes that BBVI with the STL estimator can achieve linear convergence under perfect variational family specification.

Theorem 6 (Complexity of Fixed Stepsize BBVI with STL). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the STL estimator and projected SGD with a fixed stepsize applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T-\boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 8\kappa^2\left(d+k_\varphi\right)\left(1+\frac{1}{L^2}\sqrt{\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}^*},\pi\right)}\frac{1}{\epsilon}\right)\log\left(2\Delta^2\frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta = \|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\|_2$ is the distance to the optimum and $\kappa = L/\mu$ is the condition number.

Proof. See Appendix C.8.2.

Corollary 2 (Linear Convergence of BBVI with STL). *If the variational family is perfectly specified such that $\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}}^*,\pi\right) = 0$ for $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$, then BBVI with the STL estimator converges linearly with a complexity of $\mathcal{O}\left(d\kappa^2\log(1/\epsilon)\right)$.*

Remark 9. Convergence is slowed when using a decreasing step size schedule, as shown in Theorem 12. Thus, one does not achieve a linear convergence rate under this schedule even if the variational family is perfectly specified. However, when the variational family is misspecified, this achieves a better rate of $\mathcal{O}(1/\epsilon)$ compared to the $\mathcal{O}(1/\epsilon \log 1/\epsilon)$ of Theorem 6.

Remark 10 (**Variational Family Misspecification**)**.** Under variational family misspecification, STL has an $\mathcal{O}(1/\epsilon)$ dependence on the 4th order Fisher divergence $\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi) > 0$. To compare the computational performance of CFE and STL in this setting, one needs to compare $L^{-2}\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}$ versus $\|\bar{\lambda} - \lambda^*\|_2^2$.

Remark 11. Theorem 7 also implies that the mean-field parameterization improves the dimension dependence to a complexity of $\mathcal{O}\left(\sqrt{d}\kappa^2 \log(1/\epsilon)\right)$.

3.5 Should we stick the landing?

When the variational family is misspecified, it is hard to tell *when* STL would be superior to CFE; the Fisher-Hyvärinen divergence and the posterior variance are fundamentally unrelated quantities. Furthermore, the Fisher-Hyvärinen divergence is hard to interpret apart from some relationships with other divergences (Huggins et al., 2018). Thus, we conclude by providing a characterization of the Fisher-Hyvärinen divergence.

Our final analysis will focus on Gaussian posteriors and the mean-field Gaussian family. In practice, the STL estimator becomes infeasible to use with full-rank variational families as each evaluation of the log-density $\log q_\lambda$ involves a back-substitution with a $\mathcal{O}(d^3)$ cost and numerical stability becomes a concern. Therefore, studying the effect of misspecification of mean-field is particularly relevant.

Proposition 3. *Let $\pi = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{Q}$ be the mean-field Gaussian variational family. Then, the Fisher-Hyvärinen divergence of the KL minimizer*

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathrm{D}_{\mathrm{KL}}(q, \pi)$$

is bounded as

$$\lambda_{\max}(\boldsymbol{D})^{-1} \|\boldsymbol{R}^{-1} - \mathbf{I}\|_{\mathrm{F}}^2 \leq \mathrm{D}_{\mathrm{F}}(q_*, \pi) \leq \lambda_{\min}(\boldsymbol{D})^{-1} \|\boldsymbol{R}^{-1} - \mathbf{I}\|_{\mathrm{F}}^2,$$

where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Sigma})$ and $\boldsymbol{R}$ is the correlation matrix of π such that $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$.

Proof. See Appendix C.9.

Remark 12. For Gaussians, the 4th-order Fisher-Hyvärinen divergence term in Theorem 1 can be replaced by its 2nd-order counterpart. Thus, combined with Theorem 2, the 2nd-order Fisher-Hyvärinen divergence fully characterizes the variance of STL.

Remark 13. Proposition 3 implies that, when approximating a full-rank Gaussian with a mean-field Gaussian, the value of the Fisher-Hyvärinen divergence is tightly characterized by the degree of correlation in the posterior; it will increase indefinitely as the posterior correlation matrix becomes singular.

Remark 14. We have provided a sufficient condition for the STL estimator to perform poorly compared to the CFE estimator. It is foreseeable that alternative types of model misspecification abundant in practice should yield additional sufficient conditions, *i.e.*, tail mismatch, but we leave this to future works.

4 DISCUSSIONS

Empirically Comparing Estimators From our analysis and that of Domke et al. (2023a), it is apparent that for a QV gradient estimator, α and β sufficiently characterize its behavior on log-concave posteriors: α characterizes the convergence speed, while β determines the complexity with respect to ϵ. It is conceivable that *estimating* these quantities in practical settings would provide a principled way to compare and evaluate different estimators. Previously, the signal-to-noise (SNR) ratio have been popularized by Rainforth et al. (2018), and since been used by, for example, by Fujisawa and Sato (2021); Geffner and Domke (2021); Rudner et al. (2021). In contrast to the QV coefficients, a *constant* SNR relates with convergence only through the expected strong growth condition (Schmidt and Roux, 2013; Solodov, 1998; Vaswani et al., 2019), which requires strong assumptions to hold (perfect variational family specification, strong log-concavity). The QV coefficients, α and β, on the other hand, apply to a broader range of settings.

Conclusions We have analyzed the sticking-the-landing (STL) estimator by Roeder et al. (2017). When the variational family is perfectly specified, our complexity guarantees automatically guarantees a logarithmic complexity. Also, from the results of Domke (2019) and Theorem 4 it is known that the gradient variance of CFE at the optimum depends on the mode mismatch $\|\boldsymbol{m}^* - \bar{\boldsymbol{z}}\|_2^2$ plus the variational posterior variance $\|\boldsymbol{C}^*\|_{\mathrm{F}}^2$. We show that the STL estimator instead depends on the Fisher-Hyvärinen divergence of the variational posterior. Furthermore, our work demonstrates that it is possible to rigorously show that control variates can accelerate the convergence of BBVI. It will be interesting to analyze and compare the existing control variates by Geffner and Domke (2020a); Miller et al. (2017); Wang et al. (2024).

Acknowledgements

The authors sincerely thank Jisu Oh (NCSU) for thoroughly proofreading the paper, Justin Domke (UMass Amherst) for discussions on the concurrent results, Kaiwen Wu (UPenn) for helpful discussions, Xi Wang (UMass Amherst) for pointing out a typo, and the anonymous reviewers for comments that improved the readability of the work.

K. Kim was supported by a gift from AWS AI to Penn Engineering's ASSET Center for Trustworthy AI; Y.-A. Ma was funded by the NSF Grants [NSF-SCALE MoDL-2134209] and [NSF-CCF-2112665 (TILOS)], the U.S. Department Of Energy, Office of Science, as well as CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics; J. R. Gardner was supported by NSF award [IIS-2145644].

References

Agrawal, A., Sheldon, D. R. and Domke, J. (2020) Advances in black-box VI: Normalizing flows, importance weighting, and optimization. In *Advances in Neural Information Processing Systems*, vol. 33, 17358–17369. Curran Associates, Inc. (page 2)

Alquier, P. (2021) Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions. In *Proceedings of the International Conference on Machine Learning*, vol. 193 of *PMLR*, 207–218. JMLR. (page 17)

Alquier, P. and Ridgway, J. (2020) Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, **48**, 1475–1497. (pages 1, 17)

Bakry, D. and Émery, M. (1985) Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, vol. 1123, 177–206. Berlin, Heidelberg: Springer Berlin Heidelberg. (page 7)

Bhatia, K., Kuang, N. L., Ma, Y.-A. and Wang, Y. (2022) Statistical and computational trade-offs in variational inference: A case study in inferential model selection. *arXiv Preprint arXiv:2207.11208*, arXiv. (pages 1, 17)

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112**, 859–877. (page 3)

Bottou, L. (1999) On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, 9–42. Cambridge University Press, 1 edn. (pages 1, 3)

Buchholz, A., Wenzel, F. and Mandt, S. (2018) Quasi-Monte Carlo variational inference. In *Proceedings of the International Conference on Machine Learning*, vol. 80 of *PMLR*, 668–677. JMLR. (pages 1, 17)

Challis, E. and Barber, D. (2013) Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, **14**, 2239–2286. (pages 8, 17)

Chérief-Abdellatif, B.-E., Alquier, P. and Khan, M. E. (2019) A generalization bound for online variational inference. In *Proceedings of the Asian Conference on Machine Learning*, vol. 101 of *PMLR*, 662–677. JMLR. (page 1)

Csiba, D. and Richtárik, P. (2018) Importance sampling for minibatches. *Journal of Machine Learning Research*, **19**, 1–21. (page 37)

Dieuleveut, A., Durmus, A. and Bach, F. (2020) Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, **48**, 1348–1382. (page 1)

Domke, J. (2019) Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, vol. 32, 329–338. Curran Associates, Inc. (pages 1, 3, 6, 7, 9, 17, 19, 34)

— (2020) Provable smoothness guarantees for black-box variational inference. In *Proceedings of the International Conference on Machine Learning*, vol. 119 of *PMLR*, 2587–2596. JMLR. (pages 1, 3, 4, 17, 34, 43, 44, 45, 46)

Domke, J., Garrigos, G. and Gower, R. (2023a) Provable convergence guarantees for black-box variational inference. *arXiv Preprint arXiv:2306.03638v1*, arXiv. (pages 1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 34, 37, 38)

— (2023b) Provable convergence guarantees for black-box variational inference. *arXiv Preprint arXiv:2306.03638v2*, arXiv. (pages 2, 17)

Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2019) Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, **20**, 1–42. (page 6)

Fan, K., Wang, Z., Beck, J., Kwok, J. and Heller, K. A. (2015) Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, vol. 28, 1387–1395. Curran Associates, Inc. (page 17)

Fujisawa, M. and Sato, I. (2021) Multilevel Monte Carlo variational inference. *Journal of Machine Learning Research*, **22**, 1–44. (pages 1, 3, 9, 17)

Garrigos, G. and Gower, R. M. (2023) Handbook of convergence theorems for (stochastic) gradient methods. *arXiv Preprint arXiv:2301.11235*, arXiv. (pages 1, 37)

Geffner, T. and Domke, J. (2018) Using large ensembles of control variates for variational inference. In *Advances in Neural Information Processing Systems*, vol. 31, 9960–9970. Curran Associates, Inc. (page 2)

— (2020a) Approximation Based Variance Reduction for Reparameterization Gradients. In *Advances in Neural Information Processing Systems*, vol. 33, 2397–2407. Curran Associates, Inc. (pages 2, 9)

— (2020b) A rule for gradient estimator selection, with an application to variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 108 of *PMLR*, 1803–1812. JMLR. (page 2)

— (2021) On the difficulty of unbiased alpha divergence minimization. In *Proceedings of the International Conference on Machine Learning*, vol. 139 of *PMLR*, 3650–3659. JMLR. (page 9)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. (2019) SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, vol. 97 of *PMLR*, 5200–5209. JMLR. (pages 37, 38)

Gower, R. M., Schmidt, M., Bach, F. and Richtarik, P. (2020) Variance-reduced methods for machine learning. *Proceedings of the IEEE*, **108**, 1968–1983. (pages 1, 5)

Gross, L. (1975) Logarithmic Sobolev inequalities. *American Journal of Mathematics*, **97**, 1061–1083. (page 7)

Harvey, N. J. A., Liaw, C., Plan, Y. and Randhawa, S. (2019) Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Conference on Learning Theory*, vol. 99 of *PMLR*, 1579–1613. JMLR. (page 1)

Hoffman, M. and Ma, Y. (2020) Black-box variational inference as a parametric approximation to Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, vol. 119 of *PMLR*, 4324–4341. JMLR. (pages 1, 2, 17)

Huggins, J. H., Campbell, T., Kasprzak, M. and Broderick, T. (2018) Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *arXiv Preprint arXiv:1809.09505*. (pages 3, 9)

Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709. (page 3)

Johnson, R. and Zhang, T. (2013) Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, vol. 26, 315–323. Curran Associates, Inc. (pages 1, 5)

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Machine Learning*, **37**, 183–233. (page 3)

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M. and Sugiyama, M. (2016) Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, UAI'16, 319–328. Arlington, Virginia, USA: AUAI Press. (pages 1, 17)

Khan, M. E. E., Baque, P., Fleuret, F. and Fua, P. (2015) Kullback-Leibler proximal variational inference. In *Advances in Neural Information Processing Systems*, vol. 28, 3402–3410. Curran Associates, Inc. (pages 1, 17)

Kim, K., Oh, J., Wu, K., Ma, Y. and Gardner, J. R. (2023a) On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems (to Appear)*, vol. 36. New Orleans, LA, USA: Curran Associates Inc. (pages 1, 2, 3, 4, 8, 17, 19, 38)

Kim, K., Wu, K., Oh, J. and Gardner, J. R. (2023b) Practical and matching gradient variance bounds for black-box variational Bayesian inference. In *Proceedings of the International Conference on Machine Learning*, vol. 202 of *PMLR*, 16853–16876. Honolulu, HI, USA: JMLR. (pages 1, 3, 5, 7, 8, 17, 19, 36)

Kingma, D. P. and Welling, M. (2014) Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*. Banff, AB, Canada. (page 3)

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. and Blei, D. M. (2017) Automatic differentiation variational inference. *Journal of Machine Learning Research*, **18**, 1–45. (pages 1, 2, 4, 5)

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86. (page 3)

Liu, S. and Owen, A. B. (2021) Quasi-Monte Carlo quasi-Newton in Variational Bayes. *Journal of Machine Learning Research*, **22**, 1–23. (pages 1, 17)

Ma, S., Bassily, R. and Belkin, M. (2018) The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the International Conference on Machine Learning*, vol. 80 of *PMLR*, 3325–3334. JMLR. (pages 1, 5)

Miller, A., Foti, N., D' Amour, A. and Adams, R. P. (2017) Reducing reparameterization gradient variance. In *Advances in Neural Information Processing*

Systems, vol. 30, 3708–3718. Curran Associates, Inc. (pages 2, 9)

Mohamed, S., Rosca, M., Figurnov, M. and Mnih, A. (2020) Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, **21**, 1–62. (pages 3, 17)

Moulines, E. and Bach, F. (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, vol. 24, 451–459. Curran Associates, Inc. (pages 1, 5)

Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, **19**, 1574–1609. (pages 1, 3)

Nesterov, Y. (2004) *Introductory Lectures on Convex Optimization*, vol. 87 of *Applied Optimization*. Boston, MA: Springer US. (page 1)

Otto, F. and Villani, C. (2000) Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, **173**, 361–400. (pages 3, 7)

Paisley, J., Blei, D. M. and Jordan, M. I. (2012) Variational bayesian inference with stochastic search. In *Proceedings of the International Conference on Machine Learning*, ICML'12, 1363–1370. Madison, WI, USA: Omnipress. (page 2)

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F. and Teh, Y. W. (2018) Tighter variational bounds are not necessarily better. In *Proceedings of the International Conference on Machine Learning*, vol. 80 of *PMLR*, 4277–4285. JMLR. (page 9)

Rakhlin, A., Shamir, O. and Sridharan, K. (2012) Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the International Conference on Machine Learning*, ICML'12, 1571–1578. Madison, WI, USA: Omnipress. (page 1)

Ranganath, R., Gerrish, S. and Blei, D. (2014) Black box variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 33 of *PMLR*, 814–822. JMLR. (pages 1, 2, 3, 17)

Regier, J., Jordan, M. I. and McAuliffe, J. (2017) Fast black-box variational inference through stochastic trust-region optimization. In *Advances in Neural Information Processing Systems*, vol. 30, 2399–2408. Curran Associates, Inc. (pages 1, 17)

Rezende, D. J., Mohamed, S. and Wierstra, D. (2014) Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, vol. 32 of *PMLR*, 1278–1286. JMLR. (page 3)

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

Roeder, G., Wu, Y. and Duvenaud, D. K. (2017) Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, vol. 30, 6928–6937. Curran Associates, Inc. (pages 2, 4, 5, 9)

Rudner, T. G. J., Key, O., Gal, Y. and Rainforth, T. (2021) On Signal-to-Noise Ratio Issues in Variational Inference for Deep Gaussian Processes. In *Proceedings of the International Conference on Machine Learning*, PMLR, 9148–9156. JMLR. (page 9)

Schmidt, M., Le Roux, N. and Bach, F. (2017) Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, **162**, 83–112. (pages 1, 5)

Schmidt, M. and Roux, N. L. (2013) Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv Preprint arXiv:1308.6370*, arXiv. (pages 1, 5, 9)

Shalev-Shwartz, S., Singer, Y., Srebro, N. and Cotter, A. (2011) Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, **127**, 3–30. (page 1)

Solodov, M. (1998) Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, **11**, 23–35. (page 9)

Titsias, M. and Lázaro-Gredilla, M. (2014) Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the International Conference on Machine Learning*, vol. 32 of *PMLR*, 1971–1979. JMLR. (pages 1, 2, 3, 4, 17)

Vaswani, S., Bach, F. and Schmidt, M. (2019) 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*, vol. 89 of *PMLR*, 1195–1204. JMLR. (pages 1, 5, 9)

Villani, C. (2016) *Topics in Optimal Transportation*. No. 58 in Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society. (page 7)

Wang, X., Geffner, T. and Domke, J. (2024) Joint control variate for faster black-box variational inference. In *Proceedings of The International Conference on Artificial Intelligence and Statistics (to Appear)*, PMLR. JMLR. (pages 2, 9)

Wright, S. J. and Recht, B. (2021) *Optimization for Data Analysis*. New York: Cambridge University Press. (page 5)

Xu, M., Quiroz, M., Kohn, R. and Sisson, S. A. (2019) Variance reduction properties of the reparameterization trick. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 89 of *PMLR*, 2711–2720. JMLR. (pages 3, 17)

Xu, Z. and Campbell, T. (2022) The computational asymptotics of Gaussian variational inference and the Laplace approximation. *Statistics and Computing*, **32**. (page 17)

Zegers, P. (2015) Fisher information properties. *Entropy*, **17**, 4918–4939. (page 3)

Zhang, C., Butepage, J., Kjellstrom, H. and Mandt, S. (2019) Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 2008–2026. (page 3)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes. See § 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Yes. See § 3.4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Not Applicable.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Yes. See the theorem statements and § 2
   (b) Complete proofs of all theoretical results.
   Yes. See Appendix C.
   (c) Clear explanations of any assumptions.
   Yes. See Appendix C.1 and § 2 and the main text.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Not Applicable.
   (b) The license information of the assets, if applicable.
   Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Not Applicable.
   (d) Information about consent from data providers/curators.
   Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Not Applicable.

Table of Contents

A OVERVIEW OF THEOREMS

Table 1: Overview of Results

DESCRIPTION	RESULT	SECTION
Gradient Variance Bounds		
Smoothness of the entropy with triangular scale parameterization	Proposition 2	C.3
Gradient Variance Bounds		
Upper bound for the gradient variance of the **STL** estimator with the **full-rank** parameterization	Theorem 1	C.4.2
Upper bound for the gradient variance of the **STL** estimator with the **mean-field** parameterization	Theorem 7	C.4.3
Lower bound for the gradient variance of the **STL** estimator with the **full-rank** parameterization	Theorem 2	C.5.1
Worst case lower bound (unimprovability) for the gradient variance of the **STL** estimator with the **full-rank** parameterization	Theorem 3	C.5.2
Upper bound for the gradient variance of the **CFE** estimator with the **full-rank** parameterization	Theorem 4	C.6.1
Upper bound for the gradient variance of the **CFE** estimator with the **mean-field** parameterization	Theorem 8	C.6.2
Complexity of Projected SGD		
Iteration complexity of projected SGD with a **fixed stepsize** and a gradient estimator satisfying the **QV** condition on a **strongly convex** objective function	Theorem 9	C.7.1
Iteration complexity of projected SGD with a **decreasing stepsize** schedule and a gradient estimator satisfying the **QV** condition on a **strongly convex** objective function	Theorem 10	C.7.1
Iteration complexity of projected SGD with a **fixed stepsize** and a gradient estimator satisfying the **adaptive QV** condition on a **strongly convex** objective function	Lemma 9	C.7.2
Iteration complexity of projected SGD with a **decreasing stepsize** schedule and a gradient estimator satisfying the **adaptive QV** condition on a **strongly convex** objective function	Lemma 10	C.7.2
Complexity of BBVI		
Iteration complexity of BBVI with projected SGD using a **fixed stepsize** and the **CFE** gradient estimator on a **strongly log-concave** posterior	Theorem 5	C.8.1
Iteration complexity of BBVI with projected SGD using a **decreasing stepsize** schedule and the **CFE** gradient estimator on a **strongly log-concave** posterior	Theorem 11	C.8.1
Iteration complexity of BBVI with projected SGD using a **fixed stepsize** and the **STL** gradient estimator on a **strongly log-concave** posterior	Theorem 6	C.8.2
Iteration complexity of BBVI with projected SGD using a **decreasing stepsize** schedule and the **STL** gradient estimator on a **strongly log-concave** posterior	Theorem 12	C.8.2

Table 2: Overview of Complexity Analyses of BBVI

Regularity of π					$q_{\lambda^*} = \pi$	Optimized Parameters	Gradient Estimator[1]	Iteration Complexity	Reference
μ-PL	LC	μ-SLC	L-LS	LQ					
✓	✓	✓	✓	✔	✓	scale only	exact	$\mathcal{O}\left(\log\left(L\epsilon^{-1}\right)\right)$	Hoffman and Ma, 2020
✓	✓	✓	✓	✔	✓	scale only	CFE	$\mathcal{O}\left(\kappa^2\epsilon^{-1}\right)$	Hoffman and Ma, 2020
✓	✓	✓	✓	✔		scale only	n/a[2]	$\mathcal{O}\left(L\epsilon^{-1}\right)^3$	Bhatia et al., 2022
✓	✓	✔	✔			scale only	n/a[2]	$\mathcal{O}\left(L\epsilon^{-1}\right)^3$	Bhatia et al., 2022
✔			✔			loc. & scale	CFE	$\mathcal{O}\left(L^2\kappa\epsilon^{-4}\right)$	Kim et al., 2023a
✓	✓	✔	✔			loc. & scale	CFE	$\mathcal{O}\left(\kappa^2\epsilon^{-1}\right)$	Kim et al., 2023a Domke et al., 2023a
	✔		✔			loc. & scale	CFE, STL	$\mathcal{O}\left(L^2\epsilon^{-2}\right)$	Domke et al., 2023a
✓	✓	✔	✔			loc. & scale	STL	$\mathcal{O}\left(\kappa^2\epsilon^{-1}\right)$	Domke et al., 2023b
✓	✓	✔	✔		✔	loc. & scale	STL	$\mathcal{O}\left(\kappa^2\log\epsilon^{-1}\right)$	Domke et al., 2023b
✓	✓	✔	✔			loc. & scale	STL	$\mathcal{O}\left(\kappa^2\epsilon^{-1}\right)$	Theorem 12
✓	✓	✔	✔		✔	loc. & scale	STL	$\mathcal{O}\left(\kappa^2\log\epsilon^{-1}\right)$	Theorem 6

[*] PL: Polyak-Łojasiewicz, LC: log-concave, SLC: strongly-log-concave, LQ: log-quadratic (π is Gaussian), $\kappa = L/\mu$, and $q_{\lambda_*} = \pi$ implies that "the variational family is perfectly specified" such that $\pi \in \mathcal{Q}$.
[*] Conditions implied by other stronger conditions (marked with ✔) are marked with ✓.
[*] Analyses that a-priori assumed regularity of the ELBO were omitted.
[*] The explicit dimension dependences are omitted, but in general, $\mathcal{O}(d)$ for full-rank, which is tight (Domke, 2019), and the best known for mean-field is $o(\sqrt{d})$ (Kim et al., 2023b). The algorithm of Bhatia et al. (2022) is able to trade the dimension dependence for statistical accuracy.
[1] The precise definitions of the gradient estimators are in § 2.4.
[2] This algorithm uses stochastic power method-like iterations.
[3] The per-iteration sample complexity also depends on L, d, ϵ.

B RELATED WORKS

Analyzing the Computational Properties of BBVI Since its inception by Ranganath et al. (2014); Titsias and Lázaro-Gredilla (2014), theoretical results on BBVI have been developing on two different axes: **(a)** Analyzing the regularity of the ELBO such as convexity and smoothness (Challis and Barber, 2013; Titsias and Lázaro-Gredilla, 2014), **(b)** and analyzing the variance of the Monte Carlo gradient estimators (Buchholz et al., 2018; Fan et al., 2015; Mohamed et al., 2020; Xu et al., 2019). While some convergence analyses of BBVI have been provided (Alquier, 2021; Alquier and Ridgway, 2020; Buchholz et al., 2018; Fujisawa and Sato, 2021; Khan et al., 2016, 2015; Liu and Owen, 2021; Regier et al., 2017), these works *a priori* assumed the regularity of the ELBO and the gradient estimators. Due to the difficulty of rigorously establishing these conditions, later works by Bhatia et al. (2022); Hoffman and Ma (2020) have worked with simplified or alternative implementations of BBVI. Meanwhile, Xu and Campbell (2022) showed these regularities can be realized asymptotically in high probability. In expectation, however, it was only recently that regularity conditions on the ELBO (Domke, 2020; Kim et al., 2023a) and the reparameterization gradient estimator (Domke, 2019; Kim et al., 2023b) were shown to be realizable under mild conditions without modifying the algorithms used in practice.

Concurrent Results by Domke et al. (2023b) While our work builds on top of the QV-based framework of Domke et al. (2023a), a similar convergence result on the STL estimator appeared in its later version (Domke et al., 2023b) concurrently with our work. However, our results differ in several aspects:

1. We bound the family-misspecification term T_3 with the Fisher-Hyvärinen divergence $D_{\mathrm{F}^4}(q_{\lambda^*}, \pi)$, while Domke et al. (2023b) bound it with a quadratic involving the smoothness constant of the residual function $r(z) \triangleq \log q_{\lambda^*}(z) - \log \pi(z)$.
2. For the Gaussian posterior case, the constants of Theorem 1 are tighter that of Domke et al. (2023b).
3. We also provide an upper bound for the mean-field variational family in Theorem 7.
4. We establish a lower bound on the gradient variance of STL, quantifying the tightness of the bounds.

C PROOFS

C.1 Definitions

Definition (L**-Smoothness**)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth if it satisfies

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and some $0 < L < \infty$.

Definition (μ**-Strong Convexity**)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex if it satisfies

$$\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and some $0 < \mu < \infty$.

Remark 15. We say a function f is only convex if it satisfies the strong convexity inequality with $\mu = 0$.

C.2 Auxiliary Lemmas

Lemma 2 (Domke 2019, Lemma 9). *Let $\boldsymbol{u} = (u_1, u_2, \dots, u_d)$ be a d-dimensional vector-valued random variable with zero-mean independently and identically distributed components. Then,*

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

Lemma 3. *Let $\mathcal{T}_{\boldsymbol{\lambda}} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^d$ be the location-scale reparameterization function (Definition 3). Then, for any differentiable function f, we have*

$$\left\|\nabla_{\boldsymbol{\lambda}} f\left(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)\right)\right\|_2^2 = J_{\mathcal{T}}\left(\boldsymbol{u}\right)\left\|\nabla f\left(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)\right)\right\|_2^2.$$

for any $\boldsymbol{\lambda} \in \mathbb{R}^p$ and $\boldsymbol{u} \in \mathbb{R}^d$, where $J_{\mathcal{T}}\left(\boldsymbol{u}\right) : \mathbb{R}^d \to \mathbb{R}$ is a function defined as

$$\begin{aligned}
J_{\mathcal{T}}\left(\boldsymbol{u}\right) &= 1 + \textstyle\sum_{i=1}^d u_i^2 && \text{for the full-rank and} \\
J_{\mathcal{T}}\left(\boldsymbol{u}\right) &= 1 + \sqrt{\textstyle\sum_{i=1}^d u_i^4} && \text{for the mean-field parameterizations.}
\end{aligned}$$

Proof. The result is a collection of the results of Domke (2019, Lemma 1) for the full-rank parameterization and Kim et al. (2023b, Lemma 2) for the mean-field parameterization. □

Lemma 4 (Corollary 2; Kim et al., 2023a). *Assume Assumption 1 and let $\mathcal{T}_{\boldsymbol{\lambda}} : \mathbb{R}^d \to \mathbb{R}^d$ (Definition 3) be the location-scale reparameterization function. Then, for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{R}^p$,*

$$\mathbb{E}J_{\mathcal{T}}\left(\boldsymbol{u}\right)\left\|\mathcal{T}_{\boldsymbol{\lambda}'}\left(\boldsymbol{u}\right) - \mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)\right\|_2^2 \leq C\left(d, \varphi\right)\left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\right\|_2^2,$$

where $C\left(d, \varphi\right) = d + k_\varphi$ for the full-rank and $C\left(d, \varphi\right) = 2k_\varphi\sqrt{d} + 1$ for the mean-field parameterizations.

Lemma 5 (Lemma 2; Domke, 2019). *Assume Assumption 1 and let $\mathcal{T}_{\boldsymbol{\lambda}} : \mathbb{R}^d \to \mathbb{R}^d$ (Definition 3) be the location-scale reparameterization function. Then, for the full-rank parameterization,*

$$\mathbb{E}J_{\mathcal{T}}\left(\boldsymbol{u}\right)\left\|\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right) - \bar{\boldsymbol{z}}\right\|_2^2 = C_1\left(d, \varphi\right)\left\|\boldsymbol{m} - \bar{\boldsymbol{z}}\right\|_2^2 + C_2\left(d, \varphi\right)\left\|\boldsymbol{C}\right\|_{\mathrm{F}}^2.$$

where

$$\begin{aligned}
C_1\left(d, \varphi\right) &= d + 1, & C_2\left(d, \varphi\right) &= d + k_\varphi, && \text{for the full-rank and} \\
C_1\left(d, \varphi\right) &= \sqrt{dk_\varphi} + k\sqrt{d} + 1, & C_2\left(d, \varphi\right) &= 2\kappa\sqrt{d} + 1, && \text{for the mean-field parameterizations.}
\end{aligned}$$

Proof. The result is a collection of the results of Domke (2019, Lemma 2) for the full-rank parameterization and Kim et al. (2023b, Lemma 2) for the mean-field parameterization. □

Lemma 6. *For any $a, b, c, \in \mathbb{R}$,*

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

for any $\delta > 0$.

Proof. The Peter-Paul generalization of Young's inequality states that, for $d, e \geq 0$, we have

$$de \leq \frac{\delta}{2}d^2 + \frac{\delta^{-1}}{2}e^2.$$

Applying this,

$$\begin{aligned}(a+b+c)^2 &= a^2+b^2+c^2+2ab+2ac+2bc\\ &\le a^2+b^2+c^2+2|a||b|+2|a||c|+2|b||c|\\ &\le a^2+b^2+c^2+2\left(\frac{1}{2}a^2+\frac{1}{2}b^2\right)+2\left(\frac{\delta}{2}a^2+\frac{\delta^{-1}}{2}c^2\right)+2\left(\frac{\delta}{2}b^2+\frac{\delta^{-1}}{2}c^2\right)\\ &= a^2+b^2+c^2+\left(a^2+b^2\right)+\left(\delta a^2+\delta^{-1}c^2\right)+\left(\delta b^2+\delta^{-1}c^2\right)\\ &= (2+\delta)\,a^2+(2+\delta)\,b^2+\left(1+2\delta^{-1}\right)c^2.\end{aligned}$$

□

Lemma 7. *Let $\mathcal{T}_\lambda : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^d$ be the location-scale reparameterization function (Definition 3) and $\boldsymbol{u} \sim \varphi$ satisfy Assumption 1. Then,*

$$\mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\left(\mathcal{T}_\lambda(\boldsymbol{u})+\boldsymbol{z}\right) = (d+1)\left(\boldsymbol{m}+\boldsymbol{z}\right)$$

for any $\boldsymbol{z} \in \mathbb{R}^d$.

Proof.

$$\begin{aligned}\mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\left(\mathcal{T}_\lambda(\boldsymbol{u})+\boldsymbol{z}\right) &= \mathbb{E}\left(1+\|\boldsymbol{u}\|_2^2\right)\left(\boldsymbol{C}\boldsymbol{u}+\boldsymbol{m}+\boldsymbol{z}\right)\\ &= \boldsymbol{C}\mathbb{E}\left(1+\|\boldsymbol{u}\|_2^2\right)\boldsymbol{u}+\mathbb{E}\left(1+\|\boldsymbol{u}\|_2^2\right)\left(\boldsymbol{m}+\boldsymbol{z}\right)\\ &= (d+1)\left(\boldsymbol{m}+\boldsymbol{z}\right),\end{aligned}$$

where the last equality follows from Lemma 2 and Assumption 1. □

Lemma 8. *Let $\boldsymbol{A} = \operatorname{diag}(A_1,\dots,A_d) \in \mathbb{R}^{d\times d}$ be some diagonal matrix, define*

$$\boldsymbol{B} = \begin{bmatrix} L^{-1/2} & & & \\ & L^{1/2} & & \\ & & \ddots & \\ & & & L^{1/2}\end{bmatrix}, \qquad \boldsymbol{C} = L^{-1/2}\,\mathbf{I},$$

some $\boldsymbol{u} \in \mathbb{R}^d$, $\boldsymbol{m} \in \mathbb{R}^d$, and $\boldsymbol{z} \in \mathbb{R}^d$ such that $m_1 = z_1$. For $\boldsymbol{\lambda} = (\boldsymbol{m}, \boldsymbol{C})$, the expression

$$\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{A}\boldsymbol{u}+\boldsymbol{m}-\boldsymbol{z}\right)\right\|_2^2.$$

can be bounded for the following instances of $\boldsymbol{A}$:

(i) *If $\boldsymbol{A} = \boldsymbol{C}$,*

$$\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{C}\boldsymbol{u}+\boldsymbol{m}-\boldsymbol{z}\right)\right\|_2^2 = \left\|\boldsymbol{C}\boldsymbol{u}+\boldsymbol{m}-\boldsymbol{z}\right\|_2^2 + \left(L-L^{-1}\right)u_1^2,$$

(ii) *while if $\boldsymbol{A} = \mathbf{O}$,*

$$\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{m}-\boldsymbol{z}\right)\right\|_2^2 = \left\|\boldsymbol{m}-\boldsymbol{z}\right\|_2^2.$$

Proof. First notice that

$$\boldsymbol{B}^{-1}\boldsymbol{C}^{-1} = \begin{bmatrix} L & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}.$$

Denoting the 1st coordinate of $\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m}$ as $[\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m}]_1 = A_1 u_1 + m_1$, we have

$$\begin{aligned}
&\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right)\right\|_2^2 \\
&= \left\|\begin{bmatrix} L & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}\left(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right)\right\|_2^2 \\
&= \left\|\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right\|_2^2 + \left(L^2 - 1\right)\left([\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m}]_1 - z_1\right)^2 \\
&= \left\|\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right\|_2^2 + \left(L^2 - 1\right)\left(A_1 u_1 + m_1 - z_1\right)^2,
\end{aligned} \tag{2}$$

and using the fact that $m_1 = z_1$

$$= \left\|\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right\|_2^2 + \left(L^2 - 1\right) A_1^2 u_1^2. \tag{3}$$

Proof of (i) If $\boldsymbol{A} = \boldsymbol{C} = L^{-1/2}\mathbf{I}$, Eq. (3) yields,

$$\begin{aligned}
\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right)\right\|_2^2 &= \left\|\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right\|_2^2 + \left(L^2 - 1\right) L^{-1} u_1^2 \\
&= \left\|\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right\|_2^2 + \left(L - L^{-1}\right) u_1^2
\end{aligned}$$

Proof of (ii) If $\boldsymbol{A} = \mathbf{O}$, Eq. (3) yields,

$$\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{A}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{z}\right)\right\|_2^2 = \left\|\boldsymbol{m} - \boldsymbol{z}\right\|_2^2.$$

□

C.3 Smoothness Under Triangular Scale Parameterization

Proposition 2. *The entropy $\boldsymbol{\lambda} \mapsto \mathbb{H}(q_{\boldsymbol{\lambda}})$ is S-Lipschitz smooth on Λ_S.*

Proof. From the definition of the entropy of location-scale variational families, we have

$$\|\nabla_{\boldsymbol{\lambda}} \mathbb{H}(q_{\boldsymbol{\lambda}}) - \nabla_{\boldsymbol{\lambda}'} \mathbb{H}(q_{\boldsymbol{\lambda}'})\|_2^2 = \|\nabla_{\boldsymbol{C}} \log\det \boldsymbol{C} - \nabla_{\boldsymbol{C}'} \log\det \boldsymbol{C}'\|_2^2,$$

since $\boldsymbol{C}, \boldsymbol{C}' \in \mathbb{L}_{++}^d$,

$$= \left\|\nabla_{\boldsymbol{C}} \log\det(\operatorname{diag}(\boldsymbol{C})) - \nabla_{\boldsymbol{C}'} \log\det\left(\operatorname{diag}\left(\boldsymbol{C}'\right)\right)\right\|_2^2,$$

since the determinant of triangular matrices is the product of the diagonal,

$$\begin{aligned}
&= \sum_{i=1}^d \left| \frac{\partial \log C_{ii}}{\partial C_{ii}} - \frac{\partial \log C'_{ii}}{\partial C'_{ii}} \right|^2 \\
&= \sum_{i=1}^d \left| \frac{1}{C_{ii}} - \frac{1}{C'_{ii}} \right|^2 \\
&= \sum_{i=1}^d C_{ii}^{-2} \left| C'_{ii} - C_{ii} \right|^2 \left(C'_{ii}\right)^{-2},
\end{aligned}$$

and since $\sigma_{\min}(\boldsymbol{C}) \geq S^{-1/2} \Leftrightarrow C_{ii}^{-2} \leq S$ for all $i = 1, \ldots, d$,

$$\begin{aligned}
&\leq S^2 \sum_{i=1}^d \left| C'_{ii} - C_{ii} \right|^2 \\
&= S^2 \left\| \operatorname{diag}(\boldsymbol{C}) - \operatorname{diag}\left(\boldsymbol{C}'\right) \right\|_2^2 \\
&\leq S^2 \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2.
\end{aligned}$$

□

C.4 Upper Bound on Gradient Variance of STL

C.4.1 General Decomposition

Lemma 1. *Assume Assumption 2. The expected-squared norm of STL is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \leq (2+\delta)V_1 + (2+\delta)V_2 + \left(1+2\delta^{-1}\right)V_3,$$

where the terms are

$$V_1 = \mathbb{E}\, J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2$$
$$V_2 = \mathbb{E}\, J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2$$
$$V_3 = \mathbb{E}\, J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2,$$

for any $\delta > 0$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$. $J_{\mathcal{T}} : \mathbb{R}^d \to \mathbb{R}$ is a function depending on the variational family as

$$J_{\mathcal{T}}(\boldsymbol{u}) = 1 + \sum_{i=1}^d u_i^2 \quad \textit{for full-rank and}$$
$$J_{\mathcal{T}}(\boldsymbol{u}) = 1 + \sqrt{\sum_{i=1}^d u_i^4} \quad \textit{for mean-field.}$$

Proof. From the definition of the STL estimator Definition 5,

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 = \mathbb{E}\|\nabla_{\boldsymbol{\lambda}} \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\nu}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}},$$

by Lemma 3,

$$= \mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\nu}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}}$$

adding the terms $\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))$ and $\nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))$ that cancel,

$$\begin{aligned} = \mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|&\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) \\ &+ \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) \\ &+ \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2, \end{aligned}$$

applying Lemma 6,

$$\begin{aligned} \leq \mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \Big(&(2+\delta) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2 \\ &+ (2+\delta) \|\nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \\ &+ \left(1+2\delta^{-1}\right) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2 \Big), \end{aligned}$$

and distributing $J_{\mathcal{T}}$ and the expectation,

$$\begin{aligned} = (2+\delta) &\underbrace{\mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2}_{V_1} \\ + (2+\delta) &\underbrace{J_{\mathcal{T}}(\boldsymbol{u}) \mathbb{E} \|\nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2}_{V_2} \\ + \left(1+2\delta^{-1}\right) &\underbrace{\mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2}_{V_3}. \end{aligned}$$

□

C.4.2 Full-Rank Parameterization

Theorem 1. *Assume Assumption 2 and that ℓ is L-log-smooth. For the full-rank parameterization, the expected-squared norm of the STL estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \leq \alpha_{\mathrm{STL}}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 + \beta_{\mathrm{STL}}$$

where

$$\alpha_{\mathrm{STL}} = (2+\delta)\left(L^2\left(d+k_\varphi\right) + S^2(d+1)\right)$$
$$\beta_{\mathrm{STL}} = (1+2\delta^{-1})\left(2d+k_\varphi\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*},\ell)}$$

for any $\boldsymbol{\lambda},\boldsymbol{\lambda}^* \in \Lambda_S$ and any $\delta > 0$.

Proof. We analyze each of the terms in Lemma 1.

Bound on V_1 For V_1, we obtain the quadratic bound from the optimum as

$$\begin{aligned}
V_1 &= \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\left\|\nabla\log\ell\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right) - \nabla\log\ell\left(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right)\right\|_2^2 \\
&\leq L^2\,\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\left\|\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right\|_2^2 && (L\text{-log-smoothness}) && (4)\\
&\leq L^2\left(d+k_\varphi\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2. && (\text{Lemma 4}) && (5)
\end{aligned}$$

Bound on V_2 Now for,

$$V_2 = \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\left\|\nabla\log q_{\boldsymbol{\lambda}^*}\left(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right) - \nabla\log q_{\boldsymbol{\lambda}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\right\|_2^2,$$

we use the fact that, for location-scale family distributions, the log-probability density is

$$\log q_{\boldsymbol{\lambda}}(\boldsymbol{z}) = \log\varphi\left(\boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{m})\right) - \log|\boldsymbol{C}|.$$

Considering reparameterization,

$$\begin{aligned}
\log q_{\boldsymbol{\lambda}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right) &= \log\varphi\left(\boldsymbol{C}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})-\boldsymbol{m}\right)\right) - \log|\boldsymbol{C}| \\
&= \log\varphi\left(\boldsymbol{C}^{-1}\left((\boldsymbol{C}\boldsymbol{u}+\boldsymbol{m})-\boldsymbol{m}\right)\right) - \log|\boldsymbol{C}| \\
&= \log\varphi(\boldsymbol{u}) - \log|\boldsymbol{C}|.
\end{aligned}$$

This implies

$$\begin{aligned}
\nabla\log q_{\boldsymbol{\lambda}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right) &= \nabla_{\boldsymbol{\lambda}}\log\phi(\boldsymbol{u}) - \nabla\log|\boldsymbol{C}| \\
&= -\nabla\log|\boldsymbol{C}| \\
&= \nabla_{\boldsymbol{\lambda}}\mathbb{H}(q_{\boldsymbol{\lambda}}).
\end{aligned}$$

Thus,

$$\begin{aligned}
V_2 &= \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\left\|\nabla\log q_{\boldsymbol{\lambda}^*}\left(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right) - \nabla\log q_{\boldsymbol{\lambda}}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})\right)\right\|_2^2 \\
&\leq S^2\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 && (\text{Proposition 2}) && (6)\\
&= S^2\left(1+\mathbb{E}\sum_{i=1}^{d}u_i^2\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 && (\text{definition of } J_{\mathcal{T}} \text{ in Lemma 3}) \\
&= S^2(1+d)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 && (\text{Assumption 1}) && (7)
\end{aligned}$$

Bound on V_3 Finally, for V_3,

$$V_3 = \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\left\|\nabla\log\ell\left(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right) - \nabla\log q_{\boldsymbol{\lambda}^*}\left(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\right)\right\|_2^2,$$

by the definition of $J_{\mathcal{T}}$ in Lemma 3,

$$= \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right) \|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^2$$
$$= \mathbb{E}\left(1 + \|\boldsymbol{u}\|_2^2\right) \|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^2,$$

through the Cauchy-Schwarz inequality,

$$\leq \sqrt{\mathbb{E}\left(1 + \|\boldsymbol{u}\|_2^2\right)^2} \sqrt{\mathbb{E}\|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^4}$$
$$= \sqrt{\left(1 + 2\mathbb{E}\|\boldsymbol{u}\|_2^2 + \mathbb{E}\|\boldsymbol{u}\|_2^4\right)} \sqrt{\mathbb{E}\|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^4},$$

by Lemma 2 and Assumption 1, the $2\,\mathbb{E}\|\boldsymbol{u}\|_2^2$ term becomes

$$= \sqrt{1 + 2d + \mathbb{E}\|\boldsymbol{u}\|_2^4} \sqrt{\mathbb{E}\|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^4}. \tag{8}$$

Meanwhile, $\mathbb{E}\|\boldsymbol{u}\|_2^4$ follows as

$$\mathbb{E}\|\boldsymbol{u}\|_2^4 = \mathbb{E}\left(\|\boldsymbol{u}\|_2^2\right)^2 = \mathbb{E}\left(\sum_{i=1}^{d} u_i^2\right)^2 = \mathbb{E}\left(\sum_{i=1}^{d} u_i^4 + \sum_{i \neq j} u_i^2 u_j^2\right),$$

while from Assumption 1, u_i and u_j are independent for $i \neq j$. Thus

$$= \sum_{i=1}^{d} \mathbb{E}u_i^4 + \sum_{i \neq j} \mathbb{E}u_i^2\, \mathbb{E}u_j^2,$$

and by Assumption 1, we have $\mathbb{E}u_i^4 = k_\varphi$, $\mathbb{E}u_i^2 = 1$, and $\mathbb{E}u_j^2 = 1$. Therefore,

$$= dk_\varphi + 2\binom{d}{2},$$

and applying the well-known upper bound on the binomial coefficient $\binom{d}{2} \leq \left(\frac{\mathrm{e}d}{2}\right)^2$,

$$\leq dk_\varphi + 2\left(\frac{\mathrm{e}}{2}d\right)^2 = dk_\varphi + \frac{\mathrm{e}^2}{2}d^2,$$

where e is Euler's constant.

Applying this to first term in Eq. (8),

$$\sqrt{1 + 2d + \mathbb{E}\|\boldsymbol{u}\|_2^4} \leq \sqrt{1 + 2d + k_\varphi d + \frac{\mathrm{e}^2}{2}d^2} = \sqrt{\frac{\mathrm{e}^2}{2}d^2 + \left(2 + k_\varphi\right)d + 1},$$

and since $k_\varphi \geq 1$, $d \geq 1$, and $\mathrm{e}^2/2 \leq 4$,

$$\leq \sqrt{4d^2 + \left(2k_\varphi + k_\varphi\right)d + k_\varphi} = \sqrt{4d^2 + 3k_\varphi d + k_\varphi^2}$$
$$\leq \sqrt{4d^2 + 4dk_\varphi + k_\varphi^2}$$
$$= \left(2d + k_\varphi\right) \tag{9}$$

Thus, V_3 can be bounded as

$$V_3 \leq \sqrt{1 + 2d + \mathbb{E}\|\boldsymbol{u}\|_2^4} \sqrt{\mathbb{E}\|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^4},$$

applying Eq. (9),

$$\leq \left(2d + k_\varphi\right) \sqrt{\mathbb{E}\|\nabla \log \ell\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right) - \nabla \log q_{\lambda^*}\left(\mathcal{T}_{\lambda^*}(\boldsymbol{u})\right)\|_2^4}$$

applying Change-of-Variable on the score term,

$$= \left(2d + k_\varphi\right) \sqrt{\mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\lambda}^*}} \|\nabla \log \ell\left(\boldsymbol{z}\right) - \nabla \log q_{\boldsymbol{\lambda}^*}\left(\boldsymbol{z}\right)\|_2^4}$$

$$= \left(2d + k_\varphi\right) \sqrt{\mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\lambda}^*}} \|\nabla \log \pi\left(\boldsymbol{z}\right) - \nabla \log q_{\boldsymbol{\lambda}^*}\left(\boldsymbol{z}\right)\|_2^4},$$

and by the definition of the 4th order Fisher-Hyvärinen divergence,

$$\leq \left(2d + k_\varphi\right) \sqrt{\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}^*}, \ell\right)}. \tag{10}$$

Combining Eqs. (5), (7) and (10) with Lemma 1,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}\left(\boldsymbol{\lambda}\right)\|_2^2 &\leq (2+\delta)V_1 + (2+\delta)V_2 + (1+2\delta^{-1})V_3 \\
&\leq L^2(2+\delta)\left(d + k_\varphi\right)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 + S^2(2+\delta)\left(d+1\right)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 \\
&\quad + (1+2\delta^{-1})\left(2d + k_\varphi\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}^*}, \ell\right)} \\
&= (2+\delta)\left(L^2\left(d + k_\varphi\right) + S^2\left(d+1\right)\right)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|_2^2 \\
&\quad + (1+2\delta^{-1})\left(2d + k_\varphi\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}^*}, \ell\right)}.
\end{aligned}$$

□

C.4.3 Mean-Field Parameterization

Theorem 7. *Assume Assumption 2 and that ℓ is L-log-smooth. For the mean-field parameterization, the expected-squared norm of the STL estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \leq (2+\delta)\left(L^2\left(2k_\varphi\sqrt{d}+1\right)+S^2\left(\sqrt{dk_\varphi}+1\right)\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2+(1+2\delta^{-1})\left(1+\sqrt{dk_\varphi}\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*},\ell)}$$

for any $\boldsymbol{\lambda},\boldsymbol{\lambda}^* \in \Lambda_S$ and any $\delta > 0$.

Proof. Similarly with Theorem 1, we analyze each term in Lemma 1.

Bound on V_1 The process for V_1 is more or less identical to Theorem 1. Starting from Eq. (4),

$$\begin{aligned} V_1 &\leq L^2\,\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\,\|\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})-\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u})\|_2^2, \\ &\leq \left(2k_\varphi\sqrt{d}+1\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2. \qquad \text{(Lemma 4)} \end{aligned} \tag{11}$$

Bound on V_2 This is also identical to Theorem 1 apart from $J_{\mathcal{T}}$. Resuming from Eq. (6),

$$\begin{aligned} V_2 &\leq S^2\mathbb{E}J_{\mathcal{T}}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 \\ &= S^2\left(1+\mathbb{E}\sqrt{\sum_{i=1}^d u_i^4}\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2, \qquad \text{(definition of } J_{\mathcal{T}} \text{ in Lemma 3)} \\ &\leq S^2\left(1+\sqrt{\sum_{i=1}^d \mathbb{E}u_i^4}\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2, \qquad \text{(Jensen's inequality)} \\ &\leq S^2\left(1+\sqrt{dk_\varphi}\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2. \qquad \text{(Assumption 1).} \end{aligned} \tag{12}$$

Bound on V_3 The derivation for V_3 is less technical than the full-rank case. Denoting $\boldsymbol{U} = \mathrm{diag}(u_1,\ldots,u_d)$ for clarity, we have

$$\sqrt{\textstyle\sum_{i=1}^d u_i^4} = \|\boldsymbol{U}^2\|_{\mathrm{F}}. \tag{13}$$

Then,

$$V_3 = \mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\,\|\nabla\log\ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))-\nabla\log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2$$

by the definition of $J_{\mathcal{T}}$ in Lemma 3 and Eq. (13),

$$= \mathbb{E}\left(1+\|\boldsymbol{U}^2\|_{\mathrm{F}}\right)\|\nabla\log\ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))-\nabla\log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^2,$$

through the Cauchy-Schwarz inequality,

$$\leq \underbrace{\sqrt{\mathbb{E}\left(1+2\|\boldsymbol{U}^2\|_{\mathrm{F}}+\|\boldsymbol{U}^2\|_{\mathrm{F}}^2\right)}}_{T_①}\sqrt{\mathbb{E}\|\nabla\log\ell(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))-\nabla\log q_{\boldsymbol{\lambda}^*}(\mathcal{T}_{\boldsymbol{\lambda}^*}(\boldsymbol{u}))\|_2^4}.$$

$T_①$ follows as

$$\begin{aligned} T_① &= \sqrt{\mathbb{E}\left(1+2\|\boldsymbol{U}^2\|_{\mathrm{F}}+\|\boldsymbol{U}^2\|_{\mathrm{F}}^2\right)} \\ &= \sqrt{\mathbb{E}\left(1+2\sqrt{\textstyle\sum_{i=1}^d u_i^4}+\textstyle\sum_{i=1}^d u_i^4\right)}, \end{aligned}$$

distributing the expectation,

$$= \sqrt{1 + 2\mathbb{E}\left(\sqrt{\sum_{i=1}^{d} u_i^4}\right) + \sum_{i=1}^{d} \mathbb{E}u_i^4},$$

applying Jensen's inequality to the middle term,

$$\leq \sqrt{1 + 2\sqrt{\sum_{i=1}^{d} \mathbb{E}u_i^4} + \sum_{i=1}^{d} \mathbb{E}u_i^4},$$

and from Assumption 1,

$$= \sqrt{1 + 2\sqrt{dk_\varphi} + dk_\varphi} = \sqrt{\left(1 + \sqrt{dk_\varphi}\right)^2}$$
$$= 1 + \sqrt{dk_\varphi}. \tag{14}$$

As in the proof of Lemma 1, we obtain the 4th order Fisher-Hyvärinen divergence after Change-of-Variable. Combining this fact with Eqs. (11), (12) and (14) and Lemma 1,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 &\leq (2+\delta)V_1 + (2+\delta)V_2 + (1+2\delta^{-1})V_3 \\
&\leq L^2(2+\delta)\left(2k_\varphi\sqrt{d}+1\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 + S^2(2+\delta)\left(\sqrt{dk_\varphi}+1\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 \\
&\quad + (1+2\delta^{-1})\left(\sqrt{dk_\varphi}+1\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \ell)} \\
&= (2+\delta)\left(L^2\left(2k_\varphi\sqrt{d}+1\right) + S^2\left(\sqrt{dk_\varphi}+1\right)\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 \\
&\quad + (1+2\delta^{-1})\left(\sqrt{dk_\varphi}+1\right)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}\ell)}.
\end{aligned}$$

□

C.5 Lower Bound on Gradient Variance of STL

C.5.1 General Lower Bound

Theorem 2. *Assume Assumption 2. The expected-squared norm of the STL estimator is lower bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 \geq \mathrm{D}_{\mathrm{F}}(q_{\boldsymbol{\lambda}}, \pi) \geq \frac{2}{C_{\mathrm{LSI}}}\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\lambda}}, \pi),$$

for all $\boldsymbol{\lambda} \in \Lambda_S$ and any $0 < S < \infty$, where the last inequality holds if π is LSI.

Proof.

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{STL}}(\boldsymbol{\lambda})\|_2^2 = \mathbb{E}\|\nabla_{\boldsymbol{\lambda}} \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\nu}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}},$$

by Lemma 3,

$$= \mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\nu}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2 \Big|_{\boldsymbol{\nu}=\boldsymbol{\lambda}}$$

$$= \mathbb{E} J_{\mathcal{T}}(\boldsymbol{u}) \|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2,$$

since $J_{\mathcal{T}}(\boldsymbol{u}) \geq 1$ for both the full-rank and mean-field parameterizations,

$$\geq \mathbb{E}\|\nabla \log \ell(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \log q_{\boldsymbol{\lambda}}(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}))\|_2^2,$$

after Change-of-Variable,

$$= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\lambda}}}\|\nabla \log \ell(\boldsymbol{z}) - \nabla \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\|_2^2,$$

and since $\log \pi(\boldsymbol{z}) = \log \ell(\boldsymbol{z}) + \log Z$ for some constant $Z > 0$,

$$= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\lambda}}}\|\nabla \log \pi(\boldsymbol{z}) - \nabla \log q_{\boldsymbol{\lambda}}(\boldsymbol{z})\|_2^2$$

$$= \mathrm{D}_{\mathrm{F}}(q_{\boldsymbol{\lambda}}, \ell).$$

Finally, when the log-Sobolev inequality applies,

$$\geq \frac{2}{C_{\mathrm{LSI}}}\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\lambda}}, \pi).$$

□

C.5.2 Unimprovability

Theorem 3. *Assume Assumption 2. There exists a strongly-convex, L-log-smooth posterior and some variational parameter $\widetilde{\boldsymbol{\lambda}} \in \Lambda_L$ for all $L \geq 1$ such that*

$$\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\widetilde{\boldsymbol{\lambda}}\right)\right\|_2^2 \geq \left(L^2\left(d+k_\varphi\right)-2\left(d+1\right)\right)\left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 - 2\left(k_\varphi-1\right)\left\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\right\|_2^2,$$

where $\widetilde{\boldsymbol{\lambda}} = (\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{C}})$ and $\bar{\boldsymbol{z}}$ is a stationary point of the said log posterior.

Proof. The worst case is achieved by the following:

(i) $\log \ell$ is ill-conditioned such that the smoothness constant is large. This results in the domain Λ_L to include ill-conditioned $\boldsymbol{C}$s, which has the largest impact on the gradient variance. Furthermore,

(ii) π and q_λ need to have the least overlap in probability volume. This means the variance reduction effect will be minimal.

For Gaussians, this is equivalent to minimizing $\left\|\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}}^2$ while maximizing $\left\|\boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{F}}^2$ and $\left\|\boldsymbol{S}^{-1}\right\|_{\mathrm{F}}^2$.

We therefore choose

$$\pi = \mathcal{N}\left(\bar{\boldsymbol{z}}, \boldsymbol{\Sigma}\right) \qquad q_\lambda = \mathcal{N}\left(\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{S}}\right),$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} L^{-1} & & & \\ & L & & \\ & & \ddots & \\ & & & L \end{bmatrix}, \qquad \widetilde{\boldsymbol{S}} = L^{-1}\mathbf{I}, \quad \text{and} \quad \widetilde{\boldsymbol{m}} = \begin{bmatrix} \bar{z}_1 \\ m_2 \\ \vdots \\ m_d \end{bmatrix},$$

where $\bar{z}_1$ is the 1st element of $\bar{\boldsymbol{z}}$ such that $\widetilde{m}_1 = \bar{z}_1$. The choice of $\widetilde{m}_1 = \bar{z}_1$ is purely for clarifying the derivation. Notice that $\boldsymbol{\Sigma}$ has $d-1$ entries set as L, only one entry set as L^{-1}, and $\widetilde{\boldsymbol{S}} = \widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}$. Here, π is L^{-1}-strongly log-concave, L-log smooth, and $\widetilde{\boldsymbol{\lambda}} = \left(\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{C}}\right) \in \Lambda_L$.

General Gaussian π Lower Bound As usual, we start from the definition of the STL estimator as

$$\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\boldsymbol{\lambda}\right)\right\|_2^2 = \mathbb{E}\left\|\nabla_\lambda \log \ell\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right) - \nabla_\lambda \log q_\nu\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right)\right\|_2^2 \Big|_{\nu=\lambda}$$

by Lemma 3,

$$= \mathbb{E} J_{\mathcal{T}}\left(\boldsymbol{u}\right)\left\|\nabla \log \ell\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right) - \nabla \log q_\nu\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right)\right\|_2^2 \Big|_{\nu=\lambda},$$

since both π and q_λ are Gaussians,

$$\begin{aligned}
&= \mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\left\|\nabla \log \ell\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right) - \nabla \log q_\lambda\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)\right)\right\|_2^2 \\
&= \mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right) - \boldsymbol{S}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\boldsymbol{m}\right)\right\|_2^2 \\
&= \mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right) - \boldsymbol{S}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right) + \boldsymbol{S}^{-1}\left(\boldsymbol{m}-\bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&= \mathbb{E}\left(1+\textstyle\sum_{i=1}^d u_i^2\right)\Big(\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\boldsymbol{S}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\boldsymbol{S}^{-1}\left(\boldsymbol{m}-\bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\qquad - 2\left\langle\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right)\right\rangle \\
&\qquad + 2\left\langle\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m}-\bar{\boldsymbol{z}}\right)\right\rangle \\
&\qquad - 2\left\langle\boldsymbol{S}^{-1}\left(\mathcal{T}_\lambda\left(\boldsymbol{u}\right)-\bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m}-\bar{\boldsymbol{z}}\right)\right\rangle\Big),
\end{aligned}$$

distributing the expectation and $1 + \sum_{i=1}^{d} u_i^2$,

$$
\begin{aligned}
&= \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\boldsymbol{S}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2\right) \\
&\quad + \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\|\boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad - 2\,\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\langle\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\rangle \\
&\quad + 2\left\langle\boldsymbol{\Sigma}^{-1}\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\rangle \\
&\quad - 2\left\langle\boldsymbol{S}^{-1}\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\rangle,
\end{aligned}
$$

applying Lemmas 2 and 7 to the second term and the last two inner product terms,

$$
\begin{aligned}
&= \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\boldsymbol{S}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2\right) \\
&\quad + (d+1)\left\|\boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad - 2\,\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\langle\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\rangle \\
&\quad + 2\,(d+1)\left\langle\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\rangle \\
&\quad - 2\,(d+1)\left\langle\boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right), \boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\rangle.
\end{aligned}
$$

The last two inner products can be denoted as norms such that

$$
\begin{aligned}
&= \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\boldsymbol{S}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2\right) \\
&\quad + (d+1)\left\|\boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad - 2\,\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad + 2\,(d+1)\left\|\boldsymbol{B}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\|_2^2 - 2\,(d+1)\left\|\boldsymbol{S}^{-1}\left(\boldsymbol{m} - \bar{\boldsymbol{z}}\right)\right\|_2^2,
\end{aligned}
$$

where $\boldsymbol{B}$ is the matrix square root of $\boldsymbol{\Sigma}$ such that $\boldsymbol{B}^{-1}\boldsymbol{B}^{-1} = \boldsymbol{\Sigma}^{-1}$. The derivation so far applies to any Gaussian $\pi, q_{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda} \in \Lambda_S$ for any $S > 0$.

Worst-Case Lower Bound Now, for our worst-case example,

$$
\begin{aligned}
\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\widetilde{\boldsymbol{\lambda}}\right)\right\|_2^2 &= \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(\left\|\boldsymbol{\Sigma}^{-1}\left(\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 + \left\|\widetilde{\boldsymbol{S}}^{-1}\left(\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2\right) \\
&\quad + (d+1)\left\|\widetilde{\boldsymbol{S}}^{-1}\left(\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad - 2\,\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad + 2\,(d+1)\left\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right)\right\|_2^2 - 2\,(d+1)\left\|\widetilde{\boldsymbol{S}}^{-1}\left(\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right)\right\|_2^2,
\end{aligned}
$$

since π is μ-strongly log-concave and $\widetilde{\boldsymbol{S}}^{-1} = L\mathbf{I}$,

$$
\begin{aligned}
&\geq \mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left(L^{-2}\left\|\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right\|_2^2 + L^2\left\|\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right\|_2^2\right) \\
&\quad + (d+1)\,L^2\left\|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right\|_2^2 \\
&\quad - 2\,\mathbb{E}\left(1 + \sum_{i=1}^{d} u_i^2\right)\left\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\mathcal{T}_{\widetilde{\boldsymbol{\lambda}}}(\boldsymbol{u}) - \bar{\boldsymbol{z}}\right)\right\|_2^2 \\
&\quad + 2\,(d+1)\,\left\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right)\right\|_2^2 - 2\,(d+1)\,L^2\left\|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right\|_2^2,
\end{aligned}
$$

and grouping the terms,

$$= \underbrace{\left(L^{-2}+L^2\right)\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2-(d+1)L^2\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2}_{T_①}$$
$$\underbrace{-2\,\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right)\|_2^2}_{T_②}$$
$$\underbrace{+2\,(d+1)\,\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\right)\|_2^2}_{T_③}.$$

Lower Bound on $T_①$ For $T_①$, we have

$$T_① = \left(L^{-2}+L^2\right)\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2-(d+1)L^2\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2,$$

applying Lemma 5,

$$= \left(L^{-2}+L^2\right)\left((d+1)\,\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2+\left(d+k_\varphi\right)\|\widetilde{\boldsymbol{C}}\|_{\mathrm{F}}^2\right)-(d+1)L^2\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2,$$

and since $L^{-2}>0$ and is negligible for large Ls,

$$\geq L^2\left((d+1)\,\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2+\left(d+k_\varphi\right)\|\widetilde{\boldsymbol{C}}\|_{\mathrm{F}}^2\right)-(d+1)L^2\|\widetilde{\boldsymbol{m}}-\bar{\boldsymbol{z}}\|_2^2$$
$$= L^2\left(d+k_\varphi\right)\|\widetilde{\boldsymbol{C}}\|_{\mathrm{F}}^2. \tag{15}$$

Lower Bound on $T_②$ For $T_②$, we now use the covariance structures of our worst case through Lemma 8. That is,

$$T_② = -2\,\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right)\|_2^2.$$

Noting that $\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})=\widetilde{\boldsymbol{C}}\boldsymbol{u}+\widetilde{\boldsymbol{m}}$ by definition, we can apply Lemma 8 Item (i) as

$$= -2\,\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left(\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2+\left(L-L^{-1}\right)u_1^2\right),$$

distributing the expectation and $1+\sum_{i=1}^{d}u_i^2$,

$$= -2\left(\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2+\underbrace{\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left(L-L^{-1}\right)u_1^2}_{T_④}\right),$$

$T_④$ follows as

$$T_④ = \mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left(L-L^{-1}\right)u_1^2$$
$$= \left(L^1-L^{-1}\right)\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)u_1^2$$
$$= \left(L-L^{-1}\right)\left(\mathbb{E}u_1^2+\mathbb{E}u_1^4+\sum_{i=2}^{d}\mathbb{E}u_i^2\mathbb{E}u_1^2\right),$$

applying Lemma 2,

$$= \left(L-L^{-1}\right)\left(1+k_\varphi+d-1\right)$$
$$= \left(L-L^{-1}\right)\left(d+k_\varphi\right). \tag{16}$$

Then,

$$T_② = -2\left(\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2+T_④\right),$$

bringing Eq. (16) in,

$$= -2\left(\mathbb{E}\left(1+\sum_{i=1}^{d}u_i^2\right)\left\|\mathcal{T}_{\tilde{\lambda}}(\boldsymbol{u})-\bar{\boldsymbol{z}}\right\|_2^2+\left(L-L^{-1}\right)\left(d+k_\varphi\right)\right),$$

applying Lemma 5,

$$= -2\Big(\left(d + k_\varphi\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2 + (d+1) \left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 + \left(d + k_\varphi\right)\left(L - L^{-1}\right)\Big) \tag{17}$$

Lower Bound on $T_③$ Similarly for $T_③$, we can apply Lemma 8 Item (ii) as

$$T_③ = 2\left(d+1\right) \left\|\boldsymbol{B}^{-1}\widetilde{\boldsymbol{C}}^{-1}\left(\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\right)\right\|_2^2 = 2\left(d+1\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2. \tag{18}$$

Combining Eqs. (15), (17) and (18),

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{g}_{\mathrm{STL}}\left(\widetilde{\boldsymbol{\lambda}}\right)\right\|_2^2 &\geq T_① + T_② + T_③ \\
&\geq L^2\left(d + k_\varphi\right) \left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 - 2\Big(\left(d + k_\varphi\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2 + (d+1) \left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 + \left(d + k_\varphi\right)\left(L - L^{-1}\right)\Big) + 2\left(d+1\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2 \\
&= \left(L^2\left(d + k_\varphi\right) - 2\left(d+1\right)\right) \left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 - 2\left(k_\varphi - 1\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2 + \left(d + k_\varphi\right)\left(L - L^{-1}\right),
\end{aligned}$$

and when $L \geq 1$, we have $L - L^{-1} > 0$. Therefore, we can simply the bound as

$$\geq \left(L^2\left(d + k_\varphi\right) - 2\left(d+1\right)\right) \left\|\widetilde{\boldsymbol{C}}\right\|_{\mathrm{F}}^2 - 2\left(k_\varphi - 1\right) \|\widetilde{\boldsymbol{m}} - \bar{\boldsymbol{z}}\|_2^2.$$

□

C.6 Upper Bound on Gradient Variance of CFE

C.6.1 Full-Rank Parameterization

Theorem 4. *Assume Assumption 2 and that ℓ is L-log-smooth. For the full-rank parameterization, the expected-squared norm of the CFE estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2 \leq \alpha_{\mathrm{CFE}}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 + \beta_{\mathrm{CFE}}$$

where

$$\alpha_{\mathrm{CFE}} = L^2\left(d+k_\varphi\right)(1+\delta) + (L+S)^2$$
$$\beta_{\mathrm{CFE}} = L^2\left(d+k_\varphi\right)\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2$$

for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \in \Lambda_S$ and $\delta > 0$, where $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{z}}, \mathbf{0})$ and $\bar{\boldsymbol{z}}$ is any stationary point of f.

Proof. Following the notation of Domke et al. (2023a), we denote $\log \ell = f$. Then, starting from the definition of the variance,

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2 = \mathrm{tr}\mathbb{V}\boldsymbol{g}(\boldsymbol{\lambda}) + \|\mathbb{E}\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2,$$

and by the unbiasedness of $\boldsymbol{g}_{\mathrm{CFE}}$,

$$= \mathrm{tr}\mathbb{V}\boldsymbol{g}(\boldsymbol{\lambda}) + \|\nabla F(\boldsymbol{\lambda})\|_2^2,$$

by the definition of $\boldsymbol{g}_{\mathrm{CFE}}$ (Definition 4),

$$= \mathrm{tr}\mathbb{V}_{\boldsymbol{z}\sim q_\lambda}\left(\nabla_\lambda f(\boldsymbol{z}) + \nabla\mathbb{H}(q_\lambda)\right) + \|\nabla F(\boldsymbol{\lambda})\|_2^2.$$

We now apply the property of the variance: the deterministic components are neglected as

$$= \mathrm{tr}\mathbb{V}_{\boldsymbol{z}\sim q_\lambda}\nabla_\lambda f(\boldsymbol{z}) + \|\nabla F(\boldsymbol{\lambda})\|_2^2$$
$$\leq \mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2 + \|\nabla F(\boldsymbol{\lambda})\|_2^2. \tag{19}$$

For L-log-smooth posteriors (L-smooth f), Domke (2019, Theorem 3) show that

$$\mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2 \leq L^2\left(\left(d+k_\varphi\right)\|\boldsymbol{m}-\bar{\boldsymbol{z}}\|_2^2 + (d+1)\|\boldsymbol{C}\|_{\mathrm{F}}^2\right),$$

and since $k_\varphi \geq 1$,

$$\leq L^2\left(\left(d+k_\varphi\right)\|\boldsymbol{m}-\bar{\boldsymbol{z}}\|_2^2 + \left(d+k_\varphi\right)\|\boldsymbol{C}\|_{\mathrm{F}}^2\right)$$
$$= L^2\left(d+k_\varphi\right)\|\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}}\|_2^2,$$

which is tight.

Applying Eq. (1), we have

$$\mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2 \leq L^2\left(d+k_\varphi\right)\|\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}}\|_2^2$$
$$= L^2\left(d+k_\varphi\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*+\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2$$
$$\leq L^2\left(d+k_\varphi\right)\left((1+\delta)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2 + \left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2\right). \tag{20}$$

Now, for $\boldsymbol{\lambda} \in \Lambda_S$, Domke (2020, Theorem 1 & Lemma 12) show that the negative ELBO F is $(L+S)$-smooth as

$$\|\nabla F(\boldsymbol{\lambda})\|_2^2 = \|\nabla F(\boldsymbol{\lambda}) - \nabla F(\boldsymbol{\lambda}^*)\|_2^2 \leq (L+S)^2\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2. \tag{21}$$

Now back to Eq. (19),

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2 \leq \mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2 + \|\nabla F(\boldsymbol{\lambda})\|_2^2$$

applying Eq. (20),

$$\leq L^2 \left(d + k_\varphi\right) \left((1+\delta) \left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^* \right\|_2^2 + \left(1 + \delta^{-1}\right) \left\| \boldsymbol{\lambda}^* - \bar{\boldsymbol{\lambda}} \right\|_2^2 \right) + \left\| \nabla F \left(\boldsymbol{\lambda} \right) \right\|_2^2$$

and Eq. (21),

$$\begin{aligned}
&\leq L^2 \left(d + k_\varphi\right) \left((1+\delta) \left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^* \right\|_2^2 + \left(1 + \delta^{-1}\right) \left\| \boldsymbol{\lambda}^* - \bar{\boldsymbol{\lambda}} \right\|_2^2 \right) + \left(L + S\right)^2 \left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^* \right\|_2^2 \\
&= \left(L^2 \left(d + k_\varphi\right) (1+\delta) + \left(L + S\right)^2 \right) \left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^* \right\|_2^2 + L^2 \left(d + k_\varphi\right) \left(1 + \delta^{-1}\right) \left\| \boldsymbol{\lambda}^* - \bar{\boldsymbol{\lambda}} \right\|_2^2 .
\end{aligned}$$

□

C.6.2 Mean-Field Parameterization

Theorem 8. *Assume Assumption 2 and that ℓ is L-log-smooth. For the mean-field parameterization, the expected-squared norm of the CFE estimator is bounded as*

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2 \leq \left(\left(2k_\varphi\sqrt{d}+1\right)(1+\delta)+(L+S)^2\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2+\left(2k_\varphi\sqrt{d}+1\right)\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2.$$

for any $\boldsymbol{\lambda}\in\Lambda_S$ and $\delta\geq 0$, where $\bar{\boldsymbol{\lambda}}=(\bar{\boldsymbol{z}},\mathbf{0})$ and $\bar{\boldsymbol{z}}$ is any stationary point of f.

Proof. For the mean-field case, the only difference with Theorem 4 is the upper bound on the energy term. The key step is the mean-field part of Lemma 5, first proven by Kim et al. (2023b). The remaining steps are similar to Theorem 1 of Kim et al. (2023b). That is,

$$\mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2=\mathbb{E}\|\nabla_\lambda f(\mathcal{T}_\lambda(\boldsymbol{u}))\|_2^2,$$

applying Lemma 3,

$$=\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\|\nabla f(\mathcal{T}_\lambda(\boldsymbol{u}))\|_2^2$$
$$=\mathbb{E}J_{\mathcal{T}}(\boldsymbol{u})\|\nabla f(\mathcal{T}_\lambda(\boldsymbol{u}))-\nabla f(\bar{\boldsymbol{z}})\|_2^2,$$

from L-smoothness of $f=\log\ell$,

$$\leq L^2 J_{\mathcal{T}}(\boldsymbol{u})\,\mathbb{E}\|\mathcal{T}_\lambda(\boldsymbol{u})-\bar{\boldsymbol{z}}\|_2^2,$$

applying Lemma 5,

$$\leq L^2\left(\sqrt{dk_\varphi}+k_\varphi\sqrt{d}+1\right)\|\boldsymbol{m}-\bar{\boldsymbol{z}}\|_2^2+L^2\left(2k_\varphi\sqrt{d}+1\right)\|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

and since $k_\varphi\geq 1$, we have $k_\varphi>\sqrt{k_\varphi}$, and thus

$$\leq L^2\left(2k_\varphi\sqrt{d}+1\right)\left(\|\boldsymbol{m}-\bar{\boldsymbol{z}}\|_2^2+\|\boldsymbol{C}\|_{\mathrm{F}}^2\right)$$
$$=L^2\left(2k_\varphi\sqrt{d}+1\right)\|\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}}\|_2^2.$$

We finally apply Eq. (1) as

$$\leq L^2\left(2k_\varphi\sqrt{d}+1\right)\left((1+\delta)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2+\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2\right). \tag{22}$$

Combining this with Eqs. (19) and (21), we have

$$\mathbb{E}\|\boldsymbol{g}_{\mathrm{CFE}}(\boldsymbol{\lambda})\|_2^2=\mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2+\|\nabla F(\boldsymbol{\lambda})\|_2^2$$

and applying Eq. (22),

$$\leq\mathbb{E}_{\boldsymbol{z}\sim q_\lambda}\|\nabla_\lambda f(\boldsymbol{z})\|_2^2+(L+S)^2\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2$$
$$\leq\left(2k_\varphi\sqrt{d}+1\right)\left(L^2(1+\delta)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2+L^2\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2\right)+(L+S)^2\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2$$
$$=\left(\left(2k_\varphi\sqrt{d}+1\right)L^2(1+\delta)+(L+S)^2\right)\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|_2^2+L^2\left(2k_\varphi\sqrt{d}+1\right)\left(1+\delta^{-1}\right)\|\boldsymbol{\lambda}^*-\bar{\boldsymbol{\lambda}}\|_2^2.$$

□

C.7 Non-Asymptotic Complexity of Projected SGD

To precisely compare the computational complexity resulting from different estimators, we refine the convergence analyses of Domke et al. (2023a). Specifically, we obtain precise complexity guarantees from their "anytime convergence" statements. This type of convergence analysis, which has been popular in the ERM sample selection strategy literature (Csiba and Richtárik, 2018, §1.1), is convenient for comparing the lower-order and constant factor improvements of different gradient estimators.

C.7.1 QVC Gradient Estimator

Theorem 9 (**Strongly convex F with a fixed stepsize**). *For a μ-strongly convex $F : \Lambda \to \mathbb{R}$ on a convex set Λ with a unique global minimizer $\boldsymbol{\lambda}^* \in \Lambda$, the last iterate $\boldsymbol{\lambda}_T$ of projected SGD with a fixed stepsize satisfies $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$\gamma = \min\left(\frac{\epsilon\mu}{4\beta}, \frac{\mu}{2\alpha}, \frac{2}{\mu}\right) \quad \textit{and} \quad T \geq \max\left(\frac{4\beta}{\mu^2}\frac{1}{\epsilon}, \frac{2\alpha}{\mu^2}, \frac{1}{2}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).$$

Proof. Theorem 6 of Domke et al. (2023a) utilizes the two-stage stepsize of (Gower et al., 2019). The anytime convergence of the first stage,

$$\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq (1 - \gamma\mu)^T\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 + \frac{2\gamma\beta}{\mu}$$

corresponds to the SGD with only a fixed stepsize $\gamma < \frac{\mu}{2\alpha}$.

Here, the result follows from Lemma A.2 of Garrigos and Gower (2023) by plugging the constants

$$\alpha_0 = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2, \quad A = \frac{2\beta}{\mu}, \text{ and } \quad C = \frac{2\alpha}{\mu}, \frac{\mu}{2}.$$

□

Theorem 10 (**Strongly convex F with a decreasing stepsize schedule**). *For a μ-strongly convex $F : \Lambda \to \mathbb{R}$ on a convex set Λ with a unique global minimizer $\boldsymbol{\lambda}^* \in \Lambda$, the last iterate $\boldsymbol{\lambda}_T$ of projected SGD with a descreasing stepsize satisfies $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$\gamma_t = \min\left(\frac{\mu}{2\alpha}, \frac{4t+2}{\mu\,(t+1)^2}\right) \quad \textit{and} \quad T \geq \frac{16\beta}{\mu^2}\frac{1}{\epsilon} + \frac{8\alpha\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}.$$

Proof. Theorem 6 of Domke et al. (2023a) utilizes the two-stage stepsize of Gower et al. (2019). After T steps, with a carefully tuned stepsize of

$$\gamma_t = \min\left(\frac{\mu}{2\alpha}, \frac{4t+2}{\mu\,(t+1)^2}\right)$$

projected SGD achieves

$$\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \frac{64\alpha}{\mu^2}\frac{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2}{T^2} + \frac{32\beta}{\mu^2}\frac{1}{T}.$$

Following a similar strategy to Kim et al. (2023a), we can obtain a computational complexity by solving for the smallest T that achieves

$$\frac{64\alpha}{\mu^2}\frac{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2}{T^2} + \frac{16\beta}{\mu^2}\frac{1}{T} \leq \epsilon.$$

After re-organizing, we solve for

$$A\,T^2 + B\,T + C = 0,$$

where

$$A = \epsilon, \quad B = -\frac{16\beta}{\mu^2}, \text{ and } \quad C = -\frac{64\alpha^2}{\mu^4}\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2.$$

Since $T > 0$, the equation has a unique root

$$T = \frac{-B + \sqrt{B^2 - 4AC}}{2A},$$

applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$,

$$\leq \frac{-B + \sqrt{B^2} + \sqrt{4A\,(-C)}}{2A} = \frac{2\,(-B)}{2A} + \frac{\sqrt{4A\,(-C)}}{2A} = \frac{(-B)}{A} + \frac{\sqrt{(-C)}}{\sqrt{A}}$$

$$= \frac{16\beta}{\mu^2\epsilon} + \frac{\sqrt{\frac{64\alpha^2}{\mu^4}\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2}}{\sqrt{\epsilon}}$$

$$= \frac{16\beta}{\mu^2\epsilon} + \frac{8\alpha\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2\sqrt{\epsilon}}.$$

□

C.7.2 Adaptive QVC Gradient Estimator

As mentioned at the beginning of § 3.2, we established *adaptive* QV bounds. For the complexity guarantees for strongly convex objectives (Theorems 9 and 10), it is possible to optimize the free parameter δ in the bounds, such that they automatically adapt to other problem-specific constants. In this section, we do this for both SGD with fixed stepsize (Lemma 9) and a decreasing (Lemma 10) stepsize schedule.

Lemma 9 (**Strongly convex F with adaptive QV and Fixed Stepsize**). *For a μ-strongly convex $F : \Lambda \to \mathbb{R}$ on a convex set Λ the last iterate $\boldsymbol{\lambda}_T$ of projected SGD with a gradient estimator satisfying an adaptive QV bound (Assumption 3) is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 < \epsilon$ if*

$$\gamma = \min\left(\frac{1}{2}\frac{\mu}{\widetilde{\alpha} + 2\widetilde{\beta}\epsilon^{-1}}, \frac{2}{\mu}\right) \quad \text{and}$$

$$T \geq \frac{2}{\mu^2}\max\left(\widetilde{\alpha} + 2\widetilde{\beta}\frac{1}{\epsilon}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).$$

Proof. Recall that, for a stepsize γ and a number of steps T satisfying

$$\gamma \leq \min\left(\frac{\epsilon\mu}{4\beta}, \frac{\mu}{2\alpha}, \frac{2}{\mu}\right) \quad \text{and} \quad T \geq \max\left(\frac{4\beta}{\mu^2}\frac{1}{\epsilon}, \frac{2\alpha}{\mu^2}, \frac{1}{2}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right),$$

we can guarantee that the iterate $\boldsymbol{\lambda}_t$ can guarantee $\mathbb{E}\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_T\|_2^2 \leq \epsilon$.

We optimize the parameter δ to minimize the number of steps. That is,

$$\max\left(\frac{4\beta}{\mu^2}\frac{1}{\epsilon}, \frac{2\alpha}{\mu^2}, \frac{1}{2}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right) = \frac{2}{\mu^2}\max\left(2(1 + C^{-1}\delta^{-1})\widetilde{\beta}\frac{1}{\epsilon}, (1 + C\delta)\widetilde{\alpha}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).$$

Since the first and second arguments of the max function are monotonic with respect to δ, the optimum is unique, and achieved when the two terms are equal. That is,

$$\begin{aligned}
& 2(1 + C^{-1}\delta^{-1})\widetilde{\beta}\frac{1}{\epsilon} = (1 + C\delta)\widetilde{\alpha} \\
\Leftrightarrow \quad & \frac{2\widetilde{\beta}}{\epsilon} + \frac{2\widetilde{\beta}C^{-1}}{\epsilon}\delta^{-1} = \widetilde{\alpha} + \widetilde{\alpha}C\delta \\
\Leftrightarrow \quad & \frac{2\widetilde{\beta}}{\epsilon}\delta + \frac{2\widetilde{\beta}C^{-1}}{\epsilon} = \widetilde{\alpha}\delta + \widetilde{\alpha}C\delta^2 \\
\Leftrightarrow \quad & \widetilde{\alpha}C\delta^2 + \left(\widetilde{\alpha} - \frac{2\widetilde{\beta}}{\epsilon}\right)\delta - \frac{2\widetilde{\beta}C^{-1}}{\epsilon} = 0 \\
\Leftrightarrow \quad & \left(\widetilde{\alpha}\delta - \frac{2\widetilde{\beta}C^{-1}}{\epsilon}\right)(C\delta + 1) = 0.
\end{aligned}$$

Conveniently, we have a unique feasible solution

$$\delta = 2\frac{\widetilde{\beta}}{\widetilde{\alpha}}C^{-1}\epsilon^{-1}.$$

Thus, the optimal bound is obtained by setting $\delta = 2\frac{\widetilde{\beta}}{\widetilde{\alpha}}C^{-1}\epsilon^{-1}$, such that

$$\begin{aligned}
T &\geq \frac{2}{\mu^2}\max\left(2\beta\frac{1}{\epsilon}, \alpha, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right) \\
&= \frac{2}{\mu^2}\max\left(2\left(1 + C^{-1}\delta^{-1}\right)\widetilde{\beta}\frac{1}{\epsilon}, (1 + C\delta)\,\widetilde{\alpha}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right) \\
&= \frac{2}{\mu^2}\max\left(\widetilde{\alpha} + 2\widetilde{\beta}\frac{1}{\epsilon}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).
\end{aligned}$$

The stepsize with the optimal δ is consequently

$$\gamma \leq \min\left(\frac{\epsilon\mu}{4\beta}, \frac{\mu}{2\alpha}, \frac{2}{\mu}\right) = \min\left(\frac{\epsilon\mu}{4(1+C^{-1}\delta^{-1})\widetilde{\beta}}, \frac{\mu}{2(1+C\delta)\widetilde{\alpha}}, \frac{2}{\mu}\right) = \min\left(\frac{1}{2}\frac{\mu}{\widetilde{\alpha}+2\widetilde{\beta}\epsilon^{-1}}, \frac{2}{\mu}\right).$$

□

Lemma 10 (**Strongly convex F with adaptive QV and Decreasing Stepsize**). *For a μ-strongly convex $F : \Lambda \to \mathbb{R}$ on a convex set Λ with a unique global minimizer $\boldsymbol{\lambda}^* \in \Lambda$, the last iterate $\boldsymbol{\lambda}_T$ of projected SGD with a gradient estimator satisfying an adaptive QVC bound (Assumption 3) and a decreasing stepsize satisfies a suboptimality of $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}_*\|_2^2 < \epsilon$ if*

$$\gamma_t = \min\left(\frac{\mu}{2\widetilde{\alpha} + \sqrt{2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\epsilon^{1/4}\,\widetilde{\alpha}^{3/2}\widetilde{\beta}^{-1/2}}, \frac{4t+2}{\mu\,(t+1)^2}\right)$$

$$T \geq \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\,\frac{1}{\epsilon^{3/4}} + \frac{8\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}.$$

Proof. Recall that, for a stepsize γ and a number of steps T such that

$$\gamma_t = \min\left(\frac{\mu}{2\alpha}, \frac{4t+2}{\mu\,(t+1)^2}\right) \quad \text{and} \quad T \geq \frac{16\beta}{\mu^2}\frac{1}{\epsilon} + \frac{8\alpha\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}},$$

we can guarantee that the iterate $\boldsymbol{\lambda}_t$ can guarantee $\mathbb{E}\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_T\|_2^2 \leq \epsilon$.

We optimize the parameter δ to minimize the required number of steps T. That is, we maximize

$$\frac{16\beta}{\mu^2}\frac{1}{\epsilon} + \frac{8\sqrt{2}\,\alpha\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} = \frac{16\,(1+C\delta)\,\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{8\left(1+C^{-1}\delta^{-1}\right)\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}.$$

This is clearly a convex function with respect to δ. Thus, we only need to find a first-order stationary point

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\left(\frac{16\,(1+C\delta)\,\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{8\left(1+C^{-1}\delta^{-1}\right)\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}\right) = 0.$$

Differentiating, we have

$$\frac{16C\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} - \frac{8\,C^{-1}\delta^{-2}\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} = 0,$$

multiplying δ^2 to both sides,

$$\Leftrightarrow \quad \delta^2\frac{16C\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} - \frac{8\sqrt{2}\,C^{-1}\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} = 0.$$

Reorganizing,

$$\Leftrightarrow \quad \delta^2\frac{16C\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} = \frac{8\sqrt{2}\,C^{-1}\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}$$

$$\Leftrightarrow \quad \delta^2 = \left(\frac{\mu^2\epsilon}{16C\widetilde{\beta}}\right)\left(\frac{8C^{-1}\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}\right)$$

$$\Leftrightarrow \quad \delta^2 = \frac{C^{-2}\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{2\widetilde{\beta}}\sqrt{\epsilon},$$

and taking the square-root of both sides,

$$\Leftrightarrow \quad \delta = \frac{\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}{\sqrt{2}\,C\sqrt{\widetilde{\beta}}}.$$

Recall that the required number of iterations is

$$\begin{aligned} T &\geq \frac{16\left(1+C\delta\right)\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{8\left(1+C^{-1}\delta^{-1}\right)\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} \\ &= \underbrace{\frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon}C\delta}_{T_{①}} + \underbrace{\frac{8\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} + \frac{8\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}C^{-1}\delta^{-1}}_{T_{②}}. \end{aligned}$$

Plugging δ in, we have

$$\begin{aligned} T_{①} &= \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon}C\left(\frac{\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}{\sqrt{2}C\sqrt{\widetilde{\beta}}}\right) \\ &= \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{8\sqrt{2}}{\mu^2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{-3/4} \end{aligned}$$

$$\begin{aligned} T_{②} &= \frac{8\,\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} + \frac{8\,\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}C^{-1}\left(\frac{\sqrt{2}\,C\sqrt{\widetilde{\beta}}}{\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}\right) \\ &= \frac{8\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} + \frac{8\sqrt{2}}{\mu^2}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\,\epsilon^{-3/4}. \end{aligned}$$

Combining the results,

$$\begin{aligned} T \geq T_{①} + T_{②} &= \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{8\sqrt{2}}{\mu^2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{-3/4} \\ &\quad + \frac{8\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} + \frac{8\sqrt{2}}{\mu^2}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\,\epsilon^{-3/4} \\ &= \frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\,\epsilon^{-3/4} + \frac{8\widetilde{\alpha}\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}}. \end{aligned}$$

For the stepsize

$$\gamma = \min\left(\frac{\mu}{2\alpha}, \frac{4t+2}{\mu\left(t+1\right)^2}\right) = \min\left(\frac{\mu}{2\left(1+C\delta\right)\widetilde{\alpha}}, \frac{4t+2}{\mu\left(t+1\right)^2}\right),$$

we have

$$\begin{aligned} 2\left(1+C\delta\right)\widetilde{\alpha} &= 2\widetilde{\alpha} + 2\widetilde{\alpha}C\delta \\ &= 2\widetilde{\alpha} + 2\widetilde{\alpha}C\left(\frac{\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}{\sqrt{2}C\sqrt{\widetilde{\beta}}}\right) \\ &= 2\widetilde{\alpha} + \sqrt{2}\sqrt{\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\,\widetilde{\alpha}^{3/2}\widetilde{\beta}^{-1/2}. \end{aligned}$$

Therefore,

$$\gamma = \min\left(\frac{\mu}{2\left(1+C\delta\right)\widetilde{\alpha}}, \frac{4t+2}{\mu\left(t+1\right)^2}\right) = \min\left(\frac{\mu}{2\widetilde{\alpha} + \sqrt{2\left\|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\,\widetilde{\alpha}^{3/2}\widetilde{\beta}^{-1/2}}, \frac{4t+2}{\mu\left(t+1\right)^2}\right).$$

□

C.8 Non-Asymptotic Complexity of BBVI

C.8.1 CFE Gradient Estimator

Theorem 5 (**Complexity of Fixed Stepsize BBVI with CFE**). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the CFE estimator and projected SGD with a fixed stepsize applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 2\kappa^2 \left(d + k_\varphi + 4\right)\left(1 + 2\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right)\log\left(2\Delta^2 \frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2$, and $\kappa = L/\mu$ is the condition number.

Proof. From Theorem 4 with $S = L$, the CFE estimator satisfies adaptive QV with the constants

$$\alpha_{\text{CFE}} = L^2\left(d + k_\varphi + 4\right)(1+\delta) \qquad \text{and} \qquad \beta_{\text{CFE}} = L^2\left(d + k_\varphi\right)\left(1 + \delta^{-1}\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2.$$

Furthermore, for a μ-strongly log-concave posterior and our variational parameterization, Domke (2020, Theorem 9) show that the ELBO is μ-strongly convex.

We can thus invoke Lemma 9 with

$$\widetilde{\alpha} = L^2\left(d + k_\varphi + 4\right), \qquad \widetilde{\beta} = L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2, \quad \text{and} \quad C = 1.$$

This yields a lower bound on the number of iteration

$$\begin{aligned}
&\frac{2}{\mu^2}\max\left(\widetilde{\alpha} + 2\widetilde{\beta}\frac{1}{\epsilon},\ \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right) \\
&= \frac{2}{\mu^2}\max\left(L^2\left(d + k_\varphi + 4\right) + 2L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon},\ \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right),
\end{aligned}$$

pulling out L,

$$= \frac{2L^2}{\mu^2}\max\left(\left(d + k_\varphi + 4\right) + 2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon},\ \frac{\mu^2}{4L^2}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right),$$

and since $\frac{\mu^2}{4L^2} < \frac{1}{4}$ and the first argument is larger than 1, the max operation is redundant that

$$= \frac{2L^2}{\mu^2}\left(\left(d + k_\varphi + 4\right) + 2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right).$$

Now, using the trivial fact $d + k_\varphi < d + k_\varphi + 4$ simplifies the bound as,

$$\begin{aligned}
&< \frac{2L^2}{\mu^2}\left(d + k_\varphi + 4\right)\left(1 + 2\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right) \\
&= 2\kappa^2\left(d + k_\varphi + 4\right)\left(1 + 2\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right).
\end{aligned}$$

The optimal δ is given as

$$\delta = \frac{2}{\epsilon}\frac{\widetilde{\beta}}{\widetilde{\alpha}}C^{-1} = \frac{2}{\epsilon}\frac{L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2}{L^2\left(d + k_\varphi + 4\right)}C^{-1} = \frac{2}{\epsilon}\frac{d + k_\varphi}{d + k_\varphi + 4}\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2.$$

□

Theorem 11 (**Complexity of Decreasing Stepsize BBVI with CFE**). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the CFE estimator and projected SGD with a decreasing stepsize schedule applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 16\kappa^2 \left(d + k_\varphi + 4\right)\left(\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon} + 2\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2 \frac{1}{\epsilon^{3/4}} + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2 \frac{1}{\sqrt{\epsilon}}\right).$$

for some decreasing stepsize schedule $\gamma_1, \ldots, \gamma_T$, where $\kappa = L/\mu$ is the condition number and $\boldsymbol{\lambda}^* \in \Lambda$ is the optimal variational parameter.

Proof. From Theorem 4, the CFE estimator with $S = L$ satisfies adaptive QV with the constants

$$\alpha_{\text{CFE}} = L^2\left(d + k_\varphi + 4\right)(1+\delta) \qquad \text{and} \qquad \beta_{\text{CFE}} = L^2\left(d + k_\varphi\right)\left(1 + \delta^{-1}\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2.$$

Furthermore, for a μ-strongly log-concave posterior and our variational parameterization, Domke (2020, Theorem 9) show that the ELBO is μ-strongly convex.

We thus invoke Lemma 10 with

$$\widetilde{\alpha} = L^2\left(d + k_\varphi + 4\right), \qquad \widetilde{\beta} = L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2, \qquad \text{and} \qquad C = 1.$$

This yields a lower bound on the number of iterations:

$$\begin{aligned}
&\frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\frac{1}{\epsilon^{3/4}} + \frac{8\widetilde{\alpha}\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} \\
&= \frac{16L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{\left(L^2\left(d + k_\varphi + 4\right)\right)\left(L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\right)}\frac{1}{\epsilon^{3/4}} \\
&\quad + \frac{8L^2\left(d + k_\varphi + 4\right)\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}},
\end{aligned}$$

using the trivial bound $d + k_\varphi < d + k_\varphi + 4$,

$$\begin{aligned}
&< \frac{16L^2\left(d + k_\varphi + 4\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{L^4\left(d + k_\varphi + 4\right)\left(d + k_\varphi + 4\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2}\frac{1}{\epsilon^{3/4}} \\
&\quad + \frac{8L^2\left(d + k_\varphi + 4\right)\,\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}},
\end{aligned}$$

pulling out the $16\left(d + k_\varphi + 4\right)L^2/\mu^2$ factors,

$$\begin{aligned}
&= 16\left(d + k_\varphi + 4\right)\frac{L^2}{\mu^2}\left(\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon} + \sqrt{2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2\frac{1}{\epsilon^{3/4}} + \frac{1}{2}\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2\frac{1}{\sqrt{\epsilon}}\right) \\
&= 16\kappa^2\left(d + k_\varphi + 4\right)\left(\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon} + \sqrt{2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2\frac{1}{\epsilon^{3/4}} + \frac{1}{2}\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2\frac{1}{\sqrt{\epsilon}}\right).
\end{aligned}$$

The optimal δ is given as

$$\delta = \frac{\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}{\sqrt{2}C\sqrt{\widetilde{\beta}}} = \frac{\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\,\epsilon^{1/4}\sqrt{L^2\left(d + k_\varphi + 4\right)}}{\sqrt{2}\sqrt{L^2\left(d + k_\varphi\right)\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2^2}} = \frac{1}{\sqrt{2}}\frac{\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}}{\|\bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2}\sqrt{\frac{d + k_\varphi + 4}{d + k_\varphi}}\,\epsilon^{-1/4}.$$

□

C.8.2 STL Gradient Estimator

Theorem 6 (**Complexity of Fixed Stepsize BBVI with STL**). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the STL estimator and projected SGD with a fixed stepsize applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 8\kappa^2 (d + k_\varphi) \left(1 + \frac{1}{L^2}\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}\right) \log\left(2\Delta^2 \frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2$ is the distance to the optimum and $\kappa = L/\mu$ is the condition number.

Proof. As shown by Theorem 1, the STL estimator with $S = L$ satisfies an adaptive QV bound with the constants

$$\alpha_{\mathrm{STL}} = 2(d + k_\varphi)(2 + \delta)L^2 = 4L^2(d + k_\varphi)\left(1 + \frac{1}{2}\delta\right)$$
$$\beta_{\mathrm{STL}} = (2d + k_\varphi)(1 + 2\delta^{-1})\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}.$$

Furthermore, for a μ-strongly log-concave posterior and our variational parameterization, Domke (2020, Theorem 9) show that the ELBO is μ-strongly convex. Thus, we can fully invoke Lemma 9 with

$$\widetilde{\alpha} = 4L^2(d + k_\varphi), \qquad \widetilde{\beta} = (2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}, \quad \text{and} \quad C = \frac{1}{2}.$$

This yields a lower bound on the number of iteration

$$\frac{2}{\mu^2}\max\left(\widetilde{\alpha} + 2\widetilde{\beta}\frac{1}{\epsilon}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)$$
$$= \frac{2}{\mu^2}\max\left(4L^2(d + k_\varphi) + 2(2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}, \frac{\mu^2}{4}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right),$$

pulling out the L^2 factor,

$$= \frac{2L^2}{\mu^2}\max\left(4(d + k_\varphi) + 2\frac{1}{L^2}(2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}, \frac{\mu^2}{4L^2}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right),$$

and since $\frac{\mu^2}{4L^2} < \frac{1}{4}$ and the first argument is larger than 1 due to $k_\varphi \geq 1$, the max operation is redundant such that

$$= \frac{2L^2}{\mu^2}\left(4(d + k_\varphi) + 2\frac{1}{L^2}(2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).$$

Now, using the trivial fact $2d + k_\varphi < 2d + 2k_\varphi$ simplifies the bound as,

$$< \frac{8L^2}{\mu^2}(d + k_\varphi)\left(1 + \frac{1}{L^2}\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right)$$
$$= 8\kappa^2(d + k_\varphi)\left(1 + \frac{1}{L^2}\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}\frac{1}{\epsilon}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\frac{1}{\epsilon}\right).$$

The optimal δ is given as

$$\delta = \frac{2}{\epsilon}\frac{\widetilde{\beta}}{\widetilde{\alpha}}C^{-1} = \frac{2}{\epsilon}\frac{(2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}}{4L^2(d + k_\varphi)}2 = \frac{4}{\epsilon}\frac{\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\lambda^*}, \pi)}}{L^2}\frac{2d + k_\varphi}{d + k_\varphi}.$$

□

Theorem 12 (**Complexity of Decreasing Stepsize BBVI with STL**). *The last iterate $\boldsymbol{\lambda}_T \in \Lambda_L$ of BBVI with the STL estimator and projected SGD with a decreasing stepsize schedule applied to a μ-strongly log-concave and L-log-smooth posterior is ϵ-close to $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda_L} F(\boldsymbol{\lambda})$ such that $\|\boldsymbol{\lambda}_T - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ if*

$$T \geq 32\kappa^2 (d + k_\varphi)\left(\frac{\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \pi)}}{L^2}\frac{1}{\epsilon} + \frac{1}{\sqrt{2}}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\frac{(\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \pi))^{1/4}}{L}\frac{1}{\epsilon^{3/4}} + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2\frac{1}{\sqrt{\epsilon}}\right),$$

for some decreasing stepsize schedule $\gamma_1 \geq \ldots \geq \gamma_T$, where $\kappa = L/\mu$ is the condition number.

Proof. As shown by Theorem 1, the STL estimator with $S = L$ satisfies an adaptive QV bound with the constants

$$\alpha_{\mathrm{STL}} = 2(d + k_\varphi)(2 + \delta)L^2 = 4L^2(d + k_\varphi)\left(1 + \frac{1}{2}\delta\right)$$
$$\beta_{\mathrm{STL}} = (2d + k_\varphi)(1 + 2\delta^{-1})\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \pi)}.$$

Furthermore, for a μ-strongly log-concave posterior and our variational parameterization, Domke (2020, Theorem 9) show that the ELBO is μ-strongly convex. Thus, we can invoke Lemma 10 with

$$\widetilde{\alpha} = 4L^2(d + k_\varphi), \qquad \widetilde{\beta} = (2d + k_\varphi)\sqrt{\mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \pi)}, \qquad \text{and} \qquad C = \frac{1}{2}.$$

This yields a lower bound on the number of iterations:

$$\begin{aligned}
&\frac{16\widetilde{\beta}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{\widetilde{\alpha}\widetilde{\beta}}\frac{1}{\epsilon^{3/4}} + \frac{8\widetilde{\alpha}\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}} \\
&= \frac{16(2d + k_\varphi)\sqrt{D^*_{\mathrm{F}^4}}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{4L^2(d + k_\varphi)(2d + k_\varphi)\sqrt{D^*_{\mathrm{F}^4}}}\frac{1}{\epsilon^{3/4}} \\
&\quad + \frac{32L^2(d + k_\varphi)\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}},
\end{aligned}$$

using the the trivial bound $2d + k_\varphi < 2d + 2k_\varphi$,

$$\begin{aligned}
&< \frac{32(d + k_\varphi)\sqrt{D^*_{\mathrm{F}^4}}}{\mu^2}\frac{1}{\epsilon} + \frac{16\sqrt{2}}{\mu^2}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\sqrt{8L^2(d + k_\varphi)(d + k_\varphi)\sqrt{D^*_{\mathrm{F}^4}}}\frac{1}{\epsilon^{3/4}} \\
&\quad + \frac{32L^2(d + k_\varphi)\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}{\mu^2}\frac{1}{\sqrt{\epsilon}},
\end{aligned}$$

pulling out the $32(d + k_\varphi)L^2/\mu^2$ factors,

$$\begin{aligned}
&= 32\frac{L^2}{\mu^2}(d + k_\varphi)\left(\frac{\sqrt{D^*_{\mathrm{F}^4}}}{L^2}\frac{1}{\epsilon} + \frac{1}{\sqrt{2}}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\frac{\left(D^*_{\mathrm{F}^4}\right)^{1/4}}{L}\frac{1}{\epsilon^{3/4}} + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2\frac{1}{\sqrt{\epsilon}}\right) \\
&= 32\kappa^2(d + k_\varphi)\left(\frac{\sqrt{D^*_{\mathrm{F}^4}}}{L^2}\frac{1}{\epsilon} + \frac{1}{\sqrt{2}}\sqrt{\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2}\frac{\left(D^*_{\mathrm{F}^4}\right)^{1/4}}{L}\frac{1}{\epsilon^{3/4}} + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2\frac{1}{\sqrt{\epsilon}}\right),
\end{aligned}$$

where we have denoted $D^*_{\mathrm{F}^4} = \mathrm{D}_{\mathrm{F}^4}(q_{\boldsymbol{\lambda}^*}, \pi)$.

Also, the optimal δ is given as

$$\begin{aligned}
\delta &= \frac{\sqrt{\left\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\sqrt{\widetilde{\alpha}}}{\sqrt{2}\,C\sqrt{\widetilde{\beta}}} \\
&= \frac{\sqrt{\left\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\right\|_2}\,\epsilon^{1/4}\sqrt{4L^2\left(d + k_\varphi\right)}}{2^{-1}\sqrt{2}\sqrt{\left(2d + k_\varphi\right)\sqrt{D^*_{\mathrm{F}^4}}}} \\
&= 2\sqrt{2}L\sqrt{\left\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\right\|_2}\sqrt{\frac{d + k_\varphi}{2d + k_\varphi}}\left(D^*_{\mathrm{F}^4}\right)^{-1/2}\epsilon^{1/4} \\
&= 2\sqrt{2}L\sqrt{\frac{\left\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\right\|_2}{\mathrm{D}_{\mathrm{F}^4}\left(q_{\boldsymbol{\lambda}^*}, \pi\right)}}\sqrt{\frac{d + k_\varphi}{2d + k_\varphi}}\,\epsilon^{1/4}.
\end{aligned}$$

□

C.9 Fisher-Hyvärinen Divergence Between Gaussians

Lemma 11. *For $\pi = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q = \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{C}\boldsymbol{C}^\top\right)$, the Fisher-Hyvärinen divergence is*

$$\mathrm{D_F}(q, \pi) = \left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\right\|_2^2.$$

Proof. The result is straightforward using the reparameterization representation of the Gaussian. That is,

$$\nabla \log \pi(\boldsymbol{z}) = \nabla \log \pi(\mathcal{T}_\lambda(\boldsymbol{u})) = \boldsymbol{\Sigma}^{-1}(\mathcal{T}_\lambda(\boldsymbol{u}) - \boldsymbol{\mu}).$$

Using this, we have

$$\begin{aligned}
\mathrm{D_F}(q, \pi) &= \mathbb{E}_{\boldsymbol{z}\sim q}\left\|\nabla \log \pi(\boldsymbol{z}) - \nabla \log q(\boldsymbol{z})\right\|_2^2 \\
&= \mathbb{E}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{\mu}) - \left(\boldsymbol{C}\boldsymbol{C}^\top\right)^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{m})\right\|_2^2 \\
&= \mathbb{E}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{\mu}) - \left(\boldsymbol{C}\boldsymbol{C}^\top\right)^{-1}\boldsymbol{C}\boldsymbol{u}\right\|_2^2 \\
&= \mathbb{E}\left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{u} + \boldsymbol{m} - \boldsymbol{\mu}) - \boldsymbol{C}^{-\top}\boldsymbol{u}\right\|_2^2,
\end{aligned}$$

grouping the terms involving $\boldsymbol{C}$,

$$= \mathbb{E}\left\|\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\boldsymbol{u} + \boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\right\|_2^2,$$

expanding the quadratic,

$$= \mathbb{E}\left\|\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\boldsymbol{u}\right\|_2^2 + 2\left\langle\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\mathbb{E}\,\boldsymbol{u}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\right\rangle + \left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\right\|_2^2,$$

applying Assumption 1,

$$= \mathbb{E}\left\|\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\right\|_2^2.$$

The expectation term can be simplified as

$$\mathbb{E}\left\|\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\boldsymbol{u}\right\|_2^2 = \mathbb{E}\mathrm{tr}\left(\boldsymbol{u}^\top\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)^\top\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\boldsymbol{u}\right),$$

rotating the elements of the trace,

$$= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)^\top\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\mathbb{E}\,\boldsymbol{u}\boldsymbol{u}^\top\right),$$

applying Assumption 1,

$$\begin{aligned}
&= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)^\top\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right)\right) \\
&= \left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

□

Lemma 12. *Let $\pi = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{Q}$ be the mean-field Gaussian variational family. Then, the solution of the KL divergence minimization problem*

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathrm{D}_{\mathrm{KL}}(q, \pi),$$

where $q_* = \mathcal{N}\left(\boldsymbol{m}_*, \boldsymbol{C}_* \boldsymbol{C}_*^\top\right)$ is given as

$$\boldsymbol{m}_* = \boldsymbol{\mu}, \qquad \boldsymbol{C}_* = \operatorname{diag}(\boldsymbol{\Sigma})^{1/2}.$$

Proof. Consider that the KL divergence between Gaussian distributions is given as

$$\mathcal{L}(\boldsymbol{m}, \boldsymbol{C}) = \mathrm{D}_{\mathrm{KL}}(q, \pi) = \frac{1}{2}\left((\boldsymbol{m} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu}) + \log \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{C}\boldsymbol{C}^\top|} + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{C}^\top\right) - d\right).$$

Firstly, it is clear that $\boldsymbol{m} = \boldsymbol{m}_* = \boldsymbol{\mu}$ minimizes $\mathrm{D}_{\mathrm{KL}}(q, \pi)$ with respect to $\boldsymbol{m}$ regardless of $\boldsymbol{C}$. Then, we have

$$\mathcal{L}(\boldsymbol{m}_*, \boldsymbol{C}) = \frac{1}{2}\left(\log \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{C}\boldsymbol{C}^\top|} + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{C}^\top\right) - d\right) \propto -\log\left|\boldsymbol{C}\boldsymbol{C}^\top\right| + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\boldsymbol{C}^\top\right).$$

When $\boldsymbol{C}$ is a diagonal matrix, taking the partial derivative with respect to $\boldsymbol{C}$ yields

$$\left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{C}}\right|_{\boldsymbol{m}=\boldsymbol{m}_*} = -2\,\boldsymbol{C}^{-1} + 2\operatorname{diag}\left(\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{C}.$$

The first-order optimality condition with respect to $\boldsymbol{C}$ is then

$$(\boldsymbol{C}\boldsymbol{C})^{-1} = \operatorname{diag}\left(\boldsymbol{\Sigma}^{-1}\right).$$

Since $\boldsymbol{\Sigma}$ is always positive definite, its diagonal elements are always strictly positive. Therefore, the unique solution $\boldsymbol{C}^*$ is

$$\boldsymbol{C}_* = \operatorname{diag}(\boldsymbol{\Sigma})^{1/2}.$$

□

Proposition 3. *Let $\pi = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{Q}$ be the mean-field Gaussian variational family. Then, the Fisher-Hyvärinen divergence of the KL minimizer*

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathrm{D}_{\mathrm{KL}}(q, \pi)$$

is bounded as

$$\lambda_{\max}(\boldsymbol{D})^{-1} \|\boldsymbol{R}^{-1} - \mathbf{I}\|_{\mathrm{F}}^2 \leq \mathrm{D}_{\mathrm{F}}(q_*, \pi) \leq \lambda_{\min}(\boldsymbol{D})^{-1} \|\boldsymbol{R}^{-1} - \mathbf{I}\|_{\mathrm{F}}^2,$$

where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Sigma})$ and $\boldsymbol{R}$ is the correlation matrix of π such that $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$.

Proof. First, the Fisher-Hyvärinen divergence between Gaussians is given in Lemma 11 as

$$\mathrm{D}_{\mathrm{F}}(q, \pi) = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \boldsymbol{C}^{-\top}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu})\|_2^2.$$

Plugging the KL minimizer q_* given in Lemma 12,

$$\begin{aligned}\mathrm{D}_{\mathrm{F}}(q_*, \pi) &= \|\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \boldsymbol{C}_*^{-\top}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Sigma}^{-1}(\boldsymbol{m}_* - \boldsymbol{\mu})\|_2^2 \\ &= \|\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \boldsymbol{C}_*^{-1}\|_{\mathrm{F}}^2. \end{aligned} \tag{23}$$

From here, we can pull out a $\boldsymbol{C}_*^{-1}$ factor as

$$\|\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \boldsymbol{C}_*^{-1}\|_{\mathrm{F}}^2 = \|\boldsymbol{C}_*^{-1}(\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C} - \mathbf{I})\|_{\mathrm{F}}^2. \tag{24}$$

And from the property of the Frobenius norm,

$$\lambda_{\min}(\boldsymbol{C}_*^{-1})^2 \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{C}_*^{-1}(\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I})\|_{\mathrm{F}}^2 \leq \lambda_{\max}(\boldsymbol{C}_*^{-1})^2 \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2,$$

inverting the singular values,

$$\Leftrightarrow \quad \lambda_{\max}(\boldsymbol{C}_*)^{-2} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 \leq \|\boldsymbol{C}_*^{-1}(\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I})\|_{\mathrm{F}}^2 \leq \lambda_{\min}(\boldsymbol{C}_*)^{-2} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2,$$

by Eqs. (23) and (24),

$$\Leftrightarrow \quad \lambda_{\max}(\boldsymbol{C}_*)^{-2} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 \leq \mathrm{D}_{\mathrm{F}}(q_*, \pi) \leq \lambda_{\min}(\boldsymbol{C}_*)^{-2} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2.$$

Denoting $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Sigma})$, we know that $\boldsymbol{C}_* = \boldsymbol{D}^{1/2}$. Then,

$$\lambda_{\max}(\boldsymbol{D})^{-1} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 \leq \mathrm{D}_{\mathrm{F}}(q_*, \pi) \leq \lambda_{\min}(\boldsymbol{D})^{-1} \|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2.$$

Clearly, the behavior of the Fisher divergence is fully determined by the term

$$\|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2.$$

To further analyze this quantity, notice that the correlation matrix $\boldsymbol{R}$ is related with the covariance $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma})^{1/2} \boldsymbol{R} \,\mathrm{diag}(\boldsymbol{\Sigma})^{1/2} = \boldsymbol{C}_*\boldsymbol{R}\boldsymbol{C}_*.$$

Then, it immediately follows that

$$\|\boldsymbol{C}_*\boldsymbol{\Sigma}^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 = \|\boldsymbol{C}_*(\boldsymbol{C}_*\boldsymbol{R}\boldsymbol{C}_*)^{-1}\boldsymbol{C}_* - \mathbf{I}\|_{\mathrm{F}}^2 = \|\boldsymbol{R}^{-1} - \mathbf{I}\|_{\mathrm{F}}^2.$$

□