Provably Scalable Black-Box Variational Inference with Structured Variational Families

Joohwan Ko [*1] **Kyurae Kim** [*2] **Woo Chang Kim** [1] **Jacob R. Gardner** [2]

Abstract

Variational families with full-rank covariance approximations are known not to work well in black-box variational inference (BBVI), both empirically and theoretically. In fact, recent computational complexity results for BBVI have established that full-rank variational families scale poorly with the dimensionality of the problem compared to *e.g.* mean-field families. This is particularly critical to hierarchical Bayesian models with local variables; their dimensionality increases with the size of the datasets. Consequently, one gets an iteration complexity with an explicit $\mathcal{O}(N^2)$ dependence on the dataset size N. In this paper, we explore a theoretical middle ground *between* mean-field variational families and full-rank families: *structured* variational families. We rigorously prove that certain scale matrix structures can achieve a better iteration complexity of $\mathcal{O}(N)$, implying better scaling with respect to N. We empirically verify our theoretical results on large-scale hierarchical models.

1. Introduction

A decade has passed since black-box variational inference (BBVI; Ranganath et al., 2014; Titsias & Lázaro-Gredilla, 2014), also known as Monte Carlo variational inference and stochastic gradient variational Bayes, has emerged. Among various approaches to variational inference (VI; Jordan et al., 1999; Blei et al., 2017; Zhang et al., 2019), BBVI has been immensely successful in various fields such as statistics, machine learning, signal processing, and many more, thanks to its general black-box nature and scalability to large datasets.

One of the promises of BBVI has been that we can better model correlations in the posterior by using full-rank covariance approximations (Kucukelbir et al., 2017). That way, we should have been able to move away from mean-field (Peterson & Anderson, 1987; Hinton & van Camp, 1993), or diagonal covariance, approximations traditionally required when performing coordinate-ascent VI (CAVI; see the review by Blei et al. 2017). However, this promise has shown to be elusive. Despite their theoretical and occasional empirical superiority of "expressiveness," going full-rank does not always improve over mean-field. See, for example, the experiments in §3 by Zhang et al. (2022), §B.1 by Agrawal et al. (2020), Footnote 2 by Giordano et al. (2018).

A common explanation for the underwhelming performance of full-rank approximations has been their excessive gradient variance. Indeed, recent results by Kim et al. (2023a); Domke et al. (2023) establishing the computational complexity of BBVI confirm this intuition. Due to gradient variance, full-rank covariance approximations result in a $\mathcal{O}\left(d\kappa^2\epsilon^{-1}\right)$ iteration complexity for finding an ϵ-accurate solution on strongly log-concave posteriors with a condition number of κ. This contrasts with mean-field for which a $\mathcal{O}\left(\sqrt{d}\right)$ dimensional dependence (Kim et al., 2023b) has been established. The poor d dependence explains why the full-rank approximation is underwhelming in practice.

Furthermore, for models with *local variables* (Hoffman & Blei, 2015), the $\mathcal{O}(d)$ dimension dependence is especially concerning; for these models, the dimensionality d scales with the size of the dataset N. This means that for a model taking $\Theta(N)$ datapoint queries to evaluate the joint likelihood, the sample complexity of BBVI will be $\mathcal{O}\left(N^2\kappa^2\epsilon^{-1}\right)$, where κ also increases linearly with N for Bayesian posteriors as $\kappa = \mathcal{O}(N)$ due to posterior contraction. In terms of scalability with respect to N, this $\mathcal{O}(N^4)$ complexity highlights a clear challenge for BBVI with full-rank variational families. (Note that, with our alternative proof technique, this can be tightened to an $\mathcal{O}(N^3)$ explicit dependence.)

While mean-field is scalable, it is a crude approximation as it fails to model correlations in the posterior. Therefore, a natural question is, "Can we find a variational family more expressive than mean-field families while maintaining computational tractability?" In this paper, based on the theoretical framework of Kim et al. (2023a); Domke et al.

[*]Equal contribution [1]KAIST, Daejeon, South Korea, Republic of [2]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, U.S.A.. Correspondence to: Kyurae Kim <kyrkim@seas.upenn.edu>.

Proceedings of the 41st *International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

(2023), we answer this question by revisiting the idea of *structured scale matrices*.

Structured scale matrices, or more broadly, structured variational families, exploit structural assumptions about the factor graph of the posterior and have been a widely explored concept in VI (Saul & Jordan, 1995; Hoffman & Blei, 2015; Tan & Nott, 2018; Ranganath et al., 2016; Lin et al., 2019; for a short representative list; see Section 5 for a more extensive list). However, why and when structured families can outperform full-rank ones has not been rigorously analyzed. For instance, Tan & Nott, who first performed a detailed investigation into structured scales in BBVI, only mentioned that "unrestricted (full-rank) Gaussian variational approximation … can be prohibitively slow for large data since the number of variational parameters scales as the square of the length of $\boldsymbol{z}$." However, given the incredible success of stochastic optimization in the over-parameterized regime, simply having fewer parameters cannot fully explain why structured families work well.

Based on the framework of Domke et al. (2023); Kim et al. (2023a), we rigorously prove that, for likelihoods constituting of a finite-sum of N components, each associated to a single datapoint, structured location-scale families can improve the dimensional dependence of full-rank families. For Hierarchical Bayesian models with local variables, we provide a scale matrix structure that reduces the order of the dependence on the dataset size N. For the canonical 1-level hierarchy where the variables can be split into global and local variables, we show that a triangular scale matrix with a bordered block-diagonal structure achieves an iteration complexity of $\mathcal{O}(N)$. This, in turn, corresponds to approximating the posterior with the generative process

$$\boldsymbol{z} \sim q_\lambda(\boldsymbol{z}) \quad \text{and} \quad \boldsymbol{y}_n \sim q_\lambda(\boldsymbol{y}_n \mid \boldsymbol{z}),$$

where $\boldsymbol{z}$ are the global variables, $\boldsymbol{y}_n$ are the local variables belonging to the nth datapoint, and q_λ is some location-scale distribution. Agrawal & Domke (2021) called this, where the local variables are assumed to be conditionally independent, a "hierarchical branched distribution."

Overall, this work extends the current line of work of proving that the choice of the variational family performs a trade-off between statistical accuracy and computational efficiency (Bhatia et al., 2022; Kim et al., 2023b).

1. **General Analysis:** Theorem 2 establishes that, for finite-sum likelihoods, manipulating the structure of the scale matrix in location-scale variational families can improve the dimensional dependence of BBVI.
2. **Scalable Solution:** For hierarchical models with local variables, Corollary 2 establishes that the scale matrix structure previously proposed by Tan (2021); Tan et al. (2020) achieves an iteration complexity with a better dependence of the dataset size.
3. **Analysis of Parameterizations:** Furthermore, among different ways to parameterize structured variational families, Theorem 5 shows that "non-standardized" parameterizations rule out the convexity of the ELBO, unlike the "standardized" parameterization of Tan (2021); Tan et al. (2020), and is therefore suboptimal.

2. Preliminaries

Notation Random variables are denoted in sans-serif, vectors in bold, and matrices in bold capitals. (*i.e.,* x, $\boldsymbol{x}$, and $\boldsymbol{A}$ respectively.) Given a vector $\boldsymbol{x} \in \mathbb{R}^d$, we denote its Euclidean norm as $\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$. For a matrix $\boldsymbol{A}$, we denote its Frobenius norm as $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$, where $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^d A_{ii}$, and the ℓ_2-operator norm as $\|\boldsymbol{A}\|_{2,2}$. Lastly, $\mathbb{L}_{++}^d$ denotes the set of d-dimensional lower triangular matrices with strictly positive eigenvalues.

2.1. Variational Inference

Variational inference (VI; Jordan et al., 1999; Blei et al., 2017; Zhang et al., 2019) is a method for approximating an intractable probability distribution through optimization. In general, we aim to minimize the Kullback-Leibler (KL; Kullback & Leibler, 1951) divergence through:

$$\underset{\lambda \in \Lambda}{\text{minimize}} \;\; \mathrm{D}_{\mathrm{KL}}(q_\lambda, \pi),$$

where	D_{KL}	is the KL divergence,
	π	is the (target) posterior distribution, and
	q_λ	is the variational approximation.

Evidence Lower Bound In the context of Bayesian inference, we only have access to the joint likelihood $p(\boldsymbol{z}, \boldsymbol{x}) \propto \pi(\boldsymbol{z})$, where $\boldsymbol{z}$ are the model parameters and $\boldsymbol{x}$ is the data. This results in the KL divergence also being intractable due to the intractable normalizer. Therefore, we instead rely on minimizing a surrogate objective called the negative *evidence lower bound* (ELBO, Jordan et al., 1999) function $F(\lambda)$:

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

where	$p(\mathsf{z}, \boldsymbol{x})$	is the joint likelihood,
	$\mathbb{H}$	is the differential entropy.

Without loss of generality, we will also separately denote the negative log joint likelihood as

$$\ell(\mathsf{z}) \triangleq -\log p(\mathsf{z}, \boldsymbol{x}).$$

Under this notation, the ELBO can be represented as a *composite* optimization problem. (See the taxonomization in §2.4 by Kim et al. 2023a.)

$$F(\lambda) = f(\lambda) + h(\lambda),$$

where $f(\lambda) \triangleq \mathbb{E}_{\mathsf{z} \sim q_\lambda} \ell(\mathsf{z})$, is called the *energy* and $h(\lambda) \triangleq -\mathbb{H}(q_\lambda)$ is an entropic regularizer.

Finite-Sum Likelihood In this work, we will focus on log(-joint) likelihoods of the structure

$$\ell(\boldsymbol{z}) = \sum_{n=1}^{N} \ell_n(\boldsymbol{z}) \tag{1}$$

for $n = 1, \dots, N$, where each component likelihood ℓ_n. It is common for each component to use only subsets of the variables constituting $\boldsymbol{z}$.

2.2. Variational Family

We focus on the location-scale variational family:

Definition 1 (Location-Scale Family). Let φ be some d-variate distribution. Then, q_λ that can be equivalently represented as

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

where $\overset{d}{=}$ is equivalence in distribution, is said to be part of the location-scale family generated by the base distribution φ and the reparameterization function $\mathcal{T} : \Lambda \times \mathbb{R}^d \to \mathbb{R}^d$ defined as

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

with $\boldsymbol{\lambda} \in \Lambda \subseteq \mathbb{R}^p$ containing the parameters for forming the location $\boldsymbol{m} \in \mathbb{R}^d$ and scale $\boldsymbol{C} \in \mathbb{R}^{d \times d}$.

Since the reparameterization function is differentiable, this enables the use of the M-sample *reparameterization gradient* (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014), which is an unbiased estimator of the gradient of the energy $\nabla f(\boldsymbol{\lambda})$, defined as:

$$\boldsymbol{g}_M(\boldsymbol{\lambda}) \triangleq \frac{1}{M} \sum_{m=1}^{M} \nabla_\lambda \ell(\mathcal{T}_\lambda(\boldsymbol{u}_m)), \quad \boldsymbol{u}_1, \dots, \boldsymbol{u}_m \overset{i.i.d.}{\sim} \varphi. \tag{2}$$

The reparameterization gradient is empirically known to result in lower variance compared to other gradient estimators (Mohamed et al., 2020; Xu et al., 2019) and is presently *de-facto* standard for BBVI.

For the base distribution φ, the choice of distribution generates various other families (Titsias & Lázaro-Gredilla, 2014) used in practice: mean-field Gaussian, full-rank Gaussian, Laplace, Student-t, and many more. For our analysis, we impose mild assumptions on the base distribution φ:

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

These assumptions are satisfied by most practically used location-scale families. In summary:

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

2.3. Scale Parameterization

The parameterization of the scale matrix in location-scale families $\boldsymbol{C}$ can result in vastly different statistical and computational performance results. In principle, the only restriction is that it needs to result in a proper covariance matrix such that $\boldsymbol{C}\boldsymbol{C}^\top \succ 0$. However, how many low-rank factors we include into $\boldsymbol{C}$ corresponds to a statistical-computational trade-off (Bhatia et al., 2022; Ong et al., 2018). A common choice is to restrict $\boldsymbol{C}$ to be a diagonal matrix, resulting in the *mean-field* approximation (Peterson & Anderson, 1987; Hinton & van Camp, 1993).

Triangular Scale While mean-field results in fast convergence (Kim et al., 2023b) and stable optimization, it is a crude approximation as it ignores correlations in the posterior. Therefore, we are interested in scale matrices that are more complex than mean-field. For this, we will first restrict our interest to triangular scale matrices with strictly positive eigenvalues such that $\boldsymbol{C} \in \mathbb{L}_{++}^d$. Compared to other choices, such as the "matrix square-root" analyzed by Domke et al. (2023), this "Cholesky" parameterization has the following benefits:

(a) It results in lower gradient variance (Kim et al., 2023b),
(b) the entropy term h can be computed in $\Theta(d)$ time,
(c) the positive definiteness of $\boldsymbol{C}$ can be enforced only by manipulating the d diagonal elements, and
(d) the conditional dependence between the coordinates can easily be manipulated.

Naturally, Item (d) is essential in the context of structured variational families.

2.4. Stochastic Proximal Gradient Descent

While $F(\boldsymbol{\lambda})$ is commonly optimized using stochastic gradient descent (SGD; Robbins & Monro, 1951; Bottou, 1999; Nemirovski et al., 2009). In this work, we will focus on a "proximal" variant of SGD, which has favorable theoretical properties (Domke et al., 2023; Domke, 2020) compared to projected SGD, which is more commonly used in practice.

Proximal SGD Proximal SGD, or stochastic proximal gradient descent, aims to minimize a composite objective expressed as a sum $f + h$, where f is smooth and convex, while h might convex as well but non-smooth. Given initial parameters $\boldsymbol{\lambda}_0$ and a step size schedule $(\gamma_t)_{t=0}^{T-1}$, proximal SGD repeats the update:

$$\boldsymbol{\lambda}_{t+1} = \operatorname{prox}_{\gamma_t, h}(\boldsymbol{\lambda}_t - \gamma_t \boldsymbol{g}_M(\boldsymbol{\lambda}_t))$$

until convergence, where $\boldsymbol{g}_M(\boldsymbol{\lambda}_t)$ is a stochastic estimate of ∇f not ∇F, and prox is a *proximal operator* defined as:

$$\operatorname{prox}_{\gamma_t, h} \triangleq \arg\min_{\boldsymbol{v}} \left[h(\boldsymbol{v}) + \frac{1}{2\gamma_t} \|\boldsymbol{\lambda} - \boldsymbol{v}\|_2^2 \right].$$

The precise proximal operator we use is that by Domke (2020) and discussed in Appendix B.4.1.

Proximal SGD in BBVI Proximal SGD has been proposed for variational inference under different motiva-

tions (Altosaar et al., 2018; Khan et al., 2015; 2016; Diao et al., 2023). However, Domke (2020) first suggested proximal SGD to circumvent the fact that the ELBO is non-smooth due to the entropic regularizer h being non-smooth. While the non-smoothness of the ELBO can be solved through alternative means–*e.g.*, projected SGD, nonlinear parameterizations–projected SGD necessitate the choice of a restricted domain, and nonlinear parameterizations result in slower convergence rate (Kim et al., 2023a). Furthermore, projected and proximal SGD perform very similarly both in theory (Domke et al., 2023) and in practice. As such, while we focus on proximal SGD, most guarantees will also hold with projected SGD with the projection operator considered by Kim et al. (2024).

3. Theoretical Analysis

3.1. Fundamental Limits of Being Full-Rank

First, we will discuss what we currently know about the dimensional dependence of BBVI. We will also illustrate exactly *where* the dimensional dependence is coming from. This should make the solution more clear.

Iteration Complexity of Full-Rank Families Domke et al. (2023); Kim et al. (2023a) show that, with stochastic proximal gradient descent, full-rank variational families, and the 1-sample reparameterization gradient estimator, the iteration complexity is as follows:

Theorem 1 (Domke et al., 2023; Kim et al., 2023a)**.** *Let ℓ be μ-strongly convex and L-smooth. Then, the iteration complexity of being ϵ-close to the global minimizer with proximal SGD BBVI is*

$$\mathcal{O}\left(d\,\kappa^2\,\frac{1}{\epsilon}\log\left(\Delta_0^2\frac{1}{\epsilon}\right)\right),$$

where $\kappa = L/\mu$, $\Delta_0 = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2$ is the distance between the initial point $\boldsymbol{\lambda}_0$ and the global optimum $\boldsymbol{\lambda}^* = \arg\min_{\lambda\in\Lambda} F(\boldsymbol{\lambda})$.

Local Variables In probabilistic modeling, some Hierarchical models have latent variables, $\boldsymbol{y}_n$ for $n = 1, \dots, N$, unique to each datapoint $\boldsymbol{x}_n$ (Hoffman & Blei, 2015). For example, in mixture and mixed-membership models, each datapoint has a latent variable indicating from which mixture component it was generated. In state space models, at each time point t, the state is often not observed directly and is a latent variable local to t.

More abstractly, consider a model with local variables $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N$ and a global variable $\boldsymbol{z}$ such that $\boldsymbol{y}_n \in \mathbb{R}^{d_y}$ and $\boldsymbol{z} \in \mathbb{R}^{d_z}$. The total dimension of the model is

$$d = Nd_y + d_z.$$

Then, for these models, the complexity ends up with a dependence on the size of the dataset.

Corollary (Informal)**.** *For a Hierarchical Bayesian model with $\mathcal{O}(N)$ variables, a μ-strongly log-concave posterior, and L_n-smooth component likelihoods ℓ_n for $n = 1, \dots, N$ the iteration complexity is*

$$\mathcal{O}\left(N^3\,\frac{L_{\max}^2}{\mu^2}\,\frac{1}{\epsilon}\log\left(\Delta_0^2\frac{1}{\epsilon}\right)\right),$$

where $L_{\max} = \max\{L_1, \dots, L_n\}$.

For Bayesian models, the smoothness of the full posterior is $NL_{\max}$ since Eq. (1) is a sum, not an average as in empirical risk minimization. Therefore, the N^3 factor in the corollary follows from plugging in $\kappa = NL_{\max}/\mu$ and $d = Nd_y + d_z$ in Theorem 1.

Remark 1 (**Sample Complexity**)**.** While the cost of evaluating the gradient of the joint likelihood is at least $\Omega(N)$ datapoint queries, we will assume it is $\Theta(n)$ throughout the paper, as it is the most common case. Then, the sample complexity of BBVI with proximal SGD scales as $\mathcal{O}(N^4)$.

Clearly, full-rank families do not scale even if we ignore the $\Theta(N^2)$ storage requirement of the scale matrix.

Remark 2 (**Are fewer parameters *obviously* better?**)**.** The number of variational parameters p enters the complexity statement through Δ_0. Therefore, the dependence on the number of parameters alone is only logarithmic. This implies that simply reducing the number of parameters does not obviously improve the dimensional dependence.

Remark 3 (**Where does d come from?**)**.** The $\mathcal{O}(d)$ dimension dependence directly comes from gradient variance. (See Theorem 3 and 7 of Domke (2019).) Therefore, reducing the gradient variance is key to improving the complexity.

3.2. Complexity of BBVI on Finite-Sum Likelihoods

Before presenting our result, we generalize the notation we have introduced so far. Since we consider hierarchical models with local variables, each component ℓ_n will use a subset of the variables returned by $\mathcal{T}_\lambda$. We express this through $\mathcal{T}_\lambda^n : \mathbb{R}^d \to \mathbb{R}^{d_n}$, the parameterization function specific to ℓ_n, such that

$$\mathcal{T}_\lambda^n(\boldsymbol{u}) \triangleq \boldsymbol{C}_n\boldsymbol{u} + \boldsymbol{m}_n,$$

where $\boldsymbol{C}_n \in \mathbb{R}^{d\times d_n}$ is a subset of rows $\boldsymbol{C}$ and $\boldsymbol{m}_n \in \mathbb{R}^{d_n}$ is a subset of components of $\boldsymbol{m}$ corresponding to the coordinates of $\boldsymbol{z} = \mathcal{T}_\lambda(\boldsymbol{u})$ used by ℓ_n.

Furthermore, we are interested in the *structure* of $\boldsymbol{C}_n$ (and correspondingly $\boldsymbol{C}$). That is, whether the columns of $\boldsymbol{C}_n$ are zero or non-zero. For this, we introduce the indicator

$$\delta_{n,j} \triangleq \begin{cases} 1 & \text{if the } j\text{th column of } \boldsymbol{C}_n \text{ is non-zero} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $\delta_{n,j}$ is also implicitly encoding the structure of

ℓ_n: if ℓ_n uses a certain coordinate of $\boldsymbol{z}$, say $z_k = [\boldsymbol{z}]_k$, then $\delta_{n,k}$ *has* to be non-zero, since the original matrix $\boldsymbol{C}$ has a non-zero diagonal to qualify as a Cholesky factor. Therefore, for the case of a two-level hierarchical model, $\sum_j \delta_{n,j}$ is *at least* $d_y + d_z$, where "turning-on" additional rows allows for representing of additional correlations.

Upper Bound on the Gradient Variance We will see how this block structure affects the gradient variance.

Theorem 2. *Let ℓ_n be L_n-smooth for some $n = 1, \dots, N$ and Assumption 2 hold. Then, the gradient variance of $\boldsymbol{g}_M$ is bounded as*

$$\operatorname{tr}\mathbb{V}\,\boldsymbol{g}_M(\boldsymbol{\lambda}) \leq \frac{N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^{N} L_n^2\left(\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n\|_{\mathrm{F}}^2\right),$$

where $\bar{\boldsymbol{z}}_n$ is a stationary point of ℓ_n and

$$d^* \triangleq \max_n \sum_j \delta_{n,j}$$

is the effective dimensionality.

See the *full proof* in page 27.

Remark 4. When $\boldsymbol{C}$ is dense, such as in full-rank variational families, we have $d^* = d$. Therefore, we exactly retrieve Theorem 6 of Domke (2019).

Remark 5. $\delta_{n,j}$ is related to dimension dependence for the following reason: by setting the jth column of $\boldsymbol{C}_n$ to be *non-zero* ($\delta_{n,j} = 1$), we are effectively deciding to use u_j (the jth component of the d-dimensional vector $\boldsymbol{u}$) when computing $\mathcal{T}_{\boldsymbol{\lambda}}^n$. The exact number of components of $\boldsymbol{u}$ mixed by $\boldsymbol{C}_n$ is the "effective" dimension dependence d^*.

Remark 6. Since the gradient variance dominates the computational complexity, Theorem 2 answers Remark 2. That is, the *structure* of $\boldsymbol{C}_n$, which depends on ℓ_n, affects the complexity, not the number of parameters.

Notice that Theorem 2 is pointing towards a trivial way to improve the dimensional dependence of mean-field:

Corollary 1. *Let the posterior π factorize into independent univariate sub-posteriors such that $\pi(\boldsymbol{z}) = \prod_n^N \pi_n(z_n)$, where each π_n is L_n-log-smooth for $n = 1, \dots, N$. Then, the gradient variance of the mean-field approximation is dimension-independent.*

Remark 7. While Corollary 1 partially answers Conjecture 1 of Kim et al. (2023a), the general case for jointly-L-log-smooth posteriors remains open.

Improving the Dimension Dependence Admittedly, Corollary 1 is not very interesting since posteriors do not factorize as such for interesting problems. However, Theorem 2 does suggest that we can shape the resulting dimension dependence of realistic problems by designing the structure of $\boldsymbol{C}_n$. We will later show a specific structure for models with local variables that can improve the dependence on the number of datapoints.

Complexity of BBVI Now, based on Theorem 2, we can prove an iteration complexity bound on BBVI with proximal SGD (Section 2.4) and the reparameterization gradient (Section 2.2). The proof is based on the general results on proximal SGD by Gorbunov et al. (2020), recently refined by Garrigos & Gower (2023).

Theorem 3. *Let ℓ be μ-strongly convex and L-smooth, ℓ_n be L_n-smooth for $n = 1, \dots, N$, and Assumption 2 hold. Then, the last iterate $\boldsymbol{\lambda}_{T+1}$ of BBVI with proximal SGD and $\boldsymbol{g}_M$ is ϵ-close as $\mathbb{E}\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ to the global optimum $\boldsymbol{\lambda}^* = \arg\min F(\boldsymbol{\lambda})$ if*

$$T \geq \max\left(C_{\mathrm{var}}\frac{1}{\epsilon},\ C_{\mathrm{bias}}\right)\log\left(2\Delta_0^2\,\frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta_0 = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2$ is the distance to the optimum,

$$C_{\mathrm{var}} = 4\frac{N}{M}\left(d^* + k_\varphi\right)\textstyle\sum_{n=1}^N \kappa_n^2\left(\|\boldsymbol{m}_n^* - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n^*\|_{\mathrm{F}}^2\right)$$

$$C_{\mathrm{bias}} = 2\frac{N}{M}\left(d^* + k_\varphi\right)\textstyle\sum_{n=1}^N \kappa_n^2 + \kappa,$$

$\kappa_n = L_n/\mu$, $\kappa = L/\mu$ are the condition numbers, d^* is the effective dimensionality defined in Theorem 2, $\bar{\boldsymbol{z}}_n$ is a stationary point of ℓ_n, and $\boldsymbol{m}_n^*, \boldsymbol{C}_n^*$ are part of $\boldsymbol{\lambda}^*$.

See the *full proof* in page 30.

Remark 8. Since each evaluation of $\boldsymbol{g}_M$ takes $\Theta(NM)$ time, Theorem 3 implies a sample complexity of

$$\mathcal{O}\left(N^2\, d^* \textstyle\sum_{n=1}^N \kappa_n^2 \frac{1}{\epsilon}\log\frac{1}{\epsilon}\right),$$

or equivalently, a complexity of $\mathcal{O}\left(d^* N^3\right)$ after taking $\sum_n \kappa_n^2 \leq N \max_n \kappa_n^2$.

Remark 9. It also possible to prove a $\mathcal{O}(1/\epsilon)$ complexity using the decreasing stepsize schedules of Gower et al. (2019); Stich (2019).

3.3. Hierarchical Branched Structured Families

Equipped with the results of Section 3.2, we present a scale matrix structure that is more scalable in terms of dataset size dependence. Consider a canonical 2-level Hierarchical model with local variables $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N$ and a global variable $\boldsymbol{z}$ such that $\boldsymbol{y}_n \in \mathbb{R}^{d_y}$ and $\boldsymbol{z} \in \mathbb{R}^{d_z}$. (All local variables are assumed to have the same dimensionality.) Then,

$$\ell_n(\boldsymbol{y}_n, \boldsymbol{z}) = -\log p(\boldsymbol{x}_n, \boldsymbol{y}_n \mid \boldsymbol{z}) - \frac{1}{N}\log p(\boldsymbol{z}),$$

where $\boldsymbol{x}_n$ is the nth datapoint and $d_n = d_{\boldsymbol{z}} + d_{\boldsymbol{y}}$.

Structured Covariance Matrices We assume the variables are structured as

$$\begin{bmatrix}\boldsymbol{z}^\top & \boldsymbol{y}_1^\top & \dots & \boldsymbol{y}_n^\top & \dots & \boldsymbol{y}_N^\top\end{bmatrix}^\top, \tag{3}$$

with the noise vector $\boldsymbol{u}$ correspondingly structured as

$$\begin{bmatrix}\boldsymbol{u}_{\boldsymbol{z}}^\top & \boldsymbol{u}_{\boldsymbol{y}_1}^\top & \dots & \boldsymbol{u}_{\boldsymbol{y}_n}^\top & \dots & \boldsymbol{u}_{\boldsymbol{y}_N}^\top\end{bmatrix}^\top. \tag{4}$$

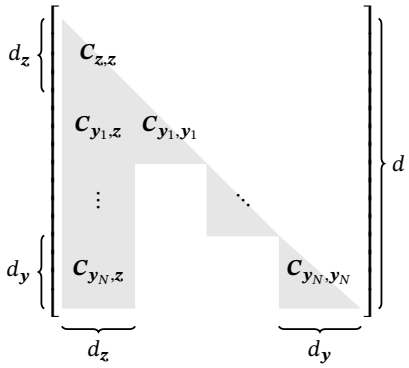


Figure 1: **Visualization of C under the proposed structure.** The colored entries are non-zero, while the white entries are filled with zeros.

Now for the $C_n \in \mathbb{R}^{(d_z+d_y)\times d}$, we propose the following structure:

$$C_n = \begin{bmatrix} C_{z,z} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ C_{y_n,z} & 0 & \dots & 0 & C_{y_n,y_n} & 0 & \dots & 0 \end{bmatrix},$$

where 0 is a zero block,
$C_{z,z}$ transforms u_z into z,
C_{y_n,y_n} transforms u_{y_n} into y_n, and
$C_{y_n,z}$ correlates u_z with y_n.

We are essentially assuming y_n and z are correlated, but $y_1, \dots, y_N$ are conditionally independent such that $(y_n \mid z) \perp\!\!\!\perp (y_m \mid z)$ for $n \neq m$. Agrawal & Domke (2021) called this a "hierarchical branched distribution" approximation. Also, combined with Eqs. (3) and (4), the full matrix C exhibits a *bordered block-diagonal structure* as visualized in Fig. 1.

Remark 10. The proposed structure has a space/storage complexity of $\Theta\left(\left(d_z d_y + d_y^2\right) N\right)$. This improves over full-rank, which has a storage complexity of $\Theta\left(\left(d_y + d_z N\right)^2\right)$.

Structured Variational Family Perspective By the property of triangular matrices, this implicitly forms a *structured variational family* of the form of

$$q(z, y_1, \dots, y_N) = q(z) \prod_{n=1}^N q(y_n \mid z).$$

Furthermore, when φ is chosen to be Gaussian, q becomes

$$q(z, y_1, \dots, y_N) = \mathcal{N}\left(z; m_z, C_{z,z} C_{z,z}^\top\right) \times \prod_{n=1}^N \mathcal{N}\left(y_n;\ m_{y_n} + C_{y_n,z} C_{z,z}^{-1}(z - m_z),\ C_{y_n,y_n} C_{y_n,y_n}^\top\right) \quad (5)$$

Note that we do not actually compute $C_{z,z}^{-1}$, and one can avoid it by directly accessing u_z. Also, Eq. (5) is identical to the structure proposed by Tan (2021, §3) and Tan et al. (2020). However, they did not consider the scalability aspect of this parameterization.

Standardized Parameterization Notice that, in Eq. (5), z is first "standardized" before interacting with y_n. (Again, this happens implicitly as we directly use u_z instead of explicitly standardizing z.) Under this parameterization, the existence of the "full" scale matrix C guarantees the ELBO to be "regular" according to Domke (2020); Titsias & Lázaro-Gredilla (2014); Challis & Barber (2013):

Theorem 4 (Theorem 9; Domke, 2020)**.** *Let Assumption 2 hold and the standardized parameterization be used. Then, if the posterior π is log-concave, then the ELBO is convex. If the posterior is also μ-strongly log-concave, the ELBO is then μ-strongly convex.*

Non-Standardized Parameterization On the other hand, it is also possible to *not* standardize z. This parameterization was considered by Agrawal & Domke (2021, Table 1):

$$q(z, y_1, \dots, y_N) = \mathcal{N}\left(z;\ m_z,\ C_{z,z} C_{z,z}^\top\right) \times \prod_{n=1}^N \mathcal{N}\left(y_n;\ m_{y_n} + A_n z,\ C_{y_n,y_n} C_{y_n,y_n}^\top\right), \quad (6)$$

where A_n is also an optimized parameter. The "expressiveness" of this parameterization is equivalent to Eq. (5). However, the loss-landscape is vastly different:

Theorem 5. *Let Assumption 2 hold. Then, under the non-standardized parameterization, there exists a strongly log-concave posterior for which the ELBO is not convex.*

See the *full proof* in page 31.

Remark 11. Even if the target posterior is strongly log-concave, the ELBO will fail to be strongly convex under the non-standardized parameterization. Therefore, convergence will be slower.

Overall, we have the following results:

Lemma 1. *The effective dimensionality d^* of the bordered block-diagonal scale matrix structure is*

$$d^* = d_z + d_y.$$

This leads to our key result:

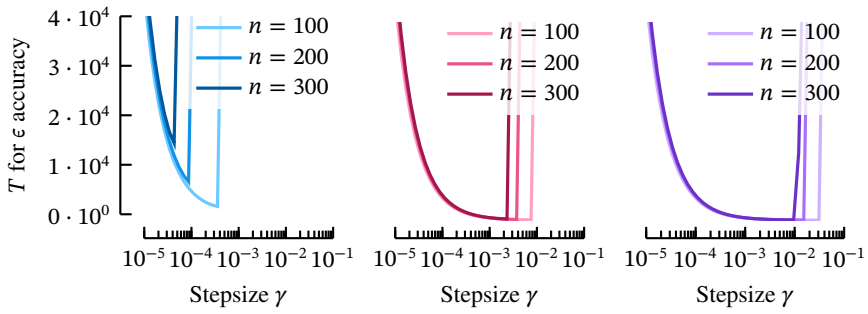
Corollary 2. *Let the assumptions of Theorem 3 hold and the structured variational family with a bordered block-diagonal structure matching that of $\ell_1, \dots, \ell_n$ be used. Then, the iteration complexity of finding an ϵ-accurate solution using BBVI with proximal SGD, g_M, and some fixed stepsize is*

$$\mathcal{O}\left(\frac{N}{M}\left(d_z + d_y + k_\varphi\right) \sum_{n=1}^N \kappa_n^2 \frac{1}{\epsilon} \log\left(2\Delta_0^2 \frac{1}{\epsilon}\right)\right).$$

Remark 12. Using the bordered block-diagonal scale matrix structure the sample complexity of BBVI is

$$\mathcal{O}\left(N^2 \sum_{n=1}^N \kappa_n^2 \frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right).$$

and equivalently $\mathcal{O}\left(N^3\right)$ after taking $\sum_{n=1}^N \kappa_n^2 \leq N \max_n \kappa_n^2$. This is a factor of N improvement from full-rank families (Remark 1).



(a) full-rank (b) structured (c) mean-field

Figure 2: **Number of iterations T required to obtain ϵ accuracy of variational families for a given stepsize γ.** structured behaves similarly to mean-field, while full-rank requires significantly more number of iterations, which also scales worse with respect to the number of datapoints n.

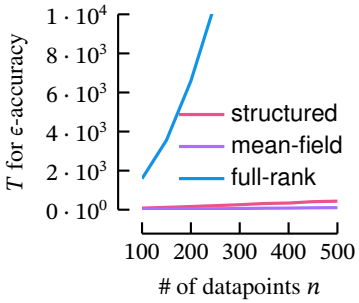


Figure 3: **Scaling of variational families with respect to the number of datapoints n.** full-rank exhibits a worst scaling than structured and mean-field.

4. Experiments

We now empirically evaluate our theoretical analysis in Section 3. Mainly, we will compare the scalability of mean-field, full-rank, and the structured variational family described in Section 3.3.

4.1. Synthetic Experiments

4.1.1. Setup

To quantitatively verify the theoretical results in Section 3, we use proximal SGD with the proximal operator described in Appendix B.4.1 to match the theory. For the target distribution, we use an isotropic Gaussian target distribution of $\ell_n(\boldsymbol{y}_n, \boldsymbol{z}) = \log \mathcal{N}\left(\boldsymbol{y}_n; 5\mathbf{1}_{d_y}, 0.1\mathbf{I}_{d_y}\right) + \log \mathcal{N}\left(\boldsymbol{z}_n; 5\mathbf{1}_{d_z}, 0.1\mathbf{I}_{d_z}\right)$ (mean 5 and variance 0.1) where we set $d_z = 5$, $d_y = 3$, and vary the "number of datapoints" n. All variational families are initialized with a standard Gaussian. We then run BBVI with $M = 8$ Monte Carlo stepsizes, 50 different stepsizes in the interval of $[10^{-6}, 1]$, and estimate the sequence of expected distance to the optimum $(r_1)_{t \geq 0}$, where $r_t \triangleq \mathbb{E}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2$. We then estimate the minimum number of iterations T required to hit ϵ accuracy such that $r_{T-1} > \epsilon$ and $r_T \leq \epsilon$. We set $\epsilon = 1$ in all cases.

4.1.2. Results

Effect of Stepsize We demonstrate the effect of stepsize on the number of iterations required to achieve ϵ-accuracy in Fig. 2. Under our setup, the distance to the optimum scales as $\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 = \Theta(n)$. Therefore, all methods show an increase in the minimum T as n increases. However, the T required by full-rank appears to increase much faster than structured and mean-field. This is because, as predicted in Section 3, the variance of full-rank also increases in n, forcing the use of a smaller stepsize γ. Clearly, the range of stepsizes full-rank achieves the ϵ accuracy threshold is much narrower and shrinks faster as n increases. In sharp contrast, structured behaves very similarly to mean-field.

Scaling w.r.t. N Now, by picking the stepsize that achieves the lowest T for hitting the ϵ accuracy target for each curve in Fig. 2, we can directly evaluate the iteration

Table 1: Models and Datasets used in the Experiments

Problem	Dimensions				# of Variational Parameters (p)		
	N	d_y	d_z	d	mean-field	structured	full-rank
rpoisson-small	1,961			1,977	3,954	35,450	1,957,230
rpoisson-middle	3,922	1	16	3,938	7,876	70,748	7,759,829
rpoisson-large	19,609			19,625	39,282	353,114	
volatility-small	262			1,605	3,210	59,544	1,290,420
volatility-middle	522	6	33	3,165	6,210	118,044	4,825,170
volatility-large	2,579			15,507	31,014	580,869	
irt-small	3,348			3,541	7,082	671,774	6,274,652
irt-middle	6,695	1	193	6,888	13,776	1,324,439	23,732,604
irt-large	33,475			33,668	67,336	6,546,539	

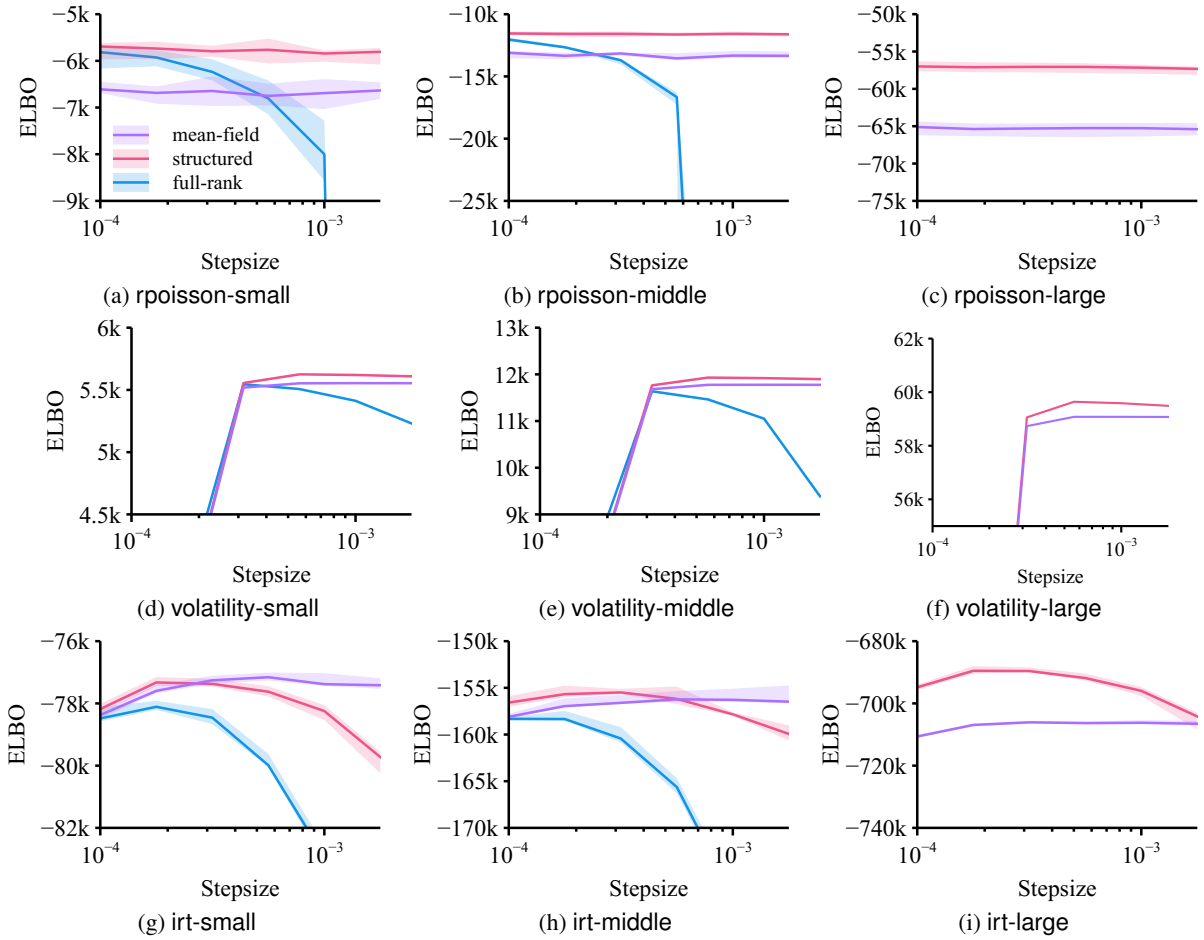


Figure 4: **ELBO at $T = 5 \times 10^4$ versus the optimizer stepsize (γ) on the considered problems with varying dataset sizes.** The solid lines are the median over 8 independent replications, while the colored bands mark the 80% empirical percentiles.

complexity bounds in Section 3. For this, we examine the scaling with respect to n, which is shown in Fig. 3. We can see that full-rank scales much worse than both mean-field and structured. In fact, full-rank exhibits a quadratic scaling while mean-field and structured exhibit a linear scaling. This confirms our theory Section 3 that structured is able to shave off a polynomial order in iteration complexity. Now, one may notice that the scaling observed in Section 3 are all a polynomial order better than what the theory in Section 3 predicts ($\mathcal{O}(N^3)$ for full-rank and $\mathcal{O}(N^2)$ for structured when taking $\sum_{n=1}^{N} \kappa_n = \Theta(N)$). This is because our target posterior is an isotropic Gaussian; there are no correlations between the gradient estimator of each component. Therefore, our theory is a polynomial-order pessimistic due to the use of Lemma 4 in Appendix B.2.

Note that, in principle, since the dimensionality of the posterior is increasing, ϵ needs to be increased at a rate of n so that we allocate an equal "accuracy" budget to each coordinate. Therefore, considering the fact that we fix ϵ irrespective of n, the linear scaling of T w.r.t. to n is actually benign.

4.2. Realistic Experiments

We will now qualitatively verify our theory on more complex models and real-world datasets. For a fixed budget of gradient evaluations, we will compare the ELBO value obtained at the end of BBVI for the three variational families. Recall that, in SGD, the stepsize γ performs a trade-off between the accuracy of the final solution versus the convergence speed. We will demonstrate that, as the theory predicts, structured provide a more favorable trade-off compared to full-rank and mean-field, especially in the large data regime (large n).

4.2.1. Setup

Stochastic Optimization While our theoretical results use proximal SGD, we use regular SGD without projection for our experiments. In practice, as long as the initialization is sensible, proximal SGD does not provide additional benefits both in practice (Kim et al., 2023a) and in theory (Domke et al., 2023), and simple SGD is the most common method for performing BBVI in practice. Furthermore, we will consider only fixed stepsizes. In practice, step-decay stepsize

schedules (Goffin, 1977) are believed to be superior to polynomial decay schedules or naive fixed stepsizes. However, the theoretical evidence for this is not yet clear. (See Wang et al. 2021; Ge et al. 2019 for related investigations.) Also, for the sake of experimentation, step-decay stepsizes are hard to control due to the additional number of configurations, such as the number of stages, amount of decay per stage, and such.

Implementation We implemented our experiments using the Turing ecosystem (Ge et al., 2018) in the Julia language (Bezanson et al., 2017). The structured covariances were implemented using the compressed sparse column (CSC) sparse matrix interface provided by the `CUDA.jl` library (Besard et al., 2019), while the sparse derivatives were implemented using the `Zygote.jl` framework (Innes, 2018). We use 8 Monte Carlo samples and the Adam optimizer (Kingma & Ba, 2015) for all problems, while the reported ELBOs are estimated using 1024 Monte Carlo samples every 100 iterations. The variational families are Gaussian such that $\varphi = \mathcal{N}(0, 1)$.

Models and Dataset We use three different Hierarchical Bayesian models, volatility, rpoisson, and irt, which have local variables. rpoisson is a robust generalized linear regression model known as the Poisson-log-normal model (Cameron & Trivedi, 2013, §4.2.4). We treat the regression coefficients and the hyperparameters as global variables, while the individual-level response, which is modeled to include log-normal noise, is treated as the local variable. volatility is a multivariate stochastic volatility model (Chib et al., 2009; Naesseth et al., 2018). We treat the hyperparameters as the global variables, while the latent volatilities are treated as the local variables. Lastly, irt is a two-parameter logistic item response theory (IRT) model (Lord et al., 2008; Wu et al., 2020). We treat the "ability" of each student as local variables, while the hyperparameters and item-level variables are treated as global variables. The full description of these models can be found in Appendix C. To evaluate the effect of dataset size, we use subsets of the full datasets, as shown in Table 1. For the initial point, we use $q_{\lambda_0} = \mathcal{N}\left(\mathbf{0}, 10^{-2}\mathbf{I}\right)$ for all experiments.

4.2.2. Results

The effect of the stepsize γ are shown in Fig. 4. Additional results can be found in Appendix D. Notice that full-rank is much more sensitive to the stepsize compared to structured and mean-field. This exactly aligns with the theory: larger gradient variance means that the size of the stationary region of SGD is wider. Therefore, the quality of the solution is much more sensitive to the stepsize. Since full-rank has the largest amount of gradient variance, followed by structured, and then mean-field, the sensitivity to the stepsize follows the same ordering.

Overall, our experimental results demonstrate that, for fixed stepsize SGD, the complexity of the variational family trades optimization speed for the statistical accuracy of the variational approximation. This re-affirms the results of Bhatia et al. (2022) on a more realistic setup.

5. Discussions

Conclusions In this work, we have theoretically investigated the limitations of full-rank variational families in BBVI. Specifically, the dimensional scaling of the gradient variance of full-rank variational families. This is particularly problematic for Bayesian models with local variables, implying that BBVI with full-rank variational families do not scale to larger datasets. Fortunately, we have rigorously shown that variational families with structured scale matrices are able to improve this scaling issue. We have evaluated this theoretical insight on large-scale Hierarchical Bayesian models and have confirmed that practice agrees with the theory. Furthermore, our analysis provides a precise quantitative analysis of how certain scale matrix structures would improve the computational complexity of BBVI.

Related Works Structured variational approximations have a long history since their use in hidden Markov models (Saul & Jordan, 1995; Ghahramani & Jordan, 1997), CAVI (Hoffman & Blei, 2015; Mimno et al., 2012; Rohde & W, 2016), fixed-form variational Bayes (Salimans & Knowles, 2013), BBVI (Ong et al., 2018; Quiroz et al., 2023; Ranganath et al., 2016; Tan, 2021; Tan et al., 2020), natural gradient VI (Lin et al., 2019), and now modern approaches such as amortized VI (Archer et al., 2015; Johnson et al., 2016; Webb et al., 2018; Sø nderby et al., 2016; Agrawal & Domke, 2021; Maaløe et al., 2016; Gao et al., 2016; Yu et al., 2022) and normalizing flows (Ambrogioni et al., 2021b;a). Our proposed covariance structure is identical to that of Tan (2021); Tan et al. (2020). But, they did not consider the computational complexity of this parameterization, for which we rigorously prove its scalability with respect to N.

Limitations of the Current Theory In the $\mathcal{O}(N^3)$ sample complexity we obtained (after making the implicit N dependence in $\sum_n \kappa_n = \mathcal{O}(N)$ explicit), an excess factor of N comes from the fact that the gradient variance analysis strategy of Domke (2019) results in an $\mathcal{O}\left(L^2\right)$ dependence on the smoothness constant $L = \mathcal{O}(N)$. We conjecture that an $\mathcal{O}(L)$ dependence is realistic, which would match that of empirical risk minimization and also imply a dataset size dependence of $\mathcal{O}(N^2)$ for hierarchical models. This also matches the following intuition: computing the gradient costs $\Theta(N)$, while, by posterior contraction, the smoothness naturally scales as $\mathcal{O}(N)$. Therefore, we conjecture a $\mathcal{O}(N^2)$ complexity. Note that this is also closely related to the fact that the complexity of BBVI currently has a $\mathcal{O}\left(\kappa^2\right)$ dependence on κ, which is believed to be loose.

Acknowledgements

J. Ko and W. Kim were supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (NRF-2022M3J6A1063021 and RS-2023-00208980); K. Kim was supported by a gift from AWS AI to Penn Engineering's ASSET Center for Trustworthy AI; J. R. Gardner was supported by NSF award [IIS-2145644].

Impact Statement

This paper presents a theoretical analysis of black-box variational inference with the goal of broadening our understanding of the algorithm and potentially contributing to its improvement. As such, the potential societal consequences of our work are inherited from those of Bayesian inference and general probabilistic modeling. However, the work itself is theoretical, and we do not expect direct societal consequences.

References

Agrawal, A. and Domke, J. Amortized variational inference for simple hierarchical models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21388–21399. Curran Associates, Inc., 2021. (pages 2, 6, 9)

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

Altosaar, J., Ranganath, R., and Blei, D. Proximity variational inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 84 of *PMLR*, pp. 1961–1969. JMLR, March 2018. (page 4)

Ambrogioni, L., Lin, K., Fertig, E., Vikram, S., Hinne, M., Moore, D., and van Gerven, M. Automatic structured variational inference. In *Proceedings of The International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 676–684. JMLR, March 2021a. (page 9)

Ambrogioni, L., Silvestri, G., and van Gerven, M. Automatic variational inference with cascading flows. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 254–263. JMLR, July 2021b. (page 9)

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. Black box variational inference for state space models. *International Conference on Learning Representations Workshops*, November 2015. (page 9)

Besard, T., Foket, C., and De Sutter, B. Effective extensible programming: Unleashing julia on GPUs. *IEEE Transactions on Parallel and Distributed Systems*, 30(4): 827–841, 2019. (page 9)

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. (page 9)

Bhatia, K., Kuang, N. L., Ma, Y.-A., and Wang, Y. Statistical and computational trade-offs in variational inference: A case study in inferential model selection. arXiv Preprint arXiv:2207.11208, arXiv, July 2022. (pages 2, 3, 9)

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. (pages 1, 2)

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

Cameron, A. C. and Trivedi, P. K. *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press, Cambridge, 2 edition, 2013. (pages 9, 32)

Challis, E. and Barber, D. Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, 14(68):2239–2286, 2013. (page 6)

Chib, S., Omori, Y., and Asai, M. Multivariate stochastic volatility. In *Handbook of Financial Time Series*, pp. 365–400. Springer, Berlin, Heidelberg, 2009. (pages 9, 33)

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The Helmholtz machine. *Neural Computation*, 7(5):889–904, September 1995. (page 32)

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 7960–7991. JMLR, July 2023. (page 4)

Domke, J. Provable gradient variance guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 32, pp. 329–338. Curran Associates, Inc., 2019. (pages 4, 5, 9, 16, 23, 27, 28)

Domke, J. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 2587–2596. JMLR, July 2020. (pages 3, 4, 6, 22, 28)

Domke, J., Gower, R., and Garrigos, G. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66289–66327. Curran Associates, Inc., 2023. (pages 1, 2, 3, 4, 8, 22)

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. (page 19)

Gao, Y., Archer, E. W., Paninski, L., and Cunningham, J. P. Linear dynamical neural population models through nonlinear embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. (page 9)

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. Preprint arXiv:2301.11235, arXiv, February 2023. (pages 5, 20, 21)

Ge, H., Xu, K., and Ghahramani, Z. Turing: A language for flexible probabilistic inference. In *Proceedings of the International Conference on Machine Learning*, volume 84 of *PMLR*, pp. 1682–1690. JMLR, 2018. (page 9)

Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. (page 9)

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. Bayesian workflow. *arXiv:2011.01808 [stat]*, November 2020. (page 33)

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1): 267–305, January 2016. (page 19)

Ghahramani, Z. and Jordan, M. I. Factorial hidden Markov models. *Machine Learning*, 29(2):245–273, November 1997. (page 9)

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

Goffin, J. L. On convergence rates of subgradient optimization methods. *Mathematical Programming*, 13(1): 329–347, December 1977. (page 9)

Gorbunov, E., Hanzely, F., and Richtarik, P. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pp. 680–690. JMLR, June 2020. (pages 5, 20)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5200–5209. JMLR, June 2019. (page 5)

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Mathematical Programming*, 188(1):135–192, July 2021. (page 19)

Hilbe, J. M. *Negative Binomial Regression*. Cambridge University Press, Cambridge, 2 edition, 2011. (page 32)

Hilbe, J. M. *COUNT: Functions, Data and Code for Count Data*, 2016. (page 32)

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Annual Conference on Computational Learning Theory*, pp. 5–13, Santa Cruz, California, United States, 1993. ACM Press. (pages 1, 3)

Hoffman, M. and Blei, D. Stochastic structured variational inference. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *PMLR*, pp. 361–369. JMLR, February 2015. (pages 1, 2, 4, 9)

Innes, M. Don't unroll adjoint: Differentiating SSA-Form programs. *CoRR*, abs/1810.07951, 2018. (page 9)

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. (page 9)

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

Khaled, A. and Richtárik, P. Better theory for SGD in the nonconvex world. *Transactions of Machine Learning Research*, 2023. (page 19)

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 319–328, Arlington, Virginia, USA, 2016. AUAI Press. (page 4)

Khan, M. E. E., Baque, P., Fleuret, F., and Fua, P. Kullback-Leibler proximal variational inference. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3402–3410. Curran Associates, Inc., 2015. (page 4)

Kim, K., Oh, J., Wu, K., Ma, Y., and Gardner, J. R. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44615–44657, New Orleans, LA, USA, December 2023a. Curran Associates Inc. (pages 1, 2, 4, 5, 8, 28, 29)

Kim, K., Wu, K., Oh, J., and Gardner, J. R. Practical and matching gradient variance bounds for black-box variational Bayesian inference. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 16853–16876, Honolulu, HI, USA, July 2023b. JMLR. (pages 1, 2, 3, 28)

Kim, K., Ma, Y., and Gardner, J. R. Linear convergence of black-box variational inference: Should we stick the landing? In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR*, pp. 235–243, Valencia, Spain, 2024. JMLR. (page 4)

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*, San Diego, California, USA, 2015. (page 9)

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, Banff, AB, Canada, April 2014. (pages 3, 32)

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

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

Lin, W., Khan, M. E., and Schmidt, M. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 3992–4002. JMLR, May 2019. (pages 2, 9)

Lord, F. M., Novick, M. R., and Birnbaum, A. *Statistical Theories of Mental Test Scores*. The Addison-Wesley Series in Behavioral Science: Quantitative Methods. Information Age Publ, Charlotte, NC, nachdr. der ausg. reading, mass. [u.a.], 1968 edition, 2008. (pages 9, 32)

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary deep generative models. In *Proceedings of The International Conference on Machine Learning*, volume 48 of *PMLR*, pp. 1445–1453. JMLR, June 2016. (page 9)

Mimno, D., Hoffman, M. D., and Blei, D. M. Sparse stochastic inference for latent Dirichlet allocation. In *Proceedings of the International Conference on Machine Learning*, ICML'12, pp. 1515–1522, Madison, WI, USA, 2012. Omnipress. (page 9)

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

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pp. 451–459. Curran Associates, Inc., 2011. (page 19)

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational sequential Monte Carlo. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 84 of *PMLR*, pp. 968–977. JMLR, March 2018. (pages 9, 33)

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009. (page 3)

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! Convergence without the bounded gradients assumption. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 3750–3758. JMLR, July 2018. (pages 19, 20)

Ong, V. M.-H., Nott, D. J., and Smith, M. S. Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, 27 (3):465–478, July 2018. (pages 3, 9)

Patz, R. J. and Junker, B. W. A straightforward approach to Markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2):146–178, June 1999. (page 32)

Peterson, C. and Anderson, J. R. A mean field theory learning algorithm for Neural Networks. *Complex Systems*, 1 (5):995–1019, 1987. (pages 1, 3)

Quiroz, M., Nott, D. J., and Kohn, R. Gaussian variational approximations for high-dimensional state space models. *Bayesian Analysis*, 18(3):989–1016, September 2023. (page 9)

Rammus. Upper bound for the trace of the product of a matrix and a positive matrix, December 2021. (page 17)

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

Ranganath, R., Tran, D., and Blei, D. Hierarchical variational models. In *Proceedings of the International Conference on Machine Learning*, volume 48 of *PMLR*, pp. 324–333. JMLR, June 2016. (pages 2, 9)

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

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

Rohde, D. and W, M. P. Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research*, 17(172):1–47, 2016. (page 9)

Salimans, T. and Knowles, D. A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, December 2013. (page 9)

Saul, L. and Jordan, M. Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. (pages 2, 9)

Sø nderby, C. K., Raiko, T., Maalø e, L., Sø nderby, S. r. K., and Winther, O. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. (page 9)

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method. Preprint arXiv:1907.04232, arXiv, December 2019. (page 5)

Tan, L. S. L. Use of model reparametrization to improve variational Bayes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):30–57, February 2021. (pages 2, 6, 9)

Tan, L. S. L. and Nott, D. J. Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28(2):259–275, March 2018. (page 2)

Tan, L. S. L., Bhaskaran, A., and Nott, D. J. Conditionally structured variational Gaussian approximation with importance weights. *Statistics and Computing*, 30(5): 1255–1272, September 2020. (pages 2, 6, 9)

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the International Conference on Machine Learning*, volume 32 of *PMLR*, pp. 1971–1979. JMLR, June 2014. (pages 1, 3, 6)

Wang, C. and Blei, D. M. A general method for robust Bayesian modeling. *Bayesian Analysis*, 13(4):1163–1191, December 2018. (page 32)

Wang, X., Magnússon, S., and Johansson, M. On the convergence of step decay step-size for stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14226–14238. Curran Associates, Inc., 2021. (page 9)

Webb, S., Golinski, A., Zinkov, R., N, S., Rainforth, T., Teh, Y. W., and Wood, F. Faithful inversion of generative models for effective amortized inference. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. (page 9)

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., and Goodman, N. Variational Item Response Theory: Fast, Accurate, and Expressive. In *Proceedings of the International Conference on Educational Data Mining*, pp. 257–268, virtual, July 2020. International Educational Data Mining Society. (pages 9, 32)

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

Yu, C., Soulat, H., Burgess, N., and Sahani, M. Structured recognition for generative models with explaining away. *Advances in Neural Information Processing Systems*, 35: 40–53, December 2022. (page 9)

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, August 2019. (pages 1, 2)

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

TABLE OF CONTENTS

1 Introduction **1**

2 Preliminaries **2**
2.1 Variational Inference . . . 2
2.2 Variational Family . . . 3
2.3 Scale Parameterization . . . 3
2.4 Stochastic Proximal Gradient Descent . . . 3

3 Theoretical Analysis **4**
3.1 Fundamental Limits of Being Full-Rank . . . 4
3.2 Complexity of BBVI on Finite-Sum Likelihoods . . . 4
3.3 Hierarchical Branched Structured Families . . . 5

4 Experiments **7**
4.1 Synthetic Experiments . . . 7
4.1.1 Setup . . . 7
4.1.2 Results . . . 7
4.2 Realistic Experiments . . . 8
4.2.1 Setup . . . 8
4.2.2 Results . . . 9

5 Discussions **9**

A Computational Resources **15**

B Proofs **15**
B.1 Definitions . . . 15
B.2 Auxiliary Lemmas . . . 16
B.3 Convergence of Stochastic Proximal Gradient Descent . . . 19
B.3.1 Overview . . . 19
B.3.2 Technical Assumptions . . . 19
B.3.3 Convergence (Lemma 8) . . . 20
B.3.4 Complexity (Corollary 3) . . . 21
B.4 Convergence of Black-Box Variational Inference of Finite-Sum Likelihoods . . . 22
B.4.1 Overview . . . 22
B.4.2 Key Lemmas . . . 23
B.4.3 Gradient Variance Bound (Theorem 2) . . . 27
B.4.4 Convex Expected Smoothness (Lemma 12) . . . 28
B.4.5 Complexity with General Location-Scale Families (Theorem 3) . . . 30
B.5 Properties of the Non-Standardized Parameterization (Theorem 5) . . . 31

C Probabilistic Models and Datasets **32**
C.1 Robust Poisson Regression . . . 32
C.2 Item Response Theory . . . 32
C.3 Multivariate Stochastic Volatility . . . 33

D Additional Experimental Results **34**
D.1 Results on rpoisson . . . 34
D.2 Results on volatility . . . 35
D.3 Results on irt . . . 36

A. Computational Resources

Table 2: Computational Resources

Type	Model and Specifications
System Topology	1 socket with 8 physical cores
Processor	1 Intel i9-11900F, 2.5 GHz (maximum 5.2 GHz) per socket
Cache	80 KB L1, 512 KB L2, and 16 MB L3
Memory	64 GiB RAM
Accelerator	1 NVIDIA GeForce RTX 3090 per node, 1.7 GHZ, 24GiB RAM

All of the experiments took approximately 1 week to run.

B. Proofs

B.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 $L > 0$.

Remark 13. Equivalently, we say a function f is L-log-smooth if $\log f$ is L-smooth.

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 $\mu > 0$.

Remark 14. Equivalently, we say a function f is only convex if it satisfies the strong convexity inequality with $\mu = 0$.

Remark 15 (**Log-Concave Measures**)**.** For a probability measure Π, we say it is μ-strongly log-concave if, in a d-dimensional Euclidean measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$, where $\mathcal{B}(\mathbb{R}^d)$ is the σ-algebra of Borel-measurable subsets of $\mathbb{R}^d$ and $\mathbb{P}$ is the Lebesgue measure, its log probability density function $x \mapsto -\log \pi(x)$ is μ-strongly convex.

Definition (**Bregman Divergence**)**.** Let $\mathcal{X}$ be a convex set. Then, we define the function

$$\mathrm{D}_\phi(\boldsymbol{x}, \boldsymbol{x}') \triangleq \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \langle \nabla\phi(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle$$

to be the Bregman divergence generated by some continuously differentiable function $\phi : \mathcal{X} \to \mathbb{R}$ convex on $\mathcal{X}$.

Definition 2 (**Subspace Identity Matrix of ℓ_n**)**.** We define $\mathbf{I}_n \in \mathbb{R}^{d \times d}$, where the entries are set as

$$[\mathbf{I}_n]_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and the } i\text{th element of } \boldsymbol{z} \text{ is used by } \ell_n \\ 0 & \text{otherwise,} \end{cases}$$

where $[\cdot]_{ij}$ denotes the ith row and jth column, which is a variation of the standard identity matrix.

Definition 2 can also be understood as a "masking matrix." That is, $\mathbf{I}_n \boldsymbol{z}$ effectively "selects" the entries of $\boldsymbol{z}$ used by ℓ_n by setting the remaining entries to 0.

B.2. Auxiliary Lemmas

Lemma 2. *Let $\boldsymbol{u} = (u_1, u_2, \dots, u_d)$ be a d-dimensional vector-valued random variable satisfying Assumption 1. Then,*

$$1.\ \mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top = \mathbf{I} \qquad \text{and} \qquad 2.\ \mathbb{E}u_i^2\,\boldsymbol{u} = \mathbf{0}$$

for any $i = 1, \dots, d$.

Proof. The first identity is derived by Domke (2019, Lemma 9).

$$\mathbb{E}u_i^2\,\boldsymbol{u} = \mathbb{E}\begin{bmatrix} u_1 & \dots & u_{i-1} & u_i^3 & u_{i+1} & \dots & u_d \end{bmatrix}^\top = \begin{bmatrix} \mathbb{E}u_1 & \dots & \mathbb{E}u_{i-1} & \mathbb{E}u_i^3 & \mathbb{E}u_{i+1} & \dots & \mathbb{E}u_d \end{bmatrix}^\top = \mathbf{0},$$

where the last equality follows from Assumption 1. □

Lemma 3. *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be some matrix and $\boldsymbol{u}$ be a d-dimensional vector-valued random variable satisfying Assumption 1. Then,*

$$\mathbb{E}\|\boldsymbol{A}\boldsymbol{u}\|_2^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{A}\boldsymbol{u}\|_2^2 &= \mathbb{E}\,\mathrm{tr}\,\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{u} && \text{(quadratic form is equal to its trace)}\\
&= \mathrm{tr}\,\boldsymbol{A}^\top\boldsymbol{A}\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top && \text{(cyclic property of trace)}\\
&= \mathrm{tr}\,\boldsymbol{A}^\top\boldsymbol{A} && \text{(Assumption 1)}\\
&= \|\boldsymbol{A}\|_{\mathrm{F}}^2 && \text{(definition of Frobenius norm).}
\end{aligned}$$

□

Lemma 4. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ be vector-variate random variables. Then, the variance of the sum is upper-bounded as*

$$\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^N \boldsymbol{x}_i\right] \leq N\sum_{i=1}^N \mathrm{tr}\mathbb{V}\,\boldsymbol{x}_i.$$

Proof.

$$\begin{aligned}
\mathrm{tr}\mathbb{V}\left[\sum_{i=1}^N \boldsymbol{x}_i\right] &= \mathbb{E}\left\|\sum_{i=1}^N(\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i)\right\|_2^2 = \mathbb{E}\sum_{i=1}^N\sum_{j=1}^N(\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i)^\top(\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j)\\
&\leq \sum_{i=1}^N\sum_{j=1}^N \frac{1}{2}\left(\mathbb{E}\|\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i\|_2^2 + \mathbb{E}\|\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j\|_2^2\right) && \text{(Young's inequality)}\\
&= \frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N \mathbb{E}\|\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i\|_2^2 + \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \mathbb{E}\|\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j\|_2^2 && \text{(Distributing the sums)}\\
&= \frac{N}{2}\sum_{i=1}^N \mathbb{E}\|\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i\|_2^2 + \frac{N}{2}\sum_{j=1}^N \mathbb{E}\|\boldsymbol{x}_j - \mathbb{E}\boldsymbol{x}_j\|_2^2 && \text{(Solving the sums)}\\
&= N\sum_{i=1}^N \mathbb{E}\|\boldsymbol{x}_i - \mathbb{E}\boldsymbol{x}_i\|_2^2\\
&= N\sum_{i=1}^N \mathrm{tr}\mathbb{V}\,\boldsymbol{x}_i.
\end{aligned}$$

□

Lemma 5. *Let $\boldsymbol{x}$ be some d-dimensional vector-variate random variables. Then, the variance is upper-bounded as*

$$\mathrm{tr}\mathbb{V}\boldsymbol{x} \leq \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

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

Proof.

$$\begin{aligned}
\mathrm{tr}\mathbb{V}\boldsymbol{x} &= \mathbb{E}\|\boldsymbol{x} - \mathbb{E}\boldsymbol{x}\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y} + \boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + 2\mathbb{E}\left\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{y} - \mathbb{E}\boldsymbol{x}\right\rangle + \|\boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|_2^2 \quad \text{(expanding the quadratic)} \\
&= \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - 2\left\langle \boldsymbol{y} - \mathbb{E}\boldsymbol{x}, \boldsymbol{y} - \mathbb{E}\boldsymbol{x}\right\rangle + \|\boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|_2^2 \quad \text{(linearity of expectation)} \\
&= \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - 2\|\boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|_2^2 + \|\boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|2^2 \\
&= \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{y} - \mathbb{E}\boldsymbol{x}\|2^2 \\
&\leq \mathbb{E}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2.
\end{aligned}$$

□

Lemma 6 (Rammus, 2021)**.** *Consider $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, where $\boldsymbol{B}$ is positive semidefinite such that $\boldsymbol{B} \succ 0$. Then,*

$$\mathrm{tr}\left(\boldsymbol{AB}\right) \leq \|\boldsymbol{A}\|_{2,2}\mathrm{tr}\left(\boldsymbol{B}\right).$$

Proof. We restate the proof of Rammus (2021) for completeness.

Consider an inner product space of matrices, where, for $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{C}^{d\times d}$, its inner product is defined as $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \mathrm{tr}\left(\boldsymbol{X}^*\boldsymbol{Y}\right)$. Here, $\boldsymbol{X}^*$ is the conjugate transpose of $\boldsymbol{X}$. This generates a p-norm as

$$\|\boldsymbol{X}\|_p \triangleq \left(\sum_i \sigma_i\left(\boldsymbol{X}\right)^p\right)^{1/p},$$

where $\sigma_i\left(\boldsymbol{X}\right)$ is the ith singular value of $\boldsymbol{X}$. This norm is known as the Schatten norm, where its limiting case $p \to \infty$ is, in fact, the ℓ_2 operator norm.

We can now apply Hölder's inequality

$$|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_p\|\boldsymbol{B}\|_q,$$

which is valid for $\frac{1}{p} + \frac{1}{q} = 1$. By choosing $p \to \infty$ and $q \to 1$, we have

$$\mathrm{tr}\left(\boldsymbol{AB}\right) = \left\langle \boldsymbol{A}^\top, \boldsymbol{B}\right\rangle \leq \|\boldsymbol{A}\|_\infty\|\boldsymbol{B}\|_1 = \|\boldsymbol{A}\|_{2,2}\|\boldsymbol{B}\|_1 = \|\boldsymbol{A}\|_\infty\mathrm{tr}\left(\boldsymbol{B}\right),$$

where the last equality follows from the fact that, for positive semidefinite matrices such as $\boldsymbol{B}$, it follows that

$$\mathrm{tr}\left(\boldsymbol{B}\right) = \sum_i \lambda_i\left(\boldsymbol{B}\right) = \sum_i \sigma_i\left(\boldsymbol{B}\right),$$

where $\lambda_i\left(\boldsymbol{B}\right)$ is the ith eigenvalue of $\boldsymbol{B}$.

□

Lemma 7. *Let δ_j be an indicator such that $\delta_j \in \{0,1\}$ depending on the index $j = 1, \dots, d$, $\boldsymbol{u}$ satisfy Assumption 1, and all expectations be respect to $\boldsymbol{u}$. Then,*

$$\left\| \mathbb{E} \textstyle\sum_j \delta_j \; u_j^2 \boldsymbol{u}\boldsymbol{u}^\top \right\|_{2,2} \;\leq\; \sum_j \delta_j + k_\varphi - 1.$$

Proof. Let us denote

$$\boldsymbol{A} = \mathbb{E} \textstyle\sum_j \delta_j \; u_j^2 \boldsymbol{u}\boldsymbol{u}^\top.$$

Then, we have

$$\begin{aligned} A_{ij} &= \mathbb{E} \textstyle\sum_k \delta_k \; u_k^2 \; (\boldsymbol{u}\boldsymbol{u}^\top)_{ij} \\ &= \textstyle\sum_k \delta_k \; \mathbb{E} u_k^2 u_i u_j. \end{aligned}$$

Especially for the diagonal elements, we have

$$\begin{aligned} A_{ii} &= \sum_k \delta_k \; \mathbb{E} u_k^2 \mathbb{E} u_i^2 \\ &= \left(\sum_{k \neq i} \delta_k \; \mathbb{E} u_k^2 \mathbb{E} u_i^2 \right) + \delta_i \mathbb{E} u_i^4 && (u_i \perp\!\!\!\perp u_k) \\ &= \left(\sum_{k \neq i} \delta_k \right) + \delta_i \, k_\varphi && (\text{Assumption 1}) \\ &\leq \left(\sum_{k \neq i} \delta_k \right) + \delta_i \, k_\varphi + \underbrace{(k_\varphi - \delta_i \, k_\varphi + \delta_i - 1)}_{\geq 0 \text{ since } k_\varphi \geq 1} \\ &= \left(\sum_k \delta_k \right) + k_\varphi - 1. \end{aligned}$$

For the off-diagonal elements,

$$\begin{aligned} A_{ij} &= \sum_k \delta_k \; \mathbb{E} u_k^2 u_i u_j \\ &= \mathbb{E} \left(\sum_{(k \neq i) \wedge (k \neq j)} \delta_k \; u_k^2 u_i u_j \right) + \mathbb{E} \delta_i \; u_i^2 u_i u_j + \mathbb{E} \delta_j \; u_j^2 u_i u_j \\ &= \left(\sum_{(k \neq i) \wedge (k \neq j)} \delta_k \; \mathbb{E} u_k^2 \mathbb{E} u_i \mathbb{E} u_j \right) + \delta_i \; \mathbb{E} u_i^3 \mathbb{E} u_j + \delta_j \; \mathbb{E} u_i \mathbb{E} u_j^3 && (u_i \perp\!\!\!\perp u_j, \, u_i \perp\!\!\!\perp u_k, \text{ and } u_j \perp\!\!\!\perp u_k) \\ &= 0. && (\text{Assumption 1}) \end{aligned}$$

Therefore, $\boldsymbol{A}$ is a diagonal matrix and the ℓ_2 matrix norm follows as the maximal diagonal element such that

$$\|\boldsymbol{A}\|_{2,2} \leq \sum_j \delta_j + k_\varphi - 1.$$

□

B.3. Convergence of Stochastic Proximal Gradient Descent

B.3.1. Overview

Consider a general class of optimization problems of the form

$$\underset{\lambda \in \Lambda}{\text{minimize}} \quad F(\lambda) \triangleq f(\lambda) + h(\lambda),$$

where f is convex and smooth while h is a possibly non-smooth but convex regularizer. Stochastic proximal gradient descent is a family of methods aimed to solve these types of problems by employing a *proximal operator* denoted as

$$\text{prox}_{\gamma,h}(\lambda) = \underset{\lambda' \in \Lambda}{\arg\min} \left[h(\lambda) + \frac{1}{2\gamma} \|\lambda - \lambda'\|_2^2 \right].$$

Naturally, we expect the proximal operator to have closed form expression that can easily be computed.

Stochastic proximal gradient descent performs a gradient descent step on f, followed by a proximal step on h such that

$$\lambda_{t+1} = \text{prox}_{\gamma_t,h}(\lambda_t - \gamma_t \boldsymbol{g}(\lambda_t)),$$

where $\boldsymbol{g}$ is some unbiased stochastic estimate of ∇f. This type of algorithm has first been studied by Duchi et al. (2011); Ghadimi et al. (2016) followed by many more.

B.3.2. Technical Assumptions

Any proximal operator needs to satisfy the following basic property:

Assumption 3 (Non-Expansiveness). The proximal operator is non-expansive such that

$$\|\text{prox}_{\gamma,h}(\lambda) - \text{prox}_{\gamma,h}(\lambda')\| \leq \|\lambda - \lambda'\|$$

for all $\lambda, \lambda' \in \Lambda$.

This assumption can be satisfied as long as h is lower semi-continuous and convex.

For the convergence guarantee of proximal SGD, we follow the "variance transfer" strategy. That is, instead of directly bounding the gradient variance on an arbitrary point λ, Moulines & Bach (2011); Nguyen et al. (2018) "transfer" the gradient variance to the global optimum λ^*, proving convergence for gradient estimators with variance that grows with $f(\lambda)$. The key assumptions are as follows:

Assumption 4 (Convex Expected Smoothness). The gradient estimator $\boldsymbol{g}$ is an unbiased estimator of f and is convex-smooth in expectation such that

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

for any $\lambda, \lambda' \in \Lambda$ and some constant $0 < \mathcal{L} < \infty$, where D_f is the Bregman divergence generated by f.

This assumption has first been used by Gower et al. (2021) and since has been popularly used throughout stochastic optimization. (See Khaled & Richtárik (2023) for an overview of conditions on the gradient variance.)

Assumption 5 (Bounded Gradient Variance). The variance of the gradient estimator $\boldsymbol{g}$ is bounded as

$$\mathrm{tr}\mathbb{V}\boldsymbol{g}(\lambda^*) \leq \sigma^2$$

for some constant $0 \leq \sigma^2 < \infty$, where $\lambda^* = \arg\min F(\lambda)$ is the global optimum.

This assumption only requires the gradient variance on the global optimum to be finite.

B.3.3. Convergence (Lemma 8)

For completeness, we restate the proof by Garrigos & Gower (2023, Theorem 11.9), which is based on the more general results of Gorbunov et al. (2020).

Lemma 8 (Gorbunov et al., 2020)**.** *Let $F = f + h$ be a composite objective on a convex domain Λ, where f is μ-strongly convex and h satisfies Assumption 3. Also, the gradient estimator $\boldsymbol{g}$ satisfies Assumption 4 and 5. The last iterate $\boldsymbol{\lambda}_T$ after T iterations of the stochastic proximal gradient descent with a constant step size γ satisfying $\gamma \in \left(0, \min\left\{\frac{1}{2\mathcal{L}}, \frac{1}{\mu}\right\}\right]$ achieves the bound on the distance to the global optimum $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda} F(\boldsymbol{\lambda})$ as*

$$\mathbb{E}\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}^*\|_2^2 \leq (1-\gamma\mu)^T \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Proof. First, the iterate at time t satisfies

$$\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*\|_2^2 = \|\mathrm{prox}_{\gamma h}(\boldsymbol{\lambda}_t - \gamma \boldsymbol{g}(\boldsymbol{\lambda}_t)) - \mathrm{prox}_{\gamma h}(\boldsymbol{\lambda}^* - \gamma\nabla f(\boldsymbol{\lambda}^*))\|_2^2,$$

and from Assumption 3,

$$\begin{aligned}
&\leq \|(\boldsymbol{\lambda}_t - \gamma \boldsymbol{g}(\boldsymbol{\lambda}_t)) - (\boldsymbol{\lambda}^* - \gamma\nabla f(\boldsymbol{\lambda}^*))\|_2^2 \\
&= \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 + \gamma^2\|\boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*)\|_2^2 - 2\gamma\langle \boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\rangle.
\end{aligned}$$

Therefore, denoting the filtration as $\mathcal{F}_t$, which is the σ-field generated by iterates up to t, the conditional expectation is bounded as

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*\|_2^2 \mid \mathcal{F}_t\right] \\
&\quad\leq \mathbb{E}\left[\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 \mid \mathcal{F}_t\right] + \gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*) \mid \mathcal{F}_t\|_2^2\right] - 2\gamma\mathbb{E}\left[\langle \boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\rangle \mid \mathcal{F}_t\right].
\end{aligned} \tag{7}$$

The second term, or the gradient variance term of Eq. (7), can be dealt with the "variance transfer" strategy pioneered by (Nguyen et al., 2018):

$$\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*)\|_2^2 \mid \mathcal{F}_t\right] \leq 2\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{\lambda}_t) - \boldsymbol{g}(\boldsymbol{\lambda}^*)\|_2^2 \mid \mathcal{F}_t\right] + 2\gamma^2\mathbb{E}\left[\|\boldsymbol{g}(\boldsymbol{\lambda}^*) - \nabla f(\boldsymbol{\lambda}^*)\|_2^2 \mid \mathcal{F}_t\right],$$

and applying Assumption 4 and 5,

$$\leq 2\gamma^2\left(2\mathcal{L}\,\mathrm{D}_f(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^*)\right) + 2\gamma^2\sigma^2. \tag{8}$$

For the third term of Eq. (7), we have

$$\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \boldsymbol{g}(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\rangle \mid \mathcal{F}_t\right] &= -2\gamma\left\langle \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{\lambda}_t) \mid \mathcal{F}_t\right] - \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\right\rangle \\
&= -2\gamma\left\langle \nabla f(\boldsymbol{\lambda}_t) - \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\right\rangle \\
&= 2\gamma\left\langle \nabla f(\boldsymbol{\lambda}_t), \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t\right\rangle + 2\gamma\left\langle \nabla f(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\right\rangle,
\end{aligned}$$

and applying the μ-strong convexity of f,

$$\leq -\gamma\mu\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 - 2\gamma\,\mathrm{D}_f(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^*). \tag{9}$$

Applying Eqs. (8) and (9) to Eq. (7) and taking the full expectation, we get

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*\|_2^2 &\leq \mathbb{E}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 + 4\gamma^2\mathcal{L}\mathbb{E}\left[\mathrm{D}_f(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^*)\right] + 2\gamma^2\sigma^2 - \gamma\mu\mathbb{E}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 - 2\gamma\mathbb{E}\left[\mathrm{D}_f(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^*)\right] \\
&= (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 + 2\gamma\left(2\gamma\mathcal{L} - 1\right)\mathbb{E}\left[\mathrm{D}_f(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^*)\right] + 2\gamma^2\sigma^2,
\end{aligned}$$

by ensuring that $2\gamma\mathcal{L} \leq 1$, we have

$$\leq (1-\gamma\mu)\,\mathbb{E}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|_2^2 + 2\gamma^2\sigma^2.$$

Provided that $1 - \gamma\mu \geq 0$, we can solve the recursion above and get

$$\mathbb{E}\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*\|_2^2 \leq (1-\gamma\mu)^t \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 + 2\gamma^2\sigma^2\sum_{k=1}^{t-1}(1-\gamma\mu)^k.$$

Since the geometric sum can be upper-bounded by

$$\sum_{k=1}^{t-1} (1-\gamma\mu)^k = \frac{1-(1-\gamma\mu)^t}{\gamma\mu} \leq \frac{1}{\gamma\mu},$$

which yields

$$\mathbb{E}\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*\|_2^2 \leq (1-\gamma\mu)^t \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

□

B.3.4. COMPLEXITY (COROLLARY 3)

Corollary 3. *Let the assumptions of Lemma 8 be satisfied. Then, the last iterate $\boldsymbol{\lambda}_{T+1}$ of proximal SGD with a fixed stepsize satisfies $\mathbb{E}\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$, where $\boldsymbol{\lambda}^* = \arg\max_{\lambda\in\Lambda} F(\boldsymbol{\lambda})$ is the global optimum, if*

$$\gamma = \min\left\{\frac{\epsilon\mu}{4\sigma^2}, \frac{2\mathcal{L}}{\mu}, \mu\right\} \quad \textit{and}$$

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

Proof. We can convert Lemma 8 to an iteration complexity guarantee for achieving an ϵ-accurate solution using Lemma A.2 of Garrigos & Gower (2023) where the constants are:

$$\alpha_0 = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2, \qquad A = \frac{2\sigma^2}{\mu}, \text{ and} \qquad C = \max\{2\mathcal{L}, \mu\}.$$

That is, we can satisfy $\mathbb{E}\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ with

$$\gamma = \min\left\{\frac{\epsilon}{2A}, \frac{1}{C}\right\} = \min\left\{\frac{\epsilon\mu}{4\sigma^2}, 2\mathcal{L}, \mu\right\}$$

and

$$\begin{aligned} T &\geq \max\left\{\frac{1}{\epsilon}\frac{2A}{\mu}, \frac{C}{\mu}\right\} \log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right) \\ &= \max\left\{\frac{1}{\epsilon}\frac{4\sigma^2}{\mu^2}, \frac{2\mathcal{L}}{\mu}, 1\right\} \log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2 \frac{1}{\epsilon}\right). \end{aligned}$$

□

B.4. Convergence of Black-Box Variational Inference of Finite-Sum Likelihoods

B.4.1. Overview

Now, we establish the complexity of BBVI applied to finite-sum likelihoods of the form of Eq. (1).

Domain As discussed in Section 2.2, we restrict our interest to location-scale variational families with a Cholesky scale parameterization. This effectively means we restrict $\boldsymbol{C}$ to be lower triangular and have only positive diagonal entries. The domain of the variational parameters is then

$$\Lambda \triangleq \{(\boldsymbol{m}, \boldsymbol{C}) \mid (\boldsymbol{m}, \boldsymbol{C}) \in \mathbb{R}^d \times \mathbb{L}_{++}^d(\mathbb{R}) \text{ and } C_{ii} > 0 \text{ for all } i = 1, \dots, d\},$$

which is convex. Furthermore, the negative entropy

$$h(\boldsymbol{\lambda}) = -\mathbb{H}(q_{\boldsymbol{\lambda}}) = -\log\det \boldsymbol{C} + \mathbb{1}_{\Lambda}(\boldsymbol{\lambda}) = -\sum_{i=1}^{d} \log C_{ii} + \mathbb{1}_{(0,+\infty)}(C_{ii}),$$

where $\mathbb{1}_A(x)$ is 0 if $x \in A$ and ∞ otherwise, is closed and convex (Domke et al., 2023, Lemma 19).

Proximal Operator Also, as noted in Section 2.4, we leverage proximal SGD, for which we provide detailed analysis in Appendix B.3 for completeness. We use the proximal operator proposed by Domke (2020), which, at each application updates the diagonal of $\boldsymbol{C}$ such that

$$\operatorname{prox}_{\gamma,h}(\boldsymbol{\lambda}) = (\boldsymbol{m}, \boldsymbol{C} + \Delta\boldsymbol{C}), \tag{10}$$

where $\boldsymbol{\lambda} = (\boldsymbol{m}, \boldsymbol{C})$ and

$$\Delta C_{ii} = \frac{1}{2}\left(\sqrt{C_{ii}^2 + 4\gamma} - C_{ii}\right),$$

$\Delta C_{ij} = 0$ for $i \neq j$. Conveniently, the complexity of applying this proximal operator is $\Theta(d)$. Furthermore, this proximal operator satisfies Assumption 3. (See the recent work by Domke et al. (2023) for more details.)

Proof Sketch The proof is based on the general analysis of proximal SGD in Appendix B.3. Specifically, we establish

1. Assumption 4 in Appendix B.4.4 and
2. Assumption 5 in Appendix B.4.3.

The key ingredients are in Appendix B.4.2.

1. Lemma 9 establishes the precise expression of the squared Jacobian of $\mathcal{T}_{\boldsymbol{\lambda}}^n$.
2. Lemma 10 connects the properties of the gradient variance with $\mathcal{T}_{\boldsymbol{\lambda}}^n$ through its Jacobian.
3. Lemma 11 resolves the randomness by directly solving the expectations.

B.4.2. Key Lemmas

Lemma 9. *The squared Jacobian of the reparameterization function $\mathcal{T}_\lambda^n$ for ℓ_n, $\boldsymbol{J}_n$, is a diagonal matrix given as*

$$\boldsymbol{J}_n(\boldsymbol{u})^\top \boldsymbol{J}_n(\boldsymbol{u}) = \left(1 + \textstyle\sum_j \delta_{n,j} u_j^2\right) \mathbf{I}_n$$

where $\mathbf{I}_n$ is the identity matrix of the subspace of ℓ_n (Definition 2) and u_j is jth element of $\boldsymbol{u}$.

Proof. Given the location-scale reparameterization function, $\mathcal{T}_\lambda$, we define the function for nth datapoint as

$$\mathcal{T}_\lambda^n(\boldsymbol{u}) = \boldsymbol{C}_n \boldsymbol{u} + \boldsymbol{m}_n,$$

where $\boldsymbol{C}_n$ and $\boldsymbol{m}_n$ has non-zero elements that are only related to the nth datapoint and makes all others zero.

For the derivatives, Domke (2019, Lemma 8) show that

$$\frac{\partial \mathcal{T}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{m}_i} = \mathbf{e}_i \quad \text{and} \quad \frac{\partial \mathcal{T}_\lambda(\boldsymbol{u})}{\partial \boldsymbol{C}_{i,j}} = \mathbf{e}_i u_j,$$

where $\mathbf{e}_i$ is the unit bases for the ith coordinate. Here, we are interested in the effect of the structure of $\boldsymbol{C}_n$.

Using the indicator δ, we have

$$\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{C}_{n,i,j}} = \delta_{n,i,j}\, \mathbf{e}_{n,i}\, u_j \tag{11}$$

where

$$\delta_{n,i,j} = \begin{cases} 1 & \text{if the } i,j\text{th element of } \boldsymbol{C}_n \text{ is non-zero} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{C}_{n,i,j}} \left(\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{C}_{n,i,j}}\right)^\top = \left(\delta_{n,i,j}\, \mathbf{e}_{n,i}\, u_j\right)\left(\delta_{n,i,j}\, \mathbf{e}_{n,i}\, u_j\right)^\top = \delta_{n,i,j}\, u_j^2 \left(\mathbf{e}_{n,i}\, \mathbf{e}_{n,i}^\top\right)$$

and thus, now denoting the Jacobian as $\boldsymbol{J}_n$, we have

$$\begin{aligned}
\boldsymbol{J}_n(\boldsymbol{u})^\top \boldsymbol{J}_n(\boldsymbol{u}) &= \left(\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{\lambda}}\right)^\top \frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \\
&= \sum_{i=1} \frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{m}_i}\left(\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{m}_i}\right)^\top + \sum_i \sum_j \frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{C}_{n,i,j}}\left(\frac{\partial \mathcal{T}_\lambda^n(\boldsymbol{u})}{\partial \boldsymbol{C}_{n,i,j}}\right)^\top \\
&= \sum_i \mathbf{e}_{n,i}\, \mathbf{e}_{n,i}^\top + \sum_i \sum_j \delta_{n,i,j}\, u_j^2 \left(\mathbf{e}_{n,i}\, \mathbf{e}_{n,i}^\top\right) \\
&= \mathbf{I}_n + \sum_i \sum_j \delta_{n,i,j}\, u_j^2 \left(\mathbf{e}_{n,i}\, \mathbf{e}_{n,i}^\top\right),
\end{aligned}$$

assuming $\delta_{n,i,j} = \delta_{n,j}$ for all n, j, then

$$\begin{aligned}
&= \mathbf{I}_n + \sum_j \delta_{n,j} u_j^2 \left(\sum_i \mathbf{e}_{n,i}\, \mathbf{e}_{n,i}^\top\right) \\
&= \mathbf{I}_n + \sum_j \delta_{n,j} u_j^2\, \mathbf{I}_n
\end{aligned}$$

where

$$\delta_{n,j} = \begin{cases} 1 & \text{if the } j\text{th column of } \boldsymbol{C}_n \text{ is non-zero} \\ 0 & \text{otherwise.} \end{cases}$$

□

Lemma 10. *Let $\ell_n : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and $\mathcal{T}^n$ be the location-scale reparameterization function for ℓ_n. Then, we have*

$$\|\nabla_{\boldsymbol{\lambda}} \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla_{\boldsymbol{\lambda}} \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\|_2^2 = \left(1 + \textstyle\sum_j \delta_{n,j} u_j^2\right) \|\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\|_2^2$$

for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ and any $\boldsymbol{u} \in \mathbb{R}^d$.

Proof.

$$\begin{aligned}
&\left\|\nabla_{\boldsymbol{\lambda}} \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla_{\boldsymbol{\lambda}} \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right\|_2^2 \\
&= \left\|\frac{\partial \mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})}{\partial \boldsymbol{\lambda}} \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \frac{\partial \mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u})}{\partial \boldsymbol{\lambda}'} \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right\|_2^2.
\end{aligned}$$

Now notice that $\frac{\partial \mathcal{T}^n_{\boldsymbol{\lambda}}}{\partial \boldsymbol{\lambda}} = \boldsymbol{J}^n$ independently of $\boldsymbol{\lambda}$. Then,

$$\begin{aligned}
&= \left\|\boldsymbol{J}^n(\boldsymbol{u}) \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \boldsymbol{J}^n(\boldsymbol{u}) \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right\|_2^2 \\
&= \left\|\boldsymbol{J}^n(\boldsymbol{u}) \left(\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right)\right\|_2^2 \\
&= \left(\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right)^\top \boldsymbol{J}^n(\boldsymbol{u})^\top \boldsymbol{J}^n(\boldsymbol{u}) \left(\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right),
\end{aligned}$$

and applying Lemma 9,

$$\begin{aligned}
&= \left(\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right)^\top \left(\left(1 + \sum_j \delta_{n,j} u_j^2\right) \mathbf{I}_n\right) \left(\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right) \\
&= \left(1 + \sum_j \delta_{n,j} u_j^2\right) \left\|\nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}}(\boldsymbol{u})) - \nabla \ell_n(\mathcal{T}^n_{\boldsymbol{\lambda}'}(\boldsymbol{u}))\right\|_2^2.
\end{aligned}$$

□

Lemma 11. *Let $\mathcal{T}_{\boldsymbol{\lambda}}^n$ be the location-scale reparameterization function for ℓ_n and let the vector-valued random variable $\boldsymbol{u} = (u_1, \dots, u_j)$ satisfy Assumption 1. Then,*

$$\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \bar{\boldsymbol{z}}_n\|_2^2 \leq \left(\sum_j \delta_{n,j} + 1\right) \|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \left(\sum_j \delta_{n,j} + k_\varphi\right) \|\boldsymbol{C}_n\|_{\mathrm{F}}^2$$

for any vector $\bar{\boldsymbol{z}}_n$ matching the output dimension of $\mathcal{T}_{\boldsymbol{\lambda}}^n$ and any $\boldsymbol{\lambda} \in \Lambda$.

Proof. First,

$$\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \bar{\boldsymbol{z}}_n\|_2^2 = \mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \|\boldsymbol{C}_n \boldsymbol{u} + \boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2,$$

expanding the square,

$$\begin{aligned}
&= \mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \left(\boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u} - 2(\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n)^\top \boldsymbol{C}_n \boldsymbol{u} + \|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2\right) \\
&= \underbrace{\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u}}_{T^{(1)}} - 2 \underbrace{\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) (\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n)^\top \boldsymbol{C}_n \boldsymbol{u}}_{T^{(2)}} \\
&\quad + \underbrace{\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2}_{T^{(3)}}.
\end{aligned} \tag{12}$$

We now solve for the individual terms $T^{(1)}$, $T^{(2)}$, and $T^{(3)}$.

Derivation of $T^{(1)}$ First,

$$\begin{aligned}
T^{(1)} &= \mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u} \\
&= \mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right) \operatorname{tr}\left(\boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u}\right) \\
&= \mathbb{E}\left(\boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u}\right) + \mathbb{E}\left(\sum_j \delta_{n,j} u_j^2\right) \operatorname{tr}\left(\boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u}\right),
\end{aligned}$$

applying Lemma 3 on the first term,

$$= \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 + \mathbb{E}\left(\sum_j \delta_{n,j} u_j^2\right) \operatorname{tr}\left(\boldsymbol{u}^\top \boldsymbol{C}_n^\top \boldsymbol{C}_n \boldsymbol{u}\right),$$

rotating the elements in the trace and pushing the scalar coefficient into the trace,

$$= \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 + \operatorname{tr}\left(\boldsymbol{C}_n^\top \boldsymbol{C}_n \mathbb{E}\left(\sum_j \delta_{n,j} u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right)\right).$$

Since we only deal with real matrices, we have $\boldsymbol{C}_n^\top \boldsymbol{C}_n \succeq 0$, enabling the use of Lemma 6 as

$$\begin{aligned}
&\leq \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 + \left\|\mathbb{E}\sum_j \delta_{n,j}\ u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right\|_{2,2} \operatorname{tr}\left(\boldsymbol{C}_n^\top \boldsymbol{C}_n\right) \\
&= \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 + \left\|\mathbb{E}\sum_j \delta_{n,j}\ u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right\|_{2,2} \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 \\
&= \left(1 + \left\|\mathbb{E}\sum_j \delta_{n,j}\ u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right\|_{2,2}\right) \|\boldsymbol{C}_n\|_{\mathrm{F}}^2.
\end{aligned} \tag{13}$$

By using Lemma 7, we have

$$\left\|\mathbb{E}\sum_j \delta_{n,j}\ u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right\|_{2,2} \leq \sum_j \delta_{n,j} + k_\varphi - 1$$

Therefore, back to Eq. (13),

$$T^{(1)} = \left(1 + \left\|\mathbb{E}\sum_j \delta_{n,j}\ u_j^2 \boldsymbol{u}\boldsymbol{u}^\top\right\|_{2,2}\right) \|\boldsymbol{C}_n\|_{\mathrm{F}}^2 \leq \left(\sum_j \delta_{n,j} + k_\varphi\right) \|\boldsymbol{C}_n\|_{\mathrm{F}}^2. \tag{14}$$

Derivation of $T^{(2)}$ Meanwhile, for $T^{(2)}$,

$$T^{(2)} = \mathbb{E}\left(1 + \sum_j \delta_{n,j}\, u_j^2\right)(\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n)\boldsymbol{C}_n \boldsymbol{u},$$

and from Assumption 1 and Lemma 2,

$$\begin{aligned} &= (\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n)\boldsymbol{C}_n \mathbb{E}\boldsymbol{u} \;+\; (\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n)\boldsymbol{C}_n\left(\sum_j \delta_{n,j}\, \mathbb{E}u_j^2 \boldsymbol{u}\right) \\ &= 0. \end{aligned} \tag{15}$$

Derivation of $T^{(3)}$ Finally,

$$\begin{aligned} T^{(3)} &= \mathbb{E}\left(1 + \sum_j \delta_{n,j}\, u_j^2\right)\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 \\ &= \left(1 + \sum_j \delta_{n,j}\, \mathbb{E}u_j^2\right)\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2, \end{aligned}$$

and from Assumption 1,

$$= \left(1 + \sum_j \delta_{n,j}\right)\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2. \tag{16}$$

Combining all the results to Eq. (12),

$$\mathbb{E}\left(1 + \sum_j \delta_{n,j}\, u_j^2\right)\left\|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \bar{\boldsymbol{z}}_n\right\|_2^2 \leq T^{(1)} - 2T^{(2)} + T^{(3)},$$

applying Eqs. (14) to (16),

$$\leq \left(\sum_j \delta_{n,j} + 1\right)\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \left(\sum_j \delta_{n,j} + k_\varphi\right)\|\boldsymbol{C}_n\|_{\mathrm{F}}^2.$$

□

Corollary 4. *Let the assumptions of Lemma 11 hold. Then,*

$$\mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right)\|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u})\|_2^2 \leq \left(\sum_j \delta_{n,j} + k_\varphi\right)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2.$$

Proof. From the linearity of $\mathcal{T}^n$, it follows that

$$\mathcal{T}_{\boldsymbol{\lambda}-\boldsymbol{\lambda}'}^n(\boldsymbol{u}) = \left(\boldsymbol{C}_n - \boldsymbol{C}_n'\right)\boldsymbol{u} + (\boldsymbol{m}_n - \boldsymbol{m}_n'),$$

where $\boldsymbol{m}_n, \boldsymbol{C}_n$ are part of $\boldsymbol{\lambda}$ and $\boldsymbol{m}_n', \boldsymbol{C}_n'$ are part of $\boldsymbol{\lambda}'$. Then, the result immediately follow from applying Lemma 11 with $\boldsymbol{z} = \boldsymbol{0}$ as

$$\begin{aligned} \mathbb{E}\left(1 + \sum_j \delta_{n,j} u_j^2\right)\|\mathcal{T}_{\boldsymbol{\lambda}-\boldsymbol{\lambda}'}^n(\boldsymbol{u})\|_2^2 &\leq \left(\sum_j \delta_{n,j} + 1\right)\left\|\boldsymbol{m}_n - \boldsymbol{m}_n'\right\|_2^2 + \left(\sum_j \delta_{n,j} + k_\varphi\right)\left\|\boldsymbol{C}_n - \boldsymbol{C}_n'\right\|_{\mathrm{F}}^2 \\ &= \left(\sum_j \delta_{n,j} + 1\right)\left\|\boldsymbol{m}_n - \boldsymbol{m}_n'\right\|_2^2 + \left(\sum_j \delta_{n,j} + k_\varphi\right)\left\|\boldsymbol{C}_n - \boldsymbol{C}_n'\right\|_{\mathrm{F}}^2, \end{aligned}$$

since the kurtosis always satisfies $k_\varphi \geq 1$,

$$\leq \left(\sum_j \delta_{n,j} + k_\varphi\right)\left(\left\|\boldsymbol{m}_n - \boldsymbol{m}_n'\right\|_2^2 + \left\|\boldsymbol{C}_n - \boldsymbol{C}_n'\right\|_{\mathrm{F}}^2\right)$$

and since adding more components into a squared ℓ_2-norm always results in an upper bound,

$$\begin{aligned} &\leq \left(\sum_j \delta_{n,j} + k_\varphi\right)\left(\|\boldsymbol{m} - \boldsymbol{m}'\|_2^2 + \|\boldsymbol{C} - \boldsymbol{C}'\|_{\mathrm{F}}^2\right) \\ &= \left(\sum_j \delta_{n,j} + k_\varphi\right)\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2. \end{aligned}$$

□

B.4.3. Gradient Variance Bound (Theorem 2)

Theorem 2. *Let ℓ_n be L_n-smooth for some $n = 1, \dots, N$ and Assumption 2 hold. Then, the gradient variance of $\boldsymbol{g}_M$ is bounded as*

$$\operatorname{tr}\mathbb{V}\,\boldsymbol{g}_M(\boldsymbol{\lambda}) \leq \frac{N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^{N} L_n^2\left(\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n\|_{\mathrm{F}}^2\right),$$

where $\bar{\boldsymbol{z}}_n$ is a stationary point of ℓ_n and

$$d^* \triangleq \max_n \textstyle\sum_j \delta_{n,j}$$

is the effective dimensionality.

Proof. The proof can be seen as a generalization of Domke (2019, Theorem 1) to the case where $\boldsymbol{C}$ is structured. Firstly,

$$\begin{aligned}
\operatorname{tr}\mathbb{V}\boldsymbol{g}_M(\boldsymbol{\lambda}) &\leq \operatorname{tr}\mathbb{V}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\nabla_{\boldsymbol{\lambda}}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}_m)\right)\right] \\
&= \frac{1}{M}\operatorname{tr}\mathbb{V}\left[\sum_{n=1}^{N}\nabla_{\boldsymbol{\lambda}}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right)\right] && (\boldsymbol{u}_1, \dots, \boldsymbol{u}_M \text{ are } i.i.d.) \\
&\leq \frac{N}{M}\sum_{n=1}^{N}\operatorname{tr}\mathbb{V}\left[\nabla_{\boldsymbol{\lambda}}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right)\right] && \text{(Lemma 4)} \\
&\leq \frac{N}{M}\sum_{n=1}^{N}\mathbb{E}\left\|\nabla_{\boldsymbol{\lambda}}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right)\right\|_2^2.
\end{aligned}$$

Now, the individual expected-squared norms can be bounded as

$$\begin{aligned}
\mathbb{E}\left\|\nabla_{\boldsymbol{\lambda}}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right)\right\|_2^2 &= \mathbb{E}\left(1 + \textstyle\sum_j \delta_{n,j}u_j^2\right)\left\|\nabla\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right)\right\|_2^2 && \text{(Lemma 10)} \\
&= \mathbb{E}\left(1 + \textstyle\sum_j \delta_{n,j}u_j^2\right)\left\|\nabla\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right) - \nabla\ell_n(\bar{\boldsymbol{z}}_n)\right\|_2^2 && (\text{since } \nabla\ell_n(\bar{\boldsymbol{z}}_n) = \boldsymbol{0}) \\
&\leq L_n^2\,\mathbb{E}\left(1 + \textstyle\sum_j \delta_{n,j}u_j^2\right)\left\|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \bar{\boldsymbol{z}}_n\right\|_2^2 && (L_n\text{-smoothness}) \\
&\leq L_n^2\left(\textstyle\sum_j \delta_{n,j} + k_\varphi\right)\left(\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n\|_{\mathrm{F}}^2\right). && \text{(Lemma 11)}
\end{aligned}$$

The sum of all the datapoints can then be bounded as

$$\begin{aligned}
\operatorname{tr}\mathbb{V}\boldsymbol{g}_M(\boldsymbol{\lambda}) &= \frac{N}{M}\sum_{n=1}^{N} L_n^2\left(\textstyle\sum_j \delta_{n,j} + k_\varphi\right)\left(\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n\|_{\mathrm{F}}^2\right) \\
&\leq \frac{N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^{N} L_n^2\left(\|\boldsymbol{m}_n - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n\|_{\mathrm{F}}^2\right).
\end{aligned}$$

□

B.4.4. Convex Expected Smoothness (Lemma 12)

Lemma 12 (Convex Expected Smoothness). *Let ℓ be μ-strongly convex and L-smooth, ℓ_n be L_n-smooth for $n = 1, \dots, N$, and Assumption 2 hold. Then, we have*

$$\mathbb{E}\|\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\|_2^2 \leq 2\left(\frac{N}{M}(d^* + k_\varphi)\sum_{n=1}^{N}\frac{L_n^2}{\mu} + L\right)\mathrm{D}_f(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$$

for any $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$, where D_f is the Bregman divergence generated by f, and $d^* = \max_n \sum_j \delta_{n,j}$ is the effective dimensionality.

Proof. The proof is a generalization of Lemma 3 by Kim et al. (2023a), which a strategy of applying smoothness (Domke, 2019) and then applying "quadratic functional growth" (Kim et al., 2023b).

First notice that if ℓ is L-smooth, then the energy is also L-smooth by virtue of Domke (2020, Theorem 1). In our case, ℓ is $NL_{\max}$ smooth. Therefore,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\|_2^2 &= \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\right] + \left\|\nabla f(\boldsymbol{\lambda}) - \nabla f(\boldsymbol{\lambda}')\right\|_2^2 \\
&\leq \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\right] + 2L\left(f(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda}') - \langle \nabla f(\boldsymbol{\lambda}'), \boldsymbol{\lambda} - \boldsymbol{\lambda}'\rangle\right) \quad (L\text{-smoothness of } \ell) \\
&= \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\right] + 2L\,\mathrm{D}_f(\boldsymbol{\lambda}, \boldsymbol{\lambda}').
\end{aligned}$$

The variance term can be bounded as

$$\begin{aligned}
\mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_M(\boldsymbol{\lambda}) - \boldsymbol{g}_M(\boldsymbol{\lambda}')\right] &= \frac{1}{M^2}\mathrm{tr}\mathbb{V}\left[\sum_{m=1}^{M}\sum_{n=1}^{N}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}_m)\right) - \ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u}_m)\right)\right] \\
&= \frac{1}{M}\mathrm{tr}\mathbb{V}\left[\sum_{n=1}^{N}\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right) - \ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u})\right)\right] && (\boldsymbol{u}_1, \dots, \boldsymbol{u}_m \text{ are } i.i.d.) \\
&\leq \frac{N}{M}\sum_{n=1}^{N}\mathrm{tr}\mathbb{V}\left[\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right) - \ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u})\right)\right] && (\text{Lemma 4}) \\
&\leq \frac{N}{M}\sum_{n=1}^{N}\mathbb{E}\left\|\ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})\right) - \ell_n\left(\mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u})\right)\right\|_2^2 \\
&= \frac{N}{M}\sum_{n=1}^{N}\mathbb{E}\left(1 + \textstyle\sum_j \delta_{n,j}u_j^2\right)\|\nabla\ell_n(\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u})) - \nabla\ell_n(\mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u}))\|_2^2 && (\text{Lemma 10}) \\
&\leq \frac{N}{M}\sum_{n=1}^{N}L_n^2\mathbb{E}\left(1 + \textstyle\sum_j \delta_{n,j}u_j^2\right)\|\mathcal{T}_{\boldsymbol{\lambda}}^n(\boldsymbol{u}) - \mathcal{T}_{\boldsymbol{\lambda}'}^n(\boldsymbol{u})\|_2^2 && (L_n\text{-smoothness}) \\
&\leq \frac{N}{M}\left\{\sum_{n=1}^{N}L_n^2\left(\textstyle\sum_j \delta_{n,j} + k_\varphi\right)\right\}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2. && (\text{Corollary 4})
\end{aligned}$$

Now, we bound $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|$ by the Bregman divergence generated by f as done by Kim et al. (2023b, Theorem 1). For this, we convert the squared distance of the variational parameters ($\boldsymbol{\lambda}$-space) into the squared distance in model parameters ($\boldsymbol{z}$-space). That is,

$$\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2 = \|\boldsymbol{C} - \boldsymbol{C}'\|_{\mathrm{F}}^2 + \|\boldsymbol{m} - \boldsymbol{m}'\|_2^2,$$

using the identity in Lemma 3,

$$\begin{aligned}
&= \mathbb{E}\left\|(\boldsymbol{C} - \boldsymbol{C}')\boldsymbol{u}\right\|_2^2 + \|\boldsymbol{m} - \boldsymbol{m}'\|_2^2 \\
&= \mathbb{E}\|\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{u}) - \mathcal{T}_{\boldsymbol{\lambda}'}(\boldsymbol{u})\|_2^2
\end{aligned}$$

by the μ-strong log-concavity of the posterior,

$$\leq \frac{2}{\mu}\mathbb{E}\left(\ell\left(\mathcal{T}_{\lambda}\left(\boldsymbol{u}\right)\right)-\ell\left(\mathcal{T}_{\lambda}\left(\boldsymbol{u}\right)\right)-\left\langle\nabla\ell\left(\mathcal{T}_{\lambda'}\left(\boldsymbol{u}\right)\right),\mathcal{T}_{\lambda}\left(\boldsymbol{u}\right)-\mathcal{T}_{\lambda'}\left(\boldsymbol{u}\right)\right\rangle\right)$$

$$=\frac{2}{\mu}\left(f\left(\lambda\right)-f\left(\lambda'\right)-\mathbb{E}\left\langle\nabla\ell\left(\mathcal{T}_{\lambda'}\left(\boldsymbol{u}\right)\right),\mathcal{T}_{\lambda}\left(\boldsymbol{u}\right)-\mathcal{T}_{\lambda'}\left(\boldsymbol{u}\right)\right\rangle\right),$$

and applying Lemma 10 by Kim et al. (2023a) on the inner product term,

$$=\frac{2}{\mu}\left(f\left(\lambda\right)-f\left(\lambda'\right)-\left\langle\nabla f\left(\lambda'\right),\lambda-\lambda'\right\rangle\right)$$

$$=\frac{2}{\mu}\mathrm{D}_{f}\left(\lambda,\lambda'\right). \tag{17}$$

Combining the results,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{g}_{M}\left(\lambda\right)-\boldsymbol{g}_{M}\left(\lambda'\right)\|_{2}^{2} &= \mathrm{tr}\mathbb{V}\left[\boldsymbol{g}_{M}\left(\lambda\right)-\boldsymbol{g}_{M}\left(\lambda'\right)\right]+2L\,\mathrm{D}_{f}\left(\lambda,\lambda'\right)\\
&=\frac{N}{M}\left\{\sum_{n=1}^{N}L_{n}^{2}\left(\textstyle\sum_{j}\delta_{n,j}+k_{\varphi}\right)\right\}\|\lambda-\lambda'\|_{2}^{2}+2L\,\mathrm{D}_{f}\left(\lambda,\lambda'\right)\\
&\leq\frac{N}{M}\left\{\sum_{n=1}^{N}L_{n}^{2}\left(\textstyle\sum_{j}\delta_{n,j}+k_{\varphi}\right)\right\}\frac{2}{\mu}\mathrm{D}_{f}\left(\lambda,\lambda'\right)+2L\,\mathrm{D}_{f}\left(\lambda,\lambda'\right) && \text{(Eq. (17))}\\
&=2\left(\frac{N}{M}\sum_{n=1}^{N}\frac{L_{n}^{2}}{\mu}\left(\textstyle\sum_{j}\delta_{n,j}+k_{\varphi}\right)+L\right)\mathrm{D}_{f}\left(\lambda,\lambda'\right), && \text{(Reorganized)}\\
&\leq 2\left(\frac{N}{M}\sum_{n=1}^{N}\frac{L_{n}^{2}}{\mu}\left(d^{*}+k_{\varphi}\right)+L\right)\mathrm{D}_{f}\left(\lambda,\lambda'\right) && (\textstyle\sum_{j}\delta_{n,j}\leq d^{*}\text{ for all }n=1,\ldots,N)\\
&=2\left(\frac{N}{M}\left(d^{*}+k_{\varphi}\right)\frac{\sum_{n=1}^{N}L_{n}^{2}}{\mu}+L\right)\mathrm{D}_{f}\left(\lambda,\lambda'\right). && \text{(Pushed-in the summation)}
\end{aligned}$$

□

B.4.5. Complexity with General Location-Scale Families (Theorem 3)

Theorem 3. *Let ℓ be μ-strongly convex and L-smooth, ℓ_n be L_n-smooth for $n = 1, \ldots, N$, and Assumption 2 hold. Then, the last iterate $\boldsymbol{\lambda}_{T+1}$ of BBVI with proximal SGD and $\boldsymbol{g}_M$ is ϵ-close as $\mathbb{E}\|\boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}^*\|_2^2 \leq \epsilon$ to the global optimum $\boldsymbol{\lambda}^* = \arg\min F(\boldsymbol{\lambda})$ if*

$$T \geq \max\left(C_{\text{var}}\frac{1}{\epsilon},\ C_{\text{bias}}\right)\log\left(2\Delta_0^2\,\frac{1}{\epsilon}\right)$$

for some fixed stepsize γ, where $\Delta_0 = \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2$ is the distance to the optimum,

$$C_{\text{var}} = 4\frac{N}{M}\left(d^* + k_\varphi\right)\textstyle\sum_{n=1}^N \kappa_n^2\left(\|\boldsymbol{m}_n^* - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n^*\|_{\text{F}}^2\right)$$
$$C_{\text{bias}} = 2\frac{N}{M}\left(d^* + k_\varphi\right)\textstyle\sum_{n=1}^N \kappa_n^2 + \kappa,$$

$\kappa_n = L_n/\mu$, $\kappa = L/\mu$ are the condition numbers, d^* is the effective dimensionality defined in Theorem 2, $\bar{\boldsymbol{z}}_n$ is a stationary point of ℓ_n, and $\boldsymbol{m}_n^*, \boldsymbol{C}_n^*$ are part of $\boldsymbol{\lambda}^*$.

Proof. First,

1. Assumption 3 is satisfied as discussed in Appendix B.4.1,
2. Assumption 4 is satisfied by Lemma 12, while
3. Assumption 5 is satisfied by Theorem 2.

Furthermore, the constants are

$$\mathcal{L} = \frac{N}{M}\left(d^* + k_\varphi\right)\frac{\sum_{n=1}^N L_n^2}{\mu} + L \qquad \text{and} \qquad \sigma^2 = \frac{N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^N L_n^2\left(\|\boldsymbol{m}_n^* - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n^*\|_{\text{F}}^2\right).$$

Therefore, we can apply the results of Lemma 8 and consequently Corollary 3, which states that an ϵ-accurate solution can be achieved by a number of iterations of at least

$$\begin{aligned}
T &\geq \max\left(\frac{4\sigma^2}{\mu^2}\frac{1}{\epsilon}, \frac{2\mathcal{L}}{\mu}, 1\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\,\frac{1}{\epsilon}\right)\\
&= \max\left(\frac{4N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^N \frac{L_n^2}{\mu^2}\left(\|\boldsymbol{m}_n^* - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n^*\|_{\text{F}}^2\right)\frac{1}{\epsilon},\ \frac{2N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^N \frac{L_n^2}{\mu^2} + \frac{L}{\mu},\ 1\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\,\frac{1}{\epsilon}\right),
\end{aligned}$$

noticing that the second entry of the max function is always larger than 1, we have

$$= \max\left(\frac{4N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^N \frac{L_n^2}{\mu^2}\left(\|\boldsymbol{m}_n^* - \bar{\boldsymbol{z}}_n\|_2^2 + \|\boldsymbol{C}_n^*\|_{\text{F}}^2\right)\frac{1}{\epsilon},\ \frac{2N}{M}\left(d^* + k_\varphi\right)\sum_{n=1}^N \frac{L_n^2}{\mu^2} + \frac{L}{\mu}\right)\log\left(2\|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|_2^2\,\frac{1}{\epsilon}\right).$$

□

B.5. Properties of the Non-Standardized Parameterization (Theorem 5)

Theorem 5. *Let Assumption 2 hold. Then, under the non-standardized parameterization, there exists a strongly log-concave posterior for which the ELBO is not convex.*

Proof. Our proof is based on a negative example of the global convexity of the energy $\boldsymbol{\lambda} \mapsto \mathbb{E}\ell\left(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)\right)$. This directly implies that the ELBO may not be convex even if the posterior is log-concave.

The non-standardized parameterization applies reparameterization as

$$\boldsymbol{z} = \boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}} + \boldsymbol{m}_{\boldsymbol{z}} \qquad \text{and} \qquad \boldsymbol{y} \mid \boldsymbol{z} = \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}} + \boldsymbol{m}_{\boldsymbol{y}} + \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{z} \tag{18}$$

Now, consider the negative log-joint likelihood $\ell(\boldsymbol{z},\boldsymbol{y}) = \|\boldsymbol{z}\|_2^2 + \|\boldsymbol{y}\|_2^2$, which corresponds to a standard Gaussian posterior. Naturally, the corresponding (normalized) posterior is strongly convex. This model can also be viewed as a 2-level hierarchical model with a single data point such that $N = 1$.

Now, the energy can be computed as

$$\begin{aligned}
\mathbb{E}\ell(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)) &= \mathbb{E}\|\boldsymbol{z}\|_2^2 + \|\boldsymbol{y}\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}} + \boldsymbol{m}_{\boldsymbol{z}}\|_2^2 + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}} + \boldsymbol{m}_{\boldsymbol{y}} + \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{z}\|_2^2 && \text{(Eq. (18))} \\
&= \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}} + \boldsymbol{m}_{\boldsymbol{z}}\|_2^2 + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}} + \boldsymbol{m}_{\boldsymbol{y}} + \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\left(\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}} + \boldsymbol{m}_{\boldsymbol{z}}\right)\|_2^2. && \text{(Eq. (18))}
\end{aligned}$$

For clarity, we will set $\boldsymbol{m}_{\boldsymbol{z}} = \boldsymbol{0}$ and $\boldsymbol{m}_{\boldsymbol{y}} = \boldsymbol{0}$. Then,

$$\mathbb{E}\ell(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)) = \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}} + \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2.$$

For the first term,

$$\mathbb{E}\|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 = \|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\|_{\mathrm{F}}^2. \qquad \text{(Assumption 1 and Lemma 3)}$$

And for the second term,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}} + \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 &= \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}}\|_2^2 + 2\mathbb{E}\left\langle \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}}, \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\right\rangle + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 \\
&= \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}}\|_2^2 + 2\left\langle \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\mathbb{E}\boldsymbol{u}_{\boldsymbol{y}}, \boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\mathbb{E}\boldsymbol{u}_{\boldsymbol{z}}\right\rangle + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 && \text{(since } \boldsymbol{u}_{\boldsymbol{y}} \perp\!\!\!\perp \boldsymbol{u}_{\boldsymbol{z}}\text{)} \\
&= \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\boldsymbol{u}_{\boldsymbol{y}}\|_2^2 + \mathbb{E}\|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}}\|_2^2 && \text{(Assumption 1)} \\
&= \|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\|_{\mathrm{F}}^2. && \text{(Assumption 1 and Lemma 3)}
\end{aligned}$$

Therefore,

$$\mathbb{E}\ell(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)) = \|\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{y}}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}_{\boldsymbol{y},\boldsymbol{z}}\boldsymbol{C}_{\boldsymbol{z},\boldsymbol{z}}\|_{\mathrm{F}}^2.$$

Now, consider the case where $d_{\boldsymbol{y}} = 1$ and $d_{\boldsymbol{z}} = 1$. Then, the scale matrices are all scalars such that

$$\mathbb{E}\ell(\mathcal{T}_{\boldsymbol{\lambda}}\left(\boldsymbol{u}\right)) = C_{z,z}^2 + C_{y,y}^2 + \left(C_{y,z}C_{z,z}\right)^2.$$

The convexity of this function with respect to $(C_{z,z}, C_{y,y}, C_{y,z})$ is equivalent to the convexity of

$$f(x,y,z) = x^2 + z^2 + x^2y^2$$

on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$. Unfortunately, this function is not convex: notice that the Hessian determinant is given as

$$\det \nabla^2 f\left(x,y,z\right) = 8x^2(1-3y^2),$$

which is negative for some $y \in \mathbb{R}_+$. Specifically, for $0 < y < 1/\sqrt{3}$. The fact that the determinant is negative implies that, for this region, some of the eigenvalues of $\nabla^2 f$ are negative, ruling out convexity.

□

C. Probabilistic Models and Datasets

C.1. Robust Poisson Regression

We use a robustified version of Poisson regression for modeling count data. This model is known as the Poisson-log-normal model (Cameron & Trivedi, 2013, §4.2.4), which is a "localized" version of regular Poisson regression (Wang & Blei, 2018, §3.2). That is, an additional hierarchy is added to model the local variations of each datapoint. A closely related model is negative binomial regression, which is obtained by setting a conjugate prior to the local noise. Here, the noise is modeled to be log-normal, resulting in a non-conjugate likelihood:

$$\begin{aligned} \eta_i &\sim \mathcal{N}\left(\boldsymbol{x}_i\boldsymbol{\beta} + \alpha, \sigma_\eta\right) \qquad i = 1, \dots, n \\ y_i &\sim \mathsf{Poisson}\left(\exp\left(\eta_i\right)\right) \quad i = 1, \dots, n, \end{aligned}$$

where $(\eta_i)_{i=1}^n$ are the local variables. The global variables are given by the priors

$$\begin{aligned} \sigma_\alpha &\sim \mathsf{Student\text{-}t}_+\left(4, 0, 1\right) \\ \sigma_\beta &\sim \mathsf{Student\text{-}t}_+\left(4, 0, 1\right) \\ \sigma_\eta &\sim \mathsf{Student\text{-}t}_+\left(4, 0, 1\right) \\ \alpha &\sim \mathcal{N}\left(0, \sigma_\alpha\right) \\ \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\beta\right). \end{aligned}$$

We use the rwm5yr German health registry doctor visit dataset (Hilbe, 2011) from the `COUNT` package in R (Hilbe, 2016).

C.2. Item Response Theory

Item response theory (IRT) is a family of models for estimating the response of humans to a set of items, often in the form of exams or questionnaires (Lord et al., 2008). BBVI has recently been shown to be very successful in estimating human ability from large-scale educational examination datasets (Wu et al., 2020). We employ the so-called two-parameter logistic model, or "2PL" model, for which the likelihood is given as

$$\begin{aligned} \mathrm{logit}_i &= \gamma_{\mathrm{item}_i}\alpha_{\mathrm{student}_i} + \beta_{\mathrm{item}_i} + \mu_\beta \\ y_i &\sim \mathsf{Bernoulli\text{-}logit}\left(\mathrm{logit}_i\right) \qquad i = 1, \dots, n. \end{aligned}$$

The global variables are given as

$$\begin{aligned} \mu_\beta &\sim \mathsf{Student\text{-}t}\left(4, 0, 1\right) \\ \sigma_\beta &\sim \mathsf{Student\text{-}t}_+\left(4, 0, 1\right) \\ \sigma_\gamma &\sim \mathsf{Student\text{-}t}_+\left(4, 0, 1\right) \\ \gamma_k &\sim \mathsf{log\text{-}normal}\left(0, \sigma_\gamma\right) \quad k = 1, \dots, K, \end{aligned}$$

while the local variables are

$$\alpha_j \sim \mathcal{N}\left(0, 1\right) \quad j = 1, \dots, J.$$

The log-normal prior on γ is inspired by Patz & Junker (1999).

While here we only consider scaling with respect to the students, we can also consider scaling with respect to the number of items by also making $q\left(\boldsymbol{\beta}\right)$ and $q\left(\boldsymbol{\gamma}\right)$ factor out. While this is less important for our dataset of choice, which has a small K, this is certainly attractive to other datasets where both K and J are large.

For the dataset, we take CritLangAcq from Wu et al. (2020). Unfortunately, this dataset is too large to fit in memory even for the mean-field approximation. Therefore, we only use a 5% random subset of the full dataset. Scaling to the full dataset will require additional strategies amortization (Kingma & Welling, 2014; Dayan et al., 1995) as done in the original work by Wu et al..

C.3. Multivariate Stochastic Volatility

Multivariate stochastic volatility (Chib et al., 2009) The likelihood is given as

$$\begin{aligned}
\boldsymbol{y}_1 &\sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}\right) \\
\boldsymbol{y}_t &\sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\phi}\left(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}\right), \boldsymbol{Q}\right) \\
\boldsymbol{x}_t &\sim \mathcal{N}\left(0, \exp\left(\boldsymbol{y}_t / 2\right)\right),
\end{aligned}$$

where $(\boldsymbol{y}_t)_{t=1}^{T}$, the latent stochastic volatilities, are the local variables. For the hyperpriors, we develop a fully Bayesian variant of the model used by Naesseth et al. (2018, §5) who perform empirical Bayes inference on the hyperparameters. Notably, following modern Bayesian modeling practice (Gelman et al., 2020), we assign a Cauchy-LKJ prior to the covariance $\boldsymbol{Q}$. The global variables are given as

$$\begin{aligned}
\boldsymbol{L}_{\boldsymbol{\Sigma}} &\sim \text{LKJ-Cholesky}\left(d_{\boldsymbol{y}}, 1\right) \\
\boldsymbol{\tau} &\sim \text{Cauchy}_{+}\left(0, 5\right) \\
\boldsymbol{L}_{\boldsymbol{Q}} &= \operatorname{diag}\left(\boldsymbol{\tau}\right) \boldsymbol{L}_{\boldsymbol{\Sigma}} \\
\boldsymbol{Q} &= \boldsymbol{L}_{\boldsymbol{Q}} \boldsymbol{L}_{\boldsymbol{Q}}^{\top} \\
\boldsymbol{\mu} &\sim \text{Cauchy}\left(0, 10\right) \\
\boldsymbol{\phi} &\sim \text{uniform}\left(-1, 1\right),
\end{aligned}$$

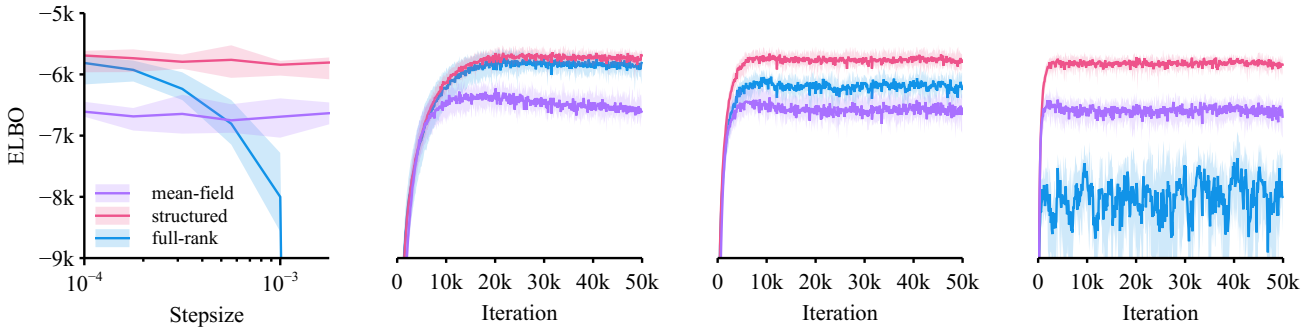
where all vector operations are elementwise.

For the datasets, we use the exchange rate ("FX") between 6 international currencies and the U.S. dollar. In particular, we use the daily closing exchange rate of EUR, JPY, GBP, AUD, CAD, and KRW over the period from 2006-05-16 to 2023-08-30. For the subsets, we deterministically slice a continuous period starting from 2006-05-16.

D. Additional Experimental Results

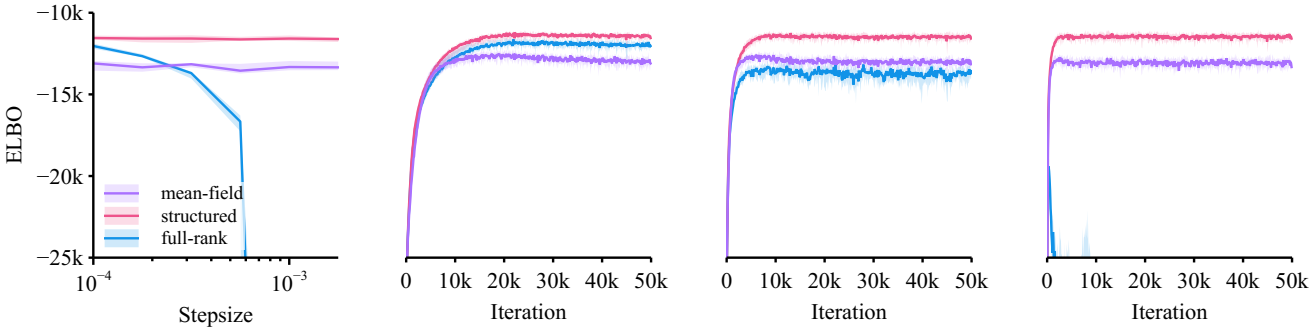
This section shows additional plots for the experimental results displayed in Section 4.2. In particular, we show convergence plots with respect to the number of iterations for a fixed stepsize. Note that all methods use the same number of gradient evaluations per iteration. Therefore, "iteration" is synonymous with "number of gradient queries."

D.1. Results on rpoisson



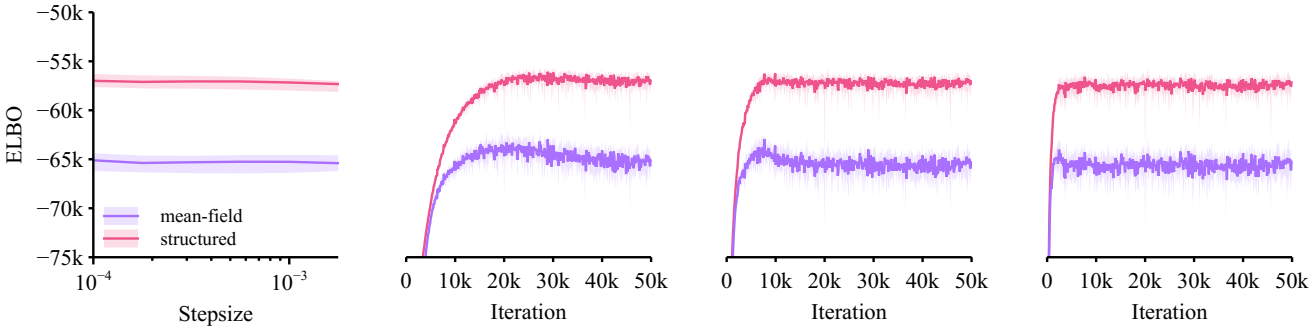
(a) $T = 50\text{k}$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

Figure 5: **ELBO versus stepsize on rpoisson-small** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications. Notice that the performance gap between **full-rank** and **structured** becomes narrower as we reduce the stepsize.



(a) $T = 50\text{k}$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

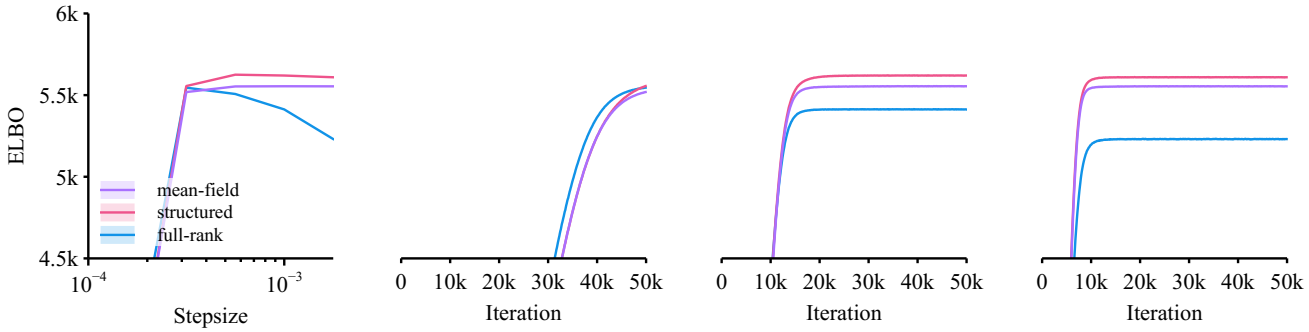
Figure 6: **ELBO versus stepsize on rpoisson-middle** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications. Notice that the performance gap between **full-rank** and **structured** becomes narrower as we reduce the stepsize.



(a) $T = 50\text{k}$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

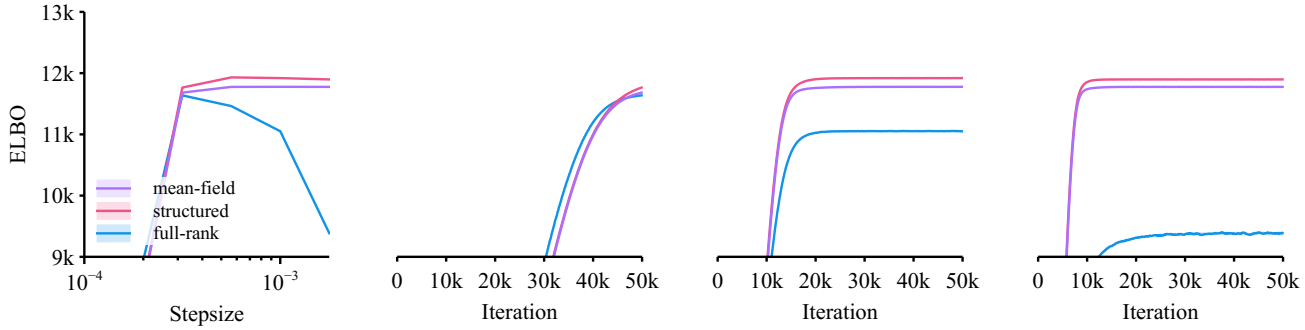
Figure 7: **ELBO versus stepsize on rpoisson-large) Full-rank** is omitted as it didn't fit in memory. The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications.

D.2. Results on volatility



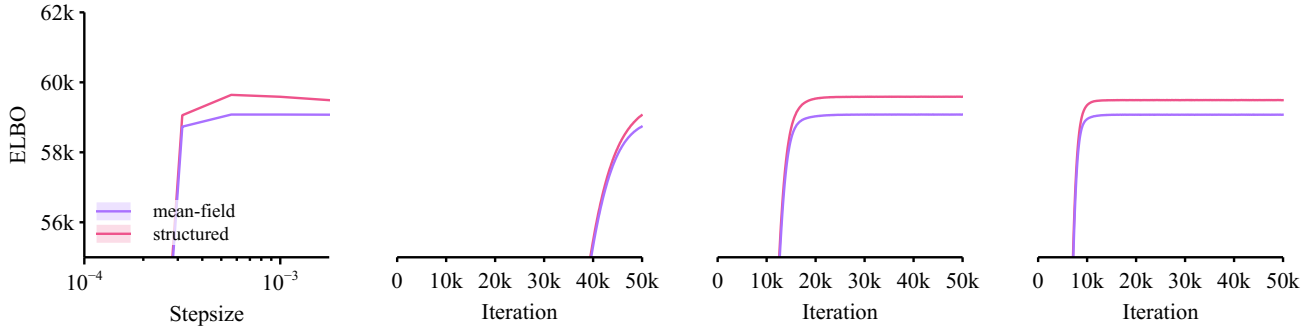
(a) $T = 50\text{k}$ (b) $\gamma = 10^{-3.5}$ (c) $\gamma = 10^{-3}$ (d) $\gamma = 10^{-2.75}$

Figure 8: **ELBO versus stepsize on volatility-small** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications.



(a) $T = 50\text{k}$ (b) $\gamma = 10^{-3.5}$ (c) $\gamma = 10^{-3}$ (d) $\gamma = 10^{-2.75}$

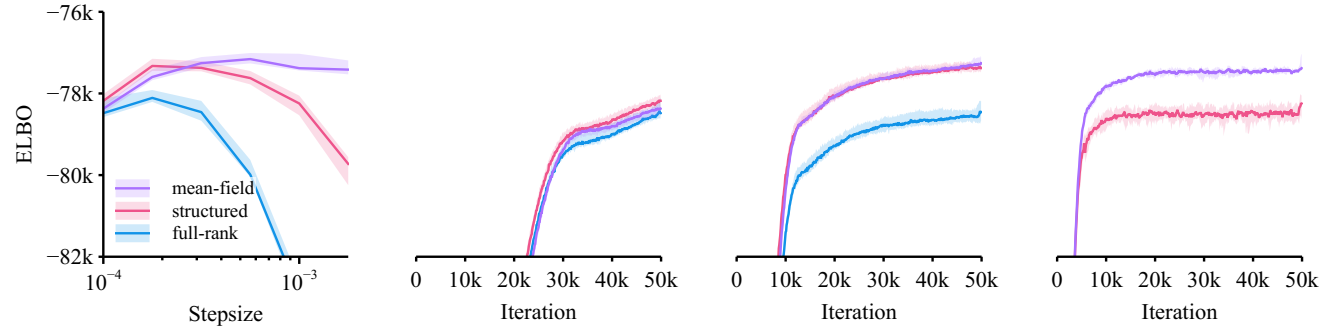
Figure 9: **ELBO versus stepsize on volatility-middle** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications.



(a) $T = 50\text{k}$ (b) $\gamma = 10^{-3.5}$ (c) $\gamma = 10^{-3}$ (d) $\gamma = 10^{-2.75}$

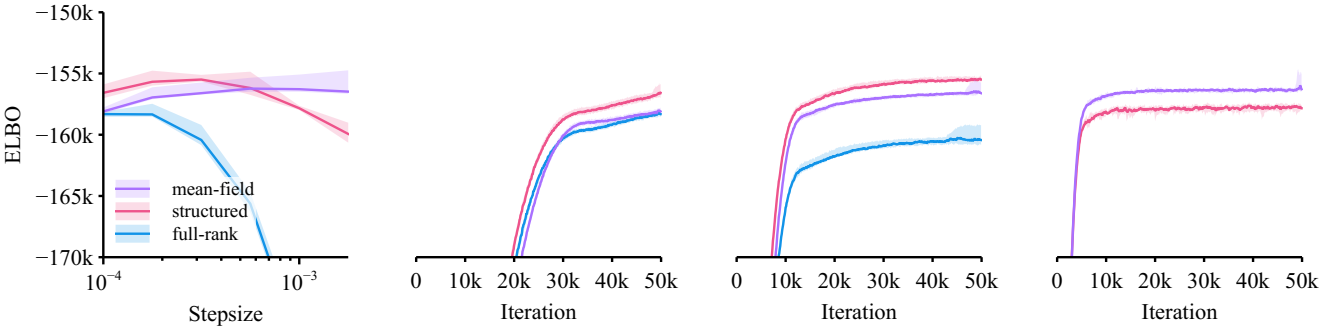
Figure 10: **ELBO versus stepsize on volatility-large.** Full-rank is omitted as it didn't fit in memory. The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications.

D.3. Results on irt



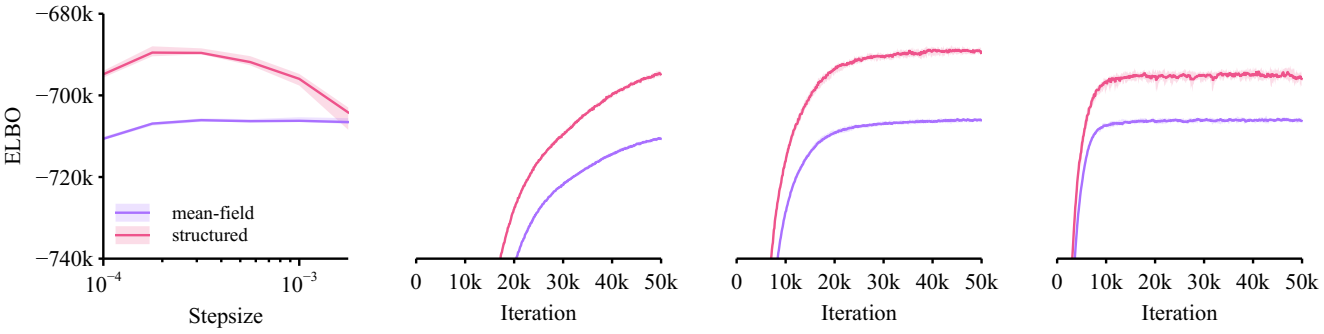
(a) $T = 50k$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

Figure 11: **ELBO versus stepsize on irt-small.** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications. Notice that **structured** performs worse than mean-field at the largest stepsize ($\gamma = 10^{-3}$), but becomes comparable as we reduce the stepsize.



(a) $T = 50k$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

Figure 12: **ELBO versus stepsize on irt-middle.** The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications. Notice that **structured** performs slightly worse than **mean-field** at the largest stepsize ($\gamma = 10^{-3}$), but becomes superior as we reduce the stepsize.



(a) $T = 50k$ (b) $\gamma = 10^{-4}$ (c) $\gamma = 10^{-3.5}$ (d) $\gamma = 10^{-3}$

Figure 13: **ELBO versus stepsize on irt-large**. Full-rank is omitted as it didn't fit in memory. The solid lines are the median, while the shaded regions are the 80% quantiles computed from 4 independent replications.