Understanding Stochastic Natural Gradient Variational Inference

Kaiwen Wu [1] **Jacob R. Gardner** [1]

Abstract

Stochastic natural gradient variational inference (NGVI) is a popular posterior inference method with applications in various probabilistic models. Despite its wide usage, little is known about the non-asymptotic convergence rate in the *stochastic* setting. We aim to lessen this gap and provide a better understanding. For conjugate likelihoods, we prove the first $\mathcal{O}(\frac{1}{T})$ non-asymptotic convergence rate of stochastic NGVI. The complexity is no worse than stochastic gradient descent (*a.k.a.* black-box variational inference) and the rate likely has better constant dependency that leads to faster convergence in practice. For non-conjugate likelihoods, we show that stochastic NGVI with the canonical parameterization implicitly optimizes a non-convex objective. Thus, a global convergence rate of $\mathcal{O}(\frac{1}{T})$ is unlikely without some significant new understanding of optimizing the ELBO using natural gradients.

1. Introduction

Given a prior $p(\mathbf{z})$ and a likelihood $p(\mathbf{y} \mid \mathbf{z})$, variational inference (VI) approximates the posterior $p(\mathbf{z} \mid \mathbf{y})$ by optimizing the evidence lower bound (ELBO) in a family of variational distributions (Blei et al., 2017). Natural gradient variational inference (NGVI), in particular, optimizes the ELBO by natural gradient descent (NGD) (Amari, 1998). Different from (standard) gradient descent that follows the steepest descent direction induced by the Euclidean distance, NGD follows the steepest descent direction induced by the KL divergence (Honkela & Valpola, 2004; Hensman et al., 2012; Hoffman et al., 2013). The folk wisdom is that the KL divergence is a better "metric" to compare distributions and thus NGD is believed to be superior than gradient descent, *a.k.a.* black-box variational inference (Ranganath et al., 2014). Indeed, NGVI as well as its variants empirically outperforms gradient descent in many cases, and thus enjoys applications in a wide range of probabilistic models. Here, we name a few exmaples: latent Dirichlet allocation topic models (Hoffman et al., 2013), Bayesian neural networks (Khan et al., 2018; Osawa et al., 2019), and large-scale Gaussian processes (Hensman et al., 2013; 2015; Salimbeni et al., 2018).

Despite its wide usage, a non-asymptotic convergence rate of NGVI in the stochastic setting is absent, even for simple conjugate likelihoods. A few convergence arguments exist in the literature, but none of them applies to any practical uses of NGVI. For example, Hoffman et al. (2013) have a convergence argument [1] by assuming the Fisher information matrix has eigenvalues bounded from below (by a positive constant) throughout the natural gradient updates. Khan et al. (2016) analyze a variant of NGVI based on Bregman proximal gradient descent by assuming the (KL) divergence is α-strongly convex, a condition that generally does not hold (at least for the KL divergence). Besides, Khan et al. (2016) did not obtain a complexity bound in the stochastic setting—they only showed convergence to a region around stationary points. Note that these assumptions do not hold in the entire domain, provably. Even if they hold in a subset of the domain, the constants in these assumptions are difficult to estimate, and might even be arbitrarily bad as the posterior distribution $p(\mathbf{z} \mid \mathbf{y})$ contracts.[2]

This work aims to lessen this gap and obtain a "clean" analysis, with minimal assumptions, that is applicable to some practical uses of stochastic NGVI. For the sake of generality, existing analyses have to use assumptions that does not hold in practice. Therefore, we pursue the opposite direction of generality—the basic setting of conjugate likelihoods, for which we establish the first $\mathcal{O}(\frac{1}{T})$ non-asymptotic convergence rate of stochastic natural gradient variational inference. This rate has the same complexity as the convergence rate of stochastic projected (and proximal) gradient descent recently studied by Domke (2020); Domke et al. (2023); Kim et al. (2023). This, along with our experiments, implies that NGVI ultimately may share the same

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States. Correspondence to: Kaiwen Wu <kaiwenwu@seas.upenn.edu>, Jacob R. Gardner <jacobrg@seas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Hoffman et al. (2013) did not give a convergence proof besides a reference to Bottou (1998).

[2]For instance, the Fisher information matrix gets increasingly close to singular as the covariance of the posterior $p(\mathbf{z} \mid \mathbf{y})$ shrinks.

complexity with other first-order methods. The empirical observation that NGVI is faster than stochastic gradient descent is likely due to a better constant dependency in the big $\mathcal{O}$ notation. Indeed, as we will see later, our convergence rate of stochastic NGVI is independent of the objective's condition number and the distance from the initialization to the optimum. Nevertheless, the constant improvement may play a huge difference in practice.

Although our convergence rate for stochastic NGVI assumes conjugate likelihoods, it is already applicable to some practical uses, including large-scale Bayesian linear regression and variational parameter learning in stochastic variational Gaussian processes (Hensman et al., 2013; 2015; Salimbeni et al., 2018). Indeed, we will show that all assumptions are strictly satisfied in practice and the constant in the convergence rate can be bounded explicitly using statistics from the training data.

For non-conjugate likelihoods, we show that the "canonical" implementation of stochastic NGVI implicitly optimizes a non-convex objective even when the likelihoods are simple log-concave distributions. Hence, the convergence behavior of stochastic NGVI with non-conjugate likelihoods is more nuanced, which might partially explain why the theoretical understanding of stochastic NGVI is lacking throughout the years. This lack of convexity implies that proving a global convergence rate of $\mathcal{O}(\frac{1}{T})$ for non-conjugate likelihoods may require new properties of the ELBO, e.g., the Polyak-Łojasiewicz inequality (Polyak et al., 1963; Lojasiewicz, 1963), in order to explain the empirical success of stochastic NGVI for non-conjugate likelihoods (e.g., Hoffman et al., 2013; Salimbeni et al., 2018).

2. Background

Notation. We use $\|\cdot\|$ to denote the vector Euclidean norm. For matrices, the same symbol $\|\cdot\|$ is overloaded to denote the spectral norm. $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. $\langle\cdot,\cdot\rangle$ denotes an inner product, whose domain is inferred from its arguments. Let $\mathsf{D}_{\mathrm{KL}}(\cdot,\cdot)$ denote the Kullback–Leibler divergence between distributions. $\mathbb{S}^d_{++}$ (and $\mathbb{S}^d_{+}$) represents the collection of all $d\times d$ symmetric positive (semi-)definite matrices. Let $\succ$ (and $\succeq$) be the partial order induced by $\mathbb{S}^d_{++}$ (and $\mathbb{S}^d_{+}$), i.e., $\mathbf{A}\succ\mathbf{B}$ if and only if $\mathbf{A}-\mathbf{B}\in\mathbb{S}^d_{++}$.

2.1. Variational Inference with Exponential Families

Suppose we have a prior $p(\mathbf{z})$ on latent variables $\mathbf{z}$ and a likelihood $p(\mathbf{y}\mid\mathbf{z})$ on observations $\mathbf{y}$. Variational inference (VI) aims to find the best approximation of the posterior $p(\mathbf{z}\mid\mathbf{y})$ inside a variational family $\mathcal{Q}$ by minimizing the Kullback–Leibler (KL) divergence

$$\operatorname*{minimize}_{q\in\mathcal{Q}} \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}),p(\mathbf{z}\mid\mathbf{y})),$$

where q is the variational distribution. This is the equivalent to minimizing the objective

$$\ell(q)=-\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}\mid\mathbf{z})]+\mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}),p(\mathbf{z})), \tag{1}$$

which is called the negative evidence lower bound (ELBO). Throughout the paper, we assume the variational family $\mathcal{Q}$ is an exponential family (which will be defined below), and the prior $p(\mathbf{z})$ is in $\mathcal{Q}$. Though, the posterior $p(\mathbf{z}\mid\mathbf{y})$ is not necessarily in $\mathcal{Q}$, unless the likelihood is conjugate: we call the likelihood $p(\mathbf{y}\mid\mathbf{z})$ conjugate (with the prior) if and only if $p(\mathbf{z}\mid\mathbf{y})\in\mathcal{Q}$. Conjugacy implies the variational approximation is exact, so long as (1) is minimized globally.

Exponential Family. A (regular and minimal) exponential family is a collection of distributions indexed by a canonical parameter $\boldsymbol{\eta}$ in the form

$$q(\mathbf{z};\boldsymbol{\eta})=h(\mathbf{z})\exp\big(\langle\phi(\mathbf{z}),\boldsymbol{\eta}\rangle-A(\boldsymbol{\eta})\big), \tag{2}$$

where h is the base measure, ϕ is the sufficient statistic, $\boldsymbol{\eta}$ is the natural parameter, and A is the log-partition function.

The set of all possible $\boldsymbol{\eta}$ that make $q(\mathbf{z};\boldsymbol{\eta})$ integrable forms an open convex set $\mathcal{D}$, called the natural parameter space. The log-partition function $A:\mathcal{D}\to\mathbb{R}$ is differentiable and strictly convex on $\mathcal{D}$. The associated expectation parameter $\boldsymbol{\omega}$ of $q(\mathbf{z};\boldsymbol{\eta})$ is defined as the expected sufficient statistic:

$$\boldsymbol{\omega}=\mathbb{E}_{q(\mathbf{z};\boldsymbol{\eta})}[\phi(\mathbf{z})]. \tag{3}$$

The set of all possible expectation parameters $\boldsymbol{\omega}$ again forms a convex set Ω, called the expectation parameter space.

The natural and expectation parameter spaces, $\mathcal{D}$ and Ω, are linked by the gradients of the log-partition function A and its convex conjugate A^*, where the differentiable and strictly convex function $A^*:\Omega\to\mathbb{R}$ is defined as

$$A^*(\boldsymbol{\omega})=\max_{\boldsymbol{\eta}\in\mathcal{D}}\langle\boldsymbol{\eta},\boldsymbol{\omega}\rangle-A(\boldsymbol{\eta}).$$

Indeed, the gradient maps $\nabla A:\mathcal{D}\to\Omega$ and $\nabla A^*:\Omega\to\mathcal{D}$ are inverses of each other. Namely, if $\boldsymbol{\eta}\in\mathcal{D}$ and $\boldsymbol{\omega}\in\Omega$ satisfy (3) representing the same distribution, then

$$\nabla A(\boldsymbol{\eta})=\boldsymbol{\omega},\quad \nabla A^*(\boldsymbol{\omega})=\boldsymbol{\eta}. \tag{4}$$

Example. A d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ has its natural parameter $\boldsymbol{\eta}=(\boldsymbol{\lambda},\boldsymbol{\Lambda})$ defined on

$$\mathcal{D}=\{\boldsymbol{\eta}=(\boldsymbol{\lambda},\boldsymbol{\Lambda})\in\mathbb{R}^d\times\mathbb{R}^{d\times d}:-\boldsymbol{\Lambda}\in\mathbb{S}^d_{++}\} \tag{5}$$

with the parameter conversion identity $\boldsymbol{\lambda}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}=-\frac{1}{2}\boldsymbol{\Sigma}^{-1}$. The expectation parameter $\boldsymbol{\omega}$ is defined on

$$\Omega=\{\boldsymbol{\omega}=(\boldsymbol{\xi},\boldsymbol{\Xi})\in\mathbb{R}^d\times\mathbb{R}^{d\times d}:\boldsymbol{\Xi}-\boldsymbol{\xi}\boldsymbol{\xi}^\top\in\mathbb{S}^d_{++}\} \tag{6}$$

with the identity $\boldsymbol{\xi}=\boldsymbol{\mu}$ and $\boldsymbol{\Xi}=\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\top$. See §A for more details, where we give explicit expressions for $A(\boldsymbol{\eta})$, $A^*(\boldsymbol{\omega})$, $\nabla A(\boldsymbol{\eta})$ and $\nabla A^*(\boldsymbol{\omega})$. For more background on exponential families, we direct readers to the monograph by Wainwright & Jordan (2008).

Remark 1 (Overload ℓ). *Let $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ be the natural and expectation parameters of the same distribution q. We have*

$$\ell(q) = \ell^{(\mathrm{n})}(\boldsymbol{\eta}) = \ell^{(\mathrm{e})}(\boldsymbol{\omega}),$$

where $\ell^{(\mathrm{n})}$ and $\ell^{(\mathrm{e})}$ are the negative ELBO as functions of the natural and expectation parameters respectively. Technically, ℓ, $\ell^{(\mathrm{n})}$ and $\ell^{(\mathrm{e})}$ are different functions with different domains and arguments. For notation simplicity, however, we will drop the superscript when the context allows—the superscript will be inferred from the argument.

2.2. Natural Gradient Descent for Variational Inference

Natural gradient variational inference (NGVI) optimizes the ELBO by natural gradient descent (NGD). It iteratively updates the natural parameter $\boldsymbol{\eta}$ of the variational distribution by taking a steepest descent step induced by the KL divergence. The yielded update rule is a preconditioned gradient descent with the Fisher information matrix (FIM).

Definition 1 (FIM). *Given a (not necessarily exponential family) distribution $q(\mathbf{z}; \boldsymbol{\eta})$ parameterized by $\boldsymbol{\eta}$, the Fisher information matrix is defined as*

$$\mathbf{F}(\boldsymbol{\eta}) = -\mathbb{E}_{q(\mathbf{z})}[\nabla^2_{\boldsymbol{\eta}} \log q(\mathbf{z}; \boldsymbol{\eta})],$$

where ∇^2 is taken w.r.t. $\boldsymbol{\eta}$. In particular, for the exponential family (2), it takes a simple form $\mathbf{F}(\boldsymbol{\eta}) = \nabla^2 A(\boldsymbol{\eta})$.

Definition 2 (NGD). *Natural gradient descent iterates*

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - \gamma_t \mathbf{F}(\boldsymbol{\eta}_t)^{-1} \nabla \ell(\boldsymbol{\eta}_t), \tag{7}$$

where γ_t is the step size and $\mathbf{F}(\boldsymbol{\eta})$ is the Fisher information matrix. The FIM-preconditioned gradient $\mathbf{F}(\boldsymbol{\eta}_t)^{-1}\nabla \ell(\boldsymbol{\eta}_t)$ is often called the "natural" gradient.

We call the update (7) the "canonical" NGD update, as NGD can be implemented in other parameters beyond the natural parameter $\boldsymbol{\eta}$. However, typically NGD converges the fastest in the natural parameterization, which will be the main focus of this paper. We will revisit other parameterizations in §5.

Explicitly inverting the FIM is inefficient, e.g., its takes $\mathcal{O}(d^6)$ time for a Gaussian due to its $d(d+1) \times d(d+1)$ size. Fortunately, the NGD update can be implemented without explicit FIM inversion for the exponential family (Raskutti & Mukherjee, 2015). Let $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ be the natural and expectation parameters of the same distribution, hence having the same ELBO value $\ell^{(\mathrm{n})}(\boldsymbol{\eta}) = \ell^{(\mathrm{e})}(\boldsymbol{\omega})$. Plugging in the identity $\boldsymbol{\omega} = \nabla A(\boldsymbol{\eta})$ as in (4), we obtain

$$\ell^{(\mathrm{n})}(\boldsymbol{\eta}) = \ell^{(\mathrm{e})}(\nabla A(\boldsymbol{\eta})).$$

Differentiating *w.r.t.* $\boldsymbol{\eta}$ on both sides gives

$$\begin{aligned}\nabla \ell^{(\mathrm{n})}(\boldsymbol{\eta}) &= \nabla^2 A(\boldsymbol{\eta}) \cdot \nabla \ell^{(\mathrm{e})}(\nabla A(\boldsymbol{\eta})) \\ &= \mathbf{F}(\boldsymbol{\eta}) \nabla \ell^{(\mathrm{e})}(\boldsymbol{\omega}),\end{aligned}$$

which implies $\mathbf{F}(\boldsymbol{\eta})^{-1}\nabla \ell^{(\mathrm{n})}(\boldsymbol{\eta}) = \nabla \ell^{(\mathrm{e})}(\boldsymbol{\omega})$: the natural gradient (of the natural parameter) is simply the gradient of the (negative) ELBO *w.r.t.* the expectation parameter. Thus, the NGD update rule (7) reduces to

$$\boldsymbol{\eta}_{t+1} = \boldsymbol{\eta}_t - \gamma_t \nabla \ell(\boldsymbol{\omega}_t), \tag{8}$$

with no explicit FIM inversion.

2.3. Natural Gradient Descent as Mirror Descent

This section reviews the connection between NGD and mirror descent (MD). This connection was discovered by Raskutti & Mukherjee (2015) and later applied to variational inference by Khan & Lin (2017); Khan et al. (2018).

Definition 3 (Bregman Divergence). *Given a differentiable and strictly convex function Φ, the associated Bregman divergence is defined as*

$$\mathsf{D}_\Phi(\mathbf{a}, \mathbf{b}) = \Phi(\mathbf{a}) - \Phi(\mathbf{b}) - \langle \nabla \Phi(\mathbf{b}), \mathbf{a} - \mathbf{b} \rangle,$$

and we call Φ the distance generating function.

Recall that A^*, as the convex conjugate of the log-partition function A, is differentiable and strictly convex on Ω. Thus, it is a valid distance generating function and induces a Bregman divergence D_{A^*} on the expectation parameters. The divergence is then used to define mirror descent (Nemirovskij & Yudin, 1983), which iteratively solves regularized first-order approximations.

Definition 4 (MD). *The mirror descent update is defined as*

$$\boldsymbol{\omega}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\omega} \in \Omega} \langle \nabla \ell(\boldsymbol{\omega}_t), \boldsymbol{\omega} \rangle + \frac{1}{\gamma_t} \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_t), \tag{9}$$

where $\gamma_t > 0$ is the step size.

Mirror descent (MD) is a generalization of gradient descent. If the Bregman divergence $\mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_t)$ in (9) is replaced with the squared Euclidean norm $\frac{1}{2}\|\boldsymbol{\omega} - \boldsymbol{\omega}_t\|^2$, we recover the familiar update rule $\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \gamma_t \nabla \ell(\boldsymbol{\omega}_t)$.

To implement MD efficiently, the minimization in (9) needs to be solved in closed-form. Taking the derivative *w.r.t.* $\boldsymbol{\omega}$ on both sides and setting it equal to zero, we obtain

$$\nabla A^*(\boldsymbol{\omega}_{t+1}) = \nabla A^*(\boldsymbol{\omega}_t) - \gamma_t \nabla \ell(\boldsymbol{\omega}_t). \tag{10}$$

Recall $\nabla A^* : \Omega \to \mathcal{D}$ maps the expectation parameter $\boldsymbol{\omega}$ of an exponential family distribution q to its natural parameter $\boldsymbol{\eta}$, i.e., $\nabla A^*(\boldsymbol{\omega}_t) = \boldsymbol{\eta}_t$ for all $t \geq 0$. Thus, the MD update (10) recovers the NGD update (8) exactly.

The discussion in this section so far is summarized in the lemma below. In particular, we will not distinguish between NGD and MD in the rest of the paper.

Lemma 1 (NGD = MD). *Suppose the NGD update (7) and the MD update (9) start from the same variational distribution q_0, i.e., $\boldsymbol{\eta}_0 = \nabla A^*(\boldsymbol{\omega}_0)$. Then, we have $\boldsymbol{\eta}_t = \nabla A^*(\boldsymbol{\omega}_t)$ for all $t \geq 0$. Namely, NGD and MD produce exactly the same sequence of variational distributions.*

We introduce a few definitions useful for later proofs. Our results in §4 are built upon casting NGD as a special case of MD and utilizing the recent developments of stochastic MD for relatively smooth and relatively strongly convex functions (Birnbaum et al., 2011; Bauschke et al., 2017; Lu et al., 2018; Hanzely & Richtárik, 2021).

Definition 5. *Let Φ be a differentiable and strictly convex function. A function f is called β-smooth relative to Φ if*

$$f(\mathbf{a}) \leq f(\mathbf{b}) + \langle \nabla f(\mathbf{b}), \mathbf{a} - \mathbf{b} \rangle + \beta \mathsf{D}_\Phi(\mathbf{a}, \mathbf{b})$$

holds for all $\mathbf{a}, \mathbf{b}$ in the domain. A function f is called α-strongly convex relative to Φ if

$$f(\mathbf{a}) \geq f(\mathbf{b}) + \langle \nabla f(\mathbf{b}), \mathbf{a} - \mathbf{b} \rangle + \alpha \mathsf{D}_\Phi(\mathbf{a}, \mathbf{b})$$

holds for all $\mathbf{a}, \mathbf{b}$ in the domain.

Relative smoothness and relative strongly convexity recover the usual definitions of smoothness and strong convexity when $\Phi(\cdot) = \frac{1}{2}\|\cdot\|^2$.

3. Stochastic Natural Gradient VI

This section discusses the implementation of natural gradient variational inference in the stochastic setting. Two types of stochasticity may arise in pratice: (a) the expected log likelihood $\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$ in the ELBO (1) does not have a closed-form for most non-conjugate likelihoods due to the intractable integral, and thus one needs to estimate it stochastically;[3] (b) the expected log likelihood is a finite sum over a large number of training data, and one needs to employ mini-batch stochastic optimization, e.g., Example 1.

Care is required when implementing the update rule (8), or equivalently the update rule (10), in the stochastic setting. Recall that the natural parameter $\boldsymbol{\eta} \in \mathcal{D}$ has a domain. In the stochastic setting, implementing the update rule (8) with stochastic gradients $\widehat{\nabla}\ell(\boldsymbol{\omega}_t) \approx \nabla\ell(\omega_t)$ does not necessarily guarantee that $\boldsymbol{\eta}_{t+1}$ stays inside the domain $\mathcal{D}$, in which case NGD breaks down.

3.1. A Sufficient Condition for Valid NGD Updates

We will give a sufficient condition on the stochastic gradient $\widehat{\nabla}\ell(\boldsymbol{\omega}_t)$ that guarantees the natural parameter $\boldsymbol{\eta}$ always stays inside the domain $\mathcal{D}$. As shown in §A, the KL divergence has a closed-form gradient *w.r.t.* the expectation parameter:

$$\nabla_{\boldsymbol{\omega}} \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})) = \boldsymbol{\eta} - \boldsymbol{\eta}_{\mathrm{p}},$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_{\mathrm{p}}$ are the natural parameters of $q(\mathbf{z})$ and $p(\mathbf{z})$ respectively. Thus, the stochasticity comes solely from stochastically estimating the expected log likelihood:

$$\widehat{\nabla}\ell(\boldsymbol{\omega}) = -\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] + \boldsymbol{\eta} - \boldsymbol{\eta}_{\mathrm{p}}.$$

Plugging it into the NGD update (8), we obtain

$$\boldsymbol{\eta}_{t+1} = (1 - \gamma_t)\boldsymbol{\eta}_t + \gamma_t\big(\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] + \boldsymbol{\eta}_{\mathrm{p}}\big).$$

Recall that the natural parameter space $\mathcal{D}$ is an open convex set. Hence, $\boldsymbol{\eta}_{t+1}$ stays in $\mathcal{D}$ provided that (a) $\gamma_t \in [0, 1]$ and (b) $\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] + \boldsymbol{\eta}_{\mathrm{p}} \in \mathcal{D}$. The first condition is satisfied if the step size γ_t is chosen properly. The second condition is more complicated, but can still be satisfied by carefully constructed stochastic gradient estimators.

3.2. Common Stochastic Gradient Estimators

This section discusses a common special case: (a) the variational family $\mathcal{Q}$ is the collection of all Gaussians; (b) the prior $p(\mathbf{z})$ is a Gaussian; and (c) the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is log-concave in $\mathbf{z}$. In the following, we give two examples of stochastic gradients. One example guarantees valid NGD updates while the other one does not.

For Gaussians, the only constraint on the natural parameter $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is that its second component is negative definite $\boldsymbol{\Lambda} \prec 0$. The sufficient condition for valid stochastic NGD updates in the previous section reduces to the following:

Remark 2. *Suppose that the variational and the prior are both Gaussians. The NGD update (8) is valid for all $t \geq 0$ in the stochastic setting if (a) the step size $\gamma_t \in [0, 1]$ and (b) the stochastic gradient of the expected log likelihood*

$$(\widehat{\nabla}_{\boldsymbol{\xi}}, \widehat{\nabla}_{\boldsymbol{\Xi}}) = \widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z})$$

has its second component negative definite $\widehat{\nabla}_{\boldsymbol{\Xi}} \prec 0$.

Automatic Differentiation. For intractable expected log likelihoods, a simple estimator for their gradients uses the reparameterization trick (Kingma & Welling, 2013; Titsias & Lázaro-Gredilla, 2014; Rezende et al., 2014) and automatic differentiation, shown in Algorithm 1. This stochastic gradient guarantees that $\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_q \log p(\mathbf{y} \mid \mathbf{z})$ is unbiased and symmetric (Murray, 2016), but does not guarantee $\widehat{\nabla}_{\boldsymbol{\Xi}}$ is negative definite. A counterexample is given in §B.

Bonnet's and Price's Gradients. Consider the gradients *w.r.t.* the mean and covariance of a Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. By the Bonnet and Price theorems (Bonnet, 1964; Price, 1958; Opper & Archambeau, 2009), they are

$$\nabla_{\boldsymbol{\mu}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] = \mathbb{E}_{q(\mathbf{z})}[\nabla_{\mathbf{z}} \log p(\mathbf{x} \mid \mathbf{z})],$$

$$\nabla_{\boldsymbol{\Sigma}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] = \frac{1}{2} \mathbb{E}_{q(\mathbf{z})}[\nabla^2_{\mathbf{z}} \log p(\mathbf{x} \mid \mathbf{z})].$$

[3] Though, the expected log likelihood does have a closed-form for some non-conjugate likelihoods. See §5 for an example.

Algorithm 1: Auto Differentiation Stochastic Gradient

Input: $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi})$, the expectation parameter of $q(\mathbf{z})$
Output: $\left(\widehat{\nabla}_{\boldsymbol{\xi}}, \widehat{\nabla}_{\boldsymbol{\Xi}}\right) = \widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$

$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\boldsymbol{\xi}, \boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top\right)$ `// conversion`
$\mathbf{C} = \texttt{cholesky}(\boldsymbol{\Sigma})$
$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{z} = \boldsymbol{\mu} + \mathbf{C}\mathbf{u}$ `//` $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
`loss` $= \log p(\mathbf{y} \mid \mathbf{z})$ `// forward pass`
`loss.backward()` `// compute` $\widehat{\nabla}_{\boldsymbol{\xi}}, \widehat{\nabla}_{\boldsymbol{\Xi}}$

Applying the chain rule through $\boldsymbol{\xi} = \boldsymbol{\mu}$ and $\boldsymbol{\Xi} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}$, and approximating the expectations with samples, we obtain a stochastic gradient $\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$:

$$\begin{aligned} \widehat{\nabla}_{\boldsymbol{\xi}} &= \frac{1}{m} \sum_{i=1}^{m} [\nabla_{\mathbf{z}} \log p(\mathbf{y} \mid \mathbf{z}_i) - \nabla^2_{\mathbf{z}} \log p(\mathbf{y} \mid \mathbf{z}_i) \cdot \boldsymbol{\mu}] \\ \widehat{\nabla}_{\boldsymbol{\Xi}} &= \frac{1}{2m} \sum_{i=1}^{m} \nabla^2_{\mathbf{z}} \log p(\mathbf{y} \mid \mathbf{z}_i) \end{aligned} \tag{11}$$

where $\mathbf{z}_i \sim q$ are *i.i.d.* samples from the variational distribution. While the stochastic gradient $\widehat{\nabla}_{\boldsymbol{\xi}}$ in (11) coincides with the reparameterization trick, the second line is not the same as the stochastic gradient by automatic differentiation in Algorithm 1: $\widehat{\nabla}_{\boldsymbol{\Xi}}$ is negative definite for all log-concave likelihoods (concavity in $\mathbf{z}$). Hence, (11) guarantees valid stochastic NGD updates provided that $\gamma_t \in [0, 1]$, and often appears in the natural gradient variational inference literature (e.g., Khan et al., 2015; Khan & Lin, 2017; Zhang et al., 2018; Lin et al., 2020).

Additional Discussion. The main goal of this section is to point out the sufficient condition for valid NGD updates in the stochastic setting, as well as its special case Remark 2. Those observations, though simple, are prerequisites for the convergence of stochastic NGVI in §4. Moreover, the Bonnet and Price stochastic gradients will be used in the experiments in §6.

We mention a few common workarounds to take advantage of automatic differentiation, even though natively applying automatic differentiation may break down stochastic NGD. Numerous *approximate* NGD methods admit valid updates in the stochastic settings (Khan et al., 2018; Osawa et al., 2019; Lin et al., 2020), with some specifically addressing the constraint on the natural parameter (Lin et al., 2020). An alternative is to parameterize the variational distribution with an unconstrained parameter, e.g., the mean and the covariance square root. Refer to Salimbeni et al. (2018) for more examples of parameterizations. As a side effect, changing the parameterization also changes the ELBO landscape and may slow down the convergence.

4. Convergence of Stochastic NGVI

Even though NGVI is known to converge in one step for conjugate likelihoods, it generally does not in the stochastic setting. This section aims to establish a convergence rate of stochastic NGVI for conjugate likelihoods. The main techniques we will use are recent developments of stochastic mirror descent for relatively smooth and strongly convex functions (Lu et al., 2018; Hanzely & Richtárik, 2021).

Definition 6 (Hanzely & Richtárik, 2021)**.** *Given the step sizes $\{\gamma_t\}_{t=0}^{\infty}$ and the iterates $\{\boldsymbol{\omega}_t\}_{t=0}^{\infty}$ generated by the updates (9), we define the gradient variance at the step t as*

$$\frac{1}{\gamma_t} \mathbb{E}[\langle \widehat{\nabla} \ell(\boldsymbol{\omega}_t) - \nabla \ell(\boldsymbol{\omega}_t), \boldsymbol{\omega}_{t+1,*} - \boldsymbol{\omega}_{t+1} \rangle \mid \boldsymbol{\omega}_t], \tag{12}$$

where $\boldsymbol{\omega}_{t+1,*} = \operatorname{argmin}_{\boldsymbol{\omega} \in \Omega} \nabla \ell(\boldsymbol{\omega}_t)^\top \boldsymbol{\omega} + \frac{1}{\gamma_t} \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_t)$ and the conditional expectation is taken over the randomness of the stochastic gradient $\widehat{\nabla} \ell(\boldsymbol{\omega}_t)$.

Note that the gradient variance (12) reduces to the familiar one $\mathbb{E}\|\widehat{\nabla} \ell(\boldsymbol{\omega}_t) - \nabla \ell(\boldsymbol{\omega}_t)\|^2$ for gradient descent updates $\boldsymbol{\omega}_{t+1,*} = \boldsymbol{\omega}_t - \gamma_t \nabla \ell(\boldsymbol{\omega}_t)$ and $\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \gamma_t \widehat{\nabla} \ell(\boldsymbol{\omega}_t)$. For mirror descent, however, (12) is a generalization that does not depend on a norm. The norm-independency is crucial for our setting. Common stochastic mirror descent analyses require the distance generating function Φ to be strongly convex *w.r.t.* a norm and then measure the gradient variance in the dual norm (e.g., Bubeck, 2015; Lan, 2020; Liu et al., 2023; Nguyen et al., 2023; Fatkhullin & He, 2024). However, as shown in §A.1, the conjugate of the log-partition function A^* is not strongly convex *w.r.t.* any norms, which prevents us from measuring the gradient variance with a norm. The absence of strong convexity in the distance generating function A^* may partially explain why a precise convergence rate of stochastic natural gradient variational inference is not developed over the years.

Lemma 2. *For conjugate likelihoods, the negative ELBO $\ell(\boldsymbol{\omega})$ is 1-smooth 1-strongly convex relative to the convex conjugate A^* of the log-partition function.*

The relative 1-smoothness and 1-strong convexity imply that the negative ELBO is a well-conditioned objective. Besides, the first-order approximation at an arbitrary $\boldsymbol{\omega}_t \in \Omega$ is exact:

$$\ell(\boldsymbol{\omega}) = \ell(\boldsymbol{\omega}_t) + \langle \nabla \ell(\boldsymbol{\omega}_t), \boldsymbol{\omega} - \boldsymbol{\omega}_t \rangle + \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_t).$$

With the exact gradient $\nabla \ell(\boldsymbol{\omega})$, the mirror descent update (9), which minimizes the first-order approximation, converges in one step with the step size $\gamma_t = 1$. However, one-step convergence is generally not possible in the stochastic setting. Next, we present a general convergence rate that holds for all conjugate likelihoods—the prior $p(\mathbf{z})$ and the likelihood $p(\mathbf{y} \mid \mathbf{z})$ are chosen such that the posterior $p(\mathbf{z} \mid \mathbf{y})$ is in the same exponential family as the prior.

Assumption 1. *The stochastic gradient* $\widehat{\nabla}\ell(\boldsymbol{\omega}_t)$

1. *respects the domain:* $\boldsymbol{\eta}_{t+1} \in \mathcal{D}$ *for all* $t \geq 0$ *in* (8);
2. *is unbiased:* $\mathbb{E}[\widehat{\nabla}\ell(\boldsymbol{\omega}_t) \mid \boldsymbol{\omega}_t] = \nabla\ell(\boldsymbol{\omega}_t)$;
3. *has bounded variance:* (12) *is bounded by* $V > 0$.

Theorem 1. *Suppose the likelihood* $p(\mathbf{y} \mid \mathbf{z})$ *is conjugate and the stochastic gradient* $\widehat{\nabla}\ell(\boldsymbol{\omega}_t)$ *satisfies Assumption 1. Running* $T+1$ *iterations of stochastic natural gradient descent with* $\gamma_t = \frac{2}{2+t}$ *generate a point* $\bar{\boldsymbol{\omega}}_{T+1}$ *that satisfies*

$$\mathbb{E}[\ell(\bar{\boldsymbol{\omega}}_{T+1})] - \min_{\boldsymbol{\omega}\in\Omega} \ell(\boldsymbol{\omega}) \leq \frac{V}{T+2}, \tag{13}$$

where $\bar{\boldsymbol{\omega}}_{T+1} = \frac{2}{(T+1)(T+2)} \sum_{t=0}^{T} (t+1)\boldsymbol{\omega}_{t+1}$. *Let* $\bar{q}_{T+1}$ *be the variational distribution represented by* $\bar{\boldsymbol{\omega}}_{T+1}$. *Then, the KL divergence to the true posterior* q^* *is bounded by*

$$\mathbb{E}[\mathsf{D}_{\mathrm{KL}}(\bar{q}_{T+1}, q^*)] \leq \frac{V}{T+2}. \tag{14}$$

We make two observations on the rate (13). First, the rate interpolates between stochastic and deterministic settings. In particular, zero variance $V = 0$ implies convergence in one step. Second, the convergence rate does not depend on the distance from the initialization q_0 to the true posterior q^*. This leads to an interesting interpretation: no matter how far away the initialization is to the true posterior, after the first iteration $\bar{\boldsymbol{\omega}}_1$ always goes to a sublevel set whose size only depends on the variance V. Both properties are due to the step size schedule $\gamma_t = \frac{2}{2+t}$, in particular $\gamma_0 = 1$. In general, linearly decreasing step sizes also guarantee convergence, but may lose these two properties.

It is not entirely clear if the conditions in Assumption 1 hold in practice at all. In particular, Assumption 1 requires the gradient variance (12), defined in a non-standard form, to be bounded. The rest of this section is devoted to this question by a case study of a common conjugate variational inference problem, where we show all conditions in Assumption 1 indeed hold in practice.

Example 1 (Bayesian Linear Regression)**.** *Consider*

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{P}), \quad p(\mathbf{y} \mid \mathbf{X}, \mathbf{z}) = \mathcal{N}(\mathbf{X}\mathbf{z}, \sigma^2\mathbf{I}),$$

where the prior $p(\mathbf{z})$ *is a zero-mean Gaussian and the label* $\mathbf{y}$ *has an independent Gaussian observation noise. The negative ELBO can be written as a finite sum*

$$\begin{aligned}\ell(q) &= -\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{z})] + \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})) \\ &= -\sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z})} \log p(y_i \mid \mathbf{x}_i, \mathbf{z}) + \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})),\end{aligned}$$

where $\{\mathbf{x}_i\}_{i=1}^n$ *are the rows of* $\mathbf{X} \in \mathbb{R}^{n\times d}$. *Without loss of generality, we assume that the variational distribution is initialized as a standard normal distribution* $q_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Data Sub-Sampling Stochastic Gradient. Each iteration samples m data points uniformly and independently:

$$\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_m}.$$

Each index i_k is independently sampled from the uniform distribution $U[n]$. The stochastic natural gradient $\widehat{\nabla}\ell(\boldsymbol{\omega})$ is

$$\nabla_{\boldsymbol{\omega}} \left[-\frac{n}{m} \sum_{k=1}^{m} \mathbb{E}_q \log p(y_{i_k} \mid \mathbf{x}_{i_k}, \mathbf{z}) + \mathsf{D}_{\mathrm{KL}}(q, p) \right], \tag{15}$$

where $p(y_{i_k} \mid \mathbf{x}_{i_k}, \mathbf{z}) = \mathcal{N}(\mathbf{z}^\top \mathbf{x}_{i_k}, \sigma^2)$ and the expectation $\mathbb{E}_{q(\mathbf{z})} \log p(y_{i_k} \mid \mathbf{x}_{i_k}, \mathbf{z})$ is computed in a closed-form. Each stochastic NGD update can be computed in $\mathcal{O}(d^2 m)$, while the closed-form posterior of Bayesian linear regression takes $\mathcal{O}(d^2 n + d^3)$ to compute. Approximating the posterior via stochastic NGD is more practical for large datasets. Indeed, it is widely used in variational Gaussian processes (e.g., Hensman et al., 2013; Salimbeni et al., 2018) where n might be too large to even fit the data into the memory.

Now we verify the conditions in Assumption 1. For each $i \in [n]$, the second component $\nabla_{\boldsymbol{\Xi}}$ of the gradient

$$\left(\nabla_{\boldsymbol{\xi}}, \nabla_{\boldsymbol{\Xi}}\right) = \nabla_{\boldsymbol{\omega}} \mathbb{E}_q \log p(y_i \mid \mathbf{x}_i, \mathbf{z})$$

is negative definite (see §C.1). By Remark 2, the stochastic gradient (15) indeed respects the domain $\mathcal{D}$ and results in valid NGD updates, as long as $0 \leq \gamma_t \leq 1$. It is clearly unbiased as each data point $\mathbf{x}_{i_k}$ is sampled uniformly. Lastly, its variance is bounded:

Lemma 3. *The stochastic gradient* (15) *satisfies*

$$\frac{1}{\gamma_t} \mathbb{E}[\langle \widehat{\nabla}\ell(\boldsymbol{\omega}_t) - \nabla\ell(\boldsymbol{\omega}_t), \boldsymbol{\omega}_{t+1,*} - \boldsymbol{\omega}_{t+1} \rangle \mid \boldsymbol{\omega}_t] \leq V_2, \tag{16}$$

where $V_2 = (\nu s_1 + \frac{1}{2}\nu^2 s_2 + 2\nu^2 b\sqrt{s_1 s_2} n + \nu^3 b^2 s_2 n^2) \frac{n^2}{\sigma^4 m}$, *with* $\nu = \max\{1, \|\mathbf{P}\|\}$, $b = \max_{1\leq i\leq n} \|y_i \mathbf{x}_i\|$, *and the empirical variances* $s_1 = \mathbb{E}_{j\sim U[n]} \left\| y_j \mathbf{x}_j - \frac{1}{n}\sum_{i=1}^n y_i \mathbf{x}_i \right\|^2$. *and* $s_2 = \mathbb{E}_{j\sim U[n]} \left\| \mathbf{x}_j \mathbf{x}_j^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right\|_{\mathrm{F}}^2$.

The constant in Lemma 3 is not necessarily tight and may be improved. Nevertheless, it serves the purpose to show that the gradient variance is bounded by a constant.

Application to Gaussian Process Regression. Our result immediately applies to stochastic variational Gaussian processes (SVGP) (Hensman et al., 2013), a popular large-scale Gaussian process regression model. SVGP training minimizes the negative ELBO of the form

$$-\int p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u}) \log p(\mathbf{y} \mid \mathbf{f}) \, \mathrm{d}\mathbf{f} \, \mathrm{d}\mathbf{u} + \mathsf{D}_{\mathrm{KL}}(q(\mathbf{u}), p(\mathbf{u})),$$

where the variational distribution is $q(\mathbf{u})$ with the likelihood

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{f}) &= \mathcal{N}(\mathbf{y}; \mathbf{0}, \sigma^2\mathbf{I}), \\ p(\mathbf{f} \mid \mathbf{u}) &= \mathcal{N}(\mathbf{f}; \mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{ff}} - \mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}}), \end{aligned}$$

and the prior $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{K_{uu}})$. Simplify the ELBO by removing terms independent of $q(\mathbf{u})$ gives

$$-\int q(\mathbf{u}) \log \mathcal{N}(\mathbf{y}; \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\mathbf{u}, \sigma^2\mathbf{I})\, \mathrm{d}\mathbf{u} + \mathsf{D}_{\mathrm{KL}}(q(\mathbf{u}), p(\mathbf{u})).$$

Hence, finding the optimal variational distribution $q(\mathbf{u})$ is equivalent to Bayesian linear regression in Example 1 with $\mathbf{X} = \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}$ and the prior covariance $\mathbf{P} = \mathbf{K_{uu}}$. Even though the optimal variational distribution q^* has a closed-form, computing it exactly needs to access the entire dataset. Besides, q^* varies after every GP hyperparameter update, and it is expensive to compute q^* exactly every iteration. Thus, a popular approach is jointly minimizing the variational parameters and the hyperparameters by mini-batch stochastic optimization. Lemma 3 together with Theorem 1 gives a convergence rate of the variational distribution in SVGP training. The convergence rate may also find applications in some collapsed variational inference methods (e.g., Hensman et al., 2012), where NGD is applied to a subset of latent variables in an conjugate exponential family.

5. ELBO Landscape

In the last section, we have seen that the (negative) ELBO $\ell(\boldsymbol{\omega})$, as a function of the expectation paremters $\boldsymbol{\omega}$, has good properties when the likelihood is conjugate (see Lemma 2). These properties are crucial for the convergence analysis. The natural question is whether the ELBO preserves these properties for non-conjugate likelihoods.

This section studies variational inference with a Gaussian prior $p(\mathbf{z})$, a Gaussian variational family $\mathcal{Q}$, and a non-Gaussian (i.e., non-conjugate) likelihood $p(\mathbf{y} \mid \mathbf{z})$. Surprisingly, we show that even when the likelihood is log-concave, the ELBO $\ell(\boldsymbol{\omega})$ is not guaranteed to be convex in the expectation parameter. This is in sharp contrast to the mean-square-root parameterization $(\mathbf{m}, \mathbf{C})$, with $\mathbf{m}$ and $\mathbf{C}$ representing the mean and the Cholesky factor respectively, used in stochastic gradient, where the ELBO is smooth and strongly convex (Domke, 2020).

Below we give two examples (with details in §E) where the negative ELBO $\ell(\boldsymbol{\omega})$ is non-convex in the expectation parameter $\boldsymbol{\omega}$, even for simple log-concave likelihoods. To show the objective is non-convex, all we need to do is to find a dataset such that the negative ELBO is non-convex.

Logistic Regression. Consider an 1-dimensional Bayesian logistic regression on the dataset $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in [-1, 1]$ and $y_i \in \{-1, 1\}$. The prior $p(w, b)$ on the weight w and the bias b is a standard Gaussian distribution. The negative ELBO $\ell(\boldsymbol{\omega})$ is

$$\mathbb{E}_{q(w,b)}\left[\sum_{i=1}^n \log\big(1 + \exp(-y_i(wx_i + b))\big)\right] + \mathsf{D}_{\mathrm{KL}}(q, p),$$

where $q(w, b)$ is a Gaussian variational distribution. Restrict the expectation parameter $\boldsymbol{\omega}$ of q on the convex subset

$$\{\boldsymbol{\omega} = (\mathbf{0}, \boldsymbol{\Xi}) : \boldsymbol{\Xi} = \mathrm{diag}(s_1, s_2), s_1 > 0, s_2 > 0\} \subseteq \Omega,$$

where the first component of $\boldsymbol{\omega}$ is zero and the second component is a diagonal matrix. If $\ell(\boldsymbol{\omega})$ was convex in $\boldsymbol{\omega}$, it would be convex in s_2 at least. Taking the second-order derivative with respect to s_2, we have

$$\nabla_{s_2}^2 \ell(\boldsymbol{\omega}) = \sum_{i=1}^n \mathbb{E}[\psi_i(1 - \psi_i)(6\psi_i^2 - 6\psi_i + 1)] + \frac{1}{2s_2^2},$$

where $\psi_i = \psi(wx_i + b)$ with ψ the sigmoid function and the expectation is taken over $(w, b) \sim q_{\boldsymbol{\omega}}$. Note that the expectation is negative in the limit:

$$\lim_{s_1, s_2 \to 0} \mathbb{E}_{q(w,b)}[\psi_i(1 - \psi_i)(6\psi_i^2 - 6\psi_i + 1)] = -\frac{1}{8}.$$

In particular, there exists an absolute constant $\delta > 0$ such that when $s_1 = s_2 = \delta$ we have

$$\psi_i(1 - \psi_i)(6\psi_i^2 - 6\psi_i + 1) < -\frac{1}{16}$$

for all $1 \le i \le n$. This implies $\nabla_{s_2}^2 \ell(\boldsymbol{\omega}) < 0$ when $n \ge 8/\delta^2$ for a particular $\boldsymbol{\omega} = (\mathbf{0}, \boldsymbol{\Xi})$ with $\boldsymbol{\Xi} = \mathrm{diag}(\delta, \delta)$.

Poisson Regression. We choose this example because of its analytical ELBO. Bayesian Poisson regression assumes that $y \mid \mathbf{x}$ follows a Poisson distribution with the expectation

$$\mathbb{E}[y \mid \mathbf{x}] = \exp(\mathbf{w}^\top \mathbf{x}).$$

The prior $p(\mathbf{w})$ on the weight $\mathbf{w}$ is a standard Gaussian distribution. Let $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi})$ be the expectation parameter of the Gaussian variational distribution q. The expected log likelihood $-\mathbb{E}_{q(\mathbf{w})} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$ is

$$-\mathbf{y}^\top \mathbf{X}\boldsymbol{\xi} + \sum_{i=1}^n \Big[\exp\big(\boldsymbol{\xi}^\top \mathbf{x}_i + \frac{1}{2}\mathbf{x}_i^\top (\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top)\mathbf{x}_i\big)\Big],$$

which is not convex in $\boldsymbol{\xi}$. Compute the Hessian of $\ell(\boldsymbol{\omega})$ *w.r.t.* $\boldsymbol{\xi}$ and evaluate it on the subset of the domain

$$\{\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi}) : \boldsymbol{\Xi} = \boldsymbol{\xi}\boldsymbol{\xi}^\top + 2\mathbf{I}\} \subseteq \Omega.$$

Then, we obtain

$$\nabla_{\boldsymbol{\xi}}^2 \ell(\boldsymbol{\omega}) \preceq \sum_{i=1}^n (\mathbf{x}_i^\top \boldsymbol{\xi})(\mathbf{x}_i^\top \boldsymbol{\xi} - 2)\mathbf{x}_i \mathbf{x}_i^\top + \nabla_{\boldsymbol{\xi}}^2 A^*(\boldsymbol{\omega}).$$

For a fixed $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi})$, there exists a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that $0 < \mathbf{x}_i^\top \boldsymbol{\xi} < 2$ for all $1 \le i \le n$. Now, consider the Hessian $\nabla_{\boldsymbol{\xi}}^2 \ell(\boldsymbol{\omega})$ on the scaled dataset $\{(c\mathbf{x}_i, y_i)\}_{i=1}^n$ evaluated at $\frac{1}{c}\boldsymbol{\xi}$. As $c \to \infty$, we have found a dataset such that the Hessian is negative.

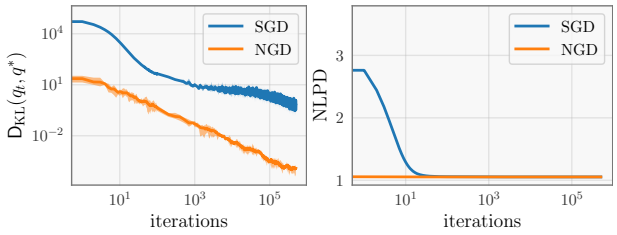

Figure 1: Mini-batch Bayesian linear regression on the Bike dataset. **Left:** The KL divergence to the optimum q^*. **Right:** The training negative log predictive density.

Recent work demonstrates that the negative ELBO is convex with a log-concave likelihood if the variational distribution is a Gaussian distribution with the mean-square-root parameterization (Domke, 2020). However, the above two examples show that it is not the case for the expectation parameter $\boldsymbol{\omega}$. Given that the canonical implementation of NGD is equivalent to mirror descent in the expectation parameter space, NGD may implicitly optimize a non-convex objective when the likelihood is non-conjugate even for simple log-concave likelihoods.

Nonetheless, the negative ELBO does have some convenient properties for log-concave likelihoods—it is not an arbitrary non-convex objective.

Proposition 1. *Suppose the prior and the variational family are both Gaussians. If the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is log-concave in $\mathbf{z}$, then the negative ELBO $\ell(\boldsymbol{\omega})$ as a function of the expectation parameter has a unique minimizer $\boldsymbol{\omega}^*$. In addition, if the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is differentiable in $\mathbf{z}$, then $\boldsymbol{\omega}^*$ is the unique stationary point of $\ell(\boldsymbol{\omega})$.*

Proposition 1 is not surprising, since there is a differentiable bijection between the expectation parameterization and the mean-square-root parameterization. The uniqueness of the minimizer and the stationary point is derived from strongly convexity in the mean-square-root parameterization. Thus, stochastic NGVI may still converge to the optimum with log-concave likelihoods despite the non-convexity.

We end this section by discussing some implications. With non-conjugate likelihoods, the negative ELBO $\ell(\boldsymbol{\omega})$ is not strongly convex nor relatively strongly convex, since it is not even convex. Strong convexity plays a crucial role in stochastic optimization. Without it, stochastic gradient descent has a convergence rate of $\mathcal{O}(1/\sqrt{T})$ under standard assumptions. This rate is improved to $\mathcal{O}(1/T)$ for strongly convex functions. The fact that stochastic NGVI is implicitly optimizing a non-convex objective implies that we may need to resort to new properties of the ELBO to prove its $\mathcal{O}(1/T)$ convergence rate for non-conjugate likelihoods, if it can achieve this rate at all. One possibility to achieve this is the Polyak-Łojasiewicz inequality (Karimi et al., 2016).

6. Numerical Simulation

This section presents supporting numerical simulations on datasets from the UCI repository (Bike and Mushroom) and MNIST (Kelly et al., 2017; LeCun et al., 1998)

6.1. Bayesian Linear Regression

Figure 1 presents Bayesian linear regression on the Bike dataset ($n = 17,389$), with a standard normal prior and a noise $\sigma^2 = 1$. The (negative) ELBO is optimized by SGD and stochastic NGD with a mini-batch size of 1000. SGD uses a step size schedule $\gamma_t = \frac{1}{10^5+t}$, a linearly decreasing schedule on the same order as Domke et al. (2023, Theorem 10). Stochastic NGD uses a step size schedule $\gamma_t = \frac{2}{2+t}$ predicted by Theorem 1.

The true posterior q^* of Bayesian linear regression has a closed-form, which allows us to plot the optimality gap in the KL divergence. In addition, we plot the negative predictive log density (NLPD) on the training set. In the log-log scale, the KL divergences to the optimal posterior of both methods decrease at the same rate, with roughly the same slope in the figure. This suggests that both methods have the same $\mathcal{O}(\frac{1}{T})$ complexity, and that stochastic NGD may be only constant times faster than SGD. Nonetheless, stochastic NGD converges very fast in the early stage. It takes SGD thousands of iterations to catch up the progress that stochastic NGD makes in the first few iterations, implying that stochastic NGD has a much better constant factor in the big $\mathcal{O}$ notation. Indeed, recall that the convergence rate in Theorem 1 only depends on the stochastic gradient variance, independent of the objective's condition number and the distance from the initialization to the optimum (see §4).

6.2. Non-Conjugate Likelihoods

Figure 2 shows Bayesian logistic regression on the Mushroom dataset ($n = 8124$) and MNIST (a subset of 1 and 7 with $n \approx 13,000$ images). Again, stochastic NGD is faster than SGD, but the improvement is less drastic compared with conjugate likelihoods. This is consistent with previous empirical observations (Salimbeni et al., 2018).

Besides faster convergence, it appears that the step size of stochastic NGD is easier to tune in practice. In most cases, the step size $\gamma = 0.1$ convergences smoothly. Sometimes $\gamma = 0.1$ is too large such that stochastic NGD oscillates in the final stage (Figure 2 right panel). Simply decreasing it to $\gamma = 0.01$ leads to smooth convergence in most cases. These observations suggest that the ELBO $\ell(\boldsymbol{\omega})$, as a function of the expectation parameter, might have a small smoothness constant. Indeed, the smoothness constant is 1 for conjugate likelihoods (recall Lemma 2). For non-conjugate likelihoods in practice, we hypothesize its smoothness constant might be close to 1 as well.

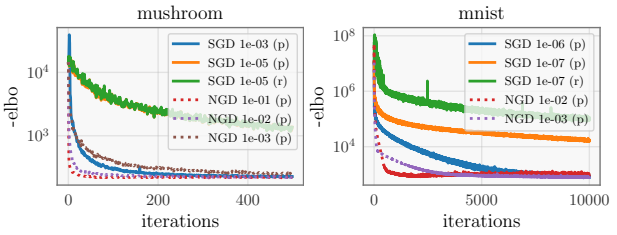

Figure 2: Bayesian logistic regression on Mushroom and MNIST. Labels with "(p)" use the stochastic gradient by the Price theorem (11). Labels with "(r)" use the stochastic gradient by the reparamerization trick.

We point out a side note that the Price stochastic gradient (11) is a high-quality gradient estimator superior to the reparameterization trick. For instance, in the special case when the log likelihood $\log p(\mathbf{y} \mid \mathbf{z})$ is a quadratic function in $\mathbf{z}$, e.g., $p(\mathbf{y} \mid \mathbf{z}) = \mathcal{N}(\mathbf{z}, \mathbf{I})$, the Price stochastic gradient $\widehat{\nabla}_{\boldsymbol{\Xi}}$ is exact and has zero variance! For general non-conjugate likelihoods, we expect the Price stochastic gradient has lower variance. Indeed, SGD has a dramatic improvement by just switching to the Price stochastic gradient.

7. Related Work

Natural gradient descent was initially proposed by Amari (1998) as a learning algorithm for multi-layer perceptrons that is believed to exploit the information geometry. Subsequently, this method has been applied to variational inference (e.g., Honkela & Valpola, 2004; Hensman et al., 2012; Hoffman et al., 2013; Khan & Lin, 2017). Recent new developments of natural gradient variational inference include generalization to mixtures of exponential families (Lin et al., 2019; Arenz et al., 2023), handling the positive definite domain constraint (Lin et al., 2020), supporting structured matrix parameterization (Lin et al., 2021), implementation via automatic differentiation (Salimbeni et al., 2018), adaptations to online learning (Chérief-Abdellatif et al., 2019), and generalization to Wasserstein statistical manifold (Chen & Li, 2020; Li & Zhao, 2023).

Outside variational inference, natural gradient descent has been applied to training (non-Bayesian) neural networks in supervised learning (e.g. Bernacchia et al., 2018; Song et al., 2018; Zhang et al., 2019). Interestingly, Zhang et al. (2018) establish a connection between training neural networks with noisy natural gradient and variational inference. For a comprehensive survey of this area, we direct readers to the monograph by Martens (2020). In particular, Martens (2020) hypothesize a $\mathcal{O}(\frac{1}{T})$ asymptotic convergence rate via an argument based on Fisher efficiency. They also gave a non-asymptotic convergence rate of $\mathcal{O}(\frac{1}{T})$ for stochastic preconditioned gradient descent with a fixed preconditioning matrix and a quadratic objective.

In addition, natural gradient descent has been applied to policy optimization in reinforcement learning, leading to natural policy gradient (Kakade, 2001). With the connection to mirror descent, there is a recent interest in this method that leads to a series of analyses (e.g., Geist et al., 2019; Shani et al., 2020; Agarwal et al., 2021; Khodadadian et al., 2021; Xiao, 2022; Yuan et al., 2023).

The natural gradient methods applied to different machine learning problems mentioned previously share a common feature: the gradient direction is preconditioned with the Fisher information matrix. Despite being coined with the same name "natural gradient", we point out a subtle difference in natural gradient variational inference. The distance generating function A^* in NGVI, namely the log-partition function's conjugate, is the negative *differential* entropy that is non-strongly convex and non-smooth, as shown in §A.1. In contrast, the distance generating function in natural policy gradient is the negative *Shannon* entropy, which is well-known to be strongly convex *w.r.t.* a norm (e.g., Bubeck, 2015, Section 4.3). As mentioned in §4, this strong convexity is a key condition for mirror descent analyses in the stochastic setting. We hope our work motivates new developments in stochastic mirror descent for non-strongly convex non-smooth distance generating functions.

8. Conclusion

Over the years, empirical observations suggest stochastic natural gradient descent (NGD) is faster than stochastic gradient descent for variational inference. To understand how fast NGD converges, we prove the first $\mathcal{O}(\frac{1}{T})$ non-asymptotic convergence rate for conjugate likelihoods. The rate appears to be tight based on experiments, suggesting that stochastic natural gradient variational inference (NGVI) may be only constant times faster than stochastic gradient descent. Nevertheless, the constant improvement could be dramatic in practice. For non-conjugate likelihoods, we show that "canonical" stochastic NGVI implicitly optimizes a non-convex objective, which suggests that a $\mathcal{O}(\frac{1}{T})$ rate is unlikely without discoveries of new properties of the ELBO.

Acknowledgements

The authors thank the anonymous reviewers for constructive feedbacks. KW would like to thank Kyurae Kim for helpful discussions in the early stage of this work. KW and JRG are supported by NSF award IIS-2145644.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There might be many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. 9

Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998. 1, 9

Arenz, O., Dahlinger, P., Ye, Z., Volpp, M., and Neumann, G. A unified perspective on natural gradient variational inference with gaussian mixture models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. 9

Bauschke, H. H., Bolte, J., and Teboulle, M. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. 4

Bernacchia, A., Lengyel, M., and Hennequin, G. Exact natural gradient in deep linear networks and its application to the nonlinear case. volume 31, 2018. 9

Birnbaum, B., Devanur, N. R., and Xiao, L. Distributed algorithms via gradient descent for Fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 127–136, 2011. 4

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

Bonnet, G. Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. In *Annales des Télécommunications*, volume 19, pp. 203–220. Springer, 1964. 4

Bottou, L. Online learning and stochastic approximations. *Online learning in neural networks*, 17(9):142, 1998. 1

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (3-4):231–357, 2015. 5, 9

Chen, Y. and Li, W. Optimal transport natural gradient for statistical manifolds with continuous sample space. *Information Geometry*, 3(1):1–32, 2020. 9

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. E. A generalization bound for online variational inference. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101, pp. 662–677. PMLR, 2019. 9

Domke, J. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 2587–2596. PMLR, 2020. 1, 7, 8

Domke, J., Gower, R. M., and Garrigos, G. Provable convergence guarantees for black-box variational inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 8, 24

Fatkhullin, I. and He, N. Taming nonconvex stochastic mirror descent with general Bregman divergence. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 3493–3501. PMLR, 2024. 5

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2160–2169. PMLR, 2019. 9

Hanzely, F. and Richtárik, P. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 79:717–766, 2021. 4, 5, 21

Hensman, J., Rattray, M., and Lawrence, N. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems*, volume 25, 2012. 1, 7, 9

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013. 1, 2, 6

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pp. 351–360, 2015. 1, 2

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013. 1, 2, 9

Honkela, A. and Valpola, H. Unsupervised variational Bayesian learning of nonlinear models. In *Advances in Neural Information Processing Systems*, volume 17, 2004. 1, 9

Kakade, S. M. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. 9

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer International Publishing, 2016. 8

Kelly, M., Longjohn, R., and Nottingham, K. The UCI machine learning repository, 2017. URL `https://archive.ics.uci.edu`. 8

Khan, M. and Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 878–887, 2017. 3, 5, 9, 13

Khan, M., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2611–2620, 2018. 1, 3, 5

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 Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016. 1

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. Curran Associates, Inc., 2015. 5

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3794–3799. IEEE Press, 2021. 9

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, 2023. 1

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *The First International Conference on Learning Representations*, 2013. 4

Lan, G. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020. 5

LeCun, Y., Cortes, C., and Burges, C. J. The MNIST database, 1998. URL https://archive.ics.uci.edu. 8

Li, W. and Zhao, J. Wasserstein information matrix. *Information Geometry*, 6(1):203–255, 2023. 9

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 36th International Conference on Machine Learning*, volume 97, pp. 3992–4002, 2019. 9

Lin, W., Schmidt, M., and Khan, M. E. Handling the positive-definite constraint in the Bayesian learning rule. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6116–6126, 2020. 5, 9

Lin, W., Nielsen, F., Emtiyaz, K. M., and Schmidt, M. Tractable structured natural-gradient descent using local parameterizations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6680–6691. PMLR, 2021. 9

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 21884–21914. PMLR, 2023. 5

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963. 2

Lu, H., Freund, R. M., and Nesterov, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. 4, 5

Martens, J. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. 9

Murray, I. Differentiation of the cholesky decomposition. *arXiv preprint arXiv:1602.07527*, 2016. 4

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983. 3

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

Nielsen, F. and Garcia, V. Statistical exponential families: A digest with flash cards. arXiv preprint arXiv:0911.4863, 2009. 13

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009. 4

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with Bayesian principles. In *Advances in neural information processing systems*, volume 32, 2019. 1, 5

Polyak, B. T. et al. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963. 2

Price, R. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958. 4

Ranganath, R., Gerrish, S., and Blei, D. Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pp. 814–822, 2014. 1

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015. 3

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1278–1286, 2014. 4

Salimbeni, H., Eleftheriadis, S., and Hensman, J. Natural gradients in practice: Non-conjugate variational inference in gaussian process models. In *International Conference on Artificial Intelligence and Statistics*, pp. 689–697, 2018. 1, 2, 5, 6, 8, 9

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020. 9

Song, Y., Song, J., and Ermon, S. Accelerating natural gradient with higher-order invariance. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4713–4722, 2018. 9

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1971–1979, 2014. 4

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. 2

Xiao, L. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282): 1–36, 2022. 9

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations*, 2023. 9

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5852–5861. PMLR, 2018. 5, 9

Zhang, G., Martens, J., and Grosse, R. B. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 9

A. Exponential Family

The following useful lemma is well-known (e.g., Nielsen & Garcia, 2009), which hints the connection between NGD and mirror descent. For the sake of completeness, we present a proof here, which largely follows Khan & Lin (2017, Lemma 2).

Lemma 4. *Let $q(\mathbf{z})$ and $q'(\mathbf{z})$ be distributions in the same exponential family. That is, they share the same base measure $h(\mathbf{z})$, sufficient statistics $\phi(\mathbf{z})$, and the log-partition function $A(\boldsymbol{\eta})$. Let $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ be the expectation parameters of $q(\mathbf{z})$ and $q'(\mathbf{z})$ respectively. Then, we have*

$$\mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}') = \mathsf{D}_A(\boldsymbol{\eta}', \boldsymbol{\eta}) = \mathsf{D}_{\mathrm{KL}}(q, q').$$

where D_A and D_{A^*} are the Bregman divergences associated with A and A^*, respectively.

Proof. The first equality is a standard property of the Bregman divergence, since $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ are dual to each other. Only the second equality needs a proof:

$$\begin{aligned}
\mathsf{D}_{\mathrm{KL}}(q, q') &= \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z}) - \log q'(\mathbf{z})] \\
&= \mathbb{E}_{q(\mathbf{z})}[\log h(\mathbf{z}) + \langle \phi(\mathbf{z}), \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta})] - \mathbb{E}_{q(\mathbf{z})}[\log h(\mathbf{z}) + \langle \phi(\mathbf{z}), \boldsymbol{\eta}' \rangle - A(\boldsymbol{\eta}')] \\
&= \left\langle \mathbb{E}_{q(\mathbf{z})}[\phi(\mathbf{z})], \boldsymbol{\eta} - \boldsymbol{\eta}' \right\rangle - A(\boldsymbol{\eta}) + A(\boldsymbol{\eta}') \\
&= A(\boldsymbol{\eta}') - A(\boldsymbol{\eta}) - \langle \boldsymbol{\omega}, \boldsymbol{\eta}' - \boldsymbol{\eta} \rangle \\
&= A(\boldsymbol{\eta}') - A(\boldsymbol{\eta}) - \langle \nabla A(\boldsymbol{\eta}), \boldsymbol{\eta}' - \boldsymbol{\eta} \rangle \\
&= \mathsf{D}_A(\boldsymbol{\eta}', \boldsymbol{\eta}),
\end{aligned}$$

where the second line uses the definition of the exponential family; the forth line uses the definition of the expectation parameter; the fifth line uses the duality between the natural and expectation parameters (4); and the last line uses the definition of the Bregman divergence. □

Let $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ be the expectation parameters of q and q' respectively. Lemma 4 gives a simple expression for the derivative of $\mathsf{D}_{\mathrm{KL}}(q, q')$ *w.r.t.* the expectation parameter $\boldsymbol{\omega}$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\omega}} \mathsf{D}_{\mathrm{KL}}(q, q') &= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\omega}} \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}') \\
&= \nabla A^*(\boldsymbol{\omega}) - \nabla A^*(\boldsymbol{\omega}') \\
&= \boldsymbol{\eta} - \boldsymbol{\eta}',
\end{aligned}$$

where the second line is a standard property of the Bregman divergence and the third line is because $\nabla A^*(\cdot)$ maps a expectation parameter to its corresponding natural parameter.

A.1. Gaussian Distributions: The Log-Partition Function

This section gives an explicit expression of the log-partition function $A(\boldsymbol{\eta})$ of the Gaussian distribution and its convex conjugate A^*. In addition, we show that the convex conjugate A^* is non-smooth and non-strongly convex.

Let $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a d-dimensional Gaussian with the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$. Its density is of the form

$$\begin{aligned}
q(\mathbf{z}; \boldsymbol{\eta}) &\propto \exp\Big(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) - \frac{1}{2}\log\det(\boldsymbol{\Sigma})\Big) \\
&= \exp\Big(\langle \mathbf{z}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \rangle + \langle \mathbf{z}\mathbf{z}^\top, -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \rangle - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log\det(\boldsymbol{\Sigma})\Big).
\end{aligned}$$

The sufficient statistic is the map $\phi : \mathbf{z} \mapsto (\mathbf{z}, \mathbf{z}\mathbf{z}^\top)$. The natural parameter $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ satisfies $\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Lambda} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$. The log-partition function $A(\cdot)$ as a function of the mean and the covariance is

$$A(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log\det(\boldsymbol{\Sigma}).$$

Plug in the relation between $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the natural parameter $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. We obtain an explicit expression of the log-partition function:

$$A(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = -\frac{1}{4}\boldsymbol{\lambda}^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda} - \frac{1}{2}\log\det(-2\boldsymbol{\Lambda}), \tag{17}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ and $\boldsymbol{\Lambda} \prec \mathbf{0}$. The convex conjugate A^* as a function of the expectation parameter $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi})$ is

$$\begin{aligned} A^*(\boldsymbol{\xi}, \boldsymbol{\Xi}) &= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^d, \boldsymbol{\Lambda}\preceq 0} \langle \boldsymbol{\xi}, \boldsymbol{\lambda}\rangle + \langle \boldsymbol{\Xi}, \boldsymbol{\Lambda}\rangle + \frac{1}{4}\boldsymbol{\lambda}^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda} + \frac{1}{2}\log\det(-2\boldsymbol{\Lambda}) \\ &= -\frac{1}{2}\log\det\big(\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top\big), \end{aligned} \tag{18}$$

where the second line solves the maximization by taking the derivative and setting it equal to zero. The constraint on the expectation parameter $(\boldsymbol{\xi}, \boldsymbol{\Xi})$ is $\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top \succ 0$.

Consider the restriction of A^* on the convex set $\{\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi}) : \boldsymbol{\xi} = \mathbf{0}, \boldsymbol{\Xi} \succ 0\}$. Clearly, A^* is already non-strongly convex and non-smooth in $\boldsymbol{\Xi}$: in one dimension $\frac{\mathrm{d}^2}{\mathrm{d}x^2}(-\log x) = \frac{1}{x^2}$ is neither lower nor upper bounded. Since the absolute value is the only norm (up to a constant) in one dimension, A^* is not strongly convex *w.r.t.* any norms. Finally, both A and A^* are non-smooth and non-strongly convex due to the duality between smoothness and strong convexity.

A.2. Gaussian Distributions: Conversion between the Natural and Expectation Parameters

This section gives explicit expressions of ∇A and ∇A^* of Gaussian distributions. These maps convert between the natural parameter $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and the expectation parameter $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi})$. Differentiating the log-partition function (17) gives

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) &= -\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda}, \\ \nabla_{\boldsymbol{\Lambda}} A(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) &= \frac{1}{4}\boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{\Lambda}^{-1} - \frac{1}{2}\boldsymbol{\Lambda}^{-1}. \end{aligned}$$

The gradient map exactly transforms the natural parameter to the expectation parameter, in that

$$\nabla_{\boldsymbol{\lambda}} A(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\xi}, \quad \nabla_{\boldsymbol{\Lambda}} A(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\Xi}.$$

Similarly, differentiating the conjugate A^* (18) gives

$$\begin{aligned} \nabla_{\boldsymbol{\xi}} A^*(\boldsymbol{\xi}, \boldsymbol{\Xi}) &= \big(\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top\big)^{-1}\boldsymbol{\xi}, \\ \nabla_{\boldsymbol{\Xi}} A^*(\boldsymbol{\xi}, \boldsymbol{\Xi}) &= -\frac{1}{2}\big(\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top\big)^{-1}, \end{aligned}$$

which transform the expectation parameter back to the natural parameter:

$$\nabla_{\boldsymbol{\xi}} A^*(\boldsymbol{\xi}, \boldsymbol{\Xi}) = \boldsymbol{\lambda}, \quad \nabla_{\boldsymbol{\Xi}} A^*(\boldsymbol{\xi}, \boldsymbol{\Xi}) = \boldsymbol{\Lambda}.$$

For a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, recall the relation between the mean/covariance and the natural parameter

$$\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \quad \boldsymbol{\Lambda} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1},$$

and the relation between the mean/covariance and the expectation parameter

$$\boldsymbol{\xi} = \boldsymbol{\mu}, \qquad \boldsymbol{\Xi} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

One can verify that these relations are indeed consistent with the maps $\nabla A(\boldsymbol{\eta}) = \boldsymbol{\omega}$ and $\nabla A^*(\boldsymbol{\omega}) = \boldsymbol{\eta}$.

B. Automatic Differentiation Stochastic Gradient Counterexample

This section gives a counterexample where estimating the gradient of the expected log likelihood

$$\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$$

using automatic differentiation shown in Algorithm 1 does not guarantee a negative definite stochastic gradient $\widehat{\nabla}_{\boldsymbol{\Xi}}$. For simplicity, we use a zero-mean Gaussian $q(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ so that $\boldsymbol{\Xi} = \boldsymbol{\Sigma}$, and a likelihood $p(\mathbf{y} \mid \mathbf{z}) = \mathcal{N}(\mathbf{z}, \mathbf{I})$. The following code produces a stochastic gradient that is not negative definite approximately 50% of the time.

```
import torch
from torch.distributions import MultivariateNormal


if __name__ == "__main__":
    d = 2

    sigma = torch.eye(d).requires_grad_()
    chol = torch.linalg.cholesky(sigma)

    u = torch.randn(d)
    z = chol @ u

    dist = MultivariateNormal(z, torch.eye(d))
    y = torch.zeros(d)

    loss = dist.log_prob(y)
    loss.backward()

    print(loss.item())
    print(sigma.grad)

    # check the diagonal gradient
    # print(-0.5 * u ** 2)

    det = torch.linalg.det(sigma.grad)

    if det > 0.:
        print("not n.d.")
    else:
        print("......")
```

C. Stochastic Gradient Variance for Bayesian Linear Regression in Example 1

In this section, we restrict ourselves to Bayesian linear regression in Example 1, and establish bounds on the stochastic gradient variance (12). Recall that the negative ELBO is a sum of the negative expected log likelihood and the KL divergence:

$$\nabla \ell(\boldsymbol{\omega}) = -\nabla_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] + \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})).$$

As discussed in §A, the KL divergence has a simple closed-form gradient available when the variational distribution $q(\mathbf{z})$ and the prior $p(\mathbf{z})$ are both in the same exponential family:

$$\nabla_{\boldsymbol{\omega}} \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})) = \boldsymbol{\eta} - \boldsymbol{\eta}_{\mathrm{p}}.$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_{\mathrm{p}}$ are the natural parameters of $q(\mathbf{z})$ and $p(\mathbf{z})$, respectively. Therefore, the stochasticity solely comes from estimating the expected log likelihood, and the stochastic gradient $\widehat{\nabla} \ell(\boldsymbol{\omega})$ admits the form

$$\widehat{\nabla} \ell(\boldsymbol{\omega}) = -\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] + \nabla_{\boldsymbol{\omega}} \mathsf{D}_{\mathrm{KL}}(q(\mathbf{z}), p(\mathbf{z})).$$

For now, we assume the stochastic gradient of the expected log likelihood $(\widehat{\nabla}_{\boldsymbol{\xi}}, \widehat{\nabla}_{\boldsymbol{\Xi}}) = \widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$ has its second component $\widehat{\nabla}_{\boldsymbol{\Xi}}$ negative definite, a sufficient condition for valid stochastic NGD updates. We will show why this is the case in the upcoming section.

The next lemma shows that the natural parameter's second component $\boldsymbol{\Lambda}_t$ is bounded away from zero throughout the NGD updates, if the stochastic gradient $\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z})$ is always negative definite.

Lemma 5. *For Bayesian linear regression in Example 1, suppose the stochastic gradient $\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_{q_t(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{z}) \preceq 0$ and the step size $0 \leq \gamma_t \leq 1$ for all $t \geq 0$. Then, we have $\boldsymbol{\Lambda}_t \preceq -\frac{1}{2\nu}\mathbf{I}$, or equivalently $-\frac{1}{2}\boldsymbol{\Lambda}_t^{-1} \preceq \nu \mathbf{I}$, throughout the NGD updates for all $t \geq 0$, where $\nu = \max\{1, \|\mathbf{P}\|\} > 0$.*

Proof. We prove it by induction. The base case $\boldsymbol{\Lambda}_0 = -\frac{1}{2}\mathbf{I} \preceq -\frac{1}{2\nu}\mathbf{I}$ satisfies the inequality trivially. For $t \geq 1$, recall the NGD update on the natural parameter

$$\boldsymbol{\eta}_{t+1} = (1-\gamma_t)\boldsymbol{\eta}_t + \gamma_t\big(\widehat{\nabla}_{\boldsymbol{\omega}} \mathbb{E}_{q_t(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z}) + \boldsymbol{\eta}_{\mathrm{p}}\big),$$

where $\boldsymbol{\eta}_{\mathrm{p}}$ is the natural parameter of the prior. This yields an update on the second component of the natural parameter:

$$\boldsymbol{\Lambda}_{t+1} = (1-\gamma_t)\boldsymbol{\Lambda}_t + \gamma_t\big(\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_{q_t(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z}) + \boldsymbol{\Lambda}_{\mathrm{p}}\big).$$

By the assumption $\widehat{\nabla}_{\boldsymbol{\Xi}} \mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z}) \preceq 0$, we have

$$\begin{aligned}
\boldsymbol{\Lambda}_{t+1} &\preceq (1-\gamma_t)\boldsymbol{\Lambda}_t + \gamma_t \boldsymbol{\Lambda}_{\mathrm{p}} \\
&\preceq (1-\gamma_t) \cdot \Big(-\frac{1}{2\nu}\Big) \cdot \mathbf{I} + \gamma_t \cdot \Big(-\frac{1}{2\nu}\Big) \cdot \mathbf{I} \\
&= -\frac{1}{2\nu}\mathbf{I},
\end{aligned}$$

where the second line uses the induction hypothesis and the fact that $\boldsymbol{\Lambda}_{\mathrm{p}} = -\frac{1}{2}\mathbf{P}^{-1} \preceq -\frac{1}{2\nu}\mathbf{I}$. □

The following lemma shows that the matrix inverse is a Lipschitz function in region bounded away from zero.

Lemma 6. *Suppose $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \preceq -\frac{1}{2\nu}\mathbf{I}$. Then, we have $\|\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}\|_{\mathrm{F}} \leq 4\nu^2 \|\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\|_{\mathrm{F}}$.*

Proof. Straightforward calculation gives a proof:

$$\begin{aligned}
\|\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}\|_{\mathrm{F}} &= \big\|\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}\big\|_{\mathrm{F}} \\
&\leq \big\|\boldsymbol{\Lambda}_1^{-1}(\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2)\boldsymbol{\Lambda}_2^{-1}\big\|_{\mathrm{F}} \\
&\leq \big\|\boldsymbol{\Lambda}_1^{-1}\big\| \cdot \|\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\|_{\mathrm{F}} \cdot \|\boldsymbol{\Lambda}_2^{-1}\| \\
&\leq 4\nu^2 \|\boldsymbol{\Lambda}_1 - \boldsymbol{\Lambda}_2\|_{\mathrm{F}},
\end{aligned}$$

where the third line uses the inequality $\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|_{\mathrm{F}}$. □

Additional Notations. Let $\boldsymbol{\eta}_t$ and $\boldsymbol{\omega}_t$ be the natural and expectation parameters of the Gaussian variational distribution q_t at the step t. Hence, we have $\boldsymbol{\omega} = \nabla A(\boldsymbol{\eta})$ and $\boldsymbol{\eta} = \nabla A^*(\boldsymbol{\omega})$. Define

$$\boldsymbol{\eta}_{t+1,*} = \boldsymbol{\eta}_t - \gamma_t \nabla \ell(\boldsymbol{\omega}_t)$$

as the natural parameter after a NGD update from $\boldsymbol{\eta}_t$ using the exact (natural) gradient $\nabla \ell(\boldsymbol{\omega}_t)$. Recall that

$$\boldsymbol{\omega}_{t+1,*} = \operatorname*{argmin}_{\boldsymbol{\omega} \in \Omega} \langle \nabla \ell(\boldsymbol{\omega}_t), \boldsymbol{\omega} - \boldsymbol{\omega}_t \rangle + \frac{1}{\gamma_t} \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_t)$$

is the expectation parameter after a mirror descent update from $\boldsymbol{\omega}_t$ using the exact gradient $\nabla \ell(\boldsymbol{\omega}_t)$. Recall the relation $\boldsymbol{\omega}_{t+1,*} = \nabla A(\boldsymbol{\eta}_{t+1,*})$ based on the equivalence of NGD and mirror descent. The components of $\boldsymbol{\eta}_{t+1,*}$ and $\boldsymbol{\omega}_{t+1,*}$, i.e.

$$\boldsymbol{\eta}_{t+1,*} = (\boldsymbol{\lambda}_{t+1,*}, \boldsymbol{\Lambda}_{t+1,*}), \qquad \boldsymbol{\omega}_{t+1,*} = (\boldsymbol{\xi}_{t+1,*}, \boldsymbol{\Xi}_{t+1,*}),$$

are marked with "$*$" in the subscript as well.

C.1. Data Sub-Sampling Stochastic Gradient

The stochastic gradient (15) uses the following the estimate of the expected log likelihood

$$\begin{aligned}
\widehat{\nabla}_{\boldsymbol{\omega}}\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] &= \nabla_{\boldsymbol{\omega}}\left[\frac{n}{m}\sum_{k=1}^{m}\mathbb{E}_{q(\mathbf{z})}\log p(y_{i_k} \mid \mathbf{x}_{i_k}, \mathbf{z})\right] \\
&= \nabla_{\boldsymbol{\omega}}\left[(-1)\cdot\frac{1}{\sigma^2}\frac{n}{m}\sum_{k=1}^{m}\mathbb{E}_{q(\mathbf{z})}\left[\frac{1}{2}(y_{i_k} - \mathbf{z}^\top\mathbf{x}_{i_k})^2\right]\right] \\
&= \nabla_{\boldsymbol{\omega}}\left[(-1)\cdot\frac{1}{\sigma^2}\frac{n}{m}\sum_{k=1}^{m}\mathbb{E}_{q(\mathbf{z})}\left[\frac{1}{2}(\mathbf{z}^\top\mathbf{x}_{i_k})^2 - y_{i_k}\mathbf{z}^\top\mathbf{x}_{i_k} + \frac{1}{2}y_{i_k}^2\right]\right] \\
&= -\frac{1}{\sigma^2}\frac{n}{m}\nabla_{\boldsymbol{\omega}}\left[\sum_{k=1}^{m}\Big(\frac{1}{2}\langle\mathbf{x}_{i_k}\mathbf{x}_{i_k}^\top, \boldsymbol{\Xi}\rangle - \langle y_{i_k}\mathbf{x}_{i_k}, \boldsymbol{\xi}\rangle\Big)\right],
\end{aligned}$$

where we note that the stochastic gradient's second component $\widehat{\nabla}_{\boldsymbol{\Xi}}\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})]$ is indeed negative definite—a requirement for the NGD updates to stay inside the domain (recall Assumption 1). We obtain a concrete expression for the data sub-sampling stochastic gradient $\widehat{\nabla}\ell(\boldsymbol{\omega}) = \big(\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}), \widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega})\big)$ as follows:

$$\begin{aligned}
\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) &= -\frac{1}{\sigma^2}\frac{n}{m}\sum_{k=1}^{m}y_{i_k}\mathbf{x}_{i_k} + \boldsymbol{\lambda} - \boldsymbol{\lambda}_{\mathrm{p}}, \\
\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) &= \frac{1}{\sigma^2}\frac{n}{m}\sum_{k=1}^{m}\frac{1}{2}\mathbf{x}_{i_k}\mathbf{x}_{i_k}^\top + \boldsymbol{\Lambda} - \boldsymbol{\Lambda}_{\mathrm{p}},
\end{aligned}$$

where $\boldsymbol{\eta} = (\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is the natural parameter of the variational distribution $q(\mathbf{z})$ and $\boldsymbol{\eta}_{\mathrm{p}} = (\boldsymbol{\lambda}_{\mathrm{p}}, \boldsymbol{\Lambda}_{\mathrm{p}})$ is the natural parameter of the prior $p(\mathbf{z})$. Meanwhile the exact gradient is

$$\begin{aligned}
\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}y_i\mathbf{x}_i + \boldsymbol{\lambda} - \boldsymbol{\lambda}_{\mathrm{p}}, \\
\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{1}{2}\mathbf{x}_i\mathbf{x}_i^\top + \boldsymbol{\Lambda} - \boldsymbol{\Lambda}_{\mathrm{p}}.
\end{aligned}$$

Roadmap. We give a brief overview before diving into the detailed proof of Lemma 3, which involves a large amount of (somewhat tedious) calculation. Lemmas 7 and 8 give bounds on the gradient variances $\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|^2 \mid \boldsymbol{\omega}_t\big]$ and $\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\|_{\mathrm{F}}^2 \mid \boldsymbol{\omega}_t\big]$ measured in the Euclidean norm. These two bounds, however, are not quite enough for the convergence proof, as the desired gradient variance (12) does not depend on a specific norm. Lemmas 10 and 11 bound $\|\boldsymbol{\xi}_{t+1,*} - \boldsymbol{\xi}_{t+1}\|$ and $\|\boldsymbol{\Xi}_{t+1,*} - \boldsymbol{\Xi}_{t+1}\|_{\mathrm{F}}$ with $\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|$ and $\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\|_{\mathrm{F}}$. Lemma 3 utilizes Lemmas 10 and 11 to reduce the gradient variance (12), a norm-independent one, to the usual gradient variance measured in the Euclidean norm, which is readily tackled by Lemma 7 and Lemma 8.

Lemma 7. *The following inequality holds:*

$$\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|^2 \mid \boldsymbol{\omega}_t\big] = \frac{n^2}{m}\frac{s_1}{\sigma^4}$$

where we recall that $s_1 = \mathbb{E}_{j\sim U[n]}\big\|y_j\mathbf{x}_j - \frac{1}{n}\sum_{i=1}^{n}y_i\mathbf{x}_i\big\|^2$ *is the variance of* $y_j\mathbf{x}_j$.

Proof. Straightforward calculation gives a proof:

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|^2 \mid \boldsymbol{\omega}_t\big] &= \frac{1}{\sigma^4}\mathbb{E}\bigg\|\frac{n}{m}\sum_{k=1}^m y_{i_k}\mathbf{x}_{i_k} - \sum_{i=1}^n y_i\mathbf{x}_i\bigg\|^2 \\ &= \frac{1}{\sigma^4}\frac{n^2}{m^2}\mathbb{E}\bigg\|\sum_{k=1}^m\Big(y_{i_k}\mathbf{x}_{i_k} - \frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i\Big)\bigg\|^2 \\ &= \frac{1}{\sigma^4}\frac{n^2}{m}\mathbb{E}_{j\sim U[n]}\bigg\|y_j\mathbf{x}_j - \frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i\bigg\|^2 \\ &= \frac{n^2}{m}\frac{s_1}{\sigma^4}\end{aligned}$$

where the third line uses the fact that i_k's are independently sampled from the uniform distribution $U[n]$. □

Lemma 8. *The following inequality holds:*

$$\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\|_{\mathrm{F}}^2 \mid \boldsymbol{\omega}_t\big] = \frac{1}{4}\frac{n^2}{m}\frac{s_2}{\sigma^4}, \tag{19}$$

where we recall that $s_2 = \mathbb{E}_{j\sim U[n]}\big\|\mathbf{x}_j\mathbf{x}_j^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\big\|_{\mathrm{F}}^2$ *is the variance of* $\mathbf{x}_j\mathbf{x}_j^\top$.

Proof. A straightforward calculation gives a proof:

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\|_{\mathrm{F}}^2 \mid \boldsymbol{\omega}_t\big] &= \mathbb{E}\bigg\|\frac{1}{2}\frac{n}{m}\sum_{k=1}^m \mathbf{x}_{i_k}\mathbf{x}_{i_k}^\top - \frac{1}{2}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\bigg\|_{\mathrm{F}}^2 \\ &= \frac{1}{4}\frac{n^2}{m^2}\mathbb{E}\bigg\|\sum_{k=1}^m\Big(\mathbf{x}_{i_k}\mathbf{x}_{i_k}^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\Big)\bigg\|_{\mathrm{F}}^2 \\ &= \frac{1}{4}\frac{n^2}{m}\mathbb{E}_{j\sim U[n]}\bigg\|\mathbf{x}_j\mathbf{x}_j^\top - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\bigg\|_{\mathrm{F}}^2 \\ &= \frac{1}{4}\frac{n^2}{m}\frac{s_2}{\sigma^4},\end{aligned}$$

where the third line uses the fact that i_k's are sampled independently from the uniform distribution $U[n]$. □

Our proof strategy is to relate the desired gradient variance (12) with the gradient variances in Lemmas 7 and 8. To establish the relation, we need to show the natural parameter's first component $\boldsymbol{\lambda}_t$ and the expectation parameter's first component $\boldsymbol{\xi}_t$ are bounded throughout the NGD updates. The trick is to observe that the natural parameter's first component $\boldsymbol{\lambda}_t$ stays in a particular region:

Lemma 9. *Define the convex set* $\mathcal{C} = \big\{\sum_{i=1}^n \rho_i y_i\mathbf{x}_i : \rho_i \geq 0, \sum_{i=1}^n \rho_i \leq n\big\}$. *Then, we have* $\boldsymbol{\lambda}_t \in \mathcal{C}$ *and* $\boldsymbol{\lambda}_{t+1,*} \in \mathcal{C}$ *throughout the NGD updates for all* $t \geq 0$.

Proof. We prove $\boldsymbol{\lambda}_t \in \mathcal{C}$ by induction. The base case $t = 0$ holds as the initialization $q_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ has $\boldsymbol{\lambda}_0 = \mathbf{0}$, with coefficients $\rho_1 = \rho_2 = \cdots = \rho_n = 0$. For $t \geq 1$, recall the update from t to $t+1$:

$$\begin{aligned}\boldsymbol{\lambda}_{t+1} &= (1-\gamma_t)\boldsymbol{\lambda}_t + \gamma_t\big(\widehat{\nabla}_{\boldsymbol{\xi}}\mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] + \boldsymbol{\lambda}_{\mathrm{p}}\big) \\ &= (1-\gamma_t)\boldsymbol{\lambda}_t + \gamma_t\widehat{\nabla}_{\boldsymbol{\xi}}\mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})],\end{aligned}$$

where the second line uses $\boldsymbol{\lambda}_{\mathrm{p}} = \mathbf{0}$ since the prior $p(\mathbf{z})$ is a zero-mean Gaussian. Recall that the stochastic gradient of the expected log likelihood at the step t is of the form

$$\widehat{\nabla}_{\boldsymbol{\xi}}\mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] = \frac{n}{m}\sum_{k=1}^m y_{i_k}\mathbf{x}_{i_k},$$

where i_k's are sampled independently and uniformly from $\{1, 2, \cdots, n\}$. The stochastic gradient $\widehat{\nabla}_{\boldsymbol{\xi}} \mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})]$ is in the convex set $\mathcal{C}$, since the sum of its coefficients is exactly n. Observe that $\boldsymbol{\lambda}_{t+1}$ is a convex combination of two points in $\mathcal{C}$, and thus stays in $\mathcal{C}$ as well. The proof is completed by an induction.

The argument for $\boldsymbol{\lambda}_{t+1,*} \in \mathcal{C}$ follows similarly, because the exact gradient of $\nabla_{\boldsymbol{\xi}} \mathbb{E}_{q_t(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] = \sum_{i=1}^n y_i \mathbf{x}_i$ is in the convex set $\mathcal{C}$ as well. □

As a result, we immediately obtain a bound on the first component of the natural parameter:

Corollary 1. *We have $\|\boldsymbol{\lambda}_t\| \le bn$ and $\|\boldsymbol{\lambda}_{t,*}\| \le bn$ for all $t \ge 0$, where $b = \max_{1 \le i \le n} \|y_i \mathbf{x}_i\|$.*

Proof. Straightforward calculation gives a proof:

$$\|\boldsymbol{\lambda}_t\| = \Big\| \sum_{i=1}^n \rho_i y_i \mathbf{x}_i \Big\| \le \sum_{i=1}^n \rho_i \|y_i \mathbf{x}_i\| \le bn.$$

The proof for $\boldsymbol{\lambda}_{t,*}$ follows the same steps. □

As a result, we also obtain a bound on the first component of the expectation parameter:

Corollary 2. *We have $\|\boldsymbol{\xi}_t\| \le \nu bn$ and $\|\boldsymbol{\xi}_{t,*}\| \le \nu bn$ for all $t \ge 0$, where $b = \max_{1 \le i \le n} \|y_i \mathbf{x}_i\|$.*

Proof. Recall the relation between the natural and expectation parameters: $\boldsymbol{\xi}_t = -\frac{1}{2} \boldsymbol{\Lambda}_t^{-1} \boldsymbol{\lambda}_t$. Recall that $0 \preceq -\frac{1}{2} \boldsymbol{\Lambda}_t^{-1} \preceq \nu \mathbf{I}$ by Lemma 5. Thus, we have $\|\boldsymbol{\xi}_t\| \le \nu \|\boldsymbol{\lambda}_t\| \le \nu bn$. □

Lemma 10. *We have $\|\boldsymbol{\xi}_{t+1,*} - \boldsymbol{\xi}_{t+1}\| \le \gamma_t \nu \|\widehat{\nabla}_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t)\| + 2\gamma_t \nu^2 bn \big\|\widehat{\nabla}_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}$.*

Proof. Straightforward calculation gives

$$\begin{aligned}
\|\boldsymbol{\xi}_{t+1,*} - \boldsymbol{\xi}_{t+1}\| &= \frac{1}{2} \|\boldsymbol{\Lambda}_{t+1}^{-1} \boldsymbol{\lambda}_{t+1} - \boldsymbol{\Lambda}_{t+1,*}^{-1} \boldsymbol{\lambda}_{t+1,*}\| \\
&= \frac{1}{2} \|\boldsymbol{\Lambda}_{t+1}^{-1} \boldsymbol{\lambda}_{t+1} - \boldsymbol{\Lambda}_{t+1}^{-1} \boldsymbol{\lambda}_{t+1,*} + \boldsymbol{\Lambda}_{t+1}^{-1} \boldsymbol{\lambda}_{t+1,*} - \boldsymbol{\Lambda}_{t+1,*}^{-1} \boldsymbol{\lambda}_{t+1,*}\| \\
&\le \frac{1}{2} \|\boldsymbol{\Lambda}_{t+1}^{-1} (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_{t+1,*})\| + \frac{1}{2} \|(\boldsymbol{\Lambda}_{t+1}^{-1} - \boldsymbol{\Lambda}_{t+1,*}^{-1}) \boldsymbol{\lambda}_{t+1,*}\|,
\end{aligned}$$

where the first line uses the relation between the natural and expectation parameters. We cope with the two terms separately.

For the first term, we have

$$\frac{1}{2} \|\boldsymbol{\Lambda}_{t+1}^{-1} (\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_{t+1,*})\| \le \nu \|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_{t+1,*}\| = \gamma_t \nu \|\widehat{\nabla}_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t)\|,$$

where the first inequality uses $-\frac{1}{2} \boldsymbol{\Lambda}_{t+1}^{-1} \preceq \nu \mathbf{I}$ by Lemma 5; the second equality uses the definition of the NGD update.

For the second term, we have

$$\begin{aligned}
\frac{1}{2} \|(\boldsymbol{\Lambda}_{t+1}^{-1} - \boldsymbol{\Lambda}_{t+1,*}^{-1}) \boldsymbol{\lambda}_{t+1,*}\| &\le \frac{1}{2} \|\boldsymbol{\Lambda}_{t+1}^{-1} - \boldsymbol{\Lambda}_{t+1,*}^{-1}\|_{\mathrm{F}} \cdot \|\boldsymbol{\lambda}_{t+1,*}\| \\
&\le 2\nu^2 \cdot \|\boldsymbol{\Lambda}_{t+1} - \boldsymbol{\Lambda}_{t+1,*}\|_{\mathrm{F}} \cdot \|\boldsymbol{\lambda}_{t+1,*}\| \\
&= 2\gamma_t \nu^2 \big\|\widehat{\nabla}_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}} \cdot \|\boldsymbol{\lambda}_{t+1,*}\| \\
&\le 2\gamma_t \nu^2 bn \big\|\widehat{\nabla}_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}},
\end{aligned}$$

where the second line uses the Lipschitz condition in Lemma 6; the third line uses the the definition of the NGD update; the last line uses Corollary 1. Summing the two bounds completes the proof. □

Lemma 11. *We have*

$$\|\boldsymbol{\Xi}_{t+1,*} - \boldsymbol{\Xi}_{t+1}\|_{\mathrm{F}} \le 2\gamma_t \nu^2 bn \big\|\widehat{\nabla}_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}} \ell(\boldsymbol{\omega}_t)\big\| + (2\gamma_t \nu^2 + 4\gamma_t \nu^3 b^2 n^2) \big\|\widehat{\nabla}_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}} \ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}.$$

Proof. Expanding the norm, we have

$$\begin{aligned}\|\boldsymbol{\Xi}_{t+1,*}-\boldsymbol{\Xi}_{t+1}\|_{\mathrm{F}} &= \Big\|-\frac{1}{2}\big(\boldsymbol{\Lambda}_{t+1,*}^{-1}-\boldsymbol{\Lambda}_{t+1}^{-1}\big)+\big(\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1,*}^{\top}-\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}^{\top}\big)\Big\|_{\mathrm{F}}\\ &\le \frac{1}{2}\Big\|\boldsymbol{\Lambda}_{t+1,*}^{-1}-\boldsymbol{\Lambda}_{t+1}^{-1}\Big\|_{\mathrm{F}}+\big\|\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1,*}^{\top}-\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}^{\top}\big\|_{\mathrm{F}},\end{aligned}$$

where the first line uses the relation between natural and expectation parameters.

For the first term, we have

$$\begin{aligned}\frac{1}{2}\big\|\boldsymbol{\Lambda}_{t+1,*}^{-1}-\boldsymbol{\Lambda}_{t+1}^{-1}\big\|_{\mathrm{F}} &\le 2\nu^2\big\|\boldsymbol{\Lambda}_{t+1,*}-\boldsymbol{\Lambda}_{t+1}\big\|_{\mathrm{F}}\\ &\le 2\gamma_t\nu^2\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}},\end{aligned}$$

where the first line uses Lemma 6; the second line uses the definition of the NGD update.

For the second term, we have

$$\begin{aligned}\big\|\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1,*}^{\top}-\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}^{\top}\big\|_{\mathrm{F}} &= \big\|\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1,*}^{\top}-\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1}^{\top}+\boldsymbol{\xi}_{t+1,*}\boldsymbol{\xi}_{t+1}^{\top}-\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}_{t+1}^{\top}\big\|_{\mathrm{F}}\\ &\le \big\|\boldsymbol{\xi}_{t+1,*}(\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1})^{\top}\big\|_{\mathrm{F}}+\big\|(\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1})\boldsymbol{\xi}_{t+1}^{\top}\big\|_{\mathrm{F}}\\ &= \|\boldsymbol{\xi}_{t+1,*}\|\cdot\|\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\|+\|\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\|\cdot\|\boldsymbol{\xi}_{t+1}\|\\ &\le 2\max\{\|\boldsymbol{\xi}_{t+1,*}\|,\|\boldsymbol{\xi}_{t+1}\|\}\cdot\|\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\|\\ &\le 2\nu bn\|\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\|\\ &\le 2\gamma_t\nu^2 bn\big\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\big\|+4\gamma_t\nu^3b^2n^2\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}\end{aligned}$$

where the third line is because $\|\mathbf{a}\mathbf{b}^{\top}\|_{\mathrm{F}}=\|\mathbf{a}\|\cdot\|\mathbf{b}\|$; the fifth line uses Corollary 2; the last line uses Lemma 10. Summing the above two parts finishes the proof. □

Now we are ready to prove the main results of this section, the variance bound of data sub-sampling stochastic gradient.

Lemma 3. *The stochastic gradient* (15) *satisfies*

$$\frac{1}{\gamma_t}\mathbb{E}[\langle\widehat{\nabla}\ell(\boldsymbol{\omega}_t)-\nabla\ell(\boldsymbol{\omega}_t),\boldsymbol{\omega}_{t+1,*}-\boldsymbol{\omega}_{t+1}\rangle\mid\boldsymbol{\omega}_t]\le V_2, \tag{16}$$

where $V_2=(\nu s_1+\frac{1}{2}\nu^2s_2+2\nu^2b\sqrt{s_1s_2}n+\nu^3b^2s_2n^2)\frac{n^2}{\sigma^4m}$, *with* $\nu=\max\{1,\|\mathbf{P}\|\}$, $b=\max_{1\le i\le n}\|y_i\mathbf{x}_i\|$, *and the empirical variances* $s_1=\mathbb{E}_{j\sim U[n]}\big\|y_j\mathbf{x}_j-\frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i\big\|^2$. *and* $s_2=\mathbb{E}_{j\sim U[n]}\big\|\mathbf{x}_j\mathbf{x}_j^{\top}-\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^{\top}\big\|_{\mathrm{F}}^2$.

Proof. Expanding the inner product inside the expectation, we need to bound the expectation of

$$\big\langle\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\big\rangle+\big\langle\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\Xi}_{t+1,*}-\boldsymbol{\Xi}_{t+1}\big\rangle.$$

For the first term $\big\langle\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\big\rangle$, applying the Cauchy-Schwarz inequality and Lemma 10 yields

$$\begin{aligned}\big\langle\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\big\rangle &\le \|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|\cdot\|\boldsymbol{\xi}_{t+1,*}-\boldsymbol{\xi}_{t+1}\|\\ &\le \gamma_t\nu\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|^2+2\gamma_t\nu^2bn\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|\cdot\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\|_{\mathrm{F}}.\end{aligned} \tag{20}$$

For the second term $\big\langle\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\Xi}_{t+1,*}-\boldsymbol{\Xi}_{t+1}\big\rangle$, applying the Cauchy-Schwarz inequality and Lemma 11 gives

$$\begin{aligned}&\big\langle\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t),\boldsymbol{\Xi}_{t+1,*}-\boldsymbol{\Xi}_{t+1}\big\rangle\le\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}\cdot\|\boldsymbol{\Xi}_{t+1,*}-\boldsymbol{\Xi}_{t+1}\|_{\mathrm{F}}\\ &\le 2\gamma_t\nu^2bn\big\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\big\|\cdot\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}+(2\gamma_t\nu^2+4\gamma_t\nu^3b^2n^2)\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}^2\end{aligned} \tag{21}$$

Summing (20) and (21), and then applying the inequality

$$\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\big\|\cdot\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}\Big]\le\sqrt{\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\big\|^2\Big]\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)-\nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}^2\Big]},$$

where the expectations are conditioned on $\boldsymbol{\omega}_t$, we obtain a bound on as follows:

$$\begin{aligned}
\mathbb{E}[\langle\widehat{\nabla}\ell(\boldsymbol{\omega}_t) - \nabla\ell(\boldsymbol{\omega}_t), \boldsymbol{\omega}_{t+1,*} - \boldsymbol{\omega}_{t+1}\rangle \mid \boldsymbol{\omega}_t] &\leq \gamma_t\nu\mathbb{E}\big[\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\|^2 \mid \boldsymbol{\omega}_t\big] \\
&+ (2\gamma_t\nu^2 + 4\gamma_t\nu^3 b^2 n^2)\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}^2 \mid \boldsymbol{\omega}_t\Big] \\
&+ 4\gamma_t\nu^2 bn\sqrt{\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega}_t)\big\|^2 \mid \boldsymbol{\omega}_t\Big]\mathbb{E}\Big[\big\|\widehat{\nabla}_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t) - \nabla_{\boldsymbol{\Xi}}\ell(\boldsymbol{\omega}_t)\big\|_{\mathrm{F}}^2 \mid \boldsymbol{\omega}_t\Big]} \\
&\leq \gamma_t\nu\cdot\frac{n^2}{m}\frac{s_1}{\sigma^4} + (2\gamma_t\nu^2 + 4\gamma_t\nu^3 b^2 n^2)\cdot\frac{1}{4}\frac{n^2}{m}\frac{s_2}{\sigma^4} + 4\gamma_t\nu^2 bn\cdot\frac{1}{2}\frac{n^2}{m}\frac{\sqrt{s_1 s_2}}{\sigma^4} \\
&= \gamma_t\nu\frac{n^2}{m}\frac{s_1}{\sigma^4} + +\frac{1}{2}\gamma_t\nu^2\frac{n^2}{m}\frac{s_2}{\sigma^4} + 2\gamma_t\nu^2 b\frac{n^3}{m}\frac{\sqrt{s_1 s_2}}{\sigma^4} + \gamma_t\nu^3 b^2\frac{n^4}{m}\frac{s_2}{\sigma^4} \\
&= \frac{1}{\sigma^4}\gamma_t(\nu s_1 + +\frac{1}{2}\nu^2 s_2 + 2\nu^2 b\sqrt{s_1 s_2}n + \nu^3 b^2 s_2 n^2)\frac{n^2}{m}
\end{aligned}$$

where the second equality is due to Lemma 7 and Lemma 8. Dividing both sides by γ_t completes the proof. □

D. Proof of the Main Theorem

Lemma 2. *For conjugate likelihoods, the negative ELBO $\ell(\boldsymbol{\omega})$ is 1-smooth 1-strongly convex relative to the convex conjugate A^* of the log-partition function.*

Proof. Let $q^*(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{y})$ be the posterior. By the definition of the negative ELBO, we have $\ell(\boldsymbol{\omega}) = \mathsf{D}_{\mathrm{KL}}(q, q^*) + C$, where $q \in \mathcal{Q}$ is the variational distribution inside an exponential family $\mathcal{Q}$ parameterized by the expectation parameter $\boldsymbol{\omega}$ and $C = p(\mathbf{y})$ is a constant (log evidence) that does not depend on q and $\boldsymbol{\omega}$.

Thanks to conjugacy, the posterior q^* is of the same form as q. By Lemma 4, $\ell(\boldsymbol{\omega}) \propto \mathsf{D}_{\mathrm{KL}}(q, q^*) = \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}^*)$. Observe that the Bregman divergence $\mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}^*)$ is trivially 1-smooth and 1-strongly convex in $\boldsymbol{\omega}$ relative to A^*. □

Below we present the main theorem, which adapts the results by Hanzely & Richtárik (2021) to stochastic natural gradient variational inference.

Theorem 1. *Suppose the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is conjugate and the stochastic gradient $\widehat{\nabla}\ell(\boldsymbol{\omega}_t)$ satisfies Assumption 1. Running $T+1$ iterations of stochastic natural gradient descent with $\gamma_t = \frac{2}{2+t}$ generate a point $\bar{\boldsymbol{\omega}}_{T+1}$ that satisfies*

$$\mathbb{E}[\ell(\bar{\boldsymbol{\omega}}_{T+1})] - \min_{\boldsymbol{\omega}\in\Omega}\ell(\boldsymbol{\omega}) \leq \frac{V}{T+2}, \tag{13}$$

where $\bar{\boldsymbol{\omega}}_{T+1} = \frac{2}{(T+1)(T+2)}\sum_{t=0}^{T}(t+1)\boldsymbol{\omega}_{t+1}$. Let $\bar{q}_{T+1}$ be the variational distribution represented by $\bar{\boldsymbol{\omega}}_{T+1}$. Then, the KL divergence to the true posterior q^* is bounded by

$$\mathbb{E}[\mathsf{D}_{\mathrm{KL}}(\bar{q}_{T+1}, q^*)] \leq \frac{V}{T+2}. \tag{14}$$

Proof. By the descent lemma of Hanzely & Richtárik (2021, Lemma 5.2), we have

$$\mathbb{E}[\ell(\boldsymbol{\omega}_{t+1})] - \ell(\boldsymbol{\omega}^*) \leq \Big(\frac{1}{\gamma_t} - 1\Big)\mathsf{D}_{A^*}(\boldsymbol{\omega}^*, \boldsymbol{\omega}_t) - \frac{1}{\gamma_t}\mathbb{E}[\mathsf{D}_{A^*}(\boldsymbol{\omega}^*, \boldsymbol{\omega}_{t+1})] + \gamma_t V.$$

Plugging in $\gamma_t = \frac{2}{2+t}$, we obtain

$$\mathbb{E}[\ell(\boldsymbol{\omega}_{t+1})] - \ell(\boldsymbol{\omega}^*) \leq \frac{1}{2}t\cdot\mathsf{D}_{A^*}(\boldsymbol{\omega}^*, \boldsymbol{\omega}_t) - \frac{1}{2}\big(t+2\big)\mathbb{E}[\mathsf{D}_{A^*}(\boldsymbol{\omega}^*, \boldsymbol{\omega}_{t+1})] + \gamma_t V.$$

Multiply the inequality by $t+1$ and sum from 0 to T. Then we have

$$\sum_{t=0}^{T}(t+1)(\ell(\boldsymbol{\omega}_{t+1}) - \ell(\boldsymbol{\omega}^*)) \leq \frac{1}{2}V\sum_{t=0}^{T}\frac{t+1}{t+2} \leq \frac{1}{2}V(T+1).$$

Dividing both sides by $\sum_{t=0}^{T}(t+1) = \frac{1}{2}(T+1)(T+2)$, and use the convexity of f, we obtain

$$\ell(\bar{\boldsymbol{\omega}}_{T+1}) \leq \frac{V}{T+2}$$

To get the convergence rate in terms of the KL divergence, notice that

$$\begin{aligned}\ell(\bar{\boldsymbol{\omega}}_{t+1}) - \ell(\boldsymbol{\omega}^*) &= \nabla\ell(\boldsymbol{\omega}^*)^\top(\boldsymbol{\omega}_{T+1} - \boldsymbol{\omega}^*) + \mathsf{D}_{A^*}(\bar{\boldsymbol{\omega}}_{T+1}, \boldsymbol{\omega}^*) \\ &= \mathsf{D}_{A^*}(\bar{\boldsymbol{\omega}}_{T+1}, \boldsymbol{\omega}^*) \\ &= \mathsf{D}_{\mathrm{KL}}(\bar{q}_{T+1}, q^*),\end{aligned}$$

where the first line is due to 1-smoothness and 1-strong convexity relative to A^*; the second line is because the optimal parameter $\boldsymbol{\omega}^*$ has zero gradient; the third line is due to Lemma 4. □

E. Missing Proofs in §5

Proposition 1. *Suppose the prior and the variational family are both Gaussians. If the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is log-concave in $\mathbf{z}$, then the negative ELBO $\ell(\boldsymbol{\omega})$ as a function of the expectation parameter has a unique minimizer $\boldsymbol{\omega}^*$. In addition, if the likelihood $p(\mathbf{y} \mid \mathbf{z})$ is differentiable in $\mathbf{z}$, then $\boldsymbol{\omega}^*$ is the unique stationary point of $\ell(\boldsymbol{\omega})$.*

Proof. Consider the set

$$\Theta = \{\boldsymbol{\theta} = \big(\boldsymbol{\mu}, \mathbf{C}\big) : \boldsymbol{\mu} \in \mathbb{R}^d, \mathbf{C} \in \mathbb{S}^d_{++}\}$$

which parameterizes all (non-degenerate) Gaussian distributions. Define $f(\boldsymbol{\theta}) = \big(\boldsymbol{\mu}, \mathbf{C}\mathbf{C}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top\big)$. Namely, f maps $\boldsymbol{\theta}$ to the expectation parameter space Ω. Thanks to the uniqueness of matrix square root, f is a bijection.

Since $\ell^{(\mathrm{mr})}$ is strongly convex in $\boldsymbol{\theta}$, it has a unique minimizer $\boldsymbol{\theta}^* \in \Theta$. Define $\boldsymbol{\omega}^* = f(\boldsymbol{\theta}^*)$. It is clear that $\boldsymbol{\omega}^* \in \Omega$ is the unique minimizer of $\ell^{(\mathrm{e})}$.

Consider the identity

$$\ell^{(\mathrm{mr})}(\boldsymbol{\theta}) = \ell^{(\mathrm{e})}(f(\boldsymbol{\theta})). \tag{22}$$

Taking the derivative of (22) on both sides, we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\ell^{(\mathrm{mr})}(\boldsymbol{\theta}) &= \left.\frac{\partial \ell^{(\mathrm{e})}}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\omega}=f(\boldsymbol{\theta})} + 2 \cdot \left.\frac{\partial \ell^{(\mathrm{e})}}{\partial \boldsymbol{\Xi}}\right|_{\boldsymbol{\omega}=f(\boldsymbol{\theta})} \cdot \boldsymbol{\mu} \\ \frac{\partial}{\partial \mathbf{C}}\ell^{(\mathrm{mr})}(\boldsymbol{\theta}) &= 2 \cdot \left.\frac{\partial \ell^{(\mathrm{e})}}{\partial \boldsymbol{\Xi}}\right|_{\boldsymbol{\omega}=f(\boldsymbol{\theta})} \cdot \mathbf{C},\end{aligned}$$

It easy to see that $\nabla\ell^{(\mathrm{mr})}(\boldsymbol{\theta}) = \mathbf{0}$ iff $\nabla\ell^{(\mathrm{e})}(f(\boldsymbol{\theta})) = \mathbf{0}$. Namely, f maps stationary points to stationary points. Since there is only one stationary point in Θ due strong convexity, there is only one stationary point in Ω as well. □

E.1. Bayesian Logistic Regression

We give a more detailed description of the non-convexity of Bayesian logistic regression. Recall that we focus on the restriction of $\ell(\boldsymbol{\omega})$ on the convex subset

$$\{\boldsymbol{\omega} = (\mathbf{0}, \boldsymbol{\Xi}) : \boldsymbol{\Xi} = \mathrm{diag}(s_1, s_2), s_1 > 0, s_2 > 0\} \subseteq \Omega.$$

Observe that $wx_i + b$ follows a Gaussian distribution $\mathcal{N}(0, x_i^2 s_1 + s_2)$. Therefore, we can use the Price theorem to take the derivative *w.r.t.* s_2. Taking the first-order derivative of $\ell(\boldsymbol{\omega})$ *w.r.t.* s_2, we have

$$\frac{\partial}{\partial s_2}\ell(\boldsymbol{\omega}) = \sum_{i=1}^{n} \mathbb{E}_{q(w,b)}[\psi_i(1-\psi_i)] + \frac{1}{2} - \frac{1}{2s_2},$$

where we use ψ_i to denote $\psi(wx_i+b)$ and ψ is the sigmoid function. Using the Price theorem again to take the second-order derivative of $\ell(\boldsymbol{\omega})$ *w.r.t.* s_2, we have

$$\frac{\partial^2}{\partial s_2^2}\ell(\boldsymbol{\omega}) = \sum_{i=1}^{n}\mathbb{E}_{q(w,b)}[\psi_i(1-\psi_i)(6\psi_i^2-6\psi_i+1)] + \frac{1}{2s_2^2},$$

Note that $\frac{\partial^2}{\partial s_2^2}\ell(\boldsymbol{\omega})$ is continuous *w.r.t.* s_1 and s_2. Moreover, we have

$$\lim_{s_1\to 0, s_2\to 0}\mathbb{E}[\psi_i(1-\psi_i)(6\psi_i^2-6\psi_i+1)] = -\frac{1}{8}.$$

Therefore, there exists a small positive constant $\delta > 0$, such that $s_1 = s_2 = \delta$ and

$$\mathbb{E}_{w\sim\mathcal{N}(0,s_1),b\sim\mathcal{N}(0,s_2)}[\psi_i(1-\psi_i)(6\psi_i^2-6\psi_i+1)] < -\frac{1}{16}.$$

Crucially, δ is an absolute constant that does not depend on i. Because all $-1 \le x_i \le 1$ are bounded, the distribution $wx_i + b \sim \mathcal{N}(0, x_i^2 s_1 + s_2)$ will shrink to zero as long as $s_1 + s_2 \to 0$, regardless of the index i. This implies that when $s_1 = s_2 = \delta$, we have

$$\sum_{i=1}^{n}\mathbb{E}_{w\sim\mathcal{N}(0,s_1),b\sim\mathcal{N}(0,s_2)}[\psi_i(1-\psi_i)(6\psi_i^2-6\psi_i+1)] < -\frac{1}{16}n.$$

Therefore, when $s_1 = s_2 = \delta$ and $n \ge \frac{8}{\delta^2}$, the second order derivative is negative

$$\frac{\partial^2}{\partial s_2^2}\ell(\boldsymbol{\omega}) < -\frac{1}{16}n + \frac{1}{2\delta^2} < 0,$$

which implies that the objective is non-convex in the expectation parameter.

E.2. Bayesian Poisson Regression

Bayesian Poisson regression assumes that $y \mid \mathbf{x}$ follows a Poisson distribution with the expectation

$$\mathbb{E}[y \mid \mathbf{x}] = \exp(\mathbf{w}^\top \mathbf{x}),$$

which gives the log likelihood

$$\log p(y \mid \mathbf{x}, \mathbf{w}) = -\log y! + y\mathbf{w}^\top\mathbf{x} - \exp(\mathbf{w}^\top\mathbf{x}).$$

We impose a Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and approximate the posterior $p(\mathbf{w} \mid \mathbf{y})$ using a Gaussian variational distribution $q(\mathbf{w})$. A nice property of the Bayesian Poisson regression is that its ELBO has a closed-form expression

$$\begin{aligned}\ell(\boldsymbol{\omega}) &= \sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{w})}[-y_i\mathbf{w}^\top\mathbf{x}_i + \exp(\mathbf{w}^\top\mathbf{x}_i)] + \mathsf{D}_{\mathrm{KL}}\big(q, p\big)\\ &= \sum_{i=1}^{n}\Big[-y_i\boldsymbol{\xi}^\top\mathbf{x}_i + \exp\Big(\boldsymbol{\xi}^\top\mathbf{x}_i + \frac{1}{2}\mathbf{x}_i^\top\big(\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top\big)\mathbf{x}_i\Big)\Big] + \mathsf{D}_{A^*}(\boldsymbol{\omega}, \boldsymbol{\omega}_0).\end{aligned}$$

The Hessian $\nabla^2_{\boldsymbol{\xi}}\ell(\boldsymbol{\omega})$ is

$$\sum_{i=1}^{n}\exp\Big(\boldsymbol{\xi}^\top\mathbf{x}_i + \frac{1}{2}\mathbf{x}_i^\top(\boldsymbol{\Xi} - \boldsymbol{\xi}\boldsymbol{\xi}^\top)\mathbf{x}_i\Big)\big(-1 + \big(1 - \mathbf{x}_i^\top\boldsymbol{\xi}\big)^2\big)\mathbf{x}_i\mathbf{x}_i^\top + \nabla^2_{\boldsymbol{\xi}}A^*(\boldsymbol{\omega}).$$

Evaluating the Hessian on the subset of the domain

$$\{\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\Xi}) \in \Omega : \boldsymbol{\Xi} = \boldsymbol{\xi}\boldsymbol{\xi}^\top + 2\mathbf{I}\},$$

we obtain the following

$$\sum_{i=1}^{n}\exp\big(\mathbf{x}_i^\top\boldsymbol{\xi} + \mathbf{x}_i^\top\mathbf{x}_i\big)\mathbf{x}_i^\top\boldsymbol{\xi}(\mathbf{x}_i^\top\boldsymbol{\xi} - 2)\mathbf{x}_i\mathbf{x}_i^\top + \nabla^2_{\boldsymbol{\xi}}A^*(\boldsymbol{\omega}).$$

With $0 < \mathbf{x}_i^\top\boldsymbol{\xi} < 2$ for all i, which can be satisfied by constructing the dataset properly, and using $\exp(\mathbf{x}_i^\top\boldsymbol{\xi} + \mathbf{x}_i^\top\mathbf{x}_i) > 1$, we can drop the exponential term. The rest of the argument follows the main paper.

F. Experimental Details

In all experiments, SGD uses the $(\mathbf{m}, \mathbf{C})$ parameterization, where $\mathbf{m}$ is the Gaussian mean and $\mathbf{C}$ is the Cholesky factor of the Gaussian covariance. We parameterize $\mathbf{C}$ as a lower triangular matrix with strictly positive diagonal entries. For SGD, we clamp the diagonal entries of $\mathbf{C}$ to make sure they are no smaller than 10^{-10}. This is effectively a projection step.

F.1. Bayesian Linear Regression

This is a Bayesian linear regression problem exactly the same as Example 1 with a standard Gaussian prior. Note that the expected log likelihood $\mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{z})$ is integrated in a closed-form. The only stochasticity comes from the mini-batch data sub-sampling. Domke et al. (2023, Theorem 7 and Theorem 10) have proved convergence for stochastic proximal (projected) gradient descent with a step size schedule $\gamma_t = \min\{\frac{\mu}{a}, \frac{1}{\mu}\frac{2t+1}{(t+1)^2}\}$. It is not easy to come up with a tight estimate of the constant a. Therefore, we pick the linearly decreasing schedule $\frac{1}{10^5+t}$ for SGD. The reason for the specific constant 10^5 in the denominator is that 10^{-5} is roughly the largest step size such that SGD does not diverge in its initial stage.

F.2. Bayesian Logistic Regression

On Mushroom, the step size of SGD is tuned from $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, while the step size of NGD is tuned from $\{5 \cdot 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}\}$. On MNIST, the step size of SGD is tuned from $\{10^{-5}, 10^{-6}, 10^{-7}\}$, while the step size of NGD is tuned from $\{10^{-1}, 10^{-2}, 10^{-3}\}$. Divergent curves (due to large step sizes) are not plotted in the graph. We use 10 samples from the variational distribution to estimate the stochastic gradient in every iteration.

Legends without the label "(p)" use the reparameterization trick to compute the stochastic gradient. For SGD with the label "(p)", we use the Price theorem as follows. First, observe the following relation between $\nabla_{\mathbf{C}}$ and $\nabla_{\mathbf{\Sigma}}$:

$$\begin{aligned}\nabla_{\mathbf{C}} \mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z}) &= 2 \cdot \nabla_{\mathbf{\Sigma}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y} \mid \mathbf{z})] \cdot \mathbf{C} \\ &= \mathbb{E}_{q(\mathbf{z})}[\nabla^2_{\mathbf{z}} \log p(\mathbf{y} \mid \mathbf{z})] \cdot \mathbf{C}.\end{aligned}$$

To obtain a stochastic gradient estimate $\widehat{\nabla}_{\mathbf{C}} \mathbb{E}_{q(\mathbf{z})} \log p(\mathbf{y} \mid \mathbf{z})$, replace the expectation with sample approximation.