On the Computational Asymptotics of Gaussian Variational Inference

by

Zuheng Xu

B.S., Sichuan University, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

September 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On the Computational Asymptotics of Gaussian Variational Inference

submitted by **Zuheng Xu** in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** in **Statistics**.

Examining Committee:

Trevor Campbell, Statistics
Supervisor

Alexandre Bouchard-Côté, Statistics
Additional Examiner

Abstract

Variational inference is a popular alternative to Markov chain Monte Carlo methods that constructs a Bayesian posterior approximation by minimizing a discrepancy to the true posterior within a pre-specified family. This converts Bayesian inference into an optimization problem, enabling the use of simple and scalable stochastic optimization algorithms. However, a key limitation of variational inference is that the optimal approximation is typically not tractable to compute; even in simple settings the problem is nonconvex. Thus, recently developed statistical guarantees—which all involve the (data) asymptotic properties of the optimal variational distribution—are not reliably obtained in practice. In this work, we provide two major contributions: a theoretical analysis of the asymptotic convexity properties of variational inference in the popular setting with a Gaussian family; and *consistent stochastic variational inference* (CSVI), an algorithm that exploits these properties to find the optimal approximation in the asymptotic regime. CSVI consists of a tractable initialization procedure that finds the local basin of the optimal solution, and a scaled gradient descent algorithm that stays locally confined to that basin. Experiments on nonconvex synthetic examples show that compared with standard stochastic gradient descent, CSVI improves the likelihood of obtaining the globally optimal posterior approximation.

Lay Summary

Bayesian inference is a statistical methodology for obtaining insights from data. Variational inference is a formulation of Bayesian inference as an optimization problem. The optimization problem is generally quite difficult to solve reliably; thus, although there is a wealth of previous work on understanding the statistical properties of the optimal solution, these guarantees are not achievable in practice. This thesis provides two major contributions: first, it provides a theoretical analysis of the computational properties of the optimization problem given a large amount of data; and then it exploits those properties to provide a simple and efficient scheme to solve the optimization problem more reliably.

Preface

This thesis is original, unpublished work by the author, Zuheng Xu, under the supervision of Professor Trevor Campbell. T. Campbell proposed the methodology in Chapter 4 and proved Theorems 3.3.2, 3.3.5 and A.2.1; aside from these, all other results, software, and experiments were contributed by the author.

Table of Contents

List of Figures

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Trevor Campbell, who provided an uncountable amount of guidance and encouragement over the last two years so that I was able to keep making progress and recover from frustrations. His enthusiasm and passion for research has constantly influenced me and motivated me to pursue a Ph.D. I feel extremely lucky to be his student and to keep working with him in the future.

I would also like to thank Prof. Alexandre Bouchard-Côté for being my second reader and for his insightful suggestions concerning my thesis. Not only that, I have really appreciated my interactions with Alex. I have learned so much from his course and reading groups, which have brought me into the world of Bayesian computation. I am also grateful to the whole department. All the free sandwiches and interesting conversations that happened in lounge have become good memories. I have to say, this is the best department in the world!

Finally, I would love to thank my family and my friends. Their unconditional support and love are, and will always be, the reason for me to move forward.

Chapter 1

Introduction

Bayesian statistical models are powerful tools for learning from data, with the ability to encode complex hierarchical dependence and domain expertise, as well as coherently quantify uncertainty in latent parameters. Unfortunately, for many modern Bayesian models, exact computation of the posterior is intractable (Blei et al., 2017, Section 2.1) and statisticians must resort to approximate inference algorithms. Currently, the most popular type of Bayesian inference algorithm in statistics is Markov Chain Monte Carlo (MCMC) (Gelfand and Smith, 1990; Hastings, 1970; Robert and Casella, 2013), which provides approximate samples from the posterior distribution supported by a comprehensive literature of theoretical guarantees (Meyn and Tweedie, 2012; Roberts and Rosenthal, 2004). However, in the setting of large-scale data, traditional MCMC methods tend to be computationally intractable due to the need to compute the full data likelihood in each step, which is required to maintain the Bayesian posterior as the stationary distribution.

Variational inference (Blei et al., 2017; Jordan et al., 1998; Wainwright and Jordan, 2008) is a popular alternative to classical MCMC methods that approximates the intractable posterior with a distribution chosen from a pre-specified family, e.g., the family of Gaussian distributions parametrized by mean and covariance. The approximating distribution is chosen by minimizing a discrepancy—such as the Kullback-Leibler (KL) divergence (Murphy, 2012, Section 2.8) or Rényi α-divergence (Van Erven and Harremos, 2014)—to the posterior distribution over the family, thus converting Bayesian inference into an optimization problem. This

formulation enables the use of simple, efficient stochastic optimization algorithms (Bottou, 2004; Robbins and Monro, 1951) that require only a subsample of the data at each iteration, avoiding computation on the entire dataset.

But despite its computational tractability, variational inference has two key limitations that prevent its widespread adoption in the statistical community. First, one must select an appropriate parametric family of distributions from which to select the variational approximation. This choice of family presents a tradeoff: a simple family typically enables the design of fast local optimization algorithms, but limits the achievable fidelity of the posterior approximation. It is in general difficult to know how limited the family is before actually optimizing; and not only that, it is also often difficult to estimate the approximation error once the optimization is complete (Huggins et al., 2020). For example, if one chooses a *mean-field* variational family in which variables are assumed to be independent, the resulting posterior approximation will typically underestimate their true posterior variances and cannot capture their covariances (Murphy, 2012, Section 21.2.2), two quantities of particular interest to statisticians. A more flexible variational family may result in a lower achievable approximation error, but the error is still not known in advance and typically results in more expensive computation. The second key limitation is that even if the family could be chosen carefully to have favourable computational properties and a global optimum with low approximation error, the optimization problem itself is typically nonconvex and the global optimum cannot be found reliably.

The key to addressing the first limitation of variational inference is to understand the minimum approximation error within a particular variational family. This is quite difficult given finite data; Han and Yang (2019) provides a non-asymptotic analysis of the optimal mean-field variational approximation, but extending these results to more general distribution families is not straightforward. However, multiple threads of research have explored the statistical properties of variational inference in an asymptotic regime by taking advantage of the limiting behavior of the Bayesian posterior. Wang and Blei (2019) exploits the asymptotic normality of the posterior distribution in a parametric Bayesian setting to show that the KL minimizer among the variational family to the posterior converges to the KL minimizer to the limiting distribution of posterior under infinite samples—a normal distribution. Alquier and

Ridgway (2020) analyze the rate of convergence of the variational approximation to a fractional posterior—a posterior with a tempered likelihood—in a high dimensional setting where the posterior itself may not have the ideal asymptotic behavior. Zhang and Gao (2020) studies the contraction rate of the variational distribution for non-parametric Bayesian inference and provides general conditions on the Bayesian model that characterizes the rate. Yang et al. (2020) and Jaiswal et al. (2019) build a framework for analyzing the statistical properties of α-Rényi variational inference, and provide sufficient conditions that guarantee an optimal convergence rate of the point estimate obtained from variational inference. But while this literature has built a comprehensive understanding of the asymptotic statistical guarantees of optimal variational posterior approximations, the nonconvexity of the optimization problem makes these guarantees difficult to obtain reliably in practice. In fact, Proposition 3.3.3 of the present work demonstrates that the problem is nonconvex even in the simple case of Gaussian variational inference with ideal asymptotic posterior behaviour. Therefore, addressing the nonconvexity of variational inference is meaningful for both computational and theoretical reasons.

In this work, we address the nonconvexity of Gaussian variational inference in the data-asymptotic regime when the posterior distribution admits asymptotic normality. However, rather than focusing on the statistical properties of the optimal Gaussian variational proxy, we investigate and exploit the asymptotic properties of the optimization problem itself (Chapter 3), and use these to design a procedure (Chapter 4) which enables tractable Gaussian variational optimization and hence makes theoretical results regarding the optimal solution applicable. We develop *consistent stochastic variational inference* (CSVI), an efficient and simple algorithm for Gaussian variational inference that guarantees the probability of achieving the global optimum converges to 1 in the limit of observed data. The two key ingredients of CSVI are the choice of initialization for the Gaussian mean (Section 4.1) and the design of a scaled projected stochastic optimization algorithm (Section 4.2). We use the mode of a smoothed posterior—the posterior distribution convolved with Gaussian noise—as the mean initialization, which can be solved by a tractable optimization formulated in Section 4.1. We show that this procedure initializes Gaussian variational inference in a local region where the optimization becomes convex with increasing sample size, and where the global optimum eventually

lies. We then show that the novel scaled projected stochastic gradient algorithm is guaranteed to stay inside this local region and eventually converge to the global optimum. Experiments on synthetic examples in Chapter 5 show that CSVI provides numerically stable and asymptotically consistent posterior approximations.

Chapter 2

Gaussian Variational Inference

In the setting of Bayesian inference considered in this paper, we are given a sequence of posterior distributions Π_n, $n \in \mathbb{N}$ each with full support on $\mathbb{R}^d$. The index n represents the amount of observed data; denote Π_0 to be the prior. We consider the problem of approximating the posterior distribution via *Gaussian variational inference*, i.e.,

$$\underset{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}}{\arg\min} \quad \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\mu, \Sigma) || \Pi_n\right) \quad \text{s.t.} \quad \Sigma \succ 0,$$

where the Kullback-Leibler divergence (Murphy, 2012, Section 2.8) is defined as

$$\mathrm{D}_{\mathrm{KL}}\left(Q||P\right) := \int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}Q$$

for any pair of probability distributions P, Q such that $Q \ll P$, and $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon-Nikodym derivative of Q with respect to P (Folland, 1999, Section 3.2). We further assume that each posterior Π_n has density π_n with respect to the Lebesgue measure, and use the standard reparametrization of Σ using the Cholesky factorization $\Sigma = n^{-1}LL^T$ to arrive at the common formulation of Gaussian variational inference

(Kucukelbir et al., 2017) that is the focus of the present work:

$$\begin{aligned}\mu_n^\star, L_n^\star = \underset{\mu\in\mathbb{R}^d, L\in\mathbb{R}^{d\times d}}{\arg\min} \quad & -n^{-1}\log\det L - \mathbb{E}\Big[n^{-1}\log\pi_n(\mu + n^{-1/2}LZ)\Big] \\ \text{s.t.} \quad & L \text{ lower triangular with positive diagonal} \\ & Z \sim \mathcal{N}(0, I).\end{aligned} \tag{2.1}$$

Denote the optimal Gaussian distribution $\mathcal{N}_n^\star := \mathcal{N}(\mu_n^\star, n^{-1}L_n^\star L_n^{\star T})$. Intuitively, this optimization problem encodes a tradeoff between maximizing the expected posterior density under the variational approximation—which tries to make L small and move μ close to the maximum point of π_n—and maximizing the entropy of the variational approximation—which prevents L from becoming too small. It crucially does not depend on the (typically unknown) normalization of π_n, which appears as an additive constant in Eq. (2.1); it is common to drop this constant and instead equivalently maximize the *expectation lower bound (ELBO)* (Blei et al., 2017). Note that there are a number of unconstrained parametrizations of the covariance matrix variable Σ (Pinheiro and Bates, 1996). We select the (unique) positive-diagonal Cholesky factor L as it makes the optimization problem Eq. (2.1) more amenable to both theoretical analysis and computational optimization.

One typically attempts to solve Eq. (2.1) using an iterative local descent optimization algorithm. In cases where the expectation in the objective is analytically tractable—e.g. in exponential family models with a mean-field variational approximation—coordinate descent is the standard approach (Bishop, 2006; Blei et al., 2017). However, the expectation is intractable in most scenarios, and one must instead rely on stochastic gradient estimates (Hoffman et al., 2013; Kingma and Welling, 2014; Kucukelbir et al., 2017; Ranganath, 2014). In particular, assuming one can interchange expectation and differentiation (see Section 3.1 for details), the quantities

$$\begin{aligned}\hat{\nabla}_{\mu,n}(\mu, L, Z) &:= -n^{-1}\nabla\log\pi_n(\mu + LZ) \\ \hat{\nabla}_{L,n}(\mu, L, Z) &:= -n^{-1}(\operatorname{diag} L)^{-1} - n^{-3/2}\operatorname{tril}\nabla\log\pi_n(\mu + n^{-1/2}LZ)Z^T,\end{aligned} \tag{2.2}$$

are unbiased estimates of the μ- and L-gradients of the objective in Eq. (2.1) given $Z \sim \mathcal{N}(0, I)$, where the functions $\operatorname{diag} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ and $\operatorname{tril} : \mathbb{R}^{d\times d} \to$

$\mathbb{R}^{d\times d}$ set the off-diagonal and upper triangular elements of their arguments to 0, respectively. These unbiased gradient estimates may be used in a wide variety of stochastic optimization algorithms (Bottou, 2004; Robbins and Monro, 1951) applied to Eq. (2.1). In this paper, we will focus on projected stochastic gradient descent (SGD) (Bubeck, 2015, Section 3.) due to its simplicity; we expect that the mathematical theory in this work extends to other related methods.

In general, Gaussian variational inference is a nonconvex optimization problem and standard iterative methods such as SGD are not guaranteed to produce a sequence of iterates that converge to $\mu_n^\star, L_n^\star$. The goal of this work is to address this limitation by developing an iterative algorithm that uses only the black-box stochastic gradient estimates from Eq. (2.2) to reliably find the globally optimal solution of Eq. (2.1).

Chapter 3

Properties of the Optimization Problem

In this section, we investigate the properties of the Gaussian variational inference optimization problem Eq. (2.1). We take advantage of the theory of statistical asymptotics to show that as we obtain more data, the optimum solution of Eq. (2.1) converges to a fixed value and the objective function becomes locally strongly convex around that fixed value.

3.1 Statistical model and assumptions

As is common in past work (Ghosal et al., 2000; Kleijn and van der Vaart, 2012; Shen and Wasserman, 2001), we take a frequentist approach to analyzing Bayesian inference. We assume that the sequence of observations are independent and identically distributed $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$ from a distribution P_{θ_0} with parameter $\theta_0 \in \mathbb{R}^d$ selected from a parametric family $\{P_\theta : \theta \in \mathbb{R}^d\}$. We further assume that for each $\theta \in \mathbb{R}^d$, P_θ has common support, has density p_θ with respect to some common base measure, and that $p_\theta(x)$ is a Lebesgue measurable function of θ for all x. Finally, we assume the prior distribution Π_0 has full support on $\mathbb{R}^d$ with density π_0 with respect to the Lebesgue measure. Thus by Bayes' rule, the posterior distribution Π_n has density proportional to the prior density times the likelihood,

i.e.,

$$\pi_n(\theta) \propto \pi_0(\theta) \prod_{i=1}^{n} p_\theta(X_i).$$

In order to develop the theory in this work, we require a set of additional technical assumptions on π_0 and p_θ given by Assumption 1. These are a collection of regularity conditions that are standard for parametric models, which guarantee that the *maximum likelihood estimate (MLE)* $\theta_{\mathrm{MLE},n} := \arg\max_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \log p_\theta(X_i)$ is well-defined and asymptotically consistent for θ_0 (Lehmann and Casella, 2006, Chapter 6, Thm 3.7), and that the Bayesian posterior distribution of $\sqrt{n}(\theta - \theta_0)$ converges in total variation to a Gaussian distribution; this is known as the *Bernstein-von Mises theorem* (van der Vaart, 2000, p. 141).

Theorem 3.1.1 (Bernstein-von Mises & MLE consistency)**.** *Under Assumption 1,*

$$\theta_{MLE,n} \overset{P_{\theta_0}}{\to} \theta_0, \quad \textit{and} \quad \mathrm{D_{TV}}\left(\Pi_n\,,\, \mathcal{N}\left(n^{-1/2}\Delta_{n,\theta_0} + \theta_0, n^{-1}H_{\theta_0}^{-1}\right)\right) \xrightarrow{P_{\theta_0}} 0, \quad (3.1)$$

where $\Delta_{n,\theta_0} = n^{-1/2} \sum_{i=1}^{n} H_{\theta_0}^{-1} \nabla \log p_\theta(X_i)$.

Assumption 1. (Regularity Conditions)

1. $\left\{P_\theta : \theta \in \mathbb{R}^d\right\}$ is an identifiable family of distributions;
2. For all x, θ, the densities π_0, p_θ are positive and twice continuously differentiable in θ;
3. For all θ, $\mathbb{E}_\theta\left[\nabla \log p_\theta(X)\right] = 0$;
4. For all θ,

   $$H_\theta := -\mathbb{E}_\theta\left[\nabla^2 \log p_\theta(X)\right] = \mathbb{E}_\theta\left[\nabla \log p_\theta(X) \nabla \log p_\theta(X)^T\right],$$

   and $H_{\theta_0} \succeq \epsilon I$ for some $\epsilon > 0$. Further, for θ, θ' in a neighbourhood of θ_0,

   $$(\theta, \theta') \to \mathbb{E}_{\theta'}\left[-\nabla^2 \log p_\theta(X)\right]$$

   is continuous in spectral norm;

5. There exists a measurable function $g(x)$ such that for θ in a neighbourhood of θ_0 and for all x,

$$\max_{i,j\in[d]} \left| \left[\nabla^2 \log p_\theta(x)\right]_{i,j} \right| < g(x), \quad \mathbb{E}_{\theta_0}[g(X)] < \infty.$$

Note that the above conditions in Assumption 1 are stronger (van der Vaart, 2000, Lemmas 7.6 and 10.6) than the usual *local asymptotic normality* (van der Vaart, 2000, Section 7) and *testability* (van der Vaart, 2000, p. 141) conditions required for asymptotic posterior concentration and Gaussianity in the Bayesian asymptotics literature. Many of the results in this work would still hold with these weaker conditions, but we prefer Assumption 1 for the present work as these conditions are much simpler to state and check in practice.

We require two conditions beyond the basic regularity conditions in Assumption 1. First, we require that the *maximum a posteriori (MAP)* estimate $\theta_{\text{MAP},n} := \arg\max_{\theta\in\mathbb{R}^d} \log \pi_n(\theta)$ converges at a $\sqrt{n}$ rate to θ_0. This is not a particularly strong assumption—the MAP estimate typically has the same asymptotic properties as the MLE—but we require this to ensure that we can initialize the optimization algorithm for variational inference in a manner that ensures convergence to the global optimum. Readers who are interested in sufficient conditions for ensuring MAP consistency may consult Bassett and Deride (2019); Dashti et al. (2013); Grendár and Judge (2009); Stefanski and Boos (2002). Second, we require control on the smoothness of the log posterior density $\log \pi_n$ asymptotically. In this work, we impose a bound on the second derivative, but we conjecture that bounds on higher-order derivatives would also suffice; see Section 3.3 for details.

Assumption 2. (MAP $\sqrt{n}$-consistency) The maximum a posteriori point satisfies

$$\|\theta_{\text{MAP},n} - \theta_0\| = O_{P_{\theta_0}}(1/\sqrt{n}).$$

Assumption 3. (Asymptotic Smoothness) There exists an $\ell > 0$ such that

$$\mathbb{P}\left(\sup_\theta \left\|n^{-1}\nabla^2 \log \pi_n(\theta)\right\|_2 > \ell\right) \to 0,$$

where the $\|\cdot\|_2$ denotes the spectral norm of matrices.

3.2 Global optimum consistency

The first important property of Gaussian variational inference is that the optimum solution $\mu_n^\star, L_n^\star$ converges in probability to θ_0, L_0 under the same conditions required for the Bernstein-von Mises theorem (Theorem 3.1.1), where θ_0 is the true data-generating parameter and L_0 is the unique positive-diagonal Cholesky factor of the inverse Fisher information matrix $H_{\theta_0}^{-1} = L_0 L_0^T$. In other words, the *global* solution of Gaussian variational inference is a statistically consistent estimator; if we can develop an algorithm that solves Eq. (2.1) reliably, we therefore have an asymptotically consistent Bayesian inference procedure. Theorem 3.2.1 makes this statement precise; the proof follows directly from a result regarding the total variation consistency of the optimal variational distribution (Wang and Blei, 2019) and the continuity of the positive-diagonal Cholesky decomposition (Schatzman, 2002, p. 295).

Theorem 3.2.1. *Under Assumption 1,*

$$\forall \epsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(\|\mu_n^\star - \theta_0\| < \epsilon,\ \|L_n^\star - L_0\| < \epsilon\right) = 1.$$

Proof. We consider the KL cost for the scaled and shifted posterior distribution. Let $\tilde{\Pi}_n$ be the Bayesian posterior distribution of $\sqrt{n}(\theta - \theta_0)$. The KL divergence measures the difference between the distributions of two random variables and is invariant when an invertible transformation is applied to both random variables (Qiao and Minematsu, 2010, Theorem 1). Note that $\tilde{\Pi}_n$ is shifted and scaled from Π_n, and that this linear transformation is invertible, so

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\mu, \Sigma)||\Pi_n\right) = \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}\left(\sqrt{n}(\mu - \theta_0), n\Sigma\right)||\tilde{\Pi}_n\right).$$

Let $\tilde{\mu}_n^\star, \tilde{\Sigma}_n^\star$ be the parameters of the optimal Gaussian variational approximation to $\tilde{\Pi}_n$, i.e.,

$$\tilde{\mu}_n^\star, \tilde{\Sigma}_n^\star = \underset{\mu\in\mathbb{R}^d, \Sigma\in\mathbb{R}^{d\times d}}{\arg\min}\ \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}(\mu, \Sigma)||\tilde{\Pi}_n\right) \quad \text{s.t.} \quad \Sigma \succ 0,$$

and let

$$\tilde{\mathcal{N}_n}^\star := \mathcal{N}\left(\tilde{\mu}_n^\star, \tilde{\Sigma}_n^\star\right) = \mathcal{N}\left(\sqrt{n}(\mu_n^\star - \theta_0), L_n^\star L_n^{\star T}\right).$$

Wang and Blei (2019, Corollary 7) shows that under Assumption 1,

$$\mathrm{D}_{\mathrm{TV}}\left(\tilde{\mathcal{N}_n}^\star, \mathcal{N}\left(\Delta_{n,\theta_0}, H_{\theta_0}^{-1}\right)\right) \overset{P_{\theta_0}}{\to} 0.$$

Convergence in total variation implies weak convergence, which then implies pointwise convergence of the characteristic function. Denote $\tilde{\phi}_n^\star(t)$ and $\phi_n(t)$ to be the characteristic functions of $\tilde{\mathcal{N}_n}^\star$ and $\mathcal{N}\left(\Delta_{n,\theta_0}, H_{\theta_0}^{-1}\right)$. Therefore

$$\forall t \in \mathbb{R}^d,\ \frac{\phi_n^\star(t)}{\phi_n(t)} = \exp\left(i(\sqrt{n}(\mu_n^\star - \theta_0) - \Delta_{n,\theta_0})^T t - \frac{1}{2} t^T \left(L_n^\star L_n^{\star T} - H_{\theta_0}^{-1}\right) t\right) \overset{P_{\theta_0}}{\longrightarrow} 1,$$

which implies

$$\mu_n^\star \overset{P_{\theta_0}}{\to} \frac{1}{\sqrt{n}}\Delta_{n,\theta_0} + \theta_0, \quad \text{and} \quad L_n^\star L_n^{\star T} \overset{P_{\theta_0}}{\to} H_0^{-1} = L_0 L_0^T.$$

Under Assumption 1, van der Vaart (2000, Theorem 8.14) states that

$$\|\Delta_{n,\theta_0} - \sqrt{n}(\theta_{\mathrm{MLE},n} - \theta_0)\| \overset{P_{\theta_0}}{\to} 0,$$

and $\theta_{\mathrm{MLE},n} \overset{P_{\theta_0}}{\to} \theta_0$ according to Eq. (3.1), yielding $\mu_n^\star \overset{P_{\theta_0}}{\to} \theta_0$.

Finally since the Cholesky decomposition defines a continuous mapping from the set of positive definite Hermitian matrices to the set of lower triangular matrices with positive diagonals (both sets are equipped with the spectral norm) (Schatzman, 2002, p. 295), we have

$$L_n^\star \overset{P_{\theta_0}}{\to} L_0.$$

□

3.3 Convexity and smoothness

The statistical consistency of the optimal parameters $\mu_n^\star, L_n^\star$ alone does not provide a complete analysis of the asymptotics of Gaussian variational inference; indeed, it is in general not tractable to actually compute or approximate the solution $\mu_n^\star, L_n^\star$, which diminishes the utility of Theorem 3.2.1 in practice. In order to make use of the consistency result, we require that solving the Gaussian variational inference problem is tractable in some sense. In this section, we investigate the tractability of Gaussian variational inference as formulated in Eq. (2.1). Since we have access only to (stochastic estimates of) the gradient of the objective function in Eq. (2.1), and projected stochastic gradient descent is known to solve optimization problems with *strongly convex* and *Lipschitz smooth* objectives (Bottou, 2004; Rakhlin et al., 2012), this amounts to investigating the convexity and smoothness of the objective function.[1]

We will begin by focusing on the expectation term in the objective of Eq. (2.1),

$$f_n : \mathbb{R}^d \to \mathbb{R}, \quad f_n(x) := -n^{-1} \log \pi_n(x) \tag{3.2}$$
$$F_n : \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}, \quad F_n(\mu, L) := \mathbb{E}\left[f_n(\mu + n^{-1/2} L Z)\right], \quad Z \sim \mathcal{N}(0, I).$$

The first main result of this section is that convexity and smoothness of the log posterior density f_n implies the same for $F_n(\mu, L)$. We begin with a generalization of the typical definitions of strong convexity and Lipschitz smoothness found in the literature (Boyd and Vandenberghe, 2004) in Definition 3.3.1, and then provide the precise theoretical statement in Theorem 3.3.2.

Definition 3.3.1 (Convexity and Smoothness)**.** Let $g : \mathcal{X} \to \mathbb{R}$ be a twice differentiable function on a convex set $\mathcal{X} \subseteq \mathbb{R}^d$, and let $D : \mathcal{X} \to \mathbb{R}^{d\times d}$ be a positive definite matrix depending on x. Then g is *D-strongly convex* if

$$\forall x \in \mathcal{X}, \quad \nabla^2 g(x) \succeq D(x),$$

[1] There are many other properties one might require of a tractable optimization problem, e.g., pseudoconvexity (Crouzeix and Ferland, 1982), quasiconvexity (Arrow and Enthoven, 1961), or invexity (Ben-Israel and Mond, 1986). We focus on convexity as it does not impose overly stringent assumptions on our theory and has stronger implications than each of the aforementioned conditions.

and g is *D-Lipschitz smooth* if

$$\forall x \in \mathcal{X}, \quad -D(x) \preceq \nabla^2 g(x) \preceq D(x).$$

Theorem 3.3.2. *Suppose f_n is D-strongly convex (-Lipschitz smooth) for positive definite matrix $D \in \mathbb{R}^{d\times d}$. Then F_n reinterpreted as a function from $\mathbb{R}^{(d+1)d} \to \mathbb{R}$—by stacking μ and each column of L into a single vector—is D'-strongly convex (-Lipschitz smooth), where*

$$D' = \operatorname{blockd}(D, n^{-1}D, \ldots, n^{-1}D) \in \mathbb{R}^{(d+1)d\times(d+1)d},$$

and blockd *creates a block-diagonal matrix out of its arguments.*

Proof. We provide a proof of the result for strong convexity; the result for Lipschitz smoothness follows the exact same proof technique. Note that if D' does not depend on x, $F_n(x)$ is D'-strongly convex if and only if $F_n(x) - \frac{1}{2}x^T D' x$ is convex. We use this equivalent characterization of strong convexity in this proof.

Note that for $Z \sim \mathcal{N}(0, I)$,

$$\mathbb{E}\left[\frac{1}{2}(\mu + n^{-1/2}LZ)^T D(\mu + n^{-1/2}LZ)\right] = \frac{1}{2}\mu^T D\mu + \frac{1}{2}\operatorname{tr} L^T(n^{-1}D)L.$$

Define $\lambda \in [0, 1]$, vectors $\mu, \mu' \in \mathbb{R}^d$, positive-diagonal lower triangular matrices $L, L' \in \mathbb{R}^{d\times d}$, and vectors $x, x' \in \mathbb{R}^{(d+1)d}$ by stacking μ and the columns of L and likewise μ' and the columns of L'. Define $x(\lambda) = \lambda x + (1-\lambda)x'$, $\mu(\lambda) = \lambda\mu + (1-\lambda)\mu'$, and $L(\lambda) = \lambda L + (1-\lambda)L'$. Then

$$\begin{aligned}
&F_n(x(\lambda)) - \frac{1}{2}x(\lambda)^T \operatorname{diag}(D, n^{-1}D, \ldots, n^{-1}D)x(\lambda) \\
=&F_n(\mu(\lambda), L(\lambda)) - \left(\frac{1}{2}\mu(\lambda)^T D\mu(\lambda) + \frac{1}{2}\operatorname{tr} L(\lambda)^T(n^{-1}D)L(\lambda)\right) \\
=&\mathbb{E}\left[n^{-1}\log\pi_n(\mu(\lambda) + n^{-1/2}L(\lambda)Z) - \frac{1}{2}(\mu(\lambda) + n^{-1/2}L(\lambda)Z)^T D(\mu(\lambda)\right. \\
&\left.+n^{-1/2}L(\lambda)Z)\right].
\end{aligned}$$

By the D-strong convexity of $n^{-1}\log\pi_n$,

$$\begin{aligned}
\leq & \lambda\left(F_n(\mu,L)-\frac{1}{2}\mu^T D\mu-\frac{1}{2}\operatorname{tr}L^T(n^{-1}D)L\right)\\
& +(1-\lambda)\left(F_n(\mu',L')-\frac{1}{2}\mu'^T D\mu'-\frac{1}{2}\operatorname{tr}L'^T(n^{-1}D)L'\right)\\
= & \lambda\left(F_n(x)-\frac{1}{2}x^T\operatorname{diag}(D,n^{-1}D,\ldots,n^{-1}D)x\right)\\
& +(1-\lambda)\left(F_n(x')-\frac{1}{2}x'^T\operatorname{diag}(D,n^{-1}D,\ldots,n^{-1}D)x'\right).
\end{aligned}$$

□

For example, if the posterior distribution Π_n is a multivariate Gaussian distribution $\mathcal{N}(\mu_n, n^{-1}\Sigma_n)$ with mean μ_n and covariance $n^{-1}\Sigma_n$, then the expectation component of the Gaussian variational inference objective function becomes

$$F_n(\mu,L)=n^{-1}\operatorname{tr}\Sigma_n^{-1}LL^T+(\mu-\mu_n)^T\Sigma_n^{-1}(\mu-\mu_n),$$

which is a jointly convex quadratic function in μ, L with Hessian matrix (for μ and columns of L stacked together in a single vector) equal to

$$\operatorname{blockd}(\Sigma_n^{-1},n^{-1}\Sigma_n^{-1},\ldots,n^{-1}\Sigma_n^{-1})\in\mathbb{R}^{(d+1)d\times(d+1)d}.$$

Combined with the convexity of the log determinant term $-n^{-1}\log\det L$ (Boyd and Vandenberghe, 2004, p.73), Gaussian variational inference for strongly convex and Lipschitz smooth log posterior density $-n^{-1}\log\pi_n$ is itself strongly convex and Lipschitz smooth in any compact set contained in the optimization domain.

However, in a typical statistical model, the posterior is typically neither Gaussian nor strongly convex. But when the Bernstein-von Mises theorem holds (van der Vaart, 2000), the posterior distribution (scaled and shifted appropriately) converges asymptotically to a Gaussian distribution. Thus, it may be tempting to think that the Bernstein von-Mises theorem implies that Gaussian variational inference should eventually become a convex optimization problem. This is unfortunately not true, essentially because Bernstein-von Mises only implies convergence to a Gaussian in total variation distance, but not necessarily in the log density function or its

gradients. The second main result in this section—Proposition 3.3.3—is a simple demonstration of the fact that the Bernstein-von Mises theorem is not sufficient to guarantee the convexity of Gaussian variational inference.

Proposition 3.3.3. *Suppose $d = 1$, f_n is differentiable to the third order for all n, that there exists an open interval $U \subseteq \mathbb{R}$ and $\epsilon > 0$ such that*

$$\sup_{\theta \in U} \frac{d^2 f_n}{d\theta^2} \leq -\epsilon,$$

and that there exists $\eta > 0$ such that

$$\sup_{\theta \in \mathbb{R}} \left| \frac{d^3 f_n}{d\theta^3} \right| \leq \eta.$$

Then there exists a $\delta > 0$ such that

$$\sup_{\sigma < \delta,\, \mu \in U} \frac{d^2}{d\mu^2} \mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}(\mu, \sigma^2) || \Pi_n \right) < 0.$$

Proof. Note that by reparameterization,

$$\arg\min_{\mu} \mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}(\mu, \sigma^2) || \Pi_n \right) = \arg\min_{\mu} \mathbb{E} \left[-n^{-1} \log \pi_n(\mu + \sigma Z) \right],$$

where $Z \sim \mathcal{N}(0, 1)$. Using a Taylor expansion,

$$\begin{aligned} &- \mathbb{E} \left[\frac{d^2}{d\mu^2} \left(-n^{-1} \log \pi_n(\mu + \sigma Z) \right) \right] \\ &= \mathbb{E} \left[-n^{-1} \log \pi_n^{(2)}(\mu) - n^{-1} \log \pi_n^{(3)}(\mu') \cdot \sigma Z \right], \end{aligned}$$

for some μ' between μ and $\mu + \sigma Z$. By the uniform bound on the third derivative and local bound on the second derivative, for any $\mu \in U$,

$$\begin{aligned} \mathbb{E} \left[-n^{-1} \log \pi_n^{(2)}(\mu) - n^{-1} \log \pi_n^{(3)}(\mu') \cdot \sigma Z \right] &\leq -\epsilon + \eta\sigma \mathbb{E} \left| Z \right| \\ &\leq -\epsilon + \eta\sigma. \end{aligned}$$

The result follows for any $0 < \delta < \epsilon/\eta$. □

Although Proposition 3.3.3 is a negative result about the global convexity of F_n, it does hint at a very useful fact: the *local* convexity (Definition 3.3.4) of F_n matches that of f_n, assuming that we control the global behaviour of f_n, e.g., through a uniform bound on the k^{th} derivative. This is essentially due to the fact that the two functions differ only by a smoothing under a standard multivariate Gaussian variable, which has finite moments of all orders. The third main result of this section, Theorem 3.3.5—which we exploit later in Chapter 4 to develop a reliable variational inference algorithm—makes the general link between the local convexity behaviour of f_n and F_n precise under global Lipschitz smoothness, i.e., a uniform bound on the Hessian.

Definition 3.3.4 (Local Strong Convexity)**.** In the same setting of Definition 3.3.1, if there exists a convex subset $\mathcal{Y} \subset \mathcal{X}$ such that the restriction of g to $\mathcal{Y}$ is ϵD-strongly convex, then g is locally ϵD-strongly convex.

Theorem 3.3.5. *Suppose there exist $\epsilon, \ell, r > 0$ and $x \in \mathbb{R}^d$ such that f_n is globally ℓI-Lipschitz smooth and locally ϵI-strongly convex in the set $\{y : \|y - x\| \leq r\}$. Define*

$$\begin{aligned} D_n &:= \text{blockd}\left(I, n^{-1}I, \ldots, n^{-1}I\right) \in \mathbb{R}^{(d+1)d \times (d+1)d} \\ \tau_n(\mu, L) &:= 1 - \chi^2_{d+2}\left(n\frac{(r^2 - 2\|\mu - x\|^2)}{2\|L\|_F^2}\right), \end{aligned} \tag{3.3}$$

where χ^2_k is the CDF of a chi-square random variable with k degrees of freedom. Then F_n reinterpreted as a function of $\mathbb{R}^{(d+1)d} \to \mathbb{R}$—by stacking μ and each column of L into a single vector—is ℓD_n-Lipschitz smooth; and is $(\epsilon - \tau_n(\mu, L) \cdot (\epsilon + \ell))D_n$-strongly convex when $\|\mu - x\|^2 \leq \frac{r^2}{2}$.

Proof. Note that we can split L into columns and express LZ as

$$LZ = \sum_{i=1}^{d} L_i Z_i,$$

where $L_i \in \mathbb{R}^p$ is the i^{th} column of L, and $(Z_i)_{i=1}^d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Denoting $\nabla^2 f_n :=$

$\nabla^2 f_n(\mu + LZ)$ for brevity, the 2^{nd} derivatives in both μ and L are

$$\begin{aligned}
\nabla^2_{\mu\mu} F_n &= \mathbb{E}\left[\nabla^2 f_n\right] \\
\nabla^2_{L_i L_j} F_n &= n^{-1}\mathbb{E}\left[Z_i Z_j \nabla^2 f_n\right] \\
\nabla^2_{\mu L_i} F_n &= n^{-1/2}\mathbb{E}\left[Z_i \nabla^2 f_n\right]
\end{aligned}$$

where we can pass the gradient and Hessian through the expectation by dominated convergence because Z has a normal distribution and f_n has ℓ-Lipschitz gradients. Stacking these together in block matrices yields the overall Hessian,

$$\begin{aligned}
A &= \left[\begin{array}{cccc} I & n^{-1/2} Z_1 I & \dots & n^{-1/2} Z_d I \end{array}\right] \in \mathbb{R}^{d \times d(d+1)} \\
\nabla^2 F_n &= \mathbb{E}\left[A^T \nabla^2 f_n A\right] \in \mathbb{R}^{d(d+1) \times d(d+1)}.
\end{aligned}$$

Since f_n has ℓ-Lipschitz gradients, for all $x \in \mathbb{R}^d$, $-\ell I \preceq \nabla^2 f_n(x) \preceq \ell I$. Applying the upper bound and evaluating the expectation yields the Hessian upper bound (and the same technique yields the corresponding lower bound):

$$\begin{aligned}
\nabla^2 F_n &= \mathbb{E}\left[A^T \nabla^2 f_n A\right] \\
&\preceq \ell \mathbb{E}\left[A^T A\right] \\
&= \ell \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & n^{-1} I & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & n^{-1} I \end{bmatrix} = \ell D_n.
\end{aligned}$$

To demonstrate local strong convexity, we split the expectation into two parts: one where $n^{-1/2} LZ$ is small enough to guarantee that $\|\mu + n^{-1/2} LZ - x\|^2 \leq r^2$, and the complement. Define

$$r_n^2(\mu, L) := n \frac{(r^2 - 2\|\mu - x\|_2^2)}{2\|L\|_F^2}.$$

Note that when $\|Z\|^2 \le r_n^2(\mu, L)$,

$$\begin{aligned}\left\|\mu + \frac{1}{\sqrt{n}} LZ - x\right\|_2^2 &\le 2\|\mu - x\|^2 + 2n^{-1}\|LZ\|^2 \\ &\le 2\|\mu - x\|^2 + 2n^{-1}\|L\|_F^2\|Z\|^2 \\ &\le r^2.\end{aligned}$$

Then we may write

$$\begin{aligned}\nabla^2 F_n = \mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 \le r_n^2(\mu, L)\right] A^T \nabla^2 f_n A\right] \\ + \mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 > r_n^2(\mu, L)\right] A^T \nabla^2 f_n A\right].\end{aligned}$$

Since f_n has ℓ-Lipschitz gradients and is locally ϵ-strongly convex,

$$\nabla^2 F_n \succeq \epsilon \cdot \mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 \le r_n^2(\mu, L)\right] A^T A\right] - \ell \cdot \mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 > r_n^2(\mu, L)\right] A^T A\right].$$

Note that $A^T A$ has entries 1 and $n^{-1}Z_i^2$ along the diagonal, as well as $n^{-1}Z_i Z_j$, $i \ne j$ and $n^{-1/2}Z_i$ on the off-diagonals. By symmetry, since Z is an isotropic Gaussian, censoring by $\mathbb{1}\left[\|Z\|^2 \le \dots\right]$ or $\mathbb{1}\left[\|Z\|^2 > \dots\right]$ maintains that the off-diagonal expectations are 0. Therefore $\mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 \le r_n^2(\mu, L)\right] A^T A\right]$ is diagonal with coefficients $1 - \alpha_n(\mu, L)$ and $n^{-1}\beta_n(\mu, L)$, and $\mathbb{E}\left[\mathbb{1}\left[\|Z\|^2 > r_n^2(\mu, L)\right] A^T A\right]$ is diagonal with coefficients $\alpha_n(\mu, L)$ and $n^{-1}\tau_n(\mu, L)$ where

$$\begin{aligned}\alpha_n(\mu, L) &= \mathbb{P}\left(\|Z\|^2 > r_n^2(\mu, L)\right) \\ \beta_n(\mu, L) &= \mathbb{E}\left[Z_1^2 \mathbb{1}\left[\|Z\|^2 \le r_n^2(\mu, L)\right]\right] = d^{-1}\mathbb{E}\left[\|Z\|_2^2 \mathbb{1}\left[\|Z\|^2 \le r_n^2(\mu, L)\right]\right] \\ \tau_n(\mu, L) &= \mathbb{E}\left[Z_1^2 \mathbb{1}\left[\|Z\|^2 > r_n^2(\mu, L)\right]\right] = d^{-1}\mathbb{E}\left[\|Z\|_2^2 \mathbb{1}\left[\|Z\|^2 > r_n^2(\mu, L)\right]\right].\end{aligned}$$

Note that $\|Z\|^2 \sim \chi_d^2$; so $\alpha_n(\mu, L) = 1 - \chi_d^2(r_n^2(\mu, L))$ and

$$\begin{aligned}\tau_n(\mu, L) &= \int_{r_n^2(\mu, L)}^{\infty} \mathbb{1}\left[x \ge 0\right] \frac{1}{2^{(d+2)/2}\Gamma((d+2)/2)} x^{\frac{d+2}{2}-1} e^{-x/2} \mathrm{d}x \\ &= 1 - \chi_{d+2}^2(r_n^2(\mu, L)) \\ \beta_n(\mu, L) &= 1 - \tau_n(\mu, L).\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\nabla^2 F_n \\
&\succeq \operatorname{diag}\big((\epsilon(1-\alpha_n(\mu,L))-\ell\alpha_n(\mu,L))I, (\epsilon n^{-1}(1-\tau_n(\mu,L))-\ell n^{-1}\tau_n(\mu,L))I, \\
&\qquad\qquad \ldots, (\epsilon n^{-1}(1-\tau_n(\mu,L)) - \ell n^{-1}\tau_n(\mu,L))I\big) \\
&= \epsilon D_n - (\epsilon+\ell)\operatorname{diag}\left(\alpha_n(\mu,L)I, n^{-1}\tau_n(\mu,L)I, \ldots, n^{-1}\tau_n(\mu,L)I\right) \\
&\succeq \epsilon D_n - (\epsilon+\ell)\operatorname{diag}\left(\tau_n(\mu,L)I, n^{-1}\tau_n(\mu,L)I, \ldots, n^{-1}\tau_n(\mu,L)I\right) \\
&= D_n\left(\epsilon - \tau_n(\mu,L)\cdot(\epsilon+\ell)\right).
\end{aligned}
$$

□

The most complicated part of the statement of Theorem 3.3.5 is the function $\tau_n(\mu, L)$, which characterizes how much the tails of f_n can influence the local strong convexity of F_n around the point x. In particular, as long as μ is close to x and $\|L\|_F$ (which modulates the effect of noise) is sufficiently small, then the argument of the χ^2_{d+2} CDF is large, so τ_n is small, so $(\epsilon - \tau_n(\mu, L) \cdot (\epsilon + \ell)) \approx \epsilon$; thus we recover local strong convexity of the same magnitude as f_n. A further note is that although Theorem 3.3.5 requires a uniform bound on the Hessian of f_n, we conjecture that a similar result would hold under the assumption of a uniform bound on the k^{th} derivative. For simplicity of the result and ease of use later on in Chapter 4, we opted for the second derivative bound.

The final step of this section is to examine when the assumptions of Theorem 3.3.5 hold. This is where statistical asymptotics provide their benefit in variational optimization: although the function f_n need not be locally strongly convex for any particular n, the probability (under the data-generating distribution) that it becomes locally strongly convex around θ_0 converges to 1 under weak conditions on the statistical model as $n \to \infty$. Lemma 3.3.6 states the result precisely, and Example 3.3.7 provides an example of a sequence of functions that individually have no guarantees regarding their convexity, but which are asymptotically locally strongly convex.

Lemma 3.3.6. *Under Assumption 1, there exist $r, \delta > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(f_n \text{ is } \delta I\text{-strongly convex in the set } B_r(\theta_0)\right) = 1,$$

where $B_r(\theta_0) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \le r\}$.

Proof. Given Assumption 1, we know f_n is twice continuously differentiable. Thus, using the second order characterization of strong convexity, it is equivalent to show the existence of $r, \delta > 0$ such that

$$\mathbb{P}\left(\forall \theta \in B_r(\theta_0), \quad \nabla^2 f_n(\theta) \succeq \delta I\right) \to 1,$$

as $n \to \infty$. Note that by Weyl's inequality

$$\begin{aligned} \nabla^2 f_n(\theta) =& \nabla^2 f_n(\theta) - H_\theta + H_\theta \\ \succeq& \lambda_{\min}\left(\nabla^2 f_n(\theta) - H_\theta\right) I + \lambda_{\min}(H_\theta) I. \end{aligned} \tag{3.4}$$

Condition 4 of Assumption 1 guarantees that $H_{\theta_0} \succeq \epsilon I$ and that there exists a $\kappa > 0$ such that H_θ is continuous in $B_\kappa(\theta_0)$. Hence there exists $0 < \kappa' \le \kappa$, such that $\forall \theta \in B_{\kappa'}(\theta_0),\ H_\theta \succeq \frac{\epsilon}{2} I$.

We then consider $\lambda_{\min}\left(\nabla^2 f_n(\theta) - H_\theta\right)$. We aim to find a $0 < r \le \kappa'$ such that $|\lambda_{\min}\left(\nabla^2 f_n(\theta) - H_\theta\right)|$ is sufficiently small. Note that for any fixed $r > 0$,

$$\begin{aligned} &\sup_{\theta \in B_r(\theta_0)} \left|\lambda_{\min}\left(\nabla^2 f_n(\theta) - H_\theta\right)\right| \\ &\le \sup_{\theta \in B_r(\theta_0)} \left\|\nabla^2 f_n(\theta) - H_\theta\right\|_2 \\ &= \sup_{\theta \in B_r(\theta_0)} \left\|\nabla^2 f_n(\theta) - \mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] + \mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] - H_\theta\right\|_2 \\ &\le \sup_{\theta \in B_r(\theta_0)} \left(\left\|\nabla^2 f_n(\theta) - \mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right]\right\|_2 + \left\|\mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] - H_\theta\right\|_2\right). \end{aligned}$$

Now we split f_n into prior and likelihood, yielding that

$$\leq \sup_{\theta \in B_r(\theta_0)} \left\| -n^{-1} \sum_{i=1}^{n} \nabla^2 \log p_\theta(X_i) - \mathbb{E}_{\theta_0} \left[-\nabla^2 \log p_\theta(X) \right] \right\|_2 \tag{3.5}$$
$$+ \sup_{\theta \in B_r(\theta_0)} \| - n^{-1} \nabla^2 \log \pi_0(\theta) \|_2 + \sup_{\theta \in B_r(\theta_0)} \left\| \mathbb{E}_{\theta_0} \left[-\nabla^2 \log p_\theta(X) \right] - H_\theta \right\|_2 .$$

Given Condition 2 of Assumption 1, for all θ, $\pi_0(\theta)$ is positive and $\nabla^2 \pi_0(\theta)$ is continuous; and further due to the compactness of $B_r(\theta_0)$, we have that

$$\forall r > 0, \quad \sup_{\theta \in B_r(\theta_0)} \| - n^{-1} \nabla^2 \log \pi_0(\theta) \|_2 \to 0, \quad \text{as } n \to \infty. \tag{3.6}$$

Then, it remains to bound the first term and the last term of Eq. (3.5). For the first term, we aim to use the uniform weak law of large numbers to show its convergence to 0. By Condition 5 of Assumption 1, there exists a $0 < r_1 \leq \kappa'$ and a measurable function g such that for all $\theta \in B_{r_1}(\theta_0)$ and for all x,

$$\max_{i,j \in [d]} \left| \left(\nabla^2 \log p_\theta(x) \right)_{i,j} \right| < g(x), \quad \mathbb{E}_{\theta_0}[g(X)] < \infty.$$

Then, by the compactness of $B_{r_1}(\theta_0)$, we can apply the uniform weak law of large numbers (Jennrich, 1969, Theorem 2), yielding that for all $i, j \in [d]$,

$$\sup_{\theta \in B_{r_1}(\theta_0)} \left| \left(-n^{-1} \sum_{i=1}^{n} \nabla^2 \log p_\theta(X_i) \right)_{i,j} - \left(\mathbb{E}_{\theta_0} \left[-\nabla^2 \log p_\theta(X) \right] \right)_{i,j} \right| \overset{P_{\theta_0}}{\to} 0.$$

Since the entrywise convergence of matrices implies the convergence in spectral norm,

$$\sup_{\theta \in B_{r_1}(\theta_0)} \left\| -n^{-1} \sum_{i=1}^{n} \nabla^2 \log p_\theta(X_i) - \mathbb{E}_{\theta_0} \left[-\nabla^2 \log p_\theta(X) \right] \right\|_2 \overset{P_{\theta_0}}{\to} 0. \tag{3.7}$$

For the last term of Eq. (3.5), by Condition 4 of Assumption 1,

$$\begin{aligned}
&\lim_{r\to 0} \sup_{\theta\in B_r(\theta_0)} \left\|\mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] - H_\theta\right\|_2 \\
&= \lim_{r\to 0} \sup_{\theta\in B_r(\theta_0)} \left\|\mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] - \mathbb{E}_{\theta}\left[-\nabla^2 \log p_\theta(X)\right]\right\|_2 \\
&\to 0.
\end{aligned}$$

Thus, there exists a sufficiently small $r_2 > 0$ such that

$$\sup_{\theta\in B_{r_2}(\theta_0)} \left\|\mathbb{E}_{\theta_0}\left[-\nabla^2 \log p_\theta(X)\right] - H_\theta\right\|_2 \le \frac{\epsilon}{8}. \tag{3.8}$$

Then, we combine Eqs. (3.6) to (3.8) and pick $r = \min(r_1, r_2) \le \kappa'$, yielding that

$$\mathbb{P}\left(\sup_{\theta\in B_r(\theta_0)} \left|\lambda_{\min}\left(\nabla^2 f_n(\theta) - H_\theta\right)\right| \le \frac{\epsilon}{4}\right) \to 1, \tag{3.9}$$

as $n \to \infty$.

Then the proof is complete. Note that we have already shown for all $\theta \in B_{\kappa'}(\theta_0), H_\theta \succeq \frac{\epsilon}{2} I$. By Eqs. (3.9) and (3.4), we conclude that for all $\delta \le \frac{\epsilon}{4}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\forall\theta \in B_r(\theta_0), \quad \nabla^2 f_n(\theta) \succeq \delta I\right) = 1.$$

□

Example 3.3.7. Let $f_n(y) = y^2 + \left(\frac{1}{n}\sum_{i=1}^n X_i\right)\cos 5y$, where $X_i \sim \mathcal{N}(0,1)$. Then

$$\left|\frac{\mathrm{d}^2 f_n}{\mathrm{d}y^2} - 2\right| = 25\left|\cos(5y)\right| \cdot \left|n^{-1}\sum_{i=1}^n X_i\right|.$$

Therefore by the law of large numbers and the fact that $|\cos(5y)| \le 1$, for any $\epsilon > 0$, the sequence $(f_n)_{n\in\mathbb{N}}$ is asymptotically $(2-\epsilon)$-strongly convex and $(2+\epsilon)$-Lipschitz smooth. Fig. 3.1 visualizes the asymptotic behaviour of f_n as n increases.

The last result of this section—Corollary 3.3.8—combines Theorems 3.3.5 and 3.2.1 and Lemma 3.3.6 to provide the key asymptotic convexity/smoothness

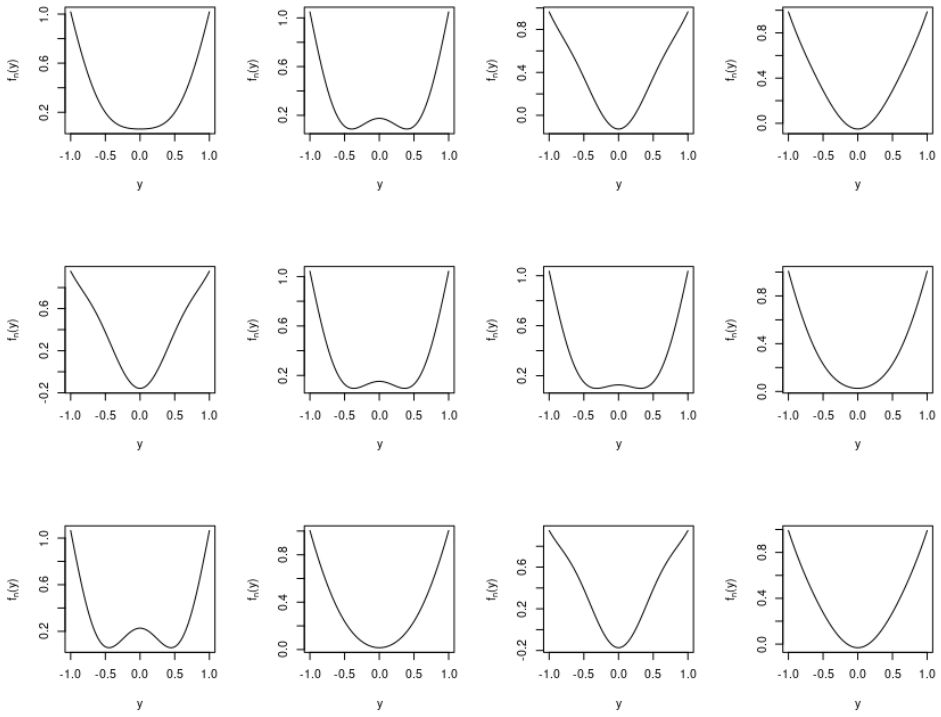

Figure 3.1: Plots of the function $f_n(y)$ from Example 3.3.7. Each row of figures represents a single realization of the sequence $(f_n)_{n\in\mathbb{N}}$ for increasing sample sizes 5, 20, 100, and 1000. Each column includes three repetitions of f_n under a single n. As n increases, the function $f_n(y)$ is more likely to be strongly convex and Lipschitz smooth with constants approaching 2.

result that we use in the development of the optimization algorithm in Chapter 4.

Corollary 3.3.8. *Suppose Assumptions 1 and 3 hold, and define D_n as in Theorem 3.3.5. Then there exist $\epsilon, \ell, r > 0$ such that F_n reinterpreted as a function of $\mathbb{R}^{(d+1)d} \to \mathbb{R}$—by stacking μ and each column of L into a single vector—satisfies*

$$\mathbb{P}\left(F_n \text{ is } \frac{\epsilon}{2}D_n\text{-strongly convex in } \mathcal{B}_{r,n} \text{ and globally } \ell D_n\text{-Lipschitz smooth}\right) \to 1,$$

as $n \to \infty$, where

$$\mathcal{B}_{r,n} = \left\{ \mu \in \mathbb{R}^d, L \in \mathbb{R}^{d\times d} \colon \|\mu - \mu_n^\star\|^2 \le \frac{r^2}{4} \text{ and } \|L - L_n^\star\|_F^2 \le 4\|I - L_n^\star\|_F^2 \right\}.$$

Proof. We begin by verifying the conditions of Theorem 3.3.5 for f_n. By Assumption 1 we know that f_n is twice differentiable. We also know that by Lemma 3.3.6, under Assumptions 1 and 3, there exist $\ell, r', \epsilon > 0$ such that

$$\mathbb{P}\left(\sup_\theta \left\| -n^{-1}\nabla^2 \log \pi_n(\theta) \right\|_2 > \ell \right) \to 0$$
$$\mathbb{P}\left(\inf_{\|\theta - \theta_0\| < r'} \lambda_{\min}\left(-n^{-1}\nabla^2 \log \pi_n(\theta) \right) < \epsilon \right) \to 0.$$

By Theorem 3.2.1 we know that $\mu_n^\star \overset{P_{\theta_0}}{\to} \theta_0$, so there exists an $r' > r > 0$ such that

$$\mathbb{P}\left(\inf_{\|\theta - \mu_n^\star\| < r} \lambda_{\min}\left(-n^{-1}\nabla^2 \log \pi_n(\theta) \right) < \epsilon \right) \to 0.$$

Therefore by Theorem 3.3.5, the probability that

$$\forall \mu, L, \quad -\ell D_n \preceq n^{-1}\nabla^2 \mathbb{E}\left[-\log \pi_n(\mu + 1/\sqrt{n} LZ) \right] \preceq \ell D_n, \tag{3.10}$$

and

$$\begin{aligned} &\text{for all } L \text{ and for } \|\mu - \mu_n^\star\|^2 < r^2/2, \\ &n^{-1}\nabla^2 \mathbb{E}\left[-\log \pi_n(\mu + n^{-1/2} LZ) \right] \succeq D_n(\epsilon - \tau_n(\mu, L) \cdot (\epsilon + \ell)), \end{aligned} \tag{3.11}$$

hold converges to 1 as $n \to \infty$, where D_n and $\tau_n(\mu, L)$ are as defined in Eq. (3.3) and $x = \mu_n^\star$. Note that the gradient and Hessian in the above expression are taken with respect to a vector in $\mathbb{R}^{d(d+1)}$ that stacks μ and each column of L into a single vector.

Then for all $(\mu, L) \in \mathcal{B}_{r,n}$, we have

$$\begin{aligned} &\|\mu - \mu_n^\star\|^2 \le r^2/4 \\ &\|L - L_n^\star\|_F^2 \le 4\|I - L_n^\star\|_F^2 \implies \|L\|_F \le 2\|I - L_n^\star\|_F + \|L_n^\star\|_F, \end{aligned}$$

yielding

$$\frac{r^2 - 2\|\mu - \mu_n^\star\|^2}{n^{-1}2\|L\|_F^2} \geq \frac{nr^2}{4\left(2\|I - L_n^\star\|_F + \|L_n^\star\|_F\right)^2}.$$

Hence $\forall(\mu, L) \in \mathcal{B}_{r,n}$, $\tau_n(\mu, L) \to 0$ as $n \to \infty$, yielding that under sufficiently large n,

$$\epsilon - \tau_n(\mu, L) \cdot (\epsilon + \ell) > \epsilon/2.$$

Therefore, the probability that for all $(\mu, L) \in \mathcal{B}_{r,n}$,

$$\frac{1}{n}\nabla^2 \mathbb{E}\left[-\log \pi_n(\mu + 1/\sqrt{n}LZ)\right] \succeq \frac{\epsilon}{2} D_n \tag{3.12}$$

converges in P_{θ_0} to 1 as $n \to \infty$.

Combining Eqs. (3.10) to (3.12), the proof is completed. □

Chapter 4

Consistent Stochastic Variational Inference (CSVI)

In this section, we use the strong convexity and smoothness results from Chapter 3 to develop an optimization algorithm—*Consistent Stochastic Variational Inference (CSVI)*—that *asymptotically solves* the Gaussian variational inference problem in Eq. (2.1), in the sense of Definition 4.0.1: the probability that the iterates converge to the global optimum converges to 1 in the asymptotic limit of observed data.

Definition 4.0.1. An iterative algorithm *asymptotically solves* a (random) sequence of optimization problems indexed by $n \in \mathbb{N}$, each with a single global optimum point $x_n^\star \in \mathbb{R}^d$, if the sequence of iterates $(x_{k,n})_{k\in\mathbb{N}}$ produced by the algorithm satisfies

$$\lim_{n\to\infty} \mathbb{P}\left(\lim_{k\to\infty} \|x_{k,n} - x_n^\star\| = 0\right) = 1.$$

As mentioned earlier, the CSVI algorithm is based on projected stochastic gradient descent (SGD) (Bubeck, 2015, Section 3.). Since we know that the expectation component F_n of the Gaussian variational inference objective function is asymptotically locally strongly convex and globally Lipschitz smooth, there are three remaining issues to address to ensure that the algorithm satisfies Definition 4.0.1. First, we need to ensure that we can initialize the optimization algorithm within the locally strongly convex region, i.e., the set $\mathcal{B}_{r,n}$ from Corollary 3.3.8. We

address this challenge by solving a smoothed version of the *maximum a posteriori (MAP)* problem, which is formulated and shown to be asymptotically tractable in Section 4.1. Second, since the gradient estimates are noisy, we need to ensure that the iterates of CSVI stay in $\mathcal{B}_{r,n}$ for all iterations so that the usual convergence guarantees of SGD apply. Finally, note that the regularization term $-n^{-1} \log \det L$ in the objective of Eq. (2.1) is not itself Lipschitz smooth, making the overall optimization problem not Lipschitz smooth. We address both the second and third issue in Section 4.2 by applying a novel scaling matrix to the gradient steps, and developing new theoretical results on the local confinement of projected stochastic gradient descent.

4.1 Initialization via smoothed MAP

The goal of this section is to design an algorithm that initializes the variational optimization within the asymptotically strongly convex local region $\mathcal{B}_{r,n}$. Note that the challenging part of this problem is to set μ, as we can simply initialize $L = I$. Since we aim to initialize μ sufficiently close to $\mu_n^\star$—and $\mu_n^\star$ converges to θ_0 per Theorem 3.2.1—we might like to find something akin to the maximum a posteriori (MAP) value of $\log \pi_n$, due to its similar convergence to θ_0. However, since $\log \pi_n$ is typically not concave, obtaining the MAP point is generally intractable.

In this section, we formulate a tractable MAP-like problem by convolving the posterior distribution with Gaussian noise prior to finding the maximum point of the log density. This Gaussian noise essentially results in smoothed log density with fewer spurious optima; hence, we denote this the *smoothed MAP problem*. In Section 4.1.1, we show that the smoothed MAP problem is asymptotically convex—and hence tractable—under conditions similar to those that guarantee the Bernstein-von Mises theorem, and that the solution is asymptotically consistent for θ_0, albeit at a slower-than-$\sqrt{n}$ rate. In Section 4.1.2, we provide a stochastic optimization algorithm that depends only on black-box access to the gradients of the original log density function. Taken together, these results demonstrate that we can tractably initialize CSVI in the asymptotically strongly convex local region $\mathcal{B}_{r,n}$ as required.

4.1.1 Smoothed MAP problem

Given the n^{th} posterior distribution Π_n, we define the *smoothed posterior* $\hat{\Pi}_n$ with smoothing variance α_n to be the θ-marginal of the generative process

$$W \sim \Pi_n, \quad \theta \sim \mathcal{N}(W, \alpha_n I).$$

Alternatively, the smoothed posterior distribution $\hat{\Pi}_n$ can be viewed as the distribution of the sum of independent realizations from Π_n and $\mathcal{N}(0, \alpha_n I)$. The probability density function $\hat{\pi}_n$ of $\hat{\Pi}_n$ is given by the convolution of π_n with a multivariate normal density,

$$\begin{aligned}\hat{\pi}_n(\theta) =& \int \frac{1}{(2\pi)^{d/2}\alpha_n^{d/2}} \exp\left(-\frac{1}{2\alpha_n}\|\theta - w\|^2\right) \pi_n(w) dw \\ =& \mathbb{E}\left[\frac{1}{(2\pi)^{d/2}\alpha_n^{d/2}} \exp\left(-\frac{1}{2\alpha_n}\|\theta - W\|^2\right)\right]. \end{aligned} \tag{4.1}$$

Given these definitions, the *smoothed MAP problem* is the MAP inference problem for the smoothed posterior distribution, i.e.,

$$\hat{\theta}_n = \underset{\theta \in \mathbb{R}^d}{\arg\min} \; -\log \mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta - W\|^2\right)\right]. \tag{4.2}$$

Gaussian smoothing is commonly used in image and signal processing (Forsyth and Ponce, 2002; Haddad and Akansu, 1991; Lindeberg, 1990; Nixon and Aguado, 2012), and has previously been applied to reduce the presence of spurious local optima in nonconvex optimization problems, making them easier to solve with local gradient-based methods (Addis et al., 2005; Mobahi, 2013). The variance α_n controls the degree of smoothing; larger values create a smoother density $\hat{\pi}_n$, at the cost of making $\hat{\pi}_n$ a poorer approximation of the original function π_n. Figure Fig. 4.1 demonstrates how increasing α_n increases the smoothing effect, resulting in fewer and flatter local optima in the objective.

In general, although intuitively reasonable, Gaussian smoothing does not typically come with strong practical theoretical guarantees, essentially because the best choice of the smoothing variance α_n is not known. Mobahi (2013) shows for

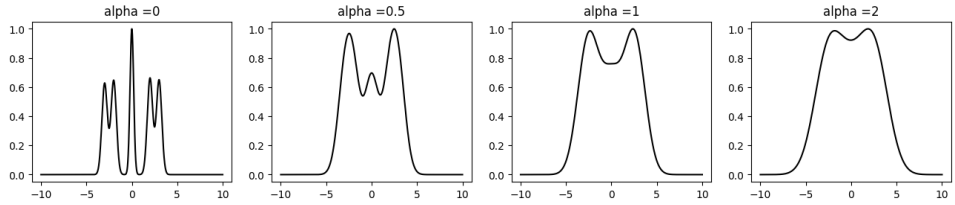

Figure 4.1: Plots of the smoothed posterior density $\hat{\pi}_n$ with increasing smoothing variance.

a continuous integrable function with quickly decaying tails (at rate $\|x\|^{-d-3}$ as $\|x\| \to \infty$), the smoothed function is strictly convex given a large enough selection of α_n. Addis et al. (2005) studies the smoothing effect of a log-concave kernel on a special type of piecewise constant function, and proves that the smoothed function is either monotonic or unimodal. To the best of our knowledge, previous analyses of smoothed optimization are not sufficient to guide the choice of α_n and do not provide bounds on the error of the smoothed optimum point versus the original.

In contrast to these previous studies, we use the asymptotic concentration of the statistical model as $n \to \infty$ to address both of these issues. In particular, Theorem 4.1.1 shows that if the sequence α_n is chosen to decrease more slowly than $n^{-1/3}$, the smoothed MAP problem is eventually strictly convex within any arbitrary compact domain; and that the solution of the smoothed MAP problem is asymptotically consistent for θ_0 at a $\sqrt{\alpha_n}^{-1}$ rate.

Theorem 4.1.1. *Suppose Assumption 1 holds and $n\alpha_n^3 \to \infty$. Then for all $M > 0$, the probability that the smoothed MAP optimization problem*

$$\min_{\|\theta - \theta_0\| \leq M} -\log \hat{\pi}_n(\theta)$$

is strictly convex converges to 1 as $n \to \infty$ under the data generating distribution. If additionally Assumption 2 holds, then

$$\|\hat{\theta}_n - \theta_0\| = O_{P_{\theta_0}}(\sqrt{\alpha_n}).$$

4.1.2 Smoothed MAP optimization

In practice, we use SGD to solve the smooth MAP problem. The gradient of the smoothed MAP objective function in Eq. (4.2) is

$$\begin{aligned}\nabla(-\log\hat{\pi}_n(\theta)) &= -\nabla\log\mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta-W\|^2\right)\right] \\ &= -\frac{\mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta-W\|^2\right)\left(-\frac{1}{\alpha_n}\right)(\theta-W)\right]}{\mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta-W\|^2\right)\right]},\end{aligned}$$

where $W \sim \Pi_n$. By change of variables and reparametrization, the gradient can be reformulated as

$$\nabla(-\log\hat{\pi}_n(\theta)) = \alpha_n^{-1/2}\frac{\mathbb{E}\left[W\pi_n\left(\theta-\alpha_n^{1/2}W\right)\right]}{\mathbb{E}\left[\pi_n\left(\theta-\alpha_n^{1/2}W\right)\right]},$$

where $W \sim \mathcal{N}(0, I)$. Note that the unknown normalization constant in π_n cancels in the numerator and denominator. We obtain stochastic estimates of the gradient using a Monte Carlo approximation of the numerator and denominator using the same samples, i.e., self-normalized importance sampling (Robert and Casella, 2013, p. 95). It is known that the variance of this gradient estimate may be quite large or even infinite; although techniques such as truncation (Ionides, 2008) and smoothing (Vehtari et al., 2015) exist to address it, we leave this issue as an open problem for future work. The resulting SGD procedure with explicit gradient estimates are shown in Algorithm 1.

4.2 Optimization via scaled projected SGD

As shown in Chapter 3 and Section 4.1, Gaussian variational inference Eq. (2.1) is locally strongly convex, globally Lipschitz smooth, and the initialization $\mu = \hat{\theta}_n$, $L = I$ is asymptotically tractable and lies in the locally convex region. We now design a stochastic optimization algorithm and prove that it asymptotically solves Eq. (2.1) per Definition 4.0.1.

Given a sequence of step sizes $(\gamma_k)_{k\in\mathbb{N}}$, $\gamma_k \geq 0$ and $(Z_k)_{k\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$, and

Algorithm 1 Smoothed MAP estimation

procedure SMOOTHEDMAP(π_n, α_n, x_0, K, S, $(\gamma_k)_{k\in\mathbb{N}}$)
  for $k = 0, 1, \ldots, K-1$ **do**
    Sample $(Z_s)_{s=1}^S \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$
    $\hat{\nabla}_k \leftarrow \alpha_n^{1/2}\Big(\sum_{s=1}^S Z_s \cdot \pi_n(x_k - \alpha_n^{1/2} Z_s)\Big) \Big/ \Big(\sum_{s=1}^S \pi_n(x_k - \alpha_n^{1/2} Z_s)\Big)$
    $x_{k+1} \leftarrow x_k - \gamma_k \hat{\nabla}_k$
  end for
  return x_K
end procedure

initialization $\mu_0 = \hat{\theta}_n, L_0 = I$, the standard stochastic gradient update applied to the Gaussian variational inference problem is

$$\mu_{k+1} \leftarrow \mu_k - \gamma_k \hat{\nabla}_{\mu,n}(\mu_k, L_k, Z_k)$$
$$L_{k+1} \leftarrow L_k - \gamma_k \hat{\nabla}_{L,n}(\mu_k, L_k, Z_k).$$

There are two major issues with this update: first, the regularization term $-\frac{1}{n}\mathrm{logdet} L$ is not globally Lipschitz smooth and has a large gradient when L is small, which is likely to produce an iterate outside the locally convex area; and second, the update might produce an infeasible iterate L with nonpositive diagonal.

We resolve the first issue by applying a scaling to the L gradient prior to the update. In particular, define the scaled L gradient matrix $\tilde{\nabla}_{L,n}(\mu, L, Z) \in \mathbb{R}^{d\times d}$ via

$$\Big[\tilde{\nabla}_{L,n}(\mu, L, Z)\Big]_{ij} = \begin{cases} \Big[\hat{\nabla}_{L,n}(\mu, L, Z)\Big]_{ij} & j \neq i \\ \frac{1}{1+(nL_{ii})^{-1}} \Big[\hat{\nabla}_{L,n}(\mu, L, Z)\Big]_{ii} & j = i, L_{ii} > 0 \\ -1 & j = i, L_{ii} = 0. \end{cases} \tag{4.3}$$

This scaling prevents the gradient of L from diverging when diagonal elements of $L \to 0$, and also creates a well-defined gradient for L at the boundary of the feasible region. Given this scaled L gradient, we resolve the second issue by employing a simple projection step after each update: we set any negative diagonal entry in the current iterate L_k to 0. SGD with these two simple modifications is presented in Algorithm 2. The final theoretical result of this work, Theorem 4.2.1, is that

Algorithm 2 Consistent stochastic Gaussian variational inference

procedure CSVI(f_n, g, $(\gamma_k)_{k\in\mathbb{N}}$, K)
  $\mu_0 \leftarrow$ `SmoothedMAP` (Algorithm 1)
  $L_0 \leftarrow I$
  for $k = 0, 1, \ldots, K-1$ **do**
    Sample $Z_k \sim \mathcal{N}(0, I)$
    $\mu_{k+1} \leftarrow \mu_k - \gamma_k \hat{\nabla}_{\mu,n}(\mu_k, L_k, Z_k)$
    $\tilde{\nabla}_{L,n} \leftarrow \hat{\nabla}_{L,n}(\mu_k, L_k, Z_k)$
    for $i = 1, \ldots, d$ **do**
      if $L_{k,ii} > 0$ **then**
        $\left[\tilde{\nabla}_{L,n}\right]_{ii} \leftarrow \frac{1}{1+(nL_{k,ii})^{-1}} \left[\tilde{\nabla}_{L,n}\right]_{ii}$
      else
        $\left[\tilde{\nabla}_{L,n}\right]_{ii} \leftarrow -1$
      end if
    end for
    $L_{k+1} \leftarrow L_k - \gamma_k \tilde{\nabla}_{L,n}$
    for $i = 1, \ldots, d$ **do**
      $L_{k+1,ii} \leftarrow \max\{0, L_{k+1,ii}\}$
    end for
  end for
  return μ_K, L_K
end procedure

Algorithm 2 asymptotically solves Gaussian variational inference given a weak condition on the step size sequence $(\gamma_k)_{k\in\mathbb{N}}$.

Theorem 4.2.1. *Suppose that we initialize $L_0 = I$ and μ_0 such that $\|\mu_0 - \hat{\theta}_n\|_2^2 \leq \frac{r^2}{32}$. Then there exists a constant $C > 0$ such that if*

$$\gamma_k = \Theta(k^{-\rho}) \text{ for some } \rho \in (0.5, 1) \quad \text{and} \quad \forall k \in \mathbb{N}, \quad 0 < \gamma_k < C, \tag{4.4}$$

then Algorithm 2 asymptotically solves Gaussian variational inference.

Chapter 5

Experiments

In this section, we compare CSVI to standard Gaussian stochastic variational inference (SVI)—i.e., reparametrized Gaussian variational inference via projected stochastic gradient descent—on one-dimensional synthetic inference problems. We run all optimization algorithms for $50{,}000$ iterations, and base the gradients for the smoothed MAP mean initialization (Algorithm 1) on 100 samples.

5.1 Synthetic Gaussian mixture

In the first experiment, we highlight the reliability of CSVI as opposed to SVI on the simple problem of approximating a Gaussian mixture target distribution Π,

$$\Pi = 0.7\mathcal{N}(0, 4) + 0.15\mathcal{N}(-30, 9) + 0.15\mathcal{N}(30, 9). \tag{5.1}$$

We set $n = 1$ and $\alpha_n = 10$ in the implementation of CSVI, and initialize the smoothed MAP optimization and the mean of SVI uniformly in the range $(-50, 50)$. For both methods, we initialize the log standard deviation $\log \sigma$ uniformly in the range $(\log 0.5, \log 10)$. We set $\gamma_k \approx 0.1/(1+k^{0.85})$ for CSVI and $\gamma_k = 0.5/(1+k)$ for SVI.

Fig. 5.1 shows the result of 10 trials of each of CSVI and SVI. The majority of the mass of the Gaussian mixture target distribution (grey) concentrates on the central mode with mean 0 and standard deviation 2; the optimal variational approximation has these same parameters. However, as shown in the plot, SVI

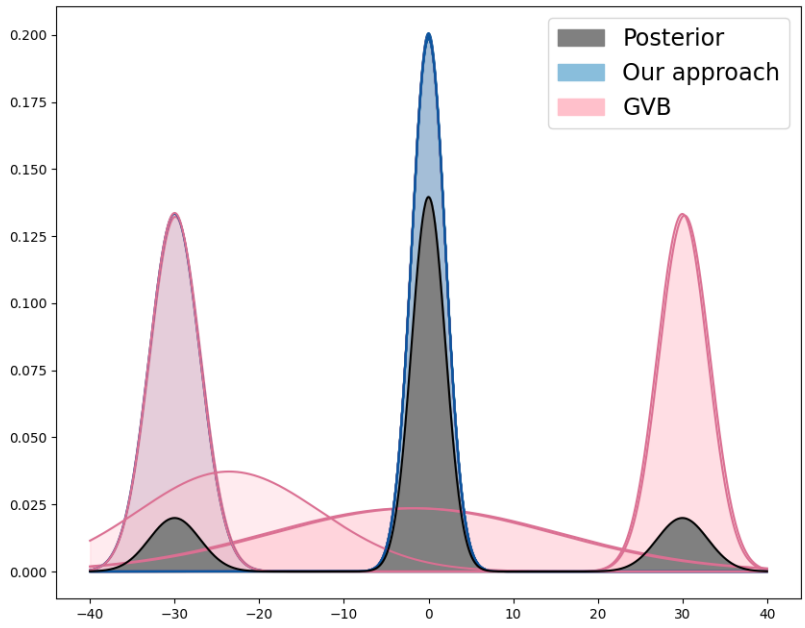

Figure 5.1: The result of running 10 trials of CSVI (blue) and SVI (pink) with the Gaussian mixture target (grey) given in Eq. (5.1). The output of CSVI reliably finds the global optimum solution corresponding to the central mixture peak; SVI often provides solutions corresponding to spurious local optima.

with randomized μ_0 often finds a spurious local optimum solution, corresponding to the two peaks centered at ± 30. On the other hand, the smoothed MAP mean initialization helps guarantee that the CSVI optimization starts close enough to the central mode that it recovers the global optimum solution in each trial. The gradient scaling also aids the stability of the algorithm; whereas SVI has unstable behaviour when σ is initialized to a small value due to the log-determinant regularization term, the scaling of CSVI in Eq. (4.3) ensures that the algorithm is stable in this region.

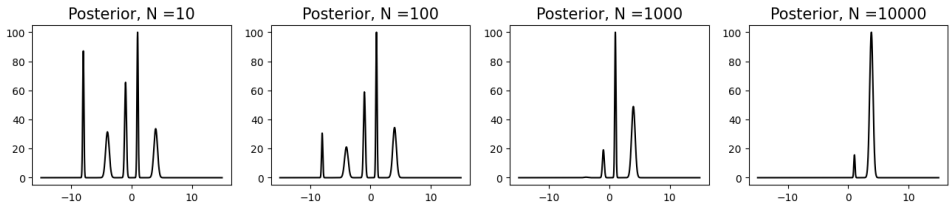

Figure 5.2: The Bayesian posterior density for increasing dataset sizes. Note the large number of spurious local optima, resulting in the unreliability of local optimization methods in variational inference.

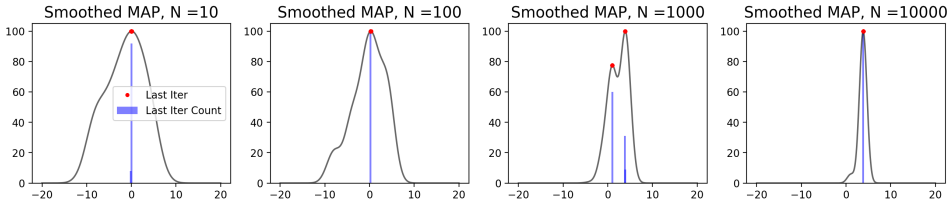

Figure 5.3: The smoothed Bayesian posterior density for the same dataset sizes as in Fig. 5.2. Black curves correspond to the smoothed posterior, red dots show local optima of the density, and the blue histogram shows the counts (over 100 trials) of the output of the smoothed MAP initialization. Note that there are fewer local optima relative to the original posterior density, and that the smoothed MAP initialization is likely to provide a mean close to that of the optimal variational distribution.

5.2 Synthetic model with a nonconvex prior

In this section, we compare the performance of CSVI and SVI on a synthetic Bayesian model across a range of observed dataset sizes ($n = 10, 100, 1000, 10000$). The model is as follows,

$$\begin{aligned}\theta &\sim \frac{1}{5}\mathcal{N}(0, 0.15^2) + \frac{1}{5}\mathcal{N}(1, 0.1^2) + \frac{1}{5}\mathcal{N}(-4, 0.3^2) + \frac{1}{5}\mathcal{N}(4, 0.3^2) \\ &\quad + \frac{1}{5}\mathcal{N}(-8, 0.1^2) \\ X_i \,|\, \theta &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 5000),\end{aligned}$$

where the data are truly generated from $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(3, 10)$. We use a smoothing constant of $\alpha_n = 10n^{-0.3}$, and set $\gamma_k = \frac{n}{10}/(1 + k^{0.85})$ for CSVI and $\gamma_k =$

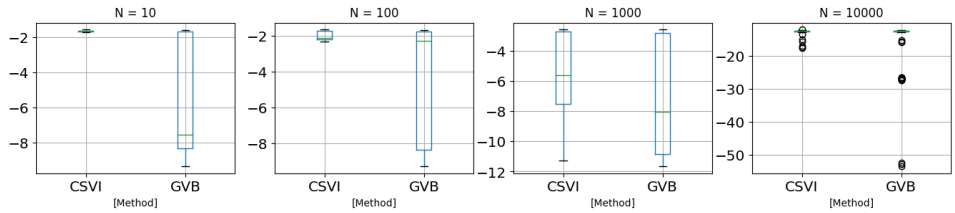

Figure 5.4: Box-plots of the final ELBO for 100 trials of CSVI and SVI.

$0.5/(1+k)$ for SVI. We construct this model to have a posterior distribution that has multiple modes and approaches a normal distribution in total variation as sample size increases. As shown in Fig. 5.2, the posterior has a number of peaks when n is small, and gradually converges to a unimodal distribution as n increases. However, even seemingly small peaks in the density may present significant local optima in the log density that prevent SVI from obtaining the global optimum, as illustrated in previous Gaussian mixture example. In contrast, Fig. 5.3 shows that the smoothed posterior tends to have fewer modes, smoothing smaller peaks and merging closeby peaks; thus the smoothed posterior can be used to find a *reliable initialization* for SVI and reduce the likelihood that SVI becomes trapped in bad local optimum. And because the smoothed MAP usually finds dominating peaks of the posterior, e.g., the larger peak in the last figure of Fig. 5.2, this initialization is usually closer to the optimal mean and thus often results in a better final variational approximation.

Fig. 5.4 provides a quantitative comparison of CSVI and SVI on this problem, confirming this intuition and demonstrating that CSVI more reliably finds low-cost variational approximations in comparison to SVI. In particular, Fig. 5.4 compares the final *expectation lower bound (ELBO)* Blei et al. (2017) for each method, which is equivalent to the negative KL divergence between posterior and variational distribution up to a normalizing constant. For comparison, we estimate ELBO using 1000 Monte Carlo samples.

The box-plots of Fig. 5.4 shows the results of running CSVI and SVI for 100 trials. Note that larger ELBO means a better approximation and a less spread box-plot represents better numerical stability. It is clear that when $n = 10000$, SVI tends to become trapped in local optima while CSVI tends to find the global optimum reliably; under all sample sizes, CSVI tends to provide a higher ELBO than SVI,

meaning a more accurate approximation. Moreover, CSVI is significantly more stable than SVI when the true posterior is peaky (e.g., $n = 10, 100$). We also find that training SVI on a peaky distribution can have diverging σ and hence yields numerical instability. CSVI, in contrast, shows great numerical stability because of the scaled gradient for σ.

It is worth noting that the performance of CSVI seems to degrade at $n = 1000$. We find that the posterior at $n = 1000$ in this example (3^{rd} figures from the left in Fig. 5.2) happens to have two big peaks in the smoothed posterior density. Thus the smoothed MAP initialization finds these two peaks with similar probability (see the blue bars of Fig. 5.3 at $n = 1000$), leading to similar final ELBO values when CSVI/SVI converges to these local optima. However, even in this pathological setting, CSVI provides a benefit over SVI, as the smoothing kernel removes other spurious local optima. Finally, it is worth pointing out that as with all local optimization methods, a careful choice of the learning rate for CSVI is important to ensure its performance. This matches the statement of Theorem 4.2.1, which guarantees that the output of CSVI is asymptotically consistent for θ_0 as long as the learning rate satisfies Eq. (4.4). However, if the learning rate is too large, CSVI may jump out of the local basin found by its initialization due to the noise in the gradient, and eventually become trapped in a spurious local optimum. But even when the learning rate is not carefully tuned, CSVI performs at least as well as SVI.

Chapter 6

Conclusion

This work provides an extensive theoretical analysis of the computational aspects of Gaussian variational inference, and uses the theory to design a general procedure that addresses the nonconvexity of the problem in the data-asymptotic regime. We show that under mild conditions, Gaussian variational optimization is locally asymptotically convex. Based on this fact, we developed consistent stochastic variational inference (CSVI), a scheme that asymptotically solves Gaussian variational inference. CSVI solves a smoothed MAP problem to initialize the Gaussian mean within the locally convex area, and then runs a scaled projected stochastic gradient descent to create iterates that converge to the optimum. The asymptotic consistency of CSVI is mathematically justified, and experimental results demonstrate the advantages over traditional SVI.

There are many avenues of further exploration for the present work. For example, we limit consideration to the case of Gaussian variational families due to their popularity; but aside from the mathematical details, nothing about the overall strategy necessarily relied on this choice. It would be worth examining other popular variational families, such as mean-field exponential families (Xing et al., 2002).

Furthermore, the current work is limited to posterior distributions with full support on $\mathbb{R}^d$—otherwise, the KL divergence variational objective is degenerate. It would be of interest to study whether variational inference using a Gaussian variational family truncated to the support of the posterior possesses the same bene-

ficial asymptotic properties and asymptotically consistent optimization algorithm as developed in the present work.

Another interesting potential line of future work is to investigate other probability measure divergences as variational objectives. For example, the chi-square divergence (Csiszár, 1967; Liese and Vajda, 1987, p. 51), Rényi α-divergence (Van Erven and Harremos, 2014), Stein discrepancy (Stein, 1972), and more (Gibbs and Su, 2002) have all been used as variational objectives. Along a similar vein, we studied the convergence properties of only a relatively simple stochastic gradient descent algorithm; other base algorithms with better convergence properties exist (Duchi et al., 2011; Kingma and Ba, 2015; Nesterov, 1983), and it may be fruitful to see if they have similar asymptotic consistency properties.

A final future direction is to investigate the asymptotic behaviour of variational inference with respect to other measures of optimization tractability. In particular, (local) pseudoconvexity (Crouzeix and Ferland, 1982), quasiconvexity (Arrow and Enthoven, 1961), and invexity (Ben-Israel and Mond, 1986; Craven and Glover, 1985) are all weaker than (local) convexity, but provide similar guarantees for stochastic optimization. These may be necessary to consider when examining other divergences as variational objectives.

Bibliography

Addis, B., Locatelli, M., and Schoen, F. (2005). Local optima smoothing for global optimization. *Optimization Methods and Software*, 20(4-5):417–437.

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

Arrow, K. and Enthoven, A. (1961). Quasi-concave programming. *Econometrica: Journal of the Econometric Society*, pages 779–800.

Bassett, R. and Deride, J. (2019). Maximum a posteriori estimators as a limit of Bayes estimators. *Mathematical Programming*, 174(1-2):129–144.

Bauschke, H. and Combettes, P. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

Ben-Israel, A. and Mond, B. (1986). What is invexity? *The ANZIAM Journal*, 28(1):1–9.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

Bottou, L. (2004). Stochastic Learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003*, pages 146–168. Springer Berlin Heidelberg.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. (2015). *Convex Optimization: Algorithms and Complexity*. Now Publishers Inc.

Craven, B. and Glover, B. (1985). Invex functions and duality. *Journal of the Australian Mathematical Society*, 39(1):1–20.

Crouzeix, J.-P. and Ferland, J. (1982). Criteria for quasi-convexity and pseudo-convexity: relationships and comparisons. *Mathematical Programming*, 23(1):193–205.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318.

Dashti, M., Law, K., Stuart, A., and Voss, J. (2013). MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159.

Folland, G. (1999). *Real Analysis: Modern Techniques and their Applications*. John Wiley & Sons.

Forsyth, D. and Ponce, J. (2002). *Computer Vision: a Modern Approach*. Prentice Hall Professional Technical Reference.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Gibbs, A. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

Grendár, M. and Judge, G. (2009). Asymptotic equivalence of empirical likelihood and Bayesian MAP. *The Annals of Statistics*, pages 2445–2457.

Haddad, R. and Akansu, A. (1991). A class of fast Gaussian binomial filters for speech and image processing. *IEEE Transactions on Signal Processing*, 39(3):723–727.

Han, W. and Yang, Y. (2019). Statistical inference in mean-field variational Bayes. *arXiv:1911.01525*.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

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

Huggins, J., Kasprzak, M., Campbell, T., and Broderick, T. (2020). Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*.

Ionides, E. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Jaiswal, P., Rao, V., and Honnappa, H. (2019). Asymptotic consistency of $\alpha-$ Rényi-approximate posteriors. *arXiv:1902.01902*.

Jennrich, R. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1998). An introduction to variational methods for graphical models. In *Learning in Graphical Models*, pages 105–161. Springer.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kleijn, B. (2004). *Bayesian asymptotics under misspecification*. PhD thesis, Vrije Universiteit Amsterdam.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

LeCam, L. (1960). *Locally Asymptotically Normal Families of Distributions*. Berkeley: University of California Press.

Lehmann, E. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner.

Lindeberg, T. (1990). Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):234–254.

Meyn, S. and Tweedie, R. (2012). *Markov Chains and Stochastic Stability*. Springer Science & Business Media.

Mobahi, H. (2013). *Optimization by Gaussian smoothing with application to geometric alignment*. PhD thesis, University of Illinois at Urbana-Champaign.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(k^2)$. In *Doklady Akademii Nauk*.

Nixon, M. and Aguado, A. (2012). *Feature Extraction and Image Processing for Computer Vision*. Academic Press.

Pinheiro, J. and Bates, D. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296.

Qiao, Y. and Minematsu, N. (2010). A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *International Coference on International Conference on Machine Learning*.

Ranganath, R. (2014). Black box variational inference. In *Advances in Neural Information Processing Systems*.

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

Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. *Probability surveys*, 1:20–71.

Schatzman, M. (2002). *Numerical Analysis: a Mathematical Introduction*. Clarendon Press. Translation: John Taylor.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714.

Stefanski, L. and Boos, D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. and Wellner, J. (2013). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv:1507.02646*.

Wainwright, M. and Jordan, M. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc.

Wang, Y. and Blei, D. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Xing, E., Jordan, M., and Russell, S. (2002). A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.

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

Appendix A

Proofs

In this appendix, we provide proofs of Theorems 4.1.1 and 4.2.1.

A.1 Proof for Theorem 4.1.1

A.1.1 Gradient and Hessian derivation

The gradient for smoothed posterior is as follows,

$$\begin{aligned}\nabla \log \hat{\pi}_n(\theta) =& \nabla \log \left\{ \mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta - W\|^2 \right) \right] \right\} \\ =& \frac{\mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta - W\|^2 \right) \left(-\frac{1}{\alpha_n} \right) (\theta - W) \right]}{\mathbb{E}\left[\exp\left(-\frac{1}{2\alpha_n}\|\theta - W\|^2 \right) \right]},\end{aligned}$$

and the Hessian matrix is given by

$$\nabla^2 \log \hat{\pi}_n(\theta) = \frac{1}{\alpha_n^2} \frac{\mathbb{E}\left[e^{\frac{-\|\theta - W\|^2 - \|\theta - W'\|^2}{2\alpha_n}} W(W - W')^T \right]}{\mathbb{E}\left[e^{-\frac{\|\theta - W\|^2}{2\alpha_n}} \right]^2} - \frac{1}{\alpha_n} I, \qquad \text{(A.1)}$$

where $W, W' \overset{\text{i.i.d.}}{\sim} \Pi_n$.

A.1.2 Proof of 1^{st} statement of Theorem 4.1.1

Proof of 1^{st} statement of Theorem 4.1.1. To show the MAP estimation for smoothed posterior is asymptotically strictly convex, we will show that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\|\theta-\theta_0\|\le M} \lambda_{\max}\left(\nabla^2 \log \hat{\pi}_n(\theta)\right) < 0 \right) = 1.$$

We focus on the first term of Eq. (A.1), and show that asymptotically it is uniformly smaller than α_n^{-1} so that the overall Hessian is negative definite. For the denominator of Eq. (A.1), define $B_n := \{W, W' : \max\{\|W' - \theta_0\|, \|W - \theta_0\|\} \le \beta_n\}$ for any sequence $\beta_n = o(\alpha_n)$. Then we have

$$\begin{aligned}
\mathbb{E}\left[e^{-\frac{\|\theta-W\|^2}{2\alpha_n}}\right]^2 &= \mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}} 1_{B_n}\right] + \mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}} 1_{B_n^c}\right] \\
&\ge \mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}} 1_{B_n}\right] \\
&\ge \mathbb{E}\left[\left(\inf_{v,v'\in B_n} e^{\frac{-\|\theta-v\|^2-\|\theta-v'\|^2}{2\alpha_n}}\right) 1_{B_n}\right] \\
&= \left(\inf_{v,v'\in B_n} e^{\frac{-\|\theta-v\|^2-\|\theta-v'\|^2}{2\alpha_n}}\right) \mathbb{P}(B_n).
\end{aligned}$$

By minimizing over $v, v' \in B_n$, the above leads to

$$\mathbb{E}\left[e^{-\frac{\|\theta-W\|^2}{2\alpha_n}}\right]^2 \ge e^{\frac{-2(\|\theta-\theta_0\|+\beta_n)^2}{2\alpha_n}} \mathbb{P}(B_n). \tag{A.2}$$

For the numerator of the first term of Eq. (A.1), since W, W' are i.i.d.,

$$\begin{aligned}
&\mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}} W(W-W')^T\right] \\
&\quad = \frac{1}{2}\mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}} (W-W')(W-W')^T\right],
\end{aligned}$$

and since $\lambda_{\max}\left((W-W')(W-W')^T\right) = \|W-W'\|^2$,

$$\lambda_{\max}\left(\mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}}W(W-W')^T\right]\right) \tag{A.3}$$
$$\leq \frac{1}{2}\mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}}\|W-W'\|^2\right].$$

With Eqs. (A.2) and (A.3), we can therefore bound the maximal eigenvalue of the Hessian matrix,

$$\lambda_{\max}\left(\nabla^2 \log \hat{\pi}_n(\theta)\right) \tag{A.4}$$
$$\leq \frac{1}{2\alpha_n^2\mathbb{P}(B_n)}\mathbb{E}\left[e^{\frac{-\|\theta-W\|^2-\|\theta-W'\|^2}{2\alpha_n}}e^{\frac{2(\|\theta-\theta_0\|+\beta_n)^2}{2\alpha_n}}\|W-W'\|^2\right] - \frac{1}{\alpha_n}.$$

We now bound the supremum of this expression over $\{\theta \in \mathbb{R}^d : \|\theta-\theta_0\| \leq M\}$. Focusing on the exponent within the expectation,

$$\begin{aligned}
&\sup_{\|\theta-\theta_0\|\leq M}\frac{1}{\alpha_n}\left[2(\|\theta-\theta_0\|+\beta_n)^2 - \|\theta-W\|^2 - \|\theta-W'\|^2\right] \\
=&\sup_{\|\theta-\theta_0\|\leq M}\frac{1}{\alpha_n}\left[2(\|\theta-\theta_0\|+\beta_n)^2 - \|\theta-\theta_0+\theta_0-W\|^2\right. \\
&\qquad\left.-\|\theta-\theta_0+\theta_0-W'\|^2\right] \\
\leq&\frac{1}{\alpha_n}\left[\left(2\beta_n^2+4M\beta_n\right) - \left(\|\theta_0-W\|^2+\|\theta_0-W'\|^2\right)\right. \\
&\left.+2M\left(\|\theta_0-W\|+\|\theta_0-W'\|\right)\right],
\end{aligned}$$

where the inequality is obtained by expanding the quadratic terms and bounding $\|\theta-\theta_0\|$ with M. We combine the above bound with Eq. (A.4) to show that $\alpha_n^2\lambda_{\max}\left(\nabla^2\log\hat{\pi}_n(\theta)\right)+\alpha_n$ is bounded above by

$$\frac{\beta_n}{2\mathbb{P}(B_n)}e^{\frac{2\beta_n^2+4M\beta_n}{\alpha_n}}\mathbb{E}\left[e^{\frac{2M\left(\|\theta_0-W\|+\|\theta_0-W'\|\right)-\left(\|\theta_0-W\|^2+\|\theta_0-W'\|^2\right)}{\alpha_n}}\frac{\|W-W'\|^2}{\beta_n}\right]. \tag{A.5}$$

By multiplying and dividing by $\exp\left(\frac{\|W-W'\|}{\sqrt{\beta_n}}\right)$, one notices that

$$\begin{aligned}\frac{\|W-W'\|^2}{\beta_n} =& \exp\left(\frac{\|W-W'\|}{\sqrt{\beta_n}}\right)\exp\left(-\frac{\|W-W'\|}{\sqrt{\beta_n}}\right)\frac{\|W-W'\|^2}{\beta_n}\\ \leq& 4e^{-2}\exp\left(\frac{\|W-\theta_0\|+\|W'-\theta_0\|}{\sqrt{\beta_n}}\right),\end{aligned}$$

where the inequality is by the fact that x^2e^{-x} maximized at $x=2$ with value $4e^{-2}$ and $\|W-W'\| \leq \|W\|+\|W'\|$. If we combine this bound with Eq. (A.5) and note that W, W' are iid, Eq. (A.5) is bounded above by

$$\frac{2e^{-2}\beta_n}{\mathbb{P}(B_n)}e^{\frac{2\beta_n^2+4M\beta_n}{\alpha_n}}\mathbb{E}\left[e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\|-\frac{1}{2\alpha_n}\|W-\theta_0\|^2}\right]^2. \tag{A.6}$$

To show that the Hessian is asymptotically negative definite, it suffices to show that Eq. (A.6) is $o_{P_{\theta_0}}(\alpha_n)$. For the terms outside the expectation, $\beta_n = o(\alpha_n)$ implies that $2e^{-2}\beta_n e^{\frac{2\beta_n^2+4M\beta_n}{\alpha_n}} = o(\alpha_n)$, and Assumption 1 and Lemma A.1.1 together imply that

$$\mathbb{P}(B_n) = \Pi_n\left(\{W : \|W-\theta_0\| \leq \beta_n\}\right)^2 \overset{P_{\theta_0}}{\to} 1,$$

so

$$\frac{2e^{-2}\beta_n}{\mathbb{P}(B_n)}e^{\frac{2\beta_n^2+4M\beta_n}{\alpha_n}} = o_{P_{\theta_0}}(\alpha_n).$$

Therefore, in order to show Eq. (A.6) is $o_{P_{\theta_0}}(\alpha_n)$, it is sufficient to show that

$$\mathbb{E}\left[e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\|-\frac{1}{2\alpha_n}\|W-\theta_0\|^2}\right] = O_{P_{\theta_0}}(1).$$

The next step is to split the expectation into two regions—$\|W-\theta_0\| \leq \beta_n$ and $\|W-\theta_0\| > \beta_n$—and bound its value within them separately.

1. When $\|W-\theta_0\| \leq \beta_n$, the exponent inside the expectation is shrinking

uniformly since $\beta_n = o(\alpha_n)$:

$$\begin{aligned}
&\mathbb{E}\left[1_{\{\|W-\theta_0\|\le\beta_n\}} e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\|-\frac{1}{2\alpha_n}\|W-\theta_0\|^2}\right] \\
&\le \mathbb{E}\left[1_{\{\|W-\theta_0\|\le\beta_n\}}\right] e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\beta_n} \\
&= O_{P_{\theta_0}}(1).
\end{aligned}$$

2. When $\|W-\theta_0\| > \beta_n$, we take the supremum over the exponent (a quadratic function), yielding $\|W-\theta_0\| = M + \alpha_n\beta_n^{-1/2}$ and the following bound,

$$\begin{aligned}
&\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\| - \frac{1}{2\alpha_n}\|W-\theta_0\|^2 \qquad \text{(A.7)} \\
&\le \sup_{\|v-\theta_0\|}\left(\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|v-\theta_0\| - \frac{1}{2\alpha_n}\|v-\theta_0\|^2\right) \\
&= \left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\left(M+\alpha_n\beta_n^{-1/2}\right) - \frac{1}{2\alpha_n}\left(M+\alpha_n\beta_n^{-1/2}\right)^2 \\
&= \frac{M^2}{2\alpha_n} + \frac{M}{\beta_n^{1/2}} + \frac{\alpha_n}{2\beta_n}.
\end{aligned}$$

This yields

$$\begin{aligned}
&\mathbb{E}\left[1_{\{\|W-\theta_0\|>\beta_n\}} e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\|-\frac{1}{2\alpha_n}\|W-\theta_0\|^2}\right] \\
&\le \Pi_n\left(\{W : \|W-\theta_0\| > \beta_n\}\right)\exp\left(\frac{M^2}{2\alpha_n} + \frac{M}{\beta_n^{1/2}} + \frac{\alpha_n}{2\beta_n}\right).
\end{aligned}$$

Note that it is always possible to choose $\beta_n = o(\alpha_n)$ with $\beta_n = \omega(\alpha_n^2)$. With this choice of β_n, the dominating term among the three of Eq. (A.7) is $\frac{M^2}{2\alpha_n}$. Then by Lemma A.1.1, there exists a sequence $\beta_n = o(\alpha_n)$ with $\beta_n = \omega(\alpha_n^2)$ such that the following holds,

$$\Pi_n(\{W : \|W-\theta_0\| > \beta_n\}) = o_{P_{\theta_0}}\left(\exp\left\{-\frac{M^2}{2\alpha_n}\right\}\right),$$

which implies

$$\mathbb{E}\left[1_{\{\|W-\theta_0\|>\beta_n\}} e^{\left(\frac{1}{\alpha_n}M+\beta_n^{-1/2}\right)\|W-\theta_0\|-\frac{1}{2\alpha_n}\|W-\theta_0\|^2}\right] = o_{P_{\theta_0}}(1).$$

This finishes the proof. □

In the last step of the above proof, we require an exponential tail bound for the posterior Π_n. We provide this in the following lemma, following the general proof strategy of van der Vaart (2000, Thm 10.3). The proof of Lemma A.1.1 involves many probability distributions; thus, for mathematical convenience and explicitness, in the proof of Lemma A.1.1 we use square bracket—$P[X]$—to denote the expectation of random variable X with respect to a probability distribution P. When taking expectation to a function of n data points $f(X_1, \ldots, X_n)$, where $(X_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_\theta$, we still write $P_\theta[f]$; and P_θ here represents the product measure.

Lemma A.1.1. *Under Assumption 1, $\alpha_n^3 n \to \infty$, there exists a sequence β_n satisfying $\beta_n = o(\alpha_n)$, $\beta_n = \omega(\alpha_n^2)$ and $\beta_n = \omega(n^{-1/2})$ such that for any fixed constant M,*

$$\Pi_n(\{W : \|W - \theta_0\| > \beta_n\}) = o_{P_{\theta_0}}\left(\exp\left\{-\frac{M^2}{2\alpha_n}\right\}\right).$$

Proof of Lemma A.1.1. In order to show that β_n satisfies the tail probability bound, it suffices to prove that

$$e^{\frac{1}{\alpha_n}} P_{\theta_0}\left[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})\right] \to 0,$$

due to Markov's inequality (we absorb the $M^2/2$ constant into α_n because it does not affect the proof). To achieve this, we take advantage of the existence of a test sequence applied from Assumption 1. By van der Vaart (2000, Lemma 10.6), given the 1st and the 2nd conditions of Assumption 1 and the fact that the parameter space $\mathbb{R}^d$ is σ-compact, there exists a sequence of tests $\phi_n : \mathcal{X}^n \to [0, 1]$, where $\mathcal{X}^n$ is the space of $(X_1, \ldots, X_n)$, such that as $n \to \infty$,

$$P_{\theta_0}[\phi_n] \to 0, \qquad \sup_{\|\theta-\theta_0\|>\varepsilon} P_\theta\left[1 - \phi_n\right] \to 0.$$

Further, by Kleijn (2004, Lemma 1.2) and van der Vaart (2000, Lemma 10.3), under Assumption 1 and the existence of the above test sequence ϕ_n, for every $M_n \to \infty$, there exists a constant $C > 0$ and another sequence of tests $\psi_n : \mathcal{X}^n \to [0, 1]$ such that for all $\|\theta - \theta_0\| > M_n/\sqrt{n}$ and sufficiently large n,

$$P_{\theta_0}[\psi_n] \leq \exp\{-Cn\}, \quad P_\theta\left[1 - \psi_n\right] \leq \exp\{-Cn(\|\theta - \theta_0\|^2 \wedge 1)\}. \tag{A.8}$$

Using ψ_n, we split the expectation as following,

$$\begin{aligned} & e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})] \\ &= \underbrace{e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})\psi_n]}_{\text{(I)}} \\ &\quad + \underbrace{e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})(1 - \psi_n)]}_{\text{(II)}}, \end{aligned}$$

and we aim to show both parts converging to 0.

For term (I), the first statement of Eq. (A.8) implies that $\exists C > 0$ such that

$$e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})\psi_n] \leq e^{\frac{1}{\alpha_n}} P_{\theta_0}[\psi_n] \leq e^{\frac{1}{\alpha_n}} e^{-nC}.$$

where the first inequality follows by $\Pi_n(\{W : \|W - \theta_0\| > \beta_n\}) \leq 1$. Since $n\alpha_n^3 \to \infty$, the last bound in the above expression converges to 0.

For term (II), we work with the shifted and scaled posterior distribution. Define $Z_n = \sqrt{n}(W - \theta_0)$ and $B_n = \{Z_n : \|Z_n\| > \sqrt{n\beta_n^2}\}$, and let $\tilde{\Pi}_0$ be the corresponding prior distribution on Z_n and $\tilde{\Pi}_n$ be the shifted and scaled posterior distribution, which yields

$$\begin{aligned} & e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}[\Pi_n(\{W : \|W - \theta_0\| > \beta_n\})(1 - \psi_n)] \\ &\quad = e^{\frac{1}{\alpha_n}} \cdot P_{\theta_0}\left[\tilde{\Pi}_n\left(B_n\right)\left(1 - \psi_n\right)\right]. \end{aligned} \tag{A.9}$$

Let U be a closed ball around 0 with a fixed radius r, then restricting $\tilde{\Pi}_0$ on U defines a probability measure $\tilde{\Pi}_0^U$, i.e., for all measurable set B, $\tilde{\Pi}_0^U(B) = \tilde{\Pi}_0(B \cap U)/\tilde{\Pi}_0(U)$. Write $P_{n,z}$ for the joint distribution of n data points $(X_1, \ldots, X_n)$

parameterized under $\theta_0 + z/\sqrt{n}$ and hence write the marginal distribution of $(X_1, \ldots, X_n)$ under $\tilde{\Pi}_0^U$ for $P_{n,U} = \int P_{n,z} \mathrm{d}\tilde{\Pi}_0^U(z)$. The densities of these distributions will be represented using lower case, e.g., $p_{n,U}(x) = \int p_{n,z}(x)\tilde{\pi}_0^U(z)\mathrm{d}z$ is the PDF of $P_{n,U}$. Here we abuse the notation that x represents $(x_1, \ldots, x_n)$.

We replace P_{θ_0} in Eq. (A.9) with $P_{n,U}$. Under Assumption 1, by van der Vaart (2000, P. 141), $P_{n,U}$ is *mutually contiguous* to P_{θ_0} (LeCam, 1960), that is, for any statistics T_n (a Borel function of X^n), $T_n \overset{P_{\theta_0}}{\to} 0$ iff $T_n \overset{P_{n,U}}{\to} 0$. Thus, considering $\tilde{\Pi}_0\left(B_n\right)\left(1-\psi_n\right)$ as the statistics T_n, the convergence to 0 of the expression in Eq. (A.9) is equivalent to

$$e^{\frac{1}{\alpha_n}} \cdot P_{n,U}\left[\tilde{\Pi}_n\left(B_n\right)\left(1-\psi_n\right)\right] \to 0.$$

Manipulating the expression of $P_{n,U}$ and $\tilde{\Pi}_n\left(B_n\right)$ (we write $\tilde{\Pi}_n\left(B_n, x\right)$ in the integral and write $\tilde{\Pi}_n\left(B_n, (X_i)_{i=1}^n\right)$ in the expectation to make the dependence of posterior on the data explicit),

$$\begin{aligned} P_{n,U}\left[\tilde{\Pi}_n\left(B_n, (X_i)_{i=1}^n\right)\left(1-\psi_n\right)\right] &= \int \tilde{\Pi}_n\left(B_n, x\right)\left(1-\psi_n\right) \mathrm{d}P_{n,U}(x) \\ &= \int \tilde{\Pi}_n\left(B_n, x\right)\left(1-\psi_n\right) p_{n,U}(x) \mathrm{d}x. \end{aligned}$$

Note that $p_{n,U}(x) = \int p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0^U(z)$,

$$= \int \tilde{\Pi}_n\left(B_n, x\right)\left(1-\psi_n\right)\left(\int p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0^U(z)\right)\mathrm{d}x.$$

Recall that for all measurable set B, $\tilde{\Pi}_0^U(B) = \tilde{\Pi}_0(B \cap U)/\tilde{\Pi}_0(U)$, thus

$$= \frac{1}{\tilde{\Pi}_0(U)} \int \tilde{\Pi}_n\left(B_n, x\right)\left(1-\psi_n\right)\left(\int_U p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)\right)\mathrm{d}x.$$

By using Bayes rule, we expand $\tilde{\Pi}_n\left(B_n, x\right) = \frac{\int \mathbb{1}[B_n] p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)}{\int p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)}$,

$$= \frac{\int (1-\psi_n)\left(\int \mathbb{1}[B_n] p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)\right)\left(\int_U p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)\right)\mathrm{d}x}{\tilde{\Pi}_0(U)\int p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)}.$$

Note that $\tilde{\Pi}_n\left(U, x\right) = \frac{\int_U p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)}{\int p_{n,z}(x)\mathrm{d}\tilde{\Pi}_0(z)}$,

$$= \frac{1}{\tilde{\Pi}_0(U)} \int \left(\int \mathbb{1}[B_n] p_{n,z}(x) \mathrm{d}\tilde{\Pi}_0(z) \right) (1 - \psi_n) \tilde{\Pi}_n\left(U, x\right) \mathrm{d}x.$$

By Fubini Theorem and $\tilde{\Pi}_n\left(U, x\right) \leq 1$,

$$\begin{aligned} &\leq \frac{1}{\tilde{\Pi}_0(U)} \int_{B_n} \left(\int (1 - \psi_n) p_{n,z}(x) \mathrm{d}x \right) \mathrm{d}\tilde{\Pi}_0(z) \\ &= \frac{1}{\tilde{\Pi}_0(U)} \int_{\{\|z\| > \sqrt{n\beta_n^2}\}} P_{n,z}[1 - \psi_n] \mathrm{d}\tilde{\Pi}_0(z). \end{aligned}$$

Note that $P_{n,z}[1 - \psi_n] \equiv P_\theta[1 - \psi_n]$ for $\theta = \theta_0 + z/\sqrt{n}$ and that $\sqrt{n\beta_n^2} \to \infty$ due to $\beta_n = \omega(n^{-1/2})$. Thus, we can use the second statement of Eq. (A.8) to bound $P_{n,z}[1 - \psi_n]$, yielding

$$\begin{aligned} &\frac{1}{\tilde{\Pi}_0(U)} \int_{\{\|z\| > \sqrt{n\beta_n^2}\}} P_{n,z}[1 - \psi_n] \mathrm{d}\tilde{\Pi}_0(z) \\ &\leq \frac{1}{\tilde{\Pi}_0(U)} \int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\{-C(\|z\|^2 \wedge n)\} \mathrm{d}\tilde{\Pi}_0(z). \end{aligned}$$

We then derive upper bounds for both the fraction and the integral to show the above is $o\left(e^{-\frac{1}{\alpha_n}}\right)$. For the fraction, we define $U_n := \left\{w \in \mathbb{R}^d : \sqrt{n}(w - \theta_0) \in U\right\}$, then

$$\tilde{\Pi}_0(U) = \Pi_0(U_n) \geq \frac{\pi^{\frac{d}{2}}}{\gamma(\frac{d}{2} + 1)} \left(n^{-1/2} r\right)^d \inf_{w \in U_n} \pi_0(w).$$

By Assumption 1, for all $w \in \mathbb{R}^d, \pi_0(w)$ is positive and continuous, and hence $\inf_{w \in U_n} \pi_0(w)$ is an increasing sequence that converges to $\pi_0(\theta_0) > 0$ as $n \to \infty$. Thus, there is a constant $D > 0$ such that for sufficiently large n,

$$\tilde{\Pi}_0(U) \geq D n^{-d/2}, \tag{A.10}$$

yielding that

$$\exists C > 0, \quad \text{s.t.} \quad \frac{1}{\tilde{\Pi}_0(U)} \leq C n^{d/2}.$$

For the integral, by splitting B_n into $\{\sqrt{n\beta_n^2} < \|z_n\| \leq k\sqrt{n}\}$ and $\{\|z_n\| > k\sqrt{n}\}$ for some positive $k < 1$,

$$\begin{aligned}
&\int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\{-C(\|z\|^2 \wedge n)\} \mathrm{d}\tilde{\Pi}_0(z) \\
&\leq \int_{\{k\sqrt{n} \geq \|z\| > \sqrt{n\beta_n^2}\}} \exp\{-C\|z\|^2\} \mathrm{d}\tilde{\Pi}_0(z) + e^{-Ck^2 n}.
\end{aligned}$$

Then by change of variable to $w = \frac{1}{\sqrt{n}} z + \theta_0$,

$$= \int_{\{k \geq \|w - \theta_0\| > \beta_n\}} \exp\{-Cn\|w - \theta_0\|^2\} \pi_0(w) \mathrm{d}w + e^{-Ck^2 n}.$$

Note that by Assumption 1, $\pi_0(w)$ is continuous for all $w \in \mathbb{R}^d$, we can choose a sufficiently small k such that $\pi_0(\theta)$ is uniformly bounded by a constant C over the region $\{k \geq \|w - \theta_0\| > \beta_n\}$. Thus, the above can be bounded above by

$$\begin{aligned}
&C \int_{\{k \geq \|w - \theta_0\| > \beta_n\}} \exp\{-Cn\|w - \theta_0\|^2\} \mathrm{d}w + e^{-Ck^2 n} \\
&= Cn^{-d/2} \int_{\{k\sqrt{n} \geq \|z\| > \sqrt{n\beta_n^2}\}} \exp\left\{-C\|z\|^2\right\} \mathrm{d}z + e^{-Ck^2 n} \\
&\leq Cn^{-d/2} \int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\left\{-C\|z\|^2\right\} \mathrm{d}z + e^{-Ck^2 n},
\end{aligned}$$

where the equality is by change of variable back to $z = \sqrt{n}(w - \theta_0)$. Then, consider the integral on RHS. Using spherical coordinates, there exists a fixed constant $D > 0$ such that

$$\begin{aligned}
\int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\left\{-C\|z\|^2\right\} \mathrm{d}z &= D \int_{\{r > \sqrt{n\beta_n^2}\}} e^{-Cr^2} r^{d-1} \mathrm{d}r \\
&= DC^{-d/2} \int_{\{s > Cn\beta_n^2\}} e^{-s} s^{\frac{d}{2} - 1} \mathrm{d}s,
\end{aligned}$$

where the second equality is by setting $s = Cr^2$. Note that the integrand of RHS is proportional to the PDF of $\Gamma(\frac{d}{2}, 1)$. Using the tail properties of the Gamma random variable (Boucheron et al., 2013, p. 28), we have that for some generic constant $D > 0$,

$$\int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\left\{-C\|z\|^2\right\} \mathrm{d}z \leq De^{-Cn\beta_n^2}.$$

Therefore, for some generic constants $C, D > 0$,

$$\begin{aligned} &\int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\{-C(\|z\|^2 \wedge n)\} \mathrm{d}\tilde{\Pi}_0(z) \\ &\leq Dn^{-d/2} e^{-Cn\beta_n^2} + e^{-Ck^2 n}. \end{aligned} \tag{A.11}$$

Then we combine Eqs. (A.10) and (A.11), yielding that for some constants $C, D > 0$ independent to n,

$$\begin{aligned} &e^{\frac{1}{\alpha_n}} \cdot P_{n,U}\left[\tilde{\Pi}_0\left(B_n\right)\left(1 - \psi_n\right)\right] \\ &\leq e^{\frac{1}{\alpha_n}} \frac{1}{\Pi_{n,0}(U)} \int_{\{\|z\| > \sqrt{n\beta_n^2}\}} \exp\{-C(\|z\|^2 \wedge n)\} \mathrm{d}\tilde{\Pi}_0(z) \\ &\leq e^{\frac{1}{\alpha_n}} C\sqrt{n}^d e^{-Cn} + e^{\frac{1}{\alpha_n}} De^{-Cn\beta_n^2}. \end{aligned}$$

Lastly, it remains to show that there exists a positive sequence β_n satisfying both $\beta_n = o(\alpha_n)$ and $\beta_n = \omega(\alpha_n^2)$ such that the RHS converges to 0. The first term always converges to 0 due to $\alpha_n^3 n \to \infty$. For the second term, we consider two different cases. If $\alpha_n = o(n^{-1/6})$, we pick $\beta_n = n^{-1/3}$, which is both $o(\alpha_n)$ and $\omega(\alpha_n^2)$. Then

$$\begin{aligned} e^{\frac{1}{\alpha_n}} De^{-Cn\beta_n^2} &= D \exp\left\{\alpha_n^{-1} - Cn^{1/3}\right\} \\ &\longrightarrow 0, \end{aligned}$$

where the convergence in the last line is by $n\alpha_n^3 \to \infty \Leftrightarrow \frac{1}{\alpha_n} = o(n^{1/3})$. If

$\alpha_n = \omega(n^{-1/6})$, we pick $\beta_n = \alpha_n^2$. Then

$$e^{\frac{1}{\alpha_n}} D e^{-Cn\beta_n^2} = D \exp\left\{\alpha_n^{-1} - Cn\alpha_n^4\right\}.$$

Since $\alpha_n = \omega(n^{-1/6})$, $\frac{1}{\alpha_n} = o(n^{1/6})$ and $n\alpha_n^4 = \omega(n^{1/3})$, yielding that the above converges to 0 as $n \to \infty$.

This completes the proof.

□

A.1.3 Proof of 2nd statement of Theorem 4.1.1

In this section, we show that the smoothed MAP estimator $(\hat{\theta}_n)$ is also a consistent estimate of θ_0, but with a convergence rate that is slower than the traditional $\sqrt{n}$. This is the case because the variance of the smoothing kernel satisfies $\alpha_n = \omega(n^{-1/3})$, and the convergence rate of $\hat{\theta}_n$ is determined by α_n via

$$\|\hat{\theta}_n - \theta_0\| = O_{P_{\theta_0}}(\sqrt{\alpha_n}). \tag{A.12}$$

Recall that $\theta_{\text{MAP},n} := \arg\max \pi_n(\theta)$ is the MAP estimator for the exact posterior, which is a $\sqrt{n}$-consistent estimate of θ_0. Thus, it is sufficient to show $\left\|\hat{\theta}_n - \theta_{\text{MAP},n}\right\| = O_{P_{\theta_0}}\left(\sqrt{\alpha_n}\right)$.

Note that $\hat{\theta}_n$ and $\theta_{\text{MAP},n}$ are maximals of stochastic process $\hat{\pi}_n(\theta)$ and $\pi_n(\theta)$ respectively, which can be studied in the framework of M-estimator (van der Vaart, 2000; van der Vaart and Wellner, 2013). A useful tool in establishing the asymptotics of M-estimators is the Argmax Continuous Mapping theorem (van der Vaart and Wellner, 2013, Lemma 3.2.1), which is introduced as follows.

Lemma A.1.2 (Argmax Continuous Mapping (van der Vaart and Wellner, 2013))**.** *Let $\{f_n(\theta)\}$ and $f(\theta)$ be stochastic processes indexed by θ, where $\theta \in \Theta$. Let $\hat{\theta}$ be a random element such that almost surely, for every open sets G containing $\hat{\theta}$,*

$$f(\hat{\theta}) > \sup_{\theta \notin G} f(\theta).$$

and $\hat{\theta}_n$ be a random sequence such that almost surely

$$f_n(\hat{\theta}_n) = \sup_{\theta \in \Theta} f_n(\theta).$$

If $\sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| = o_P(1)$ *as* $n \to \infty$*, then*

$$\hat{\theta}_n \xrightarrow{d} \hat{\theta}.$$

The proof strategy of the 2nd statement of Theorem 4.1.1 is to apply Lemma A.1.2 in a setting where f_n is $\hat{\pi}_n(\theta)$ and f is a Gaussian density. Using the Bernstein-von Mises Theorem 3.1.1, we show that $\hat{\pi}_n(\theta)$ converges uniformly to this Gaussian density, which implies that the MAP of $\hat{\pi}_n(\theta)$ converges in distribution to the MAP of this Gaussian distribution by the Argmax Continuous Mapping theorem. The detailed proof is as follows.

Proof of 2nd statement of Theorem 4.1.1. Assumption 2 guarantees $\|\theta_{\text{MAP},n} - \theta_0\| = O_{P_{\theta_0}}(1/\sqrt{n})$. Note that

$$\|\hat{\theta}_n - \theta_0\| \leq \|\hat{\theta}_n - \theta_{\text{MAP},n}\| + \|\theta_{\text{MAP},n} - \theta_0\|.$$

Since $\sqrt{\alpha_n} = \omega(1/\sqrt{n})$, in order to get Eq. (A.12), it suffices to show

$$\left\|\hat{\theta}_n - \theta_{\text{MAP},n}\right\| = O_{P_{\theta_0}}\left(\sqrt{\alpha_n}\right).$$

Thus, in this proof, we aim to show $\left\|\hat{\theta}_n - \theta_{\text{MAP},n}\right\| = O_{P_{\theta_0}}(\sqrt{\alpha_n})$ and it is sufficient to prove

$$\frac{1}{\sqrt{\alpha_n}}\left(\hat{\theta}_n - \theta_{\text{MAP},n}\right) \xrightarrow{P_{\theta_0}} 0.$$

Let $\xi = \frac{1}{\sqrt{\alpha_n}}(\theta - \theta_{\text{MAP},n})$, $\xi_n^* = \frac{1}{\sqrt{\alpha_n}}\left(\hat{\theta}_n - \theta_{\text{MAP},n}\right)$ and $t = \frac{1}{\sqrt{\alpha_n}}(w - \theta_{\text{MAP},n})$.

By expressing $\hat{\pi}_n(\theta)$, which is defined in Eq. (4.1),

$$\begin{aligned}
\xi_n^* &= \arg\max_{\xi} \hat{\pi}\left(\sqrt{\alpha_n}\xi + \theta_{\text{MAP},n}\right) \\
&= \arg\max_{\xi} \int \pi_n\left(\sqrt{\alpha_n}t + \theta_{\text{MAP},n}\right) \exp\left(-\frac{1}{2\alpha_n}\|\sqrt{\alpha_n}\xi - \sqrt{\alpha_n}t\|^2\right) \mathrm{d}t \\
&= \arg\max_{\xi} \int \alpha_n^{d/2}\pi_n\left(\sqrt{\alpha_n}t + \theta_{\text{MAP},n}\right) \exp\left(-\frac{1}{2}\|\xi - t\|^2\right) \mathrm{d}t.
\end{aligned}$$

Define

$$\begin{aligned}
f_n(\xi) &= \int \alpha_n^{d/2}\pi_n\left(\sqrt{\alpha_n}t + \theta_{\text{MAP},n}\right) \exp\left(-\frac{1}{2}\|\xi - t\|^2\right) \mathrm{d}t, \\
g_n(\xi) &= \int \phi\left(t; 0, \frac{1}{n\alpha_n}H_0^{-1}\right) \exp\left(-\frac{1}{2}\|\xi - t\|^2\right) \mathrm{d}t, \\
f(\xi) &= (2\pi)^{d/2}\phi\left(\xi; 0, I\right),
\end{aligned}$$

where $\phi(\cdot; \mu, \Sigma)$ denotes the PDF of $\mathcal{N}(\mu, \Sigma)$.

By adding and subtracting $f(\xi)$,

$$\begin{aligned}
\xi_n^* &= \arg\max_{\xi} f_n(\xi) \\
&= \arg\max_{\xi} \left\{f_n(\xi) - f(\xi) + f(\xi)\right\}.
\end{aligned}$$

We then apply Lemma A.1.2 to show $\xi_n^* \xrightarrow{d} \arg\max_{\xi} f(\xi)$. We start by verifying a condition of the argmax continuous mapping theorem that

$$\lim_{n\to\infty} \sup_{\xi} |f_n(\xi) - f(\xi)| = 0. \tag{A.13}$$

By triangle inequality, for all n,

$$\sup_{\xi} |f_n(\xi) - f(\xi)| \leq \sup_{\xi} |f_n(\xi) - g_n(\xi)| + \sup_{\xi} |g_n(\xi) - f(\xi)|. \tag{A.14}$$

Later we show both two terms on the RHS converging to 0.

For the first term. Note that $\alpha_n^{d/2}\pi_n(\sqrt{\alpha_n}t + \theta_{\text{MAP},n})$ is the probability density

function of $\Pi_{\sqrt{\alpha_n}t+\theta_{\text{MAP},n}}$, which is the posterior distribution parameterized on t. Thus, for all n,

$$
\begin{aligned}
&\sup_{\xi} |f_n(\xi) - g_n(\xi)| \\
&= \sup_{\xi} \left\{ \int \left| \alpha_n^{d/2} \pi_n(\sqrt{\alpha_n}t+\theta_{\text{MAP},n}) - \phi(t; 0, \frac{1}{n\alpha_n} H_0^{-1}) \right| \exp\left(-\frac{1}{2}\|\xi - t\|^2\right) \mathrm{d}t \right\} \\
&\leq \mathrm{D}_{\text{TV}} \left(\Pi_{\sqrt{\alpha_n}t+\theta_{\text{MAP},n}}, \mathcal{N}\left(0, \frac{1}{n\alpha_n} H_0^{-1}\right) \right),
\end{aligned}
$$

where the inequality is by $\sup_{\xi,t} \exp(-\frac{1}{2}\|\xi - t\|^2) \leq 1$. Under Assumption 1, the posterior distribution admits Bernstein-von Mises theorem (Theorem 3.1.1) that

$$
\mathrm{D}_{\text{TV}} \left(\Pi_n, \mathcal{N}\left(0, \frac{1}{n} H_0^{-1}\right) \right) = o_{P_{\theta_0}}(1).
$$

With the invariance of total variation under reparametrization, we have

$$
\mathrm{D}_{\text{TV}} \left(\Pi_{\sqrt{\alpha_n}t+\theta_{\text{MAP},n}}, \mathcal{N}\left(0, \frac{1}{n\alpha_n} H_0^{-1}\right) \right) = o_{P_{\theta_0}}(1).
$$

This shows the uniform convergence from $f_n(\xi)$ to $g_n(\xi)$.

For the second term in Eq. (A.14). Note that we can evaluate $g_n(\xi)$ since it is a convolution of two Gaussian PDFs, that is

$$
g_n(\xi) = (2\pi)^{d/2} \phi\left(\xi; 0, \frac{1}{n\alpha_n} H_0^{-1} + I\right).
$$

Comparing this to $f(\xi) = (2\pi)^{d/2}\phi(\xi; 0, I)$, one notices that $\frac{1}{n\alpha_n} H_{\theta_0}^{-1} + I \to I$ due to $\alpha_n^3 n \to \infty$. And further for Gaussian distributions, the convergence of parameters implies the uniform convergence of PDFs, yielding that

$$
\lim_{n\to\infty} \sup_{\xi} |g_n(\xi) - f(\xi)| = 0.
$$

Thus, we have Eq. (A.14) converging to 0 as $n \to \infty$.

Now we look at $f(\xi)$ with the goal to apply Lemma A.1.2 and to obtain $\xi_n^* \xrightarrow{d}$

$\arg\max_\xi f(\xi)$. Note that

$$\arg\max_{\xi} f(\xi) = 0 \quad \text{and} \quad \sup_{\xi} f(\xi) = \det(I)^{-1/2} = 1.$$

To apply Lemma A.1.2, we need to ensure that for any open set G that contains 0,

$$f(0) > \sup_{\xi \in G} f(\xi). \tag{A.15}$$

This holds by the unimodality of standard Gaussian distirbution.

Therefore, with both conditioins Eq. (A.13) and Eq. (A.15), we can apply Lemma A.1.2 to conclude that

$$\frac{1}{\sqrt{\alpha_n}} \left(\hat{\theta}_n - \theta_{\text{MAP},n} \right) \overset{P_{\theta_0}}{\to} 0.$$

This completes the proof. □

A.2 Proof for Theorem 4.2.1

Proof of Theorem 4.2.1. In this proof, we aim to apply Theorem A.2.1 with

$$\mathcal{X} = \{\mu \in \mathbb{R}^p, L \text{ lower triangular with non-negative diagonals}\},$$

which is closed and convex. Note that in the notation of this theorem, $x = (\mu^T, L_1^T, \ldots, L_p^T)^T \in \mathbb{R}^{(d+1)d}$ and $V \in \mathbb{R}^{(d+1)d \times (d+1)d}$ is set to be a diagonal matrix with entries 2 for the μ components and $r/(2\|I - L_n^\star\|_F)$ for the L components. Therefore

$$J(x) = J(\mu, L) = 4\|\mu - \mu_n^\star\|^2 + \frac{r^2}{4\|I - L_n^\star\|_F^2}\|L - L_n^\star\|_F^2.$$

This setting yields two important facts. First, by Theorems 4.1.1 and 3.2.1,

$$\hat{\theta}_n \overset{P_{\theta_0}}{\to} \theta_0 \quad \text{and} \quad \mu_n^\star \overset{P_{\theta_0}}{\to} \theta_0,$$

yielding that

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| + \|\mu_n^\star - \theta_0\| \leq \frac{r}{4\sqrt{2}}\right) \to 1, \quad \text{as } n \to \infty.$$

For $\|\mu_0 - \hat{\theta}_n\|^2 \leq \frac{r^2}{32}$, by triangle inequality, the probability that the following inequalities hold converges to 1 in P_{θ_0} as $n \to \infty$,

$$\|\mu_0 - \mu_n^\star\| \leq \|\mu_0 - \hat{\theta}_n\| + \|\hat{\theta}_n - \theta_0\| + \|\mu_n^\star - \theta_0\| \leq \frac{r}{2\sqrt{2}}.$$

Further with $L_0 = I$, $J(\mu_0, L_0) \leq \frac{3r^2}{4} \leq r^2$. Hence, if we initialize $L_0 = I$ and μ_0 such that $\|\mu_0 - \hat{\theta}_n\|_2^2 \leq \frac{r^2}{32}$,

$$\mathbb{P}\left(x_0 \in \{x : J(x) \leq r^2\}\right) \to 1, \quad \text{as } n \to \infty. \tag{A.16}$$

Second, if $J \le r^2$ then μ is close to the optimal and $\|L\|_F$ is not too large, i.e.,

$$\begin{aligned} J(\mu, L) \le r^2 &\implies \|\mu - \mu_n^\star\|^2 \le r^2/4 \\ J(\mu, L) \le r^2 &\implies \|L - L_n^\star\|_F^2 \le 4\|I - L_n^\star\|_F^2 \\ &\implies \|L\|_F \le 2\|I - L_n^\star\|_F + \|L_n^\star\|_F, \end{aligned}$$

yielding that $\{J(\mu, L) \le r^2\} \subseteq \mathcal{B}_{r,n}$. Recall that

$$\mathcal{B}_{r,n} = \left\{ \mu \in \mathbb{R}^d, L \in \mathbb{R}^{d\times d} : \|\mu - \mu_n^\star\|^2 \le \frac{r^2}{4} \text{and} \|L - L_n^\star\|_F^2 \le 4\|I - L_n^\star\|_F^2 \right\}.$$

Then by Corollary 3.3.8, under Assumptions 1 and 3, the probability of the event that

$$\begin{gathered} F_n \text{ is } \frac{\epsilon}{2} D_n\text{-strongly convex in } \{J(\mu, L) \le r^2\} \\ \text{and globally } \ell D_n\text{-Lipschitz smooth} \end{gathered} \tag{A.17}$$

converges to 1 in P_{θ_0} as $n \to \infty$.

For brevity, we make the following definitions for the rest of this proof: recall the definition of f_n, F_n in Eq. (3.2) (we state here again):

$$\begin{aligned} &I_n : \mathcal{X} \to \mathbb{R}, \qquad I_n(x) := -\frac{1}{n} \log \det L \\ &f_n : \mathbb{R}^d \to \mathbb{R}, \qquad f_n(\theta) := -\frac{1}{n} \log \pi_n(\theta) \\ &\tilde{f}_n : (\mathcal{X}, \mathbb{R}^d) \to \mathbb{R}, \qquad \tilde{f}_n(x, Z) := -\frac{1}{n} \log \pi_n \left(\mu + \frac{1}{\sqrt{n}} LZ \right) \\ &F_n : \mathcal{X} \to \mathbb{R}, \qquad F_n(x) := \mathbb{E}\left[-\frac{1}{n} \log \pi_n \left(\mu + \frac{1}{\sqrt{n}} LZ \right) \right] \\ &\phi_n := I_n + \tilde{f}_n, \qquad \Phi_n := I_n + F_n. \end{aligned}$$

Here $\phi_n(x, z)$ is the KL cost function with no expectation, and $\Phi_n(x)$ is the cost function with the expectation. To match the notation of Theorem A.2.1, we refor-

mulate the scaled gradient estimator defined in Eq. (4.3) as g_n,

$$g_n(x, Z) = \begin{cases} R_n(x)\nabla\phi_n(x, Z) & x \in \mathcal{X}^{\mathrm{o}} \\ \lim_{y\to x} R_n(y)\nabla\phi_n(y, Z) & x \in \partial\mathcal{X} \end{cases},$$

for a diagonal scaling matrix $R_n(x) \in \mathbb{R}^{d(d+1)\times d(d+1)}$. Define that $R_n(x)$ has entries 1 for the μ components, 1 for the off-diagonal L components, and $1/(1 + (nL_{ii})^{-1})$ for the diagonal L components. Note that $x \to \partial\mathcal{X}$ means that $L_{ii} \to 0$, ensuring that $g_n(x, Z)$ has entries -1 for the L_{ii}. Since Z is a standard normal random variable, under the event that $-\frac{1}{n}\log\pi_n$ has Lipschitz gradient, the gradient can be passed through the expectation so that the true gradient is defined as below,

$$G_n(x) := \mathbb{E}\left[g_n(x, Z)\right] = R_n(x)\nabla\Phi_n(x).$$

Note that the projected stochastic iteration

$$x_{k+1} = \Pi_{\mathcal{X}}\left(x_k - \gamma_k g_n(x_k, Z_k)\right), \quad k = \mathbb{N}\cup\{0\},$$

with $\Pi_{\mathcal{X}}(x) := \arg\min_{y\in\mathcal{X}} \|V(x - y)\|^2$ is equivalent to the iteration described in Algorithm 2. Note that the differentiability of ϕ_n only holds for $x \in \mathcal{X}^{\mathrm{o}}$. For the case where $L_{ii} = 0$ for some $i \in [d]$, we can use continuation via the limit $\lim_{L_{ii}\to 0} -\frac{(nL_{ii})^{-1}}{1+(nL_{ii})^{-1}} = -1$ to evaluate even though the gradient is not defined. For the following proof, we do not make special treatments to those boundary points when applying Taylor expansion and taking derivative.

Next we apply Theorem A.2.1 to carry out the proof. The rest of the proof consists two parts: **to show the confinement result (statement 2. of Theorem A.2.1)** and **to show the convergence result (statement 3. of Theorem A.2.1)**). We prove these two results under the event that Eqs. (A.16) and (A.17) hold; since the probability that these events hold converges in P_{θ_0} to 1 as $n \to \infty$, the final result holds with the same convergent probability.

We first show the confinement result by analyzing $\epsilon(x)$, $\ell^2(x)$, and $\sigma^2(r)$, which are defined in Eqs. (A.23) and (A.24) respectively. We aim to obtain that

i. We can find sufficiently small $\gamma_k > 0$ such that

$$\alpha_k = 1 + \mathbb{1}\left[J(x_k) \leq r^2\right](-2\gamma_k\epsilon(r) + 2\gamma_k^2\ell^2(r)) \in (0, 1]$$

holds for all $x \in \mathcal{X}$, i.e,

$$\forall x \in \mathcal{X} : J(x) \leq r^2, \quad 0 \leq 2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \leq 1. \tag{A.18}$$

ii. $\sigma^2(r) \to 0$ as $n \to \infty$ to guarantee the SGD iterations are eventually locally confined as $n \to \infty$ (based on Theorem A.2.1).

To show the statement i., Eq. (A.18), we start by deriving a lower bound for $2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x)$. Examine the expression,

$$\begin{aligned}
&2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \\
&= 2\gamma_k J(x)^{-1}(x - x^\star)^T V^T V R_n(x)\left(\nabla\Phi_n(x) - \nabla\Phi_n(x^\star)\right) \\
&-2\gamma_k^2 J(x)^{-1}\left(\nabla\Phi_n(x) - \nabla\Phi_n(x^\star)\right)^T R^T(x) V^T V R_n(x)\left(\nabla\Phi_n(x) - \nabla\Phi_n(x^\star)\right) \\
&= \frac{2\gamma_k}{J(x)}\left(V(x - x^\star)\right)^T V R_n(x)\left(\int\cdots\right) V^{-1}\left(V(x - x^\star)\right) \\
&- \frac{2\gamma_k^2}{J(x)}(V(x - x^\star))^T V^{-T}\left(\int\cdots\right)^T (V R_n(x))^2\left(\int\cdots\right) D^{-1}\left(V(x - x^\star)\right),
\end{aligned}$$

where $\left(\int\cdots\right) = \left(\int_0^1 \nabla^2\Phi_n((1-t)x^\star + tx)\mathrm{d}t\right)$. By splitting Φ_n into the regularization $I_n(x)$ and the expectation $F_n(x)$; and defining

$$\begin{aligned}
A(x) &:= V R_n(x)\left(\int_0^1 \nabla^2 I_n((1-t)x^\star + tx)\mathrm{d}t\right) V^{-1} \\
B(x) &:= V R_n(x)\left(\int_0^1 \nabla^2 F_n((1-t)x^\star + tx)\mathrm{d}t\right) V^{-1} \\
v(x) &:= V(x - x^\star),
\end{aligned}$$

the above expression can be written as

$$
\begin{aligned}
&2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \\
&= 2\gamma_k \frac{v(x)^T(A(x)+B(x))v(x) - \gamma_k\|(A(x)+B(x))v(x)\|^2}{\|v(x)\|^2} \\
&\geq 2\gamma_k \frac{v(x)^TA(x)v(x)+v(x)^TB(x)v(x)-2\gamma_k\|A(x)v(x)\|^2-2\gamma_k\|B(x)v(x)\|^2}{\|v(x)\|^2} \\
&\geq 2\gamma_k \left\{ \frac{v(x)^TA(x)v(x) + v(x)^TB(x)v(x) - 2\gamma_k\|A(x)v(x)\|^2}{\|v(x)\|^2} \right. \\
&\qquad \left. \frac{-2\gamma_k\|B(x)(VR_n(x)\ell D_nV^{-1})^{-1}\|^2\|VR_n(x)\ell D_nV^{-1}v(x)\|^2}{\|v(x)\|^2} \right\} \\
&= 2\gamma_k \left\{ \frac{v(x)^TA(x)v(x) + v(x)^T(B(x) - VR_n(x)\frac{\epsilon}{2}D_nV^{-1})v(x)}{\|v(x)\|^2} \right. \\
&\qquad + \frac{v(x)^T\left(VR_n(x)\frac{\epsilon}{2}D_nV^{-1}\right)v(x) - 2\gamma_k\|A(x)v(x)\|^2}{\|v(x)\|^2} \\
&\qquad \left. - \frac{2\gamma_k\|B(x)(VR_n(x)\ell D_nV^{-1})^{-1}\|^2\|DR_n(x)\ell D_nV^{-1}v(x)\|^2}{\|v(x)\|^2} \right\}
\end{aligned}
$$

Note that by Corollary 3.3.8 that $\frac{\epsilon}{2}D_n \preceq \nabla^2 F_n(x) \preceq \ell D_n$ and all the V, $R_n(x)$ are positive diagonal matrices, leading to

$$
\begin{aligned}
B(x) - VR_n(x)\frac{\epsilon}{2}D_nV^{-1} &\succeq 0I \\
\|B(x)(VR_n(x)\ell D_nV^{-1})^{-1}\|^2 &\leq 1.
\end{aligned}
$$

Thus, the above expression can be bounded below by

$$
\begin{aligned}
&2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \\
&\geq 2\gamma_k \left\{ \frac{v(x)^T A(x)v(x) + v(x)^T \left(VR_n(x)\frac{\epsilon}{2}D_nV^{-1}\right)v(x)}{\|v(x)\|^2} \right. \\
&\qquad \left. \frac{-2\gamma_k\|A(x)v(x)\|^2 - 2\gamma_k\|VR_n(x)\ell D_nV^{-1}v(x)\|^2}{\|v(x)\|^2} \right\} \\
&= \frac{2}{\|v(x)\|^2} v(x)^T \left\{ \left[\gamma_k A(x) - 2\gamma_k^2 A^2(x)\right] \right. \\
&\qquad \left. + \frac{1}{2}\left[\epsilon\gamma_k R_n(x)D_n - 4\ell^2\left(\gamma_k R_n(x)D_n\right)^2\right] \right\} v(x) \\
&\geq 2\lambda_{\min}\left(\left[\gamma_k A(x) - 2\gamma_k^2 A^2(x)\right] + \frac{1}{2}\left[\epsilon\gamma_k R_n(x)D_n - 4\ell^2\left(\gamma_k R_n(x)D_n\right)^2\right] \right).
\end{aligned}
$$

Now, notice that $A(x)$,$R_n(x)D_n$ are all diagonal matrices with non-negative entries,

$$
\begin{aligned}
\gamma_k A(x) - 2\gamma_k^2 A^2(x) &= \gamma_k A(x)\left(I - 2\gamma_k A(x)\right) \qquad \text{(A.19)} \\
\epsilon\gamma_k R_n(x)D_n - 4\ell^2\left(\gamma_k R_n(x)D_n\right)^2 &= \gamma_k R_n(x)D_n\left(\epsilon - 4\ell^2\gamma_k R_n(x)D_n\right).
\end{aligned}
$$

As long as the entries of $A(x)$, $R_n(x)D_n$ are bounded above by a constant for all n, there exists a sufficiently small constant c such that for all $\gamma_k < c$, Eq. (A.19) are both non-negative. Given that for all n and $\forall x \in \mathcal{X}$,

$$
\begin{aligned}
&A(x) = \operatorname{diag}\left(0, \cdots, \frac{(nL_{ii})^{-1}}{1 + (nL_{ii})^{-1}}\frac{1}{L_{ii}^\star}, \cdots, 0\right) \preceq \frac{1}{\min_{i\in[d]} L_{ii}^\star} I \\
&R_n(x)D_n \preceq I,
\end{aligned}
$$

we obtain the boundedness of the entries of $A(x)$, $R_n(x)D_n$. Therefore, we conclude that

$$
\forall x \in \mathcal{X}, \quad 0 \leq 2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x).
$$

It remains to show the second inequality of Eq. (A.18), i.e.,

$$\sup_{x\in\mathcal{X}:J(x)\le r^2} 2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \le 1.$$

This is true if

$$\sup_{x\in\mathcal{X}:J(x)\le r^2} \epsilon(x) \le \gamma_k^{-1}.$$

Since $\gamma_k \to 0$ as $k \to \infty$, the above holds if $\sup_{x\in\mathcal{X}:J(x)\le r^2} \epsilon(x)$ is bounded above by a constant that is independent to n. Now we consider the upper bound for $\sup_{x\in\mathcal{X}:J(x)\le r^2} \epsilon(x)$. Expanding $\epsilon(x)$,

$$\begin{aligned}
\epsilon(x) &= J(x)^{-1}(x-x^\star)^T D^T D R_n(x)\left(\nabla\Phi_n(x) - \nabla\Phi_n(x^\star)\right) \\
&= \frac{v(x)^T(A(x)+B(x))v(x)}{\|v(x)\|^2} \\
&\le \lambda_{\max}(A(x)+B(x)) \\
&= \lambda_{\max} R_n(x)^{1/2}\left(\int_0^1 \nabla^2\Phi_n((1-t)x^\star + tx)\mathrm{d}t\right) R_n(x)^{1/2}.
\end{aligned}$$

Split Φ_n into the regularization $I_n(x)$ and the expectation $F_n(x)$. For the expectation, by Corollary 3.3.8 that $\nabla^2 F_n(x) \preceq \ell D_n$ and entries of $R_n(x)$ are bounded by 1, we have

$$R_n(x)^{1/2}\nabla^2 F_n(x) R_n(x)^{1/2} \preceq \ell I,$$

and for the regularization, note that $\nabla^2 I_n$ is a diagonal matrix with 0 for μ and off-diagonals of L and L_{ii}^{-2}/n for diagonals of L, so

$$\begin{aligned}
&R_n(x)^{1/2}\left(\int_0^1 \nabla^2 I_n((1-t)x^\star + tx)\mathrm{d}t\right) R_n(x)^{1/2} \\
&= \mathrm{diag}\left(0, \cdots, \frac{(nL_{ii})^{-1}}{1+(nL_{ii})^{-1}}\frac{1}{L_{ii}^\star}, \cdots, 0\right) \\
&\preceq \frac{1}{\min_{i\in[d]} L_{ii}^\star} I
\end{aligned} \tag{A.20}$$

By the fact that $\forall i \in [d], L^{\star}_{ii} > 0$, we have Eq. (A.20) is bounded above by a constant C. Use the Weyl's inequality to bound the maximal eigenvalue of the summation of two Hermitian matrices, we conclude that

$$\sup_{x \in \mathcal{X}: J(x) \leq r^2} \epsilon(x) \leq \ell + C.$$

Therefore, we have completed the proof for statement i., Eq. (A.18).

Then we show the statement ii. by getting upper bound on $\sigma^2(r)$. Recall that $\sigma^2(r)$ is the upper bound of the fourth moment of

$$\left\| VR_n(x) \left(\nabla \phi_n(x, Z) - \nabla \Phi_n(x) \right) \right\|.$$

Since the regularizor is cancelled in this expression, we only consider the expectation part. Note that $VR_n(x)$ is a diagonal matrix with positive diagonals,

$$\begin{aligned}
&\mathbb{E}\left[\left\| VR_n(x) \left(\nabla \tilde{f}_n(x, Z) - \nabla F_n(x) \right) \right\|^4 \right]^{1/4} \\
&\leq \max_{i \in [d(d+1)]} (VR_n(x))_{ii} \mathbb{E}\left[\left\| \nabla \tilde{f}_n(x, Z) - \nabla F_n(x) \right\|^4 \right]^{1/4}.
\end{aligned}$$

Let Z_1, Z_2 be independent copies, by tower property of conditional expectation,

$$= \max_{i \in [d(d+1)]} (VR_n(x))_{ii} \mathbb{E}\left[\left\| \mathbb{E}\left[\nabla \tilde{f}_n(x, Z_1) - \nabla \tilde{f}_n(x, Z_2) | Z_1 \right] \right\|^4 \right]^{1/4}.$$

By the convexity of $\| \cdot \|^4$ and Jensen's inequality,

$$\begin{aligned}
&\leq \max_{i \in [d(d+1)]} (VR_n(x))_{ii} \mathbb{E}\left[\mathbb{E}\left[\left\| \nabla \tilde{f}_n(x, Z_1) - \nabla \tilde{f}_n(x, Z_2) \right\|^4 | Z_1 \right] \right]^{1/4} \\
&= \max_{i \in [d(d+1)]} (VR_n(x))_{ii} \mathbb{E}\left[\left\| \nabla \tilde{f}_n(x, Z_1) - \nabla \tilde{f}_n(x, Z_2) \right\|^4 \right]^{1/4}.
\end{aligned}$$

By $\left\| \nabla \tilde{f}_n(x, Z_1) - \nabla \tilde{f}_n(x, Z_2) \right\| \leq \left\| \nabla \tilde{f}_n(x, Z_1) \right\| + \left\| \nabla \tilde{f}_n(x, Z_2) \right\|$ and Minkowski's

inequality,

$$
\begin{aligned}
&\leq \max_{i\in[d(d+1)]}(VR_n(x))_{ii}\left\{\mathbb{E}\left[\|\nabla\tilde{f}_n(x,Z_1)\|^4\right]^{1/4}+\mathbb{E}\left[\|\nabla\tilde{f}_n(x,Z_1)\|^4\right]^{1/4}\right\}\\
&=2\max_{i\in[d(d+1)]}(VR_n(x))_{ii}\mathbb{E}\left[\|\nabla\tilde{f}_n(x,Z)\|^4\right]^{1/4}.
\end{aligned}
$$

Now we focus on bounding $\mathbb{E}\left[\|\nabla\tilde{f}_n(x,Z)\|^4\right]^{1/4}$. We examine $\|\nabla\tilde{f}_n(x,Z)\|$,

$$
\nabla\tilde{f}_n(x,Z)=\begin{pmatrix}\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\\ \frac{Z_1}{\sqrt{n}}\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\\ \vdots\\ \frac{Z_p}{\sqrt{n}}\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\end{pmatrix}\quad\in\mathbb{R}^{d(d+1)},
$$

yielding

$$
\|\nabla\tilde{f}_n(x,Z)\|^4=\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\right\|^4\left(1+\frac{Z^TZ}{n}\right)^2.
$$

By Cauchy-Schiwartz inequality,

$$
\mathbb{E}\left[\|\nabla\tilde{f}_n(x,Z)\|^4\right]^{1/4}\leq\mathbb{E}\left[\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\right\|^8\right]^{1/8}\mathbb{E}\left[\left(1+\frac{Z^TZ}{n}\right)^4\right]^{1/8}.
$$

We then bounds these two terms on RHS separately. We use the sub-Gaussian property of $\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\right\|$ to bound its 8^{th} moment. First notice that $\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ\right)\right\|$ is a $\max_{i\in[d]}L_{ii}\frac{\ell}{\sqrt{n}}$-Lipschitz function of Z,

$$
\begin{aligned}
&\left|\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ_1\right)\right\|-\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ_2\right)\right\|\right|\\
&\leq\left\|\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ_1\right)-\nabla f_n\left(\mu+\frac{1}{\sqrt{n}}LZ_2\right)\right\|\\
&=\left\|\int_0^1\nabla^2 f_n\left(\mu+(1-t)LZ_2/\sqrt{n}+tZ_1/\sqrt{n}\right)\mathrm{d}t\frac{L}{\sqrt{n}}(Z_1-Z_2)\right\|.
\end{aligned}
$$

Given that $\nabla^2 f_n \preceq \ell I$, the above is bounded by

$$\begin{aligned}\frac{\ell}{\sqrt{n}}\sqrt{(Z_1-Z_2)^T L^T L (Z_1-Z_2)} &\leq \frac{\ell}{\sqrt{n}}\lambda_{\max}(L^T L)\|Z_1-Z_2\| \\ &= \frac{\ell}{\sqrt{n}}\max_{i\in[d]} L_{ii}^2 \|Z_1-Z_2\|.\end{aligned}$$

Since a Lipschitz function of Gaussian noise is sub-Gaussian (Kontorovich, 2014, Thm 1), i.e., let $Z \sim \mathcal{N}(0, I_d)$, $\psi : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz, then

$$\mathbb{P}\left(|\psi(Z) - \mathbb{E}[\psi(Z)]| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{4L^2}\right).$$

Thus, $\left\|\nabla f_n\left(\mu + \frac{1}{\sqrt{n}}LZ\right)\right\|$ is $\frac{4\ell^2}{n}\max_{i\in[d]} L_{ii}^2$-sub-Gaussian. Then note that for a σ^2-sub-Gaussian random variable $X \in \mathbb{R}$, for any positive integer $k \geq 2$, $\mathbb{E}\left[|X|^k\right]^{1/k} \leq \sigma e^{1/e}\sqrt{k}$. Hence we obtain

$$\mathbb{E}\left[\left\|\nabla f_n\left(\mu + \frac{1}{\sqrt{n}}LZ\right)\right\|^8\right]^{1/8} \leq \frac{2\ell}{\sqrt{n}} e^{1/e}\sqrt{8}\max_{i\in[d]} L_{ii}.$$

Along with the fact that Gaussian random variable has arbitrary order moments,

$$\mathbb{E}\left[\left(1 + \frac{Z^T Z}{n}\right)^4\right] \leq C,$$

for some constant C, we obtain

$$\mathbb{E}\left[\|\nabla_x f_n\|^4\right]^{1/4} \leq \frac{2C^{1/4}\ell}{\sqrt{n}} e^{1/e}\sqrt{8}\max_{i\in[d]} L_{ii},$$

and hence

$$\begin{aligned}&\mathbb{E}\left[\|VR_n(x)\left(\nabla_x f_n - \nabla_x F_n\right)\|^4\right]^{1/4} \\ &\quad\leq \max_{i\in[d(d+1)]}(VR_n(x))_{ii}\frac{2C^{1/4}\ell}{\sqrt{n}} e^{1/e}\sqrt{8}\max_{i\in[d]} L_{ii}.\end{aligned}$$

Taking supremum over $J(x) \leq r^2$, the RHS is bounded above by a universal

constant, we therefore conclude that

$$\sigma^2(r) = \sup_{x\in\mathcal{X}:J(x)\leq r^2} \mathbb{E}\left[\left\|VR_n(x)\left(\nabla\phi_n(x,Z) - \nabla\Phi_n(x)\right)\right\|^4\right]^{1/4} \to 0,\ n\to\infty.$$

Therefore, with $\forall x \in \mathcal{X} : J(x) \leq r^2, 0 \leq 2\gamma_k\epsilon(x) - 2\gamma_k^2\ell^2(x) \leq 1$ and $\sigma^2(r) \to 0$ as $n \to \infty$, applying Theorem A.2.1 yields the confinement result, i.e.,

$$\mathbb{P}\left(\sup_{k\in\mathbb{N}} J(x_k) \leq r^2\right) \to 1$$

in P_{θ_0} as $n \to \infty$.

Lastly, by statement 3. of Theorem A.2.1, we prove the convergence result by checking

$$\inf_{x\in\mathcal{X},J(x)\leq r^2} \epsilon(x) > 0.$$

We use the similar way to expand the expression,

$$\begin{aligned}
\epsilon(x) &= J(x)^{-1}(x-x^\star)^T V^T V R_n(x)\left(\nabla\Phi_n(x) - \nabla\Phi_n(x^\star)\right) \\
&= \frac{v(x)^T(A(x)+B(x))v(x)}{\|v(x)\|^2} \\
&\geq \lambda_{\min}(A(x)+B(x)) \\
&= \lambda_{\min} R_n(x)^{1/2}\left(\int_0^1 \nabla^2\Phi_n((1-t)x^\star + tx)\mathrm{d}t\right) R_n(x)^{1/2}.
\end{aligned}$$

By splitting Φ_n into the regularization and the expectation, we have

$$\begin{aligned}
&R_n^{1/2}(x)\left(\int_0^1 \nabla^2 I_n((1-t)x^\star + tx)\mathrm{d}t\right) R_n(x)^{1/2} \\
&= \mathrm{diag}\left(0,\cdots,\frac{(nL_{ii})^{-1}}{1+(nL_{ii})^{-1}}\frac{1}{L_{ii}^\star},\cdots,0\right) \\
&\succeq 0I,
\end{aligned} \tag{A.21}$$

and

$$\begin{aligned} & R_n(x)^{1/2}\left(\int_0^1 \nabla^2 F_n((1-t)x^\star + tx)\mathrm{d}t\right) R_n(x)^{1/2} \\ & \geq R_n(x)^{1/2}\frac{D_n\epsilon}{2}R_n(x)^{1/2} \\ & \succeq \epsilon/2n > 0. \end{aligned} \tag{A.22}$$

We then combine Eqs. (A.21) and (A.22) and use Weyl's inequality to bound the minimal eigenvalue of the summation of two Hermitian matrices, yielding

$$\inf_{x\in\mathcal{X}, J(x)\leq r^2} \epsilon(x) > \epsilon/n > 0.$$

This gives the convergence result.

Then the proof is complete by applying Theorem A.2.1. We know that ξ_k is strictly positive. Since $\epsilon(r) > 0$ and $\ell(r)$ is bounded above, there exists $\gamma_k = \Theta(k^{-\rho}), \rho \in (0.5, 1)$ so that it satisfies the condition of the theorem. We have that $\sigma \to 0$, which makes 3. in the statement of Theorem A.2.1 become

$$\mathbb{P}\left(\limsup_{k\to\infty} \|V(x_k - x^\star)\|^2 = 0\right) \xrightarrow{p} 1, \quad n \to \infty$$

Even though D is a function of n, n is fixed as Algorithm 2 runs. Since D is invertible,

$$\mathbb{P}\left(\limsup_{k\to\infty} \|x_k - x^\star\|^2 = 0\right) \xrightarrow{P_{\theta_0}} 1, \quad n \to \infty$$

which is exactly our desired result: as the number of data $n \to \infty$, the probability that Algorithm 2 finds the optimum (as we take more iterations, $k \to \infty$) converges to 1. In other words, *variational inference gets solved asymptotically*.

□

Theorem A.2.1. *Let $\mathcal{X} \subseteq \mathbb{R}^p$ be closed and convex, $g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^p$ be a function, $G(x) := \mathbb{E}[g(x, Z)]$ for a random element $Z \in \mathcal{Z}$, $x^\star \in \mathcal{X}$ be a point in $\mathcal{X}$ such that $G(x^\star) = 0$, $V \in \mathbb{R}^{p\times p}$ be invertible, $J(x) := \|V(x - x^\star)\|^2$, and $r \geq 0$.*

Consider the projected stochastic iteration

$$x_0 \in \mathcal{X}, \quad x_{k+1} = \Pi_{\mathcal{X}}\left(x_k - \gamma_k g(x_k, Z_k)\right), \quad k = \mathbb{N} \cup \{0\},$$

with independent copies Z_k of Z, $\gamma_k \geq 0$, and $\Pi_{\mathcal{X}}(x) := \arg\min_{y \in \mathcal{X}} \|V(x-y)\|^2$. If

1. *For all $k \in \mathbb{N} \cup \{0\}$, the step sizes satisfy*

$$\forall x \in \mathcal{X} : J(x) \leq r^2, \quad 0 \leq 2\gamma_k \epsilon(x) - 2\gamma_k^2 \ell^2(x) \leq 1 \tag{A.23}$$
$$\epsilon(x) := \frac{1}{J(x)}(x - x^\star)^T V^T V \left(G(x) - G(x^\star)\right)$$
$$\ell^2(x) := \frac{1}{J(x)} \left\|V\left(G(x) - G(x^\star)\right)\right\|^2,$$

2. *For all $x \in \mathcal{X}$, $\left(\mathbb{E}\|V(g(x,Z) - G(x))\|^4\right)^{1/4} \leq \tilde{\sigma}(x)$ for $\tilde{\sigma} : \mathcal{X} \to \mathbb{R}_{\geq 0}$, and*

$$\sigma(r) := \sup_{x \in \mathcal{X} \,:\, J(x) \leq r^2} \tilde{\sigma}(x), \tag{A.24}$$

then

1. *The iterate x_k is locally confined with high probability:*

$$\mathbb{P}\left(J(x_k) \leq r^2\right) \geq \frac{\xi_k^2}{\xi_k^2 + 8\sigma(r)^2 \zeta_k}$$
$$\xi_k(r) := \max\{0, r^2 - J(x_0) - 2\sigma^2(r) \sum_{j<k} \gamma_j^2\}$$
$$\zeta_k(r) := r^2 \sum_{j<k} \gamma_j^2 + \sigma^2(r) \sum_{j<k} \gamma_j^4.$$

2. *The iterate x_k stays locally confined for all $k \in \mathbb{N}$ with high probability:*

$$\mathbb{P}\left(\sup_{k \in \mathbb{N}} J(x_k) \leq r^2\right) \geq \frac{\xi^2}{\xi^2 + 8\sigma^2(r)\zeta}$$
$$\xi(r) := \lim_{k \to \infty} \xi_k(r) \quad \zeta(r) := \lim_{k \to \infty} \zeta_k(r).$$

3. *If additionally*

$$\inf_{x\in\mathcal{X}:J(x)\le r^2} \epsilon(x) > 0 \quad \textit{and} \quad \gamma_k = \Theta(k^{-\rho}),\ \rho \in (0.5, 1),$$

the iterate x_k converges to $x^\star$ with high probability:

$$\mathbb{P}\left(\limsup_{k\to\infty} J(x_k) = 0\right) \ge \mathbb{P}\left(\sup_{k\in\mathbb{N}} J(x_k) \le r^2\right).$$

Proof. To begin, we show $\Pi_{\mathcal{X}}$ is non-expansive,

$$\|V(\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(y))\|^2 \le \|V(x-y)\|^2.$$

For all $x, y \in \mathbb{R}^p$, define $\langle x, y\rangle_V = x^T V^T V y$. Since V is invertible, $V^T V$ is symmetric and positive definite, and hence $(\mathbb{R}^p, \langle\cdot,\cdot\rangle_V)$ forms a Hilbert space. Any projection operator of a Hilbert space is non-expansive (Bauschke and Combettes, 2011, Prop. 4.4).

Note that $x^\star = \Pi_{\mathcal{X}}(x^\star)$ and the projection operation is non-expansive, expanding the squared norm yields

$$\begin{aligned}\|V(x_{k+1} - x^\star)\|^2 \le \|V(x_k - x^\star)\|^2 \\ - 2\gamma_k (x_k - x^\star)^T V^T V g(x_k, Z_k) + \gamma_k^2 \|V g(x_k, Z_k)\|^2.\end{aligned}$$

Adding and subtracting $G(x_k)$ in the second and third terms, using the elementary bound $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and defining

$$\begin{aligned}\beta_k(x) &:= -2\gamma_k (x - x^\star)^T V^T V (g(x, Z_k) - G(x)) \\ &\quad + 2\gamma_k^2 \|V(g(x, Z_k) - G(x))\|^2 - 2\gamma_k^2 \mathbb{E}\left[\|\cdot\|^2\right] \\ \epsilon(x) &:= \frac{1}{J(x)} (x - x^\star)^T V^T V \left(G(x) - G(x^\star)\right) \\ \ell^2(x) &:= \frac{1}{J(x)} \|V\left(G(x) - G(x^\star)\right)\|_2^2,\end{aligned}$$

we have that

$$J(x_{k+1}) \leq J(x_k)\left(1 - 2\gamma_k\epsilon(x_k) + 2\gamma_k^2\ell^2(x_k)\right) + \beta_k(x_k) + 2\gamma_k^2\tilde{\sigma}^2(x_k).$$

We now define the filtration of σ-algebras

$$\mathcal{F}_k = \sigma(x_1, \ldots, x_k, Z_1, \ldots, Z_{k-1}),$$

and the stopped process for $r > 0$,

$$Y_0 = J(x_0)$$
$$Y_{k+1} = \begin{cases} Y_k & Y_k > r^2 \\ J(x_{k+1}) & \text{o.w.} \end{cases}$$

Note that Y_k is $\mathcal{F}_k$-measurable, and that Y_k "freezes in place" if $J(x_k)$ ever jumps larger than r^2; so for all $t^2 \leq r^2$,

$$\begin{aligned}
\mathbb{P}\left(J(x_k) > t^2\right) &= \mathbb{P}\left(J(x_k) > t^2, Y_{k-1} > r^2\right) + \mathbb{P}\left(J(x_k) > t^2, Y_{k-1} \leq r^2\right) \\
&= \mathbb{P}\left(J(x_k) > t^2, Y_k > r^2, Y_{k-1} > r^2\right) + \mathbb{P}\left(Y_k > t^2, Y_{k-1} \leq r^2\right) \\
&\leq \mathbb{P}\left(Y_k > r^2, Y_{k-1} > r^2\right) + \mathbb{P}\left(Y_k > t^2, Y_{k-1} \leq r^2\right) \\
&\leq \mathbb{P}\left(Y_k > t^2, Y_{k-1} > r^2\right) + \mathbb{P}\left(Y_k > t^2, Y_{k-1} \leq r^2\right) \\
&= \mathbb{P}\left(Y_k > t^2\right).
\end{aligned}$$

Therefore if we obtain a tail bound on Y_k, it provides the same bound on $J(x_k)$. Now substituting the stopped process into the original recursion and collecting terms,

$$\begin{aligned}
&Y_{k+1} \\
\leq &Y_k\left(1 + \mathbb{1}\left[Y_k \leq r^2\right](-2\gamma_k\epsilon(x_k) + \gamma_k^2\ell^2(x_k))\right) + \mathbb{1}\left[Y_k \leq r^2\right]\left(\beta_k(x_k) + 2\gamma_k^2\tilde{\sigma}^2(x_k)\right) \\
\leq &Y_k\left(1 + \mathbb{1}\left[Y_k \leq r^2\right](-2\gamma_k\epsilon(x_k) + \gamma_k^2\ell^2(x_k))\right) + \mathbb{1}\left[Y_k \leq r^2\right]\beta_k(x_k) + 2\gamma_k^2\sigma^2(r).
\end{aligned}$$

Using the notation of Lemma A.2.2, set

$$\begin{aligned}
\alpha_k &= 1 + \mathbb{1}\left[Y_k \le r^2\right](-2\gamma_k \epsilon(x_k) + 2\gamma_k^2 \ell^2(x_k)) \\
\beta_k &= \mathbb{1}\left[Y_k \le r^2\right]\beta_k(x_k) \\
c_k &= 2\gamma_k^2 \sigma^2(r).
\end{aligned}$$

By the fourth moment assumption, β_k has variance bounded above by τ_k^2 conditioned on $\mathcal{F}_k$, where

$$\begin{aligned}
\tau_k^2 &= 8\gamma_k^2 \mathbb{1}\left[Y_k \le r^2\right] \|V(x_k - x^\star)\|^2 \tilde{\sigma}(x_k)^2 + 8\gamma_k^4 \mathbb{1}\left[Y_k \le r^2\right] \tilde{\sigma}^4(x_k) \\
&\le 8\gamma_k^2 r^2 \sigma^2(r) + 8\gamma_k^4 \sigma^4(r).
\end{aligned}$$

Therefore, using the descent Lemma A.2.2 and the fact that $0 \le \alpha_k \le 1$,

$$\begin{aligned}
&\mathbb{P}\left(Y_k - Y_0 > t + 2\sigma^2(r)\sum_{j<k}\gamma_j^2\right) \\
&\le \frac{8\sigma^2(r)\left(r^2\sum_{j<k}\gamma_j^2 + \sigma^2(r)\sum_{j<k}\gamma_j^4\right)}{t^2 + 8\sigma^2(r)\left(r^2\sum_{j<k}\gamma_j^2 + \sigma^2(r)\sum_{j<k}\gamma_j^4\right)}.
\end{aligned}$$

Finally let ξ_k and ζ_k be defined as

$$\begin{aligned}
\xi_k &:= \max\{0, r^2 - Y_0 - 2\sigma^2(r)\sum_{j<k}\gamma_j^2\} \\
\zeta_k &:= r^2\sum_{j<k}\gamma_j^2 + \sigma^2(r)\sum_{j<k}\gamma_j^4,
\end{aligned}$$

yielding the first result,

$$\mathbb{P}\left(Y_k > r^2\right) \le \frac{8\sigma^2(r)\zeta_k}{\xi_k^2 + 8\sigma^2(r)\zeta_k}.$$

Now since $Y_{k+1} \le r^2 \implies Y_k \le r^2$ for all $k \ge 0$, the sequence of events

$\left\{Y_k \leq r^2\right\}$ is decreasing. Therefore the second result follows from

$$\begin{aligned}\mathbb{P}\left(\bigcap_{k=0}^{\infty}\left\{Y_k \leq r^2\right\}\right) &= \lim_{k\to\infty}\mathbb{P}\left(Y_k \leq r^2\right)\\ &\geq \lim_{k\to\infty} 1 - \frac{8\sigma^2(r)\zeta_k}{\xi_k^2 + 8\sigma^2(r)\zeta_k}\\ &= \frac{\xi^2}{\xi^2 + 8\sigma^2(r)\zeta},\end{aligned}$$

where $\xi := \lim_{k\to\infty}\xi_k$ and $\zeta := \lim_{k\to\infty}\zeta_k$ (or ∞ if ζ_k diverges). Finally, we analyze the conditional tail distribution of Y_k given that it stays confined, i.e., $\forall k \geq 0$, $Y_k \leq r^2$. First define

$$0 \leq a_k := \sup_{x\in\mathcal{X}:J(x)\leq r^2} 1 - 2\gamma_k\epsilon(x) + 2\gamma_k^2\ell^2(x) \leq 1,$$

i.e., a_k is the largest possible value of α_k when $Y_k \leq r^2$. So again applying Lemma A.2.2,

$$\begin{aligned}&\mathbb{P}\left(Y_k - Y_0\prod_{j=0}^{k-1}a_j > t\prod_{j=0}^{k-1}a_j + \sum_{j=0}^{k-1}c_j\prod_{m=j+1}^{k-1}a_m \,|\, \forall k\ Y_k \leq r^2\right)\\ &= \frac{\mathbb{P}\left(Y_k - Y_0\prod_{j=0}^{k-1}a_j > t\prod_{j=0}^{k-1}a_j + \sum_{j=0}^{k-1}c_j\prod_{m=j+1}^{k-1}a_m, \forall k\ Y_k \leq r^2\right)}{\mathbb{P}\left(\forall k\ Y_k \leq r^2\right)}\\ &= \frac{\mathbb{P}\left(Y_k - Y_0\prod_{j=0}^{k-1}\alpha_j > t\prod_{j=0}^{k-1}\alpha_j + \sum_{j=0}^{k-1}c_j\prod_{m=j+1}^{k-1}\alpha_m, \forall k\ Y_k \leq r^2\right)}{\mathbb{P}\left(\forall k\ Y_k \leq r^2\right)}\\ &\leq \frac{\mathbb{P}\left(Y_k - Y_0\prod_{j=0}^{k-1}\alpha_j > t\prod_{j=0}^{k-1}\alpha_j + \sum_{j=0}^{k-1}c_j\prod_{m=j+1}^{k-1}\alpha_m\right)}{\frac{\xi^2}{\xi^2+8\sigma^2(r)\zeta}}\\ &\leq \frac{\left(\frac{8\sigma^2(r)\zeta_k}{t^2+8\sigma^2(r)\zeta_k}\right)}{\left(\frac{\xi^2}{\xi^2+8\sigma^2(r)\zeta}\right)}.\end{aligned}$$

If we set $t = s\prod_{j=0}^{k-1} a_j^{-1}$ for any $s \geq 0$, this implies that

$$\begin{aligned}
&\mathbb{P}\left(Y_k > s + Y_0\prod_{j=1}^{k-1} a_j + \sum_{j=0}^{k-1} c_j \prod_{m=j+1}^{k-1} a_m \,|\, \forall k\ Y_k \leq r^2\right) \\
&\leq \frac{\left(\frac{8\sigma^2(r)\zeta_k}{s^2\prod_{j<k} a_j^{-2} + 8\sigma^2(r)\zeta_k}\right)}{\left(\frac{\xi^2}{\xi^2+8\sigma^2(r)\zeta}\right)}.
\end{aligned}$$

Now since $\gamma_k = \Theta(k^{-\rho})$, $\rho \in (0.5, 1)$, and $\inf_{x\in\mathcal{X}:J(x)\leq r^2} \epsilon(x) > 0$, we know that $\prod_{j<k} a_j^{-2} = \Theta\left(\exp\left(Ck^{\rho'}\right)\right)$ for some $\rho', C > 0$. So therefore for any $s \geq 0$,

$$\mathbb{P}\left(Y_k > s \,|\, \forall k\ Y_k \leq r^2\right) = O\left(\exp\left(-Ck^{\rho'}\right)\right).$$

By the Borel Cantelli lemma, we have that $\mathbb{P}\left(\lim_{k\to\infty} Y_k = 0 \,|\, \forall k\ Y_k \leq r^2\right) = 1$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\lim_{k\to\infty} Y_k = 0\right) &\geq \mathbb{P}\left(\lim_{k\to\infty} Y_k = 0 \,|\, \forall k\ Y_k \leq r^2\right)\mathbb{P}\left(\forall k\ Y_k \leq r^2\right) \\
&= \mathbb{P}\left(\forall k\ Y_k \leq r^2\right),
\end{aligned}$$

and the result follows. □

Lemma A.2.2 (Descent)**.** *Suppose we are given a filtration $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$, $k \geq 0$. Let*

$$Y_{k+1} \leq \alpha_k Y_k + \beta_k + c_k, \quad k \geq 0,$$

where $Y_k \geq 0$ and $0 \leq \alpha_k \leq 1$ are $\mathcal{F}_k$-measurable, β_k is $\mathcal{F}_{k+1}$-measurable and has mean 0 and variance conditioned on $\mathcal{F}_k$ bounded above by τ_k^2, and $\tau_k^2, c_k \geq 0$ are $\mathcal{F}_0$ measurable. Then

$$\mathbb{P}\left(Y_k - Y_0\prod_{i=0}^{k-1}\alpha_i \geq t\prod_{i=0}^{k-1}\alpha_i + \sum_{i=0}^{k-1} c_i \prod_{j=i+1}^{k-1}\alpha_j\right) \leq \frac{\sum_{i=1}^{k-1}\tau_i^2}{t^2 + \sum_{i=1}^{k-1}\tau_i^2}.$$

Proof. Solving the recursion,

$$\begin{aligned}
Y_k &\leq \alpha_{k-1} Y_{k-1} + \beta_{k-1} + c_{k-1} \\
&\leq \alpha_{k-1} \left(\alpha_{k-2} Y_{k-2} + \beta_{k-2} + c_{k-2} \right) + \beta_{k-1} + c_{k-1} \\
&\leq \dots \\
&\leq Y_0 \prod_{i=0}^{k-1} \alpha_i + \sum_{i=0}^{k-1} \beta_i \prod_{j=i+1}^{k-1} \alpha_j + \sum_{i=0}^{k-1} c_i \prod_{j=i+1}^{k-1} \alpha_j.
\end{aligned}$$

So

$$\begin{aligned}
&\mathbb{P}\left(Y_k - Y_0 \prod_{i=0}^{k-1} \alpha_i - \sum_{i=0}^{k-1} c_i \prod_{j=i+1}^{k-1} \alpha_j \geq t \prod_{i=0}^{k-1} \alpha_i \right) \\
&\leq \mathbb{P}\left(\sum_{i=0}^{k-1} \beta_i \prod_{j=i+1}^{k-1} \alpha_j \geq t \prod_{i=0}^{k-1} \alpha_i \right) \\
&= \mathbb{P}\left(\sum_{i=0}^{k-1} \beta_i \prod_{j=0}^{i} \alpha_j \geq t \right).
\end{aligned}$$

By Cantelli's inequality and the fact that the i^{th} term in the sum is $\mathcal{F}_i$-measurable,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=0}^{k-1} \beta_i \prod_{j=0}^{i} \alpha_j \geq t \right) &\leq \frac{\sum_{i=1}^{k-1} \mathbb{E}\left[\beta_i^2 \prod_{j=0}^{i} \alpha_j^2 \right]}{t^2 + \sum_{i=1}^{k-1} \mathbb{E}\left[\beta_i^2 \prod_{j=0}^{i} \alpha_j^2 \right]} \\
&\leq \frac{\sum_{i=1}^{k-1} \mathbb{E}\left[\beta_i^2 \right]}{t^2 + \sum_{i=1}^{k-1} \mathbb{E}\left[\beta_i^2 \right]} \\
&\leq \frac{\sum_{i=1}^{k-1} \tau_i^2}{t^2 + \sum_{i=1}^{k-1} \tau_i^2}.
\end{aligned}$$

□