A Diffusion Process Perspective on Posterior Contraction Rates for Parameters

Wenlong Mou*, Nhat Ho†, Martin Wainwright‡, Peter Bartlett§, and Michael Jordan§

Abstract. We analyze the posterior contraction rates of parameters in Bayesian models via the Langevin diffusion process, in particular by controlling moments of the stochastic process and taking limits. Analogous to the non-asymptotic analysis of statistical M-estimators and stochastic optimization algorithms, our contraction rates depend on the structure of the population log-likelihood function, and stochastic perturbation bounds between the population and sample log-likelihood functions. Convergence rates are determined by a non-linear equation that relates the population-level structure to stochastic perturbation terms, along with a term characterizing the diffusive behavior. Based on this technique, we also prove non-asymptotic versions of a Bernstein–von Mises guarantee for the posterior. We illustrate this general theory by deriving posterior convergence rates for various concrete examples.

Key words. Bayesian inference, Diffusion processes, Posterior contraction rate, Bernstein–von Mises theorem

AMS subject classifications. 62F15, 62F12

1. Introduction.

Bayesian inference is one of the central pillars of statistics. In Bayesian analysis, we first endow the parameter space with a prior distribution chosen by modelling considerations, and then apply Bayes' rule, combining the prior with the likelihood, so as to form the posterior distribution. From a statistical perspective, this posterior is of fundamental interest, and there are various questions associated with its behavior, including its consistency as the sample size goes to infinity, and from a more refined point of view, its contraction rate in various metrics.

The earliest work on posterior consistency dates back to the seminal work of Doob [9], who exhibited conditions under which the posterior distribution is consistent for all parameters apart from a set of zero measure. Subsequent work by Freedman [13, 14] provided examples showing that this null set can be problematic for Bayesian consistency in non-parametric settings. In order to address this issue, Schwartz [40] proposed a general framework for establishing posterior consistency for both semiparametric and nonparametric models. Since then, a number of researchers have isolated conditions that are useful for studying posterior distributions [3, 54, 55].

Moving beyond posterior consistency, convergence rates for the posterior density function, along with associated parameters of models, remains an active area of research. For posterior densities, Ghosal et al. [16] gave a general testing framework for proving convergence rates for both finite and infinite dimensional models; it has been used by various researchers to analyze posterior densities for Dirichlet and nonparametric Beta mixtures [17, 18, 38, 41]. Other work [4, 58, 57] established minimax optimal rates for regression functions in nonparametric regression models. Related problems include adaptive rates for the density in nonparametric Bayesian inference [8, 15], Bayesian linear and non-linear inverse problems [33, 25], and

*Department of Statistics, University of Toronto
†Department of Statistics and Data Science, UT Austin.
‡EECS & Mathematics, Massachusetts Institute of Technology
§Department of EECS and Department of Statistics, UC Berkeley.

posterior contraction rates of density under misspecified models [24]. Other popular general frameworks for analyzing the density functions of posterior distributions include those of Shen and Wasserman [42], and Walker et al. [56].

1.1. From frequentist to Bayesian analysis. The focus of this paper is on posterior convergence rates for parameters—namely, how for parametric Bayesian models, the posterior distribution assigns mass to certain regions of the parameter space. Our contributions can be put into perspective by considering known results for M-estimators. In the world of frequentist statistics, estimators based on maximizing empirically-defined objective functions—known as M-estimators—play a central role. In the parametric setting, a generic M-estimator takes the form

$$\widehat{\theta}_n := \arg\max_{\theta \in \Theta} F_n(\theta) \quad \text{where } F_n(\theta) := \tfrac{1}{n}\textstyle\sum_{i=1}^n f(\theta; X_i), \text{ with } X_i \overset{\text{i.i.d.}}{\sim} \mathbb{P} \text{ for } i = 1, \ldots, n, \tag{1.1}$$

while the parameters θ range over some constraint set Θ, and the real-valued function f has domain $\Theta \times \mathcal{X}$. Maximum-likelihood is the archetypal example, obtained when f is the log-likelihood.

There is now a rich and well-developed theory—one which exploits ideas from both optimization theory and empirical process theory—for deriving sharp non-asymptotic bounds on the difference between the estimate $\widehat{\theta}_n$ and the maximizer θ^* of the population-level objective (e.g., see the books [52, 49, 53]). This theory leverages properties of the population-level objective $F(\theta) := \mathbb{E}[f(\theta, X)]$ where the expectation is taken with respect to $X \sim \mathbb{P}$. At a high level, there are two key steps in the analysis of an M-estimator: exploiting the structure of F, and linking the behavior of the empirical objective F_n to the population objective F. In the simplest setting, the population objective is strongly concave around its unique maximum θ^*. More generally, when F is differentiable, one can consider a condition of the following type

$$-\langle \nabla F(\theta),\, \theta - \theta^* \rangle \geq \psi(\|\theta - \theta^*\|_2), \tag{1.2a}$$

assumed to hold uniformly for all θ in a local neighborhood of θ^*. Here ψ is an increasing function on the positive real-line, with $\psi(t) = \frac{\mu}{2} t^2$ being the one obtained for a μ-strongly concave function. The second step is to relate the empirical and population objective, for instance by establishing a uniform bound on their gradients—say

$$\|\nabla F_n(\theta) - \nabla F(\theta)\|_2 \leq \zeta(\|\theta - \theta^*\|_2)\varepsilon_n, \tag{1.2b}$$

where the function ζ is again defined on the positive real line, and ε_n measures the magnitude of the noise.

When the functions F and F_n satisfy bounds of the form (1.2a) and (1.2b), it can be shown that the estimate $\widehat{\theta}_n$ satisfies a bound of the form $\|\widehat{\theta}_n - \theta^*\|_2 \precsim r_n$, where $r_n > 0$ is the largest positive solution to the inequality[1]

$$\psi(r) \leq \varepsilon_n \, \zeta(r). \tag{1.3}$$

[1] This solution exists and is unique under mild regularity conditions on the pair (ψ, ζ).

This framework is very convenient to use, since optimization theory and empirical process theory give us various tools for establishing the local growth condition (1.2a) and the stochastic perturbation bound (1.2b).

By using this framework with care, one can often obtain sharp results in terms of *problem dimension* d, in both the rate itself and sample size lower bound needed to achieve such rates. Moreover, the local growth condition (1.2a) is relatively flexible; for instance, it allows for models in which the Fisher information matrix is singular (so that the function ψ is *not* quadratic). There are many different instantiations of this general approach in past work, including various methods or establishing growth conditions and empirical process bounds [44, 34], analysis of iterative optimization algorithm [2, 12, 28, 21], as well as regularized and constrained M-estimators [27, 6].

1.2. Our contributions. Moving back to the Bayesian setup, it is natural to seek to a similarly flexible and user-friendly method for establishing finite-sample results for posterior contraction. The main contribution of this paper is to do so by using the Langevin diffusion process—a stochastic differential equation that can encode the posterior distribution—as a lens of analysis.

There are natural parallels between our mode of analysis, and deterministic analyses of optimization algorithms via differential equations [45, 43]. To provide such intuition, recall the M-estimator defined by the objective function (1.1). Under the given conditions, its optimum θ^* can be characterized as the limiting point of an *ordinary differential equation* known as the gradient flow, and the rate (1.3) via the gradient flow dynamics for population and empirical loss functions, respectively. Now consider the analogous approach for studying *not* the M-estimator, but rather (in the Bayesian set-up) the posterior distribution. It is well-known [37] that under mild regularity conditions, the posterior distribution can be represented as the stationary distribution of a *stochastic differential equation* known as the Langevin diffusion. Consequently, just as information about the M-estimator can be recovered by studying the gradient flow, we can recover information about the posterior distribution by studying the Langevin diffusion. In particular, we do so by leveraging stochastic calculus so as to control the moments of this diffusion process. At a high-level, our main results involving showing that, under assumptions of the form (1.2), the posterior convergence rate is governed by the inequality $\psi(r) \leq \varepsilon_n \zeta(r) + \frac{d}{n}$. By comparison to inequality (1.3), relevant for M-estimation, we see that this inequality includes an additional $\frac{d}{n}$ term: it characterizes the diffusive behavior (with dimension d and sample size n) induced from sampling from the Gibbs measure e^{-F_n} as opposed to taking its maximum.

With this overview in place, we now summarize the different classes of contributions that are made in this paper:

Posterior contraction under one-point strong convexity. We begin with the simplest setting, in which the population log-likelihood function is strongly concave in a global sense. Under certain regularity conditions,[2] we prove that the posterior contraction rate around the true parameter is $(d/n)^{1/2}$. Our technique allows us to specify precise non-asymptotic conditions on the sample size and other model properties under which a guarantee of this type holds. In

[2]Briefly, we require the prior distribution to be sufficiently smooth and the perturbation error between the population and empirical log-likelihood function to be well-controlled.

many practical examples, the results yield sharp dependence on the problem dimension.

Posterior contraction under weak concavity. We then relax our assumption from strongly concave to weakly concave, and prove related guarantees. Our results allow the Fisher information matrix to be degenerate, in which case the $n^{-1/2}$ convergence rate is not possible, and the contraction rate is governed by the interplay of a local growth assumption and local empirical process bounds. We illustrate these general results for two concrete classes of models: over-specified Bayesian location Gaussian mixture models and Bayesian logistic regression models.

Non-asymptotic Bernstein–von Mises (BvM) results. Our final contribution is to establish two non-asymptotic BvM results for models with non-degenerate Fisher information. We first derive a non-asymptotic upper bound on the Kullback–Leibler (KL) divergence between the posterior distribution and the limiting Gaussian distribution. Second, we prove a non-asymptotic contraction bounds for the posterior distribution that adapts to the geometry of Fisher information. The bound almost matches the tail bounds satisfied by the limiting Gaussian law.

The remainder of the paper is organized as follows. In Section 2, we set up the basic framework for Bayesian models and introduce a diffusion process that admits posterior distribution as its stationary distribution. Section 3 presents the main results whose proofs are in Section 5. Section 4 is devoted to implications to concrete examples. We conclude our work with a discussion in Section 6 while some technical proofs are in the supplementary material [31].

Notation. In the paper, the expression $a_n \gtrsim b_n$ will be used to denote $a_n \geq c b_n$ for some positive universal constant c that does not change with n. Additionally, we write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. For any $n \in \mathbb{N}$, we denote $[n] = \{1, 2, \ldots, n\}$. The notation $\mathbb{S}^{d-1}$ stands for the unit sphere, namely, the set of vectors $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$. Given a vector $\theta \in \mathbb{R}^d$ and a scalar $r > 0$, we use $\mathbb{B}(\theta, r)$ to denote the closed ball centered at θ with radius r. For any subset Θ of $\mathbb{R}^d$, $r \geq 1$, and $\varepsilon > 0$, we denote $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|_r)$ the covering number of Θ under $\|\cdot\|_r$ norm, namely, the minimum number of ε-balls under $\|\cdot\|_r$ norm to cover the entire set Θ. Given a positive-definite matrix $M \succ 0$, we use $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to denote its largest and smallest eigenvalue, respectively, and we use $\kappa(M) := \lambda_{\max}(M)/\lambda_{\min}(M)$ to denote its condition number. Finally, for any $x, y \in \mathbb{R}$, we denote $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. Given a pair of probability distributions P and Q, such that P is absolutely continuous with respect to Q. The Kullback–Leibler (KL) divergence is defined as $D_{\mathrm{KL}}(P \parallel Q) := \mathbb{E}_P\big[\log \frac{dP}{dQ}\big]$.

2. Background and problem formulation.

This section is devoted to background material along with formulation of the problems studied in this paper. We first set up the problem of studying convergence rates for posterior distributions over parameters in Subsection 2.1, and provide background on its representation as the stationary distribution of a Langevin diffusion process in Subsection 2.2. Finally, we define the population likelihood function, and introduce various smoothness conditions in Subsection 2.3.

2.1. Posterior contraction rates for parameters.

Consider a parametric family of distributions $\mathcal{P}_\Theta = \{P_\theta \mid \theta \in \Theta\}$. Throughout the paper, we assume that each distribution P_θ has density p_θ with respect to the Lebesgue measure. Let $X_1^n := (X_1, \ldots, X_n)$ be a sequence of random variables drawn i.i.d. from an underlying distribution P. In the well-specified case, we have that $P = P_{\theta^*} \in \mathcal{P}_\Theta$ for some $\theta^* \in \Theta$. However, it is important to note throughout our paper, the ground truth distribution P does not have to lie in the parametric family $\mathcal{P}_\Theta$. Instead, the posterior contraction results around the parameter θ^* hold as long as certain geometric conditions around θ^* are satisfied. These conditions are typically achieved by the parameter θ^* such that P_θ^* is the best approximation to P within the family. See Section 3 for a concrete discussion about these conditions.

Given a prior π over the parameter space, we define the log-likelihood

$$F_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i), \quad \text{along with the posterior} \quad \mathbb{Q}\left(\theta \mid X_1^n\right) := \frac{e^{nF_n(\theta)}\pi(\theta)}{\int_\Theta e^{nF_n(u)}\pi(u)du}. \tag{2.1}$$

As the sample size n increases, we expect that the posterior distribution will concentrate more of its mass over increasingly smaller neighborhoods of the true parameter θ^*. Posterior contraction rates allow us to study how quickly this concentration of mass takes place. In particular, for a given norm, we study the posterior mass of a ball of the form $\|\theta-\theta^*\| \leq \rho$ for a suitably chosen radius $\rho > 0$. For a given $\delta \in (0,1)$, our goal is to prove statements of the form $\mathbb{Q}\big(\|\theta - \theta^*\| \geq \rho(n,d,\delta) \mid X_1^n\big) \leq \delta$, with probability at least $1-\delta$ over the randomly drawn data X_1^n. Our interest is in the scaling of the radius $\rho(n,d,\delta)$ as a function of sample size n, problem dimension d, and the error tolerance δ, as well as other problem-specific parameters.

2.2. From diffusion processes to the posterior distribution. The analysis of this paper relies on a well-known connection between the posterior distribution and a particular stochastic differential equation (SDE) known as the Langevin diffusion. For a parameter $\beta > 0$, the Langevin diffusion can be written as

$$d\theta_t = -\tfrac{1}{2}\nabla U(\theta_t)dt + \tfrac{1}{\sqrt{\beta}}\, dB_t, \tag{2.2}$$

where $(B_t, t \geq 0)$ is a standard d-dimensional Brownian motion [36], and $U : \mathbb{R}^d \to \mathbb{R}$ is known as the potential function. Suppose that we impose the following regularity conditions on the potential: (a) its gradient ∇U is locally Lipschitz, and (b) its gradient satisfies the inequality $\langle \nabla U(\theta),\, \theta\rangle \geq c_1 \|\theta\|_2 - c_2$ for any $\theta \in \mathbb{R}^d$, for some strictly positive constants c_1, c_2. Under these conditions, by known results on general Langevin diffusions [1], the solution to the Langevin diffusion (2.2) exists and is unique in the strong sense. Furthermore, the density of θ_t converges in $\mathbb{L}^2$ to the stationary distribution with density proportional to $e^{-\beta U}$.

In the context of Bayesian inference, we can apply this argument to the potential function $U_n(\theta) := -F_n(\theta) - n^{-1}\log\pi(\theta)$ and $\beta = n$. Doing so will require us to verify that U_n satisfies the requisite regularity conditions. Assuming this validity, we are guaranteed that the posterior distribution $\mathbb{Q}(\theta \mid X_1^n)$ is the stationary distribution of the SDE

$$d\theta_t = \tfrac{1}{2}\nabla F_n(\theta_t)dt + \tfrac{1}{2n}\nabla \log\pi(\theta_t)dt + \tfrac{1}{\sqrt{n}}dB_t, \tag{2.3}$$

with initial condition $\theta_0 = \theta^*$. Moreover, the density of θ_t converges in $\mathbb{L}^2$ to the posterior density.

It should be noted that this SDE-based representation of the posterior underlies various algorithms for drawing samples from the posterior distribution; we refer the reader to the classical literature [47, 48] and the recent progress [7, 10, 11] for some results in this direction. In this paper, we exploit this SDE-based representation for statistical analysis (as opposed to efficient computation). In particular, by characterizing the behavior of the process $(\theta_t, t \geq 0)$ as a function of time, we can obtain bounds on the posterior distribution by taking limits. The following proposition guarantees the convergence of the moments based on a uniform-in-time moment upper bound and a convergence in total variation distance.

Proposition 2.1. *Consider a sequence of distributions $(\pi_t)_{t\geq 0}$ on $\mathbb{R}^d$ such that $d_{\mathrm{TV}}(\pi_t, \pi^*) \to 0$, and suppose that $\sup_{t\geq 0} \mathbb{E}_{\pi_t}[\|X\|_2^p] < +\infty$ and $\mathbb{E}_{\pi^*}[\|X\|_2^p] < +\infty$ for any integer $p \geq 2$. We then have $\lim_{t\to+\infty} \mathbb{E}_{\pi_t}[\|X\|_2^p] = \mathbb{E}_{\pi^*}[\|X\|_2^p]$.*

See Appendix C.1 in our supplementary material [31] for the proof of this proposition.

Given this limiting behavior, we can establish posterior contraction rates for the parameters by controlling the moments of the diffusion process $\{\theta_t\}_{t\geq 0}$. The main theoretical results of this paper are obtained by following this general roadmap.

2.3. From empirical to population likelihood. Before proceeding to our main results, let us introduce some additional definitions and conditions. A useful notion for our analysis is the population log-likelihood F. It corresponds to the limit of log-likelihood function F_n, as previously defined in equation (2.1), as the sample size n goes to infinity—viz.

$$F(\theta) := \mathbb{E}\left[\log p_\theta(X)\right], \tag{2.4}$$

where the expectation is taken with respect to $X \sim P_{\theta^*}$. Throughout the paper, we impose the following smoothness conditions on the log prior density $\log \pi$:

(A) There exists a non-negative constant $B \geq 0$ such that

$$\langle \nabla \log \pi(\theta),\, \theta - \theta^* \rangle \leq B \|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{R}^d.$$

Although the constant B in Assumption **(A)** can depend on θ^*, we suppress this dependence so as to keep the notation streamlined. When the function $\log \pi$ is globally Lipschitz (so that $\|\nabla \log \pi(\theta)\|_2$ is uniformly bounded), Assumption **(A)** is automatically satisfied. But the one-sided nature of Assumption **(A)** makes it flexible and allows many practical prior distributions. For example, given scalars $\alpha, \beta > 0$, for the prior distribution $\pi(\theta) \propto \exp(-\beta^{-1} \|\theta\|_2^\alpha)$, we have

$$\begin{aligned}\langle \nabla \log \pi(\theta),\, \theta - \theta^* \rangle &= \frac{\alpha}{\beta} \|\theta\|_2^{\alpha-2} \left\{ \langle \theta^*,\, \theta - \theta^* \rangle - \|\theta - \theta^*\|_2^2 \right\} \\ &\leq \begin{cases} 2^{\alpha-2} \frac{\alpha}{\beta} \|\theta^*\|_2^{\alpha-1} \cdot \|\theta - \theta^*\|_2 & \|\theta - \theta^*\|_2 \leq \|\theta^*\|_2\,, \\ 0 & \text{otherwise,} \end{cases}\end{aligned}$$

so that Assumption **(A)** is satisfied by $B = 2^{\alpha-2} \frac{\alpha}{\beta} \|\theta^*\|_2^{\alpha-1}$.

3. Main results. We now turn to our main results. In Subsection 3.1, we present a result (Theorem 3.1) that establishes the posterior convergence under strong concavity. Subsection 3.2 answers the same question when the population log-likelihood is only weakly concave; see the statement of Theorem 3.2. Finally, in Subsection 3.3, we pursue a more fine-grained direction by establishing the non-asymptotic Bernstein–von Mises theorems (see Proposition 3.4 and Theorem 3.5)

3.1. Posterior contraction under strong concavity. We begin with results under strong concavity conditions. For this part, the following assumptions underlie our analysis:

(S.1) There exists a scalar $\mu > 0$ such that

$$-\langle \nabla F(\theta),\, \theta^* - \theta \rangle \geq \mu \,\|\theta - \theta^*\|_2^2 \quad \text{for any } \theta \in \mathbb{R}^d.$$

(S.2) There exist non-negative functions ε_1 and ε_2 that map from $\mathbb{N} \times (0, 1]$ to $\mathbb{R}_+$ such that for any radius $r > 0$ and any $\delta \in (0, 1)$, we have

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla F_n(\theta) - \nabla F(\theta)\|_2 \leq \varepsilon_1(n, \delta) r + \varepsilon_2(n, \delta) \qquad \text{with prob. at least } 1 - \delta.$$

Assumption **(S.1)** is a standard strong concavity condition of function F around θ^*, whereas Assumption **(S.2)** provides uniform control on the gradients of the population and sample log-likelihoods. It is important to note that these assumptions, along with other assumptions to follow, *do not* require the data-generating distribution P to belong to the specified parametric class. Indeed, the results throughout this paper apply to both well-specified and mis-specified models. In the latter case, the parameter θ^* is typically the KL-projection of the true model, i.e., $\theta^* \in \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}_\theta)$.

Given the above assumptions, we are ready to state our first result regarding the posterior convergence rate of parameters for a strongly concave population log-likelihood:

Theorem 3.1. *Suppose that Assumptions* **(A)**, **(S.1)**, *and* **(S.2)** *hold. Then there is a universal constant c such that for any $\delta \in (0, 1)$ and any sample size n for which $\varepsilon_1(n, \delta) \leq \frac{\mu}{6}$, we have*

$$\mathbb{Q}\Big(\|\theta - \theta^*\|_2 \geq c\sqrt{\tfrac{d}{n\mu}} + \tfrac{B}{n\mu} + \tfrac{\varepsilon_2(n,\delta)}{\mu} + c\sqrt{\tfrac{\log(1/\delta)}{n\mu}} \;\Big|\; X_1^n \Big) \leq \delta$$

with probability $1 - \delta$, taken with respect to the random observations X_1^n.

See Subsection 5.1 for the proof of Theorem 3.1.

This result guarantees posterior convergence at the rate $(d/n)^{1/2}$ when the log-likelihood is strongly concave. To be clear, such rate of posterior contraction for the parameters can be derived from the asymptotic behavior of the posterior distribution via the classical Bernstein–von Mises theorem. However, the guarantee in Theorem 3.1 is non-asymptotic, and provides explicit dependence of the rate on other model parameters, including B and μ, both of which might vary as a function of θ^*. At the moment, we do not know whether the dependence of these parameters is optimal. This guarantee is valid as long as the error term $\varepsilon_1(n, \delta)$ is less than an absolute constant; such a bound typically holds as long as $n \gtrsim d$. In Theorem 3.5 to follow, we also provide near-optimal non-asymptotic contraction bounds on the posterior distribution that nearly match the exact shape of the posterior distribution.

Although our set-up is focused on simple sampling models, it should be noted that our method is sufficiently flexible so as to accommodate certain non-i.i.d. forms of sampling, along with mis-specified models. After the first version was posted, Mazumdar et al. [29] used a variant of this result to study the posterior contraction for Thompson sampling in contextual bandits. In their problem, the data are adaptively collected instead of being i.i.d., and the empirical process bound **(S.2)** can be verified using martingale concentration inequalities.

While this paper focuses on the contraction of posterior distribution itself, it is worth mentioning that the proof techniques of Theorem 3.1 can be extended to study the contraction behavior of discretized Langevin diffusion. In particular, by expanding the discrete-time evolution of the iterates following Subsection 5.1, we can derive recursive relations on the moment bounds for the distance between iterates and θ^* using Assumptions **(S.1)** and **(S.2)**. The solution to such recursion will lead to the rates in Theorem 3.1. This analysis tool does not depend on the ergodicity of the discretized diffusion. We defer a detailed discrete-time analysis to future work.

3.2. Posterior contraction under weak concavity.

Theorem 3.1 requires global strong concavity, which is relatively strong. In this section, we relax this assumption in two ways: we relax the growth condition locally around θ^* so as to allow for weak concavity, and the global behavior need not coincide with this local behavior. Weakly concave log-likelihoods arise for singular problems, for which the Fisher information matrix at the true parameter θ^* is rank-degenerate. Examples of such singular problems include Bayesian non-linear regression models with certain choices of link functions [30], as well as over-specified mixture models [39], in which the fitted mixture model has more components than the true mixture distribution. The mismatch between local and global concavity conditions exists not only in such models, but also in non-singular problems such as Bayesian logistic regression. We discuss implications of these examples in Section 4.

Our analysis in the weakly concave setting is based on the following assumptions:

(W.1) There exists a convex, non-decreasing function $\psi : [0, +\infty) \to \mathbb{R}$ such that

$$-\langle \nabla F(\theta), \theta - \theta^* \rangle \geq \psi(\|\theta - \theta^*\|_2) \qquad \text{for any } \theta \in \mathbb{R}^d.$$

Assumption **(W.1)** characterizes the weak concavity of the function F around the global maxima θ^*. This condition can hold when the log-likelihood is locally strongly concave around θ^* but only weakly concave in a global sense, or it can hold when the log-likelihood is weakly concave but not strongly concave. An example of the former type is the logistic regression model analyzed in Subsection 4.1, whereas an example of the latter type is given by over-specified Gaussian mixture models Subsection 4.2.

Our next assumption controls the deviation between the gradients of the population and sample likelihoods, and involves a failure probability $\delta \in (0, 1)$:

(W.2) There exist a function $\varepsilon : \mathbb{N} \times (0, 1] \mapsto \mathbb{R}_+$ and a non-decreasing function $\zeta : \mathbb{R} \to \mathbb{R}$ with that $\zeta(0) \geq 0$ such that for any radius $r > 0$, we

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla F_n(\theta) - \nabla F(\theta)\|_2 \leq \varepsilon(n, \delta)\zeta(r) \qquad \text{with prob. at least } 1 - \delta.$$

This type of localized empirical process bounds appeared in many existing literature in the study of M-estimators [50] and iterative algorithms [2, 12]. It is important to note that the bound depends on the radius r, making it possible to yield near-optimal rates in singular mixture models [12].

The previous conditions involved two functions, namely ψ and ζ. We let $\xi : \mathbb{R}_+ \to \mathbb{R}$ denote the inverse function of the strictly increasing function $r \mapsto r\zeta(r)$. Our third assumption imposes certain inequalities on these functions and their derivatives:

(W.3) The function $r \mapsto \psi(\xi(r))$ is convex, and ψ and ζ satisfy the differential inequalities

$$r\psi'(r)\zeta(r) \overset{(i)}{\geq} r\psi(r)\zeta'(r) + \psi(r)\zeta(r), \quad \text{and}$$
$$r^2\psi''(r)\zeta(r) + r\psi'(r)\zeta(r) \overset{(ii)}{\geq} 3\psi(r)\zeta(r) + r^2\psi(r)\zeta''(r) \quad \text{for all } r > 0.$$

These differential inequalities are needed controlling the moments of the diffusion process $\{\theta_t\}_{t>0}$ in equation (2.3). In our discussion of concrete examples, we provide instances for which they are satisfied.

Our result involves a certain fixed point equation that depends on the parameters and functions in our assumptions. In particular, for any tolerance parameter $\delta \in (0,1)$ and sample size n, consider the following fixed point equation in the variable $z > 0$:

$$\psi(z) = \varepsilon(n,\delta)\zeta(z)z + \tfrac{B}{n}z + \tfrac{d}{n} + \tfrac{\log(1/\delta)}{n}. \tag{3.1}$$

In order to ensure that this equation has a unique positive solution, our final assumption imposes certain condition on the growth of the functions ψ and ζ:

(W.4) The limit $\lim\inf_{z\to+\infty} \frac{\psi(z)}{z\zeta(z)}$ is strictly positive, and the sample size n and tolerance parameter $\delta \in (0,1)$ are such that $\varepsilon(n,\delta) < \lim\inf_{z\to+\infty} \frac{\psi(z)}{z\zeta(z)}$.

With this set-up, we are now ready to state our second main result:

Theorem 3.2. *Suppose that Assumptions* **(A)**, *and* **(W.1)**— **(W.3)** *hold. Then for any given sample size n and $\delta \in (0,1)$ such that Assumption* **(W.4)** *holds, equation (3.1) has a unique positive solution $z^*(n,\delta)$ such that*

$$\mathbb{Q}\Big(\|\theta - \theta^*\|_2 \geq z^*(n,\delta) \;\big|\; X_1^n \Big) \leq \delta \quad \textit{with probability } 1-\delta \textit{ w.r.t. } X_1^n. \tag{3.2}$$

See Subsection 5.2 for the proof of Theorem 3.2.

A few comments are in order. First, the convergence guarantee (3.2) depends on the weak convexity function ψ and the perturbation function ζ through the non-linear equation (3.1). In order to understand the rate, we consider the following pair of fixed-point equations

$$\psi(z) = 2\varepsilon(n,\delta)\zeta(z)z, \quad \text{with the solution } z^*_{\text{mle}}(n,\delta) \tag{3.3a}$$
$$\psi(z) = 2\tfrac{B}{n}z + 2\tfrac{d}{n} + 2\tfrac{\log(1/\delta)}{n}, \quad \text{with the solution } z^*_{\text{pop}}(n,\delta). \tag{3.3b}$$

It is easy to see that $z^*(n,\delta) \leq \max\{z^*_{\text{mle}}(n,\delta), z^*_{\text{pop}}(n,\delta)\}$.[3] This establishes that the posterior contraction rates in Theorem 3.2 are fundamentally determined by two sources of errors: on the one hand, it is known (see e.g. [50]) that the solution $z^*_{\text{mle}}(n,\delta)$ to equation (3.3a) determines (up to constant factors) the rate of convergence for the maximal likelihood estimator; on the

[3]Suppose the converse is true. We have $\psi(z^*(n,\delta)) > 2\varepsilon(n,\delta)\zeta(z^*(n,\delta))z^*(n,\delta)$ and $\psi(z^*(n,\delta)) > 2\frac{B}{n}z^*(n,\delta) + 2\frac{d}{n} + 2\frac{\log(1/\delta)}{n}$. Taking the average of two inequalities contradicts the fact that $z^*(n,\delta)$ is the fixed point.

other hand, the solution $z^*_{\mathrm{pop}}(n,\delta)$ to equation (3.3b) captures the diffusive behavior from the posterior distribution itself. In particular, the term $\frac{B}{n}z$ is usually negligible as $z^*_{\mathrm{pop}} \ll 1$ and $B = O(\sqrt{d})$, and the solution to the equation $\psi(z) = \frac{d+\log(1/\delta)}{n}$ essentially determines the contraction rate of the "population-level posterior" Gibbs distribution whose density is proportional to $e^{nF(\theta)}$, when the function $-\langle \nabla F(\theta),\, \theta - \theta^* \rangle$ locally behaves like $\psi(\|\theta - \theta^*\|_2)$. We suspect that such an additional term is unavoidable for posterior contraction results, and we defer a rigorous justification via asymptotic shape of the re-scaled posterior to future works.

Second, at least in general, it is not possible to compute an explicit form for the positive solution $z^*(n,\delta)$ to the non-linear equation (3.1). However, for certain forms of the function ψ and ζ, we can derive a relatively simple upper bound. For instance, given some positive parameters (α,β) such that $\alpha > \beta$, suppose that these functions are defined locally, in a interval above zero, as follows:

$$\psi(r) = r^{\alpha+1}, \quad \text{and} \quad \zeta(r) = r^{\beta} \qquad \text{for all } r \text{ in some interval } [0,\bar{r}). \tag{3.4a}$$

Moreover, suppose that the perturbation function takes the form

$$\varepsilon(n,\delta) = \sqrt{\left(d + \log(\tfrac{1}{\delta})\right)/n}. \tag{3.4b}$$

As shown in in Section 4, these particular forms arise in several statistical models, including Bayesian logistic regression and over specified Bayesian Gaussian mixture models. Under these conditions, we have the following simple upper bound:

Corollary 3.3. *Assume that the functions ψ, ζ have the local behavior (3.4a), and the perturbation term $\varepsilon(n,\delta)$ has the form (3.4b). If, in addition, the global forms of ψ and ζ satisfy Assumption* **(W.3)**, *then for sufficiently large n, the scalar $z^*(n,\delta)$ from Theorem 3.2 satisfies the bound $z^*(n,\delta) \leq c\, \left(\frac{d+\log(1/\delta)}{n}\right)^{\frac{1}{2(\alpha-\beta)}} \vee \left(\frac{d+\log(1/\delta)}{n}\right)^{\frac{1}{\alpha+1}} + \left(\frac{B}{n}\right)^{\frac{1}{\alpha}}$.*

Note that Corollary 3.3 ensures that the posterior has the following contraction property

$$\mathbb{Q}\Big(\|\theta - \theta^*\|_2 \geq c \left(\tfrac{d+\log(1/\delta)}{n}\right)^{\frac{1}{2(\alpha-\beta)} \wedge \frac{1}{\alpha+1}} + \left(\tfrac{B}{n}\right)^{\frac{1}{\alpha}} \,\Big|\, X_1^n\Big) \leq \delta \quad \text{with prob. } 1-\delta \tag{3.5}$$

with respect to the training data. The posterior convergence rate scales as $(d/n)^{\frac{1}{2(\alpha-\beta)}}$ when $\alpha \geq 2\beta + 1$, in which case the posterior contraction rates match the maximal likelihood. On the other hand, this rate becomes $(d/n)^{\frac{1}{\alpha+1}}$ when $\alpha < 2\beta + 1$, and the posterior contraction is slower than maximal likelihood, owing to its diffusive behavior.

Theorem 3.2 and Corollary 3.3 rely on global conditions **(W.1)** and **(W.2)**. Although these conditions can be verified for many practical examples (see Section 4), they can be restrictive in some cases, especially when multiple local maxima of the population-level function F exist. Using our techniques, it is possible to prove similar results under local assumptions. In particular, suppose that these assumptions hold only in a local ball $\mathbb{B}(\theta^*, r_0)$; then, the non-asymptotic contraction rates in Theorem 3.2 and Corollary 3.3 are available as long as we can show the posterior mass $\mathbb{Q}\big(\mathbb{B}(\theta^*, r_0)^c \mid X_1^n\big)$ is small with high probability. To obtain these rates, we could apply the arguments in the proof of Theorem 3.2 to a modified distribution, which matches the shape of $\mathbb{Q}(\cdot \mid X_1^n)$ inside the ball $\mathbb{B}(\theta^*, r_0)$, while exhibiting desirable growth and smoothness conditions outside. We defer the detailed arguments based on local assumptions as well as the study of the radius r_0 to future work.

3.3. Non-asymptotic Bernstein–von Mises results. In this section, we develop non-asymptotic Bernstein–von Mises results using the diffusion process (2.3). Under mild assumptions on the population-level and empirical-level landscapes, we establish the KL divergence between the posterior distribution and the limiting Gaussian distribution, as well as near-optimal shape-dependent posterior contraction results.

In order to obtain the non-asymptotic Bernstein–von Mises results, we first need the following assumptions on the second order derivatives with respect to the parameters (or equivalently Hessian matrices) of the empirical and population log-likelihoods:

(BvM.1) There exists $A > 0$ such that the population log-likelihood function F satisfies the one-point Lipschitz condition:

$$\forall \theta \in \mathbb{R}^d, \quad \left\|\nabla^2 F(\theta) - \nabla^2 F(\theta^*)\right\|_{\mathrm{op}} \leq A \left\|\theta - \theta^*\right\|_2 .$$

(BvM.2) For any $\delta > 0$, there exist non-negative functions $\varepsilon_1^{(2)}$ and $\varepsilon_2^{(2)}$ with domain $\mathbb{N} \times (0, 1]$ such that

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\|\nabla^2 F_n(\theta) - \nabla^2 F(\theta)\right\|_{\mathrm{op}} \leq \varepsilon_1^{(2)}(n, \delta) r + \varepsilon_2^{(2)}(n, \delta),$$

for any radius $r > 0$ with probability at least $1 - \delta$.

Additionally, we also impose a smoothness assumption on the prior distribution π

$$\text{(PS)} \qquad \left\|\nabla \log \pi(\theta_1) - \nabla \log \pi(\theta_2)\right\|_2 \leq L_2 \left\|\theta_1 - \theta_2\right\|_2 .$$

The first condition **(BvM.1)** is a standard smoothness condition needed to prove quantitative results about asymptotic normality (e.g., the paper [35]), and satisfied by many models such as exponential family models, location density models, as well as their mixtures and hierarchical composition. The second condition **(BvM.2)** is an empirical process condition on the Hessian matrix $\nabla^2 F_n$. This condition can usually be verified using suitable concentration bounds for each θ, as well as smoothness conditions on $\nabla^2 F_n$ used in controlling metric entropies. Both assumptions are naturally needed: the limiting Gaussian law $\mathcal{N}\big(\widehat{\theta}^{(n)}, (nH^*)^{-1}\big)$, which depends on the population-level Hessian at the point θ^*. The shape of posterior distribution, on the other hand, depends on the sample-level Hessian $\nabla^2 F_n$ in a local neighborhood of θ^*. These two conditions are needed to relate the shape of the sample-level posterior with the matrix H^*. The condition (PS) on the prior distribution is relatively mild and satisfied by many practical choices including Gaussian. As before, we note that these assumptions do not require the model to be well-specified, and our non-asymptotic Bernstein–von Mises theorems applies to the mis-specified case, where θ^* is the KL-projection of the model to this parametric class.

Consider the MAP estimate $\widehat{\theta}^{(n)} := \arg\max_{\theta \in \mathbb{R}^d} \big(F_n(\theta) + \frac{1}{n} \log \pi(\theta)\big)$. Then, we have the following upper bound on the difference between the posterior distribution of the parameters and the Gaussian distribution with mean $\widehat{\theta}^{(n)}$ and covariance matrix $(nH^*)^{-1}$, where $H^* := -\nabla^2 F(\theta^*)$.

Proposition 3.4. *Under Assumptions* **(BvM.1)**, **(BvM.2)** *and PS, suppose that* $H^* \succ 0$, *and that* $\|\widehat{\theta}^{(n)} - \theta^*\|_2 \leq \sigma\sqrt{\frac{d}{n}}$ *and* $\mathbb{E}_{\mathbb{Q}}(\|\theta - \theta^*\|_2^4 \mid X_1^n)^{1/4} \leq \sigma\sqrt{\frac{d}{n}}$ *with prob.* $1-\delta$. *Then there exists a constant* c *such that the KL divergence* $D_{KL}(\mathbb{Q}(\cdot \mid X_1^n) \parallel \mathcal{N}(\widehat{\theta}^{(n)}, (nH^*)^{-1}))$ *is at most*

$$c \cdot \frac{1}{\lambda_{\min}(H^*)}\left(\frac{A^2 d^2\sigma^4}{n} + \frac{\varepsilon_1^{(2)}(n,\delta)^2 d^2 \sigma^4}{n} + \sigma^2\left(\varepsilon_2^{(2)}(n,\delta)^2 + \frac{L_2^2}{n^2}\right)d\right) \quad \text{with prob. at least } 1-2\delta.$$

See Appendix A.2 for the proof of this claim.

A few remarks are in order. First, assuming that the problem-dependent constants (A, σ, L_2) are of constant order, and that the deviation bound scales as $\varepsilon_2^{(2)}(n,\delta) = O(1/\sqrt{n})$, Proposition 3.4 shows that the KL divergence between the posterior distribution and the Gaussian limit is of order $O(1/n)$; second, the non-asymptotic behavior of posterior distribution depends on the Hessian matrix $H^* = -\nabla^2 F(\theta^*)$. In the well-specified case where the data points X_1^n are i.i.d. samples from the distribution $\mathbb{P}_{\theta^*}$, the standard Fisher-information identity $H^* = \mathbb{E}_{\theta^*}\left[\nabla \log p_{\theta^*}(X) \nabla \log p_{\theta^*}(X)^\top\right]$ holds true, and the Bayesian credible set is asymptotically the same as the confidence set in the frequentist sense. On the other hand, in the mis-specified models where $\theta^* = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}_\theta)$, the limiting Gaussian law is $\mathcal{N}(\widehat{\theta}^{(n)}, (nH^*)^{-1})$, depending on the Hessian matrix but not the covariance of the log-likelihood. This result coincides with the asymptotic Bernstein–von Mises theorem for mis-specified parametric models [23], providing a non-asymptotic characterization. Using Pinsker's inequality and Talagrand's T_2-inequality [46], the KL divergence bound can also be transformed into bounds in term of total variation and Wasserstein-2 distances, yielding a non-asymptotic $O(1/\sqrt{n})$ rate of convergence.

We can also use the diffusion process approach to derive more fine-grained concentration bounds for the posterior distribution, with behavior matching the limiting Gaussian law. Doing so requires the following stronger version of the posterior contraction condition:

$$\left(\mathbb{E}_{\mathbb{Q}}\left[\|\theta - \theta^*\|_2^{2p} \mid X_1^n\right]\right)^{1/p} \leq \frac{\sigma^2 pd}{n}, \quad \text{for all } p > 0 \text{ with probability at least } 1-\delta. \tag{3.6}$$

In addition, we define the function

$$\mathcal{H}_n(t,\delta) := (A + \varepsilon_1^{(2)}(n,\delta))^2 \cdot \frac{\sigma^4 d^2 t^2}{n^2} + \frac{\sigma d}{n}\left(\varepsilon_2^{(2)}(n,\delta)^2 + \frac{L_2^2}{n^2} + (A + \varepsilon_1^{(2)}(n,\delta))^2 \frac{\sigma d}{n}\right),$$

which plays the role of a higher-order term. Equipped with this notation, we have:

Theorem 3.5. *Suppose that conditions* **(BvM.1)**, **(BvM.2)**, *and* (PS) *are in force, the Hessian* H^* *is strictly positive definite, and the high-probability posterior contraction condition* (3.6) *holds. Then for any* $\delta \in (0,1)$, *uniformly over all* $\omega \in (0,1)$ *and* $t > 0$, *we have*

$$\mathbb{Q}\left(\left\|\theta - \widehat{\theta}^{(n)}\right\|_{H^*}^2 \geq (1+\omega)\frac{d}{n} + c\frac{1 + \log\kappa(H^*)}{\omega}\left(\frac{t}{n} + \mathcal{H}_n(t,\delta)\right) \;\middle|\; X_1^n\right) \leq e^{-t}, \tag{3.7}$$

with probability at least $1-\delta$.

See Appendix A.1 for the proof of the theorem.

A few remarks are in order. Note that the limiting Gaussian density $\gamma_n = \mathcal{N}\big(0, (nH^*)^{-1}\big)$ satisfies a tail bound of the form $\gamma_n\Big(\|\theta - \widehat{\theta}^{(n)}\|_{H^*}^2 \geq \frac{d}{n} + \frac{t}{n}\Big) \leq e^{-t/2}$ for any $t > 0$. Unless the posterior is actually Gaussian in finite samples, it cannot satisfy this bound exactly. However, Theorem 3.5 provides a bound with near-matching behavior: note that the leading-order term scales $\frac{d}{n}$, matching the asymptotics with a pre-factor $1 + \omega$ that can be made arbitrarily close to 1 (at the expense of the other term). The $\frac{t}{n}$ dependency on the tail probability comes with a mild $\log \kappa(H^*)$ factor due to technical reasons. The bound also contains a high-order term $\mathcal{H}_n(t, \delta)$, which scales as $O(n^{-2})$. It is also worth noticing that the terms in Theorem 3.5 depend on the tail probability $\nu = e^{-t}$ only logarithmically, allowing for very small value of ν. We can therefore use equation (3.7) to construct non-asymptotic credible sets of ellipsoid shape, adapted to the geometry of local Hessian matrix H^*.

Proof outline: The proofs of both Proposition 3.4 and Theorem 3.5 rely on a first-order approximation of the gradient ∇F_n. In particular, the diffusion process (2.3) can be written in the form $d\theta_t = -\frac{1}{2}H^*(\theta_t - \widehat{\theta}^{(n)})dt + \frac{1}{2}e_n(\theta_t)dt + \frac{1}{2n}\log \pi(\theta_t)dt + \frac{1}{\sqrt{n}}dB_t$, where we have defined the linearization error $e_n(\theta) := \nabla F_n(\theta) + H^*(\theta - \theta^*)$. Under the smoothness assumption **(BvM.1)** and the empirical process bound **(BvM.2)**, one can show that $\|e_n(\theta)\|_2 \leq \|\theta - \theta^*\|_2 \cdot O(\sqrt{d/n})$ with high probability. When this error term is ignored, the diffusion process is an Ornstein–Uhlenbeck process whose stationary distribution is $\mathcal{N}\big(\widehat{\theta}^{(n)}, (nH^*)^{-1}\big)$. Therefore, given the non-asymptotic bounds on the error $e_n(\theta)$ stated above, we can provide a non-asymptotic characterization of the distance between the stationary distribution and the limiting Gaussian law. In order to prove Proposition 3.4, we use the Gaussian log-Sobolev inequality [19] to control the KL divergence, whereas proving Theorem 3.5 is based on using Itô calculus to study the growth of a Lyapunov function defined using the metric induced by H^*. Full proofs for the two results are given in Appendix A.2 and Appendix A.1, respectively.

4. Some illustrative examples.

Having developed some general theory, we now use it to derive some concrete results for two examples of interest in statistical analysis: Bayesian logistic regression and Gaussian mixture models.

4.1. Bayesian logistic regression.

Logistic regression is a classical way of modelling the relationship between a binary response variable $Y \in \{-1, +1\}$ and a vector $X \in \mathbb{R}^d$ of explanatory variables (e.g., see the book [30]). In the logistic regression model, the pair (X, Y) are related by the conditional distribution

$$\mathbb{P}\left(Y = 1 \mid X, \theta\right) = \frac{e^{\langle X, \theta \rangle}}{1 + e^{\langle X, \theta \rangle}}, \qquad \text{where } \theta \in \mathbb{R}^d \text{ is a parameter vector.} \tag{4.1}$$

Suppose that we observe a collection $Z_1^n = \{Z_i\}_{i=1}^n$ of n i.i.d paired samples $Z_i = (X_i, Y_i)$, each generated in the following way. First, the covariate vector X_i is drawn from a standard Gaussian distribution $N(0, I_d)$, and then the binary response Y_i is drawn according to the conditional distribution $\mathbb{P}\left(\cdot \mid X_i, \theta^*\right)$ from equation (4.1), where $\theta^* \in \mathbb{R}^d$ is a fixed but unknown value of the parameter vector. Given these assumptions, the sample log-likelihood function of the samples Z_1^n takes the form $F_n^R(\theta) := \frac{1}{n}\sum_{i=1}^n \left\{\log \mathbb{P}\left(Y_i \mid X_i, \theta\right) + \log \phi(X_i)\right\}$, where ϕ denotes the density of a standard normal vector. Combining this log-likelihood with a given prior π over θ yields the posterior distribution in the usual way. We assume that the prior function π satisfies Assumption **(A)**, and recall the constant B defined in this assumption. Throughout this section, we also assume that the norm $\|\theta^*\|_2$ is a universal constant independent of (n, d), and we suppress the dependence on this parameter.

With this set-up, the following result establishes the posterior convergence rate of θ around θ^*, conditionally on the observations Z_1^n.

Corollary 4.1. *For any $\delta \in (0,1)$, given $\frac{n}{\log n} \geq c' d \log(\frac{1}{\delta})$ i.i.d. samples from the Bayesian logistic regression model (4.1), we have $\mathbb{Q}\Big(\|\theta - \theta^*\|_2 \geq c\big\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{B}{n}\big\} \mid Z_1^n\Big) \leq \delta$ with probability $1-\delta$ over the data Z_1^n.*

See Appendix B.1 for the proof of this claim.

A few comments are in order. First, the result of Corollary 4.1 shows that for Bayesian logistic regression model (4.1), the posterior convergence rate for the parameter is of the order $(d/n)^{1/2}$. Furthermore, this result also gives a concrete dependence of the rate on B characterizing the degree to which the prior is concentrated away from the true parameter. By taking the standard Gaussian prior $\pi = \mathcal{N}(0, I_d)$, we have $B \leq \|\theta\|_2$, which is bounded by a universal constant independent of the pair (n, d).

It is important to note that Corollary 4.1 is valid as long as the sample size n is mildly larger than the problem dimension d (up to logarithmic factors). To our knowledge, this is the first time that a sharp non-asymptotic posterior contraction result is established in this regime.

Let us sketch how Theorem 3.2 can be applied so as to prove this corollary. Denote $F^R := \mathbb{E}[F_n^R]$ as the population-level log-likelihood function. The first step in our proof, as given in Appendix B.1, is to show that there are universal constants c, c_1, c_2 such that

$$-\langle \nabla F^R(\theta),\, \theta - \theta^* \rangle \geq c_1 \begin{cases} \|\theta - \theta^*\|_2^2, & \text{for all } \|\theta - \theta^*\|_2 \leq 1 \\ \|\theta - \theta^*\|_2, & \text{otherwise} \end{cases}, \quad \text{and} \tag{4.2a}$$

$$\sup_{\theta \in \mathbb{R}^d} \left\|\nabla F_n^R(\theta) - \nabla F^R(\theta)\right\|_2 \leq c_2 \left(\sqrt{\tfrac{d}{n}} + \sqrt{\tfrac{\log(1/\delta)}{n}} + \tfrac{\log(1/\delta)}{n}\right), \tag{4.2b}$$

for any $r > 0$ with probability $1-\delta$ as long as $\frac{n}{\log n} \geq c d \log(1/\delta)$. Using these results, we show that Assumptions **(W.1)** and **(W.2)** hold with

$$\psi(r) = c_1 \begin{cases} r^2 & \text{for all } r \in (0,1), \text{ and} \\ r & \text{otherwise} \end{cases}, \quad \text{and} \quad \zeta(r) = c_2 \quad \text{for all } r > 0. \tag{4.3}$$

We can check that the functions ψ and ζ satisfy the conditions in Assumptions **(W.3)** and **(W.4)**. Therefore, applying Theorem 3.2 to these functions yields the posterior contraction rate claimed in Corollary 4.1. See Appendix B.1 for the details.

4.2. Over-specified Bayesian Gaussian mixture models. Gaussian mixtures are widely used for modelling heterogeneous datasets; clusters in the data are naturally associated with different mixture components [26]. In fitting such models, the true number of components is generally unknown, and several approaches have been proposed to deal with this challenge. One of the most popular methods is to deliberately include a large number of components, leading to what are known as overspecified Gaussian mixture models [39]. While the behavior of posterior densities in such mixture models is relatively well-understood [17], the behavior of the posterior in terms of its parametric components is not as well understood. When the covariance matrices are known and the parameter space is bounded, the location parameters have been shown to have posterior convergence rates of the order $n^{-1/4}$ in the Wasserstein-2 metric [32]. However, neither the dependence on dimension d nor on the true number of components have been established.

In this section, we consider the behavior of overspecified Gaussian mixture models in a particular setting, and provide convergence rates for the parameters with precise dependence on the dimension d, and without requiring any boundedness assumption. In order to model the simplest form of over-specification, suppose that we fit a Bayesian location mixture model to a collection of i.i.d. samples $X_1^n = (X_1, \ldots, X_n)$ drawn from a Gaussian distribution $\mathcal{N}(\theta^*, I_d)$. (For concreteness, we set $\theta^* = 0$.) We study the behavior of the Bayesian Gaussian mixture model

$$\theta \sim \pi(\cdot), \qquad V_i \in \{-1, 1\} \overset{\text{i.i.d.}}{\sim} \text{Cat}(1/2, 1/2), \qquad X_i \mid V_i, \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(V_i\theta, I_d), \tag{4.4}$$

where Cat$(1/2, 1/2)$ stands for the categorical distribution with parameters $(1/2, 1/2)$. We assume that the prior π satisfies the smoothness condition (cf. Assumption **(A)**); one example is a Gaussian distribution (over the location parameter θ. Our goal in this section is to characterize the posterior contraction rate of the location parameter θ around θ^*.

In order to do so, we first define the sample log-likelihood function F_n^G given data X_1^n. It has the form $F_n^G(\theta) := \frac{1}{n}\sum_{i=1}^n \log\left(\frac{1}{2}\phi(X_i; -\theta, I_d) + \frac{1}{2}\phi(X_i; \theta, I_d)\right)$, where $x \mapsto \phi(x; \theta, I_d) = (2\pi)^{-d/2}e^{-\|x-\theta\|_2^2/2}$ denotes the density of multivariate Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_d)$. Similarly, the population log-likelihood function is given by

$$F^G(\theta) := \mathbb{E}_X\left[\log\left(\frac{1}{2}\phi(X; -\theta, I_d) + \frac{1}{2}\phi(X; \theta, I_d)\right)\right],$$

where the outer expectation in the above display is taken with respect to $X \sim \mathcal{N}(\theta^*, I_d)$.

In Appendix B.2, we prove that there is a universal constant $c_1 > 0$ such that

$$-\langle \nabla F^G(\theta),\, \theta - \theta^*\rangle \geq \begin{cases} c_1 \|\theta - \theta^*\|_2^4, & \text{for all } \|\theta - \theta^*\|_2 \leq \sqrt{2} \\ 4c_1\left(\|\theta - \theta^*\|_2^2 - 1\right), & \text{otherwise} \end{cases}, \tag{4.5a}$$

and moreover, there are universal constants (c, c_2) such that for any $\delta \in (0, 1)$, given a sample size $n \geq cd\log(1/\delta)$, we have

(4.5b)

$$\sup_{\theta \in \mathbb{B}(\theta^*, r)} \|\nabla F_n^G(\theta) - \nabla F^G(\theta)\|_2 \leq c_2\left(r + \tfrac{1}{\sqrt{n}}\right)\left(\sqrt{\tfrac{d}{n}} + \sqrt{\tfrac{\log(\log(n/\delta))}{n}}\right) \quad \text{with prob. } 1 - \delta.$$

Given the above results, the functions ψ and ζ in Assumptions **(W.1)** and **(W.2)** take the form

$$\psi(r) = \begin{cases} c_1 r^4, & \text{for all } 0 < r \leq \sqrt{2} \\ 4c_1\left(r^2 - 1\right), & \text{otherwise} \end{cases}, \quad \text{and} \quad \zeta(r) = r + \frac{1}{\sqrt{n}} \quad \text{for all } r > 0. \tag{4.6}$$

These functions satisfy the conditions of Assumptions **(W.3)** and **(W.4)**. Therefore, it leads to the following result regarding the posterior contraction rate of parameters under overspecified Bayesian location Gaussian mixtures (4.4):

Corollary 4.2. *Given the overspecified Bayesian location Gaussian mixture model* (4.4)*, there are universal constants* c, c' *such that given any* $\delta \in (0,1)$ *and a sample size* $n \geq c' d \log(1/\delta)$*, we have* $\mathbb{Q}\Big(\|\theta - \theta^*\|_2 \geq c \left(\frac{d}{n} + \frac{\log(\log(n/\delta))}{n}\right)^{1/4} + \left(\frac{B}{n}\right)^{1/3} \mid X_1^n \Big) \leq \delta$ *with probability* $1-\delta$ *over the data* X_1^n*. Here,* B *is the non-negative constant in Assumption* **(A)**.

See Appendix B.2 for the proof of Corollary 4.2.

The $O(n^{-1/4})$ rate of convergence in Corollary 4.2 is consistent with the previous result with location parameters in overspecified Bayesian location Gaussian mixtures [5, 22, 32], which is also known to be minimax optimal [20]. When taking the problem dimension into account, to our knowledge, the $(d/n)^{1/4}$ posterior contraction rate is a novel result, and matches existing analyses for frequentist methods [12]. Similar to the logistic regression case, Corollary 4.2 only requires the sample size n to be mildly larger than the dimension d. The non-asymptotic posterior contraction results are also established for the first time in such a regime. Finally, our result does not require the boundedness of the parameter space, in contrast to past work [5, 22, 32].

5. Proofs.

In this section, we collect the proofs of the main theorems.

5.1. Proof of Theorem 3.1.

Throughout the proof, in order to simplify notation, we omit the conditioning on the σ-field $\mathcal{F}_n := \sigma(X_1^n)$; it should be taken as given. Introduce the quantity $\alpha = \frac{1}{2}\mu - \varepsilon_1(n,\delta) > \frac{\mu}{6}$. Our proof relies on proving the following auxiliary bound

$$\frac{1}{2} e^{\alpha t} \|\theta_t - \theta^*\|_2^2 \leq \frac{1}{\sqrt{n}} M_t + U_n \frac{(e^{\alpha t} - 1)}{2\alpha}, \tag{5.1}$$

where $U_n := \frac{3B^2}{n^2} + \frac{3\varepsilon_2^2(n,\delta)}{\mu} + \frac{d}{n}$ and $M_t := \int_0^t e^{\alpha s} \langle \theta_s - \theta^*, dB_s \rangle$. By construction, the latter term is a martingale.

The proof of the bound (5.1) is given later in this section; we take it as given for the moment, and use it to prove the theorem. In order to bound the moments of martingale M_t, for any $p \geq 4$, we invoke the Burkholder–Davis–Gundy inequality (e.g., §4.4 of the book [36]) to find that

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\leq t\leq T} |M_t|^{\frac{p}{2}}\right] &\leq (pC)^{\frac{p}{4}} \mathbb{E}\left[[M]_T^{\frac{p}{4}}\right] = (pC)^{\frac{p}{4}} \mathbb{E}\left(\int_0^T e^{2\alpha s} \|\theta_s - \theta^*\|_2^2 \, ds\right)^{\frac{p}{4}} \\
&\leq (pC)^{\frac{p}{4}} \mathbb{E}\left(\sup_{0\leq t\leq T} e^{\alpha t} \|\theta_t - \theta^*\|_2^2 \int_0^T e^{\alpha s} ds\right)^{\frac{p}{4}} \leq \left(\frac{pCe^{\alpha T}}{\alpha}\right)^{\frac{p}{4}} \mathbb{E}\left(\sup_{0\leq t\leq T} e^{\alpha t} \|\theta_s - \theta^*\|_2^2\right)^{\frac{p}{4}},
\end{aligned}$$

where C is a universal constant. Therefore, we arrive at the bound

$$\mathbb{E}\left[\left(\sup_{0\le t\le T} e^{\alpha t}\|\theta_t-\theta^*\|_2\right)^p\right] \le \mathbb{E}\left(\frac{2}{\sqrt{n}}M_t\right)^{\frac{p}{2}} + \left(U_n\frac{(e^{\alpha T}-1)}{\alpha}\right)^{\frac{p}{2}}$$
$$\le \left(U_n\frac{e^{\alpha T}}{\alpha}\right)^{\frac{p}{2}} + \left(\frac{pCe^{\alpha T}}{\alpha n}\right)^{\frac{p}{4}}\mathbb{E}\left(\sup_{0\le s\le T} e^{\alpha s}\|\theta_s-\theta^*\|_2^2\right)^{\frac{p}{4}}.$$

For the right hand side of the above inequality, we can relate it to the left hand side by using Young's inequality, which is given by

$$\left(\frac{pCe^{\alpha T}}{\alpha n}\right)^{\frac{p}{4}}\mathbb{E}\left(\sup_{0\le s\le T} e^{\alpha s}\|\theta_s-\theta^*\|_2^2\right)^{\frac{p}{4}} \le \frac{1}{2}\left(\frac{pCe^{\alpha T}}{\alpha n}\right)^{\frac{p}{2}} + \frac{1}{2}\mathbb{E}\left(\sup_{0\le s\le T} e^{\alpha s}\|\theta_s-\theta^*\|_2^2\right)^{\frac{p}{2}}.$$

Putting the above results together, and let $\alpha=\frac{\mu}{2}$, we find that

$$\left(\mathbb{E}\left[\|\theta_T-\theta^*\|_2^p\right]\right)^{\frac{1}{p}} \le e^{-\alpha T}\left(\mathbb{E}\sup_{0\le t\le T}\left(e^{\alpha t}\|\theta_t-\theta^*\|_2^p\right)\right)^{\frac{1}{p}} \le C'\left(\sqrt{\frac{U_n}{\mu}}+\sqrt{\frac{2p}{n\mu}}\right),$$

for universal constant $C'>0$. Therefore, the diffusion process (2.3) satisfies the bound

$$\sup_{t\ge 0}\left(\mathbb{E}\left[\|\theta_t-\theta^*\|_2^p\right]\right)^{\frac{1}{p}} \le c\left(\sqrt{\frac{d}{\mu n}}+\frac{B}{\mu n}+\frac{\varepsilon_2(n,\delta)}{\mu}+\sqrt{\frac{p}{n\mu}}\right) \qquad \text{for any } p\ge 1.$$

Combining the above inequality with the inequality (5.3) yields the conclusion of the theorem.

5.1.1. Proof of claim (5.1). For the given choice $\alpha>0$, an application of Itô's formula yields the decomposition

$$\frac{1}{2}e^{\alpha t}\|\theta_t-\theta^*\|_2^2 = -\frac{1}{2}\int_0^t\langle\theta^*-\theta_s,\,\nabla F_n(\theta_s)e^{\alpha s}\rangle ds + \frac{1}{2n}\int_0^t\langle\theta_s-\theta^*,\,\nabla\log\pi(\theta_s)e^{\alpha s}\rangle ds$$
$$+\frac{d}{2n}\int_0^t e^{\alpha s}ds + \frac{1}{\sqrt{n}}\int_0^t e^{\alpha s}\langle\theta_s-\theta^*,\,dB_s\rangle + \frac{1}{2}\int_0^t \alpha e^{\alpha s}\|\theta_s-\theta^*\|_2^2\,ds$$
$$= J_1+J_2+J_3+J_4+J_5. \tag{5.2}$$

We begin by bounding the term J_1 in equation (5.2). Based on Assumption **(S.2)** regarding the perturbation error between F_n and F and the strong convexity of F, we have

$$J_1 = -\frac{1}{2}\int_0^t\langle\theta^*-\theta_s,\,\nabla F_n(\theta_s)e^{\alpha s}\rangle ds$$
$$\le -\frac{1}{2}\int_0^t\langle\theta^*-\theta_s,\,\nabla F(\theta_s)e^{\alpha s}\rangle ds + \frac{1}{2}\int_0^t\|\theta_s-\theta^*\|_2\|\nabla F(\theta_s)-\nabla F_n(\theta_s)\|_2 e^{\alpha s}ds$$
$$\le -\frac{1}{2}\int_0^t\mu\|\theta_s-\theta^*\|_2^2 e^{\alpha s}ds + \frac{1}{2}\int_0^t\|\theta_s-\theta^*\|_2(\varepsilon_1(n,\delta)\|\theta_s-\theta^*\|_2+\varepsilon_2(n,\delta))e^{\alpha s}ds$$
$$\le -\frac{1}{2}\int_0^t\mu\|\theta_s-\theta^*\|_2^2 e^{\alpha s}ds + \frac{1}{2}\int_0^t\|\theta_s-\theta^*\|_2^2(\varepsilon_1(n,\delta)+\mu/3)e^{\alpha s}ds + \frac{3\varepsilon_2^2(n,\delta)}{2\mu}\int_0^t e^{\alpha s}ds.$$

The second term J_2 involving prior π can be controlled in the following way:

$$J_2 = \frac{1}{2n}\int_0^t \langle \theta_s - \theta^*, \nabla \log \pi(\theta_s) e^{\alpha s}\rangle ds \leq \frac{1}{2n}\int_0^t B \left\|\theta_s - \theta^*\right\|_2 e^{\alpha s} ds$$
$$\leq \int_0^t \frac{\mu}{6}\left\|\theta_s - \theta^*\right\|_2^2 e^{\alpha s} ds + \frac{3B^2}{n^2\mu}\int_0^t e^{\alpha s} ds.$$

For the third term J_3, a direct calculation leads to

$$J_3 = \frac{d(e^{\alpha t} - 1)}{2\alpha n}.$$

Moving to the fourth term $J_4 = M_t/\sqrt{n}$, it is a martingale (since M_t is a martingale). Putting the above results together and noting that $\alpha = \frac{1}{2}\mu - \varepsilon_1(n, \delta) > \frac{\mu}{6}$, we obtain the bound

$$\frac{1}{2}e^{\alpha t}\left\|\theta_t - \theta^*\right\|_2^2 \leq \frac{1}{\sqrt{n}}M_t + U_n \frac{(e^{\alpha t} - 1)}{2\alpha}.$$

Putting together the pieces yields the claim (5.1).

5.2. Proof of Theorem 3.2. As in the proof of Theorem 3.1, we omit the conditioning on $\mathcal{F}_n := \sigma(X_1^n)$. For any $p \geq 2$, we define the functions on the positive real line $(0, \infty)$ as

$$\nu_{(p)}(r) := \psi\left(r^{\frac{1}{p-1}}\right) r^{\frac{p-2}{p-1}}, \quad \text{and} \quad \tau_{(p)}\left(r^{p-1}\zeta(r)\right) := r^{p-2}\psi(r).$$

Note that $\tau_{(p)}$ is defined implicitly; let us verify that this definition is meaningful. By Assumption **(W.2)**, the function $r \mapsto r^{p-1}\zeta(r)$ is a strictly increasing and surjective, mapping from $[0, +\infty)$ to $[0, +\infty)$. Therefore, it is invertible, which ensures that the function $\tau_{(p)}$ is well-defined.

Now we claim that for any $p \geq 2$, the functions $\nu_{(p)}$ and $\tau_{(p)}$ are convex and strictly increasing, and that furthermore, the expectation $\mathbb{E}\left[\|\theta_t - \theta^*\|_2^p\right]$ is upper bounded by the integral

$$\frac{p}{2}\int_0^t \left(- R_p(s) + \varepsilon(n, \delta)\tau_{(p)}^{-1}(R_p(s)) + \frac{B}{n}\nu_{(p)}^{-1}(R_p(s)) + \frac{p - 1 + d}{n}\nu_{(p)}^{-1}(R_p(s))^{\frac{p-2}{p-1}}\right) ds, \tag{5.3}$$

where $R_p(s) := \mathbb{E}\left[\|\theta_s - \theta^*\|_2^{p-2}\psi(\|\theta_s - \theta^*\|_2)\right]$.

Taking the above claims as given for the moment, let us now complete the proof of the theorem. Since for each finite $q \geq 1$, the process $(\theta_t : t \geq 0)$ converges in $\mathbb{L}^q$ norm, the limit $\lim_{t\to+\infty} R_p(t)$ exists. Since the functions $\tau_{(p)}$ and $\nu_{(p)}$ are convex and strictly increasing, their inverse functions are concave. Moreover, simple calculation leads to

$$\nabla_r \left(\nu_{(p)}^{-1}(r)^{\frac{p-2}{p-1}}\right) = \frac{p-2}{p-1} \cdot \frac{\nu_{(p)}^{-1}(r)^{-\frac{1}{p-1}}}{\nu'_{(p)}(\nu_{(p)}^{-1}(r))}. \tag{5.4}$$

Since $\nu_{(p)}$ is convex and increasing, the numerator is a decreasing positive function of r. Additionally, the denominator is an increasing positive function of r. Therefore, the derivative in equation (5.4) is a decreasing function of r, and the function $r \mapsto \nu_{(p)}^{-1}(r)^{\frac{p-2}{p-1}}$ is concave. Define the function

$$\phi(r) := -r + \varepsilon(n,\delta)\tau_{(p)}^{-1}(r) + \frac{B}{n}\nu_{(p)}^{-1}(r) + \frac{p-1+d}{n}\nu_{(p)}^{-1}(r)^{\frac{p-2}{p-1}},$$

and observe that ϕ is concave and $\phi(0) = 0$. Let r_* be the smallest positive solution to the equation

$$r = \varepsilon(n,\delta)\tau_{(p)}^{-1}(r) + \frac{B}{n}\nu_{(p)}^{-1}(r) + \frac{p-1+d}{n}\nu_{(p)}^{-1}(r)^{\frac{p-2}{p-1}}.$$

We then have $\phi(r) < 0$ for $r > r_*$ and $\phi(r) > 0$ for $r \in (0, r_*)$. By Lemma C.1, we have $\lim_{t\to+\infty} R_p(t) \le r_*$.

Since $\nu_{(p)}$ is a convex and strictly increasing function, Jensen's inequality implies that

$$R_p(t) = \mathbb{E}\left(\|\theta_t - \theta^*\|_2^{p-2}\,\psi(\|\theta_t - \theta^*\|_2)\right) \ge \nu_{(p)}\left(\mathbb{E}\,\|\theta_t - \theta^*\|_2^{p-1}\right). \tag{5.5}$$

Therefore, if we define $z_* := \lim_{t\to+\infty}\left(\mathbb{E}\,\|\theta_t - \theta^*\|_2^{p-1}\right)^{\frac{1}{p-1}}$, we have $z_*^{p-1} \le \nu_{(p)}^{-1}(r_*)$. Hence, we arrive at the following inequality

$$\begin{aligned} z_*^{p-2}\psi(z_*) &\le \varepsilon(n,\delta)\tau_{(p)}^{-1}\left(\nu_{(p)}(z_*^{p-1})\right) + \frac{B}{n}z_*^{p-1} + \frac{p-1+d}{n}z_*^{p-2} \\ &= \varepsilon(n,\delta)z_*^{p-1}\zeta(z_*) + \frac{B}{n}z_*^{p-1} + \frac{p-1+d}{n}z_*^{p-2}. \end{aligned}$$

As a consequence, we find that

$$\psi(z_*) \le \varepsilon(n,\delta)\zeta(z_*)z_* + \frac{B + (p-1)d}{n}.$$

In Appendix C.4 of the supplementary material [31], we prove the existence and uniqueness of the positive solution to the non-linear equation (3.1). Given this claim, replacing p by $(p+1)$ and putting the above results together yields

$$\lim_{t\to+\infty}\left(\mathbb{E}\left(\|\theta_t - \theta^*\|_2^p\right)\right)^{\frac{1}{p}} \le z_p^*,$$

where z_p^* is the unique positive solution to the following equation:

$$\psi(z) = \varepsilon(n,\delta)\zeta(z)z + \frac{B}{n}z + \frac{p+d}{n}.$$

Combining the above inequality with the inequality (5.3) yields the conclusion of the theorem.

We now return to prove our earlier claims about the behavior of the functions $\nu_{(p)}$, $\tau_{(p)}$, the moment bound (5.3), and the existence of unique positive solution to equation (3.1).

5.2.1. Structure of the function $\nu_{(p)}$.

Since ψ is a convex and strictly increasing function, by taking the second derivative, we find that

$$\nu''_{(p)}(r) = \nabla_r^2 \left(\psi \left(r^{\frac{1}{p-1}} \right) r^{\frac{p-2}{p-1}} \right)$$
$$= \frac{1}{p-1} r^{\frac{1}{p-1}-1} \psi'' \left(r^{\frac{1}{p-1}} \right) + \frac{1}{p-1} r^{-1} \left(\psi' \left(r^{\frac{1}{p-1}} \right) - r^{-\frac{1}{p-1}} \psi \left(r^{\frac{1}{p-1}} \right) \right) \geq 0$$

for all $r > 0$. As a consequence, the function $\nu_{(p)}$ is convex.

5.2.2. Structure of the function $\tau_{(p)}$.

The proof is by calculating the second derivative of the function $\tau_{(p)}$, and we make use of Assumption **(W.3)** on the functions ψ and ζ. Recall that $\tau_{(p)}(r^{p-1}\zeta(r)) = r^{p-2}\psi(r)$ for any $r > 0$. Taking derivatives with respect to r on both sides, we find that

$$\left[(p-1) r^{p-2} \zeta(r) + r^{p-1} \zeta'(r) \right] \tau'_{(p)}(r^{p-1}\zeta(r)) = (p-2) r^{p-3} \psi(r) + r^{p-2} \psi'(r).$$

Under the substitution $z = \zeta_{(p)}(r)$, we find that $\nabla_z \tau_{(p)}(z) = \frac{(p-2)\psi(r) + r\psi'(r)}{(p-1)r\zeta(r) + r^2\zeta'(r)}$.

Taking another derivative of the above term, we find that

$$\nabla_z^2 \tau_{(p)}(z) = \left(\zeta'_{(p)}(r) \right)^{-1} \frac{g(r,p)}{\left((p-1) r \zeta(r) + r^2 \zeta'(r) \right)^2},$$

where we denote

$$g(r,p) := \left[(p-1) r \zeta(r) + r^2 \zeta'(r) \right] \cdot \left[(p-1) \psi'(r) + r \psi''(r) \right]$$
$$- \left[(p-1) \zeta(r) + (p+1) r \zeta'(r) + r^2 \zeta''(r) \right] \cdot \left[(p-2) \psi(r) + r \psi'(r) \right].$$

According to Assumption **(W.3)**, the function $\tau_{(2)} = \psi_{(2)} \circ \zeta_{(2)}^{-1}$ is convex. Therefore, we have $g(r, 2) \geq 0$ for any $r > 0$. Simple algebra with first order derivative of function g with respect to parameter p leads to

$$\begin{aligned}
\nabla_p \left(g(r,p) \right) =& \zeta(r) \cdot \left[(p-1) r \psi'(r) + r^2 \psi''(r) - (p-2) \psi(r) - r \psi'(r) \right] \\
& - r \zeta'(r) \left[(p-2) \psi(r) + r \psi'(r) \right] + r \psi'(r) \cdot \left[(p-1) \zeta(r) + r \zeta'(r) \right] \\
& - \psi(r) \cdot \left[(p-1) \zeta(r) + (p+1) r \zeta'(r) + r^2 \zeta''(r) \right] \\
=& 2(p-2) \left[r \psi'(r) \zeta(r) - \psi(r) \zeta(r) - r \zeta'(r) \psi(r) \right] \\
& + \left[r^2 \zeta(r) \psi''(r) + r \psi'(r) \zeta(r) - 3 \psi(r) \zeta(r) - r^2 \psi(r) \zeta''(r) \right] \geq 0
\end{aligned}$$

for all $r > 0$. Here the last inequality follows from Assumption **(W.3)**. Therefore, the function g is increasing function in terms of p when $p \geq 2$, so that $g(r, p) \geq g(r, 2) \geq 0$ for all $r > 0$. Given this inequality, we have $\frac{d^2}{dz^2} \tau_{(p)}(z) \geq 0$ for any $z \geq 0$, $p \geq 2$, i.e., the function $\tau_{(p)}(z)$ is a convex function for $z = \zeta_{(p)}(r)$.

5.2.3. Proof of claim (5.3). For any $p \geq 2$, an application of Itô's formula yields the bound $\|\theta_t - \theta^*\|_2^p \leq \sum_{j=1}^5 T_j$, where

$$T_1 := -\frac{p}{2}\int_0^t \langle \theta^* - \theta_s,\, \nabla F(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds, \tag{5.6a}$$

$$T_2 := \frac{p}{2}\int_0^t \langle \theta^* - \theta_s,\, \nabla F(\theta_s) - \nabla F_n(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds \tag{5.6b}$$

$$T_3 := \frac{p}{2n}\int_0^t \langle \theta_s - \theta^*,\, \nabla \log \pi(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds \tag{5.6c}$$

$$T_4 := p\int_0^t \|\theta_s - \theta^*\|_2^{p-2}\, \langle \theta_s - \theta^*,\, dB_s\rangle \tag{5.6d}$$

$$T_5 := \frac{p(p-1+d)}{2n}\int_0^t \|\theta_s - \theta^*\|_2^{p-2}\, ds. \tag{5.6e}$$

We now upper bound the terms $\{T_j\}_{j=1}^5$ in terms of functionals of the quantity R_p. From the weak convexity of F guaranteed by Assumption W.1, we have

$$\mathbb{E}\,[T_1] = -\frac{p}{2}\mathbb{E}\left[\int_0^t \langle \theta^* - \theta_s,\, \nabla F(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds\right] \leq -\frac{p}{2}\int_0^t R_p(s) ds. \tag{5.7a}$$

Based on Assumption **(W.2)**, we find that

$$\begin{aligned}\mathbb{E}\,[T_2] &= \frac{p}{2}\mathbb{E}\left[\int_0^t \langle \theta^* - \theta_s,\, \nabla F(\theta_s) - \nabla F_n(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds\right] \\ &\leq \frac{p}{2}\varepsilon(n,\delta)\int_0^t \mathbb{E}\left[\|\theta_s - \theta^*\|_2^{p-1}\,\zeta(\|\theta_s - \theta^*\|_2)\right] ds.\end{aligned}$$

Since the function $\tau_{(p)}$ is convex, invoking Jensen's inequality, we obtain the following inequalities:

$$\begin{aligned}\int_0^t \mathbb{E}\left[\|\theta_s - \theta^*\|_2^{p-1}\,\zeta\left(\|\theta_s - \theta^*\|_2\right)\right] ds &\leq \int_0^t \tau_{(p)}^{-1}\mathbb{E}\left[\tau_{(p)}\left(\|\theta_s - \theta^*\|_2^{p-1}\,\zeta(\|\theta_s - \theta^*\|_2)\right)\right] ds \\ &= \int_0^t \tau_{(p)}^{-1}\left(R_p(s)\right) ds.\end{aligned}$$

In light of the above inequalities, we have

$$\mathbb{E}\,[T_2] \leq \frac{p}{2}\varepsilon(n,\delta)\int_0^t \tau_{(p)}^{-1}\left(R_p(s)\right) ds. \tag{5.7b}$$

Moving to T_3 in equation (5.6c), given Assumption **(A)** which controls the growth of prior distribution π, its expectation is bounded as

$$\begin{aligned}\mathbb{E}\,[T_3] &= \frac{p}{2n}\mathbb{E}\left[\int_0^t \langle \theta_s - \theta^*,\, \nabla \log \pi(\theta_s)\rangle \,\|\theta_s - \theta^*\|_2^{p-2}\, ds\right] \\ &\leq \frac{pB}{2n}\int_0^t \mathbb{E}\left[\|\theta_s - \theta^*\|_2^{p-1}\right] ds.\end{aligned} \tag{5.7c}$$

By exploiting the bound (5.5) along with the fact that $\nu_{(p)}$ is strictly increasing on $[0, +\infty)$, we find that

$$\int_0^t \mathbb{E}\left(\|\theta_s - \theta^*\|_2^{p-1}\right) ds \leq \int_0^t \nu_{(p)}^{-1}\left(R_p(s)\right) ds. \tag{5.7d}$$

Combining the inequalities (5.7c) and (5.7d), we have

$$\mathbb{E}\left[T_3\right] \leq \frac{pB}{2n} \int_0^t \nu_{(p)}^{-1}\left(R_p(s)\right) ds. \tag{5.7e}$$

Moving to the fourth term T_4 from equation (5.6d), we have

$$\mathbb{E}\left[T_4\right] = \mathbb{E}\left[\int_0^t \|\theta_s - \theta^*\|_2^{p-2} \langle \theta_s - \theta^*, dB_s\rangle\right] = 0, \tag{5.7f}$$

where we have used the martingale structure.

For the last term T_5, invoking Hölder's inequality and the bound (5.5), we have the moment estimate:

$$\mathbb{E}\left(\|\theta_s - \theta^*\|_2^{p-2}\right) \leq \left(\mathbb{E}\left[\|\theta_s - \theta^*\|_2^{p-1}\right]\right)^{\frac{p-2}{p-1}} \leq \nu_{(p)}^{-1}\left(R_p(s)\right)^{\frac{p-2}{p-1}}.$$

Consequently, the term T_5 can be bounded in expectation as

$$\mathbb{E}\left[T_5\right] \leq \frac{p(p-1+d)}{2n} \int_0^t \nu_{(p)}^{-1}\left(R_p(s)\right)^{\frac{p-2}{p-1}} ds. \tag{5.7g}$$

Collecting the bounds on the expectations of the terms $\{T_j\}_{j=1}^5$ from equations (5.7a)-(5.7g), respectively, yields the claim (5.3).

6. Discussion.

In this paper, we described an approach for analyzing the posterior contraction rates of parameters based on the diffusion processes. Our theory depends on two important features: the local growth of the population log-likelihood function F and stochastic perturbation bounds between the gradient of F and the gradient of its sample counterpart F_n. For strongly concave log-likelihood functions, we established posterior convergence rates for parameter estimation of the order $(d/n)^{1/2}$, valid under appropriate conditions on the perturbation error between ∇F_n and ∇F and sharp sample size requirements. On the other hand, when the population log-likelihood function is weakly concave, our analysis shows that convergence rates are more delicate: they depend on an interaction between the degree of weak convexity, and the stochastic error bounds. In this setting, we proved that the posterior convergence rate of parameter is upper bounded by the unique positive solution of a non-linear equation determined by the previous interplay. Compared to the convergence rate of MLE, the bound contains an additional term capturing the diffusive behavior of the posterior distribution. Finally, we demonstrated the utility of the diffusion process approach by deriving non-asymptotic forms of Bernstein–von Mises results for models with non-degenerate Fisher information.

Let us now discuss a few directions that arise naturally from our work. First, in the weakly convex setting, although we have established non-asymptotic posterior contraction bounds, the current results do not provide information on the shape of the asymptotic posterior distribution. For example, when F is locally strongly concave around θ^*, it is well-known from the Bernstein–von Mises theorem that the posterior distribution of parameter converges to a multivariate normal distribution centered at the maximum likelihood estimation (MLE) with the covariance matrix is given by $1/\left(nI(\theta^*)\right)$ (e.g., see the book [51], Chapter 10.2), where $I(\theta^*)$ denotes the Fisher information matrix at θ^*. When the F is only weakly concave, the Fisher information matrix $I(\theta^*)$ is degenerate, so that the posterior distribution can no longer be approximated by a multivariate Gaussian distribution. It is interesting to consider how the diffusion approach might provide insight into the shape of the posterior in this setting.

Second, the contraction rates given in this paper can give information about the over-specification of the latent variable models, thereby having potential applications for model selection. As a concrete example, for the symmetric two-component Gaussian mixture model example discussed in Subsection 4.2, the posterior distribution concentrates around $\theta^* = 0$ at a rate $O\big((d/n)^{1/4}\big)$ in the over-specified case. On the other hand, for a non-degenerate mixture with symmetric modes at θ^* and $-\theta^*$ (with $\theta^* \neq 0$), it concentrates at the usual rate $O\big((d/n)^{1/2}\big)$. Consequently, the degree of dispersion in the posterior serves as an indicator of over-specification. Furthermore, since our results are non-asymptotic, they also give guidance on how this procedure could be performed with finite sample size n. Finally, whereas this paper focused on posterior contraction for parametric models, we suspect that the diffusion process approach used here might also be fruitfully applied to non-parametric models.

Acknowledgements. This work was partially supported by NSF grant CCF-1955450 and ONR grant N00014-21-1-2842 to MJW¿

REFERENCES

[1] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin, *A simple proof of the Poincaré inequality for a large class of probability measures*, Electronic Communications in Probability, 13 (2008), pp. 60–66.

[2] S. Balakrishnan, M. J. Wainwright, and B. Yu, *Statistical guarantees for the EM algorithm: From population to sample-based analysis*, Annals of Statistics, 45 (2017), pp. 77–120.

[3] A. Barron, M. Schervish, and L. Wasserman, *The consistency of posterior distributions in nonparametric problems*, Ann. Statist, 27 (1999), pp. 536–561.

[4] A. Bhattacharya, D. Pati, , and D. B. Dunson, *Anisotropic function estimation using multi-bandwidth Gaussian processes*, Annals of Statistics, 42 (2014), pp. 352–381.

[5] J. Chen, *Optimal rate of convergence for finite mixture models*, Annals of Statistics, 23 (1995), pp. 221–233.

[6] S. Chrétien, M. Cucuringu, G. Lecué, and L. Neirac, *Learning with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature*, Journal of Machine Learning Research, 22 (2021).

[7] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79 (2017), pp. 651–676.

[8] R. de Jonge and J. H. van Zanten, *Adaptive nonparametric Bayesian inference using location-scale mixture priors*, Annals of Statistics, 38 (2010), pp. 3300–3320.

[9] J. L. Doob, *Application of the theory of martingales*, Actes du ColloqueInternational Le Calcul des Probabilités et ses applications (Lyon, 28 Juin– 3 Juillet, 1948), (1949), pp. 23–27.

[10] A. Durmus and E. Moulines, *Nonasymptotic convergence analysis for the unadjusted Langevin algorithm*, The Annals of Applied Probability, 27 (2017), pp. 1551–1587.

[11] A. Durmus and E. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2019), pp. 2854–2882.
[12] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, *Singularity, misspecification, and the convergence rate of EM*, arXiv preprint arXiv:1810.00828, (2018).
[13] D. A. Freedman, *On the asymptotic behavior of Bayes' estimates in the discrete case*, Annals of Statistics, 34 (1963), p. 1386–1403.
[14] D. A. Freedman, *On the asymptotic behavior of Bayes' estimates in the discrete case.II*, Annals of Statistics, 36 (1965), p. 454–456.
[15] C. Gao and H. H. Zhou, *Rate exact Bayesian adaptation with modified block priors*, Annals of Statistics, 44 (2016), pp. 318–345.
[16] S. Ghosal, J. K. Ghosh, and A. van der Vaart, *Convergence rates of posterior distributions*, Annals of Statistics, 28 (2000), pp. 500–531.
[17] S. Ghosal and A. van der Vaart, *Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities*, Annals of Statistics, 29 (2001), pp. 1233–1263.
[18] S. Ghosal and A. van der Vaart, *Posterior convergence rates of Dirichlet mixtures at smooth densities*, Annals of Statistics, 35 (2007), pp. 697–723.
[19] L. Gross, *Logarithmic Sobolev inequalities*, American Journal of Mathematics, 97 (1975), pp. 1061–1083.
[20] P. Heinrich and J. Kahn, *Strong identifiability and optimal minimax rates for finite mixture estimation*, Annals of Statistics, 46 (2018), pp. 2844–2870.
[21] N. Ho, K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu, *Instability, computational efficiency and statistical accuracy*, arXiv preprint arXiv:2005.11411, (2020).
[22] H. Ishwaran, L. F. James, and J. Sun, *Bayesian model selection in finite mixtures by marginal density decompositions*, Journal of the American Statistical Association, 96 (2001), pp. 1316–1332.
[23] B. J. Kleijn and A. W. van der Vaart, *The Bernstein-von-Mises theorem under misspecification*, Electronic Journal of Statistics, 6 (2012), pp. 354–381.
[24] B. J. K. Kleijn and A. W. van der Vaart, *Misspecification in infinite-dimensional Bayesian statistics*, Annals of Statistics, 34 (2006), pp. 837–877.
[25] B. T. Knapik, A. W. Van Der Vaart, and J. H. van Zanten, *Bayesian inverse problems with Gaussian priors*, The Annals of Statistics, 39 (2011), pp. 2626–2657.
[26] B. Lindsay, *Mixture Models: Theory, Geometry and Applications*, In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995.
[27] P.-L. Loh and M. J. Wainwright, *Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Advances in Neural Information Processing Systems, 26 (2013).
[28] C. Ma, K. Wang, Y. Chi, and Y. Chen, *Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*, Foundations of Computational Mathematics, 20 (2020), pp. 451–632.
[29] E. Mazumdar, A. Pacchiano, Y. Ma, M. Jordan, and P. Bartlett, *On approximate Thompson sampling with Langevin algorithms*, in International Conference on Machine Learning, PMLR, 2020, pp. 6797–6807.
[30] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall/CRC, 1989.
[31] W. Mou, N. Ho, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, *Supplementary material to "A Diffusion Process Perspective on Posterior Contraction Rates for Parameters"*, 2023. DOI: [COMPLETED BY TYPESETTER].
[32] X. Nguyen, *Convergence of latent mixing measures in finite and infinite mixture models*, Annals of Statistics, 4 (2013), pp. 370–400.
[33] R. Nickl, *Bayesian non-linear statistical inverse problems*, Lecture Notes ETH Zurich, (2022).
[34] D. M. Ostrovskii and F. Bach, *Finite-sample analysis of M-estimators using self-concordance*, Electronic Journal of Statistics, 15 (2021), pp. 326–391.
[35] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM journal on control and optimization, 30 (1992), pp. 838–855.
[36] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, vol. 293, Springer-Verlag, third ed., 1999.
[37] H. Risken, *The Fokker-Planck Equation*, Springer, 1996.
[38] J. Rousseau, *Rates of convergence for the posterior distributions of mixtures of Beta and adaptive*

nonparametric estimation of the density, Annals of Statistics, 38 (2010), pp. 146–180.

[39] J. Rousseau and K. Mengersen, *Asymptotic behaviour of the posterior distribution in overfitted mixture models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 689–710.

[40] L. Schwartz, *On Bayes procedures*, Zeitschrift fűr Wahrscheinlichkeitstheorie und Verwandte Gebiete, 4 (1965), pp. 10–26.

[41] W. Shen, S. R. Tokdar, , and S. Ghosal, *Adaptive Bayesian multivariate density estimation with Dirichlet mixtures*, Biometrika, 100 (2013), p. 623–640.

[42] X. Shen and L. Wasserman, *Rates of convergence of posterior distributions.*, Annals of Statistics, 29 (2001), pp. 687–714.

[43] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Mathematical Programming, (2021), pp. 1–70.

[44] V. Spokoiny, *Parametric estimation. finite sample theory*, The Annals of Statistics, 40 (2012), pp. 2877–2909.

[45] W. Su, S. Boyd, and E. J. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, Journal of Machine Learning Research, 17 (2016), pp. 1–43.

[46] M. Talagrand, *Transportation cost for Gaussian and other product measures*, Geometric & Functional Analysis GAFA, 6 (1996), pp. 587–600.

[47] D. Talay, *Second-order discretization schemes of stochastic differential systems for the computation of the invariant law*, Stochastics: An International Journal of Probability and Stochastic Processes, 29 (1990), pp. 13–36.

[48] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic analysis and applications, 8 (1990), pp. 483–509.

[49] S. van de Geer, *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[50] S. van de Geer, *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[51] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[52] A. W. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, NY, 1996.

[53] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.

[54] S. Walker, *On sufficient conditions for Bayesian consistency*, Annals of Statistics, 90 (2003), pp. 482–488.

[55] S. Walker, *New approaches to Bayesian consistency*, Annals of Statistics, 32 (2004), pp. 2028–2043.

[56] S. G. Walker, A. Lijoi, and I. Prunster, *On rates of convergence for posterior distributions in infinite-dimensional models*, Annals of Statistics, 35 (2007), pp. 738–746.

[57] Y. Yang and D. B. Dunson, *Bayesian manifold regression*, Annals of Statistics, 44 (2016), pp. 876–905.

[58] Y. Yang and S. T. Tokdar, *Minimax-optimal nonparametric regression in high dimensions*, Annals of Statistics, 43 (2015), pp. 652–674.

SUPPLEMENTARY MATERIALS: A Diffusion Process Perspective on Posterior Contraction Rates for Parameters

Wenlong Mou*, Nhat Ho†, Martin Wainwright‡, Peter Bartlett‡, and Michael Jordan‡

This supplementary material is devoted to the proofs deferred from the main paper. In Appendix A, we present the proofs of non-asymptotic Bernstein–von Mises theorems using tools from diffusion process theory. The proofs of our main corollaries are given in Appendix B, whereas Appendix C is devoted to the proofs of auxiliary results.

Appendix A. Proofs of non-asymptotic Bernstein–von Mises results.

In this section, we collect the proofs of Theorem 3.5 and Proposition 3.4.

A.1. Proof of Theorem 3.5.

For any fixed $T > 0$, we define the sequence of potential functions $\Phi_t : \mathbb{R}^d \to \mathbb{R}$

$$\Phi_t(\theta) := \langle \theta - \widehat{\theta}^{(n)}),\, H^* e^{H^*(t-T)}(\theta - \widehat{\theta}^{(n)}) \rangle, \quad \text{for each } t \in [0, T].$$

Once again, we consider the diffusion process

$$d\theta_t = -\nabla F_n(\theta_t)dt + \frac{1}{n}\nabla \log \pi(\theta_t)dt + dB_t,$$

with the initial condition $\theta_0 = \widehat{\theta}^{(n)}$. Using Itô's formula, for $t \in [0, T]$, we have

$$\begin{aligned}
\Phi_t(\theta_t) &= \int_0^t \frac{\partial \Phi_s}{\partial s}(\theta_s)ds - \int_0^t \langle \nabla \Phi_s(\theta_s),\, \nabla F_n(\theta_s) - \frac{\nabla \log \pi(\theta_s)}{n} \rangle ds \\
&\quad + \sqrt{\frac{2}{n}} \int_0^t \langle \nabla \Phi_s(\theta_s),\, dB_s \rangle + \frac{1}{n}\int_0^t \Delta \Phi_s(\theta_s)ds \\
&= \underbrace{\int_0^t \left(H^*(\theta_s - \widehat{\theta}^{(n)}) - \nabla F_n(\theta_s) + \frac{\nabla \log \pi(\theta_s)}{n} \right)^\top H^* e^{H^*(s-T)}(\theta_s - \widehat{\theta}^{(n)})ds}_{:= I_1(t)} \\
&\quad + \underbrace{\sqrt{\frac{2}{n}} \int_0^t (\theta_s - \widehat{\theta}^{(n)})^\top H^* e^{(s-T)H^*} dB_s}_{I_2(t)} + \underbrace{\frac{1}{n}\int_0^t \mathrm{Tr}\left(H^* e^{H^*(s-T)} \right) ds}_{I_3(t)}.
\end{aligned} \tag{A.1}$$

Note that the matrices H^* and $e^{(s-T)H^*}$ commute, so that we may write their product in an arbitrary order.

Defining the linearization error

$$\Delta_s := (A + \varepsilon_1^{(2)}(n, \delta)) \left(\|\theta_s - \theta^*\|_2 + \left\| \widehat{\theta}^{(n)} - \theta^* \right\|_2 \right) + \varepsilon_2^{(2)}(n, \delta) + \frac{L_2}{n},$$

*Department of EECS, UC Berkeley.
†Department of Statistics and Data Science, UT Austin.
‡Department of EECS and Department of Statistics, UC Berkeley.

we claim that the following bounds hold for each $t \in [0, T]$:

(A.2a)
$$I_1(t) \leq \tfrac{2+\log\kappa(H^*)}{a} \sup_{0 \leq s \leq t} \Phi_s(\theta_s) + a \int_0^t \Delta_s^2 \left(\|\theta_s - \theta^*\|_2^2 + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^2 \right) e^{-\frac{\lambda_{\min}(H^*)}{2}(s-T)} ds,$$

(A.2b)
$$\left(\mathbb{E} \sup_{0 \leq t \leq T} |I_2(t)|^p \right)^{1/p} \leq c \sqrt{\frac{p\,(1 + \log\kappa(H^*))}{n}} \left(\mathbb{E} \sup_{0 \leq t \leq T} \Phi_t(\theta_t)^{p/2} \right)^{1/p}, \quad \text{and}$$

(A.2c)
$$I_3(t) \leq \frac{d}{n}.$$

Here $c > 0$ is an universal constant. We prove all of these bounds in the subsections to follow.

Taking these bounds as given for the moment, let us complete the proof of the theorem. By Jensen's inequality, for an even integer $p \geq 2$, the moments of the integral term in equation (A.2a) can be bounded as

(A.3)
$$\mathbb{E} \left(\int_0^T \Delta_s^2 \left(\|\theta_s - \theta^*\|_2^2 + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^2 \right) e^{-\frac{\lambda_{\min}(H^*)}{2}(s-T)} ds \right)^p$$
$$\leq \left(\frac{c}{\lambda_{\min}(H^*)} \right)^{p-1} \cdot \mathbb{E} \int_0^T \Delta_s^{2p} \left(\|\theta_s - \theta^*\|_2^{2p} + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^{2p} \right) e^{-\frac{\lambda_{\min}(H^*)}{2}(s-T)} ds,$$

for a universal constant $c > 0$.

For any $\omega \in (0, 1)$, by taking supremum on both sides of the decomposition (A.1), combining with the bounds (A.2a) and (A.2c), and taking $a = c\frac{2+\log\kappa(H^*)}{\omega}$, we arrive at the inequality

$$\sup_{0 \leq t \leq T} \Phi_t(\theta_t) \leq (1 + \omega) \left(\frac{d}{n} + \sup_{0 \leq t \leq T} I_2(t) \right)$$
$$+ \frac{c(2 + \log\kappa(H^*))}{\omega} \int_0^T \Delta_t^2 \left(\|\theta_t - \theta^*\|_2^2 + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^2 \right) e^{-\frac{\lambda_{\min}(H^*)}{2}(t-T)} dt.$$

Taking p-th moment on both sides of the inequality, combining with the bounds (A.2b) and (A.3), and applying Minkowski's inequality, we arrive at the bound

$$\left(\mathbb{E} \sup_{0 \leq t \leq T} \Phi_t(\theta_t)^p \right)^{1/p} \leq (1 + \omega)\frac{d}{n} + \sqrt{\frac{cp(1 + \log\kappa(H^*))}{n}} \cdot \left(\mathbb{E} \sup_{0 \leq t \leq T} \Phi_t(\theta_t)^p \right)^{\frac{1}{2p}}$$
$$+ \frac{c(2 + \log\kappa(H^*))}{\omega\lambda_{\min}(H^*)} \left(\sup_{0 \leq t \leq T} \mathbb{E} \left[\Delta_t^{2p} \left(\|\theta_t - \theta^*\|_2^{2p} + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^{2p} \right) \right] \right)^{1/p}.$$

Substituting with the definition of the last term, and applying Young's inequality, we find that

$$\left(\mathbb{E} \sup_{0 \leq t \leq T} \Phi_t(\theta_t)^p \right)^{1/p} \leq (1 + \omega)\frac{d}{n} + c\frac{1 + \log\kappa(H^*)}{\omega} \left(\frac{p}{n} + \frac{\mathcal{H}_n(p, \delta)}{\lambda_{\min}(H^*)} \right),$$

where the high-order term $\mathcal{H}_n(p,\delta)$ is defined as

$$\begin{aligned}\mathcal{H}_n(p,\delta) :=& (A+\varepsilon_1^{(2)}(n,\delta))^2 \left(\mathbb{E}_{\mathbb{Q}} \left\|\theta-\theta^*\right\|_2^{4p}\right)^{1/p} \\ &+ \left\|\widehat{\theta}^{(n)}-\theta^*\right\|_2^2 \left(\varepsilon_2^{(2)}(n,\delta)^2 + \frac{L_2^2}{n^2} + (A+\varepsilon_1^{(2)}(n,\delta))^2 \left\|\widehat{\theta}^{(n)}-\theta^*\right\|_2^2\right).\end{aligned}$$

Putting together the pieces yields the conclusion of the theorem.

A.1.1. Proof of claim (A.2a). We first bound the term $I_1(t)$. Noting the defining identity $\nabla F_n(\widehat{\theta}^{(n)}) + \frac{1}{n}\nabla \log \pi(\widehat{\theta}^{(n)}) = 0$, we have the following bound:

$$\begin{aligned}&\left\|H^*(\theta_s - \widehat{\theta}^{(n)}) - \nabla F_n(\theta_s) + \nabla \log \pi(\theta_s)/n\right\|_2 \\ &= \left\|\int_0^1 \left(H^* - \nabla^2 F_n\big(\gamma\theta_s + (1-\gamma)\widehat{\theta}^{(n)}\big) + \nabla^2 \log \pi\big(\gamma\theta_s + (1-\gamma)\widehat{\theta}^{(n)}\big)/n\right)(\theta_s - \widehat{\theta}^{(n)})d\gamma\right\|_2 \\ &\leq \int_0^1 \left\|H^* - \nabla^2 F_n\big(\gamma\theta_s + (1-\gamma)\widehat{\theta}^{(n)}\big) + \nabla^2 \log \pi\big(\gamma\theta_s + (1-\gamma)\widehat{\theta}^{(n)}\big)/n\right\|_{\mathrm{op}} \cdot \left\|\theta_s - \widehat{\theta}^{(n)}\right\|_2 d\gamma.\end{aligned}$$

By Assumptions **(BvM.1)**, **(BvM.2)**, and (PS), for any $\theta \in \mathbb{R}^d$, we have the bound

$$\begin{aligned}&\left\|H^* - \nabla^2 F_n(\theta) + \nabla^2 \log \pi(\theta)/n\right\|_{\mathrm{op}} \\ &\qquad \leq \left\|H^* - \nabla^2 F(\theta)\right\|_{\mathrm{op}} + \left\|\nabla^2 F(\theta) - \nabla^2 F_n(\theta)\right\|_{\mathrm{op}} + \left\|\nabla^2 \log \pi(\theta)/n\right\|_{\mathrm{op}} \\ &\qquad \leq A\left\|\theta-\theta^*\right\|_2 + \varepsilon_1^{(2)}(n,\delta)\left\|\theta-\theta^*\right\|_2 + \varepsilon_2^{(2)}(n,\delta) + \frac{L_2}{n}.\end{aligned}$$

Substituting into the bound for $I_1(t)$, for any $a>0$, we have that

$$\begin{aligned}I_1(t) \leq& \int_0^t \left\|(H^*)^{1/2} e^{H^*(s-t)/2}\right\|_{\mathrm{op}} \\ &\quad \times \left\|H^* - \nabla^2 F_n(\theta_s) + \nabla^2 \log \pi(\theta_s)/n\right\|_2 \left\|\theta_s - \widehat{\theta}^{(n)}\right\|_2 \sqrt{\Phi_s(\theta_s)}ds \\ \leq& a^{-1} \sup_{0\leq s\leq t} \Phi_s(\theta_s) \cdot \int_0^t \left\|(H^*)^{1/2} e^{H^*(s-T)/4}\right\|_{\mathrm{op}}^2 ds \\ &\quad + a\int_0^t \left\|H^* - \nabla^2 F_n(\theta_s)^2 + \nabla^2 \log \pi(\theta_s)/n\right\|_{\mathrm{op}}^2 \cdot \left\|\theta_s - \widehat{\theta}^{(n)}\right\|_2^2 \left\|e^{H^*(s-T)/4}\right\|_{\mathrm{op}}^2 ds \\ \leq& \frac{2+\log\kappa(H^*)}{a} \sup_{0\leq s\leq t} \Phi_s(\theta_s) \\ &\quad + a\int_0^t \Delta_s^2 \left(\left\|\theta_s - \theta^*\right\|_2^2 + \left\|\widehat{\theta}^{(n)} - \theta^*\right\|_2^2\right) e^{-\frac{\lambda_{\min}(H^*)}{2}(s-T)}ds.\end{aligned}$$

Therefore, claim (A.2a) follows.

A.1.2. Proof of claim (A.2b). Note that $I_2(t)$ is a martingale with respect to the Brownian filtration. Applying the Burkholder-Gundy-Davis inequality for an arbitrary $p \geq 2$ yields

$$\begin{aligned}
\left(\mathbb{E} \sup_{0\leq t\leq T} |I_2(t)|^p\right)^{1/p} &\leq c\sqrt{\frac{p}{n}}\left(\mathbb{E}\left(\int_0^T \left\|H^* e^{(t-T)H^*}(\theta_t - \widehat{\theta}^{(n)})\right\|_2^2 dt\right)^{\frac{p}{2}}\right)^{1/p} \\
&\leq C\sqrt{\frac{p}{n}}\left(\mathbb{E}\left(\int_0^T \|(H^*)^{1/2} e^{\frac{t-T}{2}H^*}\|_{\mathrm{op}}^2 \Phi_t(\theta_t) dt\right)^{\frac{p}{2}}\right)^{1/p} \\
&\leq c\sqrt{\frac{p}{n}}\left(\mathbb{E} \sup_{0\leq t\leq T} \Phi_t(\theta_t)^{p/2}\right)^{1/p} \cdot \sqrt{\int_0^T \|(H^*)^{1/2} e^{\frac{t-T}{2}H^*}\|_{\mathrm{op}}^2 dt}.
\end{aligned}$$

We now observe that

$$\|(H^*)^{1/2} e^{\frac{t-T}{2}H^*}\|_{\mathrm{op}}^2 = \|H^* e^{(t-T)H^*}\|_{\mathrm{op}} = \max_{i\in[d]}\left(\lambda_i(H^*) e^{(t-T)\lambda_i(H^*)}\right).$$

Taking the time integral leads to the bound

$$\begin{aligned}
\int_0^T \|(H^*)^{1/2} e^{\frac{t-T}{2}H^*}\|_{\mathrm{op}}^2 dt &\leq \int_0^{+\infty} \max_{i\in[d]}\left(\lambda_i(H^*) e^{-t\lambda_i(H^*)}\right) dt \\
&\leq \underbrace{\int_0^{+\infty} \max_{\lambda_{\min}(H^*)\leq\lambda\leq\lambda_{\max}(H^*)}\left(\lambda e^{-t\lambda}\right) dt}_{=:J}.
\end{aligned}$$

We now split the integral J into three parts, thereby obtaining

$$\begin{aligned}
J &\leq \int_0^{\lambda_{\max}(H^*)^{-1}} \lambda_{\max}(H^*) e^{-t\lambda_{\max}(H^*)} dt \\
&\quad + \int_{\lambda_{\max}(H^*)^{-1}}^{\lambda_{\min}(H^*)^{-1}} \frac{dt}{et} + \int_{\lambda_{\min}(H^*)^{-1}}^{+\infty} \lambda_{\min}(H^*) e^{-t\lambda_{\min}(H^*)} dt \\
&\leq 1 + \frac{1}{e}\log\frac{\lambda_{\max}(H^*)}{\lambda_{\min}(H^*)}. \qquad \text{(A.4)}
\end{aligned}$$

Denote $\kappa(M) := \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$ for a positive definite matrix M. Collecting the above inequalities, we find that the term $I_2(t)$ is upper bounded as

$$\left(\mathbb{E} \sup_{0\leq t\leq T} |I_2(t)|^p\right)^{1/p} \leq c\sqrt{\frac{p\,(1+\log\kappa(H^*))}{n}}\left(\mathbb{E} \sup_{0\leq t\leq T} \Phi_t(\theta_t)^{p/2}\right)^{1/p}$$

for a universal constant $c > 0$. This completes the proof of the claim (A.2b).

A.1.3. Proof of claim (A.2c). Finally, the term $I_3(t)$ is straightforward to upper bound as

$$I_3(t) \leq \frac{1}{n}\mathrm{Tr}\left(H^* \int_0^T e^{H^*(s-T)} ds\right) \leq \frac{1}{n}\mathrm{Tr}\left(H^* \int_0^{+\infty} e^{-sH^*} ds\right) = \frac{d}{n},$$

which establishes the claim (A.2c).

A.2. Proof of Proposition 3.4. We introduce the shorthand $\mu := \mathcal{N}\big(\widehat{\theta}^{(n)}, (nH^*)^{-1}\big)$ for the target density. Since $H^* \succ 0$, the Gaussian log-Sobolev inequality implies that

$$D_{\mathrm{KL}}(\mathbb{Q}(\cdot \mid X_1^n) \parallel \mu) \le \frac{1}{n\lambda_{\min}(H^*)} \int_{\mathbb{R}^d} \|\nabla \log \mathbb{Q}(\theta \mid X_1^n) - \nabla \log \mu(\theta)\|_2^2 \; \mathbb{Q}(d\theta \mid X_1^n). \tag{A.5}$$

Since μ is a Gaussian density, we find that

$$\nabla \log \mu(\theta) = -nH^*(\theta - \widehat{\theta}^{(n)}).$$

For the posterior density $\mathbb{Q}(\cdot \mid X_1^n)$, we note that

$$\begin{aligned}\nabla \log \mathbb{Q}(\theta|X_1^n) &= -n\nabla F_n(\theta) + \nabla \log \pi(\theta) \\ &= \int_0^1 \Big(-n\nabla^2 F_n(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)}) + \nabla^2 \log \pi(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)})\Big) \\ &\qquad\qquad \times (\theta - \widehat{\theta}^{(n)}) d\gamma.\end{aligned}$$

Putting the above equations together yields

$$\begin{aligned}&\|\nabla \log \mathbb{Q}(\theta \mid X_1^n) - \nabla \log \mu(\theta)\|_2 \\ &\le n \int_0^1 \|\nabla^2 F_n(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)}) - H^* + \nabla^2 \log \pi(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)})/n\|_{\mathrm{op}} \cdot \Big\|\theta - \widehat{\theta}^{(n)}\Big\|_2 d\gamma.\end{aligned}$$

By Assumptions **(BvM.1)**, **(BvM.2)**, and (PS), we have the bounds

$$\begin{aligned}&\|\nabla^2 F_n(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)}) + \nabla^2 \log \pi(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)})/n - H^*\|_{\mathrm{op}} \\ &\le \|\nabla^2 F(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)}) - H^*\|_{\mathrm{op}} \\ &\qquad + \|\nabla^2 F_n(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)}) - \nabla^2 F_n(\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)})\|_{\mathrm{op}} + \frac{L_2}{n} \\ &\le A \Big\|\gamma\theta + (1-\gamma)\widehat{\theta}^{(n)} - \theta^*\Big\|_2 + \varepsilon_1^{(2)}(n,\delta) \Big\|\theta - \widehat{\theta}^{(n)}\Big\|_2 + \varepsilon_2^{(2)}(n,\delta) + \frac{L_2}{n}.\end{aligned}$$

Substituting this bound into the bound (A.5) yields

$$\begin{aligned}D_{\mathrm{KL}}(\mathbb{Q}(\cdot \mid X_1^n) \parallel \mu) \le \frac{n}{\lambda_{\min}(H^*)} &\Big(A \cdot \mathbb{E}_{\mathbb{Q}}\left[\|\theta - \theta^*\|_2^4 \mid X_1^n\right] + A \Big\|\widehat{\theta}^{(n)} - \theta^*\Big\|_2^4\Big) \\ &+ \frac{n\varepsilon_1^{(2)}(n,\delta)}{\lambda_{\min}(H^*)} \mathbb{E}_{\mathbb{Q}}\left[\Big\|\theta - \widehat{\theta}^{(n)}\Big\|_2^3 \mid X_1^n\right] \\ &+ \big(\varepsilon_2^{(2)}(n,\delta) + L_2/n\big) \cdot \mathbb{E}\left[\Big\|\theta - \widehat{\theta}^{(n)}\Big\|_2^2 \mid X_1^n\right].\end{aligned}$$

As a consequence, we obtain the conclusion of the proposition.

Appendix B. Proofs of corollaries. In this appendix, we collect the proofs of several corollaries stated in the main text and Section 4. The crux of the proofs of these corollaries involves a verification of assumptions to invoke the respective theorems. Note that the values of universal constants may change from line to line.

B.1. Proof of Corollary 4.1. We begin by verifying claim (4.2a) about the structure of the negative population log-likelihood function F^R and claim (4.2b) about the uniform perturbation error between ∇F^R and ∇F_n^R.

B.1.1. Proof of claim (4.2a). Following some algebra, we find that

$$\begin{aligned}
-F^R(\theta) &= \mathbb{E}\left[-Y\log\left(1+e^{-\langle X,\theta\rangle}\right)-(1-Y)\log\left(1+e^{\langle X,\theta\rangle}\right)\right]\\
&= -\mathbb{E}\left[\frac{1}{1+e^{-\langle X,\theta^*\rangle}}\log\left(1+e^{-\langle X,\theta\rangle}\right)+\frac{1}{1+e^{\langle X,\theta^*\rangle}}\log\left(1+e^{\langle X,\theta\rangle}\right)\right],
\end{aligned}$$

where the above expectations are taken with respect to $X\sim\mathcal{N}(0,\sigma^2 I_d)$ and $Y|X$ following probability distribution generated from logistic model (4.1). Taking the derivative of F^R with respect to θ yields

$$\begin{aligned}
&\langle \nabla F^R(\theta),\, \theta^*-\theta\rangle\\
&\qquad = \mathbb{E}\left[\left(\frac{1+e^{\langle X,\theta\rangle}}{1+e^{\langle X,\theta^*\rangle}}-\frac{1+e^{-\langle X,\theta\rangle}}{1+e^{-\langle X,\theta^*\rangle}}\right)\frac{e^{-\langle X,\theta\rangle}}{(1+e^{-\langle X,\theta\rangle})^2}\langle X,\,\theta-\theta^*\rangle\right].
\end{aligned}$$

By the mean value theorem, there exists ξ between 0 and $\langle X,\,\theta-\theta^*\rangle$ such that

$$\frac{1+e^{\langle X,\theta\rangle}}{1+e^{\langle X,\theta^*\rangle}}-\frac{1+e^{-\langle X,\theta\rangle}}{1+e^{-\langle X,\theta^*\rangle}}=\langle X,\,\theta-\theta^*\rangle\left(\frac{e^{\langle X,\theta^*\rangle+\xi}}{1+e^{\langle X,\theta^*\rangle}}+\frac{e^{-\langle X,\theta^*\rangle-\xi}}{1+e^{-\langle X,\theta^*\rangle}}\right).$$

In light of the above equality, we arrive at the following inequalities:

$$\begin{aligned}
\langle \nabla F^R(\theta),\, \theta^*-\theta\rangle &\geq \mathbb{E}\Bigg[\inf_{|\xi|\in[0,|\langle X,\theta-\theta^*\rangle|]}\left(\frac{e^{\langle X,\theta^*\rangle+\xi}}{1+e^{\langle X,\theta^*\rangle}}+\frac{e^{-\langle X,\theta^*\rangle-\xi}}{1+e^{-\langle X,\theta^*\rangle}}\right)\\
&\qquad\qquad\times\frac{e^{-\langle X,\theta\rangle}}{(1+e^{-\langle X,\theta\rangle})^2}|\langle X,\,\theta-\theta^*\rangle|^2\Bigg]\\
&\geq \mathbb{E}\left[\frac{1}{2}e^{-|\langle X,\theta-\theta^*\rangle|}\frac{e^{-\langle X,\theta\rangle}}{(1+e^{-\langle X,\theta\rangle})^2}|\langle X,\,\theta-\theta^*\rangle|^2\right]\\
&\geq \frac{1}{8}\mathbb{E}\left[e^{-|\langle X,\theta-\theta^*\rangle|-|\langle X,\theta\rangle|}|\langle X,\,\theta-\theta^*\rangle|^2\right]\\
&\geq \frac{1}{8e^4}\mathbb{E}\left[\mathbf{1}_{\{|\langle X,\theta\rangle|\leq 2,\ |\langle X,\theta-\theta^*\rangle|\leq 2\}}|\langle X,\,\theta-\theta^*\rangle|^2\right].
\end{aligned}$$

Since $X\sim\mathcal{N}(0,I_d)$, we have

$$\begin{bmatrix}\langle X,\,\theta\rangle\\ \langle X,\,\theta-\theta^*\rangle\end{bmatrix}\sim\mathcal{N}\left(0,\begin{bmatrix}\|\theta\|_2^2 & \langle\theta,\,\theta-\theta^*\rangle\\ \langle\theta,\,\theta-\theta^*\rangle & \|\theta-\theta^*\|_2^2\end{bmatrix}\right).$$

Given that result, direct calculation leads to

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\{|\langle X,\theta\rangle|\leq 2,|\langle X,\theta-\theta^*\rangle|\leq 2\}}|\langle X,\,\theta-\theta^*\rangle|^2\right)\\
&\qquad\qquad\geq \frac{c}{(1+\|\theta\|_2)(1+\|\theta-\theta^*\|_2)}\|\theta-\theta^*\|_2^2,
\end{aligned}$$

for a universal constant $c > 0$. Collecting the above results, for all θ such that $\|\theta - \theta^*\|_2 \leq 1$, we achieve that

$$\langle \nabla F^R(\theta), \theta^* - \theta \rangle \geq \frac{c}{(1 + \|\theta\|_2)(1 + \|\theta - \theta^*\|_2)} \|\theta - \theta^*\|_2^2$$
$$\geq c \frac{1}{1 + \|\theta^*\|_2} \|\theta - \theta^*\|_2^2 .$$

For θ with $\|\theta - \theta^*\|_2 > 1$, let $\widetilde{\theta} = \theta^* + \frac{\theta - \theta^*}{\|\theta - \theta^*\|_2}$. Then, we find that

$$\langle \nabla F^R(\theta), \theta^* - \theta \rangle \geq \langle \nabla F^R(\widetilde{\theta}), \theta^* - \theta \rangle \geq \frac{c}{2(1 + \|\theta^*\|_2)} \|\theta - \theta^*\|_2 ,$$

which yields the claim (4.2a).

B.1.2. Proof of the bound (4.2b). In this appendix, we prove the uniform bound (4.2b) between the empirical and population likelihood gradients. It suffices to establish the following stronger result:

$$\text{(B.1)} \qquad Z := \sup_{\theta \in \mathbb{R}^d} \left\| \nabla F_n^R(\theta) - \nabla F^R(\theta) \right\|_2 \leq c \left\{ \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right\},$$

with probability at least $1 - \delta$ for any $\frac{n}{\log n} \geq c_0 d \log(1/\delta)$ where c_0 is a universal constant.

In order to prove the claim (B.1), we exploit a concentration inequality due to Adamczak [SM1]; it gives tight tail bounds for supremum of unbounded empirical processes. Throughout our derivation, we use $\|X\|_{\psi_\alpha}$ to denote the Orlicz ψ_α norm for a random variable X, for any $\alpha \in (0, 2]$. Let us state a simplified version of a theorem due to Adamczak:

Proposition B.1 (Theorem 4, [SM1], simplified version). *Let $(x, \theta) \mapsto f(\theta; x)$ be a function with domain $\Theta \times \mathcal{X}$, and suppose that there is a function $\bar{F} : \mathcal{X} \to \mathbb{R}$ such that $|f(\theta, x)| \leq \bar{F}(x)$ for any $\theta \in \Theta$. Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$, and suppose that $\left\|\bar{F}\right\|_{\psi_\alpha} < +\infty$ for some $\alpha \leq 1$. Then the random variable $Z_n := \frac{1}{n} \sup_{\theta \in \Theta} \left|\sum_{i=1}^n f(\theta; X_i) - \mathbb{E}[f(\theta; X)]\right|$ satisfies the bound:*

$$\mathbb{P}\left(Z_n > 2\mathbb{E}[Z_n] + t\right) \leq \exp\left(-\frac{t^2}{2\mathbb{E}[\bar{F}(X)^2]}\right) + 3\exp\left(-\left(\frac{t}{c\left\|\max_{i \in [n]} \bar{F}(X_i)\right\|_{\psi_\alpha}}\right)^\alpha\right),$$

for a universal constant $c > 0$.

In order to prove the claim (B.1), we begin by writing Z as the supremum of a stochastic process. Let $\mathbb{S}^{d-1}$ denote the Euclidean sphere in $\mathbb{R}^d$, and define the stochastic process

$$Z_{u,\theta} := \left| \frac{1}{n} \sum_{i=1}^n f_{u,\theta}(X_i, Y_i) - \mathbb{E}[f_{u,\theta}(X, Y)] \right|,$$

where $f_{u,\theta}(x, y) = \dfrac{y\langle x, u\rangle e^{y\langle x, \theta\rangle}}{1 + e^{y\langle x, \theta\rangle}}$, indexed by vectors $u \in \mathbb{S}^{d-1}$ and $\theta \in \mathbb{B}(\theta^*; r)$. The outer expectation in the above display is taken with respect to (X, Y) drawn from the logistic model (4.1)

Observe that $Z = \sup_{u \in \mathbb{S}^{d-1}} \sup_{\theta \in \mathbb{R}^d} Z_{u,\theta}$. Let $\{u^1, \ldots, u^N\}$ be a 1/8-covering of $\mathbb{S}^{d-1}$ in the Euclidean norm; there exists such a set with $N \leq 17^d$ elements. By a standard discretization argument (see Chapter 6, [SM5]), we have

$$Z \leq 2 \max_{j=1,\ldots,N} \sup_{\theta \in \mathbb{R}^d} Z_{u^j,\theta}.$$

Accordingly, the remainder of our argument focuses on bounding the random variable $V := \sup_{\theta \in \mathbb{R}^d} Z_{u,\theta}$, where the vector $u \in \mathbb{S}^{d-1}$ should be understood as arbitrary but fixed. For each $u \in \mathbb{S}^{d-1}$ fixed, we note that $\bar{F}(X,Y) = |\langle X, u \rangle|$ is an envelop function for the class $(f_{u,\theta}(X,Y))_{\theta \in \mathbb{R}^d}$. Additionally, by standard tail bounds for maximum of Gaussian random variables, we know that:

$$\left\| \max_{1 \leq i \leq n} \bar{F}(X_i, Y_i) \right\|_{\psi_1} \leq \sqrt{\log n}.$$

Consequently, invoking Proposition B.1 yields that

$$V \leq 2\mathbb{E}[V] + \sqrt{\frac{2\log(1/\delta)}{n}} + \frac{c \log(1/\delta)}{n} \sqrt{\log n} \tag{B.2}$$

with probability at least $1 - \delta$.

Now define the symmetrized random variable

$$V' := \sup_{\theta \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_{\theta,u}(X_i, Y_i) \right|.$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables. By standard symmetrization arguments, we have

$$\mathbb{E}[V] \leq 2\mathbb{E}\left[V'\right].$$

We now bound the expectation of V', first over the Rademacher variables. Consider the function class

$$\mathcal{G} := \left\{ g_\theta : (x,y) \mapsto \langle x, u \rangle \varphi_\theta(x,y) \ \mid \ \theta \in \mathbb{R}^d \right\}.$$

It is clear that the function class $\mathcal{G}$ has the envelope function $\bar{G}(x) := |\langle x, u \rangle|$. We claim that the L_2-covering number of $\mathcal{G}$ can be bounded as

$$\bar{N}(t) := \sup_Q \left| \mathcal{N}\left(\mathcal{G}, \|\cdot\|_{L^2(Q)}, t \left\|\bar{G}\right\|_{L^2(Q)} \right) \right| \leq \left(\frac{1}{t}\right)^{c(d+1)} \qquad \text{for all } t > 0, \tag{B.3}$$

where $c > 0$ is a universal constant.

Let us take the claim (B.3) as given for the moment, and use it to bound the expectation of V', first over the Rademacher variables. Define the empirical expectation

$\mathbb{P}_n(\bar{G}^2) := \frac{1}{n}\sum_{i=1}^n \langle X_i,\, u\rangle^2$. Invoking Dudley's entropy integral bound (e.g., Theorem 5.22, [SM5]), we find that there are universal constants C, C' such that

$$\mathbb{E}_\varepsilon[V'] = \mathbb{E}_\varepsilon\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i,Y_i)\right|\right] \le C\sqrt{\frac{\mathbb{P}_n(\bar{G}^2)}{n}}\int_0^1 \sqrt{1+\log \bar{N}(t)}dt$$
$$\le C'\sqrt{\mathbb{P}_n(\bar{G}^2)}\sqrt{\frac{d}{n}}.$$

Up to this point, we have been conditioning on the observations $\{X_i\}_{i=1}^n$. Taking expectations over them as well yields

$$\text{(B.4)}\qquad \mathbb{E}_{\varepsilon,X_1^n}[V'] \le C'\sqrt{\frac{d}{n}}\cdot \mathbb{E}_{X_1^n}\left[\sqrt{\mathbb{P}_n(\bar{G}^2)}\right] \overset{(i)}{\le} C'\sqrt{\frac{d}{n}}\cdot\sqrt{\mathbb{E}_{X_1^n}\left[\mathbb{P}_n(\bar{G}^2)\right]} \overset{(ii)}{=} C'\sqrt{\frac{d}{n}},$$

where step (i) follows from Jensen's inequality; and step (ii) uses the fact that $\mathbb{E}_{X_1^n}[\mathbb{P}_n(\bar{G}^2)] = 1$. Putting together the bounds (B.2) and (B.4) yields the following bound with probability $1-\delta$:

$$V \le c\sqrt{\frac{d+\log\delta^{-1}}{n}} + c\frac{\log\delta^{-1}}{n}\sqrt{\log n}.$$

This probability bound holds for each $u \in \mathbb{S}^{d-1}$. By taking the union bound over the 1/8-covering set $\{u^1,\ldots,u^N\}$ of $\mathbb{S}^{d-1}$ where $N \le 17^d$ and applying above bound with $\delta' = \delta/N$, we obtain the claim (B.1) for sample size satisfying $\frac{n}{\log n} \ge cd\log(1/\delta)$.

B.1.3. Proof of claim (B.3). We consider a fixed sequence $(x_i, y_i, t_i)_{i=1}^m$ where $y_i \in \{-1, 1\}$, $x_i \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$ for $i \in [m]$. Now, we suppose that for any binary sequence $(z_i)_{i=1}^m \in \{0,1\}^m$, there exists $\theta \in \mathbb{R}^d$ such that

$$z_i = \mathbb{I}\left[\langle X_i,\, u\rangle\varphi_\theta(X_i,Y_i) \ge t_i\right] \qquad \text{for all } i \in [m].$$

Following some algebra, we find that

$$y_i x_i^T\theta - \log\frac{Y_i t_i}{\langle X_i,\, u\rangle - Y_i t_i}\begin{cases}\ge 0 & z_i = 1\\ < 0 & z_i = 0\end{cases}.$$

Consequently, the set $\{[y_i x_i, \log(Y_i t_i/(\langle X_i,\, u\rangle - Y_i t_i))]\}_{i=1}^m$ of $(d+1)$-dimensional points can be shattered by linear separators. Therefore, we have $m \le d+2$, which leads to the VC subgraph dimension of $\mathcal{G}$ to be at most $d+2$ (e.g., see the book [SM4]). As a consequence, we obtain the conclusion of the claim (B.3).

B.2. Proof of Corollary 4.2. We prove Corollary 4.2 by verifying the claims (4.5a) and (4.5b).

B.2.1. Structure of F^G. Direct algebra leads to the following equation

$$\langle \nabla F^G(\theta),\, \theta^* - \theta\rangle = \left(\theta - \mathbb{E}\left[X\tanh\left(X^\top\theta\right)\right]\right)^\top(\theta-\theta^*)$$
$$\text{(B.5)}\qquad \ge \|\theta\|_2^2 - \|\theta\|_2\left\|\mathbb{E}\left[X\tanh\left(X^\top\theta\right)\right]\right\|_2$$

where $\tanh(x) := \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$ for all $x \in \mathbb{R}$. From Theorem 2 in Dwivedi et al. [SM3], we have

$$\left\| \mathbb{E}\left[X \tanh\left(X^\top \theta\right)\right] \right\|_2 \leq \left(1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2}}\right) \|\theta\|_2$$

for all $\theta \in \mathbb{R}^d$ where $p := \mathbb{P}(|Y| \leq 1) + \frac{1}{2}\mathbb{P}(|Y| > 1)$ where $Y \sim \mathcal{N}(0, 1)$. Plugging the above inequality into equation (B.5) leads to

$$\langle \nabla F^G(\theta), \theta^* - \theta \rangle \geq \frac{p \|\theta\|_2^4}{2 + \|\theta\|_2^2} \geq \begin{cases} \frac{p}{4} \|\theta\|_2^4, & \text{for } \|\theta\|_2 \leq \sqrt{2} \\ \frac{p}{2}\left(\|\theta\|_2^2 - 1\right), & \text{otherwise} \end{cases}.$$

As a consequence, we achieve the conclusion of claim (4.5a).

B.2.2. Perturbation error between ∇F^G and ∇F_n^G. Direct calculation indicates the following equation:

$$\nabla F_n^G(\theta) - \nabla F^G(\theta) = \frac{1}{n}\sum_{i=1}^n X_i \tanh(X_i^\top \theta) - \mathbb{E}\left[X \tanh\left(X^\top \theta\right)\right].$$

The outer expectation in the above display is taken with respect to $X \sim \mathcal{N}(\theta^*, \sigma^2 I_d)$ where $\theta^* = 0$. Based on the proof argument of Lemma 1 from the paper [SM3], for each $r > 0$, we have the following concentration inequality

$$\mathbb{P}\Bigg(\sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\| \frac{1}{n}\sum_{i=1}^n X_i \tanh(X_i^\top \theta) - \mathbb{E}\left[X \tanh\left(X^\top \theta\right)\right] \right\|_2 \leq c r \sqrt{\frac{d + \log(1/\delta)}{n}} \Bigg) \geq 1 - \delta, \tag{B.6}$$

for any $\delta > 0$ as long as the sample size $n \geq c' d \log(1/\delta)$ where c and c' are universal constants. For any $M \in \mathbb{N}_+$, by the concentration bound (B.6) and the union bound, we find that

$$\mathbb{P}\Bigg(\forall r \in [2^{-M}, 1], \quad \sup_{\theta \in \mathbb{B}(\theta^*, r)} \left\| \nabla F_n^G(\theta) - F^G(\theta) \right\|_2 \leq c\, r \sqrt{\frac{d + \log(M/\delta)}{n}} \Bigg) \geq 1 - \delta. \tag{B.7}$$

On the other hand, based on the standard inequality $|\tanh(x)| \leq |x|$ for all $x \in \mathbb{R}$, we find that

$$\begin{aligned} \left\| \nabla F_n^G(\theta) - \nabla F^G(\theta) \right\|_2 &\leq \frac{1}{n}\sum_{i=1}^n \|X_i\|_2 \left|\tanh\left(X_i^\top \theta\right)\right| + \mathbb{E}\left[\|X\|_2 \left|\tanh\left(X^\top \theta\right)\right|\right] \\ &\leq \frac{1}{n}\sum_{i=1}^n \|X_i\|_2 \left|X_i^\top \theta\right| + \mathbb{E}\left[\|X\|_2 \left|X^\top \theta\right|\right] \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \|X_i\|_2^2 + \mathbb{E}\left[\|X\|_2^2\right]\right) \|\theta\|_2. \end{aligned}$$

Therefore, we have $\left\|\nabla F_n^G(\theta) - \nabla F^G(\theta)\right\|_2 \leq 2d\,\|\theta\|_2 \log(1/\delta)$ with probability $1-\delta$. By choosing $M_1 := \log(2nd)$, based on the previous bound, we obtain that

$$\mathbb{P}\left(\forall r < 2^{-M_1},\ \sup_{\theta\in\mathbb{B}(\theta^*,r)} \left\|\nabla F_n^G(\theta) - \nabla F^G(\theta)\right\|_2 \leq \frac{\log(1/\delta)}{n}\right) \geq 1-\delta. \tag{B.8}$$

Furthermore, for vector $\theta \in \mathbb{R}^d$ with large norm, by the concentration bound (B.6) combined with the union bound, for any $M' \in \mathbb{N}_+$, we find that

$$\mathbb{P}\Bigg(\forall r \in [1, 2^{M'}],\ \sup_{\theta\in\mathbb{B}(\theta^*,r)} \left\|\nabla F_n^G(\theta) - F^G(\theta)\right\|_2 \leq c\, r\, \sqrt{\frac{d+\log(M'/\delta)}{n}}\Bigg) \geq 1-\delta.$$

When r in the above bound is too large, we can simply use the fact that tanh is a bounded function. We thus have the upper bound

$$\left\|\nabla F_n^G(\theta) - \nabla F^G(\theta))\right\|_2 \leq \mathbb{E}\left[\|X\|_2\right] + \frac{1}{n}\sum_{i=1}^n \|X_i\|_2\,,$$

for any θ. Given the above bound, by choosing $M_2 := \log(2\sqrt{n})$, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\forall r > 2^{M_2},\ \sup_{\theta\in\mathbb{B}(\theta^*,r)} \left\|\nabla F_n^G(\theta) - \nabla F^G(\theta))\right\|_2 \leq r\sqrt{\frac{d+\log(1/\delta)}{n}}\right)\\
&\quad\geq \mathbb{P}\left(\mathbb{E}\left[\|X\|_2\right] + \frac{1}{n}\sum_{i=1}^n \|X_i\|_2 \leq 2^{M_2}\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \geq 1-\delta.
\end{aligned} \tag{B.9}$$

Putting the bounds (B.7), (B.8), and (B.9) together, for $n \geq c d \log(1/\delta)$, the following probability bound holds

$$\mathbb{P}\Bigg(\forall r > 0,\ \sup_{\theta\in\mathbb{B}(\theta^*,r)} \left\|\nabla F_n^G(\theta) - \nabla F^G(\theta))\right\|_2 \leq c\, r\, \sqrt{\frac{d+\log\left(\log n/\delta\right)}{n}} + \frac{\log(1/\delta)}{n}\Bigg) \geq 1-\delta,$$

which completes the proof of the claim (4.5b).

Appendix C. Proofs of the remaining auxiliary results.

In this appendix, we provide proofs of the remaining auxiliary results in the paper.

C.1. Proof of Proposition 2.1. For any $p \geq 2$, we define the quantity:

$$R_p := \sup_{p\geq 0} \left(\mathbb{E}_{\pi_t}\left[\|X\|_2^p\right]\right)^{1/p} \vee \left(\mathbb{E}_{\pi^*}\left[\|X\|_2^p\right]\right)^{1/p}$$

For any given value $\bar{R} > 0$, we note the following decomposition:

$$
\begin{aligned}
&|\mathbb{E}_{\pi_t}[\|X\|_2^p] - \mathbb{E}_{\pi^*}[\|X\|_2^p]| \\
&\leq \int_{\mathbb{B}(0,\bar{R})} |\pi_t - \pi^*| \cdot \|x\|_2^p \, dx + \int_{\mathbb{B}(0,\bar{R})^C} \pi_t(x) \, \|x\|_2^p \, dx + \int_{\mathbb{B}(0,\bar{R})^C} \pi^*(x) \, \|x\|_2^p \, dx \\
&\leq \bar{R}^p \cdot d_{\mathrm{TV}}(\pi_t, \pi^*) + \mathbb{E}_{\pi_t}\left[\|X\|_2^p \mathbf{1}_{\|X\|_2 > \bar{R}}\right] + \mathbb{E}_{\pi^*}\left[\|X\|_2^p \mathbf{1}_{\|X\|_2 > \bar{R}}\right] \\
&\leq \bar{R}^p \cdot d_{\mathrm{TV}}(\pi_t, \pi^*) + \sqrt{\mathbb{E}_{\pi_t}\left[\|X\|_2^{2p}\right]} \sqrt{\pi_t\left(\|X\|_2 > \bar{R}\right)} + \sqrt{\mathbb{E}_{\pi^*}\left[\|X\|_2^{2p}\right]} \sqrt{\pi^*\left(\|X\|_2 > \bar{R}\right)} \\
&\leq \bar{R}^p \cdot d_{\mathrm{TV}}(\pi_t, \pi^*) + 2R_{2p}^p \cdot R_2 / \bar{R}.
\end{aligned}
$$

For any $\varepsilon > 0$, take $\bar{R} := \frac{\varepsilon}{2R_{2p}^p R_2}$, we have that:

$$
\lim_{t \to +\infty} |\mathbb{E}_{\pi_t}[\|X\|_2^p] - \mathbb{E}_{\pi^*}[\|X\|_2^p]| \leq \varepsilon,
$$

which proves the claim.

C.2. A limit result. We begin with a lemma on the limiting behavior of a certain type of function. The lemma is used in the proof of Theorem 3.2 in Subsection 5.2.

Lemma C.1. *Let ϕ be a concave and continuous function on $[0, +\infty)$ with $\phi(0) = \phi(c) = 0$ for some positive constant $c > 0$. Assume furthermore that $\phi(t) < 0$ for all $t \in (c, \infty)$. Suppose that there exist two continuous functions $f, g : [0, +\infty) \to [0, +\infty)$ such that $\lim_{t \to +\infty} g(t)$ exists and $f(t) \leq \int_0^t \phi(g(s))ds$ for all $t \geq 0$. Under these conditions, we have $\lim_{t \to +\infty} g(t) \leq c$.*

Proof. Define the limit $A := \lim_{t \to +\infty} g(t)$, which exists according to the assumptions. We proceed via proof by contradiction. In particular, suppose that $A > c$. Based on the definition of A, for the positive constant $\varepsilon = (A - c)/2 > 0$, we can find a sufficiently large positive constant T such that $g(t) > A - \varepsilon$ for any $t \geq T$. Since the function ϕ is concave, with $\phi(c) = 0$ and $\phi(t) < 0$ for $t > c$, we have that ϕ is non-increasing on $[c, +\infty)$, and therefore

$$
\delta := \phi(c + \varepsilon) = - \sup_{s \geq c + \varepsilon} \phi(s) < 0.
$$

Therefore, for all $t > T$, we arrive at the following inequalities

$$
0 \leq f(t) \leq \int_0^T \phi(g(s))ds + \int_T^t \phi(g(s))ds \leq \int_0^T \phi(g(s))ds - \delta(t - T).
$$

By choosing $t = 1 + T + \delta^{-1} \int_0^T \phi(g(s))ds$, the above inequality cannot hold. This yields the desired contradiction, which completes the proof. ∎

C.3. A tail bound based on truncation. We now state an upper deviation inequality based on a truncation argument. Consider a sequence of random variables $\{Y_i\}_{i=1}^n$ satisfying the moment bounds

$$
\mathbb{E}\left[|Y_i|^q\right] \leq (aq)^{bq} \quad \text{for all } q = 1, 2, \ldots \tag{C.1}
$$

where a, b are universal constants.

Lemma C.2. *Given an i.i.d. sequence of zero-mean random variables* $\{Y_i\}_{i=1}^n$ *satisfying the moment bounds* (C.1)*, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i \geq (4a)^b\sqrt{\frac{\log 4/\delta}{n}} + \left(a\log\frac{n}{\delta}\right)^b \frac{\log 4/\delta}{n}\right) \leq \delta.$$

Proof. The proof of the lemma is a direct combination of truncation argument and Bernstein's inequality. In particular, for each $i \in [n]$, define the truncated random variable $\widetilde{Y}_i := Y_i \mathbb{I}\left[|Y_i| \leq 3(a\log\frac{n}{\delta})^b\right]$. With this definition, we have

$$\mathbb{P}\left((Y_i)_{i=1}^n \neq (\widetilde{Y}_i)_{i=1}^n\right) = \mathbb{P}\left(\max_{1\leq i\leq n}|Y_i| > 3\left(a\log\frac{n}{\delta}\right)^b\right)$$
$$\leq n\mathbb{P}\left(|Y_i| > 3\left(a\log\frac{n}{\delta}\right)^b\right) \leq \frac{\delta}{2}.$$

Therefore, it is sufficient to study a concentration behavior of the quantity $\sum_{i=1}^n \widetilde{Y}_i$. Invoking Bernstein's inequality [SM2], we obtain that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i \geq \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2(2a)^{2b} + \frac{2}{3}\varepsilon\cdot 3(a\log\frac{n}{\delta})^b}\right).$$

In order to make the RHS of the above inequality less than $\frac{\delta}{2}$, it suffices to set

$$\varepsilon = (4a)^b\sqrt{\frac{\log(4/\delta)}{n}} + \left(a\log\frac{n}{\delta}\right)^b\frac{\log(4/\delta)}{n}.$$

Collecting all of the above inequalities yields the claim. ■

C.4. Unique positive solution to equation (3.1). We now establish that equation (3.1) has a unique positive solution under the stated assumptions. Define the function

$$\vartheta(z) := \psi(z) - \left(\varepsilon(n,\delta)\zeta(z)z + \frac{Bz + d + \log(1/\delta)}{n}\right).$$

Since $\psi(0) = 0$, we have $\vartheta(0) < 0$. On the other hand, based on Assumption **(W.4)**, $\liminf_{z\to+\infty}\vartheta(z) > 0$. Therefore, there exists a positive solution to the equation $\vartheta(z) = 0$.

Recall that $\xi : \mathbb{R}_+ \to \mathbb{R}$ is an inverse function of the strictly increasing function $z \mapsto z\zeta(z)$. Therefore, we can write the function ϑ as follows:

$$\vartheta(z) = \widetilde{\vartheta}(r) := \psi(\xi(r)) - \varepsilon(n,\delta)r - \frac{B\xi(r) + d + \log(1/\delta)}{n},$$

where $r = z\cdot\zeta(z)$. Given the convexity of function $r \mapsto \psi(\xi(r))$ guaranteed by Assumption **(W.3)**, the functions $\widetilde{\vartheta}$ and ϑ are convex. Putting the above results together, there exists a unique positive solution to equation (3.1).

REFERENCES

[1] R. Adamczak, *A tail inequality for suprema of unbounded empirical processes with applications to markov chains*, Electronic Journal of Probability, 13 (2008), pp. 1000–1034.

[2] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2016.

[3] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, *Singularity, misspecification, and the convergence rate of EM*, arXiv preprint arXiv:1810.00828, (2018).

[4] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer-Verlag, New York, NY, 2000.

[5] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.