Journal of Machine Learning Research 25 (2024) 1-79 Submitted 6/23; Revised 4/24; Published 4/24

On the Computational Complexity of Metropolis-Adjusted Langevin Algorithms for Bayesian Posterior Sampling

Rong Tang MARTANG@UST.HK
Department of Mathematics
The Hong Kong University of Science and Technology

Yun Yang YY84@ILLINOIS.EDU
Department of Statistics
University of Illinois Urbana-Champaign

Editor: Pierre Alquier

Abstract

In this paper, we examine the computational complexity of sampling from a Bayesian posterior (or pseudo-posterior) using the Metropolis-adjusted Langevin algorithm (MALA). MALA first employs a discrete-time Langevin SDE to propose a new state, and then adjusts the proposed state using Metropolis-Hastings rejection. Most existing theoretical analyses of MALA rely on the smoothness and strong log-concavity properties of the target distribution, which are often lacking in practical Bayesian problems. Our analysis hinges on statistical large sample theory, which constrains the deviation of the Bayesian posterior from being smooth and log-concave in a very specific way. In particular, we introduce a new technique for bounding the mixing time of a Markov chain with a continuous state space via the s-conductance profile, offering improvements over existing techniques in several aspects. By employing this new technique, we establish the optimal parameter dimension dependence of $d^{1/3}$ and condition number dependence of κ in the non-asymptotic mixing time upper bound for MALA after the burn-in period, under a standard Bayesian setting where the target posterior distribution is close to a d-dimensional Gaussian distribution with a covariance matrix having a condition number κ. We also prove a matching mixing time lower bound for sampling from a multivariate Gaussian via MALA to complement the upper bound.

Keywords: Bayesian inference, Gibbs posterior, Large sample theory, Log-isoperimetric inequality, Metropolis-adjusted Langevin algorithms, Mixing time.

1. Introduction

Bayesian inference gains significant popularity during the last two decades due to the advance in modern computing power. As a method of statistical analysis based on probabilistic modelling, Bayesian inference allows natural uncertainty quantification on the unknown parameters via a posterior distribution. In the classical Bayesian framework, the data $X^{(n)} = \{X_1, \ldots, X_n\}$ is assumed to consist of i.i.d. samples generated from a probability distribution $p(X \,|\, \theta)$ depending on an unknown parameter θ in parameter space $\Theta \subset \mathbb{R}^d$. Domain knowledge and prior beliefs can be characterized by a probability distribution $\pi(\theta)$ over Θ called prior (distribution), which is then updated into a posterior (distribution)

$p(\theta \,|\, X^{(n)})$ by multiplying with the likelihood function

$$\mathcal{L}_n(\theta;\, X^{(n)}) := \prod_{i=1}^{n} p(X_i \,|\, \theta)$$

evaluated on the observed data $X^{(n)}$ using the Bayes theorem. The classical Bayesian framework relies on the likelihood formulation, which hinders its use in problems where the data generating model is hard to fully specify or is not our primary interest. The pseudo-posterior (Alquier et al., 2016; Ghosh et al., 2020) idea provides a more general probabilistic inference framework to alleviate this restriction by replacing the negative log-likelihood function in the Bayesian posterior with a criterion function. For example, when applied to risk minimization problems, the so-called Gibbs posteriors (Bhattacharya and Martin, 2020; Syring and Martin, 2020) use the (scaled) empirical risk function as the criterion function, thus avoiding imposing restrictive assumptions on the statistical model through a fully specified likelihood function.

Despite the conceptual appeal of Bayesian inference, its practical implementation is a notoriously difficult computational problem. For example, the posterior $p(\theta \,|\, X^{(n)})$ involves a normalisation constant that can be expressed as a multidimensional integral

$$\int_{\Theta} \mathcal{L}_n(\theta;\, X^{(n)})\, \pi(\theta)\, \mathrm{d}\theta.$$

This integral is usually analytically intractable and hard to numerically approximate, especially when the parameter dimension d is high. Different from those numerical methods for directly computing the normalisation constant, the Markov chain Monte Carlo (MCMC) algorithm (Hastings, 1970; Geman and Geman, 1984; Robert et al., 2004) constructs a Markov chain, whose simulation only requires evaluations of the likelihood ratio under a pair of parameters, such that its stationary distribution matches the target posterior distribution. Thus, MCMC provides an appealing alternative for Bayesian computation by turning the integration problem into a sampling problem that does not require computing the normalisation constant. Despite its popularity, the theoretical analysis of the computational efficiency of MCMC algorithms is mostly carried out for smooth and log-concave target distributions, and is comparatively rare in the Bayesian literature where a (pseudo-)posterior can be non-smooth and non-log-concave. In addition, precise characterizations of the computational complexity (or mixing time) and its dependence on the parameter dimension d for commonly used MCMC algorithms are important for guiding their practical designs and use.

A widely used MCMC algorithm for sampling from Bayesian posteriors is the Gibbs sampler, which generates samples from a multivariate distribution by iteratively sampling each variable from its conditional distribution, given all other variables. The Gibbs sampler is particularly efficient for Bayesian models with closed-form conditional distributions under conjugate priors. A recent theoretical study by Ascolani and Zanella (2023) provides a dimension-free mixing time bound for the Gibbs sampler when applied to certain high-dimensional Bayesian hierarchical models. However, it is important to note that each iteration of their algorithm involves the sequential sampling of each dimension of the parameter from its corresponding full conditional distribution. This means that the total number of sampling steps required for the Gibbs sampler to converge is at least linear in the parameter dimension, which is larger than our sub-linear $d^{1/3}$ scaling of the needed sampling steps for MALA. On the other hand, the per-step cost of MALA can be linear in d because of gradient computation, while that of Gibbs sampling can be much lower, especially under weak dependence (although in the worst case, computing each conditional distribution may also require $O(d)$ complexity). On a separate note, we would like to mention that although MALA has a per-iteration cost linear in d to compute the gradient, the computation across different dimensions can be parallelized. In contrast, Gibbs sampling must sequentially scan over all its components and cannot be made parallel in order to maintain the detailed balance property. Additionally, the high efficiency of the Gibbs sampler often relies on the use of conjugate priors that facilitate closed-form conditional distributions. However, for complex Bayesian models, such a conjugate prior may not exist, as is the case in Bayesian quantile regression, discussed in Yu and Moyeed (2001), or linear regression with heavy-tailed noise (like Student's t-distributions). Moreover, there are situations where people tend to use specific non-conjugate priors for particular reasons. For example, sparsity-induced priors such as the spike and slab priors (with heavy-tailed slabs) are widely used in regression analysis for facilitating variable selection. In these complicated scenarios, one might have to resort to using MALA or, more broadly, the Metropolis-Hastings (MH) algorithm, to draw samples from the Bayesian posterior.

On the other hand, the Metropolis-Hastings (MH) algorithm provides a more flexible alternative. An MH algorithm produces samples by proposing and then accepting or rejecting these proposals based on a specified acceptance criterion. A key advantage of the MH algorithm is its ability to handle Bayesian (pseudo-)posterior distributions without requiring explicit knowledge of the normalization constant or the full conditional distributions. One of the most popular MH algorithms is the Metropolis random walk (MRW), a zeroth-order method that queries the value of the target density ratio under two points per iteration. Dwivedi et al. (2019) shows that for a log-concave and smooth target density, the ε-mixing time in total variation distance (the number of iterations required to converge to an ε-neighborhood of stationary distribution in the total variation distance) for MRW is at most $\mathcal{O}\big(d\log(1/\varepsilon)\big)$. On the other hand, the $\mathcal{O}(d)$ scaling limit of Gelman et al. (1997) suggests that their linear dependence on dimension d is optimal. For a class of Bayesian pseudo-posteriors that can be non-smooth and non-log-concave, it has been shown in Belloni and Chernozhukov (2009) that as the sample size n grows to infinity while the parameter dimension d does not grow too quickly relative to n so that the pseudo-posterior satisfies a Bernstein-von Mises (asymptotic normality) result, then MRW for sampling from the target pseudo-posterior constrained on an approximate compact set with a warm start has an asymptotic total variation ε-mixing time upper bound as $\mathcal{O}_p\big(d^2\log(1/\varepsilon)\big)$.

Another prominent class of MH algorithms is the Metropolis-adjusted Langevin algorithm (MALA), which utilizes additional gradient information about the target density. Although this approach requires computing the gradient and can be costlier than zeroth-order methods that only use function evaluations, the development of automatic differentiation tools (Paszke et al., 2017; Margossian, 2019) has simplified this task for many explicit and smooth densities. These tools make the computational demands for gradient computation comparable to those for evaluating the density itself. Furthermore, it has been demonstrated that MALA tends to have a lower mixing time in comparison to the MRW. For example, Chewi et al. (2021) show that if the negative log-density (will be referred to as potential) of the target distribution is twice continuously differentiable and strongly convex, then the ε-mixing time in χ^2 divergence for MALA with a warm start scales as $\Theta(d^{1/2})$ modulo polylogarithmic factors in ε. Additionally, Roberts and Rosenthal (1998) and Chewi et al. (2021) show that the optimal dimension dependence for MALA is $d^{1/3}$ for some product measures satisfying stringent conditions like the standard Gaussian. However, for Bayesian (pseudo-)posteriors, it is common that the smoothness and strong convexity properties of the log-density assumed in literature are not satisfied. For instance, consider Bayesian quantile regression with a quantile level τ. Given a dataset $X^{(n)} = \{X_i = (\widetilde{X}_i, Y_i)\}_{i=1}^n$ consisting of covariates and response variables, the posterior distribution then takes the form of $\pi_n(\theta|X^{(n)}) \propto \exp\big(-\sum_{i=1}^n (Y_i - \widetilde{X}_i^T\theta)(\tau - \mathbf{1}(Y_i < \widetilde{X}_i^T\theta))\big)\,\pi(\theta)$, where $\mathbf{1}(\cdot)$ denotes the indicator function. An important feature of this example is that the resulting Bayesian posterior is neither differentiable owing to the discontinuity introduced by the indicator function, nor strongly log-concave. For such non-differentiable densities, we slightly extend the MALA by using any subgradient to replace the gradient in its algorithm formulation. Theoretically, it is natural to investigate:

> What is the optimal dimension (and condition number) dependence when using MALA to sample from a possibly non-smooth and non-log-concave (pseudo-)posterior density, in light of the asymptotic Gaussian nature of the posterior as predicted by statistical large sample theory?

Moreover, it would be insightful to determine to what extent we can diverge from a Gaussian distribution while preserving the dimension dependence as sampling from a Gaussian distribution, and how various factors, such as the dimensionality, sample size and density smoothness, affect the deviance of the posterior from the Gaussian distribution.

Our contributions. In this work, we show an upper bound on the ε-mixing time of MALA for sampling from a class of possibly non-smooth and non-log-concave distributions with non-product forms (c.f. Condition A for a precise definition) with an M_0-warm start (defined in Section 2.3) as $\mathcal{O}\big(\max\big\{d^{1/3}\log(\varepsilon^{-1}\log M_0),\, \log M_0\big\}\big)$, which matches (up to logarithmic terms in (M_0,ε)) the lower bound result proved in Chewi et al. (2021) that the mixing time of MALA for the standard Gaussian is at least $\mathcal{O}(d^{1/3})$. Specially, our condition requires the target distribution (after proper rescaling by the sample size n) to be close to a multivariate Gaussian subject to small perturbations. We verify that a wide class of Gibbs posteriors (Bhattacharya and Martin, 2020; Syring and Martin, 2020), including conventional Bayesian posteriors defined through likelihood functions, meets our condition under a minimal set of assumptions. In particular, our theory provides an explicit upper bound condition on the growth of parameter dimension d relative to sample size n, stated in a non-asymptotic manner, that is, $d \le c\frac{n^{\kappa_1}}{\log n}$, where κ_1 depends on the regularity of the density function (c.f. Theorem 5). Specifically, for less smooth density functions, a smaller dimension d is necessary to maintain the $d^{1/3}$ scaling of the mixing time guarantee, which is also supported by our numerical results in Section 7.

In addition, our result illustrates that the mixing time of MALA exhibits a linear dependence on the condition number κ of the covariance matrix (which may have a polynomial dependence on the dimension in some ill-conditioned cases) of the approximating multivariate Gaussian. Our bound matches the mixing time scaling of Gaussian targets with condition number κ, and is therefore optimal. For the sake of completeness, we derive a matching lower bound in Appendix A.3. In our lower bound analysis, we extend the proof of Theorem 1 in Chewi et al. (2021), which primarily focuses on a standard Gaussian target distribution. In addition, we also carefully keep track of the dependence on the condition number in our derivation, which allows us to establish a lower bound that explicitly demonstrates a linear dependence on the condition number and also matches with our upper bound.

It is worthwhile mentioning that our Condition A does not require the distance between the target posterior and the multivariate Gaussian distribution to vanish as n tends to infinity; while in the context of Bayesian posteriors, these distances indeed decay to zero under minimal assumptions on the statistical model. Therefore, our mixing time result is more generally applicable to problems beyond Bayesian posterior sampling, for example, to optimization of approximately convex functions via simulated annealing (Belloni et al., 2015), where the target distribution can deviate from being smooth and strongly log-concave by a finite amount. In such settings, the computational complexity of sampling algorithms scales as $\mathcal{O}(d^{1/3})$ with the variable dimension d under reasonably good initialization while that of a wide class of gradient-based optimization algorithms may scale exponentially (Ma et al., 2019).

Our result on the $\mathcal{O}(d^{1/3})$ dimension dependence for the mixing time of MALA after the burn-in period for the perturbed Gaussian class strengthens our understanding of sampling from non-smooth and non-log-concave distributions. It also partly fills the gap between the optimal $d^{1/3}$ mixing time for a class of sufficiently regular product distributions derived from the scaling limit approach in Roberts and Rosenthal (1998) and the $d^{1/2}$ lower bound on the class of all log-smooth and strongly log-concave distributions obtained in Chewi et al. (2021), by identifying a much larger class of distributions of practical interest that attain the optimal $d^{1/3}$ dimension dependence. Moreover, we introduce a somewhat more general average conductance argument based on the s-conductance profile in Section 3 to improve the warming parameter dependence without deteriorating the dimension dependence. More specifically, our mixing time upper bound improves upon existing results (e.g. Chewi et al., 2021) in the dependence on the warming parameter M_0 from logarithmic to doubly logarithmic (the $\log\log(M_0)$ term in Theorem 3) when $\log M_0 \leq d^{\frac{1}{3}}$, by adapting the s-conductance profile and the log-isoperimetric inequality device (Chen et al., 2020), or more generally, the log-Sobolev inequality device (Lovász and Kannan, 1999; Kannan et al., 2006), to our target distribution class. Our constraint of $d^{\frac{1}{3}}$ on $\log M_0$ can be overly strong for general target distributions in practice. For instance, in the case of distributions possessing product forms, such as a pair of isotropic Gaussians with varying means, $\log M_0$ tends to increase linearly with the dimension d. However, for Bayesian posterior with smooth density, we may leverage its asymptotic distribution to construct more effective warm starts (c.f. Lemma 4 and Corollary 7). In addition, we study a variant of MALA where the (sub-)gradient vector in the Langevin SDE is preconditioned by a matrix for capturing the local geometry, for example, the Fisher information matrix in the context of Bayesian posterior sampling, and we illustrate in our Corollaries 7 and 8 that MALA with suitable preconditioning may improve the convergence of the sampling algorithm even though the target density is non-differentiable.

Our analysis is motivated by the statistical large sample theory suggesting the Bayesian posterior to be close to a multivariate Gaussian. We develop mixing time bounds of MALA for sampling from general Gibbs posteriors (possibly with increasing parameter dimension and non-smooth criterion function) by establishing non-asymptotic Bernstein-von Mises results, applying techniques from empirical process theory, including chaining, peeling, and localization. Due to the delicate analysis in our mixing time upper bound proof that utilizes the explicit form of Gaussian distributions for bounding the acceptance probability in each step of MALA, we obtain a better dimension dependence of $d^{1/3}$ than the $d^{1/2}$ dependence derived for general smooth and log-concave densities. In addition, by utilizing our s-conductance profile technique, we can obtain a mixing time upper bound for sampling from the original Bayesian posterior instead of a truncated version considered in Belloni and Chernozhukov (2009).

Organization. The rest of the paper is organized as follows. In Section 2, we describe the background and formally formulate the theoretical problem of analyzing the computational complexity of MALA for Bayesian posterior sampling that is addressed in this work. In Section 3, we briefly review some common concepts and existing techniques for analyzing the computational complexity (in terms of mixing time) of a Markov chain, and introduce our improved technique based on s-conductance profile. In Section 4, we apply the generic technique developed in Section 3 to analyze MALA for Bayesian posterior sampling. In Section 5, we specialize the general mixing bound of MALA to the class of Gibbs posteriors, and apply it to both Gibbs posteriors with smooth and non-smooth loss functions. Section 6 sketches the main ideas in proving the MALA mixing time bound and discuss some main differences with existing proofs. Some numerical studies are provided in Section 7, where we empirically compare the convergence of MALA and MRW. All proofs and technical details are deferred to the appendices in the supplementary material.

Notation. For two real numbers, we use $a \wedge b$ and $a \vee b$ to denote the maximum and minimum between a and b. For two distributions p and q, we use $\|p-q\|_{\mathrm{TV}} = \frac{1}{2}\int |p(x)-q(x)|\,\mathrm{d}x$ to denote their the total variation distance and $\chi^2(p,\, q)$ to denote their χ^2 divergence. We use $\|\cdot\|_p$ to denote the usual vector ℓ_p norm, and suppress the subscript when $p=2$. We use $\mathbf{0}_d$ to denote the d-dimensional all zero vector, and $B_r(x)$ to denote the closed ball centered at x with radius r (under the ℓ_2 distance) in the Euclidean space; in particular, we use B_r^d to denote $B_r(\mathbf{0}_d)$ when no ambiguity may arise. We use $\mathbb{S}^d = \big\{x \in \mathbb{R}^{d+1} : \|x\| = 1\big\}$ to denote the d-dimensional sphere. We use $N_d(\mu, \Sigma)$ to denote the d-dimensional multivariate Gaussian distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, and d suppress the subscript when $d=1$. We use $\mathcal{P}(K)$ to denote the set of probability measures on a set K. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla f(x)$ to denote the d-dimensional gradient vector of f at x and $\mathrm{Hess}(f(x))$ to denote the Hessian matrix of f at x. For a matrix J, we use $|\!|\!|J|\!|\!|_{\mathrm{op}}$ and $|\!|\!|J|\!|\!|_{\mathrm{F}}$ to denote its operator norm and Frobenius norm respectively, and use $\lambda_{\max}(J)$ and $\lambda_{\min}(J)$ to denote the maximal and minimal eigenvalues of J. Throughout, C, c, C_0, c_0, C_1, $c_1, \ldots$ are generically used to denote positive constants independent of n, d whose values might change from one line to another.

2. Background and Problem Setup

We first review the Bayesian (pseudo-)posterior framework and the Metropolis-adjusted Langevin algorithm (MALA). After that, we discuss an extension of MALA to handle the case where the target density is non-smooth by using the subgradient to replace the gradient and formulate the theoretical problem to be addressed in this work.

2.1 Bayesian pseudo-posterior

A standard Bayesian model consists of a prior distribution (density) $\pi(\theta)$ over parameter space $\Theta \subset \mathbb{R}^d$ as the marginal distribution of the parameter θ and a sampling distribution (density) $p(X \,|\, \theta)$ as the conditional distribution of the observation random variable X given θ. After obtaining a collection of n observations $X^{(n)} = \{X_1, X_2, \cdots, X_n\}$ modelled as n independent copies of X given θ, we update our beliefs about θ from the prior by calculating the posterior distribution (density)

$$p(\theta \,|\, X^{(n)}) = \frac{\exp\big\{\log \pi(\theta) + \log \mathcal{L}_n(\theta;\, X^{(n)})\big\}}{\int_\Theta \exp\big\{\log \pi(\theta) + \log \mathcal{L}_n(\theta;\, X^{(n)})\big\}\,\mathrm{d}\theta}, \quad \theta \in \Theta, \tag{1}$$

where recall that $\mathcal{L}_n(\theta;\, X^{(n)}) = \prod_{i=1}^n p(X_i|\theta)$ is the likelihood function. Despite the Bayesian formulation, in our theoretical analysis, we will adopt the frequentist perspective by assuming the data $X^{(n)}$ to be i.i.d. samples from an unknown data generating distribution $\mathcal{P}^* := p(X \,|\, \theta^*)$, where θ^* will be referred to as the true parameter, or simply truth, throughout the rest of the paper.

In many real situations, practitioners may not be interested in learning the entire data generating distribution $\mathcal{P}^*$, but want to draw inference on some characteristic as a functional $\theta = \theta(\mathcal{P}^*)$ of $\mathcal{P}^*$, which alone does not fully specify $\mathcal{P}^*$. An illustrative example is the quantile regression where the goal is to learn the conditional quantile of the response given the covariates; however, the conventional Bayesian framework requires a full specification of the condition distribution by imposing extra restrictive assumptions on the model, which may lead to model misspecification and sacrifice estimation robustness. A natural idea to alleviate the limitation of requiring a well-specified likelihood function is to replace the log-likelihood function $\log \mathcal{L}_n(\theta;\, X^{(n)})$ in the usual Bayesian posterior (1) by a criterion function $\mathcal{C}_n(\theta;\, X^{(n)})$. The resulting distribution,

$$\pi_n(\theta \,|\, X^{(n)}) = \frac{\exp\big\{\log \pi(\theta) + \mathcal{C}_n(\theta;\, X^{(n)})\big\}}{\int_\Theta \exp\big\{\log \pi(\theta) + \mathcal{C}_n(\theta;\, X^{(n)})\big\}\,\mathrm{d}\theta}, \quad \theta \in \Theta, \tag{2}$$

is called the Bayesian pseudo-posterior with criterion function $\mathcal{C}_n : \Theta \times \mathcal{X}^n \to \mathbb{R}$, and we may use the shorthand $\pi_n(\cdot)$ to denote the pseudo-posterior $\pi_n(\cdot|X^{(n)})$ when no ambiguity may arise. A popular choice of a criterion function is $\mathcal{C}_n(\theta;\, X^{(n)}) = -\alpha\, n\, \mathcal{R}_n(\theta)$, where

$$\mathcal{R}_n(\theta) := n^{-1} \sum_{i=1}^n \ell(X_i,\, \theta)$$

is the empirical risk function induced from a loss function $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$, and $\alpha \in (0, \infty)$ is the learning rate parameter. The corresponding Bayesian pseudo-posterior is called the

Gibbs posterior associated with loss function ℓ in the literature (e.g. Bhattacharya and Martin, 2020; Syring and Martin, 2020). In particular, the usual Bayesian posterior (1) is a special case when the loss function is $\ell(X,\theta) = -\log p(X \,|\, \theta)$ and $\alpha = 1$. For Bayesian quantile regression, we may take the check loss function $\ell(x,q) = (q-x)\cdot\big(\tau - \mathbf{1}(q<x)\big)$ for a given quantile level $\tau \in (0,1)$, since the τ-th quantile of any one-dimensional random variable X corresponds to the population risk function minimizer $\arg\min_{q\in\mathbb{R}} \mathbb{E}[\ell(X,q)]$.

A direct computation of either the posterior $p(\theta \,|\, X^{(n)})$ or the pseudo-posterior (2) involves the normalisation constant (the denominator) as a d-dimensional integral, which is often analytically intractable unless the prior distributions form a conjugate family to the likelihood (criterion) function. In practice, Markov chain Monte Carlo (MCMC) algorithm (Hastings, 1970; Geman and Geman, 1984; Robert et al., 2004) is instead employed as an automatic machinery for sampling from the (pseudo-)posterior, whose implementation is free of the unknown normalisation constant. The aim of this paper is to provide a rigorous theoretical analysis on the computational complexity of a popular and widely used class of MCMC algorithms described below. In particular, we are interested in characterizing a sharp dependence of their mixing times on the parameter dimension in the context of Bayesian posterior sampling.

2.2 Metropolis-adjusted Langevin algorithm

Consider a generic (possibly unnormalized) density function $f(\theta) = \exp\{-U(\theta)\}$ defined on a set $\Theta \subset \mathbb{R}^d$, where $U : \Theta \to \mathbb{R}$ is called the potential (function) associated with f. For example, in the Bayesian setting with target posterior (2), we can take $U(\theta) = -\log \pi(\theta) - \mathcal{C}_n(\theta;\, X^{(n)})$. Suppose our goal is to sample from the probability distribution μ induced by f, where $\mu(A) = \frac{\int_A f(\theta)\,\mathrm{d}\theta}{\int_\Theta f(\theta)\,\mathrm{d}\theta}$ for any measurable set $A \subset \Theta$. Metropolis-adjusted Langevin algorithm (MALA), as an instance of MCMC with a special design of the proposal distribution, aims at producing a sequence of random points $\{\theta_k\}_{k\geq 0}$ in Θ such that the distribution of θ_k approaches μ as k tends to infinity, so that for sufficiently large k_0, the k_0-th iterate θ_{k_0} can be viewed as a random variable approximately sampled from the target distribution μ. In practice, every k_0 iterates from the chain can be collected (called thinning), which together form approximately independent draws from μ.

Specifically, given step size $\widetilde{h} > 0$ and initial distribution μ_0 on Θ, MALA produces $\{\theta_k\}_{k\geq 0}$ sequentially as follows: for $k = 0, 1, 2, \ldots$,

1. (**Initialization**) If $k = 0$, sample θ_0 from μ_0;

2. (**Proposal**) If $k \geq 1$, given previous state θ_{k-1}, generate a candidate point y_k from proposal distribution $N_d\big(\theta_{k-1} - \widetilde{h}\nabla U(\theta_{k-1}),\, 2\widetilde{h}\, I_d\big)$ whose density function is denoted as $Q(\theta_{k-1},\, \cdot)$, or equivalently,

$$y_k = \theta_{k-1} - \widetilde{h}\,\nabla U(\theta_{k-1}) + \sqrt{2\widetilde{h}}\, z_k, \quad \text{with } z_k \sim N_d(0,\, I_d).$$

3. (**Metropolis-Hasting rejection/correction**) Set acceptance probability $A(\theta_{k-1},\, y_k) := 1 \wedge \alpha(\theta_{k-1},\, y_k)$ with acceptance ratio statistic

$$\alpha(\theta_{k-1}, y_k) := \frac{f(y_k)\,\cdot\, Q(y_k,\, \theta_{k-1})}{f(\theta_{k-1})\cdot Q(\theta_{k-1},\, y_k)}.$$

Flip a coin and accept y_k with probability $A(\theta_{k-1},\, y_k)$ and set $\theta_k = y_k$; otherwise, set $\theta_k = \theta_{k-1}$.

It is straightforward to verify that MALA described above produces a Markov chain whose transition kernel is

$$T(\theta,\, \mathrm{d}y) = \Big(\underbrace{1 - \int_\Theta A(\theta,\, y)\, Q(\theta,\, y)\, \mathrm{d}y}_{\text{rejection probability}}\Big) \cdot \delta_\theta(\mathrm{d}y) + A(\theta,\, y)\, Q(\theta,\, y)\, \mathrm{d}y, \tag{3}$$

where δ_θ denotes the point mass measure at θ. In practice, the target density f can be non-smooth at certain point $\theta \in \Theta$, and we address this issue by replacing the gradient $\nabla U(\theta)$ with any of its subgradient $\widetilde{\nabla} U(\theta)$[1] in MALA. That means, the proposal distribution Q is being chosen as $N_d(\theta_{k-1} - \widetilde{h}\, \widetilde{\nabla} U(\theta_{k-1}), 2\widetilde{h})$ and other aspects of the MALA algorithm remain unchanged. Furthermore, MALA can be generalized by introducing a symmetric positive-definite preconditioning matrix $\widetilde{I} \in \mathbb{R}^{d\times d}$, so that the proposal Q in MALA is modified as $N_d(\theta_{k-1} - \widetilde{h}\widetilde{I}\, \widetilde{\nabla} U(\theta_{k-1}), 2\widetilde{h}\, \widetilde{I})$. It has been shown that (Girolami and Calderhead, 2011; Vacar et al., 2011) for a suitable preconditioning matrix, the resulting preconditioned MALA can help to alleviate the issue caused by the anisotropicity of the target measure. We illustrate both empirically (c.f. Appendix A.1) and theoretically (c.f. Corollary 7) that a suitable preconditioning matrix may improve the convergence of the sampling algorithm for Bayesian posteriors. As a common practice (Chen et al., 2020; Lovász and Simonovits, 1993) to simplify the analysis of MALA, in this paper, we consider the ζ-lazy version of MALA, where at each iteration, the chain is forced to remain unchanged with probability ζ. The corresponding Markov transition kernel of the ζ-lazy version of MALA is given by

$$T^\zeta(\theta, \mathrm{d}y) = \big(1 - (1-\zeta) \cdot \int_\Theta A(\theta, y)\, Q(\theta, y)\, \mathrm{d}y\big) \cdot \delta_\theta(\mathrm{d}y) + (1-\zeta) \cdot A(\theta, y) Q(\theta, y) \mathrm{d}y. \tag{4}$$

A closely related algorithm is the unadjusted Langevin algorithm (ULA, Durmus and Moulines, 2017; Cheng et al., 2018; Roberts and Tweedie, 1996; Dalalyan, 2017), which corresponds to discretization of the following Langevin stochastic differential equation (SDE),

$$\mathrm{d}X_t = -\nabla U(X_t)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t, \quad t > 0,$$

and does not have the Metropolis-Hasting correction step 3. As a consequence, the stationary distribution of ULA is of order $\mathcal{O}(\sqrt{dh})$ away from μ under several commonly used metrics (Durmus et al., 2019). Due to this error, even in the strongly log-concave scenario, unlike MALA which requires at most poly-log$(1/\varepsilon)$ iterations with a constant step size h to get one sample distributed close from μ with accuracy ε, ULA requires poly-$(1/\varepsilon)$ iterations and an ε-dependent choice of h (Durmus et al., 2019).

Another closely related algorithm is the classical Metropolis random walk (MRW), which instead uses $N_d\big(\theta_{k-1},\, 2\widetilde{h}\, I_d\big)$ without the gradient term in the proposal distribution Q. As we will see, by using the extra gradient information, the dimension dependence of the mixing time can be improved from $\mathcal{O}(d)$ (Gelman et al., 1997; Dwivedi et al., 2019) to $\mathcal{O}(d^{1/3})$ for sampling from Bayesian posteriors.

[1] A subgradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ at point $x \in \mathbb{R}^d$ is a vector $g \in \mathbb{R}^d$ such that $f(y) \geq f(x) + \langle g,\, y - x\rangle + \mathcal{O}(\|y - x\|)$ as $y \to x$

2.3 Problem setup

The goal of this paper is to characterize the mixing time of MALA for sampling from the Bayesian pseudo-posterior π_n defined in (2). Assume we have access to a *warm start* defined as follows.

Definition 1 *We say μ_0 is an M_0-warm start with respect to the stationary distribution μ, if $\mu_0(E) \leq M_0\, \mu(E)$ holds for all Borel set $E \subset \mathbb{R}^d$, and we call M_0 the warming parameter.*

We state our problem as *characterizing the ε-mixing time in χ^2 divergence of the Markov chain produced by (preconditioned) MALA starting from an arbitrary M_0-warm start μ_0 for obtaining draws from $\pi_n(\theta)$*, which is mathematically defined as the maximum of the minimal number of steps required for the chain to be within ε^2-χ^2 divergence from its stationary distribution, over M_0-warm starts, or

$$\tau_{\text{mix}}(\varepsilon, M_0) = \max\left\{\tau_{\text{mix}}(\varepsilon, \mu_0)\,:\,\mu_0 \text{ is an } M_0\text{-warm start with respect to } \pi_n\right\}$$
$$\text{with } \tau_{\text{mix}}(\varepsilon, \mu_0) = \inf\big\{k \in \mathbb{N}:\, \sqrt{\chi^2\big(\mu_k,\, \pi_n\big)} \leq \varepsilon\big\},$$

where μ_k denotes the probability distribution obtained after k steps of the Markov chain. Note that a mixing time upper bound in χ^2 divergence implies that in total variation distance since $\|p - q\|_{\text{TV}} \leq \sqrt{\chi^2(p,\, q)}$.

3. Mixing Time Bounds via s-Conductance Profile

In this section, we introduce a general technique of using s-conductance profile to bound the mixing time of a Markov chain. We first review some common concepts and previous results in Markov chain convergence analysis, and then provide an improved analysis for obtaining a sharp mixing time upper bound of MALA in this work.

Ergodic Markov chains: Given a Markov transition kernel $T(\cdot,\, \cdot)$ with stationary distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$, the ergodic flow of a set S is defined as

$$\phi(S) = \int_S \left\{\int_{S^c} T(\xi,\, \mathrm{d}y)\right\} \mu(\mathrm{d}\xi).$$

The ergodic flow captures the mass of points leaving S (i.e., $T(\xi, S^c) = \int_{S^c} T(\xi,\, \mathrm{d}y)$) on average under stationary distribution μ in one step of the Markov chain. A Markov chain is said to be ergodic if $\phi(S) > 0$ for all measurable set $S \subset \mathbb{R}^d$ with $0 < \mu(S) < 1$. Let μ_k denote the probability distribution obtained after k steps of a Markov chain. If the Markov chain is ergodic, then $\mu_k \to \mu$ as $k \to \infty$ in total variation distance; see, for example, Corollary 1.6 of Lovász and Simonovits (1993).

Conductance of Markov chain and rapid mixing: The (global) conductance of an ergodic Markov chain characterizes the least relative ratio between $\phi(S)$ and the measure $\mu(S)$ of S, and is formally defined as

$$\Phi = \inf\left\{\frac{\phi(S)}{\mu(S)}\,:\, 0 < \mu(S) \leq \frac{1}{2}\right\}.$$

A Markov chain with low conductance tends to become trapped in a subset of its states, whereas one with high conductance has more freedom to explore and transition across its entire state space. The conductance is related to the spectral gap[2] of the Markov chain via Cheeger's inequality (Cheeger, 2015), and thus can be used to characterize the convergence of the Markov chain. For example, Corollary 1.5 in Lovász and Simonovits (1993) shows that if μ_0 is an M_0-warm start with respect to the stationary distribution μ, then

$$\|\mu_k - \mu\|_{\mathrm{TV}} \leq \sqrt{M_0}\left(1 - \frac{\Phi^2}{2}\right)^k, \quad k \geq 0.$$

Furthermore, some people consider the more flexible notion of s-conductance, defined as

$$\Phi_s := \inf\left\{\frac{\phi(S)}{\mu(S) - s} \,:\, s < \mu(S) \leq \frac{1}{2}\right\}, \quad \text{for } s \in (0, 1/2),$$

which restricts the infimum over all sets with a probability greater than s. This restriction avoids including sets in the conductance bound that have poor conductance but receive negligible probability, which should be less significant to the overall mixing of the Markov chain. Specifically for sampling from Bayesian posteriors, this refined analysis allows us to focus our calculations on these "highest posterior regions" while avoiding some unwieldy tail probability regions (e.g., the region defined in Condition A.3). Using the s-conductance, Corollary 1.6 in Lovász and Simonovits (1993) proves a similar bound implying the exponential convergence of the algorithm up to accuracy level s as

$$\|\mu_k - \mu\|_{\mathrm{TV}} \leq M_0\, s + M_0\left(1 - \frac{\Phi_s^2}{2}\right)^k, \quad k \geq 0.$$

Consequently, the ε-mixing time with respect to the total variation distance of the Markov chain starting from an M_0-warm start can be upper bounded by $\frac{2}{\Phi_s^2}\log\frac{2M_0}{\varepsilon}$ if we choose $s = \frac{\varepsilon}{2M_0}$.

Conductance profile of Markov chain: Instead of controlling mixing times via a worst-case conductance bound, some recent works have introduced more refined methods based on the conductance profile. The conductance profile is defined as the following collection of conductance,

$$\Phi(v) := \inf\left\{\frac{\phi(S)}{\mu(S)} \,:\, 0 < \mu(S) \leq v\right\}, \quad \text{indexed by } v \in \left(0, \frac{1}{2}\right].$$

Note that the classic conductance constant Φ is a special case that can be expressed as $\Phi = \Phi(\frac{1}{2})$. Based on the conductance profile, Chen et al. (2020) consider the concept of Ω-restricted conductance profile for a convex set Ω, given by

$$\Phi_\Omega(v) := \inf\left\{\frac{\phi(S)}{\mu(S \cap \Omega)} \,:\, 0 < \mu(S \cap \Omega) \leq v\right\}, \quad v \in \left(0, \frac{\mu(\Omega)}{2}\right].$$

It has been shown in Chen et al. (2020) that given an M_0-warm start μ_0, if

$$\mu(\Omega) \geq 1 - \frac{\varepsilon^2}{3M_0^2} \quad \text{and} \quad \Phi_\Omega(v) \geq \sqrt{B \log\frac{1}{v}} \quad \text{for all } v \in \left[\frac{4}{M_0}, \frac{1}{2}\right],$$

[2] The spectral gap is define as $\Lambda = \inf\{\mathcal{E}(f,f)/\mathrm{Var}_\mu(f) \,:\, f \in L^2(\mu), \mathrm{Var}_\mu(f) > 0\}$, where $\mathcal{E}(f,g) = \int (f(x) - g(y))^2 T(x, \mathrm{d}y)\, \mathrm{d}\mu(x)$ is the *Dirichlet form*.

then the ε-mixing time in χ^2 divergence of the chain is bounded from above by $\mathcal{O}\big(\frac{1}{B}\log(\frac{\log M_0}{\varepsilon})\big)$. Therefore, compared with the (global) conductance, employing the technique of conductance profile may improve the warming parameter dependence in the mixing time bound from $\log M_0$ to $\log\log M_0$. This improvement from a logarithmic dependence to the double logarithmic dependence may dramatically sharpen the mixing time upper bound, since in a typical Bayesian setting M_0 may grow exponentially in the dimension d. However, one drawback of the conductance profile technique from Chen et al. (2020) is that the high probability set Ω should be constrained to be convex (Lemma 4 of Chen et al. (2020)) to bound the Ω-restricted conductance profile $\Phi_\Omega(v)$. This convexity constraint may cause $\Phi_\Omega(v)$ to have a worse dimension dependence compared with the complexity analysis using the s-conductance Φ_s.

In order to address the above issues of previous analysis, we introduce the following notion of *s-conductance profile* , which combines ideas from the s-conductance and conductance profile,

$$\Phi_s(v) := \inf\left\{\frac{\phi(S)}{\mu(S)-s}\,\middle|\, s<\mu(S)\le v\right\} \quad \text{indexed by } \; s\in\Big(0,\,\frac{1}{2}\Big) \; \text{ and } \; v\in\Big(s,\,\frac{1}{2}\Big].$$

The s-conductance profile evaluated at $v=\frac{1}{2}$ corresponds to the s-conductance that is commonly-used in previous study for analyzing the mixing time of Markov chain (Chewi et al., 2021; Dwivedi et al., 2019). We show in the following lemmas that a lower bound on the s-conductance profile can be translated into an upper bound on the mixing time in χ^2-squared divergence. We formulate here an informal result and postpone a more detailed statement to Appendix A.2.

Lemma 2 (Mixing time bound via s-conductance profile (informal)) *For any error tolerance $\varepsilon\in(0,1)$, the mixing time in χ^2 divergence of the ζ-lazy version of MALA over M_0-warm starts can be bounded as*

$$\tau_{\mathrm{mix}}(\varepsilon, M_0) \lesssim \zeta^{-1}\cdot\Big(\int_{\frac{4}{M_0}}^{\frac{1}{2}}\frac{\mathrm{d}v}{v\,\Phi_s^2(v)} + \frac{1}{\Phi_s^2(\frac{1}{2})}\log(\frac{1}{\varepsilon})\Big), \quad s=\frac{\varepsilon^2}{16M_0^2}.$$

It is worth noting that since $\Phi_s(v)$ is a decreasing function of v, by replacing $\Phi_s(v)$ with its lower bound $\Phi_s=\Phi_s(\frac{1}{2})$, one can obtain a mixing time bound via s-conductance. However, instead of simply considering the worst case, the integral $\int_{\frac{4}{M_0}}^{\frac{1}{2}}\frac{\mathrm{d}v}{v\,\Phi_s^2(v)}$ averages over $\Phi_s(v)$, offering a possible improvement in the dependence on warming parameter M_0. To establish a lower bound for the s-conductance profile, we can employ the “overlap argument” frequently used in the literature (Chewi et al., 2021; Chen et al., 2020; Belloni and Chernozhukov, 2009; Wu et al., 2022), that is, 1. prove a log-isoperimetric inequality for μ; 2. bound the total variation distance between $T(x,\cdot)$ and $T(z,\cdot)$ for any two sufficiently close points x,z in a high probability set (not necessarily convex) of μ. We leave a detailed description of this argument to Appendix A.2.

Among previous works of mixing time analysis of MALA, Chen et al. (2020) study the problem of sampling from general smooth and strongly log-concave densities, using the technique of Ω-restricted conductance profile. Their bound has a double logarithmic

$\log\log M_0$ dependence on the warmth parameter M_0 under certain regime (of step size h), and a sub-optimal $\mathcal{O}(d)$-dependence on the dimension. On the other hand, Chewi et al. (2021) study the same problem as Chen et al. (2020) and obtain a mixing time bound with an optimal $\mathcal{O}(d^{\frac{1}{2}})$-dependence, based on the s-conductance technique. However, the bound in Chewi et al. (2021) has a quadratic dependence on $\log M_0$. By utilizing our s-conductance profile argument, when $\log M_0$ and h^{-1} are not of constant order, we can improve their bounds from $h^{-1}\log(\frac{M_0}{\epsilon})$ to $\max\{h^{-1}\log(\frac{\log M_0}{\epsilon}), \log M_0\}$, where h is the step size used in Theorem 3 of Chewi et al. (2021).

4. Mixing Time of MALA

In this section, we describe our main result by providing an upper bound to the mixing time of (preconditioned) MALA for sampling from the Bayesian pseudo-posterior π_n. We consider the ζ-lazy version of MALA and assume that a warm start is accessible, which is a common assumption (e.g. Dwivedi et al., 2019; Mangoubi and Vishnoi, 2019). For example, Corollary 7 in Section 5.1 provides a construction of M_0-warm start for general Gibbs posterior with smooth criterion function, where M_0 is bounded above by an (n,d)-independent constant.

Note that the Bayesian pseudo-posterior with criterion function $\mathcal{C}_n$ can be rewritten as

$$\pi_n(\theta\,|\,X^{(n)}) = \frac{\exp\big\{-V_n\big(\sqrt{n}(\theta-\widehat{\theta})\big)\big\}}{\int_\Theta \exp\big\{-V_n\big(\sqrt{n}(\theta-\widehat{\theta})\big)\big\}\,\mathrm{d}\theta} \quad \forall\theta\in\Theta, \tag{5}$$

$$\text{where} \quad \hat{\theta} = \operatorname*{arg\,max}_{\theta\in\Theta} \mathcal{C}_n(\theta) \quad \text{and} \tag{6}$$

$$V_n(\xi) = -\mathcal{C}_n\Big(\widehat{\theta}+\frac{\xi}{\sqrt{n}}; X^{(n)}\Big) + \mathcal{C}_n\big(\widehat{\theta}\,; X^{(n)}\big) - \log\pi\Big(\widehat{\theta}+\frac{\xi}{\sqrt{n}}\Big) + \log\pi(\widehat{\theta}\,)$$

is the corresponding rescaled potential (function). In the expression of V_n, we deliberately added two terms independent of ξ so that $V_n(0)=0$ for simplifying the analysis. Motivated by the classical Bernstein-von Mises (BvM) theorem[3] (van der Vaart, 2000; Ghosh and Ramamoorthi, 2003) for Bayesian posteriors, we impose following conditions on V_n, stating that $V_n(\xi)$ is close to a quadratic form and the subgradient of $V_n(\xi)$ employed in MALA is close to a linear form, uniformly over a high probability set of the rescaled target measure $\pi_{\text{loc}} = (\sqrt{n}(\cdot-\widehat{\theta}))_{\#}\pi_n$.[4] Here π_{loc} corresponds to the measure of the localized random variable $\xi=\sqrt{n}(\theta-\widehat{\theta})$ for $\theta\sim\pi_n(\theta|X^{(n)})$, and the transformation $\sqrt{n}(\cdot-\widehat{\theta})$ makes the limiting distribution of ξ zero-centered and has constant-order variances.

Condition A: *There exists a tolerance $\varepsilon\in(0,1)$, preconditioning matrix $\widetilde{I}$, step size parameter h (rescaled by n), warming parameter M_0, numbers $R, \widetilde{\varepsilon}_0, \widetilde{\varepsilon}_1 \geq 0$, $\rho_1, \rho_2 > 0$ and a symmetric positive definite matrix $J\in\mathbb{R}^{d\times d}$ so that*

[3] When sample size n is large, the Bayesian posterior is close to the Gaussian distribution $N_d(\widehat{\theta}_{\text{MLE}}, n^{-1}\mathcal{J}^{-1})$, where $\widehat{\theta}_{\text{MLE}}$ is the maximum likelihood estimator and $\mathcal{J}$ the Fisher information matrix.

[4] We use $\mu = G_{\#}\nu$ to denote the push forward measure so that for any measurable set A, $\mu(A) = \nu(G^{-1}(A))$.

1. *for any* $\xi \in K = \{x : \|\widetilde{I}^{-1/2}x\| \leq R\}$[5]

$$\big|V_n(\xi) - \frac{1}{2}\xi^T J\xi\big| \leq \widetilde{\varepsilon}_0 \quad \textit{and} \quad \big\|\widetilde{\nabla} V_n(\xi) - J\xi\| \leq \widetilde{\varepsilon}_1,$$

   where $\widetilde{\nabla} V_n(\xi)$ *is a subgradient of* $V_n(\xi)$*;*

2. $\rho_1 I_d \preceq \widetilde{J} = \widetilde{I}^{1/2} J \widetilde{I}^{1/2} \preceq \rho_2 I_d$*;*

3. $\pi_n\big(\sqrt{n}\,\|\widetilde{I}^{-1/2}(\theta - \widehat{\theta})\| \leq R/2\big) \geq 1 - \exp(-4\widetilde{\varepsilon}_0) \cdot \frac{h\rho_1\varepsilon^2}{M_0^2}$ *and* $R \geq 8\sqrt{d/\lambda_{\min}(\widetilde{J})}\,.$

The first inequality in Condition A.1 requires that $V_n(\xi)$ can be uniformly approximated by the quadratic term $\frac{1}{2}\xi^T J\xi$ with an approximation error $\widetilde{\varepsilon}_0$. This requirement is implied by the classical BvM result, which is commonly utilized in MCMC mixing time analysis for Bayesian posterior sampling (Belloni and Chernozhukov, 2009; Ascolani and Zanella, 2023). It is noteworthy that we do not impose any smoothness or convexity constraints on $V_n(\xi)$, and the deviation characteristic $\widetilde{\varepsilon}_0$ can take any value. We also keep track of the impact of this deviation in the final mixing time bound, as reflected in Theorem 3, where we explicitly show the dependency of the mixing time on this approximation error, $\widetilde{\varepsilon}_0$. The result reveals that the mixing time exhibits an exponential dependence on $\widetilde{\varepsilon}_0$. The second inequality in Condition A.1 assumes that the subgradient of $V_n(\xi)$ can be approximated by the linear term $J\xi$ with an approximation error $\widetilde{\varepsilon}_1$. Although less standard, this condition is crucial since $\widetilde{\varepsilon}_1$ governs the efficacy of the subgradient used in MALA to adjust the proposal distribution and facilitate faster exploration of the parameter space. As we will see in Theorem 3, a small $\widetilde{\varepsilon}_1$ enables MALA, leveraging (sub)gradient information, to improve upon MRW in terms of mixing time. Condition A.2 requires the asymptotic covariance matrix J, after rescaling by the preconditioning matrix, to have its maximum eigenvalue upper-bounded by ρ_2 and its minimum eigenvalue lower-bounded by ρ_1. The condition number $\kappa = \frac{\rho_2}{\rho_1}$ serves as an indicator of how well the preconditioning matrix $\widetilde{I}$ is chosen to alleviate issues arising from the anisotropy of the target distribution. As we will see from Theorem 3, a small κ will lead to a lower mixing time. The last condition (Condition A.3) assumes that the radius R of the compact set K, considered in Condition A.1, is sufficiently large. This ensures that K is a high probability set under π_{loc}. This assumption guarantees that the region where the density π_{loc} (or π_n) deviates significantly from a Gaussian form, and is possibly non-smooth and non-log-concave, is negligible, thereby reducing the chances of the Markov chain becoming trapped in such regions.

In summary, Condition A requires the localized (rescaled) posterior $\pi_{\text{loc}} = (\sqrt{n}(\cdot - \widehat{\theta}))_{\#}\pi_n$ to be close to a Gaussian distribution $N_d(0, J^{-1})$, so that we can analyze the mixing time of MALA for sampling π_n or π_{loc} (note that the complexity for sampling from π_n with step size $\widetilde{h} = h/n$ is equivalent to that from π_{loc} with rescaled step size h) by comparing its transition kernel T expressed in (4) with the transition kernel T^{Δ} induced from the MALA for sampling the Gaussian distribution. Interestingly, we find that as long as the deviance of π_{loc} to Gaussian is sufficiently small but not necessarily diminishing as $n, d \to \infty$, some key properties (more precisely, conductance lower bound) of T^{Δ} guarantee that the fast

[5] Here the notation $A^{-1/2}$ of a symmetric positive definite matrix A means the inverse of its matrix square root $A^{1/2}$.

mixing of MALA will be inherited by T, so that the mixing time associated with T can be controlled. Using this argument, we prove a mixing time upper bound without imposing the smoothness and strongly convexity assumptions on $V_n(\xi)$ that are restrictive and commonly assumed in the literature for analyzing the convergence of MALA (Chewi et al., 2021; Chen et al., 2020). As a concrete example, under mild assumptions, Condition A holds for a broad class of Gibbs posteriors (Bhattacharya and Martin, 2020) mentioned in Section 2.1 where the criterion function $\mathcal{C}_n$ is proportional to the negative empirical risk function $\mathcal{R}_n$, as long as d is relatively small compared to n (see Lemma 18 and Lemma 19 in Appendix B.3 for details). Now we are ready to state the following theorem.

Theorem 3 (MALA mixing time upper bound) *Let π_n defined in* (5) *be the target distribution and $\zeta \in (0, \frac{1}{2}]$ be a lazy parameter. Assume Condition A holds for a tolerance ε, warming parameter M_0, sample size n, preconditioning matrix $\widetilde{I}$, rescaled step size h, and some $R > 0$, $\widetilde{\varepsilon}_1 \geq 0$, $\rho_2 \geq \rho_1 > 0$, and that there exists a small enough absolute (n, d)-independent constant c_0 so that the step size can be expressed as $\widetilde{h} = h/n$ with*

$$h = c_0 \cdot \bigg[\rho_2 \Big(d^{\frac{1}{3}} + d^{\frac{1}{4}}\Big(\widetilde{\varepsilon}_0 + \log\frac{M_0 d\kappa}{\varepsilon}\Big)^{\frac{1}{4}} + \Big(\widetilde{\varepsilon}_0 + \log\frac{M_0 d\kappa}{\varepsilon}\Big)^{\frac{1}{2}} + |||\widetilde{I}|||_{\mathrm{op}} R^2 \widetilde{\varepsilon}_1^2\Big)\bigg]^{-1}, \ \textit{where } \kappa = \frac{\rho_2}{\rho_1},$$

then the ζ-lazy version of MALA with proposal distribution $N_d(\theta_{k-1} - \widetilde{h}\widetilde{I}\,\widetilde{\nabla}U(\theta_{k-1}), 2\widetilde{h}\,\widetilde{I})$ and step size $\widetilde{h}$ has a maximal ε-mixing time in χ^2 divergence over M_0-warm starts being bounded as

$$\tau_{\mathrm{mix}}(\varepsilon, M_0) \leq \frac{C_1 \exp(4\widetilde{\varepsilon}_0)}{\zeta} \cdot \bigg\{\bigg[\rho_1^{-1} \exp(8\widetilde{\varepsilon}_0) \cdot h^{-1} \log\big(\frac{\log M_0}{\varepsilon}\big)\bigg] \vee \log M_0\bigg\}, \tag{7}$$

where C_1 is an (n, d)-independent constant.

The mixing time bound (7) is proved using the technique of s-conductance profile introduced in Section 3. A similar mixing time bound can be obtained if when consider the sampling of π_{loc} constrained on the high probability set $K = \{x : \|\widetilde{I}^{-1/2}x\| \leq R\}$, which is adopted by Belloni and Chernozhukov (2009) for analyzing the mixing time of MRW; however, our result does not require such a constraining step. According to Theorem 3, for a fixed tolerance (accuracy level) ε, the ε-mixing time is determined by the parameter dimension d, warming parameter M_0, preconditioning matrix $\widetilde{I}$, approximation errors $\widetilde{\varepsilon}_0$, $\widetilde{\varepsilon}_1$ of the potential and the gradient, radius R of the high probability set of π_{loc} and the precision matrix J of the Gaussian approximation to π_{loc}. The derived mixing time bound is exponentially dependent on $\widetilde{\varepsilon}_0$, implying that a bound that is polynomial in d can only be attained if $\widetilde{\varepsilon}_0$ is either constant-order or logarithmic in d. The fourth term $|||\widetilde{I}|||_{\mathrm{op}} R^2 \widetilde{\varepsilon}_1^2$ in the expression of h will be dominated by others once $\widetilde{\varepsilon}_1$ is sufficiently small. For example, suppose $\widetilde{I} = I_d$, $\log\frac{M_0\kappa}{\varepsilon} = \mathcal{O}(d)$ and π_{loc} has a sub-Gaussian type tail behavior, or

$$\pi_{\mathrm{loc}}\big(\|\xi\| \geq c_1(\sqrt{d} + t)\big) \leq \exp(-c_2\, t^2), \quad t > 0,$$

then we can choose the radius as $R = \mathcal{O}(\sqrt{d})$, and the term $|||\widetilde{I}|||_{\mathrm{op}} R^2 \widetilde{\varepsilon}_1^2$ will be dominated by the $\mathcal{O}(d^{\frac{1}{3}})$ term once $\widetilde{\varepsilon}_1 = \mathcal{O}(d^{-\frac{1}{3}})$. This suggests that a $d^{\frac{1}{3}}$-mixing time upper bound is achievable as long as the (sub)gradient used in MALA deviates from a linear form with approximation error at most $d^{-\frac{1}{3}}$, which is independent of the sample size. Therefore, when $d \ll n$, it is safe to fix a mini-batch dataset for computing the (sub)gradient in MALA instead of using the full batch. As another remark, our theorem also gives a tight mixing time upper bound $\mathcal{O}(d)$ of MRW by taking $\widetilde{\varepsilon}_1 = O(1)$, corresponding to the case where the gradient estimate is completely uninformative.

Our mixing time bound has a linear dependence (modulo logarithmic term) on the condition number $\kappa = \rho_2/\rho_1$, which matches the best condition number dependence for MALA under strong convexity (Wu et al., 2022) and we show the tightness of the condition number dependence in Theorem 12 of Appendix A.3. Moreover, by introducing preconditioning matrix $\widetilde{I}$, a small condition number can be obtained once $\widetilde{I}$ acts as a reasonable estimator to J^{-1}, which will lead to a faster mixing time when J is ill-conditioned. On the other hand, assume κ is bounded above by an (n, d)-independent constant and

$$\big(|\!|\!|\widetilde{I}|\!|\!|_{\mathrm{op}}\, R^2\, \widetilde{\varepsilon}_1^{2} \big) \vee \log\Big(\frac{M_0}{\varepsilon}\Big) \leq d^{\frac{1}{3}},$$

we have $\tau_{\mathrm{mix}}(\varepsilon, \mu_0) \leq C_1\, d^{\frac{1}{3}} \log(\frac{\log M_0}{\varepsilon})$. This upper bound matches the lower bound proved in Chewi et al. (2021) that the mixing time of MALA for sampling from the standard Gaussian target is at least $\mathcal{O}(d^{\frac{1}{3}})$, and it improves the warming parameter dependence from $\log M_0$ to $\log(\log M_0)$ compared with the upper bound proved in Chewi et al. (2021). Therefore, in order to attain the best achievable mixing time $\mathcal{O}(d^{\frac{1}{3}})$, we need to find a initial distribution μ_0 that is close to π_n, so that the warming parameter M_0 can be controlled. For a generic log-concave distribution, it has been shown that a warm start with warming parameter M_0 polynomial in d can be obtained with $d^{\frac{1}{2}}$ complexity, as demonstrated by Altschuler and Chewi (2023). However, efficiently obtaining a poly(d) warm start for general non-log-concave sampling problems is infeasible. Fortunately, in our Bayesian posterior sampling context, although π_n may not be log-concave, large sample asymptotic theory (refer to Section 5, for instance) ensures that π_n is approximately Gaussian. Therefore, using the Gaussian distribution $N_d(\widehat{\theta}, n^{-1}\widetilde{I})$, constrained on a compact set, as the initialization μ_0, is a natural choice. To support this initialization scheme, the following lemma provides an upper bound for the corresponding warming parameter M_0.

Lemma 4 (Warming parameter control) *Suppose Condition A is satisfied. For any compact set $K \subset \mathbb{R}^d$, the initial distribution as*

$$\mu_0 = N_d(\widehat{\theta}, n^{-1}\widetilde{I})|_{\{\theta\,:\,\sqrt{n}(\theta-\widehat{\theta})\in K\}}$$

is M_0-warm with respect to π_n, where

$$\log M_0 \leq -\log \pi_n\big(\{\theta \,:\, \sqrt{n}(\theta - \widehat{\theta}) \in K\}\big) \,+\, \sup_{\xi\in K}\big|\, \xi^T(\widetilde{I}^{-1} - J)\xi\big| \,+\, 2\cdot \sup_{\xi\in K}\big|V_n(\xi) - \frac{1}{2}x^T Jx\big|.$$

In order to construct a feasible warm start using Lemma 4, it is necessary to compute the maximizer $\widehat{\theta}$ of the criterion function $\mathcal{C}_n(\theta)$. An inaccurate approximation of $\widehat{\theta}$ may cause the warming parameter M_0 to grow linearly with the sample size n, a similar observation also noted in studies by Ascolani and Zanella (2023); Belloni and Chernozhukov (2009). While it is generally challenging to obtain solutions for non-convex optimization problems, there are cases where optimizing a nearly quadratic function can be much easier compared to sampling from a nearly Gaussian distribution. A specific example is Bayesian quantile regression, where the estimation of $\hat{\theta}$ can be efficiently achieved using linear programming techniques. Our theoretical results also suggest that under Condition A, we can control the warming parameter M_0 in MALA by choosing a reasonable estimator $\widetilde{I}$ for the inverse asymptotic covariance matrix J^{-1} of π_{loc}. For instance, if $\widetilde{I}$ is chosen to be the identity matrix and J has a bounded operator norm, then $\log M_0$ should be of order $\mathcal{O}(d)$. Furthermore, in Bayesian Gibbs posterior sampling, where the loss function ℓ is continuously twice differentiable, a viable option for approximating J^{-1} could be the plug-in estimator:

$$\widetilde{I} = \left\{ \frac{1}{|S|} \sum_{i \in S} \mathrm{Hess}_{\theta}(\ell(X_i, \widehat{\theta})) \right\}^{-1},$$

where S is a subset of $1, 2, \cdots, n$, and $\mathrm{Hess}_{\theta}(\ell(x, \theta))$ denotes the Hessian matrix of $\ell(x, \cdot)$ evaluated at θ. Notably, since the warming parameter M_0 can be of order $\mathcal{O}(d^{1/3})$ for achieving the best possible mixing time, it is feasible to compute the plug-in estimator using only a mini-batch of data, the size of which depends solely on the dimension, rather than the full dataset. Further details can be found in Corollary 7.

According to Lemma 4 and Theorem 3, a reasonably good approximation $\widetilde{I}$ to matrix J in Condition A will improve both the mixing time of MALA after burn-in period and the initialization affecting the burn-in. For completeness, we also provide an experiment in Appendix A.1 for investigating the impact of the preconditioning matrix and initial distribution on the performance of MALA. However, in some complicated problems, especially when $\log \pi_{\mathrm{loc}}$ is not differentiable, a good estimator for the matrix J may not be easy to construct. One possible strategy is to use adaptive MALA (Atchadé, 2006), where the preconditioner $\widetilde{I}$ and step size h are updated in each iteration by using the history draws. It has been empirically shown in Atchadé (2006) that adaptive MALA outperforms non-adaptive counterparts in many interesting applications. We leave a rigorous theoretical analysis of adaptive MALA as a future direction.

5. Sampling from Gibbs Posteriors

Recall from Section 2.1 that a Gibbs posterior is a Bayesian pseudo-posterior defined in (2) with the criterion function $\mathcal{C}_n(\theta;\, X^{(n)}) = -\alpha\, n\, \mathcal{R}_n(\theta)$, where α is an (n, d)-independent positive learning rate and $\mathcal{R}_n(\theta) := n^{-1} \sum_{i=1}^{n} \ell(X_i,\, \theta)$ is the empirical risk function induced from a loss function $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$. In this section, we first provide generic conditions under which Condition A for Theorem 3 can be verified for the the Gibbs posterior so that the mixing time bound of the corresponding MALA can be applied. After that, we specialize the result to two representative cases: Gibbs posterior with a generic smooth loss function, and Gibbs posterior in Bayesian quantile regression where the check loss function is non-smooth.

Firstly, we make the following conditions on the population level risk function $\mathcal{R}(\theta) = \mathbb{E}[\ell(X, \theta)]$. Recall that $\theta^* = \arg\min_{\theta \in \Theta} \mathcal{R}(\theta)$ denotes the true parameter. The key idea is that although the sample level risk function (i.e. empirical risk function) $\mathcal{R}_n$ is allowed to be non-smooth, but as the sample size n grows, it becomes closer and closer to the population level risk function $\mathcal{R}(\theta)$, which can be properly analyzed if smooth.

Condition B.1 (Risk function): *For (n,d)-independent constants $(C', C, r) > 0$ and $(\gamma_0, \gamma_1, \gamma_2) \geq 0$:*

1. *$\mathcal{R}(\theta)$ is twice differentiable with mixed partial derivatives of order two being uniformly bounded by C on $B_r(\theta^*)$; for any $\theta \in \Theta$, $\mathcal{R}(\theta) - \mathcal{R}(\theta^*) \geq C'\, d^{-\gamma_0}\, (d^{-\gamma_1} \wedge \|\theta - \theta^*\|^2)$.*
2. *Let $\mathcal{H}_\theta$ denote the Hessian of $\mathcal{R}$ at θ. For any $\theta \in B_r(\theta^*)$, $\|\!|\mathcal{H}_\theta - \mathcal{H}_{\theta^*}\|\!|_{\mathrm{op}} \leq C\, d^{\gamma_2} \|\theta - \theta^*\|$.*

Condition B.1 imposes two requirements. Firstly, the population level risk function $\mathcal{R}(\cdot)$ must possess a unique global minimizer θ^*. This condition ensures that when the empirical risk $\mathcal{R}_n$ in the Gibbs posterior is substituted with $\mathcal{R}$, the resulting distribution $\pi^*(\theta) \propto \exp(-\alpha\, n\, \mathcal{R}(\theta))\pi(\theta)$ will be unimodal, thereby preventing the Markov chain from getting stuck in any local mode. Note that this condition is equivalent to the identifiability of the parameter in the model, and therefore is natural to assume. Secondly, the risk function should exhibit sufficient smoothness and local strong convexity in the vicinity of θ^*. This property enables a reliable Gaussian approximation for the local shape of $\pi^*(\theta)$ around θ^*, which is again a standard assumption and holds when the Fisher information matrix is not singular. Next, we introduce the following assumption of Lipschitz continuity for the loss function ℓ.

Condition B.2 (Loss function): *There exist (n,d)-independent constants $C > 0$ and $\gamma \geq 0$ such that for any $x \in \mathcal{X}$ and $(\theta,\, \theta') \in \Theta^2$, it holds that $|\ell(x,\theta) - \ell(x,\theta')| \leq C\, d^\gamma\, \|\theta - \theta'\|$.*

If the loss function has uniformly bounded derivatives with respect to θ, that is, $\left|\frac{\partial \ell(X,\theta)}{\partial \theta_j}\right| \leq C$ holds for any $j \in [d]$, $x \in \mathcal{X}$, and $\theta \in \Theta$, where C is a constant independent of n and d, then Condition B.2 holds with $\gamma = \frac{1}{2}$. Next, we introduce a function $g : \mathcal{X} \times \Theta \to \mathbb{R}^d$ that satisfies the following conditions.

Condition B.3 (Subgradient of loss function): *There exist some (n,d)-independent constants $(C, r, \beta_1) > 0$ and $(\gamma_3, \gamma_4) \geq 0$ so that:*

1. *For any $\theta \in B_r(\theta^*)$, it holds $\mathbb{E}[g(X,\theta)] = \nabla \mathcal{R}(\theta)$ and $\sup_{x \in \mathcal{X}} \|g(x,\theta)\| \leq C\, d^\gamma$, where γ is the same as that defined in Condition B.2.*
2. *Let $d_n^g(\theta,\theta') = \sqrt{n^{-1}\sum_{i=1}^n \|g(X_i,\theta) - g(X_i,\theta')\|^2}$ be a pseudo-metric in Θ.[6] The logarithm of the ε-covering number of $B_r(\theta^*)$ with respect to d_n^g is upper bounded by $C\, d \log(\frac{nd}{\varepsilon})$.*
3. *For any $v \in \mathbb{S}^{d-1}$ and $\theta,\, \theta' \in B_r(\theta^*)$, it holds that $\mathbb{E}\big[\big(v^T g(X,\theta) - v^T g(X,\theta')\big)^2\big] \leq C\, d^{\gamma_3}\, \|\theta - \theta'\|^{2\beta_1}$ and $\mathbb{E}\big[\big(\ell(X,\theta) - \ell(X,\theta') - g(X,\theta')(\theta - \theta')\big)^2\big] \leq C d^{\gamma_3}\, \|\theta - \theta'\|^{2+2\beta_1}$.*
4. *Let $\Delta_{\theta^*} = \mathbb{E}[\, g(X,\theta^*)\, g(X,\theta^*)^T]$ be the covariance matrix of the "score vector" $g(X,\theta^*)$. It holds that $\mathcal{H}_{\theta^*}^{-1} \Delta_{\theta^*} \mathcal{H}_{\theta^*}^{-1} \preceq C d^{\gamma_4}\, I_d$.*

[6] $d_n^g(\theta,\theta) = 0$ and d_n^g satisfies the symmetric property and triangle inequality, but can be zero for two distinct points.

Conditions B.3.1 relaxes the pointwise differentiability requirement for the loss function $\ell(x, \theta)$ with respect to θ. In fact, in many statistical applications, the expectation in the population-level risk function $\mathcal{R}(\theta) = \mathbb{E}[\ell(X, \theta)]$ has the smoothing effect of rendering $\mathcal{R}$ to be twice differentiable. For instance, we can choose $g(x, \cdot)$ as the gradient (or any subgradient) of $\ell(x, \cdot)$ for $x \in \mathcal{X}$ when ℓ is (or not) differentiable. Moreover, the boundedness assumption on the covering number in Condition B.3.2 allows us to uniformly control the random fluctuation of the empirical mean $\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta)$ away from the gradient of $\mathcal{R}(\theta)$. Condition B.3.3 can be interpreted as "smooth" assumptions on the loss function at the population level, quantified by β_1: by taking expectations with respect to the data X, the first term controls the Lipschitz constant of $g(X, \cdot)$, while the second term controls the remainder term of the first-order Taylor expansion of $\ell(X, \cdot)$, where the gradient is replaced with $g(X, \cdot)$. Condition B.3.4 assumes the boundedness of the operator norm of the matrix $\mathcal{H}_{\theta^*}^{-1}\Delta_{\theta^*}\mathcal{H}_{\theta^*}^{-1}$. This matrix represents the limiting covariance matrix for the sampling distribution of the empirical risk minimizer $\widehat{\theta}$, scaled by the sample size, i.e., $\sqrt{n}(\widehat{\theta} - \theta)$ converges in distribution to $N_d(0, \mathcal{H}_{\theta^*}^{-1}\Delta_{\theta^*}\mathcal{H}_{\theta^*}^{-1})$. This assumption allows us to provide an explicit bound on the deviance of $\widehat{\theta}$, which represents the asymptotic mean of the Gibbs posterior, from θ^*. It is important to highlight that Conditions B.1-B.3 can cover the common scenario where the loss function is continuously twice differentiable (see Corollary 7). Furthermore, these conditions also apply to more general cases with non-smooth loss functions, such as quantile regression (see Corollary 8).

Additionally, we assume the following smoothness condition for the prior distribution and compactness of the parameter space.

Condition B.4 (Prior and parameter space): *There exist positive (n, d)-independent constants (C, r) so that the parameter space Θ satisfies $B_r(\theta^*) \subset \Theta \subset [-C, C]^d$, and for any $\theta \in \Theta$, $\|\nabla(\log \pi)(\theta)\| \leq C\sqrt{d}$.*

The posterior density is defined to be zero for values of θ outside the parameter space Θ, ensuring that MALA rejects any proposed states that go beyond the boundaries of Θ. The assumption of compactness for the parameter space is primarily for technical convenience and is commonly made in Bayesian literature (Kleijn and van der Vaart, 2012; Yang and He, 2012). However, it is possible to relax this requirement by assuming the exponential tail behavior of the prior distribution, which will only incur extra logarithmic terms in the final result. Finally, we made the following conditions to the preconditioning matrix $\widetilde{I}$.

Condition C (Preconditioning matrix): *There exist some (n, d)-independent constants C so that the preconditioning matrix $\widetilde{I}$ satisfies that*

1. $|||\widetilde{I}^{-1}|||_{\mathrm{op}}|||\widetilde{I}|||_{\mathrm{op}} \leq C|||\mathcal{H}_{\theta^*}|||_{\mathrm{op}}|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}$;

2. $|||\widetilde{I}|||_{\mathrm{op}}|||(\widetilde{I}^{\frac{1}{2}}H_{\theta^*}\widetilde{I}^{\frac{1}{2}})^{-1}|||_{\mathrm{op}} \leq C\,|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}$.

The requirement for the preconditioning matrix $\widetilde{I}$ holds when $\widetilde{I}$ and its inverse has constant-order eigenvalues, such as the identity matrix that is conventionally used in MALA. On the other hand, it can also cover the case when $\widetilde{I}$ acts as a reasonable estimator to $\mathcal{H}_{\theta^*}^{-1}$ (i.e, $\widetilde{I}^{1/2}\mathcal{H}_{\theta^*}\widetilde{I}^{1/2}$ and its inverse has constant-order eigenvalues).

We now state the following theorem that provides a mixing time bound for sampling from a Gibbs posterior using MALA. Note that the (sub)gradient g is used for constructing the proposal in each step MALA.

Theorem 5 (Complexity of MALA for Bayesian sampling) *Consider sampling from the Bayesian Gibbs posteriors where $\mathcal{C}_n(\theta;\, X^{(n)}) = -n\,\alpha\,\mathcal{R}_n(\theta)$. Under Conditions B.1-B.4 and Condition C, consider positive numbers ρ_1, ρ_2, warming parameter M_0 and tolerance ε satisfying (1) $\rho_1 I_d \preceq \widetilde{I}^{1/2}\mathcal{H}_{\theta^*}\widetilde{I}^{1/2} \preceq \rho_2 I_d$; (2) $\log(\frac{M_0}{\varepsilon}) \leq C_1\,(d^{\gamma_5} + \log n)$ for (n,d)-independent constants C_1 and $\gamma_5 \geq 1$. There exists a constant κ_1 depends only on $(\beta_1, \gamma, \gamma_0, \gamma_1, \cdots, \gamma_5)$ so that if $d \leq c\frac{n^{\kappa_1}}{\log n}$ for a small enough constant c, then with probability at least $1 - n^{-1}$, the mixing time bound (7) in Theorem 3 holds for $\widetilde{\varepsilon}_0 = 1$ and*

$$h = c_0 \cdot \left[\rho_2\Big(d^{\frac{1}{3}} + d^{\frac{1}{4}}\big(\log\frac{M_0 d\kappa}{\varepsilon}\big)^{\frac{1}{4}} + \big(\log\frac{M_0 d\kappa}{\varepsilon}\big)^{\frac{1}{2}}\Big)\right]^{-1}, \text{ where } \kappa = \frac{\rho_2}{\rho_1},$$

where c_0 is an (n,d)-independent constant.

Remark 6 *Theorem 5 is proved by verifying Condition A for Bayesian Gibbs posteriors. The parameter κ_1 sets an upper bound on how the dimensionality of the parameter space d can grow in relation to the sample size n. A smaller κ_1 value implies that a larger dataset is necessary for the target posterior to be well-approximated by a Gaussian distribution. The expression for κ_1 is given by:*

$$\begin{aligned}\kappa_1 = &\frac{1}{1+2\gamma+6\gamma_0+4\gamma_2+\gamma_4} \wedge \frac{\beta_1}{1+\gamma_3+[(2\gamma_0)\vee((\gamma_5+\gamma_0)(1+\beta_1))]} \wedge \frac{1}{\gamma_0+\gamma_1+\gamma_5} \\ &\wedge \frac{1}{2\gamma+2\gamma_0+2\gamma_1+[2\vee(1+\gamma_4)]} \wedge \frac{1}{3\gamma_5+\gamma_0+[(2\gamma)\vee(\gamma_4+2\gamma_2+\gamma_0)\vee(2\gamma_2+2\gamma_0)]}.\end{aligned} \tag{8}$$

From the expression, κ_1 tends to be smaller if the loss function exhibits low smoothness, that means, β_1 is small. The classical proof of the Gaussian approximation of Bayesian posteriors with smooth likelihoods is based on the Taylor expansion of the likelihood function around $\widehat{\theta}$ (e.g. see Ghosh and Ramamoorthi, 2003). For the general non-smooth cases, we instead apply the Taylor expansion to the population level risk function $\mathcal{R}$ and use chaining and localization techniques in the empirical process theory to relate it to the sample version. Moreover, we keep track of the parameter dimension dependence, making Theorem 5 adaptable to more general cases under increasing dimension.

5.1 Gibbs posterior with smooth loss function

One representative example of Gibbs posterior satisfying Conditions B.1-B.4 is the one equipped with a smooth loss function. More specifically, we need Condition B.1 for the local convexity of the risk function, Condition B.4 for the smoothness of the prior and the following smoothness condition to the loss function.

Condition B.3' (Smoothness of loss function): *There exist some (n,d)-independent constants $C > 0$ and $(\gamma, \gamma_2, \gamma_3, \gamma_4) \geq 0$ so that (1) the loss function is twice differentiable so that for any $x \in \mathcal{X}$ and $\theta \in \Theta$, $\|\nabla_\theta \ell(x,\theta)\| \leq Cd^{\gamma}$; $|||\mathrm{Hess}_\theta(\ell(x,\theta))|||_{\mathrm{op}}^2 \leq Cd^{\gamma_3}$;*[7] *and*

[7] We use $\nabla_\theta \ell(x,\theta)$ and $\mathrm{Hess}_\theta(\ell(x,\theta))$ to denote the gradient and Hessian matrix of $\ell_x(\cdot) = \ell(x,\cdot)$ evaluated at θ, respectively.

for any $\theta, \theta' \in \Theta$, $|||\mathrm{Hess}_\theta(\ell(x,\theta)) - \mathrm{Hess}_\theta(\ell(x,\theta'))|||_{\mathrm{op}} \leq C\, d^{\gamma_2}\|\theta-\theta'\|$*; (2) let* $\Delta_{\theta^*} = \mathbb{E}[\nabla_\theta \ell(X,\theta^*)\nabla_\theta \ell(X,\theta^*)^T]$*, then* $\mathcal{H}_{\theta^*}^{-1}\Delta_{\theta^*}\mathcal{H}_{\theta^*}^{-1} \preceq C\, d^{\gamma_4} I_d$*.*

Corollary 7 (Sampling from smooth posteriors) *Consider the Bayesian Gibbs posterior with loss function* ℓ*. Suppose (1) Conditions B.1, B.3' and B.4 hold; (2) the warming parameter* M_0 *and tolerance* ε *satisfying* $\log(\frac{M_0}{\varepsilon}) \leq C_1\,(d^{\gamma_5} + \log n)$ *for* (n,d)*-independent constants* C_1 *and* $\gamma_5 \geq 1$*; (3)* $d \leq c\frac{n^{\kappa_1}}{\log n}$ *for a small enough constant* c*, where* κ_1 *is defined in* (8) *with* $\beta_1 = 1$*. Then there exists an* (n,d)*-independent constant* c_0 *so that it holds with probability at least* $1-n^{-1}$ *that*

1. *consider the identity preconditioning matrix* $\widetilde{I} = I_d$*. the mixing time upper bound* (7) *holds for any* $\rho_1 \leq \rho_2$ *so that* $\rho_1 I_d \preceq \mathcal{H}_{\theta^*} \preceq \rho_2 I_d$*,* $\log(\frac{\rho_1}{\rho_2}) \leq C_1 d^{\gamma_5}$ *and*

$$h = c_0 \cdot \Big[\rho_2 \cdot \Big(d^{\frac{1}{3}} + d^{\frac{1}{4}}\big(\log\frac{M_0 d}{\varepsilon}\big)^{\frac{1}{4}} + \big(\log\frac{M_0 d}{\varepsilon}\big)^{\frac{1}{2}}\Big)\Big]^{-1};$$

2. *consider the inverse empirical Hessian matrix* $\widetilde{I} = \big(|S|^{-1}\sum_{i\in S}\mathrm{Hess}_\theta(\ell(X_i,\widehat{\theta}))\big)^{-1}$*, where* $S \subset \{1,2,\cdots,n\}$ *with* $|S| \geq C_2\, d^{\gamma_3+2\gamma_0+7/3}$ *for a large enough* (n,d)*-independent constant* C_2*, then the mixing time upper bound* (7) *holds with* $\rho_1 = \frac{1}{2}$ *and*

$$h = c_0 \cdot \Big[\Big(d^{\frac{1}{3}} + d^{\frac{1}{4}}\big(\log\frac{M_0 d}{\varepsilon}\big)^{\frac{1}{4}} + \big(\log\frac{M_0 d}{\varepsilon}\big)^{\frac{1}{2}}\Big)\Big]^{-1};$$

moreover, let $\mu_0 = N_d(\widehat{\theta}, n^{-1}\widetilde{I})\big|_{\{\theta:\sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})\|\leq 3c_1\sqrt{d}\}}$*, where* c_1 *is a constant so that* $c_1 \geq 3 \vee \sup\limits_{i\in[d],j\in[d]} \frac{\partial^2\mathcal{R}(\theta^*)}{\partial\theta_i\partial\theta_j}$*, then* μ_0 *is* M_0*-warm with respect to* π_n *with* $\log M_0 \leq d^{\frac{1}{3}}$*.*

When the Hessian matrix $\mathcal{H}_{\theta^*}$ is ill-conditioned, introducing the preconditioning matrix $\widetilde{I} = \big(|S|^{-1}\sum_{i\in S}\mathrm{Hess}_\theta(\ell(X_i,\widehat{\theta}))\big)^{-1}$ may lead to a faster mixing. Furthermore, if the tolerance satisfying $\log(\frac{1}{\varepsilon}) = \mathcal{O}(d^{\frac{1}{3}})$, then the second statement of Corollary 7 can lead to an optimal mixing time bound $\mathcal{O}\big(d^{\frac{1}{3}}\log(\frac{1}{\varepsilon})\big)$.

5.2 Bayesian quantile regression

We consider Bayesian quantile regression as a representative example where the loss function is non-smooth. Specifically, in quantile regression (Koenker and Bassett, 1978), for a fixed $\tau \in (0,1)$, the τ^{th} quantile $q_\tau(Y|\widetilde{X})$ of the response $Y \in \mathbb{R}$ given the covariates $\widetilde{X} \in \mathbb{R}^d$ is modelled as $q_\tau(Y|\widetilde{X}) = \widetilde{X}^T\theta^*$. Here we consider the homogeneous case where the error $e = Y - \widetilde{X}^T\theta^*$ is independent of the covariates $\widetilde{X}$. Given a set of n i.i.d. samples $X^{(n)} = \{X_i = (\widetilde{X}_i, Y_i)\}_{i\in[n]}$, the quantile regression solves the following convex optimization problem:

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\sum_{i=1}^n \Big[(Y_i - \widetilde{X}_i^T\theta)\cdot\big(\tau - \mathbf{1}(Y_i < \widetilde{X}_i^T\theta)\big)\Big],$$

where the loss function $\ell_q\big((\widetilde{X},Y),\theta\big) = (Y - \widetilde{X}^T\theta)\cdot\big(\tau - \mathbf{1}(Y < \widetilde{X}^T\theta)\big)$ is referred to as the check loss. The minimization of the check loss function is equivalent to the maximization of a likelihood function formed by combining independently distributed asymmetric Laplace densities (Yu and Moyeed, 2001). The posterior for Bayesian quantile regression can thus be formed by assuming a (possibly misspecified) asymmetric Laplace distribution (ALD) for the response, which is

$$\pi_n(\theta) \propto \exp\big(-n\,\mathcal{R}_n(\theta)\big)\,\pi(\theta), \quad \theta \in \mathbb{R}^d,$$

with $\pi(\theta)$ being a prior on Θ and $\mathcal{R}_n(\theta) = n^{-1}\sum_{i=1}^n \ell_q(X_i,\theta)$ being the empirical risk function. Furthermore, by adding a multiplier $\alpha > 0$ to the likelihood, we can obtain the Gibbs (or tempered) posterior.

Since the loss function $\ell_q(X,\theta)$ for quantile regression is not differentiable when $Y = \widetilde{X}^T\theta$, in order to sampling from the Gibbs posterior associated with Bayesian quantile regression using the (preconditioned) MALA, we need to consider the subgradient of ℓ_q with respect to θ, given by

$$g(X,\theta) = \big(\mathbf{1}(Y < \widetilde{X}^T\theta) - \tau\big)\,\widetilde{X}, \quad X = (\widetilde{X}, Y), \quad \theta \in \mathbb{R}^d.$$

The following corollary quantifies the computational complexity for sampling from π_n using MALA. We first state the required conditions.

Condition D.1: *There exist (n,d)-independent constants $(C,C') > 0$ and $(\alpha_0,\alpha_1) \geq 0$ such that (1) the support $\mathcal{X}$ of the covariates $\widetilde{X}$ is included in $[-C,C]^d$; (2) for any $v \in \mathbb{S}^{d-1}$, $\mathbb{E}|\widetilde{X}^T v|^2 \geq C'd^{-\alpha_0}$ and $\mathbb{E}|\widetilde{X}^T v|^3 \leq Cd^{\alpha_1}$.*

Condition D.2: *Let $f_e(\cdot)$ denote the probability density function of the homogeneous error $e = Y - \widetilde{X}^T\theta^*$, then there exist (n,d)-independent constants $(C,C') > 0$ such that (1) $\int_{-\infty}^0 f_e(z)dz = \tau$; (2) $f_e(0) > C'$ and $\sup_{e\in\mathbb{R}^d} f_e(e) \leq C$; (3) for any $e_1, e_2 \in \mathbb{R}$, $|f_e(e_1) - f_e(e_2)| \leq C|e_1 - e_2|$.*

Condition D.1 assumes the compactness of the covariate space and the positive definiteness of the gram matrix $\mathbb{E}[\widetilde{X}\widetilde{X}^T]$. Condition D.2 introduces several regularity conditions on the distribution of the error $e = Y - \widetilde{X}^T\theta^*$: (1) The error term e is independent of the covariates. (2) The model is correctly specified, meaning that $\widetilde{X}^T\theta^*$ corresponds to the τ-th quantile of the response variable Y given $\widetilde{X}$. (3) The density function $f_e(\cdot)$ of the error term is positive at the origin and Lipschitz continuous. Under the assumption of homogeneous errors, the limiting covariance matrix of the posterior distribution of interest is given by $n^{-1}(f_e(0)\cdot\mathbb{E}[\widetilde{X}\widetilde{X}^T])^{-1}$. In this case, a natural choice for the preconditioning matrix is the inverse of the empirical Gram matrix, denoted as $\widetilde{I} = \big(|S|^{-1}\sum_{i\in S}\widetilde{X}_i\widetilde{X}_i^T\big)^{-1}$ where $S \subset \{1,2,\cdots,n\}$. It is worth noting that similar analyses can be carried out for the case of heterogeneous errors, but the limiting covariance matrix will be more complex.

Corollary 8 (Sampling from non-smooth posteriors) *Suppose Conditions D.1, D.2, and B.4 are satisfied, and the warming parameter M_0 and tolerance ε satisfying $\log(\frac{M_0}{\varepsilon}) \leq C_1\,(d^{\alpha_2} + \log n)$ for (n,d)-independent constants C_1 and $\alpha_2 \geq 1$. Assume $d \leq c(\frac{n^{\widetilde{\alpha}}}{\log n})$ with $\widetilde{\alpha} = \frac{1}{2+4\alpha_1+7\alpha_0} \wedge \frac{1}{2+3\alpha_0+2\alpha_1+3\alpha_2}$ and a small enough constant c, and let the inverse empirical Gram matrix $\widetilde{I} = \big(|S|^{-1}\sum_{i\in S}\widetilde{X}_i\widetilde{X}_i^T\big)^{-1}$ be the preconditioning matrix, where $S \subset \{1,2,\cdots,n\}$ with $|S| \geq C_2\,d^{\alpha_1+2\alpha_0+3/2}\log n$ for a large enough (n,d)-independent* *constant C_2, then it holds with probability larger than $1-\frac{1}{n}$ that that the mixing time upper bound (7) is true with $\rho_1 = \frac{1}{2} f_e(0)$ and*

$$h = c_0 \cdot \left[f_e(0) \cdot \Big(d^{\frac{1}{3}} + d^{\frac{1}{4}} \big(\log \frac{M_0 d}{\varepsilon}\big)^{\frac{1}{4}} + \big(\log \frac{M_0 d}{\varepsilon}\big)^{\frac{1}{2}} \Big) \right]^{-1}$$

with c_0 being an (n,d)-independent constant.

Corollary 8 illustrates the implications of applying our theory to non-smooth posteriors. A key observation is that in the large-sample regime, although the potential function associated with the Bayesian posterior may be non-smooth, its population-level counterpart is smooth (as per Condition B.3). This allows MALA, using sub-gradients, to effectively sample from non-smooth posteriors. Moreover, while our theory is applicable to non-smooth posteriors, the smoothness of the posterior density function influences its convergence to a Gaussian limit as n grows, as captured by the parameter β_1 in Condition B.3. A posterior with higher smoothness (or larger β_1) will converge more rapidly to a Gaussian distribution, as demonstrated in Lemma 18, which in turn leads to an improved (higher) acceptance rate of MALA. For example, in Bayesian quantile regression, the smoothness parameter β_1 is at most $\frac{1}{2}$. In contrast, for posterior densities with smooth loss functions, β_1 can be taken as 1. Interestingly, our theoretical result also leads a practical guideline: when applying MALA to sample from a less smooth Bayesian posterior densities, a relatively larger sample size n is needed to maintain the sampling efficiency. Otherwise, if n is not sufficiently large relative to the dimension, the non-smoothness of the Bayesian posterior can result in a lower acceptance rate and slower mixing times for MALA; see our simulation results in Section 7 for some empirical evidence.

6. Proof Sketch of Theorem 3

In this section, we provide a sketched proof about how to utilize the general machinery of s-conductance profile developed in Section 3 to analyze the mixing time of MALA under Condition A. We consider the identity preconditioning matrix (i.e. $\widetilde{I} = I_d$) in this sketch for simplicity, and the case for general preconditioning matrix can be proved by considering the transformation $G(\theta) = \sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta - \widehat{\theta})$, see Appendix B.1 for further details.

Let $T_x^\zeta(\mathrm{d}y) = T^\zeta(x, \mathrm{d}y)$ denote the Markov transition kernel of the ζ-lazy version of MALA for sampling from π_{loc} as described in Section 4 with rescaled step size h. To apply Lemma 2, we first need to establish a log-isoperimetric inequality, which is a property of π_{loc} alone and is not specific to MALA. This step can be done by adapting existing proofs of a log-isoperimetric inequality for Gaussians (e.g. Lemma 16 of Chen et al. (2020)) to π_{loc} via a perturbation analysis (see Lemma 14 and its proof in the appendix for details). Second, we need to apply an overlap argument for bounding the total variation distance between $T_x^\zeta(\cdot)$ and $T_z^\zeta(\cdot)$ for x and z satisfying $\|x-z\| \leq C\sqrt{h}$ and belonging to a high probability set E under π_{loc}. This step utilizes the structure and properties of MALA algorithm, and we briefly sketch its proof below (details can be found in Lemma 15 in the appendix) and discuss its difference from existing proofs.

We construct a high probability set as $E = \{\xi \in B_{R/2}^d : \big|\xi^T \widetilde{J}^3 \xi - \mathrm{tr}(\widetilde{J}^2)\big| \leq r_d\} \cap \{\xi \in B_{R/2}^d : \big|\xi^T \widetilde{J}^2 \xi - \mathrm{tr}(\widetilde{J})\big| \leq r_d/\rho_2\}$, where the value of r_d makes $\pi_{\mathrm{loc}}(E) \geq 1 - 2\frac{h\rho_1\varepsilon^2}{M_0^2}$ based on the last property of Condition A (details can be found in Lemma 21). Recall the acceptance probability $A(x,y) = 1 \wedge \frac{\pi_{\mathrm{loc}}(y)\,Q(y,x)}{\pi_{\mathrm{loc}}(x)\,Q(x,y)}$ and denotes $\overline{A}(x,y) = 1 \wedge \frac{\overline{\pi}(y)\,Q(y,x)}{\overline{\pi}(x)\,Q(x,y)}$ with $\overline{\pi}$ being the density of the Gaussian $N_d(0, J^{-1})$. By comparing π_{loc} and $\overline{\pi}$ using Condition A, we can get the following inequality:

$$\begin{aligned}
&\|T_x^\zeta - T_z^\zeta\|_{TV} \leq 1 - (1-\zeta)\int_{B_R^d} \min\Big(A(x,y)Q(x,y), A(z,y)Q(z,y)\Big)\,\mathrm{d}y \\
&\leq 1 - \frac{1}{2}(1-\zeta)\exp(-2\widetilde{\varepsilon}_0)\cdot\bigg(\int_{B_R^d}\overline{A}(x,y)Q(x,y)\,\mathrm{d}y + \int_{B_R^d}\overline{A}(z,y)Q(z,y)\,\mathrm{d}y \\
&\qquad - \int_{B_R^d}\big|\overline{A}(x,y)Q(x,y) - \overline{A}(z,y)Q(z,y)\big|\,\mathrm{d}y\bigg) \\
&\leq 1 - (1-\zeta)\exp(-2\widetilde{\varepsilon}_0)\cdot\Big(1 - \int_{B_R^d} Q(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y - \int_{B_R^d} Q(z,y)(1-\overline{A}(z,y))\,\mathrm{d}y \\
&\qquad - \|Q_x - Q_z\|_{\mathrm{TV}} - \frac{1}{2}\int_{(B_R^d)^c} Q(x,y)\,\mathrm{d}y - \frac{1}{2}\int_{(B_R^d)^c} Q(z,y)\,\mathrm{d}y\Big).
\end{aligned} \tag{9}$$

We will separately bound the terms on the right hand side of (9) as follows. The last term $\frac{1}{2}\int_{(B_R^d)^c} Q(x,y)\,\mathrm{d}y + \frac{1}{2}\int_{(B_R^d)^c} Q(z,y)\,\mathrm{d}y$ can be upper bounded by $\frac{1}{6}$ using the condition of R in Condition A. For the remaining terms, let Q_x denote the probability measure with density function $Q(x,\cdot)$, now we use Condition A by comparing Q_x with the proposal distribution

$$Q_x^\Delta := N_d(x - hJx,\, 2hI_d)$$

of MALA for sampling from the Gaussian $N_d(0, J^{-1})$, leading to

$$\begin{aligned}
\int_{B_R^d} Q(x,y)\left(1 - \overline{A}(x,y)\right)\mathrm{d}y \leq 2\,\|Q_x - Q_x^\Delta\|_{\mathrm{TV}} &+ \int_{\mathbb{R}^d}\Big|Q^\Delta(x,y) - \frac{\overline{\pi}(y)Q^\Delta(y,x)}{\overline{\pi}(x)}\Big|\,\mathrm{d}y \\
&+ \int_{B_R^d}\Big|\frac{\overline{\pi}(y)Q^\Delta(y,x)}{\overline{\pi}(x)} - \frac{\overline{\pi}(y)Q(y,x)}{\overline{\pi}(x)(x)}\Big|\,\mathrm{d}y,
\end{aligned} \tag{10}$$

where we use $Q^\Delta(x,\cdot)$ to denote the density function of Q_x^Δ. It then can be proved using Condition A and Pinsker's inequality after some careful calculations (see Lemmas 22 and 23 in the appendix) that

$$\int_{B_R^d} Q(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y + \int_{B_R^d} Q(z,y)(1-\overline{A}(z,y))\,\mathrm{d}y + \|Q_x - Q_z\|_{\mathrm{TV}} \leq 1/3.$$

Our proof of Lemma 22 for bounding $\int_{\mathbb{R}^d}\left|Q^\Delta(x,y) - \overline{\pi}(y)Q^\Delta(y,x)/\overline{\pi}(x)\right|\mathrm{d}y$ is technically similar to that of Proposition 38 in Chewi et al. (2021) for bounding the mixing time of MALA with a standard Gaussian target (i.e. $\overline{\pi} = N_d(0, I_d)$). The non-trivial part in our analysis lies in keeping track of the dependence on the maximal and minimal eigenvalues of J. Finally, we can obtain

$$\|T_x^\zeta - T_z^\zeta\|_{TV} \leq 1 - \frac{1-\zeta}{2}\exp(-2\widetilde{\varepsilon}_0).$$

With the lower bound on $\pi_{\mathrm{loc}}(E)$ and the upper bound on $\|T_x^\zeta - T_z^\zeta\|_{TV}$, we are then able to apply the s-conductance profile argument to control the mixing time.

Remark 9 *It is worth mentioning that the analysis in Chen et al. (2020) requires the high probability set, which is set E in our case, to be convex. This requirement will deteriorate the d dependence of the mixing time bound since $\|T_x^\zeta - T_z^\zeta\|_{\mathrm{TV}}$ for $x, z \in E$ can no longer be controlled under a large step size h as ours. This motivates us to introduce the more flexible notion of* s-conductance profile *that extends the commonly used conductance profile (Goel et al., 2006; Chen et al., 2020) and s-conductance (Lovász and Simonovits, 1993). Analysis based on the s-conductance profile leads to a better warming parameter dependence than that obtained in Chewi et al. (2021); Belloni and Chernozhukov (2009) without affecting our obtained dimension dependence (based on s-conductance). A complete proof of this theorem is included in Appendix B.1. Similar analysis can also be carried over for analyzing general smooth and strictly log-concave densities to improve the warming parameter dependence (e.g. Chewi et al., 2021; Belloni and Chernozhukov, 2009).*

7. Numerical Study

In this section, we conduct an empirical study to explore how the performance of MALA varies across different dimensions and sample sizes when targeting different Bayesian posteriors.

7.1 Set up

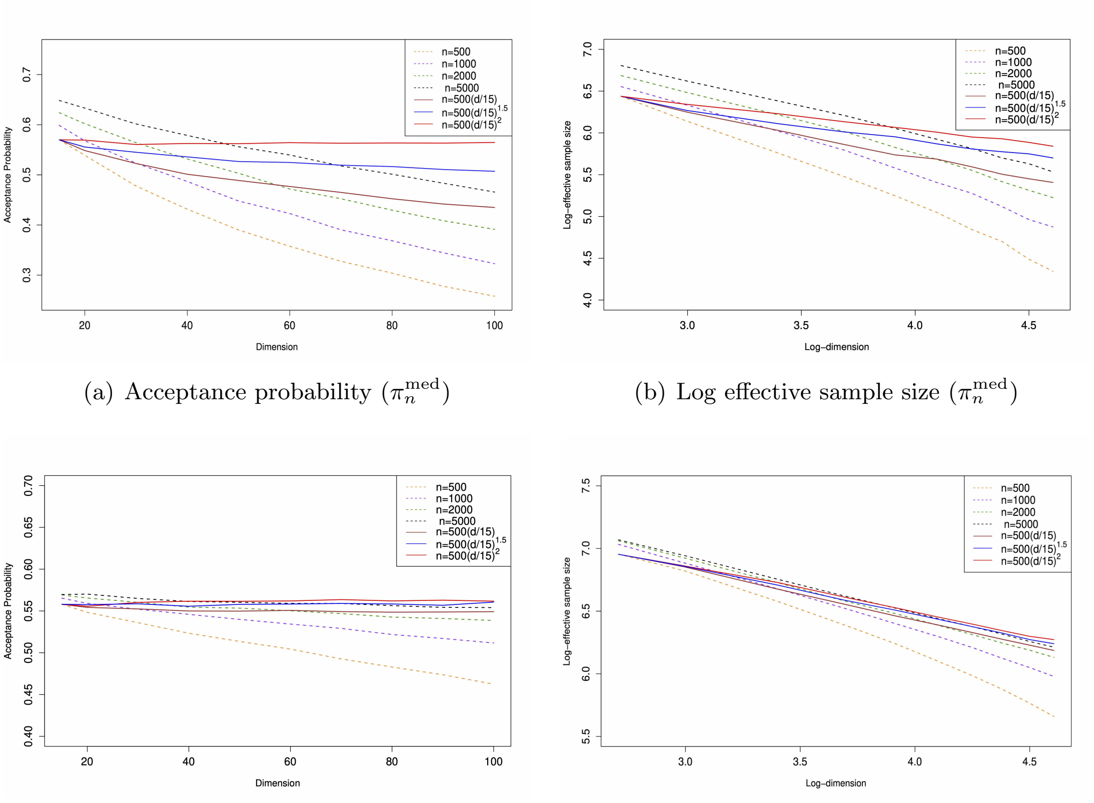
We carry out the experiment using two examples: Bayesian linear regression and Bayesian median regression. For Bayesian linear regression, the corresponding Bayesian posterior is given by:

$$\pi_n^{\mathrm{mean}}(\theta \,|\, X^{(n)}) \propto \exp\Big(-\frac{1}{2}\sum_{i=1}^n \big\|Y_i - \widetilde{X}_i^T\theta\big\|^2\Big)\,\pi(\theta), \quad \theta \in \mathbb{R}^d.$$

For Bayesian median regression, the Bayesian posterior is given by

$$\pi_n^{\mathrm{med}}(\theta \,|\, X^{(n)}) \propto \exp\Big(-\frac{1}{2}\sum_{i=1}^n \big|Y_i - \widetilde{X}_i^T\theta\big|\Big)\,\pi(\theta), \quad \theta \in \mathbb{R}^d.$$

We choose the parameter dimension d from the set $\{15, 20, 30, 40, \cdots, 100\}$ and sample size n from $\{500, 1000, 2000, 5000, 500(d/15), 500(d/15)^{3/2}, 500(d/15)^2\}$. The covariates $\widetilde{X}$ are generated from a multivariate Gaussian distribution with zero mean and identity covariance matrix. For Bayesian linear regression, we generate a random error variable e follows a standard normal distribution, and for Bayesian median regression, e follows a Laplace distribution with location parameter $\mu = 0$ and scale parameter $b = 2$. The response variable Y is given by $Y = \widetilde{X}^T\theta^* + e$ with $\theta^* = (1, 1, \cdots, 1)$. We consider the parameter space $\Theta = [-100, 100]^d$ and the prior is chosen to be a uniform distribution over Θ. We then use MALA to sample from the Bayesian posterior π_n^{mean} and π_n^{med}.



(a) Acceptance probability (π_n^{med})

(b) Log effective sample size (π_n^{med})

(c) Acceptance probability (π_n^{mean})

(d) Log effective sample size (π_n^{mean})

Figure 1: Plots (a) and (c) report the average acceptance probabilities of MALA when sampling from the posterior in Bayesian quantile regression (denoted as π_n^{med}) and Bayesian linear regression (denoted as π_n^{mean}) respectively, across various sample sizes (n) and dimensions (d), with the step size $\widetilde{h} = cd^{-\frac{1}{3}}n^{-1}$. Plots (b) and (d) present the relationship between the logarithm of the effective sample size and the logarithm of the dimension for sampling from π_n^{med} and π_n^{mean} respectively. As we can see, when the sample size increases with the dimension at a rate of d^2, by choosing steps sizes with scaling $d^{-\frac{1}{3}}n^{-1}$, the acceptance probabilities roughly remain constant and the change in the logarithmic effective sample sizes exhibit slopes close to $-\frac{1}{3}$ for both examples of Bayesian linear regression and median regression. On the other hand, when n remains a constant, in both cases, the acceptance probabilities will decrease as d becomes larger, and the changes in the logarithmic effective sample size exhibit slopes smaller than $-\frac{1}{3}$. However, compared to π_n^{med}, the decreases in acceptance probability and effective sample size are much slower for sampling from π_n^{mean}. In particular, when n increases with d at a linear rate, the acceptance probabilities for π_n^{mean} roughly remain constant, while there is obvious decrease in the acceptance probabilities for π_n^{med}.

7.2 Results

In general, estimating the mixing time of a Markov chain is a challenging task. Instead, we utilize the effective sample size (Gelman et al., 1995) as a metric to assess the mixing of MALA. The effective sample size of N Markov samples, denoted as N_{eff}, quantifies the amount of information lost due to correlations in the chain, and plays a role similar to the number of independent draws in the standard central limit theorem (Brooks et al., 2011). The effective sample size of a sequence is formally defined in terms of the autocorrelations within the sequence at different lags, i.e., $N_{\text{eff}} = \frac{N}{1+2\sum_{t=1}^{\infty}\rho_t}$, with ρ_t being the autocorrelation at lag t. Details of the estimation of N_{eff} can be found in Section 11.5 of Gelman et al. (2013). It is worth noting that, theoretically, the ratio $\frac{N}{N_{\text{eff}}}$ can be controlled by the inverse of the spectral gap (Kloeckner, 2019), which governs the convergence of the Markov chain.

Taking into account our theoretical findings regarding the convergence of MALA with an appropriate warm start, we compute the effective sample sizes after a burn-in period of 1000 iterations, totaling 5000 iterations. We choose the step size $\widetilde{h} = c_1 d^{-\frac{1}{3}} n^{-1}$, where $c_1 = 4.28$ for Bayesian median regression and $c_1 = 1.39$ for Bayesian linear regression. These choices of c_1 ensure that the overall acceptance probability in each example closely approximates 0.574 as suggested by Roberts and Rosenthal (1998). The preconditioning matrix $\widetilde{I}$ is chosen to be the identity matrix.

Figure 1 present the trends of the average acceptance probability and the logarithm of the effective sample size when sampling from π_n^{med} and π_n^{mean}, considering varying sample sizes and dimensions. When n remains unchanged for varying d, we observe a decrease in the acceptance probability as d grows larger in both cases. Additionally, the trends of the logarithmic effective sample size exhibit slopes smaller than $-\frac{1}{3}$. The reason for this phenomenon is that, the deviance $\widetilde{\varepsilon}_0$ of the target posterior from the Gaussian distribution, stated in Theorem 3, will increase with d when the sample size remains unchanged. Consequently, when d is sufficiently large, the mixing time will deviate significantly from $\mathcal{O}(d^{\frac{1}{3}})$ and the acceptance probability will decrease rapidly when employing a step size of order $d^{-\frac{1}{3}} n^{-1}$. Another interesting observation is that the decreases in acceptance probability and effective sample size are much slower when sampling from π_n^{mean} compared to sampling from π_n^{med}. One factor results in this phenomenon can be the smoothness of the loss function used in π_n^{mean}, which aids the convergence of the Gibbs posterior to the Gaussian distribution. Specifically, Lemma 18 in Appendix B.3 demonstrates that a Gibbs posterior with a smooth loss function will converge to a Gaussian distribution with a rate of $\mathcal{O}(n^{-1/2})$ for a fixed d, while the Gibbs posterior used in Bayesian quantile regression approaches a Gaussian distribution at a rate of $\mathcal{O}(n^{-1/4})$. Therefore, under the same n and d, the approximation error $\widetilde{\varepsilon}_0$ for π_n^{mean} is much smaller than π_n^{med}. Additionally, we can see from Figure 1 that, for achieving a constant acceptance probability and effective sample size at an order of $d^{-\frac{1}{3}}$ when d ranges from 15 to 100, the condition $d = \mathcal{O}(\sqrt{n})$ is required for sampling from π_n^{med}, while the condition $d = \mathcal{O}(n)$ suffices for sampling from π_n^{mean}.

8. Conclusion and Discussion

In this paper, we studied the sampling complexity of Bayesian (pseudo-)posteriors using MALA under large sample size, covering cases where the posterior density is non-smooth and/or non-log-concave. A variant of MALA that includes a preconditioning matrix was also considered. While our analysis for the preconditioned MALA suggests an adaptive MALA with a data-driven preconditioning matrix may be preferable, its rigorous theoretical analysis may leave as our future work. When applying our main result to Bayesian inference, we mainly considered the Gibbs posterior, while similar analysis may carry over to other types of Bayesian pseudo-posterior, such as Bayesian empirical likelihood (Lazar, 2003), and we leave this for future research. Another challenge lies in constructing a suitable warm start that satisfies $\log M_0 \leq d^{\frac{1}{3}}$. Obtaining a warm start efficiently for general non-log-concave sampling can be challenging. However, the asymptotic Gaussian nature of the Bayesian posterior may aid in the construction of such a warm start, and it is possible to develop specific algorithms tailored to particular problems that leverage the Gaussian asymptotics. For instance, in Bayesian quantile regression, one can determine the point estimator $\hat{\theta}$ using linear programming and utilize the Gaussian asymptotic properties of the posterior to construct initializations. A more detailed exploration of this topic is left for future research.

References

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL `http://jmlr.org/papers/v17/15-290.html`.

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv preprint arXiv:2302.10249*, 2023.

Filippo Ascolani and Giacomo Zanella. Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models, 2023.

Yves F Atchadé. An adaptive version for the Metropolis Adjusted Langevin Algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8(2):235–254, 2006. URL `https://doi.org/10.1007/s11009-006-8550-0`.

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009.

Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265. PMLR, 2015.

Indrabati Bhattacharya and Ryan Martin. Gibbs posterior inference on multivariate quantiles. *arXiv preprint arXiv:2002.01052*, 2020.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Jeff Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in analysis*, pages 195–200. Princeton University Press, 2015.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–72, 2020. URL `http://jmlr.org/papers/v21/19-441.html`.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 15–19 Aug 2021. URL `https://proceedings.mlr.press/v134/chewi21a.html`.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR, 07–10 Jul 2017. URL `https://proceedings.mlr.press/v65/dalalyan17a.html`.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183): 1–42, 2019. URL `http://jmlr.org/papers/v20/19-306.html`.

James M Flegal, Murali Haran, and Galin L Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260, 2008.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL `https://books.google.com.hk/books?id=ZXL6AQAAQBAJ`.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1): 110–120, 1997.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, pages 721–741, 1984.

Abhik Ghosh, Tuhin Majumder, and Ayanendranath Basu. General robust bayes pseudo-posterior: Exponential convergence results with applications, 2020.

J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer New York, New York, NY, 2003. URL `https://link.springer.com/book/10.1007/b97842`.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. doi: https://doi.org/10.1111/j.1467-9868.2010.00765.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x`.

Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11(none):1 – 26, 2006. doi: 10.1214/EJP.v11-300. URL `https://doi.org/10.1214/EJP.v11-300`.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

Ravi Kannan, László Lovász, and Ravi Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability and Computing*, 15(4):541–570, 2006.

B.J.K. Kleijn and A.W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none):354 – 381, 2012. doi: 10.1214/12-EJS675. URL `https://doi.org/10.1214/12-EJS675`.

Benoît Kloeckner. Effective berry–esseen and concentration bounds for Markov chains with a spectral gap. *The Annals of Applied Probability*, 29(3):1778–1807, 2019.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913643`.

M. R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer New York, New York, NY, 2008.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL `https://doi.org/10.1214/aos/1015957395`.

Nicole A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90(2):319–326, 06 2003. ISSN 0006-3444. doi: 10.1093/biomet/90.2.319. URL `https://doi.org/10.1093/biomet/90.2.319`.

László Lovász and Ravi Kannan. Faster mixing via average conductance. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 282–287, 1999.

L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993. doi: https://doi.org/10.1002/rsa.3240040402. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.3240040402`.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116 (42):20881–20885, 2019.

Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 2259–2293. PMLR, 2019.

Charles C Margossian. A review of automatic differentiation and its efficient implementation. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1305, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 2004.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Vivekananda Roy. Convergence diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020.

Nicholas Syring and Ryan Martin. Gibbs posterior concentration rates under sub-exponential type losses. *arXiv preprint arXiv:2012.04505*, 2020.

Cornelia Vacar, Jean-Françis Giovannelli, and Yannick Berthoumieu. Langevin and hessian with fisher approximation stochastic sampling for parameter estimation of structured covariance. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3964–3967, 2011. doi: 10.1109/ICASSP.2011.5947220.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270):1–63, 2022.

Yunwen Yang and Xuming He. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131, 2012. ISSN 00905364, 21688966. URL `http://www.jstor.org/stable/41713667`.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(01)00124-9. URL `https://www.sciencedirect.com/science/article/pii/S0167715201001249`.

Appendix

We summarize some necessary notation and definitions in the appendix. We use $\mathbf{1}_A$ to denote the indicator function of a set A so that $\mathbf{1}_A(x) = 1$ if $x \in A$ and zero otherwise. For two sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to mean $a_n \leq Cb_n$ and $a_n \geq Cb_n$, respectively, for some constant $C > 0$ independent of n, d. In addition, $a_n \asymp b_n$ means that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold, and $a_n = \mathcal{O}(b_n)$ if $a_n \lesssim b_n$; $a_n = \Theta(b_n)$ if $a_n \asymp b_n$. We use $\mathbf{N}(\mathcal{F}, d_n, \varepsilon)$ to denote the ε-covering number of $\mathcal{F}$ with respect to pseudo-metric d_n. Throughout, C, c, C_0, c_0, C_1, c_1, ... are generically used to denote positive constants independent of n, d whose values might change from one line to another. We denote $\mathcal{L}^2(\pi)$ to be the space of square integrable functions under measure π. For a transition kernel $T : \Theta \times \mathcal{B}(\Theta) \to \mathbb{R}$ of a reversible Markov chain with invariant distribution π, where $\mathcal{B}(\Theta)$ is the Borel-sigma algebra on Θ, the Dirichlet form $\mathcal{E} : \mathcal{L}_2(\pi) \times \mathcal{L}_2(\pi) \to \mathbb{R}$ associated with the transition kernel T is given by $\mathcal{E}(g, h) = \frac{1}{2}\int_{x,y\in\Theta^2}(g(x) - h(y))^2 T(x, \mathrm{d}y)\pi(\mathrm{d}x)$.

Table of Contents

Appendix A. Additional Results

A.1 Additional Simulation

In this section, we carry out experiment using Bayesian linear regression with the following posterior

$$\pi_n^{\mathrm{mean}}(\theta|X^{(n)}) \propto \exp\Big(-\frac{1}{2}\sum_{i=1}^n \|Y_i - \widetilde{X}_i^T\theta\|^2\Big)\pi(\theta), \quad \theta \in \mathbb{R}^d.$$

for exploring the impact of the preconditioning matrix and initial distribution in MALA. We set the sample size $n = 2000$ and choose the parameter dimension d from set $\{10, 15, 20, 30, 50\}$. The covariates $\widetilde{X}$ are generated from a multivariate Gaussian distribution with zero mean and the covariance matrix Σ given by a diagonal matrix with elements

$$\Big(\underbrace{\sqrt{d}, \sqrt{d}, \cdots, \sqrt{d}}_{\left[\frac{d}{2}\right]}, \underbrace{\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \cdots, \frac{1}{\sqrt{d}}}_{d-\left[\frac{d}{2}\right]}\Big).$$

We consider two choices for the preconditioning matrix: one is the inverse (mini-batch) empirical gram matrix $\widehat{\Sigma}_m = (m^{-1}\sum_{j=1}^m \widetilde{X}_j\widetilde{X}_j^T)^{-1}$, which is an estimator to the covariance matrix $n^{-1}\Sigma^{-1}$ of the posterior rescaled by n. Here, we consider values of $m = \{200, 500, 2000\}$. The other choice is the standard identity matrix. For the initial distribution, we also consider two options: one is $\mathcal{N}(\widehat{\theta}, n^{-1}\widehat{\Sigma}_m)$ with $\widehat{\theta}$ being the regression point estimator, as suggested in Corollary 7; and another choice is the standard normal distribution $\mathcal{N}(0, I_d)$. Figure 2 displays the minimum number of iteration required for achieving a Gelman-Rubin statistic smaller than 1.1, which is a common-used rule for determining the burn-in period (Flegal et al., 2008; Roy, 2020; Gelman and Rubin, 1992). We observe that choosing the initial distribution as $\mathcal{N}(\widehat{\theta}, n^{-1}\widehat{\Sigma}_m)$ allows the chain to converge in a very short period, whereas using $\mathcal{N}(0, I_d)$ requires a much longer time for convergence. Furthermore, we note that the mini-batch size does not significantly affect the required burn-in period, as choosing $m = 200$ is sufficient for fast convergence.

Figure 3 illustrates the largest step size allowed for achieving an average acceptance probability close to 0.57, as well as the effective sample size, after a total number of 5000 iterations with a burn-in period of 1000. We observe that utilizing the inverse empirical

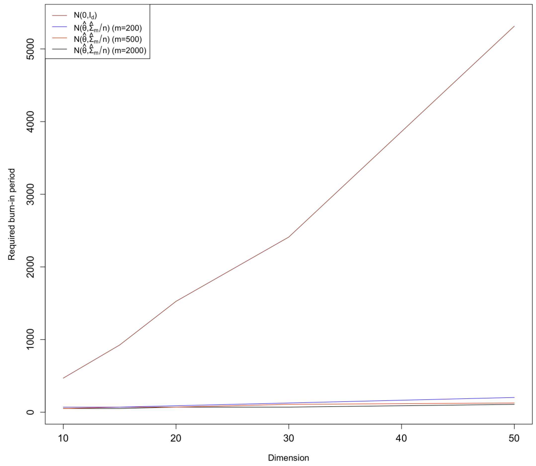


Figure 2: The figure shows the minimal burn-in period required to attain a Gelman-Rubin statistic below 1.1. It compares two scenarios: MALA with an initial distribution of $\mathcal{N}(0, I_d)$ and a preconditioning matrix of I_d, and MALA with an initial distribution of $\mathcal{N}(\widehat{\theta}, \widehat{\Sigma}_m/n)$ and a preconditioning matrix of $\hat{\Sigma}_m$. We can see the utilization of $\hat{\Sigma}_m$ for constructing the initial distribution and preconditioning matrix can significantly accelerates the convergence of MALA.

gram matrix enables a larger step size and leads to a larger effective sample size. Additionally, we find that the best performance is achieved when the batch size m is chosen to be equal to the sample size n. This is because a larger batch size provides a better estimator for $\Sigma^{-1} = (\mathbb{E}[\widetilde{X}\widetilde{X}])^{-1}$, resulting in a rescaled covariance matrix $\widehat{\Sigma}_m^{\frac{1}{2}}(\mathbb{E}[\widetilde{X}\widetilde{X}])^{-1}\widehat{\Sigma}_m^{\frac{1}{2}}$ with a smaller condition number. However, when $d \leq 20$, choosing $m = 500$ instead of using the full batch does not result in significant loss in performance.

A.2 Lemmas Related to s-conductance Profile

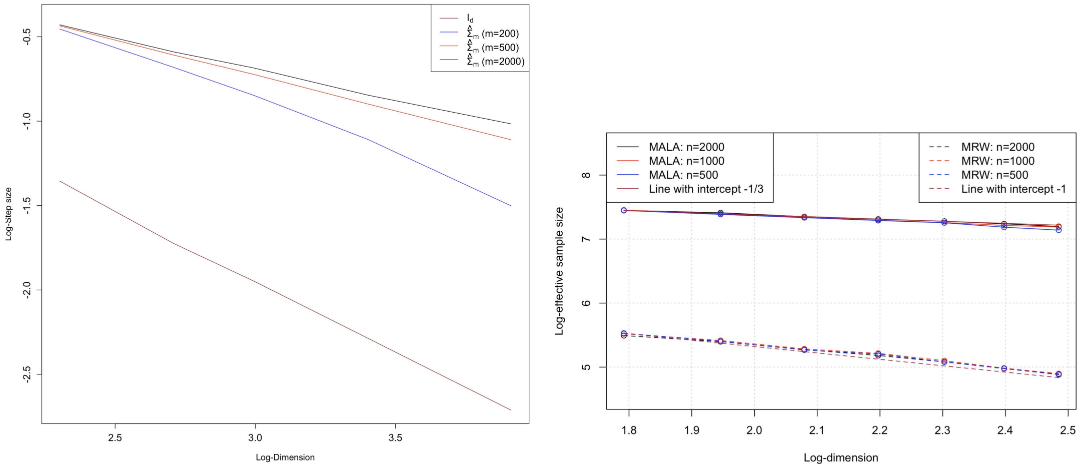
Lemma 10 (Mixing time bound via s-conductance profile) *Consider a reversible,*[8] *irreducible,*[9] *ζ-lazy*[10] *and smooth Markov chain*[11] *with stationary distribution μ. For any error tolerance $\varepsilon \in (0, 1)$, the maximal mixing time in χ^2 divergence of the chain over*

[8] A Markov chain with transition kernel $T : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ and stationary distribution μ is called reversible if $\mu(\mathrm{d}x)T(x, \mathrm{d}y) = \mu(\mathrm{d}y)T(y, \mathrm{d}x)$ holds for any $x, y \in \mathcal{X}$.

[9] A Markov chain with transition kernel $T : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ is irreducible if for all $x, y \in \mathcal{X}$, there is a natural number $k > 0$ so that $T^k(x, \mathrm{d}y) > 0$, where T^k is the k-step transition kernel.

[10] A Markov chain is said to be ζ-lazy if at each iteration, the chain is forced to stay at previous iterate with probability ζ. The laziness of Markov chain is also assumed in previous analysis based on s-conductance (Lovász and Simonovits, 1993) and conductance profile (Chen et al., 2020).

[11] We say that the Markov chain satisfies the smooth chain assumption if its transition probability function T can be expressed in the form $T(x, \mathrm{d}y) = \theta(x, y)\,\mathrm{d}y + \alpha_x\delta_x(\mathrm{d}y)$ where θ is a transition density function and δ_x is the Dirac measure at x.



(a) Log step size

(b) Log effective sample size

Figure 3: Plot (a) illustrates the logarithm of the maximum step size allowed to achieve an average acceptance probability close to 0.57 for various preconditioning matrices and dimensions. Plot (b) illustrates the relationship between the logarithm of the effective sample size and the logarithm of the dimension. The results demonstrate that choosing the preconditioning matrix based on the inverse of the empirical gram matrix enables the use of larger step sizes and leads to a higher effective sample size. Additionally, the disparity between different cases for various values of m becomes more pronounced as the dimension d increases. This is because the approximation error of $\widehat{\Sigma}_m$ increases with higher dimensions, necessitating a larger batch size for accurate estimation.

M_0-warm starts can be bounded as

$$\tau_{\text{mix}}(\varepsilon, M_0) \le \frac{16}{\zeta} \int_{\frac{4}{M_0}}^{\frac{1}{2}} \frac{\mathrm{d}v}{v\,\Phi_s^2(v)} + \frac{64}{\zeta} \int_{\frac{1}{2}}^{\frac{4\sqrt{2}}{\varepsilon}} \frac{\mathrm{d}v}{v\,\Phi_s^2(\frac{1}{2})},$$

where $s = \frac{\varepsilon^2}{16M_0^2}$.

We can calculate the second term of the upper bound above explicitly as $\frac{64}{\zeta\Phi_s^2(\frac{1}{2})} \log\big(\frac{8\sqrt{2}}{\xi}\big)$. The next lemma shows that the s-conductance profile can be lower bounded given one can: 1. prove a log-isoperimetric inequality for μ; 2. bound the total variation distance between $T(x,\cdot)$ and $T(z,\cdot)$ for any two sufficiently close points x, z in a high probability set (not necessarily convex) of μ, which will be referred to as the overlap argument.

Lemma 11 (s-conductance profile lower bound) *Consider a Markov chain with Markov transition kernel T and stationary distribution μ. Given a tolerance $\varepsilon \in (0,1)$ and warming parameter M_0, if there are two sets K, E, and positive numbers λ, ψ, ω so that*

1. *the probability measure of μ constrained on K, denoted as $\mu|_K(\cdot) = \frac{\mu(\cdot \cap K)}{\mu(K)}$, satisfies the following log-isoperimetric inequality:*

$$\mu|_K(S_3) \geq \lambda \cdot t \cdot \min\left\{\mu|_K(S_1),\, \mu|_K(S_2)\right\} \cdot \sqrt{\log\Big(1 + \frac{1}{\min\left\{\mu|_K(S_1),\, \mu|_K(S_2)\right\}}\Big)}\,,$$

   for any partition[12] *$K = S_1 \cup S_2 \cup S_3$ satisfying $\inf_{x \in S_1, z \in S_2} \|x - z\| \geq t$;*

2. *for any $x, z \in E$, if $\|x - z\| \leq \psi$, then $\|T(x,\cdot) - T(z,\cdot)\|_{\mathrm{TV}} \leq 1 - \omega$;*

3. *it holds that $\mu(E) \geq 1 - (\lambda\psi \wedge 1)\,\frac{\varepsilon^2}{256M_0^2}$ and $\mu(K) \geq 1 - (\lambda\psi \wedge 1)\,\frac{\varepsilon^2}{256M_0^2}$;*

then the s-conductance profile $\Phi_s(v)$ with $s = \frac{\varepsilon^2}{16M_0^2}$ can be bounded from below by

$$\Phi_s(v) \geq \frac{\omega}{4}\,\min\left\{1,\, \frac{\lambda\,\psi}{9}\sqrt{\log\big(1 + \frac{1}{v}\big)}\right\}.$$

By combining this lemma with Lemma 10, we obtain that if the assumptions in Lemma 11 hold, then the mixing time of the chain can be bounded as

$$\tau_{\mathrm{mix}}(\varepsilon, M_0) \leq \frac{C_1}{\zeta\omega^2}\,\log M_0 + \frac{C_1}{\zeta\omega^2}\,\lambda^{-2}\psi^{-2}\log(\log M_0) + \frac{C_1}{\zeta\omega^2}\,\lambda^{-2}\psi^{-2}\log\frac{1}{\varepsilon}, \tag{11}$$

for some universal constant C_1. Therefore, the problem of bounding the mixing time can be converted to verify the assumptions in Lemma 11.

A.3 Lower Bound of Mixing Time

Theorem 12 (MALA mixing time lower bound) *Consider a positive definite preconditioning matrix $\widetilde{I} \in \mathbb{R}^{d\times d}$, and the target distribution defined as a multivariate normal $\overline{\pi} = \mathcal{N}(0, J^{-1})$, where $J \in \mathbb{R}^{d\times d}$ is a covariance matrix with $\widetilde{I}^{\frac{1}{2}} J \widetilde{I}^{\frac{1}{2}} = \mathrm{diag}(\rho_2, \rho_2, \cdots, \rho_2, \rho_1)$. Assume $1 \leq \kappa = \frac{\rho_2}{\rho_1} \leq c_1 \cdot d^{c_2}$ for some $c_1, c_2 > 0$. Then there exists an integer N that depends only on c_1, c_2 and universal constants c_3, c_4 such that for any $d > N$, $M_0 \geq 2$, step size $h > 0$ and tolerance $\varepsilon \in (0,1)$, the $\frac{1}{2}$-lazy version of preconditioned MALA for sampling from $\overline{\pi}$ has the following mixing time lower bound in χ^2 divergence*

$$\tau_{\mathrm{mix}}(\varepsilon, M_0) \geq c_3\,\kappa\left(\frac{d}{\log(d\kappa)}\right)^{\frac{1}{3}}\log\left(\frac{c_4}{\varepsilon}\right).$$

A proof of Theorem 12 is provided in Appendix B.3, part of which is adapted from Chewi et al. (2021); Wu et al. (2022). Note that the worst-case construction used in Wu et al. (2022) does not satisfy our condition A. As a result, our lower bound has a different dimension dependence of $d^{1/3}$ than that in Wu et al. (2022) of $d^{1/2}$. Additionally, unlike Theorem 1 of Chewi et al. (2021), which considers a standard Gaussian target distribution (i.e., $\kappa = 1$), our lower bound has an explicit linear dependence on the condition number. From Theorem 3 and Theorem 12, we can see that when $\log\big(\frac{M_0\kappa}{\varepsilon}\big) = \mathcal{O}(d^{\frac{1}{3}})$, our mixing time upper bound and lower bound match up to some logarithmic terms of (d, κ) and a double logarithmic term of M_0.

[12] $\bigcup_{j=1}^J A_j$ forms a partition of set Ω means $\Omega = \bigcup_{j=1}^J A_j$ and $\{A_j\}_{j=1}^J$ are mutually disjoint.

Appendix B. Proof of Main Results

B.1 Proof of Theorem 3 (MALA mixing time upper bound)

Note that combined with Lemma 10, if the assumptions in Lemma 11 holds, we have

$$\begin{aligned}\tau_{\mathrm{mix}}(\varepsilon, \mu_0) &\leq \frac{C}{\zeta\omega} \int_{\frac{4}{M_0}}^{\frac{1}{2}} \frac{1}{v}\,\mathrm{d}v + \frac{C}{\zeta\omega} \int_{\frac{4}{M_0}}^{\frac{1}{2}} \lambda^{-2}\psi^{-2} \frac{1}{v\log(1+\frac{1}{v})}\,\mathrm{d}v + \frac{C}{\zeta\omega} \int_{\frac{1}{2}}^{\frac{4\sqrt{2}}{\varepsilon}} \lambda^{-2}\psi^{-2}\frac{1}{v}\,\mathrm{d}v \\ &\leq \frac{C_1}{\zeta\omega}\log M_0 + \frac{C_1}{\zeta\omega}\lambda^{-2}\psi^{-2}\log(\log M_0) + \frac{C_1}{\zeta\omega}\lambda^{-2}\psi^{-2}\log\frac{1}{\varepsilon},\end{aligned} \tag{12}$$

where the last inequality follows equation (18) of Chen et al. (2020). Now it remains to verify the assumptions in Lemma 11. Fix a lazy parameter $\zeta \in (0, \frac{1}{2}]$. Consider a linear transformation $G : \mathbb{R}^d \to \mathbb{R}^d$ defined as $G(\theta) = \sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta - \widehat{\theta})$, and let $\widetilde{\mu}_k = G_{\#}\mu_k$ denote the push forward measure of G by μ_k for $k \in \mathbb{N}$ and $\widetilde{\pi}_{\mathrm{loc}}$ denote the push forward measure of G by π_n. Then it holds that

$$M_0 = \sup_{A:\pi_n(A)>0} \frac{\mu_0(A)}{\pi_n(A)} = \sup_{A:\widetilde{\pi}_{\mathrm{loc}}(A)>0} \frac{\widetilde{\mu}_0(A)}{\widetilde{\pi}_{\mathrm{loc}}(A)}.$$

Moreover, by the invariability of χ^2 measure to linear transformation, we have $\chi^2(\mu_k, \pi_n) = \chi^2(\widetilde{\mu}_k, \widetilde{\pi}_{\mathrm{loc}})$. Define $\widetilde{Q}(\xi, \cdot)$ be the density function of the multivarite normal $N_d(\xi - h\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n(\widetilde{I}^{\frac{1}{2}}\xi), 2h\, I_d)$, and the corresponding Markov transition kernel

$$\widetilde{T}(\xi, \mathrm{d}y) = \left[1 - (1-\zeta)\cdot\int \widetilde{A}(\xi, y)\widetilde{Q}(\xi, y)\,\mathrm{d}y\right]\mathbf{1}_{\xi}(\mathrm{d}y) + (1-\zeta)\cdot\widetilde{Q}(\xi, y)\widetilde{A}(\xi, y)\,\mathrm{d}y$$

with

$$\widetilde{A}(\xi, y) = 1 \wedge \frac{\widetilde{\pi}_{\mathrm{loc}}(y)\widetilde{Q}(y, \xi)}{\widetilde{\pi}_{\mathrm{loc}}(\xi)\widetilde{Q}(\xi, y)}.$$

We have the following lemma.

Lemma 13 *For any $k \in \mathbb{N}$, $\widetilde{\mu}_k = G_{\#}\mu_k$ is the probability distribution obtained after k steps of a Markov chain with transition kernel $\widetilde{T}$ and initial distribution $\widetilde{\mu}_0$.*

It remains to calculate the mixing time of $\widetilde{\mu}_k$ converging to $\widetilde{\pi}_{\mathrm{loc}}$, which is equivalent to verify the assumptions in Lemma 11 for Markov transition kernel $\widetilde{T}(\xi, \cdot)$ with stationary distribution $\widetilde{\pi}_{\mathrm{loc}}$. Recall $K = \{x : \|\widetilde{I}^{-\frac{1}{2}}x\| \leq R\}$. By Condition A, firstly we have

$$\begin{aligned}&\sup_{\widetilde{\xi}\in B_R^d}\left|V_n(\widetilde{I}^{\frac{1}{2}}\widetilde{\xi}) - \frac{1}{2}\widetilde{\xi}^T\widetilde{I}^{\frac{1}{2}}J\widetilde{I}^{\frac{1}{2}}\widetilde{\xi}\right| = \sup_{\xi\in K}\left|V_n(\xi) - \frac{1}{2}\xi^T J\xi\right| \leq \widetilde{\varepsilon}_0; \\ &\sup_{\widetilde{\xi}\in B_R^d}\left\|\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n(\widetilde{I}^{\frac{1}{2}}\xi) - \widetilde{I}^{\frac{1}{2}}J\widetilde{I}^{\frac{1}{2}}\xi\right\| = \sup_{\xi\in K}\left\|\widetilde{I}^{\frac{1}{2}}(\widetilde{\nabla}V_n(\xi) - J\xi)\right\| \leq \widetilde{\varepsilon}_1 \|\!|\widetilde{I}^{\frac{1}{2}}|\!\|_{\mathrm{op}},\end{aligned}$$

and $\widetilde{\pi}_{\mathrm{loc}}(\widetilde{\xi} \in B_{R/2}^d) = \pi_n(\|\sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta - \widehat{\theta})\| \leq R/2) \geq 1 - \exp(-4\widetilde{\varepsilon}_0)\cdot\frac{h\rho_1\varepsilon^2}{M_0^2}$. We then verify the log-isoperimetric inequality in the following lemma.

Lemma 14 *Let $\widetilde{K} = B^d_{R/2}$, consider any measurable partition form $\widetilde{K} = S_1 \cup S_2 \cup S_3$ such that $\inf_{x\in S_1, z\in S_2} \|x - z\| \geq t$, we have*

$$\widetilde{\pi}_{\mathrm{loc}}|_{\widetilde{K}}(S_3) \geq \frac{\sqrt{\rho_1}}{2} t \exp(-4\widetilde{\varepsilon}_0) \min\{\widetilde{\pi}_{\mathrm{loc}}|_{\widetilde{K}}(S_1), \widetilde{\pi}_{\mathrm{loc}}|_{\widetilde{K}}(S_2)\} \log^{\frac{1}{2}}\left(1+\frac{1}{\min\{\widetilde{\pi}_{\mathrm{loc}}|_{\widetilde{K}}(S_1), \widetilde{\pi}_{\mathrm{loc}}|_{\widetilde{K}}(S_2)\}}\right).$$

We then show that $\|\widetilde{T}(x,\cdot) - \widetilde{T}(y,\cdot)\|_{\mathrm{TV}}$ can be bounded with high probability in the following lemma.

Lemma 15 *There exists a set E so that $\widetilde{\pi}_{\mathrm{loc}}(E) \geq 1 - \exp(-4\widetilde{\varepsilon}_0) \cdot \frac{2\varepsilon^2 h\rho_1}{M_0^2}$ and for any $x, z \in E$ with $\|x - z\| \leq \frac{\sqrt{h}}{3}$, we have $\|\widetilde{T}(x,\cdot) - \widetilde{T}(z,\cdot)\|_{\mathrm{TV}} \leq 1 - \frac{\exp(-2\widetilde{\varepsilon}_0)}{4}$.*

Thus the first and second assumptions in Lemma 11 holds with $\lambda = \frac{\sqrt{\rho_1}}{2}\exp(-4\widetilde{\varepsilon}_0)$, $\psi = \frac{\sqrt{h}}{3}$ and $\omega = \frac{\exp(-2\widetilde{\varepsilon}_0)}{4}$. Moreover, for the third assumption in Lemma 11, by $h\rho_1 \leq c_0 d^{-\frac{1}{3}}$, for small enough c_0, we have

$$\exp(-4\widetilde{\varepsilon}_0) \cdot \frac{2\varepsilon^2 h\rho_1}{M_0^2} \leq \frac{\sqrt{2h}}{24}\frac{\sqrt{\rho_1}}{2}\exp(-4\widetilde{\varepsilon}_0)\frac{\varepsilon^2}{256 M_0^2}.$$

Thus all the assumptions in Lemma 11 are satisfied. The desired result then follows from equation (12).

B.2 Proof of Theorem 12 (MALA mixing time lower bound)

Without loss of generality, we assume $\widetilde{I} = I_d$. Otherwise, similar as the proof of Theorem 3, we could transform the measures μ_k and $\overline{\pi}$ by the scale matrix $\widetilde{I}^{-\frac{1}{2}}$, and study the convergence of the transformed measures. We utilize the following lower bound on the χ^2-divergence via Dirichlet form.

Lemma 16 *(Corollary 7 of Wu et al. (2022)) Le T be the transition kernel of a reversible Markov chain with invariant distribution $\overline{\pi}$. For any $\varepsilon > 0$ and any initial distribution $\mu_0 \ll \overline{\pi}$ satisfying $\chi^2(\mu_0, \overline{\pi}) < \infty$, let $h_0 = \frac{\mathrm{d}\mu_0}{\mathrm{d}\overline{\pi}}$, if $\mathcal{E}(h_0,h_0)/\chi^2(\mu_0,\overline{\pi}) \leq \frac{1}{4}$ with $\mathcal{E}(\cdot,\cdot)$ being the Dirichlet form associated with T, then its mixing time in χ^2-divergence has a lower bound*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0) \geq \frac{1}{4}\left(\frac{\mathcal{E}(h_0,h_0)}{\chi^2(\mu_0,\overline{\pi})}\right)^{-1}\log\left(\frac{\chi^2(\mu_0,\overline{\pi})}{\varepsilon^2}\right).$$

Then, we state the following lemma for bounding $\mathcal{E}(h_0,h_0)/\chi^2(\mu_0,\overline{\pi})$.

Lemma 17 *Consider the target distribution $\overline{\pi} = N_d(0, J^{-1})$ with $J = \mathrm{diag}(\rho_2, \rho_2, \cdots, \rho_2, \rho_1)$ and $1 \leq \kappa = \frac{\rho_2}{\rho_1} \leq c_1 \cdot d^{c_2}$, then*

1. *There exists a 2-warm initial distribution μ_0 with $\chi^2(\mu_0,\overline{\pi}) \geq \frac{1}{5}$ so that for any $h \in (0, \frac{1}{\rho_1})$, denote $h_0 = \frac{\mathrm{d}\mu_0}{\mathrm{d}\overline{\pi}}$, then for any $\zeta \in [0,1]$, the term $\mathcal{E}(h_0,h_0)/\chi^2(\mu_0,\overline{\pi})$ under the ζ-lazy version MALA transition kernel with step size h satisfies*

$$\frac{\mathcal{E}(h_0,h_0)}{\chi^2(\mu_0,\overline{\pi})} \leq 60\rho_1 h.$$

2. *When $M_0 \geq 2$, there exists an M_0-warm initial distribution μ_0' with $\chi^2(\mu_0', \overline{\pi}) = M_0 - 1$ and a constant N that depends only on c_1, c_2 so that when $d \geq N$, denote $h_0 = \frac{\mathrm{d}\mu_0'}{\mathrm{d}\overline{\pi}}$, for any $h \in (\frac{8(\log(d\kappa))^{\frac{1}{3}}}{\rho_2 d^{\frac{1}{3}}}, \infty)$, for any $\zeta \in [0,1]$, the term $\mathcal{E}(h_0, h_0)/\chi^2(\mu_0', \overline{\pi})$ under the ζ-lazy version MALA transition kernel with step size h satisfies*
$$\frac{\mathcal{E}(h_0, h_0)}{\chi^2(\mu_0', \overline{\pi})} \leq \frac{8}{\kappa d}.$$

So when $d \geq N \vee 3$, if $h > \frac{8(\log(d\kappa))^{\frac{1}{3}}}{\rho_2 d^{\frac{1}{3}}}$, we have
$$\sup_{2-\text{warm } \mu_0} \tau_{\text{mix}}(\varepsilon, \mu_0) \geq \frac{\kappa d}{46} \log(\frac{1}{\varepsilon^2}) \geq \frac{\kappa d^{\frac{1}{3}}}{46} \log(\frac{1}{\varepsilon^2});$$
when $h \leq \frac{8(\log(d\kappa))^{\frac{1}{3}}}{\rho_2 d^{\frac{1}{3}}}$, we have $\rho_1 h < 1$ and thus,
$$\sup_{2-\text{warm } \mu_0} \tau_{\text{mix}}(\varepsilon, \mu_0) \geq \frac{1}{240} \rho_1^{-1} h^{-1} \log(\frac{1}{5\varepsilon^2}) \geq \frac{\kappa d^{\frac{1}{3}}}{1920(\log(d\kappa))^{\frac{1}{3}}} \log(\frac{1}{5\varepsilon^2}).$$

Proof is completed.

B.3 Proof of Theorem 5 (Complexity of MALA for Bayesian sampling)

Without loss of generality, we can assume the learning rate $\alpha = 1$, as otherwise we can take $\ell(X, \theta) = \alpha \cdot \ell(X, \theta)$. We only need to verify that the Assumptions in Theorem 3 holds for the Bayesian Gibbs posterior. We state the following Lemmas to verify Condition A.

Lemma 18 *Let $\kappa_2 = \frac{\beta_1}{\gamma_3 + \beta_1(1+\gamma_4) + 2\gamma_0 - \gamma_4} \wedge \frac{1}{1+2\gamma+2\gamma_2+4\gamma_0} \wedge \frac{1}{2+2(\gamma+\gamma_0+\gamma_1)}$. Under Conditions B.1-B.4, if $d \leq c(\frac{n}{\log n})^{\kappa_2}$ for a small enough constant c, then there exist (n,d)-independent constants c_1, C, C_1 so that it holds with probability at least $1 - c_1 n^{-2}$ that for any $\xi \in \mathbb{R}^d$ with $1 \leq \|\xi\| \leq C\sqrt{n}$,*
$$\begin{aligned}\Big|V_n(\xi) - \frac{\xi^T \mathcal{H}_{\theta^*} \xi}{2}\Big| \leq C_1 \Big(& d^{1+\gamma} \|\xi\| \frac{\log n}{\sqrt{n}} + d^{\gamma_2} \|\xi\|^3 \frac{1}{\sqrt{n}} + d^{\frac{1+\gamma_4}{2}+\gamma_2} \|\xi\|^2 \sqrt{\frac{\log n}{n}} \\ &+ d^{\frac{1+\gamma_3}{2}} \|\xi\|^{1+\beta_1} \frac{\sqrt{\log n}}{n^{\beta_1/2}}\Big); \\ \Big\|\widetilde{\nabla} V_n(\xi) - \mathcal{H}_{\theta^*}\xi\Big\| \leq C_1 \Big(& d^{1+\gamma} \frac{\log n}{\sqrt{n}} + d^{\gamma_2} \|\xi\|^2 \frac{1}{\sqrt{n}} + d^{\frac{1+\gamma_4}{2}+\gamma_2} \|\xi\| \sqrt{\frac{\log n}{n}} \\ + d^{\frac{1+\gamma_3}{2}} \|\xi\|^{\beta_1} \frac{\sqrt{\log n}}{n^{\beta_1/2}}\Big) &\text{ with } \widetilde{\nabla} V_n(\xi) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g\Big(X_i, \frac{\xi}{\sqrt{n}} + \widehat{\theta}\Big) - \frac{1}{\sqrt{n}} \nabla(\log \pi)\Big(\frac{\xi}{\sqrt{n}} + \widehat{\theta}\Big).\end{aligned}$$

We provide in the following lemma a tail inequality for the Gibbs posterior π_n.

Lemma 19 *Under Condition B.1-B.4. when $d \leq c\frac{n^{\kappa_3}}{\log n}$ for a small enough constant c, where*

$$\kappa_3 = \frac{\beta_1}{1+\gamma_3+[(2\gamma_0)\vee((1+\gamma_0)(1+\beta_1))]} \wedge \frac{1}{3+\gamma_0+((2\gamma)\vee(\gamma_4+2\gamma_2+\gamma_0)\vee(2\gamma_2+2\gamma_0))}$$
$$\wedge \frac{1}{1+2\gamma+6\gamma_0+4\gamma_2+\gamma_4} \wedge \frac{1}{2\gamma+2\gamma_0+2\gamma_1+(2\vee(1+\gamma_4))},$$

then there exist (n,d)-independent constants c_1, c_2, c_3 so that it holds with probability at least $1-c_1n^{-2}$ that

$$\pi_n\Big(\sqrt{n}\|\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})\| \geq |||\widetilde{I}^{-\frac{1}{2}}|||_{\mathrm{op}} \vee \frac{3(\sqrt{d}+t)}{\sqrt{\lambda_{\min}(\widetilde{J})}}\Big) \leq \exp(-t^2)+c_2\exp\big(-c_3 n\, d^{-\gamma_0}(d^{-\gamma_1}\wedge d^{-2\gamma_0-2\gamma_2})\big),$$

where $\widetilde{J} = \widetilde{I}^{\frac{1}{2}}\mathcal{H}_{\theta^*}\widetilde{I}^{\frac{1}{2}}$.

By Condition B.1, we have $|||\mathcal{H}_{\theta^*}|||_{\mathrm{op}} \leq C\,d$ and $|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}} \leq C\,d^{\gamma_0}$. Moreover, since $|||\widetilde{I}^{-1}|||_{\mathrm{op}}|||\widetilde{I}|||_{\mathrm{op}} \leq C|||\mathcal{H}_{\theta^*}|||_{\mathrm{op}}|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}$ and $|||\widetilde{I}|||_{\mathrm{op}}|||(\widetilde{I}^{\frac{1}{2}}H_{\theta^*}\widetilde{I}^{\frac{1}{2}})^{-1}|||_{\mathrm{op}} \leq C\,|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}$, we can obtain that there exists constants C_2, C_3 so that for any $R = |||\widetilde{I}^{-\frac{1}{2}}|||_{\mathrm{op}} \vee \frac{3(\sqrt{d}+t)}{\sqrt{\lambda_{\min}(\widetilde{J})}}$ with $t \geq 0$, and set $K = \{x : \|\widetilde{I}^{-1/2}x\| \leq R\}$, we have $K \subseteq \{x : \|x\| \leq C_2 d^{\frac{1+\gamma_0}{2}} + C_3 t d^{\frac{\gamma_0}{2}}\}$. Then by Lemma 18, for any $t = C_1\,(d^{\frac{\gamma_5}{2}} + \sqrt{\log n})$ (note that $\gamma_5 \geq 1$), we can find a constant c so that when $d \leq c\frac{n^{\kappa_1}}{\log n}$, we have

$$|||\widetilde{I}|||_{\mathrm{op}}\, R^2 \sup_{\xi\in K}\|\widetilde{\nabla}V_n(\xi) - \mathcal{H}_{\theta^*}\xi\Big\|^2 \leq d^{\frac{1}{3}}.$$

So in this case the step size parameter $\widetilde{h} = h/n$ in Theorem 3 satisfies

$$h \leq c_0 \cdot \Big[\rho_2\Big(2d^{\frac{1}{3}} + d^{\frac{1}{4}}\big(\log\frac{M_0 d\kappa}{\varepsilon}\big)^{\frac{1}{4}} + \big(\log\frac{M_0 d\kappa}{\varepsilon}\big)^{\frac{1}{2}}\Big)\Big]^{-1}$$

Then by the assumption $\log(\frac{M_0\rho_2}{\varepsilon\rho_1}) \leq C_1\,(d^{\gamma_5} + \log n)$, using Lemma 19 and $d \leq c\frac{n^{\kappa_1}}{\log n}$, we can obtain that there exists a large enough C_1 so that for $t = C_1\,(d^{\frac{\gamma_5}{2}} + \sqrt{\log n})$, $\pi_n(K) = \pi_n\Big(\sqrt{n}\|\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})\| \geq |||\widetilde{I}^{-\frac{1}{2}}|||_{\mathrm{op}} \vee \frac{3(\sqrt{d}+t)}{\sqrt{\lambda_{\min}(\widetilde{J})}}\Big) \geq 1 - \frac{h\rho_1\varepsilon^2}{M_0^2}$. So the Assumptions in Theorem 3 are satisfied.

Appendix C. Proof of Lemmas for Theorem 3 and Theorem 12

C.1 Proof of Lemma 10

Fix an arbitrary $\varepsilon > 0$. Suppose $\tau_{\mathrm{mix}}(\sqrt{2}\varepsilon, \mu_0) > N = \int_{\frac{4}{M_0}}^{\frac{1}{2}} \frac{16\mathrm{d}v}{\zeta\cdot v\Phi_s^2(v)} + \int_{\frac{1}{2}}^{\frac{4}{\varepsilon}} \frac{64\mathrm{d}v}{\zeta\cdot v\Phi_s^2(\frac{1}{2})}$. Then for any $k \leq N$, $\chi^2(\mu_k, \mu) > 2\varepsilon^2$, where we use $\chi^2(\cdot,\cdot)$ to denote the χ^2 divergence, μ_k to denote the distribution in k step of the Markov chain and $\mu \in \mathcal{P}(\mathbb{R}^d)$ to denote the stationary distribution. Then we will prove by contradiction that if $N < \tau_{\mathrm{mix}}(\sqrt{2}\varepsilon, \mu_0)$, then when

$k = N$, $\chi^2(\mu_k, \mu) \le 2\varepsilon^2$, which oppositely implies $N \ge \tau_{\mathrm{mix}}(\sqrt{2}\varepsilon, \mu_0)$. Our proof is based on the strategy used in Chen et al. (2020). We first introduce the following related notations. For a measurable set $S \subseteq \mathbb{R}^d$ and positive numbers ε, M_0 , the (ε, M_0)-spectral gap for the set S is defined as

$$\Lambda_{\varepsilon,M_0}(S) := \inf_{g \in c^+_{\varepsilon,M_0}(S)} \frac{\mathcal{E}(g,g)}{\mathrm{Var}_\mu(g)}$$

where

$$c^+_{\varepsilon,M_0}(S) := \left\{ g \in L_2(\mu) \mid \mathrm{supp}(g) = \{x : g(x) > 0\} \subset S,\, 0 \le g \le M_0,\, \mathrm{Var}_\mu(g) \ge \varepsilon^2 \right\},$$

and

$$\mathcal{E}(g,g) = \frac{1}{2}\int (g(x) - g(y))^2 T(x, \mathrm{d}y)\mu(\mathrm{d}x),$$

with $T(x, \mathrm{d}y)$ denoting the Markov transition kernel. Moreover, we can define the (ε, M_0, s)-spectral profile $\overline{\Lambda}^{\varepsilon,M_0}_s$ as

$$\overline{\Lambda}^{\varepsilon,M_0}_s(v) := \inf_{\mu(S) \in (s,v]} \Lambda_{\varepsilon,M_0}(S).$$

Define the ratio density

$$h_k(x) = \frac{\mu_k(x)}{\mu(x)}.$$

Note that

$$\mathbb{E}_\mu[h_k] = 1 \quad \text{and} \quad \chi^2(\mu_k, \mu) = \mathrm{Var}_\mu(h_k),$$

and $h_k(x) \le M_0$ for all $k \ge 0$ (see for example, equation (64) of Chen et al. (2020)). By tracking the proof of Lemma 11 in Chen et al. (2020), it suffices to show that for any $k \le N$,

$$2\,\mathcal{E}(h_k, h_k) \ge \mathrm{Var}_\mu(h_k)\overline{\Lambda}^{\varepsilon,M_0}_s\Big(\frac{4}{\mathrm{Var}_\mu(h_k)}\Big), \tag{13}$$

and

$$\overline{\Lambda}^{\varepsilon,M_0}_s(v) \ge \begin{cases} \frac{\Phi_s^2(v)}{16} & \text{for all } v \in \left[\frac{4}{M_0}, \frac{1}{2}\right]; \\ \frac{\Phi_s^2(\frac{1}{2})}{64} & \text{for all } v \in (\frac{1}{2}, \infty), \end{cases} \tag{14}$$

with $s = \frac{\varepsilon^2}{16M_0^2}$. We first prove claim (13). Define $\gamma_k = \frac{\mathrm{Var}_\mu(h_k)}{4\mathbb{E}_\mu[h_k]} = \frac{\mathrm{Var}_\mu(h_k)}{4}$. Then for any $k \le N$,

$$\begin{aligned}
\mathrm{Var}_\mu\left((h_k - \gamma_k)_+\right) &= \mathbb{E}_\mu\left[((h_k - \gamma_k)_+)^2\right] - \left(\mathbb{E}_\mu\left[(h_k - \gamma_k)_+\right]\right)^2 \\
&\overset{(i)}{\ge} \mathbb{E}_\mu[h_k^2] - 2\gamma_k\mathbb{E}_\mu[h_k] - (\mathbb{E}_\mu[h_k])^2 \\
&= \mathrm{Var}_\mu(h_k) - 2\gamma_k\mathbb{E}_\mu[h_k] \\
&= \frac{1}{2}\mathrm{Var}_\mu(h_k) \ge \varepsilon^2,
\end{aligned}$$

where $(x)_+ = \max\{0, x\}$, (i) is due to $((a-b)_+)^2 \le a^2 - 2ab$, $(a-b)_+ \le a$, and the last inequality is due to the assumption that $N < \tau_{\mathrm{mix}}(\sqrt{2}\varepsilon, \mu_0)$; moreover, since for any $x \in \mathbb{R}^d$,

$0 \leq h_k(x) \leq M_0$, we can get $(h_k - \gamma_k)_+ \in c^+_{\varepsilon, M_0}(\{h_k > \gamma_k\})$, which leads to

$$\mathcal{E}(h_k, h_k) \overset{(ii)}{\geq} \mathcal{E}((h_k - \gamma_k)_+, (h_k - \gamma_k)_+) \geq \operatorname{Var}_\mu((h_k - \gamma_k)_+) \cdot \inf_{f \in c^+_{\varepsilon, M_0}(\{h_k > \gamma_k\})} \frac{\mathcal{E}(f,f)}{\operatorname{Var}_\mu(f)}, \tag{15}$$

where (ii) follows from the fact that $(a-b)^2 = (a - c - (b-c))^2 \geq ((a-c)_+ - (b-c)_+)^2$. Furthermore, We have for any $k \leq N$,

$$M_0^2\, \mu(h_k \geq \gamma_k) \geq \mathbb{E}_\mu[((h_k - \gamma_k)_+)^2] \geq \operatorname{Var}_\mu((h_k - \gamma_k)_+) \geq \frac{1}{2} \operatorname{Var}_\mu(h_k) \geq \varepsilon^2.$$

On the other hand, by applying Markov's inequality, we also have

$$\mu(h_k \geq \gamma_k) \leq \frac{\mathbb{E}_\mu[h_k]}{\gamma_k} = \frac{4}{\operatorname{Var}_\mu(h_k)}.$$

Thus by equation (15), we can get for $s = \frac{\varepsilon^2}{16 M_0^2}$,

$$\mathcal{E}(h_k, h_k) \geq \frac{1}{2} \operatorname{Var}_\mu(h_k) \overline{\Lambda}_s^{\varepsilon, M_0}\Big(\frac{4}{\operatorname{Var}_\mu(h_k)}\Big).$$

Then we prove claim (14). For $v \in \big[\frac{4}{M_0}, \frac{1}{2}\big]$, fix any $A \subset \mathbb{R}^d$ with $s < \mu(A) \leq v$ and $g \in c^+_{\frac{\varepsilon}{2}, M_0}(A)$. Then by

$$\begin{aligned}
&\mathbb{E}_\mu\Big[\int (g^2(x) - g^2(y))_+ T(x, \mathrm{d}y)\Big] \\
&= \mathbb{E}_\mu\Big[\int (g^2(x) - g^2(y)) \mathbf{1}(g^2(x) > g^2(y)) T(x, \mathrm{d}y)\Big] \\
&= \mathbb{E}_\mu\Big[\int \int_0^{+\infty} \mathbf{1}(g^2(y) \leq t < g^2(x))\, \mathrm{d}t\, T(x, \mathrm{d}y)\Big] \\
&= \int_0^{+\infty} \mathbb{E}_\mu\Big[\int \mathbf{1}(g^2(y) \leq t < g^2(x)) T(x, \mathrm{d}y)\Big]\, \mathrm{d}t,
\end{aligned}$$

let $H_t = \{x \in \mathbb{R}^d : g^2(x) > t\}$, we have

$$\begin{aligned}
&\int \int |g^2(x) - g^2(y)| T(x, \mathrm{d}y) \mu(\mathrm{d}x) \\
&\geq \int \int (g^2(x) - g^2(y))_+ T(x, \mathrm{d}y) \mu(\mathrm{d}x) \\
&= \int_0^{+\infty} \mathbb{E}_\mu\Big[\int \mathbf{1}(g^2(y) \leq t < g^2(x)) T(x, \mathrm{d}y)\Big]\, \mathrm{d}t \\
&= \int_0^{+\infty} \int_{x \in H_t} T(x, H_t^c) \mu(\mathrm{d}x)\, \mathrm{d}t.
\end{aligned}$$

Let $t^* = \sup\{t \geq 0 : \mu(H_t) > s\}$, note that t^* always exists as otherwise, $\mu(g(x) = 0) \geq 1 - s$ and thus $\operatorname{Var}_\mu(g) \leq M_0^2 s = \frac{\varepsilon^2}{16}$, which is contradictory to the requirement that $\operatorname{Var}_\mu(g) \geq \frac{\varepsilon^2}{4}$.

Then

$$\begin{aligned}
&\int\int |g^2(x)-g^2(y)|T(x,\mathrm{d}y)\mu(\mathrm{d}x)\\
&\geq \int_0^{t^*}\int_{x\in H_t} T(x,H_t^c)\mu(\mathrm{d}x)\,\mathrm{d}t + \int_{t^*}^{+\infty}\int_{x\in H_t} T(x,H_t^c)\mu(\mathrm{d}x)\,\mathrm{d}t\\
&\geq \int_0^{t^*}(\mu(H_t)-s)\,\mathrm{d}t\cdot\Phi_s(\mu(A))\\
&= \Big(\mathbb{E}_\mu[g^2]-\int_{t^*}^{M_0^2}\mu(H_t)\,\mathrm{d}t - st^*\Big)\cdot\Phi_s(\mu(A))\\
&\overset{(ii)}{\geq} \Big(\mathbb{E}_\mu[g^2]-\frac{\varepsilon^2}{8}\Big)\cdot\Phi_s(\mu(A))\\
&\overset{(iii)}{\geq} \frac{1}{2}\mathbb{E}_\mu[g^2]\Phi_s(\mu(A)),
\end{aligned}$$

where (ii) uses the fact that $t^*\leq M_0^2$ and when $t>t^*$, $\mu(H_t)\leq s=\frac{\varepsilon^2}{16M_0^2}$ and (iii) uses $\mathbb{E}_\mu[g^2]\geq \mathrm{Var}_\mu(g)\geq\frac{\varepsilon^2}{4}$. Moreover, since

$$\begin{aligned}
&\int\int |g^2(x)-g^2(y)|T(x,\mathrm{d}y)\mu(\mathrm{d}x)\\
&\leq \sqrt{\int\int (g(x)-g(y))^2T(x,\mathrm{d}y)\mu(\mathrm{d}x)}\cdot\sqrt{\int\int (g(x)+g(y))^2T(x,\mathrm{d}y)\mu(\mathrm{d}x)}\\
&\leq \sqrt{2\mathcal{E}(g,g)}\cdot\sqrt{\int\int (2g^2(x)+2g^2(y))T(x,\mathrm{d}y)\mu(\mathrm{d}x)}\\
&= \sqrt{2\,\mathcal{E}(g,g)}\cdot\sqrt{4\,\mathbb{E}_\mu[g^2]},
\end{aligned}$$

we have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_\mu[g^2]\cdot\Phi_s(\mu(A))\leq\sqrt{2\,\mathcal{E}(g,g)}\cdot\sqrt{4\,\mathbb{E}_\mu[g^2]}\\
\Rightarrow\ &\frac{\mathcal{E}(g,g)}{\mathrm{Var}_\mu(g)}\geq\frac{\Phi_s^2(\mu(A))}{16}.
\end{aligned}$$

Taking infimum over $A\subset\mathbb{R}^d$ with $s<\mu(A)\leq v$ and $g\in c^+_{\frac{\varepsilon}{2},M_0}(A)$, we have

$$\overline{\Lambda}_s^{\varepsilon,M_0}(v)\geq\ \overline{\Lambda}_s^{\frac{\varepsilon}{2},M_0}(v)\geq\inf_{s<\mu(A)\leq v}\frac{\Phi_s^2(\mu(A))}{16}\geq\frac{\Phi_s^2(v)}{16}. \tag{16}$$

For the case $v>\frac{1}{2}$, consider any $A\subset\mathbb{R}^d$ with $\mu(A)>\frac{1}{2}$ and $g\in c^+_{\varepsilon,M_0}(A)$. Let $0\leq\gamma\leq M_0$ be the number such that

$$s<\mu(\{g>\gamma\})\vee\mu(\{g<\gamma\})\leq\frac{1}{2}.$$

γ always exists as otherwise, there exists $0\leq\widetilde{\gamma}\leq M_0$ such that $\mu\{g=\widetilde{\gamma}\}\geq 1-2s$, which leads to $\mathrm{Var}_\mu(g)\leq\mathbb{E}_\mu[(g-\widetilde{\gamma})^2]\leq 2M_0^2s<\varepsilon^2$, and this causes contradiction. We first consider the case that $\mu(\{g>\gamma\})\wedge\mu(\{g<\gamma\})>s$. We have

$$\mathcal{E}(g,g)=\mathcal{E}((g-\gamma),(g-\gamma))\geq\mathcal{E}((g-\gamma)_+,(g-\gamma)_+)+\mathcal{E}((g-\gamma)_-,(g-\gamma)_-).$$

Since for any function $h \geq 0$ with $\mu(\mathrm{supp}(h)) \leq \frac{1}{2}$, using Cauchy-Schwarz inequality, it holds that

$$\mathbb{E}_\mu[h^2] = \int_{x \in \mathrm{supp(h)}} h^2(x)\mu(x)dx \geq \frac{(\mathbb{E}_\mu[h])^2}{\mu(\mathrm{supp}(h))} \geq 2\,(E_\mu[h])^2,$$

which leads to

$$\mathrm{Var}_\mu(h) \geq \frac{1}{2}\mathbb{E}_\mu[h^2].$$

Since $\varepsilon^2 \leq \mathrm{Var}_\mu(g) \leq \mathbb{E}_\mu[(g-\gamma)^2]$ and $\mathbb{E}_\mu[(g-\gamma)^2] = \mathbb{E}_\mu[(g-\gamma)_+^2] + \mathbb{E}_\mu[(g-\gamma)_-^2]$, w.l.o.g, we can assume $\mathbb{E}_\mu[(g-\gamma)_+^2] \geq \frac{\mathbb{E}_\mu[(g-\gamma)^2]}{2} \geq \frac{\varepsilon^2}{2}$. Then taking $h = (g-\gamma)_+$, we can obtain

$$\begin{aligned}
\mathcal{E}(g,g) &\geq \mathcal{E}((g-\gamma)_+, (g-\gamma)_+) \\
&\geq \mathbb{E}_\mu[(g-\gamma)_+^2] \cdot \frac{\mathcal{E}((g-\gamma)_+, (g-\gamma)_+)}{2\,\mathrm{Var}_\mu((g-\gamma)_+)} \\
&\overset{(i)}{\geq} \frac{1}{4}\mathrm{Var}_\mu(g) \cdot \inf_{\mu(S)\in(s,\frac{1}{2}]} \inf_{f \in c^+_{\frac{\varepsilon}{2},M_0}(S)} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\mu(f)} \qquad (17) \\
&\overset{(ii)}{\geq} \frac{1}{64}\mathrm{Var}_\mu(g)\Phi_s^2(\frac{1}{2}),
\end{aligned}$$

where (i) uses $\mathbb{E}_\mu[(g-\gamma)_+^2] \geq \frac{\mathbb{E}_\mu[(g-\gamma)^2]}{2} \geq \frac{\mathrm{Var}_\mu(g)}{2}$ and $\mathrm{Var}_\mu((g-\gamma)_+) \geq \frac{1}{2}\mathbb{E}_\mu[(g-\gamma)_+^2] \geq \frac{\varepsilon^2}{4}$, and (ii) uses (16). Then we consider the case that $\mu(\{g > \gamma\}) \wedge \mu(\{g < \gamma\}) \leq s < \mu(\{g > \gamma\}) \vee \mu(\{g < \gamma\})$. W.l.o.g, we can assume $\mu(\{g > \gamma\}) > s$. Then we can obtain

$$\begin{aligned}
\mathbb{E}_\mu[(g-\gamma)_+^2] &= \mathbb{E}_\mu[(g-\gamma)^2] - \mathbb{E}_\mu[(g-\gamma)_-^2] \\
&\geq \mathbb{E}_\mu[(g-\gamma)^2] - M_0^2 s = \mathbb{E}_\mu[(g-\gamma)^2] - \frac{\varepsilon^2}{8} \geq \frac{\mathbb{E}_\mu[(g-\gamma)^2]}{2},
\end{aligned}$$

where the last inequality is due to $\mathbb{E}_\mu[(g-\gamma)^2] \geq \mathrm{Var}_\mu(g) \geq \varepsilon^2$. We can then obtain the desired result by taking infimum over $A \subset \mathbb{R}^d$ with $\mu(A) > \frac{1}{2}$ and $g \in c^+_{\varepsilon,M_0}(A)$ in (17).

C.2 Proof of Lemma 11

The proof follows from the standard conductance argument in Chewi et al. (2021); Belloni and Chernozhukov (2009); Dwivedi et al. (2019); Chen et al. (2020). Let $s = \frac{\varepsilon^2}{16M_0^2}$, and let S be any measurable set of $\mathbb{R}^d$ with $s \leq \mu(S) \leq v \leq \frac{1}{2}$. Define the following subsets:

$$\begin{aligned}
S_1 &:= \{x \in S | T(x, S^c) \leq \frac{\omega}{2}\}, \\
S_2 &:= \{x \in S^c | T(x, S) \leq \frac{\omega}{2}\}, \\
S_3 &:= (S_1 \cup S_2)^c,
\end{aligned}$$

Then same as the analysis in Chewi et al. (2021), if $\mu(S_1)\le \mu(S)/2$ or $\mu(S_2)<\mu(S^c)/2$, then by the fact that μ is stationary w.r.t the transition kernel T, we have

$$\begin{aligned}\int_S T(x,S^c)\mu(\mathrm{d}x) &= \int T(x,S)\mu(\mathrm{d}x) - \int_S T(x,S)\mu(\mathrm{d}x)\\ &= \int_{S^c} T(x,S)\mu(\mathrm{d}x) \ge \frac{\omega}{2}\cdot \max\{\mu(S\cap S_1^c),\mu(S^c\cap S_2^c)\}\\ &\ge \frac{\omega\cdot\mu(S)}{4}.\end{aligned}$$

Then when $\mu(S_1)\wedge\mu(S_2)\ge \frac{\mu(S)}{2}$, consider $x\in E\cap S_1$ and $z\in E\cap S_2$, then $\|T_x-T_z\|_{\mathrm{TV}}\ge T(z,S^c)-T(x,S^c)\ge 1-\omega$, thus $\|x-z\|\ge\psi$, which implies that $\inf_{x\in E\cap S_1, z\in E\cap S_2}\|x-z\|\ge\psi$. Then consider sets $E\cap K\cap S_1$ and $E\cap K\cap S_2$ in the log-isoperimetric inequality of $\mu|_K$, we can obtain that

$$\begin{aligned}\mu|_K(((E\cap K\cap S_1)\cup(E\cap K\cap S_2))^c) &\ge \lambda\cdot\psi\cdot\min\{\mu|_K(E\cap K\cap S_1),\mu|_K(E\cap K\cap S_2)\}\\ &\quad\cdot\log^{\frac{1}{2}}\Big(1+\frac{1}{\min\{\mu|_K(E\cap K\cap S_1),\mu|_K(E\cap K\cap S_2)\}}\Big)\\ &\ge \lambda\cdot\psi\cdot\min\{\mu(E\cap K\cap S_1),\mu(E\cap K\cap S_2)\}\\ &\quad\cdot\log^{\frac{1}{2}}\Big(1+\frac{1}{\min\{\mu(E\cap K\cap S_1),\mu(E\cap K\cap S_2)\}}\Big),\end{aligned}$$

where the last inequality is due to the fact that the function $x\log^{\frac{1}{2}}(1+\frac{1}{x})$ is an increasing function. W.l.o.g, we can assume $\mu(E\cap K\cap S_1)\le\mu(E\cap K\cap S_2)$, then by $((E\cap K\cap S_1)\cup(E\cap K\cap S_2))^c\subseteq E^c\cup K^c\cup S_3$ and $\mu(E^c)\le(\lambda\psi\wedge 1)\frac{\varepsilon^2}{256M_0^2}=\frac{(\lambda\psi\wedge 1)s}{16}$, $\mu(K^c)\le\frac{(\lambda\psi\wedge1)s}{16}$, we can obtain

$$\begin{aligned}&\mu(S_3)+\frac{16\lambda\psi s}{127}\\ &\ge \mu(S_3)+\frac{\mu(K^c)+\mu(E^c)}{\mu(K)}\\ &\ge\frac{\mu(S_3)+\mu(E^c)}{\mu(K)}\\ &\ge \mu|_K(((E\cap K\cap S_1)\cup(E\cap K\cap S_2))^c)\\ &\ge \lambda\cdot\psi\cdot\mu(E\cap K\cap S_1)\cdot\log^{\frac{1}{2}}\Big(1+\frac{1}{\mu(E\cap K\cap S_1)}\Big),\\ &\overset{\text{(i)}}{\ge} \lambda\cdot\psi\cdot(\frac{\mu(S)}{4}+\frac{s}{4}-\frac{s}{8})\log^{\frac{1}{2}}\Big(1+\frac{1}{\frac{\mu(S)}{4}+\frac{s}{4}-\frac{s}{8}}\Big)\\ &\ge\lambda\cdot\psi\cdot\frac{\mu(S)}{4}\log^{\frac{1}{2}}\Big(1+\frac{4}{\mu(S)}\Big),\end{aligned}$$

where (i) uses $\mu(E\cap K\cap S_1)\ge\mu(S_1)-\mu(E^c)-\mu(K^c)$, $\mu(S_1)\ge\frac{\mu(S)}{2}\ge\frac{s}{2}$ and the function $x\log^{\frac{1}{2}}(1+\frac{1}{x})$ is an increasing function. Then by $\mu(S)\ge s$, we can obtain

$$\mu(S_3)\ge\lambda\cdot\psi\cdot\frac{\mu(S)}{9}\log^{\frac{1}{2}}\Big(1+\frac{4}{\mu(S)}\Big),$$

hence

$$\int_S T(x,S^c)\mu(\mathrm{d}x) \geq \frac{1}{2}\Big(\int_S T(x,S^c)\mu(\mathrm{d}x) + \int_{S^c} T(x,S)\mu(\mathrm{d}x)\Big)$$
$$\geq \frac{\omega}{4}\mu(S_3) \geq \frac{\omega\cdot\lambda\cdot\psi}{36}\cdot\mu(S)\log^{\frac{1}{2}}\Big(1+\frac{4}{\mu(S)}\Big),$$

which leads to

$$\frac{\int_S T(x,S^c)\mu(\mathrm{d}x)}{\mu(S)} \geq \frac{\omega\cdot\lambda\cdot\psi}{36}\cdot\log^{\frac{1}{2}}\Big(1+\frac{4}{\mu(S)}\Big) \geq \frac{\omega\cdot\lambda\cdot\psi}{36}\cdot\log^{\frac{1}{2}}\Big(1+\frac{1}{v}\Big).$$

Then combining with the result for the first case, we can obtain a lower bound of

$$\frac{\omega}{4}\ \min\Big\{1, \frac{\lambda\cdot\psi}{9}\sqrt{\log\big(1+\frac{1}{v}\big)}\Big\}$$

on s-conductance profile $\Phi_s(v)$ with $s = \frac{\varepsilon^2}{16M_0^2}$.

C.3 Proof of Lemma 13

Recall the transition kernel associated with μ_k,

$$T(\theta,\mathrm{d}y) = \Big[1-(1-\zeta)\cdot\int A(\theta,y)Q(\theta,y)\,\mathrm{d}y\Big]\,\delta_\theta(\mathrm{d}y) + (1-\zeta)\cdot Q(\theta,y)A(\theta,y)\,\mathrm{d}y$$

with

$$A(\theta,y) = 1\wedge\frac{\pi_n(y)Q(y,\theta)}{\pi_n(\theta)Q(\theta,y)};\quad Q(\theta,\cdot) = N_d\Big(\theta-\frac{h}{\sqrt{n}}\widetilde{I}\widetilde{\nabla}V_n\big(\sqrt{n}(\theta-\widehat{\theta})\big), \frac{2h}{n}\,\widetilde{I}\Big).$$

Then given $\xi\in\mathbb{R}^d$, the distribution of $G_{\#}T(\xi,\cdot)$ is

$$\begin{aligned} &T^*(\theta,\mathrm{d}z)\\ &= \Big[1-(1-\zeta)\cdot\int Q^*(\theta,z)A\big(\theta,\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{z}{\sqrt{n}}\big)\,\mathrm{d}z\Big]\,\delta_{\sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})}\big(\mathrm{d}z\big)\\ &+(1-\zeta)\cdot Q^*(\theta,z)A\big(\theta,\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{z}{\sqrt{n}}\big)\,\mathrm{d}z, \end{aligned}$$

where $Q^*(\theta,\cdot)$ is the density function of $N_d(\sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})-h\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n\big(\sqrt{n}(\theta-\widehat{\theta})\big),2hI_d\big)$. Then by the fact that

$$\frac{Q\big(\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{z}{\sqrt{n}},\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{\xi}{\sqrt{n}}\big)}{Q\big(\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{\xi}{\sqrt{n}},\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{z}{\sqrt{n}}\big)} = \exp\Big(-\frac{1}{4h}\Big(\|\xi-z+h\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n(\widetilde{I}^{\frac{1}{2}}z)\|^2-\|z-\xi+h\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n(\widetilde{I}^{\frac{1}{2}}\xi)\|^2\Big)\Big)$$
$$=\frac{\widetilde{Q}(z,\xi)}{\widetilde{Q}(\xi,z)},$$

we have

$$T^*\Big(\widehat{\theta}+\widetilde{I}^{\frac{1}{2}}\frac{\xi}{\sqrt{n}},\mathrm{d}z\Big) = \Big[1-(1-\zeta)\cdot\int\widetilde{A}(\xi,z)\widetilde{Q}(\xi,z)\,\mathrm{d}z\Big]\,\mathbf{1}_\xi(\mathrm{d}z)+(1-\zeta)\cdot\widetilde{Q}(\xi,z)\widetilde{A}(\xi,z)\mathrm{d}z = \widetilde{T}(\xi,\mathrm{d}z).$$

Thus when $\widetilde{\mu}_{k-1} = G_{\#}\mu_{k-1}$, we have $\widetilde{\mu}_k = G_{\#}\mu_k$. Then combine with the fact that $\widetilde{\mu}_0 = G_{\#}\mu_0$, we can obtain by induction that $\widetilde{\mu}_k = G_{\#}\mu_k$ for $k\in\mathbb{N}$.

C.4 Proof of Lemma 14

To begin with, we consider the following lemma stated in Chen et al. (2020).

Lemma 20 *(Lemma 16 of Chen et al. (2020)) Let γ denote the density of the standard Gaussian distribution $\mathcal{N}\left(0, \sigma^2 I_d\right)$, and let μ be a distribution with density $\mu = q \cdot \gamma$, where q is a log-concave function. Then for any partition S_1, S_2, S_3 of $\mathbb{R}^d$, we have*

$$\mu\left(S_3\right) \geq \frac{d\left(S_1, S_2\right)}{2\sigma} \min\left\{\mu\left(S_1\right), \mu\left(S_2\right)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\mu\left(S_1\right), \mu\left(S_2\right)\right\}}\right).$$

We first consider the case $\widetilde{J} = I_d$ where recall $\widetilde{J} = \widetilde{I}^{\frac{1}{2}} J \widetilde{I}^{\frac{1}{2}}$. Then define $\overline{\pi} = N(0, I_d)|_{\widetilde{K}}$, by the fact that $\widetilde{K} = B^d_{R/2}$ is a convex set and $\mathbf{1}_{\widetilde{K}}$ is a log-concave function, using lemma 20, we can obtain that for any partition S_1, S_2, S_3 of $\widetilde{K}$, we have

$$\overline{\pi}\left(S_3\right) \geq \frac{d\left(S_1, S_2\right)}{2} \min\left\{\overline{\pi}\left(S_1\right), \overline{\pi}\left(S_2\right)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\overline{\pi}\left(S_1\right), \overline{\pi}\left(S_2\right)\right\}}\right).$$

Then recall $\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(\xi) = \frac{\mathbf{1}_{\widetilde{K}} \exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))}{\int_{\widetilde{K}} \exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi)) \mathrm{d}\xi}$, using the fact that $\sup_{\widetilde{\xi} \in B^d_R} \left| V_n(\widetilde{I}^{\frac{1}{2}}\widetilde{\xi}) - \frac{1}{2}\widetilde{\xi}^T \widetilde{J}\widetilde{\xi} \right| \leq \widetilde{\varepsilon}_0$, we can obtain that for any measurable set $S \subseteq \widetilde{K}$, we have

$$\exp(-2\widetilde{\varepsilon}_0) \leq \frac{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S)}{\overline{\pi}(S)} = \frac{\int_{S\cap\widetilde{K}} \exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\mathrm{d}\xi \int_K \exp(-\frac{1}{2}\xi^T\xi)\mathrm{d}\xi}{\int_{S\cap K} \exp(-\frac{1}{2}\xi^T\xi)\mathrm{d}\xi \int_K \exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\mathrm{d}\xi} \leq \exp(2\widetilde{\varepsilon}_0).$$

Thus

$$\begin{aligned}
&\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S_3) \geq \exp(-2\widetilde{\varepsilon}_0)\overline{\pi}(S_3) \\
&\geq \frac{d(S_1, S_2)}{2} \exp(-2\widetilde{\varepsilon}_0) \min\left\{\overline{\pi}\left(S_1\right), \overline{\pi}\left(S_2\right)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\overline{\pi}\left(S_1\right), \overline{\pi}\left(S_2\right)\right\}}\right) \\
&\overset{(i)}{\geq} \frac{d(S_1, S_2)}{2} \exp(-4\widetilde{\varepsilon}_0) \min\left\{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_1\right), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_2\right)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\exp(-2\widetilde{\varepsilon}_0)\min\left\{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_1\right), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_2\right)\right\}}\right) \\
&\geq \frac{d(S_1, S_2)}{2} \exp(-4\widetilde{\varepsilon}_0) \min\left\{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_1\right), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_2\right)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_1\right), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}\left(S_2\right)\right\}}\right),
\end{aligned} \tag{18}$$

where (i) uses the fact that $x\log^{\frac{1}{2}}(1+\frac{1}{x})$ is an increasing function. For the general case where $\widetilde{J}$ is not necessary an identity matrix, we can define $K' = \widetilde{J}^{\frac{1}{2}}\widetilde{K} = \{x = \widetilde{J}^{\frac{1}{2}}y : y \in \widetilde{K}\}$, and $\lambda = \widetilde{J}^{\frac{1}{2}}\xi$, where ξ is a random variable with density $\pi_{\text{loc}}|_{\widetilde{K}}$. Thus λ has a density

$$\pi_\lambda(\lambda) = \frac{\mathbf{1}_{K'}(\lambda)\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\widetilde{J}^{-\frac{1}{2}}\lambda))}{\int_{K'} \exp(-V_n(\widetilde{I}^{\frac{1}{2}}J^{-\frac{1}{2}}\lambda))\mathrm{d}\lambda},$$

Moreover, for any $\lambda \in K'$, it holds that

$$\left|V_n(\widetilde{I}^{\frac{1}{2}}\widetilde{J}^{-\frac{1}{2}}\lambda) - \frac{1}{2}\lambda^T\lambda\right| \leq \widetilde{\varepsilon}_0.$$

Then for any partition S_1, S_2, S_3 of $\widetilde{K}$, let

$$\begin{aligned}\widetilde{S_1} &= \widetilde{J}^{\frac{1}{2}} S_1;\\ \widetilde{S_2} &= \widetilde{J}^{\frac{1}{2}} S_2;\\ \widetilde{S_3} &= \widetilde{J}^{\frac{1}{2}} S_3.\end{aligned}$$

Then by the positive definiteness of $\widetilde{J}$, $(\widetilde{S_1}, \widetilde{S_2}, \widetilde{S_3})$ forms a partition for K', and

$$d(\widetilde{S_1}, \widetilde{S_2}) \geq \sqrt{\rho_1}\, d(S_1, S_2).$$

Since K' is a convex set, by applying π_λ to statement (18), we can obtain

$$\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S_3) = \pi_\lambda(\widetilde{S}_3) \geq \frac{d(\widetilde{S}_1, \widetilde{S}_2)}{2} \exp(-4\widetilde{\varepsilon}_0) \min\left\{\pi_\lambda(\widetilde{S}_1), \pi_\lambda(\widetilde{S}_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\pi_\lambda(\widetilde{S}_1), \pi_\lambda(\widetilde{S}_2)\right\}}\right)$$

$$\geq \frac{\sqrt{\rho_1}}{2} d(S_1, S_2) \exp(-4\widetilde{\varepsilon}_0) \min\left\{\pi_{\text{loc}}|_{\widetilde{K}}(S_1), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S_1), \widetilde{\pi}_{\text{loc}}|_{\widetilde{K}}(S_2)\right\}}\right).$$

Proof is completed.

C.5 Proof of Lemma 15

We first construct the high probability set E as follows: let

$$r_d = \left(\sqrt{c' d\left(\log\left(\frac{M_0^2}{\varepsilon^2 h \rho_1}\right) + \widetilde{\varepsilon}_0\right)\rho_2^2}\right) \vee \left(c'\left(\log\left(\frac{M_0^2}{\varepsilon^2 h \rho_1}\right) + \widetilde{\varepsilon}_0\right)\rho_2^2\right),$$

and $\widetilde{J} = \widetilde{I}^{\frac{1}{2}} J \widetilde{I}^{\frac{1}{2}}$. We define $E = \{\xi \in B^d_{R/2} : \left|\xi^T \widetilde{J}^3 \xi - \operatorname{tr}(\widetilde{J}^2)\right| \leq r_d\} \cap \{\xi \in B^d_{R/2} : \left|\xi^T \widetilde{J}^2 \xi - \operatorname{tr}(\widetilde{J})\right| \leq r_d/\rho_2\}$. By the choice of h, when c_0 is small enough, it holds that

$$h \leq \sqrt{c_0} \cdot \left\{\left(\rho_2{}^{-\frac{1}{3}}(\rho_2^2 d + r_d)^{-\frac{1}{3}}\right) \wedge (r_d)^{-\frac{1}{2}}\right\}.$$

Now we show that E is indeed a high probability set in the following lemma. Note that all the following lemmas in this subsection are under Assumptions in Theorem 3.

Lemma 21 *Consider* $E = \{\xi \in B^d_{R/2} : \left|\xi^T \widetilde{J}^3 \xi - \operatorname{tr}(\widetilde{J}^2)\right| \leq r_d\} \cap \{\xi \in B^d_{R/2} : \left|\xi^T \widetilde{J}^2 \xi - \operatorname{tr}(\widetilde{J})\right| \leq r_d/\rho_2\}$*. If* $r_d = \left(\sqrt{c' d \log\left(\frac{M_0^2}{\varepsilon^2 h \rho_1}\right)\rho_2^2}\right) \vee \left(c' \log\left(\frac{M_0^2}{\varepsilon^2 h \rho_1}\right)\rho_2^2\right)$ *for a sufficiently large enough constant* c'*, then* $\widetilde{\pi}_{\text{loc}}(E) \geq 1 - \exp(-4\widetilde{\varepsilon}_0) \cdot \frac{2\varepsilon^2 h \rho_1}{M_0^2}$*.*

We now show that for any $x, z \in E$ with $\|x - z\| \leq \frac{\sqrt{h}}{3}$, the total variation distance between $\widetilde{T}_x = \widetilde{T}(x, \cdot)$ and $\widetilde{T}_z = \widetilde{T}(z, \cdot)$ can be upper bounded by $1 - \frac{\exp(-2\widetilde{\varepsilon}_0)}{4}$. For any $x, z \in E$, we consider the following decomposition:

$$\begin{aligned}
&\|\widetilde{T}_x - \widetilde{T}_z\|_{TV} \\
&= \frac{1}{2}\int |\widetilde{T}(x,y) - \widetilde{T}(z,y)|\,\mathrm{d}y \\
&= \frac{1}{2}\widetilde{T}_x(\{x\}) + \frac{1}{2}\widetilde{T}_z(\{z\}) + \frac{1}{2}\int_{\mathbb{R}^d\setminus\{x,z\}} |\widetilde{T}(x,y) - \widetilde{T}(z,y)|\,\mathrm{d}y \\
&= \frac{1}{2} - \frac{1-\zeta}{2}\int_{\mathbb{R}^d} \widetilde{Q}(x,y)\widetilde{A}(x,y)\,\mathrm{d}y + \frac{1}{2} - \frac{1-\zeta}{2}\int_{\mathbb{R}^d} \widetilde{Q}(z,y)\widetilde{A}(z,y)\,\mathrm{d}y \\
&\quad + \frac{1-\zeta}{2}\int_{\mathbb{R}^d} |\widetilde{Q}(x,y)\widetilde{A}(x,y) - \widetilde{Q}(z,y)\widetilde{A}(z,y)|\,\mathrm{d}y \\
&= 1 - (1-\zeta)\int_{\mathbb{R}^d} \min\left(\widetilde{A}(x,y)\widetilde{Q}(x,y), \widetilde{A}(z,y)\widetilde{Q}(z,y)\right)\,\mathrm{d}y \\
&\le 1 - (1-\zeta)\int_{B_R^d} \min\left(\widetilde{A}(x,y)\widetilde{Q}(x,y), \widetilde{A}(z,y)\widetilde{Q}(z,y)\right)\,\mathrm{d}y
\end{aligned}$$

Recall that

$$\widetilde{A}(x,y) = 1 \wedge \frac{\widetilde{\pi}_{\mathrm{loc}}(y)\widetilde{Q}(y,x)}{\widetilde{\pi}_{\mathrm{loc}}(x)\widetilde{Q}(x,y)},$$

where $\widetilde{\pi}_{\mathrm{loc}}(x) \propto \exp(-V_n(\widetilde{I}^{\frac{1}{2}}x))$ and

$$\sup_{x\in B_R^d} \left|V_n(\widetilde{I}^{\frac{1}{2}}x) - \frac{1}{2}x^T\widetilde{J}x\right| \le \widetilde{\varepsilon}_0.$$

Define $\overline{\pi}$ as the density function of $N_d(0,\widetilde{J}^{-1})$, we have

$$\frac{\widetilde{\pi}_{\mathrm{loc}}(y)}{\widetilde{\pi}_{\mathrm{loc}}(x)} = \frac{\exp(-V_n(\widetilde{I}^{\frac{1}{2}}y))}{\exp(-V_n(\widetilde{I}^{\frac{1}{2}}y))} \ge \exp(-2\widetilde{\varepsilon}_0)\cdot\frac{\exp(-\frac{1}{2}y^T\widetilde{J}y)}{\exp(-\frac{1}{2}x^T\widetilde{J}x)} = \exp(-2\widetilde{\varepsilon}_0)\cdot\frac{\overline{\pi}(y)}{\overline{\pi}(x)}.$$

Therefore, denote

$$\overline{A}(x,y) = 1 \wedge \frac{\overline{\pi}(y)\widetilde{Q}(y,x)}{\overline{\pi}(x)\widetilde{Q}(x,y)},$$

we have

$$\widetilde{A}(x,y) \ge 1 \wedge \frac{\exp(-2\widetilde{\varepsilon}_0)\cdot\overline{\pi}(y)\widetilde{Q}(y,x)}{\overline{\pi}(x)\widetilde{Q}(x,y)} \ge \exp(-2\widetilde{\varepsilon}_0)\cdot\overline{A}(x,y).$$

We can then derive

$$\begin{aligned}
&\|\widetilde{T}_x - \widetilde{T}_z\|_{TV} \\
&\le 1 - (1-\zeta)\int_{B_R^d} \min\left(\widetilde{A}(x,y)\widetilde{Q}(x,y), \widetilde{A}(z,y)\widetilde{Q}(z,y)\right)\,\mathrm{d}y \\
&\le 1 - (1-\zeta)\exp(-2\widetilde{\varepsilon}_0)\cdot\int_{B_R^d} \min\left(\overline{A}(x,y)\widetilde{Q}(x,y), \overline{A}(z,y)\widetilde{Q}(z,y)\right)\,\mathrm{d}y \\
&= 1 - \frac{1}{2}(1-\zeta)\exp(-2\widetilde{\varepsilon}_0)\cdot\Bigg(\int_{B_R^d} \overline{A}(x,y)\widetilde{Q}(x,y)\,\mathrm{d}y + \int_{B_R^d} \overline{A}(z,y)\widetilde{Q}(z,y)\,\mathrm{d}y \\
&\quad - \int_{B_R^d} \left|\overline{A}(x,y)\widetilde{Q}(x,y) - \widetilde{A}(z,y)\widetilde{Q}(z,y)\right|\,\mathrm{d}y\Bigg)
\end{aligned} \tag{19}$$

Then consider the inequality:

$$\begin{aligned}\int_{B_R^d} |\widetilde{Q}(x,y)\overline{A}(x,y) - \widetilde{Q}(z,y)\overline{A}(z,y)|\,\mathrm{d}y \le &\int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y \\ &+ \int_{B_R^d} \widetilde{Q}(z,y)(1-\overline{A}(z,y))\,\mathrm{d}y + 2\|\widetilde{Q}_x - \widetilde{Q}_z\|_{\mathrm{TV}},\end{aligned}$$

where we use $\widetilde{Q}_x$ to denote the probability measure with density function $\widetilde{Q}(x,\cdot)$. Moreover, consider the equation:

$$\begin{aligned}\int_{B_R^d} \overline{A}(x,y)\widetilde{Q}(x,y)\,\mathrm{d}y &= \int_{B_R^d} (\overline{A}(x,y)-1)\widetilde{Q}(x,y)\,\mathrm{d}y + \int_{B_R^d} \widetilde{Q}(x,y)\,\mathrm{d}y \\ &= 1 - \int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y - \int_{(B_R^d)^c} \widetilde{Q}(x,y)\,\mathrm{d}y.\end{aligned}$$

Combined with (20), we can obtain

$$\begin{aligned}&\|\widetilde{T}_x - \widetilde{T}_z\|_{TV} \\ &\le 1 - (1-\zeta)\exp(-2\widetilde{\varepsilon}_0)\cdot\Bigg(1 - \int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y - \int_{B_R^d} \widetilde{Q}(z,y)(1-\overline{A}(z,y))\,\mathrm{d}y \\ &\qquad - \|\widetilde{Q}_x - \widetilde{Q}_z\|_{\mathrm{TV}} - \frac{1}{2}\int_{(B_R^d)^c} \widetilde{Q}(x,y)\,\mathrm{d}y - \frac{1}{2}\int_{(B_R^d)^c} \widetilde{Q}(z,y)\,\mathrm{d}y\Bigg)\end{aligned} \tag{20}$$

Consider the proposal distribution of MALA for sampling from the Gaussian $\overline{\pi} := N_d(0, \widetilde{J}^{-1})$,

$$Q_x^{\Delta}(\cdot) = N_d(x - h\widetilde{J}x, 2hI_d),$$

whose density is denoted as $Q^{\Delta}(x,\cdot)$. Then $\|\widetilde{Q}_x - \widetilde{Q}_z\|_{\mathrm{TV}} \le \|\widetilde{Q}_x - Q_x^{\Delta}\|_{\mathrm{TV}} + \|Q_x^{\Delta} - Q_z^{\Delta}\|_{\mathrm{TV}} + \|\widetilde{Q}_z - Q_z^{\Delta}\|_{\mathrm{TV}}$ can be upper bounded by Pinsker's inequality, that is, for any $x \in B_R^d$,

$$\|\widetilde{Q}_x - Q_x^{\Delta}\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\frac{h^2\|\widetilde{I}^{\frac{1}{2}}\widetilde{\nabla}V_n(\widetilde{I}^{\frac{1}{2}}x) - \widetilde{J}x\|^2}{2h}} \le \frac{\sqrt{h}\widetilde{\varepsilon}_1 |||\widetilde{I}^{\frac{1}{2}}|||_{\mathrm{op}}}{2\sqrt{2}},$$

and for any $x, z \in B_R^d$

$$\|Q_x^{\Delta} - Q_z^{\Delta}\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\frac{\|(I - h\widetilde{J})(x-z)\|^2}{2h}} \le \frac{\|x-z\|}{2\sqrt{2h}}.$$

Therefore, when $\|x - z\| \le \frac{\sqrt{h}}{3}$ and $\sqrt{h}\widetilde{\varepsilon}_1|||\widetilde{I}^{\frac{1}{2}}|||_{\mathrm{op}} \le \frac{\sqrt{2}}{36}$, we have

$$\|\widetilde{Q}_x - \widetilde{Q}_z\|_{\mathrm{TV}} \le \frac{\|x-z\|}{2\sqrt{2h}} + \frac{\sqrt{h}\widetilde{\varepsilon}_1|||\widetilde{I}^{\frac{1}{2}}|||_{\mathrm{op}}}{\sqrt{2}} < \frac{1}{6}.$$

For the term of $\int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y$, we use Condition A by comparing Q_x with Q_x^Δ, leading to the following decomposition:

$$\begin{aligned}
&\int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y\\
&\leq \int_{B_R^d} \Big|\widetilde{Q}(x,y) - \frac{\overline{\pi}(y)\widetilde{Q}(y,x)}{\overline{\pi}(x)}\Big|\,\mathrm{d}y\\
&\leq 2\|\widetilde{Q}_x - Q_x^\Delta\|_{\mathrm{TV}} + \underbrace{\int \left|Q^\Delta(x,y) - \frac{\overline{\pi}(y)Q^\Delta(y,x)}{\overline{\pi}(x)}\right|\mathrm{d}y}_{\text{(A)}} + \underbrace{\int_{B_R^d}\left|\frac{\overline{\pi}(y)Q^\Delta(y,x)}{\overline{\pi}(x)} - \frac{\overline{\pi}(y)\widetilde{Q}(y,x)}{\overline{\pi}(x)}\right|\mathrm{d}y}_{\text{(B)}}.
\end{aligned}$$

We then state the following lemma for bounding the term (A).

Lemma 22 *Consider the choice of (rescaled) step size h in Theorem 3, then when c_0 is small enough and $x \in E$, it holds that*

$$\int \left|Q^\Delta(x,y) - \frac{\overline{\pi}(y)Q^\Delta(y,x)}{\overline{\pi}(x)}\right|\mathrm{d}y \leq \frac{1}{24}.$$

Our proof of Lemma 22 is technically similar to that of Proposition 38 in Chewi et al. (2021) for bounding the mixing time of MALA with a standard Gaussian target (i.e. $\overline{\pi} = N_d(0, I_d)$). The non-trivial part in our analysis lies in keeping track of the dependence on the maximal and minimal eigenvalues of J. We then bound the term (B) by the following lemma.

Lemma 23 *Consider the choice of (rescaled) step size h in Theorem 3, then when c_0 is small enough, for any $x \in E$, it holds that*

$$\int_{B_R^d} \left|Q^\Delta(y,x) - Q(y,x)\right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)}\,\mathrm{d}y \leq \frac{1}{72}.$$

Thus when $\|x-z\| \leq \frac{\sqrt{h}}{3}$ and $\sqrt{h}\widetilde{\varepsilon}_1 \|\!|\widetilde{I}^{\frac{1}{2}}\|\!|_{\mathrm{op}} \leq \frac{\sqrt{2}}{36}$,

$$\begin{aligned}
&\int_{B_R^d} \widetilde{Q}(x,y)(1-\overline{A}(x,y))\,\mathrm{d}y\\
&\leq 2\|\widetilde{Q}_x - Q_x^\Delta\|_{\mathrm{TV}} + \frac{1}{24} + \frac{1}{72}\\
&\leq \sqrt{\frac{h}{2}}\widetilde{\varepsilon}_1 \|\!|\widetilde{I}^{\frac{1}{2}}\|\!|_{\mathrm{op}} + \frac{1}{18} \leq \frac{1}{12}.
\end{aligned} \tag{21}$$

Finally, since for any $x \in E \subset B_{R/2}^d$,

$$\begin{aligned}
\int_{(B_R^d)^c} \widetilde{Q}(x,y)\,\mathrm{d}y &\leq \int_{(B_R^d)^c} Q^\Delta(x,y)\,\mathrm{d}y + 2\,\|\widetilde{Q}_x - Q_x^\Delta\|_{\mathrm{TV}}\\
&\leq \mathbb{E}_{u\in N_d(0,I_d)}\Big[\mathbf{1}\big(\|u\| \geq \frac{R}{2\sqrt{2h}}\big)\Big] + \frac{\sqrt{h}\widetilde{\varepsilon}_1 \|\!|\widetilde{I}^{\frac{1}{2}}\|\!|_{\mathrm{op}}}{\sqrt{2}}.
\end{aligned}$$

Since $R \geq 8\sqrt{d/\lambda_{\min}(\widetilde{J})}$, when the constant c_0 in h is small enough, we can obtain

$$\int_{(B_R^d)^c} \widetilde{Q}(x,y)\,\mathrm{d}y \leq \frac{1}{6}.$$

Then combined with the bound in equation (21) and decomposition (20), we can obtain that when c_0 is small enough, for any $x, z \in E$ with $\|x - z\| < \frac{\sqrt{h}}{3}$ and $\zeta \in (0, \frac{1}{2}]$, it holds that

$$\begin{aligned}
&\|\widetilde{T}_x - \widetilde{T}_z\|_{TV} \\
&\leq 1 - (1-\zeta)\exp(-2\widetilde{\varepsilon}_0) \cdot \bigg(1 - \int_{B_R^d} \widetilde{Q}(x,y)(1 - \overline{A}(x,y))\,\mathrm{d}y - \int_{B_R^d} \widetilde{Q}(z,y)(1 - \overline{A}(z,y))\,\mathrm{d}y \\
&\qquad - \|\widetilde{Q}_x - \widetilde{Q}_z\|_{\mathrm{TV}} - \frac{1}{2}\int_{(B_R^d)^c} \widetilde{Q}(x,y)\,\mathrm{d}y - \frac{1}{2}\int_{(B_R^d)^c} \widetilde{Q}(z,y)\,\mathrm{d}y\bigg) \\
&\leq 1 - \frac{1-\zeta}{2}\exp(-2\widetilde{\varepsilon}_0) \\
&\leq 1 - \frac{\exp(-2\widetilde{\varepsilon}_0)}{4}.
\end{aligned}$$

C.6 Proof of Lemma 21

We can write $\widetilde{\pi}_{\mathrm{loc}}$ as

$$\widetilde{\pi}_{\mathrm{loc}}(\xi) = \frac{\frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))}{\int \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\,\mathrm{d}\xi}.$$

Then

$$\begin{aligned}
1 - \widetilde{\pi}_{\mathrm{loc}}(E) \leq{}& \frac{\int_{\left\{\xi \in B_{R/2}^d \,:\, |\xi^T \widetilde{J}^3 \xi - \mathrm{tr}(\widetilde{J}^2)| > r_d\right\}} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\,\mathrm{d}\xi}{\int \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\,\mathrm{d}\xi} \\
&+ \frac{\int_{\left\{\xi \in B_{R/2}^d \,:\, |\xi^T \widetilde{J}^2 \xi - \mathrm{tr}(\widetilde{J})| > r_d/\rho_2\right\}} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\,\mathrm{d}\xi}{\int \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}}\exp(-V_n(\widetilde{I}^{\frac{1}{2}}\xi))\,\mathrm{d}\xi} \\
&+ \widetilde{\pi}_{\mathrm{loc}}(\|\xi\| > R/2).
\end{aligned}$$

Then for the denominator, as

$$\sup_{\widetilde{\xi} \in B_R^d} \left|V_n(\widetilde{I}^{\frac{1}{2}}\widetilde{\xi}) - \frac{1}{2}\widetilde{\xi}^T \widetilde{I}^{\frac{1}{2}} J \widetilde{I}^{\frac{1}{2}}\widetilde{\xi}\right| \leq \widetilde{\varepsilon}_0,$$

when $R \geq 8(\frac{d}{\lambda_{\min}(\widetilde{J})})^{\frac{1}{2}}$, we can obtain that

$$
\begin{aligned}
&\int \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}} \exp(V_n(\widetilde{I}^{-\frac{1}{2}}\xi)) \,\mathrm{d}\xi \\
&\geq \int_{B_R^d} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\xi^T \widetilde{J}\xi}{2}) \exp(\frac{\xi^T \widetilde{J}\xi}{2} - V_n(I^{\frac{1}{2}}\xi)) \,\mathrm{d}\xi \\
&\geq \exp(-\widetilde{\varepsilon}_0) \int_{B_R^d} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\xi^T \widetilde{J}\xi}{2}) \,\mathrm{d}\xi \\
&\geq \frac{1}{2} \exp(-\widetilde{\varepsilon}_0).
\end{aligned}
$$

Furthermore, by Bernstein's inequality (see for example, Theorem 2.8.2 of Vershynin (2018)), for $x \sim N_d(0, \Sigma)$, it holds that

$$
\mathbb{P}(\big|\|x\|^2 - \mathrm{tr}(\Sigma)\big| \geq t) \leq 2\exp(-\frac{1}{8}(\frac{t^2}{\|\!|\Sigma\|\!|_{\mathrm{F}}^2} \wedge \frac{t}{\|\!|\Sigma\|\!|_{\mathrm{op}}})) \tag{22}
$$

We can then obtain

$$
\begin{aligned}
\pi_{\mathrm{loc}}(E) \geq 1 - 2\exp(2\widetilde{\varepsilon}_0) &\int_{\left\{|\xi^T \widetilde{J}^3\xi - \mathrm{tr}(\widetilde{J}^2)| > r_d\right\}} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\xi^T \widetilde{J}\xi}{2}) \,\mathrm{d}\xi \\
&- 2\exp(2\widetilde{\varepsilon}_0) \int_{\left\{|\xi^T \widetilde{J}^2\xi - \mathrm{tr}(\widetilde{J})| > r_d/\rho_2\right\}} \frac{\sqrt{\det(\widetilde{J})}}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\xi^T \widetilde{J}\xi}{2}) \,\mathrm{d}\xi - \frac{\varepsilon^2 h\rho_1}{M_0^2} \\
&\geq 1 - \exp(-4\widetilde{\varepsilon}_0) \cdot \frac{2\varepsilon^2 h\rho_1}{M_0^2},
\end{aligned}
$$

where the last inequality is due to the Bernstein's inequality in (22).

C.7 Proof of Lemma 22

Recall $\overline{\pi} = N_d(0, \widetilde{J}^{-1})$ and $Q^\Delta(x, \cdot)$ be the density of $N_d(x - h\widetilde{J}x, 2hI_d)$, we have

$$
\begin{aligned}
&\int \left| Q^\Delta(x,y) - \frac{\overline{\pi}(y) Q^\Delta(y,x)}{\overline{\pi}(x)} \right| \mathrm{d}y \\
&= \int \frac{1}{(4\pi h)^{\frac{d}{2}}} \left| \exp\left(-\frac{\|y - x + h\widetilde{J}x\|^2}{4h}\right) - \exp\left(\frac{x^T\widetilde{J}x - y^T\widetilde{J}y}{2}\right) \exp\left(-\frac{\|x - y + h\widetilde{J}y\|^2}{4h}\right) \right| \mathrm{d}y \\
&= \int \frac{1}{(4\pi h)^{\frac{d}{2}}} \exp\left(-\frac{\|y - x + h\widetilde{J}x\|^2}{4h}\right) \left| 1 - \exp\left(\frac{h^2\|\widetilde{J}x\|^2 - h^2\|\widetilde{J}y\|^2}{4h}\right) \right| \mathrm{d}y,
\end{aligned}
$$

let $u = \frac{y - x + h\widetilde{J}x}{\sqrt{2h}}$ in the above integral, then consider $u \sim N_d(0, I_d)$ and let

$$
\mathcal{A} = \left\{ u \in \mathbb{R}^d : \frac{1}{4} \left| 2h^2\|\widetilde{J}u\|^2 + 2\sqrt{2}h^{\frac{3}{2}} x^T\widetilde{J}^2u - 2\sqrt{2}h^{\frac{5}{2}} x^T\widetilde{J}^3u + h^3x^T\widetilde{J}^4x - 2h^2x^T\widetilde{J}^3x \right| \leq \frac{1}{49} \right\}.
$$

We can then obtain

$$
\begin{aligned}
&\int \left| Q^{\Delta}(x,y) - \frac{\overline{\pi}(y) Q^{\Delta}(y,x)}{\overline{\pi}(x)} \right| \mathrm{d}y \\
&= \mathbb{E}_u \left[\left| 1 - \exp\left(\frac{-h^2 \|\sqrt{2h}\widetilde{J}u + \widetilde{J}x - h\widetilde{J}^2 x\|^2 + h^2 \|\widetilde{J}x\|^2}{4h} \right) \right| \right] \\
&= \mathbb{E}_u \left[\left| 1 - \exp\left(-\frac{1}{4} \left(2h^2 \|\widetilde{J}u\|^2 + 2\sqrt{2} h^{\frac{3}{2}} x^T \widetilde{J}^2 u - 2\sqrt{2} h^{\frac{5}{2}} x^T \widetilde{J}^3 u + h^3 x^T \widetilde{J}^4 x - 2h^2 x^T \widetilde{J}^3 x \right) \right) \right| \right] \\
&\leq \left\{ \mathbb{E}_u \left[\left| 1 - \exp\Big(-\frac{1}{4} \Big(2h^2 \|\widetilde{J}u\|^2 + 2\sqrt{2} h^{\frac{3}{2}} x^T \widetilde{J}^2 u - 2\sqrt{2} h^{\frac{5}{2}} x^T \widetilde{J}^3 u + h^3 x^T \widetilde{J}^4 x - 2h^2 x^T \widetilde{J}^3 x \Big) \Big) \right| \right. \right. \\
&\quad \left. \left. \cdot \mathbf{1}_{\mathcal{A}}(u) \right] \right\} + \left\{ \mathbb{E}_u \left[\mathbf{1}_{\mathcal{A}^c}(u) \right] \right\} + \left\{ \exp\left(-\frac{1}{4} h^3 x^T \widetilde{J}^4 x \right) \sqrt{\mathbb{E}_u \left[\mathbf{1}_{\mathcal{A}^c}(u) \right]} \right. \\
&\quad \left. \cdot \Big(\mathbb{E}_u \Big[\exp(-3h^2 (u^T \widetilde{J}^2 u - x^T \widetilde{J}^3 x)) \Big] \cdot \mathbb{E}_u \Big[\exp(3\sqrt{2} h^{\frac{3}{2}} x^T \widetilde{J}^2 u) \Big] \cdot \mathbb{E}_u \Big[\exp(3\sqrt{2} h^{\frac{5}{2}} x^T \widetilde{J}^3 u) \Big] \Big)^{\frac{1}{6}} \right\},
\end{aligned}
\tag{23}
$$

where the last inequality uses Hölder inequality. The first term of the right hand side of equation (23) can be upper bound by $\exp(1/49) - 1 \leq 1/48$. For the second and third term, by (1) $h \leq \sqrt{c_0} \rho_2^{-\frac{1}{3}} (\mathrm{tr}(\widetilde{J}^2) + r_d)^{-\frac{1}{3}}$ and $h \leq \sqrt{c_0} r_d^{-\frac{1}{2}}$ with $r_d = \Big\{ \Big(\sqrt{c' \log \frac{M_0^2}{\varepsilon^2 h \rho_1}} \|\!|\widetilde{J}^2|\!\|_{\mathrm{F}} \Big) \vee \Big(c' \log \frac{M_0^2}{\varepsilon^2 h \rho_1} \rho_2^2 \Big) \Big\} \wedge (\rho_2^3 \|K\|^2)$ and $\|K\| \geq C(\frac{d}{\rho_1})^{\frac{1}{2}}$; (2) $x \in E = \{x \in K : \left| x^T \widetilde{J}^3 x - \mathrm{tr}(\widetilde{J}^2) \right| \leq r_d\}$, it holds that

$$
h^3 x^T \widetilde{J}^4 x \leq h^3 \rho_2 x^T \widetilde{J}^3 x \leq h^3 \rho_2 (r_d + \mathrm{tr}(\widetilde{J}^2)) \leq c_0^{\frac{3}{2}}.
$$

Moreover, since for a Gaussian random variable $\bar{u} \sim N(0, \sigma^2)$, it holds that

$$
\begin{aligned}
&\mathbb{E} \exp(t\bar{u}) = \exp(\frac{\sigma^2 t^2}{2}) \\
&\mathbb{E} \exp(-t^2 \bar{u}^2) = \frac{1}{\sqrt{1 + 2t^2 \sigma^2}} \quad |t| < \sqrt{\frac{1}{2\sigma^2}}.
\end{aligned}
$$

We can get

$$
\begin{aligned}
&\mathbb{E}_u \Big[\exp(t^2 h^2 (x^T \widetilde{J}^3 x - \|\widetilde{J}u\|^2)) \Big] \\
&\leq \exp(t^2 h^2 (x^T \widetilde{J}^3 x - \mathrm{tr}(\widetilde{J}^2))) \prod_{j=1}^{d} \frac{1/\sqrt{1 + 2t^2 h^2 \lambda_j(\widetilde{J}^2)}}{\exp\big(- t^2 h^2 \lambda_j(\widetilde{J}^2) \big)} \\
&\leq \exp(t^2 h^2 r_d) \cdot \prod_{j=1}^{d} \big(1 + C\, t^4 h^4 (\lambda_j(\widetilde{J}^2))^2 \big) \\
&\leq \exp(t^2 c_0) \exp(C t^4 h^4 \|\!|\widetilde{J}^2|\!\|_{\mathrm{F}}^2) \\
&\leq \exp(t^2 c_0 + t^4 C\, c_0^2), \quad |t| \leq \sqrt{\frac{1}{4h^2 \rho_2(\widetilde{J}^2)}},
\end{aligned}
$$

where the last inequality uses $h \le \sqrt{c_0}\rho_2^{-\frac{1}{3}}(\mathrm{tr}(\widetilde{J}^2)+r_d)^{-\frac{1}{3}} \le \sqrt{c_0}\rho_2^{-\frac{1}{3}}(\mathrm{tr}(\widetilde{J}^2))^{-\frac{1}{3}} \le \sqrt{c_0}\|\widetilde{J}^2\|_F^{-\frac{1}{2}}$, and

$$\mathbb{E}_u\left[\exp(th^{\frac{3}{2}}x^T\widetilde{J}^2u)\right] \le \exp\left(\frac{1}{2}t^2h^3\|x^T\widetilde{J}^2\|^2\right) \le \exp\left(\frac{1}{2}t^2h^3\rho_2(\mathrm{tr}(\widetilde{J}^2)+r_d)\right) \le \exp\left(\frac{1}{2}c_0^{\frac{3}{2}}t^2\right);$$

$$\mathbb{E}_u\left[\exp(th^{\frac{5}{2}}x^T\widetilde{J}^3u)\right] \le \exp\left(\frac{1}{2}t^2h^5\|x^T\widetilde{J}^3\|^2\right) \le \exp\left(\frac{1}{2}t^2h^5\rho_2^3(\mathrm{tr}(\widetilde{J}^2)+r_d)\right) \le \exp(\frac{1}{2}c_0^{\frac{5}{2}}t^2),$$

where the last inequality uses $h \le \sqrt{c_0}\rho_2^{-\frac{1}{3}}(\mathrm{tr}(\widetilde{J}^2)+r_d)^{-\frac{1}{3}} \le \sqrt{c_0}\rho_2^{-1}$. Then by Markov inequality, we can obtain that

$$\mathbb{P}_u\left(|h^{\frac{3}{2}}x^T\widetilde{J}^2u| \ge \frac{1}{96\sqrt{2}}\right) \le 2\inf_{t>0}\exp\left(\frac{1}{2}c_0^{\frac{3}{2}}t^2 - \frac{t}{96\sqrt{2}}\right) = 2\exp\left(-\frac{1}{2\cdot(96\sqrt{2})^2c_0^{\frac{3}{2}}}\right);$$

$$\mathbb{P}_u\left(|h^{\frac{5}{2}}x^T\widetilde{J}^3u| \ge \frac{1}{96\sqrt{2}}\right) \le 2\inf_{t>0}\exp\left(\frac{1}{2}c_0^{\frac{5}{2}}t^2 - \frac{t}{96\sqrt{2}}\right) = 2\exp\left(-\frac{1}{2\cdot(96\sqrt{2})^2c_0^{\frac{5}{2}}}\right).$$

Also, by Bernstein's inequality in (22), we have

$$\begin{aligned}\mathbb{P}_u\left(h^2\left|\|\widetilde{J}u\|^2 - x^T\widetilde{J}^3x\right| \ge \frac{1}{96}\right) &\le P_u\left(\left|\|\widetilde{J}u\|^2 - \mathrm{tr}(\widetilde{J}^2)\right| \ge \frac{1}{96h^2} - r_d\right)\\ &\le P_u\left(\left|\|\widetilde{J}u\|^2 - \mathrm{tr}(\widetilde{J}^2)\right| \ge \frac{1}{h^2}(\frac{1}{96} - c_0)\right)\\ &\le 2\exp\left(-\frac{1}{c'}\left(\frac{\frac{1}{96}-c_0}{h^2\rho_2^2} \wedge \frac{(\frac{1}{96}-c_0)^2}{h^4\|\widetilde{J}^2\|_F^2}\right)\right)\\ &\le 2\exp\left(-\frac{1}{c'}\left(\frac{\frac{1}{96}-c_0}{c_0} \wedge \frac{(\frac{1}{96}-c_0)^2}{c_0^2}\right)\right),\end{aligned}$$

where the last inequality uses $h \le \sqrt{c_0}\|\widetilde{J}^2\|_F^{-\frac{1}{2}}$. Therefore, when c_0 is small enough, we have

$$\begin{aligned}&\mathbb{E}_u\left[\mathbf{1}_{\mathcal{A}^c}(u)\right]\\ &\le \mathbb{P}_u\Big(h^2|\|\widetilde{J}u\|^2 - x^T\widetilde{J}^3x| \ge \frac{1}{96}\Big) + \mathbb{P}_u\Big(|h^{\frac{3}{2}}x^T\widetilde{J}^2u| \ge \frac{1}{96\sqrt{2}}\Big) + \mathbb{P}_u\Big(|h^{\frac{5}{2}}x^T\widetilde{J}^3u| \ge \frac{1}{96\sqrt{2}}\Big)\\ &\le 2\exp\Big(-\frac{\frac{1}{96}-c_0}{c'c_0}\Big) + 2\exp\Big(-\frac{1}{2\cdot(96\sqrt{2})^2c_0^{\frac{3}{2}}}\Big) + 2\exp\Big(-\frac{1}{2\cdot(96\sqrt{2})^2c_0^{\frac{5}{2}}}\Big)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_u\left[\mathbf{1}_{\mathcal{A}^c}(u)\right] + \exp\left(-\frac{1}{4}h^3x^T\widetilde{J}^4x\right)\sqrt{\mathbb{E}_u\left[\mathbf{1}_{\mathcal{A}^c}(u)\right]}\\ &\cdot\left(\mathbb{E}_u\left[\exp(-3h^2(\|\widetilde{J}u\|^2 - x^T\widetilde{J}^3x))\right]\cdot\mathbb{E}_u\left[\exp(3\sqrt{2}h^{\frac{3}{2}}x^T\widetilde{J}^2u)\right]\cdot\mathbb{E}_u\left[\exp(3\sqrt{2}h^{\frac{5}{2}}x^T\widetilde{J}^3u)\right]\right)^{\frac{1}{6}}\\ &\le \frac{1}{48}.\end{aligned}$$

We can then obtain the desired result by combining all pieces.

C.8 Proof of Lemma 23

We first write

$$
\begin{aligned}
&\int_{B_R^d} \left| Q^\Delta(y,x) - \widetilde{Q}(y,x) \right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)} \, \mathrm{d}y \\
&= \int_{B_R^d} \left| 1 - \frac{\widetilde{Q}(y,x)}{Q^\Delta(y,x)} \right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x) \, \mathrm{d}y \\
&= \int_{B_R^d} \left| 1 - \exp\left(\frac{-\|x - y + h\widetilde{I}^{\frac{1}{2}} \widetilde{\nabla} V_n(\widetilde{I}^{\frac{1}{2}} y)\|^2 + \|x - y + h\widetilde{J} y\|^2}{4h} \right) \right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x) \, \mathrm{d}y.
\end{aligned}
$$

Since $h \leq \sqrt{c_0} \rho_2^{-\frac{1}{3}} (\mathrm{tr}(\widetilde{J}^2) + r_d)^{-\frac{1}{3}} \leq \sqrt{c_0} \rho_2^{-1}$ and $h\rho_2 \|\!|\widetilde{I}|\!\|_{\mathrm{op}} R^2 \widetilde{\varepsilon}_1^2 \leq c_0$, when c_0 is sufficiently small, we have for any $x \in E$ and $y \in B_R^d$,

$$
\begin{aligned}
&\frac{\left| -\|x - y + h\widetilde{\nabla} \widetilde{I}^{\frac{1}{2}} \widetilde{\nabla} V_n(\widetilde{I}^{\frac{1}{2}} y)\|^2 + \|x - y + h\widetilde{J} y\|^2 \right|}{4h} \\
&= \frac{\left| h(\widetilde{J} y + \widetilde{I}^{\frac{1}{2}} \widetilde{\nabla} V_n(\widetilde{I}^{\frac{1}{2}} y))^T (\widetilde{J} y - \widetilde{I}^{\frac{1}{2}} \widetilde{\nabla} V_n(\widetilde{I}^{\frac{1}{2}} y)) + 2(x-y)^T (\widetilde{J} y - \widetilde{I}^{\frac{1}{2}} \widetilde{\nabla} V_n(\widetilde{I}^{\frac{1}{2}} y)) \right|}{4} \\
&\leq \frac{h\big(2\rho_2 R + \|\!|\widetilde{I}^{\frac{1}{2}}|\!\|_{\mathrm{op}} \widetilde{\varepsilon}_1\big) \|\!|\widetilde{I}^{\frac{1}{2}}|\!\|_{\mathrm{op}} \widetilde{\varepsilon}_1 + 2\|x - y\| \|\!|\widetilde{I}^{\frac{1}{2}}|\!\|_{\mathrm{op}} \widetilde{\varepsilon}_1}{4} \\
&\leq \frac{\sqrt{c_0}}{4} \big(3 + \frac{2\|x - y\|}{R\sqrt{h\rho_2}}\big).
\end{aligned}
$$

Thus we can bound

$$
\int_{B_R^d} \left| Q^\Delta(y,x) - \widetilde{Q}(y,x) \right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)} \, \mathrm{d}y \leq \int_{B_R^d} \Big(\exp\Big(\frac{\sqrt{c_0}}{4} \big(3 + \frac{2\|x - y\|}{R\sqrt{h\rho_2}}\big) \Big) - 1 \Big) \frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x) \, \mathrm{d}y.
$$

Furthermore, by Lemma 22, we can get

$$
\int \frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x) \, \mathrm{d}y \leq \int Q^\Delta(x,y) \, \mathrm{d}y + \int \left| Q^\Delta(x,y) - \frac{\overline{\pi}(y) Q^\Delta(y,x)}{\overline{\pi}(x)} \right| \mathrm{d}y \leq \frac{25}{24},
$$

which leads to

$$
\begin{aligned}
&\int_{B_R^d} \left| Q^\Delta(y,x) - \widetilde{Q}(y,x) \right| \frac{\overline{\pi}(y)}{\overline{\pi}(x)} \, \mathrm{d}y \\
&\leq \frac{25}{24} \int_{B_R^d} \Big(\exp\Big(\frac{\sqrt{c_0}}{4} \big(3 + \frac{2\|x - y\|}{R\sqrt{h\rho_2}}\big) \Big) - 1 \Big) \frac{\frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x)}{\int \frac{\overline{\pi}(y)}{\overline{\pi}(x)} Q^\Delta(y,x) \, \mathrm{d}y} \, \mathrm{d}y \\
&= \frac{25}{24} \int_{B_R^d} \Big(\exp\Big(\frac{\sqrt{c_0}}{4} \big(3 + \frac{2\|x - y\|}{R\sqrt{h\rho_2}}\big) \Big) - 1 \Big) N_d\big((I + h^2 \widetilde{J})^{-1}(x - h\widetilde{J} x), 2h(I + h^2 \widetilde{J})^{-1}\big) \, \mathrm{d}y,
\end{aligned}
$$

where the last inequality is due to $\frac{\overline{\pi}(y)}{\overline{\pi}(x)}Q^{\Delta}(y,x) \propto \exp(-\frac{y^T(I+h^2J)y-2y^T(x-h\widetilde{J}x)}{4h})$. Consider $u \sim N_d(0, I_d)$, for sufficiently small c_0, we have

$$\begin{aligned}
&\int_{B_R^d} \Big(\exp\Big(\frac{\sqrt{c_0}}{4}\big(3 + \frac{2\|x-y\|}{R\sqrt{h\rho_2}}\big)\Big) - 1\Big) N_d\big((I+h^2\widetilde{J})^{-1}(x-h\widetilde{J}x), 2h(I+h^2\widetilde{J})^{-1}\big)\,\mathrm{d}y \\
&\le \mathbb{E}_{u\sim N_d(0,I_d)}\Big[\exp\Big(\frac{\sqrt{c_0}}{4}\Big(3 + \frac{2\|(I+h^2\widetilde{J})^{-1}(x-h\widetilde{J}x) - x + \sqrt{2h}(I+h^2\widetilde{J})^{-\frac{1}{2}}u\|}{R\sqrt{h\rho_2}}\Big)\Big) - 1\Big] \\
&\overset{(i)}{\le} \mathbb{E}_{u\sim N_d(0,I_d)}\Big[\exp\Big(\frac{\sqrt{c_0}}{4}\Big(3 + 2c_0^{\frac{1}{4}} + \frac{2\sqrt{2}}{R\sqrt{\rho_2}}\|u\|\Big)\Big) - 1\Big] \\
&\le \frac{81}{80}\cdot \mathbb{E}_{u\sim N_d(0,I_d)}\Big[\exp\Big(\frac{\sqrt{2c_0}}{2R\sqrt{\rho_2}}\|u\|\Big)\Big] - 1 \\
&\le \frac{81}{80}\cdot \sqrt{\mathbb{E}_{u\sim N_d(0,I_d)}\Big[\exp\Big(\frac{c_0}{2R^2\rho_2}\|u\|^2\Big)\Big]} - 1 \\
&\le \frac{81}{80}\exp(\frac{c_0 d}{2R^2\rho_2}) - 1 \\
&\overset{(ii)}{\le} \frac{1}{75},
\end{aligned}$$

where (i) is due to $\|(I+h^2\widetilde{J})^{-1}(x-h\widetilde{J}x) - x\| \le h^2\rho_2\|x\| + h\rho_2\|x\| \le 2\sqrt{h\rho_2}c_0^{\frac{1}{4}}\|x\| \le \sqrt{h\rho_2}c_0^{\frac{1}{4}}R$, and (ii) is due to $R \ge 8\sqrt{d/\lambda_{\min}(\widetilde{J})}$. Thus we can obtain $\int_{B_R^d}\Big|Q^{\Delta}(y,x) - \widetilde{Q}(y,x)\Big| \frac{\overline{\pi}(y)}{\overline{\pi}(x)}\,\mathrm{d}y \le \frac{1}{72}$.

C.9 Proof of Lemma 17

C.9.1 Proof of statement (1) of Lemma 17

Define the following compact supported function $k: \mathbb{R} \to \mathbb{R}$:

$$k(t) = \begin{cases} 2(t-t^3) & t \in (-1,1), \\ 0 & \text{otherwise.} \end{cases}$$

Then consider a initial distribution with density function $\mu_0(x) = (1 + k(\sqrt{\rho_1}x_d))\cdot\overline{\pi}(x)$. This constriction guarantees that

$$\begin{aligned}
\chi^2(\mu_0, \overline{\pi}) &= \sqrt{\frac{\rho_1}{2\pi}}\int_{-\sqrt{\frac{1}{\rho_1}}}^{\sqrt{\frac{1}{\rho_1}}} k^2(\sqrt{\rho_1}x_d)\exp(-\frac{\rho_1}{2}x_d^2)\,\mathrm{d}x_d \\
&= \sqrt{\frac{1}{2\pi}}\int_{-1}^{1} k^2(t)\exp(-\frac{1}{2}t^2)\,\mathrm{d}t \in (0.2, 0.21),
\end{aligned}$$

$$\sup_{x\in\mathbb{R}^d}\frac{\mu_0(x)}{\overline{\pi}(x)} = 1 + \sup_{t\in(-1,1)} k(t) < 2,$$

and

$$|h_0(x) - h_0(y)| = \big|\frac{\mu_0(x)}{\overline{\pi}(x)} - \frac{\mu_0(y)}{\overline{\pi}(y)}\big| = |k(\sqrt{\rho_1}x_d) - k(\sqrt{\rho_1}y_d)| \le 2\sqrt{\rho_1}|x_d - y_d|.$$

Therefore, the spectral gap of this initialization is controlled by

$$\begin{aligned}\frac{\mathcal{E}(h_0,h_0)}{\chi^2(\mu_0',\overline{\pi})} &\leq 10\rho_1\cdot\mathbb{E}_{x\in\overline{\pi},y\in T(x,\cdot)}\big[(x_d-y_d)^2\big]\\ &\leq 10\rho_1\cdot\mathbb{E}_{x\in\overline{\pi},y\in N_d(x-h\rho_1 x,2hI_d)}\big[(x_d-y_d)^2\big]\\ &= 10\rho_1\cdot\mathbb{E}_{x_d\in N(0,1/\rho_1),\xi\in N(0,1)}\big[(h\rho_1x_d-\sqrt{2h}\xi)^2\big]\\ &\leq 20m^2h^2+40mh\leq 60mh.\end{aligned}$$

C.9.2 Proof of Statement (2) of Lemma 17

Denote sets

$$\begin{aligned}K_2&=\Big\{x\in\mathbb{R}^d\,:\,\big|x^TJ^3x-\mathrm{tr}(J^2)\big|\leq(5|||J^2|||_{\mathrm{F}})\vee(24|||J^2|||_{\mathrm{op}})\Big\};\\ K_3&=\Big\{x\in\mathbb{R}^d\,:\,\big|x^TJ^4x-\mathrm{tr}(J^3)\big|\leq(5|||J^3|||_{\mathrm{F}})\vee(24|||J^3|||_{\mathrm{op}})\Big\};\\ K_4&=\Big\{x\in\mathbb{R}^d\,:\,\big|x^TJ^6x-\mathrm{tr}(J^5)\big|\leq(5|||J^5|||_{\mathrm{F}})\vee(24|||J^5|||_{\mathrm{op}})\Big\},\end{aligned}$$

To control the probability of the above events, we utilize the following Bernstein's inequality: for $x\in N_d(0,\Sigma)$,

$$\mathbb{P}(\big|\|x\|^2-\mathrm{tr}(\Sigma)\big|\geq t)\leq 2\exp\Big(-\frac{1}{8}\big(\frac{t^2}{|||\Sigma|||_{\mathrm{F}}^2}\wedge\frac{t}{|||\Sigma|||_{\mathrm{op}}}\big)\Big),\tag{24}$$

which leads to

$$\mathbb{P}\left(\big|\|x\|^2-\mathrm{tr}(\Sigma)\big|\geq\big(\sqrt{8\lambda}|||\Sigma|||_F\big)\vee\big(8\lambda|||\Sigma|||_{\mathrm{op}}\big)\right)\leq 2\exp(-\lambda).$$

Therefore, for $x\sim N_d(0,J^{-1})$, the probability of events $x\in K_2\cap K_3\cap K_4$ is inside the interval of $(0.7,1)$. Then let $K_1\subset\mathbb{R}^d$ be an arbitrary measurable set so that the probability of events $x\in K=K_1\cap K_2\cap K_3\cap K_4$ is equal $\frac{1}{M_0}$ (notice that $M_0\geq 2$ and $\frac{1}{M_0}\leq\frac{1}{2}<0.7$, therefore such a set K_1 exists). Then consider a initial distribution with density function $\mu_0'(x)=\frac{\overline{\pi}(x)\mathbf{1}_K(x)}{\mathbb{E}_{\overline{\pi}}[\mathbf{1}_K(x)]}$, it holds that

$$\chi^2(\mu_0',\overline{\pi})=\frac{1}{\mathbb{E}_{\overline{\pi}}[\mathbf{1}_K(x)]}-1=M_0-1,$$

and

$$\sup_{x\in\mathbb{R}^d}\frac{\mu_0(x)}{\overline{\pi}(x)}=\frac{1}{\mathbb{E}_{\overline{\pi}}[\mathbf{1}_K(x)]}=M_0.$$

Then denote for bounding the spectral gap $\frac{\mathcal{E}(h_0,h_0)}{\chi^2(\mu_0',\overline{\pi})}$ with $h_0=\frac{\mathrm{d}\mu_0'}{\mathrm{d}\overline{\pi}}$, we claim it suffices to show the following claim: denote $Q^{\Delta}(x,)$ to be the density function of $N_d(x-hJx,2hI_d)$, then for any $x\in K$, there exists a set $G_x\subset\mathbb{R}^d$ so that

$$\frac{\overline{\pi}(y)Q^{\Delta}(y,x)}{\overline{\pi}(x)Q^{\Delta}(x,y)}\leq\exp(-16\log(\kappa d)),\quad\forall\,y\in G_x,\tag{25}$$

and

$$\int_{G_x} Q^{\Delta}(x,y)\,\mathrm{d}y \geq 1 - \frac{3}{\kappa d}. \tag{26}$$

Indeed, under claim (25) and (26), we have

$$\begin{aligned}
\frac{\mathcal{E}(h_0,h_0)}{\chi^2(\mu_0',\overline{\pi})} &= \frac{M_0^2 \cdot \mathbb{E}_{x\in\overline{\pi},y\in T(x,\cdot)}\big[(\mathbf{1}_K(x)-\mathbf{1}_K(y))^2\big]}{2(M_0-1)} \\
&\leq \frac{M_0^2}{M_0-1}\int_K\int_{K^c}\min\Big\{1,\frac{\overline{\pi}(y)Q^{\Delta}(y,x)}{\overline{\pi}(x)Q^{\Delta}(x,y)}\Big\}\overline{\pi}(x)Q^{\Delta}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&\leq \frac{M_0^2}{M_0-1}\int_{x\in K}\Big(\int_{G_x}\frac{\overline{\pi}(y)Q^{\Delta}(y,x)}{\overline{\pi}(x)Q^{\Delta}(x,y)}Q^{\Delta}(x,y)\,\mathrm{d}y + \int_{G_x^c} Q^{\Delta}(x,y)\,\mathrm{d}y\Big)\overline{\pi}(x)\,\mathrm{d}x \\
&\leq \frac{M_0}{M_0-1}\sup_{x\in K}\Big(\sup_{y\in G_x}\frac{\overline{\pi}(y)Q^{\Delta}(y,x)}{\overline{\pi}(x)Q^{\Delta}(x,y)} + \int_{G_x^c} Q^{\Delta}(x,y)\Big)\,\mathrm{d}y \\
&\leq \frac{8}{\kappa d},
\end{aligned}$$

where the last inequality uses claim (25) and (26). Now we show the desired claim. First note that

$$\frac{\overline{\pi}(y)Q^{\Delta}(y,x)}{\overline{\pi}(x)Q^{\Delta}(x,y)} = \exp\Big(\frac{h\cdot(x^TJ^2x - y^TJ^2y)}{4}\Big).$$

Let $u = \frac{y-x+hJx}{\sqrt{2h}}$, then for $y \in N_d(x-hJx, 2hI_d)$, we have $u \in N_d(0, I_d)$. Therefore, it suffice to show that for any $x \in K$, there exists a set $G_x' \in \mathbb{R}^d$ so that $\mathbb{E}_{N_d(0,I_d)}[\mathbf{1}_{G_x'}(u)] \geq 1 - \frac{3}{\kappa d}$ and

$$\frac{\overline{\pi}(\sqrt{2h}u + x - hJx)Q^{\Delta}(\sqrt{2h}u + x - hJx, x)}{\overline{\pi}(x)Q^{\Delta}(x, \sqrt{2h}u + x - hJx)} \leq \exp(-16\log(\kappa d)), \quad \forall\, u \in G_x'.$$

Denote the sets

$$\begin{aligned}
\mathcal{G}_x^1 &= \{u\in\mathbb{R}^d \,:\, \|Ju\|^2 - x^TJ^3x \geq -\big(\sqrt{8\log(\kappa d)}+5\big)\rho_2^2\sqrt{d}\}; \\
\mathcal{G}_x^2 &= \{u\in\mathbb{R}^d \,:\, x^TJ^2u \geq -\sqrt{\log(\kappa d)\cdot\big(10\rho_2^3\sqrt{d}+2\rho_2^3 d\big)}\}; \\
\mathcal{G}_x^3 &= \{u\in\mathbb{R}^d \,:\, x^TJ^3u \leq \sqrt{\log(\kappa d)\cdot\big(10\rho_2^5\sqrt{d}+2\rho_2^5 d\big)}\}.
\end{aligned}$$

Then under $G_x' = \mathcal{G}_x^1\cap\mathcal{G}_x^2\cap\mathcal{G}_x^3$, we have

$$\begin{aligned}
&\frac{\overline{\pi}(\sqrt{2h}u + x - hJx)Q^{\Delta}(\sqrt{2h}u + x - hJx, x)}{\overline{\pi}(x)Q^{\Delta}(x, \sqrt{2h}u + x - hJx)} \\
&= \exp\Big(-\frac{1}{4}\Big(2h^2\|Ju\|^2 + 2\sqrt{2}h^{\frac{3}{2}}x^TJ^2u - 2\sqrt{2}h^{\frac{5}{2}}x^TJ^3u + h^3x^TJ^4x - 2h^2x^TJ^3x\Big)\Big) \\
&\leq \exp\Big(-\frac{1}{4}\Big(h^3x^TJ^4x - 2h^2\big(\sqrt{8\log(\kappa d)}+5\big)\rho_2^2\sqrt{d} \\
&\qquad - 2\sqrt{2}\sqrt{\log(\kappa d)}\big(h^{\frac{3}{2}}\sqrt{10\rho_2^3\sqrt{d}+2\rho_2^3 d} + h^{\frac{5}{2}}\sqrt{10\rho_2^5\sqrt{d}+2\rho_2^5 d}\big)\Big)\Big).
\end{aligned}$$

Then there exists a universal constant N_1 so that when $d \geq N_1$, for any $x \in K$ and $u \in G'_x$,

$$\begin{aligned}
&\frac{\overline{\pi}(\sqrt{2h}u + x - hJx)Q^{\Delta}(\sqrt{2h}u + x - hJx, x)}{\overline{\pi}(x)Q^{\Delta}(x, \sqrt{2h}u + x - hJx)} \\
&\leq \exp\Big(- \frac{1}{4}\big(h^3 x^T J^4 x - 6h^2\sqrt{\log(\kappa d)}\rho_2^2\sqrt{d} - 5\sqrt{\log(\kappa d)}\big(h^{\frac{3}{2}}\sqrt{\rho_2^3 d} + h^{\frac{5}{2}}\sqrt{\rho_2^5 d}\big)\big)\Big) \\
&\leq \exp\Big(- \frac{1}{4}\big(h^3 x^T J^4 x - 6h^2\sqrt{\log(\kappa d)}\rho_2^2\sqrt{d} - 5\sqrt{\log(\kappa d)}\big(h^{\frac{3}{2}}\sqrt{\rho_2^3 d} + h^{\frac{5}{2}}\sqrt{\rho_2^5 d}\big)\big)\Big) \\
&\leq \exp\big(- 32\log(\kappa d) + 96\log^{\frac{7}{6}}(\kappa d)d^{-\frac{1}{6}} + 227\log^{\frac{4}{3}}(\kappa d)d^{-\frac{1}{3}}\big).
\end{aligned}$$

Therefore, use $\kappa \leq c_1 \cdot d^{c_2}$ there exists N_2 that depends only on c_1, c_2 so that when $d \geq N_2$, for any $x \in K$ and $u \in G'_x$

$$\frac{\overline{\pi}(\sqrt{2h}u + x - hJx)Q^{\Delta}(\sqrt{2h}u + x - hJx, x)}{\overline{\pi}(x)Q^{\Delta}(x, \sqrt{2h}u + x - hJx)} \leq \exp\big(- 16\log(\kappa d)\big).$$

Now we control the probability $u \in G'_x$. Firstly by Bernstein's inequality, for $u \in N_d(0, I_d)$, we have

$$\mathbb{P}\left(u^T J^2 u - \mathrm{tr}(J^2) \geq -\big((\sqrt{8\log(\kappa d)}|||J^2|||_{\mathrm{F}}) \vee (8\log(\kappa d)|||J^2|||_{\mathrm{op}})\big)\right) \leq 1 - \frac{1}{d\kappa}.$$

So there exists a universal constant N_3 so that when $d \geq N_3$, it holds with probability at least $1 - \frac{1}{d\kappa}$ that

$$\begin{aligned}
\|Ju\|^2 - x^T J^3 x &\geq \mathrm{tr}(J^2) - (\sqrt{8\log(\kappa d)}|||J^2|||_{\mathrm{F}}) \vee (8\log(\kappa d)|||J^2|||_{\mathrm{op}}) - \mathrm{tr}(J^2) - (5|||J^2|||_{\mathrm{F}}) \vee (24|||J^2|||_{\mathrm{op}}) \\
&\geq -\big(\sqrt{8\log(\kappa d)} + 5\big)\rho_2^2\sqrt{d}.
\end{aligned}$$

Moreover, since for any $t \in \mathbb{R}$ and $x \in K$,

$$\begin{aligned}
\mathbb{E}[\exp(t x^T J^2 u)] &= \exp\left(\frac{1}{2}t^2 x^T J^4 x\right) \\
&\leq \exp\left(\frac{1}{2}\Big(\mathrm{tr}(J^3) + \big((5|||J^3|||_{\mathrm{F}}) \vee (24|||J^3|||_{\mathrm{op}})\big)\Big)t^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\exp(t x^T J^3 u)] &= \exp\left(\frac{1}{2}t^2 x^T J^6 x\right) \\
&\leq \exp\left(\frac{1}{2}\Big(\mathrm{tr}(J^5) + \big((5|||J^5|||_{\mathrm{F}}) \vee (24|||J^5|||_{\mathrm{op}})\big)\Big)t^2\right),
\end{aligned}$$

by Markov inequality, there exists a universal constant N_4 so that when $d \geq N_4$, it holds with probability at least $1 - \frac{2}{\kappa d}$ that

$$x^T J^2 u \geq -\sqrt{\log(\kappa d) 2\rho_2^3 d + 10\rho_2^3\sqrt{d}}$$

and

$$x^T J^3 u \leq \sqrt{\log(\kappa d) 2\rho_2^5 d + 10\rho_2^5\sqrt{d}}.$$

We can then obtain the desired result by combining all pieces.

Appendix D. Proof of Lemmas for Theorem 5

D.1 Proof of Lemma 18

Without loss of generality, we can assume the learning rate $\alpha = 1$, as otherwise we can take $\ell(X, \theta) = \alpha \cdot \ell(X, \theta)$. To begin with, we provide in the following lemma some localized "maximal" type inequalities that control the supreme of empirical processes to deal with the non-smoothness of the loss function. All the following lemmas in this subsection are under Condition B.1-B.4.

Lemma 24 *There exist positive constants c and r such that it holds with probability larger than $1 - n^{-2}$ that ,*

1. *For any $\theta, \theta' \in B_r(\theta^*)$, $\Big\|\frac{1}{n}\sum_{i=1}^n g(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n g(X_i,\theta') - \mathbb{E}[g(X,\theta)] + \mathbb{E}[g(X,\theta')]\Big\| \leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}} \|\theta - \theta'\|^{\beta_1} + \frac{\log n}{n} d^{1+\gamma}\right)$.*
2. *For any $\theta, \theta' \in \Theta$, $\Big|\frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta') - \mathbb{E}[\ell(X,\theta)] + \mathbb{E}[\ell(X,\theta')]\Big| \leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1}{2}+\gamma} \|\theta - \theta'\| + \frac{\log n}{n} d^{\frac{3}{2}+\gamma}\right)$.*
3. *For any $\theta, \theta' \in B_r(\theta^*)$, $\Big|\frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta') - \frac{1}{n}\sum_{i=1}^n g(X_i,\theta')(\theta - \theta') - \mathbb{E}[\ell(X,\theta)] + \mathbb{E}[\ell(X,\theta')] + \mathbb{E}[g(X,\theta')(\theta-\theta')]\Big| \leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}} \|\theta - \theta'\|^{\beta_1+1} + \frac{\log n}{n} d^{1+\gamma}\|\theta-\theta'\| + (\frac{\log n}{n})^2\right)$.*

Recall $V_n(\xi) = n\left(\mathcal{R}_n(\widehat{\theta} + \frac{\xi}{\sqrt{n}}) - \mathcal{R}_n(\widehat{\theta})\right) + \log \pi(\widehat{\theta} + \frac{\xi}{\sqrt{n}}) - \log \pi(\widehat{\theta})$, in order to bound the difference between $V_n(\xi)$ and $\frac{\xi^T \mathcal{H}_{\theta^*} \xi}{2}$ using Lemma 24, we should first prove that $\widehat{\theta}$ is close to θ^*. Define a first order approximate to $\widehat{\theta}$: $\widehat{\theta}^\diamond = \theta^* - \frac{1}{n}\sum_{i=1}^n \mathcal{H}_{\theta^*}^{-1} g(X_i, \theta^*)$, we have the following lemma for bounding the difference between $\widehat{\theta}^\diamond$ and θ^*.

Lemma 25 *It holds with probability larger than $1 - n^{-2}$ that*

$$\|\widehat{\theta}^\diamond - \theta^*\| \leq C\, d^{\frac{1+\gamma_4}{2}} \sqrt{\frac{\log n}{n}} + C\, d^{1+\gamma_0+\gamma} \frac{\log n}{n}.$$

And we resort to the following lemma that provides an upper bound on the ℓ_2 distance between $\widehat{\theta}$ and $\widehat{\theta}^\diamond$.

Lemma 26 *There exists a small enough positive constant c such that when $d \leq c\Big((\frac{n}{\log n})^{\frac{1}{2+2(\gamma+\gamma_0+\gamma_1)}} \wedge (\frac{n}{\log n})^{\frac{1}{1+2\gamma+2\gamma_2+4\gamma_0}}\Big)$, then it holds with probability larger than $1 - c \cdot n^{-2}$ that*

$$\left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta})\right\| \leq C\, d^{1+\gamma} \frac{\log n}{n} + C\, d^{\frac{1+\gamma_3}{2}} (\frac{\log n}{n})^{\frac{1}{2}+\beta_1}; \tag{27}$$

$$\|\widehat{\theta}-\widehat{\theta}^{\diamond}\| \leq C\, d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})+\gamma_0}\big(\frac{\log n}{n}\big)^{\frac{1+\beta_1}{2}} + C\, d^{1+\gamma\vee(\gamma_2+\gamma_4)+\gamma_0}\frac{\log n}{n} + C\left(d^{\frac{1+\gamma_3}{2}+\gamma_0}\sqrt{\frac{\log n}{n}}\right)^{\frac{1}{1-\beta_1}}. \tag{28}$$

By $\sup_{x\in\mathcal{X}}\|g(X,\theta^*)\| \leq C\, d^{\gamma}$, we have $|||\mathcal{H}_{\theta^*}^{-1}\mathbb{E}[g(X,\theta^*)g(X,\theta^*)^T]\mathcal{H}_{\theta}^{-1}|||_{\mathrm{op}} \leq C_1\, d^{2\gamma}|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}^2 \leq C_2\, d^{2\gamma+2\gamma_0}$, which leads to $\gamma_4 \leq 2\gamma_0 + 2\gamma$. Then by Lemma 25 and Lemma 26, when

$$d \leq c\left(\big(\frac{n}{\log n}\big)^{\frac{\beta_1}{\gamma_3+\beta_1(1+\gamma_4)+2\gamma_0-\gamma_4}} \wedge \big(\frac{n}{\log n}\big)^{\frac{1}{1+2\gamma+2\gamma_2+4\gamma_0}} \wedge \big(\frac{n}{\log n}\big)^{\frac{1}{2+2(\gamma+\gamma_0+\gamma_1)}}\right),$$

it holds with probability larger than $1-c_1\cdot n^{-2}$ that

$$\|\widehat{\theta}-\theta^*\| \leq C\, d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}. \tag{29}$$

We can now derive (high probability) upper bound to the term of $|V_n(\xi)-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}|$ over $1\leq\|\xi\|\leq C\sqrt{n}$. Consider the following decomposition:

$$\begin{aligned}
&\Big|V_n(\xi)-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\Big| \\
&\leq \Big|n\big(\mathcal{R}_n(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}_n(\widehat{\theta})\big)-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\Big| + \Big|\log\pi(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\log\pi(\widehat{\theta})\Big| \\
&\leq n\,\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i,\widehat{\theta})\frac{\xi}{\sqrt{n}}\Big| + n\,\Big|\mathcal{R}_n(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}_n(\widehat{\theta})-\frac{1}{n}\sum_{i=1}^{n} g(X_i,\widehat{\theta})\frac{\xi}{\sqrt{n}}-\Big(\mathcal{R}(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}(\widehat{\theta}) \\
&-\mathbb{E}g(X,\widehat{\theta})\frac{\xi}{\sqrt{n}}\Big)\Big| + n\,\Big|\mathcal{R}(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}(\widehat{\theta})-\mathbb{E}[g(X,\widehat{\theta})]\frac{\xi}{\sqrt{n}}-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\Big| + C\sqrt{d}\cdot\frac{\|\xi\|}{\sqrt{n}}.
\end{aligned}$$

The first term can be bounded by Lemma 26, that is

$$\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i,\widehat{\theta})\frac{\xi}{\sqrt{n}}\Big| \leq C\,\frac{\|\xi\|}{\sqrt{n}}\left[d^{1+\gamma}\frac{\log n}{n} + d^{\frac{1+\gamma_3}{2}}\big(\frac{\log n}{n}\big)^{\frac{1}{2}+\beta_1}\right];$$

for the second term, by the third statement of Lemma 24, we can obtain that

$$\begin{aligned}
&\Big|\mathcal{R}_n(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}_n(\widehat{\theta})-\frac{1}{n}\sum_{i=1}^{n} g(X_i,\widehat{\theta})\frac{\xi}{\sqrt{n}}-\Big(\mathcal{R}(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}(\widehat{\theta})-\mathbb{E}g(X,\widehat{\theta})\frac{\xi}{\sqrt{n}}\Big)\Big| \\
&\leq C\left[d^{1+\gamma}\frac{\log n}{n}\frac{\|\xi\|}{\sqrt{n}} + \sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}\big(\frac{\|\xi\|}{\sqrt{n}}\big)^{1+\beta_1} + \big(\frac{\log n}{n}\big)^2\right];
\end{aligned}$$

for the third term, by the twice differentiability of $\mathcal{R}(\theta)$ and Lipschitzness of $\mathcal{H}_\theta$, we can obtain that

$$
\begin{aligned}
&\left|\mathcal{R}(\widehat{\theta}+\frac{\xi}{\sqrt{n}})-\mathcal{R}(\widehat{\theta})-\mathbb{E}[g(X,\widehat{\theta})]\frac{\xi}{\sqrt{n}}-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2n}\right| \\
&\leq \frac{\|\xi\|^2}{2n}\sup_{\xi\in K}\left\|\!\left|\mathcal{H}_{\widehat{\theta}+\frac{\xi}{\sqrt{n}}}-\mathcal{H}_{\theta^*}\right|\!\right\|_{\mathrm{op}} \\
&\leq C\,\frac{\|\xi\|^2}{n}d^{\gamma_2}\left(d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}+\frac{\|\xi\|}{\sqrt{n}}\right) \\
&= C\,\frac{\|\xi\|^3}{n^{\frac{3}{2}}}d^{\gamma_2}+C\,\frac{\|K\|^2}{n}\sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_4}{2}+\gamma_2}.
\end{aligned}
$$

Therefore, by combining all these result, when $1\leq\|\xi\|\leq c\sqrt{n}$ for a small enough c, we can obtain that

$$
\begin{aligned}
\left|V_n(\xi)-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\right| &\leq C\,d^{1+\gamma}\|\xi\|\frac{\log n}{\sqrt{n}}+C\,d^{\frac{1+\gamma_3}{2}}\|\xi\|^{1+\beta_1}n^{-\frac{\beta_1}{2}}\sqrt{\log n} \\
&\quad+C\,d^{\frac{1+\gamma_4}{2}+\gamma_2}\|\xi\|^2\sqrt{\frac{\log n}{n}}+C\,d^{\gamma_2}\|\xi\|^3n^{-\frac{1}{2}}.
\end{aligned}
$$

For the second statement, since when $1\leq\|\xi\|\leq c\sqrt{n}$ for a small enough c,

$$
\begin{aligned}
&\|\widetilde{\nabla}V_n(\xi)-\mathcal{H}_{\theta^*}\xi\| \\
&=\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\frac{\xi}{\sqrt{n}}+\widehat{\theta})-\frac{1}{\sqrt{n}}\nabla[\log\pi](\frac{\xi}{\sqrt{n}}+\widehat{\theta})-\mathcal{H}_{\theta^*}\xi\right\| \\
&\leq\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i,\widehat{\theta})\right\|+\sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^n g(X_i,\frac{\xi}{\sqrt{n}}+\widehat{\theta})-\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})-\mathbb{E}[g(X,\frac{\xi}{\sqrt{n}}+\widehat{\theta})]+\mathbb{E}[g(X,\widehat{\theta})]\right\| \\
&+\sqrt{n}\left\|\mathbb{E}[g(X,\frac{\xi}{\sqrt{n}}+\widehat{\theta})]-\mathbb{E}[g(X,\widehat{\theta})]-\mathcal{H}_{\theta^*}\xi\right\|+\left\|\frac{1}{\sqrt{n}}\nabla[\log\pi](\frac{\xi}{\sqrt{n}}+\widehat{\theta})\right\|.
\end{aligned}
$$

Then by the first statement of Lemma 24, Lemma 26, the twice-differentiability of $\mathcal{R}(\theta)$ and Lipschitz continuity of $\mathcal{H}_\theta$. Similar to analysis for the first statement, we can obtain that for any $1\leq\|\xi\|\leq c\sqrt{n}$,

$$
\begin{aligned}
&\|\widetilde{\nabla}V_n(\xi)-H_{\theta^*}\xi\| \\
&\leq C\,\sqrt{n}\left[d^{1+\gamma}\frac{\log n}{n}+d^{\frac{1+\gamma_3}{2}}(\frac{\log n}{n})^{\frac{1}{2}+\beta_1}\right]+C\,\left(d^{\frac{1+\gamma_3}{2}}\sqrt{\log n}(\frac{\|\xi\|}{\sqrt{n}})^{\beta_1}+d^{1+\gamma}\frac{\log n}{\sqrt{n}}\right) \\
&+C\,\left(d^{\gamma_2}\frac{\xi\|^2}{\sqrt{n}}+d^{\frac{1+\gamma_4}{2}+\gamma_2}\sqrt{\log n}\frac{\|\xi\|}{\sqrt{n}}\right)+C\,\sqrt{\frac{d}{n}} \\
&\leq C\,d^{1+\gamma}\frac{\log n}{\sqrt{n}}+C\,d^{\frac{1+\gamma_3}{2}}\|\xi\|^{\beta_1}n^{-\frac{\beta_1}{2}}\sqrt{\log n}+C\,d^{\frac{1+\gamma_4}{2}+\gamma_2}\|\xi\|\sqrt{\frac{\log n}{n}}+C\,d^{\gamma_2}\|\xi\|^2n^{-\frac{1}{2}}.
\end{aligned}
$$

D.2 Proof of Lemma 19

Without loss of generality, we can assume the learning rate $\alpha = 1$, as otherwise we can take $\ell(X,\theta) = \alpha \cdot \ell(X,\theta)$. Denote $K = \big\{\xi : \|\widetilde{I}^{-1/2}\xi\| \le \|\!|\widetilde{I}^{-\frac{1}{2}}\|\!|_{\mathrm{op}} \vee \frac{3(\sqrt{d}+t)}{\sqrt{\lambda_{\min}(\widetilde{J})}}\big\}$. Then

$$\pi_n(\sqrt{n}(\theta - \widehat{\theta}) \in K^c) = \frac{\int_{K^c} \exp(-V_n(\xi))\, \mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}} \det(\mathcal{H}_{\theta^*})}{\int \exp(-V_n(\xi))\, \mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}} \det(\mathcal{H}_{\theta^*})}$$

Denote $K_1 = K^c \cap \{\xi : \|\xi\| \le c_1 d^{-\gamma_0-\gamma_2}\sqrt{n}\}$ and $K_2 = K^c \cap \{\xi : \|\xi\| \ge c_1 d^{-\gamma_0-\gamma_2}\sqrt{n}\}$. When $\xi \in K_1$, we have $\|\xi\| \ge \frac{\|\!|\widetilde{I}^{-\frac{1}{2}}\|\!|_{\mathrm{op}}}{\|\!|\widetilde{I}^{-\frac{1}{2}}\|\!|_{\mathrm{op}}} = 1$. So by Lemma 18 and the fact that

$$\xi^T \mathcal{H}_{\theta^*}\xi = (\widetilde{I}^{-\frac{1}{2}}\xi)^T \widetilde{I}^{\frac{1}{2}} \mathcal{H}_{\theta^*} \widetilde{I}^{\frac{1}{2}} \widetilde{I}^{-\frac{1}{2}}\xi \ge \lambda_{\min}(\widetilde{J})\|\widetilde{I}^{-\frac{1}{2}}\xi\|^2 \ge 9(\sqrt{d}+t)^2;$$

and

$$\xi^T \mathcal{H}_{\theta^*}\xi \ge \lambda_{\min}(\mathcal{H}_{\theta^*})\|\xi\|^2 \ge d^{-\gamma_0}\|\xi\|^2,$$

we can verify that when $d \le c\frac{n^{\kappa_3}}{\log n}$ for small enough c and $K_1 = K^c \cap \{\xi : \|\xi\| \le c_1 d^{-\gamma_0-\gamma_2}\sqrt{n}\}$ for a small enough c_1, it holds that

$$V_n(\xi) \ge \frac{\xi^T \mathcal{H}_{\theta^*}\xi}{4}, \quad \xi \in K_1.$$

So we have

$$\begin{aligned}
&\int_{K_1} \exp(-V_n(\xi))\, \mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}} \det(\mathcal{H}_{\theta^*}) \\
&\le 2^{\frac{d}{2}} (2\pi)^{-\frac{d}{2}} \det\big(\frac{\mathcal{H}_{\theta^*}}{2}\big) \int_{K_1} \exp\big(-\frac{\xi^T \mathcal{H}_{\theta^*}\xi}{4}\big)\, \mathrm{d}\xi \\
&\le 2^{\frac{d}{2}} \cdot \mathbb{P}_{\chi^2(d)}(\|x\| \ge 4(\sqrt{d}+t)^2) \\
&\le \exp(-t^2 - \frac{1}{4}),
\end{aligned}$$

where the last inequality uses the tail inequality of χ^2 distribution with d degree of freedom (see for example, Lemma 1 of Laurent and Massart (2000)).

For $\xi \in K_2$ and $\theta = \widehat{\theta} + \frac{\xi}{\sqrt{n}}$, we have

$$\|\hat{\theta} - \theta\| \ge c_1 d^{-\gamma_0-\gamma_2}.$$

Moreover, by equation (29) which states that $\|\widehat{\theta} - \theta^*\| \lesssim d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}$, when $d \le c\frac{n^{\kappa_3}}{\log n}$ for small enough c, we have

$$\|\theta - \theta^*\| \ge \frac{c_1}{2} d^{-\gamma_0-\gamma_2}.$$

Therefore, by the second statement of Lemma 24, we can conclude

$$\mathcal{R}(\theta) - \mathcal{R}(\widehat{\theta}) = \mathcal{R}(\theta) - \mathcal{R}(\theta^*) + \mathcal{R}(\theta^*) - \mathcal{R}(\widehat{\theta}) \ge C d^{-\gamma_0}(d^{-\gamma_1} \wedge \|\theta - \theta^*\|^2) - C_1 d^{\gamma + \frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}};$$

and

$$|\mathcal{R}_n(\theta) - \mathcal{R}_n(\widehat{\theta}) - \mathcal{R}(\theta) + \mathcal{R}(\widehat{\theta})| \leq C_2\sqrt{\frac{\log n}{n}} d^{\frac{1}{2}+\gamma}\|\theta - \widehat{\theta}\| + C_2\frac{\log n}{n} d^{\frac{3}{2}+\gamma}.$$

Then if (1) $c_1 d^{-\gamma_0-\gamma_2}\sqrt{n} \leq \|\hat{\theta} - \theta\| \leq d^{-\frac{\gamma_1}{2}}$, we have

$$\begin{aligned}\mathcal{R}_n(\theta) - \mathcal{R}_n(\widehat{\theta}) &\geq \mathcal{R}(\theta) - \mathcal{R}(\widehat{\theta}) - |\mathcal{R}_n(\theta) - \mathcal{R}_n(\widehat{\theta}) - \mathcal{R}(\theta) + \mathcal{R}(\widehat{\theta})| \\ &\geq C\, c_1 d^{-3\gamma_0-2\gamma_2} - C_1 d^{\gamma+\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}} - C_2\sqrt{\frac{\log n}{n}} d^{\frac{1}{2}+\gamma-\frac{\gamma_1}{2}} - C_2\frac{\log n}{n} d^{\frac{3}{2}+\gamma} \\ &\geq \frac{C\, c_1}{2} d^{-3\gamma_0-2\gamma_2},\end{aligned}$$

where the last inequality uses $d \leq c\frac{n^{\kappa_3}}{\log n}$ for small enough c; when (2) $\|\hat{\theta} - \theta\| \geq d^{-\frac{\gamma_1}{2}}$, then by $\Theta \subset [-C, C]^d$, we can get

$$\begin{aligned}\mathcal{R}_n(\theta) - \mathcal{R}_n(\widehat{\theta}) &\geq \mathcal{R}(\theta) - \mathcal{R}(\widehat{\theta}) - |\mathcal{R}_n(\theta) - \mathcal{R}_n(\widehat{\theta}) - \mathcal{R}(\theta) + \mathcal{R}(\widehat{\theta})| \\ &\geq C\, c_1 d^{-\gamma_1-\gamma_0} - C_1 d^{\gamma+\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}} - C_2\sqrt{\frac{\log n}{n}} d^{1+\gamma} - C_2\frac{\log n}{n} d^{\frac{3}{2}+\gamma} \\ &\geq \frac{C\, c_1}{2} d^{-\gamma_1-\gamma_0},\end{aligned}$$

where the last inequality uses $d \leq c\frac{n^{\kappa_3}}{\log n}$ for small enough c. So we can obtain that when $\xi \in K_2$,

$$V_n(\xi) = n\left(\mathcal{R}_n(\widehat{\theta}+\frac{\xi}{\sqrt{n}}) - \mathcal{R}_n(\widehat{\theta})\right) - \left(\pi(\widehat{\theta}+\frac{\xi}{\sqrt{n}}) - \pi(\widehat{\theta})\right) \geq \frac{C\, c_1}{4}\cdot n \cdot d^{-\gamma_0}\big(d^{-\gamma_1} \wedge d^{-2\gamma_0-2\gamma_2}\big).$$

Thus using $d \leq c\frac{n^{\kappa_3}}{\log n}$, we have

$$\begin{aligned}&\int_{K_2} \exp(-V_n(\xi))\,\mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}}\det(\mathcal{H}_{\theta^*}) \\ &\leq \exp\Big(-\frac{d}{2}\log(2\pi) + \frac{d}{2}\log\big(\|\mathcal{H}_{\theta^*}\|_{\mathrm{op}}\big)\Big)\cdot \exp\Big(\frac{C\, c_1}{4}\cdot n\cdot d^{-\gamma_0}\big(d^{-\gamma_1} \wedge d^{-2\gamma_0-2\gamma_2}\big)\Big) \\ &\leq \exp\Big(\frac{C\, c_1}{8}\cdot n\cdot d^{-\gamma_0}\big(d^{-\gamma_1} \wedge d^{-2\gamma_0-2\gamma_2}\big)\Big).\end{aligned}$$

It remains to bound the denominator $\int \exp(-V_n(\xi))\,\mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}}\det(\mathcal{H}_{\theta^*})$, we have

$$\begin{aligned}&\int \exp(-V_n(\xi))\,\mathrm{d}\xi \cdot (2\pi)^{-\frac{d}{2}}\det(\mathcal{H}_{\theta^*}) \\ &\geq (2\pi)^{-\frac{d}{2}}\det(\mathcal{H}_{\theta^*})\int_{\|\xi\|\leq 4\sqrt{d/\lambda_{\min}(\mathcal{H}_{\theta^*})}} \exp\big(-\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\big)\,\mathrm{d}\xi \\ &\qquad \cdot \sup_{\|\xi\|\leq 4\sqrt{d/\lambda_{\min}(\mathcal{H}_{\theta^*})}} \exp\big(\frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2} - V_n(\xi)\big) \\ &\geq \exp(-\frac{1}{4}),\end{aligned}$$

where the last inequality uses $\lambda_{\min}(\mathcal{H}_{\theta^*}) \geq C\, d^{-\gamma_0}$, $d \leq c\frac{n^{\kappa_3}}{\log n}$ and the statements of Lemma 18. We can then obtain the desired results by combining all pieces.

D.3 Proof of Lemma 24

We first prove the first statement. It's equivalent to show that it holds with probability larger than $1-\frac{1}{3n^2}$ that for any $\theta,\theta'\in B_r(\theta^*)$ and $v\in\mathbb{S}^{d-1}$,

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta)-\frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta')-\mathbb{E}[v^Tg(X,\theta)]+\mathbb{E}[v^Tg(X,\theta')]\Big|\\&\le c\Big(\sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}\|\theta-\theta'\|^{\beta_1}+\frac{\log n}{n}d^{1+\gamma}\Big).\end{aligned}$$

Consider a minimal $\frac{3}{n}$-covering set $\mathcal{A}$ of $\mathbb{S}^{d-1}$ such that $\mathcal{A}\subset\mathbb{S}^{d-1}$, then $\log|\mathcal{A}|\le d\log n$. For any $v\in\mathcal{A}$, define the function class

$$\mathcal{G}_v=\{d^{-\gamma}(v^Tg(\cdot,\theta)-v^Tg(\cdot,\theta')):\theta,\theta'\in B_r(\theta^*)\}.$$

Let $\overline{\mathcal{G}}_v=\{af:a\in[0,1],f\in\mathcal{G}_v\}$ be the star hull of $\mathcal{G}_v$. Then since $\sup_{x\in\mathcal{X},\theta\in B_r(\theta^*)}\|g(x,\theta)\|\le C\,d^\gamma$, it holds that $\sup_{f\in\overline{\mathcal{G}}_v,x\in\mathcal{X}}|f(x)|\le 2C$. Consider the local Rademacher complexity associated with $\overline{\mathcal{G}}_v$,

$$\overline{R}_n(\delta;\overline{\mathcal{G}}_v)=\mathbb{E}_{X^{(n)}}\mathbb{E}_\varepsilon\Bigg[\sup_{\substack{f\in\overline{\mathcal{G}}_v\\ \mathbb{E}f^2\le\delta^2}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_if(X_i)\Big|\Bigg],$$

where ε_i are i.i.d. samples from Rademacher distribution, i.e., $\mathbb{P}(\varepsilon_i=1)=\mathbb{P}(\varepsilon_i=-1)=0.5$. We will use the following uniform law, which is a special case of Theorem 14.20 of Wainwright (2019), to prove the desired result.

Lemma 27 *(Wainwright (2019), Theorem 14.20) Given a uniformly 1-bounded function class $\mathcal{F}$ that is star shaped around 0, let $(\delta^*)^2\ge\frac{c}{n}$ be any solution to the inequality $\overline{R}_n(\delta;\mathcal{F})\le\delta^2$, then we have*

$$\sup_{f\in\mathcal{F}}\frac{\Big|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\Big|}{\sqrt{\mathbb{E}[f(X)^2]}+\delta^*}\le 10\delta^*$$

with probability greater than $1-c_1\exp(-c_2\,n\cdot(\delta^*)^2)$.

Next we will use Dudley's inequality (see for example, Theorem 5.22 of Wainwright (2019)) to determine the critical radius δ^* in Lemma 27. For $f,f':\mathcal{X}\to\mathbb{R}$, define the pseudometric

$$d_n(f,f')=\sqrt{\frac{1}{n}\sum_{i=1}^n(f(X_i)-f'(X_i))^2}.$$

Then by uniformly boundness of functions in class $\overline{\mathcal{G}}_v$, we can obtain that

$$\begin{aligned}&\log\mathbf{N}(\overline{\mathcal{G}}_v,d_n,\varepsilon)\\&\le\log\frac{4C}{\varepsilon}+\log\mathbf{N}(\mathcal{G}_v,d_n,\frac{\varepsilon}{2})\\&\le\log\frac{4C}{\varepsilon}+\log\mathbf{N}(\mathcal{B}_r(\theta^*),d_n^g,\frac{d^\gamma\varepsilon}{2})\\&\le C_1\,d\log\frac{n}{\varepsilon},\end{aligned}$$

where recall that $\mathbf{N}(\mathcal{F}, d_n, \varepsilon)$ denote the ε-covering number of class $\mathcal{F}$ w.r.t pseudo-metric d_n. Let

$$\begin{aligned}
r_n^2 = & \sup_{\substack{f,f'\in\mathcal{G}_v \\ \mathbb{E}[f^2],\mathbb{E}[f'^2]\le\delta^2}} d_n^2(f,f') \\
&\le 4 \sup_{\substack{f\in\mathcal{G}_v \\ \mathbb{E}[f^2]\le\delta^2}} \frac{1}{n}\sum_{i=1}^n f^2(X_i) \\
&\le 8 \sup_{\substack{f\in\mathcal{G}_v \\ \mathbb{E}[f^2]\le\delta^2}} \frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}f(X))^2 + 8\delta^2.
\end{aligned}$$

Then by (3.84) of Wainwright (2019), we can obtain that $\mathbb{E}[r_n^2] \le C\,\delta^2 + C\,\mathcal{R}_n(\delta)$. Choose $\delta^* = c\,d^{\frac{1}{2}}\sqrt{\frac{\log n}{n}}$, then by Dudley's inequality,

$$\begin{aligned}
\overline{\mathcal{R}}_n(\delta^*) &\le C\,\frac{1}{\sqrt{n}}\mathbb{E}\int_0^{r_n} d^{\frac{1}{2}}\sqrt{\log\frac{n}{\varepsilon}}d\varepsilon \\
&= C\,\frac{1}{\sqrt{n}}\mathbb{E}\int_0^1 r_n d^{\frac{1}{2}}\sqrt{\log\frac{n}{\varepsilon r_n}}d\varepsilon \\
&= C\,\mathbb{E}\left[\frac{1}{\sqrt{n}}\int_0^1 r_n d^{\frac{1}{2}}\sqrt{\log\frac{n}{\varepsilon r_n}}d\varepsilon\cdot\mathbf{1}(r_n < n^{-\frac{1}{2}})\right] + C\,\mathbb{E}\left[\frac{1}{\sqrt{n}}\int_0^1 r_n d^{\frac{1}{2}}\sqrt{\log\frac{n}{\varepsilon r_n}}d\varepsilon\cdot\mathbf{1}(r_n > n^{-\frac{1}{2}})\right] \\
&\le C\,d^{\frac{1}{2}}\frac{\sqrt{\log n}}{n} + C\,\mathbb{E}\Big[\frac{1}{\sqrt{n}}\int_0^1 r_n d^{\frac{1}{2}}\sqrt{\log\frac{n^{\frac{3}{2}}}{\varepsilon}}d\varepsilon\Big] \\
&\le C_1\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}}\sqrt{\delta^{*2}+\overline{\mathcal{R}}_n(\delta^*)}.
\end{aligned}$$

Then if $\overline{\mathcal{R}}_n(\delta^*) > (\delta^*)^2$, we can obtain that $\overline{\mathcal{R}}_n(\delta^*) \le 2C_1^2\,d\frac{\log n}{n} \le 2C_1^2 c^{-2}\,\delta^{*2}$. thus when c is large enough, δ^* solves the inequality $\overline{\mathcal{R}}_n(\delta^*) \le (\delta^*)^2$. Then by Lemma 27 and the assumption that $\sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}\big[(v^Tg(X,\theta)-v^Tg(X,\theta'))\big]^2 \le C\,d^{\gamma_3}\|\theta-\theta'\|^{2\beta_1}$, there exists a constant C such that it holds with probability larger than $1-\exp(-4d\log n)$ that for any $\theta,\theta'\in B_r(\theta^*)$,

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta') - \mathbb{E}v^Tg(X,\theta) + \mathbb{E}v^Tg(X,\theta')\Big| \\
&\le C\Big(\sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}\|\theta-\theta'\|^{\beta_1} + \frac{\log n}{n}d^{1+\gamma}\Big).
\end{aligned}$$

By the fact that $\log|\mathcal{A}| \le d\log n$, it holds with probability larger than $1-\exp(-3d\log n)$ that for any $v\in\mathcal{A}$ and $\theta,\theta'\in B_r(\theta^*)$,

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n v^Tg(X_i,\theta') - \mathbb{E}v^Tg(X,\theta) + \mathbb{E}v^Tg(X,\theta')\Big| \\
&\le C\Big(\sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}\|\theta-\theta'\|^{\beta_1} + \frac{\log n}{n}d^{1+\gamma}\Big).
\end{aligned}$$

Moreover, for any $\widetilde{v} \in \mathbb{S}^{d-1}$, there exists $v \in \mathcal{A}$ so that $\|v - \widetilde{v}\| \leq \frac{3}{n}$, hence for any $\theta, \theta' \in B_r(\theta^*)$,

$$\sup_{v\in\mathbb{S}^{d-1}}\Big|\frac{1}{n}\sum_{i=1}^{n}\widetilde{v}^T g(X_i,\theta) - \frac{1}{n}\sum_{i=1}^{n}\widetilde{v}^T g(X_i,\theta') - \mathbb{E}\widetilde{v}^T g(X,\theta) + \mathbb{E}\widetilde{v}^T g(X,\theta')\Big|$$
$$= \sup_{v\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} v^T g(X_i,\theta) - \frac{1}{n}\sum_{i=1}^{n} v^T g(X_i,\theta') - \mathbb{E}v^T g(X,\theta) + \mathbb{E}v^T g(X,\theta')\Big| + \mathcal{O}(\frac{d}{\sqrt{n}}).$$

Then, it follows that it holds with probability larger than $1 - \exp(3d\log n) \geq 1 - \frac{1}{3n^2}$ that

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) - \frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta') - \mathbb{E}g(X,\theta) + \mathbb{E}g(X,\theta')\Big\|$$
$$= \sup_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n} v^T g(X_i,\theta) - \frac{1}{n}\sum_{i=1}^{n} v^T g(X_i,\theta') - \mathbb{E}v^T g(X,\theta) + \mathbb{E}v^T g(X,\theta')\right|$$
$$\leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\|\theta - \theta'\|^{\beta_1} + \frac{\log n}{n} d^{1+\gamma}\right).$$

The proof of the first statement is then completed. For the second statement, by the assumption that for any $\theta, \theta' \in \Theta$ and $x \in \mathcal{X}$, $|\ell(X,\theta) - \ell(X,\theta')| \leq Cd^\gamma\|\theta - \theta'\|$, we can obtain that for any $\theta, \theta' \in \Theta$

$$\mathbb{E}\big[(\ell(X,\theta) - \ell(X,\theta'))^2\big] \leq C^2 d^{2\gamma}\|\theta - \theta'\|^2,$$

and

$$\sup_{x\in\mathcal{X}}|\ell(X,\theta) - \ell(X,\theta')| \leq Cd^\gamma(\|\theta\| + \|\theta'\|) \leq C_1\, d^{\frac{1}{2}+\gamma}.$$

We can therefore prove the second statement using the same strategy as the first statement. For the third statement, define $\delta_n = (\frac{\log n}{n} d^{-\frac{3}{2}}) \wedge ((\frac{\log n}{n})^{\frac{3}{2}} d^{-\frac{1+\gamma_3}{2}})^{\frac{1}{1+\beta_1}}$. For $k = 0, 1, \cdots, \lfloor \log_2 \frac{2r}{\delta_n} \rfloor + 1$, we define the set

$$\mathcal{A}_k = \begin{cases} \{\theta, \theta' \in B_r(\theta^*) : \|\theta - \theta'\| \leq \delta_n\} & k = 0; \\ \{\theta, \theta' \in B_r(\theta^*) : 2^{k-1}\delta_n < \|\theta - \theta'\| \leq 2^k\delta_n\} & k = 1, 2, \cdots \lfloor \log_2 \frac{2r}{\delta_n} \rfloor; \\ \{\theta, \theta' \in B_r(\theta^*) : 2^{k-1}\delta_n < \|\theta - \theta'\| \leq 2r\} & k = \lfloor \log_2 \frac{2r}{\delta_n} \rfloor + 1. \end{cases}$$

Then $\{\theta, \theta' \in B_r(\theta^*)\} = \sum_{k=1}^{\log_2\lfloor \frac{2r}{\delta_n}\rfloor + 1} \mathcal{A}_k$. Fix an integer $0 \leq k \leq \lfloor \log_2 \frac{2r}{\delta_n} \rfloor + 1$, we consider the function set

$$\mathcal{L}_k = \Big\{\frac{1}{2^k\delta_n} d^{-\gamma}(\ell(\cdot,\theta) - \ell(\cdot,\theta') - g(\cdot,\theta')(\theta - \theta')) : (\theta,\theta') \in \mathcal{A}_k\Big\}.$$

Then there exists a constant c such that for any $f \in \mathcal{L}_k$, it holds that $\sup_{x\in\mathcal{X}}|f(x)| \leq c$ and $\mathbb{E}[f^2(X)] \leq c\frac{1}{2^{2k}\delta_n^2} d^{-2\gamma} d^{\gamma_3}(2^k\delta_n)^{2+2\beta_1} \leq c\, d^{\gamma_3 - 2\gamma}(2^k\delta_n)^{2\beta_1} \leq 4c\, d^{\gamma_3-2\gamma}(2^{k-1}\delta_n)^{2\beta_1}$. Then consider the star hull $\overline{\mathcal{L}}_k$ of $\mathcal{L}_k$, by (1) $d \lesssim n^{\kappa_2}$; (2) the Lipschitzness of ℓ; (3) the bound on the ε-covering number of $B_r(\theta^*)$w.r.t d_n^g, it holds that

$$\begin{aligned}&\log \mathbf{N}(\overline{\mathcal{L}}_k, d_n, \varepsilon)\\ &\leq \log\frac{2c}{\varepsilon} + \log \mathbf{N}(\mathcal{L}_k, d_n, \varepsilon)\\ &\leq C\, d \log\frac{n}{\varepsilon}.\end{aligned}$$

Then similar as the proof of the first statement, we can use Dudley's inequality and Lemma 27 to obtain that there exists a constant c such that it holds with probability at least $1-\frac{1}{3n^3}$ that for any $(\theta,\theta')\in\mathcal{A}_k$,

$$\begin{aligned}&\bigg|\frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta') - \frac{1}{n}\sum_{i=1}^n g(X_i,\theta')(\theta-\theta')\\ &-\Big(\mathbb{E}\ell(X,\theta) - \mathbb{E}\ell(X,\theta') - \mathbb{E}g(X,\theta')(\theta-\theta')\Big)\bigg|\\ &\leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\cdot(2^{k-1}\delta_n)^{\beta_1+1} + \frac{\log n}{n} d^{1+\gamma}\cdot(2^{k-1}\delta_n)\right)\\ &\leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\cdot(\|\theta-\theta'\|+\delta_n)^{\beta_1+1} + \frac{\log n}{n} d^{1+\gamma}\cdot(\|\theta-\theta'\|+\delta_n)\right)\\ &\leq 4C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\|\theta-\theta'\|^{\beta_1+1} + \frac{\log n}{n} d^{1+\gamma}\|\theta-\theta'\| + (\frac{\log n}{n})^2\right).\end{aligned}$$

Then by $\log_2\frac{r}{\delta_n}\lesssim \log n$, consider the intersection of the above events for $k=0,1,\cdots,\lfloor\log_2\frac{r}{\delta_n}\rfloor+1$, we can obtain the desired result.

D.4 Proof of Lemma 25

Recall $\widehat{\theta}^{\diamond} = \theta^* - n^{-1}\sum_{i=1}^n \mathcal{H}_{\theta^*}^{-1} g(X_i,\theta^*)$, then by $\mathbb{E}[g(X,\theta^*) = \nabla\mathcal{R}(\theta^*) = 0$, we have

$$\begin{aligned}\|\widehat{\theta}^{\diamond}-\theta^*\| &= \|\frac{1}{n}\sum_{i=1}^n \mathcal{H}_{\theta^*}^{-1} g(X_i,\theta^*) - \mathbb{E}[\mathcal{H}_{\theta^*}^{-1} g(X,\theta^*)]\|\\ &= \sup_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^n v^T\mathcal{H}_{\theta^*}^{-1} g(X_i,\theta^*) - \mathbb{E}[v^T\mathcal{H}_{\theta^*}^{-1} g(X,\theta^*)]\right|.\end{aligned}$$

It remains to derive a high probability bound of the supremum of the above empirical process. Consider a minimal $\frac{3}{n}$-covering set $\mathcal{A}$ of S^{d-1} such that $A\subset\mathbb{S}^{d-1}$, then $\log|\mathcal{A}|\leq d\log n$. Fix an arbitrary $v\in\mathbb{S}^{d-1}$, then by the assumption that (1) $\mathcal{H}_{\theta^*}^{-1}\mathbb{E}[g(X_i,\theta^*)^T g(X_i,\theta^*)]\mathcal{H}_{\theta^*}^{-1}\preceq Cd^{\gamma_4}I_d$; (2) for any $\theta\in\Theta$, $\mathcal{R}(\theta)-\mathcal{R}(\theta^*)\geq C'd^{-\gamma_0}(d^{-\gamma_1}\wedge\|\theta-\theta^*\|^2)$, which leads to $\mathcal{H}_{\theta^*}\succeq C'd^{-\gamma_0}I_d$; (3) $\sup_{x\in\mathcal{X}}\|g(X,\theta^*)\|\leq Cd^{\gamma}$, we can obtain

$$\sup_{\substack{X\in\mathcal{X}\\ v\in\mathbb{S}^{d-1}}} |v^T\mathcal{H}_{\theta^*}^{-1} g(X,\theta^*)| \leq CC'd^{\gamma+\gamma_0},$$

and

$$\sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}[v^T\mathcal{H}_{\theta^*}^{-1} g(X,\theta^*)]^2\leq Cd^{\gamma_4}.$$

Therefore using Bernstein-type bound (see for example, Proposition 2.10 of Wainwright (2019)), we can get there exists a constant c such that it holds with probability larger than $1-\exp(3d\log n)$ that,

$$\left|\frac{1}{n}\sum_{i=1}^{n} v^T\mathcal{H}_{\theta^*}^{-1}g(X_i,\theta^*)-\mathbb{E}v^T\mathcal{H}_{\theta^*}^{-1}g(X,\theta^*)\right|\le C\,(d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}+d^{1+\gamma+\gamma_0}\frac{\log n}{n}).$$

Moreover, for any $\widetilde{v}\in\mathbb{S}^{d-1}$, there exists $v\in\mathcal{A}$ so that $\|v-\widetilde{v}\|\le\frac{3}{n}$, hence for any $\theta,\theta'\in B_r(\theta^*)$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{v}^T\mathcal{H}_{\theta^*}^{-1}g(X_i,\theta^*)-\mathbb{E}\widetilde{v}^T\mathcal{H}_{\theta^*}^{-1}g(X,\theta^*)\right| &\le\left|\frac{1}{n}\sum_{i=1}^{n}v^T\mathcal{H}_{\theta^*}^{-1}g(X_i,\theta^*)-\mathbb{E}v^T\mathcal{H}_{\theta^*}^{-1}g(X,\theta^*)\right|\\ &\quad+\mathcal{O}(d^{\gamma_0+\gamma}\frac{\log n}{n}).\end{aligned}$$

Thus by a simple union bound, it holds with probability larger than $1-\exp(2d\log n)>1-\frac{1}{n^2}$ that

$$\sup_{v\in\mathbb{S}^{d-1}}\left|\frac{1}{n}\sum_{i=1}^{n}v^T\mathcal{H}_{\theta^*}^{-1}g(X_i,\theta^*)-\mathbb{E}v^T\mathcal{H}_{\theta^*}^{-1}g(X,\theta^*)\right|\le 2C\,(d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}+d^{1+\gamma+\gamma_0}\frac{\log n}{n}).$$

We can thus obtain that it holds with probability larger than $1-n^{-2}$ that

$$\|\widehat{\theta}^{\diamond}-\theta^*\|\le C\,d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}+C\,d^{1+\gamma+\gamma_0}\frac{\log n}{n}.$$

D.5 Proof of Lemma 26

Firstly by $\mathcal{R}_n(\widehat{\theta})\le\mathcal{R}_n(\theta^*)$ and $\mathcal{R}(\theta)-\mathcal{R}(\theta^*)\ge C'd^{-\gamma_0}(d^{-\gamma_1}\wedge\|\theta-\theta^*\|^2)$, we can obtain that

$$C'd^{-\gamma_0}(d^{-\gamma_1}\wedge\|\widehat{\theta}-\theta^*\|^2)\le\mathcal{R}(\widehat{\theta})-\mathcal{R}(\theta^*)\le\mathcal{R}(\widehat{\theta})-\mathcal{R}(\theta^*)-\mathcal{R}_n(\widehat{\theta})+\mathcal{R}_n(\theta^*).$$

It follows from the second statement of Lemma 24 that

$$d^{-\gamma_0}(d^{-\gamma_1}\wedge\|\widehat{\theta}-\theta^*\|^2)\le C\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma}\|\widehat{\theta}-\theta^*\|+C\,\frac{\log n}{n}d^{\frac{3}{2}+\gamma}.$$

If $\|\widehat{\theta}-\theta^*\|\ge d^{-\frac{\gamma_1}{2}}$, then

$$d^{-\gamma_0-\gamma_1}\le C\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma}\|\widehat{\theta}-\theta^*\|+C\,\frac{\log n}{n}d^{\frac{3}{2}+\gamma}.$$

On the other hand, as $\widehat{\theta}\in\Theta\subseteq[-C,C]^d$, we have $\|\widehat{\theta}-\theta^*\|\le 2C\sqrt{d}$, we can then obtain that when $d\le c(\frac{n}{\log n})^{\frac{1}{2+2(\gamma+\gamma_0+\gamma_1)}}$,

$$\begin{aligned}\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma}\|\widehat{\theta}-\theta^*\|+\frac{\log n}{n}d^{\frac{3}{2}+\gamma} &\le 2Cd^{1+\gamma}\sqrt{\frac{\log n}{n}}+\frac{\log n}{n}d^{\frac{3}{2}+\gamma}\\ &\le 2C\sqrt{c}\,d^{-\gamma_0-\gamma_1}+c\,d^{-\frac{1}{2}-\gamma-2(\gamma_0+\gamma_1)},\end{aligned}$$

which will cause contradiction when c is sufficiently small. Hence we have $\|\widehat{\theta}-\theta^*\| < d^{-\frac{\gamma_1}{2}}$ and thus

$$d^{-\gamma_0}\|\widehat{\theta}-\theta^*\|^2 \le C\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma}\|\widehat{\theta}-\theta^*\| + C\,\frac{\log n}{n}d^{\frac{3}{2}+\gamma},$$

which leads to $\|\widehat{\theta}-\theta^*\| \le C_1\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma+\gamma_0}$. We will first show the first statement of Lemma 26 and use the statement to improve the dependence of d in the bound of $\sqrt{\frac{\log n}{n}}d^{\frac{1}{2}+\gamma+\gamma_0}$.

By $\mathcal{R}_n(\widehat{\theta}) \le \mathcal{R}_n(\widetilde{\theta})$ for any $\widetilde{\theta} \in B_r(\theta^*)$, we can obtain that

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})\\
&\le \mathcal{R}_n(\widetilde{\theta}) - \mathcal{R}_n(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})\\
&\le \left|\mathcal{R}_n(\widetilde{\theta}) - \mathcal{R}_n(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta}) - \mathcal{R}(\widetilde{\theta}) + \mathcal{R}(\widehat{\theta}) + \mathbb{E}[g(X,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})]\right|\\
&+ \left|\mathcal{R}(\widetilde{\theta}) - \mathcal{R}(\widehat{\theta}) - \mathbb{E}[g(X,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})]\right|.
\end{aligned}$$

The first term can be bounded using the third statement of Lemma 24, that is

$$\begin{aligned}
&\left|\mathcal{R}_n(\widetilde{\theta}) - \mathcal{R}_n(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta}) - \mathcal{R}(\widetilde{\theta}) + \mathcal{R}(\widehat{\theta}) + \mathbb{E}[g(X,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})]\right|\\
&\le C\sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}\|\widehat{\theta}-\widetilde{\theta}\|^{\beta_1+1} + C\,\frac{\log n}{n}d^{1+\gamma}\|\widehat{\theta}-\widetilde{\theta}\| + C\,(\frac{\log n}{n})^2.
\end{aligned}$$

The second term can be bounded using the twice differentiability of $\mathcal{R}$ around θ^*,

$$\left|\mathcal{R}(\widetilde{\theta}) - \mathcal{R}(\widehat{\theta}) - \mathbb{E}[g(X,\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})]\right| \le \frac{1}{2}\sup_{c\in[0,1]}\left\|\left|\mathcal{H}_{c\widetilde{\theta}+(1-c)\widehat{\theta}}\right\|\right|_{\mathrm{op}}\|\widehat{\theta}-\widetilde{\theta}\|^2 \le C\,d\|\widehat{\theta}-\widetilde{\theta}\|^2.$$

where the last inequality is due to the assumption that the mixed partial derivatives of $\mathcal{R}(\theta)$ up to order two are uniformly bounded by an (n,d)-independent constant on $\mathcal{B}_r(\theta^*)$. Then we choose $\widetilde{\theta} = \widehat{\theta} - t\frac{\sum_{i=1}^n g(X_i,\widehat{\theta})}{\|\sum_{i=1}^n g(X_i,\widehat{\theta})\|}$ for a $t>0$ that will be chosen later. Thus

$$\begin{aligned}
&C_1\,t\left\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})\right\| \le \sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}t^{\beta_1+1} + \frac{\log n}{n}d^{1+\gamma}t + (\frac{\log n}{n})^2 + dt^2\\
&\Rightarrow C_1\,\left\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})\right\| \le \sqrt{\frac{\log n}{n}}d^{\frac{1+\gamma_3}{2}}t^{\beta_1} + \frac{\log n}{n}d^{1+\gamma} + (\frac{\log n}{n})^2/t + dt.
\end{aligned}$$

Choose $t = \frac{\log n}{n}$, we have it holds with probability at least $1-n^{-2}$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta})\right\| \le C\,d^{1+\gamma}\frac{\log n}{n} + C\,d^{\frac{1+\gamma_3}{2}}(\frac{\log n}{n})^{\frac{1}{2}+\beta_1}.$$

For the second statement, recall $\widehat{\theta}^{\diamond} = \theta^* - \frac{1}{n}\sum_{i=1}^n \mathcal{H}_{\theta^*}^{-1} g(X_i, \theta)$. By Lemma 25 and the assumption that $d \leq c(\frac{n}{\log n})^{\frac{1}{2+2(\gamma+\gamma_0+\gamma_1)}}$, we can obtain $\|\widehat{\theta}^{\diamond} - \theta^*\| \leq C\, d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}$. We claim that it suffices to show that

$$\left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta}^{\diamond})\right\| \leq C\, d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})}(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} + C\, d^{1+\gamma\vee(\gamma_2+\gamma_4)}\frac{\log n}{n} \tag{30}$$

holds with probability at least $1 - cn^{-2}$. Indeed, under the above statement, we have

$$\begin{aligned}
&\|\mathbb{E}[g(X, \widehat{\theta})] - \mathbb{E}[g(X, \widehat{\theta}^{\diamond})]\| \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta}^{\diamond}) - \mathbb{E}[g(X, \widehat{\theta})] + \mathbb{E}[g(X, \widehat{\theta}^{\diamond})]\right\| \\
&\quad + \left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta})\right\| + \left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta}^{\diamond})\right\| \\
&\leq C\left(\sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\|\widehat{\theta} - \widehat{\theta}^{\diamond}\|^{\beta_1} + d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})}(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} + d^{1+\gamma\vee(\gamma_2+\gamma_4)}\frac{\log n}{n}\right),
\end{aligned}$$

where the last inequality follows from the first statement of Lemma 24. On the other hand, by the Lipschitzness of $\mathcal{H}_\theta$ around θ^*, we can obtain that,

$$\begin{aligned}
&\|\mathbb{E}[g(X, \widehat{\theta})] - \mathbb{E}[g(X, \widehat{\theta}^{\diamond})]\| \\
&\geq \|\mathcal{H}_{\theta^*}(\widehat{\theta} - \widehat{\theta}^{\diamond})\| - \|\mathbb{E}[g(X, \widehat{\theta})] - \mathbb{E}[g(X, \widehat{\theta}^{\diamond})] - H_{\theta^*}(\widehat{\theta} - \widehat{\theta}^{\diamond})\| \\
&= \|\mathcal{H}_{\theta^*}(\widehat{\theta} - \widehat{\theta}^{\diamond})\| - \sup_{v\in\mathbb{S}^{d-1}}\left|\mathbb{E}[v^T g(X, \widehat{\theta})] - \mathbb{E}[v^T g(X, \widehat{\theta}^{\diamond})] - v^T H_{\theta^*}(\widehat{\theta} - \widehat{\theta}^{\diamond})\right| \\
&\geq \rho_1(\mathcal{H}_{\theta^*})\|\widehat{\theta} - \widehat{\theta}^{\diamond}\| - \sup_{v\in\mathbb{S}^{d-1}}\sup_{t\in(0,1)}\left|v^T(\mathcal{H}_{t\widehat{\theta}^{\diamond}+(1-t)\widehat{\theta}} - \mathcal{H}_{\theta^*})(\widehat{\theta} - \widehat{\theta}^{\diamond})\right| \\
&\geq \rho_1(\mathcal{H}_{\theta^*})\|\widehat{\theta} - \widehat{\theta}^{\diamond}\| - C\left(d^{\frac{1}{2}+\gamma+\gamma_2+\gamma_0}\sqrt{\frac{\log n}{n}}\|\widehat{\theta} - \widehat{\theta}^{\diamond}\|\right),
\end{aligned}$$

where the last inequality uses $\|\widehat{\theta} - \theta^*\| \leq C_1\sqrt{\frac{\log n}{n}} d^{\frac{1}{2}+\gamma+\gamma_0}$ and $\|\widehat{\theta}^{\diamond} - \theta^*\| \leq C\, d^{\frac{1+\gamma_4}{2}}\sqrt{\frac{\log n}{n}}$ with $\gamma_4 \leq 2(\gamma_0 + \gamma)$. Hence when $d \leq c(\frac{n}{\log n})^{\frac{1}{1+2\gamma+2\gamma_2+4\gamma_0}}$ for a sufficiently small c, we can obtain that

$$C_1\, d^{-\gamma_0}\|\widehat{\theta} - \widehat{\theta}^{\diamond}\| \leq \sqrt{\frac{\log n}{n}} d^{\frac{1+\gamma_3}{2}}\|\widehat{\theta} - \widehat{\theta}^{\diamond}\|^{\beta_1} + d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})}(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} + d^{1+\gamma\vee(\gamma_2+\gamma_4)}\frac{\log n}{n},$$

which leads to

$$\|\widehat{\theta} - \widehat{\theta}^{\diamond}\| \leq C\left(d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})+\gamma_0}(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} + d^{1+\gamma\vee(\gamma_2+\gamma_4)+\gamma_0}\frac{\log n}{n} + \left(d^{\frac{1+\gamma_3}{2}+\gamma_0}\sqrt{\frac{\log n}{n}}\right)^{\frac{1}{1-\beta_1}}\right).$$

Now we show equation (30), using the first statement of Lemma 24, we can obtain that

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n g(X_i, \widehat{\theta}^{\diamond}) - \frac{1}{n}\sum_{i=1}^n g(X_i, \theta^*) - \mathbb{E}g(X, \widehat{\theta}^{\diamond}) + \mathbb{E}g(X, \theta^*)\right\| \\
&\leq C\,(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})} + C\,\frac{\log n}{n} d^{1+\gamma}.
\end{aligned}$$

Moreover, by the Lipschitz continuity of $\mathcal{H}_\theta$ around θ^*, we can obtain that

$$\|\mathbb{E}g(X,\widehat{\theta}^\diamond) - \mathbb{E}g(X,\theta^*) - \mathcal{H}_{\theta^*}(\widehat{\theta}^\diamond - \theta^*)\| \le d^{\gamma_2}\|\widehat{\theta}^\diamond - \theta^*\|^2 \le C\, d^{1+\gamma_4+\gamma_2}\frac{\log n}{n}.$$

Therefore, combined with the fact that $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*) + \mathcal{H}_{\theta^*}(\widehat{\theta}^\diamond - \theta^*) = 0$, we can obtain that it holds with probability at least $1 - cn^{-2}$ that

$$\begin{aligned}
\Big\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta}^\diamond)\Big\| &= \Big\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta}^\diamond) - \frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*) - \mathcal{H}_{\theta^*}(\widehat{\theta}^\diamond - \theta^*)\Big\| \\
&\le \Big\|\frac{1}{n}\sum_{i=1}^n g(X_i,\widehat{\theta}^\diamond) - \frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*) - \mathbb{E}g(X,\widehat{\theta}^\diamond) + \mathbb{E}g(X,\theta^*)\Big\| \\
&\quad + \|\mathbb{E}g(X,\widehat{\theta}^\diamond) - \mathbb{E}g(X,\theta^*) - \mathcal{H}_{\theta^*}(\widehat{\theta}^\diamond - \theta^*)\| \\
&\le C\, d^{\frac{1+\gamma_3}{2}+\beta_1(\frac{1+\gamma_4}{2})}(\frac{\log n}{n})^{\frac{1+\beta_1}{2}} + C\, d^{1+\gamma\vee(\gamma_2+\gamma_4)}\frac{\log n}{n}.
\end{aligned}$$

Appendix E. Proof of Remaining Results

E.1 Proof of Lemma 4

Let $\pi_{\mathrm{loc}} = [\sqrt{n}(\cdot - \widehat{\theta})]_{\#}\pi_n$ and $\mu_{\mathrm{loc}} = [\sqrt{n}(\cdot - \widehat{\theta})]_{\#}\mu_0$. We can bound

$$\begin{aligned}
M_0 &= \sup_{A\,:\,\pi_n(A)>0}\frac{\mu_0(A)}{\pi_n(A)} \\
&\overset{(i)}{=} \sup_{A\subset K\,:\,\pi_{\mathrm{loc}}(A)>0}\frac{\mu_{\mathrm{loc}}(A)}{\pi_{\mathrm{loc}}(A)} \\
&= \sup_{A\subset K\,:\,\pi_{\mathrm{loc}}(A)>0}\frac{\mu_{\mathrm{loc}}(A)}{\pi_{\mathrm{loc}}|_K(A)}\cdot\frac{1}{\pi_{\mathrm{loc}}(K)} \\
&\le \sup_{x\in K}\left[\frac{\int_K \exp(-\frac{1}{2}x^TJx)\,\mathrm{d}x\exp(-\frac{1}{2}x^T\widetilde{I}^{-1}x)}{\int_K \exp(-\frac{1}{2}x^T\widetilde{I}^{-1}x)\,\mathrm{d}x\exp(-\frac{1}{2}x^TJx)}\cdot\frac{\int_K \exp(-V_n(x))\,\mathrm{d}x\exp(-\frac{1}{2}x^TJx)}{\int_K \exp(-\frac{1}{2}x^TJx)\,\mathrm{d}x\exp(-V_n(x))}\right]\cdot\frac{1}{\pi_{\mathrm{loc}}(K)} \\
&\le \sup_{x\in K}\frac{\int_K \exp(-\frac{1}{2}x^TJx)\,\mathrm{d}x\exp(-\frac{1}{2}x^T\widetilde{I}^{-1}x)}{\int_K \exp(-\frac{1}{2}x^T\widetilde{I}^{-1}x)\,\mathrm{d}x\exp(-\frac{1}{2}x^TJx)}\cdot\sup_{x\in K}\frac{\int_K \exp(-V_n(x))\,\mathrm{d}x\exp(-\frac{1}{2}x^TJx)}{\int_K \exp(-\frac{1}{2}x^TJx)\,\mathrm{d}x\exp(-V_n(x))}\cdot\frac{1}{\pi_{\mathrm{loc}}(K)},
\end{aligned}$$

where (i) uses $\mu_{\mathrm{loc}}(K) = 0$. Since for any function pair f_1, f_2, it holds that

$$\int_K f_1(x)\,\mathrm{d}x\cdot\sup_{x\in K}\frac{f_2(x)}{f_1(x)} \ge \int_K f_1(x)\frac{f_2(x)}{f_1(x)}\,\mathrm{d}x = \int_K f_2(x)\,\mathrm{d}x,$$

we can obtain that

$$M_0 \le \sup_{x\in K}\exp(|x^T(\widetilde{I}^{-1} - J)x|)\cdot\sup_{x\in K}\exp\big(2\big|V_n(x) - \frac{1}{2}x^TJx\big|\big)\cdot\frac{1}{\pi_{\mathrm{loc}}(K)}.$$

E.2 Proof of Corollary 7

We first verify that under Condition B.3', Condition B.2 and Condition B.3 holds, where the function g in Condition B.3 is chosen as the gradient $\nabla_\theta \ell$. Condition B.2 and B.3.1 directly follows from the assumption that $\|\nabla_\theta \ell(x,\theta)\| \leq C d^{\gamma}$. For Condition B.3.2, since $\vert\!\vert\!\vert \mathrm{Hess}_\theta(\ell(x,\theta))\vert\!\vert\!\vert_{\mathrm{op}}^2 \leq C d^{\gamma_3}$, we have for any $x \in \mathcal{X}$ and $\theta \in \Theta$,

$$\|\nabla_\theta \ell(x,\theta) - \nabla_\theta \ell(x,\theta')\| \leq \sqrt{C d^{\gamma_3}} \|\theta - \theta'\|$$

and thus

$$d_n^g(\theta,\theta') \leq \sqrt{C d^{\gamma_3}} \|\theta - \theta'\|.$$

Then the covering number condition for d_n^g follows from the fact that the ε-covering number of unit d-ball is bounded by $(\frac{3}{\varepsilon})^d$. Condition B.3.3 directly follows from the assumption that $\vert\!\vert\!\vert \mathrm{Hess}_\theta(\ell(x,\theta))\vert\!\vert\!\vert_{\mathrm{op}}^2 \leq C d^{\gamma_3}$. Condition B.3.4 follows from the assumption that $\mathcal{H}_{\theta^*}^{-1} \Delta_{\theta^*} \mathcal{H}_{\theta^*}^{-1} \preceq C\, d^{\gamma_4} I_d$ with $\Delta_{\theta^*} = \mathbb{E}[\nabla_\theta \ell(X,\theta^*) \nabla_\theta \ell(X,\theta^*)^T]$. Then the first statement directly follows from Theorem 5. For the second statement, we first verify that $\widetilde{I}^{-1} = |S|^{-1} \sum_{i \in S} \mathrm{Hess}_\theta(\ell(X_i, \widehat{\theta}))$ is a reasonable estimator to $\mathcal{H}_{\theta^*}$ in the following lemma.

Lemma 28 *Under assumptions in Corollary 7, let $m = |S|$, it holds with probability larger than $1 - n^{-2}$ that*

$$\vert\!\vert\!\vert \widetilde{I}^{-1} - \mathcal{H}_{\theta^*} \vert\!\vert\!\vert_{\mathrm{op}} \leq C\,\big(d^{\frac{\gamma_3+1}{2}} \sqrt{\frac{\log n}{m}}\big) \vee \big(d^{\frac{\gamma_3+2}{2}} \frac{\log n}{m}\big) \vee \big(d^{\frac{1+\gamma_4}{2}+\gamma_2} \sqrt{\frac{\log n}{n}}\big).$$

Then since $\vert\!\vert\!\vert \mathcal{H}_{\theta^*}^{-1} \vert\!\vert\!\vert \leq C\, d^{\gamma_0}$, $d \leq c \frac{n^{\kappa_1}}{\log n}$ and $m \geq C_2\, d^{\gamma_3 + 2\gamma_0 + \frac{7}{3}}$, we have

$$\vert\!\vert\!\vert \widetilde{I} \vert\!\vert\!\vert_{\mathrm{op}} \leq 2 C d^{\gamma_0},$$

and

$$\begin{aligned}
\vert\!\vert\!\vert \widetilde{I}^{\frac{1}{2}} \mathcal{H}_{\theta^*} \widetilde{I}^{\frac{1}{2}} - I_d \vert\!\vert\!\vert_{\mathrm{op}} &\leq \vert\!\vert\!\vert \widetilde{I} \vert\!\vert\!\vert_{\mathrm{op}} \vert\!\vert\!\vert \widetilde{I}^{-1} - \mathcal{H}_{\theta^*} \vert\!\vert\!\vert_{\mathrm{op}} \\
&\leq C\, d^{\gamma_0} \big(d^{\frac{\gamma_3+1}{2}} \sqrt{\frac{\log n}{m}}\big) \vee \big(d^{\frac{\gamma_3+2}{2}} \frac{\log n}{m}\big) \vee \big(d^{\frac{1+\gamma_4}{2}+\gamma_2} \sqrt{\frac{\log n}{n}}\big) \\
&\leq \frac{1}{2},
\end{aligned}$$

which leads to

$$\frac{1}{2} I_d \preceq \widetilde{I}^{\frac{1}{2}} \mathcal{H}_{\theta^*} \widetilde{I}^{\frac{1}{2}} \preceq 2 I_d.$$

Then by

$$\mathcal{H}_{\theta^*} = \widetilde{I}^{-\frac{1}{2}} \big(\widetilde{I}^{\frac{1}{2}} \mathcal{H}_{\theta^*} \widetilde{I}^{\frac{1}{2}}\big) \widetilde{I}^{-\frac{1}{2}},$$

we have

$$\begin{aligned}
&\vert\!\vert\!\vert \widetilde{I} \vert\!\vert\!\vert_{\mathrm{op}} \leq 2 \vert\!\vert\!\vert \mathcal{H}_{\theta^*}^{-1} \vert\!\vert\!\vert_{\mathrm{op}}; \\
&\vert\!\vert\!\vert \widetilde{I}^{-1} \vert\!\vert\!\vert_{\mathrm{op}} \leq 2 \vert\!\vert\!\vert \mathcal{H}_{\theta^*} \vert\!\vert\!\vert_{\mathrm{op}}.
\end{aligned}$$

Thus the requirements for the preconditioning matrix $\widetilde{I}$ in Theorem 5 are satisfied with $\rho_2 = 2$ and $\rho_1 = \frac{1}{2}$. Finally, we will control the warming parameter using Lemma 4.

Recall $\mu_0 = N_d(\widehat{\theta}, n^{-1}\widetilde{I})\big|_{\{\theta:\sqrt{n}\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})\|\le 3\sqrt{c_1 d}\}}$, where c_1 is a constant so that $c_1 \geq 9 \vee \sup\limits_{i\in[d],j\in[d]} \frac{\partial^2\mathcal{R}(\theta^*)}{\partial\theta_i\partial\theta_j}$. By

$$\|\!|\widetilde{I}^{-\frac{1}{2}}\|\!|_{\mathrm{op}} \leq \sqrt{2}\|\!|\mathcal{H}_{\theta^*}^{\frac{1}{2}}\|\!|_{\mathrm{op}} \leq \sqrt{2d \sup_{i\in[d],j\in[d]} \frac{\partial^2\mathcal{R}(\theta^*)}{\partial\theta_i\partial\theta_j}} \leq \sqrt{2c_1 d},$$

and Lemma 19, we can obtain that

$$\pi_n\Big(\sqrt{n}\|\widetilde{I}^{-\frac{1}{2}}(\theta-\widehat{\theta})\| \leq 2\sqrt{c_1 d}\Big) \geq 1 - \exp(-1).$$

Moreover, consider $K = \{\xi : \widetilde{I}^{-\frac{1}{2}}\xi \leq 2\sqrt{c_1 d}\}$, then for any $\xi \in K$, we have

$$\|\xi\| \leq 2\|\!|\widetilde{I}^{\frac{1}{2}}\|\!|_{\mathrm{op}}\sqrt{c_1 d} \leq 2\sqrt{2c_1 d}\|\!|\mathcal{H}_{\theta^*}^{-\frac{1}{2}}\|\!|_{\mathrm{op}} \leq c_2 d^{\frac{1+\gamma_0}{2}}.$$

Then by Lemma 3, when $d \leq c\frac{n^{\kappa_1}}{\log n}$ for a small enough c, for any $\xi \in K$, we have

$$\Big|V_n(\xi) - \frac{\xi^T\mathcal{H}_{\theta^*}\xi}{2}\Big| \leq \frac{1}{2}.$$

In addition, for any $\xi \in K$, we have

$$\begin{aligned}
\sup_{\xi\in K}\big|\xi^T(\widetilde{I}^{-1} - \mathcal{H}_{\theta^*})\xi\big| &= \sup_{\|\xi\|\leq 2\sqrt{c_1 d}}\big|\xi^T(I_d - \widetilde{I}^{\frac{1}{2}}\mathcal{H}_{\theta^*}\widetilde{I}^{\frac{1}{2}})\xi\big| \\
&\leq 2c_1 d\|\!|I_d - \widetilde{I}^{\frac{1}{2}}\mathcal{H}_{\theta^*}\widetilde{I}^{\frac{1}{2}}\|\!|_{\mathrm{op}} \\
&\leq 2c_1 C\, d^{\gamma_0+1}\big(d^{\frac{\gamma_3+1}{2}}\sqrt{\frac{\log n}{m}}\big) \vee \big(d^{\frac{\gamma_3+2}{2}}\frac{\log n}{m}\big) \vee \big(d^{\frac{1+\gamma_4}{2}+\gamma_2}\sqrt{\frac{\log n}{n}}\big) \\
&\leq d^{\frac{1}{3}},
\end{aligned}$$

where the last inequality uses $d \leq c\frac{n^{\kappa_1}}{\log n}$ and $m \geq C_2\, d^{\gamma_3+2\gamma_0+\frac{7}{3}}$. The desired result then follows from Lemma 4.

E.3 Proof of Lemma 28

Since $\mathbb{E}[\widetilde{I}^{-1}] = \mathcal{H}_{\widehat{\theta}}$, we have

$$\|\!|\widetilde{I}^{-1} - \mathcal{H}_{\theta^*}\|\!|_{\mathrm{op}} \leq \|\!|\widetilde{I}^{-1} - \mathcal{H}_{\widehat{\theta}}\|\!|_{\mathrm{op}} + \|\!|\mathcal{H}_{\widehat{\theta}} - \mathcal{H}_{\theta^*}\|\!|_{\mathrm{op}}.$$

The second term can be bounded using Condition B.1.2 and equation (27) in the proof of Lemma 18, that is

$$\|\!|\mathcal{H}_{\widehat{\theta}} - \mathcal{H}_{\theta^*}\|\!|_{\mathrm{op}} \leq C\, d^{\gamma_2}\|\widehat{\theta} - \theta^*\| \leq C\, d^{\frac{1+\gamma_4}{2}+\gamma_2}\sqrt{\frac{\log n}{n}}.$$

The first term can be bounded using Bernstein's inequality. Let $m = |S|$, for $v, v' \in \mathbb{S}^{d-1}$ and $\theta, \theta' \in B_r(\theta^*)$, we have

$$\begin{aligned}
&\sqrt{m^{-1}\sum_{i\in S}\big(v^T\mathrm{Hess}_\theta(\ell(X_i,\theta))v - v'^T\mathrm{Hess}_\theta(\ell(X_i,\theta'))v'\big)^2} \\
&\leq C\sqrt{d}\|v - v'\| + C\, d^{r_1}\|\theta - \theta'\|.
\end{aligned} \tag{31}$$

Then consider $\mathcal{N}_v$ and $\mathcal{N}_\theta$ to be the minimal n^{-1} and $n^{-1}d^{-r_1}$ covering set of $\mathbb{S}^{d-1}$ and $B_r(\theta^*)$, then $\log|\mathcal{N}_v| \leq C\, d \log n$ and $\log|\mathcal{N}_\theta| \leq C\, d\log n$. Using the fact that

$$\sup_{\theta\in B_r(\theta^*),\, X\in\mathcal{X}} \left|\!\left|\!\left| \mathrm{Hess}_\theta(\ell(X,\theta)) \right|\!\right|\!\right|_{\mathrm{op}} \leq C\, d^{\frac{\gamma_3}{2}};$$

$$\sup_{\theta\in B_r(\theta^*)\, v\in\mathbb{S}^{d-1}} \mathbb{E}\big[(v^T\mathrm{Hess}_\theta(\ell(X,\theta))^2\big] \leq \sup_{\substack{\theta,\theta'\in B_r(\theta^*),\\ v\in\mathbb{S}^{d-1}}} \mathbb{E}\Big[\frac{(v^T\nabla\ell(X,\theta) - v^T\nabla\ell(X,\theta'))^2}{\|\theta-\theta'\|^2}\Big] \leq C\, d^{\gamma_3},$$

we can get by Bernstein's inequality and a simple union bound argument that it holds with probability at least $1-n^{-c}$ that for any $v\in\mathcal{N}_v$ and $\theta\in\mathcal{N}_\theta$,

$$\sup_{\theta\in B_r(\theta^*)} \sup_{v\in\mathbb{S}^{d-1}} \left(v^T(m^{-1}\sum_{i\in S}\mathrm{Hess}_\theta(\ell(X_i,\theta)) - \mathcal{H}_\theta)v^T \right) \leq C\,\big(d^{\frac{\gamma_3+1}{2}}\sqrt{\frac{\log n}{m}}\big) \vee \big(d^{\frac{\gamma_3+2}{2}}\frac{\log n}{m}\big).$$

E.4 Proof of Corollary 8

We will first check that Conditions B.1-B.3 hold for the quantile regression example under Condition D.1 and D.2. Consider the loss function

$$\ell(X,\theta) = (Y - \widetilde{X}^T\theta)(\tau - \mathbf{1}(Y < \widetilde{X}^T\theta)),$$

and its subgradient

$$g(X,\theta) = (\mathbf{1}(Y<\widetilde{X}^T\theta) - \tau)\widetilde{X}.$$

Then we can write

$$\mathcal{R}(\theta) = \mathbb{E}[\ell(X,\theta)] = \mathbb{E}\big[\tau\,(Y-\widetilde{X}^T\theta)\big] - \mathbb{E}\Big[\int_{-\infty}^{\widetilde{X}^T\theta-\widetilde{X}^T\theta^*} (\varepsilon + \widetilde{X}^T\theta^* - \widetilde{X}^T\theta) f_e(\varepsilon)d\varepsilon\Big].$$

Taking derivative of $\mathcal{R}$ w.r.t θ, we can obtain

$$\nabla\mathcal{R}(\theta) = -\tau\cdot\mathbb{E}[\widetilde{X}] + \mathbb{E}[\mathbf{1}(Y<\widetilde{X}^T\theta)\widetilde{X}] = \mathbb{E}g(X,\theta).$$

Thus,

$$\mathcal{H}_\theta = \mathbb{E}[f_e(\widetilde{X}^T\theta - \widetilde{X}^T\theta^*)\widetilde{X}\widetilde{X}^T].$$

Then for $\theta\in B_{c/\sqrt{d}}(\theta^*)$ with a small enough c, it holds that

$$\frac{f_e(\widetilde{X}^T\theta - \widetilde{X}^T\theta^*)}{f_e(0)} \geq \frac{1}{2}.$$

Then by the fact that $\nabla\mathcal{R}(\theta^*) = 0$ and $\mathbb{E}[\widetilde{X}\widetilde{X}^T] \succeq C' d^{-\alpha_0} I_d$, we can obtain that for any $\theta\in B_{\frac{c}{\sqrt{d}}}(\theta^*)$,

$$\mathcal{R}(\theta) - \mathcal{R}(\theta^*) \geq C_1\, d^{-\alpha_0}\|\theta-\theta^*\|^2;$$

on the other hand, for any $\theta\in B_{\frac{c}{\sqrt{d}}}(\theta^*)^c$,

$$\mathcal{R}(\theta) - \mathcal{R}(\theta^*) \geq \mathcal{R}\Big(\theta^* + \frac{c(\theta-\theta^*)}{\sqrt{d}\|\theta-\theta^*\|}\Big) - \mathcal{R}(\theta^*) \geq C_1\, d^{-\alpha_0-1},$$

hence for any $\theta \in \mathbb{R}^d$,

$$\mathcal{R}(\theta) - \mathcal{R}(\theta^*) \geq C_1\, d^{-\alpha_0}(d^{-1} \wedge \|\theta - \theta^*\|^2).$$

Moreover, for any $\theta \in \Theta$ and $v \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
|v^T(\mathcal{H}_\theta - \mathcal{H}_{\theta^*})v| &\leq v^T\mathbb{E}\left[\left|f_e(\widetilde{X}^T\theta - \widetilde{X}^T\theta^*) - f_e(0)\right|\widetilde{X}\widetilde{X}^T\right]v \\
&\leq C\,\mathbb{E}\left[|\widetilde{X}^T(\theta - \theta^*)|v^T\widetilde{X}\widetilde{X}^Tv\right] \\
&\leq C\,\|\theta - \theta^*\|\mathbb{E}\left(\left|\widetilde{X}(\theta - \theta^*)/\|\theta - \theta^*\|\right|^3\right)^{\frac{1}{3}}(\mathbb{E}|v^T\widetilde{X}|^3)^{\frac{2}{3}} \\
&\leq C\,d^{\alpha_1}\|\theta - \theta^*\|,
\end{aligned}$$

where the last inequality uses the assumption that $\sup_{\eta\in\mathbb{S}^{d-1}}\mathbb{E}[\eta^T\widetilde{X}] \leq Cd^{\alpha_1}$. Thus we have Condition B.1 holds with $\gamma_0 = \alpha_0$, $\gamma_1 = 1$, $\gamma_2 = \alpha_1$. For Condition B.2, by $\mathcal{X} = \mathrm{supp}(\widetilde{X}) \subseteq [-C, C]^d$, we can obtain $\|g(X,\theta)\| \leq C\sqrt{d}$, thus for any θ, θ', $|\ell(X,\theta) - \ell(X,\theta')| \leq C\sqrt{d}\|\theta - \theta'|$ and Condition B.2 and Condition B.3.1 hold with $\gamma = \frac{1}{2}$. For Condition B.3, since for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
\sqrt{\frac{1}{n}\sum_{i=1}^n\|g(X_i,\theta) - g(X_i,\theta')\|^2} &= \sqrt{\frac{1}{n}\sum_{i=1}^n\|\widetilde{X}_i\|^2(\mathbf{1}(Y < \widetilde{X}_i^T\theta) - \mathbf{1}(Y < \widetilde{X}_i^T\theta'))^2} \\
&= \sqrt{d}\sqrt{\frac{1}{n}\sum_{i=1}^n(\mathbf{1}(Y < \widetilde{X}_i^T\theta) - \mathbf{1}(Y < \widetilde{X}_i^T\theta'))^2},
\end{aligned}$$

by Lemma 9.8 and Lemma 9.12 of Kosorok (2008), the function class $\mathcal{F} = \{\mathbf{1}(Y \leq \theta^T\widetilde{X}), \theta \in \Theta\}$ is a VC-class with VC-dimension being bouned by $d + 3$, then using Theorem 8.3.18 of Vershynin (2018) on the covering number's upper bound via VC dimension, we can verify Condition B.3.2.

For Condition B.3.3, since for any $v \in \mathbb{S}^{d-1}$ and $\theta, \theta' \in \Theta$,

$$\begin{aligned}
\mathbb{E}(v^Tg(X,\theta) - v^Tg(X,\theta'))^2 &= \mathbb{E}[(\mathbf{1}(Y < \widetilde{X}_i^T\theta) - \mathbf{1}(Y < \widetilde{X}_i^T\theta'))^2(v^T\widetilde{X})^2] \\
&= \mathbb{E}\left[(v^T\widetilde{X})^2\int_{\widetilde{X}^T\theta\wedge\widetilde{X}^T\theta'}^{\widetilde{X}^T\theta\vee\widetilde{X}^T\theta'} f(y - \widetilde{X}^T\theta^*|\widetilde{X})\,\mathrm{d}y\right] \\
&\leq C\,\mathbb{E}\left[(v^T\widetilde{X})^2|\widetilde{X}^T\theta - \widetilde{X}^T\theta'|\right] \\
&\leq C\,\|\theta - \theta'\|\sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}|v^T\widetilde{X}|^3 \leq C\,d^{\alpha_1}\|\theta' - \theta\|;
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\big[(\ell(X,\theta) - \ell(X,\theta') - g(X,\theta')(\theta - \theta'))^2\big] \\
&= \mathbb{E}\big[\big(-(Y - \widetilde{X}^T\theta)\mathbf{1}(Y < \widetilde{X}_i^T\theta) + (Y - \widetilde{X}^T\theta')\mathbf{1}(Y < \widetilde{X}_i^T\theta') - \mathbf{1}(Y < \widetilde{X}_i^T\theta')\widetilde{X}^T(\theta - \theta')\big)^2\big] \\
&= \mathbb{E}\left[\int_{\widetilde{X}^T\theta\wedge\widetilde{X}^T\theta'}^{\widetilde{X}^T\theta\vee\widetilde{X}^T\theta'}(y - \widetilde{X}^T\theta)^2 f(y - \widetilde{X}\theta^*|\widetilde{X})\,\mathrm{d}y\right] \\
&\leq C\,\mathbb{E}|\widetilde{X}^T\theta - \widetilde{X}^T\theta'|^3 \leq C\,d^{\alpha_1}\|\theta' - \theta\|^3.
\end{aligned}$$

Thus Condition B.3.3 holds with $\gamma_3 = \alpha_1$ and $\beta_1 = \frac{1}{2}$. For condition B.3.4, since

$$\mathbb{E}[g(X,\theta^*)g(X,\theta^*)^T] = \mathbb{E}\big[(\tau^2 + \mathbf{1}(Y < \widetilde{X}^T\theta) - 2\tau\mathbf{1}(Y < \widetilde{X}^T\theta))\widetilde{X}\widetilde{X}^T\big] = (\tau - \tau^2)\mathbb{E}[\widetilde{X}\widetilde{X}^T],$$

and $J = \mathcal{H}_{\theta^*} = f_e(0)\mathbb{E}[\widetilde{X}\widetilde{X}^T]$, we have

$$(\mathbb{E}[\widetilde{X}\widetilde{X}^T])^{\frac{1}{2}} J^{-1} (\mathbb{E}[\widetilde{X}\widetilde{X}^T])^{\frac{1}{2}} = f_e(0)^{-1} I_d,$$

and thus $\gamma_4 = \gamma_0$.

Now we verify that the requirements of the $\widetilde{I}$ in Theorem 5 are satisfied. Recall $\widetilde{I}^{-1} = \frac{1}{|S|}\sum_{i\in S} X_i X_i^T$, in order to show that $|||\widetilde{I}^{-\frac{1}{2}} J^{-1} \widetilde{I}^{-\frac{1}{2}}|||_{\mathrm{op}} \vee |||\widetilde{I}^{\frac{1}{2}} J \widetilde{I}^{\frac{1}{2}}|||_{\mathrm{op}}$ is bounded above by a constant, we will derive upper bound to the term of $|||\widetilde{I}^{\frac{1}{2}}(\mathbb{E}[\widetilde{X}\widetilde{X}^T])\widetilde{I}^{\frac{1}{2}} - I_d|||_{\mathrm{op}}$. Let $m = |S|$, similar as the proof for Lemma 28, we can obtain it holds with probability larger than $1 - \frac{1}{n^2}$ that

$$\begin{aligned}\Big|\Big|\Big| n^{-1}\sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^T - \mathbb{E}[\widetilde{X}\widetilde{X}^T]\Big|\Big|\Big|_{\mathrm{op}} &\le C \sup_{v\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}|v^T\widetilde{X}|^4}\, d^{\frac{1}{2}} \sqrt{\frac{\log n}{m}} + d^2\frac{\log n}{m} \\ &\le C\, d^{\frac{3}{4}+\frac{\alpha_1}{2}} \sqrt{\frac{\log n}{m}} + d^2 \frac{\log n}{m},\end{aligned}$$

where the last inequality is due to $\sup_{v\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}|v^T\widetilde{X}|^4} \le C\, d^{\frac{1}{4}} \sup_{v\in\mathbb{S}^{d-1}} \sqrt{\mathbb{E}|v^T\widetilde{X}|^3} \le C\, d^{\frac{1+2\alpha_1}{4}}$. Then by $\mathbb{E}[\widetilde{X}\widetilde{X}^T] \succeq C' d^{-\alpha_0} I_d$, and $m \ge C_2\, d^{\alpha_1 + 2\alpha_0 + 3/2}\log n$, we can obtain

$$|||\widetilde{I}|||_{\mathrm{op}} \le \frac{2}{C'} d^{\alpha_0}$$

Thus we have

$$|||\widetilde{I}^{\frac{1}{2}}(\mathbb{E}[\widetilde{X}\widetilde{X}^T])\widetilde{I}^{\frac{1}{2}} - I_d|||_{\mathrm{op}} \le |||\widetilde{I}|||_{\mathrm{op}} |||\widetilde{I}^{-1} - (\mathbb{E}[\widetilde{X}\widetilde{X}^T])|||_{\mathrm{op}} \le C_1\, d^{\alpha_0 + \frac{3+2\alpha_1}{4}} \sqrt{\frac{\log n}{m}},$$

which leads to

$$\frac{1}{2} I_d \preceq \widetilde{I}^{\frac{1}{2}}(\mathbb{E}[XX^T])\widetilde{I}^{\frac{1}{2}} \preceq 2I_d,$$

Thus

$$\frac{1}{2} f_e(0) I_d \preceq \widetilde{I}^{\frac{1}{2}} \mathcal{H}_{\theta^*} \widetilde{I}^{\frac{1}{2}} \preceq 2 f_e(0) I_d$$

Furthermore, by

$$\mathcal{H}_{\theta^*} = f_e(0)\cdot \widetilde{I}^{-\frac{1}{2}}\big(\widetilde{I}^{\frac{1}{2}}(\mathbb{E}[XX^T])\widetilde{I}^{\frac{1}{2}}\big)\widetilde{I}^{-\frac{1}{2}},$$

we have

$$\begin{aligned}|||\widetilde{I}|||_{\mathrm{op}} &\le 2f_e(0)|||\mathcal{H}_{\theta^*}^{-1}|||_{\mathrm{op}}; \\ |||\widetilde{I}^{-1}|||_{\mathrm{op}} &\le \frac{2}{f_e(0)}|||\mathcal{H}_{\theta^*}|||_{\mathrm{op}}.\end{aligned}$$

We can then obtain that the requirements for the preconditioning matrix $\widetilde{I}$ in Theorem 5 are satisfied with $\rho_2 = 2f_e(0)$ and $\rho_1 = \frac{1}{2}f_e(0)$.