Part III Essays Easter 2023

E782H

Attach to Front of Essay

Essay Number:

41

Essay Title:

Sampling from High-Dimensional Posterior Distributions

MAY 2, 2024

ABSTRACT

This essay aims to review the literature on the theoretical guarantees for the mixing times of Markov Chain Monte Carlo Methods, particularly the Langevin Monte Carlo. We first review the Langevin Monte Carlo Algorithms for the case of strongly log-concave posterior densities and the polynomial time guarantees on the mixing times of the algorithm for a given precision level [1]. We would then, go beyond the non-concave setting explaining ideas in [2], [3]. We would consider the example of generalised linear models and explain the polynomial time mixing guarantees and the proofs for the same.

Contents

1 Introduction

In the realm of statistical inference and computational modeling, Markov Chain Monte Carlo (MCMC) methods have emerged as indispensable tools for approximating complex probability distributions and solving intricate computational problems. The main idea of MCMC methods is to generate an ergodic Markov Chain $(\vartheta_k)_{k\in\mathbb{N}}$ whose laws $\mathcal{L}(\vartheta_k)$ approximate its invariant measure on $\mathbb{R}^p$. In the context of Bayesian Inference, the invariant measure is set to be the posterior measure, given by

$$\pi(\theta|Z^{(n)}) = \frac{e^{\ell_n(\theta)}\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)}\pi(\theta)d\theta} \propto e^{\ell_n(\theta)}\pi(\theta), \qquad \theta \in \mathbb{R}^p,$$

where θ is the parameter of interest, Θ is the parameter space, $\pi(\theta)$ is the prior density function on the parameter, $\ell_n(\theta) : \mathbb{R}^p \mapsto \mathbb{R}$ is the log-likelihood function based on the data $Z^{(n)}$, which is generated using a ground truth $\theta_0 \in \ell^2(\mathbb{N})$ and $\pi(\theta|Z^{(n)})$ is the density of the posterior measure, arising from the observations.

However, for high-dimensional statistical models where $p \asymp n^\rho, \rho > 0$, the computational hardness in terms of the mixing time of the Markov chain $(\vartheta_k)_{k\in\mathbb{N}}$ typically scales exponentially in the model dimension p, sample size n, and the accuracy ϵ. In this essay, we present the works in [1], [2], and [3], which try to address this problem in different settings. [1] establishes nonasymptotic bounds for the error of approximating the target distribution using the Langevin Monte Carlo (LMC) algorithm and its variants for the case of a smooth, strongly log-concave posterior density with lipschitz gradients. Using the nonasymptotic bounds on the approximation error, the author, Dalalyan establishes easy-to-apply rules for choosing the parameters in the LMC algorithm, which in turn leads to polynomial dependence on the key parameters of the number of iterations of the LMC algorithm required to reach within ϵ-Total Variation neighborhood of the target measure. Dalalyan also extends this idea to the case of non-strongly log-concave posterior densities by 'strongly-convexifying' the negative log-posterior density. [2] develops mathematical techniques to address the issue for the non-concave likelihood $\ell_n(\theta)$ setting with a Gaussian prior. The authors, Nickl and Wang use the local geometric properties of the statistical model with tools from Bayesian nonparametrics to justify polynomial time feasibility of the LMC algorithm for the PDE model with the Schrödinger equation. The key idea is to rely on the Fisher information for providing a natural statistical notion of curvature of the log-likelihood function near θ_0. [3] extends the proof strategy developed in [2] to more general high-dimensional statistical models. The author, Altmeyer then considers concrete examples of density estimation and non-parametric regression, (with unbounded, possibly non-lipschitz regression functions) to apply the polynomial mixing times obtained for the general model.

In this essay, we present the work in [1] to give non-asymptotic bounds on the error of the LMC algorithm for the case of strongly log-concave posterior with lipschitz gradients in section 4, and then extend it to the nonstrongly log-concave posterior in section 5. We then present the proofs in [3] by taking a particular example of a Generalised Linear Model in sections 6 and 7. We then simulate the LMC algorithm in the case of a logistic regression GLM to qualitatively study the convergence properties of the same.

2 Notation

We write $a \lesssim b$ if $a \leq Cb$ for a universal constant C, and $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. We denote I_p as the $p \times p$ Identity matrix, $\mathbb{1}_A$ as the Indicator of any event A. For a measurable space $(\mathcal{X}, \mathcal{A})$, equipped with a measure $\nu_\mathcal{X}$, we write $L^p(\mathcal{X}), 1 \leq p \leq \infty$, the space of p-integrable, $\mathcal{A}$-measurable functions with respect to $\nu_\mathcal{X}$ and the norm is given by $\|.\|_{L^p}$. We denote $\ell^p(\mathbb{N})$ the space of p-

summable sequences with norm $||.||_{\ell^p}$ and $||.|| = ||.||_{\ell^2}$. Let $||.||_2$m $||.||_\infty$ denote the Euclidean norm and sup norm in $\mathbb{R}^p$, respectively. For a matrix $Q \in \mathbb{R}^{p\times p}$, let $||Q||_{\text{op}}$ be the operator norm, that is

$$||Q||_{\text{op}} := \sup_{b:||b||\leq 1} ||Qb||.$$

For functions $f, g : \mathbb{R} \mapsto \mathbb{R}$, we define the convolution as

$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau.$$

For two Borel probability measures μ_1, μ_2 on $\mathbb{R}^p$, with finite second moments, we denote the squared-Wasserstein distance as

$$W_2^2(\mu_1, \mu_2) = \inf_{\mu\in\Gamma(\mu_1,\mu_2)} \int_{\mathbb{R}^p\times\mathbb{R}^p} ||\theta - \theta'||^2 d\mu(\theta, \theta'),$$

where $\Gamma(\mu_1, \mu_2)$ is the set of all couplings of μ_1 and μ_2. Let $\mathcal{B}(\mathbb{R}^p)$ denote the σ-algebra of Borel sets of $\mathbb{R}^p$. The total variation norm of a signed measure μ is given by $||\mu||_{TV} = \sup_{A\in\mathcal{B}(\mathbb{R}^p)} |\mu(A)|$, the Kullback-Leibler and χ^2 divergences between the measures μ_1 and μ_2 are respectively defined by

$$\text{KL}(\mu_1||\mu_2) = \int_{\mathcal{X}} \log\left(\frac{d\mu_1}{d\mu_2}(x)\right)\mu_1(dx), \qquad \chi^2(\mu_1||\mu_1) = \int_{\mathcal{X}} \left(\frac{d\mu_1}{d\mu_2}(x) - 1\right)^2 \mu_2(dx).$$

The minimal and maximal eigenvalues of a positive symmetric matrix Σ are denoted by $\lambda_{\min}(\Sigma)$, $\lambda_{\max}(\Sigma)$. Denote the space of k-times differentiable real-valued functions as $C^k(\mathcal{X})$. On $C^1(\mathcal{X})$, we define the norm $||f||_{C^1} = \sup_{x\in\mathcal{X}} |f(x)| + \sup_{x\in\mathcal{X}} |f'(x)|$, and similarly, the C^2-norm is defined on $C^2(\mathcal{X})$. The gradient and Hessian of a real-valued function $f : \mathbb{R}^p \mapsto \mathbb{R}$ are denoted by ∇f and $\nabla^2 f$ respectively. We say that f is lipschitz if it has finite lipschitz norm

$$||f||_{\text{Lip}} := \sup_{x\neq y, x,y\in\mathbb{R}^p} \frac{|f(x) - f(y)|}{||x - y||}.$$

We say that f is globally m_f-strongly concave and has Λ_f-Lipschitz gradients for $m_f, \Lambda_f > 0$, if for all $\theta, \theta' \in \mathbb{R}^p$

$$\begin{aligned} ||\nabla f(\theta) - \nabla f(\theta')||_2 \leq \Lambda_f ||\theta - \theta'||_2, \\ f(\theta') \leq f(\theta) + (\theta' - \theta)^T \nabla f(\theta) - \frac{m_f}{2}||\theta - \theta'||_2^2. \end{aligned}$$

3 Langevin diffusion

A Langevin diffusion is a stochastic differential equation describing how a system evolves when subjected to a combination of deterministic and random forces. Consider a p-dimensional density π which is non-zero everywhere and differentiable so that $\nabla \log \pi(\theta)$ is well defined. Then the Langevin L_t diffusion is defined by the p-dimensional stochastic differential equation

$$dL_t = \nabla \log \pi(L_t) + \sqrt{2} dW_t,$$

where W_t is a p-dimensional Brownian motion. When π is suitably smooth, it can be shown that L_t has π as a stationary measure, that is $\pi(B) = \int_{\mathbb{R}} \mathbb{P}_L^t(x, B)\pi(dx)$ for all Borel sets $B \in \mathcal{B}(\mathbb{R})$ for all t, and also that

$$||\mathbb{P}_L^t(\theta, .) - \pi||_{TV} \to 0,$$

where $\mathbb{P}_L^t(\theta, A) = \mathbb{P}(L_t \in A|L_0 = \theta)$. More concretely, we state without proof, Theorem 2.1 of [4].

Theorem 1. *Suppose that $\nabla \log \pi(\theta)$ is continuously differentiable and that, for some $N, a, b \leq \infty$,*

$$\nabla \log \pi(\theta) \cdot \theta \leq a||\theta||_2^2 + b, \qquad ||\theta||_2 > N. \tag{1}$$

Then the Langevin diffusion L_t satisfies the following:

(i) The diffusion is non-explosive, μ^{Leb}-irreducible, aperiodic, strong Feller and all compact sets are small.

(ii) The measure π is invariant for L and, moreover, for all θ,

$$||\mathbb{P}_L^t(\theta, .) - \pi||_{TV} \to 0.$$

Note that the gradient $\nabla \log \pi(\theta)$ can be computed even if the density $\pi(\theta)$ is known only up to a scalar factor, that is only $\pi(\theta_1)/\pi(\theta_2)$ is known for $\theta_1, \theta_2 \in \mathbb{R}^p$. This is of particular interest in posterior inference in Bayesian statistics. This motivates the Langevin Monte Carlo algorithm, which can be seen as a discrete approximation of the Langevin diffusion.

3.1 Langevin Monte Carlo

The Langevin Monte Carlo (LMC) algorithm produces a Markov chain $\{\vartheta_k\}_{k\in\mathbb{N}}$ which is the Euler discretization of a continuous time Langevin diffusion $\{L_t : t \geq 0\}$, which has the target density $\pi(\theta|Z^{(n)})$ as the invariant measure.

$$\vartheta_{k+1} = \vartheta_k + \gamma\nabla \log \pi(\vartheta_k|Z^{(n)}) + \sqrt{2\gamma}\zeta_{k+1}, \qquad \vartheta_0 = \theta_{\text{init}}. \tag{2}$$

where $\zeta \sim N(0, I_p)$ is called the innovation term, γ is the step size and θ_{init} is the initialisation of the algorithm. The Markov chain can be seen to be μ^{Leb}-irreducible and weak Feller, given that $\nabla \log \pi(.|Z^{(n)})$ is continuous. But other properties like ergodicity are sensitive to the choice of step size parameter γ. But as shown in [1], if the log posterior $\log \pi(.|Z^{(n)})$ is globally m-strongly concave and has Λ-Lipschitz gradients, we can ensure the non-transience of the Markov Chain $(\vartheta_k)_{k\in\mathbb{N}}$ by simply choosing $\gamma \leq 1/\Lambda$.

4 Strongly log-concave posterior

We now consider the case of strongly log-concave posterior densities with lipschitz gradients and establish some non-asymptotic convergence guarantees for the LMC algorithm. The first thing we establish here is the geometric ergodicity of the Langevin diffusion, that is, the total-variation distance between the Langevin diffusion L_t and the posterior measure $\Pi(\theta|Z^{(n)})$ decreases exponentially in time if the posterior measure is m-strongly concave and has Λ-Lipschitz gradients.

Lemma 1. *If the target distribution $\pi(\theta|Z^{(n)})$ is m-strongly concave, then for any probability density ν, $\forall t \geq 0$*

$$\|\nu\mathbb{P}_L^t - \pi(\theta|Z^{(n)})\|_{TV} \leq \frac{1}{2}\chi^2(\nu||\pi(\theta|Z^{(n)}))^{1/2}e^{-tm/2}.$$

Something to note is that the results we establish in this section hold for any strongly-concave target distribution π, but since we are interested in these results in the setting of posterior inference, we will use the target measure as a posterior measure $\Pi(\theta|Z^{(n)})$.

Proof of Lemma 1 is based on the geometric ergodicity of L_t and can be found in Lemma 1 of [1].

We then establish the non-explosivity of the LMC iterates ϑ_k for the case when $\gamma \leq 1/\Lambda$ in the form of the following proposition

Proposition 1. *Let the function f be continuously differentiable on $\mathbb{R}^p$ and is m-strongly concave with Λ-Lipschitz gradients. Let $f^* = \sup_{\theta\in\mathbb{R}^p} f(\theta)$. Then, for every $\gamma \leq 1/\Lambda$, we have*

$$\mathbb{E}\left[f^* - f(\vartheta_k)\right] \leq (1 - m\gamma)^k\mathbb{E}\left[f^* - f(\vartheta_0)\right] + \frac{\Lambda p}{m}. \tag{3}$$

Note that when $\gamma \leq 1/\Lambda$, the term $(1 - m\gamma) \geq 0$ since using Taylor's series expansion, it is easy to see that $m \leq \Lambda$ and hence $1 - m\gamma \geq 1 - \Lambda\gamma \geq 0$. Also using the strongly-concave and lipschitz gradients assumption, we can further show that

$$\mathbb{E}\left[\|\vartheta_k - \theta^*\|_2^2\right] \leq \frac{\Lambda}{m}\mathbb{E}\left[\|\vartheta_0 - \theta^*\|_2^2\right] + \frac{2\Lambda p}{m^2}, \tag{4}$$

where θ^* is the point of the global maximiser of f. This follows from the previous proposition and the strong concavity of f. As a result, we get that the LMC iterates are bounded in L^2 and hence non-explosive when $\delta \leq 1/\Lambda$.

Proof of Proposition 1: Let us denote the shorthand $f^k = f(\vartheta_k)$ and $\nabla f^k = \nabla f(\vartheta_k)$. Using a Taylor expansion, we get that

$$\begin{aligned} f^{k+1} &\geq f^k + (\nabla f^k)^T(\vartheta_{k+1} - \vartheta_k) - \frac{M}{2}\|\vartheta_{k+1} - \vartheta_k\|_2^2 \\ &= f^k + \gamma\|\nabla f^k\|_2^2 + \sqrt{2\gamma}(\nabla f^k)^T\zeta^{(k+1)} - \frac{\Lambda}{2}\|\gamma\nabla f^k + \sqrt{2\gamma}\zeta^{(k+1)}\|_2^2. \end{aligned}$$

Using the LMC dynamics in (2). Now taking expectations on both sides

$$\begin{aligned} \mathbb{E}\left[f^{k+1}\right] &\geq \mathbb{E}[f^k] + \gamma\mathbb{E}[\|\nabla f^k\|_2^2] - \frac{\Lambda}{2}\gamma^2\mathbb{E}[\|\nabla f^k\|_2^2] - \Lambda\gamma p \\ &= \mathbb{E}[f^k] + \frac{1}{2}\gamma(2 - \Lambda\gamma)\mathbb{E}[\|\nabla f^k\|_2^2] - \Lambda\gamma p. \end{aligned} \tag{5}$$

Now using the inequality $\|\nabla f(\theta)\|_2^2 \geq 2m(f^* - f(\theta))$, we get that

$$\mathbb{E}[f^{k+1}] \geq \mathbb{E}[f^k] + m\gamma(2 - \Lambda\gamma)\mathbb{E}[f^* - f^k] - \Lambda\gamma p.$$

By rearranging the terms and adding $\mathbb{E}f^*$ on both sides, with $\kappa = m\gamma(2 - \Lambda\gamma) \in (0, 1)$ for any $\gamma \in (0, 2/\Lambda)$, we get

$$\mathbb{E}[f^* - f^{k+1}] \leq (1 - \kappa)\mathbb{E}[f^* - f^k] + \Lambda\gamma p.$$

This implies that,

$$\begin{aligned}\mathbb{E}[f^* - f^{k+1}] &\leq (1 - \kappa)^{k+1}\mathbb{E}[f^* - f(\vartheta_0)] + \Lambda\gamma p(1 + ... + (1 - \kappa)^k),\\ &\leq (1 - \kappa)^{k+1}\mathbb{E}[f^* - f(\vartheta_0)] + \Lambda\gamma p\kappa^{-1}.\end{aligned}$$

Hence, we get that,

$$\mathbb{E}[f^* - f^k] \leq (1 - m\gamma(2 - \Lambda\gamma))^k\mathbb{E}[f^* - f(\vartheta_0)] + \frac{\Lambda\gamma p}{m\gamma(2 - \Lambda\gamma)}.$$

Now for $\gamma \leq 1/\Lambda$, $(2 - \Lambda\gamma) \geq 1$ and hence the desired result in (3) follows. □

A key step in analyzing the behavior of the LMC algorithm is to construct a continuous-time diffusion $\bar{L}_t$ such that the distribution of the LMC-iterates $(\vartheta_k)_{k\in\mathbb{N}}$ coincide with $(\bar{L}_{k\gamma})_{k\in\mathbb{N}}$. Then we can upper bound the distance between the random variables $\bar{L}_{K\gamma}$ and $L_{K\gamma}$ by the distance between the distributions of the continuous-time processes $\{L_t : t \in [0, K\gamma]\}$ and $\{\bar{L}_t : t \in [0, K\gamma]\}$. More concretely, we define $\bar{L}_t$ as

$$d\bar{L}_t = \nabla f(\bar{L}_{\lfloor t/\gamma\rfloor\gamma})dt + \sqrt{2}dW_t, \qquad \bar{L}_0 = L_0 = \theta_{\text{init}}. \tag{6}$$

By the integrating the above equation on the interval $[k\gamma, (k + 1)\gamma]$, we get that

$$\bar{L}_{(k+1)\gamma} - \bar{L}_{k\gamma} = \gamma\nabla f(L_{k\gamma}) + \sqrt{2\gamma}Z_{k+1},$$

where $Z_{k+1} = (W_{(k+1)\gamma} - W_{k\gamma})/\sqrt{\gamma}$. Now since W_t is a p-dimensional Brownian motion, $(Z_k)_{k\in\mathbb{N}}$ is a sequence of iid standard Gaussian random vectors. Hence, the distribution of the random vectors $(\vartheta_0, \vartheta_1, ..., \vartheta_K)$ and $(\bar{L}_0, \bar{L}_1, ..., \bar{L}_K)$ coincide. We will use this construction again to prove the exit probability of the Langevin diffusion from a local curvature region $\mathcal{B}$ in Theorem 9.

4.1 Nonasymptotic bounds on the error of the LMC algorithm

The error of the LMC algorithm is determined by two types of errors: The error of approximating the posterior $\Pi(\theta|Z^{(n)})$ by the distribution of the Langevin diffusion L_t, and the error of approximating the continuous-time Langevin diffusion by its discretization $\bar{L}_t$ in (6). As established in the Lemma 1, the first type of error decays exponentially in T. The second type of error would be small if $\gamma \to 0$. But then the number of LMC iterates required to reach time T given by $T/\gamma \to \infty$. Also, each iteration of the LMC algorithm has a computational cost of $O(p)$. For a given T, the computational complexity of the algorithm would be $O(pT/\gamma)$ which decreases with an increase in γ. Hence we need to establish a trade-off between the computational complexity and the approximation error. We first establish an upper bound on the second type of approximation error in the following lemma.

Lemma 2. *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a function with Λ-Lipschitz gradients, and $\theta^* \in \mathbb{R}^p$ be a stationary point. For any $T > 0$, let $\mathbb{P}_L^{\theta_{init},T}$ and $\mathbb{P}_{\bar{L}}^{\theta_{init},T}$, respectively be the distributions of the Langevin* *diffusion L_t and its approximation $\bar{L}_t$ on the space of all continuous paths on $[0,T]$ with values in $\mathbb{R}^p$ and a fixed initial value θ_{init}. Then, if $\gamma \leq \frac{1}{\alpha\Lambda}$ with $\alpha \geq 1$, it holds that*

$$KL\left(\mathbb{P}_L^{\theta_{init},K\gamma}||\mathbb{P}_{\bar{L}}^{\theta_{init},K\gamma}\right) \leq \frac{\Lambda^3\gamma^2\alpha}{12(2\alpha-1)}(\|\theta_{init}-\theta^*\|_2^2+2K\gamma p)+\frac{pK\Lambda^2\gamma^2}{4}. \tag{7}$$

Now let us set $T = K\gamma$. Also, to simplify the term $\|\theta_{\text{init}}-\theta^*\|_2^2$ in (7), we assume that the initial value θ_{init} is drawn at random from a Gaussian distribution with mean θ^* and covariance $\Lambda^{-1}I_p$. that is $\nu \sim N(\theta^*, \Lambda^{-1}I_p)$. Now using (7) and the convexity of the Kullback-Leibler divergence, we get that for all $K \geq \alpha$ and $\gamma \leq \frac{1}{\alpha\Lambda}$

$$\text{KL}(\nu\mathbb{P}_L^T||\nu\mathbb{P}_{\bar{L}}^T) \leq \frac{p\Lambda^2\gamma^2\alpha}{12(2\alpha-1)}+\frac{p\Lambda^3T\gamma^2\alpha}{6(2\alpha-1)}+\frac{p\Lambda^2T\gamma}{4} \leq \frac{p\Lambda^2T\gamma\alpha}{2(2\alpha-1)}. \tag{8}$$

Proof of Lemma 2: We use an inequality from [5], which says that if for some $B>0$ and nonanticipative drift function $b: C(\mathbb{R}_+,\mathbb{R}^p)\mapsto\mathbb{R}^p$ satisfies the inequality $\|b(\bar{L},t)\|_2 \leq B(1+\|\bar{L}\|_\infty), \forall t \in [0,K\gamma]$, (which holds since the gradient of the posterior is Λ-Lipschitz) then the Kullback-Leibler divergence between the distributions of the processes $\{L_t : t\in[0,K\gamma]\}$ and $\{\bar{L}_t : t\in[0,K\gamma]\}$ with the initial value θ_{init} is given by

$$\text{KL}\left(\mathbb{P}_L^{\theta_{\text{init}},K\gamma}||\mathbb{P}_{\bar{L}}^{\theta_{\text{init}},K\gamma}\right) = \frac{1}{4}\int_0^{K\gamma}\mathbb{E}[\|\nabla f(\bar{L}_t)+b_t(\bar{L})\|_2^2]dt.$$

This gives us,

$$\begin{aligned}\text{KL}\left(\mathbb{P}_L^{\theta_{\text{init}},T}||\mathbb{P}_{\bar{L}}^{\theta_{\text{init}},T}\right) &= \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\gamma}^{(K+1)\gamma}\mathbb{E}[\|\nabla f(\bar{L}_t)-\nabla f(\bar{L}_{k\gamma})\|_2^2]dt,\\ &\leq \frac{\Lambda^2}{4}\sum_{k=0}^{K-1}\int_{k\gamma}^{(K+1)\gamma}\mathbb{E}[\|\bar{L}_t-\bar{L}_{k\gamma}\|_2^2]dt.\end{aligned}$$

Since ∇f is Λ-Lipschitz. Now using the Langevin dynamics (2), we get

$$\begin{aligned}\text{KL}\left(\mathbb{P}_L^{\theta_{\text{init}},T}||\mathbb{P}_{\bar{L}}^{\theta_{\text{init}},T}\right) &\leq \frac{\Lambda^2}{4}\sum_{k=0}^{K-1}\int_{k\gamma}^{(K+1)\gamma}\left(\mathbb{E}[\|\nabla f(\bar{L}_{k\gamma})\|_2^2(t-k\gamma)^2]+2p(t-k\gamma)\right)dt,\\ &= \frac{\Lambda^2\gamma^3}{12}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_k)\|_2^2]+\frac{pK\Lambda^2\gamma^2}{4}.\end{aligned}$$

Now, we will upper bound the term $\gamma\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\vartheta_k)\|_2^2]$ using the proof of Proposition 1. Using (5), and that $2-M\gamma \geq (2\alpha-1)/\alpha$, we get that

$$\frac{\gamma(2\alpha-1)}{2\alpha}\mathbb{E}[\|\nabla f^k\|_2^2] \leq \mathbb{E}[f^{k+1}-f^k]+\Lambda\gamma p.$$

Since the above inequality holds for all $k\in\mathbb{N}$, we sum it for $k=0,...,K-1$, and using $f^K \leq f^*$, we get

$$\gamma\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f^k\|_2^2] \leq \frac{2\alpha}{2\alpha-1}\mathbb{E}[f^*-f^0]+\frac{2\alpha\Lambda K\gamma p}{2\alpha-1}.$$

Now using $2\mathbb{E}[f^* - f^0] \leq \Lambda\mathbb{E}[\|\theta_{\text{init}} - \theta^*\|_2^2]$, we get that

$$\mathrm{KL}\left(\mathbb{P}_L^{\theta_{\text{init}},T} || \mathbb{P}_{\tilde{L}}^{\theta_{\text{init}},T}\right) \leq \frac{\Lambda^2\gamma\alpha}{12(2\alpha-1)}(\|\theta_{\text{init}} - \theta^*\|_2^2 + 2K\gamma p) + \frac{pK\Lambda^2\gamma^2}{4}.$$

Hence, we recover the claim of Lemma 2. □

Using the bounds on both types of errors, we establish an upper bound on the error of approximating the posterior $\Pi(\theta|Z^{(n)})$ in the next theorem.

Theorem 2. *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a m-strongly concave function function with Λ-Lipschitz gradients and θ^* be its global maximum point. Assume that for some $\alpha \geq 1$, we have $\gamma \leq \frac{1}{\alpha\Lambda}$ and $K \geq \alpha$. Then, for any time horizon $T = K\gamma$, the total variation distance between the target distribution $\Pi(\theta|Z^{(n)})$ and its approximation $\nu\mathbb{P}_\vartheta^K$, the K-th iterate of the LMC algorithm with initial distribution $\nu \sim N(\theta^*, \Lambda^{-1}I_p)$ satisfies*

$$\|\nu\mathbb{P}_\vartheta^K - \Pi(\theta|Z^{(n)})\|_{TV} \leq \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{\Lambda}{m}\right) - \frac{Tm}{2}\right\} + \left\{\frac{p\Lambda^2T\gamma\alpha}{4(2\alpha-1)}\right\}^{1/2}. \tag{9}$$

Something to note here is that the right-hand side of (9) can go to infinity when the time horizon T goes to infinity with fixed step size γ. Hence, the upper bound is not sharp for large values of T.

Proof of Theorem 2. Using the triangle inequality, we have:

$$\|\nu\mathbb{P}_\vartheta^K - \Pi(\theta|Z^{(n)})\|_{TV} = \|\nu\mathbb{P}_{\tilde{L}}^{K\gamma} - \Pi(\theta|Z^{(n)})\|_{TV} \leq \|\nu\mathbb{P}_L^T - \Pi(\theta|Z^{(n)})\|_{TV} + \|\nu\mathbb{P}_{\tilde{L}}^T - \nu\mathbb{P}_L^T\|_{TV}.$$

The first term on the right-hand side is called the first type error originating from the finiteness of time, since if the time T were infinite, this error would be zero. The second term on the right-hand side is the second type error which is caused by discrete time we can simulate the diffusion in.

We can use Lemma 1 to get that

$$\|\nu\mathbb{P}_L^T - \Pi(\theta|Z^{(n)})\|_{TV} \leq \frac{1}{2}\left(\chi^2(\nu|\pi(\theta|Z^{(n)}))\right)^{1/2} e^{-Tm/2}.$$

Now we use Lemma 5 of [1] to upper bound the right-hand side by

$$\|\nu\mathbb{P}_L^T - \Pi(\theta|Z^{(n)})\|_{TV} \leq \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{\Lambda}{m}\right) - \frac{Tm}{2}\right\}.$$

For the second type error, we use (8) and Pinsker's inequality in [6](Problem 3.18a), that is for any measures $\mathbb{P}$, $\mathbb{Q}$,

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \leq \sqrt{2\mathrm{KL}(\mathbb{P}||\mathbb{Q})}, \tag{10}$$

and hence (9) follows. □

Corollary 1. *Let $p \geq 2$, f be as in Theorem 2, and $\epsilon \in (0, 1/2)$ be a target precision level. Let the time horizon T and the step-size γ be defined by*

$$T = \frac{4\log(1/\epsilon) + p\log(\Lambda/m)}{2m}, \qquad \gamma = \frac{\epsilon^2(2\alpha-1)}{\Lambda^2Tp\alpha}, \tag{11}$$

where $\alpha = (1 + \Lambda pT\epsilon^{-2})/2$. Then the output of the K-step LMC algorithm, with $K = \lceil T/\gamma \rceil$, satisfies $\|\nu\mathbb{P}_\vartheta^K - \Pi(\theta|Z^{(n)})\|_{TV} \leq \epsilon$.

Proof of Corollary 1. The choice of T and γ, implies that both the terms in the RHS of (9) are bounded by $\epsilon/2$. Additionally, we see that $\alpha = (1 + \Lambda p T \epsilon^{-2})/2 \geq 1$, and $\gamma = \frac{\epsilon^2(2\alpha-1)}{\Lambda^2 T p \alpha} \leq \frac{1}{\alpha\Lambda}$ and $K \geq \alpha$ and hence the result follows by Theorem 2. □

The value of α is intentionally chosen so that the factor $(2\alpha + 1)/\alpha$ is close to 2. It helps halve the running time of the algorithm as compared to a value of α close to 1.

Now we can use Corollary 1 to establish the polynomial time mixing guarantees and guidance for choosing the step size and number of iterations for the LMC algorithm for the strictly log-concave posterior to achieve a prescribed error rate. To achieve an error smaller then ϵ, it is enough to perform $K = O(T^2 p/\epsilon^2) = O(\epsilon^{-2}(p^3 + p\log^2(1/\epsilon)))$ evaluations of the gradient of f.

5 Extension to the non Strongly log-concave posterior

The theoretical guarantees established in the previous section are valid for strongly log-concave posteriors $\Pi(\theta|Z^{(n)})$ with lipschitz gradients. They can also be extended to nonstrongly log-concave posteriors by approximating the target density by a strongly log-concave density and applying the LMC algorithm to the latter instead of the former.

Lets assume we want to sample from $\pi(\theta|Z^{(n)}) \propto \exp f(\theta)$, where f is a concave function of θ, and has Λ-Lipschitz gradients. Assume that for every $R \in [0, \infty]$, there exists $m_R \geq 0$ such that $\nabla^2 f(\theta) \preceq -m_R I_p$ for every $\theta \in B = B_R(\theta_1) = \{\theta \in \mathbb{R}^p : \|\theta - \theta_1\|_2 \leq R\}$, where θ_1 is an arbitrarily fixed point in $\mathbb{R}^p$. If f is nonstrongly concave, then $m_\infty = 0$. We introduce a strongly concave log-density $\tilde{f}$, with a tuning parameter $\kappa > 0$

$$\tilde{f}(\theta) = f(\theta) - \frac{\kappa}{2}(\|\theta - \theta_1\|_2 - R)\mathbb{1}_{B^C}(\theta). \tag{12}$$

The function $\tilde{f}$ is $\tilde{m}$-strongly concave and has $\tilde{\Lambda}$-Lipschitz gradients with $\tilde{m} = \min(m_{2R}, m_\infty + 0.5\kappa)$ and $\tilde{\Lambda} = \Lambda + \kappa$. Let us denote the density $\tilde{\pi}(\theta) \propto \exp \tilde{f}(\theta)$. We would expect the measures $\Pi(\theta|Z^{(n)})$ and $\tilde{\Pi}(\theta|Z^{(n)})$ to be close when R is large and κ is small. The measure $\tilde{\Pi}(.|Z^{(n)})$ can be seen as the strong concave surrogate measure for which we have already established polynomial sampling guarantees in the previous section. We now formalise the distance between $\Pi(\theta|Z^{(n)})$ and $\tilde{\Pi}(\theta|Z^{(n)})$ in the next lemma.

Lemma 3. *Let f and $\tilde{f}$ be two functions such that $f(\theta) \geq \tilde{f}(\theta)$ for all $\theta \in \mathbb{R}^p$ and both e^f and $e^{\tilde{f}}$ are integrable. Then the Kullback-Leibler divergence between the distribution $\pi(\theta) \propto e^{f(\theta)}$ and $\tilde{\pi}(\theta) \propto e^{\tilde{f}(\theta)}$ can be upper bounded by*

$$\mathit{KL}(\Pi(\theta|Z^{(n)})||\tilde{\Pi}(\theta|Z^{(n)})) \leq \frac{1}{2}\int_{\mathbb{R}^p} (\tilde{f}(\theta) - f(\theta))^2 \pi(\theta|Z^{(n)}) d\theta. \tag{13}$$

Proof of Lemma 3. Using the formula for KL-divergence, we get

$$\mathrm{KL}(\Pi(\theta|Z^{(n)})||\tilde{\Pi}(.|Z^{(n)})) = \int_{\mathbb{R}^p} (f(\theta) - \tilde{f}(\theta))\pi(\theta|Z^{(n)})d\theta + \log \int_{\mathbb{R}^p} e^{\tilde{f}(\theta) - f(\theta)} \pi(\theta|Z^{(n)}) d\theta.$$

Now by using inequalities $\log u \leq u - 1$ and $e^{-u} \leq 1 - u + u^2/2$ for all $u \geq 0$, we can upper bound the second term in the right-hand side by

$$\begin{aligned}\log \int_{\mathbb{R}^p} e^{\tilde{f}(\theta)-f(\theta)} \pi(\theta|Z^{(n)}) d\theta &\leq \int_{\mathbb{R}^p} e^{\tilde{f}(\theta)-f(\theta)} \pi(\theta|Z^{(n)}) d\theta - 1 \\ &\leq -\int_{\mathbb{R}^p} (f(\theta) - \tilde{f}(\theta)) \pi(\theta|Z^{(n)}) d\theta + \frac{1}{2} \int_{\mathbb{R}^p} (\tilde{f}(\theta) - f(\theta))^2 \pi(\theta|Z^{(n)}) d\theta.\end{aligned}$$

Hence the claim of the lemma follows. □

For our construction of the strongly concave $\tilde{f}$, we have using Pinsker's inequality(10) that $\|\tilde{\Pi}(.|Z^{(n)}) - \Pi(.|Z^{(n)})\|_{TV} \leq \sqrt{\frac{1}{2} KL(\Pi(.|Z^{(n)})||\tilde{\Pi}(.|Z^{(n)}))} \leq \frac{\kappa}{4}\sqrt{\int_{B^C}(\|\theta - \theta_1\|_2 - R)^4 \pi(\theta|Z^{(n)}) d\theta}$. By choosing κ to be small enough, we can ensure that $\|\tilde{\Pi}(.|Z^{(n)}) - \Pi(.|Z^{(n)})\|_{TV} \leq \epsilon/2$. We can now establish a convergence result for a nonstrongly log-concave posterior using Theorem 2.

Corollary 2. *Let f be twice differentiable function satisfying $-\Lambda I_p \leq \nabla^2 f(\theta) \leq -m_R I_p$ for every $\theta \in B_R(\theta_1)$ and for every $R \in [0, \infty]$. Let $\epsilon \in (0, 1/2)$ be a target precision level. Assume that for some known value μ_R, we have that $\int_{B_R(\theta_1)^C} (\|\theta - \theta_1\| - R)^4 \pi(\theta|Z^{(n)}) d\theta \leq p^2 \mu_R^2$. Set $\kappa = \frac{2\epsilon}{p\mu_R}$. Set the time horizon T and the step-size γ as follows:*

$$T = \frac{4 \log(1/\epsilon) + p \log(\tilde{\Lambda}/\tilde{m})}{2\tilde{m}}, \qquad \gamma = \frac{\epsilon^2}{4\tilde{\Lambda}^2 T p}.$$

Then the output of the K-step LMC algorithm applied to the approximation $\tilde{f}$ as in (12), with $K = \lceil T/\gamma \rceil$, satisfies $\|\nu \mathbb{P}_{\vartheta}^K - \Pi(\theta|Z^{(n)})\|_{TV} \leq \epsilon$.

If we look at the result for $R = 0$, that is the strongly log-concave posterior case, we need $K = O(p^5 \epsilon^{-4} \log^2(\max(p, \epsilon^{-1})))$, which is significantly deteriorated than what we got in the previous section in terms of dependence on p and ϵ. Corollary 2 is more helpful in the case of concave functions f that are strongly concave in a neighborhood of their maximum point θ^*.

6 Further extension to non-log-concave posterior

To further extend the theory to more general problems, [2] and [3] extended the idea of surrogate posterior $\tilde{f}$ developed in the previous section for nonstronly log-concave densities. Since we do not have any concavity or local concavity in the general case, we can not establish anything affirmatively. Hence, we can only establish probabilistic bounds for the mixing time of the LMC algorithm. We demonstrate that with high probability under the law generating $Z^{(n)}$, the posterior target measure $\Pi(\theta|Z^{(n)})$ is locally log-concave on a region in $\mathbb{R}^p$ where most of its mass concentrates. The idea is to rely on the Fisher information to provide a natural statistical notion of curvature for the log-likelihood near the truth θ_0. We say that the posterior measure, contracts around a point $\theta_{*,p}$ with high probability. We then use the local log-concavity to define a surrogate posterior which is globally log-concave with lipschitz gradients. We then show that a localized Langevin-Type algorithm, when initialized into the region of log-concavity, possesses polynomial time convergence guarantees in high-dimensional models using results developed for strongly log-concave posteriors. The existence of a suitable initializer inside the region of local concavity is postulated, and finding one in polynomial time needs to be studied for different problems separately.

We study the example of generalised linear models and demonstrate the proof strategy for the same.

7 Generalised linear models

We establish the polynomial sampling guarantees in the generalised linear models(GLMs) setting as discussed in [3]. GLMs are extremely popular methods in statistical analysis that allow us to model different distributions of independent responses. They were first introduced by Nelder and Wedderburn in [7].

Let $\Theta \subset \ell^2(\mathbb{N})$ be a parameter space containing $\mathbb{R}^p$. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space equipped with a measure $\nu_{\mathcal{X}}$ and let ξ be a probability measure on $\mathbb{R}$. Denote the orthonormal basis of $L^2(\mathcal{X})$ as $(e_k)_{k\geq 1}$, and $\Phi(\theta) = \sum_{k=1}^{\infty} \theta_k e_k$ be a function in $L^2(\mathcal{X})$ with parameter $\theta = (\theta_k)_{k\geq 1}$ and let $g : \mathcal{I} \mapsto \mathcal{R}$ be an invertible and continuous link function on interval $\mathcal{I} \subset \mathbb{R}$. Set $\nu = \xi \otimes \nu_{\mathcal{X}}$

Suppose we have n independent observations $Z^{(n)}$ generated from $\mathbb{P}_\theta^n = \otimes_{i=1}^n \mathbb{P}_\theta$, $\theta \in \Theta$, and the law of responses Y_i follows a one-parameter exponential family with $g(\mathbb{E}[Y_i|X_i]) = \Phi(\theta)(X_i)$. Let $p_{\mathcal{X}}$ denote the $\nu_{\mathcal{X}}$ densities of the covariates X_i, then the ν-densities $p_\theta : \mathbb{R} \times \mathcal{X} \mapsto \mathbb{R}$ are given by

$$p_\theta(y, x) = \exp(yb(\theta)(x) - A(b(\theta)(x)))p_{\mathcal{X}}(x), \tag{14}$$

where $A(t) = \log \int_{\mathbb{R}} e^{yt} d\xi(y)$ and the natural parameter is given by the function

$$b(\theta) = (A')^{-1} \circ g^{-1} \circ \Phi(\theta). \tag{15}$$

To perform posterior inference for GLMs, we assume a certain regularity α with the true parameter $\theta_0 \in h^\alpha(\mathbb{N})$ where $h^\alpha(\mathbb{N}) = \{\theta \in \ell^2(\mathbb{N}) : \|\theta\|_\alpha^2 = \sum_{k=1}^{\infty} k^{2\alpha}\theta_k^2 < \infty\}$. We assume a popular 'sieve' prior on the first p coefficients of θ_0 given by

$$\theta \in \mathbb{R}^p, \theta \sim \Pi = N(0, n^{-1/(2\alpha+1)}\Sigma_\alpha), \Sigma_\alpha = \text{diag}(1, 2^{-2\alpha}, ..., p^{-2\alpha}). \tag{16}$$

It assigns decreasing variances to the first p components of θ_0. The log prior density is given by

$$\log \pi(\theta) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\det(n^{-1/(2\alpha+1)}\Sigma_\alpha) - \frac{1}{2}n^{1/(2\alpha+1)}\theta^T\Sigma_\alpha^{-1}\theta.$$

Also the gradient $\nabla \log \pi(\theta)$ is given by

$$\nabla \log \pi(\theta) = -n^{1/2\alpha+1}\Sigma_\alpha^{-1}\theta.$$

Now to check for concavity, we need to show that for $\theta, \theta' \in \mathbb{R}^p$, there exists $m_\pi > 0$ such that

$$\begin{aligned}
0 &\leq \log \pi(\theta) - \log \pi(\theta') + (\theta' - \theta)^T \nabla \log \pi(\theta) - \frac{m_\pi}{2}\|\theta - \theta'\|^2, \\
&\leq -\theta^T\Sigma_\alpha^{-1}\theta - \theta'^T\Sigma_\alpha^{-1}\theta' + 2(\theta - \theta')\Sigma_\alpha^{-1}\theta - m_\pi n^{-1/(2\alpha+1)}\|\theta - \theta'\|^2, \\
&\leq \|(\theta - \theta')\Sigma_\alpha^{-1/2}\|^2 - m_\pi n^{-1/(2\alpha+1)}\|\theta - \theta'\|^2.
\end{aligned}$$

This gives us $m_\pi n^{-1/(2\alpha+1)} \leq 1$, that is $m_\pi = n^{1/(2\alpha+1)}$. Also, since $\|\nabla \log \pi(\theta) - \nabla \log \pi(\theta')\| \leq p^{2\alpha} n^{1/(2\alpha+1)} \|\theta - \theta'\|$ (Since Σ_α^{-1} is a diagonal matrix with diagonal entries in $[1, p^{2\alpha}]$). We get that the prior log density $\log \pi(\theta)$ is m_π-strongly concave and has Λ_π-Lipschitz gradients with $m_\pi = n^{1/(2\alpha+1)}$, $\Lambda_\pi = n^{1/(2\alpha+1)} p^{2\alpha}$. The posterior measure $\Pi(\theta|Z^{(n)})$, $\theta \in \mathbb{R}^p$ arising from the data $Z^{(n)}$ using Bayes formula is given by:

$$\pi(\theta|Z^{(n)}) = \frac{\Pi_{i=1}^n p_\theta(Y_i, X_i)\pi(\theta)}{\int_\Theta \Pi_{i=1}^n p_\theta(Y_i, X_i)\pi(\theta) d\theta} \propto \exp\left(\ell_n(\theta) - n^{1/(2\alpha+1)} \|\theta\|_\alpha^2/2\right). \tag{17}$$

Note here that the posterior measure is known up to a scalar. The log-likelihood is independent of $p_\mathcal{X}$ and up to θ independent additive constants, given by

$$\ell_n(\theta) = \sum_{i=1}^n (Y_i b(\theta)(X_i) - A(b(\theta)(X_i))). \tag{18}$$

The log posterior is given as (up to scaling factors and θ independent terms) $\ell_n(\theta) - n^{1/(2\alpha+1)} \|\theta\|_\alpha^2/2$ and is strongly concave if ℓ_n is, and has lipschitz gradients if ℓ_n does. Unfortunately, these properties do not hold generally.

7.1 Local Curvature

We determine a high-dimensional and statistically informative set of parameters where the curvature of ℓ_n can be quantified depending on n and p. We consider the ball $\mathcal{B}$ of radius η around $\theta_{*,p} = (\theta_{0,1}, ..., \theta_{0,p})$, the $\mathbb{R}^p$-projection of $\theta_{*,p}$, which we assume to be a good approximation of θ_0, that is

$$\mathcal{B} = \{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \leq \eta\},$$

where $0 < \eta \leq 1$. We will establish later in Property 2 that the eigenvalues of $-\nabla^2 \ell_n(\theta)$ up to a scalar factor lie in the interval $[n, np^{1/2}]$ on $\mathcal{B}$ on the event ϱ, with high $\mathbb{P}_{\theta_0}^n$-probability as soon as the map b in (15) is uniformly bounded on $\mathcal{B}$. We also assume that we can initialize the LMC algorithm inside $\mathcal{B}$, that is $\theta_{\text{init}} \in \mathcal{B}$ using some grid-search algorithm. We hence have a region $\mathcal{B}$, where we have a local log-concavity with high probability. We use this region to construct the surrogate posterior in the next subsection.

We now state some sufficient conditions under which we have local log-concavity and polynomial mixing times for the LMC iterates with high probability.

Condition 1. *Suppose that $\theta_0 \in h^\alpha(\mathbb{N})$, $\alpha > 1$ and let $p \leq Cn^{1/(2\alpha+1)}$, $C > 0$. The radius of $\mathcal{B}$ is $\eta = p^{-1/2}$, and $\theta_{*,p}$, θ_{init} are such that*

$$\|\theta_0 - \theta_{*,p}\| \leq c_0 n^{-\alpha/(2\alpha+1)}, \; \|\theta_{init} - \theta_{*,p}\| \leq \eta/8,$$

for $c_0 > 0$. The design is bounded in the sense that $c_\mathcal{X}^{-1} \leq p_\mathcal{X}(x) \leq c_\mathcal{X}$ for all $x \in \mathcal{X}$ and some $c_\mathcal{X} > 0$, and the basis functions are uniformly bounded, i.e. $\sup_{k \geq 1} \sup_{x \in \mathcal{X}} |e_k(x)| \leq c_\mathcal{X}$.

7.2 The surrogate Posterior

To prove the results for a non-concave log-likelihood, we first construct a globally concave log-likelihood $\tilde{\ell}_n : \mathbb{R}^p \mapsto \mathbb{R}$ such that it agrees with ℓ_n on the set

$$\begin{aligned} \tilde{\mathcal{B}} &= \{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \leq 3\eta/8\} \subset \mathcal{B}, \\ \tilde{\ell}_n(\theta) &= \nu(\|\theta - \theta_{\text{init}}\|/\eta)(\ell_n(\theta) - \ell_n(\theta_{\text{init}})) + \ell_n(\theta_{\text{init}}) - K\nu_\eta(\|\theta - \theta_{\text{init}}\|), \end{aligned} \tag{19}$$

with $K > 0$ and the 'cut-off function' $\nu : \mathbb{R}^p \mapsto [0, 1]$ and the globally convex function $\nu_\eta : \mathbb{R}^p \mapsto [0, \infty)$, $\nu_\eta(t) = (\phi_{\eta/8} * \gamma_\eta)(t)$, where * is the convolution product, and where $\nu(t) = \mathbb{1}_{\{t\leq 3/4\}}(t)$ and $\gamma_\eta(t) = (t - 5\eta/8)^2 \mathbb{1}_{\{t\geq 5\eta/8\}}(t)$. We have $\phi_u(x) = \phi(x/u)/u$, $u > 0$, is a mollifier for some smooth function $\phi : \mathbb{R} \mapsto [0, \infty)$ with support in $[-1, 1]$, satisfying $\phi(-x) = \phi(x)$, $\int_{\mathbb{R}} \phi(x)dx = 1$. Using the surrogate log-likelihood, we define the surrogate posterior measure $\tilde{\Pi}(\theta|Z^{(n)})$ with density

$$\tilde{\pi}(\theta|Z^{(n)}) = \frac{e^{\tilde{\ell}_n(\theta)\theta}}{\int_\Theta e^{\tilde{\ell}_n(\theta)\pi(\theta)d\theta}} \propto e^{\tilde{\ell}_n(\theta)}\pi(\theta). \tag{20}$$

As we will see in the proof of the next theorems, particularly in Theorem 7, the surrogate posterior agrees with the posterior on $\tilde{\mathcal{B}}$ and is $\tilde{m}$-strongly concave with $\tilde{\Lambda}$-Lipschitz gradients with high $\mathbb{P}^n_{\theta_0}$-probability, with $\tilde{m} \asymp n$ and $\tilde{\Lambda} \asymp np^{1/2}$ (on the event ϱ where we have local concavity of the log-likelihood). We can then use the existing results from [8], [1] (previous sections), and [2] for strongly log-concave posteriors, to establish polynomial time convergence of LMC iterates $(\tilde{\vartheta}_k)_{k\in\mathbb{N}}$ to the surrogate posterior measure $\tilde{\Pi}$,

$$\begin{aligned} \tilde{\vartheta}_{k+1} &= \tilde{\vartheta}_k + \gamma\nabla \log \tilde{\pi}(\tilde{\vartheta}_k|Z^{(n)}) + \sqrt{2\gamma}\zeta_{k+1}, \qquad \tilde{\vartheta}_0 = \theta_{\text{init}}, \\ &= \tilde{\vartheta}_k + \gamma\left(\nabla\tilde{\ell}_n(\tilde{\vartheta}_k) - n^{1/(2\alpha+1)}\Sigma_\alpha\tilde{\vartheta}_k\right) + \sqrt{2\gamma}\zeta_{k+1}. \end{aligned} \tag{21}$$

Since the posterior measure puts most of its mass in $\tilde{\mathcal{B}}$, and ℓ_n and $\tilde{\ell}_n$ agree on this set, we expect the invariant measure of the Markov chain $(\tilde{\vartheta}_k)_{k\in\mathbb{N}}$ to be close to the true posterior measure $\Pi(\theta|Z^{(n)})$, and the global concavity gives us fast mixing.

Example 3. *The link function g is called the canonical link function if $g = (A')^{-1}$. In this case, $b(\theta) = \Phi(\theta)$, is linear in θ. Since A is convex, $\ell_n(\theta)$ turns out to be concave. For logistic regression with $A(x) = \log(1 + e^x)$.*

$$\ell_n(\theta) = \sum_{i=1}^{n}(Y_i\Phi(\theta)(X_i) - \log(1 + e^{\Phi(\theta)(X_i)})).$$

Its partial derivative is given as

$$\frac{\partial\ell_n(\theta)}{\partial\theta_j} = \sum_{i=1}^{n}\left(Y_i - \frac{1}{1 + e^{-\Phi(\theta)(X_i)}}\right)e_j(X_i).$$

And,

$$A''(x) = \frac{e^x}{(1 + e^x)^2} \leq \frac{1}{4}.$$

Now for $\theta_1, \theta_2 \in \mathbb{R}^p$, we have

$$\begin{aligned}
|\nabla \ell_n(\theta_1) - \nabla \ell_n(\theta_2)| &= \sum_{j=1}^{p} \sum_{i=1}^{n} (A'(b(\theta_2)(X_i)) - A'(b(\theta_1)(X_i))) e_j(X_i) \\
&\leq \frac{1}{4} \sum_{i=1}^{n} |b(\theta_1)(X_i) - b(\theta_2)(X_i)| \\
&\leq \frac{n c_{\mathcal{X}}}{4} \|\theta_1 - \theta_2\|_1 \\
&\leq \frac{n\sqrt{p} c_{\mathcal{X}}}{4} \|\theta_1 - \theta_2\|_2,
\end{aligned}$$

using Condition 1, $\|\theta_1 - \theta_2\|_1 \leq \sqrt{p}\|\theta_1 - \theta_2\|_2$ and that A' is lipschitz with $\|A'\|_{lip} = 1/4$. Hence, $\ell_n(\theta)$ is strongly concave, and $\nabla \ell_n(\theta)$ is uniformly Lipschitz with $\Lambda = \frac{n\sqrt{p}c_{\mathcal{X}}}{4}$, since A is strongly concave and A'' is bounded. For a non-canonical link function, we can see that these properties would not hold.

7.3 Nonasymptotic bounds on the error of the LMC algorithm

We now state and prove Theorems 4-6 of [3]. We denote the law of the Markov chain by $\mathbb{P}$. We establish exponential concentration inequalities for the approximation of the posterior function by ergodic averages under $\mathbb{P}$, for a sufficiently small step-size γ, a burn-in time $J_{\text{in}} \geq 1$ and a precision level $\epsilon > 0$. The precision ϵ is bounded from below due to the discretization of the continuous-time Langevin diffusion. The results hold non-asymptotically but are more informative asymptotically as $n \to \infty$.

Theorem 4. *Suppose data arise in a GLM with coordinate densities (14) and $g \in C^3(\mathcal{I})$. Let Condition 1 be satisfied and let $(\tilde{\vartheta}_k)_{k\geq 1}$ be the Markov Chain with iterates (21). For $c > 0$ suppose that $\gamma \leq cn^{-1}p^{-\frac{1}{2}}$,*

$$\epsilon \geq c \max(e^{-n^{1/(2\alpha+1)}/2}, \gamma^{1/2}p, \gamma p^{3/2} n^{1/2}) \ , \ J_{in} = \tfrac{\log(c\epsilon^2)}{\log(1-cn\gamma)},$$

Then there exist $c_1, c_2, c_3 > 0$ such that for all $J \geq 1$ with $\mathbb{P}^n_{\theta_0}$ - probability at least $1 - c_2\exp(-c_1 n^{1/(2\alpha+1)})$ the following holds:

1. *We have*
$$W_2\left(\mathcal{L}(\tilde{\vartheta}_{J+J_{in}}), \Pi(\theta|Z^{(n)})\right) \leq \epsilon.$$

2. *For any lipschitz function $f : \mathbb{R}^p \mapsto \mathbb{R}$, with $\|f\|_{Lip} = 1$,*
$$\mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{in}}^{J+J_{in}} f(\tilde{\vartheta}_k) - \int_{\Theta} f(\theta) d\Pi(\theta|Z^{(n)})\right| > \epsilon\right) \leq 2\exp\left(-c_3 \frac{\epsilon^2 n^2 J\gamma}{1 + 1/(nJ\gamma)}\right).$$

The result from Theorem 4 implies that if γ^{-1} depends polynomially on p, n, then the number of iterations necessary to approximate posterior functions at precision ϵ, grows at most polynomially in n, p and ϵ^{-1}, that is $J + J_{\text{in}} = O(n^{\rho}p^{\rho'}\epsilon^{\rho''})$, $\rho, \rho', \rho'' > 0$ with high $\mathbb{P}^n_{\theta_0} \times \mathbb{P}$-probability as

$$\mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{\text{in}}}^{J+J_{\text{in}}} f(\tilde{\vartheta}_k) - \int_\Theta f(\theta)d\Pi(\theta|Z^{(n)})\right| > \epsilon\right) \leq \alpha,$$

would hold if,

$$\begin{aligned}
2\exp\left(-c_3\frac{\epsilon^2 n^2 J\gamma}{1+1/(nJ\gamma)}\right) &\geq \alpha, \\
\implies \frac{\epsilon^2 n^2 J\gamma}{1+1/(nJ\gamma)} &\geq c'_\alpha, \\
\implies J^2\epsilon^2 n^2\gamma - Jc'_\alpha - \frac{c'_\alpha}{n\gamma} &\geq 0, \\
\implies J &\geq \frac{c'_\alpha + \sqrt{{c'_\alpha}^2 + 4\epsilon^2 n c'_\alpha}}{2\epsilon^2 n^2\gamma}. \qquad (22)
\end{aligned}$$

Since the right-hand side of (22) is polynomial in n, p, ϵ, as long as γ^{-1} depends polynomially on n, p, the above argument holds. As contrasted with Corollary 1, the results here hold with high $\mathbb{P} \times \mathbb{P}^n_{\theta_0}$-probability, instead of being definite. Also, the dependence on ϵ^{-2} still holds, while we have a dependence on n as well, which we didn't have before.

Another similarity between Theorem 4 and Theorem 2 is the $\gamma \leq cn^{-1}p^{-1/2}$ and $\gamma \leq 1/(\alpha\Lambda)$ since $\Lambda \asymp np^{1/2}$ as we will show later in Theorem 7.

Theorem 5. *Under the assumptions of Theorem 4 suppose that* $\gamma \leq cn^{-1}p^{-1}$, $J_{out} = c'e^{c''n^{1/(2\alpha+1)}}$ *for* $c', c'' > 0$ *and* $p \leq C(\log n)^{-(2\alpha+1/2)} n^{\frac{2\alpha}{2\alpha+1}\cdot\frac{1}{2\alpha+1/2}}$. *Then there exist* $c_1, c_2, c_3, c_4 > 0$ *such that with* $\mathbb{P}^n_{\theta_0}$*-probability at least* $1 - c_2\exp(-c_1 n^{1/(2\alpha+1)})$ *for all* $J + J_{in} \leq J_{out}$

$$\mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{in}}^{J+J_{in}} f(\vartheta_k) - \int_\Theta f(\theta)d\Pi(\theta|Z^{(n)})\right| > \epsilon\right)$$

$$\leq c_4\exp\left(-c_3\min\left(\frac{\epsilon^2 m^2 J\gamma}{1+1/(mJ\gamma)}, \frac{n^{1/2\alpha+1}}{1-e^{-m(J+J_{in})\gamma}}, \frac{n^{1/(2\alpha+1)}}{\gamma\Lambda^2/m}\right)\right).$$

Now using $f(\theta) = \theta$, in Theorem 5, we get the next Theorem.

Theorem 6. *Under the assumptions of Theorem 5 there exist* $c_1, c_2, c_3 > 0$ *such that with* $\mathbb{P}^n_{\theta_0} \times \mathbb{P}$*-probability at least* $1 - c_2\exp(-c_1 n^{1/(2\alpha+1)})$ *for all* $J + J_{in} \leq J_{out}$

$$\|\frac{1}{J}\sum_{k=1+J_{in}}^{J+J_{in}} \vartheta_k - \theta_0\| \leq c_3 n^{-\alpha/(2\alpha+1)} + \epsilon.$$

Note here that $n^{-\alpha/(2\alpha+1)}$ is the rate at which the p-dimensional approximation $\theta_{*,p}$ of θ_0 holds.

7.4 Proofs

We divide the proof into several steps. These steps are as follows:

1. The Posterior distribution 'contracts' around a p-dimensional projection of θ_0, $\theta_{*,p}$ at the rate $\delta_n = n^{-\alpha/(2\alpha+1)}$ (23), and a small ball condition holds for the 'normalizing factors' (24).

2. The posterior distribution satisfies the local boundedness and local curvature conditions with high probability on the event ϱ (Property 2).

3. The surrogate posterior agrees with the posterior on the region $\tilde{\mathcal{B}}$, is strongly concave and has Lipschtiz gradients on the high probability event ϱ (Theorem 7).

4. The surrogate posterior also contracts at the same rate δ_n around $\theta_{*,p}$ on the high-probability event $\tilde{\varrho}$ (Proposition 2).

5. The surrogate posterior is close to the posterior distribution in the sense of Wasserstein distance on the high probability event $\tilde{\varrho}$ (Theorem 8).

6. The LMC iterates for the surrogate posterior $(\tilde{\vartheta}_k)_{k\in\mathbb{N}}$ are close to the true functionals with high probability after a burn-in rate J_{in}, that is Theorem 4

7. The LMC iterates take on an average, exponentially in n many steps $J_{out} \gg J_{\text{in}}$ to exit the region $\mathcal{B}$ of local curvature on the event ϱ (Theorem 9).

8. The proof of Theorems 5 and 6 follows from Steps 6 and 7.

7.4.1 Steps 1 and 2

We define the posterior contraction step more formally in the form of Property 1 below:

Property 1. *The data $Z^{(n)}$ is distributed according to the law $\mathbb{P}^n_{\theta_0}$ with $\theta_0 \in \Theta$. There exist $c_0 > 0$ and $\theta_{*,p} \in \mathbb{R}^p$ with $\|\theta_{*,p}\| \leq c_0$ and a positive sequence $\delta_n \to 0$ such that for some $\beta \geq 1$, any $c > 0$ and any sufficiently large L there are $C_1, C_2, C_3, C_4 > 0$ with*

$$\mathbb{P}^n_{\theta_0}\left(\Pi\left(\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\|^\beta > L\delta_n | Z^{(n)}\right) \geq e^{-cn\delta_n^2}\right) \leq C_2 e^{-C_1 n\delta_n^2}, \tag{23}$$

$$\mathbb{P}^n_{\theta_0}\left(\int_{\|\theta-\theta_{*,p}\|\leq\delta_n} e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\theta)d\theta \leq e^{-C_3 n\delta_n^2}\right) \leq e^{-C_4 n\delta_n^2}. \tag{24}$$

If a posterior distribution follows (23), then we say that the Posterior distribution contracts around the point $\theta_{*,p}$ at the rate of δ_n.

Property 2. *There exist $0 < \eta \leq 1$, an event ϱ with $\mathbb{P}^n_{\theta_0}(\varrho) \geq 1 - c' e^{-cn\delta_n^2}$ for $c, c' > 0$, and a region $\mathcal{B} = \{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \leq \eta\}$ such that $\theta \mapsto \ell_n(\theta) \in C^2(\mathcal{B})$, $\mathbb{P}^n_{\theta_0}$ - almost surely and such that for $c_{\max} \geq c_{\min} > 0$, $\kappa_1, \kappa_2, \kappa_3 \geq 0$ the following holds on ϱ:*

1. *(local boundedness)* $\|\nabla \ell_n(\theta_{*,p})\| \leq c_{\max} n\delta_n p^{\kappa_1}$ *and* $\sup_{\theta\in\mathcal{B}} \|\nabla^2 \ell_n(\theta)\|_{op} \leq c_{\max} n p^{\kappa_2}$.

2. *(local curvature)* $\inf_{\theta\in\mathcal{B}} \lambda_{\min}(-\nabla^2 \ell_n(\theta)) \geq c_{\min} n p^{-\kappa 3}$.

Lemma 4. *Suppose the data arise in a GLM with coordinate densities (14), and Condition 1 is satisfied. Then Properties 1 and 2 are satisfied with $\delta_n = n^{-\alpha/(2\alpha+1)}$, $\beta = 1$, $\kappa_1 = 0$, $\kappa_2 = 1/2$, and $\kappa_3 = 0$.*

Proof of Lemma 4. Let us begin by noting a few properties of the maps $\Phi(\theta)$, $b(\theta)$, $A(t)$, $A'(t)$ etc. We first note that Φ is an isometry. This means that the range of Φ on a convex hull $R = \mathrm{conv}(\mathcal{B} \cup \{\theta_0\})$ and $x \in \mathcal{X}$, is a bounded subset of $\mathbb{R}$. Also, recall that, $b(\theta) = (A')^{-1} \circ g^{-1} \circ \Phi(\theta)$, and since A is smooth and convex, and $g \in C^3(\mathcal{I})$ is invertible, this means there exists a constant $M > 0$ such that $\sup_{\theta \in R} \|b(\theta)\|_{L^\infty} \leq M$, $\sup_{\theta \in R} \|f(b(\theta))\|_{L^\infty} \leq M$ for $f \in \{A, A', A'', A'''\}$ and $\inf_{\theta \in R} \|A''(b(\theta))\|_{L^\infty} \geq 1/M$ (since A is convex).

We write shorthand $b_\theta^x = b(\theta)(x)$ and $(X, Y) \sim (X_i, Y_i)$. Now the bounded design assumption in Condition 1 implies that for $\theta, \theta' \in \ell^2(\mathbb{N})$

$$\|b(\theta) - b(\theta')\|_{L^2}^2 \lesssim \mathbb{E}_{\theta_0}(b_\theta^X - b_{\theta'}^X)^2 \lesssim \|b(\theta) - b(\theta')\|_{L^2}^2, \tag{25}$$

Since the design of the covariates is bounded from both sides. We also know that since Φ is an isometry, the following hold for any θ, θ': $\|\Phi(\theta) - \Phi(\theta')\|_{L^2} = \|\theta - \theta'\|_2$ For any $v \in \mathbb{R}^p$, with $\|v\| = 1$,

$$\begin{gathered}\|v^T \nabla \Phi(\theta)\|_{L^\infty} = \|\sum_{k=1}^p v_k e_k\|_{L^\infty} \lesssim \|v\|_{L^1} \leq p^{1/2}\|v\| = p^{1/2} \\ 1 \lesssim \|v^T \nabla \Phi(\theta)\|_{L^2} = \|\sum_{k=1}^p v_k e_k\|_{L^2} \lesssim 1\end{gathered} \tag{26}$$

and $\nabla^2 \Phi(\theta) = 0$. Since, $\theta \mapsto \Phi(\theta)(x) \in C^2(B)$ for every $x \in \mathcal{X}$ implies that $\theta \mapsto b_\theta^x \in C^2(\mathcal{B})$. Also since $(A')^{-1} \circ g^{-1}$ is smooth on a compact set $\mathcal{B}$, implies that the properties in (26), transfer to the map b, that is they hold with Phi replaced with b. Also, for $\theta, \theta' \in \ell^2(\mathbb{N})$ with $\|\theta\|_\alpha, \|\theta'\|_\alpha \leq r$, there exists $c_r > 0$, such that $\|b(\theta)\|_{L^\infty}, \|b(\theta')\|_{L^\infty} \leq c_r$, such that

$$\begin{gathered}c_r^{-1}\|\theta - \theta'\| \leq \|b(\theta) - b(\theta')\|_{L^2} \leq c_r\|\theta - \theta'\|, \\ \|b(\theta_0) - b(\theta_{*,p})\| \leq c_0 \delta_n.\end{gathered} \tag{27}$$

We can also see that

$$\begin{gathered}\|\nabla^2 b(\theta)\|_{L^\infty(\mathcal{X}, \mathbb{R}^{p \times p})} \lesssim p, \\ \|\nabla^2 b(\theta) - \nabla^2 b(\theta')\|_{L^\infty(\mathcal{X}, \mathbb{R}^{p \times p})} \lesssim p^{3/2}\|\theta - \theta'\|, \\ \|v^T \nabla^2 b(\theta) v\| \lesssim 1.\end{gathered} \tag{28}$$

Also the likelihood $\ell_n(\theta)$ satisfies the following

$$\begin{aligned}\ell(\theta) &= Y b_\theta^X - A(b_\theta^X), \\ \mathbb{E}_{\theta_0} Y &= \mathbb{E}_{\theta_0} A'(b_\theta^X), \\ v^T \nabla \ell(\theta) &= (Y - A'(b_\theta^X)) v^T \nabla b_\theta^X, \\ v^T \nabla^2 \ell(\theta) v &= (Y - A'(b_\theta^X)) v^T \nabla^2 b_\theta^X v - A''(b_\theta^X)(v^T \nabla b_\theta^X)^2.\end{aligned} \tag{29}$$

To prove the posterior contraction for the GLM case, we are going to use Theorem 40 from [3]. We need to verify the assumptions of Theorem 40, namely

(i) For all $\theta \in \mathcal{B}_{n,r} = \{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \le \delta_n, \|\theta\|_\alpha \le r\}$ and all $q \ge 2$

$$\mathbb{E}_{\theta_0}(\ell(\theta_0) - \ell(\theta)) \le c_r \delta_n^2, \ \mathbb{E}_{\theta_0}|\ell(\theta_0) - \ell(\theta)|^q \le (q!/2)\delta_n^2 c_r^q.$$

(ii) For all $\theta, \theta' \in \ell^2(\mathbb{N})$ with $\|\theta\|_\alpha, \|\theta'\|_\alpha \le r$, there exists $\beta \ge 1$,

$$c_r^{-1}\|\theta - \theta'\|^\beta \le h(\theta, \theta') \le c_r\|\theta - \theta'\|,$$

where $h(\theta, \theta')$ is the Hellinger distance given by

$$h^2(\theta, \theta') = \int_{\mathcal{Z}} \left(\sqrt{p_\theta(z)} - \sqrt{p_{\theta'}(z)}\right)^2 d\nu(z).$$

It is a type of f-divergence used to quantify the similarity between two probability distributions.

Proof of (i). For the first inequality, we use a Taylor expansion of A at b_θ^X, and using (29), we have

$$\begin{aligned}
|\mathbb{E}_{\theta_0}(\ell(\theta_0) - \ell(\theta))| &= |\mathbb{E}_{\theta_0}\left(A'(b_{\theta_0}^X))(b_{\theta_0}^X - b_\theta^X) + A(b_\theta^X) - A(b_{\theta_0}^X)\right)| \\
&= \left|\mathbb{E}_{\theta_0}\left(A''(b_{\theta'}^X)\frac{(b_{\theta_0}^X - b_\theta^X)^2}{2}\right)\right| \\
&\le \frac{M}{2}\mathbb{E}_{\theta_0}(b_{\theta_0}^X - b_\theta^X)^2 \\
&\lesssim \|b(\theta_0) - b(\theta)\|_{L^2}^2 \qquad \text{(Using (25))} \\
&\le \|b(\theta_0) - b(\theta_{*,p})\|_{L^2}^2 + \|b(\theta_{*,p}) - b(\theta)\|_{L^2}^2 \text{Using Triangle Inequality} \\
&\lesssim \|b(\theta_0) - b(\theta_{*,p})\|_{L^2}^2 + \|\theta_{*,p} - \theta\|_{L^2}^2 \lesssim \delta_n^2.
\end{aligned}$$

Using (27) and $\theta \in \mathcal{B}$) in the last line. For the second part of the inequality, let $\lambda \in \mathbb{R}$ be sufficiently small such that

$$\mathbb{E}_{\theta_0}[\exp(\lambda Y)|X] = \exp(\lambda \mathbb{E}_{\theta_0} Y) = \exp(\lambda A'(b_{\theta_0}^X)) = \exp(A(\lambda + b_{\theta_0}^X) - A(b_{\theta_0}^X)) \le c,$$

for some $c > 0$. This implies $\mathbb{E}_{\theta_0}[\exp(\lambda|Y|)|X] \le 2c$. Hence, we can upper bound the moments of $|Y|$ using some λ depending constant $c_\lambda \ge 0$

$$\mathbb{E}_{\theta_0}[|Y|^q|X] \le c_\lambda^q, \tag{30}$$

for $q \ge 2$.

$$\begin{aligned}
\mathbb{E}_{\theta_0}|Y(b_{\theta_0}^X - b_\theta^X)|^q &= \mathbb{E}_{\theta_0}\left[\mathbb{E}_{\theta_0}[|Y(b_{\theta_0}^X - b_\theta^X)|^q|X]\right] \\
\le c_\lambda^q \mathbb{E}_{\theta_0}|b_{\theta_0}^X - b_\theta^X|^q &\lesssim c_\lambda^q\|b_{\theta_0}^X - b_\theta^X\|_{L^2}^2 \lesssim c_\lambda^q \delta_n^2 \qquad \text{(Using 25)}
\end{aligned}$$

□

Proof of (ii). The hellinger distance $h(\theta, \theta')$ can be rewritten as

$$\begin{aligned}
h^2(\theta, \theta') &= \int_{\mathcal{Z}} (p_\theta(z) + p_{\theta'}(z) - 2\sqrt{p_\theta(z)p_{\theta'}(z)})d\nu(z) \\
&= 2 - \int_{\mathcal{Z}} 2\sqrt{p_\theta(z)p_{\theta'}(z)}d\nu(z) \\
&= 2(1 - \mathbb{E}_{\theta_0}e^{-x}),
\end{aligned}$$

where $x = (A(b_\theta^X) + A(b_{\theta'}^X))/2 - A(\frac{b_\theta^X + b_{\theta'}^X}{2}) \geq 0$ by the convexity of A. We can rewrite

$$\begin{aligned} x &= \frac{A(b_\theta^X) - A(\frac{b_\theta^X + b_{\theta'}^X}{2})}{2} + \frac{A(b_{\theta'}^X) - A(\frac{b_\theta^X + b_{\theta'}^X}{2})}{2} \\ &= (b_\theta^X - b_{\theta'}^X)/4 \int_{t=0}^1 \left(A'(\frac{t}{2}(b_\theta^X - b_{\theta'}^X)) - A'(-\frac{t}{2}(b_\theta^X - b_{\theta'}^X)) \right) dt \\ &= (b_\theta^X - b_{\theta'}^X)^2/4 \int_{t=0}^1 t \int_{t'=0}^1 A'' \left(\left(\frac{1-t}{2} + tt' \right) (b_\theta^X - b_{\theta'}^X) \right) dt' dt. \end{aligned}$$

Hence, we have that $(b_\theta^X - b_{\theta'}^X)^2 \lesssim x \lesssim (b_\theta^X - b_{\theta'}^X)^2$, hence by (27), we have that $\|\theta - \theta'\|^2 \lesssim x \lesssim \|\theta - \theta'\|^2$. Now since $\|\theta\|_\alpha, \|\theta'\|_\alpha \leq r$, we have that for $0 \leq x \leq c'(c_r)$, by convexity of e^{-x},

$$e^{-x} \leq e^{-(x/c')c'} \leq (x/c')e^{-c'} + (1 - x/c')e^0 = \frac{e^{-c'} - 1}{c'} x + 1.$$

Using the above, we have that $h^2(\theta, \theta') \geq 2\frac{1-e^{-c'}}{c'} \mathbb{E}_{\theta_0} x$. Using a Taylor series expansion, we also have that $h^2(\theta, \theta') \leq 2\mathbb{E}_{\theta_0} x$. Hence, we get the desired result with $\beta = 1$.

□

To verify the local boundedness, and local curvature for the Posterior distribution, we will use Theorem 28 of [3], the proof of which is based on Bernstein Inequality and Metric entropy bounds. This boils down to checking some boundedness, curvature, and growth conditions, namely

(i) Growth Conditions

$$p \leq Cn\delta_n^2, \qquad C \max(\delta_n p^{\kappa_2}, \eta \delta_n p^{\kappa_4}, \delta_n^2 p^{\kappa_4}) \log n \leq p^{\kappa_3}$$

(ii) Local Mean Boundedness,

$$\begin{aligned} |\mathbb{E}_{\theta_0} v^T \nabla \ell(\theta_{*,p}, Z_i)| &\leq C_1 \delta_n, \\ \text{For all } q \geq 2 \quad \mathbb{E}_{\theta_0} |v^T \nabla \ell(\theta_{*,p}, Z_i)|^q &\leq (q!/2) C_1^q p^{(q-2)/2}, \\ \sup_{\theta \in \mathcal{B}} \mathbb{E}_{\theta_0} |v^T \nabla^2 \ell(\theta, Z_i) v|^q &\leq (q!/2) C_1^q p^{1+3(q-2)/2}, \\ \sup_{\theta, \theta' \in \mathcal{B}} \mathbb{E}_{\theta_0} |v^T \nabla^2 (\ell(\theta, Z_i) - \ell(\theta', Z_i)) v|^q &\leq (q!/2) C_1^q p^{3q/2} \|\theta - \theta'\|^q. \end{aligned}$$

(iii) Local Curvature,

$$\inf_{\theta \in \mathcal{B}} \lambda_{\min}(\mathbb{E}_{\theta_0}[-\nabla^2 \ell(\theta, Z_i)]) \geq C_2.$$

Proof of (i). We are going to verify the growth conditions for $\kappa_1 = 0, \kappa_2 = 1/2, \kappa_3 = 0, \kappa_4 = 3/2, \delta_n = n^{-\alpha/(2\alpha+1)}$.

By Condition 1, we have that for $\alpha > 1$

$$p \leq Cn^{1/(2\alpha+1)} = Cnn^{-2\alpha/(2\alpha+1)} = Cn\delta_n^2.$$

We also have

$$\delta_n p^{1/2} \log n \leq C n^{-\alpha/(2\alpha+1)} n^{1/2(2\alpha+1)} \log n = C n^{(1-2\alpha)/2(2\alpha+1)} \log n \leq C', \tag{31}$$

for some $C' > 0$, since $\alpha > 1$ from Condition 1. Also since $\eta = p^{-1/2}$,

$$\eta \delta_n p^{3/2} \log n = \delta_n p \log n \leq C n^{(1-\alpha)/(2\alpha+1)} \log n \leq C'', \tag{32}$$

for some $C'' > 0$. We also have

$$\delta_n^2 p^{3/2} \log n \leq n^{(3-4\alpha)/(2\alpha+1)} \log n \leq C''', \tag{33}$$

for some $C''' > 0$. Equations (31), (32) and (33) imply that

$$C \max(\delta_n p^{1/2}, \eta \delta_n p^{3/2}, \delta_n^2 p^{3/2}) \log n \leq 1.$$

□

Proof of (ii). Using the Equations (29)

$$\left| \mathbb{E}_{\theta_0} v^T \nabla \ell(\theta_{*,p}) \right| = \left| \mathbb{E}_{\theta_0} \left[(A'(b^X_{\theta_0}) - A'(b^X_{\theta_{*,p}})) v^T \nabla b^X_{\theta_{*,p}} \right] \right|.$$

Now by the Cauchy-Schwarz Inequality and by a Taylor Series expansion of $A'(b^X_\theta)$ at $b^X_{\theta_0}$, we see that

$$\left| \mathbb{E}_{\theta_0} v^T \nabla \ell(\theta_{*,p}) \right| \lesssim \| b^X_{\theta_0} - b^X_{\theta_{*,p}} \|_{L^2} \| v^T \nabla b(\theta_{*,p}) \|_{L^2} \lesssim \delta_n,$$

using the inequalities in (26) in the last line. For $q \geq 2$ and $\theta \in \mathcal{B}$, we have

$$\mathbb{E}_{\theta_0} | v^T \nabla \ell(\theta_{*,p}) |^q = \mathbb{E}_{\theta_0} | (Y - A'(b^X_{\theta_{*,p}})) v^T \nabla b^X_{\theta_{*,p}} |^q$$

Using (29) in the last line. Now since, $A'(b^X_{\theta_{*,p}})$ is bounded and using (30)

$$\begin{aligned}
\mathbb{E}_{\theta_0} | v^T \nabla \ell(\theta_{*,p}) |^q &\leq \tilde{c}^q_\lambda \mathbb{E}_{\theta_0} | v^T \nabla b^X_{\theta_{*,p}} |^q \\
&\leq \tilde{c}^q_\lambda (\| v^T \nabla b(\theta_{*,p}) \|_{L^\infty})^{q-2} (\| v^T \nabla b(\theta_{*,p}) \|_{L^2})^2 \\
&\lesssim \tilde{c}^q_\lambda p^{(q-2)/2},
\end{aligned}$$

using (26) in the last line. Similarly using (29) and (26)

$$\begin{aligned}
\mathbb{E}_{\theta_0} | v^T \nabla^2 \ell(\theta_{*,p}) v |^q &= \mathbb{E}_{\theta_0} | (Y - A'(b^X_{\theta_{*,p}})) v^T \nabla^2 b^X_{\theta_{*,p}} v - A''(b^X_{\theta_{*,p}}) (v^T \nabla b^X_{\theta_{*,p}})^2 |^q \\
&\leq \tilde{c}^q_\lambda (\mathbb{E}_{\theta_0} | v^T \nabla^2 b^X_{\theta_{*,p}} v |^q + \mathbb{E}_{\theta_0} | v^T \nabla b^X_{\theta_{*,p}} |^{2q}) \\
&\lesssim \tilde{c}^q_\lambda p^{q-2} \\
&\lesssim \tilde{c}^q_\lambda p^{1+3(q-2)/2}.
\end{aligned}$$

This implies that

$$\sup_{\theta \in \mathcal{B}} \mathbb{E}_{\theta_0} | v^T \nabla^2 \ell(\theta, Z_i) v |^q \lesssim \tilde{c}^q_\lambda p^{1+3(q-2)/2}.$$

We also have,

$$\mathbb{E}_{\theta_0}|v^T\nabla^2(\ell(\theta,Z_i)-\ell(\theta',Z_i))v|^q = \mathbb{E}_{\theta_0}\Big|(A'(b^X_{\theta'}) - A'(b^X_\theta))v^T\nabla^2 b^X_\theta v - (Y - A'(b^X_{\theta'}))v^T$$

$$(\nabla^2 b^X_\theta - \nabla^2 b^X_{\theta'})v + (A''(b^X_{\theta'}) - A''(b^X_\theta))(v^T\nabla b^X_\theta)^2 - A''(b^X_{\theta'})((v^T\nabla b^X_\theta)^2 - (v^T\nabla b^X_{\theta'})^2)\Big|^q$$

$$\begin{aligned}
&\lesssim c^q p^{q/2}\left(\mathbb{E}_{\theta_0}|v^T\nabla^2 b^X_\theta v|^q + \mathbb{E}_{\theta_0}|v^T\nabla b^X_\theta|^{2q}\right)\|\theta-\theta'\|^q + \tilde{c}^q_\lambda \mathbb{E}_{\theta_0}|v^T(\nabla^2 b^X_\theta - \nabla^2 b^X_{\theta'})v|^q \\
&\lesssim c^q p^{q/2}\left(\|v^T\nabla^2 b(\theta)v\|^{q-2}_{L^\infty}\|v^T\nabla^2 b(\theta)v\|^2_{L^2} + \|v^T\nabla b(\theta)\|^{2q-2}_{L^\infty}\|v^T\nabla b(\theta)\|^2_{L^2}\right)\|\theta-\theta'\|^q \\
&+ \tilde{c}^q_\lambda \mathbb{E}_{\theta_0}|v^T(\nabla^2 b^X_\theta - \nabla^2 b^X_{\theta'})v|^q \\
&\lesssim \tilde{c}^q_\lambda p^{3q/2}\|\theta-\theta'\|^q,
\end{aligned}$$

using (28) and (26) in the last line. Hence, we get that

$$\sup_{\theta,\theta'\in\mathcal{B}} \mathbb{E}_{\theta_0}|v^T\nabla^2(\ell(\theta,Z_i)-\ell(\theta',Z_i))v|^q \lesssim \tilde{c}^q_\lambda p^{3q/2}\|\theta-\theta'\|^q.$$

□

Proof of (iii). Using (29), we get

$$\mathbb{E}_{\theta_0}[v^T\nabla^2\ell(\theta)v] = \mathbb{E}_{\theta_0}[(A'(b^X_{\theta_0}) - A'(b^X_\theta))v^T\nabla^2 b^X_\theta v - A''(b^X_\theta)(v^T\nabla b^X_\theta)^2].$$

Again, using Taylor's expansion in the first term and Cauchy-Schwarz Inequality, we can upper-bound the expression by

$$\mathbb{E}_{\theta_0}[v^T\nabla^2\ell(\theta)v] \leq M(\|b(\theta_0)-b(\theta)\|_{L^2}\|v^T\nabla^2 b(\theta)v\|_{L^2} + \|v^T\nabla b(\theta)\|^2_{L^2}).$$

Now using (26), we know that:

$$\|b(\theta_0)-b(\theta)\|_{L^2} \leq \|b(\theta_0)-b(\theta_{*,p})\|_{L^2} + \|b(\theta_{*,p})-b(\theta)\|_{L^2} \lesssim \eta.$$

Using (28) and that $0<\eta<1$, we get

$$\begin{aligned}
\mathbb{E}_{\theta_0}[v^T\nabla^2\ell(\theta)v] &\lesssim \eta, \\
-\mathbb{E}_{\theta_0}[v^T\nabla^2\ell(\theta)v] &\geq C', \\
\inf_{\theta\in\mathcal{B}}\lambda_{\min}(\mathbb{E}_{\theta_0}[-\nabla^2\ell(\theta,Z_i)]) &\geq C_2.
\end{aligned}$$

□

□

7.4.2 Step 3

We now show the following properties of the surrogate posterior

Theorem 7. *Given $K \geq 60c_{\max}\|\nu\|_{C^2}n(1+p^{1/2})$, the following holds for the GLM surrogate posterior from (19) on the event ϱ (from Property 2):*

(i) $\ell_n(\theta) = \tilde{\ell}_n(\theta)$ *for all* $\theta \in \tilde{\mathcal{B}} = \{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \leq 3\eta/8\}$.

(ii) $\tilde{\ell}_n$ *is* $\tilde{m}$*-strongly concave and has* $\tilde{\Lambda}$*-Lipschitz gradients with* $\tilde{\Lambda} = 7K$ *and* $\tilde{m} = c_{min}n$.

In particular, the log surrogate posterior density $\nabla \log \tilde{\pi}(.|Z^{(n)})$ *is* m*-strongly concave and has* Λ*-Lipschitz gradients with* $\Lambda = 7K + \Lambda_\pi$, $m = c_{\min}n + m_\pi$

Proof of theorem 7. For Part (i), we know that from condition 1, $\|\theta_{\text{init}} - \theta_{*,p}\| \leq \eta/8$. Hence by triangle inequality on $\tilde{\mathcal{B}}$,

$$\begin{aligned}\|\theta - \theta_{\text{init}}\| &\leq \|\theta - \theta_{*,p}\| + \|\theta_{*,p} - \theta_{\text{init}}\| \leq \eta/2, \\ \tilde{\ell}_n(\theta) &= \nu(\|\theta - \theta_{\text{init}}\|)(\ell_n(\theta) - \ell_n(\theta_{\text{init}})) + \ell_n(\theta_{\text{init}}) - K\nu_\eta(\|\theta - \theta_{\text{init}}\|)\end{aligned}$$

Notice that , $\nu_\eta(\|\theta - \theta_{\text{init}}\|)$ vanishes and $\nu(\|\theta - \theta_{\text{init}}\|/\eta) = 1$ on $\tilde{\mathcal{B}}$, hence part (i) follows.

For Part (ii), we consider the event ϱ on which the Property 2 holds. Consider $V = \{\theta \in \mathbb{R}^p : \|\theta - \theta_{\text{init}}\| \leq 3\eta/4\} \subset \mathcal{B}$. Again on V, $\nu_\eta(\|\theta - \theta_{\text{init}}\|)$ vanishes and $\nu(\|\theta - \theta_{\text{init}}\|/\eta) = 1$, i.e. ℓ_n and $\tilde{\ell}_n$ agree. Hence we have (using property 2),

$$\inf_{\theta \in V} \lambda_{\min}\left(-\nabla^2 \tilde{\ell}_n(\theta)\right) \geq \inf_{\theta \in \mathcal{B}} \lambda_{\min}\left(-\nabla^2 \ell_n(\theta)\right) \geq c_{\min}n.$$

Using Lemma B.5 (which is a simple application of the chain rule) and the proof of Proposition B.6 (Equations 189 and 190) in [2], with $\lambda_{\max}(I) = 1$, we get that

$$\begin{aligned}\|\nabla\nu_\eta(\|\theta - \theta_{\text{init}}\|/\eta)\| &\leq \|\nu\|_{C^1}\eta^{-1}, \\ \|\nabla^2\nu(\|\theta - \theta_{\text{init}}\|/\eta)\|_{\text{op}} &\leq 4\|\nu\|_{C^2}\eta^{-2} \qquad \text{for all } \theta \in \mathbb{R}^p, \\ \lambda_{\min}(\nabla^2\nu_\eta(\|\theta - \theta_{\text{init}}\|)) &\geq 1/3, \\ \|\nabla^2\nu_\eta(\|\theta - \theta_{\text{init}}\|)\|_{\text{op}} &\leq 6.\end{aligned} \tag{34}$$

Now using an upper bound on the Taylor's expansion of $\ell_n(\theta)$, we get that

$$\begin{aligned}\sup_{\theta \in \mathcal{B}} |\ell_n(\theta) - \ell_n(\theta_{*,p})| &\leq \|\nabla\ell_n(\theta_{*,p})\|\eta + \sup_{\theta \in \mathcal{B}} \|\nabla^2\ell_n(\theta)\|_{\text{op}}\eta^2/2 \\ &\leq c_{\max}n(\delta_n\eta + p^{1/2}\eta^2/2),\end{aligned}$$

using Property 2 in the last line. Again using an upper bound on the Taylor's expansion of $\nabla\ell_n(\theta)$, we get

$$\begin{aligned}\sup_{\theta \in \mathcal{B}} \|\nabla\ell_n(\theta)\| &\leq \|\nabla\ell(\theta_{*,p})\| + \sup_{\theta \in \mathcal{B}} \|\nabla^2\ell_n(\theta)\|_{\text{op}}\eta \\ &\leq c_{\max}n(\delta_n + \eta p^{1/2}).\end{aligned}$$

Using Property 2 in the last line. Now by Triangle Inequality, we have

$$\begin{aligned}\sup_{\theta\in\mathcal{B}} |\ell_n(\theta) - \ell_n(\theta_{\text{init}})| &\le \sup_{\theta\in\mathcal{B}} |\ell_n(\theta) - \ell_n(\theta_{*,p})| + |\ell_n(\theta_{*,p}) - \ell_n(\theta_{\text{init}})| \\ &\le 2c_{\max} n(\delta_n \eta + p^{1/2}\eta^2/2).\end{aligned} \tag{35}$$

By an application of chain rule, triangle inequality, last 3 inequalities above and that $\nu_\eta(\|\theta-\theta_{\text{init}}\|/\eta)$ vanishes outside $\mathcal{B}$, we get

$$\|\nabla^2[\nu(\|\theta-\theta_{\text{init}}\|/\eta)(\ell_n(\theta) - \ell_n(\theta_{\text{init}}))]\|_{\text{op}}$$

$$\begin{aligned}&= \|\nabla^2\nu(\|\theta-\theta_{\text{init}}\|/\eta) + \nabla^2\ell_n(\theta) + 2\nabla\nu_\eta(\|\theta-\theta_{\text{init}}\|/\eta)\nabla\ell_n(\theta)\|_{\text{op}} \\ &\le \sup_{\theta\in\mathcal{B}}(\|\nabla^2\nu(\|\theta-\theta_{\text{init}}\|/\eta)\||\ell_n(\theta) - \ell_n(\theta_{\text{init}})| + 2\|\nabla\nu(\|\theta-\theta_{\text{init}}\|/\eta)\|\|\|\nabla\ell_n(\theta)\| \\ &\qquad + |\nu(\|\theta-\theta_{\text{init}}\|/\eta)|\|\nabla^2\ell_n(\theta)\|_{\text{op}}) \\ &\le \sup_{\theta\in\mathcal{B}}\left(4\|\nu\|_{C^2}\eta^{-2} \times 2c_{\max}n(\delta_n\eta + p^{1/2}\eta^2/2) + 2\|\nu\|_{C^1}\eta^{-1}c_{\max}n(\delta_n + \eta p^{1/2}) - \lambda_{\min}(-\nabla^2\ell_n(\theta))\|v\|_{C^2}\right) \\ &\le 8\|\nu\|_{C^2}c_{\max}n(\delta_n\eta^{-1} + p^{1/2}/2) + 2\|\nu\|_{C^1}c_{\max}n(\delta_n\eta^{-1} + p^{1/2}) \\ &\le 10c_{\max}\|\nu\|_{C^2}(\eta^{-1}\delta_n + p^{1/2}) \\ &\le K/6.\end{aligned} \tag{36}$$

Using $\|\nu\|_{C^1} \le \|\nu\|_{C^2}$, and condition 1 implying $\eta = p^{-1/2} \ge \delta_n$,

$$\begin{aligned}\inf_{\theta\in V^C}\lambda_{\min}\left(\nabla^2\tilde{\ell}_n(\theta)\right) &\ge \sup_{\theta\in\mathbb{R}^p}\|\nabla^2(\nu(\|\theta-\theta_{\text{init}}\|/\eta)\ell_n(\theta))\|_{\text{op}} + K\lambda_{\min}(\nabla^2\nu_\eta(\|\theta-\theta_{\text{init}}\|)) \\ &\ge \sup_{\theta\in\mathbb{R}^p}\|\nabla^2(\nu(\|\theta-\theta_{\text{init}}\|/\eta)\ell_n(\theta))\|_{\text{op}} + K/3 \ge K/6.\end{aligned}$$

Using (34) and 36 in the last line. Also since $K/6 \ge 10c_{\max}\|\nu\|_{C^2}n(1+p^{1/2}) \ge 10c_{\min}n\|\nu\|_{C^2} \ge c_{\min}n$, we have

$$\begin{aligned}\inf_{\theta\in\mathbb{R}^p}\left(-\nabla^2\tilde{\ell}_n(\theta)\right) &\ge \min\left(\inf_{\theta\in V}\left(-\nabla^2\tilde{\ell}_n(\theta)\right), \inf_{\theta\in V^C}\left(-\nabla^2\tilde{\ell}_n(\theta)\right)\right) \\ &\ge \min(c_{\min}n, K/6) \\ &\ge c_{\min}n.\end{aligned}$$

Thus, we have shown that the surrogate posterior is globally strongly log-concave with $\tilde{m} = c_{\min}n$. For the gradient-lipschitz bound, we have

$$\begin{aligned}\frac{\|\nabla(\tilde{\ell}_n(\theta) - \tilde{\ell}_n(\theta'))\|}{\|\theta-\theta'\|} &\le \sup_{\theta\in\mathbb{R}^p}\|\nabla^2\tilde{\ell}_n(\theta)\|_{\text{op}} \\ &\le \sup_{\theta\in\mathbb{R}^p}\|\nabla^2(\nu(\|\theta-\theta_{\text{init}}\|/\eta)\ell_n(\theta))\|_{\text{op}} + K\sup_{\theta\in\mathbb{R}^p}\|\nabla^2\nu_\eta(\|\theta-\theta_{\text{init}}\|)\|_{\text{op}} \\ &\le K/6 + 6K \\ &\le 7K.\end{aligned}$$

Using (34) in the second last line. □

7.4.3 Step 4

We prove the posterior contraction of the surrogate posterior in the following proposition.

Proposition 2. *For the GLM model, there exist for any $c > 0$ and any sufficiently large $L > 0$ constants $c_1, c_2 > 0$ with*

$$\mathbb{P}^n_{\theta_0}\left(\tilde{\Pi}\left(\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| > L\delta_n | Z^{(n)}\right) \geq e^{cn\delta_n^2}\right) \leq c_2 e^{-c_1 n\delta_n^2}.$$

An interesting thing to note here is that the proof of this Proposition does not rely on Hellinger tests as we used in the proof of Property 1. The proof only relies on the log-concavity of the surrogate posterior measure.

Proof of Proposition 2. Let $L \geq 1$ and define the ball

$$\mathcal{U} = \left\{\theta \in \mathbb{R}^p : \|\theta - \theta_{*,p}\| \leq L\delta_n\right\}.$$

Let $\mathcal{D}^C$ be the event in (24) and consider $\bar{\mathcal{D}} = \{\Pi(\mathcal{U}|Z^{(n)}) \geq 1 - e^{-cn\delta_n^2}/4\}$. Due to the posterior contraction of Π from Property 1, we know that events $\mathcal{D}$ and $\bar{\mathcal{D}}$ occur with high probability. Let us consider the high probability event $\tilde{\varrho} = \varrho \cap \mathcal{D} \cap \bar{\mathcal{D}}$. Now by taking n large enough, we can have

$$L\delta_n \leq (\log n)^{-1}\eta \leq 3\eta/8.$$

This implies by Theorem 7 (i), we have that

$$\ell_n(\theta) = \tilde{\ell}_n(\theta) \quad \forall \theta \in \mathcal{U}, \tag{37}$$

for $C_1 = L\frac{c_{\min}}{4c_{\max}}$ and $L \geq 4c_{\max}/c_{\min}$ such that $C_1 \geq 1$. Let $\theta \in \mathcal{U}^C$ such that $\|\theta - \theta_{*,p}\| > (4C_1 c_{\max}/c_{\min})\delta_n$.

From Theorem 7 (i), we see that $\ell_n(\theta_{*,p}) = \tilde{\ell}_n(\theta_{*,p})$ and $\nabla\ell_n(\theta_{*,p}) = \nabla\tilde{\ell}_n(\theta_{*,p})$ (As ℓ_n and $\tilde{\ell}_n$ agree on $\tilde{\mathcal{B}}$, and $\theta_{*,p} \in \tilde{\mathcal{B}}$, hence the limits should also agree on $\theta_{*,p}$). Now by a Taylor's expansion of $\tilde{\ell}_n$ around $\theta_{*,p}$ and Cauchy Schwarz inequality

$$\begin{aligned}\tilde{\ell}_n(\theta) - \tilde{\ell}_n(\theta_{*,p}) &\leq \|\ell_n(\theta_{*,p})\|\|\theta - \theta_{*,p}\| + 1/2\|\nabla^2\tilde{\ell}_n(\theta')\|\|\theta - \theta_{*,p}\|^2 \\ &\leq c_{\max} n\delta_n\|\theta - \theta_{*,p}\| - 1/2 c_{\min} n\|\theta - \theta_{*,p}\|^2 \\ &\leq (c_{\min}/4)n\|\theta - \theta_{*,p}\|^2 - (c_{\min}/2)n\|\theta - \theta_{*,p}\|^2 \\ &= -(1/4)c_{\min} n\|\theta - \theta_{*,p}\|^2 \\ &< -(4C_1^2 c_{\max}^2/c_{\min})n\delta_n^2.\end{aligned}$$

Hence, we have

$$\tilde{\ell}_n(\theta) \leq -(4C_1^2 c_{\max}^2/c_{\min})n\delta_n^2 + \ell_n(\theta_{*,p}). \tag{38}$$

Since $\ell_n(\theta)$ and $\tilde{\ell}_n(\theta)$ agree on $\mathcal{U}$ and $\theta : \|\theta - \theta_{*,p}\| \leq \delta_n \subset \mathcal{U}$, we get on the event $\varrho \cap \mathcal{D}$,

$$\begin{aligned}\tilde{\pi}(\theta|Z^{(n)}) &= \frac{e^{\tilde{\ell}_n(\theta)-\ell_n(\theta_0)}\pi(\theta)}{\int_\Theta e^{\tilde{\ell}_n(\theta)-\ell_n(\theta_0)}\pi(\theta)d\theta} \\ &\leq \frac{e^{\tilde{\ell}_n(\theta)-\ell_n(\theta_0)}\pi(\theta)}{\int_{\|\theta-\theta_{*,p}\|\leq\delta_n} e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\theta)d\theta} \\ &\leq e^{C_4 n\delta_n^2} e^{\tilde{\ell}_n(\theta)-\ell_n(\theta)}\pi(\theta),\end{aligned}$$

using (24) on event $\mathcal{D}$. Hence, we have

$$\tilde{\pi}(\theta|Z^{(n)}) \leq e^{-4(C_1^2 c_{\max}^2/c_{\min}-C_4)n\delta_n^2} e^{\ell_n(\theta_{*,p})-\ell_n(\theta_0)}\pi(\theta). \tag{39}$$

Now using (39), we get that on a high probability event $\tilde{\varrho}$

$$\begin{aligned}\mathbb{P}_{\theta_0}^n\left(\tilde{\Pi}(\mathcal{U}^C|Z^{(n)}) > e^{-cn\delta_n^2}, \tilde{\varrho}\right) &\leq \mathbb{P}_{\theta_0}^n\left(e^{-4(L^2c_{\min}/4-C_4-c)n\delta_n^2} e^{\ell_n(\theta_{*,p})-\ell_n(\theta_0)}\int_{\mathcal{U}^C}\pi(\theta)d\theta > 1/2\right) \\ &\leq 2e^{-4(L^2c_{\min}/4-C_4-c)n\delta_n^2}\mathbb{E}_{\theta_0}{}^n[e^{\ell_n(\theta_{*,p})-\ell_n(\theta_0)}] \\ &\leq 2e^{-4(L^2c_{\min}/4-C_4-c)n\delta_n^2},\end{aligned} \tag{40}$$

since under θ_0, $\ell_n(\theta_0) \geq \ell_n(\theta_{*,p})$. Now by taking L large enough such that $L^2c_{\min}/4 - C_4 - c > 0$, the probability decays exponentially, hence the surrogate posterior also contracts around $\theta_{*,p}$ at the rate $\delta_n = n^{-\alpha/(2\alpha+1)}$. □

7.4.4 Step 5

Theorem 8. *For the GLM model, there exist $c, c' > 0$ and an event $\tilde{\varrho}$ with $\mathbb{P}_{\theta_0}^n(\tilde{\varrho}) \geq 1 - c'e^{-cn\delta_n^2}$ on which $W_2^2\left(\tilde{\Pi}(.|Z^{(n)}), \Pi(\theta|Z^{(n)})\right) \leq e^{-n\delta_n^2}$.*

Proof of Theorem 8. In (40), taking n and L large, we can ensure that on $\tilde{\varrho}$, $\Pi(\mathcal{U}^C|Z^{(n)}) \leq e^{-cn\delta_n^2}/2$ and $\tilde{\Pi}(\mathcal{U}^C|Z^{(n)}) \leq e^{-cn\delta_n^2}/2$. Also using (2), on $\tilde{\varrho}$, we have for $C_2 > 0$

$$\pi(\theta|Z^{(n)}) = \frac{e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)-\ell_n(\theta_0)}\pi(\theta)d\theta} \leq e^{C_2 n\delta_n^2} e^{\ell_n(\theta)-\ell_n(\theta_0)}. \tag{41}$$

By applying Theorem 6.15 of [9], we can upper bound the squared Wasserstein distance between the surrogate posterior and the posterior by

$$\begin{aligned}W_2^2\left(\tilde{\Pi}(.|Z^{(n)}), \Pi(\theta|Z^{(n)})\right) &\leq 2\int_{\mathbb{R}^p}\|\theta-\theta_{*,p}\|^2|\tilde{\pi}(\theta|Z^{(n)}) - \pi(\theta|Z^{(n)})|d\theta \\ &\leq \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3,\end{aligned}$$

where we have,

$$\mathcal{I}_1 = \int_{\mathcal{U}} \|\theta - \theta_{*,p}\|^2 \left|\tilde{\pi}(\theta|Z^{(n)}) - \pi(\theta|Z^{(n)})\right| d\theta,$$
$$\mathcal{I}_2 = \int_{\mathcal{U}^C} \|\theta - \theta_{*,p}\|^2 \tilde{\pi}(\theta|Z^{(n)}) d\theta,$$
$$\mathcal{I}_3 = \int_{\mathcal{U}} \|\theta - \theta_{*,p}\|^2 \pi(\theta|Z^{(n)}) d\theta.$$

We will show that each of these terms exceeds $e^{-n\delta_n^2}/3$ with an exponentially small $\mathbb{P}_{\theta_0}^n$ probability and the proof follows from that.

We know using (37) that for $\theta \in \mathcal{U}$, $p_n \pi(\theta|Z^{(n)}) = \tilde{\pi}(\theta|Z^{(n)})$ for some normalising constant $p_n > 0$. Now since $\Pi(\mathcal{U}|Z^{(n)}) \leq 1$ and $\tilde{\Pi}(\mathcal{U}|Z^{(n)}) \leq 1$, we have

$$p_n \geq p_n \Pi(\mathcal{U}|Z^{(n)}) = \tilde{\Pi}(\mathcal{U}|Z^{(n)}) \geq 1 - e^{-cn\delta_n^2}/2,$$

also,

$$p_n^{-1} \geq p_n^{-1} \tilde{\Pi}(\mathcal{U}|Z^{(n)}) = \Pi(\mathcal{U}|Z^{(n)}) \geq 1 - e^{-cn\delta_n^2}/2.$$

These inequalities imply that

$$1 - e^{-cn\delta_n^2}/2 \leq p_n \leq (1 - e^{-cn\delta_n^2}/2)^{-1},$$
$$-\frac{e^{-cn\delta_n^2}/2}{1 - e^{-cn\delta_n^2}/2} \leq 1 - p_n \leq e^{-cn\delta_n^2}/2.$$

Now for large enough n, the radius of $\mathcal{U}$, $L\delta_n \leq 1/3$, which implies

$$\begin{aligned}
\mathcal{I}_1 &\leq L^2 \delta_n^2 \int_{\mathcal{U}} \left|\tilde{\pi}(\theta|Z^{(n)}) - \pi(\theta|Z^{(n)})\right| d\theta \\
&= L^2 \delta_n^2 \int_{\mathcal{U}} \left|p_n \pi(\theta|Z^{(n)}) - \pi(\theta|Z^{(n)})\right| d\theta \\
&\leq (1/3)|1 - p_n| \Pi(\mathcal{U}|Z^{(n)}) \\
&\leq e^{-cn\delta_n^2}/3.
\end{aligned}$$

Now using the Cauchy-Schwarz inequality, and (39), with $C_1 \geq 1$, $4C_1^2 c_{\max}^2 / c_{\min} - C_4 \geq 0$, we have

$$\begin{aligned}
\mathcal{I}_2^2 &\leq \tilde{\Pi}(\mathcal{U}^C|Z^{(n)}) \int_{\mathcal{U}^C} \|\theta - \theta_{*,p}\|^4 \tilde{\pi}(\theta|Z^{(n)}) d\theta \\
&\leq e^{-cn\delta_n^2} e^{\ell_n(\theta_{*,p}) - \ell_n(\theta_0)} \int_{\Theta} \|\theta - \theta_{*,p}\|^4 \pi(\theta) d\theta,
\end{aligned} \tag{42}$$

using (39) in the last line. Using Markov Inequality and Fubini's theorem, we have

$$\begin{aligned}
\mathbb{P}_{\theta_0}^n \left(\mathcal{I}_2 > e^{-n\delta_n^2}/3, \tilde{\varrho}\right) &= \mathbb{P}_{\theta_0}^n \left(\mathcal{I}_2^2 > e^{-2n\delta_n^2}/9, \tilde{\varrho}\right) \\
&\lesssim e^{(2-c)n\delta_n^2} \mathbb{E}_{\theta_0}{}^n \left(e^{\ell_n(\theta_{*,p}) - \ell_n(\theta_0)}\right) \int_{\Theta} \|\theta - \theta_{*,p}\|^4 \pi(\theta) d\theta.
\end{aligned}$$

Now note that $\|\theta_{*,p}\| \leq c_0$, by the property 1, and the Gaussian prior (16) has uniformly bounded fourth moments. We get

$$\mathbb{P}^n_{\theta_0}\left(\mathcal{I}_2 > e^{-n\delta_n^2}/3, \tilde{\varrho}\right) \lesssim e^{(2-c)n\delta_n^2}.$$

By taking $c > 2$, the probability decays exponentially. Similar result can be proved for $\mathcal{I}_3$ since also $\Pi(\mathcal{U}^C|Z^{(n)}) \leq e^{-cn\delta_n^2}/2$. □

7.4.5 Step 6

Proof of Theorem 4. For the strong log-concave surrogate posterior, we can now apply Proposition A.4 of [2] (an application of Theorem 5 of [8]) with $\theta_{\max}$ as the unique maximizer of $\tilde{\Pi}(\theta|Z^{(n)})$. This gives us

$$W_2^2(\mathcal{L}(\tilde{\vartheta}_k), \tilde{\Pi}(.|Z^{(n)})) \leq 2(1 - m\gamma/2)^k \left(\|\theta_{\text{init}} - \theta_{\max}\|^2 + p/m\right) + B(\gamma)/2, \tag{43}$$

$$k \geq 0, \qquad B(\gamma) = 36\frac{\gamma p \Lambda^2}{m^2} + 12\frac{\gamma^2 D \Lambda^4}{m^3}.$$

We will now show that $\|\theta_{\text{init}} - \theta_{\max}\|^2 \leq c_w$, where $c_w = C(c_0, c_{\max}, c_{\min})$. Let $\tilde{\theta}_{\max}$ be the unique maximizer of the strongly concave map $\tilde{\ell}_n$. Using triangle inequality, we have

$$\|\theta_{\text{init}} - \theta_{\max}\| \leq \|\theta_{\text{init}} - \theta_{*,p}\| + \|\theta_{*,p} - \tilde{\theta}_{\max}\| + \|\tilde{\theta}_{\max} - \theta_{\max}\|. \tag{44}$$

The first term $\|\theta_{\text{init}} - \theta_{*,p}\| \leq \eta/8$ by Condition 1. Also if $\|\tilde{\theta}_{\max} - \theta_{*,p}\| > (4C_1 c_{\max}/c_{\min})\delta_n$, then we have using (38) $\tilde{\ell}_n(\tilde{\theta}_{\max}) - \tilde{\ell}_n(\theta_{*,p}) \leq -(4C_1^2 c_{\max}/c_{\min}) n\delta_n^2 < 0$, which contradicts the fact that $\tilde{\theta}_{\max}$ is the maximizer of $\tilde{\ell}_n$. Hence, we have that $\|\tilde{\theta}_{\max} - \theta_{*,p}\| \leq (4C_1 c_{\max}/c_{\min})\delta_n \leq c_w$. This also gives us $\|\tilde{\theta}_{\max}\| \lesssim 1$. Now if $\tilde{\theta}_{\max} = \theta_{\max}$, we have the desired result already. Consider the case $\tilde{\theta}_{\max} \neq \theta_{\max}$, we have

$$\begin{aligned}
0 &\leq \tilde{\ell}_n(\tilde{\theta}_{\max}) - \tilde{\ell}_n(\theta_{\max}) \\
&= \log\tilde{\pi}(\tilde{\theta}_{\max}|Z^{(n)}) - \log\tilde{\pi}(\theta_{\max}|Z^{(n)}) + \log\pi(\theta_{\max}) - \log\pi(\tilde{\theta}_{\max}) \\
&\leq \log\tilde{\pi}(\tilde{\theta}_{\max}|Z^{(n)}) - \log\tilde{\pi}(\theta_{\max}|Z^{(n)}) + (\nabla\log\pi(\tilde{\theta}_{\max}) - \nabla\log\pi())^T(\theta_{\max} - \tilde{\theta}_{\max}).
\end{aligned}$$

Since is the maximizer of $\log\pi(\theta)$ and hence $\nabla\log\pi() = 0$. Now using that $\log\pi(\theta)$ is concave and has Λ_π-Lipschitz gradients

$$\begin{aligned}
0 &\leq \log\tilde{\pi}(\tilde{\theta}_{\max}|Z^{(n)}) - \log\tilde{\pi}(\theta_{\max}|Z^{(n)}) + \Lambda_\pi\|\theta_{\max} - \tilde{\theta}_{\max}\|\|\tilde{\theta}_{\max}\| \\
&\lesssim -\frac{m}{2}\|\tilde{\theta}_{\max} - \theta_{\max}\|^2 + \Lambda_\pi\|\tilde{\theta}_{\max} - \theta_{\max}\|.
\end{aligned}$$

Using that the surrogate posterior is m-strongly concave using Theorem 7. Hence, $\|\tilde{\theta}_{\max} - \theta_{\max}\| \lesssim \Lambda_\pi/m \lesssim 1$. Hence using (44), we get that $\|\theta_{\text{init}} - \theta_{\max}\|^2 \leq c_w$.

Now using triangle inequality for Wasserstein distance, we get that

$$\begin{aligned}
W_2(\mathcal{L}(\tilde{\vartheta}_k), \Pi(.|Z^{(n)})) &\leq W_2(\mathcal{L}(\tilde{\vartheta}_k), \tilde{\Pi}(.|Z^{(n)})) + W_2(\Pi(.|Z^{(n)}), \tilde{\Pi}(.|Z^{(n)})), \\
W_2^2(\mathcal{L}(\tilde{\vartheta}_k), \Pi(.|Z^{(n)})) &\leq 2(W_2^2(\mathcal{L}(\tilde{\vartheta}_k), \tilde{\Pi}(.|Z^{(n)})) + W_2^2(\Pi(.|Z^{(n)}), \tilde{\Pi}(.|Z^{(n)}))) \\
&\leq 2e^{-n\delta_n^2} + 4(1 - m\gamma/2)^k\,(c_w + p/m) + B(\gamma),
\end{aligned}$$

using (44) and Theorem 8. Now for $k \geq J_{\text{in}}$, we have that

$$4(1-m\gamma/2)^k(c_w+p/m) \leq 4e^{\log(1-m\gamma/2)\frac{\log e\epsilon^2}{\log(1-cn\gamma)}} \leq 4c'\epsilon^2,$$

since $m \gtrsim n$. Also $B(\gamma) = 36\frac{\gamma p\Lambda^2}{m^2} + 12\frac{\gamma^2 D\Lambda^4}{m^3} \lesssim \gamma p^2 + \gamma^2 p^3 n$ and $\epsilon \geq c\max(e^{-n^{1/(2\alpha+1)}/2}, \gamma^{1/2}p, \gamma p^{3/2}n^{1/2})$ implies that

$$\epsilon^2 \geq c'' \max(B(\gamma), e^{-n\delta_n^2}),$$

since $\delta_n = n^{1/(2\alpha+1)}$. Now by choosing appropriate c' and c'', we have that

$$W_2^2(\mathcal{L}(\tilde{\vartheta}_k), \Pi(.|Z^{(n)})) \leq \epsilon^2,$$

which gives us the first part of Theorem 4. For the second part, we will apply Proposition A.3 of [2] and the claim follows. □

7.4.6 Step 7

Since $\theta_{\text{init}} \in \tilde{\mathcal{B}}$, let us define the exit time from the ball $\tilde{\mathcal{B}}$ for the LMC iterates for the surrogate posterior by $\tau = \inf\{k \geq 1 : \tilde{\vartheta}_k \notin \tilde{\mathcal{B}}\}$, that is the first time the Markov chain $(\tilde{\vartheta}_k)_{k\in\mathbb{N}}$ exits the region $\tilde{\mathcal{B}}$, where ℓ_n and $\tilde{\ell}_n$ agree. We now quantize the exit time τ with high $\mathbb{P}^n_{\theta_0}$-probability in the following theorem.

Theorem 9. *For the GLM model, there exist $c_1, c_2, c_3, c_4 > 0$ such that on an event $\bar{\varrho}$ with $\mathbb{P}^n_{\theta_0}(\bar{\varrho}) \geq 1 - c_1 e^{c_2 n\delta_n^2}$ for all $J \geq 1$*

$$\mathbb{P}(\tau \leq J) \leq c_3 p \exp\left(-c_4 \frac{\eta^2 m}{p(1-e^{-mJ\gamma})}\right) + c_1 J p \exp\left(-c_2 \frac{\eta^2 m^2}{\gamma p \Lambda^2}\right).$$

Now with a polynomial computational budget, that is $J \asymp n^\rho, \rho > 0$, since $\eta^2 m^2/(\gamma p\Lambda^2) \gtrsim n^{\rho'}, \rho' > 0$, we will stay within the region $\mathcal{B}$ of local curvature with high $\mathbb{P}^n_{\theta_0} \times \mathbb{P}$-probability.

Proof of Theorem 9. For the proof of this theorem, we will restrict to the high probability event ϱ. We will extend the probability space to support a p-dimensional Brownian motion $(W_t)_{t\geq 0}$ with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ and consider the Langevin diffusion process with gradient potential $f(\theta) = \log\tilde{\pi}(\theta|Z^{(n)})$. We assume that the filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfies the usual conditions. We will upper bound the probability of Markov chain iterates exiting $\mathcal{B}$ with the probability of its continuous interpolation exiting $\tilde{\mathcal{B}}$. Let us define two p-dimensional stochastic differential equations with the same initial point $L_0 = \bar{L}_0$

$$\begin{aligned} dL_t &= \nabla f(L_t)dt + \sqrt{2}dW_t, \\ d\bar{L}_t &= \nabla f(\bar{L}_{\lfloor t/\gamma\rfloor\gamma})dt + \sqrt{2}dW_t. \end{aligned} \tag{45}$$

Since the function f is strongly m-concave and has Λ-Lipschitz gradients on the event ϱ, (45) has a unique strong solution $(L_t)_{t\geq 0}$ adapted to the filtration $(\mathcal{F}_t)_{t\geq 0}$. The LMC iterates $\tilde{\vartheta}_k$ can be seen as a discretisation of the process $\bar{L}_t$ since, their laws ($\mathcal{L}$) are the same, i.e. $\mathcal{L}(\bar{L_\gamma}, ..., \bar{L}_{J\gamma}) = \mathcal{L}(\vartheta_1, ..., \vartheta_J)$. Thus

$$\mathbb{P}(\tau \leq J) = \mathbb{P}\left(\sup_{k=1,,J} \|\bar{L}_{k\gamma} - \theta_{*,p}\| > 3\eta/8\right).$$

Using triangle inequality, we can consider

$$\|\bar{L}_{k\gamma} - \theta_{*,p}\| \leq \|L_{k\gamma} - \theta_{*,p}\| + \|L_{k\gamma} - \bar{L}_{k\gamma}\|.$$

We will now compute the probabilities for both parts separately. More precisely, we will show that

$$\begin{aligned}\mathbb{P}\left(\sup_{0\leq t\leq J\gamma} \|L_t - \theta_{*,p}\| > x + \eta/8\right) &\leq c'p\exp\left(-c\frac{x^2 m}{p(1-e^{-mJ\gamma})}\right),\\ \mathbb{P}\left(\sup_{k=1,..,J} \|L_{k\gamma} - \bar{L}_{k\gamma}\| > 3\eta/16\right) &\leq c'p\exp\left(-c\frac{x^2 m}{p(1-e^{-mJ\gamma})}\right) + c'Jp\exp\left(-c\frac{\eta^2 m^2}{\gamma p\Lambda^2}\right).\end{aligned} \tag{46}$$

We can then conclude the proof since $\mathbb{P}\left(\sup_{k=1,,J} \|\bar{L}_{k\gamma} - \theta_{*,p}\| > 3\eta/8\right) \leq \mathbb{P}\left(\sup_{0\leq t\leq J\gamma} \|L_t - \theta_{*,p}\| + \sup_{k=1,..,J} \|L_{k\gamma} - \bar{L}_{k\gamma}\| > 3\eta/8\right)$.

For the first part of (46), we apply Ito's formula to the function $\theta \mapsto \|\theta - \theta_{*,p}\|$. It is valid since the Brownian motion (hence the diffusion), doesn't hit $\theta_{*,p}$ almost surely. Using Ito's formula, we get

$$\|L_t - \theta_{*,p}\| = \int_0^t \left[\frac{L_s - \theta_{*,p}}{\|L_s - \theta_{*,p}\|}.\nabla f(L_s) + \frac{1}{2}\frac{p-1}{\|L_s - \theta_{*,p}\|}\right] ds + \sqrt{2}\tilde{W}_t,$$

with $\tilde{W}_t = \int_0^t (L_s - \theta_{*,p})/\|L_s - \theta_{*,p}\| dW_s$ a scalar Brownian motion due to Levy's characterization since its quadratic variation matches that of a scalar Brownian motion. Also

$$\begin{aligned}(\theta - \theta_{*,p})^T \nabla f(\theta) &\leq (\theta - \theta_{*,p})^T \nabla f(\theta_{*,p}) - (m/2)\|\theta - \theta_{*,p}\|^2\\ &= (\theta - \theta_{*,p})^T(-(m/2)(\theta - \bar{\theta})),\end{aligned}$$

where $\bar{\theta} = \theta_{*,p} - (2/m)\nabla f(\theta_{*,p})$, due to strong m- concavity of f. This motivates us to consider a p-dimensional Ornstein-Uhlenbeck process with a closed form solution starting at $V_0 = \theta_{\text{init}}$,

$$dV_t = -(m/2)(V_t - \bar{\theta})dt + \sqrt{2}d\tilde{W}_t.$$

By comparing the drift components, we infer that

$$\|L_t - \theta_{*,p}\| \leq \|V_t - \theta_{*,p}\| \quad \mathbb{P}\text{-almost surely.} \tag{47}$$

The explicit solution for V_t is given by

$$\begin{aligned}V_t &= \theta_{\text{init}} e^{-(m/2)t} + \bar{\theta}(1 - e^{-(m/2)t}) + \sqrt{2/m}\tilde{W}_{1-e^{-mt}},\\ &= \theta_{*,p} + (\theta_{\text{init}} - \theta_{*,p})e^{-(m/2)t} - (2/m)\nabla f(\theta_{*,p})(1 - e^{-mt/2}) + \sqrt{2/m}\tilde{W}_{1-e^{-mt}}.\end{aligned}$$

Now from property 2, $\|\nabla \ell_n(\theta_{*,p})\| \le c_{\max} n\delta_n = c_{\max} n^{\frac{\alpha+1}{2\alpha+1}}$, $m \ge c_{\min} n$, $\eta m \ge C' n^{\frac{4\alpha+1}{2(2\alpha+1)}}$. For n large enough, $\|\nabla \ell_n(\theta_{*,p})\| \le \eta m/32$. Also since $m_\pi = n^{1/(2\alpha+1)}$, $\Lambda_\pi = n^{1/(2\alpha+1)} p^{2\alpha}$, $\|\nabla \log \pi(\theta_{*,p})\| \le \Lambda_\pi \|\theta_{*,p}\| \le c_0 n^{1/(2\alpha+1)} p^{2\alpha} \le \eta m/32$ for n large enough. Hence,

$$\|\nabla f(\theta_{*,p})\| \le \|\nabla \ell_n(\theta_{*,p})\| + \|\nabla \log \pi(\theta_{*,p})\| \le \eta m/16. \tag{48}$$

Now since $\|\theta_{\text{init}} - \theta_{*,p}\| \le \eta/8$, we have

$$\|\theta_{\text{init}} - \theta_{*,p}\| e^{-mt/2} + (2/m)\|\nabla f(\theta_{*,p})\|(1 - e^{-mt/2}) \le \eta/8 + (2/m)\eta m/16 \le \eta/8.$$

Using the last equation, we get

$$\mathbb{P}\left(\sup_{0\le t\le J\gamma} \|L_t - \theta_{*,p}\| > x + \eta/8\right) \le \mathbb{P}\left(\sup_{0\le t\le J\gamma} \|\tilde{W}_{1-e^{-mt}}\| > x\sqrt{m/2}\right).$$

Now we take the coordinates $(\tilde{W}_{i,t})_{t\ge 0}$ of the Brownian motion individually and apply a union bound

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le t\le J\gamma} \|L_t - \theta_{*,p}\| > x + \eta/8\right) &\le p\mathbb{P}\left(\sup_{0\le s\le 1-e^{-mJ\gamma}} |\tilde{W}_{1,s}| > x\sqrt{m/2p}\right) \qquad (49)\\ &\le p\frac{\sqrt{2(1-e^{-mJ\gamma})}}{x\sqrt{m/(2p\pi)}} \exp\left(-\frac{x^2 m}{4p\pi(1-e^{mJ\gamma})}\right),\end{aligned}$$

using the result on exit time of a scalar Brownian motion from an interval in [10](Remark 2.8.3). The result now follows by noting that $x\sqrt{m/p} \gtrsim p^{-1}\sqrt{n} \ge cn^{-1/(2\alpha+1)}$, hence the first term can be bounded uniformly by a constant.

For the second part of (46), we use Lemma 22 of [8] with the strongly convex function $U = -f$ and $\kappa = (2m\Lambda)/(m+\Lambda)$, $\epsilon = \kappa/4$ such that $\forall k \ge 1$, we have

$$\|L_{k\gamma} - \bar{L}_{k\gamma}\|^2 \le (1-\gamma\kappa/2)\|L_{(k-1)\gamma} - \bar{L}_{(k-1)\gamma}\|^2 + (\gamma + 2/\kappa)\int_{(k-1)\gamma}^{k\gamma} \|(L_s) - \nabla f(L_{(k-1)\gamma})\|^2 ds.$$

Applying this step recursively and using $L_0 = \bar{L}_0$ gives us

$$\|L_{k\gamma} - \bar{L}_{k\gamma}\|^2 \le (\gamma + 2/\kappa)\sum_{i=1}^{k}(1-\gamma\kappa/2)^{\kappa-i}\int_{(i-1)\gamma}^{i\gamma} \|\nabla f(L_s) - \nabla f(L_{(i-1)\gamma})\|^2 ds.$$

Noting that ∇f is Λ-Lipschitz, we get

$$\|L_{k\gamma} - \bar{L}_{k\gamma}\|^2 \le \frac{2}{\kappa\gamma}(\gamma + 2/\kappa)(1-(1-\gamma\kappa/2)^k)\Lambda^2 \sup_{0\le t\le k\gamma}\|L_t - L_{\lfloor t/\gamma\rfloor\gamma}\|^2\gamma.$$

Since L solves (47), also, $m = c_{\min}n + n^{1/(2\alpha+1)}$, $\Lambda \ge cnp^{1/2} + n^{1/(2\alpha+1)}p^{2\alpha}$. In the setting of Theorem 4, we have $\gamma \le cn^{-1}p^{-1/2} \le \Lambda^{-1}$, $\kappa = m\frac{2\Lambda}{m+\Lambda} \ge m$, $\Lambda/\kappa \le 1/m^2$. Hence, we have,

$$\begin{aligned}\|L_{k\gamma} - \bar{L}_{k\gamma}\| &\le \sqrt{2(\gamma/\kappa + 2/\kappa^2)}\Lambda \sup_{0\le t\le k\gamma}\|L_t - L_{\lfloor t/\gamma\rfloor\gamma}\|\\ &\le (\sqrt{6}\Lambda/m)\sup_{0\le t\le k\gamma}\left(\|\nabla f(L_s)\|ds + \sqrt{2}\|W_t - W_{\lfloor t/\gamma\rfloor\gamma}\|\right)\\ &\le (\sqrt{6}\Lambda/m)\sup_{0\le t\le k\gamma}(\|\nabla f(L_s) - \nabla f(\theta_{*,p})\|ds + \gamma\|\nabla f(\theta_{*,p})\|) + (\sqrt{12}\Lambda/m)V,\end{aligned}$$

is obtained by adding and substracting the term $\nabla f(\theta_{*,p})$, with $V = \sup_{0\le t\le k\gamma} \|W_t - W_{\lfloor t/\gamma\rfloor\gamma}\|$. Now again using that ∇f is Λ-Lipschitz and that $\|\nabla f(\theta_{*,p})\| \le \eta m/8$ for n large enough as in (48), we get

$$\|L_{k\gamma} - \bar{L}_{k\gamma}\| \le (\sqrt{6}\Lambda^2\gamma/m) \sup_{0\le t\le J\gamma} \|L_t - \theta_{*,p}\| + \sqrt{6}\Lambda\gamma\eta/8 + (\sqrt{12}\Lambda/m)V. \tag{50}$$

Also note $\sqrt{6}\Lambda\gamma\eta/8 \le \eta/16$ and $\eta m/(16\sqrt{6}\Lambda^2\gamma) - \eta/8 \ge \eta/16$ since $\gamma \le \Lambda^{-1}$, we get

$$V \sim \gamma^{1/2} \sup_{0\le t\le J} \|W_t - W_{\lfloor t\rfloor}\| \le \gamma^{1/2}p^{1/2} \max_{1\le I\le p, 1\le k\le J} \sup_{k-1\le t\le k} |W_{i,t} - W_{i,\lfloor t\rfloor}|,$$

since $\|x\|_2 \le p^{1/2}\|x\|_\infty$. Now consider the expression in the second part of (46)

$$\mathbb{P}\left(\sup_{k=1,..,J} \|L_{k\gamma} - \bar{L}_{k\gamma}\| > 3\eta/16\right)$$

$$\begin{aligned}
&\le \mathbb{P}\left((\sqrt{6}\Lambda^2\gamma/m) \sup_{0\le t\le J\gamma} \|L_t - \theta_{*,p}\| > \eta/8\right) + \mathbb{P}\left((\sqrt{12}\Lambda/m)V > 2\eta/16\right) \\
&\le c'p \exp\left(-c\frac{\eta^2 m}{p(1-e^{-mJ\gamma})}\right) + Jp\mathbb{P}\left(\sup_{k-1\le t\le k} |W_{i,t} - W_{i,\lfloor t\rfloor}| > \eta m/(16\sqrt{12\gamma p}\Lambda)\right),
\end{aligned}$$

using the first part of (46) and union bound on i, k. Now we can use the result for the exit time of a scalar Brownian motion as in (49) and conclude with $\eta\sqrt{m/p} \gtrsim n^{-1/(2\alpha+1)}$ and $\gamma \lesssim m/\Lambda^2$. □

7.4.7 Step 8

Proof of Theorem 5. Consider the stopping time τ defined in Theorem 9. The probability of any event A can be upper bound by:

$$\mathbb{P}(A) \le \mathbb{P}(A, \tau > J + J_{\text{in}}) + \mathbb{P}(\tau \le J + J_{\text{in}}).$$

Now for all $k \le J + J_{\text{in}}$, the LMC iterates for the posterior, and the surrogate posterior coincide that is $\vartheta_k = \tilde{\vartheta}_k$ since we do not exit the region $\tilde{\mathcal{B}}$, on which both the densities agree. Hence, we have

$$\mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{\text{in}}}^{J+J_{\text{in}}} f(\vartheta_k) - \int_\Theta f(\theta)d\Pi(\theta|Z^{(n)})\right| > \epsilon\right)$$

$$\begin{aligned}
&\le \mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{\text{in}}}^{J+J_{\text{in}}} f(\vartheta_k) - \int_\Theta f(\theta)d\Pi(\theta|Z^{(n)})\right| > \epsilon, \tau > J + J_{\text{in}}\right) + \mathbb{P}\left(\tau \le J + J_{\text{in}}\right) \\
&= \mathbb{P}\left(\left|\frac{1}{J}\sum_{k=1+J_{\text{in}}}^{J+J_{\text{in}}} f(\tilde{\vartheta}_k) - \int_\Theta f(\theta)d\Pi(\theta|Z^{(n)})\right| > \epsilon, \tau > J + J_{\text{in}}\right) + \mathbb{P}\left(\tau \le J + J_{\text{in}}\right) \\
&\le 2\exp\left(-c_3\frac{\epsilon^2 n^2 J\gamma}{1 + 1/(nJ\gamma)}\right) + c_3 p \exp\left(-c_4\frac{\eta^2 m}{p(1-e^{-m(J+J_{\text{in}})\gamma})}\right) + c_1(J + J_{\text{in}})p \exp\left(-c_2\frac{\eta^2 m^2}{\gamma p\Lambda^2}\right),
\end{aligned}$$

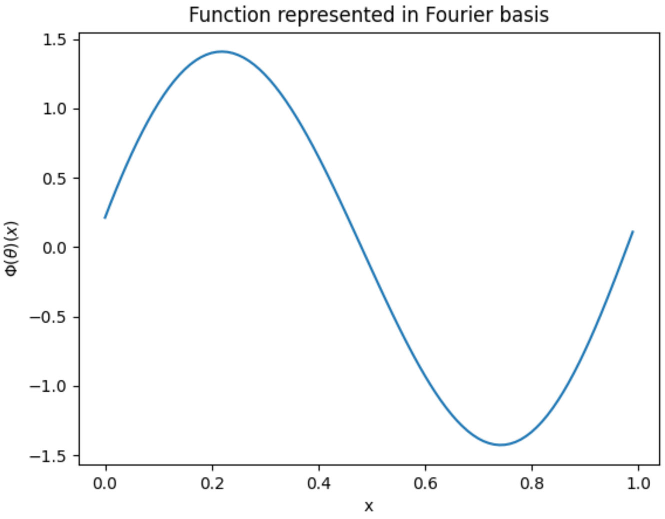

Figure 1: $\phi(\theta_0)(x)$ on a Fourier basis

using Theorems 4 and 9 in the last line. Now using $m \gtrsim n$, $\eta = p^{-1/2}$, $p \le cn^{1/(2\alpha+1)}$, $\gamma \le cn^{-1}p^{-1/2}$, we have $\eta^2 m/p \le c'n^{(2\alpha-1)/(2\alpha+1)}$, $(J + J_{\text{in}})pe^{-c_2\eta^2 m/(\gamma p \Lambda^2)} \lesssim e^{\log(J+J_{\text{in}})}e^{\frac{1}{(2\alpha+1)}\log n}e^{-n^{(2\alpha-1)/(2\alpha+1)}} \lesssim e^{-C'n^{1/(2\alpha+1)}}$ and $p\exp\left(-c_4\frac{\eta^2 m}{p(1-e^{-m(J+J_{\text{in}})\gamma})}\right) \lesssim e^{\frac{1}{2\alpha+1}\log n}e^{\left(-c'\frac{n^{(2\alpha-1)/(2\alpha+1)}}{(1-e^{-m(J+J_{\text{in}})\gamma})}\right)} \lesssim \exp\left(c''\frac{n^{1/(2\alpha+1)}}{1-e^{-m(J+J_{\text{in}})}}\right)$. Hence, the result follows. □

8 Numerical Experiments

We consider an example of non-parametric Logistic Regression GLM with $\theta_0 \in h^\alpha(\mathbb{N})$ for $\alpha = 2$. We take the function $\Phi(\theta) : [0,1] \mapsto \mathbb{R}$ on a Fourier basis e_k, that is

$$\Phi(\theta)(x) = \sqrt{2}\sum_{k=0}^{\infty}(\theta_{2k+1}\sin(2\pi kx) + \theta_{2k}\cos(2\pi kx)),$$

where

$$e_k(x) = \begin{cases} \sqrt{2}\sin(2\pi kx), & \text{if } k \text{ is even,} \\ \sqrt{2}\cos(2\pi kx), & \text{otherwise.} \end{cases}$$

We note that the Fourier basis is bounded and orthonormal on $\mathcal{X} = [0,1]$, that is $|e_k(x)| \le \sqrt{2}$ and $\int_{\mathcal{X}} e_k(x)e_j(x) = \mathbb{1}_{\{k=j\}}$. We take the design $p_{\mathcal{X}}(x)$ to be uniform on $\mathcal{X}$. Hence the bounded design assumption is satisfied.

Now let us choose θ_0 such that $\sum_{k=1}^{\infty} k^4\theta_{0,k}^2 < \infty$. Our candidate for θ_0 is

$$\theta_{0,k} = 1/k^3.$$

For this choice of θ_0, the map $\Phi(\theta_0)(x)$ can be seen in Figure 1

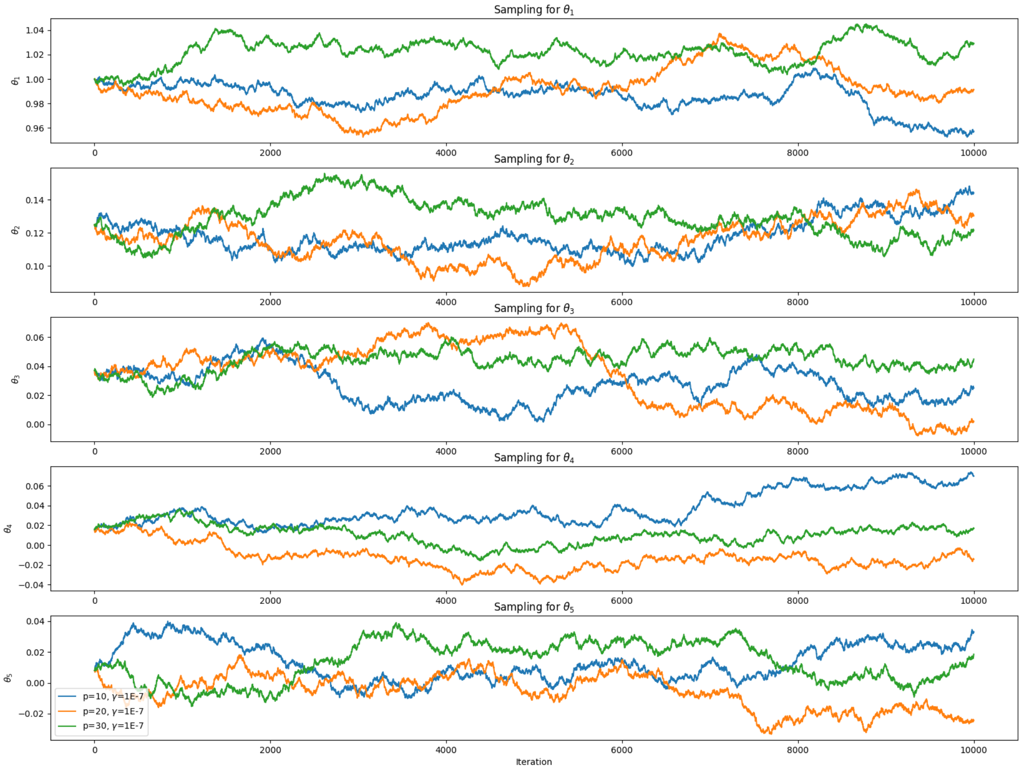

Figure 2: LMC iterates for different values of the dimension p

We generate $\theta = X^{(1000)}$, $n = 1000$ samples uniformly distributed in $[0, 1]$. We then generate the responses Y_i using the map $Y_i = \text{Bern}(p(X_i))$, where $\text{Bern}(p)$ is a Bernoulli random variable with success probability p, $p(x) = (A') \circ \Phi(\theta_0)(x)$ where $A'(x) = \frac{1}{1+e^{-x}}$. We then take a p-dimensional projection of θ_0 for different values of p. For simplicity, initialize the LMC algorithm at $\theta_{\text{init}} = \theta_{*,p}$. We run the LMC iterates for 10,000 iterations with $\gamma = 1E - 7$ for different values of p. We plot the first 5 dimensions of θ in Figures 2 and 3. The code for the simulation can be found here.

The LMC iterates are given by

$$\vartheta_{k+1} = \vartheta_k + \gamma \left(\nabla \ell_n(\vartheta_k) - n^{1/(2\alpha+1)} \Sigma_\alpha \vartheta_k \right) + \sqrt{2\gamma} \zeta_{k+1} \tag{51}$$

We observe that for $p = 30$, the posterior means are higher than $p = 10$ or $p = 20$. This could be because, for $p = 30$, the LMC algorithm would take more steps to converge. As we have already seen in Example 3, the likelihood $\ell_n(\theta)$ is m-strongly concave and has Λ-Lipschitz gradients. Hence, the posterior is also $m + m_\pi$- strongly concave with $\Lambda + \Lambda_\pi$-Lipschitz gradients, with $m \lesssim n$, $\Lambda = \frac{n\sqrt{p}c\mathcal{X}}{4}$, $m_\pi = n^{1/(2\alpha+1)}$, and $\Lambda_\pi = n^{1/(2\alpha+1)} p^{2\alpha}$. From Corollary 1, we have that for

$$T = \frac{4 \log(1/\epsilon) + p \log((\Lambda + \Lambda_\pi)/m)}{2m}, \qquad \gamma = \frac{\epsilon^2 (2\alpha - 1)}{(\Lambda + \Lambda_\pi)^2 T p \alpha},$$

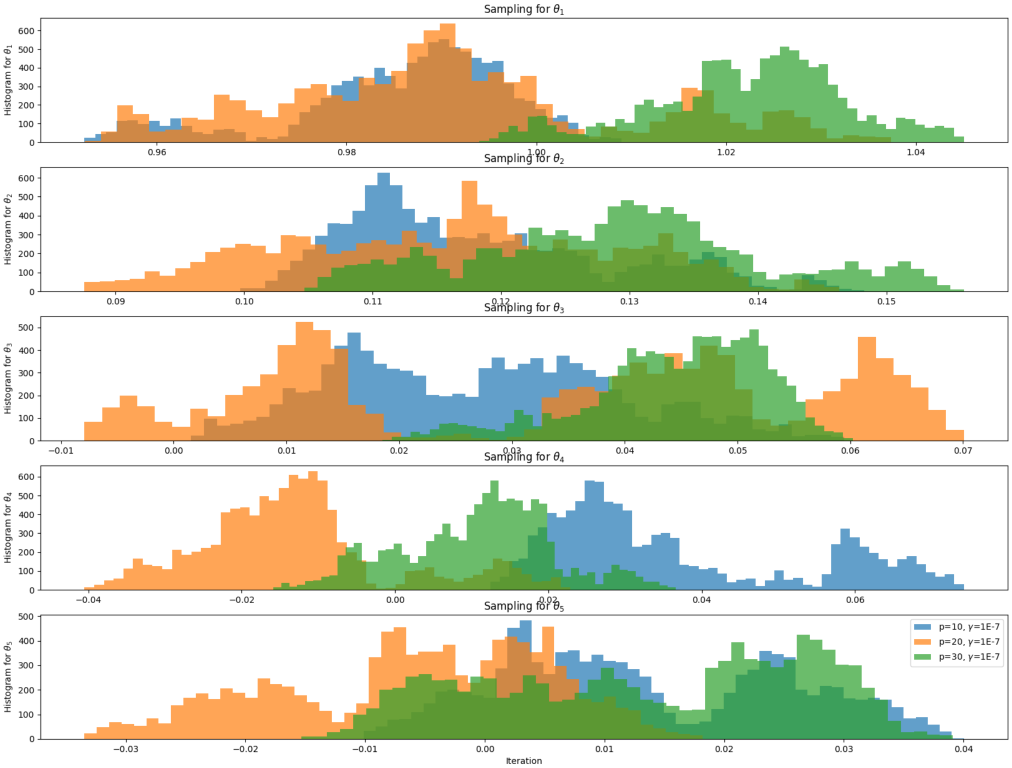

Figure 3: LMC iterates for different values of the dimension p

where $\alpha = (1 + (\Lambda + \Lambda_\pi)pT\epsilon^{-2})/2$, the J-step output of the LMC algorithm, with $J = \lceil T/\gamma \rceil$ is within ϵ total variation distance of the posterior measure. We get

$$J = \frac{(\Lambda + \Lambda_\pi)^2 p}{4\epsilon^2 m^2} \frac{\alpha}{2\alpha - 1} \left(4\log(1/\epsilon) + p\log((\Lambda + \Lambda_\pi)/m)\right)$$
$$= O(\frac{p^3 \log p}{\epsilon^2} + \frac{p^2 \log(1/\epsilon)}{\epsilon^2}). \tag{52}$$

Since the posterior $\pi(\theta|Z^{(n)})$ is already strongly concave with lipschitz gradients, the surrogate posterior $\tilde{\pi}(\theta|Z^{(n)})$ in 20 is the same as the posterior $\pi(\theta|Z^{(n)})$ and $\mathbb{P}^{\ltimes}_{\theta_{\nvdash}}(\varrho) = 1$. Hence, we can directly use Theorem 4, and obtain

$$J \geq \frac{c'_\alpha + \sqrt{{c'_\alpha}^2 + 4\epsilon^2 n c'_\alpha}}{2\epsilon^2 n^2 \gamma},$$

to get within ϵ distance of the true posterior functional, with $\mathbb{P}^n_{\theta_0} \times \mathbb{P}$-probability $\geq \alpha$. Now taking the best possible step-size $\gamma = cn^{-1}p^{-1/2}$, we get

$$J \geq \frac{p^{1/2}}{2\epsilon^2 n} \left(c'_\alpha + \sqrt{{c'_\alpha}^2 + 4\epsilon^2 n c'_\alpha}\right).$$

Now using the condition on ϵ from Theorem 4, we have

$$J = O(\frac{p^{3/2}}{\epsilon^2 n}),$$

which is an improvement compared to (52). Hence, the results from sections 6 and 7 often result in computation time improvements even for the case of strongly log-concave posteriors.

9 Summary

Through this essay, we have shown the feasibility of the Langevin Monte Carlo Algorithm in various settings in Bayesian inference. To go beyond the strongly log-concave posterior case, using Fisher information as a statistical estimate of local curvature becomes key. Under mild-regularity assumptions, we establish a region $\mathcal{B}$, on which the posterior density is strictly log-concave with high probability. However, this comes at the cost of probabilistic error bounds, instead of a definitive one. We can upper bound the error of the LMC algorithm with the threshold in a polynomial number of iterations (in n, p .and ϵ^{-1}) with high probability. Since the proof strategy is quite general, the ideas can be used to establish these guarantees for other high-dimensional statistical models.

References

[1] A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," 2016.

[2] R. Nickl and S. Wang, "On polynomial-time computation of high-dimensional posterior measures by langevin-type algorithms," 2022.

[3] R. Altmeyer, "Polynomial time guarantees for sampling based posterior inference in high-dimensional generalised linear models," 2022.

[4] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, pp. 341–363, 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:18787082

[5] A. Dalalyan and A. Tsybakov, "Sparse regression learning by aggregation and langevin monte-carlo," *Journal of Computer and System Sciences*, vol. 78, no. 5, p. 1423–1443, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jcss.2011.12.023

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011. [Online]. Available: https://books.google.co.uk/books?id=2gsLkQlb8JAC

[7] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972. [Online]. Available: http://www.jstor.org/stable/2344614

[8] A. Durmus and E. Moulines, "High-dimensional bayesian inference via the unadjusted langevin algorithm," *Bernoulli*, vol. 25, pp. 2854–2882, 11 2019.

[9] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2016. [Online]. Available: https://books.google.co.uk/books?id=5p8SDAEACAAJ

[10] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, ser. Graduate Texts in Mathematics (113) (Book 113). Springer New York, 1991. [Online]. Available: https://books.google.co.uk/books?id=ATNy_Zg3PSsC