Sampling Can Be Faster Than Optimization

Yi-An Ma[a], Yuansi Chen[b], Chi Jin[a], Nicolas Flammarion[a], and Michael I. Jordan*[a, b]

[a]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

[b]Department of Statistics, University of California, Berkeley, CA 94720

November 21, 2018

Abstract

Optimization algorithms and Monte Carlo sampling algorithms have provided the computational foundations for the rapid growth in applications of statistical machine learning in recent years. There is, however, limited theoretical understanding of the relationships between these two kinds of methodology, and limited understanding of relative strengths and weaknesses. Moreover, existing results have been obtained primarily in the setting of convex functions (for optimization) and log-concave functions (for sampling). In this setting, where local properties determine global properties, optimization algorithms are unsurprisingly more efficient computationally than sampling algorithms. We instead examine a class of *nonconvex* objective functions that arise in mixture modeling and multi-stable systems. In this nonconvex setting, we find that the computational complexity of sampling algorithms scales linearly with the model dimension while that of optimization algorithms scales exponentially.

Machine learning and data science are fields that blend computer science and statistics so as to solve inferential problems whose scale and complexity require modern computational infrastructure. The algorithmic foundations on which these blends have been based repose on two general computational strategies, both which have their roots in mathematics—optimization and Markov chain Monte Carlo (MCMC) sampling. Research on these strategies has mostly proceeded separately, with research on optimization focused on estimation and prediction problems, and with research on sampling focused on tasks that require uncertainty estimates, such as forming confidence intervals and conducting hypothesis tests. There is a trend, however, towards the use of common methodological elements within the two strands of research. In particular, both strands have focused on the use of gradients and stochastic gradients—rather than function values or higher-order derivatives—as providing a useful compromise between the computational complexity of individual algorithmic steps and the overall rate of convergence. Empirically, the effectiveness of this compromise is striking. But the relative paucity of theoretical research linking optimization and sampling has limited the flow of ideas; in particular, the rapid recent advance of theory for optimization [34, see, e.g.,] has not yet translated into a similarly rapid advance of the theory for sampling. Accordingly, machine learning has remained limited in its inferential scope, with little concern for estimates of uncertainty.

Theoretical linkages have begun to appear in recent work [13, 15, 14, 11, 9, 17, 30, 31, see, e.g.,], where tools from optimization theory have been used to establish rates of convergence—notably including dimension dependence—for MCMC sampling. The overall message from these results is that sampling is slower than optimization—a message which accords with the folk wisdom that sampling approaches are warranted only if there is need for the stronger inferential outputs that they provide. These results are, however, obtained in the setting of *convex* functions. For convex functions, global properties can be assessed via local information. Not surprisingly, gradient-based optimization is well suited to such a setting.

*jordan@cs.berkeley.edu

Input: $\mathbf{x}^0$, stepsizes $\{h^k\}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - h^k \nabla U(\mathbf{x}^k) + \xi$
  if $\dfrac{p\left(\mathbf{x}^k|\mathbf{x}^{k+1}\right) p^*(\mathbf{x}^k)}{p\left(\mathbf{x}^{k+1}|\mathbf{x}^k\right) p^*\left(\mathbf{x}^{k+1}\right)} < u$ **then**
   $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k$ ▷ Metropolis Adjustment
Return $\mathbf{x}^K$

Figure 1: The (Metropolis adjusted) Langevin algorithm is a gradient-based MCMC algorithm. In each step, one simulates $\xi \sim \mathcal{N}(0, 2h^k\mathbb{I})$ and $u \sim \mathcal{U}[0,1]$ a uniform random variable between 0 and 1. The conditional distribution $p\left(\mathbf{x}^k|\mathbf{x}^{k+1}\right)$ is the normal distribution centered at $\mathbf{x}^k - h^k\nabla U(\mathbf{x}^k)$ and p^* is the target distribution. Without the Metropolis adjustment step, the algorithm is called the unadjusted Langevin algorithm (ULA). Otherwise, it is called the Metropolis Adjusted Langevin Algorithm (MALA).

Our focus is the *nonconvex* setting. We consider a broad class of problems that are strongly convex outside of a bounded region, but nonconvex inside of it. Such problems arise, for example, in a Bayesian mixture model problems with a proper prior [33, 32], and in the noisy multi-stable models that are common in statistical physics [24, 25]. We find that when the nonconvex region has a constant and nonzero radius in $\mathbb{R}^d$, the MCMC methods converge to ϵ accuracy in $\widetilde{\mathcal{O}}\left(d/\epsilon\right)$ or $\widetilde{\mathcal{O}}\left(d^2 \ln\left(1/\epsilon\right)\right)$ steps whereas any optimization approach converges in more than $\widetilde{\mathcal{O}}\left((1/\epsilon)^d\right)$ steps. Thus, for this class of problems, sampling is more effective than optimization.

We obtain these polynomial convergence results for the MCMC algorithms in the nonconvex setting by working in continuous time and separating the problem into two subproblems: given the target distribution we first exploit the properties of a weighted Sobolev space endowed with that target distribution to obtain convergence rates for the continuous dynamics, and we then discretize and find the appropriate step size to retain those rates for the discretized algorithm. This general framework allows us to strengthen recent results in the MCMC literature [18, 7, 10, 29] and examine a broader class of algorithms including the celebrated Metropolis-Hastings method.

1 Polynomial Convergence of MCMC Algorithms

The *Langevin algorithm* is a family of gradient-based MCMC sampling algorithms [42, 41, 16]. We present pseudocode for two variants of the algorithm in Algorithm 1, and, by way of comparison, we provide pseudocode for classical gradient descent in Algorithm 2. The variant of the Langevin algorithm which does not include the "if" statement is referred to as the *unadjusted Langevin algorithm* (ULA); as can be seen, it is essentially the same as gradient descent, differing only in its incorporation of a random term $\xi \sim \mathcal{N}(0, 2h^k\mathbb{I})$ in the update. The variant that includes the "if" statement is referred to as the *Metropolis Adjusted Langevin Algorithm* (MALA); it is the standard Metropolis-Hastings algorithm applied to the Langevin setting.

We consider sampling from a smooth target distribution p^* that is strongly log-concave outside of a region. We prove convergence of the Langevin sampling algorithms for this target, establishing a convergence rate. Given an error tolerance $\epsilon \in (0,1)$ and an initial distribution p^0, define the ϵ-mixing time in total variation distance as

$$\tau(\epsilon; p^0) = \min\left\{k \middle| \left\|p^k - p^*\right\|_{\mathrm{TV}} \leq \epsilon\right\}.$$

Gradient Descent

Input: $\mathbf{x}^0$, stepsizes $\{h^k\}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - h^k \nabla U(\mathbf{x}^k)$
Return $\mathbf{x}^K$

Figure 2: Gradient descent is a classical gradient-based optimization algorithm which updates $\mathbf{x}$ along the negative gradient direction.

Theorem 1. *For $p^* \propto e^{-U}$, we assume that U is m-strongly convex outside of a region of radius R and L-Lipschitz smooth (see the Supplement for a formal statement of the assumptions). Let $\kappa = L/m$ denote the condition number of U. Consider Algorithm 1 with initialization $p^0 = \mathcal{N}\left(0, \frac{1}{L}\mathbb{I}_d\right)$ and error tolerance $\epsilon \in (0,1)$. Then ULA satisfies*

$$\tau_{ULA}(\epsilon, p^0) \leq \mathcal{O}\left(e^{32LR^2}\kappa^2 \frac{d}{\epsilon^2} \ln\left(\frac{d}{\epsilon^2}\right)\right). \tag{1}$$

For MALA,

$$\tau_{MALA}(\epsilon, p^0) \leq \mathcal{O}\left(e^{16LR^2}\kappa^{3/2} d^{1/2} \left(d \ln \kappa + \ln\left(\frac{1}{\epsilon}\right)\right)^{3/2}\right). \tag{2}$$

Comparing (9) with (10), we see that the Metropolis adjustment improves the mixing time of ULA to a logarithmic dependence in ϵ, while sacrificing a factor of dimension d. Comparing (9) and (10) with previous results in the literature that provide upper bounds on the mixing time of ULA and MALA for strongly convex potentials U [13, 15, 14, 11, 9, 17, 30, 31], we find that the local nonconvexity results in an extra factor of $e^{\mathcal{O}\left(LR^2\right)}$. Thus, when the Lipschitz smoothness L and radius of the nonconvex region R satisfy LR^2 is $\mathcal{O}(\log d)$, the computational complexity is polynomial in dimension d.

Our proof of Theorem 3 involves a two-step framework that applies more widely than our specific setting. We first use properties of $p^* \propto e^{-U}$ to establish linear convergence of a continuous stochastic process that underlies Algorithm 1. We then discretize, finding an appropriate step size for the algorithm to converge to the desired accuracy. These two parts can be tackled independently. In this section, we provide an overview of the first part of the argument in the case of the MALA algorithm. The details, as well as a presentation of the second part of the argument, are provided in the Supplement.

Letting $t = \sum_k h^k$, assumed finite, a standard limiting process yields the following stochastic differential equation (SDE) as a continuous-time limit of Algorithm 1: $\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$, where B_t is a Brownian motion. To assess the rate of convergence of this SDE, we make use of the Kullback-Leibler (KL) divergence, which upper bounds the total variation distance and allows us to obtain strong convergence guarantees that include dimension dependence. Denoting the probability distribution of $\mathbf{X}_t$ as p_t, we obtain (see the derivation in the Supplement) the following time derivative of the divergence of p_t to the target distribution p^*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t \parallel p^*) = -\mathbb{E}_{p_t}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{x})}{p^*(\mathbf{x})}\right)\right\|^2\right]. \tag{3}$$

The property of $p^* \propto e^{-U}$ that we require to turn this time derivative into a convergence rate is that it satisfies a *log-Sobolev inequality*. Considering the Sobolev space defined by the weighted L^2 norm: $\int g(\mathbf{x})^2 p^*(\mathbf{x})\mathrm{d}\mathbf{x}$, we say that p^* satisfies a log-Sobolev inequality if there exists a constant $\rho > 0$ such that for any smooth function g on $\mathbb{R}^d$, satisfying $\int_{\mathbb{R}^d} g(\mathbf{x})p^*(\mathbf{x})\mathrm{d}\mathbf{x} = 1$, we have:

$$\int g(\mathbf{x}) \ln g(\mathbf{x}) \cdot p^*(\mathbf{x})\mathrm{d}\mathbf{x} \leq \frac{1}{2\rho} \int \frac{\|\nabla g(\mathbf{x})\|^2}{g(\mathbf{x})} p^*(\mathbf{x})\mathrm{d}\mathbf{x}.$$

The largest ρ for which this inequality holds is said to be the *log-Sobolev constant* for the objective U. We denote it as ρ_U. Taking $g = p_t/p^*$, we obtain:

$$\mathrm{KL}(p_t \parallel p^*) = \mathbb{E}_{p_t}\left[\ln\left(\frac{p_t(\mathbf{x})}{p^*(\mathbf{x})}\right)\right] \leq \frac{1}{2\rho_U}\mathbb{E}_{p_t}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{x})}{p^*(\mathbf{x})}\right)\right\|^2\right]. \tag{4}$$

Note the resemblance of this bound to the Polyak-Łojasiewicz condition [39] used in optimization theory for studying the convergence of smooth and strongly convex objective functions—in both cases the difference from the current iterate to the optimum is upper bounded by the norm of the gradient squared. Combining (3) with (4), we derive the promised linear convergence rate for the continuous process:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t \parallel p^*) \leq -2\rho_U \mathrm{KL}(p_t \parallel p^*).$$

In the Supplement, we present similar results for the ULA algorithm, again using the KL divergence.

The next step is to bound ρ_U in terms of the basic smoothness and local nonconvexity assumptions in our problem. We first require an approximation result:

Lemma 1. *For U m-strongly convex outside of a region of radius R and L-Lipschitz smooth, there exists $\hat{U} \in C^1(\mathbb{R}^d)$ such that $\hat{U}$ is $m/2$ strongly convex on $\mathbb{R}^d$, and has a Hessian that exists everywhere on $\mathbb{R}^d$. Moreover, we have* $\sup\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right) - \inf\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right) \leq 16LR^2$.

The proof of this lemma is presented in the Supplement. The existence of the smooth approximation established in this lemma can now be used to bound the log-Sobolev constant using standard results.

Proposition 1. *For $p^* \propto e^{-U}$, where U is m-strongly convex outside of a region of radius R and L-Lipschitz smooth,*

$$\rho_U \geq \frac{m}{2}e^{-16LR^2}. \tag{5}$$

Proof For $m/2$-strongly convex $\hat{U} \in C^1(\mathbb{R}^d)$ whose Hessian $\nabla^2\hat{U}(\mathbf{x})$ exists everywhere on $\mathbb{R}^d$, the distribution $e^{-\hat{U}(\mathbf{x})}$ satisfies the Bakry-Emery criterion [3] for a strongly log-concave density, which yields:

$$\rho_{\hat{U}} \geq \frac{m}{2}. \tag{6}$$

We use the Holley-Stroock theorem [20] to obtain:

$$\rho_U \geq \frac{m}{2}e^{-\left|\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\right|} \geq \frac{m}{2}e^{-16LR^2}. \tag{7}$$

■

We see from this proof outline that our approach enables one to adapt existing literature on the convergence of diffusion processes [27, 44, 45] to work out suitable log-Sobolev bounds and thereby obtain sharp convergence rates in terms of distance measures such as the KL divergence and total variation. The detailed proof also reveals that the log-Sobolev constant ρ_U is largely determined by the global qualities of U where most of the probability mass is concentrated; local properties of U have limited influence on ρ_U.

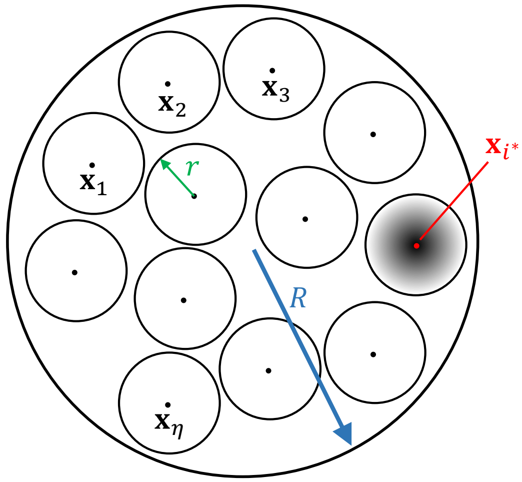

Figure 3: Depiction of an instance of $U(\mathbf{x})$, inside the region of radius R. that attains the lower bound.

2 Exponential Dependence on Dimension for Optimization

It is well known that finding global minima of a general nonconvex optimization problem is NP-hard [21]. Here we demonstrate that it is also hard to find an ϵ approximation to the optimum for a Lipschitz smooth, locally nonconvex objective function U.

Specifically, we consider a general iterative algorithm family $\mathcal{A}$ which, at every step k, is allowed to query not only the function value of U but also its derivatives up to any fixed order at a chosen point $\mathbf{x}^k$. Thus the algorithm has access to the vector $(\{U(\mathbf{x}^k), \nabla U(\mathbf{x}^k), \cdots, \nabla^n U(\mathbf{x}^k)\}$, for any fixed $n \in \mathcal{N})$. Moreover, the algorithm can use the entire query history to determine the next point $\mathbf{x}^{k+1}$, and it can do so randomly or deterministically. In the following theorem, we prove that the number of iterations for any algorithm in $\mathcal{A}$ to approximate the minimum of U is necessarily exponential in the dimension d.

Theorem 2 (Lower bound for optimization)**.** *For any $R > 0$, $L \geq 2m > 0$, and $\epsilon \leq \mathcal{O}(LR^2)$, there exists an objective function, $U : \mathbb{R}^d \to \mathbb{R}$, which is m-strongly convex outside of a region of radius $2R$ and L-Lipschitz smooth, such that any algorithm in $\mathcal{A}$ requires at least $K = \Omega((LR^2/\epsilon)^{d/2})$ iterations to guarantee that $\min_{k\leq K} |U(\mathbf{x}^K) - U(\mathbf{x}^*)| < \epsilon$ with constant probability.*

We remark that Theorem 4 is an information-theoretic result based on the class of iterative algorithms $\mathcal{A}$ and the forms of the queries to this class. It is thus an unconditional statement that does not depend on conjectures such as $P \neq NP$ in complexity theory.

A depiction of an example that achieves this computational lower bound is provided in Fig. 4. The idea is that we can pack exponentially many balls of radius less than $R/3$ inside a region of radius R. We can randomly assign the minimum $\mathbf{x}^*$ to one of the balls, assigning a larger constant value to the other balls. We show that the number of queries needed to find the specific ball containing the minimum is exponential in d. Moreover, the difference from $U(\mathbf{x}^*)$ to any other point outside of the ball is $\mathcal{O}(LR^2)$, which can be significant.

This example suggests that the lower-bound scenario will be realized in cases in which regions of attraction are small around a global minimum and behavior within each region of attraction is relatively autonomous. This phenomenon is not uncommon in multi-stable physical systems. Indeed, in non-equilibrium statistical physics, there are examples where the global behavior of a system can be treated approximately as a set of local behaviors within stable regimes plus Markov transitions among stable regimes [19]. In such cases, when the regions of attraction are small, the computational complexity to find the global minimum can be combinatorial. In Sec. 3, we explicitly demonstrate that this combinatorial complexity holds for a Gaussian mixture model.

2.1 Why Can't One Optimize in Polynomial Time Using the Langevin Algorithm?

Consider the rescaled density function $q_\beta^* \propto e^{-\beta U}$. A line of research beginning with simulated annealing [23] uses a sampling algorithm to sample from q_β^*, doing so for increasing values of β, and uses the resulting samples to approximate $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x})$. In particular, simply returning one of the samples obtained for sufficiently large β will yield an output that is close to the optimum with high probability. This suggests the following natural question: Can we use the Langevin algorithm to generate samples from q_β^*, and thereby obtain an approximation to $\mathbf{x}^*$ in a number of steps polynomial in d?

In the following Corollary 2, we demonstrate that this is *not* possible: We need $\beta = \widetilde{\Omega}\left(d/\epsilon\right)$ so that a sample $\mathbf{x}$ from q_β^* will satisfy $\|\mathbf{x}-\mathbf{x}^*\| \le \epsilon$ with constant probability. This requires the Lipschitz smoothness of U to scale with d, which in turn cause the number of iterations required to scale exponentially with d, as established in (9) and (10)).

Corollary 1. *There exists an objective function U that is m-strongly convex outside of a region of radius $2R$ and L-Lipschitz smooth, such that, for $\hat{\mathbf{x}} \sim q_\beta^*$, it is necessary that $\beta = \widetilde{\Omega}\left(d/\epsilon\right)$ in order to have $U(\hat{\mathbf{x}}) - U\left(\mathbf{x}^*\right) < \epsilon$ with constant probability. Moreover, the number of iterations required for the Langevin algorithms to achieve $U(\mathbf{x}^K) - U\left(\mathbf{x}^*\right) < \epsilon$ with constant probability is $K = e^{\widetilde{\mathcal{O}}\left(d\cdot LR^2/\epsilon\right)}$.*

It should be noted that this upper bound for the Langevin algorithms agrees with the lower bound for optimization algorithms in Theorem 4 up to a factor of LR^2/ϵ in the exponent. This is because in the optimization complexity lower bound, we are considering the least number of queries of function values and n-th order derivatives required to obtain a result that is within ϵ of $\mathbf{x}^*$. A hypothetical optimization algorithm can therefore query one point to determine whether one region contains the global minimum or not. When using the Langevin algorithms, $\widetilde{\mathcal{O}}\left(LR^2/\epsilon\right)$ steps are required to explore each local region to a constant level of confidence.

3 Parameter Estimation from Gaussian Mixture Model: Sampling versus Optimization

We have seen that for problems with local nonconvexity, the computational complexity for the Langevin algorithm is polynomial in dimension whereas it is exponential in dimension for optimization algorithms. These are, however, worst-case guarantees. It is important to consider whether they also hold for natural statistical problem classes and for specific optimization algorithms. In this section, we study the Gaussian mixture model, comparing Langevin sampling and the popular expectation-maximization (EM) optimization algorithm.

Consider the problem of inferring the mean parameters of a Gaussian mixture model, $\boldsymbol{\mu} = \{\mu_1, \cdots, \mu_M\} \in \mathbb{R}^{d\times M}$, when N data points are sampled from that model. Letting $\mathbf{y} = \{y_1, \cdots, y_N\}$ denote the data, we have:

$$p\left(y_n|\boldsymbol{\mu}\right) = \sum_{i=1}^M \frac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}(y_n-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(y_n-\mu_i)\right) + \left(1-\sum_{i=1}^M \lambda_i\right) p_0(y_n), \tag{8}$$

where Z_i are normalization constants and $\sum_{i=1}^M \lambda_i \le 1$. $p_0(y_n)$ represents general constraints on the data (e.g., data may be distributed inside a region or may have sub-Gaussian tail behavior). The objective function is given by the log posterior distribution: $U(\boldsymbol{\mu}) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right)$. Assume data are distributed in a bounded region ($\|y_n\| \leq R$) and take $p_0(y_n) = \mathbb{1}\left\{\|y_n\| \leq R\right\}/Z_0$.

We prove in the Supplement that for a suitable choice of the prior $p(\boldsymbol{\mu})$ and weights $\{\lambda_i\}$, the objective function is Lipschitz smooth and strongly convex for $\|\boldsymbol{\mu}\| \geq 2R\sqrt{M}$. Therefore, taking $MR^2 = \mathcal{O}(\log d)$, the ULA and MALA algorithms converge to ϵ accuracy within $K \leq \widetilde{\mathcal{O}}\left(d^3/\epsilon\right)$ and $K \leq \widetilde{\mathcal{O}}\left(d^3 \ln^2\left(1/\epsilon\right)\right)$ steps, respectively.

The EM algorithm updates the value of $\boldsymbol{\mu}$ in two steps. In the expectation (E) step a weight is computed for each data point and each mixture component, using the current parameter value $\boldsymbol{\mu}_k$. In the maximization (M) step the value of $\boldsymbol{\mu}_{k+1}$ is updated as a weighted sample mean (see the Supplement for a more detailed description). It is standard to initialize the EM algorithm by randomly selecting M data points (sometimes with small perturbations) to form $\boldsymbol{\mu}_0$. We demonstrate in the Supplement that under the condition that $MR^2 = \mathcal{O}(\log d)$, there exists a dataset $\{y_1, \cdots, y_N\}$ and covariances $\{\Sigma_1, \cdots, \Sigma_M\}$, such that the EM algorithm requires more than $K \geq \min\{\mathcal{O}(d^{1/\epsilon}), \mathcal{O}(d^d)\}$ queries to converge if one initializes the algorithm close to the given data points. That is, for big ϵ, the computational complexity of the EM algorithm depends on d with arbitrarily high order (depending on ϵ); for small ϵ, the computational complexity of the EM algorithm scales exponentially with d. The latter case corresponds to our lower bound in Theorem 4 when taking the radius of the convex region of $\boldsymbol{\mu}$ to scale with $\sqrt{\log d}$. Therefore, it is significantly harder for the EM algorithm to converge if we initialize the algorithm close to the given data points.

It can be seen that many mixture models with strongly log-concave priors fall into the assumed class of distributions with local nonconvexity. If data are distributed relatively close to each other, sampling these distributions can often be easier than searching for their global minima. This scenario is also common in the setting of the noisy multi-stable models arising in statistical physics (e.g., where the negative log likelihood is the potential energy of a classical particle system in an external field [25]) and related fields.

4 Discussion

We have shown that there is a natural family of nonconvex functions for which sampling algorithms have polynomial complexity in dimension whereas optimization algorithms display exponential complexity. The intuition behind these results is that computational complexity for optimization algorithms depends heavily on the local properties of the objective function U. This is consistent with a related phenomenon that has been studied in the optimization—local strong convexity near the global optimum can improve the convergence rate of convex optimization [2]). On the other hand, sampling complexity depends more heavily on the global properties of U. This is also consistent with existing literature; for example, it is known that the dimension dependence of the Langevin algorithm upper bounds deteriorates when U changes from strongly convex to weakly convex. This corresponds to the fact that the sub-Gaussian tails for strongly log-concave distributions are easier to explore than the sub-exponential tails for log-concave distributions.

A scrutiny of the relative scale between radius of the nonconvex region R and the dimension d is interesting (for constant Lipschitz smoothness L): when R is a constant or less than $\mathcal{O}(\log d)$, sampling is generally easier than optimization; when $R \leq \sqrt{d}$, the convergence upper bound for sampling is still slightly smaller than the optimization complexity lower bound; when $R > \sqrt{d}$, the comparison is indeterminate; and the converse is true if $R > d$.

The relatively rapid advance of the theory of gradient-based optimization in recent years has been due in part to the development of lower bounds, of the kind exhibited in our Theorem 4, for broad classes of algorithms. It is of interest to develop such lower bounds for MCMC algorithms, particularly bounds that capture dimension dependence. It is also of interest to develop both lower bounds and upper bounds for other forms of nonconvexity. For example, there has been recent work studying strongly dissipative functions [40]. Here the worst-case convergence bounds have exponential dependence on the dimension, but $p^* \propto e^{-U}$ has a sub-Gaussian tail; thus, further exploration of this setting may yield milder conditions on U that allow MCMC algorithms to have polynomial convergence rates.

Appendices

Appendix A Assumptions on the Objective Function U

Assumptions on $U : \mathbb{R}^d \to \mathbb{R}$ (local nonconvexity):

1. $U(\mathbf{x})$ is L-Lipschitz smooth and its Hessian exists $\forall \mathbf{x} \in \mathbb{R}^d$.

   That is: $U \in C^1(\mathbb{R}^d)$, $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, $\|\nabla U(\mathbf{x}) - \nabla U(\mathbf{z})\| \leq L \|\mathbf{x} - \mathbf{z}\|$; $\forall \mathbf{x} \in \mathbb{R}^d$, $\nabla^2 U(\mathbf{x})$ exists.

2. $U(\mathbf{x})$ is m-strongly convex for $\|\mathbf{x}\| > R$.

   That is: $V(\mathbf{x}) = U(\mathbf{x}) - \dfrac{m}{2} \|\mathbf{x}\|_2^2$ is convex on $\Omega = \mathbb{R}^d \setminus \mathbb{B}(0, R)$ [1]. We then follow the definition of convexity on nonconvex domains [38, 46] to require that $\forall \mathbf{x} \in \Omega$, any convex combination of $\mathbf{x} = \lambda_1 \mathbf{x}_1 + \cdots + \lambda_k \mathbf{x}_k$ with $\mathbf{x}_1, \cdots, \mathbf{x}_k \in \Omega$ satisfy:

   $$V(\mathbf{x}) \leq \lambda_1 V(\mathbf{x}_1) + \cdots + \lambda_k V(\mathbf{x}_k).$$

   We further denote the "condition number" of U on Ω as $\kappa = L/m$.

3. For convenience, let $\nabla U(0) = 0$ (i.e., 0 is a local extremum).

Appendix B Proofs for Sampling

Theorem 3. *For $p^* \propto e^{-U}$, we assume that U satisfies the local nonconvexity Assumptions 1–3. Consider unadjusted Langevin algorithm (ULA) and Metropolis adjusted Langevin algorithm (MALA) with initialization $p^0 = \mathcal{N}\left(0, \frac{1}{L}\mathbb{I}_d\right)$ and error tolerance $\epsilon \in (0,1)$. Then ULA satisfies*

$$\tau_{ULA}(\epsilon, p^0) \leq \mathcal{O}\left(e^{32LR^2} \kappa^2 \frac{d}{\epsilon^2} \ln\left(\frac{d}{\epsilon^2}\right)\right). \tag{9}$$

For MALA,

$$\tau_{MALA}(\epsilon, p^0) \leq \mathcal{O}\left(e^{16LR^2} \kappa^{3/2} d^{1/2} \left(d \ln \kappa + \ln\left(\frac{1}{\epsilon}\right)\right)^{3/2}\right). \tag{10}$$

We first prove the following Lemma 2 to obtain the log-Sobolev inequality (Proposition 2 below). We then prove convergence of ULA and MALA respectively in Sec. B.2 and B.3.

B.1 Log-Sobolev Inequality

Proposition 2. *For $p^* \propto e^{-U}$ where U satisfies Assumptions 1–3 in Appendix A,*

$$\rho_U \geq \frac{m}{2} e^{-16LR^2}. \tag{11}$$

Proof First note that for $m/2$-strongly convex $\hat{U} \in C^1(\mathbb{R}^d)$ with $\nabla^2 \hat{U}(\mathbf{x})$ exists on the entire $\mathbb{R}^d$, distribution $e^{-\hat{U}(\mathbf{x})}$ satisfies the Bakry-Emery criterion [3] for strongly log concave density and have:

$$\rho_{\hat{U}} \geq \frac{m}{2}. \tag{12}$$

[1] Here we have denoted $\mathbb{B}(0, R)$ to be the closed ball of radius R centered at 0.

Next we invoke Lemma 2 that such $\hat{U}$ exists and satisfies $\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\leq 16LR^2$.

Then we use the result from Holley-Stroock [20] and obtain:

$$\rho_U \geq \frac{m}{2}e^{-\left|\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\right|} \geq \frac{m}{2}e^{-16LR^2}. \tag{13}$$

■

Lemma 2. *For U satisfying Assumptions 1–3, there exists $\hat{U}\in C^1(\mathbb{R}^d)$ with its Hessian exists everywhere on $\mathbb{R}^d$, and $\hat{U}$ being $m/2$ strongly convex on $\mathbb{R}^d$, such that* $\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\leq 16LR^2$.

Proof of Lemma 2 Similar to Assumptions 1–3, denote $\Omega=\mathbb{R}^d\setminus\mathbb{B}(0,R)$. Also denote $\tilde{U}(\mathbf{x})=U(\mathbf{x})-\frac{m}{4}\left\|\mathbf{x}\right\|^2$.

We follow [46] to construct $\hat{U}(\mathbf{x})-\frac{m}{4}\left\|\mathbf{x}^2\right\|\in C^1(\mathbb{R}^d)$ with Hessian defined on $\mathbb{R}^d$ so that it's convex on $\mathbb{R}^d$ and differs from $\tilde{U}(\mathbf{x})$ less than $16LR^2$.

First we define function V as the convex extension [43] of $\tilde{U}$ on the domain of Ω:

$$V(\mathbf{x})=\inf_{\{\mathbf{x}_i\}\subset\Omega,\{\lambda_i\left|\sum_i\lambda_i=1\right.\}}\left\{\sum_{i=1}^{l}\lambda_i\tilde{U}(\mathbf{x}_i)\right\},\quad \forall\mathbf{x}\in\Omega^{co}=\mathbb{R}^d. \tag{14}$$

Then $V(\mathbf{x})$ is convex on the entire $\mathbb{R}^d$. Also, since $\tilde{U}(\mathbf{x})$ is convex in Ω, $V(\mathbf{x})=\tilde{U}(\mathbf{x})$ for $\mathbf{x}\in\Omega$. By Lemma 3, we also know that $\forall\mathbf{x}\in\mathbb{B}(0,R)$, $\inf_{\bar{\mathbf{x}}=R}\tilde{U}(\bar{\mathbf{x}})\leq V(\mathbf{x})\leq\sup_{\bar{\mathbf{x}}=R}\tilde{U}(\bar{\mathbf{x}})$.

Next we construct $\tilde{V}(\mathbf{x})$ to be a smoothing of V on $\mathbb{B}\left(0,\frac{4}{3}R\right)$. Let $\phi\geq 0$ be a smooth function supported on the ball $\mathbb{B}(0,\delta)$ where $\delta=\frac{m}{L}\frac{R}{1600}<\frac{R}{6}$ such that $\int\phi(\mathbf{x})\mathrm{d}\mathbf{x}=1$. Define

$$\tilde{V}(\mathbf{x})=\int V(\mathbf{y})\phi(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{y}=\int V(\mathbf{x}-\mathbf{y})\phi(\mathbf{y})\mathrm{d}\mathbf{y}. \tag{15}$$

Then $\tilde{V}$ is a smooth and convex function on $\mathbb{R}^d$. The second expression in (15) implies that $\tilde{V}(\mathbf{x})$ is $\frac{m}{2}$-strongly convex in $\mathbb{R}^d\setminus\mathbb{B}\left(0,R+\delta\right)\supset\mathbb{B}\left(0,\frac{3}{2}R\right)\setminus\mathbb{B}\left(0,\frac{4}{3}R\right)$. Also note that the definition of $\tilde{V}$ implies that $\forall\left\|\mathbf{x}\right\|<\frac{4}{3}R$,

$$\inf_{\|\bar{\mathbf{x}}\|<\frac{4}{3}R+\delta}V(\bar{\mathbf{x}})\leq\tilde{V}(\mathbf{x})\leq\sup_{\|\bar{\mathbf{x}}\|<\frac{4}{3}R+\delta}V(\bar{\mathbf{x}}).$$

And by Lemma 3,

$$\inf_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{4}{3}R+\delta\right)\setminus\mathbb{B}(0,R)}\tilde{U}(\bar{\mathbf{x}})\leq\tilde{V}(\mathbf{x})\leq\sup_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{4}{3}R+\delta\right)\setminus\mathbb{B}(0,R)}\tilde{U}(\bar{\mathbf{x}}),\quad\forall\left\|\mathbf{x}\right\|<\frac{4}{3}R. \tag{16}$$

Finally, we construct the auxiliary function $\hat{U}(\mathbf{x})$:

$$\hat{U}(\mathbf{x})-\frac{m}{4}\left\|\mathbf{x}^2\right\|=\begin{cases}\tilde{U}(\mathbf{x}), & \|\mathbf{x}\|>\frac{3}{2}R\\ \alpha(\mathbf{x})\tilde{U}(\mathbf{x})+(1-\alpha(\mathbf{x}))\tilde{V}(\mathbf{x}), & \frac{4}{3}R<\|\mathbf{x}\|<\frac{3}{2}R\\ \tilde{V}(\mathbf{x}), & \|\mathbf{x}\|<\frac{4}{3}R\end{cases}, \tag{17}$$

where $\alpha(\mathbf{x}) = \frac{1}{2}\cos\left(\frac{36\pi}{17}\frac{\|\mathbf{x}\|^2}{R^2} - \frac{64\pi}{17}\right) + \frac{1}{2}$. Here we know that $\tilde{U}(\mathbf{x})$ is $\frac{m}{2}$-strongly convex and smooth in $\mathbb{R}^d \setminus \mathbb{B}(0,R)$; $\tilde{V}(\mathbf{x})$ is $\frac{m}{2}$-strongly convex and smooth in $\mathbb{R}^d \setminus \mathbb{B}\left(0, \frac{4}{3}R\right)$. Hence for $\frac{4}{3}R < \|\mathbf{x}\| < \frac{3}{2}R$,

$$
\begin{aligned}
&\nabla^2\left(\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\|\right) \\
=&\nabla^2\tilde{U}(\mathbf{x}) + \nabla^2\left((1-\alpha(\mathbf{x}))(\tilde{V}(\mathbf{x}) - \tilde{U}(\mathbf{x}))\right) \\
=&\alpha(\mathbf{x})\nabla^2\tilde{U}(\mathbf{x}) + (1-\alpha(\mathbf{x}))\nabla^2\tilde{V}(\mathbf{x}) \\
&-\nabla^2\alpha(\mathbf{x})\left(\tilde{V}(\mathbf{x}) - \tilde{U}(\mathbf{x})\right) - 2\nabla\alpha(\mathbf{x})\left(\nabla\tilde{V}(\mathbf{x}) - \nabla\tilde{U}(\mathbf{x})\right)^T \\
\succeq&\frac{m}{2}\mathbb{I} - \nabla^2\alpha(\mathbf{x})\left(\tilde{V}(\mathbf{x}) - \tilde{U}(\mathbf{x})\right) - 2\nabla\alpha(\mathbf{x})\left(\nabla\tilde{V}(\mathbf{x}) - \nabla\tilde{U}(\mathbf{x})\right)^T.
\end{aligned}
$$

Note that for $\frac{4}{3}R < \|\mathbf{x}\| < \frac{3}{2}R$,

$$
\left\|\nabla\tilde{V}(\mathbf{x}) - \nabla\tilde{U}(\mathbf{x})\right\| = \int \left\|\nabla\tilde{U}(\mathbf{x}-\mathbf{y}) - \nabla\tilde{U}(\mathbf{x})\right\|\phi(\mathbf{y})\mathrm{d}\mathbf{y} \le L\delta.
$$

$$
\tilde{V}(\mathbf{x}) - \tilde{U}(\mathbf{x}) = \int\left(\tilde{U}(\mathbf{x}-\mathbf{y}) - \tilde{U}(\mathbf{x})\right)\phi(\mathbf{y})\mathrm{d}\mathbf{y} \le \frac{3}{2}LR\delta.
$$

Therefore, when $\frac{4}{3}R < \|\mathbf{x}\| < \frac{3}{2}R$,

$$
\nabla^2\left(\hat{U}(\mathbf{x}) - \frac{1}{4}\left\|\mathbf{x}^2\right\|\right) \succeq \frac{m}{2}\mathbb{I} - 3\pi\frac{L\delta}{R}\mathbb{I} - 54\pi^2\frac{L\delta}{R}\mathbb{I} \succeq \left(\frac{m}{2} - 800\frac{L\delta}{R}\right)\mathbb{I}.
$$

Since $\delta = \frac{m}{L}\frac{R}{1600}$, $\nabla^2\left(\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\|\right)$ is positive semi-definite for $\frac{4}{3}R < \|\mathbf{x}\| < \frac{3}{2}R$. Hence $\nabla^2\left(\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\|\right)$ is positive semi-definite on the entire $\mathbb{R}^d$, and $\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\|$ is convex on $\mathbb{R}^d$.

From (16), we know that for $\|\mathbf{x}\| \le \frac{3}{2}R$,

$$
\inf_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)\setminus\mathbb{B}(0,R)} \tilde{U}(\bar{\mathbf{x}}) \le \hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\| \le \sup_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)\setminus\mathbb{B}(0,R)} \tilde{U}(\bar{\mathbf{x}}).
$$

Therefore,

$$
\begin{aligned}
&\sup\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right) - \inf\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right) \\
=&\sup\left(\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\| - \tilde{U}(\mathbf{x})\right) - \inf\left(\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\| - \tilde{U}(\mathbf{x})\right) \\
\le&2\left(\sup_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)\setminus\mathbb{B}(0,R)} \tilde{U}(\bar{\mathbf{x}}) - \inf_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)\setminus\mathbb{B}(0,R)} \tilde{U}(\bar{\mathbf{x}})\right) \\
\le&2\left(\sup_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)} \tilde{U}(\bar{\mathbf{x}}) - \inf_{\bar{\mathbf{x}}\in\mathbb{B}\left(0,\frac{3}{2}R+\delta\right)} \tilde{U}(\bar{\mathbf{x}})\right).
\end{aligned}
$$

Since U is L-smooth, $\tilde{U}$ is $\left(L+\frac{m}{2}\right)$-smooth and $\nabla\tilde{U}(0)=0$. Hence

$$\left|\tilde{U}(\mathbf{x})-\tilde{U}(0)-<\mathbf{x},\nabla\bar{U}(0)>\right|\leq\left(\frac{L}{2}+\frac{m}{4}\right)\|\mathbf{x}\|_2^2.$$

So for $\forall\,\|\mathbf{x}\|\leq\left(\frac{3}{2}R+\delta\right)$,

$$\sup_{\bar{\mathbf{x}}\in\mathbb{B}\left(\frac{3}{2}R+\delta\right)}\tilde{U}(\bar{\mathbf{x}})-\inf_{\bar{\mathbf{x}}\in\mathbb{B}\left(\frac{3}{2}R+\delta\right)}\tilde{U}(\bar{\mathbf{x}})\leq 8LR^2.$$

Hence

$$\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\leq 16LR^2.$$

$\blacksquare$

Lemma 3. *For function V defined in* (14), $\forall\mathbf{x}\in\mathbb{B}(0,R)$, $\inf_{\|\bar{\mathbf{x}}\|=R}\tilde{U}(\bar{\mathbf{x}})\leq V(\mathbf{x})\leq\sup_{\|\bar{\mathbf{x}}\|=R}\tilde{U}(\bar{\mathbf{x}})$.

Proof of Lemma 3 First, from the definition of V inside $\mathbb{B}(0,R)$:

$$\begin{aligned}V(\mathbf{x})&=\inf_{\{\mathbf{x}_i\}\subset\Omega,\{\lambda_i\,|\,\sum_i\lambda_i=1\}}\left\{\sum_{i=1}^{l}\lambda_i\tilde{U}(\mathbf{x}_i)\right\}\\&\leq\inf_{\{\mathbf{x}_i\}\subset\partial\Omega,\{\lambda_i\,|\,\sum_i\lambda_i=1\}}\left\{\sum_{i=1}^{l}\lambda_i\tilde{U}(\mathbf{x}_i)\right\}\\&\leq\sup_{\|\bar{\mathbf{x}}\|=R}\tilde{U}(\bar{\mathbf{x}}),\quad\forall\mathbf{x}\in\mathbb{B}(0,R),\end{aligned}$$

where the first inequality follows from the fact that $\partial\Omega\subset\Omega$ and that any $\mathbf{x}\in\mathbb{B}(0,R)$ can be represented as a convex combination of elements of $\partial\Omega$.

Next we prove that $\forall\mathbf{x}\in\mathbb{B}(0,R)$, $V(\mathbf{x})\geq\inf_{\|\bar{\mathbf{x}}\|=R}\tilde{U}(\bar{\mathbf{x}})$. Assume that at $\mathbf{x}\in\mathbb{B}(0,R)$, $V(\mathbf{x})$ is equal to a linear combination of $\{\mathbf{x}_i\}\subset\Omega=\mathbb{R}^d\setminus\mathbb{B}(0,R)$: $V(\mathbf{x})=\sum_i\lambda_i\tilde{U}(\mathbf{x}_i)$. We hereby prove that for any $\mathbf{x}_j\in\{\mathbf{x}_i\}$, such that $\|\mathbf{x}_j\|>R$, there exists a new convex combination $\{\mathbf{x}_i\}\bigcup\{\bar{\mathbf{x}}_j\}\setminus\{\mathbf{x}_j\}$ with $\|\bar{\mathbf{x}}_j\|=R$, such that $V(\mathbf{x})\geq\tilde{\lambda}_j\tilde{U}(\bar{\mathbf{x}}_j)+\sum_{i\neq j}\tilde{\lambda}_i\tilde{U}(\mathbf{x}_i)$.

$\exists\lambda_j<\bar{\lambda}_j<1$, such that $\bar{\mathbf{x}}_j$ defined below is a linear combination of $\mathbf{x}$ and $\mathbf{x}_j$ satisfying $\|\bar{\mathbf{x}}_j\|=R$:

$$\bar{\mathbf{x}}_j=\frac{1-\bar{\lambda}_j}{1-\lambda_j}\mathbf{x}+\frac{\bar{\lambda}_j-\lambda_j}{1-\lambda_j}\mathbf{x}_j.$$

Then $\bar{\mathbf{x}}_j$ is a convex combination of $\{\mathbf{x}_i\}$:

$$\bar{\mathbf{x}}_j=\bar{\lambda}_j\mathbf{x}_j+\left(\frac{1-\bar{\lambda}_j}{1-\lambda_j}\right)\left(\sum_{i\neq j}\lambda_i\mathbf{x}_i\right),$$

and since U is convex on Ω,

$$\tilde{U}(\bar{\mathbf{x}}_j)\leq\bar{\lambda}_j\tilde{U}(\mathbf{x}_j)+\left(\frac{1-\bar{\lambda}_j}{1-\lambda_j}\right)\left(\sum_{i\neq j}\lambda_i\tilde{U}(\mathbf{x}_i)\right).$$

On the other hand, we can reexpress $\mathbf{x}$ as a convex combination of $\{\mathbf{x}_i\} \bigcup \{\bar{\mathbf{x}}_j\} \setminus \{\mathbf{x}_j\}$:

$$\mathbf{x} = \frac{\lambda_j}{\bar{\lambda}_j}\bar{\mathbf{x}}_j + \left(1 - \frac{\lambda_j}{\bar{\lambda}_j}\frac{1-\bar{\lambda}_j}{1-\lambda_j}\right)\left(\sum_{i \neq j} \lambda_i \mathbf{x}_i\right) = \tilde{\lambda}_j \bar{\mathbf{x}}_j + \sum_{i \neq j} \tilde{\lambda}_i \mathbf{x}_i,$$

and that

$$\begin{aligned} V(\mathbf{x}) = \sum_i \lambda_i \tilde{U}(\mathbf{x}_i) &\geq \frac{\lambda_j}{\bar{\lambda}_j}\tilde{U}(\bar{\mathbf{x}}_j) + \left(1 - \frac{\lambda_j}{\bar{\lambda}_j}\frac{1-\bar{\lambda}_j}{1-\lambda_j}\right)\left(\sum_{i \neq j} \lambda_i \tilde{U}(\mathbf{x}_i)\right) \\ &= \tilde{\lambda}_j \tilde{U}(\bar{\mathbf{x}}_j) + \sum_{i \neq j} \tilde{\lambda}_i \tilde{U}(\mathbf{x}_i). \end{aligned}$$

Using an inductive argument, we obtain that $\forall \mathbf{x} \in \mathbb{B}(0, R)$, $V(\mathbf{x})$ is bigger than or equal to a certain convex combination of $\tilde{U}(\bar{\mathbf{x}}_i)$, where $\{\bar{\mathbf{x}}_i\} \subset \partial\Omega$. Therefore, $\forall \mathbf{x} \in \mathbb{B}(0, R)$, $V(\mathbf{x}) \geq \inf_{\|\bar{\mathbf{x}}\|=R} \tilde{U}(\bar{\mathbf{x}})$. ■

B.2 Proof of ULA Convergence Rate ((9) of Theorem 3)

Proof of (9) **of Theorem 3** We first quantify the convergence of distribution p following a stochastic process to the stationary distribution p^* through the Kullback-Leibler divergence (KL-divergence), $F(p)$:

$$F(p) = \int p(\mathbf{x}) \ln \left(\frac{p(\mathbf{x})}{p^*(\mathbf{x})}\right) \mathrm{d}\mathbf{x},$$

if $p(\mathbf{x})$ is absolutely continuous with respect to $p^*(\mathbf{x})$; and $F(p) = \infty$ otherwise. Then we use Pinsker inequality to bound the total variation norm, for two densities p and p^*

$$\|p - p^*\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(p \parallel p^*)} = \sqrt{2F(p)}.$$

Here we take the process as a discretized Langevin dynamics:

$$\mathbf{X}_{(k+1)h} = \mathbf{X}_{kh} - \nabla U(\mathbf{X}_{kh})h + \sqrt{2}(B_{(k+1)h} - B_{hk}), \tag{18}$$

which is equivalent to defining for $kh < t \leq (k+1)h$:

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_{kh})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{19}$$

For dynamics within $kh < t \leq (k+1)h$, we have from the Girsanov theorem [35] that $\mathbf{X}_t$ admits a density function p_t with respect to the Lebesgue measure. This density function can also be represented as $p_t(\mathbf{x}) = \int p_{kh}(\mathbf{y})p(\mathbf{x}, t|\mathbf{y}, kh)\mathrm{d}\mathbf{y}$, where $p(\mathbf{x}, t|\mathbf{y}, kh)$ is the solution to the following Kolmogorov forward equation in the weak sense [37]

$$\frac{\partial p(\mathbf{x}, t|\mathbf{y}, kh)}{\partial t} = \nabla^T \left(\nabla p(\mathbf{x}, t|\mathbf{y}, kh) + \nabla U(\mathbf{y})p(\mathbf{x}, t|\mathbf{y}, kh)\right),$$

where $p(\mathbf{x}, t|\mathbf{y}, kh)$ and its derivatives are defined via $P_t(f) = \int f(\mathbf{x})p(\mathbf{x}, t|\mathbf{y}, kh)\mathrm{d}\mathbf{x}$ as a functional over the space of smooth bounded functions on $\mathbb{R}^d$. It can be further derived [9] that the time derivative of the KL-Divergence along p_t is

$$\frac{d}{dt}F(p_t) = -\mathbb{E}\left[\left\langle \nabla \ln \left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right), \nabla \ln p_t(\mathbf{X}_t) + \nabla U(\mathbf{X}_{kh})\right\rangle\right],$$

where the expectation is taken with respect to the joint distribution of $\mathbf{X}_t$ and $\mathbf{X}_{kh}$. Hence

$$
\begin{aligned}
\frac{d}{dt}F(p_t) =& -\mathbb{E}\left[\left\langle \nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right), \nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right) + (\nabla U(\mathbf{X}_{kh}) - \nabla U(\mathbf{X}_t))\right\rangle\right] \\
=& -\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right)\right\|^2\right] + \mathbb{E}\left[\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right), \nabla U(\mathbf{X}_t) - \nabla U(\mathbf{X}_{kh})\right].
\end{aligned}
$$

For the second term, we use Young's inequality:

$$
\begin{aligned}
&\mathbb{E}\left[\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right), \nabla U(\mathbf{X}_t) - \nabla U(\mathbf{X}_{kh})\right] \\
\leq& \frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right)\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\nabla U(\mathbf{X}_t) - \nabla U(\mathbf{X}_{kh})\right\|^2\right] \\
\leq& \frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right)\right\|^2\right] + \frac{L^2}{2}\mathbb{E}\left[\left\|\mathbf{X}_t - \mathbf{X}_{kh}\right\|^2\right].
\end{aligned}
$$

Now we bound $\mathbb{E}\left[\left\|\mathbf{X}_t - \mathbf{X}_{kh}\right\|^2\right]$ using Lipschitz smoothness of U (define $\tau = t - kh \in (0, h]$):

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}_t - \mathbf{X}_{kh}\right\|^2\right] \\
\leq& \mathbb{E}\left[\left\|-\nabla U(\mathbf{X}_{kh})\tau + \sqrt{2}(B_{(k+1)h} - B_{hk})\right\|^2\right] \\
\leq& \mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\nabla U(\mathbf{x})\right\|^2\right]\tau^2 + 2d\tau \\
\leq& \mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\mathbf{x}\right\|^2\right]L^2\tau^2 + 2d\tau.
\end{aligned}
$$

Therefore, plugging in the bounds and using the log-Sobolev inequality proved in Proposition 2, we get for $kh < t \leq (k+1)h$:

$$
\begin{aligned}
&\frac{d}{dt}F(p_t) \\
\leq& -\frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right)\right\|^2\right] + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\mathbf{x}\right\|^2\right] + dL^2\tau \\
=& -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_t}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{x})}{p^*(\mathbf{x})}\right)\right\|^2\right] + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\mathbf{x}\right\|^2\right] + dL^2\tau \\
\leq& -\rho_U F(p_t) + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\mathbf{x}\right\|^2\right] + dL^2\tau. \qquad (20)
\end{aligned}
$$

From Lemma 5, we know that $\mathbb{E}_{\mathbf{x}\sim p_0}\left[\left\|\mathbf{x}\right\|_2^2\right] = \frac{d}{L} \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$. Combined with Lemma 4, we obtain that when $h \leq \frac{1}{4}\frac{\rho_U}{L^2}$, $\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\left\|\mathbf{x}\right\|_2^2\right] \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$ for any $k \in \mathbb{N}^+$. Therefore, for $h \leq \frac{1}{4}\frac{\rho_U}{L^2}$,

$$
\frac{d}{dt}F(p_t) \leq -\rho_U\left(F(p_t) - 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} - 256h^2\rho_U\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 - h\frac{L^2}{\rho_U}d\right).
$$

Using Gronwall's inequality,

$$\begin{aligned}
&F(p_{(k+1)h}) - 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} - 256h^2\rho_U\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 - h\frac{L^2}{\rho_U}d\\
\le& e^{-\rho_U h}\left(F(p_{kh}) - 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} - 256h^2\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 - h\frac{L^2}{\rho_U}d\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&F(p_{kh}) - 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} - 256h^2\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 - h\frac{L^2}{\rho_U}d\\
\le& e^{-\rho_U hk}\left(F(p_0) - 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} - 256h^2\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 - h\frac{L^2}{\rho_U}d\right)\\
&+ 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} + 256h^2\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 + h\frac{L^2}{\rho_U}d\\
\le& e^{-\rho_U hk}F(p_0) + 8h^2\frac{L^4}{\rho_U^2}d\ln\frac{2L}{m} + 256h^2\frac{L^4}{\rho_U^2}\frac{L^2}{m^2}LR^2 + h\frac{L^2}{\rho_U}d.
\end{aligned}$$

To make $F(p_{kh}) < \epsilon^2$, we take:

$$h = \frac{\rho_U}{4L^2}\min\left\{\frac{\epsilon^2}{d}, \sqrt{\frac{\epsilon^2}{2d\ln\frac{2L}{m} + 64\frac{L^2}{m^2}LR^2}}\right\} = \mathcal{O}\left(e^{-16LR^2}\frac{m}{L^2}\cdot\min\left\{\frac{\epsilon^2}{d}, \frac{m}{L}\frac{\epsilon}{\sqrt{LR^2}}\right\}\right). \tag{21}$$

Therefore, combining (21) with Lemma 5, we know that whenever

$$k \ge \mathcal{O}\left(e^{32LR^2}\frac{L^2}{m^2}\ln\left(\frac{F(p_0)}{\epsilon^2}\right)\cdot\max\left\{\frac{d}{\epsilon^2}, \frac{L}{m}\frac{\sqrt{LR^2}}{\epsilon}\right\}\right) = \mathcal{O}\left(e^{32LR^2}\frac{L^2}{m^2}\ln\left(\frac{d}{\epsilon^2}\right)\cdot\max\left\{\frac{d}{\epsilon^2}, \frac{L}{m}\frac{\sqrt{LR^2}}{\epsilon}\right\}\right), \tag{22}$$

$F(p_{kh}) < \frac{1}{2}\epsilon^2$. Using Pinsker inequality, we obtain

$$\|p_{kh} - p^*\|_{\mathrm{TV}} \le \sqrt{2F(p_{kh})} \le \epsilon.$$

Focusing on the dimension dependency, we obtain that the computation complexity scales as

$$k = \mathcal{O}\left(e^{32LR^2}\frac{L^2}{m^2}\frac{d}{\epsilon^2}\ln\left(\frac{F(p_0)}{\epsilon^2}\right)\right).$$

■

Lemma 4. *For p_t following (19), if $\mathbb{E}_{\mathbf{x}\sim p_0}\left[\|\mathbf{x}\|_2^2\right] \le \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$, and $h \le \frac{1}{4}\frac{\rho_U}{L^2}$, then for all $k\in\mathbb{N}^+$,*

$$\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|_2^2\right] \le \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2.$$

Lemma 5. *For*

$$p_0(\mathbf{x}) = \left(\frac{L}{2\pi}\right)^{d/2}\exp\left(-\frac{L}{2}||\mathbf{x}||^2\right)$$

and p^* following Assumptions 1–3,

$$F(p_0) = KL(p_0 \,\|\, p^*) = \int p_0(\mathbf{x})\ln\left(\frac{p_0(\mathbf{x})}{p^*(\mathbf{x})}\right)\mathrm{d}\mathbf{x} \le \frac{d}{2}\ln\frac{2L}{m} + 32\frac{L^2}{m^2}LR^2; \tag{23}$$

$$\mathbb{E}_{\mathbf{x}\sim p_0}\left[\|\mathbf{x}\|_2^2\right] = \frac{d}{L}; \tag{24}$$

and

$$\mathbb{E}_{\mathbf{x}\sim p^*}\left[\|\mathbf{x}\|_2^2\right] \leq \frac{4d}{\rho_U}\ln\frac{2L}{m} + \frac{128}{\rho_U}\frac{L^2}{m^2}LR^2. \tag{25}$$

B.2.1 Supporting Proofs for (9) of Theorem 3: Bounded Variance and $F(p_0)$

Proof of Lemma 4 Consider proof by induction. First assume that for some $k \geq 0$, for all $t = 0, h, \cdots, kh$, $\mathbb{E}_{\mathbf{x}\sim p_t}\left[\|\mathbf{x}\|_2^2\right] \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$. Then consider bounding $\mathbb{E}_{\mathbf{x}\sim p_t}\left[\|\mathbf{x}\|_2^2\right]$ for $kh < t \leq (k+1)h$, where p_t follows (19):

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_{kh})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \tag{26}$$

To bound $\mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\|\mathbf{x}_t\|_2^2\right]$, we choose an auxiliary random variable $\mathbf{x}^*$ following the law of p^* and couples optimally with $x_t \sim p_t$: $(\mathbf{x}_t, \mathbf{x}^*) \sim \gamma \in \Gamma_{opt}(p_t, p^*)$. Then using Young's inequality, and the bound for $\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right]$

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\|\mathbf{x}_t\|_2^2\right] =& \mathbb{E}_{(\mathbf{x}_t,\mathbf{x}^*)\sim\gamma}\left[\|\mathbf{x}^* + (\mathbf{x}_t - \mathbf{x}^*)\|_2^2\right] \\
\leq& 2\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right] + 2\mathbb{E}_{(\mathbf{x}_t,\mathbf{x}^*)\sim\gamma}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] \\
=& \frac{8d}{\rho_U}\ln\frac{2L}{m} + \frac{256}{\rho_U}\frac{L^2}{m^2}LR^2 + 2W_2^2(p_t, p^*).
\end{aligned} \tag{27}$$

Using the generalized Talagrand inequality [36] for Lipschitz smooth p^* with log-Sobolev constant ρ_U,

$$W_2^2(p_t, p^*) \leq \frac{2}{\rho_U}\mathrm{KL}(p_t \parallel p^*). \tag{28}$$

On the other hand, we know from (20) that for $F(p_t) = \mathrm{KL}(p_t \parallel p^*)$ (denote $\tau = t - kh$),

$$\frac{d}{dt}F(p_t) \leq -\rho_U F(p_t) + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|^2\right] + dL^2\tau.$$

Plugging in the step size $\tau \leq h \leq \frac{1}{4}\frac{\rho_U}{L^2}$ and the inductive assumption that $\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|^2\right] \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$, we obtain:

$$\frac{d}{dt}F(p_t) \leq -\rho_U F(p_t) + \frac{\rho_U}{4}d\ln\frac{2L}{m} + 8\rho_U\frac{L^2}{m^2}LR^2 + \frac{\rho_U}{4}d.$$

Without loss of generality, assume that $L \geq 2m$. Then

$$\frac{d}{dt}F(p_t) \leq -\rho_U\left(F(p_t) - \frac{d}{2}\ln\frac{2L}{m} - 8\frac{L^2}{m^2}LR^2\right).$$

Using Gronwall's inequality, we obtain:

$$
\begin{aligned}
F(p_{(k+1)h}) - \frac{d}{2}\ln\frac{2L}{m} - 8\frac{L^2}{m^2}LR^2 \leq & e^{-\rho_U h}\left(F(p_{kh}) - \frac{d}{2}\ln\frac{2L}{m} - 8\frac{L^2}{m^2}LR^2\right) \\
\leq & e^{-\rho_U h(k+1)}\left(F(p_0) - \frac{d}{2}\ln\frac{2L}{m} - 8\frac{L^2}{m^2}LR^2\right) \\
\leq & e^{-\rho_U h(k+1)}F(p_0) \\
\leq & F(p_0).
\end{aligned}
$$

Therefore, combining with (23) in Lemma 5,

$$
\begin{aligned}
F(p_{(k+1)h}) \leq & F(p_0) + \frac{d}{2}\ln\frac{2L}{m} + 8\frac{L^2}{m^2}LR^2 \\
\leq & d\ln\frac{2L}{m} + 40\frac{L^2}{m^2}LR^2. \qquad (29)
\end{aligned}
$$

Plugging (29) into (27) and (28), we finish the inductive proof:

$$
\mathbb{E}_{\mathbf{x}\sim p_{(k+1)h}}\left[\|\mathbf{x}\|_2^2\right] \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2.
$$

■

Proof of (23) **of Lemma 5** We want to bound $F(p_0) = \int p_0(\mathbf{x})\ln\left(\frac{p_0(\mathbf{x})}{p^*(\mathbf{x})}\right)\mathrm{d}\mathbf{x}$, where $p^*(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ and $p_0 = \left(\frac{L}{2\pi}\right)^{d/2}\exp\left(-\frac{L}{2}||\mathbf{x}||^2\right)$. First define $\bar{U}(\mathbf{x}) = U(\mathbf{x}) - U(0)$. Then

$$
p^*(\mathbf{x}) = \exp\left(-\bar{U}(\mathbf{x})\right) \Big/ \int \exp\left(-\bar{U}(\mathbf{x})\right)\mathrm{d}\mathbf{x}.
$$

By Assumptions 1 and 3, $\bar{U}(\mathbf{x}) \leq \frac{L}{2}\|\mathbf{x}\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d$. We also prove in the following that $\bar{U}(\mathbf{x}) \geq \frac{m}{4}\|\mathbf{x}\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d \setminus \mathbb{B}\left(0, \frac{8L}{m}R\right)$; and $\bar{U}(\mathbf{x}) \geq -\frac{L}{2}\|\mathbf{x}\|^2$, $\forall \mathbf{x} \in \mathbb{B}\left(0, \frac{8L}{m}R\right)$.

The latter case follows directly from Assumptions 1 and 3. For the former case, $\|\mathbf{x}\| \geq \frac{8L}{m}R$. Then define $\mathbf{y} = \frac{R}{\|\mathbf{x}\|}\mathbf{x}$. Since $\|\mathbf{y}\| = R$,

$$
< \nabla U(\mathbf{y}), \mathbf{y} > \geq -LR^2.
$$

Because any convex combination of $\mathbf{x}$ and $\mathbf{y}$ belongs to the set $\mathbb{R}^d \setminus \mathbb{B}(0, R)$, where U is m-strongly convex,

$$
\begin{aligned}
U(\mathbf{x}) - U(\mathbf{y}) \geq & < \nabla U(\mathbf{y}), \mathbf{x} - \mathbf{y} > + \frac{m}{2}\|\mathbf{x} - \mathbf{y}\|^2 \\
= & \left(\frac{\|\mathbf{x}\|}{R} - 1\right) < \nabla U(\mathbf{y}), \mathbf{y} > + \frac{m}{2}\left(\frac{\|\mathbf{x}\|}{R} - 1\right)^2 \\
\geq & -\left(\frac{\|\mathbf{x}\|}{R} - 1\right)LR^2 + \frac{m}{2}\left(\frac{\|\mathbf{x}\|}{R} - 1\right)^2 \\
\geq & \frac{m}{4}\|\mathbf{x}\|^2 + LR^2,
\end{aligned}
$$

since $\|\mathbf{x}\| \geq \frac{8L}{m}R$. Again, using Assumptions 1 and 3, $U(\mathbf{y}) \geq -\frac{L}{2}R^2$, which leads to the result that $U(\mathbf{x}) \geq \frac{m}{4}\|\mathbf{x}\|^2$.

Therefore, $U(\mathbf{x}) \geq \frac{m}{4}\|\mathbf{x}\|^2 - 32\frac{L^2}{m^2}LR^2$ and

$$\begin{aligned}
-\ln p^*(\mathbf{x}) =& \bar{U}(\mathbf{x}) + \ln \int \exp\left(-\bar{U}(\mathbf{x})\right) \mathrm{d}\mathbf{x} \\
\leq& \frac{L}{2}\|\mathbf{x}\|^2 + \ln \int \exp\left(-\frac{m}{4}\|\mathbf{x}\|^2 + 32\frac{L^2}{m^2}LR^2\right) \mathrm{d}\mathbf{x} \\
=& \frac{L}{2}\|\mathbf{x}\|^2 + \frac{d}{2}\ln\frac{4\pi}{m} + 32\frac{L^2}{m^2}LR^2.
\end{aligned}$$

Hence

$$-\int p_0(\mathbf{x}) \ln p^*(\mathbf{x}) \mathrm{d}\mathbf{x} \leq 32\frac{L^2}{m^2}LR^2 + \frac{d}{2}\ln\frac{4\pi}{m} + \frac{d}{2}.$$

We can also calculate that

$$\int p_0(\mathbf{x}) \ln p_0(\mathbf{x}) \mathrm{d}\mathbf{x} = -\frac{d}{2}\ln\frac{2\pi}{L} - \frac{d}{2}.$$

Therefore,

$$\begin{aligned}
F(p_0) =& \int p_0(\mathbf{x}) \ln p_0(\mathbf{x}) \mathrm{d}\mathbf{x} - \int p_0(\mathbf{x}) \ln p^*(\mathbf{x}) \mathrm{d}\mathbf{x} \\
\leq& 32\frac{L^2}{m^2}LR^2 + \frac{d}{2}\ln\frac{2L}{m}.
\end{aligned}$$

■

Proof of (24) **of Lemma 5** It is straightforward to calculate that $\mathbb{E}_{p_0}\left[\|\mathbf{x}\|_2^2\right] = \mathrm{trace}\left(\frac{1}{L}\mathbb{I}\right) = \frac{d}{L}$.

It is worth noting that the choice of the initial condition p_0 can be flexible. For example, if we choose $\mathbf{x}_0 \sim \mathcal{N}\left(0, \frac{1}{m}\mathbb{I}\right)$, then $F(p_0) \leq 32\frac{L^2}{m^2}LR^2 + \frac{d}{2}\cdot\frac{L}{m}$ and $\mathbb{E}_{p_0}\left[\|\mathbf{x}\|_2^2\right] = \frac{d}{m} \leq 48R^2 + \frac{4d}{m}$ (resulting in merely an extra $\log\frac{L}{m}$ term in the computation complexity). ■

Proof of (25) **of Lemma 5** To bound $\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right]$, we choose an auxiliary random variable $\mathbf{x}_0$ following the law of p_0 and couples optimally with $x^* \sim p^*$: $(\mathbf{x}^*, \mathbf{x}_0) \sim \gamma \in \Gamma_{opt}(p^*, p_0)$. Then using Young's inequality,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right] =& \mathbb{E}_{(\mathbf{x}^*,\mathbf{x}_0)\sim\gamma}\left[\|\mathbf{x}_0 + (\mathbf{x}^* - \mathbf{x}_0)\|_2^2\right] \\
\leq& 2\mathbb{E}_{\mathbf{x}_0\sim p_0}\left[\|\mathbf{x}_0\|_2^2\right] + 2\mathbb{E}_{(\mathbf{x}^*,\mathbf{x}_0)\sim\gamma}\left[\|\mathbf{x}^* - \mathbf{x}_0\|_2^2\right] \\
=& \frac{2d}{L} + 2W_2^2\left(p^*, p_0\right).
\end{aligned}$$

Using the generalized Talagrand inequality [36] for Lipschitz smooth p^* with log-Sobolev constant ρ_U,

$$W_2^2\left(p^*, p_0\right) \leq \frac{2}{\rho_U}\mathrm{KL}(p_0 \parallel p^*).$$

On the other hand, we know from (23) that

$$\mathrm{KL}(p_0 \parallel p^*) \leq \frac{d}{2}\ln\frac{2L}{m} + 32\frac{L^2}{m^2}LR^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right] \leq & \frac{2d}{L} + \frac{2d}{\rho_U}\ln\frac{2L}{m} + \frac{128}{\rho_U}\frac{L^2}{m^2}LR^2 \\ \leq & \frac{4d}{\rho_U}\ln\frac{2L}{m} + \frac{128}{\rho_U}\frac{L^2}{m^2}LR^2.\end{aligned}$$

■

B.3 Proof of MALA Convergence Rate ((10) of Theorem 3)

Proof of (10) Our proof of Theorem 10 is based on the following two lemmas. The first one characterizes the convergence of MALA under a warm starting distribution. The second one shows that the initial distribution $\mathcal{N}\left(0, \frac{1}{2L}\mathbb{I}_d\right)$ is $\mathcal{O}(e^d)$-warm. Let us first define the warm start.

Definition 6 (Warm start)**.** *Given a scalar $\theta > 0$, an initial distribution with density p^0 is said to be β-warm with respect to the stationary distribution with density p^* if*

$$\forall \mathbf{x} \in \mathbb{R}^d, \frac{p^0(\mathbf{x})}{p^*(\mathbf{x})} \leq \beta.$$

Lemma 7. *Assume $p^*(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ where U satisfies the local nonconvexity Assumptions 1–3. Then the MALA with a β-warm distribution with density p^0 and error tolerance $\epsilon \in (0,1)$, satisfies*

$$\tau(\epsilon, p^0) \leq \mathcal{O}\left(e^{16LR^2}\cdot\ln\left(\frac{2\beta}{\epsilon}\right)\cdot\max\left\{r\left(\frac{2\beta}{\epsilon}\right)\kappa^{3/2}d^{1/2}, \kappa d\right\}\right). \tag{30}$$

Lemma 8. *The initial distribution $\mathcal{N}\left(0, \frac{1}{L}\mathbb{I}_d\right)$ is $e^{16LR^2}(2\kappa)^{d/2}$-warm with respect to the target distribution p^*.*

Theorem 10 directly follows by combining Lemma 7 and Lemma 8. ■

Proof of Lemma 7 At a high level, the proof closely follows the proof of Theorem 1 in [17]. We replace their Lemma 1 with Lemma 12 to establish that for an appropriate choice of stepsize, the MALA updates have large overlap inside the high probability ball. Lemma 11 allows us to obtain a lower bound on the conductance. Finally applying the Lovasz lemma, we obtain convergence guarantees.

In order to start the proof, we first introduce conductance related notions for a general Markov chain. Consider an ergodic Markov chain defined by a transition operator $\mathcal{T}$, and let Π denote its stationary distribution. We define the ergodic flow from A to its complement A^c

$$\phi(A) = \int_A \mathcal{T}_{\mathbf{u}}(A^c)p^*(\mathbf{u})\mathrm{d}\mathbf{u}.$$

For each scalar $s \in (0, 1/2)$, we define the s-conductance

$$\Phi_s = \inf_{\Pi(A)\in(s,1-s)} \frac{\phi(A)}{\min\left\{\Pi(A) - s, \Pi(A^c) - s\right\}}.$$

The notation $\mathcal{T}_{\mathbf{u}}$ is the shorthand for the distribution $\mathcal{T}(\delta_{\mathbf{u}})$ obtained by applying the transition operator to a dirac distribution concentrated on $\mathbf{u}$.

For a Markov chain with θ-warm start initial distribution Π_0, having s-conductance Φ_s, Lovász and Simonovits [28] proved its convergence

$$\left\|\mathcal{T}^k(\Pi_0) - \Pi\right\|_{\mathrm{TV}} \leq \theta s + \theta\left(1 - \frac{\Phi_s^2}{2}\right)^k \leq \theta s + \theta e^{-k\Phi_s^2/2} \text{ for any } s \in (0, \frac{1}{2}). \tag{31}$$

We will apply this result for s small by cutting off the probability mass outside a Euclidean ball. We define radius

$$r(s) = 2 + \sqrt{2}e^{8LR^2}\ln^{0.5}(d/s) + 7R/\sqrt{d/m}, \tag{32}$$

and the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(0, r(s)\sqrt{\frac{d}{m}}\right). \tag{33}$$

We define the appropriate stepsize.

$$\tilde{w}(s,\gamma) = \min\left\{\frac{\sqrt{\gamma}}{8\sqrt{2}r(s)}\frac{\sqrt{m}}{L\sqrt{dL}}, \quad \frac{\gamma}{96\alpha_\gamma}\frac{1}{Ld}, \quad \frac{\gamma^{2/3}}{26(\alpha_\gamma r^2(s))^{1/3}}\frac{1}{L}\left(\frac{m}{Ld^2}\right)^{1/3}\right\}, \tag{34}$$

$$\text{where} \quad \alpha_\gamma = 1 + 2\sqrt{\log(16/\gamma)} + 2\log(16/\gamma). \tag{35}$$

Applying Lemma 12 with $h = \tilde{w}(s,\gamma)$, for $\mathbf{x}, \mathbf{y} \in \mathcal{R}_s$ and $\|\mathbf{x} - \mathbf{y}\|_2 \leq \Delta = \gamma\sqrt{h}/4$, we obtain

$$\begin{aligned}\|\mathcal{T}_{\mathbf{x}} - \mathcal{T}_{\mathbf{y}}\|_{\mathrm{TV}} &\leq \|\mathcal{T}_{\mathbf{x}} - \mathcal{P}_{\mathbf{x}}\|_{\mathrm{TV}} + \|\mathcal{P}_{\mathbf{x}} - \mathcal{P}_{\mathbf{x}}\|_{\mathrm{TV}} + \|\mathcal{P}_{\mathbf{x}} - \mathcal{T}_{\mathbf{y}}\|_{\mathrm{TV}} \\ &\leq \frac{\sqrt{2}\gamma}{4} + \frac{\gamma}{8} + \frac{\sqrt{2}\gamma}{4} \\ &\leq \gamma. \end{aligned} \tag{36}$$

Applying Lemma 11 with $\mathcal{K} = \mathcal{R}_s$ in combination with Lemma 9, Lemma 10 and Lemma 12, we obtain that for stepsize $h \in (0, \tilde{w}(s,\gamma)]$, the s-conductance is lower bounded.

$$\Phi_s \geq \frac{(1-\gamma)\cdot(1-s)^2\cdot\gamma\sqrt{h}\cdot\sqrt{\rho_U/10}}{256}.$$

Now we can conclude by making appropriate choice of s and γ. Letting $s = \dfrac{\epsilon}{2\beta}$ and $\gamma = \dfrac{1}{2}$, we obtain

$$\Phi_s \geq \mathcal{O}(\sqrt{\rho_U h}).$$

Plugging this conductance expression into the result of Lovász and Simonovits (31), with Π_0 the distribution with density p^0 and Π the stationary distribution with density p^*, we obtain that

$$\left\|\mathcal{T}^k(\Pi_0) - \Pi\right\|_{\mathrm{TV}} \leq \beta\frac{\epsilon}{2\beta} + \beta e^{-k\Phi_s^2/2} \leq \epsilon, \text{ for } k \geq \mathcal{O}\left(\frac{1}{\rho_U h}\cdot\ln\left(\frac{2\beta}{\epsilon}\right)\right),$$

where

$$\rho_U \geq \frac{m}{2}e^{-16LR^2},$$

and

$$h = \mathcal{O}\left(\min\left\{\frac{1}{L\cdot r(\frac{2\beta}{\epsilon})\kappa^{1/2}d^{1/2}}, \frac{1}{Ld}\right\}\right).$$

■

Lemma 9. *For any $s \in (0, \frac{1}{2})$, we have $\Pi(\mathcal{R}_s) \geq 1 - s$.*

Lemma 10. *If the density p^* satisfies the log-Sobolev inequality with constant ρ_U, then it also satisfies the following isoperimetric inequality with constant $c = (\rho_U/10)^{1/2}$: For any A and B open disjoint subsets of $\mathbb{R}^d$, $C = \mathbb{R}^d \setminus (A \cup B)$, Π being the probability measure for p^*, we have*

$$\Pi(A) \geq c \cdot d(A, B)\Pi(A)\Pi(B), \tag{37}$$

where $d(A, B) = \min_{\mathbf{x} \in A, y \in B} \|\mathbf{x} - \mathbf{y}\|_2$, is the set distance with Euclidean metric on $\mathbb{R}^d$.

Lemma 11. *Let $\mathcal{K}$ be a convex set such that $\|\mathcal{T}_{\mathbf{x}} - \mathcal{T}_{\mathbf{y}}\|_{TV} \leq \gamma$ whenever $\mathbf{x}, \mathbf{y} \in \mathcal{K}$ and $\|\mathbf{x} - \mathbf{y}\|_2 \leq \Delta$. Π satisfies the partition type isoperimetric inequality (37) with constant ρ. Then for any measurable partition A_1 and A_2 of $\mathbb{R}^d$, we have*

$$\int_{A_1} \mathcal{T}_{\mathbf{u}}(A_2) p^*(\mathbf{u}) \mathrm{d}\mathbf{u} \geq \frac{\rho}{8} \min\left\{1, \frac{\Delta \cdot (1-\gamma) \cdot \Pi^2(\mathcal{K})}{8}\right\} \min\left\{\Pi(A_1 \cap \mathcal{K}), \Pi(A_2 \cap \mathcal{K})\right\}. \tag{38}$$

Lemma 12. *For any step size $h \in \left(0, \frac{1}{L}\right]$, the MALA proposal distribution satisfies the bound*

$$\sup_{\substack{\mathbf{x}, \mathbf{y} \in \mathbb{R}^d \\ \mathbf{x} \neq \mathbf{y}}} \frac{\left\|\mathcal{P}_{\mathbf{x}}^{MALA(h)} - \mathcal{P}_{\mathbf{y}}^{MALA(h)}\right\|_{TV}}{\|\mathbf{x} - \mathbf{y}\|_2} \leq \sqrt{\frac{2}{h}}. \tag{39a}$$

Moreover, given scalars $s \in (0, 1/2)$ and $\gamma \in (0, 1)$, then the MALA proposal distribution for any stepsize $h \in \left(0, \tilde{w}(s, \gamma)\right]$ satisfies the bound

$$\sup_{\mathbf{x} \in \mathcal{R}_s} \left\|\mathcal{P}_{\mathbf{x}}^{MALA(h)} - \mathcal{T}_{\mathbf{x}}^{MALA(h)}\right\|_{TV} \leq \frac{\gamma}{8}, \tag{39b}$$

where the truncated ball $\mathcal{R}_s$ was defined in (33).

B.3.1 Supporting Proofs for (10) of Theorem 3

Proof of Lemma 8 The starting distribution $\mathcal{N}\left(0, \frac{1}{L}\mathbb{I}_d\right)$ has density

$$p^0(\mathbf{x}) = \left(\frac{L}{2\pi}\right)^{d/2} e^{-\frac{L\|\mathbf{x}\|^2}{2}}.$$

Taking the ratio with respect to the stationary distribution, we have

$$\begin{aligned}
\frac{p^0(\mathbf{x})}{p^*(\mathbf{x})} &= \frac{p^0(\mathbf{x})}{\frac{1}{\int e^{-U(\mathbf{x})}\mathrm{d}\mathbf{x}} e^{-U(\mathbf{x})}} \\
&\leq e^{16LR^2} (2\kappa)^{d/2} \cdot \exp\left(-L\|\mathbf{x}\|^2/2 + U(\mathbf{x})\right) \\
&\leq e^{16LR^2} (2\kappa)^{d/2}.
\end{aligned}$$

The first inequality is because, according to Lemma 2, we have

$$\int e^{-U(\mathbf{x})}\mathrm{d}\mathbf{x} \leq e^{16LR^2} \cdot \int e^{-\frac{m\|\mathbf{x}\|^2}{4}} \mathrm{d}\mathbf{x} = e^{16LR^2} \left(\frac{m}{4\pi}\right)^{d/2}.$$

■

Proof of Lemma 9 This lemma relies on the concentration of the stationary distribution p^* around 0. The concentration follows from the log-Sobolev constant shown in Proposition 2. The following lemma is a classical way to obtain concentration from the log-Sobolev inequality is based on Herbst argument (e.g. see Section 2.3 in [26]).

Lemma 13. *If* Π *satisfies a log-Sobolev inequality with constant* ρ *then every* 1*-Lipschitz function* f *is integrable with respect to* Π *and satisfies the concentration inequality*

$$\mathbb{P}_{\mathbf{x}\sim\Pi}\left[f(\mathbf{x}) > \mathbb{E}_{\Pi}\left[f\right] + t\right] \leq e^{-\rho t^2/2}.$$

Applying this lemma with f being the projection to each coordinate and using union bound, we obtain that

$$\mathbb{P}_{\mathbf{x}\sim\Pi}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2 > \frac{2td}{\rho_U}\right] \leq de^{-t}.$$

We define $\mathcal{B}_1 = \mathbb{B}\left(\mathbb{E}[\mathbf{x}], \sqrt{2\log(\frac{d}{s})\frac{d}{\rho_U}}\right)$. Taking $t = \log(\frac{d}{s})$, we obtain that

$$\Pi(\mathcal{B}_1) \geq 1 - s.$$

Using the results proved in Lemma 4, we can also turn this concentration around the mean to the concentration around 0. According to Lemma 4, we have

$$\mathbb{E}_{\mathbf{x}\sim\Pi}\|\mathbf{x}\|_2^2 \leq 48R^2 + \frac{4d}{m}.$$

Using Jensen's inequality, we obtain

$$\|\mathbb{E}_{\mathbf{x}\sim\Pi}[\mathbf{x}]\|_2 \leq \mathbb{E}_{\mathbf{x}\sim\Pi}\|\mathbf{x}\|_2 \leq \sqrt{\mathbb{E}_{\mathbf{x}\sim\Pi}\|\mathbf{x}\|_2^2} \leq \sqrt{48R^2 + \frac{4d}{m}}.$$

We define $\mathcal{B}_2 = \mathbb{B}\left(0, \sqrt{48R^2 + \frac{4d}{m}} + \sqrt{2\log(\frac{d}{s})\frac{d}{\rho_U}}\right)$. We deduce that

$$\mathcal{B}_1 \subset \mathcal{B}_2 \subset \mathcal{R}_s.$$

As a result, we obtain $\Pi(\mathcal{R}_s) \geq \Pi(\mathcal{B}_1) \geq 1 - s$ as claimed. ■

Proof of Lemma 10 Lemma 10 shows that log-Sobolev inequality implies isoperimetric inequality with constants of the same order. It is pretty standard. Since we can't find a complete proof in the literature, we provide it for completeness. p^* satisfies the following log-Sobolev inequality, for any smooth $g : \mathbb{R}^d \to \mathbb{R}$.

$$2\rho_U\left[\int_{\mathbb{R}^d} g\ln g \mathrm{d}\Pi - \int_{\mathbb{R}^d} g\mathrm{d}\Pi \cdot \ln\left(\int_{\mathbb{R}^d} g\mathrm{d}\Pi\right)\right] \leq \int_{\mathbb{R}^d}\frac{\|\nabla g\|_2^2}{g}\mathrm{d}\Pi. \tag{40}$$

where

$$\mathrm{d}\Pi(\mathbf{x}) = p^*(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Replacing g with g^2 in (40), for $g : \mathbb{R}^d \mapsto \mathbb{R}$, we obtain the equivalent form

$$2\rho_U \mathrm{Ent}\left(g^2\right) \leq \int_{\mathbb{R}^d}\|\nabla g\|_2^2 \mathrm{d}\Pi,$$

where

$$\mathrm{Ent}_{p^*}(g^2) = \left[\int_{\mathbb{R}^d} g^2\ln g^2 \mathrm{d}\Pi - \int_{\mathbb{R}^d} g^2\mathrm{d}\Pi \cdot \ln\left(\int_{\mathbb{R}^d} g^2\mathrm{d}\Pi\right)\right].$$

It is well known that the log-Sobolev inequality implies the following Poincaré inequality with the same constant (e.g. [5]). For any smooth $g : \mathbb{R}^d \to \mathbb{R}$, we have

$$\rho_U \mathrm{Var}_{p^*}(g) \leq \int_{\mathbb{R}^d} \|\nabla g\|_2^2 \, \mathrm{d}\Pi,$$

where

$$\mathrm{Var}_{p^*}(g) = \int_{\mathbb{R}^d} g^2 \mathrm{d}\Pi - \left(\int_{\mathbb{R}^d} g \mathrm{d}\Pi \right)^2.$$

This implication is based on the fact that the gradient operator is invariant to translation (i.e. for $c \in \mathbb{R}$, $\nabla(f + c) = \nabla f$) and

$$\mathrm{Ent}\left((f + c)^2 \right) \to 2\mathrm{Var}(f), \text{ as } c \to \infty.$$

Next, Buser [8] provides a lower bound on the isoperimetric constant based on the Poincaré constant. We denote h the isoperimetric constant defined as

$$h = \sup_{A \subset \mathbb{R}^d, \text{ open}} \frac{\Pi^+(\partial A)}{\min \Pi(A), 1 - \Pi(A)}, \tag{41}$$

where $\Pi^+(\partial A) = \lim_{\delta \to 0} \frac{\Pi(A+\delta) - \Pi(A)}{\delta}$. Then Buser [8] shows that

$$h \geq \left(\frac{\rho_U}{10} \right)^{1/2}.$$

Finally, it is easy to show that the infinitesimal version of the isoperimetric inequality in (41) is equivalent to the partition version (see e.g. [6] Proposition 11.1 and [4]). Let A and B be open disjoint subsets of $\mathbb{R}^d$, $C = \mathbb{R}^d \setminus (A \cup B)$, then

$$\Pi(C) \geq \left(\frac{\rho_U}{10} \right)^{1/2} d(A, B) \Pi(A) \Pi(B). \tag{42}$$

■

In the following, we provide useful lemmas for proving Lemma 7.

Proof of Lemma 11 The proof of this lemma follows directly from the proof of Lemma 2 in [17]. The main difference in the setting is that the target distribution is no longer log-concave, however, the proof follows because the log-concavity was never used in the proof of this lemma. It is sufficient to replace the isoperimetric inequality with ours in (42). ■

Proof of Lemma 12 We prove the two claims in this Lemma separately. In order to simplify notation, we drop the superscript from our notations of distributions $\mathcal{T}_{\mathbf{x}}^{\mathrm{MALA}(h)}$ and $\mathcal{P}_{\mathbf{x}}^{\mathrm{MALA}(h)}$. ■

Proof of Lemma 39a We first apply the Pinsker inequality [12] to bound the total variation distance via KL-divergence.

$$\|\mathcal{P}_{\mathbf{x}} - \mathcal{P}_{\mathbf{y}}\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(\mathcal{P}_{\mathbf{x}} \parallel \mathcal{P}_{\mathbf{y}})}.$$

Since our proposals before applying Metropolis filters follow multivariate normal distributions, we obtain closed form expressions for the KL divergence.

$$\begin{aligned}
\|\mathcal{P}_{\mathbf{x}} - \mathcal{P}_{\mathbf{y}}\|_{\mathrm{TV}} &\le \sqrt{2\mathrm{KL}(\mathcal{P}_{\mathbf{x}} \parallel \mathcal{P}_{\mathbf{y}})} \\
&= \frac{\|\Pi_{\mathbf{x}} - \Pi_{\mathbf{y}}\|_2}{\sqrt{2h}} \\
&= \frac{\|(\mathbf{x} - h\nabla U(\mathbf{x})) - (\mathbf{y} - h\nabla U(\mathbf{y}))\|_2}{\sqrt{2h}}.
\end{aligned}$$

Here we use the smoothness without using the convexity to bound the last term. We have

$$\begin{aligned}
\|(\mathbf{x} - h\nabla U(\mathbf{x})) - (\mathbf{y} - h\nabla U(\mathbf{y}))\|_2 &= \left\| \int_0^1 \left[\mathbb{I}_d - h\nabla^2 U(\mathbf{x} + t(\mathbf{x} - \mathbf{y}))\right] (\mathbf{x} - \mathbf{y})\mathrm{d}t \right\|_2 \\
&\le \int_0^1 \left\| \left[\mathbb{I}_d - h\nabla^2 U(\mathbf{x} + t(\mathbf{x} - y))\right] (\mathbf{x} - \mathbf{y}) \right\|_2 \mathrm{d}t \\
&\le \sup_{t\in[0,1]} \|\!|\mathbb{I}_d - h\nabla^2 U(\mathbf{x} + t(\mathbf{x} - \mathbf{y}))|\!\|_2 \, \|\mathbf{x} - \mathbf{y}\|_2 \\
&\le 2\,\|\mathbf{x} - \mathbf{y}\|_2\,.
\end{aligned}$$

The last inequality follows from the fact that $\nabla^2 U(\mathbf{z}) \preceq L\mathbb{I}_d$ for all $\mathbf{z} \in \mathbb{R}^d$. Note that we lose a 2 factor without using the convexity. ■

Proof of Lemma 39b We denote $p_{\mathbf{x}}$ the density corresponding to the proposal distribution $\mathcal{P}_{\mathbf{x}} = \mathcal{N}\left(\mathbf{x} - h\nabla U(\mathbf{x}), 2h\mathbb{I}_d\right)$. We have

$$\begin{aligned}
\|\mathcal{P}_{\mathbf{x}} - \mathcal{T}_{\mathbf{x}}\|_{\mathrm{TV}} &= \frac{1}{2}\left(\mathcal{T}_{\mathbf{x}}\left(\{\mathbf{x}\}\right) + \int_{\mathbb{R}^d} p_{\mathbf{x}}(\mathbf{z})\mathrm{d}\mathbf{z} - \int_{\mathbb{R}^d} \min\left\{1, \frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})}\right\} p_{\mathbf{x}}(\mathbf{z})\mathrm{d}\mathbf{z}\right) \\
&= \frac{1}{2}\left(2 - 2\int_{\mathbb{R}^d} \min\left\{1, \frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})}\right\} p_{\mathbf{x}}(\mathbf{z})\mathrm{d}\mathbf{z}\right) \\
&\le 1 - \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_{\mathbf{x}}}\left[\min\left\{1, \frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})}\right\}\right].
\end{aligned}$$

Applying Markov inequality, we know that

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_{\mathbf{x}}}\left[\min\left\{1, \frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})}\right\}\right] \ge \alpha\mathbb{P}\left[\frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})} \ge \alpha\right] \quad \text{for all } \alpha \in (0, 1].$$

It is sufficient to derive a high probability lower bound for the ratio $\dfrac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})}$. Plugging the fact that $p^*(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$ and $p_{\mathbf{x}}(\mathbf{z}) \propto \exp\left(-\|\mathbf{x} - h\nabla U(\mathbf{x}) - \mathbf{z}\|_2^2/(4h)\right)$, we have

$$\frac{p^*(\mathbf{z})\cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x})\cdot p_{\mathbf{x}}(\mathbf{z})} = \exp\left(\frac{4h\left(U(\mathbf{x}) - U(\mathbf{z})\right) + \|\mathbf{z} - \mathbf{x} + h\nabla U(\mathbf{x})\|_2^2 - \|\mathbf{x} - \mathbf{z} + h\nabla U(\mathbf{z})\|_2^2}{4h}\right).$$

We then lower bound the term in the numerator of the exponent, without using the convexity of U.

$$\begin{aligned}
&4h\left(U(\mathbf{x}) - U(\mathbf{z})\right) + \|\mathbf{z} - \mathbf{x} + h\nabla U(\mathbf{x})\|_2^2 - \|\mathbf{x} - \mathbf{z} + h\nabla f(\mathbf{z})\|_2^2 \\
&= 2h\underbrace{\left(U(\mathbf{x}) - U(\mathbf{z}) - (\mathbf{x} - \mathbf{z})^\top \nabla U(\mathbf{x})\right)}_{M_1} + 2h\underbrace{\left(U(\mathbf{x}) - U(\mathbf{z}) - (\mathbf{x} - \mathbf{z})^\top \nabla U(\mathbf{z})\right)}_{M_2} + h^2\underbrace{\left(\|\nabla U(\mathbf{x})\|_2^2 - \|U(\mathbf{z})\|_2^2\right)}_{M_3}.
\end{aligned}$$

Using the fact that U is smooth, we have

$$M_1 \geq -\frac{L}{2} \left\| \mathbf{x} - \mathbf{z} \right\|_2^2 \text{ and } M_2 \geq -\frac{L}{2} \left\| \mathbf{x} - \mathbf{z} \right\|_2^2.$$

Again using the smoothness, we have

$$M_3 = \left\| \nabla f(\mathbf{x}) \right\|_2^2 - \left\| \nabla f(\mathbf{z}) \right\|_2^2 = \langle \nabla f(\mathbf{x}) + \nabla f(\mathbf{z}), \nabla f(\mathbf{x}) - \nabla f(\mathbf{z}) \rangle \geq -\left(2 \left\| \nabla f(\mathbf{x}) \right\|_2 + L \left\| \mathbf{x} - \mathbf{z} \right\|_2\right) L \left\| \mathbf{x} - \mathbf{z} \right\|_2.$$

Combining the bounds M_1, M_2, M_3, we have established that

$$\frac{p^*(\mathbf{z}) \cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x}) \cdot p_{\mathbf{x}}(\mathbf{z})} \geq \exp \underbrace{\left(-\frac{1}{2} L \left\| \mathbf{x} - \mathbf{z} \right\|_2^2 - \frac{h}{4} \left(2L \left\| \mathbf{x} - \mathbf{z} \right\|_2 \left\| \nabla f(\mathbf{x}) \right\|_2 + L^2 \left\| \mathbf{x} - \mathbf{z} \right\|_2^2 \right) \right)}_{T}.$$

In addition, using the fact that $\mathbf{z}$ is a proposal, we have

$$\left\| \mathbf{x} - \mathbf{z} \right\|_2 = \left\| h \nabla U(\mathbf{x}) + \sqrt{2h} \xi \right\|_2 \leq h \left\| f(\mathbf{x}) \right\|_2 + \sqrt{2h} \left\| \xi \right\|_2.$$

Simplifying and using the fact that $Lh \leq 1$, we obtain

$$T \geq -2Lh^2 \left\| \nabla U(\mathbf{x}) \right\|_2^2 - 3Lh \left\| \xi \right\|_2^2 - Lh\sqrt{h} \left\| \nabla U(\mathbf{x}) \right\|_2 \left\| \xi \right\|_2.$$

Since $\mathbf{x} \in \mathcal{R}$, we can bound the gradient roughly

$$\left\| \nabla U(\mathbf{x}) \right\|_2 = \left\| \nabla U(\mathbf{x}) - \nabla U(\mathbf{x}^*) \right\|_2 \leq L \left\| \mathbf{x} - \mathbf{x}^* \right\|_2 \leq L \sqrt{\frac{d}{m}} r(s) =: \mathcal{D}_s.$$

$\left\| \xi \right\|_2^2$ is bounded via standard χ^2-variable tail bound. We have

$$\mathbb{P}\left[\left\| \xi \right\|_2^2 \leq d\alpha_\epsilon \right] \geq 1 - \frac{\epsilon}{16},$$

for $\alpha_\epsilon = 1 + 2\sqrt{\log(16/\epsilon)} + 2\log(16/\epsilon)$. The choice of $\tilde{w}$ guarantees that for $h \leq \tilde{w}$, we have

$$Lh^2 \mathcal{D}_s^2 \leq \frac{\epsilon}{128}, Lhd\alpha_\epsilon \leq \frac{\epsilon}{96}, \text{ and } Lh\sqrt{h}\mathcal{D}_s \sqrt{d\alpha_\epsilon} \leq \frac{\epsilon}{64}.$$

Combining all these bound, we obtain

$$\mathbb{P}\left[T \geq -\frac{\epsilon}{16} \right] \geq 1 - \frac{\epsilon}{16}.$$

Using the fact that $e^{-\epsilon/16} \geq 1 - \epsilon/16$, we have

$$\mathbb{E}_{z \sim \mathcal{P}_{\mathbf{x}}} \left[1, \frac{p^*(\mathbf{z}) \cdot p_{\mathbf{z}}(\mathbf{x})}{p^*(\mathbf{x}) \cdot p_{\mathbf{x}}(\mathbf{z})} \right] \geq 1 - \frac{\epsilon}{8}, \text{ for any } \epsilon \in (0,1), \text{ and } h \leq \tilde{w}.$$

■

Appendix C Proofs for Optimization

We denote $\tilde{\nabla} U(\mathbf{x}) = \{ \nabla^n U(\mathbf{x}) | n \in \mathcal{N} \}$ as shorthand for all n-th order derivative at point $\mathbf{x}$. We consider iterative algorithm class $\mathcal{A}_\infty$ operating on a function $U : \mathbb{R}^d \to \mathbb{R}$ whose iterates has following form:

$$\mathbf{x}_t = g_t(\zeta, \tilde{\nabla} U(\mathbf{x}_0), \ldots, \tilde{\nabla} U(\mathbf{x}_{t-1}))$$

where g_t is a mapping to $\mathbb{R}^d$. ζ is a random variable sampled from uniform distribuion over $[0,1]$ (indepedent of U), and it contains infinitely many random bits. We note standard optimization algorithms (either deterministic or randomized) which utilize gradient information or any p-th order information all fall in to this class of algorithms $\mathcal{A}_\infty$.

Theorem 4 (Lower bound for optimization). *For any $R > 0$, $L \geq 2m > 0$, probability $0 < p \leq 1$, and $\epsilon \leq \dfrac{LR^2}{32\pi^2 + 16\pi}$, there exists an objective function U satisfying the local nonconvexity Assumptions 1–3 with constants L, m, and $2R$, such that any algorithm in $\mathcal{A}_\infty$ requires at least $\Omega\left(p \cdot \left(LR^2/\epsilon\right)^{d/2}\right)$ iterations to guarantee $P\left(\min_{\tau \leq t} |U(\mathbf{x}_\tau) - U(\mathbf{x}^*)| < \epsilon\right) \geq p$.*

C.1 Proof of Theorem 4

We constructively prove Theorem 4 by defining such a $U(\mathbf{x})$ in what follows. We first make use of the following lemma about packing numbers.

Lemma 14 (Packing number). *For $R > r > 0$, denote $\eta = \left\lfloor \left(\dfrac{R-r}{2r}\right)^d \right\rfloor$. Then there exists set $\mathbb{X}_\eta = \{\mathbf{x}_1, \cdots \mathbf{x}_\eta\}$, s.t. $\bigcup_{i=1}^{\eta} \mathbb{B}(\mathbf{x}_i, r) \subset \mathbb{B}(0, R)$, and $\mathbb{B}(\mathbf{x}_i, r) \bigcap \mathbb{B}(\mathbf{x}_j, r) = \emptyset, \forall i \neq j$.*

As shown in Fig. 4, this Lemma 14 guarantees the existence of the set $\{\mathbf{x}_1, \cdots \mathbf{x}_\eta\}$ so that η balls of radius r centered at $\mathbf{x}_\eta$ are contained inside the larger ball of radius R without intersecting with each other.

We hereby construct $U(\mathbf{x})$ that gives the lower bound. If $\epsilon \geq \dfrac{LR^2}{18\pi^2 + 9\pi}$, then

$$T \geq 1 \geq p \cdot \left\lfloor \left(\frac{R}{2}\sqrt{\frac{L}{2\pi^2 + \pi}} \cdot \frac{1}{\sqrt{\epsilon}} - \frac{1}{2}\right)^d \right\rfloor, \quad \forall 0 < p \leq 1.$$

Otherwise, take $r = \sqrt{(2\pi^2 + \pi)\epsilon / L}$ in Lemma 14. Then we have

$$\eta = \left\lfloor \left(\frac{R-r}{2r}\right)^d \right\rfloor = \left\lfloor \left(\frac{R}{2}\sqrt{\frac{L}{2\pi^2 + \pi}} \cdot \frac{1}{\sqrt{\epsilon}} - \frac{1}{2}\right)^d \right\rfloor \geq 1,$$

such that there exists set $\mathbb{X}_\eta = \{\mathbf{x}_1, \cdots \mathbf{x}_\eta\}$ satisfying $\bigcup_{i=1}^{\eta} \mathbb{B}(\mathbf{x}_i, r) \subset \mathbb{B}(0, R)$ and $\forall i \neq j, \mathbb{B}(\mathbf{x}_i, r) \bigcap \mathbb{B}(\mathbf{x}_j, r) = \emptyset$. Choose $i^* \in \{1, \cdots, \eta\}$ uniformly at random. Let

$$U(\mathbf{x}) = \begin{cases} \dfrac{Lr^2}{4\pi^2 + 2\pi} \cos\left(\dfrac{\pi}{r^2}\left(||\mathbf{x} - \mathbf{x}_{i^*}||_2^2 - r^2\right)\right) - \dfrac{Lr^2}{4\pi^2 + 2\pi}, & ||\mathbf{x} - \mathbf{x}_{i^*}||_2 \leq r \\ 0, \quad ||\mathbf{x} - \mathbf{x}_{i^*}||_2 > r, ||\mathbf{x}||_2 \leq R \\ m\left(||\mathbf{x}||_2 - R\right)^2, ||\mathbf{x}||_2 > R. \end{cases} \tag{43}$$

Lemma 15 (Lipschitz smoothness and strong convexity). *Let $L \geq 2m$. Then $U(\mathbf{x})$ is L-Lipschitz smooth and when $||\mathbf{x}||_2 > 2R$, $U(\mathbf{x})$ is m-strongly convex.*

Now we prove that $\forall\ 0 < p \leq 1$, for any algorithm that inputs $\{U(\mathbf{x}), \nabla U(\mathbf{x}), \cdots, \nabla^n U(\mathbf{x})\}, \forall n \in \mathcal{N}$, $\forall\ \epsilon < \dfrac{LR^2}{18\pi^2 + 9\pi}$, at least $T \geq p \cdot \eta$ steps are required so that $P\left(|U(\mathbf{x}^T) - U(\mathbf{x}^*)| < \epsilon\right) \geq p$.

Note that for any $\mathbf{x}^t \notin \mathbb{B}(\mathbf{x}_{i^*}, r)$, $|U(\mathbf{x}^t) - U(\mathbf{x}^*)| \geq \dfrac{Lr^2}{2\pi^2 + \pi} = \epsilon$. Therefore, probability that $U(\mathbf{x}^t)$ is ϵ close to $U(\mathbf{x}^*)$ is smaller than the probability of $\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r)$:

$$\begin{aligned} P(|U(\mathbf{x}^t) - U(\mathbf{x}^*)| < \epsilon) &\leq P(\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r)) \\ &\leq P\left(\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r) \,\middle|\, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right). \end{aligned} \tag{44}$$

We first assume that $\forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$, then prove that breaking this assumption cannot obtain a better rate of convergence.

1. Assume that $\forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$. From the definition of $U(\mathbf{x})$, (43), we know that $\forall j \in \{1, \cdots, \eta\}, j \ne i^*$, $\forall \mathbf{x} \in \mathbb{B}(\mathbf{x}_j, r)$, $U(\mathbf{x}) = 0$, $\nabla U(\mathbf{x}) = 0$, $\cdots$, $\nabla^n U(\mathbf{x}) = 0$. Hence $\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_j, r)$, $j \ne i^*$ only contains information that $i^* \in \{1, \cdots \eta\} \setminus \{j\}$. Since i is chosen uniformly at random from $\{1, \cdots, \eta\}$, for $T \le \eta$

$$P\left(\mathbf{x}^T \notin \mathbb{B}(\mathbf{x}_{i^*}, r) \,\middle|\, \forall t < T, \mathbf{x}^t \in \bigcup_{\substack{j=1\\ j\ne i^*}}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \ge \frac{\eta - T}{\eta - (T-1)}.$$

   Therefore,

$$\begin{aligned} &P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) = \emptyset \,\middle|\, \forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ \ge& \frac{\eta-1}{\eta} \frac{\eta-2}{\eta-1} \cdots \frac{\eta - T}{\eta - (T-1)} = \frac{\eta - T}{\eta}. \end{aligned} \tag{45}$$

   This implies: the probability that first passage time into set $\mathbb{B}(\mathbf{x}_{i^*}, r)$ is less than or equal to T is:

$$\begin{aligned} &P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) \ne \emptyset \,\middle|\, \forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ =& 1 - P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) = \emptyset \,\middle|\, \forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ \le& 1 - \frac{\eta - T}{\eta} = \frac{T}{\eta}. \end{aligned} \tag{46}$$

   Therefore,

$$\begin{aligned} p &\le P(|U(\mathbf{x}^T) - U(\mathbf{x}^*)| < \epsilon) \\ &\le P\left(\mathbf{x}^T \in \mathbb{B}(\mathbf{x}_{i^*}, r) \,\middle|\, \mathbf{x}^T \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\le P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) \ne \emptyset \,\middle|\, \forall t \le T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\le \frac{T}{\eta}, \end{aligned} \tag{47}$$

$$T \ge p \cdot \eta.$$

2. Suppose there exists an algorithm that output $\{\mathbf{x}_1, \cdots, \mathbf{x}^T\}$, where $\exists\, t \le T$, $\mathbf{x}^t \notin \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$ and finds $\mathbf{x}_{i^*} + r\mathbb{B}$ with probability p within less than $p \cdot \eta$ steps. Then design a corresponding algorithm that outputs $\{\mathbf{x}_1, \cdots, \mathbf{x}^T\} \setminus \left\{\mathbf{x} \middle| \mathbf{x} \notin \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right\}$ so that $\forall\, t \le T$, $\mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$, and $\mathbb{B}(\mathbf{x}_{i^*}, r)$ is found with probability p within less than $p \cdot \eta$ steps. But this contradicts with 1.

C.1.1 Supporting Proofs for Theorem 4

Proof of Lemma 14 (Packing number) Let $\mathcal{P}(r, \mathbb{B}(0,R), ||\cdot||_2)$ be the r-packing number of $\mathbb{B}(0,R)$; and $\mathcal{C}(r, \mathbb{B}(0,R), ||\cdot||_2)$ be the r-covering number of $\mathbb{B}(0,R)$. One can follow the properties of packing and covering numbers to proved that: $\mathcal{P}(r, \mathbb{B}(0,R), ||\cdot||_2) \geq \mathcal{C}(r, \mathbb{B}(0,R), ||\cdot||_2) \geq \left\lfloor \left(\frac{R}{r}\right)^d \right\rfloor$. Therefore, number of non-intersecting r-balls that can be contained in an $\mathbb{B}(0,R)$ is $\mathcal{P}(2r, \mathbb{B}(0,R-r), ||\cdot||_2) \geq \left\lfloor \left(\frac{R-r}{2r}\right)^d \right\rfloor$. ■

Proof of Lemma 14 (Lipschitz smoothness and strong convexity) We first prove that when $||\mathbf{x}||_2 \leq R$, $U(\mathbf{x})$ is L-Lipschitz smooth. We then prove that when $||\mathbf{x}||_2 > R$, $U(\mathbf{x})$ is $2m$-Lipschitz smooth. At last we prove that $U(\mathbf{x})$ is m-strongly convex for $||\mathbf{x}||_2 > 2R$. Since $L \geq 2m$, this proves Lemma 15.

- Define $U_1(\mathbf{x}) = \cos\left(\frac{\pi}{r^2}\left(||\mathbf{x}-\mathbf{x}_i||_2^2 - r^2\right)\right)$. Then $U(\mathbf{x}) = \frac{Lr^2}{4\pi^2+2\pi}\left(U_1(\mathbf{x}) - 1\right)$ when $||\mathbf{x}-\mathbf{x}_i||_2 \leq r$.

  Hessian of U_1 is:

$$
\begin{aligned}
H[U_1](\mathbf{x}) = &-\frac{4\pi^2}{r^4}\cos\left(\frac{\pi}{r^2}\left(||\mathbf{x}-\mathbf{x}_i||_2^2 - r^2\right)\right)(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}} \\
&-\frac{2\pi}{r^2}\sin\left(\frac{\pi}{r^2}\left(||\mathbf{x}-\mathbf{x}_i||_2^2 - r^2\right)\right)\mathbb{I}.
\end{aligned}
$$

  We first note that $\left|\left|(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\right|\right|_2 = ||\mathbf{x}-\mathbf{x}_i||_2^2 \leq r^2$. Hence,

$$
\begin{aligned}
||H[U_1](\mathbf{x})||_2 \leq &\left|\left|\frac{4\pi^2}{r^4}\cos\left(\frac{\pi}{r^2}\left(||\mathbf{x}-\mathbf{x}_i||_2^2 - r^2\right)\right)(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\right|\right|_2 \\
&+\left|\left|\frac{2\pi}{r^2}\sin\left(\frac{\pi}{r^2}\left(||\mathbf{x}-\mathbf{x}_i||_2^2 - r^2\right)\right)\mathbb{I}\right|\right|_2 \\
=&\frac{4\pi^2+2\pi}{r^2}.
\end{aligned}
$$

  Therefore, when $||\mathbf{x}-\mathbf{x}_i||_2 \leq r$, $U(\mathbf{x}) = \frac{Lr^2}{4\pi^2+2\pi}\left(U_1(\mathbf{x}) - 1\right)$ is L-Lipschitz smooth.

  When $||\mathbf{x}-\mathbf{x}_i||_2 > r$ and $||\mathbf{x}||_2 \leq R$, $U(\mathbf{x}) = 0$ is also L-Lipschitz smooth, which leads to the result that $U(\mathbf{x})$ is L-Lipschitz smooth for $||\mathbf{x}||_2 \leq R$.

- Define $U_2(\mathbf{x}) = \left(||\mathbf{x}||_2 - R\right)^2$. Then $U(\mathbf{x}) = mU_2(\mathbf{x})$ when $||\mathbf{x}||_2 > R$.

$$
H[U_2](\mathbf{x}) = 2\left(1 - \frac{R}{||\mathbf{x}||_2}\right)\mathbb{I} + \frac{2R}{||\mathbf{x}||_2^3}\mathbf{x}\mathbf{x}^{\mathrm{T}}.
$$

  Similar to above, it can be proven that $||\mathbf{x}\mathbf{x}^{\mathrm{T}}||_2 = ||\mathbf{x}||_2^2$. Hence $||H[U_2](\mathbf{x})||_2 \leq 2\left|1 - \frac{R}{||\mathbf{x}||_2}\right| + 2\frac{R}{||\mathbf{x}||_2} = 2$. Therefore, $mU_2(\mathbf{x})$ is $2m$-Lipschitz smooth for $||\mathbf{x}||_2 > R$.

- Define

$$
U_3(\mathbf{x}) = \begin{cases} U_2(\mathbf{x}), \ \|\mathbf{x}\|_2 > 2R \\ \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{1}{2}R^2, \ \|\mathbf{x}\|_2 \leq 2R \end{cases}.
$$

  Then

$$
H\left[U_3(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_2^2\right] = \begin{cases} \left(1 - \frac{2R}{||\mathbf{x}||_2}\right)\mathbb{I} + \frac{2R}{||\mathbf{x}||_2^3}\mathbf{x}\mathbf{x}^{\mathrm{T}}, \ \|\mathbf{x}\|_2 > 2R \\ 0, \ \|\mathbf{x}\|_2 \leq 2R \end{cases}.
$$

For any $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y}^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{y} = (\mathbf{y}^{\mathrm{T}}\mathbf{x})^2 \geq 0$. Therefore all eigenvalues of $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ are bigger than or equal to 0. Since $\mathbb{I}$ can be simultaneously diagonalized with $\mathbf{x}\mathbf{x}^{\mathrm{T}}$, $H\left[U_3(\mathbf{x}) - \frac{1}{2}\|x\|_2^2\right] \succeq \left(1 - \frac{2R}{||\mathbf{x}||_2}\right)\mathbb{I} \succeq 0$ when $||\mathbf{x}||_2 > 2R$. When $||\mathbf{x}||_2 \leq 2R$, $H\left[U_3(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_2^2\right] = 0$. Also note that $U_3(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_2^2$ is continuously differentiable. Hence $U_3(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_2^2$ is convex.

On the other hand, $U(\mathbf{x}) = mU_3(\mathbf{x})$ when $||\mathbf{x}||_2 > 2R$. Following Assumption 2, this implies that $U(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex on $\mathbb{R}^d \setminus \mathbb{B}(0, 2R)$. Therefore, $U(\mathbf{x})$ is m-strongly convex on $\mathbb{R}^d \setminus \mathbb{B}(0, 2R)$.

■

C.2 Proof of Corollary 2

Corollary 2. *There exists an objective function U that is m-strongly convex outside of a region of radius $2R$ and L-Lipschitz smooth, such that for $\hat{\mathbf{x}} \sim q_\beta^*$, it is required that $\beta = \widetilde{\Omega}\left(d/\epsilon\right)$ to have $U(\hat{\mathbf{x}}) - U\left(\mathbf{x}^*\right) < \epsilon$ for a constant probability. Moreover, number of iterations required for the Langevin algorithms is $K = e^{\widetilde{\mathcal{O}}\left(d \cdot LR^2/\epsilon\right)}$ to guarantee that $U(\mathbf{x}^K) - U\left(\mathbf{x}^*\right) < \epsilon$ for a constant probability.*

To use Langevin algorithm to attain optimal value with probability p, we separate the optimization problem into two: one is to find a parameter β such that $\hat{\mathbf{x}} \sim q_\beta^* \propto e^{-\beta U}$ has probability p of being close to the optimum $\mathbf{x}^*$ (i.e., $P(U(\hat{\mathbf{x}}) - U\left(\mathbf{x}^*\right) < \epsilon) \geq p$); another is to sample from a distribution q_β^K after K-th iteration so that it is δ-close to q_β^*, for $\delta \leq p/2$ in TV distance. Then by the definition of TV distance, $\mathbf{x}^K \sim q_\beta^K$ will have probability $p/2$ of being close to the optimum $\mathbf{x}^*$.

Proof of Corollary 2 We take U as the one defined in (43) and similarly take $r = \sqrt{(2\pi^2 + \pi)\epsilon/L}$. Then $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbb{R}^d} U(\mathbf{x}) = \mathbf{x}_{i^*}$ and $\min_{\mathbf{x} \in \mathbb{R}^d} U(\mathbf{x}) = -\frac{Lr^2}{2\pi^2 + \pi} = -\epsilon$. For $U(\hat{\mathbf{x}}) - U\left(\mathbf{x}^*\right) < \epsilon$, it is required that $\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq r$.

If $\hat{\mathbf{x}}$ follows the law of q_β^*, then denote the associated probability measure $\mathrm{d}\Pi_\beta^* = q_\beta^* \mathrm{d}\hat{\mathbf{x}}$. We then estimate the probability that $\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x}^*, r)$

$$
\begin{aligned}
P\left(\|\hat{\mathbf{x}}-\mathbf{x}^*\| \leq r\right) =&\Pi_\beta^*\big(\mathbb{B}(\mathbf{x}^*, r)\big) \\
=&\frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}+\int_{\mathbb{B}(0,R)\backslash\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}+\int_{\mathbb{R}^d\backslash\mathbb{B}(0,R)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
=&\frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}+\int_{\mathbb{B}(0,R)\backslash\mathbb{B}(\mathbf{x}^*,r)} 1\ \mathrm{d}\mathbf{x}+\int_{\mathbb{R}^d\backslash\mathbb{B}(0,R)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
=&\frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} \left(e^{-\beta U(\mathbf{x})}-1\right)\mathrm{d}\mathbf{x}+\int_{\mathbb{B}(0,R)} 1\ \mathrm{d}\mathbf{x}+\int_{\mathbb{R}^d\backslash\mathbb{B}(0,R)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
\leq&\frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R)} 1\ \mathrm{d}\mathbf{x}} \\
\leq&\frac{e^{-\min_{\|\mathbf{x}-\mathbf{x}^*\|\leq r}\beta U(\mathbf{x})}\int_{\mathbb{B}(\mathbf{x}^*,r)} 1\ \mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R)} 1\ \mathrm{d}\mathbf{x}} \\
=&e^{\beta\epsilon}\frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} 1\ \mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R)} 1\ \mathrm{d}\mathbf{x}} \\
=&e^{\beta\epsilon}\left(\frac{r}{R}\right)^d.
\end{aligned}
\tag{48}
$$

To obtain that $P(U(\hat{\mathbf{x}})-U(\mathbf{x}^*)<\epsilon) = P\left(\|\hat{\mathbf{x}}-\mathbf{x}^*\| \leq r\right) \geq p$, we need that

$$
e^{\beta\epsilon}\left(\frac{r}{R}\right)^d \geq p.
$$

Therefore,

$$
\beta \geq \frac{1}{\epsilon}\ln p + \frac{d}{\epsilon}\ln\left(\frac{R}{r}\right) = \frac{1}{\epsilon}\ln p + \frac{1}{2}\frac{d}{\epsilon}\ln\left(\frac{1}{2\pi^2+\pi}\frac{LR^2}{\epsilon}\right).
$$

To use the Langevin algorithms to search for optimum, we are actually using $\mathbf{x}^K$, which follows the sampled distribution q_β^K at K-th step. And we are taking K large enough so that $\left\|q_\beta^K - q_\beta^*\right\|_{\mathrm{TV}} \leq \delta$, for $\delta \leq p/2$. Then, for a large enough K, we can have

$$
\begin{aligned}
&\left|P\left(\left\|\mathbf{x}^K-\mathbf{x}^*\right\| \leq r\right) - P\left(\|\hat{\mathbf{x}}-\mathbf{x}^*\| \leq r\right)\right| \\
=&\left|\Pi_\beta^K\big(\mathbb{B}(\mathbf{x}^*, r)\big) - \Pi_\beta^*\big(\mathbb{B}(\mathbf{x}^*, r)\big)\right| \\
\leq&\sup_A\left|\Pi_\beta^K(A)-\Pi_\beta^*(A)\right| \\
=&\left\|q_\beta^K - q_\beta^*\right\|_{\mathrm{TV}} \\
\leq&\delta,
\end{aligned}
\tag{49}
$$

which guarantees that $P\left(\left\|\mathbf{x}^K-\mathbf{x}^*\right\| \leq r\right) \geq p/2$.

We directly obtain from Theorem 3 that for the objective function βU with Lipschitz constant $\beta L \geq \frac{L}{\epsilon}\ln p + \frac{d}{2}\frac{L}{\epsilon}\ln\left(\frac{1}{2\pi^2+\pi}\frac{LR^2}{\epsilon}\right)$, we need to iterate $e^{\widetilde{\mathcal{O}}\left(d\cdot LR^2/\epsilon\right)}$ steps to guarantee convergence.

■

Appendix D Proofs for Gaussian Mixture Models

Consider the problem of inferring mean parameters $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_M) \in \mathbb{R}^{d\times M}$ in a Gaussian mixture model with M mixtures from N data $\mathbf{y} = (y_1, \cdots, y_N)$:

$$p\left(y_n|\boldsymbol{\mu}\right) = \sum_{i=1}^{M} \frac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}(y_n - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (y_n - \mu_i)\right) + \left(1 - \sum_{i=1}^{M} \lambda_i\right) p_0(y_n), \tag{50}$$

where Z_i are the normalization constants and $\sum_{i=1}^{M} \lambda_i \leq 1$. For succinctness, we consider in this section the cases where covariances Σ_i are isotropic and uniform across all mixture components: $\Sigma_i = \Sigma = \sigma^2 \mathbb{I}$. $p_0(y_n)$ represents crude observations of the data (e.g., data may be distributed inside a region or may have sub-Gaussian tail behavior). The objective function is given by the log posterior distribution: $U(\boldsymbol{\mu}) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right)$. Assume data are distributed in a bounded region ($\|y_n\| \leq R$) and take $p_0(y_n) = \mathbb{1}\left\{\|y_n\| \leq R\right\}/Z_0$ to describe this observation.

We also take the prior to be

$$p(\boldsymbol{\mu}) \propto \exp\left(-m\left(\|\boldsymbol{\mu}\|_F - \sqrt{M}R\right)^2 \mathbb{1}\left\{\|\boldsymbol{\mu}\|_F \leq \sqrt{M}R\right\}\right). \tag{51}$$

D.1 Proofs for Smoothness

Fact 1. *For the Gaussian mixture model defined in* (50)*, define*

$$\alpha = \frac{1}{\sigma^2} \max\left\{ 2 \sup_{\mu \in \{\mu_1, \cdots, \mu_M\}} \sum_{n=1}^{N} \frac{\|\mu - y_n\|^2}{\sigma^2} \exp\left(-||\mu - y_n||^2/2\sigma^2\right), \sup_{\mu \in \{\mu_1, \cdots, \mu_M\}} \sum_{n=1}^{N} \exp\left(-||\mu - y_n||^2/2\sigma^2\right) \right\}. \tag{52}$$

If we take $\lambda_i = \dfrac{\frac{l}{\alpha} Z_i}{Z_0 + \frac{l}{\alpha}\sum_{j=1}^{M} Z_j}$*, then the log-likelihood* $-\sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right)$ *is l-Lipschitz smooth.*

Proof of Fact 1 Define the mixture components: $W_{i,n} = \dfrac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right)$ and $C_n = \left(1 - \sum_{i=1}^{M} \lambda_i\right) p_0(y_n)$. Since all the data $\{y_n\}$ are distributed in $\mathbb{B}(0, R)$, $p_0(y_n) = \dfrac{1}{Z_0}\mathbb{1}\left\{\|y_n\| \leq R\right\} = \dfrac{1}{Z_0}$. We can plug in the expression of $\lambda_i = \dfrac{\frac{l}{\alpha} Z_i}{Z_0 + \frac{l}{\alpha}\sum_{j=1}^{M} Z_j}$ and obtain for any $n = 1, \cdots, N$:

$$\begin{aligned}
C_n = C =& \frac{1}{Z_0}\left(1 - \sum_{i=1}^{M} \lambda_i\right) \\
=& \frac{1}{Z_0}\left(1 - \frac{\frac{l}{\alpha}\sum_{i=1}^{M} Z_i}{Z_0 + \frac{l}{\alpha}\sum_{j=1}^{M} Z_j}\right) \\
=& \frac{1}{Z_0 + \frac{l}{\alpha}\sum_{j=1}^{M} Z_j}.
\end{aligned} \tag{53}$$

Then we can use C to simplify the expression of λ_i for $i = 1, \cdots, M$:

$$\lambda_i = \frac{l}{\alpha} C Z_i.$$

We also represent the objective function as:

$$U(\boldsymbol{\mu}) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log\left(\sum_{i=1}^{M} W_{i,n} + C\right),$$

and define

$$\gamma_{i,n} = \frac{W_{i,n}}{\sum_{k=1}^{M} W_{k,n} + C}.$$

One can find that

$$-\nabla_{\mu_i} \log p\left(y_n|\boldsymbol{\mu}\right) = \frac{W_{j,n}}{\sum_{j=1}^{M} W_{j,n} + C} \frac{\mu_i - y_n}{\sigma^2} = \gamma_{i,n} \frac{\mu_i - y_n}{\sigma^2},$$

and

$$-\nabla^2_{\mu_i,\mu_j} \log p\left(y_n|\boldsymbol{\mu}\right) = \begin{cases} \dfrac{\gamma_{i,n}}{\sigma^2}\mathbb{I} + (\gamma_{i,n}^2 - \gamma_{i,n})\dfrac{(\mu_i - y_n)(\mu_i - y_n)^T}{\sigma^4}, & i = j \\ \gamma_{i,n}\gamma_{j,n}\dfrac{(\mu_i - y_n)(\mu_j - y_n)^T}{\sigma^4}, & i \neq j \end{cases}.$$

Therefore,

$$\begin{aligned} & -\mathbf{v}^T \nabla^2_{\boldsymbol{\mu}^2} \log p\left(y_n|\boldsymbol{\mu}\right) \mathbf{v} \\ = & \sum_{i=1}^{M} \frac{\gamma_{i,n}}{\sigma^2} v_i^T v_i - \sum_{i=1}^{M} \gamma_{i,n} \left[v_i^T \left(\frac{\mu_i - y_n}{\sigma^2}\right)\right]^2 \\ & + \sum_{i=1}^{M}\sum_{j=1}^{M} \gamma_{i,n}\gamma_{j,n} \left[v_i^T \left(\frac{\mu_i - y_n}{\sigma^2}\right)\right] \left[v_j^T \left(\frac{\mu_j - y_n}{\sigma^2}\right)\right]. \end{aligned}$$

Since $\sum_{i=1}^{M} \gamma_{i,n} = \sum_{i=1}^{M} \dfrac{W_{i,n}}{\sum_{k=1}^{M} W_{k,n} + C} \leq 1$,

$$\begin{aligned} & \left|\sum_{i=1}^{M}\sum_{j=1}^{M} \gamma_{i,n}\gamma_{j,n} \left[v_i^T \left(\frac{\mu_i - y_n}{\sigma^2}\right)\right] \left[v_j^T \left(\frac{\mu_j - y_n}{\sigma^2}\right)\right]\right| \\ \leq & \frac{1}{2} \sum_{i=1}^{M}\sum_{j=1}^{M} \gamma_{i,n}\gamma_{j,n} \left(\left[v_i^T \left(\frac{\mu_i - y_n}{\sigma^2}\right)\right]^2 + \left[v_j^T \left(\frac{\mu_j - y_n}{\sigma^2}\right)\right]^2\right) \\ \leq & \gamma_{i,n} \left[v_i^T \left(\frac{\mu_i - y_n}{\sigma^2}\right)\right]^2. \end{aligned}$$

Therefore

$$\operatorname{diag}\left(\frac{\gamma_{i,n}}{\sigma^2}\left(1 - 2\frac{\|\mu_i - y_n\|^2}{\sigma^2}\right)\mathbb{I}\right) \preceq \nabla^2_{\boldsymbol{\mu}^2} \log p\left(y_n|\boldsymbol{\mu}\right) \preceq \operatorname{diag}\left(\frac{\gamma_{i,n}}{\sigma^2}\mathbb{I}\right).$$

Since $\{W_{i,n}\}$ are positive,

$$\gamma_{i,n} = \frac{W_{i,n}}{\sum_{j=1}^{M} W_{j,n} + C} \leq \frac{W_{i,n}}{C} = \frac{\lambda_i}{C Z_i} \exp\left(-||\mu_i - y_n||^2 / 2\sigma^2\right).$$

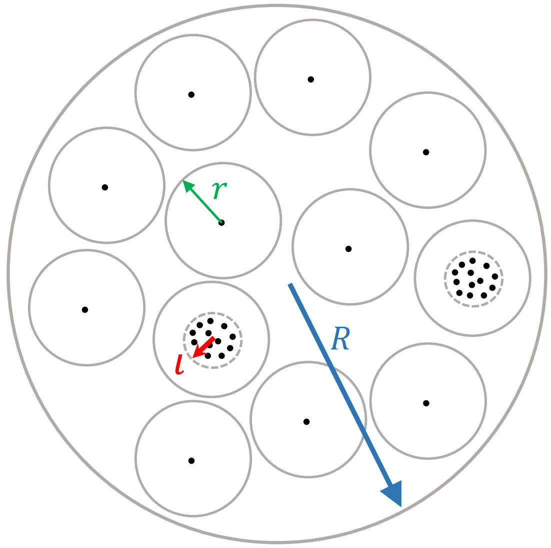

Figure 4: Cartoon of an exemplified dataset.

Since

$$\alpha = \frac{1}{\sigma^2} \max \left\{ 2 \sup_{\mu} \sum_{n=1}^{N} \frac{\|\mu - y_n\|^2}{\sigma^2} \exp\left(-||\mu - y_n||^2/2\sigma^2\right), \sup_{\mu} \sum_{n=1}^{N} \exp\left(-||\mu - y_n||^2/2\sigma^2\right) \right\}. \tag{54}$$

and

$$\lambda_i = \frac{l}{\alpha} C Z_i, \tag{55}$$

log-likelihood $-\sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right)$ is l-Lipschitz smooth. It can be seen from (54) that if one uses a loose upper bound for α, we can simply take λ_i to be $\frac{l}{2} \frac{C Z_i \sigma^2}{N}$. ■

D.2 Proofs for the EM Algorithm

We prove in the following Lemma that there exists a dataset $(y_1, \cdots, y_N)$ and variance σ^2 with the previous setting that takes $K \geq \min\{\mathcal{O}(d^{1/\epsilon}), \mathcal{O}(d^d)\}$ steps for the EM algorithm to converge if one initializes the algorithm close to the given data points.

Lemma 16. *Let the objective function $U(\boldsymbol{\mu}) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right)$ with prior $p(\boldsymbol{\mu})$ and likelihood $p\left(y_n|\boldsymbol{\mu}\right)$ defined in* (51) *and* (50)*. Take the parameters λ_i so that the log-likelihood is Lipschitz smooth with Lipschitz constant $L = 1/16$, strong convexity constant $m = 1/64$ outside of region with radius $R = 1/2$, and number of mixtures $M = \log_2 d$. Then there exists a dataset $(y_1, \cdots, y_N)$ and variance σ^2 so that the EM algorithm will take $K \geq \min\{\mathcal{O}(d^{1/\epsilon}), \mathcal{O}(d^d)\}$ queries to converge to $\mathcal{O}(\epsilon)$ close to the optimum if one randomly initializes the algorithm* 0.01 *close to the given data points.*

Proof of Lemma 16 shares similar traits as that in [22, 1].

Directly invoking Theorem 3, we know that the Langevin algorithms converge within $K \leq \widetilde{\mathcal{O}}\left(d^3/\epsilon\right)$ and $K \leq \widetilde{\mathcal{O}}\left(d^3 \ln^2\left(1/\epsilon\right)\right)$ steps, respectively.

Proof of Lemma 16 Consider a dataset with N number of d-dimensional data points, $y_n \in \mathbb{R}^d$, $n = 1, \cdots, N$, described below. We suppose that it is modeled with $M < N$ mixture components in the Gaussian mixture model (50).

For the first $N - 9M$ points, let $\|y_n\| \leq 0.45$, and $\|y_k - y_l\| \geq 0.11$, where $n, k, l \in \{1, \cdots, N - 9M\}$ and $k \neq l$. From Lemma 14, we know that when $N \leq 2^d$, this setting is feasible. For the next $9M$ points, first select M different indices $\{i_1, \cdots, i_M\}$ from $\{1, \cdots, N - 9M\}$ uniformly at random. Then for $n \in \{N - 9M + 9(k-1) + 1, \cdots, N - 9M + 9k\}$ ($k \in \{1, \cdots, M\}$), $\|y_n - y_{i_k}\| \leq \sigma/2$.

By this setting, $\forall y_n$, $\|y_n\| \leq 0.5$. Furthermore, when $n, \hat{n} \in \{N-9M+9(k-1)+1, \cdots, N-9M+9k\} \cup \{i_k\}$, $\|y_n - y_{\hat{n}}\| \leq \sigma/2$; otherwise, $\|y_n - y_{\hat{n}}\| \geq 0.1$ for $n \neq \hat{n}$.

Since it can be observed that all the data are distributed in $\mathbb{B}(0, 0.5)$, we let $p_0(y_n) = \dfrac{1}{Z_0} \mathbb{1}\{\|y_n\| \leq 0.5\} = \dfrac{\Gamma(d/2+1)}{(2\pi)^{d/2}} \mathbb{1}\{\|y_n\| \leq 0.5\}$. Inclusion of p_0 provides a better description of the data, since they are mostly distributed uniformly in $\mathbb{B}(0, 0.5)$, with some concentrated around the chosen M centers. Then according to (51), we set the prior to be:

$$p(\boldsymbol{\mu}) \propto \exp\left(-\frac{\left(\|\boldsymbol{\mu}\|_F - \sqrt{M}/2\right)^2}{64} \mathbb{1}\left\{\|\boldsymbol{\mu}\|_F \leq \sqrt{M}/2\right\}\right),$$

where $\|\boldsymbol{\mu}\|_F = \sqrt{\sum_{i=1}^M \|\mu_i\|_2^2}$. Note that in this setting, the positions of local minima are exactly the same as the Gaussian mixture model that does not include prior observation $p_0(y)$ and prior belief $p(\boldsymbol{\mu})$.

We take $\lambda_i = \dfrac{1}{64\alpha} C Z_i$ (using notations defined in (55) and (53)). Then the objective function defined via the log posterior:

$$\begin{aligned}
U(\boldsymbol{\mu}) &= -\log p(\boldsymbol{\mu}) - \sum_{n=1}^N \log p\left(y_n|\boldsymbol{\mu}\right) \\
&= -\log p(\boldsymbol{\mu}) - \sum_{n=1}^N \log\left(\sum_{i=1}^M \frac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right) + C\right) \\
&= \frac{\left(\|\boldsymbol{\mu}\|_2 - \sqrt{M}/2\right)^2}{64} \mathbb{1}\left\{\|\boldsymbol{\mu}\|_2 \leq \sqrt{M}/2\right\} \\
&\quad - \sum_{n=1}^N \log\left(\sum_{i=1}^M \frac{1}{64\alpha} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right) + 1\right) + \widetilde{C}
\end{aligned} \tag{56}$$

has Lipschitz smoothness $L \leq 1/32$. In what follows, we take $\sigma = \sigma = \dfrac{0.01}{\sqrt{\log_2 N}}$.

It can be seen that α in (54) is bounded as: $\alpha \leq \dfrac{50}{\sigma^2}$. Then $\lambda_i = \dfrac{1}{3200} C Z_i \sigma^2$. It can also be checked that the objective function U is also $m \geq 1/64$ strongly convex for $\|\boldsymbol{\mu}\|_F \geq \sqrt{M}$.

We then estimate number of fixed points for $\|\boldsymbol{\mu}\|_F \leq \sqrt{M}/2$ when running the EM algorithm. If we run the EM algorithm starting with $\|\boldsymbol{\mu}^{(t)}\|_F \leq \sqrt{M}/2$, we first compute the weights for each component using old value $\boldsymbol{\mu}^{(t)}$ (in E step):

$$\gamma_{i,n}^{(t)} = \frac{\frac{\lambda_i}{Z_i}\exp\left(-\frac{1}{2}||y_n-\mu_i^{(t)}||^2/\sigma^2\right)}{\sum_{j=1}^M \frac{\lambda_j}{Z_j}\exp\left(-\frac{1}{2}||y_n-\mu_j^{(t)}||^2/\sigma^2\right)+C}$$
$$= \frac{\frac{\sigma^2}{3200}\exp\left(-\frac{1}{2}||y_n-\mu_i^{(t)}||^2/\sigma^2\right)}{\sum_{j=1}^M \frac{\sigma^2}{3200}\exp\left(-\frac{1}{2}||y_n-\mu_j^{(t)}||^2/\sigma^2\right)+1}. \tag{57}$$

We then update $\boldsymbol{\mu}$ (in M step):

$$\mu_i^{(t+1)} = \sum_{n=1}^N \frac{\gamma_{i,n}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} y_n.$$

We prove in Lemma 17 that $\forall y_{n_i}, n_i \in \{1,\cdots,N/2\}$, if $\|\mu_i^{(0)} - y_{n_i}\| \le 0.01$, then $\|\mu_i^{(\tau)} - y_{n_i}\| \le 0.01$, $\forall \tau > 0$. Therefore, any M combinations of $N-9M$ data points is a fixed point for $\boldsymbol{\mu}$.

Lemma 17. *Suppose we run the EM algorithm with the dataset specified in the beginning of Sec. D.2 for T steps. If we initialized each component of $\boldsymbol{\mu}$ with $\|\mu_i^{(0)} - y_{n_i}\| \le 0.01$ for $n_i \in \{1,\cdots,N/2\}$, then $\|\mu_i^{(\tau)} - y_{n_i}\| \le 0.01$, $\forall \tau > 0$.*

We note that the global minima $\boldsymbol{\mu}^* = (\mu_1^*,\cdots,\mu_M^*)$ will have $\forall i \in \{1,\cdots,M\}$, $\mu_i^* \in \bigcup_{k=1}^M \Omega_k$, where we denote $\Omega_k = \{N-9M+9(k-1)+1,\cdots,N-9M+9k\} \cup \{i_k\}$. It can also be checked from (56) that the difference ϵ between the global minima and any local minimum $\bar{\boldsymbol{\mu}}$ that has $\exists i \in \{1,\cdots,M\}$, s.t. $\bar{\mu}_i \notin \bigcup_{k=1}^M \Omega_k$ scales with N as $\epsilon = \mathcal{O}(\sigma^2) = \mathcal{O}\left(\frac{1}{\log_2 N}\right)$. Therefore, if one randomly initialize from the dataset, to attain global minima with probability p, at least $K = p \cdot \begin{pmatrix} N \\ M \end{pmatrix} \Big/ \begin{pmatrix} 10M \\ M \end{pmatrix} \ge p \cdot \left(\frac{N}{10M}\right)^M$ re-initializations are required. Let $N \gg M$. Then the number of re-initializations are of order $K = \mathcal{O}(p \cdot N^M)$.

Note that we have taken $M = \log_2 d$. For $\epsilon > \mathcal{O}(1/d)$, take $N = \mathcal{O}\left(2^{1/\epsilon}\right)$. Then $T = \mathcal{O}\left(d^{1/\epsilon}\right)$. For $\epsilon \le \mathcal{O}(1/d)$, take $N = 2^d$. Then $T = \mathcal{O}(d^d)$. So $T = \min\left\{\mathcal{O}\left(d^{1/\epsilon}\right), \mathcal{O}(d^d)\right\}$. ■

Proof of Lemma 17 We prove for each component μ_i using induction over $t \in \{0,\cdots,\tau\}$. First assume that $\|\mu_i^{(t)} - y_{n_i}\| \le 0.01$.

Then we observe from (57) that $\forall i, n$,

$$\gamma_{i,n}^{(t)} = \left(\sum_{j=1}^M \exp\left(\frac{1}{2}||y_n-\mu_i||^2/\sigma^2 - \frac{1}{2}||y_n-\mu_j||^2/\sigma^2\right) + \frac{3200}{\sigma^2}\exp\left(\frac{1}{2}||y_n-\mu_i||^2/\sigma^2\right)\right)^{-1}.$$

Since $\sum_{j=1}^M \exp\left(-\frac{1}{2}||y_n-\mu_j||^2/\sigma^2\right) \le M \le 3200/\sigma^2$,

$$\frac{\sigma^2}{6400}\exp\left(-\frac{1}{2}||y_n-\mu_i||^2/\sigma^2\right) \le \gamma_{i,n}^{(t)} \le \frac{\sigma^2}{3200}\exp\left(-\frac{1}{2}||y_n-\mu_i||^2/\sigma^2\right). \tag{58}$$

Therefore, when $\|\mu_i^{(t)} - y_n\| \le 0.01$, $\gamma_{i,n}^{(t)} \ge \frac{\sigma^2}{6400} N^{-1/2}$; when $\|\mu_i^{(t)} - y_n\| \le 0.015$, $\gamma_{i,n}^{(t)} \ge \frac{\sigma^2}{6400} N^{-9/8}$; when $\|\mu_i^{(t)} - y_n\| \ge 0.1$, $\gamma_{i,n}^{(t)} \le \frac{\sigma^2}{3200} N^{-50}$.

- For $n_i \in \{1, \cdots, N-9M\} \setminus \{i_1, \cdots, i_M\}$,

$$\begin{aligned}\|\mu_i^{(t+1)} - y_{n_i}\| \le & \frac{\gamma_{i,n_i}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{n_i} - y_{n_i}\| + \frac{\sum_{\tilde{n}\neq n_i} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{n_i}\| \\ = & \frac{\sum_{\tilde{n}\neq n_i} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{n_i}\|.\end{aligned}$$

Since $\|\mu_i^{(t)} - y_{n_i}\| \le 0.01$ and $\|\mu_i^{(t)} - y_{\hat{n}}\| \ge 0.1$, $\forall \hat{n} \neq n_i$ (and that $N \ge 2$),

$$\frac{\sum_{\tilde{n}\neq n_i} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \le \frac{\sum_{\hat{n}\neq n_i} \gamma_{i,\hat{n}}^{(t)}}{\gamma_{i,n_i}^{(t)}} \le \frac{N \cdot \frac{\sigma^2}{3200} N^{-50}}{\frac{\sigma^2}{6400} N^{-1/2}} \le 10^{-10}. \tag{59}$$

Hence

$$\begin{aligned}\|\mu_i^{(t+1)} - y_{n_i}\| \le & \frac{\sum_{\tilde{n}\neq n_i} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{n_i}\| \\ \le & 2 \cdot 10^{-10} \sup_{\hat{n}} \|y_{\hat{n}}\| \le 10^{-10} \le 0.01.\end{aligned}$$

- Denote $\Omega_k = \{N - 9M + 9(k-1) + 1, \cdots, N - 9M + 9k\} \cup \{i_k\}$. For $n_i \in \Omega_k$, $\forall k \in \{1, \cdots, M\}$,

$$\|\mu_i^{(t+1)} - y_{n_i}\| \le \|\mu_i^{(t+1)} - y_{i_k}\| + \|y_{n_i} - y_{i_k}\| \le \|\mu_i^{(t+1)} - y_{i_k}\| + \frac{\sigma}{2}.$$

And

$$\|\mu_i^{(t+1)} - y_{i_k}\| \le \left\| \sum_{\tilde{n}\in\Omega_k} \frac{\gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \ (y_{\tilde{n}} - y_{i_k}) \right\| + \sum_{\tilde{n}\notin\Omega_k} \frac{\gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{i_k}\|.$$

Define

$$y_{i_k}^{avg} = \frac{\sum_{\tilde{n}\in\Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}\in\Omega_k} \gamma_{i,\hat{n}}^{(t)}} y_{\tilde{n}}.$$

Then

$$\|\mu_i^{(t+1)} - y_{i_k}\| \le \frac{\sum_{\tilde{n}\in\Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \left\| y_{i_k}^{avg} - y_{i_k} \right\| + \frac{\sum_{\tilde{n}\notin\Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^N \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{i_k}\|.$$

Since $\sup_{\tilde{n}\in\Omega_k} \|y_{\tilde{n}} - y_{i_k}\| \le \sigma/2$, $\|y_{i_k}^{avg} - y_{i_k}\| \le \sigma/2$. And for any $\tilde{n} \in \Omega_k$, we use induction assumption and $\sup_{\tilde{n}\in\Omega_k} \|y_{\tilde{n}} - y_{i_k}\| \le \sigma/2$ to obtain that

$$\|\mu_i^{(t)} - y_{\tilde{n}}\| \le \|\mu_i^{(t)} - y_{n_i}\| + \|y_{n_i} - y_{i_k}\| + \|y_{i_k} - y_{\tilde{n}}\| \le 0.1 + \frac{\sigma}{2} + \frac{\sigma}{2} \le 0.015.$$

Hence $\gamma_{i,\tilde{n}}^{(t)} \geq \frac{\sigma^2}{4N} N^{-9/8}$. Then similar to (59),

$$\frac{\sum_{\tilde{n} \notin \Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^{N} \gamma_{i,\hat{n}}^{(t)}} \leq \frac{\sum_{\tilde{n} \notin \Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n} \in \Omega_k} \gamma_{i,\hat{n}}^{(t)}} \leq 10^{-10}.$$

Therefore,

$$\begin{aligned}\|\mu_i^{(t+1)} - y_{n_i}\| \leq & \frac{\sum_{\tilde{n} \in \Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^{N} \gamma_{i,\hat{n}}^{(t)}} \left\|y_{i_k}^{avg} - y_{i_k}\right\| + \frac{\sum_{\tilde{n} \notin \Omega_k} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^{N} \gamma_{i,\hat{n}}^{(t)}} \|y_{\tilde{n}} - y_{i_k}\| + \frac{\sigma}{2} \\ \leq & \|y_{i_k}^{avg} - y_{i_k}\| + 10^{-10} \cdot 1 + \frac{\sigma}{2} \leq \sigma + 10^{-10} \leq 0.01.\end{aligned}$$

It follows from induction that if $\|\mu_i^{(0)} - y_{i_k}\| \leq 0.01$, then $\|\mu_i^{(\tau)} - y_{i_k}\| \leq 0.01$, $\forall \tau > 0$. ∎

References

[1] C. Améndola, M. Drton, and B. Sturmfels. Maximum likelihood estimates for Gaussian mixtures are transcendental. In *International Conference on Mathematical Aspects of Computer and Information Sciences*, pages 579–590. 2015.

[2] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J Mach Learn Res*, 15:595–627, 2014.

[3] D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. 1985.

[4] S. G. Bobkov. *On Isoperimetric Constants for Log-Concave Probability Distributions*, pages 81–88. Springer, Berlin Heidelberg, 2007.

[5] S. G. Bobkov and P. Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *J Theor Probab*, 19(2):289–336, 2006.

[6] S. G. Bobkov and B. Zegarlinski. *Entropy Bounds and Isoperimetry*. Number 829. American Mathematical Soc, 2005.

[7] N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. arXiv:1805.00452, 2018.

[8] P. Buser. A note on the isoperimetric constant. *Ann Sci Éc Norm Supér*, 15(2):213–230, 1982.

[9] X. Cheng and P. L. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, pages 186–211, 2018.

[10] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv:1805.01648, 2018.

[11] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 300–323, 2018.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 2012.

[13] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J Royal Stat Soc B*, 79(3):651–676, 2017.

[14] A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. arXiv:1710.00095, 2017.

[15] A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. arXiv:1605.01559, 2016.

[16] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann Appl Probab*, 27(3):1551–1587, 06 2017.

[17] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! arXiv:1801.02309, 2018.

[18] A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. arXiv:1703.01617, 2017.

[19] H. Ge and H. Qian. Landscapes of non-gradient dynamics without detailed balance: Stable limit cycles and multiple attractors. *Chaos*, 22(2):023140, 2012.

[20] R. Holley and D. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *J Stat Phys*, 46(5):1159–1194, 1987.

[21] P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3–4):142–336, 2017.

[22] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 4116–4124. 2016.

[23] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[24] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284, 1940.

[25] L. D. Landau and E. M. Lifshitz. *Statistical Physics*. Oxford, Pergamon, 3rd edition, 1980.

[26] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. 1999.

[27] M. Ledoux. The geometry of Markov diffusion generators. *Ann Fac Sci Toulouse Math*, 9(6):305–366, 2000.

[28] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Struct Alg*, 4(4):359–412, 1993.

[29] M. B. Majka, A. Mijatović, and Lukasz Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. arXiv:1808.07105, 2018.

[30] O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. arXiv:1708.07114, 2017.

[31] O. Mangoubi and N. K. Vishnoi. Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. arXiv:1802.08898, 2018.

[32] J.-M. Marin, K. Mengersen, and C. P. Robert. *Bayesian Modelling and Inference on Mixtures of Distributions*. Springer-Verlag, New York, 2005.

[33] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, Chichester, 2000.

[34] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Boston, 2004.

[35] B. Øksendal. *Stochastic Differential Equations*. Springer, Berlin, 6th edition, 2003.

[36] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J Funct Anal*, 173(2):361–400, 2000.

[37] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer, New York, 2014.

[38] H. J. M. Peters and P. P. Wakker. Convex functions on non-convex domains. *Econ Lett*, 22(2):251–255, 1986.

[39] B. T. Polyak. Gradient methods for minimizing functionals. *Zh Vychisl Mat Mat Fiz*, 3(4):643–653, 1963.

[40] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1674–1703, 2017.

[41] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodol Comput Appl Probab*, 4:337–357, 2002.

[42] P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *J Chem Phys*, 69(10):4628, 1978.

[43] A. Uhlmann. Roofs and convexity. *Entropy*, 12:1799–1832, 2010.

[44] C. Villani. *Optimal Transport: Old and New*. Wissenschaften. Springer, Berlin, 2009.

[45] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. arXiv:1802.08089, 2018.

[46] M. Yan. Extension of convex function. *J Convex Anal*, 21(4):965–987, 2014.