PNAS
www.pnas.org

Supplementary Information for

Sampling Can Be Faster Than Optimization

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion and Michael I. Jordan

Michael I. Jordan.
E-mail: jordan@cs.berkeley.edu

This PDF file includes:

Supplementary text
Figs. S1 to S3
References for SI reference citations

Supporting Information Text

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

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

   That is: $V(\mathbf{x}) = U(\mathbf{x}) - \frac{m}{2} \|\mathbf{x}\|_2^2$ is convex on $\Omega = \mathbb{R}^d \setminus \mathbb{B}(0, R)$*. We then follow the definition of convexity on nonconvex domains (1, 2) 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 $\mathbf{x}_1, \cdots, \mathbf{x}_k \in \Omega$ satisfies:

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

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

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

B. Proofs for Sampling.

Theorem 1. *For $p^* \propto e^{-U}$, we assume that U satisfies the local nonconvexity Assumptions 1–3. Consider the unadjusted Langevin algorithm (ULA) and the 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*

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

For MALA,

$$\tau_{MALA}(\epsilon, p^0) \le \mathcal{O}\left(\frac{e^{40LR^2}}{m} \kappa^{3/2} d^{1/2} \left(d \ln \kappa + \ln\left(\frac{1}{\epsilon}\right)\right)^{3/2}\right). \quad [2]$$

Remark 1. *Assumptions 1–3 can be shown to imply that the nonconvex region will have small probability mass in high dimensions. The theorem quantifies the consequences of this small mass on ULA and MALA, showing essentially that their mixing time is not perturbed qualitatively by the nonconvexity. It is the coupling of this result with the exponential complexity of optimization, as shown in Theorem 2, that is our main result. The assumptions have been chosen to make this comparison as simple as possible. But it is noteworthy that we can weaken the assumptions and still obtain rapid mixing for ULA and MALA. In particular, note that we assumed that the Lipschitz parameter L is uniformly bounded by a constant over the entire $\mathbb{R}^d$. This assumption is in fact not necessary in our proofs. Indeed, we can allow the Lipschitz parameter $\widetilde{L}$ and strong convexity parameter $\widetilde{m}$ outside of the region $\mathbb{B}(0, R)$ to scale with the dimension d (while U is still L-Lipschitz smooth inside $\mathbb{B}(0, R)$ and L does not scale with d). In that setup, the probability mass inside the nonconvex region $\mathbb{B}(0, R)$ no longer shrinks as a function of d.*

Moreover, in that setup we can repeat the constructive proof in Lemma 1 (via choosing a smaller smoothing radius $\delta = \mathcal{O}(\kappa R/d)$) and demonstrate that $\rho_U \ge L e^{-16LR^2}$. It follows that the computational complexity for ULA becomes (in terms of dimension d and accuracy ϵ): $\mathcal{O}(d^3/\epsilon^2)$, where the extra d^2 factor is due to the fact that the step size h scales inversely with $\widetilde{L}^2 = \mathcal{O}(d^2)$. A similar result holds for MALA.

This more general setup highlights the value of our general approach to analyzing MCMC algorithms via the properties of weighted Sobolev spaces. It naturally allows us to combine convergence rates for sampling strongly log-concave posteriors and those for sampling smooth posteriors in a bounded region. Indeed, our upper bounds on convergence rates generalize existing results for strongly log-concave posteriors (3–10) and also strengthen recent work using the Wasserstein metric to the KL divergence (11–14).

We begin by proving the basic log-Sobolev inequality that underlies our results. We then prove convergence of ULA and MALA respectively in Sec. B.2 and B.4.

* Here we let $\mathbb{B}(0, R)$ denote the closed ball of radius R centered at 0.

B.1. Log-Sobolev Inequality.

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

$$\rho_U \geq \frac{m}{2} e^{-16LR^2}. \tag*{[3]}$$

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 (15) for strongly log concave density and have:

$$\rho_{\hat{U}} \geq \frac{m}{2}. \tag*{[4]}$$

Next we invoke Lemma 1 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 a result from Holley-Stroock (16) and obtain:

$$\rho_U \geq \frac{m}{2} e^{-\left|\sup\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right) - \inf\left(\hat{U}(\mathbf{x}) - U(\mathbf{x})\right)\right|} \geq \frac{m}{2} e^{-16LR^2}. \tag*{[5]}$$

$\blacksquare$

Lemma 1. *For U satisfying Assumptions 1–3, there exists $\hat{U} \in C^1(\mathbb{R}^d)$ with a Hessian that exists everywhere on $\mathbb{R}^d$, and $\hat{U}$ that is $m/2$-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 1 Similar to Assumptions 1–3, denote $\Omega = \mathbb{R}^d \setminus \mathbb{B}(0, R)$. Also denote $\tilde{U}(\mathbf{x}) = U(\mathbf{x}) - \frac{m}{4}\|\mathbf{x}\|^2$.

We follow (2) 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 is convex on $\mathbb{R}^d$ and differs from $\tilde{U}(\mathbf{x})$ less than $16LR^2$.

First we define the function V as the convex extension (17) of $\tilde{U}$ from domain Ω to its convex hull Ω^{co}:

$$V(\mathbf{x}) = \inf_{\substack{\{\mathbf{x}_i\} \subset \Omega, \\ \left\{\lambda_i \middle| \sum_i \lambda_i = 1\right\}, \\ \text{s.t.}, \sum_i \lambda_i \mathbf{x}_i = \mathbf{x}}} \left\{\sum_{i=1}^{l} \lambda_i \tilde{U}(\mathbf{x}_i)\right\}, \quad \forall \mathbf{x} \in \Omega^{co} = \mathbb{R}^d. \tag*{[6]}$$

$V(\mathbf{x})$ is convex on the entire domain $\mathbb{R}^d$. Also, since $\tilde{U}(\mathbf{x})$ is convex in Ω, $V(\mathbf{x}) = \tilde{U}(\mathbf{x})$ for $\mathbf{x} \in \Omega$. By Lemma 2, 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

$$\tilde{V}(\mathbf{x}) = \int V(\mathbf{y})\phi(\mathbf{x} - \mathbf{y}) \mathrm{d}\mathbf{y} = \int V(\mathbf{x} - \mathbf{y})\phi(\mathbf{y}) \mathrm{d}\mathbf{y}. \tag*{[7]}$$

Then $\tilde{V}$ is a smooth and convex function on $\mathbb{R}^d$. The second expression in Eq. (7) 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 $\forall \|\mathbf{x}\| < \frac{4}{3}R$,

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

And by Lemma 2,

$$\inf_{\bar{\mathbf{x}} \in \mathbb{B}\left(0, \frac{4}{3}R + \delta\right) \setminus \mathbb{B}(0, R)} \tilde{U}(\bar{\mathbf{x}}) \leq \tilde{V}(\mathbf{x}) \leq \sup_{\bar{\mathbf{x}} \in \mathbb{B}\left(0, \frac{4}{3}R + \delta\right) \setminus \mathbb{B}(0, R)} \tilde{U}(\bar{\mathbf{x}}), \quad \forall \|\mathbf{x}\| < \frac{4}{3}R. \tag*{[8]}$$

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

$$\hat{U}(\mathbf{x}) - \frac{m}{4}\left\|\mathbf{x}^2\right\| = \begin{cases} \tilde{U}(\mathbf{x}), \quad \|\mathbf{x}\| > \dfrac{3}{2}R \\ \alpha(\mathbf{x})\tilde{U}(\mathbf{x}) + (1-\alpha(\mathbf{x}))\tilde{V}(\mathbf{x}), \quad \dfrac{4}{3}R < \|\mathbf{x}\| < \dfrac{3}{2}R \\ \tilde{V}(\mathbf{x}), \quad \|\mathbf{x}\| < \dfrac{4}{3}R \end{cases}, \qquad [9]$$

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 $\mathbb{R}^d \backslash \mathbb{B}\left(0, R\right)$; $\tilde{V}(\mathbf{x})$ 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 $\frac{4}{3}R < \|\mathbf{x}\| < \frac{3}{2}R$,

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

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

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 Eq. (8), we know that for $\|\mathbf{x}\| \leq \frac{3}{2}R$,

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

Therefore,

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

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

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

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 2. *For function V defined in* Eq. (6)*, $\forall \mathbf{x} \in \mathbb{B}(0, R)$, $\inf_{\|\bar{\mathbf{x}}\|=R} \tilde{U}(\bar{\mathbf{x}}) \leq V(\mathbf{x}) \leq \sup_{\|\bar{\mathbf{x}}\|=R} \tilde{U}(\bar{\mathbf{x}})$.*

Proof of Lemma 2 First, from the definition of V inside $\mathbb{B}(0, R)$:

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

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

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

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, $\forall \mathbf{x} \in \mathbb{B}(0,R)$, $V(\mathbf{x}) \geq \inf_{\|\bar{\mathbf{x}}\|=R} \tilde{U}(\bar{\mathbf{x}})$. ■

For reader's convenience, we state the Holley-Stroock lemma in the following.

Lemma 3 (Holley-Stroock)**.** *For probability densities $p \propto e^{-U}$ and $\hat{p} \propto e^{-\hat{U}}$, assume $\hat{p}$ has log-Sobolev constant $\rho_{\hat{U}}$. Then if U is a bounded perturbation of $\hat{U}$, log-Sobolev constant ρ_U for p satisfy:*

$$\rho_U \geq \rho_{\hat{U}} e^{-\left|\sup\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)-\inf\left(\hat{U}(\mathbf{x})-U(\mathbf{x})\right)\right|}. \qquad [10]$$

*B.2. Proof of ULA Convergence Rate (*Eq. (1) *of Theorem 1).*

Proof of Eq. (1) **of Theorem 1** We first quantify the convergence of a stochastic process to a stationary distribution p^* via the Kullback-Leibler divergence (KL-divergence), $F(p)$:

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

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

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

for two densities p and p^*.

Here we take the process whose convergence is to be determined as a discretized Langevin dynamics:

$$\mathbf{X}_{(k+1)h} = \mathbf{X}_{kh} - \nabla U(\mathbf{X}_{kh})h + \sqrt{2}(B_{(k+1)h} - B_{hk}), \qquad [11]$$

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

$$\mathrm{d}\mathbf{X}_t = -\nabla U(\mathbf{X}_{kh})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t. \qquad [12]$$

For dynamics within $kh < t \leq (k+1)h$, we have from the Girsanov theorem (18) 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 (19):

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

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 established (7) 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

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

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

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

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

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

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

$$\begin{aligned}
&\frac{d}{dt}F(p_t) \\
&\leq -\frac{1}{2}\mathbb{E}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{X}_t)}{p^*(\mathbf{X}_t)}\right)\right\|^2\right] + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|^2\right] + dL^2\tau \\
&= -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_t}\left[\left\|\nabla \ln\left(\frac{p_t(\mathbf{x})}{p^*(\mathbf{x})}\right)\right\|^2\right] + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|^2\right] + dL^2\tau \\
&\leq -\rho_U F(p_t) + \frac{L^4\tau^2}{2}\mathbb{E}_{\mathbf{x}\sim p_{kh}}\left[\|\mathbf{x}\|^2\right] + dL^2\tau. \qquad [13]
\end{aligned}$$

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

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

Therefore,

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

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

$$h = \frac{\rho_U}{4L^2} \min\left\{ \frac{\epsilon^2}{d}, \sqrt{\frac{\epsilon^2}{2d\ln\frac{2L}{m} + 64\frac{L^2}{m^2}LR^2}} \right\} = \mathcal{O}\left(e^{-16LR^2}\frac{m}{L^2} \cdot \min\left\{ \frac{\epsilon^2}{d}, \frac{m}{L}\frac{\epsilon}{\sqrt{LR^2}} \right\} \right). \tag{14}$$

Therefore, combining Eq. (14) with Lemma 5, we know that whenever

$$k \geq \mathcal{O}\left(e^{32LR^2}\frac{L^2}{m^2}\ln\left(\frac{F(p_0)}{\epsilon^2}\right) \cdot \max\left\{ \frac{d}{\epsilon^2}, \frac{L}{m}\frac{\sqrt{LR^2}}{\epsilon} \right\} \right) = \mathcal{O}\left(e^{32LR^2}\frac{L^2}{m^2}\ln\left(\frac{d}{\epsilon^2}\right) \cdot \max\left\{ \frac{d}{\epsilon^2}, \frac{L}{m}\frac{\sqrt{LR^2}}{\epsilon} \right\} \right), \tag{15}$$

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

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

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* Eq. (12)*, if $\mathbb{E}_{\mathbf{x}\sim p_0}\left[\|\mathbf{x}\|_2^2\right] \leq \frac{16d}{\rho_U}\ln\frac{2L}{m} + \frac{512}{\rho_U}\frac{L^2}{m^2}LR^2$, and $h \leq \frac{1}{4}\frac{\rho_U}{L^2}$, then for all $k \in \mathbb{N}^+$,*

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

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,

$$F(p_0) = KL(p_0 \parallel p^*) = \int p_0(\mathbf{x})\ln\left(\frac{p_0(\mathbf{x})}{p^*(\mathbf{x})}\right)\mathrm{d}\mathbf{x} \leq \frac{d}{2}\ln\frac{2L}{m} + 32\frac{L^2}{m^2}LR^2; \tag{16}$$

$$\mathbb{E}_{\mathbf{x}\sim p_0}\left[\|\mathbf{x}\|_2^2\right] = \frac{d}{L}; \tag{17}$$

and

$$\mathbb{E}_{\mathbf{x}\sim p^*}\left[\|\mathbf{x}\|_2^2\right] \leq \frac{4d}{\rho_U}\ln\frac{2L}{m} + \frac{128}{\rho_U}\frac{L^2}{m^2}LR^2. \tag{18}$$

B.3. Supporting Proofs for Eq. (1) *of Theorem 1: 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 Eq. (12):

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

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 $\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right]$

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\|\mathbf{x}_t\|_2^2\right] &= \mathbb{E}_{(\mathbf{x}_t,\mathbf{x}^*)\sim\gamma}\left[\|\mathbf{x}^* + (\mathbf{x}_t - \mathbf{x}^*)\|_2^2\right] \\ &\leq 2\mathbb{E}_{\mathbf{x}^*\sim p^*}\left[\|\mathbf{x}^*\|_2^2\right] + 2\mathbb{E}_{(\mathbf{x}_t,\mathbf{x}^*)\sim\gamma}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] \\ &= \frac{8d}{\rho_U}\ln\frac{2L}{m} + \frac{256}{\rho_U}\frac{L^2}{m^2}LR^2 + 2W_2^2(p_t, p^*). \end{aligned} \tag{20}$$

Using the generalized Talagrand inequality (20) for Lipschitz smooth p^* with log-Sobolev constant ρ_U,

$$W_2^2\left(p_t, p^*\right) \leq \frac{2}{\rho_U} \mathrm{KL}(p_t \parallel p^*). \tag{21}$$

On the other hand, we know from Eq. (13) 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 Eq. (16) in Lemma 5,

$$\begin{aligned} F(p_{(k+1)h}) &\leq F(p_0) + \frac{d}{2}\ln\frac{2L}{m} + 8\frac{L^2}{m^2}LR^2 \\ &\leq d\ln\frac{2L}{m} + 40\frac{L^2}{m^2}LR^2. \end{aligned} \tag{22}$$

Plugging Eq. (22) into Eq. (20) and Eq. (21), 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 Eq. (16) **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 $\|\mathbf{y}\| = R$,

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

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,

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

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

Therefore, $U(\mathbf{x}) \geq \dfrac{m}{4} \|\mathbf{x}\|^2 - 32 \dfrac{L^2}{m^2} LR^2$ 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 Eq. (17) **of Lemma 5** It is straightforward to calculate that $\mathbb{E}_{p_0}\left[\|\mathbf{x}\|_2^2\right] = \operatorname{trace}\left(\dfrac{1}{L}\mathbb{I}\right) = \dfrac{d}{L}$.

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

Proof of Eq. (18) **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 (20) 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 Eq. (16) that

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

Therefore,

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

■

*B.4. Proof of MALA Convergence Rate (*Eq. (2) *of Theorem 1).*

Proof of Eq. (2) Our proof of Theorem 2 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 θ-warm with respect to the stationary distribution with density p^* if*

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

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 θ-warm distribution with density p^0 and error tolerance $\epsilon \in (0,1)$, satisfies*

$$\tau(\epsilon, p^0) \le \mathcal{O}\left(\frac{e^{32LR^2}}{m}\cdot\ln\left(\frac{2\theta}{\epsilon}\right)\cdot\max\left\{r\left(\frac{2\theta}{\epsilon}\right)\kappa^{3/2}d^{1/2}, \kappa d\right\}\right). \qquad [23]$$

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

Theorem 2 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 (8). 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 (21) proved its convergence

$$\left\|\mathcal{T}^k(\Pi_0) - \Pi\right\|_{\mathrm{TV}} \le \theta s + \theta\left(1 - \frac{\Phi_s^2}{2}\right)^k \le \theta s + \theta e^{-k\Phi_s^2/2} \text{ for any } s \in (0, \frac{1}{2}). \qquad [24]$$

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

$$r(s) = 2 + \sqrt{2}e^{8LR^2}\ln^{0.5}(d/s) + 7R/\sqrt{d/m}, \qquad [25]$$

and the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(0, r(s)\sqrt{\frac{d}{m}}\right). \tag{26}$$

We define the appropriate stepsize.

$$\tilde{w}(s,\gamma) = \min\left\{\frac{\sqrt{\gamma}}{8\sqrt{2}r(s)}\frac{\sqrt{m}}{L\sqrt{dL}},\quad \frac{\gamma}{96\alpha_\gamma}\frac{1}{Ld},\quad \frac{\gamma^{2/3}}{26(\alpha_\gamma r^2(s))^{1/3}}\frac{1}{L}\left(\frac{m}{Ld^2}\right)^{1/3}\right\}, \tag{27}$$

$$\text{where}\quad \alpha_\gamma = 1 + 2\sqrt{\log(16/\gamma)} + 2\log(16/\gamma). \tag{28}$$

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

$$\begin{aligned}\|\mathcal{T}_\mathbf{x} - \mathcal{T}_\mathbf{y}\|_{\mathrm{TV}} &\leq \|\mathcal{T}_\mathbf{x} - \mathcal{P}_\mathbf{x}\|_{\mathrm{TV}} + \|\mathcal{P}_\mathbf{x} - \mathcal{P}_\mathbf{y}\|_{\mathrm{TV}} + \|\mathcal{P}_\mathbf{y} - \mathcal{T}_\mathbf{y}\|_{\mathrm{TV}}\\ &\leq \frac{\sqrt{2}\gamma}{4} + \frac{\gamma}{8} + \frac{\sqrt{2}\gamma}{4}\\ &\leq \gamma. \end{aligned} \tag{29}$$

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.

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

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

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

Plugging this conductance expression into the result of Lovász and Simonovits Eq. (24), with Π_0 the distribution with density p^0 and Π the stationary distribution with density p^*, we obtain that

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

where

$$\rho_U \geq \frac{m}{2}e^{-16LR^2},\ \text{and}\quad h = \mathcal{O}\left(\min\left\{\frac{1}{L\cdot r(\frac{2\theta}{\epsilon})\kappa^{1/2}d^{1/2}}, \frac{1}{Ld}\right\}\right).$$

This concludes the proof of this lemma. ■

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 ρ_U: 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*

$$\Pi(A) \geq \rho_U \cdot d(A,B)\Pi(A)\Pi(B), \tag{30}$$

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* Eq. (30) *with constant ρ. Then for any measurable partition A_1 and A_2 of $\mathbb{R}^d$, we have*

$$\int_{A_1}\mathcal{T}_\mathbf{u}(A_2)p^*(\mathbf{u})\mathrm{d}\mathbf{u} \geq \frac{\rho}{8}\min\left\{1, \frac{\Delta\cdot(1-\gamma)\cdot\Pi^2(\mathcal{K})}{8}\right\}\min\left\{\Pi(A_1\cap\mathcal{K}), \Pi(A_2\cap\mathcal{K})\right\}. \tag{31}$$

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

$$\sup_{\substack{\mathbf{x},\mathbf{y}\in\mathbb{R}^d \\ \mathbf{x}\neq\mathbf{y}}} \frac{\left\|\mathcal{P}_{\mathbf{x}}^{MALA(h)} - \mathcal{P}_{\mathbf{y}}^{MALA(h)}\right\|_{TV}}{\|\mathbf{x}-\mathbf{y}\|_2} \leq \sqrt{\frac{2}{h}}. \tag{32a}$$

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

$$\sup_{\mathbf{x}\in\mathcal{R}_s} \left\|\mathcal{P}_{\mathbf{x}}^{MALA(h)} - \mathcal{T}_{\mathbf{x}}^{MALA(h)}\right\|_{TV} \leq \frac{\gamma}{8}, \tag{32b}$$

where the truncated ball $\mathcal{R}_s$ was defined in Eq. (26)*.*

Remark 2. *It can be seen that the constraint on the step size h originates from* Eq. (32b)*, where the difference between the proposal and transition distributions are bounded by the acceptance rate (see proof of* Eq. (32b)*). The resulting step size scaling with respect to the dimension d is $h = \tilde{w}(s,\gamma) = \mathcal{O}(d^{-1})$ under our current assumption. In a celebrated work (22), with extra assumptions on higher order smoothness and decomposability of the target distribution p^*, the log-acceptance rate was expanded to higher orders and a much better scaling of $h = \mathcal{O}(d^{-1/3})$ was obtained. It would be of great theoretical interest to understand whether such scaling can be achieved without the decomposability assumption on p^*.*

B.5. Supporting Proofs for Eq. (2) of Theorem 1.

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

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

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

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

The first inequality is because, according to Lemma 1, we have

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

■

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 1. 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 (23)).

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*

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

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

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

156 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 $g : \mathbb{R}^d \to \mathbb{R}$.

$$2\rho_U \left[\int_{\mathbb{R}^d} g\ln g \mathrm{d}\Pi - \int_{\mathbb{R}^d} g\mathrm{d}\Pi \cdot \ln\left(\int_{\mathbb{R}^d} g\mathrm{d}\Pi\right)\right] \leq \int_{\mathbb{R}^d} \frac{\|\nabla g\|_2^2}{g} \mathrm{d}\Pi. \quad [33]$$

where

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

Replacing g with g^2 in Eq. (33), 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. (24)). For any smooth $g : \mathbb{R}^d \to \mathbb{R}$, we have

$$\rho_U \mathrm{Var}_{p^*}(g) \leq \int_{\mathbb{R}^d} \|\nabla g\|_2^2 \mathrm{d}\Pi, \quad [34]$$

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, we show that the isoperimetric constant can by lower bounded by the Poincaré constant. We denote Ψ the isoperimetric constant defined as

$$\Psi = \sup_{A\subset\mathbb{R}^d,\text{ open}} \frac{\Pi^+(\partial A)}{\min \Pi(A), 1 - \Pi(A)}, \quad [35]$$

where $\Pi^+(\partial A) = \lim_{\delta \to 0} \frac{\Pi(A+\delta)-\Pi(A)}{\delta}$. Taking a sequence of smooth $\{g_k\}_{k=1,\ldots\infty}$ with limit the indicator function of A in equation Eq. (34), we obtain[†]

$$\Psi \geq \rho_U.$$

Finally, it is easy to show that the infinitesimal version of the isoperimetric inequality in Eq. (35) is equivalent to the partition version (see e.g. (26) Proposition 11.1 and (27)). Let A and B be open disjoint subsets of $\mathbb{R}^d$, $C = \mathbb{R}^d \setminus (A \cup B)$, then

$$\Pi(C) \geq \rho_U \cdot d(A,B)\Pi(A)\Pi(B). \qquad [36]$$

$\blacksquare$

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 (8). 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 Eq. (36). $\blacksquare$

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 $\mathcal{P}_{\mathbf{x}}^{\mathrm{MALA}(h)}$. $\blacksquare$

Proof of Eq. (32a) We first apply the Pinsker inequality (28) to bound the total variation distance via KL-divergence.

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

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

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

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

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

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. $\blacksquare$

Proof of Eq. (32b) We denote $p_{\mathbf{x}}$ the density corresponding to the proposal distribution $\mathcal{P}_{\mathbf{x}} = \mathcal{N}(\mathbf{x} - h\nabla U(\mathbf{x}), 2h\mathbb{I}_d)$. We have

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

[†] Note that Buser's inequality (25) (Theorem 1.2), which would give $h \geq (\rho_U/10)^{1/2}$, does not directly apply here because of the possible negative curvature.

Applying Markov inequality, we know that

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

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

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

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

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

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

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

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

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

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

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

■

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:

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

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 2 (Lower bound for optimization)**.** *For any $R > 0$, $L \geq 2m > 0$, probability $0 < p \leq 1$, and $\epsilon \leq \dfrac{LR^2}{64\,(2\pi^2 + \pi)}$, there exists an objective function U satisfying the local nonconvexity Assumptions 1–3 with constants L, m, and R, such that any algorithm in $\mathcal{A}_\infty$ requires at least $\left\lfloor \left(\frac{R}{4}\sqrt{\frac{L}{2\pi^2+\pi}} \cdot \frac{1}{\sqrt{\epsilon}} - \frac{1}{2}\right)^d \right\rfloor = \Omega\left(p \cdot \left(LR^2/\epsilon\right)^{d/2}\right)$ 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 2. We constructively prove Theorem 2 by defining such a $U(\mathbf{x})$ in what follows. We first make use of the following lemma about packing numbers. Again we denote $\mathbb{B}(0, R)$ as the closed ball of radius R centered at 0.

Lemma 14 (Packing number)**.** *For $R > r > 0$, denote $\eta = \left\lfloor \left(\frac{R-r}{2r}\right)^d \right\rfloor$. 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. S1, 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 $\epsilon \geq \dfrac{LR^2}{36(2\pi^2 + \pi)}$, then

$$T \geq 1 \geq p \cdot \left\lfloor \left(\frac{R}{4}\sqrt{\frac{L}{2\pi^2+\pi}} \cdot \frac{1}{\sqrt{\epsilon}} - \frac{1}{2}\right)^d \right\rfloor, \quad \forall 0 < p \leq 1.$$

Otherwise, take $r = \sqrt{(2\pi^2 + \pi)\epsilon/L}$ in Lemma 14. Then we have the r-packing number inside $\mathbb{B}(0, R/2)$ to be

$$\eta = \left\lfloor \left(\frac{R/2 - r}{2r}\right)^d \right\rfloor = \left\lfloor \left(\frac{R}{4}\sqrt{\frac{L}{2\pi^2+\pi}} \cdot \frac{1}{\sqrt{\epsilon}} - \frac{1}{2}\right)^d \right\rfloor \geq 1,$$

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

$$U(\mathbf{x}) = \begin{cases} \dfrac{Lr^2}{4\pi^2 + 2\pi} \cos\left(\dfrac{\pi}{r^2}\left(||\mathbf{x} - \mathbf{x}_{i^*}||_2^2 - r^2\right)\right) - \dfrac{Lr^2}{4\pi^2 + 2\pi}, & ||\mathbf{x} - \mathbf{x}_{i^*}||_2 \leq r \\ 0, \quad ||\mathbf{x} - \mathbf{x}_{i^*}||_2 > r, ||\mathbf{x}||_2 \leq R/2 \\ m\left(||\mathbf{x}||_2 - R/2\right)^2, ||\mathbf{x}||_2 > R/2. \end{cases} \qquad [37]$$

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 $\{U(\mathbf{x}), \nabla U(\mathbf{x}), \cdots, \nabla^n U(\mathbf{x})\}, \forall n \in \mathcal{N}$, $\forall\ \epsilon < \dfrac{LR^2}{36(2\pi^2 + \pi)}$, 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 $\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r)$:

$$\begin{aligned} P(|U(\mathbf{x}^t) - U(\mathbf{x}^*)| < \epsilon) &\leq P(\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r)) \qquad [38] \\ &\leq P\left(\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_{i^*}, r) \,\middle|\, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right). \end{aligned}$$

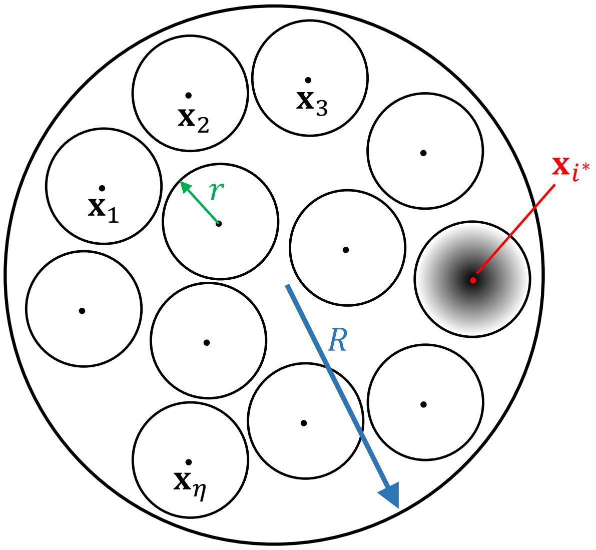

Fig. S1. A depiction of $U(\mathbf{x})$ inside $\mathbb{B}(0, R)$.

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

1. Assume that $\forall t \leq T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$. From the definition of $U(\mathbf{x})$, Eq. (37), we know that $\forall j \in \{1, \cdots, \eta\}, j \neq i^*$, $\forall \mathbf{x} \in \mathbb{B}(\mathbf{x}_j, r)$, $U(\mathbf{x}) = 0$, $\nabla U(\mathbf{x}) = 0$, $\cdots$, $\nabla^n U(\mathbf{x}) = 0$. Hence $\mathbf{x}^t \in \mathbb{B}(\mathbf{x}_j, r)$, $j \neq i^*$ only contains information that $i^* \in \{1, \cdots \eta\} \setminus \{j\}$. Since i is chosen uniformly at random from $\{1, \cdots, \eta\}$, for $T \leq \eta$

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

   Therefore,

$$\begin{aligned} &P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) = \emptyset \,\middle|\, \forall t \leq T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\geq \frac{\eta-1}{\eta}\frac{\eta-2}{\eta-1}\cdots\frac{\eta-T}{\eta-(T-1)} = \frac{\eta-T}{\eta}. \end{aligned} \qquad [39]$$

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

$$\begin{aligned} &P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) \neq \emptyset \,\middle|\, \forall t \leq T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &= 1 - P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) = \emptyset \,\middle|\, \forall t \leq T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\leq 1 - \frac{\eta - T}{\eta} = \frac{T}{\eta}. \end{aligned} \qquad [40]$$

   Therefore,

$$\begin{aligned} p &\leq P(|U(\mathbf{x}^T) - U(\mathbf{x}^*)| < \epsilon) \\ &\leq P\left(\mathbf{x}^T \in \mathbb{B}(\mathbf{x}_{i^*}, r) \,\middle|\, \mathbf{x}^T \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\leq P\left(\{\mathbf{x}^1, \cdots, \mathbf{x}^T\} \bigcap \mathbb{B}(\mathbf{x}_{i^*}, r) \neq \emptyset \,\middle|\, \forall t \leq T, \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)\right) \\ &\leq \frac{T}{\eta}, \end{aligned} \qquad [41]$$

$$T \geq p \cdot \eta.$$

2. Suppose there exists an algorithm that output $\{\mathbf{x}_1, \cdots, \mathbf{x}^T\}$, where $\exists\ t \leq T,\ \mathbf{x}^t \notin \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$ 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 $\forall\ t \leq T,\ \mathbf{x}^t \in \bigcup_{j=1}^{\eta} \mathbb{B}(\mathbf{x}_j, r)$, 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.2. Supporting Proofs for Theorem 2.

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 15 (Lipschitz smoothness and strong convexity) We first prove that when $||\mathbf{x}||_2 \leq R/2$, $U(\mathbf{x})$ is L-Lipschitz smooth. We then prove that when $||\mathbf{x}||_2 > R/2$, $U(\mathbf{x})$ is $2m$-Lipschitz smooth. At last we prove that $U(\mathbf{x})$ is m-strongly convex for $||\mathbf{x}||_2 > R$. 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 $U_2(\mathbf{x}) = (||\mathbf{x}||_2 - R/2)^2$. Then $U(\mathbf{x}) = mU_2(\mathbf{x})$ when $||\mathbf{x}||_2 > R/2$.

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

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

- Define

$$U_3(\mathbf{x}) = \begin{cases} U_2(\mathbf{x}), \ \|\mathbf{x}\|_2 > R \\ \frac{1}{2}\|\mathbf{x}\|_2^2 + \frac{1}{8}R^2, \ \|\mathbf{x}\|_2 \leq R \end{cases}.$$

  Then

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

  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 $\mathbf{x}\mathbf{x}^{\mathrm{T}}$, $H\left[U_3(\mathbf{x}) - \frac{1}{2}\|x\|_2^2\right] \succeq \left(1 - \frac{R}{||\mathbf{x}||_2}\right)\mathbb{I} \succeq 0$ when $||\mathbf{x}||_2 > R$. When $||\mathbf{x}||_2 \leq R$, $H\left[U_3(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_2^2\right] = 0$. 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 $||\mathbf{x}||_2 > R$. Following Assumption 2, this implies that $U(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex on $\mathbb{R}^d \setminus \mathbb{B}(0, R)$. Therefore, $U(\mathbf{x})$ is m-strongly convex on $\mathbb{R}^d \setminus \mathbb{B}(0, R)$.

■

C.3. Proof of Corollary 1.

Corollary 1. *There exists an objective function U that is m-strongly convex outside of a region of radius R 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 1 We take U as the one defined in Eq. (37) 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 $\min_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x}) = -\dfrac{Lr^2}{2\pi^2 + \pi} = -\epsilon$. 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 $\hat{\mathbf{x}} \in \mathbb{B}\left(\mathbf{x}^*, r\right)$

$$\begin{aligned}
P\left(\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq r\right) &= \Pi_\beta^*\Big(\mathbb{B}\left(\mathbf{x}^*, r\right)\Big) \\
&= \frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x} + \int_{\mathbb{B}(0,R/2)\backslash\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x} + \int_{\mathbb{R}/\nvDash^d\backslash\mathbb{B}(0,R/2)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
&= \frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x} + \int_{\mathbb{B}(0,R/2)\backslash\mathbb{B}(\mathbf{x}^*,r)} 1 \ \mathrm{d}\mathbf{x} + \int_{\mathbb{R}/\nvDash^d\backslash\mathbb{B}(0,R/2)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
&= \frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(\mathbf{x}^*,r)} \left(e^{-\beta U(\mathbf{x})} - 1\right)\mathrm{d}\mathbf{x} + \int_{\mathbb{B}(0,R/2)} 1 \ \mathrm{d}\mathbf{x} + \int_{\mathbb{R}/\nvDash^d\backslash\mathbb{B}(0,R/2)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}} \\
&\leq \frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} e^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R/2)} 1 \ \mathrm{d}\mathbf{x}} \\
&\leq \frac{e^{-\min_{\|\mathbf{x}-\mathbf{x}^*\|\leq r} \beta U(\mathbf{x})} \int_{\mathbb{B}(\mathbf{x}^*,r)} 1 \ \mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R/2)} 1 \ \mathrm{d}\mathbf{x}} \\
&= e^{\beta\epsilon} \frac{\int_{\mathbb{B}(\mathbf{x}^*,r)} 1 \ \mathrm{d}\mathbf{x}}{\int_{\mathbb{B}(0,R/2)} 1 \ \mathrm{d}\mathbf{x}} \\
&= e^{\beta\epsilon}\left(\frac{2r}{R}\right)^d.
\end{aligned} \qquad [42]$$

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

$$e^{\beta\epsilon}\left(\frac{2r}{R}\right)^d \geq p.$$

Therefore,

$$\beta \geq \frac{1}{\epsilon}\ln p + \frac{d}{\epsilon}\ln\left(\frac{R}{2r}\right) = \frac{1}{\epsilon}\ln p + \frac{1}{2}\frac{d}{\epsilon}\ln\left(\frac{1}{4(2\pi^2 + \pi)}\frac{LR^2}{\epsilon}\right).$$

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

$$\begin{aligned}
&\left|P\left(\left\|\mathbf{x}^K-\mathbf{x}^*\right\|\le r\right)-P\left(\|\hat{\mathbf{x}}-\mathbf{x}^*\|\le r\right)\right|\\
&=\left|\Pi_\beta^K\big(\mathbb{B}\left(\mathbf{x}^*,r\right)\big)-\Pi_\beta^*\big(\mathbb{B}\left(\mathbf{x}^*,r\right)\big)\right|\\
&\le\sup_A\left|\Pi_\beta^K(A)-\Pi_\beta^*(A)\right|\\
&=\left\|q_\beta^K-q_\beta^*\right\|_{\mathrm{TV}}\\
&\le\delta,
\end{aligned}\tag{43}$$

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

We directly obtain from Theorem 1 that for the objective function βU with Lipschitz constant $\beta L\ge\dfrac{L}{\epsilon}\ln p+\dfrac{d}{2}\dfrac{L}{\epsilon}\ln\left(\dfrac{1}{4(2\pi^2+\pi)}\dfrac{LR^2}{\epsilon}\right)$, we need to iterate $e^{\widetilde{\mathcal{O}}\left(d\cdot LR^2/\epsilon\right)}$ steps to guarantee convergence.

■

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 $\mathbf{y}=(y_1,\cdots,y_N)$:

$$p\left(y_n|\boldsymbol{\mu}\right)=\sum_{i=1}^M\frac{\lambda_i}{Z_i}\exp\left(-\frac{1}{2}(y_n-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(y_n-\mu_i)\right)+\left(1-\sum_{i=1}^M\lambda_i\right)p_0(y_n),\tag{44}$$

where Z_i are the normalization constants and $\sum_{i=1}^M\lambda_i\le1$. 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 ($\|y_n\|\le R$) and take $p_0(y_n)=\mathbb{1}\left\{\|y_n\|\le R\right\}/Z_0$ to describe this observation.

We also take the prior to be

$$p(\boldsymbol{\mu})\propto\exp\left(-m\left(\|\boldsymbol{\mu}\|_F-\sqrt{M}R\right)^2\mathbb{1}\left\{\|\boldsymbol{\mu}\|_F\ge\sqrt{M}R\right\}\right).\tag{45}$$

D.1. Proofs for Smoothness.

Fact 1. *For the Gaussian mixture model defined in* Eq. (44)*, define*

$$\alpha=\frac{1}{\sigma^2}\max\left\{2\sup_{\mu\in\{\mu_1,\cdots,\mu_M\}}\sum_{n=1}^N\frac{\|\mu-y_n\|^2}{\sigma^2}\exp\left(-||\mu-y_n||^2/2\sigma^2\right),\sup_{\mu\in\{\mu_1,\cdots,\mu_M\}}\sum_{n=1}^N\exp\left(-||\mu-y_n||^2/2\sigma^2\right)\right\}.\tag{46}$$

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: $W_{i,n}=\dfrac{\lambda_i}{Z_i}\exp\left(-\dfrac{1}{2}||y_n-\mu_i||^2/\sigma^2\right)$ 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 $\mathbb{B}(0,R)$, $p_0(y_n)=\dfrac{1}{Z_0}\mathbb{1}\left\{\|y_n\|\le R\right\}=\dfrac{1}{Z_0}$. We can plug in the expression of

$\lambda_i = \dfrac{\dfrac{l}{\alpha} Z_i}{Z_0 + \dfrac{l}{\alpha}\sum_{j=1}^M Z_j}$ and obtain for any $n = 1, \cdots, N$:

$$\begin{aligned} C_n = C &= \frac{1}{Z_0}\left(1 - \sum_{i=1}^M \lambda_i\right) \\ &= \frac{1}{Z_0}\left(1 - \frac{\dfrac{l}{\alpha}\sum_{i=1}^M Z_i}{Z_0 + \dfrac{l}{\alpha}\sum_{j=1}^M Z_j}\right) \\ &= \frac{1}{Z_0 + \dfrac{l}{\alpha}\sum_{j=1}^M Z_j}. \end{aligned} \qquad [47]$$

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

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

For any vector $\mathbf{v}$,

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

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,

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

Since

$$\alpha=\frac{1}{\sigma^2}\max\left\{2\sup_{\mu}\sum_{n=1}^{N}\frac{\|\mu-y_n\|^2}{\sigma^2}\exp\left(-\|\mu-y_n\|^2/2\sigma^2\right),\sup_{\mu}\sum_{n=1}^{N}\exp\left(-\|\mu-y_n\|^2/2\sigma^2\right)\right\}. \tag*{[48]}$$

and

$$\lambda_i=\frac{l}{\alpha}CZ_i, \tag*{[49]}$$

log-likelihood $-\sum_{n=1}^{N}\log p\left(y_n|\boldsymbol{\mu}\right)$ is l-Lipschitz smooth. It can be seen from Eq. (48) that if one uses a loose upper bound for α, we can simply take λ_i to be $\dfrac{l}{2}\dfrac{CZ_i\sigma^2}{N}$. ■

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 $K\ge\min\{\mathcal{O}(d^{1/\epsilon}),\mathcal{O}(d^d)\}$ 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* Eq. (45) *and* Eq. (44). *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 $K\ge\min\{\mathcal{O}(d^{1/\epsilon}),\mathcal{O}(d^d)\}$ 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 (29, 30).

Directly invoking Theorem 1, we know that the Langevin algorithms converge within $K\le\widetilde{\mathcal{O}}\left(d^3/\epsilon\right)$ and $K\le\widetilde{\mathcal{O}}\left(d^3\ln^2\left(1/\epsilon\right)\right)$ 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 Eq. (44).

For the first $N-9M$ points, let $\|y_n\|\le0.45$, and $\|y_k-y_l\|\ge0.11$, where $n,k,l\in\{1,\cdots,N-9M\}$ and $k\ne l$. From Lemma 14, we know that when $N\le2^d$, 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 $n\in\{N-9M+9(k-1)+1,\cdots,N-9M+9k\}$ $(k\in\{1,\cdots,M\})$, $\|y_n-y_{i_k}\|\le\sigma/2$.

By this setting, $\forall y_n$, $\|y_n\|\le0.5$. Furthermore, when $n,\hat{n}\in\{N-9M+9(k-1)+1,\cdots,N-9M+9k\}\cup\{i_k\}$, $\|y_n-y_{\hat{n}}\|\le\sigma/2$; otherwise, $\|y_n-y_{\hat{n}}\|\ge0.1$ for $n\ne\hat{n}$. We depict a cartoon of this dataset in Fig. S2.

Since it can be observed that all the data are distributed in $\mathbb{B}(0,0.5)$, we let $p_0(y_n)=\dfrac{1}{Z_0}\mathbb{1}\left\{\|y_n\|\le0.5\right\}=\dfrac{\Gamma(d/2+1)}{(2\pi)^{d/2}}\mathbb{1}\left\{\|y_n\|\le0.5\right\}$. 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 Eq. (45), we set the prior to be:

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

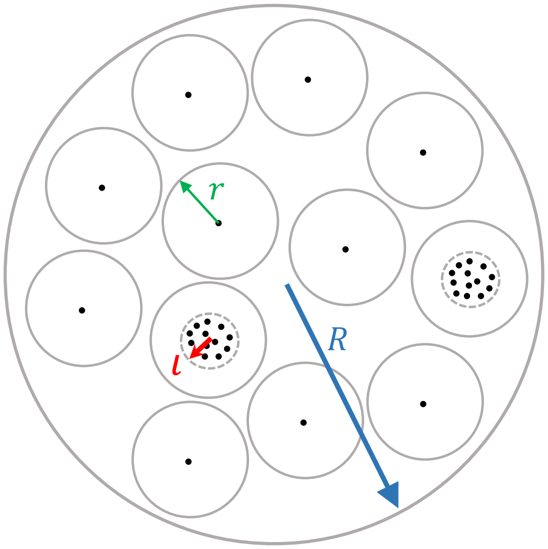

Fig. S2. A depiction of an example dataset.

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 $\lambda_i = \frac{1}{64\alpha} C Z_i$ (using notations defined in Eq. (49) and Eq. (47)). Then the objective function defined via the log posterior:

$$\begin{aligned}
U(\boldsymbol{\mu}) &= -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log p\left(y_n|\boldsymbol{\mu}\right) \\
&= -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N} \log\left(\sum_{i=1}^{M} \frac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right) + C\right) \\
&= \frac{\left(\|\boldsymbol{\mu}\|_F - \sqrt{M}/2\right)^2}{64} \mathbb{1}\left\{\|\boldsymbol{\mu}\|_F \geq \sqrt{M}/2\right\} \\
&\quad - \sum_{n=1}^{N} \log\left(\sum_{i=1}^{M} \frac{1}{64\alpha} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right) + 1\right) + \widetilde{C}
\end{aligned} \tag{50}$$

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

It can be seen that α in Eq. (46) is bounded as: $\alpha \leq \frac{50}{\sigma^2}$. Then $\lambda_i = \frac{1}{3200} C Z_i \sigma^2$. 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):

$$\begin{aligned}
\gamma_{i,n}^{(t)} &= \frac{\frac{\lambda_i}{Z_i} \exp\left(-\frac{1}{2}||y_n - \mu_i^{(t)}||^2/\sigma^2\right)}{\sum_{j=1}^{M} \frac{\lambda_j}{Z_j} \exp\left(-\frac{1}{2}||y_n - \mu_j^{(t)}||^2/\sigma^2\right) + C} \\
&= \frac{\frac{\sigma^2}{3200} \exp\left(-\frac{1}{2}||y_n - \mu_i^{(t)}||^2/\sigma^2\right)}{\sum_{j=1}^{M} \frac{\sigma^2}{3200} \exp\left(-\frac{1}{2}||y_n - \mu_j^{(t)}||^2/\sigma^2\right) + 1}.
\end{aligned} \tag{51}$$

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 $\|\mu_i^{(0)} - y_{n_i}\| \leq 0.01$, then $\|\mu_i^{(\tau)} - y_{n_i}\| \leq 0.01$, $\forall \tau > 0$. 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 $\|\mu_i^{(0)} - y_{n_i}\| \leq 0.01$ for $n_i \in \{1, \cdots, N/2\}$, then $\|\mu_i^{(\tau)} - y_{n_i}\| \leq 0.01$, $\forall \tau > 0$.*

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 Eq. (50) 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 $K = p \cdot \left(\begin{array}{c} N \\ M \end{array}\right) \Big/ \left(\begin{array}{c} 10M \\ M \end{array}\right) \geq p \cdot \left(\frac{N}{10M}\right)^M$ 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 $T = \min\left\{\mathcal{O}\left(d^{1/\epsilon}\right), \mathcal{O}(d^d)\right\}$. ■

Remark 3. *It can be similarly proven that the gradient descent algorithm with its stepsize tuned according to the Lipschitz smoothness has the same behavior if initialized randomly from the dataset.*

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 Eq. (51) that $\forall i, n$,

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

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

$$\frac{\sigma^2}{6400} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right) \le \gamma_{i,n}^{(t)} \le \frac{\sigma^2}{3200} \exp\left(-\frac{1}{2}||y_n - \mu_i||^2/\sigma^2\right). \tag{52}$$

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

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

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

$$\frac{\sum_{\tilde{n} \ne n_i} \gamma_{i,\tilde{n}}^{(t)}}{\sum_{\hat{n}=1}^{N} \gamma_{i,\hat{n}}^{(t)}} \le \frac{\sum_{\tilde{n} \ne n_i} \gamma_{i,\tilde{n}}^{(t)}}{\gamma_{i,n_i}^{(t)}} \le \frac{N \cdot \frac{\sigma^2}{3200} N^{-50}}{\frac{\sigma^2}{6400} N^{-1/2}} \le 10^{-10}. \tag{53}$$

Hence

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

- 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 $\sup\limits_{\tilde{n}\in\Omega_k}\|y_{\tilde{n}} - y_{i_k}\| \le \sigma/2$, $\|y_{i_k}^{avg} - y_{i_k}\| \le \sigma/2$. And for any $\tilde{n}\in\Omega_k$, we use induction assumption and $\sup\limits_{\tilde{n}\in\Omega_k}\|y_{\tilde{n}} - y_{i_k}\| \le \sigma/2$ 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 $\gamma_{i,\tilde{n}}^{(t)} \ge \dfrac{\sigma^2}{4N}N^{-9/8}$. Then similar to Eq. (53),

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

Therefore,

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

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

E. Detailed Experimental Settings for Gaussian Mixture Models. We consider the same problem as that in Supplement D 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 points $\mathbf{y} = (y_1, \cdots, y_N)$:

$$p\left(y_n|\boldsymbol{\mu}\right) = \sum_{i=1}^{M}\frac{\lambda_i}{Z_i}\exp\left(-\frac{1}{2}(y_n - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(y_n - \mu_i)\right) + \left(1 - \sum_{i=1}^{M}\lambda_i\right)p_0(y_n), \tag{54}$$

where the covariances Σ_i are isotropic and uniform across all mixture components: $\Sigma_i = \Sigma = \sigma^2\mathbb{I}$. The constant mixture $p_0(y_n) = \mathbb{1}\left\{\|y_n\| \le R\right\}/Z_0$ represents crude observations of the data, which are distributed in a bounded region: $\|y_n\| \le R$. The objective function is given by the log posterior distribution: $U(\boldsymbol{\mu}) = -\log p(\boldsymbol{\mu}) - \sum_{n=1}^{N}\log p\left(y_n|\boldsymbol{\mu}\right)$, where we take the prior to be

$$p(\boldsymbol{\mu}) \propto \exp\left(-m\left(\|\boldsymbol{\mu}\|_F - \sqrt{M}R\right)^2 \mathbb{1}\left\{\|\boldsymbol{\mu}\|_F \ge \sqrt{M}R\right\}\right). \tag{55}$$

Similar to the setting in Supplement D.2, we take $\dfrac{\lambda_i}{Z_i}$ to be $\sigma^2/1000$, where the variance $\sigma = 1/\sqrt{d}$, so that the mixtures are well separated from each other.

We consider a synthetic dataset, $\{y_1, \cdots, y_N\}$, with sparse entries: only $\lfloor\log_2 d\rfloor$ of the entries in each data point y_n are nonzero. Indices of the nonzero entries are uniformly distributed over the set $\{1, \cdots, d\}$. All the nonzero entries follow a uniform distribution on $[-1, 1]$. Also assume that the number of mixtures $M = \lfloor\log_2 d\rfloor$. Hence the radius containing the data $R = 2\sqrt{M\lfloor\log_2 d\rfloor} = 2\lfloor\log_2 d\rfloor$. We generate $N = 2^d$ data points following this rule.

We let the dimension d range from 2 to 32 and recorded the number of gradient entries required for EM (with random initialization from the data) and ULA to converge. The results were averaged over 20 trials of experiments. When

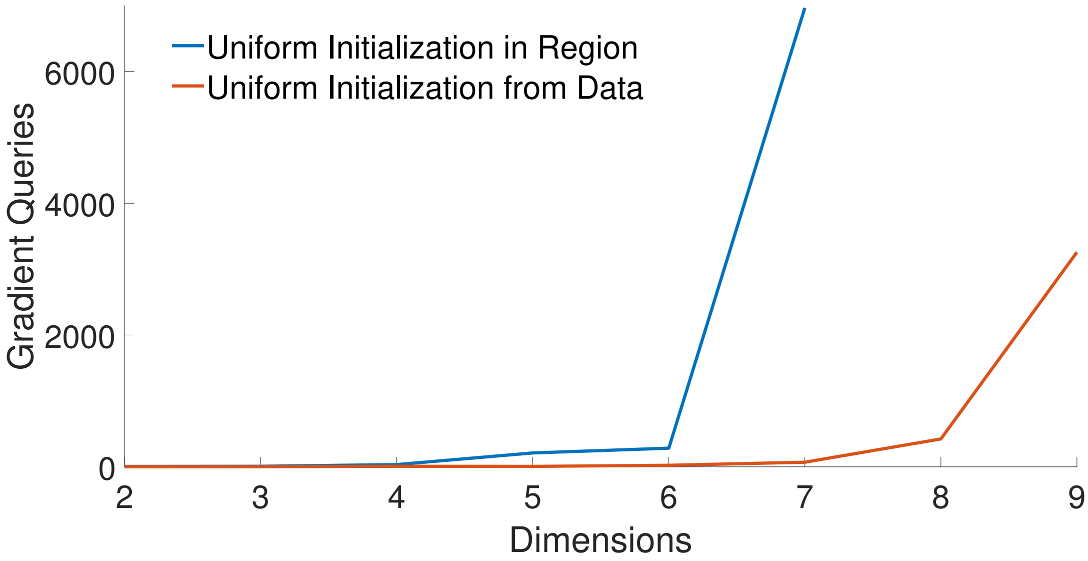

Fig. S3. Experimental results: scaling of the number of gradient queries required for EM with random initialization uniformly in the ball of radius R and uniformly from the data.

dimension $d \geq 10$, too many gradient queries are required for EM to converge, so that an accurate estimate of convergence time is not available.

For EM, we measured its accuracy in terms of the objective function value U and require $U(\boldsymbol{\mu}_K) - U(\widehat{\boldsymbol{\mu}^*}) < 10^{-6}$ to conclude that $\boldsymbol{\mu}_K$ has converged close enough to $\boldsymbol{\mu}^*$. For ULA, we measured its accuracy in terms of both the expected objective function value $\mathbb{E}\left[U(\boldsymbol{\mu})\right]$ (or equivalently the cross entropy between the sampled distribution and the posterior) and the expected mean parameters $\mathbb{E}\left[\boldsymbol{\mu}\right]$. We required both $\left|\frac{1}{K}\sum_{k=1}^{K} U(\boldsymbol{\mu}_k) - \widehat{\mathbb{E}_{p^*}\left[U(\boldsymbol{\mu})\right]}\right| < 10^{-6}$ and $\left\|\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_k - \widehat{\mathbb{E}_{p^*}\left[\boldsymbol{\mu}\right]}\right\|_F < 10^{-3}$ (which are of comparable scales) to assess the convergence of the sampling algorithm.

To estimate the reference value $\widehat{\boldsymbol{\mu}^*} \in \mathbb{R}^{d \times M}$, we run EM 1000 times longer than the number of required steps found for the previous experiment with dimension $d-1$. If estimates from 20 different initializations differed by less than 10^{-8}, we accepted $\widehat{\boldsymbol{\mu}^*}$. Otherwise, we increased the number of steps by 10 times. We similarly estimated $\widehat{\mathbb{E}_{p^*}\left[U(\boldsymbol{\mu})\right]}$ and $\widehat{\mathbb{E}_{p^*}\left[\boldsymbol{\mu}\right]}$ by long runs of ULA (also 1000 times longer than the number of required steps found for dimension $d-1$). If estimates from 20 different initializations differed by less than 10^{-8} for $\widehat{\mathbb{E}_{p^*}\left[U(\boldsymbol{\mu})\right]}$ and 10^{-5} for $\widehat{\mathbb{E}_{p^*}\left[\boldsymbol{\mu}\right]}$, we accepted the estimates. Otherwise, we increased the number of steps by 10 times.

We also compared EM with random initialization uniformly in the ball of radius R against that with uniform initialization from the data points. We observed in Fig. S3 that initializing uniformly in the ball of radius R leads to poorer convergence, implying that there are more local minima of U than merely those nearby the data.

References

1. Peters HJM, Wakker PP (1986) Convex functions on non-convex domains. *Econ Lett* 22(2):251–255.
2. Yan M (2014) Extension of convex function. *J Convex Anal* 21(4):965–987.
3. Dalalyan AS (2017) Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J Royal Stat Soc B* 79(3):651–676.
4. Durmus A, Moulines E (2016) Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. arXiv:1605.01559.
5. Dalalyan AS, Karagulyan AG (2017) User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. arXiv:1710.00095.
6. Cheng X, Chatterji NS, Bartlett PL, Jordan MI (2018) Underdamped Langevin MCMC: A non-asymptotic analysis in *Proceedings of the 31st Conference on Learning Theory (COLT)*. pp. 300–323.
7. Cheng X, Bartlett PL (2018) Convergence of Langevin MCMC in KL-divergence in *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*. pp. 186–211.
8. Dwivedi R, Chen Y, Wainwright MJ, Yu B (2018) Log-concave sampling: Metropolis-Hastings algorithms are fast! arXiv:1801.02309.
9. Mangoubi O, Smith A (2017) Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. arXiv:1708.07114.
10. Mangoubi O, Vishnoi NK (2018) Dimensionally tight running time bounds for second-order Hamiltonian Monte Carlo. arXiv:1802.08898.
11. Eberle A, Guillin A, Zimmer R (2017) Couplings and quantitative contraction rates for Langevin dynamics. arXiv:1703.01617.
12. Bou-Rabee N, Eberle A, Zimmer R (2018) Coupling and convergence for Hamiltonian Monte Carlo. arXiv:1805.00452.
13. Cheng X, Chatterji NS, Abbasi-Yadkori Y, Bartlett PL, Jordan MI (2018) Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv:1805.01648.
14. Majka MB, Mijatović A, Szpruch L (2018) Non-asymptotic bounds for sampling algorithms without log-concavity. arXiv:1808.07105.
15. Bakry D, Emery M (1985) Diffusions hypercontractives in *Séminaire de Probabilités XIX 1983/84*. pp. 177–206.
16. Holley R, Stroock D (1987) Logarithmic Sobolev inequalities and stochastic Ising models. *J Stat Phys* 46(5):1159–1194.
17. Uhlmann A (2010) Roofs and convexity. *Entropy* 12:1799–1832.
18. Øksendal B (2003) *Stochastic Differential Equations.* (Springer, Berlin), 6th edition.
19. Pavliotis GA (2014) *Stochastic Processes and Applications.* (Springer, New York).

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

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

22. Roberts GO, Rosenthal JS (1998) Optimal scaling of discrete approximations to Langevin diffusions. *J Royal Stat Soc B* 60(1):255–268.

23. Ledoux M (1999) Concentration of measure and logarithmic Sobolev inequalities in *Séminaire de Probabilités XXXIII.* pp. 120–216.

24. Bobkov SG, Tetali P (2006) Modified logarithmic Sobolev inequalities in discrete settings. *J Theor Probab* 19(2):289–336.

25. Buser P (1982) A note on the isoperimetric constant. *Ann Sci École Normale Sup* 15(2):213–230.

26. Bobkov SG, Zegarlinski B (2005) *Entropy Bounds and Isoperimetry.* (American Mathematical Soc) No. 829.

27. Bobkov SG (2007) *On Isoperimetric Constants for Log-Concave Probability Distributions*, eds. Milman VD, Schechtman G. (Springer, Berlin Heidelberg), pp. 81–88.

28. Cover TM, Thomas JA (2012) *Elements of Information Theory.* (Wiley, New York).

29. Jin C, Zhang Y, Balakrishnan S, Wainwright MJ, Jordan MI (2016) Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences in *Advances in Neural Information Processing Systems (NIPS) 29.* pp. 4116–4124.

30. Améndola C, Drton M, Sturmfels B (2015) Maximum likelihood estimates for Gaussian mixtures are transcendental in *International Conference on Mathematical Aspects of Computer and Information Sciences.* pp. 579–590.