CS395T: Continuous Algorithms, Part XVII
Langevin algorithms

Kevin Tian

1 W_2^2 convergence of unadjusted Langevin

In Part XVI, we gave tools for understanding the convergence of the Langevin dynamics,

$$\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t. \tag{1}$$

For instance, we gave a simple coupling argument showing that when the target stationary distribution $\pi^\star \propto \exp(-V)$ (Theorem 1, Part XVI) is strongly logconcave, then the Langevin dynamics converge linearly in W_2^2 (Theorem 2, Part XVI). Moreover, using tools from Markov semigroup theory, we established that when the stationary distribution satisfies weaker functional inequalities such as Poincaré or log-Sobolev, the Langevin dynamics (1) actually converge under stronger error metrics such as χ^2 or D_{KL}. Unfortunately, these results do not immediately lead to implementable algorithms, because they only hold in continuous time.

Our goal in this lecture is now to give an introduction to convergence guarantees for discrete-time approximate implementations of the Langevin dynamics. In this and the following section, we will specifically focus on the *unadjusted Langevin algorithm* (ULA), which samples $\mathbf{x}_0$ from a starting distribution π_0, and for a step size $\eta > 0$, iterates[1]

$$\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} - \eta\nabla V(\mathbf{x}^{(k)}) + \sqrt{2\eta}\boldsymbol{\xi}^{(k)}, \text{ where } \boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d). \tag{2}$$

The motivation for considering (2), a forward Euler discretization of the Langevin dynamics, is that it only requires one query to ∇V, as opposed to running (1) which would require an unbounded number of queries. This is entirely analogous to the relationship between gradient descent (a discrete-time algorithm) and its continuous-time counterpart, gradient flow.

It is straightforward to check that the iteration (2) is equivalently induced by the SDE

$$\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_0)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t \tag{3}$$

up to time $t = \eta$, initialized at $\mathbf{x}_0 \leftarrow \mathbf{x}^{(k)}$. In other words, rather than the position-dependent drift $\nabla V(\mathbf{x}_t)$ typically used in the Langevin dynamics, ULA uses a constant drift $\nabla V(\mathbf{x}_0)$. In this sense, the (discrete-time) ULA is simply an Euler discretization of the (continuous-time) Langevin dynamics, just as gradient descent is an Euler discretization of gradient flow (Part II).

Our strategy for analyzing the convergence of (2) under strong logconcavity, when the error metric is W_2^2, is then fairly straightforward. We first use rapid convergence of the continuous-time Langevin dynamics as in Theorem 2, Part XVI, and then bound the discretization error through a coupling argument. We introduce two standard helper claims which help in our analysis.

Lemma 1. *Let $\pi^\star \propto \exp(-V)$, where $V : \mathbb{R}^d \to \mathbb{R}$ is L-smooth. Then,*

$$\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\|\nabla V(\mathbf{x})\|_2^2\right] \le Ld.$$

[1] In this lecture, for consistency with Part XIII, we use superscripts to denote an iteration count for ULA, to contrast with subscripts which are used to indicate the passage of time.

Proof. Shifting V by a constant so $\int \exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} = 1$ and integrating by parts,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\pi^\star}\left[\|\nabla V(\mathbf{x})\|_2^2\right] &= -\int \langle \nabla V(\mathbf{x}), \nabla \exp(-V(\mathbf{x}))\rangle \,\mathrm{d}\mathbf{x} \\ &= \int \nabla\cdot(\nabla V(\mathbf{x}))\exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} \\ &= \int \mathrm{Tr}\left(\nabla^2 V(\mathbf{x})\right)\exp(-V(\mathbf{x}))\mathrm{d}\mathbf{x} \le Ld.\end{aligned}$$

□

Lemma 2. *Let $\{\mathbf{x}_t\}_{t\in[0,\eta]}$ follow (1), where $V:\mathbb{R}^d\to\mathbb{R}$ is L-smooth and $\eta \le \frac{1}{3L}$. Then,*

$$\mathbb{E}\left[\|\mathbf{x}_\eta - \mathbf{x}_0\|_2^2\right] \le 6\eta^2\mathbb{E}\left[\|\nabla V(\mathbf{x}_0)\|_2^2\right] + 12\eta d.$$

Proof. By using $\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|_2^2 \le 3\|\mathbf{a}\|_2^2 + 3\|\mathbf{b}\|_2^2 + 3\|\mathbf{c}\|_2^2$, we have for any $t\in[0,\eta]$,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_0\|_2^2\right] &= \mathbb{E}\left[\left\|-\int_0^t \nabla V(\mathbf{x}_s)\mathrm{d}s + \sqrt{2}\boldsymbol{B}_t\right\|_2^2\right] \\ &\le 3t^2\mathbb{E}\left[\|\nabla V(\mathbf{x}_0)\|_2^2\right] + 3\mathbb{E}\left[\left\|\int_0^t (\nabla V(\mathbf{x}_s) - \nabla V(\mathbf{x}_0))\mathrm{d}s\right\|_2^2\right] + 6\mathbb{E}\left\|\boldsymbol{B}_t\right\|_2^2 \\ &\le 3t^2\mathbb{E}\left[\|\nabla V(x_0)\|_2^2\right] + 3t\mathbb{E}\left[\int_0^t \|\nabla V(x_s) - \nabla V(\mathbf{x}_0)\|_2^2\,\mathrm{d}s\right] + 6\mathbb{E}\left\|\boldsymbol{B}_t\right\|_2^2 \\ &\le 3\eta^2\mathbb{E}\left[\|\nabla V(\mathbf{x}_0)\|_2^2\right] + 3\eta L^2\mathbb{E}\left[\int_0^t \|\mathbf{x}_s - \mathbf{x}_0\|_2^2\,\mathrm{d}s\right] + 6\eta d.\end{aligned} \tag{4}$$

The second-to-last inequality was due to Cauchy-Schwarz, i.e., for $\{\mathbf{v}_s\}_{s\in[0,t]}\subset\mathbb{R}^d$,

$$\begin{aligned}\left\|\int_0^t \mathbf{v}_s\mathrm{d}s\right\|_2^2 &= \int_0^t\int_0^t \langle \mathbf{v}_s, \mathbf{v}_{s'}\rangle\,\mathrm{d}s\mathrm{d}s' \\ &\le \int_0^t\int_0^t \left(\frac{1}{2}\|\mathbf{v}_s\|_2^2 + \frac{1}{2}\|\mathbf{v}_{s'}\|_2^2\right)\mathrm{d}s\mathrm{d}s' = t\int_0^t \|\mathbf{v}_s\|_2^2\,\mathrm{d}s,\end{aligned} \tag{5}$$

and the last inequality in (4) used our smoothness assumption. Therefore, the conclusion follows from a variant of Grönwall's inequality (Fact 1, Part II), which states that if $\{\Phi_t\}_{t\in[0,\eta]}$ satisfies the integral inequality $\Phi_t \le C_1 + C_2\int_0^t \Phi_s\mathrm{d}s$, then $\Phi_\eta \le C_1\exp(C_2\eta)$. We apply this to $\Phi_t := \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_0\|_2^2]$ and use the assumption on η, yielding the claim:

$$\mathbb{E}\left[\|\mathbf{x}_\eta - \mathbf{x}_0\|_2^2\right] \le \exp\left(3\eta^2L^2\right)\left(3\eta^2\mathbb{E}\left[\|\nabla V(\mathbf{x}_0)\|_2^2\right] + 6\eta d\right) \le 6\eta^2\mathbb{E}\left[\|\nabla V(\mathbf{x}_0)\|_2^2\right] + 12\eta d.$$

□

We can now analyze the discretization error of one step of the unadjusted Langevin algorithm.

Lemma 3. *Let $V:\mathbb{R}^d\to\mathbb{R}$ be L-smooth and μ-strongly convex. Let $\mathbf{x}_0\sim\pi_0$, let $\{\mathbf{x}_t\}_{t\in[0,\eta]}$ follow (3), and let π_η denote the law of $\mathbf{x}_\eta$. Then, for $\eta \le \frac{\mu}{10L^2}$,*

$$W_2^2(\pi_\eta,\pi^\star) \le \left(1 - \frac{\mu\eta}{2}\right)W_2^2(\pi_0,\pi^\star) + \frac{32\eta^2L^2d}{\mu}.$$

Proof. We first introduce some simplifying notation. Let $\{\bar{\mathbf{x}}_t\}_{t\in[0,\eta]}$ follow (1), starting from $\bar{\mathbf{x}}_0 = \mathbf{x}_0$, and with law $\bar{\pi}_t$ at time $t\in[0,\eta]$. Then the proof of Theorem 2, Part XVI shows that

$$W_2^2(\bar{\pi}_\eta,\pi^\star) \le \exp\left(-2\mu\eta\right)W_2^2(\pi_0,\pi^\star). \tag{6}$$

Next, applying Lemma 2 (with $\eta \leftarrow t$ for each $t \in [0, \eta]$), and using the coupling γ_η of π_η and $\bar{\pi}_\eta$ which share a copy of Brownian motion driving the respective SDEs, shows that

$$W_2^2(\pi_\eta, \bar{\pi}_\eta) \le \mathbb{E}_{(\mathbf{x}_\eta, \bar{\mathbf{x}}_\eta) \sim \gamma_\eta} \left[\|\mathbf{x}_\eta - \bar{\mathbf{x}}_\eta\|_2^2 \right] = \mathbb{E} \left[\left\| \int_0^\eta (\nabla V(\mathbf{x}_t) - \nabla V(\mathbf{x}_0)) \mathrm{d}t \right\|_2^2 \right]$$
$$\le \eta L^2 \mathbb{E} \left[\int_0^\eta \|\mathbf{x}_t - \mathbf{x}_0\|_2^2 \, \mathrm{d}t \right] \le 6\eta^4 L^2 \mathbb{E} \left[\|\nabla V(\mathbf{x}_0)\|_2^2 \right] + 12\eta^3 L^2 d.$$

In the second-to-last inequality, we again used (5) and smoothness, and in the last inequality, we gained a factor of η by using Lemma 2 at each time $t \in [0, \eta]$. We further have, for the optimal coupling $\gamma \in \mathcal{C}(\pi_0, \pi^\star)$ realizing $W_2^2(\pi_0, \pi^\star)$,

$$\begin{aligned} \mathbb{E} \left[\|\nabla V(\mathbf{x}_0)\|_2^2 \right] &\le 2\mathbb{E}_{\mathbf{x}^\star \sim \pi^\star} \left[\|\nabla V(\mathbf{x}^\star)\|_2^2 \right] + 2\mathbb{E}_{(\mathbf{x}_0, \mathbf{x}^\star) \sim \gamma} \left[\|\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}^\star)\|_2^2 \right] \\ &\le 2Ld + 2L^2 \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}^\star) \sim \gamma} \left[\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 \right] = 2Ld + 2L^2 W_2^2(\pi_0, \pi^\star), \end{aligned} \tag{7}$$

using Lemma 1. Combining the above displays and using $\eta L \le \frac{1}{10}$ implies

$$W_2^2(\pi_\eta, \bar{\pi}_\eta) \le 16\eta^3 L^2 d + 12\eta^4 L^4 W_2^2(\pi_0, \pi^\star). \tag{8}$$

Finally, because any three vectors $\mathbf{x}_\eta \sim \pi_\eta, \bar{\mathbf{x}}_\eta \sim \bar{\pi}_\eta, \mathbf{x}_\eta^\star \sim \pi^\star$ satisfy

$$\left\|\mathbf{x}_\eta - \mathbf{x}_\eta^\star\right\|_2^2 \le (1 + \mu\eta) \left\|\bar{\mathbf{x}}_\eta - \mathbf{x}_\eta^\star\right\|_2^2 + \left(1 + \frac{1}{\mu\eta}\right) \|\mathbf{x}_\eta - \bar{\mathbf{x}}_\eta\|_2^2,$$

we combine (6) and (8) to obtain the conclusion:

$$\begin{aligned} W_2^2(\pi_\eta, \pi^\star) &\le (1 + \mu\eta) \, W_2^2(\bar{\pi}_\eta, \pi^\star) + \left(1 + \frac{1}{\mu\eta}\right) W_2^2(\pi_\eta, \bar{\pi}_\eta) \\ &\le (1 + \mu\eta) \exp\left(-2\mu\eta\right) W_2^2(\pi_0, \pi^\star) + \left(1 + \frac{1}{\mu\eta}\right) \left(16\eta^3 L^2 d + 12\eta^4 L^4 W_2^2(\pi_0, \pi^\star)\right) \\ &\le \left(1 - \frac{\mu\eta}{2}\right) W_2^2(\pi_0, \pi^\star) + \frac{32\eta^2 L^2 d}{\mu}. \end{aligned}$$

□

By iterating upon Lemma 3, we obtain a convergence rate for the unadjusted Langevin algorithm in the W_2^2 error metric. As we will discuss in Section 4, this analysis can be slightly improved.

Theorem 1 (W_2^2 convergence of unadjusted Langevin)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $\pi^\star \propto \exp(-V)$, $\kappa := \frac{L}{\mu}$, $\epsilon \in (0, 1)$. Let $\mathbf{x}^{(0)} \leftarrow \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} V(\mathbf{x})$, and consider iterating (2) for $0 \le k < K$ with $\eta = \frac{\epsilon^2 \mu}{128 L^2 d}$. Then, if $\pi^{(K)}$ denotes the law of $\mathbf{x}^{(K)}$,*

$$\mu W_2^2 \left(\pi^{(K)}, \pi^\star\right) \le \epsilon^2, \text{ for } K \ge \frac{256\kappa^2 d}{\epsilon^2} \log\left(\frac{4d}{\epsilon^2}\right).$$

Proof. Let $\pi^{(k)}$ denote the law of $\mathbf{x}^{(k)}$ for all $0 \le k \le K$, and recall that Lemma 5, Part XVI shows that $W_2^2(\pi^{(0)}, \pi^\star) \le \frac{2d}{\mu}$. Moreover, applying Lemma 3 with $\pi_0 \leftarrow \pi^{(k)}$ and $\pi_\eta \leftarrow \pi^{(k+1)}$ shows

$$W_2^2 \left(\pi^{(k+1)}, \pi^\star\right) \le \left(1 - \frac{\mu\eta}{2}\right) W_2^2 \left(\pi^{(k)}, \pi^\star\right) + \frac{32\eta^2 L^2 d}{\mu},$$

for each $0 \le k < K$. Recursing upon this guarantee yields

$$W_2^2 \left(\pi^{(K)}, \pi^\star\right) \le \exp\left(-\frac{\mu\eta K}{2}\right) W_2^2 \left(\pi^{(0)}, \pi^\star\right) + \frac{32\eta^2 L^2 d}{\mu} \cdot \frac{2}{\mu\eta},$$

where we summed a geometric sequence, and our choices of η, K give the claim. □

We remark that we use the more natural error metric μW_2^2 in Theorem 1 as opposed to W_2^2, as it is a scale-invariant quantity in the strong logconcavity parameter μ, and is directly comparable to $D_{\mathrm{KL}}(\cdot\|\pi^\star)$ via the Otto-Villani theorem, i.e., Lemma 13, Part XVI), which states

$$\begin{aligned} \frac{\mu}{2} W_2^2(\pi, \pi^\star) \le D_{\mathrm{KL}}\left(\pi \| \pi^\star\right), \\ \text{if } \pi^\star \text{ satisfies a log-Sobolev inequality with constant } \frac{1}{\mu}. \end{aligned} \tag{9}$$

We also showed in Section 5.2, Part XVI that μ-strong logconcavity implies such a log-Sobolev inequality holds. In the following section, we give an alternative analysis of (2) which shows that we can directly achieve bounds on $D_{\mathrm{KL}}(\pi^{(K)}\|\pi^\star)$, strengthening Theorem 1 as implied by (9).

2 D_{KL} convergence of unadjusted Langevin

Our goal in this section is to give a discrete-time analog of Lemma 10, Part XVI developed by [VW19], which shows rapid convergence of $D_{\mathrm{KL}}(\pi_t\|\pi^\star)$ along the Langevin dynamics when $\pi^\star$ satisfies a log-Sobolev inequality. As in Section 1, the simplest way to measure discretization error is in the W_2^2 metric, as we have already developed such tools (e.g., Lemma 2). We will use the Otto-Villani theorem (9) to relate these W_2^2 errors back to the function value of interest, i.e., $D_{\mathrm{KL}}\left(\cdot\|\pi^\star\right)$. We again start by analyzing the change in KL divergence of the law of an iterate after one step of ULA, which runs the Euler-discretized SDE (3) for time η.

Lemma 4. *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and suppose $\pi^\star \propto \exp(-V)$ satisfies a log-Sobolev inequality with constant $\frac{1}{\mu}$. Let $\mathbf{x}_0 \sim \pi_0$, let $\{\mathbf{x}_t\}_{t\in[0,\eta}$ follow (3), and let π_η denote the law of $\mathbf{x}_\eta$. Then for $\eta \le \frac{\mu}{10L^2}$,*

$$D_{\mathrm{KL}}\left(\pi_\eta \| \pi^\star\right) \le \left(1 - \frac{\mu\eta}{2}\right) D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right) + 9\eta^2 L^2 d.$$

Proof. Throughout this proof, let $\pi_{0t} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ be the density corresponding to the joint law of $(\mathbf{x}_0, \mathbf{x}_t)$, for all $t \in [0, \eta]$. We also use the notation $\pi_{0|t}(\mathbf{x}_0 \mid \mathbf{x}_t)$ to mean the conditional distribution of $\mathbf{x}_0$ given $\mathbf{x}_t$, and similarly define $\pi_{t|0}(\mathbf{x}_t \mid \mathbf{x}_0)$, such that

$$\pi_{0t}(\mathbf{x}_0, \mathbf{x}_t) = \pi_0(\mathbf{x}_0)\pi_{t|0}(\mathbf{x}_t \mid \mathbf{x}_0) = \pi_t(\mathbf{x}_t)\pi_{0|t}(\mathbf{x}_0 \mid \mathbf{x}_t). \tag{10}$$

Our first step is to derive a continuity equation (in the sense of Lemma 6, Part XVI) for the SDE (3). By using the Fokker-Planck equation (Proposition 3, Part XVI), we have that

$$\frac{\partial}{\partial t}\pi_{t|0}(\mathbf{x} \mid \mathbf{x}_0) = \nabla \cdot \left(\nabla V(\mathbf{x}_0)\pi_{t|0}(\mathbf{x} \mid \mathbf{x}_0)\right) + \Delta \pi_{t|0}(\mathbf{x} \mid \mathbf{x}_0).$$

Therefore, averaging over $\mathbf{x}_0 \sim \pi_0$, we have

$$\begin{aligned} \frac{\partial}{\partial t}\pi_t(\mathbf{x}) &= \int \left(\nabla \cdot \left(\nabla V(\mathbf{x}_0)\pi_{t|0}(\mathbf{x} \mid \mathbf{x}_0)\right) + \Delta \pi_{t|0}(\mathbf{x} \mid \mathbf{x}_0)\right) \pi_0(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \\ &= \int \left(\nabla \cdot \left(\nabla V(\mathbf{x}_0)\pi_{0t}(\mathbf{x}_0, \mathbf{x})\right) + \Delta \pi_{0t}(\mathbf{x}_0, \mathbf{x})\right) \mathrm{d}\mathbf{x}_0 \\ &= \nabla \cdot \left(\pi_t(\mathbf{x}) \int \pi_{0|t}(\mathbf{x}_0 \mid \mathbf{x}) \nabla V(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0\right) + \Delta \pi_t(\mathbf{x}) \\ &= \nabla \cdot \left(\pi_t(\mathbf{x}) \mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}\left[\nabla V(\mathbf{x}_0) \mid \mathbf{x}_t = \mathbf{x}\right]\right) + \Delta \pi_t(\mathbf{x}) \\ &= \nabla \cdot \left(\pi_t(\mathbf{x}) \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right) + \nabla \cdot \left(\pi_t(\mathbf{x}) \mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}\left[\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}) \mid \mathbf{x}_t = \mathbf{x}\right]\right). \end{aligned} \tag{11}$$

Comparing to Eq. (16), Part XVI, we see that the continuity equations differ only by a term that looks like $\mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}[\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}) \mid \mathbf{x}_t = \mathbf{x}]$. At this point, our proof is very similar to Lemma

10, Part XVI, except we use the tools from Section 1 to bound the discretization error. Concretely,

$$\begin{aligned}
\frac{\partial}{\partial t} D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right) &= \frac{\partial}{\partial t}\left(\int \pi_t(\mathbf{x}) \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \mathrm{d}\mathbf{x}\right) \\
&= \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \nabla \cdot \left(\pi_t(\mathbf{x}) \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right) \mathrm{d}\mathbf{x} \\
&+ \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \nabla \cdot \left(\pi_t(\mathbf{x}) \mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}\left[\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}) \mid \mathbf{x}_t = \mathbf{x}\right]\right) \mathrm{d}\mathbf{x} \\
&= -\int \left\|\nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\|_2^2 \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&- \int \left\langle \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right), \mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}\left[\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}) \mid \mathbf{x}_t = \mathbf{x}\right] \right\rangle \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\leq -\frac{1}{2} \int \left\|\nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\|_2^2 \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&+ \frac{1}{2} \int \left\|\mathbb{E}_{\mathbf{x}_0 \sim \pi_{0|t}}\left[\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x}) \mid \mathbf{x}_t = \mathbf{x}\right]\right\|_2^2 \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\leq -\frac{1}{2} \int \left\|\nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\|_2^2 \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} + \frac{1}{2} \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}) \sim \pi_{0t}}\left[\left\|\nabla V(\mathbf{x}_0) - \nabla V(\mathbf{x})\right\|_2^2\right],
\end{aligned} \tag{12}$$

where the second line again used $\frac{\partial}{\partial t} \int \pi_t(\mathbf{x}) \mathrm{d}\mathbf{x} = \frac{\partial}{\partial t} 1 = 0$ and substituted (11), the fourth line used integration by parts, the sixth line used $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2$, and the last line used Jensen's inequality. Now, by plugging in the log-Sobolev inequality (in the form of Lemma 8, Part XVI) into (12), as well as our bound from Lemma 2,

$$\frac{\partial}{\partial t} D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right) \leq -\mu D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right) + 3\eta^2 L^2 \mathbb{E}\left[\left\|\nabla V(\mathbf{x}_0)\right\|_2^2\right] + 6\eta L^2 d.$$

Moreover, using the bound (7) with Talagrand's transportation inequality (9) shows

$$\mathbb{E}\left[\left\|\nabla V(\mathbf{x}_0)\right\|_2^2\right] \leq 2Ld + 2L^2 W_2^2(\pi_0, \pi^\star) \leq 2Ld + \frac{4L^2}{\mu} D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right).$$

Combining the above two displays and using our bound on η finally yields

$$\begin{aligned}
&\frac{\partial}{\partial t} D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right) \leq -\mu D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right) + 9\eta L^2 d + \frac{12\eta^2 L^4}{\mu} D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right) \\
\implies &\frac{\partial}{\partial t}\left(\exp\left(\mu t\right) D_{\mathrm{KL}}\left(\pi_t \| \pi^\star\right)\right) \leq \exp\left(\mu t\right)\left(9\eta L^2 d + \frac{12\eta^2 L^4}{\mu} D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right)\right).
\end{aligned}$$

The conclusion then follows from integrating and using our choice of η:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\pi_\eta \| \pi^\star\right) &\leq \exp\left(-\mu\eta\right)\left(D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right) + \eta \exp(\mu\eta)\left(9\eta L^2 d + \frac{12\eta^2 L^4}{\mu} D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right)\right)\right) \\
&\leq \left(1 - \frac{\mu\eta}{2}\right) D_{\mathrm{KL}}\left(\pi_0 \| \pi^\star\right) + 9\eta^2 L^2 d.
\end{aligned}$$

□

At this point, the same recursion as used in Theorem 1 (with slightly different parameters), using the one-step guarantee in Lemma 4 rather than Lemma 3, yields our desired convergence rate.

Theorem 2 (D_{KL} convergence of unadjusted Langevin)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and suppose $\pi^\star \propto \exp(-V)$ satisfies a log-Sobolev inequality with constant $\frac{1}{\mu}$, and let $\kappa := \frac{L}{\mu}$, $\epsilon \in (0, 1)$. Let $\mathbf{x}^{(0)} \sim \pi_0$, and consider iterating the update (2) for $0 \leq k < K$ with $\eta = \frac{\epsilon^2 \mu}{72 L^2 d}$. Then, if $\pi^{(K)}$ denotes the law of $\mathbf{x}^{(K)}$,*

$$D_{\mathrm{KL}}\left(\pi^{(K)} \| \pi^\star\right) \leq \frac{\epsilon^2}{2} \text{ for } K \geq \frac{144\kappa^2 d}{\epsilon^2} \log\left(\frac{4 D_{\mathrm{KL}}\left(\pi^{(0)} \| \pi^\star\right)}{\epsilon^2}\right).$$

Proof. As in the proof of Theorem 1, applying Lemma 4 for K iterations yields

$$D_{\mathrm{KL}}\left(\pi^{(K)}\|\pi^\star\right) \le \exp\left(-\frac{\mu\eta K}{2}\right) D_{\mathrm{KL}}\left(\pi^{(0)}\|\pi^\star\right) + 9\eta^2 L^2 d \cdot \frac{2}{\mu\eta}$$
$$\le \exp\left(-\frac{\mu\eta K}{2}\right) D_{\mathrm{KL}}\left(\pi^{(0)}\|\pi^\star\right) + \frac{\epsilon^2}{4} \le \frac{\epsilon^2}{2}.$$

□

As discussed at the end of Section 1, the assumptions made in Theorem 2 are actually weaker than those in Theorem 1, since strong logconcavity implies a log-Sobolev inequality (but not the other way around). Moreover, Theorem 2 implies Theorem 1 up to constants, via (9). The reason for the scaling $\frac{\epsilon^2}{2}$ in Theorem 2 is that Pinsker's inequality then shows $D_{\mathrm{TV}}(\pi^{(K)}, \pi^\star) \le \epsilon$ as well.

3 Metropolis-adjusted Langevin algorithm

In this section, we show give *high-accuracy* convergence rates for Langevin-based algorithms in discrete time, by making use of the *Metropolis-Hastings* filter (Eq. (5), Part XV). Unlike the unadjusted Langevin algorithms analyzed in Sections 1 and 2, the *Metropolis-adjusted Langevin algorithm* (MALA) actually has a stationary distribution of $\pi^\star \propto \exp(-V)$, rather than a biased approximation thereof. This lets us use tools from Part XV to give gradient query complexities depending polylogarithmically on $\frac{1}{\epsilon}$, where ϵ is a specified distance bound from $\pi^\star$. In contrast, Theorems 1 and 2 required querying ∇V a number of times scaling polynomially in $\frac{1}{\epsilon}$.

We now define the MALA iteration for sampling from $\pi^\star \propto \exp(-V)$, parameterized by a step size $\eta > 0$. Following notation from Section 2.1, Part XV, the proposal distribution from $\mathbf{x} \in \mathbb{R}^d$ is

$$\mathcal{P}_{\mathbf{x}} = \mathcal{N}\left(\mathbf{x} - \eta\nabla V(\mathbf{x}), 2\eta\mathbf{I}_d\right). \tag{13}$$

Observe that this proposal distribution is simply the distribution of $\mathbf{x}^{(k+1)}$ induced by one step of ULA (2), starting from $\mathbf{x}^{(k)} \leftarrow \mathbf{x}$. Next, the MALA transition density is the result of applying a Metropolis-Hastings correction to (13), i.e.,

$$\mathcal{T}_{\mathbf{x}}(\mathbf{y}) = \mathcal{P}_{\mathbf{x}}(\mathbf{y}) \min\left(1, \frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})}\right), \text{ for all } \mathbf{y} \in \mathbb{R}^d, \ \mathbf{y} \neq \mathbf{x}. \tag{14}$$

To analyze the convergence of MALA, specified by transitions (14), we leverage Sections 2.2 and 3, Part XV, specifically Propositions 1 and 3. These results give a recipe for achieving an error bound of ϵ in χ^2 assuming three conditions have been met.

1. We can sample from π_0, a β-warm distribution for $\pi^\star$.
2. The stationary distribution $\pi^\star$ satisfies an isoperimetric inequality.
3. There is a set $\Omega \subset \mathbb{R}^d$ with large stationary measure $\pi^\star(\Omega) \ge 1 - \frac{\epsilon^2}{3\beta^2}$ such that for all sufficiently close pairs of $\mathbf{x}, \mathbf{x}' \in \Omega$, we have $D_{\mathrm{TV}}(\mathcal{T}_{\mathbf{x}}, \mathcal{T}_{\mathbf{x}'}) \le \frac{1}{2}$.

Under the latter two conditions above, Proposition 3, Part XV proves a conductance bound over Ω, and this can be put into Proposition 1, Part XV with the first condition above to give an overall mixing time bound. We now address each condition, beginning with the warm start.

Lemma 5. *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $\kappa := \frac{L}{\mu}$, $\epsilon \in (0, 1)$. For $\mathbf{x}^\star := \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} V(\mathbf{x})$, $\pi_0 := \mathcal{N}(\mathbf{x}^\star, \frac{1}{L}\mathbf{I}_d)$ is β-warm with respect to $\pi^\star$, for $\beta = \kappa^{\frac{d}{2}}$.*

Proof. By smoothness and strong convexity, we have for all $\mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|_2^2 \le f(\mathbf{x}) \le f(\mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^\star\|_2^2.$$

Thus, by plugging in the exact formula for π_0, we derive

$$\begin{aligned}\frac{\pi_0(\mathbf{x})}{\pi^\star(\mathbf{x})} &= \frac{\int \exp(-f(\mathbf{y}))\mathrm{d}\mathbf{y}}{\exp(-f(\mathbf{x}))} \cdot \frac{\exp\left(-\frac{L}{2}\left\|\mathbf{x}-\mathbf{x}^\star\right\|_2^2\right)}{(2\pi L^{-1})^{\frac{d}{2}}} \\ &\le \frac{\int \exp(-f(\mathbf{x}^\star) - \frac{\mu}{2}\left\|\mathbf{y}-\mathbf{x}^\star\right\|_2^2)\mathrm{d}\mathbf{y}}{\exp(-f(\mathbf{x}^\star) - \frac{L}{2}\left\|\mathbf{x}-\mathbf{x}^\star\right\|_2^2)} \cdot \frac{\exp\left(-\frac{L}{2}\left\|\mathbf{x}-\mathbf{x}^\star\right\|_2^2\right)}{(2\pi L^{-1})^{\frac{d}{2}}} \\ &= \frac{\int \exp(-\frac{\mu}{2}\left\|\mathbf{y}-\mathbf{x}^\star\right\|_2^2)\mathrm{d}\mathbf{y}}{(2\pi L^{-1})^{\frac{d}{2}}} = \frac{(2\pi\mu^{-1})^{\frac{d}{2}}}{(2\pi L^{-1})^{\frac{d}{2}}} = \kappa^{\frac{d}{2}}.\end{aligned}$$

□

Next, we provide a useful isoperimetric bound for *strongly logconcave* densities.

Lemma 6. *Let $\pi^\star$ be a μ-strongly logconcave density on $\mathbb{R}^d$, and let S_1, S_2, S_3 partition $\mathbb{R}^d$. Then,*

$$\frac{\pi^\star(S_3)}{(\min_{\mathbf{x}\in S_1, \mathbf{y}\in S_2}\left\|\mathbf{x}-\mathbf{y}\right\|_2)} \ge \frac{\pi^\star(S_1)\pi^\star(S_2)}{2}\sqrt{\mu\log\left(1+\frac{1}{\min\left(\pi^\star(S_1),\pi^\star(S_2)\right)}\right)}. \tag{15}$$

Proof. We describe the main ideas behind the proof here, deferring full details to Lemma 16, [CDWY20]. The idea is to use a localization argument (e.g., Proposition 4, Part XV) to show that it suffices to prove (15) for $\pi^\star = \mathcal{N}(x, \frac{1}{\mu})$, restricted to a one-dimensional subspace of $\mathbb{R}^d$.

The rough strategy for doing so is to first express (15) using linear inequality constraints on $\pi^\star$, similarly to Lemma 3, Part XV. Then, the localization lemma (Proposition 4, Part XV) implies that it suffices to prove (15) for all log-linear densities in one dimension, convolved with $\mathcal{N}(0, \frac{1}{\mu})$. As discussed after Proposition 4, Part XV, the localization lemma applies here because strongly logconcave densities are closed under restricting to convex sets.

Log-linear densities in one dimension convolved with $\mathcal{N}(0, \frac{1}{\mu})$ are just recentered Gaussians of the form $\mathcal{N}(x, \frac{1}{\mu})$. For such densities, it is a classical fact that (15) holds. We can first restrict to the case of $S_1 = (-\infty, a]$, $S_3 = (a, b)$, and $S_2 = [b, \infty)$ without loss of generality, by a similar partitioning argument as used in Lemma 3, Part XV. Then, a bound of the form (15) holds by an averaging argument over S_3, and standard tail bounds over Gaussian random variables.

To provide some intuition for this, it turns out that the extreme case is where $a \to b > 0$ (i.e., $S_2 = [b, \infty)$ has smaller mass under $\pi^\star$, and S_3 approaches a single point, so the left-hand side of (15) is roughly the density at b). We can estimate using Mill's inequality that

$$\pi^\star(S_2) = \Pr_{\xi\sim\mathcal{N}(0,\frac{1}{\mu})}[\xi > b] \approx \frac{1}{b\sqrt{\mu}}\exp\left(-\frac{\mu b^2}{2}\right).$$

A straightforward calculation now shows that both sides of (15) scale as

$$\sqrt{\mu}\exp\left(-\frac{\mu b^2}{2}\right).$$

□

Note that Lemma 6 proves that the *isoperimetric constant* (Definition 4, Part XV) of any μ-strongly logconcave $\pi^\star$ is $\Omega(\sqrt{\mu})$. However, we can say more: the isoperimetry improves at smaller scales of $\min(\pi^\star(S_1), \pi^\star(S_2))$ (i.e., small sets "expand" more). This turns out to significantly sharpen mixing time bounds in certain applications, see the discussion after Proposition 1, Part XV.

We next provide an Ω with enough regularity to ensure that $\mathcal{T}_{\mathbf{x}}$, $\mathcal{T}_{\mathbf{x}'}$ have significant overlap, for nearby $\mathbf{x}, \mathbf{x}' \in \Omega$. In our case it is enough to take Ω to be a sufficiently large ℓ_2 norm ball.

Lemma 7. *Let $\pi^\star \propto \exp(-V)$ be a μ-strongly logconcave density on $\mathbb{R}^d$, and let $\mathbf{x}^\star := \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} V(\mathbf{x})$. For all $\delta \in (0,1)$, we have*

$$\Pr_{\mathbf{x}\sim\pi^\star}\left[\left\|\mathbf{x}-\mathbf{x}^\star\right\| > \sqrt{\frac{2d}{\mu}} + \sqrt{\frac{2\log(\frac{1}{\delta})}{\mu}}\right] \le \delta.$$

Proof. Recall that $\mathbb{E}\left\|\mathbf{x}-\mathbf{x}^\star\right\|_2 \leq (\frac{2d}{\mu})^{1/2}$, by Lemma 5, Part XVI and Jensen's inequality. Also, $\pi^\star$ satisfies a log-Sobolev inequality with constant $\frac{1}{\mu}$ (Theorem 4, Part XVI), so the conclusion now follows from Lemma 11, Part XVI, because $\left\|\mathbf{x}-\mathbf{x}^\star\right\|_2$ is a 1-Lipschitz function. □

Lemma 8. *In the setting of Lemma 7, let V be L-smooth, let $R > 0$, and let $\Omega := \mathbb{B}(\mathbf{x}^\star, R)$. Assume that $LR^2 \geq d$. Then if $\eta \leq \frac{1}{500L^2R^2d}$ in (13), for all $\mathbf{x}, \mathbf{x}' \in \Omega$, we have*

$$D_{\mathrm{TV}}\left(\mathcal{T}_{\mathbf{x}}, \mathcal{T}_{\mathbf{x}'}\right) \leq \frac{1}{2} \text{ if } \left\|\mathbf{x}-\mathbf{x}'\right\|_2 \leq \frac{\sqrt{\eta}}{5}.$$

Proof. We claim that such nearby $\left\|\mathbf{x}-\mathbf{x}'\right\|_2 \leq$ satisfy

$$D_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{x}}, \mathcal{P}_{\mathbf{x}'}\right) \leq \frac{1}{6}, \tag{16}$$

and further, that all $\mathbf{x} \in \Omega$ satisfy

$$D_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{x}}, \mathcal{T}_{\mathbf{x}}\right) \leq \frac{1}{6}. \tag{17}$$

Combining (16) and (17) then gives the desired claim $D_{\mathrm{TV}}(\mathcal{T}_{\mathbf{x}}, \mathcal{T}_{\mathbf{x}'}) \leq D_{\mathrm{TV}}(\mathcal{T}_{\mathbf{x}}, \mathcal{P}_{\mathbf{x}}) + D_{\mathrm{TV}}(\mathcal{P}_{\mathbf{x}}, \mathcal{P}_{\mathbf{x}'}) + D_{\mathrm{TV}}(\mathcal{P}_{\mathbf{x}'}, \mathcal{T}_{\mathbf{x}'}) \leq \frac{1}{2}$. To prove (16), we have for $\mathcal{P}_{\mathbf{x}}, \mathcal{P}_{\mathbf{x}'}$ defined in (13) that

$$\begin{aligned} D_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{x}}, \mathcal{P}_{\mathbf{x}'}\right) &\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{P}_{\mathbf{x}} \| \mathcal{P}_{\mathbf{x}'}\right)} \\ &= \frac{\left\|(\mathbf{x}-\eta\nabla V(\mathbf{x})) - (\mathbf{x}'-\eta\nabla V(\mathbf{x}'))\right\|_2}{\sqrt{8\mu}} \\ &\leq \frac{(1+\eta L)\left\|\mathbf{x}-\mathbf{x}'\right\|_2}{\sqrt{8\eta}} \leq \frac{\left\|\mathbf{x}-\mathbf{x}'\right\|_2}{\sqrt{2\eta}} \leq \frac{1}{6}, \end{aligned} \tag{18}$$

where the first line in (18) used Pinsker's inequality (i.e., a continuous-valued analog of Lemma 6 and Remark 5, Part III), the second line used an exact formula for the KL divergence between multivariate Gaussians, and the third line used smoothness of V and our bounds on η, $\left\|\mathbf{x}-\mathbf{x}'\right\|_2$.

We are left to show (17). First, note that a draw from $\mathcal{P}_{\mathbf{x}}$ can be decomposed as $\mathbf{x}-\eta\nabla V(\mathbf{x})+\sqrt{2\eta}\boldsymbol{\xi}$, for $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. We claim that with probability $\geq \frac{11}{12}$, if d is sufficiently large,

$$\left\|\boldsymbol{\xi}\right\|_2 \leq \sqrt{2d}, \tag{19}$$

which follows from standard chi-squared tail bounds (see e.g., the discussion after Theorem 2, Part VI). Next, observe that drawns from $\mathcal{P}_{\mathbf{x}}$ and $\mathcal{T}_{\mathbf{x}}$ can be coupled as long as the Metropolis-Hastings filter is not applied. Thus, we have

$$D_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{x}}, \mathcal{T}_{\mathbf{x}}\right) = 1 - \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[\min\left(1, \frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})}\right)\right], \text{ for } \mathbf{y} := \mathbf{x}-\eta\nabla V(\mathbf{x})+\sqrt{2\eta}\boldsymbol{\xi}. \tag{20}$$

Finally, we claim that whenever $\boldsymbol{\xi}$ satisfies (19), we have that

$$\frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})} \geq \frac{11}{12}, \tag{21}$$

This would conclude the proof, because by Markov's inequality,

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[\min\left(1, \frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})}\right)\right] \geq \frac{11}{12} \Pr_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)}\left[\frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})} \geq \frac{11}{12}\right] \geq \left(\frac{11}{12}\right)^2 \geq \frac{5}{6},$$

which plugged into (20) gives our claim (17). To see (21), we have by a direct calculation that

$$\frac{\pi^\star(\mathbf{y})\mathcal{P}_{\mathbf{y}}(\mathbf{x})}{\pi^\star(\mathbf{x})\mathcal{P}_{\mathbf{x}}(\mathbf{y})} = \exp\left(\frac{-V(\mathbf{y}) - \frac{\left\|\mathbf{x}-\mathbf{y}+\eta\nabla V(\mathbf{y})\right\|_2^2}{4\eta}}{-V(\mathbf{x}) - \frac{\left\|\mathbf{x}-\mathbf{y}+\eta\nabla V(\mathbf{x})\right\|_2^2}{4\eta}}\right).$$

Thus to prove (21) we need to show that

$$V(\mathbf{x}) - V(\mathbf{y}) \geq -\frac{1}{24},\ \|\mathbf{x} - \mathbf{y} + \eta\nabla V(\mathbf{x})\|_2^2 - \|\mathbf{x} - \mathbf{y} + \eta\nabla V(\mathbf{y})\|_2^2 \geq -\frac{\eta}{6}. \tag{22}$$

To obtain these bounds, first note that under (19) and $\|\mathbf{x} - \mathbf{x}^\star\|_2 \leq R$, we have from smoothness that $\|\mathbf{x} - \mathbf{y}\|_2 \leq \eta LR + 3\sqrt{\eta d}$, so

$$\begin{aligned} V(\mathbf{x}) - V(\mathbf{y}) &\geq \langle \nabla V(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \\ &\geq -\eta L^2 R^2 - 3LR\sqrt{\eta d} - \frac{\eta^2 L^3 R^2}{2} - \frac{9\eta L d}{2} \geq -\frac{1}{24}, \end{aligned}$$

from our choice of parameters. Similarly, we can verify that

$$\begin{aligned} \|\mathbf{x} - \mathbf{y} + \eta\nabla V(\mathbf{x})\|_2^2 - \|\mathbf{x} - \mathbf{y} + \eta\nabla V(\mathbf{y})\|_2^2 &\geq 2\eta\langle \nabla V(\mathbf{x}) - \nabla V(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle - \eta^2\|\nabla V(\mathbf{y})\|_2^2 \\ &\geq -\left(2\eta + \eta^2 L\right) L\|\mathbf{x} - \mathbf{y}\|_2^2 - \eta^2 L^2 R^2 \geq -\frac{\eta}{6}. \end{aligned}$$

□

All that is left in our MALA analysis is combining these results within Proposition 1, Part XV.

Theorem 3 (Convergence of Metropolis-adjusted Langevin)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $\pi^\star \propto \exp(-V)$, $\kappa := \frac{L}{\mu}$, $\epsilon \in (0,1)$. Let $\mathbf{x}^{(0)} \sim \mathcal{N}(\mathbf{x}^\star, \frac{1}{L}\mathbf{I}_d)$ where $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} V(\mathbf{x})$, and consider iterating the Markov chain with transitions (14) for K iterations. Then if $\pi^{(K)}$ denotes the law of $\mathbf{x}^{(K)}$, the K^{th} Markov chain iterate,*

$$\chi^2\left(\pi^{(K)}\|\pi^\star\right) \leq \epsilon,\ \text{if } \eta = \frac{1}{10^5 L\kappa d(d\log(\kappa) + \log(\frac{1}{\epsilon}))},\ K = \Omega\left(\kappa^2 d^2 \log^2\left(\frac{d\log(\kappa)}{\epsilon}\right)\right).$$

Proof. Throughout this proof, let $\beta = \kappa^{\frac{d}{2}}$ be the warmness parameter given by Lemma 5. Our goal is to apply Proposition 1, Part XV, so the next step is to choose a set Ω with good conductance properties, and with stationary measure $\geq 1 - \frac{\epsilon^2}{3\beta^2}$. Letting $\delta = \frac{\epsilon^2}{3\beta^2}$, Lemma 7 shows we may choose $\Omega = \mathbb{B}(\mathbf{x}^\star, R)$ for

$$R := 10\sqrt{\frac{d\log(\kappa) + \log\left(\frac{1}{\epsilon}\right)}{\mu}} \geq \sqrt{\frac{2d}{\mu}} + \sqrt{\frac{2\log(\frac{1}{\delta})}{\mu}}.$$

Note that by plugging Lemmas 6 and 8 into Proposition 3, Part XV, we have proven that for $\eta = \frac{1}{500L^2R^2d}$, and all $\tau \in (0, \frac{1}{2}\pi^\star(\Omega))$, that following Eq. (11), Part XV,

$$\Phi_\Omega(\tau) \geq \min\left(\frac{1}{8}, \frac{\sqrt{\eta\mu\log(1 + \frac{1}{\tau})}}{640}\right).$$

By plugging this into Proposition 1, Part XV (see the discussion thereafter), we conclude that it is enough to take (for a sufficiently large constant)

$$K = \Omega\left(\kappa^2 d^2 \log^2\left(\frac{d\log(\kappa)}{\epsilon}\right)\right) = \Omega\left(\frac{1}{\eta\mu}\log\log(\beta) + \log(\beta) + \frac{1}{\eta\mu}\log\left(\frac{1}{\epsilon}\right)\right).$$

□

Remark 1. *The analysis of MALA in this section can be sharpened significantly. By being more careful about the decomposition of the rejection probability as in Lemma 8, [DCWY19, CDWY20] proved that $\approx \kappa d + \kappa^{1.5}\sqrt{d}$ iterations of MALA suffice for high-accuracy mixing. This was later improved to $\approx \kappa d$ by [LST20], who proved a good conductance bound over a more complex set Ω where both iterate norms and gradient norms are small. Finally, [CLA+21] directly established a spectral gap of $\mathrm{poly}(\kappa)\sqrt{d}$ for MALA with an appropriate step size. This result on its own gives mixing times scaling as $\approx \sqrt{d}$ only if a polynomially-warm start is provided (recall that the natural strategy in Lemma 5 only yields $\beta \approx \exp(d)$). However, combined with other tools developed in the literature since, the [CLA+21] strategy yields the sharpest-known mixing bound on MALA.*

4 The frontier

The results presented in this lecture are suggestive of the following natural question: what is the minimum number of gradient queries needed to sample from $\pi^\star \propto \exp(-V)$ to accuracy ϵ (in an appropriate error metric), when $V : \mathbb{R}^d \to \mathbb{R}$ obeys specified regularity conditions (e.g., L-smoothness and μ-strong convexity)? Recall that in Part II, we gave matching upper and lower bounds for this question in the context of optimizing V to additive error ϵ, under a variety of regularity conditions, e.g., Lipschitzness, smoothness, and strong convexity.

This question was addressed first in [Dal17] for the family of L-smooth and μ-strongly convex potentials $V : \mathbb{R}^d \to \mathbb{R}$, parameterizing a target density $\pi^\star \propto \exp(-V)$. Here we primarily discuss existing sampling theory for this family, parameterized by $\kappa := \frac{L}{\mu}$. Unfortunately, our current understanding of even this basic landscape is somewhat murky, with various (sometimes incomparable) guarantees. We restrict our discussion to the following error metrics, for some $\epsilon \in (0, 1)$.

- Sampling from within ϵ in total variation distance to $\pi^\star$.
- Sampling from within ϵ^2 in KL divergence to $\pi^\star$ (or, the α-Rényi divergence for $\alpha > 1$).
- Sampling from within $\sqrt{\mu}\epsilon$ in W_2 to $\pi^\star$.

To compare these different notions of convergence, recall that Pinsker's inequality shows that convergence in KL divergence implies convergence in total variation, and further, that the Rényi divergences are monotone nondecreasing in α. Moreover, when V is μ-strongly convex, the Otto-Villani theorem (Lemma 13, Part XVI) says that convergence in KL divergence implies convergence in W_2. Thus, the relative strength of these metrics is

$$\{D_{\mathrm{TV}}, \sqrt{\mu} W_2\} \lesssim \sqrt{D_{\mathrm{KL}}} \lesssim \sqrt{D_\alpha}.$$

Low-accuracy guarantees. Gradient query complexities for samplers are typically parameterized by a function of $\kappa, d, \frac{1}{\epsilon}$, and the *low-accuracy* regime allows for a polynomial dependence on $\frac{1}{\epsilon}$. In this setting, algorithmic improvements largely come from either considering alternatives to the Langevin dynamics with the potential for faster mixing (e.g., the *underdamped Langevin dynamics* [CCBJ18]), or from considering alternative discretization strategies to the standard naïve forward Euler strategy [FLO21, BRM25]. Under a W_2 notion of convergence, the state-of-the-art result is by [SL19], who used both of these strategies to achieve a gradient query complexity of

$$\approx \frac{\kappa^{\frac{7}{6}} d^{\frac{1}{6}}}{\epsilon^{\frac{1}{3}}} + \frac{\kappa d^{\frac{1}{3}}}{\epsilon^{\frac{2}{3}}},$$

via a second-order discretization approach they introduced, the *randomized midpoint method*. It is standard to prioritize improvements to the d dependence in this regime (after all, the structural assumption is that κ is bounded). Thus, the [SL19] rate scales with $d^{\frac{1}{3}}$ for W_2 convergence.

What is the situation for stronger metrics, e.g., D_{KL} and D_α? Currently, it is only known how to obtain query complexities of $\sqrt{d} \cdot \mathrm{poly}(\kappa, \frac{1}{\epsilon})$ in KL divergence [ZCL+23] or $\sqrt{d} \cdot \mathrm{poly}(\kappa, \frac{1}{\epsilon}, \alpha)$ in Rényi divergence of order $\alpha > 1$ [AC24]. These results are based on analyses of Euler discretizations of the underdamped Langevin dynamics. There are certain challenges involved in the analysis of the randomized midpoint method under these stronger metrics, though there is certainly potential for a $d^{\frac{1}{3}}$-type convergence result for all Rényi divergences. On the lower bound side, there is evidence that current techniques may be bottlenecked at gradient query complexities of $\approx d^{\frac{1}{3}}$ [CLW21]; however, fully algorithm-independent lower bounds in this regime remain lacking.

High-accuracy guarantees. As mentioned previously (Remark 1), [CLA+21] gave a roadmap towards an $\approx \sqrt{d} \cdot \mathrm{poly}(\kappa, \log(\frac{1}{\epsilon}))$ gradient query complexity: provide a $\beta = \mathrm{poly}(d, \kappa, \frac{1}{\epsilon})$-warm start for MALA, using this number of queries. The situation was simplified by [LST21], who gave a generic reduction, showing that all query complexities in the high-accuracy regime scaled at most linearly in κ without loss of generality. Finally, [AC24] showed how Rényi divergence guarantees from low-accuracy samplers could be converted into a warmness bound, and using a new analysis of the underdamped Langevin dynamics that they provide, gave an end-to-end algorithm using $\approx \kappa\sqrt{d}$ queries to sample to high accuracy in Rényi divergence. On the other hand, current techniques yield no better than lower bounds of $\approx \sqrt{\kappa} \log(d)$ [CdDPL+24]. The question of optimal gradient query algorithms for well-conditioned sampling remains an exciting open problem.

Source material

Portions of this lecture are based on reference material in [Che24], as well as the author's own experience working in the field.

References

[AC24] Jason M. Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *J. ACM*, 71(3):24, 2024.

[BRM25] Nawaf Bou-Rabee and Milo Marsden. Unadjusted hamiltonian mcmc with stratified monte carlo time integration. *The Annals of Applied Probability*, 35(1):360–392, 2025.

[CCBJ18] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 2018.

[CdDPL+24] Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. *J. ACM*, 71(4):29:1–29:42, 2024.

[CDWY20] Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:92:1–92:72, 2020.

[Che24] Sinho Chewi. *Log-Concave Sampling*. 2024.

[CLA+21] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1260–1300. PMLR, 2021.

[CLW21] Yu Cao, Jianfeng Lu, and Lihan Wang. Complexity of randomized algorithms for underdamped langevin dynamics. *Communications in Mathematical Sciences*, 19(7):1827–1853, 2021.

[Dal17] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[DCWY19] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *J. Mach. Learn. Res.*, 20:183:1–183:42, 2019.

[FLO21] James Foster, Terry J. Lyons, and Harald Oberhauser. The shifted ODE method for underdamped langevin MCMC. *CoRR*, abs/2101.03446, 2021.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized hamiltonian monte carlo. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2565–2597. PMLR, 2020.

[LST21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2993–3050. PMLR, 2021.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 2098–2109, 2019.

[VW19] Santosh S. Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Pro-*

cessing Systems 32: Annual Conference on Neural Information Processing Systems 2019, pages 8092–8104, 2019.

[ZCL+23] Matthew Shunshi Zhang, Sinho Chewi, Mufan (Bill) Li, Krishna Balasubramanian, and Murat A. Erdogdu. Improved discretization analysis for underdamped langevin monte carlo. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 36–71. PMLR, 2023.