

Discretizing Langevin Dynamics (LD)

— Langevin Monte Carlo (LMC)

Refs:

1. S. Chewi, Log-concave sampling, (book draft), Nov. 2024.
2. A. Wibisono, Sampling as optimization in the space of measures: the Langevin Dynamics as a composite optimization problem. COLT 2018.
3. A. Dalalyan and A. Tsybakov, Sparse regression learning by aggregation and Langevin Monte Carlo, Journal of Comput. System Science, 2012.
4. S. Vempala and A. Wibisono, Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices, NeurIPS 2019.
5. S. Chewi, M. Erdogdu, M. Li, R. Shen and M. Zhang, Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev, arXiv: 2112.12662, 2021.

(I closely follow [1])

So far, we have discussed sampling algorithms that have evolved in continuous time. In practice, time is discretized. This discretization leads to an error in the distribution of samples. This note quantifies this discretization error for Langevin Dynamics.

Setting and Notation:

$$\pi_*(x) \sim e^{-V(x)}, \quad x \in \mathbb{R}^d$$

$$dx_t = -\nabla V(x_t) dt + \sqrt{2} dB_t \quad (1D)$$

continuous-time evolution

The discretization of the above is done through the Euler-Maruyama discretization, which is the "natural" time discretization $t \rightarrow kh$, $h = \text{step size}$, $k \in \{0, 1, 2, \dots\}$. Care is taken to make sure that the Brownian motion discretization continues to be independent of current state.

$$\tilde{X}_{(k+1)h} = \tilde{X}_{kh} - h \nabla V(\tilde{X}_{kh}) + \sqrt{2} \underbrace{(B_{(k+1)h} - B_{kh})}_{\sim N(0, h), \text{ indep. of } \tilde{X}_{kh}}$$

Langevin Monte Carlo (LMC)

discrete-time evolution.

In the second proof (Girsanov Thm approach), we will need to embed this discrete-time evolution back into continuous time. The mapping ensures that:

- (i) At each continuous time when $t = kh$, the discrete evolution equals the value of the embedding, and
- (ii) Between timesteps, the drift term is linearly interpolated and the Brownian motion naturally evolves as below:

$$\tilde{X}_t = \tilde{X}_{\lfloor \frac{t}{h} \rfloor h} - (t - \lfloor \frac{t}{h} \rfloor h) \nabla V(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}) + \sqrt{2} (B_t - B_{\lfloor \frac{t}{h} \rfloor h})$$

continuous-time embedding of LMC.

We present two proofs for deriving the discretization error between LD and LMC. There are four proofs in [1] (all of which are illuminating to understand the machinery of discretization). In these notes, we only present two of the proofs. The assumptions on $V(\cdot)$ differ across these proofs.

① Proof via Wasserstein Coupling (4.1 in [1])

In this proof, we assume that $V(x)$ is strongly convex and smooth, i.e.,

$$\alpha I_d \leq \nabla^2 V(x) \leq \beta I_d, \quad 0 < \alpha < \beta < \infty$$

Lemma 1: $\pi_*(x) \propto e^{-V(x)}, \quad x \in \mathbb{R}^d$

(a) Suppose $\nabla^2 V(x) \geq \alpha I_d$, and $x_* = \arg\min_x V(x)$.

$$\text{Then, } E_{\pi} [\|x - x_*\|^2] \leq d/\alpha$$

(b) Suppose $\nabla^2 V(x) \leq \beta I_d$.

$$\text{Then } E_\pi[\|\nabla V(x)\|^2] \leq \beta d.$$

Pf: Recall for the Langevin Dynamics, we have

$$\mathcal{L} f(x) = \Delta f(x) - \langle \nabla V(x), \nabla f(x) \rangle,$$

i.e.,

$$\mathcal{L}(f)(x) = \sum_{i=1}^d \frac{\partial^2 f(x)}{\partial x_i^2} - \sum_{i=1}^d \frac{\partial V(x)}{\partial x_i} \cdot \frac{\partial f(x)}{\partial x_i}$$

and $E_\pi[\mathcal{L} f(x)] = 0.$

(a) Let $f(x) = \frac{1}{2} \|x - x_*\|^2$. Then,

$$0 = E_\pi[\mathcal{L} f(x)] = d - E_\pi[\langle \nabla V(x), (x - x_*) \rangle]$$

$$\leq d - \alpha E_\pi[\|x - x_*\|^2]$$

$\downarrow$
 $V(\cdot)$ is α -strongly convex

⑥ Let $f(x) = V(x)$. Then,

$$0 = E_{\pi}[\Delta f(x)] = E_{\pi}[\Delta V(x) - \|\nabla V(x)\|^2]$$

$$\leq \beta d - E_{\pi}[\|\nabla V(x)\|^2]$$

$\downarrow$
 $V(\cdot)$ is β -smooth

□

ASIDE: Discussion on Gradient Descent

Recall standard **gradient descent** for $V(\cdot)$ α -strongly convex and β -smooth; with iterates, $\{x_0, x_1, \dots\}$, i.e.,

$$x_{k+1} = x_k - h \nabla V(x_k), \quad h > 0, \\ k = 0, 1, 2, \dots$$

For any $y \in \mathbb{R}^d$,

$$\|x_{k+1} - y\|^2 \leq (1 - \alpha h) \|x_k - y\|^2 \\ - 2h (V(x_k) - V(y)) + h^2 \|\nabla V(x_k)\|^2$$

Further V is β -smooth, hence for any $h \leq 1/\beta$,

$$\|\nabla V(x_k)\|^2 \leq \frac{2}{h} (V(x_k) - V(x_{k+1})), \text{ and}$$

$$\|x_{k+1} - y\|^2 \leq (1 - \alpha h) \|x_k - y\|^2 - 2h (V(x_{k+1}) - V(y)).$$

$\therefore y = x_*$, and $h = 1/\beta$ implies:

$$\|x_N - x_*\| \leq \varepsilon \text{ in } N \leq O\left(\frac{\alpha}{\beta} \ln\left(\frac{\|x_0 - x_*\|}{\varepsilon}\right)\right)$$

iterations.

2-Wasserstein distance: μ, ν are prob measures over $\mathbb{R}^d$, i.e., $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

$W_2(\mu, \nu)$ is the 2-Wasserstein distance, defined as:

$$W_2^2(\mu, \nu) = \inf \int \|x - y\|^2 \gamma(dx, dy)$$

$$\left\{ \begin{array}{l} \gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d), \\ \mu(A) = \int_A \int \gamma(dx, dy) \\ \nu(B) = \int \int_B \gamma(dx, dy) \end{array} \right\}$$

joint dist over $x \sim \mu, y \sim \nu$

hence forth:

$C(\mu, \nu) \rightarrow$

i.e., minimize the average squared Euclidean distance between x, y , over all joint dist over (x, y) s.t. marginal of X is μ , and marginal of Y is ν .

We can characterize $\frac{1}{2} W_2^2(\mu, \nu)$ using the dual formulation:

Define $D(\mu, \nu) = \left\{ f(x), g(y) : x \in \mathbb{R}^d, y \in \mathbb{R}^d, \int |f| d\mu < \infty, \int |g| d\nu < \infty, \right.$

$$\left. \begin{aligned} f(x) + g(y) &\leq \frac{\|x - y\|^2}{2} \\ \forall x, y &\quad \mu \text{ a.s.}, \nu \text{ a.s.} \end{aligned} \right\}$$

These are the dual feasible functions.

Then, from the fundamental theorem of optimal transport (see Thm 1.3.8 in [1]), we have

$$\frac{1}{2} W_2^2(\mu, \nu) = \sup_{(f, g) \in D(\mu, \nu)} \left(E_{\mu} [f(x)] + E_{\nu} [g(y)] \right)$$

OT Duality

Furthermore, the optimal dual pair for the above, (f_*, g_*) can be explicitly characterized. When we later discuss flow models, we will revisit optimal transport (to discuss the optimal transport problem and Brenier's thm).

Back to LMC: We now consider the measure induced by LD and that by LMC. Note that the stat. measure $\pi(x)$ for LD is $\frac{e^{-V(x)}}{Z}$, where

Z is the normalizing constant. Suppose $X_0 \sim \pi$, then $\forall t$, $X_t \sim \pi$ for the LD process:

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t, \quad t \geq 0.$$

For the LMC discrete time evolution,

$$\tilde{X}_{(k+1)h} = \tilde{X}_{kh} - h \nabla V(\tilde{X}_{kh}) + \sqrt{2} (B_{(k+1)h} - B_{kh})$$

define μ_{kh} by: $\tilde{X}_{kh} \sim \mu_{kh}$.

Theorem 1 (Thm 4.1.2 in [1]): V s.t.

$$\alpha \text{Id} \preceq \nabla^2 V \preceq \beta \text{Id}, \text{ and}$$

$h \leq \frac{1}{\beta K}$, where $K = \left(\frac{\alpha}{\beta}\right)$. Then, for any $K \geq 0$,

$$W_2(\mu_{kh}, \pi) \leq e^{-\frac{\alpha kh}{2}} W_2(\mu_0, \pi) + O\left(\frac{\beta \sqrt{dh}}{\alpha}\right)$$

Further, for $h \approx \frac{\varepsilon^2}{\beta K d}$, and $\varepsilon \in [0, \sqrt{d}]$, we have:

$$\sqrt{d} W_2(\mu_{Nh}, \pi) \leq \varepsilon \text{ after } N = O\left(\frac{K^2 d}{\varepsilon^2} \ln\left(\frac{\sqrt{d} W_2(\mu_0, \pi)}{\varepsilon}\right)\right) \text{ iterations.}$$

Aside: If we have access to $x^* = \arg \min_x V(x)$, then, choosing $\mu_0 = \delta_{x^*}$ results in $\sqrt{d} W_2(\mu_0, \pi) \leq \sqrt{d}$
(see [1], Remark 4.1.4 for more discussion.)

Before we prove Theorem 1, we show the following intermediate results.

Lemma 2 (Gronwall's Lemma): $g : [0, T] \mapsto \mathbb{R}$ a differentiable function, s.t. $\exists c > 0$ with

$$g'(t) \leq c g(t) \quad \forall t \in [0, T]$$

Then: $g(t) \leq g(0) e^{ct} \quad \forall t \in [0, T]$

PF: See Lemma 1.2.20 in [1]. $\square$

Note: We sometimes see the integral form: Suppose that $g(t) \leq b + c \int_0^t g(s) ds$. Then, $g(t) \leq b e^{ct}$, $t \geq 0$.

Lemma 3: Let $\{Z_t, t \geq 0\}$ be the LD with initial condition $Z_0 \sim \pi_0$ (note, this is not initialized with the stat. measure). Let π_t be the measure induced by Z_t . (i.e., density of Z_t). Further, suppose that $\nabla V(x)$ is β -Lipschitz (V is β -smooth; note we are not using α -strong convexity here). Then, for any $0 \leq t \leq \frac{1}{3\beta}$,

$$E[\|Z_t - Z_0\|^2] \leq 8t^2 E[\|\nabla V(Z_0)\|^2] + 8dt.$$

Proof of Lemma 3:

$$\begin{aligned} E[\|Z_t - Z_0\|^2] &= E\left[\left\| -\int_0^t \nabla V(Z_s) ds + \sqrt{2} B_t \right\|^2\right] \\ &\stackrel{(a+b)^2 \leq 2a^2 + 2b^2}{\leq} 2E\left[\left\| -\int_0^t \nabla V(Z_s) ds \right\|^2\right] + 4 \underbrace{E[\|B_t\|^2]}_{dt} \end{aligned}$$

Aside:
$$\begin{aligned} E\left[\left\| \int_0^t f(Z_s) ds \right\|^2\right] &= E\left[t^2 \cdot \left\| \frac{1}{t} \int_0^t f(Z_s) ds \right\|^2\right] \\ &= t^2 \cdot E_Z \left[\left\| E_{\tau \sim \text{Unif}[0,t]} [f(Z_\tau)] \right\|^2 \right] \end{aligned}$$

$$E[R^2] \geq (E[R])^2 \leq t^2 E_Z [E_T [\|F(z_T)\|^2]]$$

$$= t^2 E_Z \left[\frac{1}{t} \int_0^t \|F(z_s)\|^2 ds \right]$$

$$= t E_Z \left[\int_0^t \|F(z_s)\|^2 ds \right]$$

$$\therefore E[\|z_t - z_0\|^2] \leq 2t \int_0^t E[\|\nabla V(z_s)\|^2] ds + 4dt$$

$V(\cdot)$ β -smooth $\Leftrightarrow \nabla V(\cdot)$ is β -Lipschitz.

$$\therefore \|\nabla V(z_s)\| \leq \|\nabla V(z_0)\| + \beta \|z_s - z_0\|.$$

$$\Rightarrow \leq 4\beta^2 t \int_0^t E[\|z_s - z_0\|^2] ds + 4t^2 E[\|\nabla V(z_0)\|^2] + 4dt$$

∴ From integral form of Gronwall's Lemma (Lem 2),

$$E[\|z_t - z_0\|^2] \leq (4t^2 E[\|\nabla V(z_0)\|^2] + 4dt) e^{4\beta^2 t^2}$$

$\hookrightarrow \leq 1/3\beta$
 $\leq 1/3\beta$ 

Lemma 4: Suppose that the LMC and the LD are both initialized with the same initial condn $X_0 \sim \mu_0$. Then,

$$W_2^2(\mu_h, \pi_h) \leq 3\beta^2 h^4 E[\|\nabla V(X_0)\|^2] + 4\beta^2 d h^3$$

where:

$$\tilde{X}_h = X_0 - h \nabla V(X_0) + \sqrt{2} B_h, \quad \left(\begin{array}{l} \text{1-step} \\ \text{of LMC} \end{array} \right)$$

$\downarrow \mu_h$ $\downarrow \mu_0 = \pi_0$ (same IC) $\updownarrow$ same B.M.

$$X_h = X_0 - \int_0^h \nabla V(X_s) ds + \sqrt{2} B_h, \quad \left(\begin{array}{l} h \text{ time interval} \\ \text{of LD} \end{array} \right)$$

$\downarrow \pi_h$

Proof: $X_0 = \tilde{X}_0$ (coupled IC), both from dist. μ_0 .

Now, observe that by choosing $X_0 \sim \mu_0$, we generate a joint dist γ_h over $(\tilde{X}_h, X_h)$ such that by construction, $\tilde{X}_h \sim \mu_h$, and $X_h \sim \pi_h$.

$$\begin{aligned} \therefore W_2^2(\mu_h, \pi_h) &= \inf_{\gamma \in C(\mu_h, \pi_h)} \int \|x - y\|^2 \gamma(dx, dy) \\ &\leq \int \|x - y\|^2 \gamma_h(dx, dy) \end{aligned}$$

$$= E_{x_0 \sim \mu_0} [\|\tilde{X}_h - X_h\|^2].$$

$$\leq E_{x_0 \sim \mu_0} \left[\left\| \int_0^h \nabla V(X_s) ds - h \nabla V(X_0) \right\|^2 \right]$$

see Aside
in Lemma 3

$$\leq h \int_0^h E_{x_0 \sim \mu_0} [\|\nabla V(X_s) - \nabla V(X_0)\|^2] ds$$

$\nabla V(\cdot)$ is
 β -Lipschitz

$$\leq \beta^2 h \int_0^h E_{x_0 \sim \mu_0} [\|X_s - X_0\|^2] ds$$

$$\hookrightarrow \leq 8h^2 E [\|\nabla V(X_0)\|^2] + 8dt$$

from Lemma 3

Lemma 3

$$\leq 3\beta^2 h^4 E [\|\nabla V(X_0)\|^2] + 4\beta^2 dh^3 \quad \square$$

Proof of Theorem 1: We use the same Brownian motion B_t to drive both the LD and the LMC.

We recursively construct a coupling between μ_{kh} and

π_{kh} , where we use the same Brownian motion to drive both the LMC and the LD. We initialize the LD with its stat. dist $\pi \propto e^{-V(\cdot)}$; thus, $\pi_t = \pi \forall t \geq 0$ for the LD. For the LMC, we initialize with $\tilde{X}_0 \sim \mu_0$.

We initially create a coupling between $\pi_0 = \pi$ and μ_0 ; this joint is denoted by γ_0 . This corresponds to the optimal joint that leads to the initial 2-Wasserstein distance of $W_2(\mu_0, \pi)$. We push this coupling forward using the common BM from time $0 \rightarrow h$. We then reset the coupling at time h (Π^h at times kh) and analyze blocks of time $[kh, (k+1)h]$, with the coupling at beginning times kh being optimal w.r.t $W_2(\cdot, \cdot)$. Thus the analysis below propagates forward in the expectations but not along the sample path.

$$\tilde{X}_{(k+1)h} = \tilde{X}_{kh} - h \nabla V(X_{kh}) + \sqrt{2} (B_{(k+1)h} - B_{kh}), \text{ (LMC)}$$

$$X_t = X_{kh} - \int_{kh}^t \nabla V(X_s) ds + \sqrt{2} (B_t - B_{kh}), \text{ (LD)}$$

$$t \in [kh, (k+1)h]$$

When evolved over the time interval $t \in [kh, (k+1)h]$, the processes differ in two ways: (i) the dynamics are different as above, and (ii) the initial condition at time $t=kh$ differ ($\tilde{X}_{kh}$ vs. X_{kh}).

We thus introduce a third dynamic that evolves like the LD, but is initialized according to the LMC:

$$\bar{X}_t = \tilde{X}_{kh} - \int_{kh}^t \nabla V(\bar{X}_s) ds + \sqrt{2} (B_t - B_{kh})$$

$$t \in [kh, (k+1)h].$$

(Auxiliary Langevin Dynamics - ALD)

Then, we have (analogous to the Proof in Lemma 4)

$$W_2(\mu_{(k+1)h}, \pi) = \sqrt{\inf_{\rho \in C(\mu_{(k+1)h}, \pi)} E_{\rho} [\|\tilde{X}_{(k+1)h} - X_{(k+1)h}\|^2]}$$

$$\leq \sqrt{E_{\hat{\mathbb{P}}_{k+1}} [\|\tilde{X}_{(k+1)h} - X_{(k+1)h}\|^2]}$$

where (abusing notation) $\hat{\mathcal{P}}_{k+1}$ is the joint dist. over the triplet $(\tilde{X}_{(k+1)h}, X_{(k+1)h}, \bar{X}_{(k+1)h})$ that is induced by the initial conditions of $(\tilde{X}_{kh}, X_{kh}, \bar{X}_{kh})$ and the common Brownian motion driving these three processes. Below, we use $\hat{\mathcal{P}}_{k+1}$ to also denote the marginal over any single or pair of these three processes. Note that we pushed the coupling forward in time from $t=kh$ to $t=(k+1)h$.

$$\begin{aligned} \text{Thus, } W_2(\mu_{(k+1)h}, \pi) &\leq \sqrt{E_{\hat{\mathcal{P}}_{k+1}}[\|\tilde{X}_{(k+1)h} - X_{(k+1)h}\|^2]} \\ &\leq \sqrt{E_{\hat{\mathcal{P}}_{k+1}}[\|\tilde{X}_{(k+1)h} - \bar{X}_{(k+1)h}\|^2]} + \sqrt{E_{\hat{\mathcal{P}}_{k+1}}[\|\bar{X}_{(k+1)h} - X_{(k+1)h}\|^2]} \\ &\quad \text{common IC (i)} \qquad \qquad \qquad \text{common dynamics (ii)} \end{aligned}$$



(ii) above corresponds to a pair of LD that are coupled by a common BM, but have different initial conditions. Also $V(\cdot)$ is α -strongly convex and β -smooth.

We use the following fact (for proof, see Theorem 1.4.10 in [1])

Fact (Thm 1.4.10): Suppose $\nabla^2 V \geq \alpha \text{Id}$, and (Z_t, Z'_t) , $t \geq 0$ are two Langevin Dynamics with common B_t , but initial conditions (Z_0, Z'_0) respectively, then,

$$E[\|Z_t - Z'_t\|^2] \leq e^{-2\alpha t} E[\|Z_0 - Z'_0\|^2].$$

Discussion: The proof in [1] uses the interpretation of LD as a Wasserstein Gradient Flow, develops optimal transport theory over Riemannian manifolds (Brenier's thm, Otto-Villani Thm, etc.), the LSI through this approach leading to the above result.

I believe we can do the same using the M.P. approach we studied to derive the LSI for KL divergence (see Note 7), and use the connection between KL divergence and 2-Wasserstein distance for processes

when LSI constant is α , $K_L(\mu \parallel \pi) \leq \frac{\alpha^2}{2} W_2^2(\mu, \pi)$.

(Known as Talagrand's T_2 transportation cost inequality;
also see notes on this in the advanced prob. class).

Back to proof of Theorem 1 (Bounding (ii)).

$$(ii)^2 = E_{\hat{P}_{kh}}[\|\bar{X}_{(k+1)h} - X_{(k+1)h}\|^2] \stackrel{\text{Fact above}}{\leq} e^{-2\alpha h} E_{\hat{P}_k}[\|\bar{X}_k - X_k\|^2]$$

common dynamics

optimally coupled
at start of each block of time; see pp. 16

$$\stackrel{\text{Fact above}}{=} e^{-2\alpha h} \cdot W_2^2(\mu_{kh}, \pi)$$

$$(i)^2 = E_{\hat{P}_{kh}}[\|\tilde{X}_{(k+1)h} - \bar{X}_{(k+1)h}\|^2] \stackrel{\text{Lemma 4}}{\leq} 3\beta^2 h^4 E_{\hat{P}_k}[\|\nabla V(X_k)\|^2]$$

common IC

$$+ 4\beta^2 d h^3$$

$V(\cdot)$ is
 β -Lipschitz

$$\leq 3\beta^4 h^4 E_{\hat{P}_k}[\|X_k - \tilde{X}_k\|^2] + 3\beta^2 h^4 E_{\pi}[\|\nabla V(x)\|^2]$$

$$+ 4\beta^2 d h^3$$

$V(\cdot)$ is
 β -Lipschitz;
 Lemma 1 (b)

$$\leq 3\beta^4 h^4 W_2^2(\mu_{kh}, \pi) + 3\beta^3 h^4 d + 4\beta^2 d h^3$$

Finally, from $\textcircled{*}$ (pp. 18), we put these two bound together (and also for $a, b \geq 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$)

$$W_2(\mu_{(k+1)h}, \pi) \leq e^{-\alpha h} W_2(\mu_{kh}, \pi) +$$

(h small; $h \ll 1/\beta$).

$$O(\beta^2 h^2 W_2(\mu_{kh}, \pi) + \beta \sqrt{d} h^{3/2})$$

Further assuming $h \ll \alpha/\beta^2$, $e^{-\alpha h} + O(\beta^2 h^2) \leq e^{-\alpha h/2}$

$$\therefore W_2(\mu_{(k+1)h}, \pi) \leq e^{-\alpha h/2} W_2(\mu_{kh}, \pi) + O(\beta \sqrt{d} h^{3/2}).$$

Iterate this recursive coupling bound to get:

$$W_2(\mu_{Nh}, \pi) \leq e^{-\frac{\alpha Nh}{2}} W_2(\mu_0, \pi) + O\left(\frac{\beta \sqrt{d} h}{\alpha}\right)$$

$\square$