Statistics 291: Lecture 9 (February 20th, 2024)

Langevin dynamics

Instructor: Mark Sellke

Scribe: Xiaodong Yang

1 Langevin dynamics

In today's lecture, we will switch our attention a little bit to Langevin dynamics. This is widely used for physical time evolution and is a candidate sampling algorithm. In general, we will prove the fast mixing of Langevin dynamics for strongly log-concave distributions, or spherical spin glasses at high temperature. Next class will discuss all temperature, stability and concentration on short time scales.

Definition 1.1 (Langevin dynamics on $\mathbb{R}^N$)**.** Given a Hamiltonian function $H_N : \mathbb{R}^N \to \mathbb{R}$, the corresponding Langvein dynamics is a continuous-time stochastic process which solves the following Ito equation,

$$\mathrm{d}X_t = \sqrt{2}\mathrm{d}B_t + \beta\nabla H(X_t)\mathrm{d}t.$$

Existence and uniqueness of its solutions are readily available as long as ∇H is Lipschitz and the initial state X_0 is determined.

A few remarks follow immediately.

- The Hamiltonian is strongly log-concave if $-O(1)I_N \le \nabla^2 H(x) \le -\lambda I_N$ for all $x \in \mathbb{R}^N$.
- Langevin dynamics can also be defined just using B_t through ODEs. To be preciser, given B_t, let Y_t solve the ODE,
$$\dot{Y}_t = \beta\nabla H(Y_t + \sqrt{2}B_t),$$
then Y_t coincides with $X_t - \sqrt{2}B_t$.
- The stationary measure of this continuous-time Markov process is the Gibbs measure $\mu_\beta(\mathrm{d}x) = \frac{e^{\beta H(x)}\mathrm{d}x}{Z(\beta)}$. With $X_0 \sim p_0(x)$, denote $p_t(x)$ as the law of X_t given by Langevin dynamics. Then Fokker-Planck equation implies
$$\left.\frac{\mathrm{d}p_t(x)}{\mathrm{d}t}\right|_{t=0} = -\mathrm{div}_x\left[\beta\nabla H(x)p_0(x)\right] + \Delta_x[p_0(x)],$$
where $\Delta_x = \sum_{i=1}^N \left(\frac{\partial}{\partial x_i}\right)^2$ corresponding to the heat kernel in heat equations. The first divergence term can be further expanded to
$$-\mathrm{div}_x\left[\beta\nabla H(x)p_0(x)\right] = -\mathrm{div}_x\left(\beta\nabla H(x)\right)p_0(x) - \beta\langle\nabla H(x), \nabla p_0(x)\rangle,$$

where the second term can be understood as a transport term. If we plug in $p_t = \mu_\beta$,

$$\Delta_x \left[\frac{e^{\beta H(x)}}{Z(\beta)} \right] = \frac{1}{Z(\beta)} \mathrm{div}_x \left(\nabla_x e^{\beta H(x)} \right) = \frac{1}{Z(\beta)} \mathrm{div}_x \left(\beta e^{\beta H(x)} \nabla_x H(x) \right).$$

Therefore, μ_β indeed solves the Fokker-Planck equation $\frac{\mathrm{d}p_t(x)}{\mathrm{d}t}\Big|_{t=0} = 0$, so we should believe it to be the stationary measure.

1.1 Fast mixing for strongly log-concave distributions

We firstly define the metric of the Wasserstein 2-distance, which quantifies mixing rates.

Definition 1.2. For any two probability measures $\nu, \tilde{\nu}$ on $\mathbb{R}^N$, we define the Wasserstein 2-distance to be

$$W_2(\nu, \tilde{\nu}) = \sqrt{\inf_{\pi} \mathbb{E}^{\pi} \|X - Y\|^2},$$

where the infimum is taken over all possible couplings $\pi(x, y)$, which are defined as probability measures on $(\mathbb{R}^N)^2$ with corresponding marginals $\pi_x = \nu, \pi_y = \tilde{\nu}$. The optimal π is called “optimal transport coupling".

Theorem 1.3. *Let $H : \mathbb{R}^N \to \mathbb{R}$ be λ-strongly concave. Fix $x_0, y_0 \in \mathbb{R}^N$, and run Langevin dynamics with shared B_t respectively,*

$$\begin{aligned} \mathrm{d}X_t &= \sqrt{2}\mathrm{d}B_t + \beta \nabla H(X_t)\mathrm{d}t, \quad X_0 = x_0, \\ \mathrm{d}Y_t &= \sqrt{2}\mathrm{d}B_t + \beta \nabla H(Y_t)\mathrm{d}t, \quad Y_0 = y_0. \end{aligned} \tag{1}$$

Then for any $t \geq 0$, there holds

$$W_2 \left[\mathrm{law}(X_t), \mathrm{law}(Y_t)\right] \leq \mathbb{E} \|X_t - Y_t\|^2 \leq e^{-2\lambda t} \|X_0 - Y_0\|^2.$$

To prove this fast mixing result, we incorporate the following two standard lemmas from convex optimization.

Lemma 1.4. *If $H : \mathbb{R}^N \to \mathbb{R}$ is λ-strongly concave, then for any $x, y \in \mathbb{R}^N$,*

$$H(x) - H(y) \leq \langle x - y, \nabla H(y) \rangle - \frac{\lambda \|x - y\|^2}{2}.$$

Proof. If $\lambda = 0$, concavity directly implies $H(x) \leq H(y) + \langle x - y, \nabla H(y) \rangle$. If not, we find $H(x) + \frac{\lambda}{2}\|x\|^2$ to be concave, and

$$\begin{aligned} & H(x) + \frac{\lambda}{2}\|x\|^2 \leq H(y) + \frac{\lambda}{2}\|y\|^2 + \left\langle x - y, \nabla H(y) + \lambda y \right\rangle, \\ \Rightarrow \quad & H(x) - H(y) \leq \langle x - y, \nabla H(y) \rangle - \frac{\lambda \|x - y\|^2}{2}. \end{aligned}$$

□

Corollary 1.5. *Under the same condition, we also have*

$$\langle x - y, \nabla H(x) - \nabla H(y) \rangle \leq -\lambda \|x - y\|^2.$$

Proof. By Lemma 1.4, we have

$$\begin{aligned} H(x) - H(y) &\leq \langle x - y, \nabla H(y) \rangle - \frac{\lambda \|x - y\|^2}{2}, \\ H(y) - H(x) &\leq \langle y - x, \nabla H(x) \rangle - \frac{\lambda \|x - y\|^2}{2}. \end{aligned}$$

Sum these two inequalities up to obtain the final result. □

Proof of Theorem 1.3. By subtracting the two terms in (1), we have

$$\mathrm{d}(X_t - Y_t) = \beta\left(\nabla H(X_t) - \nabla H(Y_t)\right)\mathrm{d}t.$$

Then it follows by chain rule that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|X_t - Y_t\|^2 &= 2\left\langle X_t - Y_t, \frac{\mathrm{d}}{\mathrm{d}t}(X_t - Y_t)\right\rangle \\ &= 2\beta\left\langle X_t - Y_t, \nabla H(X_t) - \nabla H(Y_t)\right\rangle \\ &\le -2\beta\lambda\|X_t - Y_t\|^2.\end{aligned}$$

This is sufficient for Theorem 1.3. □

If $Y_0 \sim \mu_\beta$ is directly taken from the stationary measure, then every Y_t is also distributed as μ_β. After applying Gronwall's inequality, we obtain

$$W_2(\mathrm{law}(X_t), \mu_\beta) \le e^{-\beta\lambda t} W_2(\delta_{X_0}, \mu_\beta).$$

The right handed term is decreasing exponentially with respect to $t \to \infty$, and is referred to "fast mixing".

2 Langvein dynamics for spin glasses

We begin this section by defining spherical Langvein dynamics as follows.

Definition 2.1. Let $H : S_N \to \mathbb{R}$ be a Hamiltonian function. (In most cases it should be naturally defined on the whole space $\mathbb{R}^N$, as the p-spin models.) The corresponding spherical Langvein dynamics is a continuous-time stochastic process which solves the following Ito equation,

$$\mathrm{d}X_t = \sqrt{2}P_{X_t}^{\perp}\mathrm{d}B_t + \left(\beta\nabla_{\mathrm{rad}}H(X_t) - \frac{N-1}{N}X_t\right)\mathrm{d}t. \tag{2}$$

where $P_x^{\perp} = I_N - \frac{xx^\top}{N}$ is a rank-$(N-1)$ projection matrix.

If we apply Ito's formula immediately,

$$\mathrm{d}\|X_t\|^2 = \left[2(N-1) - 2\left\langle X_t, \frac{N-1}{N}X_t\right\rangle\right]\mathrm{d}t = 0\mathrm{d}t.$$

So the dynamics always stay on the sphere. The diffusive term cancels out due to $P_x^{\perp}$. If $H \equiv 0$, this process is referred to as the so-called *spherical* Brownian motion.

Definition 2.2. Suppose H_N is "C-bounded of order 2" in the following sense: there exists a universal constant $C > 0$ such that

— $\sup_{x\in X_N} |H_N(x)| \le CN$;

— $\sup_{x\in X_N} \|\nabla H_N(x)\| \le C\sqrt{N}$;

— $\sup_{x\in X_N} \|\nabla^2 H_N(x)\|_{op} \le C$.

In the case of pure p-spin models, we have seen before that H_N is C-bounded with probability $1 - e^{-N}$ for some fixed C.

2.1 Fast mixing at high temperature

This section discusses fast mixing of (2) at high temperature $\beta \le O(1/C)$. The high-level idea is that the sphere's own curvature already forces the process to mix fast enough. All our conditions (Definition 2.2 and $\beta \le O(1/C)$) are ensuring that the Hamiltonian does not blend the curvature too much.

Lemma 2.3. *For any $x, y \in S_N$, there hold*

$$\|\nabla H_N(x) - \nabla H_N(y)\| \le C\|x-y\|$$
$$\|\nabla_{\text{rad}} H_N(x) - \nabla_{\text{rad}} H_N(y)\| \le 3C\|x-y\|.$$

Proof. The fist part is immediate from the Hessian bound,

$$\begin{aligned}\|\nabla H_N(x) - \nabla H_N(y)\| &= \max_{\|z\|=1} \langle \nabla H_N(x) - \nabla H_N(y), z\rangle \\ &= \max_{\|z\|=1} \int_0^1 \langle y-x, \nabla^2 H_N(ya + x(1-a))\rangle \mathrm{d}a \\ &\le \|y-x\| \cdot C\|z\| = C\|y-x\|.\end{aligned}$$

The second part is then due to $\|P_x^\perp\|_{op} = \|P_y^\perp\|_{op} = 1$ and

$$\|P_x^\perp - P_y^\perp\|_{op} = \frac{1}{N}\left\|xx^\top - yy^\top\right\|_{op} \le \frac{2\|x-y\|}{\sqrt{N}}.$$

Consequently,

$$\|\nabla_{\text{rad}} H_N(x) - \nabla_{\text{rad}} H_N(y)\| \le \|P_x^\perp - P_y^\perp\|_{op}\|\nabla H_N(x)\| + \|P_y^\perp\|_{op}\|\nabla H_N(x) - \nabla H_N(y)\| \le 3C\|x-y\|.$$

□

Finally, we are ready to give the theorem of fast mixing for the spherical Langevin dynamics at high temperature.

Theorem 2.4. *If H_N is C-bounded of order 2 as defined in Definition 2.2 and $\beta \le \frac{1}{30C}$, then using some B_t coupling, we are able to establish that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|X_t - Y_t\|^2 \le -\frac{1}{10}\|X_t - Y_t\|^2, \tag{3}$$
$$W_2\left[\text{law}(X_t), \text{law}(Y_t)\right] \le e^{-t/20}\|X_0 - Y_0\|^2. \tag{4}$$

Proof. Firstly, let's consider the case $H_N \equiv 0$, which is also called spherical Brownian motion. By subtracting the equations, we obtain

$$\mathrm{d}(X_t - Y_t) = \sqrt{2}\left(P_{X_t}^\perp - P_{Y_t}^\perp\right)\mathrm{d}B_t - \frac{N-1}{N}(X_t - Y_t)\,\mathrm{d}t.$$

By Ito's formula,

$$\begin{aligned}\mathrm{d}\|X_t - Y_t\|^2 &= \left[\frac{2\|Y_tY_t^\top - X_tX_t^\top\|_F^2}{N^2} - 2\frac{N-1}{N}\|X_t - Y_t\|^2\right]\mathrm{d}t + 2\sqrt{2}\left\langle P_{X_t}^\perp - P_{Y_t}^\perp, X_t - Y_t\right\rangle \mathrm{d}B_t \\ &\le \left[O\left(\frac{\|X_t - Y_t\|^2}{N}\right) - 2\frac{N-1}{N}\|X_t - Y_t\|^2\right]\mathrm{d}t - 2\sqrt{2}\left(1 - \frac{X_t^\top Y_t}{N}\right)(Y_t + X_t)\,\mathrm{d}B_t.\end{aligned}$$

After taking expectation, it translates to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|X_t - Y_t\|^2 \le -\frac{1}{2}\mathbb{E}\|X_t - Y_t\|^2.$$

To proceed, we add back the gradient term $\beta\nabla H$ and go over the same procedures,

$$\mathrm{d}(X_t - Y_t) = \sqrt{2}\left(P_{X_t}^{\perp} - P_{Y_t}^{\perp}\right)\mathrm{d}B_t + \left[\beta\left(\nabla_{\mathrm{rad}} H_N(X_t) - \nabla_{\mathrm{rad}} H_N(Y_t)\right) - \frac{N-1}{N}(X_t - Y_t)\right]\mathrm{d}t.$$

Then it follows that

$$\begin{aligned}\mathrm{d}\|X_t - Y_t\|^2 = &\left[\frac{2\|Y_tY_t^\top - X_tX_t^\top\|_F^2}{N^2} - 2\frac{N-1}{N}\|X_t - Y_t\|^2 + 2\beta\langle\nabla_{\mathrm{rad}} H_N(X_t) - \nabla_{\mathrm{rad}} H_N(Y_t), X_t - Y_t\rangle\right]\mathrm{d}t \\ &+ 2\sqrt{2}\left\langle P_{X_t}^{\perp} - P_{Y_t}^{\perp}, X_t - Y_t\right\rangle \mathrm{d}B_t \\ \le &\left[O\left(\frac{\|X_t - Y_t\|^2}{N}\right) - 2\frac{N-1}{N}\|X_t - Y_t\|^2 + 6\beta C\|X_t - Y_t\|^2\right]\mathrm{d}t - 2\sqrt{2}\left(1 - \frac{X_t^\top Y_t}{N}\right)(Y_t + X_t)\,\mathrm{d}B_t.\end{aligned}$$

If $\beta \le 1/(30C)$, we should arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\|X_t - Y_t\|^2 \le -\frac{1}{4}\mathbb{E}\|X_t - Y_t\|^2.$$

Some constant dependency can be surely improved, but in spirit, β should be scaled $O(1/C)$. $\square$