Langevin Monte Carlo without log-concavity

Candidacy Exam

Leello Dadi

August 31, 2021

The problem

The goal is to sample from a distribution π on the d dimensional space $\mathbb{R}^d$.

$$p_\pi(x) = \frac{\exp(-f(x))}{Z}$$

- $f \in \mathcal{C}^1$ is called the *potential*.
- We will not assume convexity of f, i.e, log-concavity of p_π.
- We have query access to f and ∇f.

First approach: Random Walk Metropolis

Initiate a random walk at $X_0 \in \mathbb{R}^d$,

$$X_{k+1} = X_k + \sqrt{2\eta} Z_{k+1}$$

where $\eta > 0$, and $(Z_k)_k$ is i.i.d $\mathcal{N}(0, I_d)$.

Query f to apply a Metropolis filter: Accept uphill steps, randomize acceptance of downhill steps.

Inform with the gradient

Initiate a random walk at $X_0 \in \mathbb{R}^d$,

$$X_{k+1} = X_k + \sqrt{2\eta} Z_{k+1} - \eta \nabla f(X_k)$$

where $\eta > 0$, and $(Z_k)_k$ is i.i.d $\mathcal{N}(0, I_d)$.

- Introduced as a "Gradient biasing method" [Rossky et al '1978].
- We will study these iterates for a fixed step-size η, with no Metropolis filter.

Outline

Langevin Monte Carlo (LMC)

For some $X_0 \sim \pi_0$,

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{2\eta} Z_{k+1} \tag{LMC}$$

1. Convergence of LMC [Vempala and Wibisono '19]
2. Why can convergence time be exponential in dimension ? [Tzen, Liang, Raginsky '18]
3. LMC in the real world [Song and Ermon '19]

Convergence of LMC

Defining a measure of success :

- Goal is to measure some expectation $\mathbb{E}_\pi[\phi(X)]$, for ϕ bounded.
- To estimate this quantity using our chain, we need

$$\sup_{\phi \text{ 1−bounded}} |\mathbb{E}[\phi(X_k)] - \mathbb{E}_\pi[\phi(X)]| \leftarrow \text{ small}$$

Convergence of LMC

Defining a measure of success :

- Goal is to measure some expectation $\mathbb{E}_\pi[\phi(X)]$, for ϕ bounded.
- To estimate this quantity using our chain, we need

$$\sup_{\phi \text{ 1−bounded}} |\mathbb{E}[\phi(X_k)] - \mathbb{E}_\pi[\phi(X)]| \leq \sqrt{\frac{1}{2}\mathrm{KL}(X_k \| \pi)}$$

- So we will track the evolution of KL divergence from π.
- Recall that

$$\mathrm{KL}(\nu || \pi) = \int_{\mathbb{R}^d} p_\nu(x) \log \frac{p_\nu(x)}{p_\pi(x)} dx \quad \text{and} \quad I(\nu || \pi) = \int_{\mathbb{R}^d} p_\nu(x) \left\| \nabla \log \frac{p_\nu(x)}{p_\pi(x)} \right\|^2 dx.$$

Evolution of KL from iteration k to $k+1$

- How does $\mathrm{KL}(\mathrm{X}_{k+1} \| \pi)$ relate to $\mathrm{KL}(\mathrm{X}_k \| \pi)$?

Langevin Monte Carlo (LMC)

$$\mathrm{X}_{k+1} = \mathrm{X}_k - \eta \nabla f(\mathrm{X}_k) + \sqrt{2\eta}\mathrm{Z}_{k+1} \tag{LMC}$$

Continuous interpolation of the discrete chain

A single step of LMC is the *exact* solution of the following constant drift Stochastic Differential Equation (SDE) at time $t = \eta$:

$$\begin{cases} dX_t = -\nabla f(\mathrm{X}_k) dt + \sqrt{2} dB_t \\ X_0 = \mathrm{X}_k \end{cases}.$$

where $(B_t)_t$ is a standard Brownian motion.

- Track $t \mapsto \mathrm{KL}(X_t \| \pi)$ from $t = 0$ to $t = \eta$.

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.
- We can track KL using $\frac{d}{dt}\mathrm{KL}(X_t \| \pi)$.

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.
- We can track KL using $\frac{d}{dt}\mathrm{KL}(X_t \| \pi)$.
- Vempala and Wibisono observe that

$$\frac{d}{dt}\mathrm{KL}(X_t \| \pi) \leq -\frac{3}{4} I(X_t \| \pi) + \mathbb{E}\left[\|\nabla f(X_t) - \nabla f(\mathrm{X}_k)\|_2^2\right]$$

where I is the relative Fisher information.

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.
- We can track KL using $\frac{d}{dt}\mathrm{KL}(X_t \| \pi)$.
- Vempala and Wibisono observe that (with L-smoothness)

$$\frac{d}{dt}\mathrm{KL}(X_t \| \pi) \leq -\frac{3}{4} I(X_t \| \pi) + L^2 \mathbb{E}\left[\|X_t - \mathrm{X}_k\|_2^2\right]$$

where I is the relative Fisher information.

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.
- We can track KL using $\frac{d}{dt}\mathrm{KL}(X_t \| \pi)$.
- Vempala and Wibisono observe that (with L-smoothness)

$$\frac{d}{dt}\mathrm{KL}(X_t \| \pi) \leq -\frac{3}{4} I(X_t \| \pi) + L^2 \mathbb{E}\left[\|X_t - \mathrm{X}_k\|_2^2\right] \leq -C\mathrm{KL}(X_t \| \pi) + D$$

- If we had such a bound for some $C, D > 0$, then by applying classic tools (Grönwall), we would obtain

$$\mathrm{KL}(\mathrm{X}_{k+1} \| \pi) \leq e^{-C\eta} \mathrm{KL}(\mathrm{X}_k \| \pi) + \text{bias}$$

Using tools from SDE literature

- We gain access to SDE tools : the evolution of the density p_t of X_t, given by $\frac{\partial}{\partial t} p_t$ is known.
- We can track KL using $\frac{d}{dt}\mathrm{KL}(X_t \| \pi)$.
- Vempala and Wibisono observe that (with L-smoothness)

$$\frac{d}{dt}\mathrm{KL}(X_t \| \pi) \leq -\frac{3}{4} I(X_t || \pi) + L^2 \mathbb{E}\left[\|X_t - \mathrm{X}_k\|_2^2\right] \leq -C\mathrm{KL}(X_t \| \pi) + D$$

- **By assuming that** $-I(X_t || \pi) \leq -C\mathbf{KL}(X_t \| \pi)$, V&W show that it is possible to obtain such a bound, and prove that

$$\mathrm{KL}(\mathrm{X}_{k+1} \| \pi) \leq e^{-C\eta}\mathrm{KL}(\mathrm{X}_k \| \pi) + \text{bias}$$

Main result of VW'19

Assumption: log-Sobolev Inequality

There exists a constant $\alpha > 0$ such that, for any probability measure ν where $\mathrm{KL}(\nu\|\pi) < \infty$, we have

$$\mathrm{KL}(\nu\|\pi) \leq \frac{1}{2\alpha} I(\nu\|\pi) \tag{LSI}$$

The biggest α is called *the log-Sobolev constant* of π.

Assumption: Smoothness

The gradient of f is L-Lipschitz : $\forall x, y, \quad \|\nabla f(x) - \nabla f(z)\| \leq L\|x - y\|$

Evolution of KL at each step

LMC with step size $\eta > 0$ verifies at each step :

$$\mathrm{KL}(\mathrm{X}_{k+1}\|\pi) \leq e^{-\alpha\eta}\mathrm{KL}(\mathrm{X}_k\|\pi) + 6\eta^2 dL^2,$$

The log-Sobolev Inequality

Assumption: log-Sobolev Inequality

There exists a constant $\alpha > 0$ such that, for any probability measure ν where $\mathrm{KL}(\nu\|\pi) < \infty$, we have

$$\mathrm{KL}(\nu\|\pi) \leq \frac{1}{2\alpha} I(\nu\|\pi) \tag{LSI}$$

The biggest α is called *the log-Sobolev constant* of π.

- Convenient tool to show sub-Gaussian concentration (like Hoeffding etc).
- **For this talk**, think of LSI as equivalent to sub-Gaussianity : LSI "$\equiv$" light tails
- Contains all distributions with strongly-convex potentials f [Bakry, Émery '85].
- Stable under bounded perturbations of f [Holley, Stroock '87].

Stability under Lipschitz functions

Figure: Images from BigGAN [Brock, Donahue, Simonyan ICLR'19]

Preservation under Lipschitz functions [VW, Lemma 16]

Let $G : \mathbb{R}^p \to \mathbb{R}^d$ be a differentiable L-Lipschitz function, and π admit a LSI constant α, then $G(X)$ for $X \sim \pi$ admits a LSI constant of at least α / L^2.

aggregated across all devices, rather than a single device as in standard implementations. Spectral Normalization (Miyato et al., 2018) is used in both **G** and **D**, following SA-GAN (Zhang et al., 2018).

Distributions captured by GANs (whose noise input is Gaussian) admit a LSI constant.

Convergence Bound

Convergence bound [VW '19, Theorem 1]

Under $LSI(\alpha)$, to get ϵ close to π in KL, LMC needs a number of steps of the order of

$$O\left(\frac{L^2}{\alpha^2}\frac{d}{\epsilon}\right)$$

given an appropriate starting distribution on X_0 .

- LSI only requires good behavior at the tails.
- Scaling problem parameters, we can transform an optimization problem into a sampling problem.
- Non-convex optimization should cost $\left(\frac{1}{\epsilon}\right)^d$ to ϵ-approximate global minima [Nemirovsky, Yudin '83].

Behavior of the trajectory of LMC

Tzen, Liang, Raginsky study LMC to optimize an empirical risk approximation.

- The potential takes the form

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(x, \xi_i)$$

  where $(\xi_i)_i$ are n i.i.d samples from a data distribution.
- They assume ∇f and $\nabla^2 f$ are Lipschitz, and $\pi \propto e^{-f}$ verifies LSI.
- Using LMC to optimize introduces a parameter β, *the inverse temperature*:

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{\frac{2\eta}{\beta}} Z_{k+1}$$

- With the added parameter β, we are targeting the measure $\pi_\beta \propto e^{-\beta f}$.

Main Result

- Pick a non-degenerate local minimum w of f.
- Pick $\epsilon > 0$ small enough.
- Initialize within an ϵ neighborhood of w.

Main Result of [Tzen, Liang, Raginsky '18]

For any $\delta \in [0, 1]$, for any $T > 0$, there is a choice of $\eta > 0$ scaling with $\frac{1}{T}$ and a choice of β scaling with $\log(T)$ such that

$$\mathbb{P}(\text{Iterates escape } \epsilon \text{ ball around } w \text{ before } \frac{T}{\eta}) \leq \delta$$

- For a time of my choice T, I can set parameters so that LMC is trapped around a local minimum.

How LMC behaves vs What I can make LMC do

- Dependence on T of the parameters feels like cheating. "Choose your own adventure" –Tzen.

How LMC behaves vs What I can make LMC do

- Dependence on T of the parameters feels like cheating. "Choose your own adventure" –Tzen.
- Ideally, β would be set beforehand: choose β so that $\pi_\beta \propto e^{-\beta f}$ is sufficiently concentrated, THEN, ask how long will local minima trap LMC for ?

How LMC behaves vs What I can make LMC do

- Dependence on T of the parameters feels like cheating. "Choose your own adventure" –Tzen.
- Ideally, β would be set beforehand: choose β so that $\pi_\beta \propto e^{-\beta f}$ is sufficiently concentrated, THEN, ask how long will local minima trap LMC for ?
- In the literature [Menz, Schlichting '14], given β (large enough), the LSI constant α_β of π_β verifies
$$\frac{1}{\alpha_\beta} \lesssim \mathbb{E}[\tau]$$
where $\mathbb{E}[\tau]$ is the (worst) mean exit time from one local minimum to another.

How LMC behaves vs What I can make LMC do

- Dependence on T of the parameters feels like cheating. "Choose your own adventure" –Tzen.
- Ideally, β would be set beforehand: choose β so that $\pi_\beta \propto e^{-\beta f}$ is sufficiently concentrated, THEN, ask how long will local minima trap LMC for ?
- In the literature [Menz, Schlichting '14], given β (large enough), the LSI constant α_β of π_β verifies

$$\frac{1}{\alpha_\beta} \lesssim \mathbb{E}[\tau]$$

  where $\mathbb{E}[\tau]$ is the (worst) mean exit time from one local minimum to another.
- *LSI is measuring how hard it is to jump from mode to mode.*

LMC in the real world

Consider π to be a natural distribution, like natural images of a given size.

Langevin Monte Carlo (LMC)

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{2\eta} Z_{k+1} \tag{LMC}$$

- We need the existence of the *score* ∇f.
- No reason to believe the real world admits a density, let alone a positive differentiable one.
- Good news: Convolution with a Gaussian confers all the necessary regularity.

Learning the score of a perturbed distribution

For $X \sim \pi$, $N \sim \mathcal{N}(0, \sigma^2 I)$, we consider π_σ the law of $Y = X + N$.

Score Matching [Song&Ermon '19]

We can parametrize a vector field using a neural network: $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Train it to minimize

$$\min_{\theta \in \Theta} \mathbb{E}_{\pi_\sigma} \left[\| s_\theta(Y) - \nabla \log p_{\pi_\sigma}(Y) \|_2^2 \right]$$

Learning the score of a perturbed distribution

For $X \sim \pi$, $N \sim \mathcal{N}(0, \sigma^2 I)$, we consider π_σ the law of $Y = X + N$.

Score Matching [Song&Ermon '19]

We can parametrize a vector field using a neural network: $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Train it to minimize

$$\min_{\theta \in \Theta} \mathbb{E}_{\pi_\sigma} \left[\| s_\theta(Y) - \nabla \log p_{\pi_\sigma}(Y) \|_2^2 \right]$$

- The true (perturbed) score is unknown.

Learning the score of a perturbed distribution

For $X \sim \pi$, $N \sim \mathcal{N}(0, \sigma^2 I)$, we consider π_σ the law of $Y = X + N$.

Score Matching [Song&Ermon '19]

We can parametrize a vector field using a neural network: $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Train it to minimize

$$\min_{\theta \in \Theta} \mathbb{E}_{\pi_\sigma} \left[\| s_\theta(Y) - \nabla \log p_{\pi_\sigma}(Y) \|_2^2 \right]$$

- The true (perturbed) score is unknown.
- Luckily, with a few interchanges we obtain the *equivalent* loss [Vincent '10]

$$\min_{\theta \in \Theta} \mathbb{E}_{X \sim \pi,\ N \sim \mathcal{N}(0, \sigma^2 I)} \left[\| s_\theta(X + N) - \frac{N}{\sigma^2} \|_2^2 \right]$$

The need for multiple noise scales

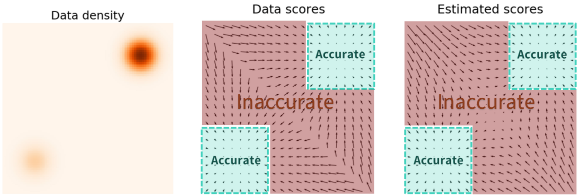


Figure: Inaccurate estimation in low density regions [Song blog]

The need for multiple noise scales

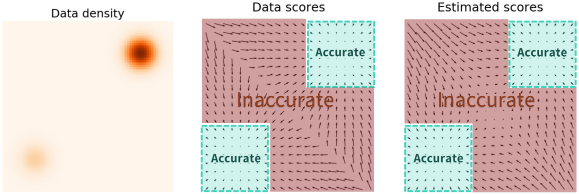


Figure: Inaccurate estimation in low density regions [Song blog]

- Idea [SE '19] : Perturb with a decreasing sequence of noise scales:

$$\sigma_1 > \sigma_2 > \cdots > \sigma_L$$

The need for multiple noise scales

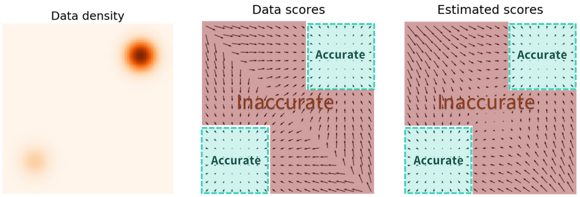


Figure: Inaccurate estimation in low density regions [Song blog]

- Idea [SE '19] : Perturb with a decreasing sequence of noise scales:

$$\sigma_1 > \sigma_2 > \cdots > \sigma_L$$

- Learn the L scores jointly with a *Noise Conditional Neural Score Network*

$$(x, \sigma) \mapsto s_\theta(x, \sigma)$$

such that $x \mapsto s_\theta(x, \sigma_i)$ approximates the score of π_{σ_i}.

Annealed Langevin Dynamics

- At first the only scores I can trust are $x \mapsto s_\theta(x, \sigma_1)$
- Closer to the modes of π_{σ_1} I can start trusting the scores $x \mapsto s_\theta(x, \sigma_2)$, and so on ...

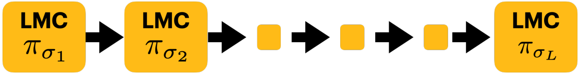


Figure: Annealed Langevin Dynamics [SE '19]

Annealed Langevin Dynamics

- At first the only scores I can trust are $x \mapsto s_\theta(x, \sigma_1)$
- Closer to the modes of π_{σ_1} I can start trusting the scores $x \mapsto s_\theta(x, \sigma_2)$, and so on ...

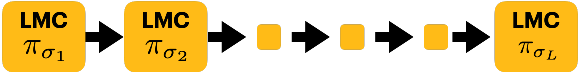
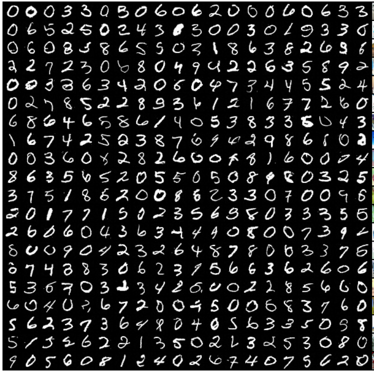


Figure: Annealed Langevin Dynamics [SE '19]

- Each i-th LMC is run for T steps, with a fixed step-size η_i, and $\eta_i = \gamma \eta_{i+1}$ (decreased by a multiplicative factor)

Results



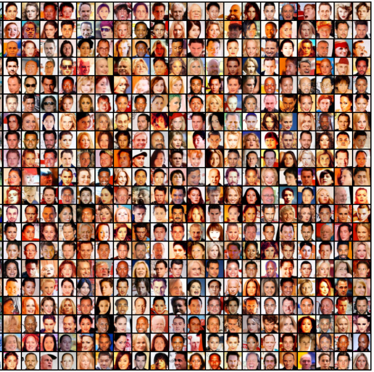


Figure: MNIST samples

Figure: CIFAR-10 samples

Figure: CelebA samples

Quantitative metrics : Achieved the best *Inception Score* on CIFAR-10, a metric that values clarity and coverage of all classes (can be fooled by single ouput per class).

A very small time scale

- The results were achieved for a very small $T = 100$ iterations of LMC per noise scale with $\eta = 10^{-5}$. This *is* small.

A very small time scale

- The results were achieved for a very small $T = 100$ iterations of LMC per noise scale with $\eta = 10^{-5}$. This *is* small.
- Consider simply going from $\mathcal{N}(0, \sigma_1)$ to $\mathcal{N}(0, \sigma_2)$, with the **diffusion**, in **closed form**, we have :

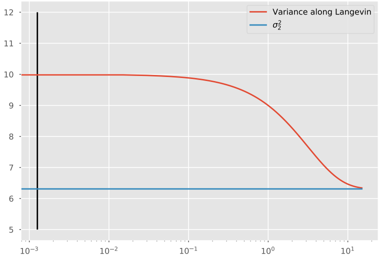


Figure: Evolution of the variance following Langevin Diffusion

Are we even sampling ?

- Isn't outputting the modes enough to fool us ? (Good FID scores suggest otherwise)
- Is Annealed Langevin Dynamics completely different from LMC ?

Are we even sampling ?

- Isn't outputting the modes enough to fool us ? (Good FID scores suggest otherwise)
- Is Annealed Langevin Dynamics completely different from LMC ?
- Is it wrong to think that each LMC block must mix for it to work ?

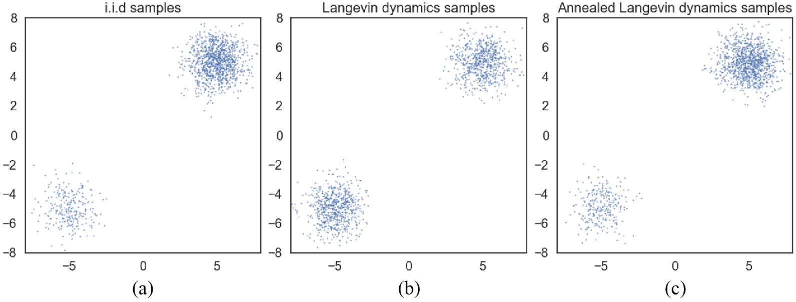


Figure: Annealed Langevin Dynamics does better than LMC on a mixture

Conclusion

- [Vempala, Wibisono '19] We can sample under just LSI and smoothness.
- [Tzen, Liang, Raginsky '18]: But we can be trapped by local minima for a very long time. Depends on the dimension dependence of the LSI constant.
- [Song, Ermon '19]: A Langevin like scheme appears successful.

Question

What can be said about the dimension dependence of the LSI constant of natural images ?

- The modes of π_σ are shallow.
- Manifold hypothesis ? [Block, Mrouef, Rakhlin, Ross '20]

- We are not sampling.
- Annealed Langevin Dynamics is drastically different from LMC.