Generative Models

Saeed Saremi

Assigned reading: 14.3[1]

November 7, 2024

[1] Lecture 1 focuses on the Langevin algorithm for sampling, with the primary source being the lecture notes.

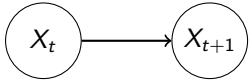
BROAD OUTLINE

Lecture 1: LANGEVIN MCMC uses the SCORE FUNCTION

Lecture 2: Learn the SCORE FUNCTION via DENOISING

OUTLINE

- sampling and generative modeling
  - score-based generative modeling
- Langevin Markov chain Monte carlo
  - Langevin Markov chain



$$X_{t+1}|X_t \sim \mathcal{N}(X_t + h\nabla \log p(X_t), 2h\, I_d)$$

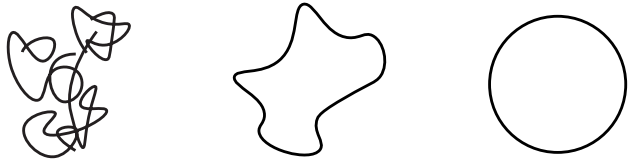
  - random walk
  - 1D and 2D demonstrations of Langevin MCMC
  - some analogies with SGD
- smoothing with Gaussian noise: geometric and algebraic view

- We are given some i.i.d. draws

$$\{x_1, \ldots, x_n\}$$

from an unknown distribution p_X in $\mathbb{R}^d$. We like to draw independent samples from p_X.

- A very effective strategy to tackle this problem is smoothing via Gaussian noise:



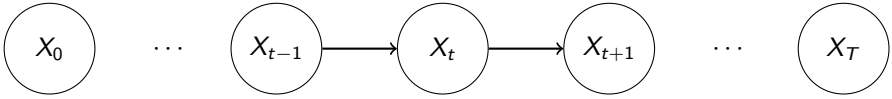
(a) X (b) $Y = X + \sigma N$ (c) $N \sim \mathcal{N}(0, I)$

- The score function, $g = \nabla \log p$, will end up playing a central role.

- We are given a distribution

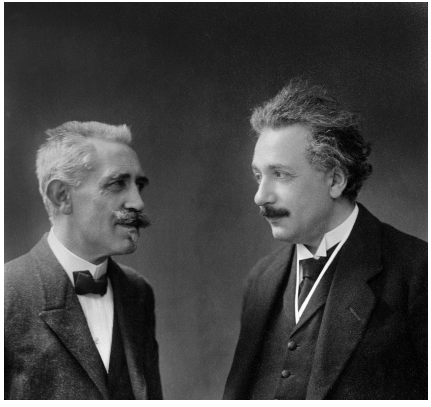
$$p(x) = \frac{e^{-f(x)}}{Z},$$

  where $Z = \int \exp(-f(x))dx$ is the ("impossible to compute") normalizing constant.
- We would like to draw independent samples from p_X.
- An effective strategy is to "cook up" a Markov chain such that it converges to p_X:

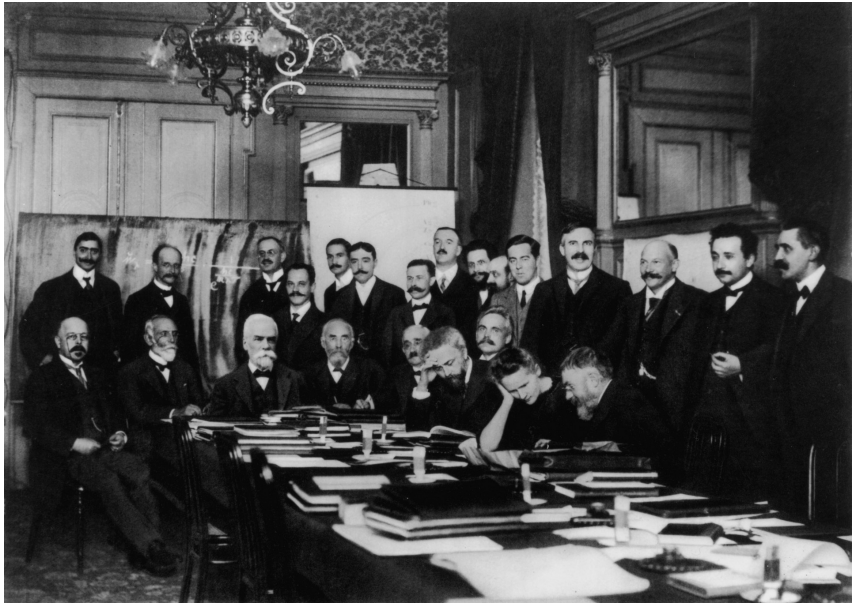


- Therefore for any $X_0 \sim \pi$, $X_T \sim p$ as T becomes large.
  – We say the Markov chain mixes fast if the "large T" is not too large.

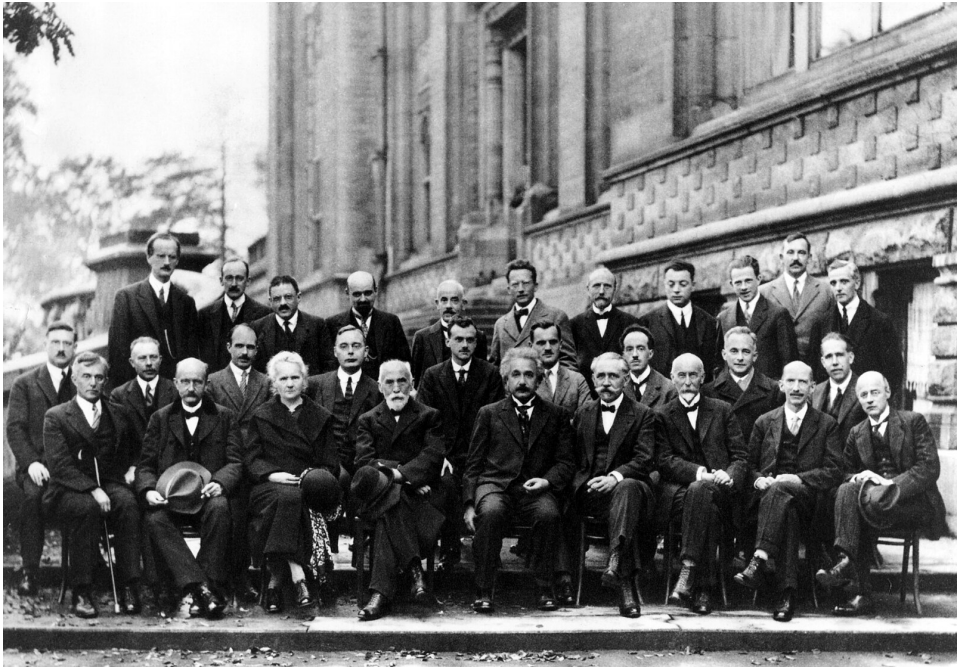
PAUL LANGEVIN (1872 – 1946)



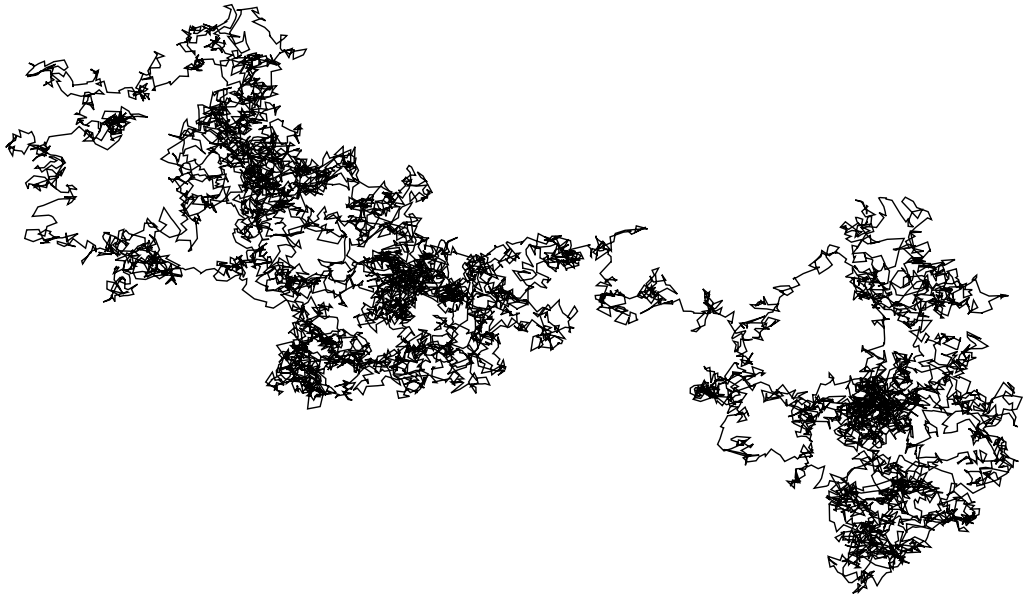
PAUL LANGEVIN (1872 – 1946)



PAUL LANGEVIN (1872 – 1946)

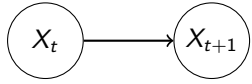


⌨ RANDOM WALK (BROWNIAN MOTION (WIENER PROCESS)), $X_{t+1}|X_t \sim \mathcal{N}(X_t, h\, I_d)$



- The art of MCMC is coming up with "good" transition kernels $p(x_{t+1}|x_t)$.
- One celebrated example is unadjusted[2] Langevin algorithm, where

$$x_{t+1} = x_t + h \underbrace{\nabla \log p(x_t)}_{\text{score function}} + \sqrt{2h}\varepsilon_t, \ \varepsilon_t \sim N(0, I)$$



$$X_{t+1}|X_t \sim \mathcal{N}(X_t + h\nabla \log p(X_t), 2h\, I)$$

- If you run a Langevin chain for a long time it converges to a sample from $p(x)$.

[2]Called "unadjusted" since all moves are accepted. We will not discuss Markov chains with accept/reject!

SCORE FUNCTION[3]

- ALL generative models can be viewed from the perspective of how they deal with the normalizing constant $Z(\theta)$, also call the partition function:

$$p_\theta(x) = \frac{e^{-f_\theta(x)}}{Z(\theta)},$$

where

$$Z(\theta) = \int e^{-f_\theta(x)} dx$$

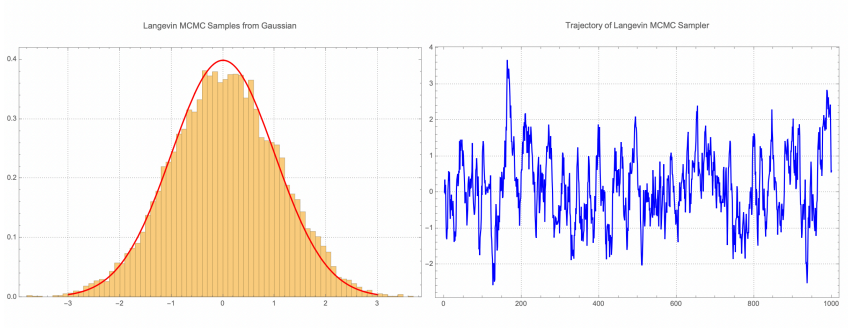
- In score-based generative models we deal with $Z(\theta)$ via the score function:

$$g_\theta(x) = \nabla_x \log p_\theta(x) = -\nabla_x f_\theta(x) - \cancel{\nabla_x \log Z(\theta)}$$

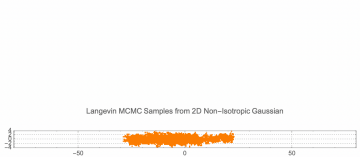
[3] We will return to this in the next lecture.

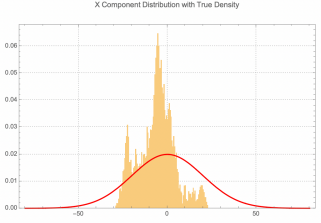
⌨ 1D GAUSSIAN

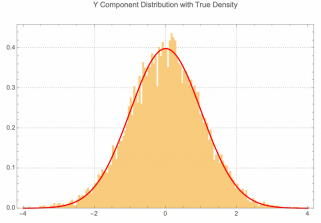
Demonstrate the "tuning" of Langevin MCMC (and the tradeoffs).

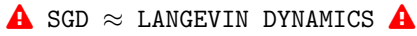


Demonstrate that sampling non-isotropic Gaussian is harder, controlled by $\kappa = (\sigma_{\max}/\sigma_{\min})^2$.








SGD ≈ LANGEVIN DYNAMICS

📢 GAUSSIAN NOISE SMOOTHS ANY DISTRIBUTION, MAKING IT EASIER TO SAMPLE.

(a) X

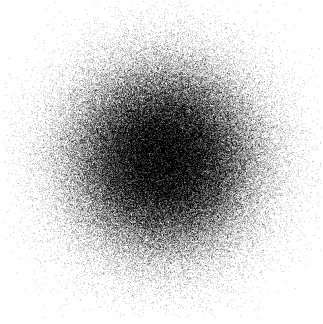
(b) $Y = X + \sigma N$

(c) $\hat{x}(y) = y + \sigma^2 \nabla \log p(y)$

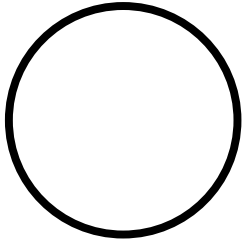
GEOMETRIC VIEW

GAUSSIAN DISTRIBUTION IN HIGH DIMENSIONS

- $X_{1:d} \sim \mathcal{N}(0, \sigma^2 I_d)$ in high dimensions is concentrated on S_{d-1} with radius $\sigma\sqrt{d}$
- one-line proofs:
  - $\|X_{1:d}\|^2 = \sum_{i=1}^{d} X_i^2 = d \cdot \frac{1}{d} \sum_{i=1}^{d} X_i^2 \approx d\,\mathbb{E}[X^2] = d\sigma^2 \Rightarrow \|X_{1:d}\| \approx \sigma\sqrt{d}$
  - `x = torch.randn(n, d).pow(2).sum(1).sqrt()/math.sqrt(d)`



(b) $d = 2$



(c) $d \gg 1$

ISOTROPIC GAUSSIAN NOISE MAKES THE DATA MANIFOLD MORE SPHERICAL.

ALGEBRAIC VIEW

ADDITIVE NOISE = CONVOLUTION

- Consider additive (Gaussian) noise

$$Y = X + N, \text{ where } N \sim \mathcal{N}(0, \sigma^2 I).$$

- Equivalently,

$$y|x \sim \mathcal{N}(x, \sigma^2 I)$$

Show

$$p_Y(y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \int p_X(x)\, exp\Big(-\frac{\|y-x\|^2}{2\sigma^2}\Big)\, dx$$

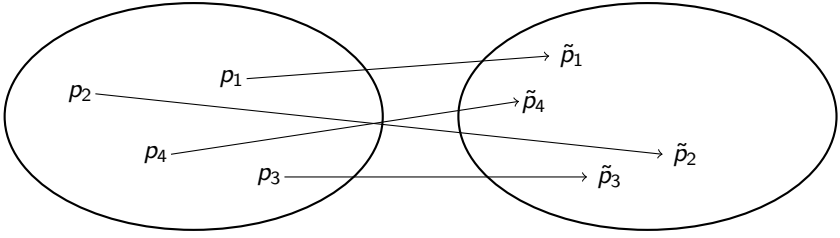
This is written more compactly, using the convolution notation:

$$p_Y = p_X * p_N$$

- Recall the definition of Fourier transform

$$\tilde{p}(\omega) = \int e^{i\omega^\top x} p(x) dx$$

- Fourier transform is a bijection:



- Fourier transform of a probability distribution is named characteristic function.

ADDITIVE GAUSSIAN NOISE (THE FOURIER ANGLE)

- Recall the definition of convolution of two functions $p_3 = p_1 * p_2$:

$$(p_1 * p_2)(y) = \int p_1(x) p_2(y - x) dx$$

- In Fourier space, convolution=multiplication: $\tilde{p}_3(\omega) = \tilde{p}_1(\omega)\tilde{p}_2(\omega)$
- The Fourier transform of the Gaussian distribution is of particular interest!

$$\tilde{p}_N(\omega) = \exp\left(-\frac{\sigma^2}{2}\|\omega\|^2\right)$$

- Gaussian noise $Y = X + N$ has the effect of smoothing out high frequencies in p_X:

$$\tilde{p}_Y(\omega) = \tilde{p}_X(\omega) \exp\left(-\frac{\sigma^2}{2}\|\omega\|^2\right),$$

since for large $\sigma\|\omega\| \gg 1$, we have $\tilde{p}_Y(\omega) \ll \tilde{p}_X(\omega)$.