Sampling

Yin Tat Lee

August 13, 2019

In this lecture, we study the problem of sampling $x \propto e^{-f(x)}$. Namely, sample $x \in \mathbb{R}^d$ according to the distribution $e^{-f(x)} / \int_{\mathbb{R}^d} e^{-f(x)} dx$.

1 Toy Example $d = 1$

For $d = 1$ case, the problem can be solved as follows:

- Sample u from the uniform distribution from $[0, 1]$.
- Let $p(x) = e^{-f(x)} / \int_{-\infty}^{\infty} e^{-f(x)} dx$ and $P(x) = \int_{-\infty}^{x} p(x) dx$.
- Output $P^{-1}(u)$.

Theorem. *The algorithm above correctly samples $x \propto e^{-f(x)}$.*

Proof. Let x be an sample. This is because the cumulative distribution of x is equals to the target cumulative distribution:

$$\mathbf{P}(x \le t) = \mathbf{P}(P^{-1}(u) \le t) = \mathbf{P}(u \le P(t)) = P(t) = \int_{-\infty}^{x} p(x) dx.$$

□

Exercise. Extend the algorithm above to higher dimensions $d > 1$.

This algorithm is inefficient in higher dimension. For many functions, this is not even the best choice for $d = 1$. For Gaussian distribution, there are better algorithms such as Box–Muller transform and Ziggurat algorithm.

2 Why Sampling?

There are many applications for sampling, such as Bayesian statistics, convex geometry, robust convex optimization. First of all, we note that sampling is a more general problem than optimization.

Exercise. Suppose we can sample any function f in polynomial time. Show that we can find the minimizer of any function in polynomial time.

On the other hand, for smooth function, one can sometimes approximate e^{-f} by a Gaussian distribution via expanding f at the minimum. Look at Laplace's method in Wikipedia for more information.

Exercise. Using the fact that $N! = \int_0^{\infty} e^{-x} x^N dx$, shows that $N! \sim \sqrt{2\pi N} N^N e^{-N}$.

2.1 Toy application

Imagine you have a new design of personal website[1], and you want more visitors to download your beloved papers under the new design. As a scientist, you did a comparison test on both versions for few days and see this:

- Old version: Visits 1135. Downloads 5.
- New version: Visits 1149. Downloads 17.

Seems the new version is better. But how can we know if this is just due to random fluctuations and that we need to run a longer test?

Suppose the number of downloads k satisfies the Binomial distribution $B(n, r)$. Recall the Bayes' theorem that

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)}.$$

The event B is we observe the number of downloads k in n downloads and we want to know the probability of event A: the download rate is r. Since we have no prior information about the distribution of r (the download rate), we simply assume r is uniform between $[0, 1]$. Hence, the probability distribution of the download rate r is

$$\begin{aligned}\mathbf{P}(r|k, n) &= \frac{\mathbf{P}(k, n|r)\mathbf{P}(r)}{\mathbf{P}(k, n)} \\ &= \frac{\binom{n}{k} r^k (1-r)^{n-k}}{\int_0^1 \binom{n}{k} r^k (1-r)^{n-k} dp} \\ &\propto r^k (1-r)^{n-k}.\end{aligned}$$

We can simply draw the distribution of $\mathbf{P}(r|k = 5, n = 1135)$ and $\mathbf{P}(r|k = 17, n = 1149)$ and see that $\mathbf{P}(r_{\text{new}} > r_{\text{old}}) > 0.99$. This shows that the new version is indeed better.

For this model, one can do all calculation exactly without sampling. In reality, a model can depends on many parameters. Sampling from the parameters (via the Bayes' fromula) allows you to approximately answer many questions without doing a crazy multi-dimensional integration, which is often computationally infeasible.

Compare to simply finding the best parameters to fit the existing data, sampling gives you a lot more information such as how confident is the result and how much is the variance. For example, Henri did an interesting experiment on this[2]. He trained a neural network on MNIST (a dataset on images from 0 to 9) and give the network some letter to recognize. Instead of returning low confidence result, his neural network gives high confidence that the given letters are some numbers.

There are many researchers/professors whose daily jobs are just creating models and applying sampling techniques to answer questions about data. The goal of this lecture is merely to show you some theory background about sampling.

2.2 Theory Applications

In general, sampling is useful whenever we need to explore the domain. Here are some theory results that crucially relies on the fact we can sample x according to $e^{-f(x)}$ in polynomial time whenever f is convex.

Explore Space

As in the last lecture, we see that bandit problem is about explore and exploit. For simple problems, one can explore the space via just uniformly sample the state space. But when the state space is infinite, it is much more difficult.

[1] This example is modified from http://mathamy.com/using-pymc-to-analyze-ab-testing-data.html
[2] https://henripal.github.io/blog/langevin

Theorem 1 (Convex Bandit[3]). *Given a sequence of convex functions f_i defined on $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ with value between 0 and 1. At step i, we need to output a point x_i and receive the loss $f_i(x_i)$. The only information we are given about f_i is the loss $f_i(x_i)$ received. There is an algorithm that can find a sequence of x_i adaptively such that*

$$\sum_{i=1}^{T} f_i(x_i) \leq \min_x \sum_{i=1}^{T} f_i(x) + O(\sqrt{T}(d \log T)^{O(1)}).$$

Robust to adversary

Another benefit of sampling compared to optimization is that the output is more robust compared to noise and hence useful for optimizing noisy function.

Theorem 2 (Convex Optimization using Function Value[4]). *Given an approximately convex function f. Suppose that there is a convex function g such that $|f(x) - g(x)| \leq \epsilon$. We can find x such that $f(x) \leq \min_x f(x) + O(\epsilon n)$ in polynomial time. Furthermore, $O(\epsilon n)$ is the best possible informatically.*

Extract global information

Sampling is also useful for extracting global information about convex set.

Theorem 3 (Volume Computation[5]). *Given a convex set $K \subset \mathbb{R}^d$. We can compute γ such that $\gamma = (1 \pm \epsilon)\text{vol}K$ in time polynomial in d/ϵ.*

Finally, we note that informatically best algorithm on convex optimization result uses sampling, such as the center of gravity methods, interior point method for universal barrier function.

3 Continuous Langevin Dynamic

In this and next section, we discuss a process for sampling from e^{-f} called Langevin Dynamic. This process can also be viewed as a stochastic version of gradient descent. While gradient descent corresponds to an ordinary differential equation (ODE), stochastic gradient descent corresponds to a stochastic differential equation (SDE):

$$\begin{aligned} dx_t &= -\nabla f(x_t)dt + \sqrt{2}dW_t, \\ x_0 &= \text{initial point}. \end{aligned}$$

Here $f : \mathbb{R}^n \to \mathbb{R}$ is the function in e^{-f} we want to sample, x_t is the random variable at time t and dW_t is infinitesimal Brownian motion. We first discuss the continuous version here only for simplicity. In the next section, we will analyze the discrete version.

3.1 Notation: Basics of SDE

We will first explain a little bit about SDE notation.

Definition 4. Brownian motion as the unique random process satisfying:

- $W_0 = 0$.
- W_t is continuous.
- $W_{t_1} - W_{t_2}$ is independent to $W_{t_2} - W_{t_3}$ for any $t_1 \geq t_2 \geq t_3$.

[3] https://arxiv.org/abs/1607.03084
[4] https://arxiv.org/pdf/1501.07242.pdf
[5] https://dl.acm.org/citation.cfm?id=102783

- $W_t - W_s \sim N(0, t-s)$.

We will also abuse W_t to denote any m dimensional Brownian motion where $W_{t,i}$ is an independent copy of Brownian motion for each $i \in [m]$.

Now, consider SDE of the form

$$dx_t = \mu(x_t)dt + \sigma(x_t)dW_t$$

where $x_t \in \mathbb{R}^d$, $\mu(x_t) \in \mathbb{R}^d$, $\sigma(x_t) \in \mathbb{R}^{d\times m}$ and $dW_t \in \mathbb{R}^m$. We call μdt as the drift term and σdW as the diffusion term. To define it informally, consider the following discrete process

$$\begin{aligned} x_{t+h} &= x_t + \mu(x_t)(t+h-t) + \sigma(x_t)(W_{t+h} - W_t) \\ &= x_t + h\mu(x_t) + \sqrt{h}\sigma(x_t)\zeta_t \end{aligned}$$

with ζ_t sampled independently from $N(0, I)$. When we take the step size $h \to 0$, this discrete process converges to the continuous one.

Beware that the diffusion term scales with $\sqrt{h}$ instead of h. To see why, we consider the simplest SDE:

$$dx_t = dW_t, x_0 = 0.$$

The solution of this SDE is simply $x_t = W_t$. If we *incorrectly* discretize the equation by $x_{t+h} = x_t + h\zeta_t$. Since ζ_t are independent, we have that

$$\mathrm{Var}x_{t+h} = \mathrm{Var}x_t + h^2.$$

Summing up $\frac{1}{h}$ steps, we have

$$\mathrm{Var}x_t = \mathrm{Var}x_0 + \frac{t}{h} \cdot h^2 = th.$$

Taking $h \to 0$, we would get $\mathrm{Var}x_t = 0$. The limit is deterministic because the cancellation of random variables. By a similar argument, one can show that $\sqrt{h}$ is the only scaling that converges to a random process.

A key difference between ODE and SDE is the chain rule. Suppose we have a ODE $\frac{dx_t}{dt} = f(x_t)$. Then, chain rule shows that

$$\frac{dg(x_t)}{dt} = g'(x_t)\frac{dx_t}{dt} = g'(x_t)f(x_t).$$

In comparison, the chain rule for SDE, call Ito's lemma, has one more extra term:

Lemma 5 (Ito's lemma)**.** *For any process $x_t \in \mathbb{R}^d$ satisfying $dx_t = \mu(x_t)dt + \sigma(x_t)dW_t$ where $\mu(x_t) \in \mathbb{R}^d$ and $\sigma(x_t) \in \mathbb{R}^{d\times m}$, we have that*

$$df(x_t) = \nabla f(x_t)^\top \mu(x_t)dt + \nabla f(x_t)^\top \sigma(x_t)dW_t + \frac{1}{2}\mathrm{tr}(\sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t))dt.$$

Proof: (Intuition only). Consider

$$x_{t+h} = x_t + h\mu(x_t) + \sqrt{h}\sigma(x_t)\zeta_t.$$

Then, we have

$$\begin{aligned} f(x_{t+h}) =& f(x_t) + \nabla f(x_t)^\top (x_{t+h} - x_t) + \frac{1}{2}(x_{t+h} - x_t)^\top \nabla^2 f(x_t)(x_{t+h} - x_t) \\ =& f(x_t) + h\nabla f(x_t)^\top \mu(x_t) + \sqrt{h}\nabla f(x_t)^\top \sigma(x_t)\zeta_t \\ &+ \frac{h}{2}\zeta_t^\top \sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t)\zeta_t + O(h^{1.5}). \end{aligned}$$

For the terms involving h, we only care about its expected value. The variance of h related term are too small at the limit. Note that

$$\begin{aligned} \mathbf{E}\zeta_t^\top \sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t)\zeta_t &= \mathbf{E}\mathrm{tr}\sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t)\zeta_t\zeta_t^\top \\ &= \mathrm{tr}\sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t). \end{aligned}$$

Hence, we have

$$\begin{aligned} f(x_{t+h}) =& f(x_t) + h\left(\nabla f(x_t)^\top \mu(x_t) + \frac{1}{2}\mathrm{tr}\sigma(x_t)^\top \nabla^2 f(x_t)\sigma(x_t)\right) \\ &+ \sqrt{h}\nabla f(x_t)^\top \sigma(x_t)\zeta_t + O(h^{1.5}) + O(h \text{ mean 0 terms}) \end{aligned}$$

This explains the formula when we take $h \to 0$. □

Exercise 6. (for students who do options trading). Assume the stock price S follows a geometric Brownian motion, namely

$$dS = \mu S dt + \sigma S dW_t.$$

Derive the fair pricing for an European option. Hints: a pricing is correct if one can eliminate the risk of the option by buying and selling the underlying asset in some right way.

See the Black–Scholes equation in the Wikipedia for the answer for the above exercise. As with the chain rule, there are numerous other applications of Ito's lemma.

3.2 Correctness: Fokker–Planck equation

Now, we show that this process converges to e^{-f} in continuous time. The proof relies on the following general theorem about the distribution induced by an SDE.

Theorem 7 (Fokker–Planck equation)**.** *For any process $x_t \in \mathbb{R}^n$ satisfying $dx_t = \mu(x_t)dt + \sigma(x_t)dW_t$ where $\mu(x_t) \in \mathbb{R}^n$ and $\sigma(x_t) \in \mathbb{R}^{n\times m}$ with the initial point x_0 drawn from p_0. Then, the distribution p_t of x_t satisfies the equation*

$$\frac{dp_t}{dt} = -\sum_i \frac{\partial}{\partial x_i}(\mu(x)_i p_t(x)) + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[(D(x))_{ij}p_t(x)\right]$$

where $D(x) = \sigma(x)\sigma(x)^\top$.

Proof. For any smooth function ϕ, we have that

$$\mathbf{E}_{x\sim p_t}\phi(x) = \mathbf{E}\phi(x_t).$$

Taking derivatives on the both sides with respect to t and using Itô's lemma (Lemma 5), we have that

$$\begin{aligned} \int \phi(x)dp_t(x)dx &= \mathbf{E}\left(\nabla\phi(x_t)^\top\mu(x_t)dt + \nabla\phi(x_t)^\top\sigma(x_t)dW_t + \frac{1}{2}\mathrm{tr}(\sigma(x_t)^\top\nabla^2\phi(x_t)\sigma(x_t))dt\right) \\ &= \mathbf{E}\left(\nabla\phi(x_t)^\top\mu(x_t)dt + \frac{1}{2}\mathrm{tr}(\nabla^2\phi(x_t)D(x_t))dt\right). \end{aligned}$$

Using $x_t \sim p_t$, we have that

$$\int \phi(x)\frac{dp_t}{dt}dx = \int \nabla\phi(x)^\top\mu(x)p_t(x) + \frac{1}{2}\mathrm{tr}(\nabla^2\phi(x)D(x))p_t(x)dx.$$

Integrating by parts,

$$\int \nabla\phi(x)^\top\mu(x)p_t(x)dx = -\int \phi(x)\sum_i \frac{\partial}{\partial x_i}(\mu_i(x)p_t(x))dx.$$

Similarly, integrating by parts twice gives

$$\begin{aligned} \int \mathrm{tr}(\sigma(x)^\top\nabla^2\phi(x)\sigma(x))p_t(x)dx &= \int \mathrm{tr}(\nabla^2\phi(x)\sigma(x)\sigma(x)^\top)p_t(x)dx \\ &= \sum_{i,j}\int \phi(x)\frac{\partial^2}{\partial x_i\partial x_j}\left[(D(x))_{ij}p_t(x)\right]dx. \end{aligned}$$

Hence,

$$\int \phi(x) \left[\frac{dp_t}{dt} + \sum_i \frac{\partial}{\partial x_i}(\mu(x)_i p_t(x)) - \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} \left[(D(x))_{ij} p_t(x) \right] \right] dx = 0$$

for any smooth ϕ. Therefore, we have the conclusion of the lemma.

We apply the Fokker–Planck equation to the Langevin dynamics. □

Theorem 8. *For any smooth function f, the density proportional to $F = e^{-f}$ is stationary for the Langevin dynamics.*

Proof. The Fokker–Planck equation (Theorem 7) shows that the distribution p_t of x_t satisfies

$$\frac{dp_t}{dt} = \sum_i \frac{\partial}{\partial x_i}(\frac{\partial f(x)}{\partial x_i} p_t(x)) + \sum_i \frac{\partial^2}{\partial x_i^2} \left[p_t(x) \right]. \tag{3.1}$$

We can verify that $p_t(x) \propto e^{-f(x)}$ is a solution. □

■ 3.3 Mixing Time[6]

Here we show that Langevin dynamics is simply gradient descent for the function $F(\rho) = D_{\mathrm{KL}}(\rho \| \nu)$ on the Wasserstein space where $\nu = e^{-f(x)} / \int e^{-f(y)} dy$. For this, we first define the Wasserstein space.

Definition 9. The Wasserstein space $P_2(\mathbb{R}^n)$ on $\mathbb{R}^n$ is the manifold on the set of probability measures on $\mathbb{R}^n$ such that the shortest path distance of two measures x, y in this manifold is exactly equal to the Wasserstein distance between x and y.

Lemma 10. *For any $p \in P_2(\mathbb{R}^n)$ and $v \in T_p P_2(\mathbb{R}^n)$, we can write $v(x) = \nabla \cdot (p(x) \nabla \lambda(x))$ for some function λ on $\mathbb{R}^n$. Furthermore, the length of v in this metric is given by*

$$\|v\|_p^2 = \mathbf{E}_{x \sim p} \|\nabla \lambda(x)\|^2.$$

Proof. Let $p \in P_2(\mathbb{R}^n)$ and $v \in T_p P_2(\mathbb{R}^n)$. We will show that any change of density v can be represented by a vector field c on $\mathbb{R}^n$ as follows: Consider the process $x_0 \sim p$ and $\frac{d}{dt} x_t = c(x_t)$. Let p_t be the density of the distribution of x_t. To compute $\frac{d}{dt} p_t$, we follow the same idea as in the proof as Theorem 7. For any smooth function ϕ, we have that $\mathbf{E}_{x \sim p_t} \phi(x) = \mathbf{E} \phi(x_t)$. Taking derivatives on the both sides with respect to t, we have that

$$\int \phi(x) \frac{d}{dt} p_t(x) dx = \int \nabla \phi(x)^\top c(x) p_t(x) dx = - \int \nabla \cdot (c(x) p_t(x)) \phi(x) dx$$

where we used integration by parts at the end. Since this holds for all ϕ, we have that

$$\frac{dp_t(x)}{dt} = -\nabla \cdot (p_t(x) c(x)).$$

Since we are interested only in the vector fields that generate minimum movement in Wasserstein distance, we consider the optimization problem

$$\min_{-\nabla \cdot (pc) = v} \frac{1}{2} \int p(x) \|c(x)\|^2 dx$$

where we can think v is the change of p_t. Let $\lambda(x)$ be the Lagrangian multiplier of the constraint $-\nabla \cdot (pc) = v$. Then, the problem becomes

$$\begin{aligned} &\min_c \frac{1}{2} \int p(x) \|c(x)\|^2 dx - \int \lambda(x) \nabla \cdot (p(x) c(x)) dx. \\ =& \min_c \frac{1}{2} \int p(x) \|c(x)\|^2 dx + \int \nabla \lambda(x)^\top c(x) \cdot p(x) dx. \end{aligned}$$

[6] Will skip in the lecture

Now, we note that the problem is a pointwise optimization problem with the minimizer is given by

$$c(x) = -\nabla\lambda(x).$$

This proves that any vector fields that generate minimum movement in Wasserstein distance is a gradient field. Also, we have that $v(x) = \nabla \cdot (p(x)\nabla\lambda(x))$. Note that the right hand side is an elliptical differential equation and hence for any v with $\int v(x)dx = 0$, there is an unique solution $\lambda(x)$. Therefore, we can write $v(x) = \nabla \cdot (p(x)\nabla\lambda(x))$ for some $\lambda(x)$.

Next, we note that the movement is given by

$$\|v\|_p^2 = \int p(x)\|c(x)\|^2 dx = \mathbf{E}_{x\sim p}\|\nabla\lambda(x)\|^2.$$

□

Now, we show that Langevin Dynamics is simply gradient descent on KL distance under the Wasserstein space.

Theorem 11. *Let ρ_t be the density of the distribution produced by Langevin Dynamics for the target distribution $\nu = e^{-f(x)}/\int e^{-f(y)}dy$. Then, we have that*

$$\frac{d\rho}{dt} = \mathrm{argmin}_{v\in T_p P_2(\mathbb{R}^n)} \langle \nabla F(\rho), v\rangle_p + \frac{1}{2}\|v\|_p^2.$$

Namely, ρ_t follows continuous gradient descent in the density space for the function $F(\rho) = D_{\mathrm{KL}}(\rho\|\nu)$ under the Wasserstein metric.

Proof. For any function c, the optimization problem of interest satisfies

$$\min_{\delta=\nabla\cdot(\rho\nabla\lambda)} \langle c, \delta\rangle + \frac{1}{2}\int \rho(x)\|\nabla\lambda(x)\|^2 dx = \min_{\nabla\lambda} -\int \rho(x)\cdot\nabla c(x)^\top \nabla\lambda(x)dx + \frac{1}{2}\int \rho(x)\|\nabla\lambda(x)\|^2 dx.$$

Solving the right hand side, we have $\nabla c = \nabla\lambda$ and hence $\delta = \nabla\cdot(\rho\nabla c)$. Now, we note that $\nabla F(\rho) = \log\frac{\rho}{\nu} - 1$. Therefore,

$$\begin{aligned}\frac{d\rho}{dt} &= \nabla\cdot(\rho\nabla(\log\frac{\rho}{\nu} - 1)) \\ &= \nabla\cdot(\rho\nabla\log\frac{\rho}{\nu}) \\ &= \nabla\cdot(\rho\nabla f) + \Delta\rho\end{aligned}$$

which is exactly equal to (3.1). □

To analyze this continuous descent on the Wasserstein space, we first prove that continuous gradient descent converges exponentially whenever F is strongly convex.

Lemma 12. *Let F be a function satisfying "Gradient Dominance":*

$$\|\nabla F(x)\|_x^2 \geq \alpha\cdot(F(x) - \min_y F(y)) \quad \textit{for all } x \tag{3.2}$$

on the manifold with the metric $\|\cdot\|_x$ where ∇ is the gradient on the manifold. Then, the process $dx_t = -\nabla F(x_t)dt$ converges exponentially, i.e., $F(x_t) - \min_y F(y) \leq e^{-\alpha t}(F(x_0) - \min_y F(y))$.

Proof. We write

$$\frac{d}{dt}(F(x) - \min_y F(y)) = \langle \nabla F(x_t), \frac{dx_t}{dt}\rangle_{x_t} = -\|\nabla F(x_t)\|_{x_t}^2 \leq -\alpha(F(x) - \min_y F(y)).$$

The conclusion follows. □

Finally, we note that the log-Sobolev inequality for the density ν can be re-stated as the condition (3.2).

Lemma 13. *Fix a density ν. Then the log-Sobolev inequality, namely, for every smooth function g,*

$$2\int \|\nabla g\|^2 \, d\nu \geq \alpha \int g(x)^2 \log g(x)^2 \, d\nu$$

implies the condition (3.2).

Proof. Take $g(x) = \sqrt{\frac{\rho(x)}{\nu(x)}}$, the log-Sobolev inequality shows that

$$\frac{1}{2}\int \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx \geq \alpha \cdot \int \rho(x) \log \frac{\rho(x)}{\nu(x)} \, dx \text{ for all } \rho.$$

As we calculate in Theorem 11, we have that

$$\|\nabla F(\rho)\|_\rho^2 = \int \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx.$$

Therefore, this is exactly the condition (3.2) with coefficient 2α. □

Combining Lemma 13 and Lemma 12, we have the following result:

Theorem 14. *Let f be a smooth function with log-Sobolev constant α. Then the Langevin dynamics*

$$dx_t = -\nabla f(x)dt + \sqrt{2}dW_t$$

converges exponentially in KL-divergence to the density $\nu(x) \propto e^{-f(x)}$ with mixing rate $O(\frac{1}{\alpha})$, i.e., $KL(x_t, \nu) \leq e^{-2\alpha t} KL(x_0, \nu)$.

We note that many distributions have log-Sobolev constant larger than 0, including logconcave distributions.

4 Discrete Langevin Dynamic

In this section, we discuss the discrete Langevin Dynamic:

$$x_{t+h} = x_t - \nabla f(x_t)h + \sqrt{2h}\zeta_t \tag{4.1}$$

where $\zeta_t \sim N(0, I)$.

4.1 Relation with the stochastic gradient descent

Many functions in machine learning is of the form

$$f(x) = \frac{1}{n}\sum_{i\in[n]} f_i(x).$$

Often, we minimize such functions via stochastic gradient descent:

$$x_{t+h} = x_t - h\nabla f_i(x)$$

where i is randomly sampled from $[n]$. We can rewrite the equation above as

$$x_{t+h} = x_t - h\nabla f(x) - h(\nabla f_i(x) - \nabla f(x)).$$

Note that $h(\nabla f_i(x) - \nabla f(x))$ is a mean 0 term some variance. If we ends our algorithm with some positive learning rate, we are essentially running a discrete Langevin Dynamic with the diffusion term $\sqrt{2h}\zeta_t$ replaced by some other diffusion depending on the function f.

4.2 Metropolis–Hastings algorithm

In general, the stationary distribution for (4.1) is not $e^{-f(x)}$. However, there is a technique that corrects the stationary distribution. More general, we can think (4.1) is a Markov chain (but with infinitely many states). The following Lemma gives a sufficient condition for a Markov chain having a stationary distribution π.

Lemma 15. *Given a Markov chain $p(x \to y)$ and a distribution π. If $\pi(x)p(x \to y) = \pi(y)p(y \to x)$ for any $x \neq y$, then we have that π is a stationary distribution of the chain.*

Proof. Let $P\pi$ is the distribution after one step of the Markov chain starting from π. Then, we have that

$$(P\pi)_y = \sum_x \pi(x)p(x \to y) = \sum_x \pi(y)p(y \to x) = \pi(y)$$

where we used that $p(y \to \cdot)$ is a probability distribution. Hence, we have $P\pi = \pi$, namely, π is a stationary distribution. □

When the condition $\pi(x)p(x \to y) = \pi(y)p(y \to x)$ is violated, say if $p(x \to y) > \frac{\pi(y)p(y \to x)}{\pi(x)}$, then we can simply reject the sample y with some appropriate probability and this is called the Metropolis–Hastings algorithm.

Theorem 16. *Given a Markov chain $p(x \to y)$ and a target distribution π. We define a new Markov chain q as follows:*

- *Sample y according to $p(x \to y)$.*
- *Go to y with probability $\min(1, \frac{\pi(y)p(y \to x)}{\pi(x)p(x \to y)})$. Otherwise, stay at x.*

Then, π is a stationary distribution of q.

Proof. For any $y \neq x$, we have that

$$\begin{aligned}\pi(x)q(x \to y) &= \pi(x)p(x \to y)\min(1, \frac{\pi(y)p(y \to x)}{\pi(x)p(x \to y)}) \\ &= \min(\pi(x)p(x \to y), \pi(y)p(y \to x)).\end{aligned}$$

Similarly, we have

$$\pi(y)q(y \to x) = \min(\pi(x)p(x \to y), \pi(y)p(y \to x)).$$

Hence, we have $\pi(x)q(x \to y) = \pi(y)q(y \to x)$ for all $x \neq y$. Lemma 15 shows that π is a stationary distribution of q. □

Now, we can apply Metropolis–Hastings on the discrete Langevin Dynamic and get an algorithm that converges to e^{-f}.

Algorithm 1: Metropolis-adjusted Langevin algorithm (MALA)

Input: starting point x.
Repeat T times:

1. Compute $y = x - \nabla f(x)h + \sqrt{2h}\zeta$ where $\zeta \sim N(0, I)$.
2. Compute $\alpha = \min(1, \frac{\exp(-f(y))\exp(-\frac{1}{4h}\|y - x + \nabla f(x)h\|^2)}{\exp(-f(x))\exp(-\frac{1}{4h}\|x - y + \nabla f(y)h\|^2)})$.
3. Set $x \leftarrow y$ with probability α.

return x.

■ 5 Hamiltonian Monte Carlo

Metropolis-adjusted Langevin algorithm looks practical in the sense that it is a small variant of gradient descent with just one extra step of accepting samples. Unfortunately, it does not work well in practice because whenever the step size is large, it will reject all the samples. To see this, let us use this to sample Gaussian distribution.

Lemma 17. *For $f(x) = \frac{1}{2}\|x\|^2$ with $x \in \mathbb{R}^d$, when $h \geq n^{-1/3}$, the MALA accepted only $\exp(-O(h^2 n))$ portions of samples, which is exponentially small.*

Proof. Recall the acceptance probability is given by:

$$\begin{aligned}
\alpha =& \frac{\exp(-\frac{1}{2}\|y\|^2 - \frac{1}{4h}\|y - (1-h)x\|^2)}{\exp(-\frac{1}{2}\|x\|^2 - \frac{1}{4h}\|x - (1-h)y\|^2)} \\
=& \frac{\exp(-\frac{1}{2}\|y\|^2 - \frac{1}{4h}\|y\|^2 + \frac{1-h}{2h} y^\top x - \frac{(1-h)^2}{4h}\|x\|^2)}{\exp(-\frac{1}{2}\|x\|^2 - \frac{1}{4h}\|x\|^2 + \frac{1-h}{2h} y^\top x - \frac{(1-h)^2}{4h}\|y\|^2)} \\
=& \exp((1 - \frac{h}{4})\|x\|^2 - (1 - \frac{h}{4})\|y\|^2)
\end{aligned}$$

Now, using $y = x - \nabla f(x)h + \sqrt{2h}\zeta$, we have that

$$\begin{aligned}
\frac{\log \alpha}{1 - \frac{h}{4}} &= \|x\|^2 - \|x - \nabla f(x)h + \sqrt{2h}\zeta\|^2 \\
&= \|x\|^2 - \|(1-h)x + \sqrt{2h}\zeta\|^2 \\
&= (2h - h^2)\|x\|^2 - 2\sqrt{2h} x^\top \zeta + 2h\|\zeta\|^2.
\end{aligned}$$

Suppose now x is already close to normal distribution. Then, we have that $\|x\|^2 \sim n \pm O(\sqrt{n})$. Similarly, we have $\|\zeta\|^2 = n \pm O(\sqrt{n})$ and $x^\top \zeta = \pm O(\sqrt{n})$. Hence, we have

$$\log \alpha = \pm O(\sqrt{hn}) - O(h^2 n).$$

Note that when $h \gg n^{-1/3}$, then we have that $\log \alpha = -O(h^2 n) \ll 0$. In this case, the Metropolis-adjusted Langevin algorithm only accepted $\exp(-O(h^2 n))$ portions of samples, which is exponentially small. □

Since the step size is $h \sim n^{-1/3}$, this means that it requires $n^{1/3}$ steps to just sample from Gaussian distribution. It turns out that Gaussian distribution is the easiest kind of distribution for MALA because the Hessian of f is constant. When the Hessian is not a constant function, the step size need to be even smaller, such as $n^{-1/2}$ even in practice.

It turns out there is a better algorithm for sampling called Hamiltonian Monte Carlo. Unfortunately, I am lazy to type it up. So, please see the white board.