SAMPLING.

Monday, March 23, 2020 4:11 PM

$f: \mathbb{R}^n \to \mathbb{R}$

Sample $x \in \mathbb{R}^n$ with prob $\propto e^{-f(x)}$.

e.g. $f(x) = \|x\|^2$ Gaussian

$$f(x) = \begin{cases} 1 & x \in K \\ \infty & x \notin K. \end{cases}$$

Intractable in general.

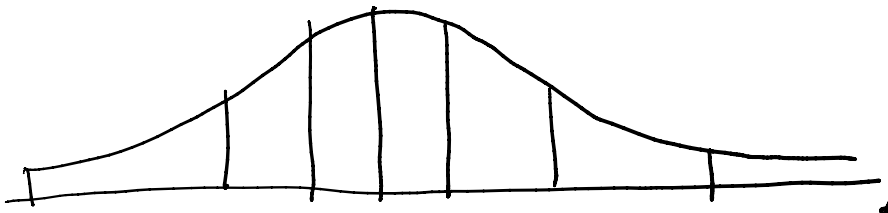
What classes of functions have polytime algorithms?

P1 How to sample from $\nu(x) \propto e^{-\frac{x^T A x}{2}} \cdot \mathbb{1}_{\{x \geq 0\}}$?

$e^{-\frac{x^T A x}{2}}$ is easy: get $y \sim N(0, I)$

apply $x = A^{-\frac{1}{2}} y$ then $x \sim N(0, A^{-1})$.

$N(0, I)$: generate each $x_i \sim N(0,1)$



many numerically efficient methods.

many numerically efficient methods

What about P1 ?

Idea: use a randomized process

$dX_t = -\nabla f(X_t)dt$ ⟶ to go to a local min of f.

$$\boxed{dX_t = -\nabla f(X_t)dt + \sqrt{2}\, dW_t}$$ Langevin Dynamics or Diffusion

(dW_t: infinitesimal Brownian motion.)

(discrete time

$X^{k+1} = X^k - \nabla f(X_k)\cdot h + \sqrt{2h}\, Z$ $\quad Z \sim N(0,1)$.)

Thm! For any smooth f, the density $\propto e^{-f}$ is stationary for Langevin Diffusion.

This theorem follows from a very general analysis of stochastic processes of this form.

SDE Stochastic Differential Equation

$$dX_t = \mu(t)\,dt + \sigma(t)\,dW_t$$

m

$$dX_t = \nu(t)\,dt + \sigma(t)\,dW_t$$

$X_t, \nu_t \in \mathbb{R}^n \qquad \sigma(t) \in \mathbb{R}^{n \times m} \qquad dW_t \in \mathbb{R}^m.$

Thm 2 [Fokker-Planck] $X \sim p_0$

$$\frac{dp_t(x)}{dt} = -\sum_i \frac{\partial}{\partial x_i}\left(\nu_i(t)\, p_t(x)\right) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left(D_{ij}(t)\, p_t(x)\right)$$

$$= -\nabla\cdot(p\nu) + \frac{1}{2}\nabla\cdot(\nabla\cdot(pD))$$

$$D(t) = \sigma(t)\sigma(t)^T$$

Pf (of Thm 1). For stationary p_t, with $D = 2I$

$$0 = \frac{dp_t(x)}{dt} = -\sum_i \frac{\partial}{\partial x_i}\left(-p_t(x)\frac{\partial f}{\partial x_i}(x)\right) + \sum_i \frac{\partial^2}{\partial x_i^2}(p_t(x))$$

$$0 = \sum_i p_t(x)\frac{\partial^2 f(x)}{\partial x_i^2} + \frac{\partial p_t(x)}{\partial x_i}\frac{\partial f(x)}{\partial x_i} + \frac{\partial^2}{\partial x_i^2} p_t(x)$$

OR

$$0 = \sum_i \frac{\partial}{\partial x_i}\left(p_t(x)\frac{\partial}{\partial x_i} f(x) + \frac{\partial}{\partial x_i} p_t(x)\right)$$

$$= \sum_i \frac{\partial}{\partial x_i}\left(p_t(x)\left(\frac{\partial f(x)}{\partial x_i} + \frac{\partial}{\partial x_i}\log p_t(x)\right)\right)$$

$$= \sum_i \frac{\partial}{\partial x_i} \left(p_t(x) \left(\frac{\partial V}{\partial x_i} + \frac{\partial}{\partial x_i} \log p_t \right) \right)$$

$$0 = \sum_i \frac{\partial}{\partial x_i} \left(p_t(x) \frac{\partial}{\partial x_i} \log \left(\frac{p_t(x)}{e^{-f(x)}} \right) \right)$$

$p_t(x) = c \cdot e^{-f(x)}$ is a solution.

We will review the proof of F-P at the end.

What is the rate of convergence of LD?

First in continuous time. How to compare distributions?

$p_0 \rightsquigarrow p_t \rightsquigarrow p \propto e^{-f}$.

$d_W(p, q)$ Wasserstein distance metric $\|\cdot\|$ on domain of p.

$$= \min_{\Pi : \text{coupling of } X \sim p, Y \sim q} \mathbb{E}(\|X - Y\|^2)$$

marginal of Π in X is p, in Y is q.

More familiar $d_{TV}(p, q) = \frac{1}{2} \int |p(x) - q(x)| \, dx$.

and KL-divergence (asymmetric)

$$d_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Thm 3 Assume f is ν-strongly convex. Then LD starting at X_0, Y_0 can be coupled s.t. $\mathbb{E}(\|X_t - Y_t\|^2) \le e^{-2\nu t} \|X_0 - Y_0\|^2$

Cor. LD converges to e^{-f}.

Pf. Take $Y_0 \sim e^{-f}$ then $Y_t \sim e^{-f}$.

Pf (of Thm 3). Use identity coupling.

$$\left.\begin{aligned} dX_t &= -\nabla f(X_t)dt + \sqrt{2}\, dW_t \\ dY_t &= -\nabla f(Y_t)dt + \sqrt{2}\, dW_t \end{aligned}\right\} \text{equal.}$$

$$\frac{d}{dt}(X_t - Y_t) = -(\nabla f(X_t) - \nabla f(Y_t))$$

So $\frac{d}{dt}\|X_t - Y_t\|^2 = 2\left\langle X_t - Y_t, \frac{d}{dt}(X_t - Y_t)\right\rangle$

$= \cdots \le -2\nu \|X_t - Y_t\|^2 \rangle$

$$\frac{d}{dt} = -2 \langle \nabla f(X_t) - \nabla f(Y_t),\ X_t - Y_t \rangle$$

Using strong convexity,

$$f(X_t) - f(Y_t) \geq \langle \nabla f(Y_t),\ X_t - Y_t \rangle + \frac{\mu}{2} \| X_t - Y_t \|^2$$

$$f(Y_t) - f(X_t) \geq \langle \nabla f(X_t),\ Y_t - X_t \rangle + \frac{\mu}{2} \| X_t - Y_t \|^2$$

$$\langle \nabla f(X_t) - \nabla f(Y_t),\ X_t - Y_t \rangle \geq \mu \| X_t - Y_t \|^2$$

$$\Rightarrow \quad \frac{d}{dt} \| X_t - Y_t \|^2 \leq -2\mu \| X_t - Y_t \|^2$$

i.e. $$\frac{d}{dt} \log \| X_t - Y_t \|^2 \leq -2\mu$$

$$\Rightarrow \quad \| X_t - Y_t \|^2 \leq e^{-2\mu t} \| X_0 - Y_0 \|^2 .$$

Back to Fokker-Planck and SDE

Given $$dX_t = \mu(X_t)\, dt + \sigma(t)\, dW_t$$

Given $dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t$

What can we say about $df(X_t)$?

In usual calculus, this is the chain rule,

$$df(X_t) = \nabla f(X_t)\, dX_t$$

Which comes from Taylor expansion of f

$$f(X_t + h) = f(X_t) + \nabla f(X_t)\, h + \frac{1}{2} h^T \nabla^2 f(X_t)\, h + \dots$$

$h \to 0$ only second term matters in

$$\frac{f(X_t+h) - f(X_t)}{h} \to \nabla f(X_t).$$

But with an SDE, we have both dt and dW_t.

Note that $W_t - W_0 \sim N(0, t)$

$$\mathbb{E}(W_t^2) = t$$

$$dW_t = W_{dt}$$

$$\mathbb{E}((dW_t)^2) = \mathbb{E}(W_{dt}^2) = dt.$$

So,

$$df(X_t) = \nabla f(X_t)\, dX_t + \frac{1}{2} dX_t^T \nabla^2 f(X_t)\, dX_t$$

so,

$$df(X_t) = \nabla f(X_t)\, dX_t + \frac{1}{2} dX_t \nabla^2 f(X_t)\, dX_t$$

$$= \nabla f(X_t)\, \mu(X_t)\, dt + \nabla f(X_t)\, \sigma(t)\, dW_t$$

$$+ \frac{1}{2} \mu(X_t)^T \nabla^2 f(X_t)\, \mu(X_t) (dt)^2 \longrightarrow 0$$

$$+ \mu(X_t)^T \nabla^2 f(X_t)\, \sigma(t)\, dW_t\, dt \longrightarrow 0$$

$$+ \frac{1}{2} dW_t^T \sigma(t)^T \nabla^2 f(X_t)\, \sigma(t)\, dW_t$$

$$df(X_t) = \nabla f(X_t)^T \mu(X_t)\, dt + \frac{1}{2} \nabla f(X_t)^T \sigma(t)\, dW_t$$

$$+ \frac{1}{2} \left\langle \nabla^2 f(X_t), \sigma(t)\sigma(t)^T \right\rangle dt$$

Itô's rule : chain rule for stochastic calculus.

Now let's prove F-P.

Pf. (of Thm 2) ϕ smooth function

$$\mathbb{E}_{p_t}(\phi(x)) = E_p(\phi(X_t))$$

... t ... both sides.

Take derivatives wrt t on both sides.

$$\int \phi(x)\, d p_t(x)\, dx = \mathbb{E}\Big(\nabla\phi(X_t)\cdot dX_t + \frac{1}{2}\langle \nabla^2\phi(X_t), \sigma(X_t)\sigma(X_t)^T\rangle dt\Big)$$

$$= \mathbb{E}\Big(\underbrace{\nabla\phi(X_t)^T r(X_t)\, dt}_{①} + \underbrace{\nabla\phi(X_t)\sigma(X_t)\, dW_t}_{} + \frac{1}{2}\langle \quad \underbrace{}_{②} \quad\rangle dt\Big)$$

$\mathbb{E}(dW_t) = 0$

Note that $X_t \sim p_t$. So,

$$① = \mathbb{E}_p\big(\nabla\phi(X_t)^T r(X_t)\, dt\big) = \mathbb{E}_{p_t}\big(\nabla\phi(x) r(x) dt\big)$$

$$= \int \nabla\phi(x)^T r(x)\, p_t(x) dt \cdot dx$$

integrate by parts

$$= \Big[\phi(x) r(x) p_t(x)\Big] - \int \phi(x) \sum_i \frac{\partial\big(r_i(x) p_t(x)\big)}{\partial x_i} dt\, dx$$

$$= 0 .$$

$$② = \frac{1}{2}\int \langle \nabla^2\phi(x), \sigma(x)\sigma(x)^T \rangle p_t(x)dt$$

integrate by parts

$$= \frac{1}{2}\int \langle \nabla\phi(x), \sum_i \frac{\partial}{\partial x_i}\left(p_t(x)(\sigma(x)\sigma(x)^T)_i\right) dt\ dx$$

$$= \frac{1}{2}\int \phi(x) \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left(p_t(x)(\sigma(x)\sigma(x)^T)_{ij}\right) dt$$

$\therefore$

$$\int \phi(x)\Big[\underbrace{\frac{dp_t(x)}{dt} + \sum_i \frac{\partial}{\partial x_i}\left(p_t(x)\mu_i(x)\right) - \frac{1}{2}\sum \frac{\partial^2}{\partial x_i \partial x_j}\left(p_t(x)(\sigma(x)\sigma(x)^T)_{ij}\right)}_{=0}\Big] dx$$

$\forall$ smooth ϕ

$= 0$. ✓

$\Rightarrow$

There is a more direct connection to gradient descent.

We start at p_0

$-f$

$\propto p$.

we want to end at $q \propto e^{-t}$.

Can we phrase this as an OPT problem?

YES.

Min $d(p_t, q)$

What distance to use?
is the gradient implementable?

One nice answer.

Use $d_{KL}(p_t \| q)$ minimized when $p_t = q$

with Wasserstein metric as distance between measures.

Then GD is LD!