GD continued

Saturday, January 11, 2025 7:52 PM

yhw $\nabla f(x) = 0 \Rightarrow x$ is global min for convex f.

Lem. $f(x) - f(y) \le \|\nabla f(x)\|_2 \|x - y\|_2$ for convex f.

Pf: $f(y) - f(x) \ge \langle \nabla f(x), y - x \rangle$

$$f(x) \le f(y) - \langle \nabla f(x), y - x \rangle$$
$$\le f(y) + \|\nabla f(x)\|_2 \|y - x\|_2 .$$

Th. $x^{(k+1)} = x^{(k)} - \frac{1}{L} \nabla f(x^{(k)})$ achieves

$$f(x^{(k)}) - f(x^*) \le \frac{2L}{k+4} \|x^{(0)} - x^*\|_2^2$$

Pf: $f(x^{(k+1)}) = f\left(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)})\right)$

$$\le f(x^{(k)}) - \frac{1}{2L} \|\nabla f(x^{(k)})\|_2^2$$

$$f(x^{(k+1)}) - f(x^*) \le f(x^{(k)}) - f(x^*) - \frac{1}{2L} \frac{\left(f(x^{(k)}) - f(x^*)\right)^2}{\|x^{(k)} - x^*\|_2^2}$$

ε_{k+1} ε_k

$$\varepsilon_{k+1} \le \varepsilon_k - \frac{1}{2L} \left(\frac{\varepsilon_k}{R}\right)^2$$

$$\frac{1}{\varepsilon_{k+1}} - \frac{1}{\varepsilon_k} = \frac{\varepsilon_k - \varepsilon_{k+1}}{\varepsilon_{k+1} \varepsilon_k} \le \frac{1}{2L} \frac{\varepsilon_k^2}{R^2} \frac{}{\varepsilon_k^2} = \frac{1}{2LR^2}$$

$$\frac{1}{\varepsilon_{k+1}} - \frac{1}{\varepsilon_k} = \frac{\varepsilon_k - \varepsilon_{k+1}}{\varepsilon_{k+1}\varepsilon_k} \le \frac{1}{2L} \frac{\varepsilon_k}{R^2} \frac{}{\varepsilon_k^2} \cdot \frac{}{2LR^2}$$

$$\varepsilon_0 = f(x^{(0)}) - f(x^*) \le \langle \nabla f(x^*), x^{(0)} - x^* \rangle + \frac{L}{2}\|x^{(0)} - x^*\|_2^2$$

$$= \frac{L}{2}\|x^{(0)} - x^*\|_2^2 \le \frac{LR^2}{2}$$

$$\frac{1}{\varepsilon_k} \le \frac{2}{LR^2} + \frac{k}{2LR^2} = \frac{k+4}{2LR^2}$$

$$\Rightarrow \varepsilon_k \le \frac{2LR^2}{k+4}$$

$$R = \max_{x:\, f(x) \le f(x^{(0)})} \|x - x^*\|_2$$

In fact, we can set $R = \|x^{(0)} - x^*\|_2$

Lemma. For $h \le \frac{2}{L}$ $\|x^{(k+1)} - x^*\|_2 \le \|x^{(k)} - x^*\|_2$.

Pf. $\|x^{(k+1)} - x^*\|_2^2 = \|x^{(k)} - h\nabla f(x^{(k)}) - x^*\|_2^2$

$$= \|x^{(k)} - x^*\|_2^2 + h^2\|\nabla f(x^{(k)})\|_2^2 - 2h\langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle$$

$$\nabla f(x^{(k)}) = \underbrace{\nabla f(x^*)}_{0} + \int_0^1 \langle \nabla^2 f(x^* + t(x^{(k)} - x^*)),\ x^{(k)} - x^* \rangle \, dt$$

$$\langle \nabla f(x^{(k)}),\ x^{(k)} - x^* \rangle = (x^{(k)} - x^*)^T \underbrace{\int_0^1 \nabla^2 f(x^* + t(x^{(k)} - x^*))\, dt}_{H \succcurlyeq 0} \quad (x^{(k)} - x^*)$$

$$H^2 \succcurlyeq L H \qquad \geq \frac{1}{L} \| H(x^{(k)} - x^*) \|_2^2$$

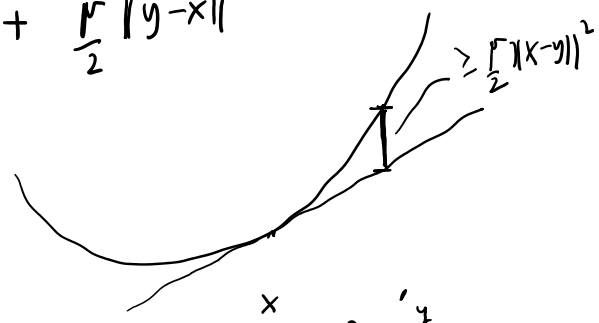
$$= \frac{1}{L} \| \nabla f(x^{(k)}) \|_2^2$$

$$\| x^{(k+1)} - x^* \|_2^2 \leq \| x^{(k)} - x^* \|_2^2 + \left(h^2 - \frac{2h}{L} \right) \| \nabla f(x^{(k)}) \|_2^2$$

$h \geq \frac{2}{L} \Rightarrow$ ≥ 0.

Strongly convex.

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



Lemma. The following are equivalent. ($\mu > 0$)

(1). $\forall t \quad f((1-t)x + ty) \leq (1-t) f(x) + t f(y) - \frac{\mu}{2} t(1-t) \| x - y \|^2$

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

(3) $\nabla^2 f(x) \succcurlyeq \mu I$

(3) $\nabla^2 f(x) \succeq \nu I$

Pf. Define $g(x) = f(x) - \frac{\nu}{2}\|x\|^2$. Apply equivalent definitions of convexity to g.

Lemma For f ν-strongly convex,

$$\|\nabla f(x)\|^2 \geq 2\nu\,(f(x) - f(y)).$$

Pf. $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\nu}{2}\|y-x\|^2$

$$f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle - \frac{\nu}{2}\|y-x\|^2$$

$$\leq \max_z \; \langle \nabla f(x), z\rangle - \frac{\nu}{2}\|z\|^2$$

At optimal z, $\nabla f(x) = \nu z$

So $f(x) - f(y) \leq \frac{1}{2\nu}\|\nabla f(x)\|_2^2$.

Thm. $f(x^{(k)}) - f(x^*) \leq \left(1 - \frac{\nu}{L}\right)^k \left(f(x^{(0)}) - f(x^*)\right)$

Pf. $f(x^{(k+1)}) - f(x^*) \leq f(x^{(k)}) - f(x^*) - \frac{1}{2L}\|\nabla f(x^{(k)})\|_2^2$

$$\leq f(x^{(k)}) - f(x^*) - \frac{2\nu}{2L}\left(f(x^{(k)}) - f(x^*)\right)$$

$$\leq \left(1 - \frac{\nu}{L}\right)\left(f(x^{(k)}) - f(x^*)\right)$$

$$\leq \left(1 - \frac{\mu}{L}\right)\left(f(x^{(t)}) - f(x^*)\right).$$

f		# steps
general	$\|\nabla f(x)\| \leq \varepsilon$	$\frac{2L\,(f(x^0) - f(x^*))}{\varepsilon^2}$
convex	$f(x) \leq f^* + \varepsilon$	$\frac{2L\,\|x^{(0)} - x^*\|_2^2}{\varepsilon}$
μ-strongly convex	$f(x) \leq f^* + \varepsilon$	$\frac{L}{\mu}\log\left(\frac{f(x^{(0)}) - f(x^*)}{\varepsilon}\right)$

} different measures of convergence.

Note that for $f(x) = \frac{1}{2}\|x\|^2$

$\|\nabla f(x)\| \leq \varepsilon$ is $\|x\| \leq \varepsilon$

vs. $f(x) \leq f^* + \varepsilon$ is $\|x\|^2 \leq \varepsilon$

So the rates are the same.

Sampling

Recall goal: given access to f, sample according to e^{-f}.

Ex. 1. $f(x) = \frac{\|x\|^2}{2}$ $\quad e^{-\frac{\|x\|^2}{2}}$ $\quad N(0, I)$

Ex. 2 $f(x) = \begin{cases} 0 & x \in K \\ \infty & x \notin K \end{cases}$ $\quad e^{-f} = \mathbb{1}_K$

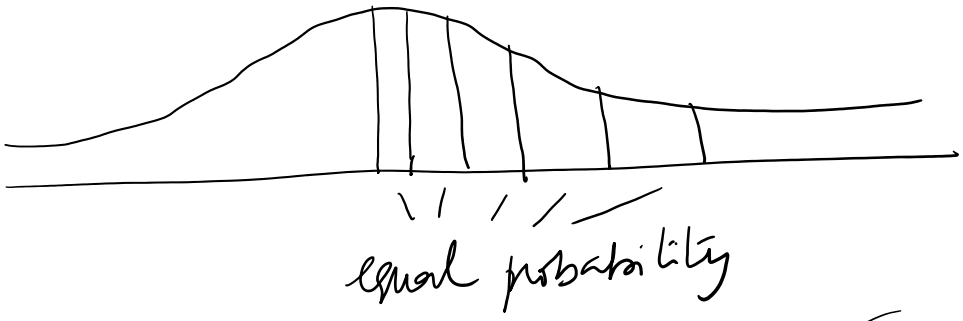
uniformly sample K.

To sample $N(0, I)$

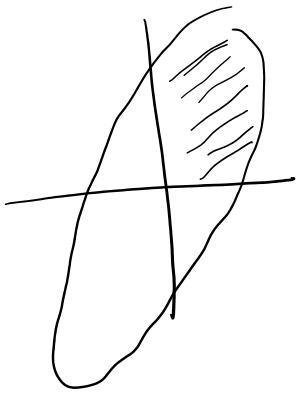
Observe that this is independent $x_i \sim N(0,1)$.

Many numerical methods to sample $N(0,1)$

e.g.,



How about $e^{-\frac{x^T A x}{2}} \cdot \mathbb{1}_{x \geq 0}$?



without the constraint

$e^{-\frac{x^T A x}{2}}$ is $N(0, A^{-1})$

Gaussian with covariance A^{-1}

$y = A^{\frac{1}{2}} x \qquad y^T y = x^T A x$

$\mathbb{E}(yy^T) = \mathbb{E}(A^{\frac{1}{2}} x x^T A^{\frac{1}{2}}) = I$

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

How about with constraint $x \geq 0$?

How about with constraint $x \geq 0$?

Rejection sampling (sample $x \sim N(0, A^{-1})$ output x if $x \geq 0$)

is too expensive.

Need an algorithm.

Back to 6D. Can view as a continuous time Algorithm.

Gradient flow

$$dX_t = -\nabla f(X_t)\, dt$$

$$\therefore d(f(X_t) - f(X_0)) = \langle \nabla f(X_t), dX_t \rangle$$

$$= -\| \nabla f(X_t) \|_2^2 \, dt$$

Under μ-strong convexity,

$$d(f(X_t) - f(X^*)) \leq -2\mu (f(X_t) - f(X^*))\, dt$$

$$f(X_t) - f(X^*) \leq e^{-2\mu t} (f(X_0) - f(X^*))$$

How about Sampling?

need to introduce randomness.

e.g.

To sample uniformly, start at i

$i \to i-1$ | $i \to i+1$ & [illegible]

To sample uniformly, start at i

w.p. $\frac{1}{2}$ → $i-1$

——— → $i+1$

stay if $i-1 \notin [n]$ or $i+1$

$p^{(k+1)}(i) = p^{(k)}(i-1)\cdot\frac{1}{2} + p^{(k)}(i+1)\frac{1}{2}$

$p(i) = \frac{1}{n}$ is stationary.

$$dX_t = -\nabla f(X_t)\,dt$$

(LD) $$\boxed{dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t}$$

"Brownian motion". infinitesimal Gaussian

$W_t \sim N(0, t)$

Langevin Diffusion.

Algorithm? Fix $h > 0$

$$X^{(k+1)} = X^{(k)} - h\nabla f(X^{(k)}) + \sqrt{2}\,Z$$

$Z \sim N(0, hI)$

$\sqrt{2h}\,\bar{Z}$ $\bar{Z} \sim N(0, I)$

Thm 1. For continuously differentiable f, with $\int e^{-f} < \infty$, density $\nu \propto e^{-f}$ is stationary for LD.

More generally $dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t$ is a Stochastic Differential Equation (SDE).

$X_t, \mu \in \mathbb{R}^n$, $\sigma \in \mathbb{R}^{n\times m}$, $W_t \in \mathbb{R}^m$

Thm [Fokker-Planck]

$$\frac{d p_t}{dt} = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\mu(x_t)_i p_t\right) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}\left(D(x_t)_{ij} p_t\right)$$

$$\left(D(x_t) = \sigma(x_t)\sigma(x_t)^T\right.$$

$$= -\nabla\cdot\left(p_t \mu(x_t)\right) + \frac{1}{2}\nabla\cdot\left(\nabla\cdot\left(p_t D(x_t)\right)\right)$$

$$= -\nabla\cdot\left(p_t \mu(x_t)\right) + \frac{1}{2}\Delta\left(p_t D(x_t)\right).$$

Pf. (of Thm 1). $\mu(x_t)_i = -\nabla f(x_t)_i = \frac{-\partial f(x)}{\partial x_i}$

$\sigma(x_t) = \sqrt{2}\, I$

$D(x_t)_{ij} = \begin{cases} 2 & \text{if } i = j \\ 0 & i \neq j \end{cases}$

So at stationary p

$$0 = \frac{dp}{dt} = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(-\frac{\partial f(x)}{\partial x_i} p(x)\right) + \frac{1}{2}\sum_i \frac{\partial^2}{\partial x_i^2}\left(2\, p(x)\right)$$

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

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

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

And one can see that $p(x) = C e^{-f(x)}$

And now we see that $p(x) = Ce^{-f}$ is a solution.

Q. Rate of convergence?

How to measure it?

$$d_{TV}(p,q) = \frac{1}{2}\int_\Omega |p(x)-q(x)|\,dx = \sup_{A\subseteq\Omega} p(A)-q(A)$$

$$\chi^2(p,q) = \int \left(\frac{p(x)}{q(x)}-1\right)^2 q(x)\,dx = \mathbb{E}_q\left(\left(\frac{p}{q}-1\right)^2\right)$$

$$d_{KL}(p,q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx = \mathbb{E}_p\left(\log\frac{p}{q}\right)$$

$$W_2(p,q) = \min_{\pi:\text{ coupling of } X\sim p,\, Y\sim q} \mathbb{E}\left(\|X-Y\|^2\right)$$

$\|\cdot\|$ some norm on Ω.

coupling is a 1-1 map $\pi(X,Y)$ joint distribution

marginal $\pi(\cdot, Y)$ is q

marginal $\pi(X,\cdot)$ is p.

Thm $W_2(p_t, e^{-f}) \le e^{-2\nu t} W_2(p_0, e^{-f})$

for ν-strongly convex f.

Lemma: X Y_0 X_t Y_t Langevin Diffusions for f ν-strongly

Lemma. X_0, Y_0. X_t, Y_t Langevin Diffusions for f μ-strongly convex be coupled s.t. $\mathbb{E}(\|X_t - Y_t\|^2) \le e^{-2\mu t} \|X_0 - Y_0\|^2$

Pf.

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$
$$dY_t = -\nabla f(Y_t)\,dt + \sqrt{2}\,dW_t$$

couple with same W_t!

$$d(X_t - Y_t) = -(\nabla f(X_t) - \nabla f(Y_t))\,dt$$

$$d\|X_t - Y_t\|^2 = -2\langle \nabla f(X_t) - \nabla f(Y_t), X_t - Y_t \rangle\,dt$$

By μ-strong convexity,

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

$$\langle \nabla f(X_t) - \nabla f(Y_t), X_t - Y_t \rangle \ge \mu \|X_t - Y_t\|_2^2$$

$$\therefore \quad d\|X_t - Y_t\|^2 \le -2\mu \|X_t - Y_t\|^2\,dt$$

$$d \log \|X_t - Y_t\|^2 \le -2\mu\,dt$$

$$\|X_t - Y_t\|^2 \le e^{-2\mu t} \|X_0 - Y_0\|^2$$

$d \log z \le a\,dt$

$\log\left(\frac{z}{z_0}\right) \le at$

$z \le e^{at} z_0.$

Next time: Proof of F-P and hence of limiting distribution.

Main tool: A chain rule for stochastic variables!