Lecture 14 Stochastic Gradient and Variance Reduction

Sunday, October 16, 2016 10:25 PM

- Stochastic Gradient descent
  - least squares

$$\min \|y - Ax\|^2$$

$$\frac{1}{2n}\sum_{i=1}^{n} (y_i - \langle a_i, x\rangle)^2, \quad a_i \in \mathbb{R}^d$$

For simplicity assume $\|a_i\| = 1$

$$y_i = \langle a_i, x^*\rangle + \varepsilon_i \quad \left(\begin{array}{l} \sum \varepsilon_i a_i = 0 \\ |\varepsilon_i| \le \sigma \end{array}\right)$$

  - can rewrite

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T M (x - x^*)$$

where $M = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T$

  - SGD for least squares

$$f_i(x) = \frac{1}{2}(y_i - \langle a_i, x\rangle)^2$$

pick random i

$$x^{t+1} = x^t - \eta \nabla f_i(x)$$

$$= x^t + \eta (y_i - \langle a_i, x^t\rangle) a_i$$

  - Analyzing SGD

$$x^{t+1} = x^t - \eta \nabla f_i(x^t) = x^t - \eta(\nabla f(x^t) + \xi_i)$$

ξ_i independent of x^t. $\mathbb{E}[\xi_i] = 0$

Let $r_t = \mathbb{E}[\|x^t - x^*\|^2]$

$$r_{t+1}^2 = r_t^2 - 2\mathbb{E}[\eta\langle\nabla f(x^t) + \xi_i, x^t - x^*\rangle] + \eta^2\mathbb{E}[\|\nabla f(x^t) + \xi_i\|^2]$$

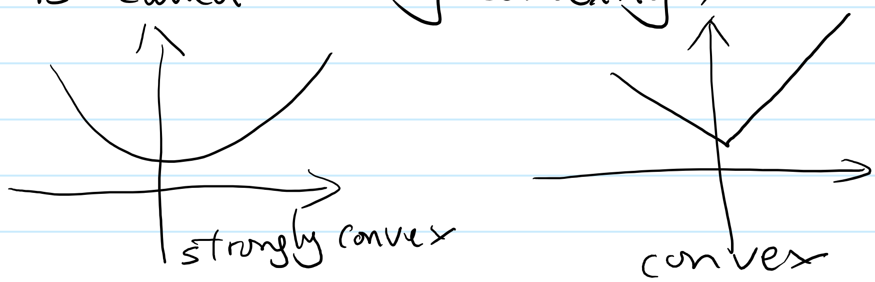
$$= r_t^2 - 2\eta\langle\nabla f(x^t), x^t - x^*\rangle + \eta^2\mathbb{E}[(y_i - \langle a_i, x^t\rangle)^2]$$

$$= r_t^2 - 2\eta(x^t - x^*)M(x^t - x^*) + 2\eta^2 f(x^t)$$

$$= r_t^2 - 2\eta(x^t - x^*)M(x^t - x^*) + 2\eta^2\left(f(x^*) + \frac{1}{2}(x^t - x^*)M(x^t - x^*)\right)$$

- Suppose $\sigma_{min}(M) = \mu$

(this is called Strong Convexity)



$$r_{t+1}^2 \leq r_t^2 - \underbrace{(2\eta - \eta^2)\mu r_t^2}_{A} + \underbrace{2\eta^2 f(x^*)}_{B}$$

we want term A to dominate term B!

set $\eta \ll \dfrac{\mu r_t^2}{f(x^*)}$ works

in that case

$$r_{t+1}^2 \le r_t^2 (1-\eta)$$
$$\le r_t^2 \left(1 - \frac{\mu r_t^2}{f(x^*)}\right)$$

again we solve the recursion and get

$$r_t^2 = \frac{f(x^*)}{\mu t}$$ (if the initial point is close enough)

in the best case, $\mu = \frac{1}{d}$

(because $tr(M) = \frac{1}{n} \sum \|a_i\|^2 = 1$)

so we can hope to get reasonably close after $O(d)$ iterations.

- System of linear equations

$$r_{t+1}^2 \le r_t^2 - \underbrace{(2\eta - \eta^2)\mu r_t^2}_{A} + \underbrace{2\eta^2 f(x^*)}_{B}$$

we had to use a small η to let A dominate B.

what if $f(x^*) = 0$? (this means $y_i = \langle a_i, x^* \rangle$)

then we can choose $\eta = 1$ and

$$r_{t+1}^2 \le (1-\mu) r_t^2$$

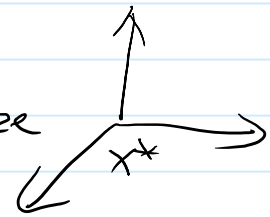
$$\Longrightarrow \quad r_t^2 \le (1-\mu)^t r_0^2$$

r_t^2 decreases by a constant factor every $\frac{1}{\mu}$ iterations!

- Variance Reduction

- idea: Previously, things did not work because

$$\nabla f_i(x^*) \neq 0$$

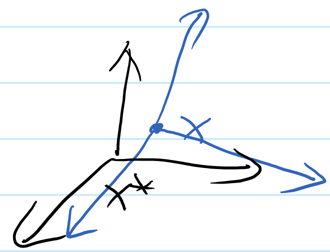


- if we choose a large step size will go away even if we are already at x^*!

- idea: if x is very close to x^*

$$\nabla f_i(x) \approx \nabla f_i(x^*)$$

Fix $\tilde{x}^0 = x$, pick i randomly



$$\tilde{x}^{t+1} = \tilde{x}^t - \eta\Big(\underbrace{\nabla f_i(\tilde{x}^t) - \nabla f_i(x)}_{\text{variance reduction}} + \nabla f(x)\Big)$$

make sure $\mathbb{E}[\,] = \nabla f(\tilde{x}^t)$