LECTURE – 7

Note Title 28-Jun-19

Stochastic Variance Reduced Gradient (SVRG)

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$ E.g., $f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2$

Each f_i is L-Smooth & Convex $\quad L = \max_i \|a_i\|^2$

f is μ-Strongly convex. $\quad \mu = \sigma_{min}\left(\frac{1}{n}\sum_i a_i a_i^T\right)$

GD: $O\left(\underbrace{n \cdot d}_{\text{one } \nabla f \text{ Computation}} \cdot \underbrace{\frac{L}{\mu} \log\frac{1}{\epsilon}}_{\#\text{ iterations}}\right)$ Computations.

SGD: $O\left(d \cdot \frac{\sigma \cdot \|x_0 - x^*\|}{\epsilon^2} \right)$

- σ: Std. dev. in the Stochastic gradients
- d: One ∇f_i Computation.
- $\frac{\sigma \cdot \|x_0 - x^*\|}{\epsilon^2}$: # iterations

In most cases: $\sigma \ll n \cdot \frac{L}{\mu}$

	Dep. on prob. param.	Desired acc.
SGD	Usually better	
GD		Usually better

→ Best of both worlds?

the slow convergence rate of SGD $\left(\frac{1}{\sqrt{t}}\right)$ is due to large variance in the stochastic gradients even close to the optimal point.

$$= \frac{1}{n}\sum_i \nabla f_i(x^*)$$

$x^* = \arg\min_x f(x)$ then $\nabla f(x^*) = 0$

but $\nabla f_i(x^*) \neq 0$ for most i.

Variance reduction step

Pick x_0

Compute $\nabla f(x_0) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_0)$

Do $t = 1, \cdots T$

Pick 'i' uniformly at random from $1, \cdots n$

$$x_t = x_{t-1} - \eta \left[\nabla f_i(x_{t-1}) - \nabla f_i(x_0) + \nabla f(x_0) \right].$$

Suppose $x_0 = x^*$

$$x_1 = x_0 - \eta \left[\nabla f_i(x_0) - \nabla f_i(x_0) + \nabla f(x_0) \right]$$

$$= x_0 = x^*$$

Contrast with SGD step: $x_t = x_{t-1} - \eta \nabla f_i(x_{t-1})$.

$$\mathop{\mathbb{E}}_{i} \left[\nabla f_i(x_0) - \nabla f(x_0) \right] = \mathop{\mathbb{E}}_{i} \left[\nabla f_i(x_0) \right] - \nabla f(x_0) = \nabla f(x_0) - \nabla f(x_0) = 0.$$

Recall that the convergence guarantee of SGD depends on the variance in stochastic gradients.

$$\sigma^2 \triangleq \mathbb{E}\left[\left\| \underbrace{\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0)}_{\text{Stochastic gradient}} - \underbrace{\nabla f(x_t)}_{\substack{\text{True gradient} \\ = \mathbb{E}[\text{Stoch. grad.}]}} \right\|^2 \right]$$

$$\left(\because (a+b)^2 \leq 2a^2 + 2b^2 \right)$$

$$\leq 2\,\mathbb{E}\left[\left\| \nabla f_i(x_t) - \nabla f_i(x_0) \right\|^2 \right] + 2\,\mathbb{E}\left[\left\| \nabla f(x_0) - \nabla f(x_t) \right\|^2 \right]$$

$$\overset{\text{Jensen's ineq}}{\leq} 4\,\mathbb{E}\left[\left\| \nabla f_i(x_t) - \nabla f_i(x_0) \right\|^2 \right]$$

$\left\{ \mathbb{E}[\|w\|^2] \geq \|\mathbb{E}[w]\|^2 \right.$ Jensen's inequality $\left. \right\}$

$x^* = \arg\min_x f(x)$

$((a+b)^2 \le 2a^2+2b^2)$

$$\mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x_0)\|^2\right] \le 2\left[\mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x^*)\|^2\right] + \nabla f_i(x^*) - \nabla f_i(x^*)\right] + \mathbb{E}\left[\|\nabla f_i(x_0) - \nabla f_i(x^*)\|^2\right].$$

$$\mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2\right] \le \frac{2L}{3}\left[f(x) - f(x^*)\right].$$

Goal: Bound ↑

Approach: $g_i(x) = f_i(x) - \langle \nabla f_i(x^*), x \rangle$

- $g_i(x)$: L-smooth, convex
- $f_i(x)$: L-smooth, convex
- $\langle \nabla f_i(x^*), x \rangle$: linear

$$\nabla g_i(x^*) = \nabla f_i(x^*) - \nabla f_i(x^*) = 0$$

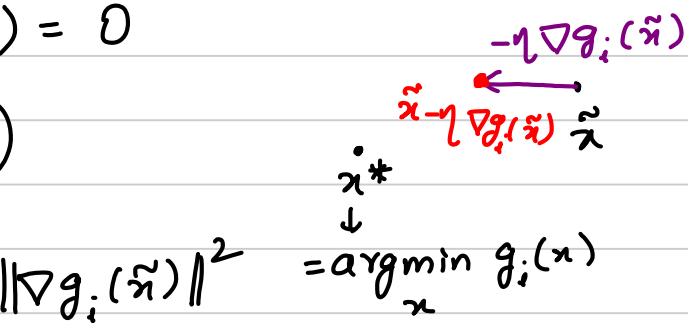
$$\Rightarrow g_i(x^*) \overset{\text{Conv. of } g_i}{\leq} g_i\left(\tilde{x} - \eta \nabla g_i(\tilde{x})\right)$$

$$\overset{L\text{-Smooth}}{\leq} g_i(\tilde{x}) - \underbrace{\eta\left(2 - \frac{\eta L}{2}\right)}\underbrace{\|\nabla g_i(\tilde{x})\|^2}$$

$$\eta = \frac{1}{L} \longrightarrow \leq g_i(\tilde{x}) - \frac{3}{2L}\|\nabla g_i(\tilde{x})\|^2$$

$$\|\nabla g_i(\tilde{x})\|^2 \leq \left[g_i(\tilde{x}) - g_i(x^*)\right]\frac{2L}{3}.$$

$$\|\nabla f_i(\tilde{x}) - \nabla f_i(x^*)\|^2 \leq \left[g_i(\tilde{x}) - g_i(x^*)\right]\frac{2L}{3}$$



Taking an average,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{x})-\nabla f_i(x^*)\|^2 \le \frac{1}{n}\cdot\frac{2L}{3}\sum_{i=1}^{n}\Big[\underbrace{g_i(\tilde{x})-g_i(x^*)}_{f_i(\tilde{x})-\langle\nabla f_i(x^*),\tilde{x}\rangle}\Big]$$

$$= \frac{2L}{3n}\sum_{i=1}^{n} f_i(\tilde{x}) - f_i(x^*) - \langle\nabla f_i(x^*),\ \tilde{x}-x^*\rangle$$

$$= \frac{2L}{3}\Big\{\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{x}) - \frac{1}{n}\sum_{i=1}^{n} f_i(x^*)$$

$$-\Big\langle\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^*),\ \tilde{x}-x^*\Big\rangle$$

$$= \frac{2L}{3}\left\{ f(\tilde{x}) - f(x^*) - \underbrace{\langle \nabla f(x^*), \tilde{x} - x^* \rangle}_{0} \right\}$$

Lemma: $\mathbb{E}\left[\| \nabla f_i(\tilde{x}) - \nabla f_i(x^*) \|^2 \right] \leq \frac{2L}{3} \left[f(\tilde{x}) - f(x^*) \right]$.

Furthermore,

$$\sigma^2 \triangleq \mathbb{E}\left[\| \nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0) - \nabla f(x_t) \|^2 \right]$$

[2nd & 3rd page Computation]

$$\leq 8 \left\{ \mathbb{E}\left[\| \nabla f_i(x_t) - \nabla f_i(x^*) \|^2 \right] + \mathbb{E}\left[\| \nabla f_i(x_0) - \nabla f_i(x^*) \|^2 \right] \right\}.$$

$$\leq \frac{16L}{3} \left[f(x_t) - f(x^*) + f(x_0) - f(x^*) \right].$$

SVRG : $x_{t+1} = x_t - \eta \left[\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0) \right]$

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] = \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta \left\langle \mathbb{E}[\quad], x_t - x^* \right\rangle + \eta^2 \mathbb{E}\left[\|\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0)\|^2\right]$$

$$= \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta \left\langle \nabla f(x_t), x_t - x^* \right\rangle + \eta^2 \|\nabla f(x_t)\|^2 + \eta^2 \sigma^2$$

[By L-Smoothness of $f(\cdot)$, $\|\nabla f(x_t)\|^2 \le L\left[f(x_t) - f(x^*)\right]$]

$$\le \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta \left[f(x_t) - f(x^*)\right] + \eta^2 L \left(f(x_t) - f(x^*)\right)$$

$$+\eta^2 \cdot \frac{16L}{3}\left[f(x_t)-f(x^*)+f(x_0)-f(x^*)\right]$$

$$\mathbb{E}\left[\|x_{t+1}-x^*\|^2\right] \le \mathbb{E}\left[\|x_t-x^*\|^2\right] - \eta\left(2-\frac{19\eta L}{3}\right)\left(f(x_t)-f(x^*)\right)$$

$$+\frac{16\eta^2 L}{3}\left(f(x_0)-f(x^*)\right).$$

Telescoping and averaging,

$$\frac{1}{T+1}\sum_{t=0}^{T}\left[\mathbb{E}[f(x_t)]-f(x^*)\right] \le \frac{\|x_0-x^*\|^2}{\eta\left(2-\frac{19\eta L}{3}\right)(T+1)} + \frac{16\eta L}{3\left(2-\frac{19\eta L}{3}\right)}\left[f(x_0)-f(x^*)\right].$$

f is μ-Str. cvx. : $f(x_0)-f(x^*) \ge \frac{\mu}{2}\|x_0-x^*\|^2$.

Av. exp. Suboptimality $\leq$ $\left[\underbrace{\frac{2}{\mu\eta\left(2-\frac{19\eta L}{3}\right)(T+1)}}_{\frac{L}{\mu}=K} + \underbrace{\frac{16\eta L}{3\left(2-\frac{19\eta L}{3}\right)}}_{\leq \frac{1}{10}}\right] \underbrace{\{f(x_0)-f(x^*)\}}_{\text{Initial Suboptimality}}$

$\eta = \frac{1}{100 \cdot L}$

$\approx \left[\frac{2K}{T+1} + \frac{1}{10}\right]$ [Initial Suboptimality].

$T > 6 \cdot K \Rightarrow$ Av. exp. Suboptimality $\leq \frac{1}{2}$ Initial Subopt.

Pick x_0

For epochs $= 1, \ldots \log(\frac{1}{\epsilon})$.

SVRG epoch:

- Compute $\nabla f(x_0) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_0)$ ← $O(nd)$
- For $t = 0, \ldots T = 5n$ → $O(nd)$
  - Pick i u.a.r. from $1 \ldots n$.
  - $x_{t+1} = x_t - \eta \left[\nabla f_i(x_t) - \nabla f_i(x_0) + \nabla f(x_0) \right]$

Reset $x_0 = x_1, \ldots x_{T+1}$ uniformly at random.

$$O\left(\underbrace{(nd + nd)}_{} \underbrace{\log \frac{1}{\epsilon}}_{} \right)$$

$$f_i(x) = (a_i^T x - b_i)^2$$

$$O\left((nd + kd) \log \frac{1}{\epsilon} \right)$$

$$\downarrow \quad \max_i \|a_i\|^2$$

$$\frac{}{\mu}$$

Importance Sampling

i w.p. $p_i \propto \|a_i\|^2$.

$$f(x) = \sum_{i=1}^{n} \underbrace{p_i}_{} \underbrace{\left(\frac{1}{np_i} (a_i^T x - b_i)^2 \right)}$$

$$\text{Smoothness}\left(\frac{1}{np_i} (a_i^T x - b_i)^2 \right) = \frac{\sum_{j=1}^{n} \|a_j\|^2}{n}$$

$$p_i = \frac{\|a_i\|^2}{\sum_{j=1}^{n} \|a_j\|^2}$$

Uniform sampling

$$K_{us} = \frac{\max_i \|a_i\|^2}{\mu}$$

Weighted Sampling

$$K_{ws} = \frac{\sum_{j=1}^{n} \|a_j\|^2}{n\mu}$$

$$K_{ws} \leq K_{us}$$