Lecture 19: **Stochastic gradient descent** E1 260

(Contd)

- Variance reduction (SVRG)
  - Algorithm
  - Convergence

TA 2nd session: Nov. 3rd (Wed.) 18:00 – 19:00

Smooth and strongly convex functions

ERM: $\underset{\underline{x}}{\text{minimize}} \quad f(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

- f_i is β - smooth
- f is α - strongly convex

Example:

$$f(x) = \frac{1}{n} \|A\underline{x} - \underline{y}\|_2^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{(\underline{a}_i^T \underline{x} - y_i)^2}_{f_i(x)}$$

- SGD needs small step size
- no self tuning ($\tilde{g}_x \not\to 0$ as $\underline{x} \to \underline{x}^*$)
- mini-batch reduces variance, but does not yield self tuning

SVRG: Stochastic variance reduced gradient

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad ; \quad \tilde{g}_x = \nabla f_I(x)$$

$$I \sim \text{unif}(1, \dots, n)$$

Key observation:

Any x and z

$$\tilde{g}_x = \nabla f_I(x) - \left(\nabla f_I(z) - \nabla f(z)\right)$$

(Recentered gradient)

is a stochastic gradient.

$z \simeq x \longrightarrow 0$ $\qquad$ $z = x^* \longrightarrow 0$

- unbiasedness:

$$E_I[\tilde{g}_x] = E[\nabla f_I(x)] - \left(E_I[\nabla f_I(z)] - \nabla f(z)\right)$$

$$= \nabla f(x) - \cancel{\nabla f(z)} + \cancel{\nabla f(z)}$$

$$= \nabla f(x).$$

- Reducing variance by recentering: z is a history point x^{old} and ∇f is the full gradient

The algorithm:

- Operates in epochs:

≫ In the k^{th} epoch: take a snapshot of the current iterate $x_k^{old} = y_k$ and compute the batch gradient $\nabla f(\underline{x}_k^{old})$

≫ Inner loop:

$$\underline{x}_k^{t+1} = \underline{x}_k^t - \eta \{ \nabla f_{i_t}(\underline{x}_k^t) - (\nabla f_{i_t}(\underline{x}_k^{old}) - \nabla f(\underline{x}_k^{old}) \}$$

Outer loop:

k^{th} iteration

Set $\underline{x}_1 = \underline{y}_k$

Inner loop: for $t = 1, \ldots, T$ do

$I \sim \text{unif}(1, \ldots, n)$

$$\underline{x}_{t+1} = \underline{x}_t - \eta \{ \nabla f_I(\underline{x}_t) - (\nabla f_I(\underline{y}_k) - \nabla f(\underline{y}_k)) \}$$

update: $\underline{y}_{k+1} = \frac{1}{T} \sum_{t=1}^{T} \underline{x}_t$

- Batch gradient is computed once per epoch (expensive)

- Inner loop: Requires the same effort as SGD to compute $\nabla f_I(x_t)$

- Take advantage of both worlds: batch and SGD

Variance reduction lemma:

Let $\{f_i\}$ be β- smooth and $I \sim \text{unif}[1, \dots, n)$

$$\mathbb{E}_I \left[\underbrace{\| \nabla f_I(\underline{x}) - \nabla f_I(\underline{x}^*) \|_2^2}_{} \right] \leq 2\beta \left[f(\underline{x}) - f(\underline{x}^*) \right]$$

diff. goes to zero as $x \to x^*$
and nothing about $\nabla f_I(x)$

Proof:

$$g_i(\underline{x}) = f_i(x) - \left[f_i(\underline{x}^*) + \nabla f_i^T(\underline{x}^*)(\underline{x} - \underline{x}^*) \right] \geq 0$$

[$f_i(x)$ is convex ; $\Rightarrow$ $g_i(\underline{x})$ is convex]

Recall : if h is convex and β- Smooth

$$h(\underline{y}) \leq h(\underline{x}) + \nabla h^T(\underline{x})(\underline{y} - \underline{x}) + \frac{\beta}{2} \| \underline{x} - \underline{y} \|_2^2$$

$$y := \underline{x} - \frac{1}{\beta} \nabla h(\underline{x})$$

$$\Rightarrow \quad h\left(\underline{x} - \frac{1}{\beta}\nabla h(\underline{x})\right) \le h(\underline{x}) + \nabla h^{\top}(\underline{x})\left(-\frac{1}{\beta}\nabla \underline{h}(\underline{x})\right) + \frac{\beta}{2}\left\| -\frac{1}{\beta}\nabla \underline{h}(\underline{x})\right\|_2^2$$

$$= h(\underline{x}) - \frac{1}{2\beta}\|\nabla h(\underline{x})\|_2^2$$

Apply to $g_i(\underline{x})$:

$$0 \le g_i\left(x - \frac{1}{\beta}\nabla g_i(x)\right) \le g_i(\underline{x}) - \frac{1}{2\beta}\|\nabla g_i(\underline{x})\|_2^2$$

$$\Rightarrow \quad -g_i(\underline{x}) \le -\frac{1}{2\beta}\|\nabla g_i(\underline{x})\|_2^2$$

$$\Rightarrow \quad \|\nabla g_i(\underline{x})\|_2^2 \le 2\beta\, g_i(\underline{x})$$

Substitute for $\nabla g_i(x) = \nabla f_i(\underline{x}) - \nabla f_i(x^*)$

$$\Rightarrow \quad \| \nabla f_i(\underline{x}) - \nabla f_i(\underline{x}^*) \|_2^2 \leq 2\beta \left[f_i(\underline{x}) - (f_i(x^*) - \nabla f_i^T(\underline{x}^*)(\underline{x} - \underline{x}^*) \right]$$

$$\mathbb{E}\left[\| \nabla f_I(x) - \nabla f_I(\underline{x}^*) \|_2^2 \right]$$

$$\leq 2\beta \left[\underbrace{\mathbb{E}\left[f_I(\underline{x})\right.}\; - f_I(\underline{x}^*) \right] + \mathbb{E}\left[\underbrace{\nabla f_I^T(\underline{x}^*)}_{\nabla f(\underline{x}^*)} (\underline{x} - \underline{x}^*) \right]$$

$$0$$

$$\mathbb{E}\left[f_I(\underline{x}) \right] = \sum_{i=1}^{n} \frac{1}{n} f_i(x)$$

$$= f(\underline{x})$$

$$\mathbb{E}\left[\| \nabla f_I(x) - \nabla f_I(\underline{x}^*) \|_2^2 \right] \leq 2\beta \left(f(\underline{x}) - f(\underline{x}^*) \right)$$

Convergence analysis of SVRG:

Let f be α strongly convex and $\{f_i\}$ be β-smooth, then SVRG with a fixed step size $\eta = \frac{1}{10\beta}$ and inner loop size

$T = 10\left(\beta/\alpha\right)$ [β/α : condition number]

Then, after $s+1$ epochs (outer loop)

$$\mathbb{E}\left[f(\underline{y}_{s+1}) - f(\underline{x}^*)\right] \leq 0.9^s \left(f(\underline{y}_1) - f(\underline{x}^*)\right)$$

- Linear convergence
- Taking no. of inner loop iteration as a factor (β/α); convergence does n't depend of β/α; unlike GD.

Proof:

$$\mathbb{E}\left[f(\underline{y}_{s+1}) - f(\underline{x}^*)\right] \le 0.9\left(f(\underline{y}_s) - f(\underline{x}^*)\right)$$

Recall:

$$\underline{y}_{s+1} = \frac{1}{T}\sum_{t=1}^{T} \underline{x}_t \; ; \; \underline{x}_t \text{ is from the } s^{th} \text{ epoch. } (\underline{x}_t^s)$$

Similarly, let us use $\underline{y}$ (instead of $\underline{y}_{s+1}$)

$$\|\underline{x}_{t+1} - \underline{x}^*\|_2^2 = \|\underline{x}_t - \eta\left[\nabla f_{i_t}(\underline{x}_t) - (\nabla f_{i_t}(\underline{y}) - \nabla f(\underline{y})\right] - \underline{x}^*\|_2^2$$

$$= \underbrace{\|\underline{x}_t - \underline{x}^*\|_2^2}_{A} + \underbrace{\eta^2\|\underline{u}_t\|^2}_{B} - \underbrace{2\eta\,\underline{u}_t^T(\underline{x}_t - \underline{x}^*)}_{C}$$

$$\underline{u}_t = \nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{y}) + \nabla f(\underline{y})$$

- Recall (B) from SGD is the variance term was not going to zero. So we took small η

$$\circledast \quad \mathbb{E}_I\left[\|\underline{u}_t\|_2^2\right] = \mathbb{E}_I\left[\|\nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{y}) + \nabla f(\underline{y}) - \nabla f_{i_t}(\underline{x}^*) + \nabla f_{i_t}(\underline{x}^*)\|_2^2\right]$$

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

$$\le 2\,\mathbb{E}_I\left[\|\nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{x}^*)\|_2^2\right]$$

$$+ 2\,\mathbb{E}_I\left[\|\nabla f_{i_t}(\underline{y}) - \nabla f(\underline{y}) - \nabla f_{i_t}(\underline{x}^*)\|_2^2\right]$$

$\mathbb{E}\left[\nabla f_{i_t}(\underline{y}) - \nabla f(\underline{y}) - \nabla f_{i_t}(\underline{x}^*)\right] = 0$ and $\mathbb{E}\left[\|\underline{z} - \mathbb{E}(\underline{z})\|^2\right] \le \mathbb{E}\left[\|\underline{z}\|^2\right]$

$$\circledast \quad \mathbb{E}_I\left[\|\underline{u}_t\|_2^2\right] \le 2\,\mathbb{E}_I\left[\|\nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{x}^*)\|_2^2\right] + 2\,\mathbb{E}_I\left[\|\nabla f_{i_t}(\underline{y}) - \nabla f_{i_t}(\underline{x}^*)\|_2^2\right]$$

Recall variance reduction lemma:

$$\mathbb{E}_I\left[\|\nabla f_I(\underline{x}) - \nabla f_I(\underline{x}^*)\|_2^2\right] \leq 2\beta\left[f(\underline{x}) - f(\underline{x}^*)\right]$$

$$(*) \quad \mathbb{E}_I\left[\|\underline{u}_t\|_2^2\right] \leq 4\beta\left[f(\underline{x}_t) - f(\underline{x}^*) + f(y) - f(\underline{x}^*)\right]$$

Now, let work with (C):

$$2\eta\, u_t^T(\underline{x}_t - \underline{x}^*)$$

$$\mathbb{E}\left[2\eta\, \underline{u}_t^T(\underline{x}_t - \underline{x}^*)\right] = 2\eta\, \mathbb{E}[\underline{u}_t]^T(\underline{x}_t - \underline{x}^*)$$

$$= 2\eta\, \nabla f^T(\underline{x})(\underline{x}_t - \underline{x}^*)$$

(Convexity) $$\geq 2\eta\left(f(\underline{x}_t) - f(\underline{x}^*)\right)$$

Combining everything:

$$
\begin{aligned}
\mathbb{E}_{I}\left[\|\underline{x}_{t+1} - \underline{x}^{*}\|\right] &\leq \|\underline{x}_{t} - \underline{x}^{*}\|_2^2 \\
&\quad + 4\beta\eta^2\left[f(\underline{x}_t) - f(\underline{x}^{*}) + f(\underline{y}) - f(\underline{x}^{*})\right] \\
&\quad - 2\eta\left[f(\underline{x}_t) - f(\underline{x}^{*})\right] \\
&\leq \|\underline{x}_t - \underline{x}^{*}\|_2^2 \\
&\quad - 2\eta(1-2\beta\eta)\left[f(\underline{x}_t) - f(\underline{x}^{*})\right] \\
&\quad + 4\eta^2\beta\left[f(\underline{y}) - f(\underline{x}^{*})\right]
\end{aligned}
$$

Iterating this inequality:

$$\mathbb{E}_I\left[\|x_{t+1} - x^*\|_2^2\right] \le \|x_1 - x^*\|_2^2 - 2\eta(1-2\beta\eta)\,\mathbb{E}\left[\sum_{k=1}^{T} f(x_k) - f(x^*)\right] + 4\eta^2\beta \cdot T \cdot \left[f(y) - f(x^*)\right]$$

- we have $x_1 = y$ (our initialization)
- f is α strongly convex

$$\|y - x^*\|_2^2 \le \frac{2}{\alpha}\, f(y) - f(x^*)$$

$$2\eta\left(1 - 2\beta\eta\right)\left(\mathbb{E}\, f\left(\frac{1}{T}\sum_k x_k\right) - f(x^*)\right) \le \left(\frac{2}{\alpha} + 4\eta^2\beta T\right) \cdot \frac{1}{T}\left[f(y) - f(x^*)\right]$$

$$\Rightarrow \mathbb{E}\, f(\underline{y}_{s+1}) - f(\underline{x}^*) \leq \frac{\left(\frac{2}{\alpha} + 4\eta^2\beta T\right)\cdot\frac{1}{T}}{2\eta\,(1-2\beta\eta)}\cdot\left[f(\underline{y}_s) - f(\underline{x}^*)\right]$$

$$= 0.9\left[f(\underline{y}_s) - f(\underline{x}^*)\right]$$

with

$$\eta = \frac{1}{10\beta} \quad \text{and} \quad T = 10\cdot\left(\beta/\alpha\right)$$

GD vs. SGD vs. SVRG

GD: $$x_{t+1} = x_t - \eta \nabla f(x_t) \quad : \quad [\, n \text{ grad. comp.}]$$

SGD: $$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t) \quad : \quad [\, 1 \text{ grad. comp}]$$

SVRG: $$x_{t+1} = x_t - \eta \left[\nabla f_{i_t}(x_t) - \left(\nabla f_{i_t}(y) - \nabla f(y) \right) \right] \quad : \quad [\, n + T \text{ grad comp.}]$$

$$\approx n + 10 \cdot (\beta / \alpha)$$

depends on the condn. number

GD vs. SGD vs. SVRG: ε- accuracy

GD: $O\left(\left(\beta/\alpha\right)\log\left(1/\varepsilon\right)\right)$ iterations

Momentum $O\left(\sqrt{\beta/\alpha}\,\log\left(1/\varepsilon\right)\right)$

So $O\left(n\cdot\beta/\alpha\,\log\left(1/\varepsilon\right)\right)$ grad. computations

SGD: $O\left(\frac{1}{\alpha\varepsilon}\right)$ iterations

$= O\left(\frac{1}{\alpha\varepsilon}\right)$ grad. computations

SVRG: $O\left(\log\left(1/\varepsilon\right)\right)$ iteration

$O\left(\left(n+\beta/\alpha\right)\log\left(1/\varepsilon\right)\right)$ grad. computation