• Last time: SGD

- #iter(ε) worse than GD
- cost/iter $\approx$ $n \times$ faster than GD

Today's Lecture:

Q: Can we have the best of both worlds, i.e.,
convergence as fast as GD
and $O(1)$ grad. computation/iter like SGD

A: SVRG

We would like to understand what causes SGD to have slower rates than GD

We will revisit the finite sum setup:

$$f(w) = \frac{1}{n} \sum_i f_i(w)$$

Quick comparison:

- SGD on λ-str. cvx:

$$\mathbb{E}\|w_{k+1}-w^*\|^2 \le (1-\gamma\lambda)\mathbb{E}\|w_k-w^*\|^2+\gamma^2\mathbb{E}\|\nabla f_{s_k}(w_k)\|^2$$

[Due to bound from last lecture] $\le (1-\gamma\lambda)^{k+1}\|w_0-w^*\|^2 + \frac{\gamma}{\lambda}M^2$

- GD on λ-str cvx + β-smooth

Due to smoothnes we have:

$$\|\nabla f(w_k)\|^2 = \|\nabla f(w_k) - \nabla f(w^*)\|^2$$
$$\le \beta^2\|w_k-w^*\|^2$$

Hence a simple conv. rate bound gives:

$$\|w_{k+1}-w^*\|^2 \le (1-\gamma\lambda)\|w_k-w^*\|^2+\gamma^2\beta^2\|w_k-w^*\|^2$$
$$\le (1-\gamma\lambda+\gamma^2\beta^2)^{k+1}\cdot\|w_k-w^*\|^2$$

Observe that SGD rates look like:

$$C_1^K \cdot \|w_0 - w^*\|^2 + V$$

And for GD : $C_2^k \|w_0 - w^*\|^2$

The "V"-term causes worse rates in SGD

Q: Why? Can we fix it?

Remark: We can't take advantage of smoothness in SGD since

$$\|\nabla f_{s_k}(w_k)\|^2 \leq B_{s_k} \|w_K - w^*_{s_k}\|^2$$

That is smoothness gives us an upper bound, but only with respect to the global opt of a single function and in general

$$\arg\min_w f_i(w) \neq \arg\min_w \sum f_i(w)$$

However, we can do a trick:

$$
\begin{aligned}
\|\nabla f_{s_k}(w_k)\|^2 &\leq \|\nabla f_{s_k}(w_k) - \nabla f_{s_k}(w^*) + \nabla f_{s_k}(w^*)\|^2 \\
&\leq 2\|\nabla f_{s_k}(w_k) - \nabla f_{s_k}(w^*)\|^2 + 2\|\nabla f_{s_k}(w^*)\|^2 \\
&\leq \underbrace{2 \cdot \beta_{s_k} \|w_k - w^*\|}_{A} + \underbrace{2 \cdot \sigma_{s_k}}_{B}
\end{aligned}
$$

A looks like the term in GD and B measures how large the grad. of f_{s_k} is at the global min of $\sum_i f_i(w)$.

Remark:

When $A \geq B \Rightarrow$ SGD is in the linear rate regime

i.e. variance decays with # iter.

__What we want:__ A variant of SGD, e.g.

$$w_{k+1} = w_k - \gamma \cdot g_k(w_k)$$

such that:

- $E g_k = \nabla f$
- g_k is "cheap" on average
- __$A \geq B$ always__

This is possible!

SVRG:
Stochastic variance reduced gradient method.

In SVRG we choose g_k as follows:

$$g_k(w) = \nabla f_{s_k}(w) \overbrace{- \nabla f_{s_k}(w_0) + \underbrace{\nabla f(w_0)}_{\text{full grad.}}}^{\text{This term will allow the } A \geq B \text{ property}}$$

Lets bound the variance of g_k

$$\mathbb{E}\|g_k(w)\|^2 = \mathbb{E}\|g_{k}(w) \pm \nabla f_{s_k}(w^*)\|^2$$

$$= \mathbb{E}\|\nabla f_{s_k}(w^*) - \nabla f_{s_k}(w_0) + \nabla f(w_0) \pm \nabla f_{s_k}(w^*)\|^2$$

$$\le 2\mathbb{E}\|\nabla f_{s_k}(w_k) - \nabla f_{s_k}(w^*)\|^2 +$$
$$2\mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*) - \nabla f(w_0)\|^2$$

$$\le 2\beta\, \mathbb{E}\|w_k - w^*\|^2 + (\quad \cdot \quad)$$

Observe now that

$$\mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*) - \nabla f(w_0)\|^2$$
$$= \mathbb{E}\|\underbrace{\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*)}_{X} \underbrace{- \nabla f(w_0) + \nabla f(w^*)}_{\mathbb{E}X}\|$$
$$= \mathbb{E}\| X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$$

$$= \mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*)\|^2$$
$$\le \beta\, \mathbb{E}\|w_0 - w^*\|^2$$

With the above "update rule" we then get:

If f is λ-str. cvx and each f_i is β-smooth

$$\mathbb{E}\|w_{k+1}-w^*\|^2 \le \mathbb{E}\|w_k - x^*\|^2 - (2\gamma\lambda + 2\gamma^2\beta)\mathbb{E}\|w_k - w^*\|^2$$
$$+ 2\gamma^2\beta^2\,\mathbb{E}\|w_0 - w^*\|^2$$

$$\vdots$$

$$\le \underbrace{(1-2\gamma\lambda+2\gamma^2\beta)^{k+1}}_{1/4}\|w_0-w^*\|^2 + \underbrace{2(k+1)\gamma^2\beta^2}_{1/4}\|w_0-w^*\|^2$$

set $\quad 2(k+1)\gamma^2\beta^2 = 1/4$

$$\Rightarrow \quad \gamma = O(1)\cdot\frac{1}{\beta\cdot k}$$

set $\quad (1-2\gamma\lambda+2\gamma^2\beta)^{k+1} = 1/4$

$$\Rightarrow \left(1-\frac{c\cdot\lambda}{\beta\cdot k}+\frac{c'}{\beta k^2}\right)^{k+1} = 1/4$$

Setting $k = O(1)\cdot B^2/\lambda^2$ gives the above

Hence if $\quad \gamma = O(1)\,\frac{\lambda}{\beta^2}, \; k = O(1)\cdot B^2/\lambda^2$

$$\Rightarrow \quad \mathbb{E}\|w_T - w^*\|^2 \le 0.5\,\|w_0 - w^*\|^2$$

Remarks:

- The above only decreases dist. to opt by a constant factor
- The step
  $g_k(w) = \nabla f_{s_k}(w) - \nabla f_{s_k}(w_0) + \nabla f(w_0)$
  costs 1 full grad, i.e., its complexity is proportional to GD

Solution to above:

Repeat in "epochs"

SVRG:

$y = w_0$

$t = 0$

for epoch $= 1 : E$

$\quad g = \nabla f(y)$

$\quad$ for $s = 1 : S$

$\quad\quad s_t \sim \text{unif}\{1, \dots, n\}$

$\quad\quad w_{t+1} = w_t - \gamma \left(\nabla f_{s_t}(w_t) - \nabla f_{s_t}(y) + g \right)$

$\quad\quad t = t+1$

$\quad y = w_{K-1}$

Observe that the cost of $g = \nabla f(y)$ becomes amortized

Also overall rate

$$\mathbb{E}\|w_E - w^*\|^2 \leq 0.5^E \|w_0 - w^*\|$$

similar to GD

Hence,

\# epochs to ε accuracy: $O(\log(1/\varepsilon))$

\# iterations/epoch : $K = O(\beta^2/\lambda^2)$

computational cost/epoch: 1 full grad
S "small" grads

Overall complexity of SVRG

$$O\left(\log(1/\varepsilon) \cdot \text{cost}(\nabla f) + \frac{\beta^2}{\lambda^2} \log(1/\varepsilon) \cdot \frac{\text{cost}(\nabla f)}{n} \right)$$

Compare to GD: $O\left(\frac{\beta^2}{\lambda^2} \log(1/\varepsilon) \cdot \text{cost}(\nabla f) \right)$.

Remarks:

- SVRG has linear rate of convergence like GD and small amortized cost/iter like SGD
- However we have to tune more hyperparam i.e, stepsize + length of each epoch.

Open Problems:

- What happens if we first run vanilla SGD and then full GD or SVRG
- Adaptively change estimate $g_k(w)$ can be coarser in the begining and finer towards the end
- What is the best way to chose g_k to optimize rates?
- Performance on non-convex problems is questionable. why?