- 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

   \[ O(1) \] 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:

- **S UD on 2-str. cvx:**

  \[\|w_{k+1} - w^*\|^2 \leq (1 - \gamma) \|w_k - w^*\|^2 + 2 \|E\| \|^2\|f_{sk}(w_k)\|^2\]

  Due to bound \[
  \leq (1 - \gamma) \|w_0 - w^*\|^2 + \frac{\delta}{\alpha} M^2
  \]

- **UD on 2-str. cvx + B-smooth**

Due to smoothness we have:

\[
\|\nabla f(w_k)\|^2 = \|\nabla f(w_k) - \nabla f(w^*)\|^2 \leq \beta^2 \|w_k - w^*\|^2
\]

Hence a simple convex rate bound gives:

\[
\|w_{k+1} - w^*\|^2 \leq (1 - \gamma) \|w_k - w^*\|^2 + \frac{\beta^2}{\alpha} \|w_k - w^*\|^2 \leq (1 - \gamma^2 + \frac{\beta^2}{\alpha}) \|w_k - w^*\|^2
Observe that SGD rates look like:

$$C_1 \cdot \| \omega_0 - \omega^* \|^2 + \nu$$

And for GD:

$$C_2 \cdot \| \omega_0 - \omega^* \|^2$$

The "\( \nu \)"-term causes worse rates in SGD.

Q: Why? Can we fix it?

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

$$\| \nabla f(w_k) \|^2 \leq B_{sk} \| w_k - w^* \|^2$$

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

$$\arg \min_w f(w) \neq \arg \min_w \Sigma f_i(w)$$
However, we can do a trick:

\[
\|\nabla f_{sk}(w_k)\|^2 \leq \|\nabla f_{sk}(w_k) - \nabla f_{sk}(w^*) + \nabla f_{sk}(w^*)\|^2 \\
\leq 2\|\nabla f_{sk}(w_k) - \nabla f_{sk}(w^*)\|^2 + 2\|\nabla f_{sk}(w^*)\|^2 \\
\leq 2B_{sk} \|w_k - w^*\| + 2.6_{sk} \\
\text{A} + \text{B}
\]

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

**Remark:**

When A>B => SGD is in the linear rate regime i.e. variance decays with #iter.
What we want: A variant of SAG, e.g.

\[ w_{k+1} = w_k - g \cdot g_k(w_k) \]

such that:

- \( E g_k = \nabla f \)
- \( g_k \) is "cheap" on average
- \( A \Rightarrow B \) always

This is possible!

SVRG: Stochastic variance reduced gradient method.

In SVRG we choose \( g_k \) as follows:

\[ g_k(w) = \nabla f_{sk}(w) - \nabla f_{sk}(w_0) + \nabla f(w_0) \]

This term will allow the \( A \Rightarrow B \) property.

full grad.
Let's bound the variance of \( g_k \)

\[
\mathbb{E} \| g_k (w) \|^2 = \mathbb{E} \| g_k (w) + \nabla f_{sk}(w) \|^2 \\
= \mathbb{E} \| \nabla f_{sk}(w^*) - \nabla f_{sk}(w) + f(w) + \nabla f(w) \|^2 \\
\leq 2 \mathbb{E} \| \nabla f_{sk}(w) - \nabla f_{sk}(w^*) \|^2 + \\
2 \mathbb{E} \| \nabla f_{sk}(w) - \nabla f_{sk}(w^*) - \nabla f(w) \|^2 \\
\leq 2 \nu \| f_{sk} - g_{sk} \|^2 + (\cdot)
\]

Observe now that

\[
\mathbb{E} \| \nabla f_{sk}(w) - \nabla f_{sk}(w^*) - \nabla f(w) \|^2 \\
= \mathbb{E} \| \nabla f_{sk}(w) - \nabla f_{sk}(w^*) - \nabla f(w) + \nabla f(w^*) \|^2 \\
\times \mathbb{E} X \\
= \mathbb{E} \| X \|^2 \leq \mathbb{E} \| X \|^2 \\
= \mathbb{E} \| \nabla f_{sk}(w) - \nabla f_{sk}(w^*) \|^2 \\
\leq \nu \mathbb{E} \| w^* - w \|^2
\]
With the above "update rule" we then get:

If \( f \) is strongly convex and each \( f_i \) is \( \beta \)-smooth

\[
\|E\|w_{k+1} - w^*\|^2 \leq \|E\|w_k - x^*\|^2 - (2\gamma + 2\gamma^2 \beta) \|E\|w_k - w^*\|^2
\]

\[
+ 2\gamma^2 \beta^2 \|E\|w_0 - w^*\|^2
\]

\[
\leq \left( 1 - 2\gamma + 2\gamma^2 \beta \right)^{k+1} \|w_0 - w^*\|^2 + 2(k+1)\gamma^2 \beta^2 \|w_0 - w^*\|^2
\]

\[
\leq \left( 1 - 2\gamma + 2\gamma^2 \beta \right)^{k+1} \left( \|w_0 - w^*\|^2 \right)
\]

\[
\frac{1}{4} \quad \frac{1}{4}
\]

set \( 2(k+1)\gamma^2 \beta^2 = \frac{1}{4} \)

\[
\Rightarrow \gamma = \mathcal{O}(1) \cdot \frac{1}{\beta \cdot k}
\]

set \( (1 - 2\gamma + 2\gamma^2 \beta)^{k+1} = \frac{1}{4} \)

\[
\Rightarrow 1 - \frac{c_0^2}{\beta \cdot k} + \frac{c_1^2}{\beta k^2}
\]

setting \( k = \mathcal{O}(1) \cdot B^2/\gamma^2 \) gives the above

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

\[
\Rightarrow \|E\|w_{T} - w^*\|^2 \leq 0.5 \|w_0 - w^*\|^2
\]
Remarks:  
- The above only decreases dist. to opt by a constant factor.
- The step
  \[ g_k(w) = \nabla f_k(w) - \nabla f_k(w_0) + \nabla f(w) \]
  costs 1 full grad, i.e., its complexity is proportional to 2D.

Solution to above:

Repeat in "epochs".

SVRG:

\[ y = w_0 \]
\[ t = 0 \]

for epoch = Δ: E

\[ q = \nabla f(y) \]

for s = 1: S

\[ s+ \sim \text{unif } \{1, \ldots, y\} \]

\[ w_{t+1} = w_t - \gamma (\nabla f_{s+1}(w_t) - \nabla f_s(y) + q) \]

\[ t = t + 1 \]

\[ y = w_{k-1} \]
Observe that the cost of \( q = \nabla f(y) \) becomes amortized

Also overall rate

\[ \| E \|_2^2 \leq 0.5^E \| w_0 - w^* \|_2^2 \]

Similar to GD

Hence,

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

\# iterations/epoch: \( K = O(\frac{B^2}{\varepsilon^2}) \)

Computational cost/epoch: A full grad + \( n \) "small" grads

Overall complexity of SVRG

\[ O\left( \log(1/\varepsilon) \cdot \text{cost}(\nabla f) + \frac{B^2}{\varepsilon^2} \log(1/\varepsilon) \cdot \frac{\text{cost}(\nabla f)}{n} \right) \]

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

- SVRG has linear rate of convergence like GD and small amortized cost/iter like SGD.

- However, we have to tune more hyperparameters, i.e., stepsize and 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 beginning and finer towards the end.

- What is the best way to choose $g_k$ to optimize rates?

- Performance on non-convex problems is questionable. Why?