ECE 901: Large-scale Machine Learning and Optimization **Spring 2018**

Lecture 7 — February 13

Lecturer: Dimitris Papailiopoulos *Scribe: Jiefeng Chen & Zihang Meng*

Note: These lecture notes are still rough, and have only have been mildly proofread.

7.1 Review

Last time we learn about SGD. Compared with GD, it has the following properties:

- number of iterations to get accuracy ϵ is more than GD.
- Cost per update is less than GD ($\approx$ n times faster than GD).

Now we have a question: Can we have small iteration complexity and also fast convergence? The answer is yes. We can use SVRG.

7.2 Convergence Rate of GD and SGD

We would like to understand what causes SGD to have slower rate than GD. We will revisit the finite sum setup:

$$f(w) = \frac{1}{n}\sum_{p=1}^{n} f_p(w) \tag{7.1}$$

Now we will compare the convergence rate of SGD with that of GD.

- SGD on λ-strong convexity function $f(w)$:

$$\begin{aligned}
\mathbb{E}\|w_{k+1}-w^*\|^2 &\le (1-\gamma\lambda)\mathbb{E}\|w_k-w^*\|^2 + \gamma^2\mathbb{E}\|\nabla f_{s_k}(w_k)\|^2 && (7.2)\\
&\le (1-\gamma\lambda)\mathbb{E}\|w_k-w^*\|^2 + \gamma^2 M^2 && (7.3)\\
&\cdots && (7.4)\\
&\le (1-\gamma\lambda)^{k+1}\|w_0-w^*\|^2 + \frac{\gamma}{\lambda}M^2 && (7.5)
\end{aligned}$$

- GD on λ-strong convexity and B-smooth function $f(w)$:

$$\begin{aligned}
\|w_{k+1}-w^*\|^2 &\le (1-\gamma\lambda)\|w_k-w^*\|^2 + \gamma^2\|\nabla f(w_k)\|^2 && (7.6)\\
&\le (1-\gamma)\|w_k-w^*\|^2 + \gamma^2\beta^2\|w_k-w^*\|^2 && (7.7)\\
&= (1-\gamma\lambda+\gamma^2\beta^2)\|w_k-w^*\|^2 && (7.8)\\
&\le (1-\gamma\lambda+\gamma^2\beta^2)^{k+1}\|w_0-w^*\|^2 && (7.9)
\end{aligned}$$

(7.2) and (7.6) are the bounds from last lecture.
We can observe that the convergence rate of SGD look like this:

$$C_1^k \|w_0 - w^*\|^2 + V \tag{7.10}$$

And for GD, it looks like this:

$$C_2^k \|w_0 - w^*\|^2 \tag{7.11}$$

The V-term causes worse rates in SGD. This is because we cannot take advantage of smoothness in SGD since

$$\|\nabla f_{s_k}(w_k)\|^2 \leq \beta_{s_k} \|w_k - w_{s_k}^*\|^2 \tag{7.12}$$

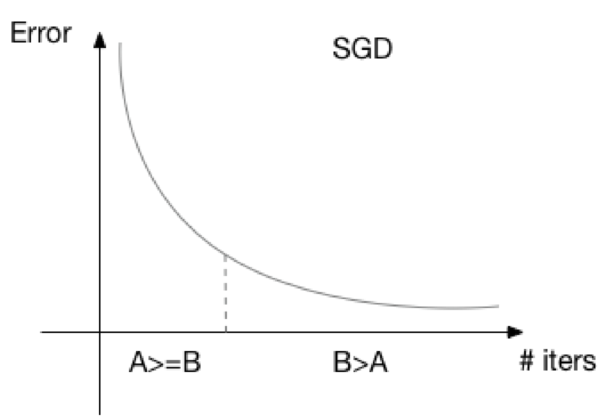
The smoothness gives us an upper bound, but only with respect to the global optimization of a single function and in general

$$arg\min_w f_i(w) \neq arg\min_w \sum_i f_i(w) \tag{7.13}$$

However, we can do a trick:

$$\begin{aligned}
\|\nabla f_{s_k}(w_k)\|^2 &= \|(\nabla f_{s_k}(w_k) - \nabla f_{s_k}(w^*)) + \nabla f_{s_k}(w^*)\|^2 && (7.14)\\
&\leq 2\|\nabla f_{s_k}(w_k) - \nabla f_{s_k}(w^*)\|^2 + 2\|\nabla f_{s_k}(w^*)\|^2 && (7.15)\\
&\leq 2\beta \|w_k - w^*\|^2 + 2\|\nabla f_{s_k}(w^*)\|^2 && (7.16)
\end{aligned}$$

Note that $\nabla f_{s_k}(w^*) \neq 0$. From (7.14) to (7.15), we use the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. Let $A = 2\beta\|w_k - w^*\|^2$ and $B = 2\|\nabla f_{s_k}(w^*)\|^2$. A looks like the term in GD and B measures how large the gradient of f_{s_k} is at the global minimum of $\sum_i f_i(w)$. Note that when $A \geq B$, SGD is in the linear rate regime. (i.e. variance decays with number of iterations).



What we want is a variant of SGD, e.g. $w_{k+1} = w_k - \gamma g_k(w_k)$ (First-order update) such that we have following properties:

- A good converge rate($A \geq B$ is always true).
- Fast update(g_k is "cheap" on average).
- $\mathbb{E}[g_k(w_k)] = \nabla f(w_k)$.

And this is possible. We can use SVRG which will be introduced in the next section.

7.3 Stochastic Variance Reduced Gradient (SVRG)

We can let

$$g_k(w) = \nabla f_{s_k}(w) - \nabla f_{s_k}(w_0) + \nabla f(x_0) \tag{7.17}$$

Then

$$\mathbb{E}[g_k(w)] = \nabla f(w) - \nabla f(w_0) + \nabla f(w_0) = \nabla f(w) \tag{7.18}$$

So this satisfy $\mathbb{E}[g_k(w_k)] = \nabla f(w_k)$.

Lemma: suppose each f_i is λ-strongly convex, then

$$\mathbb{E}\|w_{k+1} - w^*\|^2 \leq (1 - \lambda\gamma)\mathbb{E}\|w_k - w^*\|^2 + \gamma^2\mathbb{E}\|g_k(w_k)\|^2 \tag{7.19}$$

The last term on the right is the "variance".

Let's bound the $\mathbb{E}\|g_k(w)\|^2$.

$$\begin{aligned}
\mathbb{E}\|g_k(w)\|^2 &= \mathbb{E}\|\nabla f_{s_k}(w) - \nabla f_{s_k}(w_0) + \nabla f(w_0) + \nabla f_{s_k}(w^*) - \nabla f_{s_k}(w^*)\|^2 && (7.20)\\
&\leq 2\mathbb{E}\|\nabla f_{s_k}(w) - \nabla f_{s_k}(w^*)\|^2 + 2\mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*) - \nabla f(w_0)\|^2 && (7.21)
\end{aligned}$$

Let the first term be A, the second term be B

$$A \leq 2\beta^2\mathbb{E}\|w - w^*\|^2 \tag{7.22}$$

$$B = 2\mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*) - \nabla f(w_0) + \nabla f(w^*)\|^2 \tag{7.23}$$

Since we have: $2\mathbb{E}\|x - E(x)\|^2 \leq 2\mathbb{E}\|x\|^2$ and $\mathbb{E}[\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*)] = -\nabla f(w^*) + \nabla f(w_0)$. So

$$B \leq 2\mathbb{E}\|\nabla f_{s_k}(w_0) - \nabla f_{s_k}(w^*)\|^2 \leq 2\beta^2\|w_0 - w_k\|^2 \tag{7.24}$$

Then if we suppose that all f is λ-strongly convex, β-smoothness

$$\begin{aligned}
\mathbb{E}\|w_{k+1} - w^*\|^2 &\leq (1 - \gamma\lambda)\mathbb{E}\|w_k - w^*\|^2 + 2\gamma^2\beta^2\mathbb{E}\|w_k - w^*\|^2 + 2\gamma^2\beta^2\|w_0 - w^*\|^2 && (7.25)\\
&\leq (1 - \gamma\lambda + 2\gamma^2\beta^2)^{(k+1)} + 2(k+1)\gamma^2\beta^2\|w_0 - w^*\|^2 && (7.26)
\end{aligned}$$

We want $(1 - \gamma\lambda + 2\gamma^2\beta^2)^{(k+1)} \leq 1/4$ and $2(k+1)\gamma^2\beta^2 \leq 1/4$.
So we can set $k = O(\frac{\beta^2}{\gamma^2})$ and $\gamma = O(1)\frac{\lambda}{\beta^2}$, then we have

$$\mathbb{E}\|w_k - w^*\|^2 \leq \frac{1}{2}\|w_k - w^*\|^2 \tag{7.27}$$

Notice that the above decreasing rate is a constant factor, so we need to do SVRG in "Epochs". The algorithm is showed below:

Algorithm 1 Doing SVRG in Epochs

for epoch=1:E **do**
    g ← ∇ f(y)
    for s=1:S **do**
        $s_t \sim unif\{1, ..., n\}$
        $w_{t+1} \leftarrow w_t - \gamma(\nabla f_{s_t}(w_t) - \nabla f_{s_t}(y) + g)$
        $t \leftarrow t + 1$
    $y \leftarrow w_{k-1}$

If so, we have:

$$\mathbb{E}\|w_E - w^*\|^2 \leq (\frac{1}{2})^E\|w_0 - w^*\|^2 \tag{7.28}$$

It behaves like "Linear Convergence" (like GD)
The cost is:

$$O(log(\frac{1}{\epsilon})) * cost(\nabla f + \frac{\beta^2}{\lambda^2}log(1/\epsilon) + \frac{cost(\nabla f)}{n}) \tag{7.29}$$

7.4 Discuss

For SVRG, there are some issues:

- More hyper-parameters to tune.
- Seems to not do as well on non-convex functions.

And there are some open problems:

1. What happens if we run SGD for a while, then do GD or SVRG?
2. Can $g_k(w)$ be adaptively chosen?
3. How do we pick g_k to minimize number of iterations?
4. Why is SVRG not as good on non-convex functions?