DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: Merlin Chang, Marialena Sfyraki Feb 1, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 8: SGD and Variance Reduction

1 SGD and Variance

Recall a stochastic optimization algorithm in general can be expressed as follows:

Algorithm 1 (General) Stochastic Gradient Descent

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute a stochastic gradient g_k that satisfies $\mathbb{E}[g_k] = \nabla F(x_k)$
3: $x_{k+1} = x_k - \eta g_k$.
4: **end for**

SGD has a slower convergence rate compared to GD, see figure 1 for the illustration. Since at each iteration, we have some randomness of computing the stochastic gradient, which can cause fluctuation during updates. Specifcially, even when the update is close to an optimal point and the full gradient is close to zero, the stochastic gradient may not be close to zero, which slows down the progress.

On the other hand, recall that we derived the iteration complexity of SGD in last lecture:

Theorem 1. *Let $F(x) = \mathbb{E}_z[f(x; z)] : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Consider the update*

$$x_{k+1} = x_k - \eta g_k,$$

where $\mathbb{E}_z[g_k] = \nabla F(x_k)$. Suppose $x_* = \arg\min_x F(x)$ exists and the initial distance is bounded, i.e., $\|x_1 - x_*\| \leq D$. Then,

$$\frac{1}{K}\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\big) \leq \frac{\eta}{2K}\left(\sum_{k=1}^{K} E[\|g_k\|_2^2]\right) + \frac{\|x_1 - x_*\|_2^2}{2\eta K},$$

where $\bar{x}_K := \frac{1}{K}\sum_{k=1}^{K} x_k$.

Does the upper bound reflect the fact that the variance of stochastic gradients slows down the progress? Notice that the term in red is actually the variance in disguise (to be elaborated soon). Since g_k does not reduce as we approach the optimal point, we were only able to bound $E[\|g_k\|_2^2] \leq G$ with a constant. This causes SGD to have $O\left(\frac{1}{\sqrt{K}}\right)$ rate under optimal tuning of η.

(a) SGD (b) GD

Figure 1: Illustration of the progress

Lemma 1. $\mathbb{E}\left[\|g_k\|_2^2\right]$ *is an upper bound of the variance of the stochastic gradient.*

Proof. The variance of g_k can be simplified in the following way:

$$\begin{aligned}
\mathrm{Var}(g_k) &= \mathbb{E}\left[\|g_k - \mathbb{E}(g_k)\|_2^2\right] \\
&= \mathbb{E}\left[\|g_k\|_2^2 - 2{g_k}^\top \mathbb{E}(g_k) + \|\mathbb{E}(g_k)\|_2^2\right] \\
&= \mathbb{E}\left[\|g_k\|_2^2\right] - 2\|\mathbb{E}(g_k)\|_2^2 + \|\mathbb{E}(g_k)\|_2^2 \qquad \text{by linearity of expectation.} \\
&= \mathbb{E}\left[\|g_k\|_2^2\right] - \|\mathbb{E}(g_k)\|_2^2 \\
&\leq \mathbb{E}\left[\|g_k\|_2^2\right]
\end{aligned}$$

□

2 Stochastic Variance Reduced Gradient (SVRG)

To continue, we are going to consider the finite-sum problem;

$$\min_{x\in\mathbb{R}^d} F(x), \text{ where } F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1}$$

The algorithm for SVRG is the following:

Algorithm 2 Stochastic Variance Reduced Gradient Method (SVRG)

Set $s = 1$. Init $v_1 = x_1$. Learning rate η.
for stage $s = 1, 2, \ldots, S$ **do**
  for iteration $k = 1, 2, \ldots, K$ **do**
    Randomly pick a sample $i_k \in [n]$.
    Set $g_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)$. (variance reduction)
    Update $x_{k+1} = x_k - \eta g_k$.
  end for
  Update the snapshot $v_{s+1} = \frac{1}{K}\sum_{k=1}^{K} x_k$.
  Set $x_1 = v_{s+1}$
end for

Some things to help decipher the algorithm:

- There are 2 iterative loops, inner one is the same K iteration loop, the outer one is looping over S.
- The full gradient $\nabla F(v_s)$ is used to compute the stochastic gradient g_k, however, since it is only computed with respect to v_s, a snapshot at each outer loop iteration, we do not need to compute the full gradient K times, and we can just compute $\nabla F(v_s)$ at the beginning of each stage.
- We update the snapshot of v_{s+1} at the end of each stage s, which is the average of all x_k computed in the K iterations of the inner loop.

Recall. A stochastic gradient g_k for a function F at x_k in its expectation is the full gradient, $\mathbb{E}_z[g_k] = \nabla F(x_k)$.
We can see that if we set $g_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)$, then g_k is still a stochastic gradient. This is evident where

$$\begin{aligned}
\mathbb{E}\left[\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\right] &= \mathbb{E}\left[\nabla f_{i_k}(x_k)\right] - \mathbb{E}\left[\nabla f_{i_k}(v_s)\right] + \mathbb{E}\left[\nabla F(v_s)\right] \\
&= \mathbb{E}\left[\nabla f_{i_k}(x_k)\right] - \nabla F(v_s) + \nabla F(v_s) \\
&= \mathbb{E}\left[\nabla f_{i_k}(x_k)\right] \\
&= \nabla F(x_k)
\end{aligned}$$

Theorem 2. *Suppose each $f_i(\cdot)$ is L-smooth and μ-strongly convex. Setting SVRG with $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$. Then, at each stage s,*

$$F(v_{s+1}) - F(x_*) \leq \frac{3}{4}\left(F(v_s) - F(x_*)\right)$$

where $x_* \in \arg\min_x F(x)$.

The proof of this theorem will be in Section 2.2.
In order to get an ϵ-gap, the size of S (number of stages) should be in the order of $O(\log \frac{1}{\epsilon})$. Since we can repeatedly apply Theorem 1:

$$F(v_{s+1}) - F(x_*) \leq \left(\frac{3}{4}\right)^S \big(F(v_1) - F(x_*)\big) \leq \epsilon$$

$$S = \frac{4}{3} \log \frac{\big(F(v_1) - F(x_*)\big)}{\epsilon} \approx O(\log \frac{1}{\epsilon})$$

2.1 Computation Cost of SVRG

If we set $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$, we can see that within SVRG,

- Total number of stochastic gradient computations:

$$2 \times K \times S = 2 \times 64\frac{L}{\mu} \times S = O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right).$$

- Total number of full gradient computations:

$$1 \times S = O\left(\log \frac{1}{\epsilon}\right).$$

- Total number of stochastic gradient computations:

$$O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right) + n \times O\left(\log \frac{1}{\epsilon}\right) = O\left(\left(\frac{L}{\mu} + n\right) \log \frac{1}{\epsilon}\right).$$

2.1.1 SVRG vs. GD

Recall that for gradient descent, in order to get an ϵ-gap for a function that is L-smooth and μ-strongly convex, the number of iterations we need is

$$K = O\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$$

If we compare the runtime between SVRG and GD, we see that

$$\frac{\text{runtime of SVRG}}{\text{runtime of GD}} = \frac{(\frac{L}{\mu} + n) \log \frac{1}{\epsilon}}{\frac{L}{\mu} \log \frac{1}{\epsilon} \times n}$$

We want the above fraction to be much less than 1, thus the condition in which SVRG is faster than GD is when

$$\left(\frac{L}{\mu} + n\right) \ll n\frac{L}{\mu} \;\Leftrightarrow\; n \ll (n-1)\frac{L}{\mu},$$

in other words, when the condition number of F is greater than 1, SVRG will be faster than GD. This is almost always true in practice.

2.1.2 SVRG vs. SGD

Recall that for SGD, in order to get an ϵ-gap for function that is L-smooth and μ-strongly convex, the asymptotic lower bound of the number of iterations we need is

$$K = \Omega\left(\frac{1}{\epsilon}\right)$$

If we compare the runtime between SVRG and SGD, we see that

$$\frac{\text{runtime of SVRG}}{\text{runtime of SGD}} = \frac{\left(\frac{L}{\mu} + n\right)\log\frac{1}{\epsilon}}{\frac{1}{\epsilon} \times 1}$$

Similarly, the condition for SVRG to be faster than SGD is when

$$\left(\frac{L}{\mu} + n\right)\log\frac{1}{\epsilon} \ll \frac{1}{\epsilon} \Leftrightarrow \left(\frac{L}{\mu} + n\right) \ll \frac{\frac{1}{\epsilon}}{\log\frac{1}{\epsilon}}.$$

One way to interpret the above condition is that the condition number($\frac{L}{\mu}$) and the number of samples(n) is bounded by the inverse of ϵ, which implies that we want to use SVRG when the target ϵ is smaller comparatively to $\kappa_F + n$, where $\kappa_F = \frac{L}{\mu}$.

2.2 Proof of Theorem 2

We will proceed by introducing a few lemmas required for the proof, ultimately see why setting g_k the way in the SVRG algorithm, it will reduce as we move close to optimal.

Lemma 2. *(**Variance Expression**) For any random variable $Y \in \mathbb{R}^d$*

$$Var(Y) = \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right] = \mathbb{E}[\|Y\|_2^2] - \left(\mathbb{E}[\|Y\|]\right)^2 \leq \mathbb{E}[\|Y\|_2^2].$$

Lemma 3. *(**Expected Optimality Gap**) If each $f_i(\cdot)$ is L-smooth convex, then*

$$\mathbb{E}\left[\|\nabla f_{i_k}(x) - \nabla f_{i_k}(x_*)\|^2\right] \leq 2L\left(F(x) - F(x_*)\right)$$

Proof. The proof is left as an exercise in HW3. □

Lemma 4. *(**Variance bound**)*

$$\mathbb{E}\left[\|g_k\|_2^2\right] \leq 4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right).$$

Proof. Recall that $g_k := \nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)$, thus

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&= \mathbb{E}\left[\|\nabla f_{i_k}(x_k) + \left[\nabla f_{i_k}(x_*) - \nabla f_{i_k}(x_*)\right] - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right], \text{ terms cancels} \\
&= \mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) + \nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*)\|_2^2\right] \\
&\quad + 2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \text{ triangle inequality}
\end{aligned}$$

We will look at the term in red first.

$$\begin{aligned}
&2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&= 2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) - \left(\nabla F(x_*) - \nabla F(v_s)\right)\|_2^2\right], \quad \text{since } \nabla F(x_*) = \mathbf{0}
\end{aligned}$$

If we denote

$$Y := \nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s),\ \mathbb{E}[Y] := \nabla F(x_*) - \nabla F(v_s)$$

then, using lemma 1, we can see that

$$\begin{aligned}
&2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s) - \left(\nabla F(x_*) - \nabla F(v_s)\right)\|_2^2\right] \\
&= 2\mathbb{E}\left[Y - \mathbb{E}[Y]\|_2^2\right] \\
&\leq \mathbb{E}[\|Y\|_2^2] \\
&= 2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s)\|_2^2\right].
\end{aligned}$$

Using lemma 3, both terms can be bounded quite easily.

$$\begin{aligned}
&\mathbb{E}[\|g_k\|^2] := \mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(v_s) + \nabla F(v_s)\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*)\|_2^2\right] \\
&\quad + 2\mathbb{E}\left[\|\nabla f_{i_k}(x_*) - \nabla f_{i_k}(v_s)\|_2^2\right] \\
&\leq 4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(v_s) - F(x_*)\right).
\end{aligned}$$

□

Now that we have a bound on the g_k in terms of the optimality gap, we can plug it back into the iteration complexity formula for the generic SGD algorithm. Recall the iteration complexity for SGD is

$$\frac{1}{K}\sum_{k=1}^{K}\left(F(x_k) - F(x_*)\right) \leq \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|_2^2}{2\eta K}. \tag{2}$$

Since F is strongly convex, we know that

$$F(x_1) \geq F(x_*) + \langle \nabla F(x_*), x_1 - x_* \rangle + \frac{\mu}{2}\|x_1 - x_*\|_2^2,$$

and this implies

$$\frac{2}{\mu}(F(x_1) - F(x_*)) \geq \|x_1 - x_*\|_2^2.$$

We can then plug the bound on the square distance between x_1 and x_* back into equation (2), which yields

$$\frac{1}{K}\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\big) \leq \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{F(x_1) - F(x_*)}{\eta\mu K}.$$

By lemma 4, we can plug the variance bound into the first term of the RHS and get

$$\frac{\eta}{2K}\sum_{k=1}^{K}\mathbb{E}[\|g_k\|^2] = \frac{\eta}{2K}\sum_{k=1}^{K}\Big(4L\left(F(x_k) - F(x_*)\right) + 4L\left(F(x_1) - F(x_*)\right)\Big).$$

Then, combining simplified terms and with some algebra we can arrive at

$$\frac{1}{K}\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\big) \leq \frac{2\eta L + \frac{1}{\eta\mu K}}{1 - 2\eta L}\big(F(x_1) - F(x_*)\big) \tag{3}$$

By construction, we have set $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$, then plugging them in results in

$$F(\bar{x}_K) - F(x_*) \leq \frac{1}{K}\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\big) \leq \frac{3}{4}\left(F(x_1) - F(x_*)\right),$$

thus completes the proof for Theorem 2, where the first inequality is be Jensen's inequality.

Remark. Since $\bar{x}_k$ is used to initialize x_1 and the snapshot v_{s+1} in the next stage, the above means that the optimality gap $\delta_s := F(v_s) - F(x_*)$ is shrinking within a constant factor in each stage.

Bibliographic notes

SVRG was proposed by Johnson and Zhang [1].

References

[1] Rie Johnson and Tong Zhang Accelerating Stochastic Gradient Descent using Predictive Variance Reduction NeurIPS 2013