Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Gurpreet Singh
Email: gurpreet.s@columbia.edu

SCRIBE
15

Convergence Analysis of SVRG

1. Recap

In the last lecture we discussed Stochastic Variance Reduced Gradients (SVRG) approach to finite sum minimization. The algorithm for SVRG is given in algorithm box 1.

Algorithm 1: Stochastic Variance Reduced Gradients

1. Start with arbitrary $\mathbf{x}_0^{(0)}$
2. For $k = 0, 1 \ldots K - 1$
   (a) Set $\mathbf{x}_0 = \mathbf{x}_0^{(k)}$, compute $\nabla f(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} \nabla g_i(\mathbf{x}_0)$
   (b) For $t = 0, 1 \ldots T - 1$:
      i. Sample i_t uniformly at random from $\{1, 2, \ldots, n\}$.
      ii. Set
      $$\mathbf{x}_{t+1} \;=\; \mathbf{x}_t - \eta\left(\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right)$$
   (c) Set $\mathbf{x}_0^{(k+1)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$
3. Output $x_0^{(K)}$.

To motivate SVRG, note that the gradient estimator used in the inner loop is unbiased:

Lemma 15.0.1 We have

$$\mathop{\mathbb{E}}_{i_t \sim [n]}\left[\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \,\middle|\, \mathbf{x}_t, \mathbf{x}_{t-1} \ldots, \mathbf{x}_0\right] \;=\; \nabla f(\mathbf{x}_t)$$

Proof. We can write the expectation as

$$\begin{aligned}\mathop{\mathbb{E}}_{i_t\sim[n]}\left[\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\,\middle|\,\mathbf{x}_t, \mathbf{x}_{t-1}\ldots, \mathbf{x}_0\right] &= \frac{1}{n}\sum_{i=1}^{n}\big(\nabla g_i(\mathbf{x}_t) - \nabla g_i(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\nabla g_i(\mathbf{x}_t) - \frac{1}{n}\sum_{i=1}^{n}\nabla g_i(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\\ &= \frac{1}{n}\sum_{i=1}^{n}\nabla g_i(\mathbf{x}_t)\\ &= \nabla f(\mathbf{x}_t)\end{aligned}$$

□

In the next section, we will look at the convergence of SVRG.

2. Convergence of SVRG

Fix an epoch k (i.e. condition on $\mathbf{x}_0^{(k)}$).

Following the same route as in the case of vanilla GD, we can write

$$2\eta\nabla f(\mathbf{x}_t)^{\mathrm{T}}\left(\mathbf{x}_t - \mathbf{x}^*\right) \le \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|^2 \mid \mathbf{x}_t\right] - \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2 + \eta^2\mathbb{E}\left[\left\|\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right\|^2\right]$$

Also, using the strong convexity of f, we can write

$$2\eta\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right) \le (1-\eta\alpha)\,\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|^2 \mid \mathbf{x}_t\right] - \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2 + \eta^2\mathbb{E}\left[\left\|\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right\|^2\right]$$

Taking expectation on both sides, we have

$$2\eta\mathbb{E}\left[\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\right] \le (1-\eta\alpha)\,\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right\|^2\right]$$

Summing over $t = 0\ldots T-1$, we get

$$\sum_{t=0}^{T} 2\eta\mathbb{E}\left[\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\right] \le (1-\eta\alpha)\,\mathbb{E}\left[\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_T - \mathbf{x}^*\right\|^2\right] - \eta\alpha\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|^2\right] + \sum_{t=0}^{T-1}\eta^2\mathbb{E}\left[\left\|\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right\|^2\right]$$

Dropping the negative terms, we get

$$\sum_{t=0}^{T} 2\eta\mathbb{E}\left[\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right)\right] \le (1-\eta\alpha)\,\mathbb{E}\left[\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|^2\right] + \sum_{t=0}^{T-1}\eta^2\mathbb{E}\left[\left\|\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\right\|^2\right]$$

Since the gradient estimator is unbiased, we can write the above as

$$2\eta\mathbb{E}\left[\sum_{t=0}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \;\leq\; (1-\eta\alpha)\,\mathbb{E}\left[\,\|\,\mathbf{x}_0 - \mathbf{x}^*\,\|^2\right] + \eta^2\sum_{t=0}^{T-1}\mathbb{E}\left[\,\|\,\nabla f(\mathbf{x}_t)\,\|^2\right] + \sum_{t=0}^{T-1}\eta^2\mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\,\|^2\right]$$

Again following the vanilla analysis, we can say that if $\eta \leq \frac{1}{\beta}$, then we have

$$\frac{\eta}{2}\mathbb{E}\left[\,\|\,\nabla f(\mathbf{x}_t)\,\|^2\right] \;\leq\; \mathbb{E}\left[\,f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\,\right]$$

Putting this in the original equation, we get

$$2\eta\mathbb{E}\left[\sum_{t=0}^{T} f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \;\leq\; (1-\eta\alpha)\,\mathbb{E}\left[\,\|\,\mathbf{x}_0 - \mathbf{x}^*\,\|^2\right] + \sum_{t=0}^{T-1}\eta^2\mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\,\|^2\right]$$

Now notice that for a random vector X, we always have $\mathbb{E}\left[\,\|\,X - \mathbb{E}\,[\,X\,]\,\|^2\right] \;\leq\; \mathbb{E}\left[\,\|\,X\,\|^2\right]$. Therefore, we have

$$\mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{t}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\,\|^2 \;|\, \mathbf{x}_t, \mathbf{x}_0\right] \;\leq\; \mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0)\,\|^2 \;|\, \mathbf{x}_0, \mathbf{x}_t\right]$$
$$\implies \mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{t}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\,\|^2\right] \;\leq\; \mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}_0)\,\|^2\right]$$

Also, we can write $\|\,\mathbf{x} - \mathbf{y}\,\|^2 \;\leq\; 2(\|\,\mathbf{x}\,\|^2 + \|\,\mathbf{y}\,\|)^2$. Therefore, we have

$$\mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{t}_t) - \nabla g_{i_t}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) - \nabla f(\mathbf{x}_t)\,\|^2\right] \;\leq\; \mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_t) - \nabla g_{i_t}(\mathbf{x}^*)\,\|^2\right] + \mathbb{E}\left[\,\|\,\nabla g_{i_t}(\mathbf{x}_0) - \nabla g_{i_t}(\mathbf{x}^*)\,\|^2\right]$$

Suppose we define $\tilde{g}_i(\mathbf{x}) = g_i(\mathbf{x}) - \nabla g(\mathbf{x}^*)^{\mathrm{T}}\,(\mathbf{x} - \mathbf{x}^*)$. Then, we have

$$\nabla\tilde{g}_i(\mathbf{x}) \;=\; \nabla g_i(\mathbf{x}) - \nabla g_i(\mathbf{x}^*)$$
$$\nabla^2\tilde{g}_i(\mathbf{x}) \;=\; \nabla^2 g_i(\mathbf{x})$$

The above implies that $\tilde{g}_i$ is also β-smooth. Therefore, for any $\mathbf{y} = \mathbf{x} - \frac{1}{\beta}\nabla g_i(\mathbf{x})$, we can write

$$\tilde{g}_i(\mathbf{y}) \;\leq\; \tilde{g}_i(\mathbf{x}) - \frac{1}{2\beta}\,\|\,\nabla\tilde{g}_i(\mathbf{x})\,\|^2$$

Also, one can claim that $\mathbf{x}^*$ is a minimizer of $\tilde{g}_i$. This is because the gradient vanishes at this point and $\tilde{g}_i$ is convex, so $\mathbf{x}^*$ is a global minimizer. Hence,

$$\tilde{g}_i(\mathbf{x}^*) \;\leq\; \tilde{g}_i(\mathbf{x}) - \frac{1}{2\beta}\,\|\,\nabla\tilde{g}_i(\mathbf{x})\,\|^2,$$

or, equivalently using the definition of $\tilde{g}_i$, we have

$$\|\nabla g_i(\mathbf{x}) - \nabla g_i(\mathbf{x}^*)\|^2 \leq 2\beta(g_i(\mathbf{x}) - \nabla g_i(\mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*) - g_i(\mathbf{x}^*)).$$

Using this, we can write

$$\mathop{\mathbb{E}}_{i\sim[\,n\,]}\left[\,\|\,\nabla g_i(\mathbf{x}) - \nabla g_i(\mathbf{x}^*)\,\|^2\right] \;\leq\; 2\beta\mathop{\mathbb{E}}_{i\sim[\,n\,]}\left[\,g_i(\mathbf{x}) - \nabla g_i(\mathbf{x}^*)^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^*\right) - g_i(\mathbf{x}^*)\,\right]$$
$$=\; 2\beta\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right)$$

The above equality uses the fact that $\mathop{\mathbb{E}}_{i\sim[\,n\,]}\left[\,\nabla g_i(\mathbf{x}^*)\,\right] = \nabla f(\mathbf{x}*) = 0$.

We can use this bound to simplify the original bound we arrived at.

$$
\begin{aligned}
2\eta\mathbb{E}\left[\sum_{t=0}^{T} f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \;&\leq\; (1-\eta\alpha)\,\mathbb{E}\left[\,\left\|\,\mathbf{x}_0 - \mathbf{x}^*\,\right\|^2\,\right] \\
&\quad + \sum_{t=0}^{T-1} 4\beta\eta^2 \left\{\mathbb{E}\left[\,f(\mathbf{x}_t) - f(\mathbf{x}^*)\,\right] + \mathbb{E}\left[\,f(\mathbf{x}_0) - f(\mathbf{x}^*)\,\right]\right\}
\end{aligned}
$$

We will continue the analysis in the next lecture.