CSC 576: Stochastic Gradient "Descent" Algorithm

Ji Liu
Department of Computer Sciences, University of Rochester

April 29, 2015

1 Introduction

Consider the following least squares problem

$$\min_{x\in\mathbb{R}^N} \quad f(x) := \frac{1}{2M}\sum_{m=1}^{M}(a_m^T x - b_m)^2 \equiv \frac{1}{2M}\|Ax-b\|^2, \tag{1}$$

where M could be infinite. To solve this problem, we can apparently use the gradient descent algorithm we learned before by updating

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

where

$$\nabla f(x_k) = \frac{1}{M}\sum_{m=1}^{M} a_m(a_m^T x_k - b_m).$$

It is easy to verify that the computation complexity per iteration is $O(MN)$. However, in some scenarios, the gradient descent algorithm or other gradient based methods is inefficient or unavalabile for two reasons:

- When the number of data samples is large, that is, M is large, the complexity is substantial per iteration;
- In the online learning scenario, the data (a_m, b_m) is coming one by one, which means that the gradient is not computable.

To avoid to evaluate the full gradient per iteration, the stochastic gradient "descent" method only uses a small portion of data to compute an approximate gradient, which is called "stochastic gradient", for updating x per iteration:

$$x_{k+1} = x_k - \gamma_k \underbrace{a_m(a_m^T x_k - b_m)}_{\text{stochastic gradient}}$$

where m is randomly selected from $\{1, 2\cdots, M\}$ per iteration. One can easily verify that

$$\mathbb{E}_m(a_m(a_m^T x_k - b_m)) = \nabla f(x_k)$$

Consider a general optimization problem

$$\min_x \quad f(x). \tag{2}$$

The general stochastic gradient "descent" (SGD) algorithm is updating x by

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

where g_k is a vector (called stochastic gradient) satisfying $\mathbb{E}(g_k) = \nabla f(x_k)$. Note that SGD is not a real "descent" algorithm, because it does not guarantee to decrease the objective function value in every iteration. For example, consider

$$f(x) = \frac{1}{2}(f_1(x) + f_2(x))$$

where $f_1(x) = \frac{1}{2}(x-2)^2$ and $f_2(x) = \frac{1}{2}(x+1)^2$. The optimal solution is $x^* = 0.5$. Assume that we start from $x_0 = 0$. It is easy to see that the decreasing direction is positive. Applying the SGD algorithm, if we sample the subfunction $f_2(x)$, we obtain

$$x_1 = x_0 - \gamma(x_0 + 1) = -\gamma.$$

Since the steplenght γ is positive, no matter how to choose γ, we have $f(x_1) > f(x_0)$. Therefore, SGD is not a real descent algorithm.

2 Convergence Rate

SGD essentially uses the inaccurate gradient per iteration. Since there is no free food, what is the cost by using approximate gradient? The answer is that the convergence rate is slower than the gradient descent algorithm.

Theorem 1. *Assume that* $\mathbb{E}(\|g_k\|^2) \leq G^2$ *and* $f(x)$ *is strongly convex, that is, there exists a positive number* $l > 0$ *satisfying*

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle + \frac{l}{2}\|x - y\|^2.$$

Choose the steplength $\gamma_k = \frac{1}{lk}$. *We have*

$$\mathbb{E}(\|x_k - x^*\|^2) \leq \frac{\max\{\|x_1 - x^*\|^2, G^2/l^2\}}{k}$$

where x^* *is the optimal solution to* (2).

Proof. From the strong convexity, we have

$$f(x^*) - f(x_k) \geq \langle \nabla f(x_k), x^* - x_k \rangle + \frac{l}{2}\|x_k - x^*\|^2$$
$$f(x_k) - f(x^*) \geq \langle \nabla f(x^*), x_k - x^* \rangle + \frac{l}{2}\|x_k - x^*\|^2.$$

Summarizing both inequalities gives

$$\langle \nabla f(x_k) - \nabla f(x^*),\ x_k - x^* \rangle = \langle \nabla f(x_k),\ x_k - x^* \rangle \geq l\|x_k - x^*\|^2. \tag{3}$$

We have

$$\begin{aligned}
\mathbb{E}(\|x_{k+1} - x^*\|^2) =& \mathbb{E}(\|x_k - \gamma_k g_k - x^*\|^2) \\
=& \mathbb{E}(\|x_k - x^*\|^2) - 2\gamma_k \mathbb{E}\langle g_k, x_k - x^* \rangle + \gamma_k^2 \mathbb{E}(\|g_k\|^2) \\
\leq& \mathbb{E}(\|x_k - x^*\|^2) - 2\gamma_k \mathbb{E}\langle \nabla f(x_k), x_k - x^* \rangle + \gamma_k^2 G^2
\end{aligned} \tag{4}$$

Applying (3), it follows

$$\begin{aligned}
\mathbb{E}(\|x_{k+1} - x^*\|^2) \leq& \mathbb{E}(\|x_k - x^*\|^2) - 2l\gamma_k \mathbb{E}(\|x_k - x^*\|^2) + \gamma_k^2 G^2 \\
=& (1 - 2l\gamma_k)\mathbb{E}(\|x_k - x^*\|^2) + \gamma_k^2 G^2.
\end{aligned} \tag{5}$$

We prove the convergence rate by induction. First it is easy to see that

$$\|x_1 - x^*\|^2 \leq \frac{\max\{\|x_1 - x^*\|^2, G^2/l^2\}}{1}.$$

Then we assume that the convergence rate holds with k. Next we only need to show that it holds with $k+1$. Denote $L = \max \|x_1 - x^*\|^2, G^2/l^2$. From (5), we have

$$\begin{aligned}
\mathbb{E}(\|x_{k+1} - x^*\|^2) &\leq \left(1 - \frac{2}{k}\right) \mathbb{E}(\|x_k - x^*\|^2) + \frac{1}{l^2 k^2} G^2 \\
&\leq \left(1 - \frac{2}{k}\right) \frac{L}{k} + \frac{1}{l^2 k^2} G^2 \\
&\leq \left(\frac{1}{k} - \frac{2}{k^2}\right) L + \frac{L}{k^2} \\
&\leq \left(\frac{1}{k} - \frac{1}{k^2}\right) L \\
&\leq \frac{L}{k+1}.
\end{aligned}$$

It completes the proof. □

Sometimes, people use $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ as the output, which is more robust than using x_k. A similar convergence rate $O(1/k)$ can be guaranteed.

We discussed the convergence rate for strongly convex case above. Next we consider the general convex function $f(x)$. The convergence rate is $O(1/\sqrt{k})$, slightly worse than the strongly convex case.

Theorem 2. *Assume that $\mathbb{E}(\|g_k\|^2) \leq G^2$, $\mathbb{E}\|x_k - x^*\|^2 \leq D^2$, and $f(x)$ is convex. Choose the steplength $\gamma_k = \frac{c}{\sqrt{k}}$ where c is predefined positive number. We have*

$$f(\bar{x}_k) - f^* \leq \frac{c^{-1}D^2 + c\sqrt{\frac{k+1}{k}}G^2}{\sqrt{k}},$$

where x^* is the optimal solution to (2), f^* is short for $f(x^*)$, and $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$.

Proof. From the convexity of $f(x)$, we have

$$\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*.$$

Considering (5), we have

$$\begin{aligned}
&\mathbb{E}(\|x_{k+1} - x^*\|^2) \leq \mathbb{E}(\|x_k - x^*\|^2) - 2\gamma_k(f(x_k) - f^*) + \gamma_k^2 G \\
\Rightarrow &2(f(x_k) - f^*) \leq \gamma_k^{-1}\mathbb{E}(\|x_k - x^*\|^2) - \gamma_k^{-1}\mathbb{E}(\|x_{k+1} - x^*\|^2) + \gamma_k G^2.
\end{aligned}$$

Summarizing the inequality above from $i = 1, 2, \cdots, k$, we obtain

$$\begin{aligned}
2\sum_{i=1}^{k}(f(x_i) - f^*) \leq &\gamma_1^{-1}\mathbb{E}(\|x_1 - x^*\|^2) + \sum_{i=2}^{k}(\gamma_i^{-1} - \gamma_{i-1}^{-1})\mathbb{E}(\|x_i - x^*\|^2) + \sum_{i=1}^{k}\gamma_i G^2 \\
\leq &\gamma_1^{-1} D^2 + \sum_{i=2}^{k}(\gamma_i^{-1} - \gamma_{i-1}^{-1})D^2 + \sum_{i=1}^{k}\gamma_i G^2 \\
\leq &\frac{\sqrt{k}}{c} D^2 + c\sqrt{k+1} G^2,
\end{aligned}$$

where the last inequality uses $\sum_{i=1}^{k}\gamma_i \leq c\sqrt{k+1}$. From Jensen's inequality we have

$$\frac{1}{k}\sum_{i=1}^{k}(f(x_k) - f^*) \geq f(\bar{x}_k) - f^*,$$

which indicates that

$$f(\bar{x}_k) - f^* \leq \frac{c^{-1}\sqrt{k}D^2 + c\sqrt{k+1}G^2}{k}.$$

It completes the proof. □