Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Victor Ye Dong
Email: yd2470@columbia.edu

SCRIBE

11

Projected Gradient Descent

1 Problem Setup

The optimization problem for stochastic gradient descent is as follows

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{\xi\sim\mathcal{D}}[g(x,\xi)] \\ s.t. \quad & x \in K, \qquad K \text{ is convex} \end{aligned} \tag{1.1}$$

1.1 Algorithm

Algorithm 1: SGD

Init: Start with arbitrary $x_0 \in K$
for $t = 0, 1, 2...$ **do**
  Draw $\xi_i \sim \mathcal{D}$
  Update $x_{t+1} = \Pi_K(x_t - \eta\nabla g(x, \xi_i))$
end
return some combination of $x_0, ..., x_T$

1.2 Assumption

Variance of sgd is bounded

$$\mathbb{E}[||\nabla g(x,\xi) - \nabla f(x)||_2^2] \leq \sigma^2 \tag{1.2}$$

which is equivalent as

$$\mathbb{E}[||\nabla g(x,\xi)||^2] - ||\nabla f(x)||_2^2 \leq \sigma^2 \tag{1.3}$$

2 Analysis for L-Lipschitz f

In the previous lecture, we showed that setting the step size $\eta = \frac{D}{\sqrt{\sigma^2+L^2}\sqrt{T}}$, we obtain

$$\mathbb{E}[f(\bar{x})] - f(x^*) \leq \frac{D\sqrt{\sigma^2+L^2}}{\sqrt{T}} \tag{2.1}$$

Sanity Check : If $g(x, \xi) = f(x)$, then $\sigma = 0$. We then recover deterministic GD and its convergence rate.

3 Analysis for β-smooth f

We will only analyze the case when $K = \mathbb{R}^d$, so that no projections are necessary. Projections add a slight extra complication which is handled exactly as in the deterministic case.

Just as in the previous analysis for L-Lipschitz f, we have

$$\begin{aligned}\mathbb{E}[||x_{t+1} - x^*||^2|x_t] &= ||x_t - x^*||^2 + \eta^2 \cdot \mathbb{E}[||\nabla g(x_t)||^2] - \mathbb{E}[2\eta \nabla g(x_t)^T (x_t - x^*)] \\ &\leq ||x_t - x^*||^2 + \eta^2(||\nabla f(x_t)||^2 + \sigma^2) - 2\eta \nabla f(x_t)^T (x_t - x^*)\end{aligned} \tag{3.1}$$

By smoothness, we have

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2.$$

Since $x_{t+1} = x_t - \eta_t \nabla g(x_t, \xi_t)$ (since we don't need projections), we have

$$f(x_{t+1}) \leq f(x_t) - \eta \nabla f(x_t)^\top \nabla g(x_t, \xi_t) + \frac{\beta\eta^2}{2}\|\nabla g(x_t, \xi_t)\|^2.$$

Taking expectations on both sides conditioned on x_t, we have

$$\begin{aligned}\mathbb{E}[f(x_{t+1})|x_t] &\leq f(x) - \eta||\nabla f(x_t)||^2 + \frac{\beta\eta^2}{2}(||\nabla f(x_t)||^2 + \sigma^2) \\ &\leq f(x_t) - \frac{\eta}{2}||\nabla f(x_t)||^2 + \frac{\eta\sigma^2}{2}\end{aligned} \tag{3.2}$$

if we choose $\eta \leq \frac{1}{\beta}$.

Combining (3.1) and (3.2) and convexity property, we have that

$$\begin{aligned}\mathbb{E}[f(x_t)|x_t] - f(x^*) &\leq \nabla f(x_t)^T (x_t - x^*) \\ &\leq \frac{1}{2\eta}(||x_t - x^*||^2 + \eta^2(||\nabla f(x_t)||^2 + \sigma^2) - \mathbb{E}[||x_{t+1} - x^*||^2|x_t]) \\ &\leq \frac{1}{2\eta}(||x_t - x^*||^2 - \mathbb{E}[||x_{t+1} - x^*||^2|x_t])) + f(x_t) - \mathbb{E}[f(x_{t+1})|x_t] + \eta\sigma^2\end{aligned} \tag{3.3}$$

Reorganizing (3.3) above, we have

$$\mathbb{E}[f(x_{t+1})|x_t] - f(x^*) \leq \frac{1}{2\eta}((||x_t - x^*||^2 - \mathbb{E}[||x_{t+1} - x^*||^2|x_t]) + \eta\sigma^2 \tag{3.4}$$

Now taking expectation w.r.t. x_t to remove the conditioning, we get

$$\mathbb{E}[f(x_{t+1})] - f(x^*) \leq \frac{1}{2\eta}(\mathbb{E}[||x_t - x^*||^2] - \mathbb{E}[||x_{t+1} - x^*||^2]) + \eta\sigma^2$$

Sum up the term on both sides, we have

$$\begin{aligned}\frac{1}{T}\sum_{0}^{T-1} \mathbb{E}[f(x_{t+1})] - f(x^*) &\leq \frac{1}{2\eta T}(||x_0 - x^*||^2 - \mathbb{E}[||x_T - x^*||^2]) + \eta\sigma^2 \\ &\leq \frac{1}{2\eta T}||x_0 - x^*||^2 + \eta\sigma^2\end{aligned} \tag{3.5}$$

Since we need $\eta \leq \frac{1}{\beta}$, let us set $\eta = \frac{1}{\beta + c\sqrt{T}}$, where $c > 0$ to be determined shortly.

Let $||x_0 - x^*|| = D$. Then we have

$$\begin{aligned}\frac{1}{T}\sum_{0}^{T-1}\mathbb{E}[f(x_{t+1})] - f(x^*) &\leq \frac{1}{2\eta T}||x_0 - x^*||^2 + \eta\sigma^2 \\ &\leq \frac{(\beta + c\sqrt{T})D^2}{2T} + \frac{\sigma^2}{c\sqrt{T}} \\ &= \frac{\beta D^2}{2T} + \frac{D^2 c}{2\sqrt{T}} + \frac{\sigma^2}{c\sqrt{T}}\end{aligned} \tag{3.6}$$

Therefore, if we set $c = \frac{\sqrt{2}\sigma}{D}$, we can achieve the minimum value for the RHS, which leads to

$$\begin{aligned}\mathbb{E}[f(\bar{x})] - f(x^*) &\leq \frac{1}{T}\sum_{0}^{T-1}\mathbb{E}[f(x_{t+1})] - f(x^*) \\ &\leq \frac{\beta D^2}{2T} + \frac{D\sqrt{2}\sigma}{2\sqrt{T}}\end{aligned} \tag{3.7}$$

4 Analysis for α-strongly convex and β-smooth f

Again we will only look at the unconstrained case, i.e. $K = \mathbb{R}^d$, so that projections are not needed. Similarly as above, and using α-strong convexity, we have

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}(||x_t - x^*||^2 - \mathbb{E}[||x_{t+1} - x^*||^2|x_t]) + \frac{\eta}{2}(||\nabla f||^2 + \sigma^2) - \frac{\alpha}{2}||x_t - x^*||^2 \tag{4.1}$$

$\Rightarrow$

$$\mathbb{E}[f(x_{t+1})|x_t] - f(x^*) \leq \frac{1}{2\eta}(1 - \alpha\eta)||x_t - x^*||^2 - \frac{1}{2\eta}\mathbb{E}[||x_{t+1} - x^*||^2|x_t] + \frac{\eta}{2}\sigma^2 \tag{4.2}$$

Rearranging, and taking expectation w.r.t. x_t, we have $\Rightarrow$

$$2\eta[\mathbb{E}[f(x_{t+1})] - f(x^*)] + \mathbb{E}[||x_{t+1} - x^*||^2] \leq (1 - \alpha\eta)\mathbb{E}[||x_t - x^*||^2] + \eta^2\sigma^2 \tag{4.3}$$

Since $\mathbb{E}[f(x_{t+1})] - f(x^*)$, we have

$$\mathbb{E}[||x_{t+1} - x^*||^2] \leq (1 - \alpha\eta)\mathbb{E}[||x_t - x^*||^2] + \eta^2\sigma^2.$$

Unrolling the above inequality recursively, we get

$$\begin{aligned}\mathbb{E}[||X_T - x^*||^2] &\leq (1 - \eta\alpha)^T||x_0 - x^*||^2 + 2\eta^2\sigma^2(1 + (1 - \eta\alpha) + ... + (1 - \eta\alpha)^{T-1}) \\ &\leq (1 - \eta\alpha)^T||x_0 - x^*||^2 + \frac{\eta\sigma^2}{\alpha}\end{aligned} \tag{4.4}$$

(To be continued)