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

7

GRADIENT DESCENT

1. Gradient Descent

We analyze gradient descent assuming different conditions on f.

1.1 f is Lipschitz with constant L

Definition 1. f is Lipschitz with constant L if

$$|f(x) - f(y)| \leq L||x - y|| \tag{1.1.1}$$

which is equivalent to

$$||\nabla f(x)|| \leq L \tag{1.1.2}$$

Recall the analysis of gradient descent: we have

$$\begin{aligned} \sum_{t=0}^{T-1} f(x_t) - f(x^*) &\leq \frac{1}{2\eta}(||x_0 - x^*||^2 - ||x_T - x^*||^2) + \frac{\eta}{2}\sum_{t=0}^{T-1}||\nabla f(x_t)||^2 \\ &\leq \frac{1}{2\eta}(||x_0 - x^*||^2) + \frac{\eta}{2}TL^2 \\ &\leq \sqrt{T}L||x_0 - x^*|| \end{aligned} \tag{1.1.3}$$

where the last equality holds by setting $\eta = \frac{||x_0 - x^*||}{\sqrt{T}L}$.

By convexity of f, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \geq f(\frac{1}{T}\sum_{t=0}^{T-1} x_t) \tag{1.1.4}$$

which implies that

$$f(\bar{x}) - f(x^*) \leq \frac{||x_0 - x^*||L}{\sqrt{T}} \tag{1.1.5}$$

Then if we let $T \geq \frac{||x_0 - x^*||^2 L^2}{\epsilon^2}$, we have that

$$\begin{aligned} f(\bar{x}) - f(x^*) &\leq \frac{||x_0 - x^*||L}{\sqrt{T}} \\ &\leq \epsilon \end{aligned} \tag{1.1.6}$$

1.2 f is β-smooth

We now give a tighter analysis in the case f is β-smooth. From β-smoothness we have that

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}||y-x||^2 \tag{1.2.1}$$

If we let $x' = x - \frac{1}{\beta}\nabla f(x)$, then

$$f(x') \leq f(x) - \frac{1}{2\beta}||\nabla f(x)||^2 \tag{1.2.2}$$

Reorganizing we have

$$\frac{1}{2\beta}||\nabla f(x)||^2 \leq f(x) - f(x') \tag{1.2.3}$$

Now suppose we run gradient descent with $\eta = \frac{1}{\beta}$. Using inequality (1.2.3) with $x = x_t$ and $x' = x_{t+1}$, we get

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) + f(x^t) - f(x^{t+1}) \tag{1.2.4}$$

Then we have

$$\begin{aligned}\sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*) &\leq \frac{1}{2\eta}(||x_0 - x^*||^2 - ||x_T - x^*||^2) \\ &\leq \frac{1}{2\eta}||x_0 - x^*||^2\end{aligned} \tag{1.2.5}$$

Finally, replacing back $\eta = \frac{1}{\beta}$ and dividing on both sides by $T \Rightarrow$

$$\frac{1}{T}[\sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*)] \leq \frac{\beta}{2}(||x_0 - x^*||^2/T) \tag{1.2.6}$$

$\Rightarrow$

$$f(\frac{1}{T}\sum_{0}^{T-1} x_{t+1}) - f(x^*) \leq \frac{\beta}{2}(||x_0 - x^*||^2/T) \tag{1.2.7}$$

by Jensen's inequality since f is convex.

We can also prove a sub-optimality bound for the last iterate x_T. Note that inequality (1.2.2) implies that f monotonically decreases along the sequence $x_0, x_1, x_2, \ldots$, i.e.

$$f(x_0) \geq f(x_1) \geq f(x_2) \geq \cdots \geq f(x_T).$$

Thus,

$$f(x_T) - f(x^*) \leq \frac{1}{T}[\sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*)] \leq \frac{\beta}{2}\frac{||x_0 - x^*||^2}{T} \tag{1.2.8}$$

Then if we set

$$T \geq \frac{\beta}{2} \cdot \frac{||x_0 - x^*||^2}{\epsilon} \tag{1.2.9}$$

we can achieve ϵ-sub-optimality.

1.3 f is α-strongly convex and β-smooth

Now suppose f is α-strongly convex and β-smooth. Suppose we run gradient descent with step size $\eta = \frac{1}{\beta}$.

From the definitions, we have

$$f(x) + \nabla f(x)^T(y-x) + \frac{\alpha}{2}||x-y||^2 \leq f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}||y-x||^2 \quad (1.3.1)$$

if we set $y = x^*$ and $x = x_t$, we have

$$f(x_t) + \nabla f(x_t)^T(x^* - x_t) + \frac{\alpha}{2}||x_t - x^*||^2 \leq f(x^*) \quad (1.3.2)$$

Plugging in the bound from inequality (1.3.2) into the basic gradient descent analysis, we have

$$\begin{aligned} f(x_t) - f(x^*) &\leq \frac{1}{2\eta}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) + \frac{\eta}{2}||\nabla f(x_t)||^2 - \frac{\alpha}{2}||x_t - x^*||^2 \\ &\leq \frac{1}{2\eta}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) + (f(x_t) - f(x_{t+1})) - \frac{\alpha}{2}||x_t - x^*||^2 \end{aligned} \quad (1.3.3)$$

The second inequality above follows from the fact that $\eta = \frac{1}{\beta}$ and inequality (1.2.3) exactly as in the gradient descent analysis for β-smooth f.

Thus, we have

$$0 \leq f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) - \frac{\alpha}{2}||x_t - x^*||^2 \quad (1.3.4)$$

So,

$$||x_{t+1} - x^*||^2 \leq (1 - \eta\alpha)||x_t - x^*||^2 \quad (1.3.5)$$

Using $\eta = \frac{1}{\beta}$, we have that

$$||x_{t+1} - x^*||^2 \leq (1 - \frac{\alpha}{\beta})||x_t - x^*||^2 \quad (1.3.6)$$

$\Rightarrow$

$$||x_T - x^*||^2 \leq (1 - \frac{\alpha}{\beta})^T||x_0 - x^*||^2 \quad (1.3.7)$$

Since f is β smooth, we have

$$f(x_T) - f(x_*) \leq \nabla f(x^*)^T(x_T - x^*) + \frac{\beta}{2}||x_T - x^*||^2 \quad (1.3.8)$$

$$= \frac{\beta}{2}||x_T - x^*||^2 \quad (1.3.9)$$

$$\leq \frac{\beta}{2} \cdot (1 - \frac{\alpha}{\beta})^T||x_0 - x^*||^2 \quad (1.3.10)$$

The equality above uses the fact that $\nabla f(x^*) = 0$. Thus, after

$$\begin{aligned} T &= \frac{\log\left(\frac{2\epsilon}{D^2\beta}\right)}{-\log\left(1 - \frac{\alpha}{\beta}\right)} \\ &\approx \frac{\beta}{\alpha}\log\left(\frac{\beta D^2}{2\epsilon}\right), \quad \text{since } \log(1-x) \approx -x \end{aligned} \quad (1.3.11)$$

iterations we can achieve ϵ-sub-optimality.