COMS 4995-8: Advanced Algorithms (Spring'21) Mar 30, 2021

Lecture 21: Gradient descent: smoothness, convexity

Instructor: *Alex Andoni* Scribes: *Wenjie Wang, Lingxiao Li*

1 Overview

We will continue the context about **Gradient Descent**
The object function:

$$Let\ f : \mathbb{R}^n \to \mathbb{R}$$

$$\min_{x \in \mathbb{R}^n} f(x)$$

2 Smoothness, convexity

Theorem 1. *Taylor Expansion:*

$$f(x+\delta) = f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\delta^T \cdot \nabla^2 f(y) \cdot \delta$$

Notice y is some point nearby x.

Definition 2. *f is β-smooth: $\lambda_{max} \nabla^2 f(y) \leq \beta$*

Then we can use Taylor Expansion to get the folowing upper bound on $f(x+\delta)$:

$$f(x+\delta) \leq f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\beta ||\delta||^2$$

We want δ which minimizes the later quantity:

$$\begin{aligned}
\delta &\triangleq \text{argmin}_{\delta \in \mathbb{R}^n} f(x) + \nabla f(x)^T \cdot \delta + \frac{\beta}{2}||\delta||^2 \\
&= \text{argmin}_{\delta \in \mathbb{R}^n} \nabla f(x)^T(-\eta \cdot \nabla f(x)) + \frac{\beta}{2}\eta^2 ||\nabla f(x)||^2 \\
&= \text{argmin}_{\delta \in \mathbb{R}^n} ||\nabla f(x)||^2(-\eta + \frac{\beta}{2}\eta^2) \qquad (1) \\
&= \text{argmin}_{\delta \in \mathbb{R}^n} \frac{\beta}{2}\eta^2 - \eta \\
&= -\frac{1}{\beta} \cdot \nabla f(x)
\end{aligned}$$

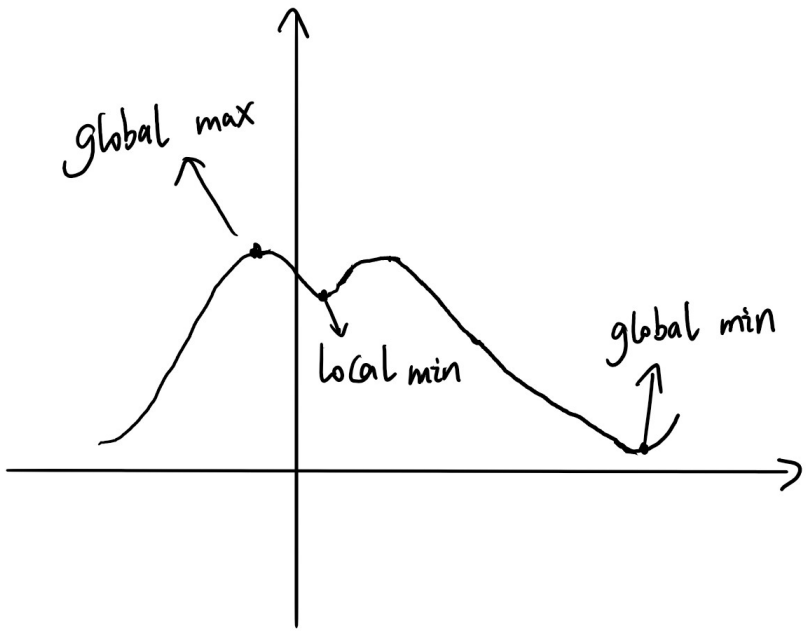
Now we can use $\delta = -\frac{1}{\beta}\nabla f(x)$ to get the upper bound on $f(x+\delta)$:

$$\begin{aligned} f(x+\delta) &\leq f(x) + \nabla f(x)^T \left(-\frac{1}{\beta}\nabla f(x)\right) + \frac{\beta}{2}\| - \frac{1}{\beta}\nabla f(x)\|^2 \\ &= f(x) + \|\nabla f(x)\|^2 \left[-\frac{1}{\beta} + \frac{1}{2\beta}\right] \\ &= f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2 \end{aligned} \tag{2}$$

if $\nabla f(x) \neq 0$, we can have $f(x+\delta) < f(x)$. It means that we are making progress and we can quantify how much progress we are making.

When $\nabla f(x) = 0$(we are not making any progress), we have following situations:

1. Global min, which is our goal
2. Local min
3. Global/local max, can be usually solved (escaped) by some random perturbations
4. Saddle point, in some directions it increases, in some other directions it decreases, or may stay at constant. Can be usually solved by some random perturbations.



Only the first situation is what we want, but it is usually easy to get out of situation 3 and 4 by some random perturbations.
In general, dealing with local minimum is a hard problem. It is non-convex optimization.
To avoid local minimum, we need another assumption: f(x) is convex.

Definition 3. *f is convex iff $\lambda_{min}\nabla^2 f(y) \geq 0$*

Fact 4. *f is convex iff $\forall x, y \in \mathbb{R}^n, \alpha \in [0,1]$, there's*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

Claim 5. *if f(x) is a convex function, then $\nabla f(x) = 0$ implies x is global minimum*

Proof: Fix x such that $\nabla f(x) = 0$, using Taylor Expansion, we can have

$$f(x+\delta) = f(x) + \nabla f(x)^T \delta + \frac{1}{2}\delta^T \nabla^2 f(y)\delta \geq f(x)$$

It means that at any other point $x + \delta$, f is at least $f(x)$, which implies that x is a global minimum.

Assumption 6. *If f is convex*

$$\nabla f(x^t) = 0$$

$\implies$ *x^t is global minimum*

$$\nabla f(x^t) \neq 0 \implies f(x^{t+1}) \leq f(x^t) - \frac{1}{2\beta}\|\nabla f(x)\|^2$$

$\implies$ *gradient descent is making progress*

Next, we need to consider how far we are from the optimal solution.
Goal: find some x' such that

$$f(x') - f(x^*) \leq \varepsilon$$

where x^* is the global minimum and x' is the $\varepsilon - approximate$ solution
We want to relate how far we are from the optimal solution to how many progress we are making. That is relate $f(x') - f(x^*)$ to $\|\nabla f(x)\|$:

$$\begin{aligned} f(x^*) &= f(x + x^* - x) \\ &= f(x) + \nabla f(x)^T (x^* - x) + (x^* - x)^T \nabla^2 f(y)(x^* - x) \\ &\geq f(x) + \nabla f(x)^T (x^* - x) \end{aligned} \tag{3}$$

After rearranging, we have:

$$\begin{aligned} \nabla f(x)^T (x - x^*) &\geq f(x) - f(x^*) \\ f(x) - f(x^*) &\leq \|\nabla f(x)\| \cdot \|x - x^*\| \\ \|\nabla f(x)\| &\geq \frac{f(x) - f(x^*)}{\|x - x^*\|} \end{aligned} \tag{4}$$

For example: if x^t s.t. $f(x^t) - f(x^*) > \epsilon \implies \|\nabla f(x)\| \geq \frac{\epsilon}{\|x - x^*\|}$

Theorem 7. *$f(x^T) - f(x^*) \leq \epsilon$ after $T = O(\beta \cdot \frac{D^2}{\epsilon})$ iterations, where $D \triangleq \max_{x: f(x) \leq f(x^0)} \|x - x^*\|$*

Proof. Let $\triangle_t = f(x^t) - f(x^*) > \epsilon$

$$\|\nabla f(x^t)\| \geq \frac{\triangle_t}{\|x^t - x^*\|} \geq \frac{\triangle_t}{D}$$

Let $T_1 = \#iterations$ until $\triangle_{T_1} \leq \frac{\triangle_0}{2}$
Before reaching $t = T_1$:

$$\|\nabla f(x^t)\| \geq \frac{\triangle_0/2}{D}$$

In each iteration, $f(x^t) - f(x^*)$ drops by $\geq \frac{1}{2\beta} \cdot \|\nabla f(x^t)\|^2 \geq \frac{1}{2\beta} \cdot \frac{\triangle_0^2}{4D^2} = \frac{\triangle_0^2}{8D^2\beta}$

$$\implies T_1 \leq \frac{\triangle_0/2}{\triangle_0^2/8D^2\beta} = \frac{4D^2\beta}{\triangle_0}$$

Let $T_2 = \#iterations$ until $\triangle_{T_1+T_2} \leq \frac{\triangle_{T_1}}{2}$. Similarly, $T_2 \leq \frac{4D^2\beta}{\triangle_{T_1}} = 4D^2\beta \cdot \frac{2}{\triangle_0}$

$$T_3 : T3 \leq 4d^2\beta \cdot \frac{4}{\triangle_0}$$

$$\vdots$$

$$T_k : T_k \leq 4D^2\beta \cdot \frac{2^k}{\triangle_0} \implies f(x^{T_1+T_2+\cdots+T_k}) - f(x^*) \leq \frac{\triangle_0}{2^k} \leq \epsilon$$

We set k so that at time $T = T_1 + T_2 + \cdots + T_k$, value:

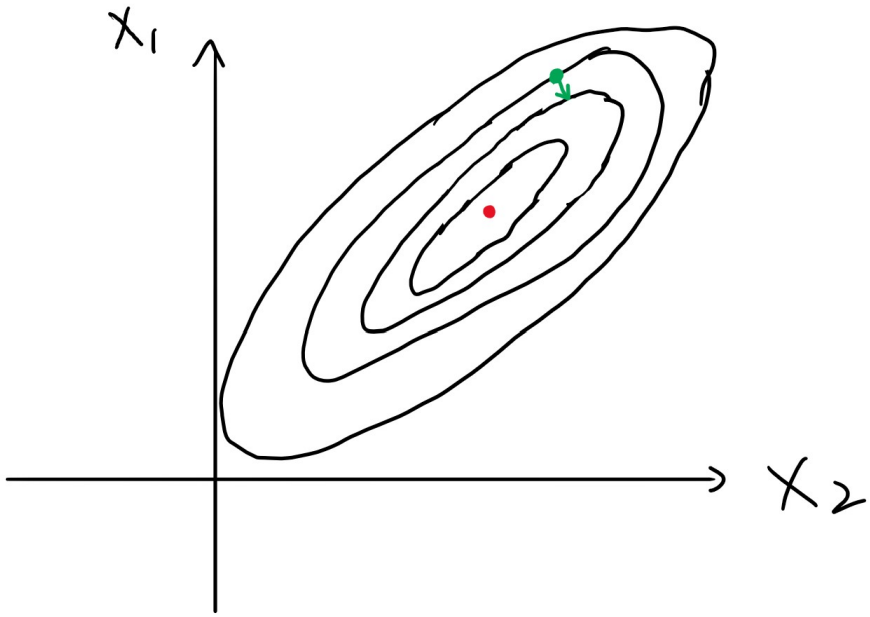
$$f(x^T) - f(x^*) \leq \epsilon,$$

i.e., $\epsilon/2 \leq \Delta_0/2^k \leq \epsilon$. Total time:

$$\begin{aligned} T &= T_1 + T_2 + \cdots + T_k \\ &= 4D^2\beta \cdot \left[\frac{1}{\triangle_0} + \frac{2}{\triangle_0} + \cdots \frac{2^k}{\triangle_0}\right] \\ &= 4D^2\beta \cdot \left[\frac{1}{\triangle_0} + \frac{2}{\triangle_0} + \cdots \frac{2}{\epsilon}\right] \\ &\leq 4D^2\beta \cdot \frac{4}{\epsilon} \end{aligned} \tag{5}$$

since the sum is geometrically increasing. Now, we have proved that when $T \leq O(\beta \cdot \frac{D^2}{\epsilon})$, $f(x^T) - f(x^*) \leq \epsilon$ □

Here is an example where lack of smoothness is bad:



Definition 8. *$\alpha -$ strongly convex for $\alpha > 0$ iff $\lambda_{min}\left(\nabla^2 f(y)\right) \geq \alpha$*