# AA Lecture 21 3/29.

## Gradient Descent

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

Taylor expansion:

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

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

(TE): $f(x+\delta) \le f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\beta \cdot \|\delta\|^2$.

$\delta$ that is min is $\delta = -\frac{1}{\beta} \cdot \nabla f(x)$.

Fix $x^t \in \mathbb{R}^d$

$$x^{t+1} = x^t - \frac{1}{\beta} \nabla f(x^t).$$

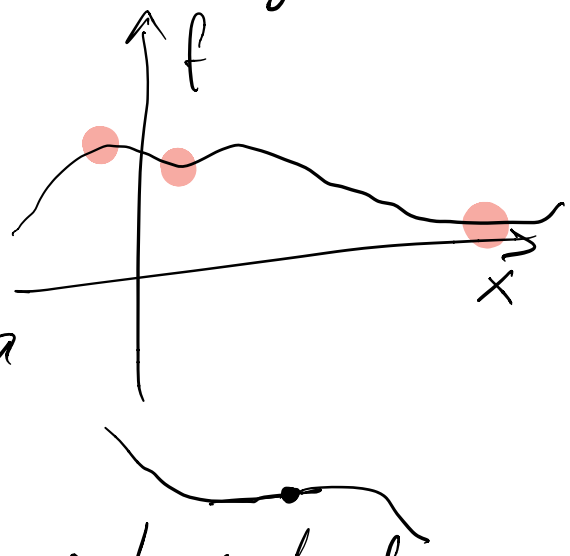
$$f(x^{t+1}) \le f(x^t) - \frac{1}{\beta} \cdot \nabla f(x^t)^T \cdot \nabla f(x^t) + \frac{\beta}{2} \cdot \frac{1}{\beta^2} \cdot \|\nabla f(x^t)\|^2$$

$$= f(x^t) - \frac{1}{2\beta} \cdot \|\nabla f(x^t)\|^2.$$

$$< f(x^t) \quad \text{if} \quad \nabla f(x^t) \neq 0.$$

# When is $\nabla f(x) = 0$?

1. Global min ← our goal!
2. Local min
3. A gl/loc. max
4. Saddle point
   (in some directions ↗
   in others ↘)



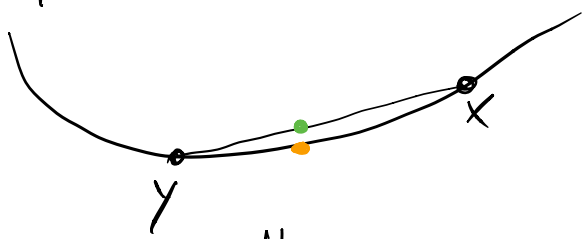

Note: usually easy to get out of #3, #4.

#2 → non-convex opt.

Def: $f$ is convex iff $\lambda_{min}(\nabla^2 f(y)) \geq 0$.

Fact: $f$ is convex iff $\forall x, y \in \mathbb{R}^d$, $\alpha \in [0,1]$

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




Claim: if $f(x)$ is convex, then

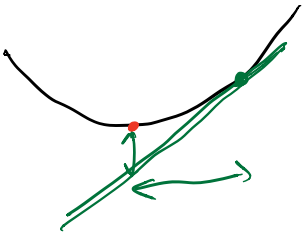
$\nabla f(x) = 0 \Rightarrow$ $x$ is global min.

pf: fix $x$ s.t. $\nabla f(x) = 0$.

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

$\nabla f(x) = 0$ $\qquad \geq 0$

$\geq f(x)$.



Assumption #2 : $f$ is convex.

$\Rightarrow \nabla f(x^t) = 0 \Rightarrow$ done! $x^t$ = gl. min.

$\nabla f(x^t) \neq 0 \Rightarrow$ GD is making progress.

$$f(x^{t+1}) \leq f(x^t) - \frac{1}{2\beta} \cdot \|\nabla f(x)\|^2.$$

Goal: find some $x'$ s.t.

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

($f(x^*)$: global min)

$x' \rightarrow$ $\varepsilon$-approx. sol.

Relate $f(x^t) - f(x^*)$ to $\|\nabla f(x)\|$:

$$f(x^*) = f\left(x + (x^* - x)\right)$$

$$\overset{TE}{=} f(x) + \nabla f(x)^T \cdot (x^* - x) + \underbrace{\frac{1}{2}(x^* - x)^T \cdot \nabla^2 f(y) \cdot (x^* - x)}$$

$$\geq f(x) + \nabla f(x)^T \cdot (x^* - x)$$

$$\nabla f(x)^T \cdot (x - x^*) \geq f(x) - f(x^*).$$

$p^T q \leq \|p\|_\alpha \|q\|_\beta$, $\frac{1}{\alpha} + \frac{1}{\beta} = 1$

Hölder ineq.

$$f(x) - f(x^*) \leq \|\nabla f(x)\| \cdot \|x - x^*\|.$$

$$\|\nabla f(x)\| \geq \frac{f(x) - f(x^*)}{\|x - x^*\|}.$$

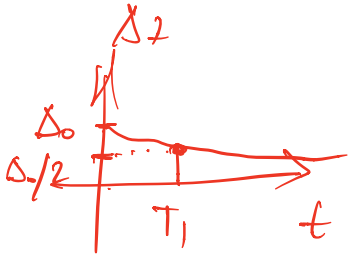
For example: if $x^t$ s.t. $f(x^t) - f(x^*) \geq \varepsilon$

$$\Rightarrow \|\nabla f(x)\| \geq \frac{\varepsilon}{\|x - x^*\|}.$$

**Thm:** $f(x^T) - f(x^*) \leq \varepsilon$ after

$$T = O\left(\beta \cdot D^2/\varepsilon\right) \text{ iterations}$$

$$D \triangleq \max \|x - x^*\|.$$
$$x: f(x) \leq f(x^0)$$




**Pf:**

Let $\Delta_t = f(x^t) - f(x^*) \geq \varepsilon$.

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

Let $T_1$ = # iterations until $\Delta_{T_1} \leq \frac{\Delta_0}{2}$.

Before reaching $t = T_1$:

$$\|\nabla f(x^t)\| \geq \frac{\Delta_0/2}{D}.$$

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

$$\geq \frac{1}{2\beta} \cdot \frac{\Delta_0^2}{4D^2} = \frac{\Delta_0^2}{8D^2\beta}.$$

$$\Rightarrow T_1 \leq \frac{\Delta_0/2}{\Delta_0^2/8D^2\beta} = 4D^2\beta/\Delta_0.$$

Let $T_2$ = # iterations until $\Delta_{T_1+T_2} \leq \frac{\Delta_{T_1}}{2}$.

Similarly, $T_2 \leq 4D^2\beta/\Delta_{T_1} = 4D^2\beta \cdot 2/\Delta_0$.

$T_3$: $T_3 \leq 4D^2\beta \cdot 4/\Delta_0$

⋮

$T_k$ : $T_k \le 4D^2\beta \cdot 2^k/\Delta_0 \Rightarrow f(x^{T_1+\cdots+T_k}) - f(x^*) \le \frac{\Delta_0}{2^k} \le \varepsilon$.

At time $T = T_1 + T_2 + \cdots + T_k$, value:

$$f(x^T) - f(x^*) \le \varepsilon$$

Total time: $T = T_1 + T_2 + \cdots + T_k$

$$= 4D^2\beta \cdot \left[ \frac{1}{\Delta_0} + \frac{2}{\Delta_0} + \cdots + \frac{2^k}{\Delta_0} \right]$$

$$= 4D^2\beta \left[ \frac{1}{\Delta_0} + \frac{2}{\Delta_0} + \cdots + \frac{1}{\varepsilon} \right]$$

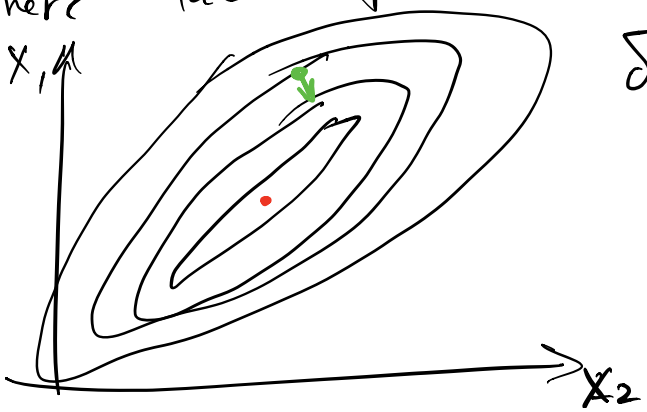
$$\le 4D^2\beta \cdot 2/\varepsilon \cdot \quad \boxtimes$$

$T \le O(\beta D^2/\varepsilon)$.

$f(x^T) - f(x^*) \le \varepsilon$.

---

Example where lack of smoothness is bad:




$\delta = -\frac{1}{\beta} \cdot \nabla$.

**Def:** $\alpha$ - strongly convex for $\alpha > 0$ iff

$$\lambda_{min}\left(\nabla^2 f(y)\right) \geq \alpha.$$