* HW 5 due Nov 4
* Today: GD for Smooth fns, Online gradient descent

Smoothness and Strong-Convexity

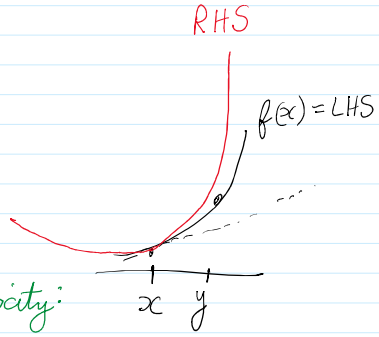
Roughly, Smoothness is an upper bound on 2nd derivative
Strong-convexity is a lower " "

β-Smoothness:

(1) $\nabla^2 f(x) \lesssim \beta$ ← all e.v.s $\leq \beta$

(2) $\|\nabla f(x) - \nabla f(y)\| \leq \beta \cdot \|x-y\|$

(3) $f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|^2$



Recall 1st-order defn of convexity:

$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

α-Strong Convexity:

(1) $\nabla^2 f(x) \gtrsim \alpha$ ← all e.v.s $\geq \alpha$

(2) $\|\nabla f(x) - \nabla f(y)\| \geq \alpha \cdot \|x-y\|$

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

E.g.: $f(x) = \|Ax-b\|^2/2$
$\nabla f(x) = Ax-b$, $\nabla^2 f(x) = A$
β = Largest e.v. of A^TA
α = Min e.v. of A^TA

What can be shown:

	Function class	# iterations (ignoring non-ϵ terms)	Step Size
Last lecture → 1)	Convex	$1/\epsilon^2$	$1/\sqrt{T}$
HW5 → 2)	Strongly-convex	$1/\epsilon$	$1/(\alpha t)$
Today → 3)	Smooth	$1/\epsilon$	$1/\beta$
4)	Str-convex + Smooth	polylog$(1/\epsilon)$	$1/\beta$

Gradient Descent for β-Smooth functions

$$x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$$

Thm: Grad descent on β-smooth convex fn and given x_0 s.t. $\|x_0 - x^*\| \le D$ gives

$$f(x_T) - f(x^*) \le \beta \frac{D^2(1+\ln T)}{2T}$$

Remark: (1) The $\ln T$ can be removed with a better analysis.
(2) This result is about x_T (and not $\frac{\sum_t x_t}{T}$)

Proof: Define potential $\phi_t := t\cdot\left[f(x_t) - f(x^*)\right]$

Want to show ϕ_T is small

$$\phi_{t+1} - \phi_t = (t+1)\left[f(x_{t+1}) - f(x^*)\right] - t\left[f(x_t) - f(x^*)\right]$$

$$= (t+1)\underbrace{\left[f(x_{t+1}) - f(x_t)\right]}_{\text{Claim 1}} + \underbrace{\left[f(x_t) - f(x^*)\right]}_{\text{Claim 2}}$$

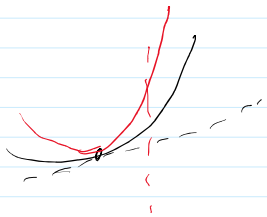
$$\le -\frac{\|\nabla f(x_t)\|^2}{2\beta} \qquad \le \frac{(t+1)}{2\beta}\|\nabla f(x_t)\|^2 + \frac{D^2\beta}{2(t+1)}$$

Assuming claims,

$$\phi_T - \phi_0 = \sum_{t=0}^{T-1}(\phi_{t+1} - \phi_t) \le \frac{D^2\beta}{2}\sum_{t=0}^{T-1}\frac{1}{t+1} \le \frac{D^2\beta}{2}(1+\ln T)$$

$$\phi_T - \phi_0 = T\left[f(x_T) - f(x^*)\right]$$

Claim 1: $f(x_{t+1}) - f(x_t) \le -\frac{\|\nabla f(x_t)\|^2}{2\beta}$



Pf: LHS $\overset{\text{smoothness}}{\le} \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$

$$= -\eta\|\nabla f(x_t)\|^2 + \frac{\beta}{2}\eta^2\|\nabla f(x_t)\|^2 = \text{RHS for } \eta = 1/\beta$$

Claim 2: $f(x_t) - f(x^*) \le \frac{t+1}{\beta}\frac{\|\nabla f(x_t)\|^2}{2} + \frac{D^2\beta}{2(t+1)}$

Claim 2: $f(x_t) - f(x^*) \le \frac{t+1}{\beta} \frac{\|\nabla f(x_t)\|^2}{2\beta} + \frac{D^2\beta}{2(t+1)}$

Pf: LHS $\overset{\text{Convexity}}{\le} \langle \nabla f(x_t), x_t - x^* \rangle \le \|\nabla f(x_t)\| \cdot \|x_t - x^*\|$

$$\le \underbrace{\|\nabla f(x_t)\| \sqrt{\frac{t+1}{\beta}}}_{a} \cdot \underbrace{\frac{D}{\sqrt{(t+1)/\beta}}}_{b}$$

Recall $(a-b)^2 \ge 0 \Rightarrow \frac{a^2+b^2}{2} \ge ab$

$$\le \frac{1}{2}\left[\|\nabla f(x_t)\|^2 \frac{t+1}{\beta} + \frac{D^2\beta}{t+1}\right] = RHS$$

Online Convex Optimization / Online learning

Example (Regression): Given T labeled samples $(a_1, b_1) \cdots (a_T, b_T)$ where $a_t \in \mathbb{R}^n$ and $b_t \in \mathbb{R}$

Find x s.t. $a_t x \approx b_t$ ← offline optim / Learning

Formally, $\min_x \sum_{t=1}^{T} (a_t x - b_t)^2 = \min_x \underbrace{\|Ax - b\|^2}_{f(x)}$

Gradient descent will find x s.t. $f(x) - f(x^*) \le \epsilon$.

What if the samples not given upfront and arrive over time?

Problem Model

1) We are given a convex body $K \subseteq \mathbb{R}^n$
2) T rounds:
   (a) Alg plays $x_t \in K$
   (b) Convex cost function $f_t : \mathbb{R}^n \to \mathbb{R}$ is revealed

E.g. $f_t(x) = (a_t x - b_t)^2$

(b) Convex cost function $f_t : \mathbb{R}^n \to \mathbb{R}$ is revealed

Goal: $\min \sum_{t=0}^{T-1} f_t(x_t)$

Benchmark: $\min_{x^* \in K} \sum_{t=0}^{T-1} f_t(x^*)$

$x_{t+1} = x_t - \eta \nabla f_t(x_t)$

Thm: Online gradient descent guarantees

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[f_t(x_t) - f_t(x^*) \right] \leq \frac{DG}{\sqrt{T}}$$

where $D = \text{Diameter}(K)$

& $G = \max_t \|\nabla f_t(x)\|$ for $x \in K$