Lec 18

Friday, October 24, 2025 6:41 AM

* Exam 2 : Median 70 ± 17
* HW 4 graded
* HW 5 out, due Nov 4 (One more HW with 4+ pbs after this)
* Recap : Convex fns and Convex Optim

Gradient Descent (assume no constraints K)

$$\min_{x \in \mathbb{R}^n} f(x)$$ ← we want x^* s.t. $\|\nabla f(x^*)\| \approx 0$

(we are not given ∇f fn explicitly to solve directly)

Since we want to find a local opt,

take a step in $-\nabla f(x)$ ← direction of steepest descent

Grad Descent Alg:

1) Start with x_0

2) In t-th step :

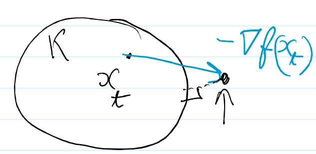
$$x_t = x_{t-1} - \underbrace{\eta}_{\text{Step size}} \nabla f(x_{t-1})$$

Exercise: $\lim_{\epsilon \to 0} \max_{y : \|y\| = \epsilon} \frac{f(x+y) - f(x)}{\epsilon} = \|\nabla f(x)\|_2$

→ We will show that after poly$(1/\epsilon)$ steps, G.D. finds ϵ-optimal soln.

Properties:
(1) G.D. is fast when ϵ is not very small.
(2) Only requires gradient oracle
(3) Well-defined (and works) even for non-convex fns to give local-opt.
(4) Similar results even hold with convex constraints K, but then we need to project onto K

Thm 1: Assume $\|\nabla f(x)\|_2 \le G$ and $\|x_0 - x^*\| \le D$.



After T steps of G.D with $\eta = \frac{D}{G\sqrt{T}}$,

$$f\left(\frac{\sum_{t=0}^{T-1} x_t}{T}\right) - f(x^*) \le \frac{GD}{\sqrt{T}} \Rightarrow \text{For } \epsilon \text{ error, } T \ge \frac{G^2 D^2}{\epsilon^2}.$$

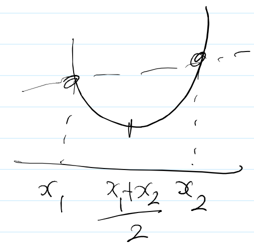
$$f\left(\frac{\sum_{t=0} x_t}{T}\right) - f(x^*) \leq \frac{GD}{\sqrt{T}} \implies \text{For } \epsilon \text{ error, } T \geq \frac{G^2 D^2}{\epsilon^2}.$$

Obs: Suffices to prove

$$\sum_{t=0}^{T-1} \Big[\overbrace{f(x_t) - f(x^*)}^{\text{Regret at step } t}\Big] \leq \frac{GD}{\sqrt{T}} \times T$$

since

$$f\left(\frac{\sum_t x_t}{T}\right) \overset{\text{Convexity}}{\leq} \sum_t \frac{f(x_t)}{T}$$



Pf: Define a potential function $\phi_t = \|x_t - x^*\|^2 \cdot \frac{1}{2\eta}$

We will show that

$$\Big[f(x_t) - f(x^*)\Big] \leq \Big(\phi_t - \phi_{t+1}\Big) + \frac{\eta G^2}{2} \qquad (1)$$

This suffices since

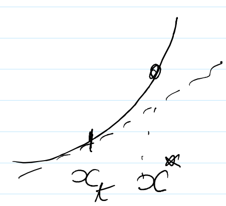
$$\sum_{t=0}^{T-1} \underbrace{\Big[f(x_t) - f(x^*)\Big]}_{\text{Regret}} \leq \overbrace{\phi_0}^{= \frac{D^2}{2\eta}} - \underbrace{\phi_T}_{\geq 0} + \frac{\eta G^2}{2} T$$

$$\leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2} = DG\sqrt{T}. \qquad \boxed{\eta = \frac{D}{G\sqrt{T}}}$$

Finally, to prove (1),

$$2\eta \cdot \phi_{t+1} = \|x_{t+1} - x^*\|^2$$

$$= \|x_t - \eta \nabla f(x_t) - x^*\|^2$$



$$= \underbrace{\|x_t - x^*\|^2}_{= \phi_t \cdot 2\eta} + \eta^2 \underbrace{\|\nabla f(x_t)\|^2}_{\leq G^2} + 2\eta \underbrace{\langle \nabla f(x_t), x^* - x_t \rangle}_{\leq f(x^*) - f(x_t)}$$

$$\implies f(x_t) - f(x^*) \leq \phi_t - \phi_{t+1} + \frac{\eta G^2}{2}.$$

∎

Takeaways and Remarks

1) Grad descent gets ϵ-approx in $\approx 1/\epsilon^2$ steps

1) Grad descent gets ϵ-approx in $\simeq 1/\epsilon^2$ steps

2) This can be shown to be optimal for 1st-order methods (grad-based)

3) we can do better if more assumptions on gradient.

Smoothness and Strong-Convexity

Roughly, smoothness is an upper bound on 2nd derivative

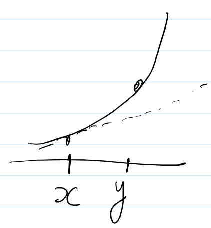
strong-convexity is a lower " "

β-Smoothness:

(1) $\nabla^2 f(x) \precsim \beta$ $\leftarrow$ 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) $\nabla^2 f(x) \succsim \alpha$ $\leftarrow$ 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.: $f(x) = \|Ax-b\|^2/2$

$\nabla f(x) = (Ax-b)$

β = Largest e.v. of A^TA

α = Min e.v. of A^TA

What can be shown:

	Function class	# iterations (ignoring non-ϵ dependency)
1)	Convex	$1/\epsilon^2$
2)	Strongly-convex	$1/\epsilon$
3)	Smooth	$1/\epsilon$
4)	Str-convex + Smooth	$\text{polylog}(1/\epsilon)$

4) Str-convex + Smooth | polylog$(1/\epsilon)$