1st lecture Convex Optimization.

We will see 3 results today.

→ Gradient Descent linear convergence for smooth, strongly convex functions.

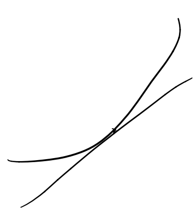
→ Gradient Descent sublinear conv for smooth convex functions. → Interlude Nesterov.

→ Stochastic Gradient Descent Convergence.

Notations / Definitions → $F(x)$ convex

if $F(\alpha x + (1-\alpha) y) \leq \alpha F(x) + (1-\alpha) F(y)$

Alternative: $F(y) \geq F(x) + \langle \nabla F(x), y-x \rangle$ $\forall y, x$.



$\gamma(\alpha)$

$$F(y) - F(x) = \int_0^1 \langle \nabla F(\gamma(\alpha)), y-x \rangle d\alpha$$

$$(1-\alpha) \int_0^\alpha f'(t)\, dt \leq \alpha \int_\alpha^1 f'(t)\, dt$$

$$(1-\alpha)\,[f(\alpha)-f(0)] \leq \alpha\,[f(1)-f(\alpha)]$$

$$(1-\alpha)\, f(\alpha) + \alpha f(\alpha) \leq \alpha f(1) + (1-\alpha) f(0)$$

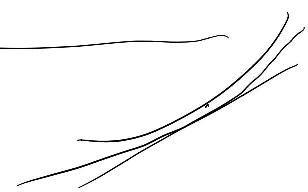
dC

So $$\frac{(1-\alpha)}{\alpha}\int_0^\alpha f'(t)\,dt \leq \int_\alpha^1 f'(t)\,dt$$

so $$\lim_{\alpha\to 0} (1-\alpha)\left[\frac{1}{\alpha}\int_0^\alpha f'(t)\,dt\right] \leq \lim_{\alpha\to 0}\int_\alpha^1 f'(t)\,dt$$

$$f'(0) \leq \int_0^1 f'(t)\,dt.$$

→ $F(x)$ strongly convex



if $$F(y) \geq F(x) + \langle \nabla F(x), y-x\rangle + \frac{\ell}{2}\|y-x\|^2$$

($HF(x)$ has eigenvalue at least ℓ). Taylor)

→ F is smoothly convex if ∇F is Lipschitz:

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x-y\|.$$

This implies by Taylor that

$$F(y) \leq F(x) + \langle \nabla F(x), y-x\rangle + \frac{L}{2}\|x-y\|^2$$

Strongly convex case:

$$\| x - t\nabla F(x) - (y - t\nabla F(y)) \| \leq \max(|1-tL|, |1-tl|) \cdot \|x-y\|.$$

$\nabla F(x)$ is Bilipschitz with constants

$\to HF(x)$ has eigenvals in (l, L)

$\Rightarrow H\,tF$ has constants (tl, tL)

$\Rightarrow I - t\nabla F$ is Lipschitz w/ constant $\max(|1-tl|, |1-tL|)$.

$\to$ say it more in terms of Hessians instead.

$$\|x_{k+1} - x_*\| \leq \| x_k - t\nabla F(x_k) - x_* - t\nabla F(x_*)\|$$

$$\leq \max(|1-tl|, |1-tL|) \|x_k - x_*\|$$

$t := \frac{2}{l+L}$ minimizes and we obtain

$\left(\frac{L-l}{L+l}\right)$; in terms of condition number $K = \frac{L}{l}$,

$$\left(\frac{K-1}{K+1}\right)^K$$

:)

Ex: $F(x) = \frac{1}{2}\|\tilde{A}x - \tilde{b}\|^2 = \frac{1}{2}x^T A x - b^T x + c$

$$DF(x_k) = A\, x_k - b \qquad A = \tilde{A}^T \tilde{A}, \quad b = \tilde{A}^T \tilde{b}$$

$$x_0 = tb$$

$$x_1 = x_0 - t\nabla F(x_0) =$$

$$= x_0 - t(A x_0 - b) = (I - tA)\, x_0 + tb$$
$$= (I - tA) x_0 + x_0 .$$

$$x_2 = x_1 - t\nabla F(x_1) = (I - tA)\big((I - tA) x_0 + x_0\big) + x_0 .$$

$$\Rightarrow x_k = \Big[\sum_{j \le k} (I - tA)^j\Big](x_0).$$

$$\frac{1}{z} = \sum (1-z)^k \qquad |z| < 1$$

A has eingenvalues (l, L)

$\Rightarrow tA$ has eigenvalues $\frac{2}{l+L}(l, L)$

$\Rightarrow I - tA$ has eigenvalues $\left(\frac{2l - l - L}{l+L}, \frac{2L - l - L}{l+L}\right)$

$\left(\frac{l-L}{l+L}, \frac{L-l}{L+l}\right)$ between $(-1, 1)$

So the series $\sum_{j \leq k} (I - tA)^j$ converges to A^{-1}, with rate $O(\|(I-tA)^k q\|) = (1 - \frac{l}{L})^k$

Q: Can we do better?

Find a k-degree polynomial $q_k(A)$ minimizing the residual error

$$\| (I - A q_k(A)) b \|$$

The polynomial correspondent form is

$$p_k(z) = 1 - z q_k(z)$$

When applied to A, small norm $\Rightarrow$ eigenvalues of A are bounded in the interval (l, L). also, we need

$$p_k(0) = 1.$$

$\hookrightarrow$ Chebyshev polynomials are optimum at that

Tmel.

<u>Lemma</u>: There exist $P_k(z)$ of degree

$O\left(\sqrt{(L/\ell)\log(1/\varepsilon)}\right)$ with $P_k(0)=1$ and

$|P_k(z)| \leq \varepsilon \quad \forall\, z \in [\ell, L]$.

quadratic savings in degree.

$$O\left(\sqrt{\frac{K-1}{K+1}}\right)^k \qquad \tilde{\rho}^k = \rho^{\sqrt{k}}$$

Moreover, Chebyshev Polynomials can be obtained recursively using two previous polynomials.

$$\boxed{\Rightarrow \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k) - \beta_k \nabla f(x_{k-1})}$$

↳ This extends to generic convex functions (Nesterov)

↳ This rate is optimal: cannot be improved

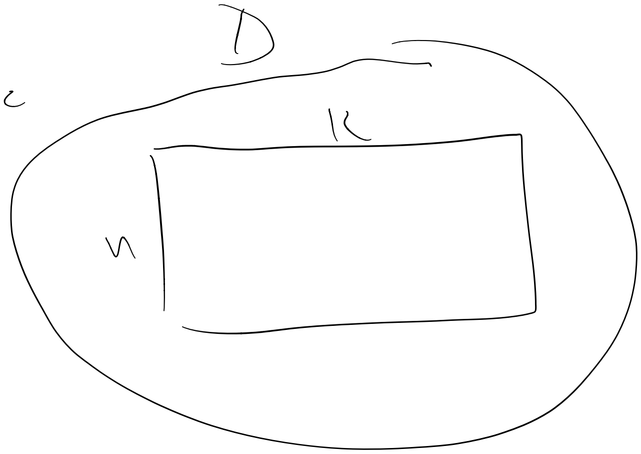
↳ Alternative interpretation of acceleration (Bubeck, Lee, Singh) (using a variant of ellipsoid method).

$$B = D^T D \in \mathbb{R}^{k \times k} \qquad D \in \mathbb{R}^{n \times k} \quad (n < k)$$

$$B = F^T \Lambda F = F^T \tilde{\Lambda}^T \tilde{\Lambda} F$$

$$D = \tilde{\Lambda} \mathcal{F}$$

$$\mathcal{F} = \mathcal{F}_N \in \mathbb{R}^{n \times k}$$



$\|z\|_1 \in \mathbb{R}^k$ $\qquad$ $\mathcal{F} \in \mathbb{R}^{k \times k}$

$\|\mathcal{F}_N z\|_1$ $\qquad$ $A \in \mathbb{R}^{k \times n}$

$$\|\mathcal{F} z\|_1 \approx \|A z\|_1 \cdot \left(\frac{k}{n}\right)$$

$\|z\|_1$