Optimization for Machine Learning
CSCI-599

Lecture 2: Gradient Descent

Sai Praneeth Karimireddy

USC – https://spkreddy.org/optmlspring2025.html

January 15, 2025

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

$$f^* = \inf_x f(x)$$

opt 1st $\nabla f(x^*) = 0$ $\quad \|\nabla f(x^*)\|_2 \le \varepsilon$

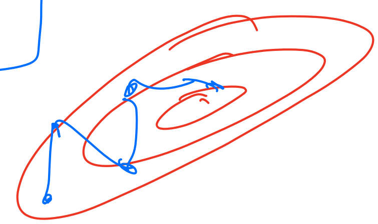
0th $f(x^*) - f^* \le 0$?

$\hat{x}$ ~~Suppose f^* exists~~

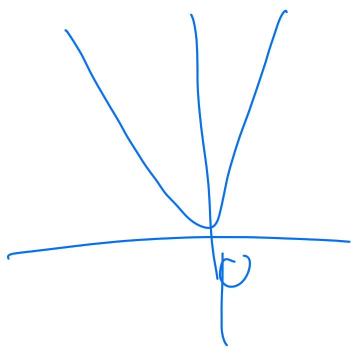
$\left(f(\hat{x}) - \underset{?}{f^*} \le \varepsilon \right)$ ✓

$\boxed{\text{Suppose } \exists\, x^* \text{ s.t. } f(x^*) = f^*,}$

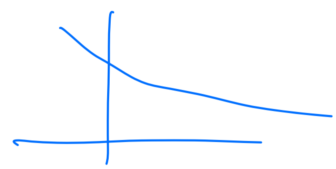
$\|\hat{x} - x^*\|_2 \le \varepsilon$



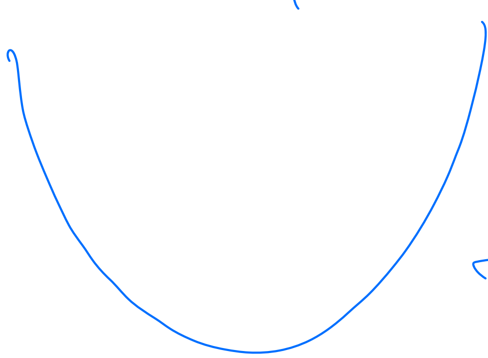
$|x|$



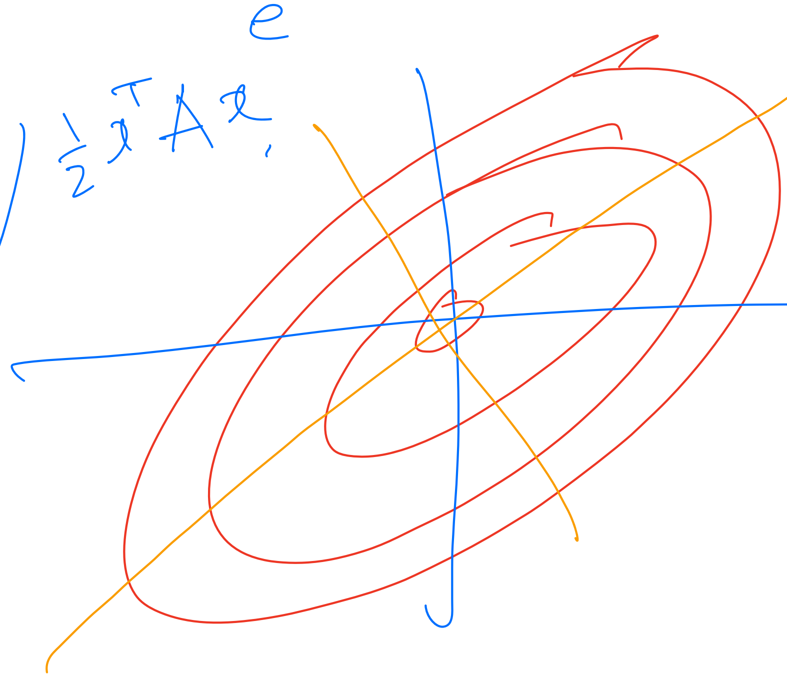
$\inf_x f(x) = x$



e^{-x}



$\frac{1}{2} x^T A x,$



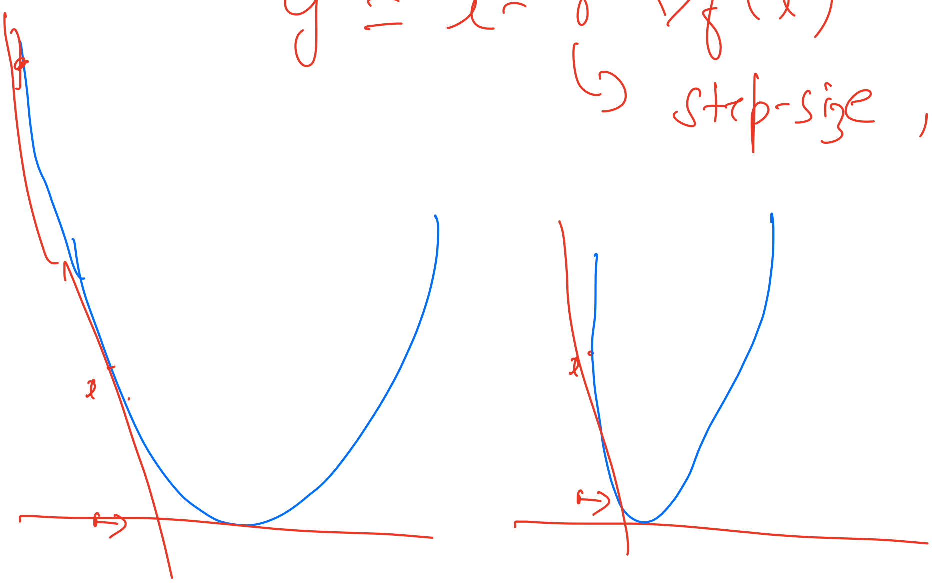
By Taylor's Expansion

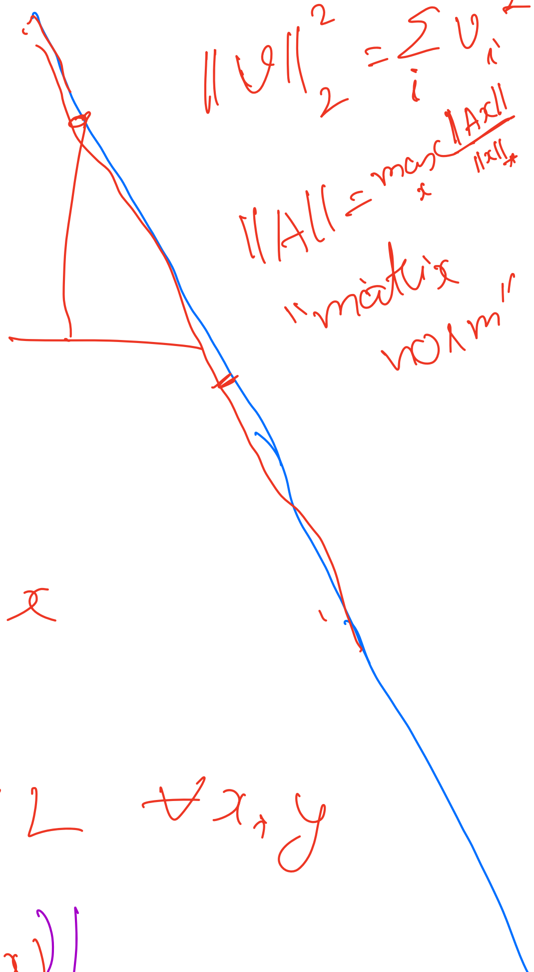
for a "close enough" y,

$$f(y) \text{ "}\approx\text{" } f(x) + \nabla f(x)^T \underbrace{(y-x)}_{\to\, -\nabla f(x)}$$

$$y = x - \gamma \nabla f(x)$$

γ ↳ step-size, learning rate





$$\|v\|_2^2 = \sum_i v_i^2$$

$$\|A\| = \max_x \frac{\|Ax\|}{\|x\|_*}$$

"matrix norm"

Def L-Smoothness

2nd order $\quad \|\nabla^2 f(x)\|_2 \le L \quad \forall x$ $\quad$ (→ norm)

1st order $\quad \dfrac{\|\nabla f(y) - \nabla f(x)\|_2}{\|y-x\|_2} \le L \quad \forall x, y$

0th order $\quad \dfrac{|f(y) - (f(x) + \nabla f(x)^T (y-x))|}{\|y-x\|_2^2} \le \dfrac{L}{2} \quad \forall x, y$

$\nabla f(y)$

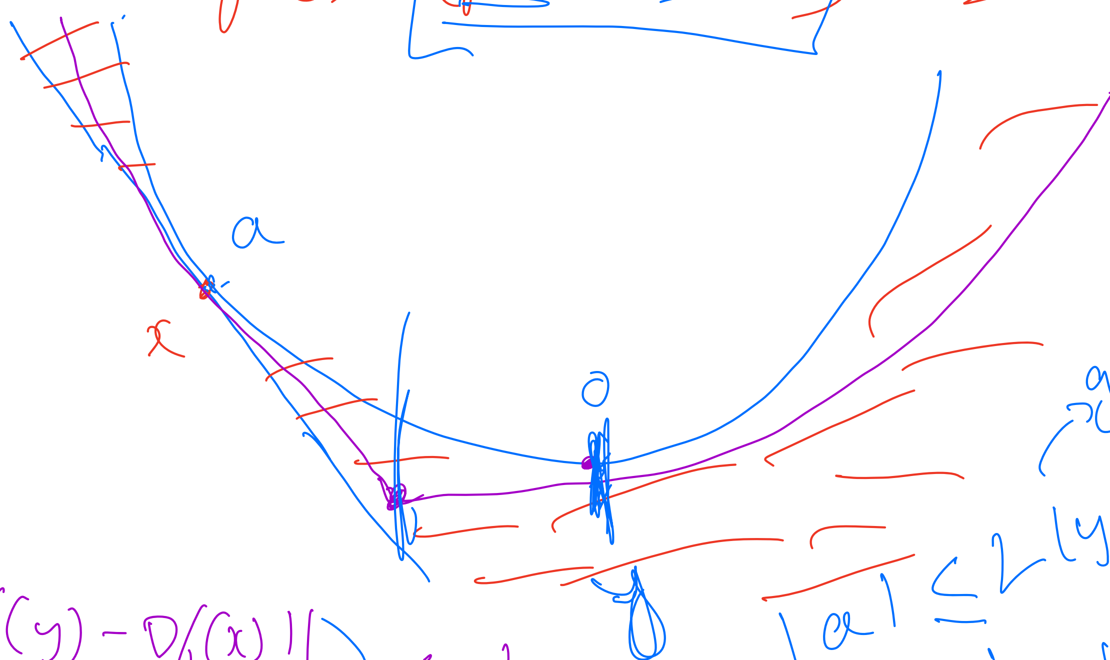
$$\frac{f(y) - f(x)}{\|y-x\|_2 \cdot \|y-x\|_2} - \frac{\nabla f(x)^T \cancel{(y-x)}}{\|y-x\|_2^2} \approx \frac{\|\nabla f(y) - \nabla f(x)\|}{\|y-x\|}$$

$$\boxed{f(y) \le \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2}_{g(y)} \quad \forall x, y \in D}$$

$\nabla g(y) = \nabla f(x) + L(y-x)$, $\nabla^2 g(y) = L$

f is cvx & L-smooth $\Rightarrow$ $\nabla g(x) = \nabla f(x)$

$$0 \le f(y) - \left(f(x) + \nabla f(x)^T (y-x)\right) \le \frac{L}{2}\|y-x\|^2$$



$$\frac{\|\nabla f(y) - \nabla f(x)\|}{\|y-x\|} \le L$$

$|a| \le L|y-x|$

$|-a-\delta| = \frac{|a+\delta|}{f(x-y)}$

$y-x \ge L$

Claim: Minimizer of the upper quadratic never "overshoots"

$\Rightarrow$ $\gamma = \frac{1}{L}$ never "overshoots"

$$\boxed{y = x - \frac{1}{L}\nabla f(x)}$$

$$f(x) = \frac{1}{2}x^T A x - b^T x$$

np.inv

np.solve

$$\nabla f(x) = \frac{1}{2}(A + A^T)x - b$$

A is symmetric & inv

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

$$x^* = A^{-1}b \qquad x_0 = 0$$

$$x_{t+1} = x_t - \gamma(Ax_t - b) \quad -① $$

Setting γ

$$\nabla^2 f(x) = A$$

$$L = \|A\|_2$$

$$\Rightarrow \gamma = \frac{1}{\|A\|_2}$$

Unrolling ① $\Rightarrow$

$$x_{t+1} = (I - \gamma A)x_t + \gamma b$$

$$x_{t+1} = \underline{\qquad\qquad\qquad}$$

$$\|x_{t+1} - A^{-1}b\|_2 \leq ???$$

Claim: if f is cvx and L-smooth

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

① $f(x_t) - f^* \leq O\left(\frac{L}{t}\right)$

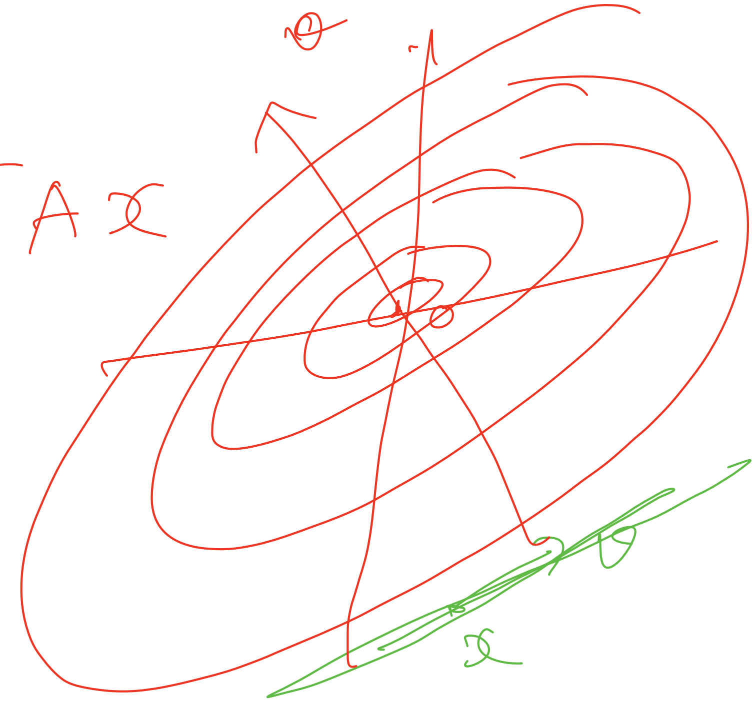
② if f μ-strongly-cvx $\Rightarrow$

$$f(x_t) - f^* \leq O\left(\exp\left(-\frac{\mu}{L} t\right)\right)$$

Crash Course on Spectral Norm

→ symmetric p.s.d. finite dim.

A $\quad d \times d$



$$f(x) = \frac{1}{2} x^T A x$$

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

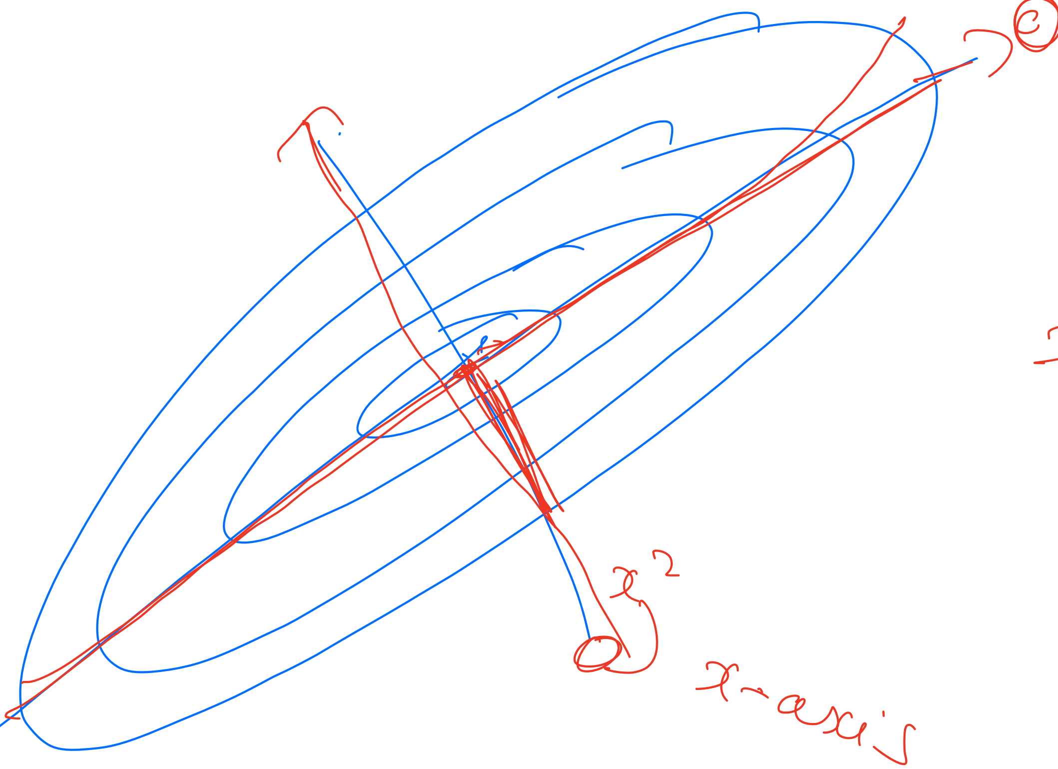
$$\nabla^2 f(x) = A$$

$$\to g_v(\lambda) = f(x + \lambda v) = \frac{1}{2} x^T A x + Ax^T v \lambda + \frac{1}{2}\lambda^2 v^T A v$$

$$\to \frac{d\, g_v(\lambda)}{d\lambda} = Ax^T v$$

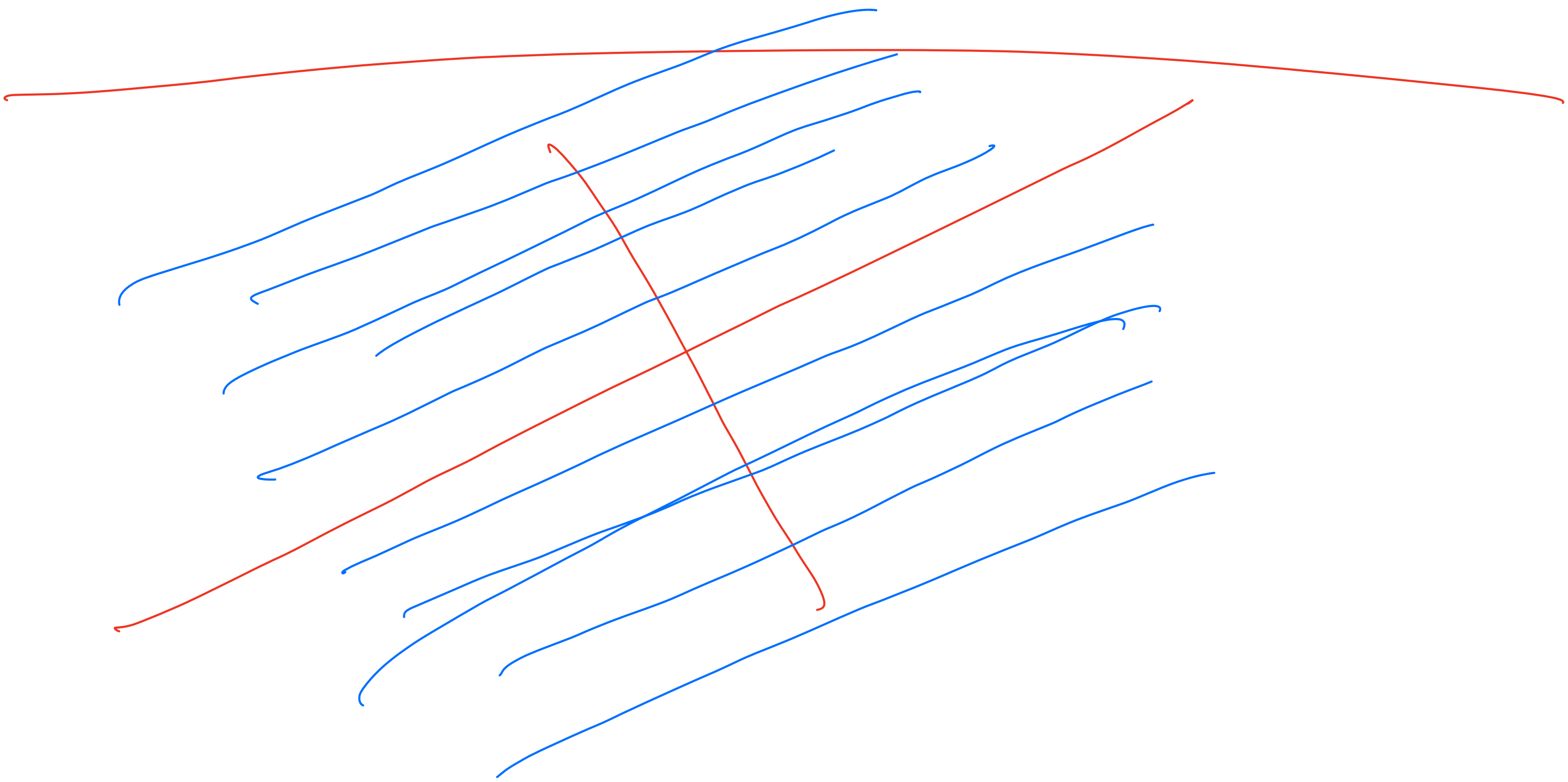
$$\frac{d^2 g_v(\lambda)}{d\lambda^2} = \boxed{v^T A v}$$

min eigenvalue



$\|A\|_2 = 2$

$= \dfrac{\text{max eigenvalue}}{\text{min eigenvalue}}$



Jan 27

A cvx function f is L-smooth if

(i) $\|\nabla^2 f(x)\|_2 \leq L \quad \forall x$

(ii) $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2 \quad \forall x, y$

(iii) $f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$

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

~~Thm~~ Claim

$$f(x_{t+1}) - f^* \leq \text{some decreasing function of } t$$

$f(x) = \frac{1}{2}x^T(A^TA)x - b^Tx$ $\quad$ A^TA is symm, pd

$x_{t+1} = x_t - \gamma(A^TAx_t - b)$ $\quad$ $\Rightarrow x^* = (A^TA)^{-1}b$

$\because \nabla f(x) = (A^TA)x - b$

at x^* $\nabla f(x^*) = 0$

Observation 1

$$f(x) - f^* = \frac{1}{2}\|A(x-x^*)\|^2 \longrightarrow A^TA(x-x^*)$$

$$= A^TAx - \cancel{A^TA(A^TA)}\,b$$

$$= A^TAx - b$$

$$\nabla f(x) = (A^TA)x - b$$

$$\|x_{t+1} - x^*\|_2^2 = \|x_t - \gamma(A^TAx_t - b) - x^*\|_2^2$$

$$= \|x_t - x^*\|^2 - 2\gamma(A^TAx_t - b)^T(x_t - x^*) + \gamma^2\|A^TAx_t - b\|_2^2$$

< 0

Attempt 1

$$-2\gamma(A^TAx_t - b)^T(x_t - (A^TA)^{-1}b)$$

$$-2\gamma(A^TAx_t - b)^T(A^TA)^{-1}(A^TA)(x_t - (A^TA)^{-1}b)$$

$$-2\gamma\boxed{(A^TAx_t - b)^T}\boxed{(A^TA)^{-1}}\boxed{(A^TAx_t - b)} > 0$$

$$\le -2\gamma\lambda_{min}\|A^TAx_t - b\|^2$$

P.d M

$M \succeq \lambda_{min} I$

$$(M - \lambda_{min} I) \succeq 0$$

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\gamma \lambda_{min} \|A^T A x_t - b\|^2 + \gamma^2 \|A^T A x_t - b\|^2$$

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

$$\|A(x_t - (A^T A)^{-1} b\|^2$$

Attempt 2

$$-2\gamma \quad (A^T A x_t - b)^T (x_t - (A^T A)^{-1} b) + \gamma^2 \|A^T A x_t - b\|^2$$

$$(A^T A x_t - b)^T (A^T A x_t - b)$$

$$= \quad A(x_t - (A^T A)^{-1} b)^T (A A^T) \; A(x_t - (A^T A)^{-1} b)$$

Strong Convexity

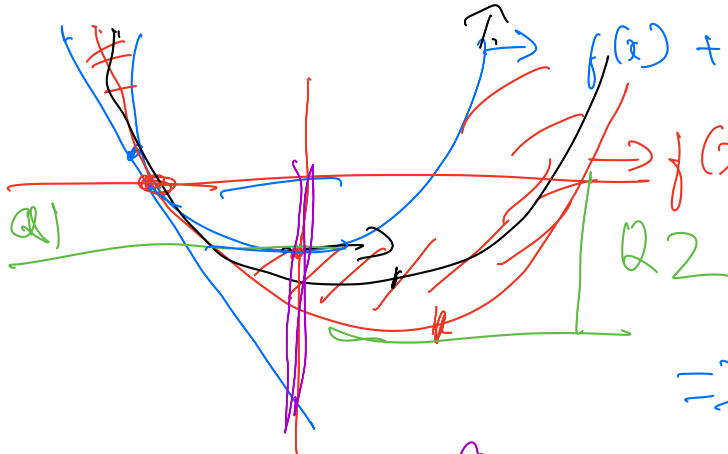
- $\nabla^2 f(x) \succeq 0$ — cvx
- $\nabla^2 f(x) \succ 0$ — strict cvx
- $\boxed{\nabla^2 f(x) \succeq \mu I}$ — μ-strongly cvx, $\mu > 0$
- $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y-x\|^2$

$\Rightarrow \|\nabla f(y) - \nabla f(x)\| \geq \mu \|y - x\|$

$\frac{L}{2}\|y-x\|^2 \Rightarrow \left| f(y) - \left[f(x) + \nabla f(x)^T (y-x) \right] \right| \geq \frac{\mu}{2}\|y-x\|^2$

f is L-smooth $\Leftarrow$ μ-strongly cvx

$y = x - \frac{1}{L}\nabla f(x)$



$\Rightarrow f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$

$\Rightarrow f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$

$\Rightarrow L \geq \mu$

$Q1 = \frac{1}{2L}\|\nabla f(x)\|^2$

$Q2 = \frac{1}{2\mu}\|\nabla f(x)\|^2$

Claim: a) $f(x_{t+1}) \le f(x_t) - Q1$

b) $f(x^*) - f(x_t) \le -Q2$ ☒

$$f(x_{t+1}) - f^* \le f(x_t) - f^* - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

$$= f(x_t) - f^* - \frac{\mu}{L}\left(\frac{1}{2\mu}\|\nabla f(x_t)\|^2\right)$$

$$\le f(x_t) - f^* - \frac{\mu}{L}\left(f(x_t) - f^*\right)$$

$$\left(f(x_{t+1}) - f^*\right) \le \left(1 - \frac{\mu}{L}\right) f(x_t) - f^*$$

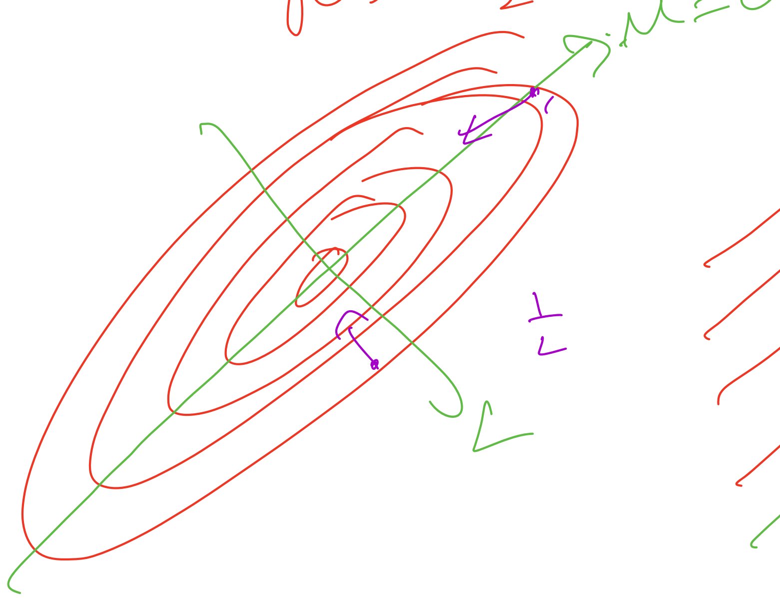
$$\le \left(1 - \frac{\mu}{L}\right)^t f(x_0) - f^*$$

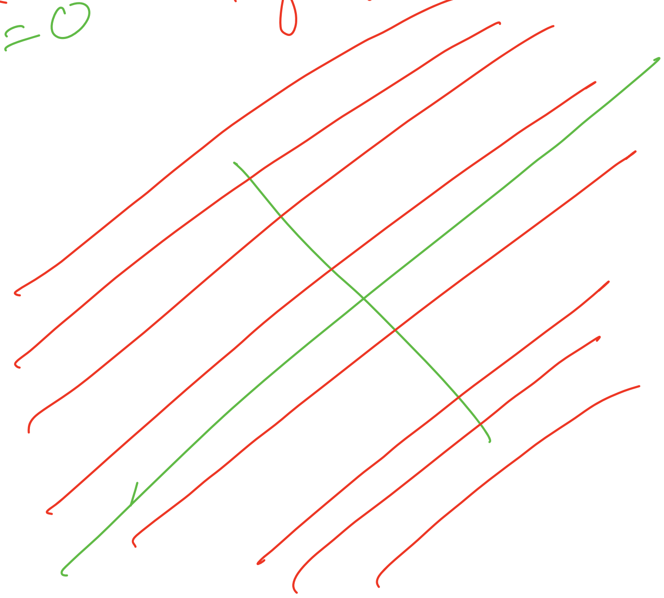
Viz

$f(x) = \frac{1}{2} x^T A x$ (A → p.s.d)

$\nabla^2 f(x) = A$

$\mu = 0$

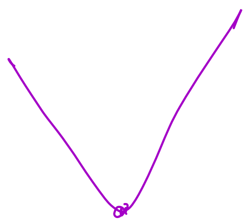




f is non-smooth, but cvx

$$\|x_{t+1} - x^*\|^2 = \|x_t - \gamma g_t - x^*\|^2$$

$$x_{t+1} = x_t - \gamma \underbrace{\nabla f(x_t)}_{= g_t}$$



$$= \|x_t - x^*\|^2 - 2\gamma \underbrace{g_t^\top (x_t - x^*)}_{\geq 0} + \gamma^2 \|g_t\|^2$$

Claim $\nabla f(x_t)^\top (x_t - x^*) \geq f(x_t) - f^*$

Proof $f(y) \geq f(x_t) + \nabla f(x_t)^\top (y - x_t)$

$y = x^*$

$$f^* \geq f(x_t) + \nabla f(x_t)^\top (x^* - x_t)$$

$$\Rightarrow \|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\gamma\left(f(x_t) - f^*\right) + \gamma^2 \underbrace{\|\nabla f(x_t)\|_2^2}_{\leq G^2}$$

$$\frac{1}{\gamma}\sum_t 2\gamma\left(f(x_t)-f^*\right) \le \frac{1}{\gamma}\left(\sum_t \|x_t - x^*\|^2 - \|x_{t+1}-x^*\|^2 + T\gamma^2 G^2\right)$$

$$= \frac{1}{\gamma}\left(\|x_0 - x^*\|^2 - \|x_{t+1} - x^*\|^2 + T\gamma^{\cancel{2}} G^2\right)$$

$$\le \frac{1}{\gamma}\|x_0 - x^*\|^2 + T\gamma G^2$$

$$\gamma = \sqrt{\frac{D^2}{TG^2}} \qquad \le \frac{D^2}{\gamma} + T\gamma G^2 = 2DG\sqrt{T}$$

$$\frac{\cancel{2}}{T}\sum_t f(x_t) - f^* \le \frac{2DG\sqrt{T}}{T} = \frac{2DG}{\sqrt{T}}$$

$$\bar{x}_T = \frac{1}{T}\sum_t x_t$$

$$f(\bar{x}_T) - f^* \le \frac{1}{T}\sum_t f(x_t) - f^* \le \frac{DG}{\sqrt{T}}$$

<u>Assumption</u> $\|x_0 - x^*\|^2 \le D$

$\|\nabla f(x_t)\|^2 \le G^2$

$$\|\nabla f(x)\|_2 \leq G \quad \forall x$$

$$|f(y) - f(x)| \leq G\|y-x\|_2$$

G-Lip

Recap

① f is L-smooth + μ-S.C

$$\Rightarrow f(x_t) - f^* \leq O\left[\left(1-\frac{\mu}{L}\right)^t\right]$$

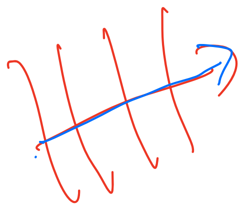
② f is G-Lip + cvx +

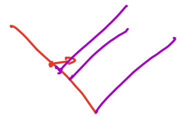
$$\|x_0 - x^*\|_2 \leq D \Rightarrow$$

$$f(x_T) - f^* \leq O\left(\frac{1}{\sqrt{T}}\right)$$

③ f is L-smooth + cvx $\Rightarrow$

$$f(x_t) - f^* \leq O\left(\frac{1}{T}\right)$$





Chapter 2

Gradient Descent

$$x^{t+1} = x^t - \nabla^2 f(x^t)^{-1} \nabla f(x^t)$$

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

$\nabla f(x^t)$

x^t

$$f(y) \approx \boxed{f(x^t) + \nabla f(x^t)^T (y - x^t)}$$

$-\nabla f(x^t)$

$$x^{t+1} = x^t - \gamma \nabla f(x^t)$$

$$f(y) \approx f(x^t) + \nabla f(x^t)^T (y - x^t) + \frac{1}{2}(y - x^t)^T \nabla^2 f(x^t)(y - x^t)$$

The Algorithm

$f(y) = f(x^*) + f'(x^*)(y - x^*) + \frac{1}{2} f''(x^*) \cdot (y - x^*)^2$

Get near to a minimum $\mathbf{x}^\star$ / close to the optimal value $f(\mathbf{x}^\star)$?

(Assumptions: $f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, has a global minimum $\mathbf{x}^\star$)

Goal: Find $\mathbf{x} \in \mathbb{R}^d$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \varepsilon.$$

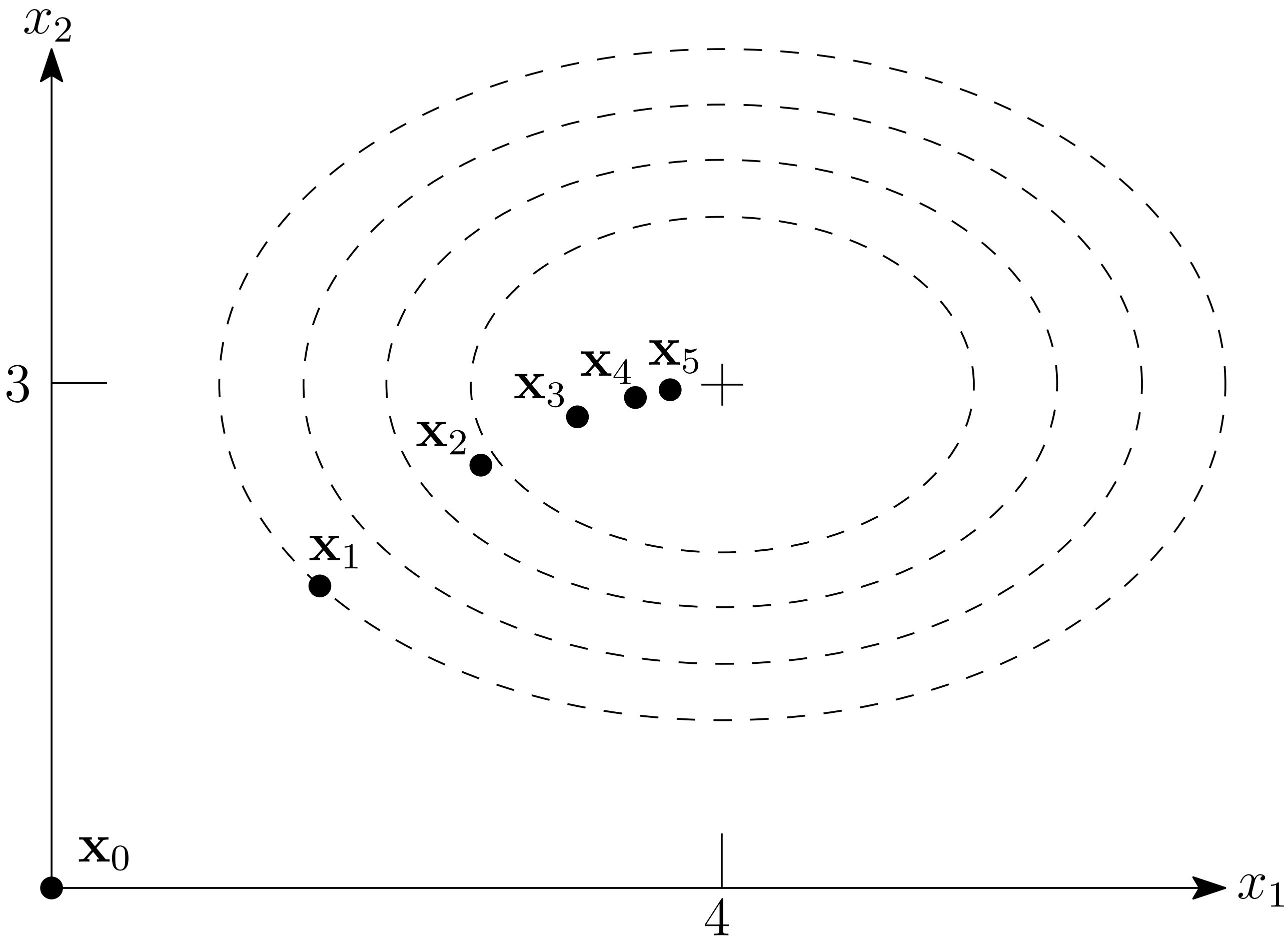
Note that there can be several global minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

Iterative Algorithm: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

for **timesteps** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.

Example



$$f(x_1, x_2) := 2(x_1 - 4)^2 + 3(x_2 - 3)^2, \mathbf{x}_0 := (0, 0), \gamma := 0.1$$

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

$x_{t+1} = x_t - \gamma g_t$

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) =$$

$$\frac{1}{2\gamma} \cdot \|x_{t+1} - x^*\|^2 = \frac{1}{2\gamma}\left(\|x_t - \gamma g_t - x^*\|^2\right)$$

$$= \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - 2\gamma\, g_t^\top(x_t - x^*) + \gamma^2\|g_t\|^2\right)$$

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to rewrite

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \quad =$$

Vanilla analysis

$g_t = \nabla f(x_t)$

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to rewrite

$$\begin{aligned}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) &= \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)\\ &= \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\cancel{\|\mathbf{x}_t - \mathbf{x}^\star\|^2} - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)\end{aligned}$$

$$\left(\cancel{\|x_{t-1} - x^*\|^2} - \cancel{\|x_t - x^*\|^2}\right)$$

$$\|x_{t-2} - x^*\|^2 - \cancel{\|x_{t-1} - x^*\|^2}$$

- Sum this up over the first T iterations:

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \;=$$

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to rewrite

$$\begin{aligned}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) &= \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right)\\ &= \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right)\end{aligned}$$

- Sum this up over the first T iterations:

$$\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0-\mathbf{x}^\star\|^2 - \|\mathbf{x}_T-\mathbf{x}^\star\|^2\right)$$

Vanilla analysis II

Use first-order characterization of convexity: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y}$

- with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

Vanilla analysis II

Use first-order characterization of convexity: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y}$

- with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$$

giving

$$\sum_{t=0}^{T-1} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq$$

Vanilla analysis II

Use first-order characterization of convexity: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y}$

- with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)$$

giving

$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{\gamma}{2} \underbrace{\sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2}_{\leq TB^2} + \frac{1}{2\gamma} \underbrace{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}_{\leq R^2},$$

an upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ over the steps

- last iterate is not necessarily the best one
- stepsize is crucial

$$\leq \frac{\gamma}{2} TB^2 + \frac{R^2}{2\gamma}$$

$$\gamma = \frac{R}{B\sqrt{T}} \Rightarrow \leq BR\sqrt{T}$$

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Assume that all gradients of f are bounded in norm.

- Equivalent to f being Lipschitz (Theorem **??**; **Exercise ??**).
- Rules out many interesting functions (for example, the "supermodel" $f(x) = x^2$)

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$. Choosing the stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

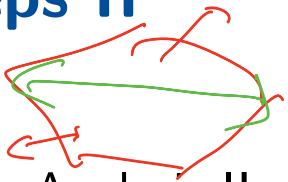
gradient descent yields

$$\boxed{\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.}$$

f is B-Lip

$f(x) = \|x\|$, $\frac{1}{2}x^T A x - b^T x$

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps II



C

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

$\min_{x \in C} f(x)$

Proof.

- Plug $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\mathbf{g}_t\| \leq B$ into Vanilla Analysis II:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \leq \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2.$$

- choose γ such that

$$q(\gamma) = \frac{\gamma}{2} B^2 T + \frac{R^2}{2\gamma}$$

  is minimized.
- Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{T}}$, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$.
- Dividing by T, the result follows. $\square$

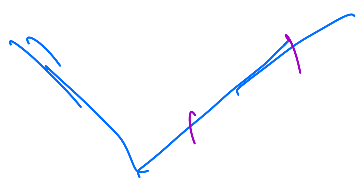
$f(x) = \|x\|_1$ is cvx

Limitations

$\|x\|_2 \le 1$

① How to be sure $\|x_0 - x^*\| \le R$ w/o knowing x^*? Need to constrain x.

② gradient may be undefined for a non-smooth f



③ How to know $\|\nabla f(x)\|_2 \le B \ \forall x$?

$f(x) = \|x\|_1$, Eg 1

$$|f(y) - f(x)| \le B\|y-x\|_2$$

$$(\|y\|_1 - \|x\|_1) \le B \|y-x\|_2$$

$$\|y\|_1 - \|x\|_1 \le \|y-x\|_1 \le \sqrt{d}\|y-x\|_2^2$$

$$\|z\|_1 \le \sqrt{d}\|z\|_2$$

Eg 2

$f(x) = \frac{1}{2}x^T A x$

$\|x\|_2 \le 1$

$\|\nabla f(x)\|_2 = \|Ax\|_2$

$\le \|A\|_2 \|x\|_2 \le \|A\|_2$

$f(x) = x^2$ $|x| \le 1$

$\nabla f(x) = 2x$

≤ 2

ToDo

1. Run GD on constrained domains.

2. Make GD work when func is not diff.

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps III

$$T \geq \frac{R^2 B^2}{\varepsilon^2} \quad \Rightarrow \quad \text{average error } \leq \frac{RB}{\sqrt{T}} \leq \varepsilon.$$

Advantages:

- dimension-independent (no d in the bound)!
- holds for both average, or best iterate

In Practice:
What if we don't know R and B? $\rightarrow$ **Exercise ??** (having to know R can't be avoided)

Smooth functions

“Not too curved”

Definition

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) over X if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

$$\forall y, x, \quad (\nabla f(y) - \nabla f(x))^T (y - x) \leq L\|y - x\|_2$$

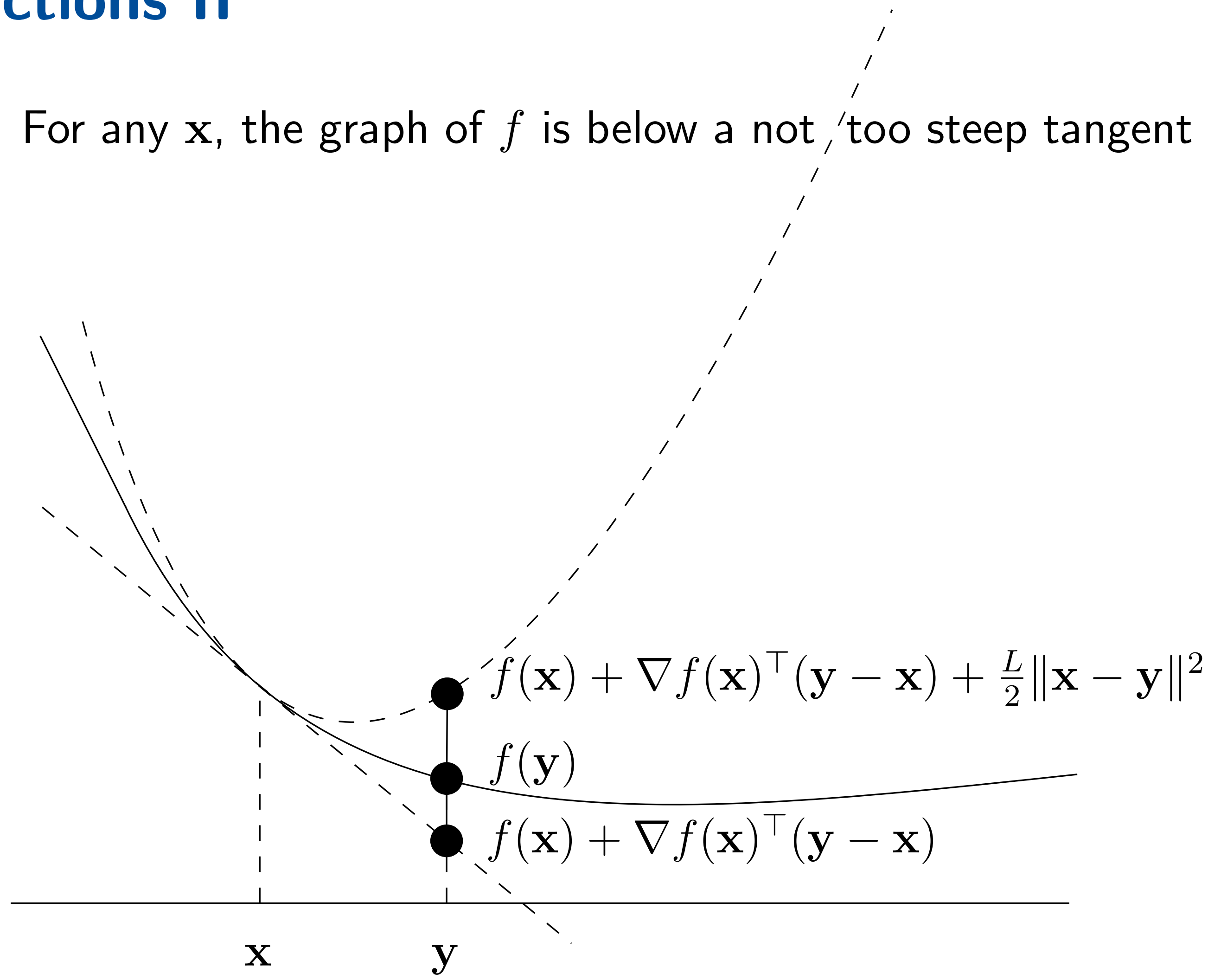
f smooth $:\Leftrightarrow$ f smooth over $\mathbb{R}^d$.

Definition does not require convexity (useful later)

$$\|\nabla^2 f(x)\|_2 \leq L \quad \forall x$$

Smooth functions II

Smoothness: For any $\mathbf{x}$, the graph of f is below a not too steep tangent paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



Smooth functions III

- In general: quadratic functions are smooth (**Exercise ??**).
- Operations that preserve smoothness (the same that preserve convexity):

Lemma (Exercise **??**)

(i) *Let* $f_1, f_2, \ldots, f_m$ *be functions that are smooth with parameters* $L_1, L_2, \ldots, L_m$*, and let* $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$*. Then the function* $f := \sum_{i=1}^m \lambda_i f_i$ *is smooth with parameter* $\sum_{i=1}^m \lambda_i L_i$*.*

(ii) *Let* f *be smooth with parameter* L*, and let* $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$*, for* $A \in \mathbb{R}^{d \times m}$ *and* $\mathbf{b} \in \mathbb{R}^d$*. Then the function* $f \circ g$ *is smooth with parameter* $L\|A\|^2$*, where is* $\|A\|$ *is the* **spectral norm** *of* A *(Definition* **??***).*

Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f (in the convex case).

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.

(i) *f is smooth with parameter L.*

(ii) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Proof in lecture slides of L. Vandenberghe, `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

Sufficient decrease

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize

$$\gamma := \frac{1}{L},$$

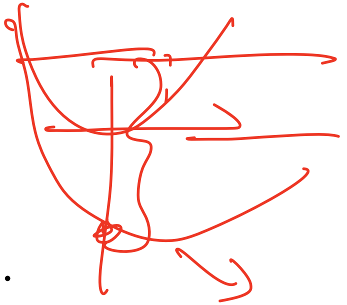
gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

Remark

More specifically, this already holds if f is smooth with parameter L over the line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.

Sufficient decrease II



$$\frac{1}{2L}\|\nabla f(x_t)\|^2$$

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

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

Proof.

Use smoothness and definition of gradient descent ($\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$):

$$f(\mathbf{x}_{t+1}) \leq f(x_t) + \nabla f(x_t)^T \underbrace{(x_{t+1} - x_t)}_{-\frac{1}{L}\nabla f(x_t)} + \frac{L}{2}\|\underbrace{x_{t+1} - x_t}_{-\frac{1}{L}\nabla f(x_t)}\|^2$$

$$= f(x_t) - \frac{1}{L}\|\nabla f(x_t)\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|^2$$

$$= f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

Sufficient decrease II

$\Longrightarrow$ $\|\nabla f(x_t)\|_2^2 \le 2L\left(f(x_t) - f(x_{t+1})\right)$

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

Proof.

Use smoothness and definition of gradient descent ($\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$):

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

□

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps II

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Vanilla Analysis II:

$$\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \leq \qquad \frac{1}{\cancel{2L}}\sum_{t=0}^{T-1}\cancel{2L}\left(\cancel{f(x_t)} - f(x_{t+1})\right)$$

$$f(x_{t+1}) - \cancel{f(x_t)}$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps II

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Vanilla Analysis II:

$$\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T).$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps III

Putting it together with $\gamma = 1/L$:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{1}{2L}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

$$\le f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

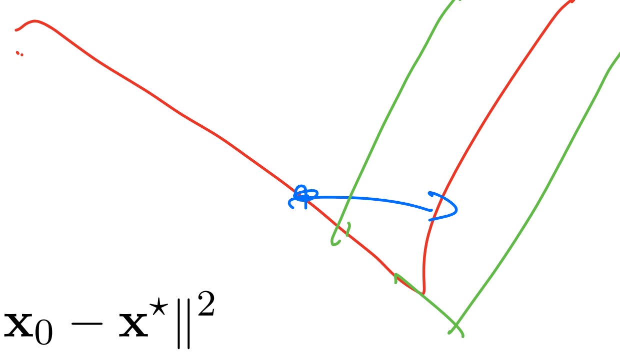
$\geq f(x^*)$

$\leq f(x_0) - f(x^*)$

Rewriting:

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{T}$$

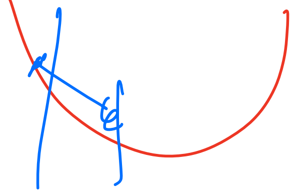
Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps III



Putting it together with $\gamma = 1/L$:

$$\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

$$\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$



Rewriting:

$f(x_t) - f(x^*) \leq \frac{1}{T}$

$$\boxed{\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^\star))} \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

$\frac{1}{T}$

$\boxed{f(x_{t+1}) \leq f(x_t)}$

$- \frac{1}{2L}\|\nabla f(x_t)\|^2$

As last iterate is the best (sufficient decrease!):

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|x_0 - x^*\|^2 \leq \frac{LR^2}{2T}$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps III

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}
\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\
&\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.
\end{aligned}$$

Rewriting:

$$\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

As last iterate is the best (sufficient decrease!):

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{T}\left(\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^\star))\right) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

□

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps IV

$R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.

$$T \geq \frac{R^2 L}{2\varepsilon} \quad \Rightarrow \quad \text{error} \leq \frac{L}{2T} R^2 \leq \varepsilon.$$

- $50 \cdot R^2 L$ iterations for error 0.01 ...
- ... as opposed to $10,000 \cdot R^2 B^2$ in the Lipschitz case

In Practice:
What if we don't know the smoothness parameter L?

$\rightarrow$ **Exercise ??**

Lemma 2.4

$$f(x) = x^T Q x + b^T x + c$$

$Q \to$ sym, psd

is $2\|Q\|_2$ smooth

Proof $\to$

$$D^2 f(x) = 2Q$$

$$\|D^2 f(x)\|_2 \le \|2Q\|_2 = 2\|Q\|_2$$

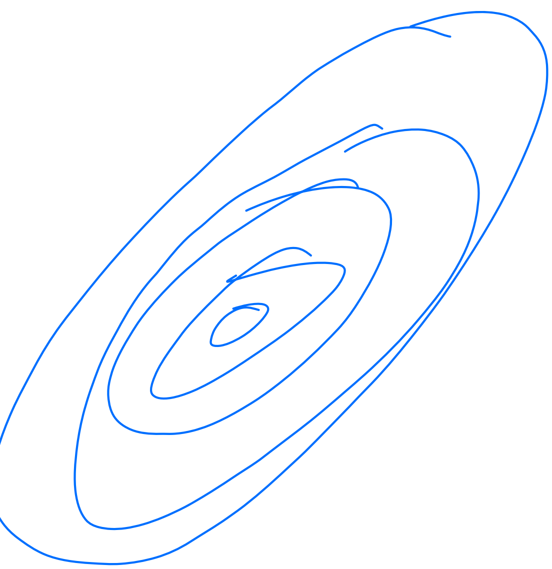
$\to$

$$\|Df(y) - Df(x)\|_2 =$$

$$\|2Qy - \cancel{b} - (2Qx - \cancel{b})\|_2$$

$$= \|2Q(y-x)\|_2$$

$$\le 2\|Q\|_2 \|y-x\|_2$$



Prove

i) $\{f_i \text{ is } L_i\text{-smooth}\}$

$$g(x) = \sum_{i=1}^{m} \lambda_i f_i \text{ is } \left(\sum_{i=1}^{m} \lambda_i L_i\right)\text{-smooth}$$

$$D^2 g(x) = \left\| \sum_{i=1}^{m} \lambda_i D^2 f(x) \right\|_2$$

$$\leq \sum_{i=1}^{m} \lambda_i \, \| D^2 f(x) \|_2$$

$$\leq \sum_{i=1}^{m} \lambda_i L_i$$

ii) $g(x) = f(Ax+b)$ is $L\|A\|_2^2$ smooth

$$Dg(x) = A^T Df(Ax+b)$$

$$\| D^2 g(x) \| = \| A^T D^2 f(Ax+b) A \|_2$$

$$= \max_v \frac{\| A^T \boxed{D^2 f(Ax+b) A v} \|_2}{\|v\|_2}$$

$$\le \max_{v} \frac{\|A\|_2 \, \|D^2 f(Ax+b)(Av)\|_2}{\|v\|_2}$$

$$\le \max_{v} \frac{\|A\|_2 \, \|D^2 f(Ax+b)\|_2 \, \|Av\|_2}{\|v\|_2}$$

$$\le \max_{v} \frac{\|A\|_2^2 \, \|D^2 f(Ax+b)\| \, \cancel{\|v\|_2}}{\cancel{\|v\|_2}}$$

$$= \|A\|_2^2 \, \|D^2 f(Ax+b)\|_2$$

Ex

$f(x) = x^4$ $\qquad x \in (-a, a)$

$D^2 f(x) = 12x^2 \le 12a^2 \quad \forall x \in (-a, a)$

$\|Df(x)\| = |4x^3| \le 4a^3$