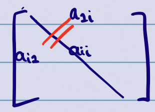
$A = A^T$ "symmetric matrix"



Claim: $A^T \cdot A$ is a symmetric matrix.

Definition: λ is an eigenvalue for M (square) such that there exists $v \in \mathbb{R}^n$ such that $Mv = \lambda v$.

Example: v is an eigenvector for I.

$$\left(I = \begin{bmatrix}1 & & 0\\ & \ddots & \\ 0 & & 1\end{bmatrix}\right). \qquad Iv = v.$$

1 is an eigenvalue for I.

Example: $M = \begin{bmatrix}2&1\\1&2\end{bmatrix}$ $\quad M - I\lambda = \begin{bmatrix}2-\lambda & 1\\ 1 & 2-\lambda\end{bmatrix}$

Solve for v, λ such that

$$Mv = \lambda v.$$
$$-\lambda v \quad -\lambda v$$
$$(M - I\lambda)v = Mv - \lambda v = 0$$

$M - Iv$ is not invertible. (v can be rescaled)

$\Rightarrow \det(M - I\lambda) = 0$.

$$M - I\lambda = \begin{bmatrix}2-\lambda & 1\\ 1 & 2-\lambda\end{bmatrix}$$

$$\det(M - I\lambda) = (2-\lambda)^2 - 1 = 0.$$
$$\lambda^2 - 4\lambda + 3 = 0$$
$$(\lambda - 3)(\lambda - 1) = 0$$

$\Rightarrow \lambda = 3$; $\lambda = 1$ are eigenvalues

$\downarrow \qquad \downarrow$

$\begin{pmatrix}1\\1\end{pmatrix} \quad \begin{pmatrix}1\\-1\end{pmatrix}$ associated eigenvectors

Fact: If M is symmetric and its entries are real, then its eigenvalues are real and its eigenvectors are orthonormal. (when ℓ_2-normed)

($u_1, \dots, u_d$ are eigenvectors of $M \in \mathbb{R}^{d\times d}$)

$\Rightarrow i, j \ (i \neq j)$: $\langle u_i, u_j \rangle = 0$

i : $\|u_i\|_2 = 1$

Fact: If M ($\mathbb{R}^{d\times d}$) is symmetric and real-valued, then

$$M = \sum_{i=1}^{d} \lambda_i \underbrace{u_i u_i^T}_{\text{outer product}}$$

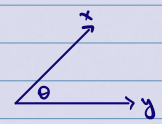
If $\lambda_i \geq 0$ for all i, then M is "symmetric positive semidefinite." $\equiv x^T M x \geq 0$ for any x.

$$x^T M x = x^T \left(\sum_{i=1}^{d} \lambda_i u_i u_i^T\right) x = \sum_{i=1}^{d} \lambda_i \underline{x^T u_i}\,\underline{u_i^T x}$$
$$= \sum_{i=1}^{d} \lambda_i \langle u_i, x\rangle^2 \geq 0.$$

Fact: $A \in \mathbb{R}^{n\times d}$. $\Rightarrow A^T A$ is symmetric positive semidefinite "psd"

Cauchy-Schwarz Inequality: $x, y \in \mathbb{R}^d$

$$\Rightarrow \langle x, y\rangle \leq \|x\|_2 \cdot \|y\|_2$$



$$\langle x, y\rangle = \|x\|_2 \cdot \|y\|_2 \cdot \underbrace{\cos\theta_{xy}}_{\leq 1}$$
$$\leq \|x\|_2 \cdot \|y\|_2$$

For matrices: $\|Mx\|_2 \leq \lambda_{max}(M) \cdot \|x\|_2$

"maximum eigenvalue"

(M is psd)

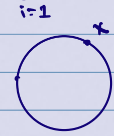
Proof: $M = \sum_{i=1}^{d} \lambda_i u_i u_i^T$

$$\|Mx\|_2^2 = x^T M^T M x = x^T\left(\sum_{i=1}^{d} \lambda_i^2 u_i u_i^T\right) x$$

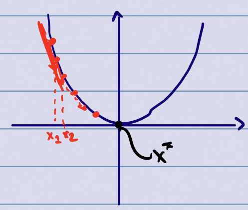
$(\|y\|_2^2 = \langle y, y\rangle)$ $\qquad = \sum_{i=1}^{d} \lambda_i^2 \langle u_i, x\rangle^2$

Assume $\|x\|_2 = 1$. $\qquad \leq \lambda_{max}^2$.

$$\Rightarrow \sum_{i=1}^{d} \langle u_i, x\rangle^2 = 1$$



Gradient descent: what was special about quadratics?



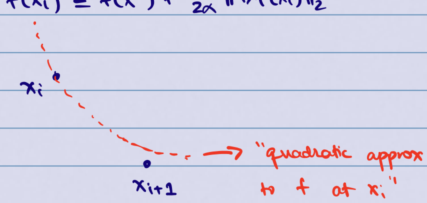
1. In each step, you make a large improvement $(f(x_t) - f(x^*) \downarrow)$
2. Small derivative $\Rightarrow$ small fn value

Definition: A function f is "nice" with respect to $x_1, \dots, x_T$ if:

1. $f(x_{i+1}) \leq f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{\beta}{2}\|x_{i+1} - x_i\|_2^2$
2. $f(x_i) \leq f(x^*) + \frac{1}{2\alpha}\|\nabla f(x_i)\|_2^2$

any quadratic satisfies these.

(1)



(2) $f(x_i) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x_i)\|_2^2$

(suppose $\|\nabla f(x_i)\|_2^2 \leq 2\alpha \cdot \varepsilon$)

$$\Rightarrow f(x_i) - f(x^*) \leq \frac{1}{2\alpha} \cdot \|\nabla f(x_i)\|_2^2 \leq \varepsilon$$

Theorem: If f was "nice" with respect to $x_1, \dots, x_T$, and $x_1, \dots, x_T$ are the iterates from gradient descent, then

$$f(x_{i+1}) - f(x^*) \leq [f(x_i) - f(x^*)] \cdot \left[1 - \frac{\alpha}{\beta}\right].$$

Proof: A function f is "nice" with respect to $x_1, \dots, x_T$ if:

1. $f(x_{i+1}) \le f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{\beta}{2} \|x_{i+1} - x_i\|_2^2$
2. $f(x_i) \le f(x^*) + \frac{1}{2\alpha} \|\nabla f(x_i)\|_2^2$

GD update rule: $x_{i+1} = x_i - \eta \cdot \nabla f(x_i)$

(1) implies: $f(x_{i+1}) \le f(x_i) + \langle \nabla f(x_i), -\eta \cdot \nabla f(x_i) \rangle + \frac{\beta}{2} \cdot \|-\eta \nabla f(x_i)\|_2^2$

$$= f(x_i) - \eta \cdot \|\nabla f(x_i)\|_2^2 + \frac{\beta}{2} \cdot \eta^2 \|\nabla f(x_i)\|_2^2$$

$$= f(x_i) - \|\nabla f(x_i)\|_2^2 \left(\eta - \frac{\beta}{2}\eta^2\right)$$

(2) $\|\nabla f(x_i)\|_2^2 \ge 2\alpha(f(x_i) - f(x^*))$

$$\le f(x_i) - 2\alpha[f(x_i) - f(x^*)]\left[\eta - \frac{\beta}{2}\eta^2\right]$$

$$f(x_{i+1}) - f(x^*) \le f(x_i) - f(x^*) - 2\alpha[f(x_i) - f(x^*)]\left[\eta - \frac{\beta}{2}\eta^2\right]$$

$$\delta_i := f(x_i) - f(x^*)$$

$$\delta_{i+1} \le \delta_i - 2\alpha \cdot \delta_i \left(\eta - \frac{\beta}{2}\eta^2\right) = \delta_i\left[1 - 2\alpha\left(\eta - \frac{\beta}{2}\eta^2\right)\right]$$

It is enough to minimize $1 - 2\alpha\left(\eta - \frac{\beta}{2}\eta^2\right)$ over η.

$\eta = \frac{1}{\beta}$ is enough.

$$\Longrightarrow 1 - 2\alpha\left(\eta - \frac{\beta}{2}\eta^2\right) = 1 - 2\alpha\left(\frac{1}{\beta} - \frac{\beta}{2} \cdot \frac{1}{\beta^2}\right) = 1 - \frac{\alpha}{\beta}.$$

$$\Longrightarrow \delta_{i+1} \le \delta_i \left(1 - \frac{\alpha}{\beta}\right) \quad \square.$$

Theorem: If f is convex and iterates $x_1, \dots, x_T$ are s.t.

$f(x_{i+1}) \le f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{\beta}{2} \|x_{i+1} - x_i\|_2^2$

and, if you know x_0 such that for all x with $f(x) \le f(x_0)$, you had

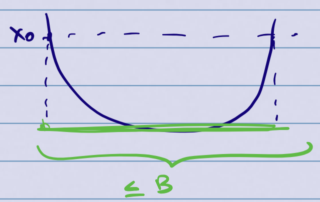
$$\|x_0 - x\|_2 \le B,$$

then with

$$T = O\left(\beta \cdot \frac{B^2}{\varepsilon} \cdot \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)\right)$$

we have

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



Proof: $f(x_{i+1}) \le f(x_i) - \|\nabla f(x_i)\|_2^2 \left(\eta - \frac{\beta}{2}\eta^2\right)$

$$\boxed{\text{Run GD on } \quad g(x) := f(x) + \frac{\varepsilon}{2B^2} \cdot \|x_0 - x\|_2^2}$$

Plan: approximately minimizing $g \Longrightarrow$ approximately minimizing f.

$y^* := \operatorname{argmin}_{y'} g(y')$

y: $\quad g(y) \le g(y^*) + \frac{\varepsilon}{2}$.

$$f(y) \le g(y) \le g(y^*) + \frac{\varepsilon}{2} = f(y^*) + \frac{\varepsilon}{2B^2} \cdot \|x_0 - y^*\|_2^2 + \frac{\varepsilon}{2}$$

$$\le f(y^*) + \frac{\varepsilon}{2B^2} \cdot B^2 + \frac{\varepsilon}{2} = f(y^*) + \varepsilon \quad (1)$$

$$\underbrace{f(y^*) \le f(y^*) + \frac{\varepsilon}{2B^2} \|x_0 - y^*\|_2^2}_{g(y^*)} \le f(x^*) + \frac{\varepsilon}{2B^2} \|x_0 - x^*\|_2^2$$

$$\le f(x^*) + \frac{\varepsilon}{2B^2} \cdot B^2 = f(x^*) + \varepsilon/2 \quad (2)$$

Plug (2) into (1).

$$f(y) \underset{(1)}{\le} f(y^*) + \varepsilon \underset{(2)}{\le} (f(x^*) + \varepsilon) + \varepsilon/2 = f(x^*) + \frac{3\varepsilon}{2}$$

Next: Approximately minimize g.

Remember: If g was "nice" with respect to $x_1, \dots, x_T$, and $x_1, \dots, x_T$ are the iterates from gradient descent, then

$$g(x_{i+1}) - g(x^*) \le [g(x_i) - g(x^*)] \cdot \left[1 - \frac{\alpha'}{\beta'}\right] \le [g(x_i) - g(x^*)] \cdot \exp\left(-\frac{\alpha'}{\beta'}\right)$$

$1 - x \le e^{-x}$

We can choose $\alpha' = \frac{\varepsilon}{B^2}$

$$\beta' = \beta + \frac{\varepsilon}{B^2}.$$

$\Longrightarrow$ After $\frac{\beta'}{\alpha'} \log\left(\frac{g(x_0) - g(x^*)}{\varepsilon}\right)$ steps, we had

$$g(x_{i+1}) - g(x^*) \le \varepsilon.$$

$$\frac{\beta'}{\alpha'} = 1 + \beta \cdot \frac{B^2}{\varepsilon}$$

$\Longrightarrow$ After $O\left(\left(1 + \beta \cdot \frac{B^2}{\varepsilon}\right) \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)\right)$ steps of gradient descent on g, you can ensure

$$f(x_T) \le f(x^*) + \varepsilon. \quad \square$$

Remark: Running GD directly on f gets the same rate.

Remark:

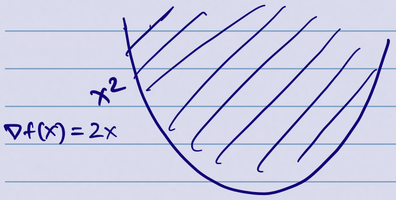
1. $f(x_{i+1}) \le f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{\beta}{2} \|x_{i+1} - x_i\|_2^2$
2. $f(x_i) \le f(x^*) + \frac{1}{2\alpha} \|\nabla f(x_i)\|_2^2$

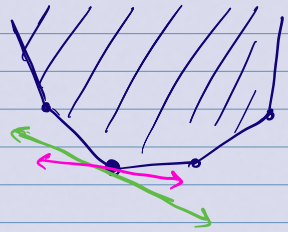
Definition: f is β-smooth if for all x: for all y: $|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \le \frac{\beta}{2} \|x - y\|_2^2$ (Check that β-smooth $\Longrightarrow$ quadratic UB)

Definition: f is α-PL (Polyak–Łojasiewicz) if for all x:

$$f(x) \le f(x^*) + \frac{1}{2\alpha} \|\nabla f(x)\|_2^2.$$

(Strong convexity $\Longrightarrow$ PL)





Two convex sets

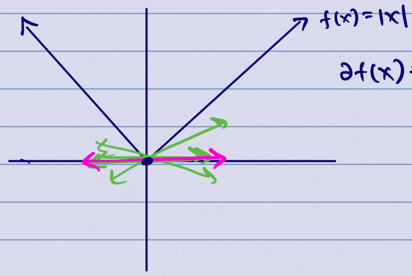
<u>Definition</u>: A function is convex if for every x, there is some $\nabla f(x)$ such that for all y

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

The set of all valid $\nabla f(x)$ is called the subgradient and it is denoted as $\underline{\partial f(x)}$.

$$\partial f(x) := \{ \nabla f(x) \mid \text{for all } y,\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \}$$

$f(x) = |x|$.



$$\partial f(x) = \begin{cases} \operatorname{sign}(x) & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

<u>Fact</u>: At x^*, $0 \in \partial f(x)$.

<u>Exercise</u>: Calculate $\partial f(x)$ when $f(x) = \|x\|_1 = \sum_{i=1}^{d} |x_i|$.

LASSO: $f_\lambda(w) = \|Xw - y\|_2^2 + \alpha \cdot \|w\|_1$

$X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$. $X = \begin{bmatrix} — x_1 — \\ \vdots \\ — x_n — \end{bmatrix}$

$n \ll d$.

w only depends on very few features (sparse).