Lecture 14:

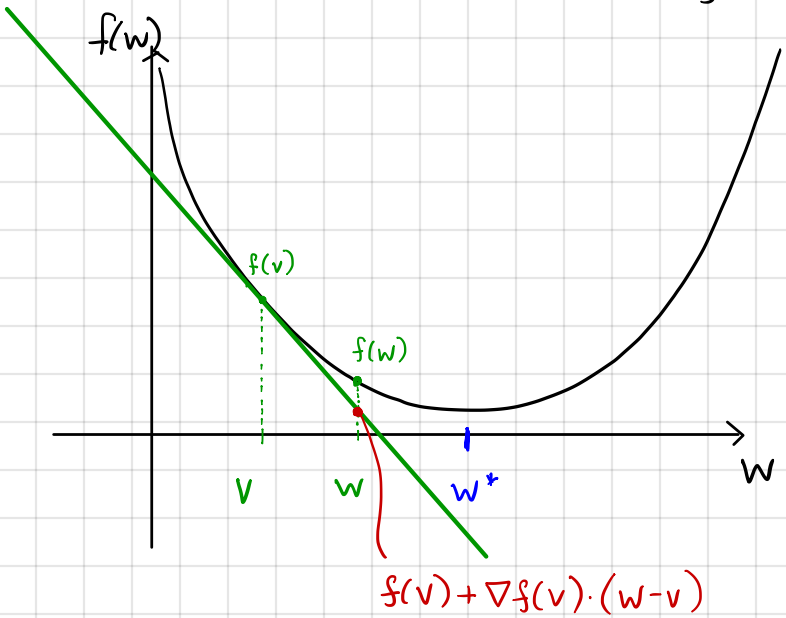
Basic Convex Optimization

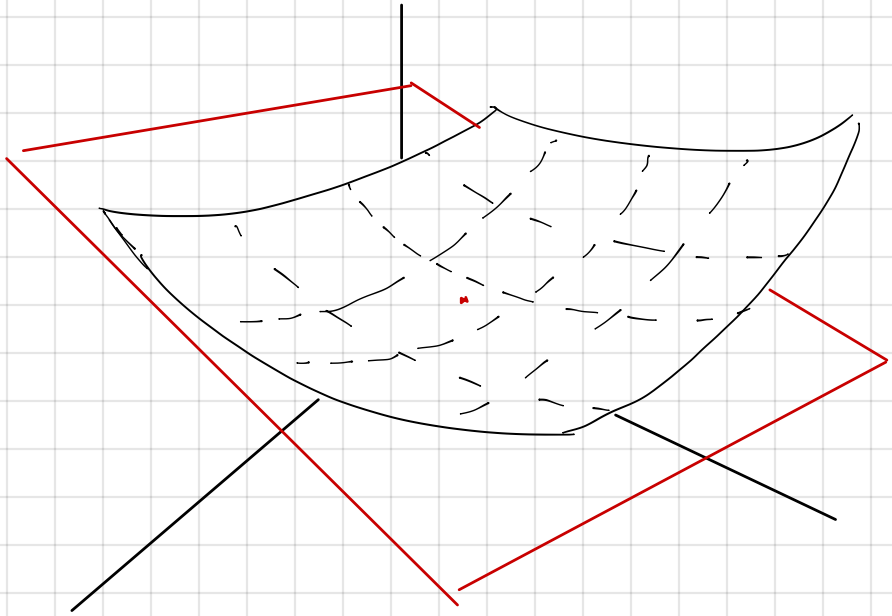
Basic convex optimization

Goal: find $\underline{w}^* = \arg\min_{\underline{w}} f(\underline{w})$ when f is a convex function.

A function is convex if $f(\underline{w}) \geq f(\underline{v}) + \nabla f(\underline{v})^T (\underline{w} - \underline{v})$

- i.e. if it's $\geq$ all its tangents





Ex. $f(\underline{w}) = \|\underline{y} - X\underline{w}\|_2^2$. We know $\underline{w}^* = (X^T X)^{-1} X^T \underline{y}$

Gradient descent finds this point iteratively.

- avoids computing matrix inverse
- generalizes to many other problems.

Gradient:

if $f(\underline{w}) = \underline{y}^T\underline{y} - 2\underline{w}^T X^T \underline{y} + \underline{w}^T X^T X \underline{w}$, then $\nabla_{\underline{w}} f = 0 - 2X^T\underline{y} + 2X^T X\underline{w}$

Gradient descent starts with initial guess $\underline{w}^{(1)}$, and then repeatedly takes steps in the direction of the negative gradient.

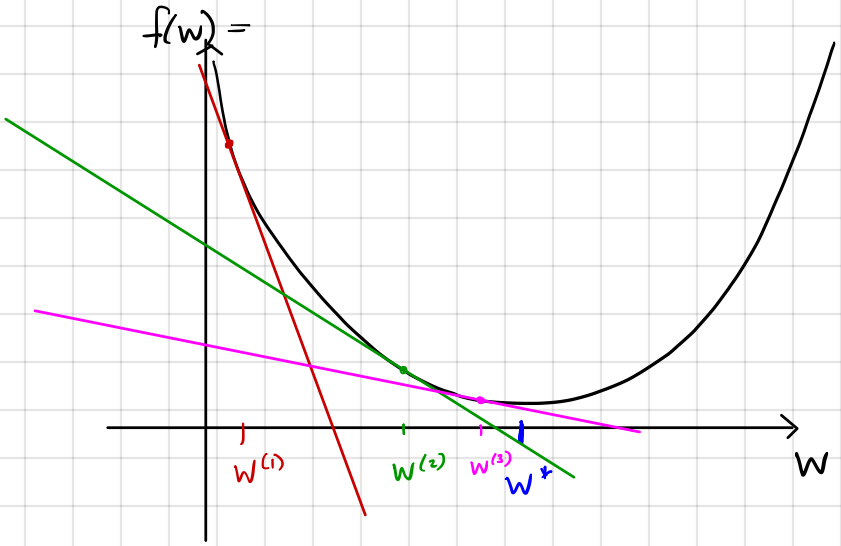
for $k = 1, 2, 3, \ldots$

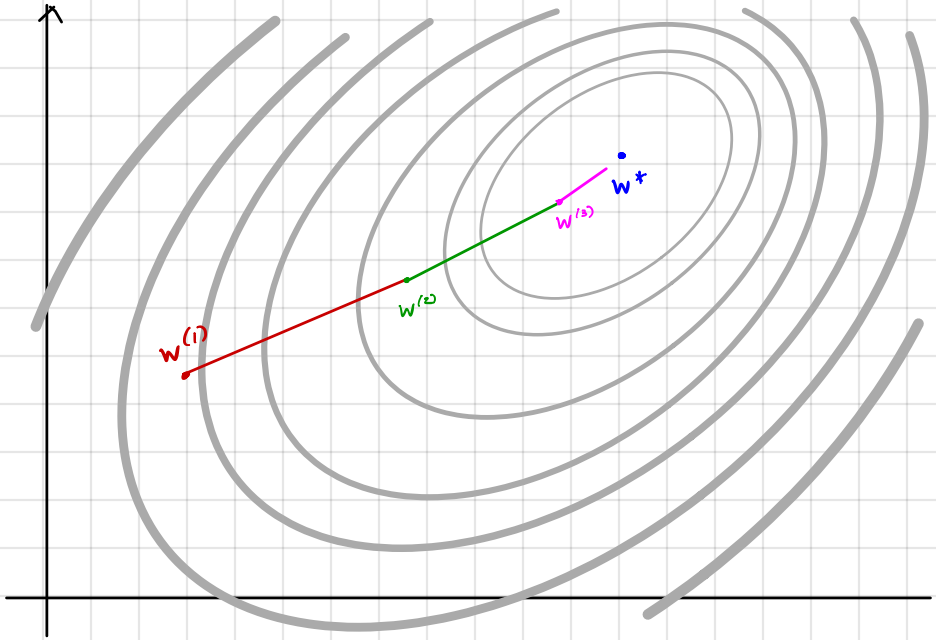
$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - 2\tau(X^T X \underline{w}^{(k)} - X^T\underline{y})$$

$$= \underline{w}^{(k)} - 2\tau X^T(X\underline{w}^{(k)} - \underline{y})$$

if $\| \underline{w}^{(k+1)} - \underline{w}^{(k)} \|_2 < \varepsilon$, then BREAK

$\tau > 0$ is step size





More generally:

want to minimize $f(\underline{w})$
initialize with $\underline{w}^{(1)}$
for $k = 1, 2, 3, \ldots$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau \nabla_{\underline{w}} f \Big|_{\underline{w} = \underline{w}^{(k)}}$$

if $\| \underline{w}^{(k)} - \underline{w}^{(k+1)} \| < \varepsilon$, then BREAK

Does this work?

Convergence for $f(\underline{w}) = \|X\underline{w} - y\|_2^2$

want $\|Xw^{(k+1)} - y\|_2^2 < \|Xw^{(k)} - y\|_2^2$

recall $w^{(k+1)} = w^{(k)} - 2\tau X^T(Xw^{(k)} - y)$

$$\Rightarrow \|Xw^{(k+1)} - y\|_2^2 = \|X\left(w^{(k)} - 2\tau X^T(Xw^{(k)} - y)\right) - y\|_2^2$$

$$= \|X\underline{w}^{(k)} - y - 2\tau XX^T(Xw^{(k)} - y)\|_2^2$$

$$= \|Xw^{(k)} - y\|_2^2 - 4\tau\underbrace{(Xw^{(k)} - y)^T\left(XX^T(Xw^{(k)} - y)\right)}_{= \|X^T(Xw^{(k)} - y)\|_2^2} + 4\tau^2\underbrace{\|XX^T(Xw^{(k)} - y)\|_2^2}_{\leq \|X\|_{op}^2 \|X^T(Xw^{(k)} - y)\|_2^2}$$

$$\|Xw^{(k+1)} - y\|_2^2 \leq \|Xw^{(k)} - y\|_2^2 + 4\tau\left(\tau\|X\|_{op}^2\|X^T(Xw^{(k)} - y)\|_2^2 - \|X^T(Xw^{(k)} - y)\|_2^2\right)$$

$$= \|Xw^{(k)} - y\|_2^2 + 4\tau\|X^T(Xw^{(k)} - y)\|_2^2\left(\tau\|X\|_{op}^2 - 1\right)$$

$\Rightarrow$ if $\tau\|X\|_{op}^2 - 1 < 0$ $\left(\tau < 1/\|X\|_{op}^2\right)$, then $\|Xw^{(k+1)} - y\|_2^2 < \|Xw^{(k)} - y\|_2^2$

if $\underline{w}^{(1)} = 0$ and $\tau < 1/\|X\|_{op}^2$, then

$$\underline{w}^{(k)} \longrightarrow (X^TX)^{-1}X^Ty \quad \text{as} \quad k \longrightarrow \infty$$

Another perspective:

$$w^{(k+1)} = w^{(k)} + \tau(X^T y - X^T X w^{(k)})$$
$$= w^{(k)} + \tau X^T X\left[(X^T X)^{-1} X^T y - w^{(k)}\right]$$
$$= w^{(k)} - \tau X^T X (w^{(k)} - w^*)$$

subtract w^* from both sides:

$$v^{(k+1)} = w^{(k+1)} - w^* = w^{(k)} - w^* - \tau X^T X (w^{(k)} - w^*)$$
$$= v^{(k)} - \tau X^T X v^{(k)}$$

$$\Rightarrow v^{(k+1)} = (I - \tau X^T X) v^{(k-1)}$$
$$= (I - \tau X^T X)^{k-1} v^{(1)}$$

for this sequence to go to zero, need all the singular values of $(I - \tau X^T X)$ to be less than 1 in magnitude

this happens if

$$\tau < \frac{1}{\sigma_{max}^2(X^T X)}$$

$$I - \tau X^T X$$
$$= I - \tau U \Sigma V^T V \Sigma^T U^T$$
$$= U(I - \tau \Sigma \Sigma^T) U^T$$

$\Rightarrow$ singular values are $1 - \tau\sigma_1^2, 1 - \tau\sigma_2^2, \ldots$

$\Rightarrow$ need $\max_i |1 - \tau\sigma_i^2| < 1$

$\Rightarrow$ need $|1 - \tau\sigma_{max}^2| < 1$

$\Rightarrow$ need $\tau < 1/\sigma_{max}^2$

Where do gradient descent updates come from?

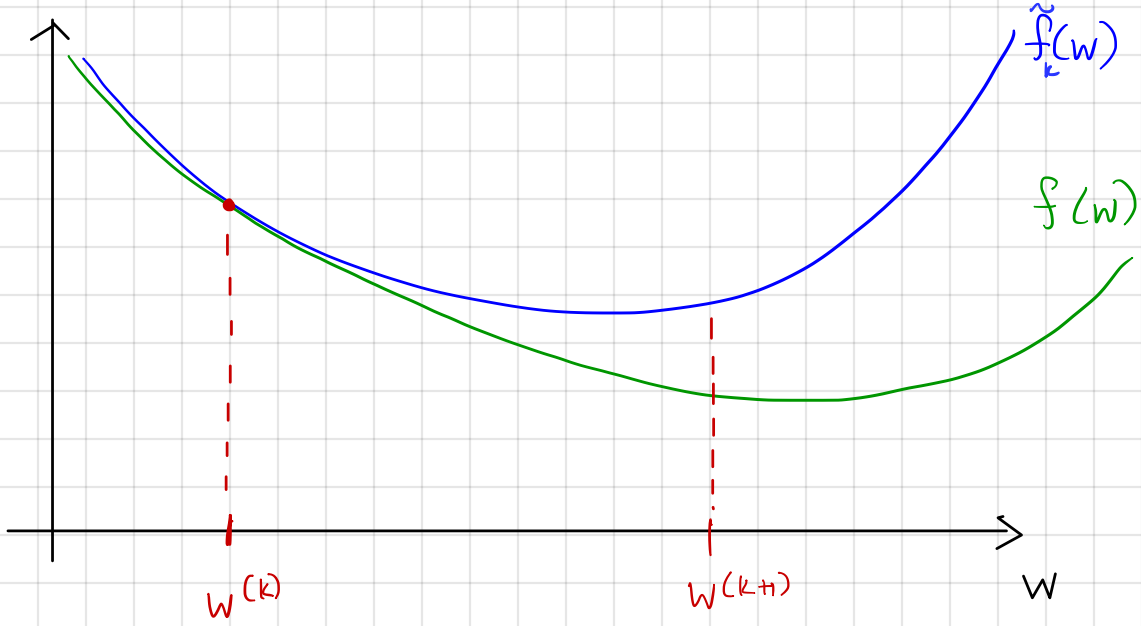
Consider $f(w) = \|y - Xw\|_2^2 = \|y - Xw^{(k)} + Xw^{(k)} - Xw\|_2^2$

$= \|y - Xw^{(k)}\|_2^2 + 2(y - Xw^{(k)})^T X(w^{(k)} - w) + \|Xw^{(k)} - Xw\|_2^2$

$\leq \underbrace{\|y - Xw^{(k)}\|_2^2}_{\text{same value regardless of } w} + 2(y - Xw^{(k)})^T X(w^{(k)} - w) + \|X\|_{op}^2 \|w^{(k)} - w\|_2^2$

Let τ be a step size, assume $\tau < 1/(2\|X\|_{op}^2)$ $\left(\|X\|_{op}^2 < \frac{1}{2\tau}\right)$

$\Rightarrow f(w) \leq C + 2(y - Xw^{(k)})^T X(w^{(k)} - w) + \frac{1}{2\tau}\|w^{(k)} - w\|_2^2 =: \tilde{f}_k(w)$



$f(w^{(k)}) = \tilde{f}_k(w^{(k)})$

$f(w) \leq \tilde{f}_k(w)$

Choose $w^{(k+1)}$ to minimize $\tilde{f}_k$

Aside: $\|Xw\|_2^2 \leq \|X\|_{op}^2 \|w\|_2^2$ where $\|X\|_{op}$ = max singular value of X.

because $\|Xw\|_2^2 = \|U\Sigma V^T w\|_2^2$

$= \|\Sigma V^T w\|_2^2$

$= \sum_i \sigma_i^2 (V^T w)_i^2$

$\leq \sigma_{max}^2 \sum_i (V^T w)_i^2$

$= \sigma_{max}^2 \|V^T w\|_2^2$

$= \sigma_{max}^2 \|w\|_2^2$

$$\hat{w}_{KH} = \underset{w}{\operatorname{argmin}} \quad 2(y - Xw^{(k)})^T X(w^{(k)} - w) + \frac{1}{2\tau}\|w^{(k)} - w\|_2^2$$

$$= \underset{w}{\operatorname{argmin}} \quad \underbrace{2 \cdot 2\tau(y - Xw^{(k)})^T X}_{=: v^T}(w^{(k)} - w) + \|w^{(k)} - w\|_2^2$$

Let $v := 2\tau X^T(y - Xw^{(k)})$ — independent of w!

$$= \underset{w}{\operatorname{argmin}} \quad 2v^T(w^{(k)} - w) + \|w^{(k)} - w\|_2^2$$

$$= \underset{w}{\operatorname{argmin}} \quad \|v + w^{(k)} - w\|_2^2 - \|v\|_2^2$$

$$= \underset{w}{\operatorname{argmin}} \quad \|v + w^{(k)} - w\|_2^2 \quad = \quad w^{(k)} + v = w^{(k)} + 2\tau X^T(y - Xw^{(k)}) = \text{Gradient Descent Step !!!!!}$$

$$= w^{(k)} - 2\tau X^T(Xw^{(k)} - y)$$