Lecture 13

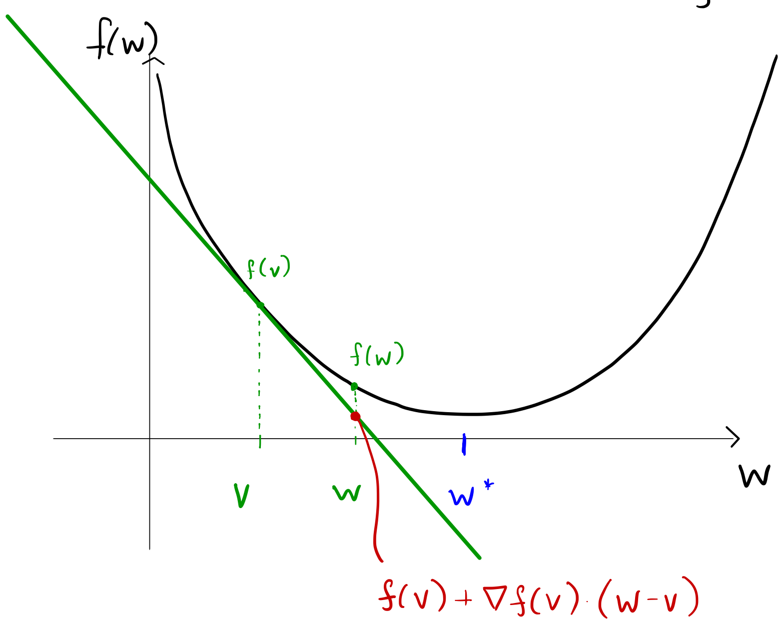
(Stochastic) Gradient Descent

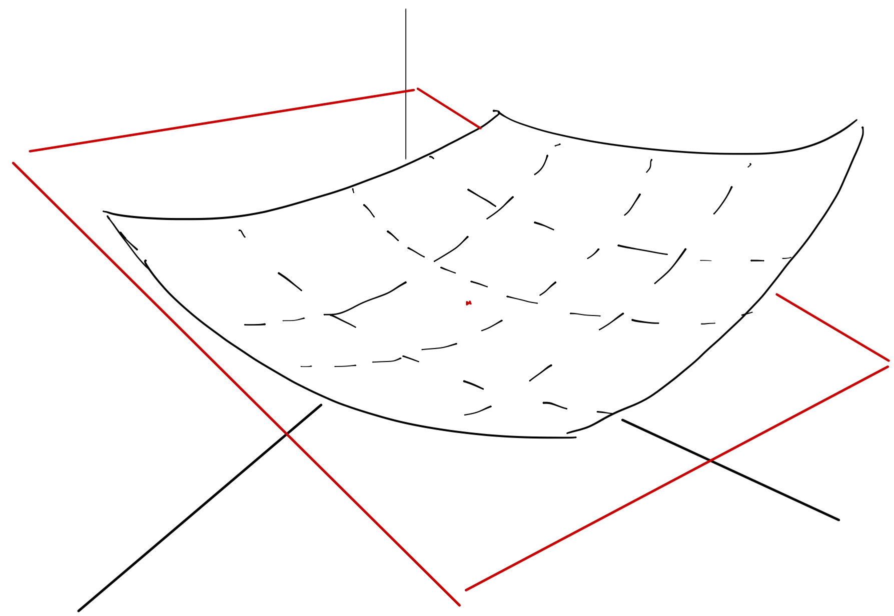
Basic convex optimization

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 $\underline{w}^* = (X^TX)^{-1}X^T\underline{y}$

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.

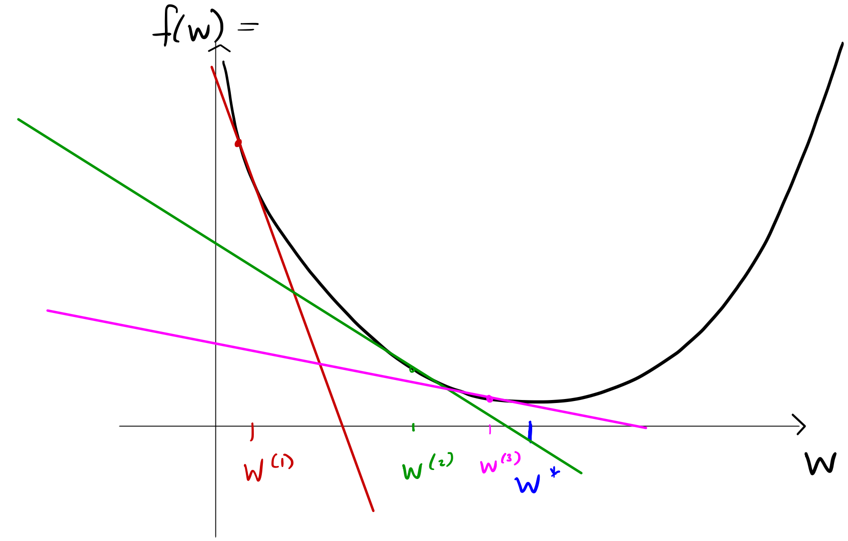
for $k = 1, 2, 3, \ldots$

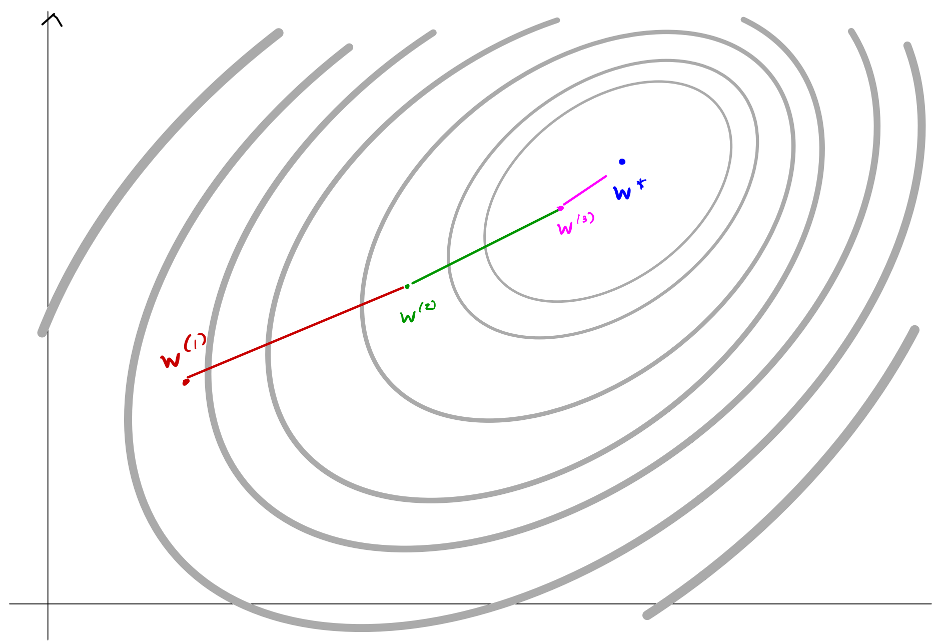
$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - 2\tau\left(X^T X \underline{w}^{(k)} - X^T\underline{y}\right)$$

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

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

$\tau > 0$ is step size (sometimes called learning rate)





More generally:

want to minimize $f(\underline{w})$

initialize with $\underline{w}^{(1)}$

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

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

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

Convergence of gradient descent for least squares

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

Optional

Subtract w^* from both sides:

$$\underbrace{w^{(k+1)} - w^*}_{\underline{e}^{(k+1)}} = \underbrace{w^{(k)} - w^*}_{\underline{e}^{(k)}} - \tau X^T X\big(\underbrace{w^{(k)} - w^*}_{\underline{e}^{(k)}}\big)$$

$$e^{(k+1)} = e^{(k)} - \tau X^T X e^{(k)}$$

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

We want $\underline{e}^{(k)} \to 0$ (ie. $\underline{w}^{(k)} \to \underline{w}^*$) as $k \to \infty$

$$\|e^{(k)}\| = \|(I - \tau X^T X) e^{(k-1)}\| \leq \underbrace{\sigma_{max}(I - \tau X^T X)}_{\text{this must be } < 1} \|e^{(k-1)}\|$$

if $X = U\Sigma V^T$, then $(I - \tau X^T X) = VV^T - \tau V \Sigma^T \Sigma V^T = V(I - \tau \Sigma^T \Sigma) V^T$

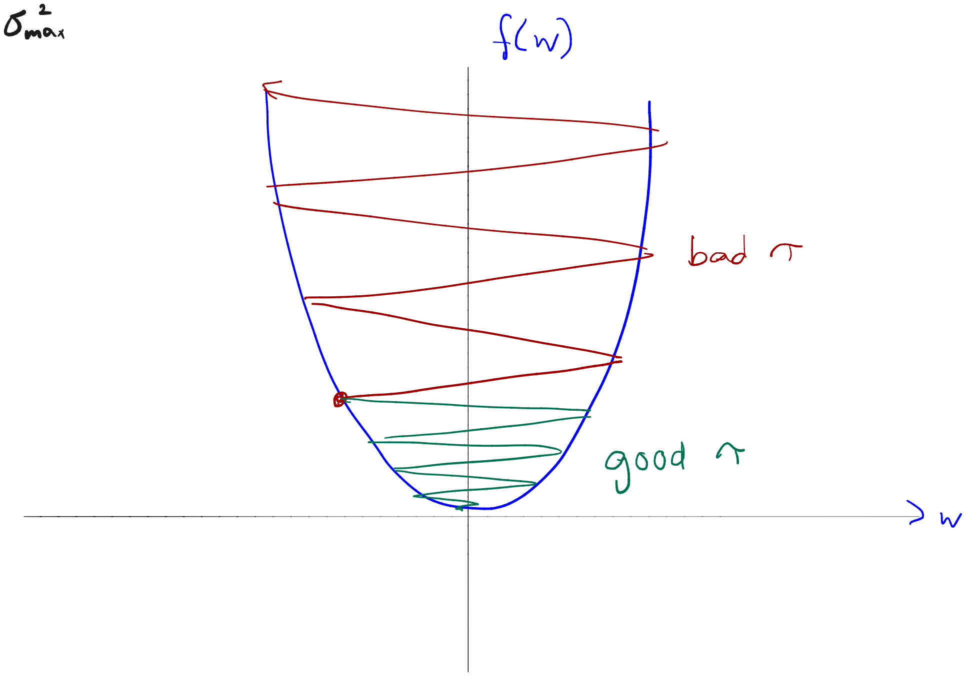
so max singular value of $I - \tau X^T X$ is $\max_i |1 - \tau \sigma_i^2|$

This is < 1 if $|1 - \tau \sigma^2_{max}(X)| < 1$ or if $\boxed{\tau < \frac{1}{\sigma^2_{max}(X)}}$

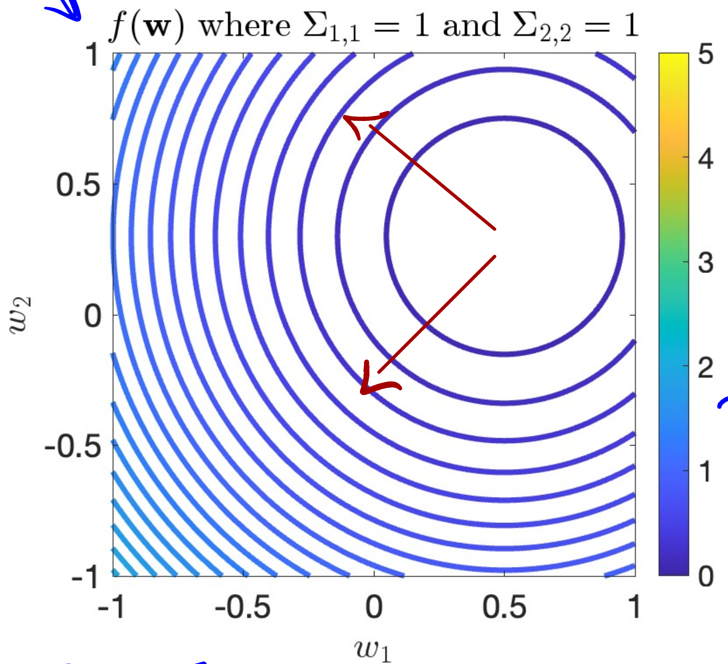
Step size τ needs to be small enough to ensure we find the minimum but not so small that it takes forever.

If f has more curvature, we need a smaller step size.

For convex problems like least squares, best $\tau \approx \frac{1}{\sigma_{max}^2}$

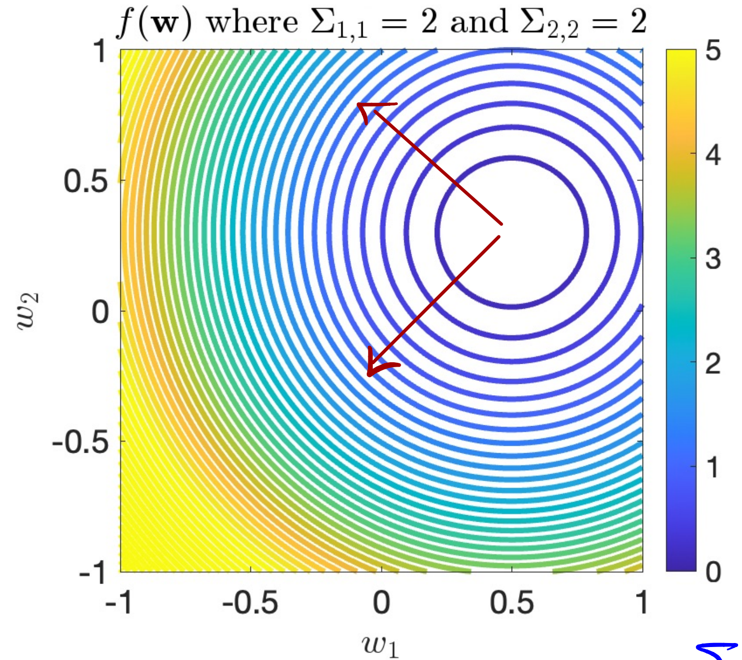


$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\qquad f(\underline{w}) = \| \underline{y} - X\underline{w} \|_2^2, \quad X = U \underline{\Sigma} V^T$

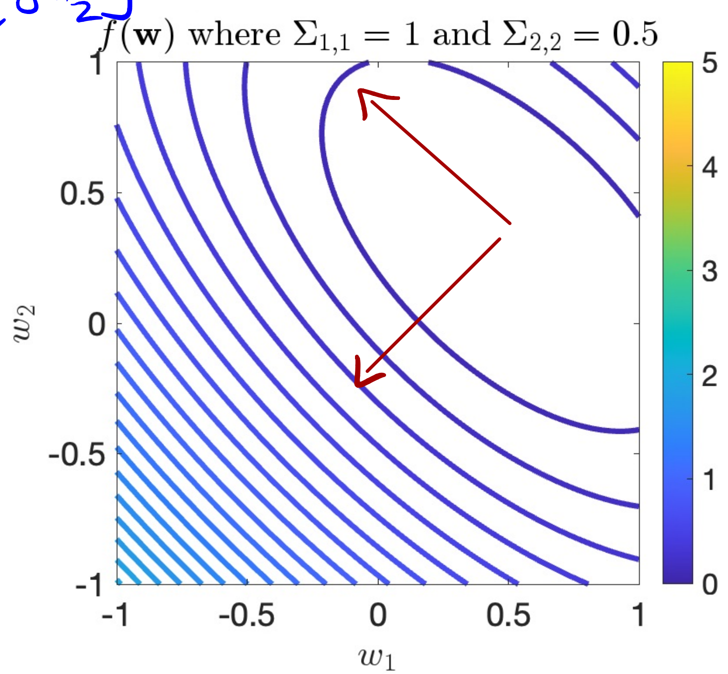


~ value of $f(w)$

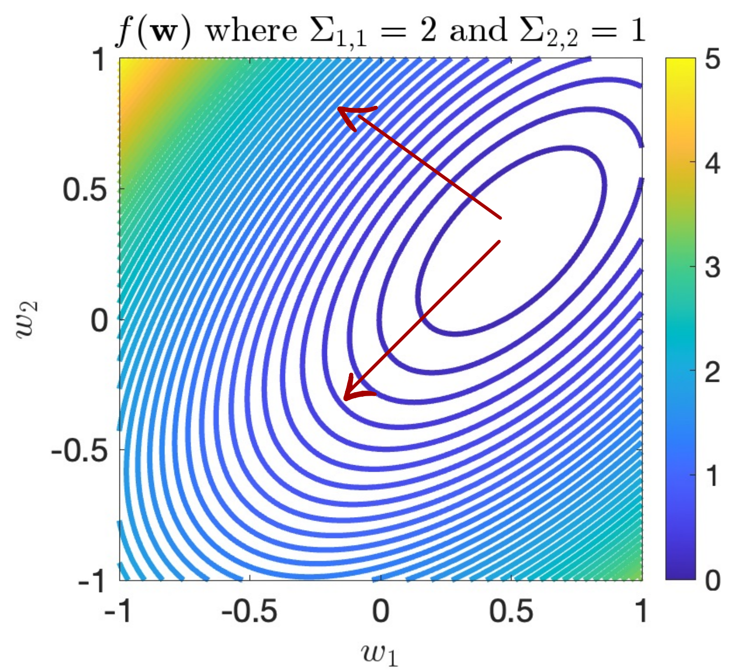
$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



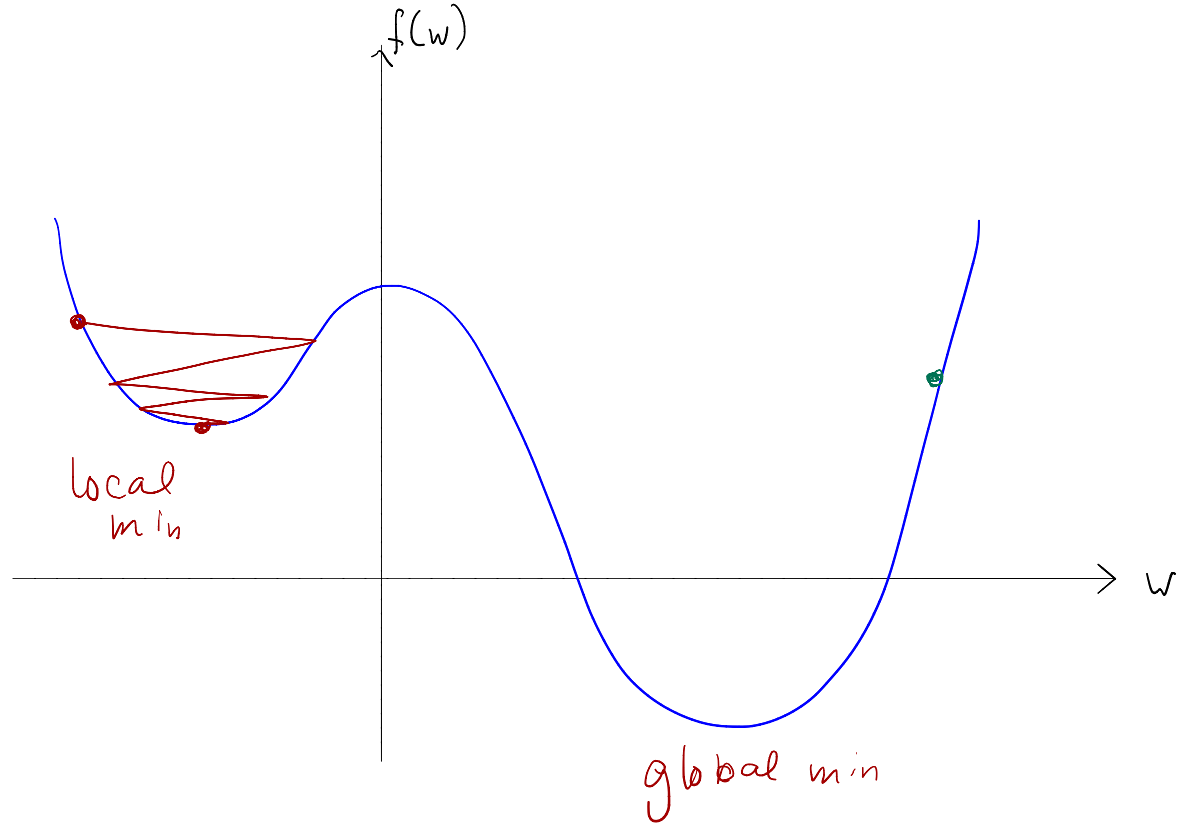
$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$



$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$



gradient descent will find a global minimizer of a $\boxed{\text{convex}}$ f no matter what we choose for $\underline{w}^{(1)}$ — guaranteed. But if f is non-convex, the result of gradient descent will depend heavily on where we place $\underline{w}^{(1)}$.



Gradient Descent for SVM

– more efficient methods exist!

$$\hat{\underline{\alpha}} = \underset{\underline{\alpha}}{\arg\min} \sum_{i=1}^{n} \Big(1 - y_i \sum_j \alpha_j K(\underline{x}_i, \underline{x}_j)\Big)_+ + \lambda \sum_i \sum_j \alpha_i \alpha_j K(\underline{x}_i, \underline{x}_j)$$

$$\ell(\underline{w}) = \sum_{i=1}^{n} (1 - y_i \underline{x}_i^T \underline{w})_+$$

$$\nabla_{\underline{w}} \ell = \sum_{i=1}^{n} \mathbb{I}\{y_i \underline{x}_i^T \underline{w} < 1\} (-y_i \underline{x}_i)$$

$$= \underset{\underline{\alpha}}{\arg\min} \underbrace{\sum_i (1 - y_i \underline{k}_i^T \underline{\alpha})_+ + \lambda \underline{\alpha}^T K \underline{\alpha}}_{f(\underline{\alpha})}$$

where $\underline{k}_i = i^{th}$ column of K

$$\Rightarrow \nabla_{\underline{\alpha}} f = \sum_i \mathbb{I}\{y_i \underline{k}_i^T \underline{\alpha} < 1\} (-y_i \underline{k}_i) + 2\lambda K \underline{\alpha}$$

can solve for $\underline{\alpha}$ using gradient descent:

$\eta > 0$ = step size

$\underline{\alpha}^{(0)}$ = initial guess

for $i = 1, 2, \ldots$

$$\underline{\alpha}^{(i)} = \underline{\alpha}^{(i-1)} - \eta \nabla_{\underline{\alpha}} f \Big|_{\underline{\alpha}^{(i-1)}}$$

$$= \underline{\alpha}^{(i-1)} - \eta \Big[\sum_{i=1}^{n} \mathbb{I}\{y_i \underline{k}_i^T \underline{\alpha}^{(i-1)} < 1\} (-y_i \underline{k}_i) + 2\lambda K \underline{\alpha}^{(i-1)}\Big]$$

if $\|\underline{\alpha}^{(i)} - \underline{\alpha}^{(i-1)}\| < \varepsilon$ then STOP

Stochastic Gradient Descent

Recall gradient descent:

$$\underline{w}^* = \underset{\underline{w}}{\operatorname{argmin}} f(w)$$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau \nabla f(\underline{w}^{(k)})$$

Imagine $\boxed{f(\underline{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\underline{w})}$

e.g. if $f(\underline{w}) = \frac{1}{n}\|y - X\underline{w}\|_2^2$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle x_i, \underline{w}\rangle)^2$$

$$\Rightarrow f_i(\underline{w}) = (y_i - \langle \underline{x}_i, \underline{w}\rangle)^2$$

Then Gradient Descent =

$$\underline{w}^{(t+1)} = w^{(t)} - \tau \sum_{i=1}^{n} \nabla f_i(\underline{w}^{(t)})$$

Now:

@ iteration t. choose $i \in \{1, 2, \ldots, n\}$

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} - \tau \nabla f_{i_t}(\underline{w}^{(t)})$$

- each iteration easier/faster to compute
- need more iterations

How to choose i_k?

A. cyclical ("Incremental Gradient Descent")

$i_t = t \bmod n$

e.g. $n = 3$: i_t's $= 1, 2, 3, 1, 2, 3, 1, 2, \ldots$

B. random permutations (common in practice)

every n rounds, reshuffle

e.g. $n=3$: i_t's $= \underbrace{1, 3, 2}_{\text{epoch 1}}, \underbrace{3, 1, 2}_{\text{epoch 2}}, \underbrace{2, 1, 3}_{\text{epoch 3}}, \ldots$

C. choose i_t uniformly at random

"stochastic gradient descent" (easier to analyze than random perturbations)

$i_t \sim \text{unif}(1, \ldots, n)$

e.g. $n=3$: i_t's $= 1, 3, 3, 2, 3, 1, 2, 2, 2, \ldots$

note: expected value $\mathbb{E}[\nabla f_{i_t}(w)] = \nabla f(\underline{w})$

Ex: $f(\underline{w}) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \langle \underline{x}_i, \underline{w} \rangle)^2 + \lambda \|\underline{w}\|_2^2$

$$f_i(w) = (y_i - \langle \underline{x}_i, \underline{w} \rangle)^2 + \lambda \|\underline{w}\|_2^2$$

check: $\frac{1}{n}\sum_{i=1}^{n} f_i(\underline{w}) = f(\underline{w})$

$$\nabla f_i(\underline{w}) = -2(y_i - \langle \underline{x}_i, \underline{w} \rangle)\underline{x}_i + 2\lambda \underline{w}$$

SGD: $\underline{w}^{(t+1)} = \underline{w}^{(t)} + 2\tau(y_{i_t} - \langle \underline{x}_{i_t}, \underline{w}^{(t)} \rangle)\underline{x}_{i_t} - 2\tau\lambda \underline{w}^{(t)}$

Mini-batch SGD:

1. randomly divide n samples into K batches

   e.g., $n = 12$, $k = 3$

   $\mathcal{B}_1 = \{1, 4, 6, 10\}$

   $\mathcal{B}_2 = \{3, 5, 9, 12\}$

   $\mathcal{B}_3 = \{2, 7, 8, 11\}$

2. for $k = 1, 2, \ldots, K$

   Let $f_k(\underline{w}) = \frac{K}{n}\sum_{i \in \mathcal{B}_k} f_i(\underline{w})$

   Compute batch gradient $\nabla_{\underline{w}} f_k$

   Update $\hat{\underline{w}}^{(t+1)} = \hat{\underline{w}}^{(t)} - \tau \nabla_{\underline{w}} f_k(\hat{\underline{w}}^{(t)})$

   $t = t+1$

3. If $\|\hat{w}^{(t)} - \hat{w}^{(t-K)}\|_2^2 < \varepsilon$, BREAK. otherwise, go to step 1.

Optional
Notes

Where do gradient descent updates come from?

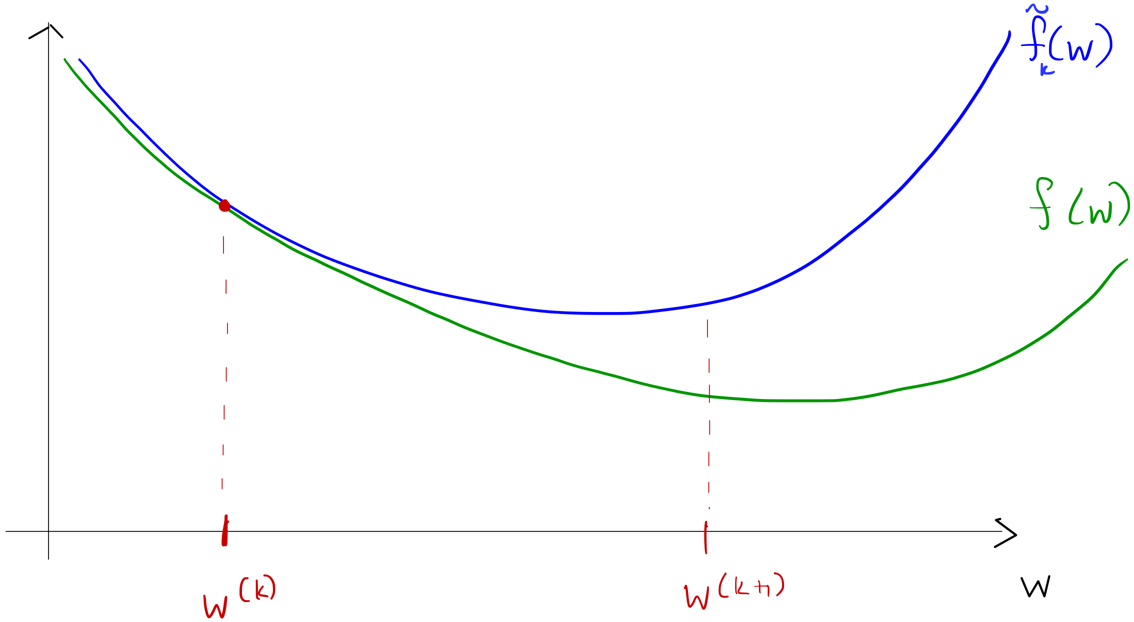
Consider $f(w) = \|y - Xw\|_2^2 = \|y - Xw^{(k)} + Xw^{(k)} - Xw\|_2^2$

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

$$\leq \underbrace{\|y - Xw^{(k)}\|_2^2}_{\text{same value regardless of } w} + 2\left(y - Xw^{(k)}\right)^T X\left(w^{(k)} - w\right) + \underbrace{\|X\|_{op}^2}_{\text{maximum singular value of } X \text{ squared}, \ \sigma_1^2} \|w^{(k)} - w\|_2^2$$

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

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



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

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

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

aside: $\|Xw\|_2^2 \leq \|X\|_{op}^2 \|w\|_2^2$

where $\|X\|_{op}$ = max singular value of X.

because
$$\|Xw\|_2^2 = \|U\Sigma V^T w\|_2^2$$
$$= \|\Sigma V^T w\|_2^2$$
$$= \sum_i \sigma_i^2 (V^T w)_i^2$$
$$\leq \sigma_{max}^2 \sum_i (V^T w)_i^2$$
$$= \sigma_{max}^2 \|V^T w\|_2^2$$
$$= \sigma_{max}^2 \|w\|_2^2$$

$$\hat{w}_{KH} = \arg\min_{w} \; 2(y - Xw^{(k)})^T X(w^{(k)} - w) + \frac{1}{2\tau}\|w^{(k)} - w\|_2^2$$

$$= \arg\min_{w} \; 2\,\underbrace{2\tau(y - Xw^{(k)})^T X}_{=:\, v^T}(w^{(k)} - w) + \|w^{(k)} - w\|_2^2$$

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

— independent of w !

$$= \arg\min_{w} \; 2v^T(w^{(k)} - w) + \|w^{(k)} - w\|_2^2$$

$$= \arg\min_{w} \; \|v + w^{(k)} - w\|_2^2 - \|v\|_2^2$$

$$= \arg\min_{w} \; \|v + w^{(k)} - w\|_2^2 = w^{(k)} + v = w^{(k)} + 2\tau X^T(y - Xw^{(k)})$$ = Gradient Descent Step !!!!!

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

Does this work?

$(a-b)^2 = a^2 - 2ab + b^2$

$\|a-b\|^2 = \|a\|^2 - 2a^Tb + \|b\|^2$

$a = Xw^{(k)} - y$

$b = 2\tau XX^T(Xw^{(k)}-y)$

$\boxed{\text{know: } \|Xa\|_2 \leq \|X\|_{op}\|a\|_2}$

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

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

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

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

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

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

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

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

$\Rightarrow$ 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

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