Lecture 17: Iterative Solvers (continued)

Least squares: $\hat{w} = (X^T X)^{-1} X^T y$

Gradient descent:

start w/ initial guess $\hat{w}^{(0)}$

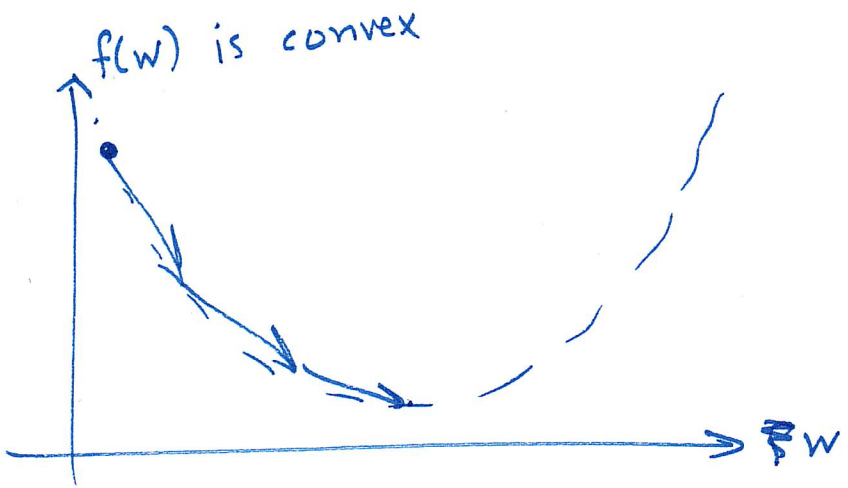
for $k = 0, 1, 2, \ldots$

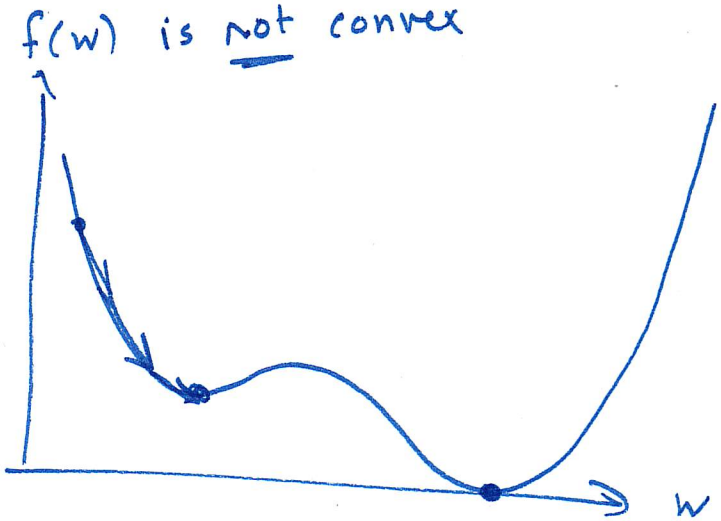
$$\hat{w}^{(k+1)} = \hat{w}^{(k)} - \underbrace{\tau}_{\text{step size}} \underbrace{X^T (X \hat{w}^{(k)} - y)}_{\text{gradient } \nabla_w f}$$

$$\underset{w}{\text{minimize}} \underbrace{\|y - Xw\|_2^2}_{f(w)}$$

$f(w)$ is convex

if $\tau < \frac{2}{\|X\|_{op}^2}$, then $\hat{w}^{(k)} \xrightarrow[k \to \infty]{} (X^T X)^{-1} X^T y$





Regularized Regression

$$\hat{w} = \underset{w}{\operatorname{argmin}} \underbrace{\|y - Xw\|_2^2}_{\text{squared error loss measures data fit}} + \underbrace{\lambda}_{\substack{\text{tuning parameter} \\ \lambda > 0}} \underbrace{r(w)}_{\text{regularizer}}$$

- $r(w) = \|w\|_2^2$ (Ridge)
- $r(w) = \|w\|_1 = \sum_i |w_i|$ (Lasso)

Let $f(w) = \|y - Xw\|_2^2 + \lambda r(w)$

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

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

does not depend on w
is Constant C with respect to w

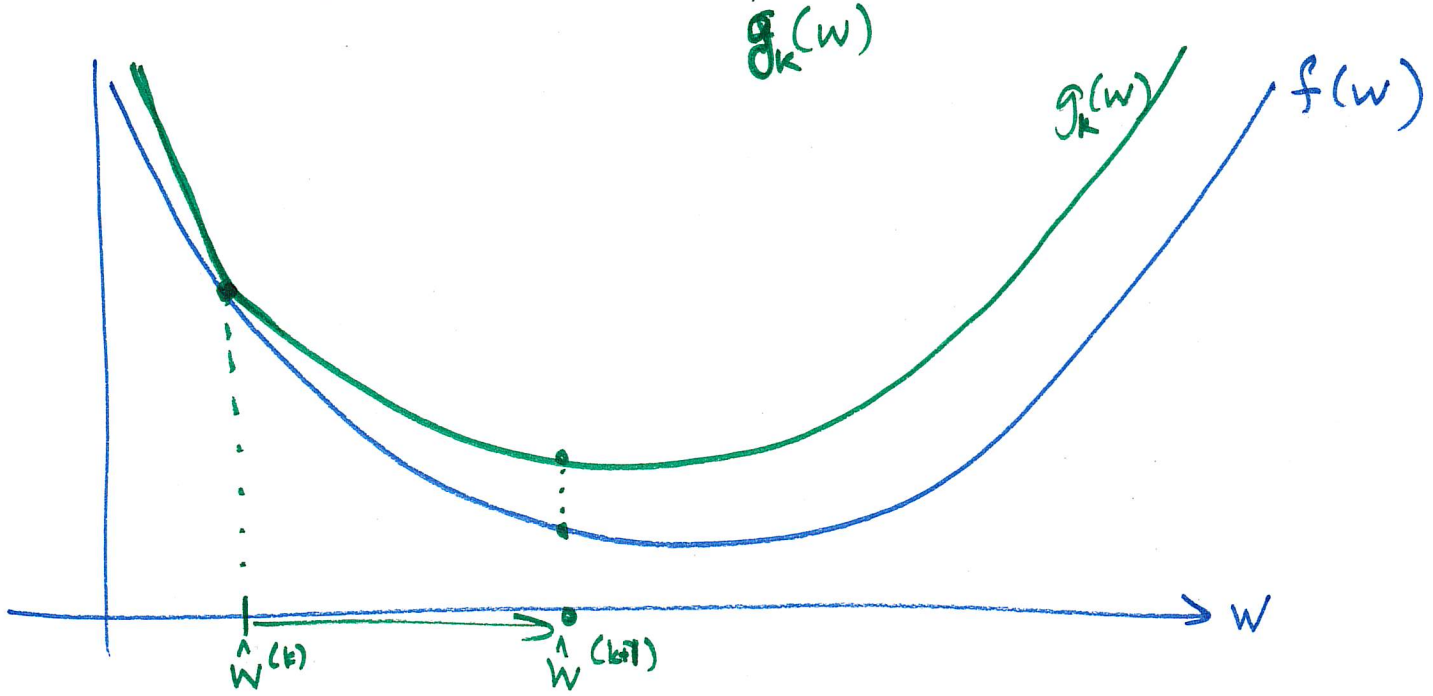
$\|Xa\|_2 \leq \|X\|_{op} \|a\|_2$

$\|X(\hat{w}^{(k)} - w)\|_2^2 \leq \|X\|_{op}^2 \|\hat{w}^{(k)} - w\|_2^2$

$f(w) \leq C + \|X\|_{op}^2 \|\hat{w}^{(k)} - w\|_2^2 + 2(y - X\hat{w}^{(k)})^T X(\hat{w}^{(k)} - w) + \lambda r(w)$

Let $\tau > 0$ be a step size, and assume $\tau < \frac{1}{\|X\|_{op}^2}$

$$f(w) \leq \underbrace{C + 2(y - X\hat{w}^{(k)})^T X(\hat{w}^{(k)} - w) + \tau^{-1}\|\hat{w}^{(k)} - w\|^2 + \lambda r(w)}_{g_k(w)}$$



$f(w) \leq g_k(w)$

$f(\hat{w}^{(k)}) = g_k(\hat{w}^{(k)})$

choose $\hat{w}^{(k+1)}$ to minimize bound on $f(w)$

$g_k(w)$

$$\hat{w}^{(k+1)} = \arg\min_w \; 2(y - X\hat{w}^{(k)})^T X(\hat{w}^{(k)} - w) + \tau^{-1}\|\hat{w}^{(k)} - w\|_2^2 + \lambda r(w)$$

$$= \arg\min_w \; 2\tau \underbrace{(y - X\hat{w}^{(k)})^T}_{= v^T} X(\hat{w}^{(k)} - w) + \|\hat{w}^{(k)} - w\|_2^2 + \lambda\tau r(w)$$

let $v = \tau X^T(y - X\hat{w}^{(k)})$ (v does not depend on w)

$$\hat{w}^{(k+1)} = \arg\min_w \; 2v^T(\hat{w}^{(k)} - w) + \|\hat{w}^{(k)} - w\|_2^2 + \lambda\tau r(w)$$

Complete the square

$$= \arg\min_w \; \|v + \hat{w}^{(k)} - w\|_2^2 - \|v\|_2^2 + \lambda\tau r(w)$$

$$= \arg\min_w \; \|v + \hat{w}^{(k)} - w\|_2^2 + \lambda\tau r(w)$$

Let $\hat{z}^{(k)} = v + \hat{w}^{(k)}$

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

$= \underbrace{\hat{w}^{(k)} - \tau X^T (X\hat{w}^{(k)} - y)}$

$=$ Landweber iteration/step

$=$ Grad. descent on squared error

$$\hat{w}^{(k+1)} = \underset{w}{\operatorname{argmin}} \; \| \hat{z}^{(k)} - w \|_2^2 + \lambda \tau r(w)$$

Algorithm (Proximal Gradient Descent)

- choose $\hat{w}^{(0)}$ (initial guess)
- for $k = 0, 1, 2, \ldots$

$$\hat{z}^{(k)} = \hat{w}^{(k)} - \tau X^T (X\hat{w}^{(k)} - y) \qquad \text{(GD)}$$

$$\hat{w}^{(k+1)} = \underset{w}{\operatorname{argmin}} \; \| \hat{z}^{(k)} - w \|_2^2 + \lambda \tau r(w) \qquad \text{(regularize)}$$

- check convergence: if $\| \hat{w}^{(k)} - \hat{w}^{(k+1)} \| < \varepsilon$, break

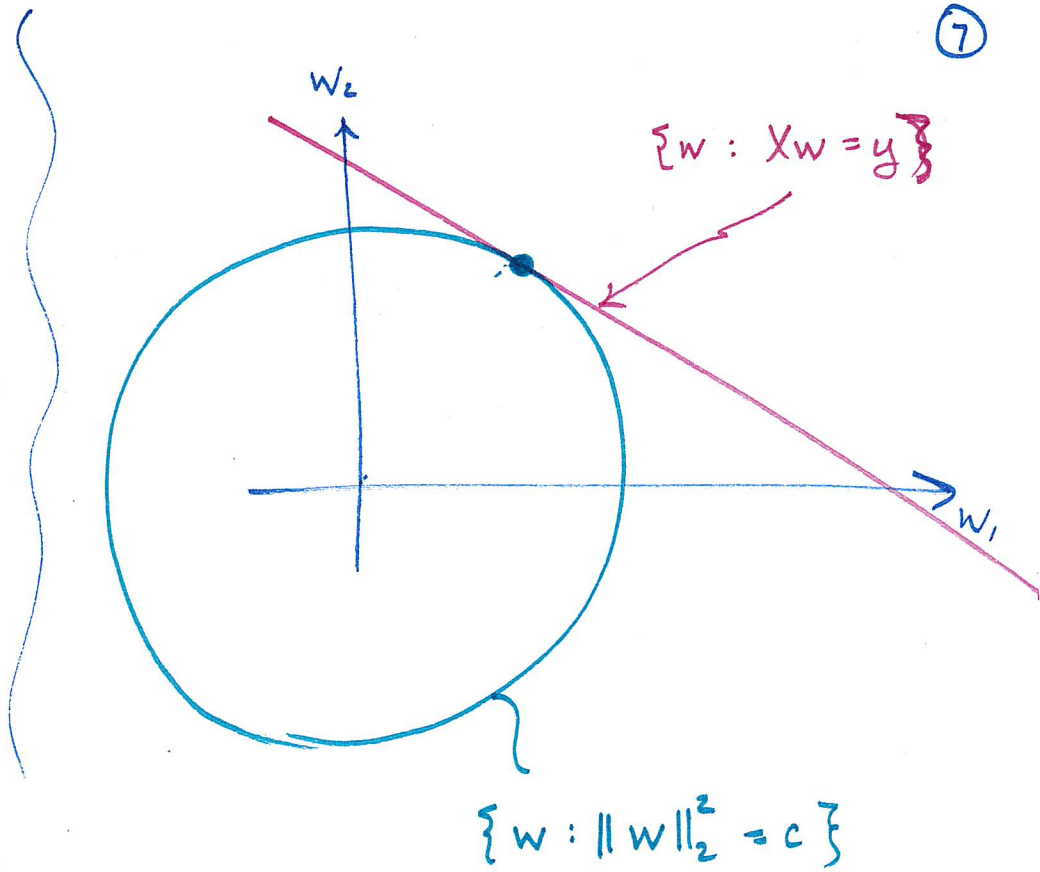
Ex: $r(w) = \|w\|^2$

$$\hat{w}^{(k+1)} = \underset{w}{\operatorname{argmin}} \underbrace{\|\hat{z}^{(k)} - w\|_2^2 + \lambda\tau\|w\|_2^2}_{g(w)}$$

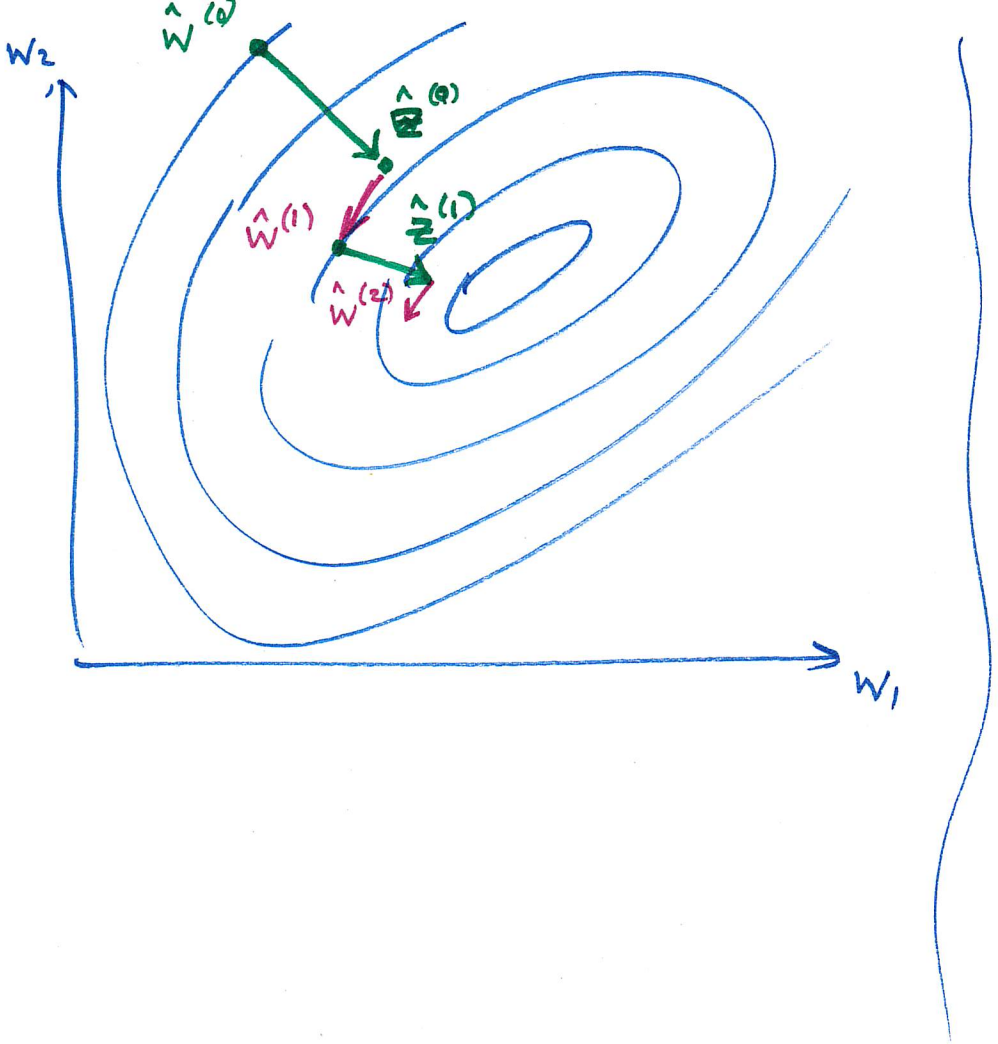
$$\nabla_w g = -2(\hat{z}^{(k)} - w) + 2\lambda\tau w = 0$$

$$\Rightarrow (2 + 2\lambda\tau) w = 2\hat{z}^{(k)}$$

$$\hat{w}^{(k+1)} = \frac{\hat{z}^{(k)}}{1 + \lambda\tau}$$



$$\Rightarrow \hat{w}^{(k+1)} = \frac{1}{1 + \lambda\tau}\left(\hat{w}^{(k)} - \tau X^T(Xw^{(k)} - y)\right)$$


w_2
$\hat{w}^{(0)}$
$\hat{\nabla}\hat{E}^{(0)}$
$\hat{w}^{(1)}$
$\hat{\nabla}\hat{E}^{(1)}$
$\hat{w}^{(2)}$
w_1