Lecture 18: LASSO

(least absolute selection + shrinkage operator)

no Willett office hours this week!
(Extra TA office hours.)

obs : $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$
or
$\mathbb{R}^p \times \{-1, +1\}$ for $i=1, \ldots, n$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix} \in \mathbb{R}^{n \times p}$$

Goal: w s.t. $y \approx Xw$

① Least squares

check is X is full rank ($\text{rank}(X) = p \leq n$)

$$\hat{w}_{LS} = \arg\min_{w} \|y - Xw\|_2^2$$

$$= \boxed{(X^TX)^{-1}X^Ty}$$

Problem: some singular values of X may be small

② Ridge regression:

$$\boxed{\hat{w}_R = \arg\min_{w} \|y - Xw\|_2^2 + \lambda\|w\|_2^2}$$

$$= \boxed{(X^TX + \lambda I)^{-1}X^Ty}$$

- eigenvalues of $(X^TX + \lambda I)$ greater than λ
- generally all elements $\hat{w}_R \neq 0$

$$X = U\Sigma V^T \Rightarrow X^TX = V\Sigma^2V^T \Rightarrow X^TX + \lambda I = V\Sigma^2V^T + \lambda VV^T = V\underbrace{(\Sigma^2 + \lambda I)}_{\text{eigenvals of } X^TX + \lambda I}V^T$$

$$\Rightarrow (X^TX + \lambda I)^{-1} = V(\Sigma^2 + \lambda I)^{-1}V^T$$

alternative: $\hat{w}_{R2} = \underset{w}{\arg\min} \; \|w\|_2^2$

subject to $\|y - Xw\| < \varepsilon$

for any λ in $\hat{w}_\lambda$, there is some ε so that

$\hat{w}_{R2} = \hat{w}_\lambda$

3 sparse estimation

$$\|w\|_0 = \sum_{i=1}^{p} \mathbb{1}\{w_i \neq 0\} = \#\text{ of nonzero elements in } w$$

$=$ "ℓ_0 norm" not really a norm because

$\|aw\|_0 \neq a\|w\|_0$

e.g. $w = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \\ 3 \end{bmatrix} \Rightarrow \|w\|_0 = 3$

$$\hat{w}_s = \underset{w}{\operatorname{argmin}} \; \|w\|_0 \quad \text{subject to} \quad \|y - Xw\| < \varepsilon$$

or

$$= \underset{w}{\operatorname{argmin}} \; \|y - Xw\| + \lambda \|w\|_0$$

$$y = \begin{bmatrix} 2 \\ 3 \\ 4.1 \\ 4.9 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 3 & 2 \\ 3 & 4 & 4 \\ 4 & 5 & 4 \end{bmatrix}$$

$$\hat{w}_{LS} = \begin{bmatrix} -.05 \\ .95 \\ .1 \end{bmatrix} \qquad \underset{(\lambda = 10^{-2})}{\hat{w}_R} = \begin{bmatrix} -.03 \\ .93 \\ .11 \end{bmatrix} \qquad \hat{w}_s = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

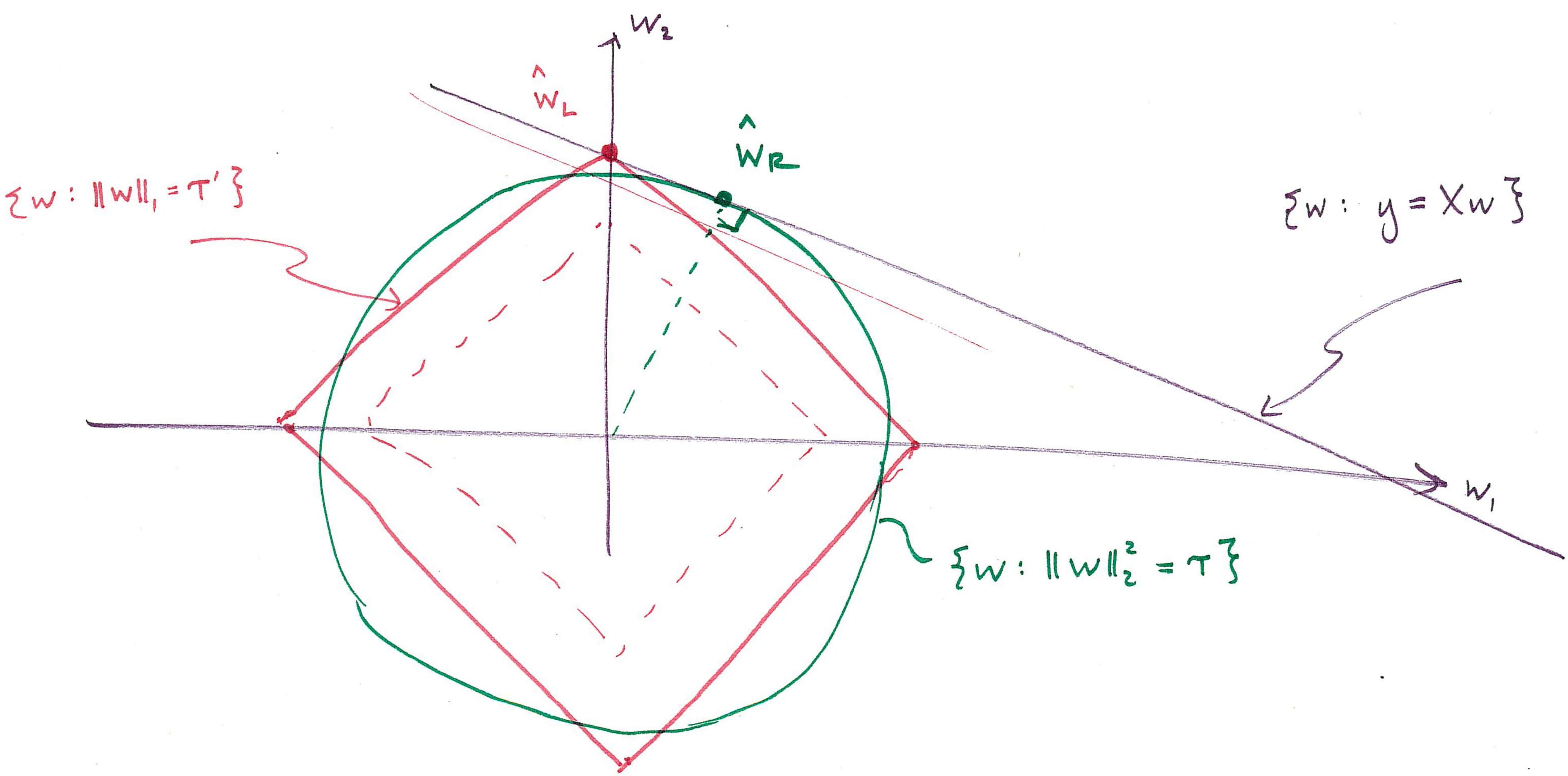
Problem: <u>very</u> hard to compute

Feature: "interpretable" — small # of explanatory variables

5 LASSO:

$$\hat{w}_L = \underset{w}{\operatorname{argmin}} \|w\|_1 \quad \text{subject to} \quad \|y - Xw\| < \varepsilon$$

$$\boxed{\hat{w}_L = \underset{w}{\operatorname{argmin}} \|y - Xw\|_2^2 + \lambda \|w\|_1}$$



measures of performance:

A prediction error

$\| X w^* - X \hat{w} \|$ — Ridge ~~Rigde~~ is great

B model selection error

$\| w^* - \hat{w} \|$ — Lasso can be effective

---

debiasing:

$\hat{S}$ = support $(\hat{w}_L)$ = locations of nonzeros in $\hat{w}_L$

$X_{\hat{S}}$ = cols of X in $\hat{S}$

$\hat{w} = (X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T y$

Proximal gradient algorithm

$$\arg\min_{w} \|y - Xw\|_2^2 + \lambda r(w)$$

choose initial $\hat{w}^{(0)}$

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

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

$$* \quad \hat{w}^{(k+1)} = \arg\min_{w} \|z^{(k)} - w\|_2^2 + \lambda \tau r(w) \quad \leftarrow$$

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

end

eg. $r(w) = \|w\|_1 = \sum_i |w_i|$

$$* \quad \hat{w} = \arg\min_{w} \|z - w\|_2^2 + \lambda\tau \cancel{r(w)} \|w\|_1$$

$$= \arg\min_{w} \underbrace{\sum_i \left[(z_i - w_i)^2 + \lambda\tau |w_i| \right]}_{\text{separable}}$$

separable $\Rightarrow$ solve for each w_i independently!

$$\hat{w}_i = \underset{w}{\operatorname{argmin}} \; (z_i - w_i)^2 + \lambda\tau |w_i| \quad , \qquad \lambda, \tau > 0$$

case 1: $z_i > 0$

$\Rightarrow \underline{\hat{w}_i \geq 0}$

$$\hat{w}_i = \underset{w}{\operatorname{argmin}} \; (z_i - w_i)^2 + \lambda\tau w_i$$

$$\frac{d}{dw_i} = -2(z_i - w_i) + \lambda\tau = 0$$

$$\Rightarrow \lambda\tau = 2z_i - 2w_i$$

$$2w_i = 2z_i - \lambda\tau$$

$$w_i = z_i - \frac{\lambda\tau}{2}$$

if $z_i > \lambda\tau/2$, then $\hat{w}_i = z_i - \frac{\lambda\tau}{2}$

if $z_i < \lambda\tau/2$, then $\hat{w}_i = 0$

$$\hat{w}_i = \left(z_i - \frac{\lambda\tau}{2}\right)_+$$

case 2: $z_i < 0$

$\Rightarrow \underline{\hat{w}_i \leq 0}$

$$\hat{w}_i = \underset{w}{\operatorname{argmin}} \; (z_i - w_i)^2 - \lambda\tau w_i$$

$$\frac{d}{dw_i} = -2(z_i - w_i) - \lambda\tau = 0$$

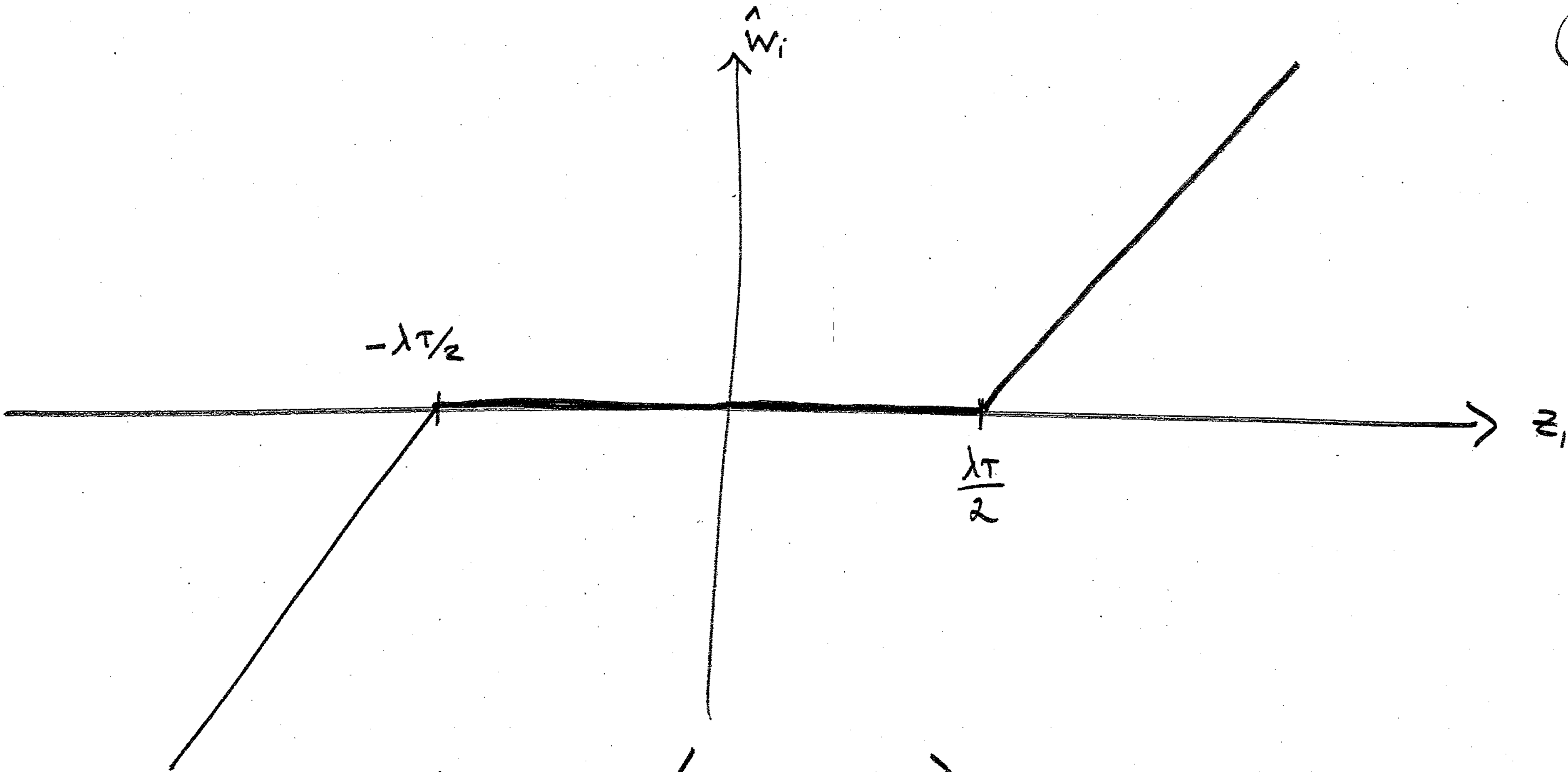
$$\Rightarrow \lambda\tau + 2z_i = 2w_i$$

$$\Rightarrow w_i = z_i + \frac{\lambda\tau}{2}$$

if $z_i < -\lambda\tau/2$, $w_i = z_i + \frac{\lambda\tau}{2}$

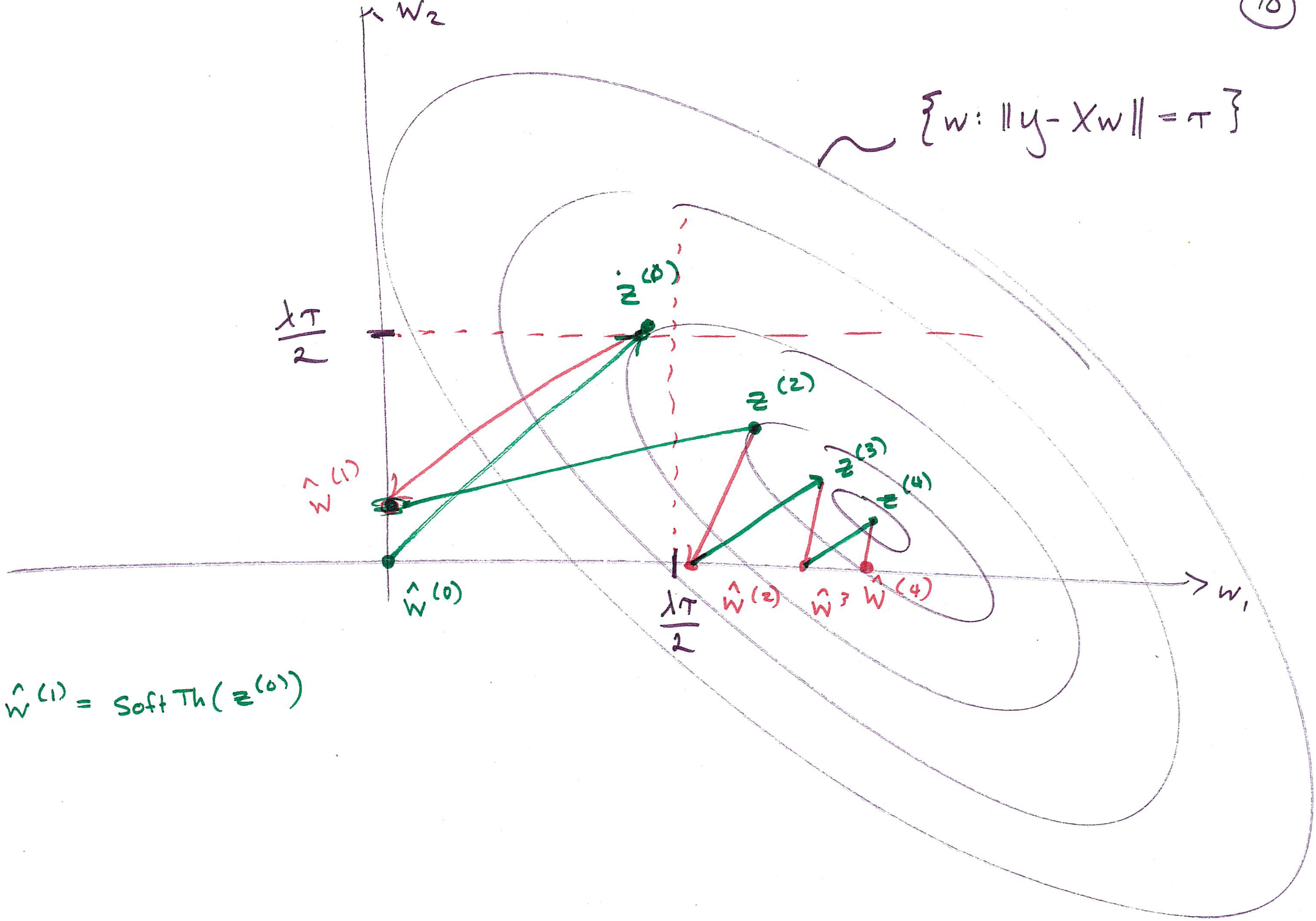
$z_i > -\lambda\tau/2$, $w_i = 0$

$$\hat{w}_i = \left[z_i + \frac{\lambda\tau}{2}\right]_- = -\left(|z_i| - \lambda\tau/2\right)_+$$



$$\hat{w}_i = \left(|z_i| - \frac{\lambda\tau}{2} \right)_+ \cdot \operatorname{sign}(z_i)$$

$$= \text{SoftThreshold}\left(z_i, \lambda\tau/2 \right)$$



$\hat{w}^{(1)} = \text{Soft Th}(z^{(0)})$