§ Proximal Gradient Descent.

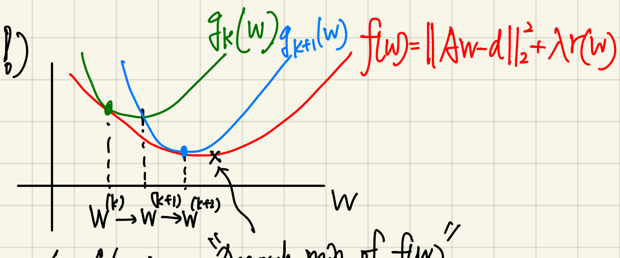
* 2 steps of Proximal GD: LS GD → regularize.
* Application: Ridge Regression.

• Proximal GD solves regularized LS problems

$$\min_w \underbrace{\|Aw-d\|_2^2 + \lambda r(w)}_{f(w)} \quad \begin{cases} r(w): \text{regularizer} \\ \lambda: \text{tuning parameter } (\lambda>0) \end{cases}$$

⊛ Some common "Convex regularizer"

- Ridge (Tikhonov): $r(w) = \|w\|_2^2 = \sum_{i=1}^{M} w_i^2$
- Lasso (ℓ_1): $r(w) = \|w\|_1 = \sum_{i=1}^{M} |w_i|$ (not differentiable!)



⊛ Idea: find a $g_k(w)$ s.t. $g_k(w)$ "touches" $f(w)$

(i) Solves easier min problem

(ii) Simple for separable regularizer (★) $r(w) = \sum_i h_i(w_i)$ (No cross terms $w_i w_j$)

minimize $g_k(w) \Rightarrow f(w)$ decreases.

$\Rightarrow$ find $g_k(w)$ s.t. $f(w) \le g_k(w)$, $g_k(w^{(k)}) = f(w^{(k)})$
(f lies below)

"Find a "nice" convex upperbound"

Define step size: $0 < \tau < \frac{1}{\|A\|_{op}^2} \Rightarrow \frac{1}{\tau} > \|A\|_{op}^2$

Consider $f(w) = \|d - Aw\|_2^2 + \lambda r(w)$

$$= \|d - Aw^{(k)} + Aw^{(k)} - Aw\|_2^2 + \lambda r(w)$$

expand.

$$= \underbrace{\|d - Aw^{(k)}\|_2^2}_{C_k} + \underbrace{\|A(w^{(k)}-w)\|_2^2}_{\le \|A\|_{op}^2 \|w^{(k)}-w\|_2^2} + 2\underbrace{(d-Aw^{(k)})'A}_{\equiv V_k'}(w^{(k)}-w) + \lambda r(w)$$

$$\le C_k + \|A\|_{op}^2 \|w^{(k)} - w\|_2^2 + 2V_k'(w^{(k)}-w) + \lambda r(w)$$

$$\le C_k + \frac{1}{\tau}\|w^{(k)} - w\|_2^2 + 2V_k'(w^{(k)}-w) + \lambda r(w) \equiv g_k(w)$$

Note that $g_k(w)$ is separable if $r(w)$ separable (no $w_i w_j$ terms) $\Rightarrow g_k(w) = C_k + \sum_{i=1}^{M} g_i(w_i)$

Solution. Find $W^{(k+1)} = \arg\min_{w} g_k(w)$,

where $g_k(w) = C_k + \frac{1}{\tau}\|W^{(k)} - w\|_2^2 + 2V_k'(W^{(k)} - w) + \lambda r(w)$

$$g_k(w) = C_k + \frac{1}{\tau}\|W^{(k)} - w\|_2^2 + 2V_k'(W^{(k)} - w) + \lambda r(w)$$

$$\Rightarrow \tau g_k(w) = \tau C_k + \underbrace{\underbrace{(W^{(k)} - w)'(W^{(k)} - w)}_{\|\ \|_2^2} + 2\tau V_k'(W^{(k)} - w)}_{\text{Complete square}} + \lambda\tau r(w).$$

$$= \tau C_k + (\tau V_k + (W^{(k)} - w))'(\tau V_k + (W^{(k)} - w)) - \tau^2 V_k' V_k + \lambda \tau r(w)$$

$$= \tau C_k + ((\tau V_k + W^{(k)}) - W)'((\tau V_k + W^{(k)}) - W) - \tau^2 V_k' V_k + \lambda\tau r(W)$$

$$\equiv (Z^{(k)} - W)'(Z^{(k)} - W) + \lambda\tau r(W) + \text{const.} = \|Z^{(k)} - W\|_2^2 + \lambda\tau r(w) + \text{const.}$$

$$\Rightarrow \boxed{W^{(k+1)} = \arg\min_{w} \|Z^{(k)} - w\|_2^2 + \lambda\tau r(w)},$$

Where $Z^{(k)} \equiv \tau V_k + W^{(k)}$

$= W^{(k)} + \tau A'(d - AW^{(k)})$ exactly. (Landweber).

$= W^{(k)} - \tau A'(AW^{(k)} - d)$ ← which is "Gradient Descent Iteration"!.

- Sum up: Proximal Gradient Descent alternates LS GD & Regularization

Algorithm. (Proximal GD)

Set $W^{(0)} = 0$, τ s.t. $0 < \tau < \frac{1}{\|A\|_{op}^2}$ ← Initialization

for $k = 1, 2, \cdots$ (to converge)

$Z^{(k)} = W^{(k)} - \tau A'(AW^{(k)} - d)$ ← original LS Gradient Descent

$W^{(k+1)} = \arg\min_{w} \|Z^{(k)} - w\|_2^2 + \lambda\tau r(w)$ ← Regularize.

if $\|W^{(k+1)} - W^{(k)}\| < \varepsilon$: ← check if $W^{(k+1)}$ converges to $W^{(k)}$

break.

⊛ Regularization is simple if $r(w)$ separable.

$\Rightarrow$ if $r(w) = \sum_{i=1}^{M} h_i(w_i)$, then $W^{(k+1)} = \arg\min_{w_i, i=1,\dots,M} \sum_{i=1}^{M} \left((Z_i^{(k)} - W_i)^2 + \lambda\tau h(W_i)\right)$

which is M scalar minimizations.

Example. (Ridge/Tikhonov in Proximal GD)

$$f(w) = \|d - Aw\|_2^2 + \lambda \|w\|_2^2$$

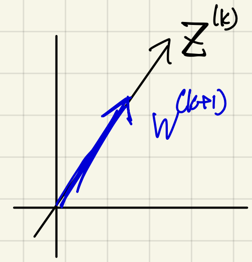
$h(w_i) = \|w\|_2^2 \Rightarrow \sum_i w_i^2$

1° LS GD: $z^{(k)} = w^{(k)} - \tau A^T (Aw^{(k)} - d)$

2° Regularization: $w^{(k+1)} = \underset{w_i, i=1,\dots,M}{\operatorname{argmin}} \sum_{i=1}^{M} (z_i^{(k)} - w_i)^2 + \lambda\tau w_i^2$

$\Rightarrow w_i^{(k+1)} = \frac{1}{1+\lambda\tau} z_i^{(k)}$ (solves FOC [w_i])

$\Rightarrow \boxed{w^{(k+1)} = \underbrace{\frac{1}{1+\lambda\tau}}_{(<1)} z^{(k)}}$ shrinking toward origin



§ LASSO Regression.

(*) Search for sparse solutions.
(*) ℓ_1-norm regularization (LASSO).

Consider. $A\mathbf{w} = [a_1 \cdots a_M]\begin{bmatrix} w_1 \\ \vdots \\ w_M \end{bmatrix} = \sum_{i=1}^{M} w_i a_i$ ⇒ with $w_\ell \approx 0$ it implies a_ℓ not important.

⇒ If only a few w_ℓ's are "non-zero" (important)

then we have "sparse $\mathbf{w}$"

$\|\mathbf{w}\|_0 \equiv \sum_{i=1}^{M} \mathbb{1}\{w_i \neq 0\}$ (counting non-zero w_ℓ).

(*) ℓ_0-norm is NOT norm as $\|a\mathbf{w}\|_0 \neq a\|\mathbf{w}\|_0$

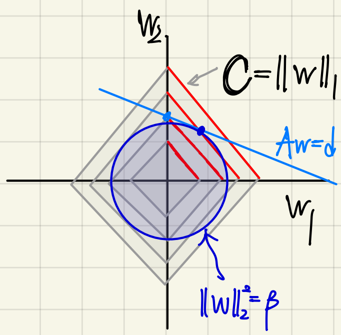
Consider. $\min_{w} \|\mathbf{w}\|_0$ s.t. $\|A w - d\|_2^2 < \varepsilon$.

⊗ Problem: $\|\mathbf{w}\|_0$ not convex ⇒ "computationally intractable".

- Convex Relaxation gives tractable problem

$$\min_{w} \underbrace{\|\mathbf{w}\|_1}_{\text{Convex}} \text{ s.t. } \|Aw-d\|_2^2 < \varepsilon.$$

Least Absolute Selection & Shrinkage Operator (LASSO)



$C = \|w\|_1 = \sum_{i=1}^{M}|w_i|$: $\xrightarrow{\text{2D case.}}$ $|w_1|+|w_2| = C$

In 1st quad: $w_1 + w_2 = C$

Now consider ① $\min \|w\|_1$ s.t. $Aw = d$. ⇒ "Corner" on $\|w\|_1$ sparse solutions.

② $\min \|w\|_2^2$ s.t. $Aw = d$. ⇒ "circular" $\|w\|_2^2$ (less-likely corner) non-sparse solutions

- LASSO is a regularized LS problem.

LASSO: $\min_{w} \|w\|_1$ s.t. $Aw - d < \varepsilon$. $\overset{\text{eqv.}}{\sim}$ $\min_{w} \|Aw-d\|_2^2 + \lambda\|w\|_1$ for some λ. ε.

$\left(\overset{\text{eqv.}}{\sim} \min_{w} \|w\|_1 + \frac{1}{\lambda}\|Aw-d\|_2^2\right)$

LASSO.	$w_L = \arg\min_{w} \|Aw-d\|_2^2 + \lambda\|w\|_1$	: Sparse w_L ; Can have small model error ($w_{opt} - w_L$.); Iterative solution method.
Ridge	$w_R = \arg\min_{w} \|Aw-d\|_2^2 + \lambda\|w\|_2^2$	: non-sparse w_R ; Can have small prediction error ($\|Aw_{opt} - Aw_R\|_2^2$) ; Solve in closed form.

- LASSO & Feature Selection.

$$W_L = \underset{w}{\mathrm{argmin}} \|Aw-d\|_2^2 + \lambda\|w\|_1.$$

1° Selection: $S_L = \{ i : [W_L]_i \neq 0 \}$ (the non-zero W_{L_i})

2° $AW_L = \sum_{i=1}^{M} a_i [W_L]_i = \sum_{i \in S_L} a_i [W_L]_i$

3° Debiasing: $A_L = \{ a_i : i \in S_L \}$ (debias A by those a_i's selected by LASSO)

4° Re-solve LS problem: $\underline{\hat{W}_L = \underset{w}{\mathrm{argmin}} \| A_L w - d \|_2^2} = (A_L' A_L)^{-1} A_L' d.$ avoids $\|w\|_1$ shrinkage.

§ LASSO & Proximal GD.

- l_1-regularized LS problem can be solved by Proximal GD.

$\begin{cases} \text{GD.} \\ \text{Regularization ("shrinkage")} \end{cases}$

$\min_w \|Aw-d\|_2^2 + \lambda\|w\|_1$ encourages sparse sol'n.

Apply "Proximal GD" (since no close form)

1° $z^{(k)} = w^{(k)} - \tau A'(Aw^{(k)}-d)$ ← LS GD

☆ 2° $w^{(k+1)} = \arg\min_w \|z^{(k)}-w\|_2^2 + \tau\lambda\|w\|_1$ ← regularization.

- Regularization steps involves "scalar minimizations"

$$\min_w \|z^{(k)}-w\|_2^2 + \tau\lambda\|w\|_1 \Rightarrow \min_{w_i,\, i=1,\dots,M} \sum_{i=1}^{M} (z_i^{(k)}-w_i)^2 + \lambda\tau|w_i| \quad (\lambda\tau>0)$$

Case ①: $w_i \geq 0$ (1st quad)

$$\Rightarrow \min_{w_i} (z_i-w_i)^2 + \lambda\tau w_i \Rightarrow [w_i]: 0 = -2(z_i-w_i)+\lambda\tau \Rightarrow w_i = z_i - \tfrac{1}{2}\lambda\tau$$

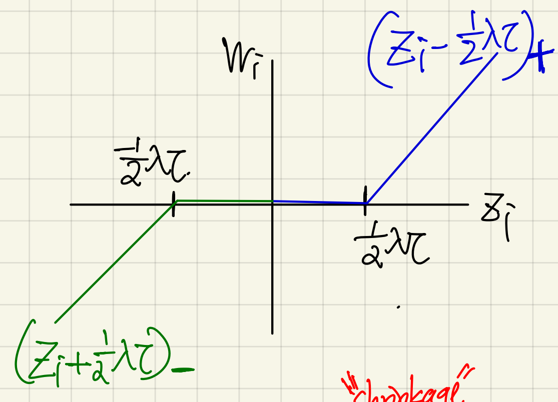
So, $w_i = \begin{cases} z_i - \frac{1}{2}\lambda\tau, & \text{if } z_i > \frac{1}{2}\lambda\tau \\ 0, & \text{if } z_i < \frac{1}{2}\lambda\tau. \end{cases}$ $\qquad w_i = (z_i - \tfrac{1}{2}\lambda\tau)_+$

Case ②: $w_i \leq 0$ (3rd quad)

$$\Rightarrow \min_{w_i} (z_i-w_i)^2 - \lambda\tau w_i \Rightarrow [w_i]$$

So, $w_i = \begin{cases} 0, & \text{if } z_i > -\frac{1}{2}\lambda\tau \\ z_i + \frac{1}{2}\lambda\tau, & \text{if } z_i < -\frac{1}{2}\lambda\tau \end{cases}$

$$w_i = (z_i + \tfrac{1}{2}\lambda\tau)_-$$



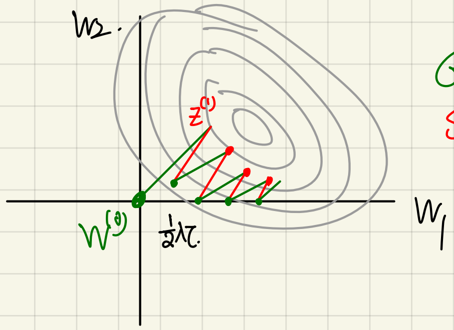
"Soft threshold"

"shrinkage"

$$w_i = \begin{cases} z_i - \frac{1}{2}\lambda\tau, & z_i \in (\frac{1}{2}\lambda\tau, \infty) \\ 0, & z_i \in (-\frac{1}{2}\lambda\tau, \frac{1}{2}\lambda\tau) \\ z_i + \frac{1}{2}\lambda\tau, & z_i \in (-\infty, -\frac{1}{2}\lambda\tau) \end{cases}$$

$$= (|z_i| - \tfrac{1}{2}\lambda\tau)_+ \operatorname{sign}\{z_i\}$$

- It alternates Descent & Shrinkage (soft thresholding)



Gradient Descent

Shrinkage. (most likely send W back to 0)