1.

(a) F LASSO is a biased estimator, so it cannot prevent bias

(b) F $\hat{\theta}_{Lasso}$ could have negative coordinates.

(c) T For instance, when X is not full rank, it's possible to have many Lasso solutions.

(d) F it is recommended because different features may have different magnitudes. Otherwise, the regularization is unfair

(e) F when $n>d$ we invert $\underset{d \times d}{(X^TX+\lambda I_d)^{-1}}$, which is the original ridge regression

(f) T Gram matrix is a PSD

(g) T $K(x,z)=(x^Tz+1)^P$, P can be any large as it does not affect computational complexity

(h) F $K(x,z)=(x^Tz+1)^1$

(i) T in documentation, it says it often faster to fit a whole path than compute a single fit

1. $\min\limits_{\theta \in R^d} \frac{1}{n} \|X\theta - y\|_2^2 + w\|\theta\|_1$

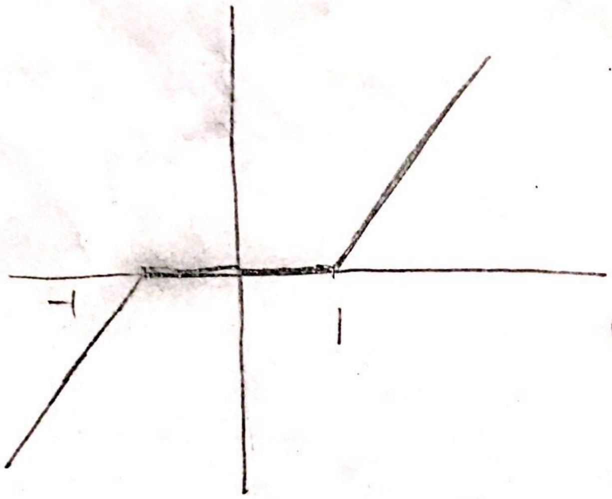
$= \min\limits_{\theta \in R^d} \frac{1}{n} (\|X\theta - y\|_2^2 + nw\|\theta\|_1)$

let $w = \frac{\lambda}{n}$

$= \min\limits_{\theta \in R^d} \frac{1}{n} (\|X\theta - y\|_2^2 + \lambda\|\theta\|_1)$

$= \min\limits_{\theta \in R^d} \|X\theta - y\|_2^2 + \lambda\|\theta\|_1$

2.



non-decreasing

3.

$$\min_{\theta_j} \sum_{i=1}^{n}\left(\theta_j x_{ij} + \sum_{k=1,k\neq j}^{d} \theta_k x_{ik} - y_i\right)^2 + \lambda \sum_{k=1,k\neq j}^{d} |\theta_k| + \lambda|\theta_j|$$

is the loss function when fixing all θ excepe θ_j

let this function be g and $\alpha = \theta_j$ we get

$$g(\alpha) = \sum_{i=1}^{n}\left(a x_{ij} + \sum_{k=1;k\neq j}^{d} \theta_k x_{ik} - y_i\right)^2 + \lambda|\alpha| + \lambda \sum_{k=1;k\neq j}^{d} |\theta_k|$$

4. $$\frac{\partial g}{\partial \alpha} = 2\sum_{i=1}^{n} x_{ij}\left(a x_{ij} + \sum_{k=1;k\neq j}^{d} \theta_k x_{ik} - y_i\right) + \lambda \quad \text{if } a > 0$$

$$\frac{\partial g}{\partial \alpha} = 2\sum_{i=1}^{n} x_{ij}\left(a x_{ij} + \sum_{k=1;k\neq j}^{d} \theta_k x_{ik} - y_i\right) - \lambda \quad \text{if } a < 0$$

5. if $a^* > 0$

$$\frac{\partial g}{\partial a} = 2\sum_{i=1}^{n} a x_{ij}^2 + 2\sum_{i=1}^{n} x_{ij}\Big(\underbrace{\sum_{k=1,k\neq j}^{d} \theta_k x_{ik} - y_i}_{= -C_j}\Big) + \lambda = 0$$

$$a^* \cdot a_j - C_j + \lambda = 0$$

$$a^* = \frac{1}{a_j}(C_j - \lambda)$$

6. if $a^* < 0$

$$\frac{\partial g}{\partial a} = 2\sum_{i=1}^{n} a x_{ij}^2 + 2\sum_{i=1}^{n} x_{ij}\Big(\underbrace{\sum_{k=1,k\neq j}^{d} \theta_k x_{ik} - y_i}_{= -C_j}\Big) - \lambda = 0$$

$$a^* a_j - C_j - \lambda = 0$$

$$a^* = \frac{1}{a_j}(C_j + \lambda)$$

if $a^* = 0$

$$-\lambda\alpha - C_j = 0 \qquad \beta \in [-1, 1] \text{ slope}$$

$$\beta = \frac{-C_j}{\lambda} \in [-1, 1] \Rightarrow C_j \in [-\lambda, \lambda]$$

7\. Based on (5) (6)

for a^* is positive, we need $c_j - \lambda > 0$

$$c_j > \lambda$$

for a^* is negative, We need $c_j + \lambda < 0$

$$c_j < -\lambda$$

$$8.\ D^{+}(g)(\alpha) = \lim_{\varepsilon \to 0, \varepsilon > 0} = \frac{\sum_{i=1}^{n} (A + \varepsilon X_{ij})^2 + \lambda|\alpha+\varepsilon| - \lambda|\alpha| - \sum_{i=1}^{n} (A)^2}{\varepsilon}$$

$$\text{let } A = \left(\alpha X_{ij} + \sum_{k=1, k\neq j}^{d} \theta_k X_{ik} - y_i\right)^2 \quad = \lim_{\varepsilon \to 0, \varepsilon > 0} \frac{2\sum_{i=1}^{n} \varepsilon X_{ij}(A) + \sum_{i=1}^{n} \varepsilon^2 X_{ij}^2 + \lambda\alpha + \lambda\varepsilon - \lambda\alpha}{\varepsilon}$$

$$= \lim_{\varepsilon \to 0, \varepsilon > 0} 2\sum_{i=1}^{n} X_{ij} A + \varepsilon \sum_{i=1}^{n} X_{ij}^2 + \lambda$$

$$= -C_j + \lambda$$

$$D^-(g)(\alpha) = \lim_{\varepsilon>0,\varepsilon\to 0} \frac{g(\alpha-\varepsilon) - g(\alpha)}{\varepsilon}$$

$$= \lim_{\varepsilon>0,\varepsilon\to 0} \frac{\sum_{i=1}^{n}(A-\varepsilon X_{ij})^2 + \lambda|\alpha-\varepsilon| - \lambda|\alpha| - \sum_{i=1}^{n} A^2}{\varepsilon}$$

$$= \lim_{\varepsilon>0,\varepsilon\to 0} \frac{\sum_{i=1}^{n} -2A\varepsilon X_{ij} + \sum_{i=1}^{n} \varepsilon^2 X_{ij}^2 + \varepsilon\lambda}{\varepsilon}$$

$$= \lim_{\varepsilon>0,\varepsilon\to 0} -2\sum_{i=1}^{n} X_{ij}A + \varepsilon\sum_{i=1}^{n} X_{ij}^2 + \lambda$$

$$= -2\sum_{i=1}^{n} X_{ij}\left(\alpha X_{ij} + \sum_{k=1, k\neq j}^{d} \theta_k X_{ik} - y_i\right) + \lambda$$

$$= C_j + \lambda$$

9. $D^+(g)(\alpha^*) = \lambda - C_j \geq 0$ since $\lambda > 0$ and $C_j \in [-\lambda, \lambda]$

$D^-(g)(\alpha^*) = C_j + \lambda \geq 0$ since $C_j \in [-\lambda, \lambda]$

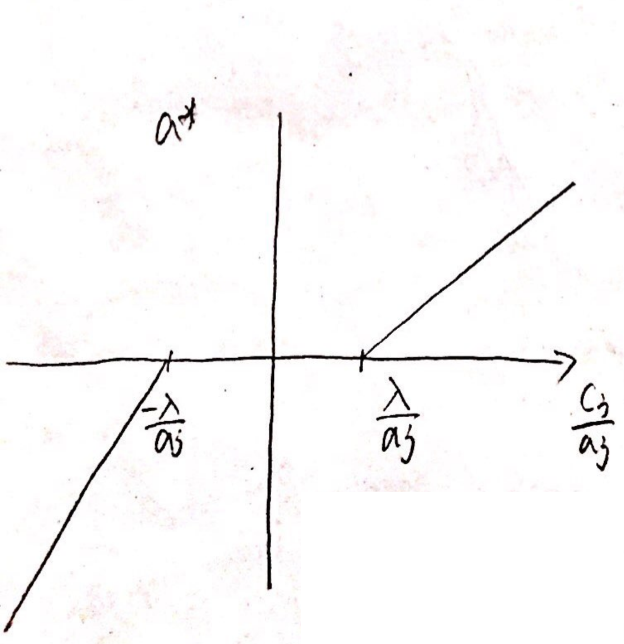
by result (4), $\alpha^* = 0$ satisfies the minimizer of g.

10. $\theta_j = \text{soft}\left(\frac{c_j}{a_j}, \frac{\lambda}{a_j}\right)$

in Algorithm 1

$a_j = 2\sum_{i=1}^{n} X_{ij}^2 \geq 0$

if not all $X_{ij} = 0$, we can assume $a_j > 0$



Based on Q5 and Q6 and Q8

$$a^* = \begin{cases} \frac{c_j}{a_j} - \frac{\lambda}{a_j} & c_j > \lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ \frac{c_j}{a_j} + \frac{\lambda}{a_j} & c_j < -\lambda \end{cases}$$

then a^* can be written as

$$a^* = \begin{cases} \frac{c_j}{a_j} - \frac{\lambda}{a_j} & \text{if } \frac{c_j}{a_j} > \frac{\lambda}{a_j} \\ 0 & \text{if } \frac{c_j}{a_j} \in \left[-\frac{\lambda}{a_j}, \frac{\lambda}{a_j}\right] \\ \frac{c_j}{a_j} + \frac{\lambda}{a_j} & \text{if } \frac{c_j}{a_j} < -\frac{\lambda}{a_j} \end{cases}$$

which is just the definition of

$\text{soft}\left(\frac{c_j}{a_j}, \frac{\lambda}{a_j}\right)$