MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2018**

Problem Set 2

Student Name: Sebastiani Aguirre-Navarro

Problem 1 1.1

Let

$$\sum_{i=1}^{n} x_i x_i^\top = I_{dxd}$$

or $X^\top X = I_{dxd}$ Also, let

$$\begin{aligned} L(w) &= \frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 + 2\lambda||w||_1 \\ &= \frac{1}{n}(Xw - Y)^\top (Xw - Y) + 2\lambda||w||_1 \end{aligned} \tag{2.1}$$

Then, the solution that minimizes $L(w)$ is

$$\begin{aligned} 0 &= \frac{\partial}{\partial w}L(w) = \frac{2}{n}X^T(Xw - Y) + 2\lambda\frac{\partial}{\partial w}||w||_1 \\ 0 &= X^\top Xw - X^\top Y + \lambda n\frac{\partial}{\partial w}||w||_1 \end{aligned} \tag{2.2}$$

Since $X^\top X = I$,

$$w = X^\top Y - \lambda n\frac{\partial}{\partial w}||w||_1$$

Considering the component w^j

$$\begin{aligned} w^j &= \sum_{i=1}^{n} x_i^j y_i - \lambda n\frac{\partial}{\partial w}||w||_1 \\ &= y^i - \lambda n\frac{\partial}{\partial w^j}\sum_{i=1}^{d}|w^j| \\ &= y^i - \lambda n\frac{\partial}{\partial w^j}|w^j| \\ &= y^i(1 - \frac{\lambda n}{y^j}\frac{\partial}{\partial w^j}|w^j|) \end{aligned} \tag{2.3}$$

The derivative of $|w^j|$ is

$$\frac{\partial}{\partial w^j}|w^j| = \begin{cases} 1, & \text{if } w^j > 0 \\ -1, & \text{if } w^j < 0 \end{cases} = sign(w^j) \tag{2.4}$$

and in the case $w^j = 0$, then the weight is trivially 0. But since λ and n are positive, $sign(w^j) = sign(y^j) = \frac{y^j}{|y^i|}$ Then

$$w^j = 0 \text{ or } w^j = y^j(1 - \frac{\lambda n}{y^j}\frac{y^j}{|y^j|}) = y^j(1 - \frac{\lambda n}{|y^j|})$$

therefore,

$$w^j = y^j \max\{0, 1 - \frac{\lambda n}{|y^j|}\}$$

1.2

Now, assume

$$\sum_{i=1}^{n} x_i x_i^T = diag(\sigma_1^2, \ldots, \sigma_d^2)$$

or

$$X^\top X = \Sigma_{dxd}^2$$

Then, using the same derivative as above,

$$\begin{aligned} 0 &= X^\top X w - X^\top Y + \lambda n \frac{\partial}{\partial w}||w||_1 \\ \Sigma^2 w &= X^\top Y - \lambda n \frac{\partial}{\partial w}||w||_1 \\ w &= \Sigma^{-2}(X^\top Y + \lambda n \frac{\partial}{\partial w}||w||_1) \end{aligned} \tag{2.5}$$

Then,

$$\begin{aligned} w^j &= \sum_{i=1}^{n} \frac{1}{\sigma_j^2}(x_i^j y_i) + \Sigma^{-2}\lambda n \frac{\partial}{\partial w}||w||_1 \\ &= \frac{1}{\sigma_j^2} y^j + \frac{\lambda n}{\sigma_j^2}\frac{\partial}{\partial w}||w||_1 \end{aligned} \tag{2.6}$$

Using the same reasoning, the $sign(w^j) = sign(y^j)$ since $\sigma_j^2 > 0$, then

$$w^j = \frac{y^j}{\sigma_j^2} \max\{0, 1 - \frac{\lambda n}{|y^j|}\}$$

1.3

Let

$$X = U\Sigma V^\top$$

and consider the problem

$$\min_{w \in \mathcal{R}^d}\{\frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 + 2\lambda||V^\top w||\}$$

Let $\hat{w} = V^\top w$, and

$$\begin{aligned} L(\hat{w}) &= \frac{1}{n}(Xw - Y)^\top (Xw - Y) + 2\lambda n||\hat{w}|| \\ &= \frac{1}{n}(U\Sigma V^\top w - Y)^\top (U\Sigma V^\top w - Y) + 2\lambda n||\hat{w}|| \\ &= \frac{1}{n}(U\Sigma\hat{w} - Y)^\top (U\Sigma\hat{w} - Y) + 2\lambda n||\hat{w}|| \end{aligned} \tag{2.7}$$

Then,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\hat{w}} L(\hat{w}) = \frac{2}{n}\Sigma U^\top (U\Sigma\hat{w} - Y) + 2\lambda n \frac{\partial}{\partial\hat{w}}||\hat{w}|| \\ 0 &= \Sigma^2\hat{w} - \Sigma U^\top Y + \lambda n \frac{\partial}{\partial\hat{w}}||\hat{w}|| \\ \Sigma^2\hat{w} &= \Sigma U^\top Y - \lambda n \frac{\partial}{\partial\hat{w}}||\hat{w}|| \\ \hat{w} &= \Sigma^{-1} U^\top Y - \lambda n \Sigma^{-2} \frac{\partial}{\partial\hat{w}}||\hat{w}|| \\ V^\top w &= \Sigma^{-1} U^\top Y - \lambda n \Sigma^{-2} \frac{\partial}{\partial\hat{w}}||\hat{w}|| \\ w &= V\Sigma^{-1} U^\top Y - \lambda n V \Sigma^{-2} \frac{\partial}{\partial\hat{w}}||\hat{w}|| \end{aligned} \tag{2.8}$$

following the same reasoning as the previous problems

$$\begin{aligned} \frac{\partial}{\partial\hat{w}}||\hat{w}|| &= sign(\hat{w}) \\ &= sign(\frac{u_i^\top Y}{\sigma_i}) \\ &= \frac{\frac{u_i^\top Y}{\sigma_i}}{|\frac{u_i^\top Y}{\sigma_i}|} \end{aligned} \tag{2.9}$$

Finally,

$$w = \sum_{i=1}^{m} \frac{1}{\sigma_i} u_i^\top Y \max\{0, 1 - \frac{\lambda n}{\sigma_i^2} \frac{1}{|\frac{u_i^\top Y}{\sigma_i}|}\} v_i$$

1.4

Let us reformulate the minimization problem as such

$$\min_{w\in\mathcal{R}^d} \frac{1}{n}(Xw - Y)^\top (Xw - Y) + \lambda||w||_1$$

Where $E(w) = \frac{1}{n}(Xw - Y)^\top (Xw - Y)$ is differentiable, convex function and $R(w) = \lambda||w||_1$ Then,

$$\frac{\partial}{\partial w} E(w) = \frac{2}{n} X^\top (Xw - Y)$$

Then,

$$\begin{aligned} w_{t+1} &= w_t - \gamma \frac{\partial}{\partial w} E(w) - \gamma \frac{\partial}{\partial w} R(W) \\ &= w_t - \frac{2\gamma}{n} X^\top (X w_t - Y) - \gamma\lambda \frac{\partial}{\partial w} ||w||_1 \\ &= w_t - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i (w_t^\top x_i - y_i) - \gamma\lambda ||w_t||_1 \end{aligned} \tag{2.10}$$

Let us consider a component w_{t+1}^j

$$w_{t+1}^j = w_t^j - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i^j (w_t^\top x_i - y_i) - \gamma\lambda \frac{\partial}{\partial w^j} |w_t^j|$$

If $w_t^j > 0$, then

$$w_t^j - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i^j (w_t^\top x_i - y_i) > \gamma\lambda$$

Such that the w_{t+1}^j stays in the positive region, If $w_t^j < 0$, then

$$w_t^j - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i^j (w_t^\top x_i - y_i) < -\gamma\lambda$$

Such that the w_{t+1}^j stays in the negative region For convergence, γ has to be chosen accordingly. Because of the subgradient of the l1-norm, if $w_t^j = 0$, then

$$-\gamma\lambda \leq w_t^j - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i^j (w_t^\top x_i - y_i) \leq \gamma\lambda$$

. Therefore

$$w_{t+1} = prox_{\gamma\lambda||\cdot||_1}(w_t - \frac{2\gamma}{n} \sum_{i=1}^{n} x_i^j (w_t x_i - y_i))$$

Problem 2 2.1

Let us consider

$$\min_{w \in \mathcal{R}^d} ||Xw - Y||^2$$

and $w_t = w_{t-1} - 2\gamma X^T (X w_{t-1} - Y)$ and $w_0 = 0$ Proving $w_t = X^\top c_t$ by induction, starting with the base case

$$\begin{aligned} w_1 &= w_0 - 2\gamma X^\top (X w_0 - Y) \\ w_1 &= X^\top (2\gamma Y) \end{aligned} \tag{2.11}$$

and let $c_1 = 2\gamma Y$, such that $w_1 = X^\top c_1$ is of the form to be proved. Assume that

$$w_t = X^\top c_t$$

is true, then

$$\begin{aligned} w_{t+1} &= w_t - 2\gamma X^\top (Xw_t - Y) \\ &= X^\top c_t - 2\gamma X^\top (XX^\top c_t - Y) \\ &= X^\top (c_t - 2\gamma (XX^\top c_t - Y)) \\ &= X^\top c_{t+1} \end{aligned} \tag{2.12}$$

Therefore, the recursive definition for c_t is

$$c_t = c_{t-1} - 2\gamma (XX^\top c_{t-1} - Y)$$

This proof by induction is concluded.

2.2

Consider

$$\begin{aligned} &\min_{w \in \mathcal{R}^d} ||Xw - Y||^2 \\ 0 &= 2X^\top (Xw - Y) \\ 0 &= 2X^\top Xw - Y \\ X^\top Xw &= X^\top Y \\ w &= (X^\top X)^{-1} X^\top Y \\ &= X^\top (XX^\top)^{-1} Y \end{aligned} \tag{2.13}$$

Then $w = X^\top c$ and $c = (XX^\top)^{-1} Y$ Then

$$\min_{c \in \mathcal{R}^n} ||XX^\top c - Y||^2$$

substituting w for the result obtained earlier. Now let us differentiate with respect to c

$$\frac{\partial}{\partial c} (XX^\top - Y)^\top (XX^\top - Y) = X^\top X (XX^\top c - Y) \tag{2.14}$$

Then, gradient descent for c_t is

$$c_t = c_{t-1} - \gamma X^\top X (XX^\top c_{t-1} - Y)$$

Considering that X is $n \times d$, if $d < n$ then the gradient descent rule given by the representer theorem is more efficient. If $n < d$ then the gradient descent rule given by differentiating with respect to c is more efficient.

2.3

Let us consider now a general loss convex function

$$\min_{w \in \mathcal{R}^d} l(w^\top x, y)$$

Then

$$\begin{aligned} w_{t+1} &= w_t - \gamma \frac{\partial}{\partial w} l(w^\top x, y) \\ &= w_t - \gamma X^\top \frac{\partial}{\partial u} l(u, y) \\ &= X^\top c_{t+1} \end{aligned} \tag{2.15}$$

and for

$$\min_{c\in\mathcal{R}^n} l(XX^\top c, y)$$

Differentiating with respect c

$$c_{t+1} = c_t - \gamma X^\top X \frac{\partial}{\partial u} l(u, y) \tag{2.16}$$

2.4
Let us consider

$$w_t = w_{t-1} - \gamma x_t(x_t^\top w_{t-1} - y_t)$$

Assume $w_0 = 0$, then the base case is

$$\begin{aligned} w_1 &= w_0 - \gamma x_1(x_1^\top w_0 - y_1) \\ &= -\gamma x_1(-y_1) \\ &= \gamma x_1 y_1 \end{aligned} \tag{2.17}$$

However, $x_1 = X^\top e_1$, where e_t is the standard basis where it is 1 in position t and 0 elsewhere, of size$n$. Then

$$w_1 = X^\top(\gamma e_1 y_1) = X^\top c_1$$

Now, assume $w_t = X^\top c_t$ is true, then

$$\begin{aligned} w_{t+1} &= w_t - \gamma x_{t+1}(x_{t+1}^\top w_t - y_{t+1}) \\ &= X^\top c_t - \gamma x_{t+1}(x_{t+1}^\top X^\top c_t - y_t) \end{aligned} \tag{2.18}$$

Using the same reasoning, $x_{t+1} = X^\top e_{t+1}$, then

$$\begin{aligned} w_{t+1} &= X^\top c_t - \gamma X^\top e_{t+1}(e_{t+1}^\top XX^\top c_t - y_t) \\ &= X^\top(c_t - \gamma e_{t+1}(e_{t+1}^\top XX^\top c_t - y_t)) \\ &= X^\top c_{t+1} \end{aligned} \tag{2.19}$$

By induction, $w_t = X^\top c_t$ is true. The definition for c_t is,

$$c_t = c_{t-1} - \gamma e_t(e_t^\top XX^\top c_{t-1} - y_t)$$

This stochastic gradient descent version is less computationally expensive than the two previous versions since the dot product is happening between one vector and the data matrix as opposed to all the dot products calculated from the kernel matrices in the previous versions.

Problem 3 3.1
The solution for the least squares is $w = (X_M^\top X_M)^{-1} X_M^\top Y$. Since $X_M = XV_M = U_M \Sigma_M$

$$\begin{aligned} w &= V_M^\top (X^\top X)^{-1} V_M V_M^\top X^\top Y \\ &= \Sigma_M^{-1} U_M^\top Y \end{aligned} \tag{2.20}$$

Since $x \in \mathcal{R}^d$, then to map it to $\mathcal{R}^m$ we need $X_M = V_M^\top X$ then

$$\begin{aligned} f_M(x) &= w^\top V_M^\top x \\ &= (\Sigma_M^{-1} U_M^\top Y) V_M^\top x \\ &= \sum_{j=1}^{M} \frac{1}{\sigma_j} u_j^\top Y v_j^\top x \end{aligned} \tag{2.21}$$

And $w = V_M \Sigma^{-1} U_M^\top$

3.2

Let $C_M = \sum_{j=1}^{M} \sigma_j^2 v_j v_j^T$ and let $\tilde{w}_M^\top x = C_M^\dagger X^\top Y x$. C_M can also be written as $C_M = V_M \Sigma_M^2 V_M^\top$. Then,

$$\begin{aligned} C_M^\dagger &= (C_M^\top C_M)^{-1} C_M^\top = (V_M \Sigma^4 V_M^\top)^\top V_M \Sigma^2 V_M^\top \\ &= V_M \Sigma_M^{-2} V_M^\top \end{aligned} \tag{2.22}$$

Then

$$\tilde{w}_M^\top = V_M \Sigma_M^{-2} V_M^\top V_M \Sigma_M U_M^\top Y = V_M \Sigma_M^{-1} U_M^\top Y$$

Then,

$$\tilde{w}_M^\top x = \sum_{j=1}^{M} \frac{1}{\sigma_j} u_j^\top Y v_j^\top x = f_M(x)$$

Which was to be proven.

3.3

Consider

$$w_M = V_M \Sigma_M^{-1} U_M^\top Y$$

Then

$$\begin{aligned} f_M(x) &= x^\top w_M = x^\top V_M \Sigma_M^{-1} U_M^\top Y \\ &= x^\top X^\top U_M \Sigma_M^{-1} \Sigma_M^{-1} U_M^\top Y \\ &= x^\top X^\top U_M \Sigma_M^{-2} U_M^T Y \end{aligned} \tag{2.23}$$

However, $x^T X^T$ is a series of inner products between x and the entries of X. The product $U_M \Sigma_M^{-2} U_M^\top$ represents $XX^\top$ diagonalized, which is the Kernel matrix. Therefore,

$$\begin{aligned} f_M(x) &= \sum_{i=1}^{n} k(x, x_i) \frac{1}{\sigma_i^2} u_i (u_i^\top Y) \\ &= \sum_{i=1}^{n} k(x, x_i) (c_M)_i \end{aligned} \tag{2.24}$$