

HW5: Question 2.1 - 2

Ian Dover

March 2023

1 Part 1

Solve the following optimization problem:

$$\min_w ||y - Xw||_2^2 + \frac{\lambda}{2} ||w||_2^2$$

where $y, w \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times n}$.

Drive a closed-form solution for w . A closed form solution for w involves determining the value for w which minimizes the loss function. Set the partial with respect to w to 0:

$$\frac{\partial \left[\frac{1}{2} ||y - Xw||_2^2 + \frac{\lambda}{2} ||w||_2^2 \right]}{\partial w} = 0$$

Use the Chain rule to solve this equation:

$$2 \times \frac{1}{2} \times -X^T \times ||y - Xw||_2 + \frac{\lambda}{2} \times 2 \times ||w||_2 = 0$$
$$-X^T \times ||y - Xw||_2 + \lambda ||w||_2 = 0$$

For the purposes of minimization, we can set the L2-Norm to be parenthesis:

$$-X^T(y - Xw) + \lambda(w) = 0$$

$$-X^T y + X^T X w + \lambda w = 0$$

$$-X^T y + (X^T X + \lambda I) w = 0$$

Solving for w , we get:

$$w = \frac{X^T y}{(X^T X + \lambda I)}$$

Rewriting this, we get:

$$w = X^T y \times (X^T X + \lambda I)^{-1}$$

The objective is to minimize this function with respect to the i th element of w . We must rewrite this equation to take that into account:

$$\begin{aligned}
& -X^T \left[(y_i - X_i w_i) + (y_j - X_j w_j) \right] + \lambda I(w_i + w_j) = 0 \\
& -X_i^T (y_i - X_i w_i) - X_j^T (y_j - X_j w_j) + \lambda I(w_i + w_j) = 0 \\
& -X_i^T y_i + X_i^T X_i w_i - X_j^T y_j + X_j^T X_j w_j + \lambda I(w_i + w_j) = 0 \\
& -X^T y + X_i^T X_i w_i + X_j^T X_j w_j + \lambda I(w_i + w_j) = 0 \\
& -X^T y + X_i^T X_i w_i + X_j^T X_j w_j + \lambda I w_i + \lambda I w_j = 0 \\
& -X^T y + X_j^T X_j w_j + \lambda I w_j + (X_i^T X_i + \lambda I) w_i = 0
\end{aligned}$$

Solving for w_i , we get:

$$\begin{aligned}
w_i &= -\frac{-X^T y + X_j^T X_j w_j + \lambda I w_j}{(X_i^T X_i + \lambda I)} \\
w_i &= \frac{X^T y - X_j^T X_j w_j - \lambda I w_j}{(X_i^T X_i + \lambda I)} \\
w_i &= (X^T y - X_j^T X_j w_j - \lambda I w_j) \times (X_i^T X_i + \lambda I)^{-1}
\end{aligned}$$

We can now use this new formulation for w_i to drive updates in the Coordinate Descent algorithm:

Coordinate Descent Algorithm

01. Select $x_0 \in \mathbf{R}^n$
02. for $k = 1 : K$ do
03. for $i = 1 : n$ do
04. $w_j \leftarrow \left[w_1^k, w_2^k, \dots, w_{i-1}^k, w_{i+1}^{k-1}, \dots, w_n^{k-1} \right]$
05. $w_i = (X^T y - X_j^T X_j w_j - \lambda I w_j) \times (X_i^T X_i + \lambda I)^{-1}$
06. end for
07. end for

This is taken from the common definition of Coordinate Gradient Descent:

Coordinate Descent Algorithm

```

01. Select  $x_0 \in \mathbf{R}^n$ 
02. for  $k = 1 : K$  do
03.     for  $i = 1 : n$  do
04.         Compute  $w_i^k = \underset{w_i}{\operatorname{argmin}} f(w_1^k, w_2^k, \dots, w_{i-1}^k, w_i, w_{i+1}^{k-1}, \dots, w_n^{k-1})$ 
05.     end for
06. end for

```

2 Part 2

Solve the following optimization problem:

$$\min_w \frac{1}{2} \|y - Xw\|_2^2 + \lambda_1 |w|_1 + \frac{\lambda_2}{2} \|w\|_2^2$$

where $\lambda_1 = 0.05$ and $\lambda_2 = 0.01$

For this problem, we must find the value of $w_i \in w$ that minimizes the equation:

$$w_i^k = \underset{w_i}{\operatorname{argmin}} \left\{ \frac{1}{2} \|y - Xw\|_2^2 + \lambda_1 |w|_1 + \frac{\lambda_2}{2} \|w\|_2^2 \right\}$$

This minimization problem is taking by finding the derivative with respect to w_i and setting it to 0:

$$\frac{\partial \left[\frac{1}{2} \|y - Xw\|_2^2 + \lambda_1 |w|_1 + \frac{\lambda_2}{2} \|w\|_2^2 \right]}{\partial w_i} = 0$$

$$X_i^T \|y - Xw\|_2 + \lambda_1 I + \lambda_2 I \|w\|_2 = 0$$

$$X_i^T (y - Xw)_2 + \lambda_1 I + \lambda_2 I(w) = 0$$

$$X_i^T y - X_i^T Xw + \lambda_1 I + \lambda_2 Iw = 0$$

Break the w into $\{w_i, w_j\}$ where (1) $w \in \mathbb{R}^n$, (2) $w_i \in w$ and $w_i \in \mathbb{R}^1$, (3) $w_j \subset w$ and $w_j \in \mathbb{R}^{n-1}$, and (4) $i \neq j$

$$X_i^T y - X_i^T X_i w_i - X_i^T X_j w_j + \lambda_1 I + \lambda_2 I w_i + \lambda_2 I w_j = 0$$

$$X_i^T y - X_i^T X_j w_j + \lambda_1 I + \lambda_2 I w_j + (\lambda_2 I - X_i^T X_i) w_i = 0$$

Let us re-order the elements:

$$X_i^T y - X_i^T X_j w_j + \lambda_2 I w_j + \lambda_1 I + (\lambda_2 I - X_i^T X_i) w_i = 0$$

Solve for w_i :

$$w_i = -\frac{X_i^T y - X_i^T X_j w_j + \lambda_2 I w_j + \lambda_1 I}{\lambda_2 I - X_i^T X_i}$$

$$w_i = -\frac{X_i^T (y - X_j w_j) + \lambda_2 I w_j + \lambda_1 I}{\lambda_2 I - X_i^T X_i}$$

$$w_i = -\frac{X_i^T (y - X_j w_j) + \lambda_2 I w_j}{\lambda_2 I - X_i^T X_i} + \frac{\lambda_1 I}{\lambda_2 I - X_i^T X_i}$$

Now we can set the value of γ_i for soft-thresholding:

$$a = -\frac{X_i^T (y - X_j w_j) + \lambda_2 I w_j}{\lambda_2 I - X_i^T X_i} = (X_i^T (y - X_j w_j) + \lambda_2 I w_j) \times (\lambda_2 I - X_i^T X_i)^{-1}$$

and

$$\gamma_i = \frac{\lambda_1 I}{\lambda_2 I - X_i^T X_i} = (\lambda_1 I) \times (\lambda_2 I - X_i^T X_i)^{-1}$$

where

$$w_i = a + \gamma_i$$

The soft-thresholding is defined by:

$$S_{\gamma_i}(a) = \begin{cases} a - \gamma_i & a > \gamma_i \\ 0 & -\gamma_i < a < \gamma_i \\ a + \gamma_i & a < -\gamma_i \end{cases}$$

The soft-thresholding will be used in each update within the Coordinate Descent algorithm:

Coordinate Descent Algorithm

01. Select $w_0 \in \mathbf{R}^n$
02. for $k = 1 : K$ do
03. for $i = 1 : n$ do
04. $w_j \leftarrow [w_1^k, w_2^k, \dots, w_{i-1}^k, w_{i+1}^{k-1}, \dots, w_n^{k-1}]$
05. $a \leftarrow (X_i^T (y - X_j w_j) + \lambda_2 I w_j) \times (\lambda_2 I - X_i^T X_i)^{-1}$
06. $\gamma_i \leftarrow (\lambda_1 I) \times (\lambda_2 I - X_i^T X_i)^{-1}$
07. $w_i^k = \begin{cases} a - \gamma_i & a > \gamma_i \\ 0 & -\gamma_i < a < \gamma_i \\ a + \gamma_i & a < -\gamma_i \end{cases}$
08. end for
09. end for