Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: October 30, 2020

QUESTION 1

Solution 1

Define

$$f(\boldsymbol{w}) = \sum_{n=1}^{N} |y_n - \boldsymbol{w}^T\boldsymbol{x}_n| + \lambda||\boldsymbol{w}||_1 \tag{1}$$

where $\lambda > 0,\ y_n \in \mathbb{R},\ \boldsymbol{x}_n \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{w} \in \mathbb{R}^{D\times 1}$

We want to show that $f(\boldsymbol{w})$ is convex, i.e,

$$f(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2) \leq tf(\boldsymbol{w}_1) + (1-t)f(\boldsymbol{w}_2) \quad \forall\, t \in [0,1],\ \boldsymbol{w}_1,\ \boldsymbol{w}_2 \in \mathbb{R}^{D\times 1} \tag{2}$$

Note that,

$$f(\boldsymbol{w}) = \sum_{n=1}^{N} |y_n - \boldsymbol{w}^T\boldsymbol{x}_n| + \lambda||\boldsymbol{w}||_1 = \sum_{n=1}^{N} |y_n - \boldsymbol{x}_n^T\boldsymbol{w}| + \lambda \sum_{d=1}^{D} |\boldsymbol{w}_d|$$

Now consider,

$$\begin{aligned}
&f(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)\\
&= \sum_{n=1}^{N} \mid y_n - \boldsymbol{x}_n^T(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2) \mid + \lambda \sum_{d=1}^{D} \mid t(\boldsymbol{w}_1)_d + (1-t)(\boldsymbol{w}_2)_d \mid\\
&= \sum_{n=1}^{N} \mid y_n - t\boldsymbol{x}_n^T\boldsymbol{w}_1 - (1-t)\boldsymbol{x}_n^T\boldsymbol{w}_2 \mid + \lambda \sum_{d=1}^{D} \mid t(\boldsymbol{w}_1)_d + (1-t)(\boldsymbol{w}_2)_d \mid\\
&= \sum_{n=1}^{N} \mid (t + (1-t))y_n - t\boldsymbol{x}_n^T\boldsymbol{w}_1 - (1-t)\boldsymbol{x}_n^T\boldsymbol{w}_2 \mid + \lambda \sum_{d=1}^{D} \mid t(\boldsymbol{w}_1)_d + (1-t)(\boldsymbol{w}_2)_d \mid\\
&= \sum_{n=1}^{N} \mid t(y_n - \boldsymbol{x}_n^T\boldsymbol{w}_1) + (1-t)(y_n - \boldsymbol{x}_n^T\boldsymbol{w}_2) \mid + \lambda \sum_{d=1}^{D} \mid t(\boldsymbol{w}_1)_d + (1-t)(\boldsymbol{w}_2)_d \mid\\
&\leq \sum_{n=1}^{N} [\, \mid t(y_n - \boldsymbol{x}_n^T\boldsymbol{w}_1) \mid + \mid (1-t)(y_n - \boldsymbol{x}_n^T\boldsymbol{w}_2) \mid \,] + \lambda \sum_{d=1}^{D} [\, \mid t(\boldsymbol{w_1})_d \mid + \mid (1-t)(\boldsymbol{w_2})_d \mid \,]\\
&\quad = t \sum_{n=1}^{N} \mid (y_n - \boldsymbol{x}_n^T\boldsymbol{w}_1) \mid + (1-t) \sum_{n=1}^{N} \mid (y_n - \boldsymbol{x}_n^T\boldsymbol{w}_2) \mid \ + \lambda[\, t \sum_{d=1}^{D} \mid (\boldsymbol{w_1})_d \mid + (1-t) \sum_{d=1}^{D} \mid (\boldsymbol{w_2})_d \mid \,]\\
&\quad = t \cdot [\, \sum_{n=1}^{N} \mid y_n - \boldsymbol{x}_n^T\boldsymbol{w}_1 \mid + \lambda \sum_{d=1}^{D} |(\boldsymbol{w}_1)_d| \,] + (1-t) \cdot [\, \sum_{n=1}^{N} \mid y_n - \boldsymbol{x}_n^T\boldsymbol{w}_2 \mid + \lambda \sum_{d=1}^{D} |(\boldsymbol{w}_2)_d| \,]\\
&\quad = tf(\boldsymbol{w}_1) + (1-t)f(\boldsymbol{w}_2)
\end{aligned}$$

Hence, we've shown that $f(\boldsymbol{w})$ is convex

Subgradient

A **subgradient** of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}_0 \in \mathbb{R}^n$ is any $\boldsymbol{g} \in \mathbb{R}^n$ such that

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \boldsymbol{g}^T(\boldsymbol{x} - \boldsymbol{x}_0) \quad \forall\, \boldsymbol{x} \in \mathrm{dom}(f)$$

Collection of all subgradients of f at $\boldsymbol{x}_0$ is denoted by $\partial f(\boldsymbol{x}_0)$, which is called **subdifferential**

We know that for $h : \mathbb{R} \to \mathbb{R} \ni x \mapsto |x|$, we have

$$\partial h(x) = \mathrm{sign}(x) \quad \text{if } x \neq 0 \qquad \text{and} \qquad \partial h(0) = [-1, 1] \tag{3}$$

where $\forall\, x \neq 0$

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Let

$$F : \mathbb{R}^n \to \mathbb{R} \quad \ni \quad \boldsymbol{x} \mapsto ||\boldsymbol{x}||_1 \quad \left(= \sum_{i=1}^{n} |x_i| \right)$$

Then, using (3), we can see that, if $\boldsymbol{g} \in \partial F(\boldsymbol{x})$, then

$$g_i = \mathrm{sign}(x_i) \quad \text{if } x_i \neq 0 \qquad \text{and} \qquad g_i \in [-1, 1] \quad \text{if } x_i = 0$$

Hence, we must have

$$\partial F(\boldsymbol{x}) \supseteq \{\boldsymbol{g} \in \mathbb{R}^n \mid \boldsymbol{g} = \sum_{x_i \neq 0} \mathrm{sign}(x_i)\ \boldsymbol{e}_i^D + \sum_{x_j = 0} c_j\ \boldsymbol{e}_j^D \quad \text{with } c_j \in [-1, 1]\} \tag{4}$$

where $\boldsymbol{e}_j^D = (0_1,\ \ldots 0_{i-1}, 1, 0_{i+1} \ldots 0_D) \in \mathbb{R}^{D \times 1}$

We wish to calculate the subgradient vector of

$$\begin{aligned} f(\boldsymbol{w}) &= \sum_{n\,=\,1}^{N} |y_n - \boldsymbol{w}^T \boldsymbol{x}_n| + \lambda ||\boldsymbol{w}||_1 \\ &= \sum_{n\,=\,1}^{N} |y_n - \boldsymbol{x}_n^T \boldsymbol{w}| + \lambda ||\boldsymbol{w}||_1 \\ &= \sum_{n\,=\,1}^{N} |\boldsymbol{x}_n^T \boldsymbol{w} - y_n| + \lambda ||\boldsymbol{w}||_1 \end{aligned} \tag{5}$$

Using subdifferential calculus, we have

$$
\begin{aligned}
\partial f(\boldsymbol{w}) &= \partial(\sum_{n\,=\,1}^{N} |\boldsymbol{x}_n^T \boldsymbol{w} - y_n| + \lambda ||\boldsymbol{w}||_1 \ ) \\
&= \partial \sum_{n\,=\,1}^{N} |\boldsymbol{x}_n^T \boldsymbol{w} - y_n| + \lambda \ \partial ||\boldsymbol{w}||_1 \\
&= \sum_{n\,=\,1}^{N} \partial |\boldsymbol{x}_n^T \boldsymbol{w} - y_n| + \lambda \ \partial ||\boldsymbol{w}||_1 \\
&= \sum_{n\,=\,1}^{N} \boldsymbol{x_n} \partial |t_n| + \lambda \ \partial ||\boldsymbol{w}||_1 \ \text{ where } t_n = \boldsymbol{x}_n^T \boldsymbol{w} - y_n \\
&\supseteq \{ \sum_{n=1}^{N} c_n \ \boldsymbol{x_n} + \lambda \sum_{w_i \neq 0} \text{sign}(w_i) \ \boldsymbol{e}_i^D + \lambda \sum_{w_j = 0} k_j \ \boldsymbol{e}_j^D \ \} \qquad (6) \\
&\text{where } c_n = \text{sign}(t_n) \ \text{ if } t_n \neq 0 \quad \text{else } c_n \in [-1, 1], \ \text{ and } k_j \in [-1, 1]
\end{aligned}
$$

We can always choose $c_n = 0$ when $t_n = 0$ and $k_j = 0$ to get the required **(sub)gradient** vector

$$
\boldsymbol{g} = \sum_{\boldsymbol{x}_n^T \boldsymbol{w} - y_n \neq 0} \text{sign}(\boldsymbol{x}_n^T \boldsymbol{w} - y_n) \ \boldsymbol{x_n} + \lambda \sum_{w_i \neq 0} \text{sign}(w_i) \ \boldsymbol{e}_i^D \ \ (\in \mathbb{R}^{D \times 1}) \qquad (7)
$$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: October 30, 2020

QUESTION 2

Solution 2

Note that,

$$\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T x_i)^2 = (\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w})$$

where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, $X \in \mathbb{R}^{N\times D}$ and $\boldsymbol{w} \in \mathbb{R}^{D\times 1}$

We want to replace x_i by $\tilde{x_i}$ where $\tilde{x_i} = x_i \circ m_i$, where $m_{id} \in \{0,1\}$ and $m_{id} \sim \text{Bern}(p)$

Let us define $M \in \mathbb{R}^{N\times D}$ where $M_{ij} \in \{0,1\}$ and $M_{ij} \sim \text{Bern}(p)$

Then similar to above, we can essentially write

$$\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T \tilde{x_i})^2 = (\boldsymbol{y} - (X\circ M)\boldsymbol{w})^T(\boldsymbol{y} - (X\circ M)\boldsymbol{w}) \tag{1}$$

For convenience, let $L = X \circ M$ and consider,

$$\begin{aligned}
&E_M[(\boldsymbol{y} - (X\circ M)\boldsymbol{w})^T(\boldsymbol{y} - (X\circ M)\boldsymbol{w})]\\
&= E_M[(\boldsymbol{y} - L\boldsymbol{w})^T(\boldsymbol{y} - L\boldsymbol{w})]\\
&= E_M[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T L^T \boldsymbol{y} + \boldsymbol{w}^T L^T L\boldsymbol{w}]\\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T E_M(L^T)\boldsymbol{y} + \boldsymbol{w}^T E(L^T L)\boldsymbol{w}\\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T (E_M(L))^T\boldsymbol{y} + \boldsymbol{w}^T E(L^T L)\boldsymbol{w}
\end{aligned} \tag{2}$$

Now, note that,

$$\begin{aligned}
[E_M(L)]_{ij} &= E_M(L_{ij})\\
&= E(X_{ij}M_{ij})\\
&= pX_{ij}\\
\Rightarrow E_M(L) &= pX\\
\Rightarrow 2\boldsymbol{w}^T(E_M(L))^T\boldsymbol{y} &= 2p\boldsymbol{w}^T X^T \boldsymbol{y}
\end{aligned} \tag{3}$$

Also, note that,

$$\begin{aligned}
[E_M(L^TL)]_{ij} &= E_M[(L^TL)_{ij}] \\
&= E_M[\sum_{k=1}^{N}(L^T)_{ik}L_{kj}] \\
&= E_M[\sum_{k=1}^{N} L_{ki}L_{kj}] \\
&= E_M[\sum_{k=1}^{N} X_{ki}M_{ki}X_{kj}M_{kj}] \\
&= E_M[\sum_{k=1}^{N} X_{ki}X_{kj}M_{ki}M_{kj}] \\
&= \sum_{k=1}^{N} X_{ki}X_{kj}E_M[M_{ki}M_{kj}] \qquad (4)
\end{aligned}$$

When $i = j$, $E_M[M_{ki}M_{kj}] = E_M[M_{ki}^2] = V_M(M_{ki}) + E(M_{ki})^2 = p(1-p) + p^2 = p$

and when $i \neq j$, $E_M[M_{ki}M_{kj}] = E_M[M_{ki}]E_M[M_{kj}] = p^2$

Therefore, using these observations, we have

When $i = j$, $[E_M(L^TL)]_{ij} = p\sum\limits_{k=1}^{N} X_{ki}^2 = p\sum\limits_{k=1}^{N}(X^T)_{ik}X_{ki} = p(X^TX)_{ii} = p(X^TX)_{ij}$

and when $i \neq j$, $[E_M(L^TL)]_{ij} = p^2\sum\limits_{k=1}^{N} X_{ki}X_{kj} = p^2\sum\limits_{k=1}^{N}(X^T)_{ik}X_{kj} = p^2(X^TX)_{ij}$

Therefore,

$$[E_M(L^TL)]_{ii} = p(X^TX)_{ii} \quad \text{and} \quad [E_M(L^TL)]_{ij} = p^2(X^TX) \;\; (i \neq j) \qquad (5)$$

Using $(2), (3)$ and (5), we get

$$\begin{aligned}
&E_M[(\boldsymbol{y} - (X \circ M)\boldsymbol{w})^T(\boldsymbol{y} - (X \circ M)\boldsymbol{w})] \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T(E_M(L))^T\boldsymbol{y} + \boldsymbol{w}^TE(L^TL)\boldsymbol{w} \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2p\boldsymbol{w}^TX^T\boldsymbol{y} + \boldsymbol{w}^TE_M(L^TL)\boldsymbol{w} \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2p\boldsymbol{w}^TX^T\boldsymbol{y} + p^2\boldsymbol{w}^TX^TX\boldsymbol{w} - p^2\boldsymbol{w}^TX^TX\boldsymbol{w} + \boldsymbol{w}^TE_M(L^TL)\boldsymbol{w} \\
&= (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + \boldsymbol{w}^T(E_M(L^TL) - p^2X^TX]\boldsymbol{w} \qquad (6)
\end{aligned}$$

Now note that, from (5),

$$[E_M(L^TL) - p^2X^TX]_{ii} = p(1-p)(X^TX)_{ii} \text{ and } [E_M(L^TL) - p^2X^TX]_{ij} = 0 \;\; (i \neq j)$$

and therefore,

$$E_M(L^TL) - p^2X^TX = p(1-p)\text{diag}(X^TX) \tag{7}$$

Finally, using (6) and (7), we obtain

$$\begin{aligned} & E_M[(\boldsymbol{y} - (X \circ M)\boldsymbol{w})^T(\boldsymbol{y} - (X \circ M)\boldsymbol{w})] \\ & = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + \boldsymbol{w}^T(E_M(L^TL) - p^2X^TX]\boldsymbol{w} \\ & = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)\boldsymbol{w}^T\text{diag}(X^TX)\boldsymbol{w} \\ & = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)[\text{diag}(X^TX)^{1/2}]^T[\text{diag}(X^TX)^{1/2}]\boldsymbol{w} \\ & = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)[\text{diag}(X^TX)^{1/2}\boldsymbol{w}]^T[\text{diag}(X^TX)^{1/2}\boldsymbol{w}] \end{aligned} \tag{1}$$

Hence, we get

$$E_M[\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\tilde{x_i})^2] = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)[\text{diag}(X^TX)^{1/2}\boldsymbol{w}]^T[\text{diag}(X^TX)^{1/2}\boldsymbol{w}] \tag{9}$$

Therefore,

$$\begin{aligned} & \arg\,\min_{\boldsymbol{w}}[E_M[\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\tilde{x_i})^2]] \\ & = \arg\,\min_{\boldsymbol{w}}[(\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)[\text{diag}(X^TX)^{1/2}\boldsymbol{w}]^T[\text{diag}(X^TX)^{1/2}\boldsymbol{w}]] \end{aligned} \tag{10}$$

which is equivalent to **Ridge Regression** and the required regularized loss function is

$$L(\boldsymbol{w}) = (\boldsymbol{y} - pX\boldsymbol{w})^T(\boldsymbol{y} - pX\boldsymbol{w}) + p(1-p)[\text{diag}(X^TX)^{1/2}\boldsymbol{w}]^T[\text{diag}(X^TX)^{1/2}\boldsymbol{w}] \tag{11}$$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: October 30, 2020

QUESTION 3

Solution 3

$$\begin{aligned}
\mathrm{TR}[(\boldsymbol{Y}-\boldsymbol{XW})^T(\boldsymbol{Y}-\boldsymbol{XW})] &= \sum_{i=1}^{M}[(\boldsymbol{Y}-\boldsymbol{XW})^T(\boldsymbol{Y}-\boldsymbol{XW})]_{ii} \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}[(\boldsymbol{Y}-\boldsymbol{XW})^T]_{ik}(\boldsymbol{Y}-\boldsymbol{XW})_{ki} \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}(\boldsymbol{Y}-\boldsymbol{XW})_{ki}(\boldsymbol{Y}-\boldsymbol{XW})_{ki} \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}[(\boldsymbol{Y}-\boldsymbol{XW})_{ki}]^2 \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}[\boldsymbol{Y}_{ki}-(\boldsymbol{XW})_{ki}]^2 \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}\left[y_{ki}-(\boldsymbol{X}[k,:]\boldsymbol{W}[:,i])\right]^2 \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}\left[y_{ki}-(\boldsymbol{x}_k^T\boldsymbol{w}_i)\right]^2 \\
&= \sum_{i=1}^{M}\sum_{k=1}^{N}\left[y_{ki}-\boldsymbol{w}_i^T\boldsymbol{x}_k\right]^2 \\
&= \sum_{k=1}^{N}\sum_{i=1}^{M}\left[y_{ki}-\boldsymbol{w}_i^T\boldsymbol{x}_k\right]^2 \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M}\left[y_{nm}-\boldsymbol{w}_m^T\boldsymbol{x}_n\right]^2
\end{aligned}$$

Hence, we have shown that,

$$\sum_{n=1}^{N}\sum_{m=1}^{M}\left[y_{nm}-\boldsymbol{w}_m^T\boldsymbol{x}_n\right]^2 = \mathrm{TR}[(\boldsymbol{Y}-\boldsymbol{XW})^T(\boldsymbol{Y}-\boldsymbol{XW})] \tag{1}$$

Now, we assume that,

$$\boldsymbol{W} = \boldsymbol{B}\boldsymbol{S} \tag{2}$$

where $\boldsymbol{B} \in \mathbb{R}^{D\times K}$ and $\boldsymbol{S} \in \mathbb{R}^{K\times M}$

Note that, $\forall\, m \in \{1, 2 \ldots M\}$

$$\begin{aligned}\boldsymbol{w}_m &= \boldsymbol{W}[:, m]\\ &= (\boldsymbol{B}\boldsymbol{S})[:, m]\\ &= \boldsymbol{B}S[:, m]\\ &= \sum_{i=1}^{K} s_{im}B[:, i] \qquad (3)\end{aligned}$$

Hence, we note that, the columns of $\boldsymbol{W}$, that is, $\boldsymbol{w}_m \;\; (m = 1, 2 \ldots M)$ can be written as a linear combination of the K columns of $\boldsymbol{B}$ (4)

Using (1) and (2), out optimization problem now becomes

$$\{\hat{\boldsymbol{B}}, \hat{\boldsymbol{S}}\} = \text{arg min}_{\boldsymbol{B},\boldsymbol{S}}\text{TR}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})] \tag{5}$$

We will try to learn $\boldsymbol{B}$ and $\boldsymbol{S}$ using **Alternating Optimization**

Define

$$\begin{aligned}L(\boldsymbol{B}, \boldsymbol{S}) &= \text{TR}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})]\\ &= \text{TR}[(\boldsymbol{Y}^T - \boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})]\\ &= \text{TR}(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S})\\ &= \text{TR}(\boldsymbol{Y}^T\boldsymbol{Y}) - \text{TR}(\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y}) - \text{TR}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}) + \text{TR}(\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S})\\ &= \text{TR}(\boldsymbol{Y}^T\boldsymbol{Y}) - 2\;\text{TR}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}) + \text{TR}(\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}) \qquad (6)\end{aligned}$$

Given $\boldsymbol{B}^t$ and $\boldsymbol{S}^t$, we will first try to solve the following two sub-problems:

$$\boldsymbol{B}^{t+1} = \text{arg min}_{\boldsymbol{B}}L(\boldsymbol{B}, \boldsymbol{S}^t) \tag{6.1}$$

and

$$\boldsymbol{S}^{t+1} = \text{arg min}_{\boldsymbol{S}}L(\boldsymbol{B}^{t+1}, \boldsymbol{S}) \tag{6.2}$$

$$\begin{aligned}\nabla_{\boldsymbol{B}}L(\boldsymbol{B}, \boldsymbol{S}) &= \nabla_{\boldsymbol{B}}[\text{TR}(\boldsymbol{Y}^T\boldsymbol{Y}) - 2\;\text{TR}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}) + \text{TR}(\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S})]\\ &= -2\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{S}^T + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^T \qquad (6.3)\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{S}}L(\boldsymbol{B}, \boldsymbol{S}) &= \nabla_{\boldsymbol{S}}[\text{TR}(\boldsymbol{Y}^T\boldsymbol{Y}) - 2\;\text{TR}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}) + \text{TR}(\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S})]\\ &= -2\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S} \qquad (6.4)\end{aligned}$$

Using (6.3) and (6.1) we have

$$\begin{aligned} &\nabla_{\boldsymbol{B}} L(\boldsymbol{B}, \boldsymbol{S}) = 0 \\ \Rightarrow &-2\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{S}^T + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^T = 0 \\ \Rightarrow &\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^T = \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{S}^T \\ \Rightarrow &\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{S}^T)^{-1} \\ \Rightarrow &\text{arg min}_{\boldsymbol{B}} L(\boldsymbol{B}, \boldsymbol{S}^t) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{S}^t)^T(\boldsymbol{S}^t(\boldsymbol{S}^t)^T)^{-1} \end{aligned} \tag{6.5}$$

Using (6.4) and (6.2) we have

$$\begin{aligned} &\nabla_{\boldsymbol{S}} L(\boldsymbol{B}, \boldsymbol{S}) = 0 \\ \Rightarrow &-2\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S} = 0 \\ \Rightarrow &\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}\boldsymbol{S} = \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} \\ \Rightarrow &\boldsymbol{S} = (\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{Y} \\ \Rightarrow &\boldsymbol{S} = [(\boldsymbol{X}\boldsymbol{B})^T\boldsymbol{X}\boldsymbol{B}]^{-1}(\boldsymbol{X}\boldsymbol{B})^T\boldsymbol{Y} \\ \Rightarrow &\text{arg min}_{\boldsymbol{S}} L(\boldsymbol{B}^{t+1}, \boldsymbol{S}) = [(\boldsymbol{X}\boldsymbol{B}^{t+1})^T\boldsymbol{X}\boldsymbol{B}^{t+1}]^{-1}(\boldsymbol{X}\boldsymbol{B}^{t+1})^T\boldsymbol{Y} \end{aligned} \tag{6.6}$$

Hence, the required Alternating Optimization algorithm is:

Optimization problem:

$$\{\hat{\boldsymbol{B}}, \hat{\boldsymbol{S}}\} = \text{arg min}_{\boldsymbol{B},\boldsymbol{S}} L(\boldsymbol{B}, \boldsymbol{S}) = \text{arg min}_{\boldsymbol{B},\boldsymbol{S}} \text{TR}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{S})]$$

1. Initialise $\boldsymbol{S}^0,\ t = 0$
2. Update

$$\begin{aligned} \boldsymbol{B}^{t+1} &= \text{arg min}_{\boldsymbol{B}} L(\boldsymbol{B}, \boldsymbol{S}^t) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{S}^t)^T(\boldsymbol{S}^t(\boldsymbol{S}^t)^T)^{-1} \end{aligned} \tag{7}$$

$$\begin{aligned} \boldsymbol{S}^{t+1} &= \text{arg min}_{\boldsymbol{S}} L(\boldsymbol{B}^{t+1}, \boldsymbol{S}) \\ &= [(\boldsymbol{X}\boldsymbol{B}^{t+1})^T\boldsymbol{X}\boldsymbol{B}^{t+1}]^{-1}(\boldsymbol{X}\boldsymbol{B}^{t+1})^T\boldsymbol{Y} \end{aligned} \tag{8}$$

3. $t = t + 1$
4. Goto Step 2 if not yet converged.

Therefore, (7) and (8) provide the required updates for Alternating Optimization algorithm

Since we have $K < \min\{D, M\}$, it becomes clear from (7) and (8) that the subproblem

$$\boldsymbol{B}^{t+1} = \text{arg min}_{\boldsymbol{B}} L(\boldsymbol{B}, \boldsymbol{S}^t)$$

is computationally more demanding and difficult to solve than the subproblem

$$\boldsymbol{S}^{t+1} = \text{arg min}_{\boldsymbol{S}} L(\boldsymbol{B}^{t+1}, \boldsymbol{S})$$

as in (7) we need to invert $\boldsymbol{X}^T\boldsymbol{X}$ which is a $D \times D$ matrix
whereas in (8) we need to invert $(\boldsymbol{X}\boldsymbol{B})^T\boldsymbol{X}\boldsymbol{B}$ which is only a $K \times K$ matrix with $K < D$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: October 30, 2020

QUESTION 4

Solution 4

For a given loss function $L(\boldsymbol{w})$ of unknown weight vector $\boldsymbol{w} \in \mathbb{R}^{D\times 1}$, the **Newton's method** minimises the second order approximation of $L(\boldsymbol{w})$, that is, in order to get the next updated weight vector $\boldsymbol{w^{t+1}}$ from $\boldsymbol{w^t}$, we solve

$$\boldsymbol{w^{t+1}} = \text{arg min}_{\boldsymbol{w}}[L(\boldsymbol{w^t}) + \nabla L(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w^t})^T\nabla^2 L(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t)] \tag{1*}$$

Let

$$f(\boldsymbol{w}) = L(\boldsymbol{w^t}) + \nabla L(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w^t})^T\nabla^2 L(\boldsymbol{w}^t)(\boldsymbol{w}-\boldsymbol{w}^t) \tag{2}$$

Note that, $L : \mathbb{R}^D \to \mathbb{R} \ni \boldsymbol{w} \mapsto L(\boldsymbol{w})$

Hence

$$\nabla L(\boldsymbol{w}) \in \mathbb{R}^{1\times D} \text{ and } \nabla^2 L(\boldsymbol{w}) \in \mathbb{R}^{D\times D} \text{ which is the } \textbf{Hessian matrix} \tag{3**}$$

For convenience, denote $\nabla L(\boldsymbol{w}^t)$ by g^t and denote $\nabla^2 L(\boldsymbol{w}^t)$ by H^t and note that $(H^t)^T = H^t$

From (2) we have

$$\begin{aligned} f(\boldsymbol{w}) &= L(\boldsymbol{w^t}) + \nabla L(w^t)(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w^t})^T\nabla^2 L(\boldsymbol{w}^t)(w-w^t) \\ &= L(\boldsymbol{w}^t) + g^t(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^t)^T H^t(\boldsymbol{w}-\boldsymbol{w}^t) \\ &= L(\boldsymbol{w^t}) + g^t(\boldsymbol{w}-\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w}^T H^t\boldsymbol{w} - 2(\boldsymbol{w}^t)^T H^t\boldsymbol{w} + (\boldsymbol{w}^t)^T H^t\boldsymbol{w}^t) \end{aligned}$$

$$\begin{aligned} &\Rightarrow \nabla f(\boldsymbol{w}) = 0 + g^t + \frac{1}{2}(2\boldsymbol{w}^T H^t - 2(\boldsymbol{w}^t)^T H^t + 0) \\ &\Rightarrow \nabla f(\boldsymbol{w}) = g^t + \boldsymbol{w}^T H^t - (\boldsymbol{w}^t)^T H^t \end{aligned}$$

$$\begin{aligned} &\nabla f(\boldsymbol{w}) = 0 \\ &\Rightarrow \boldsymbol{w}^T H^t = (\boldsymbol{w}^t)^T H^t - g^t \\ &\Rightarrow \boldsymbol{w}^T = (\boldsymbol{w}^t)^T - g^t(H^t)^{-1} \\ &\Rightarrow \boldsymbol{w} = \boldsymbol{w^t} - (H^t)^{-1}(g^t)^T \quad \text{using } [(H^t)^{-1}]^T = [(H^t)^T]^{-1} = (H^t)^{-1} \\ &\Rightarrow \boldsymbol{w} = \boldsymbol{w^t} - (H^t)^{-1}[\nabla L(\boldsymbol{w^t})]^T \end{aligned} \tag{4}$$

* Contrary to what is used in slides, I have used $\nabla L(\boldsymbol{w})$ instead of $[L(\boldsymbol{w})]^T$ because of (3**)
** Convention followed in Calculus on Manifolds, Michael Spivak, Theorem 2-7

Using (1$*$) and (4) we get

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - (H^t)^{-1}[\nabla L(\boldsymbol{w}^t)]^T \tag{6}$$

Newton Method's update for the given model

We are given that,

$$\begin{aligned}
L(\boldsymbol{w}) &= \frac{1}{2}[(\boldsymbol{y} - X\boldsymbol{w})^T(\boldsymbol{y} - X\boldsymbol{w}) + \lambda \boldsymbol{w}^T\boldsymbol{w}] && (7)\\
&= \frac{1}{2}[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{w} + \boldsymbol{w}^T X^T X\boldsymbol{w} + \lambda \boldsymbol{w}^T\boldsymbol{w}]\\
\Rightarrow \nabla L(\boldsymbol{w}) &= \frac{1}{2}[0 - 2\boldsymbol{y}^T X + 2\boldsymbol{w}^T X^T X + 2\lambda \boldsymbol{w}^T]\\
\Rightarrow \nabla[L(\boldsymbol{w^t})]^T &= X^T X\boldsymbol{w}^t + \lambda\boldsymbol{w}^t - X^T\boldsymbol{y} && (8)\\
\Rightarrow \nabla^2 L(\boldsymbol{w}) &= X^T X + \lambda I_D\\
\Rightarrow \nabla^2 L(\boldsymbol{w^t}) &= X^T X + \lambda I_D && (9)
\end{aligned}$$

Using (6), (8) and (9), we get the Newton Method's update for our model as follows:

$$\begin{aligned}
\boldsymbol{w}^{t+1} &= \boldsymbol{w}^t - (X^T X + \lambda I_D)^{-1}[X^T X\boldsymbol{w}^t + \lambda\boldsymbol{w}^t - X^T\boldsymbol{y}]\\
&= \boldsymbol{w}^t - (X^T X + \lambda I_D)^{-1}[(X^T X + \lambda I_D)\boldsymbol{w}^t - X^T\boldsymbol{y}]\\
&= (X^T X + \lambda I_D)^{-1}X^T y && (10)
\end{aligned}$$

(10) gives the required update for Newton's Method.
We see that this update is independent of the input $\boldsymbol{w}^t$

Infact, for input $\boldsymbol{w}^0$, we get $\boldsymbol{w}^1 = (X^T X + \lambda I_D)^{-1}X^T y$
which is the closed form solution of **Ridge Regression** given in (7).

Hence, in the case of Ridge Regression, Newton's Method converges just after **one step**!

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: October 30, 2020

QUESTION 5

Solution 5

We are rolling a six faced die N times.
$N_i = \#$ of times i'th face is obtained
$\pi_i =$ probability of showing i'th face on a dice roll

Note that $\sum_{i=1}^{6} N_i = N$ and $\sum_{i=1}^{6} \pi_i = 1$ where $\pi_i \in (0, 1)$

Likelihood

Let the outcome of N dice rolls be represented as $Y = \{y_1, y_2 \ldots, y_N\}$ where $y_i \in \{1, 2, \ldots 6\}$. Assuming that this data is Independent and Identically Distributed (IID), we have the following likelihood:

$$p(Y|\boldsymbol{\pi}) = \prod_{i=1}^{6} \pi_i^{N_i} \tag{1}$$

where, $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_6\}$

Prior

Note that we have, $\pi_i \in (0, 1)$ and $\sum_{i=1}^{6} \pi_i = 1$

Hence, an appropriate prior conjugate to our likelihood above would be the **Dirichlet** prior given by:

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{6} \pi_i^{\alpha_i - 1} \tag{2}$$

where, $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_6\}$ is a vector of hyperparameters which we assume to be fixed and

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{6} \Gamma\alpha_i}{\Gamma(\sum_{i=1}^{6} \alpha_i)} \tag{3}$$

MLE Estimation

We have,

$$\begin{aligned} LL(\boldsymbol{\pi}) &= \log(p(Y|\boldsymbol{\pi})) \\ &= \log(\prod_{i=1}^{6} \pi_i^{N_i}) \\ &= \sum_{i=1}^{6} N_i \log(\pi_i) \end{aligned} \tag{4}$$

We know that, $\hat{\boldsymbol{\pi}}_{MLE} = \arg\max_{\boldsymbol{\pi}} LL(\boldsymbol{\pi})$ with respect to the constrain $\sum_{i=1}^{6} \pi_i = 1$ where $\hat{\boldsymbol{\pi}}_{MLE} = (\hat{\pi_1}, \hat{\pi_2} \dots \hat{\pi_6})$

We will solve this using **Lagrange multiplier** as follows:

$$l(\boldsymbol{\pi}, \lambda) = \sum_{i=1}^{6} N_i \log(\pi_i) + \lambda(1 - \sum_{i=1}^{6} \pi_i) \tag{5}$$

$$l_\lambda = 0 \Rightarrow 1 - \sum_{i=1}^{6} \pi_i = 0 \Rightarrow \sum_{i=1}^{6} \pi_i = 1 \tag{5.1}$$

$$l_{\pi_j} = 0 \Rightarrow \frac{N_j}{\pi_j} - \lambda = 0 \Rightarrow N_j = \lambda \pi_j \tag{5.2}$$

From (5.1) and (5.2) we have

$$\sum_{j=1}^{6} N_j = \lambda \sum_{j=1}^{6} \pi_j \Rightarrow N = \lambda$$

Substituting this back into (5.2) we get

$$\hat{\pi_j} = \frac{N_j}{N} \quad \forall\, j = 1, 2 \dots 6$$

Hence, we have

$$\hat{\boldsymbol{\pi}}_{MLE} = (\frac{N_1}{N}, \frac{N_2}{N} \dots \frac{N_6}{N}) \tag{6}$$

MAP Estimation

Using (4), we have,

$$\begin{aligned} LL(\boldsymbol{\pi}) + \log(p(\boldsymbol{\pi})) &= \sum_{i=1}^{6} N_i \log(\pi_i) + \log(p(\boldsymbol{\pi})) \\ &= \sum_{i=1}^{6} N_i \log(\pi_i) + \log(\frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{6} \pi_i^{\alpha_i - 1}) \\ &\quad \text{(ignoring any constants, we get)} \\ &= \sum_{i=1}^{6} (N_i + \alpha_i - 1) \log(\pi_i) \qquad (7) \end{aligned}$$

We know that, $\hat{\boldsymbol{\pi}}_{MAP} = \arg\max_{\boldsymbol{\pi}} (LL(\boldsymbol{\pi}) + \log(p(\boldsymbol{\pi}))$ with respect to the constrain $\sum\limits_{i=1}^{6} \pi_i = 1$ where $\hat{\boldsymbol{\pi}}_{MAP} = (\hat{\pi_1}, \hat{\pi_2} \ldots \hat{\pi_6})$

We will solve this using **Lagrange multiplier** as follows:

$$l(\boldsymbol{\pi}, \lambda) = \sum_{i=1}^{6} (N_i + \alpha_i - 1) \log(\pi_i) + \lambda(1 - \sum_{i=1}^{6} \pi_i) \qquad (8)$$

$$l_\lambda = 0 \Rightarrow 1 - \sum_{i=1}^{6} \pi_i = 0 \Rightarrow \sum_{i=1}^{6} \pi_i = 1 \qquad (8.1)$$

$$l_{\pi_j} = 0 \Rightarrow \frac{N_j + \alpha_j - 1}{\pi_j} - \lambda = 0 \Rightarrow N_j + \alpha_j - 1 = \lambda \pi_j \qquad (8.2)$$

From (8.1) and (8.2) we have $\sum_{j=1}^{6} (N_j + \alpha_j - 1) = \lambda \sum\limits_{j=1}^{6} \pi_j \Rightarrow N + \sum\limits_{j=1}^{6} \alpha_j - 6 = \lambda$

Substituting this back into (8.2) we get

$$\hat{\pi_j} = \frac{N_j + \alpha_j - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6} \qquad \forall \; j = 1, 2 \ldots 6$$

Hence, we have

$$\hat{\boldsymbol{\pi}}_{MAP} = (\frac{N_1 + \alpha_1 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6}, \frac{N_2 + \alpha_2 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6} \ldots \frac{N_6 + \alpha_6 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6}) \qquad (9)$$

Posterior

We have,

$$\begin{aligned} p(\boldsymbol{\pi}|Y) &\propto p(Y|\boldsymbol{\pi})p(\boldsymbol{\pi}) \\ &\propto \prod_{i=1}^{6} \pi_i^{N_i} \times \frac{1}{B(\boldsymbol{\alpha})} \times \prod_{i=1}^{6} \pi_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{6} \pi_i^{N_i + \alpha_i - 1} \\ &= \mathrm{Dir}(\pi|\boldsymbol{n} + \boldsymbol{\alpha}) \end{aligned} \tag{10}$$

which gives the required posterior distribution where,

$\boldsymbol{n} = \{N_1, N_2 \dots N_6\}$ and hence, $\boldsymbol{n} + \boldsymbol{\alpha} = \{N_1 + \alpha_1, N_2 + \alpha_2 \dots N_6 + \alpha_6\}$

MAP and MLE Estimation from Posterior

We can obtain MAP Estimation from Posterior using its **mode**.
The mode of $\mathrm{Dir}(\pi|\boldsymbol{n} + \boldsymbol{\alpha})$ where $\boldsymbol{n} + \boldsymbol{\alpha} = \{N_1 + \alpha_1, N_2 + \alpha_2 \dots N_6 + \alpha_6\}$ is given by

$$\textbf{Mode} :(\frac{N_1 + \alpha_1 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6}, \frac{N_2 + \alpha_2 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6} \dots \frac{N_6 + \alpha_6 - 1}{N + \sum\limits_{i=1}^{6} \alpha_i - 6}) \qquad (= \hat{\boldsymbol{\pi}}_{MAP})$$

which is precisely what we obtained in (9)

We can further obtain the MLE Estimation from this by using uniform prior, that is, by substituting above $\alpha_i = 1 \ \forall \, i = 1, 2 \dots 6$ to get

$$\hat{\boldsymbol{\pi}}_{MLE} = (\frac{N_1}{N}, \frac{N_2}{N} \dots \frac{N_6}{N})$$

Q When can we expect MAP solution to be better than MLE?
Since we are quite confident about the distribution of our prior, we can certainly say that MAP solution will always be better than MLE for this model because MAP takes into account regularization, which MLE doesn't. Thus, to prevent overfitting, MAP will be better.