AKSHATA BHAT
abhat6@wisc.edu
CS532 Homework 1

① Let's calculate the j^{th} component of $\nabla_\beta(\beta^T A \beta)$

$$\frac{\partial}{\partial \beta_j}(\beta^T A \beta) = \frac{\partial}{\partial \beta_j} \sum_{k,l} (\beta_k a_{kl} \beta_l)$$

$$A = (a_{kl})$$

$$\beta^T A \beta = (\beta_1 \ \beta_2 \ \cdots \ \beta_p) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & \cdots & & \\ \vdots & & \ddots & \\ & & & a_{pp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\Rightarrow \frac{\partial}{\partial \beta_j}(\beta^T A \beta) = \frac{\partial}{\partial \beta_j}\left(\sum_{k \neq j} a_{kj} \beta_k \beta_j + \sum_{k \neq j} a_{jk} \beta_k \beta_j + a_{jj} \beta_j^2 \right)$$

$$= \sum_{k \neq j} a_{kj} \beta_k + \sum_{k \neq j} a_{jk} \beta_k + 2 a_{jj} \beta_j$$

$$= \underbrace{\sum_{k \neq j} a_{kj} \beta_k + a_{jj} \beta_j}_{} + \underbrace{\sum_{k \neq j} a_{jk} \beta_k + a_{jj} \beta_j}_{}$$

$$(a_{1j} \ a_{2j} \ \cdots \ a_{pj}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \qquad (a_{j1} \ a_{j2} \cdots a_{jp}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$= A^T \beta \qquad\qquad = A\beta$$

Hence, when A is not symmetric,

$$\boxed{\nabla_\beta(\beta^T A \beta) = A\beta + A^T \beta}$$

∴ When A is symmetric $A^T = A$,

so $\nabla_\beta(\beta^T A \beta) = A\beta + A\beta = 2A\beta$

② $X = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 2 \\ 2 & 3 & 5 \end{bmatrix}$ $y = \begin{bmatrix} 0 \\ 4 \\ 10 \end{bmatrix}$

a) $X^T X = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 2 & 5 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 2 \\ 2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 12 \\ 7 & 10 & 17 \\ 12 & 17 & 29 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 12 \\ 7 & 10 & 17 \\ 12 & 17 & 29 \end{bmatrix}$

Here, Column 3 is a linear combination of column 1 and 2.

$C3 = C1 + C2$

Hence, rank $(X^T X) < 3$, $X^T X$ is not invertible

Also, to find $\min_\beta \|y - X\beta\|_2^2$, lets calculate

$X^T X \beta = X^T y$

$\Rightarrow \begin{bmatrix} 5 & 7 & 12 \\ 7 & 10 & 17 \\ 12 & 17 & 29 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 24 \\ 34 \\ 58 \end{bmatrix}$

Let's perform Guassian Elimination on the augmented matrix

$\left[\begin{array}{ccc|c} 5 & 7 & 12 & 24 \\ 7 & 10 & 17 & 34 \\ 12 & 17 & 29 & 58 \end{array}\right]$

After series of row-wise operations, we get

$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right]$

Only 2 pivotal points which also implies rank $(X^T X) = 2$

$\Rightarrow \beta_1 + \beta_3 = 2$

$\beta_2 + \beta_3 = 2$

Hence $\beta_{OLS} = \begin{bmatrix} 2-\beta_3 \\ 2-\beta_3 \\ \beta_3 \end{bmatrix}$ β_3 is a free variable.

for any value of β_3, it will result in the same error.

Hence, multiple minimisers exist

Problem 3 $\rightarrow$

(3) $y_i = x_i^T \beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$

Given $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ \|y - X\beta\|_2^2 + \lambda \beta_2^2 + \frac{\lambda}{2} \beta_3^2 \}$

Maximum Likelihood Estimate of β is given by
(a posteriori)

$$L(X,Y) = P(\beta)\, P(X,y)$$

$$= P(\beta) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)$$

$\sigma = 1$

$$\Rightarrow P(\beta) \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2\right)$$

$$\Rightarrow \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 + \log P(\beta)\right)$$

maximising above equation, is similar to

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T\beta) - \log P(\beta) \qquad -(1)$$

$$= \min_{\beta} \|y - X\beta\|_2^2 - 2\log P(\beta)$$

We need to find the distribution of β, such that

$$-2\log P(\beta) \approx \lambda \beta_2^2 + \frac{\lambda}{2} \beta_3^2 \quad \text{(minimising is same)}$$

if we consider $\beta \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} \infty & 0 & 0 \\ 0 & 1/\lambda & 0 \\ 0 & 0 & 2/\lambda \end{bmatrix}}_{=\Sigma} \right)$

$$P(\beta) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(\frac{-1}{2} \beta^T \Sigma^{-1} \beta\right)$$

$$= C \cdot \exp\left(\frac{-1}{2} [\beta_1, \beta_2\ \beta_3] \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda/2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}\right)$$

$$= C \cdot \exp\left(\frac{-1}{2} [0 + \lambda \beta_2^2 + \frac{\lambda}{2} \beta_3^2]\right)$$

Substituting $P(\beta)$ in ①

$$\Rightarrow \min_{\beta} \|y - X\beta\|_2^2 - 2\left[\log c - \frac{1}{2}\left[\lambda\beta_2^2 + \frac{\lambda}{2}\beta_3^2\right]\right]$$

$$= \min_{\beta} \|y - X\beta\|_2^2 + \lambda\beta_2^2 + \frac{\lambda}{2}\beta_3^2$$

Hence, $\beta \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \infty & 0 & 0 \\ 0 & 1/\lambda & 0 \\ 0 & 0 & 2/\lambda \end{bmatrix}\right)$

Problem 4 $\rightarrow$

(4) $x_i \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right)$ $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ $y_i = Ax_i$

We know that, $E[A\varepsilon] = AE[\varepsilon]$

and $Cov(A\varepsilon) = A\,Cov(\varepsilon)A^T$

a)

$$E(y_i) = E[Ax_i] = AE[x_i] = A\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$Cov(y_i) = Cov(Ax_i) = A\,Cov(x_i)A^T$$

$$= \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

$$Cov(y_i) = \begin{bmatrix} 9 & 19 \\ 19 & 41 \end{bmatrix}$$

b) Now, $x_i \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right)$

$$E(y_i) = E[Ax_i] = AE[x_i] = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$$

$$Cov(y_i) = A\,Cov(x_i)\,A^T$$

Since the covariance matrix of x_i is same as (a), $cov(y_i) = \begin{bmatrix} 9 & 19 \\ 19 & 41 \end{bmatrix}$

```
% Solution to HW1 - Problem 5
X = mvnrnd([0;0;0], [2,0,0;0,4,0;0,0,10],100); % Generating X for 100 samples
E = normrnd(0, sqrt(12), [100,1]); % Generating error for 100 samples
B = [0;1;-1]; % Regression coefficients
Y = X*B+E;

% (5a)
B_ols = inv(X'*X)*X'*Y; % Calculating regression parameters using ordinary least squares
disp("5a)")
disp("B_ols : ")
disp(B_ols)
z = 1.645; % For constructing 90% confidence interval, z = 1.645
sd = sqrt(12);
%sqrt(sum((Y - X*B_ols).^2)/100); % Calcuating standard deviation from the sample
v = diag(inv(X'*X)) ;
disp("90% Confidence Interval for B (B_1, B_2 and B_3): ")
CI = [B_ols - (z * (sd * sqrt(v))),  B_ols + (z * (sd * sqrt(v)))];
disp(CI)
disp("Yes, the interval convers B_1");
disp(" ")
% (5b)
disp("5b)")
count = 0;
for i=1:500
    E = normrnd(0, sqrt(12), [100,1]);
    Y = X*B+E;
    B_ols = inv(X'*X)*X'*Y;
    sd = sqrt(12); % sqrt(sum((Y - X*B_ols).^2)/100); % Calcuating standard deviation from
the sample
    v = diag(inv(X'*X));
    CI = [B_ols - (z * (sd * sqrt(v))),  B_ols + (z * (sd * sqrt(v)))];
    if B(1) > CI(1,1) && B(1) < CI(1,2)
        count = count + 1;
    end
end
disp("Percentage of time confidence interval covers B_1 is")
disp((count*100)/500)

% (5c)
disp("5c)")
E = -6 + (6+6)*rand(100,1);
B_ols = inv(X'*X)*X'*Y; % Calculating regression parameters using ordinary least squares
z = 1.645; % For constructing 90% confidence interval, z = 1.645
sd = sqrt(12); % sqrt(sum((Y - X*B_ols).^2)/100); % Calcuating standard deviation from the
sample
v = diag(inv(X'*X));
CI = [B_ols - (z * (sd * sqrt(v))),  B_ols + (z * (sd * sqrt(v)))];
count = 0;
for i=1:500
    E = -6 + (6+6)*rand(100,1);
    Y = X*B+E;
    B_ols = inv(X'*X)*X'*Y;
    sd = sqrt(12); % sqrt(sum((Y - X*B_ols).^2)/100); % Calcuating standard deviation from
the sample
    v = diag(inv(X'*X)) ;
    CI = [B_ols - (z * (sd * sqrt(v))),  B_ols + (z * (sd * sqrt(v)))];
    if B(1) > CI(1,1) && B(1) < CI(1,2)
        count = count + 1;
    end
```

```
end
disp("Percentage of time confidence interval covers B_1 is")
disp((count*100)/500)
```

```
5a)
B_ols :
   -0.1862
    0.8574
   -1.2067

90% Confidence Interval for B (B_1, B_2 and B_3):
   -0.5682    0.1957
    0.5353    1.1796
   -1.3993   -1.0141

Yes, the interval convers B_1

5b)
Percentage of time confidence interval covers B_1 is
   89.4000

5c)
Percentage of time confidence interval covers B_1 is
   88.4000
```

```
% Solution to Problem 6
data = readtable('bikeshare.csv');
Y = data.cnt;
X = [ones(size(Y)), data.holiday, data.temp, data.hum, data.windspeed];

%Ordinary Least Squares
disp('6a)') %Ordinary Least Sqaures
B_ols = inv(X'*X)*X'*Y;
disp('B_olsf for [intercept, holiday, temperature, humidity, windspeed]')
disp(B_ols)
disp("Coefficients B_2, B_4 and B_5 for attributes holiday, humidity and windspeed are nega
tive, which indicate that")
disp("usage of bike is low when there is a holiday, when humidity and windspeed is high. We
 also see that maginitude ")
disp("of B_3 is high, indicating that temperature has a high positive correlation with bike
 usage - probably people ")
disp("like to bike during summers. Hence coefficients seem convincing.")
disp(" ")

disp("6b)")
%Ridge Regression
s = size(X);
I = eye(s(2));

disp('Ridge Regression with lambda = 0.1')
lambda = 0.1;
B_ridge = inv((X'*X) + (lambda * I))*X'*Y

disp('Ridge Regression with lambda = 1')
lambda = 1;
B_ridge = inv((X'*X) + (lambda * I))*X'*Y

disp('Ridge Regression with lambda = 10')
lambda = 10;
B_ridge = inv((X'*X) + (lambda * I))*X'*Y

disp('Ridge Regression with lambda = 100')
lambda = 100;
B_ridge = inv((X'*X) + (lambda * I))*X'*Y

disp('Ridge Regression with lambda = 1000')
lambda = 1000;
B_ridge = inv((X'*X) + (lambda * I))*X'*Y
disp(" ")
disp(" - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -  ")
```

```
6a)
B_olsf for [intercept, holiday, temperature, humidity, windspeed]
   1.0e+03 *

    4.1156
   -0.6136
    6.6103
   -3.1093
   -4.8085

Coefficients B_2, B_4 and B_5 for attributes holiday, humidity and windspeed are negative,
```

```
which indicate that
usage of bike is low when there is a holiday, when humidity and windspeed is high. We also
see that maginitude
of B_3 is high, indicating that temperature has a high positive correlation with bike usage
 - probably people
like to bike during summers. Hence coefficients seem convincing.

6b)
Ridge Regression with lambda = 0.1

B_ridge =

   1.0e+03 *

    4.0568
   -0.6101
    6.5955
   -3.0493
   -4.6627

Ridge Regression with lambda = 1

B_ridge =

   1.0e+03 *

    3.6454
   -0.5814
    6.4484
   -2.5952
   -3.6448

Ridge Regression with lambda = 10

B_ridge =

   1.0e+03 *

    2.6060
   -0.4050
    5.0866
   -0.7355
   -0.9642

Ridge Regression with lambda = 100

B_ridge =

   1.0e+03 *

    2.3612
   -0.0637
    2.2269
    1.0719
    0.2401

Ridge Regression with lambda = 1000

B_ridge =

   1.0e+03 *
```

```
    1.4655
    0.0262
    0.8656
    0.8852
    0.2555


 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```