CSC412: Assignment 1

Due on Friday, Feb 8, 2018

Zhongtian Ouyang
1002341012

(collaborate with Yihao Ni)

Problem 1

The proves are for discrete variables. For continuous variables, just change sum to integration and everything else should be almost the same. (a)
For two independent variables X, Y:
$P(X \cap Y) = P(X)P(Y)$
$E[XY] = \sum_x \sum_y xyP(x,y) = \sum_x \sum_y xyP(x)P(y) = (\sum_x xP(x))(\sum_y yP(y)) = E[X]E[Y]$
$Cov(X,Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$

(b)

$$\begin{aligned} E[X+\alpha Y] &= \sum_x \sum_y (x+\alpha y)P(x,y) \\ &= \sum_x \sum_y xP(x,y) + \sum_x \sum_y \alpha yP(x,y) \\ &= \sum_x xP(x) + \alpha \sum_y yP(y) \\ &= E[x] + \alpha E[y] \end{aligned} \tag{1}$$

$$\begin{aligned} V[X+\alpha Y] &= E[((X+\alpha Y) - E[X+\alpha Y])^2] \\ &= E[((X-\mu_x) + (\alpha Y - \alpha\mu_y))^2] \\ &= E[((X-\mu_x) + \alpha(Y-\mu_y))^2] \\ &= E[(X-\mu_x)^2 + 2(X-\mu_x)\alpha(Y-\mu_y) + \alpha^2(Y-\mu_y)^2] \\ &= V[X] + 2\alpha Cov(X,Y) + \alpha^2 V[Y] \\ &= V[X] + \alpha^2 V[Y] \end{aligned} \tag{2}$$

Problem 2

(a)
For continuous random variables, its pdf at some value can be greater than 1 as long as the area under the curve sums to 1.

(b)

$$f(x|\mu = 0, \sigma^2 = \frac{1}{100}) = \frac{1}{\sqrt{2\pi\frac{1}{100}}}e^{-\frac{(x-0)^2}{2\frac{1}{100}}} = \frac{1}{\sqrt{\frac{\pi}{50}}}e^{-50x^2}$$

(c)

$$f(0|\mu = 0, \sigma^2 = \frac{1}{100}) = \frac{1}{\sqrt{\frac{\pi}{50}}}e^{-50*0^2} = 3.9894$$

(d)
The probability that X=0 is 0.

Problem 3

(a)

$$r = x^T y = \sum_{i=1}^{m} x_i * y_i$$
$$\frac{\partial r}{\partial x_i} = y_i$$
$$\frac{\partial x^T y}{\partial x} = y$$

(b)

$$r = x^T x = \sum_{i=1}^{m} x_i * x_i = \sum_{i=1}^{m} x_i^2$$
$$\frac{\partial r}{\partial x_i} = 2 * x_i$$
$$\frac{\partial x^T x}{\partial x} = 2x$$

(c)

$$r = x^T A,\ r_i = \sum_{j=1}^{m} x_j a_{ji}$$
$$\frac{\partial r}{\partial x} = J = \begin{bmatrix} \frac{\partial r_1}{\partial x_1} & ... & \frac{\partial r_1}{\partial x_m} \\ & \vdots & \\ \frac{\partial r_m}{\partial x_1} & ... & \frac{\partial r_m}{\partial x_m} \end{bmatrix} = \begin{bmatrix} a_{11} & ... & a_{m1} \\ & \vdots & \\ a_{1m} & ... & a_{mm} \end{bmatrix}$$
$$\frac{\partial x^T A}{\partial x} = J = A^T$$

(d)

let $r = x^T A$, $y = x$

$z = x^T A x = ry = (r_1 x_1 + ... + r_m x_m) = (a_{11} x_1 + ... + a_{m1} x_m) x_1 + ... + (a_{1m} x_1 + ... + a_{mm} x_m) x_m$:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial x} = x^T A^T + r = x^T A^T + x^T A = x^T (A + A^T)$$

Problem 4

(a)
$Y = X\beta + \epsilon$ where ϵ is the difference between Y and $X\beta$, the noise from variance. $E[\epsilon|X] = 0$

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^TY \\ &= (X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= \beta + (X^TX)^{-1}X^T\epsilon \end{aligned} \tag{3}$$

$$E[\hat{\beta}] = E[\beta + (X^TX)^{-1}X^T\epsilon] = \beta + (X^TX)^{-1}E[X^T\epsilon] = \beta$$

$$\begin{aligned} V[\hat{\beta}] &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\ &= E[((X^TX)^{-1}X^T\epsilon)((X^TX)^{-1}X^T\epsilon)^T] \\ &= E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^TE[\epsilon\epsilon^T]X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned} \tag{4}$$

(b)
Likelihood function for β is:

$$\begin{aligned} L(Y|X, \beta, \sigma^2I) &= \prod_{i=1}^{n} \frac{1}{det(2\pi \sum)^{\frac{1}{2}}} e^{-\frac{1}{2}(y_i-\mu_i)^T \sum^{-1}(y_i-\mu_i)} \\ &= \prod_{i=1}^{n} \frac{1}{det(2\pi\sigma^2I)^{\frac{1}{2}}} e^{-\frac{1}{2}(y_i-x_i\beta)^T(\sigma^2I)^{-1}(y_i-x_i\beta)} \\ &= \prod_{i=1}^{n} \frac{1}{det(2\pi\sigma^2I)^{\frac{1}{2}}} e^{-\frac{1}{2}\sigma^{-2}I(y_i-x_i\beta)^T(y_i-x_i\beta)} \\ &= \frac{1}{det(2\pi\sigma^2I)^{\frac{n}{2}}} e^{\sum_{i=1}^{n} -\frac{1}{2}\sigma^{-2}I(y_i-x_i\beta)^T(y_i-x_i\beta)} \\ &= \frac{1}{det(2\pi\sigma^2I)^{\frac{n}{2}}} e^{-\frac{1}{2}\sigma^{-2}I\sum_{i=1}^{n}(y_i-x_i\beta)^T(y_i-x_i\beta)} \\ &= \frac{1}{det(2\pi\sigma^2I)^{\frac{n}{2}}} e^{-\frac{1}{2}\sigma^{-2}I(Y-X\beta)^T(Y-X\beta)} \end{aligned} \tag{5}$$

When we find an β that minimize $\sum_{i=1}^{n}(y_i - x_i\beta)^2$, since $\sum_{i=1}^{n}(y_i - x_i\beta)^2 = (Y - X\beta)^T(Y - X\beta)$, such β also minimize $(Y - X\beta)^T(Y - X\beta)$. Therefore, such β would maximize the term $e^{-\frac{1}{2}\sigma^{-2}I(Y-X\beta)^T(Y-X\beta)}$ and maxmize the likelihood function.
An β that maxmize the likelihood function would also be minimizing the square error
Without even using a log likelihood trick, we can conclude that minimizing square error is equivalent to maximizin the likelihood.

(c)

$$\begin{aligned}\sum_{i=1}^{n}(y_i - x_i\beta)^2 &= (Y - X\beta)^T(Y - X\beta) \\ &= Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta\end{aligned} \tag{6}$$

(d)

$$\begin{aligned}\frac{\partial}{\partial\beta}(Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta) &= \frac{\partial Y^TY}{\partial\beta} - \frac{\partial 2\beta^TX^TY}{\partial\beta} + \frac{\partial\beta^TX^TX\beta}{\partial\beta} \\ &= 0 - 2Y^TX + \beta^T(X^TX + (X^TX)^T) \\ &= 0 - 2Y^TX + 2\beta^T(X^TX)\end{aligned} \tag{7}$$

Since both $2Y^TX$ and $2\beta^T(X^TX)$ vectors, $0-2Y^TX+2\beta^T(X^TX)$ and $0-2X^TY+2(X^TX)\beta$ are the same except one is row vector, the other is column vector. To find $\beta = \hat{\beta}$ minimize the error, we set derivative equals to zero

$$\begin{aligned}-2X^TY + 2(X^TX)\hat{\beta} = 0 \\ X^TY = (X^TX)\hat{\beta} \\ \hat{\beta} = (X^TX)^{-1}X^TY\end{aligned}$$

Problem 5

(a)

$$argmax_{\hat{\beta}}\frac{P(y|\beta=\hat{\beta})P(\beta=\hat{\beta})}{P(y)} = argmax_{\hat{\beta}}P(y|\beta=\hat{\beta})P(\beta=\hat{\beta}) = argmax_{\hat{\beta}}lnP(y|\beta=\hat{\beta})+lnP(\beta=\hat{\beta})$$

$$P(y|\beta=\hat{\beta}) = \prod_{i=1}^{n}\frac{1}{det(2\pi\sum)^{\frac{1}{2}}}e^{-\frac{1}{2}(y_i-\mu_i)^T\sum^{-1}(y_i-\mu_i)}$$

$$\begin{aligned} lnP(y|\beta=\hat{\beta}) &= -\frac{n^2}{2}ln(2\pi)-\frac{n}{2}ln(det(\sum))-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu_i)^T(\sum)^{-1}(y_i-\mu_i) \\ &= -\frac{n^2}{2}ln(2\pi)-\frac{n}{2}ln(det(\sigma^2 I))-\frac{1}{2}\sum_{i=1}^{n}(y_i-x_i\beta)^T(\sigma^2 I))^{-1}(y_i-x_i\beta) \\ &= -\frac{n^2}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^{2n}det(I))-\frac{1}{2}\sum_{i=1}^{n}(y_i-x_i\beta)^T\sigma^{-2}I(y_i-x_i\beta) \\ &= -\frac{n^2}{2}ln(2\pi)-\frac{n^2}{2}ln(\sigma^2)-\frac{1}{2}\sigma^{-2}I(Y-X\beta)^T(Y-X\beta) \end{aligned} \tag{8}$$

$$P(\beta=\hat{\beta}) = \frac{1}{det(2\pi\sum)^{\frac{1}{2}}}e^{-\frac{1}{2}(\beta-0)^T(\sum)^{-1}(\beta-0)} = \frac{1}{2\pi^{\frac{m}{2}}det(\sum)^{\frac{1}{2}}}e^{-\frac{1}{2}\beta^T(\sum)^{-1}\beta}$$

$$\begin{aligned} lnP(\beta=\hat{\beta}) &= -\frac{m}{2}ln(det(2\pi))-\frac{1}{2}ln(det(\sum))-\frac{1}{2}\beta^T(\sum)^{-1}\beta \\ &= -\frac{m}{2}ln(det(2\pi))-\frac{1}{2}ln(det(\tau^2 I))-\frac{1}{2}\beta^T(\tau^2 I)^{-1}\beta \\ &= -\frac{m}{2}ln(det(2\pi))-\frac{m}{2}ln(\tau^2)-\frac{1}{2}\tau^{-2}I\beta^T\beta \end{aligned} \tag{9}$$

$$F = lnP(y|\beta=\hat{\beta})+lnP(\beta=\hat{\beta})$$

$$\begin{aligned} \frac{\partial}{\partial\beta}F &= 0+0+\frac{\partial}{\partial\beta}(-\frac{1}{2}\sigma^{-2}I(Y-X\beta)^T(Y-X\beta))+0+0+\frac{\partial}{\partial\beta}(-\frac{1}{2}\tau^{-2}I\beta^T\beta) \\ &= -\frac{1}{2}\sigma^{-2}I(-2X^TY+2(X^TX)\beta)-\frac{1}{2}\tau^{-2}I(2\beta) \\ &= -\frac{1}{2}\sigma^{-2}I(-2X^TY+2(X^TX)\beta)-\tau^{-2}I\beta \end{aligned} \tag{10}$$

let $\beta = \hat{\beta}$ such that $\frac{\partial}{\partial \beta} F = 0$

$$\begin{aligned}
0 &= -\frac{1}{2}\sigma^{-2} I(-2X^T Y + 2(X^T X)\hat{\beta}) - \tau^{-2} I \hat{\beta} \\
0 &= \sigma^{-2} I X^T Y - \sigma^{-2} I (X^T X)\hat{\beta} - \tau^{-2} I \hat{\beta} \\
(\sigma^{-2} I (X^T X) + \tau^{-2} I)\hat{\beta} &= \sigma^{-2} I X^T Y \\
\hat{\beta} &= (\sigma^{-2} I (X^T X) + \tau^{-2} I)^{-1} \sigma^{-2} I X^T Y \\
\hat{\beta} &= (\sigma^{2} I^{-1} \sigma^{-2} I (X^T X) + \sigma^{2} I^{-1} \tau^{-2} I)^{-1} X^T Y \\
\hat{\beta} &= (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} X^T Y \\
\hat{\beta}_{MAP} &= (X^T X + \lambda I)^{-1} X^T Y
\end{aligned} \tag{11}$$

(b)
Modified X and Y:

$$\bar{X} = \begin{bmatrix} x_{11} & ... & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{nm} \\ \sqrt{\lambda} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sqrt{\lambda} \end{bmatrix}, \quad \bar{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{aligned} \bar{X}^T\bar{X} &= \begin{bmatrix} x_{11} & ... & x_{n1} & \sqrt{\lambda} & ... & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ x_{1m} & ... & x_{nm} & 0 & ... & \sqrt{\lambda} \end{bmatrix} \begin{bmatrix} x_{11} & ... & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{nm} \\ \sqrt{\lambda} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sqrt{\lambda} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 + \lambda & \sum_{i=1}^n x_{i2}x_{i1} & ... & \sum_{i=1}^n x_{im}x_{i1} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 + \lambda & ... & \sum_{i=1}^n x_{im}x_{i2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i1}x_{im} & \sum_{i=1}^n x_{i2}x_{im} & ... & \sum_{i=1}^n x_{im}^2 + \lambda \end{bmatrix} \end{aligned} \tag{12}$$

$$\begin{aligned} X^TX + \lambda I &= \begin{bmatrix} x_{11} & ... & x_{n1} \\ \vdots & & \vdots \\ x_{1m} & ... & x_{nm} \end{bmatrix} \begin{bmatrix} x_{11} & ... & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{nm} \end{bmatrix} + \begin{bmatrix} \lambda & & 0 \\ & \ddots & \\ 0 & ... & \lambda \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i2}x_{i1} & ... & \sum_{i=1}^n x_{im}x_{i1} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & ... & \sum_{i=1}^n x_{im}x_{i2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i1}x_{im} & \sum_{i=1}^n x_{i2}x_{im} & ... & \sum_{i=1}^n x_{im}^2 \end{bmatrix} + \begin{bmatrix} \lambda & & 0 \\ & \ddots & \\ 0 & ... & \lambda \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 + \lambda & \sum_{i=1}^n x_{i2}x_{i1} & ... & \sum_{i=1}^n x_{im}x_{i1} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 + \lambda & ... & \sum_{i=1}^n x_{im}x_{i2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i1}x_{im} & \sum_{i=1}^n x_{i2}x_{im} & ... & \sum_{i=1}^n x_{im}^2 + \lambda \end{bmatrix} \end{aligned} \tag{13}$$

For $\bar{X}^T\bar{Y}$, since the last m columns of $\bar{X}^T$ are the added rows of $\bar{X}$ and last m rows of $\bar{Y}$ are the added value 0, $\bar{X}^T\bar{Y} = X^TY$
Therefore, the following is true:

$$(\bar{X}^T\bar{X})^{-1}\bar{X}^T\bar{Y} = (X^TX + \lambda I)^{-1}X^TY$$

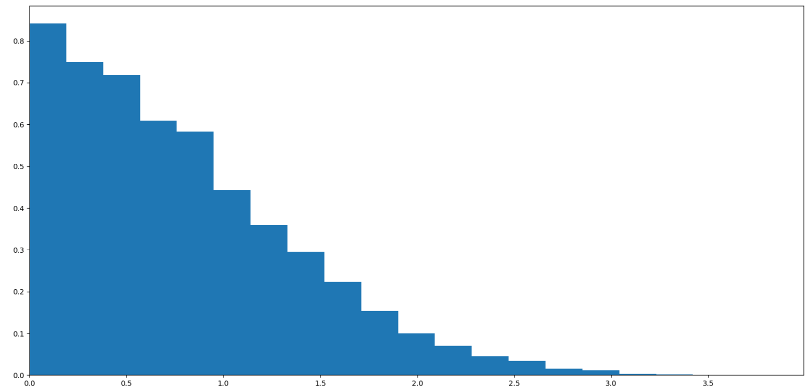
This shows that ridge regression with X and Y is equivalent to computing maximum likelihood estimate of β using the modified $\bar{X}$ and $\bar{Y}$

Problem 6

1.

$$\begin{aligned} distance\ of\ x\ from\ origin &= \sqrt{(x_1-0)^2+(x_2-0)^2+...+(x_D-0)^2} \\ &= \sqrt{x_1^2+x_2^2+...+x_D^2} \\ &= \sqrt{x^Tx} \end{aligned} \tag{14}$$

2.
From the histogram below, we can see that most samples will be near the origin.



3.
From the histograms in Figure 1 for different dimensions, we can observe that as the dimensionality of the Gaussian increases, the expected distance of the samples from the Gaussian's mean increase, shift away from 0.

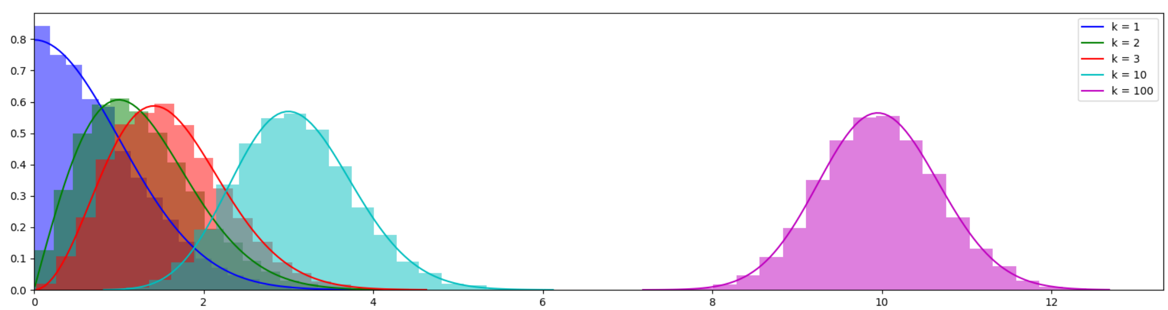


Figure 1: Question 3 & 4

4.
The lines in Figure 1 are the pdfs of the chi distribution with k = {1,2,3,10,100}

5.
$(x_a - x_b) \sim N(0_D, 2I_D)$
Because of euclidean distance and standard deviation is $\sqrt{2}$ of the previous. $Y = \sqrt{2}X$, $g^{-1}(Y) = (1/\sqrt{2})Y$, $f_y(Y) = (1/\sqrt{2})f_x((1/\sqrt{2})Y)$
Again, as the dimesionality increases, the distance between samples from a Gaussian increases.The mean shifts from 0 by about $\sqrt{2}$ times of the previous question.

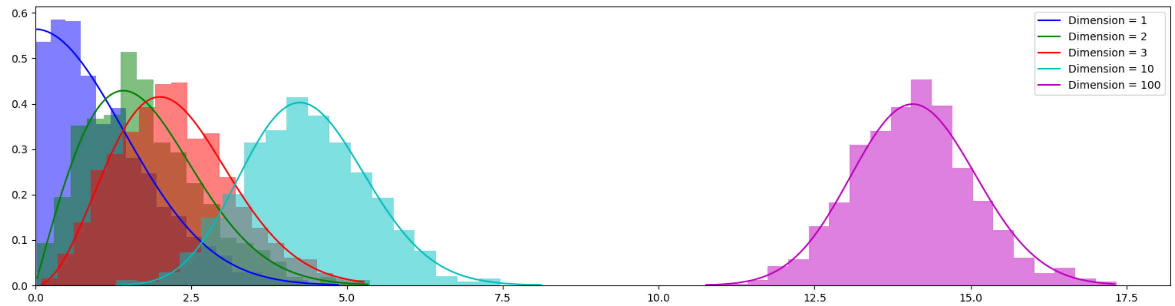


Figure 2: Question5

6.
The log-likelihood increases as α approaching 0.5, reach the maximum, and then decrease as α approaching 1. The shape is identical for all dimensions. However, as the dimension increases, the overall value of log-likelihood decreases by a large amount. A higher log-likelihood for the interplolated points is not necessrily better. It is not a good idea to linearly interpolate between samples from a high dimensional Gaussian
7.
The log-likelihood is mostly flat for $\alpha \in [0, 1]$. Polar interpolation is more suitable because the interpolates it provides are uniform and contain same level of information for high dimensional Gaussians.

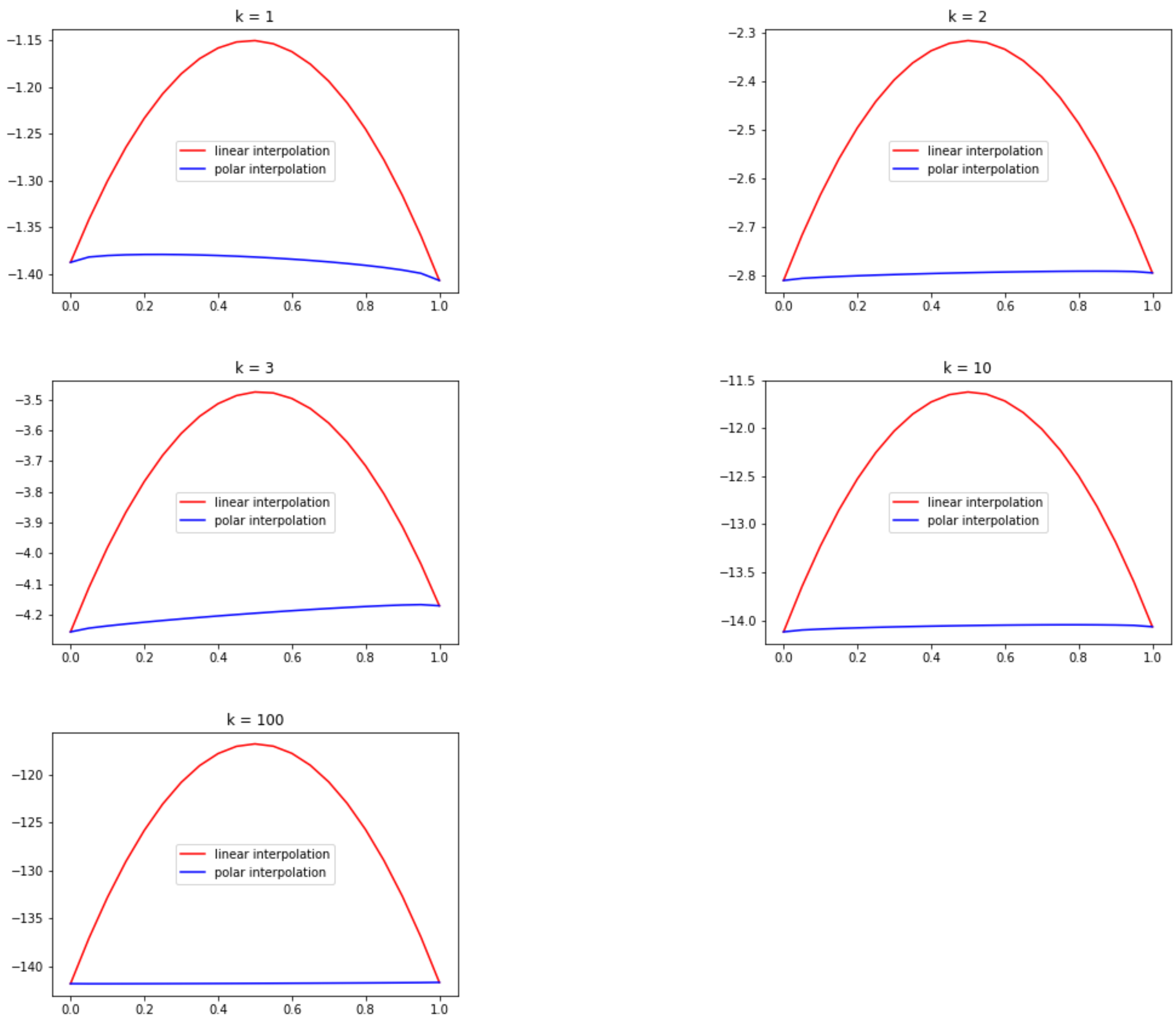


Figure 3: Q6 & 7