Department of Electrical Engineering
School of Science and Engineering

EE212 Mathematical Foundations for Machine Learning and Data Science

ASSIGNMENT 3 – SOLUTIONS

Due Date: 23:55, Saturday. August 8, 2020 (Submit online on LMS)
Format: 12 problems, for a total of 100 marks
Instructions:

- You are not allowed to submit a group assignment. Each student must submit his/her own hand-written assignment, scanned in a single PDF document.
- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. Anybody found guilty would be subjected to disciplinary action in accordance with the university rules and regulations.

Problem 1 (7 marks)

Suppose the $m \times n$ matrix A has linearly independent columns, and b is an m-vector. Let $\hat{x} = A^\dagger b$ denote the least squares approximate solution of $Ax = b$.

(a) [**2 marks**] Show that for any n-vector x, $(Ax)^T b = (Ax)^T(A\hat{x})$, i.e., the inner product of Ax and b is the same as the inner product of Ax and $A\hat{x}$. *Hint.* Use $(Ax)^T b = x^T(A^T b)$ and $(A^T A)\hat{x} = A^T b$.

(b) [**2 marks**] Show that when $A\hat{x}$ and b are both nonzero, we have

$$\frac{(A\hat{x})^T b}{\|A\hat{x}\|\|b\|} = \frac{\|A\hat{x}\|}{\|b\|}$$

The left-hand side is the cosine of the angle between $A\hat{x}$ and b. *Hint.* Apply part (a) with $x = \hat{x}$.

(c) [**3 marks**] The choice $x = \hat{x}$ minimizes the distance between Ax and b. Show that $x = \hat{x}$ also minimizes the angle between Ax and b. (You can assume that Ax and b are nonzero.) *Remark.* For any positive scalar α, $x = \alpha\hat{x}$ also minimizes the angle between Ax and b.

Solution:

(a) To see this, we note that

$$(Ax)^T b = x^T(A^T b) = x^T(A^T A\hat{x}) = (Ax)^T(A\hat{x})$$

using the normal equations $(A^T A)\hat{x} = A^T b$

(b) We apply part (a) with $x = \hat{x}$ to get $(A\hat{x})^T b = \|A\hat{x}\|^2$. Divide this by $\|A\hat{x}\|\|b\|$ to get the equality.

(c) We have to show that for any x, we have

$$\cos^{-1}\left(\frac{(A\hat{x})^T b}{\|A\hat{x}\|\|b\|}\right) \leq \cos^{-1}\left(\frac{(Ax)^T b}{\|Ax\|\|b\|}\right)$$

provided $Ax \neq 0$. Using part (a) this is the same as

$$\frac{(A\hat{x})^T(A\hat{x})}{\|A\hat{x}\|\|b\|} \geq \frac{(Ax)^T(A\hat{x})}{\|Ax\|\|b\|}$$

We can cancel the term $\|b\|$ from each side, and re-arrange this as

$$\|A\hat{x}\|\|Ax\| \geq (Ax)^T(A\hat{x})$$

which is true by the Cauchy–Schwarz inequality.

Problem 2 (8 marks)

Suppose that A has linearly independent columns, so $\hat{x} = A^\dagger b$ minimizes $\|Ax - b\|^2$. In this exercise, we explore an iterative method, due to the mathematician Lewis Richardson, that can be used to compute $\bar{x}$. We define $x^{(1)} = 0$ and for $k = 1, 2, \ldots,$

$$x^{(k+1)} = x^{(k)} - \mu A^T(Ax^{(k)} - b)$$

where μ is a positive parameter, and the superscripts denote the iteration number. This defines a sequence of vectors that converge to $\bar{x}$ provided μ is not too large; the choice $\mu = 1/\|A\|^2$, for example, always works. The iteration is terminated when $A^T(Ax^{(k)} - b)$ is small enough, which means the least squares optimality conditions are almost satisfied. To implement the method we only need to multiply vectors by A and by A^T. If we have efficient methods for carrying out these two matrix-vector multiplications, this iterative method can be faster than other methods (although it does not give the exact solution). Iterative methods are often used for very large scale least squares problems.

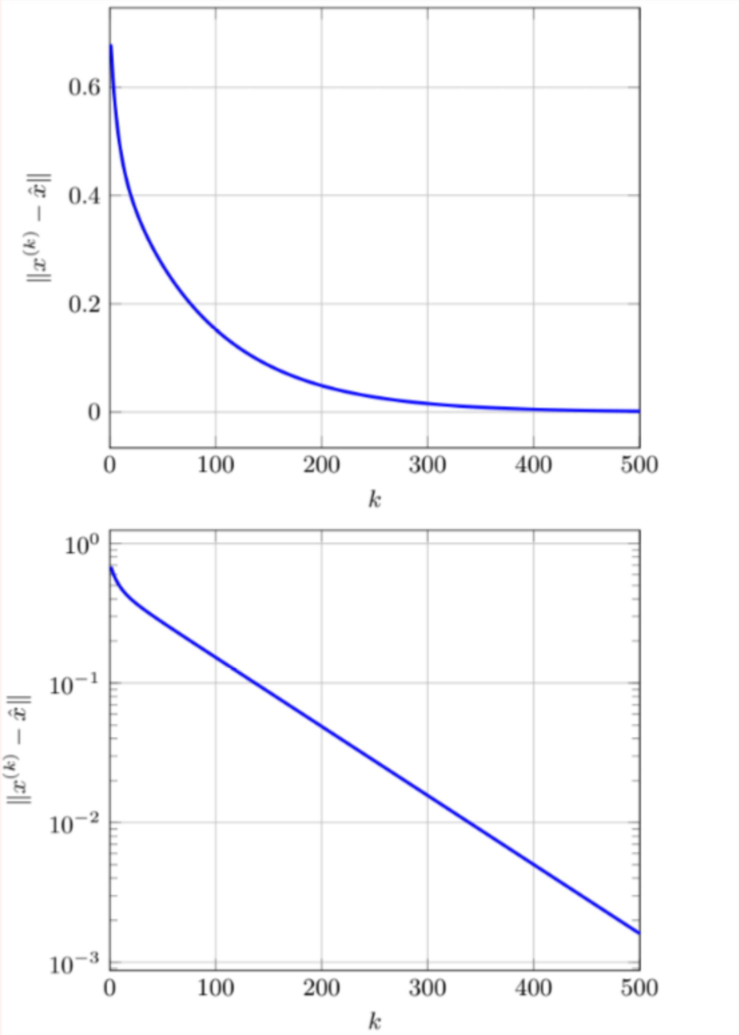
(a) **[3 marks]** Show that if $x^{(k+1)} = x^{(k)}$, we have $x^{(k)} = \hat{x}$.

(b) **[5 marks]** Generate a random 20×10 matrix A and 20-vector b, and compute $\hat{x} = A^\dagger b$. Run the Richardson algorithm with $\mu = 1/\|A\|^2$ for 500 iterations, and plot $\|x^{(k)} - \hat{x}\|$ to verify that $x^{(k)}$ appears to be converging to $\hat{x}$.

Solution:

(a) If $x^{(k+1)} = x^{(k)}$, then $A^T(Ax^{(k)} - b) = 0$, so $x^{(k)} = \hat{x}$. (This is because $\hat{x}$ is the unique solution of the normal equations $A^T(Ax - b) = 0$.)

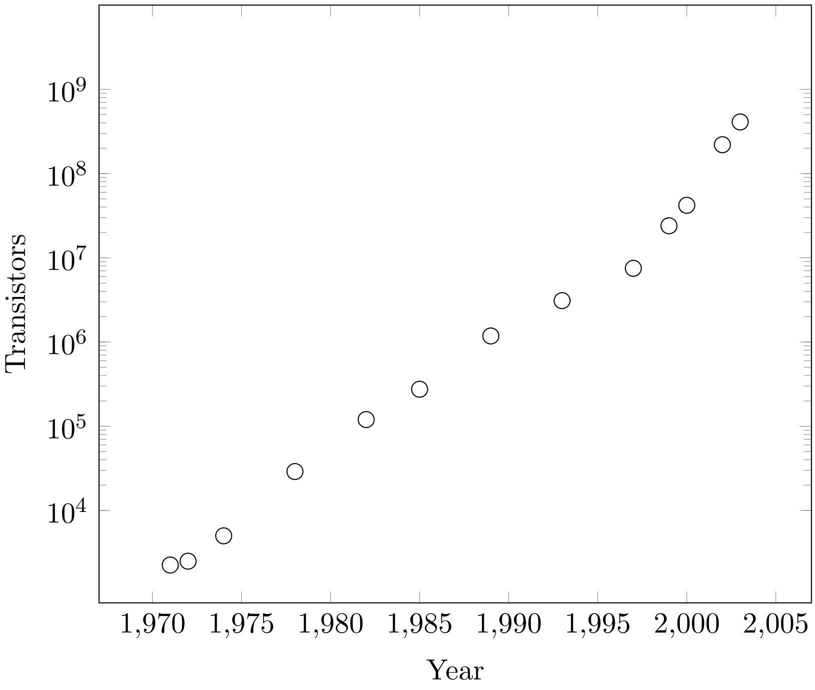
(b) The plots show the error $\|x^{(k)} - \hat{x}\|$ on a linear and logarithmic scale:



As evident from the plots, $\|x^{(k)} - \hat{x}\|$ converges to 0 as k increases, i.e., $x^{(k)}$ converges to $\hat{x}$.

Problem 3 (10 marks)

The figure and table below show the number of transistors N in 13 microprocessors, and the year of their introduction.



Year	Transistors
1971	2,250
1972	2,500
1974	5,000
1978	29,000
1982	120,000
1985	275,000
1989	1,180,000
1993	3,100,000
1997	7,500,000
1999	24,000,000
2000	42,000,000
2002	220,000,000
2003	410,000,000

The plot gives the number of transistors on a logarithmic scale. Find the least squares straight-line fit of the data using the model

$$\log_{10} N \approx \theta_1 + \theta_2(t - 1970)$$

where t is the year and N is the number of transistors. Note that $_1$ is the model'prediction of the log of the number of transistors in 1970, and 10^{θ_2} gives the model's prediction of the fractional increase in number of transistors per year.

(a) **[5 marks]** Find the coefficients θ_1 and θ_2 that minimize the RMS error on the data, and give the RMS error on the data. Plot the model you find along with the data points.

(b) **[3 marks]** Use your model to predict the number of transistors in a microprocessor introduced in 2015. Compare the prediction to the IBM Z13 microprocessor, released in 2015, which has around 4×10^9 transistors.

(c) **[2 marks]** Compare your result with Moore's law, which states that the number of transistors per integrated circuit roughly doubles every one and a half to two years.

The computer scientist and Intel corporation co-founder Gordon Moore formulated the law that bears his name in a magazine article published in 1965.

Solution:

(a) We minimize the RMS error by minimizing the sum of the squares of the prediction errors,

$$\sum_{k=1}^{13}(\log_{10} n_k = \theta_1 - (t_k - 1970)\theta_2)^2 = \|A\theta - b\|^2,$$

where

$$A = \begin{bmatrix} 1 & t_1 - 1970 \\ 1 & t_2 - 1970 \\ \vdots & \vdots \\ 1 & t_{13} - 1970 \end{bmatrix}, \qquad b = \begin{bmatrix} \log_{10} N_1 \\ \log_{10} N_2 \\ \vdots \\ \log_{10} N_{13} \end{bmatrix}$$

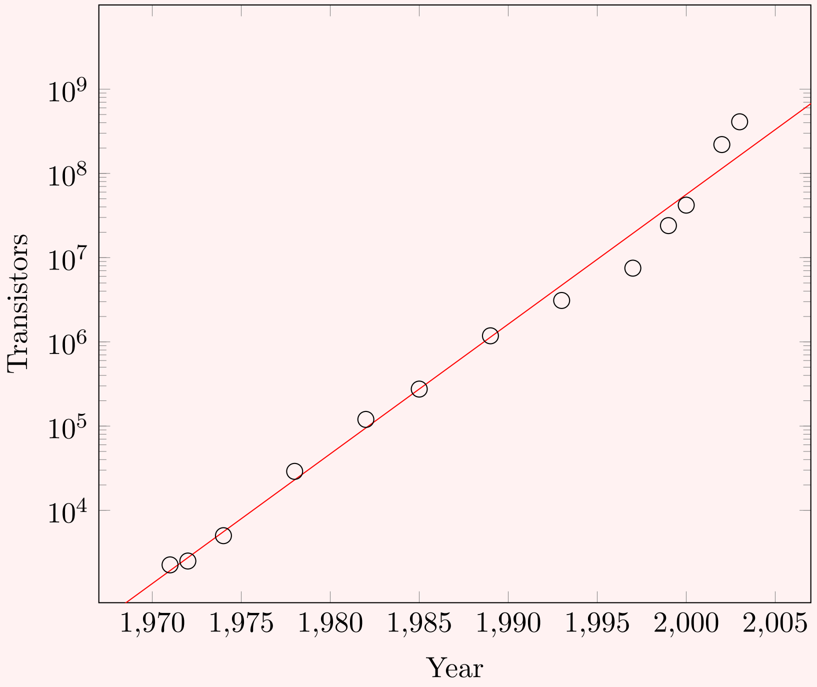
The solution is

$$\hat{\theta}_1 = 3.13, \qquad \hat{\theta}_2 = 0.154$$

The RMS error of this model is

$$\frac{1}{\sqrt{13}}\|A\hat{\theta} - b\| = 0.20$$

This means that we can expect our prediction of $\log_{10} N$ to typically be off by around 0.20. This corresponds to a prediction typically off by around a factor of $10^{0.2} = 1.6$. The straight-line fit is shown in the following figure.



(b) The predicted number of transistors in 2015 is

$$10^{\theta_1+\theta_2(2015-1970)} \approx 1.14 \times 10^{10}$$

This prediction is about a factor of 3 off from the IBM Z13 processor. That is around two standard deviations off from the prediction, which is reasonable. (Although in general we would not expect extrapolations to have the same error as observed on the training data set.)

(c) In our model the number of transistors doubles approximately every $\log_{10} 2/\theta_2 = 1.95$ years, which is consistent with Moore's law.

Problem 4 (5 marks)

Five different models are fit using the same training data set, and tested on the same (separate) test set (which has the same size as the training set). The RMS prediction errors for each model, on the training and test sets, are reported below. Comment briefly on the results for each model. You might mention whether the model's predictions are good or bad, whether it is likely to generalize to unseen data, or whether it is over-fit. You are also welcome to say that you don't believe the results, or think the reported numbers are fishy.

Solution:

(a) This is a good model, and likely will generalize.

(b) This is a bad model, but will likely generalize.

Model	Train RMS	Test RMS
A	1.355	1.423
B	9.760	9.165
C	5.033	0.889
D	0.211	5.072
E	0.633	0.633

(c) Something is wrong, or you are outrageously lucky, Probably the former.

(d) The model is over-fit

(e) These results are suspicious, since it's unlikely that the train and test RMS errors would be so close. For example, maybe the model was accidentally tested on the training set. If the numbers are correct, then this is a very good model, and would likely generalize.

Problem 5 (5 marks)

For the function $f(x_1, x_2, x_3) = x_1^2 + 2x_2^3 + 3x_3^4$, find ∇f.

Solution: $\nabla f = 2x_1 + 6x_2^2 = 12x_3^3$

Problem 6 (5 marks)

Consider the following functions:

$$\begin{aligned} f_1(x) &= \sin(x_1)\cos(x_2), \qquad && x \in \mathbb{R}^2 \\ f_2(x, y) &= x^T y, && x, y \in \mathbb{R}^n \end{aligned}$$

(a) What are the dimensions of $\frac{\partial f_i}{\partial x}$?

(b) Compute the Jacobians.

Solution:

(a) $\frac{\partial f_1}{\partial x}$ is 2 dimensional
$\frac{\partial f_2}{\partial x}$ is n dimensional

(b) $J(f_1) = \cos(x_1)\cos(x_2) - \sin(x_1)\sin(x_2)$
$J(f_2) = y^T$

Problem 7 (10 marks)

An experiment consists of tossing two dice.

(a) [**2 marks**] Find the sample space S.

(b) [**3 marks**] Find the event A that the sum of the dots on the dice is greater than 7.

(c) [**3 marks**] Find the event B that the sum of the dots on the dice is greater than 10.

(d) [**2 marks**] Find the event C that the sum of the dots on the dice is greater than 12.

Solution:

(a) For this experiment, the sample space S consists to 36 points:

$$S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$$

where i represents the number of dots appearing on one die and j represents the number of dots appearing on the other die.

(b) The event A consists of 15 points:

$$A = \{(2, 6), (3, 5), (3, 6), (4, 4), (4, 5), (4, 6), (5, 3), \\ (5, 4), (5, 5), (5, 6), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

(c) The event B consists of 3 points:

$$B = \{(5, 6), (6, 5), (6, 6)\}$$

(d) The event C is an impossible event, i.e. $C = \emptyset$.

Problem 8 (10 marks)

A company producing electric relays has three manufacturing plants producing 50, 30, and 20 percent, respectively, of its product. Suppose that the probabilities that a relay manufactured by these plants is defective are 0.02, 0.05, and 0.01, respectively.

(a) [**5 marks**] If a relay is selected at random from the output of the company, what is the probability that it is defective?

(b) [**5 marks**] If a relay selected at random is found to be defective, what is the probability that it was manufactured by plant 2?

Solution:

(a) Let B be the event that the relay is defective, and let A_i be the event that the relay is manufactured by plant $i (i = 1, 2, 3)$. The desired probability is $\mathbf{P}(B)$. We have

$$\begin{aligned}\mathbf{P}(B) &= \sum_{i=1}^{3} \mathbf{P}(B|A_i)\mathbf{P}(A_i) \\ &= (0.02)(0.5) + (0.05)(0.3) + (0.01)(0.2) = 0.027 \qquad (1)\end{aligned}$$

(b) The desired probability is $\mathbf{P}(A_2|B)$. Using the result from part (a), we obtain

$$\mathbf{P}(A_2|B) = \frac{\mathbf{P}(B|A_2)\mathbf{P}(A_2)}{\mathbf{P}(B)} = \frac{(0.05)(0.3)}{0.027} = 0.556$$

Problem 9 (10 marks)

A coin is tossed twice. Alice claims that the event of two heads is at least as likely if we know that the first toss is a head than if we know that at least one of the tosses is a head. Is she right? Does it make a difference if the coin is fair or unfair? How can we generalize Alice's reasoning?

Solution: Let A be the event that the first toss is a head and let B be the event that the second toss is a head. We must compare the conditional probabilities $\mathbf{P}(A \cap B|A)$ and $\mathbf{P}(A \cap B|A \cup B)$. We have

$$\mathbf{P}(A \cap B|A) = \frac{\mathbf{P}((A \cap B) \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)},$$

and

$$\mathbf{P}(A \cap B|A \cup B) = \frac{\mathbf{P}((A \cap B) \cap (A \cup B))}{\mathbf{P}(A \cup B)} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A \cup B)}$$

Since $\mathbf{P}(A \cup B) \geq \mathbf{P}(A)$, the first conditional probability above is at least as large, so Alice is right, regardless of whether the coin is fair or not. In the case where the coin is fair, that is, if all four outcomes HH, HT, TH, TT are equally likely, we have

$$\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{1/4}{1/2} = \frac{1}{2}, \quad \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A \cup B)} = \frac{1/4}{3/4} = \frac{1}{3},$$

A generalisation of Alice's reasoning is that if A, B and C are events such hat $B \subset C$ and $A \cap B = A \cap C$ (for example, if $A \subset B \subset C$), then the event A is more likely if we know that B has occurred than if we know that C has occurred.

Problem 10 (10 marks)

The MIT soccer team has 2 games scheduled for one weekend. MIT has a 0.4 probability of not losing the first game, and a 0.6 probability of not losing the second game. If it does not lose, the team has a 50% chance of a win, and a 50% chance of a tie, independently of all other weekend events. MIT will receive 2 points for a win, 1 for a tie, and a 0 for a loss. Let X be the number of points the MIT team earns over the weekend. Find the PMF of X.

Solution: This requires enumeration of the possibilities and straightforward computation:

$$\begin{aligned}
&\mathbf{P}(X = 0) = 0.6 \cdot 0.4 = 0.24,\\
&\mathbf{P}(X = 1) = 0.4 \cdot 0.5 \cdot 0.4 + 0.6 \cdot 0.5 \cdot 0.6 = 0.26\\
&\mathbf{P}(X = 2) = 0.4 \cdot 0.5 \cdot 0.4 + 0.6 \cdot 0.5 \cdot 0.6 + 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.32\\
&\mathbf{P}(X = 3) = 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 + 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.12\\
&\mathbf{P}(X = 4) = 0.4 \cdot 0.5 \cdot 0.6 \cdot 0.5 = 0.06\\
&\mathbf{P}(X > 4) = 0
\end{aligned}$$

Problem 11 (10 marks)

Find the PDF, the mean, and the variance of the random variable X with CDF

$$F_X(x) = \begin{cases} 1 - \frac{a^3}{x^3} & \text{if } x \geq a, \\ 0 & \text{if } x < a, \end{cases}$$

where a is a positive constant.

Solution: We have

$$f_X(x) = \frac{dF_X}{dx}(x) = \begin{cases} 3a^3x^{-4} & \text{if } x \geq a, \\ 0 & \text{if } x < a, \end{cases}$$

Also

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^{\infty} x \cdot 3a^3x^{-4} dx = 3a^3 \int_a^{\infty} x^{-3} dx = 3a^3 \left(-\frac{1}{2}x^{-2} \right) \Big|_a^{\infty} = \frac{3a}{2}.$$

Finally, we have

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_a^{\infty} x^2 \cdot 3a^3x^{-4} dx = 3a^3 \int_a^{\infty} x^{-2} dx = 3a^3 \left(-x^{-1} \right) \Big|_a^{\infty} = 3a^2,$$

so the variance is

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = 3a^2 - \left(\frac{3a}{2} \right)^2 = \frac{3a^2}{4}.$$

Problem 12 (5 marks)

A certain bag of fertilizer advertises that it contains 7.25 kg, but the amounts these bags actually contain is normally distributed with a mean of 7.4 kg and a standard deviation of 0.15 kg.

The company installed new filling machines, and they wanted to perform a test to see if the mean amount in these bags had changed. Their hypotheses were $H_0 : \mu = 7.4$ kg vs $H_a : \mu \neq 7.4$ kg (where μ is the true mean weight of these bags filled by the new machines).

They took a random sample of 50 bags and observed a sample mean and standard deviation of $\overline{x} = 7.36$ kg and $s_x = 0.12$ kg. They calculated that these results had a P-value of approximately 0.02.

What conclusion should be made using a significant level of $\alpha = 0.05$? Choose the correct answer and provide a brief justification.

(a) Fail to reject H_0.

(b) Reject H_0 and accept H_a

(c) Accept H_0

Solution: Option (b) is correct.
P-value $< \alpha \Rightarrow$ reject $H_0 \Rightarrow$ accept H_a
P-value $\geq \alpha \Rightarrow$ fail to reject H_0
Since the P-value of 0.02 is smaller than $\alpha = 0.05$, sample results as far or farther than $\overline{x} = 7.36$ are unlikely to happen by random chance alone when H_0 is true. In other words, if these machines were filling with a mean of 7.4 kg, there is about a 2% chance of getting a sample mean as far or farther away than 7.36 kg or 7.44 kg. This random chance probability is lower than the significance level $\alpha = 0.05$, so the results are unusual enough for us to reject H_0.

— End of Assignment —