LUMS
A Not-for-Profit University

Department of Electrical Engineering
School of Science and Engineering

EE212 Mathematical Foundations for Machine Learning and Data Science

ASSIGNMENT 3

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 (10 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) **[3 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) **[3 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) **[4 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.

Problem 2 (10 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 $k = 1, 2, \ldots,$

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

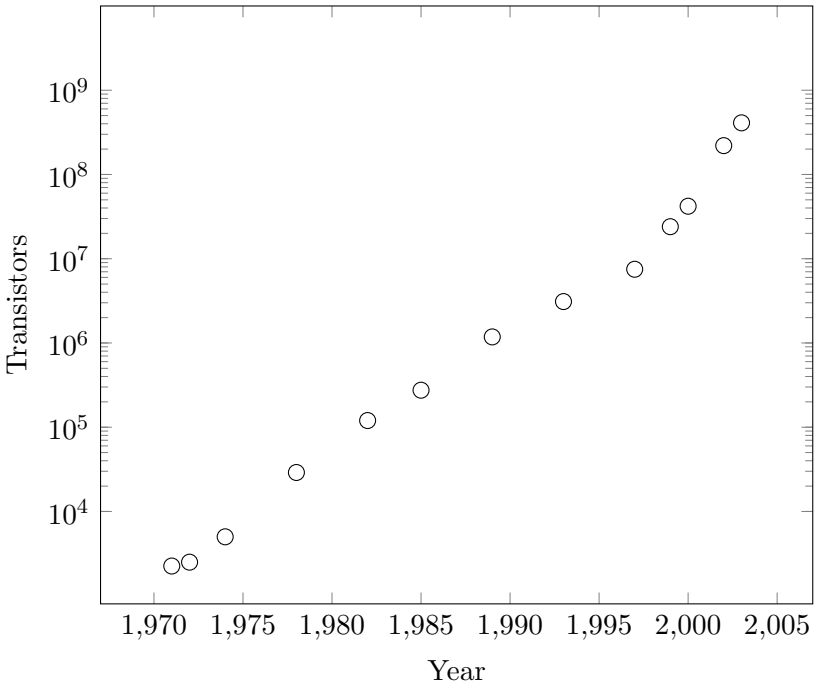
where μ is a positive parameter, and the superscripts denote the iteration number. This defines a sequence of vectors that converge to $\bar{x}$ provided μ is not too large; the choice $\mu = 1/\|A\|^2$, for example, always works. The iteration is terminated when $A^T(Ax^{(k)} - b)$ is small enough, 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) **[4 marks]** Show that if $x^{(k+1)} = x^{(k)}$, we have $x^{(k)} = \hat{x}$.

(b) **[6 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 $\hat{x}$.

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 θ_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.

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.

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

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.

Problem 6 (5 marks)

Consider the following functions:

$$f_1(x) = \sin(x_1)\cos(x_2), \quad x \in \mathbb{R}^2$$
$$f_2(x, y) = x^T y, \quad x, y \in \mathbb{R}^n$$

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

(b) Compute the Jacobians.

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.

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?

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 i the coin is fair or unfair? How can we generalize Alice's reasoning?

Problem 10 (10 marks)

Your favourite soccer team has 2 games scheduled for one weekend. Team 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. Your team 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 team earns over the weekend. Find the PMF of X.

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.

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

— End of Assignment —