Machine Learning Homework1
Lachin Naghashyar - 98110179

1.1 Assume that x is a vector and A is a square matrix. Show that $\frac{\partial x^T A x}{\partial x} = x^T \left(A + A^T\right)$

Proof: suppose that A is a $n \times n$ matrix and let $x = \left(x_1, x_2, \ldots, x_n\right)^T$. This way, $f(x) = x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$ will be an scaler and hence its derivative with respect to x is defined by $\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^T$. Hence we have:

$$\frac{\partial f}{\partial x_k} = \sum_{i \neq k} \frac{\partial \sum_{j=1}^{n} a_{ij} x_i x_i}{\partial x_k} + \frac{\sum_{j=1}^{n} a_{kj} x_k x_j}{\partial x_k} = \sum_{i \neq k} a_{ik} x_i + \sum_{j \neq k} a_{kj} x j + 2 a_{kk} x_k$$
$$= \sum_{i=1}^{n} a_{ik} x_i + \sum_{j=1}^{n} a_{kj} x_i = \sum_{i=1}^{n} \left(a_{ik} + a_{ki}\right) x_i = \left(x^T \left(A + A^T\right)\right)_k$$

Which is the $k'th$ element of the vector $x^T(A + A^T)$ and using this, we can derive the result.

Note that since $f(x) = x^T A x$ is a scaler, $trace(x^T A x) = x^T A x$ and the result is the same. Also for a symmetric matrix A, we will have $A^T = A$ and thus result becomes $2x^T A$.

1.2 Suppose $\lambda_1, \ldots, \lambda_n$ are eigenvalues of matrix A. Prove:
- $\lambda_1 + \ldots + \lambda_n = trace(A)$
- $\lambda_1 \ldots \lambda_n = det(A)$

Proof: We know that the solutions to the characteristic equation of A are the eigenvalues of A and hence we can write its characteristic polynomial in two ways:

- $p(\lambda) = \det(A - \lambda I) = (-1)^n \left(\lambda^n + c_{n-1}\lambda^{n-1} + \ldots + c_1 \lambda + c_0\right)$
- $p(\lambda) = (-1)^n \left(\lambda - \lambda_1\right) \ldots \left(\lambda - \lambda_n\right)$

Now, to derive the above results, we're going to use the fact that these two polynomials are the same and hence the corresponding coefficients should match.

First consider the constant term in each of them. It's equal to $(-1)^n c_0$ in the first one and $(-1)^n \lambda_1 \ldots \lambda_n$ in the other one. Moreover, we can substitute $\lambda = 0$ in the first equation and have $p(0) = det(0 - A) = det(-A) = (-1)^n det(A) = (-1)^n c_0$. By using these equalities, we can derive the result $\lambda_1 \ldots \lambda_n = det(A)$.

For the second part we should consider the coefficient of the term λ^{n-1} which is $(-1)^{n-1}(\lambda_1 + \ldots + \lambda_n)$ (using Vieta's formula) in one and $(-1)^n c_{n-1}$ in the other one.

I Using the formula provided in Wikipedia for calculating the determinant of a matrix, we have $det(A) = \Sigma_{\sigma \in S_n}(sgn(\sigma)\Pi_{i=1}^{n} a_{i,\sigma_i})$ where S_n denotes all the permutations of set $\{1,2,\ldots,n\}$. Moreover, since we are interested in the coefficient of λ^{n-1}, by expanding $A - \lambda I$, we will only be able to have λ^{n-1} if we choose all the elements of the permutations from the diagonal or else the most we can get is the power of $n-2$. Also if we know $n-1$ elements of the permutation, the last element will also be known. Hence, we can derive the coefficient of λ^{n-1} using $(\lambda - a_{1,1})(\lambda - a_{2,2})\ldots(\lambda - a_{n,n})$. Which turns out to be $(-1)^n(a_{1,1} + \ldots + a_{n,n})$.
Therefore, we have $-c_{n-1} = (\lambda_1 + \ldots + \lambda_n) = (a_{1,1} + \ldots + a_{n,n}) = trace(A)$

1.3 Suppose X and Y are independent normal random variables with mean μ and variance 1 , where $\mu \sim Uni(0,1)$.

$$f_\mu(t) = \begin{cases} 1, & \text{t} \in [0,1] \\ 0, & \text{elsewhere} \end{cases}$$

- Find the joint distribution of μ, X, Y . $\left(f_{\mu,X,Y}(t\,.\,x\,.\,y)\right)$.
- Find the MAP estimate of μ.

Proof:

$$\xrightarrow{\text{X and Y independent}} f_{\mu,X,Y}(t,x,y) = f_\mu(t) f_{X|\mu}\left(x\,|\,t\right) f_{Y|\mu}\left(y\,|\,t\right)$$

$$\xrightarrow{\text{X and Y are normal}} \rightarrow f_{X|\mu}\left(x\,|\,t\right) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-t)}{2}}, f_{Y|\mu}\left(y\,|\,t\right) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-t)}{2}}$$

$$\rightarrow f_{\mu,X,Y}\left(t,x,y\right) = \begin{cases} \dfrac{1}{2\pi} e^{\dfrac{-(x-t)^2 - \left(y-t\right)^2}{2}} & t \in \left[0,1\right] \\ 0 \text{ o.w.} \end{cases}$$

Also for the MAP estimation we want to find a θ that for a given x and y, maximizes $f_{X,Y|\mu}(x,y\,|\,\theta)f_\mu(\theta)$.
Which is equivalent to $A = f_{X,Y,\mu}(X = x, Y = y, \mu = \theta)$ and in order to find the maximum of it, we take derivative with respect to θ. And since A is 0 when θ is not in $[0,1]$, we need to consider $\dfrac{e^{\frac{-(x-\theta)^2 - (y-\theta)^2}{2}}}{2\pi}$ and if we take its derivative with respect to θ, we will have $\dfrac{(-2\theta + x + y)e^{\frac{-(x-\theta)^2 - (y-\theta)^2}{2}}}{2\pi}$ and in order for it to be zero, we should have $-2\theta + x + y = 0$ so we should have $\theta = \dfrac{(x+y)}{2}$.

1.4 Calculate the L1 norm, the Euclidean (L2) norm and the Maximum(L infinity) norm of the following matrix :

$$A = \begin{bmatrix} 5 & -4 & 2 \\ -1 & 2 & 3 \\ -2 & 1 & 0 \end{bmatrix}$$

Proof:

$$\| A \|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} \left| Aij \right| = \max\left(8{,}7{,}5\right) = 8$$

$$\| A \|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| aij \right|_n^2} = \sqrt{25 + 16 + 4 + 1 + 4 + 9 + 4 + 1} = 8$$

$$\left| \left| A \right| \right|_\infty = \max_{i=1,\dots,n} \sum_{j=1} \left| Aij \right| = max\left(11{,}6{,}3\right) = 11$$

$$\| A \|_2 = \text{largest eigen value of } \left(A^T A\right)^{\frac{1}{2}}$$

$$A^T A = \begin{bmatrix} 5-1-2 \\ -4 \quad 2 \quad 1 \\ 2 \quad 3 \quad 0 \end{bmatrix} \begin{bmatrix} 5-42 \\ -1 \quad 23 \\ -2 \quad 10 \end{bmatrix} = \begin{bmatrix} 30-24 \quad 7 \\ -24 \quad 21-2 \\ 7 \quad -2 \ 13 \end{bmatrix}$$

$$\text{characteristic polynomial}: -\lambda^3 + 64\lambda^2 - 664\lambda + 225$$

$$\text{Eigenvalues}: \lambda_1 \approx 51, \lambda_2 \approx 12 \cdot 5, \lambda_3 \approx 0.35$$

$$\rightarrow \| A \|_2 = \sqrt{51} = 7.14$$

2.1 Let the joint probability density function of random variables X be :

$$f(x) = \begin{cases} 2x - 2, & 1 < x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

Find $\mathbb{E}\left(X^3 + 2X - 7\right)$.

Proof:

$$E\left(g(x)\right) = \int_{-\infty}^{\infty} g(x) f(x)\, dx \rightarrow E\left(x^3 + 2x - 7\right) = \int_{-\infty}^{\infty} \left(x^3 + 2x - 7\right)(2x - 2)\, dx$$

$$= \int^{2} \left(x^3 + 2x - 7\right)(2x - 2)\, dx = \int_1^2 \left(2x^4 + 4x^2 - 18x - 2x^3 + 14\right) dx$$

$$= \left(\frac{x^5}{5} + 4\frac{x^3}{3} - 9x^2 - \frac{x^4}{2} + 14x\right)\Big|_1^2 = \frac{37}{30}$$

2.2 Suppose $X_1, X_2, \ldots, X_n$ are iid random variables. Find the probability density function of $Y_1 = \max\left[X_1, X_2, \ldots, X_n\right]$, $Y_2 = \min\left[X_1, X_2, \ldots, X_n\right]$.

Proof:

$$F_{Y_1}(y) = P\left(Y_1 \le y\right) = P\left(X_1 \le y, X_2 \le y, \ldots, X_n \le y\right) \xrightarrow{\text{independent}}$$
$$= \prod_{i=1}^{n} P\left(X_i \le y\right) = \prod_{i=1}^{n} F_{X_i}(y) \xrightarrow{\text{same dist}} F_x(y)^n$$
$$\rightarrow f_{Y_1}(y) = F_{Y_1}(y)' = nf(y)\, F_X(y)^{n-1}$$

$$F_{Y_2}(y) = P\left(Y_2 \le y\right) = 1 - P\left(Y_2 > y\right) = 1 - P(X_1 > y, \ldots, X_n > y)$$
$$P\left(X_1 > y, \ldots, X_n > y\right) = \prod_{i=1}^{n} P\left(X_i > y\right) = \prod_{i=1}^{n}\left(1 - F_{X_i}(y)\right) = \left(1 - F_X(y)\right)^n$$
$$\rightarrow F_{Y_2}(y) = 1 - \left[1 - F_X(y)\right]^n, f_{Y_2}(y) = F_{Y_2}(y)' = n\left[1 - F_X(y)\right]^{n-1} f_X(y)$$

2.3 Prove that if P is a full rank matrix, matrices M and $\mathrm{P}^{-1}\mathrm{MP}$ have the same set of eigenvalues.

Proof: we prove this in two steps:

$$1. \forall \lambda \in \left\{\text{eigenvalues of M}\right\} : \lambda \in \left\{\text{eigenvalues of } P^{-1}MP\right\}$$
$$\text{if } \exists A : MA = \lambda A \rightarrow \text{ since P is full rank,there exists B such that} : B = P^{-1}A \rightarrow A = PB$$
$$\rightarrow MPB = \lambda PB \xrightarrow{\times P^{-1}} P^{-1}MPB = \lambda P^{-1}PB = \lambda B \rightarrow \lambda \in \left\{\text{eigenvalues of } P^{-1}MP\right\}$$
$$2. \forall \lambda \in \left\{\text{eigenvalues of } P^{-1}MP\right\} : \lambda \in \left\{\text{eigenvalues of M}\right\}$$
$$\text{if } \exists A : P^{-1}MPA = \lambda A \xrightarrow{\times P} MPA = \lambda PA \rightarrow MB = \lambda B \rightarrow \lambda \in \left\{\text{eigenvalues of M}\right\}$$