The Australian National University
School of Computing

Semester 2, 2024
Assignment 2, Part 1, Theory

COMP3670/6670: Introduction to Machine Learning

Due Date. 5:00pm, Sept 25th, 2024

Maximum credit. 10

Marking Rubric. The entire assignment will be graded out of a total of 10 credits.

In Exercise 1, you will need to prove four conditions (I,II,III,IV) are satisfied for each part. In Exercise 1, your proof for each conditions will sum up to the total credits (2+3) assigned to Exercise 1.

For each Exercise 1.1 conditions I,II,III,IV, Exercise 1.2 conditions I,II,III,IV, for Exercise 2, and 3, we will follow the grading rubric listed below.

100% You will receive 100% of the available credits for a question if your answer is completely correct **and** you show all the correct steps **and** your steps are clearly laid out.

50% You will receive 50% of the available credits for a question if your answer is not correct **or** you show mostly correct steps **and** your steps are clearly laid out.

0% You will otherwise receive 0% of the available credits for a question.

Sample partial solutions will be released with the grades after the submission deadline.

Exercise 1 — Matrix Norm (2 + 3 credits)

We learnt vector norm from the lectures, and it natural to think about its generalisation, **matrix norm**. We call a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ a matrix norm, if for all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}, c \in \mathbb{R}$, the following conditions are satisfied:

I. $\|\mathbf{A}\| \geq 0$ with equality attained if and only if $\mathbf{A}$ is a zero matrix

II. $\|c\mathbf{A}\| \leq |c|\|\mathbf{A}\|$

III. $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

IV. $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$

Consider the following problems:

1. Suppose $\|\cdot\|$ is a matrix norm on $\mathbb{R}^{n\times n}$. Prove whether the function $\|\cdot\|_{\mathbf{C}} : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined by $\|\mathbf{A}\|_{\mathbf{C}} = \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\|$ is a matrix norm for arbitrary ~~invertible~~ $\mathbf{A} \in \mathbb{R}^{n\times n}$ and invertible $\mathbf{C} \in \mathbb{R}^{n\times n}$[1]
.

   Solution: prove (1) 0.25pts, (2) 0.25pts, (3) 0.5pts, (4) 1pt

[1] Updated for clarity, 05/09/2024

1. **Proof of Condition I: Non-negativity** (1) 0.25pts

We need to show that $\|\mathbf{A}\|_{\mathbf{C}} \geq 0$ and $\|\mathbf{A}\|_{\mathbf{C}} = 0$ if and only if $\mathbf{A} = 0$.

Since $\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\|$ is a matrix norm, $\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\| \geq 0$. Therefore, $\|\mathbf{A}\|_{\mathbf{C}} \geq 0$.

If $\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\| = 0$, then $\mathbf{C}^{-1}\mathbf{A}\mathbf{C} = 0$, and since $\mathbf{C}$ is invertible, it follows that $\mathbf{A} = 0$. Conversely, if $\mathbf{A} = 0$, then $\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\| = 0$. Hence, $\|\mathbf{A}\|_{\mathbf{C}} = 0$ if and only if $\mathbf{A} = 0$.

2. **Proof of Condition II: Absolute homogeneity** (2) 0.25pts

We need to show that $\|c\mathbf{A}\|_{\mathbf{C}} = |c|\|\mathbf{A}\|_{\mathbf{C}}$ for any scalar $c \in \mathbb{R}$.

Using the definition of $\|\mathbf{A}\|_{\mathbf{C}}$, we have:

$$\|c\mathbf{A}\|_{\mathbf{C}} = \|\mathbf{C}^{-1}(c\mathbf{A})\mathbf{C}\| = \|c(\mathbf{C}^{-1}\mathbf{A}\mathbf{C})\|$$

Since $\|\cdot\|$ is a matrix norm, it satisfies the property $\|cM\| = |c|\|M\|$, which gives us:

$$\|c(\mathbf{C}^{-1}\mathbf{A}\mathbf{C})\| = |c|\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\| = |c|\|\mathbf{A}\|_{\mathbf{C}}$$

Therefore, $\|c\mathbf{A}\|_{\mathbf{C}} = |c|\|\mathbf{A}\|_{\mathbf{C}}$.

3. **Proof of Condition III: Triangle inequality** (3) 0.5pts

We need to show that $\|\mathbf{A}+\mathbf{B}\|_{\mathbf{C}} \leq \|\mathbf{A}\|_{\mathbf{C}} + \|\mathbf{B}\|_{\mathbf{C}}$.

Using the definition of $\|\mathbf{A}\|_{\mathbf{C}}$, we have:

$$\|\mathbf{A}+\mathbf{B}\|_{\mathbf{C}} = \|\mathbf{C}^{-1}(\mathbf{A}+\mathbf{B})\mathbf{C}\| = \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C} + \mathbf{C}^{-1}\mathbf{B}\mathbf{C}\|$$

Since $\|\cdot\|$ is a matrix norm, it satisfies the triangle inequality:

$$\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C} + \mathbf{C}^{-1}\mathbf{B}\mathbf{C}\| \leq \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\| + \|\mathbf{C}^{-1}\mathbf{B}\mathbf{C}\| = \|\mathbf{A}\|_{\mathbf{C}} + \|\mathbf{B}\|_{\mathbf{C}}$$

Therefore, $\|\mathbf{A}+\mathbf{B}\|_{\mathbf{C}} \leq \|\mathbf{A}\|_{\mathbf{C}} + \|\mathbf{B}\|_{\mathbf{C}}$.

4. **Proof of Condition IV: Sub-multiplicativity** (4) 1pt

We need to show that $\|\mathbf{A}\mathbf{B}\|_{\mathbf{C}} \leq \|\mathbf{A}\|_{\mathbf{C}}\|\mathbf{B}\|_{\mathbf{C}}$.

Using the definition of $\|\mathbf{A}\|_{\mathbf{C}}$, we have:

$$\|\mathbf{A}\mathbf{B}\|_{\mathbf{C}} = \|\mathbf{C}^{-1}\mathbf{A}\mathbf{B}\mathbf{C}\| = \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\mathbf{C}^{-1}\mathbf{B}\mathbf{C}\| = \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\|\|\mathbf{C}^{-1}\mathbf{B}\mathbf{C}\|$$

Since $\|\cdot\|$ is a matrix norm, it satisfies the sub-multiplicative property:

$$\|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\mathbf{C}^{-1}\mathbf{B}\mathbf{C}\| \leq \|\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\|\|\mathbf{C}^{-1}\mathbf{B}\mathbf{C}\|$$

Therefore, $\|\mathbf{AB}\|_{\mathbf{C}} \leq \|\mathbf{A}\|_{\mathbf{C}}\|\mathbf{B}\|_{\mathbf{C}}$.

In conclusion, since $\|\cdot\|_{\mathbf{C}}$ satisfies all four properties, it is a matrix norm.

2. A function $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ is defined as $\|\mathbf{A}\| = \max_{i\in\{1\cdots n\}} \sum_{j=1}^{n} |A_{ij}|$. The expression computes the maximum value of the column-sum of absolute values ($|\cdot|$) of the entries for all rows . Prove whether this function is a matrix norm.

Solution: prove (1) 0.25pts, (2) 0.25pts, (3) 1pt, (4) 1.5pts

1. **Proof of Condition I: Non-negativity** (1) 0.25pts

We need to show that $\|\mathbf{A}\| \geq 0$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$.

Firstly, since $\|\mathbf{A}\|$ is defined as the maximum of sums of absolute values of matrix entries, it is trivially non-negative, i.e., $\|\mathbf{A}\| \geq 0$.

Secondly, if $\|\mathbf{A}\| = 0$, then all elements in matrix $\mathbf{A}$ must be zero (as the absolute values of the entries sum to zero), implying $\mathbf{A} = 0$. Conversely, if $\mathbf{A} = 0$, it is clear that $\|\mathbf{A}\| = 0$.

Thus, $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$.

2. **Proof of Condition II: Absolute homogeneity** (2) 0.25pts

We need to show that $\|c\mathbf{A}\| = |c|\|\mathbf{A}\|$ for any scalar $c \in \mathbb{R}$.

Using the definition of $\|\mathbf{A}\|$, we have:

$$\|c\mathbf{A}\| = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |cA_{ij}| = |c| \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |A_{ij}|$$

Thus, $\|c\mathbf{A}\| = |c|\|\mathbf{A}\|$, which satisfies the property of absolute homogeneity.

3. **Proof of Condition III: Triangle inequality** (3) 1pt

We need to show that $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

Using the definition of $\|\mathbf{A}\|$, we have:

$$\|\mathbf{A} + \mathbf{B}\| = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |A_{ij} + B_{ij}|$$

By the triangle inequality for real numbers, we know that:

$$|A_{ij} + B_{ij}| \leq |A_{ij}| + |B_{ij}|$$

Therefore, we can write:

$$\sum_{j=1}^{n} |A_{ij} + B_{ij}| \le \sum_{j=1}^{n} |A_{ij}| + \sum_{j=1}^{n} |B_{ij}|$$

Taking the maximum over all i, we get:

$$\|\mathbf{A} + \mathbf{B}\| \le \|\mathbf{A}\| + \|\mathbf{B}\|$$

Thus, the triangle inequality holds for this norm.

4. **Proof of Condition IV: Sub-multiplicativity** (4) 1.5pts

We need to show that $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$.

Using the definition of matrix multiplication, we have:

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Thus, the norm $\|\mathbf{AB}\|$ is given by:

$$\|\mathbf{AB}\| = \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} \left| \sum_{k=1}^{n} A_{ik} B_{kj} \right|$$

By applying the triangle inequality to the sum inside the absolute value, we have:

$$\left| \sum_{k=1}^{n} A_{ik} B_{kj} \right| \le \sum_{k=1}^{n} |A_{ik}||B_{kj}|$$

Thus, we can write:

$$\|\mathbf{AB}\| \le \max_{i \in \{1,\ldots,n\}} \sum_{k=1}^{n} |A_{ik}| \max_{k \in \{1,\ldots,n\}} \sum_{j=1}^{n} |B_{kj}|$$

This is equivalent to:

$$\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$$

Thus, the sub-multiplicativity condition is satisfied.

In conclusion, since $\|\cdot\|$ satisfies all four conditions, it is a matrix norm.

To prove a function is a matrix norm, you need to show all the conditions I-IV listed above are satisfied for arbitrary matrix in the domain.

Exercise 2 **Positive Definite Matrices** (2 credits)

Given that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite, it is known that any diagonal entry of $\mathbf{A}$, A_{ii} is positive. Prove whether matrix $\mathbf{B}$ with $B_{ij} = A_{ij}(A_{ii}A_{jj})^k$ is also positive definite, where $k \in \mathbb{R}$.

[Hint] You might want to try to convert between the vector form and the scalar form to solve this question.

Solution: We are given that matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite, and we need to prove that matrix $\mathbf{B}$, defined by $B_{ij} = A_{ij}(A_{ii}A_{jj})^k$, is also positive definite for any $k \in \mathbb{R}$.

1. **Understanding the quadratic form** (0.5pts)

We start by using the quadratic form for a matrix $\mathbf{A}$, where we know that for a positive definite matrix $\mathbf{A}$, we have:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j$$

Now, for matrix $\mathbf{B}$, the quadratic form $\mathbf{x}^T \mathbf{B} \mathbf{x}$ is given by:

$$\mathbf{x}^T \mathbf{B} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i B_{ij} x_j = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} (A_{ii} A_{jj})^k x_j$$

2. **Rewriting the quadratic form in terms of diagonal scaling** (1pts)

We can factor out the diagonal terms involving A_{ii}^k and A_{jj}^k. Let's define a diagonal matrix $\text{diag}(\mathbf{A})^k$ with entries $(A_{ii})^k$. Now, the quadratic form can be rewritten as:

$$\mathbf{x}^T \mathbf{B} \mathbf{x} = \mathbf{x}^T \left(\text{diag}(\mathbf{A})^k \mathbf{A} \text{diag}(\mathbf{A})^k \right) \mathbf{x}$$

This expression can be seen as applying diagonal scaling to both sides of matrix $\mathbf{A}$. Let's define a new vector $\mathbf{y} = \text{diag}(\mathbf{A})^k \mathbf{x}$, so the quadratic form becomes:

$$\mathbf{x}^T \mathbf{B} \mathbf{x} = \mathbf{y}^T \mathbf{A} \mathbf{y}$$

3. **Proving positive definiteness of $\mathbf{B}$** (0.5pts)

Since $\mathbf{A}$ is positive definite, we know that for any non-zero vector $\mathbf{y}$, we have:

$$\mathbf{y}^T \mathbf{A} \mathbf{y} > 0$$

Therefore, $\mathbf{x}^T \mathbf{B} \mathbf{x} = \mathbf{y}^T \mathbf{A} \mathbf{y} > 0$ for any non-zero vector $\mathbf{x}$, which proves that $\mathbf{B}$ is positive definite.

Hence, matrix $\mathbf{B}$ is positive definite for any $k \in \mathbb{R}$.

Exercise 3 **Matrix Calculus** (3 credits)

Consider the following function of $\mathbf{X} \in \mathbb{R}^{n\times m}$:

$$f(\mathbf{X}) = \exp\left\{-\frac{1}{2}\mathbf{tr}\left[(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{V})\right]\right\}$$

where $\mathbf{M} \in \mathbb{R}^{n\times m}$, $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\mathbf{V} \in \mathbb{R}^{m\times m}$, and $\mathbf{tr}$ is the trace. $\mathbf{U}, \mathbf{V}$ are positive definite.

Find the gradient of f, $\nabla_{\mathbf{X}} f$. Cite the rules you use and show all your steps clearly.

[Hint] You may apply the rules from the lecture and other identities like the ones here.

[Note] Beware of the difference between the numerator layout and denominator layout in material you come across. We are using the numerator layout convention in this course.

Solution: We are asked to find the gradient of the function

$$f(\mathbf{X}) = \exp\left(-\frac{1}{2}\,\mathrm{tr}\left[(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\,\mathbf{U}^{-1}\,(\mathbf{X}-\mathbf{M})\,\mathbf{V}\right]\right),$$

where $\mathbf{M} \in \mathbb{R}^{n\times m}$, $\mathbf{U} \in \mathbb{R}^{n\times n}$, and $\mathbf{V} \in \mathbb{R}^{m\times m}$ are positive definite matrices.

1. **Applying the chain rule correctly** (0.5pts)

We first apply the chain rule to differentiate $f(\mathbf{X})$ with respect to $\mathbf{X}$. Define the intermediate function Y as:

$$Y = -\frac{1}{2}\,\mathrm{tr}\left[(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\,\mathbf{U}^{-1}\,(\mathbf{X}-\mathbf{M})\,\mathbf{V}\right]$$

Using the chain rule, we get:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = \exp(Y)\,\nabla_{\mathbf{X}} Y$$

2. **Finding the derivative of the trace expression with respect to $\mathbf{X}$** (2pts)

Let $\mathbf{Z} = \mathbf{X} - \mathbf{M}$. Now, the expression for Y becomes:

$$Y = -\frac{1}{2}\,\mathrm{tr}\left[\mathbf{Z}^{\mathrm{T}}\,\mathbf{U}^{-1}\,\mathbf{Z}\,\mathbf{V}\right]$$

We need to compute $\nabla_{\mathbf{X}} Y$. Since $\mathbf{M}$ is constant, we have $\nabla_{\mathbf{X}}\mathbf{Z} = \mathbf{I} \otimes \mathbf{I}$, where $\otimes$ denotes the Kronecker product. Thus:

$$\nabla_{\mathbf{X}} Y = \nabla_{\mathbf{Z}} Y$$

Applying the matrix calculus rule:

$$\nabla_{\mathbf{Z}} \operatorname{tr}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z}\mathbf{B}\right) = \mathbf{A}\mathbf{Z}\mathbf{B} + \left(\mathbf{A}\mathbf{Z}\mathbf{B}\right)^{\mathrm{T}},$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices, we obtain:

$$\begin{aligned}\nabla_{\mathbf{Z}} Y &= -\frac{1}{2}\left[\mathbf{U}^{-1}\mathbf{Z}\mathbf{V} + \left(\mathbf{U}^{-1}\mathbf{Z}\mathbf{V}\right)^{\mathrm{T}}\right] \\ &= -\frac{1}{2}\left[\mathbf{U}^{-1}\mathbf{Z}\mathbf{V} + \mathbf{V}\mathbf{Z}^{\mathrm{T}}\mathbf{U}^{-1}\right]\end{aligned}$$

Since $\mathbf{U}$ and $\mathbf{V}$ are symmetric (positive definite), and assuming that $\mathbf{Z}$ is a real matrix, we can simplify the expression:

$$\nabla_{\mathbf{Z}} Y = -\mathbf{U}^{-1}\mathbf{Z}\mathbf{V}$$

3. **Final result** (0.5pts)

Substituting back into the expression for the gradient of $f(\mathbf{X})$, we have:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = \exp\left(Y\right)\left(-\mathbf{U}^{-1}\mathbf{Z}\mathbf{V}\right)$$

Since $f(\mathbf{X}) = \exp\left(Y\right)$, we can write:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = -f(\mathbf{X})\, \mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{V}$$

Thus, the gradient of $f(\mathbf{X})$ is:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = -\exp\left(-\frac{1}{2}\operatorname{tr}\left[(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\,\mathbf{U}^{-1}\,(\mathbf{X}-\mathbf{M})\,\mathbf{V}\right]\right)\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{V}$$

Note: If the answer differs only by a transpose, full marks should be awarded. Some may write the gradient as:

$$\nabla_{\mathbf{X}} f(\mathbf{X}) = -f(\mathbf{X})\, \mathbf{V}(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\mathbf{U}^{-1}$$

This form is acceptable for this assignment since:

$$\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{V} = \left(\mathbf{V}(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\mathbf{U}^{-1}\right)^{\mathrm{T}}$$

Since the transpose of a gradient is still acceptable, and the expression may vary depending on the conventions used, both forms are correct, and full marks should be given if the answer differs only by a transpose.

Acknowledgement of Assignment 2, Part 2: Please ensure that in addition to submitting the answers to these exercises, you complete Assignment 2: Part 2 out of 2 (timed online assessment, 50mins). The timed online assessment deadline is 5pm, Sep 25, 2024.

Acknowledgement of Academic Integrity By submitting Assignment 2 you attest to understand and commit to the pledge described according to university policy and detailed in Assignment 1.