ML2025 Fall Homework Assignment 1
Handwritten

Yu-Cheng Lin
b11901152@ntu.edu.tw

Lu-Fang Chiang
b12901140@ntu.edu.tw

September 2025

Tools You Need to Know/Learn

- Rank, Nullity, Vector Space, Inner Product Space
- Determinant, Trace, Inverse Matrix, Adjoint Matrix
- Symmetric Matrix, Orthogonal Matrix, Diagonalization
- Positive Definite, Positive Semidefinite
- Eigenvalue Decomposition, Singular Value Decomposition (SVD)

You are expected to know (or learn through this homework) the definitions and usages of the above concepts, and are encouraged to discuss with TAs during TA hour if you encounter difficulties.

Homework Policy

- The homework is graded out of 100 points, with up to 15 additional bonus points available (only awarded for completely correct solutions).
- The official submission deadline will follow the schedule announced on NTU COOL.
- Homework may be handwritten or typed (e.g. using LaTeX), but must be submitted in **PDF format**.
- If you discuss the homework with classmates, you should state their student IDs in your submission.
- Plagiarism is strictly prohibited. Serious violations will be dealt with according to NTU regulations.

Problem 1 (Linear Algebra Recap) (30 pts)

In the following problems, we begin by recapping key topics from undergraduate linear algebra. We then proceed to derive several fundamental theorems and lemmas, which will play a central role in the forthcoming lectures. In the following problem, we assume V is an inner product space over $\mathbb{R}$ with $\dim V = n$.

Definition 1.1. (Symmetric transformation) A linear transformation $\mathbf{T} : V \to V$ is called symmetric if

$$\langle \mathbf{Tu}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{Tv} \rangle, \quad \forall\, \mathbf{u}, \mathbf{v} \in V.$$

If $\mathbf{A}$ is the matrix representation of $\mathbf{T}$ and $V \cong \mathbb{R}^n$, this condition is equivalent to

$$\mathbf{u}^\top \mathbf{Av} \;=\; \mathbf{u}^\top \mathbf{A}^\top \mathbf{v}, \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Definition 1.2. (positive semidefinite) A symmetric linear transformation $\mathbf{T} : V \to V$ is called positive semidefinite if

$$\langle \mathbf{Tu}, \mathbf{u} \rangle \;\geq\; 0, \quad \forall\, \mathbf{u} \in V.$$

Equivalently, for the matrix representation $\mathbf{A}$ of $\mathbf{T}$, this condition is equivalent to $\mathbf{A}$ is symmetric and

$$\mathbf{u}^\top \mathbf{Au} \;\geq\; 0, \quad \forall\, \mathbf{u} \in \mathbb{R}^n.$$

(a) (10 pts) Prove the following statements are equivalent.

(i) $\mathbf{A}$ is positive semidefinite.

(ii) $\mathbf{A}$ is symmetric and all of the eigenvalues of $\mathbf{A}$ are all non-negative.

(iii) There exists some square matrix $\mathbf{B}$ such that $\mathbf{A} = \mathbf{B}^\mathsf{T}\mathbf{B}$.

(Hint: The symmetric matrix is orthogonally diagonalizable).

In the following section, we will prove an important theorem in machine learning, namely the Singular Value Decomposition (SVD).

Theorem 1 (Singular Value Decomposition)**.** For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, there exist orthogonal matrices $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$, and a diagonal matrix $\Sigma \in \mathbb{R}^{m \times n}$ with nonnegative diagonal entries such that

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top.$$

Consider $\mathbf{A}^\top\mathbf{A}$. It is symmetric and hence orthogonally diagonalizable. Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal set of eigenvectors of $\mathbf{A}^\top\mathbf{A}$ with $\mathbf{A}^\top\mathbf{A}\,\mathbf{v}_i = \lambda_i \mathbf{v}_i$

(b) (3 pts) Prove that λ_i are nonnegative $\forall i = 1, \cdots, n$.

(c) (5 pts) Prove that $\text{Null}(\mathbf{A}^\mathsf{T}\mathbf{A}) = \text{Null}(\mathbf{A})$

Suppose $\sigma_i \coloneqq \sqrt{\lambda_i}$ for each i, and let $r = \text{rank}(\mathbf{A})$. By reordering the set $\{\sigma_1, \ldots, \sigma_n\}$, we may assume that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

By problem (c), we know $\sigma_r > 0$ and $\sigma_k = 0 \quad \forall k \in \{r+1, \cdots, n\}$.

(d) (5 pts) Prove that $\{\mathbf{Av}_1, \ldots, \mathbf{Av}_r\}$ is an orthogonal set.

(e) (7 pts) Let $\mathbf{u}_i = \dfrac{\mathbf{Av}_i}{\sigma_i}$, complete the proof the SVD.
(Hint: You may start with $\mathbf{A}\,(\mathbf{v}_1, \ldots, \mathbf{v}_n) = (\sigma_1\mathbf{u}_1, \ldots, \sigma_r\mathbf{u}_r, 0, \cdots, 0)$ and using Gram–Schmidt process. Don't forget to show that $\mathbf{U}, \mathbf{V}$ are orthogonal matrices.)

Problem 2 (Definition of Derivative as Linear Operator) (10 pts)

We now introduce the rigorous definition of the derivative in finite-dimensional vector spaces. The key idea is that the derivative at a point is a linear operator that best approximates the change of the function.

General definition (Fréchet derivative). Let V, W be finite-dimensional real inner product spaces, and let $f : V \to W$. We say that f is differentiable at $\mathbf{x} \in V$ if there exists a linear operator

$$Df(\mathbf{x}) : V \to W$$

such that

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\| f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})[\mathbf{h}] \|_W}{\|\mathbf{h}\|_V} = 0,$$

where $\|\cdot\|_V$ and $\|\cdot\|_W$ denote the norms on V and W respectively. The operator $Df(\mathbf{x})$ is called the derivative (or Jacobian operator) of f at $\mathbf{x}$.

Remark (scalar-valued functions). If $W = \mathbb{R}$, then $Df(\mathbf{x}) : V \to \mathbb{R}$ is a linear transformation. In this case the definition is equivalent to

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = Df(\mathbf{x})[\mathbf{h}] + o(\|\mathbf{h}\|_V), \qquad \|\mathbf{h}\| \to 0.$$

If there exists a vector $g \in V$ such that

$$Df(\mathbf{x})[\mathbf{h}] = \langle g, \mathbf{h} \rangle, \quad \forall \mathbf{h} \in V,$$

we call this vector g the gradient of f at $\mathbf{x}$, denoted $\nabla f(\mathbf{x})$. It turns out that a fundamental result in mathematics (the Riesz representation theorem) guarantees that such a vector always exists and is unique (you don't need to know the details in this course, but you can look it up if you're interested).

Coordinate form.

- If $V = \mathbb{R}^n$, $W = \mathbb{R}$ with the standard Euclidean inner product, then
$$Df(\mathbf{x})[d\mathbf{x}] = \nabla f(\mathbf{x})^{\mathsf{T}} d\mathbf{x}.$$
- If $V = \mathbb{R}^{m \times n}$, $W = \mathbb{R}$ with the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle_F = \operatorname{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{B})$, then
$$Df(\mathbf{X})[d\mathbf{X}] = \langle \nabla f(\mathbf{X}), d\mathbf{X} \rangle_F = \operatorname{tr}\left(\nabla f(\mathbf{X})^{\mathsf{T}} d\mathbf{X}\right).$$

Chain rule. If $f : V \to W$ and $g : U \to V$ are differentiable at $\mathbf{x} \in U$, then their composition $h = f \circ g : U \to W$ is differentiable at $\mathbf{x}$ and satisfies

$$Dh(\mathbf{x}) = Df(g(\mathbf{x})) \circ Dg(\mathbf{x}).$$

In the scalar-valued case $f : \mathbb{R} \to \mathbb{R}$ and $g : V \to \mathbb{R}$, this reduces to the familiar chain rule

$$\nabla(f \circ g)(\mathbf{x}) = f'(g(\mathbf{x}))\, \nabla g(\mathbf{x}).$$

All the definitions above are just to make sure our calculus is mathematically solid. For the following problems, please use this rigorous way of thinking — that is, start from the small-increment approximation when writing your proofs (which means you can still take derivatives/gradients in the usual intuitive way).

(a) (3 pts) Given $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$, show that

$$\nabla \|\mathbf{x} - \mathbf{a}\|_2 = \frac{\partial \|\mathbf{x} - \mathbf{a}\|_2}{\partial \mathbf{x}} = \frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|_2}.$$

(b) (3 pts) Given $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^n$, show that

$$\nabla\left(\mathbf{a}^{\mathsf{T}}\mathbf{X}\mathbf{b}\right) = \frac{\partial \mathbf{a}^{\mathsf{T}}\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^{\mathsf{T}}.$$

(c) (4 pts) Given $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$, show that

$$\nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}\right) = \frac{\partial\left(\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}\right)}{\partial \mathbf{x}} = \left(\mathbf{A} + \mathbf{A}^{\mathsf{T}}\right)\mathbf{x}.$$

Problem 3 (Matrix Calculus) (20 pts)

In this problem, you need to compute the derivative of a scalar function f with respect to a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. For convenience, we often denote this differential simply as

$$df = Df(\mathbf{A})[d\mathbf{A}] = \sum_{i,j} \frac{\partial f}{\partial a_{ij}}\, da_{ij} = \operatorname{tr}\left(\nabla f(\mathbf{A})^{\mathsf{T}} d\mathbf{A}\right).$$

There are several conventions for the matrix gradient (mainly differing by a transpose); we follow the previous convention, consistent with the differential identity and the "summation over i, j" view of df, and define

$$\nabla f(\mathbf{A}) = \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \cdots & \frac{\partial f}{\partial a_{nn}} \end{bmatrix}$$

(a) (10 pts) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. Show that

$$\frac{\partial \det(\mathbf{A})}{\partial \mathbf{A}} = \det(\mathbf{A}) \left(\mathbf{A}^{-1}\right)^{\mathsf{T}}.$$

Hint: Recall from high school that for a 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

where the matrix on the right (before dividing by $\det(\mathbf{A})$) is constructed by taking cofactors and then transposing — this is the adjugate matrix. Similarly, for 3×3 matrices, you may have seen the cofactor expansion formula for the determinant and the adjugate-based formula for the inverse.

In general, for an $n \times n$ matrix $\mathbf{A}$, the cofactor matrix is

$$\mathbf{C} = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{bmatrix}, \qquad C_{ij} = (-1)^{i+j} M_{ij},$$

where M_{ij} is the determinant of the $(n-1) \times (n-1)$ minor (delete row i, column j). The transpose of $\mathbf{C}$ is called the adjugate matrix:

$$\operatorname{adj}(\mathbf{A}) = \mathbf{C}^{\mathsf{T}}.$$

One always has the identity

$$\mathbf{A} \operatorname{adj}(\mathbf{A}) = \det(\mathbf{A})\, \mathbf{I}.$$

Use this identity and the definition of the determinant to derive the derivative formula.

(b) (10 pts) Prove that

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial a_{ij}} = \mathbf{e}_j^{\mathsf{T}} \mathbf{A}^{-1} \mathbf{e}_i,$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a (non-singular) matrix, and $\mathbf{e}_j$ is the unit vector along the j-th axis (e.g. $\mathbf{e}_3 = [0, 0, 1, 0, ..., 0]^{\mathsf{T}}$). It is common to write the formula as

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial \mathbf{A}} = \left(\mathbf{A}^{-1}\right)^{\mathsf{T}}.$$

Hint: You can continue directly from part (a), apply the chain rule to $\log(\det(\mathbf{A}))$.

Alternative hint (prove (b) ⇒ (a)). You may also start from Cramer's rule: for $\mathbf{A}\mathbf{x} = \mathbf{e}_i$,

$$x_j = \frac{\det(\mathbf{A}_j)}{\det(\mathbf{A})},$$

where $\mathbf{A}_j$ is obtained from $\mathbf{A}$ by replacing its j-th column with $\mathbf{e}_i$. By differentiating $\log\det(\mathbf{A})$ along coordinate directions and applying Cramer's rule, you can derive part (b). If you proceed this way and then recover part (a) from (b) via the chain rule, you will still receive full credit.

Problem 4 (Closed-Form Linear Regression Solution) (20 + 15 pts)

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n$. Denote $\mathbf{X}_i \in \mathbb{R}^{1\times d}$ as the i-th row of $\mathbf{X}$, with the following interpretations:

- If the linear model has the bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_m, b]^\mathsf{T}$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}, 1]$, namely $d = m + 1$.
- If the linear model has no bias term, then write $\boldsymbol{\theta} = [w_1, \cdots, w_d]^\mathsf{T}$ and $\mathbf{X}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}]$, namely $d = m$.

(a) (20 pts) Without the bias term, consider the L^2-regularized loss function:

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_\mathrm{i}\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2, \quad \lambda > 0,\ \kappa_i > 0 \text{ for all } i.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta}^* = \left(\boldsymbol{X}^\mathsf{T}\boldsymbol{K}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{K}\boldsymbol{y}$, where

$$\boldsymbol{K} = \begin{bmatrix} \kappa_1 & & 0 \\ & \ddots & \\ 0 & & \kappa_n \end{bmatrix}$$

is a diagonal matrix and $\boldsymbol{I}$ is the $d \times d$ identical matrix.

(b) (Bonus) (15 pts) With the bias term, the L^2-regularized loss function becomes

$$\sum_i \kappa_i \left(y_i - \boldsymbol{X}_\mathrm{i}\boldsymbol{\theta}\right)^2 + \lambda \sum_j w_j^2, \quad \lambda > 0,\ \kappa_i > 0 \text{ for all } i.$$

Show that the optimal solution that minimizes the loss function is $\boldsymbol{\theta}^* = [\boldsymbol{w}^{\star\mathsf{T}}, b^\star]^\mathsf{T}$, where

$$\boldsymbol{w}^\star = \left(\tilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{K}\tilde{\boldsymbol{X}} + \lambda\boldsymbol{I} - \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\tilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{K}\boldsymbol{e}\boldsymbol{e}^\mathsf{T}\boldsymbol{K}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^\mathsf{T}\boldsymbol{K}\left(\boldsymbol{y} - \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\boldsymbol{e}\boldsymbol{e}^\mathsf{T}\boldsymbol{K}\boldsymbol{y}\right),$$

$$b^\star = \frac{1}{\mathrm{Tr}\left(\boldsymbol{K}\right)}\left(\boldsymbol{e}^\mathsf{T}\boldsymbol{K}\boldsymbol{y} - \boldsymbol{e}^\mathsf{T}\boldsymbol{K}\tilde{\boldsymbol{X}}\boldsymbol{w}^\star\right)$$

for which $\boldsymbol{e} = [1\ \ldots\ 1]^\mathsf{T}$ denotes the all one vector, $\boldsymbol{X} = [\tilde{\boldsymbol{X}}\boldsymbol{e}]$, where $\tilde{\mathbf{X}}$ is the dataset of the no bias term. $\mathrm{Tr}\left(\boldsymbol{K}\right)$ is the trace of the matrix $\boldsymbol{K}$, and that $\mathbf{K}$ and $\mathbf{I}$ are defined as in (a).

Problem 5 (Noise and Regularization) (20 pts)

Consider the linear model $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$, where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{\mathbf{w},b}(x) = \mathbf{w}^\mathsf{T}\mathbf{x} + b$$

Given dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, if the inputs $\mathbf{x}_i \in \mathbb{R}^k$ are contaminated with input noise $\boldsymbol{\eta}_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ss}(\mathbf{w}, b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}\left(f_{\mathbf{w},b}(\mathbf{x}_i + \boldsymbol{\eta}_i) - y_i\right)^2\right]$$

where the expectation is taken over the randomness of input noises $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_N$. Additionally, the inputs $(\mathbf{x}_i)$ and the input noise $(\boldsymbol{\eta}_i)$ are independent.

Now assume the input noises $\eta_i = [\eta_{i,1}, \eta_{i,2}, ..., \eta_{i,k}]^\mathsf{T}$ are random vectors with zero mean $\mathbb{E}[\eta_{i,j}] = 0$, and the covariance between components is given by

$$\mathbb{E}[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1 & \text{, if } i = i' \\ 0 & \text{, otherwise.} \end{cases}$ denotes the Kronecker delta.

(a) (20 pts) Please show that

$$\tilde{L}_{ss}(\mathbf{w}, b) = \frac{1}{2N}\sum_{i=1}^{N}\left(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i\right)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L^2-regularization term on the weights.

(Hint: $\|\mathbf{x}\|^\mathbf{2} = \mathbf{x}^\mathsf{T}\mathbf{x} = \mathbf{tr}(\mathbf{x}\mathbf{x}^\mathsf{T})$ and the square of a vector is dot product with itself)