Westlake University
Department of Artificial Intelligence, SOE
Tao Lin & Kaicheng YU
https://github.com/LINs-lab/course_machine_learning

Problem Set Lab 01 (graded), Feb. 18, 2025 (Mathematical Foundation of Machine Learning)

Goals. The goals of this lab are to:

- Familiarize yourself with the mathematical foundations of the machine learning course.

Submission instructions:

- Please submit a PDF file to canvas.
- Deadline: 23.59 on Mar. 02, 2025

Review of Linear Algebra

Problem 1 (Idempotent Matrices and Rank Inequality):

Given A and B are idempotent $n \times n$ matrices (i.e., $A^2 = A$ and $B^2 = B$), and they commute with each other ($AB = BA$):

1. Prove that $A + B - AB$ is also an idempotent matrix.
2. Further prove that

$$\operatorname{rank}(A + B - AB) \leq \operatorname{rank}(A) + \operatorname{rank}(B)\,. \tag{1}$$

Solution 1 (Idempotent Matrices and Rank Inequality):

The rank of an idempotent matrix equals its trace, so $\operatorname{rank}(A) = \operatorname{tr}(A)$ and $\operatorname{rank}(B) = \operatorname{tr}(B)$. Compute $\operatorname{tr}(A + B - AB) = \operatorname{tr}(A) + \operatorname{tr}(B) - \operatorname{tr}(AB)$. Since $AB = BA$ and both A and B can be simultaneously diagonalized (into block matrices of the form $\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$), we have $\operatorname{tr}(AB) \geq 0$. Therefore:

$$\operatorname{rank}(A + B - AB) = \operatorname{tr}(A + B - AB) \leq \operatorname{tr}(A) + \operatorname{tr}(B) = \operatorname{rank}(A) + \operatorname{rank}(B)\,. \tag{2}$$

Problem 2 (Diagonalizability and Eigenvalues of a Linear Transformation):

Let V be a four-dimensional vector space, and let $T : V \to V$ be a linear transformation satisfying $T^3 - 2T^2 + T - 2I = 0$, where I is the identity transformation on V. Prove that T is diagonalizable and determine all its eigenvalues.

Solution 2 (Diagonalizability and Eigenvalues of a Linear Transformation):

Substitute T with λ:

$$\lambda^3 - 2\lambda^2 + \lambda - 2 = 0\,. \tag{3}$$

Using the Rational Root Theorem, possible roots are $\pm 1, \pm 2$. When $\lambda = 2$:

$$2^3 - 2(2)^2 + 2 - 2 = 8 - 8 + 2 - 2 = 0\,. \tag{4}$$

Thus, $\lambda = 2$ is a root. Factor the polynomial:

$$(\lambda - 2)(\lambda^2 + 1) = 0\,. \tag{5}$$

Remaining roots:

$$\lambda = i \quad \text{and} \quad \lambda = -i\,. \tag{6}$$

The minimal polynomial splits into distinct linear factors over the complex field:

$$(\lambda - 2)(\lambda - i)(\lambda + i)\,. \tag{7}$$

Since all eigenvalues are distinct and V is four-dimensional, T has four linearly independent eigenvectors (taking multiplicity into account).

Problem 3 (Existence of Real Matrix Roots for Positive Eigenvalue Matrices):

If a real matrix A has all eigenvalues as positive real numbers, then for any positive integer m, there exists a real matrix B such that $B^m = A$.

Solution 3 (Existence of Real Matrix Roots for Positive Eigenvalue Matrices):

Define $B = \exp\left(\frac{1}{m}\log(A)\right)$, we have:

$$B^m = \left[\exp\left(\frac{1}{m}\log(A)\right)\right]^m = \exp\left(\log(A)\right) = A\,. \tag{8}$$

Problem 4 (Eigenvalue Equivalence under Commutator-like Condition):

Let A and B be $n \times n$ square matrices satisfying

$$AB - BA = A - B\,. \tag{9}$$

Then, A and B have the same eigenvalues.

Solution 4 (Eigenvalue Equivalence under Commutator-like Condition):

Assume λ is an eigenvalue of A with corresponding eigenvector v, i.e.,

$$Av = \lambda v\,. \tag{10}$$

Applying the given relation to v, we have

$$ABv - BAv = Av - Bv\,. \tag{11}$$

Calculate each term:

$$ABv = A(Bv)\,. \tag{12}$$

$$BAv = B(Av) = B(\lambda v) = \lambda Bv\,. \tag{13}$$

Substituting these into the equation yields

$$A(Bv) - \lambda Bv = \lambda v - Bv\,. \tag{14}$$

Rearranging terms, we obtain

$$A(Bv) = \lambda v + (\lambda - 1)Bv\,. \tag{15}$$

Assume $Bv = \mu v$, where μ is an eigenvalue of B. Then,

$$A(Bv) = A(\mu v) = \mu Av = \mu\lambda v\,. \tag{16}$$

Substituting back, we get

$$\mu\lambda v = \lambda v + (\lambda - 1)\mu v\,. \tag{17}$$

Since v is non-zero, we can divide both sides by v:

$$\mu\lambda = \lambda + \mu\lambda - \mu\,. \tag{18}$$

Simplifying, we find

$$0 = \lambda - \mu \quad \Rightarrow \quad \mu = \lambda \,. \tag{19}$$

Therefore, λ is also an eigenvalue of B. Since λ was an arbitrary eigenvalue of A, it follows that all eigenvalues of A are eigenvalues of B. By symmetry of the argument, all eigenvalues of B are also eigenvalues of A. Thus, A and B have the same eigenvalues.

Problem 5 (Matrix Determinant and Commutator):

Let A and B be two $n \times n$ matrices satisfying the equation

$$AB - BA = A \,. \tag{20}$$

Prove that $\det(A) = 0$.

Solution 5 (Matrix Determinant and Commutator):

Assume A is invertible, i.e., $\det(A) \neq 0$. Then, A^{-1} exists. Multiply both sides of the equation $AB - BA = A$ on the left by A^{-1}:

$$A^{-1}AB - A^{-1}BA = A^{-1}A \,. \tag{21}$$

Simplifying, we get:

$$B - A^{-1}BA = I \,, \tag{22}$$

where I is the identity matrix. Take the trace of both sides:

$$\mathrm{tr}(B - A^{-1}BA) = \mathrm{tr}(I) \,, \tag{23}$$

Using the cyclic property of trace ($\mathrm{tr}(A^{-1}BA) = \mathrm{tr}(BAA^{-1}) = \mathrm{tr}(B)$), the left side simplifies to:

$$\mathrm{tr}(B) - \mathrm{tr}(B) = 0 \,. \tag{24}$$

The right side is:

$$\mathrm{tr}(I) = n \,. \tag{25}$$

This leads to the contradiction:

$$0 = n \,. \tag{26}$$

Since n is a positive integer, the assumption that A is invertible is false. Therefore, A is singular, which means:

$$\det(A) = 0 \,. \tag{27}$$

Review of Probability Theory

Problem 6 (Moment Bound for a Standard Normal Random Variable):

For a standard normal random variable X, there exists a constant C such that for all $p > 1$,

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq C\sqrt{p} \,. \tag{28}$$

Solution 6 (Moment Bound for a Standard Normal Random Variable):

For a standard normal random variable X, the p-th moment is given by:

$$\mathbb{E}\left[|X|^p\right] = 2^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \,, \tag{29}$$

where Γ denotes the Gamma function. Stirling's approximation provides an asymptotic expression for the Gamma function for large arguments:

$$\Gamma(z) \sim \sqrt{2\pi}\, z^{z-1/2} e^{-z}, \quad \text{as } z \to \infty \,. \tag{30}$$

Let us set $z = \frac{p+1}{2}$. Then, for large p ,

$$\Gamma\left(\frac{p+1}{2}\right) \sim \sqrt{2\pi}\left(\frac{p+1}{2}\right)^{\frac{p}{2}} e^{-\frac{p+1}{2}}. \tag{31}$$

For simplicity, and since p is large, we can approximate $p+1 \approx p$, yielding:

$$\Gamma\left(\frac{p+1}{2}\right) \sim \sqrt{2\pi}\left(\frac{p}{2}\right)^{\frac{p}{2}} e^{-\frac{p}{2}}. \tag{32}$$

Substituting the Stirling approximation into the expression for $\mathbb{E}\left[|X|^p\right]$:

$$\mathbb{E}\left[|X|^p\right] \sim 2^{p/2} \cdot \frac{\sqrt{2\pi}\left(\frac{p}{2}\right)^{\frac{p}{2}} e^{-\frac{p}{2}}}{\sqrt{\pi}} = 2^{p/2} \cdot \sqrt{2}\left(\frac{p}{2}\right)^{\frac{p}{2}} e^{-\frac{p}{2}}. \tag{33}$$

Simplifying further:

$$\mathbb{E}\left[|X|^p\right] \sim 2^{p/2} \cdot \sqrt{2} \cdot \left(\frac{p}{2}\right)^{\frac{p}{2}} e^{-\frac{p}{2}} = \sqrt{2}\left(\frac{p}{e}\right)^{\frac{p}{2}}. \tag{34}$$

Taking the p-th root of both sides:

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \sim \left(\sqrt{2}\left(\frac{p}{e}\right)^{\frac{p}{2}}\right)^{1/p} = 2^{1/(2p)}\left(\frac{p}{e}\right)^{1/2}. \tag{35}$$

As $p \to \infty$, $2^{1/(2p)} \to 1$, so for sufficiently large p,

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \lesssim \sqrt{\frac{p}{e}}. \tag{36}$$

Thus, there exists a constant $C \geq \frac{1}{\sqrt{e}}$ such that

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq C\sqrt{p}. \tag{37}$$

The above asymptotic analysis holds for large p. For smaller values of p, the moments $\mathbb{E}\left[|X|^p\right]^{1/p}$ can be explicitly computed or bounded, and they are finite.

Problem 7 (Bounded Random Variable and Exponential Expectation):

Let X be a bounded random variable with $\mathbb{E}[X] = 0$ and $|X|_\infty \leq a$ for some $a > 0$. Prove that

$$\mathbb{E}\left[e^X\right] \leq \cosh(a). \tag{38}$$

Solution 7 (Bounded Random Variable and Exponential Expectation):

We have:

$$-a \leq X \leq a \quad \text{and} \quad \mathbb{E}[X] = 0. \tag{39}$$

Assume X takes values a and $-a$ with probabilities p and $1-p$, respectively. Then:

$$p \cdot a + (1-p)(-a) = 0 \Longrightarrow 2p - 1 = 0 \Longrightarrow p = \frac{1}{2}. \tag{40}$$

We compute the expectation of exponential:

$$\mathbb{E}\left[e^X\right] = \frac{1}{2}e^a + \frac{1}{2}e^{-a} = \cosh(a). \tag{41}$$

Since e^x is convex, by Jensen's Inequality,

$$\mathbb{E}\left[e^X\right] \leq \cosh(a). \tag{42}$$

The maximum is achieved when X is concentrated at $\pm a$ with equal probability.

Problem 8 (Almost Sure Convergence of Scaled Random Walks):

Let X be a random variable with distribution $P(X = 1) = P(X = -1) = \frac{1}{2}$. Define the partial sum $S_n = X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) copies of X. For any $\alpha > \frac{1}{2}$, prove that

$$P\left(\lim_{n\to\infty} \frac{S_n}{n^\alpha} = 0\right) = 1\,. \tag{43}$$

Solution 8 (Almost Sure Convergence of Scaled Random Walks):

Since $\mathbb{E}[X_i] = 0$ and $\mathsf{Var}(X_i) = 1$, the variance of the partial sum is $\mathsf{Var}(S_n) = n$. For any $\epsilon > 0$,

$$P\left(|S_n| > n^\alpha\right) \leq \frac{\mathsf{Var}(S_n)}{n^{2\alpha}} = \frac{n}{n^{2\alpha}} = n^{1-2\alpha}\,. \tag{44}$$

Since $\alpha > \frac{1}{2}$, the exponent $1 - 2\alpha < 0$, implying $P\left(|S_n| > n^\alpha\right)$ decays polynomially. Consider the series

$$\sum_{n=1}^{\infty} P\left(|S_n| > n^\alpha\right) \leq \sum_{n=1}^{\infty} n^{1-2\alpha}\,. \tag{45}$$

The series converges because $1 - 2\alpha < -1$ when $\alpha > \frac{1}{2}$. By the First Borel-Cantelli Lemma, since the sum is finite, the probability that infinitely many events $\{|S_n| > n^\alpha\}$ occur is zero. With probability 1, only finitely many n satisfy $|S_n| > n^\alpha$. Therefore,

$$\lim_{n\to\infty} \frac{S_n}{n^\alpha} = 0 \quad \text{almost surely}\,. \tag{46}$$

Problem 9 (Probability Bound for the Standardized Sum of Uniform Random Variables):

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables uniformly distributed on the interval $(-1, 1)$. For any $r > 0$, prove that

$$P\left(\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} < r\right) > 1 - \frac{1}{3r^2}\,. \tag{47}$$

Solution 9 (Probability Bound for the Standardized Sum of Uniform Random Variables):

Since each X_i is uniformly distributed on $(-1, 1)$,

$$\mathbb{E}[X_i] = 0\,. \tag{48}$$

The variance of X_i is

$$\mathsf{Var}(X_i) = \frac{(b-a)^2}{12} = \frac{(1-(-1))^2}{12} = \frac{1}{3}\,. \tag{49}$$

For the sum $S_n = X_1 + X_2 + \cdots + X_n$,

$$\mathbb{E}[S_n] = 0 \quad \text{and} \quad \mathsf{Var}(S_n) = n \cdot \mathsf{Var}(X_i) = \frac{n}{3}\,. \tag{50}$$

The standardized sum is

$$\frac{S_n}{\sqrt{n}}\,. \tag{51}$$

Its variance is

$$\mathsf{Var}\left(\frac{S_n}{\sqrt{n}}\right) = \frac{\mathsf{Var}(S_n)}{n} = \frac{1}{3}\,. \tag{52}$$

Chebyshev's inequality states that for any $k > 0$,

$$P\left(\left|\frac{S_n}{\sqrt{n}}\right| \geq k\right) \leq \frac{\mathsf{Var}\left(\frac{S_n}{\sqrt{n}}\right)}{k^2} = \frac{1}{3k^2}\,. \tag{53}$$

Setting $k = r$, we have

$$P\left(\left|\frac{S_n}{\sqrt{n}}\right| \geq r\right) \leq \frac{1}{3r^2}\,. \tag{54}$$

Therefore,

$$P\left(\frac{S_n}{\sqrt{n}} < r\right) \geq 1 - P\left(\left|\frac{S_n}{\sqrt{n}}\right| \geq r\right) \geq 1 - \frac{1}{3r^2}\,. \tag{55}$$

Problem 10 (Central Limit Theorem and Standardized Sum Convergence):

Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i) > 0$. Define the standardized sum:

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\,. \tag{56}$$

Then, as $n \to \infty$, the distribution of Z_n converges to the standard normal distribution $\mathcal{N}(0, 1)$. Formally, for all real numbers z:

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)\,, \tag{57}$$

where $\Phi(z)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Solution 10 (Central Limit Theorem and Standardized Sum Convergence):

Define the standardized sum:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \quad \text{where } S_n = \sum_{i=1}^{n} X_i\,. \tag{58}$$

The characteristic function of Z_n is:

$$\phi_{Z_n}(t) = \mathbb{E}\left[e^{itZ_n}\right] = \mathbb{E}\left[e^{it\frac{S_n - n\mu}{\sigma\sqrt{n}}}\right]\,. \tag{59}$$

Since X_i are i.i.d., the characteristic function of S_n is:

$$\phi_{S_n}(t) = \left(\phi_X(t)\right)^n\,, \tag{60}$$

where $\phi_X(t)$ is the characteristic function of a single X_i. Expand $\phi_X(t)$ around $t = 0$ using Taylor's theorem:

$$\phi_X(t) = 1 + it\mu - \frac{t^2\sigma^2}{2} + o(t^2)\,. \tag{61}$$

Therefore,

$$\phi_{Z_n}(t) = \left[1 + i\left(\frac{t}{\sigma\sqrt{n}}\right)\mu - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n\,. \tag{62}$$

Using the limit $\lim_{n\to\infty}\left(1 + \frac{a}{n} + \frac{b}{n} + o\left(\frac{1}{n}\right)\right)^n = e^{a+b}$, we have:

$$\lim_{n\to\infty} \phi_{Z_n}(t) = \lim_{n\to\infty}\left[1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{1}{n}\right)\right]^n = e^{-\frac{t^2}{2}}\,. \tag{63}$$

This is the characteristic function of the standard normal distribution $\mathcal{N}(0, 1)$. By Levy's Continuity Theorem, if the characteristic functions $\phi_{Z_n}(t)$ converge pointwise to $e^{-\frac{t^2}{2}}$, then Z_n converges in distribution to $\mathcal{N}(0, 1)$.