DA 5001/6400 (July-Nov 2024): HW0

IIT Madras

Due Date: August 9, 2024 at 11:59 PM

Instructions

- The maximum score on the homework is 50 marks, including extra credit exercises.
- List the names of students you collaborated with for the homework. Also, provide citations to any books, notes, or web resources you used for any problem. Failure to do so will be considered as a violation of the honour code.
- If you used an LLM for help in one of the exercises, you must specify the name of the LLM and the exact prompt used.

Honour Code

Please read the full honour code on the course webpage and write "I ACCEPT THE HONOUR CODE" in your submission. Your submission will not be graded if you do not accept the honour code.

0 Review of Linear Algebra, Calculus, Probability

Ungraded Exercise: While this exercise is ungraded, we strongly recommend that you solve the following problems from the Mathematics for Machine Learning (MML) book[1] to review your mathematical fundamentals:

1. Problem 3.1 (page 96)
2. Problem 4.11 (page 138)
3. Problems 5.2 and 5.3 (page 170)
4. Problem 5.8, parts a and b (page 171)
5. Problem 6.12, all parts (page 223-224)
6. (Optional) Problem 6.13 (page 224).

[1] Available at `https://mml-book.github.io/book/mml-book.pdf`

1 Gaussian Log-Likelihood Ratio (5 marks)

Let p denote the probability density function (PDF) of the one-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and q denote the PDF of $\mathcal{N}(\mu, \sigma^2)$ the same variance σ^2. Let the random variable X be distributed according to p. Find the PDF of the random variable $\log(p(X)/q(X))$. What is its distribution?

2 Markov and Chebyshev inequalities (5 marks)

1. For a non-negative random variable X, prove that

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}\,.$$

   You may assume that the variable is continuous for this proof (although Markov's inequality holds more generally).

2. For a random variable X with $\text{Var}(X) < \infty$ (i.e., it has finite variance), prove that

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq \frac{\text{Var}(X)}{t^2}\,.$$

Hint: For Markov's inequality, note that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \geq \int_{t}^{\infty} x\, f_X(x)\, dx$$

for any $t > 0$, where $f_X(x)$ is the PDF of X. Then, argue that the right side is at least $t \cdot \mathbb{P}(X > t)$ to prove Markov's inequality. Chebyshev's inequality follows from Markov's inequality – how?

3 Moment Generating Functions (5 marks)

Recall that the moment generating function (MGF) of a random variable X is defined as $\phi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ whenever the expectation exists and is finite. Find the MGF of:

1. The Gaussian distribution $\mathcal{N}(0, \sigma^2)$.
2. The chi-squared distribution with 1 degree of freedom. That is, if $X \sim \mathcal{N}(0, 1)$, find $\mathbb{E}[e^{\lambda X^2}]$.

Context: Note that Markov's inequality required only the first moment and the right side scales as $1/t$, while Chebyshev's inequality, which required the 2nd moment), scales as $1/t^2$, which is smaller than $1/t$. If the MGF exists, then we get for any $\lambda > 0$, we have

$$\mathbb{P}(X > t) = \mathbb{P}(e^{\lambda X} > e^{\lambda t}) \leq e^{-\lambda t}\ \phi_X(\lambda)\,.$$

The right side here scales as $e^{-\lambda t}$ which is asymptotically much smaller than even $1/t^2$, meaning that we get a tighter bound. If the MGF exists at larger λ, this also gives a tighter bound. Of the examples above, which do you think gives the tightest right-hand side?

4 Convexity of the KL Divergence (2 + 4 + 4 = 10 marks)

Let $p = (p_1, \ldots, p_k)$ and $q = (q_1, \ldots, q_k)$ be two discrete distributions over k items. Recall that the Kullback-Leibler (KL) divergence between discrete distributions p and q is defined as

$$D_{\mathrm{KL}}(p\|q) = \sum_{i=1}^{k} p_i \log(p_i/q_i)\,.$$

Intuitively, it is large if p and q are very "different" from each other. In this exercise, we will prove that it is jointly convex in (p, q).

Convexity Recall that a function f is convex if

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

for all x_1, x_2 in the domain of f and all $\lambda \in [0, 1]$. See the figure below: convexity means the blue curve is below the orange line between x_1 and x_2.

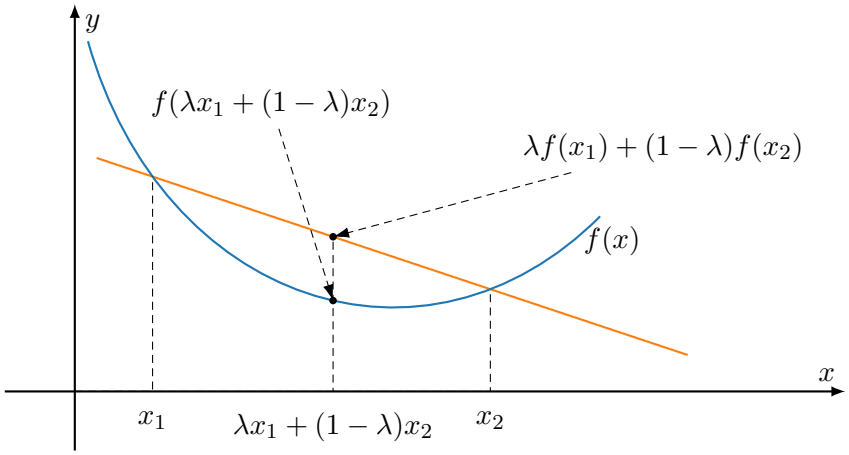

Figure credits: Stack Exchange

1. Show that the negative entropy function $h : (0, 1) \to \mathbb{R}$ given by $h(x) = x \log(x)$ is convex. All logarithms are natural logarithms.

   Hint: You may use the following equivalent condition of convexity: a twice differentiable function h is convex if and only if its second derivative $h''(x) \ge 0$ is non-negative for all x in the domain of h.

2. Use the convexity of h to establish the log-sum inequality: for any $a_1, a_2, b_1, b_2 > 0$, we have

$$(a_1 + a_2) \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \le a_1 \log\left(\frac{a_1}{b_1}\right) + a_2 \log\left(\frac{a_2}{b_2}\right)\,.$$

   Hint: We can rewrite the right side as

$$b_1 \cdot h\left(\frac{a_1}{b_1}\right) + b_2 \cdot h\left(\frac{a_2}{b_2}\right) = (b_1 + b_2)\left(\lambda \cdot h\left(\frac{a_1}{b_1}\right) + (1-\lambda)\cdot h\left(\frac{a_2}{b_2}\right)\right)\,,$$

   where we take $\lambda = b_1/(b_1 + b_2) \in (0, 1)$. Now apply the convexity of h.

3. Prove that the function $(p, q) \mapsto D_{\mathrm{KL}}(p\|q)$ is convex. That is, for distributions p, p', q, q' and $\lambda \in [0, 1]$, show that

$$D_{\mathrm{KL}}(\lambda p + (1-\lambda)p' \| \lambda q + (1-\lambda)q') \leq \lambda\, D_{\mathrm{KL}}(p\|q) + (1-\lambda) D_{\mathrm{KL}}(p'\|q')\,.$$

Hint: Use the log-sum inequality per component.

5 KL divergence between multivariate Gaussians (5 marks)

Let $P = \mathcal{N}(\mu_1, \Sigma)$ and $Q = \mathcal{N}(\mu_2, \Sigma)$ be two multivariate Gaussian distributions with means $\mu_1, \mu_2 \in \mathbb{R}^d$ and equal covariance $\Sigma \in \mathbb{S}^d_{++}$.[2] Calculate the KL divergence $D_{\mathrm{KL}}(P\|Q)$.

6 Gradient Descent on Quadratic Functions ($5 \times 2 = 10$ marks)

Consider the quadratic function $f : \mathbb{R}^d \to \mathbb{R}$ given by

$$f(\theta) = \frac{1}{2}\theta^\top A\theta + b^\top \theta\,,$$

for some positive definite matrix $A \in \mathbb{S}^d_{++}$ and vector $b \in \mathbb{R}^d$. Starting from some fixed $\theta_0 \in \mathbb{R}^d$, gradient descent with a learning rate $\gamma > 0$ produces the sequence $(\theta_t)_{t=0}^\infty$ given by

$$\theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t)\,,$$

where ∇f is the gradient of our quadratic function f with respect to its argument θ.

1. Find $\nabla f(\theta)$.
2. Show that f is a strictly convex function. Thus, its global minimizer $\theta_\star$ is unique and is the solution of the equation $\nabla f(\theta_\star) = 0$. Then, find $\theta_\star$
3. Show that the sequence $(\theta_0)_{t=0}^\infty$ produced by gradient descent satisfies

$$\theta_t - \theta_\star = (I - \gamma A)(\theta_{t-1} - \theta_\star)\,.$$

   where I is the identity matrix in d dimensions. Thus, conclude that $\theta_t - \theta_\star = (I - \gamma A)^t(\theta_0 - \theta_\star)$.
4. Let $\lambda_1 = \lambda_{\max}(A)$ denote the largest eigenvalue of A. If the learning rate satisfies $\gamma \leq 1/\lambda_1$, show that

$$\lim_{t\to\infty} \|\theta_t - \theta_\star\|_2 = 0\,.$$

5. Conversely, if the learning rate $\gamma > 2/\lambda_1$ is large, show that

$$\lim_{t\to\infty} \|\theta_t - \theta_\star\|_2 = \infty\,.$$

[2] Note that $\mathbb{S}^d_+$ denotes the set of positive semi-definite matrices. That is, $A \in \mathbb{S}^d_+$ if A is symmetric and $x^\top Ax \geq 0$ for all vectors $x \in \mathbb{R}^d$. If the above inequality is strict, then we say that A is positive definite and denote it as $A \in \mathbb{S}^d_{++}$.

Context: With gradient descent or stochastic gradient descent, we usually want to use the largest learning rate under which we still have convergence; this is called the **divergent learning rate**. The above argument shows that $1/\lambda_1$ is the divergent learning rate for this problem.

Hint for the last two parts: The eigenvectors of A form a basis of $\mathbb{R}^d$ – it is convenient to analyze the iterations in that basis. Concretely, let $A = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$ be the eigenvalue decomposition of A with eigenvalues $\lambda_1 > \cdots > \lambda_d > 0$.

To represent our iterates in this eigen basis, define $\delta_{t,i} = u_i^\top(\theta_t - \theta_\star)$; this is the component of $\theta_t - \theta_\star$ along the eigenvector δ_t. Next, prove that $\delta_{t,i} = (1 - \gamma\lambda_i)^t \delta_{0,i}$. In other words, gradient descent on the objective is equivalent to running one-dimensional gradient descent along each eigenvector.

Now complete the proof from here: Part 4 requires that $\delta_{t,i} \to 0$ *for all* i. while part 5 requires showing that $\delta_{t,i} \to \infty$ *for some* i.

7 Membership Inference: Toy Problem ($5 \times 2 = 10$ marks)

Membership inference is a privacy attack in which an adversary tries to guess if a data point appeared in the training dataset of a model. We will perform membership inference on a toy problem, where our function of the data is simply a sum (rather than a model trained on this data). This toy problem is also a useful building block for auditing rigorous privacy guarantees — we will discuss this later in the course.

Membership inference game Consider the following game:

- The learner has access to two datasets $D_0 = \{x_1, \ldots, x_n\}$ and $D_1 = D_0 \cup \{z\}$ of d-dimensional vectors. These two datasets differ in only one item z.
- The learner picks dataset D from either D_0 or D_1 uniformly at random and reveals its sum $f(D) = \sum_{y \in D} y$.
- The privacy adversary observers the value $f(D)$ and the datapoint z.
- The adversary must guess whether $D = D_0$ or $D = D_1$. In other words, the adversary must guess if $z \in D$ or $z \notin D$.

The adversary's task can be cast as a statistical hypothesis test

$$H_0 : D = D_0$$
$$H_1 : D = D_1 \,.$$

The purpose of this exercise is to bound the type-I error (i.e. the false positive rate, or FPR) and type-II error (i.e. one minus the true positive rate or $1 - \text{TPR}$) of the adversary as a function of the problem parameters.

We consider the case when the data point z is random; in particular, $z \sim \mathcal{N}(0, \sigma^2 I_d)$ is component-wise i.i.d. Gaussian with variance σ^2 (and I_d is the $d \times d$ identity matrix). This situation shows up in testing the privacy property of models where we insert special datapoints known as "canaries" in order to test the properties of the model.

Questions Now, suppose that $x_1, \ldots, x_n \sim \mathcal{N}(0, I_d)$ are component-wise i.i.d. standard Gaussian. Express the answers to following questions in terms of the dimension d and variance σ^2 of z.

1. Calculate $\mathbb{E}[z^\top f(D_0)]$ and $\mathbb{E}[z^\top f(D_1)]$.
2. Calculate $\mathbb{E}\big[(z^\top f(D_0))^2\big]$ and $\mathbb{E}\big[(z^\top f(D_1))^2\big]$.

   Context: The above inequalities suggest that we should reject the null hypothesis H_1 if $z^\top f(D)$ is larger than some threshold t. Next, we will bound the type-I and type-II errors.
3. Next, we wish to bound the error of the adversary. Assume that $z^\top f(D_0)$ and $z^\top f(D_1)$ are normally distributed with their respective means and variances as computed above. Plot the trade-off between the type-I error $\mathbb{P}(z^\top f(D_0) > t)$ and the type-II error $\mathbb{P}(z^\top f(D_1) < t)$ while varying the threshold t.[3] You may use $d = 10^4$ and $\sigma = 1$.
4. Plot the minimum total error $\min_t\{\mathbb{P}(z^\top f(D_0) > t) + \mathbb{P}(z^\top f(D_1) < t)\}$ versus the dimension d while σ^2 is fixed. You may use $d \in \{10^1, 10^2, \ldots, 10^7\}$ for the x-axis. Repeat this for $\sigma^2 \in \{0.1, 0.25, 1, 4, 10\}$.
5. What are the type-I and type-II errors in the limit $d \to \infty$ of high dimension? Although it is not necessary, you may use the tail inequality $\mathbb{P}_{Z\sim\mathcal{N}(0,1)}(Z > t) \le e^{-t^2/2}$ for any $t > 0$.

Extra Credit 1: Maximum of the Uniform Distribution (1+2+2 = 5 marks)

Let $X_1, \ldots, X_n \sim \text{Unif}(0, a)$ be n i.i.d. samples distributed between 0 and $a > 0$. Let $Y = \max\{X_1, \ldots, X_n\}$ denote the largest sample.

1. Compute the mean of Y.
2. Compute the variance of Y.
3. Suppose that the upper limit a is unknown and we use Y to estimate a. Evaluate its squared error $\mathbb{E}(Y - a)^2$.

Context: A related approach was used in World War 2 to estimate the enemy capability. If interested, read about the German Tank Problem on Wikipedia.

[3] That is, plot the parametric curve $(x(t), y(t))$ where $x(t) = \mathbb{P}(z^\top f(D_0) > t)$ and $y(t) = \mathbb{P}(z^\top f(D_1) < t)$ while varying $t \in \mathbb{R}$.