# CS/ECE/STAT-861: Theoretical Foundations of Machine Learning

**University of Wisconsin–Madison, Fall 2025** *Instructor: Kirthevasan Kandasamy*

Homework 0. Due 09/12/2025, 11.59 pm

---

Homework 0 serves as a diagnostic tool to assess your background knowledge and mathematical maturity for this course. It also familiarizes you with course expectations as you advance to more complex topics. While you are not expected to know all the solutions right away, you should be able to solve most of the questions with reasonable effort, using references as needed. Future homework assignments will be more challenging than Homework 0.

**Instructions:**

1. Homework is due on Canvas by 11:59 pm on the due date. Please plan to submit well before the deadline. Refer to the course website for policies on late submission.
2. Homework must be typeset using appropriate software, such as LaTeX. Handwritten and scanned submissions will **not** be accepted.
3. Your solutions will be evaluated on correctness, clarity, and conciseness.
4. Unless otherwise specified, you may use any result we have already proved in class. Clearly state which result you are using.
5. Solutions to some of the problems may be found in the recommended textbook or other resources. Unless stated otherwise, you should attempt the problems on your own. You may **not** search the internet for solutions or use LLM-based tools.
6. If you use any external references, please cite them in your submission.
7. **Collaboration:** You may collaborate in groups of size up to 3 to solve problems indicated by a star ($\star$). You may not collaborate on other problems. If you collaborate, please indicate your collaborators at the beginning of the problem. Even if you collaborate, *you must write the solution in your own words*.

# 1 Estimating the mean of a normal distribution

We are given $n$ independent samples $S = \{X_1, \ldots, X_n\}$, where each $X_i$ is sampled from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *unknown* mean $\mu$, but known variance $\sigma^2$. We wish to estimate the mean $\mu$. An estimator $\widehat{\mu} : \mathbb{R}^n \to \mathbb{R}$ is a function of the data collected. We will quantify the performance of an estimator via the risk $R$, which is simply the expected squared error of the estimator,

$$R(\widehat{\mu}, \mu) = \mathbb{E}[(\widehat{\mu}(S) - \mu)^2].$$

Here, the expectation is with respect to the randomness in the data.

1. **[2 pts]** *(Bias–variance decomposition)* First, show that the following holds for any estimator $\widehat{\mu}$,

$$R(\widehat{\mu}, \mu) = \underbrace{(\mathbb{E}[\widehat{\mu}(S)] - \mu)^2}_{\text{bias}} + \underbrace{\mathbb{E}[(\widehat{\mu}(S) - \mathbb{E}[\widehat{\mu}(S)])^2]}_{\text{variance}}.$$

2. **[2 pts]** *(Sample mean)* The most natural estimator for $\mu$ is the *sample mean* $\widehat{\mu}_{\mathrm{SM}}$, defined below.

$$\widehat{\mu}_{\mathrm{SM}}(S) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

   Using the result of part 1, show that $R(\widehat{\mu}_{\mathrm{SM}}, \mu) = \sigma^2/n$.

3. **[3 pts]** *(Concentration)* While the risk measures how well an estimator does in expectation, sometimes we also wish to know that $\widehat{\mu}_{\mathrm{SM}}$ is within some margin of error $\epsilon$ of the true mean $\mu$ with high probability. Prove the following result for any $\epsilon > 0$:

$$\mathbb{P}(|\widehat{\mu}_{\mathrm{SM}}(S) - \mu| > \epsilon) \leq 2 \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

   where the probability $\mathbb{P}$ is with respect to the randomness in the data.

   You may use the following facts about normal random variables:

   - If $X_1, \ldots, X_n$ are normal, then so is $\sum_{i=1}^n X_i$. (You will need to compute the mean and variance.)
   - If $X$ are normal, then so is $aX$ for any $a \in \mathbb{R}$. (You will need to compute the mean and variance.)
   - If $Z \sim \mathcal{N}(0, 1)$ is a standard normal random variable, then $\mathbb{P}(Z > \epsilon) \leq e^{-\epsilon^2/2}$, and $\mathbb{P}(Z < -\epsilon) \leq e^{-\epsilon^2/2}$.

4. **[2 pts]** *(A hypothesis test)* Suppose it was known that $\mu \in \{0, \Delta\}$ for some $\Delta > 0$. A hypothesis test $\psi$ chooses either 0 or $\Delta$ based on a dataset $S$. Consider the following hypothesis test:

$$\text{Choose } \psi(S) = 0 \text{ if } \widehat{\mu}_{\mathrm{SM}}(S) \leq \frac{\Delta}{2}, \qquad \text{else choose } \psi(S) = \Delta.$$

   Let $\mathbb{P}_\mu$ denote probabilities when the true mean is $\mu$. Show that if $n \geq \frac{8\sigma^2}{\Delta^2} \log(1/\delta)$, we have

$$\mathbb{P}_0(\psi(S) = 0) \geq 1 - \delta, \qquad \mathbb{P}_\Delta(\psi(S) = \Delta) \geq 1 - \delta.$$

   **N.B.** This shows that $\mathcal{O}\left(\sigma^2 \Delta^{-2} \log(1/\delta)\right)$ samples are sufficient to separate two normal distributions that are $\Delta$ apart. In class, we will show that at least $\Omega\left(\sigma^2 \Delta^{-2} \log(1/\delta)\right)$ samples are also necessary.

5. **[2 pts]** *(An alternative estimator)* The sample mean is just one of several possible estimators for $\mu$. Student A proposes the following alternative estimator $\widehat{\mu}_\alpha$ with some parameter $\alpha \in (0, 1)$,

$$\widehat{\mu}_\alpha(S) = \frac{\alpha}{n} \sum_{i=1}^{n} X_i.$$

   Using the result from part 1, compute the risk of the estimator $\widehat{\mu}_\alpha$. Note that, unlike the sample mean, the risk of $\widehat{\mu}_\alpha$ depends on the true mean $\mu$.

6. **[2 pts]** Show that there exists at least one value for $\mu$ such that $\widehat{\mu}_\alpha$ is a strictly better estimator than $\widehat{\mu}_{\mathrm{SM}}$. That is, there exists $\mu \in \mathbb{R}$, such that, for all $\alpha \in (0,1)$, we have $R(\widehat{\mu}_\alpha, \mu) < R(\widehat{\mu}_{\mathrm{SM}}, \mu)$.

7. **[4 pts]** *(Maximum risk)* Despite the result from part 6, student B is not satisfied with student A's proposition, as an estimator should perform well for all values of $\mu$, and not just for one value of $\mu$. In particular, she argues that the worst-case risk over all $\mu$ should be small. She proposes the following criterion, the *maximum risk* $\mathcal{R}$, as a way to measure how well an estimator performs.

$$\mathcal{R}(\widehat{\mu}) = \sup_{\mu \in \mathbb{R}} R(\widehat{\mu}, \mu) = \sup_{\mu \in \mathbb{R}} \mathbb{E}[(\widehat{\mu} - \mu)^2].$$

   (a) Compute $\mathcal{R}(\widehat{\mu}_{\mathrm{SM}})$ and $\mathcal{R}(\widehat{\mu}_\alpha)$.

   (b) Based on the above answers, which estimator would you choose?

8. **[5 pts]** *(Maximum risk over a bounded domain)* Suppose we had prior knowledge that $\mu \in [0,1]$. While student A agrees with student B's criterion, she argues that we should modify the definition of the maximum risk to incorporate this prior knowledge. She proposes the following definition instead:

$$\mathcal{R}'(\widehat{\mu}) = \sup_{\mu \in [0,1]} R(\widehat{\mu}, \mu) = \sup_{\mu \in [0,1]} \mathbb{E}[(\widehat{\mu} - \mu)^2].$$

   (a) Compute $\mathcal{R}(\widehat{\mu}_{\mathrm{SM}})$ and $\mathcal{R}(\widehat{\mu}_\alpha)$, the maximum risk for the two estimators discussed above.

   (b) Is there any particular value of $\alpha$ (possibly dependent on $n$ and $\sigma$) for which $\mathcal{R}'(\widehat{\mu}_\alpha) < \mathcal{R}'(\widehat{\mu}_{\mathrm{SM}})$?

   (c) Based on the above answer, which estimator would you choose? Intuitively, explain the discrepancy in the conclusions in part 7 and part 8.

**N.B.** In class, we will study minimax optimality, where our goal is to design algorithms with the smallest maximum risk across a class of distributions. We will begin with simple mean estimation problems, and then apply these techniques to regression, classification, density estimation, online learning, and bandits throughout the course.

## 2 Understanding exploration–exploitation trade-offs ⋆

Consider the following game which proceeds over $T$ rounds. You have access to two normal distributions $\nu^{(1)} = \mathcal{N}(\mu_1, \sigma^2)$ and $\nu^{(2)} = \mathcal{N}(\mu_2, \sigma^2)$, where $\sigma^2$ is known but $\mu_1, \mu_2 \in [0,1]$ are not. On each round $t$, you have the opportunity to earn a monetary reward by drawing a sample from either distribution. You choose $I_t \in \{1,2\}$, where $I_t = i$ corresponds to drawing a sample $X_t$ from $\nu_i = \mathcal{N}(\mu_i, \sigma^2)$, and earning $\$X_t$. If $X_t < 0$, you should pay that amount instead. Your total cumulative reward, over $T$ rounds is $\sum_{t=1}^T X_t$. We will measure how well we perform via our *average regret*, defined below:

$$R_T = \max\{\mu_1, \mu_2\} - \frac{1}{T}\sum_{t=1}^T X_t.$$

We wish to design an algorithm whose average regret vanishes[1] with $T$ in expectation, i.e $\mathbb{E}[R_T] \to 0$ as $T \to \infty$.

**Algorithm:** A student proposes the following simple algorithm. First sample each of the distributions $N$ times (where $N < T/2$). Then, for the remaining $T - 2N$ rounds, sample the distribution with the highest observed sample mean

[1] Intuitively, if we knew *a priori* which mean was larger, we will always pull the arm with the highest mean and have $\mathbb{E}[R_T] = 0$ as $\frac{1}{T}\mathbb{E}[\sum_t X_t] = \max\{\mu_1, \mu_2\}$. If $\mathbb{E}[R_T] \to 0$, this means we are able to learn which of the two distributions has a larger mean and converge towards the correct answer as we collect more samples.

using the $N$ samples. That is, $I_t$ is chosen as follows:

$$I_t = \begin{cases} 1 & \text{if } t \le N, \\ 2 & \text{if } N+1 \le t \le 2N, \\ 1 & \text{if } t > 2N \text{ and } \widehat{\mu}_1 \ge \widehat{\mu}_2, \\ 2 & \text{if } t > 2N \text{ and } \widehat{\mu}_1 < \widehat{\mu}_2. \end{cases}$$

$$\text{where, } \widehat{\mu}_1 = \frac{1}{N}\sum_{t=1}^{N} X_t, \qquad \widehat{\mu}_2 = \frac{1}{N}\sum_{t=N+1}^{2N} X_t,$$

For what follows, let $\Delta = |\mu_1 - \mu_2|$ denote the gap between the two means.

1. **[5 pts]** *(Regret decomposition)* Establish the following identity for the expected average regret:

$$\mathbb{E}[R_T] = \frac{N\Delta}{T} + \frac{(T-2N)\Delta}{T}\Phi\left(-\Delta\sqrt{\frac{N}{2\sigma^2}}\right)$$

   Here, $\Phi(x) = \mathbb{P}_{Z\sim\mathcal{N}(0,1)}(Z < x)$ is the CDF of the standard normal distribution.

2. **[2 pts]** Using the result from part 1 and the fact that $\mu_1, \mu_2 \in [0,1]$, show the following upper bound on the expected average regret.

$$\mathbb{E}[R_T] \le \frac{N}{T} + \Delta \exp\left(\frac{-N\Delta^2}{4\sigma^2}\right).$$

3. **[3 pts]** Use the result in part 2 to show the following upper bound.

$$\mathbb{E}[R_T] \le \frac{N}{T} + C\frac{\sigma}{\sqrt{N}}, \qquad \text{where, } C = \sqrt{2}e^{-1/2}.$$

   **Hint:** Consider the function $f(x) = \log(x) - \alpha x^2$, where $\alpha > 0$. What is the maximizer of $f$?

4. **[2 pts]** *(An optimal choice of $N$.)* Specify a choice for $N$, depending only on $\sigma$ and $T$, so that the upper bound in part 3 is minimized. Are you able to achieve $\mathbb{E}[R_T] \to 0$ as $T \to \infty$? If so, at what rate does it go to zero?

5. **[2 pts]** *(Exploration–exploitation trade-off.)* Let $N^\star$ denote the optimal choice in part 4. In words, explain what would happen had we chosen $N \ll N^\star$ or $N \gg N^\star$.

**N.B.** Later in class, we will study several models for adaptive decision-making. The model discussed in this question is an example of a *stochastic bandit*, which is one paradigm for decision-making. In bandit settings, we often have to trade-off between *exploration* (learning about the environment) and *exploitation* (leveraging what we have learned to maximize rewards). The above algorithm is a simple, albeit sub-optimal, approach that separates exploration (the first $2N$ rounds) from exploitation (the remaining $T - 2N$ rounds). In class, we will look at better algorithms to manage this trade-off which have faster rates of convergence.

# 3 Short questions from Chapter 0

## 3.1 Basic inequalities

1. **[5 pts]** Let $\frac{1}{\infty} = 0$. Show that the following statements are true for all $x \in \mathbb{R}^d$ and $p, q$ such that $1 \le p < q \le \infty$:

$$\textit{(i)}\ \|x\|_q \le \|x\|_p, \qquad \text{and} \qquad \textit{(ii)}\ \frac{1}{d^{1/p}}\|x\|_p \le \frac{1}{d^{1/q}}\|x\|_q.$$

   **Hint.** Consider using Hölder's inequality for *(ii)*.

2. **[2 pts]** Let $\mathcal{B}_p = \{x; \|x\|_p \le 1\}$ and let $\widetilde{\mathcal{B}}_p = \{x; d^{-1/p}\|x\|_p \le 1\}$. Draw a picture illustrating $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_\infty, \widetilde{\mathcal{B}}_1, \widetilde{\mathcal{B}}_2, \widetilde{\mathcal{B}}_\infty$ in $\mathbb{R}^2$, all in the same figure.

## 3.2 Bayesian inference

1. **[3 pts]** Let a parameter of interest $\mu$ have prior distribution $\mu \sim \mathcal{N}(\mu_0, \tau^2)$. Given $\mu$, we observe $S = \{X_1, \ldots, X_n\}$ with independent draws $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (with known $\sigma^2$). Find the posterior distribution of $\mu$ given $S$.

## 3.3 Covering and packing numbers

1. **[4 pts]** *(Packing-covering sandwich)* Let $\mathcal{X}$ be a set and let $\rho$ be a pseudo-metric defined on $\mathcal{X}$. Show that for any $\epsilon > 0$ and $A \subset \mathcal{X}$, we have $M(2\epsilon, A, \rho) \leq N(\epsilon, A, \rho) \leq M(\epsilon, A, \rho)$.
2. **[2 pts]** Let $A \subset \mathbb{R}^d$ and $\epsilon > 0$. Let $N_p(\epsilon, A) = N\left(\epsilon, A, \frac{1}{d^{1/p}}\|\cdot\|_p\right)$ denote the covering number under the metric $\rho(x, y) = \frac{1}{d^{1/p}}\|x - y\|_p$. Show that, $N_1(A, \epsilon) \leq N_2(A, \epsilon) \leq \cdots \leq N_\infty(A, \epsilon)$.