Homework 0: review of calculus, linear algebra, probability and statistics

Hugo Chardon
Department of Statistics, UC Berkeley

DUE DATE: Monday, February 9th, 2026 (11:00am)

Instructions:

- Questions marked with a cup of coffee (☕) are harder and are for students enrolled in Stat 254. Students from 154, don't hesitate to try! All efforts appreciated.
- Submit your answers as a single pdf on gradescope. You can use whatever tool you like to produce the pdf (LaTeX, markdown, quarto documents, Jupyter, scanned handwritten notes for mathematical problems (make them easy to read!), etc...).

1 Extreme values and gradients

1.1 Computing gradients

Recall from the lecture notes the following alternative, more "intrinsic" definition of gradients than the one using partial derivatives.

Definition 1 (Intrinsic definition of gradient)**.** Let $F : \mathbb{R}^d \to \mathbb{R}$ and let $\theta_0 \in \mathbb{R}^d$. F is said to be *differentiable* at θ_0 if there exists a vector $g(\theta_0)$ such that for every $h \in \mathbb{R}^d$ (think of this as a small increment),

$$F(\theta_0 + h) = F(\theta_0) + \langle g(\theta_0), h\rangle + o(h)\,, \tag{1}$$

where $o(h)$ denotes (in Landau notation) a quantity that goes to 0 faster than $\|h\|$ as $\|h\|$ goes to 0, namely a quantity $\varepsilon(h)$ such that $\varepsilon(h)/\|h\| \underset{h\to 0}{\to} 0$. If such a vector $g(\theta_0)$ exists, then we call it *gradient of F at θ_0*, denoted by $\nabla F(\theta_0)$.

Problem 1. Use Definition 1 to compute the gradient of the following functions. You may use the following method. Fix $\theta_0 \in \mathbb{R}^d$, and compute $F(\theta_0 + h)$ for every $h \in \mathbb{R}^d$. Then, identify the (unique) vector that plays the role of $g(\theta_0)$ from (1).

1. Let $u \in \mathbb{R}^d$ and let $F : \theta \mapsto \langle u, \theta\rangle$.
2. $F : \theta \mapsto \|\theta\|$
3. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix and let $F : \theta \mapsto \langle A\theta, \theta\rangle$.
4. Let $A \in \mathbb{R}^{n\times d}$ for integers $n, d \geqslant 1$. Let $y \in \mathbb{R}^n$ and let $F : \theta \mapsto \|A\theta - y\|^2$.

1.2 Extrema

This part is taken from Prof. R. Giordano's Stat 154/254 homework, past semesters.

Let K be some set, and $f : K \to \mathbb{R}$. The expression $\inf_{x\in K} f(x)$ or simply $\inf_K f$ denotes the largest lower bound for f over K, that is, the largest real number m such that for all $x \in K$, $f(x) \geqslant m$. The symbol "inf" stands for "infimum". Analogously, $\sup_{x\in K} f(x)$ denotes the smallest upper bound for f over K, and is pronounced "supremum".

The infimum is like the minimum and the supremum is like the maximum, with the important difference that the infimum and supremum always exist, whereas the minimum and maximum might not. (For this to be true, we must allow for the supremum and infimum to be infinite, which we will in this class.)

The notation

$$\operatorname*{argmin}_{x\in K} f(x) \ (= \operatorname*{argmin}_K f) \quad \text{and} \quad \operatorname*{argmax}_{x\in K} f(x) \ (= \operatorname*{argmax}_K f)$$

denote, respectively, the set of all points $x \in K$ at which $f(x)$ achieves its minimum in K, and that where f achieves its maximum. If no such points exist, these sets are empty.

Problem 2. Here we compute some extrema.

1. Suppose f achieves its minimum in K, meaning there exists some $x_0 \in K$ such that $f(x_0) \leqslant f(x)$ for all $x \in K$. When this happens, show that the infimum is the same as the minimum, i.e. $\inf_{x\in K} f(x) = \min_{x\in K} f(x)$.

2. Suppose that $\operatorname*{argmax}_{x\in K} f(x)$ is non-empty. Show that the supremum is the same as the maximum, i.e. $\sup_{x\in K} f(x) = \max_{x\in K} f(x)$.

3. Find $\inf_{x\in\mathbb{R}} \exp(x)$. Show that there is no x_0 such that $\exp(x_0) = \inf_{x\in\mathbb{R}} \exp(x)$, so $\min_{x\in\mathbb{R}} \exp(x)$ does not exist.

4. Let $U = (0, 1)$ denote the open unit interval and let $f : U \to \mathbb{R}, x \mapsto x$ be the identity function. Find $\sup_U f$, and show that $\max_U f$ does not exist. What minor modification in the definition of the domain U could we make that would make the maximum exist?

5. Let $f : x \mapsto 1$ be a constant function. Find $\operatorname*{argmin}_{x\in(0,1)} f(x)$.

6. (☕) Identify a function defined on all of $[0, 1]$ for which $\sup_{x\in[0,1]} f(x) = 1$ but for which the maximum does not exist.

7. (☕) Let $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and find $\operatorname*{argmin}_{x\in\mathbb{R}^2} \langle Ax, x\rangle$.

2 Linear algebra

In this section we review some linear algebra related to linear regression.

Recall the linear least-squares problem. We have an unknown probability distribution P on $\mathbb{R}^d \times \mathbb{R}$ and an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with common distribution P. We define for every $\theta \in \mathbb{R}^d$,

$$L(\theta) = \mathbf{E}[(Y - \langle\theta, X\rangle)^2], \ (X, Y) \sim P\,, \quad \text{and} \quad \widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle\theta, X_i\rangle)^2\,,$$

respectively the population and empirical risks for the quadratic loss. Define also

$$\theta^* \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} L(\theta)\,, \quad \text{and} \quad \widehat{\theta}_n \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \widehat{L}_n(\theta)\,. \tag{2}$$

Finally, define the population and empirical covariance matrices of the design X as

$$\Sigma = \mathbf{E}[XX^\top], \quad \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top .$$

Problem 3.

1. Check that Σ and $\widehat{\Sigma}_n$ are both positive semidefinite (PSD).
2. Prove that θ^* is unique if and only if $\Sigma \succ 0$ (positive definite).
3. The random vector X, or its distribution P_X, is said to be *non-degenerate* if there is no strict subspace $V \subsetneq \mathbb{R}^d$ (a linear subspace which is not $\mathbb{R}^d$ itself) such that $P_X(V) = \mathbf{P}(X \in V) = 1$. Intuitively, it means that X "is really d-dimensional". Show that X is not degenerate if and only if $\Sigma \succ 0$.
4. **From now on, and until the end of the problem, we assume that** $n \geqslant d$. The random vector X, or its distribution P_X, is said to be ***strictly*** *non-degenerate* if for any hyperplane $H \subset \mathbb{R}^d$, $P_X(H) = \mathbf{P}(X \in H) = 0$. Show that P_X is strictly non-degenerate if and only if $\widehat{\Sigma}_n \succ 0$ almost surely.
5. Show that the vectors $X_1, \dots, X_n$ almost surely span $\mathbb{R}^d$. (In short, and in general if we were not assuming $n \geqslant d$ this amounts to $\operatorname{rank}(X_1, \dots, X_n) = \min(n, d)$ almost surely.)
6. Conclude that the OLS $\widehat{\theta}_n$ is almost surely uniquely defined if and only if P_X is strictly non-degenerate.
7. (☕) What can be said about linear regression when $n < d$? Does the solution of the OLS as a minimizer still make sense? If so, describe the obtained predictor.

3 Probability and statistics

Problem 4.

1. For $\theta \in \mathbb{R}^d$, let P_θ denote the multivariate Gaussian distribution on $\mathbb{R}^d$ with mean θ and identity covariance, $P_\theta = \mathsf{N}(\theta, I_d)$.
   (a) Write the density p_θ of this distribution.
   (b) Let $\theta^* \in \mathbb{R}^d$ and let $X_1, \dots, X_n$ be an i.i.d. sample from P_{θ^*}. Compute the maximum-likelihood estimator (MLE) $\widehat{\theta}_n$ of θ^*.
   (c) Show that $\widehat{\theta}_n$ is consistent and asymptotically normal, and give the covariance matrix of the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta^*)$.
2. Now, assume a well-specified, Gaussian linear model with fixed design: let $x_1, \dots, x_n \in \mathbb{R}^d$ be fixed design vectors and $Y_i = \langle \theta^*, x_i \rangle + \varepsilon_i$ for $i \in \{1, \dots, n\}$, where θ^* is an unknown fixed vector in $\mathbb{R}^d$ and ε_i are i.i.d. with distribution $\mathsf{N}(0, \sigma^2)$, The level of noise σ^2 is considered to be known (we don't seek to estimate it, we are only interested in estimating θ^*). Compute the MLE $\widehat{\theta}_n^{\mathsf{ML}}$ of θ^* and compare with the usual least-squares estimator (2).

Problem 5 (Hoeffding's inequality)**.** Let $n \geqslant 1$ and $X, X_1, \dots, X_n$ be i.i.d. random variables taking value in the interval $[0, 1]$. Let

$$S_n = X_1 + \cdots + X_n \quad \text{and} \quad \mu_n = S_n/n .$$

For every random variable Z, define the functions $M_Z(\lambda) = \mathbf{E}[\exp(\lambda Z)]$ and $\psi_Z(\lambda) = \log(M_Z(\lambda))$. M_Z is known as the moment-generating function, or Laplace transform of Z. The goal of this

problem is to prove a concentration inequality for μ_n, that is, letting $\delta \in (0,1)$ denote a failure level, an inequality of the form:

$$\mathbf{P}\big(|\mu_n - \mathbf{E}X| \leqslant r(\delta, n)\big) \geqslant 1 - \delta\,, \tag{3}$$

where $r(\delta, n)$ is a rate that typically decreases with δ and n (r increases for a higher desired level of condidence and decreases as the sample size increases).

1. Check that $M_{X-\mathbf{E}X}(\lambda)$ is defined for every $\lambda \in \mathbb{R}$.

2. (Symmetrization lemma) Let X' be an independent copy of X. This means that X' has the same distribution as X but is independent of X. Let also ε be a Rademacher variable (*i.e.*, $\mathbf{P}(\varepsilon = 1) = \mathbf{P}(\varepsilon = -1) = 1/2$) independent of both X and X'. Recall Jensen's inequality: let φ be a convex function and Z an integrable random variable. Then $\varphi(\mathbf{E}[Z]) \leqslant \mathbf{E}[\varphi(Z)]$ (this remains true if $\mathbf{E}[\varphi(Z)] = +\infty$). The conditional version is also valid (replacing the expectation by a conditional expectation in the previous inequality). Let also $\lambda > 0$ be fixed for this question.

   (a) Argue that $\varepsilon \cdot (X - X')$ has the same distribution as $X - X'$.

   (b) Noting that $X' = \mathbf{E}[X'|X]$ (why?), use the conditional Jensen inequality to show that

   $$\mathbf{E}\big[\exp\big\{\lambda(X - \mathbf{E}X)\big\}\big] \leqslant \mathbf{E}\big[\exp\big\{\lambda\varepsilon(X - X')\big\}\big]\,. \tag{4}$$

   The name of the lemma comes from the fact that we bound from above the Laplace transform of $X - \mathbf{E}X$ by that of the symmetric random variable $\varepsilon \cdot (X - X')$.

3. Still with ε denoting a Rademacher variable, check that for all $\lambda \in \mathbb{R}$,

$$\mathbf{E}[\exp(\lambda\varepsilon)] = \frac{e^{\lambda} + e^{-\lambda}}{2} = \cosh(\lambda)\,.$$

4. Using the power series for the hyperbolic cosine, show that for all $\lambda \in \mathbb{R}$,

$$\mathbf{E}[\exp(\lambda\varepsilon)] \leqslant e^{\lambda^2/2}\,. \tag{5}$$

5. Show that for all λ,

$$\mathbf{E}\big[\exp\big\{\lambda(X - \mathbf{E}X)\big\}\big] \leqslant e^{\lambda^2/2}\,, \tag{6}$$

   i.e., that $\psi_{X-\mathbf{E}X}(\lambda) \leqslant \lambda^2/2$. (Hint: use the symmetrization lemma (4), then condition on both variables X and X' and use (5)).

6. Let $\lambda \geqslant 0$ and $t \geqslant 0$. Show that

$$\mathbf{P}(S_n - \mathbf{E}S_n > t) \leqslant \exp\big(\psi_{S_n - \mathbf{E}S_n}(\lambda) - \lambda t\big)\,. \tag{7}$$

7. Check that for all λ, $\psi_{S_n - \mathbf{E}S_n}(\lambda) = n\psi_{X-\mathbf{E}X}(\lambda)$. Use (6) to further bound the right-hand side of (7) and deduce that

$$\mathbf{P}(S_n - \mathbf{E}S_n > t) \leqslant \exp\Big(-\frac{t^2}{2n}\Big)\,. \tag{8}$$

8. Conclude: deduce an inequality of the form (3) involving $\mu_n = S_n/n$ that holds for every $\delta \in (0,1)$.

9. (☕) Compare the rate that you obtained in the last question with what you would get from the central limit theorem. To do so, check what the tail profile of the standard distribution is like. To that end, let $G \sim \mathsf{N}(0,1)$ and bound from above, for $x \geqslant 0$,

$$\mathbf{P}(G > t) = \int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \mathrm{d}t \,.$$

(Hint: integrate by parts).

10. (☕) Can Hoeffding's inequality (8) be improved? By which we mean: is it possible to find a constant $C > 1/2$ such that

$$\mathbf{P}(S_n - \mathbf{E}S_n > t) \leqslant \exp\Big(- C\frac{t^2}{n}\Big) \,?$$

(Hint: this amounts to bounding the MGF as $\mathbf{E}\big[\exp\big\{\lambda(X - \mathbf{E}X)\big\}\big] \leqslant e^{K^2\lambda^2}$ with a better constant K, that is $K^2 \leqslant 1/2$).