STAT547U Diagnostic Problem Set

Your Name

The purpose of this problem set is to ensure that you are fluent with linear algebra and basic probability theory, which will be necessary for the course. If these problems feels extremely difficult, then you will likely find this course to be technically overwhelming.

There are 3 questions (broken into subquestions) in this problem set for a total of 60 pts.

A quick note on notation. Variable names should use the following convention:

- Scalars (deterministic and random) will be represented by lowercase/non-bold letters (e.g. a, θ, etc.).
- Vectors (deterministic and random) will be represented by lowercase/bold letters (e.g. $\boldsymbol{a}$, $\boldsymbol{\theta}$, etc.).
- Matrices (deterministic and random) will be represented by uppercase/bold letters (e.g. $\boldsymbol{A}$, $\boldsymbol{\Theta}$, etc.).

---

1) Introduction to Functional Analysis (20 pts)

Let $\mu(\boldsymbol{x}_i)$ be a probability distribution over $\mathbb{R}^d$ with mean 0 and covariance $\boldsymbol{\Sigma}$; i.e.

$$\mathbb{E}[\boldsymbol{x}_i] = 0, \qquad \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] = \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}.$$

Consider the space of linear $\mathbb{R}^d \to \mathbb{R}$ functions $\mathcal{H} := \{f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x} : \boldsymbol{\theta} \in \mathbb{R}^d\}$.

1. (4 points. L2 inner products) For any two functions $f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$ and $f'(\boldsymbol{x}) = \boldsymbol{\theta}'^\top \boldsymbol{x}$, define the $L2$ inner product $\langle f, f' \rangle_{\mathcal{H}}$ as
$$\langle f, f' \rangle_{L2} := \mathbb{E}[f(\boldsymbol{x}) f'(\boldsymbol{x})].$$
Show that this inner product is equal to $\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}'$.

2. (4 points. Orthonormal functions) An orthonormal basis of $\mathcal{H}$ is a set of functions $\{\phi_1, \ldots, \phi_d\}$ such that
   - any function $f(\boldsymbol{x}) \in \mathcal{H}$ can be written as the linear combination of $\phi_1, \ldots, \phi_d$ and
   - $\mathbb{E}[\phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x})] = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

   Construct an orthonormal basis of $\mathcal{H}$—$\phi_1(\boldsymbol{x}), \ldots, \phi_d(\boldsymbol{x})$—in terms of $(\lambda_1, \boldsymbol{v}_1), \ldots, (\lambda_d, \boldsymbol{v}_d)$—the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$.

   (Hint: This answer should take you ≈ 5 lines.)

3. (4 points. $\mathcal{H}$ inner products) For any two functions $f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$ and $f'(\boldsymbol{x}) = \boldsymbol{\theta}'^\top \boldsymbol{x}$, define the $\mathcal{H}$ inner product $\langle f, f' \rangle_{\mathcal{H}}$ as
$$\langle f, f' \rangle_{L2} := \sum_{i=1}^{d} \tfrac{1}{\sigma_i} \langle f, \phi_i \rangle_{L2} \langle f', \phi_i \rangle_{L2},$$
where $\{\phi_1, \ldots, \phi_d\}$ is any orthonormal basis of $\mathcal{H}$, and $\sigma_1, \ldots, \sigma_d$ are scalars such that, for any $\boldsymbol{z} \in \mathbb{R}^d$ that is fixed (i.e. not random),
$$\int \phi_i(\boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{z}) d\mu(\boldsymbol{x}) = \mathbb{E}[\phi_i(\boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{z})] = \sigma_i \boldsymbol{z}$$

Show that this inner product is equal to $\boldsymbol{\theta}^\top \boldsymbol{\theta}'$.

(Hint: This proof should take you ≈ 6 lines.)

4. (4 points. $\mathcal{H} \subset L2$.) Now assume that $d \to \infty$. Assuming that the eigenvalues of $\boldsymbol{\Sigma}$ are summable (i.e. $\lim_{d\to\infty} \sum_{i=1}^{d} \lambda_i < \infty$), prove that—for any $f \in \mathcal{H}$—$\|f\|_{L2} < \infty$ if $\|f\|_{\mathcal{H}} < \infty$, where

$$\|f\|_{L2} := \sqrt{\langle f, f \rangle_{L2}}, \qquad \|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}.$$

(Note: if you're smart with your linear algebra, this proof should take you ≈ 4 lines.)

5. (4 points. $L2 \not\subset \mathcal{H}$.) Show that the converse is not true. I.e., as $d \to \infty$, construct a function where $\|f\|_{L2} < \infty$ but $\|f\|_{\mathcal{H}} = \infty$. (You should continue to assume that $\lim_{d\to\infty} \sum_{i=1}^{d} \lambda_i < \infty$.)

---

2) Ridge(less) Linear Regression (20 pts)

Assume that we are performing linear regression to predict a real-valued response variable $y \in \mathbb{R}$ from a d-dimensional input variable $\boldsymbol{x} \in \mathbb{R}^d$ using the parameter vector $\boldsymbol{\theta}^*$:

$$y = \boldsymbol{x}^\top \boldsymbol{\theta}^*$$

Given training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, the ridge regression estimator $\hat{\boldsymbol{\theta}}_\lambda$ is given by:

$$\hat{\boldsymbol{\theta}}_\lambda = \left(\lambda \boldsymbol{I} + \boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{1}$$

- $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is a matrix where each *row* represents a training input (i.e. $\boldsymbol{X}^\top = \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n \end{bmatrix}$),
- $\boldsymbol{y} \in \mathbb{R}^n$ is the concatenation of the training responses, and
- $\lambda > 0$ is the ridge constant.

For all problems, assume that $\boldsymbol{X}$ is full rank..

1. (5 points.) Using Woodbury's matrix inversion lemma, prove that $\hat{\boldsymbol{\theta}}_\lambda$ is equal to

$$\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{X}^\top \left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{y}. \tag{2}$$

(Hint: after applying Woodbury, consider rewriting $\boldsymbol{I}$ as $(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{X}^\top)$.)

2. (3 points.) We now have two formula for $\hat{\boldsymbol{\theta}}_\lambda$: Eq. (1) and Eq. (2). From a computational perspective, which formula is preferable when $n > d$? When $d < n$? Justify your answer in ≈ 2 sentences. (Hint: think about the matrices you have to invert.)

3. (3 points.) Consider the scenario where $d > n$; i.e. we have more features than training data. This scenario is often referred to as the *overparameterized regime*. Assuming that $\boldsymbol{X}$ is full rank, show that the ridge estimator *interpolates* the training data as $\lambda \to 0$; i.e.

$$\lim_{\lambda \to 0} \boldsymbol{X} \hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{y}.$$

This proof should take you ≈ 2 lines.

4. (3 points.) Now consider the scenario where $n < d$; i.e. we have more training data than features. This scenario is often referred to as the *underparameterized regime*. Assuming that $\boldsymbol{X}$ is full rank, show that
$$\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{U}^\top,$$
where $\boldsymbol{U}$ is some $n \times d$ orthonormal matrix.

5. (3 points.) In ≈ 3 sentences, derive the eigenvalues and the corresponding eigenvectors of $\boldsymbol{U}\boldsymbol{U}^\top$.

6. (3 points.) Putting the last two results together, argue why—in general—$\hat{\boldsymbol{\theta}}_\lambda$ does not interpolate the training data in the *underparameterized ridgeless regime* ($\lambda = 0$, $n > d$).

---

3) Bias-Variance Tradeoff (20 pts)

Now imagine that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the previous problem are i.i.d. samples from a distribution $\mu(\boldsymbol{x})$ and $y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\theta}^*, \sigma^2)$ for each $i \in [1, n]$. Assume that the y_i are conditionally independent given $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\theta}}_\lambda$ be the ridge estimator given in Eq. (1) (or, equivalently, in Eq. (2)).

Consider a new (independent) test point $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \sim \mu(\boldsymbol{x})$ and $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\theta}^*, \sigma^2)$, the *risk* is defined as
$$\mathcal{R} := \mathbb{E}\left[\left(\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{x}^\top \boldsymbol{\theta}^*\right)^2\right].$$

1. (8 points.) Decompose $\mathcal{R}$ into two components that represent *squared bias* and *variance* of $\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_\lambda$. (Hint: consider adding and subtracting $\boldsymbol{x}^\top \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda]$ inside the parentheses.)

2. (6 points.) Show that the squared bias of $\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_\lambda$ is 0 when $\lambda = 0$ and $n > d$. (This proof should take 3 lines when you write $\mathbb{E}[\hat{\boldsymbol{\theta}}_{\lambda=0}]$ as $\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}]$ and rewrite $\boldsymbol{y}$ in terms of $\boldsymbol{X}$.)

3. (8 points.) Show that bias term of is nonzero when $d > n$ and $\lambda \to 0$. For this problem you can make the following assumptions:

   - Limits and expectations can be interchanged— i.e. you can assume that $\lim_{\lambda \to 0} \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda] = \mathbb{E}[\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}]$. No need to dive into the measure theory!
   - $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] := \boldsymbol{\Sigma}$ is positive definite.

   (There are many ways to prove this statement. Use your linear algebra and probability skills. Be creative!)