Fall 2022 *Instructor: Ashwin Pananjady*

7750: Mathematical Foundations of Machine Learning

Linear algebra and probability for data analysis

Homework 2

Released: Sep 8 *Due: Sep 20, 11:59pm ET*

Note. All external sources and collaborators must be acknowledged in your submission. As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

Objective. To get comfortable with reasoning about Hilbert spaces and linear approximation.

Resources. Lectures, notes, and modules posted on and before Sep 14.

Notation: Capital boldface letters will typically be matrices, and small boldface letter will be vectors.

Problem 1 (Exercises from lecture). 30 points: In this problem, we will walk you through a few proofs of facts that were mentioned in lecture or notes but not explicitly proved. Parts (a-d) are about Cauchy sequences and completeness. Part (e) is about inner products and the so-called parallelogram law. Parts (f-h) help you verify that the space of finite-variance random variables can be viewed as an inner product space.

(a) Recall that we considered the vector space of real-valued, continuous functions on $[0, 1]$, denoted by $\mathcal{C}[0, 1]$. Equip this space with the standard inner product, with $\langle f, g\rangle := \int_0^1 f(t)g(t)dt$.

Also recall the sequence of functions $(f_n)_{n=1}^{\infty}$ defined as follows

$$f_n(t) = \begin{cases} 0 & \text{if } 0 \le t \le \frac{1}{2} - \frac{1}{2n} \\ 2nt + (1-n) & \text{if } \frac{1}{2} - \frac{1}{2n} < t \le \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < t \le 1. \end{cases}$$

Show that for any finite n, f_n is a continuous function.

(b) Now consider any pair of positive integers $n < m$, and evaluate the norm $\|f_n - f_m\|$. Your expression should be explicit and depend on n and m (recall that the norm is the one induced by the inner product).

(c) Using your expression above, show that the sequence (f_n) is a Cauchy sequence. Conclude that the inner product space introduced above is not complete by thinking about f_∞.

(d) (BONUS:) How might you define the "completion" of the inner product space above so that the resulting space is complete (and hence a Hilbert space)? Note that this will require extra reading that you don't need to do in principle, just if you're interested.

(e) Show that a norm $\|\cdot\|$ defined on a vector space $\mathcal{S}$ is induced by an inner product if and only if the following parallelogram law is satisfied for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$:

$$2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 = \|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2.$$

One of these directions is trickier than the other.

(f) Consider the space of all real-valued random variables with finite-variance. Show that this is a vector space, and call it $\mathcal{S}$.

(g) Note that for $X, Y \in \mathcal{S}$, we may view $\mathbb{E}[XY]$ as a function mapping $\mathcal{S} \times \mathcal{S} \to \mathbb{R}$. Show that this is a valid inner product on this vector space.

(h) For any pair of finite variance random variables (X, Y), the *conditional expectation* $\mathbb{E}[Y|X]$ is a function of X that is known to satisfy the following property: for all functions[1] ϕ

$$\mathbb{E}[(Y - \mathbb{E}[Y|X])\phi(X)] = 0.$$

Using this definition, prove that the mean squared error $\mathbb{E}[(Y - \phi(X))^2]$ of estimating Y from X is minimized by choosing $\phi(X) = \mathbb{E}[Y|X]$. I.e., the conditional expectation minimizes the mean squared error of estimation.

Hint: Think about how we proved the orthogonality principle without necessarily trying to formally define a subspace.

Problem 2 (Gram matrices and Gram–Schmidt). 20 points:

(a) As you know, a square $N \times N$ matrix $\mathbf{G}$ is *invertible* if

$$\mathbf{x}_1 \neq \mathbf{x}_2 \;\Leftrightarrow\; \mathbf{G}\mathbf{x}_1 \neq \mathbf{G}\mathbf{x}_2.$$

That is, $\mathbf{G}\mathbf{x}$ is different for every different $\mathbf{x}$. In other words, if you can show that $\mathbf{G}\mathbf{x} = \mathbf{0}$ only if $\mathbf{x} = \mathbf{0}$, then you have shown that $\mathbf{G}$ is invertible.

Let $\mathbf{v}_1, \ldots, \mathbf{v}_N$ be N linearly independent vectors in a Hilbert space, and let $\mathcal{T} = \mathsf{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. Show that if $\mathbf{z} \in \mathcal{T}$ and $\langle \mathbf{v}_n, \mathbf{z} \rangle = 0$ for all $n = 1, \ldots, N$, then it must be true that $\mathbf{z} = \mathbf{0}$.

(b) Show that if $\mathbf{v}_1, \ldots, \mathbf{v}_N$ are N linearly independent vectors in a Hilbert space, then the Gram matrix

$$\mathbf{G} = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \cdots & \langle \mathbf{v}_N, \mathbf{v}_1 \rangle \\ \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & & \langle \mathbf{v}_N, \mathbf{v}_2 \rangle \\ \vdots & & \ddots & \vdots \\ \langle \mathbf{v}_1, \mathbf{v}_N \rangle & \cdots & & \langle \mathbf{v}_N, \mathbf{v}_N \rangle \end{bmatrix},$$

is invertible. (Hint: use part (a).)

[1] In reality you need a measurability condition that we will ignore.

(c) Let us now explore an algorithm that takes a basis for a subspace and produces and orthonormal basis for that same subspace. Let $\mathbf{v}_1, \ldots, \mathbf{v}_N$ be a basis for a subspace $\mathcal{T}$ of a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ with induced norm $\| \cdot \|$. Define

$$\psi_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|},$$

then for $k = 2, \ldots, N$,

$$\begin{aligned} \mathbf{u}_k &= \mathbf{v}_k - \sum_{\ell=1}^{k-1} \langle \mathbf{v}_k, \psi_\ell \rangle \psi_\ell, \\ \psi_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}. \end{aligned}$$

Argue that for the $\mathbf{u}_2$ produced above that $\|\mathbf{u}_2\| > 0$, and so ψ_2 is well defined.

(d) Argue that $\mathsf{span}\{\psi_1, \psi_2\} = \mathsf{span}\{\mathbf{v}_1, \mathbf{v}_2\}$, and show that ψ_1 and ψ_2 are orthonormal. Hence $\{\psi_1, \psi_2\}$ is an orthonormal basis for $\mathsf{span}\{\mathbf{v}_1, \mathbf{v}_2\}$.

(e) Use induction to show that $\{\psi_1, \ldots, \psi_N\}$ is an orthonormal basis for $\mathcal{T}$. Part of this argument will be to ensure that $\mathbf{u}_k \neq \mathbf{0}$.

Problem 3. (Linear approximation with "bump" functions). 30 points: In this problem, we will develop the computational framework for approximating a function on $[0, 1]$ using scaled and shifted version of the classic bell-curve bump:

$$\phi(t) = e^{-t^2}.$$

Fix an integer $N > 0$ and define $\phi_k(t)$ as

$$\phi_k(t) = \phi\left(\frac{t - (k - 1/2)/N}{1/N}\right) = \phi\left(Nt - k + 1/2\right)$$

for $k = 1, 2, \ldots, N$. The $\{\phi_k(t)\}$ are a basis for the subspace

$$\mathcal{T}_N = \mathsf{span}\left\{\phi_k\right\}_{k=1}^{N}.$$

(a) For a fixed value of N, we can plot all of the $\phi_k(t)$ on the same set of axes. You can do this in Python using:

```
import numpy as np
import matplotlib.pyplot as plt

phi = lambda z: np.exp(-z**2)
t = np.linspace(0,1,1000)

plt.figure(1)
plt.clf()
for kk in range(N):
plt.plot(t, phi(N*t - (kk + 1) + 0.5))
```

Do this for $N = 10$ and $N = 25$ and one more value of N (of your choosing) and turn in your plots.

(b) Since $\{\phi_k\}$ is a basis for $\mathcal{T}_N$, we can write any $\mathbf{y} \in \mathcal{T}_N$ as

$$y(t) = \sum_{k=1}^{N} a_k \phi_k(t)$$

for some set of coefficients $a_1, \ldots, a_N \in \mathbb{R}^N$. In Python, we can plot $y(t)$ using

```
y = np.zeros(1000)
for jj in range(N):
y = y+a[jj]*phi(N*t - (jj + 1) + 0.5)
plt.figure()
plt.plot(t,y)

y = np.zeros(1000)
```

Do this for $N = 4$, and $a_1 = -1/2, a_2 = 3, a_3 = 2, a_4 = -1$ and submit a plot.

(c) Define the function $f(t)$ on $[0, 1]$ as

$$f(t) = \begin{cases} 4t & 0 \leq t < 1/4 \\ -4t + 2 & 1/4 \leq t < 1/2 \\ -\sin(14\pi t) & 1/2 \leq t \leq 1 \end{cases}.$$

Write a function that finds the closest point $\hat{\mathbf{f}}$ in $\mathcal{T}_N$ to $\mathbf{f}$ for any fixed N. By "closest point", we mean that $\hat{x}(t)$ is the solution to

$$\mathsf{minimize}_{\mathbf{y}\in\mathcal{T}_N} \ \|\mathbf{f} - \mathbf{y}\|_{L_2([0,1])}, \quad \|\mathbf{f} - \mathbf{y}\|^2_{L_2([0,1])} = \int_0^1 |f(t) - y(t)|^2 dt.$$

Turn in your code and four plots; one of which has $f(t)$ and $\hat{f}(t)$ plotted on the same set of axes for $N = 5$, and then repeat for $N = 10, 20$, and 50.

Hint: You can create a function pointer for $f(t)$ in Python by using

```
f = lambda z: (z < .25)*(4*z) + (z >= 0.25)*(z < 0.5)*(-4*z+2) - \
 (z>= 0.5)*np.sin(14*np.pi*z)
```

and then calculate the continuous-time inner product $\langle \mathbf{x}, \phi_k \rangle$ in Python with

```
import scipy.integrate as integrate
f_phik = lambda z: f(z)*phi(N*z - jj + 0.5)
integrate.quad(f_phik, 0, 1)
```

You can use similar code to calculate the entries of the Gram matrix $\langle \phi_j, \phi_k \rangle$. (There is a way to calculate the $\langle \phi_j, \phi_k \rangle$ analytically if you'd like — think about what happens when you convolve a bump with itself.)

Problem 4. (Finite dimensional linear regression to predict disease progression). 20 points In this problem, you will run linear regressions on a data set (`diabetes.csv`) containing $d = 10$ predictors (including 7 blood serum measurements) of a response `prog` (diabetes progression) in $n = 442$ patients. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the matrix formed by stacking all these predictors, and $\mathbf{y} \in \mathbb{R}^n$ denote the corresponding responses. You may use any Python package for this problem, but just doing standard linear algebra operations with numpy should suffice.

(a) Import the data and *standardize* each predictor, i.e., when you look at a particular predictor on all the samples and view it as a vector (i.e. a column vector of $\mathbf{X}$), you want its empirical mean to be 0 and its empirical variance to be 1.

(b) Perform linear regression on this data by writing down the least squares solution **with the intercept**.

(c) Perform linear regression on the data **without an intercept**. Compare your solution to the previous part and explain what just happened. We would like a justification rooted in linear algebra arguments.

(d) Let us now explore how to obtain a purely data-dependent estimate of the *test error*. Split your data set (at random) into two equal portions, use the first half to fit a linear regression as in part (b), and use the second half to measure the *test* error in predicting their corresponding response values. Do this 100 times and compute the average test error.

(e) Repeat part (d) with the linear regressions computed without intercepts.

(f) Depending on your observations in the previous two parts, which linear model—with or without intercepts—do you think provides a better representation of this data set?