College of Engineering
School of Aerospace Engineering

ISYE 7750: Mathematical Foundations of Machine Learning

Homework 2

Professor:
Ashwin Pananjady
Gatech ISYE Professor

Student:
Tomoki Koike
Gatech AE MS Student

September 18, 2022

Table of Contents

I Problem One

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

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

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.

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

Solution:
(a) Let $a = \frac{0}{2} - \frac{1}{2n}$ then

$$\begin{aligned} &\lim_{t \to a^+} f_n(t) = 0 \\ &\lim_{t \to a^-} f_n(t) = \lim_{t \to a} (2nt + (1-n)) = 2n\left(\frac{1}{2} - \frac{1}{2n}\right) + 1 - n = 0 \end{aligned}$$

Similarly,

$$\begin{aligned} &\lim_{t \to \frac{1}{2}^+} f_n(t) = \lim_{t \to \frac{1}{2}} (2nt + (1-n)) = n + 1 - n = 1 \\ &\lim_{t \to \frac{1}{2}^-} f_n(t) = 1 \end{aligned}$$

These limits hold for all n in the domain of $[0, 1]$. Hence, $f_n(t)$ is continuous in the domain of $[0, 1]$.

■

(b) Given positive integers $n < m$, the norm is

$$\|f_n - f_m\|_2 = \left(\int_0^1 (f_n - f_m)^2 dt\right)^{\frac{1}{2}}$$

Thus, first we calculate the inside of $(\cdot)^{\frac{1}{2}}$

$$\begin{aligned} &\int_0^{\frac{1}{2}-\frac{1}{2n}} (f_n - f_m)^2 dt + \int_{\frac{1}{2}-\frac{1}{2n}}^{\frac{1}{2}-\frac{1}{2m}} (f_n - f_m)^2 dt + \int_{\frac{1}{2}-\frac{1}{2m}}^{\frac{1}{2}} (f_n - f_m)^2 dt + \int_{\frac{1}{2}}^{1} (f_n - f_m)^2 dt \\ &= \int_{\frac{1}{2}-\frac{1}{2n}}^{\frac{1}{2}-\frac{1}{2m}} (2nt + 1 - n)^2 dt + \int_{\frac{1}{2}-\frac{1}{2m}}^{\frac{1}{2}} (2nt + 1 - n - 2mt - 1 + m)^2 dt \\ &= \frac{(m-n)^3}{6m^3 n} + \frac{(m-n)^2}{6m^3} \\ &= \frac{(m-n)^2}{6m^2 n} \end{aligned}$$

Hence,

$$\boxed{\|f_n - f_m\|_2 = \sqrt{\frac{(m-n)^2}{6m^2 n}} = \frac{m-n}{m\sqrt{6n}}.}$$

(c) From the previous problem (b) we can prove that this sequence is Cauchy since

$$\|f_n - f_m\|^2 = \frac{m-n}{m\sqrt{6n}} = \frac{1}{\sqrt{6n}} - \frac{n}{m\sqrt{6n}} < \frac{1}{\sqrt{6n}} \xrightarrow{n,m\to\infty} 0.$$

Now if $n, m \to \infty$ this means that $\frac{1}{2} - \frac{1}{2n} \to \frac{1}{2}$ and this makes f_n discontinuous which contradicts our proof of continuity in (a). Hence, from a contradictory example we can show that the inner product space is not complete and is not a Banach space.

■

(d) Using Bolzano–Weierstrass theorem, we can state that every bounded sequence has a convergent subsequence meaning that the Euclidean inner product space is closed and bounded [1].

(e) ($\Longrightarrow$) Suppose $\mathcal{S}$ is induced by an inner product space then

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 &= \langle x+y, x+y\rangle + \langle x-y, x-y\rangle \\ &= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle + \langle x,x\rangle - \langle x,y\rangle - \langle y,x\rangle + \langle y,y\rangle \\ &= 2\langle x,x\rangle + 2\langle y,y\rangle \\ &= 2\|x\|^2 + 2\|y\|^2 .\end{aligned}$$

($\Longleftarrow$) Assuming the parallelogram law we have that

$$\begin{aligned}&\|x+y\|^2 - \|x-y\|^2 \\ &\quad= \langle x+y, x+y\rangle - \langle x-y, x-y\rangle \\ &\quad= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle - \langle x,x\rangle - \langle -y,x\rangle - \langle x,-y\rangle - \langle -y,-y\rangle \\ &\quad= 4\langle x,y\rangle .\end{aligned}$$

Hence,

$$\langle x,y\rangle = \frac{\|x+y\|^2 - \|x-y\|^2}{4}. \tag{I.1}$$

We prove each property of the inner product space for this expression. First, from the equation I.1, it is obvious that $\langle x,y\rangle = \langle y,x\rangle$ and $\|x\| = \sqrt{\langle x,x\rangle}$. Second, we show that $\langle x+y, z\rangle = \langle x,z\rangle + \langle y,z\rangle$. By the parallelogram law we have that

$$2\|x+z\|^2 + 2\|y\|^2 = \|x+y+z\|^2 + \|x-y+z\|^2$$

Now we can rewrite the right hand side as

$$\begin{aligned}\|x+y+z\|^2 &= \|x\|^2 + \|y\|^2 + \|x+z\|^2 + \|y+z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|y-x+z\|^2 \\ \|x+y-z\|^2 &= \|x\|^2 + \|y\|^2 + \|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|x-y-z\|^2 - \frac{1}{2}\|y-x-z\|^2 \\ &= \|x\|^2 + \|y\|^2 + \|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|y-x+z\|^2\end{aligned}$$

Subtract the second equation from the first one in the equations right above and we have

$$\|x+y+z\|^2 - \|x+y-z\|^2 = \|x+z\|^2 + \|y+z\|^2 + \|x-z\|^2 + \|y-z\|^2 .$$

Then from equation I.1 and substituting in the above to the first line below, we have

$$\begin{aligned}\langle x+y, z\rangle &= \frac{1}{4}\left(\|x+y+z\|^2 - \|x+y-z\|^2\right)\\ &= \frac{1}{4}(\|x+z\|^2 + \|x-z\|^2) + \frac{1}{4}(\|y+z\|^2 + \|y-z\|^2)\\ &= \langle x, z\rangle + \langle y, z\rangle\,.\end{aligned}$$

For the third step we prove that $\langle \lambda x, y\rangle = \lambda \langle x, y\rangle \quad \forall\ \lambda \in \mathbb{R}$. This hold for $\lambda = -1$ and by the previous step and induction we have that $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle$ for all $\lambda \in \mathbb{N}$, thus for all $\lambda \in \mathbb{Z}$. Now if $\lambda = \frac{p}{q}$ with $p, q \in \mathbb{Z}, q \neq 0$ we get with $x' = \frac{x}{q}$ that

$$q\langle \lambda x, y\rangle = q\langle px', y\rangle = p\langle qx', y\rangle = p\langle x, y\rangle\,.$$

Dividing the above gives us what we desired, $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \forall\ \lambda \in \mathbb{Q}$. (The source referenced to solve this problem [3, 2]).

■

(f) A random variable supports the additive property if $\mathbf{X}_1, \mathbf{Y}_1 \in \mathcal{S}$ and $\mathbf{X}_1 = \mathbf{X}_2$ as well as $\mathbf{Y}_1 = \mathbf{Y}_2$ then $\mathbf{X}_1 + \mathbf{Y}_1 = \mathbf{X}_2 + \mathbf{Y}_2$. Also it agrees with the multiplicative property since $1 \cdot \mathbf{X}_1 = \mathbf{X}_1$ and $c\mathbf{X}_1 = c\mathbf{X}_2$.

■

(g) Let $X, Y, Z \in \mathcal{S}$ and $\lambda \in \mathbb{R}$. The conjugate symmetry is clear since

$$\mathbb{E}\left[\mathbf{XY}\right] = \mathbb{E}\left[\mathbf{YX}\right].$$

Then the linearity is obvious

$$\begin{gathered}\mathbb{E}\left[(\mathbf{X}+\mathbf{Y})\mathbf{Z}\right] = \mathbb{E}\left[\mathbf{XZ}+\mathbf{YZ}\right] = \mathbb{E}\left[\mathbf{XZ}\right] + \mathbb{E}\left[\mathbf{YZ}\right]\\ \mathbb{E}\left[\lambda\mathbf{X}\right] = \lambda\mathbb{E}\left[\mathbf{X}\right].\end{gathered}$$

Now to prove the positive definiteness we use the fact that the variance of a random variable is strictly positive, so

$$\begin{gathered}\mathrm{Var}\left[\mathbf{X}\right] = \mathbb{E}\left[\mathbf{X}^2\right] - \mathbb{E}\left[\mathbf{X}\right]^2 \geq 0\\ \therefore \mathbb{E}\left[\mathbf{X}^2\right] \geq \mathbb{E}\left[\mathbf{X}\right]^2 \geq 0\end{gathered}$$

■

(h) We want to show that

$$\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right] = \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbf{Y} - \phi(\mathbf{X}))^2\right].$$

To show this we solve the following

$$\begin{aligned}
&\underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbf{Y}-\phi(\mathbf{X}))^2\right] \\
&\quad= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]+\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right] \\
&\quad= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[((\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])^2-2(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))+(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right] \\
&\quad= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(-2(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))+(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right] \\
&\quad= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(-2(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-2(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])\phi(\mathbf{X})+(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right]
\end{aligned}$$

We are given that

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

thus,

$$\begin{aligned}
\underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbf{Y}-\phi(\mathbf{X}))^2\right] &= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(-2(\mathbf{Y}-\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right])\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]+(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right] \\
&= \underset{\phi(\mathbf{X})}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]-\phi(\mathbf{X}))^2\right] \\
&\geq 0
\end{aligned}$$

Hence, the global minimizer is when

$$\phi(\mathbf{X})=\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right].$$

■

II Problem Two

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

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

(b) Show that if $\mathbf{v}_1, \dots, \mathbf{v}_N$ 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).)

(c) Let us now explore an algorithm that takes a basis for a subspace and produces and orthonormal basis for that same subspace. Let $\mathbf{v}_1, \dots, \mathbf{v}_N$ 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 $k = 2, \dots, N$,

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

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 $\{\psi_1, \dots, \psi_N\}$ is an orthonormal basis for $\mathcal{T}$. Part of this argument will be to ensure that $\mathbf{u}_k \neq \mathbf{0}$.

Solution:
(a) Since $\mathbf{z} \in \mathcal{T}$, $\mathbf{z}$ can be represented as a linear combination of the vectors $\mathbf{v}_1, ..., \mathbf{v}_N$. Therefore,

$$\langle \mathbf{v}_n, \mathbf{z} \rangle = \left\langle \mathbf{v}_n, \sum_{j=1}^{N} \alpha_j \mathbf{v}_j \right\rangle = \sum_{j=1}^{N} \alpha_j \langle \mathbf{v}_j, \mathbf{v}_n \rangle$$

Now since the linear independence of the two vectors $\mathbf{v}_j$ and $\mathbf{v}_n$ does not imply that the inner product of those two are zero, which means that for $\langle v_n, z \rangle = 0$ to be true it must be that $\alpha_j = 0$, and therefore, the vector $\mathbf{z} = 0$.

■

(b) Let $\mathbf{v}_n = \sum_{k=1}^{n} a_{ki} \mathbf{e}_k$ where $\{\mathbf{e}_1, ..., \mathbf{e}_n\}$ is an orthonormal basis. Then we can represent

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_k \sum_l a_{ki} a_{lj} \langle \mathbf{e}_k, \mathbf{e}_l \rangle = \sum_k a_{ki} a_{kj}.$$

Then it follows that if $\mathbf{A}_{ij} = [a_{ij}]$ we obtain $\mathbf{G} = \mathbf{A}^\top \mathbf{A}$. Now if $\mathbf{G}\mathbf{x} = 0$ it implies that $\mathbf{x}^\top \mathbf{G}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^\top (\mathbf{A}\mathbf{x}) = 0$. Thus, $\mathbf{A}\mathbf{x} = 0$. Now, the matrix $\mathbf{A}$ has a rank of N if and only if $\{\mathbf{v}_1, ..., \mathbf{v}_N\}$ are linearly independent. Hence, if this set is linearly independent, then $\mathbf{A}\mathbf{x} = 0$ implying $\mathbf{x} = 0$. Then it follows that $\mathbf{G}\mathbf{x} = 0$ where $\mathbf{x} = 0$ and $\mathbf{G}$ has a rank of N and is invertible.

■

(c) This algorithm is famous for being the Grand-Schmidt algorithm for a method of orthonormalizing a set of vectors. Now if $\mathbf{u}_1 = \mathbf{v}_1$, then

$$\mathbf{u}_2 = \mathbf{v}_2 - \langle \mathbf{v}_2, \psi_1 \rangle \psi_1 = \mathbf{v}_2 - \left\langle \mathbf{v}_2, \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} \right\rangle \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \mathbf{v}_2 - \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle} \mathbf{u}_1 = \mathbf{v}_2 - \mathrm{proj}_{\mathbf{u}_1}(\mathbf{v}_2)$$

In a 2-dimensional diagram, this can be represented as below.

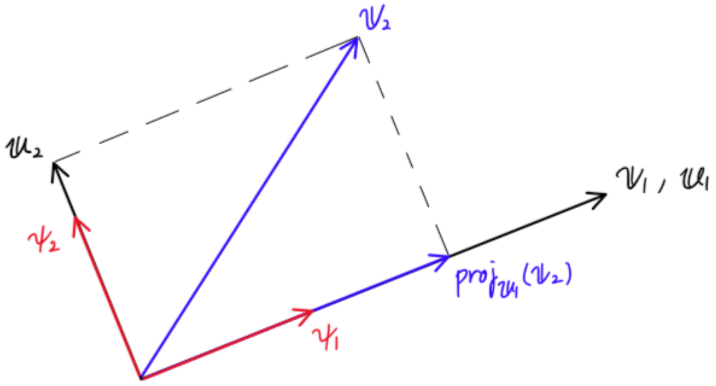


Figure 1: Visual representation of Grand-Schmidt for $\mathbf{u}_2$.

Now if $\|\mathbf{u}_2\| > 0$ it is clear that ϕ_2 is a unique vector and well-defined.

■

(d) To show $\text{span}\{\psi_1, \psi_2\} = \text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ we can show that $\text{span}\{\mathbf{u}_1, \mathbf{u}_2\} = \text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$. We have to show that $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$ can be represented as a linear combination of $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2\}$ and vice versa. From the previous problem we know that $\mathbf{u}_1 = \mathbf{v}_1$ and

$$\begin{aligned}\mathbf{u}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 \\ \mathbf{v}_2 &= \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 + \mathbf{u}_2 \\ \therefore \mathbf{v}_2 &= \alpha \mathbf{u}_1 + \mathbf{u}_2\end{aligned}$$

Thus, we can say that

$$\mathbf{u}_2 = -\alpha \mathbf{v}_1 + \mathbf{v}_2.$$

Thus, $\mathcal{V}$ and $\mathcal{U}$ both span the same subspace. To show that ϕ_1 and ϕ_2 are orthonormal we take the inner product of these vectors

$$\begin{aligned}\langle \psi_1, \psi_2 \rangle &= \left\langle \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}, \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \right\rangle = \frac{1}{\|\mathbf{u}_1\| \|\mathbf{u}_2\|} \langle \mathbf{u}_1, \mathbf{u}_2 \rangle \\ &= \frac{1}{\|\mathbf{u}_1\| \|\mathbf{u}_2\|} \left\langle \mathbf{v}_1, \mathbf{v}_2 - \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 \right\rangle \\ &= \frac{1}{\|\mathbf{u}_1\| \|\mathbf{u}_2\|} \left(\langle \mathbf{v}_1, \mathbf{v}_2 \rangle - \frac{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle \langle \mathbf{v}_1, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \right) \\ &= 0.\end{aligned}$$

■

(e) Now if we follow the algorithm we have

$$\begin{aligned}\mathbf{u}_{N+1} &= \mathbf{v}_{N+1} - \sum_{j=1}^{N} \text{proj}_{\mathbf{u}_j}(\mathbf{v}_{N+1}) \\ &= \mathbf{v}_{N+1} - \frac{\langle \mathbf{u}_1, \mathbf{v}_{N+1} \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 - \cdots - \frac{\langle \mathbf{u}_N, \mathbf{v}_{N+1} \rangle}{\langle \mathbf{u}_N, \mathbf{u}_N \rangle}\mathbf{u}_N,\end{aligned}$$

this shows that all $\mathbf{u}_i$ and $\mathbf{v}_i$ can be represented as a linear combination of each other. Futher, and sequentially it follows that $\mathbf{u}_2 \perp \mathbf{u}_1, \mathbf{u}_3 \perp \mathbf{u}_2, ..., \mathbf{u}_{N+1} \perp \mathbf{u}_N$. Thus,

$$\begin{aligned}\langle \mathbf{u}_{N+1}, \mathbf{u}_N \rangle &= \langle \mathbf{v}_{N+1}, \mathbf{u}_N \rangle - 0 - 0 - \cdots - \frac{\langle \mathbf{u}_N, \mathbf{v}_{N+1} \rangle \langle \mathbf{u}_N, \mathbf{u}_N \rangle}{\langle \mathbf{u}_N, \mathbf{u}_N \rangle} \\ &= 0.\end{aligned}$$

Thus, $\{\psi_1, ..., \psi_N\}$ are orthonormal basis of the subspace $\mathcal{T}$ by proof of induction.

■

III Problem Three

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

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

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.)

Solution:

(a) Below we have the plots for $N = 10,\ 25,\ 40.$

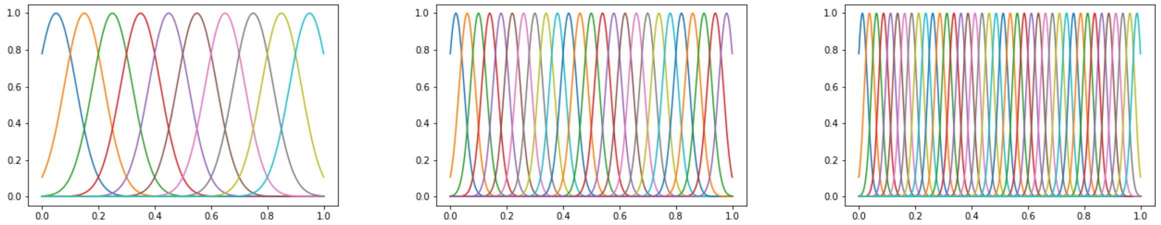


Figure 2: Plots of $\phi_k(t)$ function for different N values

(b) The plot is as follows

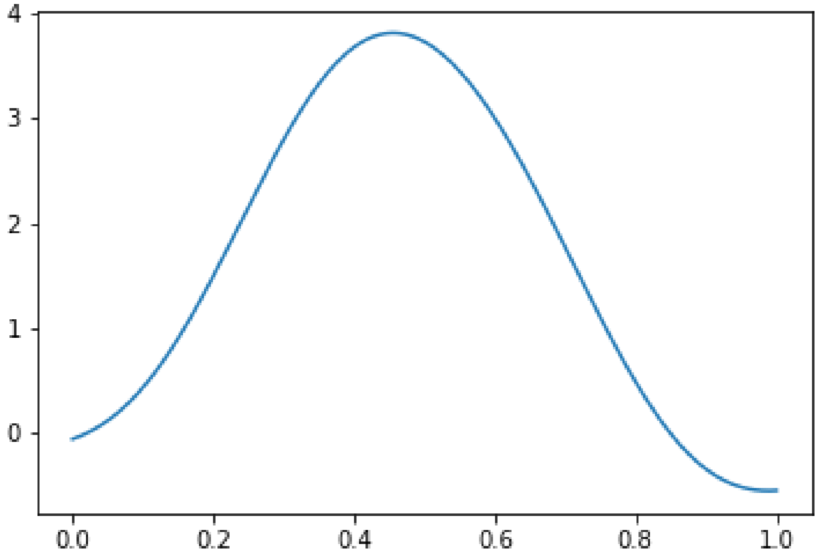
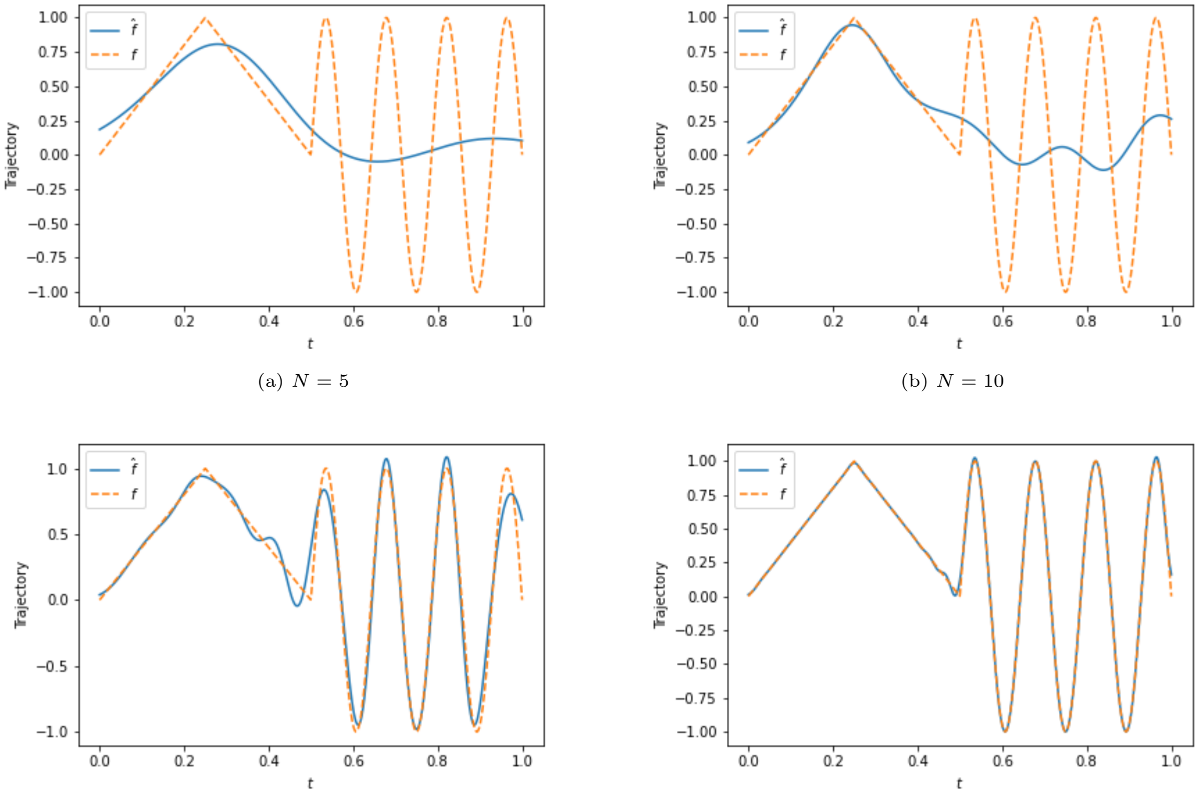


Figure 3: Plot of $y(t)$ for $N = 4$.

(c) The created plots are as follows. (The code is submitted separately)



(a) $N = 5$

(b) $N = 10$

(c) $N = 20$

(d) $N = 50$

Figure 4: $\hat{f}$ and f trajectorys using the Gram matrix.

IV Problem Four

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**.

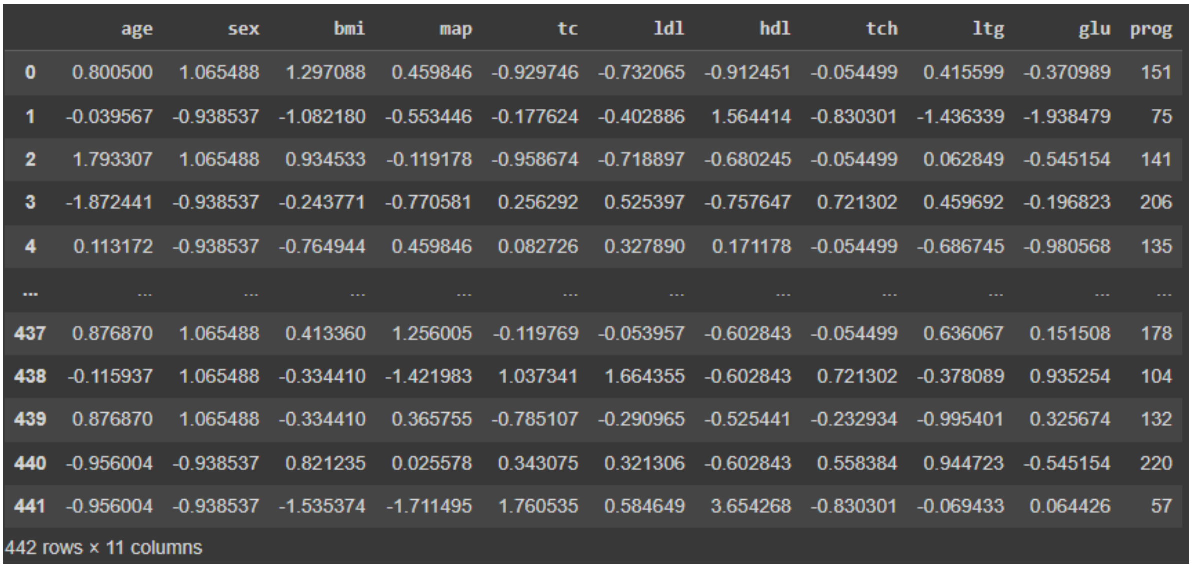
(c) Perform linear regression on the data **without an intercept**. Compare your solution to the previous part and explain what just happened. We would like a justification rooted in linear algebra arguments.

(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?

Solution:
(a) The standardize outcome is



	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	prog
0	0.800500	1.065488	1.297088	0.459846	-0.929746	-0.732065	-0.912451	-0.054499	0.415599	-0.370989	151
1	-0.039567	-0.938537	-1.082180	-0.553446	-0.177624	-0.402886	1.564414	-0.830301	-1.436339	-1.938479	75
2	1.793307	1.065488	0.934533	-0.119178	-0.958674	-0.718897	-0.680245	-0.054499	0.062849	-0.545154	141
3	-1.872441	-0.938537	-0.243771	-0.770581	0.256292	0.525397	-0.757647	0.721302	0.459692	-0.196823	206
4	0.113172	-0.938537	-0.764944	0.459846	0.082726	0.327890	0.171178	-0.054499	-0.686745	-0.980568	135
...	...	...	...	...	...	...	...	...	...	...	...
437	0.876870	1.065488	0.413360	1.256005	-0.119769	-0.053957	-0.602843	-0.054499	0.636067	0.151508	178
438	-0.115937	1.065488	-0.334410	-1.421983	1.037341	1.664355	-0.602843	0.721302	-0.378089	0.935254	104
439	0.876870	1.065488	-0.334410	0.365755	-0.785107	-0.290965	-0.525441	-0.232934	-0.995401	0.325674	132
440	-0.956004	-0.938537	0.821235	0.025578	0.343075	0.321306	-0.602843	0.558384	0.944723	-0.545154	220
441	-0.956004	-0.938537	-1.535374	-1.711495	1.760535	0.584649	3.654268	-0.830301	-0.069433	0.064426	57

442 rows × 11 columns

Figure 5: Data of the standardized predictors.

(b) The linear regression it computed using the gradient descent. Say the prediction is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b\mathbf{J}_{n,1}$$

where $\mathbf{w} \in \mathbb{R}^d$ are the weights for each predictor, b is the intercept, and $\mathbf{J}_{n,1} \in \mathbb{R}^n$ is a matrix of ones. The cost function that we want to minimize is the difference between the actual data and the prediction which is

$$\mathcal{J} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2.$$

Now to fit the data we use a gradient descent approach in which we simply take the gradients for the components that we want to find. Therefore,

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{w}} &= -\frac{2}{n}\mathbf{X}^\top\left(\mathbf{y} - \mathbf{X}\mathbf{w} - b\mathbf{J}_{n,1}\right) = -\frac{2}{n}\mathbf{X}^\top\left(\mathbf{y} - \hat{\mathbf{y}}\right) \\ \frac{\partial \mathcal{J}}{\partial b} &= -\frac{2}{n}\mathbf{J}_{n,1}^\top\left(\mathbf{y} - \mathbf{X}\mathbf{w} - b\mathbf{J}_{n,1}\right) = -\frac{2}{n}\mathbf{J}_{n,1}^\top(\mathbf{y} - \hat{\mathbf{y}}). \end{aligned}$$

In the code provided separately we use three methods to compute the coefficients and the intercept. In the first method we use the `sklearn` library which has a built-in linear regression function. In the second method, we build a linear regression class from scratch using the details provided above with an iteration of 2000 and learning rate of 0.01. For the third method we use the normal equations for linear regression as follows.

$$\mathbf{x}^\top\mathbf{x}\begin{pmatrix}\mathbf{w}\\ b\end{pmatrix} = \mathbf{x}^\top\mathbf{y}.$$

Now, with the first method we have the following values for the weight, $\mathbf{w} = \{w_1, ..., w_{10}\}$ and intercept, b

$$\begin{gathered} w_1 = -0.4621, \quad w_2 = -11.3733, \quad w_3 = 24.6936, \quad w_4 = 15.4175, \quad w_5 = -37.4332 \\ w_6 = 22.4540, \quad w_7 = 4.7522, \quad w_8 = 8.4306, \quad w_9 = 35.6459, \quad w_{10} = 3.2323 \\ b = 152.1335 \end{gathered}$$

With method 2 we get

$$\begin{gathered} w_1 = -0.3467, \quad w_2 = -11.2503, \quad w_3 = 25.0069, \quad w_4 = 15.3073, \quad w_5 = -11.8601 \\ w_6 = 2.1184, \quad w_7 = -6.6397, \quad w_8 = 5.2573, \quad w_9 = 26.0870, \quad w_{10} = 3.3208 \\ b = 152.1335 \end{gathered}$$

With method 3 we get

$$\begin{gathered} w_1 = -0.4621, \quad w_2 = -11.3733, \quad w_3 = 24.6936, \quad w_4 = 15.4175, \quad w_5 = -37.4332 \\ w_6 = 22.4540, \quad w_7 = 4.7522, \quad w_8 = 8.4306, \quad w_9 = 35.6459, \quad w_{10} = 3.2323 \\ b = 152.1335 \end{gathered}$$

The first and third methods have the exact same weight and intercept, however the second method differs. The cost for the three methods are 2860.3269, 2871.4639, 2860.3269 respectively. The first method using `sklearn` and the third method performs somewhat better then the original python code.

(c) Without the intercept we have the following weight values for each methods. Method 1:

$$\begin{gathered} w_1 = -0.4621, \quad w_2 = -11.3733, \quad w_3 = 24.6936, \quad w_4 = 15.4175, \quad w_5 = -37.4332 \\ w_6 = 22.4540, \quad w_7 = 4.7522, \quad w_8 = 8.4306, \quad w_9 = 35.6459, \quad w_{10} = 3.2323 \end{gathered}$$

Method 2:

$$\begin{gathered} w_1 = -0.3467, \quad w_2 = -11.2503, \quad w_3 = 25.0069, \quad w_4 = 15.3073, \quad w_5 = -11.8601 \\ w_6 = 2.1184, \quad w_7 = -6.6397, \quad w_8 = 5.2573, \quad w_9 = 26.0870, \quad w_{10} = 3.3208 \end{gathered}$$

Method 3:

$$\begin{gathered} w_1 = -0.4621, \quad w_2 = -11.3733, \quad w_3 = 24.6936, \quad w_4 = 15.4175, \quad w_5 = -37.4332 \\ w_6 = 22.4540, \quad w_7 = 4.7522, \quad w_8 = 8.4306, \quad w_9 = 35.6459, \quad w_{10} = 3.2323 \end{gathered}$$

We can observe that the weight values have not changed. For each methods we have a cost of 26004.9239, 26016.0609, and 26004.9239 respectively. We can see that the cost has deteriorated significantly. The intercept was responsible of moving the linear regression from the origin but regardless of the intercept the weights do not change. As you can see from the gradient descent equation in Problem (b), the partial differential of the cost function, $\mathcal{J}$ with respect to $\mathbf{w}$ does not change even if there is no intercept. Similarly, for the third method, if we want the intercept we horizontally concatenate a column of ones

$$\mathbf{x}' = \begin{bmatrix} \cdots & x_1 & \cdots & \vdots \\ & \vdots & & 1 \\ \cdots & x_n & \cdots & \vdots \end{bmatrix}$$

But when we do not have a intercept we do not do this operation and use the original **x** instead. As you can this operation does not effect the outcome of the weights but just adds an additional scalar value to evaluate for the linear regression. Still this intercept is important for a more accurate linear regression.

(d) For here on we will only use the third method for a linear regression. Let the error be

$$E = \frac{1}{n/2} \sum_{i=1}^{n} |\mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}}\mathbf{w} - b\mathbf{J}_{\frac{n}{2},1}|$$

The error for all 100 tests are

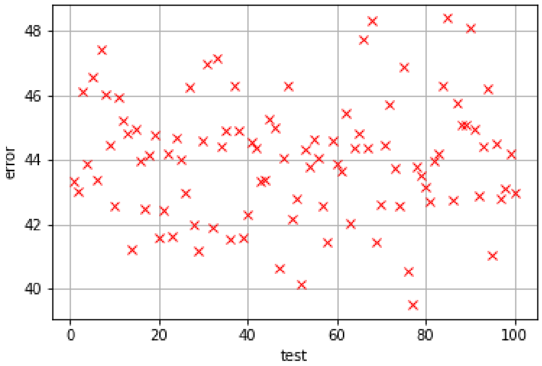


Figure 6: Errors for all 100 tests for problem (d).

and the average error is 44.0183.

(e) If we repeat this for no intercept case we have the following.

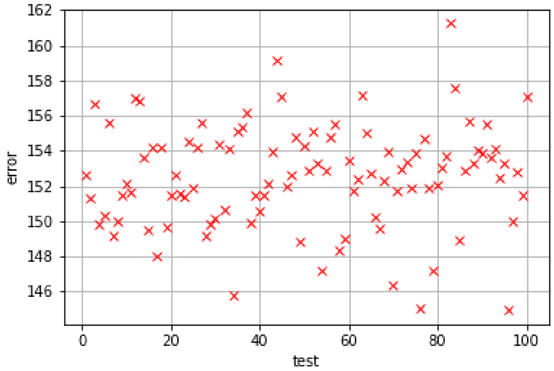


Figure 7: Errors for all 100 tests for problem (e).

and the average error is 152.5323.

(f) Undoubtedly, the linear regression with the intercept provides a better representation of the data set. From the problems (a)-(c) we have seen that the cost without the intercept is much higher than when we do have an intercept. Further, from the last 2 problems we can clearly see a significant rise in the error when we do not include an intercept for the linear regression.

References

[1] *Inner Product Space*. University of Washington. URL: https://sites.math.washington.edu/~morrow/335_17/inner%5C%20products.pdf.

[2] *Norms Induced by Inner Products and the Parallelogram Law*. Math StackExchange. 2011. URL: https://math.stackexchange.com/questions/21792/norms-induced-by-inner-products-and-the-parallelogram-law.

[3] Bruno Félix Rezende Ribeiro. *Proof of Jordan-von Neumann theorem for vector spaces over R*. FAMAT: Universidade Federal de Uberlândia. 2017. URL: https://oitofelix.github.io/academic/Jordan-von%5C%20Neumann%5C%20Theorem.pdf.