Mathematics for Machine Learning
Problem Set 1
COMS 3770: Summer A 2025

Instructor: Samuel Deng
samdeng@cs.columbia.edu
Due: 06/06/2025 11:59 PM ET.

Problem 1

Properties of inner products (20 points total). The *dot product* or *standard Euclidean inner product* is an important operation in linear algebra that takes two vectors and returns a scalar value. Recall that, given two vectors $\mathbf{v} = (v_1, \ldots, v_d)$ and $\mathbf{u} = (u_1, \ldots, u_d)$ in $\mathbb{R}^d$, their dot product is

$$\mathbf{u}^\top \mathbf{v} = u_1 v_1 + \cdots + u_d v_d = \sum_{i=1}^{d} u_i v_i.$$

The dot product is an example of an *inner product*, an operation that takes two vectors and returns a scalar value that obey three important properties. For two vectors $\mathbf{u}, \mathbf{v}$, the inner product is denoted $\langle \mathbf{u}, \mathbf{v} \rangle$. Inner products on $\mathbb{R}^d$ obey three properties:

- *Symmetry.* For all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$.
- *Positive definiteness.* For all $\mathbf{v} \in \mathbb{R}^d$, $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
- *Linearity.* Let $\alpha \in \mathbb{R}$ be a scalar and let $\mathbf{w} \in \mathbb{R}^d$ be another vector. Then:

$$\langle \alpha \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle.$$

Problem 1(a) [3 points]. Consider the dot product on $\mathbb{R}^2$:

$$\langle \mathbf{u}, \mathbf{v} \rangle := \mathbf{u}^\top \mathbf{v} = u_1 v_1 + u_2 v_2.$$

Prove, using the three properties above, that this is indeed an inner product.

Defining an inner product on $\mathbb{R}^d$ imbues $\mathbb{R}^d$ with notions of length and angle, which allows us to do geometry. For example, using the standard dot product, we recover the standard notion of length, the *Euclidean norm*

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{d} u_i^2} = \sqrt{\mathbf{u}^\top \mathbf{u}}.$$

However, the dot product isn't the only possible inner product in $\mathbb{R}^d$.

Problem 1(b) [3 points]. Consider the following inner product on $\mathbb{R}^2$:

$$\langle \mathbf{u}, \mathbf{v} \rangle := u_1 v_1 - (u_1 v_2 + u_2 v_1) + 2 u_2 v_2.$$

Prove, using the three properties above, that this is indeed an inner product.

Any inner product induces a notion of "length," or *norm*. Note from the second property, positive definiteness, that inner products are always nonnegative, so we can always take a square root. For any inner product $\langle \cdot, \cdot \rangle$, the *induced norm* by that inner product is defined as $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$.

Problem 1(c) [2 points]. Consider the vector $\mathbf{u} = (1, 1) \in \mathbb{R}^2$. Compute $\|\mathbf{u}\|_2$, the standard Euclidean norm. Then, using the inner product defined in Problem 1(b), compute the induced norm $\|\mathbf{u}\|$. Compare: is $\mathbf{u}$ "larger" with the standard Euclidean norm or the induced norm from Problem 1(b)?

One can prove properties of the norm just from the three properties of inner products above. Here is one example:

Problem 1(d) [2 points]. Let $\langle \cdot, \cdot \rangle$ be an arbitrary inner product, and let $\| \cdot \|$ be its induced norm. Prove that, for any two vectors $\mathbf{u}, \mathbf{v}$,

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2).$$

Now we will state and prove perhaps two of the most important properties of inner products. The *Cauchy-Schwarz Inequality* states that, for any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|. \tag{1}$$

The *triangle inequality* states that the sum of the lengths of two sides of a triangle is never greater than the third side. That is, for any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|. \tag{2}$$

Both of these properties are true for *any* inner product. The rest of this problem will walk you through a proof of these two properties for the standard Euclidean inner product.

Problem 1(e) [2 points]. Consider the following vectors in $\mathbb{R}^2$:

$$\mathbf{u} = (1, 1) \quad \mathbf{v} = (2, -1).$$

Verify that $\mathbf{u}, \mathbf{v}$ obey the Cauchy-Schwarz inequality with the standard Euclidean inner product (show all your steps). Verify that they also obey the triangle inequality (show all your steps).

There are many proofs of the Cauchy-Schwarz inequality, but we will do a proof by induction. The induction will be on the dimension d.

Problem 1(f) [2 points]. For $u, v \in \mathbb{R}$, the dot product is

$$u^\top v = uv.$$

For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, the dot product is

$$\mathbf{u}^\top \mathbf{v} = u_1 v_1 + u_2 v_2.$$

Prove the base case for the induction when $d = 1$ (i.e., Cauchy-Schwarz for $\mathbb{R}$). This should be trivial, so also prove the base case for $d = 2$:

$$|u_1 v_1 + u_2 v_2| \leq \sqrt{u_1^2 + u_2^2}\sqrt{v_1^2 + v_2^2}.$$

Now, we need to prove our induction step. Assume that Cauchy-Schwarz holds for $\mathbb{R}^d$:

$$|u_1 v_1 + \cdots + u_d v_d| \leq \sqrt{u_1^2 + \cdots + u_d^2}\sqrt{v_1^2 + \cdots + v_d^2}.$$

It suffices to show that it holds for $\mathbb{R}^{d+1}$ as well.

Problem 1(g) [2 points]. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Let $u_{d+1}, v_{d+1} \in \mathbb{R}$ be scalars. Prove the inequality

$$|\|\mathbf{u}\|\|\mathbf{v}\| + u_{d+1} v_{d+1}| \leq \sqrt{\|\mathbf{u}\|^2 + u_{d+1}^2}\sqrt{\|\mathbf{v}\|^2 + v_{d+1}^2}.$$

Hint: It may be helpful to use Problem 1(f).

Problem 1(h) [2 points]. Complete the proof by using the inequality in Problem 1(g) to show that the positive side of the Cauchy-Schwarz inequality holds for $\mathbb{R}^{d+1}$. That is, you need to prove:

$$\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\|\|\mathbf{v}\|.$$

The negative direction is slightly trickier; a valid proof of the full Equation (1) (with the absolute value) will give you **2 points extra credit.**

This completes our proof of the Cauchy-Schwarz inequality for the dot product on $\mathbb{R}^d$. The triangle inequality is a direct consequence of the Cauchy-Schwarz inequality.

Problem 1(i) [2 points]. Prove the triangle inequality, stated above in Equation (2). You may use the Cauchy-Schwarz inequality in Equation (1).

Problem 2

Linear transformations and matrices (26 points total). The property that underlies all of linear algebra is *linearity*. In this problem, we will attempt to understand the relationship between matrices and linear transformations.

Many common functions in the real world are linear. Cooking is one of them. Consider the following example. Suppose that we have $d = 7$ ingredients to make some classic NYC fare: bacon, egg, cheddar cheese, cream cheese, bagel, Kaiser roll, and lox. Consider four recipes we can make with these ingredients, represented by the vectors $\mathbf{r}, \mathbf{c}, \mathbf{b}$ and $\mathbf{l}$. The d ingredients are ordered as above; for example, to make a bacon, egg and cheese on a roll (vector $\mathbf{r}$), we need one unit each of bacon, egg, cheddar cheese, and Kaiser roll, with zero units of the other ingredients.

$$\text{bacon, egg, and cheese on roll:} \quad \mathbf{r} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \text{cream cheese on bagel:} \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\text{bacon, egg, and cheese on bagel:} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{lox sandwich on bagel:} \quad \mathbf{l} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

Suppose the vector $\mathbf{u} = (4, 4, 4, 5, 6, 2, 3)$ describes how much of each ingredient we have in supply today (four units of bacon, four units of egg, etc.).

Problem 2(a) [2 points]. We would like to use as many of our ingredients as possible to make as many of the above recipes as possible. How many of each recipe can we make with zero surplus (or shortfall) of each ingredient? Set up a system of linear equations for this question in matrix-vector form.

Problem 2(b) [2 points]. Does the system of equations in Problem 2(a) have a solution? If so, write down a solution. If not, explain why. Feel free to use `numpy` or any other numerical computing software to help you solve the system.

As we can see from the above example, matrix-vector multiplication has the nice property that, if you add the inputs, you add the outputs (if we wanted twice as many of each recipe,

we'd need exactly twice as many ingredients).

Let $T : \mathbb{R}^d \to \mathbb{R}^n$ be a function (also referred to as a "mapping" or "transformation"). Functions can be arbitrarily complicated; a function need only map inputs in $\mathbb{R}^d$ to outputs in $\mathbb{R}^n$. *Linear transformations* (a.k.a. "linear functions" or "linear maps") are restricted to obey two rules that force them to behave nicely:

$$T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y}) \quad \text{and} \quad T(\alpha\mathbf{x}) = \alpha T(\mathbf{x})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and scalars $\alpha \in \mathbb{R}$.

Problem 2(c) [8 points] Determine whether the following transformations are linear. If a function is linear, give a proof by showing the function satisfies the properties of linearity. If not, state which property of linearity fails and give a specific pair of vectors $\mathbf{x}, \mathbf{y}$ or a scalar α and vector $\mathbf{x}$ for which it fails.

- $T : \mathbb{R} \to \mathbb{R}$ defined $T(x) := 2x - 1$.
- $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined as $T(x_1, x_2) := (x_2, x_1 + x_2)$.
- $T : \mathbb{R}^d \to \mathbb{R}$ defined $T(x) := \frac{1}{d}(x_1 + \cdots + x_d)$.
- $T : \mathbb{R}^d \to \mathbb{R}$ defined $T(x_1, \ldots, x_d) := x_d - x_1$.

Taken as functions, inner products and matrix-vector products are also linear. For a given vector $\mathbf{a} \in \mathbb{R}^d$, let the function $T_\mathbf{a} : \mathbb{R}^d \to \mathbb{R}$ be defined as:

$$T_\mathbf{a}(\mathbf{x}) := \mathbf{a}^\top \mathbf{x}. \tag{3}$$

For a given matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, let the function $T_\mathbf{A} : \mathbb{R}^d \to \mathbb{R}^n$ be defined as:

$$T_\mathbf{A}(\mathbf{x}) := \mathbf{A}\mathbf{x}. \tag{4}$$

Problem 2(d) [4 points] Prove that the function defined by inner products in Equation (3) and the function defined by matrix-vector products in Equation (4) are linear transformations. For Equation (4), you may use any of the equivalent characterizations of matrix-vector multiplication shown in class.

In this way, any matrix defines a linear transformation. This is important — perhaps in your introductory linear algebra class, matrices were introduced as just a way to organize a system of linear equations, like $\mathbf{A}\mathbf{x} = \mathbf{b}$. Equation (4) tells us that we can actually think of a matrix as an object that *does something* to vectors. Given a matrix, matrix-vector multiplication is a linear transformation. Surprisingly, the reverse is true as well: *any* linear transformation has an associated matrix!

Consider the following example. Let $\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0)$, and $\mathbf{e}_3 = (0, 0, 1)$ denote the standard basis vectors in $\mathbb{R}^3$. Let $T : \mathbb{R}^3 \to \mathbb{R}^2$ be the linear transformation defined as:

$$T(x_1, x_2, x_3) := (2x_1, x_2 + x_3).$$

Problem 2(e) [1 point] Where does T map the basis vectors to? That is, compute $T(\mathbf{e}_1)$, $T(\mathbf{e}_2)$, and $T(\mathbf{e}_3)$.

Now, consider the input vector $\mathbf{x} = (3, 2, -1)$. Because $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are a basis for $\mathbb{R}^3$, we can write $\mathbf{x}$ as a linear combination of $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. Using this example, we'll try to "guess" the matrix that corresponds to T.

Problem 2(f) [1 point] Write the matrix $\mathbf{A} \in \mathbb{R}^{2\times 3}$ such that:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

for $\mathbf{x} = (3, 2, -1)$.

Hint: Write $\mathbf{x}$ as a linear combination of $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, i.e.,

$$\mathbf{x} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3, \tag{5}$$

where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ are scalars. Apply $T(\cdot)$ to both sides of Equation (5), and use linearity to get the right-hand side to be a sum of three terms.

Problem 2(f) shows us that $T(\mathbf{x})$ is just a linear combination of $T(\mathbf{e}_1)$, $T(\mathbf{e}_2)$, and $T(\mathbf{e}_3)$. It turns out that, in general, if we are given a linear transformation and want to find its corresponding matrix $\mathbf{A}$, we only need to see what that linear transformation does to the standard basis vectors.

Problem 2(g) [4 points] Prove that any linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ is given by matrix-vector multiplication by a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x},$$

where the ith column of $\mathbf{A}$ is $T(\mathbf{e}_i)$.

Together, Equation (4) and Problem 2(g) give us a central theorem of linear algebra: the equivalence of matrices and linear transformations:

(a) Any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ defines a linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ through matrix-vector multiplication:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

(b) Any linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ is given by matrix-vector multiplication by a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x},$$

where the ith column of $\mathbf{A}$ is $T(\mathbf{e}_i)$.

The claim in (b) is particularly interesting — it tells us that, *any* linear transformation can be pinned down (by a concrete box of $n \times d$ numbers) just by seeing how that transformation acts on the standard basis vectors. Just by imposing the property of linearity on functions, we can treat them as matrices which we can easily write down! This perspective on matrices as linear transformations (and vice versa) is very helpful in understanding many of the definitions and theorems of linear algebra.

One such operation that we've already studied is the *projection* operation. Informally, we compute a projection of a point onto a subspace by seeing where a perpendicular line from the point intersects the subspace. Formally, for the subspace $S \subseteq \mathbb{R}^d$, the projection $\Pi_S(\mathbf{x})$ of $\mathbf{x} \in \mathbb{R}^d$ onto S satisfies:

$$(\mathbf{x} - \Pi_S(\mathbf{x}))^\top \mathbf{u}, \quad \text{for all } \mathbf{u} \in S.$$

The theorem we proved above shows us that we can determine the exact projection matrix if we know what a transformation does to the standard basis vectors.

Problem 2(h) [2 points] Consider the linear transformation in $T : \mathbb{R}^2 \to \mathbb{R}^2$ that takes any point $\mathbf{x} \in \mathbb{R}^2$ and outputs its projection onto the x-axis, i.e. the subspace spanned by the vector $\mathbf{u} = (1, 0)$. Find the matrix $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ that corresponds to this transformation. Find the explicit rule $T(x_1, x_2)$ that corresponds to this transformation.

Hint: What does this transformation do to $\mathbf{e}_1$? What does it do to $\mathbf{e}_2$? It may help to draw a picture.

Problem 2(i) [2 points] Consider the linear transformation in $T : \mathbb{R}^2 \to \mathbb{R}^2$ that takes any point $\mathbf{x} \in \mathbb{R}^2$ and outputs its projection onto the $y = x$ line, i.e. the subspace spanned by the vector $\mathbf{u} = (1, 1)$. Find the matrix $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ that corresponds to this transformation. Find the projection of the vector $\mathbf{x} = (3, -1)$ onto this subspace.

Other properties of matrices also become more intuitive when we conceive of matrices in $\mathbb{R}^{n \times d}$ as linear transformations from $\mathbb{R}^d$ to $\mathbb{R}^n$. For example, one of the concepts we've learned is *rank*, the number of linearly independent columns of a matrix. From (b), the columns of a matrix are exactly where the standard basis vectors "land" after the associated transformation. Therefore, a matrix that is not full-rank transforms the standard basis such that some of them are linearly dependent after the transformation.

Commit the theorem you proved above to memory — it's at the very heart of linear algebra!

Problem 3

Orthonormal bases, projection, and the dot product (18 points total).

In this problem, we will study *orthonormal bases*, why they're "good" bases, and how to interpret them with respect to projections. Throughout this problem, an *ordered basis* refers to the ordered collection $\mathcal{V} := (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Order is important because basis vector $\mathbf{v}_j$ determines coordinate j of a vector in that basis. Recall that a collection of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is an orthonormal basis if:

$$\begin{aligned} \mathbf{v}_i^\top \mathbf{v}_j &= 0 \quad \text{for } i \neq j \\ \|\mathbf{v}_i\| &= 1 \quad \text{for all } i \in [n]. \end{aligned}$$

We can think of bases as different "languages" to describe vectors. Different bases give us different coordinates for the same vector, and subspaces can have many different bases. These coordinates are given by the coefficients of the vector in the linear combination of the basis vectors.

When we write a vector as a list of numbers, we implicitly write it with respect to the standard basis $\mathbf{e}_1 = (1, \ldots, 0), \ldots, \mathbf{e}_n = (0, \ldots, 1)$. For example, consider $\mathbf{x} = (1, -1) \in \mathbb{R}^2$. The standard basis in $\mathbb{R}^2$ is $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$. The coordinates $x_1 = 1$ and $x_2 = -1$ are implicitly the coefficients of this linear combination:

$$\mathbf{x} = 1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

However, given another basis, say, $\mathbf{v}_1 = (1, 1)$ and $\mathbf{v}_2 = (0, -1)$, we can write the same vector as the linear combination:

$$\mathbf{x} = 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

The coefficients of the linear combination are now $\nu_1 = 1$ and $\nu_2 = 2$, so we say that the coordinates of $\mathbf{x}$ in the ordered basis $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2)$ is $(1, 2)$. Drawing a picture may make this clear — the same vector can be expressed two different ways. Try drawing this to have a picture in mind throughout this problem!

In general, let $\mathbf{x} \in \mathbb{R}^n$ be a vector. If we have two ordered bases $\mathcal{U} := (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $\mathcal{V} := (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ of vectors in $\mathbb{R}^n$, the coordinate representation of $\mathbf{x}$ in these two bases are the coefficients of the linear combination of $\mathbf{x}$ in each of the bases. For example, given these bases, we can write $\mathbf{x}$ in two ways:

$$\mathbf{x} = \nu_1 \mathbf{v}_1 + \cdots + \nu_n \mathbf{v}_n \tag{6}$$

$$\mathbf{x} = \mu_1 \mathbf{u}_1 + \cdots + \mu_n \mathbf{u}_n \tag{7}$$

We will write these coordinates as $[\mathbf{x}]_\mathcal{U} = (\mu_1, \ldots, \mu_n)$ and $[\mathbf{x}]_\mathcal{V} = (\nu_1, \ldots, \nu_n)$, respectively.

When we write $\mathbf{x}$ without any brackets, we implicitly refer to $\mathbf{x}$ in the coordinates of the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$:

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n. \tag{8}$$

We can construct a matrix from these ordered bases by arranging them column-wise in $n \times n$ matrices as follows:

$$\mathbf{U} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \mathbf{u}_1 & \cdots & \mathbf{u}_n \\ \downarrow & \cdots & \downarrow \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ \downarrow & \cdots & \downarrow \end{bmatrix} \quad \mathbf{I} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \mathbf{e}_1 & \cdots & \mathbf{e}_n \\ \downarrow & \cdots & \downarrow \end{bmatrix}$$

Written like this, the linear combinations of Equations (6), (7), and (8) can be compactly expressed, respectively, as the matrix-vector multiplications (recall the "linear combination view" of matrix-vector multiplication from lecture):

$$\mathbf{x} = \mathbf{U}[\mathbf{x}]_{\mathcal{U}} \qquad \mathbf{x} = \mathbf{V}[\mathbf{x}]_{\mathcal{V}} \qquad \mathbf{x} = \mathbf{I}\mathbf{x}.$$

Problem 3(a) [4 points] Consider the vector $\mathbf{x} = (1, 2, -1) \in \mathbb{R}^3$. Consider the ordered bases:

$$\mathcal{U} := \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right) \quad \mathcal{V} := \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \right).$$

Find $[\mathbf{x}]_{\mathcal{U}} \in \mathbb{R}^d$ and $[\mathbf{x}]_{\mathcal{V}} \in \mathbb{R}^d$, the coordinate representations of $\mathbf{x}$ in each of these bases. State and prove whether each basis is an orthonormal basis (if not, provide a clear reason why not).

Hint: To find the coordinate representations, set up a system of linear equations in matrix-vector form and take inverses. You may use any numerical computing software, including `numpy` to take inverses and compute your final answers.

It is clear from the above that getting the coordinates of a vector in a different basis is equivalent to setting up a system of linear equations and solving it. Now, we see one reason why orthonormal bases are "nice" to work with. We have learned that if $\mathbf{U} \in \mathbb{R}^{n \times n}$ is a matrix formed by an orthonormal basis, then $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$. Therefore, finding the coordinate representation of a vector $[\mathbf{x}]_{\mathcal{U}}$ just involves taking a transpose and doing a matrix-vector multiplication:

$$[\mathbf{x}]_{\mathcal{U}} = \mathbf{U}^\top\mathbf{x}.$$

Transposes are much easier to compute than inverses. In general, if we had a basis that was not orthonormal, we may have needed to compute an inverse instead.

Now, suppose that we have a vector $\mathbf{x} \in \mathbb{R}^n$ and we want to compute its projection $\Pi_S(\mathbf{x}) \in \mathbb{R}^n$ onto a lower dimensional subspace $S \subseteq \mathbb{R}^n$, with dimension $d \leq n$. Recall that, informally, we can think of a projection as the "closest vector to $\mathbf{x}$ in the subspace":

$$\mathbf{x} \approx \Pi_S(\mathbf{x}).$$

Suppose that we know an ordered basis $\mathcal{U} := (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ of vectors in $\mathbb{R}^n$ for the subspace S. Arranging them in a matrix still gives us an inverse when multiplying on the left.

Problem 3(b) [2 points] Let $S \subseteq \mathbb{R}^3$ be a 2-dimensional subspace of $\mathbb{R}^3$. Suppose $\mathcal{U} := (\mathbf{u}_1, \mathbf{u}_2)$ is an ordered basis for S where $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthonormal, and let $\mathbf{U} \in \mathbb{R}^{3\times 2}$ be the matrix with these basis vectors as its columns. Let $\mathbf{I}_{d\times d}$ be the $d \times d$ identity matrix. Prove that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_{2\times 2}$. Show that $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_{3\times 3}$ is not necessarily true by finding an orthonormal basis that violates this equality.

From lecture, we used least squares to derive the *projection matrix* for subspaces for which we know a basis.

Problem 3(c) [4 points] Let $\mathbf{x} = (1, 2, 3, 4) \in \mathbb{R}^4$. Consider the 3-dimensional subspace $S \subseteq \mathbb{R}^4$ spanned by the basis

$$\mathcal{U} := \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right).$$

Find the projection $\Pi_S(\mathbf{x})$. Find the *projection matrix* for S, the matrix P_S such that $\Pi_S(\mathbf{x}) = P_S\mathbf{x}$.

We'll focus now on the case where we are projecting onto a one-dimensional subspace, or a line. Given a vector $\mathbf{u} \in \mathbb{R}^n$, any one-dimensional subspace can be written as

$$S_\mathbf{u} = \{\alpha\mathbf{u} : \alpha \in \mathbb{R}\},$$

all the scalar multiples of $\mathbf{u}$. Because we can determine such subspaces with just a single vector, we will write $\Pi_\mathbf{u}(\mathbf{x})$ to denote the projection of $\mathbf{x}$ onto the subspace $S_\mathbf{u}$ spanned by $\mathbf{u}$. It turns out that projection is a way to geometrically interpret the dot product.

When one first learns the dot product, it seems like a strictly algebraic operation:

$$\mathbf{u}^\top \mathbf{v} = u_1 v_1 + \cdots + u_n v_n := \sum_{i=1}^{n} u_i v_i.$$

We can view this as a matrix-vector multiplication by the matrix $\mathbf{u} \in \mathbb{R}^{1\times n}$ (with one row and n columns) and the vector $\mathbf{v} \in \mathbb{R}^n$. This motivates the usual notation for the dot product with the transpose operator.

It is usually also defined in terms of the angle θ between vectors:

$$\mathbf{u}^\top \mathbf{v} := \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta.$$

These two definitions are equivalent (we will not prove that here, but one can refer to any standard linear algebra textbook for the proof). But, at first glance, the relationship between these two equivalent definitions is mysterious. Why would the element-wise sum of products of entries between two vectors have anything to do with angle?

Let $\mathbf{u} \in \mathbb{R}^n$ be a vector. Immediately, we can consider the one-dimensional subspace $S_\mathbf{u}$ spanned this vector. This single vector is clearly a basis for this subspace, and normalizing it gives us an orthonormal basis $\mathcal{U} = \left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)$. By normalizing $\mathbf{u}$, we focus only on the "direction" that $\mathbf{u}$ defines, which is all that matters for a line.

Problem 3(d) [2 points] Prove that the subspace $S_\mathbf{u}$ has the projection matrix $P_\mathbf{u} = \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|^2}$. That is, $\Pi_\mathbf{u}(\mathbf{x}) = P_\mathbf{u}\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$.

Hint: Use the projection matrix formula derived in lecture from least squares. What is the "orthogonal matrix" in this situation?

From Problem 3(d), we can see that the projection $\Pi_\mathbf{u}(\mathbf{x})$ of any $\mathbf{x} \in \mathbb{R}^n$ onto $S_\mathbf{u}$ can be computed as:

$$\Pi_\mathbf{u}(\mathbf{x}) = P_\mathbf{u}\mathbf{x} = \frac{\mathbf{u}\mathbf{u}^\top\mathbf{x}}{\|\mathbf{u}\|^2} = \underbrace{\left(\frac{\mathbf{u}^\top\mathbf{x}}{\|\mathbf{u}\|}\right)}_{\text{coefficient}}\underbrace{\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)}_{\text{direction}}.$$

We see from the above that the (normalized with $\|\mathbf{u}\|$) dot product $\frac{\mathbf{u}^\top\mathbf{x}}{\|\mathbf{u}\|}$ is how many units in the direction $\frac{\mathbf{u}}{\|\mathbf{u}\|}$ the projected vector $\Pi_\mathbf{u}(\mathbf{x})$ is.

Now, we'll relate this back to the geometric definition of the dot product. For simplicity, first assume that $\mathbf{u}$ and $\mathbf{x}$ form an acute angle, with $0 \leq \theta \leq 90°$:

$$\mathbf{u}^\top\mathbf{x} = \|\mathbf{u}\|\|\mathbf{x}\|\cos\theta \implies \frac{\mathbf{u}^\top\mathbf{x}}{\|\mathbf{u}\|} = \|\mathbf{x}\|\cos\theta. \tag{9}$$

What is this quantity $\|\mathbf{x}\|\cos\theta$? One way to find $\|\mathbf{x}\|\cos\theta$ involves analyzing $\|\Pi_\mathbf{u}(\mathbf{x})\|$ (do this yourself!), but this doesn't give much intuition. We can get more intuition through some basic trigonometry.

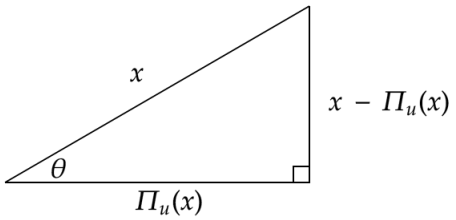
Problem 3(e) [2 points] Draw a right triangle with first leg $\mathbf{x} - \Pi_\mathbf{u}(\mathbf{x})$, second leg $\Pi_\mathbf{u}(\mathbf{x})$, and hypotenuse $\mathbf{x}$ (include your drawing in your submission). Prove, using trigonometry, that $\|\mathbf{x}\|\cos\theta = \|\Pi_\mathbf{u}(\mathbf{x})\|$.

Hint: The right triangle should come immediately from the "arrow view" of vector addition. Use the identity from basic trigonometry: $\cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$.

Finally, plugging this back into the geometric definition of the dot product in Equation (9),

$$\mathbf{u}^\top\mathbf{x} = \|\Pi_\mathbf{u}(\mathbf{x})\|\|\mathbf{u}\| \quad \text{when } 0 \leq \theta \leq 90°. \tag{10}$$

Figure 1: Triangle for Problem 3(e)



When $90° < \theta \leq 180°$, we form the right triangle with $\theta' = 180° - \theta$. Using the trigonometric identity $\cos(180° - \theta) = -\cos\theta$, we obtain:

$$\cos\theta' = \cos(180° - \theta) = -\cos\theta = \frac{\|\Pi_{\mathbf{u}}(\mathbf{x})\|}{\|\mathbf{x}\|} \implies -\|\mathbf{x}\|\cos\theta = \|\Pi_{\mathbf{u}}(\mathbf{x})\|,$$

and, by the same argument:

$$\mathbf{u}^\top\mathbf{x} = -\|\Pi_{\mathbf{u}}(\mathbf{x})\|\|\mathbf{u}\| \quad \text{when } 90° < \theta \leq 180°. \tag{11}$$

Together, Equations (10) and (11) show us that the dot product between two vectors is equivalent to projecting one of the vectors onto the other and measuring the length of that projection! Recall that we can think of a projection as the "shadow" of a vector on a subspace. In an informal sense, the dot product is the length of the "shadow" one vector casts on another. Notice that, in the above arguments, we could have also switched the places of $\mathbf{u}$ and $\mathbf{x}$ — it doesn't matter which vector we're projecting and which we're using to form the subspace.

Problem 3(f) [4 points] Let $\mathbf{u} = (2, 1) \in \mathbb{R}^2$ be a vector and consider

$$\mathbf{x}_1 = \begin{bmatrix}2\\1\end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix}1\\1\end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix}1\\-2\end{bmatrix} \quad \mathbf{x}_4 = \begin{bmatrix}-1\\2\end{bmatrix}.$$

Compute the dot products $\mathbf{u}^\top\mathbf{x}_1$, $\mathbf{u}^\top\mathbf{x}_2$, $\mathbf{u}^\top\mathbf{x}_3$ and $\mathbf{u}^\top\mathbf{x}_4$. Compute $\Pi_{\mathbf{u}}(\mathbf{x}_i)$ and $\|\Pi_{\mathbf{u}}(\mathbf{x}_i)\|$ for each of $\mathbf{x}_1, \ldots, \mathbf{x}_4$. You don't need to submit this, but drawing a picture for each of these should help your geometric intuition.

Interpreting the dot product in this way also clarifies why the dot product is so often thought of as a measure of "similarity" between vectors. The more colinear one vector is with another, the larger the "shadow" it casts. When a vector is orthogonal to another, it doesn't cast any shadow — this tracks with the definition of orthogonality: $\mathbf{u}^\top\mathbf{v} = 0$.

Problem 4

Errors in least-squares regression (18 points total).

One of the central ideas in this course is *least-squares regression.* Recall the setup from lecture. We are given n training samples with d features $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ and a vector of training labels $\mathbf{y} \in \mathbb{R}^n$. Arranged row-wise, the training samples form a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d$. For each $i \in [n]$, our goal is to make a prediction $\hat{y}_i \in \mathbb{R}$, such that $\hat{y}_i$ is as close to y_i as possible. In order to make these predictions, we need to construct a weight vector $\mathbf{w} \in \mathbb{R}^d$ such that $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i$. In matrix-vector form, we want to find $\mathbf{w} \in \mathbb{R}^d$ such that

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}.$$

The "least-squares" part of "least-squares regression" comes from how we model the approximation, "$\approx$." We want to find the $\mathbf{w} \in \mathbb{R}^d$ that minimizes a specific notion of error, the *sum of squared residuals* (also known as *mean squared error*), which we'll denote with $\text{err}(\cdot)$:

$$\text{err}(\mathbf{w}) := \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

In this problem, we will investigate the limitations of this model by examining what happens when we change how $\mathbf{x}_i$ and y_i are related.

As a warm-up, let's make sure we understand what each object in the least squares derivation is. Suppose we have the following (small) set of data from some local basketball league.

	height (in)	**workouts per week**	**score**
Aaron	80	5	27.2
Bob	72	4	20.5
Charlie	68	2	15
David	74	4	18.1
Evan	68	5	22.8

In this case, we have $n = 5$ data samples total, and we have $d = 2$ features for every sample: the number of workouts the player does each week and the height of the player in inches. The label, y, for each player is their average score in the season.

Problem 4(a) [4 points] For the basketball data above, (i) construct the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and (ii) the vector $\mathbf{y} \in \mathbb{R}^n$. Then, (iii) find $\mathbf{w} \in \mathbb{R}^d$ through least squares regression by using the ordinary least squares solution from lecture. Finally, (iv) compute the sum of squared residuals error of your solution, $\text{err}(\mathbf{w})$. You may use `numpy` or any other scientific computing utility to perform the matrix multiplications.

Now, let us consider an ideal general scenario. Suppose that, for every $i \in [n]$, there exists some $\mathbf{w}^* \in \mathbb{R}^d$ such that

$$y_i = (\mathbf{w}^*)^\top \mathbf{x}_i. \tag{12}$$

In this case, there is a perfect linear relationship between $\mathbf{x}_i$ and y_i. For example, if $d = 1$ and $w^* = 2$, then $y_i = 2x_i$. The labels just come from a relationship that looks like a line with slope 2 passing through the origin.

Problem 4(b) [2 points] Let Equation (12) apply to all samples $i = 1, \ldots, n$, let $n \geq d$, and let the columns of $\mathbf{X}$ be linearly independent. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the ordinary least squares solution. Prove that $\hat{\mathbf{w}} = \mathbf{w}^*$ and $\text{err}(\hat{\mathbf{w}}) = 0$.

In most real-world problems, there is not a perfect linear relationship between the labels and the training features, however. One way of modeling such relationships is by positing that each sample has some error unexplained by the linear relationship, $\epsilon_i \in \mathbb{R}$. In this case, there exists some $\mathbf{w}^* \in \mathbb{R}^d$, but the labels are now:

$$y_i = (\mathbf{w}^*)^\top \mathbf{x}_i + \epsilon_i. \tag{13}$$

We can collect all these errors into a vector, $\bar{\epsilon} \in \mathbb{R}^n$.

Problem 4(c) [4 points] Let Equation (13) apply to all samples $i = 1, \ldots, n$, let $n \geq d$, and let the columns of $\mathbf{X}$ be linearly independent. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the ordinary least squares solution. Supposing that $(\mathbf{X}^\top \mathbf{X}) = 2\mathbf{I}_{d \times d}$, where $\mathbf{I}_{d \times d}$ is the $d \times d$ identity matrix, prove that:

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2 = \|\bar{\epsilon}\|^2 - \frac{1}{2}\|\mathbf{X}^\top \bar{\epsilon}\|^2.$$

Hint: Expand out $\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2$ using properties of dot products. Using the property $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$ may help.

From Problem 4(c), we see that, if we don't know anything else about the errors, ϵ_i, there's not much else we can say about the optimality of $\hat{\mathbf{w}}$. We will elaborate more on Equation (13) later in the probability section of this course.

Finally, we will consider an example where the true relationship between y_i and $\mathbf{x}_i$ is nonlinear in the original features but linear in some new features we will engineer. Consider the following dataset with $d = 2$ and $n = 5$, already arranged in a data matrix and label vector:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ -1 & 2 \\ 0 & 1 \\ -2 & 2 \\ 0 & -1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 7 \\ -7 \\ -1 \\ -8 \\ -1 \end{bmatrix}. \tag{14}$$

Problem 4(d) makes clear that the true relationship is certainly not linear as in Equation (12) (if it were, the error would be 0).

Problem 4(d) [2 points] For the data in (14), find $\mathbf{w} \in \mathbb{R}^d$ through least squares regression by using the ordinary least squares solution from lecture. Also compute the sum of squared residuals error of your solution, $\text{err}(\mathbf{w})$. You may use `numpy` or any other scientific computing utility to perform the matrix multiplications.

Now, consider the following nonlinear function, $\phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$\phi(x_1, x_2) = (x_1^2, x_1 x_2, x_2^2).$$

Because $\phi(\cdot, \cdot)$ takes inputs in $\mathbb{R}^2$, we can feed it each row (sample) in our data matrix. This allows us to "transform" our data matrix to a new data matrix, $\mathbf{X}' \in \mathbb{R}^{5 \times 3}$ by applying $\phi(\cdot, \cdot)$ row by row. By doing so, we are constructing 3 new features from the $d = 2$ old features.

Problem 4(e) [4 points] Find the transformed data matrix $\mathbf{X}' \in \mathbb{R}^{5 \times 3}$ obtained by applying $\phi(\cdot, \cdot)$ to each of the 5 rows. Find $\mathbf{w} \in \mathbb{R}^d$ by least squares regression on $\mathbf{X}'$ and the original $\mathbf{y}$. Also compute the sum of squared residuals error of your solution, $\text{err}(\mathbf{w})$ (you should find that, now, $\text{err}(\mathbf{w}) = 0$). You may use `numpy` or any other scientific computing utility to perform the matrix multiplications.

It turns out that the true relationship between y_i and $\mathbf{x}_i = (x_{i1}, x_{i2})$ for the data in (14) is actually:

$$y_i = x_{i1}^2 + 2x_{i1}x_{i2} - x_{i2}^2 \quad \text{for all } i \in [n]. \tag{15}$$

By finding the feature transformation $\phi(\cdot, \cdot)$ above, we turned a problem with a nonlinear relationship into a problem where a linear model is again useful (and, in fact, perfectly fits $\mathbf{X}'$). We are back in our ideal scenario in Equation (12), but there now exists some $\mathbf{w}^* \in \mathbb{R}^d$ such that

$$y_i = (\mathbf{w}^*)^\top \phi(\mathbf{x}_i).$$

Problem 4(f) [2 points] What is $\mathbf{w}^* \in \mathbb{R}^d$ for the relationship in Equation (15) that generated our data in (14)?

Finding useful functions $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ that transform vectors to a higher dimensional space d' such that we can eventually use a linear model atop the transformed vectors is the focus of a great deal of machine learning. For instance, one can take the simplified view that the architecture of many neural networks finds some extremely complex $\phi(\cdot)$ which we can eventually apply a linear model on to make predictions.

Programming Part

Basics of linear regression in Python (18 points total). In this problem, you will familiarize yourself with the basics of running linear regression in Python on three simple examples of increasing dimensionality.

In order to start this programming part, download the file `ps1.ipynb` from Course Content on the course webpage. Your submission for this part will be the same `ps1.ipynb` file modified with your code; see HW Submission on the course webpage for additional instructions.