CS181 Section 0 - Solutions

Due: Never

The goal of these section notes is to cover some material that is mostly review for CS 181. There are a number of problems to test your understanding and readiness for the course. (*) indicates challenge sections or challenge problems. Do not worry if you cannot solve these problems as the corresponding material will not be necessary as prerequisites.

1 Linear Algebra

A great reference for this material is Sheldon Axler's *Linear Algebra Done Right*, which can be found on *Hollis*.

1.1 Scalars and Vectors

A **scalar** is a single element of the real numbers. $a \in \mathbb{R}$ is a scalar. We usually denote scalars using lowercase letters, such as a or x.

A **vector** of n dimensions is an ordered collection of n coordinates, where each coordinate is a scalar. An n-dimensional vector $\mathbf{v}$ with real coordinates is an element of $\mathbb{R}^n$. Equivalently, the coordinates specify as single point in an n-dimensional space, just like you may have seen with cartesian coordiates where $(1, 3)$ might denote a point. By default, vectors will be columns and their transposes will be rows. We write vectors in bold lowercase, and the vector itself as a column of scalars:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T.$$

This is the default format. Sometimes vectors will be in row form and their symbols may not be bolded. If you find this confusing at first please reach out to one of the course staff.
Vectors may be scaled. $a\mathbf{x}$ scales each element of $\mathbf{x}$ by scalar a so that

$$a\mathbf{x} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_n \end{bmatrix}.$$

Vectors of the same dimension may be added coordinate-wise:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

Vectors have both a **direction** and a **magnitude**. The magnitude of a vector (or its length) is typically the vector's $\mathbf{L_2}$ norm, which can be computed as the square root of the sum of the squares of the coordinates:

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

There are a number of other vector norms such as the $\mathbf{L_1}, \mathbf{L_p}, \mathbf{L}_\infty$ norms:

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|,$$

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

$$||\mathbf{x}||_\infty = \max_{i=1,\ldots,n} |x_i|.$$

Definition 1.1 (Norm)**.** We say that $\|\cdot\|$ is a norm if it satisfies the following properties:

- Triangle inequality: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.
- $\|a\mathbf{x}\| = |a| \cdot \|\mathbf{x}\|$ for a scalar a.
- $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Problem 1

(*) *Challenge*: Show that the $\mathbf{L_p}$ norms are indeed norms for $p \in [1, \infty)$ and $p = \infty$. We will mostly work with L_1 and L_2 so it is recommended you understand these two norms.

Solution: We prove by cases.

Case 1: We assume $p \in [1, \infty)$. Then $||\mathbf{x}||_p$ satisfies the properties of a norm:

- Minkowski's inequality proves that: $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$.
- $\|a\mathbf{x}\|_p = \left(\sum_{i=1}^n |ax_i|^p\right)^{1/p} = \left(\sum_{i=1}^n |a|^p|x_i|^p\right)^{1/p} = \left(|a|^p \sum_{i=1}^n |x_i|^p\right)^{1/p} = |a| \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = |a| \cdot \|\mathbf{x}\|_p$
- Assume $\|\mathbf{x}\|_p = 0$. Then $\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = 0 \implies \sum_{i=1}^n |x_i|^p = 0$. Since every term in the sum is non-negative due to the absolute value, x_i must equal 0 for all i. $\implies \mathbf{x} = \mathbf{0}$.

  Now assume $\mathbf{x} = \mathbf{0}$. Then $\|\mathbf{0}\|_p = \left(\sum_{i=1}^n |0|^p\right)^{1/p} = 0$

Case 2: We assume $p = \infty$. Then $||\mathbf{x}||_\infty$ satisfies the properties of a norm:

- $$\|\mathbf{x} + \mathbf{y}\|_\infty = \max_{i=1,\ldots,n} |x_i + y_i| \leq \max_{i=1,\ldots,n} |x_i| + |y_i| \leq \max_{i=1,\ldots,n} |x_i| + \max_{i=1,\ldots,n} |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty$$
- $$\|a\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |ax_i| = \max_{i=1,\ldots,n} |a||x_i| = |a| \max_{i=1,\ldots,n} |x_i| = |a| \cdot \|\mathbf{x}\|_\infty$$
- Assume $\|\mathbf{x}\|_\infty = 0$. Then $\max_{i=1,\ldots,n} |x_i| = 0$. Suppose for the sake of contradiction that $\mathbf{x} \neq \mathbf{0}$. Then that means either all the terms $|x_i|$ are negative, or one of the x_i's is greater than 0. Both are contradictions since the first case contradicts $|x_i|$ being non-negative, and the second is a contradiction because that would mean $\max_{i=1,\ldots,n} |x_i| \neq 0$ So by proof of contradiction, $\mathbf{x} = \mathbf{0}$.

  Now assume $\mathbf{x} = \mathbf{0}$. Then $\|\mathbf{0}\|_\infty = \max_{i=1,\ldots,n} |0| = 0$.

□

The direction of a vector can be represented using a vector of magnitude one (according to some norm):

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \begin{bmatrix} x_1/\|x\| \\ x_2/\|x\| \\ \vdots \\ x_n/\|x\| \end{bmatrix}.$$

We often use the "hat" symbol (i.e. $\hat{\mathbf{x}}$) to denote that a vector has magnitude one, or is a unit vector.
An important product between vectors of the same dimension is the **inner product** (also called dot product or scalar product). For two vectors $\mathbf{u}$ and $\mathbf{v}$, this is defined as

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^n u_i v_i.$$

It is also written as $\langle \mathbf{u}, \mathbf{v} \rangle$. We can introduce **cosine similarity** through the formula

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||_2 \, ||\mathbf{v}||_2},$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. The cosine similarity ranges from -1 (exactly opposite) to 1 (exactly the same), with 0 indicating orthogonal vectors. If $\mathbf{v}$ is a unit vector then $\mathbf{u} \cdot \mathbf{v}$ gives us the magnitude of the projection of $\mathbf{u}$ onto the direction of $\mathbf{v}$. Thus it makes sense that a vector $\mathbf{u}$ dotted with itself equals the square of its L2 norm: $\langle \mathbf{u}, \mathbf{u} \rangle = ||\mathbf{u}||_2^2$.

The **outer product** between two vectors is the matrix $\mathbf{W} = [w_{ij}]_{i,j \leq n}$ whose entries are $w_{ij} = u_i v_j$. When the two vectors are dimension n and m, respectively, their outer product is an $n \times m$ matrix.

1.2 Linear Independence

A set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is **linearly independent** if and only if the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 +, ..., +c_n\mathbf{v}_n = \mathbf{0}$ for scalars $c_1, ..., c_n$ can only be satisfied by setting $c_1, ..., c_n$ all to 0. Intuitively, it means that none of the vectors (or linear combinations of them) are parallel.

1.3 Spaces and Subspaces

A **vector space** $\mathcal{V}$ is a collection of vectors that follow several axioms regarding the properties of scaling and addition described above, and most importantly:

- $\mathbf{0} \in \mathcal{V}$
- closure under scaling: $\forall\ \mathbf{v} \in \mathcal{V}$ and scalars $a \in \mathbb{R}$, $a\mathbf{v} \in \mathcal{V}$
- closure under addition: $\forall\ \mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{u} + \mathbf{v} \in \mathcal{V}$

The most intuitive vector space and the one most relevant to the course is $\mathbb{R}^n$, the space of n-dimensional vectors. $\mathbb{R}^2$ is the 2-dimensional Cartesian plane for example.

Now we define a **basis** for a vector space. First, we define a **linear combination** of a list of vectors $(v_1, \ldots, v_m)$ as any quantity of the form:

$$a_1 v_1 + \ldots + a_m v_m \ \text{ where } a_1, \ldots, a_m \in \mathbb{R} \tag{1}$$

The **span** of $(v_1, \ldots, v_m)$ is the set of all linear combinations of $(v_1, \ldots, v_m)$. Moreover, if the span of $(v_1, \ldots, v_m)$ is equal to the vector space V, then we say that $(v_1, \ldots, v_m)$ **spans** V.

Then a **basis** of a vector space V is a list of vectors in V that both are linearly independent and also span V. For the space $\mathbb{R}^n$, the most intuitive basis, which we call the **standard basis** is the list:

$$((1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)) \tag{2}$$

The set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ form an **orthonormal basis** for $\mathcal{V}$ if they are all unit vectors (normal) and if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \forall\ i \neq j$ (orthogonal) where $\langle \cdot, \cdot \rangle$ is the inner product. The standard basis that we defined above is also an orthonormal basis.

The **dimension** of a vector space V is the number of vectors of any basis of V. Since every basis of V has the same number of vectors, this is uniquely defined.

Let $\mathcal{S}$ be a vector space. If $\mathcal{S} \subseteq \mathcal{V}$, then $\mathcal{S}$ is a **subspace** of $\mathcal{V}$. Intuitively, a subspace is a lower-dimensional space in a higher-dimensional space—think about the plane defined by the x and y axis in a 3-dimensional x, y and z space.

1.4 Scalar, Vector, and Subspace Projection

For vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and $\mathbf{v} \neq \mathbf{0}$, the **scalar projection** a of $\mathbf{u}$ onto $\mathbf{v}$ is computed as:

$$a = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||}$$

Think about this as the size of $\mathbf{u}$ along the direction of $\mathbf{v}$. Using scalar projection a, the **vector projection** $\mathbf{u}^{\parallel}$ of $\mathbf{u}$ onto $\mathbf{v}$ can be computed as:

$$\mathbf{u}^{\parallel} = a \cdot \frac{\mathbf{v}}{||\mathbf{v}||} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v}.$$

Think about this as scaling by a the unit vector in the direction of $\mathbf{v}$. For a projection onto $\mathbf{v}$, we can then write $\mathbf{u} = \mathbf{u}^{\parallel} + \mathbf{u}^{\perp}$, completing $\mathbf{u}$ with this new component $\mathbf{u}^{\perp}$. In particular, $\langle \mathbf{u}^{\parallel}, \mathbf{u}^{\perp} \rangle = 0$, and $\mathbf{u}^{\perp}$ is orthogonal to $\mathbf{v}$. It follows that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$.

Problem 2

Verify that $\langle \mathbf{u}^{\parallel}, \mathbf{u}^{\perp} \rangle = 0$ and that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$.

Solution: Notice that: $\mathbf{u}^{\perp} = \mathbf{u} - \mathbf{u}^{\parallel}$ and let $c = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}$. By construction,

$$c = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \implies c\langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle \implies \langle \mathbf{u}, \mathbf{v} \rangle - \langle c\mathbf{v}, \mathbf{v} \rangle = 0 \implies \langle \mathbf{u} - c\mathbf{v}, \mathbf{v} \rangle = 0.$$

Then:

$$\langle \mathbf{u}^{\parallel}, \mathbf{u}^{\perp} \rangle = \langle \mathbf{u}^{\parallel}, \mathbf{u} - \mathbf{u}^{\parallel} \rangle = \langle c\mathbf{v}, u - c\mathbf{v} \rangle = \langle \mathbf{u}, c\mathbf{v} \rangle - \langle c\mathbf{v}, c\mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle - c\langle c\mathbf{v}, \mathbf{v} \rangle = c\langle \mathbf{u} - c\mathbf{v}, \mathbf{v} \rangle = c * 0 = 0.$$

□

Assume $\mathbf{u} = \mathbf{u}^{\parallel}$. Then $\mathbf{u} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} = c\mathbf{v}$. So $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$. Next assume that $\mathbf{u}$ is a scaled multiple of $\mathbf{v}$. Then: $\mathbf{u}^{\parallel} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} = \frac{\langle c\mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} = \frac{c\langle \mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} = c\mathbf{v} = \mathbf{u}$.

□

1.4.1 Subspace Projections

Finally, it is possible to project a vector $\mathbf{u}$ in a vector space $\mathcal{V}$ onto a subspace $\mathcal{S}$ of $\mathcal{V}$. If the set of vectors $\{\mathbf{s}_1, ..., \mathbf{s}_m\}$ form an orthonormal basis for $\mathcal{S}$, then the **subspace projection** $\mathbf{u}^{\parallel}$ of $\mathbf{u}$ onto $\mathcal{S} = \text{span}(\mathbf{s}_1, ..., \mathbf{s}_m)$ can be expressed as the sum of the projections of $\mathbf{u}$ onto each element of the basis of $\mathcal{S}$:

$$\mathbf{u}^{\parallel} = \sum_{i=1}^{m} \frac{\langle \mathbf{u}, \mathbf{s}_i \rangle}{\langle \mathbf{s}_i, \mathbf{s}_i \rangle} \mathbf{s}_i$$

This has the properties that the vector $\mathbf{u}^{\perp} = \mathbf{u} - \mathbf{u}^{\parallel}$ is orthogonal to all vectors in $\mathcal{S}$, that $\mathbf{u} = \mathbf{u}^{\parallel}$ if and only if $\mathbf{u} \in \mathcal{S}$, and that $\mathbf{u}^{\parallel}$ is the closest vector in $\mathcal{S}$ to $\mathbf{u}$: $||\mathbf{u} - \mathbf{v}|| > ||\mathbf{u} - \mathbf{u}^{\parallel}||, \forall\ \mathbf{v} \neq \mathbf{u}^{\parallel}, \mathbf{v} \in \mathcal{S}$.

Problem 3 (Distance between a hyperplane and a point)

(*) *Challenge:* Suppose we have a hyperplane defined by $\mathbf{w}^T\mathbf{x} + w_0 = 0$. In this problem, we will derive the formula for the distance between the hyperplane and a point $\mathbf{x}'$.

(a) Imagine two points $\mathbf{x_1}$ and $\mathbf{x_2}$ on this hyperplane. Show that $\mathbf{w}$ is orthogonal to the difference $\mathbf{x_1} - \mathbf{x_2}$. Why does this imply that $\mathbf{w}$ is orthogonal to the hyperplane?

(b) Now, suppose we wish to find the distance d between a point $\mathbf{x}'$ and the our hyperplane. Let $\mathbf{x}_p$ be the projection of $\mathbf{x}'$ onto the hyperplane. Find an expression for $\mathbf{x}'$ in terms of d, $\mathbf{w}$, and w_0. (*Hint*: use the fact from (a) that $\mathbf{w}$ is perpendicular to the hyperplane.)

(c) Using your expression from (b), show that the distance d is the following:

$$d = \frac{\mathbf{w}^T\mathbf{x}' + w_0}{\|\mathbf{w}\|_2} \tag{3}$$

Solution:

(a) First notice that since $\mathbf{x_1}$ and $\mathbf{x_2}$ is on this hyperplane, by definition:

$$\mathbf{w}^T\mathbf{x_1} + w_0 = 0, \mathbf{w}^T\mathbf{x_2} + w_0 = 0 \implies \mathbf{w}^T\mathbf{x_1} = -w_0 = \mathbf{w}^T\mathbf{x_2}$$

Now we look at the dot product:

$$\langle \mathbf{w}, \mathbf{x_1} - \mathbf{x_2} \rangle = \mathbf{w}^T(\mathbf{x_1} - \mathbf{x_2}) = \mathbf{w}^T\mathbf{x_1} - \mathbf{w}^T\mathbf{x_2} = 0.$$

Therefore $\mathbf{w}$ is orthogonal to the difference $\mathbf{x_1} - \mathbf{x_2}$.

□

$\mathbf{x_1} - \mathbf{x_2}$ is a vector in the hyperplane. And since $\mathbf{x_1}, \mathbf{x_2}$ are arbitrary and we showed $\mathbf{w}$ is orthogonal to $\mathbf{x_1}, \mathbf{x_2}$, this implies that $\mathbf{w}$ is orthogonal to the hyperplane.

(b) Since $\mathbf{x}_p$ is the projection of the point onto the hyperplane, we can decompose the point $\mathbf{x}'$ as $\mathbf{x}' = \mathbf{x}_p + d\frac{\mathbf{w}}{\|\mathbf{w}\|}$. Intuitively, this decomposition is the projection of $\mathbf{x}'$ plus the distance d in the orthogonal direction of $\mathbf{w}$ and normalized.

(c) If we multiple by $\mathbf{w}^T$ on both sides of our decomposition of $\mathbf{x}'$, we obtain:

$$\mathbf{w}^T\mathbf{x}_p + d\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} = \mathbf{w}^T\mathbf{x}' \implies d = \frac{\mathbf{w}^T\mathbf{x}' + w_0}{\|\mathbf{w}\|}$$

□

1.5 Matrices

A **matrix** is a rectangular array of scalars. Primarily, an $n \times m$ matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is used to describe a **linear transformation** from m to n dimensions, where the matrix is an **operator**. To see this, note that the result of multiplying an $n \times m$ matrix and an $m \times 1$ vector is an $n \times 1$ vector. A_{ij} is the scalar found at the i^{th} row and j^{th} column. We write matrices in bold uppercase.

A typical linear transformation looks like $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times m}$. The transformation $\mathbf{A}$ is linear because $\mathbf{A}(\lambda_1\mathbf{u} + \lambda_2\mathbf{v}) = \lambda_1\mathbf{A}\mathbf{u} + \lambda_2\mathbf{A}\mathbf{v}$ for scalars λ_1 and λ_2.

1.6 Matrix Multiplication Properties

AB is a valid **matrix product** if **A** is $p \times q$ and **B** is $q \times r$, or the left matrix has same number of columns q as the right matrix has rows. The standard matrix product is defined as follow:

$$(\mathbf{AB})_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iq}b_{qj} = \sum_{k=1}^{q} a_{ik}b_{kj}; \quad i = 1, ..., p \text{ and } j = 1, ..., r.$$

In other words, $(\mathbf{AB})_{ij}$ is the dot product of the ith row of **A** with the jth column of **B**.

Properties of matrix multiplication:

- Generally not commutative: $\mathbf{AB} \neq \mathbf{BA}$
- Left/Right Distributive over addition: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
- For some scalar λ: $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} = (\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$.
- Transpose of product: $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

Problem 4

Given the matrix **X** and the vectors **y** and **z** below:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{4}$$

(a) Expand $\mathbf{Xy} + \mathbf{z}$

(b) Expand $\mathbf{y^T Xy}$

Solution:

(a)

$$\mathbf{Xy} + \mathbf{z} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_{11}y_1 + x_{12}y_2 \\ x_{21}y_1 + x_{22}y_2 \end{pmatrix} + \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_{11}y_1 + x_{12}y_2 + z_1 \\ x_{21}y_1 + x_{22}y_2 + z_2 \end{pmatrix}$$

(b)

$$\mathbf{y^T Xy} = \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1x_{11} + y_2x_{21} & y_1x_{12} + y_2x_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$= y_1^2x_{11} + y_1x_{21}y_2 + y_2y_1x_{12} + y_2^2x_{22}$$

1.7 Rank, Determinant, Inverse

The **column rank** of a matrix **A** is the **dimension** of the vector space spanned by its column vectors, i.e., the number of linearly independent columns. The **row rank** is the dimension of the space spanned by its row vectors. A fundamental result in linear algebra is that the column rank and the row rank are always equal and this number is the **rank** of a matrix. If **A** is $n \times m$, then $rank(A) \leq min(n, m)$.

A matrix is **full rank** if its rank equals the largest possible for a matrix with the same dimensions, i.e. $\min(n, m)$. For a square matrix, full rank requires all its column (or row) vectors to be linearly independent.

The **determinant** $\det(\mathbf{A})$ is defined for a square matrix $\mathbf{A}$ and is a scalar quantity with various uses. Its computation differs for square matrices of different sizes. An n-by-n square matrix may have an **inverse**. There is a matrix inverse if and only if $\mathbf{A}$ has a non-zero determinant. A square matrix that is not invertible is called **singular**. $\det(\mathbf{A})$ is also the product of the **eigenvalues** of $\mathbf{A}$ (see Section 1.9). We will denote the determinant with single bars, e.g. $\det(X) = |\mathbf{X}|$. Do not confuse $|\mathbf{X}|$ with double bars $||\mathbf{X}||$, which typically denote a norm.

A few properties of the determinant (it's okay if you understand but can't recall from memory the rest of this section):

- The determinant of a diagonal matrix is the product of its diagonal values, and in particular the determinant of the **identity matrix** $\mathbf{I}$ is 1: $|\mathbf{I}| = 1$.
- For an $n \times n$-matrix $\mathbf{A}$ and a scalar value c we have $|c\mathbf{A}| = c^n|\mathbf{A}|$.
- The determinant factors over products: $|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|$.

The **inverse** $\mathbf{A}^{-1}$ of matrix operator $\mathbf{A}$ "undoes" $\mathbf{A}$ much like multiplying by $\frac{1}{x}$ undoes multiplying by x. We have $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. $\mathbf{A}^{-1}$ exists if and only if $|\mathbf{A}| \neq 0$. In general, matrix inversion is a complicated operation, but special cases that are easy to work with come up in the machine learning literature. Often analytical solutions to systems depend on the existence of the inverse of a matrix.

Problem 5

For an invertible matrix $\mathbf{A}$ show that $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$.

Solution:

By definition,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \implies |\mathbf{A}^{-1}\mathbf{A}| = |I| \implies |\mathbf{A}^{-1}| \cdot |\mathbf{A}| = |I| \implies |\mathbf{A}^{-1}| = \frac{|I|}{|\mathbf{A}|} = |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

□

(*) The **Moore-Penrose pseudoinverse** $\mathbf{A}^+$ of $\mathbf{A}$ is a generalization of the inverse to non-square matrices, where $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. Matrix $\mathbf{A}\mathbf{A}^+$ may not be the general identity matrix but maps all column vectors of $\mathbf{A}$ to themselves.

1.8 Matrix Properties

- $\mathbf{A}^\top$ is the **transpose** of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$. This is just like flipping the two dimensions of your matrix.
- $\mathbf{A}$ is **symmetric** if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
- (*) $\mathbf{A}$ is said to be **orthogonal** if its rows and its columns are orthogonal unit vectors. Consequence: $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$ where $\mathbf{I}$ is the **identity matrix** (ones on the main diagonal and zeros elsewhere). For an orthogonal matrix $\mathbf{A}$ we have $\mathbf{A}^\top = \mathbf{A}^{-1}$.
- **Diagonal** matrices have non-zero values on the main diagonal and zeros elsewhere. Diagonal matrices are easy to take powers of because you just take the powers of the diagonal entries. Under certain conditions a matrix may be diagonalized, see **eigen-decomposition** and **SVD** below.

- A matrix is **upper-triangular** if the only non-zero values are on the diagonal or above (top right of matrix). A matrix is **lower-triangular** if the only non-zero values are on the diagonal or below (bottom left of matrix).

1.9 Eigen-Everything

Recall that a matrix $\mathbf{A}$ can be thought of as an operator. Each square matrix $\mathbf{A}$ has some set of vectors $\mathbf{x} \in \mathbb{R}^n$ in its domain that are simply mapped to a scaled version of the vector in the codomain. The matrix preserves the direction of these vectors: $\mathbf{Ax} = \lambda\mathbf{x}$ for some scalar value λ. In this case, λ is an **eigenvalue** of $\mathbf{A}$ and $\mathbf{x}$ is a corresponding **eigenvector**. Eigenvectors can also be seen as the invariant directions of the matrix.

Problem 6

An *eigenspace* of a matrix $\mathbf{A}$ is an eigenvalue λ and the set $U_\lambda = \{\mathbf{v} \mid A\mathbf{v} = \lambda\mathbf{v}\}$. Show that U_λ is a vector subspace of the span of the columns of A.

Solution:

We first show that the span of the columns (column space) of A is a vector space. The column space of $\mathbf{A}$ is the set of all possible linear combinations of the column vectors of $\mathbf{A}$. $C(\mathbf{A}) = \{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^d\}$. Clearly $C(\mathbf{A})$ satisfies all 8 properties of being a vector space over the field of $\mathbb{R}$.

(VS 1) Let $\mathbf{x}, \mathbf{y} \in C(\mathbf{A})$. Then $\mathbf{x} + \mathbf{y} = \mathbf{Ax}' + \mathbf{Ay}' = \mathbf{Ay}' + \mathbf{Ax}' = \mathbf{y} + \mathbf{x}$.

(VS 2) Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in C(\mathbf{A})$. Then $(\mathbf{x}+\mathbf{y})+\mathbf{z} = (\mathbf{Ax}'+\mathbf{Ay}')+\mathbf{Az}' = \mathbf{Ax}'+(\mathbf{Ay}'+\mathbf{Az}') = \mathbf{x}+(\mathbf{y}+\mathbf{z})$

(VS 3) These exists an element, denoted by 0, in $C(\mathbf{A})$ such that $\mathbf{x} + 0 = \mathbf{x}, \quad \forall \mathbf{x} \in C(\mathbf{A})$. That element is the 0 vector $\mathbf{0}$.

(VS 4) Let $\mathbf{x} \in C(\mathbf{A})$. Then there exists an element $-\mathbf{x} \in C(\mathbf{A})$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$. Notice that $-\mathbf{x}$ is an element of $C(\mathbf{A})$: $-\mathbf{x} = -\mathbf{Ax}' = \mathbf{A}(-\mathbf{x}') \in C(\mathbf{A})$.

(VS 5) For each element $\mathbf{x} \in C(\mathbf{A}), \mathbf{Ix} = \mathbf{IAx}' = \mathbf{Ax}' = \mathbf{x}$.

(VS 6) Let $a, b \in \mathbb{R}, \mathbf{x} \in C(\mathbf{A})$. Then: $(ab)\mathbf{x} = (ab)\mathbf{Ax}' = a(b\mathbf{Ax}') = a(b\mathbf{x})$.

(VS 7) Let $a \in \mathbb{R}, \mathbf{x}, \mathbf{y} \in C(\mathbf{A})$. Then $a(\mathbf{x} + \mathbf{y}) = a(\mathbf{Ax}' + \mathbf{Ay}') = a\mathbf{Ax}' + a\mathbf{Ay}' = a\mathbf{x} + a\mathbf{y}$.

(VS 8) Let $a, b \in \mathbb{R}, \mathbf{x} \in C(\mathbf{A})$. Then $(a + b)\mathbf{x} = (a + b)\mathbf{Ax}' = (a\mathbf{Ax}' + b\mathbf{Ax}') = a\mathbf{x} + b\mathbf{x}$.

We then show U_λ is subspace of the span of the columns of $\mathbf{A}$. Notice first that U_λ is a subset of $C(\mathbf{A})$ because $C(\mathbf{A})$ consists of all vectors of the form $\mathbf{Ax}$ and U_λ consists of vectors $\mathbf{Ax}$ that satisfies a particular condition $\mathbf{Ax} = \lambda\mathbf{x}$.

Now the following three conditions hold:

- $\mathbf{0} \in U_\lambda$ because $\mathbf{A0} = \mathbf{0} = \lambda\mathbf{0}$.
- Given $\mathbf{x}, \mathbf{y} \in U_\lambda, \mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \lambda\mathbf{x} + \lambda\mathbf{y} = \lambda(\mathbf{x} + \mathbf{y}) \implies \mathbf{x} + \mathbf{y} \in U_\lambda$.
- Given $c \in \mathbb{R}$, notice that $\mathbf{A}(c\mathbf{x}) = c\mathbf{Ax} = c\lambda\mathbf{x} = \lambda(c\mathbf{x}) \implies c\mathbf{x} \in U_\lambda$.

So we have shown that U_λ is a subspace of $C(\mathbf{A})$.

□

Eigen-decomposition: Let $\mathbf{A}$ be an $n \times n$ full-rank matrix that has n linearly independent eigenvectors $\{\mathbf{q}_i\}_{i=1}^n$. In this case, $\mathbf{A}$ can be factored into $\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$ where $\mathbf{Q}$ is $n \times n$ and has eigenvector $\mathbf{q}_i$ for its i^{th} column. $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues: $\Lambda_{ii} = \lambda_i$. This is the **eigen-decomposition** of the matrix and we say the matrix has been **diagonalized**. If a matrix $\mathbf{A}$ can be eigen-decomposed and none of its eigenvalues are 0, then $\mathbf{A}$ is **nonsingular** (i.e., it is **invertible**) and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q\Lambda}^{-1}\mathbf{Q}^{-1}$ with $\mathbf{\Lambda}_{ii}^{-1} = \frac{1}{\lambda_i}$.

Singular Value Decomposition is a useful generalization of eigen-decomposition to rectangular matrices. Let $\mathbf{A}$ be an $m \times n$ matrix. Then $\mathbf{A}$ can be factored into $\mathbf{U\Sigma V}^\top = \mathbf{U\Sigma V}^{-1}$ where

- $\mathbf{U}$ is $m \times m$ and orthogonal. The columns of $\mathbf{U}$ are called the **left-singular vectors** of $\mathbf{A}$.
- $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative real entries. The diagonal values σ_i of $\mathbf{\Sigma}$ are known as the **singular values** of $\mathbf{A}$. These are also the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$.
- $\mathbf{V}$ is an $n \times n$ orthogonal matrix. The columns of $\mathbf{V}$ are called the **right-singular** vectors of $\mathbf{A}$.

1.10 Positive Definiteness

The symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be **positive definite** if, for every non-zero vector $\mathbf{x} \in \mathbb{R}^n$, it satisfies the property

$$\mathbf{x}^\top\mathbf{Ax} > 0$$

and **positive semi-definite** if it satisfies

$$\mathbf{x}^\top\mathbf{Ax} \geq 0.$$

Problem 7

(*) Show that positive definite matrices have all eigenvalues > 0 and positive semi-definite matrices have all eigenvalues ≥ 0.

Solution:

Let $\mathbf{A}$ be a positive definite matrix with eigenvalues $\lambda_1, ..., \lambda_n$. Since $\mathbf{A}$ is symmetric, then there exists an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{QDQ}^T$ where $\mathbf{D} = \text{diag}(\lambda_1, ..., \lambda_n)$. Then:

$$\mathbf{x}^T\mathbf{QDQ}^T\mathbf{x} = (\mathbf{Q}^T\mathbf{x})^T\mathbf{DQ}^T\mathbf{x} = \mathbf{y}^T\mathbf{Dy} = \sum_{i=1}^{n} \lambda_i y_i^2 > 0.$$

Since y_i^2 is non-negative, $\implies \lambda_i > 0$ for all i.

A similar proof can be done for positive semi-definite matrices as:

$$\mathbf{x}^T\mathbf{QDQ}^T\mathbf{x} = (\mathbf{Q}^T\mathbf{x})^T\mathbf{DQ}^T\mathbf{x} = \mathbf{y}^T\mathbf{Dy} = \sum_{i=1}^{n} \lambda_i y_i^2 \geq 0.$$

Since y_i^2 is non-negative, $\implies \lambda_i \geq 0$ for all i.

□

2 Calculus

Khan Academy has good reference material for calculus and multivariable calculus. For matrix calculus see *The Matrix Cookbook* by Petersen and Pedersen, specifically sections 2.4, 2.6, and 2.7.

2.1 Differentiation

You should be familiar with single-variable differentiation, including properties like:

$$\text{Chain rule: } \frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$
$$\text{Product rule: } \frac{d}{dx}f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$
$$\text{Linearity: } \frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)$$

for scalars a and b. In multivariable calculus, a function may have some number of inputs (say n) and some number of outputs (say m). In general, there is a partial derivative for every input-output pair. This is called the **Jacobian**. The j^{th} column of the Jacobian is made up of the partial derivatives of f_j (the j^{th} output value of $\mathbf{f}$) with respect to all input elements, rows $i = 1$ to n.

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

If f is scalar-valued (has only 1 output), its derivative is a column vector we call the **gradient vector**, written as ∇f:

$$\nabla f = \frac{df(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \cdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$. This is useful for optimization.

The **Hessian** matrix is like the Jacobian but with second-order derivatives. There are many interesting optimization topics related to the Hessian.

The most important vector or matrix derivatives that we will use in CS 181 can be found on p. 8-10 of *The Matrix Cookbook* by Petersen and Pedersen. We've reproduced a few important derivatives here:

$$\frac{d\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \frac{d\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \mathbf{a}$$
$$\frac{d\mathbf{a}^\top\mathbf{X}\mathbf{b}}{d\mathbf{X}} = \mathbf{a}\mathbf{b}^\top$$
$$\frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{b}}{d\mathbf{X}} = \mathbf{b}\mathbf{a}^\top$$
$$\frac{d\mathbf{a}^\top\mathbf{X}\mathbf{a}}{d\mathbf{X}} = \frac{d\mathbf{a}^\top\mathbf{X}^\top\mathbf{a}}{d\mathbf{X}} = \mathbf{a}\mathbf{a}^\top$$
$$\frac{d\mathbf{X}}{dX_{ij}} = \mathbf{B}^{ij} \quad ***$$

*** $\mathbf{B}$ is a matrix with all zeros except for a 1 in the i, j entry.

Have you ever wondered how to differentiate the norm of a matrix? The eigenvalues? For more, see *The Matrix Cookbook*.

2.2 Optimization

Local Extrema: Recall that the local extrema of a single-variable function can be found by setting its derivative to 0. The same is true here, using the condition $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can also search for local minima numerically using gradient-based methods.

Gradient Descent (we will learn this in class): We start with an initial guess at at a useful value for a parameter $\mathbf{w}$: $\mathbf{w}_0$. Then at each step i we update our guess by going in the direction of greatest descent of a loss function (opposite the direction of the gradient vector):

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \frac{df(\mathbf{w})}{d\mathbf{w}}$$

where $\eta > 0$ is the *step size*. We stop updating $\mathbf{w}_i$ when the value of the gradient is close to 0.

Lagrange Multipliers: This technique is used to optimize a function $f(\mathbf{x})$ given some constraint $g(\mathbf{x}) = 0$. First construct what is called the **Lagrangian function** $L(\mathbf{x}, \lambda)$:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Then, set the derivative of L with respect to both $\mathbf{x}$ and λ equal to 0:

$$\nabla L_{\mathbf{x}} = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \qquad \frac{\partial L}{\partial \lambda} = g(\mathbf{x}) = 0$$

If $\mathbf{x}$ is d-dimensional, this will give you a system of $d + 1$ equations. In this way, you can solve analytically for $\mathbf{x}$ to find the optimal value of $f(\mathbf{x})$ subject to the constraint $g(\mathbf{x})$. As with unconstrained optimization, this too is intractable and gradient descent is used to make progress.

Problem 8

Solve the following vector/matrix calculus problems.

(a) Let $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. Find $\nabla f(\mathbf{x})$.

(b) Let $f(\mathbf{w}) = (1 - \mathbf{w}^T\mathbf{x})^2$. Find $\nabla f(\mathbf{w})$ where the gradient is taken with respect to $\mathbf{w}$.

(c) Let $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$ are both differentiable. Find $\nabla f(\mathbf{x})$.

(d) Let $\mathbf{A}$ be a symmetric n-by-n matrix. If $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{w}^T\mathbf{x}$, find $\nabla f(\mathbf{x})$.

Solution:

(a) $\nabla_x f(\mathbf{x}) = \mathbf{x}$ using the first derivative property above.

(b) $\nabla_w f(\mathbf{w}) = 2(1 - \mathbf{w}^T\mathbf{x}) \cdot \nabla_w(1 - \mathbf{w}^T\mathbf{x}) = 2(1 - \mathbf{w}^T\mathbf{x}) \cdot -\nabla_w \mathbf{w}^T\mathbf{x} = 2(1 - \mathbf{w}^T\mathbf{x}) \cdot \mathbf{x}$, again using the first derivative property above.

(c) $\nabla_x f(\mathbf{x}) = \nabla_x g(h(\mathbf{x})) = g'(h(\mathbf{x})) \cdot h'(\mathbf{x})$ using the Chain Rule.

(d) $\nabla_x f(\mathbf{x}) = \frac{1}{2}\nabla_x \mathbf{x}^T\mathbf{A}\mathbf{x} + \nabla_x \mathbf{w}^T\mathbf{x} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)\mathbf{x} + \mathbf{w}$. using the first derivative property above and $\frac{d\mathbf{x}^t\mathbf{B}\mathbf{x}}{d\mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$.

3 Probability Theory

Problem 9 (Example 2.3.9 from the *Stat 110* textbook)

A patient named Fred is tested for a disease called conditionitis, a medical condition that afflicts 1% of the population. The test result is positive, i.e., the test claims that Fred has the disease. Let D be the event that Fred has the disease and T be the event that he tests positive.

Suppose that the test is "95% accurate." What that means is $P(T|D) = 0.95$ and $P(T^c|D^c) = 0.95$. Find the conditional probability that Fred has conditionitis, given his positive test result.

Solution: We want to find $P(D \mid T)$. Using Bayes' Theorem, we have

$$\begin{aligned} P(D \mid T) &= \frac{P(T \mid D)P(D)}{P(T)} \\ &= \frac{0.95 \times 0.01}{P(T)} \end{aligned} \tag{5}$$

However, we also need to find $P(T)$. Law of total probability gives us

$$\begin{aligned} P(T) &= P(T \mid D)P(D) + P(T \mid D^c)P(D^c) \\ &= 0.95 \times 0.01 + 0.05 \times 0.99 \\ &= 0.059 \end{aligned} \tag{6}$$

Plugging this in gives us

$$P(D \mid T) \approx 0.161 \tag{7}$$

Surprisingly, this means that Fred likely still doesn't have the disease. This is because it's more to be a false positive, given that much of the population doesn't have the disease.

As an extension, you can think about how we might increase our confidence that Fred does or does not have the disease. Does it matter if test results are independent?

Problem 10

(*) Show that if $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$. What does this mean?

Solution: Suppose that $A \subset B$, so we can write $B = A + C$ with A and C disjoint. Thus, we have (where the first line holds since probabilities are non-negative and the second line holds by countable additivity for disjoint sets)

$$\begin{aligned} \mathbb{P}(A) &\leq \mathbb{P}(A) + \mathbb{P}(C) \\ &= \mathbb{P}(B) \end{aligned} \tag{8}$$

□

This means that given two sets of outcomes, one of which contains the other, the larger set of outcomes is at least as likely as the smaller set of outcomes. Intuitively, this makes sense since if one event happens, the containing event must also happen.

Problem 11

An example of a discrete distribution X is the result from rolling a standard, fair 6-sided die.

(a) What is the set of outcomes Ω?

(b) Calculate $\mathbb{E}(X)$ and $\text{Var}(X)$

Solution:

1. $\Omega = \{1, 2, 3, 4, 5, 6\}$ (the six numbers that can come up)
2. $\mathbb{E}(X) = 3.5, \text{Var}(X) = \frac{105}{36} \approx 2.917$

Problem 12

Verify that $\text{Var}(aX + b) = a^2\text{Var}(X)$.

We can expand the terms to get

$$\begin{aligned} \text{Var}(aX+b) &= \mathbb{E}[(aX+b)^2] - (\mathbb{E}[aX+b])^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - (a\mathbb{E}[X] + b)^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - a^2\mathbb{E}[X]^2 - 2ab\mathbb{E}[X] - b^2 \\ &= a^2(\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\ &= a^2\text{Var}(X) \end{aligned} \tag{9}$$

□

□

Problem 13

Show that if X and Y are independent then $p(x|y) = p(x)$. Interpret this.

Solution: Using the definition of conditional probability

$$\begin{aligned} p(x \mid y) &= \frac{p(x,y)}{p(y)} \\ &= \frac{p(x)p(y)}{p(y)} \\ &= p(x) \end{aligned} \tag{10}$$

□

This says that if X and Y are independent, then knowing Y happened or didn't happen tells you nothing about the probability of X.

Problem 14

Does independence imply conditional independence? Does conditional independence imply independence?

Solution: Independence and conditional independence do not imply each other. Consider rolling a six sided die and define the events $A = \{1, 2\}, B = \{2, 4, 6\}, C = \{1, 4\}$. $P(A, B) = \frac{1}{6} = \frac{1}{3}\frac{1}{2} = P(A)P(B)$, so A, B independent, but $P(A \mid C)P(B \mid C) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$, whereas $P(A, B \mid C) = 0$, so they are not conditionally independent given C.

In the other direction, consider again the outcome of a six-sided die with the events $A = \{1, 3\}, B = \{1, 2\}, C = \{1\}$. $P(A, B \mid C) = 1 = P(A \mid C)P(B \mid C)$, so A, B conditionally independent given C. However, $P(A, B) = \frac{1}{6} \neq P(A)P(B) = \frac{1}{3}\frac{1}{3}$, so A, B not independent.

Intuitively, conditional independence is about the independence of events in some "subset" of the probability space, and that doesn't tell us anything about the independence of events in the whole probability space.

Problem 15

Prove Bayes' theorem.

Solution: We use the definition of conditional probability twice. We can rewrite the definition

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \Leftrightarrow p(x, y) = p(y \mid x)p(x) \tag{11}$$

Starting with the definition of conditional probability and applying 11, we get Bayes' Theorem.

$$\begin{aligned} p(x \mid y) &= \frac{p(x, y)}{p(y)} \\ &= \frac{p(y \mid x)p(x)}{p(y)} \end{aligned} \tag{12}$$

□

Problem 16

(*) Prove these five properties. The last one is tough!

1. From the definition of covariance,

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] \\ &= \text{Cov}(Y, X) \end{aligned} \tag{13}$$

2. From the definition of covariance,

$$\begin{aligned} \text{Cov}(X, X) &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \text{Var}(X) \end{aligned} \tag{14}$$

   Since variance is non-negative, this quantity is non-negative. In more detail, since squared quantities are non-negative, and the expectation of a non-negative quantity is non-negative, this means $\text{Cov}(X, X) \geq 0$.

3. Suppose towards contradiction that X takes on a value k other than its mean and

$$0 = \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{X \in \Omega} (X - \mathbb{E}[X])^2 p(X) dX \tag{15}$$

   This means that for $X = k$, the integrand is some positive quantity. For all other values in Ω, both $(X - \mathbb{E}[X])^2$ and $p(X)$ are non-negative, so the integral must therefore evaluate to a positive quantity. This contradictions 15, so X cannot take on a value other than its mean.

4. From the definition of covariance and linearity of expectation,

$$\begin{aligned} \text{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY - \mathbb{E}[aX + bY])(Z - \mathbb{E}[Z])] \\ &= \mathbb{E}[(aX + bY - a\mathbb{E}[X] - b\mathbb{E}[Y])(Z - \mathbb{E}[Z])] \\ &= \mathbb{E}[a(X - \mathbb{E}[X])(Z - \mathbb{E}[Z]) + b(Y - \mathbb{E}[Y])(Z - \mathbb{E}[Z])] \\ &= a\mathbb{E}[(X - \mathbb{E}[X])(Z - \mathbb{E}[Z])] + b\mathbb{E}[(Y - \mathbb{E}[Y])(Z - \mathbb{E}[Z])] \\ &= a\text{Cov}(X, Z) + b\text{Cov}(Y, Z) \end{aligned} \tag{16}$$

5. Proving this in the discrete case using Cauchy-Schwarz is instructive as well, but this is my favorite general proof–it's so pretty!

   Let X, Y random variables and define $Z = X + aY$. We know variance is weakly positive, so we have (by properties of variance)

$$\text{Var}(Z) = \text{Var}(Y) + a^2\text{Var}(X) + 2a\text{Cov}(X, Y) \geq 0 \tag{17}$$

   We can view this equation as a quadratic inequality with the variable in question being a. Since we know it is weakly positive, its discriminant must be weakly negative (if its discriminant were positive, this would imply that the parabola had two distinct real roots and therefore attain a negative value somewhere). Thus, we have

$$4\text{Cov}^2(X, Y) - 4\text{Var}(X)\text{Var}(Y) \leq 0 \Rightarrow |\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)} \tag{18}$$

Problem 17

Show that for random variables X, Y that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$.

Solution: Using the definition of variance and covariance, we have

$$\begin{aligned}\text{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - E[Y]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - E[Y]^2 + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)\end{aligned} \tag{19}$$

□

Problem 18

(*) Show that for random variables $X_1, ..., X_n$ that

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{Cov}(X_i, X_j).$$

Hint: Use induction and the problem above.

Solution: We will prove this inductively. For $n = 1$, the equation just says $\text{Var}(X_1) = \text{Var}(X_1)$.

Now for the inductive step. Suppose that this equation holds for some n and we will prove it for $n+1$. Using Problem 17 (for the first line) and the symmetry and bilinearity of covariance (for the second line), we have

$$\begin{aligned}\text{Var}(X_1 + \ldots + X_n + X_{n+1}) &= \text{Var}(X_1 + \ldots + X_n) + \text{Var}(X_{n+1}) + 2\text{Cov}(X_1 + \ldots + X_n, X_{n+1}) \\ &= \text{Var}(X_1 + \ldots + X_n) + \text{Var}(X_{n+1}) + 2 \sum_{1 \le i < n} \text{Cov}(X_i, X_{n+1})\end{aligned} \tag{20}$$

Now applying the equation for n, we can break the first term down into

$$\begin{aligned}\text{Var}(X_1 + \ldots + X_n + X_{n+1}) &= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{Cov}(X_i, X_j) + \text{Var}(X_{n+1}) + 2 \sum_{1 \le i < n} \text{Cov}(X_i, X_{n+1}) \\ &= \sum_{i=1}^{n+1} \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n+1} \text{Cov}(X_i, X_j)\end{aligned} \tag{21}$$

□

Problem 19

(*) Prove Adam's law. This is quite tough so feel free to look it up on Wikipedia if needed.

Solution: Below we present the proof in the case that X, Y take on a countable number of values. The proof in the general case is more an application of measure theory and doesn't provide very much instructive value. Feel free to access it on Wikipedia.

By the definition of expectation, we have (letting Ω_x, Ω_y be the sets of values that X and Y take on).

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]] &= \mathbb{E}\left[\sum_{x\in\Omega_x} xP(x \mid Y)\right] \\ &= \sum_{y\in\Omega_y}\left[\sum_{x\in\Omega_x} xP(x \mid y)\right]P(y) \\ &= \sum_{y\in\Omega_y}\sum_{x\in\Omega_x} xP(x,y)\end{aligned} \tag{22}$$

If the sums are finite, we can switch them around. Otherwise, the full hypothesis of Adam's law states that $\mathbb{E}[X]$ is defined, which implies that $\min(\mathbb{E}[\min(X,0)], \mathbb{E}[\max(X,0)]) < \infty$. If both $\mathbb{E}[\min(X,0)]$ and $\mathbb{E}[\max(X,0)]$ are finite, then this series converges absolutely and by Fubini's theorem, we can switch the sums. If one is finite and the other is infinite, the series diverges and we can still switch the sums. (If you only understood the first sentence of this paragraph, that's perfectly fine–ignore the rest!) Switching sums gives.

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_{x\in\Omega_x}\sum_{y\in\Omega_y} xP(x,y) \\ &= \sum_{x\in\Omega_x} x\sum_{y\in\Omega_y} P(x,y) \\ &= \sum_{x\in\Omega_x} xP(x) \\ &= \mathbb{E}[X]\end{aligned} \tag{23}$$

□

Problem 20

Prove Eve's law using Adam's law.

Solution: The first line is by the definition of variance, the second line is by Adam's law, and the fourth line is by definition of variance

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 \mid Y]] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 \mid Y]] - \mathbb{E}[\mathbb{E}[X \mid Y]^2] + \mathbb{E}[\mathbb{E}[X \mid Y]^2] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2 \\ &= \mathbb{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathbb{E}[X|Y]]\end{aligned} \tag{24}$$

□

Problem 21

Using the probability density function of $X \sim \mathcal{N}(0,1)$ show that X has mean 0 and variance 1.

Hint: The PDF is $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$. For the mean, you can reason about the properties of the PDF itself to get the answer without integration techniques. For the variance, use integration by parts and the fact that the PDF itself integrates to 1.

Solution: We see that X has mean 0 since the PDF is symmetric. This means the integral cancels to 0 since $p(x) = p(-x)$.

Now to calculate the variance. $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2]$. Thus, we have, substituting $u = \frac{x}{\sqrt{2}}$ and using integration by parts

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\infty} u^2 \exp\left(-u^2\right) du \\
&= \frac{2}{\sqrt{\pi}} \left(\left[-\frac{u}{2} \exp(-u^2) \right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \exp(-u^2) du \right) \\
&= \frac{2}{\sqrt{\pi}} \frac{1}{2} \int_{-\infty}^{\infty} \exp(-u^2) du \\
&= \frac{\sqrt{\pi}}{\sqrt{\pi}} \\
&= 1
\end{aligned} \tag{25}$$

□

Problem 22

Solve the following problems:

(a) Let $Z \sim \mathcal{N}(0,1)$. Find a random variable in terms of Z that has the distribution $\mathcal{N}(-2,4)$.

(b) (*) Show that in general, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Solution:

1. $X = -2 + 2Z$.

2. First to show the expectation. We have by properties of expectation

$$\begin{aligned} \mathbb{E}[aX + b] &= a\mathbb{E}[X] + b \\ &= a\mu + b \end{aligned} \tag{26}$$

Now to show variance. By properties of variance, we have

$$\begin{aligned} \text{Var}[aX + b] &= \text{Var}[aX] \\ &= a^2\text{Var}[X] \\ &= a^2\sigma^2 \end{aligned} \tag{27}$$

□

Problem 23

A simple random walk is defined by setting $X_0 = 0$ and letting $X_{i+1} = X_i + R_i$ where the R_i's are independent variables taking value $+1$ or -1 with equal probability: $\mathbb{P}(R_i = 1) = \mathbb{P}(R_i = -1) = 1/2$. Show that simple random walk is a Markov chain.

Solution: Fix arbitrary $j_1, \ldots, j_{n+1}$. I will show that the Markov property holds. Consider $\mathbb{P}(X_{n+1} = j_{n+1} | X_n = j_n, ..., X_1 = j_1)$. If j_{n+1} is one more or one less than j_n, then this will be 0.5, and otherwise it will be 0. Similarly, if j_{n+1} is one more or one less than j_n, then $\mathbb{P}(X_{n+1} = j_{n+1} | X_n = j_n)$ will be 0.5, and otherwise it will be 0. Thus, these probabilities will be equal and the Markov property holds.

□

Problem 24

Consider the following Markovian environment. The weather is either sunny (state 1) or rainy (state 2). If it's sunny today it will be sunny tomorrow with probability 0.7 and if it's rainy today it will be rainy tomorrow with probability 0.5.

- What is the transition matrix P for this environment?
- What is the corresponding stationary distribution?
- Let's say we start with the initial distribution $\begin{pmatrix}0.5 & 0.5\end{pmatrix}$. What does the distribution look like after 1 sample of the Markov chain? After 2? After 10? Please feel free to use a computer algebra system like WolframAlpha. Do you see any similarities with the stationary distribution? Check out the concept of the *"limiting distribution"* of a Markov chain if you're interested.

Solution:

- $$P = \begin{bmatrix}0.7 & 0.3\\ 0.5 & 0.5\end{bmatrix} \tag{28}$$
- $$\pi = \begin{bmatrix}0.625\\ 0.375\end{bmatrix} \tag{29}$$
- After 1 sample of the chain we have $\begin{pmatrix}0.6 & 0.4\end{pmatrix}$. After 2 it's $\begin{pmatrix}0.62 & 0.38\end{pmatrix}$. After 10 it's the stationary distribution. Thus we see that π is both the limiting and stationary distribution. When a limiting distribution exists, it is the unique stationary distribution

Problem 25

(*) Derive the likelihood and log-likelihood functions for i.i.d. samples $y_1, ..., y_n \sim \mathcal{N}(\mu, \sigma^2)$.

Solution: First the likelihood. We have

$$\begin{aligned} L(\mu, \sigma; y_1, ..., y_n) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \end{aligned} \tag{30}$$

The log likelihood is much simpler, we get

$$\ell(\mu, \sigma; y_1, ..., y_n) = -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{31}$$

Problem 26

(*) Compute the MLE estimates for i.i.d. samples $y_1, ..., y_n \sim \mathcal{N}(\mu, \sigma^2)$.

Solution: In order to find the MLE, we differentiate the log likelihood with respect to μ and σ. First we take the derivative with respect to μ

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma; y_1, \ldots, y_n) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\mu - 2y_i \tag{32}$$

Setting this equal to 0 gives

$$2n\mu = 2\sum_{i=1}^{n} y_i \Rightarrow \mu = \frac{\sum_{i=1}^{n} y_i}{n} \tag{33}$$

This makes logical sense. The MLE for μ is the just the average value of all the y_is! Now for σ.

$$\frac{\partial}{\partial \sigma} \ell(\mu, \sigma; y_1, \ldots, y_n) = -n\sqrt{2\pi}\frac{1}{\sigma\sqrt{2\pi}} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{34}$$

Setting this equal to 0 gives

$$0 = -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n} \tag{35}$$

This also makes sense–the MLE for σ is the unadjusted sample variance!