Section 01: Solutions

Definitions

Norms
We haven't covered norms yet but they are incredibly useful, and they show up quite often! We will cover them later in more detail, but for now it is sufficient to get yourself familiar with definitions for some of the most widely used norms! For any vector v that is n-dimensional, i.e. $v \in \mathbb{R}^n$, we have the following

(a) **One-norm (ℓ_1)**: $||v||_1 = \sum_{i=1}^{n} |v_i|$

(b) **Two-norm (ℓ_2)**: $||v||_2 = \sqrt{v^T v} = \sqrt{\sum_{i=1}^{n} v_i^2}$

(c) **∞-norm**: $||v||_\infty = \max_i |v_i|$

Symmetric Matrices and the Quadratic Form
Let us define a matrix $A \in \mathbb{R}^{n \times n}$.

(a) We have that the matrix A is symmetric iff $A = A^T$

(b) The quadratic form is defined to be $x^T Ax$ for any vector $x \in \mathbb{R}^n$. The matrix A is said to be positive semi-definite if $x^T Ax \geq 0$

1. Probability Review: PDF, CDF and Expectation

The **Cumulative Density Function** (CDF) $F_X : \mathbb{R} \to [0, 1]$ of a random variable X is defined as $\mathbb{P}(X \leq x)$. The **Probability Density Function** (PDF), $f_X : \mathbb{R} \to \mathbb{R}^{\geq 0}$, of the same random variable is defined as $f_X(x) = \frac{d}{dx} F_X(x)$.

Note that the CDF can be computed from the PDF, and vice versa; e.g. $F_X = \int_{-\infty}^{x} f(x)dx$.

We can use these functions to directly compute the expectation of random variables, since the expectation is defined in terms of the PDF: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$.

These functions can also be used to compute the distribution of any one-to-one transformation $g(\cdot)$ of the random variable: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$.

Note: this section focuses on the continuous case, but equivalent formulations hold in the discrete case by replacing integration with summation.

(a) You've just started a new exercise regimen. You start on the 2nd floor of CSE1, and then make a random choice:

- With probability p_1 you run up 2 flights of stairs.
- With probability p_2 you run up 1 flight of stairs.
- With probability p_3 you walk down 1 flight of stairs.

Where $p_1 + p_2 + p_3 = 1$.

You will do two iterations of your exercise scheme (with each draw being independent). Let X be the floor you're on at the end of your exercise routine. Recall you start on floor 2.

(i) Let Y be the difference between your ending floor and your starting floor in one iteration. What is $\mathbb{E}[Y]$ (in terms of p_1, p_2, p_3)?

Solution:

Recall for a random variable $X, \mathbb{E}[X] = \sum_i x_i \cdot p_i$.
So $\mathbb{E}[Y] = 2 \cdot p_1 + 1 \cdot p_2 + (-1) \cdot p_3$

(ii) What is $\mathbb{E}[X]$ (use your answer from the previous part)

Solution:

Since we start at floor 2, we can take 2 and add the difference ($\mathbb{E}[Y]$) twice to get our expected floor at the end of the routine.
$\mathbb{E}[X] = 2 + \mathbb{E}[Y] + \mathbb{E}[Y] = 2 + 2\mathbb{E}[Y]$

(iii) You change your scheme: instead of doing two independent iterations, you decide the second iteration of your regimen will just use the same random choice as your first (in particular they are no longer independent!). Does $\mathbb{E}[X]$ change? (Optional)

Solution:

No! We can say using the same choice as the first will effectively double Y, thus by linearity of expectation, $\mathbb{E}[X] = 2 + \mathbb{E}[2Y] = 2 + 2\mathbb{E}[Y]$

Fact 1. *Let $X_{(j)}$ denote the jth order statistic in a sample of i.i.d. random variables; that is, the jth element when the items are sorted in increasing order $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$.*

The PDF of $X_{(j)}$ is given by:

$$f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x). \tag{1}$$

(b) When a sample of $2N+1$ i.i.d. random variables is observed, the $(N+1)^{\text{st}}$ smallest is called the sample median. If a sample of size 3 from a uniform distribution over $[0,1]$ is observed, find the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$. *Hint: use Fact 1.*

Solution:

We will use Fact 1. To apply Fact 1, we can note that $n = 3, j = 2$ and

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x \geq 1 \end{cases} \tag{2}$$

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases} \tag{3}$$

We can use the PDF, which we compute via (2) and (3) to compute the probability that the median lies in the specified range:

$$\mathbb{P}\Big(\frac{1}{4} \leq X_{(2)} \leq \frac{3}{4}\Big) = \int_{\frac{1}{4}}^{\frac{3}{4}} f_{X_{(2)}}(x)dx \tag{4}$$

$$= 6\int_{\frac{1}{4}}^{\frac{3}{4}} (x)(1-x)dx \qquad \text{Using Fact 1 with } n=3,\ j=2 \tag{5}$$

$$= 6\Big[\frac{x^2}{2} - \frac{x^3}{3}\Big]\Big|_{x=\frac{1}{4}}^{x=\frac{3}{4}} \tag{6}$$

$$= \frac{11}{16} \tag{7}$$

2. Linear Algebra Review

Let $X \in \mathbb{R}^{m \times n}$. X may not have full rank. We explore properties about the four fundamental subspaces of X.

2.1. Summation form v.s. Matrix form

(a) Let $w \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Let $x_i^\top$ denote each row in X and y_i in Y. Show $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

Solution:

Note $Xw - Y$ is a vector in $\mathbb{R}^n$, and the i th row has the value $(x_i^\top w - y_i)$. Without loss of generality, let P be vector of any length. By linear algebra, $\|P\|_2$ means $\sqrt{\sum_i P_i^2}$. Also note the identity $P^T P = P \cdot P = \sum_i P_i \cdot P_i = \sum_i P_i^2$. Therefore, $\|P\|_2 = \sqrt{\sum_i P_i^2} = \sqrt{P^T P}$, and thus $\|P\|_2^2 = P^T P = \sum_i P_i^2$. Now substitute $P = Xw - Y$, and we naturally get $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

(b) Let $L(w) = \|Xw - Y\|_2^2$. What is $\nabla_w L(w)$? (Hint: You can use either summation or matrix form from first sub-problem).

Solution:

Matrix:

$$\begin{aligned}
\nabla_w L(w) &= \nabla_w \|Xw - Y\|_2^2 \\
&= \nabla_w (Xw - Y)^T (Xw - Y) \\
&= X^T (Xw - Y) + X^T (Xw - Y) \\
&= 2X^T (Xw - Y)
\end{aligned}$$

Summation: For an element w_j.

$$\begin{aligned}
\frac{\partial L(w)}{\partial w_j} &= \frac{\partial}{\partial w_j} \sum_{i=1}^m (x_i^\top w - y_i)^2 && (1) \\
&= \frac{\partial}{\partial w_j} \sum_{i=1}^m \left(\left(\sum_{k=1}^n x_{ik} w_k \right) - y_i \right)^2 && (2) \\
&= \sum_{i=1}^m \frac{\partial}{\partial w_j} \left(\left(\sum_{k=1}^n x_{ik} w_k \right) - y_i \right)^2 && (3) \\
&= \sum_{i=1}^m 2 \left(\left(\sum_{k=1}^n x_{ik} w_k \right) - y_i \right) \frac{\partial}{\partial w_j} \sum_{k=1}^n x_{ik} w_k \qquad \text{[chain rule]} && (4) \\
&= \sum_{i=1}^m 2 \left(\left(\sum_{k=1}^n x_{ik} w_k \right) - y_i \right) x_{ij} && (5)
\end{aligned}$$

Note that on line 4, when evaluating $\frac{\partial}{\partial w_j} \sum_{k=1}^n x_{ik} w_k$, the summation can be decomposed to

$$\frac{\partial}{\partial w_j} \left(\left(\sum_{k \neq j} x_{ik} w_k \right) + x_{ij} w_j \right)$$

The partial derivative of $\sum_{k \neq j} x_{ik} w_k$ will evaluate to 0, since it is not in terms of w_j. The partial derivative of $x_{ij} w_j$ will evaluate to x_{ij}.

For an element w_j. So for whole vector w:

$$\nabla_w L(w) = \begin{pmatrix} \sum_{i=1}^m 2\left(\left(\sum_{k=1}^n x_{ik} w_k\right) - y_i\right) x_{i1} \\ \vdots \\ \sum_{i=1}^m 2\left(\left(\sum_{k=1}^n x_{ik} w_k\right) - y_i\right) x_{ij} \\ \vdots \\ \sum_{i=1}^m 2\left(\left(\sum_{k=1}^n x_{ik} w_k\right) - y_i\right) x_{in} \end{pmatrix} = 2X^T(Xw - Y)$$

2.2. Subspaces of X

What is the rowspace, columnspace, nullspace, and rank of $X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6. \end{pmatrix}$?

Solution:

- Rowspace is the **span** (i.e., *the set of all linear combinations*) of the rows of X. Therefore, in this example, it is the subspace of vectors of the form $(1 \cdot x + 4 \cdot y, 2 \cdot x + 5 \cdot y, 3 \cdot x + 6 \cdot y)$ for all x and y.
- Columnspace (a.k.a. Range(X)) is the span of the columns of X. In this example, it is the subspace of vectors of the form $(1 \cdot x + 2 \cdot y + 3 \cdot z, 4 \cdot x + 5 \cdot y + 6 \cdot z)$ for all x, y, and z.
- Nullspace (a.k.a. Null(X)) is the set of vectors v such that $Xv = 0$. In this example, the nullspace is the subspace spanned by $(1, -2, 1)$.
- The matrix X can be reduced to the form $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$. This matrix has submatrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which has rank 2. Observe that the third column, $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$, is in the columnspace of this first submatrix.

2.3. Connections between subspaces of X

Check the following facts.

(a) The rowspace of X is the columnspace of $X^\top$, and vice versa.

Solution:

The matrix $X^\top$ is $\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$. The rows of X are the columns of $X^\top$, and vice versa.

(b) The nullspace of X and the rowspace of X are orthogonal complements. This can be written in shorthand as Null$(X) =$ Range$(X^\top)^\perp$. This is further equivalent to saying Range$(X^\top) =$ Null$(X)^\perp$.

Solution:

A vector $v \in$ Null(X) if and only if $Xv = 0$, which is true if and only if for every row X_i of X, $\langle X_i, v \rangle = 0$. This is precisely the condition that v is perpendicular to each row of X, which is the stated claim.

(c) The nullspace of $X^\top$ is orthogonal to the columnspace of X. This can be written in shorthand as Null$(X^\top) =$ Range$(X)^\perp$.

Solution:

This is seen by applying the previous result to $X^\top$.

2.4. Linear algebra facts for linear regression

We saw in lecture on Linear Regression that the closed form expression for linear regression without an offset involves the term $(X^\top X)^{-1}$.

(a) Is it true that the matrix $X^\top X$ is always symmetric and positive semidefinite?

Solution:

Yes. Symmetry can be checked by computing the transpose. For any vector u, we have $u^\top X^\top X u = \|Xu\|_2^2 \geq 0$.

(b) State and prove the connection between the nullspace of X and the nullspace of $X^\top X$. That is, your statement should look like one of the following: $\text{Null}(X) \subseteq \text{Null}(X^\top X)$, or $\text{Null}(X) \supseteq \text{Null}(X^\top X)$ or $\text{Null}(X) = \text{Null}(X^\top X)$.

Solution:

We have, $\text{Null}(X) = \text{Null}(X^\top X)$. Let $v \in \text{Null}(X)$. Then, one can check that $X^\top X v = 0$, leading to $v \in \text{Null}(X^\top X)$, which proves $\text{Null}(X) \subseteq \text{Null}(X^\top X)$. For the other direction, let $0 \neq v \in \text{Null}(X^\top X)$. Then, $0 = v^\top X^\top X v = \|Xv\|_2^2$, which implies $v \in \text{Null}(X)$. Therefore, $\text{Null}(X^\top X) \subseteq \text{Null}(X)$, which finishes the proof.

(c) Is it true that $X^\top X$ is always invertible?

Solution:

No, this isn't always the case. Since $\text{Null}(X) = \text{Null}(X^\top X)$ (see the answer to the previous question), the matrix $X^\top X$ is not invertible if X has a non-empty nullspace.

(d) Based on the above fact about the connection between the nullspaces of X and $X^\top X$ and the expression for linear regression without an offset (that we referred to two problems above), justify the use of "tall skinny" data matrix X as opposed to a "short wide" matrix X.

Solution:

If X is "short and wide", it has a non-empty nullspace. Therefore, $X^\top X$ is not invertible.

(e) The columnspace and rowspace of $X^\top X$ are the same, and are equal to the rowspace of X. (Hint: Use the relationship between nullspace and rowspace.)

Solution:

$X^\top X$ is symmetric, and from the previous parts, we have $\text{Row}(X^\top X) = \text{Col}((X^\top X)^\top) = \text{Col}(X^\top X)$. By previous parts again, we have: $\text{Row}(X^\top X) = \text{Null}(X^\top X)^\perp = \text{Null}(X)^\perp = \text{Row}(X)$.