446 Section 01

TA: Sankar Harilal

Welcome to section!

Plans for today!

1. Welcome!
2. Intro and Icebreakers
3. Reminders and Logistics
4. How are we running this?
5. Linear Algebra Review
6. Probability Review

Who Am I?

Sankar Harilal

- ML enjoyer
- 5th time TAing this class
- Applied ML and RL research
- Hobbies:
  - Snowboarding
  - Traveling
  - Soccer

OH:

- **Tuesdays 1:30 pm - 2:30 pm (CSE2 131)**
- **Wednesdays 11:30 am - 12:30 pm (CSE2 131)**

Say your name and your favorite sport/team (or food or hobby)!

Logistics

- Lectures: Monday, Wednesday, Friday 9:30am – 10:20am in CSE2 G20
- Sections on Thursdays (usually review topics taught in lecture; occasionally, we cover topics that we don't cover in lecture).
- 5 homeworks total + midterm (week 5 – Friday, 5/1) + final (finals week)
- Homeworks are worth 40% of your grade (in total), midterm is worth 24% of your grade, and final is worth 33% of your grade.
- **New! Section participation is worth 3% of your grade**
    - **Tracked via weekly Gradescope assignment (starts week 2)**
- Website: https://courses.cs.washington.edu/courses/cse446/26wi/
- Any questions on course policy, syllabus, etc?
- **Q: What are you looking forward to in ML?**

Reminders

- Homework 0 is out, due <1 week from now (Wednesday, 4/8).
  - Meant to be a review of the prereqs for this course
- Be sure to read the course syllabus, and ask us if you have any questions.

Canvas: Lecture recordings

Gradescope: Turn-ins/regrade requests

EdStem: Questions and comments

Office hours: Questions (likely more effective than Ed since it's in-person)

Course Website: Everything else

Off the record…

- I really, *really* want you all to enjoy this class, and **especially this section.**

- ML is a difficult class, but it can be very rewarding – you get out what you put into it

I promise to be more helpful than "read the question more carefully". If you're in this section and find me at OH, I will really try my best to ensure you succeed.

Anything I can do to ensure you all have fun in this class?
We can also talk more thoroughly at the end of class

Practice

Question 2.1

Quick Matrix Algebra/Calculus Refresher

Unsure if this has been taught before to you all (when I took 208, I definitely did not learn it)

Note: For the matrix calculus section to the right, **B** is a constant, square matrix. **b** is a scalar. *b* is a scalar. x is a column vector.

Rule	Comments
$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$	order is reversed, everything is transposed
$(\mathbf{a}^T\mathbf{Bc})^T = \mathbf{c}^T\mathbf{B}^T\mathbf{a}$	as above
$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$	(the result is a scalar, and the transpose of a scalar is itself)
$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$	multiplication is distributive
$(\mathbf{a}+\mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C}+\mathbf{b}^T\mathbf{C}$	as above, with vectors
$\mathbf{AB} \neq \mathbf{BA}$	multiplication is **not** commutative

Scalar derivative			Vector derivative		
$f(x)$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}x}$	$f(\mathbf{x})$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{B}$	$\rightarrow$	$\mathbf{B}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{b}$	$\rightarrow$	$\mathbf{b}$
x^2	$\rightarrow$	$2x$	$\mathbf{x}^T\mathbf{x}$	$\rightarrow$	$2\mathbf{x}$
bx^2	$\rightarrow$	$2bx$	$\mathbf{x}^T\mathbf{Bx}$	$\rightarrow$	$2\mathbf{Bx}$

Norms

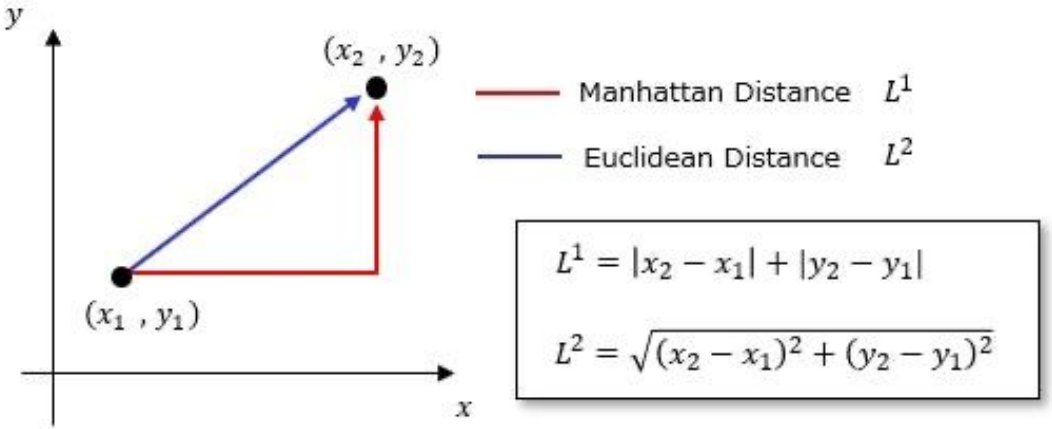
For any vector v that is n-dimensional, i.e. $v \in \mathbb{R}^n$

(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:** $||v||_\infty = \max_i |v_i|$

R^2 example



Question 2.1a, 2.1b

You are given some matrices, their shapes, and are asked to manipulate them!

$$X \in \mathbb{R}^{m \times n} \qquad w \in \mathbb{R}^n \qquad Y \in \mathbb{R}^m$$

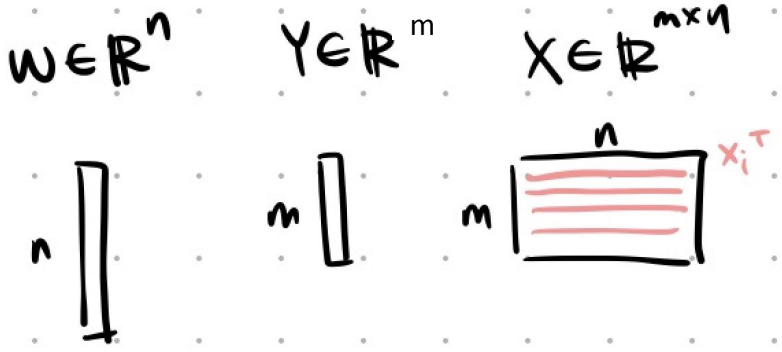
Let $x_i^\top$ denote each row in X and y_i in Y.

(a) Show $\|Xw - Y\|_2^2 = \sum_{i=1}^{m} (x_i^\top w - y_i)^2$.

(b) Let $L(w) = \|Xw - Y\|_2^2$. What is $\nabla_w L(w)$?

Context (2.1a)

Hint 1: See if you can break down the computation to be element wise (rather than dealing with entire matrices): suppose v = Xw - Y, write v_i explicitly. Then plug into the definition of $||v||_2$



Let $x_i^\top$ denote each row in X and y_i in Y.

Hint 2: The "2" superscript means square the whole norm: $\|Xw - Y\|_2^2$

This can be used on the norm equivalencies:

(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:** $||v||_\infty = \max_i |v_i|$

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 $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

$w \in \mathbb{R}^n$ $Y \in \mathbb{R}^m$ $X \in \mathbb{R}^{m \times n}$

let $r := Xw - Y \in \mathbb{R}^m$

$(m \times n)(n \times 1) \Rightarrow (m \times 1) \Rightarrow Xw - Y$

then $r_i := (Xw - Y)_i = x_i^\top w - y_i$

Apply def l_2 norm:

$$\|r\|_2 = \sqrt{r^\top r} \xrightarrow{\text{square}} \|r\|_2^2 = r^\top r = \sum_{i=1}^m r_i^2$$

↑ just plug in

$$= \sum_{i=1}^m (x_i^\top w - y_i)^2$$

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 $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

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 $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

Context (2.1b)

Hint 1: The "2" superscript means square the whole norm: $\|Xw - Y\|_2^2$

This can be used on the norm equivalencies →

(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:** $||v||_\infty = \max_i |v_i|$

Hint 2: Verify your matrix calculus is correct by ensuring the shapes of the matrices allow for valid matrix multiplication.

Scalar derivative			Vector derivative		
$f(x)$	$\rightarrow$	$\frac{df}{dx}$	$f(\mathbf{x})$	$\rightarrow$	$\frac{df}{d\mathbf{x}}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{B}$	$\rightarrow$	$\mathbf{B}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{b}$	$\rightarrow$	$\mathbf{b}$
x^2	$\rightarrow$	$2x$	$\mathbf{x}^T\mathbf{x}$	$\rightarrow$	$2\mathbf{x}$
bx^2	$\rightarrow$	$2bx$	$\mathbf{x}^T\mathbf{B}\mathbf{x}$	$\rightarrow$	$2\mathbf{B}\mathbf{x}$

Rule
$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$
$(\mathbf{a}^T\mathbf{B}\mathbf{c})^T = \mathbf{c}^T\mathbf{B}^T\mathbf{a}$
$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$
$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{A}\mathbf{C} + \mathbf{B}\mathbf{C}$
$(\mathbf{a}+\mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$
$\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$

2.1. Summation form v.s. Matrix form

(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).

$$L(w) = \|Xw - Y\|_2^2$$

$$\nabla_w L(w) = \nabla_w \left(\sqrt{(Xw-Y)^T(Xw-Y)}\right)^2$$

$$= \nabla_w (Xw-Y)^T(Xw-Y)$$

$$= \nabla_w (w^TX^T - Y^T)(Xw-Y)$$

Scalar transpose = scalar

$$= \nabla_w (w^TX^TXw - w^TX^TY - Y^TXw + Y^TY)$$

$$= \nabla_w (w^TX^TXw - 2w^TX^TY + Y^TY)$$

$$= 2X^TXw - 2X^TY$$

$$= X^T(2Xw - 2Y)$$

can take derivative wrt to either one - just be consistent

Useful gradients

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{b}) = \mathbf{b}$$

$$\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{B}\mathbf{x} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$$

if $\mathbf{B} = \mathbf{B}^T$ // symmetric
then $\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{B}\mathbf{x} = 2\mathbf{B}\mathbf{x}$

$$(X^TX)^T = X^TX$$

Question 1

Question 1a

(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.

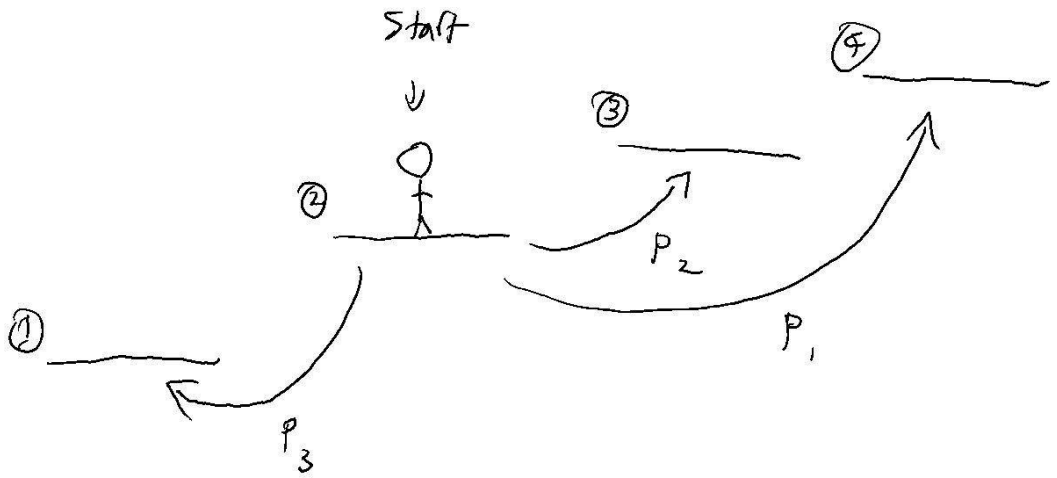
Where $p_1 + p_2 + p_3 = 1$.

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.

Context (1a)

New workout!

- Start at floor **2**
  - p_1 : +2
  - p_2 : +1
  - p_3 : -1
- Do this **twice**
  - No resetting to floor 2



Question 1a

(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)?

Question 1a

(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 $\mathbb{E}[Y] = 2 \cdot p_1 + 1 \cdot p_2 + (-1) \cdot p_3$

Question 1a

(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)

Question 1a

(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.
$\mathbb{E}[X] = 2 + \mathbb{E}[Y] + \mathbb{E}[Y] = 2 + 2\mathbb{E}[Y]$

Question 1a

(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)

Question 1a

(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, $\mathbb{E}[X] = 2 + \mathbb{E}[2Y] = 2 + 2\mathbb{E}[Y]$

Context (1b)

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:

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

Question 1b

(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.*

<u>More confusing than it needs to be…</u>

First sentence: Defining what a median is; you already know this, don't let it confuse you.

Second sentence: The actual question

Find the following, then plug & chug:

- *n* (not *N*)
- *j*
- *f(x)* (PDF)
- *F(x)* (CDF)

Hint: Both the PDF and CDF are piecewise

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

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

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

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

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

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

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

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

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

Question 2.2

Question - Q2.2

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

- 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.

Question - Q2.4a

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?

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

Questions/Chat Time!