PhD Bootcamp Day 2- Linear Algebra

Rick Presman and Emily Tallman

August 2021

Matrix Manipulation

- Addition- $(A+B)_{i,j} = A_{i,j} + B_{i,j}$
- Scalar Multiplication- $(cA)_{i,j} = cA_{i,j}$
- Transpose- $A^T_{i,j} = A'_{i,j} = A_{j,i}$
  - If $A^\intercal = A$, then A is a **symmetric** matrix
- Multiplication- If A is m-by-n and B is n-by-p, then $(AB)_{i,j} = \sum_{r=1}^{n} a_{i,r} b_{r,j}$
- Inversion- Let A^{-1} be the inverse of A, meaning $AA^{-1} = I$ Then we have that $(AB)^{-1} = B^{-1}A^{-1}$ (the shoes-socks theorem)
- Some useful properties:
  - If $a, b \in \mathbb{R}^n$, then $a^\intercal b = b^\intercal a$ because both sides of the equation are scalars.
  - If we denote the columns of $A \in \mathbb{R}^{m \times n}$ by $\mathbf{a}_j$, then $Ax = \sum_{j=1}^{n} x_j \mathbf{a}_j$.
  - If Λ is a diagonal matrix, then $A\Lambda A^\intercal = \sum_{j=1}^{n} \lambda_j \mathbf{a}_j \mathbf{a}_j^\intercal$

Matrix Calculus

- Derivatives of matrices/vectors are taken element-wise, i.e.
  - $\left(\frac{\partial \boldsymbol{X}}{\partial y}\right)_{i,j} = \frac{\partial x_{i,j}}{\partial y}$
  - $\left(\frac{\partial y}{\partial \boldsymbol{X}}\right)_{i,j} = \frac{\partial y}{\partial x_{i,j}}$
- $\partial(\boldsymbol{X}^\intercal) = (\partial \boldsymbol{X})^\intercal$
- Let $A \in \mathbb{R}^{n\times n}$ (not necessarily symmetric), and $x, b \in \mathbb{R}^n$. Then
  - $\partial_x(b^\intercal x) = b$
  - $\partial_x(x^\intercal Ax) = (A + A^\intercal)x$
  - If A is symmetric, then $\partial_x(x^\intercal Ax) = 2Ax$

Vector Spaces

A vector space over a field F is set V with the following properties for any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ and scalars $a, b \in F$:

- Commutativity: $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$
- Associativity: $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$
- Additive identity: $\exists 0 \in V$ such that $0 + \boldsymbol{x} = \boldsymbol{x}$
- Additive inverses: $\exists -\boldsymbol{x} \in V$ such that $-\boldsymbol{x} + \boldsymbol{x} = \boldsymbol{x}$
- Associativity scalar mult: $a(b\boldsymbol{x}) = (ab)\boldsymbol{x}$
- Distributivity of scalar sums: $(a + b)\boldsymbol{x} = a\boldsymbol{x} + b\boldsymbol{x}$
- Distributivity of vector sums: $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$
- Scalar mult identity: $\exists 1 \in F$ such that $1\boldsymbol{x} = \boldsymbol{x}$

Vector Spaces Continued

- The **span** of a set of vectors $\mathcal{S}$ over a field F is defined as the set of linear combinations of the vectors. Specifically

$$\text{span}(\mathcal{S}) = \{\sum_{i=1}^{n} \lambda_i v_i, n \in \mathbb{N}, v_2 \in \mathcal{S}, \lambda_i \in F\}$$

- A **basis** B is a set of vectors in V with two properties:
  - Linear independence: for any subset $v_1, ..., v_m \in B$, $c_1 v_1 + \cdots + c_m v_m = 0 \implies c_1, ..., c_m = 0$
  - Spanning: for any $v \in V$, $\exists c_1, c_2, ..., c_n \in F, v_1, v_2, ..., v_n \in F$ such that $v = c_1 v_1 + \cdots + c_n v_n$.
    Equivalently, $\text{span}(v_1, ..., v_n) = V$
- The **dimension** of a vector space V is equal to the cardinality of its basis.

Vector Spaces Continued

- Change of basis: Let $v_1, ..., v_n$ be the old basis of V, and $w_1, ..., w_n$ be the new basis. We can write $w_j = \sum_{i=1}^{n} a_{ij} v_i$ for all j. Let A be the matrix where column j is made of each of the a_{ij} corresponding to w_j. If

$$z = c_1 v_1 + ... + c_n v_n = b_1 w_1 + ... + b_n w_n$$

meaning the c_i and the b_i are the coordinates in the old and new bases respectively, then we can have that

$$c_i = \sum_{j=1}^{n} a_{i,j} b_j$$

or equivalently $c = Ab$, and $b = A^{-1} c$.

Linear Transformations

- Suppose V, W are vector spaces (both over $\mathbb{F}$). Then we call a function $T : V \to W$ a **linear transformation/map** if for all $u, v \in V, \alpha \in,\ T(\alpha u + v) = \alpha T(u) + T(v)$.
- A linear transformation that is also a bijection (one-to-one and onto) is called a **(linear) isomorphism.**
- Intuitively, a linear transformation preserves the "linear" structure between spaces: it doesn't break the vector space operations when moving from one space to another. Two spaces are isomorphic if they are the same and only differ in labels.
- If we choose bases $\{b_v\}_{v \in V}, \{b_w\}_{w \in W}$, then there is a matrix realization $M(T)_{b_v, b_w}$ of T, and the columns of the matrix are the images of b_v under T.
- Exercise: Suppose $V = W = \mathbb{R}^2$ and $\mathbb{F} = \mathbb{R}$. Let $T(v_1, v_2) = (v_1 + v_2, v_2)$ be the shearing transform. Show that it is a linear transformation, and write down its matrix realization.

Column Space, Row Space, Kernel

- Let M be a matrix. Then the **column space** $C(M)$ is the span of the columns of M. and it is equal to the image of the linear transformation T associated with M.
- More concretely, if $w \in C(M) \subset W$, then there exists $v \in V$ such that $w = Mv$.
- The **row space** is the column space of $M^\intercal$.
- The **rank** of a matrix M is the dimension of the column space (or the row space).
- The **kernel/null space** of M is defined as the preimage of 0: $N(M) = \{v \in V : Mv = 0\}$.
- Rank-Nullity Theorem: $V = \text{rank}(M) + \dim N(M)$.
- Exercise: For the shearing transformation, find a vector in the column space. Can you find a vector not in the column space? Why or why not?

Norms

- Let V be a vector space over $\mathbb{R}$. A **norm** $\|\cdot\| : V \to \mathbb{R}$ is an $\mathbb{R}$-valued function that satisfies the following properties:
    - $\|v\| \geq 0$ for all $v \in V$
    - $\|v\| = 0$ if and only if $v = 0$
    - $\|\lambda v\| = |\lambda|\|v\|$ for all $v \in V$ and $\lambda \in \mathbb{R}$
    - $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$
- We call the pair $(V, \|\cdot\|)$ a **normed space**.
- Intuitively, norms measure the size of vectors.
- Example: Let $V = \mathbb{R}^n$. The (usual) ℓ_2-norm is defined by $\|v\|_2 := \sqrt{\sum_{j=1}^n v_j^2} = (v^\mathsf{T} v)^{1/2}$.
- Equivalence of Norms: Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are two norms. Then there exist constants $c, C > 0$ such that

$$c\|\cdot\|_a \leq \|\cdot\|_b \leq C\|\cdot\|_a$$

- This last fact tells us that we can use whichever norm is convenient for a problem.

Inner Products

- Let V be a vector space over $\mathbb{R}$. An **inner product** $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ is an $\mathbb{R}$-valued function that satisfies the following properties:
    - $\langle \alpha v_1 + v_2, w \rangle = \alpha \langle v_1, w \rangle + \langle v_2, w \rangle$ for all $\alpha \in \mathbb{R}$ and $v_1, v_2, w \in V$
    - $\langle v, w \rangle = \langle w, v \rangle$
    - $\langle v, v \rangle \geq 0$ for all $v \in V$
    - $\langle v, v \rangle = 0$ if and only if $v = 0$
- We call the pair $(V, \langle \cdot, \cdot \rangle)$ an **inner product space**.
- Intuitively, inner products measure angles between angles.
- Example: Let $V = \mathbb{R}^n$. Then the (usual) dot product is defined by $\langle v, w \rangle_{\text{dot}} := v^\mathsf{T} w = \sum_{j=1}^n v_j w_j$.
- Inner products naturally induce norms: $\|x\| := \sqrt{\langle x, x \rangle}$.
- Example: $\|x\|_2 = \sqrt{x^\mathsf{T} x}$
- Cauchy-Schwarz: For $v, w \in \mathbb{R}^n$, $|\langle v, w \rangle| \leq \|v\|_2 \|w\|_2$, where $\langle v, w \rangle := v^\mathsf{T} w$.

Orthogonality

- Informally, a fancy word for perpendicular, see for example, this 2010 Supreme Court case
- Formally, suppose $(V, \langle\cdot,\cdot\rangle)$ is an inner product space. We say $v, w \in V$ are **orthogonal** if $\langle v, w\rangle = 0$.
- Unless specified otherwise, you can generally assume the dot product for the inner product, which gives the definition $w^\intercal v = v^\intercal w = 0$.
- A matrix A is **orthogonal** if $A^\intercal A = AA^\intercal = I$, or equivalently $A^\intercal = A^{-1}$
- We say $v, w \in V$ are **orthonormal** if $w^\intercal v = 0$ and $\|v\| = \|w\| = 1$
- In particular, the previous definitions can be combined to say that the columns/rows of an orthogonal matrix are orthonormal

Orthogonality Continued

- Let us take a subspace $W \subset V$. Then we define the **orthogonal complement** $W^{\perp} = \{v \in V, w \in W : \langle w, v \rangle = w^{\intercal} v = 0\}$
- If you have seen direct sums before, you can show that for any subspace $W \subset V$, $V = W \oplus W^{\perp}$.
- Example: If $M \in \mathbb{R}^{m \times n}$, then $C(M)^{\perp} = N(M^{\intercal})$. In particular, this means we can write

$$V = C(M) \oplus C(M)^{\perp} = C(M) \oplus N(M^{\intercal})$$

Eigenvalues (Done Wrong)

- Let $M \in \mathbb{R}^{n\times n}$ be a matrix. Then an **eigenvalue** of M is a number such that $Mv = \lambda v$, where v is called an **eigenvector**.
- If M is symmetric, the **Spectral Theorem** says that there exists an **eigenvalue decomposition** (EVD) of M: $M = Q\Lambda Q^\intercal$, where Q is an orthogonal matrix (i.e., $Q^\intercal Q = QQ^\intercal = I_n$) with columns consisting of the eigenvectors of M and Λ is a diagonal matrix with the diagonal elements consisting of the eigenvalues of M.
  - The EVD can be used to find the inverse of M, since $M^{-1} = (Q\Lambda Q^\intercal)^{-1} = (Q^{-1})^{-1}\Lambda^{-1}Q^{-1} = Q\Lambda^{-1}Q^{-1}$
- Properties: Suppose λ is an eigenvalue of M. Then
  - For $n \in \mathbb{N}$, λ^n is an eigenvalue of M^n.
  - If M is nonsingular, then the above extends to $n \in \mathbb{Z}$.
  - The eigenvalues of $I + M$ are $1 + \lambda$.
  - Eigenvalues can be found using the characteristic equation: $\det(M - \lambda I) = 0$.

Eigenvalues (Done Right)

- Let $T : V \to V$ be a linear operator, and let $W \subset V$ be a subspace. We say W is **invariant** if $T(W) \subset W$. That is, once you're in W, you cannot escape under the action of T.
- Now, suppose $W = \text{span}(w)$ is invariant for some vector $w \in V$. Note this is a 1-dimensional subspace. By definition, $T(w) \in W$, so there exists some λ such that $T(w) = \lambda w$.
- Geometrically, this is to say T (or its matrix) stretches or compresses its eigenvectors.
- Therefore, the eigenvalues characterize the 1-dimensional invariant subspaces of V with respect to T.
- If all such w are linearly independent of each other, the eigenvectors form an **eigenbasis**, and the matrix is **diagaonlizable**, meaning we can write the EVD of the matrix associated with T.

Projections

- A matrix $P \in \mathbb{R}^{n \times n}$ is a **projection** if $P^2 = P$.
- A matrix $P \in \mathbb{R}^{n \times n}$ is an **orthogonal projection** if in addition to being a projection it is symmetric: $P^\intercal = P$.
- Geometrically, a projection matrix takes a vector in a higher-dimensional space and pushes it to a lower-dimensional subspace (think taking a 2D picture of a 3D object).
- Exercise: Take $v \in \mathbb{R}^2$, and define $T : \mathbb{R}^2 \to \mathbb{R}^2$ by $T(v_1, v_2) = (v_1, 0)$. Write down the matrix realization of T, and show that it is an orthogonal projection matrix. Draw a sketch of T acting on $(1, 1)$ to convince yourself of this fact.
- Exercise: Find a matrix $M \in \mathbb{R}^{2 \times 2}$ that is a projection, but not an orthogonal projection.

Projections (continued)

- The eigenvalues of a projection matrix are 0 and 1. The number of 1's is equal to rank(P). To see the first fact, note by symmetry we have an EVD: $P = Q\Lambda Q^\intercal$. By $P^2 = P$, we have $Q\Lambda^2 Q^\intercal = Q\Lambda Q^\intercal$, so $\Lambda^2 = \Lambda$. Then $\lambda_i(\lambda_i - 1) = 0$, so $\lambda_i = 0, 1$.
- Another perspective: since P is symmetric, we have an eigenbasis $\{v_i\}_{i=1}^n$ with corresponding eigenvalues λ_i. Take any vector $v \in \mathbb{R}^n$, and write $v = \sum_{j=1}^n \alpha_j v_j$. Then note that

$$Pv = \sum_{j=1}^{n} \alpha_j P v_j = \sum_{j=1} \alpha_j \lambda_j v_j = \sum_{j:\lambda_j=1} \alpha_j v_j$$

- In the above line, the projection matrix "turned off" the part of the basis with eigenvalues equal to 0. Moreover,

$$P^2 v = P\left(\sum_{j:\lambda_j=1} \alpha_j v_j\right) = \sum_{j:\lambda_j=1} \alpha_j P v_j = \sum_{j:\lambda_j=1} \alpha_j v_j$$

Determinants

- The **determinant** of a square matrix A is a complex function of its values with some nice properties, detonated $\det(A)$ or $|A|$. It can be defined in terms of the eigenvalues of the matrix, with $|A| = \prod_{i=1}^{n} \lambda_i$
- For a 2-by-2 matrix we have that

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

- Some of the nice properties
  - $|I| = 1$
  - If at least two columns or row of A are the same, $|A| = 0$
  - For n-by-n matrix A and scalar c, $|cA| = c^n|A|$
  - If A is a triangular matrix, $|A| = \prod_{i=1}^{n} a_{i,i}$, the product of the diagonal entries
  - $|A| = |A^\intercal|$
  - $|AB| = |A||B|$
  - A is invertible if and only if $|A| \neq 0$
  - If A is invertible, $|A^{-1}| = \frac{1}{|A|}$

Trace

- The **trace** of a square matrix A is a much less complex function of its values with some nice properties, equal to the sum of the diagonal elements and denoted $\text{tr}(A)$. The trace is also the sum of the eigenvalues of A, and is invariant to a change of basis.
- Some of the nice properties
  - $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
  - $\text{tr}(A^\intercal) = \text{tr}(A)$
  - $\text{tr}(AB) = \text{tr}(BA)$
  - For scalar c, $\text{tr}(cA) = c\text{tr}(A)$
  - Unlike the determinant, we do not have $\text{tr}(AB) = \text{tr}(A)\text{tr}(B)$
  - More generally, the trace is invariant under cyclic permutations, meaning

$$\text{tr}(ABCD) = \text{tr}(DABC) = \text{tr}(CDAB) = \text{tr}(BCDA)$$

Quadratic Forms

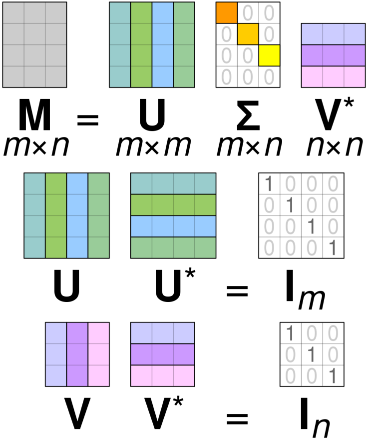
- Let $A \in \mathbb{R}^{n \times n}$ be a matrix. We refer to the expression $x^\intercal Ax$ as a **quadratic form**.
- Another way to write it is $x^\intercal Ax = \sum_{1 \leq i,j \leq n} a_{ij} x_i x_j$
- We say A a (symmetric) matrix is positive definite, or PD, (semidefinite, or PSD) if $x^\intercal Ax > 0$ ($x^\intercal Ax \geq 0$) for all $x \in \mathbb{R}^n$. We define negative definite. (semidefinite) analogously.
- The following are equivalent:
  - A is PD
  - All of the eigenvalues of A are positive
  - All of the principal minors are positive

Singular Value Decomposition (SVD)

- Problem: Suppose $A \in \mathbb{R}^{m \times n}$. How do we compute an "eigenvalue" of A, which we denote by $\lambda_i(A)$?
- Define the **singular values** $\sigma_i(A) = \sqrt{\lambda_i(A^\intercal A)}$, $i = 1, \ldots, n$.
- We drop the dependence on A when it is obvious from context: $\sigma_i(A) \equiv \sigma_i$.
- We take the convention that the singular values are arranged in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.
- Properties:
  - The singular values are all non-negative: $\sigma_i(A) \geq 0$.
  - If $\text{rank}(A) = r$, then $\sigma_i(A) > 0$ for $i = 1, \ldots, r$, and $\sigma_i(A) = 0$ for $i = r + 1, \ldots, n$.

Singular Value Decomposition (SVD)

- **Theorem:** Suppose $M \in \mathbb{R}^{m\times n}$. Then there exists $U \in \mathbb{R}^{m\times m}, \Sigma \in \mathbb{R}^{m\times n}, V \in \mathbb{R}^{n\times n}$ such that:
  - $M = U\Sigma V^{\intercal}$
  - The columns of U and V are orthonormal.



Cholesky

- Suppose $A \in \mathbb{R}^{n \times n}$ is a (symmetric) PD matrix. Then there exists a lower triangular matrix $B \in \mathbb{R}^{n \times n}$ such that $A = BB^\intercal$.
- B can be found by multiplying out $BB^\intercal$, setting it equal to A, and solving for the terms b_{ij} through a set of equations
- Somewhat interestingly, a good way to check if a matrix is positive definite is to perform a Cholesky decomposition, and if the algorithm doesn't terminate successfully, then the matrix is not PD

Gram-Schmidt and QR

- The **Gram-Schmidt Orthonormalization Process** is a finite-step algorithm that takes any set of n vectors in $\mathbb{R}^n$ and generates an orthonormal basis.
- If $A \in \mathbb{R}^{n \times n}$ is a square matrix, then the QR decomposition allows us to write $A = QR$ where $Q \in \mathbb{R}^{n \times n}$ is an orthonormal matrix and $R \in \mathbb{R}^{n \times n}$ is an upper triangular matrix.
- In particular, if we takes the n column vectors of A, then the columns of Q are the output of the Gram-Schmidt Process
- In practice, there are better ways than the Gram-Schmidt process to compute the QR decomposition
- The QR decomposition is a quick way to compute the determinant of a matrix, since we have

$$\det(A) = \det Q \det R = \prod_{i=1}^{n} r_{i,i} = \prod_{i=1}^{n} \lambda_i$$

Block Matrices

- Suppose $A \in \mathbb{R}^{m \times n}$. Fix some $p = 1, \ldots, m$ and $q = 1, \ldots, n$. We can write A in the form of a block matrix:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

  where $A_{11} \in \mathbb{R}^{p \times q}$, $A_{12} \in \mathbb{R}^{p \times (n-q)}$, $A_{21} \in \mathbb{R}^{(m-p) \times q}$, and $A_{22} \in \mathbb{R}^{(m-p) \times (n-q)}$

- Compatible block matrices can be added and multiplied by treating the appropriate blocks as "scalars"
- If $A_{21} = 0$ or $A_{12} = 0$, then the A is block upper-triangular (lower-triangular), and we have in either case

$$\det(A) = \det(A_{11}) \det(A_{22})$$

Systems of Equations

A system of m linear equations with n unknowns can be written as

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}$$

or in vector notation as

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

or in matrix notation as $Ax = b$ where

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Systems of Equations Continued

Assuming the rows of A are linearly independent, meaning there are no redundancies in the equations, we have that:

- If m (number of equations) $> n$ (number of unknowns), there are no solutions
- If $m < n$, there are infinitely many solutions
- If $m = n$, there is a single solution, given by $x = A^{-1}b$

The End

Feel free to reach out with any questions-
Emily (emily.tallman@duke.edu) & Rick (rick.presman@duke.edu)