PhD Bootcamp : Linear Algebra

Bongjung Sung

August 16, 2023

Outline

Definition 1 (Field)

A field is a set F equipped with two binary operations, called addition and multiplication. A binary operation maps a value in $F \times F$ to a unique element in F. Addition and multiplication is denoted by $+$ and $\cdot$, and satisfy the following properties.

- $a + (b + c) = (a + b) + c$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- $a + b = b + a$ and $a \cdot b = b \cdot a$.
- There exist two distinct elements 0 and 1 such that $a + 0 = a$ and $a \cdot 1 = a$ for any $a \in F$.
- For any $a \in F$, there exists a unique element in F, denoted by $-a$, such that $a + (-a) = 0$.
- For any $a \in F \setminus 0$, there exists a unique element in F, denoted by a^{-1}, such that $a \cdot a^{-1} = 1$.
- For $a, b, c \in F$, it holds that $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$.

- Consider $\mathbb{R}$, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{C}$ with usual addition and multiplication.
- $\mathbb{R}$, $\mathbb{Q}$ and $\mathbb{C}$ are field but $\mathbb{N}$ and $\mathbb{Z}$ are not.

Definition 2 (Vector Space)

A vector space V over a field F is a non-empty set with two binary operations, vector addition and scalar multiplication. The vector addition assigns two elements in V, say u and v, to a unique element w in V, denoted by $u+v$. The scalar multiplication assigns $a \in F$ and $u \in V$ to au in V. These two operations satisfy the following properties :

- For any $u, v, w \in V$, $u+(v+w)=(u+v)+w$.
- For any $u, v \in V$, $u+v=v+u$.
- There exists $0 \in V$, called a nonzero vector, such that $v+0=v$ for any $v \in V$.
- For each $v \in V$, there exists a unique vector $-v$ in V such that $v+(-v)=0$.
- $a(bv)=(ab)v$ for any $a, b \in F$ and $v \in V$.
- For $1 \in F$, $1v=v$ for all $v \in V$.
- $(a+b)v=av+bv$ for all $a, b \in F$ and $v \in V$.
- $a(v+u)=av+au$ for all $a \in F$ and $v, u \in V$.

Subspace

Definition 3 (Subspace)

The subset $W \subseteq V$ for a vector space V over F is a subspace of V if W is a vector space over F with the operations of addition and scalar multiplication defined on V.

Definition 4 (Linear Combination & Span)

For vectors $v_1, \ldots, v_n$ in a vector space V over F, we say $a_1 v_1 + \cdots + a_n v_n$ is a linear combination of $v_1, \ldots, v_n$ for $a_1, \ldots, a_n \in F$. For nonempty set S of V, we denote the set of all linear combinations of the vectors in S by $\text{span}(S)$. For convenience, we define $\text{span}(\emptyset) = \{0\}$. If $W = \text{span}(S)$, then we say S generates or spans W.

Definition 5 (Basis)

Suppose $\beta \subset V$. β is called linearly dependent if there exist a finite number of vectors $v_1, \ldots, v_n \in \beta$ and scalars $c_1, \ldots, c_n$, not all zero, such that

$$c_1 v_1 + \cdots + c_n v_n = 0.$$

Otherwise, β is linearly independent. If linearly independent subset β of V generates $W \subset V$, then β is a basis for W. If β is finite, then we say W is finite-dimensional and define its dimension by the cardinality of β.

Subspace

Theorem 1

Let V be a vector space over a field F. For a subset β of V, β is a basis for V if and only if each element in V can be uniquely expressed as a linear combination of vectors of β, i.e., for each $v \in V$ and $\beta = \{v_1, \ldots, v_n\}$, there exist unique scalars $c_1, \ldots, c_n \in V$ such that $v = c_1 v_1 + \cdots c_n v_n$.

Theorem 2

Let V be a vector space over a field F. If V is generated by a finite set S, then some subset of S is a basis for V. Hence, V has a finite basis.

Theorem 3

Let V be a vector space that is generated by a set G containing exactly n vectors, and let L be a linearly independent subset of V containing exactly m vectors. Then $m \leq n$ and there exists a subset H of G containing exactly $n - m$ vectors such that $L \cup H$ generates V.

Corollary 1

Let V be a vector space having a finite basis. Then every basis for V contains the same number of vectors.

Corollary 2

Let V be a vector space with dimension n.

- *Any finite generating set for V contains at least n vectors, and a generating set for V that contains exactly n vectors is a basis for V.*
- *Any linearly independent subset of V that contains exactly n vectors is a basis for V.*
- *Every linearly independent subset of V can be extended to a basis for V.*

Theorem 4

Let W be a subspace of a finite-dimensional vector space V. Then, W is finite-dimensional and $\dim(W) \leq \dim(V)$. Moreover, if $\dim(W) = \dim(V)$, $W = V$.

Definition 6

For a vector space V over F, the inner product is a map $<,>: V \times V \mapsto F$ such that

- $< cx + z, y >= c < x, y > + < z, y >$ for any $c \in F$ and $x, y, z \in V$.
- $\overline{< x, y >} =< y, x >$.
- $< x, x > \geq 0$ and the equality holds only when $x = 0$.

A vector space equipped with an inner product is called an inner product space. If $< v, w >= 0$ for $v, w \in V$, then v and w are orthogonal. If all elements in the subset S of V are mutually orthogonal, then we say S is orthogonal. The norm induced by an inner product $< \cdot, \cdot >$ is a map $|| \cdot || : V \mapsto \mathbb{R}$ defined by $||x|| = \sqrt{< x, x >}$. If any element in orthogonal set S has a norm 1, then the set S is orthonormal.

Theorem 5 (Gram-Schdmit Process)

Let V be an inner product space and $S = \{w_1, \ldots, w_n\}$ be a linearly independent subset of V. Define $S' = \{v_1, \ldots, v_n\}$, where $v_1 = w_1$ and

$$v_k = w_k - \sum_{j=1}^{k-1} \frac{< v_j, w_k >}{||v_j||^2} v_j$$

for $2 \leq k \leq n$. Then S' is an orthogonal set of nonzero vectors such that span(S) = span(S').

- As a consequence of Theorem 5, any finite-dimensional vector space V has an orthonormal basis.
- In fact, any orthogonal finite set is linearly independent.

Linear Map & Matrix

- Suppose V and W are finite-dimensional vector spaces over F. Then, the map $T : V \mapsto W$ is linear if $T(c_1 v_1 + v_2) = c_1 T(v_1) + v_2$
- Every linear map between finite-dimensional vector spaces has a matrix representation.
- If $\{\beta_1, \ldots, \beta_n\}$ and $\{\gamma_1, \ldots, \gamma_m\}$ are ordered bases for V and W, respectively, then for each $i = 1, \ldots, n$, there exist unique scalars $g_{i1}, \ldots, g_{im}$in F such that

$$T(\beta_j) = \sum_{i=1}^{m} g_{ij} \gamma_i.$$

- Then, we can identify T with a matrix $G = (g_{ij})$.
- Conversely, matrix can be viewed as a linear map. From now on, we restrict our vector space to $\mathbb{R}^d$ over $\mathbb{R}$. We write the space of all real $m \times n$ matrix by $\mathbb{R}^{m \times n}$. Then, it is straightforward to see that for $X \in \mathbb{R}^{m \times n}$, X can be viewed as a linear map that maps $\mathbb{R}^n$ to $\mathbb{R}^m$.

Definition 7 (Column Space & Kernel)

Suppose $X \in \mathbb{R}^{n \times p}$. Then, the column space of X, $C(X)$, is the linear subspace of $\mathbb{R}^n$ spanned by columns of X. The rank of X, rank(X), is the dimension of $C(X)$. The kernel of X, ker(X), is the set of all vectors $v \in \mathbb{R}^p$ such that $Xv = 0$. The nullity of X is the dimension of ker(X).

Definition 8 (Null Space)

The null space of X, null(X), is the set of all vectors $v \in \mathbb{R}^n$ such that $v^\top X = 0$.

Theorem 6 (Rank-Nullity Theorem)

For $X \in \mathbb{R}^{n \times p}$, rank$(X)$ + nullity$(X) = p$.

Corollary 3

For any $X \in \mathbb{R}^{n \times p}$, rank$(X) =$ rank$\left(X^\top\right)$.

Theorem 7

For $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$, rank$(AB) \leq \min\{$rank$(A)$, rank$(B)\}$.

- We call matrix $X \in \mathbb{R}^{n \times p}$ is of full-rank if the rank of X attains its maximum value, i.e., rank$(X) = \min\{n, p\}$.

Definition 9 (Transpose)

Suppose $X = (x_{ij}) \in \mathbb{R}^{m \times n}$. Then, the transpose of X, denoted by $X^\top$, is defined to be a matrix in $\mathbb{R}^{n \times m}$ such that $(X^\top)_{ij} = x_{ji}$. If $m = n$ and $X^\top = X$, we call X is symmetric.

- We denote the set of all real symmetric $d \times d$ matrix by $\mathbb{S}^d$.
- For $X \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{m \times n}$, $(XY)^\top = Y^\top X^\top$.

Definition 10 (Trace)

Suppose $X = (x_{ij}) \in \mathbb{R}^{n \times n}$. Then, the trace of X, denoted by $\text{tr}(X)$, is the sum of diagonal entries of X, i.e., $\text{tr}(X) = \sum_{i=1}^{n} x_{ii}$.

- For $X \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{m \times n}$, $\text{tr}(XY) = \text{tr}(YX)$.
- Show that there does not exist $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$ such that $AB - BA = I_n$.

Definition 11 (Determinant)

Suppose $X = (x_{ij}) \in \mathbb{R}^{n\times n}$. If $n = 2$, then the determinant of X, denoted by $\det(X)$, is

$$\det(X) = x_{11}x_{22} - x_{12}x_{21}.$$

If $n \geq 3$,

$$\det(X) = \sum_{j=1}^{n} (-1)^{i+j} x_{ij} \det M_{ij},$$

where M_{ij} is a $(n-1)\times(n-1)$ submatrix of X by omitting ith row and jth column of X.

- For $A, B \in \mathbb{R}^{n\times n}$, $\det(AB) = \det(A)\det(B)$.

Definition 12 (Inverse)

For $X \in \mathbb{R}^{n\times n}$, X^{-1} is an inverse of X in $\mathbb{R}^{n\times n}$ if $X^{-1}X = XX^{-1} = I_n$. If X has an inverse, then X is called invertible or non-singular.

Theorem 8

Suppose $X \in \mathbb{R}^{n\times n}$. Then, X is invertible if and only if det$(X) \neq 0$. Also, X is invertible if and only if rank$(X) = n$.

Theorem 9

Suppose $X = (x_{ij}) \in \mathbb{R}^{n\times n}$ is invertible. Then, its inverse X^{-1} is $Y^{\top}$, where $Y = (y_{ij})$ is defined by

$$y_{ij} = (-1)^{i+j} \mathit{det} M_{ij} / \mathit{det} X.$$

and M_{ij} is a $(n-1)\times(n-1)$ sub-matrix of X by omitting ith row and jth column of X.

Projection Matrix

Definition 13

For $P \in \mathbb{R}^{n\times n}$, P is a projection matrix if $P^2 = P$.

Definition 14

For a projection matrix $P \in \mathbb{R}^{n\times n}$, P is orthogonal if Py and $(I-P)y$ are orthogonal for $y \in \mathbb{R}^n$.

Projection Matrix

- Suppose $X \in \mathbb{R}^{n\times p}$. and assume $n \geq p$ and X is of full-rank.
- Then, $P_X = X(X^\top X)^{-1}X^\top$ is a projection matrix onto $C(X)$.
- Observe that $P_X y = y$ for all $y \in C(X)$. Also, $P_X^2 y = P_X y$ for any $y \in \mathbb{R}^n$.
- Show that P is an orthogonal projection matrix if and only if $P^\top(I - P) = 0$. Deduce that P is an orthogonal matrix if and only if P is symmetric.
- Let $P_X^\perp = I - P_X$. Then, $P_X^\perp$ is a projection matrix onto $\mathcal{N} = \text{null}(X)$. Also, $P_X^\perp P_X = 0$.
- For every $y \in \mathbb{R}^n$, there uniquely exist $u \in C(X)$ and $v \in \mathcal{N}$ such that $y = u + v$.

Definition 15 (Eigenvalue & Eigenvector)

For $X \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{C}$ is an eigenvalue of X if there exists a nonzero eigenvector $v \in \mathbb{C}$ such that $Xv = \lambda v$.

- Alternative definition of eigenvalue is that the eigenvalue is a solution of $\det(X - tI_n) = 0$.

Definition 16 (Diagonalizable matrix)

For $X \in \mathbb{R}^{n\times n}$, X is diagonalizable if there exist an invertible matrix $Q \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that $X = QDQ^{-1}$. We call such a decomposition as an eigendecomposition.

- For any $X \in \mathbb{S}^d$, X has an eigendecomposition and $X = Q\Lambda Q^\top$, where Λ is a diagonal matrix with all diagonal entries being eigenvalues of X and Q is an orthogonal matrix whose columns are corresponding eigenvectors. Moreover, eigenvalues of X are real.

Definition 17 (Positive (Semi)Definite Matrix)

A matrix $X \in \mathbb{S}^d$ is called positive semidefinite if

$$v^\top X v \geq 0$$

for all nonzero vector $v \in \mathbb{R}^d$ and positive definite if the equality is strict.

- One can show that $X \in \mathbb{S}^d$ is positive (semi) definite if and only if all eigenvalues of X are strictly positive (nonnegative).
- Furthermore, one can show that if $X \in \mathbb{R}^{n \times p}$ for $n \geq p$, $X^\top X$ is positive definite if and only if $\mathrm{rank}(X) = p$.

SVD

Definition 18

For $X \in \mathbb{R}^{m \times n}$, the left singular values of X are eigenvalues of $XX^\top$ and the left singular vectors are corresponding eigenvectors. Similarly, the right singular values of X are eigenvalues of $X^\top X$ and the right singular vectors are corresponding eigenvectors.

Theorem 10 (SVD)

For $X \in \mathbb{R}^{m \times n}$, suppose X has rank r. there exist $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ and $\Sigma \in \mathbb{R}^{m \times n}$ such that

$$X = U \Sigma V^\top$$

and $U^\top U = UU^\top = I_m$, $V^\top V = VV^\top = I_n$ and Σ is a "diagonal" matrix that takes the square root of positive singular values in decreasing order on the upper left diagonal part of Σ and 0 on other entries.

- Using SVD, one can construct $\tilde{X} \in \mathbb{R}^{n \times r}$ such that $\tilde{X}$ is of full-rank and $C(\tilde{X}) = C(X)$.

Outline

- Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$. Then, the gradient of f at $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is defined by

$$\operatorname{grad} f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d} \right),$$

  provided that all the stated partial derivatives exist.
- The Hessian matrix of f at $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is defined by

$$\operatorname{Hess} f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}.$$

Examples of Vector Calculus

- For $A \in \mathbb{R}^{d\times d}$, suppose $f(x) = x^\top Ax$. Then, $\mathsf{grad} f(x) = (A + A^\top)x$ and $\mathsf{Hess} f(x) = A + A^\top$. In case $A \in \mathbb{S}^d$, $\mathsf{grad} f(x) = 2Ax$ and $\mathsf{Hess} f(x) = 2A$.
- Suppose $f(x) = \log\left(\frac{\exp(\beta^\top x)}{(1+\exp(\beta^\top x))^2}\right)$. Then, $\mathsf{grad} f(x) = \beta - 2\frac{\exp(\beta^\top x)}{1+\exp(\beta^\top x)}\beta$ and $\mathsf{Hess} f(x) = -2\frac{\exp(\beta^\top x)}{(1+\exp(\beta^\top x))^2}\beta\beta^\top$.
- Suppose $f(x) = \log(1 + x^\top x)$. Then, $\mathsf{grad} f(x) = 2x/(1 + x^\top x)$ and $\mathsf{Hess} f(x) = 2/(1 + x^\top x) \cdot I_d - 4/(1 + x^\top x)^2 \cdot xx^\top$.

- Suppose $f : \mathbb{R}^{d\times d} \mapsto \mathbb{R}$. Then, the gradient of f at $X = (x_{ij}) \in \mathbb{R}^{d\times d}$ is defined by

$$\nabla f(X) = (\frac{\partial f}{\partial x_{ij}}),$$

provided that all the stated partial derivatives exist.

- There are various ways to define the Hessian matrix of f. Viewing f as a function of d^2 entries, one can define the Hessian matrix of f as usual in vector calculus or using a tensor.

- Suppose $f(X) = \log\det(X)$. Then, $\nabla f(X) = (X^{-1})^\top$. If f is defined on $\mathbb{S}^d$, note that $\nabla f(X) = m(X^{-1})$, where $m(X) = 2X - \text{diag}(X)$.
- Suppose $f(X) = \text{tr}(AX)$. Then, $\nabla f(X) = A$. If $X \in \mathbb{S}^d$, $\nabla f(X) = m(A)$.
- Suppose $f(X) = \sum \text{eig}\,(X)$. Then, $\nabla f(X) = I$. Also, if $f(X) = \prod \text{eig}(X)$, $\nabla f(X) = \det(X)\left(X^{-1}\right)^\top$. If $X \in \mathbb{S}^d$, $\nabla f(X) = \det(X) m(X^{-1})$.

Outline

- Consider the matrix

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A and D are square matrices of appropriate size.

Theorem 11

Suppose D is invertible. Then, X is invertible if and only if D's Schur's complement $A - BD^{-1}C$ is also invertible. Furthermore,

$$\mathit{det}(X) = \mathit{det}(D)\mathit{det}(A - BD^{-1}C).$$

Corollary 4

Suppose D is invertible. If X is invertible, then

$$X^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

Theorem 12

Suppose $\Sigma \in \mathbb{R}^{p \times p}$ *is symmetric and*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix}.$$

Then, Σ *is positive definite if and only if* Σ_{22} *and its Schur's complement* $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top$ *are positive definite.*

- Suppose

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix} \right).$$

- Show that the conditional distribution of $Y_1|Y_2$ is

$$\mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top).$$

Woodbury's Formula

Theorem 13 (Matrix Determinant Lemma)

Suppose A is invertible, and u and v are column vectors. Then,

$$\mathit{det}(A + uv^\top) = (1 + v^\top A^{-1} u)\mathit{det}(A).$$

Theorem 14 (Generalization of Matrix Determinant Lemma)

Suppose $A \in \mathbb{R}^{n\times n}$ is invertible and $U, V \in \mathbb{R}^{n\times k}$. Then,

$$\mathit{det}(A + UV^\top) = \mathit{det}(A)\mathit{det}(I_k + V^\top A^{-1} U).$$

Also, if $W \in \mathbb{R}^{k\times k}$ is invertible,

$$\mathit{det}(A + UWV^\top) = \mathit{det}(A)\mathit{det}(W)\mathit{det}(W^{-1} + V^\top A^{-1} U).$$

Theorem 15 (Woobury's Formula)

Suppose $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times k}$, and $V \in \mathbb{R}^{k\times n}$. Then,

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

provided that all the stated inverses exist.

Kronecker Product

Definition 19

Suppose $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ and $B = (b_{ij}) \in \mathbb{R}^{p\times q}$. Then, the Kronecker product $A \bigotimes B$ is the $pm \times nq$ block matrix defined by

$$A \bigotimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Properties of Kronecker Product

- $A \bigotimes (B + C) = A \bigotimes B + A \bigotimes C$, $(A + B) \bigotimes C = A \bigotimes C + B \bigotimes C$.
- For $c \in \mathbb{R}$, $(cA) \bigotimes B = A \bigotimes (cB) = c(A \bigotimes B)$.
- $(A \bigotimes B)^\top = A^\top \bigotimes B^\top$.
- $(A \bigotimes B)(C \bigotimes D) = (AC) \bigotimes (BD)$.
- $\text{vec}(AYB^\top) = (B \bigotimes A)\text{vec}(Y)$.

Application of Kronecker Product : one-factor ANOVA

- Suppose $y_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\theta_j, \sigma^2)$, where $\sigma^2 > 0$ is unknown, $i = 1, \ldots, r$, and $j = 1, \ldots, p$.
- Observe that the given model can be also written as

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{r1} & \cdots & y_{rp} \end{bmatrix} = \begin{bmatrix} \theta_1 & \cdots & \theta_p \\ \vdots & \ddots & \vdots \\ \theta_1 & \cdots & \theta_p \end{bmatrix} + E, \tag{1}$$

  where the entries of $r \times p$ matrix E are random samples from $\mathcal{N}(0, \sigma^2)$.
- Express the model in (1) using $y = \text{vec}(Y)$, $e = \text{vec}(E)$, and Kronecker product.
- Using the model in form of Kronecker product, derive the OLS estimate of $\theta = (\theta_1, \ldots, \theta_p)^\top$.
- Deduce the fitted values of y and residuals from your OLS estimate.