

Section 01: Suplimentary Material for Expectation

Prepared by: Shuyang Hou

Notation and Assumptions

Throughout these notes, X , Y , Z , and $X_1, \dots, X_n$ are real-valued random variables. The coefficients a , b , c , and $a_1, \dots, a_n$ are deterministic constants in $\mathbb{R}$. All expectations appearing below are assumed to exist and be finite. Whenever variance or covariance is used, assume that the relevant random variables have finite second moments.

1. Linearity of Expectation

The most important rule is

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

Independence is not required. This identity holds even when X and Y are dependent.

1.1. Basic Forms

Linearity gives the following familiar identities:

$$\begin{aligned}\mathbb{E}[c] &= c, \\ \mathbb{E}[aX] &= a\mathbb{E}[X], \\ \mathbb{E}[X + Y] &= \mathbb{E}[X] + \mathbb{E}[Y], \\ \mathbb{E}[X - Y] &= \mathbb{E}[X] - \mathbb{E}[Y].\end{aligned}$$

More generally, for a finite weighted sum,

$$\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i \mathbb{E}[X_i].$$

The random variables need not be independent or identically distributed. However, the coefficients a_i must be deterministic constants; a random coefficient cannot generally be pulled outside an expectation.

1.2. Example: The Sample Mean

Suppose each X_i has the same expectation μ , and define

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then

$$\mathbb{E}[\bar{X}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

This calculation requires equal expectations, not independence.

2. Products, Independence, and Covariance

2.1. When Can We Factor an Expectation?

If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

More generally, if $X_1, \dots, X_n$ are *mutually independent*, then

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Without an appropriate additional assumption, a product cannot be factored this way. Linearity alone does not justify factoring a product.

2.2. Counterexample Without Independence

Let

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}, \quad Y = X.$$

Then

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0, \quad \mathbb{E}[XY] = \mathbb{E}[X^2] = 1.$$

Consequently,

$$\mathbb{E}[XY] = 1 \neq 0 = \mathbb{E}[X]\mathbb{E}[Y].$$

2.3. The General Identity

For random variables with finite second moments,

$$\boxed{\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \text{Cov}(X, Y).}$$

Thus, covariance measures the difference between the expectation of a product and the product of the expectations.

Proof. Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. By the definition of covariance,

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Expand the product and apply linearity:

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_Y \mathbb{E}[X] - \mu_X \mathbb{E}[Y] + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_X \mu_Y. \end{aligned}$$

The quantities μ_X and μ_Y are deterministic constants, so they can be pulled outside the expectation. Rearranging gives

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \text{Cov}(X, Y).$$

No independence assumption was used. □

Independence is sufficient, but not necessary. With finite second moments, the identity $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ holds exactly when $\text{Cov}(X, Y) = 0$. Independence implies zero covariance, but zero covariance does not in general imply independence.

3. Expectations of Functions

3.1. Apply the Function First, Then Average

For a measurable function g with $\mathbb{E}[|g(X)|] < \infty$, its expectation can be computed using the distribution of X :

$$\boxed{\mathbb{E}[g(X)] = \begin{cases} \sum g(x)\mathbb{P}(X = x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f_X(x) dx, & \text{if } X \text{ has density } f_X. \end{cases}}$$

In general, the operations of applying a function and taking an expectation cannot be interchanged. In other words,

$$\mathbb{E}[g(X)] = g(\mathbb{E}[X]) \quad \text{is not a general rule.}$$

An important exception is an affine function $g(x) = ax + b$:

$$\mathbb{E}[g(X)] = \mathbb{E}[aX + b] = a\mathbb{E}[X] + b = g(\mathbb{E}[X]).$$

3.2. Second Moments and Variance

A particularly important identity is

$$\boxed{\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2.}$$

Equivalently,

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

For the random variable that takes 1 and -1 with equal probability,

$$\mathbb{E}[X^2] = 1, \quad (\mathbb{E}[X])^2 = 0.$$

Remember the distinction:

$$\underbrace{\mathbb{E}[X^2]}_{\text{square first, then average}} \quad \text{versus} \quad \underbrace{(\mathbb{E}[X])^2}_{\text{average first, then square}}.$$

4. Indicator Random Variables

For an event A , define its indicator by

$$\mathbf{1}_A = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Its expectation is

$$\boxed{\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).}$$

Indeed,

$$\mathbb{E}[\mathbf{1}_A] = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A).$$

4.1. Counting Events

If N counts how many of the events $A_1, \dots, A_n$ occur, then

$$N = \sum_{i=1}^n \mathbf{1}_{A_i}.$$

By linearity,

$$\boxed{\mathbb{E}[N] = \sum_{i=1}^n \mathbb{P}(A_i).}$$

The events do not need to be independent. This is a useful way to turn an expected-count problem into a collection of probability calculations.

5. Order Properties and Inequalities

5.1. Nonnegativity and Monotonicity

If $X \geq 0$ almost surely, meaning with probability 1, then

$$\mathbb{E}[X] \geq 0.$$

If $X \leq Y$ almost surely, then

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

In particular, if $a \leq X \leq b$ almost surely, then

$$a \leq \mathbb{E}[X] \leq b.$$

Another useful inequality is

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

5.2. Jensen's Inequality

If g is convex on an interval containing the values of X , then

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

If g is concave, the inequality is reversed:

$$g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)].$$

For example, $g(x) = x^2$ is convex, so

$$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2].$$

Although nonlinear functions cannot generally be moved outside an expectation, convexity or concavity can provide a useful comparison.

6. Conditional Expectation

6.1. The Law of Total Expectation

The law of total expectation states that

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]].$$

When Y is discrete, this becomes

$$\mathbb{E}[X] = \sum_{y: \mathbb{P}(Y=y)>0} \mathbb{E}[X | Y = y] \mathbb{P}(Y = y).$$

Intuitively, first average within each group determined by Y , and then average the group means using the probabilities of the groups.

6.2. Conditional Linearity

Conditional expectation also satisfies linearity:

$$\mathbb{E}[aX + bZ | Y] = a\mathbb{E}[X | Y] + b\mathbb{E}[Z | Y].$$

As with ordinary linearity, no independence assumption is required.

6.3. Taking Out What Is Known

A function of the conditioning variable can be pulled outside a conditional expectation:

$$\mathbb{E}[h(Y)X \mid Y] = h(Y)\mathbb{E}[X \mid Y],$$

provided h is measurable and the relevant products are integrable. The reason is that, once Y is given, the value of $h(Y)$ is known. This does *not* mean that a random variable can generally be pulled outside an unconditional expectation.

Identities between conditional expectations are understood to hold almost surely.

Section 01: Linear Algebra and Matrix Calculus

Prepared by: Shuyang Hou

1. Notation and Basic Identities

All matrices and vectors are real. Vectors are column vectors unless transposed. Throughout the linear algebra sections,

$$X \in \mathbb{R}^{m \times n}, \quad w \in \mathbb{R}^n, \quad Y \in \mathbb{R}^m, \quad r = \text{rank}(X).$$

Write the rows of X as $x_i^\top$ and its columns as c_j :

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix} = (c_1 \quad \cdots \quad c_n).$$

Thus $x_i \in \mathbb{R}^n$, $c_j \in \mathbb{R}^m$, and

$$(Xw)_i = x_i^\top w, \quad Xw = \sum_{j=1}^n w_j c_j \in \mathbb{R}^m.$$

The first expression views multiplication through rows; the second views it as a linear combination of columns.

1.1. Products, Transposes, and Trace

For compatible dimensions,

$$\begin{aligned} (AB)_{ij} &= \sum_k A_{ik} B_{kj}, & (AB)^\top &= B^\top A^\top, \\ (A+B)^\top &= A^\top + B^\top, & (A^\top)^\top &= A. \end{aligned}$$

Matrix multiplication is associative and distributive, but generally not commutative: AB and BA need not be equal, and one may be undefined.

For $A \in \mathbb{R}^{k \times k}$, its trace is $\text{tr}(A) = \sum_{i=1}^k A_{ii}$. Useful identities are

$$\text{tr}(AB) = \text{tr}(BA), \quad \text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB).$$

Here every displayed product must be well-defined and square. Cyclic permutations are allowed; arbitrary reorderings are not.

1.2. Vector Norms and the Frobenius Norm

For $v \in \mathbb{R}^n$, the norms used in the handout are

$$\|v\|_1 = \sum_{i=1}^n |v_i|, \quad \|v\|_2 = \sqrt{v^\top v}, \quad \|v\|_\infty = \max_i |v_i|.$$

In particular,

$$\|v\|_2^2 = v^\top v = \sum_i v_i^2.$$

For a matrix W , the Frobenius norm is the square root of the sum of squared entries:

$$\|W\|_F^2 = \sum_{i,j} W_{ij}^2 = \text{tr}(W^\top W).$$

1.3. Symmetry and Positive Definiteness

A square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^\top$. Its quadratic form is the scalar $v^\top A v$. For a *symmetric* real matrix, we use the definitions

$$\begin{aligned} A \succeq 0 &\iff v^\top A v \geq 0 \quad \text{for every } v \in \mathbb{R}^n, \\ A \succ 0 &\iff v^\top A v > 0 \quad \text{for every nonzero } v \in \mathbb{R}^n. \end{aligned}$$

These mean positive semidefinite (PSD) and positive definite (PD), respectively. The quantifier “for every” is essential. A PSD matrix can be singular; a PD matrix is invertible.

2. The Four Fundamental Subspaces and Rank

2.1. Span, Basis, and Dimension

The span of $v_1, \dots, v_k$ is

$$\text{span}\{v_1, \dots, v_k\} = \left\{ \sum_{i=1}^k \alpha_i v_i : \alpha_i \in \mathbb{R} \right\}.$$

A basis is a linearly independent spanning set. The dimension of a subspace is the number of vectors in any basis.

2.2. Definitions

The four subspaces associated with X are [?, ?]

$$\begin{aligned} \text{Row}(X) &= \text{span}\{x_1, \dots, x_m\} = \{X^\top u : u \in \mathbb{R}^m\} \subseteq \mathbb{R}^n, \\ \text{Col}(X) &= \text{span}\{c_1, \dots, c_n\} = \{X v : v \in \mathbb{R}^n\} \subseteq \mathbb{R}^m, \\ \text{Null}(X) &= \{v \in \mathbb{R}^n : X v = 0\}, \\ \text{Null}(X^\top) &= \{u \in \mathbb{R}^m : X^\top u = 0\}. \end{aligned}$$

The column space is also called the range. The nullspace is also called the kernel. The space $\text{Null}(X^\top)$ is the *left nullspace*, since $X^\top u = 0$ is equivalent to $u^\top X = 0$. In defining the row space, rows are identified with their transposes so that its elements are column vectors in $\mathbb{R}^n$.

Rank is the common dimension of the row and column spaces:

$$\boxed{\text{rank}(X) = \dim \text{Row}(X) = \dim \text{Col}(X) = r.}$$

It equals the maximum number of independent columns, the maximum number of independent rows, and the number of pivots in row-echelon form. Consequently,

$$0 \leq r \leq \min(m, n), \quad \text{rank}(X^\top) = \text{rank}(X).$$

Full column rank means $r = n$; full row rank means $r = m$. The term *full rank* alone means $r = \min(m, n)$.

2.3. Dimensions and Rank–Nullity

The rank–nullity identities are

$$\boxed{\text{rank}(X) + \dim \text{Null}(X) = n, \quad \text{rank}(X) + \dim \text{Null}(X^\top) = m.}$$

In particular, the four spaces have the following locations and dimensions:

Subspace	Ambient space	Dimension
$\text{Row}(X)$	$\mathbb{R}^n$	r
$\text{Null}(X)$	$\mathbb{R}^n$	$n - r$
$\text{Col}(X)$	$\mathbb{R}^m$	r
$\text{Null}(X^\top)$	$\mathbb{R}^m$	$m - r$

Row space and column space have equal dimensions, but they are not generally the same subspace. They may even live in different spaces.

2.4. Orthogonal Complements

For a subspace $S \subseteq \mathbb{R}^k$, define

$$S^\perp = \{z \in \mathbb{R}^k : z^\top s = 0 \text{ for every } s \in S\}.$$

Transposition exchanges row and column spaces:

$$\text{Row}(X) = \text{Col}(X^\top), \quad \text{Col}(X) = \text{Row}(X^\top).$$

The fundamental orthogonality relations are

$$\begin{array}{ll} \text{Null}(X) = \text{Row}(X)^\perp, & \text{Row}(X) = \text{Null}(X)^\perp, \\ \text{Null}(X^\top) = \text{Col}(X)^\perp, & \text{Col}(X) = \text{Null}(X^\top)^\perp. \end{array}$$

Proof. A vector v satisfies $Xv = 0$ exactly when $x_i^\top v = 0$ for every row $x_i^\top$. This is equivalent to being orthogonal to every linear combination of the rows, so $\text{Null}(X) = \text{Row}(X)^\perp$. Apply the same argument to $X^\top$ for the second pair. The reverse identities follow from $(S^\perp)^\perp = S$ in finite dimensions. $\square$

Therefore,

$$\mathbb{R}^n = \text{Row}(X) \oplus \text{Null}(X), \quad \mathbb{R}^m = \text{Col}(X) \oplus \text{Null}(X^\top).$$

Here $\oplus$ denotes a direct sum; these particular sums are orthogonal. Every $w \in \mathbb{R}^n$ has a unique decomposition

$$w = w_{\text{row}} + w_{\text{null}}, \quad Xw = Xw_{\text{row}}.$$

Thus the nullspace component is invisible to multiplication by X . Restricted to $\text{Row}(X)$, the map $w \mapsto Xw$ is a bijection onto $\text{Col}(X)$: its kernel is $\text{Row}(X) \cap \text{Null}(X) = \{0\}$, and both spaces have dimension r .

2.5. Worked Example from the Handout

Consider

$$X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \text{rref}(X) = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}.$$

There are two pivots, so $r = 2$. We obtain

$$\begin{aligned} \text{Row}(X) &= \text{span} \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \right\} \subseteq \mathbb{R}^3, \\ \text{Col}(X) &= \text{span} \left\{ \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 5 \end{pmatrix} \right\} = \mathbb{R}^2, \\ \text{Null}(X) &= \text{span} \left\{ \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \right\}, \\ \text{Null}(X^\top) &= \{0\} \subseteq \mathbb{R}^2. \end{aligned}$$

For the nullspace calculation, $Rv = 0$ gives $v_1 = v_3$ and $v_2 = -2v_3$. The dimensions are 2, 2, 1, 0 for row space, column space, nullspace, and left nullspace, respectively. The row space is the plane

$$\{v \in \mathbb{R}^3 : v_1 - 2v_2 + v_3 = 0\},$$

whose normal vector spans the nullspace.

2.6. Connection to Solving Linear Systems

For a given $b \in \mathbb{R}^m$,

$$Xw = b \text{ is solvable} \iff b \in \text{Col}(X) \iff b \perp \text{Null}(X^\top).$$

If w_0 is one solution, all solutions are

$$w = w_0 + v, \quad v \in \text{Null}(X).$$

Indeed, $Xw = Xw_0$ exactly when $X(w - w_0) = 0$. A consistent system has a unique solution exactly when $\text{rank}(X) = n$. Every $b \in \mathbb{R}^m$ is attainable exactly when $\text{rank}(X) = m$. Thus existence and uniqueness are different questions.

2.7. Additional Rank Properties

For compatible matrices A and B ,

$$\begin{aligned}\text{Col}(AB) &\subseteq \text{Col}(A), \\ \text{Row}(AB) &\subseteq \text{Row}(B), \\ \text{rank}(AB) &\leq \min\{\text{rank}(A), \text{rank}(B)\}.\end{aligned}$$

The first two inclusions follow by viewing the columns and rows of a product as linear combinations [1]. For matrices A, B of the same size,

$$\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B),$$

since $\text{Col}(A + B) \subseteq \text{Col}(A) + \text{Col}(B)$. If $P \in \mathbb{R}^{m \times m}$ and $Q \in \mathbb{R}^{n \times n}$ are invertible, then

$$\text{rank}(PXQ) = \text{rank}(X).$$

This follows by applying the product-rank bound both to PXQ and to $X = P^{-1}(PXQ)Q^{-1}$.

3. How to Determine Whether a Matrix Is Invertible

3.1. Equivalent Conditions for a Square Matrix

A square matrix $A \in \mathbb{R}^{n \times n}$ is invertible if a matrix A^{-1} exists with

$$A^{-1}A = AA^{-1} = I_n.$$

The inverse, when it exists, is unique. The following conditions are all equivalent [1][2]:

- A is invertible;
- $\text{rank}(A) = n$;
- $\text{Null}(A) = \{0\}$;
- the columns of A are linearly independent;
- the rows of A are linearly independent;
- $\text{Col}(A) = \text{Row}(A) = \mathbb{R}^n$;
- $\text{rref}(A) = I_n$;
- $\det(A) \neq 0$;
- 0 is not an eigenvalue of A ;
- $Aw = b$ has exactly one solution for every $b \in \mathbb{R}^n$.

The nullspace, rank, and system-solving conditions are linked by rank-nullity. The eigenvalue condition follows because $Av = 0$ with $v \neq 0$ is exactly the definition of a zero eigenvalue. No diagonalizability assumption is needed.

Important: being square is necessary for an ordinary two-sided inverse, but being square alone does not guarantee invertibility. For a non-square matrix, discuss full row rank, full column rank, or a one-sided inverse instead.

3.2. Useful Special Cases

A diagonal or triangular square matrix is invertible exactly when all of its diagonal entries are nonzero, since its determinant is their product. An orthogonal square matrix Q satisfies $Q^\top Q = I$, so $Q^{-1} = Q^\top$. A symmetric PD matrix is invertible, because $Av = 0$ for nonzero v would contradict $v^\top Av > 0$.

Symmetry and positive semidefiniteness alone do not imply invertibility: $\text{diag}(1, 0)$ is symmetric PSD but singular. Conversely, invertibility does not imply positive definiteness: $\text{diag}(1, -1)$ is invertible but indefinite.

For invertible square matrices of the same size,

$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A^\top)^{-1} = (A^{-1})^\top.$$

There is no general identity $(A + B)^{-1} = A^{-1} + B^{-1}$.

3.3. Gram Matrices: Symmetry, Nullspaces, and Rank

For any $X \in \mathbb{R}^{m \times n}$, define

$$G = X^\top X \in \mathbb{R}^{n \times n}, \quad H = XX^\top \in \mathbb{R}^{m \times m}.$$

Both matrices are symmetric PSD. For example,

$$G^\top = G, \quad v^\top Gv = \|Xv\|_2^2 \geq 0.$$

The key nullspace identities from the handout, and their transpose counterparts, are

$$\boxed{\text{Null}(X^\top X) = \text{Null}(X), \quad \text{Null}(XX^\top) = \text{Null}(X^\top).}$$

Proof. If $Xv = 0$, then $X^\top Xv = 0$. Conversely, if $X^\top Xv = 0$, then

$$0 = v^\top X^\top Xv = \|Xv\|_2^2,$$

so $Xv = 0$. Replace X by $X^\top$ for the second identity. □

By rank-nullity and orthogonal complements,

$$\begin{aligned} \text{rank}(X^\top X) &= \text{rank}(XX^\top) = \text{rank}(X), \\ \text{Row}(X^\top X) &= \text{Col}(X^\top X) = \text{Row}(X), \\ \text{Row}(XX^\top) &= \text{Col}(XX^\top) = \text{Col}(X). \end{aligned}$$

For example, $\text{Row}(X^\top X) = \text{Null}(X^\top X)^\perp = \text{Null}(X)^\perp = \text{Row}(X)$; symmetry then gives its column space.

3.4. When Is a Gram Matrix Invertible?

The precise tests are

$$\boxed{X^\top X \text{ invertible} \iff X^\top X \succ 0 \iff \text{rank}(X) = n,}$$

$$\boxed{XX^\top \text{ invertible} \iff XX^\top \succ 0 \iff \text{rank}(X) = m.}$$

These follow because $v^\top X^\top Xv = \|Xv\|_2^2$ vanishes exactly on $\text{Null}(X)$ [?, 3].

If $m < n$, then $\text{rank}(X) \leq m < n$, so $X^\top X$ is always singular. If $m \geq n$, full column rank is possible but not guaranteed. For example, the tall matrix

$$X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix}$$

has dependent columns, so its Gram matrix is singular. **Tall is a shape condition; full column rank is the actual test.**

For the handout's 2×3 example,

$$X^\top X = \begin{pmatrix} 17 & 22 & 27 \\ 22 & 29 & 36 \\ 27 & 36 & 45 \end{pmatrix}, \quad XX^\top = \begin{pmatrix} 14 & 32 \\ 32 & 77 \end{pmatrix}.$$

The first has rank $2 < 3$ and is singular. The second has determinant $14 \cdot 77 - 32^2 = 54 \neq 0$ and is invertible.

3.5. One-Sided Inverses (Supplement)

If X has full column rank, then

$$X_{\text{left}}^{-1} = (X^\top X)^{-1} X^\top, \quad X_{\text{left}}^{-1} X = I_n.$$

If X has full row rank, then

$$X_{\text{right}}^{-1} = X^\top (X X^\top)^{-1}, \quad X X_{\text{right}}^{-1} = I_m.$$

The reverse products are projection matrices, not generally identity matrices. A strictly rectangular matrix cannot have a two-sided inverse [2].

4. Common Matrix and Vector Differentiation Formulas

4.1. Fix the Gradient Convention First

For a scalar-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, define its gradient as a column vector:

$$\nabla_w f = \begin{pmatrix} \partial f / \partial w_1 \\ \vdots \\ \partial f / \partial w_n \end{pmatrix}, \quad \mathrm{d}f = (\nabla_w f)^\top \mathrm{d}w.$$

For $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, define the Jacobian by

$$(J_g)_{ij} = \frac{\partial g_i}{\partial w_j}, \quad J_g \in \mathbb{R}^{m \times n}, \quad \mathrm{d}g = J_g \mathrm{d}w.$$

For a scalar function F of a matrix $W \in \mathbb{R}^{p \times q}$,

$$(\nabla_W F)_{ij} = \frac{\partial F}{\partial W_{ij}}, \quad \mathrm{d}F = \mathrm{tr}((\nabla_W F)^\top \mathrm{d}W).$$

Thus the gradient has the same shape as the variable. Unless stated otherwise, matrix entries are treated as independent coordinates, and all other quantities are fixed [4].

4.2. Differential Rules

For compatible differentiable matrix-valued quantities U, V , a constant scalar a , and constant matrices A, B ,

$$\begin{aligned} \mathrm{d}(U + V) &= \mathrm{d}U + \mathrm{d}V, \\ \mathrm{d}(aU) &= a\mathrm{d}U, \\ \mathrm{d}(UV) &= (\mathrm{d}U)V + U(\mathrm{d}V), \\ \mathrm{d}(U^\top) &= (\mathrm{d}U)^\top, \\ \mathrm{d}(AWB) &= A(\mathrm{d}W)B, \\ \mathrm{d}\mathrm{tr}(U) &= \mathrm{tr}(\mathrm{d}U). \end{aligned}$$

The product rule preserves the order of the factors.

For $f(w) = \ell(g(w))$, the chain rule is

$$\nabla_w f = J_g(w)^\top \nabla_z \ell(z) \Big|_{z=g(w)}.$$

In particular, if $g(w) = Xw - Y$, then $J_g = X$. For vector functions $u(w)$ and $v(w)$ of equal output dimension,

$$\nabla_w (u(w)^\top v(w)) = J_u(w)^\top v(w) + J_v(w)^\top u(w).$$

4.3. Scalar Functions of a Vector

For constant $a, b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, and $c \in \mathbb{R}$,

$$\begin{aligned} \nabla_w c &= 0, \\ \nabla_w (a^\top w) &= a, \\ \nabla_w (w^\top w) &= 2w, \\ \nabla_w \|w - a\|_2^2 &= 2(w - a), \\ \nabla_w (w^\top A w) &= (A + A^\top)w, \\ \nabla_w^2 (w^\top A w) &= A + A^\top. \end{aligned}$$

If A is symmetric, these simplify to

$$\boxed{\nabla_w(w^\top Aw) = 2Aw, \quad \nabla_w\left(\frac{1}{2}w^\top Aw - b^\top w\right) = Aw - b.}$$

Do not use $2Aw$ without checking symmetry.

Derivation of the quadratic-form gradient. Using the product rule,

$$\begin{aligned} d(w^\top Aw) &= (dw)^\top Aw + w^\top Adw \\ &= w^\top (A^\top + A)dw \\ &= ((A + A^\top)w)^\top dw. \end{aligned}$$

Compare with $df = (\nabla_w f)^\top dw$. □

For constant $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^m$,

$$\begin{aligned} \nabla_w \|Xw - Y\|_2^2 &= 2X^\top (Xw - Y), \\ \nabla_w^2 \|Xw - Y\|_2^2 &= 2X^\top X. \end{aligned}$$

More generally, if $S \in \mathbb{R}^{m \times m}$ is constant and symmetric,

$$\nabla_w ((Xw - Y)^\top S (Xw - Y)) = 2X^\top S (Xw - Y).$$

For nonsymmetric S , replace $2S$ by $S + S^\top$.

4.4. Norms That Are Not Everywhere Differentiable (Optional)

For $w \neq 0$,

$$\nabla_w \|w\|_2 = \frac{w}{\|w\|_2}.$$

The Euclidean norm is not differentiable at $w = 0$, although its square is differentiable everywhere. For the ℓ_1 norm, the derivative in coordinate i is $\text{sign}(w_i)$ when $w_i \neq 0$. At zero coordinates, use a subgradient instead:

$$g \in \partial \|w\|_1 \iff g_i \in \begin{cases} \{1\}, & w_i > 0, \\ \{-1\}, & w_i < 0, \\ [-1, 1], & w_i = 0. \end{cases}$$

For a convex function, g is a subgradient at w if $f(z) \geq f(w) + g^\top (z - w)$ for every z . A subgradient is not an ordinary gradient at a nondifferentiable point.

4.5. Scalar Functions of a Matrix (Supplement)

Let $W, C \in \mathbb{R}^{p \times q}$, $a \in \mathbb{R}^p$, $b \in \mathbb{R}^q$, $A \in \mathbb{R}^{p \times p}$, and $B \in \mathbb{R}^{q \times q}$. All quantities except W are constant. Then

$$\begin{aligned} \nabla_W \text{tr}(C^\top W) &= C, \\ \nabla_W (a^\top W b) &= ab^\top, \\ \nabla_W \frac{1}{2} \|W - C\|_F^2 &= W - C, \\ \nabla_W \text{tr}(W^\top A W) &= (A + A^\top)W, \\ \nabla_W \text{tr}(W^\top A W B) &= A W B + A^\top W B^\top. \end{aligned}$$

For square W , also

$$\nabla_W \text{tr}(W) = I, \quad \nabla_W \text{tr}(W^2) = 2W^\top.$$

Notice the distinction between $\nabla_W \text{tr}(W^2) = 2W^\top$ and $\nabla_W \text{tr}(W^\top W) = 2W$.

For constant $L \in \mathbb{R}^{s \times p}$, $R \in \mathbb{R}^{q \times t}$, and $T \in \mathbb{R}^{s \times t}$,

$$\nabla_W \frac{1}{2} \|LWR - T\|_F^2 = L^\top (LWR - T) R^\top.$$

To derive it, set $E = LWR - T$. Then

$$\begin{aligned} d\left(\frac{1}{2} \|E\|_F^2\right) &= \text{tr}(E^\top dE) \\ &= \text{tr}(E^\top L(dW)R) \\ &= \text{tr}((L^\top ER^\top)^\top dW). \end{aligned}$$

Read off the coefficient using the matrix-gradient convention.

4.6. Inverses and Determinants (Optional Reference)

Let $A \in \mathbb{R}^{n \times n}$ be invertible, and write $A^{-\top} = (A^{-1})^\top$. Then

$$\begin{aligned} d(A^{-1}) &= -A^{-1}(dA)A^{-1}, \\ d \det(A) &= \det(A) \text{tr}(A^{-1}dA), \\ d \log |\det(A)| &= \text{tr}(A^{-1}dA). \end{aligned}$$

The corresponding scalar-function gradients are

$$\nabla_A \det(A) = \det(A) A^{-\top}, \quad \nabla_A \log |\det(A)| = A^{-\top}.$$

When $\det(A) > 0$, the absolute value may be omitted. For symmetric PD A , the Frobenius gradient of $\log \det(A)$ is A^{-1} . These are matrix gradients, not derivatives with respect to a vector containing only the independent upper-triangular entries.

The inverse rule follows by differentiating $AA^{-1} = I$:

$$(dA)A^{-1} + Ad(A^{-1}) = 0.$$

Multiplying on the left by A^{-1} gives the result. Do not replace $-A^{-1}(dA)A^{-1}$ by $-A^{-2}dA$ unless the required factors commute. Further identities appear in [4].

5. Putting Everything Together: Linear Regression

5.1. Summation Form and Matrix Form

Using the handout's notation,

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

Expanding the quadratic gives

$$\begin{aligned} L(w) &= (Xw - Y)^\top (Xw - Y) \\ &= w^\top X^\top Xw - 2Y^\top Xw + Y^\top Y. \end{aligned}$$

Because $X^\top X$ is symmetric,

$$\nabla_w L(w) = 2X^\top (Xw - Y), \quad \nabla_w^2 L(w) = 2X^\top X \succeq 0.$$

The coordinate calculation gives the same result:

$$\frac{\partial L}{\partial w_j} = 2 \sum_{i=1}^m (x_i^\top w - y_i) X_{ij}.$$

The dimensions are a useful check: $X^\top$ is $n \times m$, the residual is in $\mathbb{R}^m$, and the gradient is in $\mathbb{R}^n$ [?].

5.2. Normal Equations and Uniqueness

Because L is a convex quadratic, a zero gradient characterizes a global minimum. Setting the gradient to zero gives the normal equations

$$X^\top X \hat{w} = X^\top Y.$$

These equations always have at least one solution because $X^\top Y \in \text{Row}(X) = \text{Col}(X^\top X)$. If $\text{rank}(X) = n$, the solution is unique:

$$\hat{w} = (X^\top X)^{-1} X^\top Y.$$

If $\text{rank}(X) < n$, this inverse formula is unavailable, but least squares still has minimizers. If $\hat{w}_0$ is one minimizer, then all minimizers are

$$\hat{w}_0 + v, \quad v \in \text{Null}(X).$$

Indeed, all solutions of the normal equations differ by an element of $\text{Null}(X^\top X) = \text{Null}(X)$. Thus the coefficients need not be unique, while the fitted vector $X\hat{w}$ on the given design matrix is unique.

5.3. Geometric Interpretation

Let $\hat{Y} = X\hat{w}$ and $e = Y - \hat{Y}$. The normal equations imply

$$X^\top e = 0, \quad e \in \text{Null}(X^\top) = \text{Col}(X)^\perp.$$

Therefore,

$$Y = \underbrace{\hat{Y}}_{\in \text{Col}(X)} + \underbrace{e}_{\in \text{Null}(X^\top)}$$

is the orthogonal decomposition of Y into a fitted component and a residual [3]. When X has full column rank,

$$\hat{Y} = PY, \quad P = X(X^\top X)^{-1}X^\top, \quad P^\top = P, \quad P^2 = P.$$

This P is the orthogonal projector onto $\text{Col}(X)$.

5.4. Wide Matrices and Ridge Regression (Optional)

A wide matrix does not make regression impossible. If $m < n$ and X has full row rank, every $Y \in \mathbb{R}^m$ can be fitted exactly. There are infinitely many coefficient vectors, but the unique one with minimum Euclidean norm is

$$\hat{w}_{\min} = X^\top (X X^\top)^{-1} Y.$$

It lies in $\text{Row}(X)$; every other solution adds an orthogonal nullspace component and therefore increases the squared norm.

Another approach is ridge regression, with a penalty on all coordinates:

$$L_\lambda(w) = \|Xw - Y\|_2^2 + \lambda \|w\|_2^2, \quad \lambda > 0.$$

Then

$$\begin{aligned} \nabla_w L_\lambda(w) &= 2X^\top (Xw - Y) + 2\lambda w, \\ \nabla_w^2 L_\lambda(w) &= 2(X^\top X + \lambda I_n). \end{aligned}$$

For every nonzero v ,

$$v^\top (X^\top X + \lambda I_n) v = \|Xv\|_2^2 + \lambda \|v\|_2^2 > 0.$$

Consequently the matrix is PD regardless of the rank of X , giving

$$\hat{w}_\lambda = (X^\top X + \lambda I_n)^{-1} X^\top Y.$$

This changes the objective; it is not simply a different formula for unregularized least squares.

References

- [1] Gilbert Strang / MIT OpenCourseWare. *Multiplication and Inverse Matrices*. 18.06SC Linear Algebra, Fall 2011, lecture summary.
- [2] Gilbert Strang / MIT OpenCourseWare. *Left and Right Inverses; Pseudoinverse*. 18.06SC Linear Algebra, Fall 2011, lecture summary.
- [3] Gilbert Strang / MIT OpenCourseWare. *Projection Matrices and Least Squares*. 18.06SC Linear Algebra, Fall 2011, lecture summary. MIT course materials: <https://ocw.mit.edu/courses/18-06sc-linear-algebra-fall-2011/>.
- [4] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*, November 15, 2012, Section 2. <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.