Linear Algebra Review (cont.)

Finishing the toolkit for the semester

CSCI 567 — Machine Learning • Discussion 02 • Sep 4, 2026

Outline

1. Linear algebra review, continued: norms, span/rank, determinant, matrix calculus (carried over from Discussion 01), then the essentials of eigenvalues and PSD matrices — just what Lecture 02's linear-regression analysis needs

Discussion 01 ran to time at the cyclic property of the trace — today's linear algebra picks up exactly there. The evaluation-metrics material moves to Discussion 04 (alongside the HW1 review); PyTorch/Colab setup is deferred to a dedicated PyTorch/NumPy session later in the semester — the homeworks carry the hands-on practice for now.

1. Linear algebra review, continued

Norms, rank, determinant, matrix calculus, PSD

Norms: what makes something a valid “length”

A function $f : \mathbb{R}^n \to \mathbb{R}$ is a norm if, for all $x, y \in \mathbb{R}^n$:

Non-negativity: $f(x) \geq 0$

Definiteness: $f(x) = 0 \iff x = 0$

Homogeneity: $f(tx) = |t| f(x)$

Triangle inequality: $f(x + y) \leq f(x) + f(y)$

Examples of norms

$$\ell_2 \text{ (Euclidean): } \|x\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{x^\top x}$$

$$\ell_1\text{: } \|x\|_1 = \sum_i |x_i|$$

$$\ell_\infty\text{: } \|x\|_\infty = \max_i |x_i|$$

ℓ_2 and ℓ_1 return in Lecture 03 as ridge and lasso regularizers.

The ℓ_p family, and the Frobenius norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \qquad p \geq 1$$

Matrix norm (Frobenius): $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)}$

Span, linear independence, rank

Span$(\{x_1, \ldots, x_n\}) = \{v : v = \sum_i \alpha_i x_i,\ \alpha_i \in \mathbb{R}\}$ — column span = column space

Linearly dependent: some x_n is a combination of the rest, $x_n = \sum_{i<n} \alpha_i x_i$

Rank: largest number of linearly independent columns (= rows); $\mathrm{rank}(A) \leq \min(m, n)$

Rank and inverse: properties

$\mathrm{rank}(A) = \mathrm{rank}(A^\top)$; for $A \in \mathbb{R}^{m \times p}, B \in \mathbb{R}^{p \times n}$: $\mathrm{rank}(AB) \leq \min(\mathrm{rank}A, \mathrm{rank}B)$

A^{-1} is the unique matrix with $A^{-1}A = I_n = AA^{-1}$; exists iff A is full rank ($\mathrm{rank}(A) = n$)

Determinant: geometric intuition

$|A|$ is the (signed) **volume** of the parallelepiped spanned by A's columns, $S = \{v : v = \sum_i \alpha_i a_i,\ 0 \leq \alpha_i \leq 1\}$

2×2: $\left|\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right| = a_{11}a_{22} - a_{12}a_{21}$.

Determinant: the recursive formula

Let $A_{\setminus i,\setminus j}$ be A with row i and column j deleted (cofactor expansion along any row i, or any column j):

$$|A| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \, |A_{\setminus i,\setminus j}| \quad (\forall i)$$

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \, |A_{\setminus i,\setminus j}| \quad (\forall j)$$

Base case: $|a_{11}| = a_{11}$ for a 1×1 matrix.

Determinant: properties

$$|A| = |A^\top|, \quad |AB| = |A||B|$$

$$|A| = 0 \iff A \text{ is singular}$$

non-singular: $|A^{-1}| = 1/|A|$

Exercises

Which identities are **not** correct (inverses exist, multiplications legal)?

A. $(AB)^{-1} = B^{-1}A^{-1}$
B. $(I + A)^{-1} = I - A$
C. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
D. $(AB)^{\top} = A^{\top}B^{\top}$

Also: for $x \in \mathbb{R}^n$, what is $\mathrm{rank}(xx^{\top})$?

(B and D are false in general; $\mathrm{rank}(xx^{\top}) = 1$ *for* $x \neq 0$*.)*

Matrix calculus

The tools Lecture 02 needs directly

The gradient

For $f : \mathbb{R}^n \to \mathbb{R}$ (scalar output, vector input):

$$\nabla_x f(x) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$$

Linearity carries over: $\nabla_x(f+g) = \nabla_x f + \nabla_x g$, $\nabla_x(tf) = t\nabla_x f$.

Gradient, Hessian, Jacobian

Hessian ($f : \mathbb{R}^n \to \mathbb{R}$): $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, entries $\partial^2 f / \partial x_i \partial x_j$ — symmetric

Jacobian ($f : \mathbb{R}^n \to \mathbb{R}^m$, vector output): $\nabla_x f(x) \in \mathbb{R}^{m \times n}$, row i is $\nabla_x f_i(x)^\top$

Gradient of a linear function, derived

$f(x) = b^\top x = \sum_i b_i x_i$ for fixed $b \in \mathbb{R}^n$. Then

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_i b_i x_i = b_k.$$

$$\nabla_x \, b^\top x = b$$

Same shape as single-variable calculus: $\frac{\partial (ax)}{\partial x} = a$.

Jacobian of a linear function, derived

$f(x) = Ax$ for fixed $A \in \mathbb{R}^{m \times n}$, so $f_i(x) = a_i^\top x$.

$$\nabla_x f_i(x) = a_i \quad \Longrightarrow \quad \nabla_x f(x) = \begin{pmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{pmatrix} = A$$

Gradient of a quadratic function, step by step

$$f(x) = x^\top A x = \sum_{i,j} A_{ij} x_i x_j \text{ for fixed } A \in \mathbb{R}^{n \times n}.$$

Product rule on $f(x) = g(x)^\top x$ with $g(x) = A^\top x$:

$$\nabla_x f(x) = \nabla_x g(x)^\top x + \nabla_x x^\top g(x) = (A^\top)^\top x + I\, A^\top x = (A + A^\top)x$$

$$\nabla_x\, x^\top A x = (A + A^\top)x$$

Hessian: $\nabla_x^2 f(x) = A + A^\top$ (symmetric A: simplifies to $2Ax$ and $2A$).

Exercises

$f(x) = x^\top Ax + b^\top x$ for $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$. What is $\nabla_x f(x)$?

(Sum rule on the two pieces already derived: $(A + A^\top)x + b$.)

$f(A) = x^\top Ax$ for fixed $x \in \mathbb{R}^n$. What is $\partial f / \partial A$?

($f(A) = \sum_{i,j} x_i x_j A_{ij}$, so $\partial f / \partial A_{ij} = x_i x_j$: $\nabla_A f(A) = xx^\top$.)

Why this toolkit, right now

Lecture 02 sets $\nabla_{\mathbf{w}} \hat{\mathcal{R}}(\mathbf{w}) = 0$ to solve linear regression in closed form — exactly the gradient-of-quadratic-form rule just derived

every gradient descent algorithm this course builds computes $\nabla g(\mathbf{w})$ and moves against it: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla g(\mathbf{w})$

$X^\top X$'s positive semi-definiteness (next frames) and PCA's eigendecomposition (Lecture 06) both build directly on today

Eigenvalues

key idea: sometimes matrices behave simply along **special directions** — are there vectors whose direction does not change under A?

A nonzero vector v is an **eigenvector** of a matrix A if

$$Av = \lambda v.$$

In this case, λ is the **eigenvalue** of v.

Theorem. Every real $n \times n$ matrix has exactly n eigenvalues in $\mathbb{C}$, counting multiplicity (i.e., repetitions)

Symmetric matrices & quadratic forms

Symmetric matrices: from now, we will assume that matrices A are symmetric, meaning $A = A^\top$.

their eigenvalues are all **real**

eigenvectors can be chosen **orthogonal**

A **quadratic form** is a function of the form

$$f(x) = x^\top A x.$$

Appears in least squares, ridge regression, Gaussian likelihoods, etc.

Positive semidefinite (PSD) matrices

a symmetric matrix A is **positive semidefinite (PSD)** if

$$x^\top Ax \geq 0 \quad \forall x$$

Intuition: Ax never "pushes against" the vector x.

Equivalent characterization: symmetric A is PSD if and only if all eigenvalues are ≥ 0

PSD matrices: examples, and positive definiteness

Common examples of PSD matrices:

covariance matrices, Gram matrices
$X^\top X$, the Hessian of a convex loss, ...

a **positive definite** matrix A is symmetric with
$x^\top Ax > 0 \; \forall x \neq 0$ (equivalently, all eigenvalues > 0)

Why this matters, this same week

Lecture 02's convexity check, on *every* loss in this course, is exactly today's PSD test: $\nabla^2 \hat{\mathcal{R}}(\mathbf{w}) \succeq 0$ everywhere $\Leftrightarrow$ $\hat{\mathcal{R}}$ is convex

MSE's Hessian is $X^\top X$ — a Gram matrix, always PSD, exactly the example on this deck's PSD-examples slide

eigenvalues return in Lecture 06: PCA finds the eigenvectors of a covariance matrix, ranked by their eigenvalues

References

- Content this section: Vatsal Sharan, CSCI 567 Discussions 1 and 2 – Linear Algebra Review I and II (Spring 2026), used with permission (norms through matrix calculus carried over from our Discussion 01; the eigenvalue/PSD frames are Discussion 2 material, reduced).
- Vatsal Sharan's Discussion 1 slides are adapted from Nandita Bhaskhar's discussion slides for CS229 at Stanford.

Recap

norms, rank/determinant, matrix calculus, and the PSD test are exactly the tools Lecture 02's linear-regression analysis leans on — $X^\top X$ PSD is why MSE is convex and its stationary point is the global minimum

eigenvalues return in Lecture 06 (PCA); the PSD test returns whenever a convexity check does

hands-on NumPy/PyTorch practice lives in the homeworks for now — a dedicated PyTorch/NumPy session comes later in the semester