CS 480: Math Refresher

Tutorial Session

Haochen Sun

January 16, 2026

Quiz Review

Problem: Compute the derivative of $f(x) = e^{ax}$, where $a \in \mathbb{R}$.

Quiz Review

Problem: Compute the derivative of $f(x) = e^{ax}$, where $a \in \mathbb{R}$.

Solution: Using the property $\frac{d}{du} e^u = e^u$ and the **chain rule**:

Quiz Review

Problem: Compute the derivative of $f(x) = e^{ax}$, where $a \in \mathbb{R}$.

Solution: Using the property $\frac{d}{du}e^u = e^u$ and the **chain rule**:

$$\begin{aligned} f'(x) &= \frac{d}{dx}\left(e^{ax}\right) \\ &= e^{ax} \cdot \frac{d}{dx}(ax) \\ &= a \cdot e^{ax} \end{aligned}$$

The Chain Rule: Definition

Given a composite function $y = f(g(x))$, let $u = g(x)$. The derivative of y with respect to x is:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

In **prime notation**, this is expressed as:

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

Note: This rule allows us to decompose complex gradients into a product of simpler local derivatives.

Exercises

Find the derivative $f'(x)$ for the following functions:

1. $f(x) = (x - a)^2$
2. $f(x) = (g(x) - a)^2$
3. $f(x) = \ln(g(x))$
4. $f(x) = e^{-\frac{x^2}{2}}$

What is a Vector?

A **vector** $\mathbf{x} \in \mathbb{R}^d$ is an ordered list of d real numbers.

- In ML, we typically represent vectors as **column vectors**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- **Geometric View:** A point in d-dimensional space or an arrow from the origin $(0, \ldots, 0)$ to that point.
- **Data View:** A feature vector representing one sample (e.g., $x_1 =$ age, $x_2 =$ income).

Vector Norms

A **norm** $\|\mathbf{x}\|$ measures the "length" or "size" of a vector.

1. L_1 **Norm** (Manhattan): Sum of absolute values.

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

2. L_2 **Norm** (Euclidean): Standard distance from origin.

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$$

3. L_∞ **Norm** (Max): The largest absolute component.

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

Dot Product and Orthogonality

- **Dot Product:** For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the dot product is:

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta$$

- **Orthogonality:** Two vectors are **orthogonal** ($\mathbf{x} \perp \mathbf{y}$) if:

$$\mathbf{x}^\top \mathbf{y} = 0$$

This means they are perpendicular to each other ($\theta = 90^\circ$).

Hyperplanes

A **hyperplane** in $\mathbb{R}^d$ is a set of points defined by:

$$\mathcal{H} = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} = b\}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the **normal vector** (perpendicular to the plane) and $b \in \mathbb{R}$ is the bias.

Key Properties:

- If $b = 0$, the hyperplane **passes through the origin**.
- If $b \neq 0$, the hyperplane is shifted away from the origin.
- The vector $\mathbf{w}$ determines the orientation of the hyperplane.

Halfspaces

A hyperplane divides the space $\mathbb{R}^d$ into two **halfspaces**.

An **closed halfspace** is defined as:

$$\mathcal{S} = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} \geq b\}$$

- In Machine Learning, we often use this for **binary classification**:
  - Class 1: $\mathbf{w}^\top \mathbf{x} - b \geq 0$
  - Class 2: $\mathbf{w}^\top \mathbf{x} - b < 0$
- The hyperplane $\mathbf{w}^\top \mathbf{x} = b$ acts as the **decision boundary**.

What is a Matrix?

A **matrix** $A \in \mathbb{R}^{m \times n}$ is a rectangular array of real numbers with m rows and n columns:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

In Machine Learning:

- Often, a matrix represents a **dataset** $X \in \mathbb{R}^{n \times d}$, where:
  - Each **row** $\mathbf{x}_i^\top$ is a data sample (e.g., one person).
  - Each **column** is a feature (e.g., height, weight).
- **Transpose:** $A^\top$ swaps rows and columns ($a_{ij}^\top = a_{ji}$).

Matrix-Vector Product

Given $A \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, the product $\mathbf{y} = A\mathbf{x}$ results in $\mathbf{y} \in \mathbb{R}^m$.

Two ways to view the product:

1. **Row View:** Each y_i is the dot product of the i-th row of A with $\mathbf{x}$:
$$y_i = \mathbf{a}_{i,\cdot}^\top \mathbf{x}$$
2. **Column View (Linear Combination):** $\mathbf{y}$ is a weighted sum of the columns of A:
$$A\mathbf{x} = x_1 \mathbf{a}_{\cdot,1} + x_2 \mathbf{a}_{\cdot,2} + \cdots + x_n \mathbf{a}_{\cdot,n}$$

This view is crucial for understanding concepts like **span** and **column space**.

Matrix-Matrix Product

Given $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their product $C = AB$ is an $m \times p$ matrix.

The element at row i and column j is:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Important Properties:

- **Dimension Match:** Inner dimensions must agree: $(m \times \underline{n}) \times (\underline{n} \times p)$.
- **Non-commutative:** In general, $AB \neq BA$.
- **Associative:** $A(BC) = (AB)C$.
- **Distributive:** $A(B + C) = AB + AC$.

The Gradient

For a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$, the **gradient** $\nabla f(\mathbf{x})$ is the vector of all first-order partial derivatives:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \in \mathbb{R}^d$$

Key Intuitions:

- The gradient points in the direction of the **steepest ascent**.
- In ML, we use $-\nabla f(\mathbf{x})$ to perform **Gradient Descent**.

The Jacobian

When we have a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, its derivative is represented by the **Jacobian matrix** $J \in \mathbb{R}^{m \times n}$:

$$J = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- The i-th **row** of the Jacobian is the transpose of the gradient of the i-th component function: $\nabla f_i(\mathbf{x})^\top$.
- It describes how every component of the output changes with respect to every component of the input.

The Hessian

The **Hessian matrix** $\nabla^2 f(\mathbf{x})$ contains the second-order partial derivatives of a scalar function $f : \mathbb{R}^d \to \mathbb{R}$:

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \implies H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- **Symmetry:** If the second derivatives are continuous, H is symmetric ($H = H^\top$).
- **Curvature:** While the gradient tells us the direction of slope, the Hessian tells us the **curvature** (how the slope is changing).

Multivariate Chain Rule

Suppose $\mathbf{y} = \mathbf{g}(\mathbf{x})$ and $\mathbf{z} = \mathbf{f}(\mathbf{y})$. To find the derivative of the composition $\mathbf{f}(\mathbf{g}(\mathbf{x}))$ with respect to $\mathbf{x}$:

In Matrix Form:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

- This is a **matrix multiplication** of two Jacobians!
- Dimensions: If $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^k, \mathbf{z} \in \mathbb{R}^m$, then:

$$\underbrace{J_{comp}}_{m \times n} = \underbrace{J_f}_{m \times k} \times \underbrace{J_g}_{k \times n}$$

- This is exactly how **Backpropagation** works in Deep Learning.

Multivariate Chain Rule: Exercise 1

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

Multivariate Chain Rule: Exercise 1

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

Solution: Let $\mathbf{u} = A\mathbf{x} - \mathbf{b}$. Then $f = \|\mathbf{u}\|_2^2 = \mathbf{u}^\top \mathbf{u}$.

- By the chain rule: $\nabla_{\mathbf{x}} f = \left(\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\right)^\top \nabla_{\mathbf{u}} f$
- We know: $\nabla_{\mathbf{u}}(\mathbf{u}^\top \mathbf{u}) = 2\mathbf{u}$ and $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = A$

Multivariate Chain Rule: Exercise 1

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

Solution: Let $\mathbf{u} = A\mathbf{x} - \mathbf{b}$. Then $f = \|\mathbf{u}\|_2^2 = \mathbf{u}^\top \mathbf{u}$.

- By the chain rule: $\nabla_{\mathbf{x}} f = \left(\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\right)^\top \nabla_{\mathbf{u}} f$
- We know: $\nabla_{\mathbf{u}}(\mathbf{u}^\top \mathbf{u}) = 2\mathbf{u}$ and $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = A$

Substituting these back in:

$$\begin{aligned} \nabla f(\mathbf{x}) &= A^\top (2\mathbf{u}) \\ &= 2A^\top (A\mathbf{x} - \mathbf{b}) \end{aligned}$$

Multivariate Chain Rule: Exercise 2

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|_2^2}{2}}$.

Multivariate Chain Rule: Exercise 2

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|_2^2}{2}}$.

Solution: Let $g(\mathbf{x}) = -\frac{1}{2}\|\mathbf{x}\|_2^2 = -\frac{1}{2}\mathbf{x}^\top\mathbf{x}$. Then $f = e^{g(\mathbf{x})}$.

- We know from scalar calculus: $\frac{d}{dg}e^g = e^g$.
- We know from vector calculus: $\nabla_{\mathbf{x}}\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{x}\right) = -\mathbf{x}$.

Multivariate Chain Rule: Exercise 2

Problem: Find $\nabla f(\mathbf{x})$ for $f(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|_2^2}{2}}$.

Solution: Let $g(\mathbf{x}) = -\frac{1}{2}\|\mathbf{x}\|_2^2 = -\frac{1}{2}\mathbf{x}^\top\mathbf{x}$. Then $f = e^{g(\mathbf{x})}$.

- We know from scalar calculus: $\frac{d}{dg}e^g = e^g$.
- We know from vector calculus: $\nabla_{\mathbf{x}}\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{x}\right) = -\mathbf{x}$.

Applying the chain rule:

$$\begin{aligned}\nabla f(\mathbf{x}) &= e^{g(\mathbf{x})} \cdot \nabla_{\mathbf{x}} g(\mathbf{x}) \\ &= e^{-\frac{\|\mathbf{x}\|_2^2}{2}} \cdot (-\mathbf{x}) \\ &= -\mathbf{x}e^{-\frac{\|\mathbf{x}\|_2^2}{2}}\end{aligned}$$

Subspaces

A **subspace** $\mathcal{V}$ of $\mathbb{R}^n$ is a subset that is closed under linear combinations. For any $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and $c, d \in \mathbb{R}$:

$$c\mathbf{u} + d\mathbf{v} \in \mathcal{V}$$

Key Properties:

- Every subspace must contain the **zero vector 0**.
- Common examples: A line or a plane passing through the origin.
- The **span** of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is the smallest subspace containing them.

Basis and Orthogonal Basis

A **basis** for a subspace $\mathcal{V}$ is a set of vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ that:

1. Are **linearly independent**.
2. **Span** the subspace $\mathcal{V}$.

Orthogonal and Orthonormal Basis:

- **Orthogonal Basis:** A basis where every pair of vectors is orthogonal ($\mathbf{b}_i^\top \mathbf{b}_j = 0$ for $i \neq j$).
- **Orthonormal Basis:** An orthogonal basis where every vector has unit length ($\|\mathbf{b}_i\|_2 = 1$).

Why it matters: Any vector $\mathbf{v} \in \mathcal{V}$ can be uniquely written as:

$$\mathbf{v} = c_1\mathbf{b}_1 + c_2\mathbf{b}_2 + \cdots + c_k\mathbf{b}_k$$

If the basis is orthonormal, the coefficients are simple dot products: $c_i = \mathbf{v}^\top \mathbf{b}_i$.

Orthogonal Matrices

A square matrix $Q \in \mathbb{R}^{n \times n}$ is **orthogonal** if its columns are orthonormal:

$$Q^\top Q = QQ^\top = I$$

This implies $Q^{-1} = Q^\top$.

Properties:

- **Preserves Length:** $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for any vector $\mathbf{x}$.
- **Preserves Angles:** $(Q\mathbf{x})^\top (Q\mathbf{y}) = \mathbf{x}^\top \mathbf{y}$.
- **Geometry:** Geometrically, multiplying by Q represents a **rotation** or **reflection** of the coordinate system.

Symmetric Matrices

A square matrix A is **symmetric** if it is equal to its transpose:

$$A = A^\top \quad \text{or} \quad a_{ij} = a_{ji}$$

Common Examples in ML:

- **Covariance Matrix:** $\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top]$, which describes the spread of data.
- **Hessian Matrix:** H, the matrix of second-order partial derivatives (as discussed).

Positive (Semi-)Definite Matrices

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **Positive Semi-Definite (PSD)**, denoted $A \succeq 0$, if for all $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x}^\top A \mathbf{x} \geq 0$$

It is **Positive Definite (PD)**, denoted $A \succ 0$, if $\mathbf{x}^\top A \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

Why this matters for ML:

- **Convexity:** If the Hessian of a function is PSD everywhere, the function is **convex**.
- **Local Minima:** At a critical point ($\nabla f = \mathbf{0}$), if the Hessian is PD, the point is a **strict local minimum**.
- **Variance:** Covariance matrices are always PSD.

Eigenvalues and Eigenvectors

For a square matrix $A \in \mathbb{R}^{n \times n}$, a non-zero vector $\mathbf{v}$ is an **eigenvector** if:

$$A\mathbf{v} = \lambda\mathbf{v}$$

where $\lambda \in \mathbb{R}$ (or $\mathbb{C}$) is the corresponding **eigenvalue**.

Intuition:

- Multiplying $\mathbf{v}$ by A only **scales** the vector; it does not change its direction.
- The eigenvalue λ tells us the scaling factor.

Symmetric Matrices: Spectral Theorem

Symmetric matrices ($A = A^\top$) have special properties that are fundamental to Machine Learning:

1. **Real Eigenvalues:** All eigenvalues $\lambda_1, \ldots, \lambda_n$ are real numbers ($\lambda_i \in \mathbb{R}$).
2. **Orthogonal Eigenvectors:** Eigenvectors corresponding to distinct eigenvalues are **orthogonal**.
   - We can always find an **orthonormal basis** of eigenvectors for $\mathbb{R}^n$.

Eigendecomposition: Every symmetric matrix can be factored as:

$$A = Q\Lambda Q^\top$$

where Q is an **orthogonal matrix** of eigenvectors and Λ is a **diagonal matrix** of eigenvalues.

Eigenvalues and Definiteness

We can characterize the definiteness of a symmetric matrix A entirely by its eigenvalues:

- **Positive Definite ($A \succ 0$):** All $\lambda_i > 0$.
- **Positive Semi-Definite ($A \succeq 0$):** All $\lambda_i \geq 0$.
- **Indefinite:** Has both positive and negative eigenvalues.

Exercise: Eigenvalues of AB and BA

Problem: Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$. If $\lambda \neq 0$ is an eigenvalue of AB, prove that it is also an eigenvalue of BA.

Exercise: Eigenvalues of AB and BA

Problem: Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$. If $\lambda \neq 0$ is an eigenvalue of AB, prove that it is also an eigenvalue of BA.

Proof: By definition, if λ is an eigenvalue of AB, there exists a non-zero eigenvector $\mathbf{v} \in \mathbb{R}^m$ such that:

$$(AB)\mathbf{v} = \lambda \mathbf{v} \tag{1}$$

Exercise: Eigenvalues of AB and BA

Problem: Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$. If $\lambda \neq 0$ is an eigenvalue of AB, prove that it is also an eigenvalue of BA.

Proof: By definition, if λ is an eigenvalue of AB, there exists a non-zero eigenvector $\mathbf{v} \in \mathbb{R}^m$ such that:

$$(AB)\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

Multiply both sides on the left by B:

$$B(AB)\mathbf{v} = B(\lambda\mathbf{v})$$

Exercise: Eigenvalues of AB and BA

Problem: Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$. If $\lambda \neq 0$ is an eigenvalue of AB, prove that it is also an eigenvalue of BA.

Proof: By definition, if λ is an eigenvalue of AB, there exists a non-zero eigenvector $\mathbf{v} \in \mathbb{R}^m$ such that:

$$(AB)\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

Multiply both sides on the left by B:

$$B(AB)\mathbf{v} = B(\lambda\mathbf{v}) \implies (BA)(B\mathbf{v}) = \lambda(B\mathbf{v}) \tag{2}$$

Exercise: Eigenvalues of AB and BA

Problem: Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$. If $\lambda \neq 0$ is an eigenvalue of AB, prove that it is also an eigenvalue of BA.

Proof: By definition, if λ is an eigenvalue of AB, there exists a non-zero eigenvector $\mathbf{v} \in \mathbb{R}^m$ such that:

$$(AB)\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

Multiply both sides on the left by B:

$$B(AB)\mathbf{v} = B(\lambda\mathbf{v}) \implies (BA)(B\mathbf{v}) = \lambda(B\mathbf{v}) \tag{2}$$

Observe that $B\mathbf{v} \neq \mathbf{0}$ (why?)