CSC 311: Introduction to Machine Learning
Tutorial 9 - Math Review

Slides by: Haonan Duan
Instructors: Rahul G. Krishnan and Alice Gao

University of Toronto, Fall 2022
Outline

- Gradients of multivariate functions
- Matrix decomposition
Gradients of vector-valued functions
Vector-valued functions

For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a vector $\mathbf{x} = [x_1, \cdots, x_n]^T \in \mathbb{R}^n$, the corresponding vector of function values is given as:

$$f(\mathbf{x}) = [f_1(\mathbf{x}) \cdots f_m(\mathbf{x})] \in \mathbb{R}^m$$

where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$. 
Partial derivatives

The partial derivative of a vector-valued function \( f : \mathbb{R}^n \to \mathbb{R}^m \) with respect to \( x_i \in \mathbb{R} \) is given as:

\[
\frac{\partial f}{\partial x_i} = \left[ \frac{\partial f_1}{\partial x_i} \ldots \frac{\partial f_m}{\partial x_i} \right] \in \mathbb{R}^m
\]
The collection of all first-order partial derivatives of a vector-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called the Jacobian. The Jacobian $\frac{\partial f(x)}{\partial x}$ is an $m \times n$ matrix, which is defined as:

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}.$$
Example

Given $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$, we define the linear vector-valued function $f$ as:

$$f(x) = Ax$$

- $Q_1$: What is the dimension of $\frac{\partial f(x)}{\partial x}$?
- $Q_2$: Compute $\frac{\partial f(x)}{\partial x}$. 
Since $f : \mathbb{R}^n \to \mathbb{R}^m$, it follows that $\frac{\partial f(x)}{\partial x} \in \mathbb{R}^{m \times n}$.

The first step is to compute each entry of the Jacobian matrix, $\frac{\partial f_i}{\partial x_j}$. From the definition of the matrix decomposition, we know:

$$f_i(x) = \sum_{j=1}^{N} A_{ij} x_j$$

Then each entry $\frac{\partial f_i}{\partial x_j} = A_{ij}$. It follows that:

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = A$$
Often in machine learning, we need to take gradients of matrices with respect to other matrices. The Jacobian in this case will be a multi-dimension tensor.

For example, if we compute the gradient of an $m \times n$ matrix $A$ with respect to a $p \times q$ matrix $B$, the resulting Jacobian $J$ is a four-dimensional tensor $m \times n \times p \times q$. Each entry $J_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}$. 
Exercise

Given a matrix \( \mathbf{R} \in \mathbb{R}^{m \times n} \). We define:

\[
f(\mathbf{R}) = \mathbf{R}^T \mathbf{R}
\]

- **Q1**: What is the dimension of \( \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \)?
- **Q2**: Compute \( \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \).
Recommended resources

Matrix decomposition
We can decompose an integer into its prime factors, e.g., \(12 = 2 \times 2 \times 3\).

Similarly, matrices can be decomposed into product of other matrices.

Examples are Eigendecomposition, SVD, Schur decomposition, LU decomposition, . . . .
An eigenvector of a square matrix $A$ is a nonzero vector $v$ such that multiplication by $A$ only changes the scale of $v$:

$$Av = \lambda v$$

The scalar $\lambda$ is known as the eigenvalue.

If $v$ is an eigenvector of $A$, so is any rescaled vector $sv$. Moreover, $sv$ still has the same eigenvalue. Thus, we constrain the eigenvector to be of unit length.
Compute eigenvalues - characteristic polynomial

- Eigenvalue equation of matrix $A$:

$$Av = \lambda v$$
$$\lambda v - Av = 0$$
$$(\lambda I - A)v = 0$$

- If nonzero solution for $v$ exists, then it must be the case that:

$$det(\lambda I - A) = 0$$

- Unpacking the determinant as a function of $\lambda$, we get a polynomial, called the characteristic polynomial:

$$P_A(\lambda) = det(\lambda I - A) = \lambda^n + c_{n-1}\lambda^{n-1} + \lambda + c_0$$

- Compute eigenvalues of $A \rightarrow$ solve $P_A(\lambda) = 0$
Exercise

Consider the matrix:

\[
A = \begin{bmatrix}
2 & 1 \\
1 & 2 \\
\end{bmatrix}
\]

- What is the characteristic polynomial of \( A \)?
- What are the eigenvalues of \( A \)?
- What are the associated eigenvectors?
Eigendecomposition

- Every symmetric (hermitian) matrix of dimension $n$ has a set of (not necessarily unique) $n$ orthogonal eigenvectors. Furthermore, all eigenvalues are real.

- Every real symmetric matrix $A$ can be decomposed into real-valued eigenvectors and eigenvalues:

$$A = PDP^{-1}$$

- $P$ is an orthogonal matrix of the eigenvectors of $A$, and $D$ is a diagonal matrix of eigenvalues.
Intuitions of Eigendecomposition

- Diagonal matrix allows fast computations of their determinants, powers and inverses.
- Eigendecomposition transforms a matrix into a diagonal form by changing the basis.
Geometric intuitions of eigendecomposition

- Top-left to bottom-left: $P^{-1}$ performs a basis change.
- Bottom-left to bottom-right: $D$ performs a scaling.
- Bottom-right to top-right: $P$ undoes the basis change.
Singular Value Decomposition (SVD)

- If $A$ is not square, eigendecomposition is undefined.
- SVD is a decomposition of the form $A = UDV^T$.
- SVD is more general than eigendecomposition.
- Every real matrix has a SVD.
SVD

- If $A$ is $m \times n$, then $U$ is $m \times m$, $D$ is $m \times n$, and $V$ is $n \times n$.
- $U$ and $V$ are orthogonal matrices, and $D$ is a diagonal matrix (not necessarily square).
- Diagonal entries of $D$ are called singular values of $A$.
- Columns of $U$ are the left singular vectors, and columns of $V$ are the right singular vectors.
SVD and eigendecomposition

- SVD can be interpreted in terms of eigendecomposition.
- Left singular vectors of $A$ are the eigenvectors of $AA^T$.
- Right singular vectors of $A$ are the eigenvectors of $A^TA$.
- Nonzero singular values of $A$ are square roots of eigenvalues of $A^TA$ and $AA^T$. ($A^TA$ and $AA^T$ are semipositive definite, thus their eigenvalues are positive)
Exercise

Compute SVD of the matrix:

\[ A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \]
Exercise

Compute SVD of the matrix:

\[ A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \]

- What is \( AA^T \) and \( A^T A \)?
- Apply eigendecomposition on \( AA^T \) and \( A^T A \)
Rank-r approximation

- Given a matrix $A$, SVD allows us to find its “best” (to be defined) rank-r approximation $A_r$.
- TODO: intuition: compress parameters
- We can write $A = UDV^T$ as $A = \sum_{i=1}^{n} d_i u_i v_i^T$, where $d_i$ are sorted from the largest to the smallest.
- The rank-r approximation $A_r$ is defined as:

$$A = \sum_{i=1}^{r} d_i u_i v_i^T$$

- $A_r$ is the best approximation of rank $r$ by many norms, such as, $L_2$ norm. It means that $\|A - A_r\|_2 \leq \|A - B\|_2$ for any rank $r$ matrix $B$. 

Intro ML  (UofT)  CSC311-Tut9  25 / 26
Exercise

Fine the rank-1 approximation and rank-2 approximation of the matrix:

\[ A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \]