1 Notation and Definitions

1.1 Matrix Vector Multiplication

The application of a matrix $A \in \mathbb{R}^{n\times m}$ to a vector $v \in \mathbb{R}^m$ is the matrix vector multiplication Av. The i-th component of Av is given by the dot product of v with the i-th row of A.

1. Given $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times k}, C \in \mathbb{R}^{k\times l}$, what dimension is the vector v that is multiplied to form $ABCv$? What would the output dimension of this expression be?

2. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$. What are the dimensions of $C = (AB^T)^T$

1.2 Normed Vector Spaces

A norm $\|\cdot\|$: $V \to \mathbb{R}$ on a vector space is defined as any function satisfying:

- $\|v\| \geq 0$ and $\|v\| = 0$ iff $v = \mathbf{0}$ (Positivity)
- $\|av\| = |a|\|v\|$ for $a \in \mathbb{R}$ (Homogeneity)
- $\|v + w\| \leq \|v\| + \|w\|$ (Triangle Inequality)

1. For $p \geq 1$ we define the l^p norm of $v \in \mathbb{R}^n$ to be

$$\|v\|_p = (\sum_i |v_i|^p)^{1/p}$$

   Show that this is a norm.

1.3 Vector 2-Norms

Throughout this course, we will often see the l^2 norm. The l^2 norm is notated as $\|x\|_2$. This is called the "Euclidean norm" because it's how we conventionally calculate "Euclidean distance." It can be expressed using a summation or a dot product. All of the following are equal:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{M} x_i^2} = \sqrt{x \cdot x} = \sqrt{x^T x}$$

1. For $a = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix}$, calculate $\|a - 2\|_2^2$

2. For $z \in \mathbb{R}^2$ and $w \in \mathbb{R}^2$, expand $\|z - w\|_2^2$, writing it in terms of z_1, z_2, w_1, w_2

1.4 Vector 1-Norm

In addition to the l^2 norm, we will also use the l^1 norm denoted as $\|x\|_1$. The $l1$ norm also goes by "Manhattan Distance" and it corresponds to the sum of the magnitudes of the vectors in a space. So, it's the sum of the absolute value of each component of the vector.

$$\|x\|_1 = \sum_{i=1}^{n} \mid x_i \mid$$

1. For $a = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix}$, calculate $\|a - 3\|_1$

1.5 Intuition in L1 vs L2 Norm

Euclidean vs Mnahattan distance comes up often in machine learning, such as in ridge vs. lasso regression. Here are some high level comparisons:

1.5.1 Robustness

Robustness is resistance to outliers. The l^1 norm is considered more robust than the l^2 norm because the l^1 norm takes the cost of outliers linearly while the cost of outliers is squared when using the l^2 norm

2 Derivatives

2.1 Vector Derivatives

Given a function $y = f(\mathbf{x})$, $f : \mathbb{R}^m \to \mathbb{R}$, the derivative $\frac{\partial y}{\partial \mathbf{x}}$ is a m-dimensional **column vector** where each component is the partial derivative of y with respect to the corresponding component of $\mathbf{x}$:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_m} \end{bmatrix}$$

1. Let $y = f(\mathbf{x}) = 3x_1^2 \sin x_2$. What is $\frac{\partial y}{\partial \mathbf{x}}$?

2. Let $y = \mathbf{z}^T\mathbf{x}$ for some $\mathbf{x}, \mathbf{z} \in \mathbb{R}^m$. What is $\frac{\partial y}{\partial \mathbf{x}}$? What is $\frac{\partial y}{\partial \mathbf{z}}$?

3. Let $y = f(\mathbf{x}) = ||\mathbf{x}||_2^2$ and $\mathbf{x} \in \mathbb{R}^3$. What is $\frac{\partial y}{\partial \mathbf{x}}$? Write it in terms of x_1, x_2, x_3. Then, write it in terms of $\mathbf{x}$.

On the other hand, if we are taking the derivative of a vector $\mathbf{y} \in \mathbb{R}^m$ with respect to a scalar $x \in \mathbb{R}$, the derivative is an m-dimensional **row vector** as shown below:

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \dots & \frac{\partial y_m}{\partial x} \end{bmatrix}$$

2.2 Matrix Derivatives

Consider a vector-valued function $\mathbf{y} = f(\mathbf{x})$ where $f : \mathbb{R}^m \to \mathbb{R}^n$. Then the Jacobian is defined as

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} \dots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \dots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \ddots & \\ \frac{\partial y_1}{\partial x_m} & \frac{\partial y_2}{\partial x_m} \dots & \frac{\partial y_n}{\partial x_m} \end{bmatrix}$$

1. Let $\mathbf{y} = f(\mathbf{x}) = (x_1^2 x_2, \sin x_3)$ where $f : \mathbb{R}^3 \to \mathbb{R}^2$. What is $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$?

2. Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{k \times m}$ and $\mathbf{x} \in \mathbb{R}^m$. What is $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$?

2.3 Differentiation Rules

The following rules for vector and matrix differentiation are good to remember. Note here that **a** and **z** are vectors and M is a matrix.

1. $\frac{\partial \mathbf{a}^T \mathbf{z}}{\partial \mathbf{z}} = \mathbf{a}$
2. $\frac{\partial M \mathbf{z}}{\partial \mathbf{z}} = M^T$
3. $\frac{\partial \mathbf{z}^T M \mathbf{z}}{\partial \mathbf{z}} = (M + M^T)\mathbf{z}$

2.4 A Brief Note on Numerator Layout vs Denominator Layout

There are two different layouts to express vector/matrix derivatives, namely the numerator and the denominator layout. In this course, we use the **denominator layout**. These layouts are mostly the same and can easily be switched using transpose operations. To demonstrate this better, some examples are shown below:

	Numerator Layout	**Denominator Layout**
$\frac{\partial y}{\partial \mathbf{x}}$	1-D row vector	1-D column vector
$\frac{\partial \mathbf{y}}{\partial x}$	1-D column vector	1-D row vector
$\frac{\partial \mathbf{a}^T \mathbf{z}}{\partial \mathbf{z}}$	$\mathbf{a}^\mathbf{T}$	$\mathbf{a}$
$\frac{\partial \mathbf{M} \mathbf{z}}{\partial \mathbf{z}}$	$\mathbf{M}$	$\mathbf{M}^\mathbf{T}$

A handy way to distinguish numerator vs denominator layout is to remember that **the layout type corresponds the number of rows in the output matrix**. In a numerator layout, the output matrix has number of rows equal to the size of the numerator, while in a denominator layout, the output matrix has number of rows equal to the size of the denominator.

2.5 Chain Rule

Chain rule in matrix calculus is similar to the usual chain rule in 1-dimension, with the exception that in the **denominator layout**, the order in which we multiply the derivatives is reversed, i.e. successive derivatives are written to the left (this is opposite for the numerator layout).

For instance, in 1-dimension, if we have $y = g(h(k(x)))$, then $\frac{dy}{dx} = \frac{dg}{dh}\frac{dh}{dk}\frac{dk}{dx}$. However, if we are considering the denominator layout vector derivatives, then it is instead

$$y = \frac{\partial k}{\partial x}\frac{\partial h}{\partial k}\frac{\partial g}{\partial h}$$

Technically, order doesn't matter in 1-D cases, but it matters for vectors since the shapes have to align.

1. Let $\mathbf{A} \in \mathbb{R}^{k \times m}$, $\beta \in \mathbb{R}^m$, and $\mathbf{y} \in \mathbb{R}^k$. Find $\arg\min_\beta \|\mathbf{A}\beta - \mathbf{y}\|_2^2$ by considering $\frac{\partial \|\mathbf{A}\beta - \mathbf{y}\|_2^2}{\partial \beta} = 0$.

3 Eigendecomposition and SVD

3.1 Eigenvalues and Eigenvectors

Given a matrix $A \in \mathbb{R}^{n \times m}$, $v \in \mathbb{R}^m$ is an eigenvector corresponding to eigenvalue λ iff

$$Av = \lambda v$$

If for all $x \in \mathbb{R}^m$, $x^T A x \geq 0$, then we say A is positive semidefinite. Furthermore, a matrix is positive semidefinite iff all its eigenvalues are non-negative.

3.1.1 Eigendecomposition

Let $S \in \mathbb{R}^{m \times m}$ be a square matrix with m linearly independent eigenvectors. There must exist an eigendecomposition

$$S = U \Lambda U^{-1}$$

where the columns of U are the eigenvectors and Λ is a diagonal matrix consisting of the corresponding eigenvalues

1. Find the eigendecomposition of matrix $M = \begin{bmatrix} 4 & 3 \\ 2 & -1 \end{bmatrix}$

2. In class, you saw how adding λI to $A^T A$ helps make this matrix invertible. Demonstrate this with $\lambda = 1$ and $M = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

3.2 SVD

Given a matrix $A \in \mathbb{R}^{n \times m}$, the singular value decomposition of A factors A into three matrices:

$$A = USV^T$$

where the matrix S is diagonal with non-negative real entries and the columns of U and V are orthonormal. One advantage of using SVD to Eigendecompisition in practice is that SVD does not require a square matrix.

1. Follow the below steps to compute the SVD of

$$X = \begin{bmatrix} 4 & 4 \\ 3 & -3 \end{bmatrix}$$

   (a) Find $X^T X$ and XX^T

   (b) The singular values of X are the square roots of the eigenvalues of $X^T X$ and XX^T, which have the same eigenvalues. Find the square roots of the eigenvalues of $X^T X$ and XX^T and use them to construct S.

   (c) Now find U and V. The columns of U are eigenvectors of XX^T and the columns of V are eigenvectors of $X^T X$. Also note that the columns of both matrices should be orthonormal.