EECE 5644: Linear Algebra

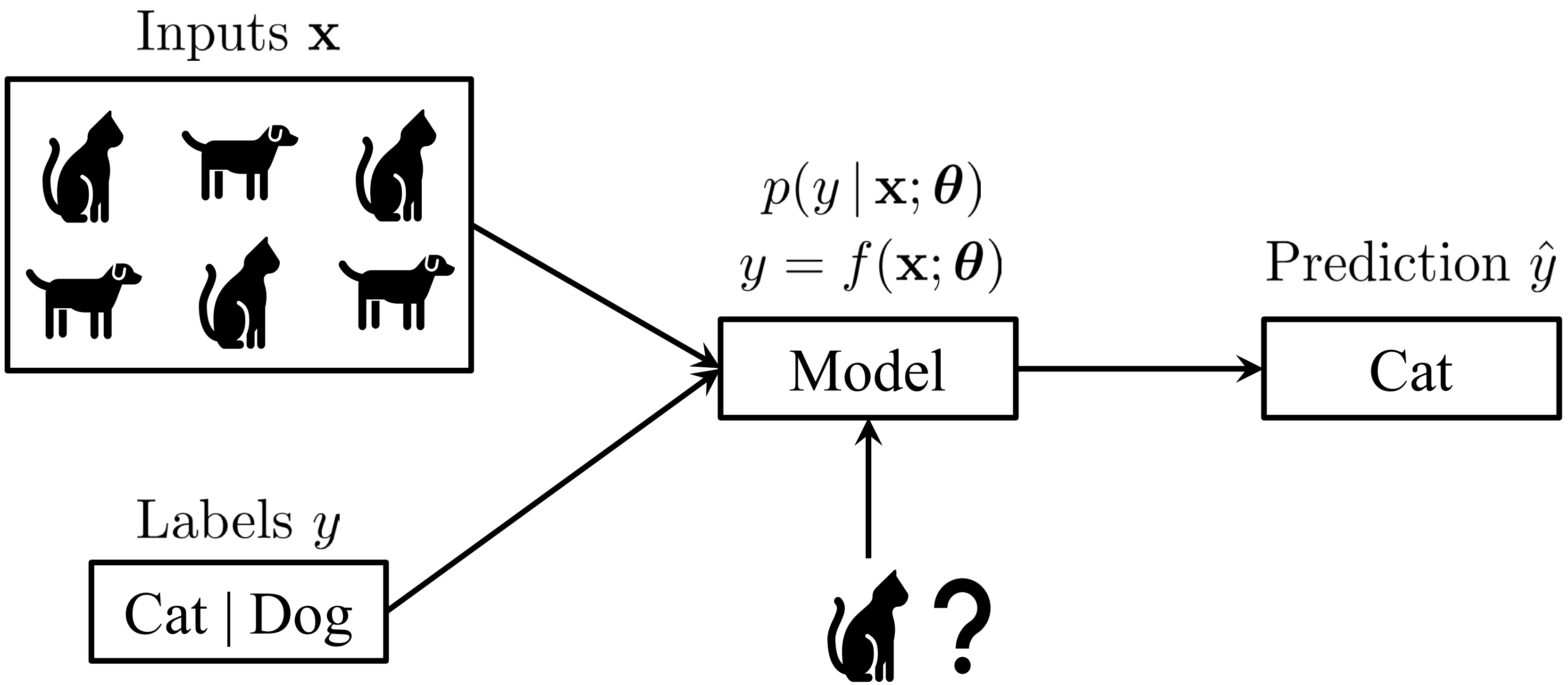
Mark Zolotas

E-mail: m.zolotas@northeastern.edu
Webpage: https://coe.northeastern.edu/people/zolotas-mark/

Tentative Course Outline (Wks. 1-2)

Topics	Dates	Assignments	Additional Reading
~~Course Overview~~ ~~Machine Learning Basics~~	~~07/05~~	**Optional Homework 0** released on Canvas on 07/08 but please do NOT submit on Canvas	~~Chpt. 1~~ ~~Murphy 2012~~
Foundations: Linear Algebra, Probability, Numerical Optimization (Gradient Descent), Regression	07/06-11		Stanford LA Review Stanford Prob. Review Chpt. 8 Murphy 2022
Quick Python Tutorial	07/12	**Homework 1** released on Canvas on 07/15 **Due 07/25**	N/A
Linear Classifier Design, Linear Discriminant Analysis and Principal Component Analysis (PCA)	07/13-14		Chpts. 9.2 & 20.1 Murphy 2022
Bayesian Decision Theory: Empirical Risk Min, Max Likelihood (ML), Max a Posteriori	07/14-15		Chpt. 2 Duda & Hart 2001 Deniz Erdogmus Notes

Intro Recap

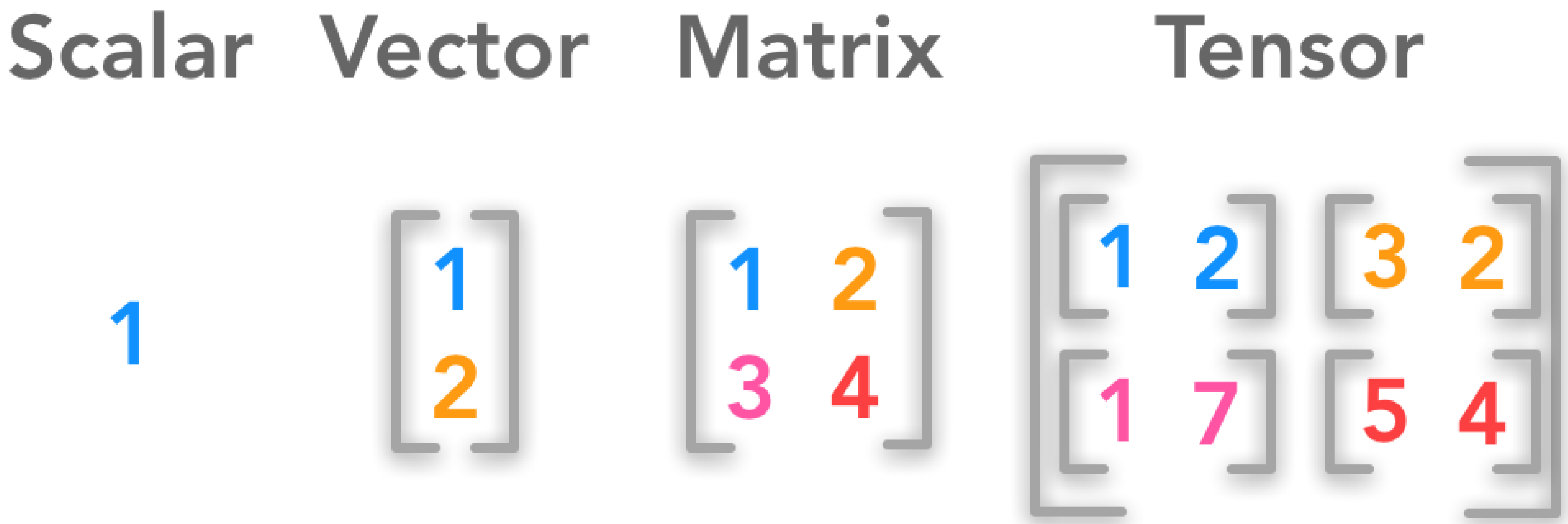


Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, N training samples

Inputs or **features** $\mathbf{x} \in \mathcal{X} = \mathbb{R}^n$

Classification: discrete valued outputs or **labels** $y \in \{1, \ldots, C\}$

Linear Algebra



Sources: https://hadrienj.github.io/posts/Deep-Learning-Book-Series-2.1-Scalars-Vectors-Matrices-and-Tensors/

Scalars

- Are **single** real numbers
- Integers $(7, 42)$, rational numbers $\left(\frac{2}{3}, 0.73\right)$, irrational numbers $\left(\sqrt{7}, \pi\right)$, etc.
- Written in lowercase and *italics*:

$$x, n, d \in \mathbb{R}$$

Vectors

- Are **1-D arrays** of numbers
- Numbers can be binary, integer, real etc.
- Written in lowercase and **bold**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \mathbf{x}^\intercal = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

Matrices

- Are **2-D array** of numbers

- Written in UPPERCASE and **bold**:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \in \mathbb{R}^{2\times 3}$$

Row

Column

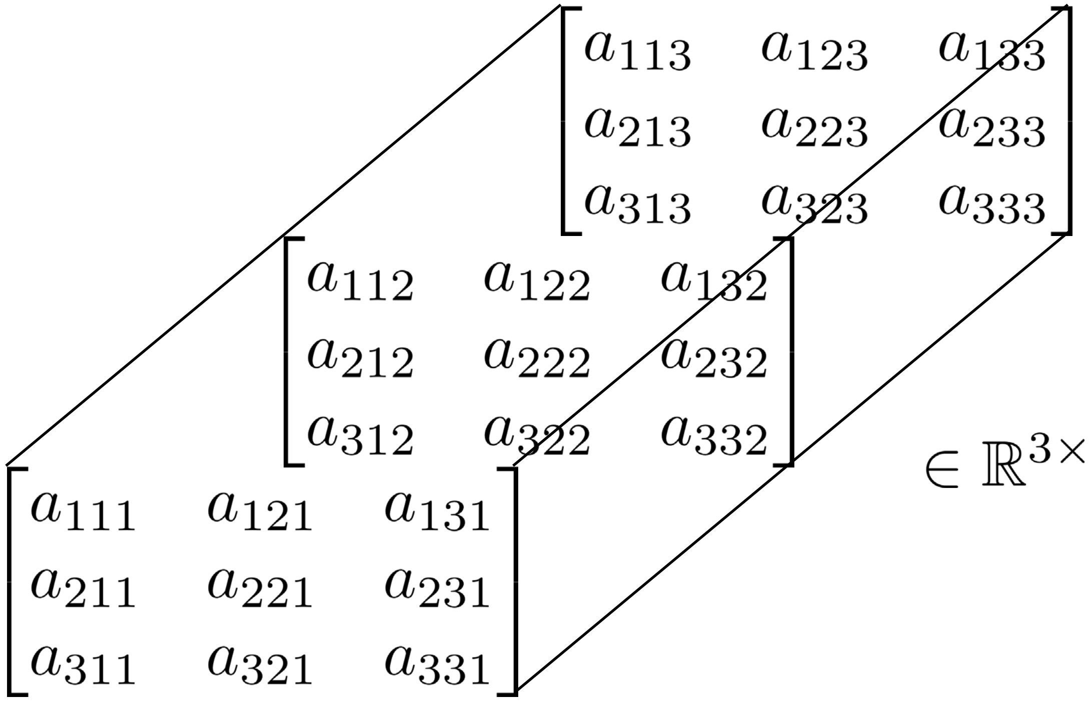
- Vectorized:

$$\text{vec}(\mathbf{A}) = \left[\mathbf{A}_{:,1}; \mathbf{A}_{:,2}; \ldots; \mathbf{A}_{:,n}\right] \in \mathbb{R}^{mn\times 1}$$

Tensors

- Are ***n*-D arrays** of numbers

- Can have $n = 0$ (scalar), $n = 1$ (vector), $n = 2$ (matrix), or $n > 2$



$\in \mathbb{R}^{3\times 3\times 3}$

*For tensors, **rank/order** refers to the number of dimensions (NOT the same meaning as for matrices)*

Matrix Transpose

- "Flipping" the rows and columns of a matrix (across main diagonal)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \in \mathbb{R}^{2\times 3} \qquad \mathbf{A}^\intercal = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

$$\mathbf{A}_{ij} = (\mathbf{A}^\intercal)_{ji}$$

- Some important properties:

$$(\mathbf{A}^\intercal)^\intercal = \mathbf{A} \qquad (\mathbf{AB})^\intercal = \mathbf{B}^\intercal\mathbf{A}^\intercal \qquad (\mathbf{A}+\mathbf{B})^\intercal = \mathbf{A}^\intercal + \mathbf{B}^\intercal$$

- **Symmetric** matrices:

$$\mathbf{A}^\intercal = \mathbf{A} \in \mathbb{R}^{n\times n}$$

Inner/Dot Product (1)

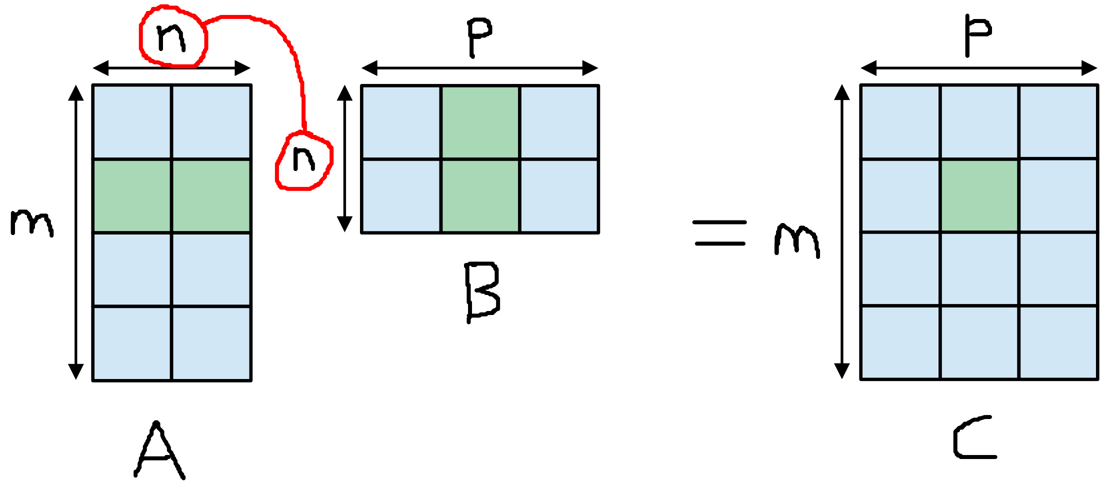
- Between vectors

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\mathsf{T}\mathbf{y} = \mathbf{y}^\mathsf{T}\mathbf{x}$$

- Between matrices $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$

$$\mathbf{C} = \mathbf{AB}$$

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



Inner/Dot Product (2)

- **Angle** between two vectors:

Outer Product

- Between two vectors: $\mathbf{x}\mathbf{y}^\intercal$
- Can have different dimensions $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$

Linear Transformations

- Linear transformation

$$f : \mathcal{X} \to \mathcal{Y} \text{ such that } f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) \text{ and } f(a\mathbf{x}) = af(\mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}$$

$$f(a_1 x_1 + \cdots + a_n x_n) = a_1 f(x_1) + \cdots a_n f(x_n)$$

Linear combo

Distributes

- Example: expectation operator $\mathbb{E}$

$$f_1, f_2 \Rightarrow f_1(f_2(x)) = f_2(f_1(x))$$

Linear Independence

Def. If no vector in a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ can be expressed as a linear combination of the others.

$$\sum_i^n a_i \mathbf{x}_i = 0 \implies a_1 = a_2 = \ldots = a_n = 0$$

$$\begin{bmatrix} | & | & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Conversely, $\mathbf{x}_n$ is *linearly dependent* on $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\}$ if there exist scalars $\{a_1, \ldots, a_{n-1}\}$ such that:

$$\mathbf{x}_n = \sum_i^{n-1} a_i \mathbf{x}_i$$

Linear Independence: Examples

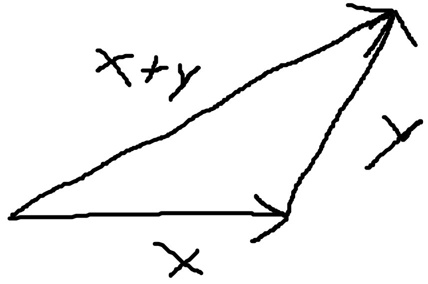
$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \qquad \begin{bmatrix} 1 & -3 \\ 1 & 2 \end{bmatrix}$$

Vector Norms (1)

- Interpretation of a vector's "length" $||\mathbf{x}||$ ← Linear ?

- Formally $||\mathbf{x}||$ is a function $\mathbb{R}^n \to \mathbb{R}$ that must satisfy 4 properties:

    1. $||\mathbf{x}|| \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ (*non-negativity*)
    2. $||\mathbf{x}|| = 0$ iff $\mathbf{x} = \mathbf{0}$ (*definiteness*)
    3. $||a\mathbf{x}|| = a||\mathbf{x}||, \forall a \in R$ (*absolute homogeneity*)
    4. $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (*triangle inequality*)



Vector Norms (2)

- General family of metrics for $\mathbf{x} \in \mathbb{R}^n$ known as L_p norm:

$$||\mathbf{x}||_p = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

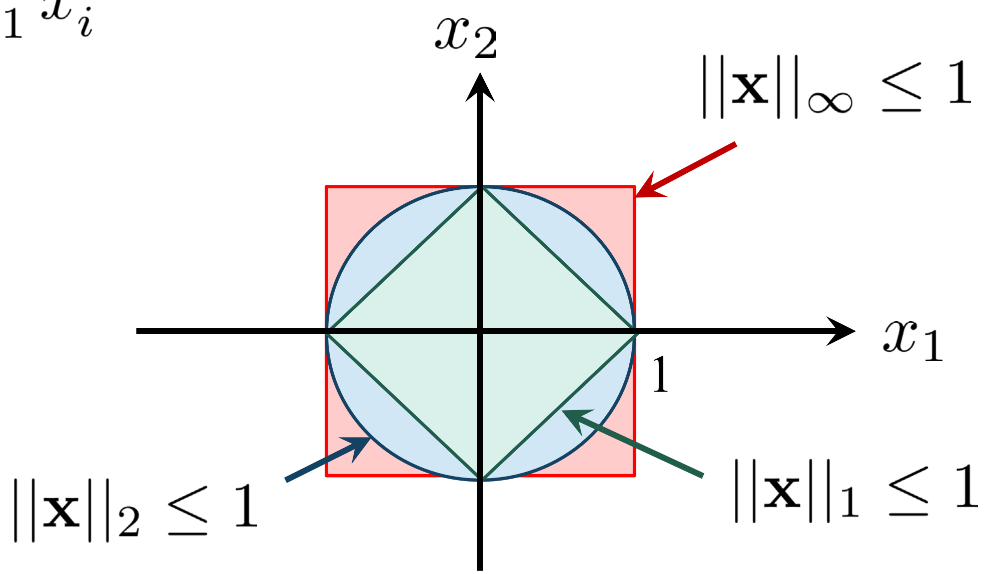
- Most popular (Euclidean): $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

- Others common in Machine Learning:

  Manhattan: $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$

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

  0-norm: $||\mathbf{x}||_0 = \text{count of non-zero } x_i$



Square Matrix Properties

- Trace: $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^\top)$$

$$\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$$

$$\mathrm{tr}(c\mathbf{A}) = c\,\mathrm{tr}(\mathbf{A})$$

? Linear op.

Cyclic permutation property: $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) = \mathrm{tr}(\mathbf{CAB})$

- Determinant: $\det(\mathbf{A})$ or $|\mathbf{A}|$

$$|\mathbf{A}| = |\mathbf{A}^\top| \qquad |c\mathbf{A}| = c^n|\mathbf{A}|$$

No

$$|\mathbf{A}| = 0 \text{ iff } \mathbf{A} \text{ singular, else } |\mathbf{A}^{-1}| = 1/|\mathbf{A}|$$

For 2×2 matrix: $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$

Trace Trick

- Trace trick to rewrite scalar dot product: $\mathbf{x}^\top \mathbf{A} \mathbf{x}$

Matrix Rank

- Dimension of the column space of $\mathbf{A} \implies$ no. of **linearly independent** cols.
- Column rank same as row rank: $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\intercal)$
- For $\mathbf{A} \in \mathbb{R}^{m \times n}$: $\text{rank}(\mathbf{A}) \leq \min(m, n)$
- **Full** $\text{rank}(\mathbf{A}) = \min(m, n)$ else rank **deficient**

$$\text{rank}\left(\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}\right) = 1 \qquad \text{rank}\left(\begin{bmatrix} 1 & -3 \\ 1 & 2 \end{bmatrix}\right) = 2$$

Special Matrices

- Diagonal (0 elsewhere)

$$\text{diag}(a_1, a_2 \ldots, a_n) = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

- Identity, $\mathbf{I}$, is a diagonal with 1s

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3 \times 3}$$

$$\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{I}\mathbf{x} = \mathbf{x} \qquad \forall \mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A}$$

- Orthogonality

Vectors

$$\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^\intercal \mathbf{y} = 0$$

Matrices

$$\mathbf{A}^\intercal \mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^\intercal$$

$$\mathbf{A}^\intercal = \mathbf{A}^{-1}$$

Set of orthogonal (column) vectors are linearly independent

Matrix Inversion

- Inverse of a square matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$, is $\mathbf{A}^{-1}$ such that: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$
- Exists iff $|\mathbf{A}| \neq 0$ (**full rank**), else referred to as a **singular** matrix
- Properties: $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$ $\quad (a\mathbf{A})^{-1} = \frac{1}{a}\mathbf{A}^{-1}$ $\quad (\mathbf{A}^{\intercal})^{-1} = (\mathbf{A}^{-1})^{\intercal}$
- For a 2×2 matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

System of Linear Equations

$$3x_1 + 8x_2 = 5$$
$$4x_1 + 11x_2 = 7$$

- Matrix representation

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{A} = \begin{bmatrix} 3 & 8 \\ 4 & 11 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

- If $\mathbf{A}^{-1}$ exists

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Eigenvalue Decomposition (EVD) – Intro

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, an *eigenvector* $\mathbf{u} \in \mathbb{R}^n$ is a non-zero vector with an associated number, called an *eigenvalue* $\lambda \in \mathbb{R}$, such that:

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

- $(\lambda, \mathbf{u})$ may be complex-valued, but **real**-valued for **symmetric** $\mathbf{A} = \mathbf{A}^\intercal$

- Properties: $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$ $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$

Eigenvalue Decomposition (EVD) – Matrix Form

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, there are $n \times (\lambda, \mathbf{u})$ pairs, which can be written as:

$$\mathbf{AU} = \mathbf{U\Lambda} \leftarrow \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

- If $\mathbf{U}^{-1}$ exists, then $\mathbf{A}$ is **"diagonalizable"** so the EVD is:

Spectral decomposition $\longrightarrow$ $\mathbf{A} = \mathbf{U\Lambda U}^{-1} = U\Lambda U^T$ if $A = A^T$

Singular Value Decomposition (SVD)

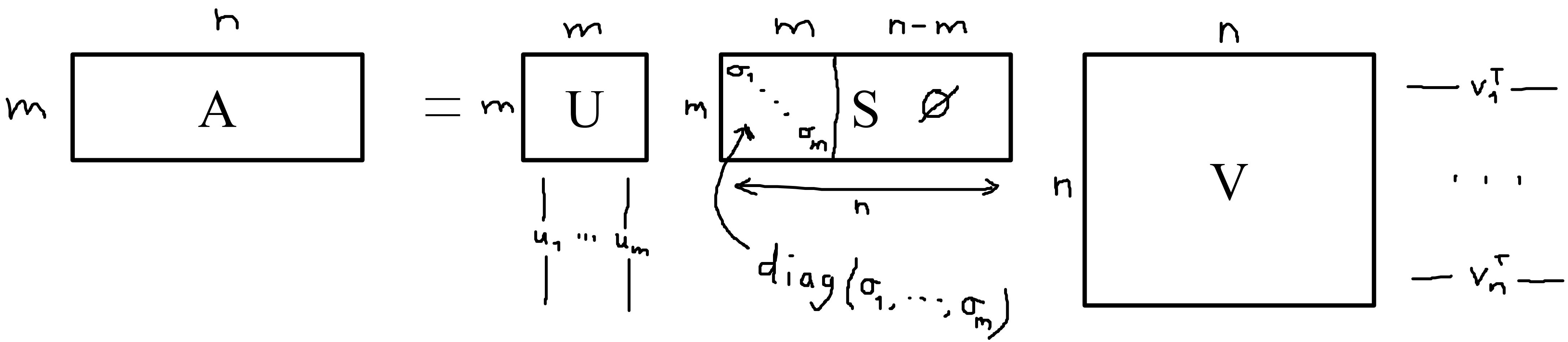
- More general case of rectangular matrices, $\mathbf{A} \in \mathbb{R}^{m \times n}$
- SVD of A: $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal$

$m \times m$ (U), $m \times n$ (S), $n \times n$ ($V^\intercal$)

Orthogonal U & V

$U^T U = I, \quad V^T V = I$

$m < n$



Coding Break

Quadratic Forms & Positive Definiteness

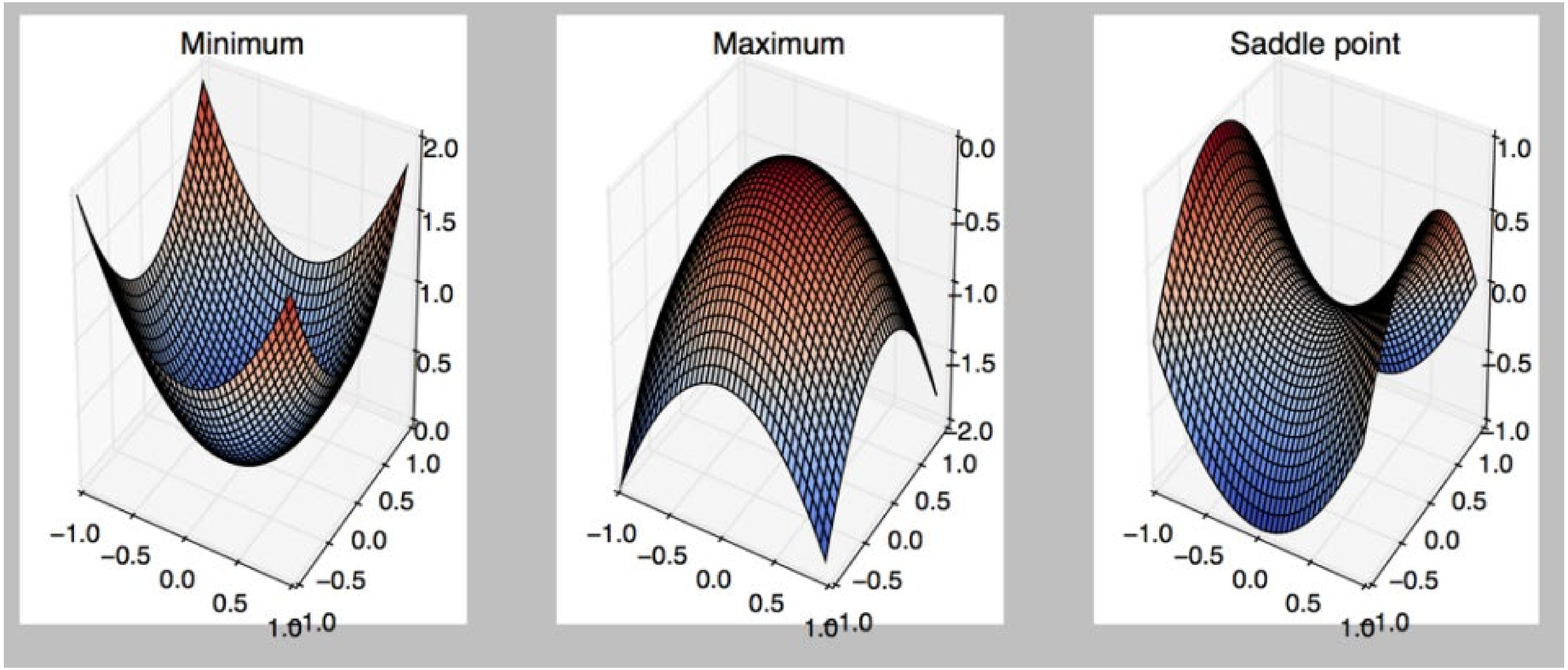
- Given $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, a **quadratic form** is the scalar:

$$\mathbf{x}^\intercal \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

- Can assume $\mathbf{A}$ is **symmetric** and:
  - *Positive definite (PD)* iff $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, denoted $\mathbf{A} > 0$
  - *Positive semidefinite (PSD)* iff $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} \geq 0, \forall \mathbf{x} \in \mathbb{R}^n$, denoted $\mathbf{A} \geq 0$
  - *Negative definite (ND)* iff $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} < 0, \forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, denoted $\mathbf{A} < 0$
  - *Negative semidefinite (NSD)* iff $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} \leq 0, \forall \mathbf{x} \in \mathbb{R}^n$, denoted $\mathbf{A} \leq 0$

Significance of Positive Definiteness

Zero gradient, and Hessian with...



All positive eigenvalues All negative eigenvalues Some positive and some negative

Sources: *Goodfellow et al, "Deep Learning", 2016*

Matrix Calculus

- **Gradient** of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$ is the vector of its partial derivatives:

$$\frac{\delta f}{\delta \mathbf{x}} = \nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \frac{\delta f}{\delta x_2} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix} \in \mathbb{R}^n$$

- Corresponding 2nd order derivatives are the symmetric **Hessian** matrix:

$$\frac{\delta^2 f}{\delta \mathbf{x}^2} = \nabla^2_{\mathbf{x}} f = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \cdots & \frac{\delta^2 f}{\delta x_1 \delta x_n} \\ \vdots & \ddots & \vdots \\ \frac{\delta^2 f}{\delta x_n \delta x_1} & \cdots & \frac{\delta^2 f}{\delta x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Summary

	Non-singular	Singular
$\mathbf{A}$ is	invertible	not invertible
Columns	independent	dependent
Rows	independent	dependent
$\det(\mathbf{A})$	$\neq 0$	$= 0$
$\mathbf{Ax} = \mathbf{0}$	one solution $\mathbf{x} = \mathbf{0}$	infinitely many solution
$\mathbf{Ax} = \mathbf{b}$	one solution	no solution or infinitely many
$\mathbf{A}$ has	n (nonzero) pivots	$r < n$ pivots
$\mathbf{A}$ has	full rank $r = n$	rank $r < n$
Column space	is all of $\mathbb{R}^n$	has dimension $r < n$
Row space	is all of $\mathbb{R}^n$	has dimension $r < n$
Eigenvalue	All eigenvalues are non-zero	Zero is an eigenvalue of $\mathbf{A}$
$\mathbf{A}^\mathsf{T}\mathbf{A}$	is symmetric positive definite	is only semidefinite
Singular value of $\mathbf{A}$	has n (positive) singular values	has $r < n$ singular values

Sources: Yuji Oyamada, https://www.slideshare.net/charmie11/comparison-singular-and-nonsingular

Resources

Linear Algebra

https://cs229.stanford.edu/section/cs229-linalg.pdf

https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Concluding Remarks

- Notation can appear scary but will get easier with practice

- Look at *"eig_svd.ipynb"* file for NumPy example of EVD & SVD:

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/foundations/eig_svd.ipynb

- Questions?