# Section 1: Probability, Statistics, & Linear Algebra review

## STATS 202: Data Mining and Analysis

Linh Tran
tranlm@stanford.edu

Department of Statistics
Stanford University

June 30, 2023

# Outline

- Linear algebra
  - Basic concepts
  - Matrix multiplication
  - Operations and Properties
  - Matrix Calculus
- Probability
  - Sample space
  - Probability function
  - Probability space
  - Random variables
- Statistics
  - Expected value
  - Moments & Moment generating functions
  - Distributions

# Linear algebra

# Basic concepts

Consider the following equations:

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \quad (1) \\ -2x_1 + 3x_2 &= 9 \quad (2) \end{aligned}$$

Let's solve for $x_1$ and $x_2$.

Consider the following equations:

$$4x_1 - 5x_2 = -13 \quad (1)$$

$$-2x_1 + 3x_2 = 9 \quad (2)$$

Let's solve for $x_1$ and $x_2$.

We can write this system of equations more compactly in matrix notation, e.g.

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (3)$$

where $\mathbf{A} = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$

# Basic concepts

Some basic notation:

- We denote a matrix with $m$ rows and $n$ columns as $\mathbf{A} \in \mathbb{R}^{m \times n}$, where each entry in the matrix is a real number.
- We denote a vector with $n$ entries as $\mathbf{x} \in \mathbb{R}^n$.
  - By convention, we typically think of a vector as a 1 column matrix.
- We denote the $i^{th}$ element of a vector $\mathbf{x}$ as $x_i$, e.g.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{4}$$

# Basic concepts

Some basic notation:

- We denote each entry in a matrix $\mathbf{A}$ by $a_{ij}$, corresponding to the $i^{th}$ row and $j^{th}$ column, e.g.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{5}$$

- We denote the *transpose* of a matrix as $\mathbf{A}^\top$, e.g.

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \tag{6}$$

# Basic concepts

Some basic notation:

- We denote the $j^{th}$ column of $\mathbf{A}$ by $\mathbf{a}_j$ or $\mathbf{A}_{\cdot j}$, e.g.

$$\mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \tag{7}$$

- We denote the $i^{th}$ row of $\mathbf{A}$ by $\mathbf{a}_i^\top$ or $\mathbf{A}_{i\cdot}$.

$$\mathbf{A} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \tag{8}$$

n.b. This isn't universal, though should be clear from its presentation and use.

# Matrix multiplication

Given two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, we can multiply them by

$$\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p} : \mathbf{C}_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik} \mathbf{B}_{kj} \tag{9}$$

n.b. The dimensions have to be compatible for matrix multiplication to be valid (e.g. the number of columns in $\mathbf{A}$ must be equal to the number of rows in $\mathbf{B}$).

# Matrix multiplication

Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the quantity $\mathbf{x}^\top \mathbf{y} \in \mathbb{R}$ (aka *dot product* or *inner product*) is a scalar given by

$$\mathbf{x}^\top \mathbf{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i \tag{10}$$

Note: For vectors, we always have that $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$. This is not generally true for matrices.

# Matrix multiplication

Given $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$, the quantity $\mathbf{x}^\top \mathbf{y} \in \mathbb{R}^{m \times n}$ (aka *outer product*) is a matrix given by

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix} \quad (11)$$

# Matrix multiplication

**Example:** Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix such that all columns are equal to some vector $\mathbf{x} \in \mathbb{R}^m$. Using outer products, we can represent $\mathbf{A}$ compactly as

$$\mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{x} & \mathbf{x} & \cdots & \mathbf{x} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} \tag{12}$$

$$= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \tag{13}$$

$$= \mathbf{x}\mathbf{1}^\top \tag{14}$$

# Matrix-vector products

Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^n$, their product is a vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$.

# Matrix-vector products

Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^n$, their product is a vector $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$.

There are two ways of interpreting this:

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix} \tag{15}$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \tag{16}$$

$$= \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \cdots + \mathbf{a}_n x_n \tag{17}$$

# Matrix-vector products

**Example:**

Define $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} -3 \\ -2 \\ -1 \end{bmatrix}$.

Calculate $\mathbf{y} = \mathbf{A}\mathbf{x}$.

# Matrix-matrix products

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

# Matrix-matrix products

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

Similar to before, we can think of this in two ways:

**Interpretation # 1**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & & | \end{bmatrix} \tag{18}$$

$$= \begin{bmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots \mathbf{a}_1^\top \mathbf{b}_p \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots \mathbf{a}_2^\top \mathbf{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_m^\top \mathbf{b}_1 & \mathbf{a}_m^\top \mathbf{b}_2 & \cdots \mathbf{a}_m^\top \mathbf{b}_p \end{bmatrix} \tag{19}$$

# Matrix-matrix products

**Interpretation # 2**

$$\mathbf{C} = \mathbf{AB} = \mathbf{A} \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_p \\ | & | & & | \end{bmatrix} \quad (20)$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{Ab}_1 & \mathbf{Ab}_2 & \cdots & \mathbf{Ab}_p \\ | & | & & | \end{bmatrix} \quad (21)$$

$$= \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \mathbf{B} = \begin{bmatrix} — & \mathbf{a}_1^\top \mathbf{B} & — \\ — & \mathbf{a}_2^\top \mathbf{B} & — \\ & \vdots & \\ — & \mathbf{a}_m^\top \mathbf{B} & — \end{bmatrix} \quad (22)$$

# Matrix multiplication properties

- Associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- Not commutative: $\mathbf{AB} \neq \mathbf{BA}$

# Matrix multiplication properties

Demonstrating *associativity*:

We just need to show that $((\mathbf{AB})\mathbf{C})_{ij} = (\mathbf{A}(\mathbf{BC}))_{ij}$:

$$
\begin{aligned}
((\mathbf{AB})\mathbf{C})_{ij} &= \sum_{k=1}^{p} (\mathbf{AB})_{ik}\mathbf{C}_{kj} = \sum_{k=1}^{p}\left(\sum_{l=1}^{n}\mathbf{A}_{il}\mathbf{B}_{lk}\right)\mathbf{C}_{kj} && (23)\\
&= \sum_{k=1}^{p}\left(\sum_{l=1}^{n}\mathbf{A}_{il}\mathbf{B}_{lk}\mathbf{C}_{kj}\right) = \sum_{l=1}^{n}\left(\sum_{k=1}^{p}\mathbf{A}_{il}\mathbf{B}_{lk}\mathbf{C}_{kj}\right) && (24)\\
&= \sum_{l=1}^{n}\mathbf{A}_{il}\left(\sum_{k=1}^{p}\mathbf{B}_{lk}\mathbf{C}_{kj}\right) = \sum_{l=1}^{n}\mathbf{A}_{il}(\mathbf{BC})_{lj} && (25)\\
&= (\mathbf{A}(\mathbf{BC}))_{ij} && (26)
\end{aligned}
$$

**The identity matrix**:

The *identity matrix*, denoted $\mathbf{I} \in \mathbb{R}^{n \times n}$ is a square matrix with 1's in the diagonal and 0's everywhere else, i.e.

$$\mathbf{I}_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{27}$$

**The identity matrix**:

The *identity matrix*, denoted $\mathbf{I} \in \mathbb{R}^{n \times n}$ is a square matrix with 1's in the diagonal and 0's everywhere else, i.e.

$$\mathbf{I}_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{27}$$

It has the property

$$\mathbf{AI} = \mathbf{A} = \mathbf{IA} \; \forall \mathbf{A} \in \mathbb{R}^{m \times n} \tag{28}$$

n.b. The dimensionality of $\mathbf{I}$ is typically inferred (e.g. $n \times n$ vs $m \times m$)

**The diagonal matrix**: The *diagonal matrix*, denoted $\mathbf{D} = diag(d_1, d_2, \ldots, d_n)$ is a matrix where all non-diagonal elements are 0, i.e.

$$\mathbf{D}_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases} \tag{29}$$

Clearly, $\mathbf{I} = diag(1, 1, ..., 1)$.

# The transpose

The *transpose* of a matrix results from "*flipping*" the rows and columns, i.e.

$$(\mathbf{A}^\top)_{ij} = \mathbf{A}_{ji} \tag{30}$$

Consequently, for $\mathbf{A} \in \mathbb{R}^{m \times n}$ we have that $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$.

Some properties:

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top \mathbf{A}^\top$
- $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

# Symmetry

A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^\top$.

It is *anti-symmetric* if $\mathbf{A} = -\mathbf{A}^\top$.

# Symmetry

A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^\top$.

It is *anti-symmetric* if $\mathbf{A} = -\mathbf{A}^\top$.

It is easy to show that $\mathbf{A} + \mathbf{A}^\top$ is symmetric and $\mathbf{A} - \mathbf{A}^\top$ is anti-symmetric. Consequently, we have that

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top) \tag{31}$$

# Symmetry

A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^\top$.

It is *anti-symmetric* if $\mathbf{A} = -\mathbf{A}^\top$.

It is easy to show that $\mathbf{A} + \mathbf{A}^\top$ is symmetric and $\mathbf{A} - \mathbf{A}^\top$ is anti-symmetric. Consequently, we have that

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top) \tag{31}$$

Symmetric matrices tend to be denoted as $\mathbf{A} \in \mathbb{S}^n$.

# Trace

The *trace* of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $tr(\mathbf{A})$ or $tr\mathbf{A}$ is the sum of the diagonal elements, i.e.

$$tr\mathbf{A} = \sum_{i=1}^{n} \mathbf{A}_{ii} \tag{32}$$

The trace has the following properties:

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $tr\mathbf{A} = tr\mathbf{A}^\top$
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $tr(\mathbf{A} + \mathbf{B}) = tr\mathbf{A} + tr\mathbf{B}$
- For $\mathbf{A} \in \mathbb{R}^{n \times n}, c \in \mathbb{R}$, $tr(c\mathbf{A}) = c\, tr\mathbf{A}$
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n} \ni \mathbf{AB} \in \mathbb{R}^{n \times n}$, $tr\mathbf{AB} = tr\mathbf{BA}$
- For $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times n} \ni \mathbf{ABC} \in \mathbb{R}^{n \times n}$, $tr\mathbf{ABC} = tr\mathbf{BCA} = tr\mathbf{CAB}$, and so on for more matrices

# Trace

**Example:** Proving that $tr\mathbf{AB} = tr\mathbf{BA}$

$$
\begin{aligned}
tr\mathbf{AB} &= \sum_{i=1}^{m} (\mathbf{AB})_{ii} = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \mathbf{A}_{ij}\mathbf{B}_{ji} \right) && (33) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{A}_{ij}\mathbf{B}_{ji} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{B}_{ji}\mathbf{A}_{ij} && (34) \\
&= \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \mathbf{B}_{ji}\mathbf{A}_{ij} \right) = \sum_{j=1}^{n} (\mathbf{BA})_{jj} && (35) \\
&= tr\mathbf{BA} && (36)
\end{aligned}
$$

# Norms

A *norm* of a vector $\mathbf{x}$, denoted $||\mathbf{x}||$ is a measure of the "*length*" of the vector. For example, the $\ell_2$-norm (aka Euclidean norm) is

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{37}$$

n.b. $||\mathbf{x}||_2^2 = \mathbf{x}^\top \mathbf{x}$, i.e. the squared norm of a vector is the dot product with itself.

# Norms

A *norm* of a vector $\mathbf{x}$, denoted $||\mathbf{x}||$ is a measure of the "*length*" of the vector. For example, the $\ell_2$-norm (aka Euclidean norm) is

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (37)$$

n.b. $||\mathbf{x}||_2^2 = \mathbf{x}^\top \mathbf{x}$, i.e. the squared norm of a vector is the dot product with itself.

**Other norms:**

- $\ell_1$-norm, i.e. $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$.
- $\ell_\infty$-norm, i.e. $||\mathbf{x}||_\infty = \max_i |x_i|$.
- $\ell_p$-norm, i.e. $||\mathbf{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$.

# Norms

Formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying four properties:

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

# Norms

Formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying four properties:

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

Norms can also be defined for matrices, e.g. The Frobenius norm,

$$||\mathbf{A}||^F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{A}_{ij}^2} = \sqrt{tr(\mathbf{A}^\top \mathbf{A})} \tag{38}$$

# Linear independence

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^m$ is *(linearly) dependent* if one of the vectors $\mathbf{x}_i$ can be represented as a linear combination of the remaining vectors, i.e.

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i \tag{39}$$

for some scalar values $\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \in \mathbb{R}$

# Linear independence

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^m$ is *(linearly) dependent* if one of the vectors $\mathbf{x}_i$ can be represented as a linear combination of the remaining vectors, i.e.

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{x}_i \tag{39}$$

for some scalar values $\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \in \mathbb{R}$

**Example:** Let

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} \tag{40}$$

Is $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ linearly independent?

# Rank

The *column rank* of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the largest subset of columns of $\mathbf{A}$ that are linearly independent.

- The column rank is always $\leq n$.

The *row rank* of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the largest subset of rows of $\mathbf{A}$ that are linearly independent.

- The row rank is always $\leq m$.

# Rank

The *column rank* of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the largest subset of columns of $\mathbf{A}$ that are linearly independent.

- The column rank is always $\leq n$.

The *row rank* of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the largest subset of rows of $\mathbf{A}$ that are linearly independent.

- The row rank is always $\leq m$.

n.b. Column rank is always equal to row rank. Thus, we refer to both as the *rank* of the matrix.

- For $\mathbf{A} \in \mathbb{R}^{m \times n}$, if $rank(\mathbf{A}) = \min(m, n)$, then $\mathbf{A}$ is said to be of *full rank*.
- For $\mathbf{A} \in \mathbb{R}^{m \times n}$, $rank(\mathbf{A}) = rank(\mathbf{A}^\top$.
- For $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$, $rank(\mathbf{AB}) \leq \min(rank(\mathbf{A}), rank(\mathbf{B}))$.
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, $rank(\mathbf{A} + \mathbf{B}) \leq rank(\mathbf{A}) + rank(\mathbf{B})$

# Matrix inverse

The *inverse* of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is denoted $\mathbf{A}^{-1}$, and is unique such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \tag{41}$$

# Matrix inverse

The *inverse* of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is denoted $\mathbf{A}^{-1}$, and is unique such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \tag{41}$$

n.b. Not all matrices have inverses (e.g. $m \times n$ matrices).

**Def:**
A is *invertible* or *non-singular* if $\mathbf{A}^{-1}$ exists.
Otherwise, it is *non-invertible* or *singular*.

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
3. $(\mathbf{A}^{-1})^{\top} = (\mathbf{A}^{\top})^{-1}$
   - This matrix is sometimes denoted $\mathbf{A}^{-\top}$

# Orthogonal Matrices

**Def:**

- A vector $\mathbf{x} \in \mathbb{R}^n$ is *normalized* if $||\mathbf{x}||_2 = 1$
- Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are *orthogonal* if $\mathbf{x}^\top \mathbf{y} = 0$
- A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is *orthogonal* or *orthonormal* if all its columns are:
  1. Orthogonal to each other
  2. Normalized

We therfore have that

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top \tag{42}$$

# Orthogonal Matrices

**Def:**

- A vector $\mathbf{x} \in \mathbb{R}^n$ is *normalized* if $||\mathbf{x}||_2 = 1$
- Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are *orthogonal* if $\mathbf{x}^\top \mathbf{y} = 0$
- A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is *orthogonal* or *orthonormal* if all its columns are:
    1. Orthogonal to each other
    2. Normalized

We therfore have that

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top \tag{42}$$

Another nice property:

$$||\mathbf{U}\mathbf{x}||_2 = ||\mathbf{x}||_2 \ \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{U} \in \mathbb{R}^{n \times n} \text{ orthogonal} \tag{43}$$

# Range

**Def:**
The *span* of a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is

$$\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \alpha_i \in \mathbb{R} \right\} \tag{44}$$

# Range

**Def:**
The *span* of a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is

$$\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \alpha_i \in \mathbb{R} \right\} \tag{44}$$

n.b. If $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is linearly independent, then $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = \mathbb{R}^n$.

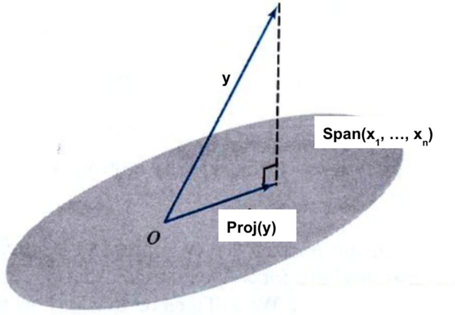
**Example:**

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{45}$$

# Projection

**Def:**
The *projection* of a vector $\mathbf{y} \in \mathbb{R}^m$ onto $\text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = \mathbb{R}^n$ is

$$\text{Proj}(\mathbf{y}; \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}) = \underset{\mathbf{v} \in \text{span}(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\})}{\arg\min} ||\mathbf{y} - \mathbf{v}||_2 \quad (46)$$



**Def:**

The *range* of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(\mathbf{A})$ is the span of the columns of $\mathbf{A}$, i.e.

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^m : \mathbf{v} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\} \tag{47}$$

Assuming that $\mathbf{A}$ is full rank and $n < m$, the projection of $\mathbf{y} \in \mathbb{R}^m$ onto $\mathcal{R}(\mathbf{A})$ is

$$\begin{aligned} \text{Proj}(\mathbf{y}; \mathbf{A}) &= \arg\min_{\mathbf{v} \in \mathcal{R}(\mathbf{A})} ||\mathbf{v} - \mathbf{y}||_2 & (48) \\ &= \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} & (49) \end{aligned}$$

# Nullspace

**Def:**
The *nullspace* of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(\mathbf{A})$ is the set of all vectors that equal 0 when multiplied by $\mathbf{A}$, i.e.

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = 0\} \tag{50}$$

Some properties:

- $\{w : w = u + v, u \in \mathcal{R}(\mathbf{A}^\top), v \in \mathcal{R}(\mathbf{A})\} = \mathbb{R}^n$
- $\mathcal{R}(\mathbf{A}^\top) \bigcap \mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$

This is referred to as *orthogonal complements*, denoted as $\mathcal{R}(\mathbf{A}^\top) = \mathcal{N}(\mathbf{A})^\perp$

**Def:**
The *determinant* of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $|\mathbf{A}|$ or det $\mathbf{A}$ is a function det: $\mathbb{R}^{n \times n} \rightarrow \mathbb{R}$.

Let $\mathbf{A}_{\setminus i, \setminus j} \in \mathbb{R}^{(n-1) \times (n-1)}$ be the matrix that results from deleting the $i^{th}$ row and $j^{th}$ column. The general (recursive) formula for the determinant is

$$\begin{aligned} |\mathbf{A}| &= \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{A}_{\setminus i, \setminus j}| \quad (\forall j \in 1, ..., n) \\ &= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{A}_{\setminus i, \setminus j}| \quad (\forall i \in 1, ..., n) \end{aligned} \tag{51}$$

Given a matrix

$$\mathbf{A} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_n^\top & — \end{bmatrix} \tag{52}$$

and a set $\mathbf{S} \subset \mathbb{R}^n$,

$$\mathbf{S} = \{\mathbf{v} \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i \mathbf{a}_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, ..., n\} \tag{53}$$

$|\mathbf{A}|$ is the volume of $\mathbf{S}$.

# Determinant

**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{54}$$

**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{54}$$

The matrix rows are:

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{55}$$

And $|\mathbf{A}| = -7$

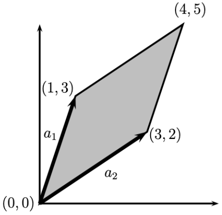
**Example:**

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \tag{54}$$

The matrix rows are:

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{55}$$

And $|\mathbf{A}| = -7$



# Determinant

Properties of determinants:

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $|\mathbf{A}| = |\mathbf{A}^\top|$
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $|\mathbf{A}| = 0$ iff $\mathbf{A}$ is singular (i.e. non-invertible).
- For $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{A}$ non-singular, $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$

# Quadratic form

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

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i (\mathbf{A}\mathbf{x})_i = \sum_{i=1}^{n} x_i \left( \sum_{j=1}^{n} \mathbf{A}_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{A}_{ij} x_i x_j \quad (56)$$

# Quadratic form

Some properties involving quadratic form:

- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is *positive definite* if for a non-zero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$
- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is *positive semi-definite* if for a non-zero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$
- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is *negative definite* if for a non-zero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A} \mathbf{x} < 0$
- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is *negative semi-definite* if for a non-zero $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top \mathbf{A} \mathbf{x} \leq 0$
- A symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ is *indefinite* if it is neither positive nor negative semidefinite

n.b. Positive definite and negative definite matrices always have full rank.

# Eigenvalues & eigenvectors

Given $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\lambda \in \mathbb{C}$ is an *eigenvalue* of $\mathbf{A}$ with corresponding *eigenvector* $\mathbf{x} \in \mathbb{C}^n$ if

$$\mathbf{Ax} = \lambda \mathbf{x} : \mathbf{x} \neq 0 \tag{57}$$

n.b. The eigenvector is (usually) normalized to have length 1

# Eigenvalues & eigenvectors

Given $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\lambda \in \mathbb{C}$ is an *eigenvalue* of $\mathbf{A}$ with corresponding *eigenvector* $\mathbf{x} \in \mathbb{C}^n$ if

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} : \mathbf{x} \neq 0 \quad (57)$$

n.b. The eigenvector is (usually) normalized to have length 1

We can write all of the eigenvector equations simultaneously as

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \quad (58)$$

where

$$\mathbf{X} \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{bmatrix}, \quad \mathbf{\Lambda} = diag(\lambda_1, ..., \lambda_n) \quad (59)$$

This implies $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$

# Eigenvalues & eigenvectors

**Some properties:**

- $tr\mathbf{A} = \sum_{i=1}^{n} \lambda_i$
- $|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$
- The rank of **A** is equal to the number of non-zero eigenvalues of **A**.
- If **A** is non-singular, then $1/\lambda_i$ is an eigenvalue of $\mathbf{A}^{-1}$ with correspondng eigenvector $\mathbf{x}_i$, i.e. $\mathbf{A}^{-1}\mathbf{x}_i = (1/\lambda_i)\mathbf{x}_i$
- The eigenvalues of a diagonal matrix $D = diag(d_1, ..., d_n)$ are just its diagonal entries $d_1, ..., d_n$

# Eigenvalues & eigenvectors

**Example**: For $\mathbf{A} \in \mathbb{S}^n$ with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$,

$$\max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{A} \mathbf{x} \text{ subject to } ||\mathbf{x}||_2^2 = 1 \tag{60}$$

is solved with $\mathbf{x}_1$ corresponding to $\lambda_1$. Similarly, it is solved with $\mathbf{x}_n$ corresponding to $\lambda_n$.

# Eigenvalues & eigenvectors

**Example:**

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ Find the eigenvalues & eigenvectors.

# Eigenvalues & eigenvectors

**Example:**

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ Find the eigenvalues & eigenvectors.

We want

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = 0 \tag{61}$$

**Example:**

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ Find the eigenvalues & eigenvectors.

We want

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = 0 \tag{61}$$

We want $det(\mathbf{A} - \lambda \mathbb{I}) = 0$.

$$\begin{aligned} det(\mathbf{A} - \lambda \mathbb{I}) &= (1-\lambda)^2 - 2^2 = \lambda^2 - 2\lambda - 3 && (62) \\ &= (\lambda - 3)(\lambda + 1) && (63) \end{aligned}$$

$\therefore \ \lambda = 3, -1.$

# Eigenvalues & eigenvectors

Finding the eigenvectors: calculating the null spaces of $(\mathbf{A} - \lambda\mathbf{I})$

$$\mathcal{N}(\mathbf{A} - 3\mathbf{I}) = \mathcal{N}\left(\begin{bmatrix}-2 & 2\\ 2 & -2\end{bmatrix}\right) = \begin{bmatrix}1\\1\end{bmatrix} \tag{64}$$

$$\mathcal{N}(\mathbf{A} + \mathbf{I}) = \mathcal{N}\left(\begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}\right) = \begin{bmatrix}1\\-1\end{bmatrix} \tag{65}$$

Finding the eigenvectors: calculating the null spaces of $(\mathbf{A} - \lambda \mathbf{I})$

$$\mathcal{N}(\mathbf{A} - 3\mathbf{I}) = \mathcal{N}\left(\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{64}$$

$$\mathcal{N}(\mathbf{A} + \mathbf{I}) = \mathcal{N}\left(\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{65}$$

Thus:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} \tag{66}$$

# Singular Value Decomposition

SVD is a way of decomposing matrices.

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r$, $\exists$
$\mathbf{\Sigma} \in \mathbb{R}^{m \times n}, \mathbf{U} \in \mathbb{R}^{m \times m}, \mathbf{V} \in \mathbb{R}^{n \times m} \ni$

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{67}$$

Notes:

- $\mathbf{\Sigma}$ is a diagonal matrix with entries $\sigma_1, ..., \sigma_r > 0$ known as *singular values*.
- $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices.
- Common uses:
  - Least squares models
  - Range, rank, null space
  - Moore-Penrose inverse

**Some intuition:**

$\mathbf{A} \in \mathbb{R}^{m \times n}$ can be thought of as a linear transformation, such that for $\mathbf{x} \in \mathbb{R}^n$,

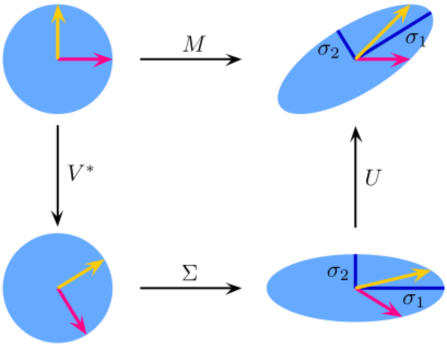
$$f(\mathbf{x}) = \mathbf{A}\mathbf{x} \tag{68}$$

**Some intuition:**

$\mathbf{A} \in \mathbb{R}^{m \times n}$ can be thought of as a linear transformation, such that for $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) = \mathbf{Ax} \tag{68}$$

SVD can be thought of as breaking this into individual steps:



Given $f : \mathbb{R}^{m \times n} \to \mathbb{R}$, the *gradient* of $f$ wrt $\mathbf{A} \in \mathbb{R}^{m \times n}$ is

$$\nabla_{\mathbf{A}} f(\mathbf{A}) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{11}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{12}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{1n}} \\ \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{21}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{22}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{m1}} & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{m2}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{mn}} \end{bmatrix} \tag{69}$$

Some properties

- $\nabla_{\mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}} g(\mathbf{x})$
- For $c \in \mathbb{R}$, $\nabla_{\mathbf{x}}(c\, f(\mathbf{x})) = c \nabla_{\mathbf{x}}(f(\mathbf{x}))$

# The Hessian

Given $f : \mathbb{R}^n \to \mathbb{R}$, the *Hessian* of $f$ wrt $\mathbf{x} \in \mathbb{R}^n$ is

$$\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \tag{70}$$

n.b. The Hessian is always symmetric, since $\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}$

# Least squares

Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m \ni b \notin \mathcal{R}(A)$, we want to find $\mathbf{x} \in \mathbb{R}^n$ as close as possible to $\mathbf{b}$ (via the Euclidean norm),

$$
\begin{aligned}
||\mathbf{Ax} - \mathbf{b}||_2^2 &= (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \qquad (71) \\
&= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \qquad (72)
\end{aligned}
$$

# Least squares

Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m \ni b \notin \mathcal{R}(A)$, we want to find $\mathbf{x} \in \mathbb{R}^n$ as close as possible to $\mathbf{b}$ (via the Euclidean norm),

$$||\mathbf{Ax} - \mathbf{b}||_2^2 = (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \tag{71}$$

$$= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \tag{72}$$

Taking the gradient wrt $\mathbf{x}$, we have

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b}) = \nabla_{\mathbf{x}}\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - \nabla_{\mathbf{x}} 2\mathbf{b}^\top \mathbf{Ax} + \nabla_{\mathbf{x}}\mathbf{b}^\top \mathbf{b} \tag{73}$$

$$= \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{A}^\top \mathbf{b} \tag{74}$$

# Least squares

Given $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m \ni b \notin \mathcal{R}(A)$, we want to find $\mathbf{x} \in \mathbb{R}^n$ as close as possible to $\mathbf{b}$ (via the Euclidean norm),

$$||\mathbf{Ax} - \mathbf{b}||_2^2 = (\mathbf{Ax} - \mathbf{b})^\top (\mathbf{Ax} - \mathbf{b}) \quad (71)$$
$$= \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b} \quad (72)$$

Taking the gradient wrt $\mathbf{x}$, we have

$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b}) = \nabla_{\mathbf{x}}\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - \nabla_{\mathbf{x}} 2\mathbf{b}^\top \mathbf{Ax} + \nabla_{\mathbf{x}}\mathbf{b}^\top \mathbf{b} \quad (73)$$
$$= \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{A}^\top \mathbf{b} \quad (74)$$

Setting this expression equal to zero and solving for $\mathbf{x}$ gives the normal equations,

$$\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} \quad (75)$$

# References

Some textbooks on linear algebra:

- *Linear Algebra (Jim Hefferon)*
- *Introduction to Applied Linear Algebra (Boyd & Vandenberghe)*
- *Linear Algebra (Cherney, Denton et al.)*
- *Linear Algebra (Hoffman & Kunze)*
- *Fundamentals of Linear Algebra (Carrell)*
- *Linear Algebra (S. Friedberg A. Insel L. Spence)*

# Probability

# Sample space

The set of all possible values is called the *sample space* $S$.

- It's the space where realizations can be produced.

# Sample space

The set of all possible values is called the *sample space* $S$.

- It's the space where realizations can be produced.

**Example**: Tossing a coin

$$S = \{Heads, Tails\} \tag{76}$$

# Sample space

The set of all possible values is called the *sample space* $S$.

- It's the space where realizations can be produced.

**Example**: Tossing a coin

$$S = \{Heads, Tails\} \tag{76}$$

More notation:

- $\emptyset$ is the *empty set*. Can be denoted as $\emptyset = \{\}$.
- $\cup_{i=1}^{\infty} B_i$ is the union of sets $B_i$. Formally,
  - $\cup_{i=1}^{\infty} B_i = \{s \in S : s \in B_i \forall\, i\}$
- $B \subseteq S$ means $B$ is a *subset* of the sample space.
- *Heads*, without curly braces, is an *element* of set $B$.
- $B^C = S \setminus B$ is the complement of set $B$

# Probability function

A *probability function* is a function $P : \mathcal{B} \to [0, 1]$, where

- $P(S) = 1$
- $P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i)$ when $B_1, B_2, \ldots$ are disjoint

# Probability function

A *probability function* is a function $P : \mathcal{B} \to [0, 1]$, where

- $P(S) = 1$
- $P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i)$ when $B_1, B_2, \ldots$ are disjoint

n.b. We can define the domain $\mathcal{B}$ many ways, e.g. $\mathcal{B} = 2^S$

# Probability function

A *probability function* is a function $P : \mathcal{B} \to [0, 1]$, where

- $P(S) = 1$
- $P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i)$ when $B_1, B_2, \ldots$ are disjoint

n.b. We can define the domain $\mathcal{B}$ many ways, e.g. $\mathcal{B} = 2^S$

**Example:** For flipping a coin, we have

$$\mathcal{B} = 2^S = \{\emptyset, \{Heads\}, \{Tails\}, \{Heads, Tails\}\} \tag{77}$$

This implies that

$$P(B) = \begin{cases} 1 & B = \{Heads, Tails\} \\ \frac{1}{2} & B = \{Heads\} \\ \frac{1}{2} & B = \{Tails\} \\ 0 & B = \emptyset \end{cases} \tag{78}$$

n.b. The power set is a *'set of sets'*

**Problem:** Power sets don't work well for $\mathbb{R}$.

# Probability function domains

**Problem:** Power sets don't work well for $\mathbb{R}$.
**Solution:** Define the domain using $\sigma-$algebra:

- $\emptyset \in \mathcal{B}$
- $B \in \mathcal{B} \Rightarrow B^C \in \mathcal{B}$
- $B_1, B_2, \ldots \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} B_i \in \mathcal{B}$

**Problem:** Power sets don't work well for $\mathbb{R}$.
**Solution:** Define the domain using $\sigma-$algebra:

- $\emptyset \in \mathcal{B}$
- $B \in \mathcal{B} \Rightarrow B^C \in \mathcal{B}$
- $B_1, B_2, \ldots \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} B_i \in \mathcal{B}$

**Example:**

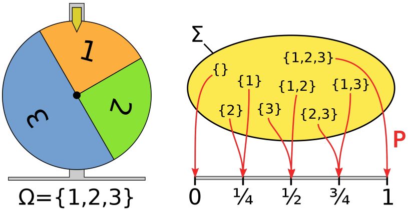
- The *discrete* $\sigma$-algebra:
  $\mathcal{B} = 2^S = \{\emptyset, \{Heads\}, \{Tails\}, \{Heads, Tails\}\}$
- The *trivial* $\sigma$-algebra: $\mathcal{B} = \emptyset \cup S = \{\emptyset, \{Heads, Tails\}\}$

n.b. For uncountable sets, we use the *Borel* $\sigma$-algebra.

**Def:**
A *probability space* is a triple $(S, \mathcal{B}, P)$.

- $S$ is the set of possible singleton events
- $\mathcal{B}$ is the set of questions to ask $P$
- $P$ maps sets into probabilities

n.b. They represent the ingredients needed to talk about probabilities



# Probability functions

Some properties of $P(\cdot)$

- $P(B) = 1 - P(B^C)$
- $P(\emptyset) = 0$, since $P(\emptyset) = 1 - P(S)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, implying that
  - $P(A \cup B) \leq P(A) + P(B)$
  - $P(A \cap B) \geq P(A) + P(B) - 1$

# Conditional probability

For events $A$ and $B$ where $P(B) > 0$, the *conditional probability* of $A$ given $B$ (denoted $P(A|B)$) is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{79}$$

**Example:** In an agricultural region with 1000 farms, we want to know if the farm has vineyards or cork trees.

| | | **Cork Trees** | |
|---|---|---|---|
| | | Yes | No |
| **Vineyard** | Yes | 200 | 50 |
| | No | 150 | 600 |

Table: Frequency counts

# Conditional probability

**Example:** In an agricultural region with 1000 farms, we want to know if the farm has vineyards or cork trees.

| | | **Cork Trees** | |
|---|---|---|---|
| | | Yes | No |
| **Vineyard** | Yes | 20% | 5% |
| | No | 15% | 60% |

Table: Joint probabilities

**Questions**:

- What is the probability of seeing cork trees in a farm with vineyards?
- Among farms with cork trees or vineyards, what is the probability of having both?

# Conditional probability

Let's assume the following joint probabilties

| | | Cork Trees | |
|---|---|---|---|
| | | Yes | No |
| **Vineyard** | Yes | 25% | 25% |
| | No | 25% | 25% |

We have that $P(A \cap B) = P(A) \cdot P(B)$, meaning that they are *independent*

# Law of total probability

Let $B_1, B_2, \ldots, B_k \in \mathcal{B}$ and $P(B_i) > 0 : i = 1, \ldots, k$. The *law of total probability* states that

$$P(A) = \sum_{i=1}^{k} P(B_i) P(A|B_i) \tag{80}$$

# Law of total probability

Let $B_1, B_2, \ldots, B_k \in \mathcal{B}$ and $P(B_i) > 0 : i = 1, \ldots, k$. The *law of total probability* states that

$$P(A) = \sum_{i=1}^{k} P(B_i) P(A|B_i) \tag{80}$$

The *conditional law of total probability* states that

$$P(A|C) = \sum_{i=1}^{k} P(B_i|C) P(A|B_i, C) \tag{81}$$

# Bayes' Theorem

Let $B_1, B_2, \ldots, B_k \in \mathcal{B}$, $P(B_i) > 0 : i = 1, \ldots, k$, and $P(A) > 0$. Then Bayes' Theorem states that for $i = 1, \ldots, k$

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{k} P(B_j)P(A|B_j)} \tag{82}$$

n.b. Can be proven using the def of conditional probability

**Example**: You test positive for disease $X$, which has 90% sensitivity and a FPR of 10%. Past genetic screening has indicated that you have a 1 in $10,000$ chance of having the disease. What is the probability of having disease $X$?

# Bayes' Theorem

**Example**: You test positive for disease $X$, which has 90% sensitivity and a FPR of 10%. Past genetic screening has indicated that you have a 1 in $10,000$ chance of having the disease. What is the probability of having disease $X$?

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} \quad (83)$$

$$= \frac{(0.9)(0.0001)}{(0.9)(0.0001) + (0.1)(0.9999)} = 0.0009 \quad (84)$$

# Bayes' Theorem

**Example**: You test positive for disease $X$, which has 90% sensitivity and a FPR of 10%. Past genetic screening has indicated that you have a 1 in $10,000$ chance of having the disease. What is the probability of having disease $X$?

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} \quad (83)$$

$$= \frac{(0.9)(0.0001)}{(0.9)(0.0001) + (0.1)(0.9999)} = 0.0009 \quad (84)$$

Notes:

- $P(B_1)$ is often referred to as the *prior* probability
- $P(B_1|A)$ is often referred to as the *posterior* probability

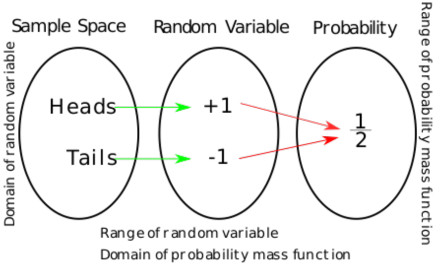
# Random variables

A *random variable* is a (Borel measureable) function
$X : S \to \mathbb{R}$

# Random variables

A *random variable* is a (Borel measureable) function
$X : S \to \mathbb{R}$
**Example**: For coin tossing, we have $X : \{Heads, Tails\} \to \mathbb{R}$, where

$$X(s) = \begin{cases} 1 & \text{if } s = Heads \\ 0 & \text{if } s = Tails \end{cases} \tag{85}$$



# Cumulative distribution function

The *cumulative distribution function* (cdf) of a random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$.

# Cumulative distribution function

The *cumulative distribution function* (cdf) of a random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$.

**Example**: For coin tossing, we have $X : \{Heads, Tails\} \to \mathbb{R}$,

where

$$X(s) = \begin{cases} 1 & \text{if } s = Heads \\ 0 & \text{if } s = Tails \end{cases} \quad (86)$$

we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} \quad (87)$$

# Cumulative distribution function

The *cumulative distribution function* (cdf) of a random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$.

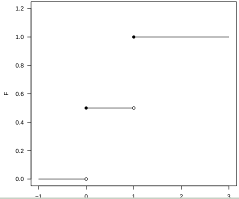
**Example**: For coin tossing, we have $X : \{Heads, Tails\} \to \mathbb{R}$,

where

$$X(s) = \begin{cases} 1 & \text{if } s = Heads \\ 0 & \text{if } s = Tails \end{cases} \quad (86)$$

we have

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} \quad (87)$$



# Cumulative distribution function

n.b. We have two ways of thinking about probabilities:

1. Probability functions
2. Cumulative distribution functions

**Question**: Which one should we use?

# Cumulative distribution function

n.b. We have two ways of thinking about probabilities:

1. Probability functions
2. Cumulative distribution functions

**Question**: Which one should we use?

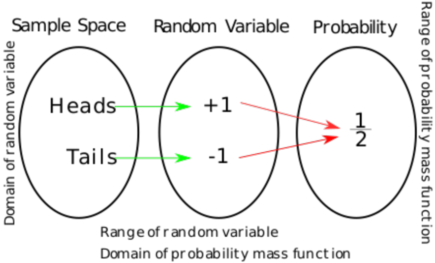
**The Correspondence Theorem**: Let $P_X(\cdot)$ and $P_Y(\cdot)$ be probability functions and $F_X(\cdot)$ and $F_Y(\cdot)$ be their associated cdfs. Then

$$P_X(\cdot) = P_Y(\cdot) \iff F_X(\cdot) = F_Y(\cdot) \tag{88}$$

# Cumulative distribution function

Some properties for cdfs:

- $\lim_{x \Rightarrow -\infty} F(x) = 0$
- $\lim_{x \Rightarrow \infty} F(x) = 1$
- $F(\cdot)$ is non-decreasing
- $F(\cdot)$ is right-continuous



# Quantile function

Let $X$ be a continuous rv and one-to-one over the the possible values of $X$. Then

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\} \tag{89}$$

Is the quantile function of $X$.

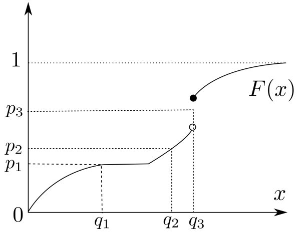
# Quantile function

Let $X$ be a continuous rv and one-to-one over the the possible values of $X$. Then

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\} \quad (89)$$

Is the quantile function of $X$. Let $X$ be a *discrete* rv and one-to-one over the the possible values of $X$. Then $F^{-1}(p)$ states that we take the smallest value of x.

**Example:**



# Nature of random variables

A random variable $X$ is

- *Discrete* if $\exists\, f_X : \mathbb{R} \to [0,1] \ni F_X(x) = \sum_{t \le x} f_X(t), x \in \mathbb{R}$
  - $f_X$ is referred to as the probability mass function (pmf)
- *Continuous* if $\exists\, f_X : \mathbb{R} \to \mathbb{R}_+ \ni F_X(x) = \int_{-\infty}^{x} f_X(t)dt, x \in \mathbb{R}$
  - $f_X$ is referred to as the probability density function (pdf).
  - n.b. We can have multiple pdf's consistent with the same cdf.
  - n.b. For any specific value of a continuous random variable, its probability is 0, i.e. $P(\{x\}) = 0\, \forall x \in \mathbb{R}$.

# Nature of random variables

A random variable $X$ is

- *Discrete* if $\exists\, f_X : \mathbb{R} \to [0,1] \ni F_X(x) = \sum_{t \leq x} f_X(t), x \in \mathbb{R}$
  - $f_X$ is referred to as the probability mass function (pmf)
- *Continuous* if $\exists\, f_X : \mathbb{R} \to \mathbb{R}_+ \ni F_X(x) = \int_{-\infty}^{x} f_X(t)dt, x \in \mathbb{R}$
  - $f_X$ is referred to as the probability density function (pdf).
  - n.b. We can have multiple pdf's consistent with the same cdf.
  - n.b. For any specific value of a continuous random variable, its probability is 0, i.e. $P(\{x\}) = 0\, \forall x \in \mathbb{R}$.

n.b. pmf's and pdf's sum to 1, i.e.

- $f : \mathbb{R} \to [0,1]$ is the pmf of a discrete RV iff $\sum_{x \in \mathbb{R}} f(x) = 1$
- $f : \mathbb{R} \to \mathbb{R}_+$ is the pdf of a continuous RV iff $\int_{-\infty}^{\infty} f(x)dx = 1$

**Example #1**: Coin tossing

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1 \end{cases} \tag{90}$$

Here, $F_X$ is a step function with pmf

$$f_X(x) = \begin{cases} \frac{1}{2} & x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases} \tag{91}$$

**Example #2**: Uniform distribution on (0,1)

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} \tag{92}$$

Here, $F_X$ is a continuous function. Two consistent pdfs include

$$f_X(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{93}$$

$$f_X(x) = \begin{cases} 1 & x \in (0, 1) \\ 0 & \text{otherwise} \end{cases} \tag{94}$$

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and $X$ is a *discrete* rv with cdf $F_X$.

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and $X$ is a *discrete* rv with cdf $F_X$.

Since the function is applied to a rv, $Y$ is also a random variable with probability function

$$f_Y(y) = P_Y(g(X) = y) = \sum_{x:g(x)=y} f_X(x) \tag{95}$$

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and $X$ is a *discrete* rv with cdf $F_X$.

Since the function is applied to a rv, $Y$ is also a random variable with probability function

$$f_Y(y) = P_Y(g(X) = y) = \sum_{x:g(x)=y} f_X(x) \tag{95}$$

**Example:**

Let $X$ be a uniform random variable on $\{-n, -n+1, ..., n-1, n\}$. Then $Y = |X|$ has mass function

$$f_Y(y) = \begin{cases} \frac{1}{2n+1} & \text{if } x = 0 \\ \frac{2}{2n+1} & \text{if } x \neq 0 \end{cases} \tag{96}$$

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and rv $X$ with cdf $F_X$.

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and rv $X$ with cdf $F_X$.

Then $Y$ is also a random variable with cdf

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int x : g(x) \leq y f_X(x) dx \tag{97}$$

We can get the probability function by taking the derivative

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y) \tag{98}$$

# Transformations of random variables

Suppose $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ and rv $X$ with cdf $F_X$.

Then $Y$ is also a random variable with cdf

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int x : g(x) \leq y f_X(x) dx \tag{97}$$

We can get the probability function by taking the derivative

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y) \tag{98}$$

**Example:**

Let $X$ be a uniform rv on $[-1, 1]$. Then $Y = X^2$ has cdf

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) = P_X(X^2 \leq y) = P_X(-y^{1/2} X \leq y^{1/2}) \\ = \int_{-y^{1/2}}^{y^{1/2}} f(x) dx = y^{1/2} \end{aligned} \tag{99}$$

and $f_Y(y) = \frac{\partial}{\partial y} F_Y(y) = \frac{1}{2y^{1/2}}$

# Affine transformations

Suppose $Y = g(X) = aX + b, a > 0, b \in \mathbb{R}$. Then

$$P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \tag{100}$$

# Affine transformations

Suppose $Y = g(X) = aX + b, a > 0, b \in \mathbb{R}$. Then

$$P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \tag{100}$$

If $a < 0$, then

$$P(Y \leq y) = P(aX+b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1-F_X\left(\frac{y-b}{a}\right) \tag{101}$$

# Affine transformations

Suppose $Y = g(X) = aX + b, a > 0, b \in \mathbb{R}$. Then

$$P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \tag{100}$$

If $a < 0$, then

$$P(Y \leq y) = P(aX+b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1-F_X\left(\frac{y-b}{a}\right) \tag{101}$$

In general, as long as the transformation $Y = g(X)$ is monotonic, then

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right| \tag{102}$$

# References

- Grinstead & Snell Chapters 1,2,4
- DeGroot & Schervish Chapters 1,2,3

# Statistics

# Expectation

The *expected value* of rv $X$ is defined as

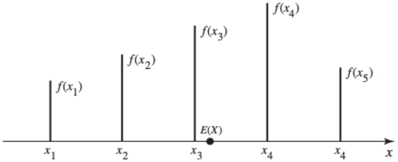
$$\mathbb{E}[X] = \begin{cases} \sum_x x f_X(x) & \text{if x is discrete} \\ \int x f_X(x) dx & \text{if x is continuous} \end{cases} \tag{103}$$

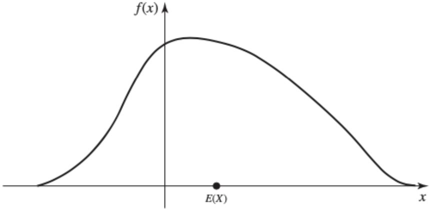
For functions $g$ of $X$,

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) f_X(x) & \text{if x is discrete} \\ \int g(x) f_X(x) dx & \text{if x is continuous} \end{cases} \tag{104}$$

n.b. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$

**Examples**:



**Important:** Expectations might not exist!

**Example:** Suppose $f_X(x) = \frac{1}{x^2}$, defined on $[1, \infty]$. Then

$$\mathbb{E}[X] = \int x f_X(x) dx = \int x \frac{1}{x^2} dx = \int \frac{1}{x} dx = \infty \qquad (105)$$

# Expectation

**Important:** Expectations might not exist!

**Example:** Suppose $f_X(x) = \frac{1}{x^2}$, defined on $[1, \infty]$. Then

$$\mathbb{E}[X] = \int x f_X(x) dx = \int x \frac{1}{x^2} dx = \int \frac{1}{x} dx = \infty \tag{105}$$

Some properties of expectations:

- Linearity: $\mathbb{E}[ag(X) + bh(X)] = \mathbb{E}[ag(X)] + \mathbb{E}[bh(X)]$
- Order preserving:
  $g(X) \leq h(X), \forall x \in \mathbb{R} \Rightarrow \mathbb{E}[g(X)] \leq \mathbb{E}[h(X)]$

# Variance

The *variance* of rv $X$ is defined as

$$var(X) = \mathbb{E}[(X - \mu)^2] : \mu = \mathbb{E}[X] \quad (106)$$

The *variance* of rv $X$ is defined as

$$var(X) = \mathbb{E}[(X - \mu)^2] : \mu = \mathbb{E}[X] \quad (106)$$

Some notes:

- If $\mathbb{E}[X]$ doesn't exist then $var(X)$ doesn't exist.
- $var(X)$ can be infinite.
- The standard deviation $\sigma$ of $X$ is $\sqrt{var(X)}$.

# Variance

With some algebra, we see that

$$\begin{aligned} var(X) &= \mathbb{E}[(X-\mu)^2] && (107) \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] && (108) \\ &= \mathbb{E}[X^2] - \mathbb{E}[2X\mu] + \mathbb{E}[\mu^2] && (109) \\ &= \mathbb{E}[X^2] - \mu^2 && (110) \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (111) \end{aligned}$$

# Variance

Some properties:

- If $X$ is bounded, then $var(X)$ exists and is finite.
- $var(X) = 0 \iff P(X = c) = 1$ for some constant $c$.
- $var(cX) = c^2 var(X)$ for some constant $c$.
- variance is linear, i.e. $var(X_1 + X_2) = var(X_1) + var(X_2)$.

# Moments

The $k^{th}$ *moment* of rv $X$ is defined as

$$\mathbb{E}[X^k] = \mu'_k : k \in \mathbb{N} \tag{112}$$

The $k^{th}$ *central/centered moment* of rv $X$ is defined as

$$\mathbb{E}[(X - \mu)^k] = \mu_k : k \in \mathbb{N} \tag{113}$$

# Moments

The *$k^{th}$ moment* of rv $X$ is defined as

$$\mathbb{E}[X^k] = \mu'_k : k \in \mathbb{N} \tag{112}$$

The *$k^{th}$ central/centered moment* of rv $X$ is defined as

$$\mathbb{E}[(X - \mu)^k] = \mu_k : k \in \mathbb{N} \tag{113}$$

Notes:

- $\mu'_k$ exists if and only if $\mathbb{E}[|X|^k] < \infty$.
- If $\mu'_k$ exists, then for all $j < k$, $\mu'_j$ also exists.
- Variance is $\mu_2$.
- *Skewness* is $\mu_3/\sigma^2$.
- *Kurtosis* is $\mu_4/\sigma^4$.

# Moments

**Example:** Suppose $X \sim N(0,1) \ni f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

$$\mu_1' = \mathbb{E}[X] = \int x f_X(x) dx = f_X(x)|_{-\infty}^{\infty} = 0 \tag{114}$$

n.b. For the normal distribution, $x f_X(x) = -\frac{\partial}{\partial x} f_X(x)$.

# Moments

**Example:** Suppose $X \sim N(0,1) \ni f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

$$\mu_1' = \mathbb{E}[X] = \int x f_X(x)dx = f_X(x)|_{-\infty}^{\infty} = 0 \qquad (114)$$

n.b. For the normal distribution, $xf_X(x) = -\frac{\partial}{\partial x} f_X(x)$.

$$\mu_2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[(X-0)^2] = \mathbb{E}[X^2] = \int x^2 f_X(x)dx \qquad (115)$$

using integration by parts, we get

$$\int x^2 f_X(x)dx = \underbrace{-xf_X(x)|_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{\infty}^{\infty} f_X(x)dx}_{=1} = 1 \qquad (116)$$

# Moment generating function

*Moment generating functions* (mgf) are used to calculate the moments of a rv. The mgf of a rv $X$ is a function $M_X : \mathbb{R} \Rightarrow \mathbb{R}_+$ such that

$$M_X(t) = \mathbb{E}[e^{tX}] : t \in \mathbb{R} \tag{117}$$

*Moment generating functions* (mgf) are used to calculate the moments of a rv. The mgf of a rv $X$ is a function $M_X : \mathbb{R} \Rightarrow \mathbb{R}_+$ such that

$$M_X(t) = \mathbb{E}[e^{tX}] : t \in \mathbb{R} \tag{117}$$

Notes:

- The mgf is a function of $t$; $X$ is integrated out by $\mathbb{E}$.
- The mgf only applies if the moments of the rv exists.
- If two rv $X, Y$ have the same mgf (i.e. $M_X(t) = M_Y(t)$), then they have the same distribution.
- Even if a rv has moments, the mgf may yield infinity (e.g. log-normal distribution).

# Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \qquad (118)$$

What happens when $t = 0$?

# Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \qquad (118)$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \qquad (119)$$

# Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \tag{118}$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \tag{119}$$

What happens when $t = 0$ for the $k^{th}$ derivative?

# Moment generating function

Taking the derivative of the mgf, we see that

$$\frac{\partial}{\partial t} M_X(t) = \frac{\partial}{\partial t} \int e^{tx} f_X(x) dx = \int x \cdot e^{tx} f_X(x) dx \qquad (118)$$

What happens when $t = 0$?

$$\int x \cdot e^{tx} f_X(x) dx = \int x f_X(x) dx = \mathbb{E}[X] \qquad (119)$$

What happens when $t = 0$ for the $k^{th}$ derivative?

$$\frac{\partial}{\partial t^k} M_X(t) = \int x^k \cdot e^{tx} f_X(x) dx \qquad (120)$$

At $t = 0$, we get $\frac{\partial}{\partial t^k} M_X(t)|_{t=0} = \mathbb{E}[X^k]$

**Evaluating the $k^{th}$ derivative at $t = 0$ gives us the $k^{th}$ moment of $X$.**

# Moment generating function

**Example:** The standard normal distribution

$$
\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] = \int e^{tX} f_X(x)dx && (121)\\
&= \int e^{tX} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx && (122)\\
&= \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) \exp\left(\frac{t^2}{2}\right) dx && (123)\\
&= \exp\left(\frac{t^2}{2}\right) \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) dx && (124)\\
&= \exp\left(\frac{t^2}{2}\right) && (125)
\end{aligned}
$$

# Moment generating function

The mgf for *affine transformations* is straight forward,
e.g. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.

**Example:** Let $X = \mu + \sigma Z : Z \sim N(0, 1)$. Then

$$M_X(t) = M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{126}$$

# Moment generating function

The mgf for *affine transformations* is straight forward,
e.g. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.

**Example:** Let $X = \mu + \sigma Z : Z \sim N(0,1)$. Then

$$M_X(t) = M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{126}$$

**Another example:**
Let $X_1, \ldots, X_n \stackrel{iid}{\sim} P_0$ and $Y = \sum_{i=1}^n X_i$. Then

$$M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[e^{t(X_1+\cdots+X_n)}] = \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right] \tag{127}$$

$$= \prod_{i=1}^n \mathbb{E}\left[e^{tX_i}\right] = \prod_{i=1}^n M_{X_i}(t) \tag{128}$$

# Distributions

Most useful distributions have names, e.g.

- Normal distribution
- Uniform distribution
- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Gamma distribution

# Normal distribution

A rv $X$ follows a *Normal distribution*, denoted as $X \sim N(\mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$, if $X$ is continuous with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : x \in \mathbb{R} \qquad (129)$$

**Note:**

If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. It follows that

- $\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu$.
- $var(X) = var(\mu + \sigma Z) = \sigma^2 var(Z) = \sigma^2$.

# Normal distribution

A rv $X$ follows a *Normal distribution*, denoted as $X \sim N(\mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$, if $X$ is continuous with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) : x \in \mathbb{R} \qquad (129)$$

**Note:**
If $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$. It follows that

- $\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu$.
- $var(X) = var(\mu + \sigma Z) = \sigma^2 var(Z) = \sigma^2$.

Most well known distribution due to:

1. Good mathematical properties
2. Often (approximately) observed in the real world (e.g. heights, weights, etc.)
3. Central limit theorem

# Central limit theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} P_0$, where $\mathbb{E}[X_i] = \mu$ and $var(X_i) = \sigma^2$.
Then

$$\lim_{n\to\infty} P\left(\frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} \leq x\right) = \Phi(x) \qquad (130)$$

where $\Phi(x)$ is the cdf for the standard normal distribution.

# Central limit theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} P_0$, where $\mathbb{E}[X_i] = \mu$ and $var(X_i) = \sigma^2$. Then

$$\lim_{n\to\infty} P\left(\frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} \leq x\right) = \Phi(x) \tag{130}$$

where $\Phi(x)$ is the cdf for the standard normal distribution.

**Example:** The sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{131}$$

The 95% CI: $\bar{X}_n \pm z_{\alpha/2}\hat{se}_n$

# Uniform distribution

A rv $X$ follows a Uniform distribution $U(a, b)$ if $X$ is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{132}$$

Under $U(a, b)$, all observations are "*equally likely*"
$\mathbb{E}[X] = \frac{a+b}{2}$, $var(X) = \frac{(b-a)^2}{12}$, and $M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$.

# Uniform distribution

A rv $X$ follows a Uniform distribution $U(a, b)$ if $X$ is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{132}$$

Under $U(a, b)$, all observations are "*equally likely*"
$\mathbb{E}[X] = \frac{a+b}{2}$, $var(X) = \frac{(b-a)^2}{12}$, and $M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$.

Note: if $X \sim U(a, b)$, then $X = (b-a)\tilde{X} + a : \tilde{X} \sim U(0, 1)$ and

$$f_{\tilde{X}}(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{133}$$

A rv $X$ follows a Bernoulli distribution $Ber(p)$ if $X$ is discrete with pmf

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{134}$$

$\mathbb{E}[X] = p$, $var(X) = p(1-p)$, and $M_X(t) = e^t p + (1-p)$.

# Binomial distribution

A rv $X$ follows a Binomial distribution $Bin(n, p)$ if $X$ is discrete with pmf

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in \{0, 1, ..., n\} \\ 0 & \text{otherwise} \end{cases} \tag{135}$$

$\mathbb{E}[X] = np$, $var(X) = np(1-p)$, and
$M_X(t) = (e^t p + (1-p))^n$.

If $X_1, ..., X_n \overset{iid}{\sim} Ber(p)$, then $Y = X_1 + \cdots + X_n$ follows $B(n, p)$.

# Negative Binomial distribution

A rv $X$ follows a Negative Binomial distribution $NB(r, p)$ if $X$ is discrete with pmf

$$f_X(x) = \begin{cases} \binom{r+x-1}{x} p^x (1-p)^r & \text{if } x \in \{0, 1, ..., n\} \\ 0 & \text{otherwise} \end{cases} \quad (136)$$

$\mathbb{E}[X] = \frac{r(1-p)}{p}$, $var(X) = \frac{r(1-p)}{p^2}$, and $M_X(t) = \left(\frac{p}{1-qe^t}\right)^r : t < \log\left(\frac{1}{q}\right)$.

When $r = 1$, we refer to it as the *Geometric distribution*.

- It has a *memoryless* property.

# Poisson distribution

A rv $X$ follows a Poisson distribution $Pois(\lambda)$ if $X$ is discrete with pmf

$$f_X(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \tag{137}$$

$\mathbb{E}[X] = \lambda$, $var(X) = \lambda$, and $M_X(t) = e^{\lambda(e^t - 1)}$.

Some notes:

- $Bin(n, p) \approx Pois(np)$ when $n$ is large and $np$ is small.
- "*Poisson Processes*" are typically used to model rates, e.g. mortality rates
  1. The number of events in each fixed time interval $t$ has a Poisson distribution with mean $\lambda t$.
  2. The number of events in each time interval is independent.

# Gamma distribution

A rv $X$ follows a Gamma distribution $Gamma(\alpha, \beta)$ if $X$ is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{138}$$

where $\Gamma(x) = \int_0^\infty t^{\alpha-1} e^{-t} dt : \alpha > 0$.

$\mathbb{E}[X] = \alpha\beta$, $var(X) = \alpha\beta^2$, and
$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} : t < \beta$.

# Gamma distribution

A rv $X$ follows a Gamma distribution $Gamma(\alpha, \beta)$ if $X$ is continuous with pdf

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{138}$$

where $\Gamma(x) = \int_0^\infty t^{\alpha-1} e^{-t} dt : \alpha > 0$.

$\mathbb{E}[X] = \alpha\beta$, $var(X) = \alpha\beta^2$, and
$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} : t < \beta$.

Notes:

- $\frac{1}{\Gamma(\alpha)\beta^\alpha}$ is often referred to as the '*normalizing constant*'.
- When $\alpha = 1$, we get the exponential distribution.

# Beta distribution

A rv $X$ follows a Beta distribution $Beta(\alpha, \beta)$ if $X$ is continuous with pdf

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{139}$$

$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$, $var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, and $M_X(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx$.

n.b. Very popular distribution in Bayesian statistics.

# Multinomial distribution

Suppose rv $\mathbf{X} = (X_1, ..., X_k)$ represents counts of $k$ different classes. Then it follows a Multinomial distribution $Multi(p_1, ..., p_k)$ if it has pdf

$$f_X(x) = \begin{cases} \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k} & x_1 \geq 0, ..., x_k \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (140)$$

where $n = \sum_{i=1}^{k} X_i$.

$\mathbb{E}[X_i] = np$, $var(X_i) = np_i(1 - p_i)$, and $Cov(X_i, X_j) = -np_i p_j$.

# Dirac delta function

While not technically a pdf, often used for e.g. mixture of discrete distributions

The Dirac delta function is defined as $\delta : \mathbb{R} \to \mathbb{R} \cup \infty \ \ni$

$$\delta(x) = \begin{cases} +\infty & x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{141}$$

and $\int_{-\infty}^{\infty} \delta(x) dx = 1$

**The sifting property:**

$$\int f(x)\delta(x-a)dx = f(a) \tag{142}$$

# Dirac delta function

**Example:** Let

$$Y = \begin{cases} 1 & \text{w.p. } \alpha \\ U(0,1) & \text{w.p. } 1-\alpha \end{cases} \tag{143}$$

Then $f_Y(y) = \alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1])$

# Dirac delta function

**Example:** Let

$$Y = \begin{cases} 1 & \text{w.p. } \alpha \\ U(0,1) & \text{w.p. } 1-\alpha \end{cases} \tag{143}$$

Then $f_Y(y) = \alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1])$

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{\infty}^{\infty} y(\alpha\delta(y-1) + (1-\alpha)\mathbb{I}(y \in [0,1]))dy & (144)\\
&= \alpha \int_{\infty}^{\infty} y(\delta(y-1)dy + (1-\alpha)\int_0^1 ydy & (145)\\
&= \alpha + (1-\alpha)\frac{y^2}{2}|_0^1 & (146)\\
&= \alpha + \frac{1-\alpha}{2} & (147)\\
&= \frac{1+\alpha}{2} & (148)
\end{aligned}$$

# References

- DeGroot & Schervish Chapters 4.1-4.5,5.1-5.9
- Grinstead & Snell Chapters 5,6