Math for ML

Week 5.1: Basic Probability Theory, Models, and Data

By: Samuel Deng

Logistics & Announcements

- HW #3 DUE TONIGHT 11:59 PM.
  - ⇒ Late deadline Thurs.
- PS #4 out on the website. (last Thurs). ← next Tues.
- I have OH. today 3PM – 5PM.
- Last one ⇒ PS 5 this Thurs. ~2 Problems. (shorter).

Lesson Overview

Probability Spaces. We'll review the basic axioms and components of probability: sample space, events, and probability measures. This allows us to ditch these notions and introduce *random variables.*

Random variables. Review of the definition of a random variable, its *distribution/law*, its PDF/PMF/CDF, and joint distributions of several RVs.

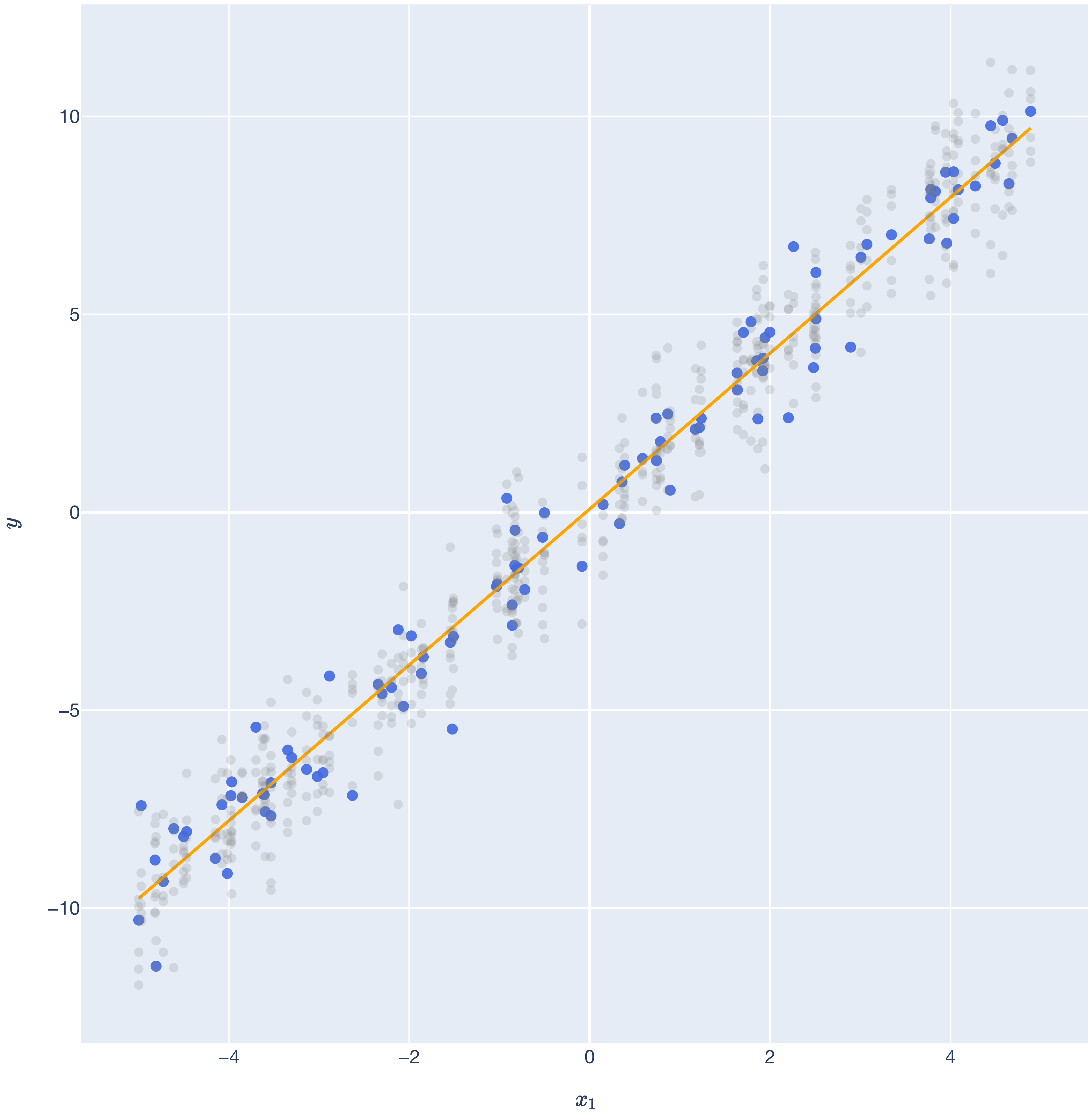
Expectation, variance, and covariance. Review of these basic summary statistics of random variables and common properties.

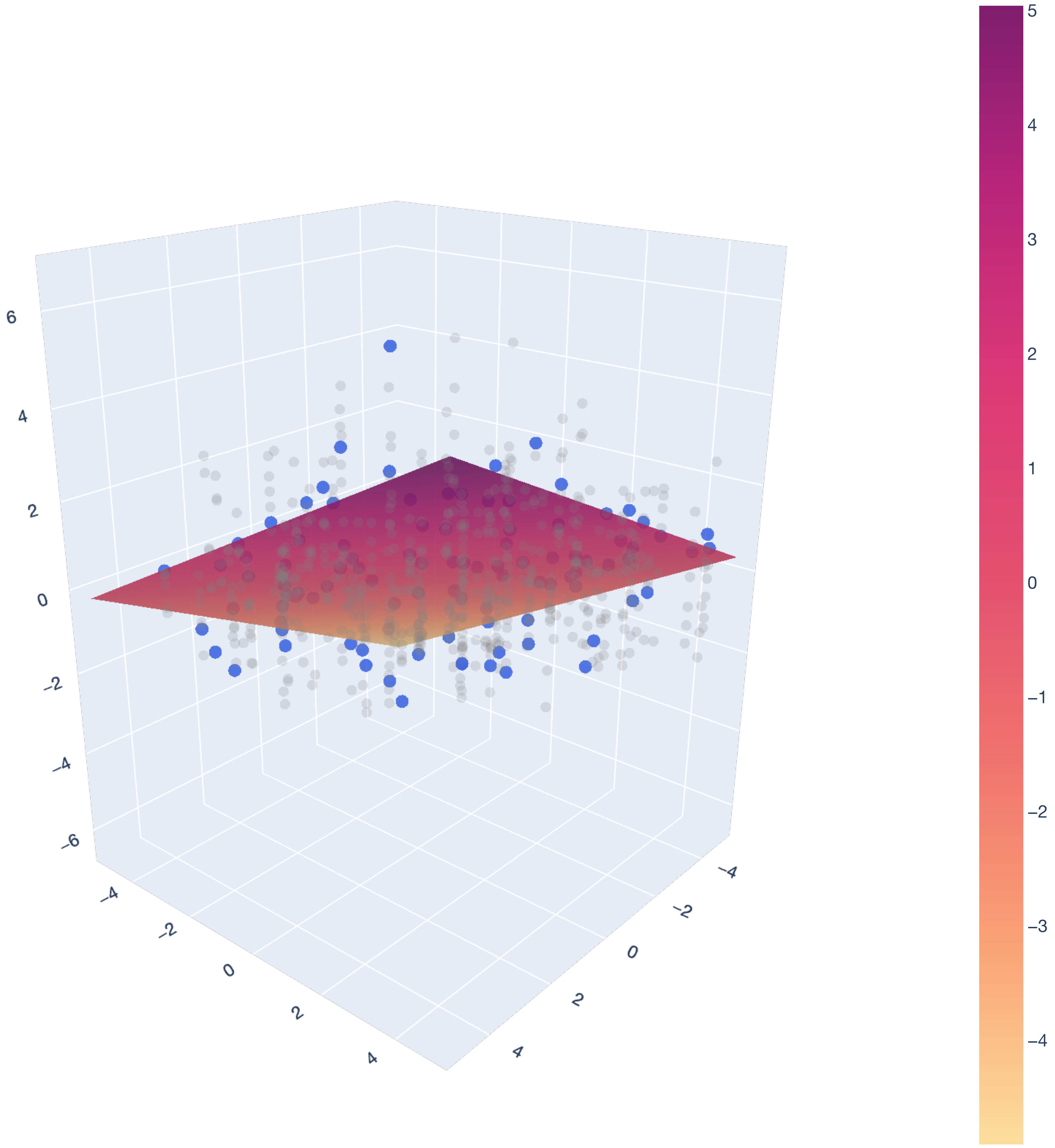
Random vectors. Introduce the idea of a *random vector*, which is just a list of multiple random variables. Discuss generalizations of expectation and variance to random vectors.

Data as random, statistical model of ML. Introduce the statistical model of ML and the random error model. Introduce *modeling assumptions.* State and prove basic statistical properties of the OLS estimator.

Lesson Overview

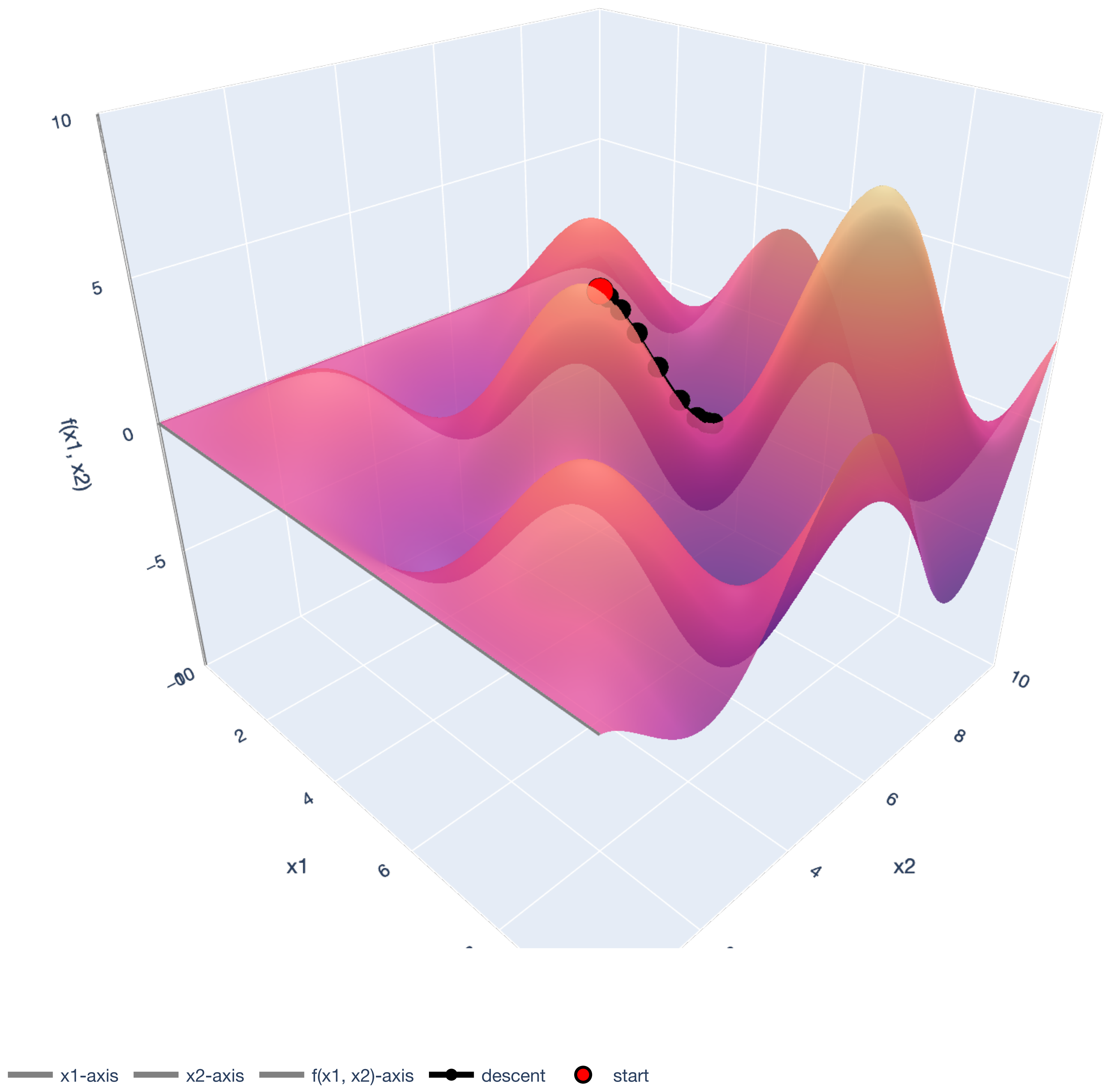
Big Picture: Least Squares

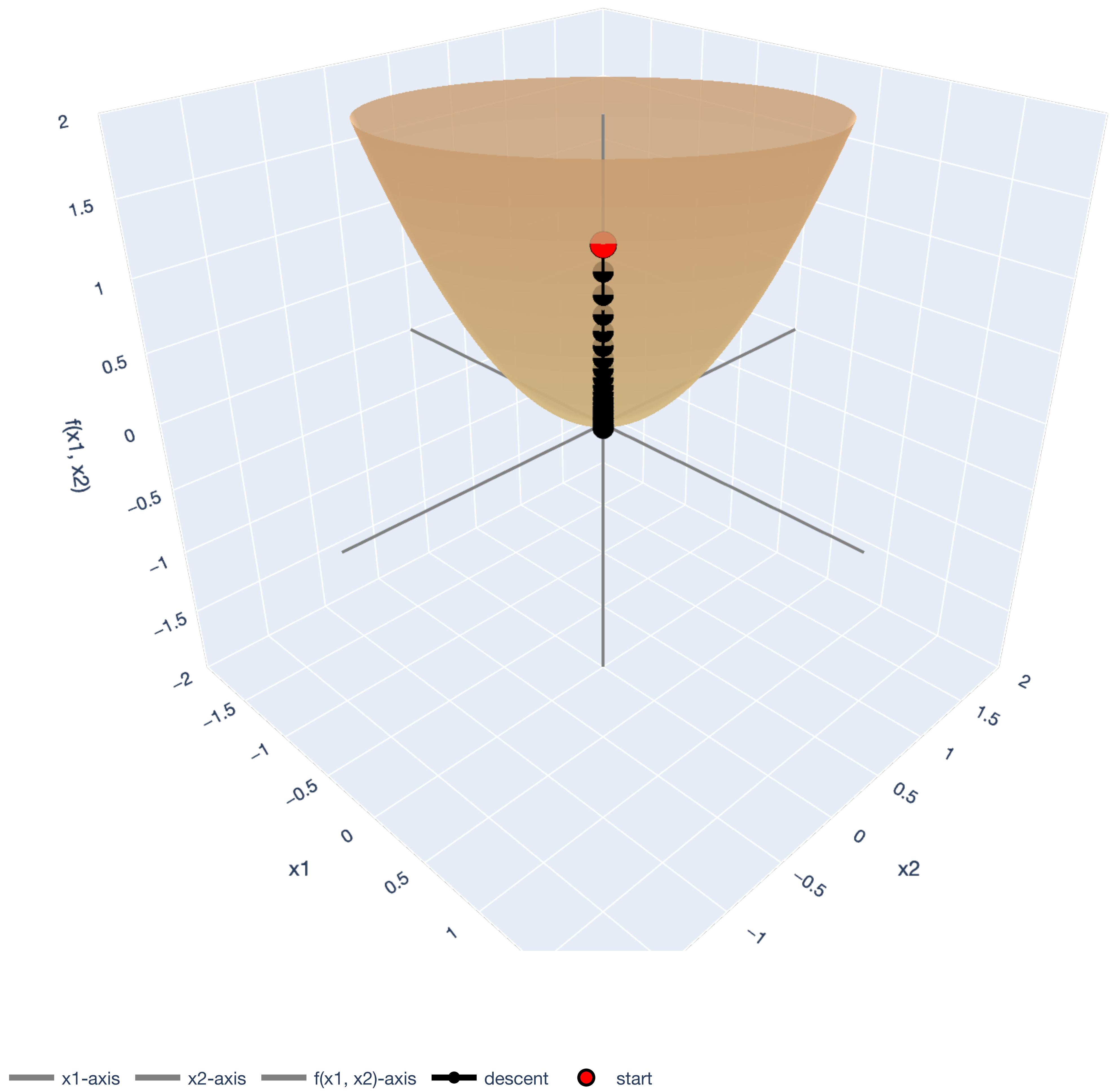




Lesson Overview

Big Picture: Gradient Descent





Motivation

Data as randomly distributed

Regression

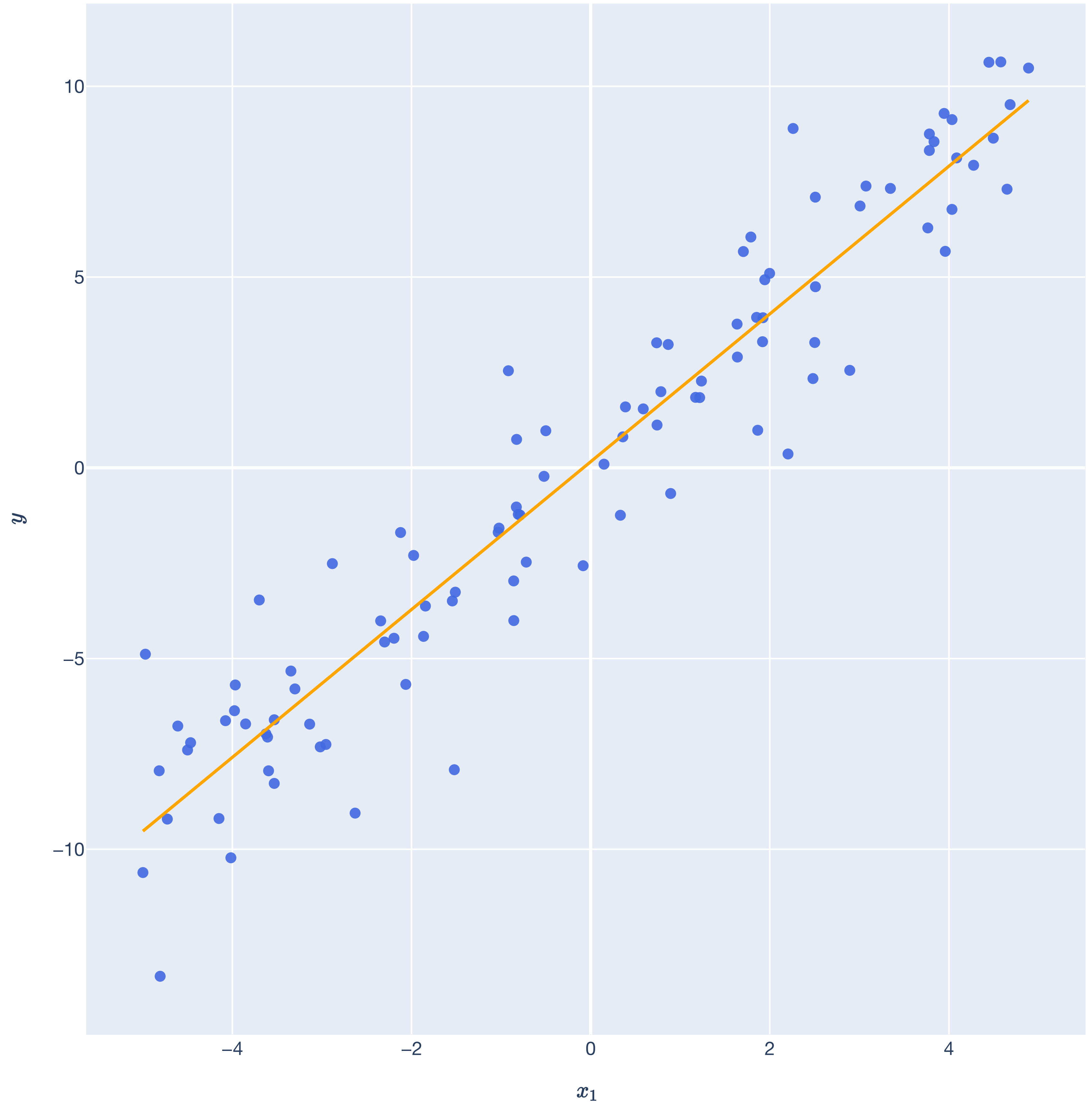
Setup

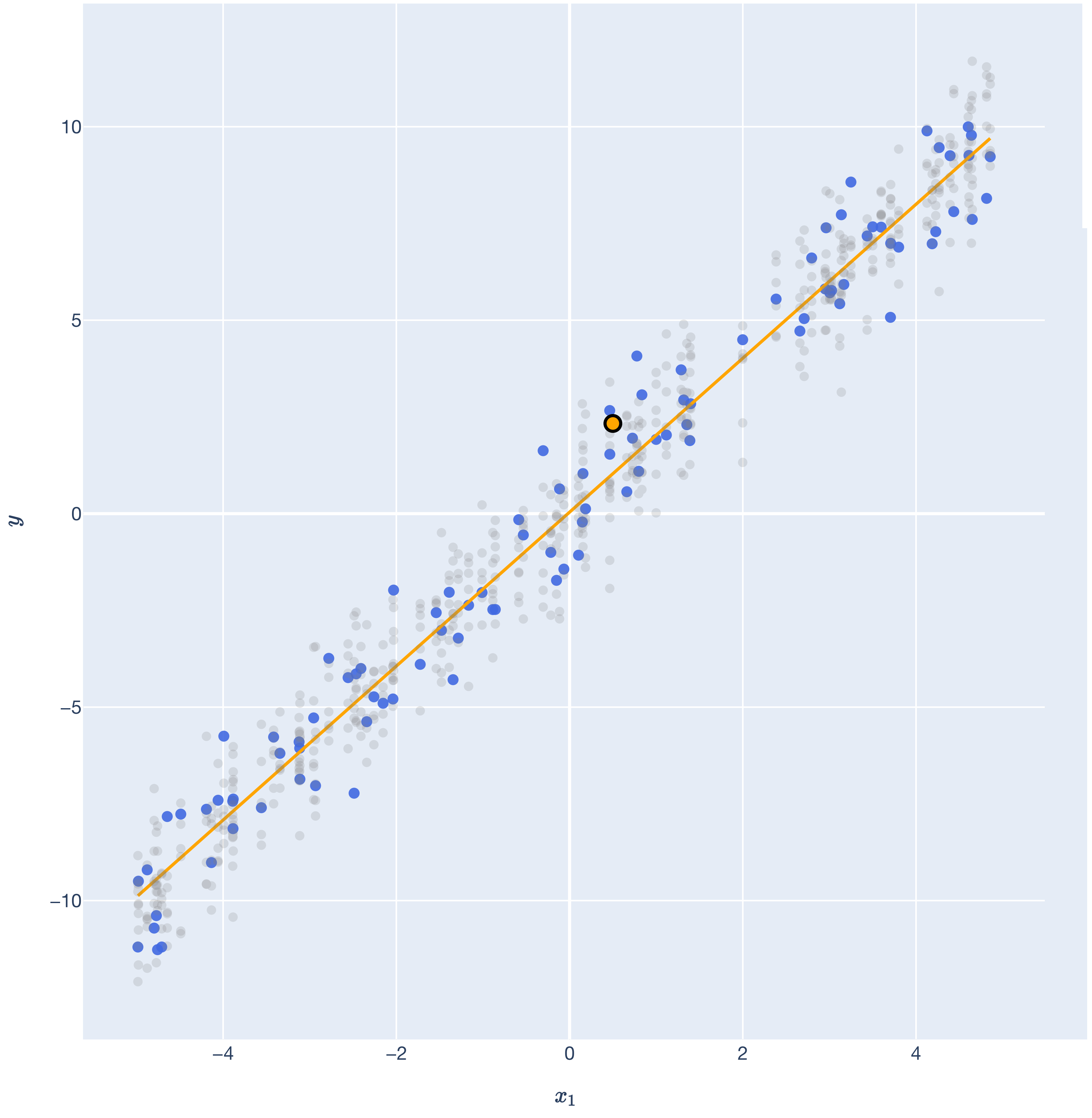
$X \in \mathbb{R}^{n \times d}$
$Y \in \mathbb{R}^{n}$

$\|Xw - y\|^2$

$\in \mathbb{R}^{d}$

Collect labeled training data $\Longrightarrow$ Fit the model $\hat{\mathbf{w}}$ $\Longrightarrow$ Generalize on new $\mathbf{x}_0$





Regression
Setup

Observed: Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

Unknown: *Weight vector* $\mathbf{w} \in \mathbb{R}^d$ with weights $w_1, \ldots, w_d$.

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

Regression
Setup

<u>Goal:</u> For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Regression

Setup

Original Goal: Given a new, unseen $(\mathbf{x}_0, y_0) \in \mathbb{R}^d \times \mathbb{R}$, we wanted to *generalize*:

$$\hat{\mathbf{w}}^\top \mathbf{x}_0 \approx y_0.$$

To do this, we fit the "training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Regression

Setup

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Least squares expanded is just:

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

Put a $1/n$ there, and it looks like we're minimizing an average…

Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

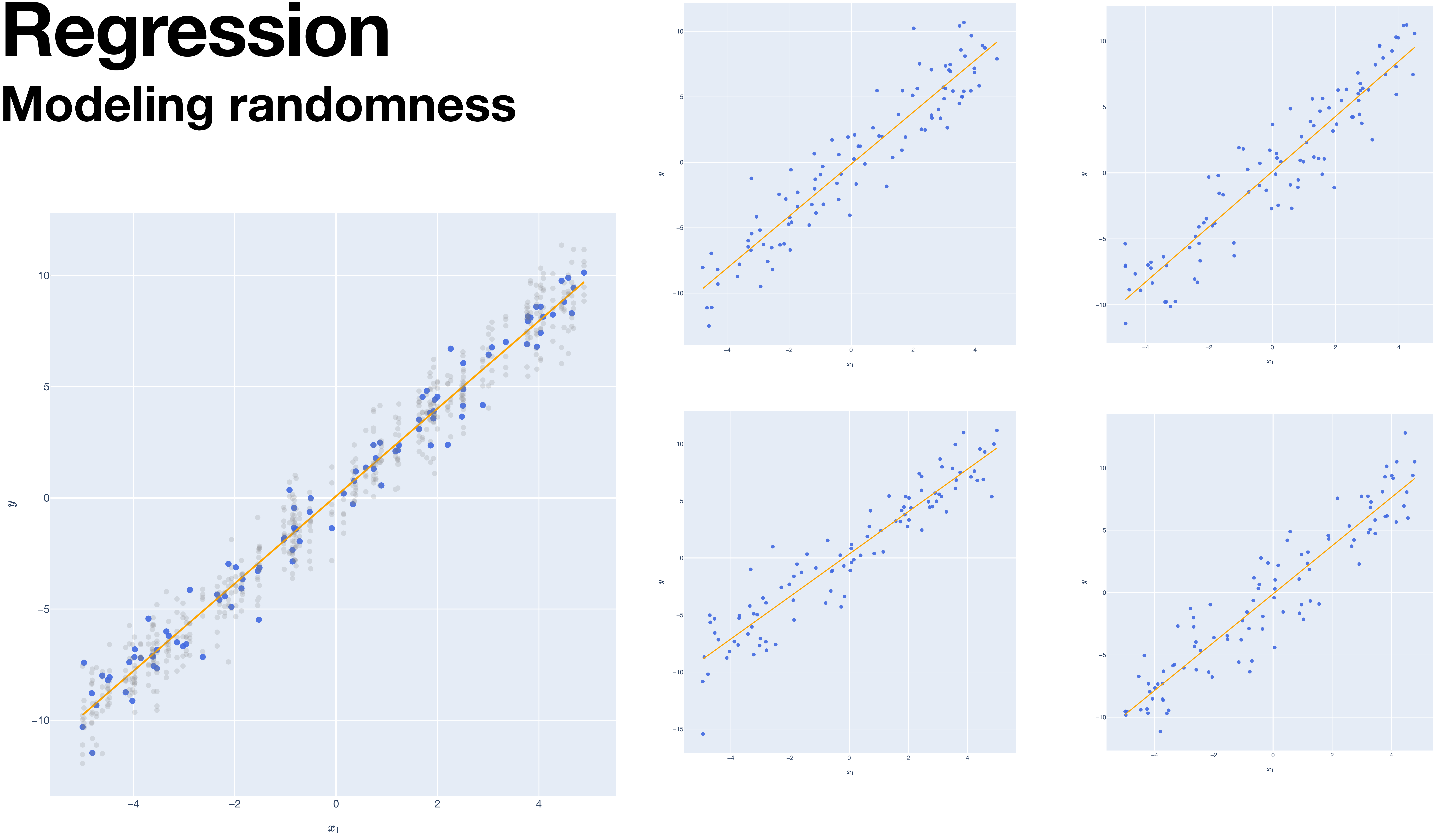
$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Regression

Modeling randomness



Probability Spaces

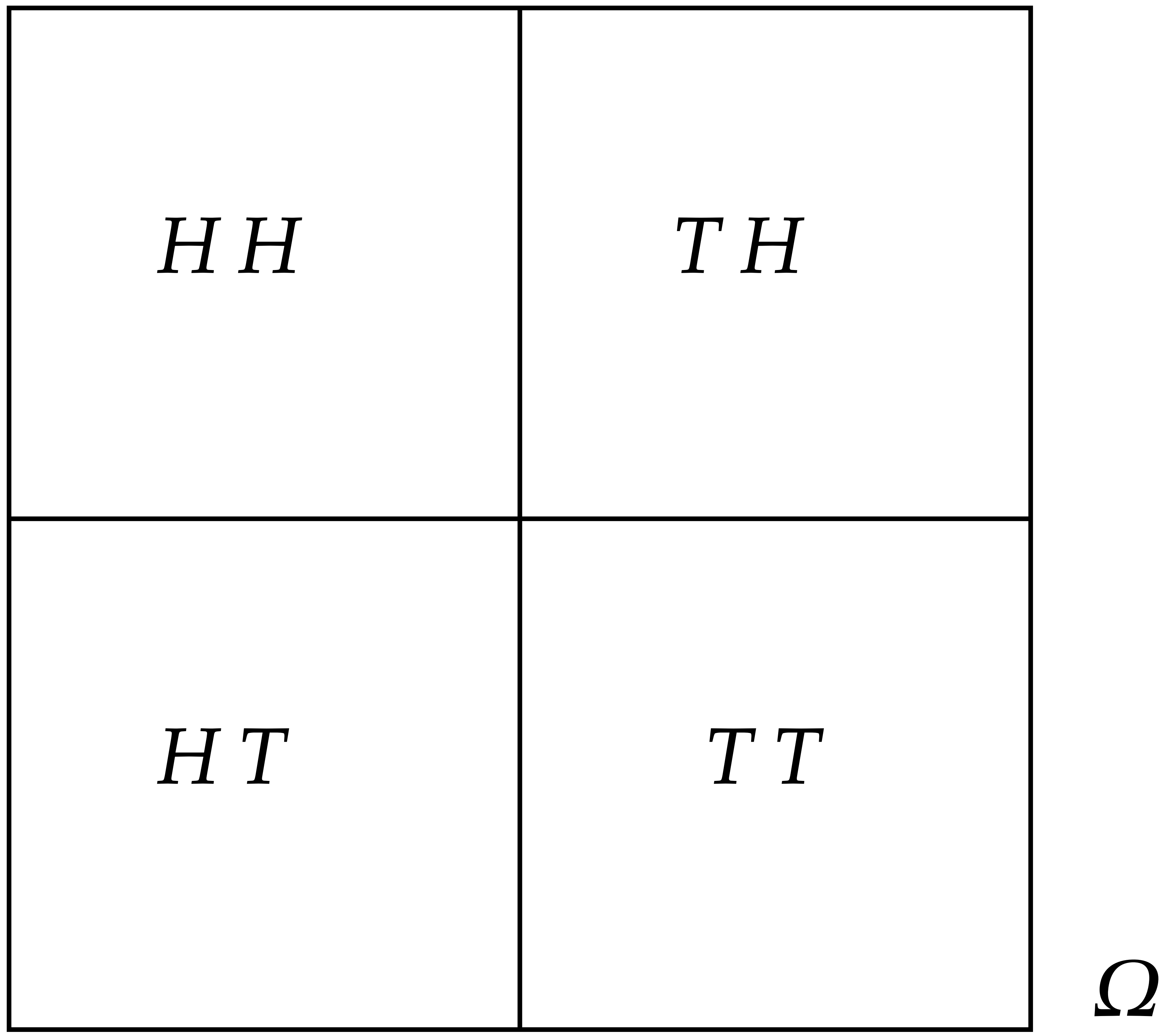
Sample Spaces, Events, and Random Variables

Sample Space

Example: Flipping 2 fair coins

Consider the following *experiment*:

Alice and Bob both have a fair coin. They each flip their coins simultaneously, and the result can be either H or T.



What are the possible outcomes of this experiment?

Sample Space

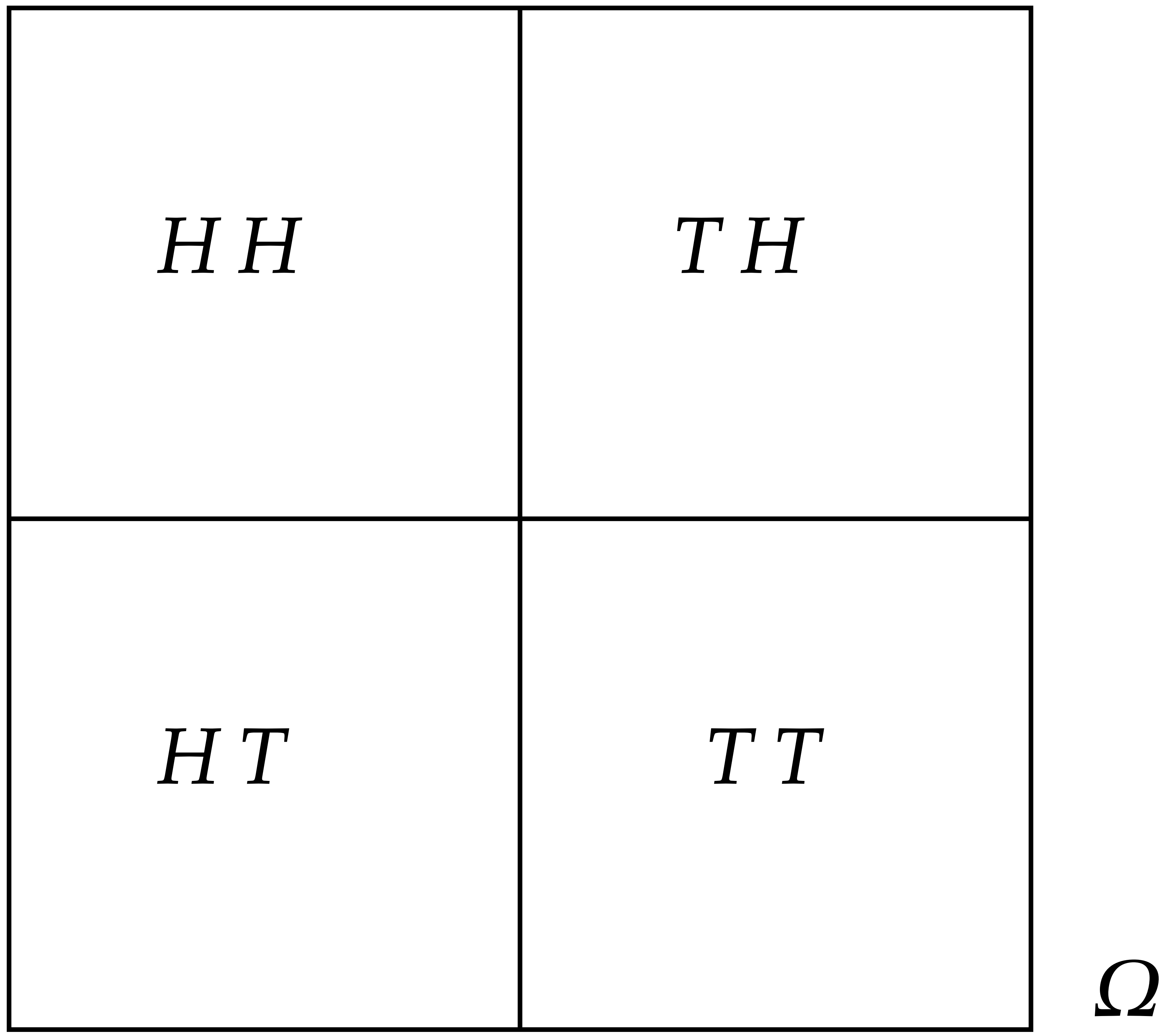
Intuition and definition

The ***sample space*** of some experiment on which we want to model probabilities is the *set* of all possible outcomes. We usually denote this Ω.

DO NOT HAVE TO BE NUMBERS!

Example:

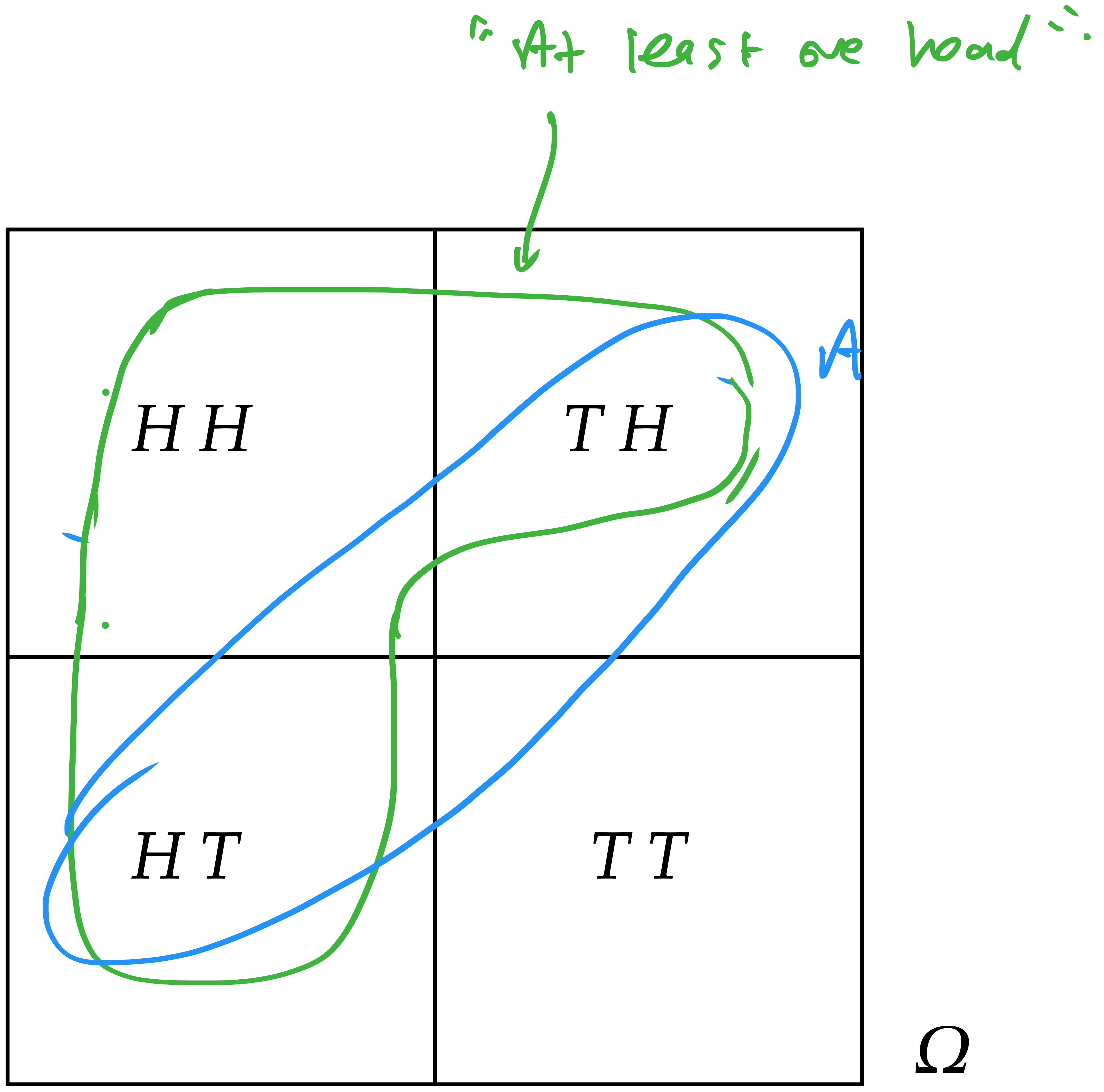
$$\Omega = \{HH, HT, TH, TT\}.$$



Events

Intuition and definition

Given a sample space Ω, an ***event*** is a subset $A \subseteq \Omega$ of outcomes. Denote a collection of events $\mathscr{A}$.



Example:

$$A = \{HT, TH\} = \{\text{"exactly 1 head"}\}$$

$$\mathscr{A} = \{\varnothing, \{HH\}, \{HT\}, \dots, \{HH, HT, TH, TT\}\}$$

$2^4 = 16$

Events

Intuition and definition

Events are subsets, so they obey the usual rules and definitions of set logic.

$A \cup B$ (union)

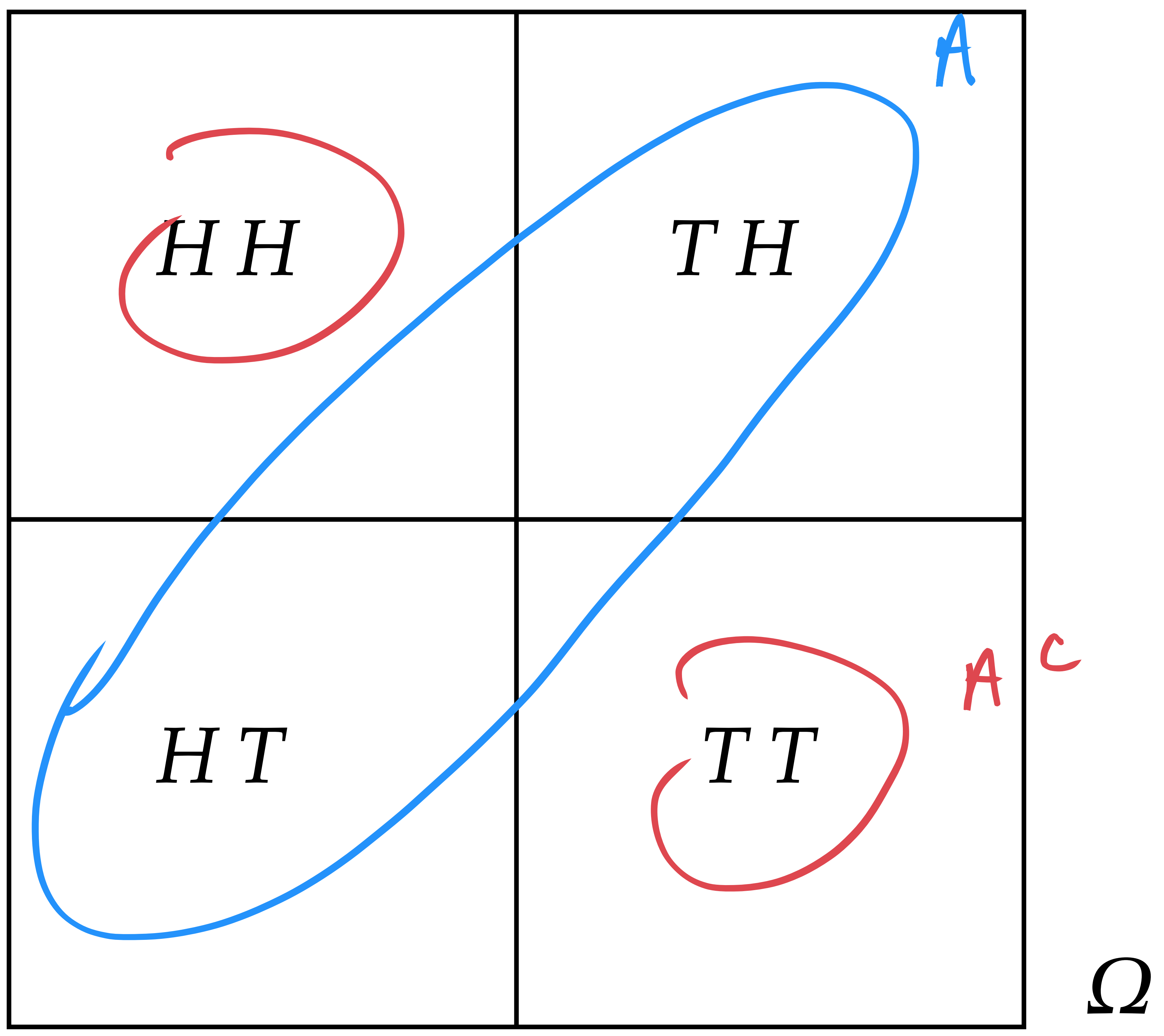
$A \cap B$ (intersection)

A^C (complement)

Example:

$A = \{HT, TH\} = \{$"exactly 1 head"$\}$

$A^C = \{HH, TT\}$



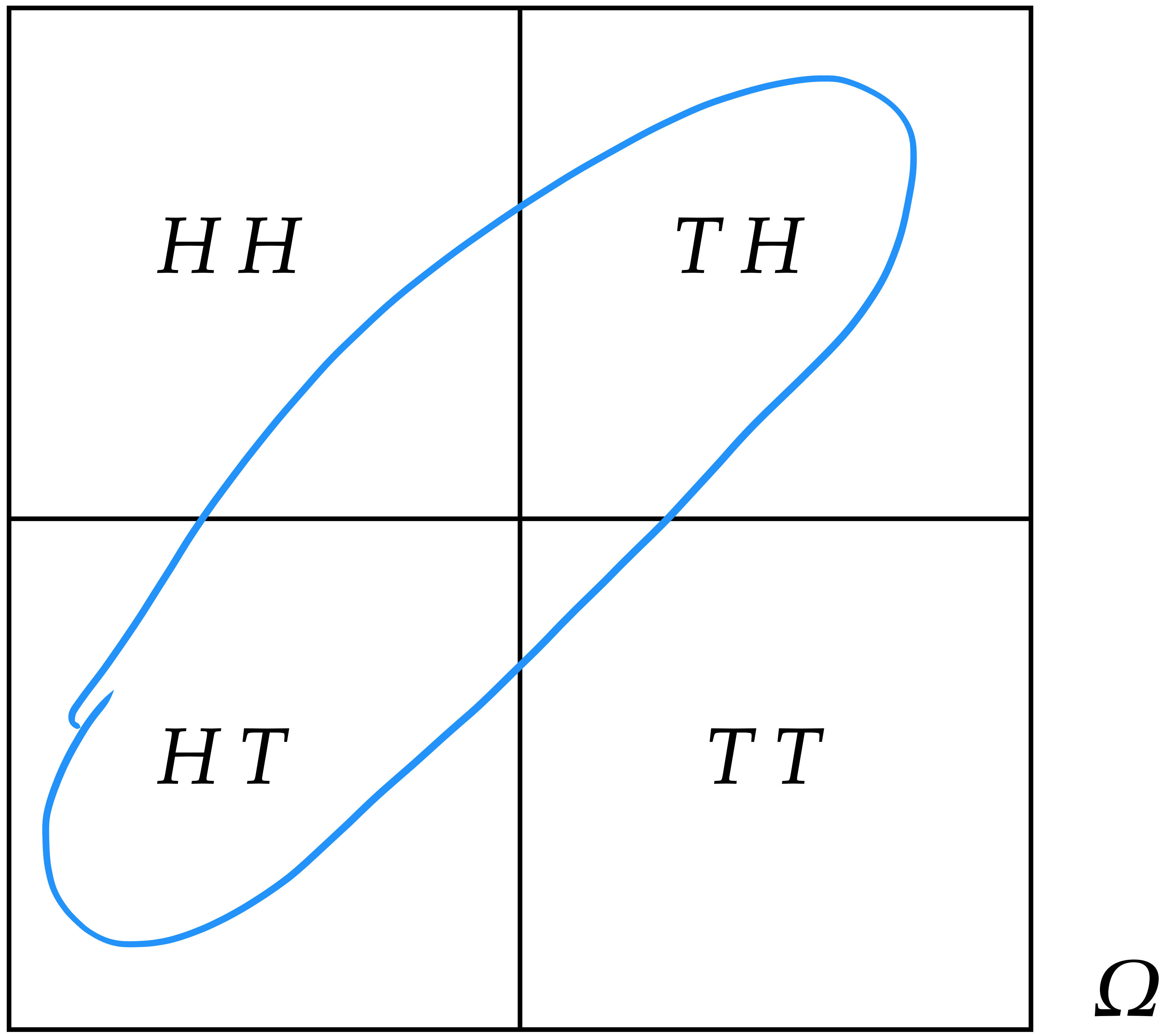
Probability Measure
Intuition and definition

A ***probability measure*** is a set function $\mathbb{P} : \mathscr{A} \to [0,1]$ mapping from sets to a number in $[0,1]$.

For an event $A \in \mathscr{A}$, we call $\mathbb{P}(A)$ the ***probability*** that event A occurs.

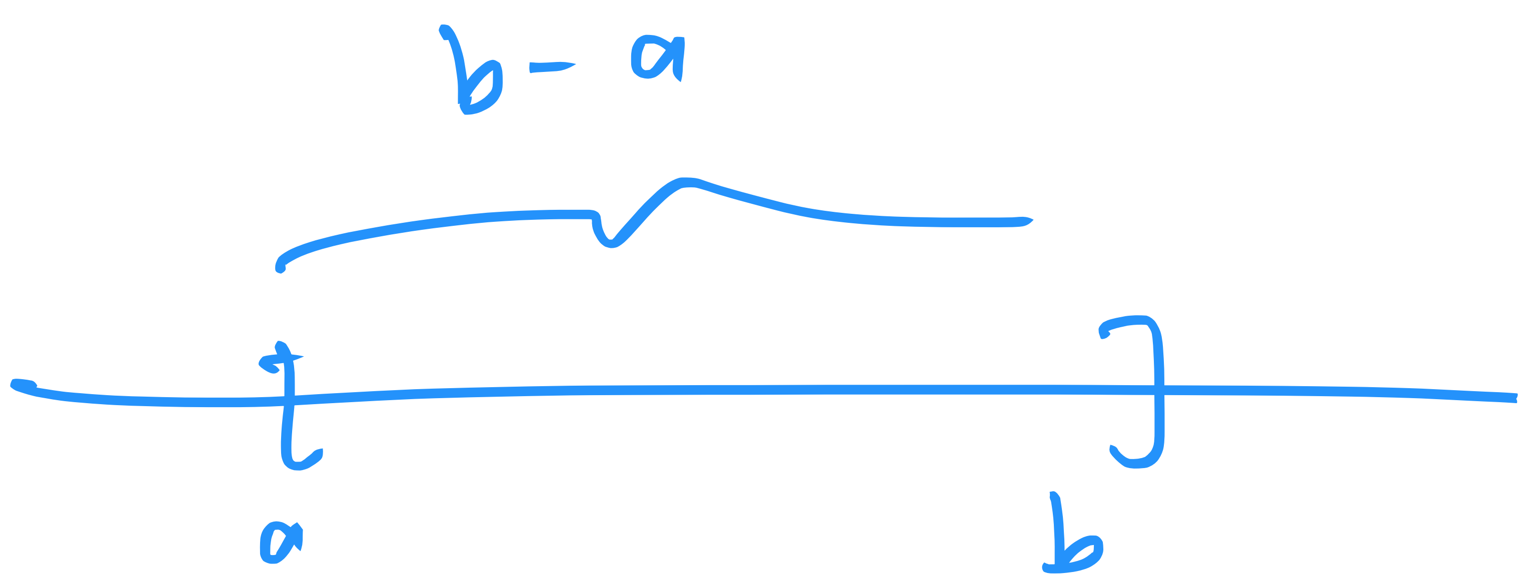
Can be interpreted as *“degree of belief”* or *“long-run frequency.”*

Or just the *“mass”* of a particular subset!



Probability Measure

Axiomatic Properties



Any valid probability measure $\mathbb{P}$ satisfies two properties:

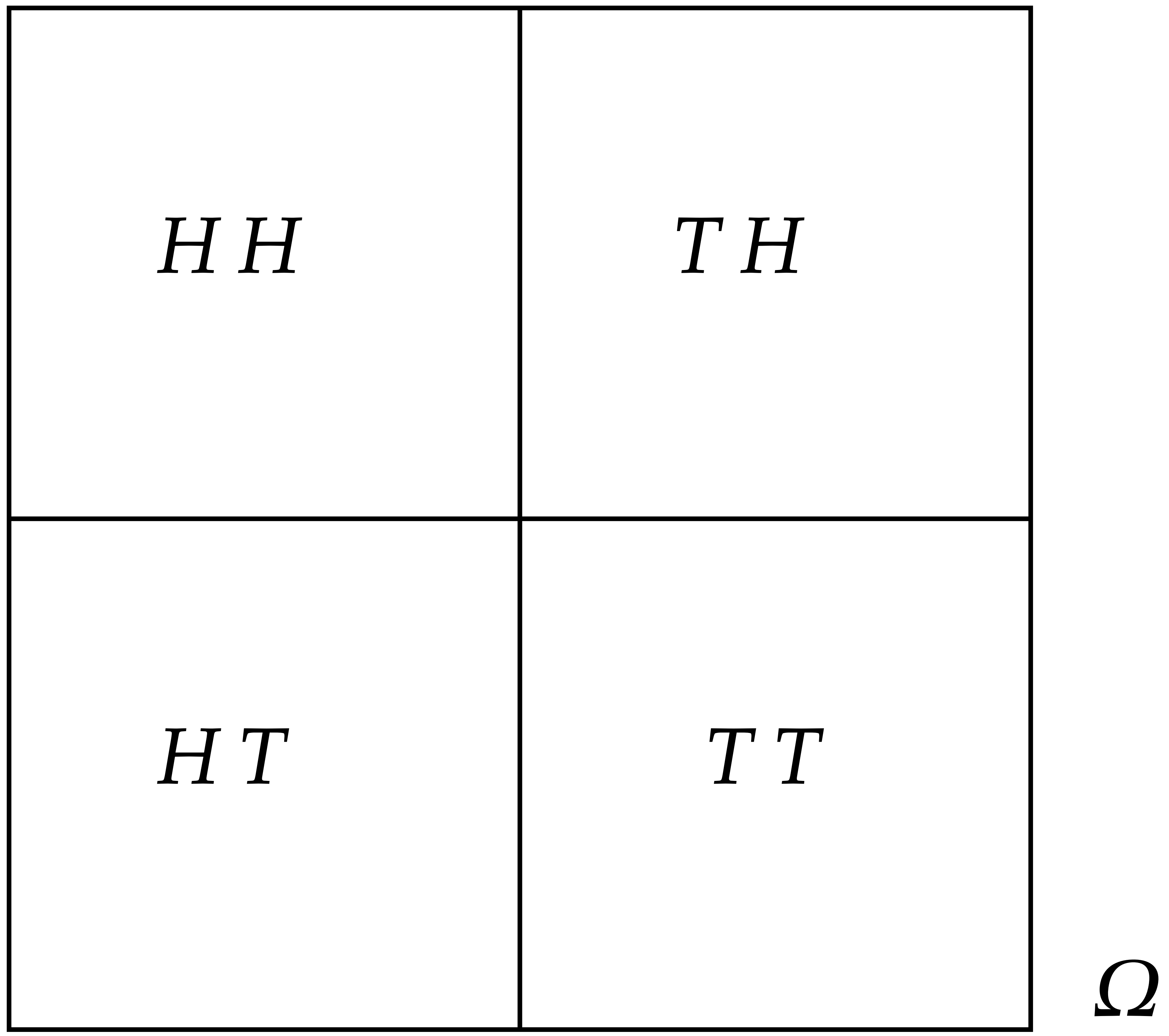
1. The measure of the entire sample space:

$$\mathbb{P}(\Omega) = 1.$$

2. For disjoint events $A_1, A_2, A_3, \ldots$

$$\mathbb{P}\left(A_1 \cup A_2 \cup A_3 \cup \ldots\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \ldots$$

also known as ***countable additivity.***



Probability Measure

Properties of probability measures

1. **Complements.** For any event $A \in \mathscr{A}$, the probability of the complement is:

$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A).$$

2. **Subsets of events.** For two events $A, B \in \mathscr{A}$, if $A \subseteq B$, then:

$$\mathbb{P}(B) \leq \mathbb{P}(A).$$

3. **Unions of events.** For any two events $A, B \in \mathscr{A}$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

4. **Union bound.** For any finite collection of events $A_1, \ldots, A_n$,

$$\mathbb{P}\left(A_1 \cup \ldots \cup A_n\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Probability Measure

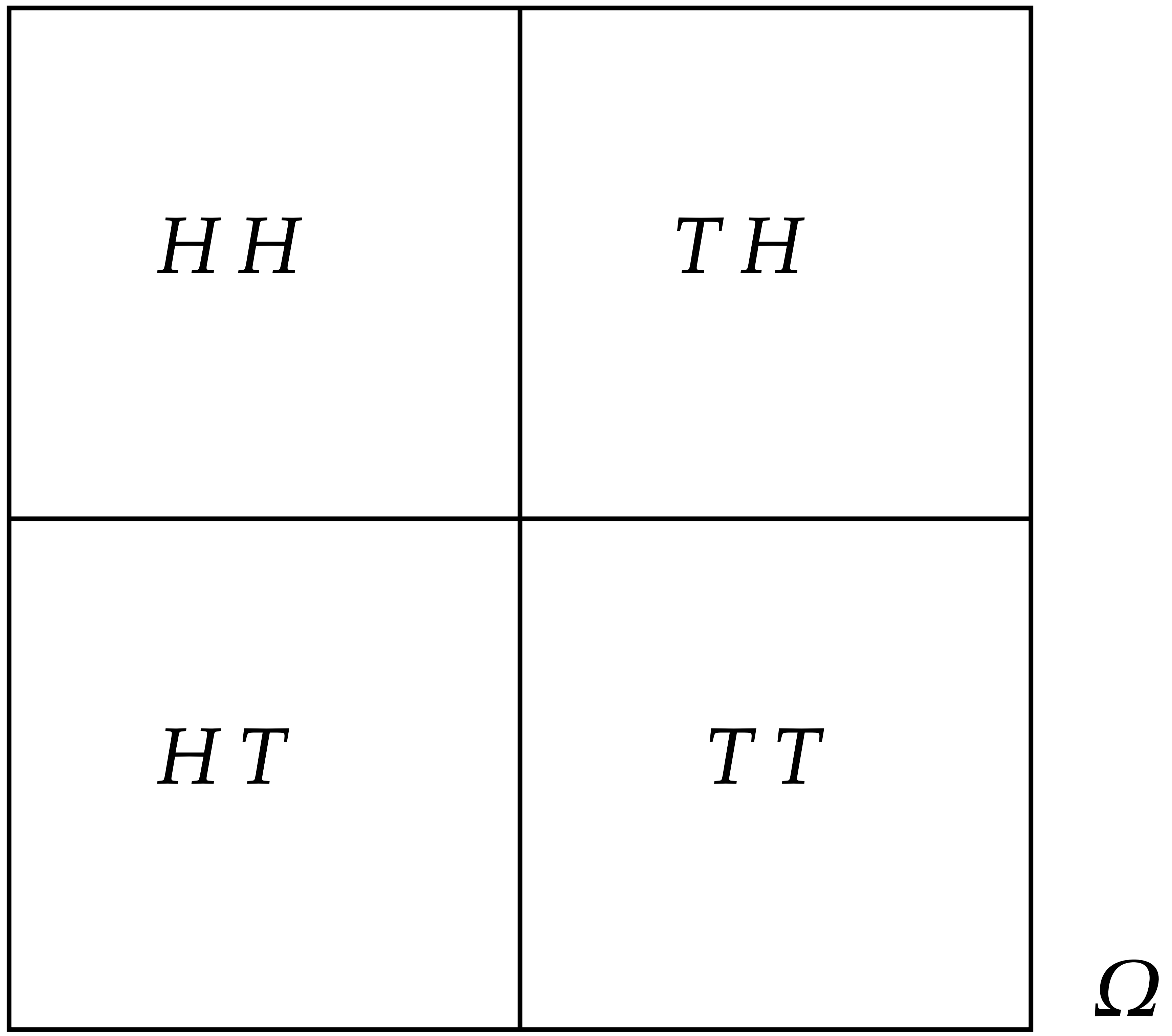
Example Measures

For discrete outcome spaces, a common way to measure probabilities is to make outcomes equally probable:

$$\mathbb{P}(\{\omega\}) = 1/|\Omega| \text{ for } \omega \in \Omega.$$

This isn't the only valid measure, e.g.

$$\mathbb{P}(\{HH\}) = 1$$



Conditional Probabilities

Intuition and definition

For events A, B, the ***conditional probability*** of B given A is:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

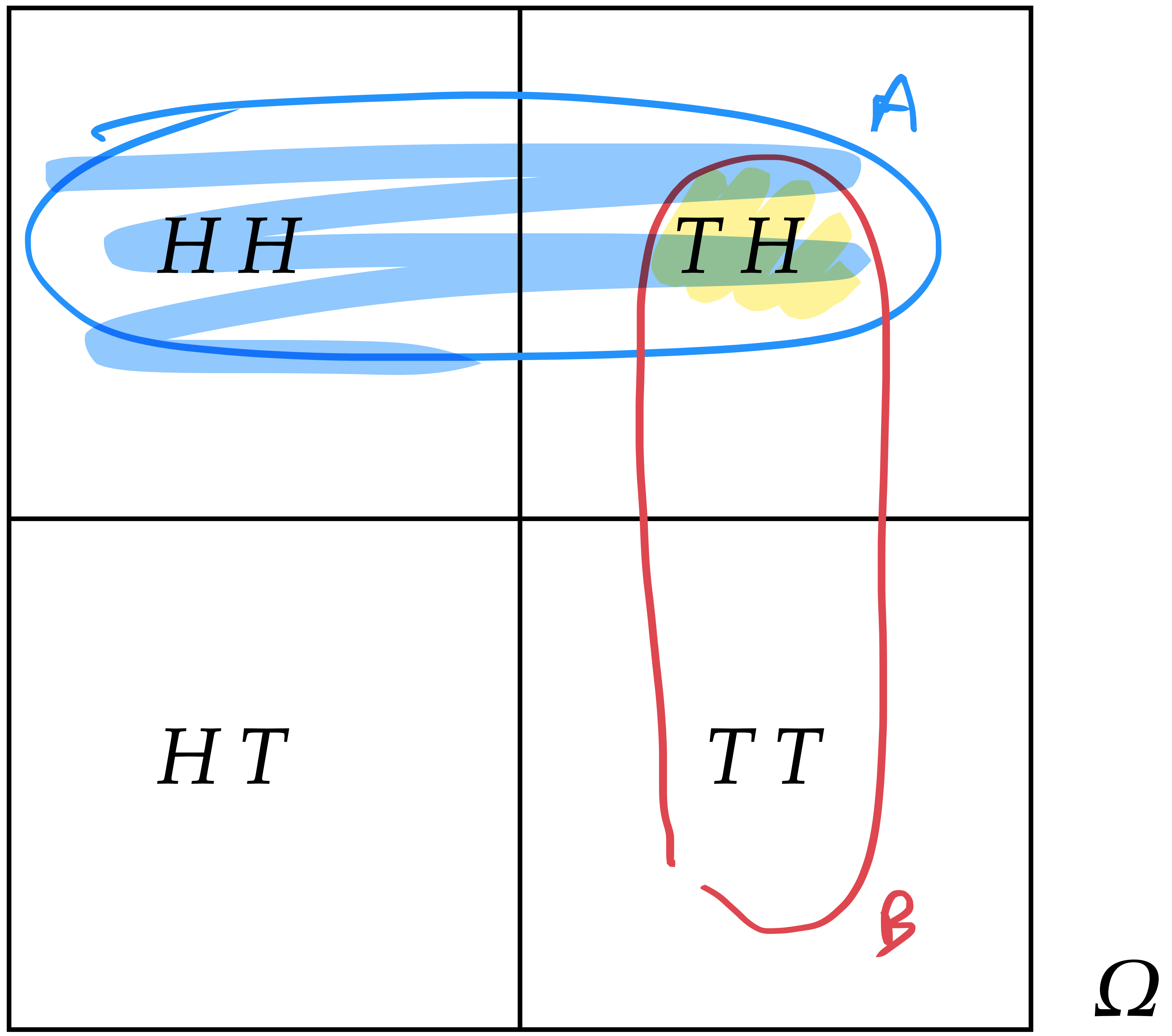
Example:

$A = \{\text{Bob's coin is } H\}$

$B = \{\text{Alice's coin is } T\}$

$C = \{\text{Alice's coin is } H\}$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/2} = 1/2$$



Conditional Probabilities

Chain Rule and Bayes' Rule

The ***chain rule*** of conditional probability is:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A).$$

This easily gives us ***Bayes' rule***:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Bayes' rule can be thought of as how we "update our beliefs."

Conditional Probabilities

Law of Total Probability

A = Alice flips Heads.
B_1 = Bob flips Tails
B_2 = Bob flips Heads

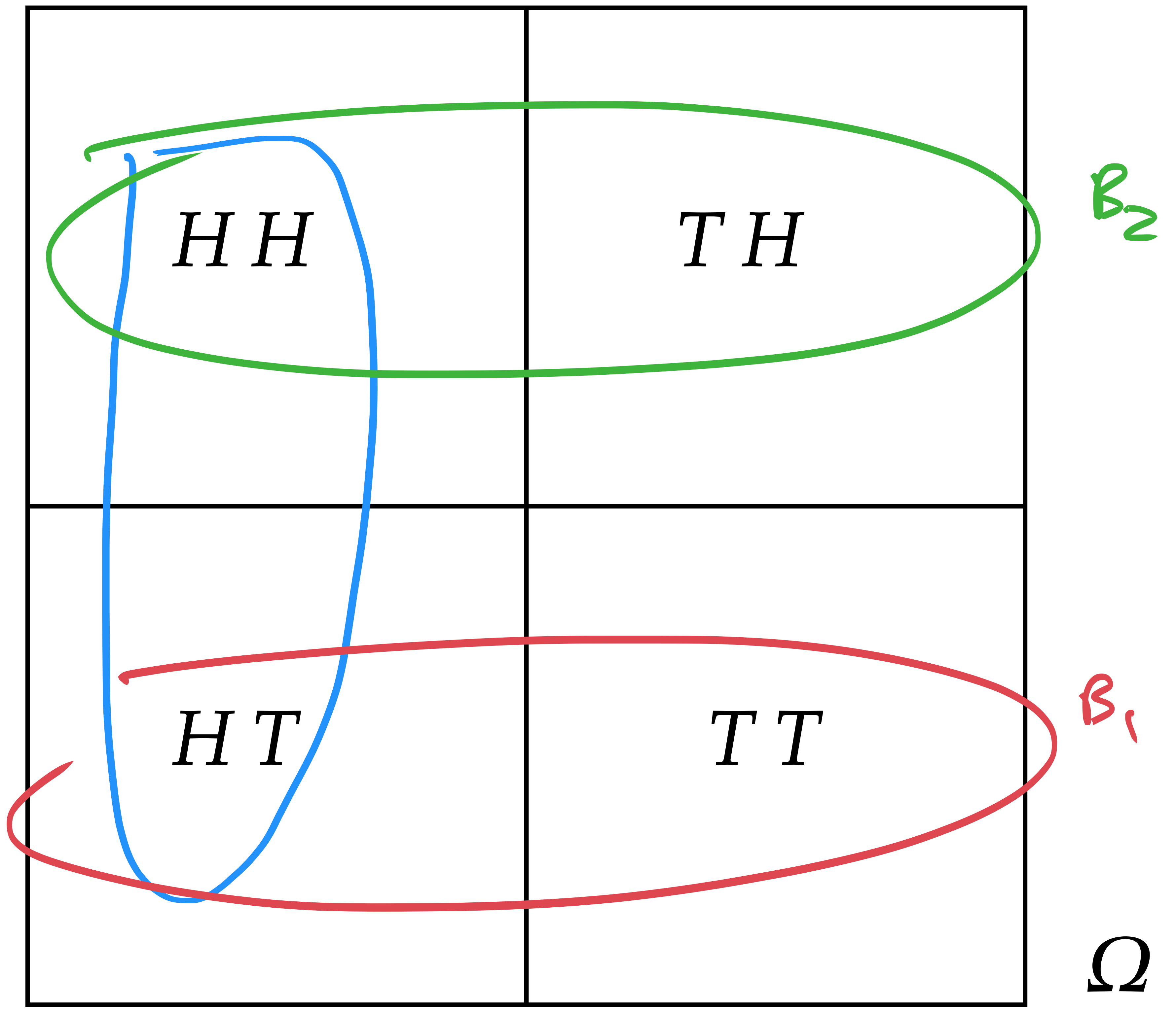
The ***law of total probability*** allows us to chop up probabilities into an exact sum of distinct events.

If $B_1, B_2, B_3, \ldots$ is a *countable* collection of events, then, for any event A:

Axiom

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$$

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$



Probability Space

Intuition and definition

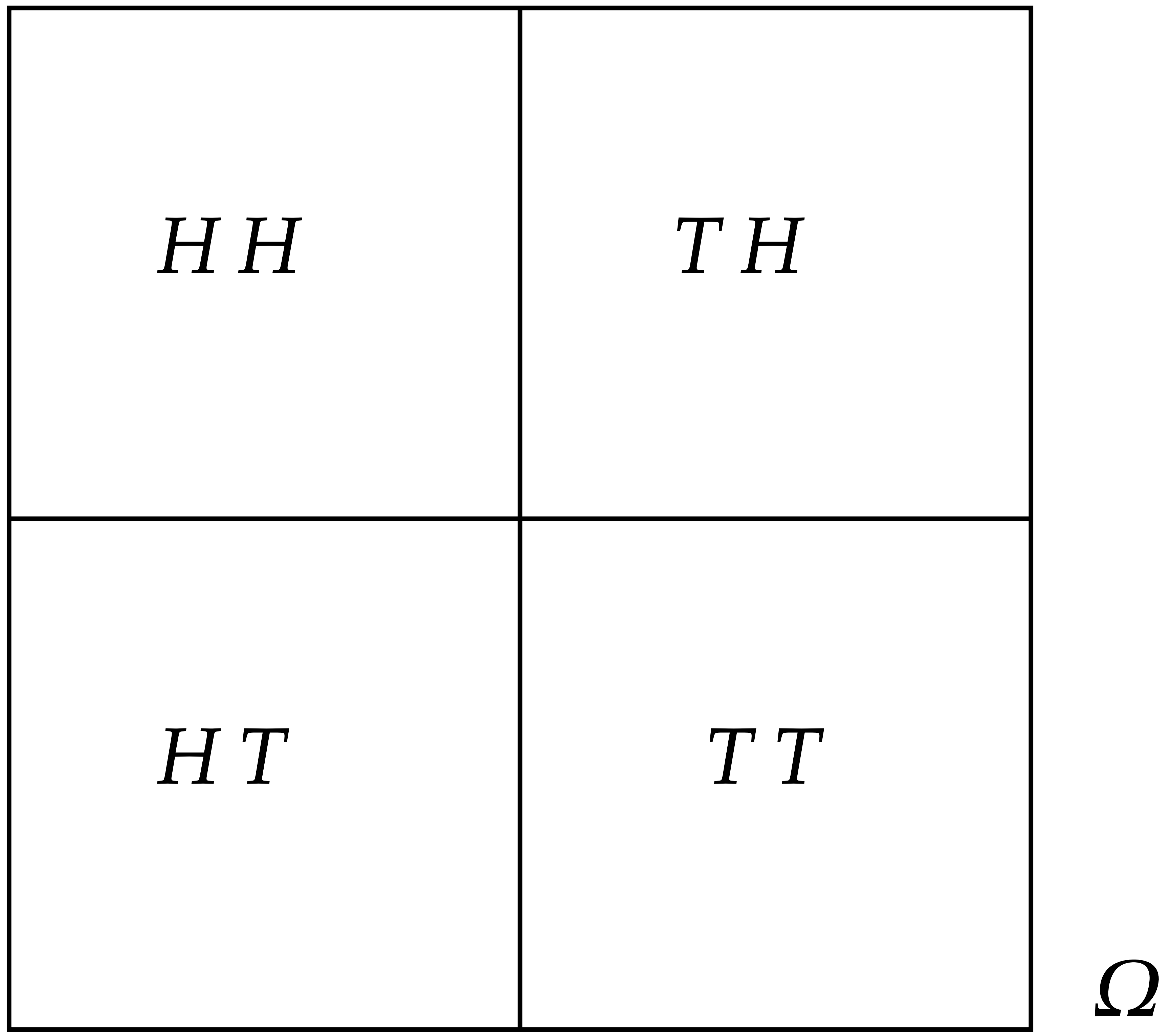
A tuple of a *sample space*, *event space* (σ-algebra), and *probability measure* $(\Omega, \mathscr{A}, \mathbb{P})$ is called a ***probability space***.

Example:

$\Omega = \{HH, HT, TH, TT\}$

$\mathscr{A} = \{\varnothing, \{HH\}, \{HT\}, \ldots, \{HH, HT, TH, TT\}\}$

$\mathbb{P}(\{\omega\}) = 1/4$ for all $\omega \in \Omega$.



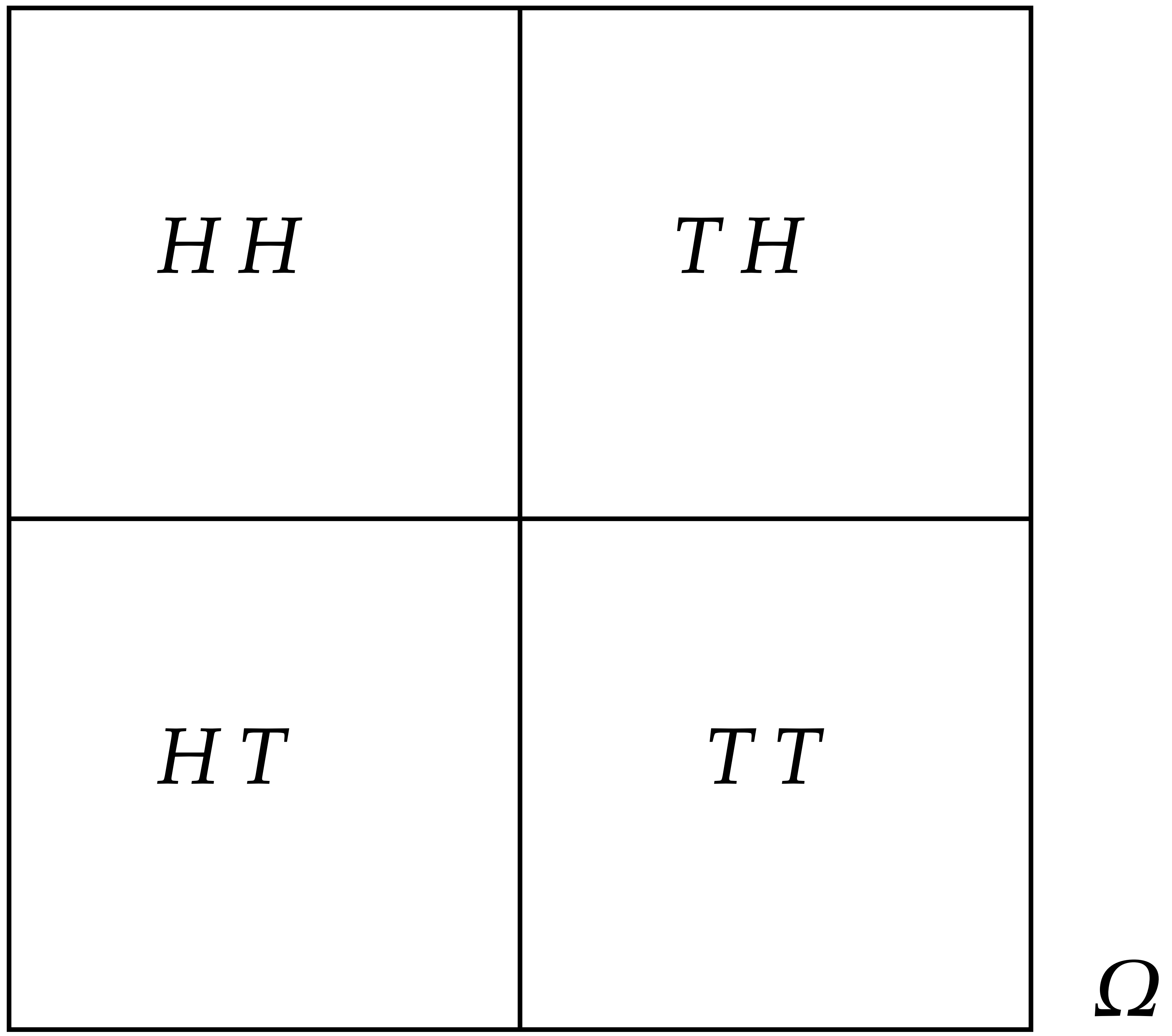
Probability Space

Intuition and definition

A tuple of a *sample space*, *event space* (σ-algebra), and *probability measure* $(\Omega, \mathscr{F}, \mathbb{P})$ is called a **probability space**.

This is "underlying"

We avoid dealing with these directly!
Instead, we use ***random variables.***



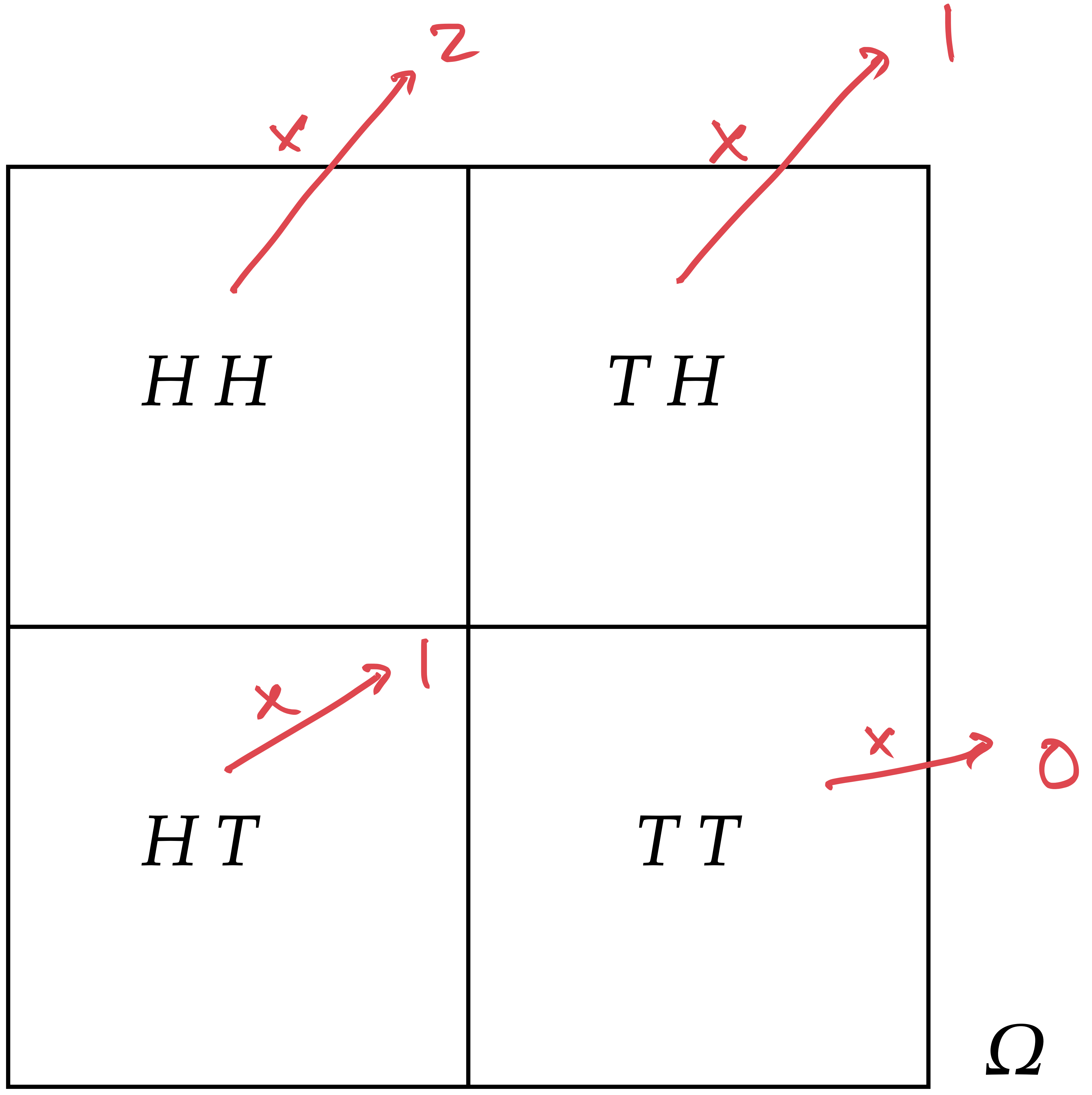
Random Variables

Example: Flipping 2 fair coins

Consider the following function:

$$X : \Omega \to \mathbb{R}$$

where $X(\omega)$ = number of heads, H.



[***Random variables***]() are *functions* that assign a numerical quantity to every outcome in the sample space.

Random Variables

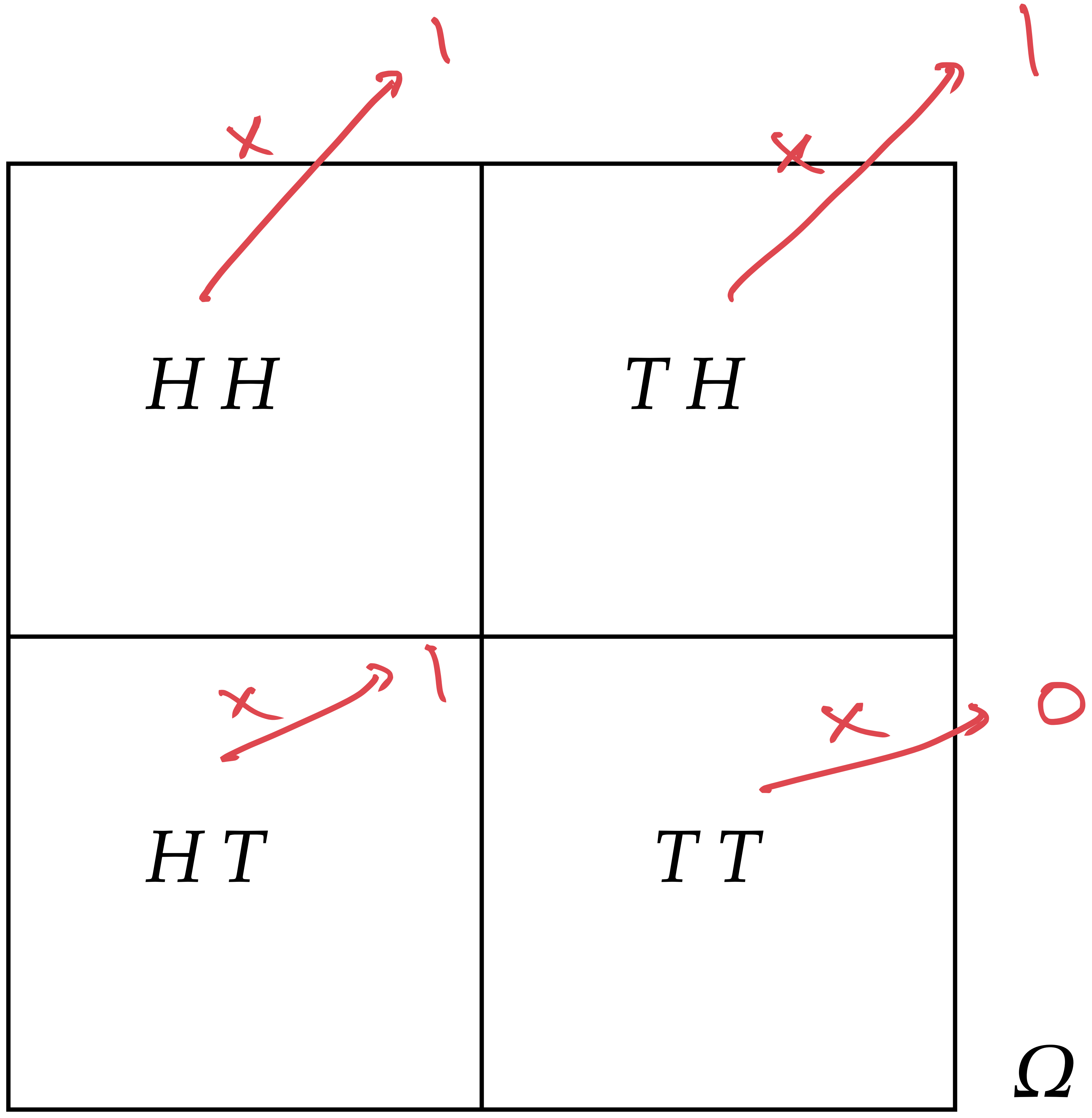
Example: Flipping 2 fair coins

Consider the following function:

$$X : \Omega \to \mathbb{R}$$

where $X(\omega) = 1$ if at least one H, and 0 otherwise.

Random variables are *functions* that assign a numerical quantity to every outcome in the sample space.



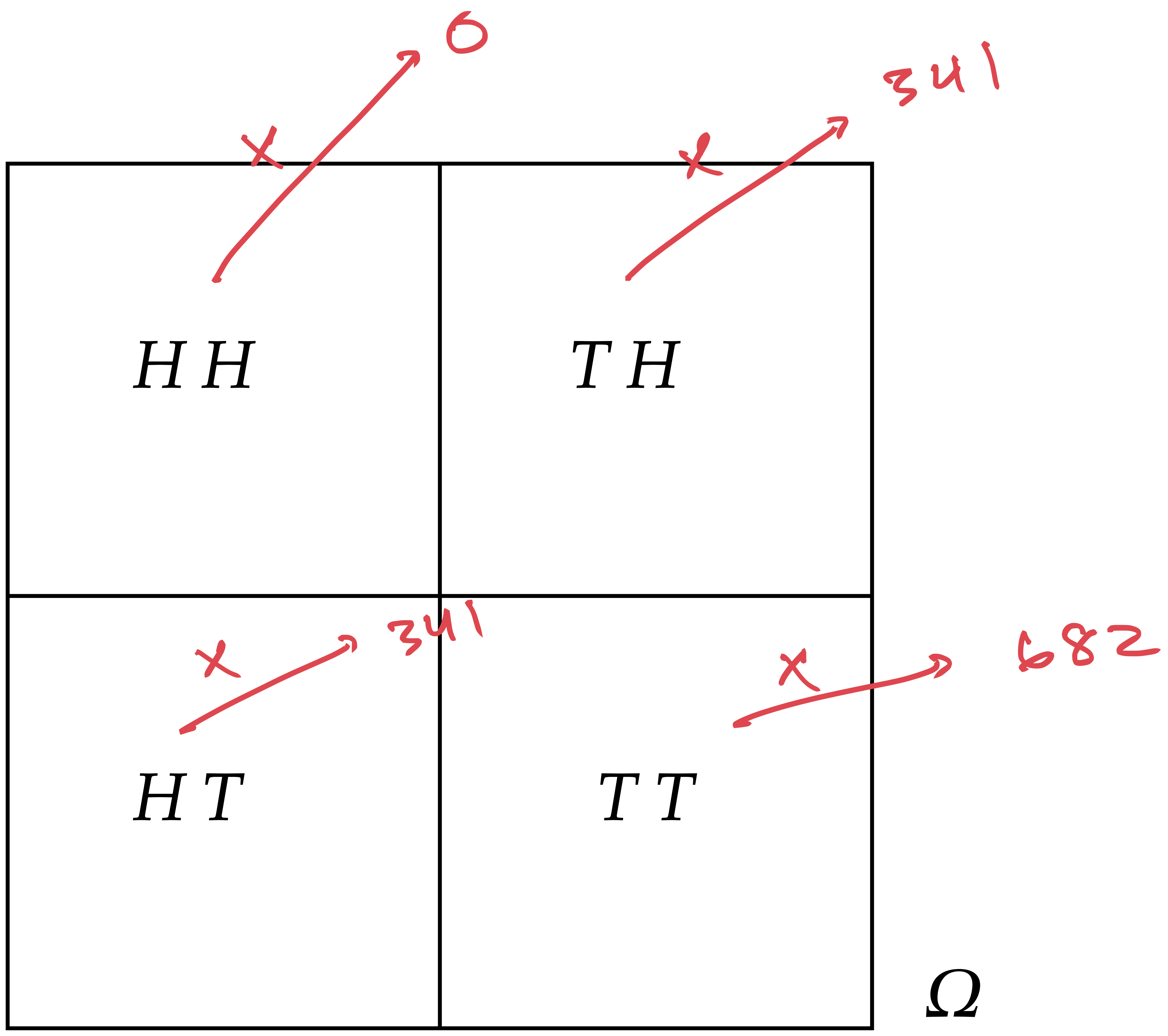
Random Variables

Example: Flipping 2 fair coins

Consider the following function:

$$X : \Omega \to \mathbb{R}$$

where $X(\omega) = 341x$ where x is the number of T.

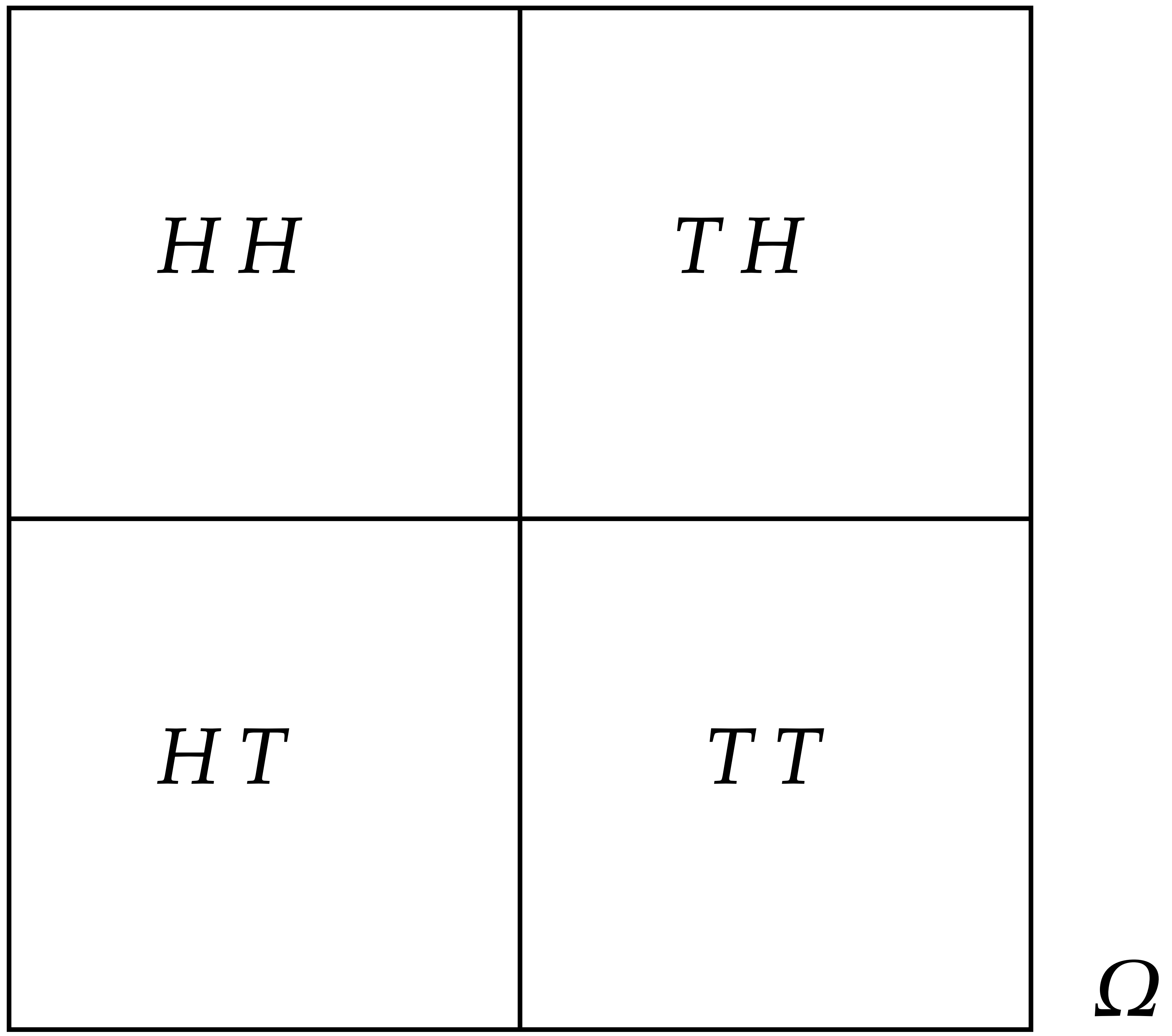


[Random variables] are *functions* that assign a numerical quantity to every outcome in the sample space.

Random Variable

Intuition and definition

A ***random variable*** is a function $X : \Omega \to \mathbb{R}$ that takes outcomes $\omega \in \Omega$ of the sample space and maps them to real values.

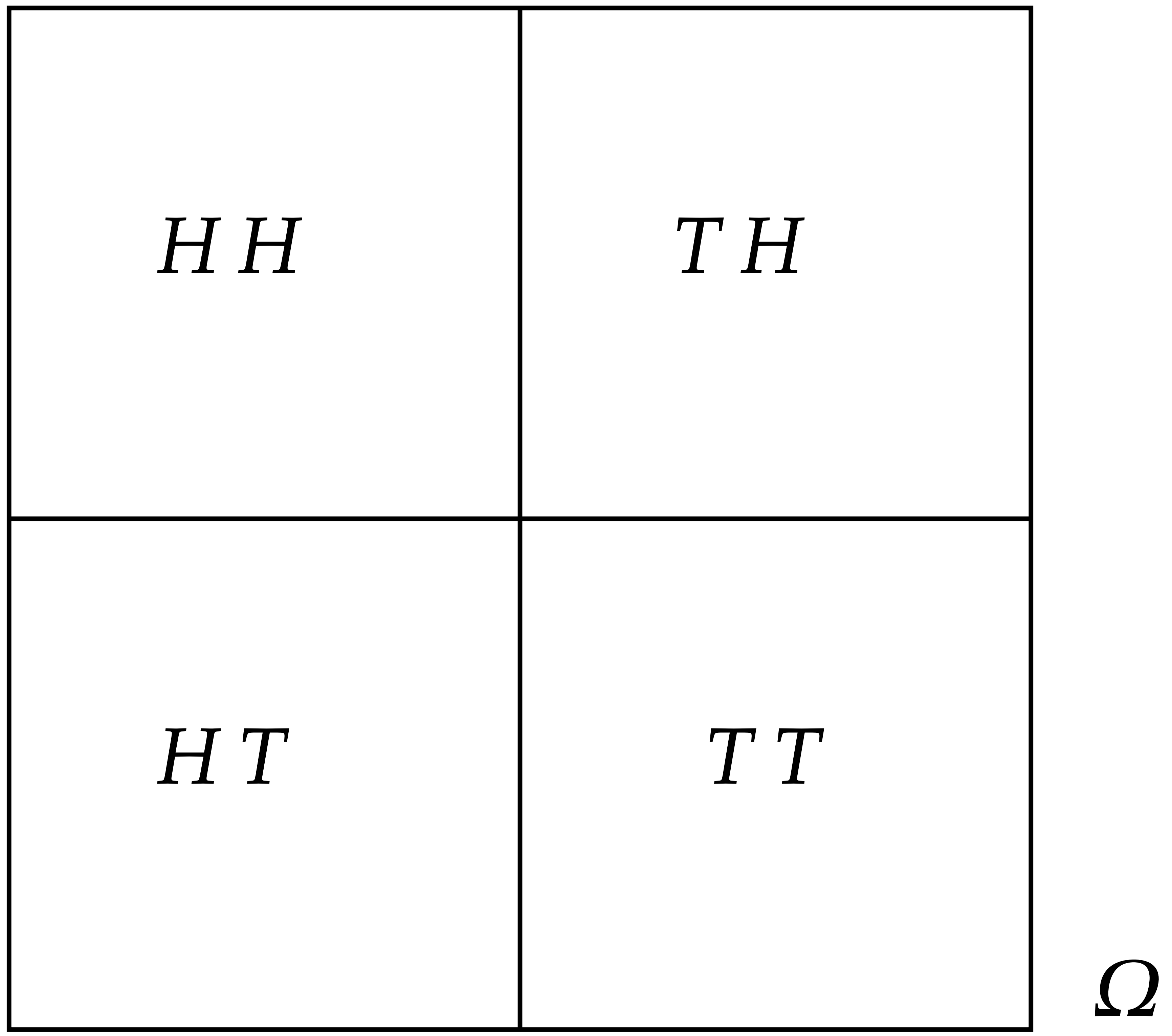


Random Variable

Intuition and definition

A ***random variable*** is a function $X : \Omega \to \mathbb{R}$ that takes outcomes $\omega \in \Omega$ of the sample space and maps them to real values.

We typically use random variables to talk about events without referencing the underlying sample space.



Random Variable

Intuition and definition

Let $X : \Omega \to \mathbb{R}$ be defined as

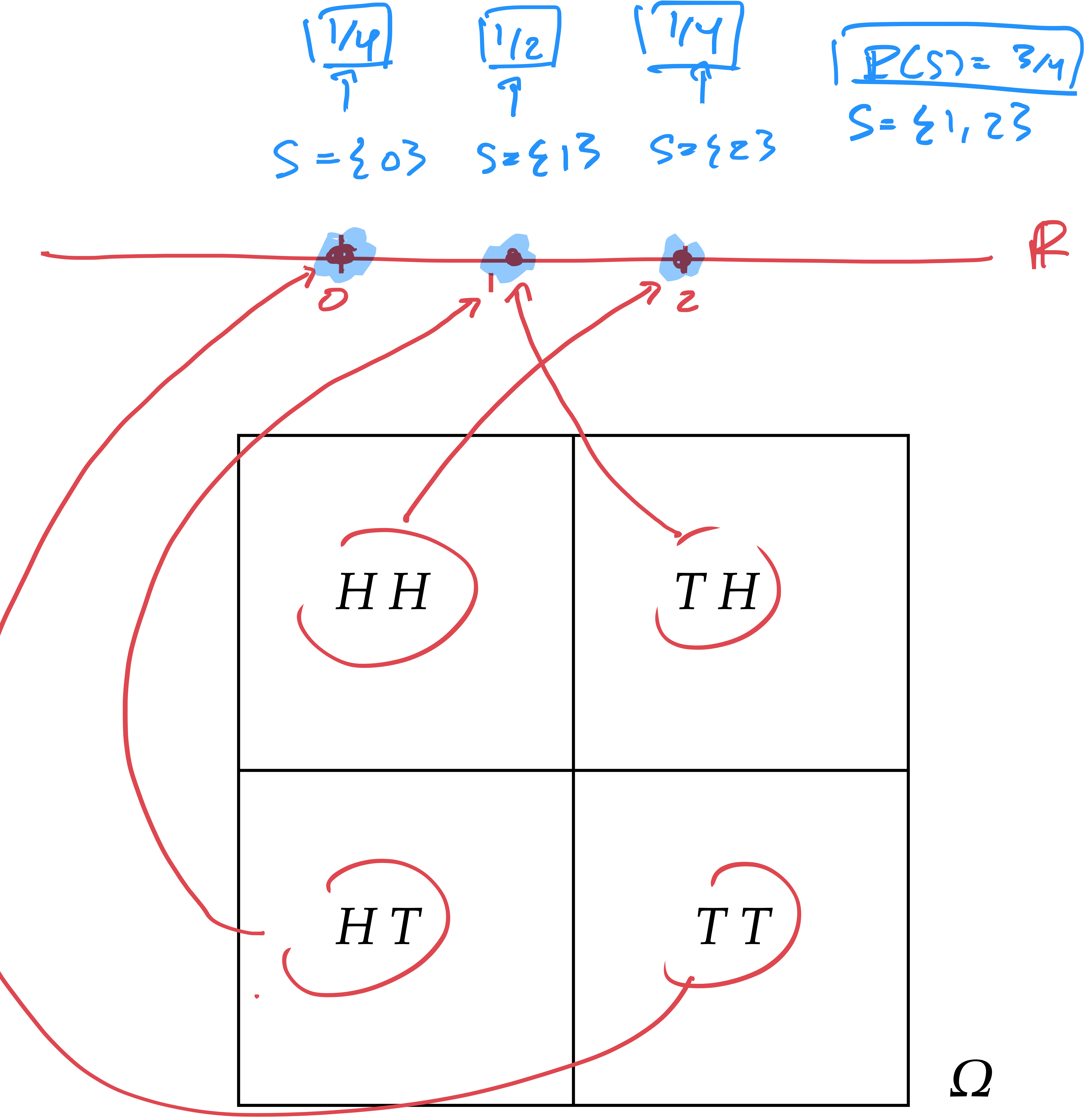
$$X(\omega) = \text{\# of heads, } H.$$

Let the underlying probability measure assign outcomes to be equally likely:

$$\mathbb{P}(\{\omega\}) = 1/4$$

Then, for any $S \subseteq \mathbb{R}$,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S).$$



$X^{-1}(\{20\}) = \emptyset$.
$X^{-1}(\{1\}) = \{TH, HT\}$

Random Variable

Intuition and definition

Let $X : \Omega \to \mathbb{R}$ be defined as

$$X(\omega) = \text{\# of heads, } H.$$

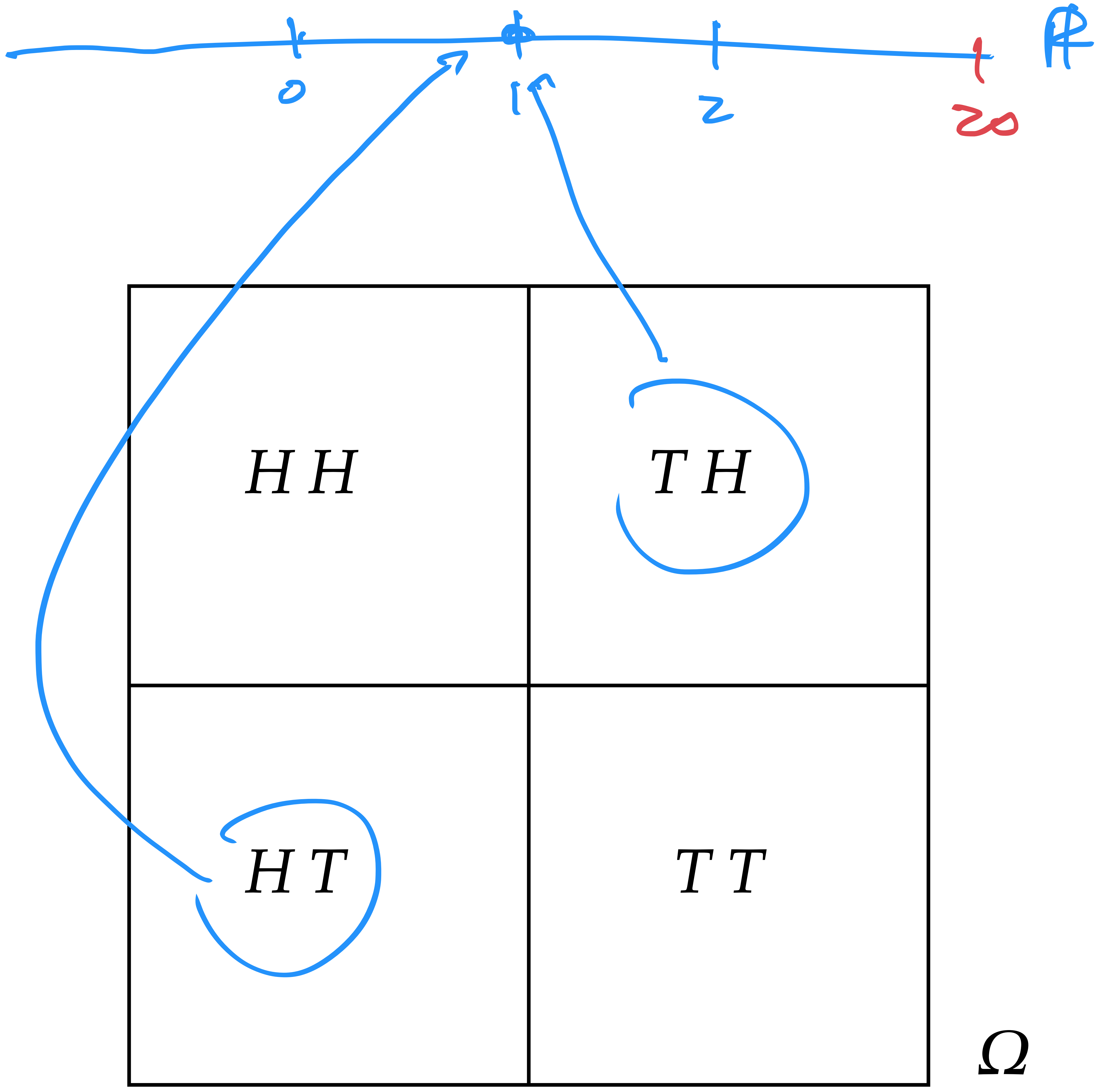
For any $S \subseteq \mathbb{R}$,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in S\})$$

Example.

What's $\mathbb{P}_X(1)$? $= 1/2$

What's $\mathbb{P}_X(20)$? $= 0$



Random Variable

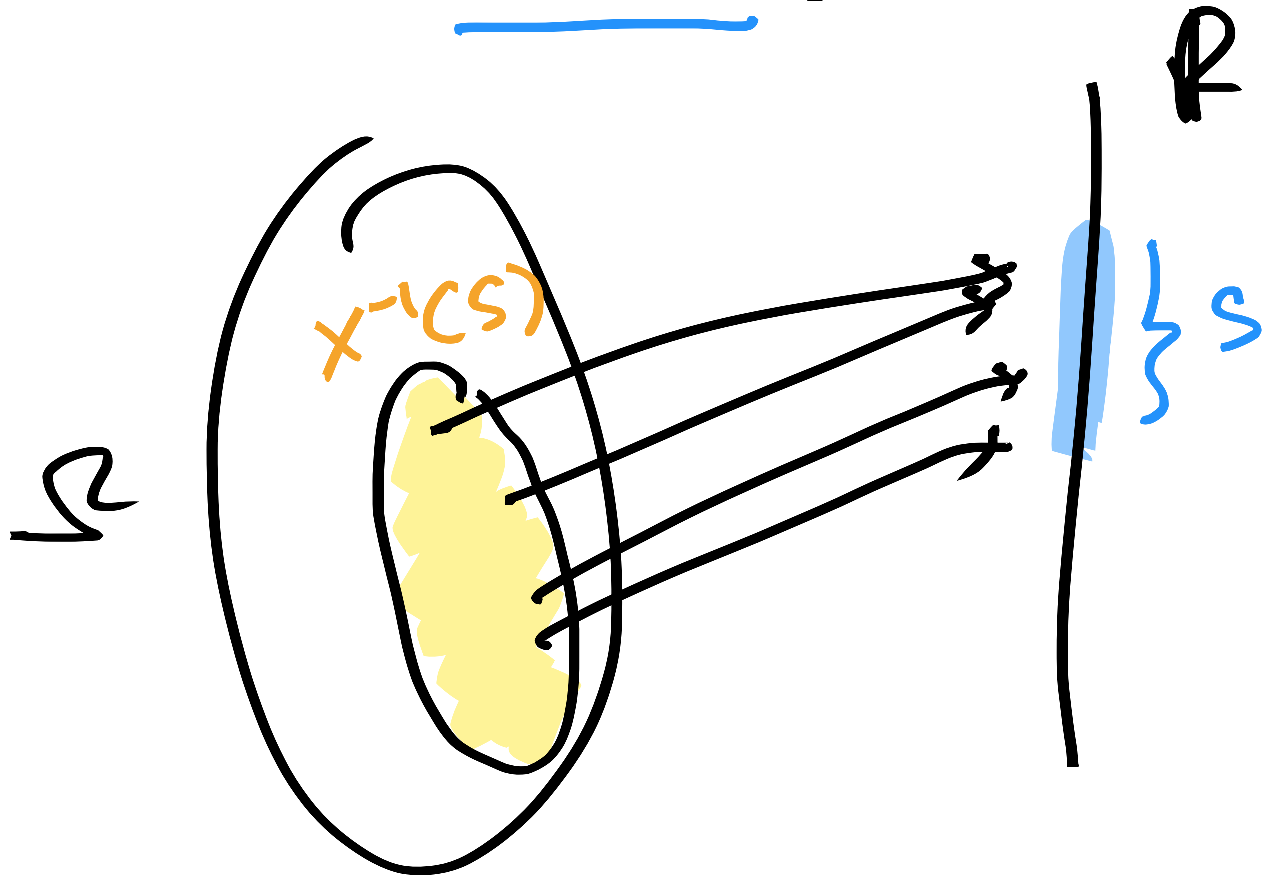
The distribution of a random variable

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be some underlying probability space.

Random variables $X : \Omega \to \mathbb{R}$ come with a ***distribution/law***, $\mathbb{P}_X$.

This implicitly defines a probability measure on $\mathbb{R}$. For $S \subseteq \mathbb{R}$,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in S\}).$$



Random Variable

The distribution of a random variable

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be some underlying probability space.

Random variables $X : \Omega \to \mathbb{R}$ come with a ***distribution/law***, $\mathbb{P}_X$.

This implicitly defines a probability measure on $\mathbb{R}$. For $S \subseteq \mathbb{R}$,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in S\}).$$

This allows us to just talk about the numbers in $\mathbb{R}$!

Probability Spaces

Putting everything together

The ***sample space*** is the set of all possible outcomes:

$$\Omega = \{HH, TH, HT, TT\}.$$

The ***event space*** (σ-algebra) is some collection of events:

$$\mathscr{A} = \{\varnothing, \{HH\}, \{TT\}, \ldots, \{HH, HT, TH, TT\}\}$$

The (underlying/base) ***probability measure*** is how we measure the "mass" of events:

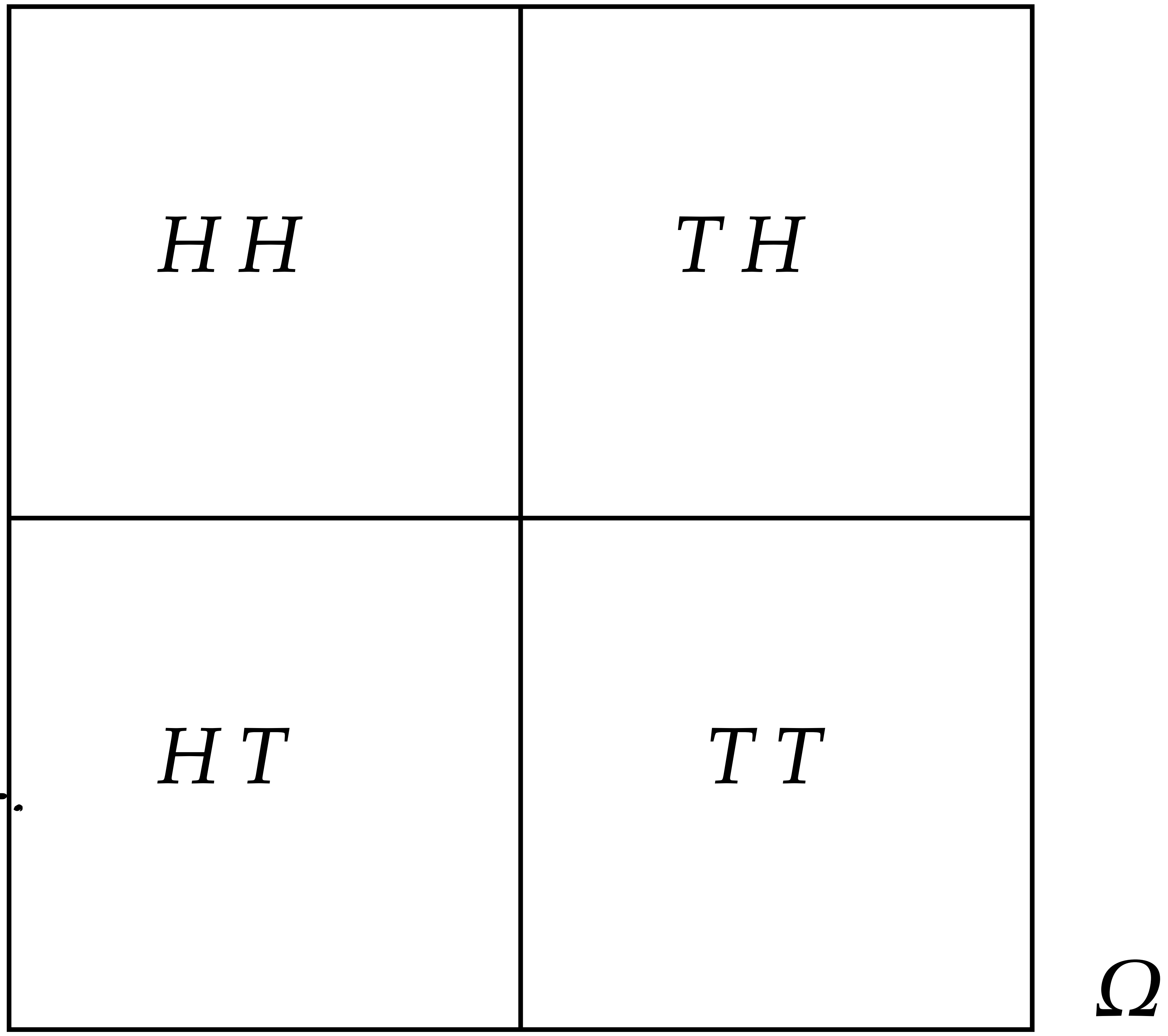
$$\mathbb{P}(\omega) = 1/4 \text{ for } \omega \in \Omega.$$

Discrete: Equally probable

Continuous: "Length" of a subset.

A ***random variable*** on $(\Omega, \mathscr{A}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ associating outcomes $\omega \in \Omega$ to numerical values in $\mathbb{R}$:

$$X(\omega) = \text{\# of heads in } \omega$$



Probability Spaces

Putting everything together

$X = \#$ of Heads

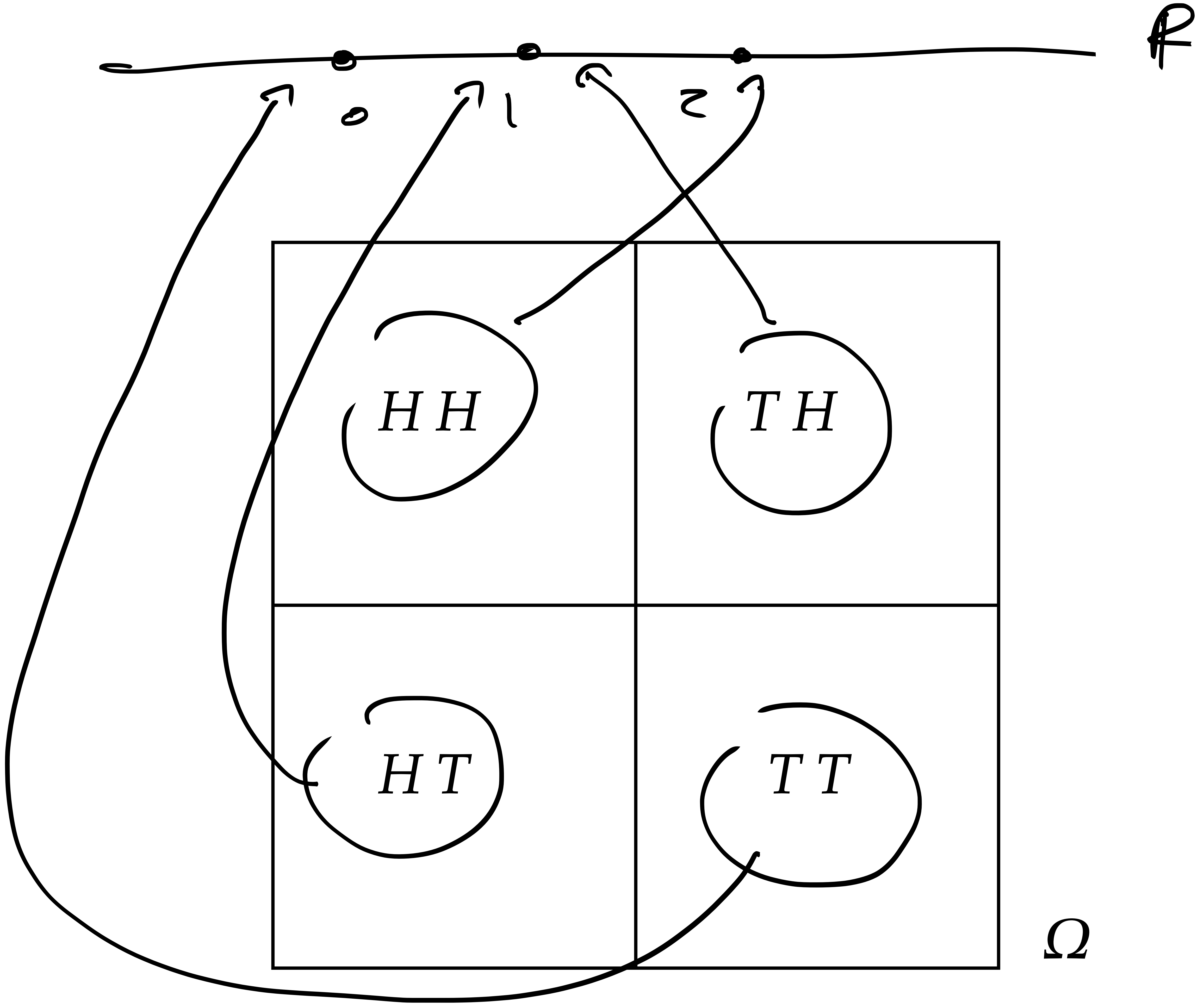
Example:

Compute $\mathbb{P}(X = 0)$: $1/4$

$\mathbb{P}(X \in \{0, 1\}) = 3/4$

Compute $\mathbb{P}(X = 1)$: $1/2$

Compute $\mathbb{P}(X = 2)$: $1/4$



Random Variables

Distributions of random variables

Cumulative Distribution Function

Intuition and definition

IMPLICIT

Let $X : \Omega \to \mathbb{R}$ be some random variable (*on an underlying probability space* $(\Omega, \mathscr{A}, \mathbb{P})$*)*.

The ***cumulative distribution function (CDF)*** of X is the function $F_X : \mathbb{R} \to [0,1]$ defined as:

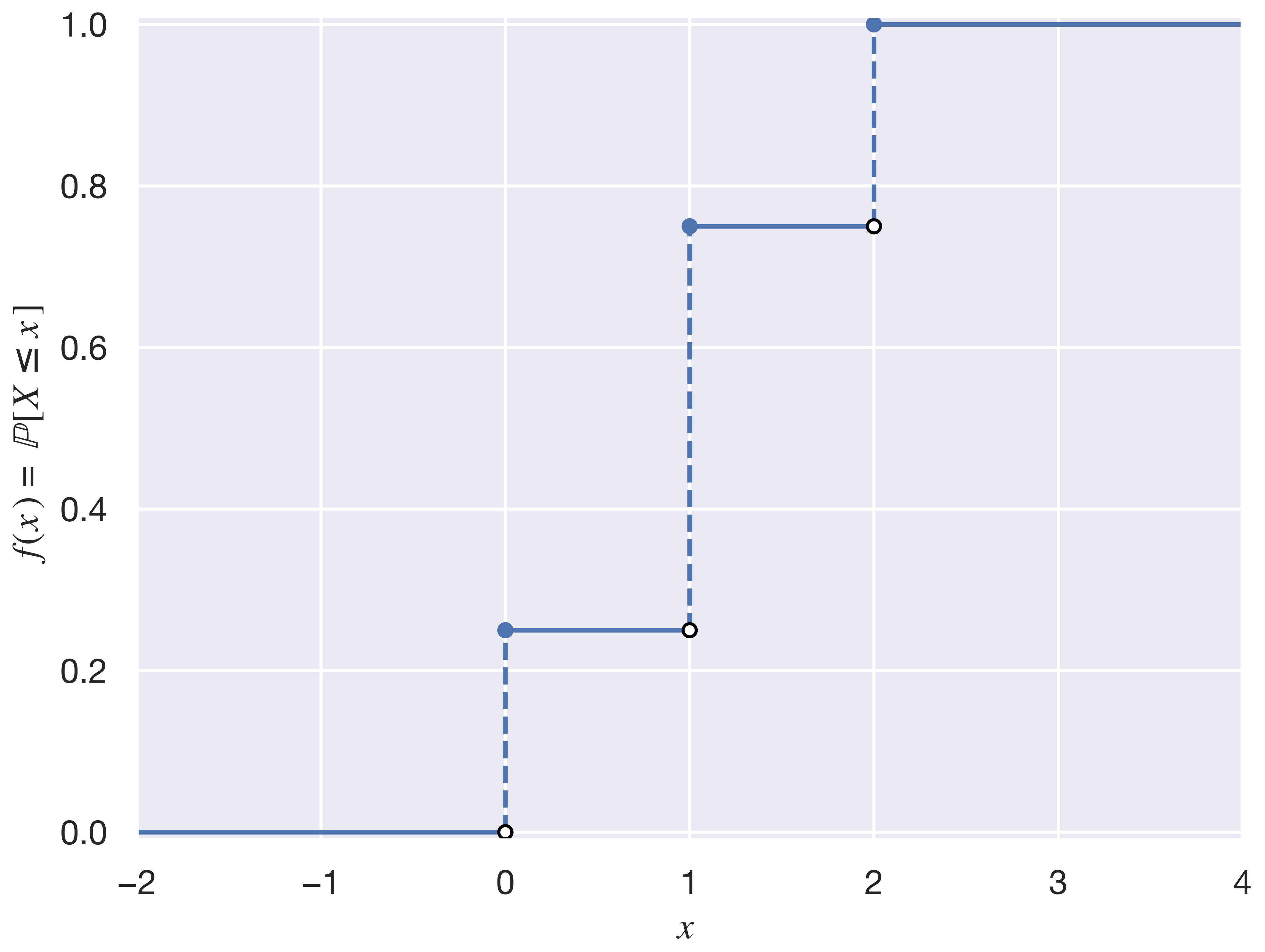
$$F_X(x) = \mathbb{P}(X \leq x)$$

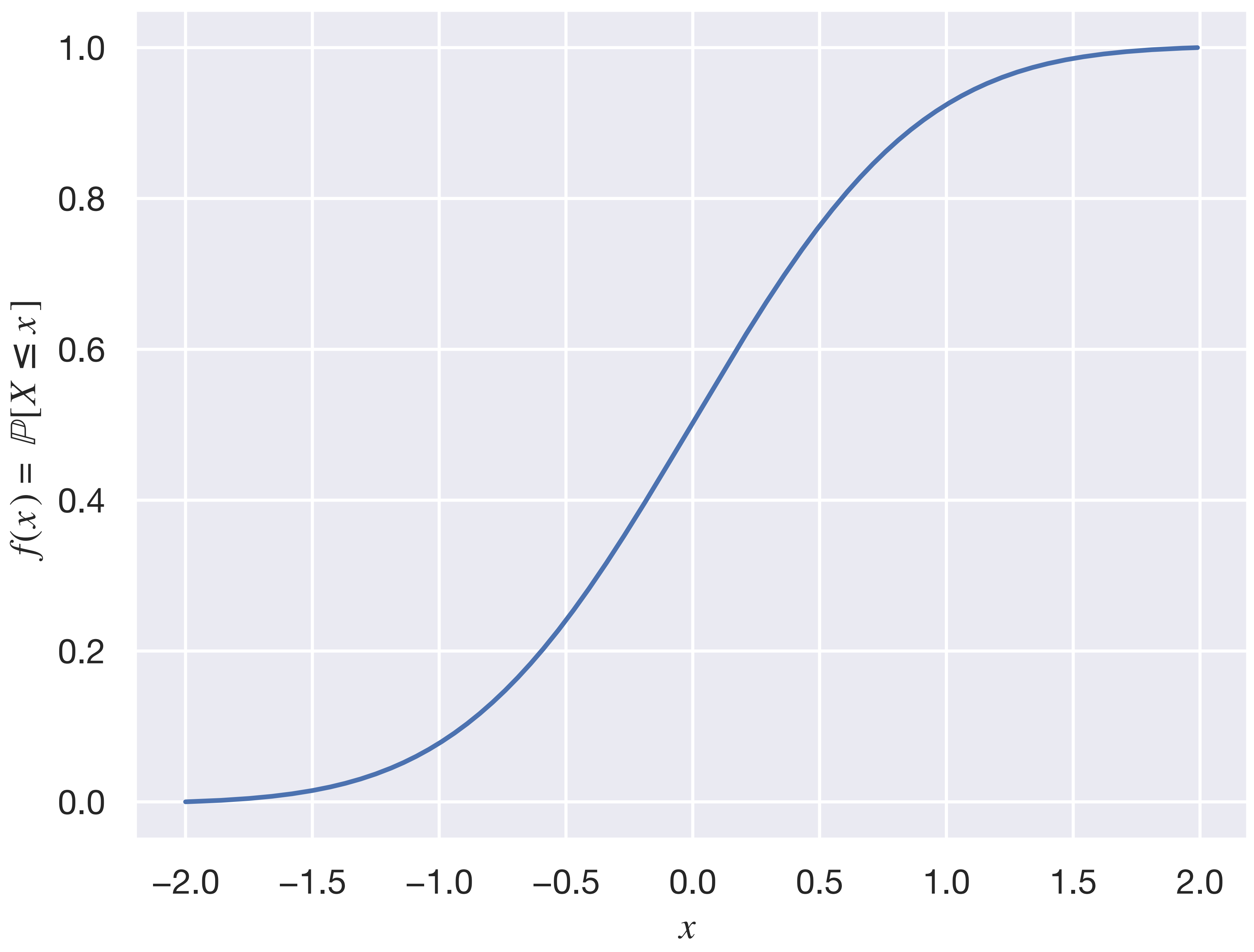
This function allows us to get probabilities in an interval:

$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$$

Cumulative Distribution Function

Examples





Cumulative Distribution Function

Properties

$F(x) = \mathbb{P}[X \leq x]$

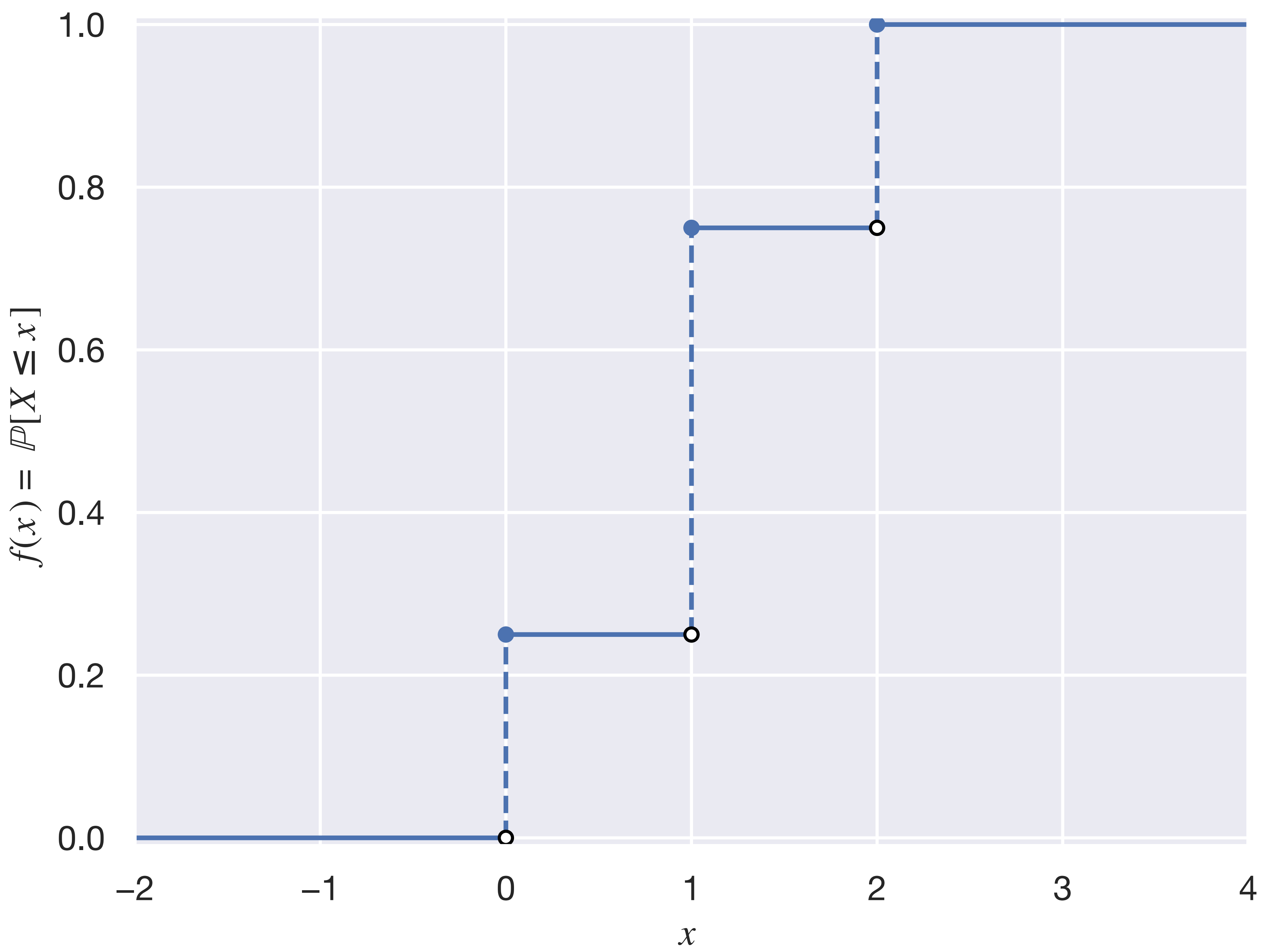
Right-continuous. Every for every point $a \in \mathbb{R}$, the CDF satisfies:

$$\lim_{x \to a+} f(x) = f(a).$$

Monotonically nondecreasing. For every $x \leq y$, $F_X(x) \leq F_X(y)$.

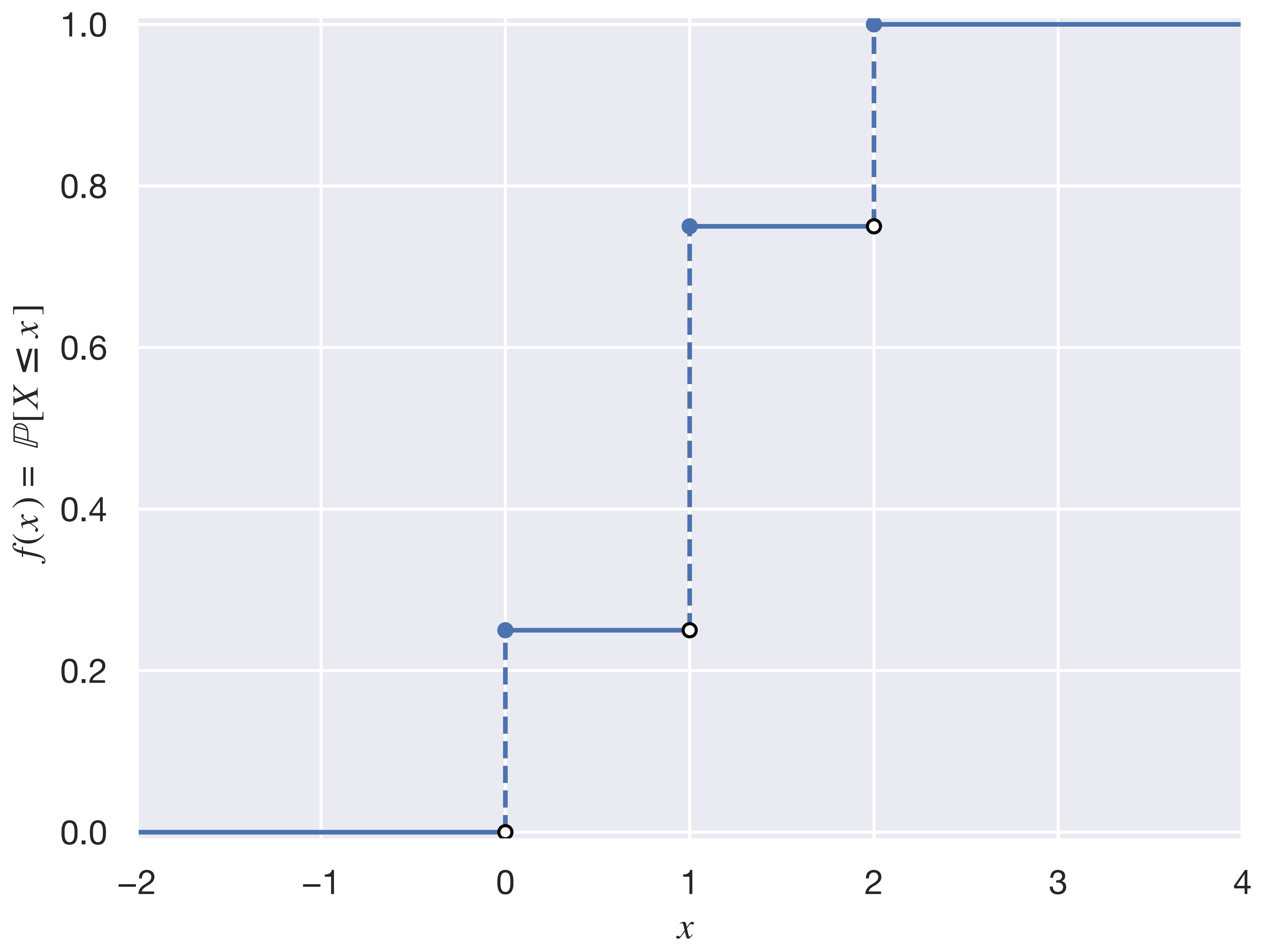
Limits at infinities. The limits at both infinities are:

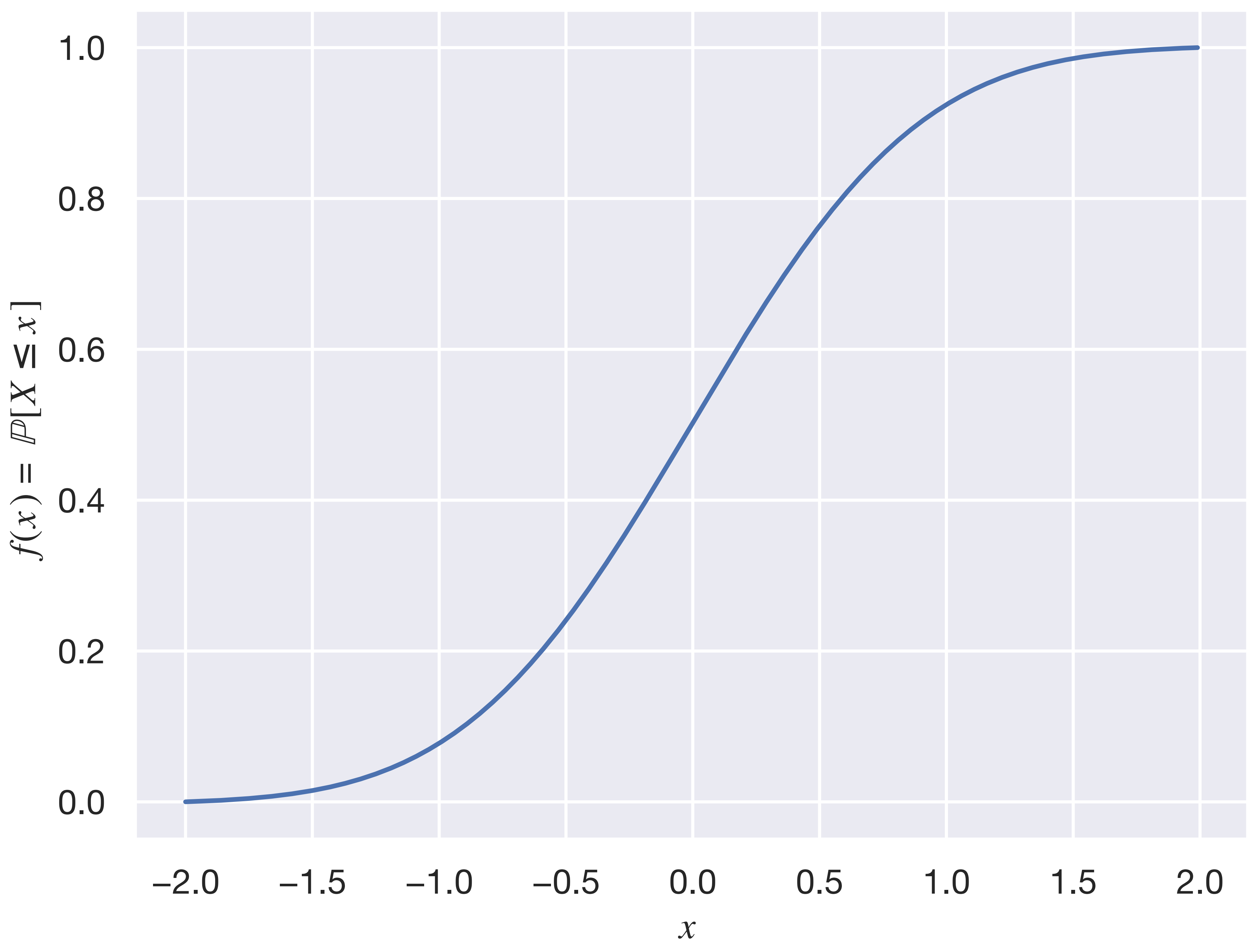
$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to \infty} F_X(x) = 1.$$



Discrete vs. Continuous RVs

Difference in CDF





Discrete RVs have “jumps” in the CDF; (absolutely) continuous RVs are smooth.

Discrete Random Variables

Intuition and definition

{1, 2, 3, 4, 5, ...}

A ***discrete random variable*** is a random variable whose range

$$X(\Omega) = \{x \in \mathbb{R} : X(\omega) = x \text{ for some } \omega \in \Omega\}$$

is *countable* or *finite*.

Example.

$X : \{HH, HT, TH, TT\} \to \mathbb{R}$ with $X(\omega)$ counting the number of heads.

$X : [0,1] \to \mathbb{R}$ defined by $X(\omega) = 0$ if $\omega < 0.5$ and $X(\omega) = 1$ otherwise.

Discrete Random Variables

Probability mass function

$\mathbb{P}[X=0] = 1/4$

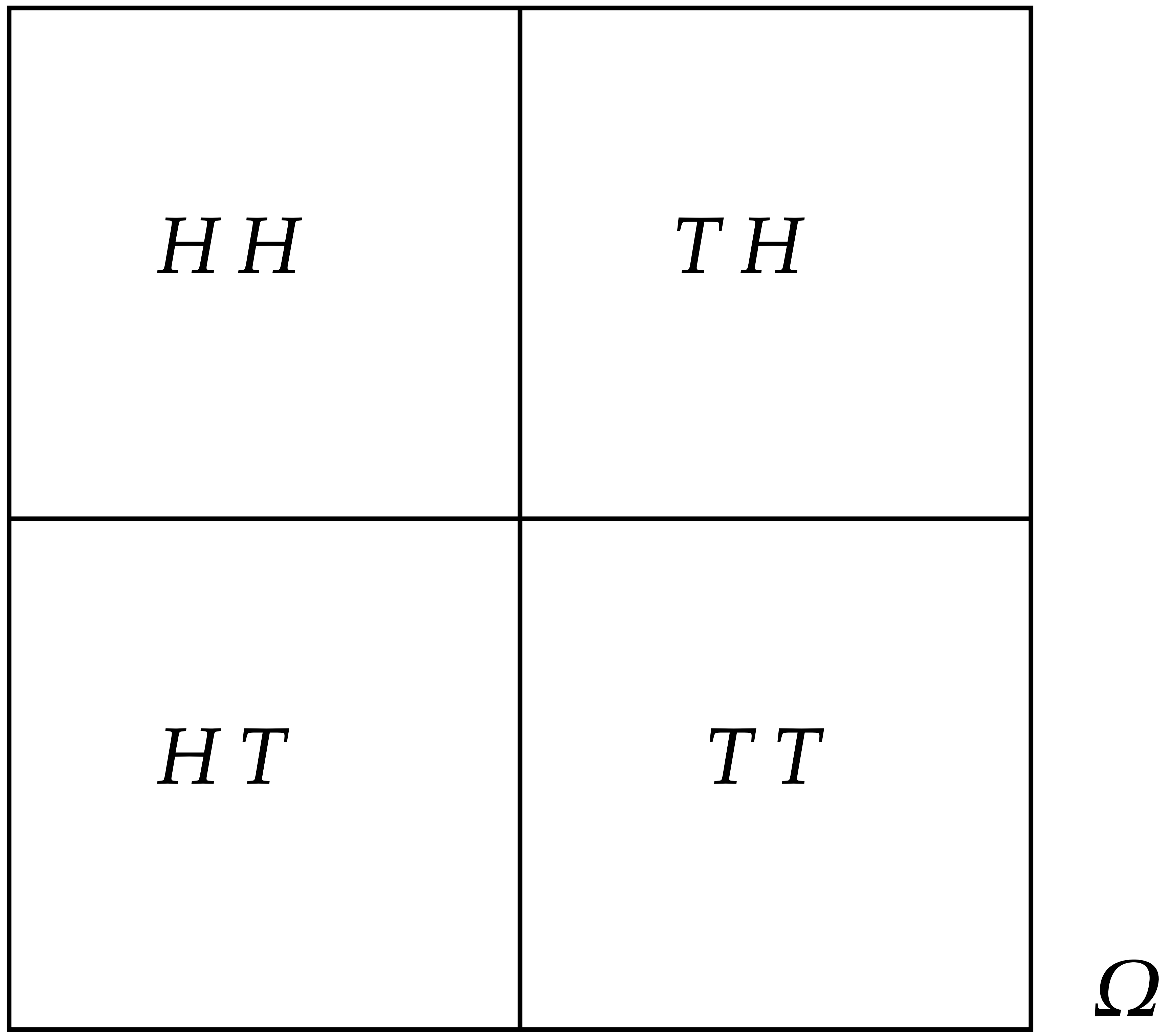
$\mathbb{P}[X=1] = 1/2$

$\mathbb{P}[X=2] = 1/4$

A discrete random variable X has a ***probability mass function (PMF)*** $p_X : \mathbb{R} \to [0,1]$ defined by:

$$p_X(x) = \mathbb{P}[X = x].$$

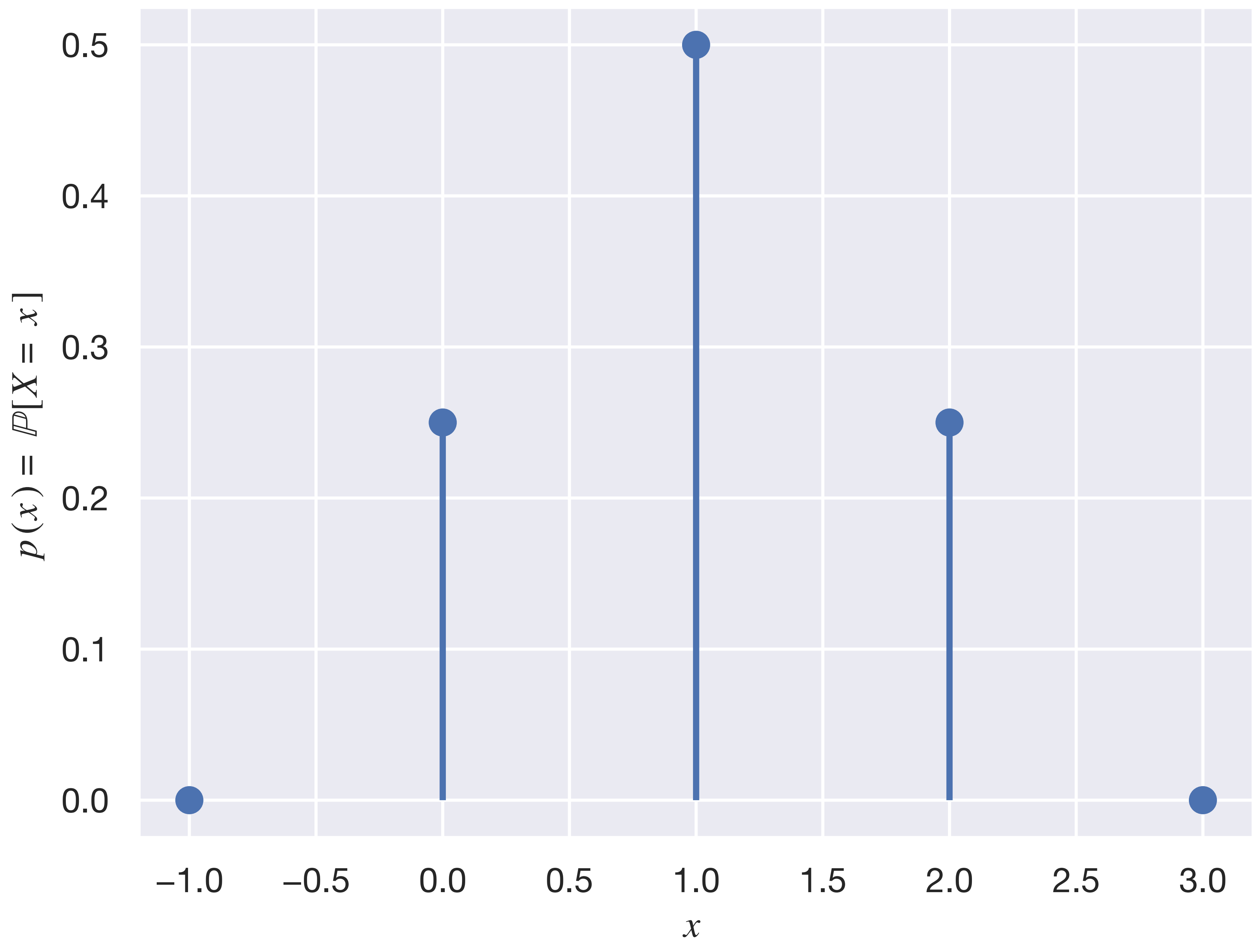
Example. What's the PMF of the RV $X : \Omega \to \mathbb{R}$ with $X(\omega)$ counting the number of heads?

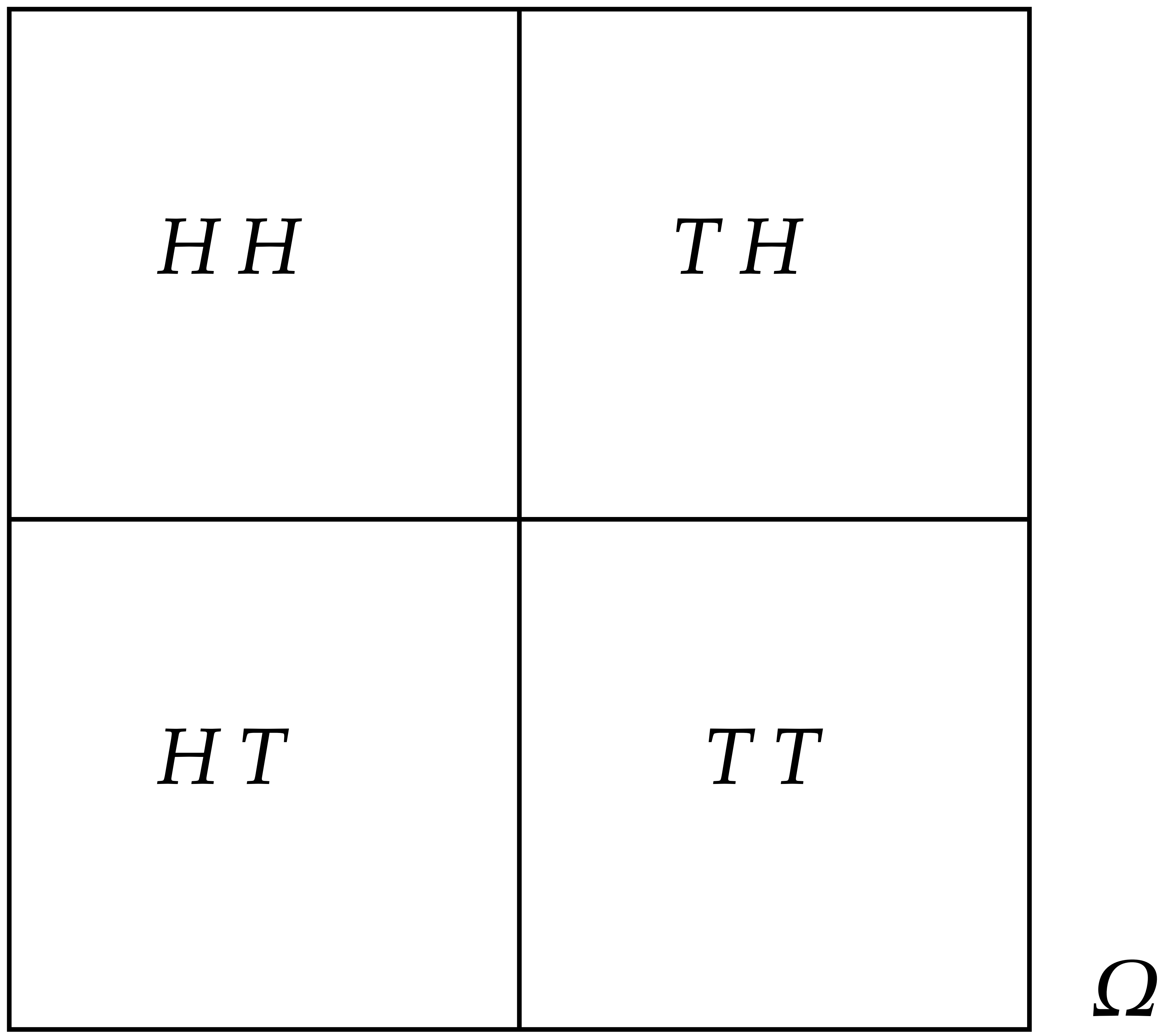


Discrete Random Variables

Example: Flipping 2 fair coins

Example. What's the PMF of the RV $X : \Omega \to \mathbb{R}$ with $X(\omega)$ counting the number of heads?





Continuous Random Variables

Intuition and definition

A ***continuous random variable*** is a random variable whose range

$$X(\Omega) = \{x \in \mathbb{R} : X(\omega) = x \text{ for some } \omega \in \Omega\}$$

is *uncountably infinite*.

For continuous random variables, the probability at any point $x \in \mathbb{R}$ is zero!

$$\mathbb{P}[X = x] = 0.$$

So there is no "probability mass function," but there is a ***probability density function.***

Continuous Random Variables

Probability density functions

A continuous random variable X has a ***probability density function (PDF)*** $p_X : \mathbb{R} \to \mathbb{R}$ (notice the output space need not be $[0,1]$) with the properties:

$$\text{For all } x \in \mathbb{R}, p_X(x) \geq 0 \text{ and } \int_{\mathbb{R}} p_X(z)dz = 1.$$

To get probabilities from the PDF:

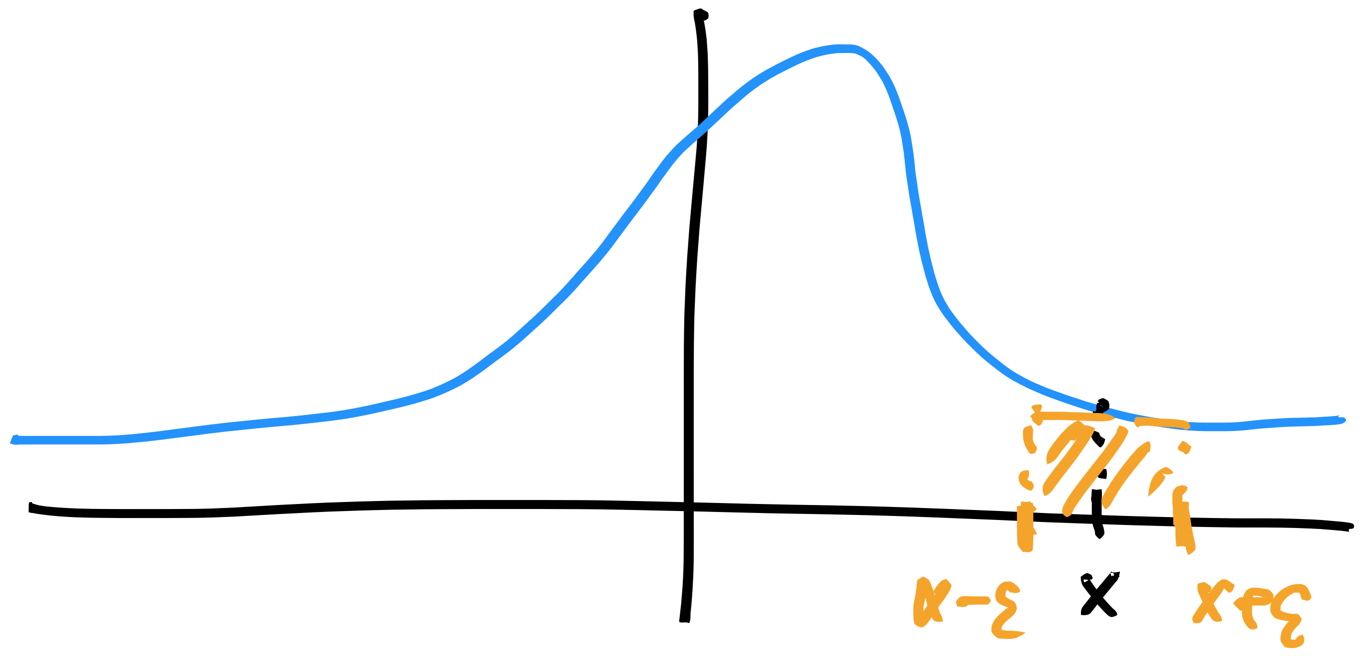
$$\mathbb{P}(a \leq X \leq b) = \int_a^b p_X(z)dz.$$

We can also obtain the CDF by the fundamental theorem of calculus:

$$p_X(x) = F'(x).$$

Continuous Random Variables

Intuition for the PDF



PDFs do NOT give probabilities.

Think of them in analogy to the physical notion of *density*:

$$\text{density} = \frac{\text{mass}}{\text{volume}}.$$

In an infinitesimally small interval, we can get a probability:

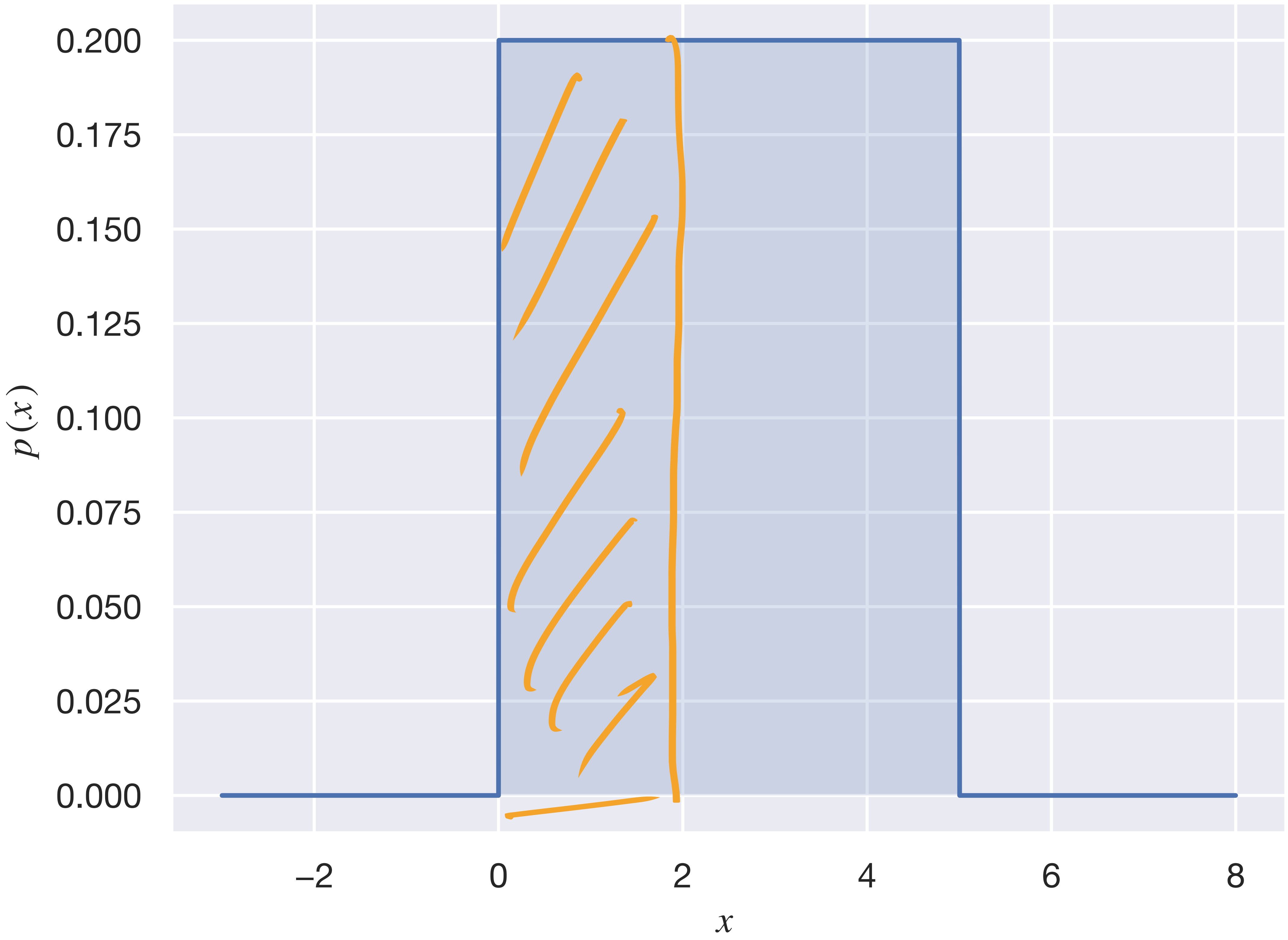
$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} p_X(z)dz \approx 2\epsilon p_X(x).$$

Continuous Random Variables

Example: Picking uniformly in the interval

- X equal density on every point in [a,b].

Example. What's the PDF of the RV $X : \Omega \to \mathbb{R}$ with the uniform random variable on $[0,5]$?



$\mathbb{P}(X \in [0,2])$

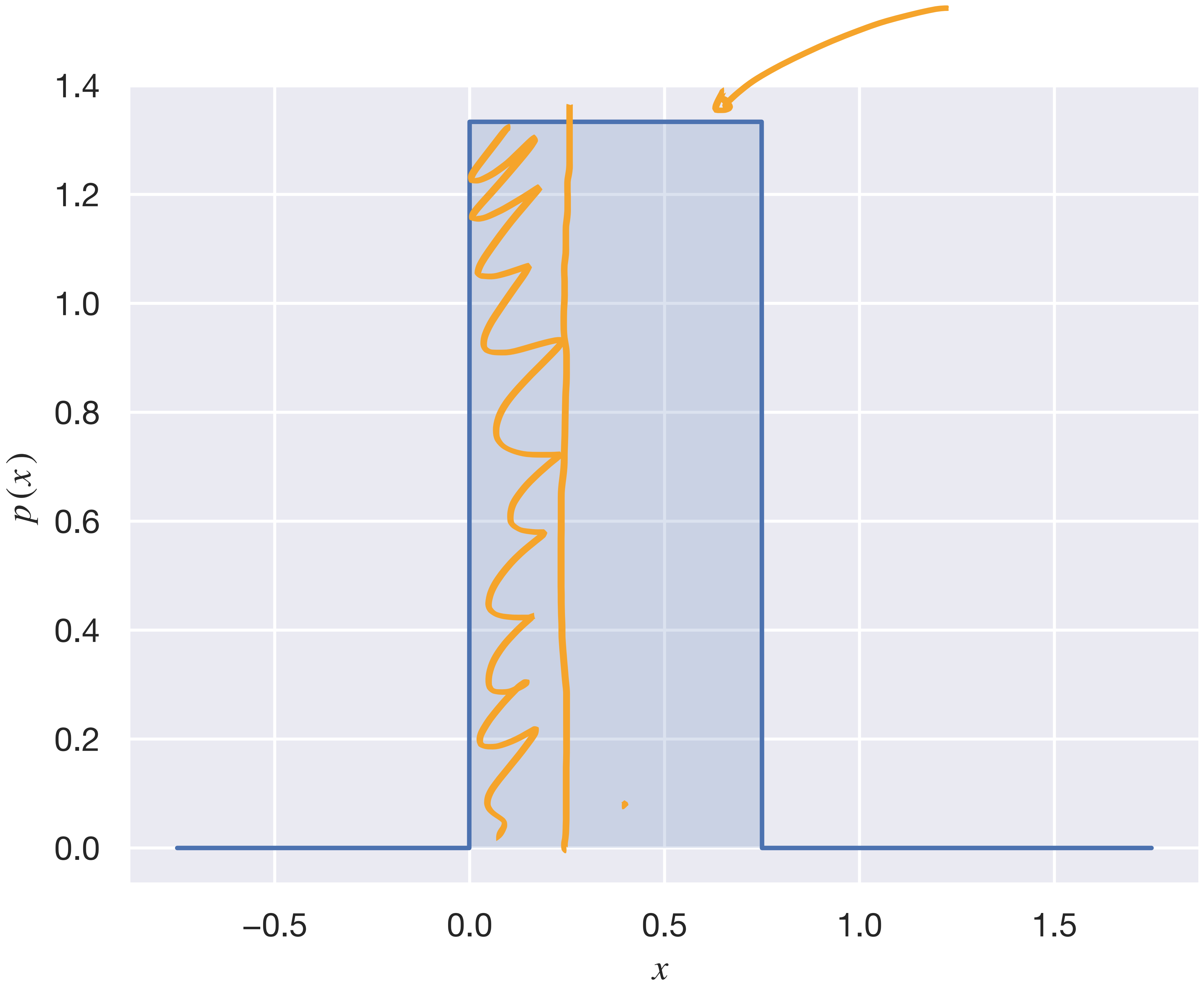
$= 2 \times 0.2 = 0.4$

Continuous vs. Discrete RVs

Example: Uniform Discrete and Uniform Continuous

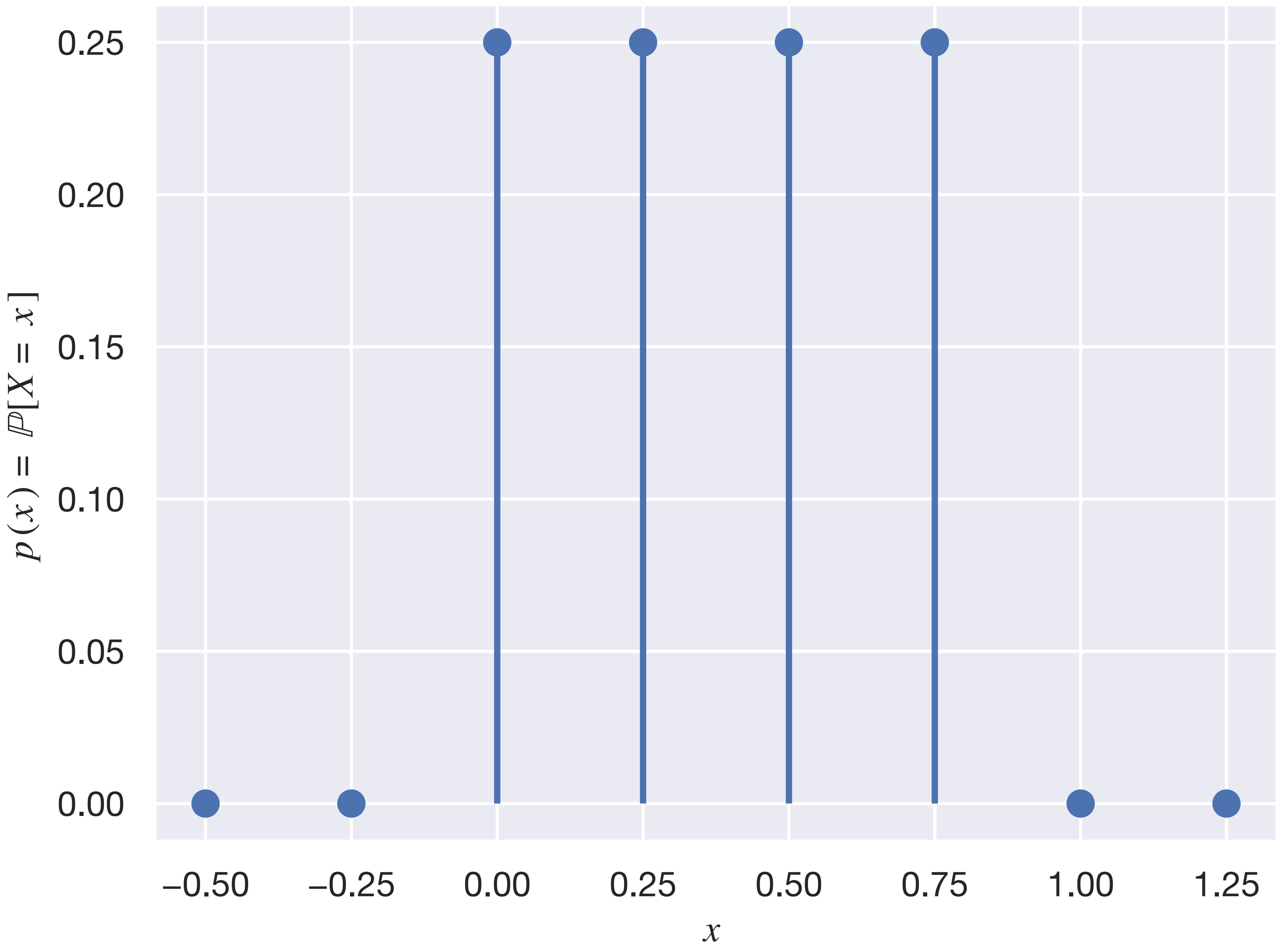
PDF

Continuous RV uniform on $[0,0.75]$.



PMF

Discrete RV uniform on $\{0,0.25,0.5,0.75\}$.



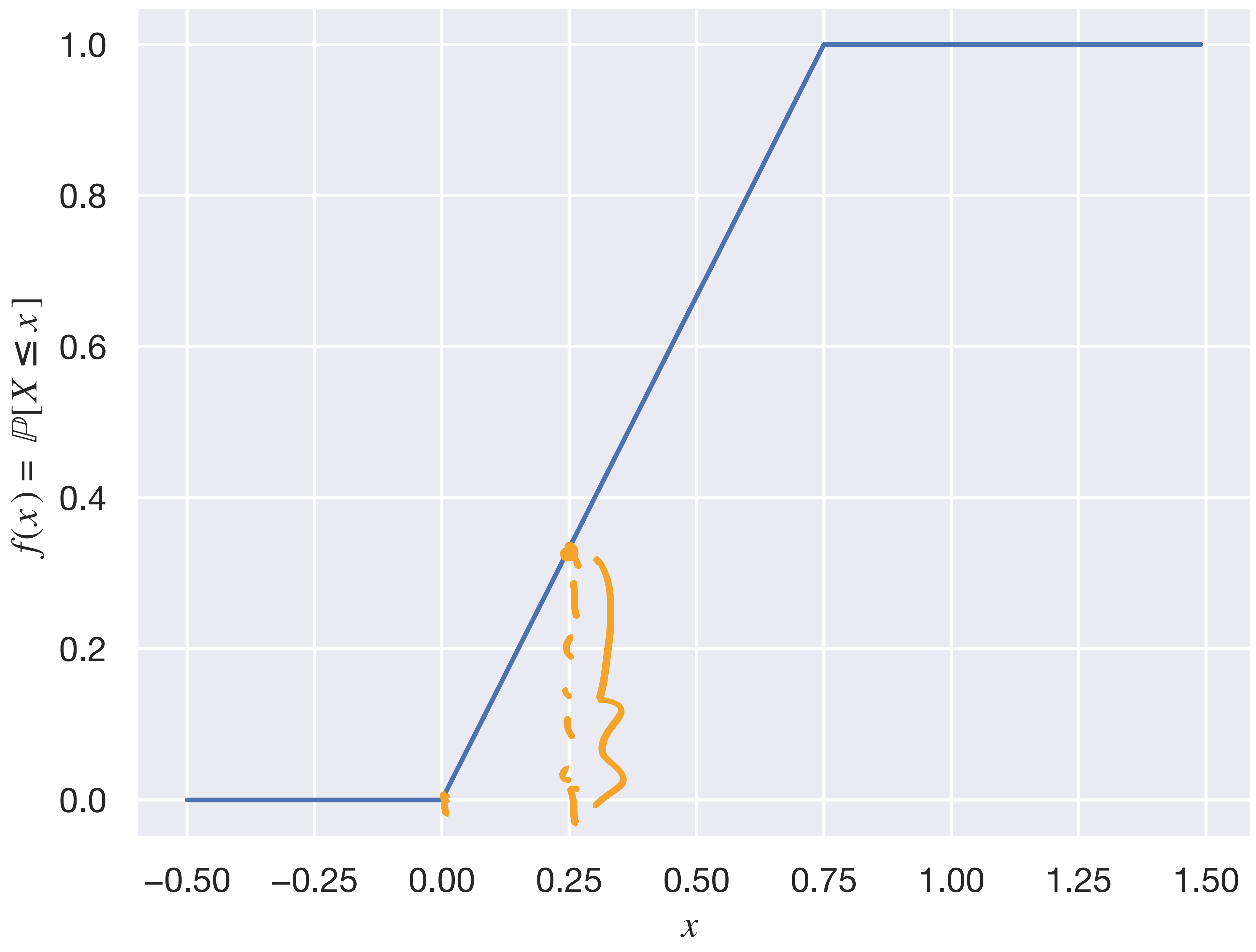
Continuous vs. Discrete RVs

Example: Uniform Discrete and Uniform Continuous

CDF

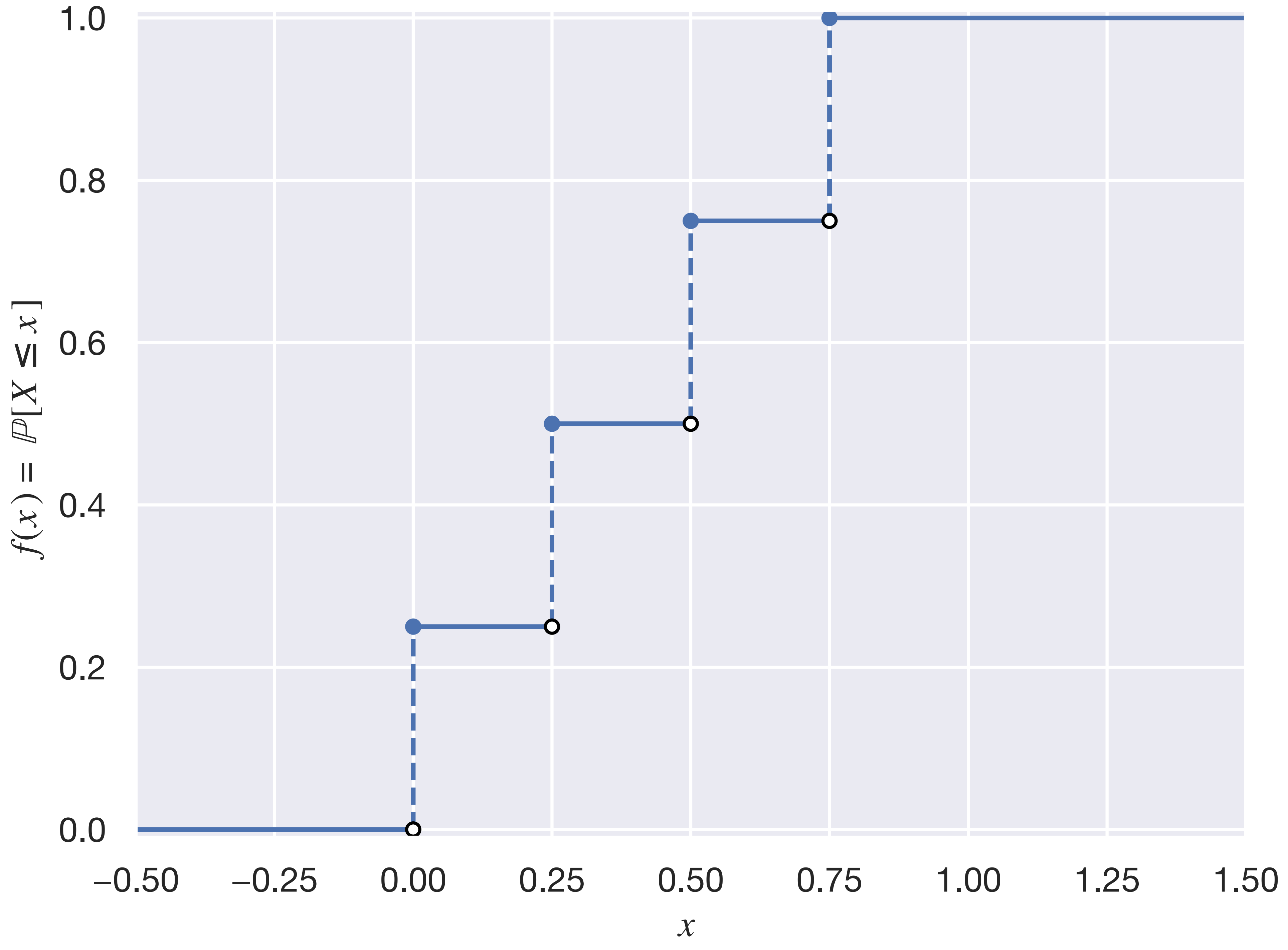
Continuous RV uniform on $[0,0.75]$.

$F(b) - F(a) = \mathbb{P}[a \le X \le b]$



CDF

Discrete RV uniform on $\{0,0.25,0.5,0.75\}$.



Continuous vs. Discrete RVs

Summary

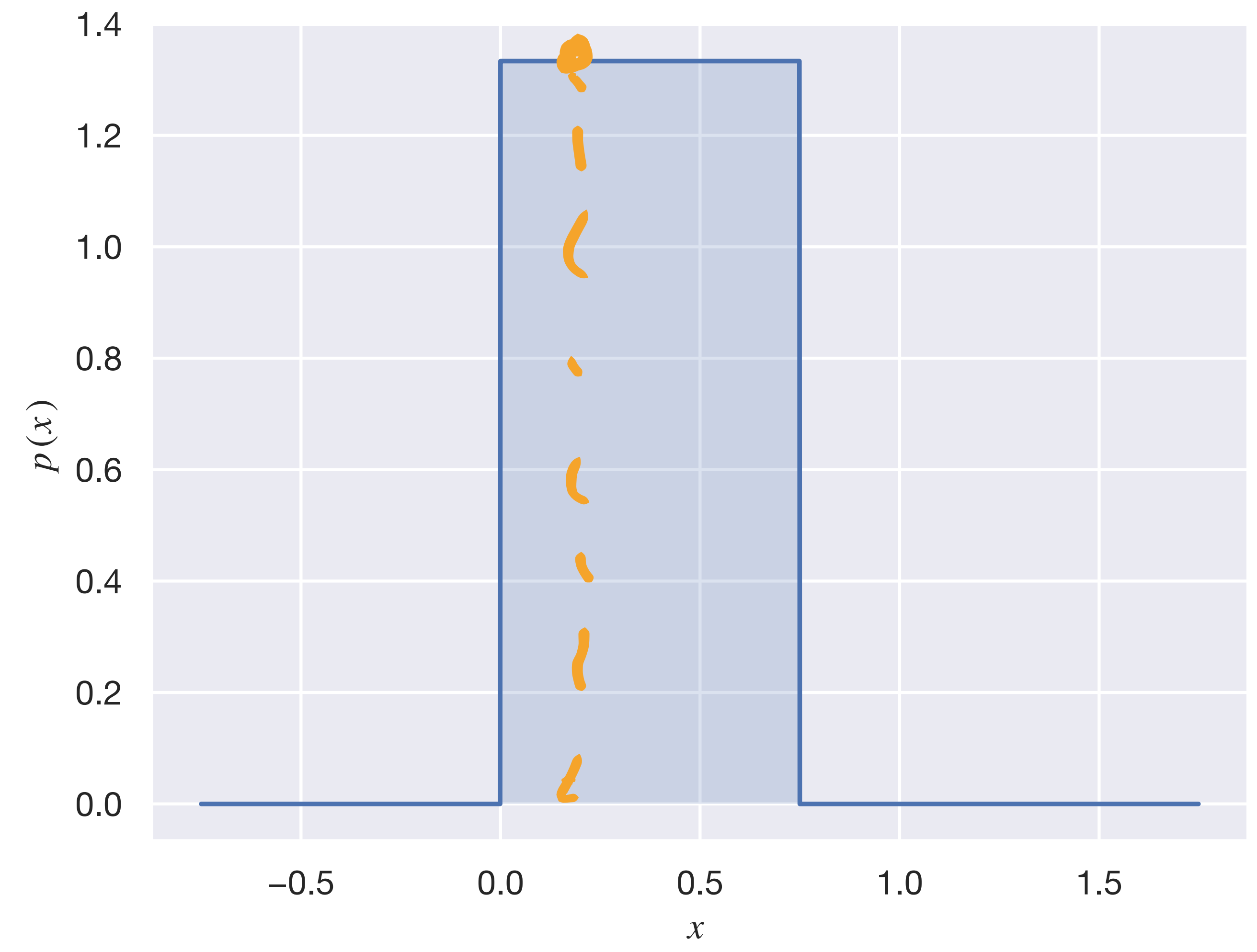
For continuous RVs,

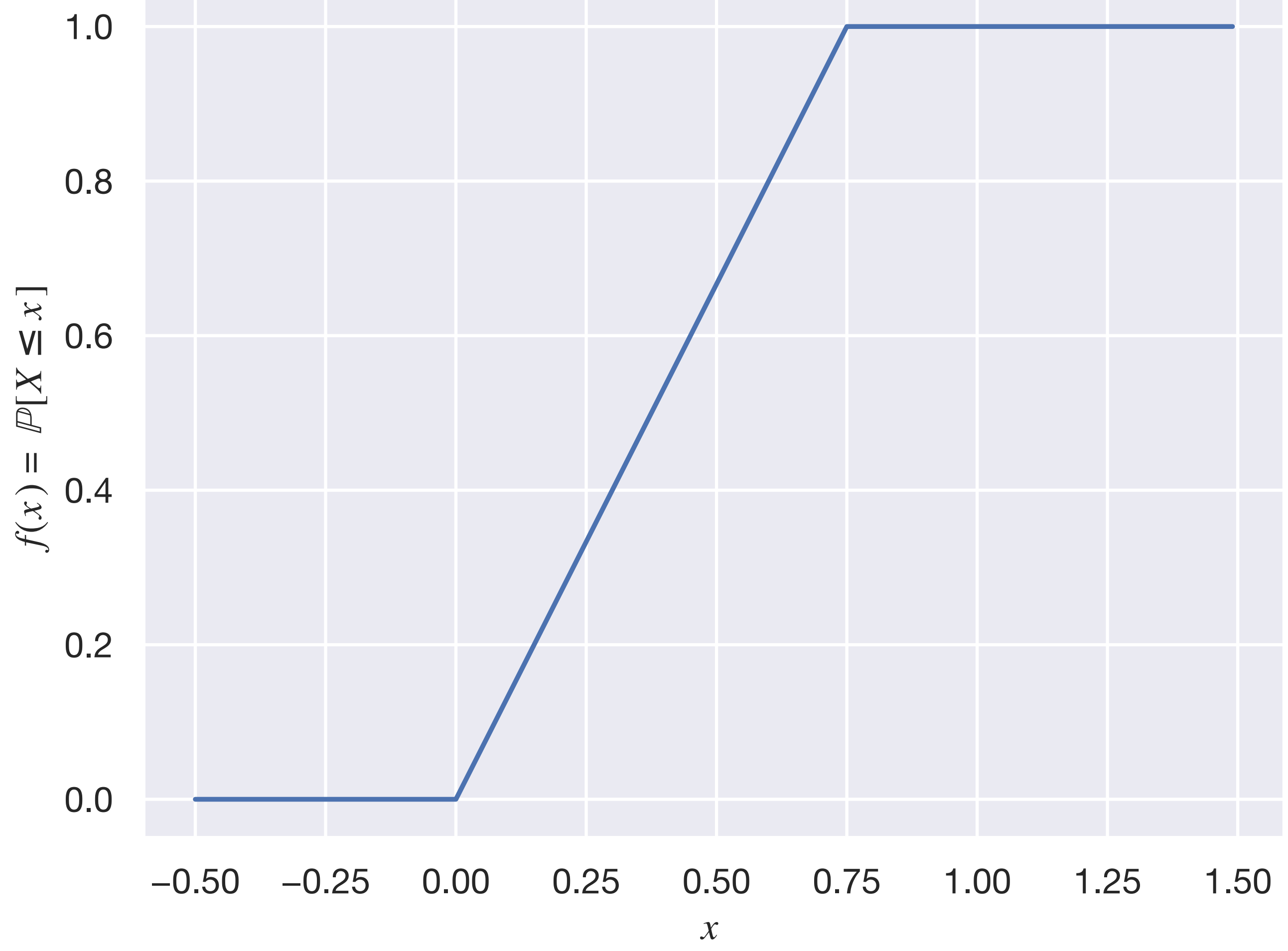
$$\mathbb{P}(X = x) = 0$$

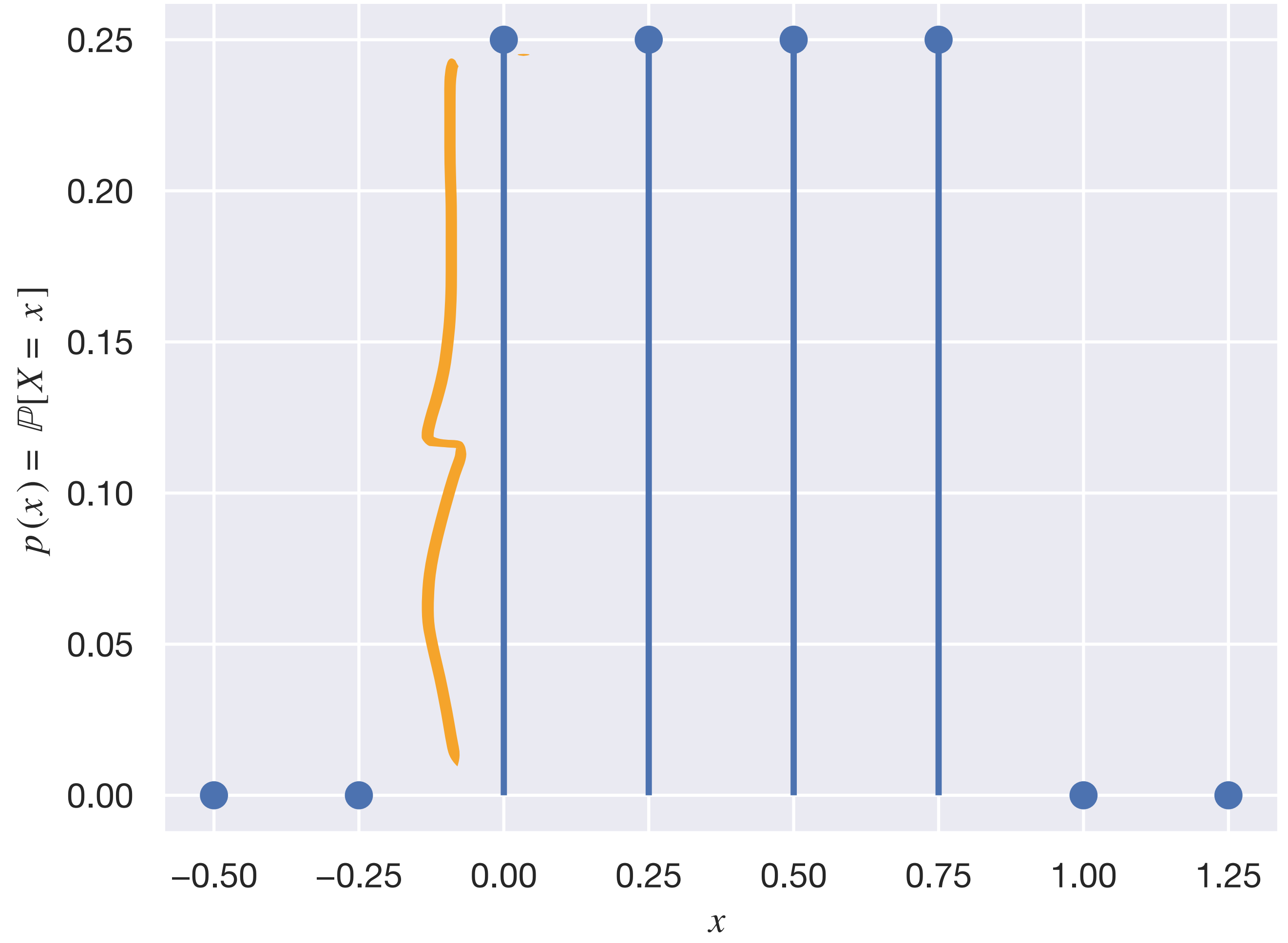
$$\mathbb{P}(b \leq X \leq b) = \int_a^b p_X(x)dx$$

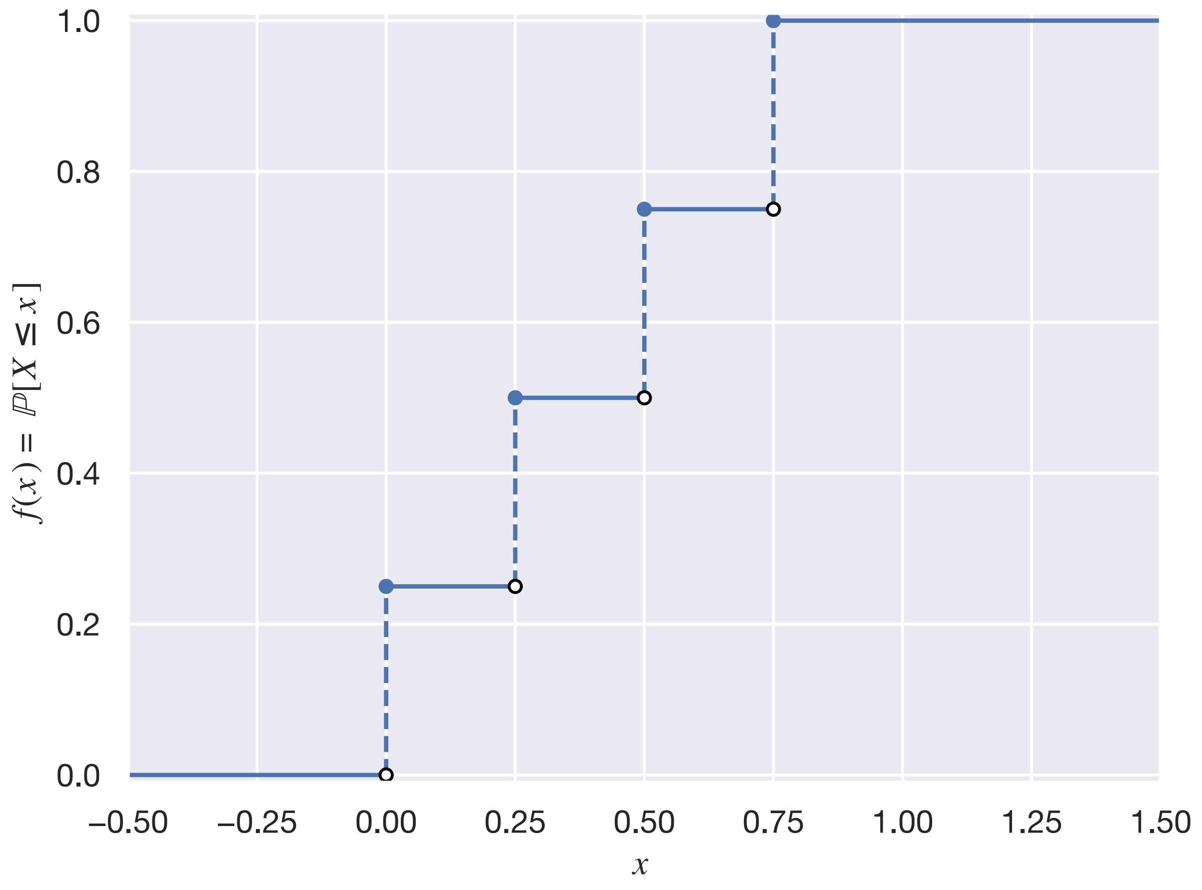
For discrete RVs,

$$\mathbb{P}(X = x) \in [0,1].$$









Random Variables

Multiple random variables

Joint Distribution

Example: Tossing coins and rolling die

Consider two experiments:

Alice tosses a fair coin, Bob tosses a fair coin. → X

Charlie rolls a fair six-sided die. → Y

Let X count the number of heads in the first experiment.

Let Y be the integer of the face of the die in the second experiment.

X \ Y	1	2	3	4	5	6
0						
1						
2						

Joint Distribution

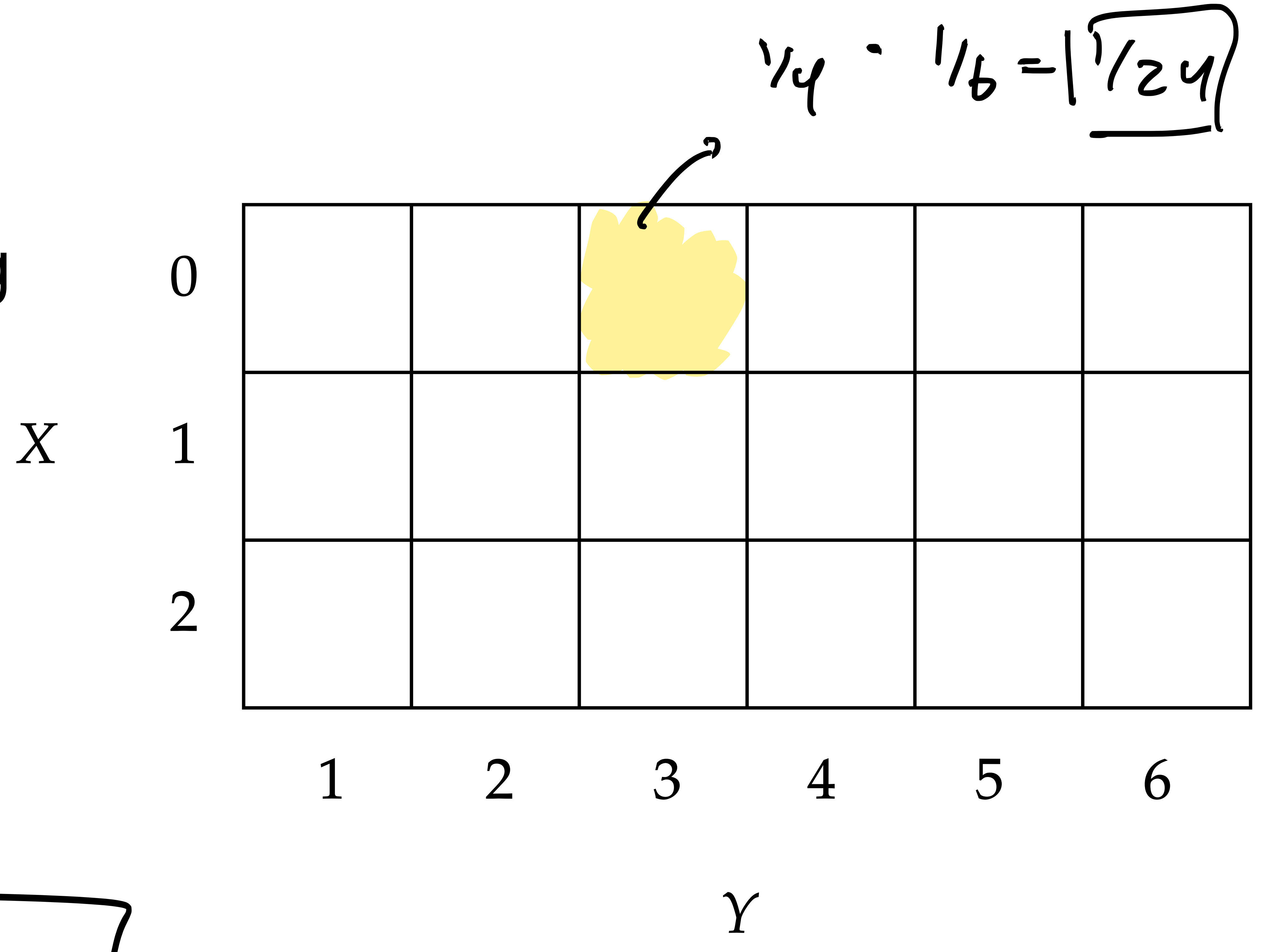
Definition

$\mathbb{P}_{X,Y}$ w/ PMF: $p_{X,Y}$

$p_{X,Y}(0,3) = \mathbb{P}[X=0, Y=3]$

Let $X_1, \ldots, X_n$ be random variables. The ***joint distribution*** of $X_1, \ldots, X_n$ is the probability distribution written $\mathbb{P}_{X_1,\ldots,X_n}$ with corresponding PMF/PDF:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$



For discrete random variables,

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

PMF

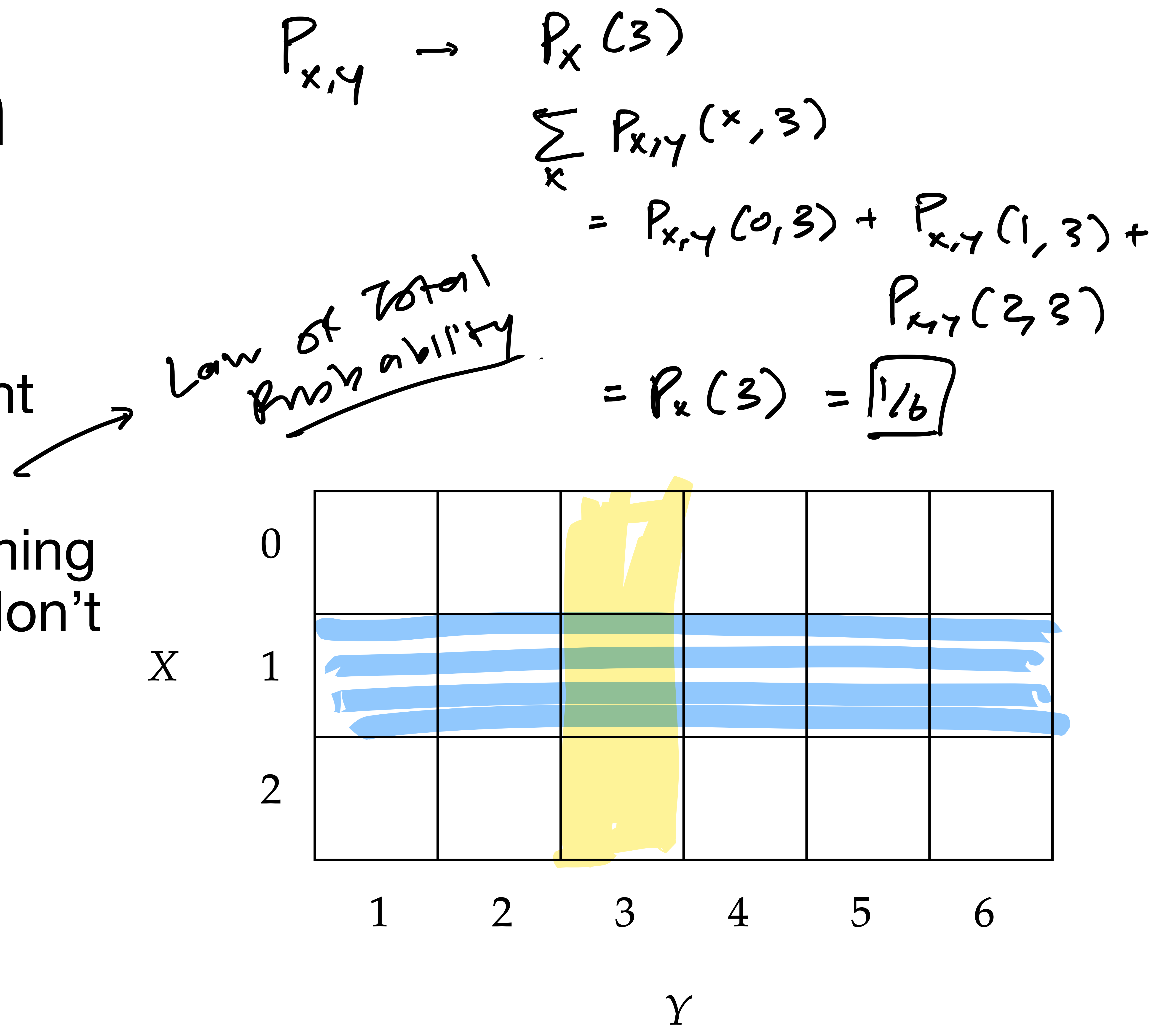
Marginal Distribution

Definition

For two random variables X, Y with joint distribution $p_{X,Y}(x, y)$, the ***marginal distribution*** of X is obtained by "summing out"/"integrating out" the variable we don't care about:

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dy$$



Conditional Distribution

Definition

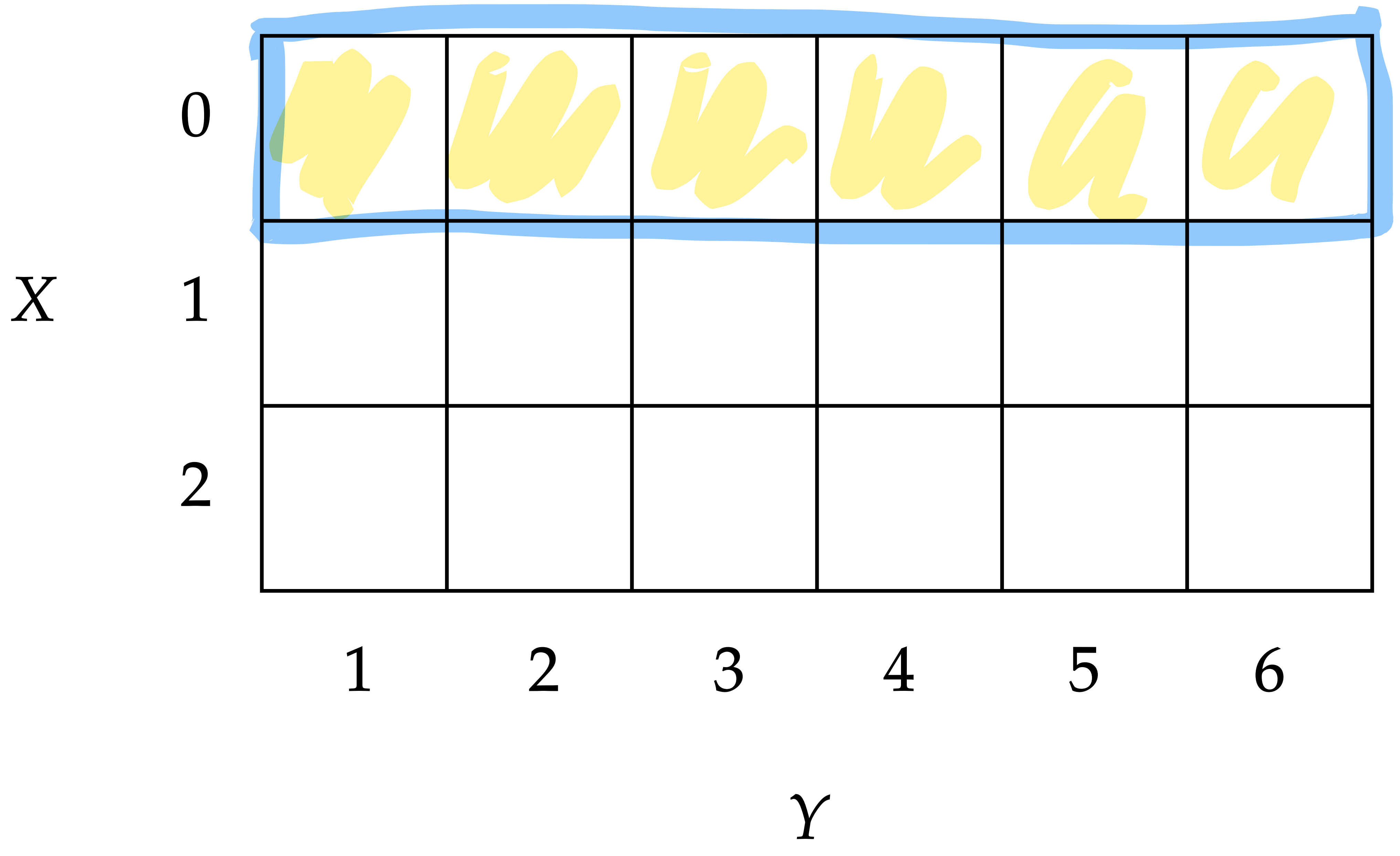
For two random variables X, Y with joint distribution $p_{X,Y}(x, y)$, the ***conditional distribution*** of X given $Y = y$ is given by *only* considering the events where $Y = y$.

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

$$P_{Y|X}(y \mid x)$$

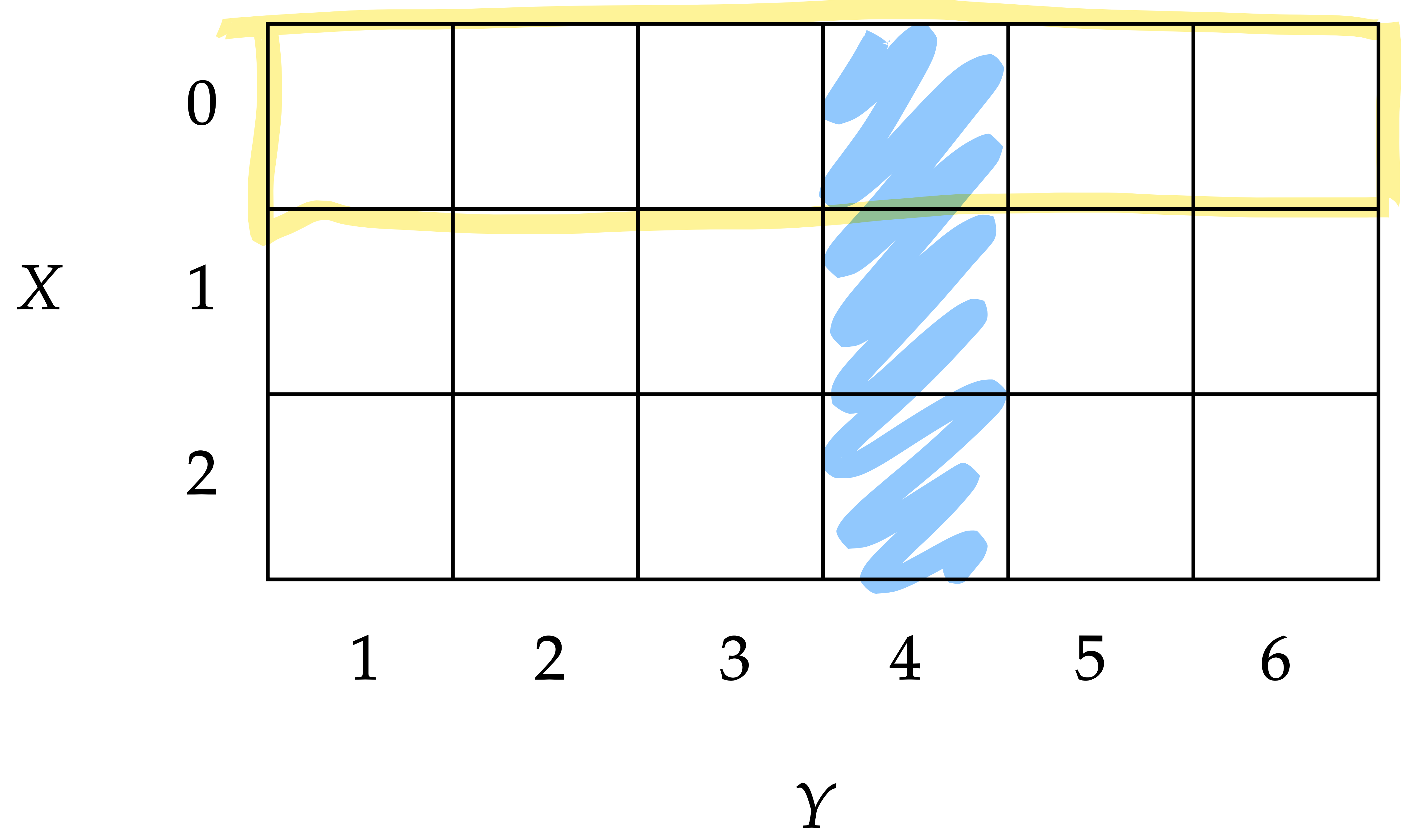
$$= P_{Y|X}(y \mid X=0)$$

$$= \frac{P_{X,Y}(0, y)}{P_X(0)} = \frac{P_{X,Y}(0,y)}{1/4} = 4\, P_{X,Y}(0,y)$$



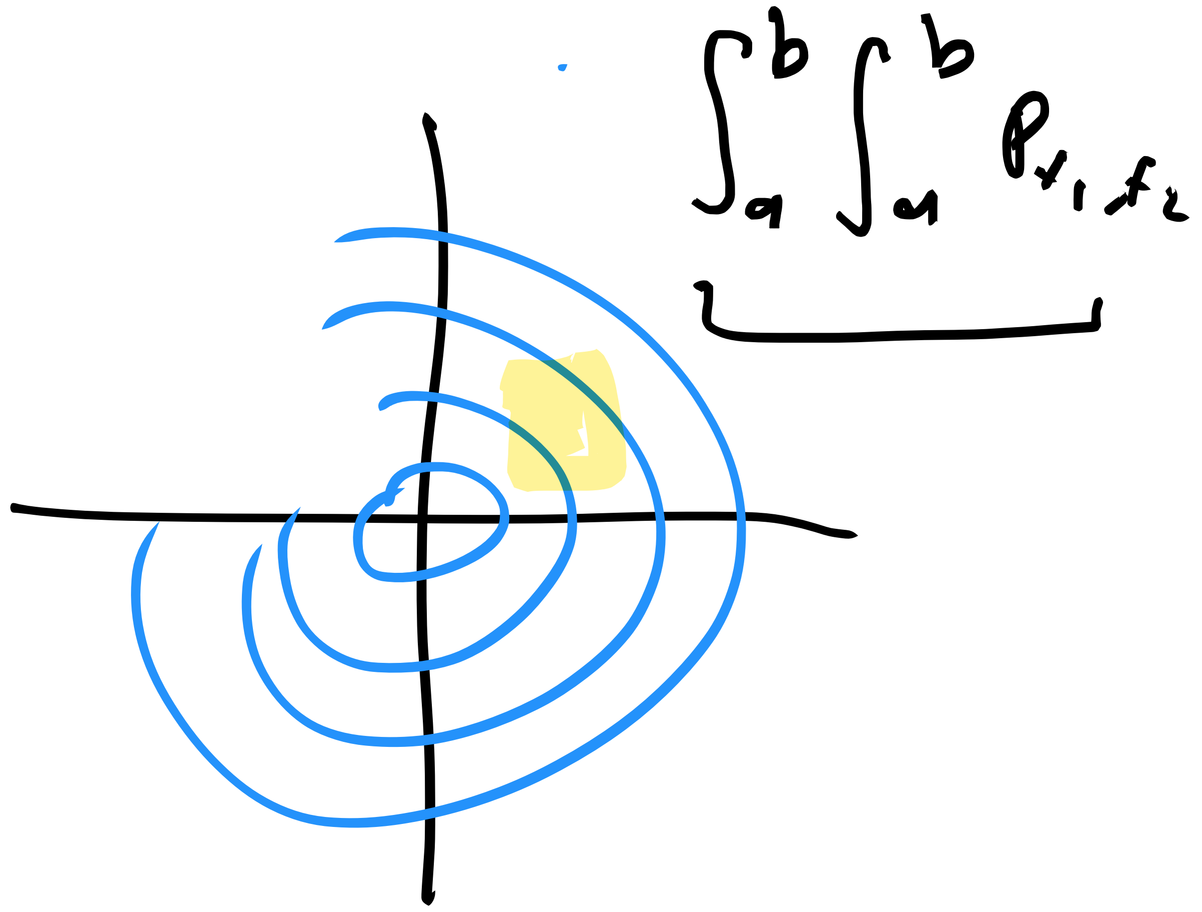
Joint Discrete Distributions

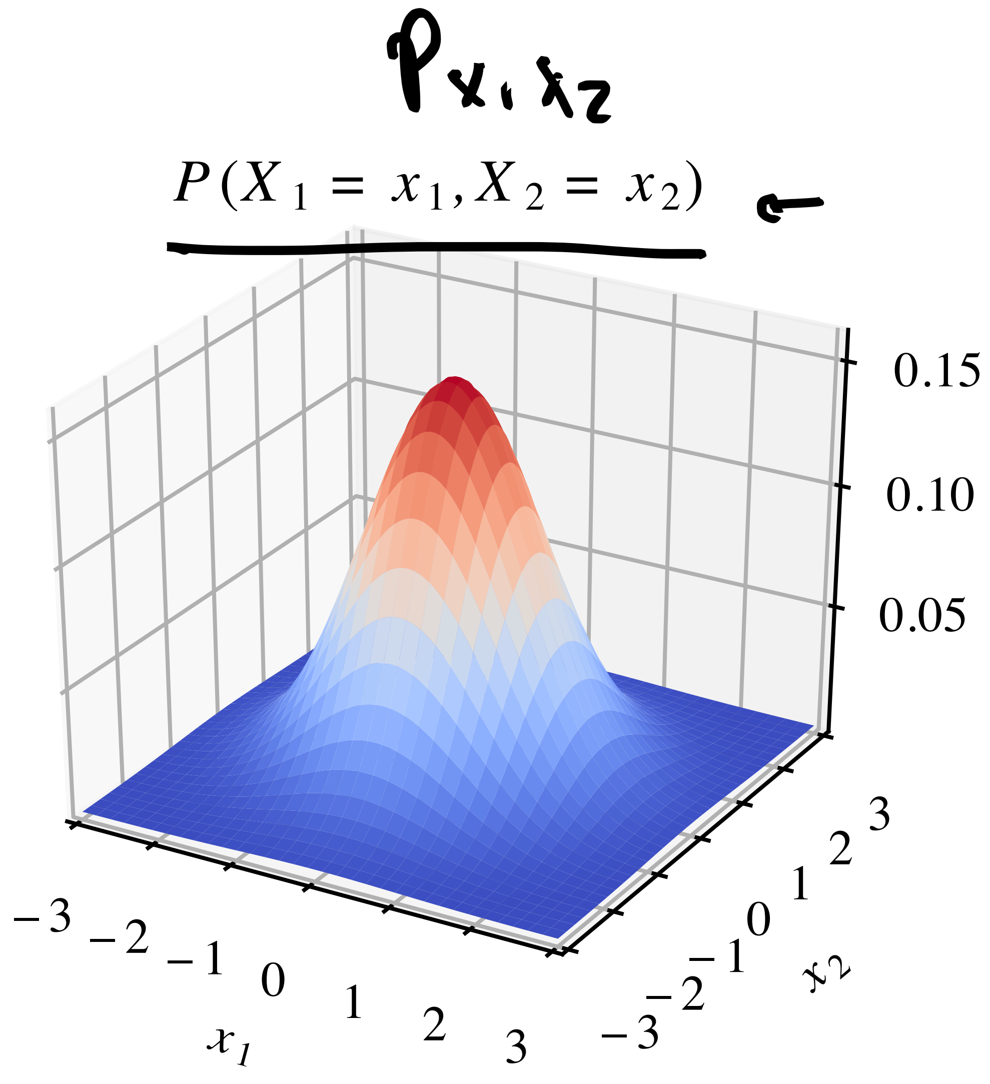
Joint, marginal, and conditional



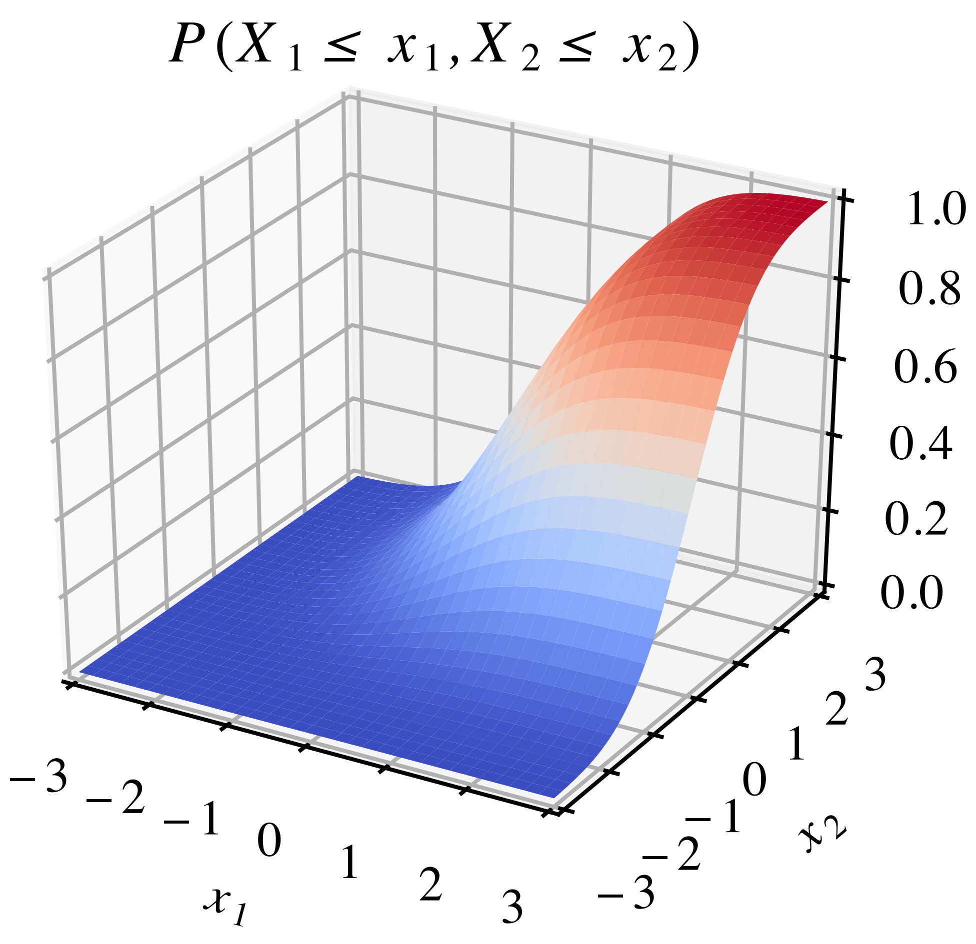
Joint Continuous Distributions

Joint, marginal, and conditional

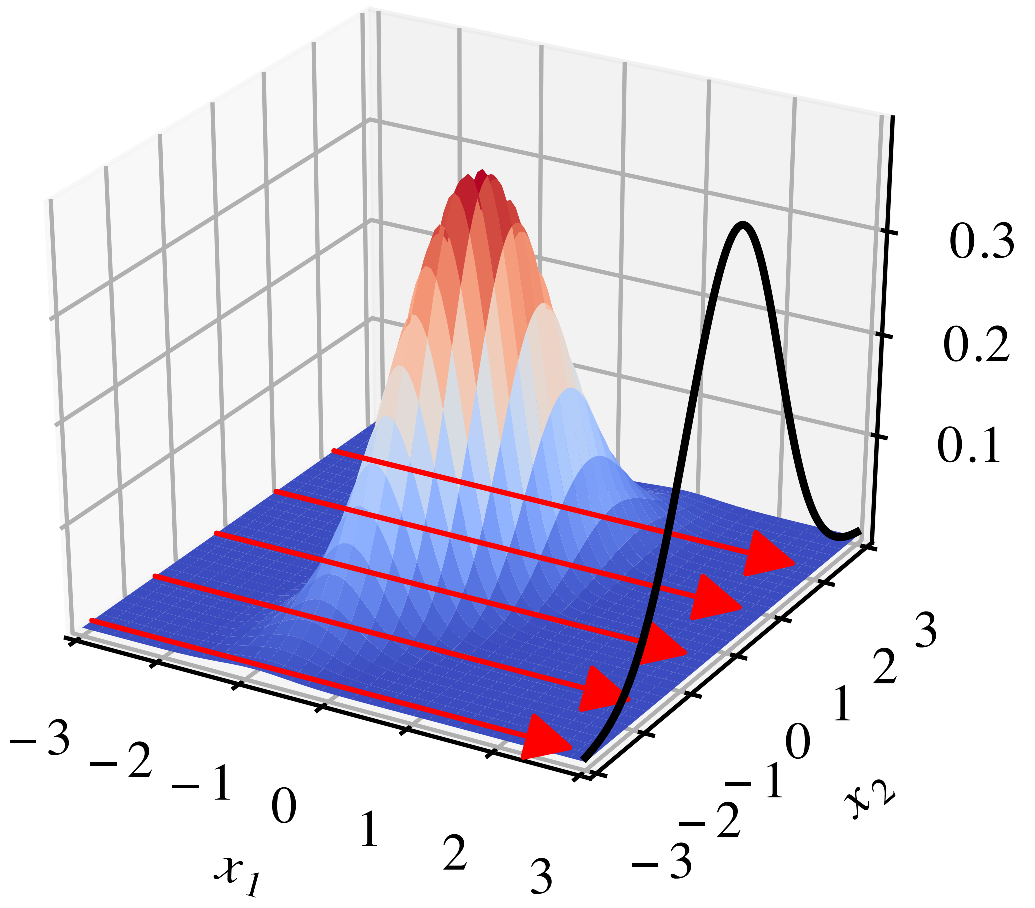




PDF.

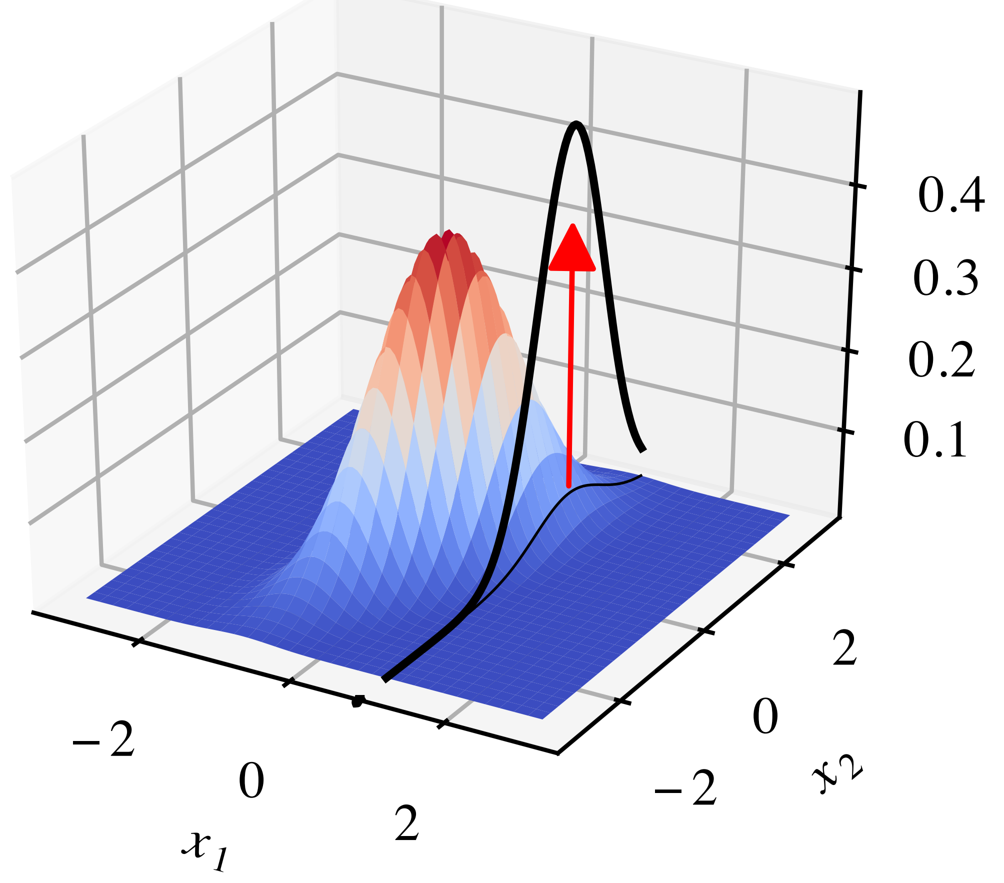


CDF.



Marginal of X_2

$\int p_{x_1, x_2}(x_1, x_2)\, dx_1$



Conditional of $X_1 = 1$

$p_{x_2 | x_1 = 1}$

Joint Distributions

Summary

Let $p_{X,Y}(x, y)$ be a joint distribution.

The ***sum rule/marginalization*** allows us to get from a joint to a marginal distribution.

$$p_X(x) = \begin{cases} \sum_y p_{X,Y}(x, y) & Y \text{ is discrete} \\ \int_{-\infty}^{\infty} p_{X,Y}(x, y) & Y \text{ is continuous} \end{cases}$$

The ***product rule/factorization*** allows us to "factor" the joint distribution into the marginal and conditional distributions.

$$p_{X,Y}(x, y) = p_{Y|X}(y \mid x)p_X(x) = p_{X|Y}(x \mid y)p_Y(y).$$

Independence

Intuition and definition

We say that two random variables X, Y are ***independent*** if their joint distribution factors into their respective distributions:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Another definition: the conditional distribution is the marginal.

$$p_{X|Y}(x \mid y) = p_X(x) \text{ and } p_{Y|X}(y \mid x) = p_Y(y).$$

Knowledge of Y doesn't affect my belief of X

Independence

Intuition and definition

We say that two random variables X, Y are ***independent*** if their joint distribution factors into their respective distributions:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Another definition: the conditional distribution is the marginal.

$$p_{X|Y}(x \mid y) = p_X(x) \text{ and } p_{Y|X}(y \mid x) = p_Y(y).$$

For more than two RVs, let $\{X_i\}_{i \in I}$ be a collection of RVs indexed by I. Then, $\{X_i\}$ are ***independent*** if, for any finite subset of indices $\{i_1, \ldots, i_k\} \in I$,

$$p_{X_{i_1},\ldots,X_{i_k}}(X_{i_1}, \ldots, X_{i_k}) = \prod_{j=1}^{k} p_{X_{i_j}}(x_{i_j}).$$

Independence

Independent and identically distributed (i.i.d.)

A collection of random variables $X_1, \ldots, X_n$ are ***independent and identically distributed (i.i.d.)*** if their joint distribution can be factored entirely:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

Very common assumption in ML!

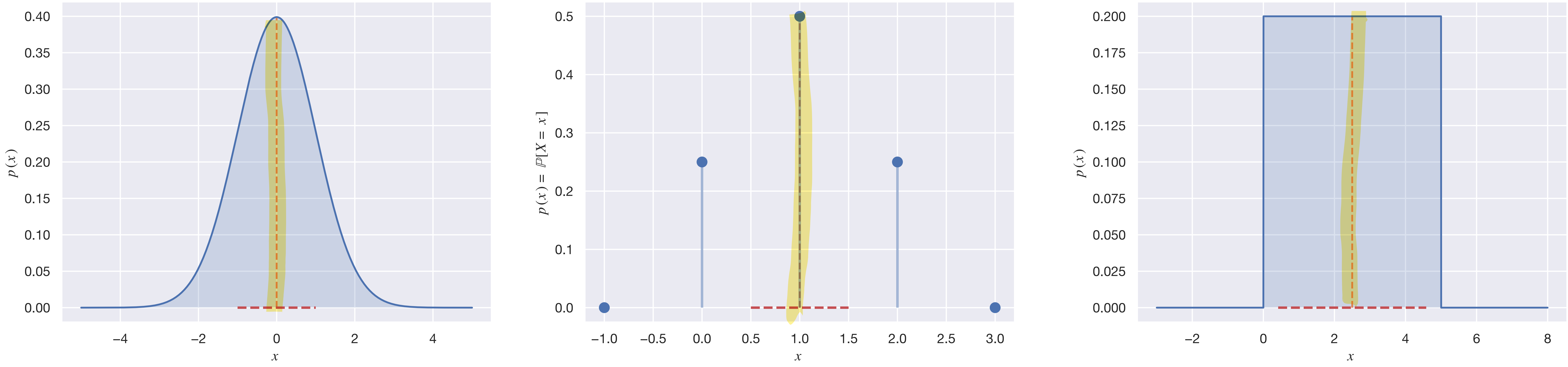
Expectation

Definition and Properties

Expected Value

Intuition

The ***expectation/expected value*** or ***mean*** of a random variable is its "center of mass."



Expected Value

Definition

WEIGHTED SUM

The ***expectation/expected value*** or ***mean*** of a random variable X is

$$\mathbb{E}[X] = \sum_{x} x p_X(x) \text{ for discrete } X$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) dx \text{ for continuous } X$$

Expected Value

Definition (Functions of RVs)

The ***expectation/expected value*** or ***mean*** of a function $g(X)$ of a random variable X is

$$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x) \text{ for discrete } X$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)dx \text{ for continuous } X$$

A function of a random variable is a random variable!

Expected Value

Properties of the expected value

Linearity. The expectation is a linear operator:

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X] \text{ and } \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y],$$

for *any* random variables X and Y (need not be independent)!

Expected Value

Properties of the expected value

$\mathbb{E}$: functions $\to \mathbb{R}$.

Linearity. The expectation is a linear operator:

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X] \text{ and } \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y],$$

for *any* random variables X and Y (need not be independent)!

Product (for indepndent RVs). For independent random variables X, Y

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

More generally, for independent $X_1, \ldots, X_n$:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

Conditional Expectation

Intuition

The ***conditional expectation*** is the "best guess" of a random variable's expectation, given an event occurs.

Depending on context, this is a *random variable* or a *function.*

$\mathbb{E}[X \mid Y = y]$ is a function $g(y) = \mathbb{E}[X \mid Y = y]$.

$\mathbb{E}[X \mid Y]$ is a random variable $g(Y)$.

Conditional Expectation

Intuition

Consider the roll of a six-sided fair die. → Equally probable

Let $X = 1$ if the roll is even, $X = 0$ otherwise.

Let $Y = 1$ if the roll is prime, $Y = 0$ otherwise.

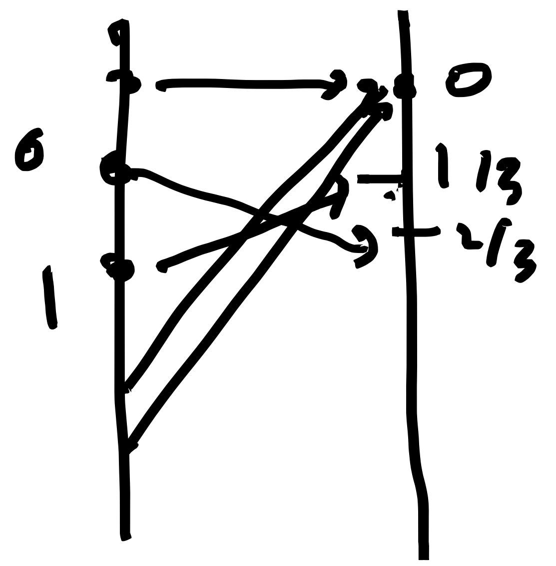
outcomes

	1	2	3	4	5	6
X	0	1	0	1	0	1
Y	0	1	1	0	1	0

What is $\mathbb{E}[X]$? $= 1/6 \cdot 0 + 1/6 \cdot 1 + \ldots + 1/6 \cdot 1 = 1/2$

What is $\mathbb{E}[X \mid Y = 1]$? $= 1/3 \cdot 1 + 1/3 \cdot 0 + 1/3 \cdot 0 = 1/3$

What is $\mathbb{E}[X \mid Y = 0]$? $= 1/3 \cdot 0 + 1/3 \cdot 1 + 1/3 \cdot 1 = 2/3$



What is $\mathbb{E}[X \mid Y = y]$ and $\mathbb{E}[X \mid Y]$?

$\mathbb{E}[X \mid Y = y] : \mathbb{R} \to \mathbb{R}$

$g(0) = 2/3 \quad g(1) = 1/3 \quad g(x) = 0$ for $x \neq 0, 1$.

Conditional Expectation

Definition (given events)

→ $\mathbb{R}$ (a number)

If A is an event and X is a discrete random variable, the ***conditional expectation*** of X given A is:

$$\mathbb{E}[X \mid A] = \sum_x \mathbb{P}_X[X = x \mid A].$$

If X, Y are discrete random variables, the ***conditional expectation*** of X given $Y = y$ is:

$$\mathbb{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y) = \sum_x x \mathbb{P}[X = x \mid Y = y].$$

← discrete

If X, Y are continuous random variables with joint density $p_{X,Y}(x, y)$, Y's marginal $p_Y(y)$ and conditional density $p_{X|Y}(x \mid y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$, the ***conditional expectation*** of X given $Y = y$ is:

← continuous

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x p_{X|Y}(x \mid y) dx.$$

Conditional Expectation

Definition (given a random variable)

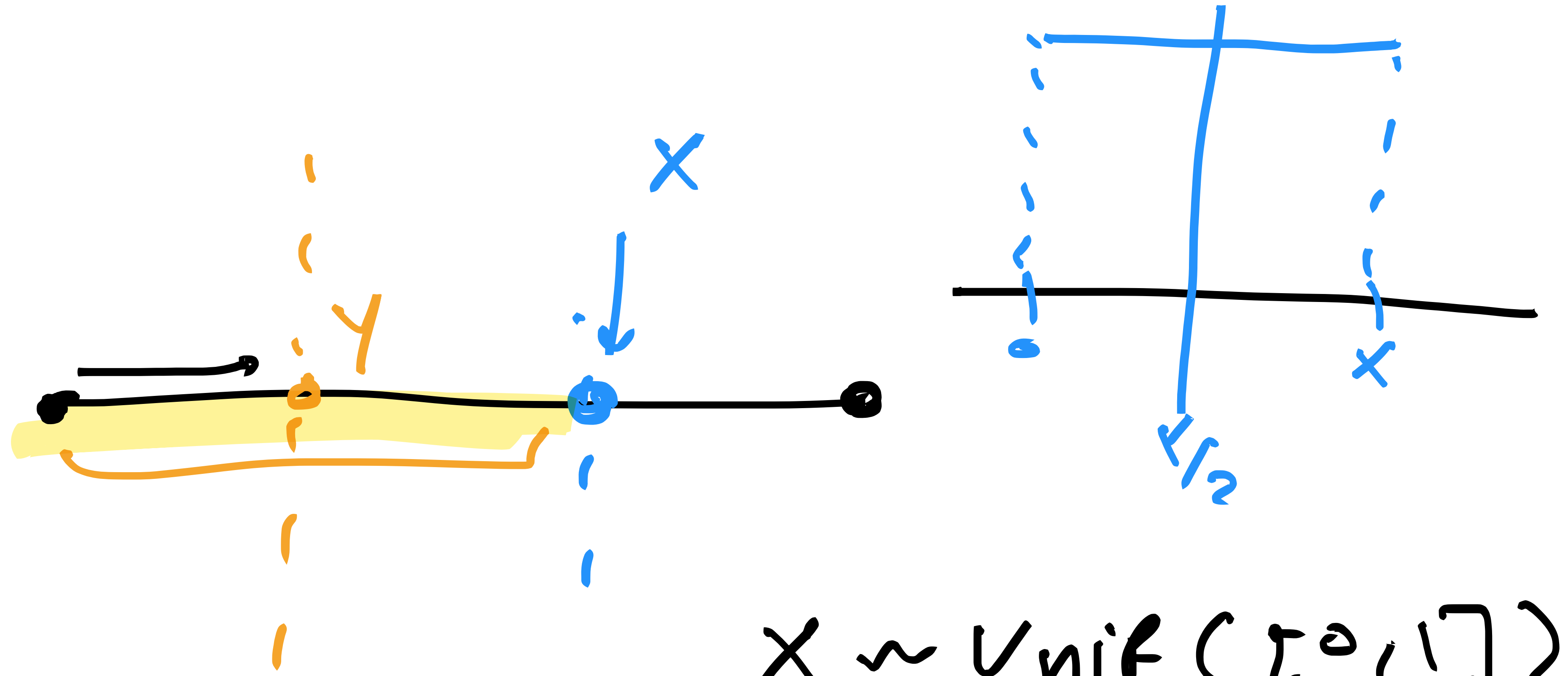
For two random variables X and Y, think of the ***conditional expectation*** of X given Y as the "best guess" of Y only using the information from X:

$\mathbb{E}[Y \mid X]$ is a *random variable* (a function $g(X)$ of the RV X).

We can obtain this random variable by figuring out the function $g(x)$ for $\mathbb{E}[Y \mid X = x]$ and then "plugging back in" the random variable $g(X)$.

Conditional Expectation

Definition (given a random variable)



$X \sim \text{Unif}([0,1])$

Example. A stick of length 1 is broken at a point X chosen uniformly at random. Given that $X = x$, choose another breakpoint Y uniformly on the interval $[0,x]$. What is the random variable $\mathbb{E}[Y \mid X]$? What is its mean?

$\mathbb{E}[Y \mid X]$

$\mathbb{E}[Y \mid X = x] \leftarrow$ "What is the expected value of Y if the stick is $[0, x]$?"

$g(x) = x/2$

$\mathbb{E}[Y \mid X] = g(X) = X/2$

$\mathbb{E}[Y] =$

$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[g(X)] = \mathbb{E}[\frac{X}{2}] = \frac{1}{2}\mathbb{E}[X]$

$= \frac{1}{2} \cdot \frac{1}{2} = 1/4$

Conditional Expectation

Properties of conditional expectation

Independence. If X is independent of Y,

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X].$$

Pulling out what's known. For any function g,

$$\mathbb{E}[h(X)Y \mid X] = h(X)\mathbb{E}[Y \mid X].$$

Linearity. For any random variables X, Y, Z and scalar $\alpha \in \mathbb{R}$,

$$\mathbb{E}[X + Y \mid Z] = \mathbb{E}[X \mid Z] + \mathbb{E}[Y \mid Z] \text{ and } \mathbb{E}[\alpha X \mid Z] = \alpha\mathbb{E}[X \mid Z].$$

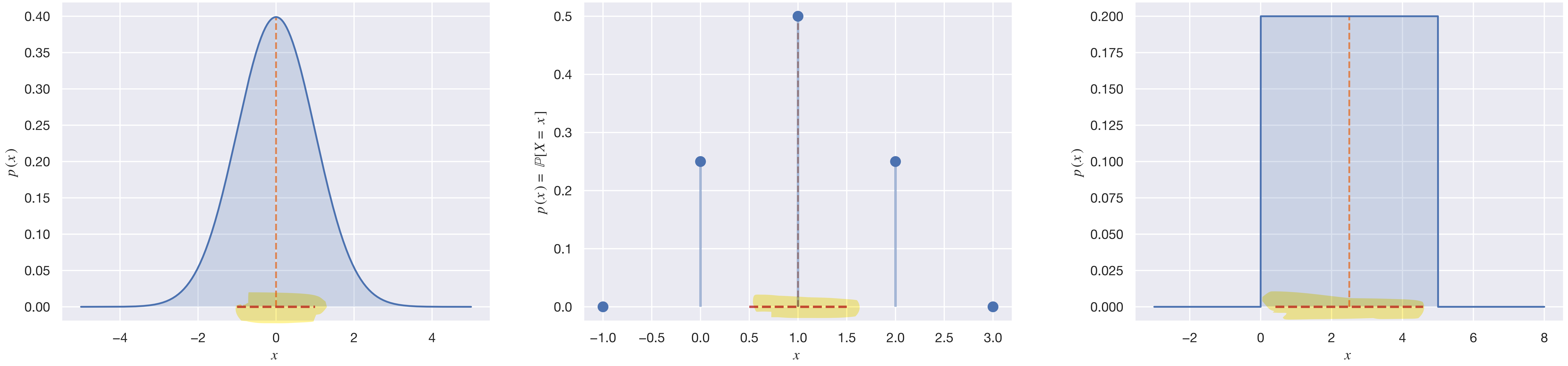
Law of total expectation/tower rule. For any random variables X, Y,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y].$$

Variance
Definition and Covariance

Variance
Intuition

The ***variance*** of a random variable is how "spread" around its expectation it is.



Variance

Definition

$\|X\|^2 = \sum (X - \mathbb{E}[X])^2$

The ***variance*** of a random variable $\mathrm{Var}(X)$ is:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

$$= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$$

Linearity

$$= \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2]$$

* constants are not random!

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2$$

This can also be written (using linearity of expectation):

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Variance
Definition

The ***variance*** of a random variable $\mathrm{Var}(X)$ is:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

This can also be written (using linearity of expectation):

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The ***standard deviation*** is $\sqrt{\mathrm{Var}(X)}$. → units we started.

Variance

Properties of variance

$\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$
$\mathrm{Var}(\beta) = 0$

The variance is *NOT* linear, but we do have, for $\alpha, \beta \in \mathbb{R}$,

$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X).$$

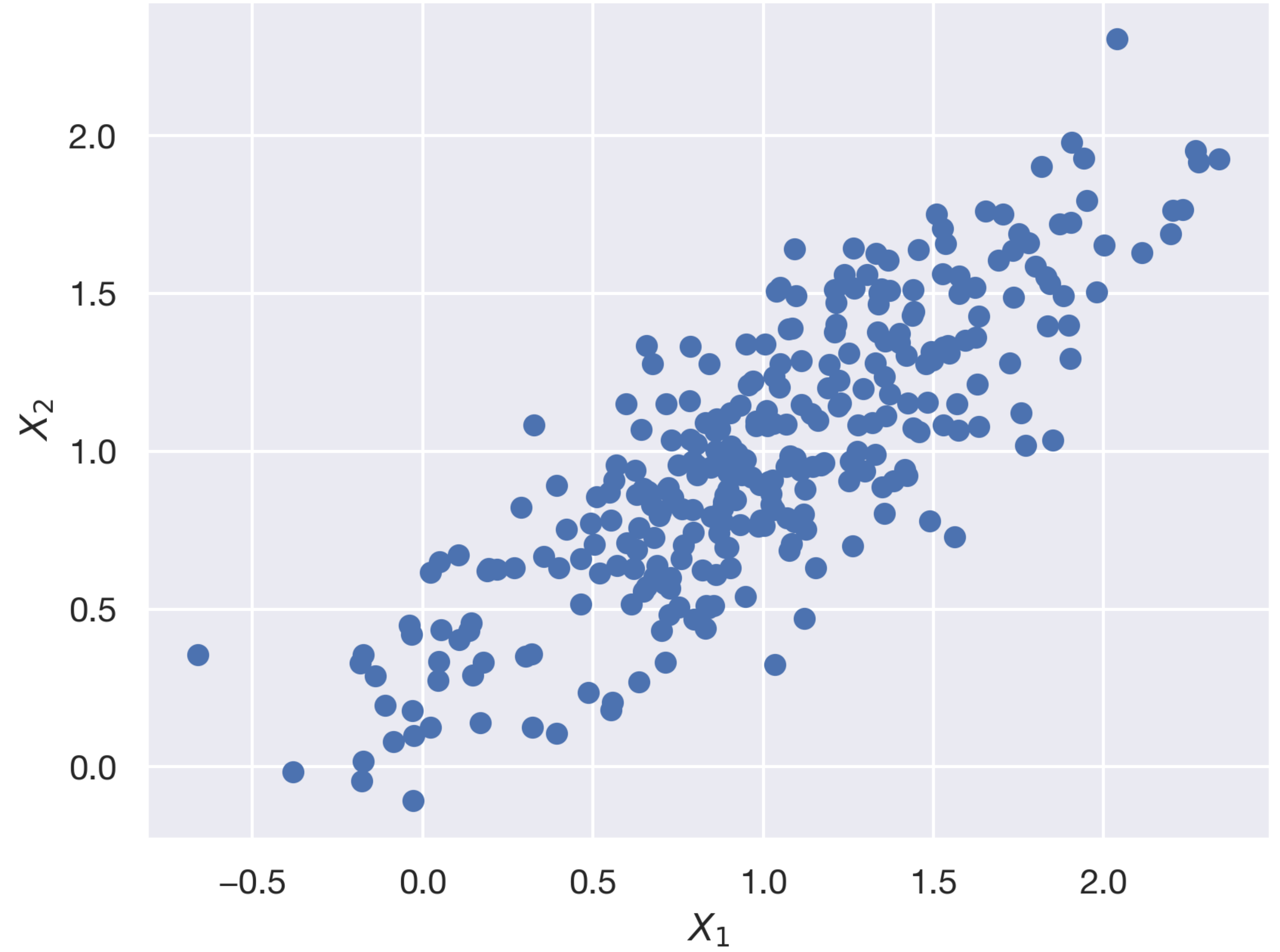
If $X_1, \ldots, X_n$ are independent (more generally, *uncorrelated*),

$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n).$$

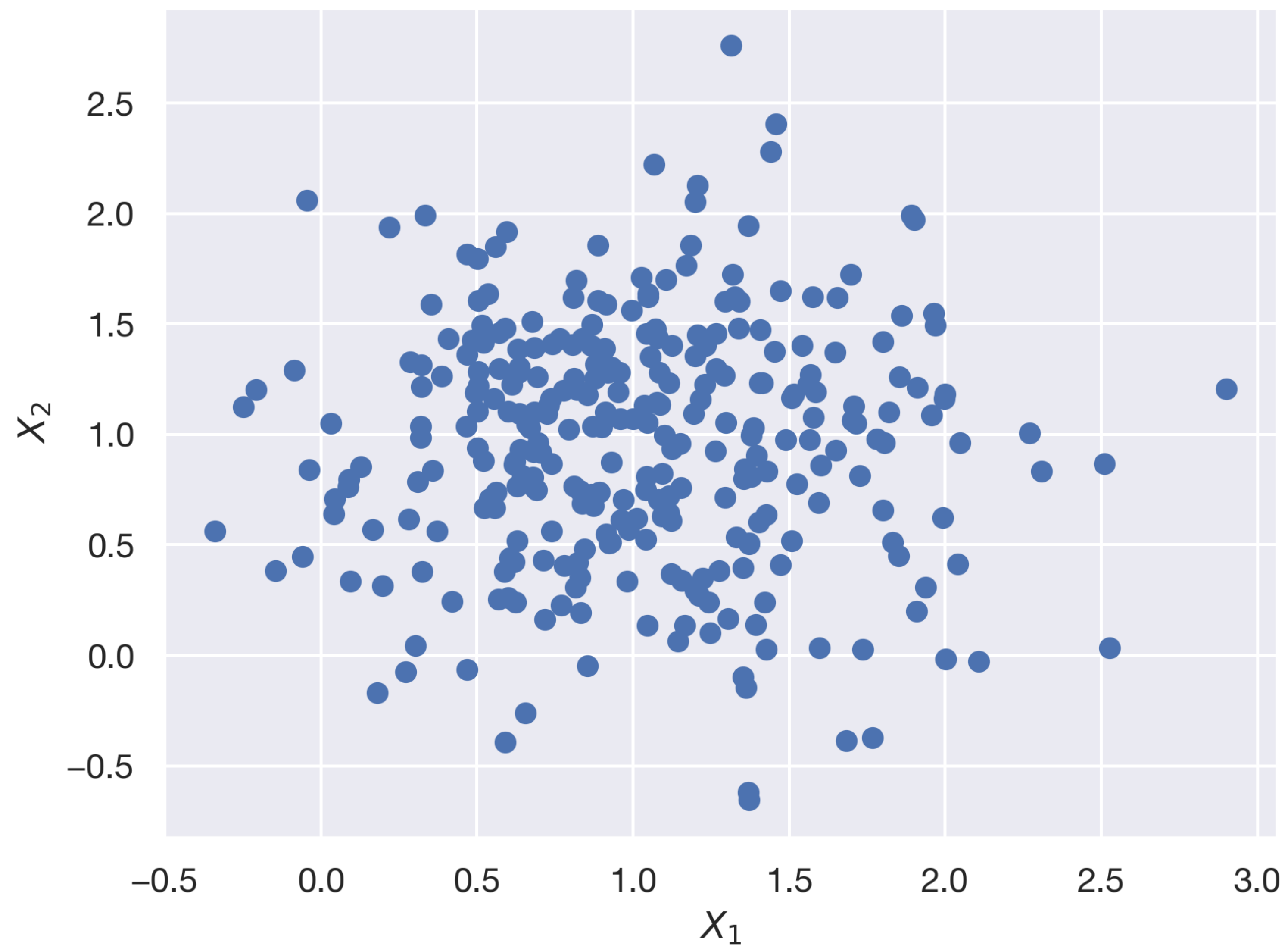
Covariance

Intuition

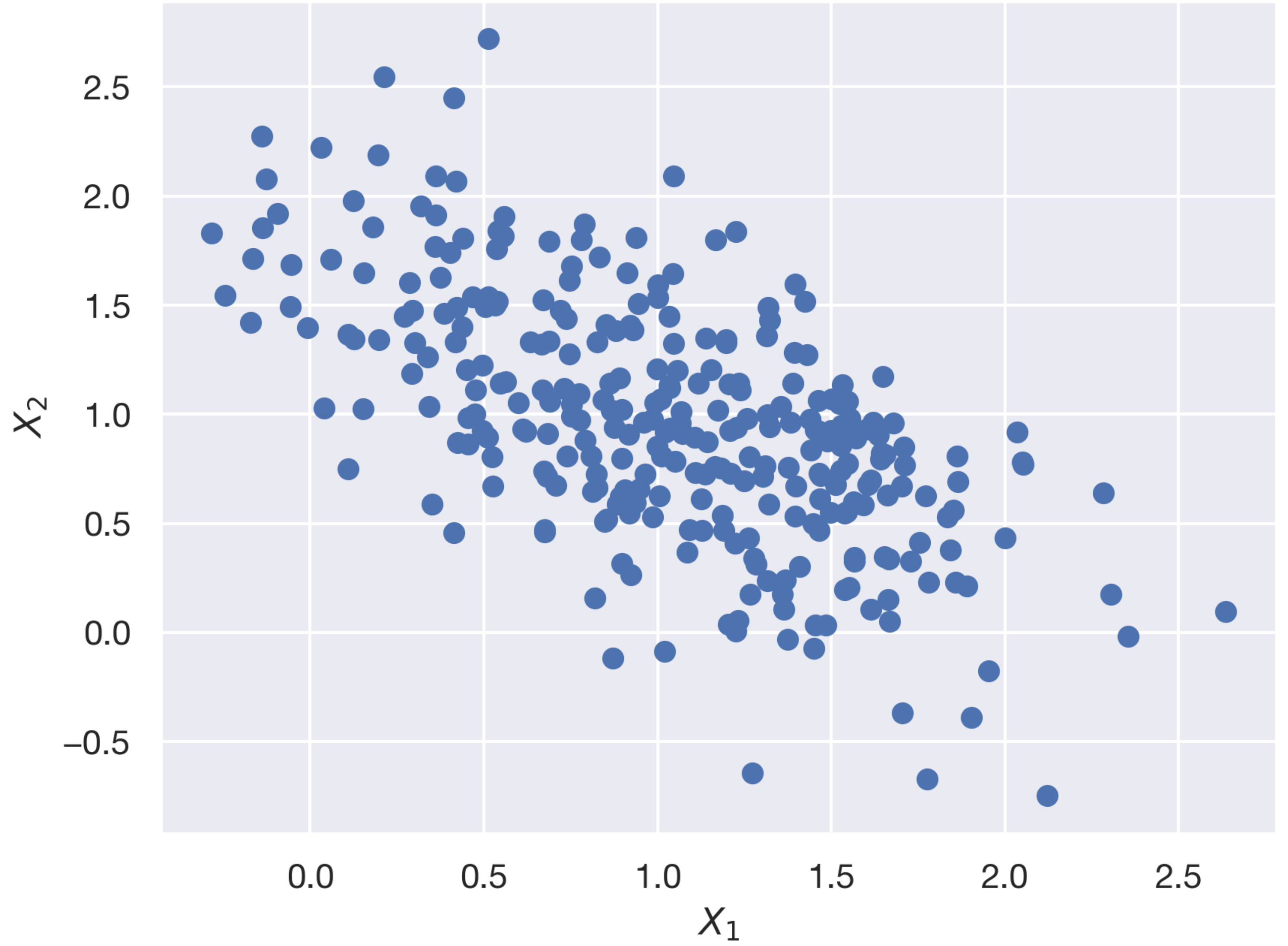
The ***covariance*** measures the linear relationship between two random variables.



$cov(X_1, X_2) > 0$



$cov(X_1, X_2) = 0$



$cov(X_1, X_2) < 0$

Covariance

Definition

cov(x,x) = E[(x - E(x))²]

The ***covariance*** of X, Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The outer expectation is over both X and Y (their joint distribution).

This can also be rewritten as:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Covariance
Definition

The ***covariance*** of X, Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The outer expectation is over both X and Y (their joint distribution).

This can also be rewritten as:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

The ***correlation*** is what we get from normalizing the covariance:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance

Properties of covariance

$X^T X.$

Covariance follows the “symmetry” property:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X).$$

Covariance follows the “bilinearity” property:

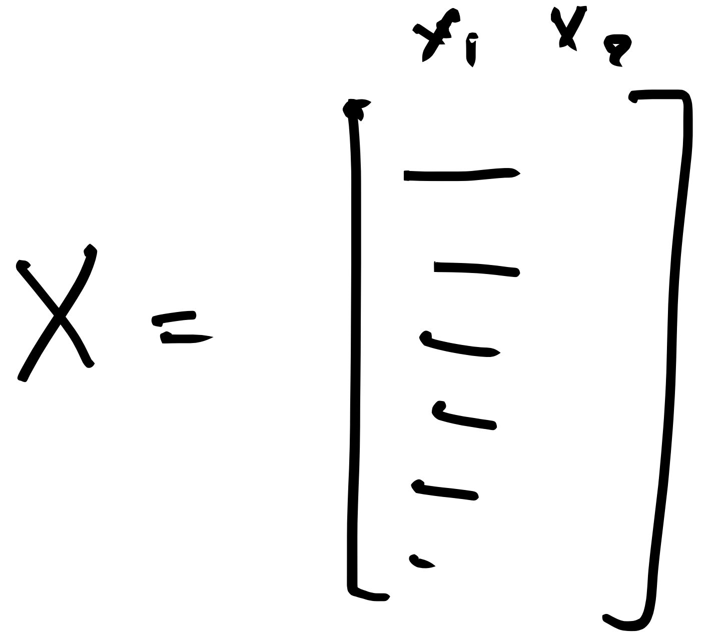
$$\mathrm{Cov}(\alpha X + \beta Y, Z) = \alpha \mathrm{Cov}(X, Z) + \beta \mathrm{Cov}(Y, Z).$$

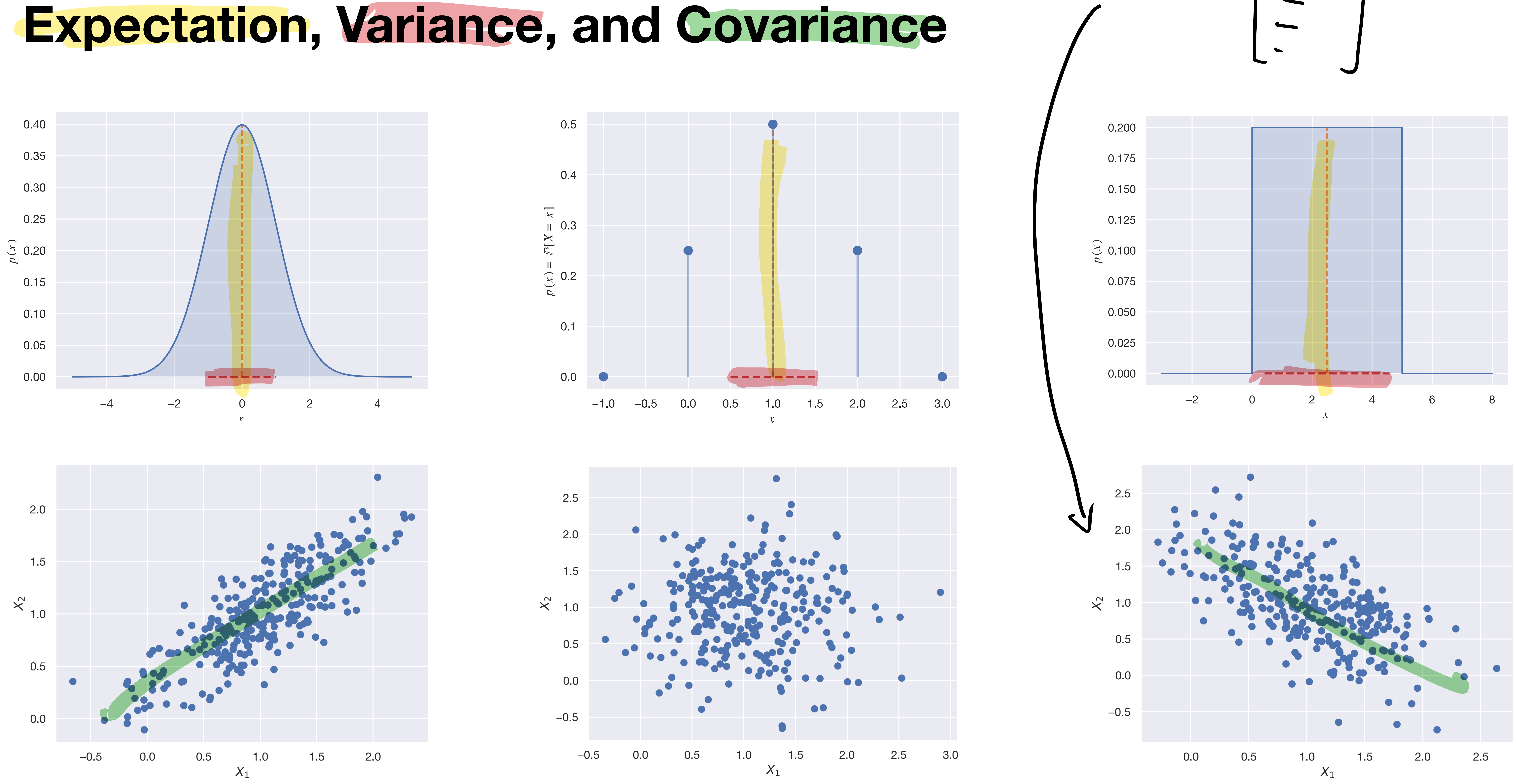
Covariance follows the “positive definiteness” property:

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \geq 0.$$

Summary Statistics

Expectation, Variance, and Covariance





Random Vectors

Multivariate Random Variables

Random Vectors

Definition

$f: \mathbb{R} \to \mathbb{R}$

$f: \mathbb{R}^d \to \mathbb{R}^n$

So far, we have only been talking about single-variable distributions.

We can talk about multivariable distributions by considering ***random vectors***:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Random Vectors

Expectation

The ***expectation*** of a random vector just comes from taking the entry-wise expectation:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

Random Vectors

Covariance Matrix

$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

The variance of a random vector generalizes to the ***covariance matrix***

In the $d = 2$ case,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}.$$

What do you notice about this matrix?

Random Vectors

Covariance Matrix

The variance of a random vector generalizes to the ***covariance matrix***

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}$$

In general, $\mathbf{\Sigma}_{i,j} = \mathrm{Cov}(X_i, X_j)$.

Random Vectors

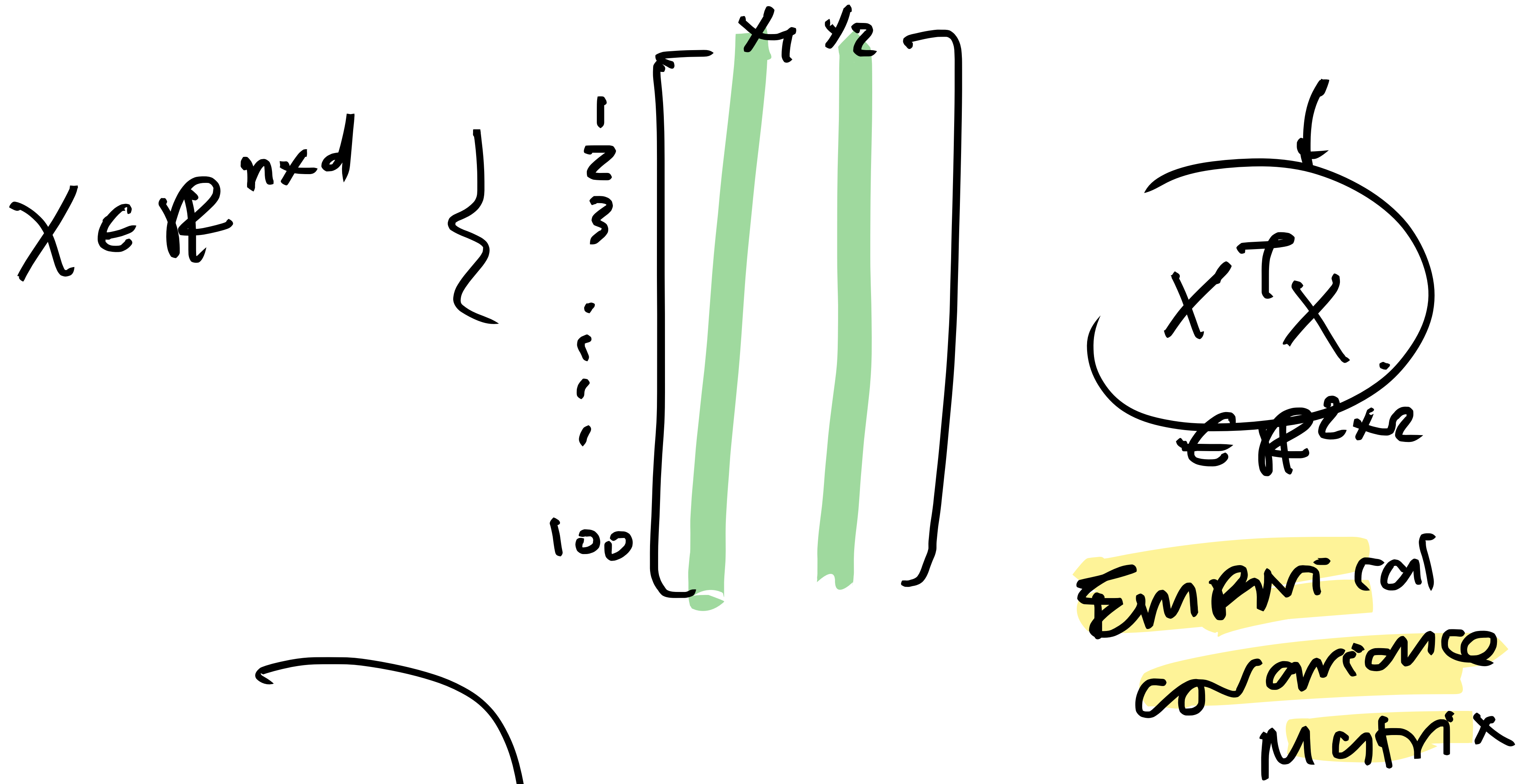
Covariance Matrix

The covariance matrix is **symmetric.**

$$\mathbf{\Sigma} = \mathbf{\Sigma}^\top.$$

The covariance matrix is also **positive semidefinite.**

$$\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

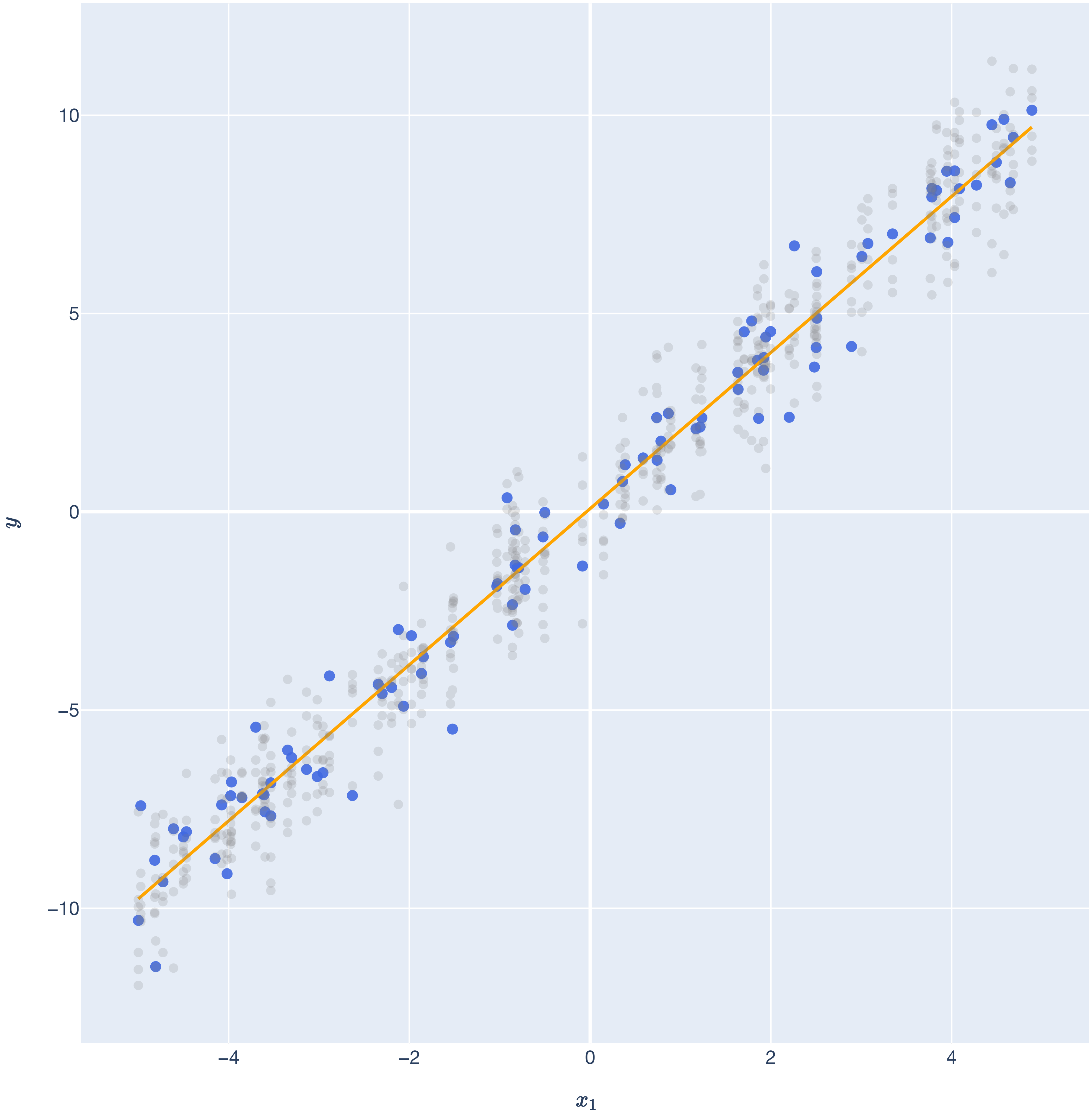


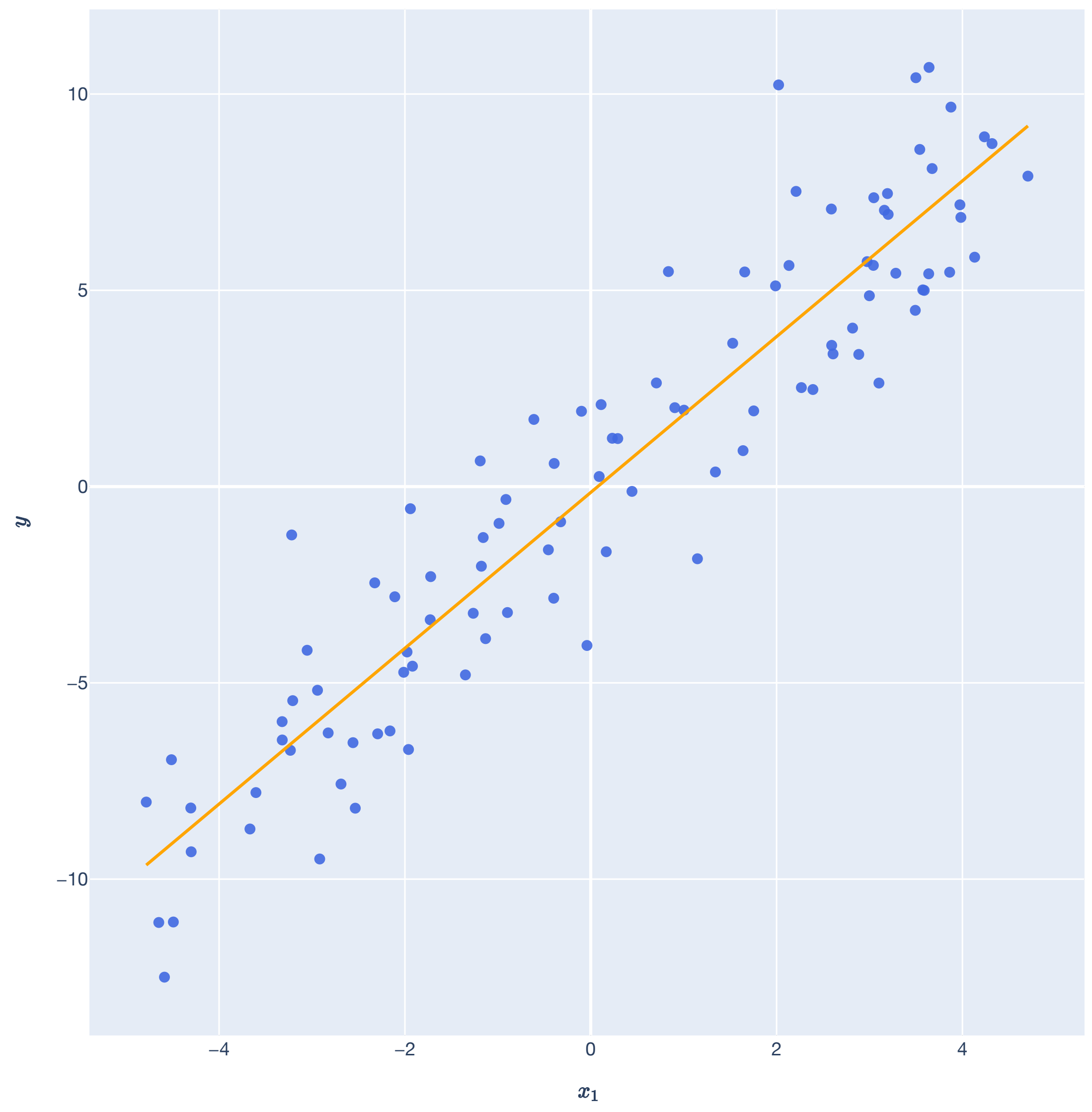
Data as random

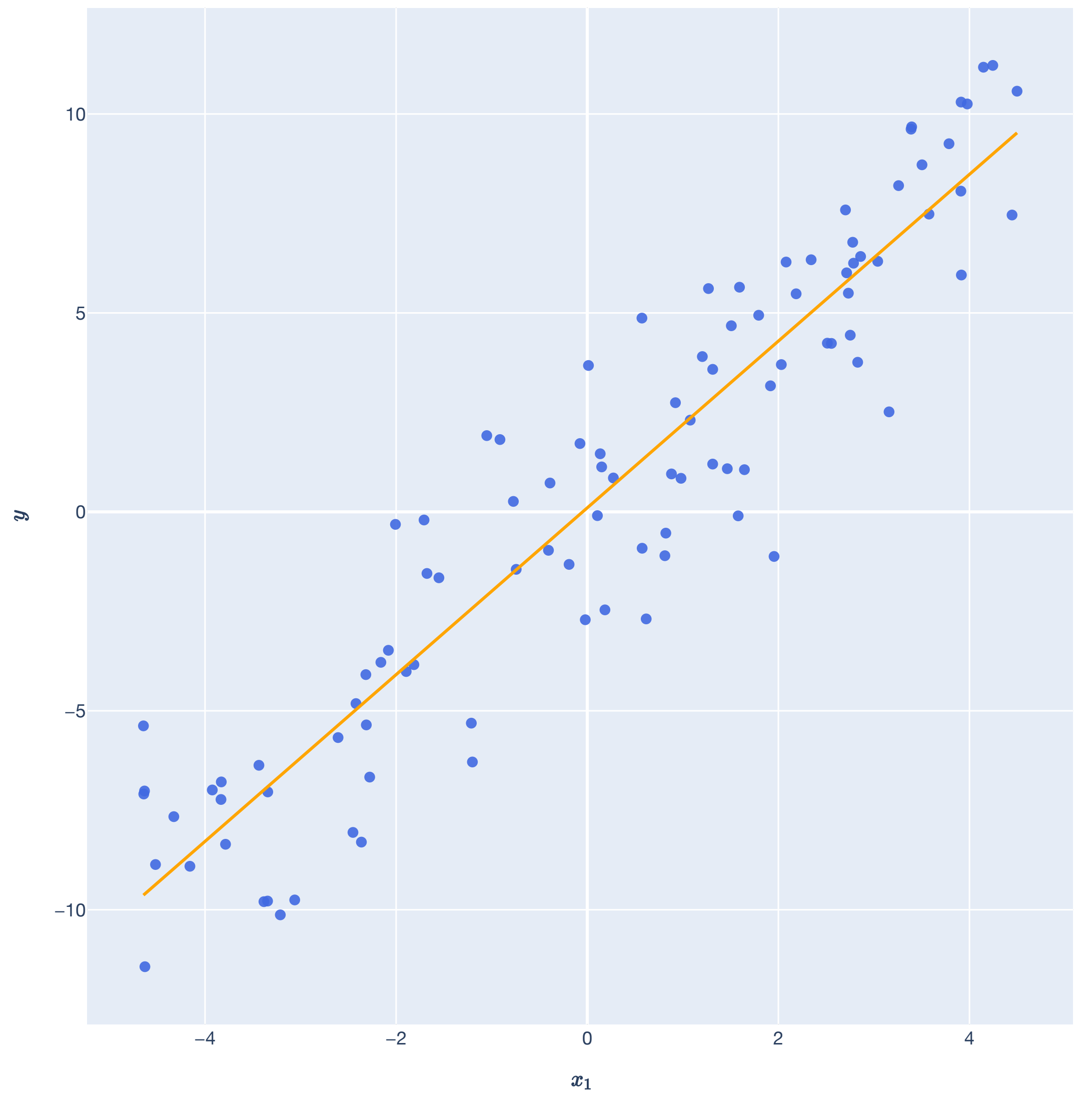
Modeling regression with probability

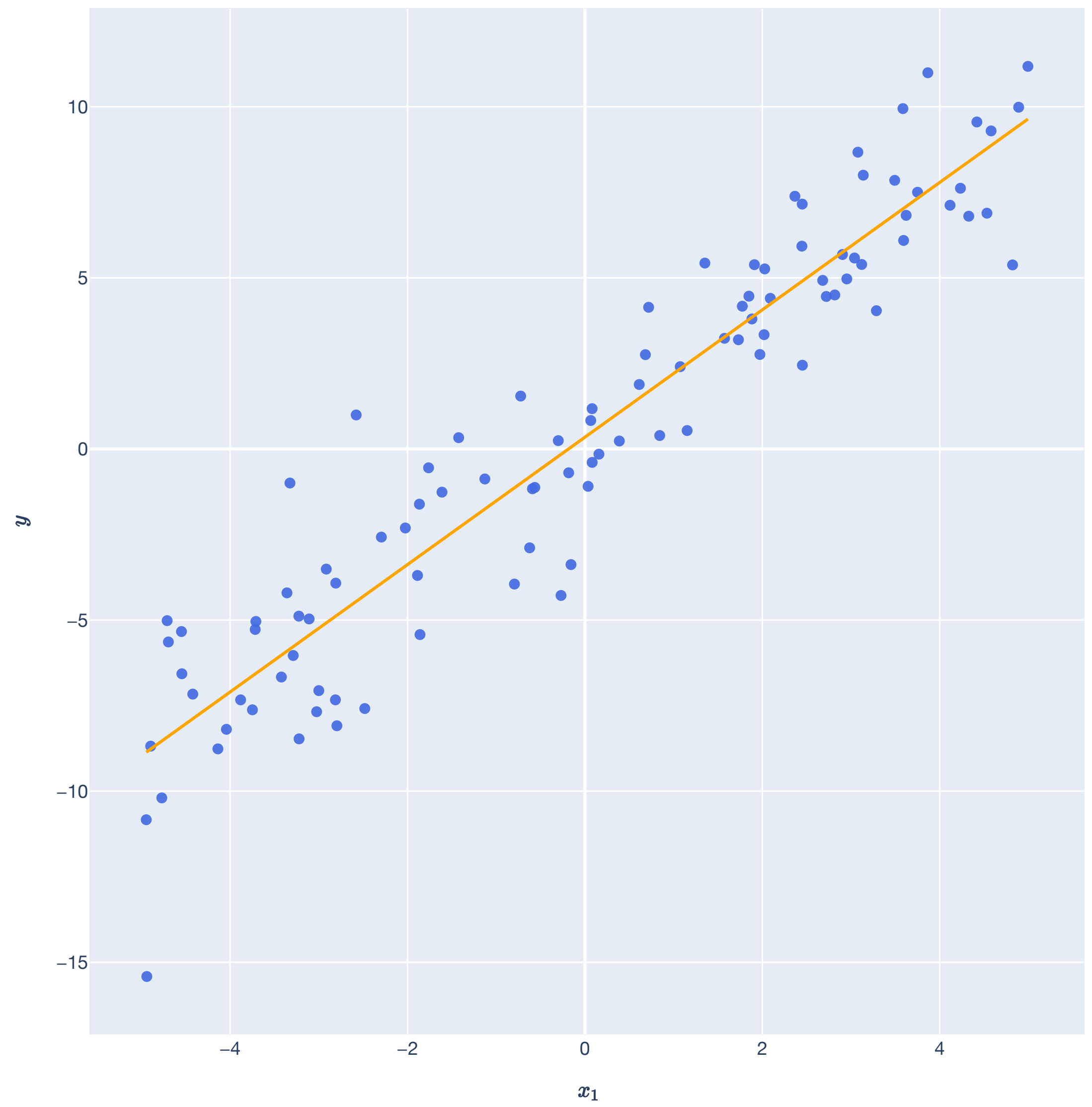
Regression

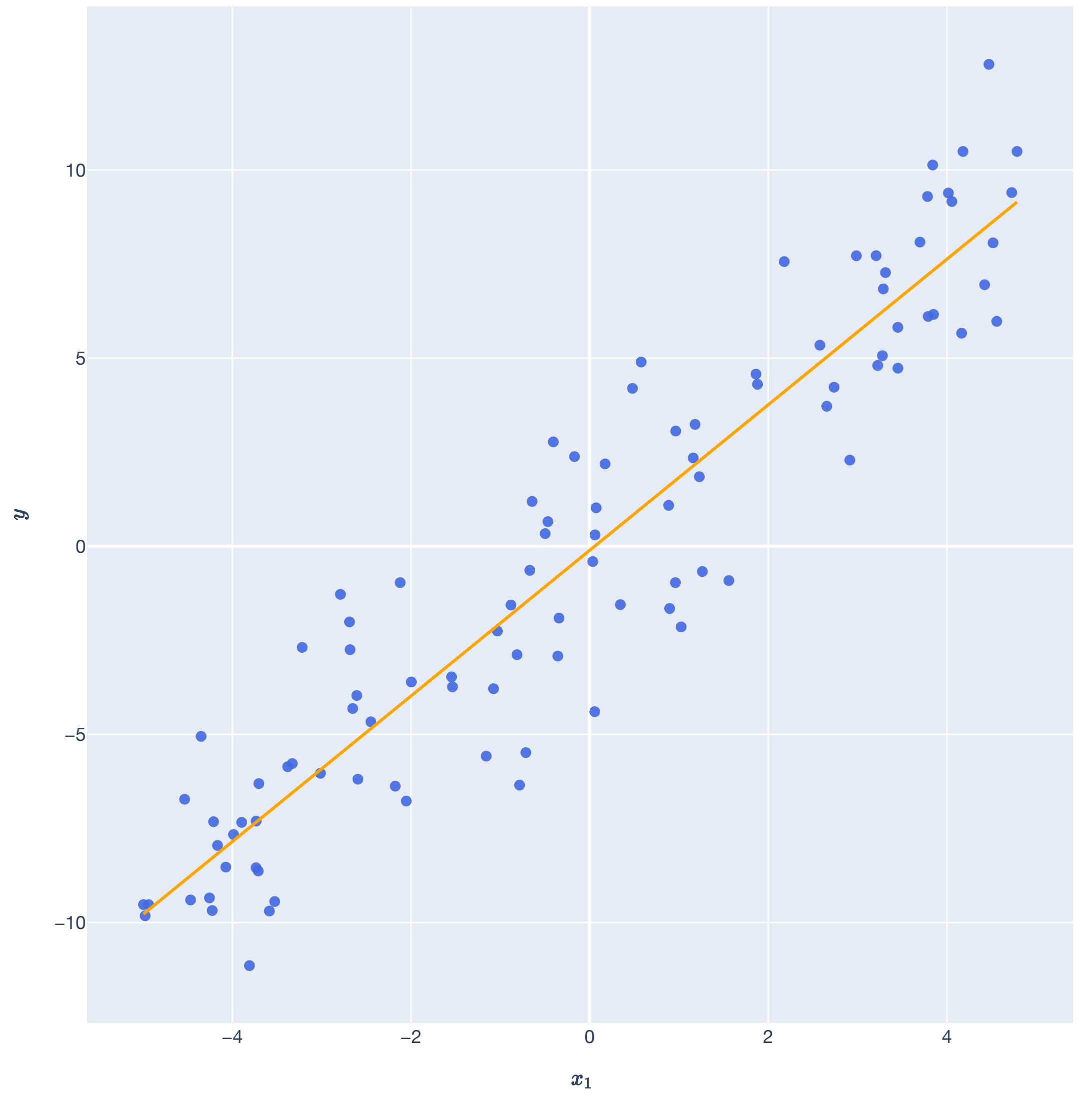
Modeling randomness











Regression

Setup (Review)

<u>Observed:</u> Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

<u>Unknown:</u> *Weight vector* $\mathbf{w} \in \mathbb{R}^d$ with weights $w_1, \ldots, w_d$.

<u>Goal:</u> For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

Regression

Setup (Review)

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Regression

Setup (Review)

Original goal:

Given a new, unseen $(\mathbf{x}_0, y_0) \in \mathbb{R}^d \times \mathbb{R}$, we wanted to *generalize*:

$$\hat{\mathbf{w}}^\top \mathbf{x}_0 \approx y_0.$$

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

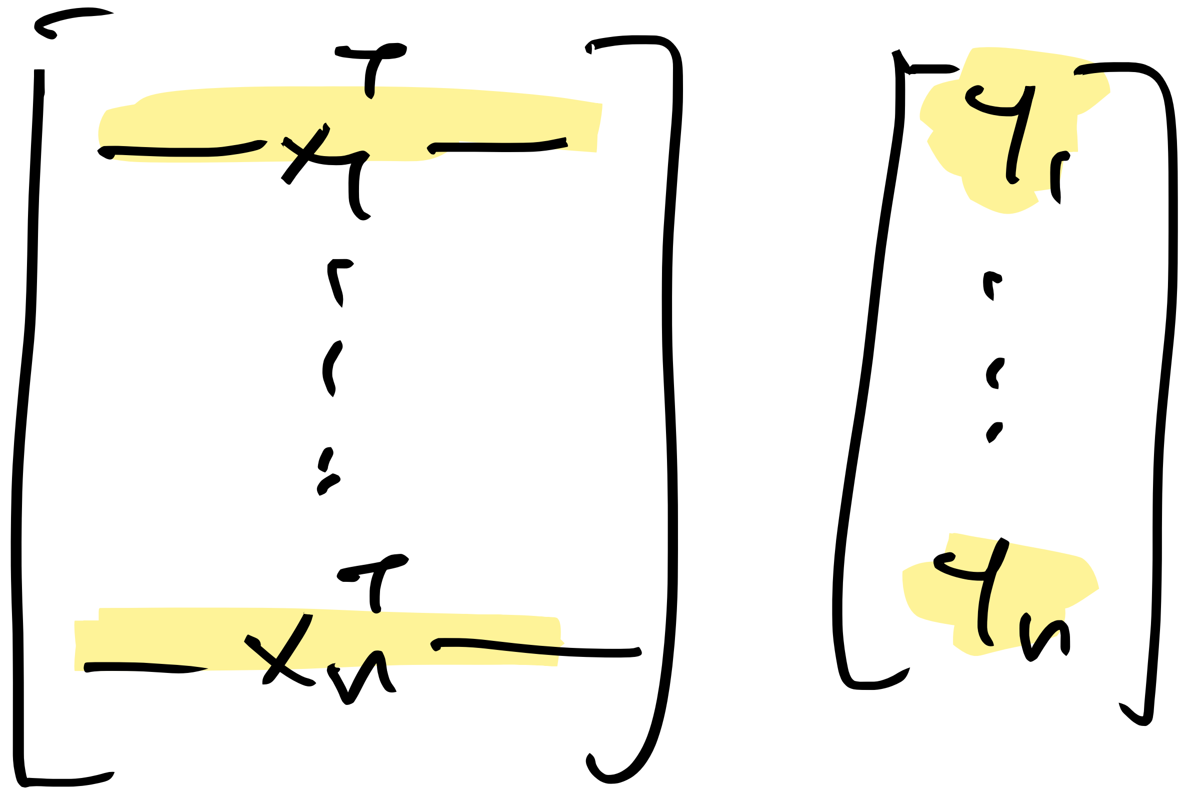
$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min}\ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Regression

Setup (Review)

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$



Least squares expanded is just:

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

Put a $1/n$ there, and it looks like we're minimizing an average…

Regression

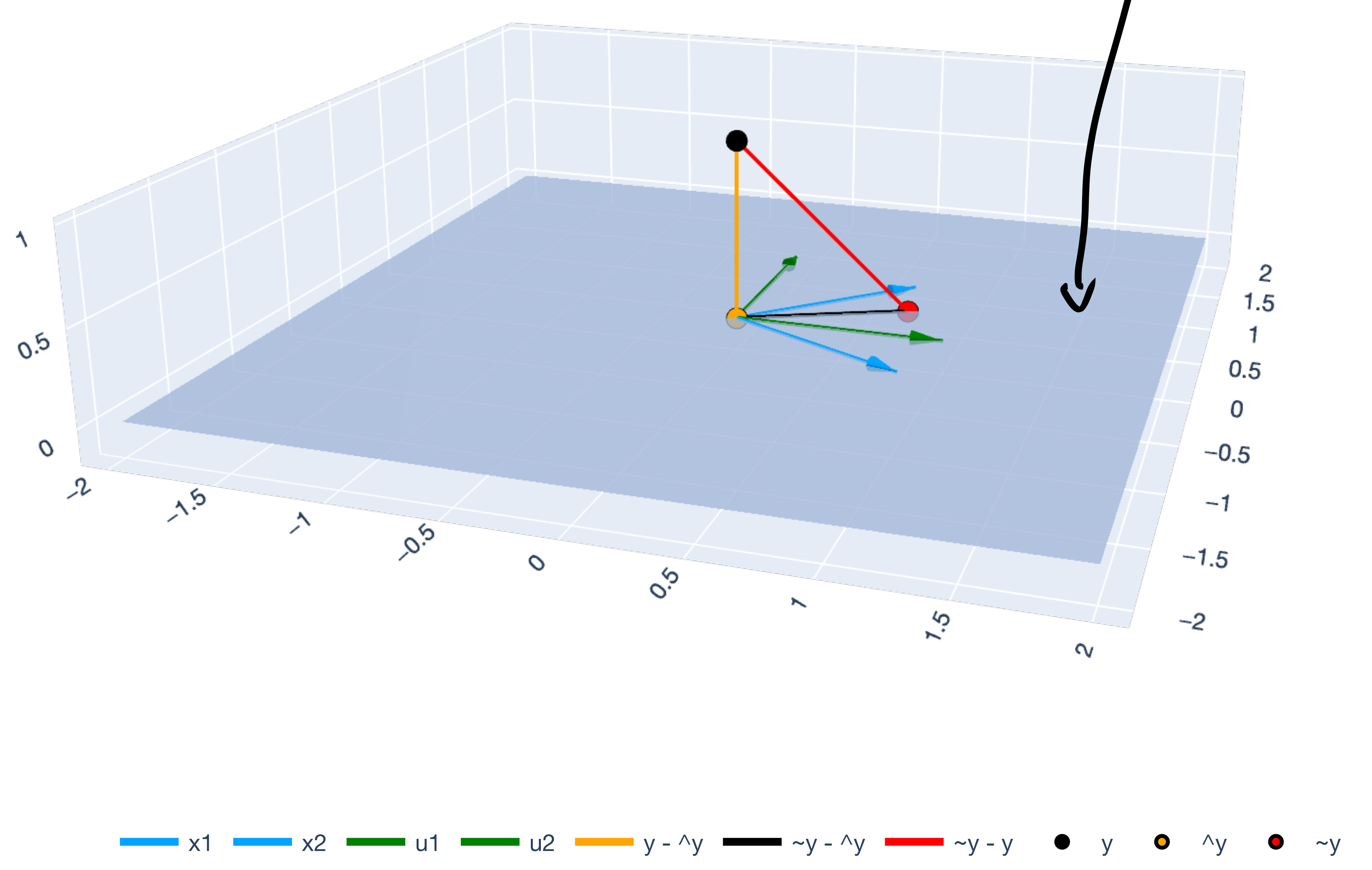
A note on $\hat{\mathbf{w}}$

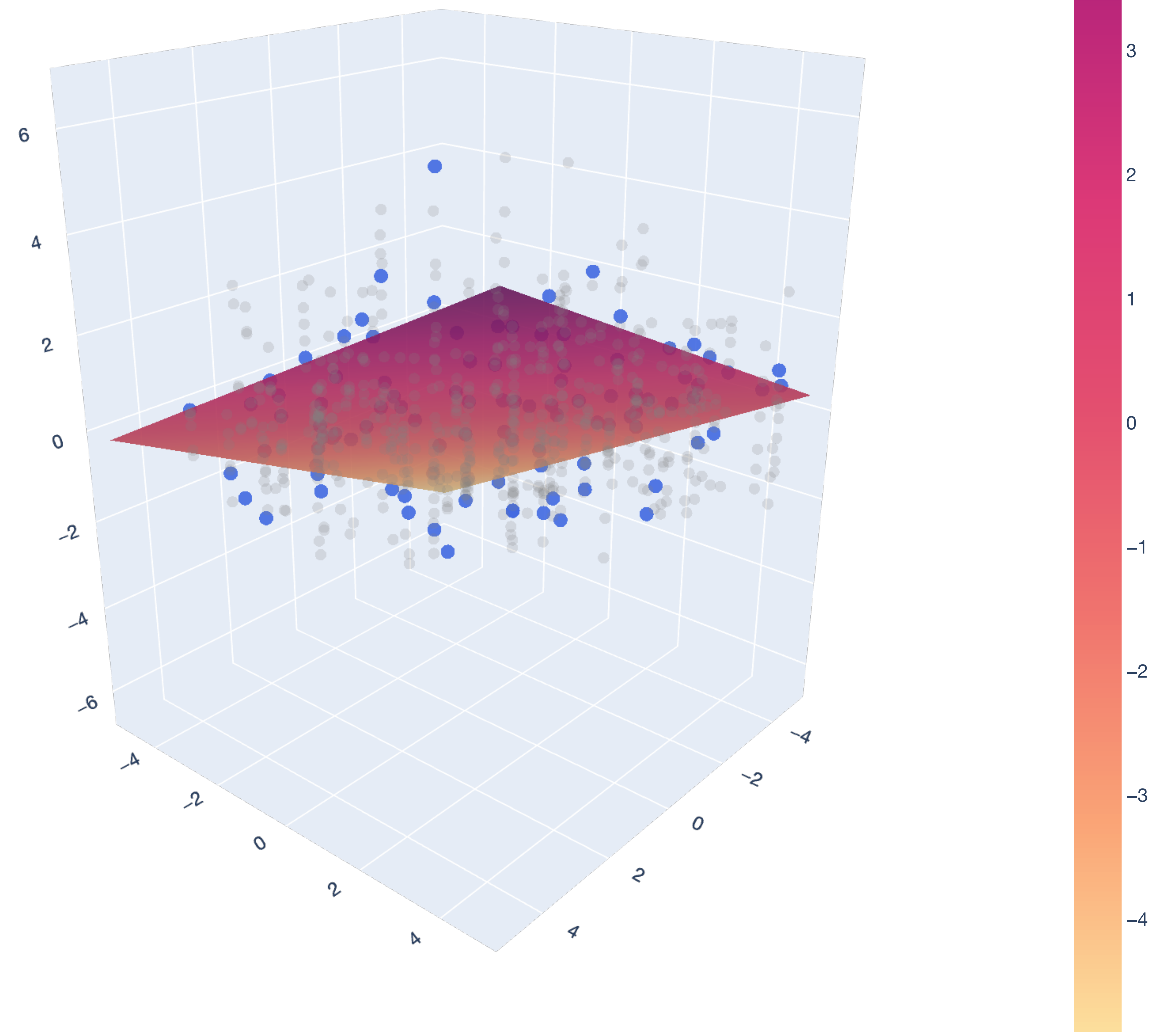
w_1, w_2

$\hat{w} \in \mathbb{R}^2$

col(X)

$w^T x = 0$





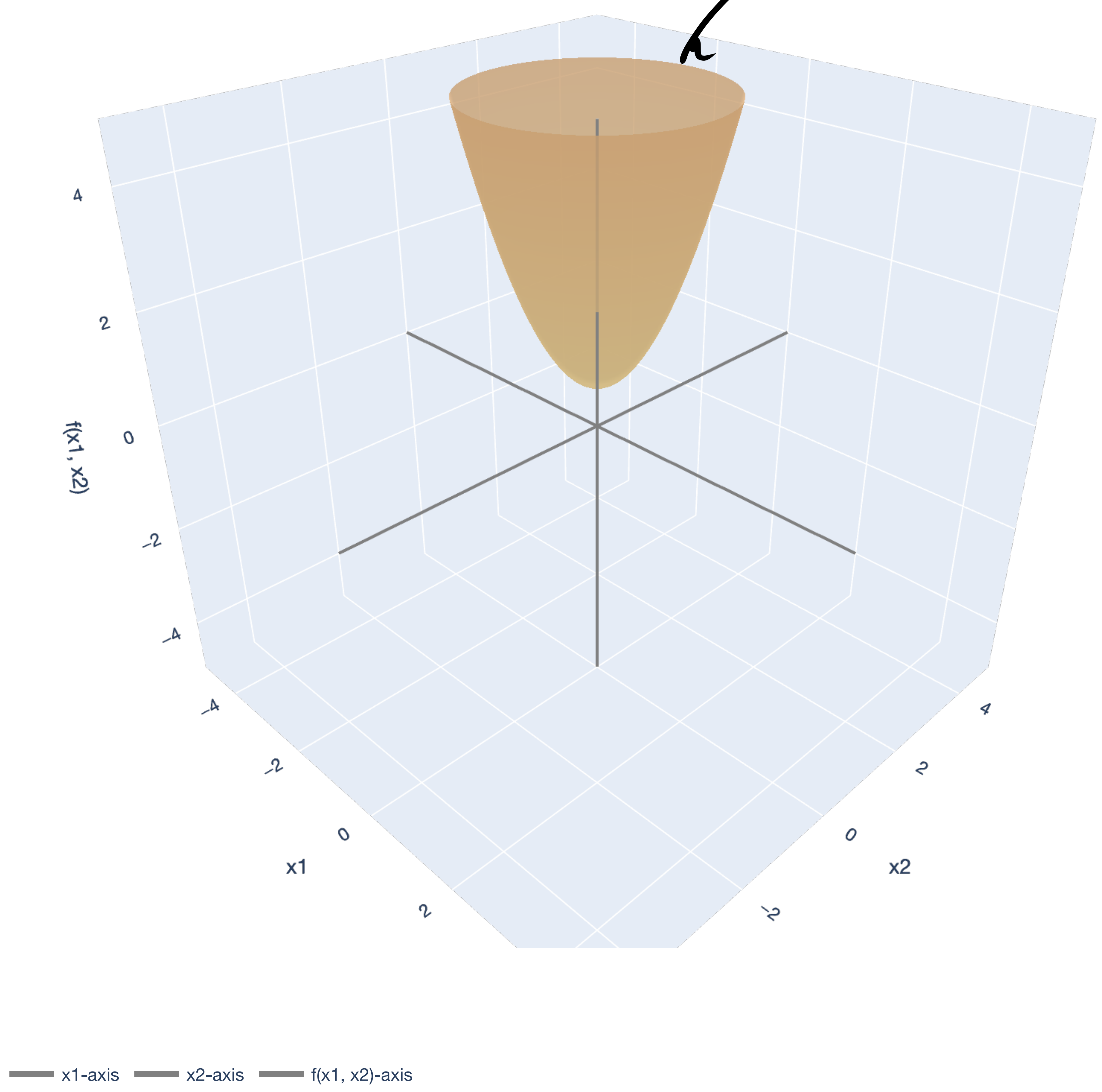
Regression

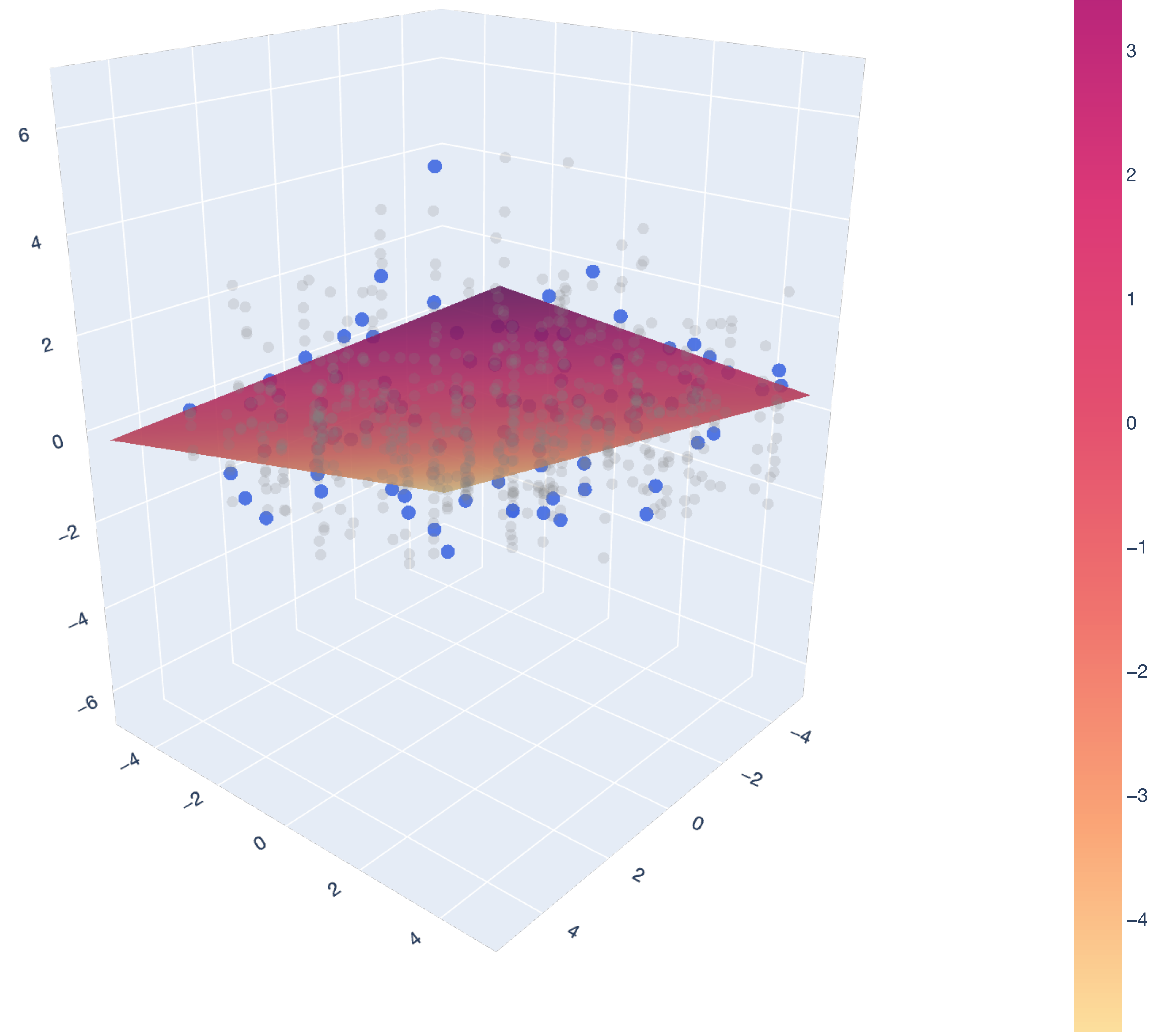
A note on $\hat{\mathbf{w}}$

$\hat{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

$\|Xw - y\|^2$

$w^T x = 0$

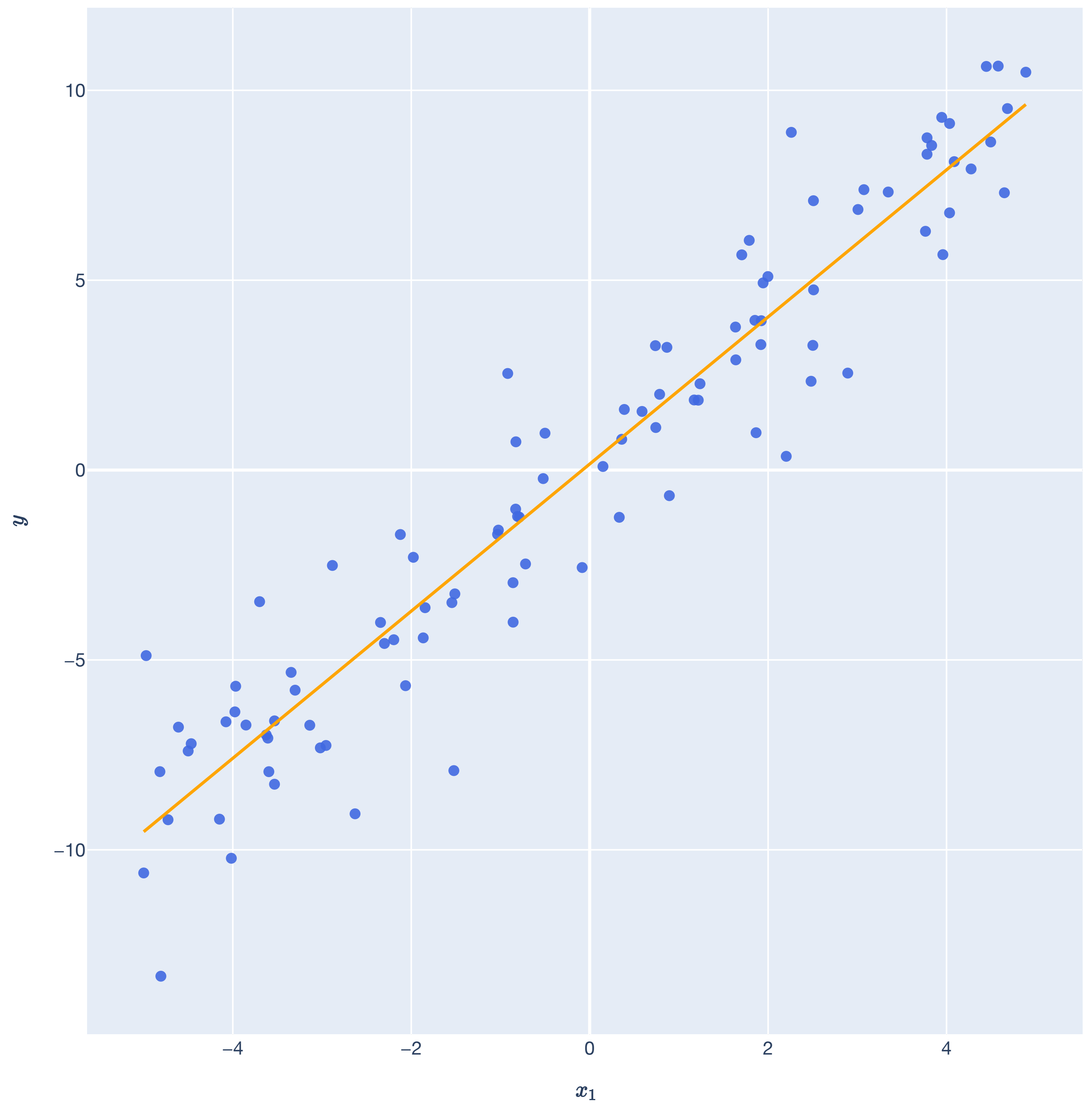


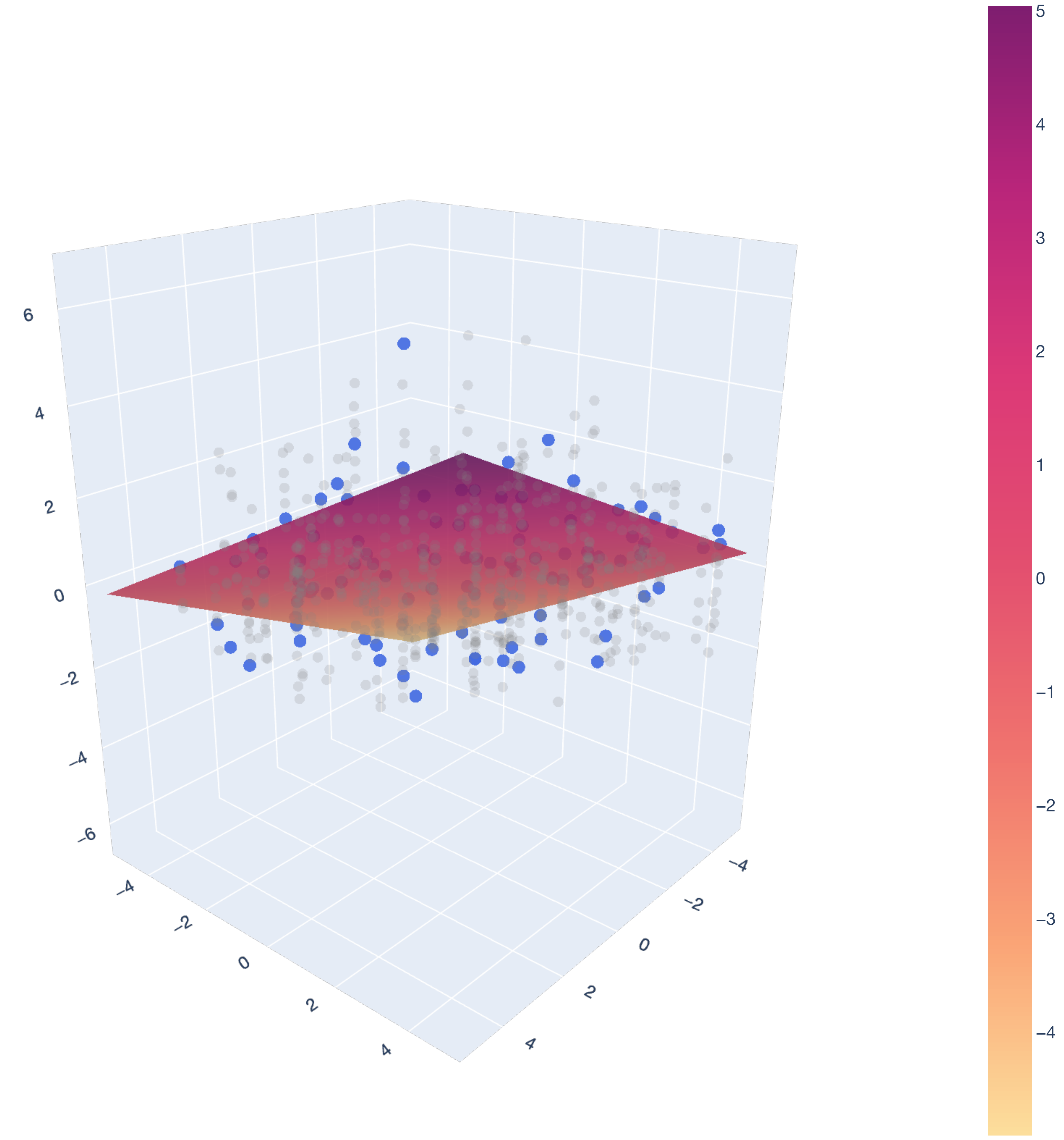


Regression

A note on $\hat{\mathbf{w}}$

$wx_1 = 0$





Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

Regression with randomness

Training examples

Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

Each entry is a random variable, think of $\mathbf{x}_i^\top \in \mathbb{R}^d$ as a d-dimensional ***random vector***.

Each label is a random variable, think of $y_i \in \mathbb{R}$ as a ***random variable***.

Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a ***joint distribution***, $\mathbb{P}_{\mathbf{x},y}$

Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

$|y - f(x)|$

Regression with randomness

Model of error (ASSUMPTION)

Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

$$y_i = f^*(\mathbf{x}_i) + \epsilon_i$$

Some *deterministic* function $f^* : \mathbb{R}^d \to \mathbb{R}$ explains as much as it can

Some *randomness* ϵ_i models the unexplained relationship, where we assume

$\mathbb{E}[\epsilon_i] = 0$ and ϵ_i is independent of $\mathbf{x}_i$.

Regression with randomness

Model of error

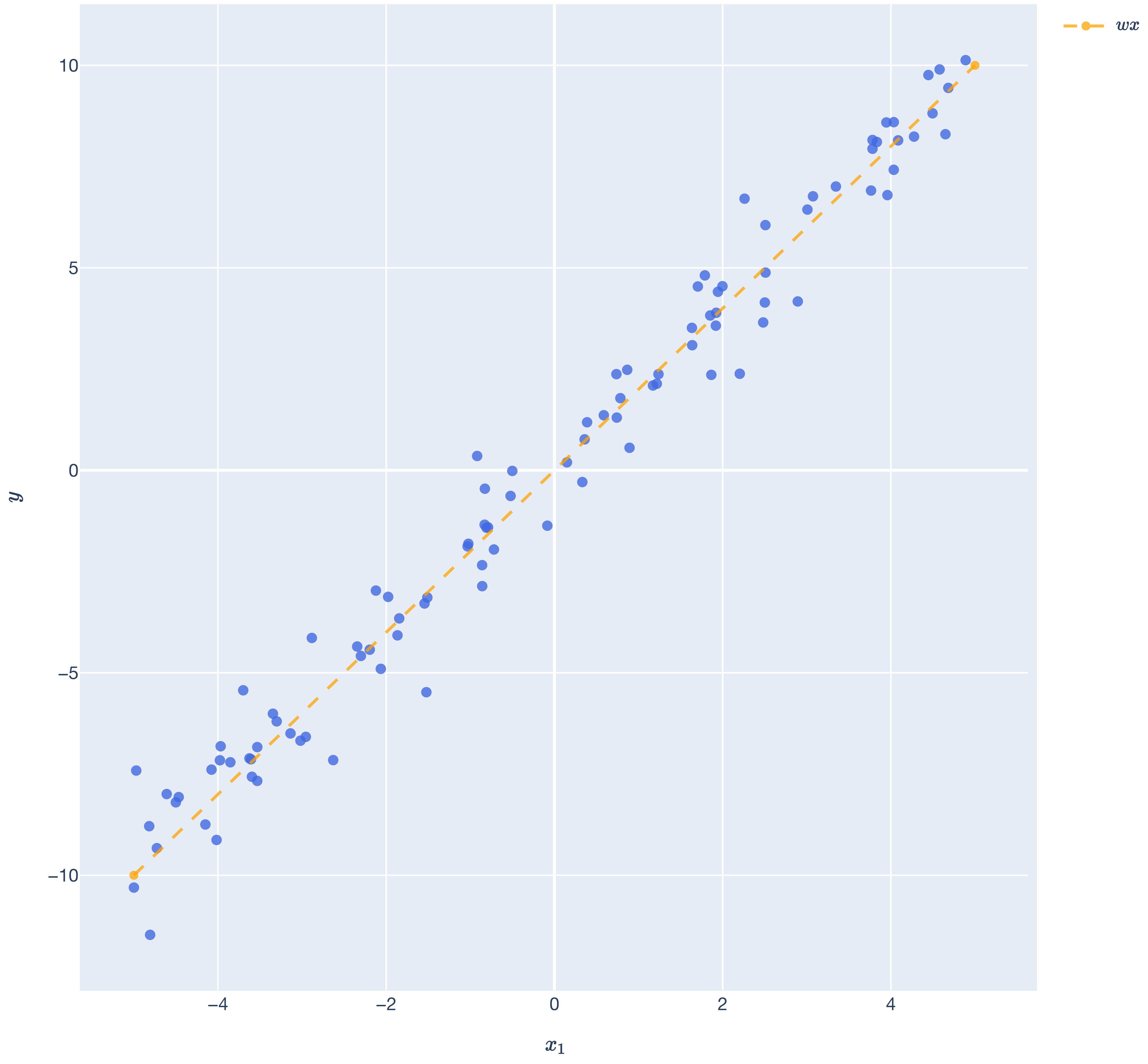
Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution,* $\mathbb{P}_{\mathbf{x},y}$

$$y_i = f^*(\mathbf{x}_i) + \epsilon_i$$

Some *deterministic* function $f^* : \mathbb{R}^d \to \mathbb{R}$ explains as much as it can

Some *randomness* ϵ_i models the unexplained relationship, where we assume

$$\mathbb{E}[\epsilon_i] = 0 \text{ and } \epsilon_i \text{ is independent of } \mathbf{x}_i.$$



Regression with randomness

Model of error

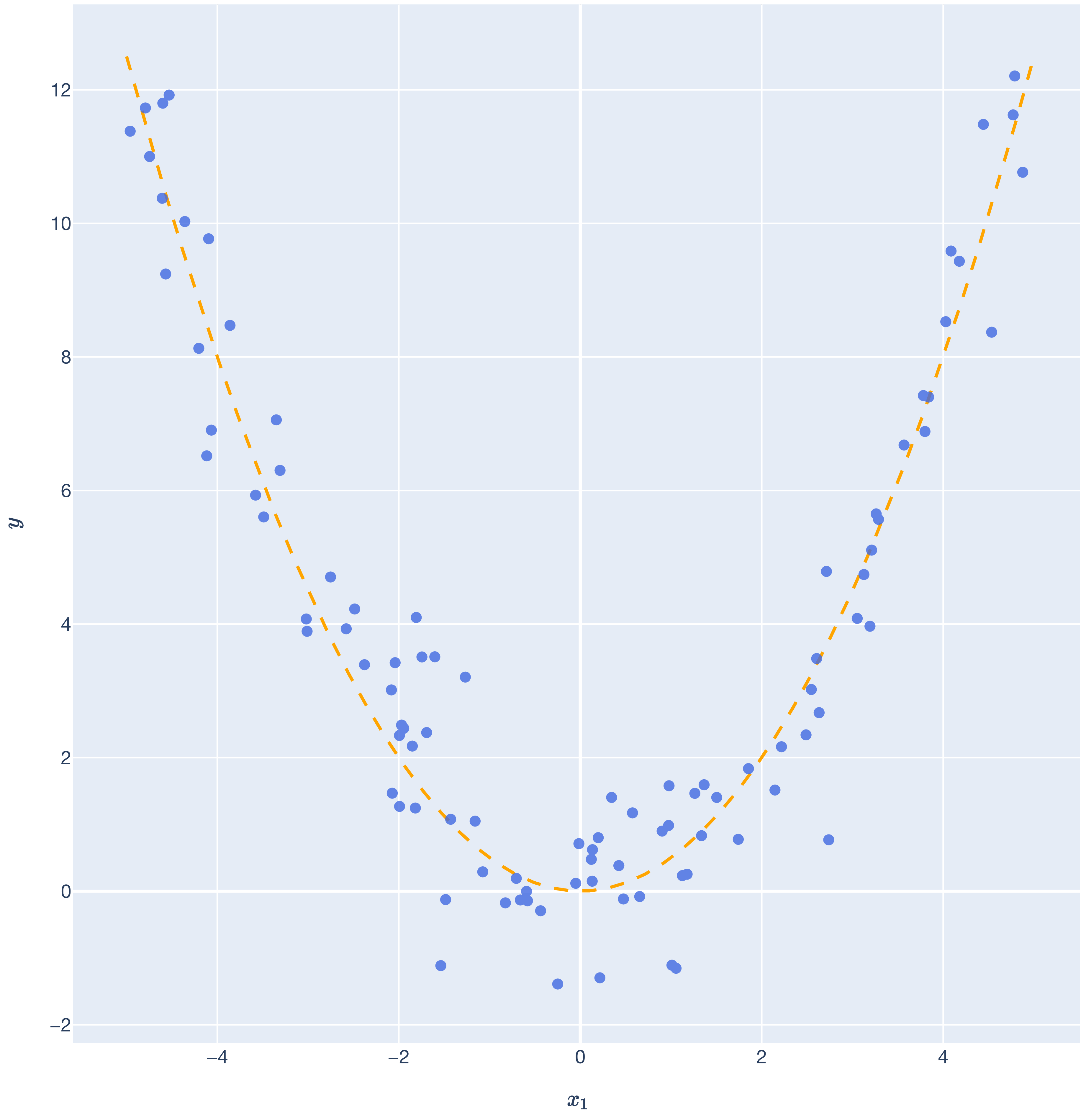
Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution,* $\mathbb{P}_{\mathbf{x},y}$

$$y_i = f^*(\mathbf{x}_i) + \epsilon_i$$

Some *deterministic* function $f^* : \mathbb{R}^d \to \mathbb{R}$ explains as much as it can

Some *randomness* ϵ_i models the unexplained relationship, where we assume

$$\mathbb{E}[\epsilon_i] = 0 \text{ and } \epsilon_i \text{ is independent of } \mathbf{x}_i.$$



Regression with randomness

Model of error

Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution,* $\mathbb{P}_{\mathbf{x},y}$

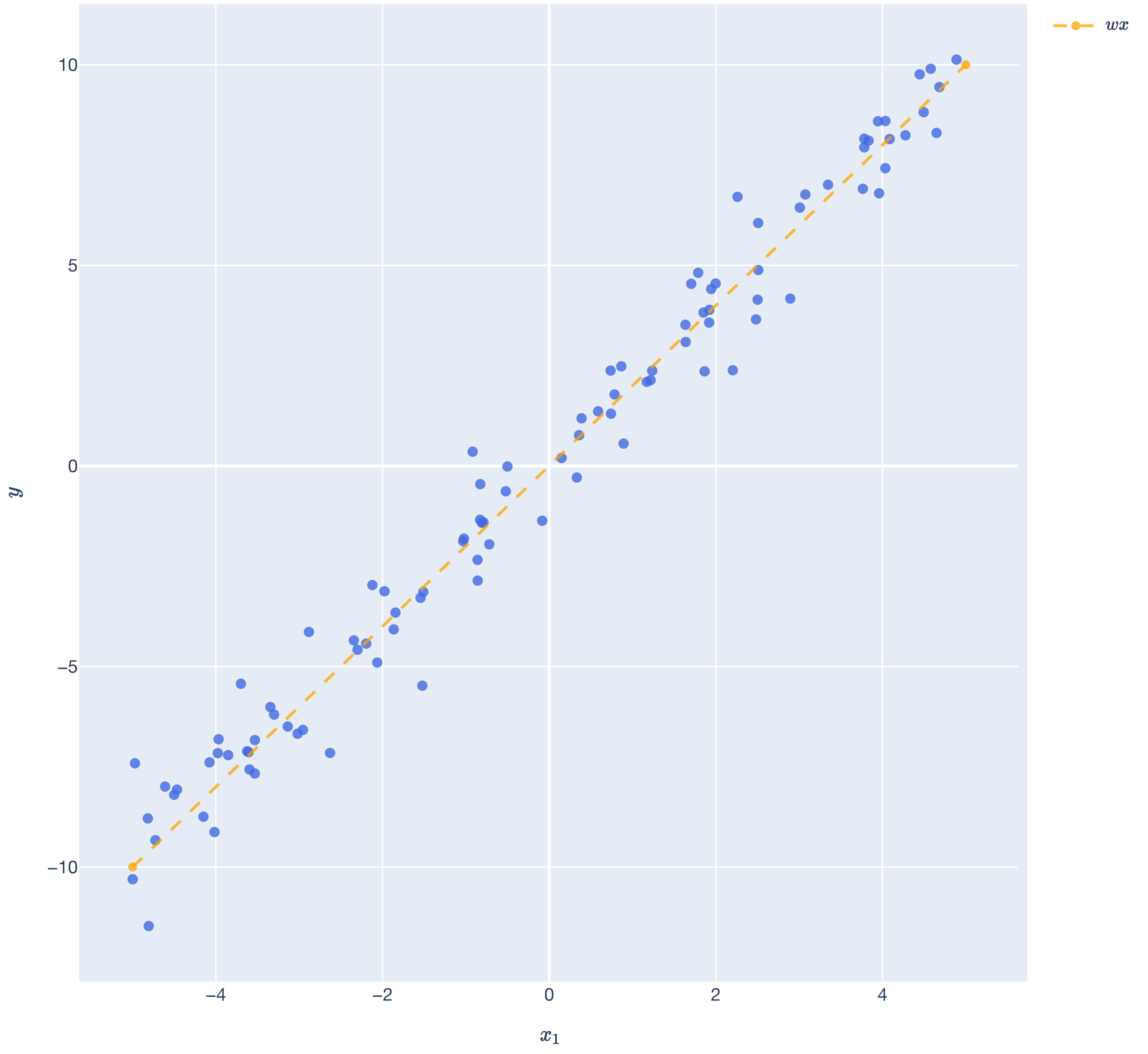
$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$$

Deterministic linear function

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}^*$$

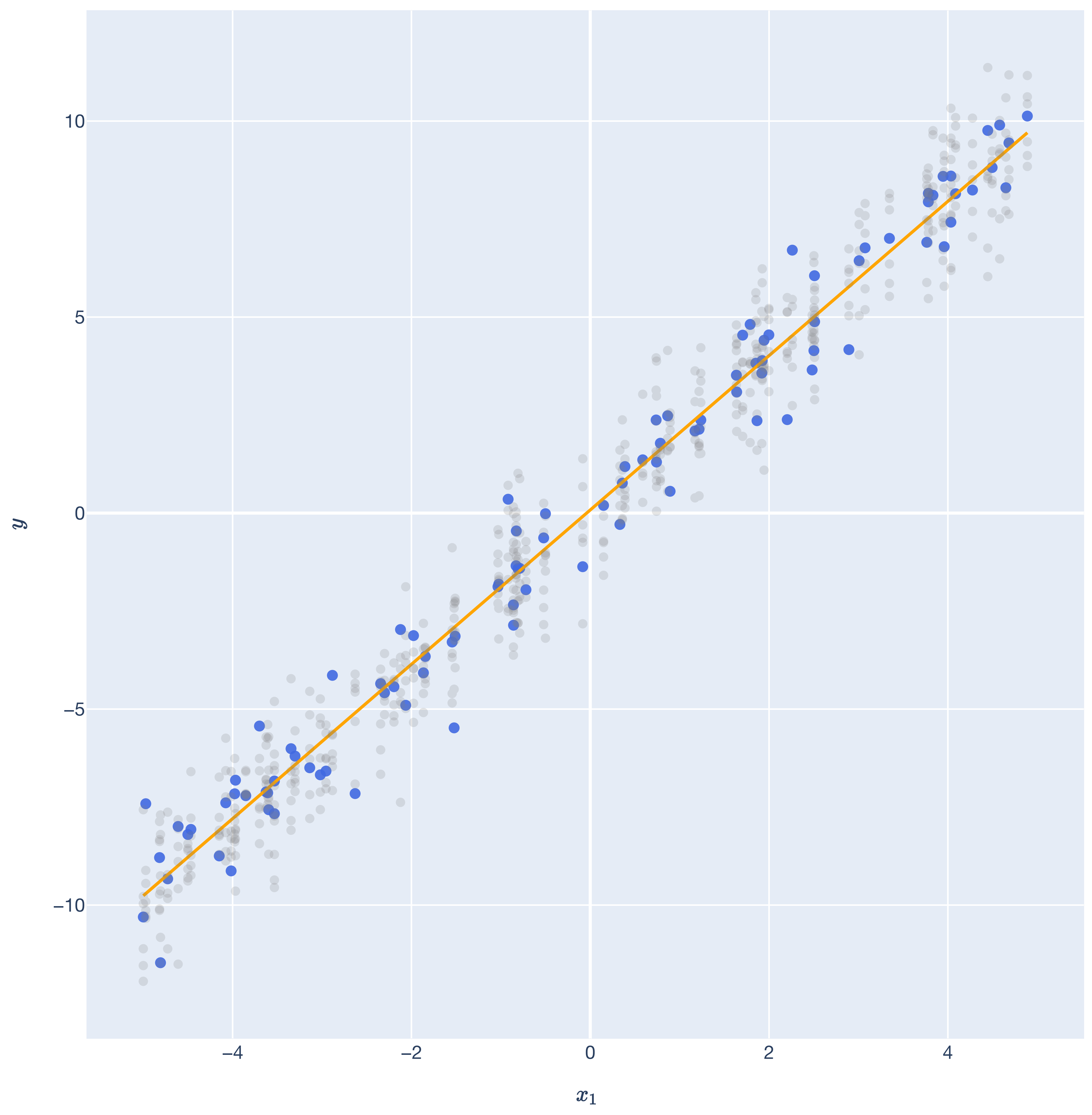
Some *randomness* ϵ_i models the unexplained relationship, where we assume

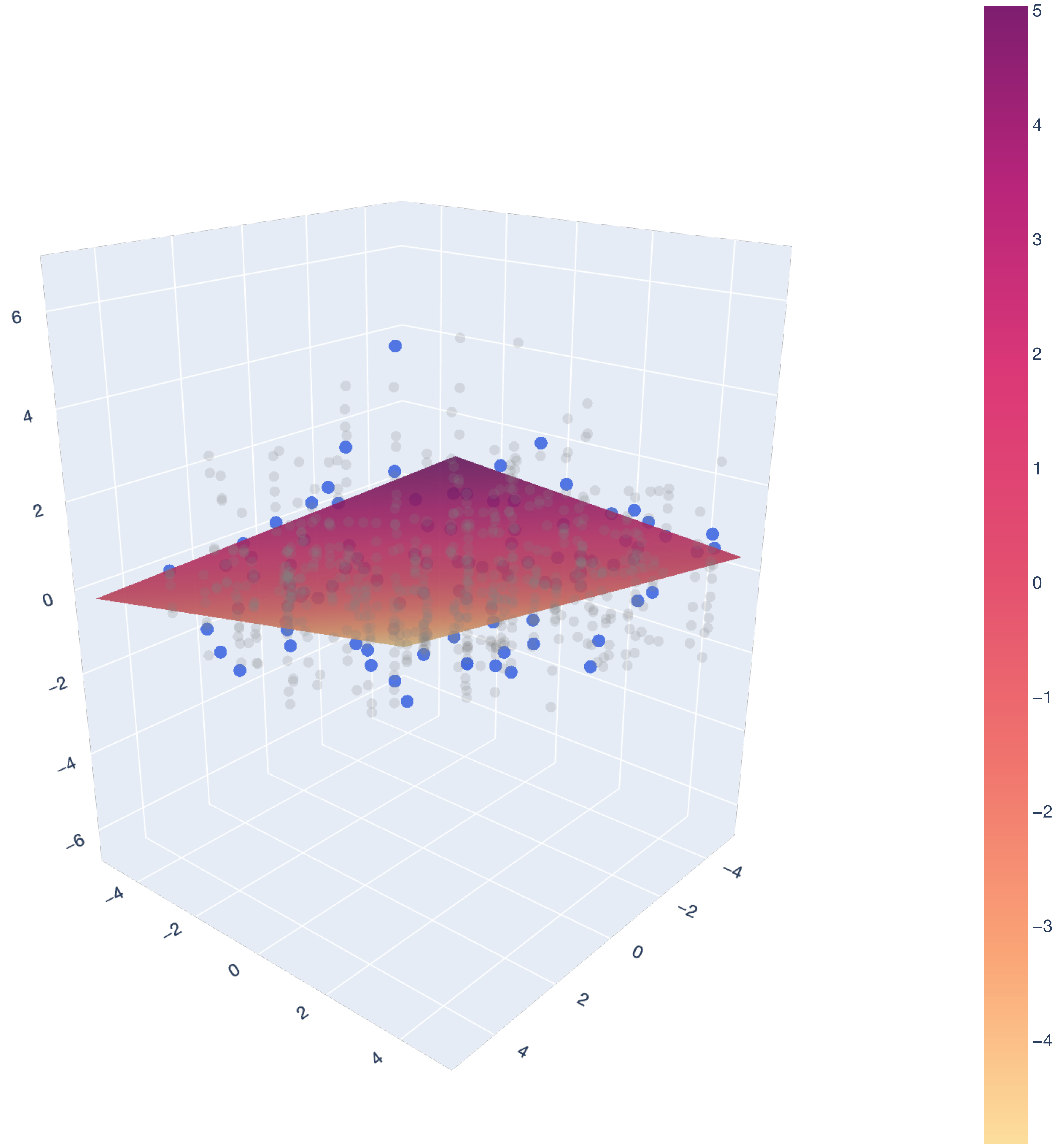
$$\mathbb{E}[\epsilon_i] = 0 \text{ and } \epsilon_i \text{ is independent of } \mathbf{x}_i.$$



Regression with randomness

Model of error





Regression with randomness

Goal, with randomness

$\hat{w} = (X^\top X)^{-1} X^\top y$

Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution,* $\mathbb{P}_{\mathbf{x},y}$

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i, \text{ where } \mathbb{E}[\epsilon_i] = 0 \text{ and } \epsilon_i \text{ is independent of } \mathbf{x}_i.$$

This gives us $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$, so we can also write:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon, \text{ where } \epsilon \in \mathbb{R}^n \text{ is a random vector.}$$

Regression with randomness

Goal, with randomness

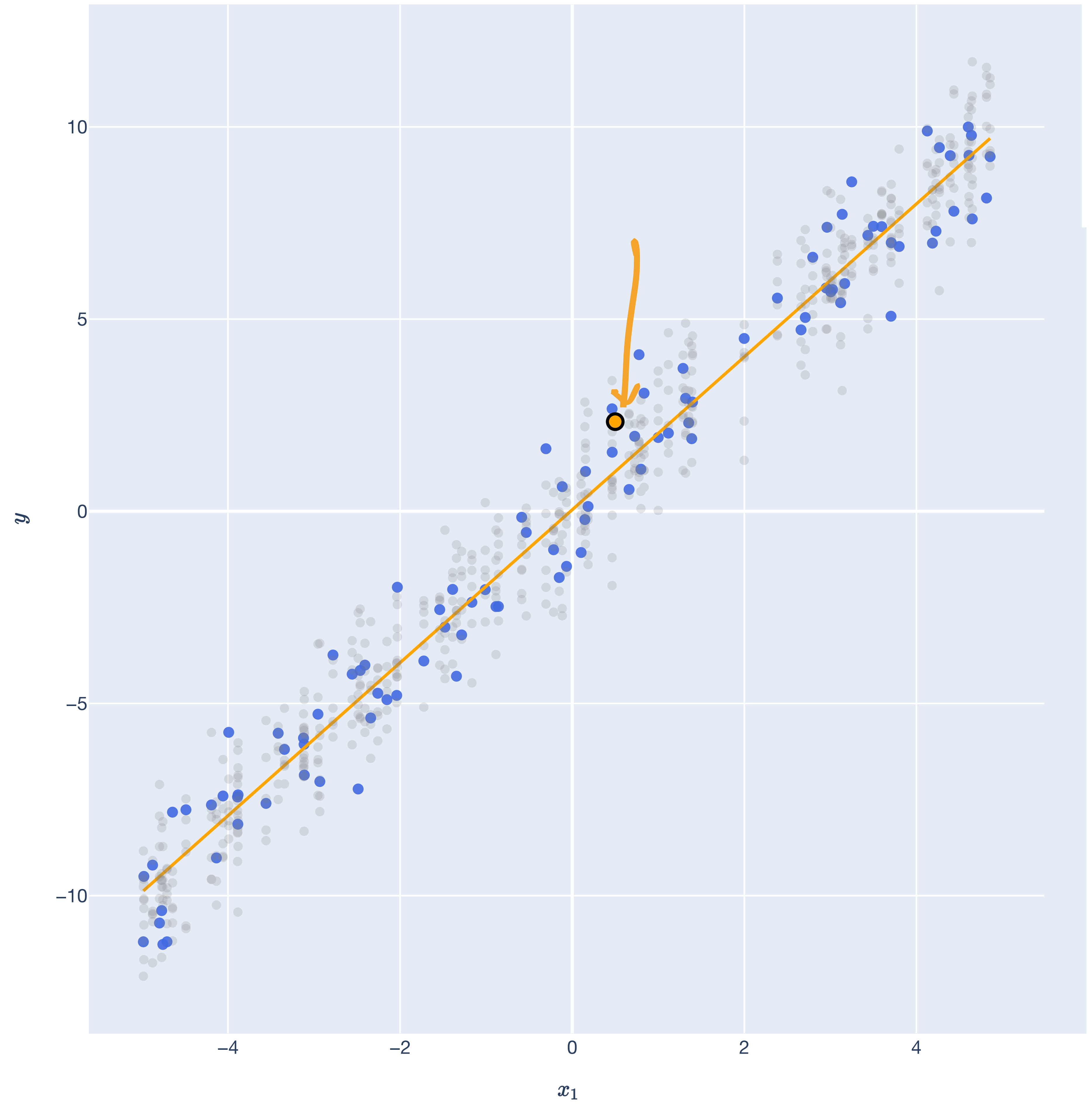
Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

We can draw a new $(\mathbf{x}_0, y_0)$ from the distribution $\mathbb{P}_{\mathbf{x},y}$.

We want to find a ***model*** $f : \mathbb{R}^d \to \mathbb{R}$ for predicting on this new example.

Notion of "badness" is **squared loss**

$$\ell(f(\mathbf{x}_0), y_0) := (y_0 - f(\mathbf{x}_0))^2.$$



Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$$\mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

Regression with randomness

Goal, with randomness

Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution,* $\mathbb{P}_{\mathbf{x},y}$

We can draw a new $(\mathbf{x}_0, y_0)$ from the distribution $\mathbb{P}_{\mathbf{x},y}$.

We want to find a linear function $f : \mathbb{R}^d \to \mathbb{R}$ for predicting on this new example:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

Notion of "badness" is **squared loss:**

$$\ell(f(\mathbf{x}_0), y_0) := (y_0 - f(\mathbf{x}_0))^2.$$

To make a decision, we care about the *expected loss* (***risk***):

$$R(f) := \mathbb{E}_{(\mathbf{x}_0, y_0)}[\underbrace{(y_0 - f(\mathbf{x}_0))^2}_{\text{RV.}}]$$

Regression

Goal, with randomness

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where ϵ is a *random variable* with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, with ϵ is independent of $\mathbf{x}$.

Draw n examples: *random matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and random vector $y \in \mathbb{R}^n$.

Ultimate goal: Find $f(\mathbf{x}) := \hat{\mathbf{w}}^\top \mathbf{x}$ that *generalizes* on a new $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$:

$$R(\hat{f}) := \mathbb{E}_{\mathbf{x}_0, y_0}[(\hat{f}(\mathbf{x}_0) - y_0)^2]$$

Intermediary goal: Find $f(\mathbf{x}) := \hat{\mathbf{w}}^\top \mathbf{x}$ that does well on the training samples:

$$\hat{R}(\hat{f}) := \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(\mathbf{x}_i) - y_i)^2.$$

Regression

Goal, with randomness

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where ϵ is a *random variable* with $\mathbb{E}[\epsilon] = 0$ and ϵ is independent of $\mathbf{x}$.

Draw n examples: *random matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and random vector $y \in \mathbb{R}^n$.

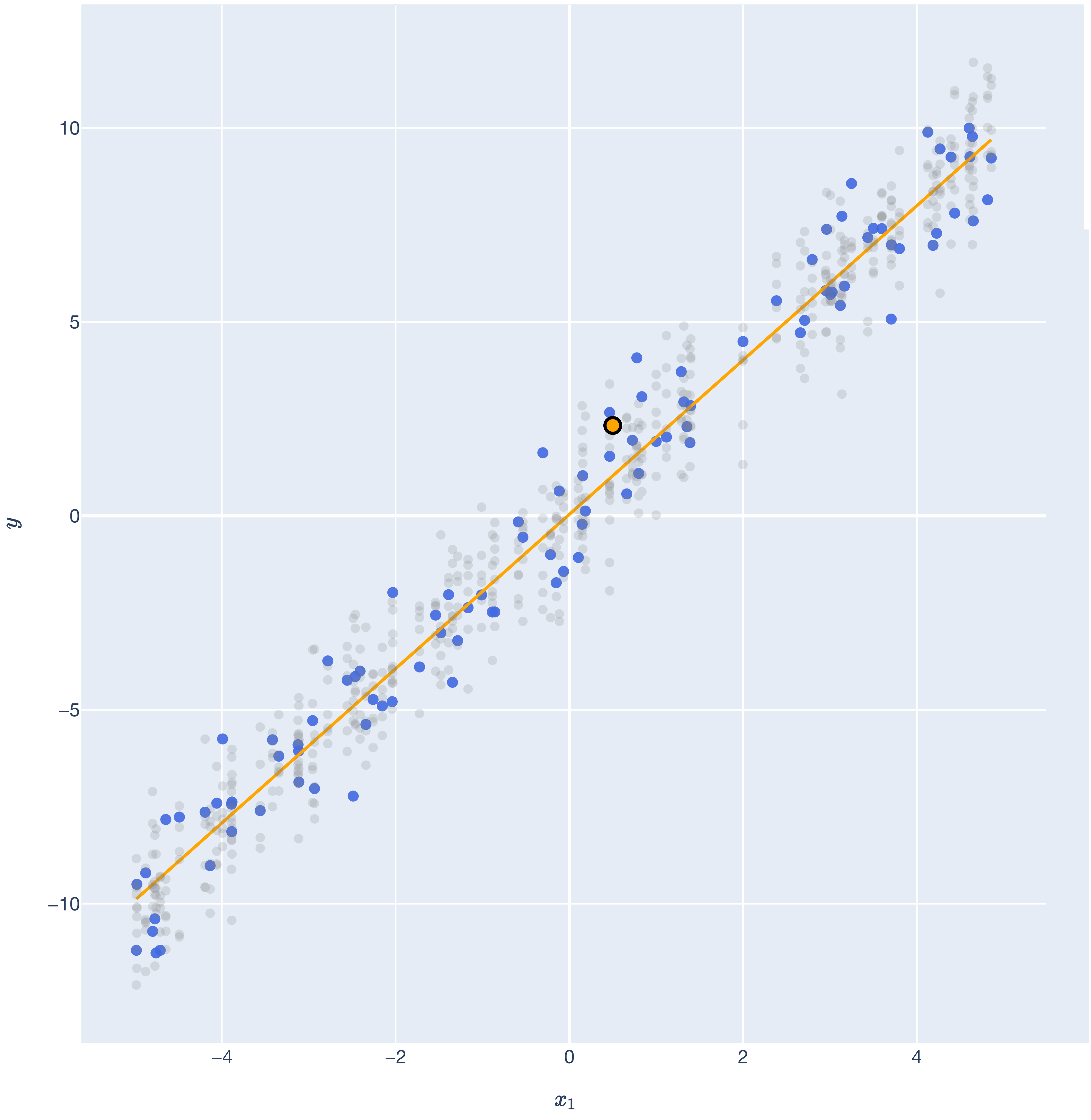
Ultimate goal: Find $f(\mathbf{x}) := \hat{\mathbf{w}}^\top \mathbf{x}$ that *generalizes* on a new $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$:

$$R(\hat{f}) := \mathbb{E}_{\mathbf{x}_0,y_0}[(\hat{f}(\mathbf{x}_0) - y_0)^2]$$

Intermediary goal: Find $f(\mathbf{x}) := \hat{\mathbf{w}}^\top \mathbf{x}$ that does well on the training samples, minimizing ***empirical risk***:

$$\hat{R}(\hat{f}) := \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(\mathbf{x}_i) - y_i)^2 = \frac{1}{n} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

This is what we've been doing all along!



Regression with randomness

Setup

Each row $\mathbf{x}_i^\top \in \mathbb{R}^d$ for $i \in [n]$ is a ***random vector***. Each $y_i \in \mathbb{R}$ is a ***random variable.*** There exists a joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$, where we draw:

$$(\mathbf{x}_i, y_i) \sim \mathbb{P}_{\mathbf{x},y}.$$

We want to find a ***model*** of the data, a function $f : \mathbb{R}^d \to \mathbb{R}$ that *generalizes* well to a newly drawn $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

To choose the model f, make the assumption that it is *linear*: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}$.

Our notion of error is the ***squared loss***:

$$\ell(f(\mathbf{x}), y) := (y - f(\mathbf{x}))^2.$$

To choose the model f, we attempt to minimize the expected squared loss, or the ***risk***:

$P_{x,y}$

$$R(f) := \mathbb{E}_{\mathbf{x},y}[(y - f(\mathbf{x}))^2] = \int (y - f(\mathbf{x}))^2 d\mathbb{P}(\mathbf{x}, y)$$

LLN.

As a substitute, we can minimize the ***empirical risk***:

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2. \quad = \|Xw - y\|^2$$

Statistics of the OLS Estimator

Bias and Variance

Statistics of the Error Model

Setup

Let $\mathbf{x} \in \mathbb{R}^d$ be a *random vector* and $y \in \mathbb{R}$ be *random variable* be drawn from the *joint distribution* $\mathbb{P}_{\mathbf{x},y}$, where

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where ϵ is a *random variable* with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, with ϵ independent of $\mathbf{x}$.

Statistics of the Error Model

Expectation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\mathbb{E}[\epsilon \mid \mathbf{x}] = 0$, because errors are independent of $\mathbf{x}$.

$$\mathbb{E}[\varepsilon \mid x] = \mathbb{E}[\varepsilon] = 0.$$

Statistics of the Error Model

Variance

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\mathbb{E}[\epsilon \mid \mathbf{x}] = 0$, because errors are independent of $\mathbf{x}$.

$\mathrm{Var}(\epsilon \mid \mathbf{x}) = \sigma^2$, because errors are independent of $\mathbf{x}$.

$\mathrm{Var}(\varepsilon | x) = \mathbb{E}[(\varepsilon - \mathbb{E}[\varepsilon])^2 | x] = \mathbb{E}[\varepsilon^2 | x] = \mathbb{E}[\varepsilon^2]$

$= \mathrm{Var}(\varepsilon) = \sigma^2$.

Statistics of the Error Model

Conditional Expectation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\mathbb{E}[\epsilon \mid \mathbf{x}] = 0$, because errors are independent of $\mathbf{x}$.

$\text{Var}(\epsilon \mid \mathbf{x}) = \sigma^2$, because errors are independent of $\mathbf{x}$.

$\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}^\top \mathbf{w}^*$, the ***regression function***.

$$\mathbb{E}[y \mid x] = \mathbb{E}[x^T w^* + \varepsilon \mid x] = \mathbb{E}[x^T w^* \mid x] + \mathbb{E}[\varepsilon \mid x]$$

$$= \mathbb{E}[x^T w^* \mid x]$$

$$= x^T w^* \cdot \mathbb{E}[1 \mid x] = \boxed{x^T w^*}. \quad \blacksquare$$

Statistics of the Error Model

Conditional Expectation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

$\mathbb{E}[\epsilon \mid \mathbf{x}] = 0$, because errors are independent of $\mathbf{x}$.

$\text{Var}(\epsilon \mid \mathbf{x}) = \sigma^2$, because errors are independent of $\mathbf{x}$.

$\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}^\top \mathbf{w}^*$, the ***regression function***.

This is the target we're aiming for!

Statistics of OLS

Using OLS to minimize empirical risk

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Find $f(\mathbf{x}) := \hat{\mathbf{w}}^\top \mathbf{x}$ that does well on training samples, minimizing ***empirical risk***:

$$\hat{R}(\hat{f}) := \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(\mathbf{x}_i) - y_i)^2 = \frac{1}{n} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

Obtain the least squares estimator the same way:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Statistics of OLS

Using OLS to minimize empirical risk

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Obtain the least squares estimator the same way:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This $\hat{\mathbf{w}} \in \mathbb{R}^d$ is a random vector now!

If we *condition on* $\mathbf{X} \in \mathbb{R}^{n \times d}$, we can get statistics on this random vector:

Statistics of OLS

Expectation

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Obtain the least squares estimator the same way:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This $\hat{\mathbf{w}} \in \mathbb{R}^d$ *is a random vector now!*

If we *condition on* $\mathbf{X} \in \mathbb{R}^{n \times d}$, we can get statistics on this random vector:

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$

as a vector

$$\begin{aligned}
\mathbb{E}[\hat{w} \mid X] &= \mathbb{E}[(X^\top X)^{-1} X^\top y \mid X] \\
&= \mathbb{E}[(X^\top X)^{-1} X^\top (X w^* + \varepsilon) \mid X] \\
&= \mathbb{E}[(X^\top X)^{-1} X^\top X w^* + (X^\top X)^{-1} X^\top \varepsilon \mid X] \\
&= \mathbb{E}[w^* + (X^\top X)^{-1} X^\top \varepsilon \mid X] \\
&= \mathbb{E}[w^* \mid X] + \mathbb{E}[(X^\top X)^{-1} X^\top \varepsilon \mid X] \\
&= w^* .
\end{aligned}$$

$\mathbb{E}[\varepsilon \mid X] = 0.$

ε is indep. from X

Statistics of OLS

Variance

$X^\top X = V \lambda V^\top$

$(X^\top X)^{-1} = V \lambda^{-1} V^\top$

$= V \begin{bmatrix} \sigma^2/\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \sigma^2/\lambda_d \end{bmatrix} V^\top$

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Obtain the least squares estimator the same way:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This $\hat{\mathbf{w}} \in \mathbb{R}^d$ *is a random vector now!*

If we *condition on* $\mathbf{X} \in \mathbb{R}^{n \times d}$, we can get statistics on this random vector:

Empirical covariance

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$.

Variance: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$. ← $Var(\epsilon) = \sigma^2$.

Statistics of OLS
Intuition

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Obtain the least squares estimator the same way:

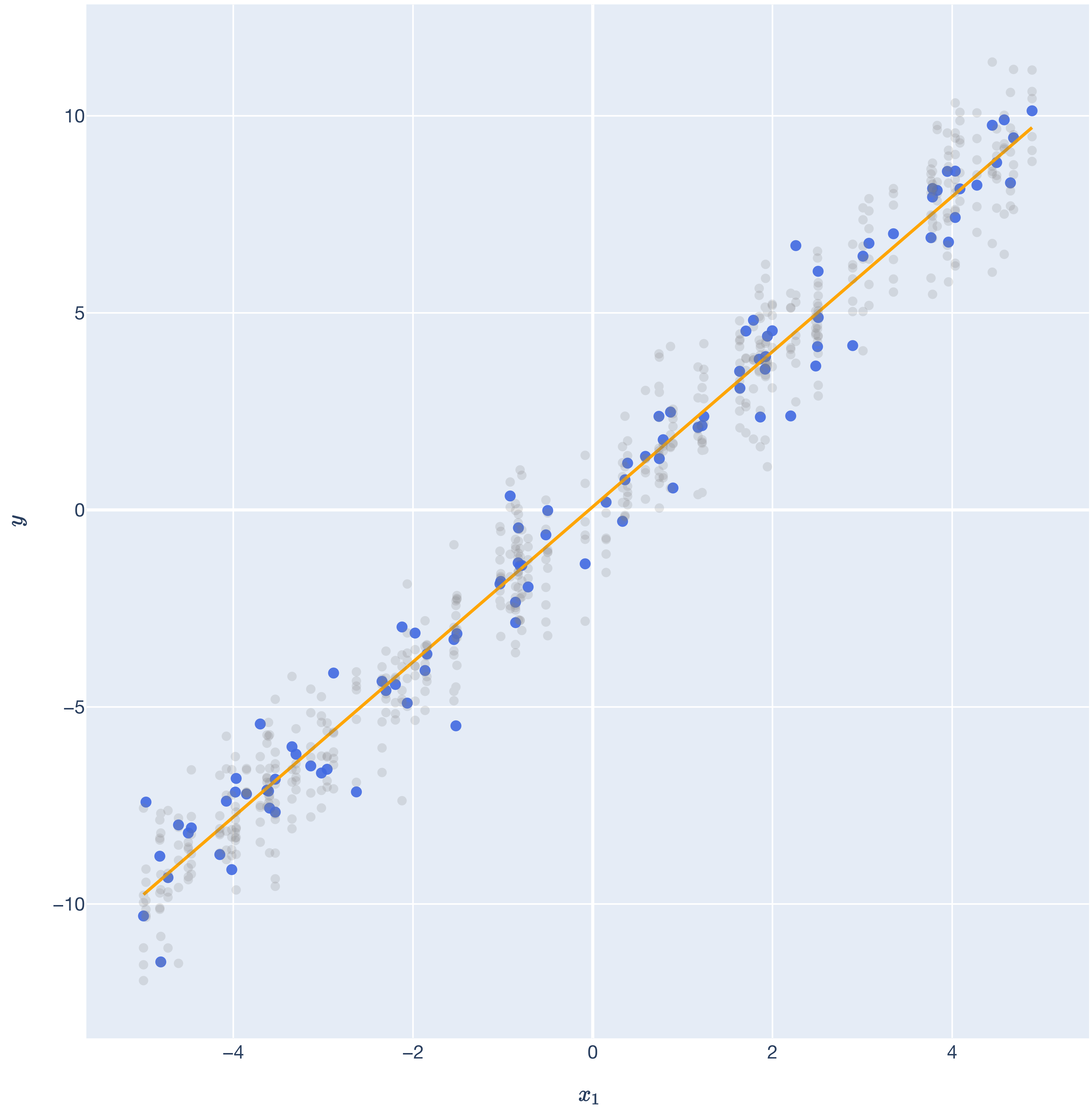
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This $\hat{\mathbf{w}} \in \mathbb{R}^d$ is a random vector now!

If we *condition on* $\mathbf{X} \in \mathbb{R}^{n \times d}$, we can get statistics on this random vector:

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$.

Variance: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$.



Statistics of OLS

Intuition

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$$

Obtain the least squares estimator the same way:

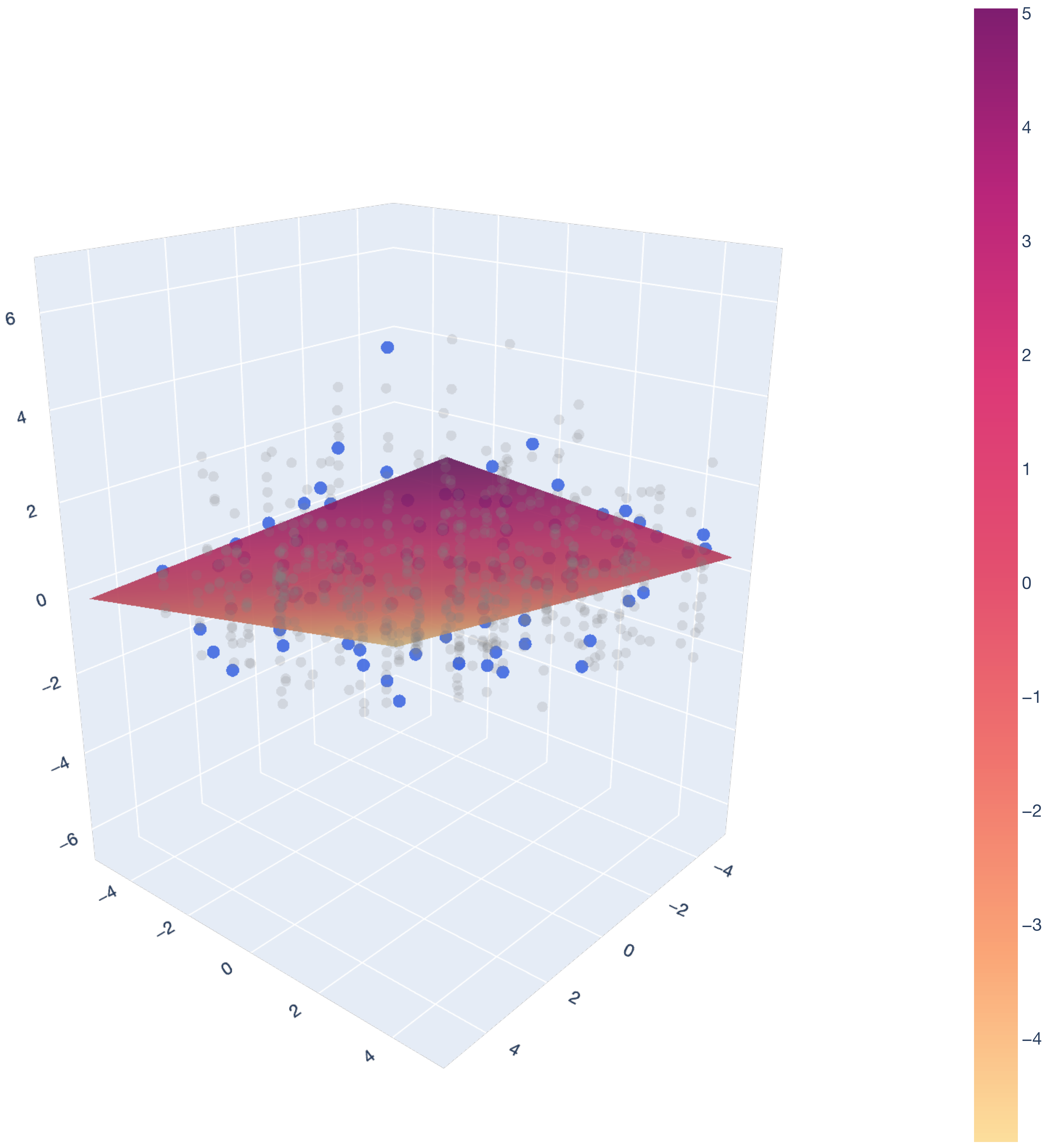
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

This $\hat{\mathbf{w}} \in \mathbb{R}^d$ is a random vector now!

If we *condition on* $\mathbf{X} \in \mathbb{R}^{n \times d}$, we can get statistics on this random vector:

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$.

Variance: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$.



Statistics of OLS

Theorem

Theorem (Statistical properties of OLS). Let $\mathbb{P}_{\mathbf{x},y}$ be a joint distribution $\mathbb{R}^d \times \mathbb{R}$ defined by the error model:

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon,$$

where $\mathbf{w}^* \in \mathbb{R}^d$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$. Suppose we construct a random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and random vector $\mathbf{y} \in \mathbb{R}^n$ by drawing n random examples $(\mathbf{x}_i, y_i)$ from $\mathbb{P}_{\mathbf{x},y}$. Then, the OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the following statistical properties:

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$.

Variance: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2$.

Recap

Lesson Overview

Probability Spaces. We'll review the basic axioms and components of probability: sample space, events, and probability measures. This allows us to ditch these notions and introduce *random variables.*

Random variables. Review of the definition of a random variable, its *distribution/law*, its PDF/PMF/CDF, and joint distributions of several RVs.

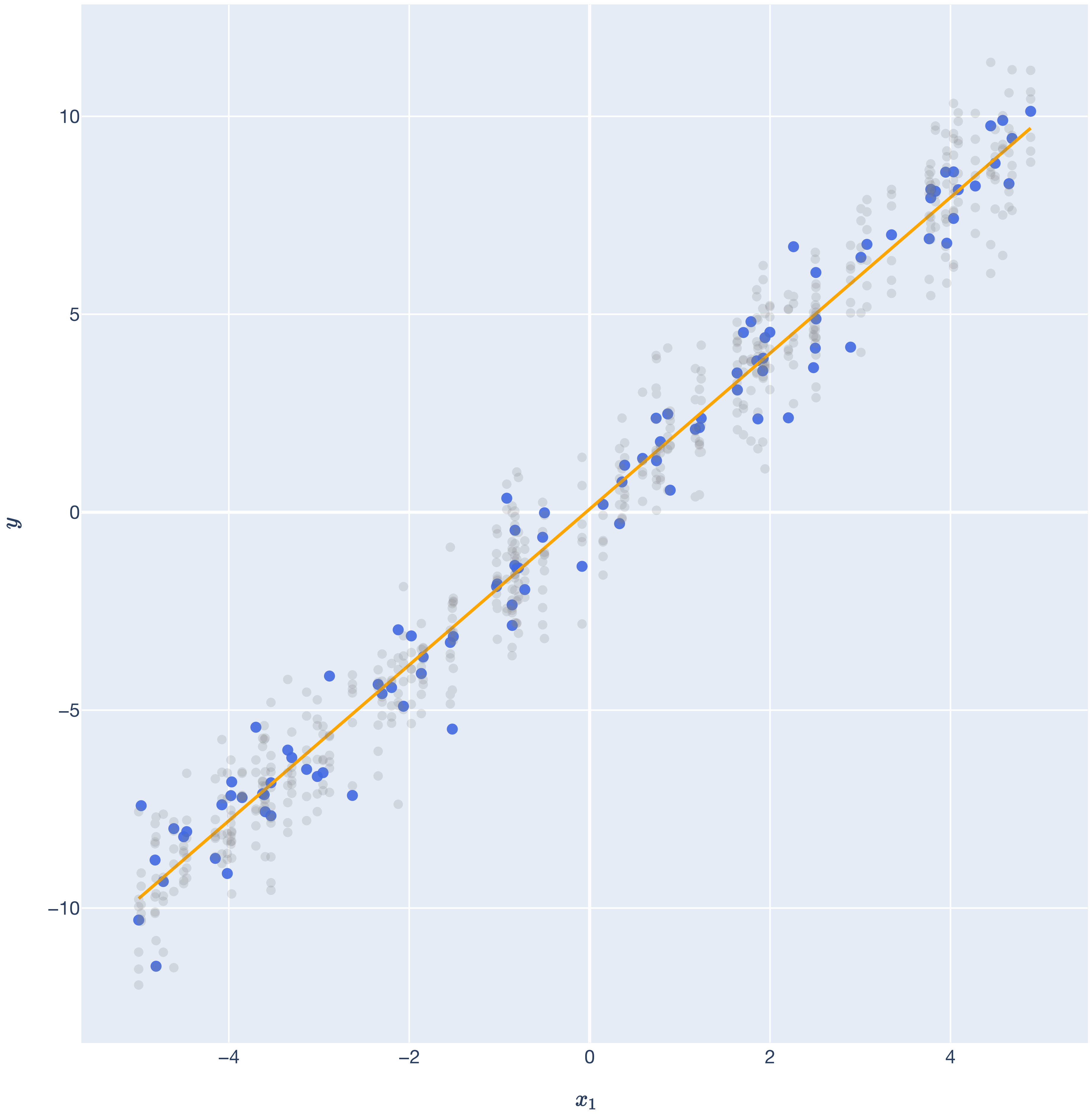
Expectation, variance, and covariance. Review of these basic summary statistics of random variables and common properties.

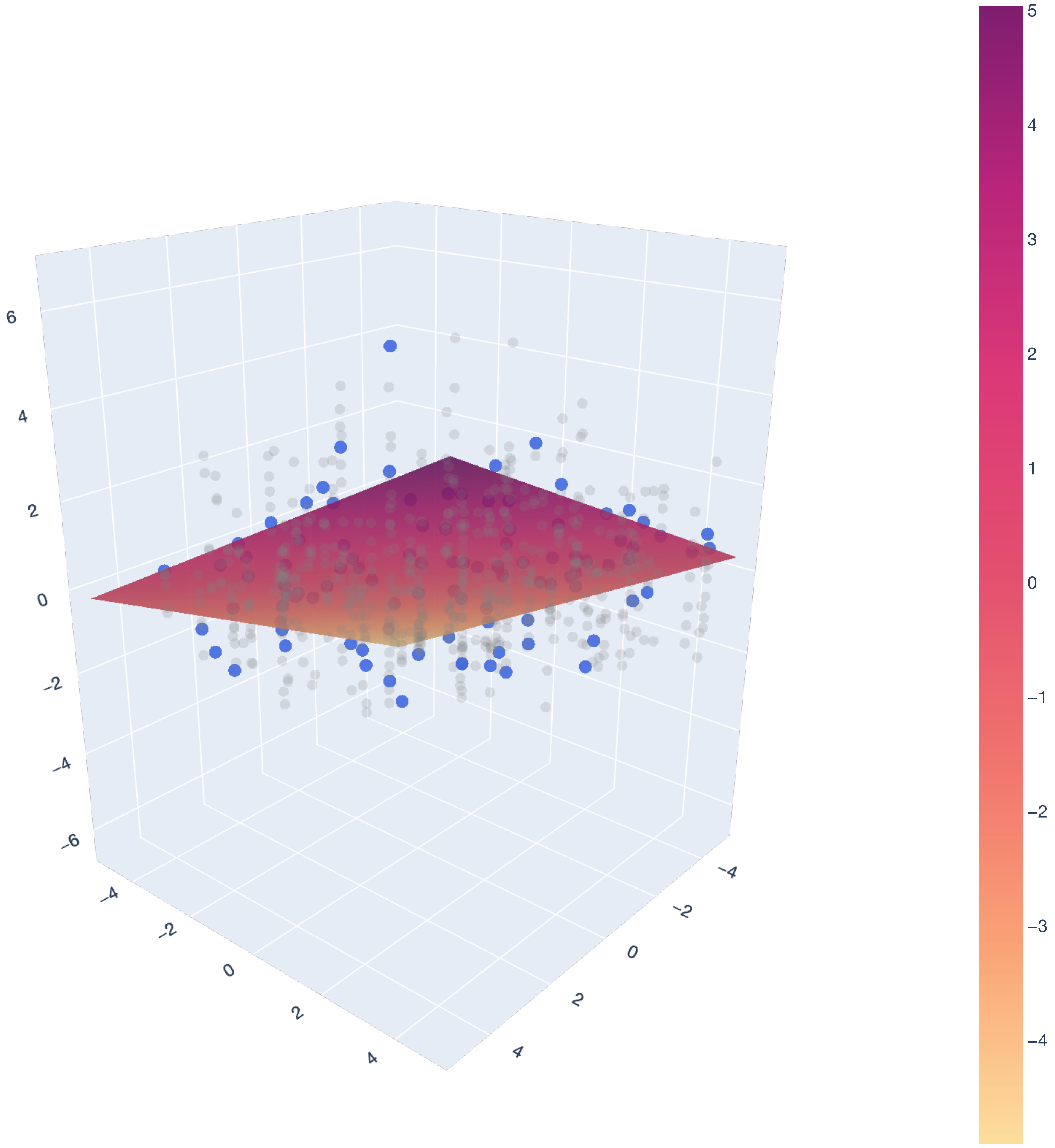
Random vectors. Introduce the idea of a *random vector*, which is just a list of multiple random variables. Discuss generalizations of expectation and variance to random vectors.

Data as random, statistical model of ML. Introduce the statistical model of ML and the random error model. Introduce *modeling assumptions.* State and prove basic statistical properties of the OLS estimator.

Lesson Overview

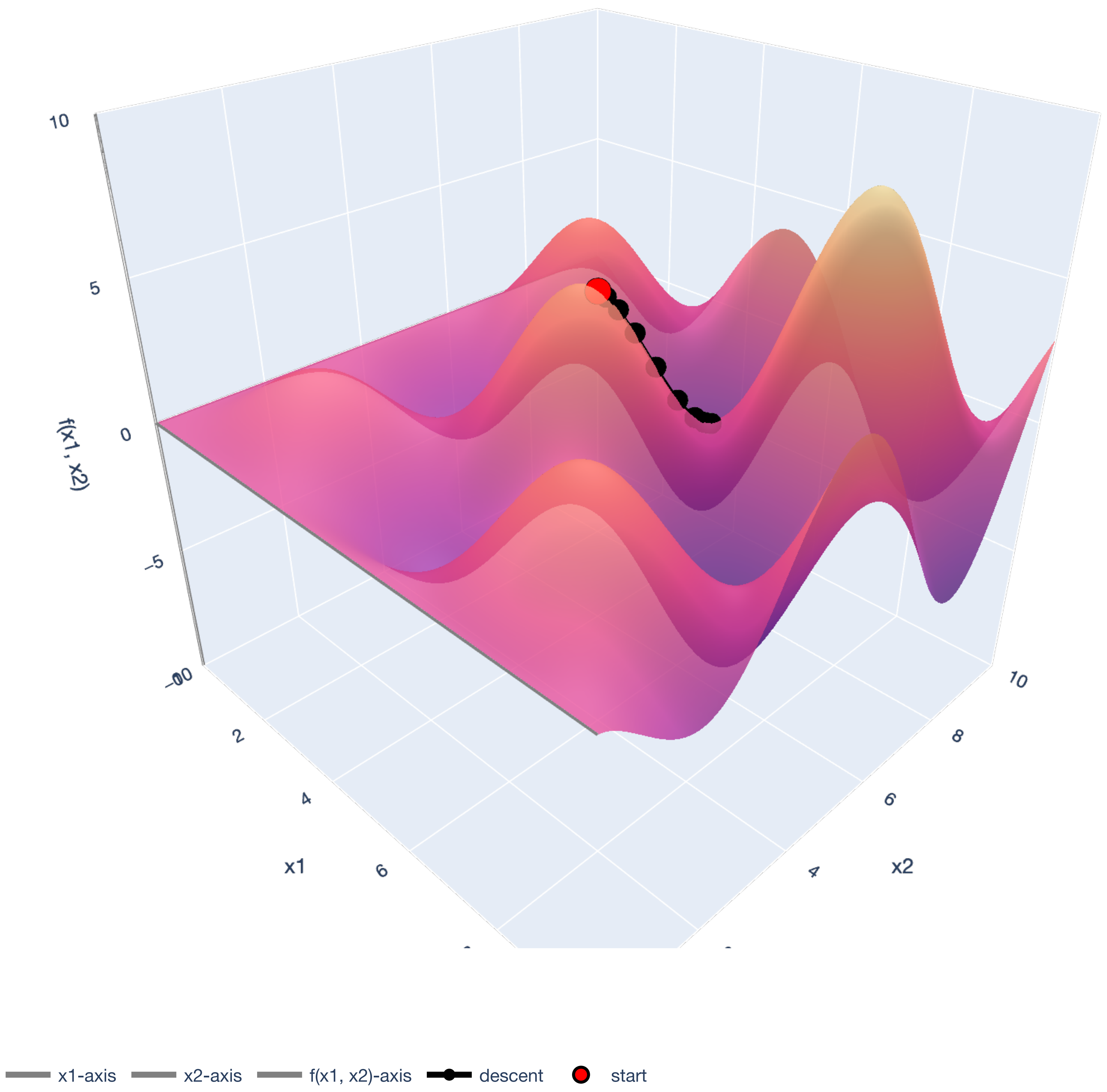
Big Picture: Least Squares

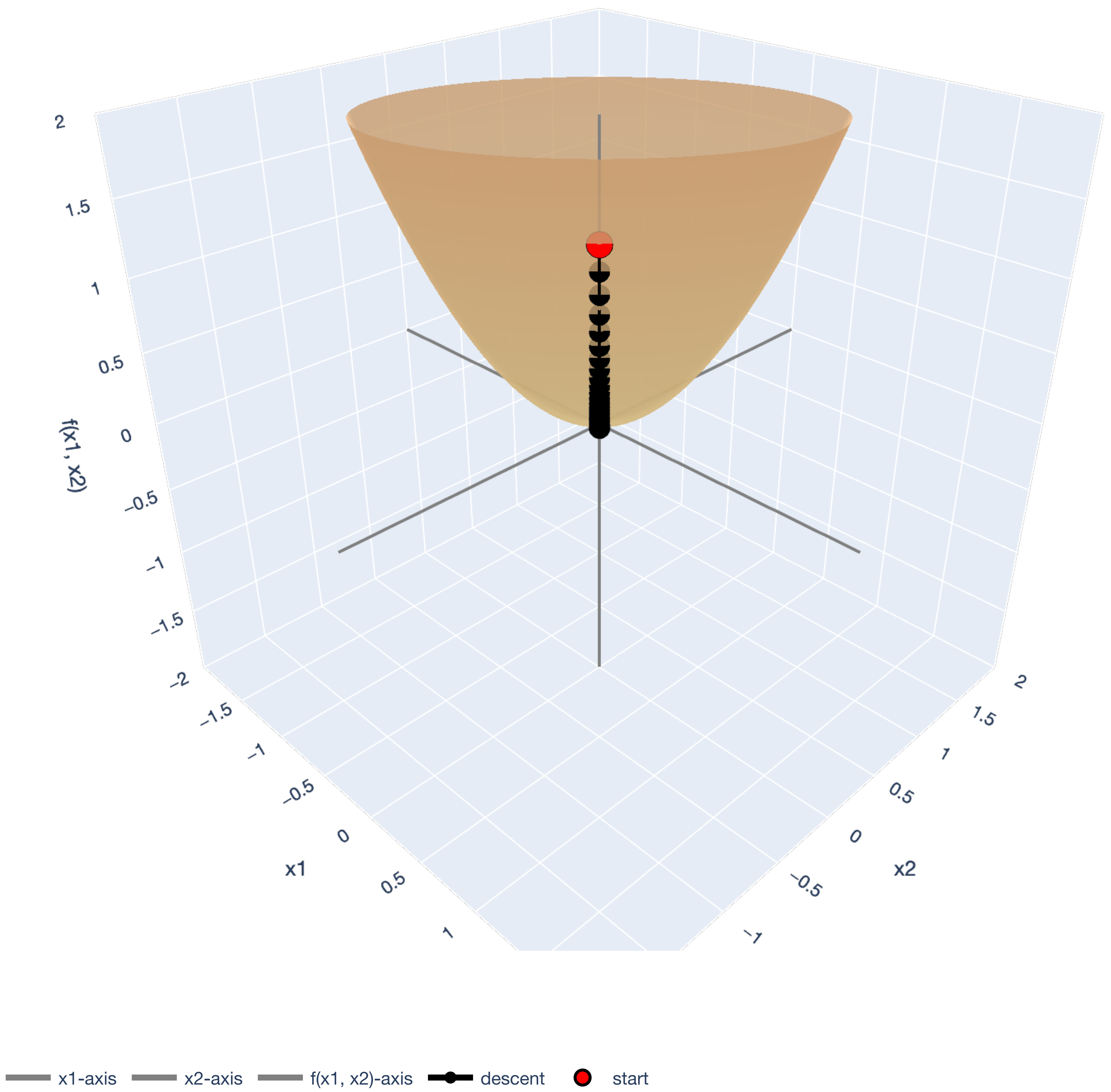




Lesson Overview

Big Picture: Gradient Descent





References

Mathematics for Machine Learning. Marc Pieter Deisenroth, A. Aldo Faisal, Cheng Soon Ong.

Elements of Statistical Learning: *Data Mining, Inference, and Prediction.* Trevor Hastie, Robert Tibshirani, Jerome Friedman.