# Math for ML

## Week 5.1: Basic Probability Theory, Models, and Data

By: Samuel Deng

# Logistics & Announcements

# 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.

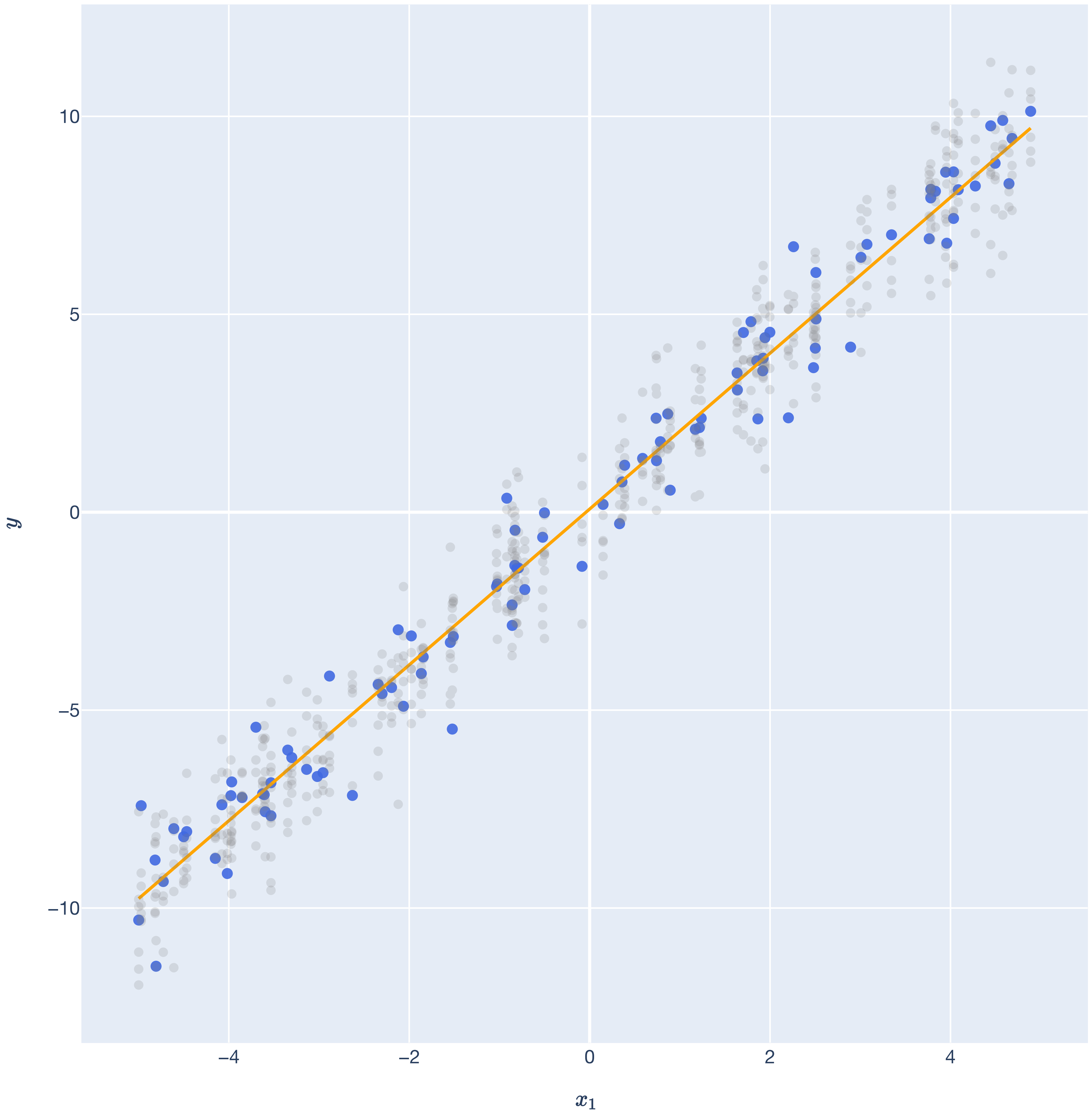
**Expectation, variance, and covariance.** Review of these basic summary statistics of random variables and common properties.

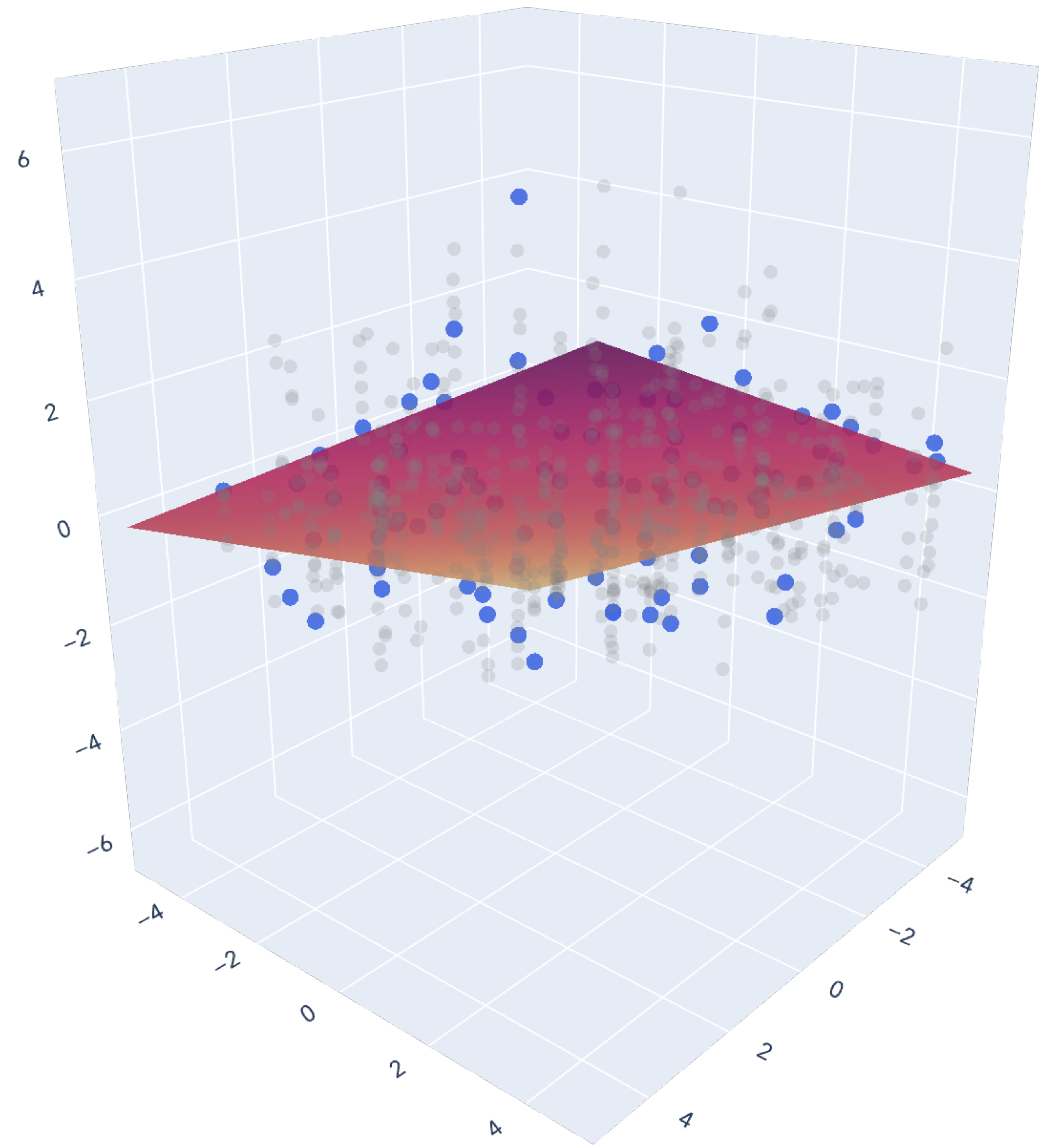
**Random vectors.** Introduce the idea of a *random vector*, which is just a list of multiple random variables. Discuss generalizations of expectation and variance to random vectors.

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

# Lesson Overview

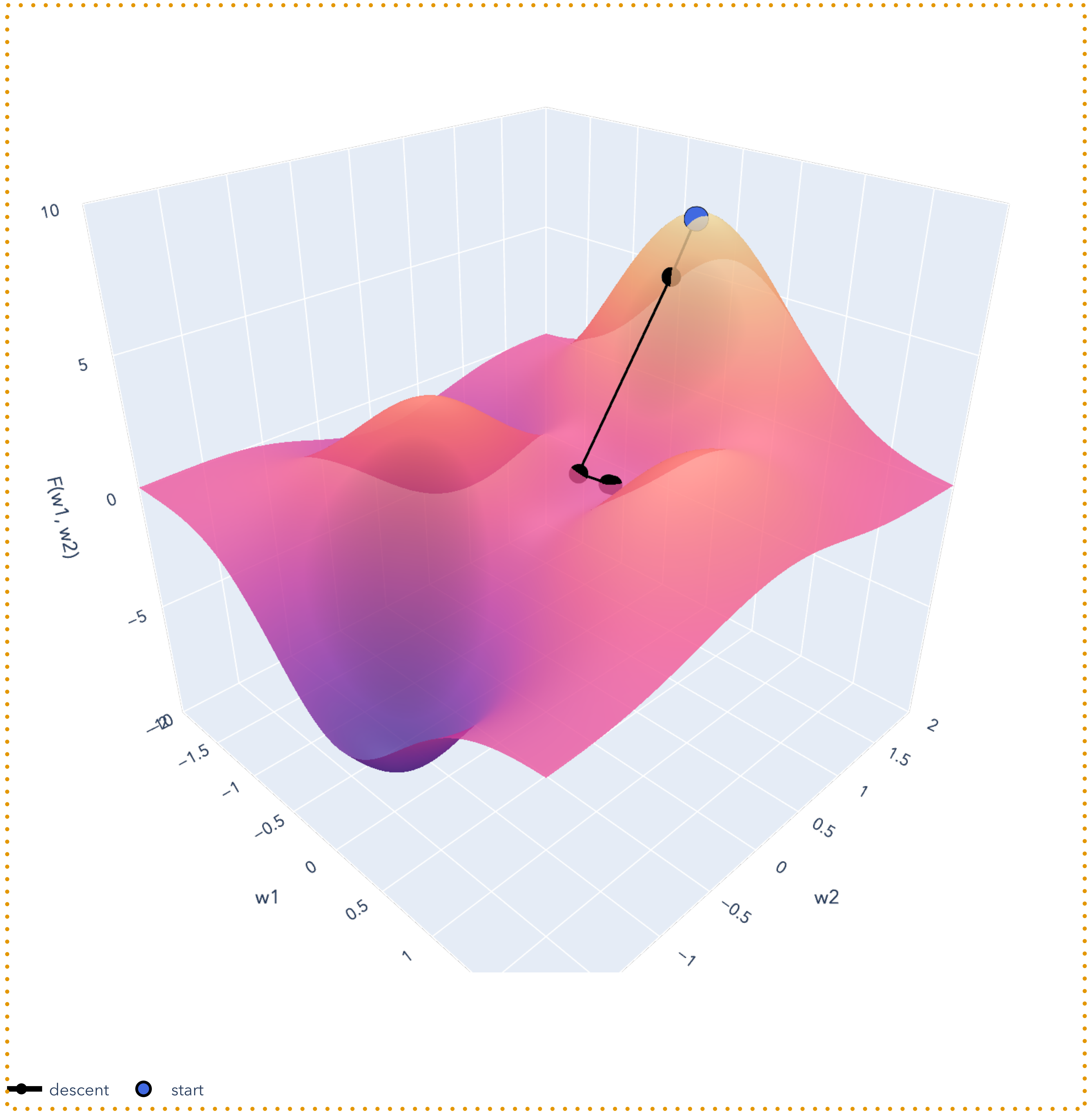
## Big Picture: Least Squares

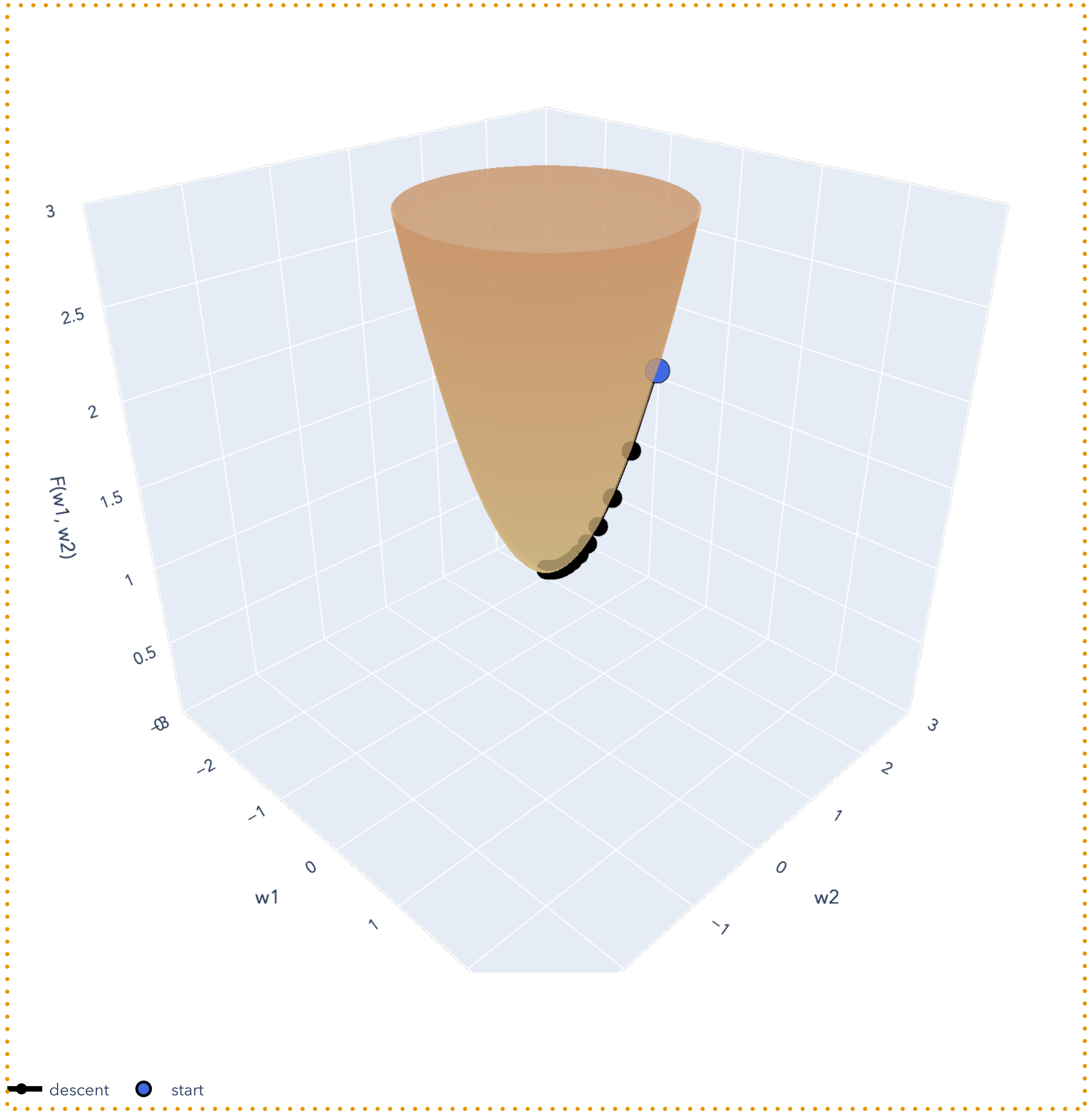


# Lesson Overview

## Big Picture: Gradient Descent



# Motivation

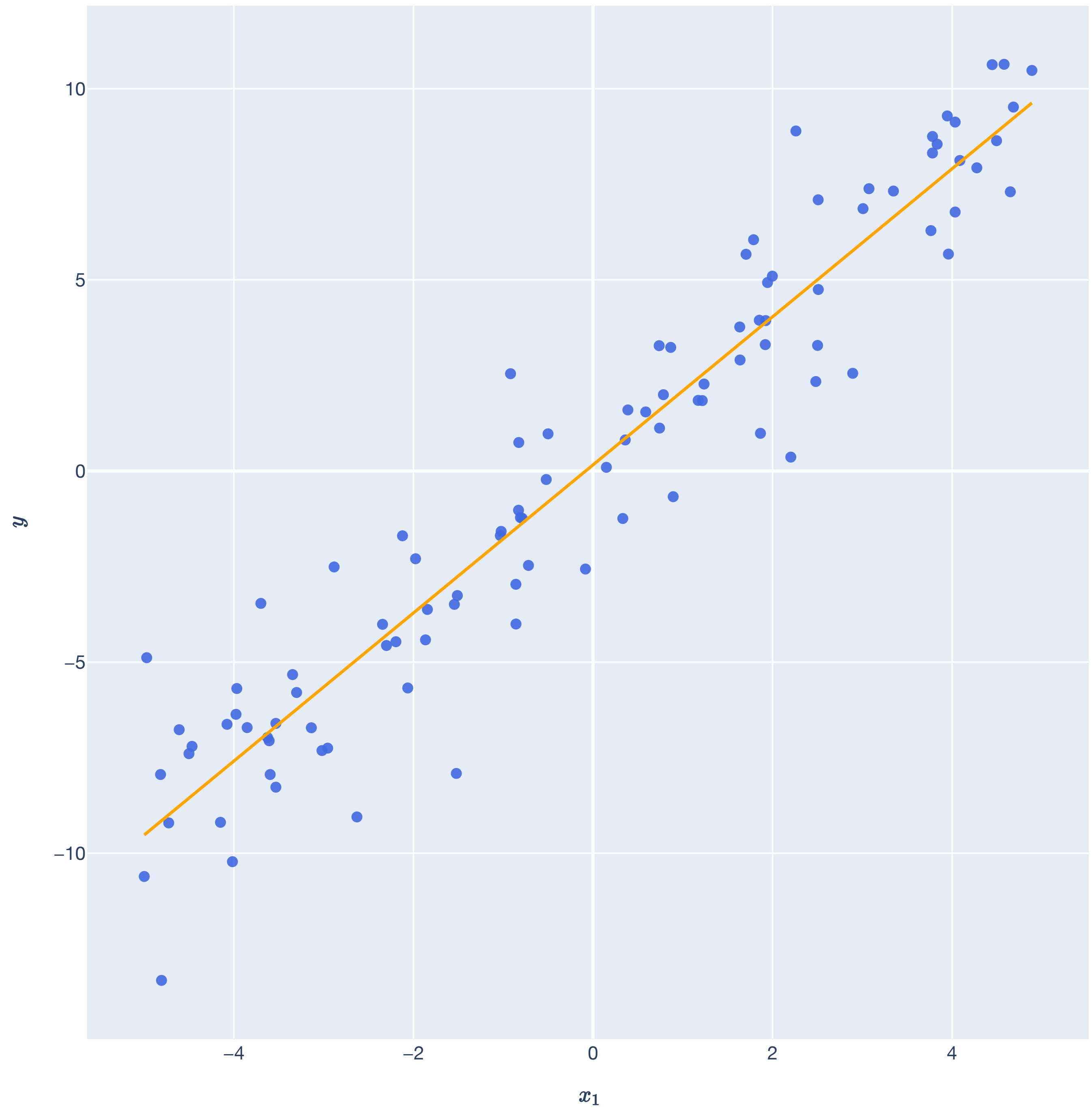
## Data as randomly distributed

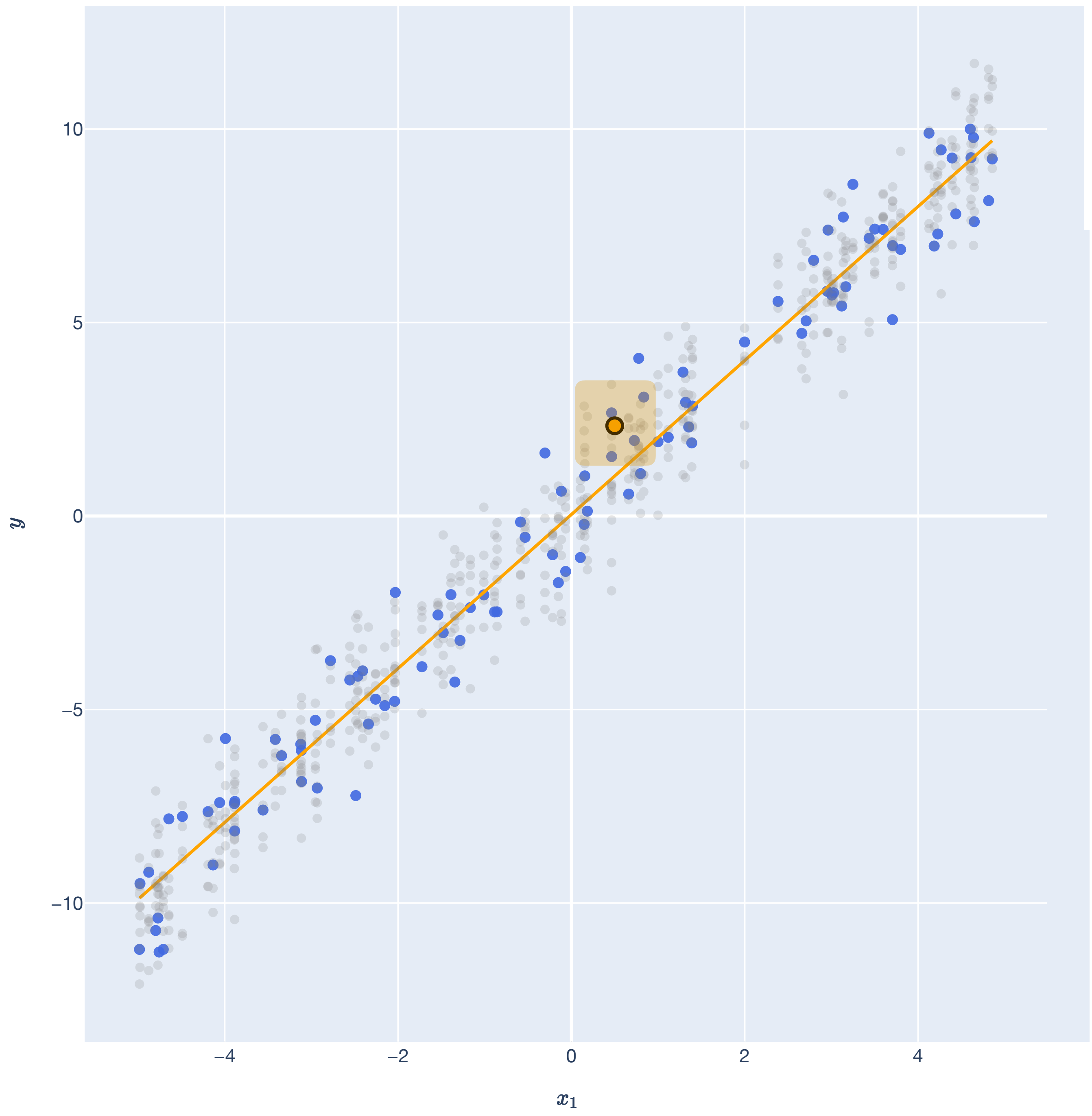
# Regression

## Setup

What we've been doing so far: Collect labeled training data $\Longrightarrow$ Fit the model $\hat{\mathbf{w}}$

Hope training data looks like new data: Generalize on new $\mathbf{x}_0$.



# Regression

## Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

<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 (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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>Ultimate Goal:</u> 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

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 $1/n$ in front, and it looks like we're minimizing an average:

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

# Regression

## Setup, with randomness

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

Random vector

Random variable

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

… from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$!

<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: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Our "model", a function that we hope "generalizes" well to new datapoint $(\mathbf{x}_0, y_0)$.

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:

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

Because we assume that $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$ as well!

# Regression

## Setup, with randomness

To choose a model (*predictor/classifier/etc.*) $f : \mathbb{R}^d \to \mathbb{R}$, we want to minimize the expected squared loss, or the risk:

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

…but we can't know the true distribution, so as a substitute, minimize the empirical risk:

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

# Regression

## Setup, with randomness

Our choice of $f$ is $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, a linear model, so risk:

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

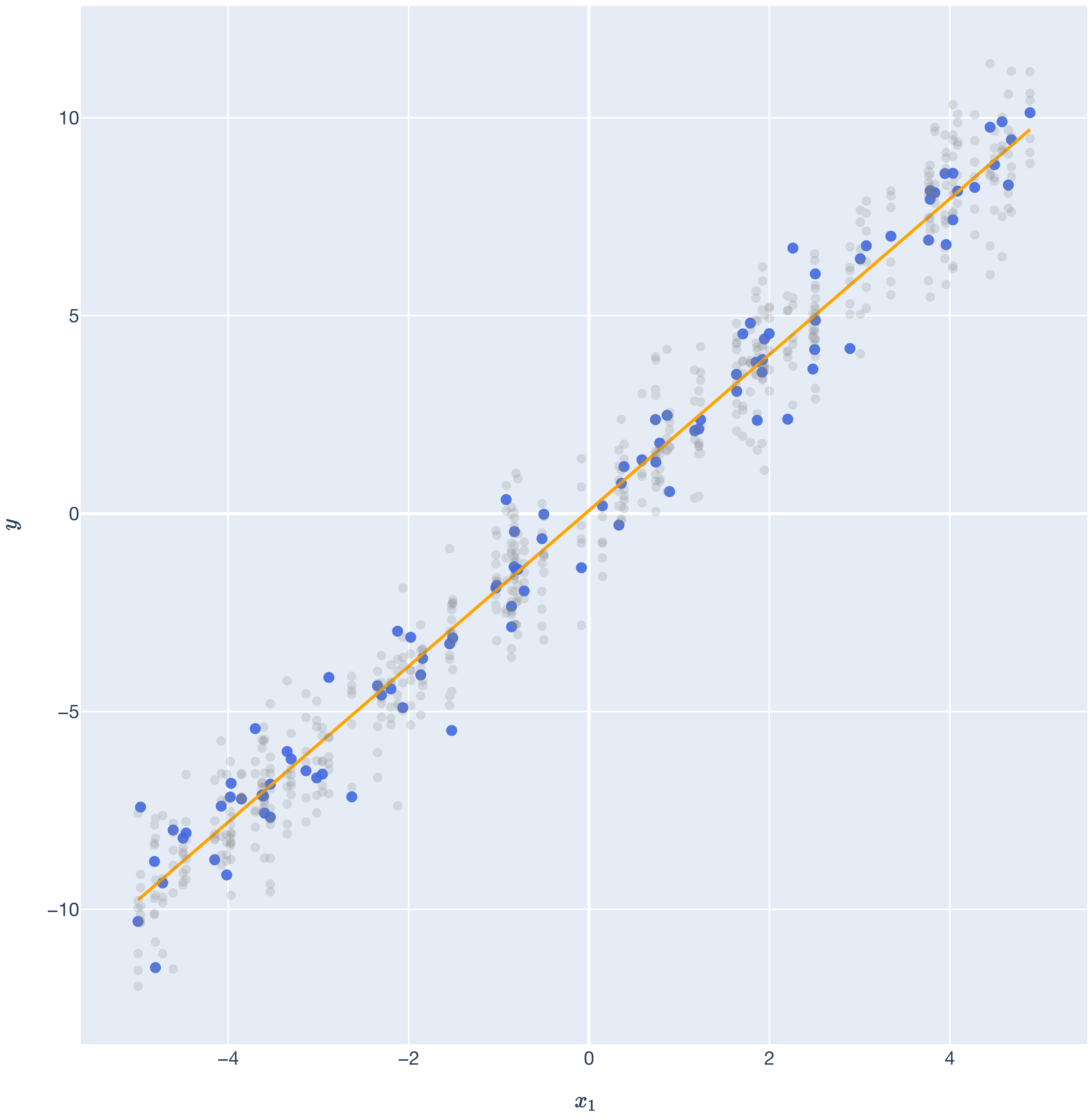
…but we can't know the true distribution, so as a substitute, minimize the empirical risk:

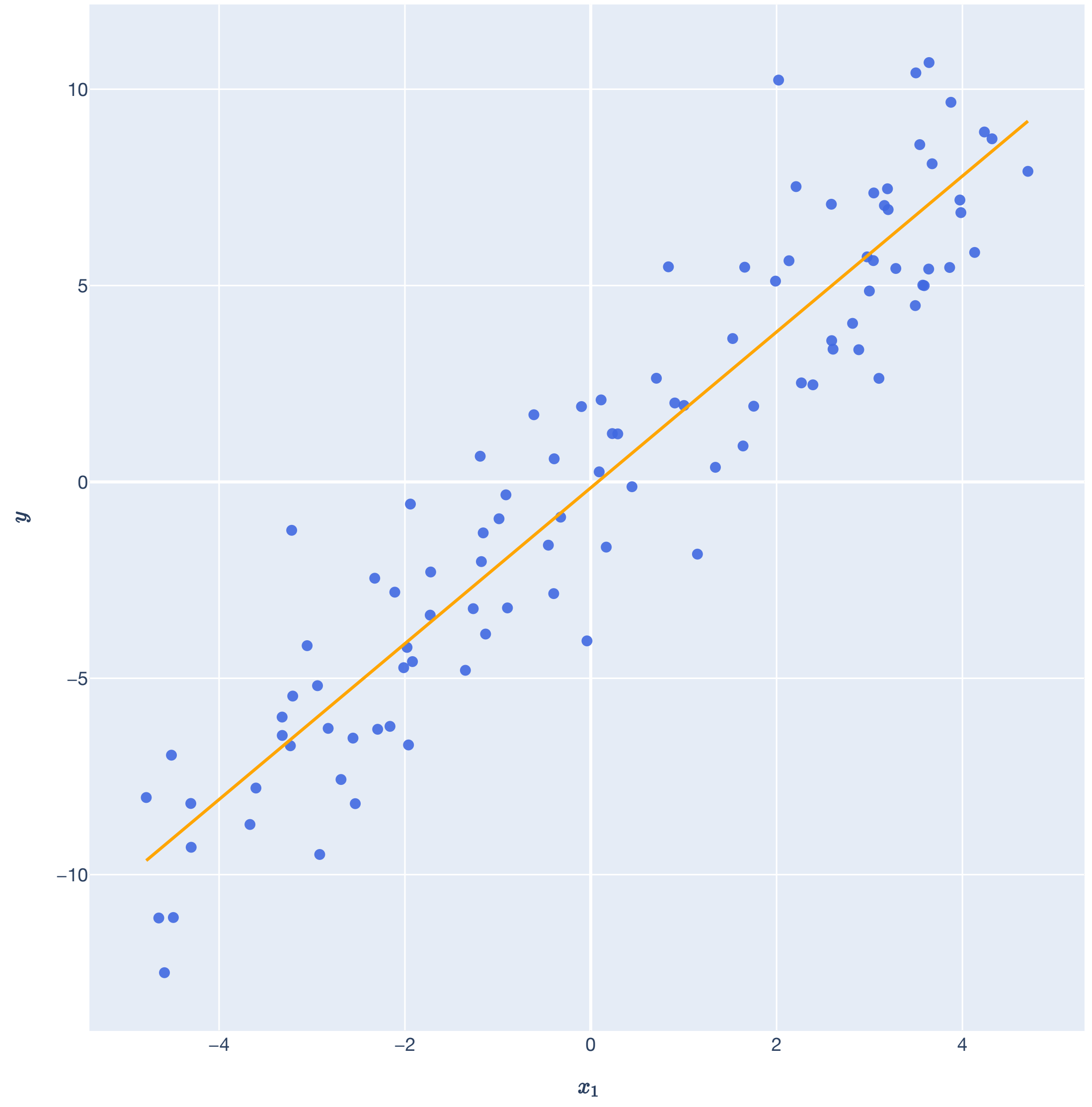
$$\hat{R}(f_{\mathbf{w}}) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

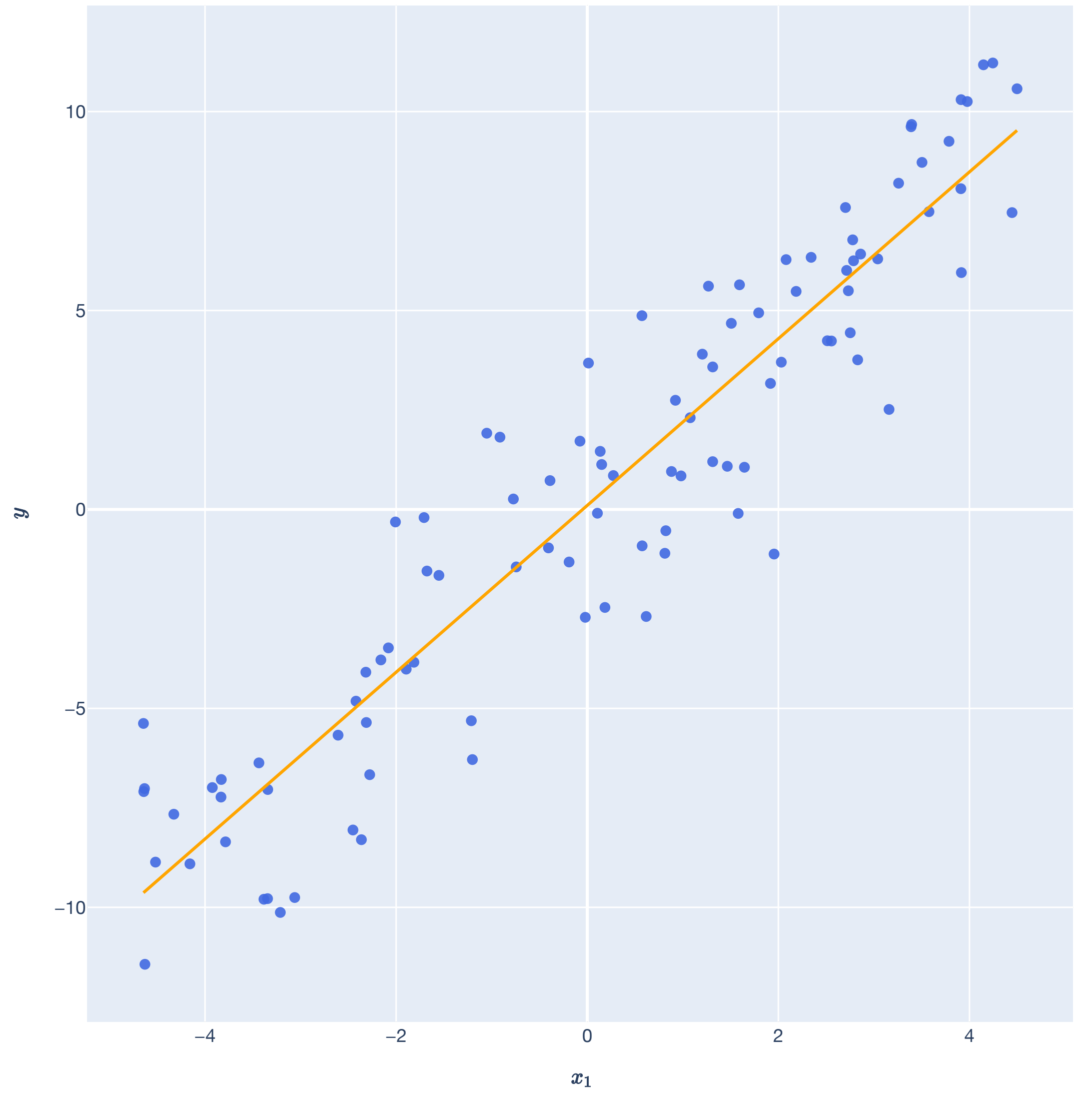
which is just our least squares error, scaled by $1/n$!

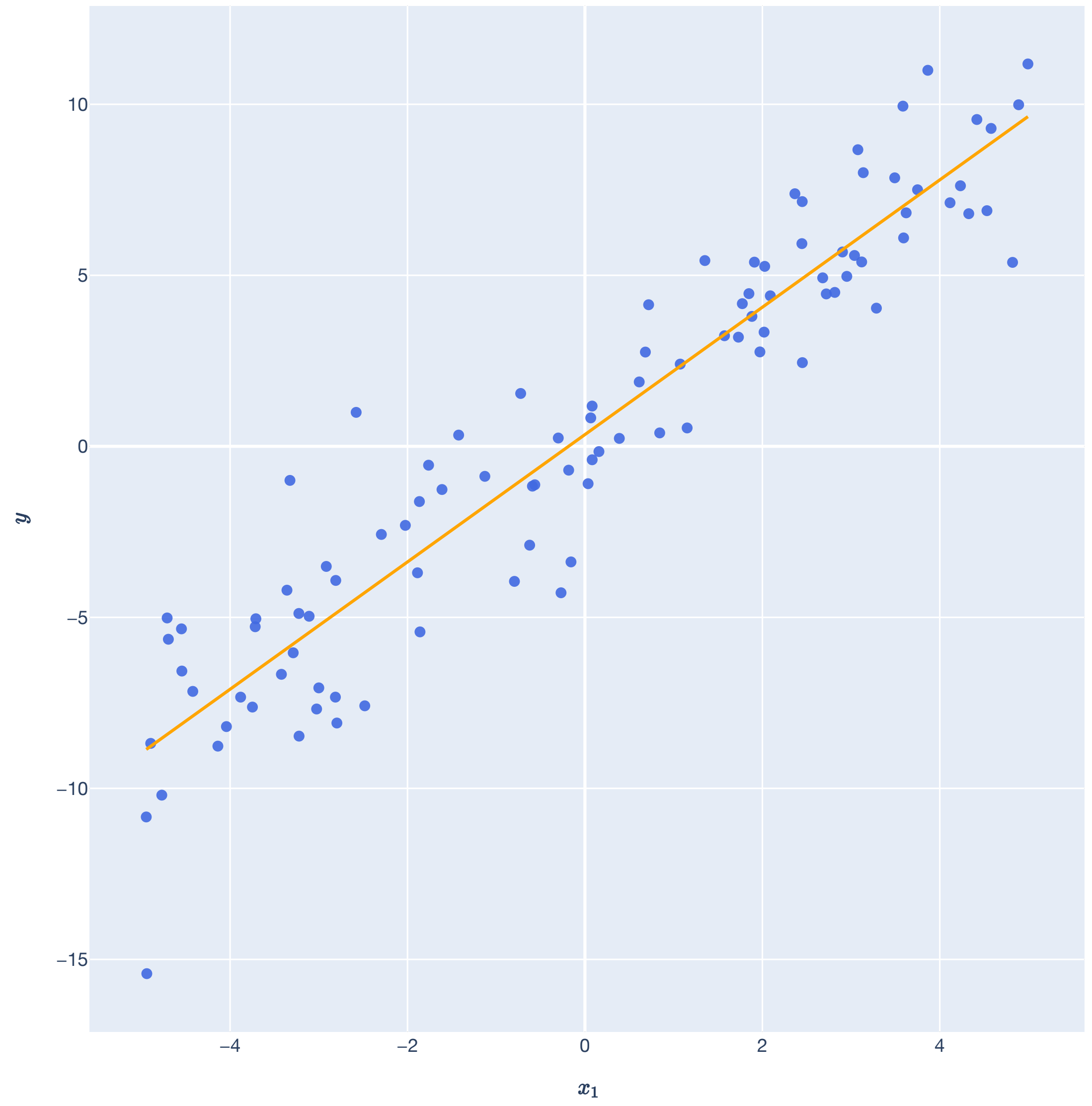
# Regression

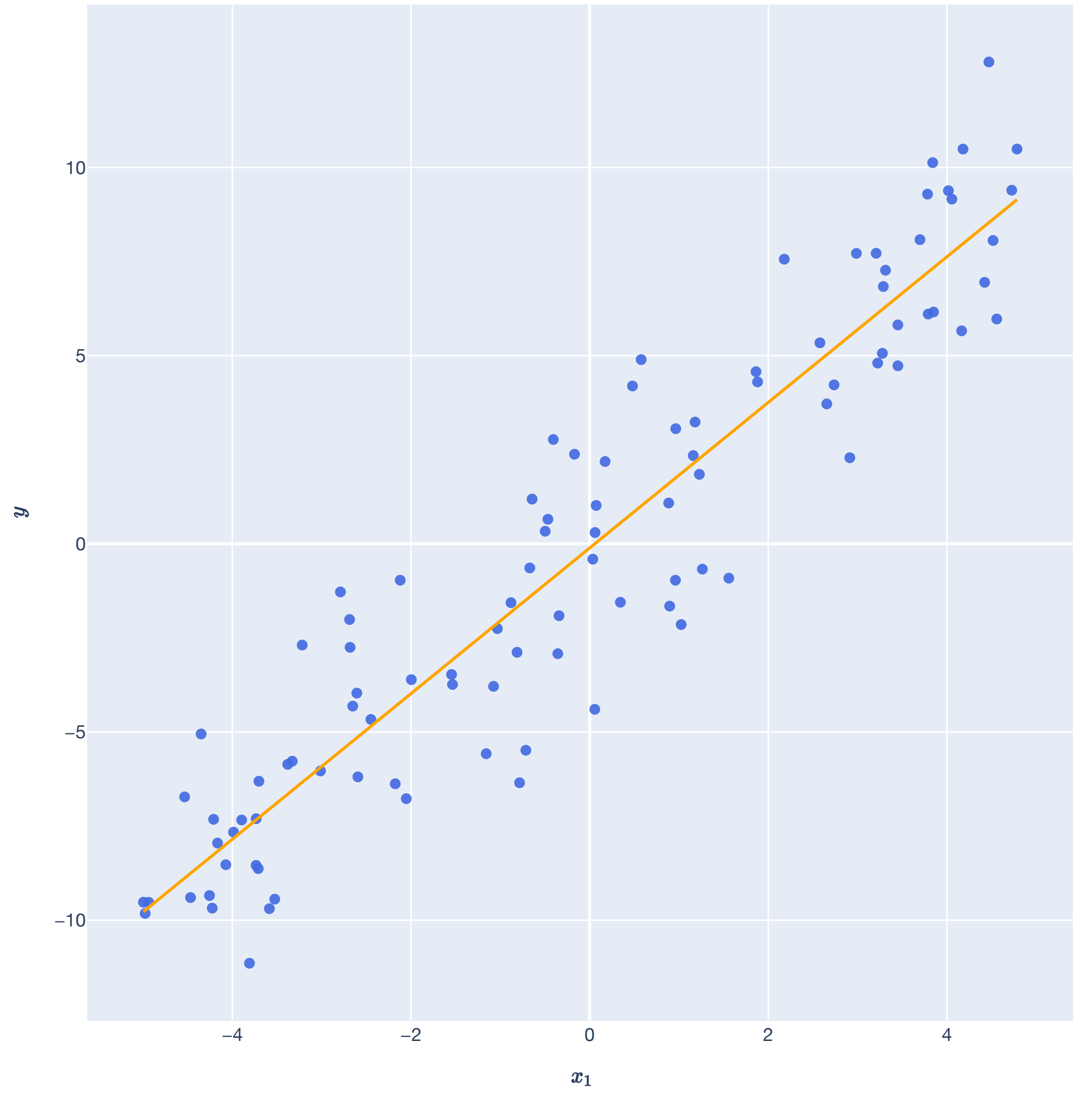
## Modeling randomness









# The "Story" of Probability and Statistics

## Modeling randomness

### 1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

$$\mathrm{y} \longleftarrow \boxed{\quad\text{nature}\quad} \longleftarrow \mathbf{x}$$

There are two goals in analyzing the data:

*Prediction.* To be able to predict what the responses are going to be to future input variables;

*Information.* To extract some information about how nature is associating the response variables to the input variables.

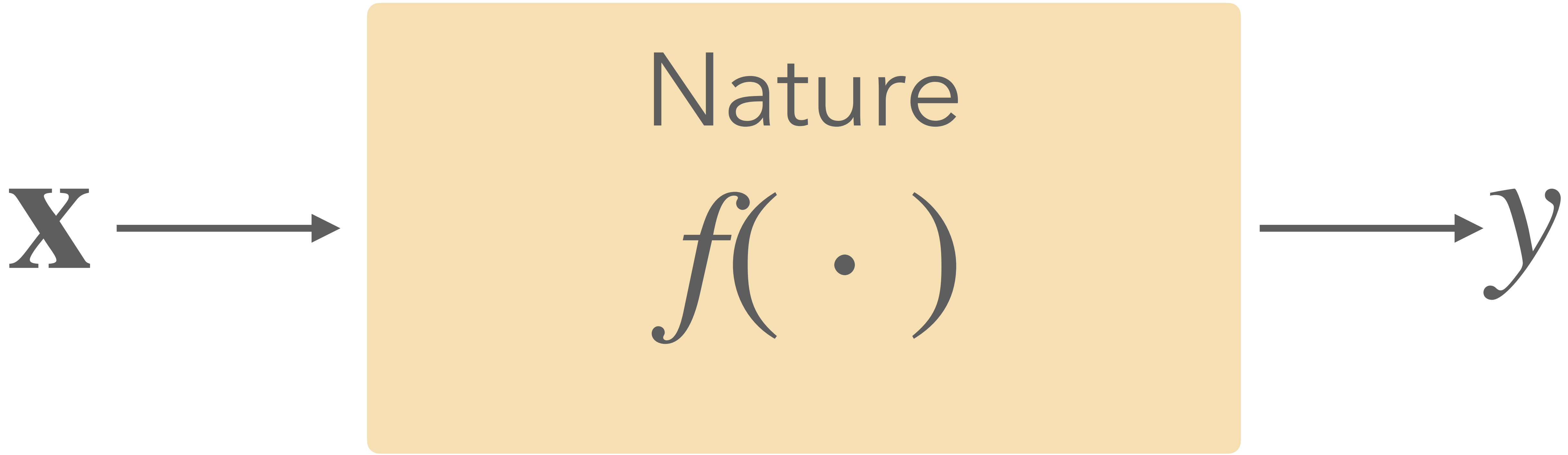
From: "Statistical Modeling: The Two Cultures" by Leo Breiman

# The "Story" of Probability and Statistics

## Modeling randomness

# The “Story” of Probability and Statistics

## Modeling randomness



# The "Story" of Probability and Statistics

## Modeling randomness

# The "Story" of Probability and Statistics

## 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 $T$.

| | |
|---|---|
| $H\,H$ | $T\,H$ |
| $H\,T$ | $T\,T$ |

$\Omega$

*What are the possible outcomes of this experiment?*

# Sample Space

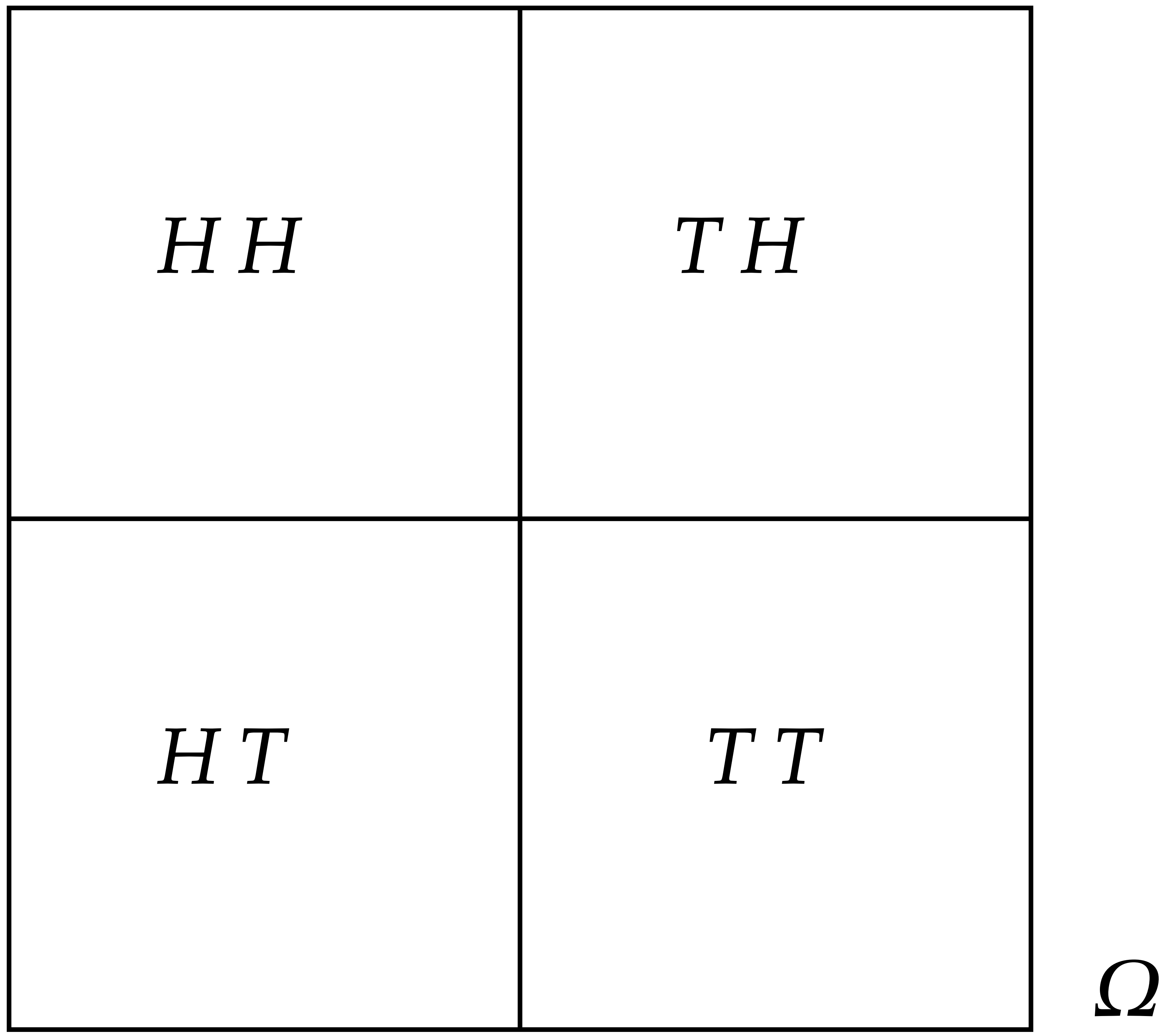
## Intuition and definition

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

Denote this $\Omega$.

**Example:**

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



# Events

## Intuition and definition

Given a sample space $\Omega$, an event is a subset $A \subseteq \Omega$ of outcomes.

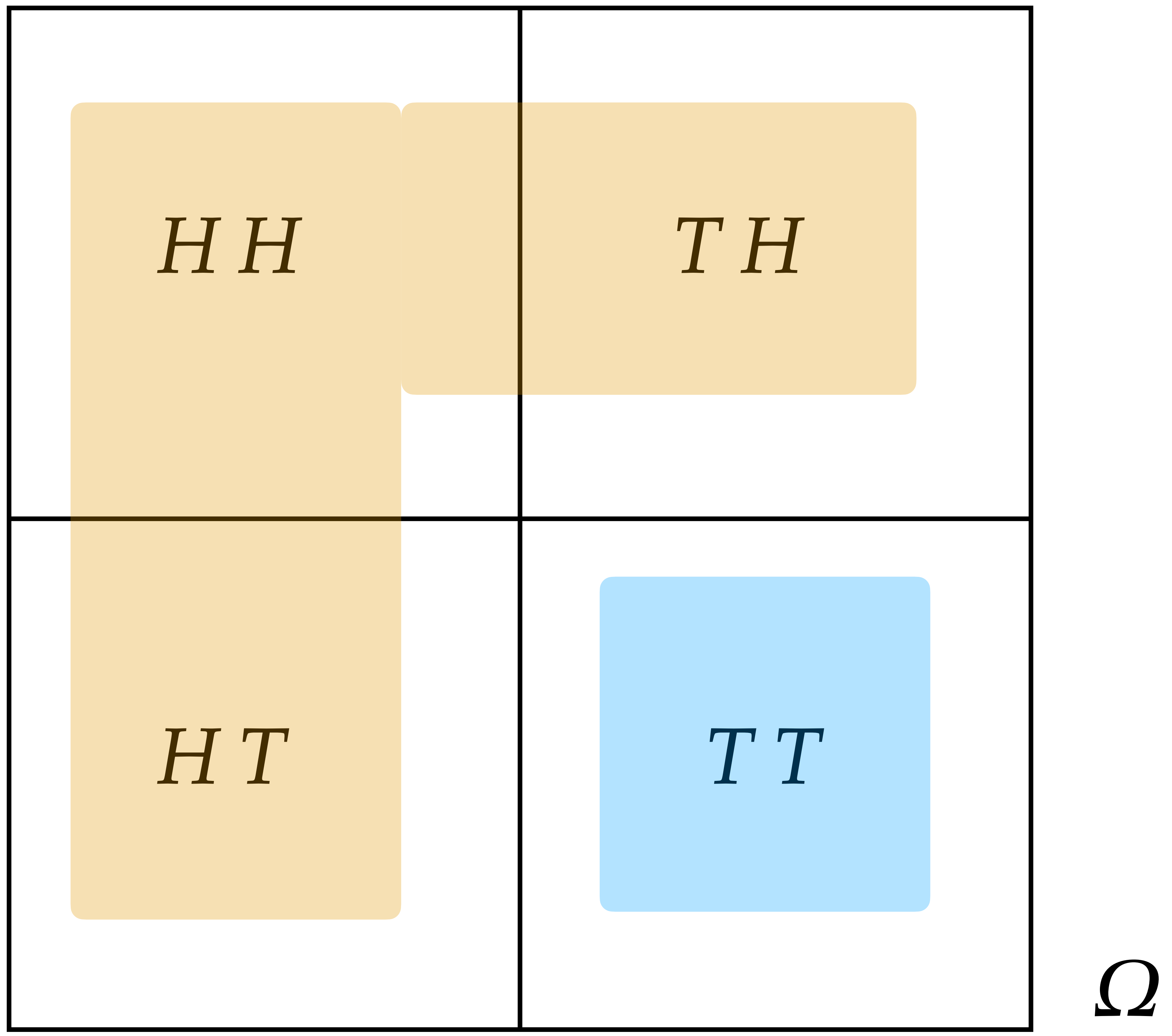
Denote a collection of events $\mathscr{A}$.

Example:

$A = \{HT, TH, HH\} = \{$"at least 1 head"$\}$

$B = \{TT\} = \{$"no heads"$\}$

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



# Events

## Intuition and definition

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

$$A \cup B \text{ (union)}$$

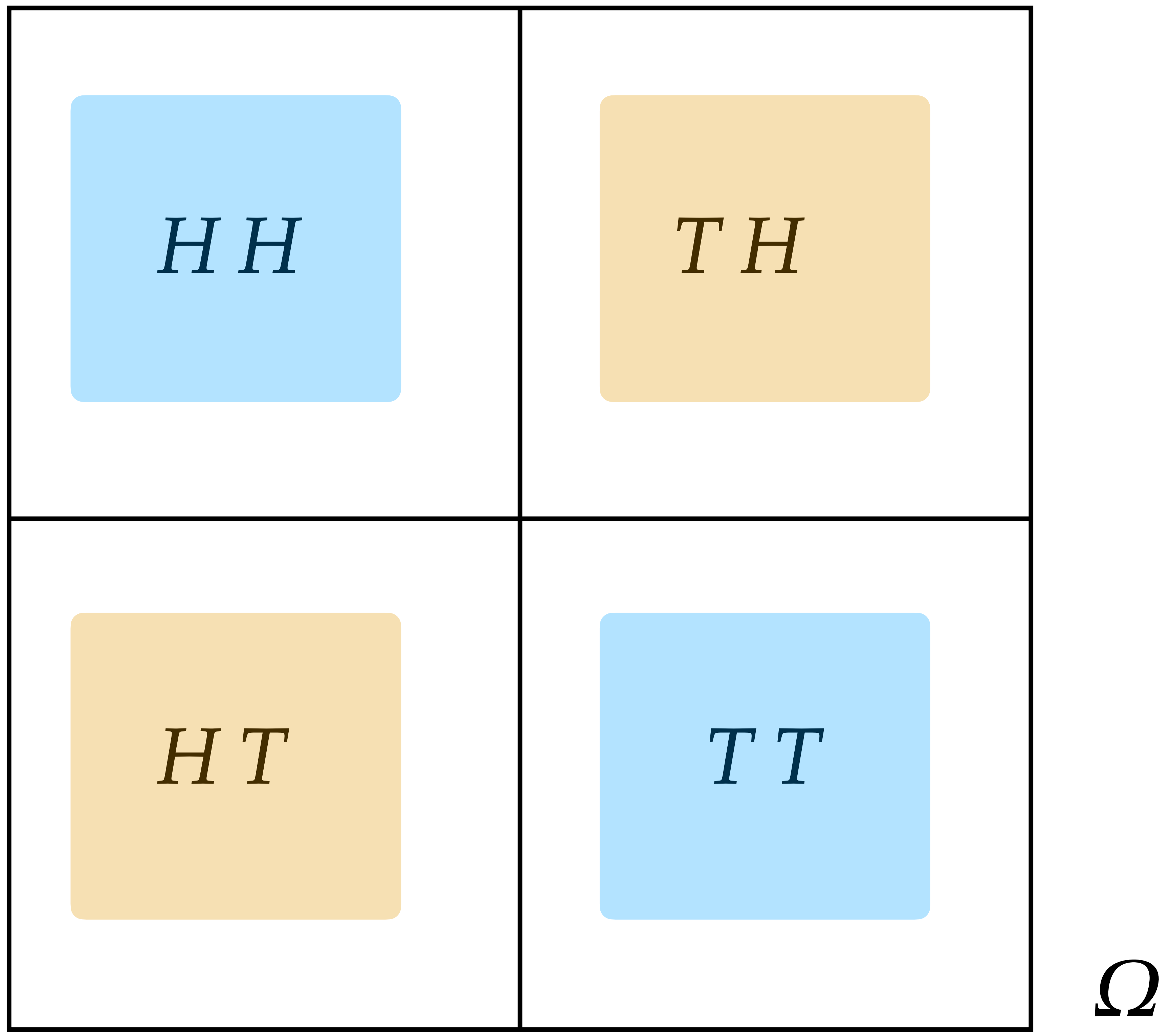
$$A \cap B \text{ (intersection)}$$

$$A^C \text{ (complement)}$$

Example:

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

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



# Probability Measure

## Intuition and definition

A probability measure is a set function

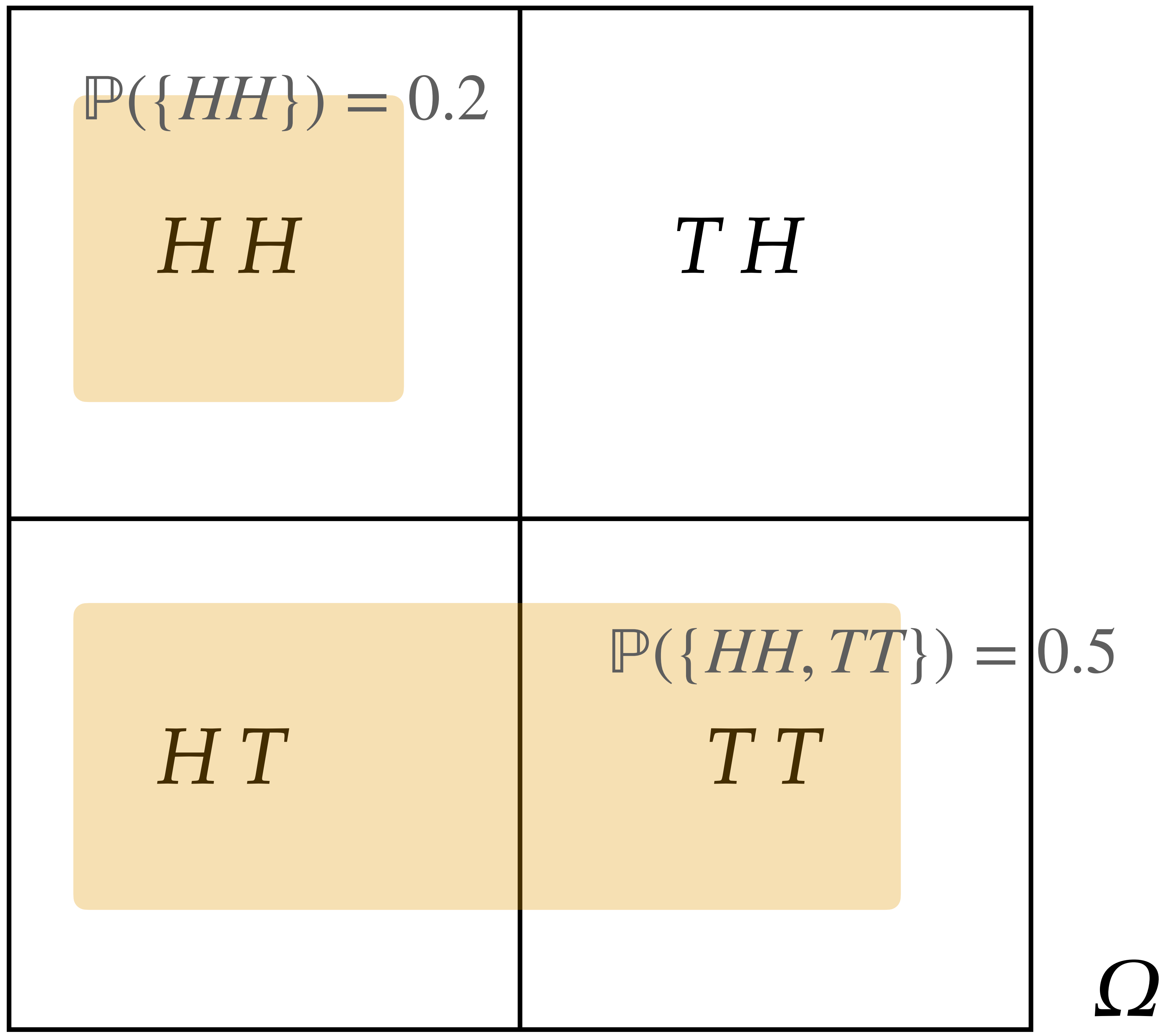
$$\mathbb{P} : \mathscr{A} \to [0,1]$$

mapping from sets to a number in $[0,1]$.

For event $A \in \mathscr{A}$, we call $\mathbb{P}(A)$ the probability of $A$.

*"degree of belief"* or *"long-run frequency."*

Or just the *"mass"* of a particular subset!



# Probability Measure

## Axiomatic Properties

A *valid* probability measure $\mathbb{P}$ satisfies:

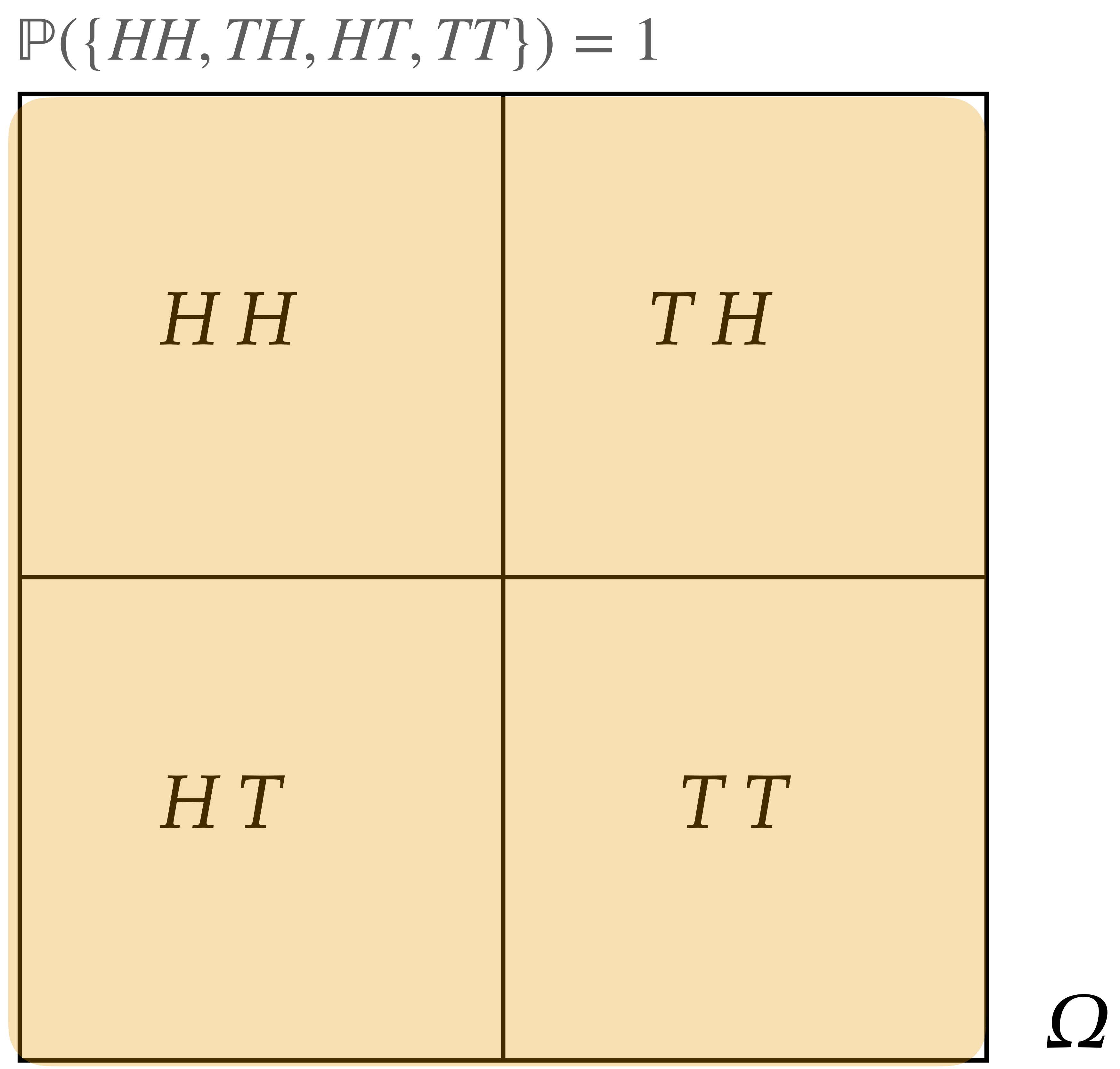
1. The measure of the entire sample space:

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

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

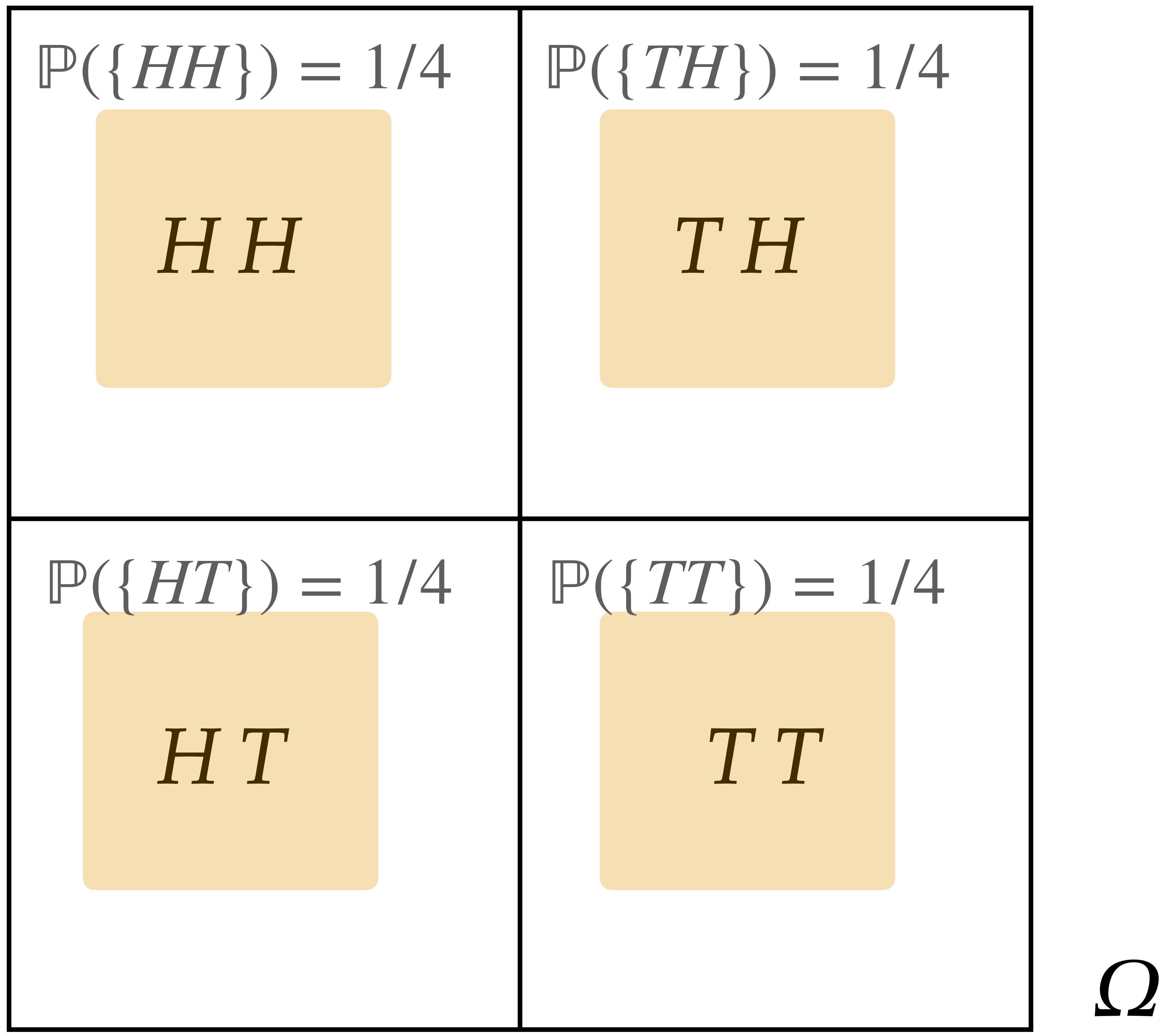
## Example Measures

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

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

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:

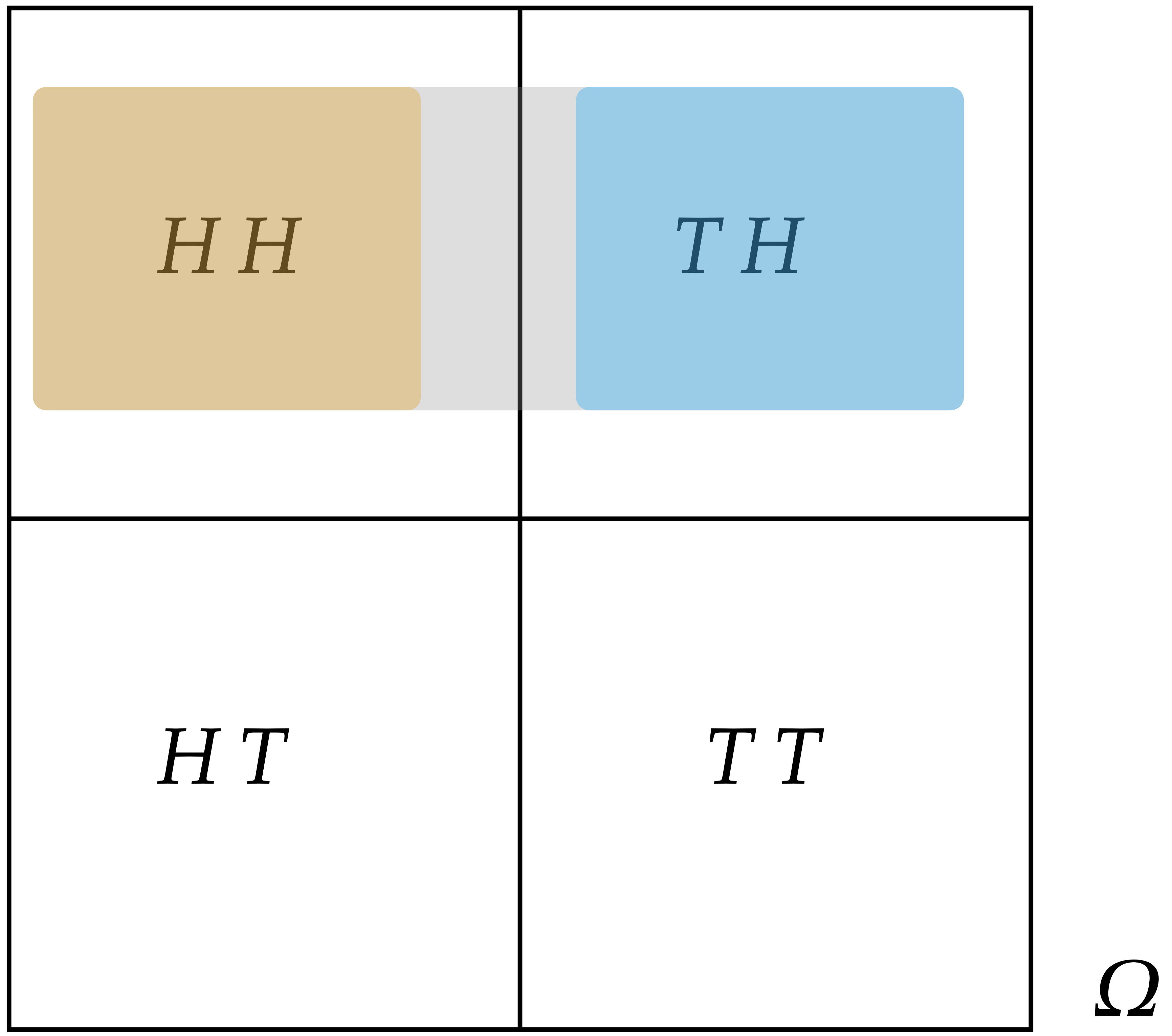
$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Example:

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

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

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



# 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

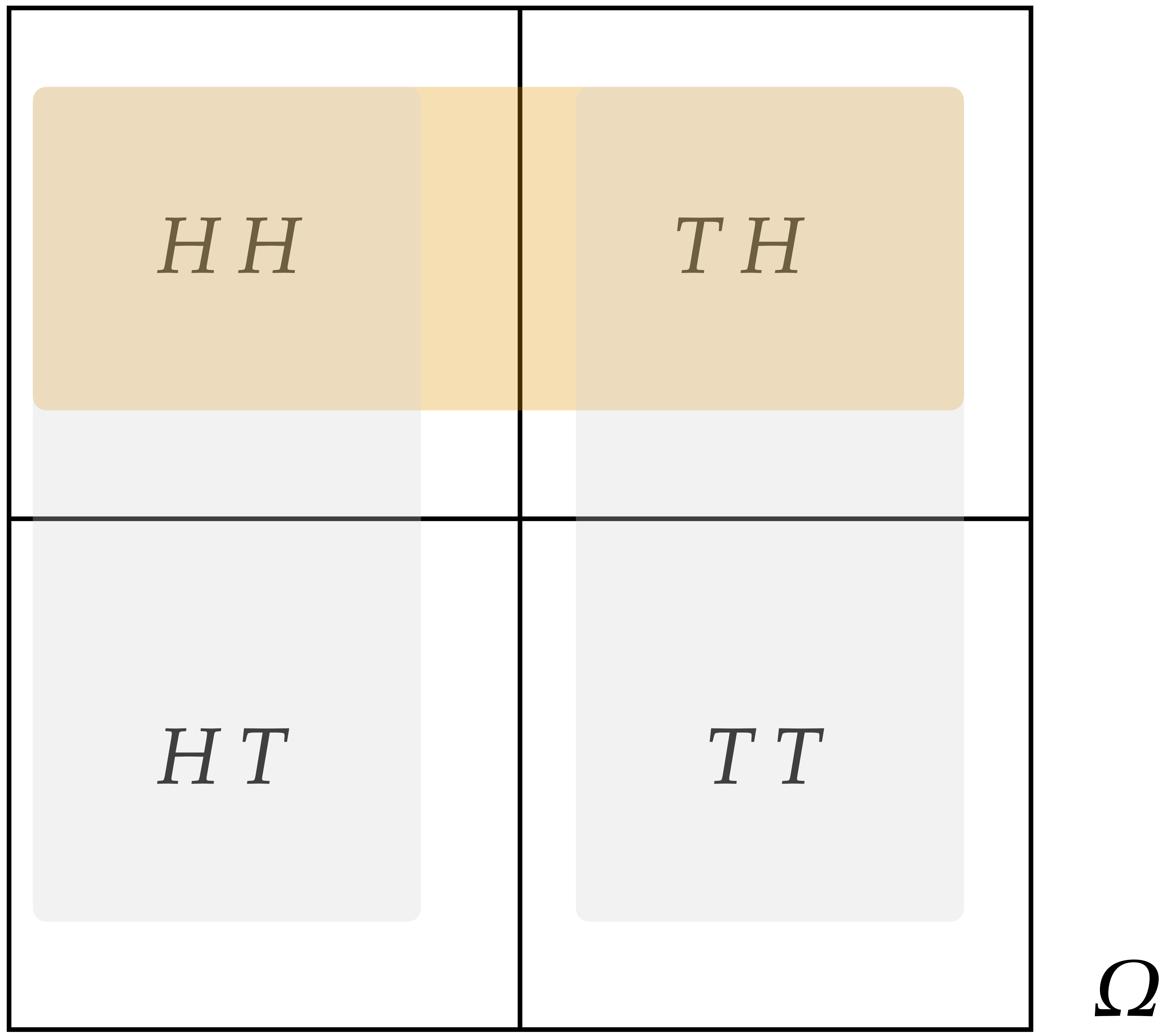
The law of total probability allows us to chop up probabilities into an exact sum of distinct events.

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

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

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

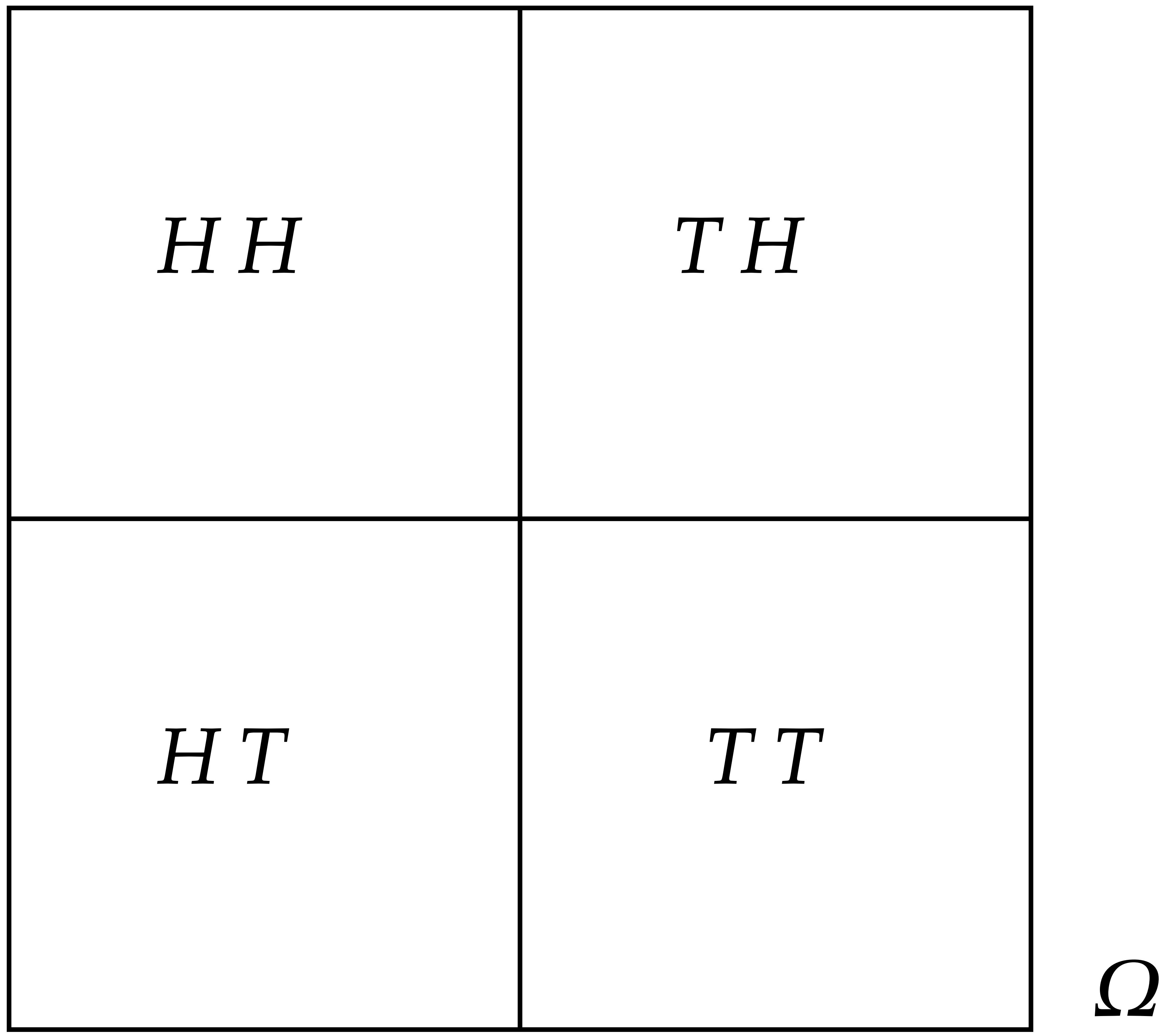
*Super useful: commit this to memory!*



# Probability Space

## Intuition and definition

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




Example:

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

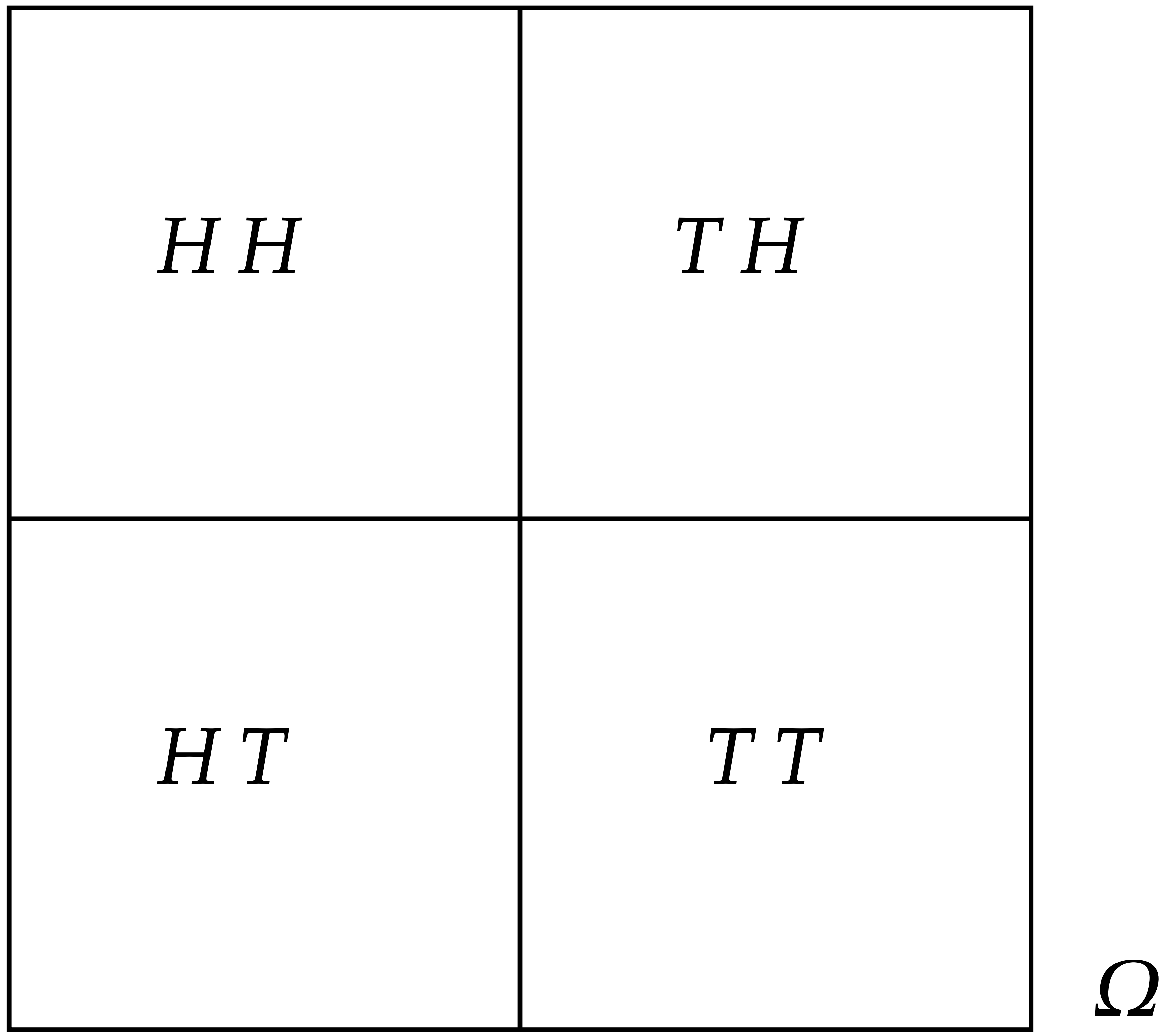
$\mathscr{A} = \{\varnothing, \{HH\}, \{HT\}, \dots, \{HH, HT, TH, TT\}\}$

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

# Probability Space

## Intuition and definition

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




*We avoid dealing with these directly! Instead, we use random variables as an "interface."*

# Random Variables

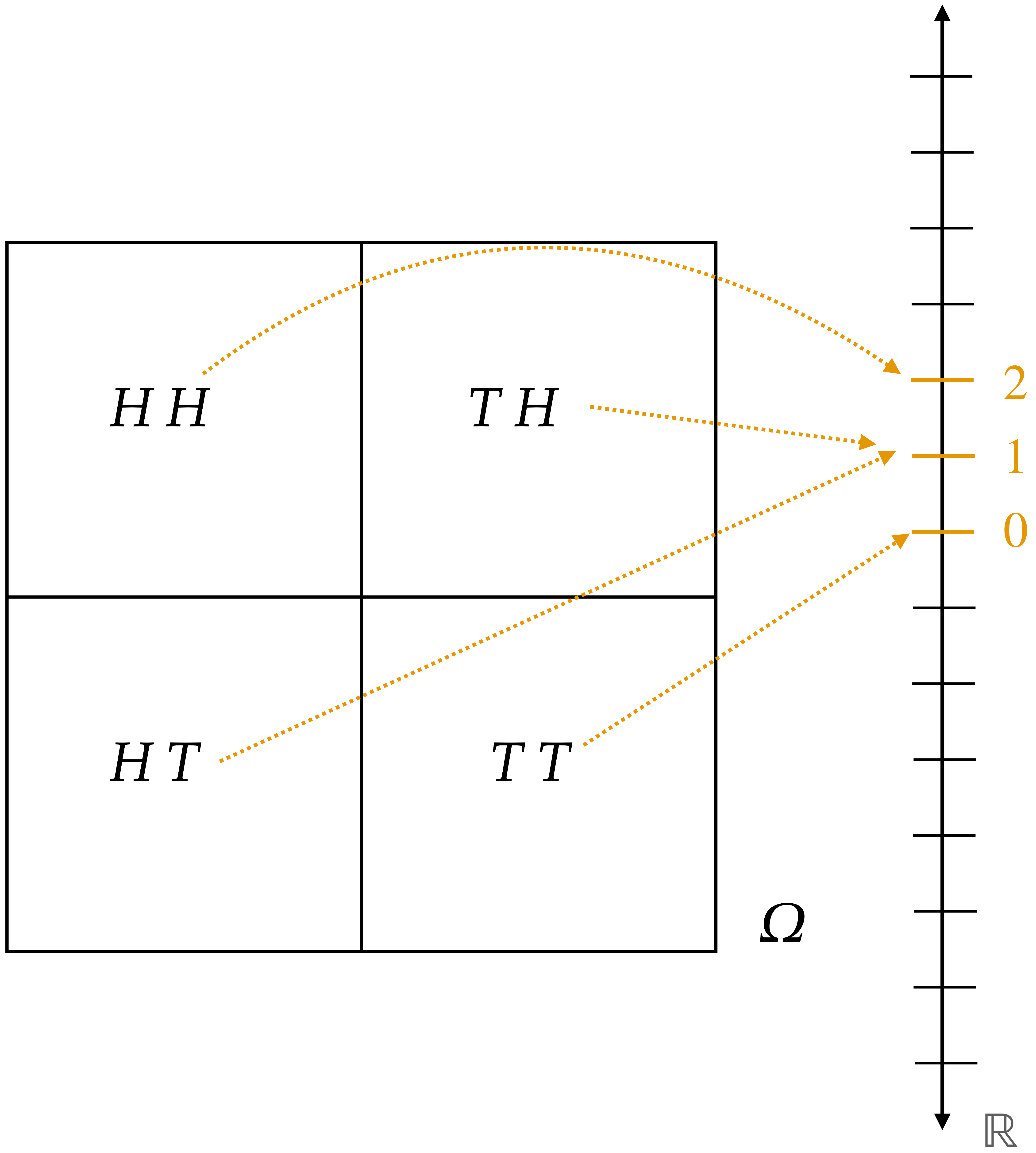
## Example: Flipping 2 fair coins

Consider the following function:

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

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

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



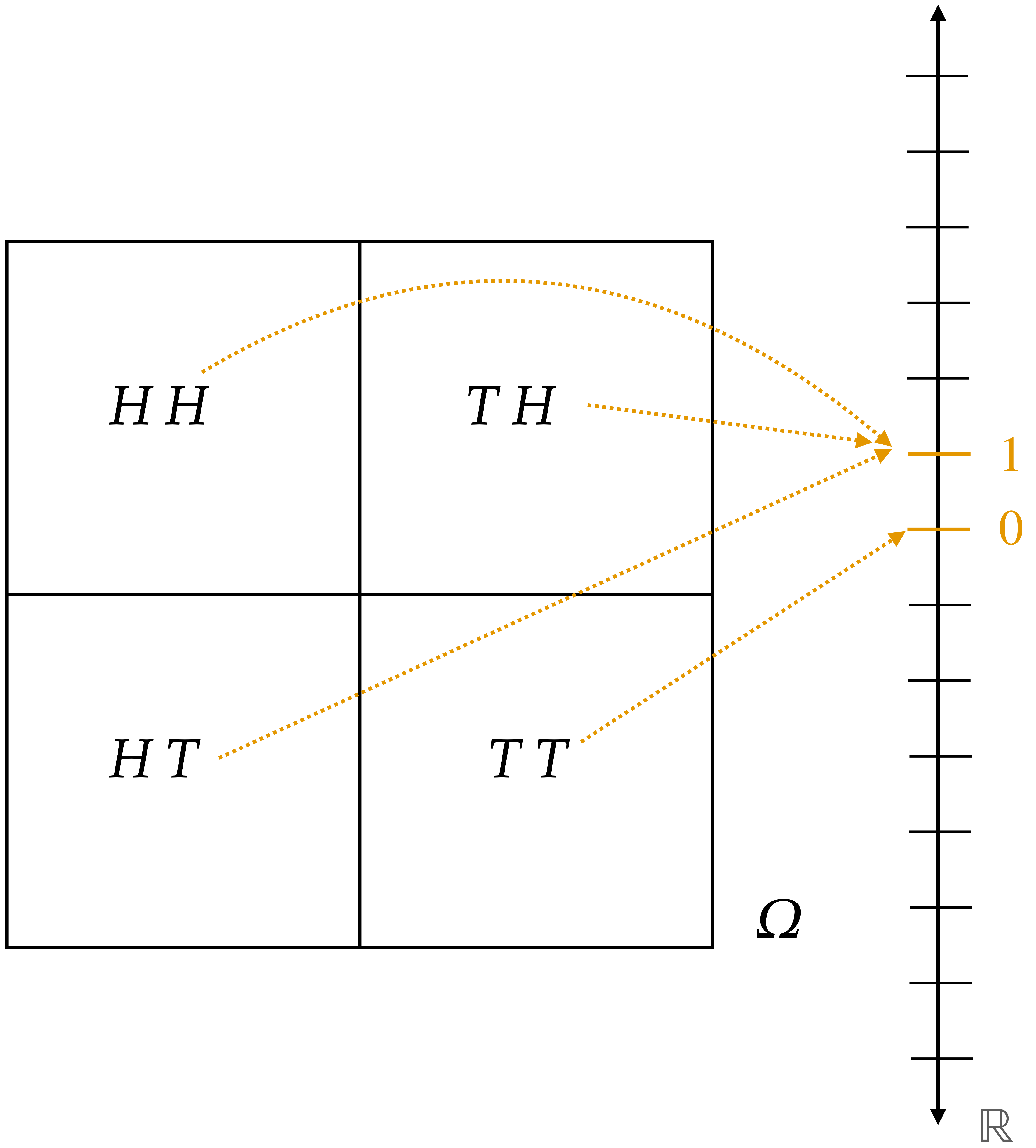
# Random Variables

## 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 number to *every* outcome in the sample space.

# Random Variables

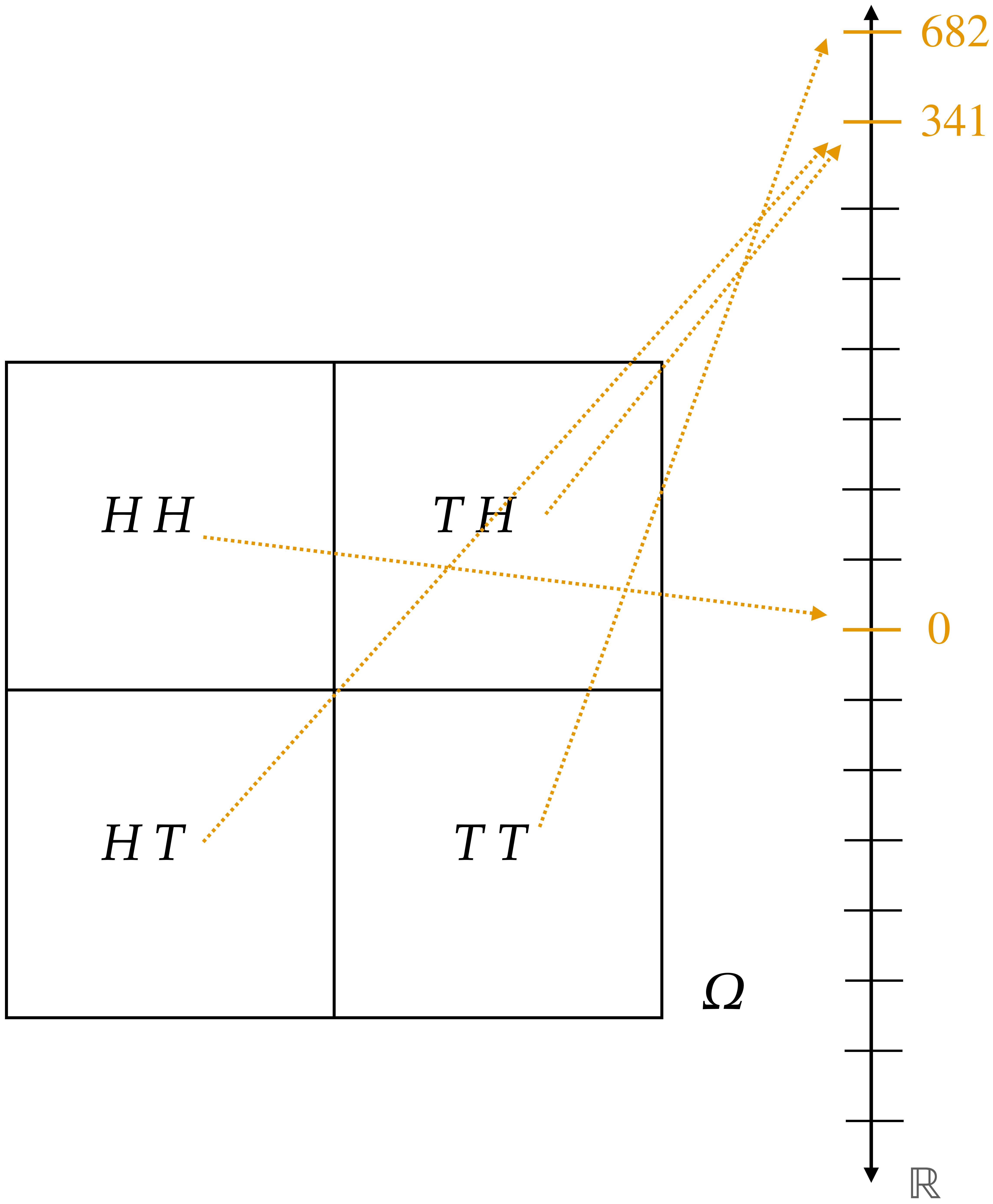
## Example: Flipping 2 fair coins

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 number to *every* outcome in the sample space.

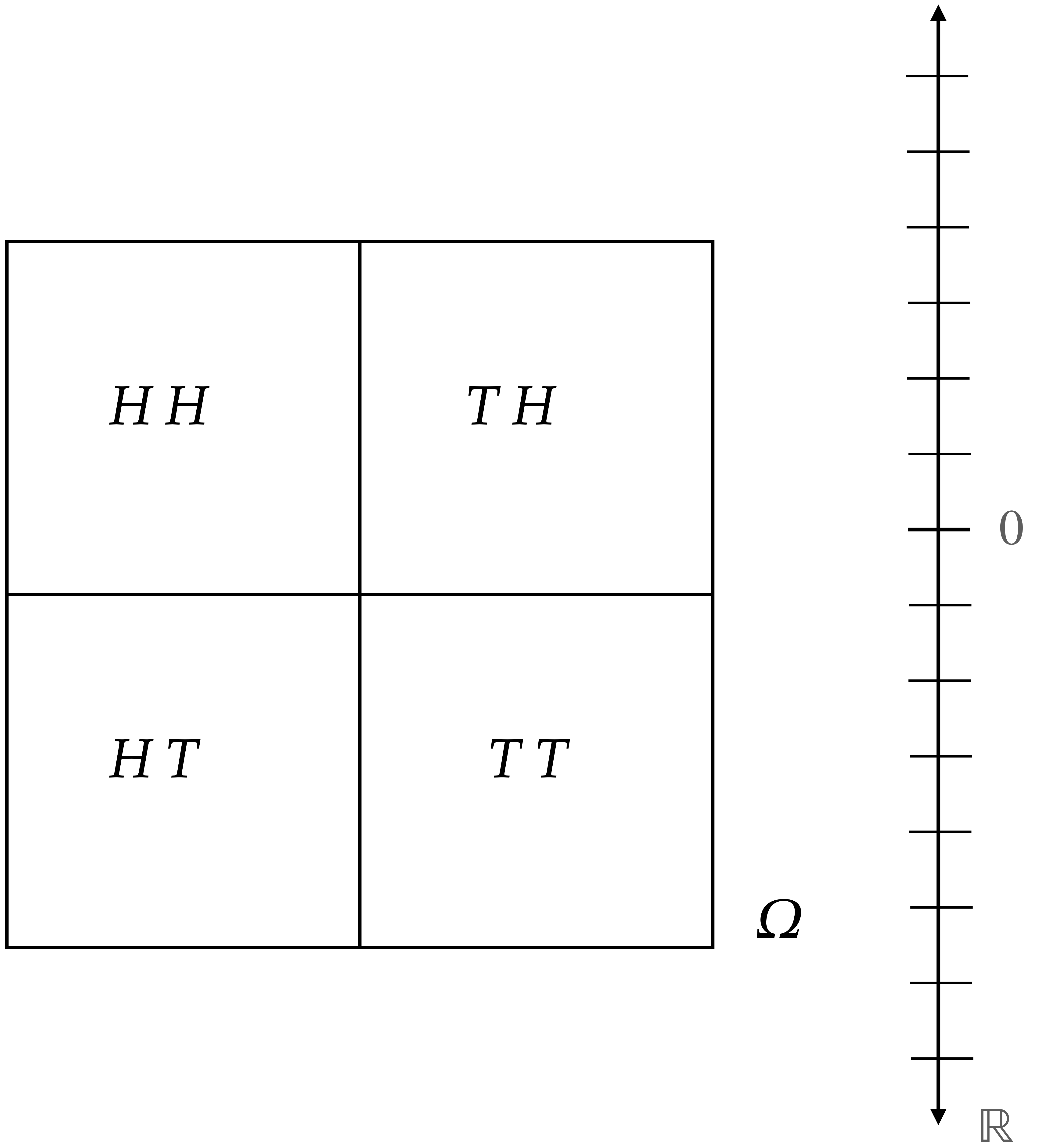


# Random Variable

## 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.

We 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

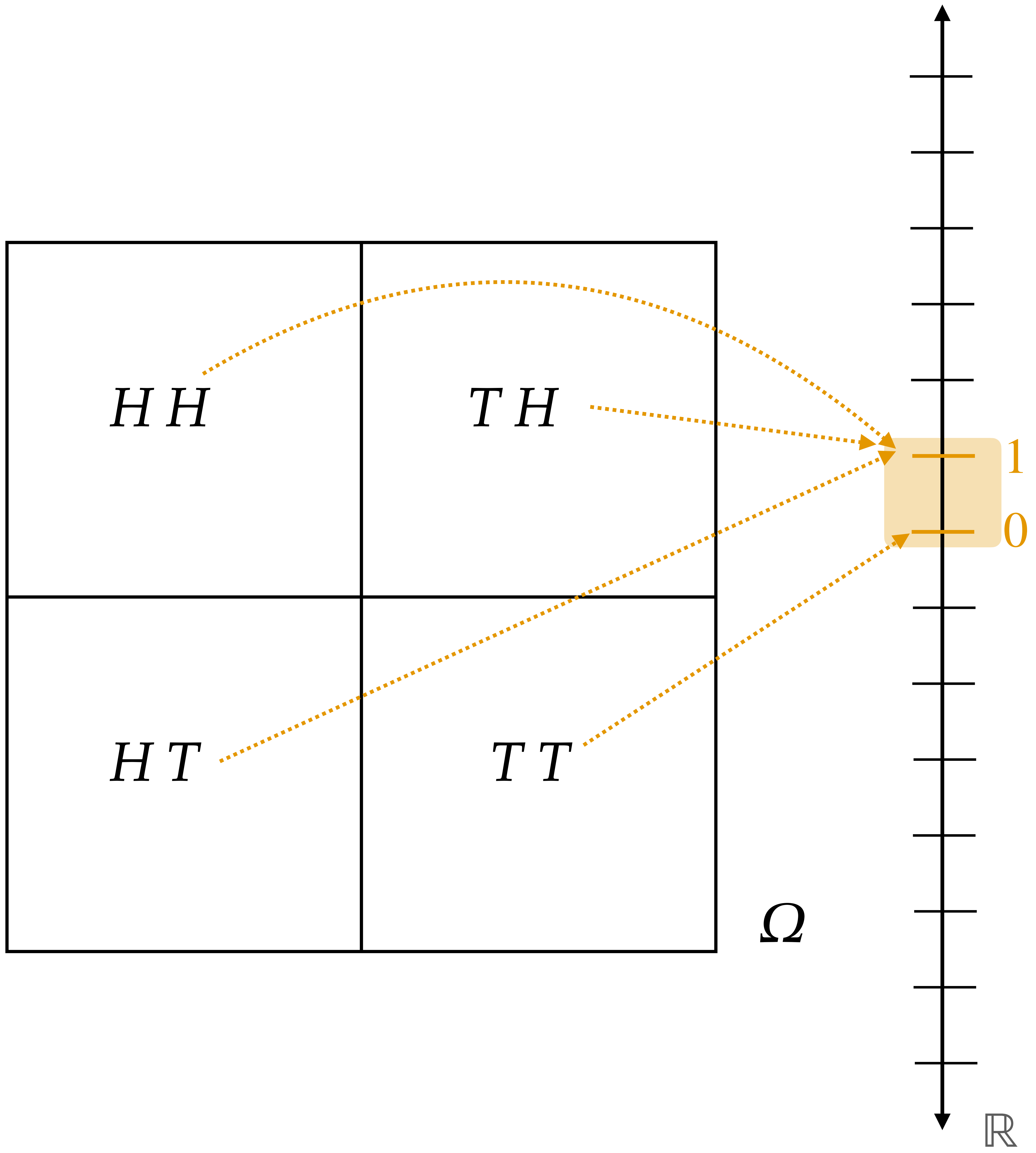
$$X(\omega) = \text{\# of heads, } H.$$

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

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

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

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



# Random Variable

## Intuition and definition

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

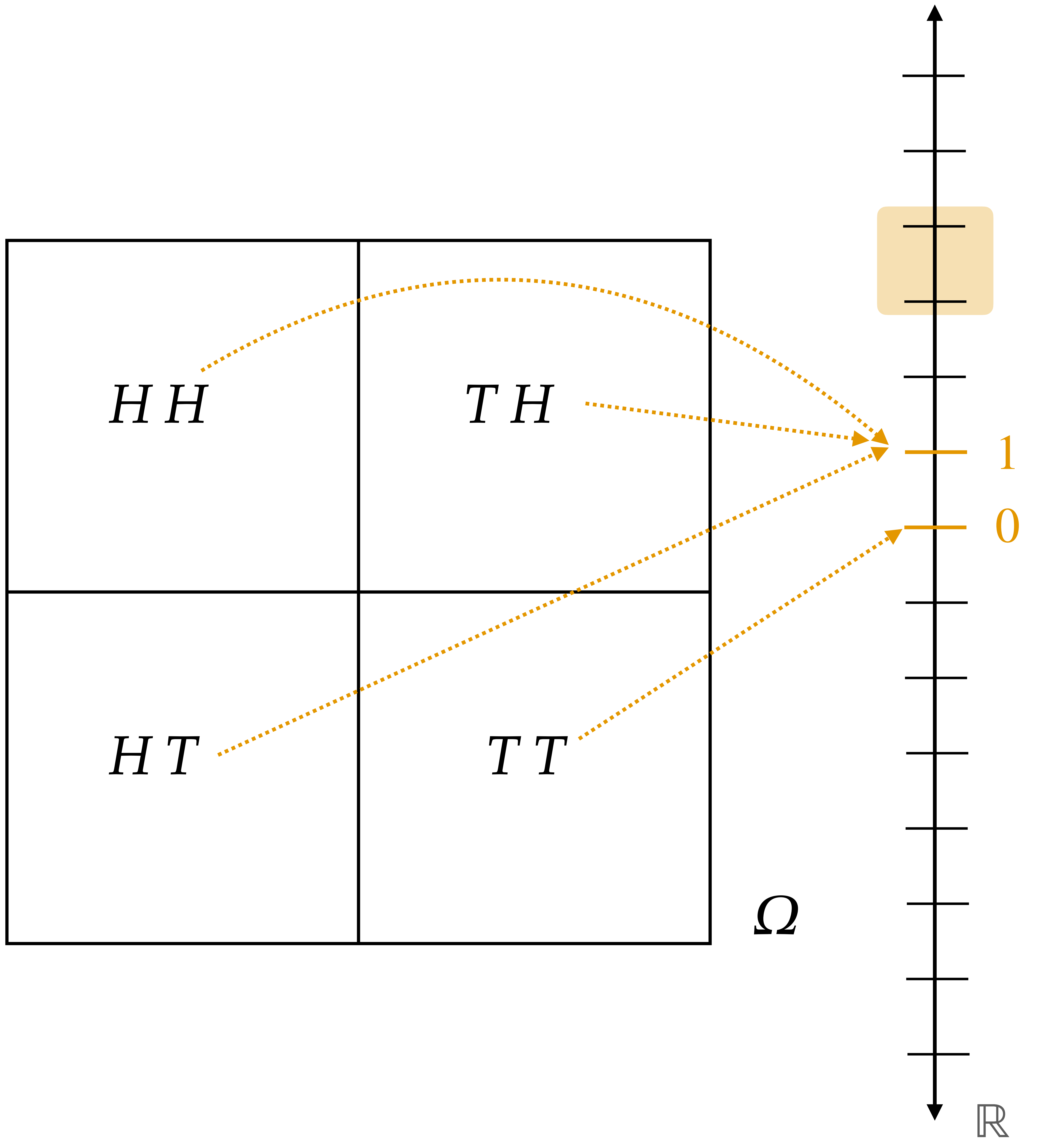
$$X(\omega) = \text{\# of heads, } H.$$

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

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

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

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



# Random Variable

## Intuition and definition

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

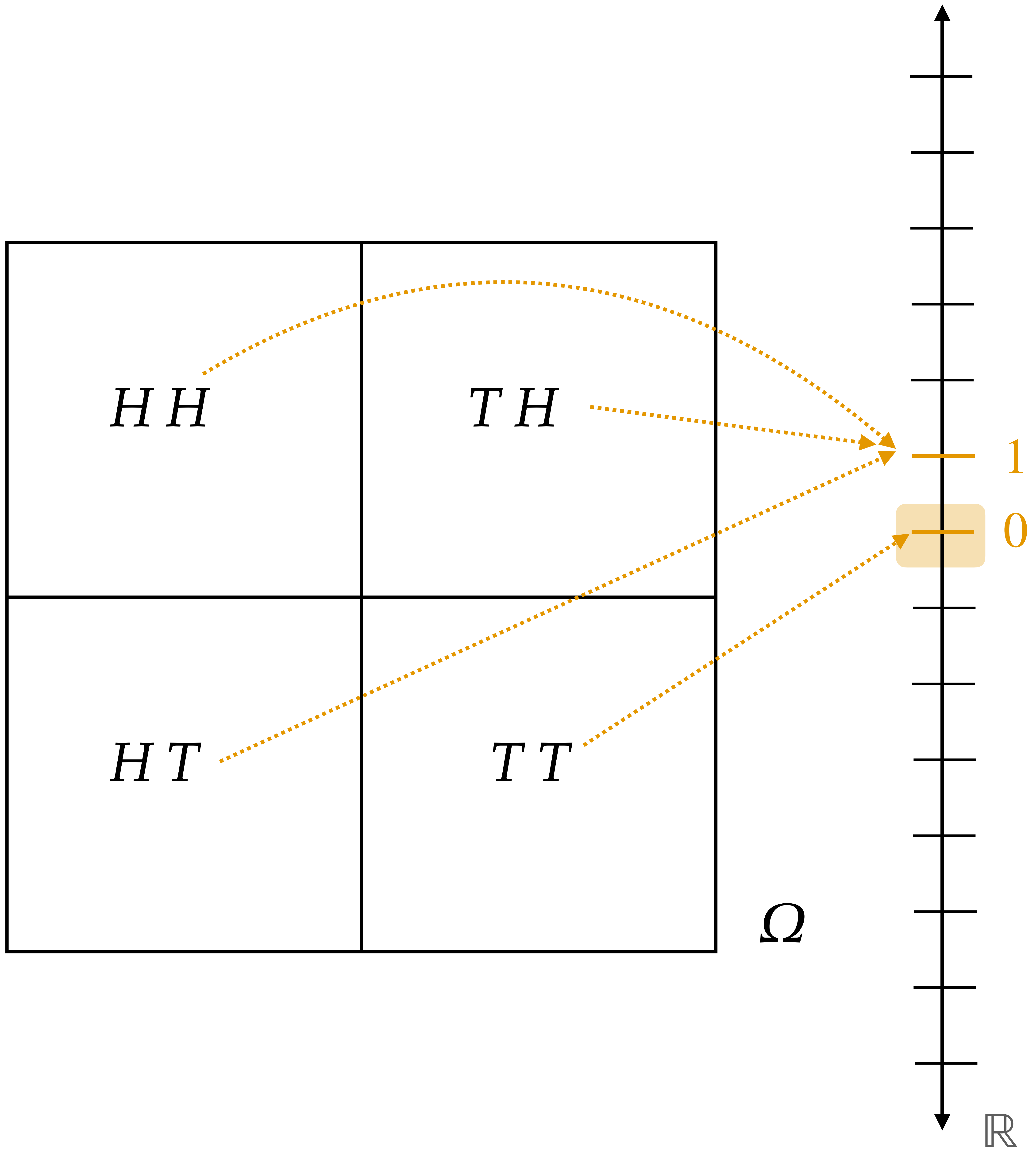
$$X(\omega) = \#\text{ of heads, } H.$$

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

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

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

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



# Random Variable

## Intuition and definition

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

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

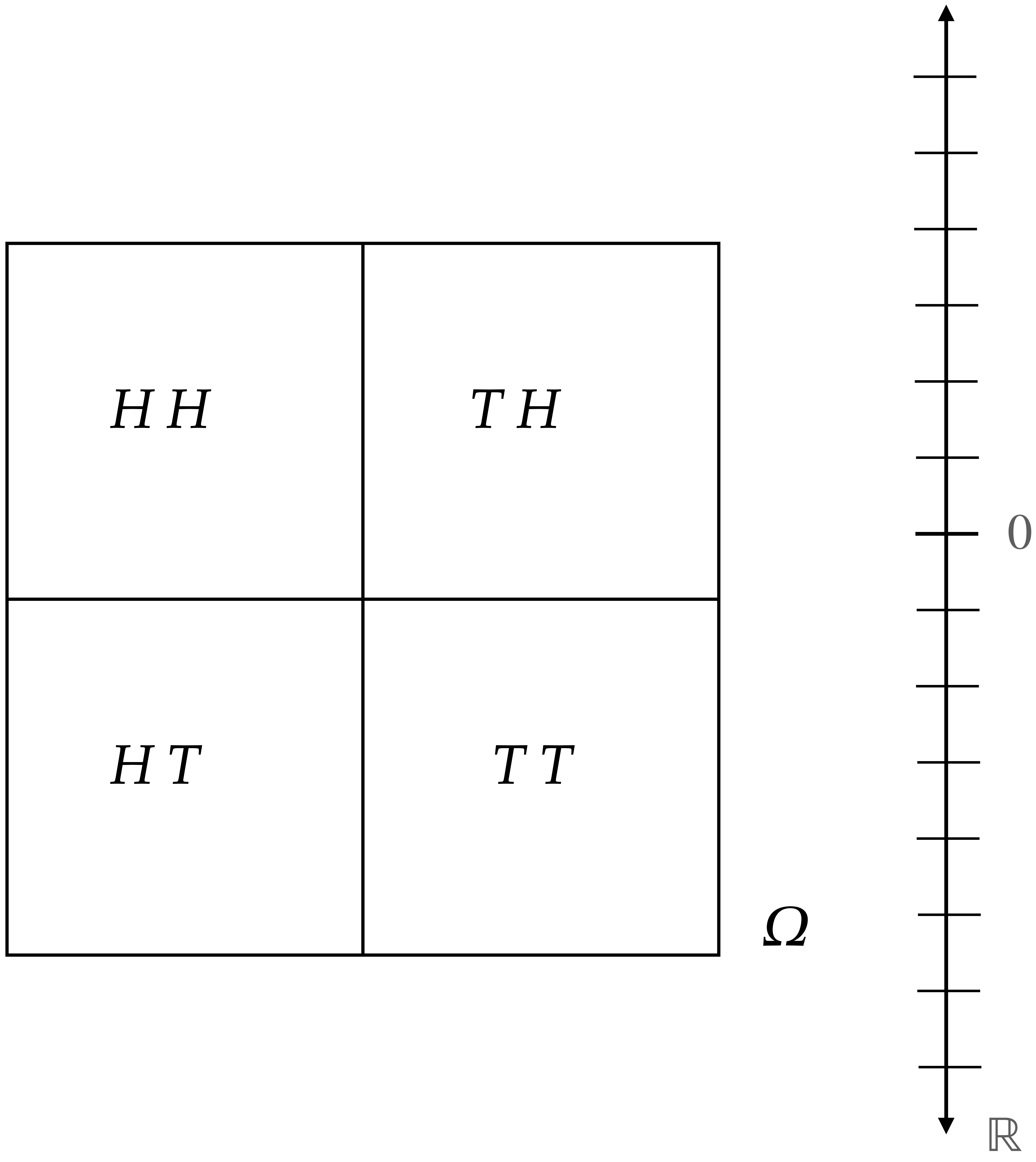
For any $S \subseteq \mathbb{R}$,

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

**Example.**

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

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



# Random Variable

## The distribution of a random variable

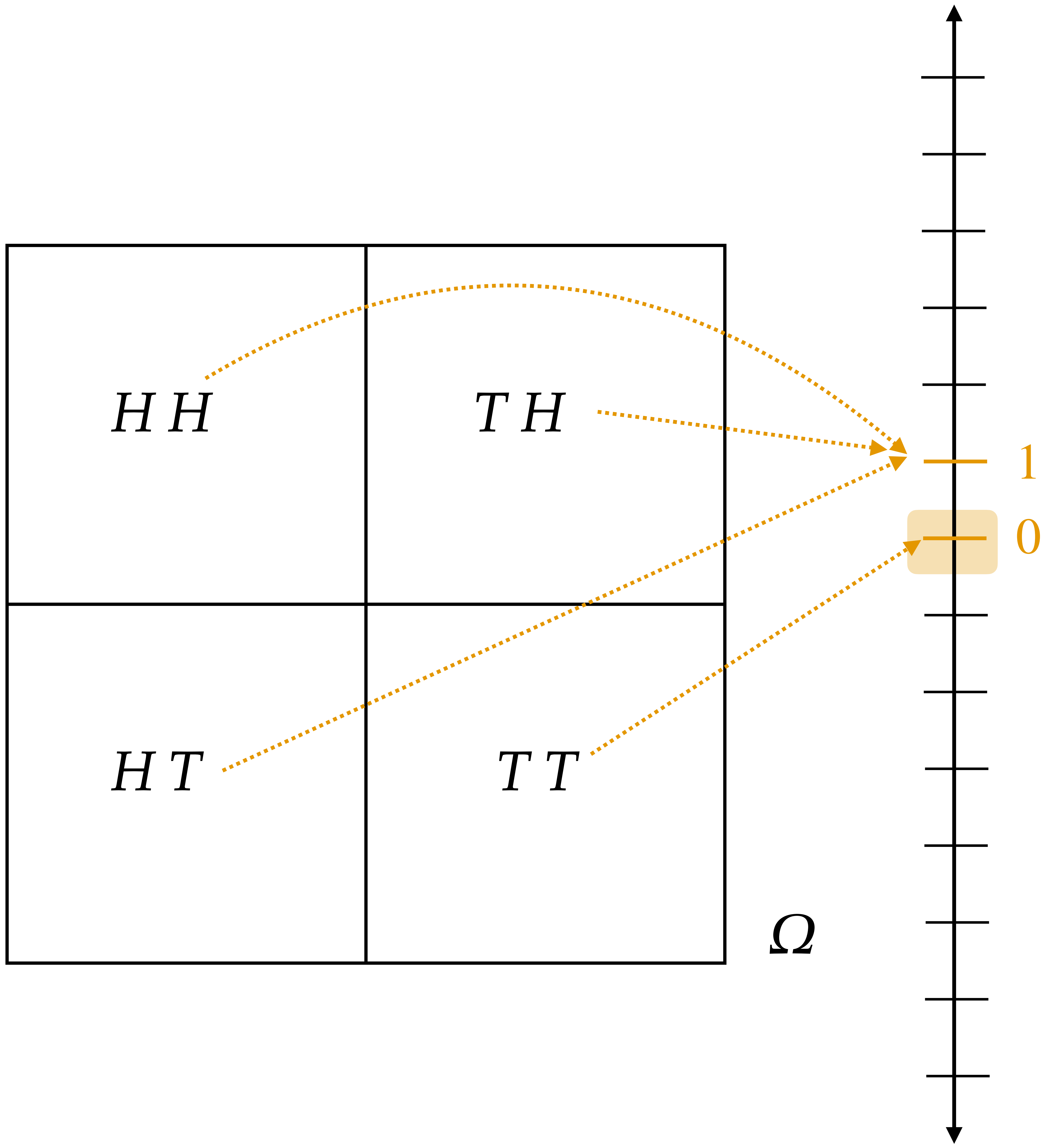
Let $(\Omega, \mathscr{A}, \mathbb{P})$ be some underlying probability space.

Random variables $X : \Omega \to \mathbb{R}$ come with a distribution/law, $\mathbb{P}_X$. This implicitly defines a 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 ($\sigma$-algebra) is some collection of events:

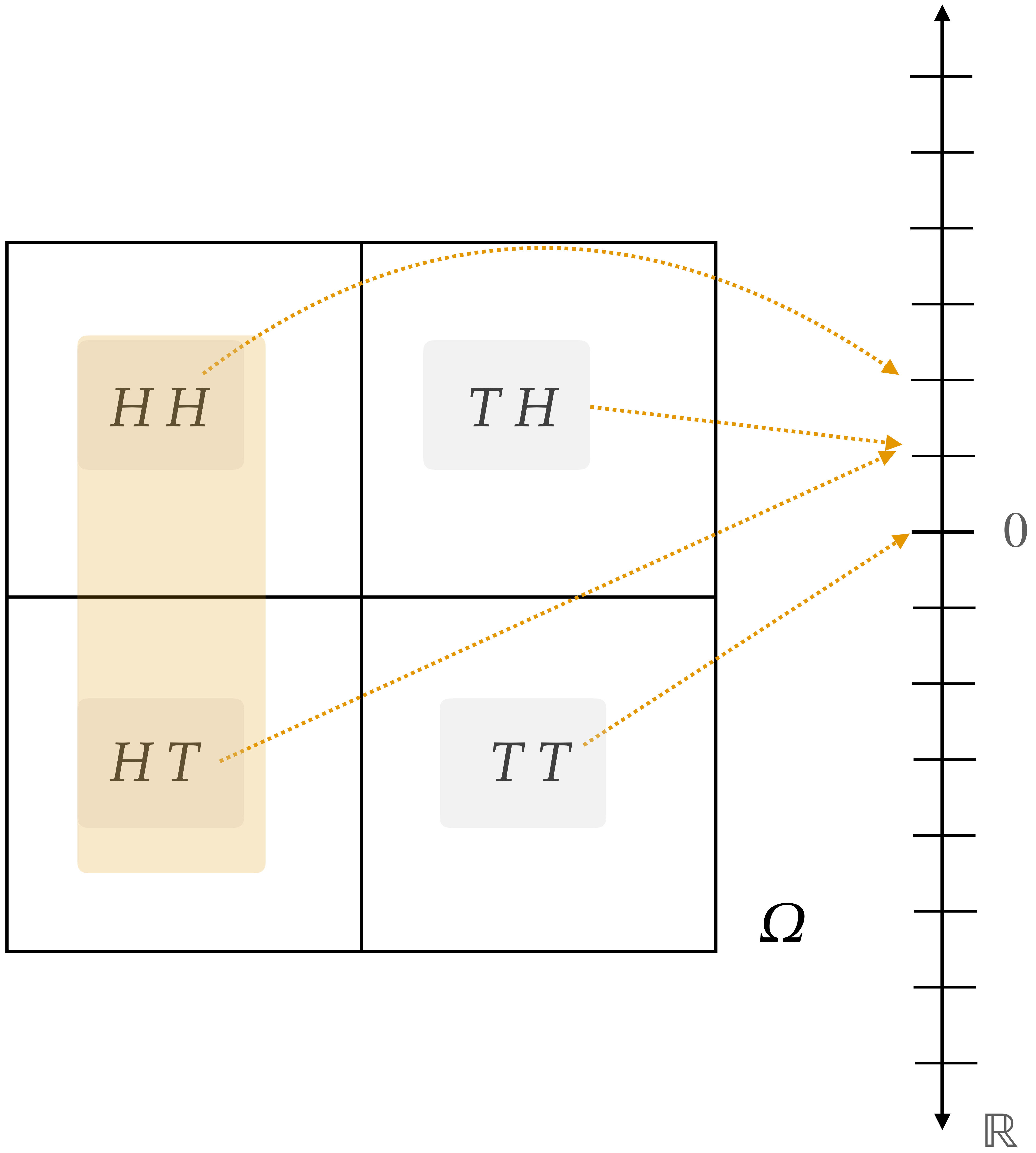
$$\mathscr{A} = \{\varnothing, \{HH\}, \{TT\}, \ldots, \{HH, HT, TH, TT\}\}$$

The probability measure is how we measure the "mass" of events:

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

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

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



# Probability Spaces

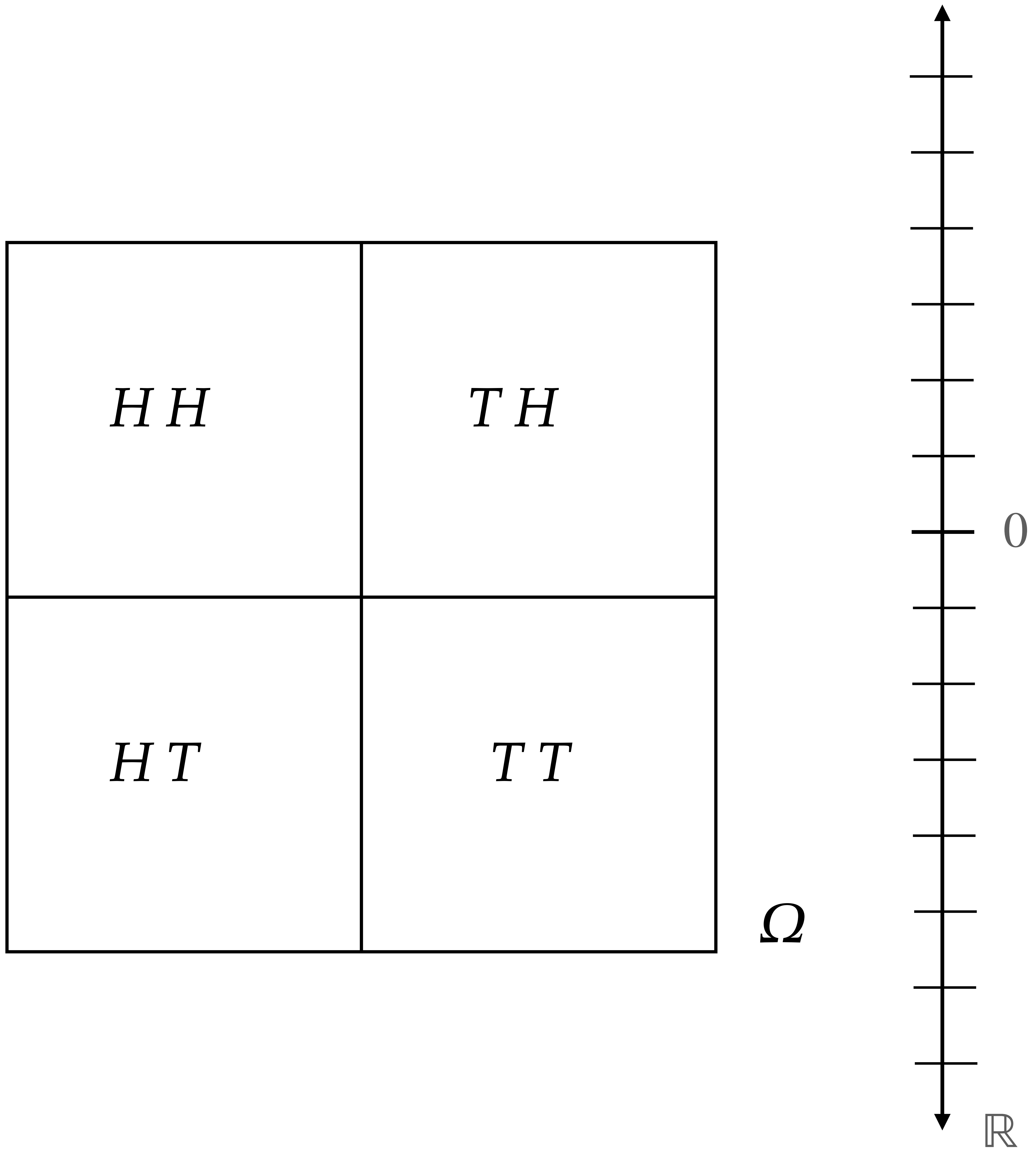
## Putting everything together

Example:

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

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

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



# Random Variables

## Distributions of random variables

# Cumulative Distribution Function

## Intuition and definition

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

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

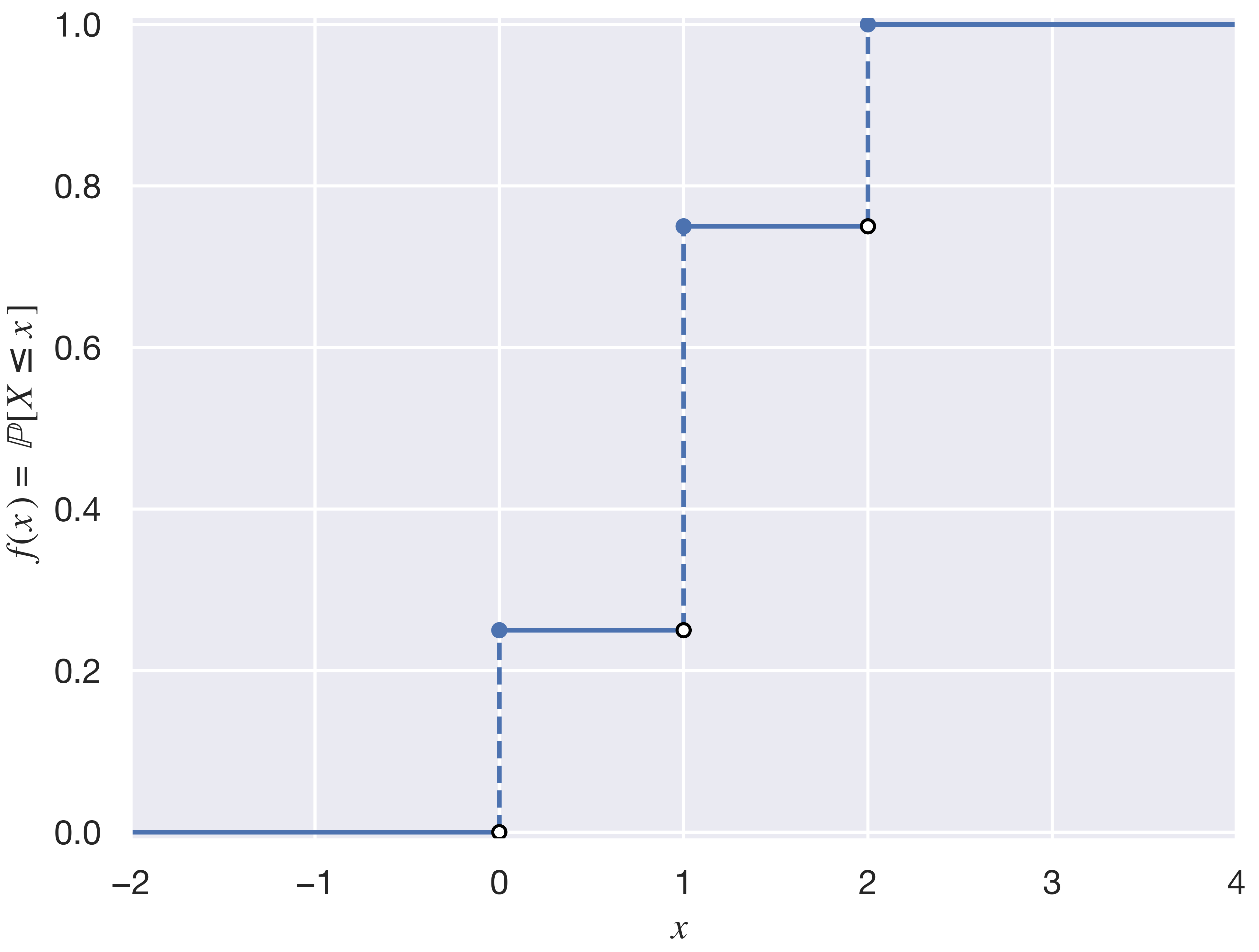
$$F_X(x) = \mathbb{P}(X \leq x)$$

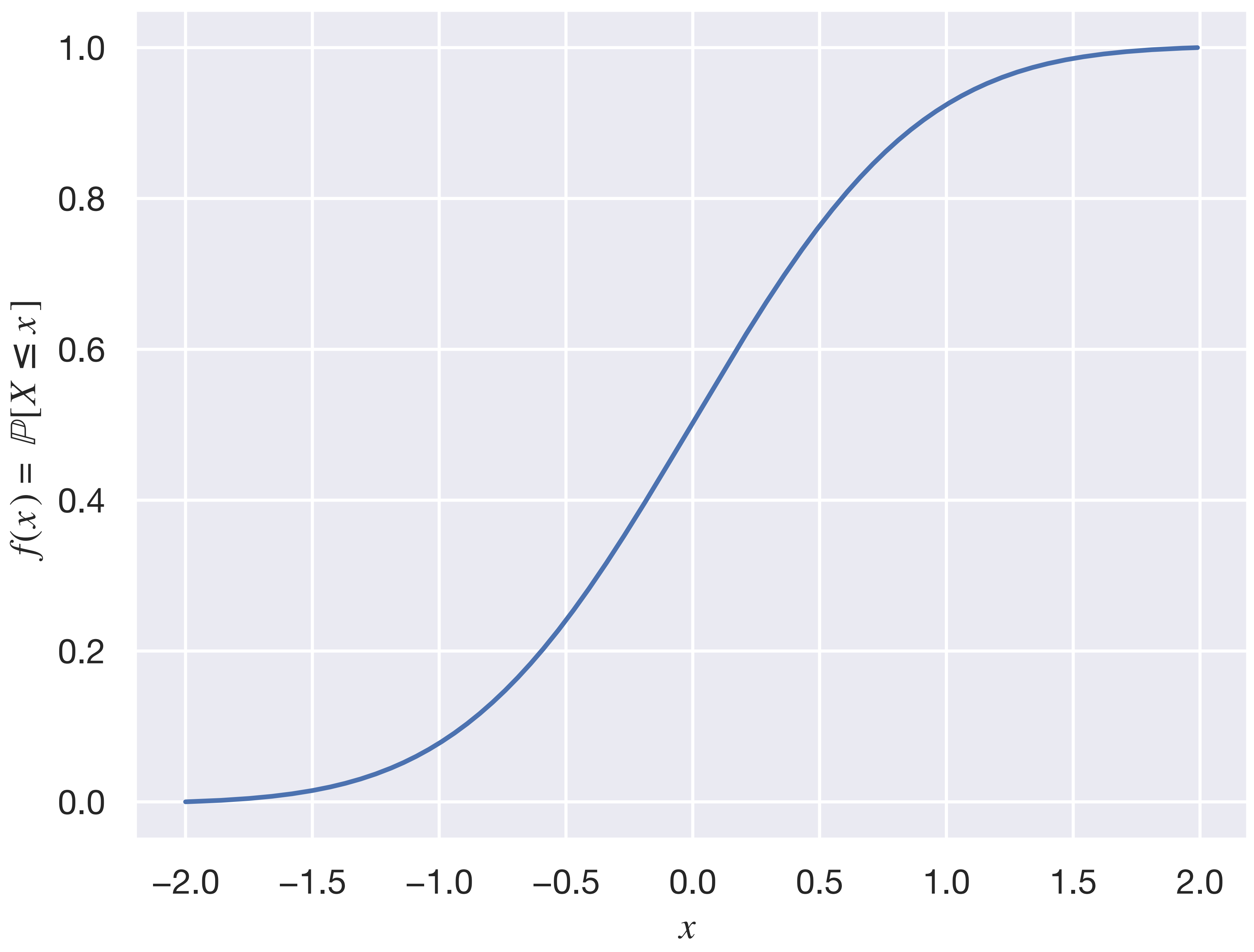
This function allows us to get probabilities in an interval:

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

# Cumulative Distribution Function

## Examples



# Cumulative Distribution Function

## Properties

Right-continuous. For any $a \in \mathbb{R}$, CDF satisfies:

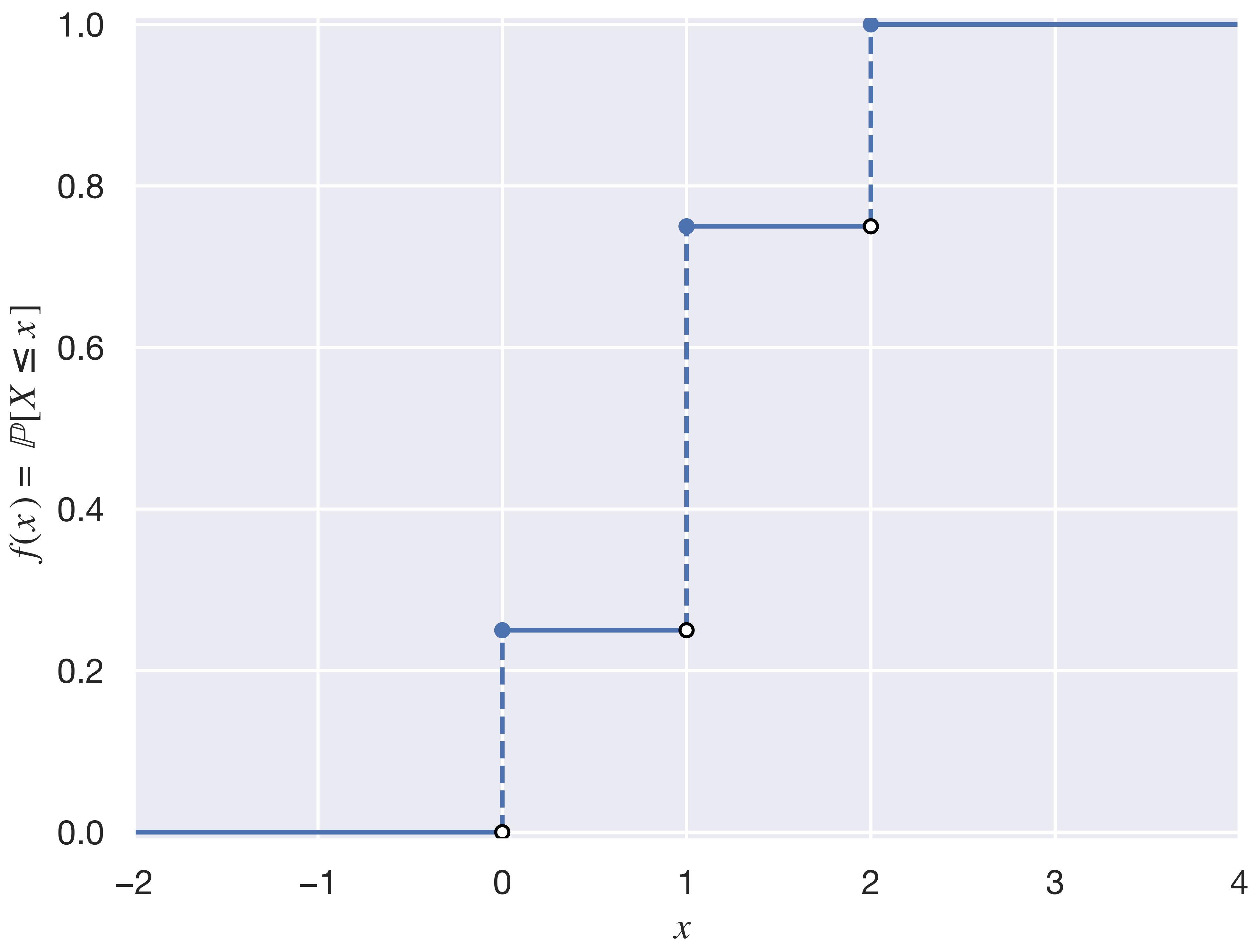
$$\lim_{x \to a+} f(x) = f(a).$$

Monotonically nondecreasing. For every $x \leq y$,

$$F_X(x) \leq F_X(y).$$

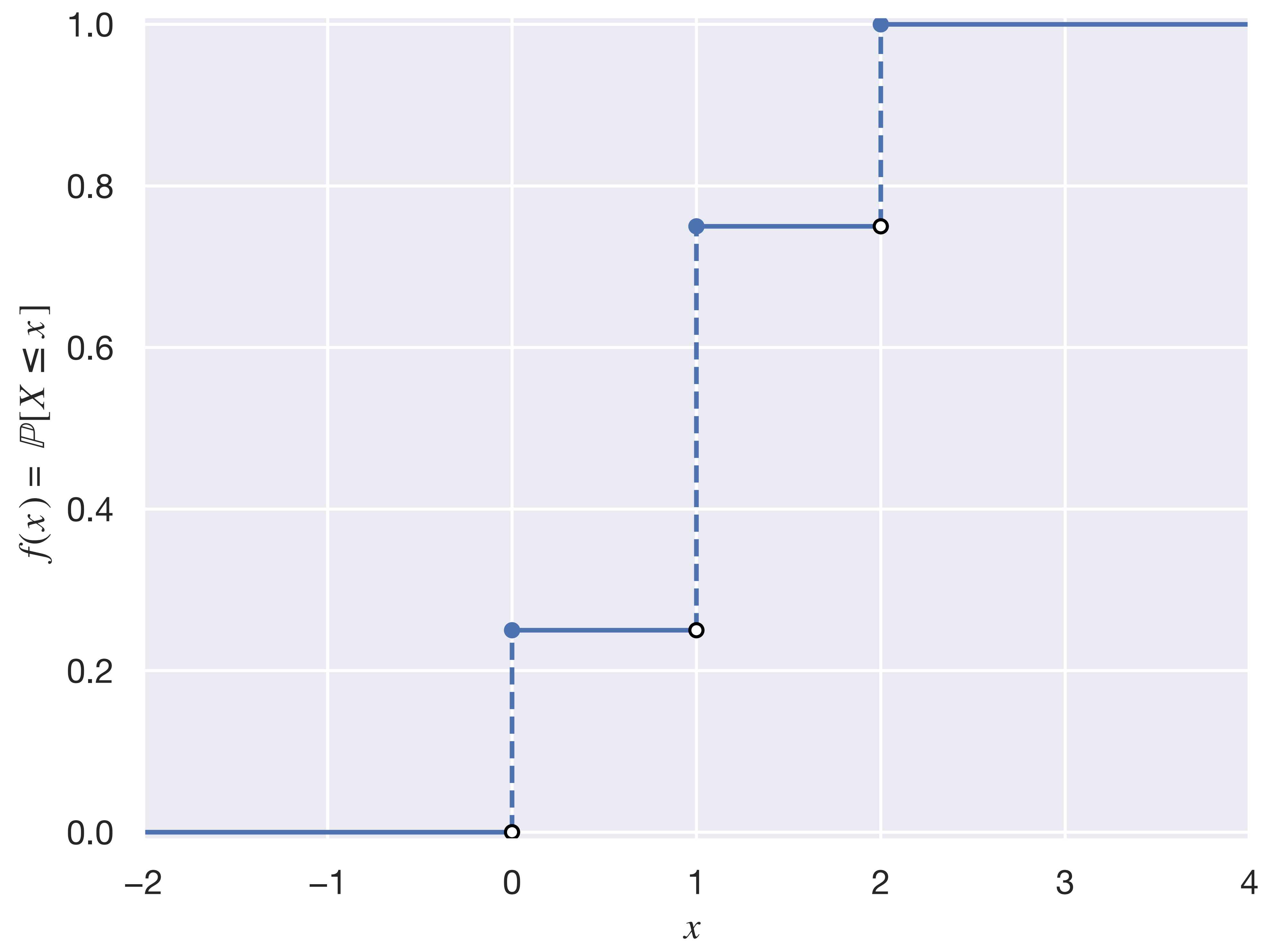
Limits at infinities. The limits at infinity are:

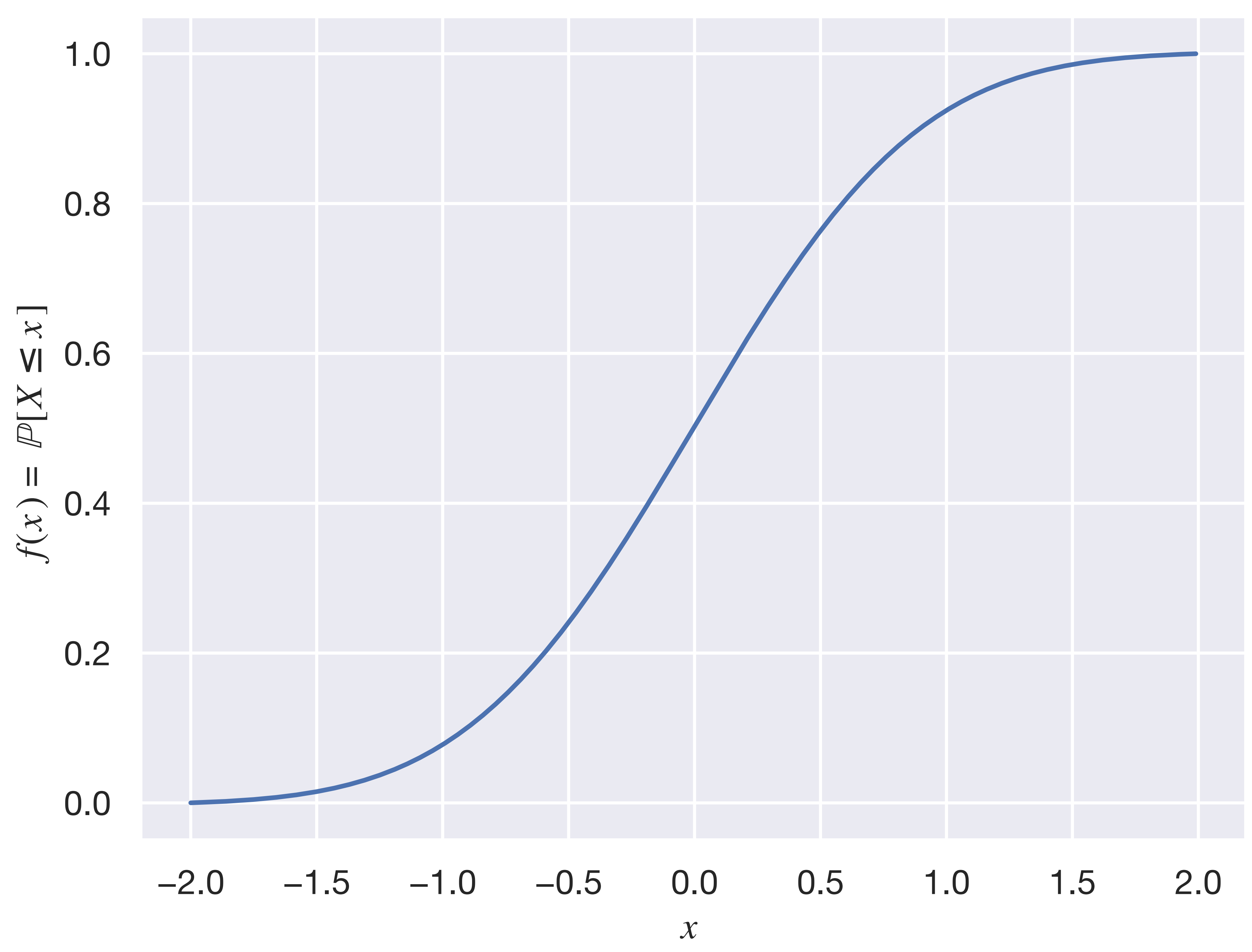
$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to \infty} F_X(x) = 1.$$



# Discrete vs. Continuous RVs

## Difference in CDF



# Discrete Random Variables

## Intuition and definition

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.

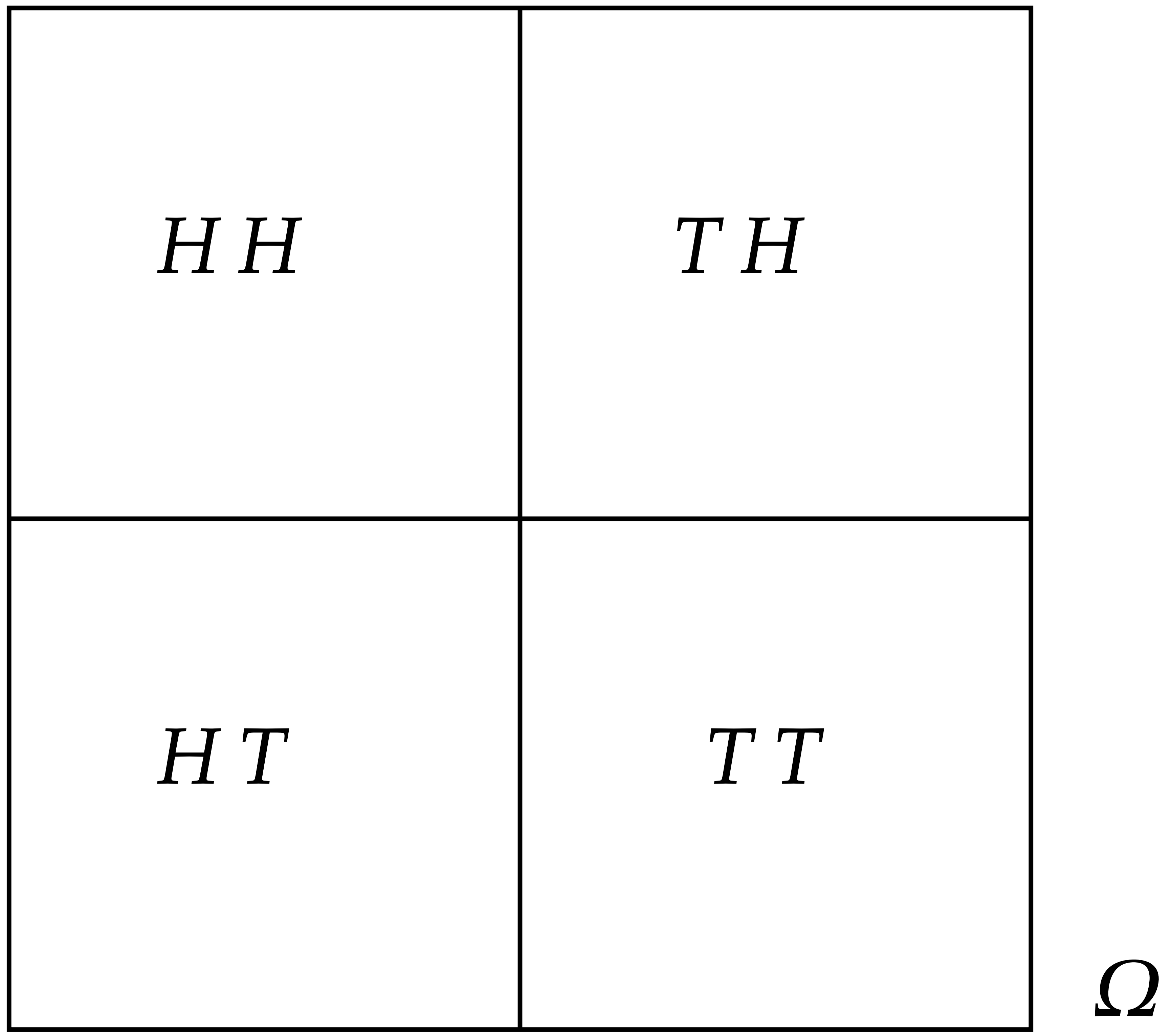
# Discrete Random Variables

## Probability mass function

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

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

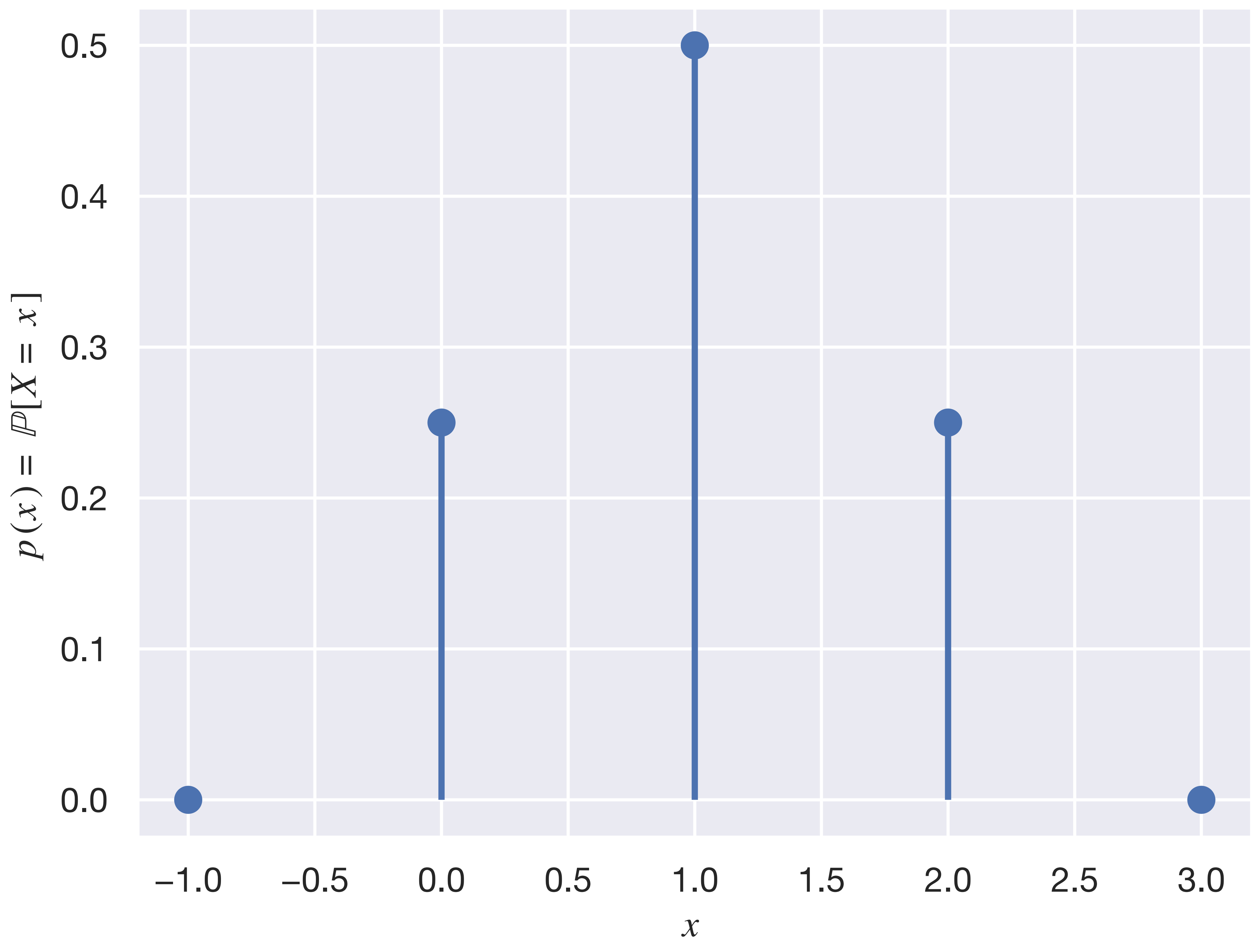
**Example.** What's the PMF of the RV $X : \Omega \to \mathbb{R}$ with $X(\omega)$ counting the number of heads?

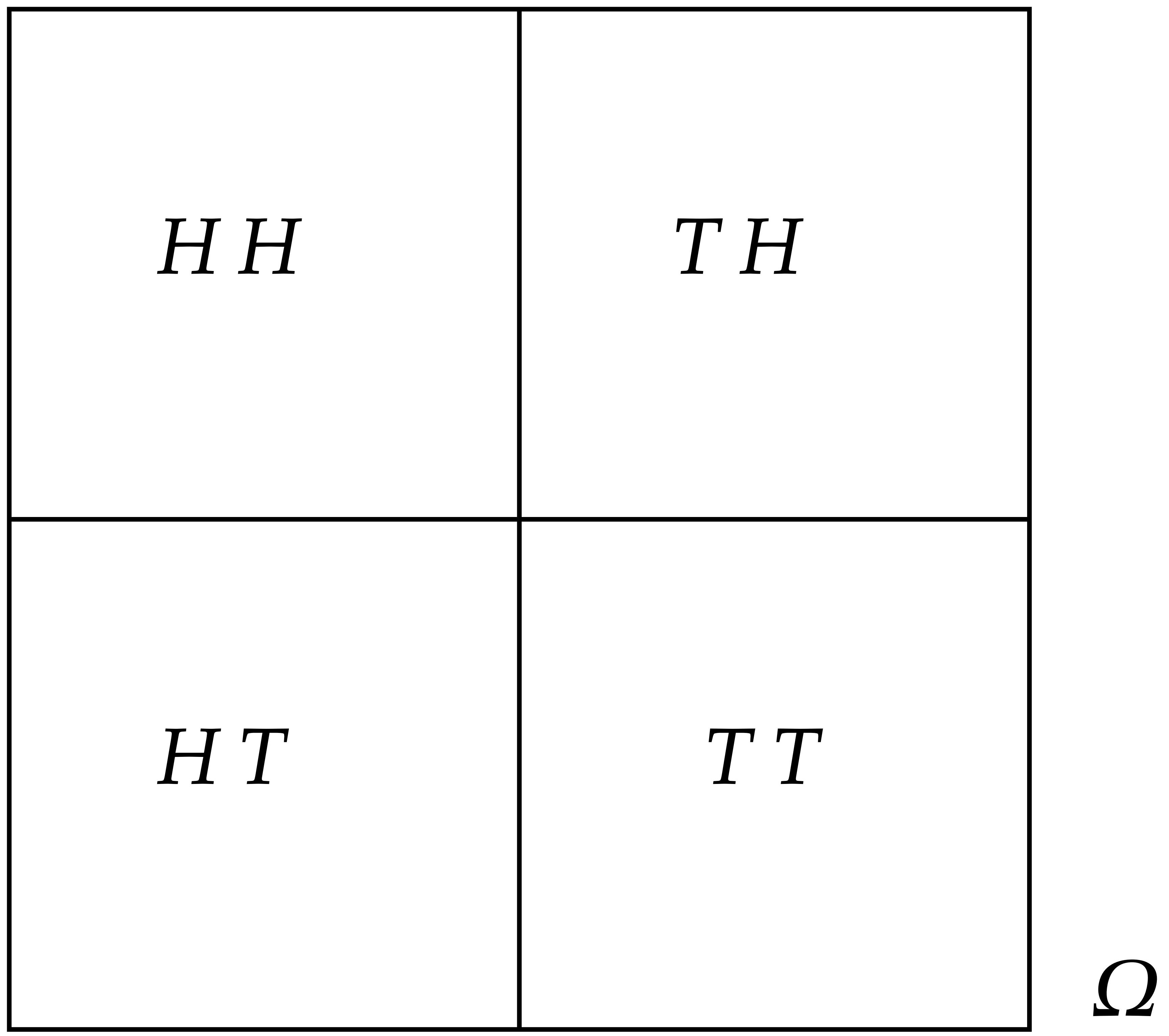


# Discrete Random Variables

## 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) $p_X : \mathbb{R} \to \mathbb{R}$

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

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 immediately give probabilities (unlike PMFs).

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

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

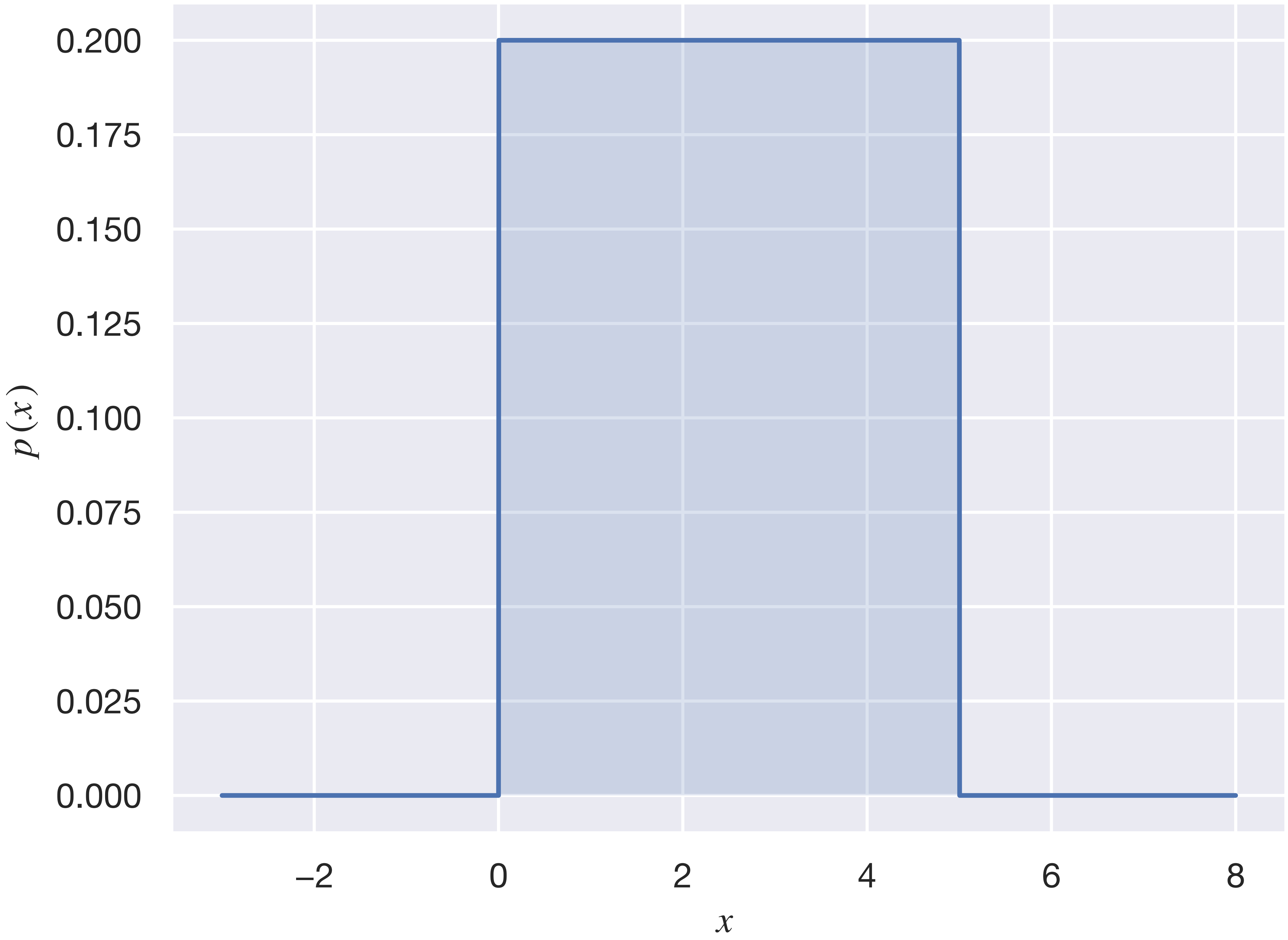
In an infinitesimally small interval, we can get a probability:

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

# Continuous Random Variables

## Example: Picking uniformly in the interval

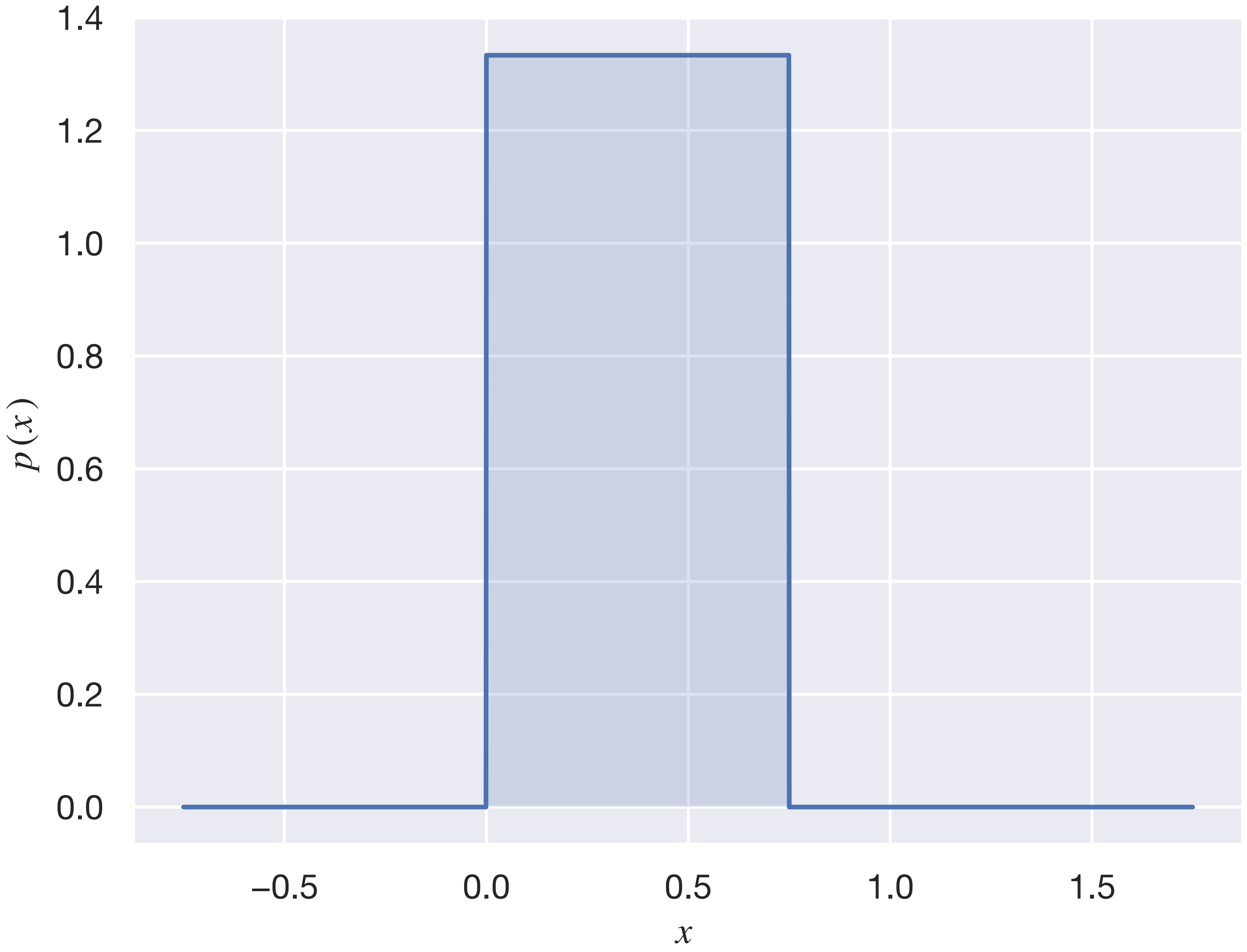
**Example.** What's the PDF of the RV $X : \Omega \to \mathbb{R}$ with the uniform random variable on $[0,5]$?



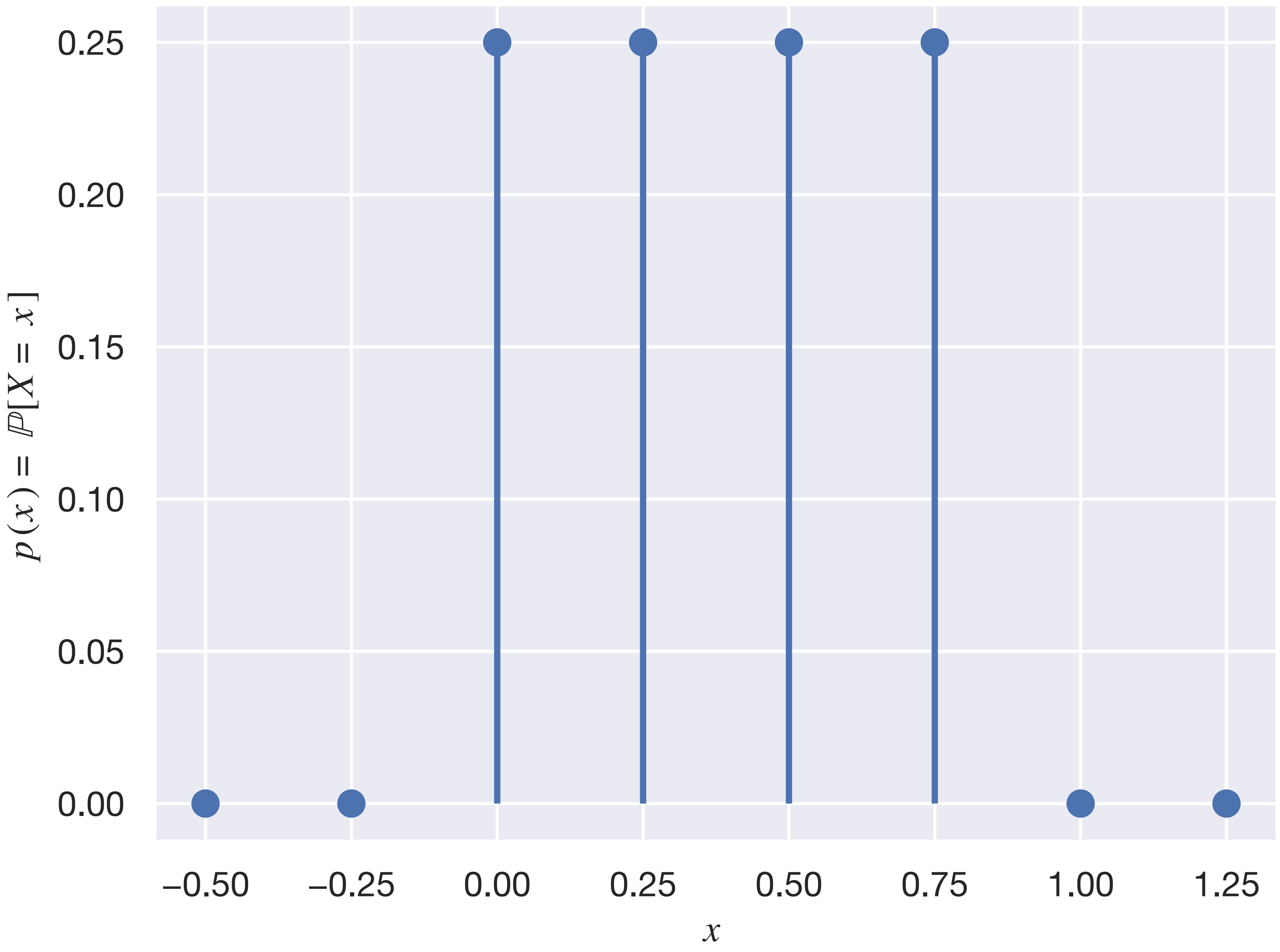
# Continuous vs. Discrete RVs

## Example: Uniform Discrete and Uniform Continuous PDFs

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



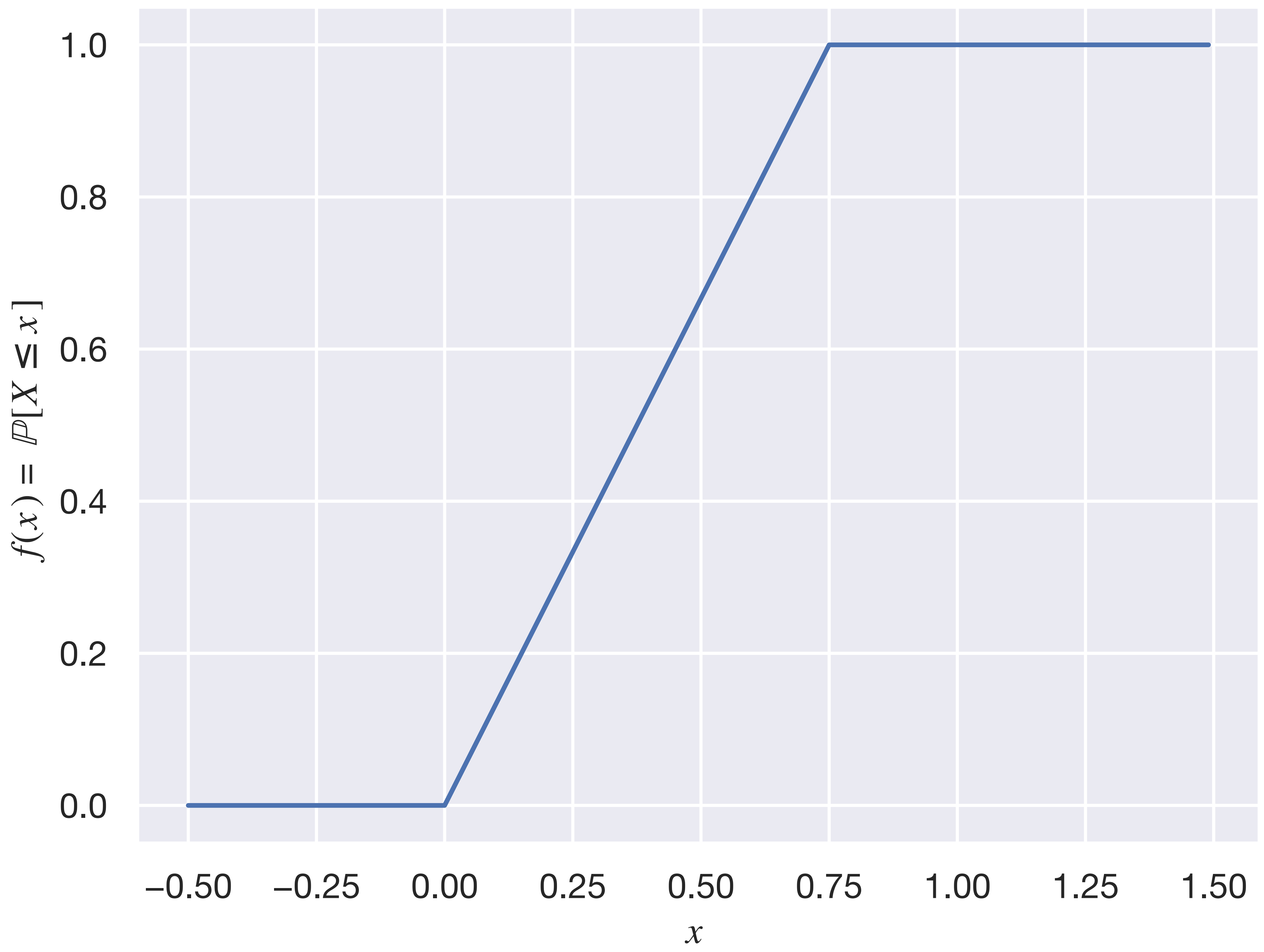

Discrete RV uniform on $\{0, 0.25, 0.5, 0.75\}$.



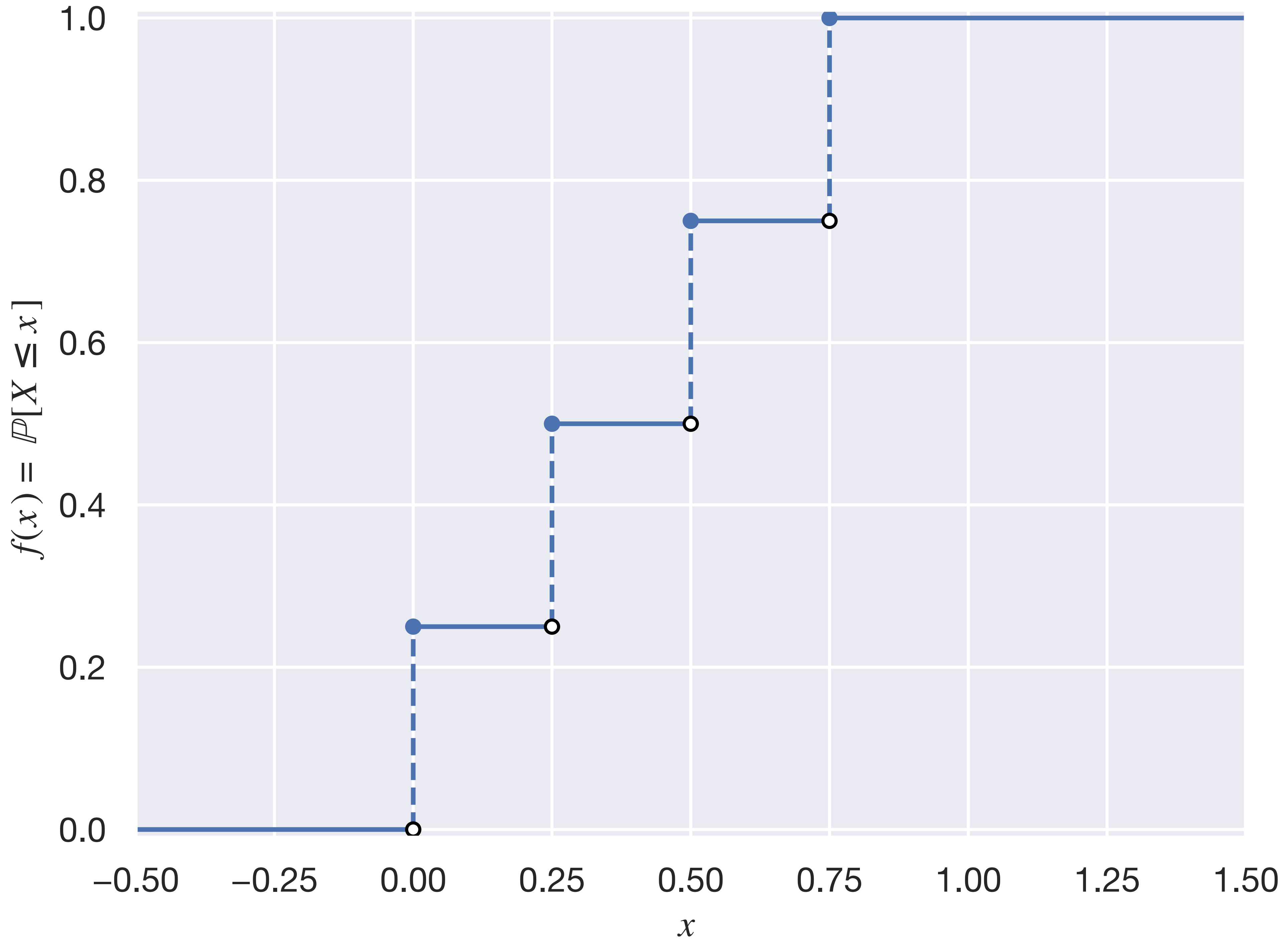
# Continuous vs. Discrete RVs

## Example: Uniform Discrete and Uniform Continuous CDFs

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




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



# Continuous vs. Discrete RVs

## Summary

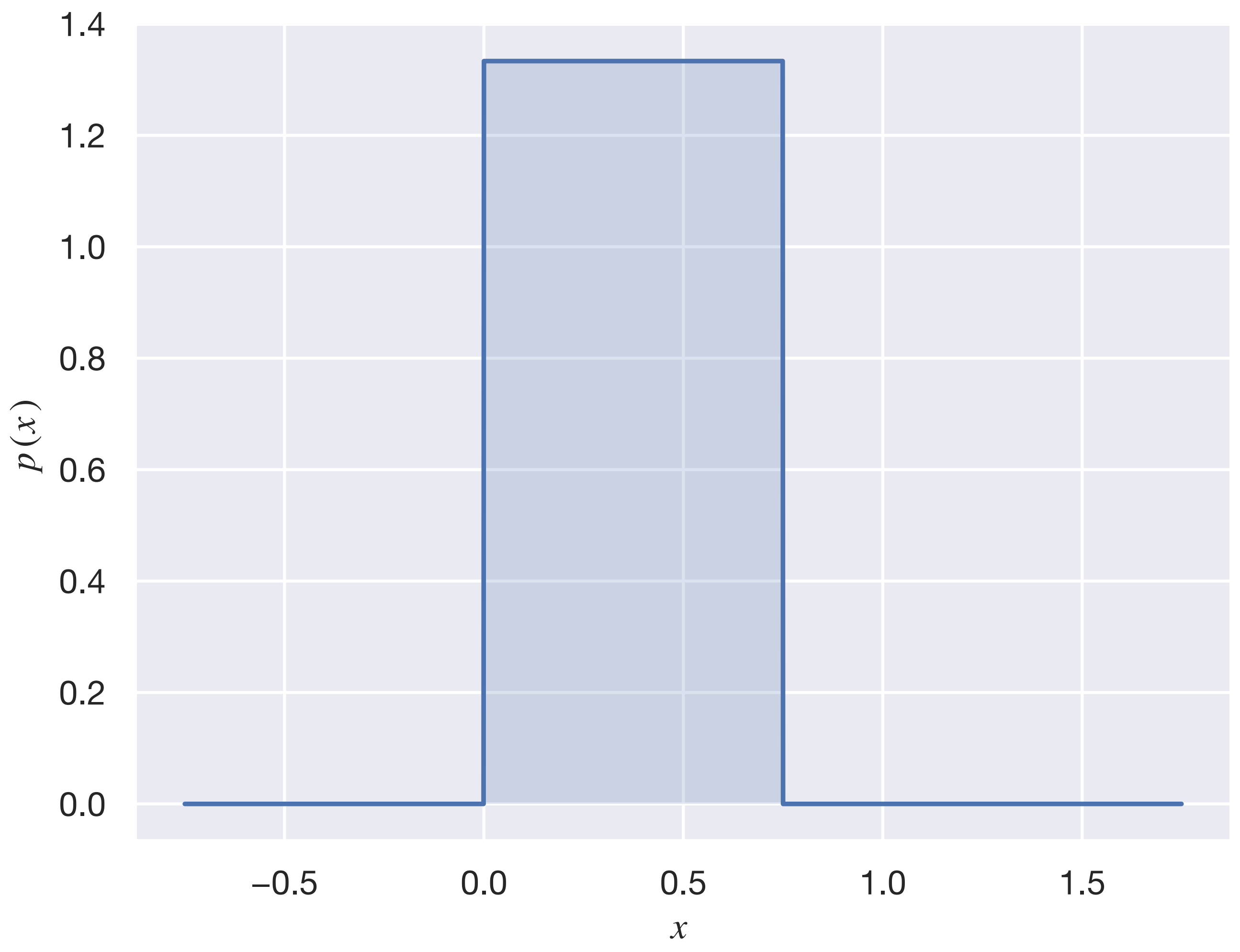
For continuous RVs,

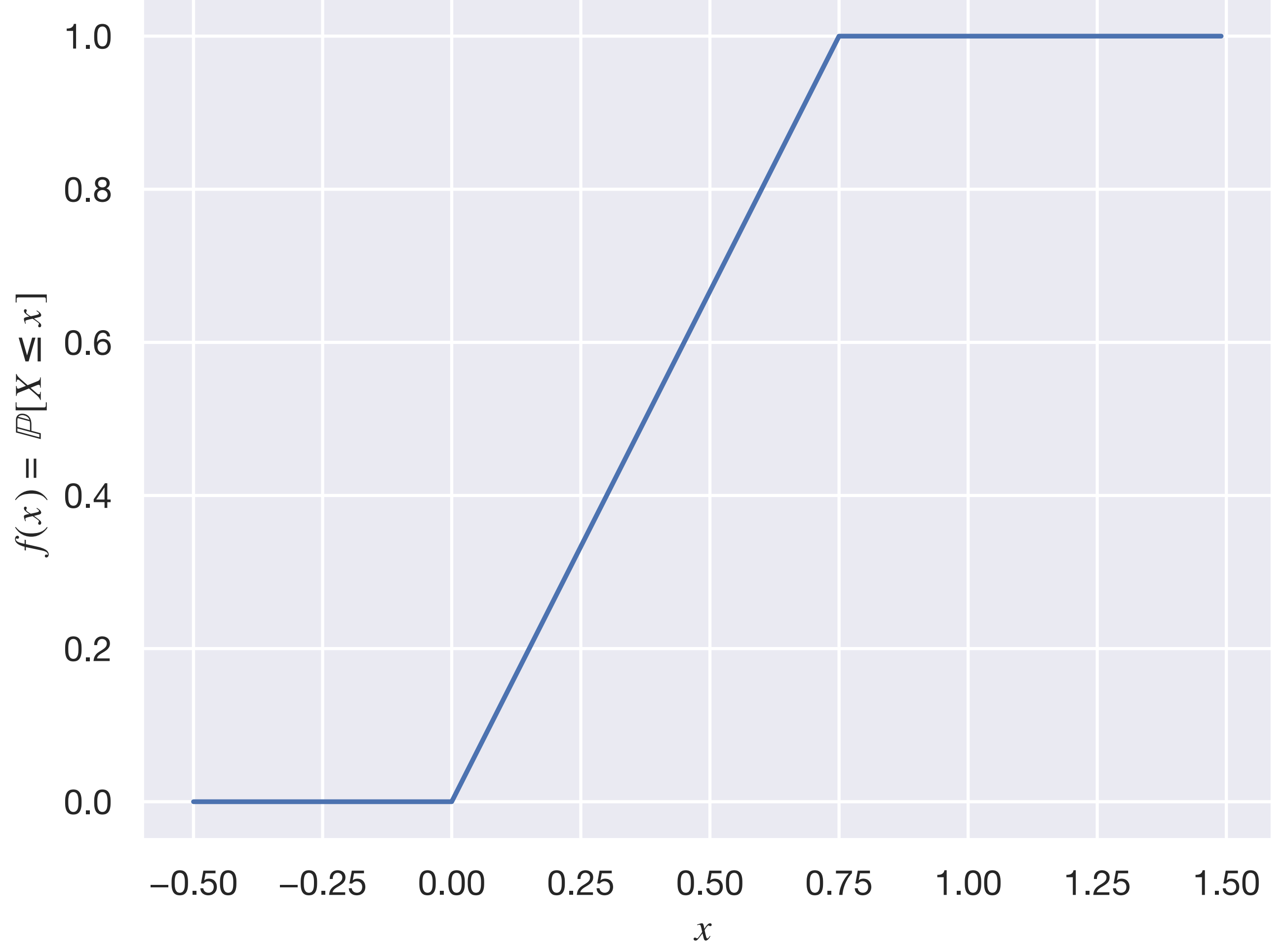
$$\mathbb{P}(X = x) = 0$$

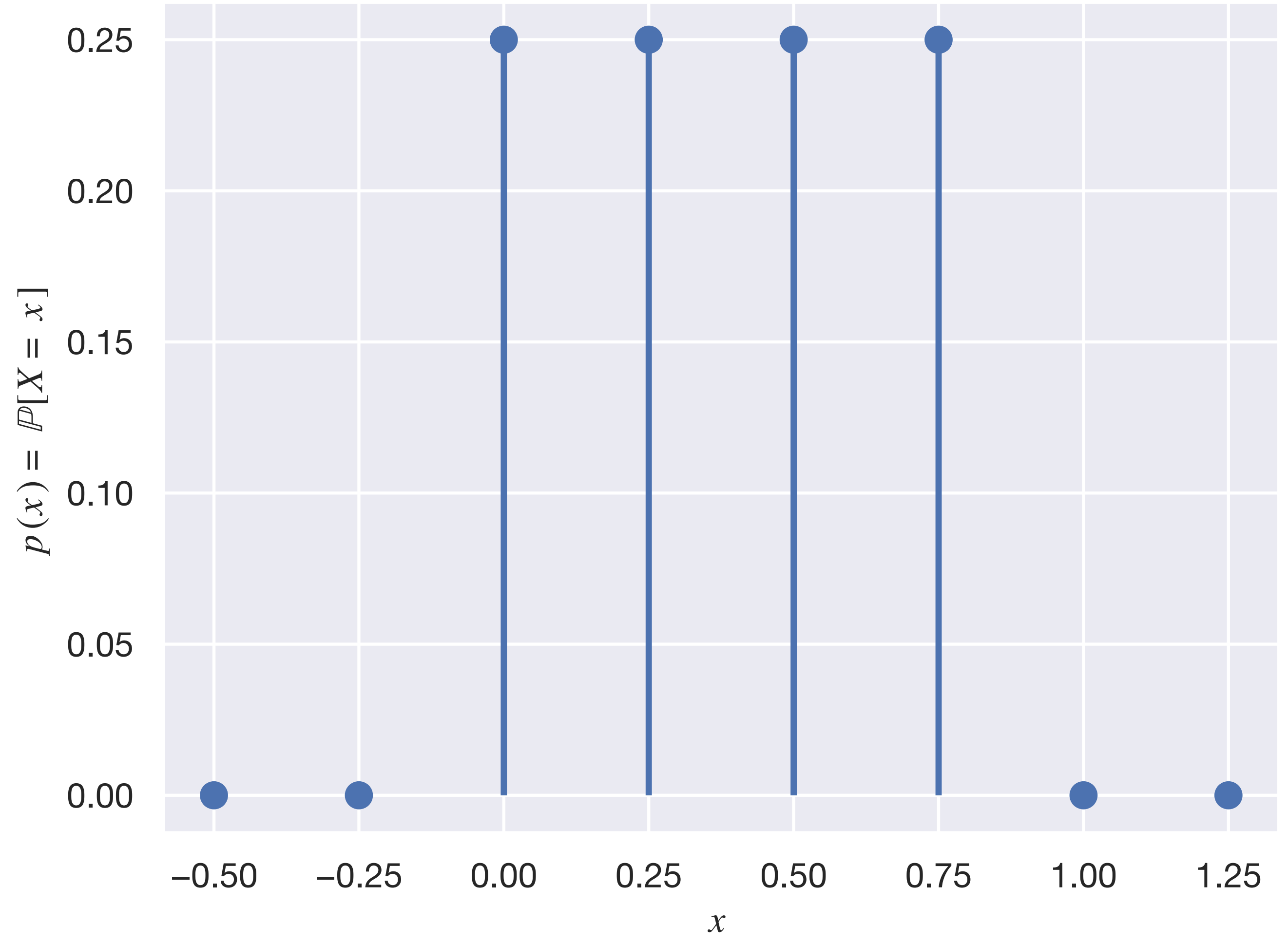
$$\mathbb{P}(b \leq X \leq b) = \int_a^b p_X(x)dx$$

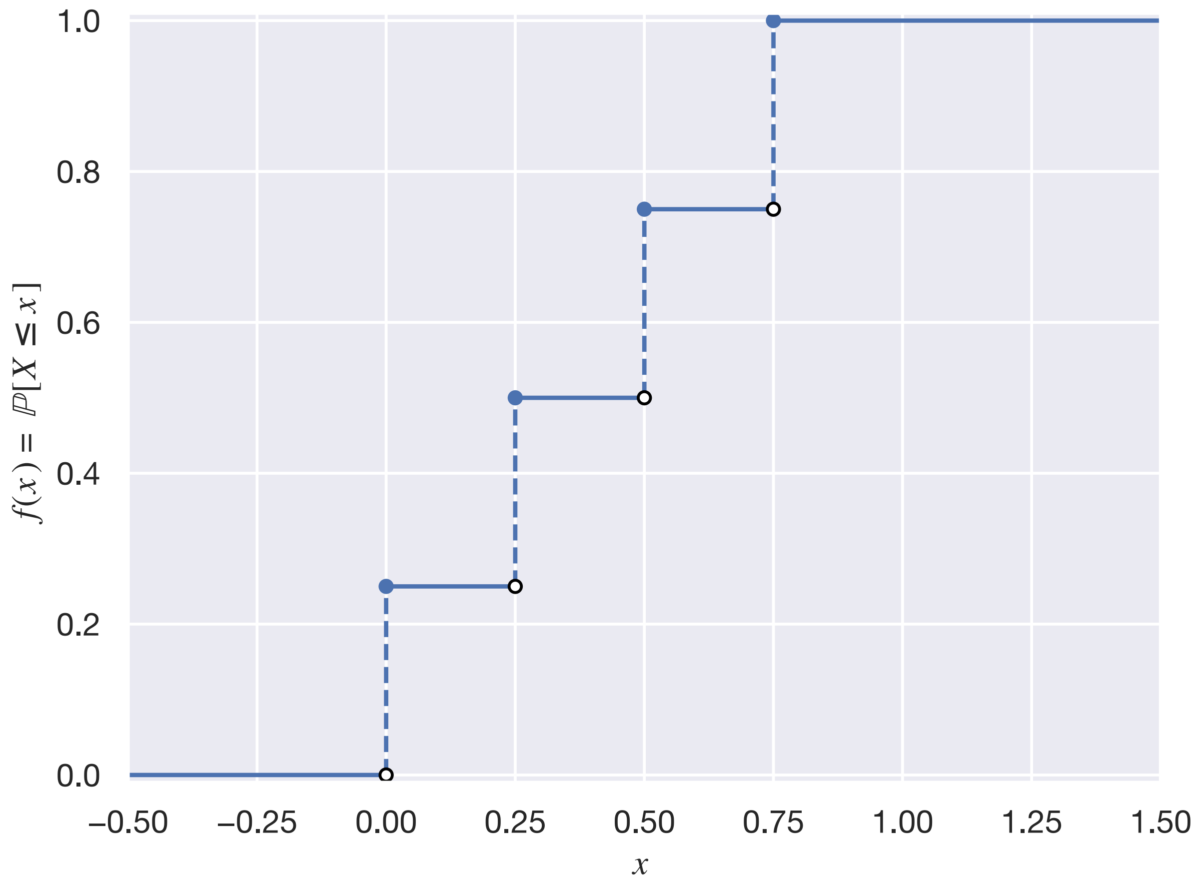
For discrete RVs,

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



# 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.

Charlie rolls a fair six-sided die.

Let $X$ count the # of heads for Alice and Bob.

Let $Y$ be the # on the face of Charlie's die.

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | | | | | | |
| 1 | | | | | | |
| 2 | | | | | | |

# Joint Distribution

## Definition

Let $X_1, \ldots, X_n$ be random variables.

The joint distribution of $X_1, \ldots, X_n$ is the probability distribution $\mathbb{P}_{X_1,\ldots,X_n}$ with corresponding PMF/PDF:

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

For discrete random variables,

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

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | | | | | | |
| 1 | | | | | | |
| 2 | | | | | | |

# Joint Distribution

## Definition

Let $X_1, \ldots, X_n$ be random variables.

The joint distribution of $X_1, \ldots, X_n$ is the probability distribution $\mathbb{P}_{X_1,\ldots,X_n}$ with corresponding PMF/PDF:

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

For discrete random variables,

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

$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}($the number of heads is $x$ and the die is $y)$

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |
| 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 |
| 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |

# Joint Distribution

## Definition

Let $X_1, \ldots, X_n$ be random variables.

The joint distribution of $X_1, \ldots, X_n$ is the probability distribution $\mathbb{P}_{X_1,\ldots,X_n}$ with corresponding PMF/PDF:

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

Joint distribution of example: $p_{X,Y}(x, y)$.

$$p_{X,Y}(1,1) = 1/12$$

$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}$(the number of heads is $x$ and the die is $y$)

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |
| 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 |
| 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |

# 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:

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

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

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |
| 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/2 |
| 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |

# 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:

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

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

| | | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |
|---|---|---|---|---|---|---|---|
| $X$ | 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |
| | 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 |
| | 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 |
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | | | | $Y$ | | | |

# Conditional Distribution

## Definition

For two random variables $X, Y$ with joint distribution $p_{X,Y}(x, y)\ldots$

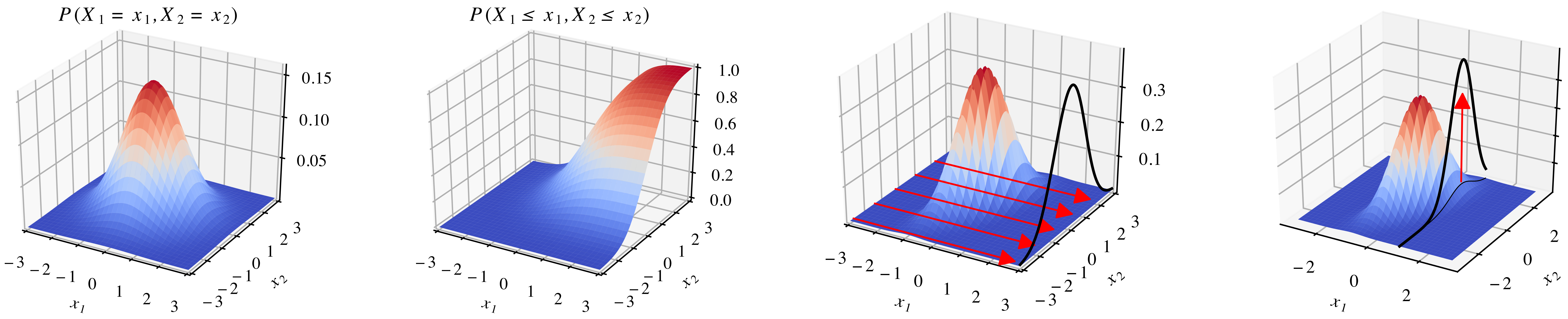
The conditional distribution of $X$ given $Y = y$ is given by *only* considering the events where $Y = y$.

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

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |
| 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/2 |
| 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |
| | | 1/6 | | | | | |

# Joint Continuous Distributions

## Joint, marginal, and conditional



# 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 "factors" the joint distribution into marginal and conditional.

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

# Independence

## Intuition and definition

Two RVs $X, Y$ 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.

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

Let $\{X_i\}_{i \in I}$ be a collection of RVs indexed by $I. \{X_i\}$ are independent if

$$p_{X_{i_1},\ldots,X_{i_k}}(X_{i_1}, \ldots, X_{i_k}) = \prod_{j=1}^{k} p_{X_{i_j}}(x_{i_j}) \text{ for any subset of indices } \{i_1, \ldots, i_k\} \in I,$$

# Independence

## Intuition and definition

Two RVs $X, Y$ 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.

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

| $X$ \ $Y$ | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| 0 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |
| 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/2 |
| 2 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/24 | 1/4 |
| | | 1/6 | | | | | |

# 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:

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

and all the $X_i$ have the same distribution.

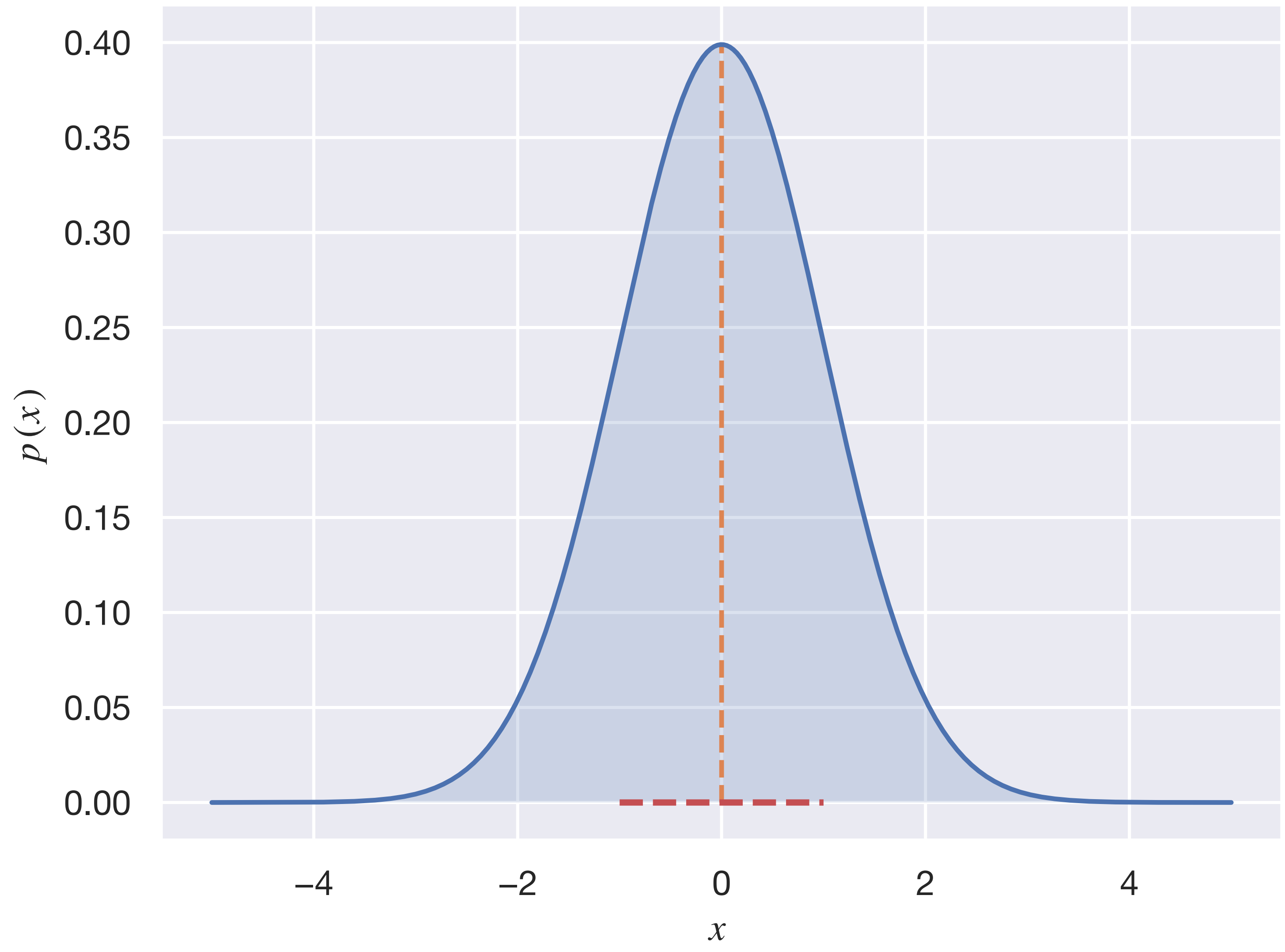
*Very common assumption in ML!*

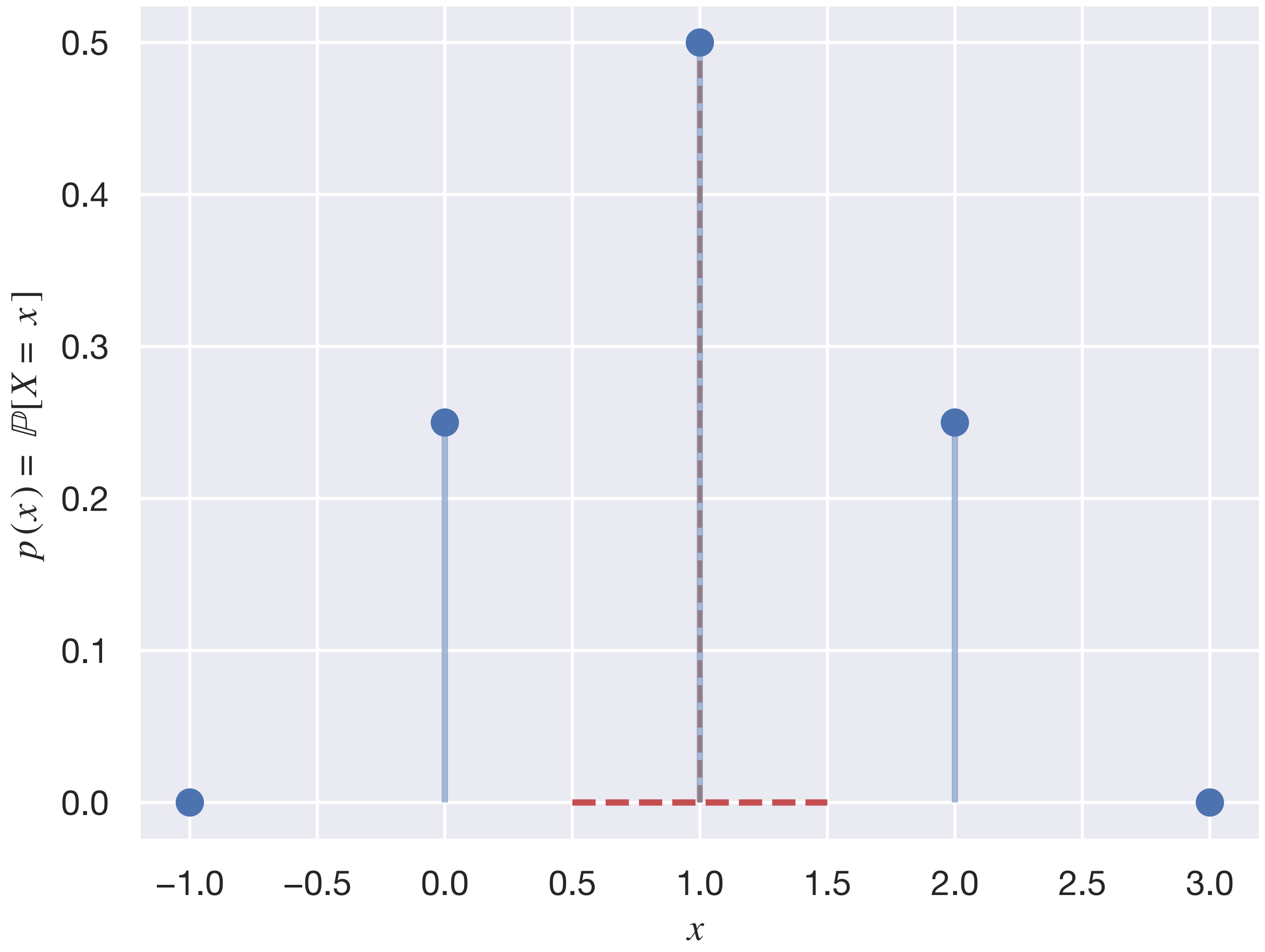
# Expectation

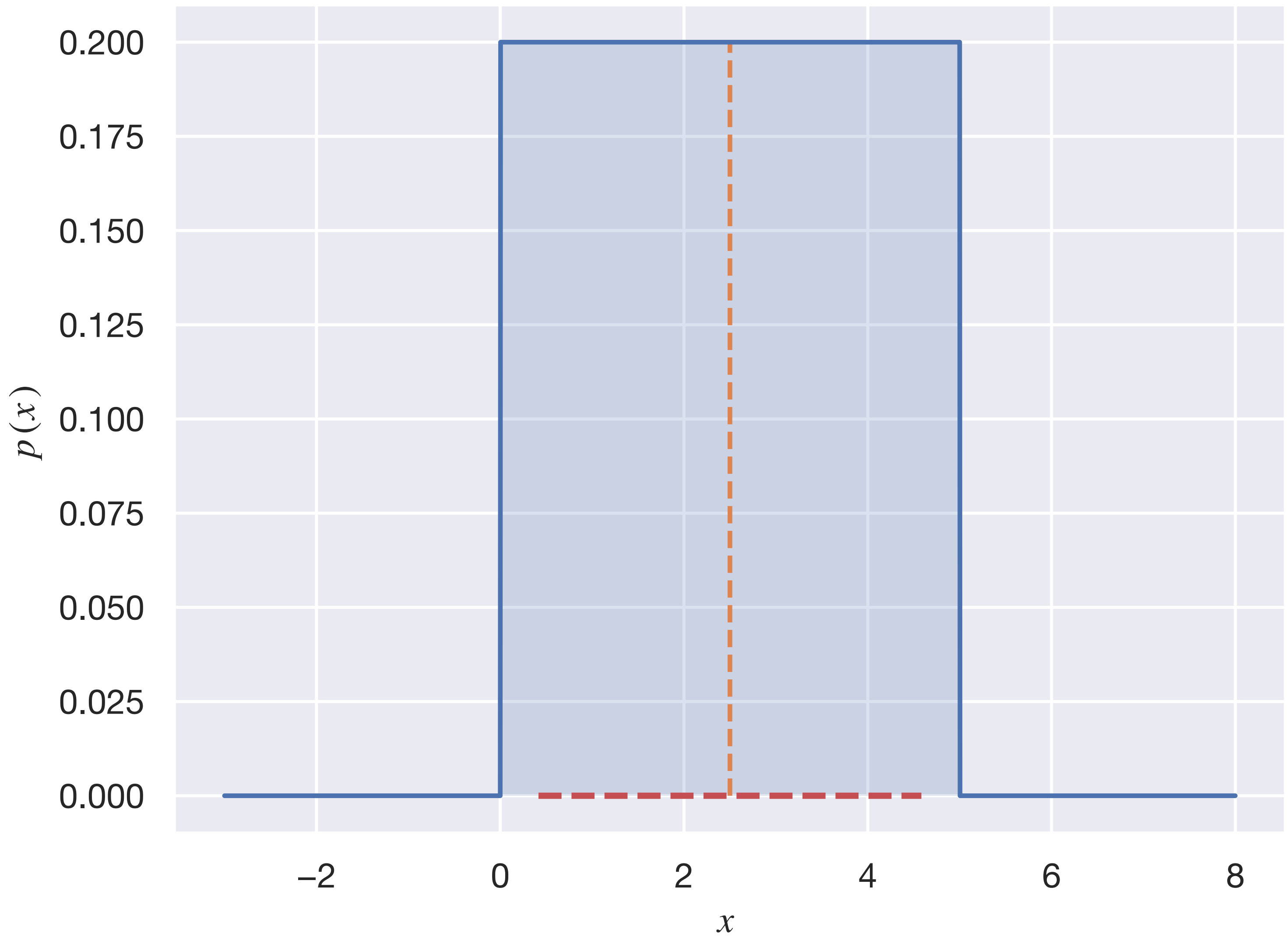
## Definition and Properties

# Expected Value
## Intuition

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



# Expected Value

## Definition

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

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

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

*A function of a random variable is a random variable!*

# Functions of RVs are RVs

## Very important fact!

*A function of a random variable is a random variable!*

# Expected Value

## Properties of the expected value

**Linearity.** The expectation is a linear operator:

$$\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X] \text{ and } \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \text{ for } \textit{any} \text{ random variables } X \text{ and } Y$$

**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 RV's expectation, given an event.

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.

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

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

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X$ | 0 | 1 | 0 | 1 | 0 | 1 |
| $Y$ | 0 | 1 | 1 | 0 | 1 | 0 |

What is $\mathbb{E}[X]$?

What is $\mathbb{E}[X \mid Y = 1]$?

What is $\mathbb{E}[X \mid Y = 0]$?

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

# Conditional Expectation

## Definition (given events)

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:

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

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:

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

# Conditional Expectation

## Definition (given a random variable)

For two RVs $X$ and $Y$, think of the conditional expectation of $Y$ given $X$ 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)

**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?

# 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 $h$,

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

**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 $X, Y$,

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

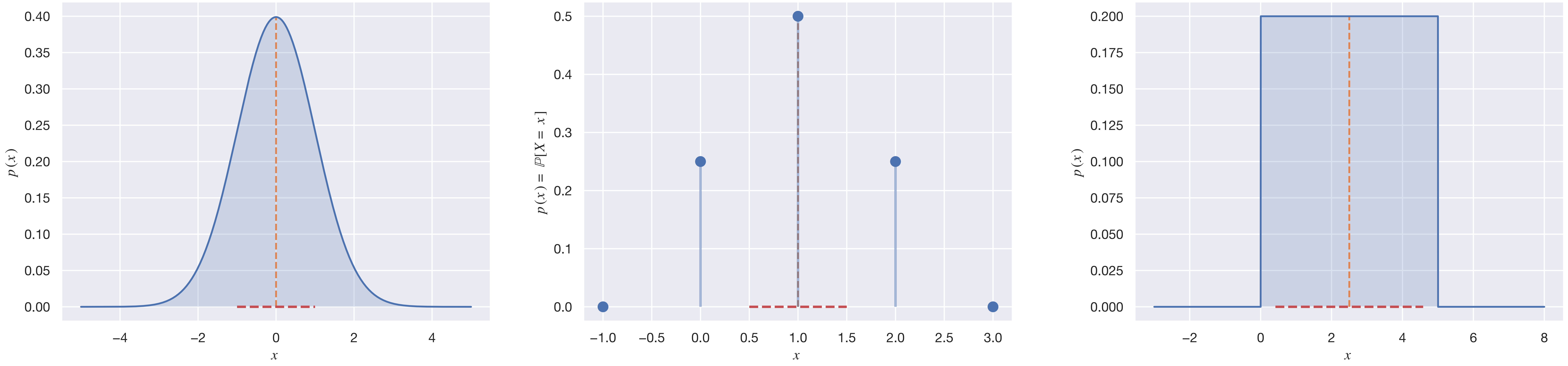
# Variance

## Definition and Covariance

# Variance

## Intuition

The variance of a random variable is how “spread” around its expectation it is.



# Variance

## Definition

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 $\sqrt{\mathrm{Var}(X)}$.

# Variance

## Properties of variance

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

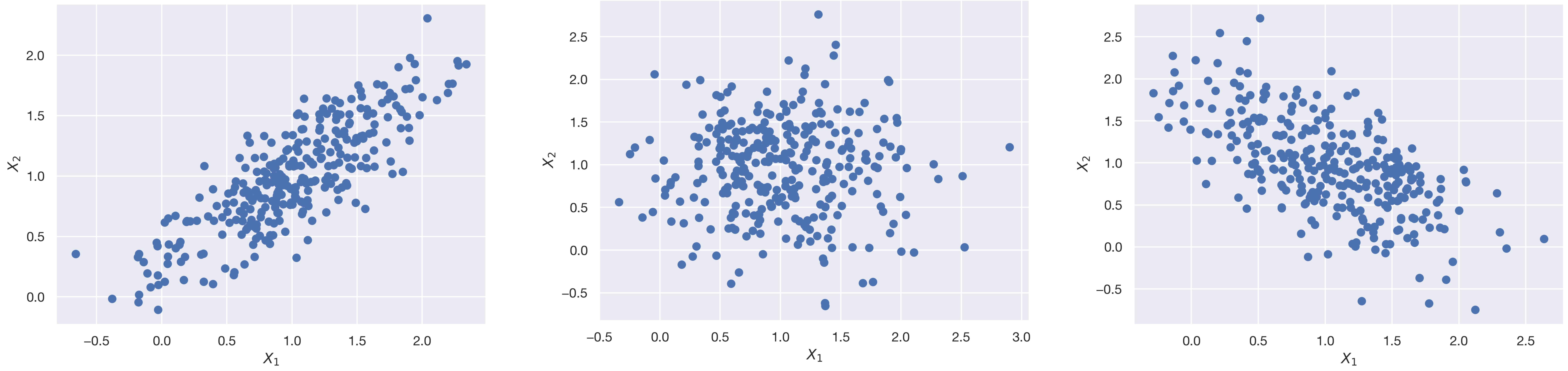
$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X).$$

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.



# Covariance

## Definition

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). Equivalently:

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

The correlation is what we get from normalizing the covariance:

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

# Covariance

## Properties of covariance

Covariance follows the “symmetry” property:

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

Covariance follows the “bilinearity” property:

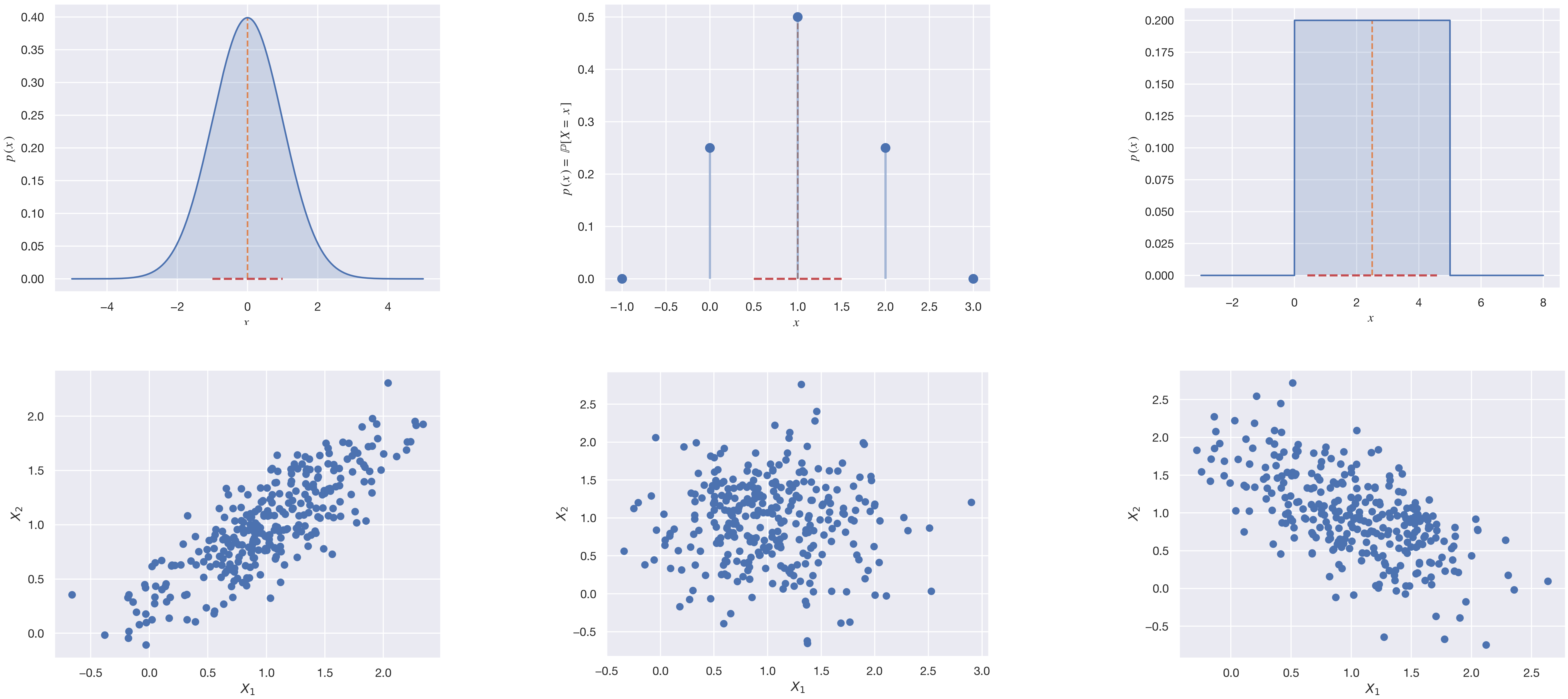
$$\mathrm{Cov}(\alpha X + \beta Y, Z) = \alpha \mathrm{Cov}(X, Z) + \beta \mathrm{Cov}(Y, Z).$$

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

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

## Variance and Covariance Matrix

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

## Variance and Covariance Matrix

The variance of a random vector generalizes to the covariance matrix

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

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

In this class, a random vector's variance *is* its covariance:

$$\mathrm{Var}(\mathbf{x}) := \boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$

# Random Vectors

## Properties of the variance

The covariance matrix is **symmetric.**

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

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

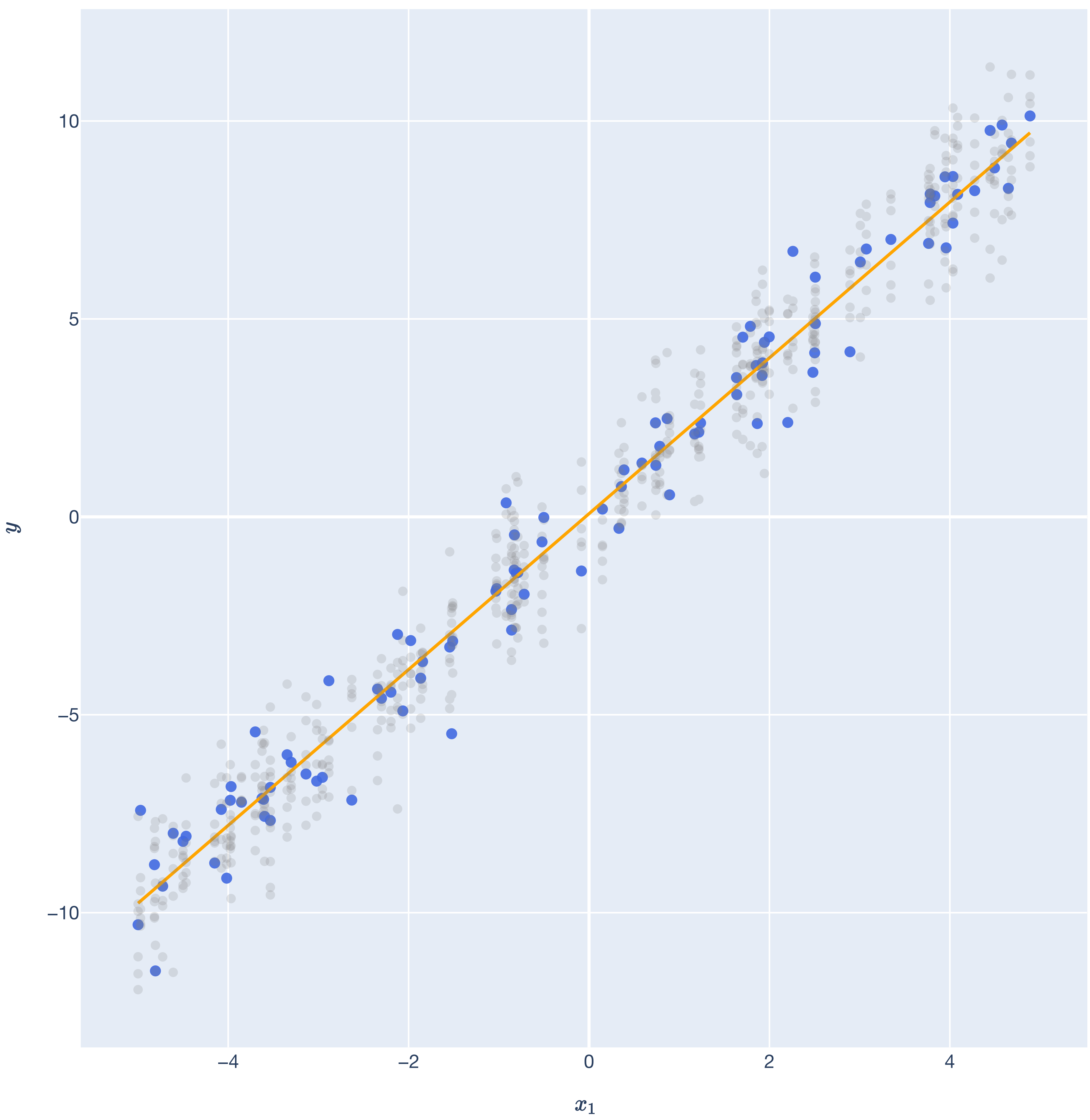
$$\mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

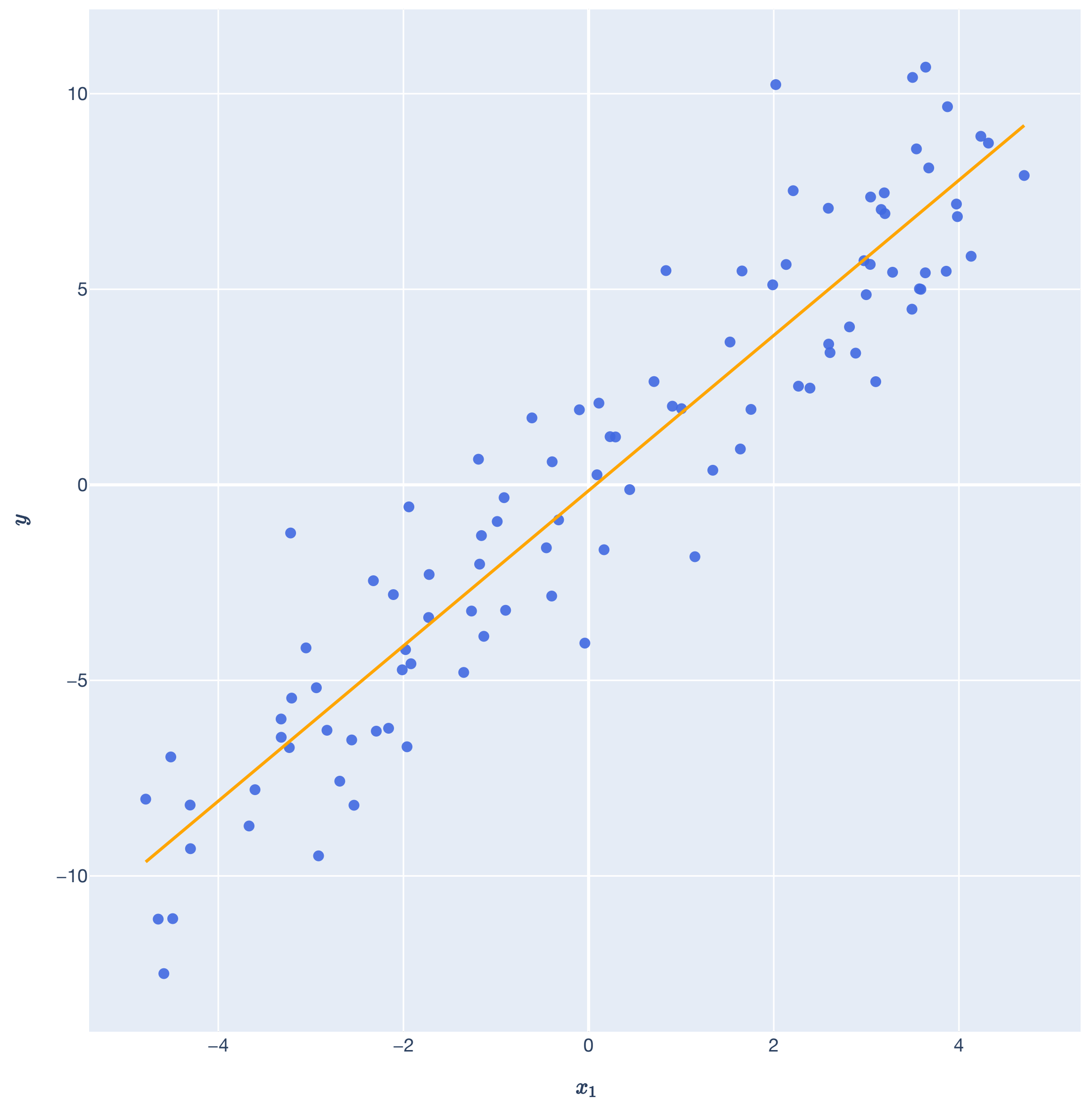
# Data as random

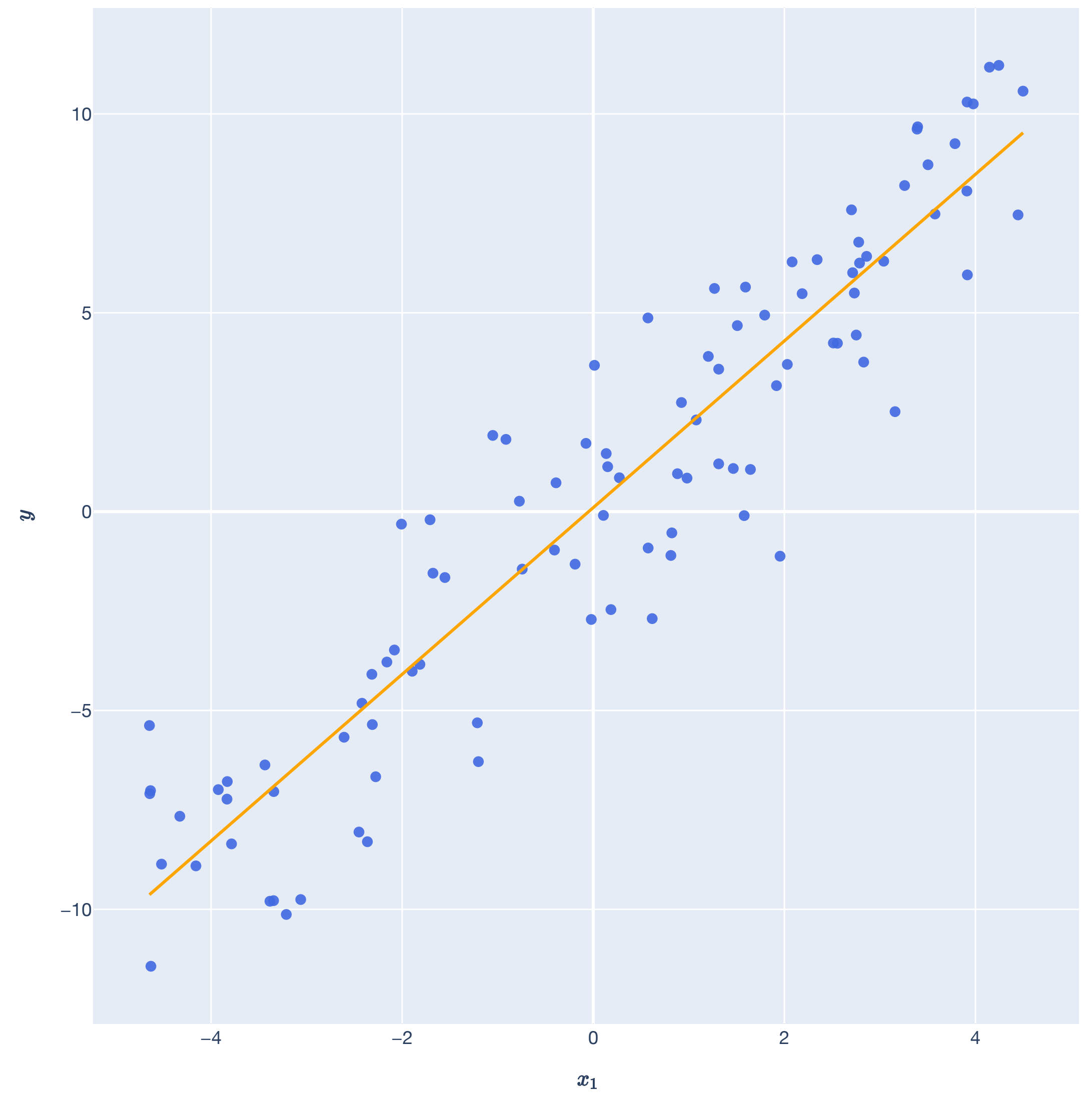
## Modeling regression with probability

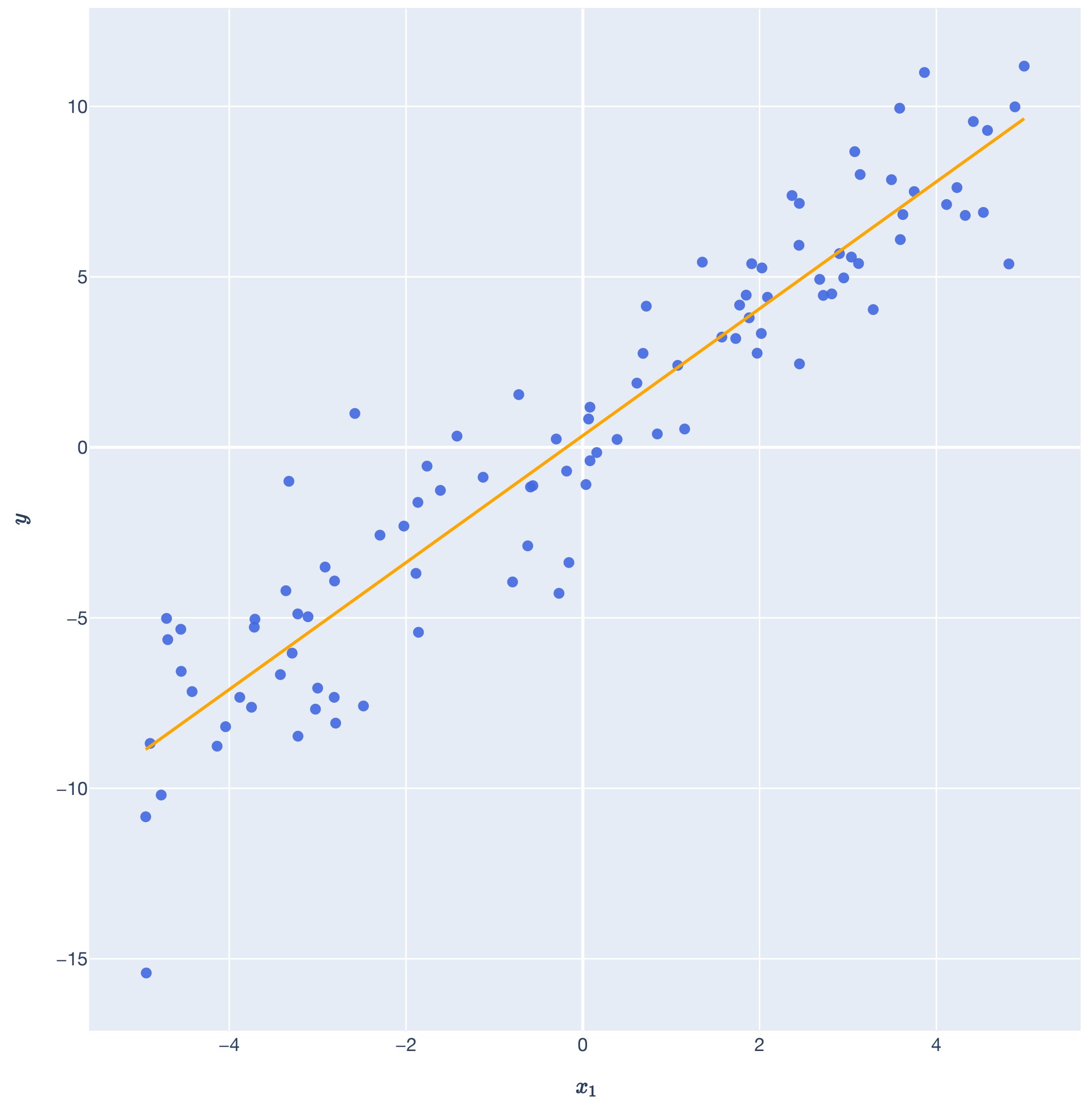
# Regression

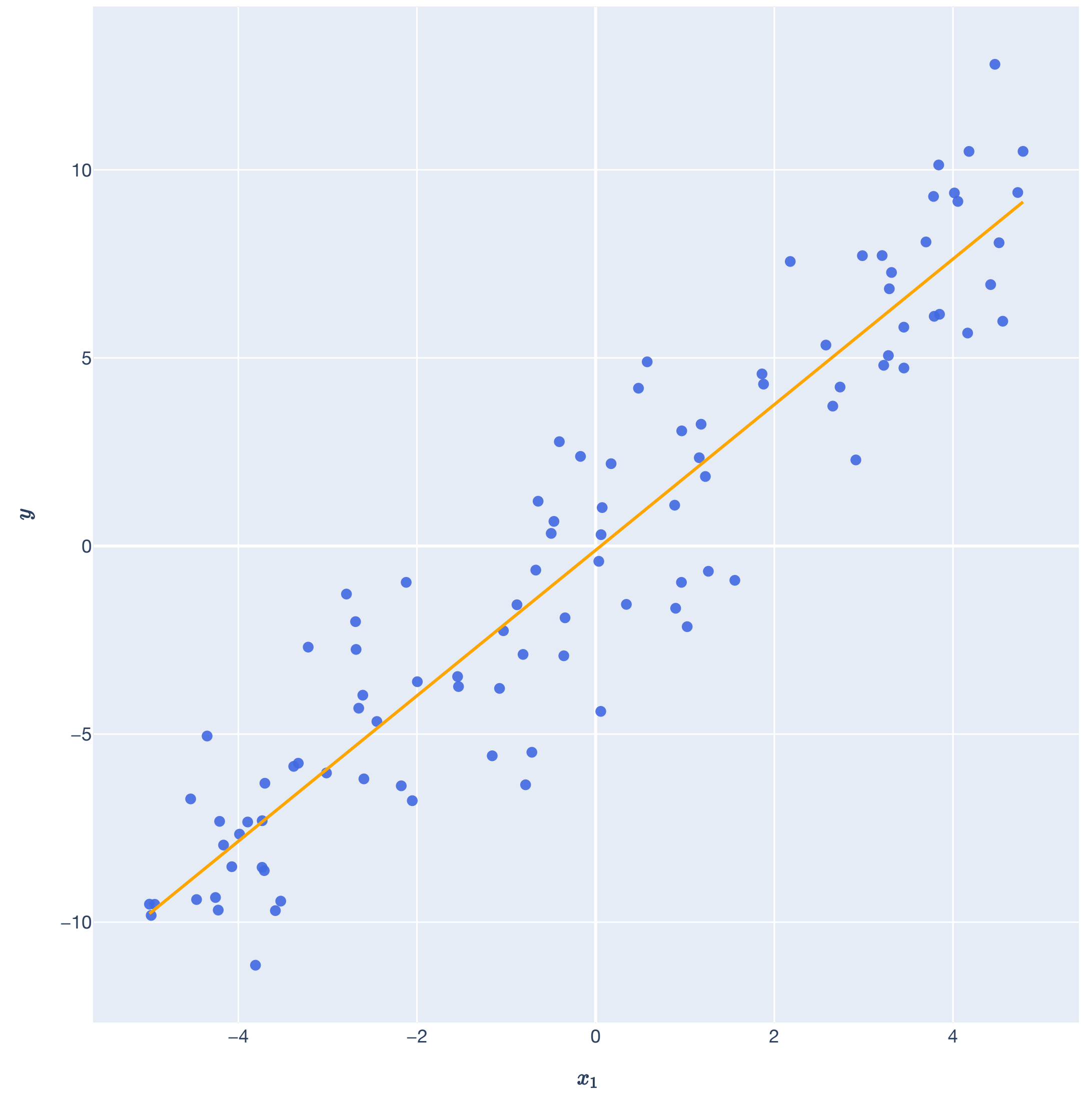
## Modeling randomness



# Regression

## Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

<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 (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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>Ultimate Goal:</u> 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*.

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

# Regression

## Setup

Least squares expanded is just:

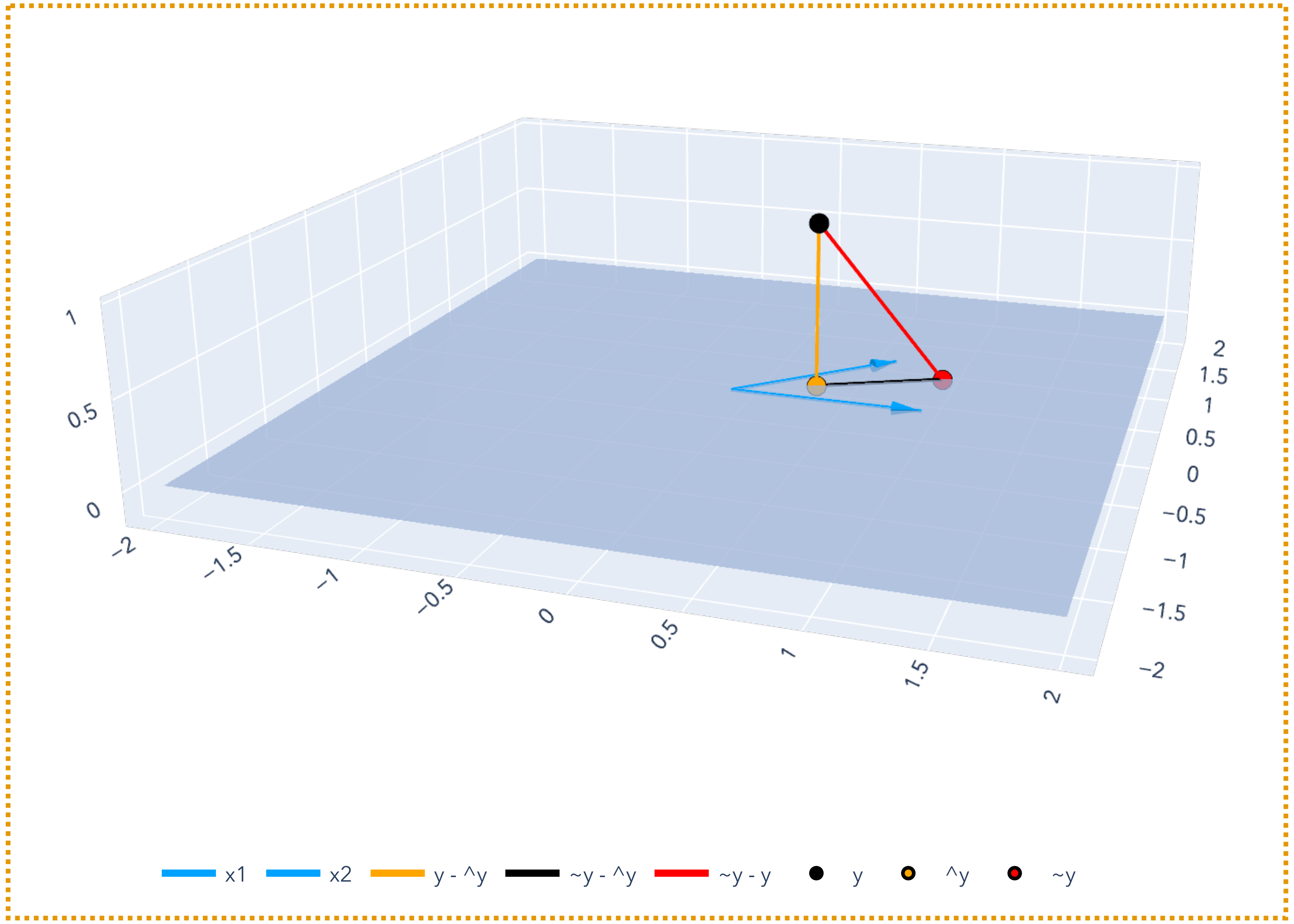
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

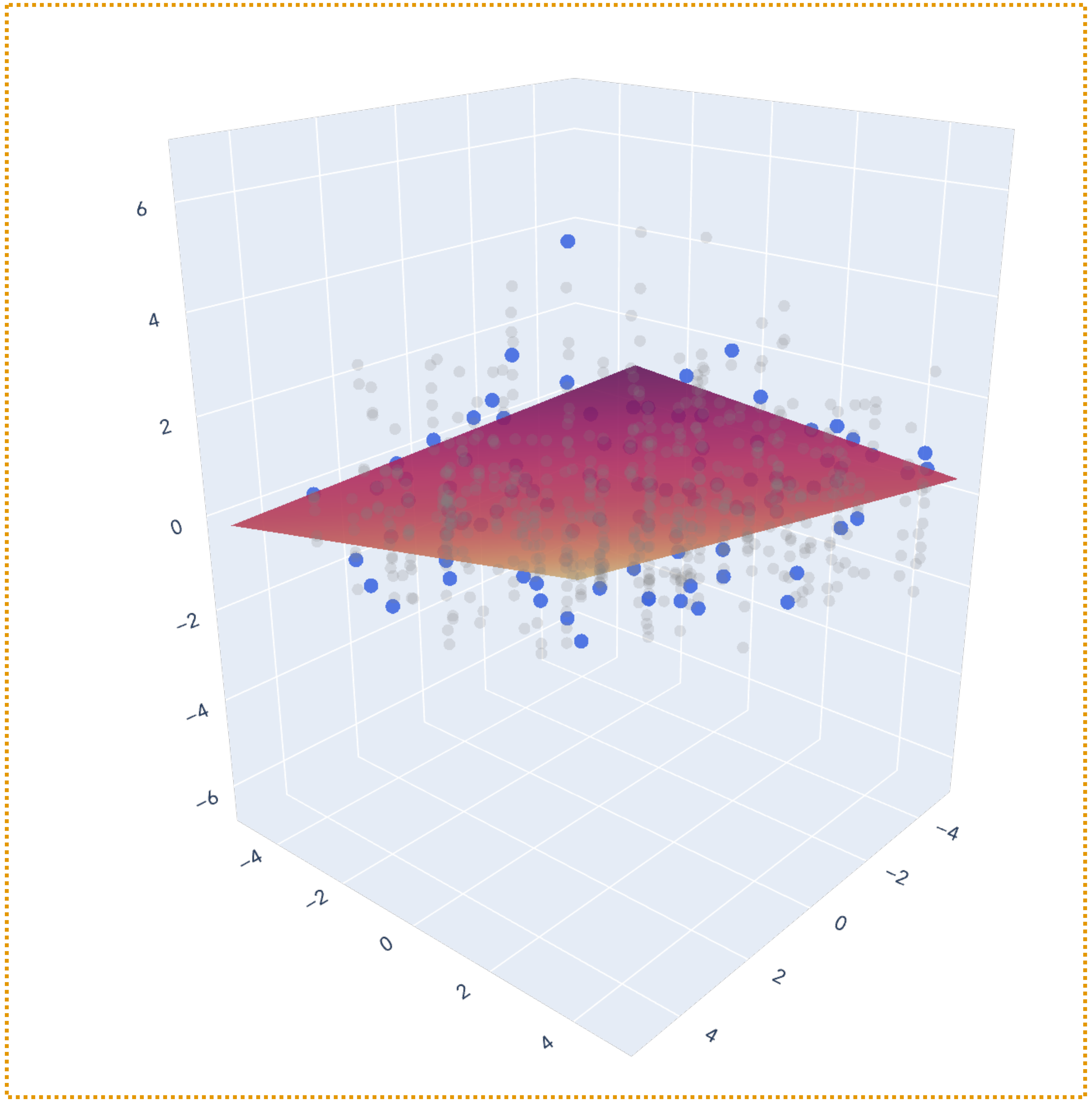
Put a $1/n$ in front, and it looks like we're minimizing an average:

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

# Regression

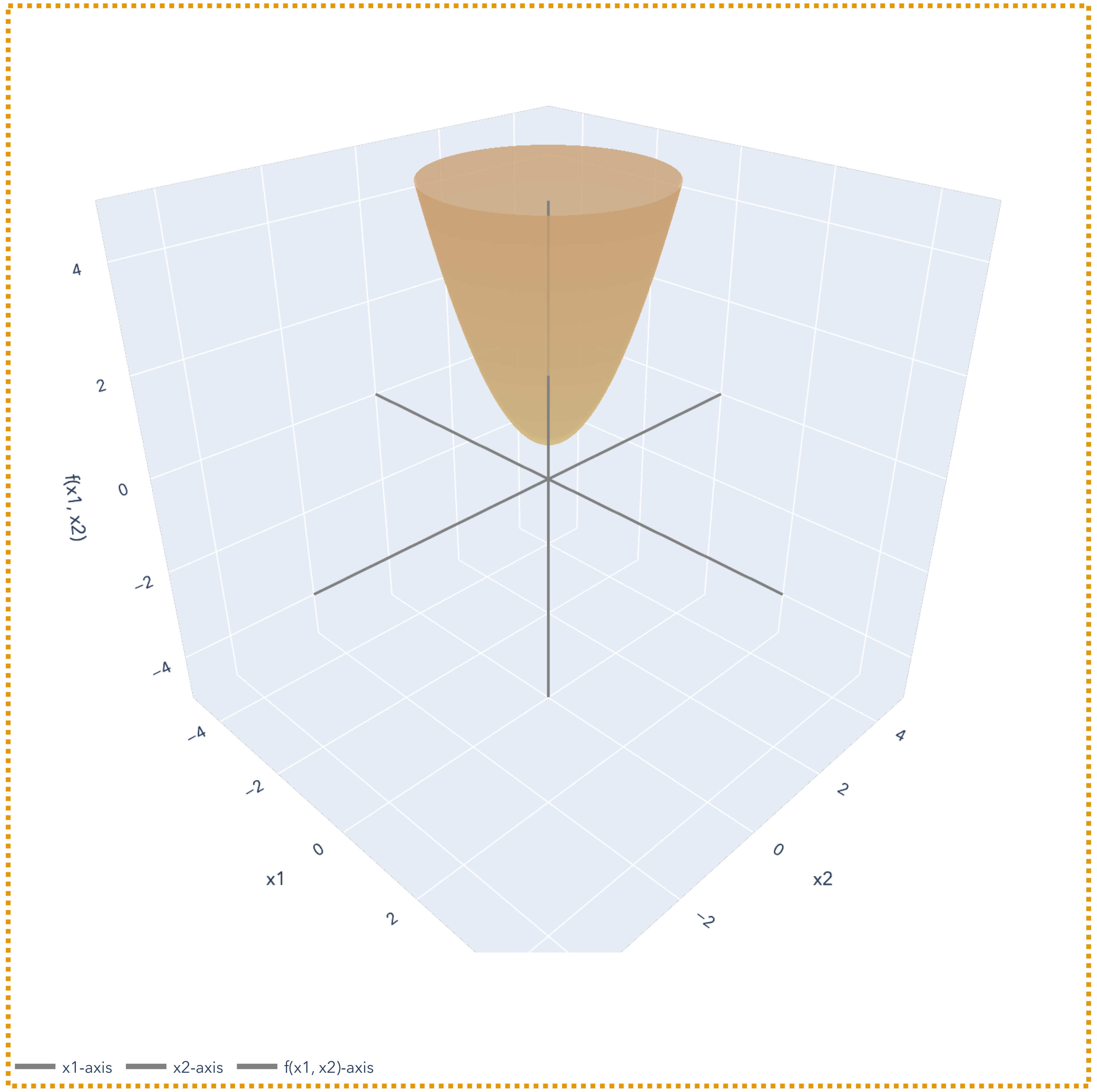
## A note on $\hat{\mathbf{w}}$

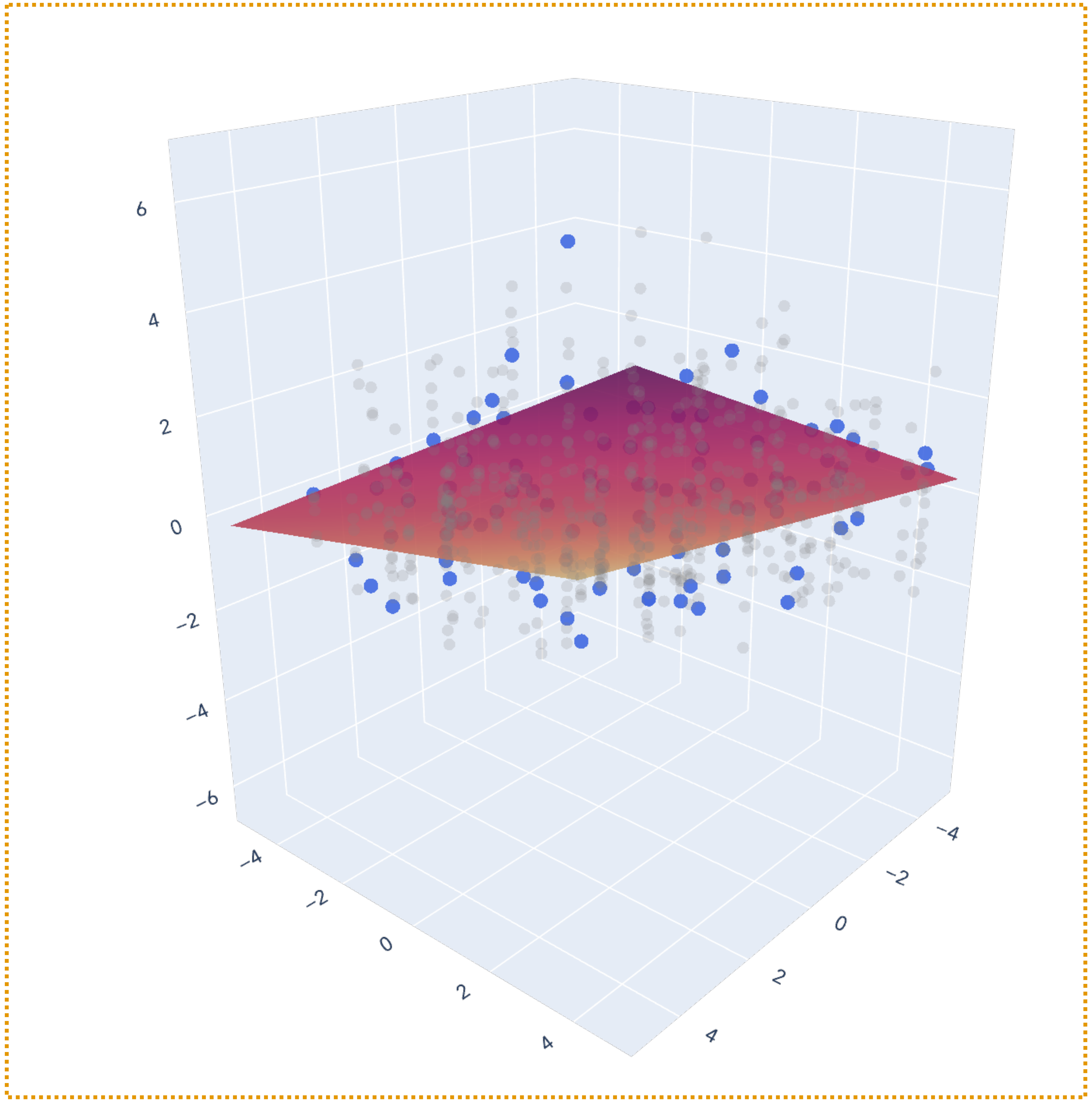


# Regression

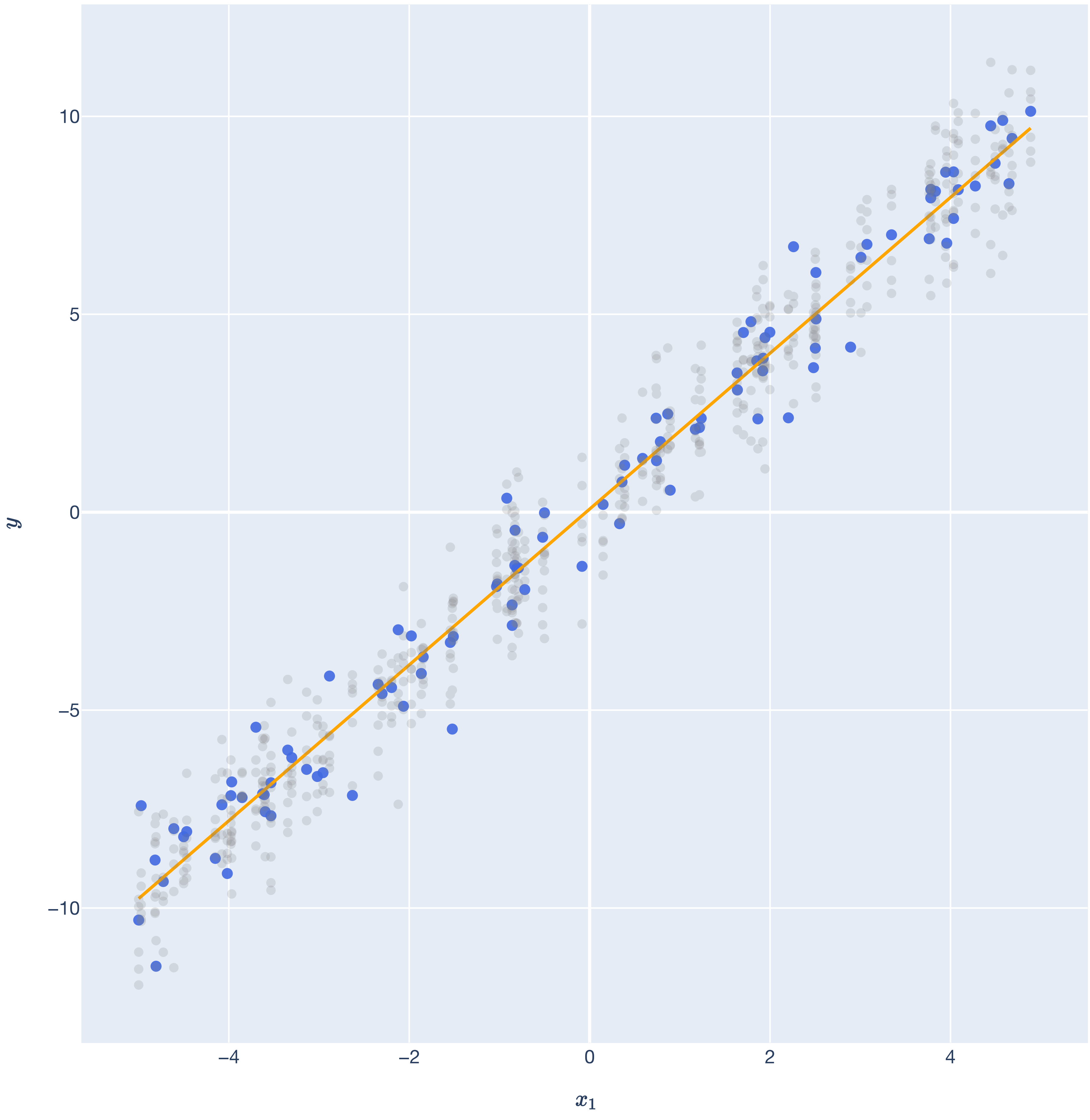
## A note on $\hat{\mathbf{w}}$

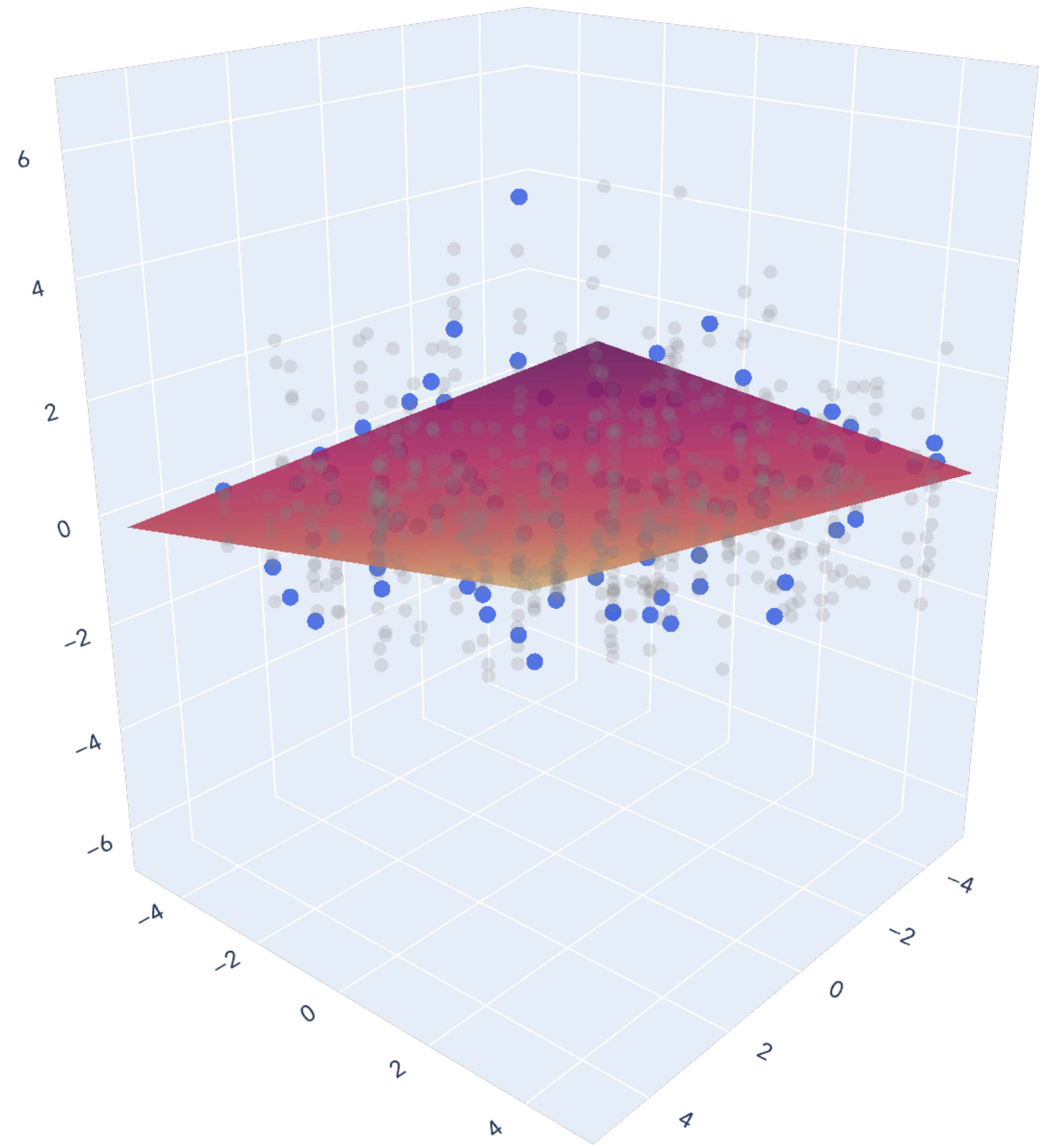


# Regression

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



# Regression

## Setup, with randomness

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

Random vector

Random variable

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

… from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$!

<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: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Our "model", a function that we hope "generalizes" well to new datapoint $(\mathbf{x}_0, y_0)$.

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:

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

Because we assume that $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$ as well!

# Regression with randomness

## Part 1: Random data from probability distribution

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.

We will now assume 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

## Part 1: Random data from probability distribution

Random vector

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}$$

Random variable

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

… from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$!

# Regression with randomness

## Part 2: Notion of error and linear model

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, unseen $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$.

Our notion of error is the squared loss:

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

In our case, $f : \mathbb{R}^d \to \mathbb{R}$ is a *linear functional* of the features.

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

# Regression with randomness

## Part 2: Notion of error and linear model

<u>Assume further:</u> $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ 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* $\epsilon_i$ 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

## Part 2: Notion of error and linear model

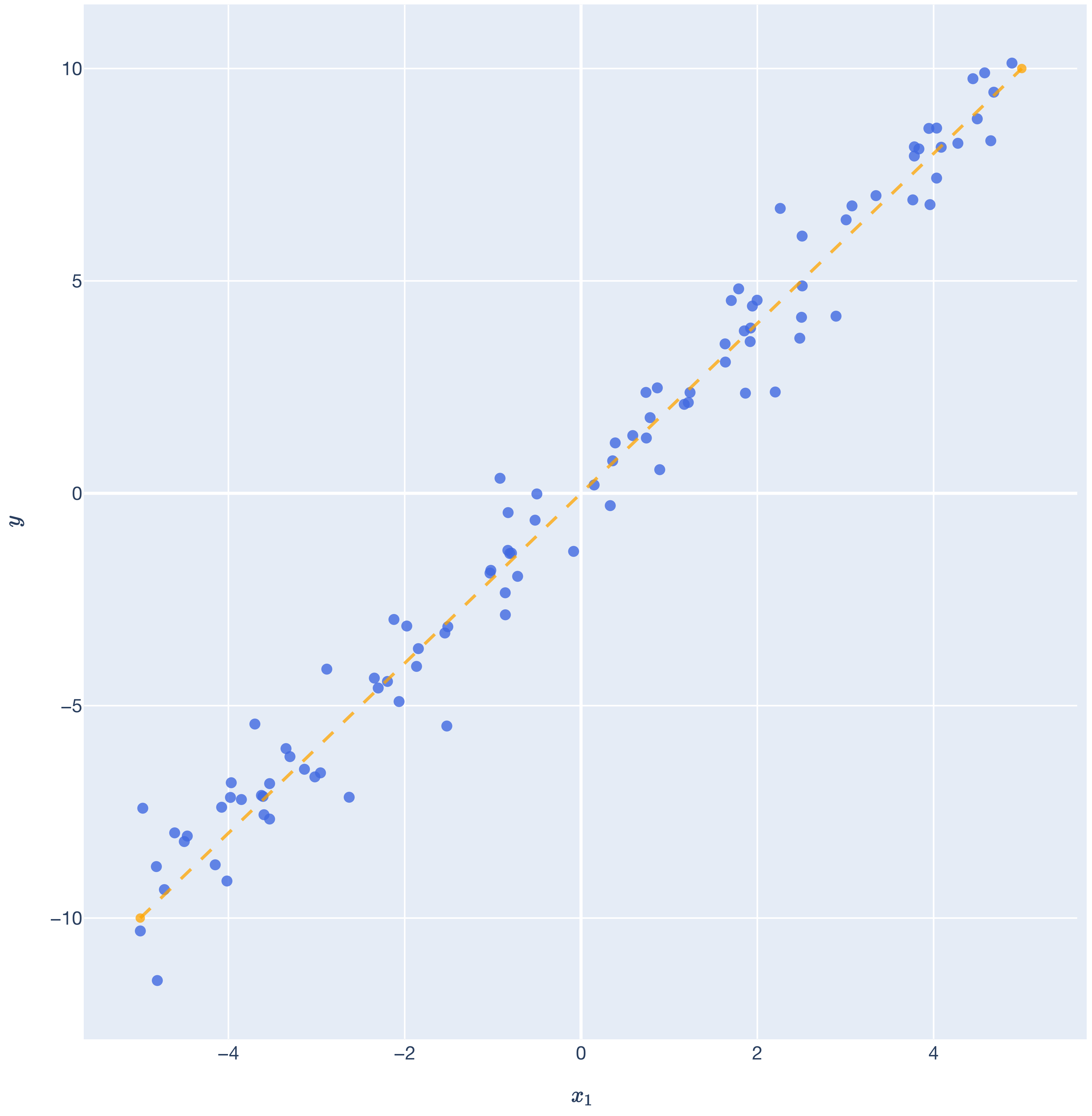
<u>Assume further:</u> $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

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

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

Some *randomness* $\epsilon_i$ 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

## Part 2: Notion of error and linear model

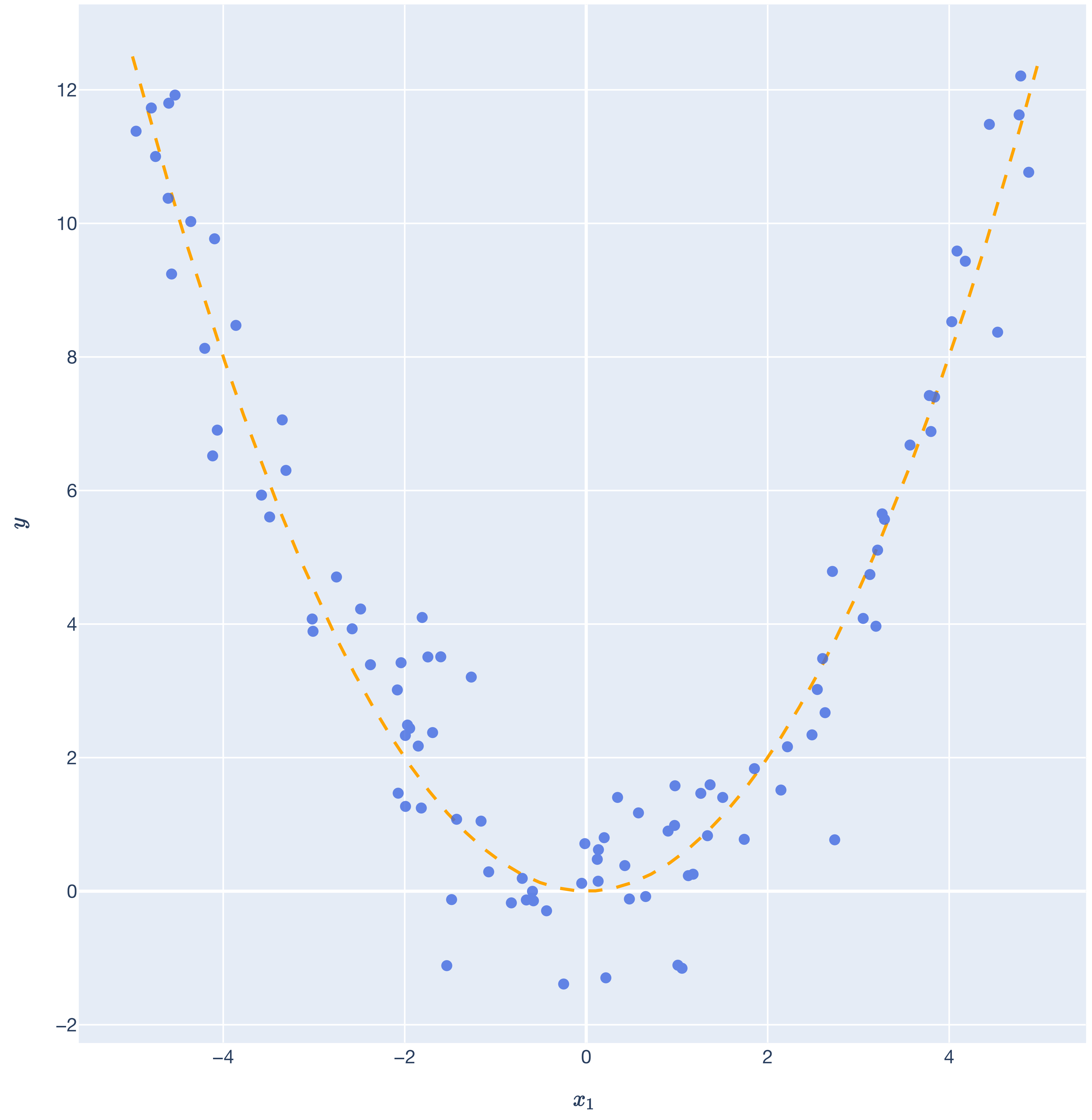
<u>Assume further:</u> $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

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

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

Some *randomness* $\epsilon_i$ 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

## Part 2: Notion of error and linear model

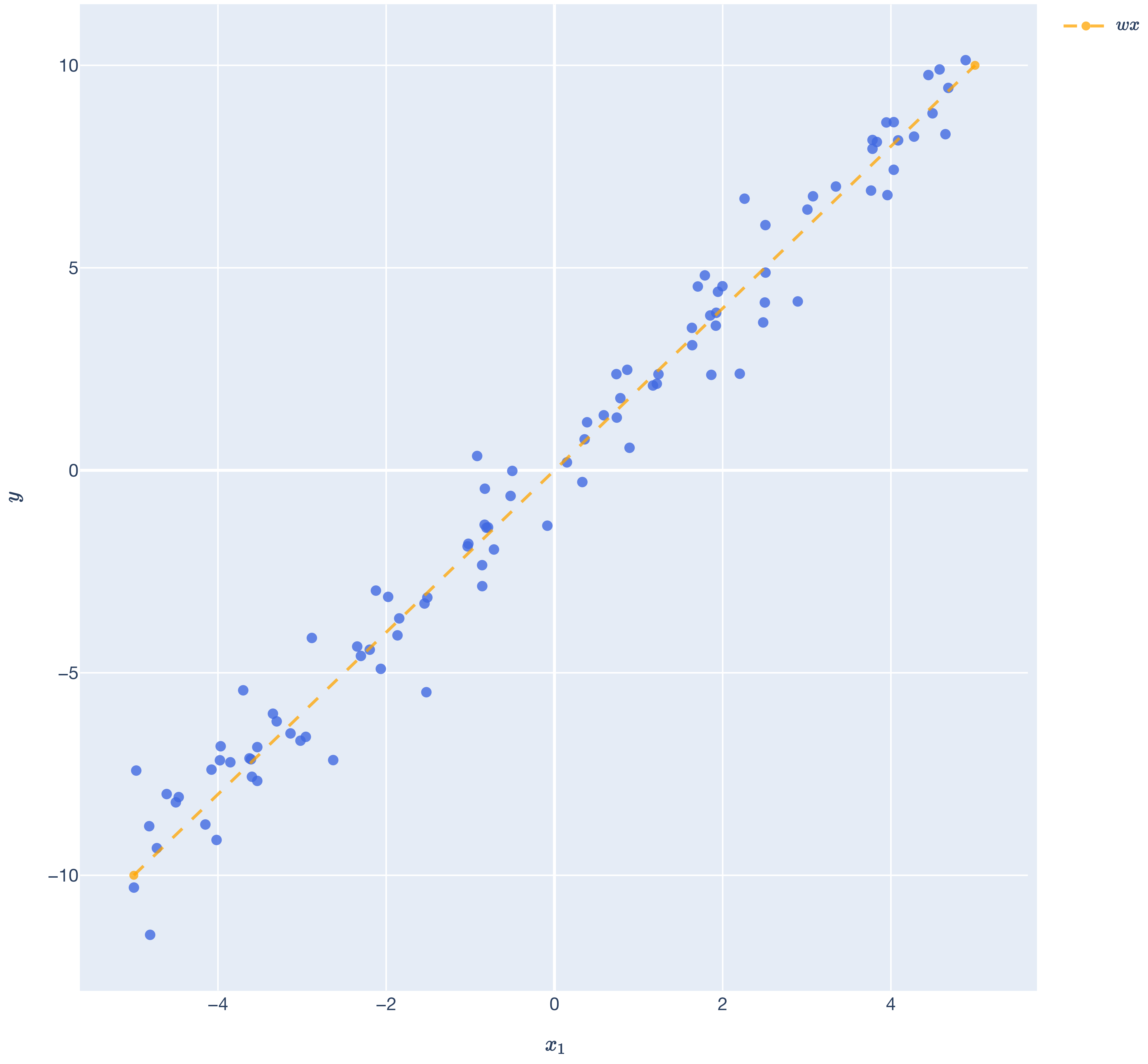
<u>Assume further:</u> $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

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

Some *linear* function $f^*_{\mathbf{w}^*}(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}^*$ explains linear relationship between $\mathbf{x}$ and $y$.

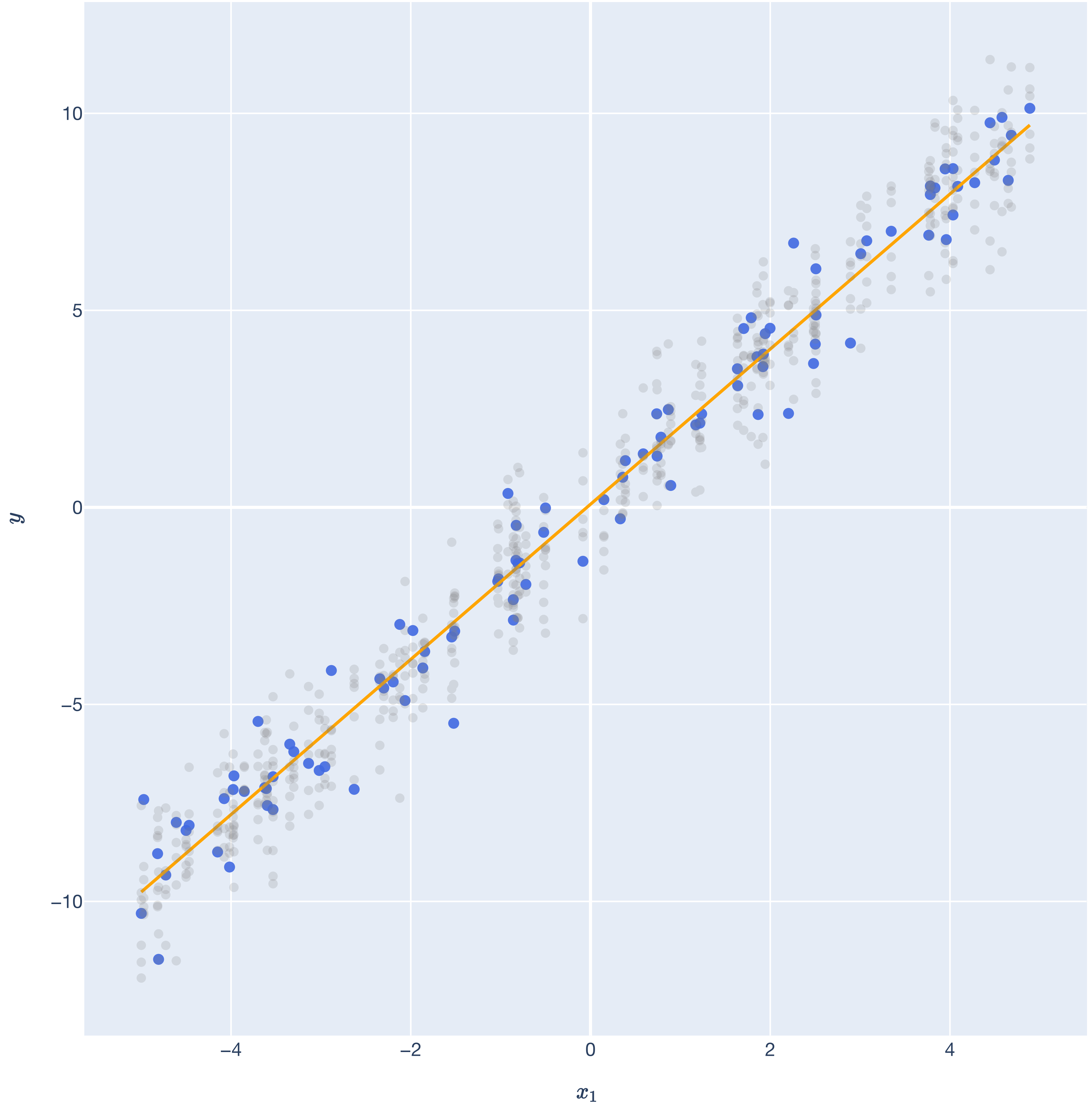
Some *randomness* $\epsilon_i$ models the unexplained relationship, where we assume

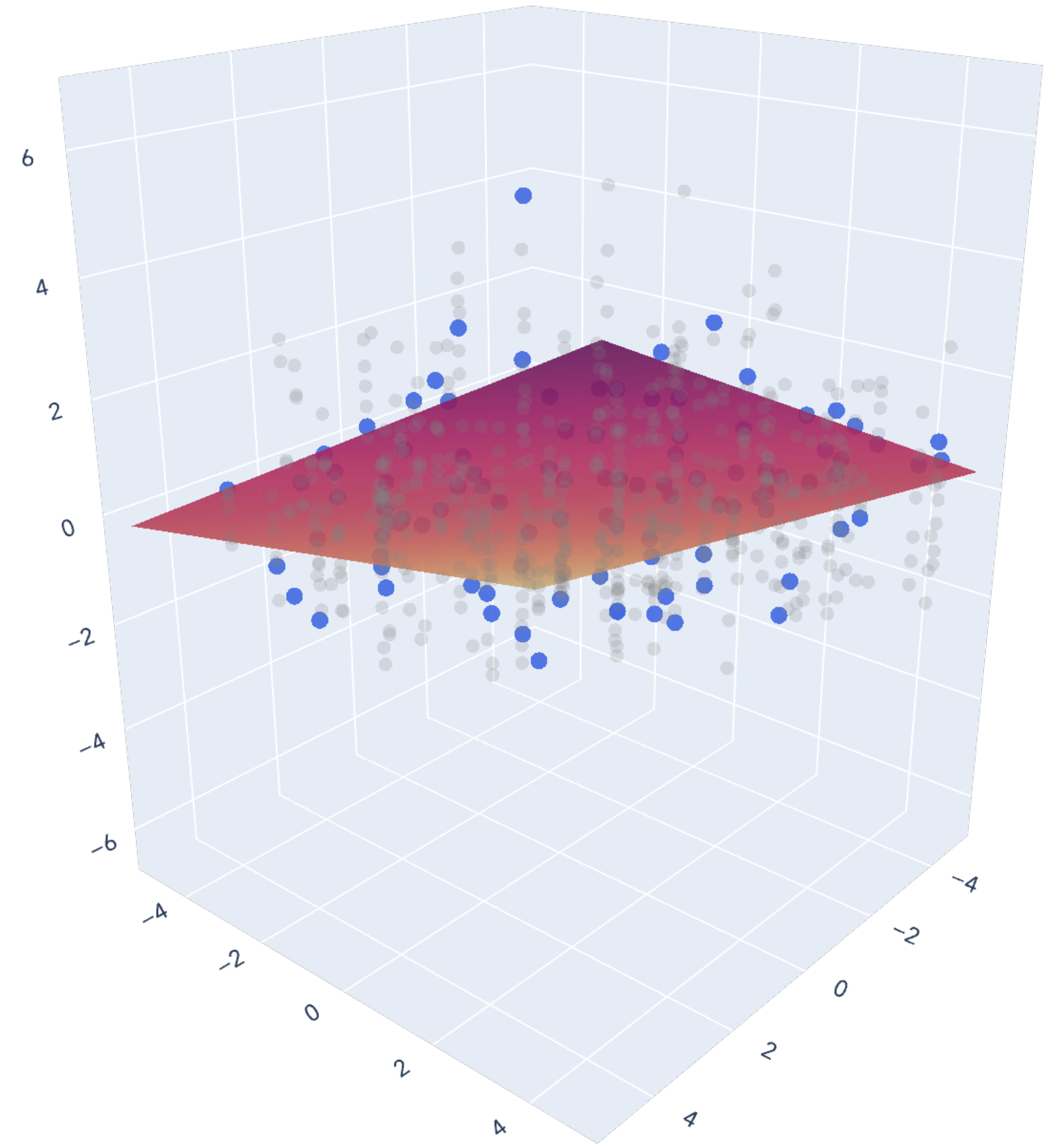
$$\mathbb{E}[\epsilon_i] = 0 \text{ and } \epsilon_i \text{ is independent of } \mathbf{x}_i.$$



# Regression with randomness

## Part 2: Notion of error and linear model



# Regression with randomness

## Part 2: Notion of error and linear model

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

## Part 2: Notion of error and linear model

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

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

# Regression with randomness

## Part 2: Notion of error and linear model

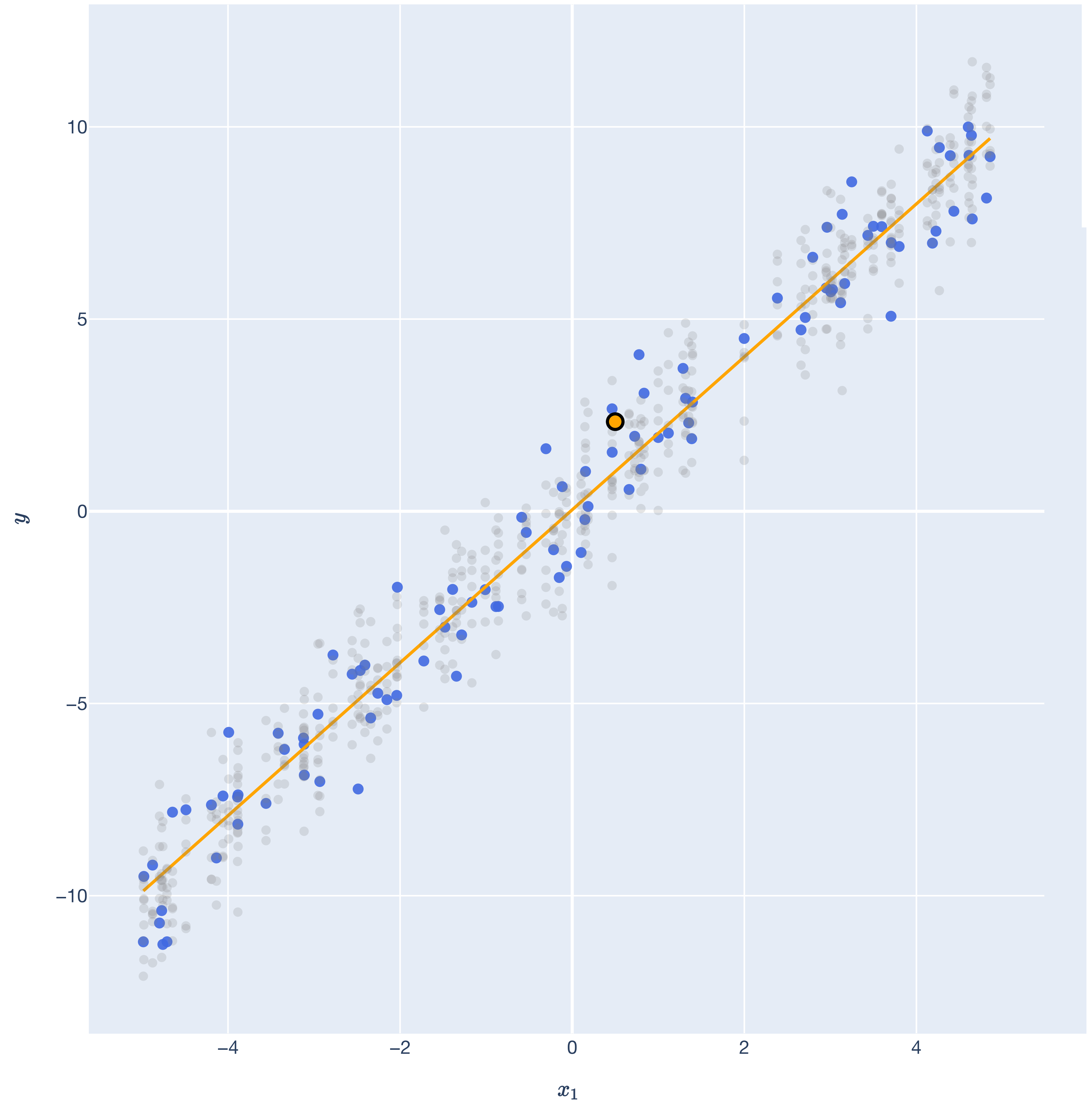
Each $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ pair is drawn from a *joint distribution*, $\mathbb{P}_{\mathbf{x},y}$

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

Our model $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ should predict well on new examples.

Notion of error for new $(\mathbf{x}_0, y_0)$ is still squared loss

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



# Regression with randomness

## Part 2: Notion of error and linear model

Random vector

Random variable

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

… from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$!

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

Our "model", a function that we hope "generalizes" well to new datapoint $(\mathbf{x}_0, y_0)$.

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

# Regression with randomness

## Part 3: Risk and empirical risk

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

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

New $(\mathbf{x}_0, y_0)$ is also from $\mathbb{P}_{\mathbf{x},y}$ and notion of error 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):

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

# Regression

## Part 3: Risk and empirical risk

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

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

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

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

*As long as $(\mathbf{x}_0, y_0)$ from same distribution, these are close as $n \to \infty$!*

# Regression

## Part 3: Risk and empirical risk

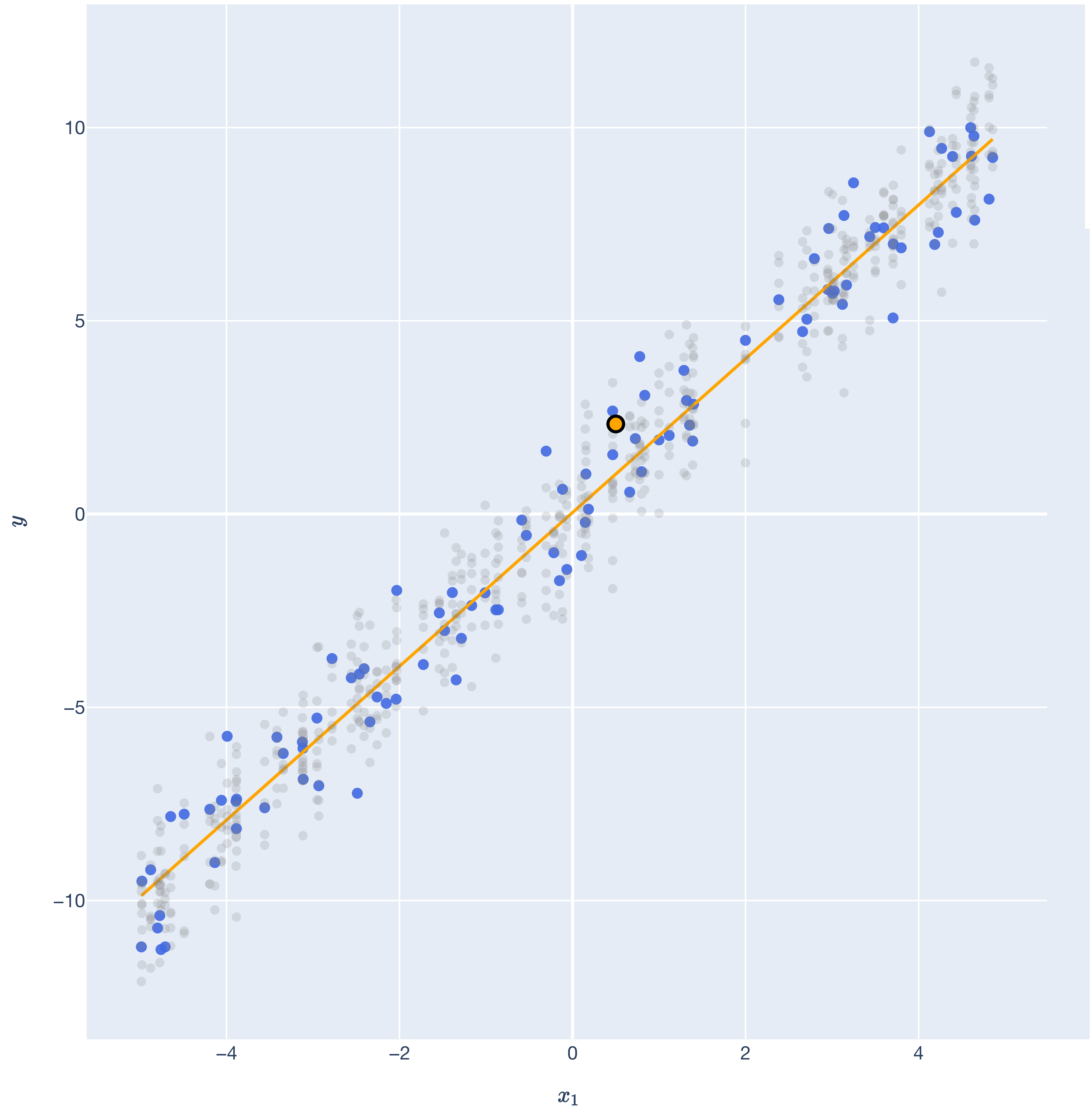
<u>Ultimate goal:</u> Find $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ that *generalizes* on a new $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$:

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

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

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

*As long as $(\mathbf{x}_0, y_0)$ from same distribution, these are close as $n \to \infty$!*



# Regression

## Part 3: Risk and empirical risk

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

Random vector

Random variable

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

… from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ over $\mathbb{R}^d \times \mathbb{R}$!

<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: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Our "model", a function that we hope "generalizes" well to new datapoint $(\mathbf{x}_0, y_0)$.

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:

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

Because we assume that $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y}$ as well!

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

# 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 $\mathbf{x}$.

$\mathrm{Var}(\epsilon \mid \mathbf{x}) = \sigma^2$, 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

## Conditional expectation and variance

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

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:

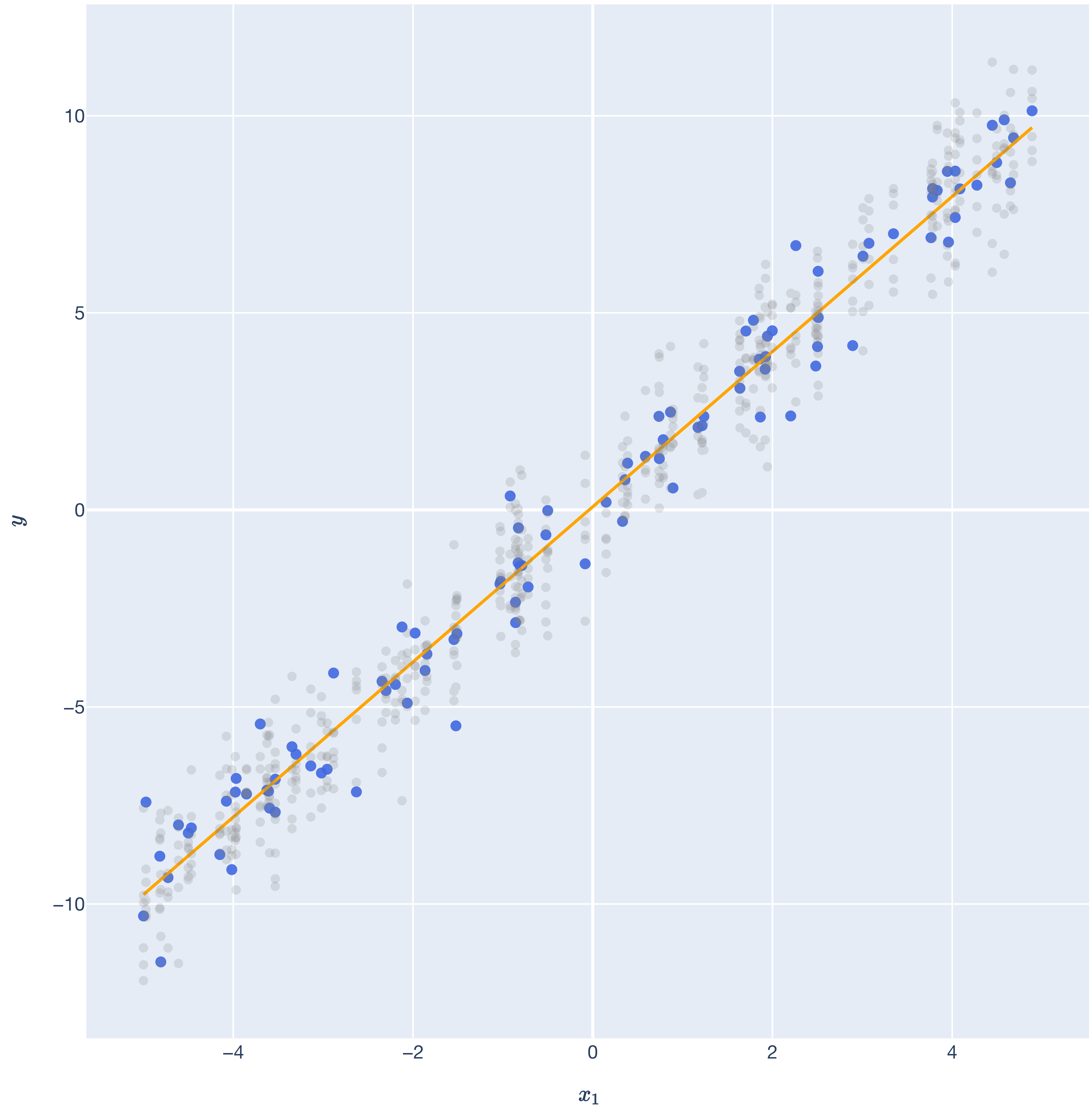
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

*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

## Intuition

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

Obtain the least squares estimator the same way:

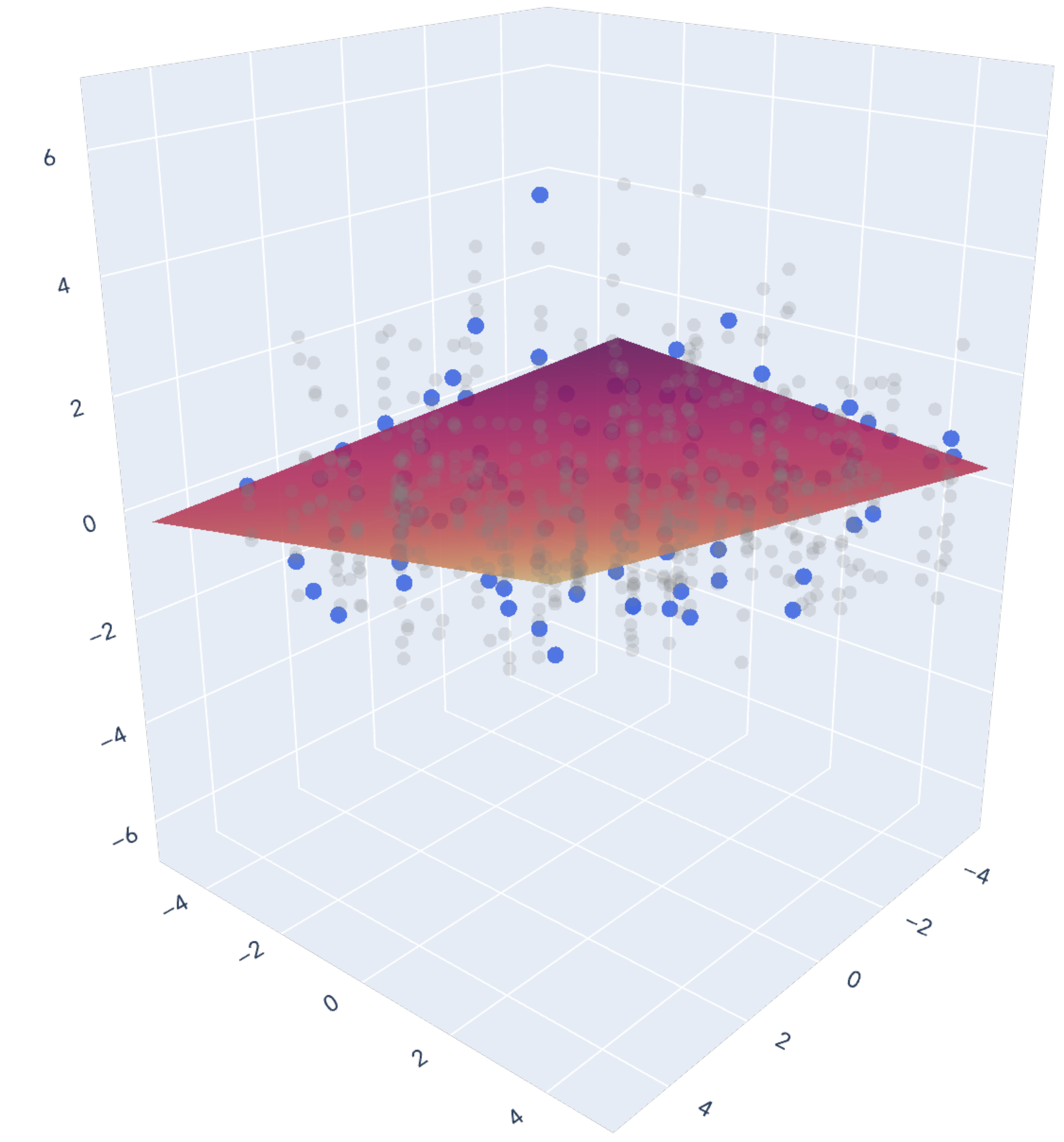
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

*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 $\mathbb{R}^d \times \mathbb{R}$ such that

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

where $\mathbf{w}^* \in \mathbb{R}^d$ and $\epsilon$ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$ 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: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$.

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

# Proof of OLS Expectation

## Use conditional expectation and error model, then tower rule

$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$ so $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$ and $\epsilon \in \mathbb{R}^n$

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

Take *conditional* expectation of OLS estimator and use the error model assumption on $\mathbf{y}$:

$$\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \mid \mathbf{X}]$$

$$= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boxed{(\mathbf{X}\mathbf{w}^* + \epsilon)} \mid \mathbf{X}]$$

Replace $\mathbf{y}$ with error model assumption.

$$= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\epsilon) \mid \mathbf{X}]$$

$$= \mathbb{E}[\boxed{\mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\epsilon \mid \mathbf{X}}] = \mathbf{w}^* \implies \boxed{\mathbb{E}[\hat{\mathbf{w}}] = \mathbb{E}[\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}]]} = \mathbf{w}^*$$

"Pulling out what's known" property of cond. expectation.

Tower rule/law of total expectation property.

# Proof of OLS Variance

## Part 1: Take conditional variance and use error model

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ so } \mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon, \text{ where } \mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{y} \in \mathbb{R}^n \text{ and } \epsilon \in \mathbb{R}^n$$

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

Take *conditional* variance of OLS estimator and use the error model assumption on $\mathbf{y}$:

Definition of OLS $\hat{\mathbf{w}}$ and error model

From previous proof: expectation of $\hat{\mathbf{w}}$.

$$\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbb{E}[(\hat{\mathbf{w}} - \mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}])(\hat{\mathbf{w}} - \mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}])^\top \mid \mathbf{X}]$$

$$= \mathbb{E}[((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \epsilon) - \mathbf{w}^*)((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \epsilon) - \mathbf{w}^*)^\top \mid \mathbf{X}]$$

$$= \mathbb{E}[(\mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon - \mathbf{w}^*)(\mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon - \mathbf{w}^*)^\top \mid \mathbf{X}]$$

$$= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon\epsilon^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}]\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$$

Cond. expectation property: taking out what's known.

# Proof of OLS Variance

## Part 2: Understand variance (covariances) of $\epsilon\epsilon^\top$

$$y_i = \mathbf{x}_i^\top \mathbf{w}* + \epsilon_i \text{ so } \mathbf{y} = \mathbf{X}\mathbf{w}* + \epsilon\text{, where } \mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n \text{ and } \epsilon \in \mathbb{R}^n$$

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

Remains to understand the expectation of the matrix $\epsilon\epsilon^\top \in \mathbb{R}^{n\times n}$:

$$\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

For $n = 3$, this looks like:

$\mathbb{E}[\epsilon_i^2] = \mathbb{E}[(\epsilon_i - 0)(\epsilon_i - 0)] = \mathrm{Cov}(\epsilon_i, \epsilon_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$

$\epsilon$ are indep. of $\mathbf{X}$ in error model.

$$\epsilon\epsilon^\top = \begin{bmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \epsilon_1\epsilon_3 \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \epsilon_2\epsilon_3 \\ \epsilon_3\epsilon_1 & \epsilon_3^2 & \epsilon_3\epsilon_3 \end{bmatrix} \text{ so } \mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}] = \mathbb{E}[\epsilon\epsilon^\top] = \begin{bmatrix} \mathbb{E}\epsilon_1^2 & \mathbb{E}\epsilon_1\epsilon_2 & \mathbb{E}\epsilon_1\epsilon_3 \\ \mathbb{E}\epsilon_2\epsilon_1 & \mathbb{E}\epsilon_2^2 & \mathbb{E}\epsilon_2\epsilon_3 \\ \mathbb{E}\epsilon_3\epsilon_1 & \mathbb{E}\epsilon_3\epsilon_2 & \mathbb{E}\epsilon_3^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

For $i \neq j$, $\mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}[(\epsilon_i - 0)(\epsilon_j - 0)] = \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ by independence

# Random Vectors

## Variance and Covariance Matrix

The variance of a random vector generalizes to the covariance matrix

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

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

In this class, a random vector's variance *is* its covariance:

$$\mathrm{Var}(\mathbf{x}) := \boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$

# Proof of OLS Variance

## Part 3: Cancel out inverse and use law of total variance

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ so } \mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon \text{, where } \mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{y} \in \mathbb{R}^n \text{ and } \epsilon \in \mathbb{R}^n$$

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

Remains to understand the expectation of the matrix $\epsilon\epsilon^\top \in \mathbb{R}^{n \times n}$:

$$\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

Therefore, $\mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}] = \sigma^2 \mathbf{I}$.

Uses inverse!

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

and $\mathrm{Var}[\hat{\mathbf{w}}] = \mathbb{E}[\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}]] + \mathrm{Var}[\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}]] = \mathbb{E}[\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}] + \mathrm{Var}[\mathbf{w}^*] = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}]$

Law of total variance (PS5).

From first part (expectation of $\hat{\mathbf{w}}$)

# Statistics of OLS

## Theorem

**Theorem (Statistical properties of OLS).** Let $\mathbb{P}_{\mathbf{x},y}$ be a joint distribution $\mathbb{R}^d \times \mathbb{R}$ such that

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

where $\mathbf{w}^* \in \mathbb{R}^d$ and $\epsilon$ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ 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: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$.

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

# 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.

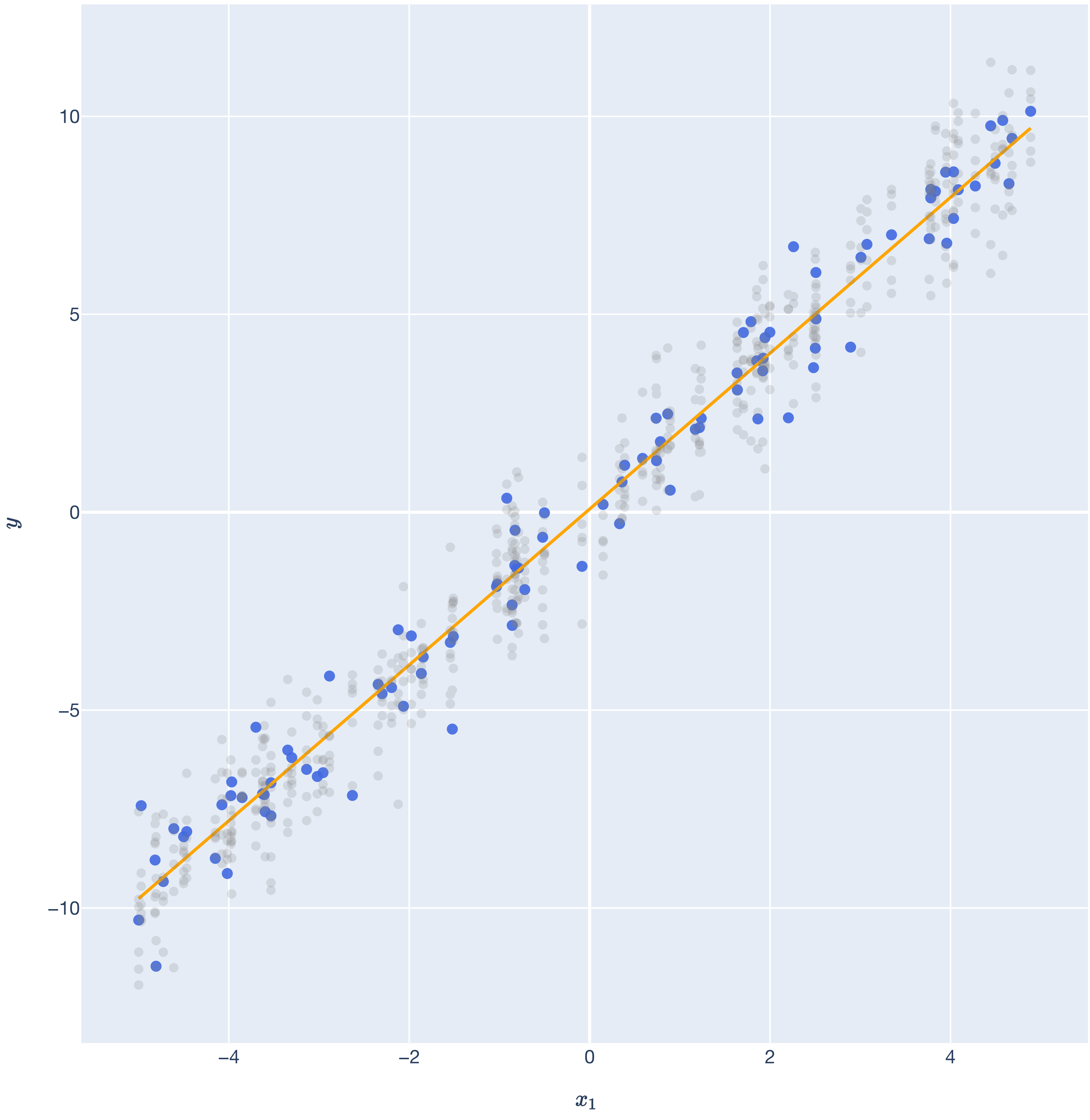
**Expectation, variance, and covariance.** Review of these basic summary statistics of random variables and common properties.

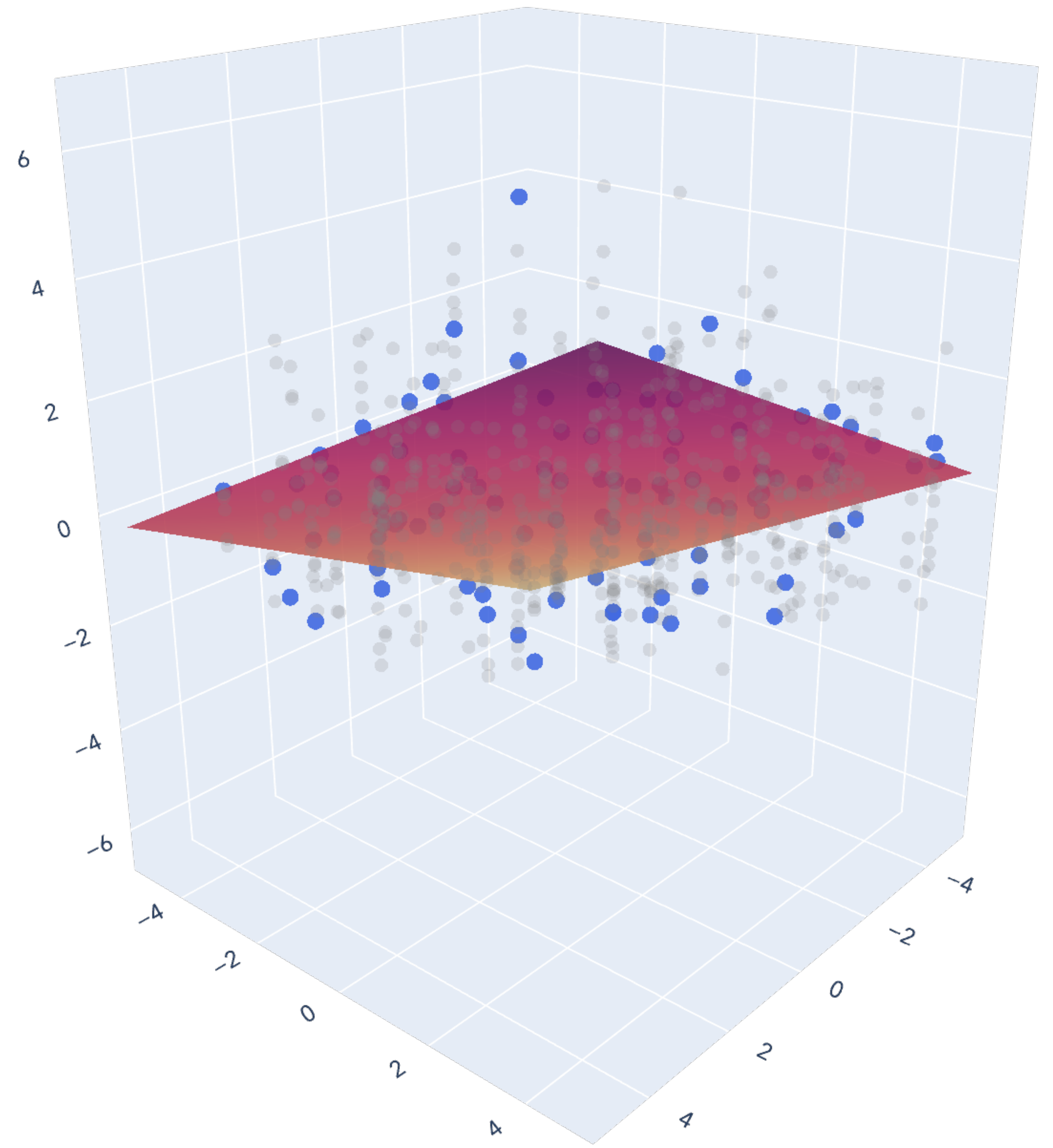
**Random vectors.** Introduce the idea of a *random vector*, which is just a list of multiple random variables. Discuss generalizations of expectation and variance to random vectors.

**Data as random, statistical model of ML.** Introduce a 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