

# 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.

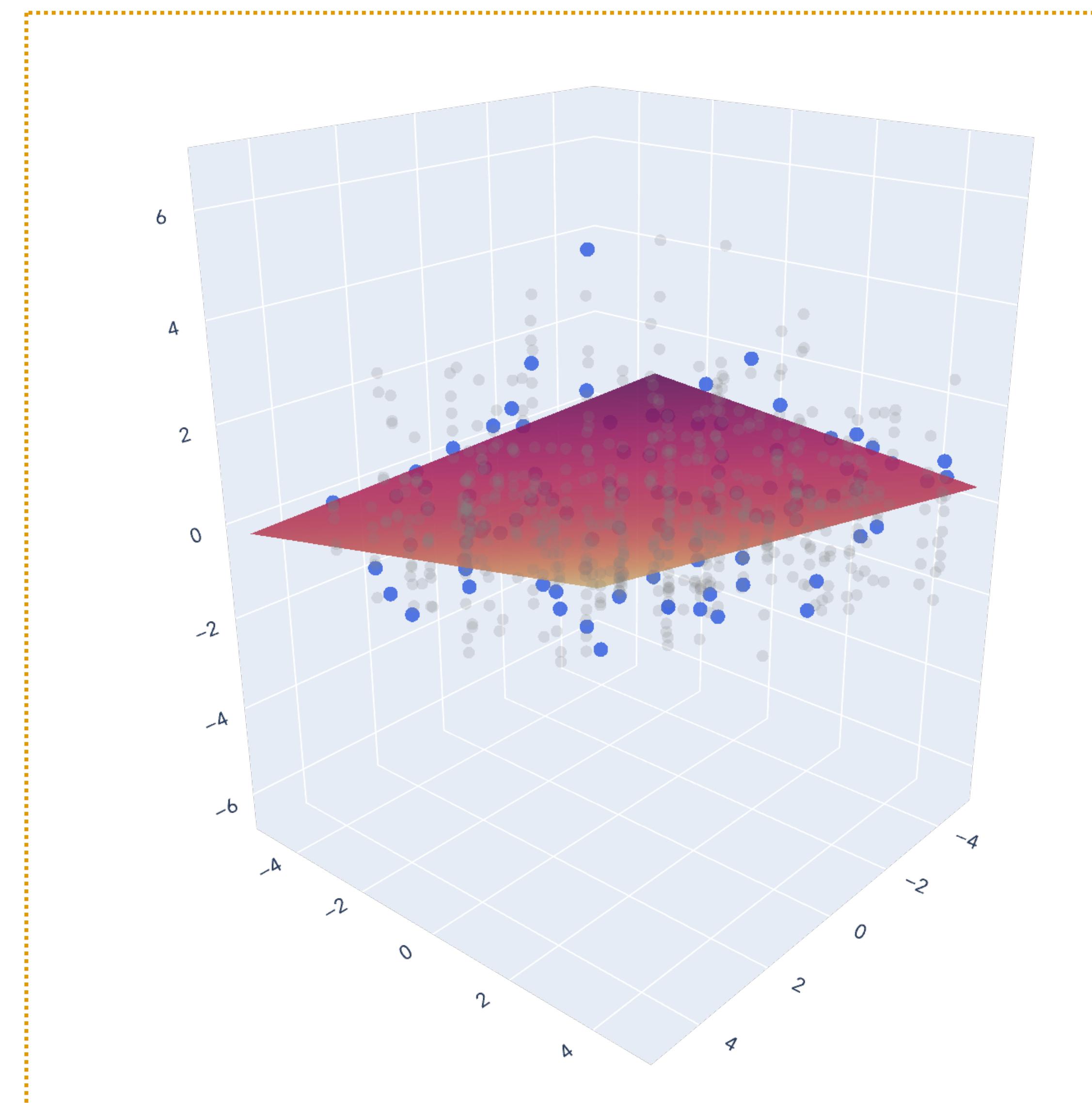
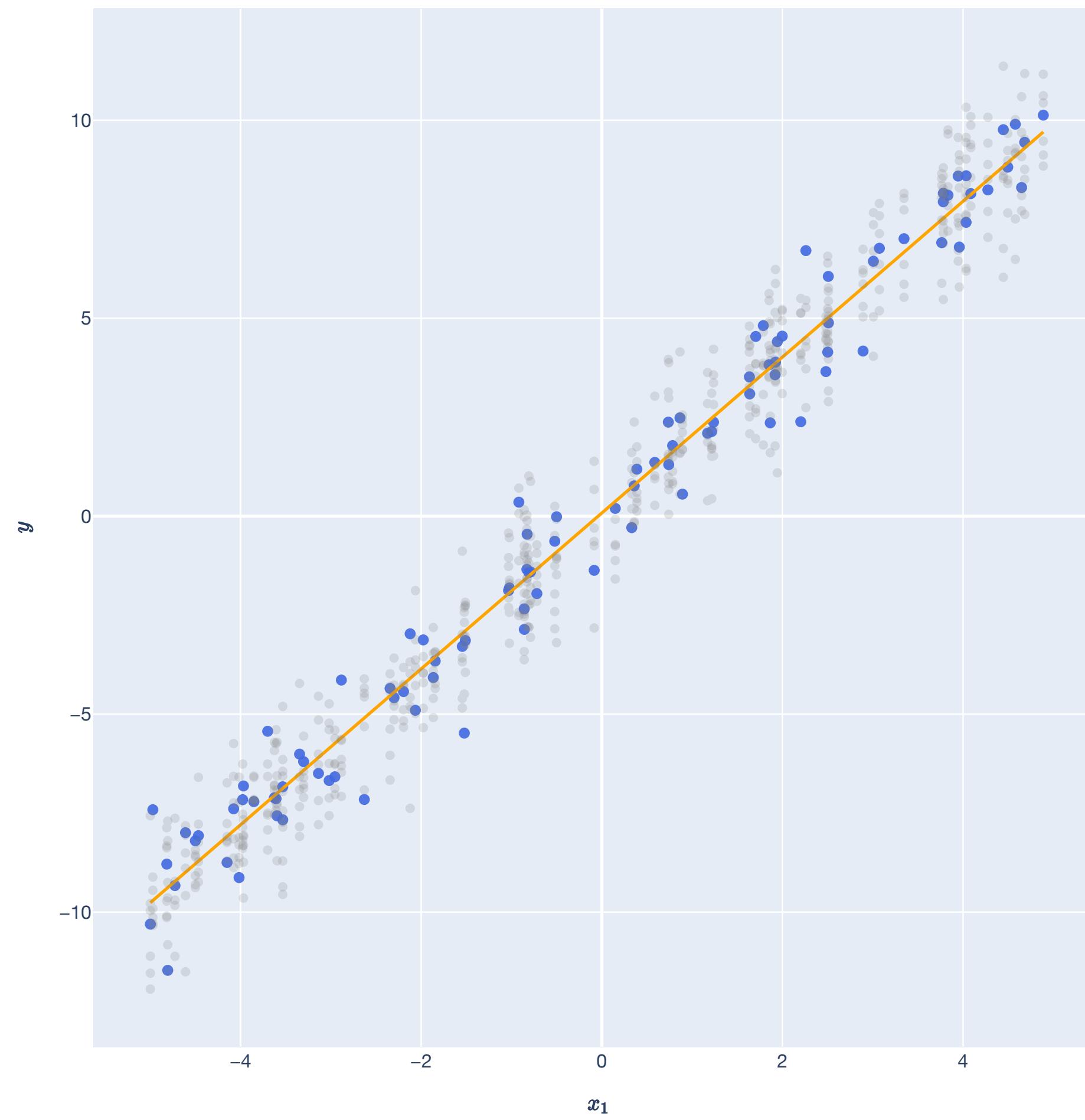
**Expectation, variance, and covariance.** Review of these basic summary statistics of random variables and common properties.

**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.

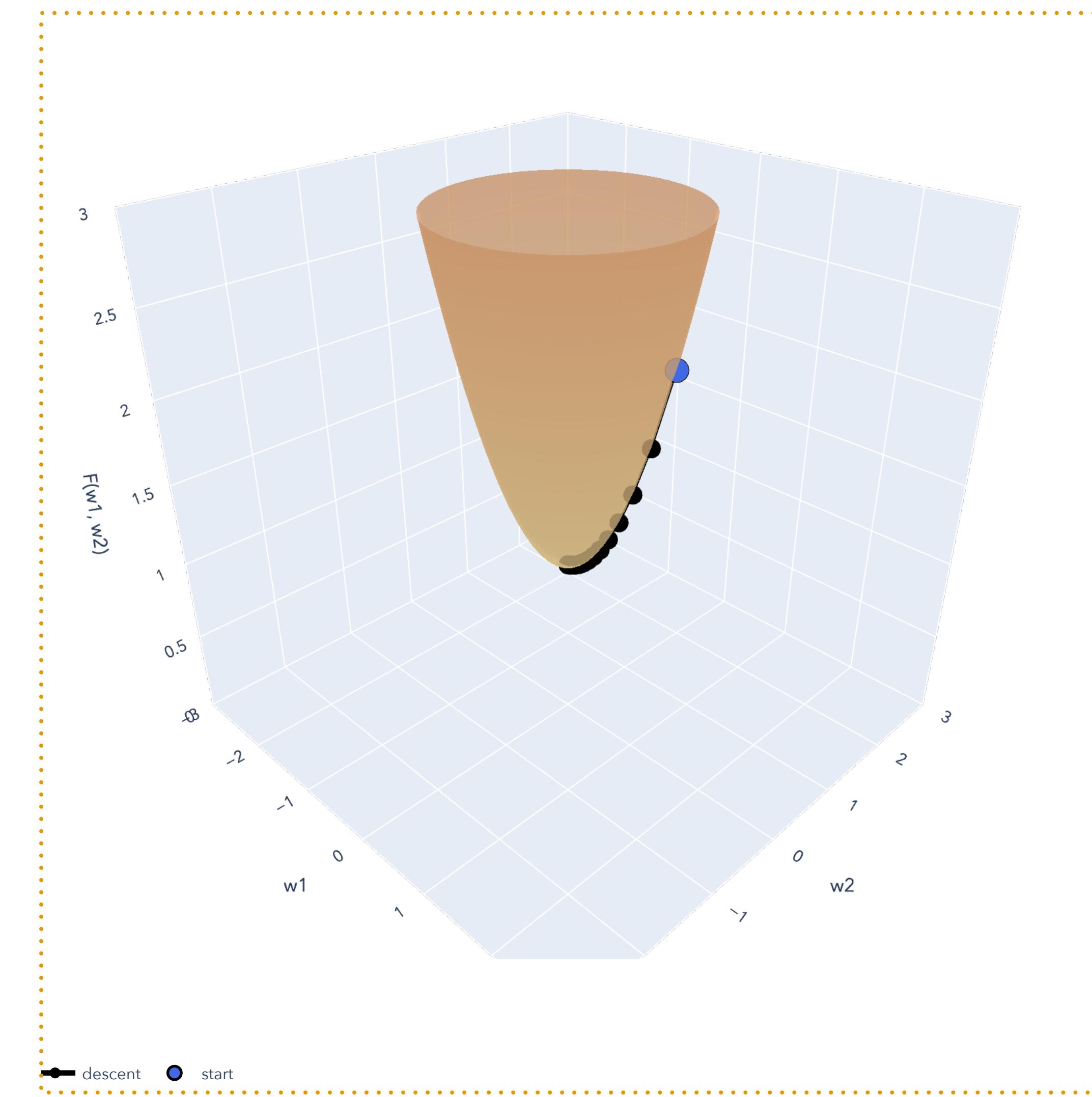
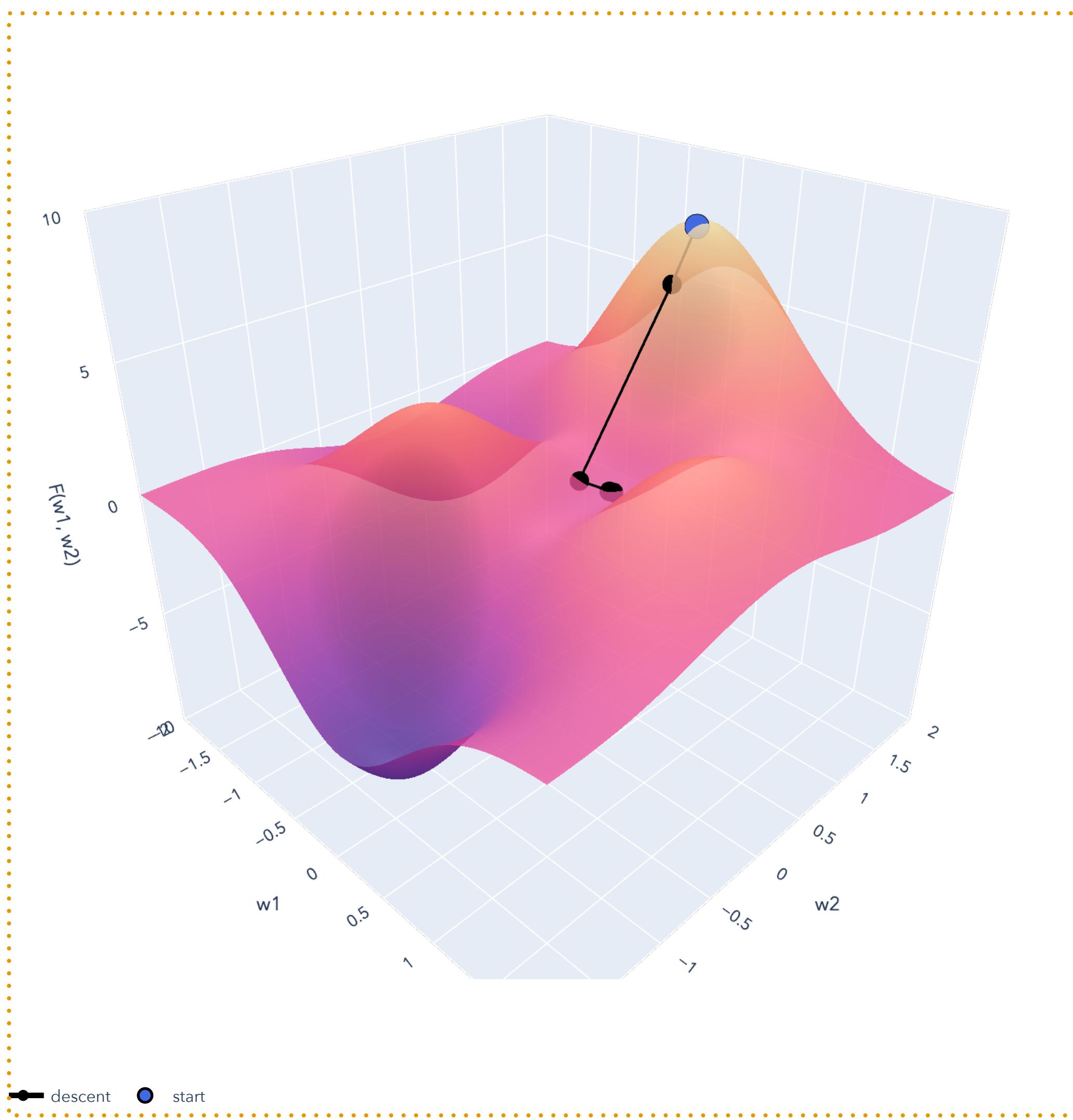
# Lesson Overview

## Big Picture: Least Squares



# Lesson Overview

## Big Picture: Gradient Descent



# Motivation

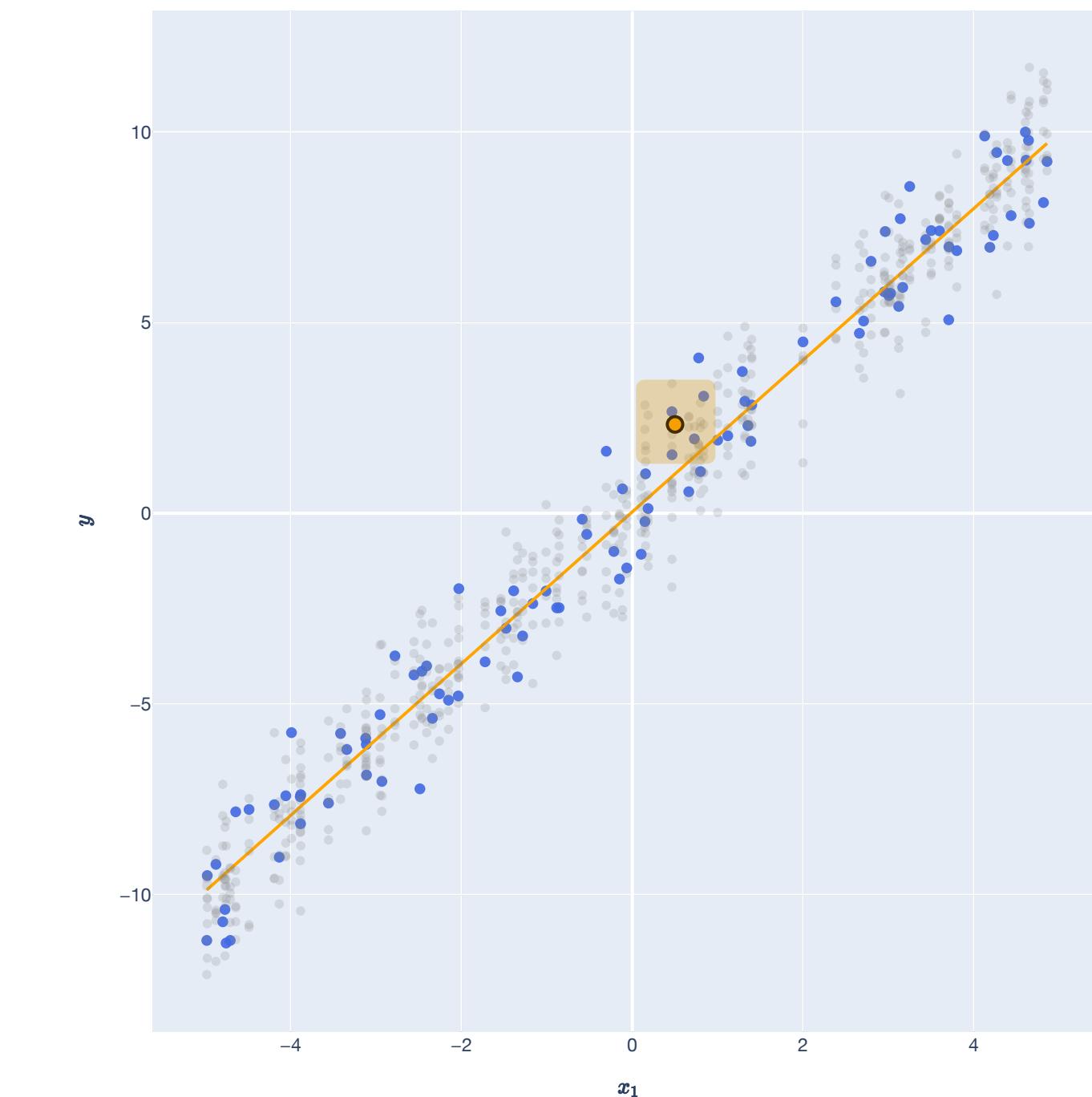
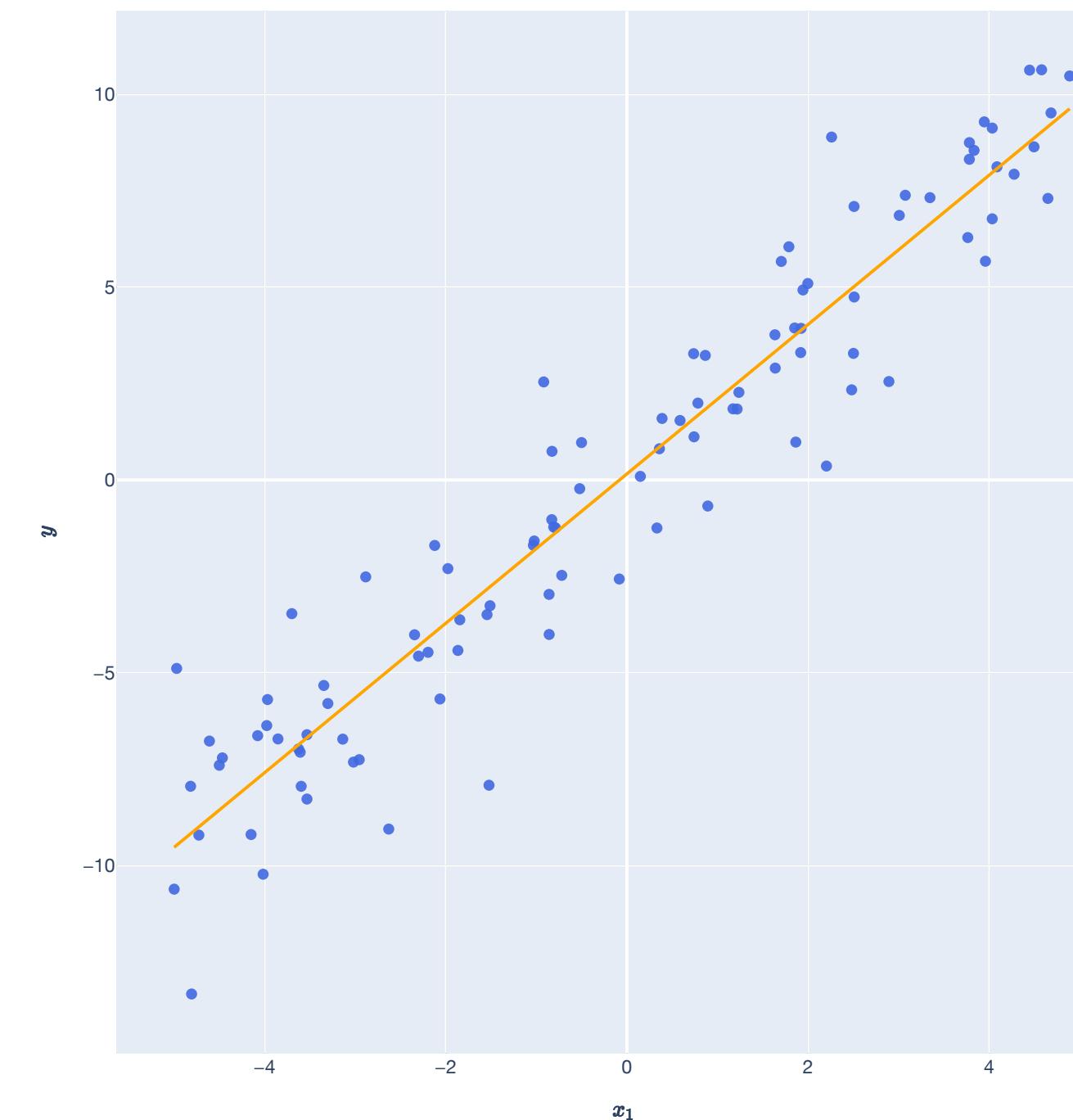
Data as randomly distributed

# Regression

## Setup

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

Hope training data looks like new data: Generalize on new  $x_0$ .



# Regression

## Setup (Example View)

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

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

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

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

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)

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

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

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

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

Ultimate 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

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

Observed: 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  $\mathbf{x}_i \in \mathbb{R}^d$

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

Random variable  $y_i$

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

... 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  $w_1, \dots, w_d$ .

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

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.*)  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , 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:

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

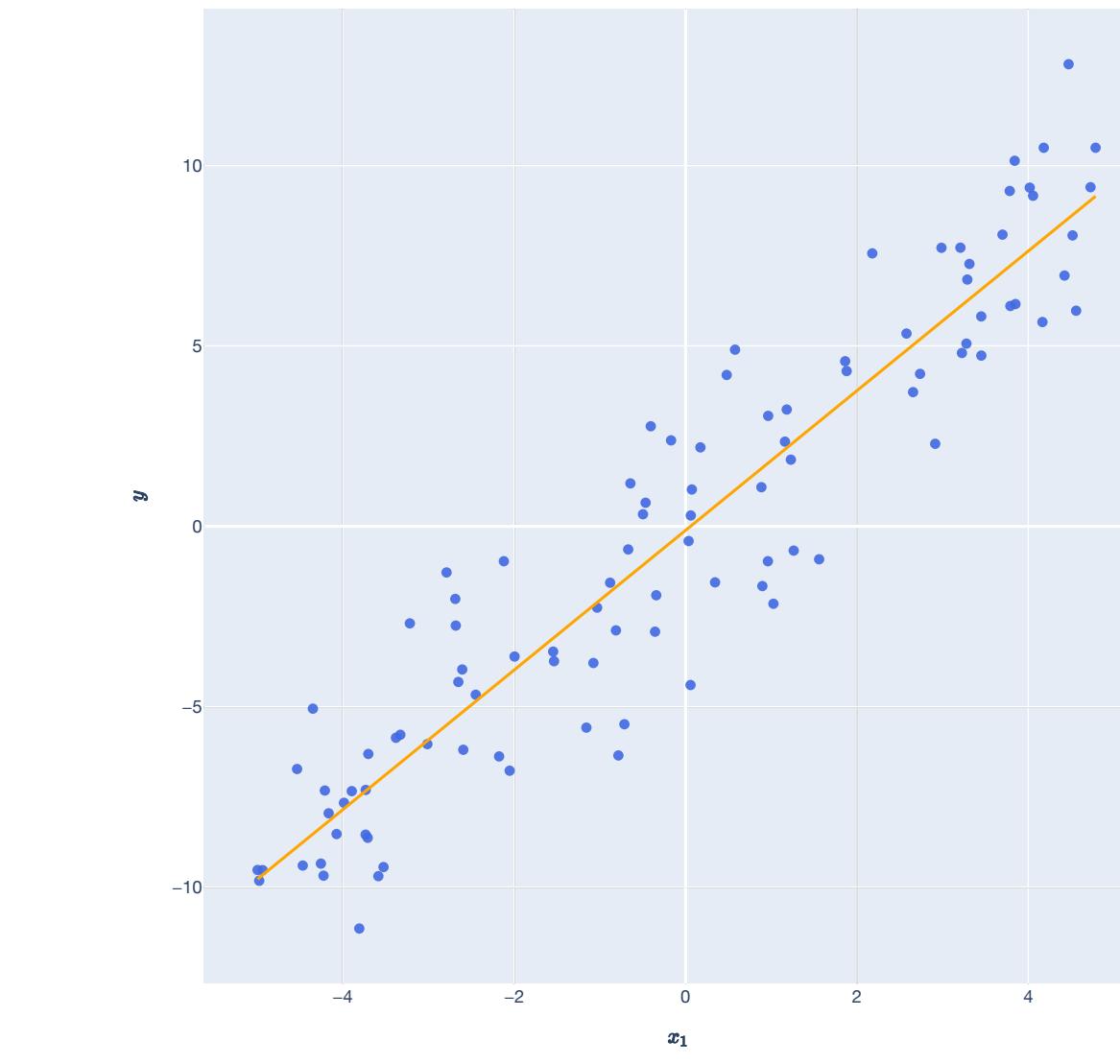
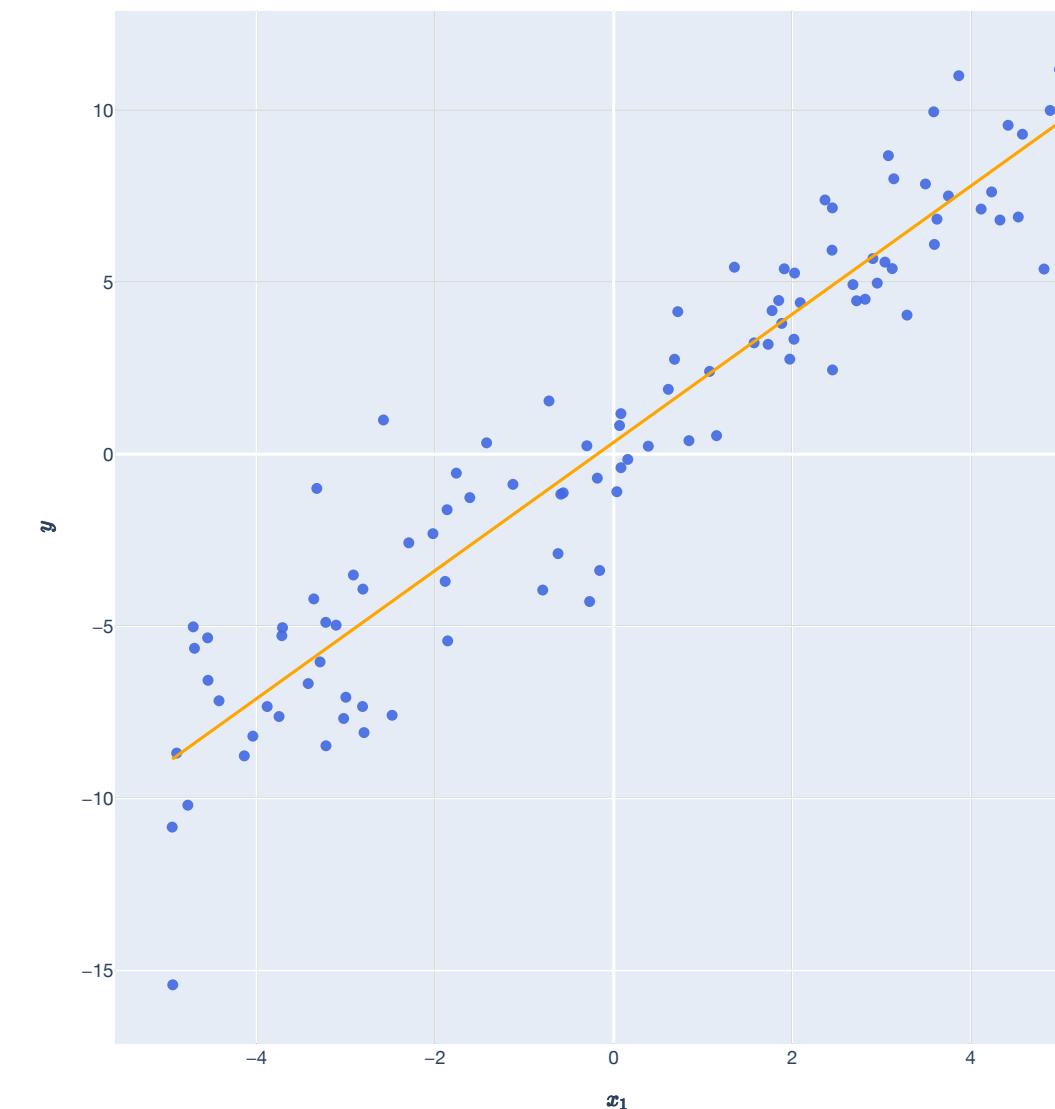
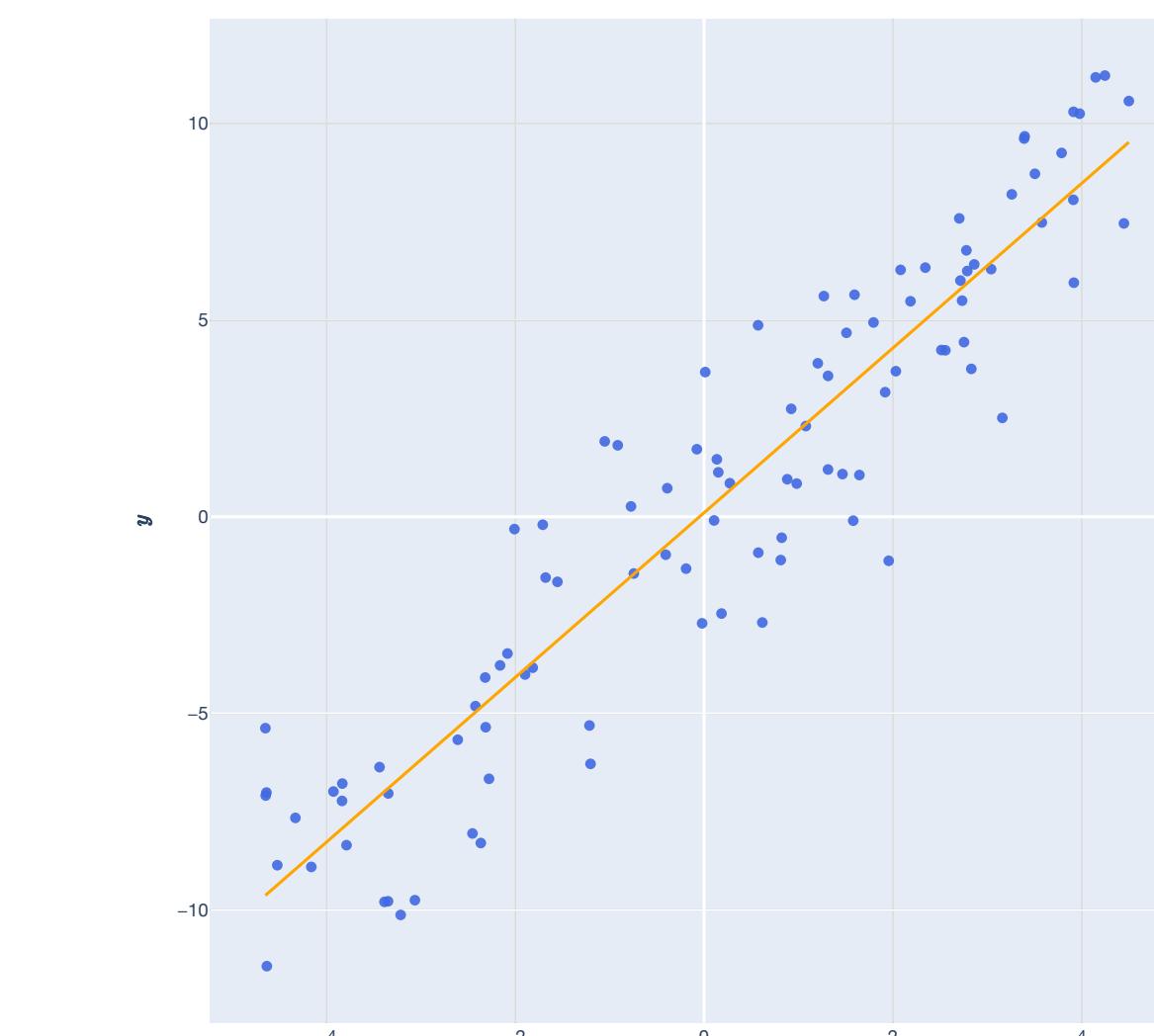
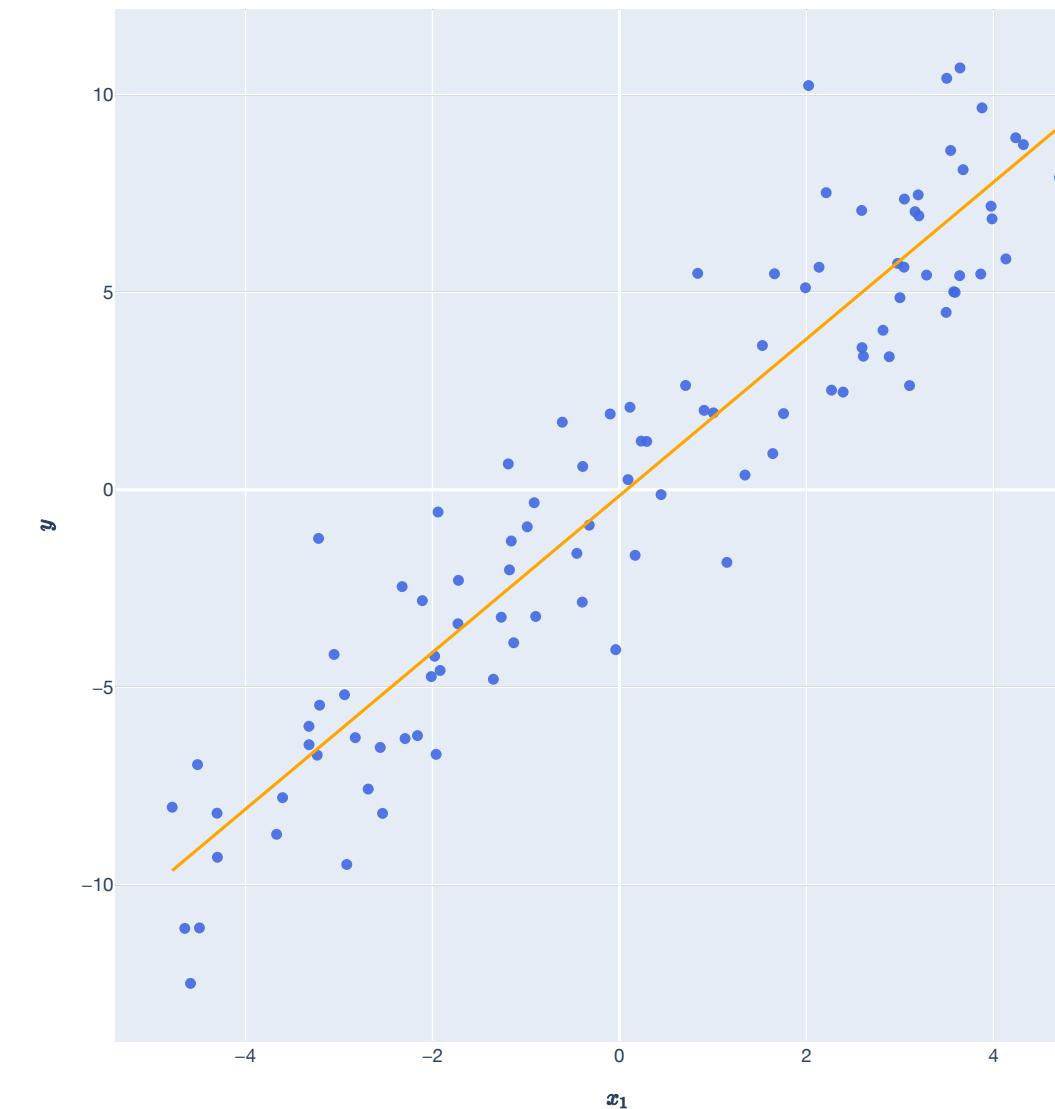
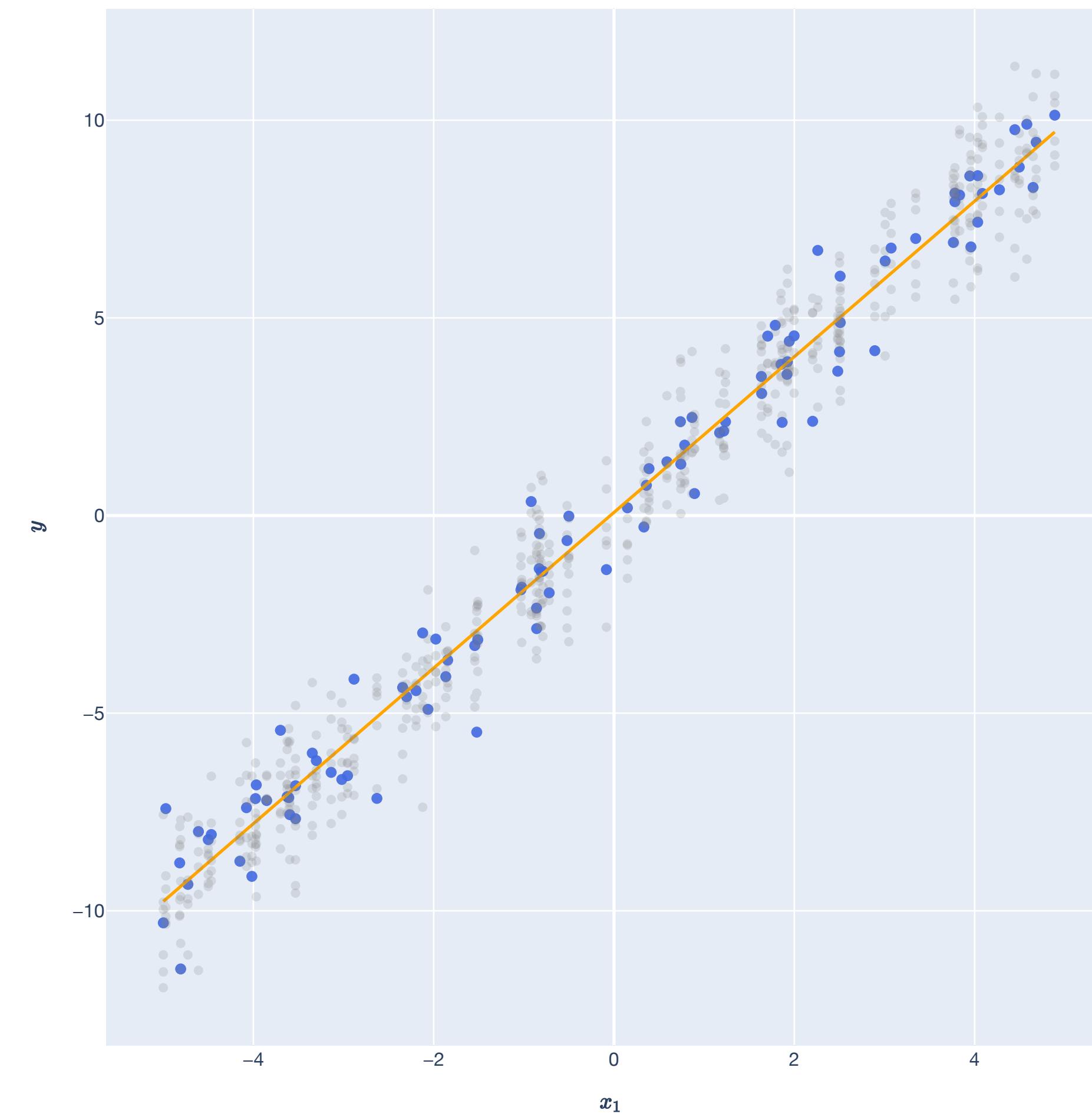
...but we can't know the true distribution, so as a substitute, minimize the empirical risk:

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

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

# Regression

## Modeling randomness

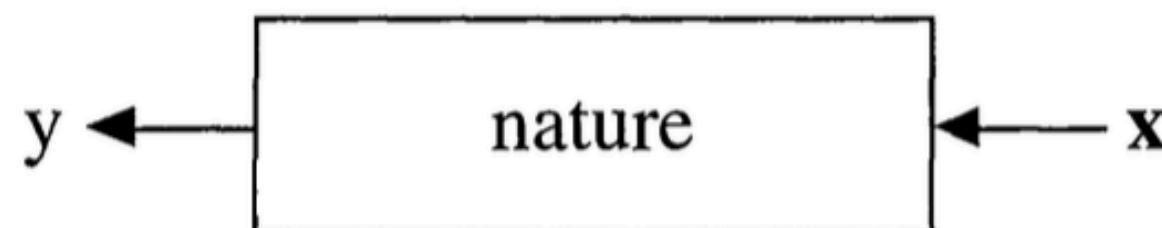


# The “Story” of Probability and Statistics

## 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  $\mathbf{x}$  (independent variables) go in one side, and on the other side the response variables  $\mathbf{y}$  come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



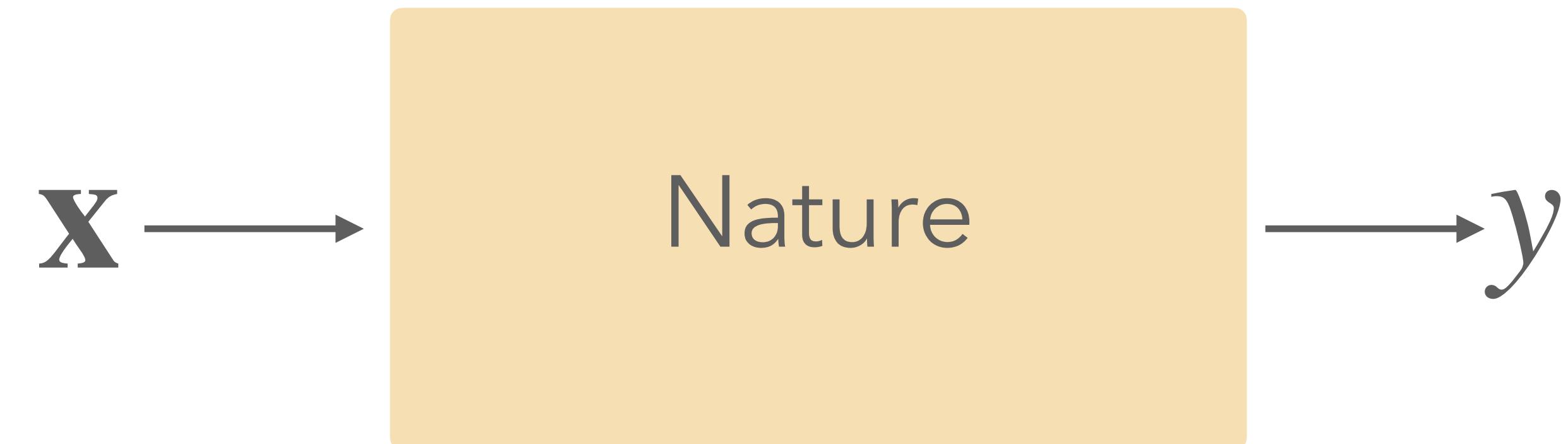
There are two goals in analyzing the data:

*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.

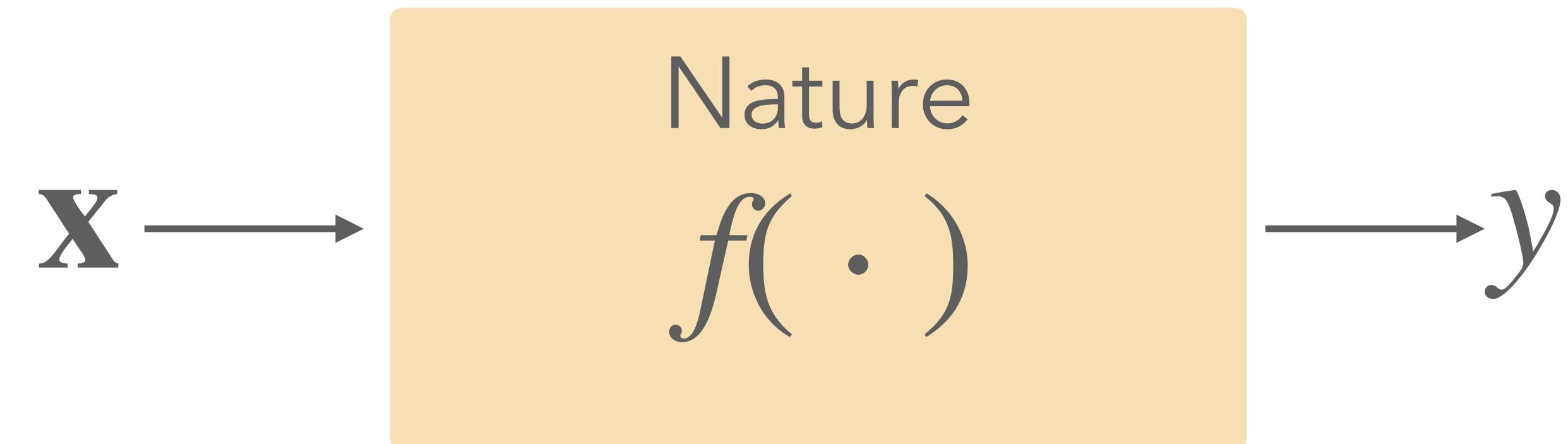
# The “Story” of Probability and Statistics

Modeling randomness



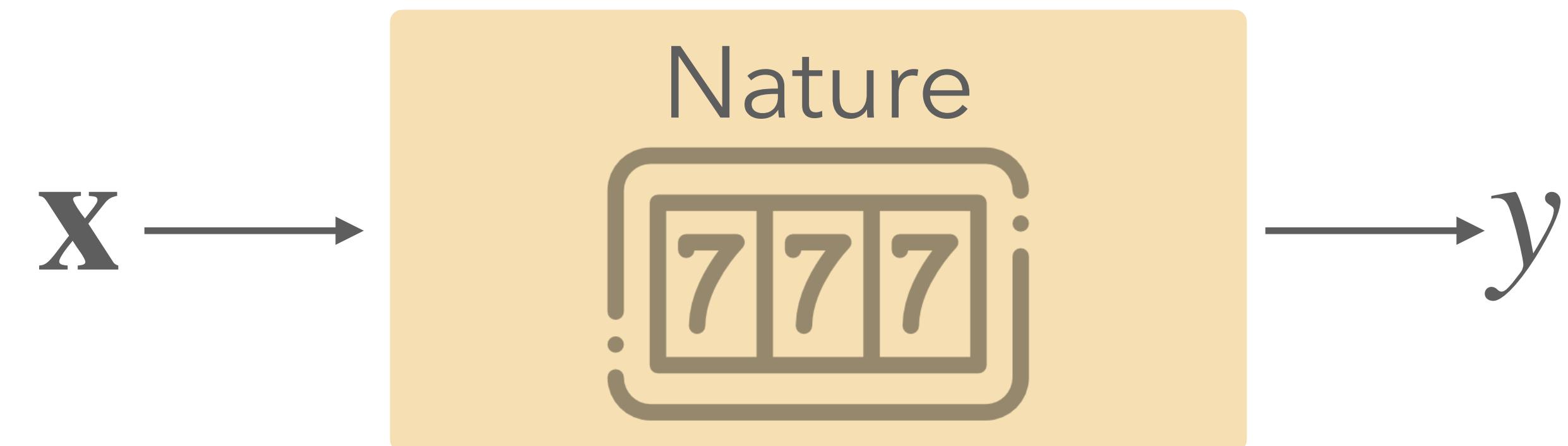
# The “Story” of Probability and Statistics

Modeling randomness



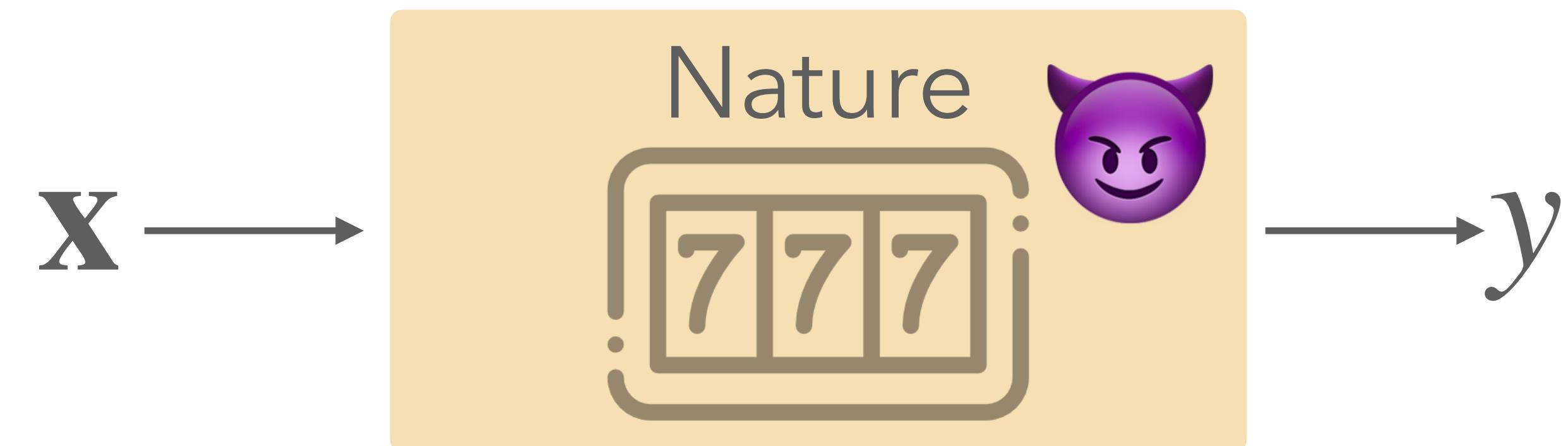
# The “Story” of Probability and Statistics

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  $T$ .

*What are the possible outcomes of this experiment?*

$H H$	$T H$
$H T$	$T T$

$\Omega$

# 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:

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

$HH$	$TH$
$HT$	$TT$
$\Omega$	

# Events

## Intuition and definition

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

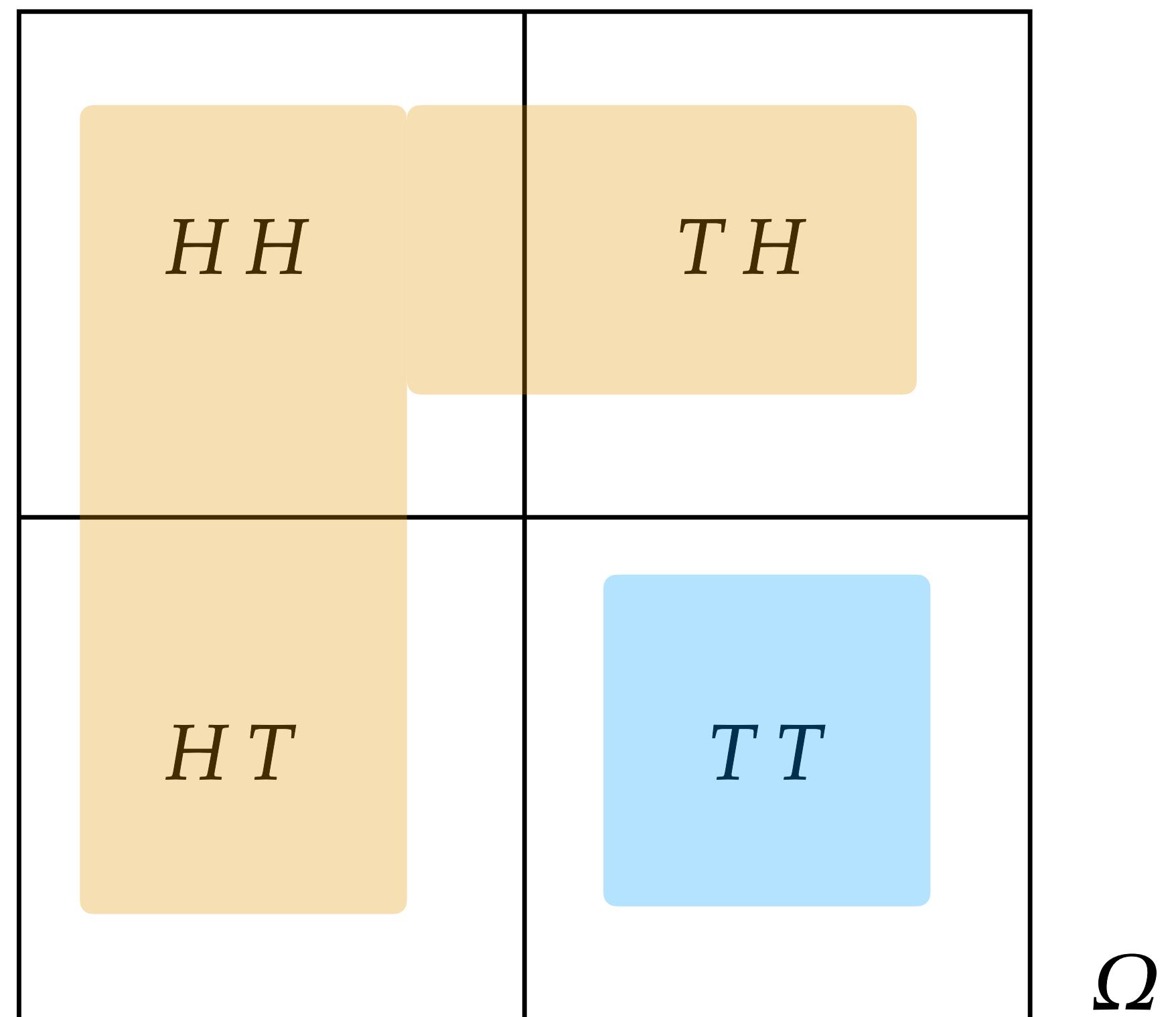
Denote a collection of events  $\mathcal{A}$ .

Example:

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

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

$\mathcal{A} = \{\emptyset, \{HH\}, \{HT\}, \dots, \{HH, HT, TH, TT\}\}$



# Events

## Intuition and definition

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

$A \cup B$  (union)

$A \cap B$  (intersection)

$A^C$  (complement)

Example:

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

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

$HH$	$TH$
$HT$	$TT$

$\Omega$

# Probability Measure

## Intuition and definition

A probability measure is a set function

$$\mathbb{P} : \mathcal{A} \rightarrow [0,1]$$

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

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

“degree of belief” or “long-run frequency.”

Or just the “mass” of a particular subset!

$\mathbb{P}(\{HH\}) = 0.2$	$HH$	$TH$	$\Omega$
	$HT$	$TT$	$\Omega$

# 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  $A_1, A_2, A_3, \dots$

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \dots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \dots$$

also known as countable additivity.

$$\mathbb{P}(\{HH, TH, HT, TT\}) = 1$$

$HH$	$TH$
$HT$	$TT$
	$\Omega$

# Probability Measure

## Properties of probability measures

1. **Complements.** For any event  $A \in \mathcal{A}$ , the probability of the complement is:  
 $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$ .
2. **Subsets of events.** For two events  $A, B \in \mathcal{A}$ , if  $A \subseteq B$ , then:  $\mathbb{P}(B) \leq \mathbb{P}(A)$ .
3. **Unions of events.** For any two events  $A, B \in \mathcal{A}$ ,  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ .
4. **Union bound.** For any finite collection of events  $A_1, \dots, A_n$ ,

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

# Probability Measure

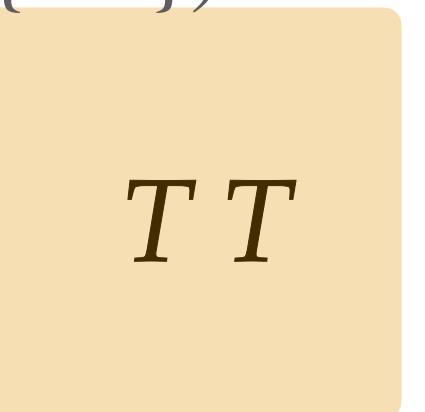
## Example Measures

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

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

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

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

$\mathbb{P}(\{HH\}) = 1/4$ 	$\mathbb{P}(\{TH\}) = 1/4$ 
$\mathbb{P}(\{HT\}) = 1/4$ 	$\mathbb{P}(\{TT\}) = 1/4$ 

$\Omega$

# Conditional Probabilities

## Intuition and definition

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

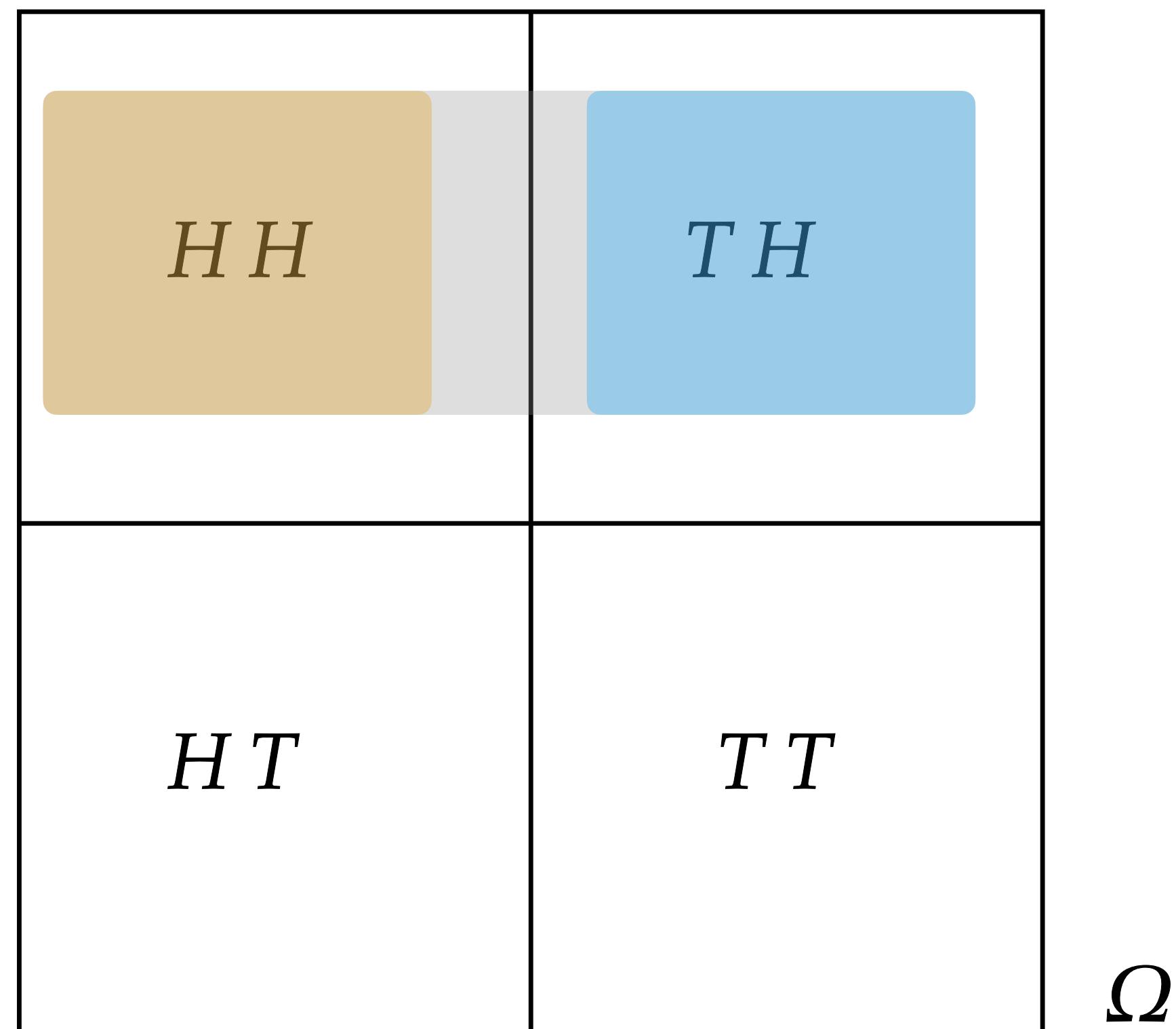
$$\mathbb{P}(B | A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

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:

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

This easily gives us Bayes' rule:

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

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

# Conditional Probabilities

## Law of Total Probability

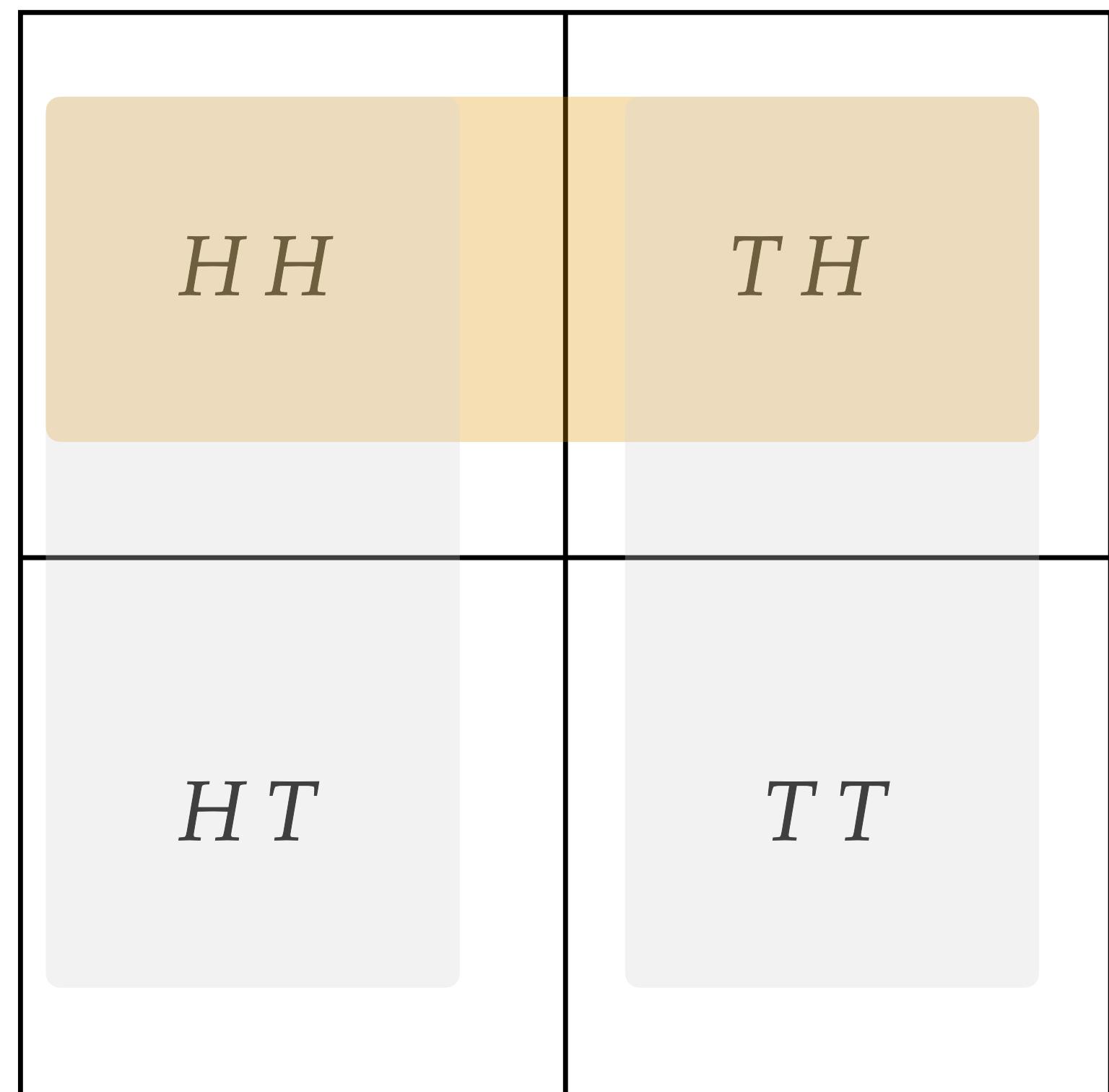
The [law of total probability](#) allows us to chop up probabilities into an exact sum of distinct events.

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

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

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

*Super useful: commit this to memory!*



# Probability Space

## Intuition and definition

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

Example:

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

$$\mathcal{A} = \{\emptyset, \{HH\}, \{HT\}, \dots, \{HH, HT, TH, TT\}\}$$

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

$HH$	$TH$
$HT$	$TT$
	$\Omega$

# Probability Space

## Intuition and definition

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

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

$H H$	$T H$
$H T$	$T T$
$\Omega$	

# Random Variables

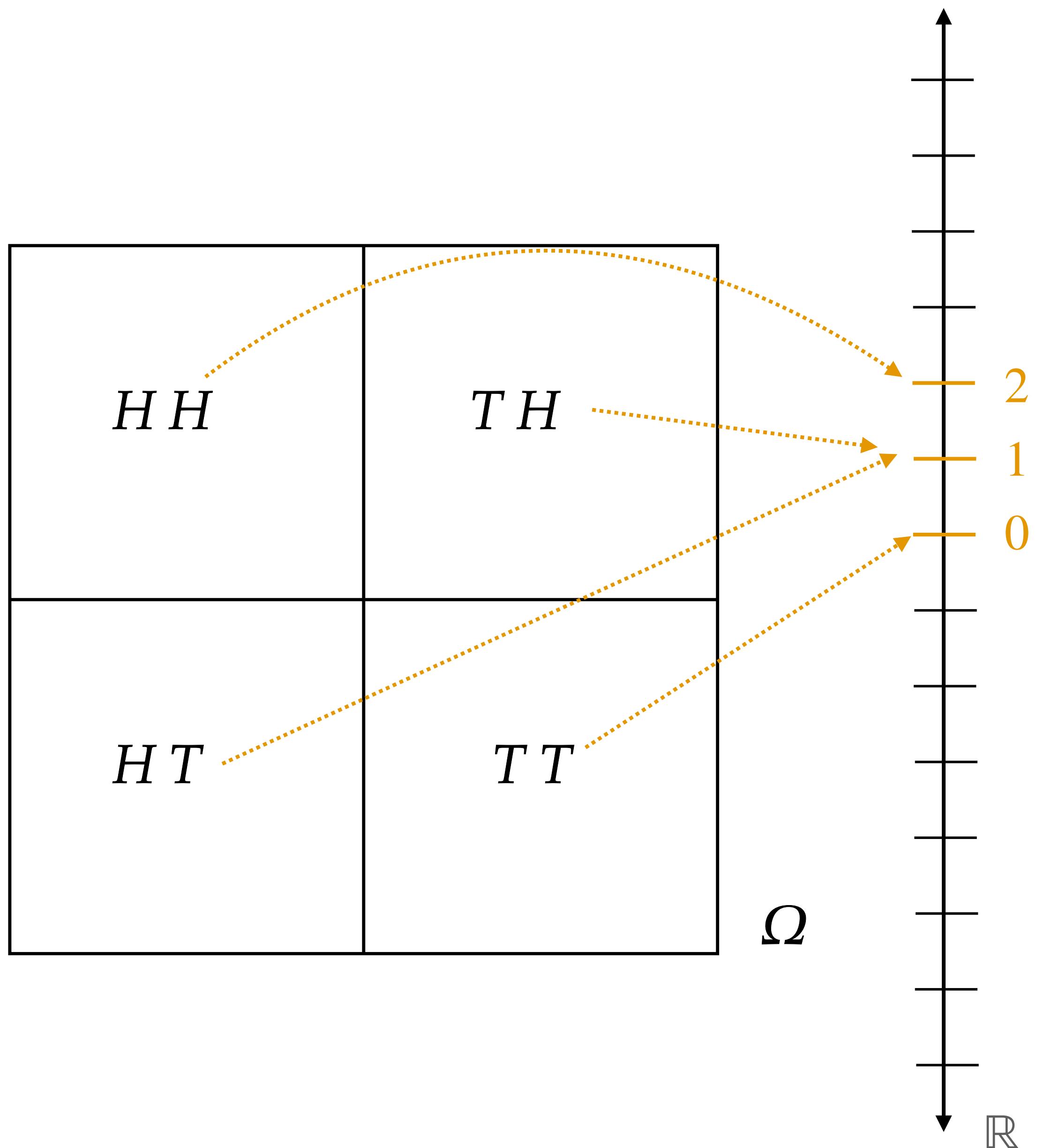
Example: Flipping 2 fair coins

Consider the following function:

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

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

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



# Random Variables

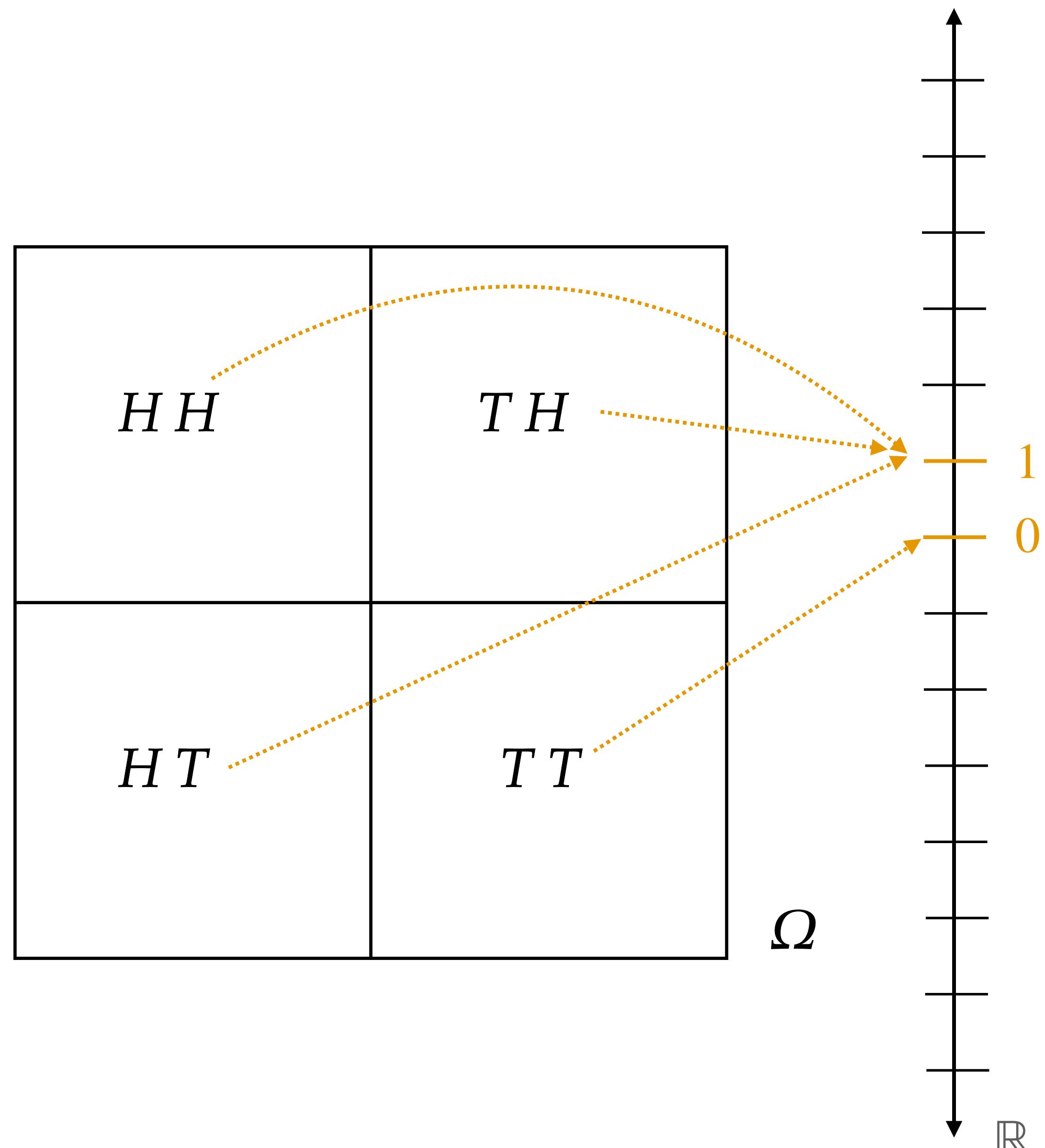
Example: Flipping 2 fair coins

Consider the following function:

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

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

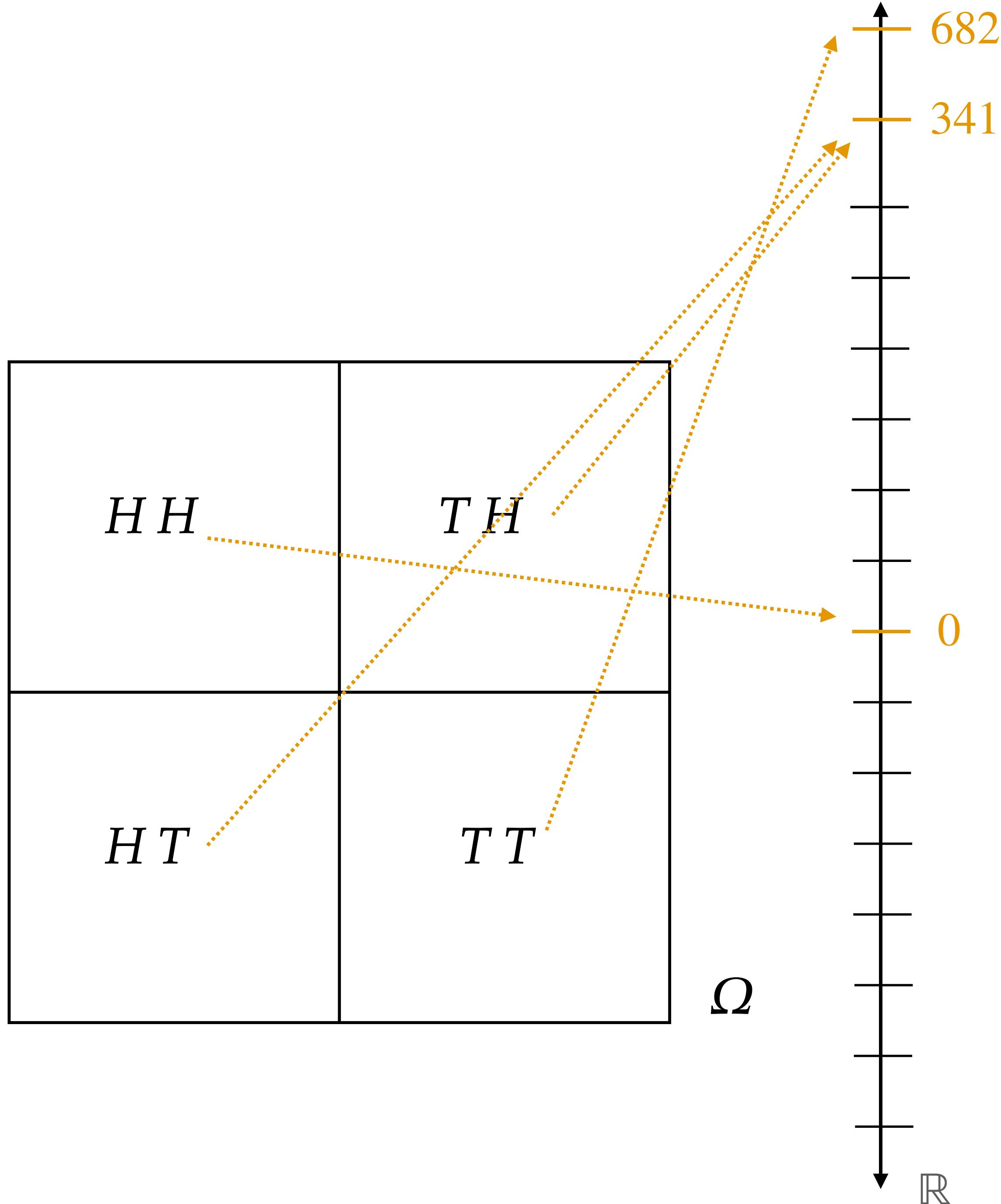
Example: Flipping 2 fair coins

Consider the following function:

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

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.

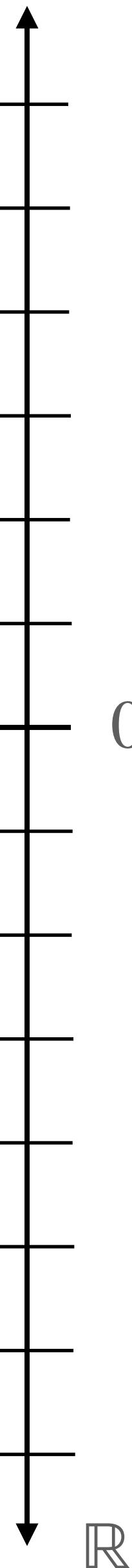
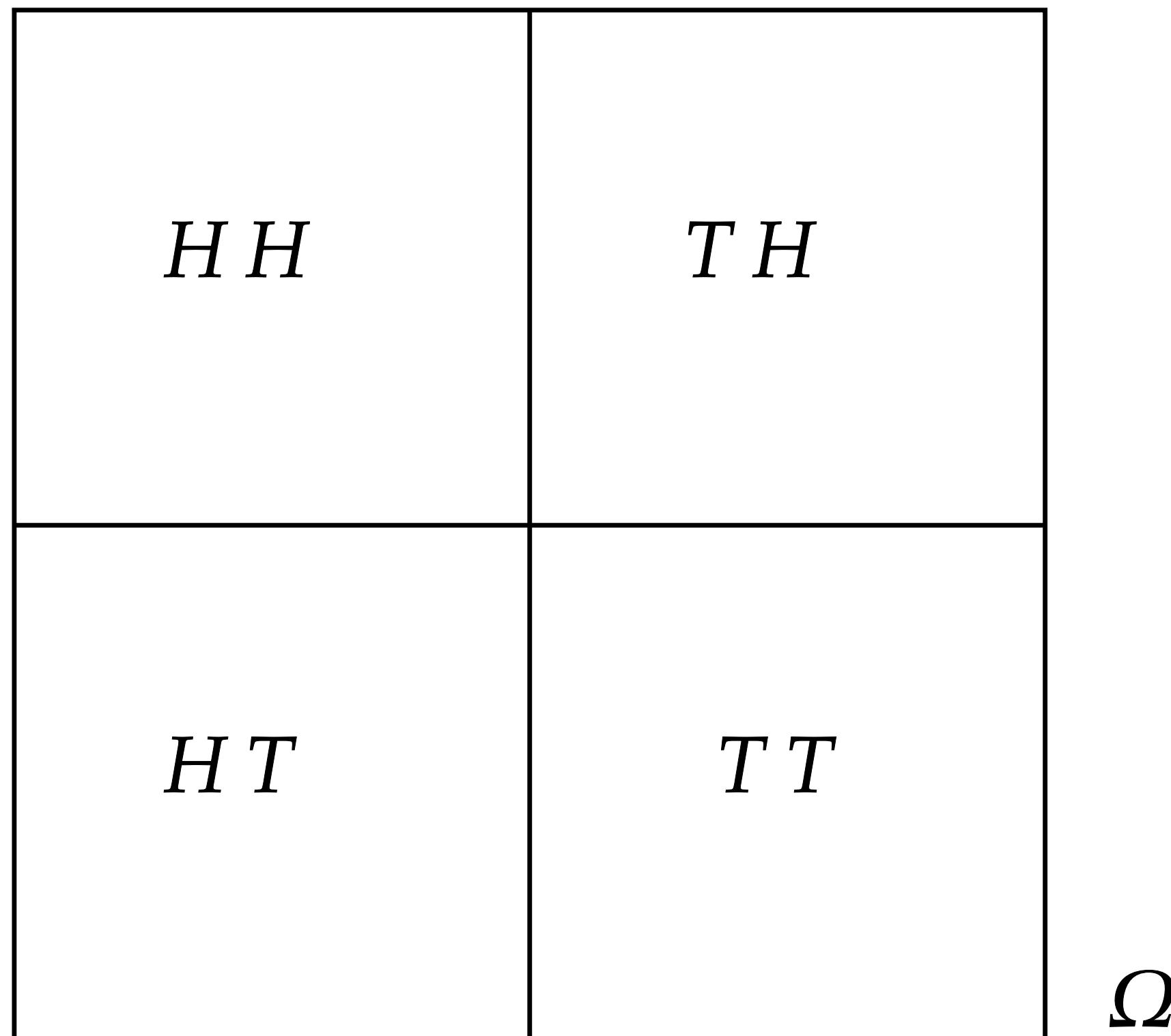


# Random Variable

## Intuition and definition

A random variable is a function  $X : \Omega \rightarrow \mathbb{R}$  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  $X : \Omega \rightarrow \mathbb{R}$  be defined as

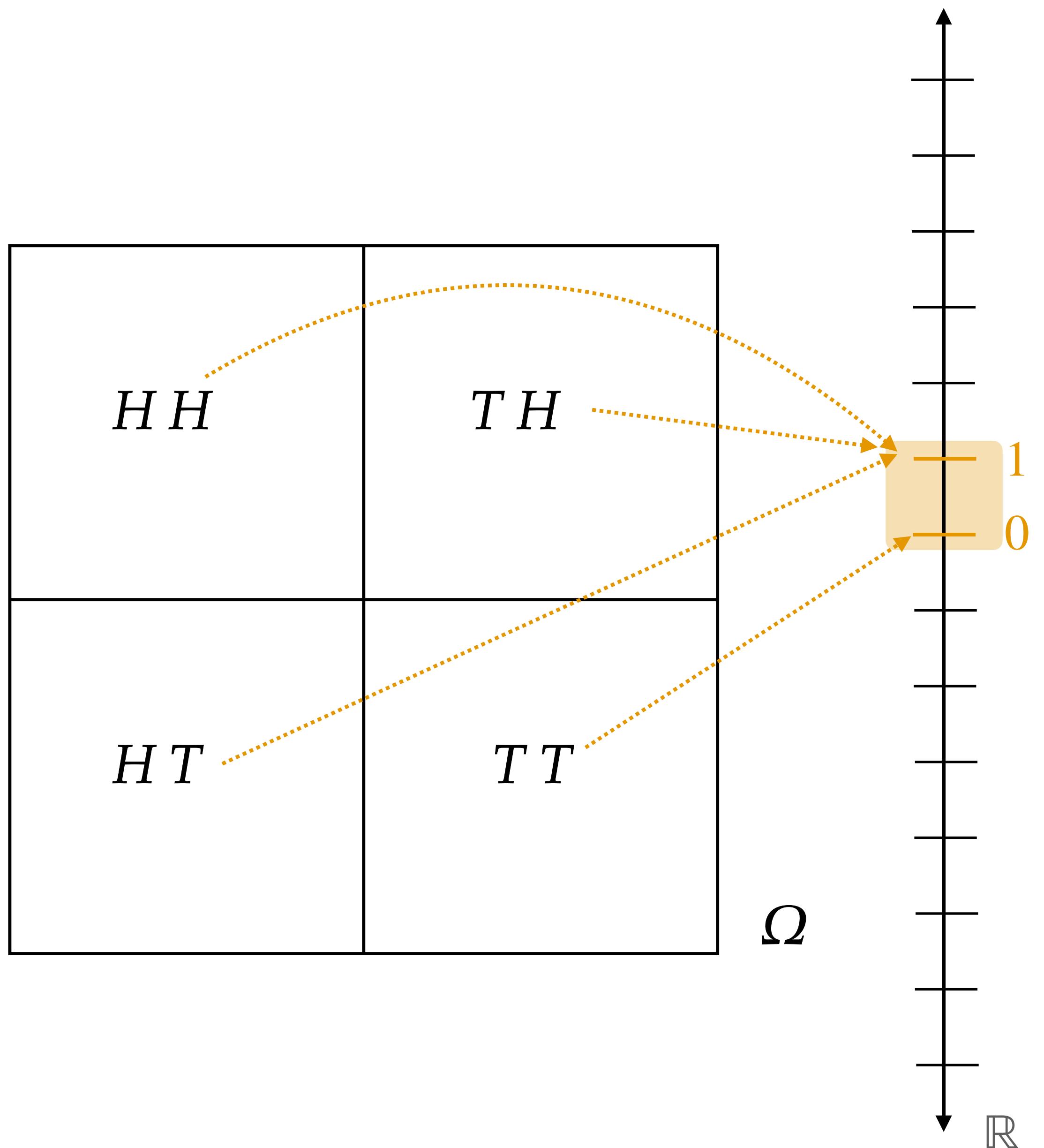
$$X(\omega) = \# \text{ of heads, } H.$$

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

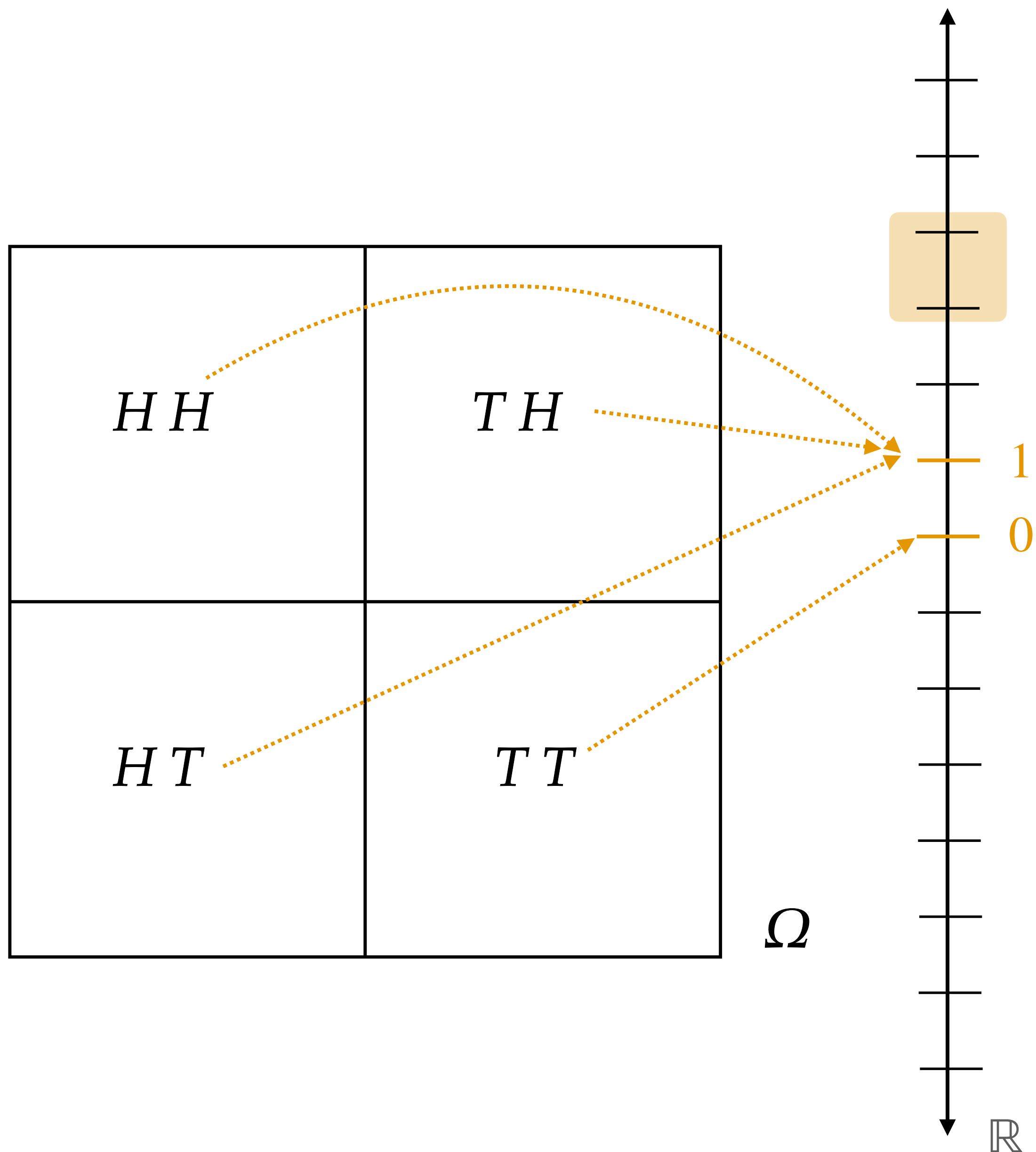
$$X(\omega) = \# \text{ of heads, } H.$$

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

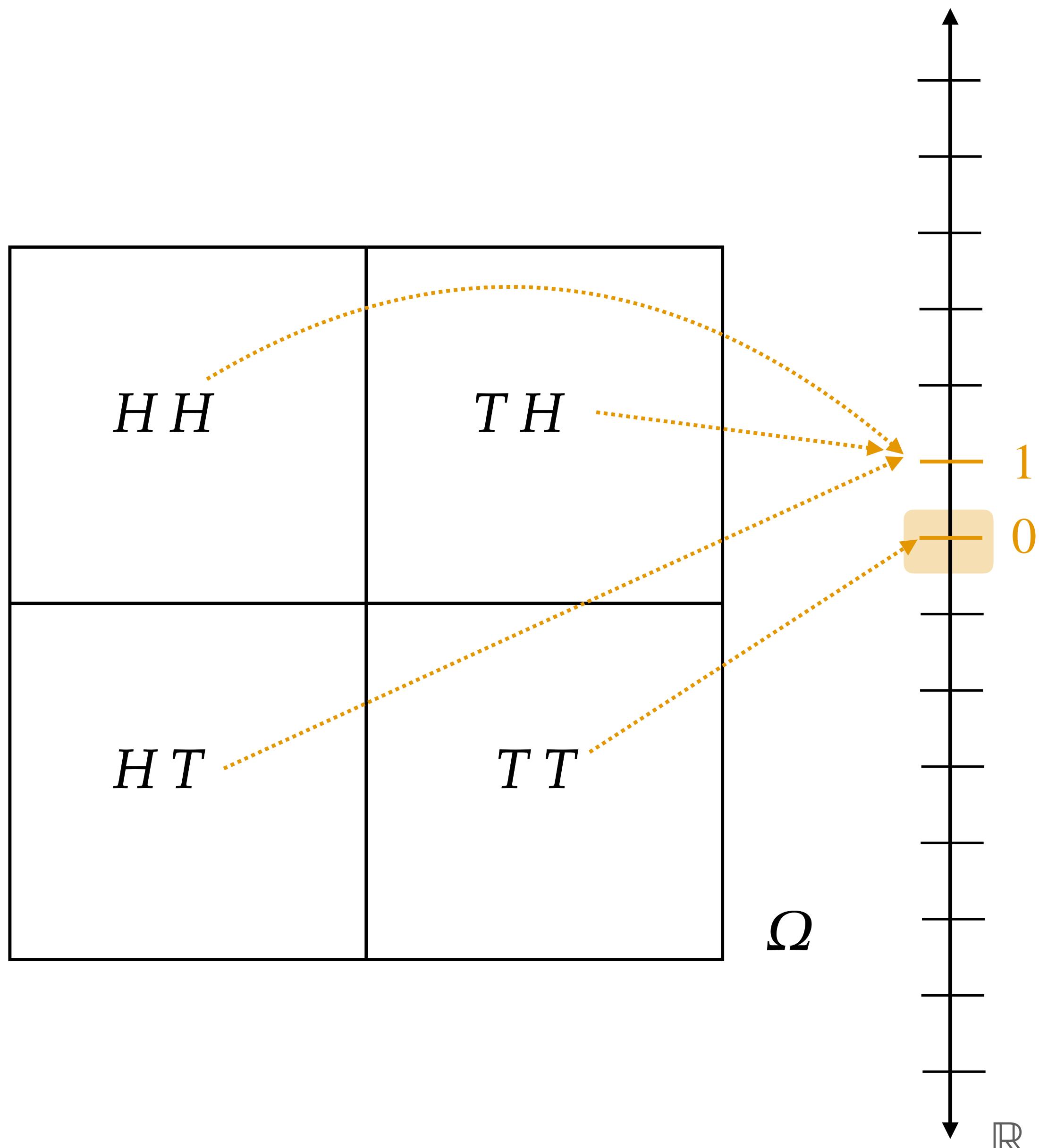
$$X(\omega) = \# \text{ of heads, } H.$$

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

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

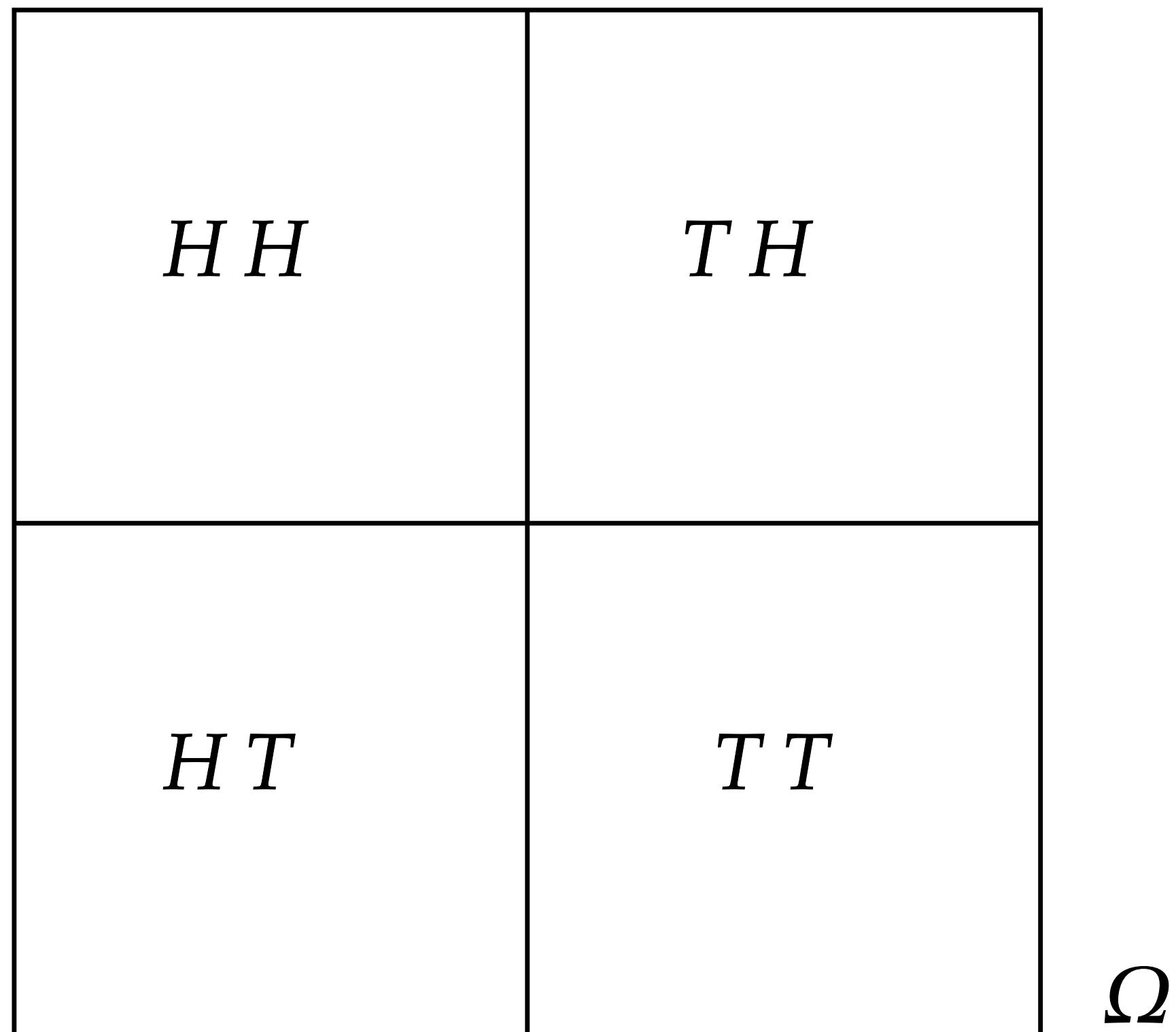
For any  $S \subseteq \mathbb{R}$ ,

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

Example.

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

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



# Random Variable

The distribution of a random variable

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

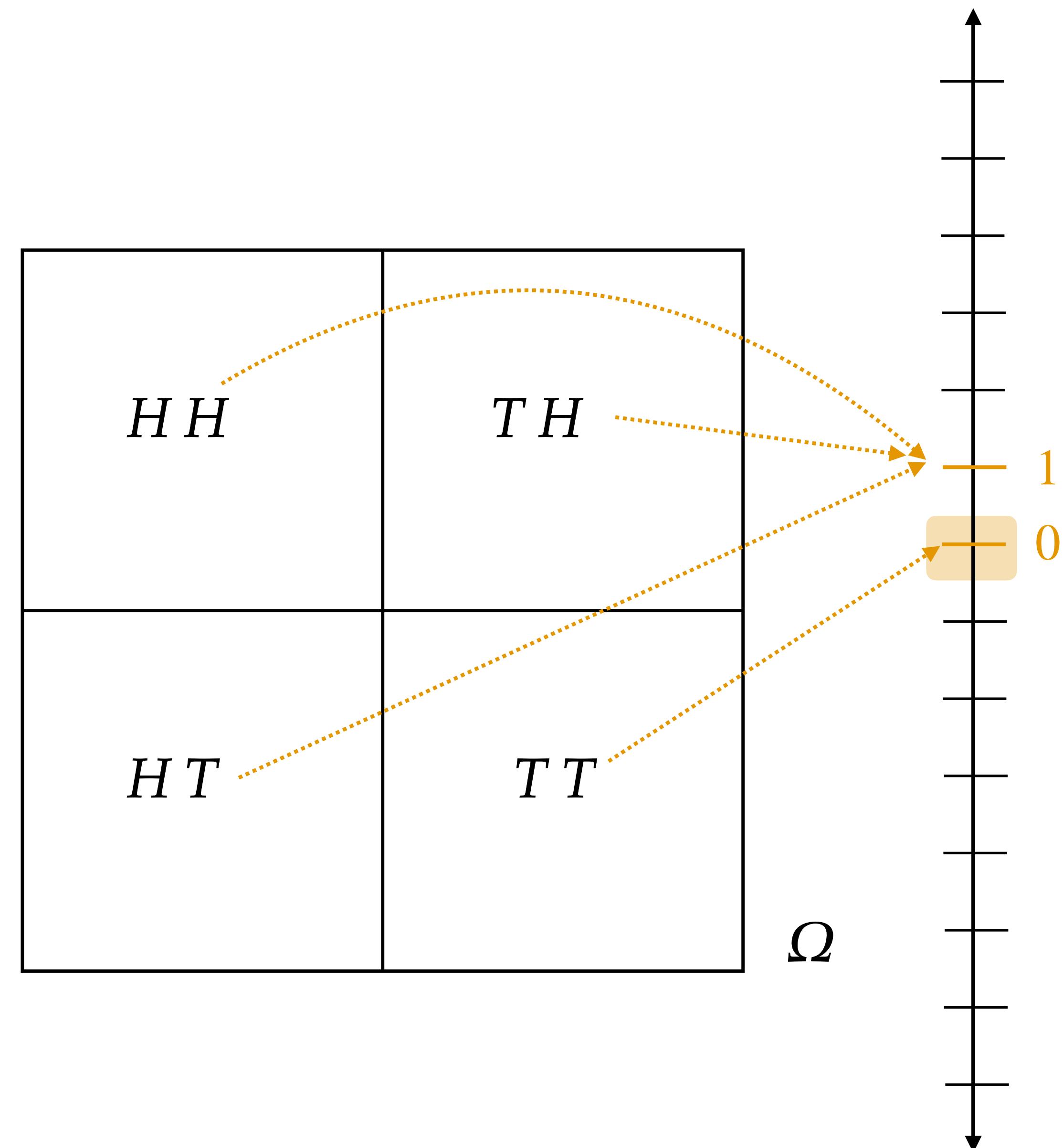
Random variables  $X : \Omega \rightarrow \mathbb{R}$  come with a distribution/  
law,  $\mathbb{P}_X$ . This implicitly defines a measure on  $\mathbb{R}$ .

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:

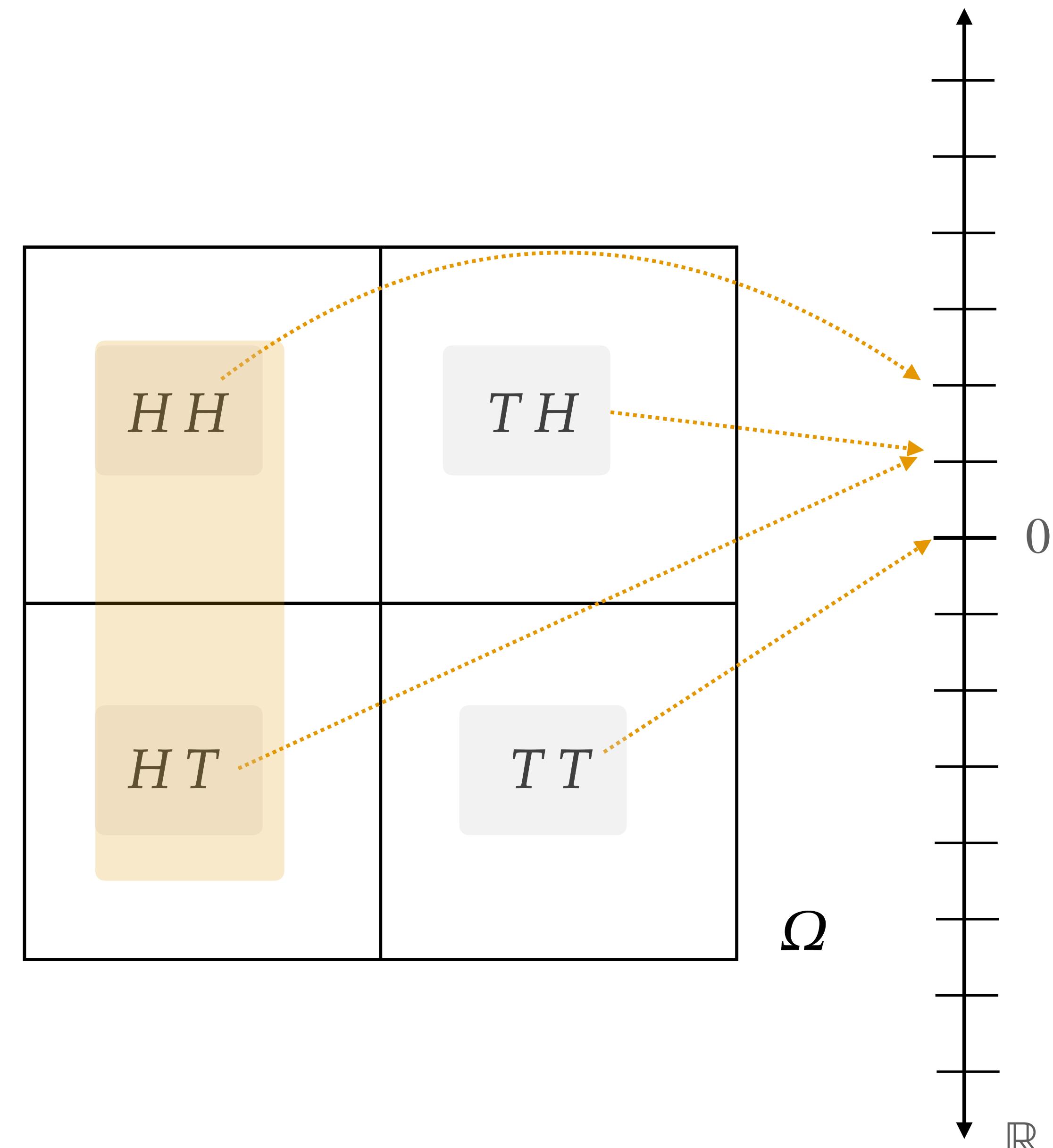
$$\mathcal{A} = \{\emptyset, \{HH\}, \{TT\}, \dots, \{HH, HT, TH, TT\}\}$$

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  $(\Omega, \mathcal{A}, \mathbb{P})$  is a function  $X : \Omega \rightarrow \mathbb{R}$  associating outcomes  $\omega \in \Omega$  to numbers in  $\mathbb{R}$ :

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



# Probability Spaces

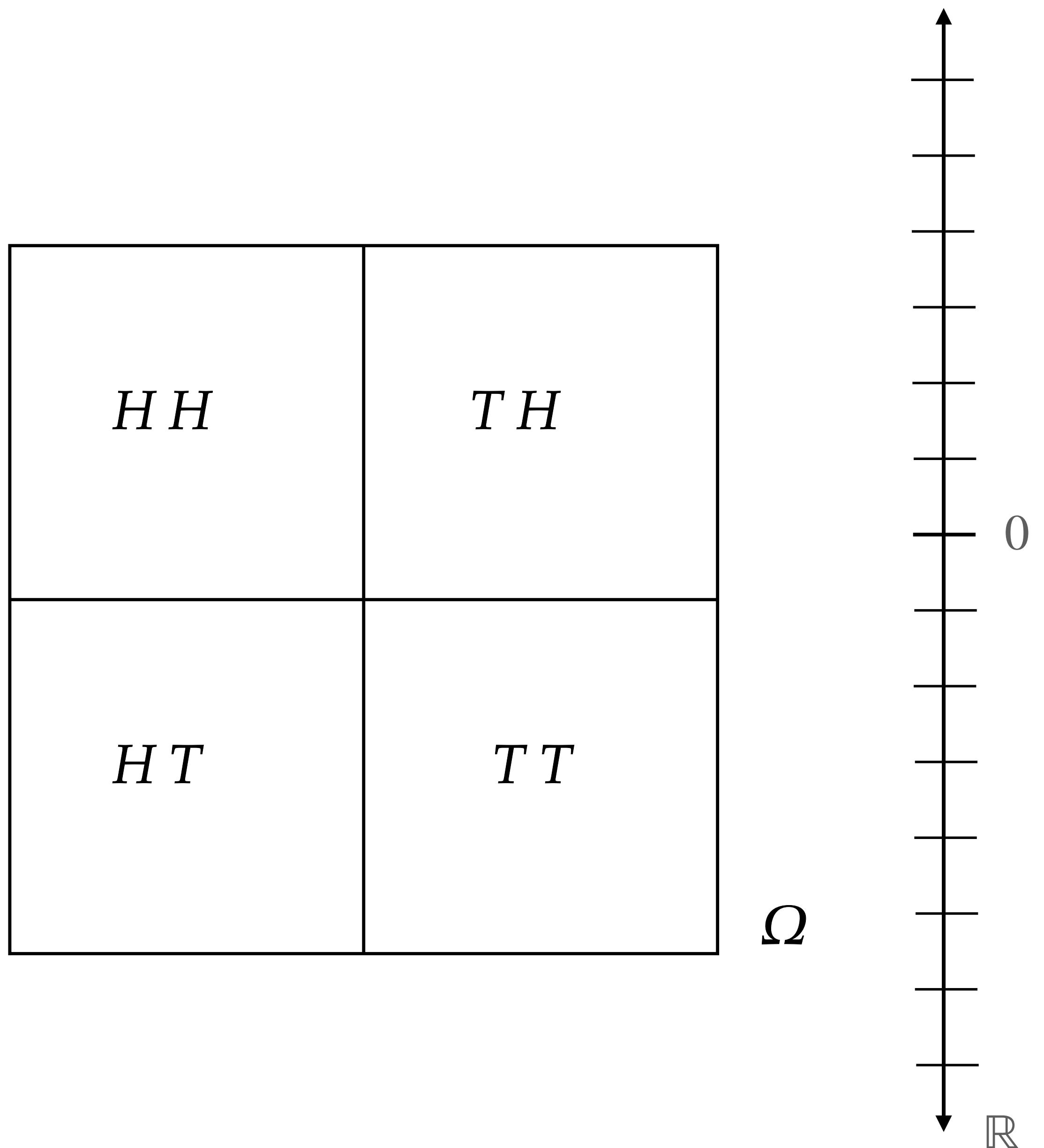
Putting everything together

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

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

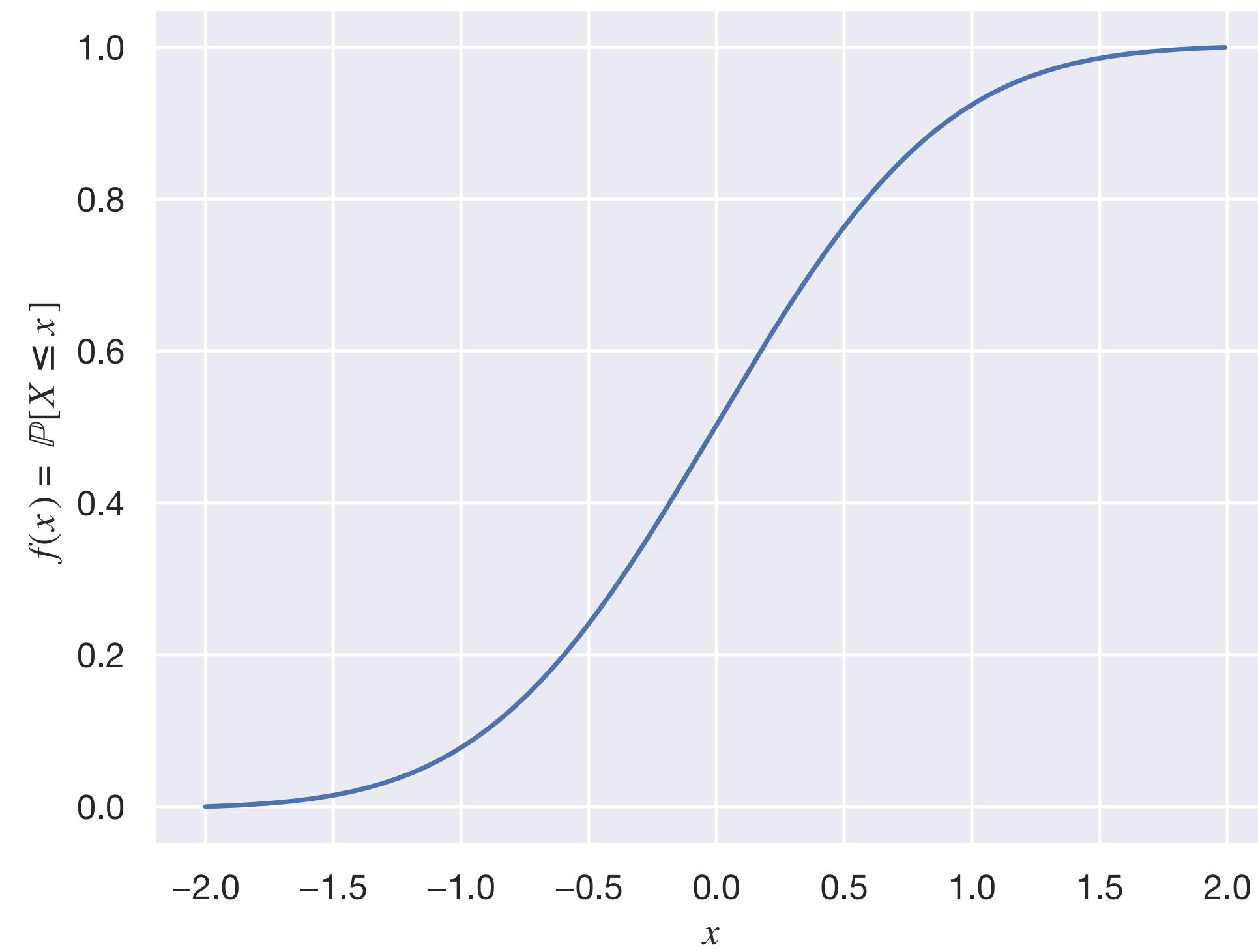
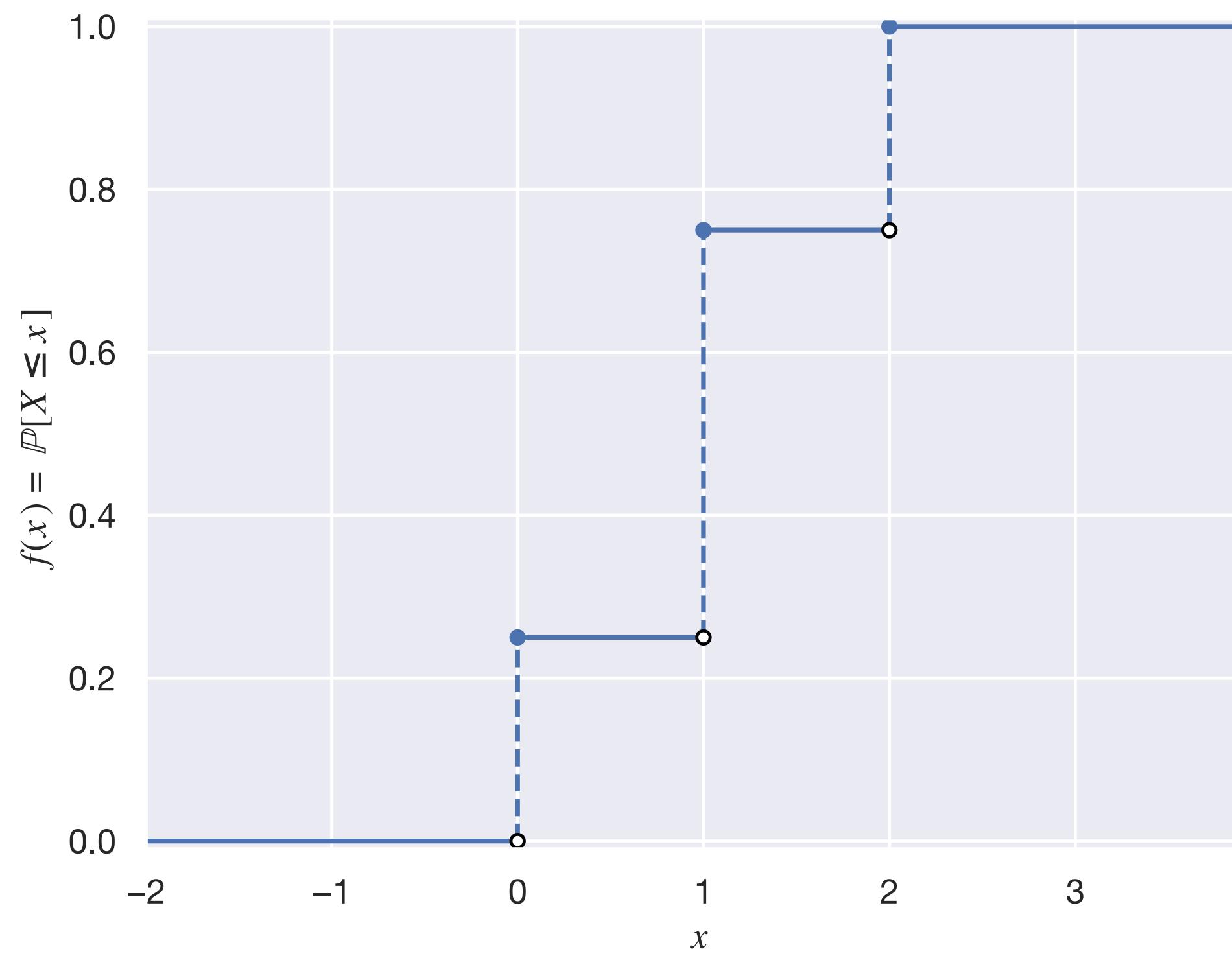
$$F_X(x) = \mathbb{P}(X \leq x)$$

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:

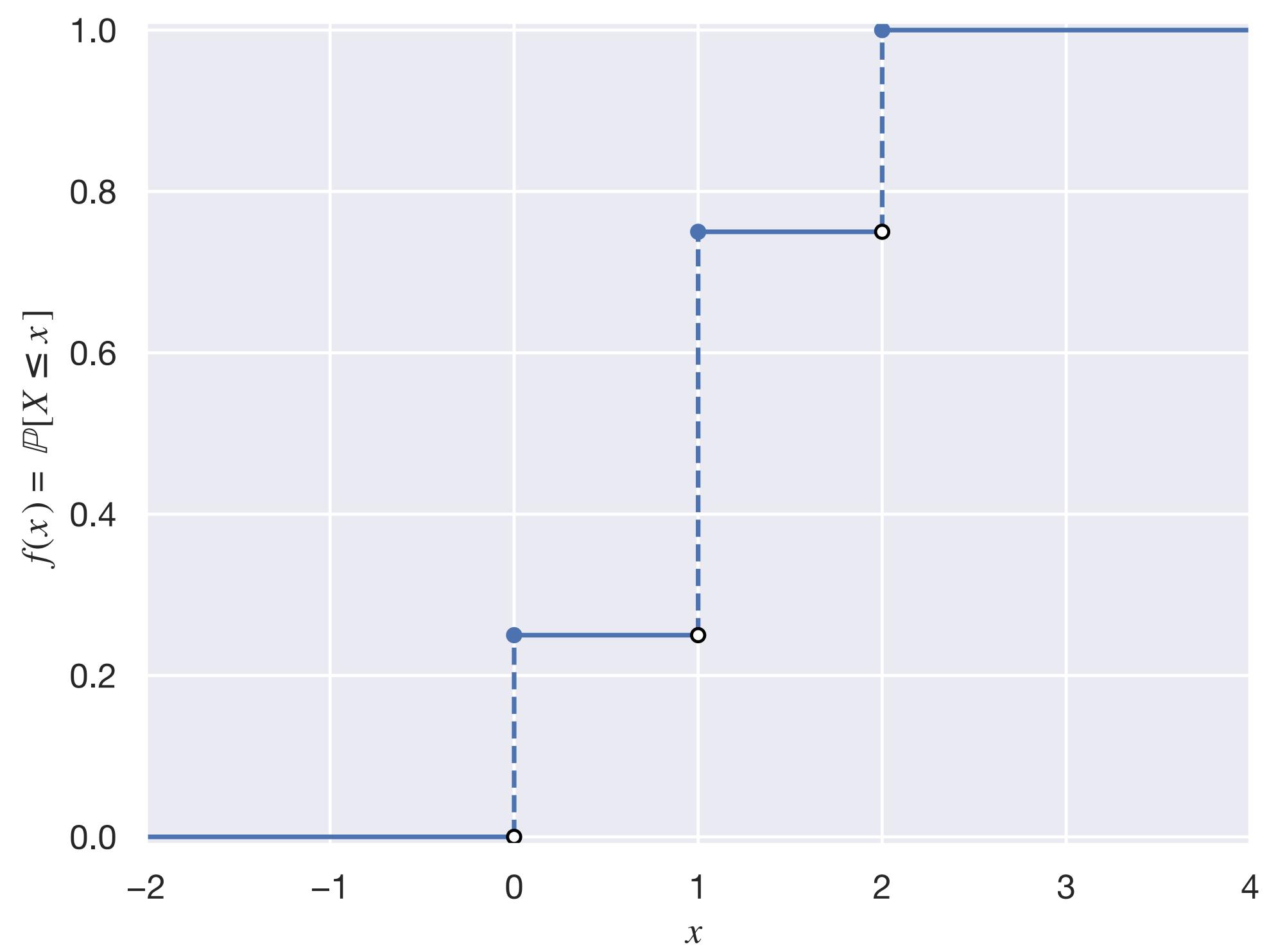
$$\lim_{x \rightarrow a^+} f(x) = f(a).$$

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

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

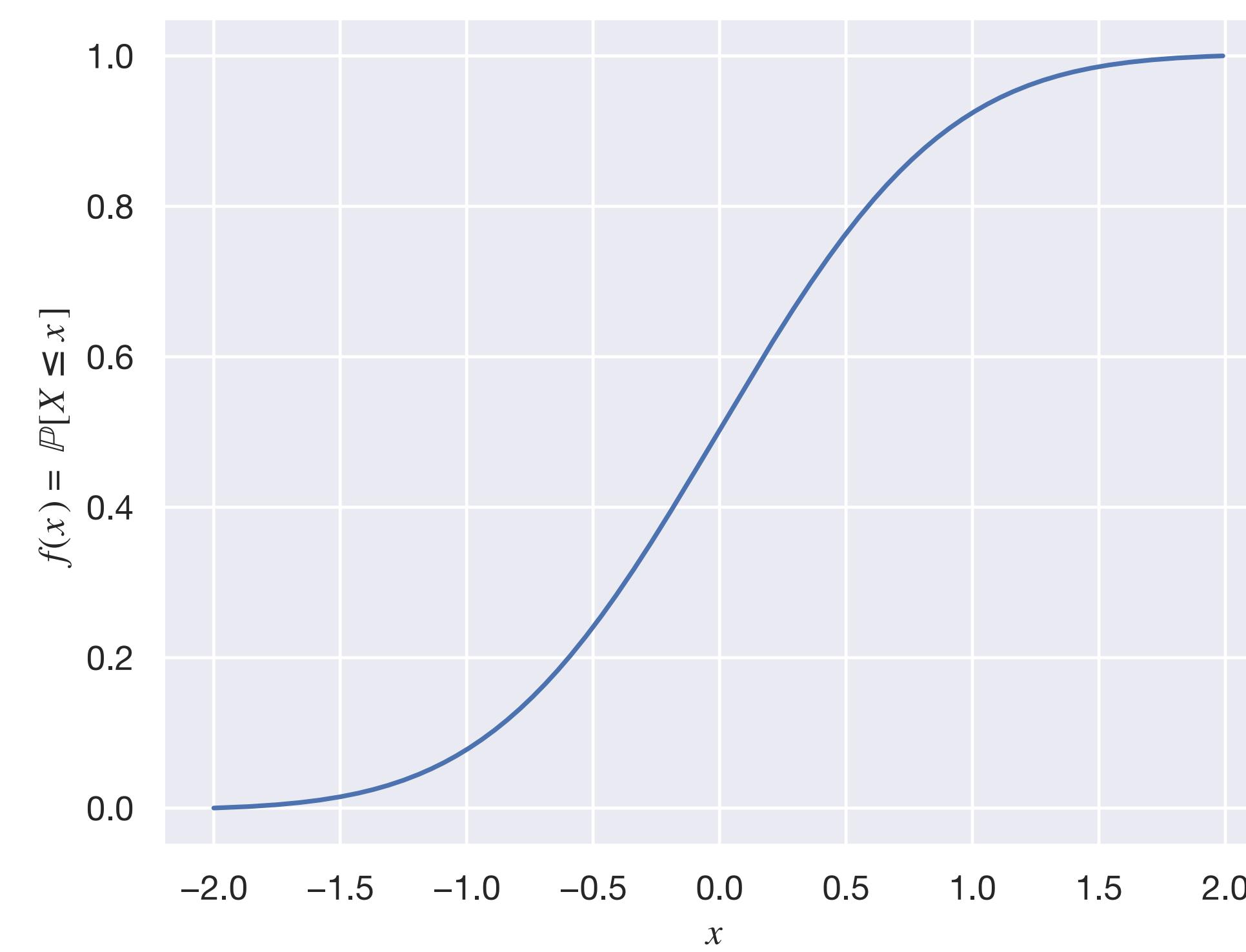
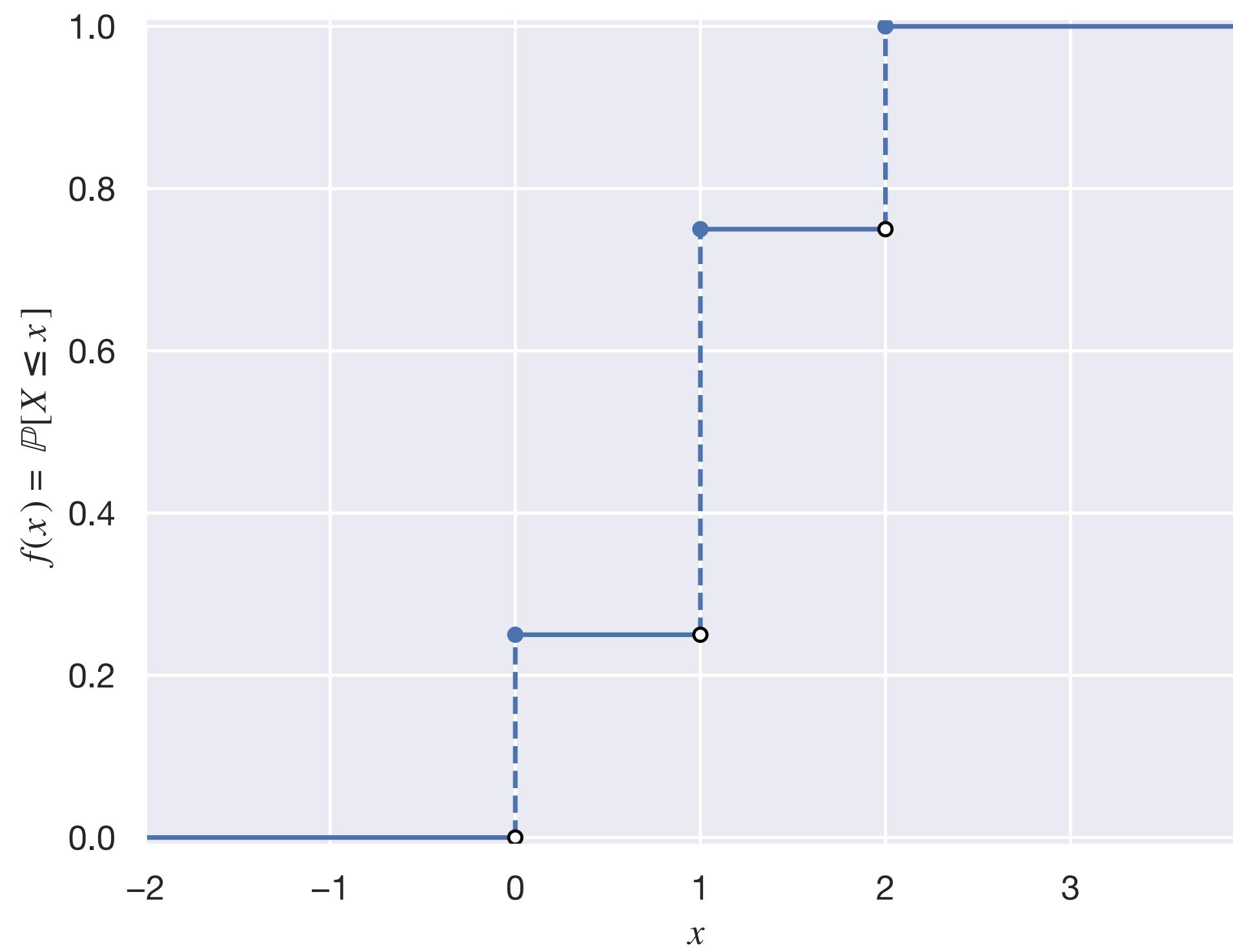
Limits at infinities. The limits at infinity are:

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



# 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.

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

$X : [0,1] \rightarrow \mathbb{R}$  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)  $p_X : \mathbb{R} \rightarrow [0,1]$  defined by:

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

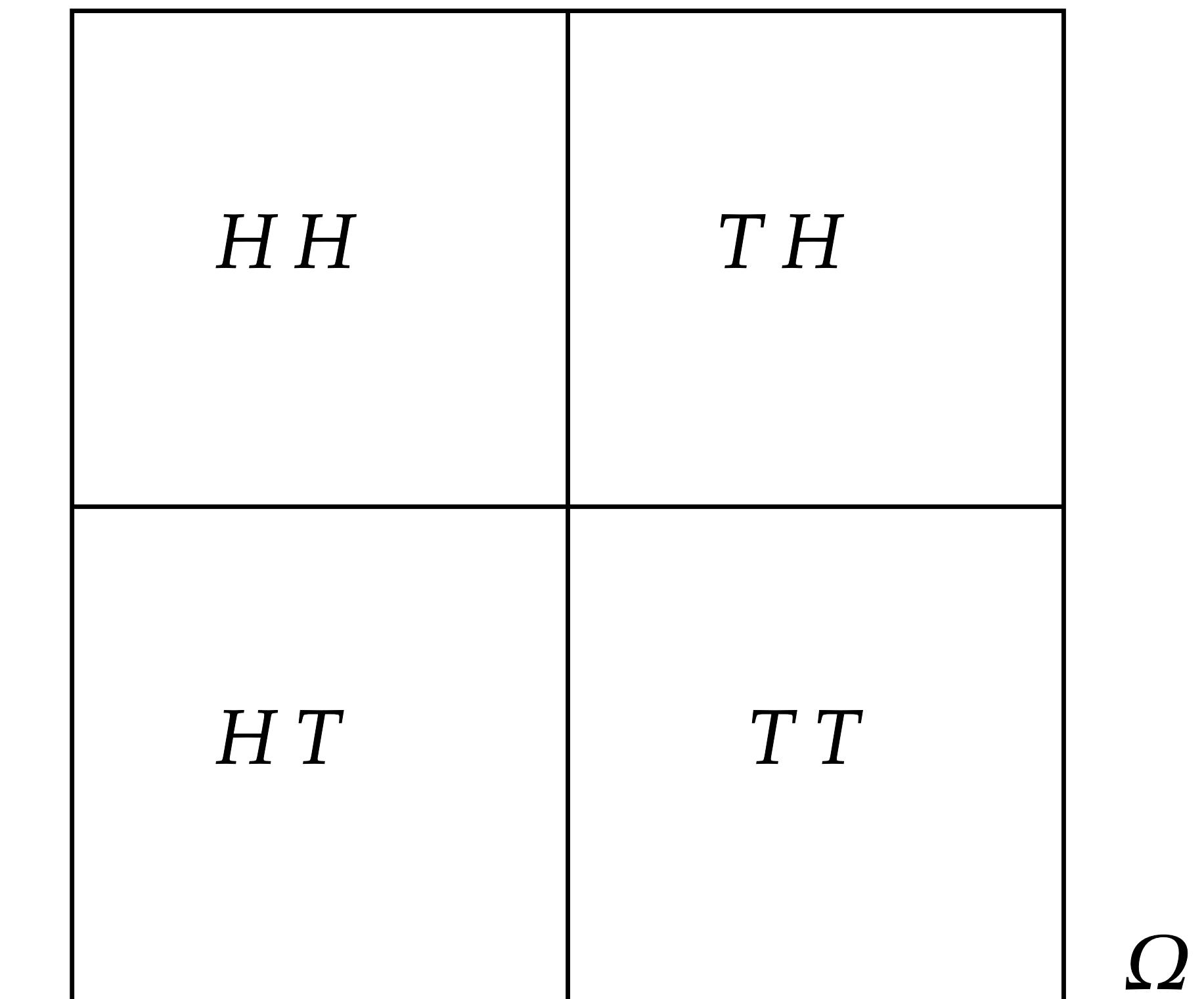
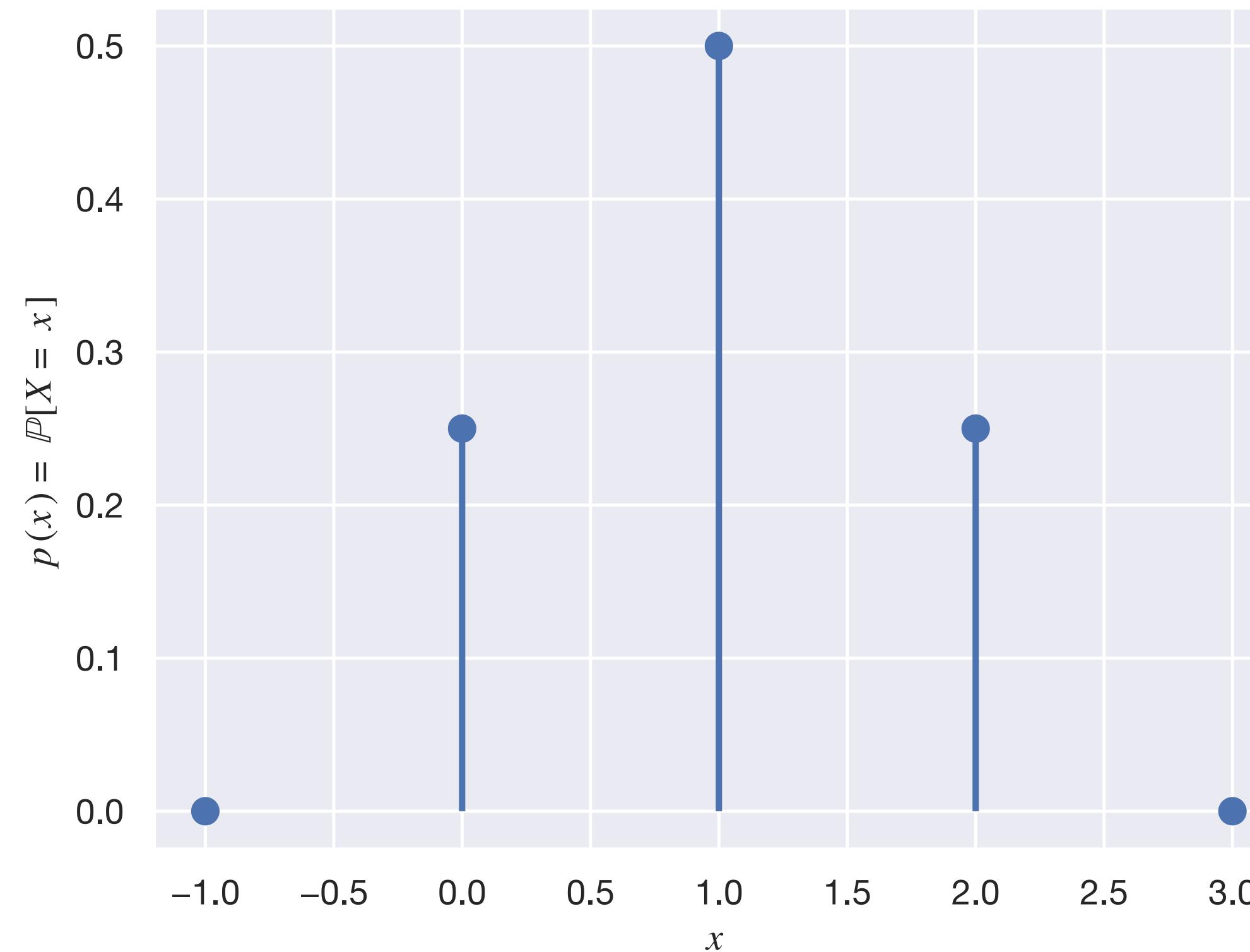
Example. What's the PMF of the RV  $X : \Omega \rightarrow \mathbb{R}$  with  $X(\omega)$  counting the number of heads?

$H H$	$T H$
$H T$	$T T$
$\Omega$	

# Discrete Random Variables

Example: Flipping 2 fair coins

Example. What's the PMF of the RV  $X : \Omega \rightarrow \mathbb{R}$  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)  $p_X : \mathbb{R} \rightarrow \mathbb{R}$

For all  $x \in \mathbb{R}$ ,  $p_X(x) \geq 0$  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*:

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

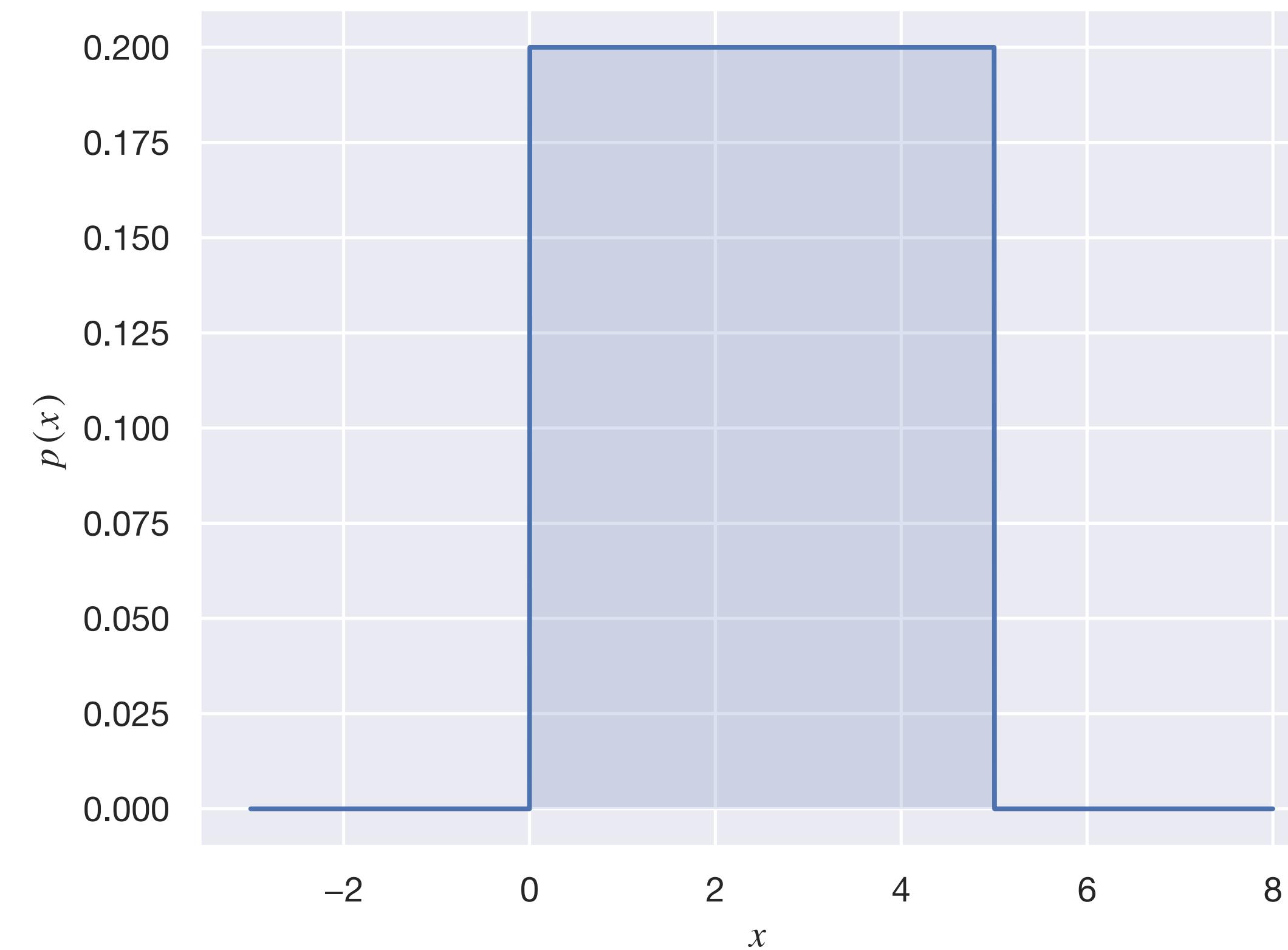
In an infinitesimally small interval, we can get a probability:

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

# Continuous Random Variables

Example: Picking uniformly in the interval

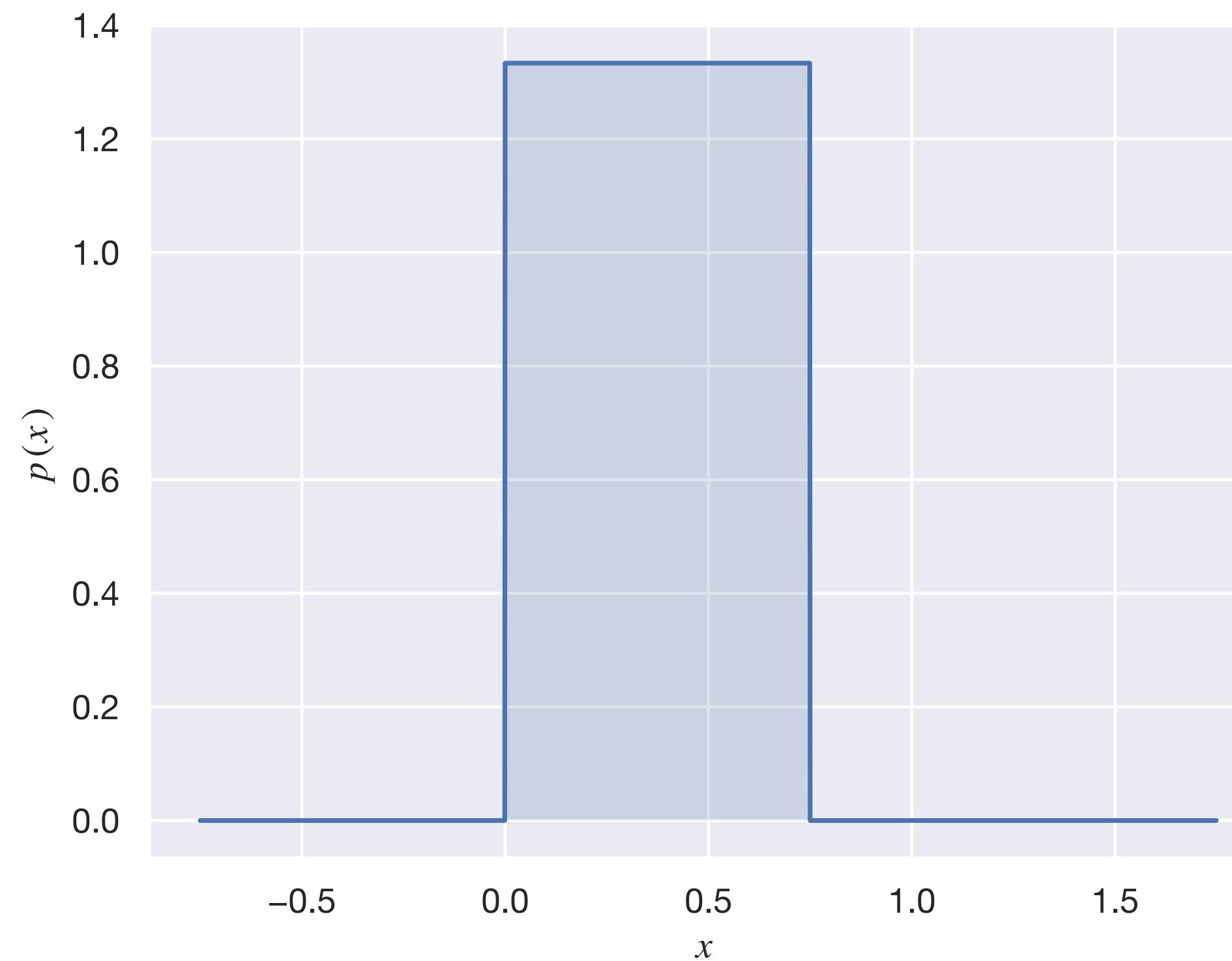
Example. What's the PDF of the RV  $X : \Omega \rightarrow \mathbb{R}$  with the uniform random variable on  $[0,5]$ ?



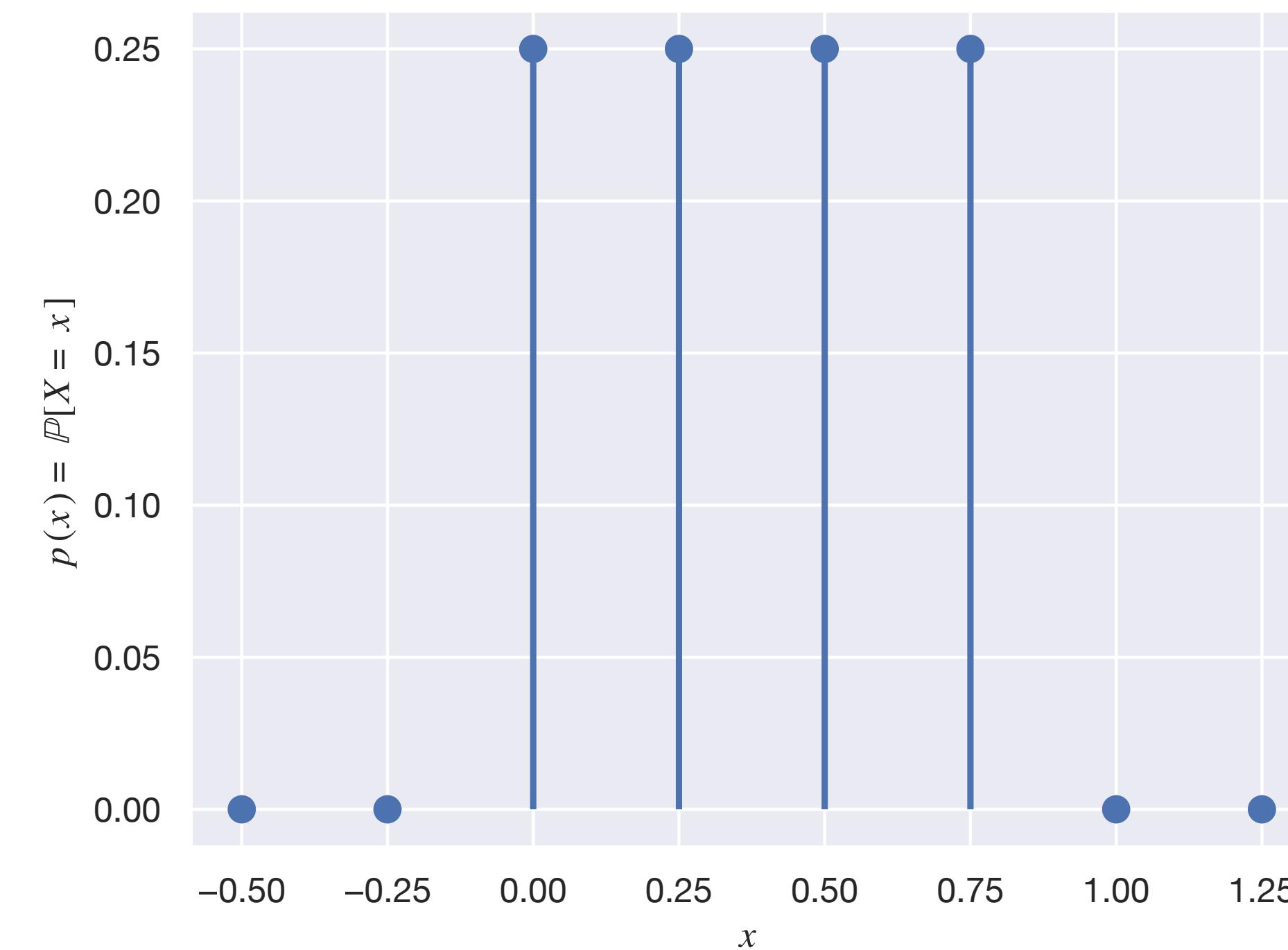
# Continuous vs. Discrete RVs

Example: Uniform Discrete and Uniform Continuous PDFs

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



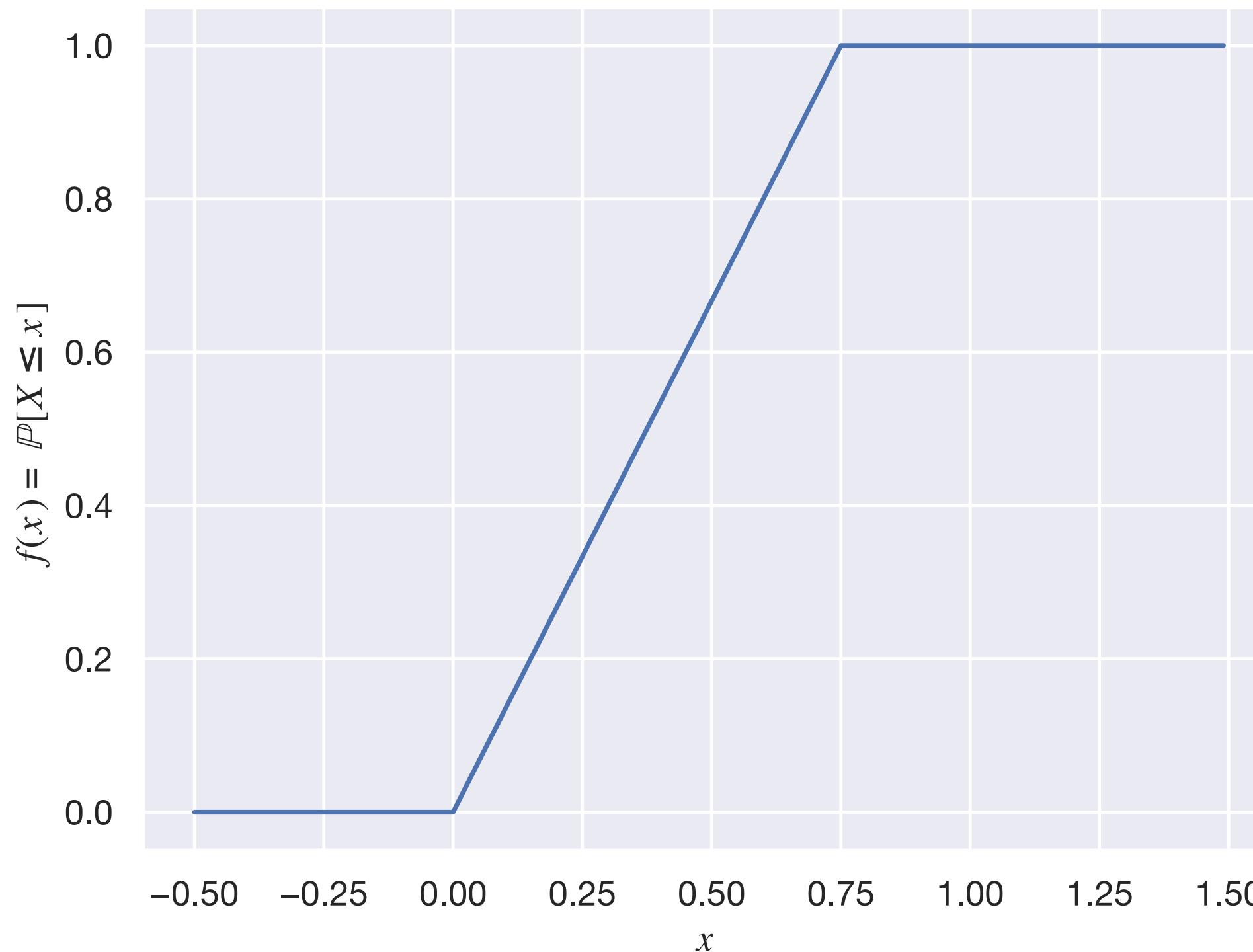
Discrete RV uniform on  $\{0,0.25,0.5,0.75\}$ .



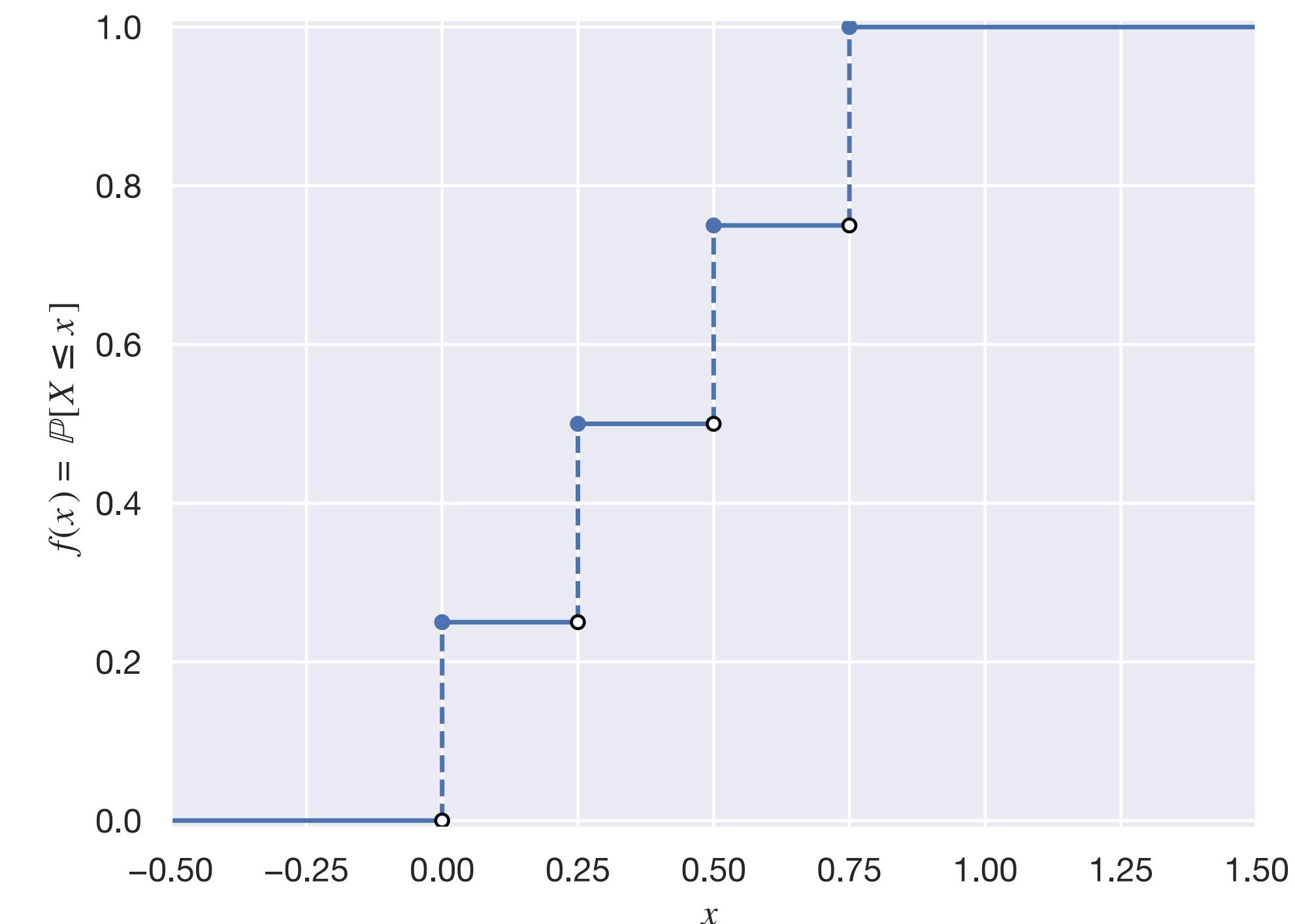
# Continuous vs. Discrete RVs

Example: Uniform Discrete and Uniform Continuous CDFs

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



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



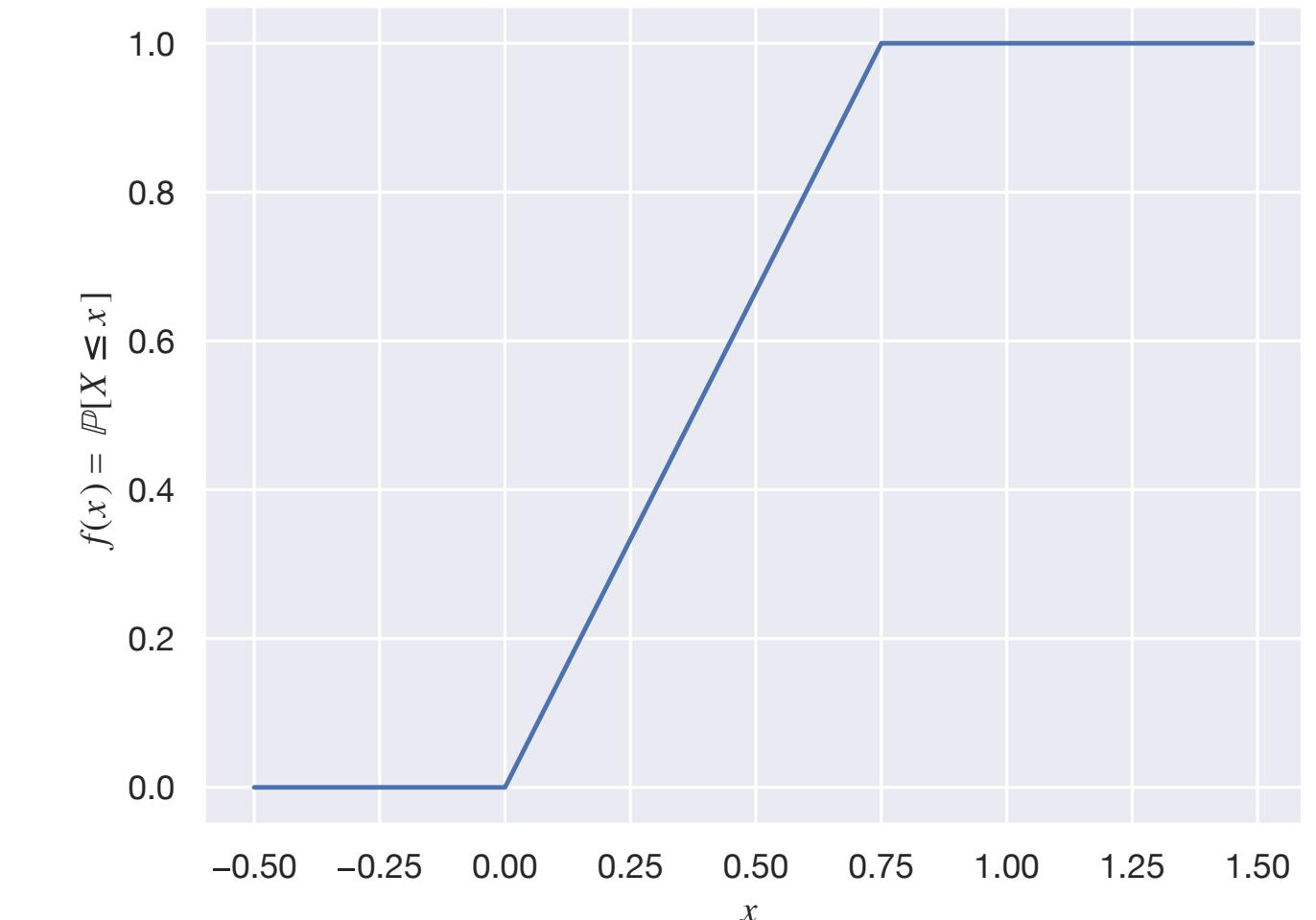
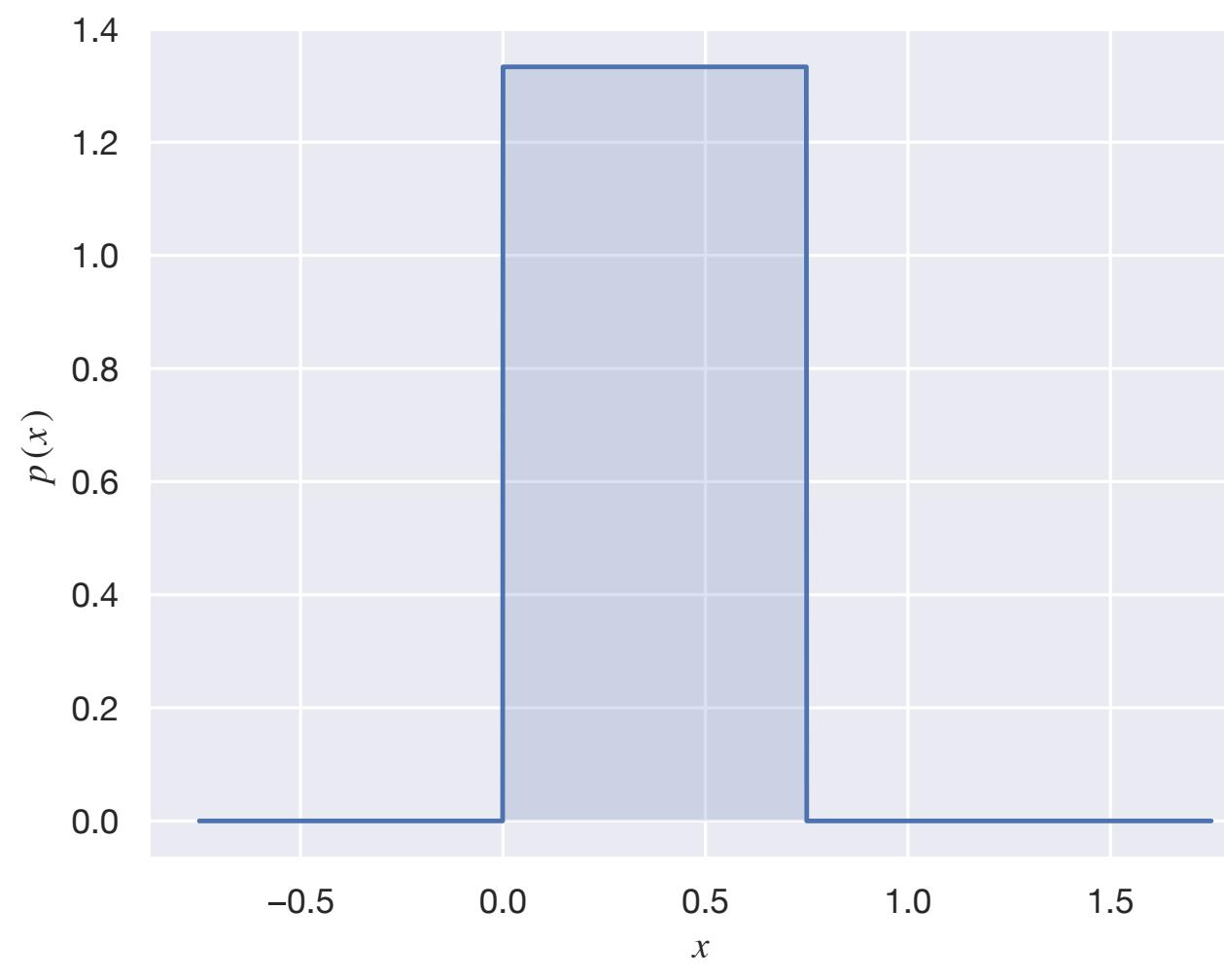
# Continuous vs. Discrete RVs

## Summary

For continuous RVs,

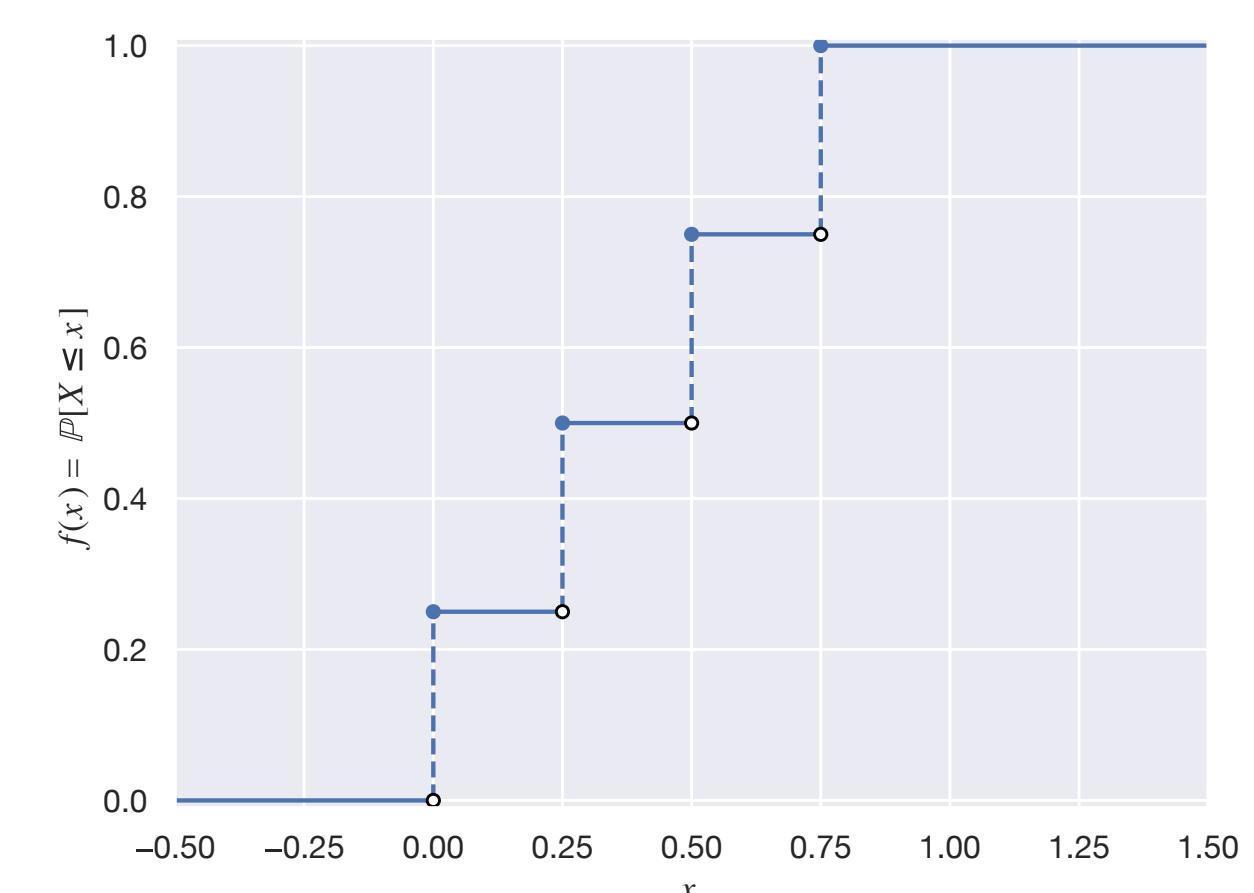
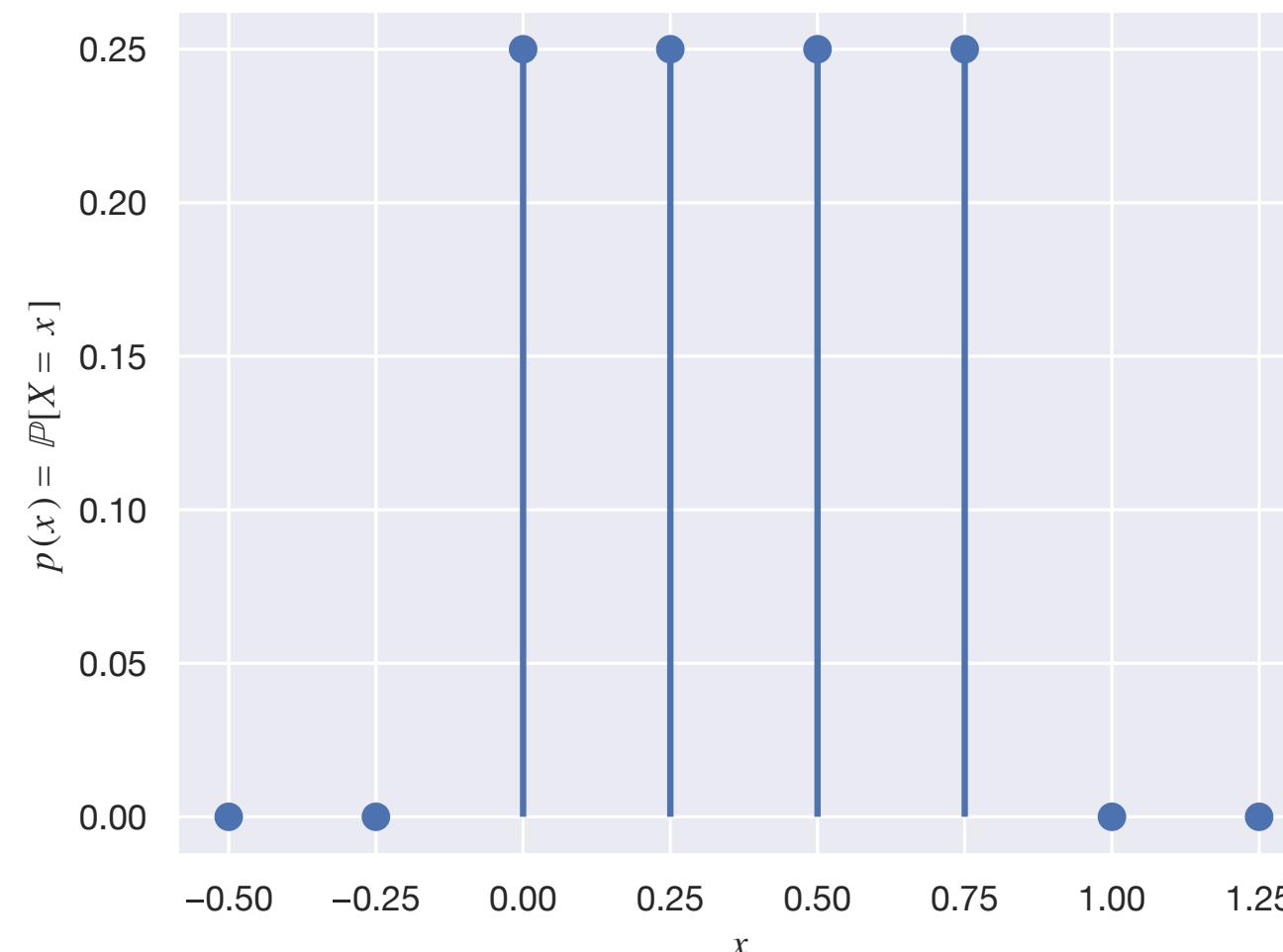
$$\mathbb{P}(X = x) = 0$$

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



For discrete RVs,

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



# Random Variables

Multiple random variables

# Joint Distribution

# Example: Tossing coins and rolling die

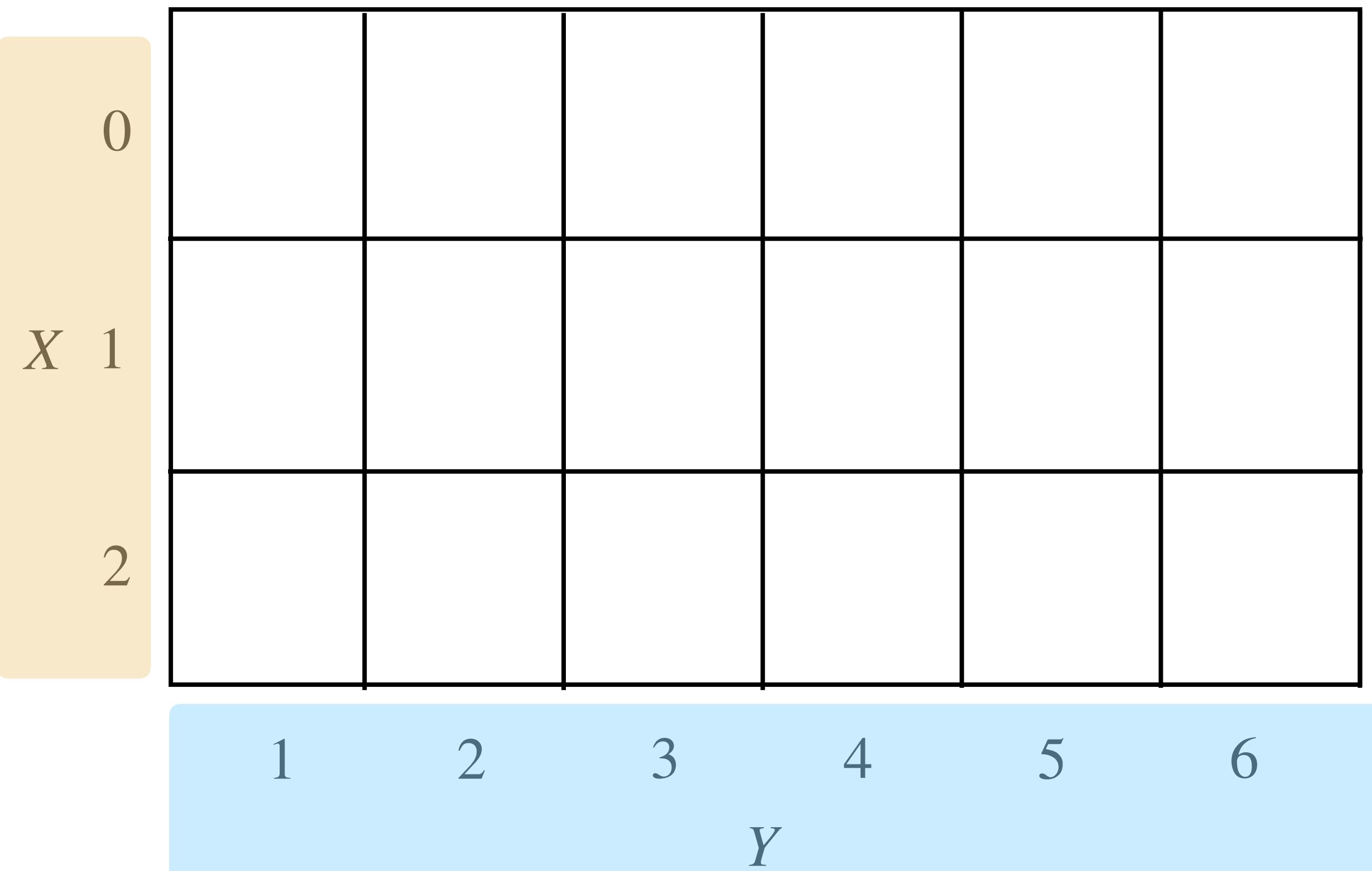
# *Consider two experiments:*

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.



# Joint Distribution

## Definition

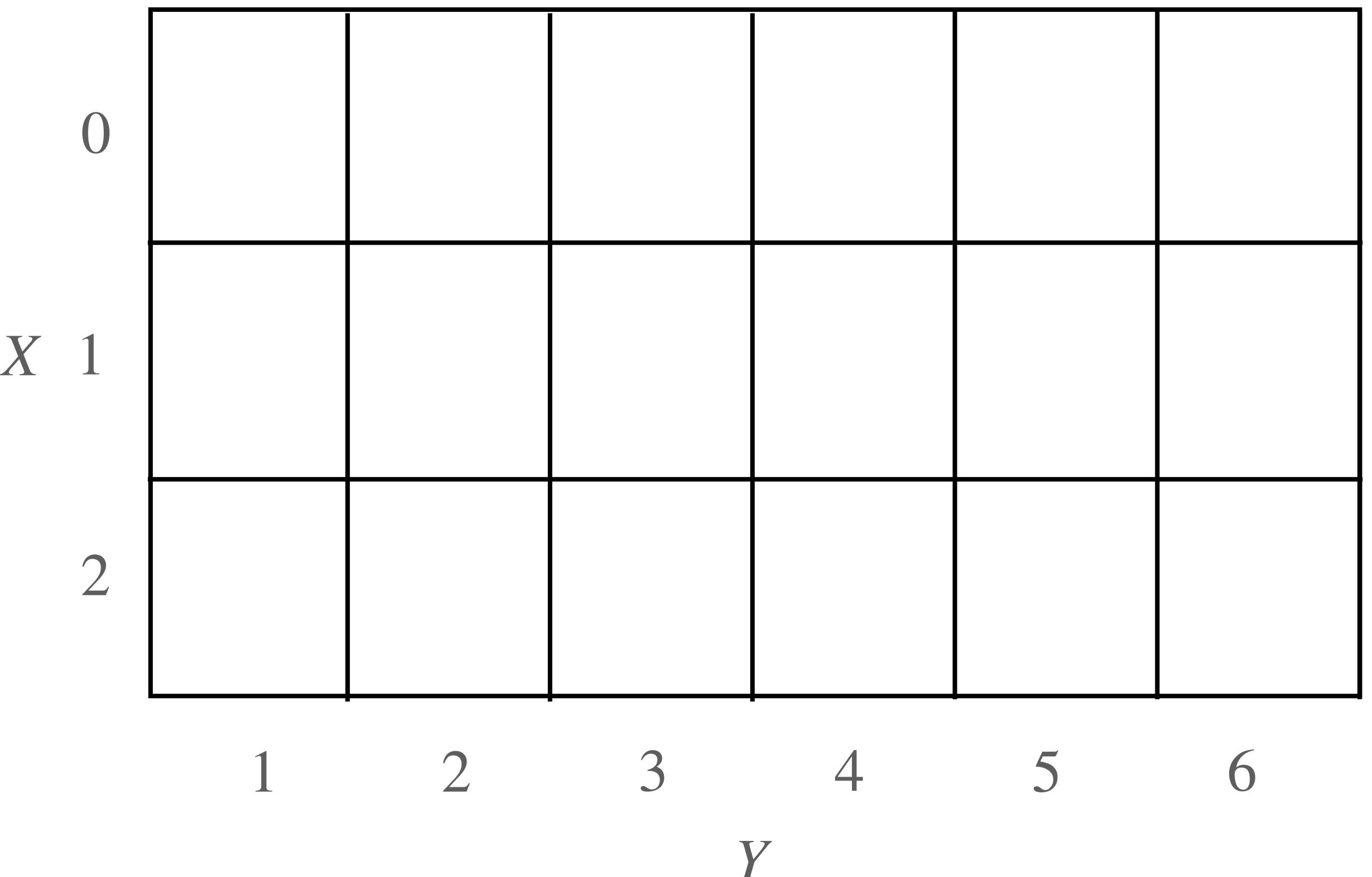
Let  $X_1, \dots, X_n$  be random variables.

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

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

For discrete random variables,

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



# Joint Distribution

## Definition

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

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

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

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

For discrete random variables,

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

	0	1/24	1/24	1/24	1/24	1/24	1/24
X	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

# Joint Distribution

## Definition

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

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

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

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

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

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

	0	1/24	1/24	1/24	1/24	1/24	1/24
X	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

# Marginal Distribution

# Definition

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

The marginal distribution of  $X$  is obtained by “summing out”/“integrating out” the variable we don’t care about:

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

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

# Marginal Distribution

## Definition

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

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
0		1/24	1/24	1/24	1/24	1/24	1/24
	X	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

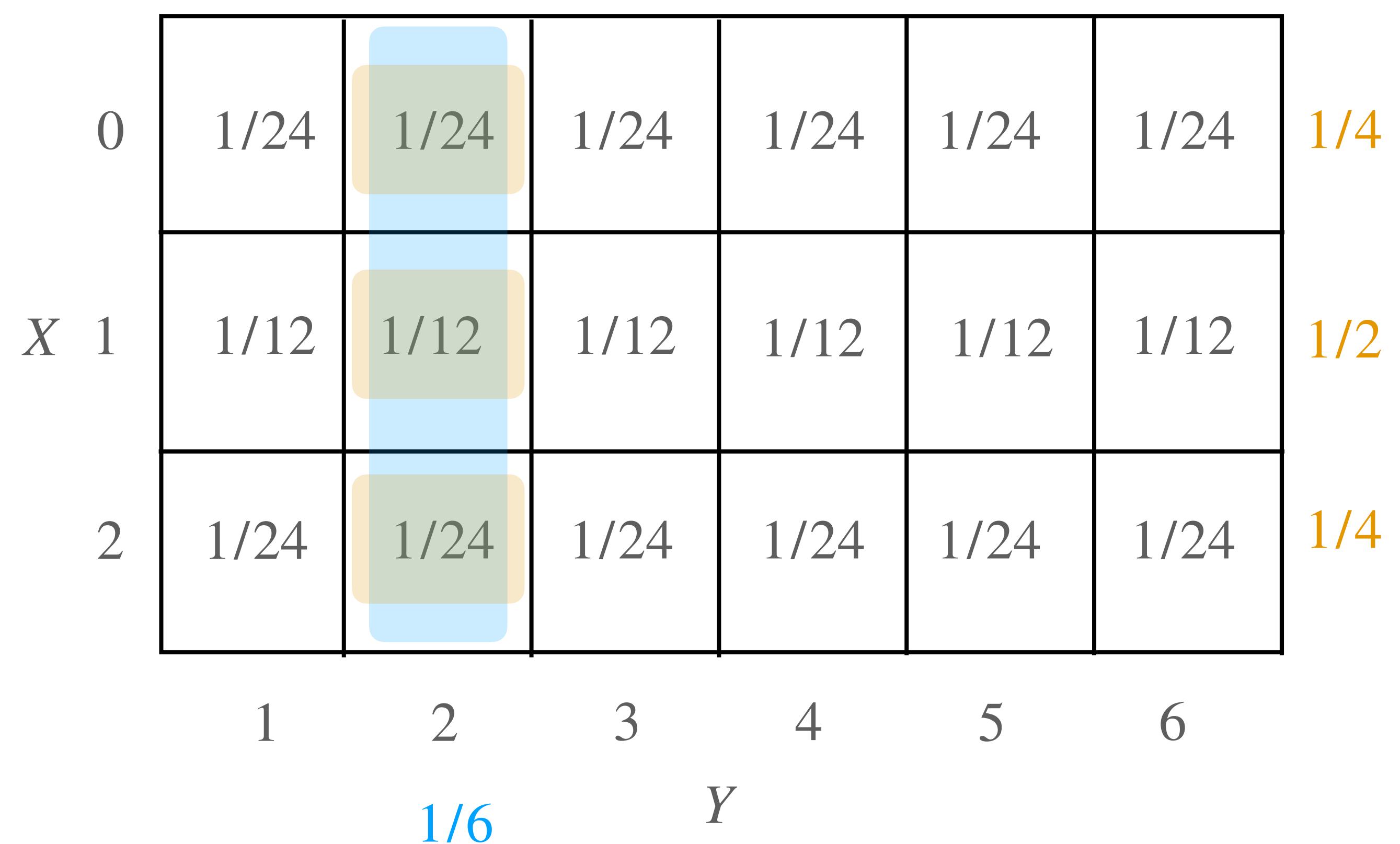
# Conditional Distribution

## Definition

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

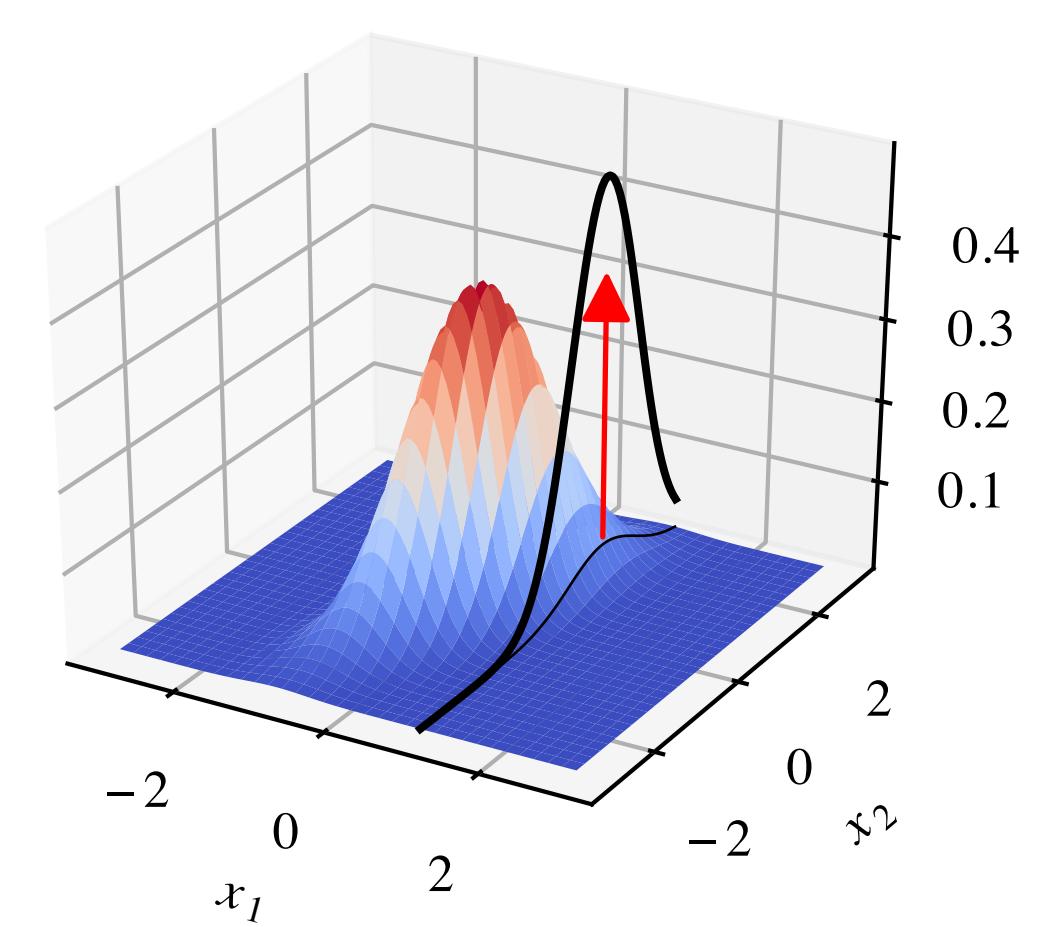
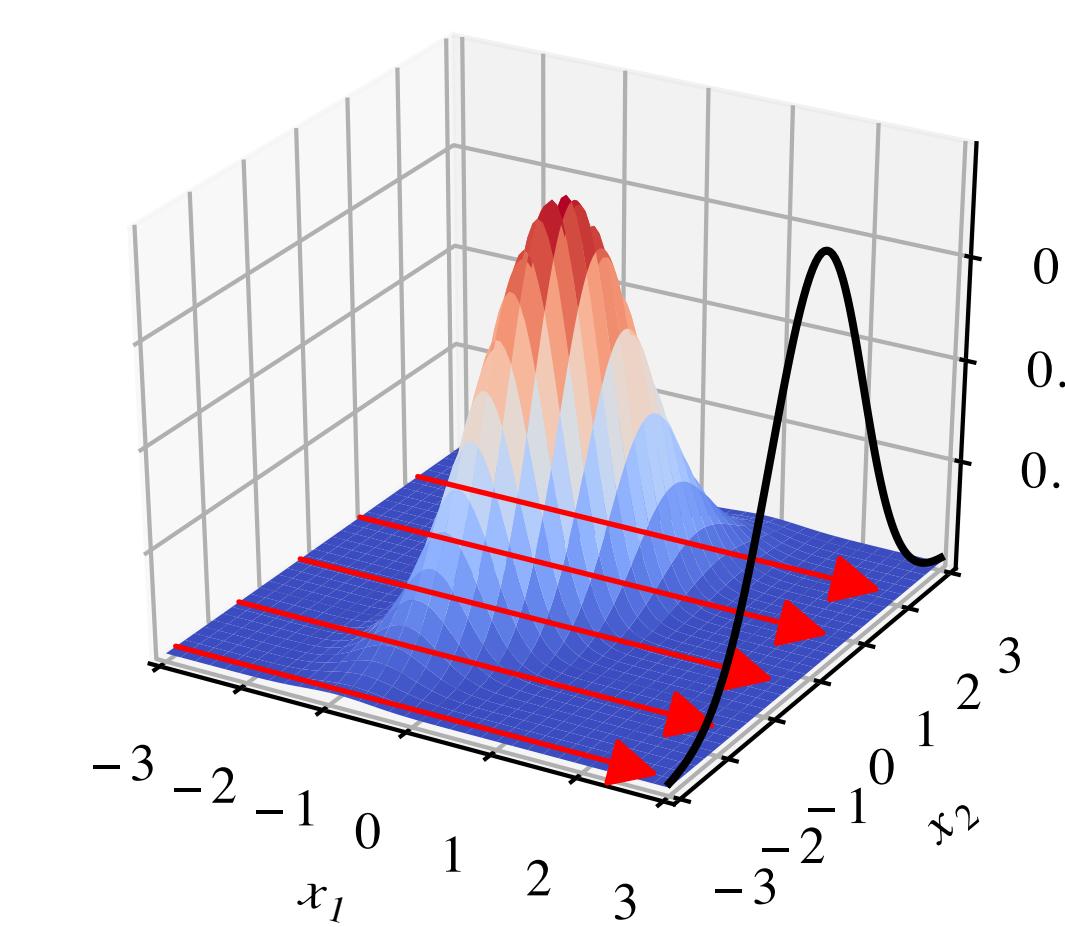
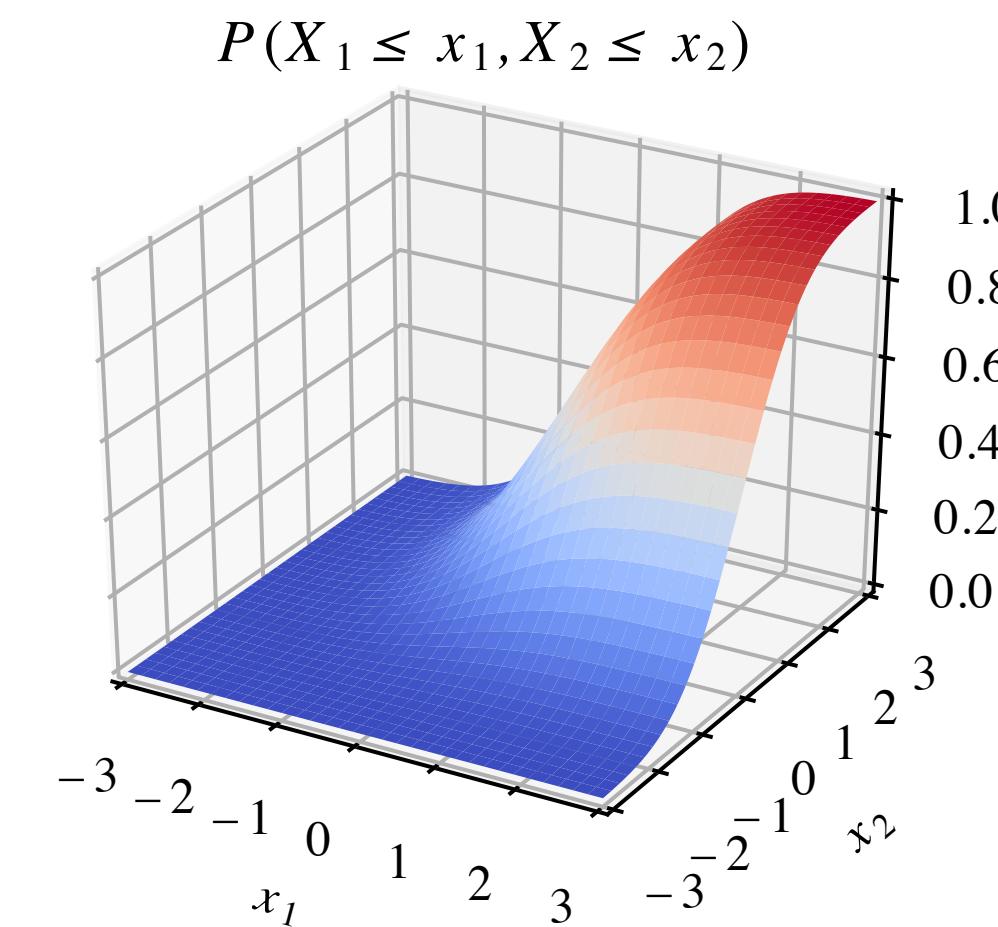
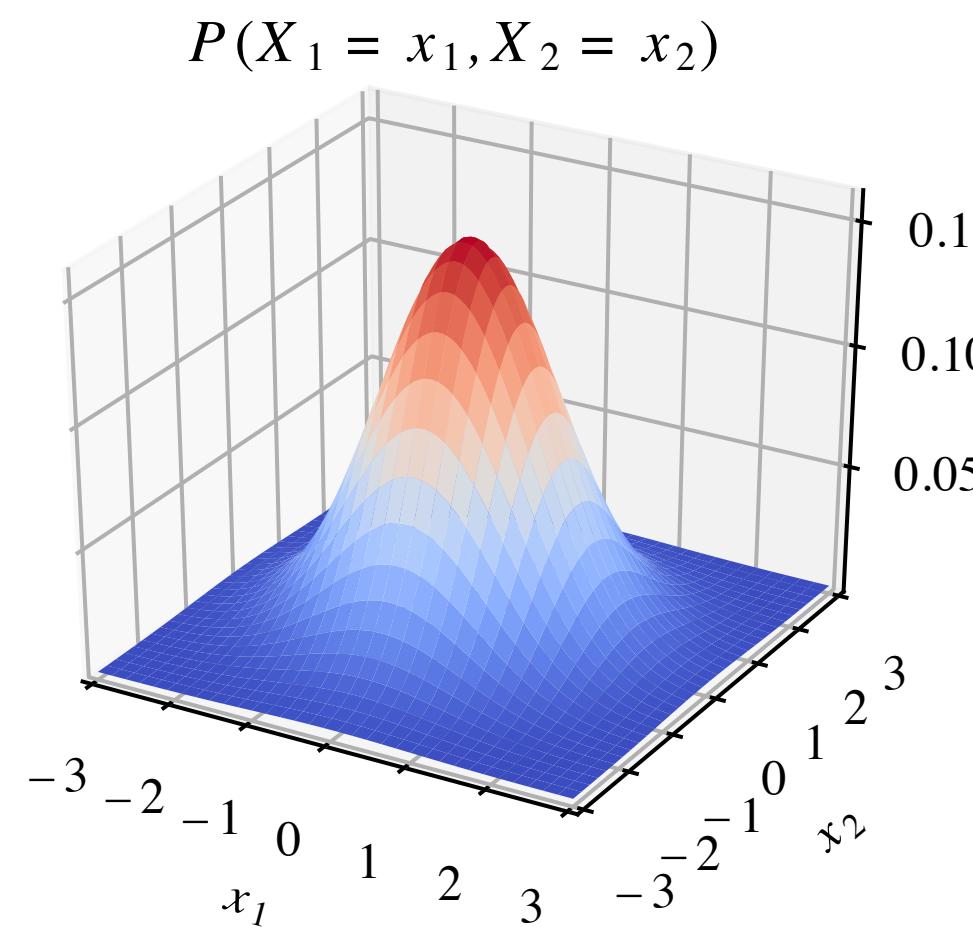
The conditional distribution of  $X$  given  $Y = y$  is given by *only* considering the events where  $Y = y$ .

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



# 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.

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

# 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

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

# Independence

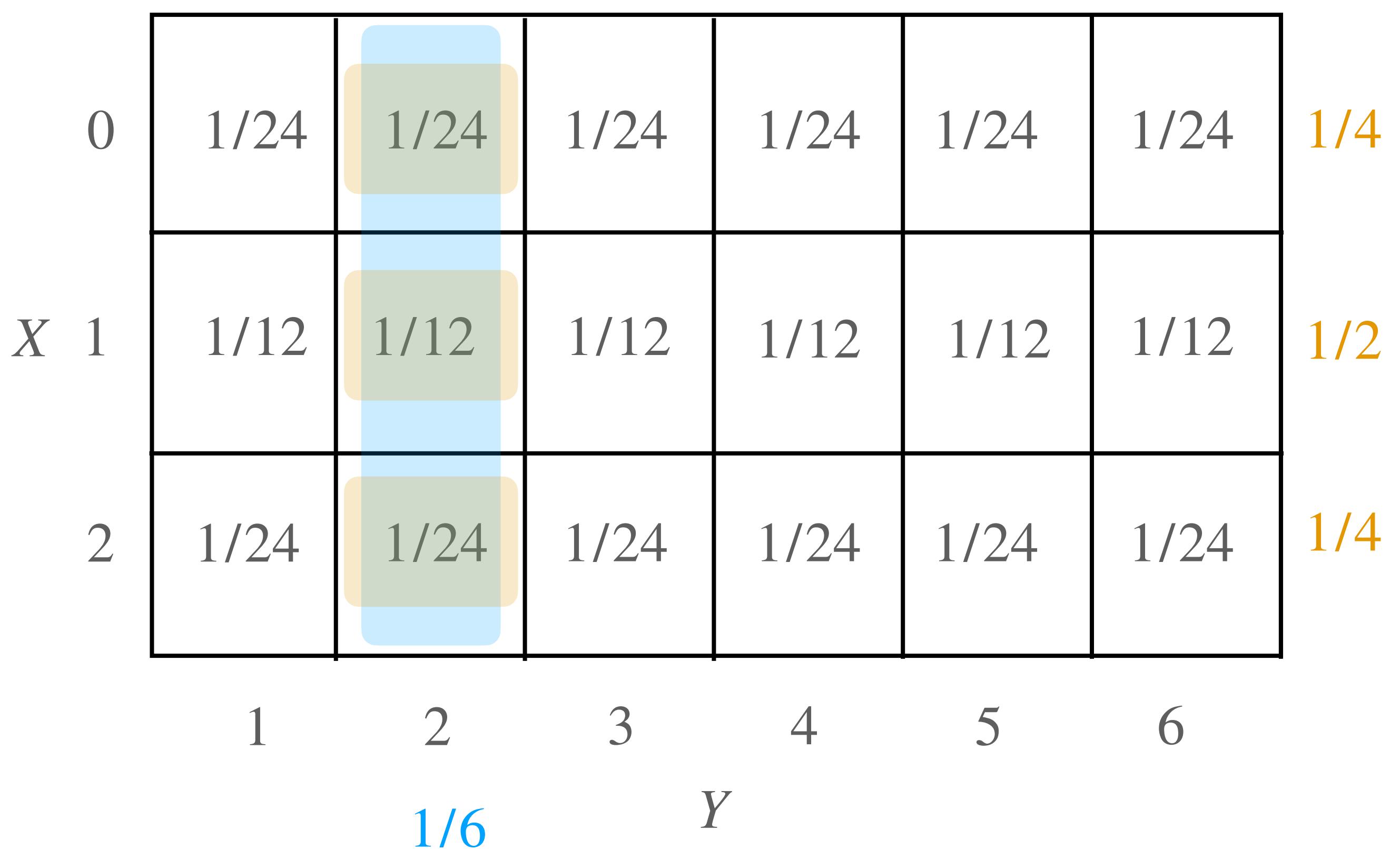
## Intuition and definition

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.

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



# Independence

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

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

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

and all the  $X_i$  have the same distribution.

*Very common assumption in ML!*

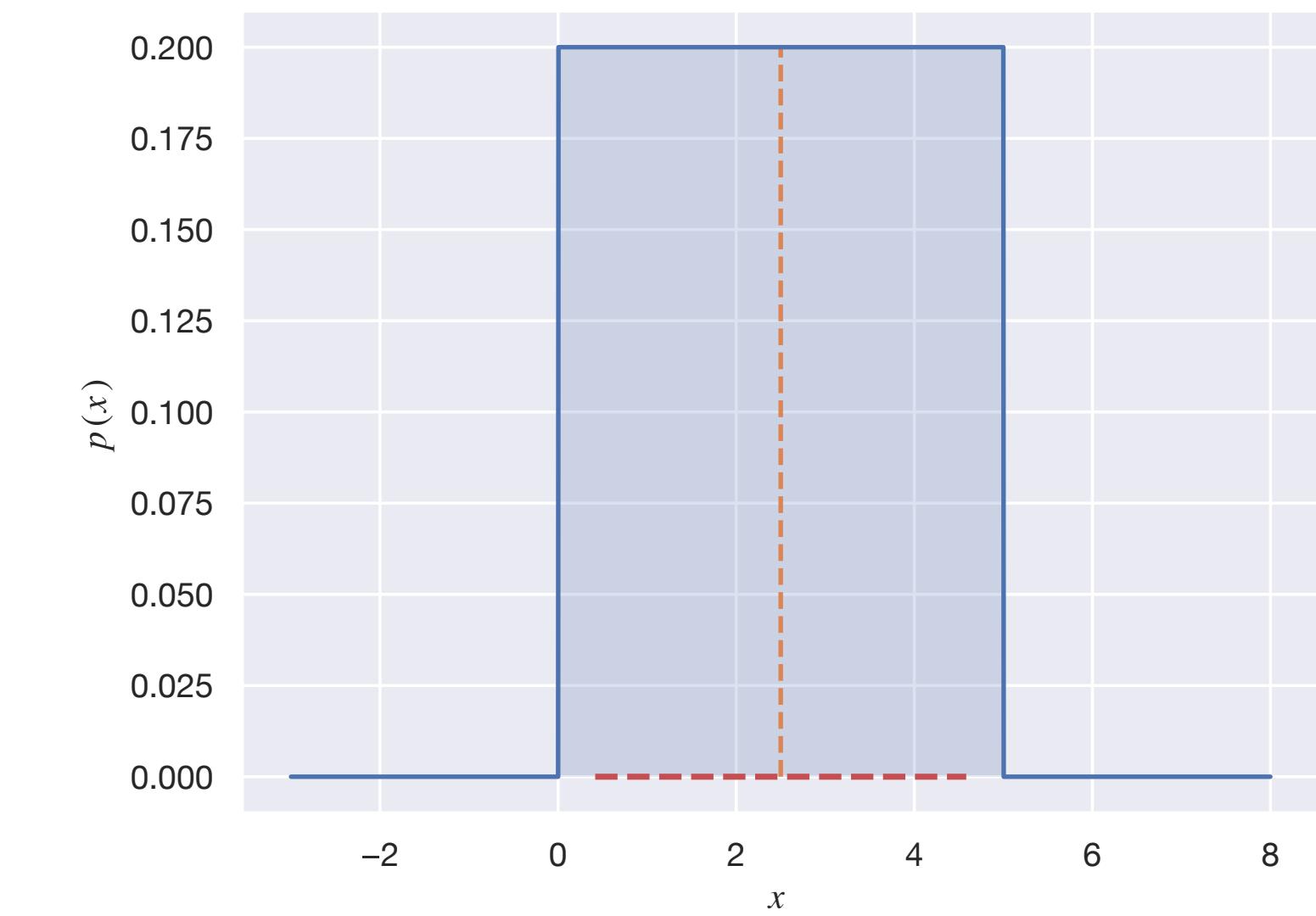
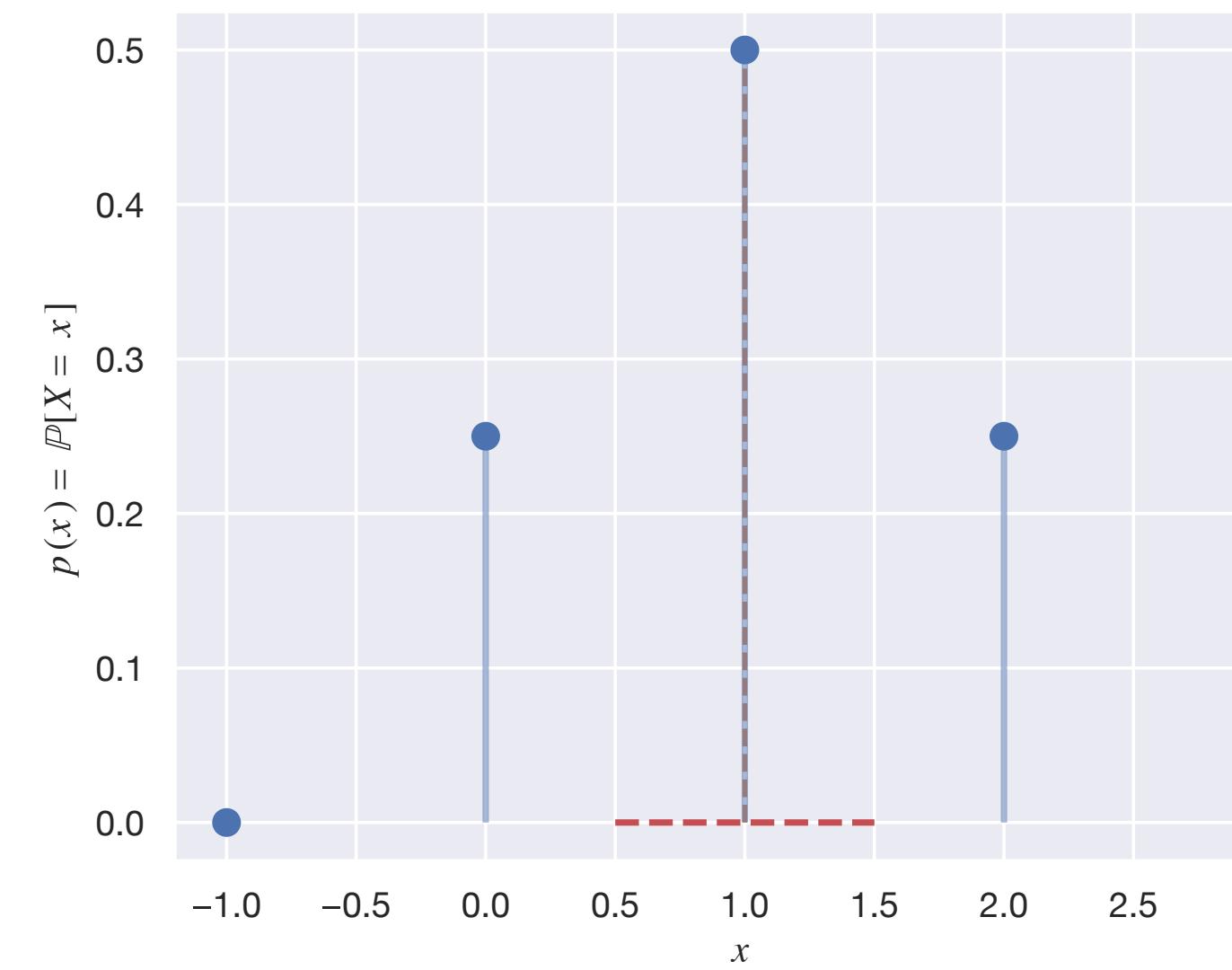
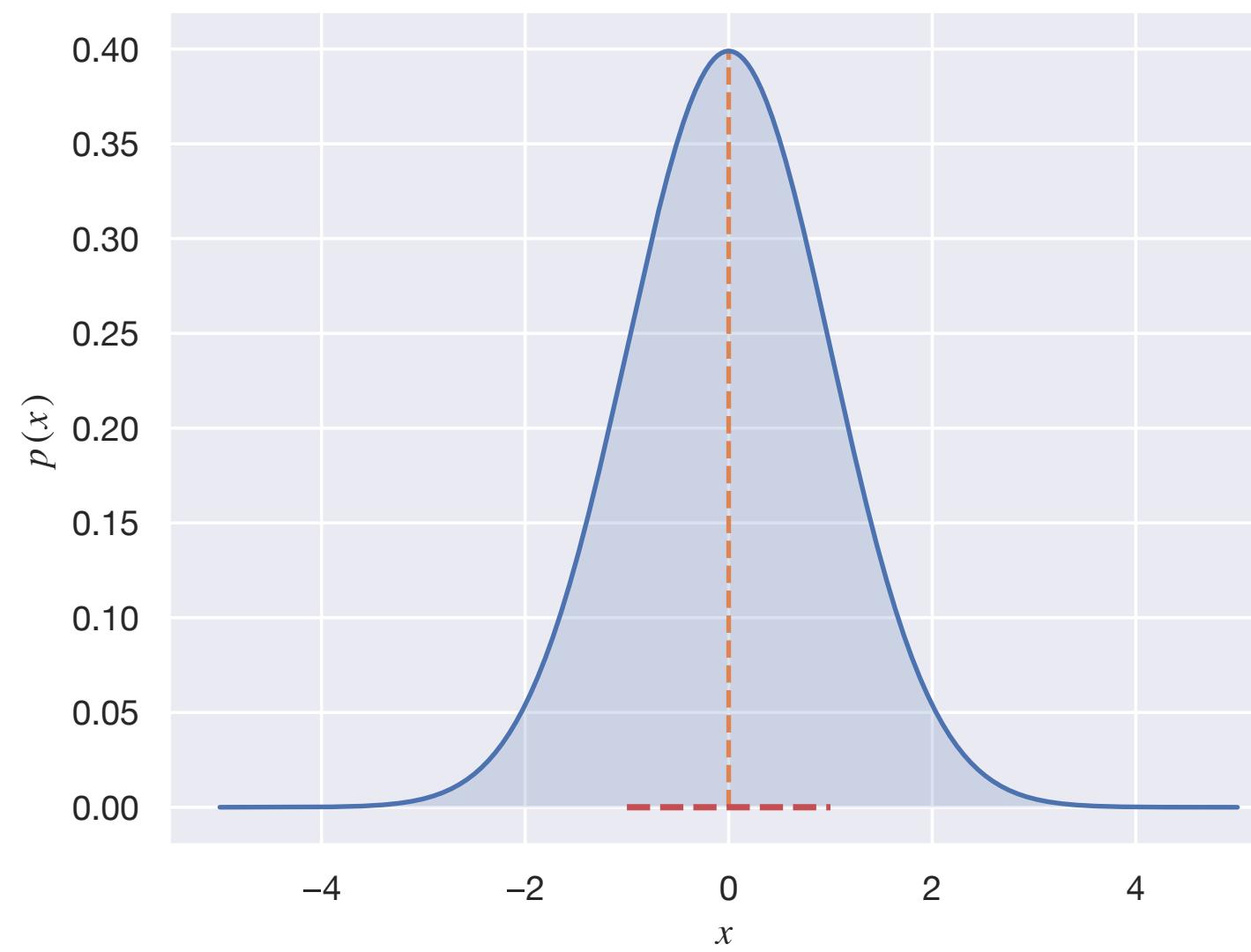
# Expectation

## 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

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

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

*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:

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

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

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

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

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

# 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*.

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

$\mathbb{E}[X | Y]$  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.

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

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

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

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

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

# 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:

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

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

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

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

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

# 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  $X$ :

$\mathbb{E}[Y | X]$  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  $\mathbb{E}[Y | X = x]$  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  $\mathbb{E}[Y | X]$ ?

What is its mean?

# Conditional Expectation

## Properties of conditional expectation

**Independence.** If  $X$  is independent of  $Y$ ,

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

**Pulling out what's known.** For any function  $h$ ,

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

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

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

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

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

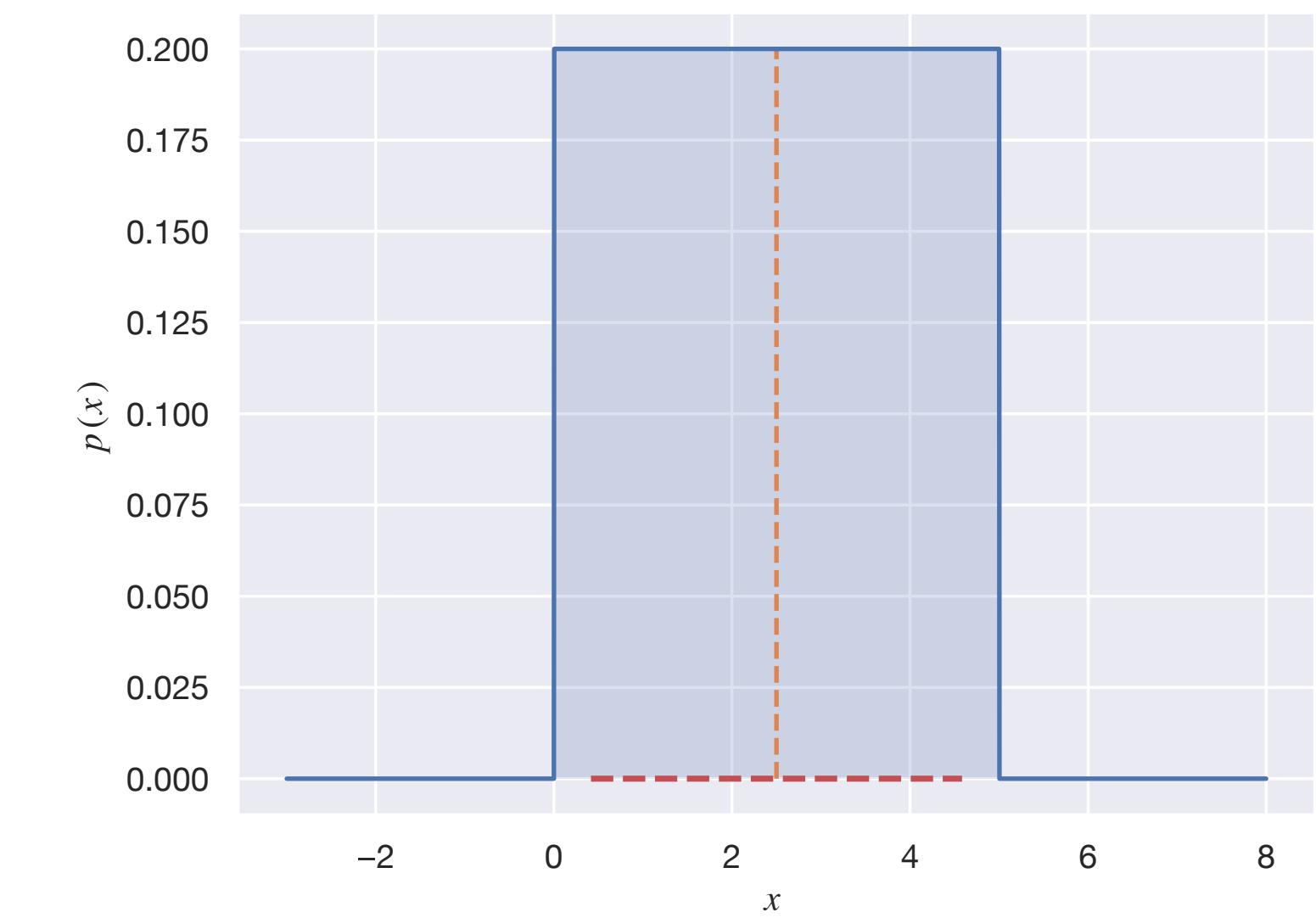
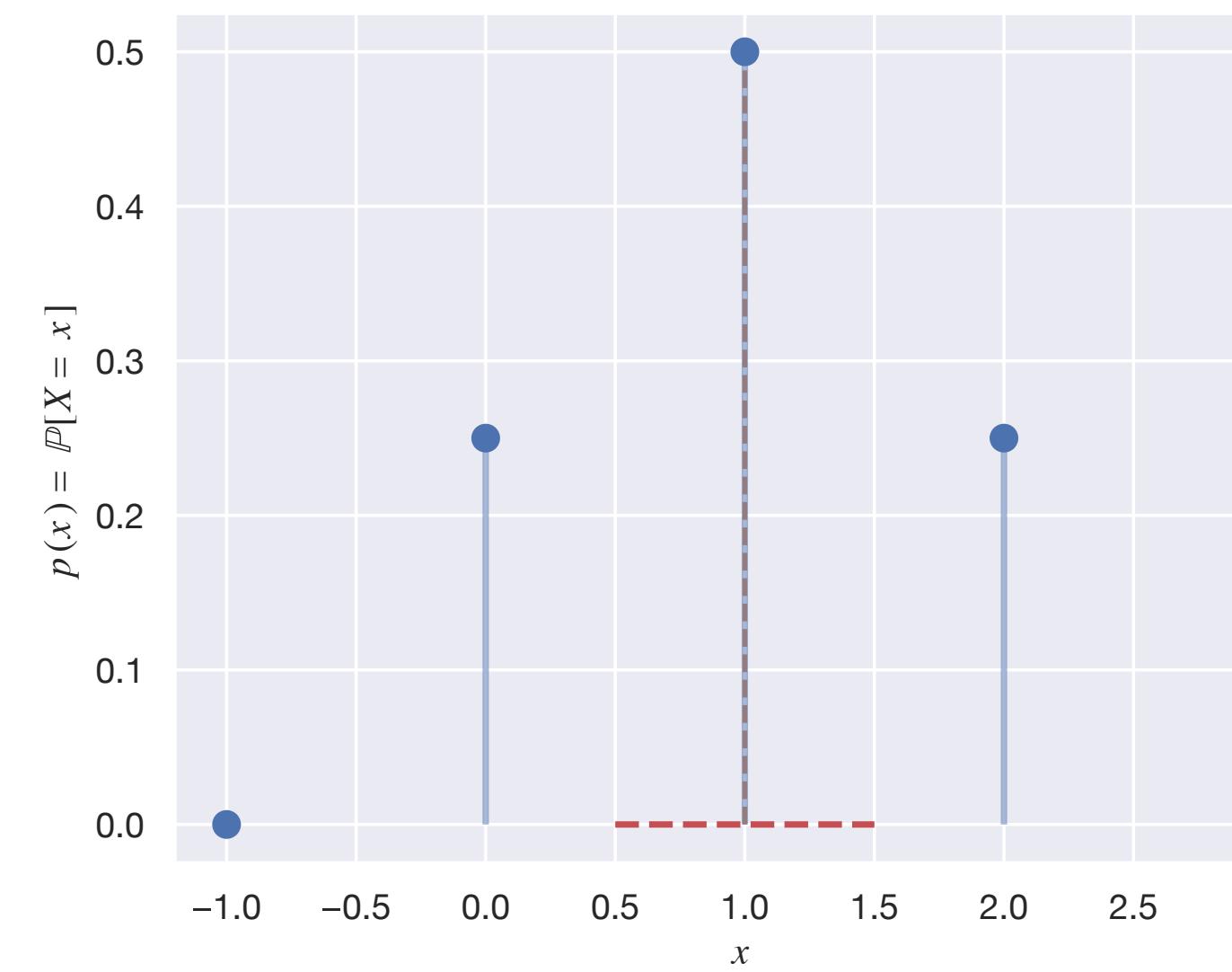
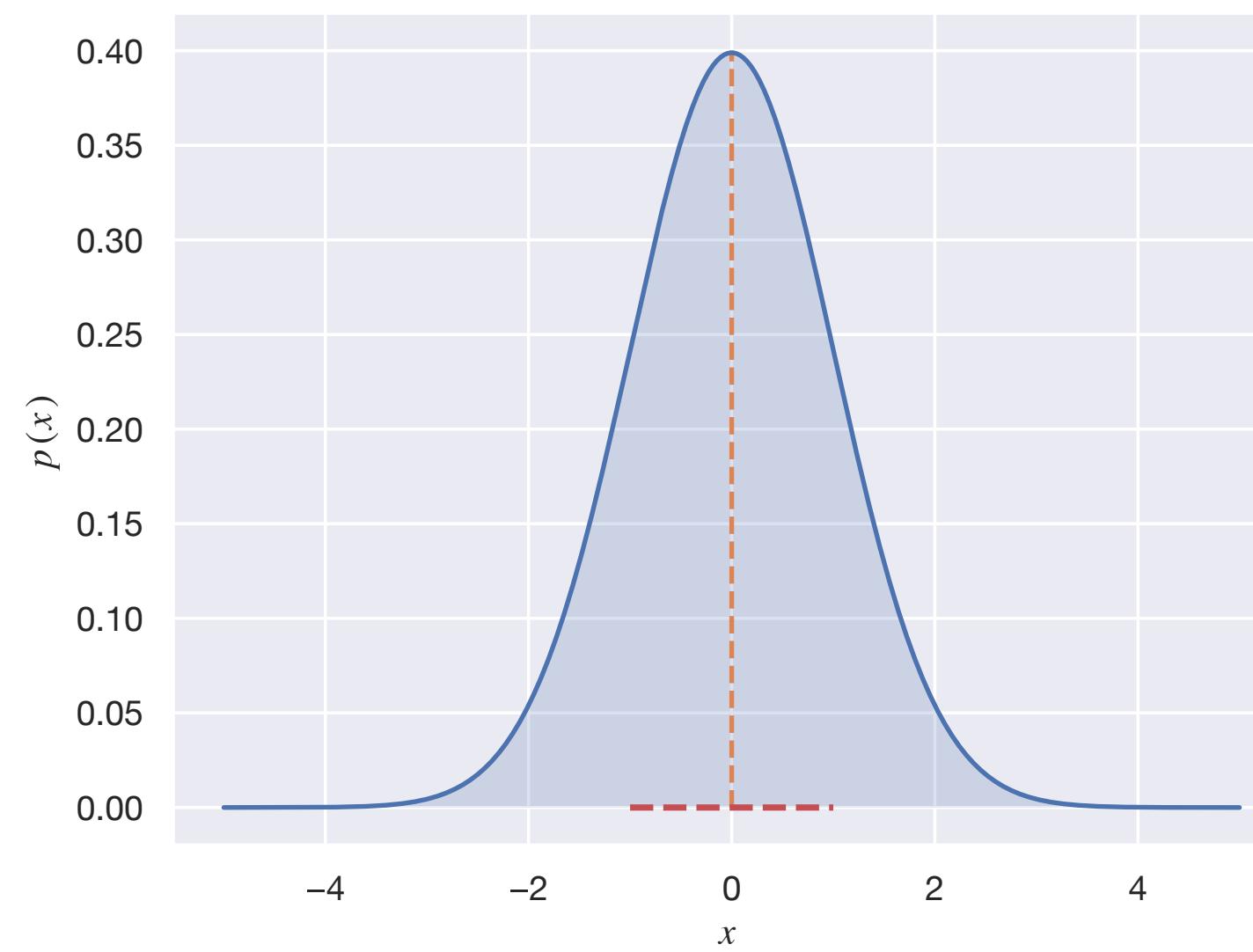
# Variance

## 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  $\text{Var}(X)$  is:

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

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

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

The standard deviation is  $\sqrt{\text{Var}(X)}$ .

# Variance

## Properties of variance

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

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

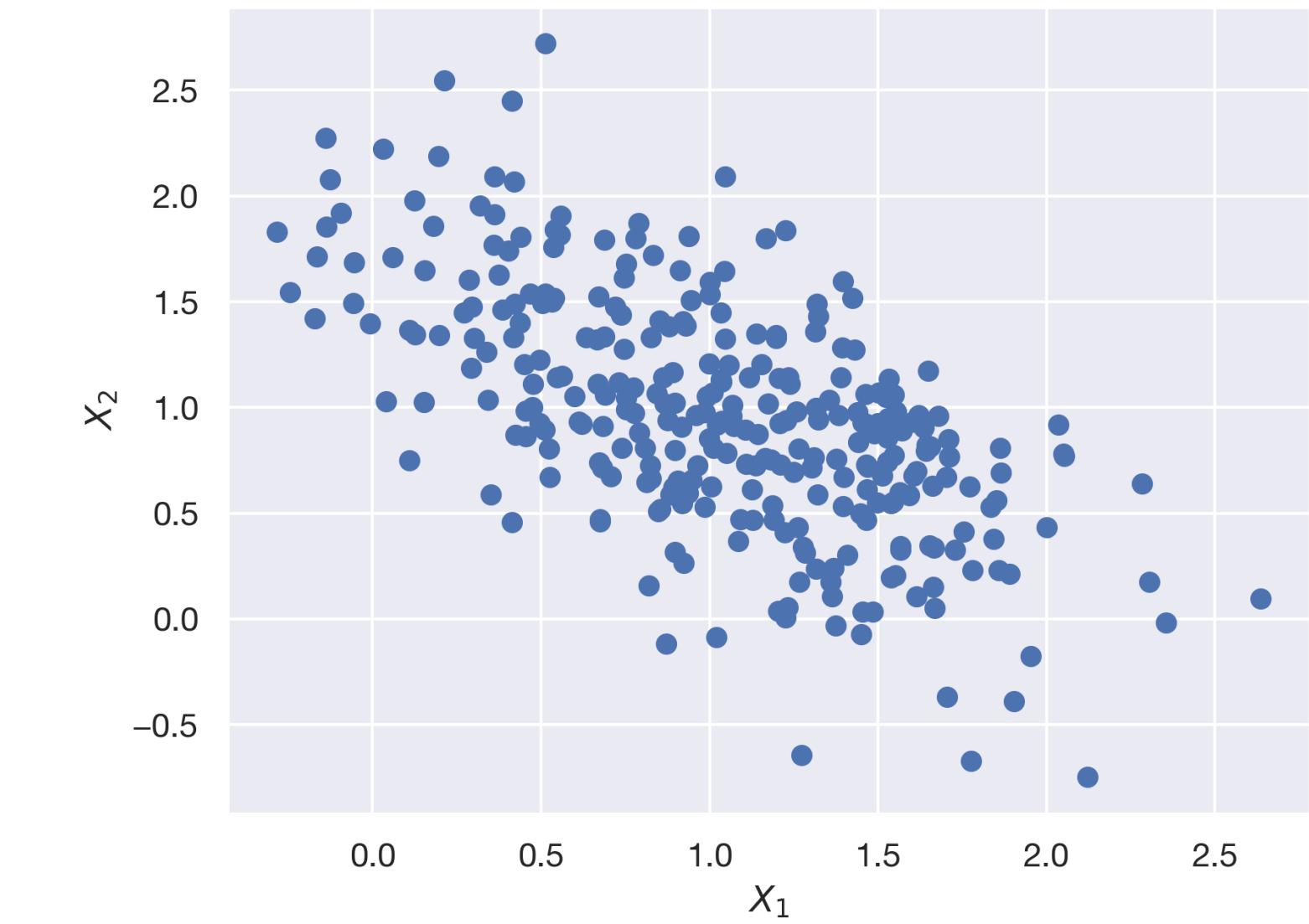
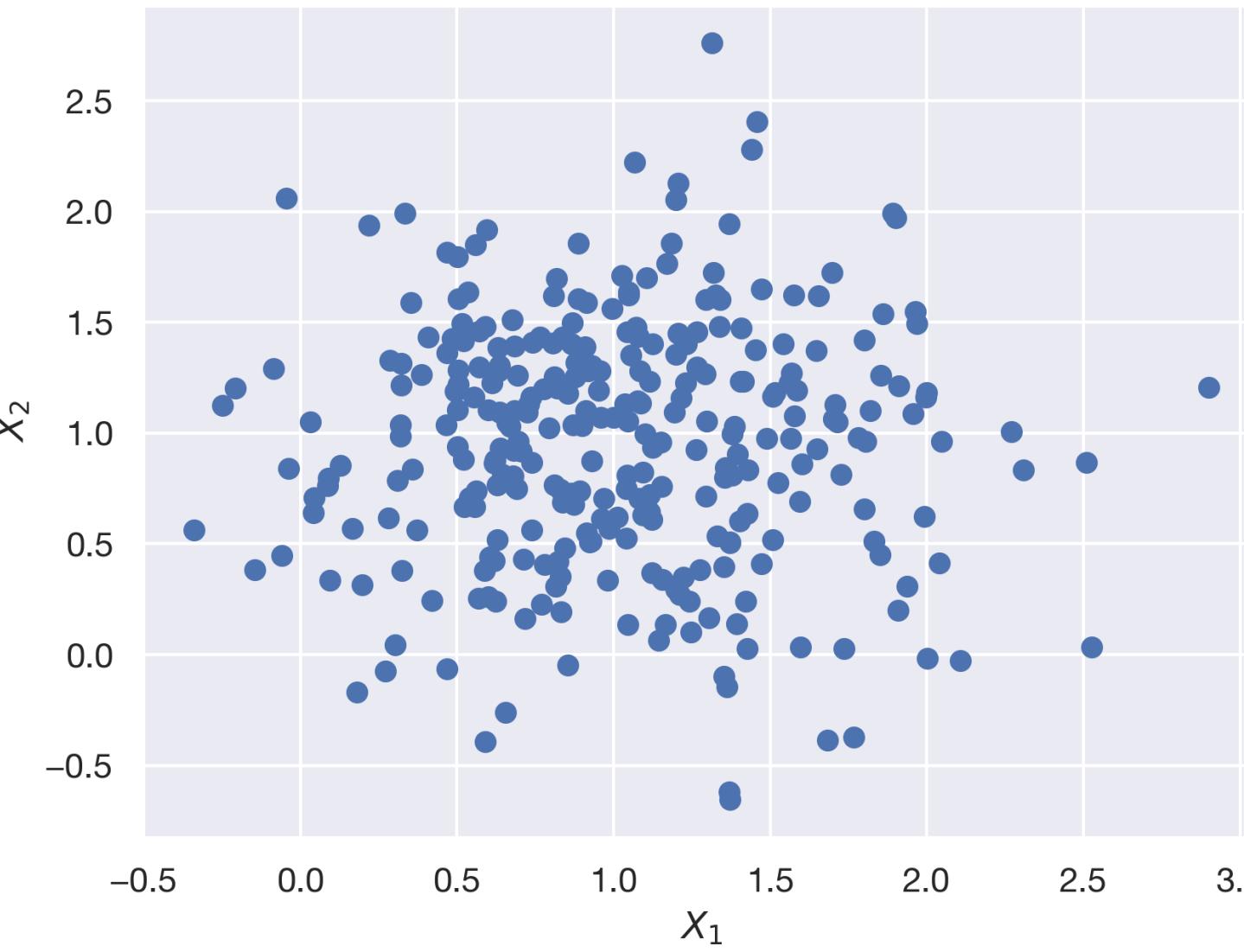
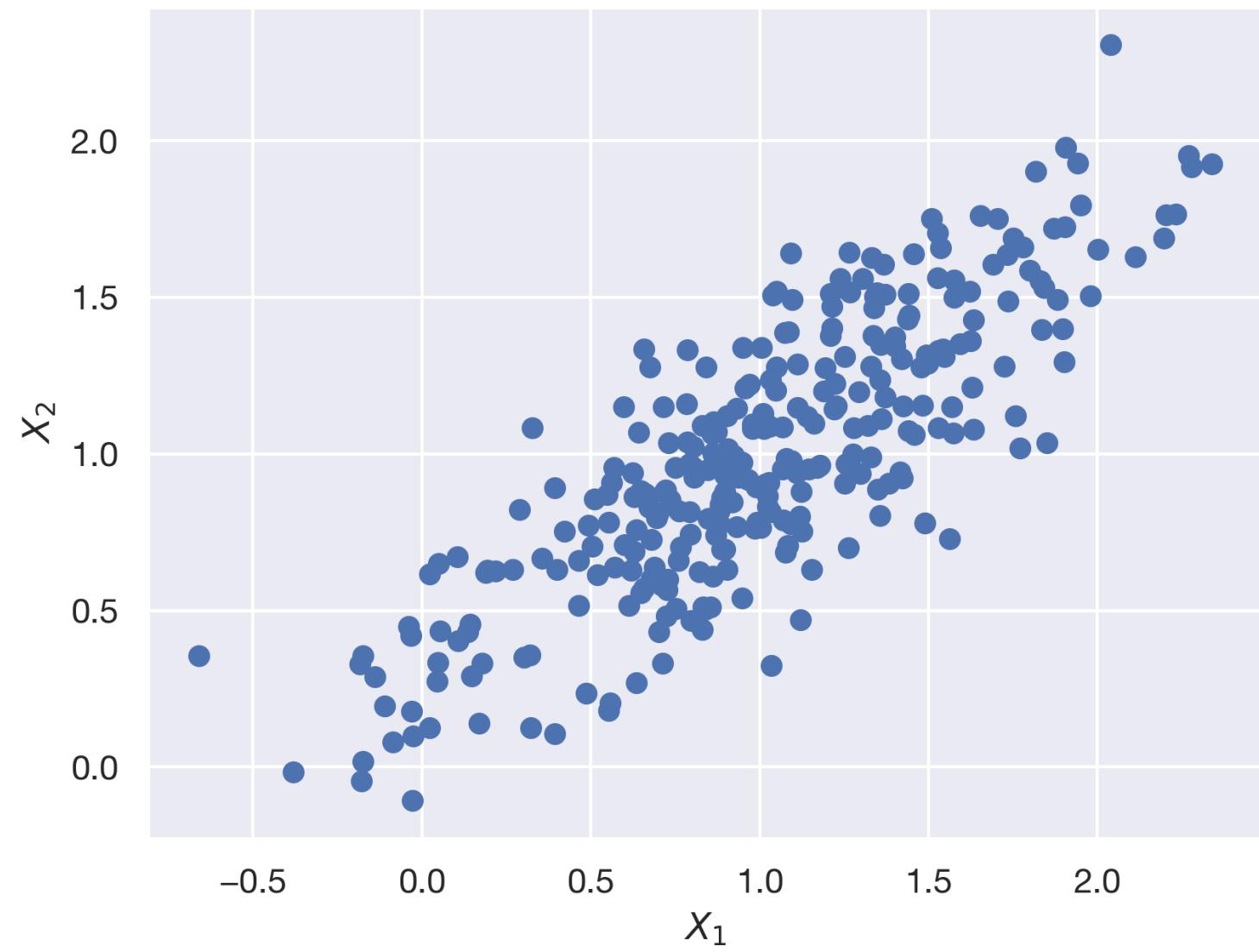
If  $X_1, \dots, X_n$  are independent (more generally, *uncorrelated*),

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

# Covariance

## Intuition

The covariance measures the linear relationship between two random variables.



# 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). Equivalently:

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

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

Covariance follows the “symmetry” property:

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

Covariance follows the “bilinearity” property:

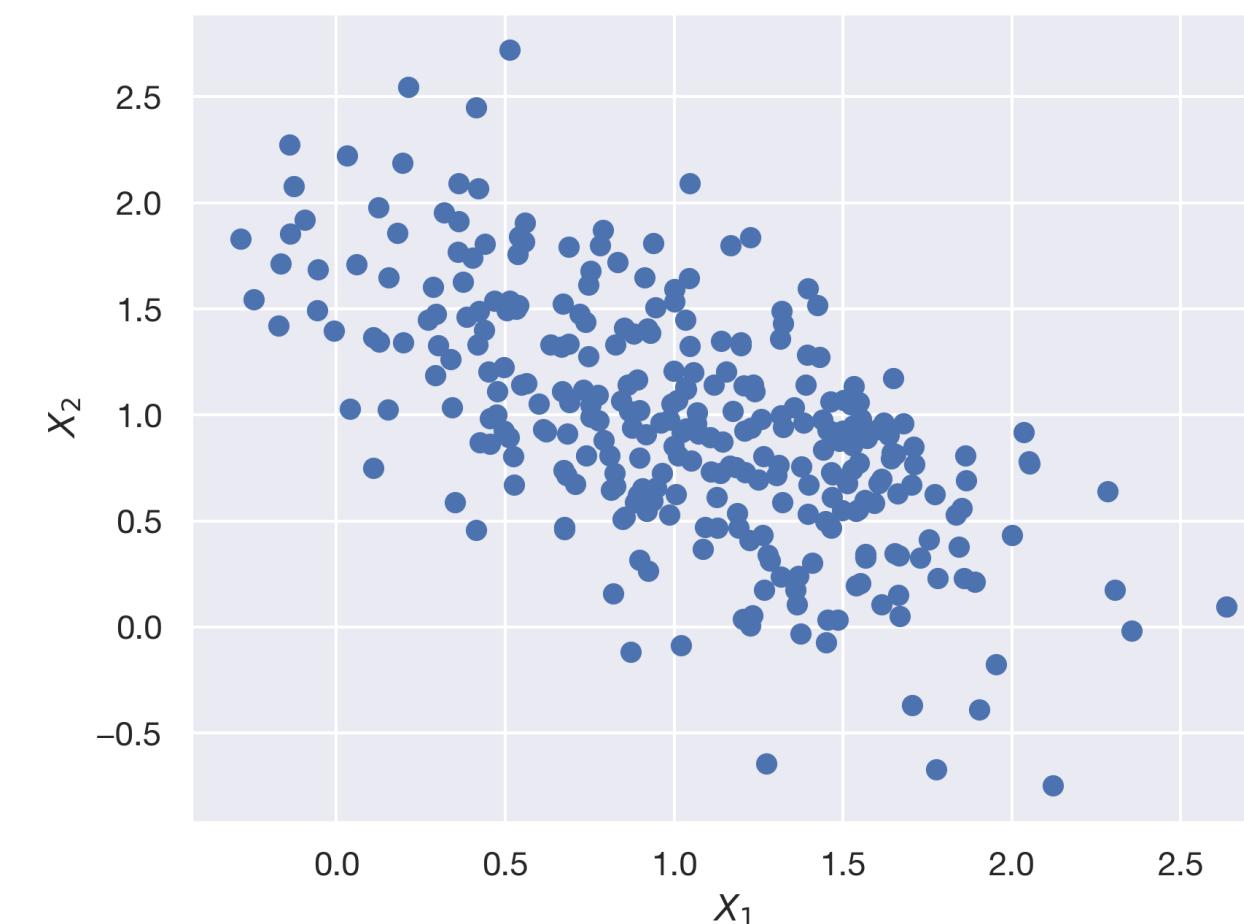
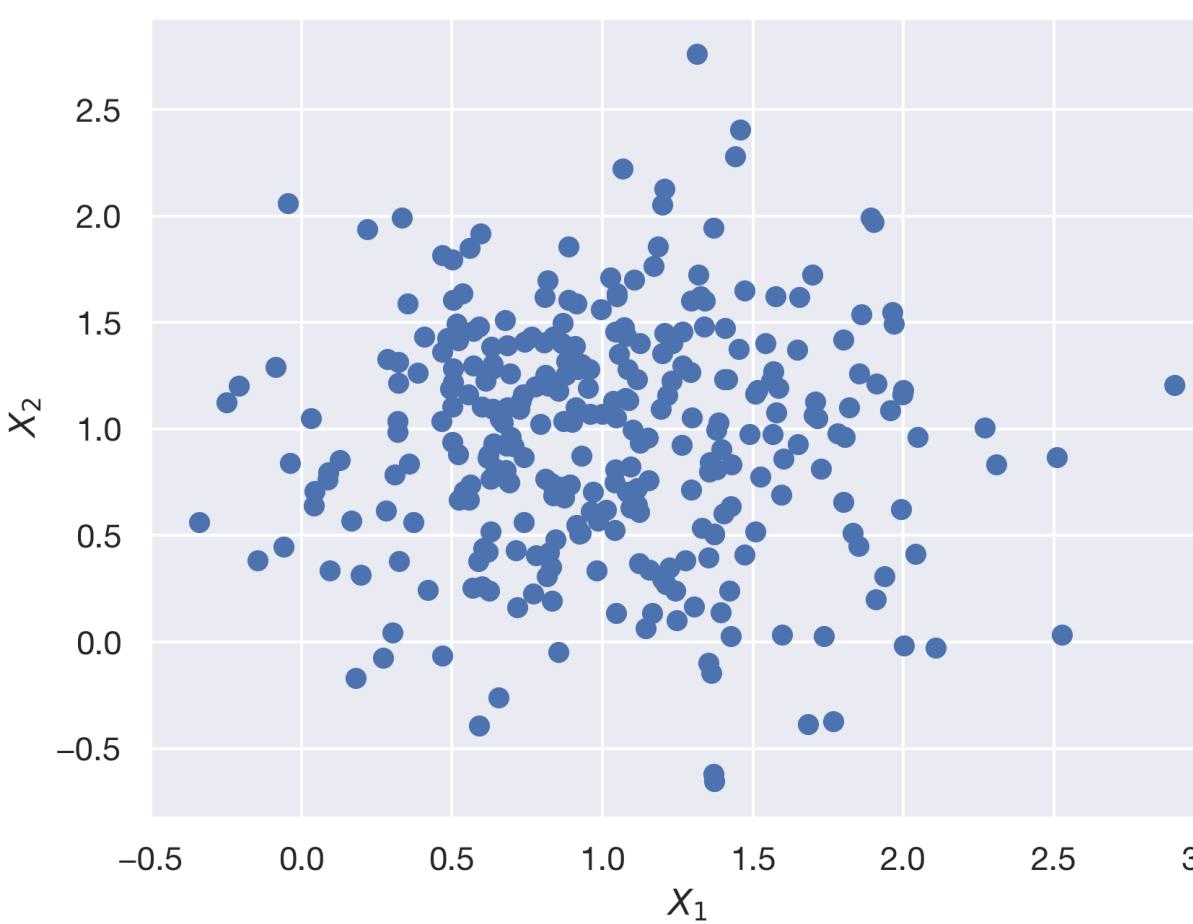
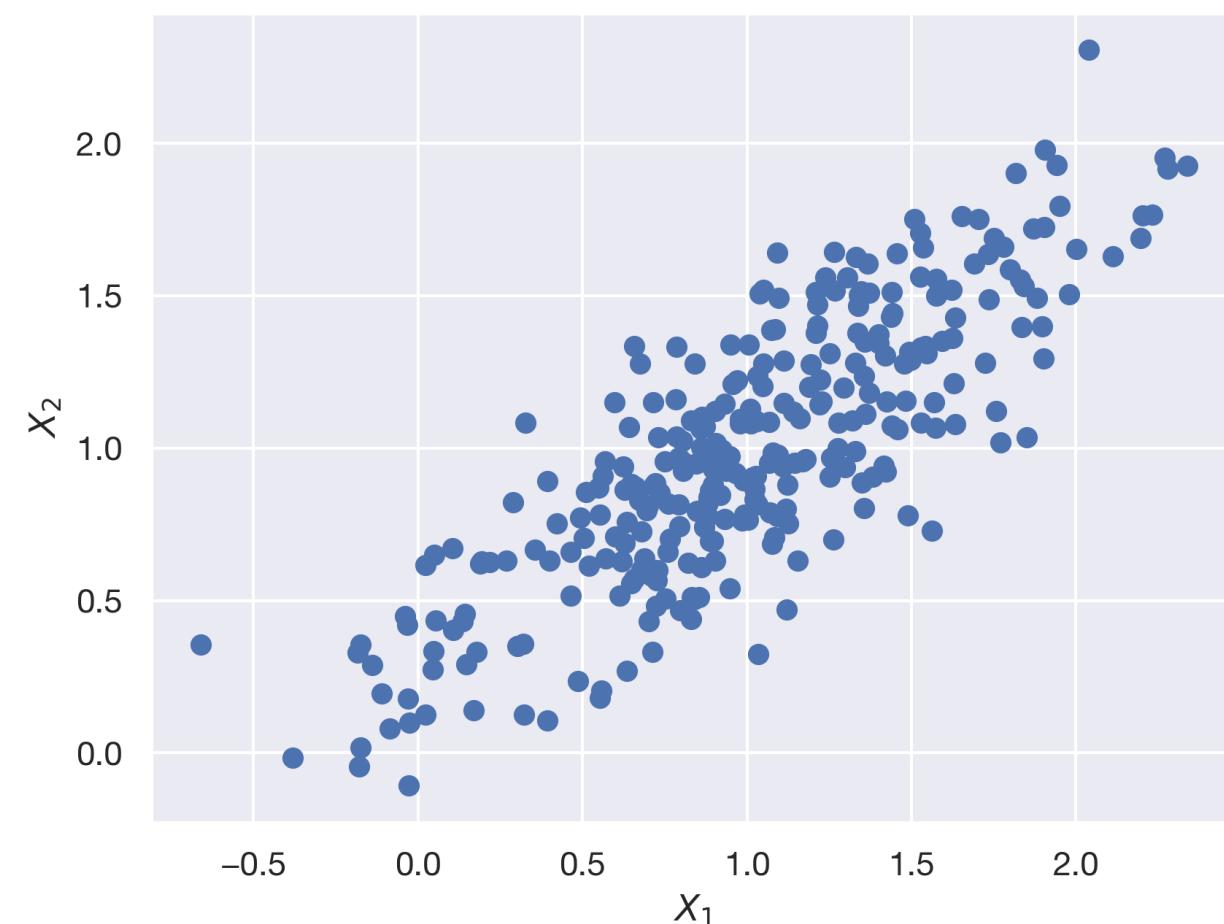
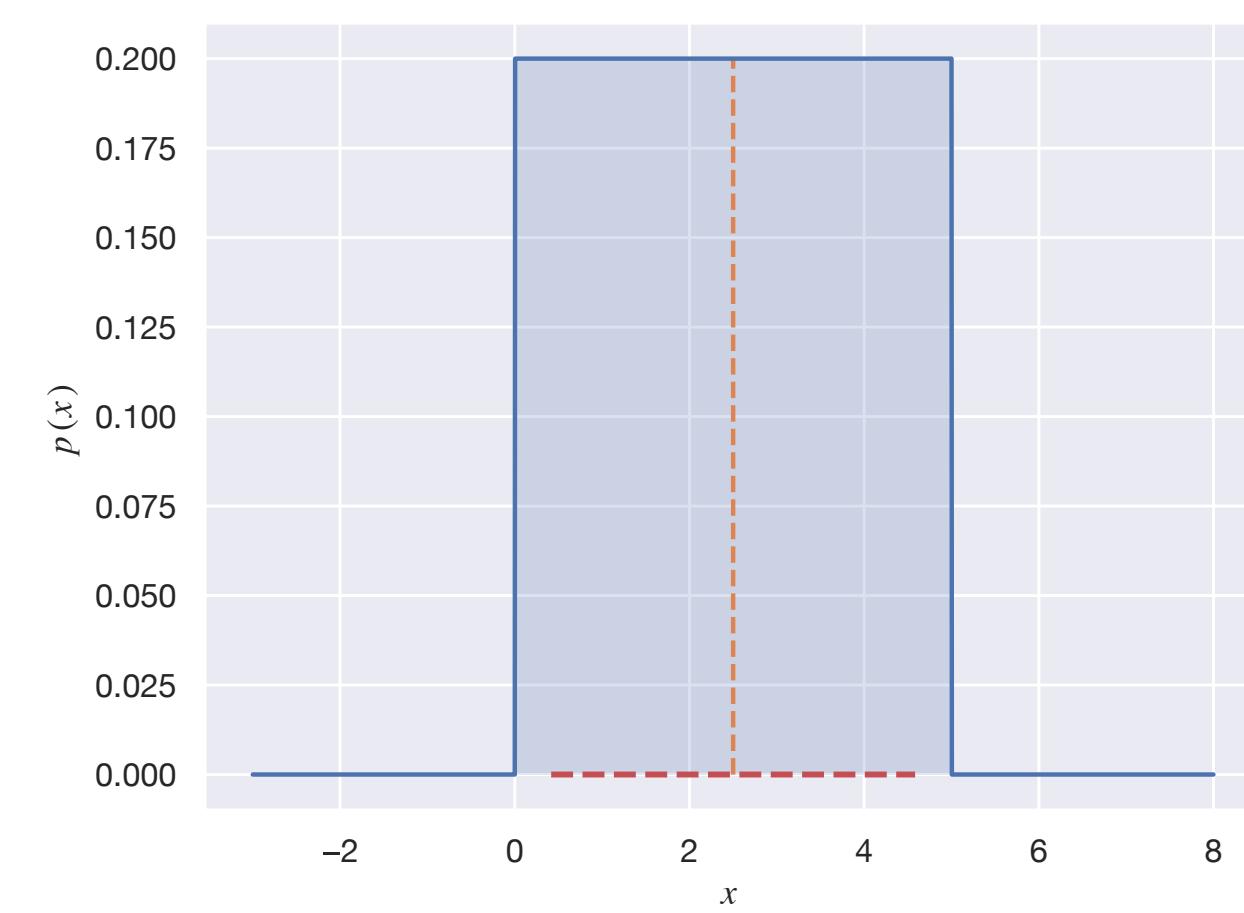
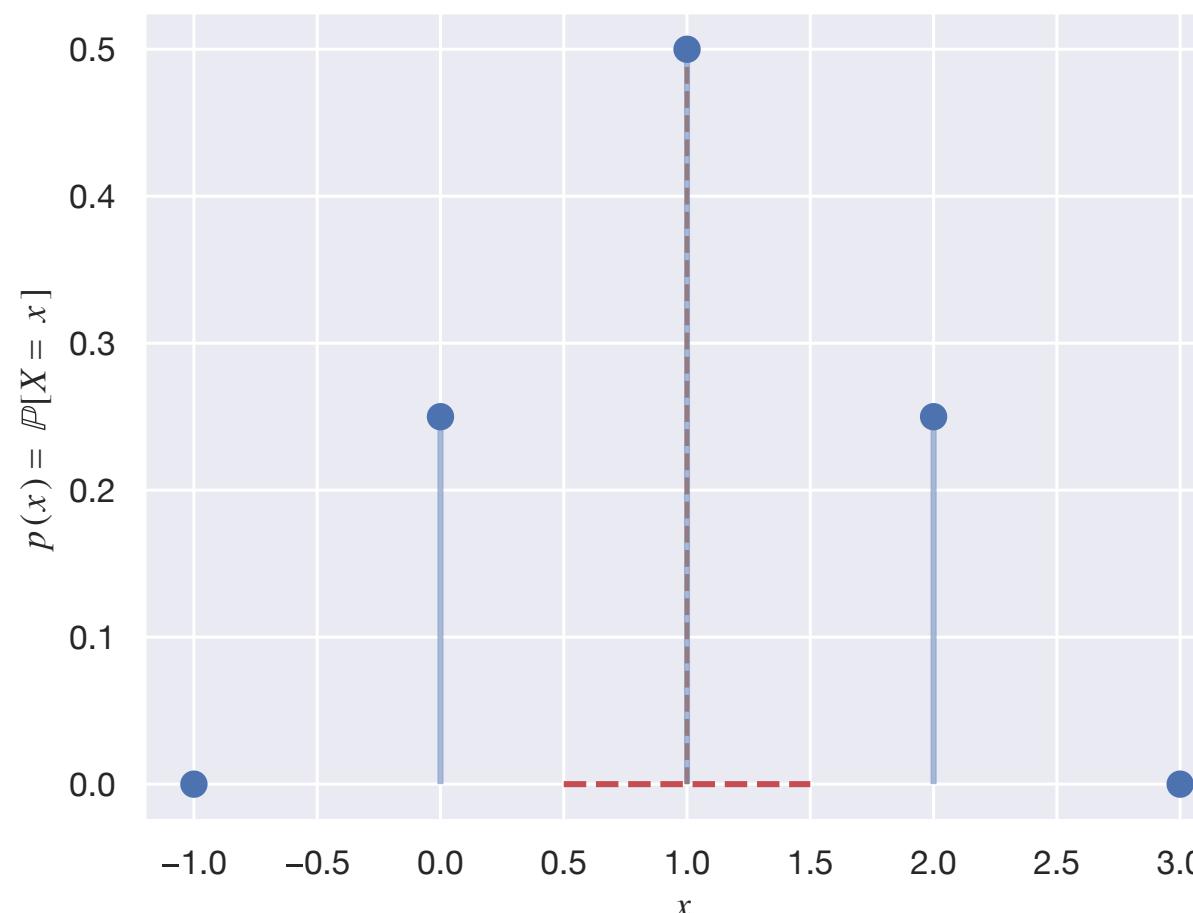
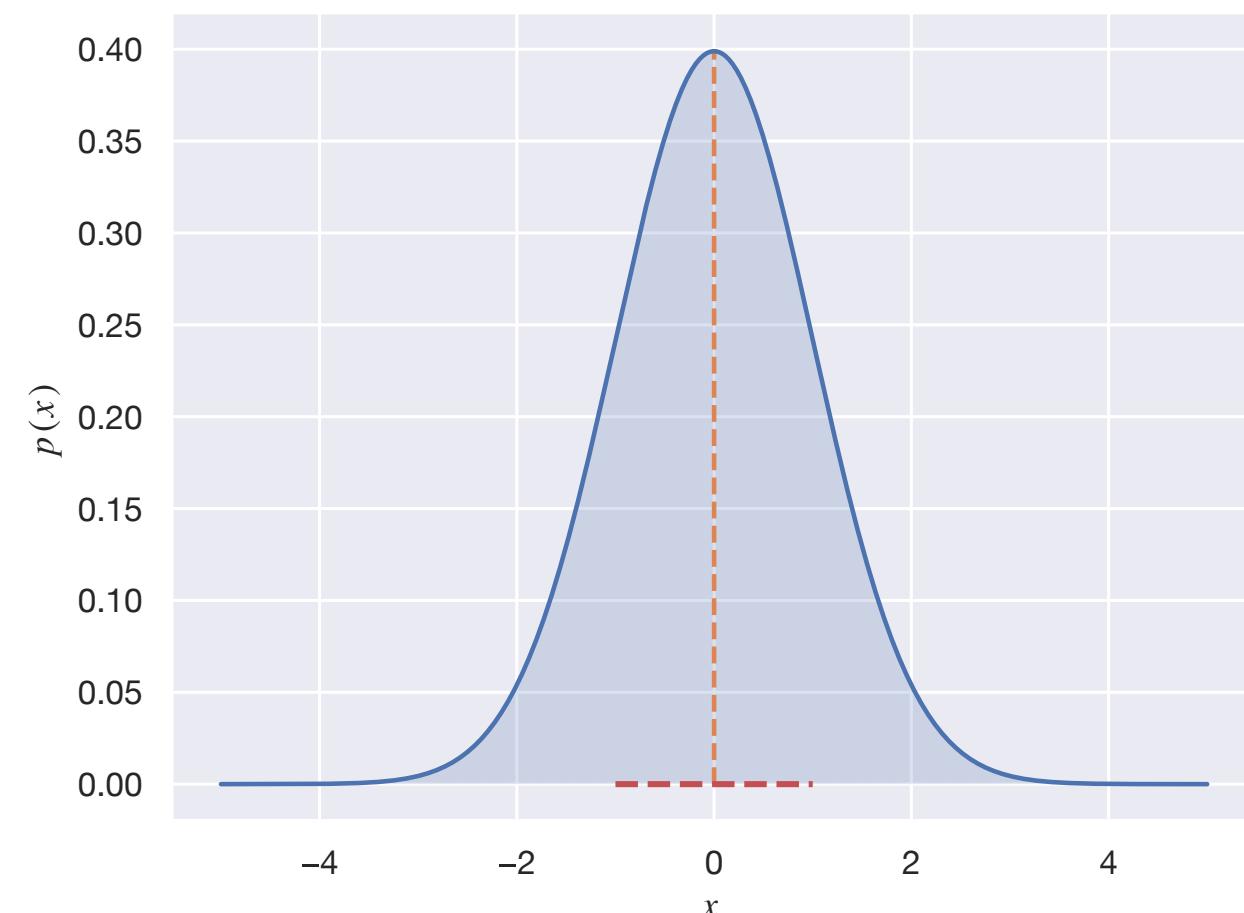
$$\text{Cov}(\alpha X + \beta Y, Z) = \alpha \text{Cov}(X, Z) + \beta \text{Cov}(Y, Z).$$

Covariance follows the “positive definiteness” property:

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

# 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

## Covariance Matrix

The variance of a random vector generalizes to the [covariance matrix](#)

In the  $d = 2$  case,

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

*What do you notice about this matrix?*

# Random Vectors

## Covariance Matrix

The variance of a random vector generalizes to the covariance matrix

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

In general,  $\Sigma_{i,j} = \text{Cov}(X_i, X_j)$ .

# Random Vectors

## Covariance Matrix

The covariance matrix is **symmetric**.

$$\Sigma = \Sigma^T.$$

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

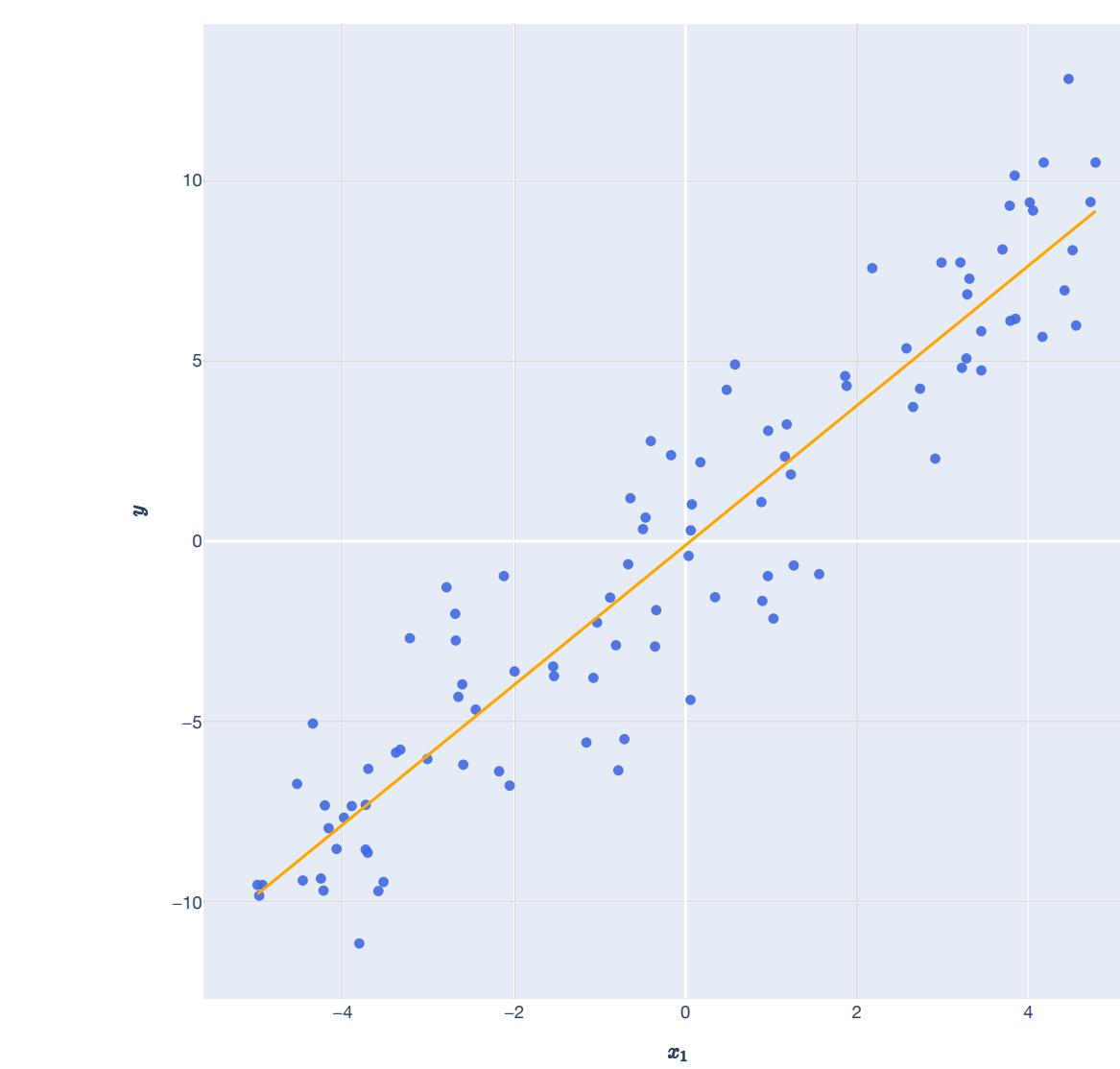
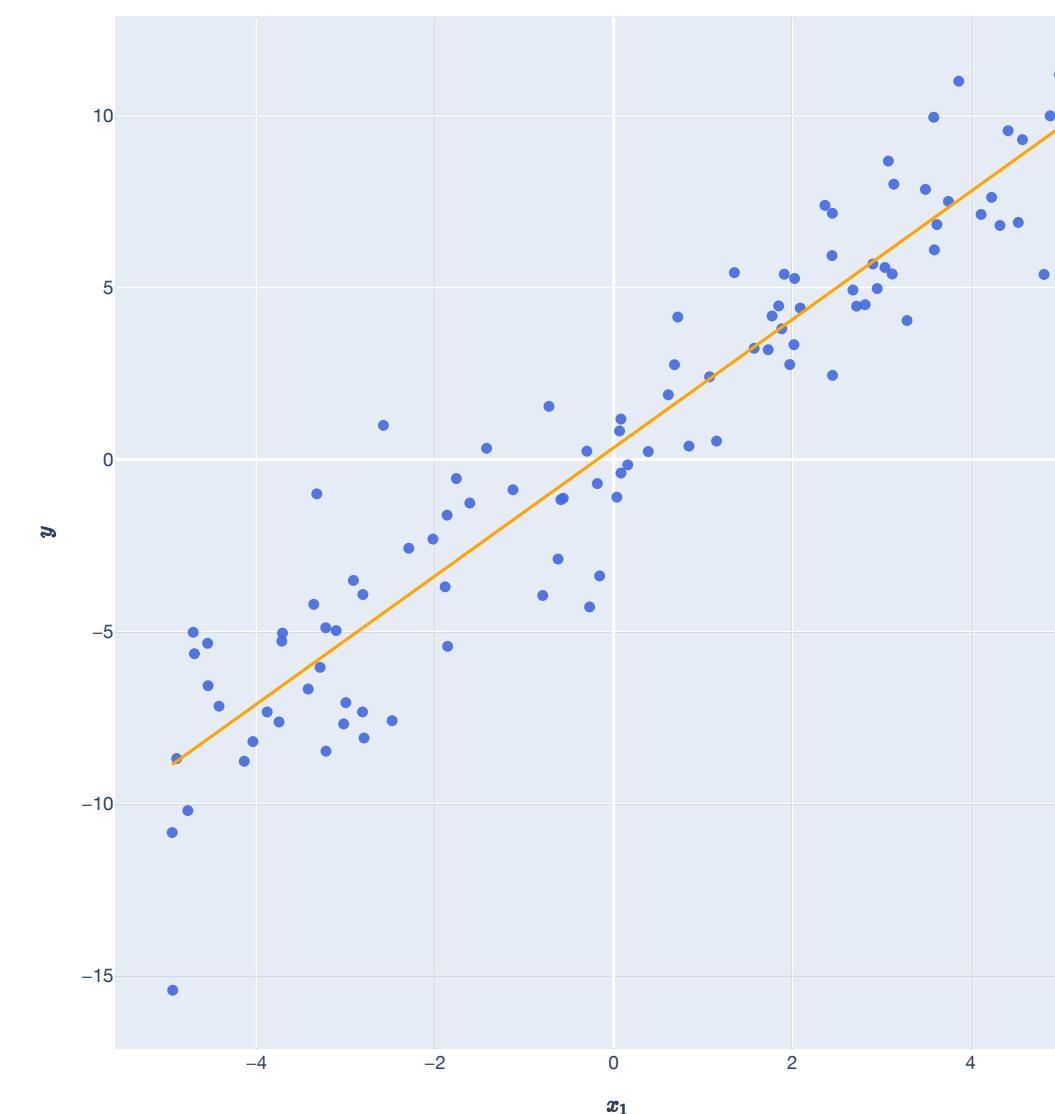
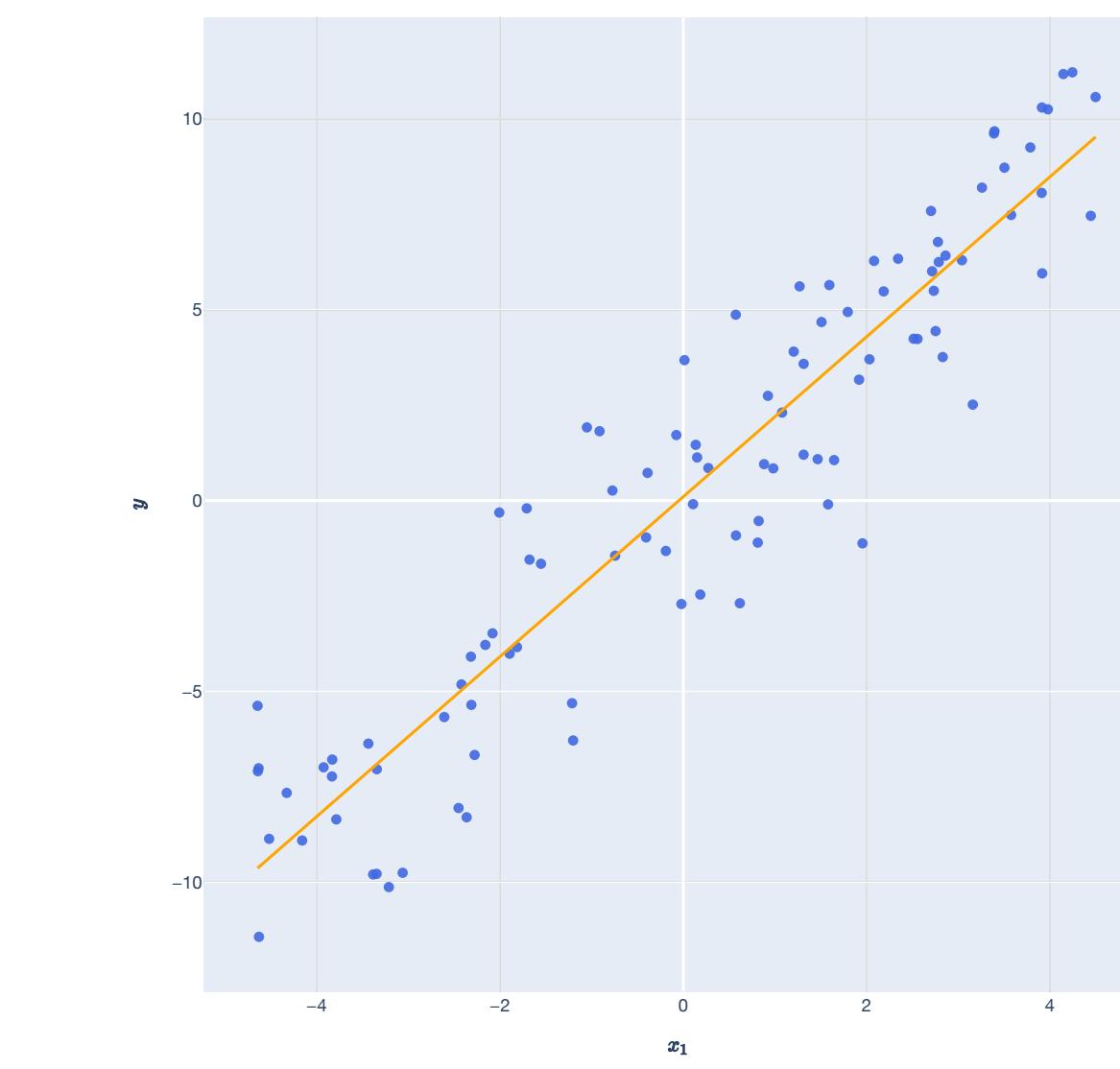
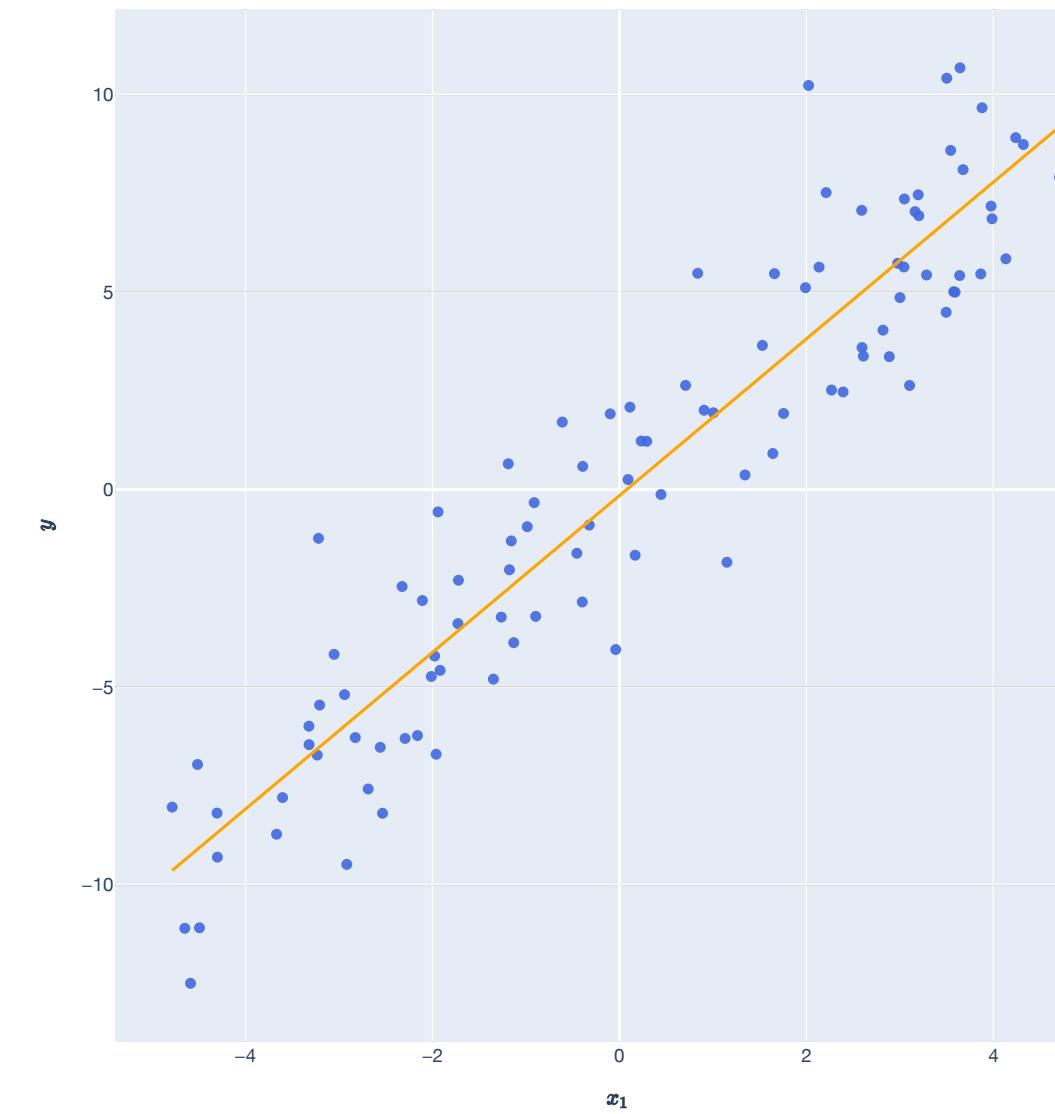
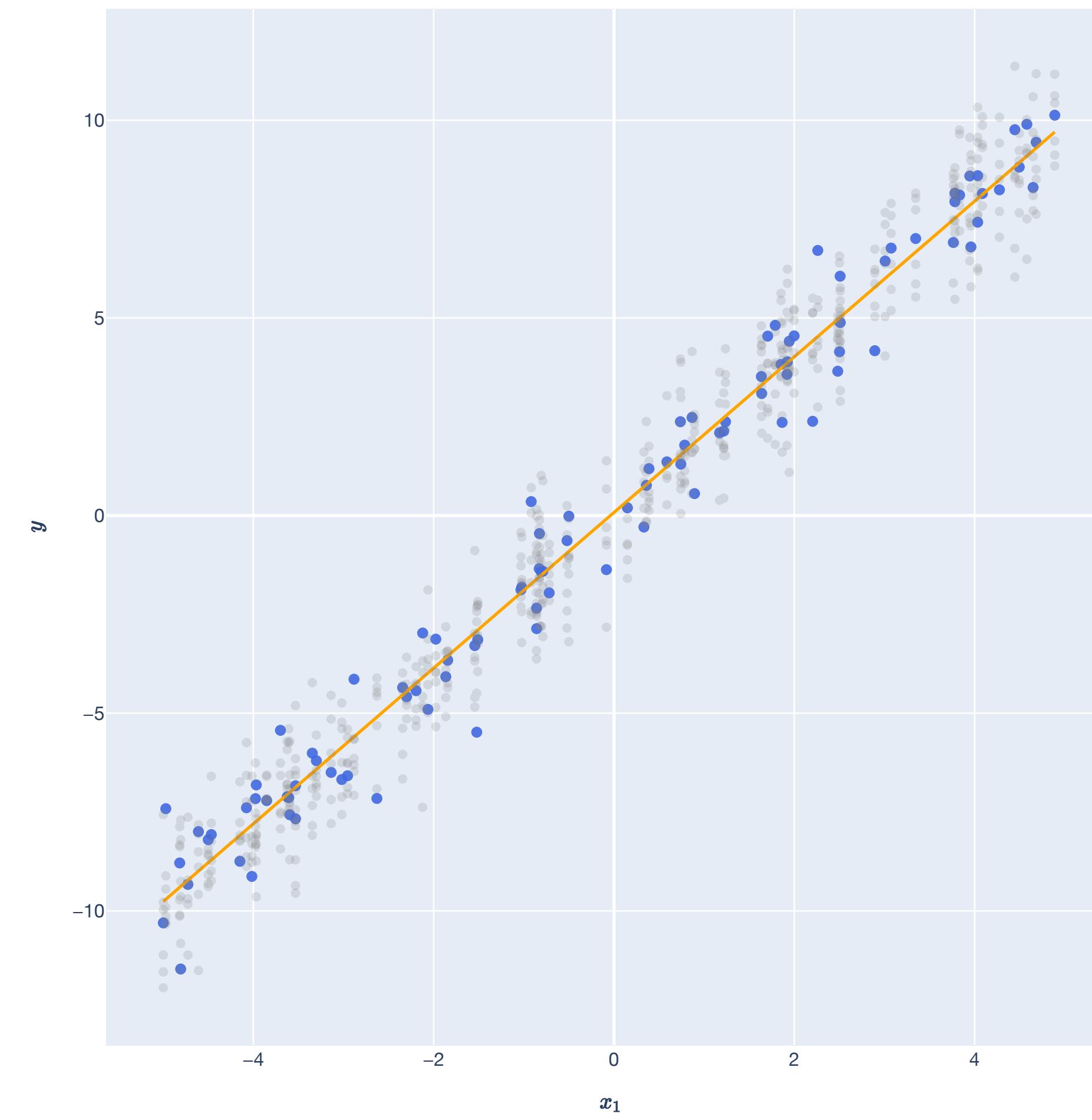
$$\mathbf{x}^T \Sigma \mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

# Data as random

Modeling regression with probability

# Regression

## Modeling randomness



# Regression

## Setup (Example View)

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

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

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

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

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)

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

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

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

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

Ultimate 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

Least squares expanded is just:

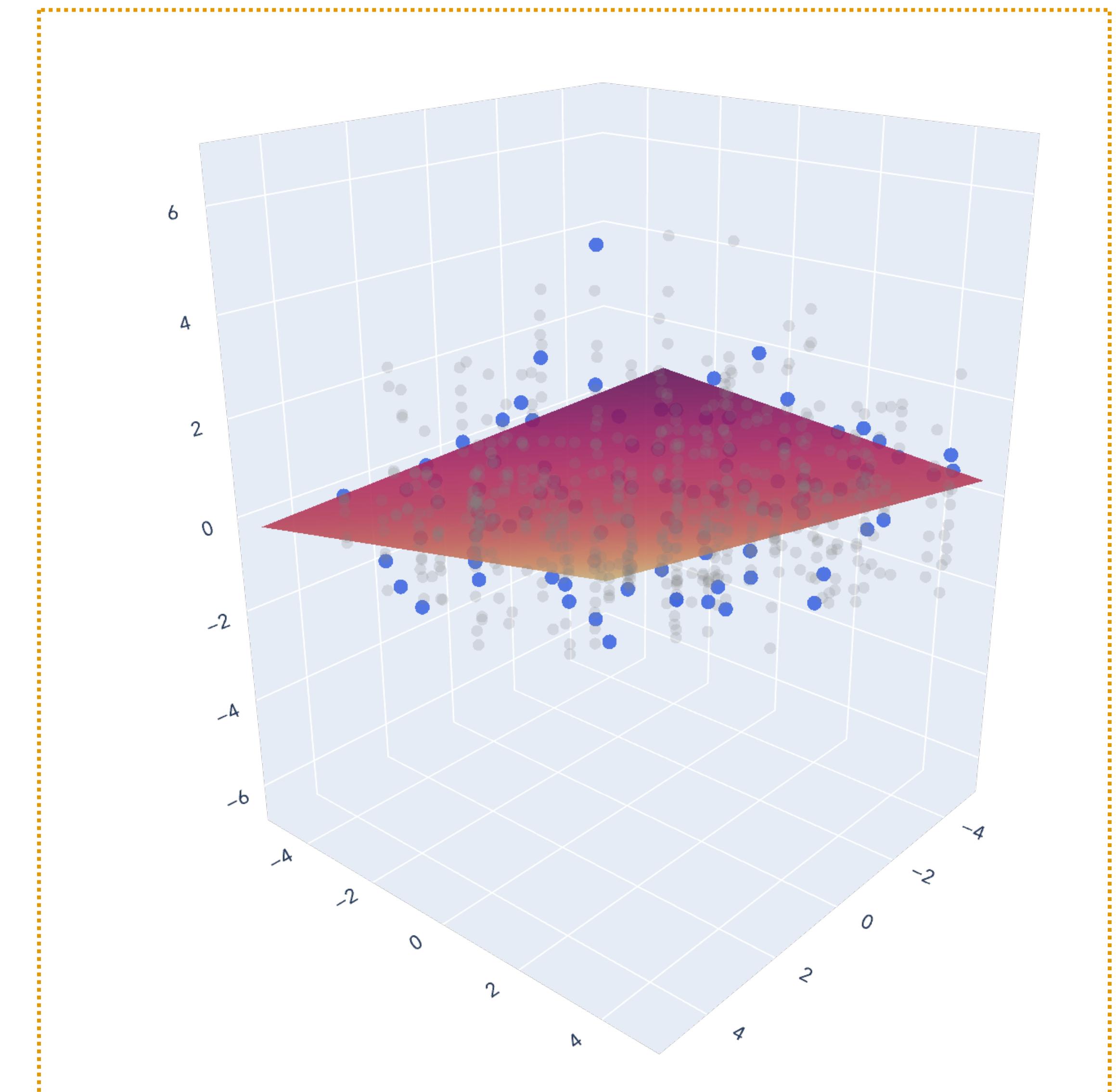
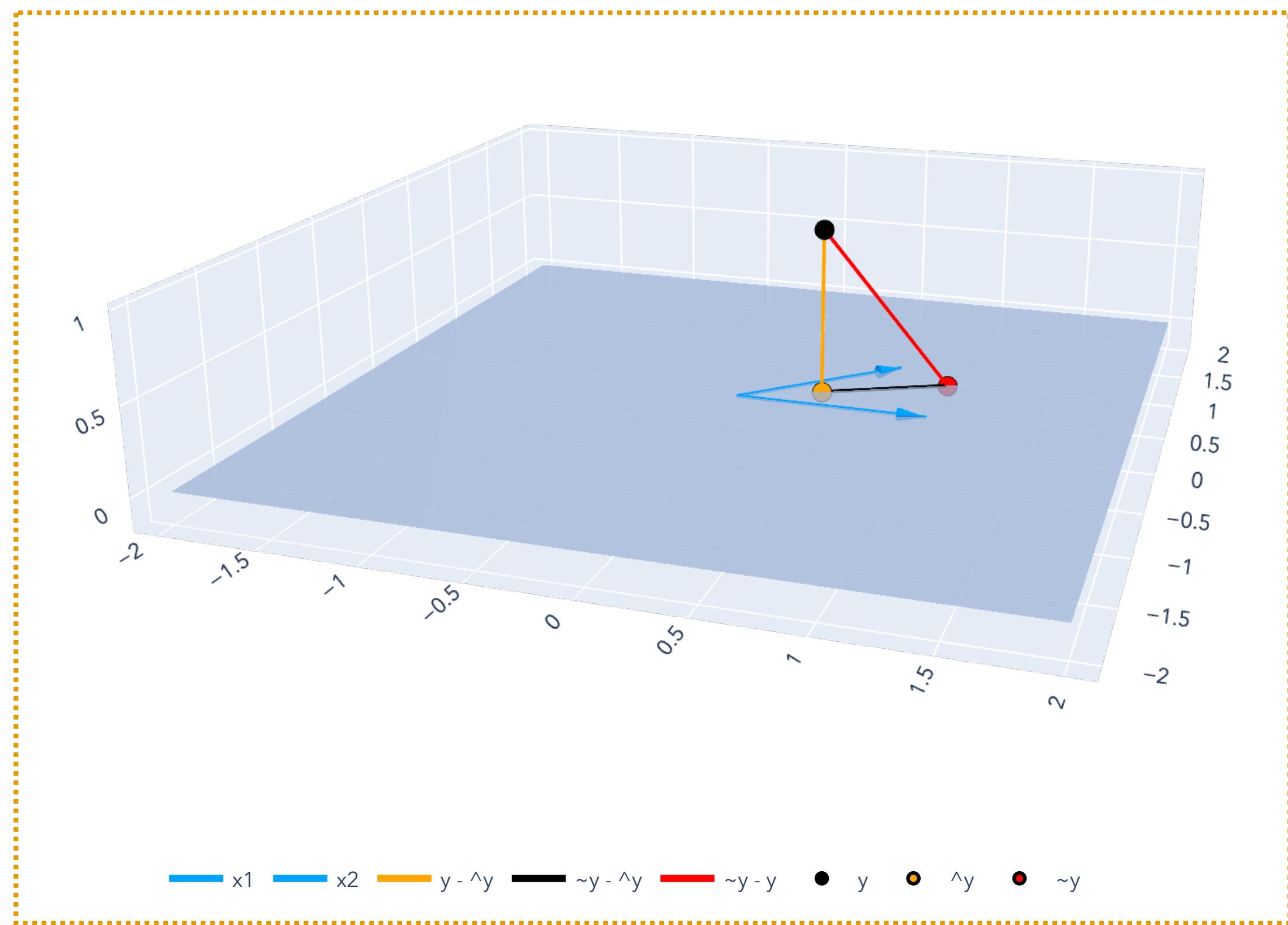
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \sum_{i=1}^n (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

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

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

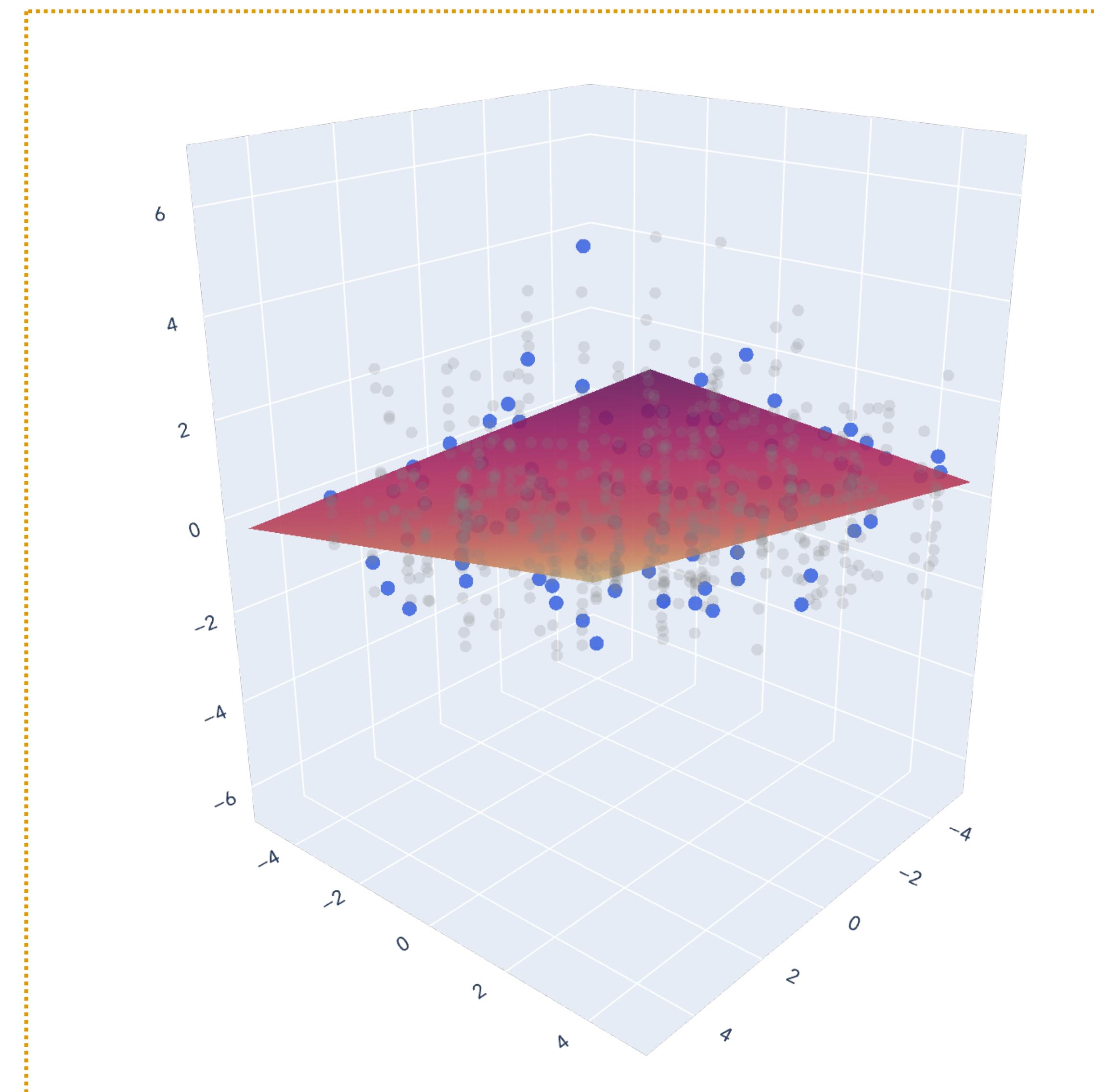
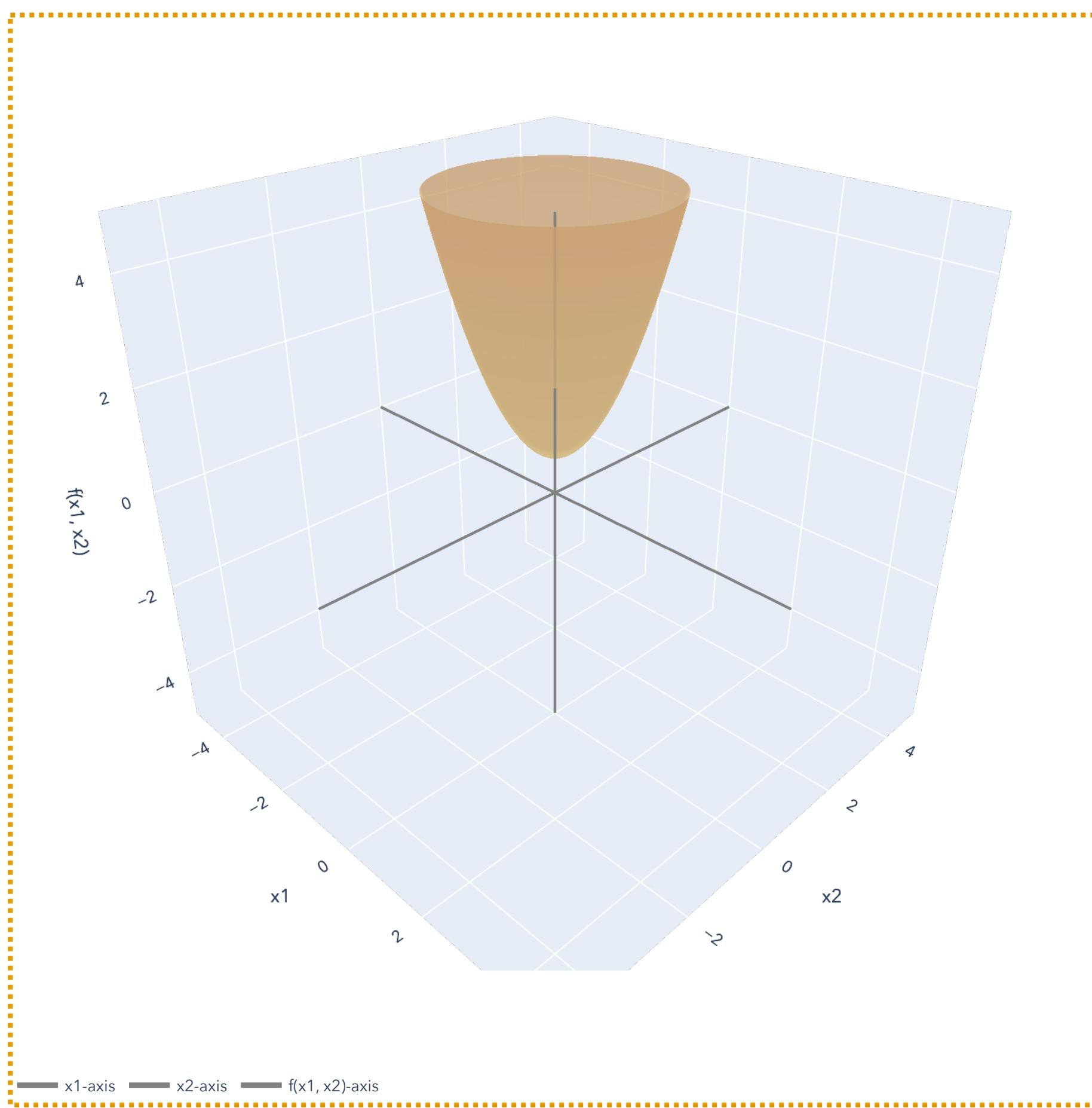
# Regression

A note on  $\hat{w}$



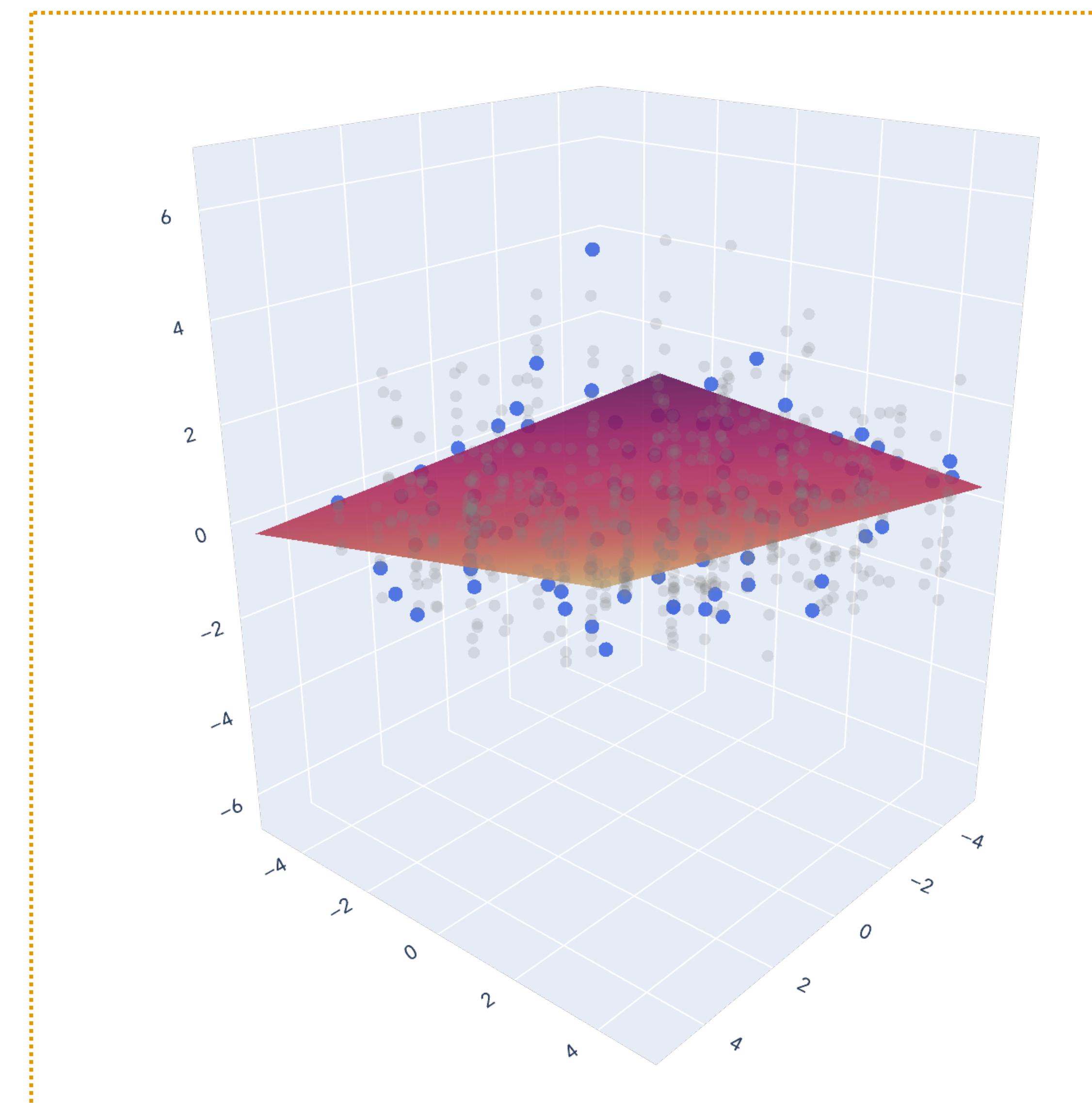
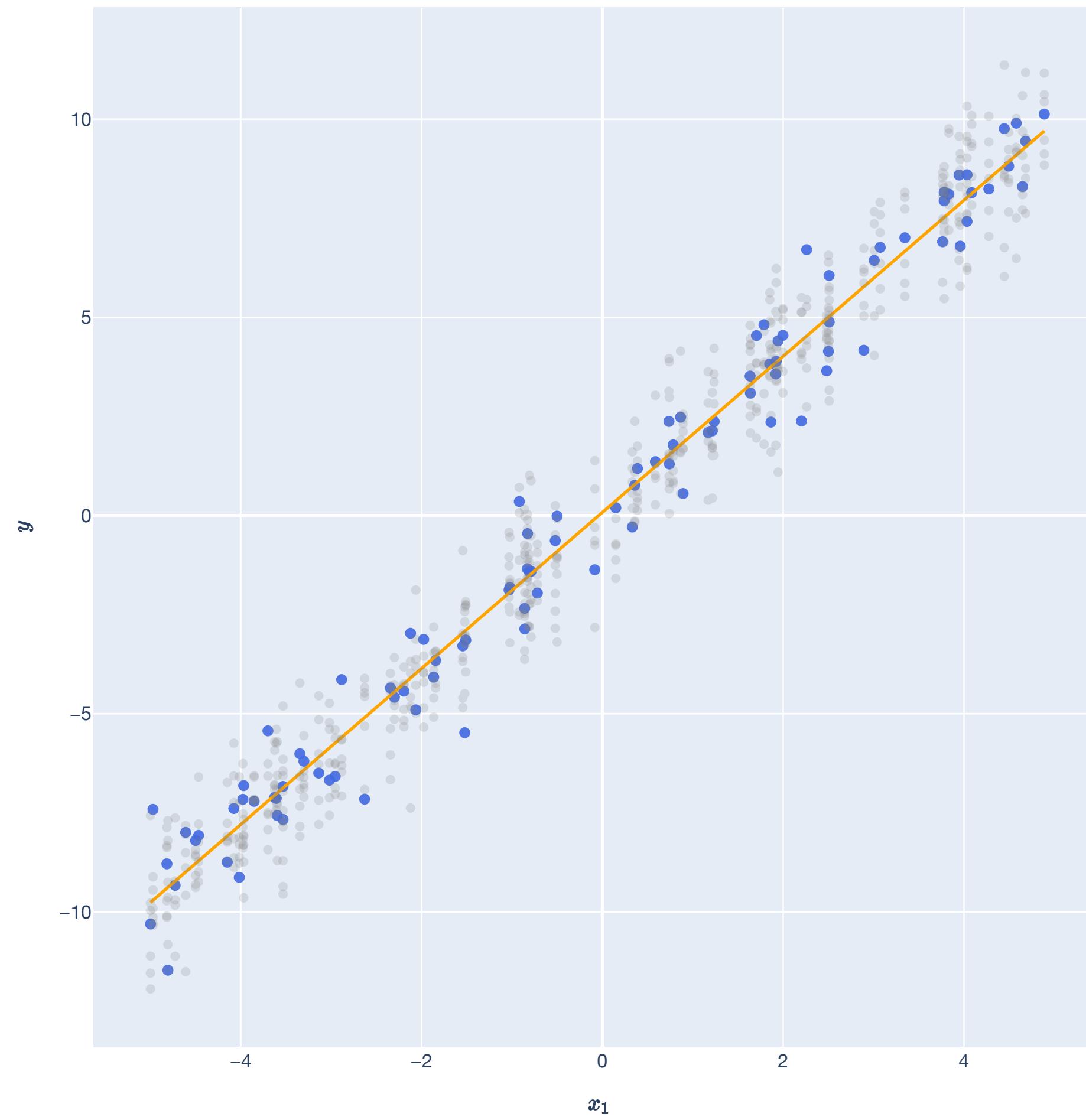
# Regression

A note on  $\hat{w}$



# Regression

A note on  $\hat{w}$



# Regression

## Setup, with randomness

Observed: 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  $\mathbf{x}_i \in \mathbb{R}^d$

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

Random variable  $y_i$

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

... 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  $w_1, \dots, w_d$ .

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

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      Random variable

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

... 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  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  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,  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  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

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

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

Some *deterministic* function  $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$  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

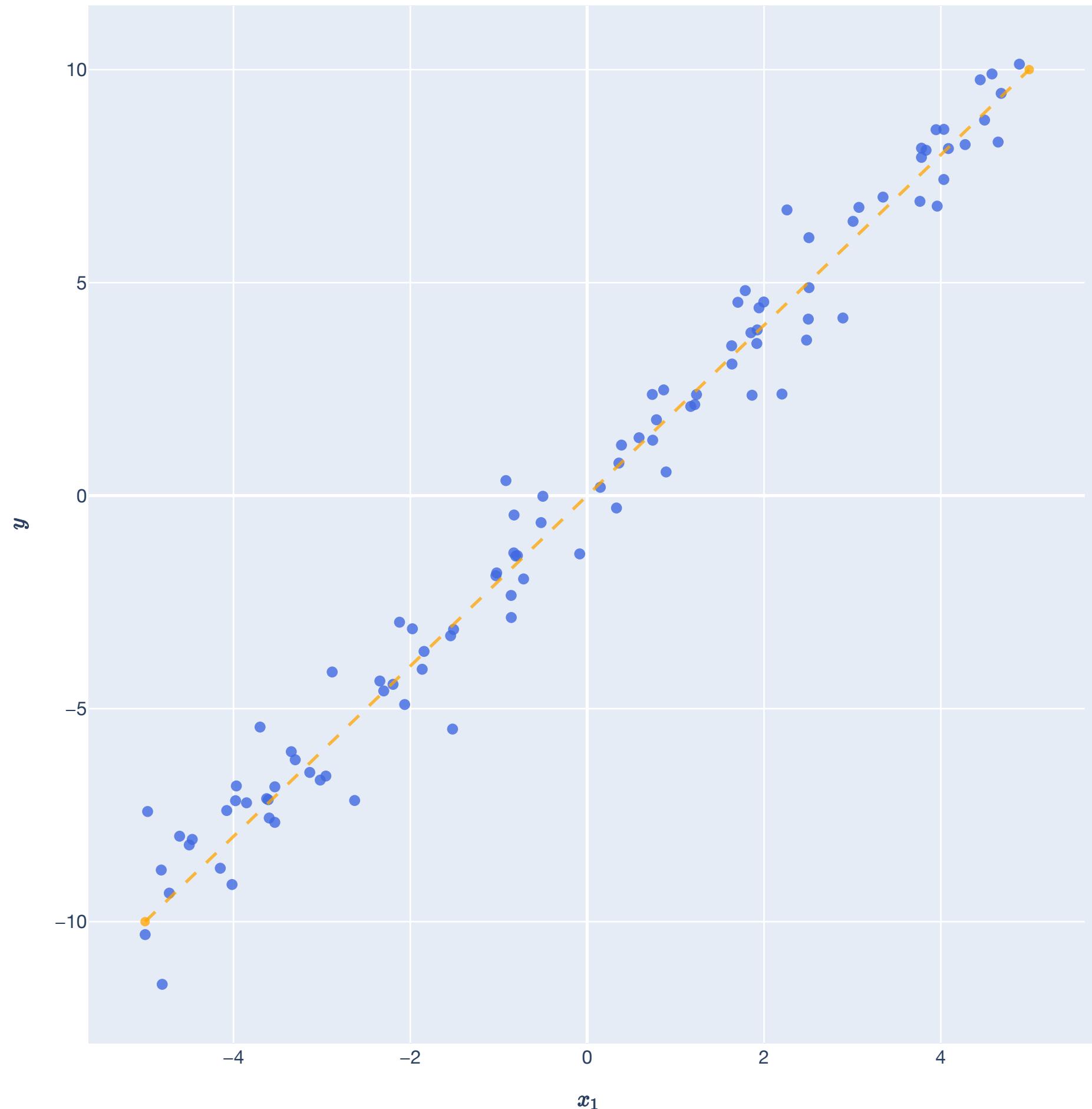
Assume further:  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$  pair is drawn from a joint distribution,  $\mathbb{P}_{\mathbf{x},y}$

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

Some *deterministic* function  $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$  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

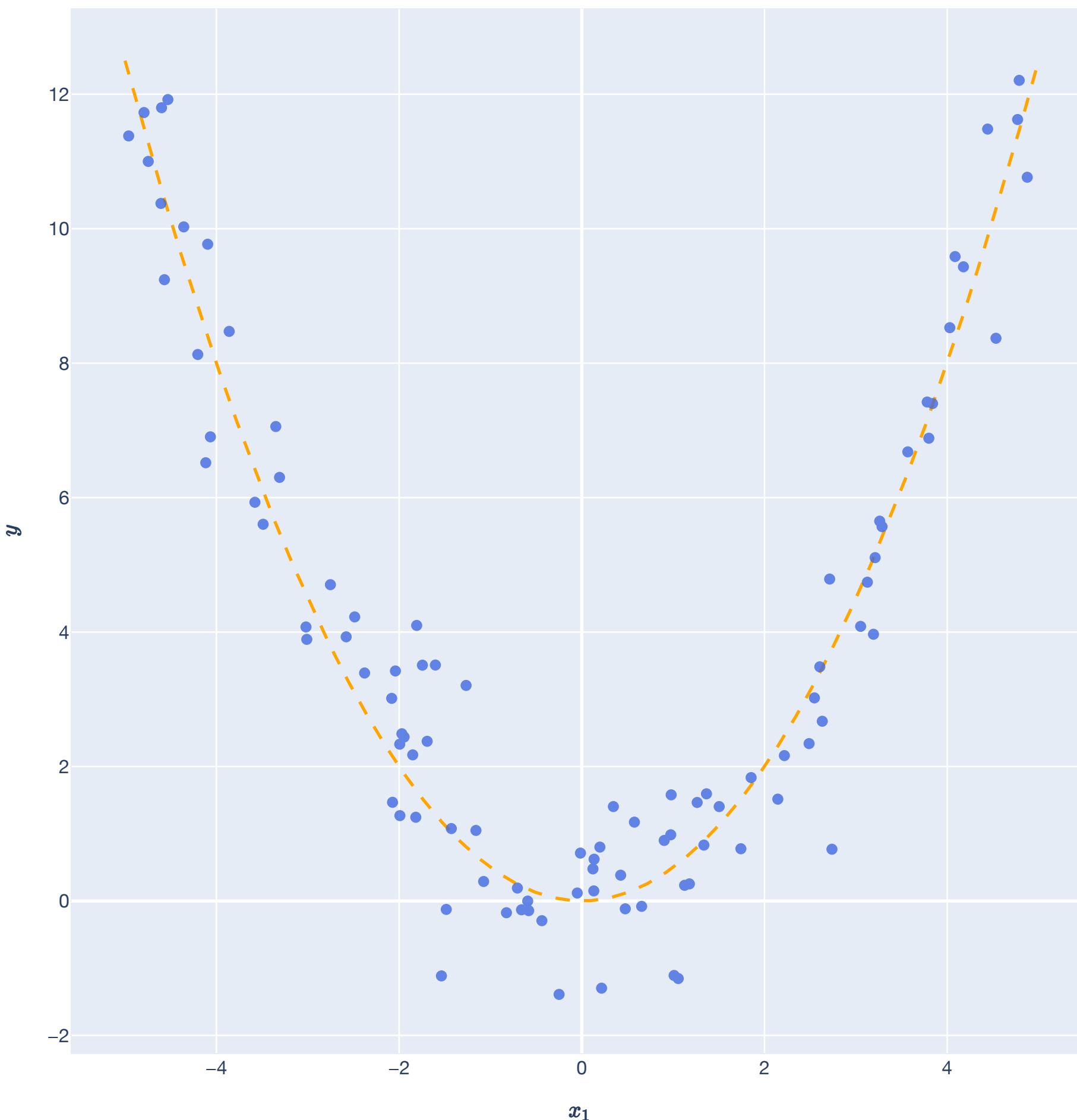
Assume further:  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$  pair is drawn from a joint distribution,  $\mathbb{P}_{\mathbf{x},y}$

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

Some deterministic function  $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$  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

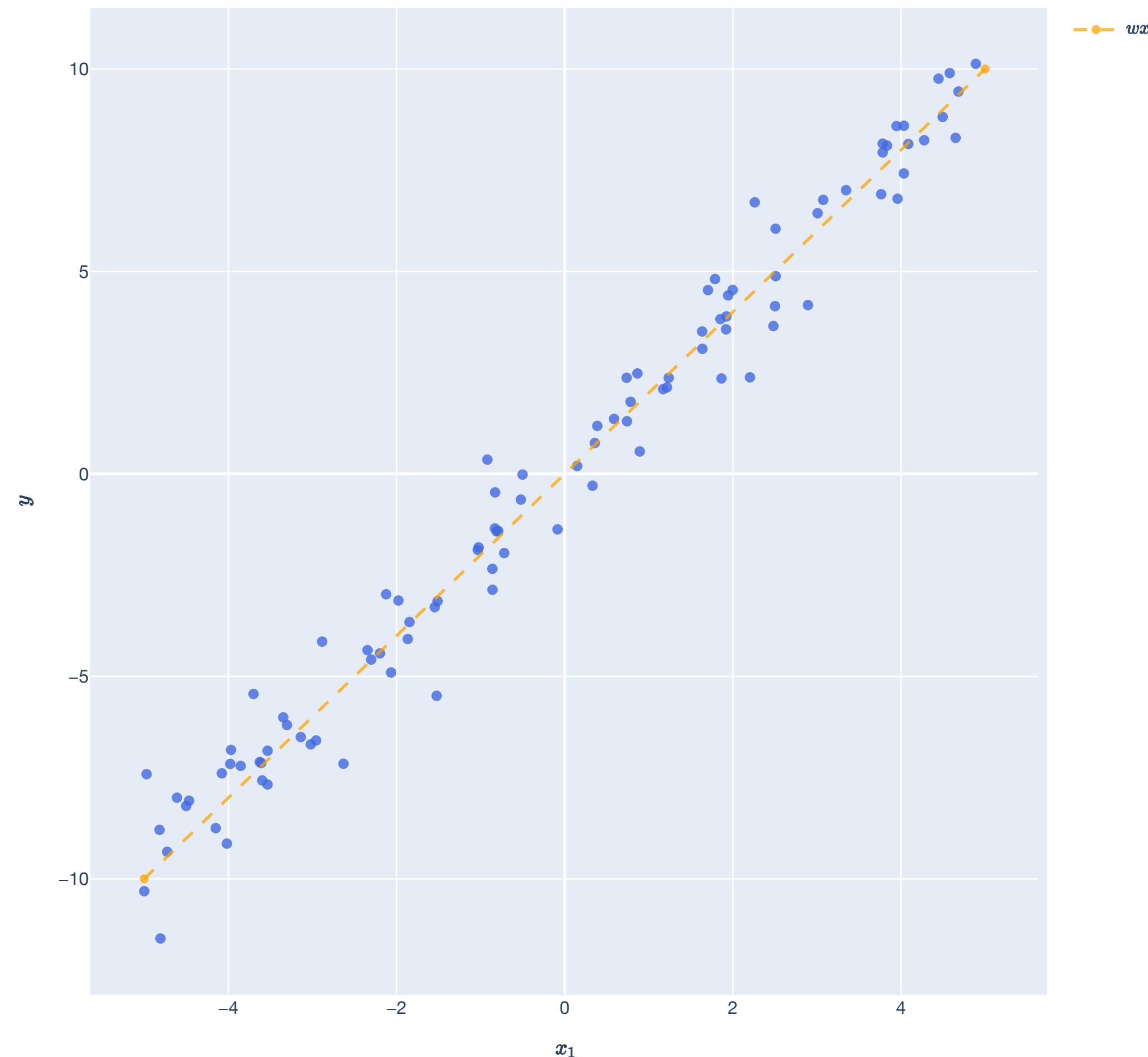
Assume further:  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$  pair is drawn from a joint distribution,  $\mathbb{P}_{\mathbf{x}, y}$

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

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

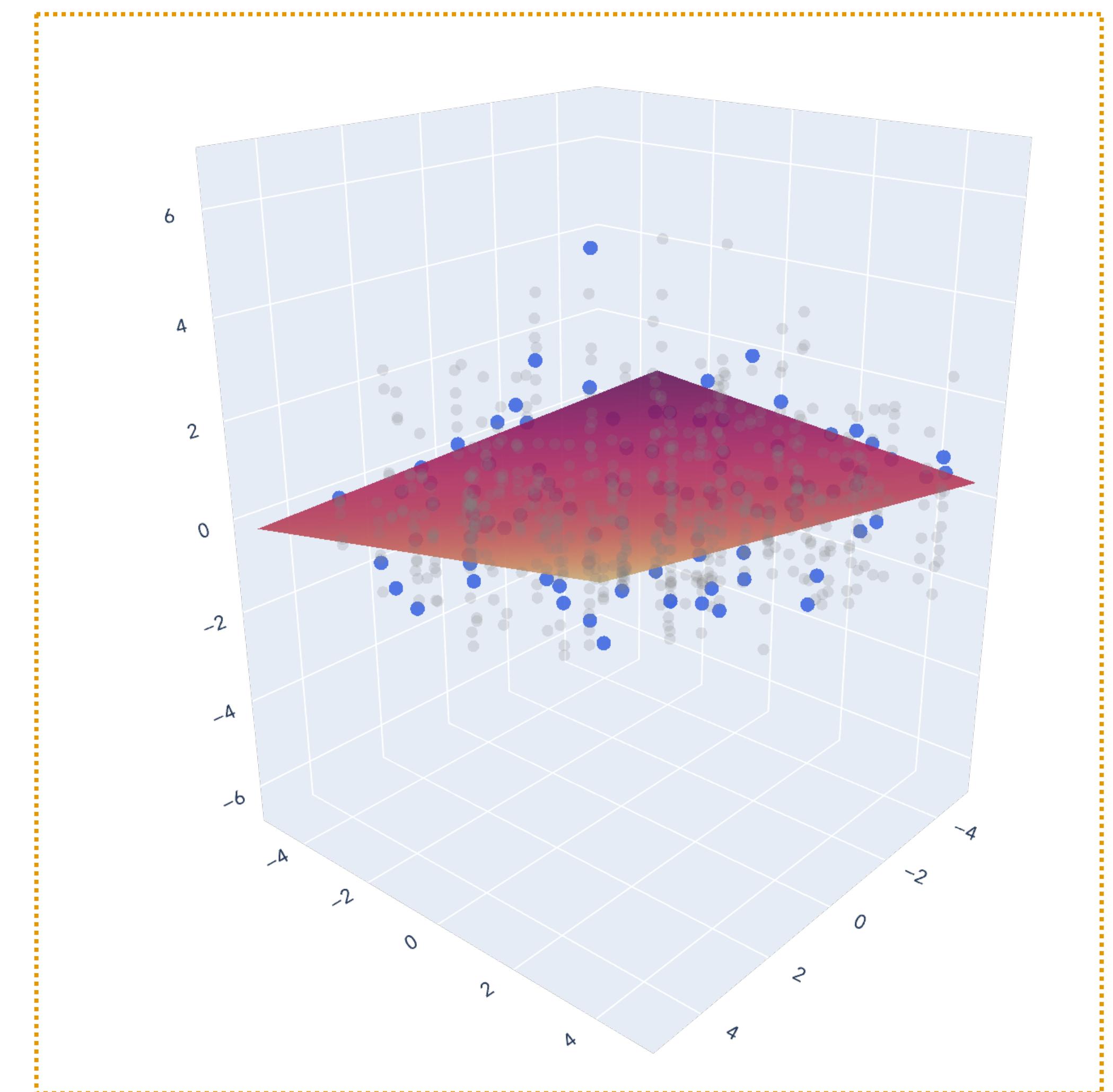
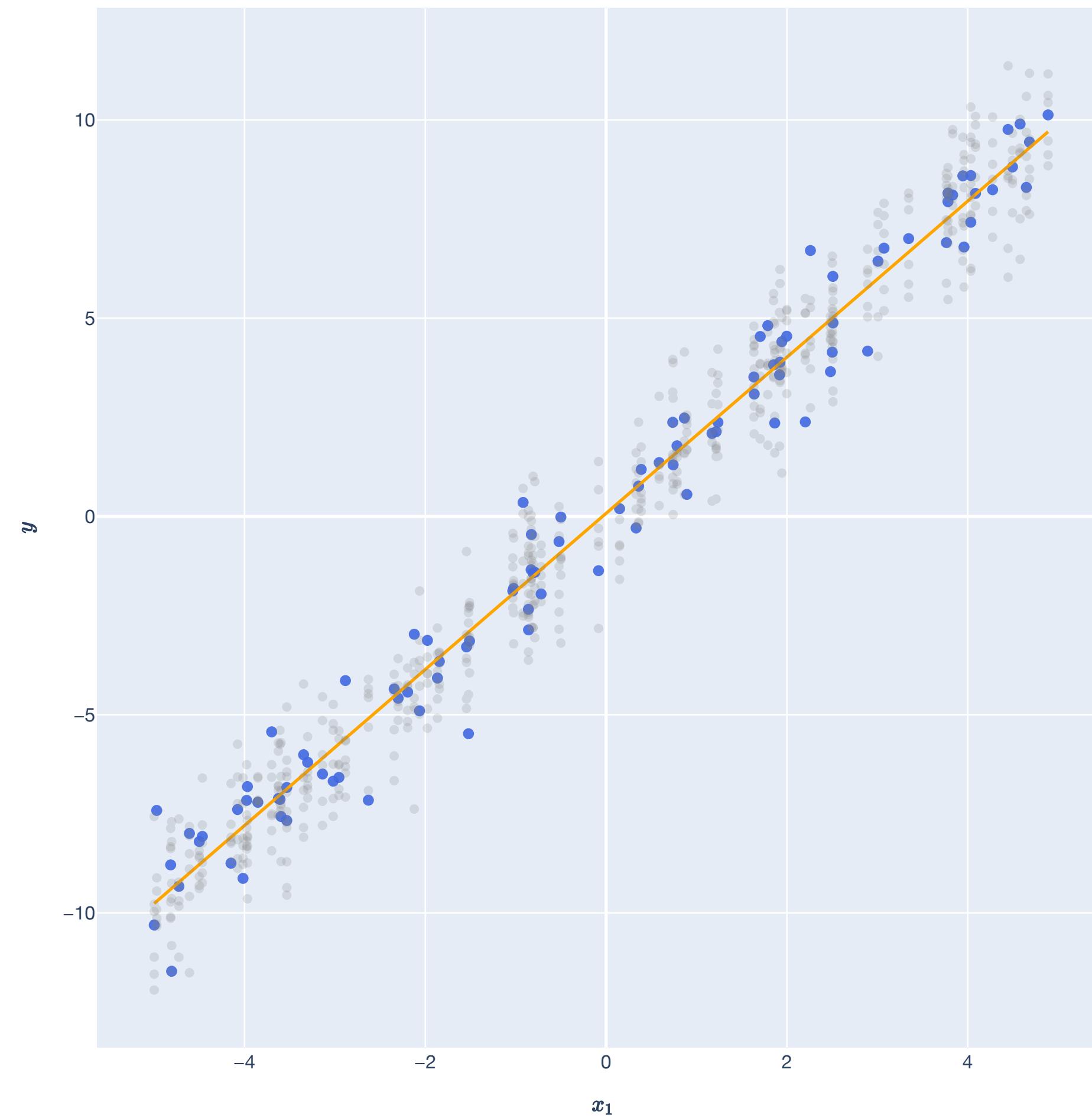
Some *randomness*  $\epsilon_i$  models the unexplained relationship, where we assume

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



# 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:

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

# Regression with randomness

## Part 2: Notion of error and linear model

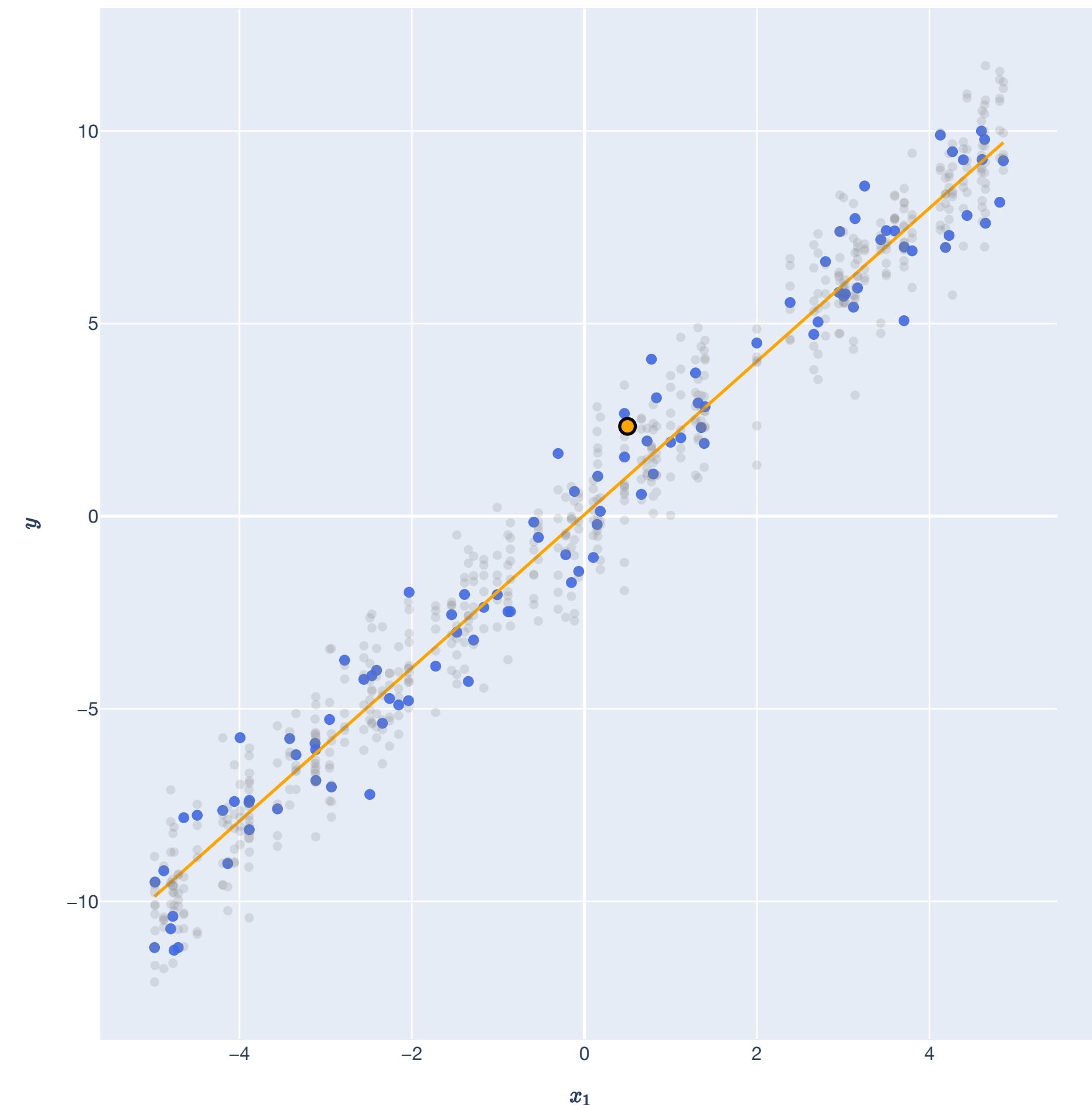
Each  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$  pair is drawn from a *joint distribution*,  $\mathbb{P}_{\mathbf{x},y}$

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

Our model  $\hat{f}: \mathbb{R}^d \rightarrow \mathbb{R}$  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  $\mathbf{X} = \begin{bmatrix} \leftarrow \mathbf{x}_1^\top \rightarrow \\ \vdots \\ \leftarrow \mathbf{x}_n^\top \rightarrow \end{bmatrix}$  Random variable  $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ , where  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ . ... 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  $w_1, \dots, w_d$ .

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

# 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  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  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

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

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

Intermediary goal: 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  $n \rightarrow \infty$ !

# Regression

## Part 3: Risk and empirical risk

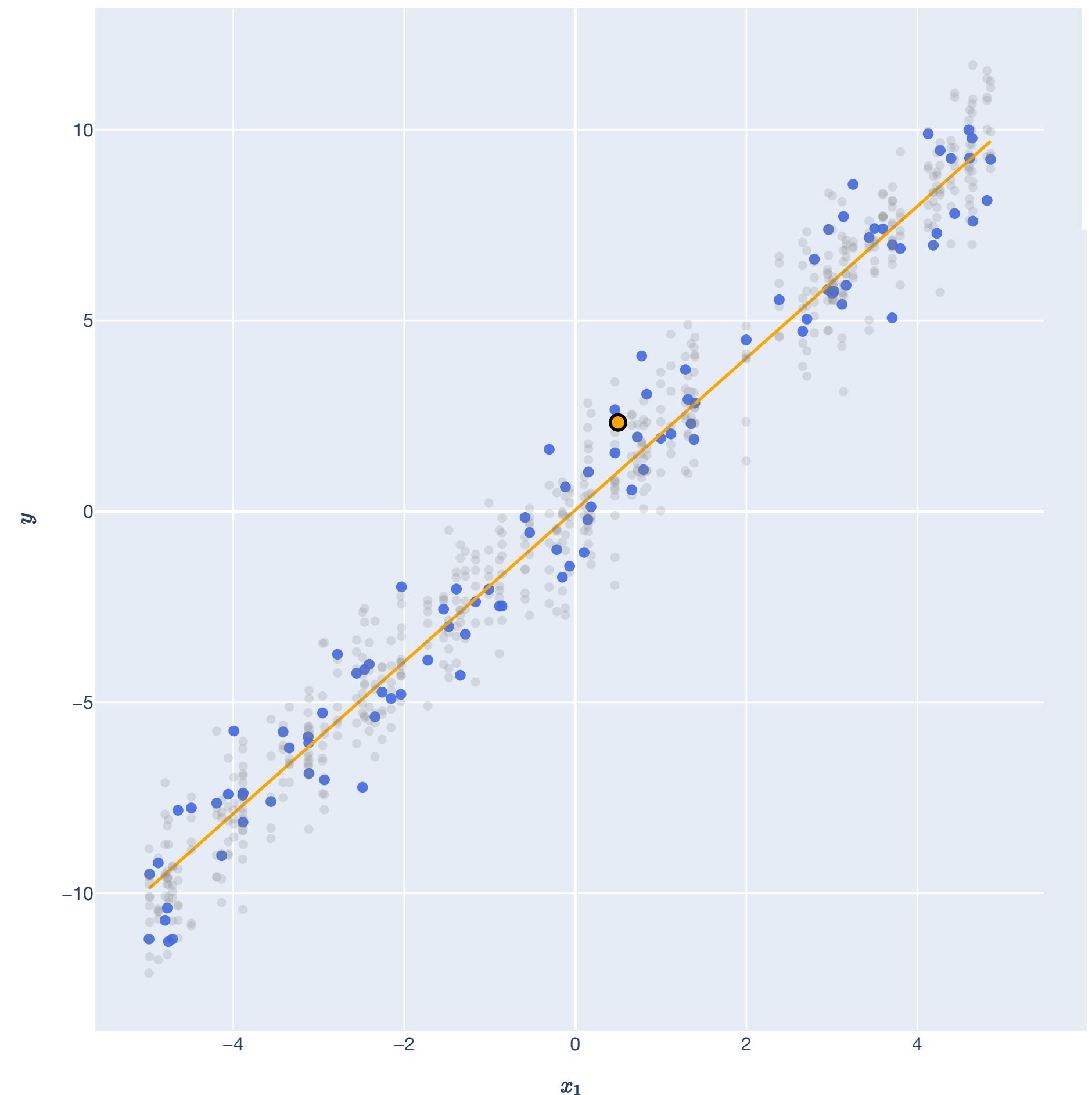
Ultimate goal: Find  $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$  that generalizes on a new  $(\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x}, y}$ :

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

Intermediary goal: 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  $n \rightarrow \infty!$



# Regression

## Part 3: Risk and empirical risk

Observed: 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  $\mathbf{x}_i$

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

Random variable  $y_i$

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

... 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  $w_1, \dots, w_d$ .

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

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:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}. \quad \text{Because we assume that } (\mathbf{x}_0, y_0) \sim \mathbb{P}_{\mathbf{x},y} \text{ as well!}$$

# 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  $\text{Var}(\epsilon) = \sigma^2$ , with  $\epsilon$  independent of  $\mathbf{x}$ .

# Statistics of the Error Model

## Conditional Expectation

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

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

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

$\mathbb{E}[y | \mathbf{x}] = \mathbf{x}^\top \mathbf{w}^*$ , 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:  $\mathbb{E}[\hat{\mathbf{w}} | \mathbf{X}] = \mathbf{w}^*$ .

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

# Statistics of OLS

## Intuition

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

Obtain the least squares estimator the same way:

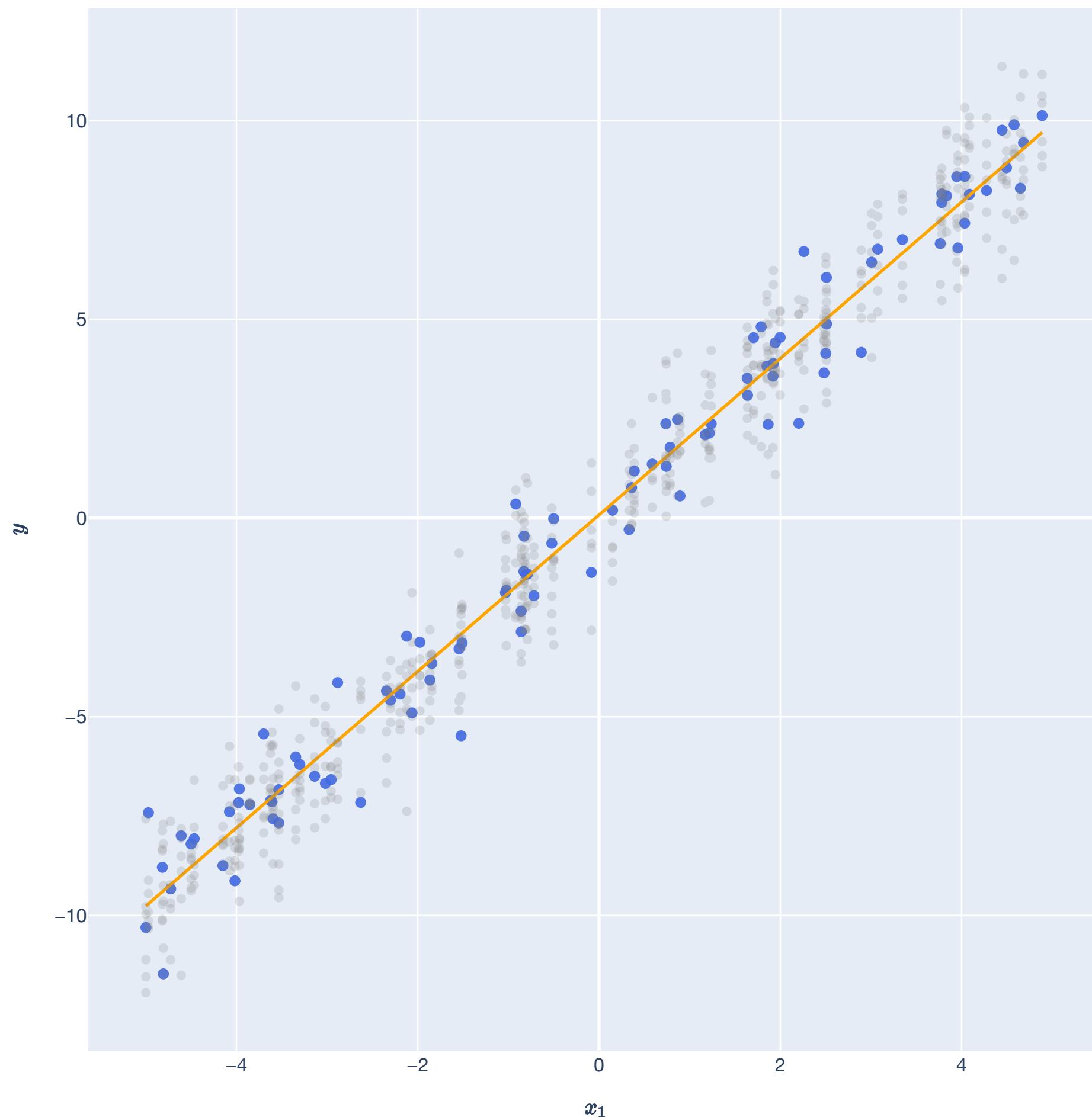
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

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

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



# Statistics of OLS

## Intuition

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

Obtain the least squares estimator the same way:

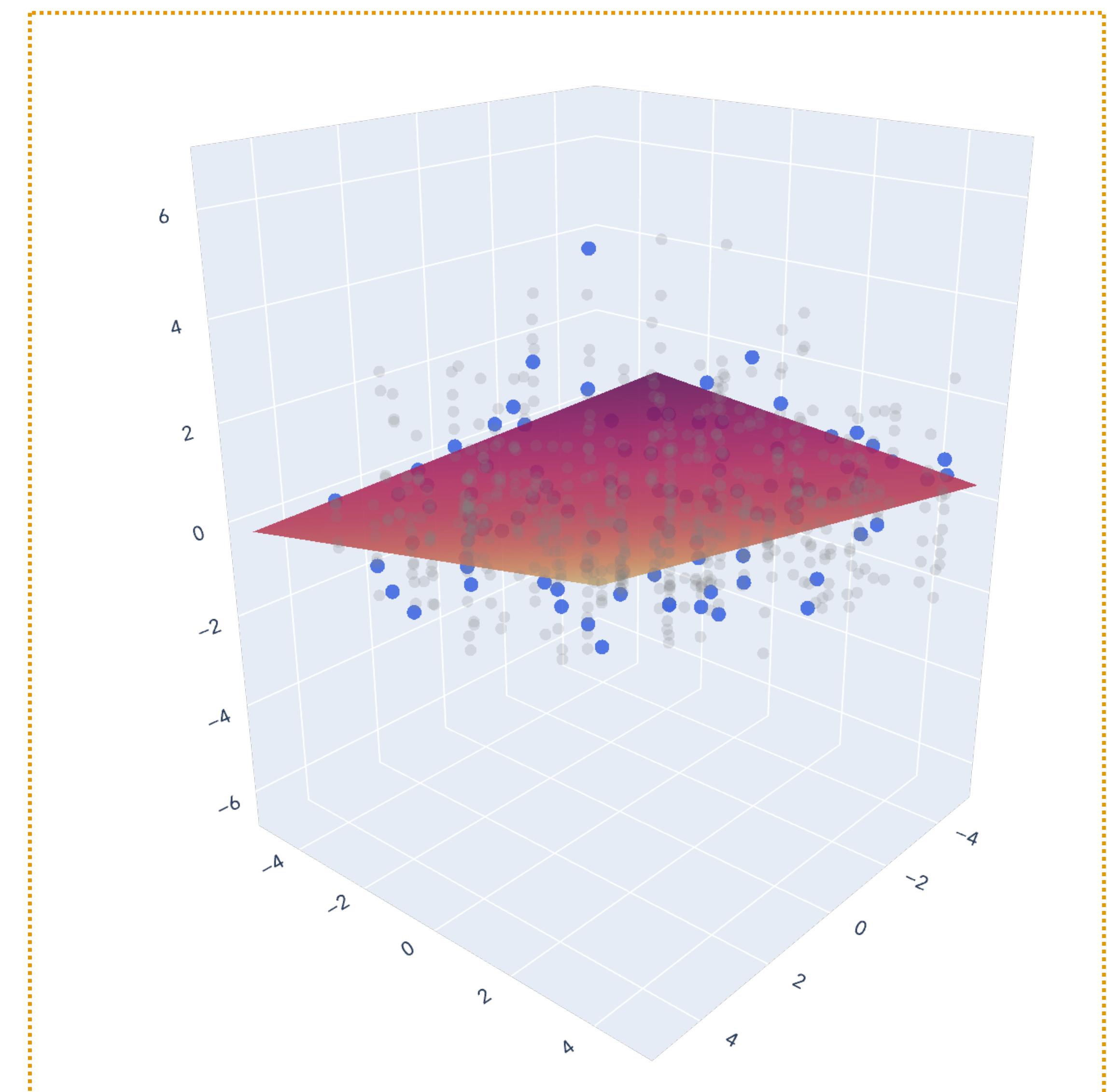
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

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

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



# 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  $\text{Var}(\epsilon) = \sigma^2$ , 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:  $\mathbb{E}[\hat{\mathbf{w}} | \mathbf{X}] = \mathbf{w}^*$ .

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

# 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.

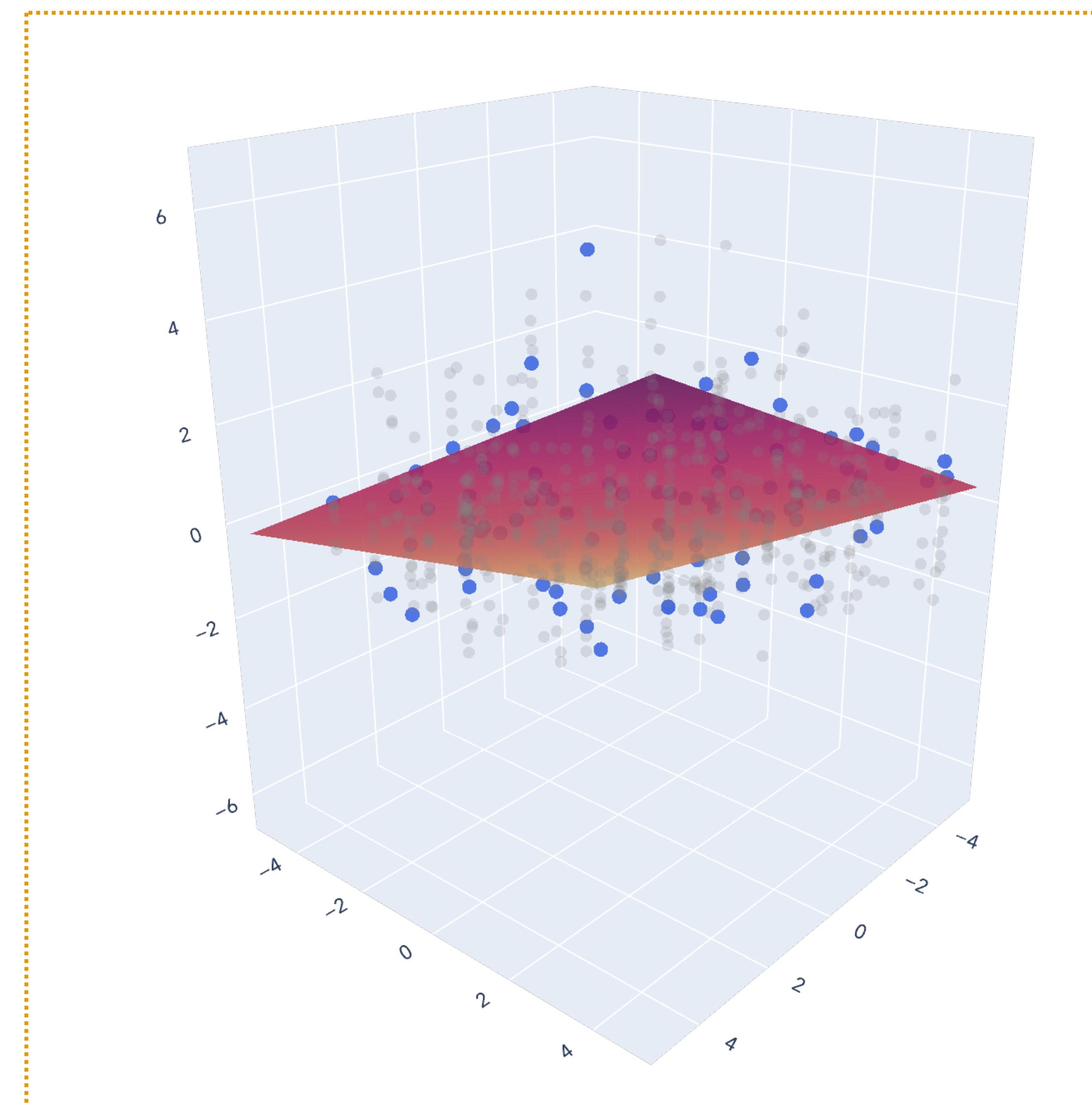
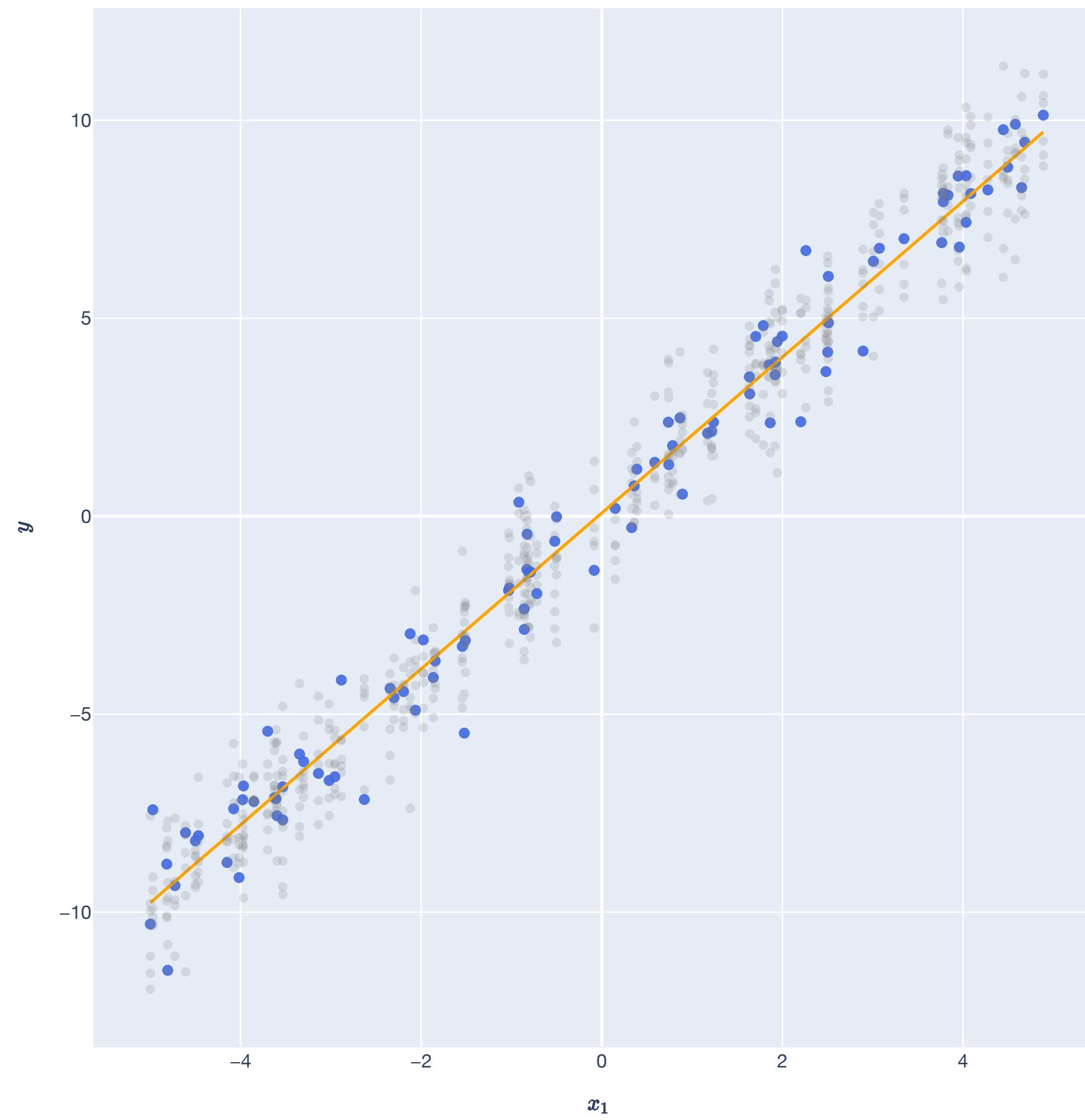
**Expectation, variance, and covariance.** Review of these basic summary statistics of random variables and common properties.

**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 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

