Math for ML

Week 5.2: Bias, Variance, and Statistical Estimators

By: Samuel Deng

Logistics & Announcements

- PS4 is out, due next Tues. 11:59 PM.
- PS5 released tomorrow (LAST PROBLEM SET! 2 problems)
- Paper Reading Project; Released tomorrow. (1-2 pages)

* COURSE EVALUATIONS!

Lesson Overview

Law of Large Numbers. The LLN allows us to move from probability to statistics (reasoning about an *unknown* data generating process using data from that process).

Statistical estimators. We define a *statistical estimator*, which is a function of a collection of random variables (data) aimed at giving a "best guess" at some unknown quantity from some probability distribution.

Bias, variance, and MSE. Two important properties of statistical estimators are their *bias* and *variance*, which are measures of how good the estimator is at guessing the target. These form the estimator's MSE.

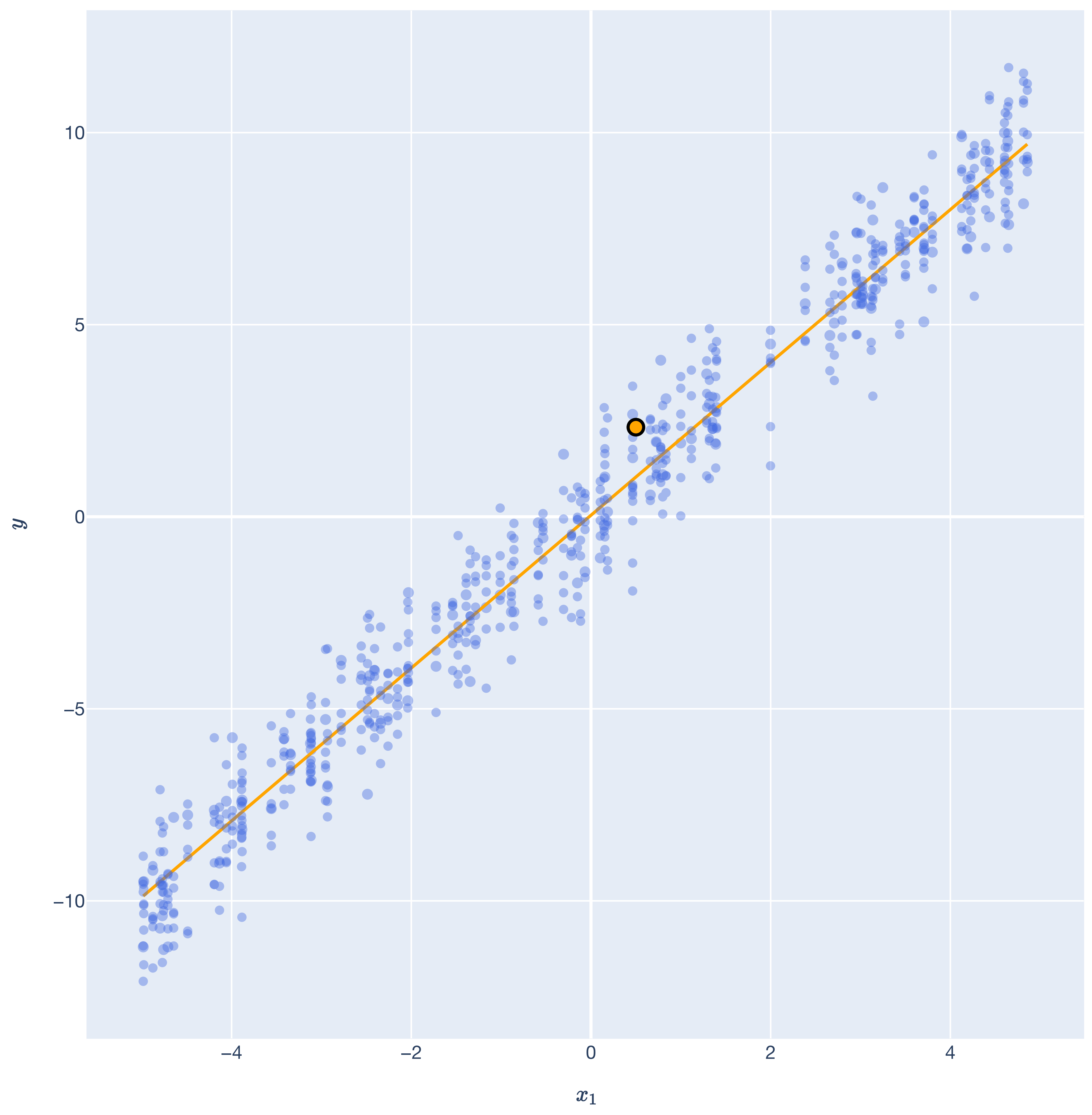
Stochastic gradient descent (SGD). Gradient descent needs to take a gradient over all n training examples, which may be large; SGD *estimates* the gradient to speed up the process.

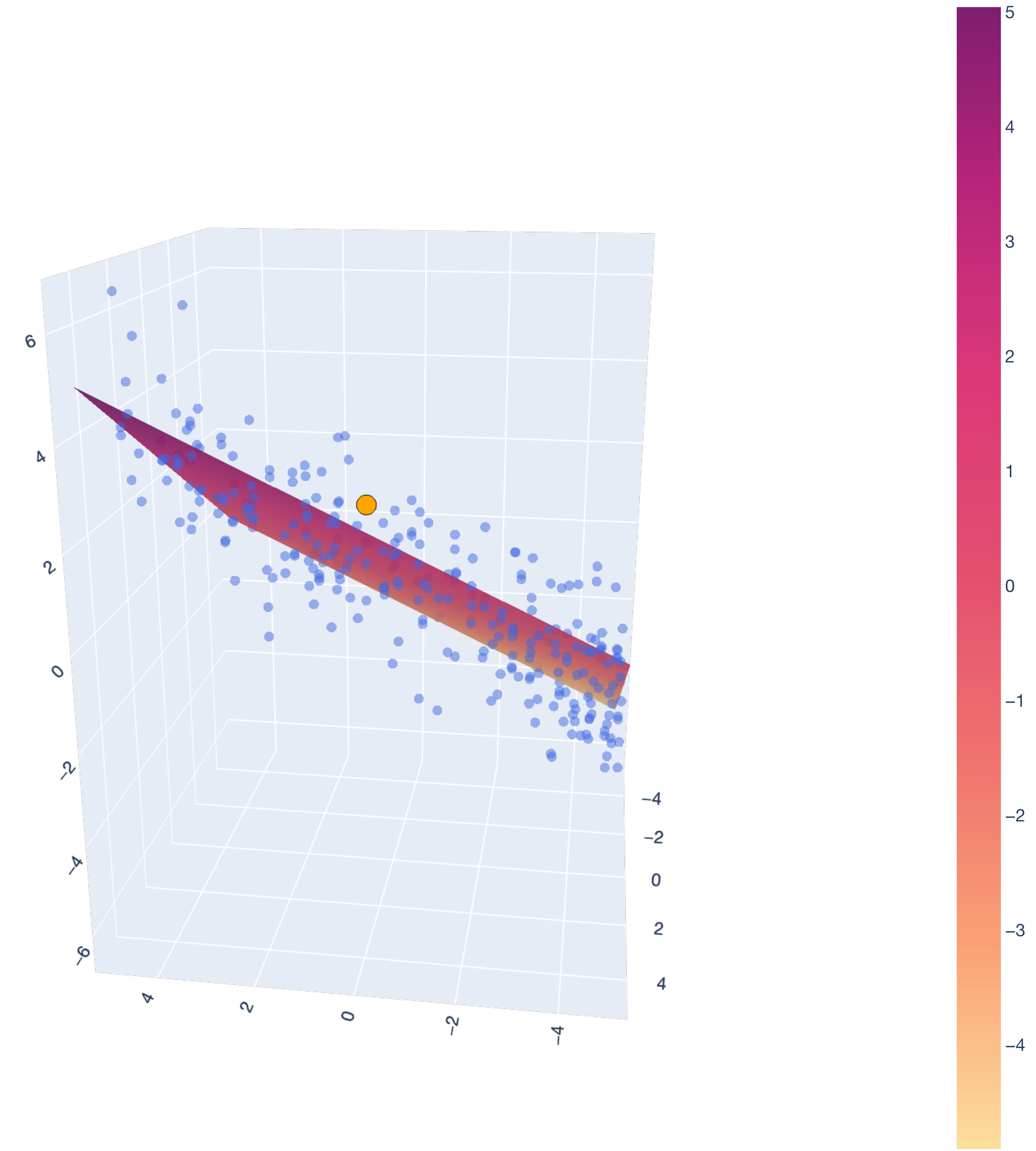
Gauss-Markov Theorem. We show that OLS is the minimum variance estimator in the class of all unbiased, linear estimators.

Statistical analysis of OLS risk. We analyze the *risk* of OLS — how well it's expected to do on future examples drawn from the same distribution it was trained on.

Lesson Overview

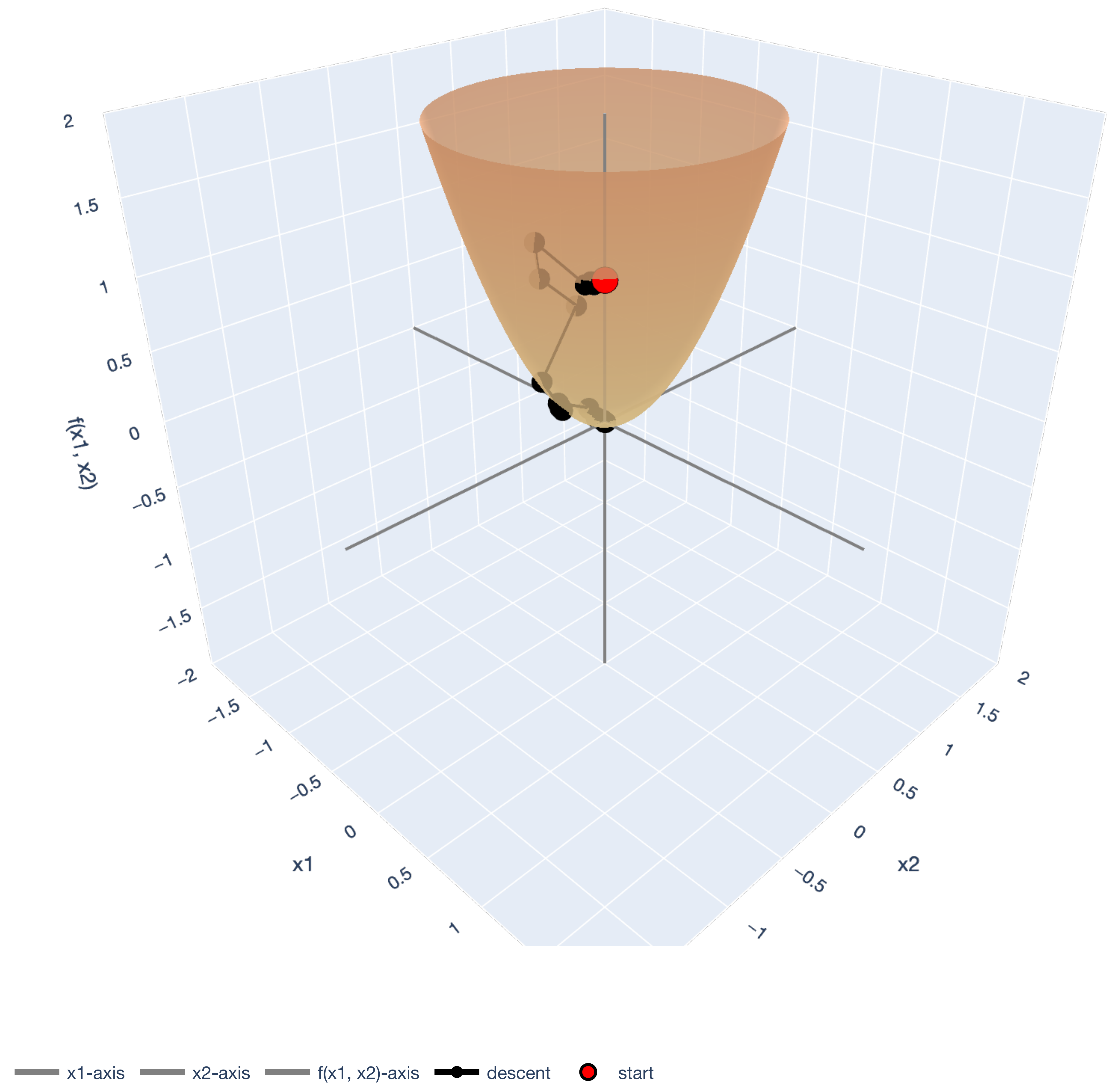
Big Picture: Least Squares

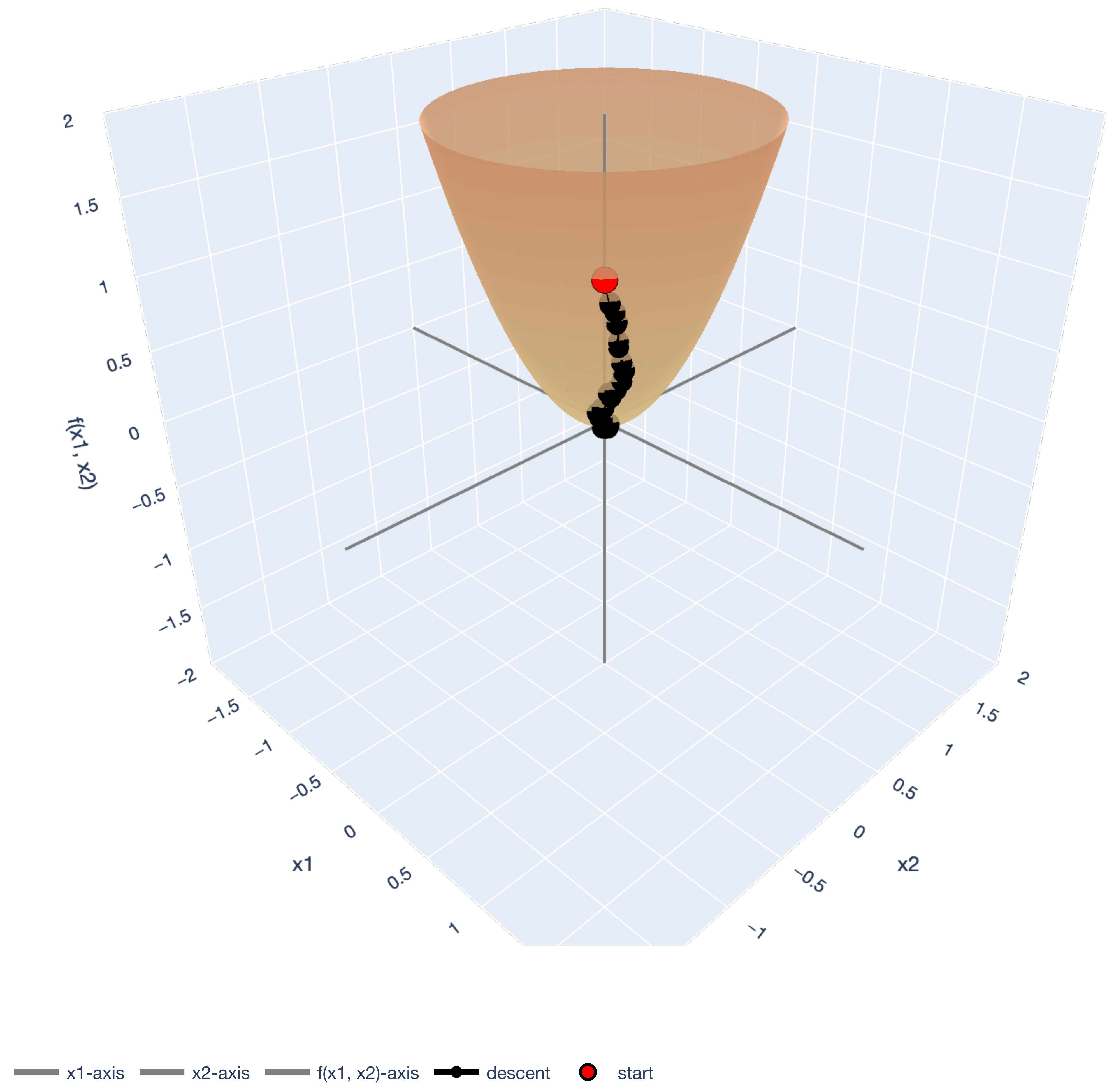




Lesson Overview

Big Picture: Gradient Descent





Law of Large Numbers

+ CLT

Theorem and Statistical Estimation 101

Statistical Estimation

Intuition

In ***probability theory***, we assumed we knew some data generating process (as a *distribution*) $\mathbb{P}_{\mathbf{x}}$, and we analyzed observed data under that process.

$$\mathbb{P}_{\mathbf{x}} \implies \mathbf{x}_1, \ldots, \mathbf{x}_n.$$

Statistics can be thought of as the "reverse process." We see some data and we try to make inferences about the process that generated the data.

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \implies \mathbb{P}_{\mathbf{x}}$$

Statistical Estimation

Intuition

In ***probability theory***, we assumed we knew some data generating process (as a *distribution*) $\mathbb{P}_{\mathbf{x}}$, and we analyzed observed data under that process.

$$\mathbb{P}_{\mathbf{x}} \implies \mathbf{x}_1, \ldots, \mathbf{x}_n.$$

Statistics can be thought of as the "reverse process." We see some data and we try to make inferences about the process that generated the data.

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \implies \mathbb{P}_{\mathbf{x}}$$

In order to do so, we need to formalize the notion that "collecting a lot of data" gives us a peek at the underlying process!

Law of Large Numbers

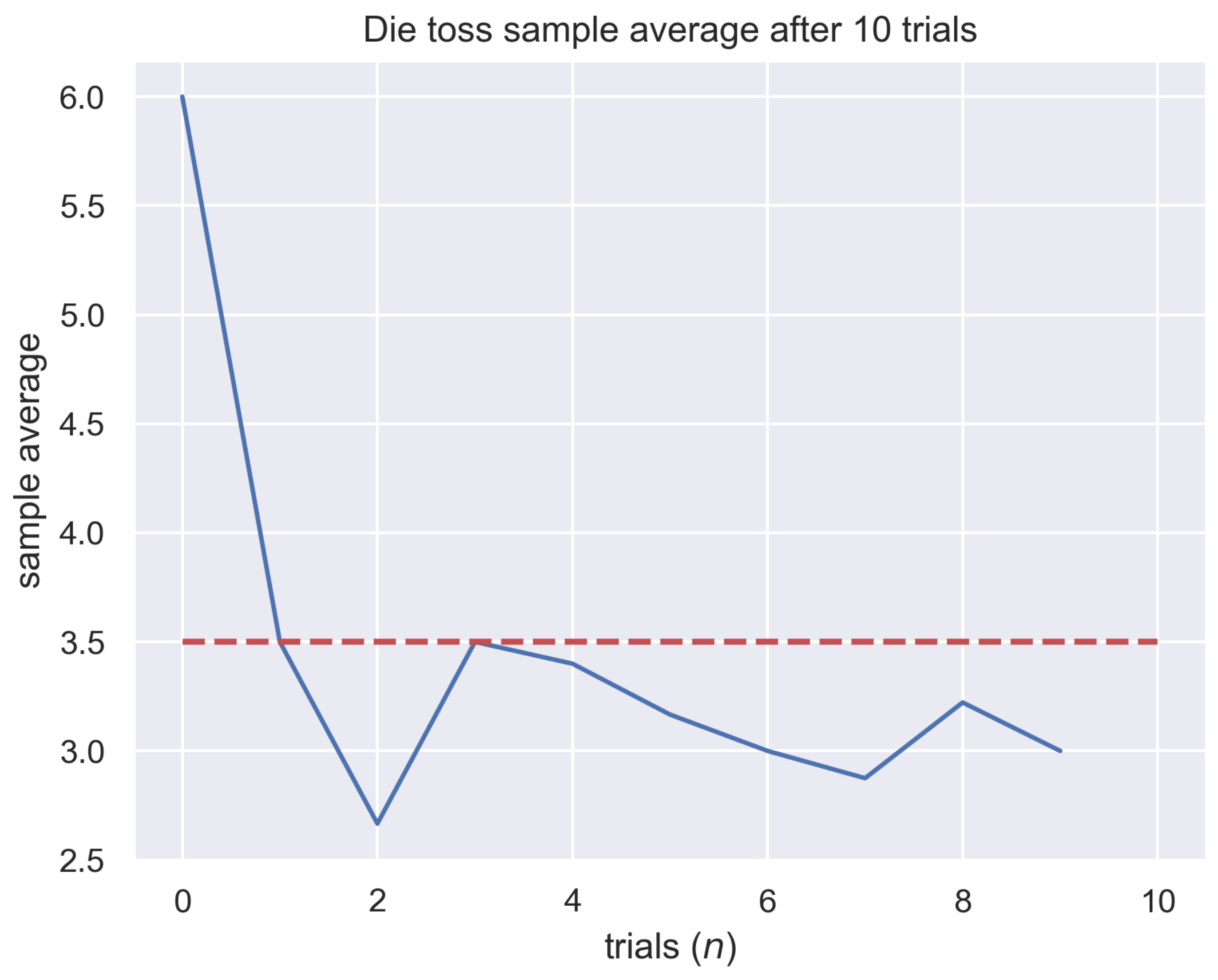
Intuition

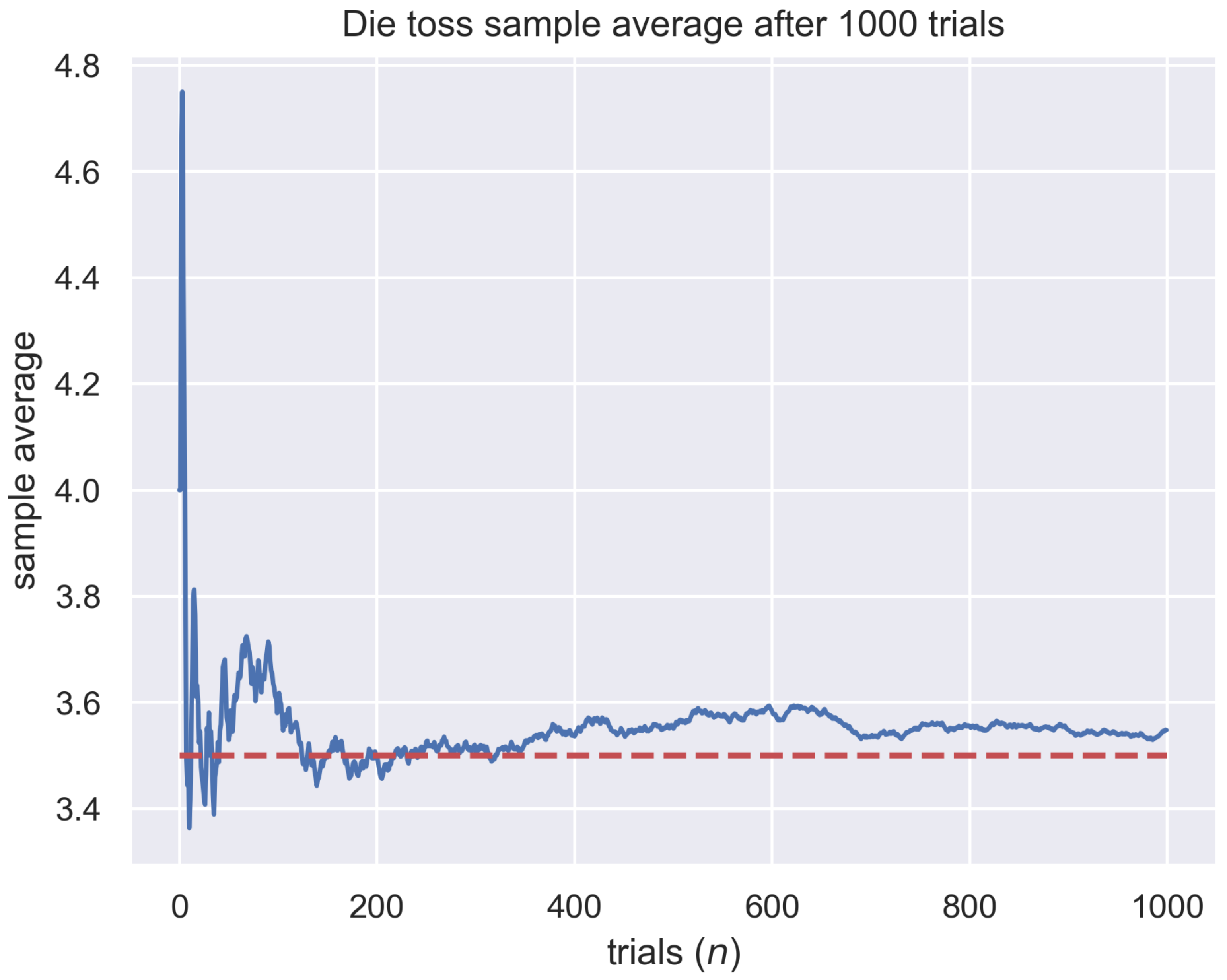
X is a RV for a 6-sided die

$\Rightarrow E[X] = 1/6 \cdot 1 + 1/6 \cdot 2 + \dots + 1/6 \cdot 6 = \boxed{3.5}$.

Averages of a *large* number of random samples converge to their mean. $\boxed{\frac{1}{n}\sum_{i=1}^{n} x_i}$

Example. The average die roll after many trials is expected to be close to 3.5.





Independence

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

$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots$

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:

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

eg $X_1, X_2, X_3, \ldots$ ← outcomes of a coin flip

$\mathbb{P}[X_i = 0] = \mathbb{P}[X_i = 1] = 1/2$

Very common assumption in ML!

Law of Large Numbers

Theorem Statement

Theorem (Weak Law of Large Numbers). Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with finite mean $\mu := \mathbb{E}[X_i]$. Let their *sample average* be denoted as

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \epsilon\right) = 1.$$

0.01

$\overline{X}_n$ is close to the mean.

Sample Avg. close to the mean.

0.01

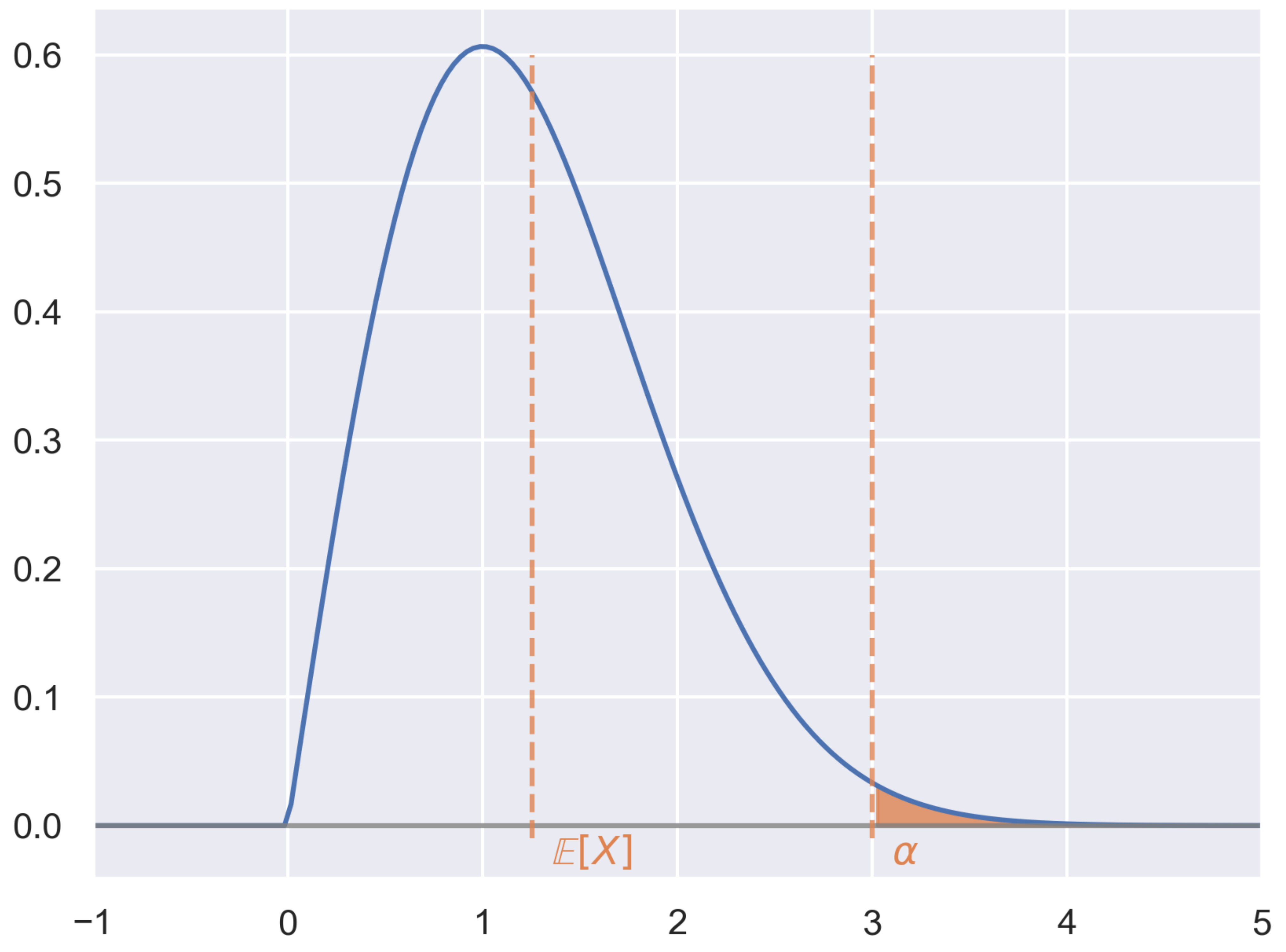
This type of convergence is also called ***convergence in probability***.

Markov's Inequality

Statement and Proof

Theorem (Markov's Inequality). Let X be any nonnegative random variable and suppose that $\mathbb{E}[X]$ exists. For any $\alpha > 0$,

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$



Markov's Inequality

Statement and Proof

$\mathbb{E}[X] \geq \alpha \, \mathbb{P}[X > \alpha]$

For any $\alpha > 0$,

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof.

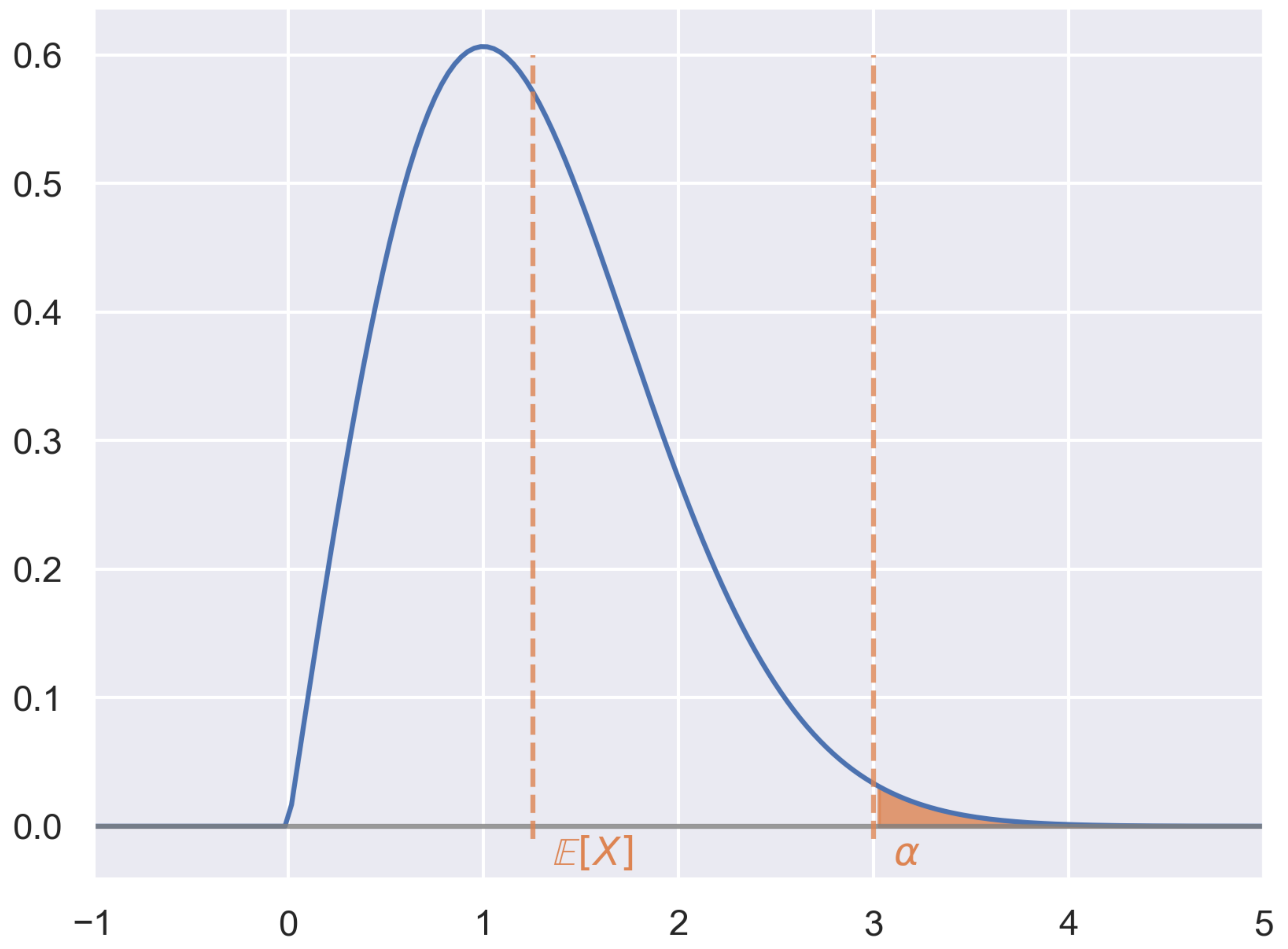
Because $X > 0$,

$$\mathbb{E}(X) = \int_0^\infty x p_X(x) dx = \int_0^\alpha x p_X(x) dx + \int_\alpha^\infty x p_X(x) dx$$

≥ 0

$$\geq \int_\alpha^\infty x p_X(x) dx \geq \alpha \int_\alpha^\infty p_X(x) dx = \alpha \mathbb{P}(X > \alpha)$$

$x \geq \alpha$

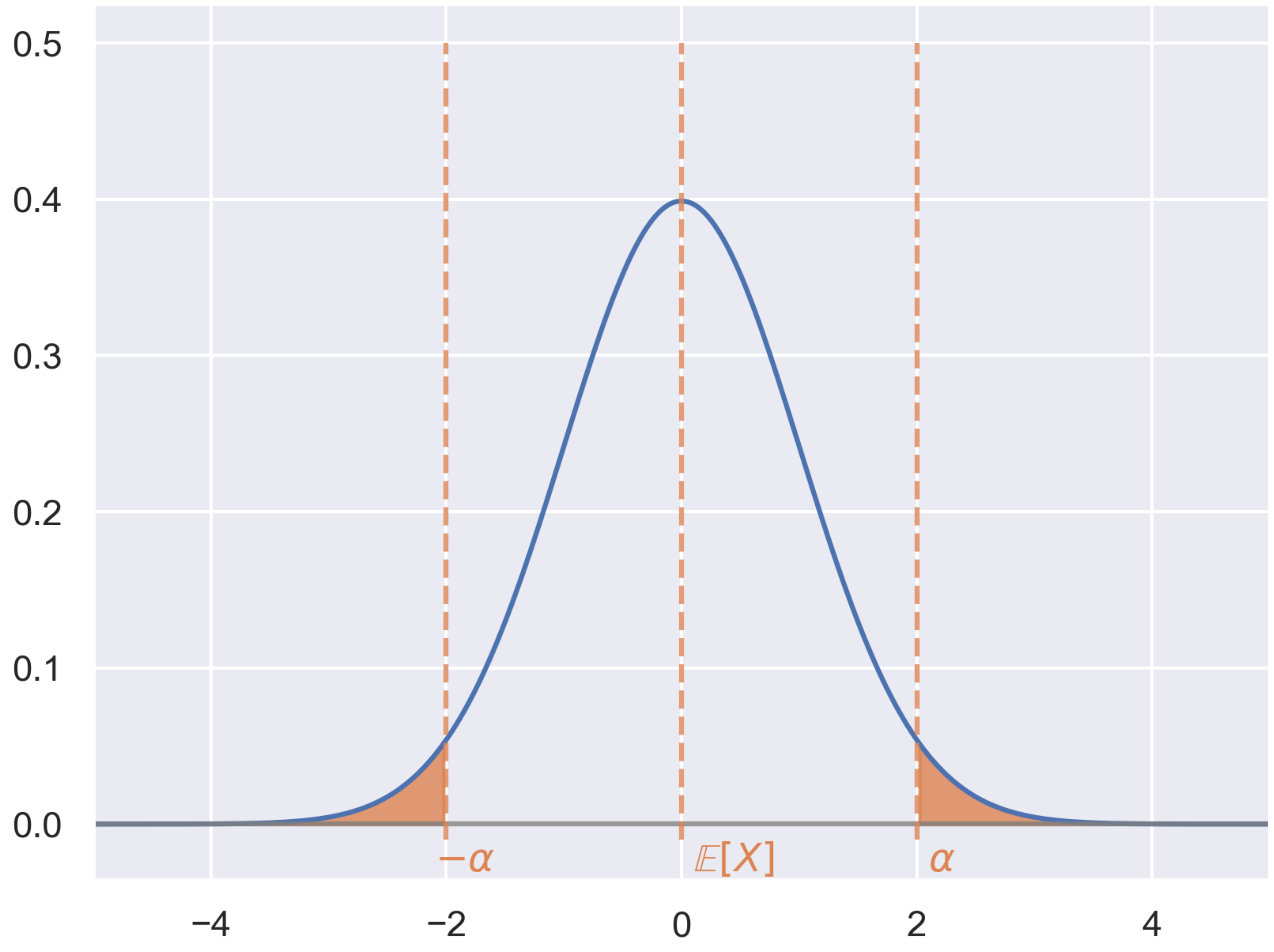


Chebyshev's Inequality

Statement and Proof

Theorem (Chebyshev's Inequality).
Let X be any arbitrary random variable, and let $\mu := \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X)$. Then,

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}.$$



Chebyshev's Inequality

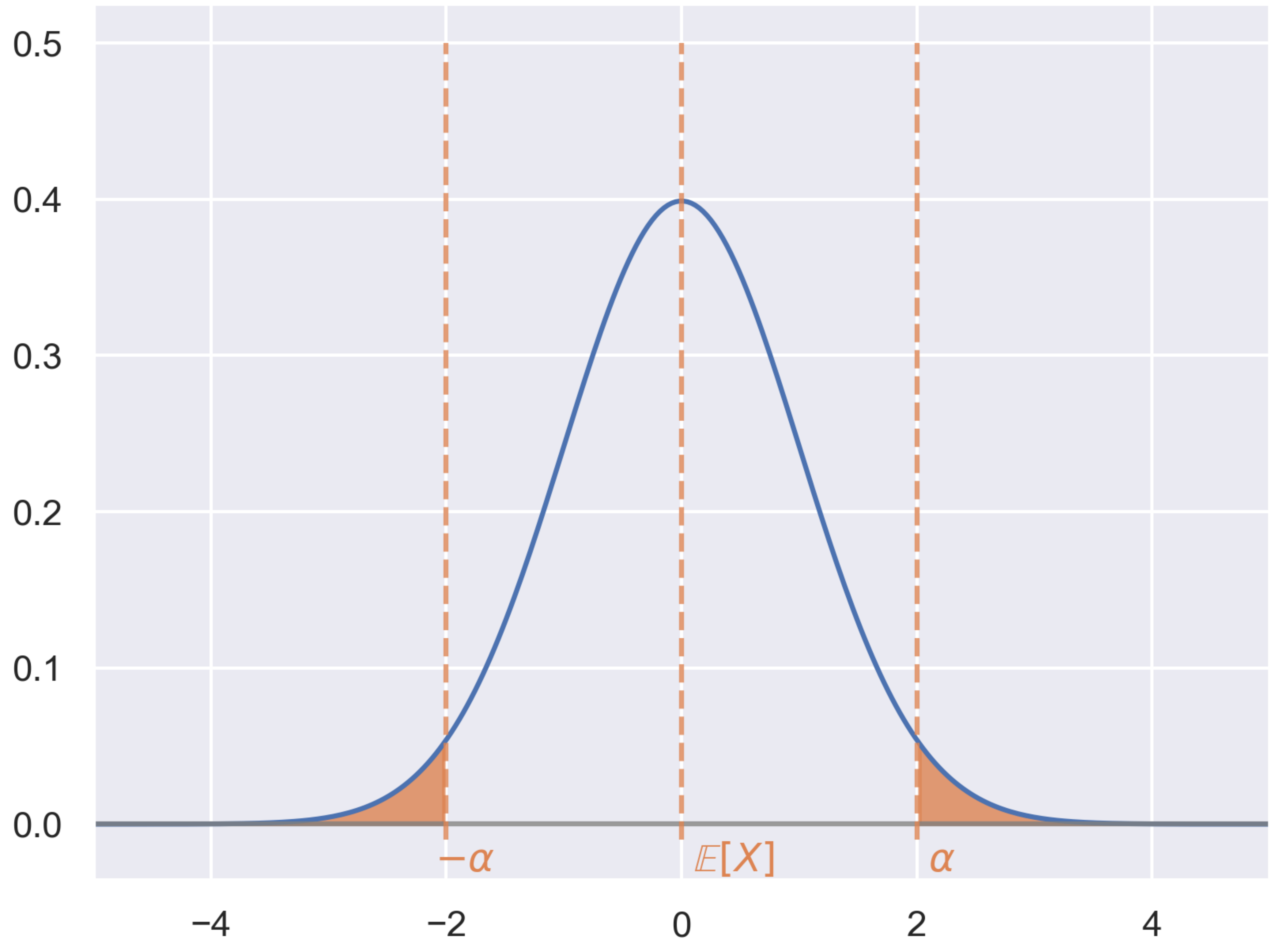
Statement and Proof

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}.$$

Proof.

Apply Markov's inequality to the random variable $|X - \mu|^2$:

$$\mathbb{P}(|X - \mu| \geq \alpha) = \mathbb{P}(|X - \mu|^2 \geq \alpha^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{\alpha^2} = \frac{\sigma^2}{\alpha^2}.$$



Law of Large Numbers

Proof

Let $X_1, \ldots, X_n$ be i.i.d. with their *sample average* denoted as

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \epsilon\right) = 1.$$

Proof (simplified version with $\sigma^2 < \infty$).

Assuming $\sigma^2 < \infty$, apply Chebyshev's inequality to $\overline{X}_n$:

$$\mathbb{P}(\left|\overline{X}_n - \mu\right| > \epsilon) \leq \frac{\mathrm{Var}(\overline{X}_n - \mu)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

complement. $\overline{X}_n - \mu$

$\alpha, \beta \in \mathbb{R}$

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

$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$ for indep. rand. vars.

$\mathrm{Var}(\overline{X}_n - \mu)$

$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)$

$= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right)$ (indep.)

$= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i)$ (identically dist. $\mathrm{Var}(X_i) = \sigma^2$.)

$= \frac{1}{n^2}\sum_{i=1}^n \sigma^2$

$= \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$

Sample Average

Definition

For i.i.d. random variables $X_1, \ldots, X_n$, their ***sample average/sample mean/ empirical mean*** is the quantity:

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Law of Large Numbers

Example: Mean Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Law of Large Numbers

Example: Mean Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i = \text{ average frequency of heads}$$

Law of large numbers states that for *any* $\epsilon > 0$*, no matter how small*:

$$\lim_{n \to \infty} \mathbb{P}(\left| \overline{X}_n - 1/2 \right| < \epsilon) = 1$$

Law of Large Numbers

Example: Mean Estimator for Coins

$\mathbb{E}[\overline{X}_n] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Law of large numbers states that for *any* $\epsilon > 0$, *no matter how small*:

$$\lim_{n\to\infty} \mathbb{P}(\left|\overline{X}_n - 1/2\right| < \epsilon) = 1$$

$\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X] - 1/4 = 1/2 - 1/4 = 1/4$

We can quantify this more exactly with Chebyshev's inequality:

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} = \frac{1}{4n}$$

Therefore, using Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}(0.4 \le \overline{X}_n \le 0.6) &= \mathbb{P}(\left|\overline{X}_n - \mu\right| \le 0.1) \\ &= 1 - \mathbb{P}(\left|\overline{X}_n - \mu\right| > 0.1) \\ &\ge 1 - \frac{1}{4n(0.1)^2} = 1 - \frac{25}{n}\end{aligned}$$

Chebyshev.

Law of Large Numbers

Example: Mean Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Law of large numbers states that for *any $\epsilon > 0$, no matter how small*:

$$\lim_{n\to\infty} \mathbb{P}(\left| \bar{X}_n - 1/2 \right| < \epsilon) = 1$$

From the previous slide:

$$\mathbb{P}(0.4 \leq \bar{X}_n \leq 0.6) \geq 1 - \frac{25}{n}.$$

So, for example, for $n = 100$ flips, the probability that the frequency of heads is between 0.4 and 0.6 is at least 0.75.

Law of Large Numbers

Example: Mean Estimator for Coins

$$\overline{X}_{100} = \frac{1}{100}\sum_{i=1}^{100} x_i$$

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

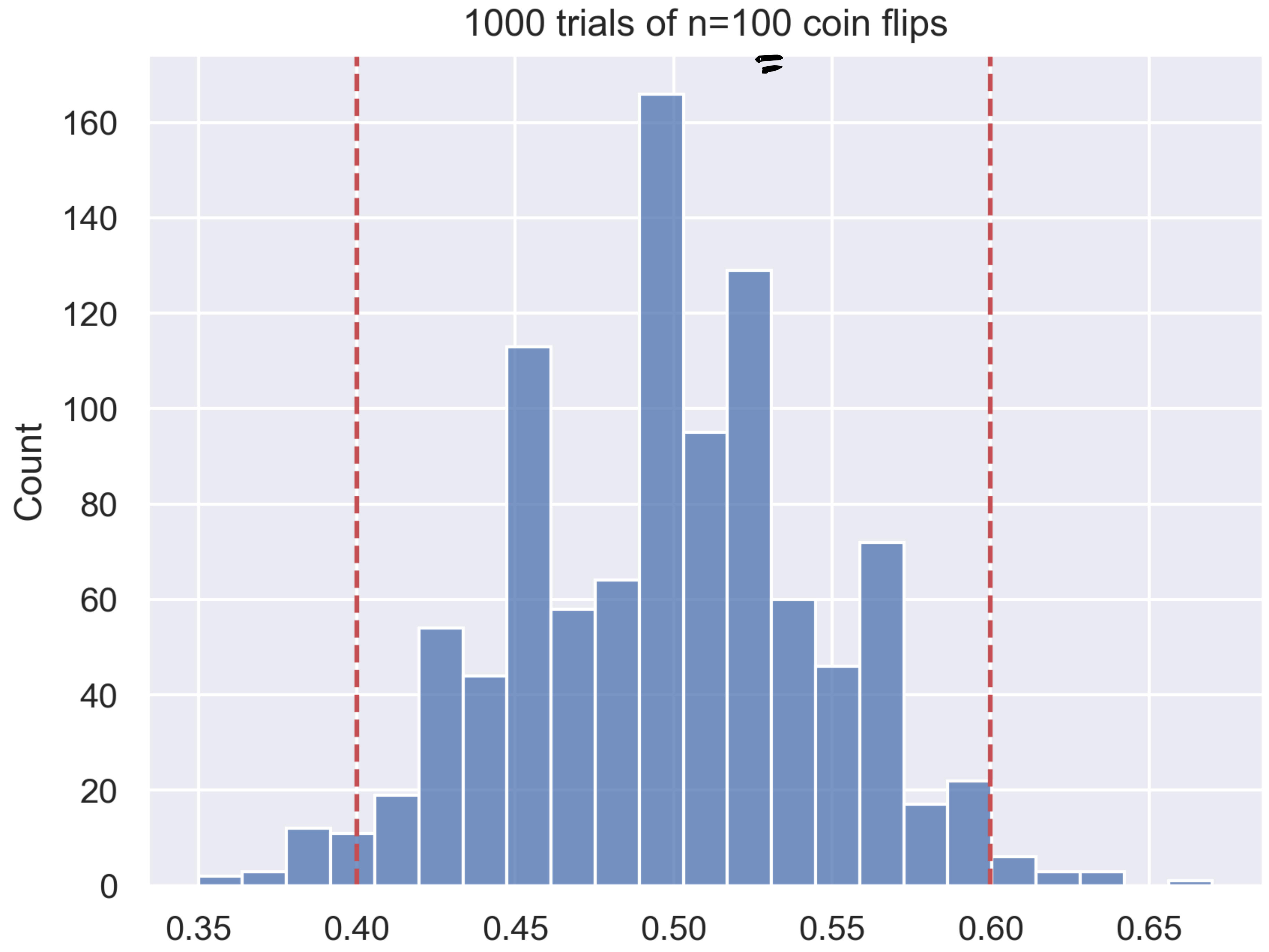
Law of large numbers states that for *any* $\epsilon > 0$*, no matter how small*:

$$\lim_{n\to\infty} \mathbb{P}(\left|\overline{X}_n - 1/2\right| < \epsilon) = 1$$

From the previous slide:

$$\mathbb{P}(0.4 \leq \overline{X}_n \leq 0.6) \geq 1 - \frac{25}{n}.$$

So, for example, for $n = 100$ flips, the probability that the frequency of heads is between 0.4 and 0.6 is at least 0.75.



Statistical Estimation
Intuition

In a nutshell:

Make some assumptions about data that we're to collect.

(*i.i.d. assumption*).

Collect as much data as we can about the phenomenon.

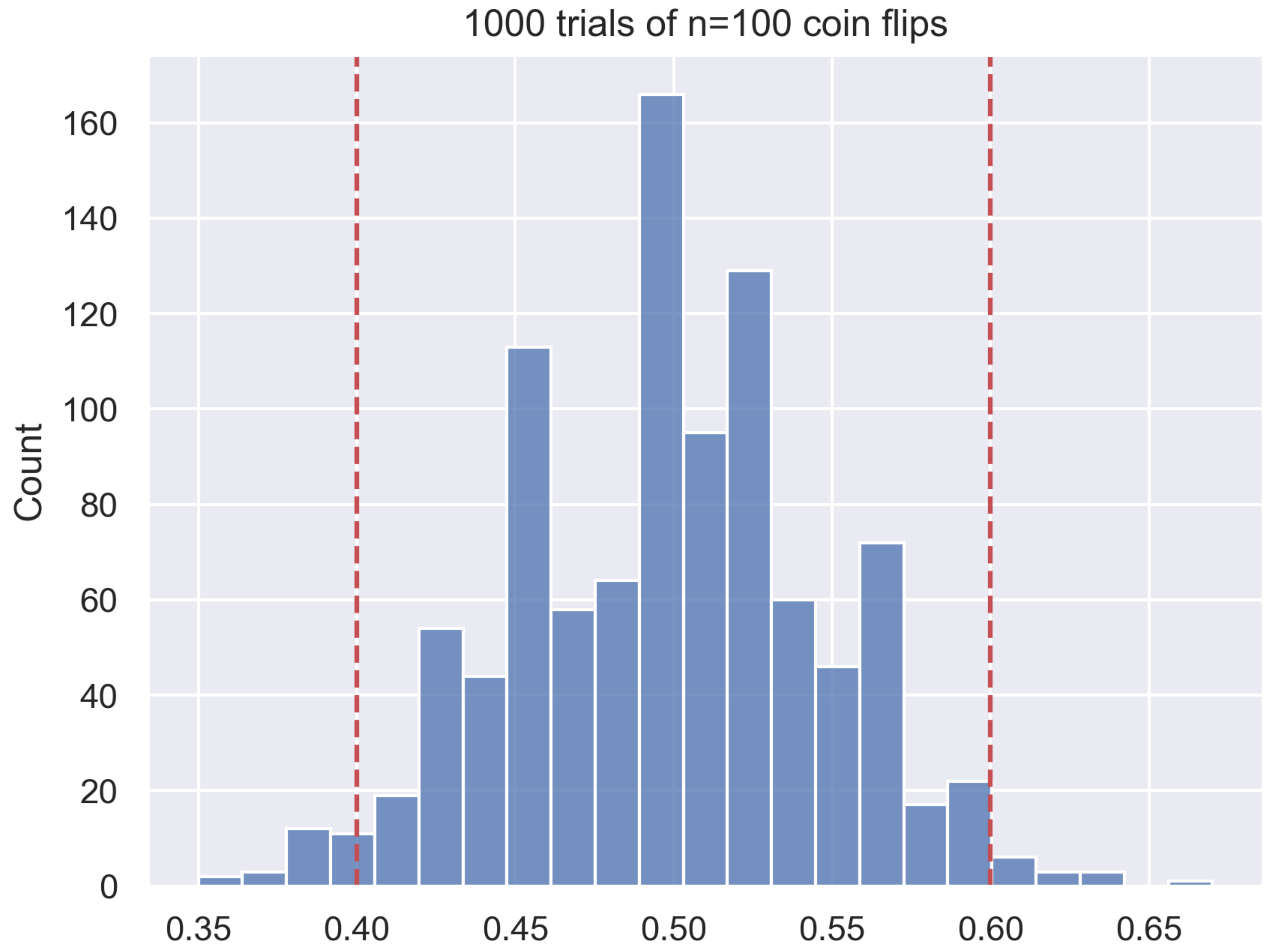
($n = 100$ *coin flips*).

Use the data to derive characteristics (***statistics***) about how the data were generated

(the *true* mean $\mathbb{E}[X_i] = 0.5$)

via some ***estimator.***

$$(\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i)$$



Generalization

Intuition

Statistics/statistical inference concerns drawing conclusions about data that we've already been given.

Generalization is a big concern in machine learning — we also want to describe *future* data well.

Key link:

If the future data comes from the same distribution as our past data, then we can hope to generalize by describing our past data well!

Statistical Estimators

Definition and examples

Statistical Estimator

Intuition

A ***(statistical) estimator*** is a “best guess” at some (unknown) quantity of interest (the ***estimand***) using observed data.

> We will only concern ourselves with ***point estimation***, where we want to estimate a single, fixed quantity of interest (as opposed to, say, an interval).

The quantity doesn’t have to be a single number; it could be, for example, a fixed vector, matrix, or function.

Statistical Estimator

Definition

Let $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from some distribution $\mathbb{P}_X$. An ***estimator*** $\hat{\theta}_n$ of some fixed, unknown parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n).$$

Defined similarly for random vectors.

Statistical Estimator

Definition

Let $X_1, \ldots, X_n$ be n i.i.d. random variables drawn from some distribution $\mathbb{P}_X$. An **estimator** $\hat{\theta}_n$ of some fixed, unknown parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n).$$

Defined similarly for random vectors.

Importantly: statistical estimators are functions of random variables, so they are *themselves* random variables!

Statistical Estimator

Example: Mean Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$. $= \mathbb{E}[X]$

Estimator: $\hat{\theta}_n = \overline{X}_n := \dfrac{1}{n} \sum_{i=1}^{n} X_i$.

$\hat{\theta}_n(X_1, \ldots, X_n)$

Statistical Estimator

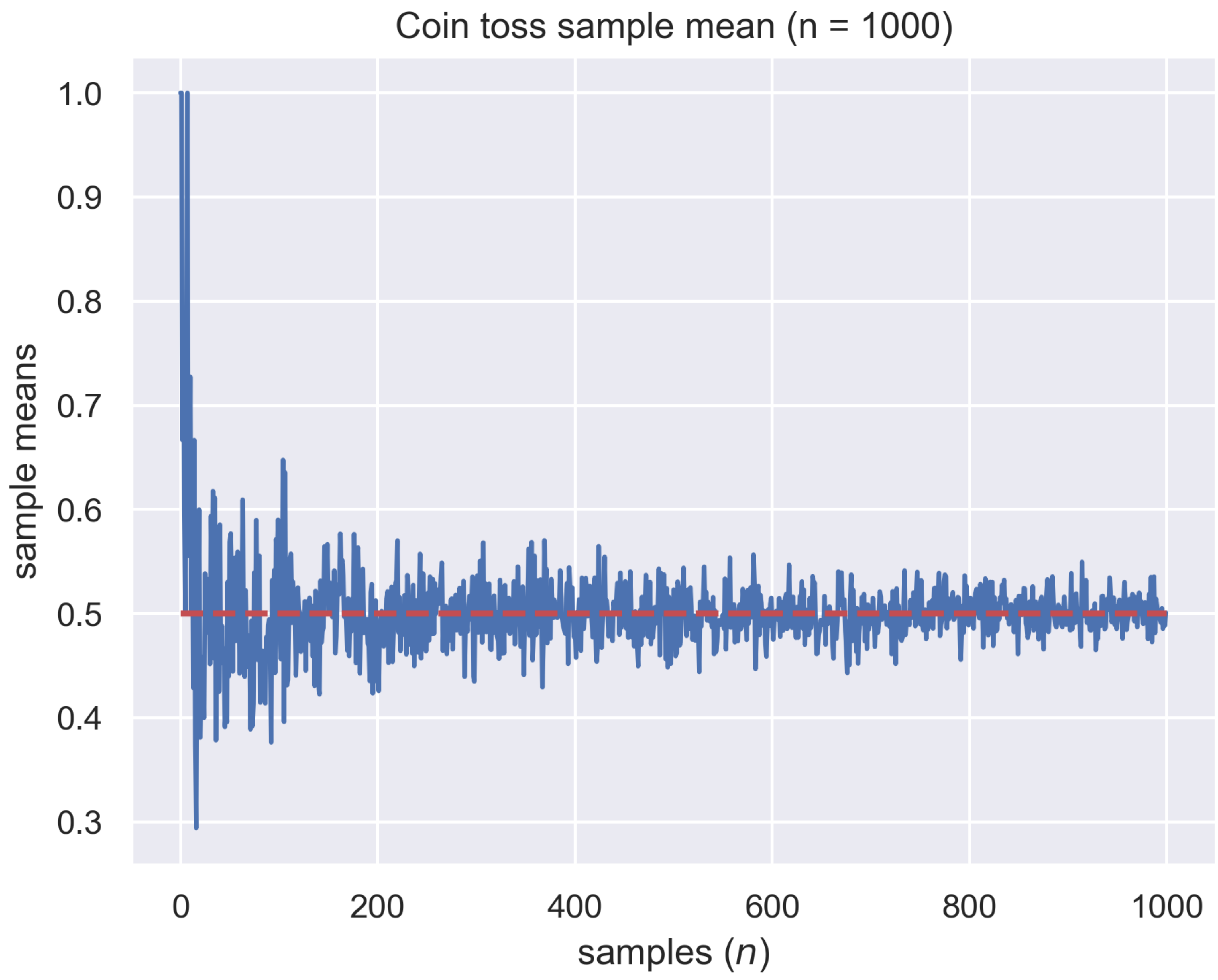
Example: Estimating coin flip

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$.



Statistical Estimator

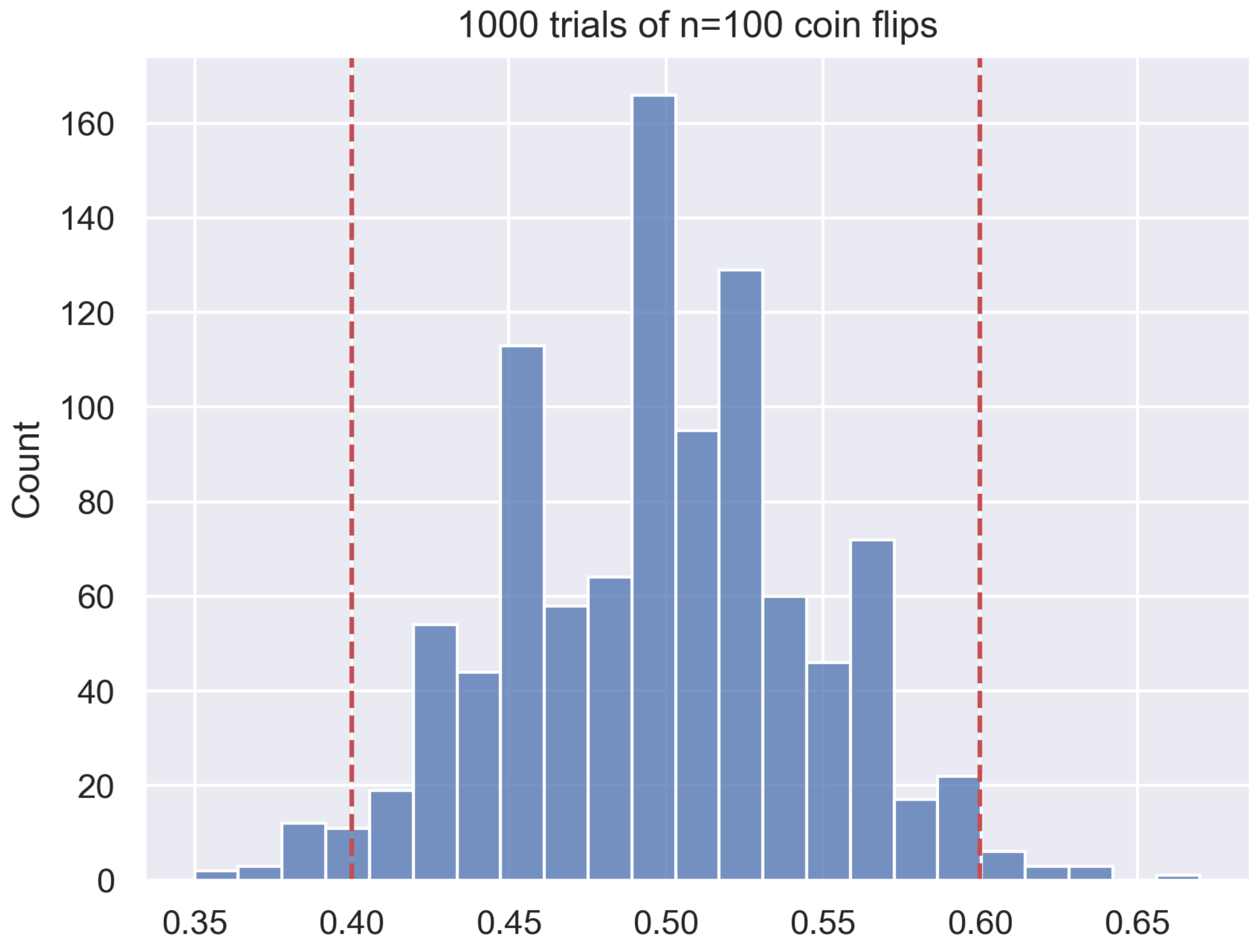
Example: Estimating coin flip

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.



Statistical Estimator

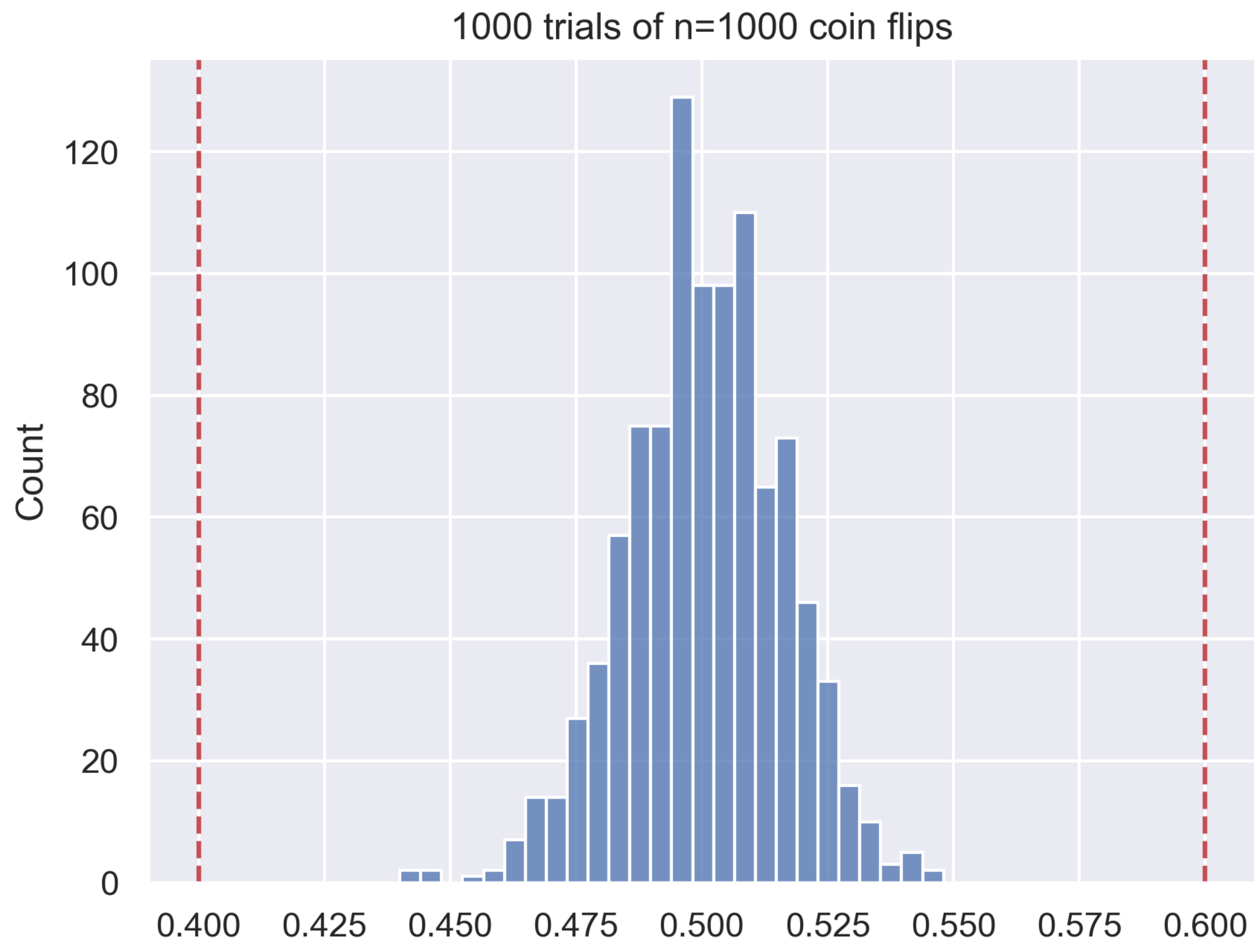
Example: Estimating coin flip

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \dfrac{1}{n} \sum_{i=1}^{n} X_i$.



Statistical Estimator

Example: Variance Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mathrm{Var}(X_i) = (1/2)(1 - 1/2) = 1/4$.

Estimator: $\hat{\theta}_n = S_n^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ (*biased* sample variance).

Statistical Estimator

Example: Variance Estimator for Coins

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mathrm{Var}(X_i) = (1/2)(1 - 1/2) = 1/4$.

Estimator: $\hat{\theta}_n = s_n^2 := \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ (*unbiased* sample variance).

Statistical Estimator

Example: Variance Estimation

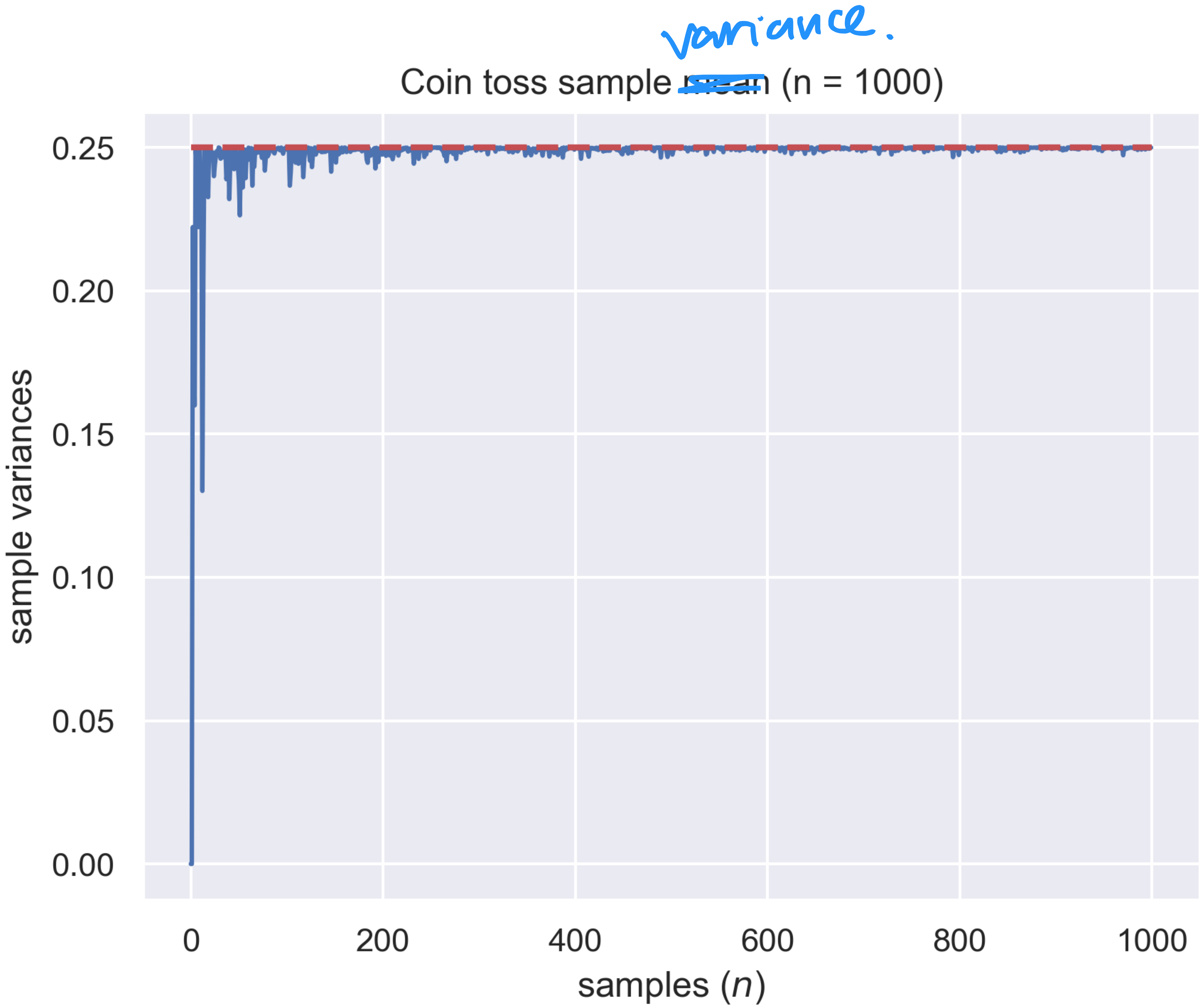
Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

Suppose we independently toss n coins, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mathrm{Var}(X_i) = (1/2)(1 - 1/2) = 1/4$.

Estimator: $\hat{\theta}_n = s_n^2 := \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

(*unbiased* sample variance).



Statistical Estimator

Example: Mean Estimator for Dice

Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

Suppose we independently roll n dice, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$.

Statistical Estimator

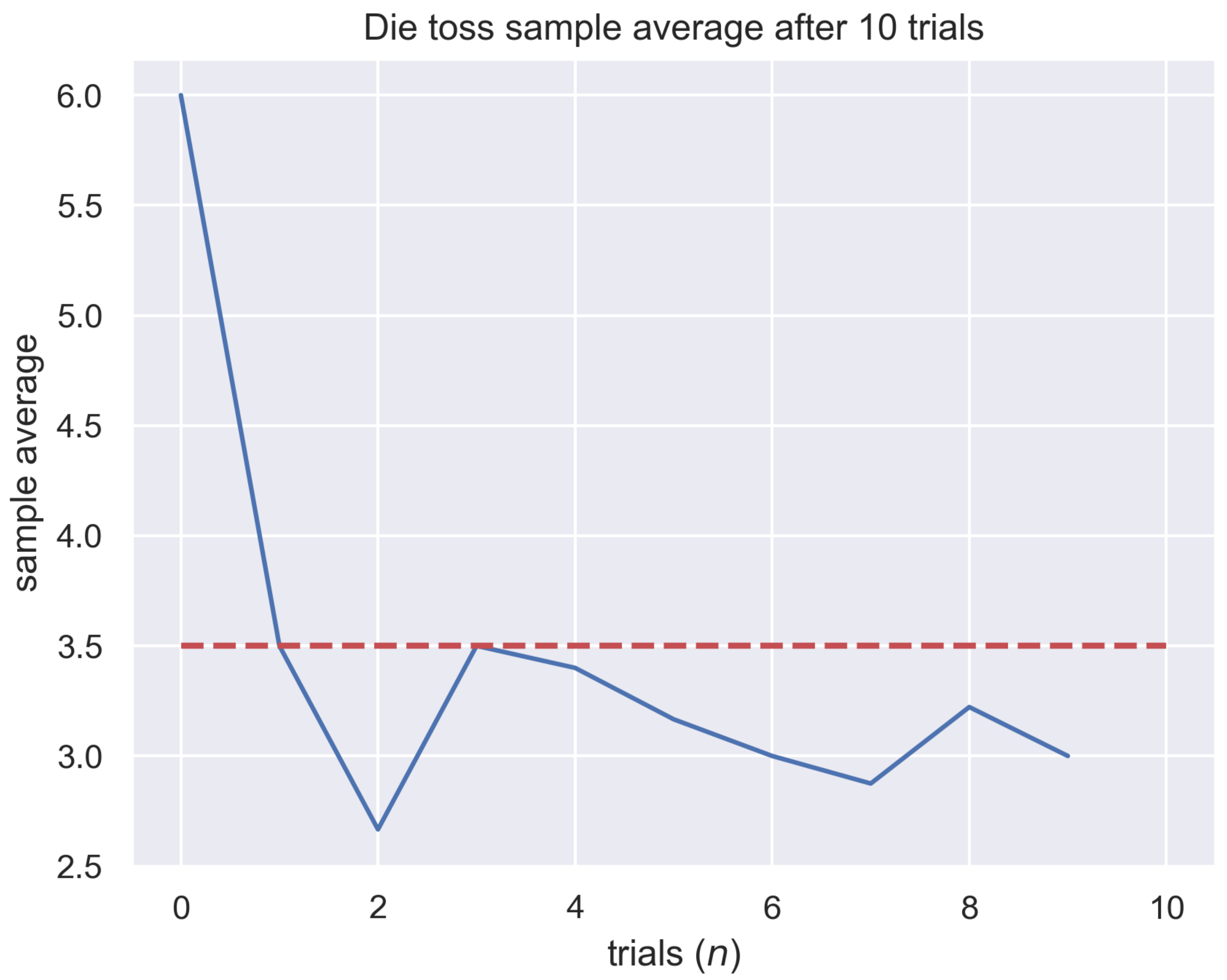
Example: Mean Estimator for Dice

Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

Suppose we independently roll n dice, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$.



Statistical Estimator

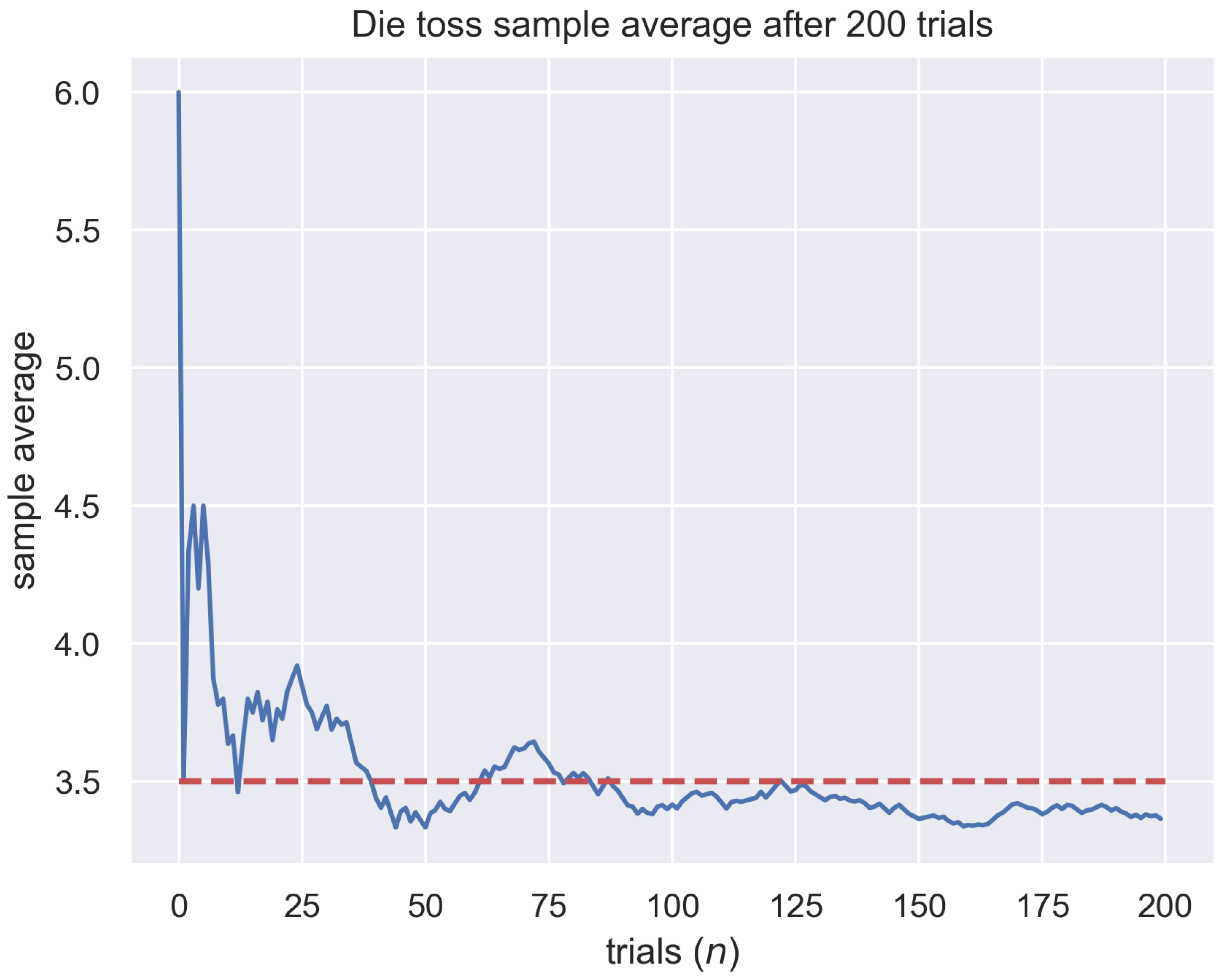
Example: Mean Estimator for Dice

Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

Suppose we independently roll n dice, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \dfrac{1}{n}\sum_{i=1}^{n} X_i$.



Statistical Estimator

Example: Mean Estimator for Dice

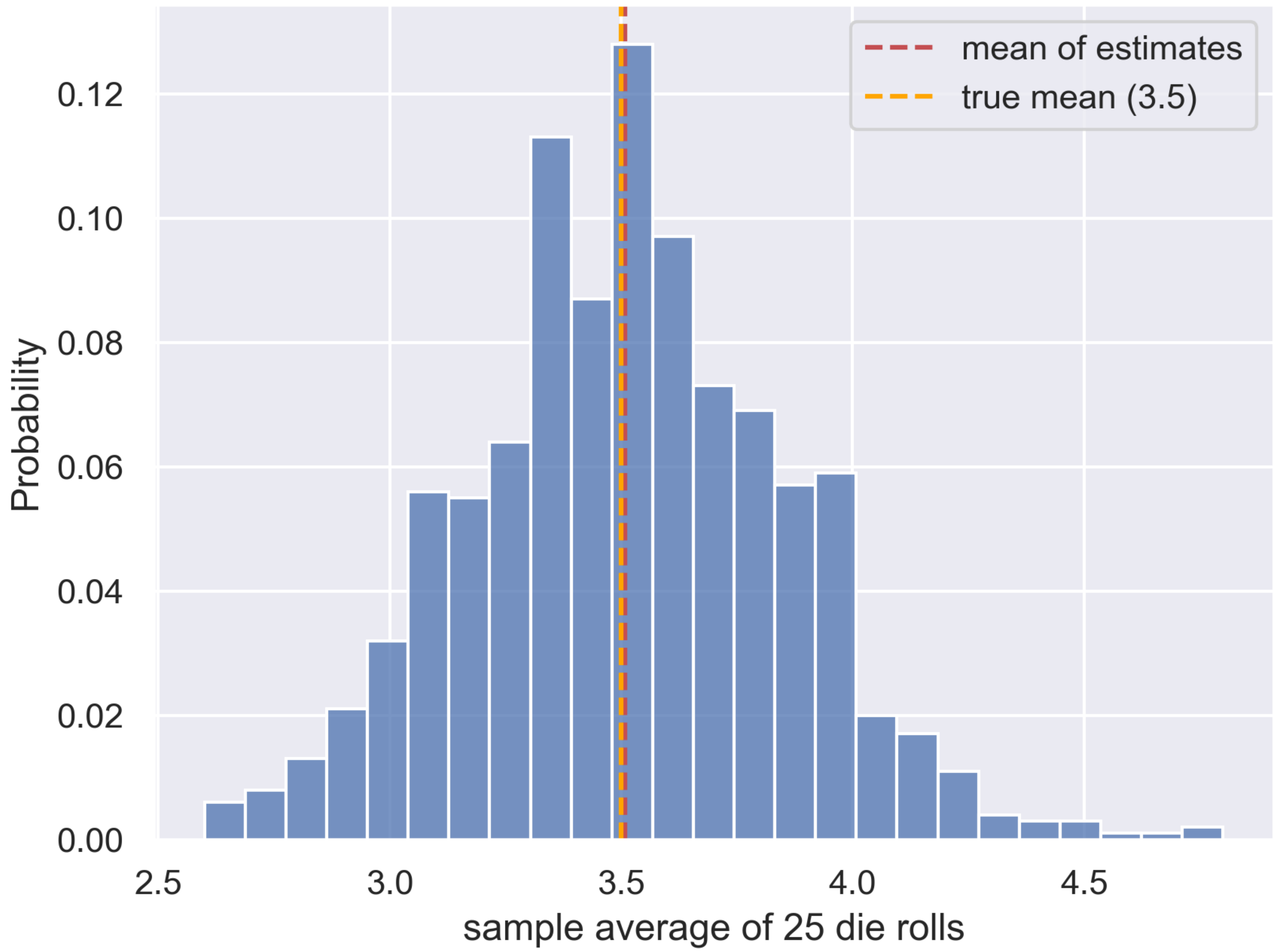
Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

Suppose we independently roll n dice, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.

The estimator is itself a random variable!



Statistical Estimator

Example: Mean Estimator for Dice

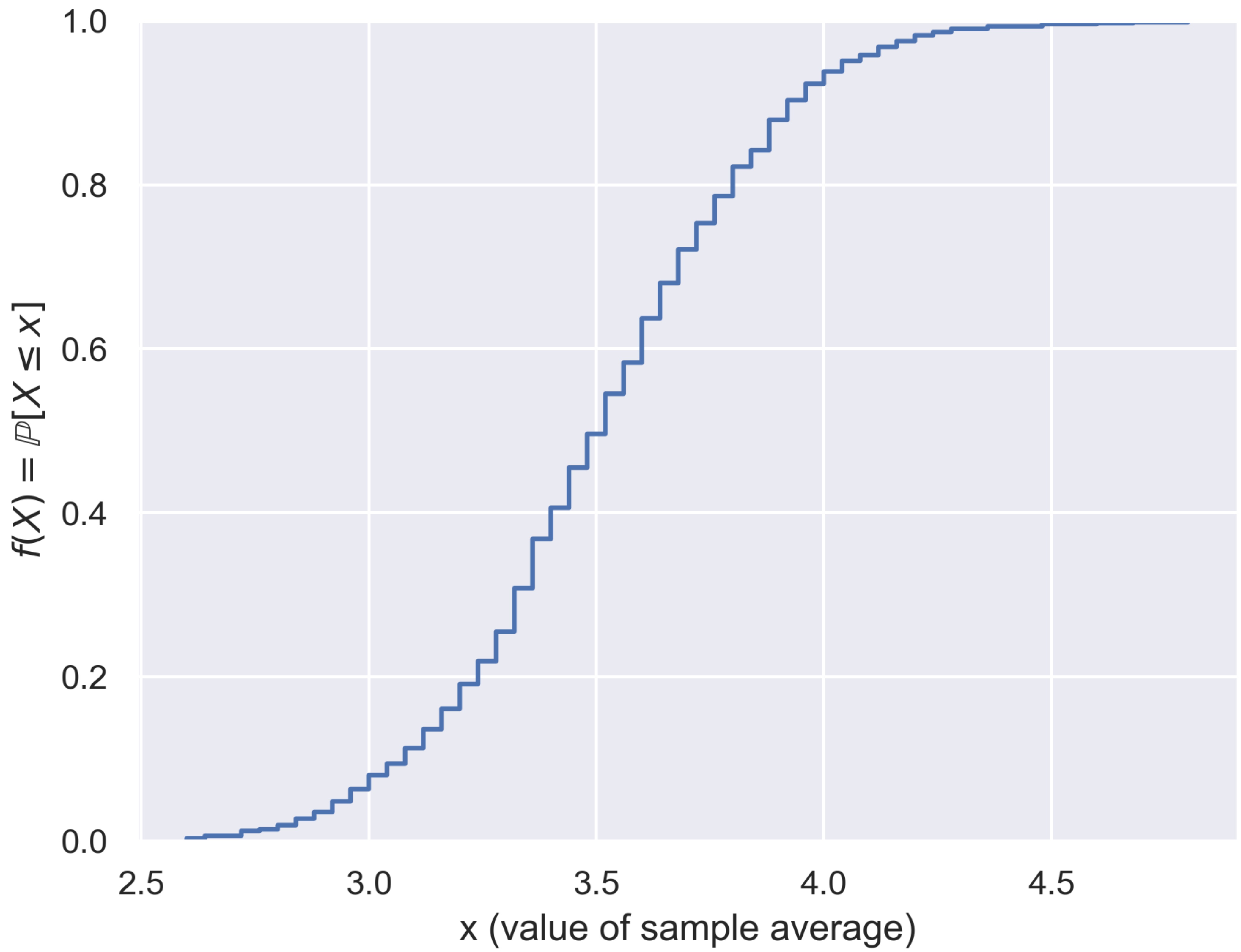
Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

Suppose we independently roll n dice, obtaining RVs $X_1, \ldots, X_n$.

Estimand: $\theta = \mu$.

Estimator: $\hat{\theta}_n = \overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$.

The estimator is itself a random variable!



Statistical Estimator

Example: OLS Estimator

$X = (1, 1)$

$\varepsilon = 1$

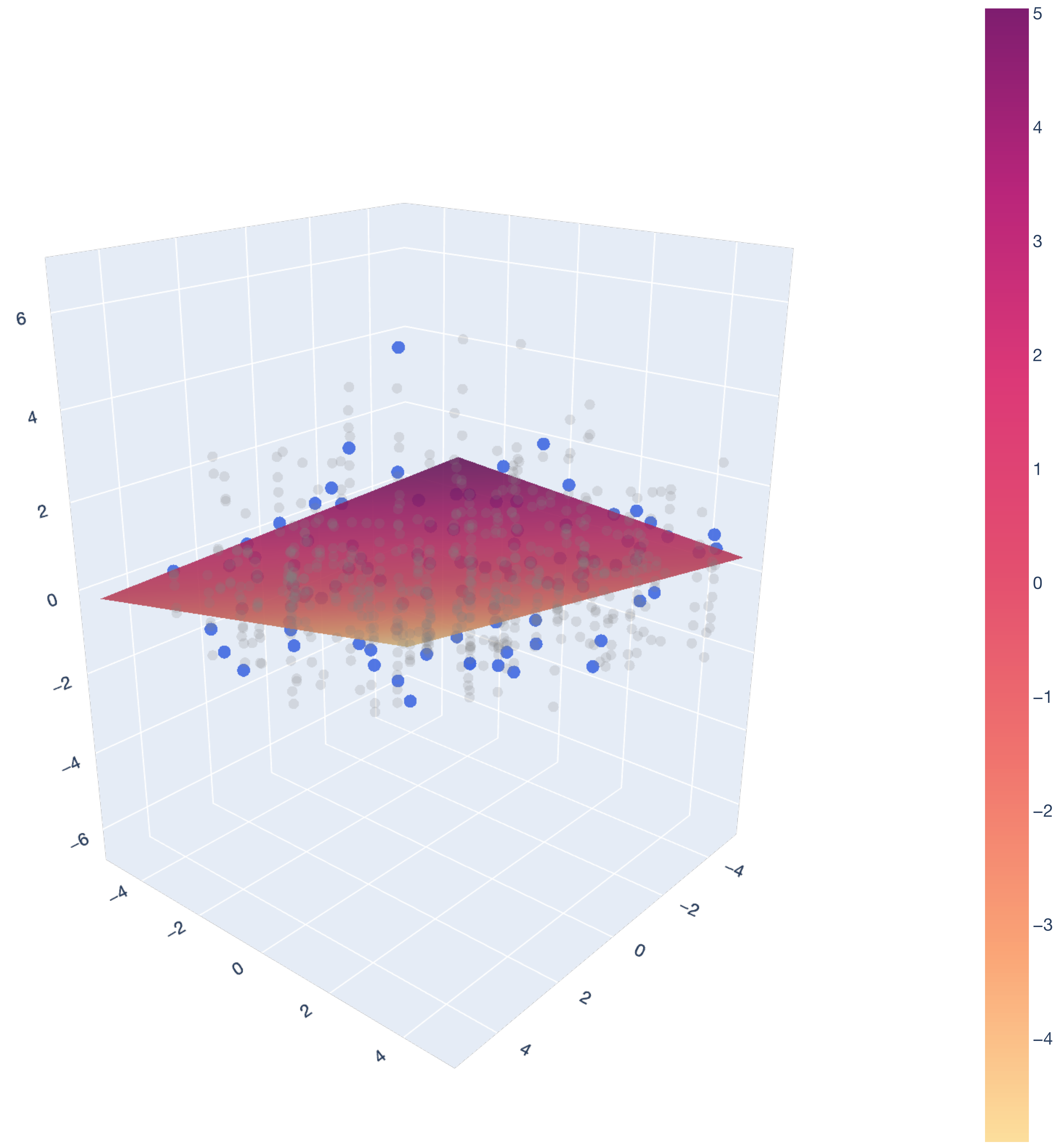
Example. Let $(\mathbf{x}_1, y_1)\ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. samples from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ with the error model:

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

where $\mathbf{w}^* \in \mathbb{R}^d$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ independent from $\mathbf{x}^*$.

Estimand: $\theta = \mathbf{w}^*$.

Estimator: $\hat{\theta}_n = \hat{\mathbf{w}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ are constructed from the samples row-wise.



Statistical Estimator

Example: Ridge Estimator

Example. Let $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. samples from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ with the error model:

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

where $\mathbf{w}^* \in \mathbb{R}^d$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ independent from $\mathbf{x}^*$.

Estimand: $\theta = \mathbf{w}^*$.

Estimator: $\hat{\theta}_n = \hat{\mathbf{w}}_{ridge} = (\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ are constructed from the samples row-wise and $\gamma > 0$ is the *regularization parameter*.

Statistical Estimators

Variance and bias

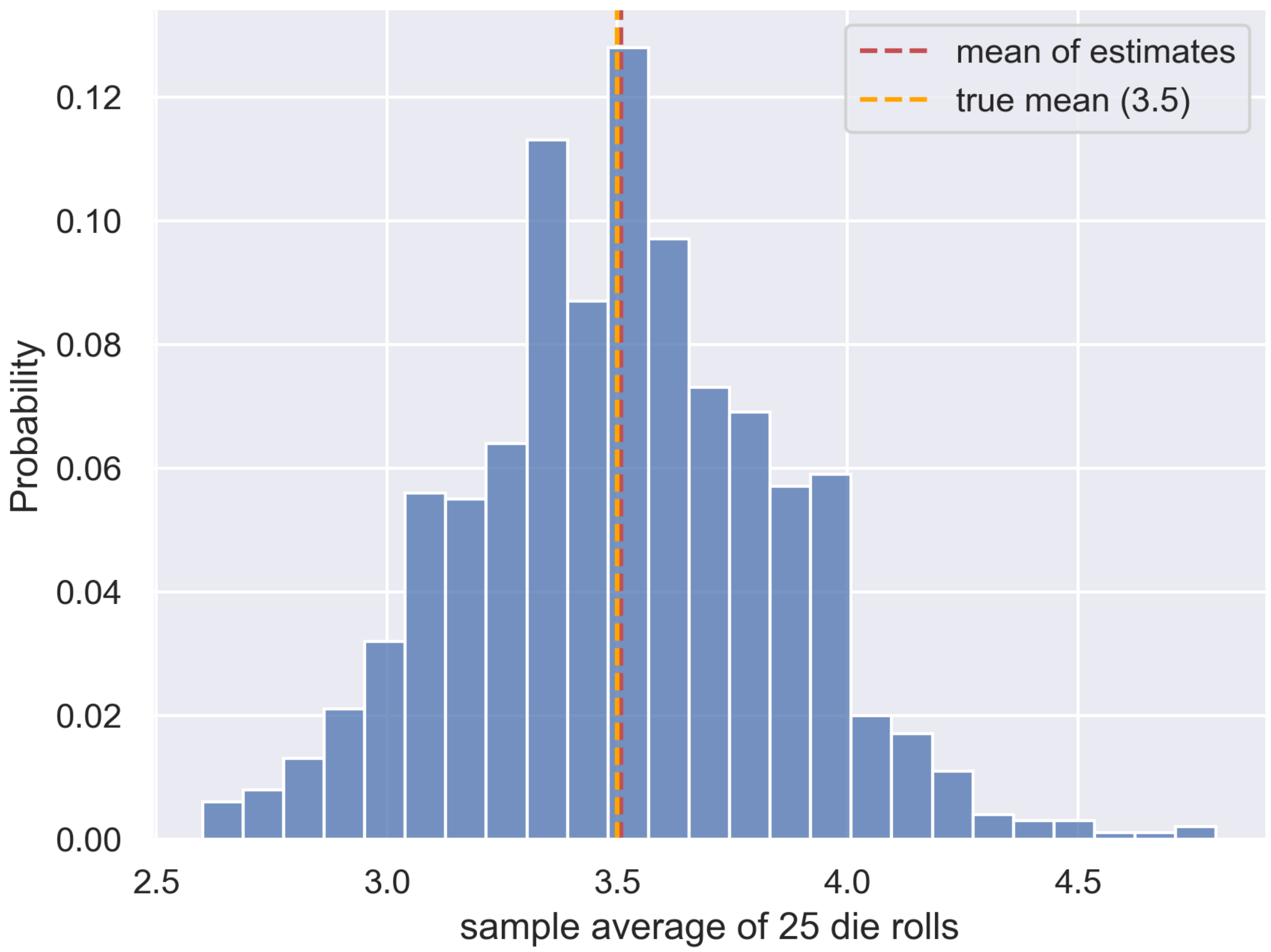
Statistical Estimator

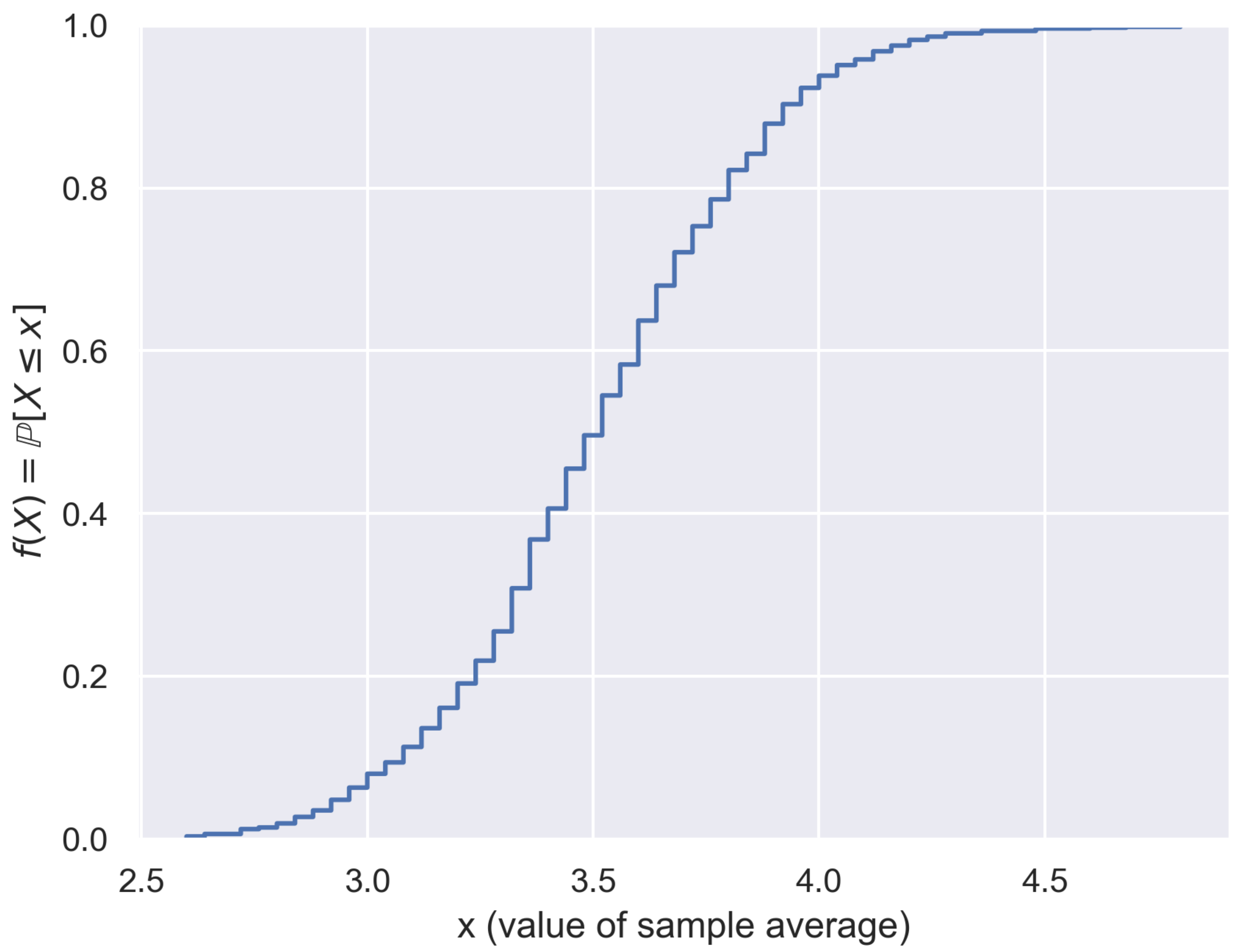
Random Variables

$$\frac{1}{25}\sum_{i=1}^{25} X_i$$

Remember that statistical estimators are random variables!

Below, the mean estimator $\overline{X}_n$ of $n = 25$ dice rolls $X_1, \ldots, X_{25}$.

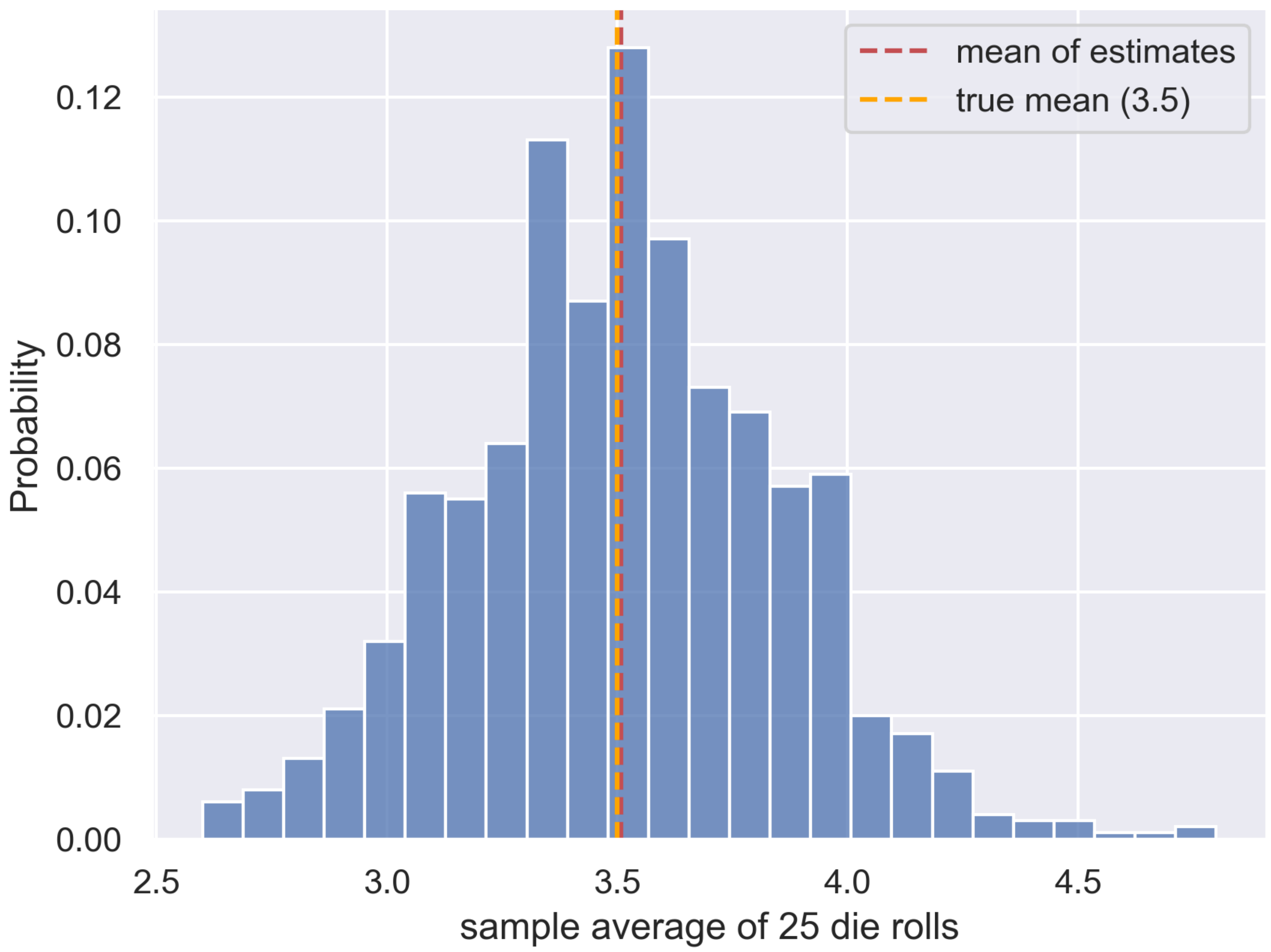


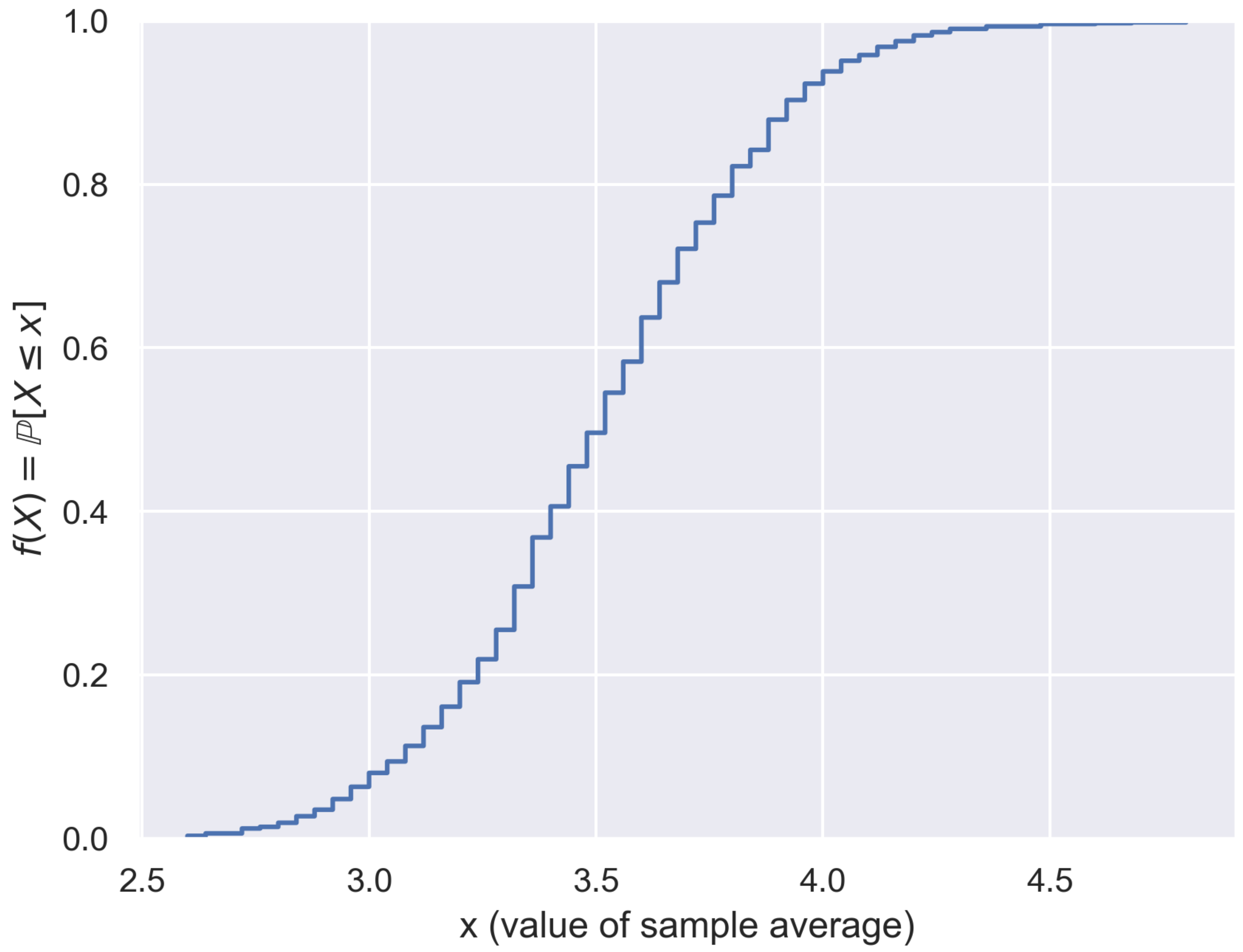


Statistical Estimator
Random Variables

Remember that statistical estimators are random variables!

What are the properties of estimators as random variables?





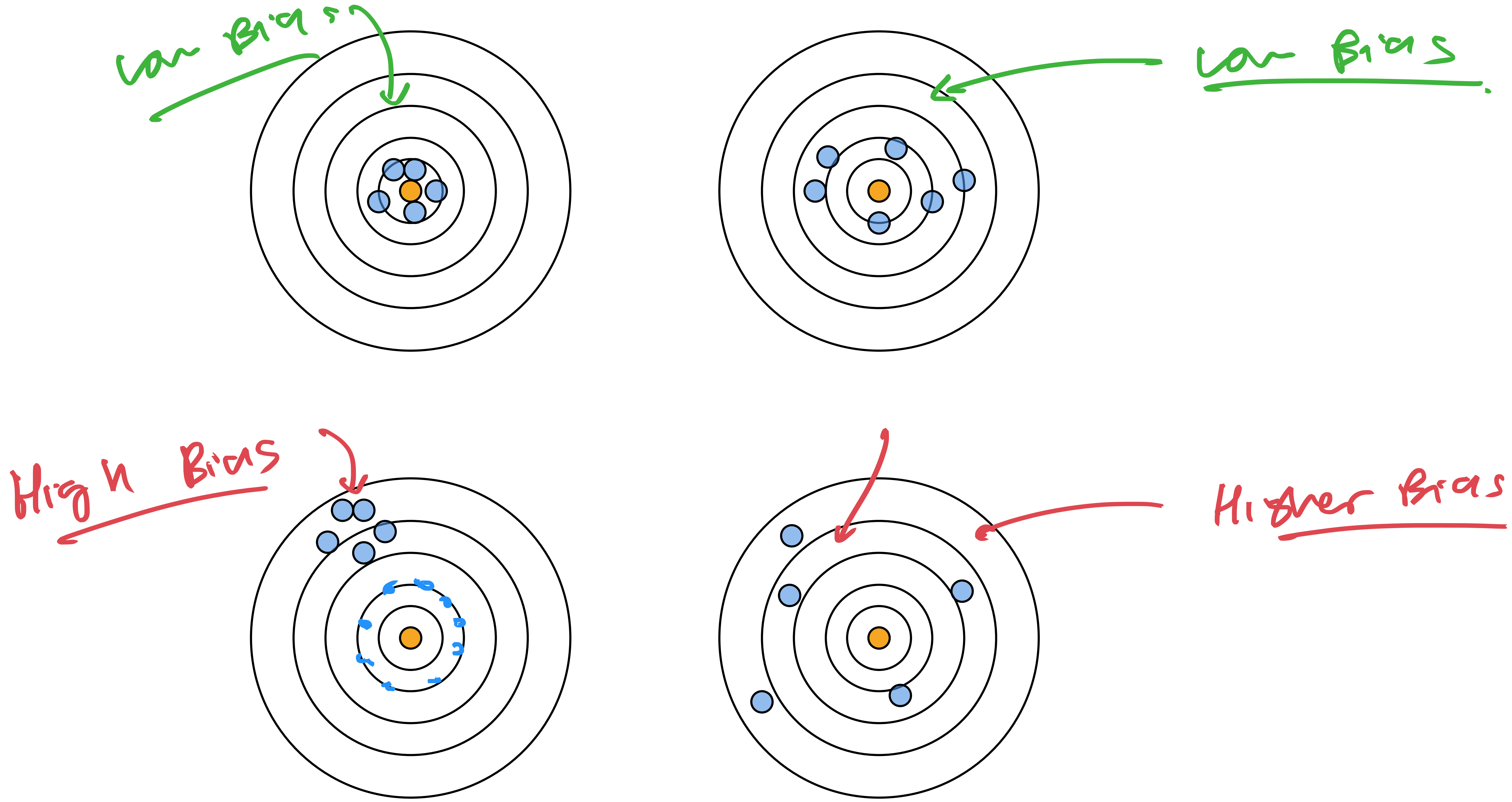
Bias of Estimators

Intuition

Blue = Realizations of the Estimator $\hat{\theta}_n$.
orange = Estimand, θ.

METAPHOR

The ***bias*** of an estimator is “how far off” it is from its estimand.



Bias of Estimators

Definition

Let $\hat{\theta}_n$ be an estimator for the estimand θ. The ***bias*** of $\hat{\theta}_n$ is defined as:

$$\text{Bias}(\hat{\theta}_n) := \mathbb{E}[\hat{\theta}_n] - \theta.$$

We say that an estimator is ***unbiased*** if $\mathbb{E}[\hat{\theta}_n] = \theta$. $\Longrightarrow$ Bias $= 0$.

Bias of Estimators

Example: Constant Estimator

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the bias of the estimator

$$\hat{\theta}_n = 1?$$

$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[1] = 1$$

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \mu = \boxed{1 - \mu}.$$

Bias of Estimators

Example: Single Sample Estimator

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the bias of the estimator

$$\hat{\theta}_n = X_n?$$

$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[X_n] = \boxed{\mu}.$$

Unbiased!

Bias of Estimators

Example: Sample Mean

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the bias of the estimator

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i?$$

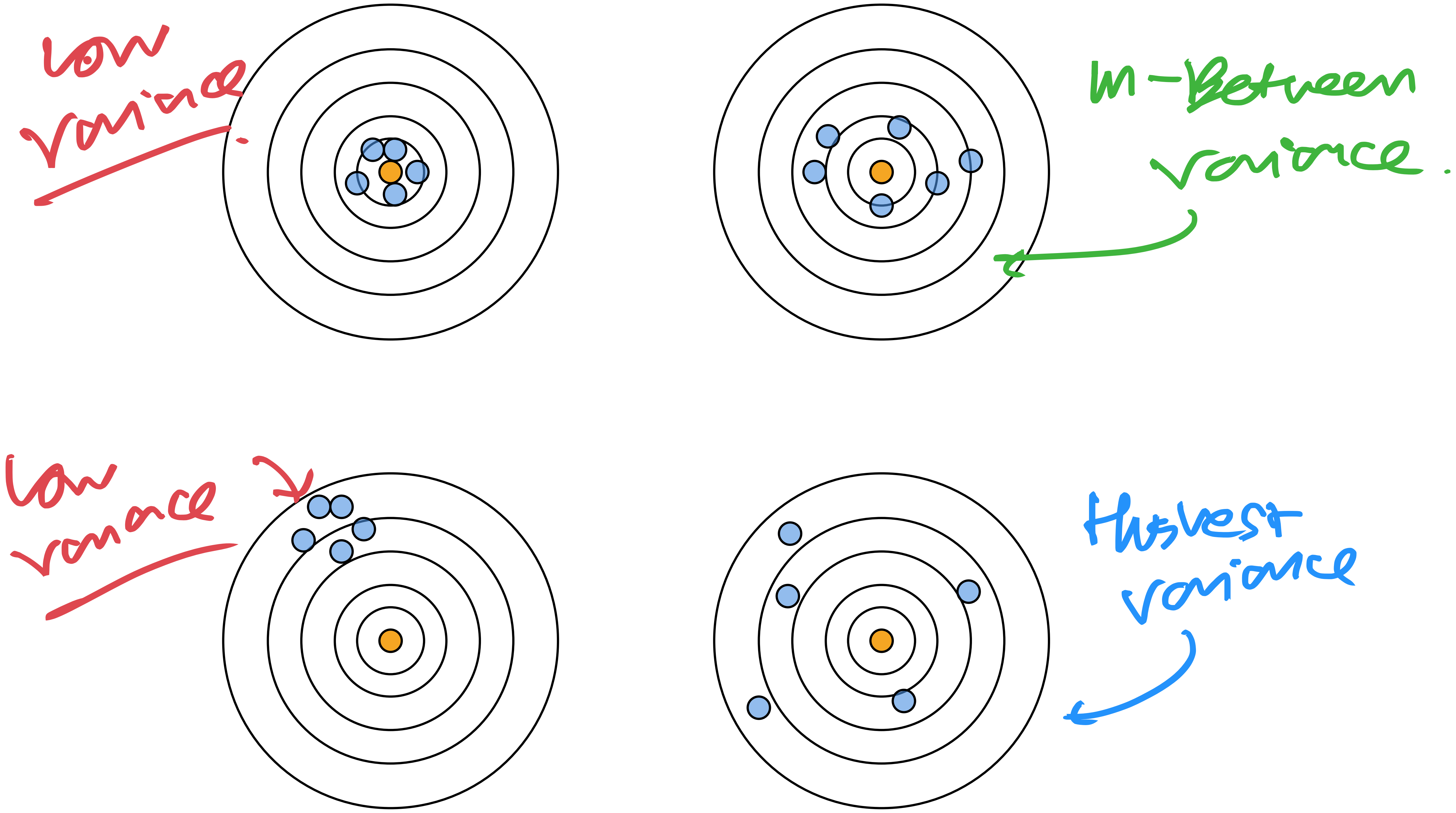
$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{n}{n} \mu = \boxed{\mu}.$$

Unbiased!

Variance of Estimators

Intuition

The ***variance*** of an estimator is simply its variance, as a random variable. This is the "spread" of the estimates from the whatever the estimator's mean is.



Variance of Estimators

Definition

The ***variance*** of an estimator $\hat{\theta}_n$ is simply its variance, as a random variable:

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

The ***standard error*** of an estimator is simply its standard deviation:

$$\mathrm{se}(\hat{\theta}_n) := \sqrt{\mathrm{Var}(\hat{\theta}_n)}.$$

Variance of Estimators

Definition

The ***variance*** of an estimator $\hat{\theta}_n$ is simply its variance, as a random variable:

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

The ***standard error*** of an estimator is simply its standard deviation:

$$\mathrm{se}(\hat{\theta}_n) := \sqrt{\mathrm{Var}(\hat{\theta}_n)}.$$

Notice: The variance of an estimator *does not* concern its estimand.

Variance of Estimators

Example: Constant Estimator

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the variance of the estimator

$$\hat{\theta}_n = 1?$$

$$\text{Var}(\hat{\theta}_n) = \boxed{0}$$

Variance of Estimators

Example: Single Sample Estimator

variance $\sigma^2 := \text{Var}(X_i)$

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the variance of the estimator

$$\hat{\theta}_n = X_n?$$

$$\text{Var}(\hat{\theta}_n) = \text{Var}(X_n) = \boxed{\sigma^2}.$$

Variance of Estimators

Example: Sample Mean

Example. Consider i.i.d. random variables $X_1, \ldots, X_n$ with mean $\mu := \mathbb{E}[X_i]$. Suppose we are estimating the mean, μ. What's the variance of the estimator

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i?$$

$$\begin{aligned}
Var(\hat{\theta}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{n}{n^2}\sigma^2 = \boxed{\frac{\sigma^2}{n}}
\end{aligned}$$

independent

identically distributed

Statistics of OLS

Theorem

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

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

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

$\mathbb{E}[\mathbb{E}[x \mid y]] = \mathbb{E}[x]$

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

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

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

$= \mathbb{E}[(X^T X)^{-1}]\, \sigma^2$

Bias and Variance of OLS

Corollaries from Theorem

d=2: covariance matrix

var((x_1, x_2)).

= Σ = [var(x_1) cov(x_1, x_2) ; cov(x_1, x_2) var(x_2)].

Under the error model:

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

OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the following statistical properties:

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

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

This implies that, as an estimator of $\mathbf{w}^*$ ← Estimand.

$$\mathrm{Bias}(\hat{\mathbf{w}}) = 0$$

cov matrix $(\hat{w})$ = $\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \in \mathbb{R}^{d \times d}$

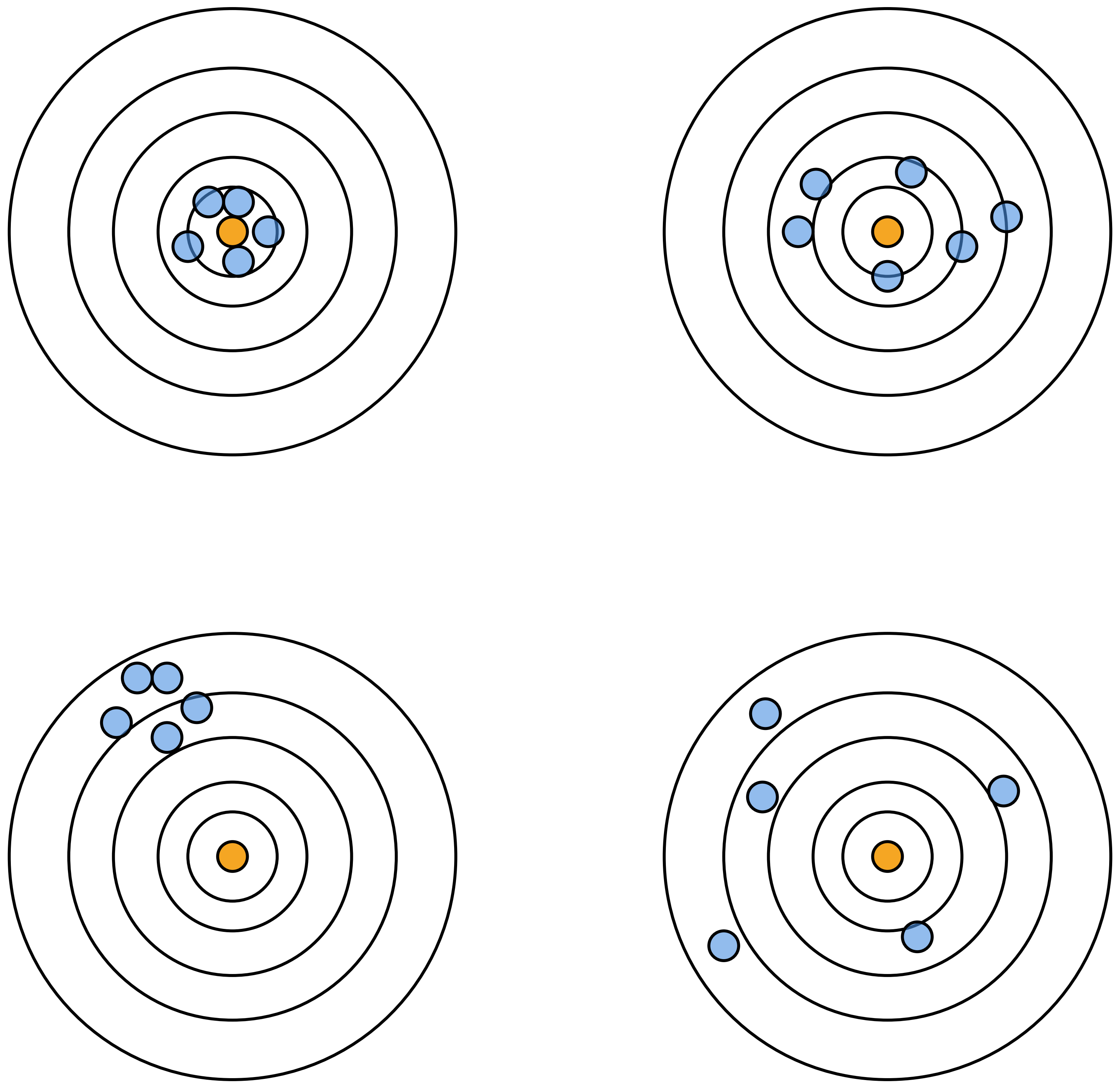
Bias vs. Variance of Estimators

Summary

For an estimator $\hat{\theta}_n$ of the unknown estimand θ, its ***bias*** and ***variance*** are:

$$\text{Bias}(\hat{\theta}_n) := \mathbb{E}[\hat{\theta}_n] - \theta$$

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



Mean Squared Error

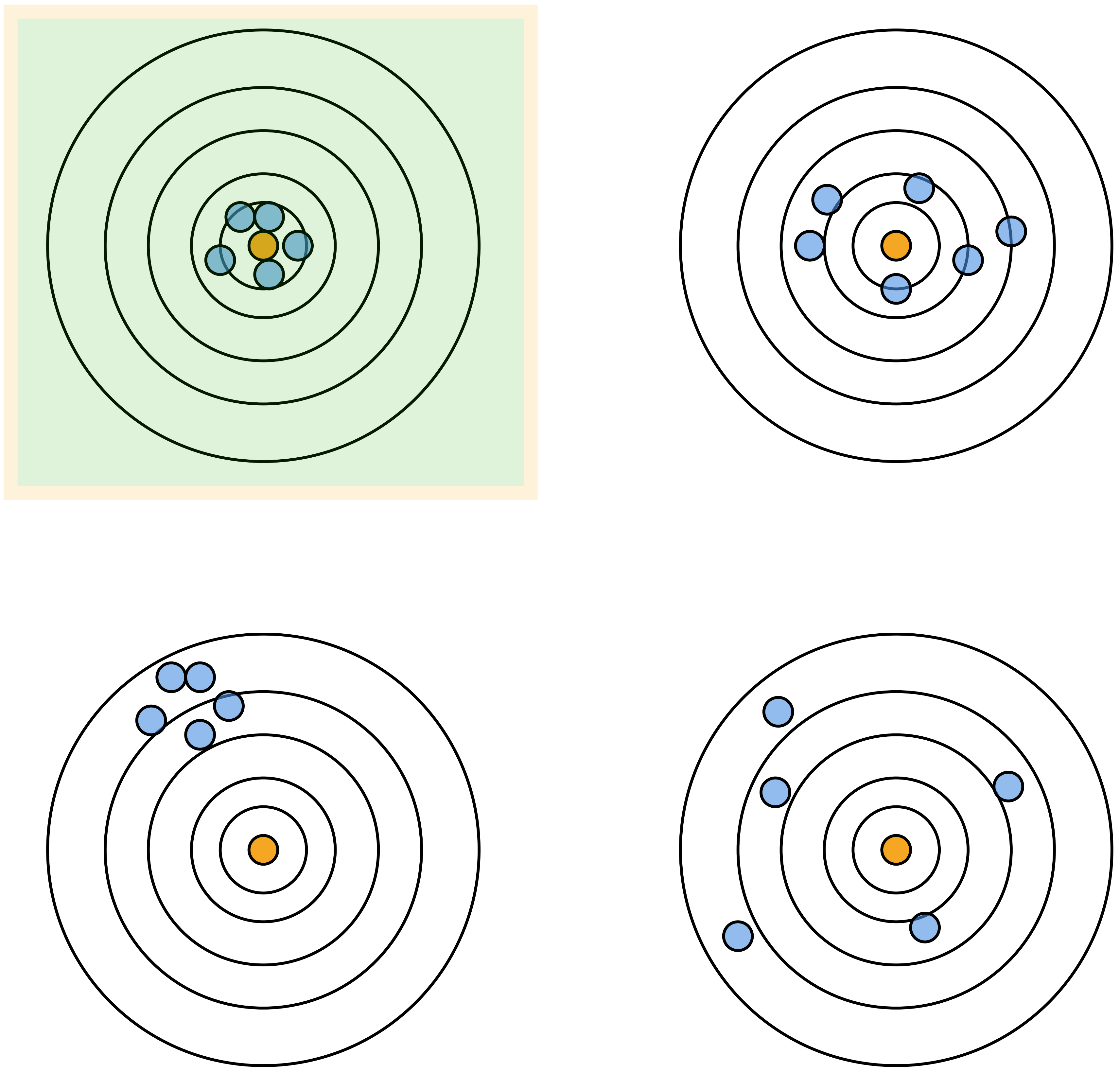
Bias-Variance Tradeoff

Mean Squared Error

Intuition

Intuitively, the best kind of estimator $\hat{\theta}_n$ should have low bias *and* low variance.

And it shouldn't be "too far" from the estimate, in a *distance* sense.



Mean Squared Error

Definition

The ***mean squared error*** of an estimator $\hat{\theta}_n$ of an estimand θ is:

$$\text{MSE}(\hat{\theta}_n) := \mathbb{E}[(\hat{\theta}_n - \theta)^2].$$

This is a common assessment of the quality of an estimator.

Bias-Variance Decomposition

Theorem Statement

Theorem (Bias-Variance Decomposition of MSE). Let $\hat{\theta}_n$ be an estimator of some estimand θ. The ***bias-variance decomposition*** of the mean squared error of $\hat{\theta}_n$ is:

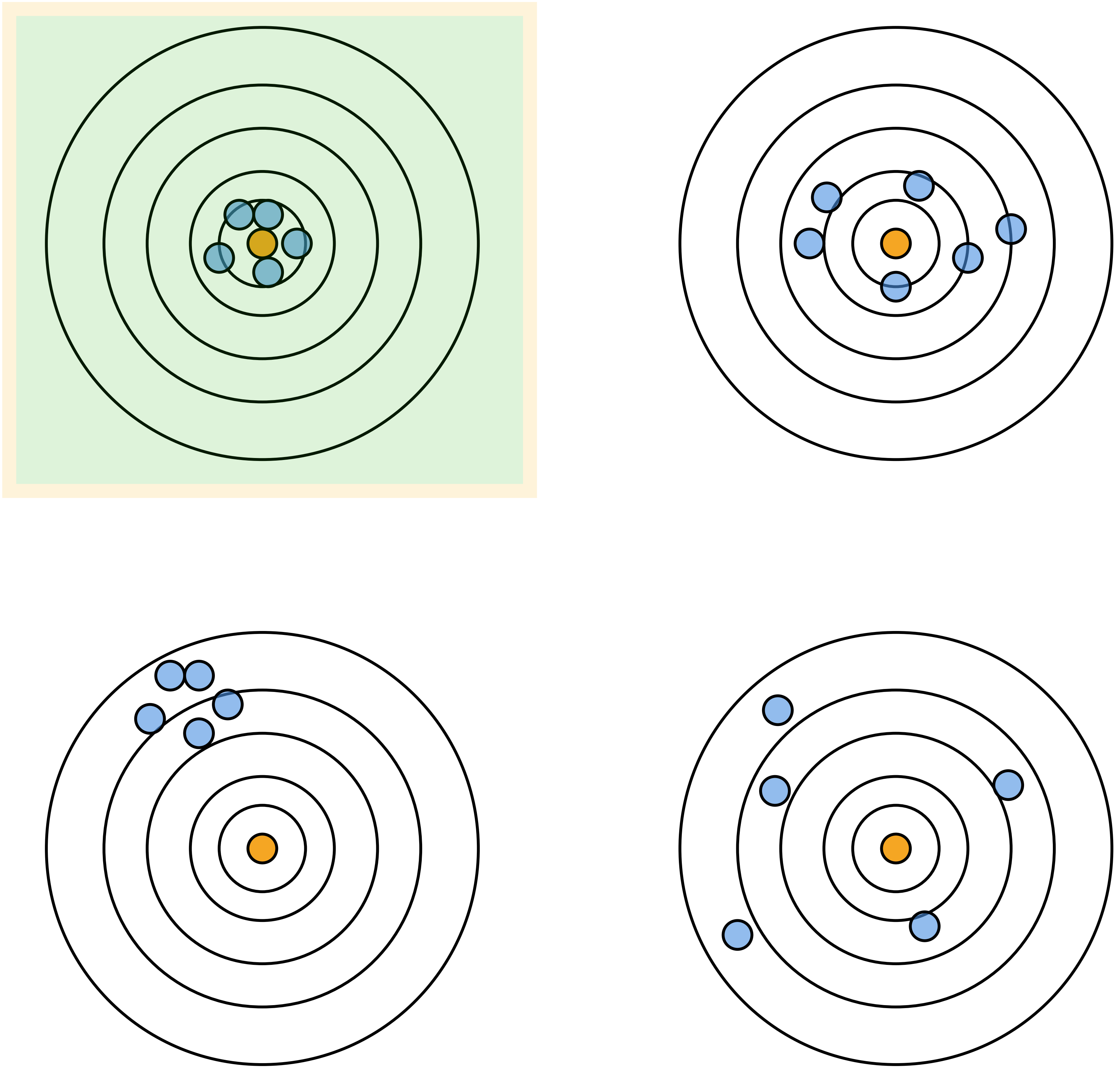
$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n).$$

Bias-Variance Decomposition

Theorem Statement

Theorem (Bias-Variance Decomposition of MSE). Let $\hat{\theta}_n$ be an estimator of some estimand θ. The ***bias-variance decomposition*** of the mean squared error of $\hat{\theta}_n$ is:

$$\mathrm{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \mathrm{Bias}(\hat{\theta}_n)^2 + \mathrm{Var}(\hat{\theta}_n)$$



Bias-Variance Decomposition

Proof

Want to show: $\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \mathrm{Bias}(\hat{\theta}_n)^2 + \mathrm{Var}(\hat{\theta}_n)$

Let $\overline{\theta}_n := \mathbb{E}[\hat{\theta}_n]$. Then:

$$\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2]$$

Bias-Variance Decomposition

Proof

Want to show: $\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$

Let $\overline{\theta}_n := \mathbb{E}[\hat{\theta}_n]$. Then:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}_n - \theta)^2] &= \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)^2] + 2(\overline{\theta}_n - \theta)\mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)] + \mathbb{E}[(\overline{\theta}_n - \theta)^2]\end{aligned}$$

Bias-Variance Decomposition

Proof

Want to show: $\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$

Let $\bar{\theta}_n := \mathbb{E}[\hat{\theta}_n]$. Then:

$$
\begin{aligned}
\mathbb{E}[(\hat{\theta}_n - \theta)^2] &= \mathbb{E}[(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta}_n - \bar{\theta}_n)^2] + 2(\bar{\theta}_n - \theta)\mathbb{E}[(\hat{\theta}_n - \bar{\theta}_n)] + \mathbb{E}[(\bar{\theta}_n - \theta)^2] \\
&= (\bar{\theta}_n - \theta)^2 + \mathbb{E}[(\hat{\theta}_n - \bar{\theta}_n)^2] \\
&= (\mathbb{E}[\hat{\theta}_n] - \theta)^2 + \mathbb{E}[(\hat{\theta}_n - \bar{\theta}_n)^2] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)
\end{aligned}
$$

→ 0 Because $\mathbb{E}[\hat{\theta}_n] = \bar{\theta}_n$ and linearity.

Bias-Variance Decomposition

Example: Coin Flip Mean Estimator

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

What is the mean squared error of $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$?

$\text{Bias}(\overline{X}_n) = 0$

$\text{var}(\overline{X}_n) = \frac{\sigma^2}{n} = \frac{1}{4n}.$

Bias-Variance Decomposition

Example: Coin Flip Mean Estimator

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

What is the mean squared error of $\overline{X}_n := \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i$?

$$\mathrm{MSE}(\overline{X}_n) = \mathrm{Bias}(\overline{X}_n)^2 + \mathrm{Var}(\overline{X}_n)$$

Bias-Variance Decomposition

Example: Coin Flip Mean Estimator

Example. Let X_i be a random variable denoting the outcome of a single fair coin toss, with $X_i = 0$ for tails and $X_i = 1$ for heads. Clearly, $\mu := \mathbb{E}[X_i] = 1/2$.

What is the mean squared error of $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$?

$$\text{MSE}(\overline{X}_n) = \text{Bias}(\overline{X}_n)^2 + \text{Var}(\overline{X}_n)$$

$$\text{Bias}(\overline{X}_n) = 0$$

$$\text{Var}(\overline{X}_n) = \frac{1}{4n}$$

$\Rightarrow$ $\boxed{1/4n}$.

Statistics of OLS

Mean Squared Error of OLS Estimator

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

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

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

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

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

Bias: $\mathrm{Bias}(\hat{\mathbf{w}}) = 0$**, Variance:** $\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \implies \mathrm{MSE}(\hat{\mathbf{w}}) = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}]$

Stochastic Gradient Descent

Estimators for the gradient

Gradient Descent

Algorithm

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$. Initial point $\mathbf{x}_0 \in \mathbb{R}^d$. Step size $\eta \in \mathbb{R}$.

For $t = 1,2,3,\ldots$

Compute: $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1})$.

If $\nabla f(\mathbf{x}_t) = 0$ or $\mathbf{x}_t - \mathbf{x}_{t-1}$ is sufficiently small, then **return** $f(\mathbf{x}_t)$.

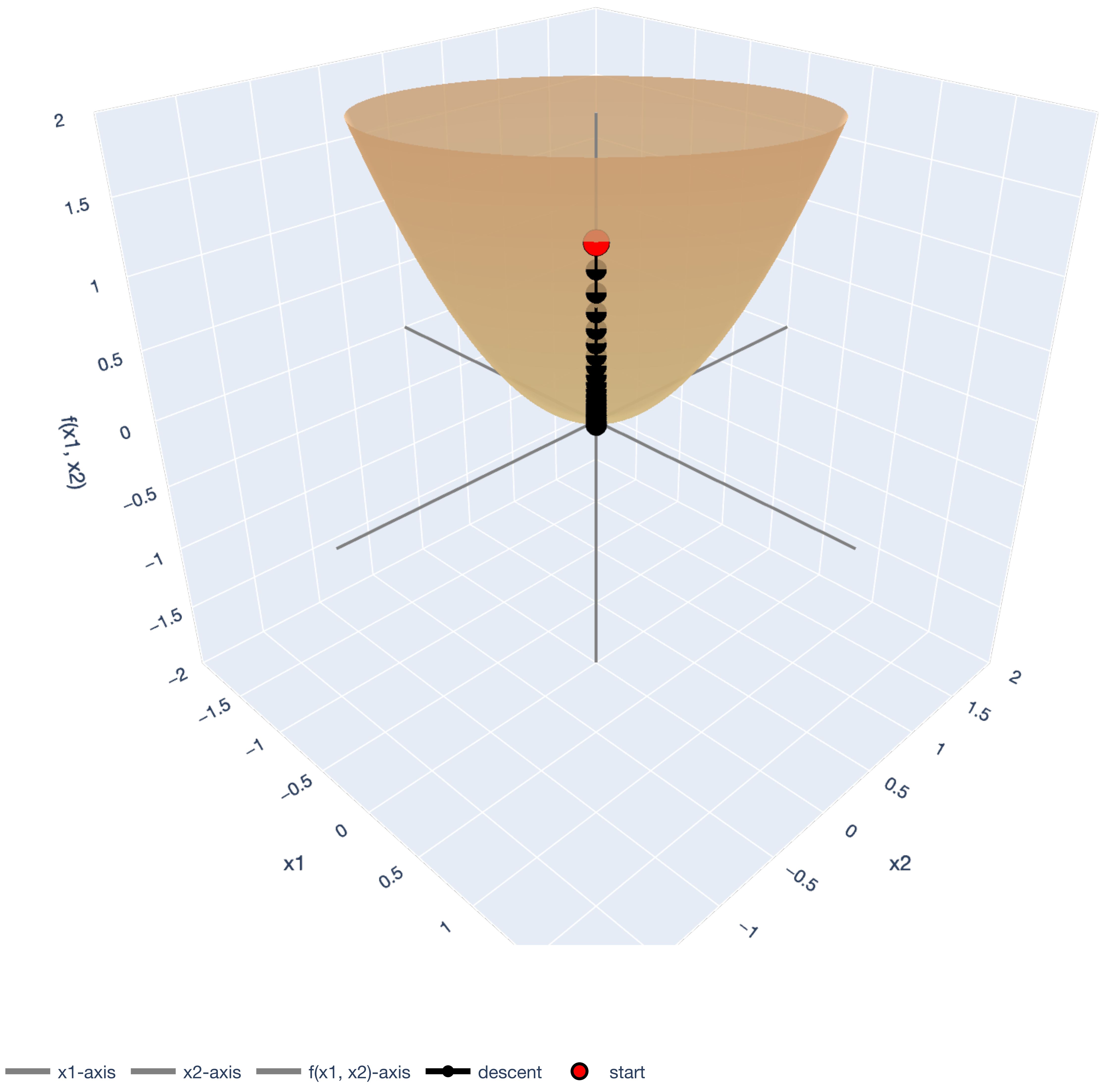
Gradient Descent

Algorithm for OLS

Make an initial guess $\mathbf{w}_0$.

For $t = 1,2,3,\ldots$

- Compute:
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta\mathbf{X}^\top\left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$.
- Stopping condition: If $\|\mathbf{w}_t - \mathbf{w}_{t-1}\| \leq \epsilon$, then return $f(\mathbf{w}_t)$.



What's the problem?

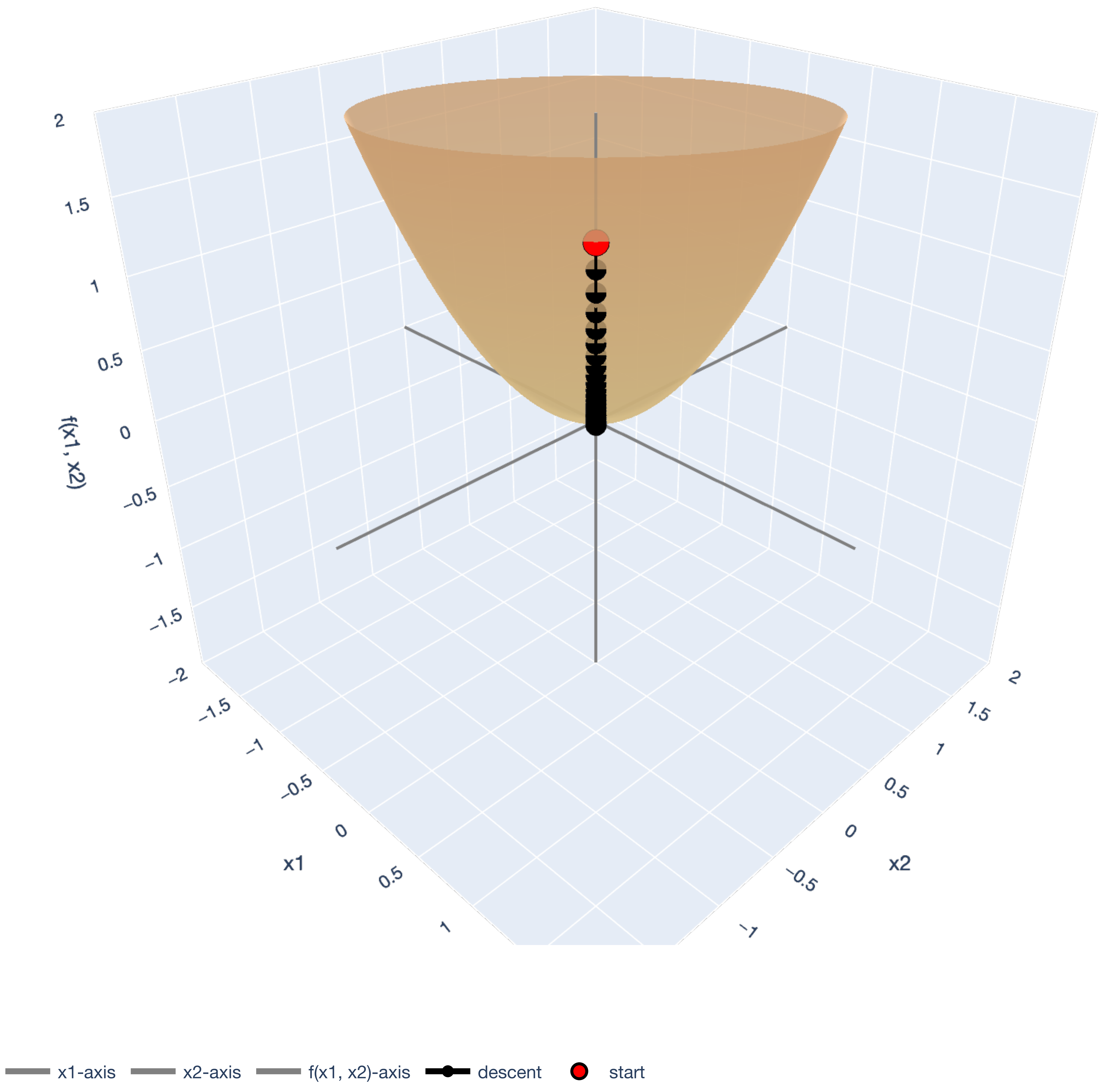
Update Step for OLS

Compute:

$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta\mathbf{X}^\top\left(\mathbf{X}\mathbf{w} - \mathbf{y}\right).$$

This could be expensive for large datasets! (n)

$$\nabla f(w) = \nabla_w \|Xw - y\|^2$$



Stochastic Gradient Descent (SGD)

Intuition

In general, the *objective function* we do gradient descent on typically looks like:

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$$

Let us consider the *average* in this case. For OLS, adding the $1/n$ out front, we have:

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

When we take a gradient, we take it over the *entire* dataset (all n examples):

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

Stochastic Gradient Descent (SGD)

Intuition

When we take a gradient, we take it over the *entire* dataset (all n examples):

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

Idea: What if we just randomly sampled an example i uniformly from $\{1, \ldots, n\}$ and only took the gradient with respect to that example?

$$i \sim \text{Unif}([n]) \Longrightarrow \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

Stochastic Gradient Descent (SGD)

Intuition

In ***stochastic gradient descent*** we replace the gradient over the entire dataset

$$\nabla f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

with an *estimator* of the gradient: $\widehat{\nabla f(\mathbf{w})}$.

Single-sample SGD: Sample a single example i uniformly from $1,\ldots,n$ and take the gradient:

$$\widehat{\nabla f(\mathbf{w})} = \nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

Minibatch SGD: Sample a batch of k examples $B = \{i_1, \ldots, i_k\}$uniformly from all k-subsets of $1,\ldots,n$ and take the gradient:

$$\widehat{\nabla f(\mathbf{w})} = \nabla_{\mathbf{w}}\frac{1}{k}\sum_{j=1}^{k} (\mathbf{w}^\top \mathbf{x}_{i_j} - y_{i_j})^2$$

Gradient Estimator

Unbiased Estimate of the Gradient

Let's try to find the statistical properties of the gradient estimator…

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

Estimator: Sample a single example i uniformly from $1, \ldots, n$ and take the gradient:

$$\widehat{\nabla f(\mathbf{w})} = \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

Gradient Estimator

Unbiased Estimate of the Gradient

Let's try to find the statistical properties of the gradient estimator…

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

Estimator: Sample a single example i uniformly from $1, \ldots, n$ and take the gradient:

$$\widehat{\nabla f(\mathbf{w})} = \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

Bias: The randomness is over the uniform sample, so:

$$\mathbb{E}[\widehat{\nabla f(\mathbf{w})}] = \sum_{i=1}^{n} \frac{1}{n} \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 \implies \mathrm{Bias}(\widehat{\nabla f(\mathbf{w})}) = 0$$

Stochastic Gradient Descent

Single-sample SGD for OLS

Input: Initial point $\mathbf{w}_0 \in \mathbb{R}^d$. Step size $\eta \in \mathbb{R}$.

For $t = 1,2,3,\ldots$

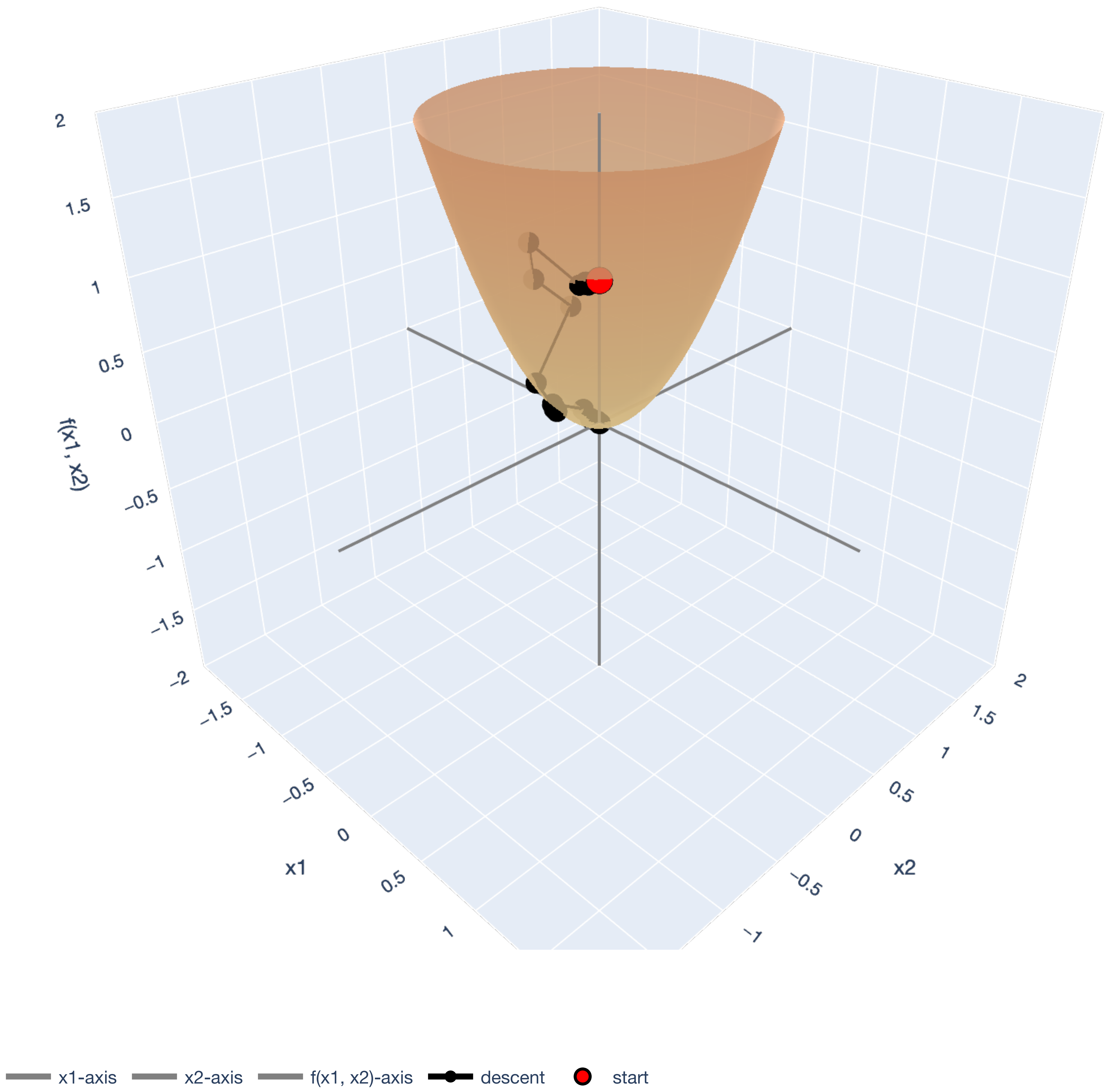
Sample i uniformly from $1,\ldots,n$.

Compute:

$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \eta \widehat{\nabla f(\mathbf{w})} = \mathbf{w}_{t-1} - \eta \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

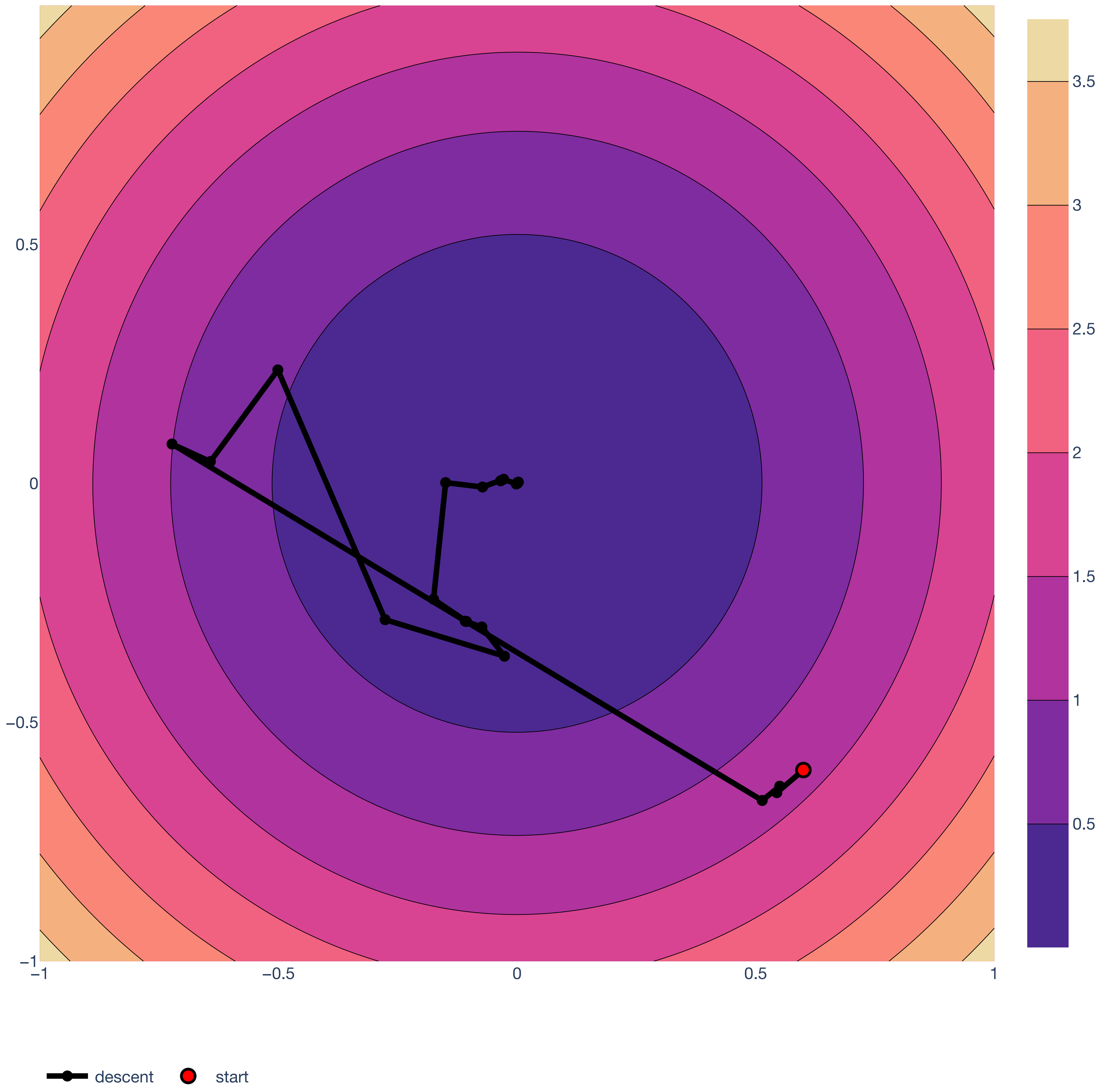
$2(w^T x_i - y_i)\, x_i$

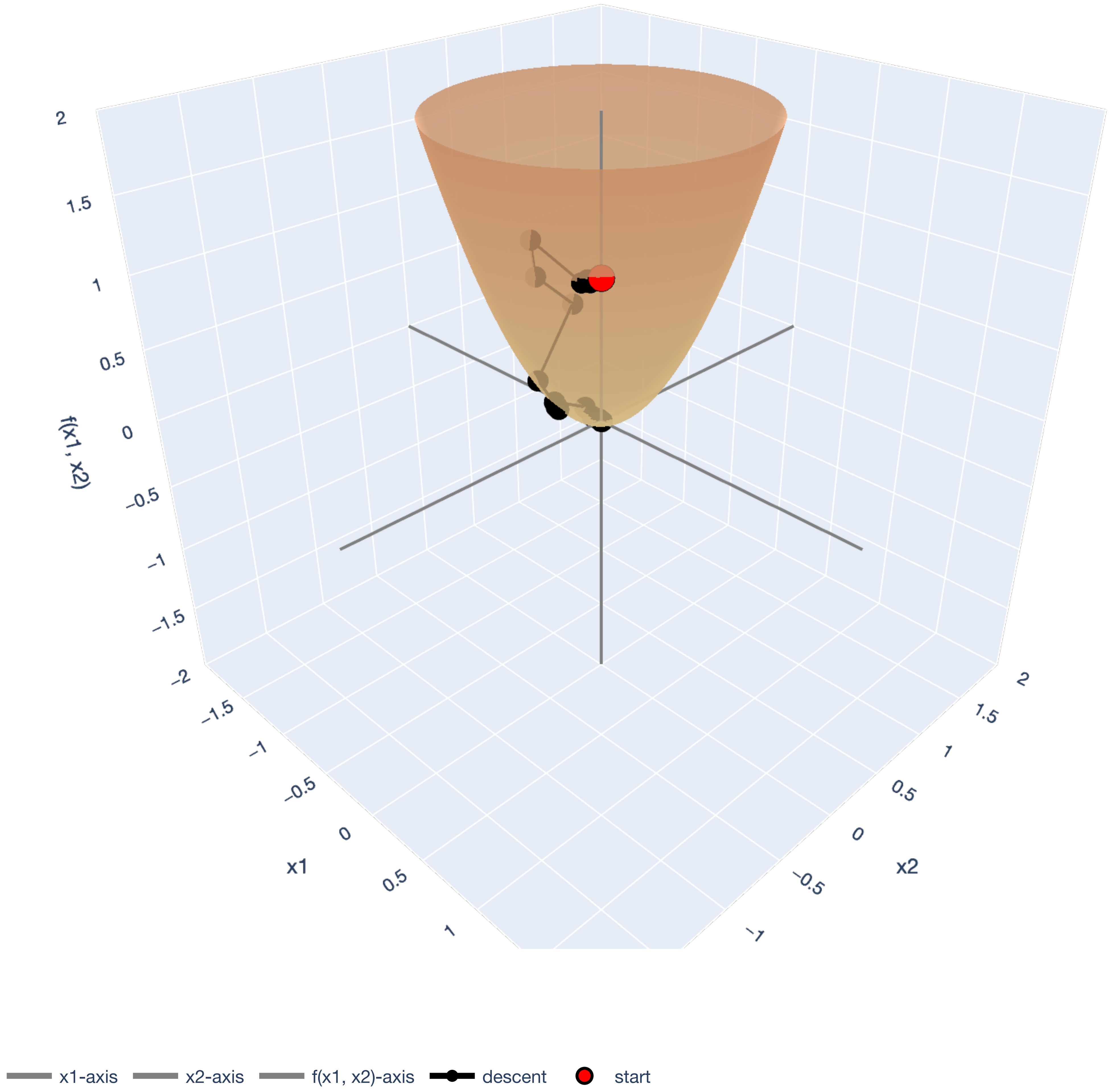
If $\mathbf{w}_t - \mathbf{w}_{t-1}$ is sufficiently small, then **return** $\frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{y}\|^2$.



Stochastic Gradient Descent

Single-sample SGD for OLS





Stochastic Gradient Descent

Minibatch SGD

Input: Initial point $\mathbf{w}_0 \in \mathbb{R}^d$. Step size $\eta \in \mathbb{R}$. Mini-batch size $1 \leq k \leq n$.

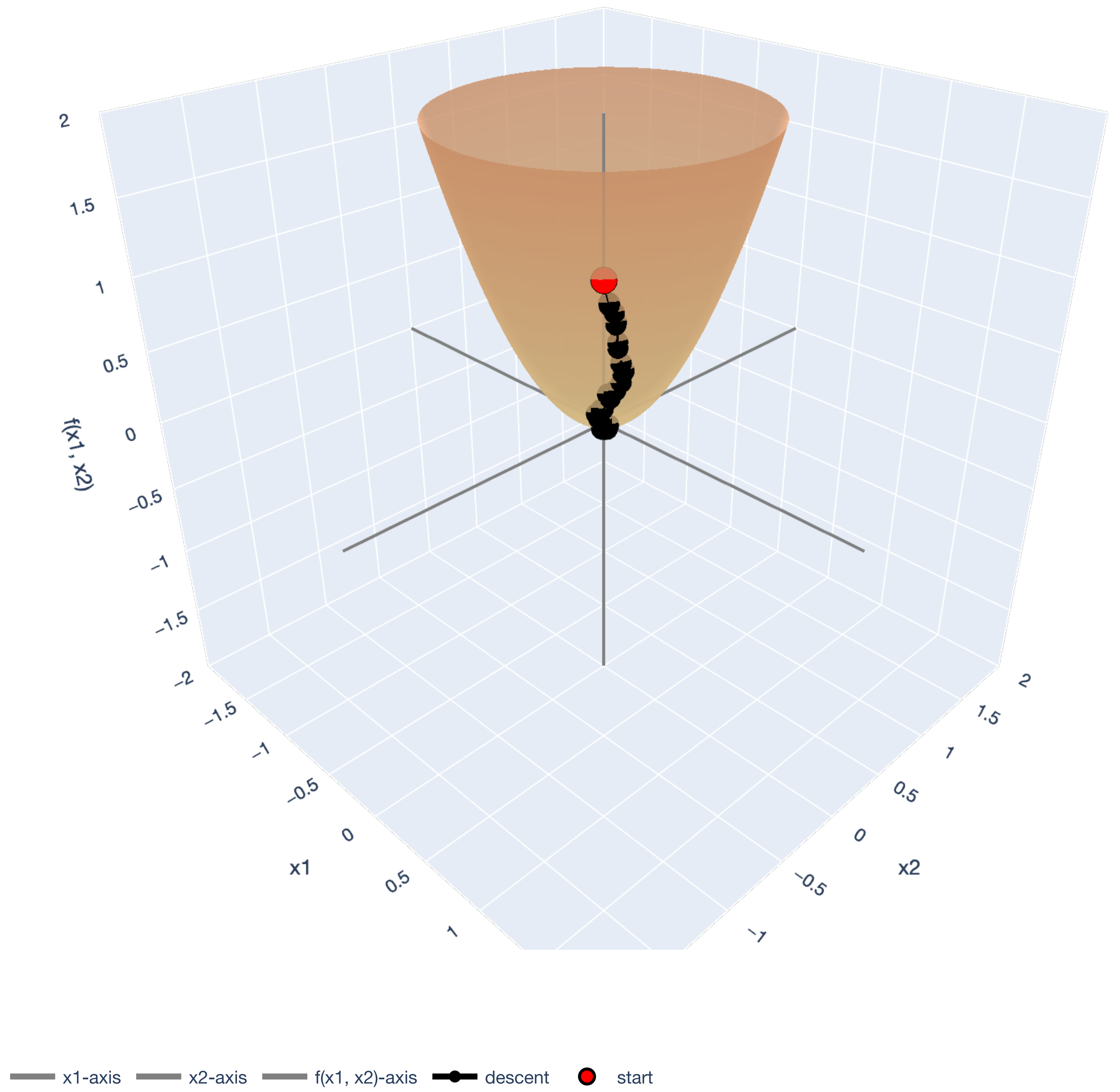
For $t = 1,2,3,\ldots$

Sample $B = \{i_1, \ldots, i_k\}$ uniformly from all k-subsets of $\{1,\ldots,n\}$.

Compute:

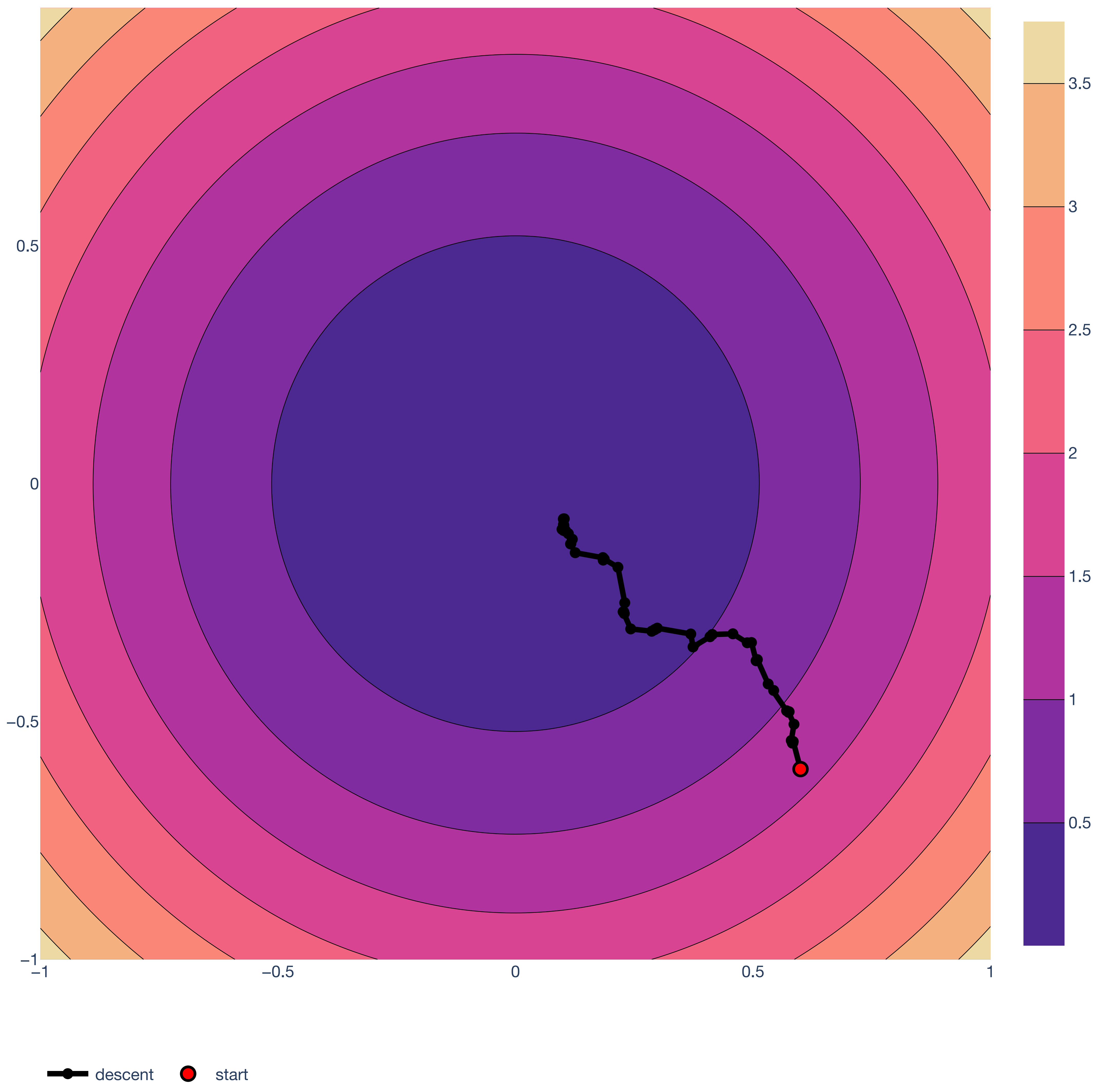
$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \eta \widehat{\nabla f(\mathbf{w})} = \mathbf{w}_{t-1} - \frac{\eta}{k} \sum_{j=1}^{k} \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_{i_j} - y_{i_j})^2$$

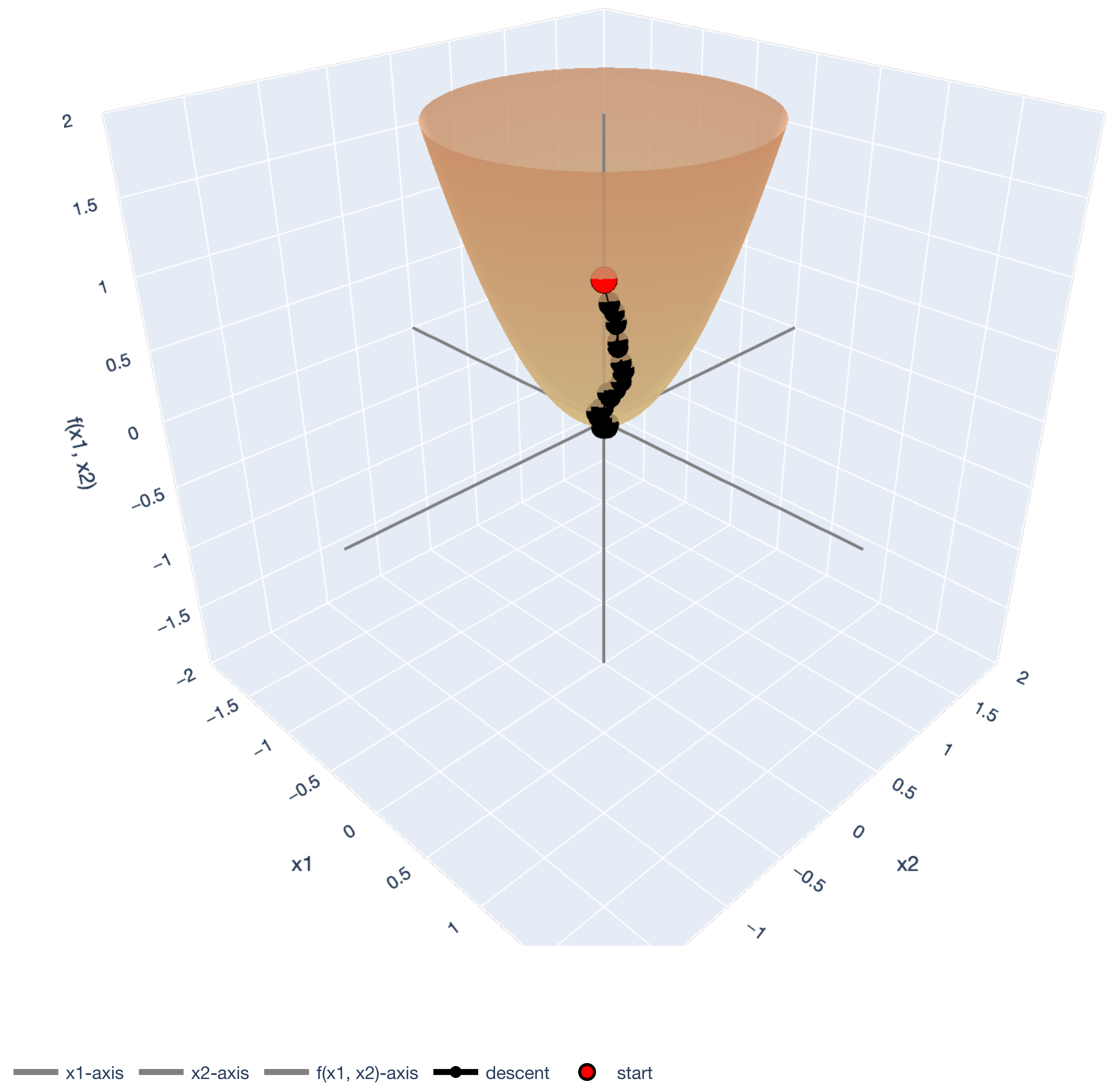
If $\mathbf{w}_t - \mathbf{w}_{t-1}$ is sufficiently small, then **return** $\frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{y}\|^2$.



Stochastic Gradient Descent

Minibatch SGD





Gauss-Markov Theorem

OLS as “optimal”

“Optimality” of OLS

Intuition

We evaluate statistical estimators $\hat{\theta}_n$ through their bias and variance, which make up their mean squared error:

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{Bias}(\hat{\theta}_n)^2 + \mathrm{Var}(\hat{\theta}_n).$$

In what sense is OLS optimal (compared to other possible estimators), with respect to bias and variance?

Gauss-Markov Theorem

Intuition

$y_i = (w^*)^T x_i + \varepsilon_i$

Recall our model of errors:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon, \text{ where } \mathbb{E}[\epsilon] = \mathbf{0} \text{ and } \mathrm{Var}(\epsilon_i) = \sigma^2 < \infty.$$

We will claim that the OLS estimator

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

has the *lowest variance* within the class of *linear, unbiased* estimators.

Gauss-Markov Theorem

Fixed Design Assumption

Recall our model of errors:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon, \text{ where } \mathbb{E}[\epsilon] = \mathbf{0} \text{ and } \mathrm{Var}(\epsilon_i) = \sigma^2 < \infty.$$

We will assume that $\mathbf{X} \in \mathbb{R}^{n \times d}$ is *fixed* to make our derivation easier (we can also avoid this by taking *conditional* expectations/variances with respect to $\mathbf{X}$).

Note: This still means that $\mathbf{y}$ is random because ϵ is random.

Gauss-Markov Theorem

Linear Estimator

$\hat{w} = \begin{bmatrix} \hat{w}_1 \\ \vdots \\ \hat{w}_d \end{bmatrix}$

Recall our model of errors:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon, \text{ where } \mathbb{E}[\epsilon] = \mathbf{0} \text{ and } \mathrm{Var}(\epsilon_i) = \sigma^2 < \infty.$$

We want to estimate $\mathbf{w}^*$, using $\mathbf{X}$ and $\mathbf{y}$. A ***linear estimator*** of entry w_i^* is a linear combination of $y_1, \ldots, y_n$:

$$\hat{w}_i^* = c_{1i}y_1 + \ldots + c_{ni}y_n.$$

$\hat{w} = Cy$

$C \in \mathbb{R}^{n \times d}$

The OLS estimator is clearly a linear estimator:

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

C

Gauss-Markov Theorem

"Greater Than" for Matrices

Random vector: $\mathrm{Var}(w) \in \mathbb{R}^{d\times d}$.

We need to compare the variances of random vectors, $\mathrm{Var}(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$.

Recall that, for random vectors, $\mathrm{Var}(\mathbf{w})$ is given by a positive semidefinite *covariance matrix.* For PSD matrices, the ***Loewner order*** imposes an ordering:

$$\mathbf{A} \leq \mathbf{B} \text{ means that } \mathbf{B} - \mathbf{A} \text{ is PSD.}$$

$$\mathbf{A} < \mathbf{B} \text{ means that } \mathbf{B} - \mathbf{A} \text{ is positive definite.}$$

$a \leq b \Rightarrow b - a \geq 0$

$a < b \Rightarrow b - a > 0$

They are ordered in the sense that their quadratic forms obey the ordering:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \mathbf{x}^\top \mathbf{B}\mathbf{x}.$$

Gauss-Markov Theorem

Theorem Statement

Theorem (Gauss-Markov Theorem). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be fixed and let $\mathbf{y} \in \mathbb{R}^n$ be given entry-wise by the linear error model:

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

where $\epsilon \in \mathbb{R}^n$ is a random vector with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$ and each ϵ_i is independent. Let $\tilde{\mathbf{w}} \in \mathbb{R}^d$ be any linear estimator of $\mathbf{w}^*$, with entries:

$$\tilde{w}_i = c_{1i}y_1 + \ldots + c_{ni}y_n,$$

$\Rightarrow \tilde{w} = Cy$ for some $C \in \mathbb{R}^{d \times n}$

such that $\tilde{\mathbf{w}}$ is unbiased, i.e. $\mathbb{E}[\tilde{\mathbf{w}}] = \mathbf{w}^*$. Then, the OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ has variance (and, thus, mean squared error) no larger than $\tilde{\mathbf{w}}$:

$$\mathrm{Var}(\hat{\mathbf{w}}) = \mathrm{Var}(\tilde{\mathbf{w}}) + \mathbf{A}, \text{ where } \mathbf{A} \in \mathbb{R}^{d \times d} \text{ is some PSD matrix.}$$

$\mathrm{Var}(\hat{w}) \leq \mathrm{Var}(\tilde{w})$ in Loewner order

Gauss-Markov Theorem

Proof

Step 1: Formally state the “other” linear estimator.

Suppose that $\tilde{\mathbf{w}} \in \mathbb{R}^d$ is another linear estimator of $\mathbf{w}^*$. We can write it as:

$$\tilde{\mathbf{w}} = \mathbf{C}\mathbf{y}, \text{ where } \mathbf{C} \in \mathbb{R}^{n \times d}.$$

Without loss of generality, let:

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D} \text{ where } \mathbf{D} \in \mathbb{R}^{d \times n}.$$

Gauss-Markov Theorem

Proof

Step 2: We know that $\tilde{\mathbf{w}}$ is an unbiased estimator, so enforce $\mathbb{E}[\tilde{\mathbf{w}}] = \mathbf{w}^*$.

Calculate the expectation of $\tilde{\mathbf{w}}$.

C   Error model

$$\begin{aligned}
\mathbb{E}[\tilde{\mathbf{w}}] &= \mathbb{E}[\mathbf{C}\mathbf{y}] \\
&= \mathbb{E}\left[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})(\mathbf{X}\mathbf{w}^* + \epsilon)\right] && \text{(Step 1)} \\
&= ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})\mathbf{X}\mathbf{w}^* + ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})\mathbb{E}[\epsilon] \\
&= ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})\mathbf{X}\mathbf{w}^* && \mathbb{E}[\epsilon] = 0 \\
&= \mathbf{w}^* + \mathbf{D}\mathbf{X}\mathbf{w}^*
\end{aligned}$$

→ 0

But because we assumed $\tilde{\mathbf{w}}$ is unbiased, → $\mathbb{E}[\tilde{w}] = w^*$

$$\mathbf{w}^* + \mathbf{D}\mathbf{X}\mathbf{w}^* = \mathbf{w}^* \implies \mathbf{D}\mathbf{X} = \mathbf{0}.$$

Gauss-Markov Theorem

Proof

$$\mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\mathbf{X}\mathbf{w}^* + \epsilon) = \mathrm{Var}(\epsilon) = \mathbb{E}[\epsilon\epsilon^\top]$$

Step 3: Using the fact that $\mathbf{DX} = \mathbf{0}$ from Step 2, show $\mathrm{Var}(\hat{\mathbf{w}}) \leq \mathrm{Var}(\tilde{\mathbf{w}})$.

Finally, let's analyze the variance of $\tilde{\mathbf{w}}$:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\mathbf{w}}) &= \mathrm{Var}(\mathbf{C}\mathbf{y}) && \tilde{w}\ \text{def.} \\
&= \mathbf{C}\,\mathrm{Var}(\mathbf{y})\mathbf{C}^\top && \text{variance of random vector: } \mathbb{E}[xx^\top] \\
&= \sigma^2 \mathbf{C}\mathbf{I}_{n\times n}\mathbf{C}^\top && (\epsilon_i \text{ are independent}) \\
&= \sigma^2((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{D})(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}^\top) && \text{distribute} \\
&= \sigma^2((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top + \mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^\top) \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{D}\mathbf{X})^\top + \sigma^2\mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}^\top && \text{distributing} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}^\top && \text{(Step 2)} \quad \mathbf{DX} = 0 \\
&= \mathrm{Var}(\hat{\mathbf{w}}) + \sigma^2\mathbf{D}\mathbf{D}^\top && \text{(Variance of OLS estimator)}
\end{aligned}
$$

PSD $\implies \mathrm{Var}(\hat{w}) \leq \mathrm{Var}(\tilde{w})$

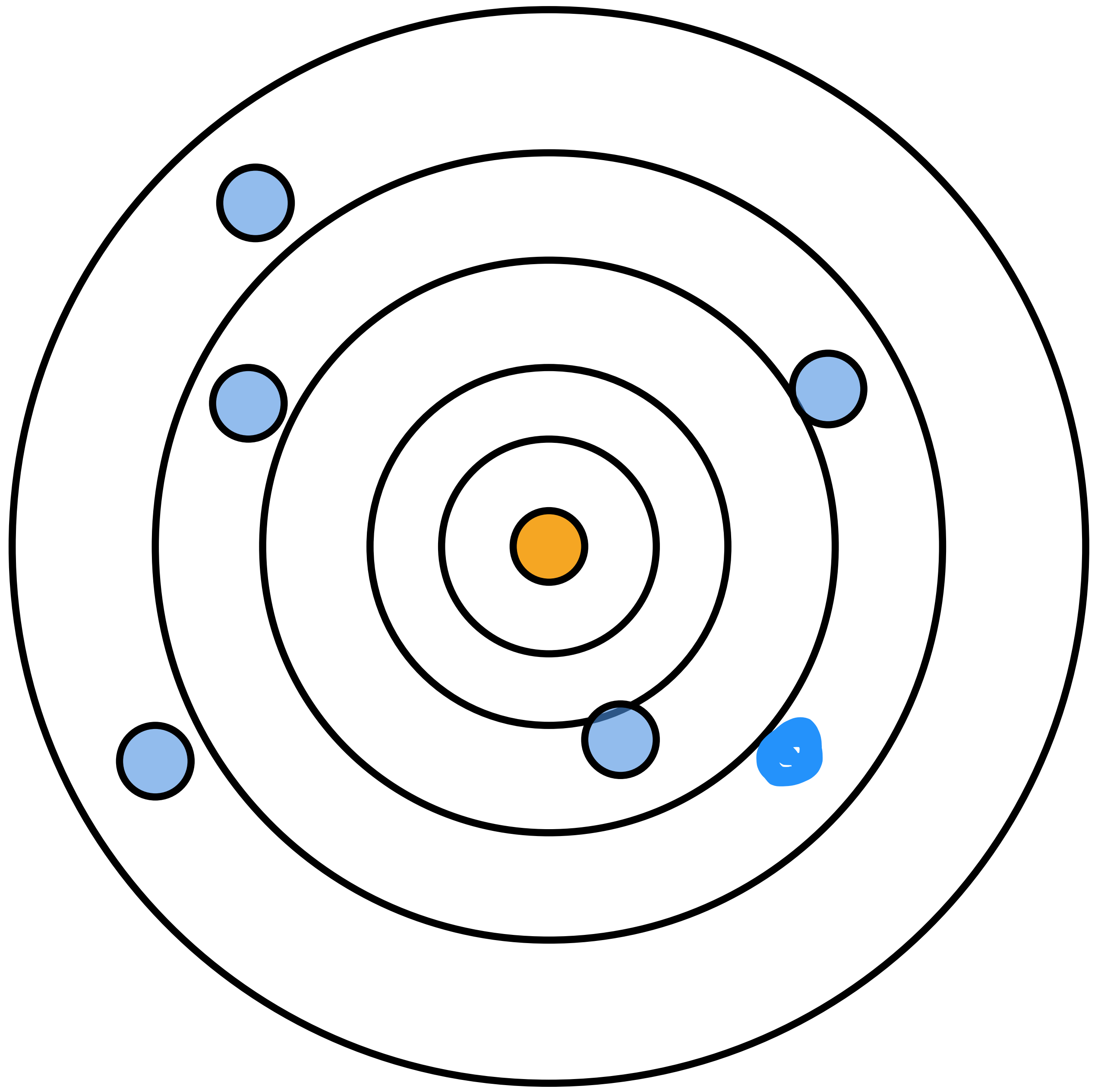
Mean Squared Error

Trading bias for reduction in variance

The Gauss-Markov Theorem states that $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the smallest variance out of *all* linear estimators with no bias.

Recall the MSE is how we evaluate an estimator:

$$\mathrm{MSE}(\hat{\mathbf{w}}) = \mathrm{Bias}(\hat{\mathbf{w}})^2 + \mathrm{Var}(\hat{\mathbf{w}}).$$



But unbiasedness might not always be a good thing if the variance is high!

Mean Squared Error

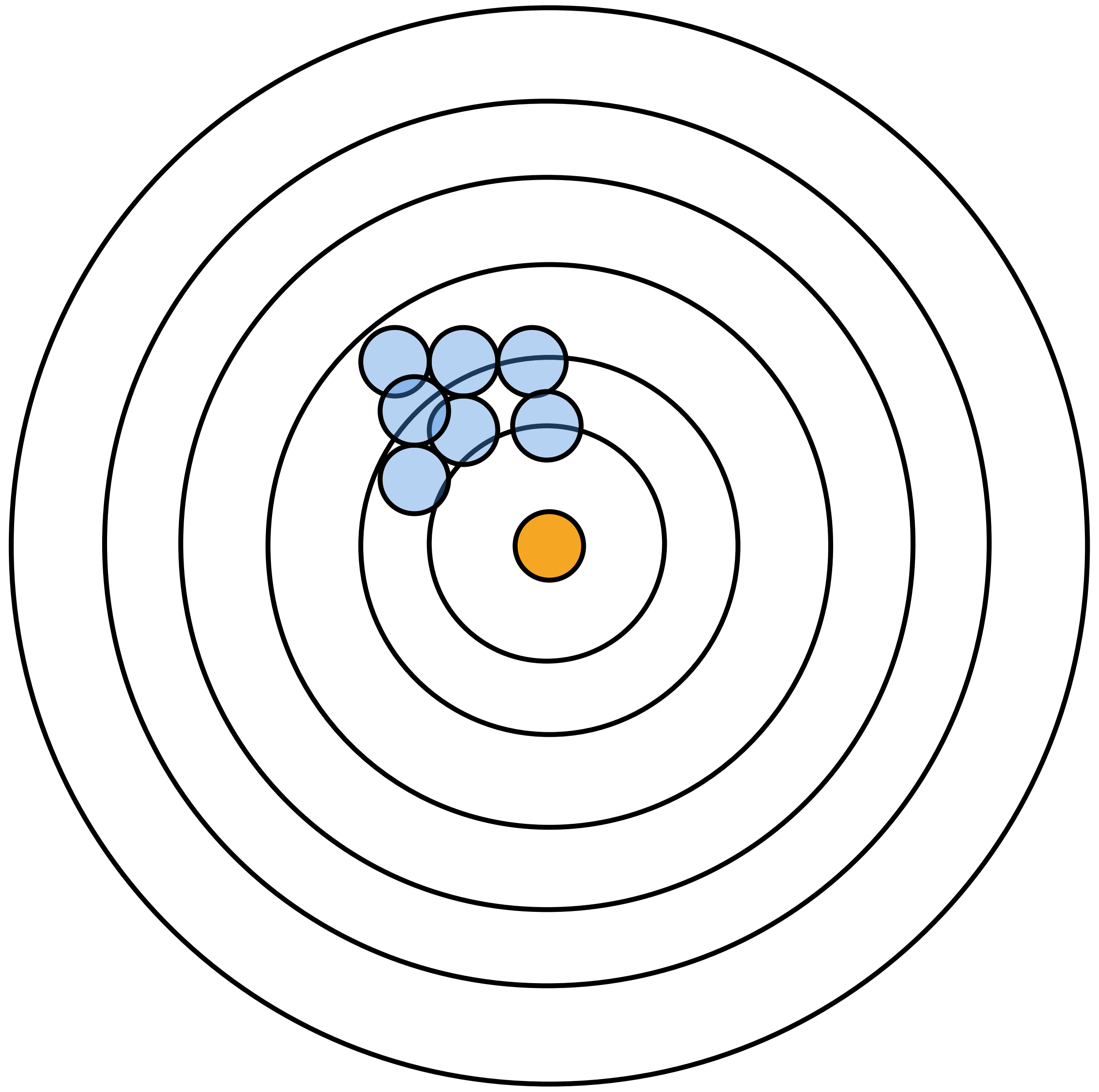
Trading bias for reduction in variance

The Gauss-Markov Theorem states that $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ has the smallest variance out of *all* linear estimators with no bias.

Recall the MSE is how we evaluate an estimator:

$$\text{MSE}(\hat{\mathbf{w}}) = \text{Bias}(\hat{\mathbf{w}})^2 + \text{Var}(\hat{\mathbf{w}}).$$

Can we trade a bit of bias for a reduction in variance?



Mean Squared Error

Trading bias for reduction in variance

minimize $\|Xw - y\|^2$, $\|w\| \le 1$ / 2

$f(w) = \|Xw - y\|^2 \Downarrow f(w) = \|Xw - y\|^2 + \gamma \|w\|^2$

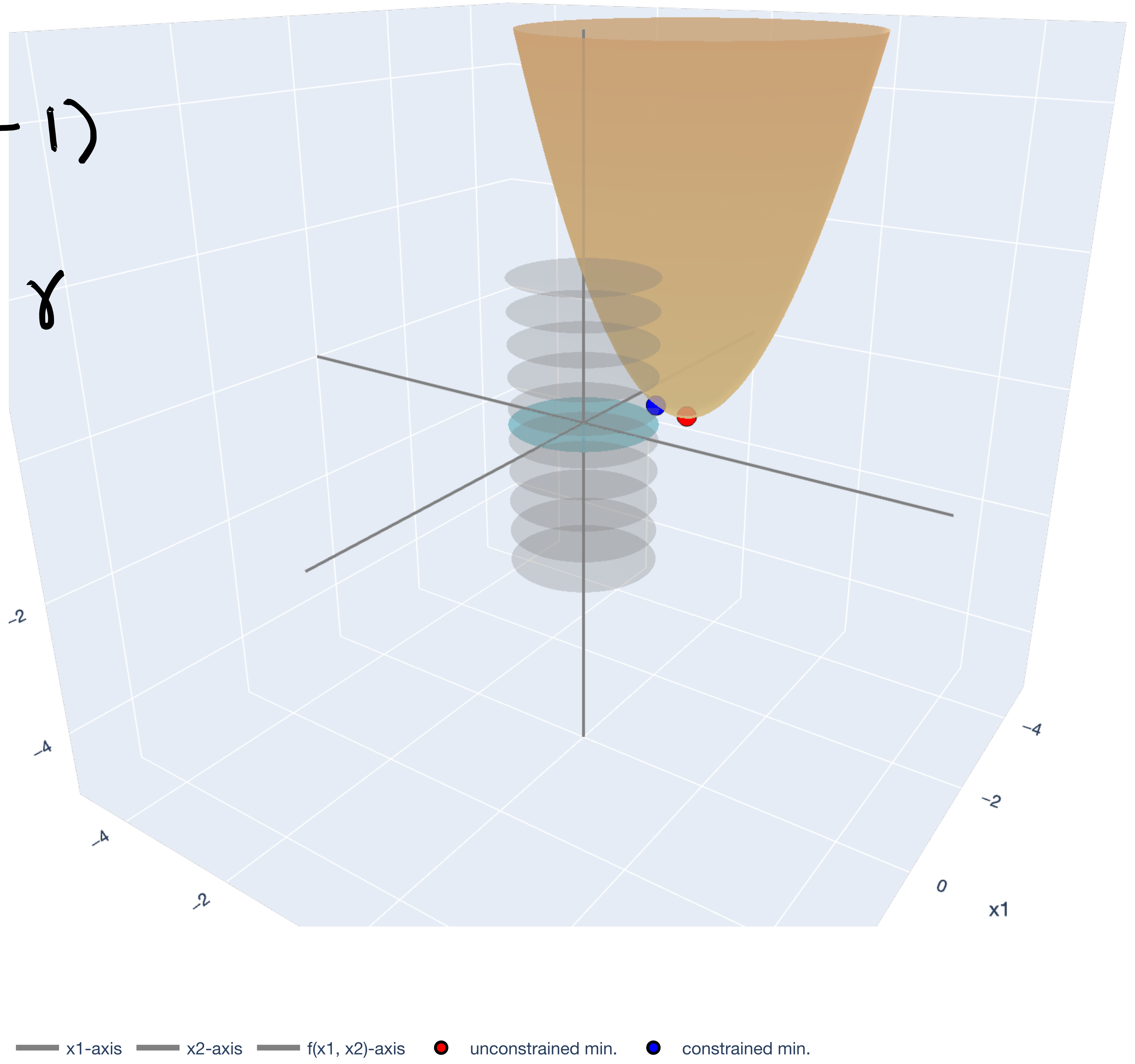
The *ridge regression* estimator: $L(w, \gamma) = \|Xw - y\|^2 + \gamma(\|w\| - 1) = \|Xw - y\|^2 + \gamma\|w\| - \gamma$

$$\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

for $\gamma > 0$ does exactly that! The γ parameter controls the bias-variance tradeoff.

Bias comes from "shrinking" the $\hat{\mathbf{w}}$ coefficients to zero.

Variance reduction comes from constraining the coefficients to preferably come from a constrained ball.



Regression

Statistical analysis of risk

Statistics of OLS

Theorem

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

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

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

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

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

Bias and Variance of OLS

Corollaries from Theorem

Under the error model:

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

OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the following statistical properties:

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

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

This implies that, as an estimator of $\mathbf{w}^*$,

$$\mathrm{Bias}(\hat{\mathbf{w}}) = 0$$

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

Regression

Setup, with randomness

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

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

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

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

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

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

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

This is what we've been doing!

Statistical Analysis of Risk

Breaking down generalization error

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

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

Statistical Analysis of Risk

Breaking down generalization error

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

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

This was the notion of ***risk*** or ***generalization error*** — how well we do on a new, randomly drawn example.

Can we analyze this in terms of OLS?

Statistical Analysis of Risk

Breaking down generalization error

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

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

$$\implies R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0)^2]$$

What is random in the above expectation?

$\mathbf{x}_0$ is random because it's a new example $\mathbf{x}_0 \sim \mathbb{P}_{\mathbf{x}}$.

y_0 is random because it's a new label $y_0 \sim \mathbb{P}_y$.

$\hat{\mathbf{w}}$ is random because it depends on the training data $\mathbf{X}$ and $\mathbf{y}$.

Statistical Analysis of Risk

Law of Total Expectation

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

$$R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0)^2].$$

Let $\mathbf{X}, \mathbf{y}$ be randomly drawn training data, which the estimator $\hat{\mathbf{w}}$ depends on.
By the *tower rule/law of total expectation*:

$$R(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{x}_0}\left[\mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right]\right]$$

Statistical Analysis of Risk

Law of Total Expectation

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

$$R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0)^2].$$

Let $\mathbf{X}, \mathbf{y}$ be randomly drawn training data, which the estimator $\hat{\mathbf{w}}$ depends on.
By the *tower rule/law of total expectation*:

$$R(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{x}_0}\left[\mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right]\right]$$

Training Data randomness (X, Y)

x_0's randomness from $\mathbb{P}_x$.

Let's analyze this quantity!

y_0 randomness (ε's randomness)

Statistical Analysis of Risk

Analyzing the risk

$$R(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{x}_0}\left[\mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right]\right]$$

Denote: $R(\hat{\mathbf{w}} \mid \mathbf{x}_0) := \mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right]$

Statistical Analysis of Risk

Analyzing the risk

$$
\begin{aligned}
R(\hat{\mathbf{w}} \mid \mathbf{x}_0) := \mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right] \\
= \mathrm{Var}(y_0 \mid x_0) + \mathbb{E}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - \mathbb{E}[\hat{\mathbf{w}}^\top \mathbf{x}_0]\right)^2\right] + \left(\mathbb{E}[\hat{\mathbf{w}}^\top \mathbf{x}_0] - \mathbf{x}_0^\top \mathbf{w}*\right)^2 \\
= \mathrm{Var}(y_0 \mid x_0) + \mathrm{Var}(\hat{\mathbf{w}}^\top \mathbf{x}_0) + \mathrm{Bias}(\hat{\mathbf{w}}^\top \mathbf{x}_0)^2 \\
= \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0 \sigma^2\right]
\end{aligned}
$$

Note: We are conditioning on $\mathbf{x}_0$, so the only random quantity in the last term is $\mathbf{X}^\top \mathbf{X}$.

Statistical Analysis of Risk

Analyzing the risk

$$
\begin{aligned}
R(\hat{\mathbf{w}} \mid \mathbf{x}_0) &:= \mathbb{E}_{y_0}\left[\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0\right)^2 \mid y_0\right] \mid \mathbf{x}_0\right] \\
&= \mathrm{Var}(y_0 \mid x_0) + \mathbb{E}\left[\left(\hat{\mathbf{w}}^\top \mathbf{x}_0 - \mathbb{E}[\hat{\mathbf{w}}^\top \mathbf{x}_0]\right)^2\right] + \left(\mathbb{E}[\hat{\mathbf{w}}^\top \mathbf{x}_0] - \mathbf{x}_0^\top \mathbf{w}*\right)^2 \\
&= \mathrm{Var}(y_0 \mid x_0) + \mathrm{Var}(\hat{\mathbf{w}}^\top \mathbf{x}_0) + \mathrm{Bias}(\hat{\mathbf{w}}^\top \mathbf{x}_0)^2 \\
&= \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0 \sigma^2\right]
\end{aligned}
$$

Note: We are conditioning on $\mathbf{x}_0$, so the only random quantity in the last term is $\mathbf{X}^\top \mathbf{X}$.

Statistical Analysis of Risk

Analyzing the risk

$R(\hat{\mathbf{w}} \mid \mathbf{x}_0) = \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0\sigma^2\right]$ from the previous slide.

Consider the ***empirical covariance matrix*** $\frac{1}{n}(\mathbf{X}^\top\mathbf{X})$. If n is large and $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, then $\mathbf{X}^\top\mathbf{X} \to n\Sigma$, where $\Sigma := \mathrm{Var}(\mathbf{x}) \in \mathbb{R}^{d\times d}$ is the covariance matrix of the features.

$$R(\hat{\mathbf{w}} \mid \mathbf{x}_0) = \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0\sigma^2\right] = \sigma^2 + \frac{\sigma^2}{n}\mathbf{x}_0^\top\Sigma^{-1}\mathbf{x}_0$$

Now, take the expectation over all $\mathbf{x}_0 \sim \mathbb{P}_{\mathbf{x}}$:

$$\mathbb{E}_{\mathbf{x}_0}[R(\hat{\mathbf{w}} \mid \mathbf{x}_0)] = \sigma^2 + \frac{\sigma^2}{n}\mathbb{E}_{\mathbf{x}_0}\left[\mathbf{x}_0^\top\Sigma^{-1}\mathbf{x}_0\right]$$

Trace

Definition and the "trace trick"

For a $d \times d$ square matrix $\mathbf{A}$, the ***trace*** of $\mathbf{A}$, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of its diagonal entries:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{d} a_{ii} = a_{11} + \ldots + a_{dd}.$$

"Trace trick:" For any quadratic form $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$,

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

Statistical Analysis of Risk

Analyzing the risk

$R(\hat{\mathbf{w}} \mid \mathbf{x}_0) = \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0 \sigma^2\right]$ from the previous slide.

Consider the ***empirical covariance matrix*** $\frac{1}{n}(\mathbf{X}^\top \mathbf{X})$. If n is large and $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, then $\mathbf{X}^\top \mathbf{X} \to n\Sigma$, where $\Sigma := \mathrm{Var}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is the covariance matrix of the features.

$$R(\hat{\mathbf{w}} \mid \mathbf{x}_0) = \sigma^2 + \mathbb{E}\left[\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0 \sigma^2\right] = \sigma^2 + \frac{\sigma^2}{n} \mathbf{x}_0^\top \Sigma^{-1} \mathbf{x}_0$$

Now, take the expectation over all $\mathbf{x}_0 \sim \mathbb{P}_{\mathbf{x}}$:

$$R(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{x}_0}[R(\hat{\mathbf{w}} \mid \mathbf{x}_0)] = \sigma^2 + \frac{\sigma^2}{n} \mathbb{E}_{\mathbf{x}_0}\left[\mathbf{x}_0^\top \Sigma^{-1} \mathbf{x}_0\right]$$

Using the “trace trick,”

$$R(\hat{\mathbf{w}}) = \sigma^2 + \frac{\sigma^2}{n} \mathbf{E}_{\mathbf{x}_0}\left[\mathrm{tr}\left(\Sigma^{-1} \mathbf{x}_0 \mathbf{x}_0^\top\right)\right] = \sigma^2 + \frac{\sigma^2}{n} \mathrm{tr}\left(\Sigma^{-1} \mathbb{E}[\mathbf{x}_0 \mathbf{x}_0^\top]\right) = \sigma^2 + \frac{\sigma^2}{n} \mathrm{tr}\left(\Sigma^{-1} \Sigma\right) = \sigma^2 + \frac{\sigma^2 d}{n}$$

Statistical Analysis of Risk

Theorem Statement

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

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

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

Then, the OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has risk:

$$R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x}_0 - y_0)^2] = \sigma^2 + \frac{\sigma^2 d}{n}.$$

Risk and MSE

Theorem Statement

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

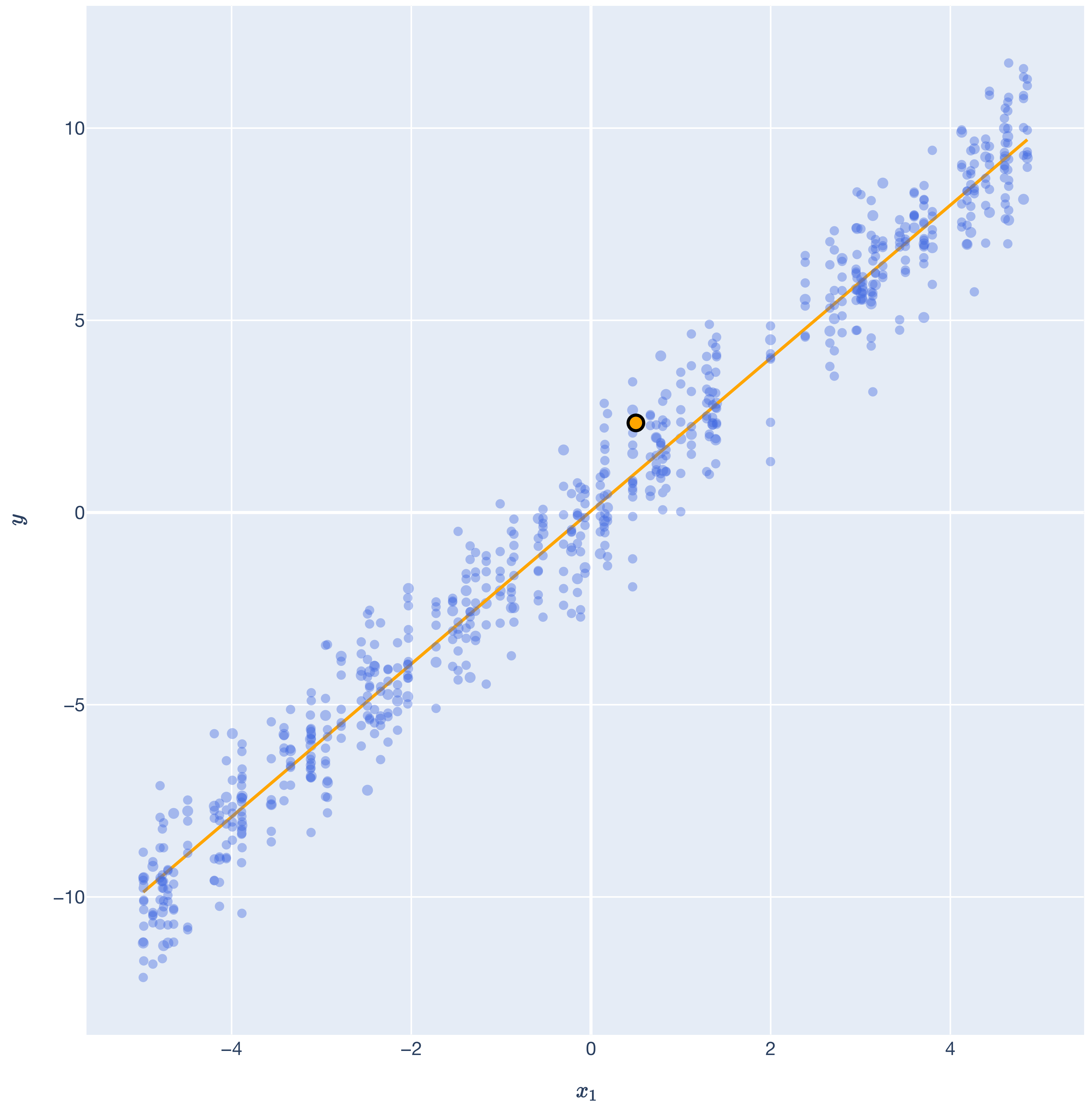
$$y = f(\mathbf{x}) + \epsilon,$$

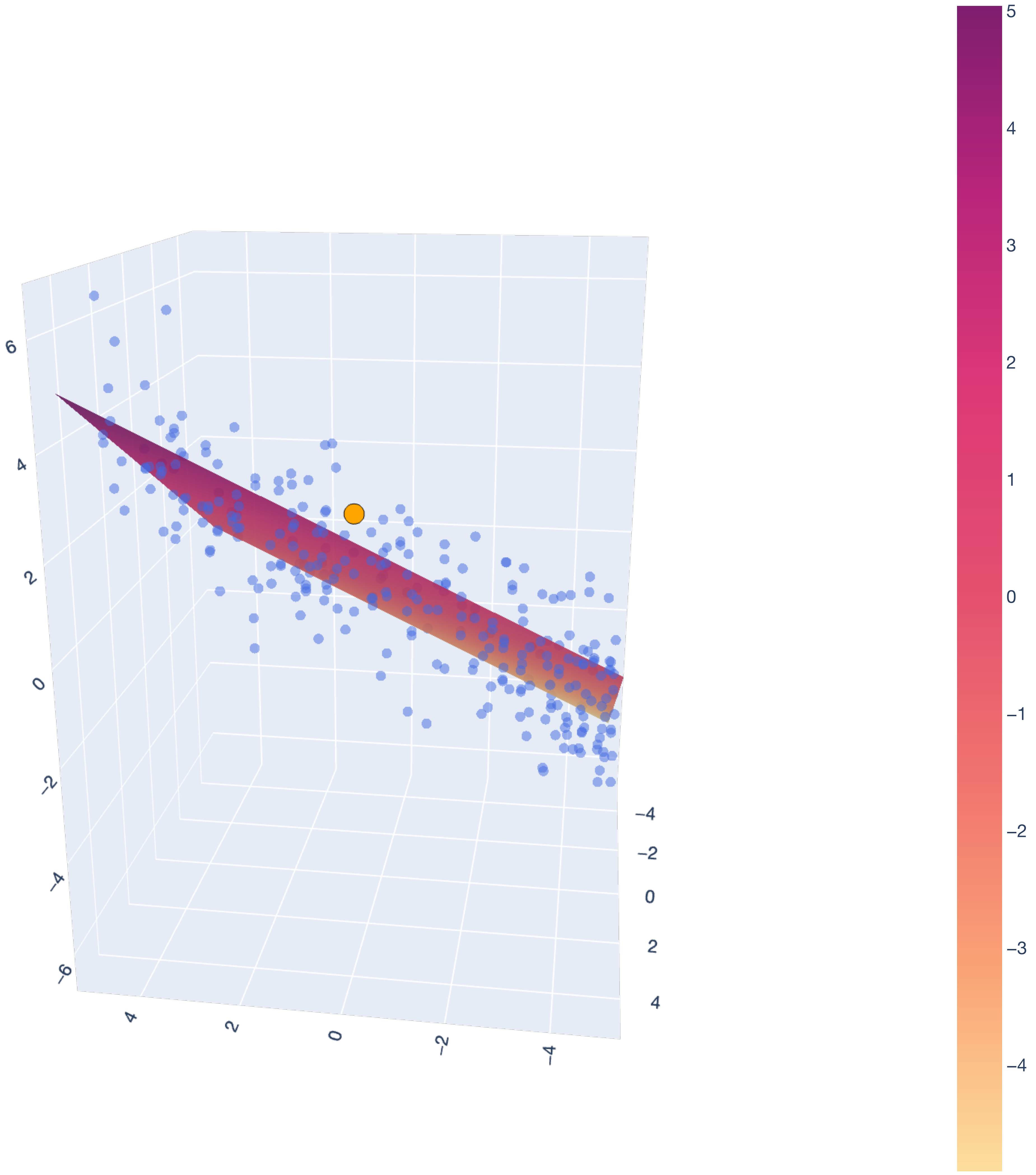
where $f : \mathbb{R}^d \to \mathbb{R}$ and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, independent of $\mathbf{x}$. Consider any linear predictor, $\tilde{f}(\mathbf{x}) = \tilde{\mathbf{w}}^\top \mathbf{x}$, where $\tilde{\mathbf{w}}$ depends on random training data $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Then, for a random $\mathbf{x}_0$, the predictor $\tilde{f}(\mathbf{x}_0)$ is an estimator of $f(\mathbf{x}_0)$, and its risk is:

$$R(\tilde{\mathbf{w}}) = \sigma^2 + \mathrm{MSE}(\tilde{f}(\mathbf{x}_0)).$$

Risk of OLS

$d = 1$ and $d = 2$





Recap

Lesson Overview

Law of Large Numbers. The LLN allows us to move from probability to statistics (reasoning about an *unknown* data generating process using data from that process).

Statistical estimators. We define a *statistical estimator*, which is a function of a collection of random variables (data) aimed at giving a "best guess" at some unknown quantity from some probability distribution.

Bias, variance, and MSE. Two important properties of statistical estimators are their *bias* and *variance*, which are measures of how good the estimator is at guessing the target. These form the estimator's MSE.

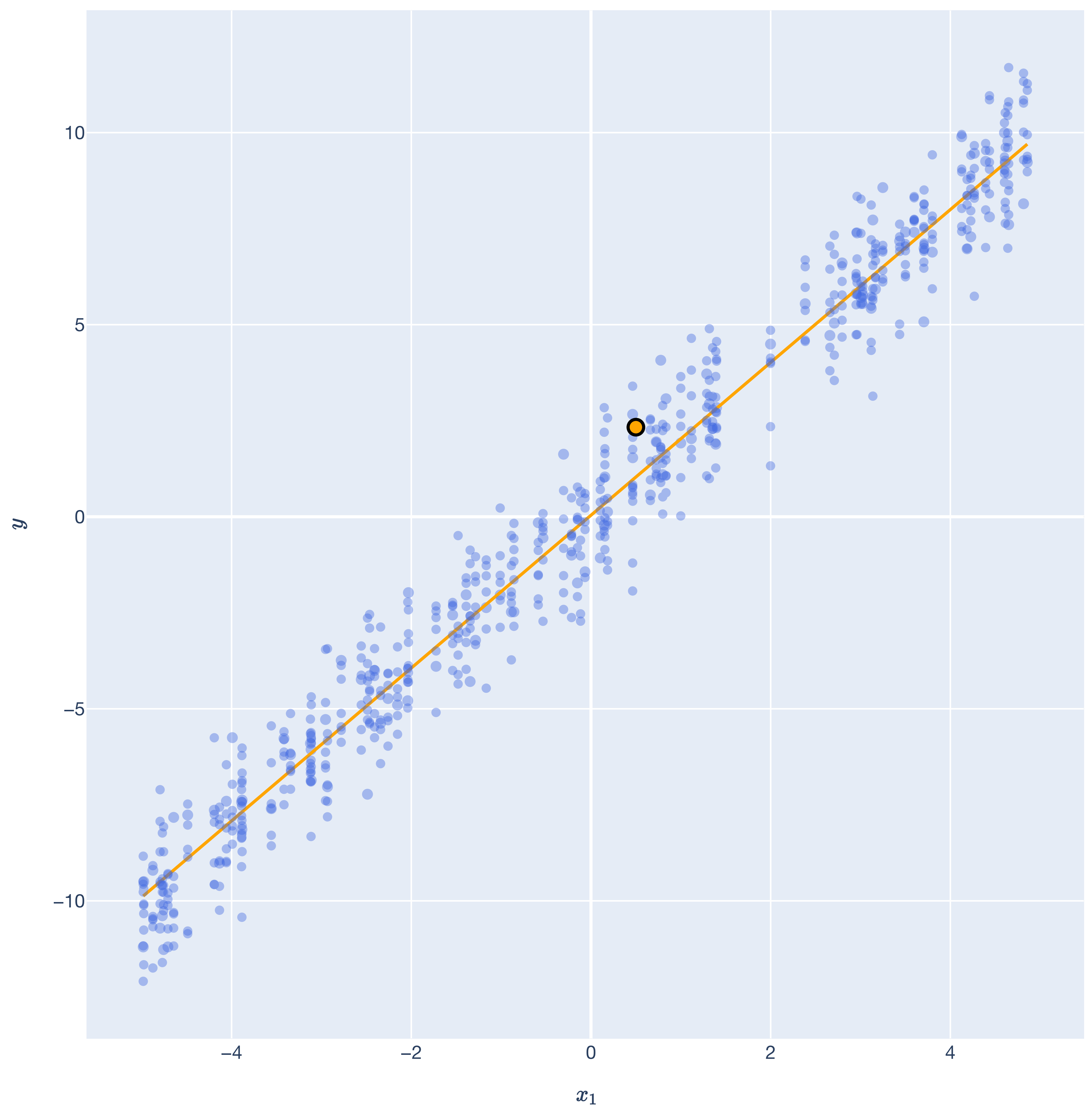
Stochastic gradient descent (SGD). Gradient descent needs to take a gradient over all n training examples, which may be large; SGD *estimates* the gradient to speed up the process.

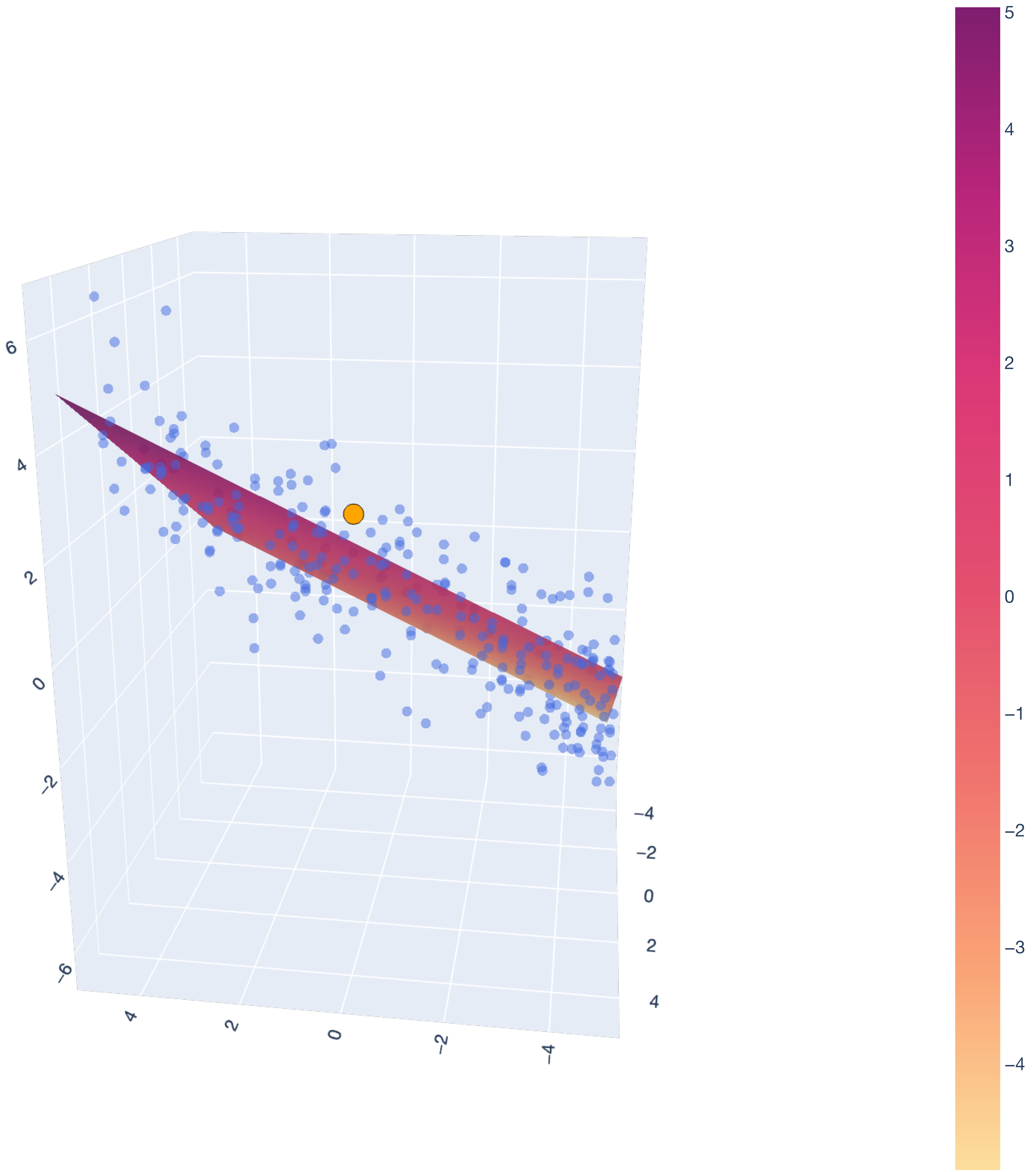
Gauss-Markov Theorem. We show that OLS is the minimum variance estimator in the class of all unbiased, linear estimators.

Statistical analysis of OLS risk. We analyze the *risk* of OLS — how well it's expected to do on future examples drawn from the same distribution it was trained on.

Lesson Overview

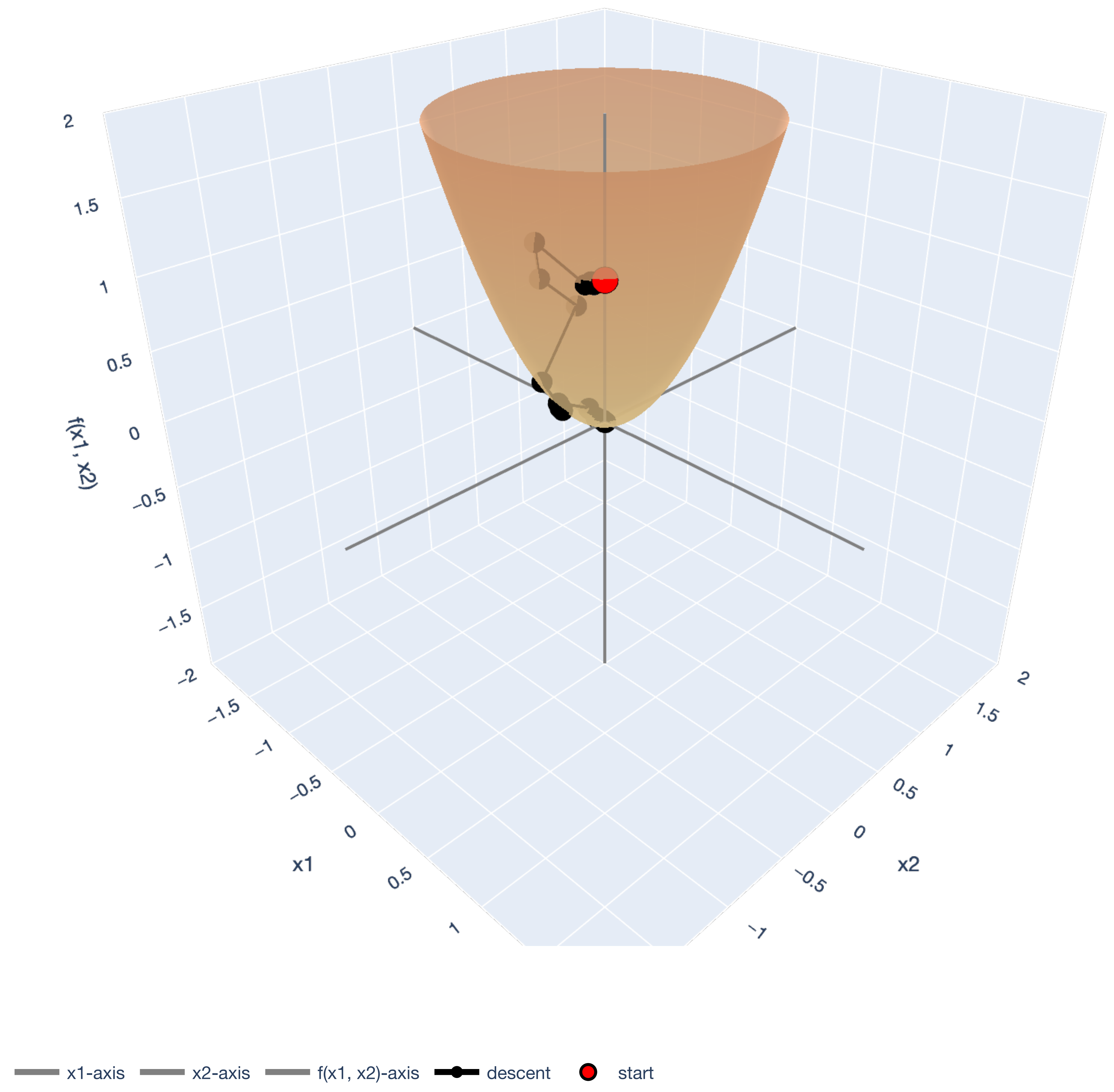
Big Picture: Least Squares

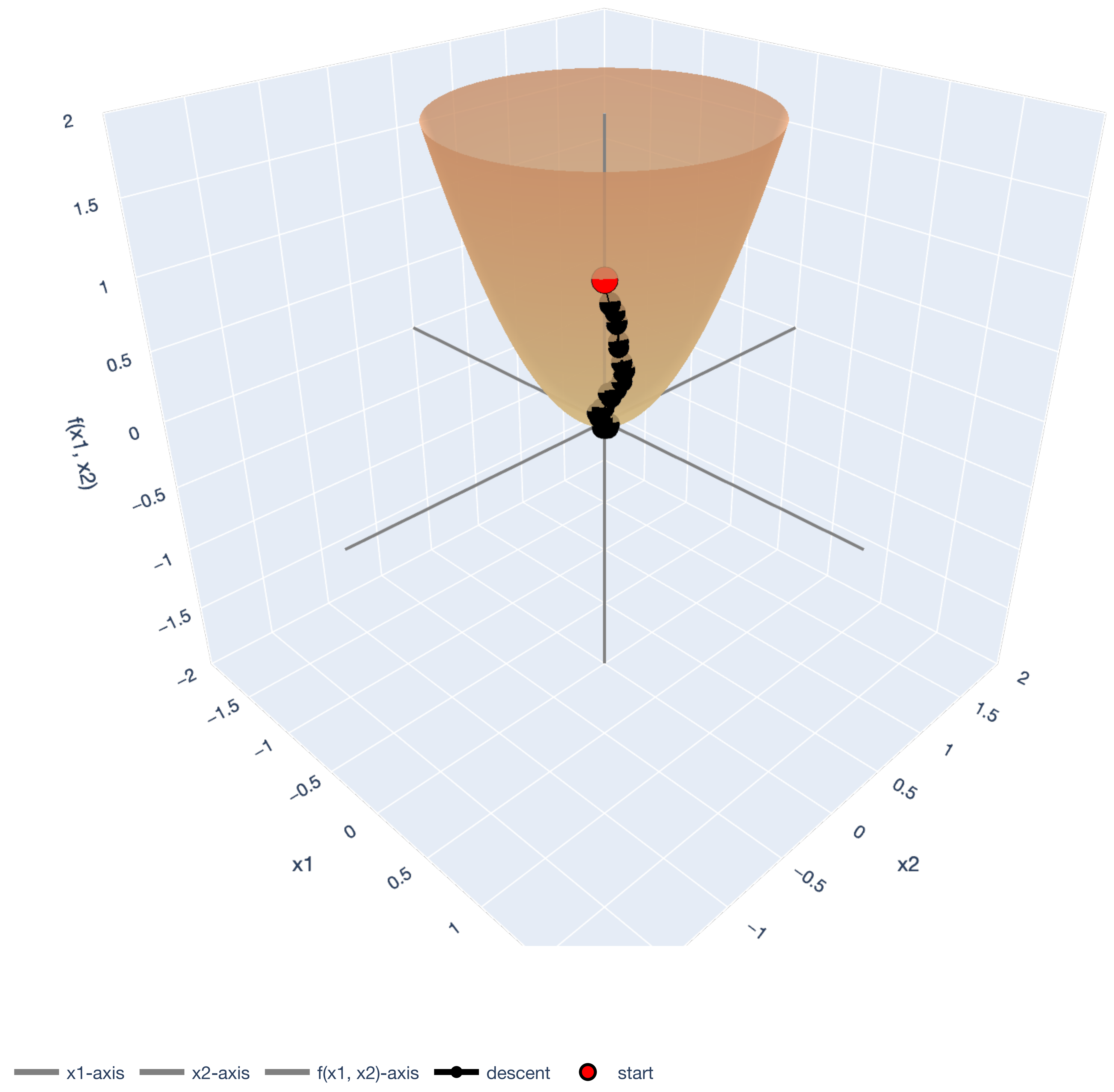




Lesson Overview

Big Picture: Gradient Descent





References

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

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