Math for ML

Week 5.2: Bias, Variance, and Statistical Estimators

By: Samuel Deng

Logistics & Announcements

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.

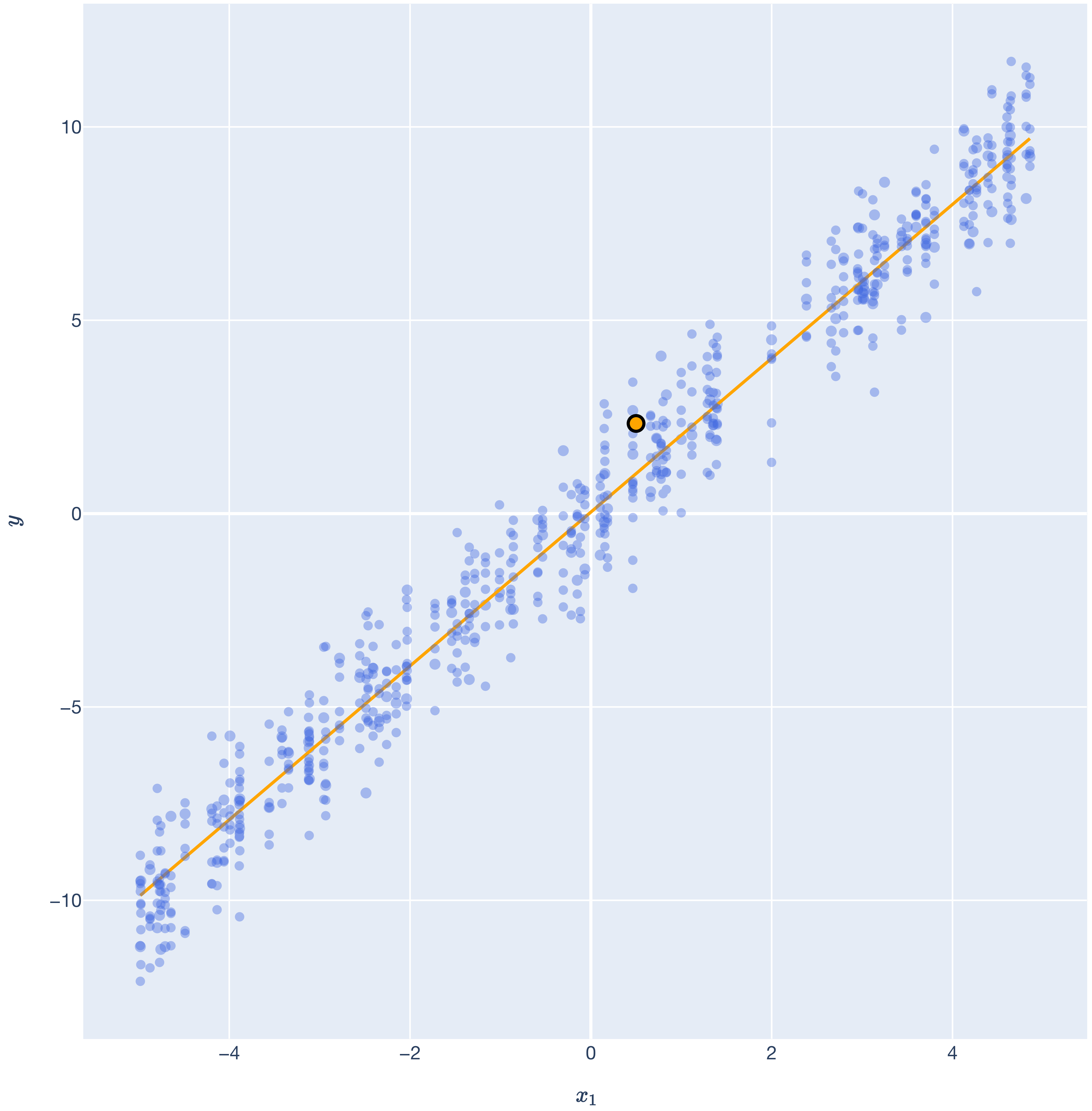
Bias, variance, and MSE. Two important properties of statistical estimators are their *bias* and *variance*, which are measures of how good the estimator is at guessing the target. These form the estimator’s MSE.

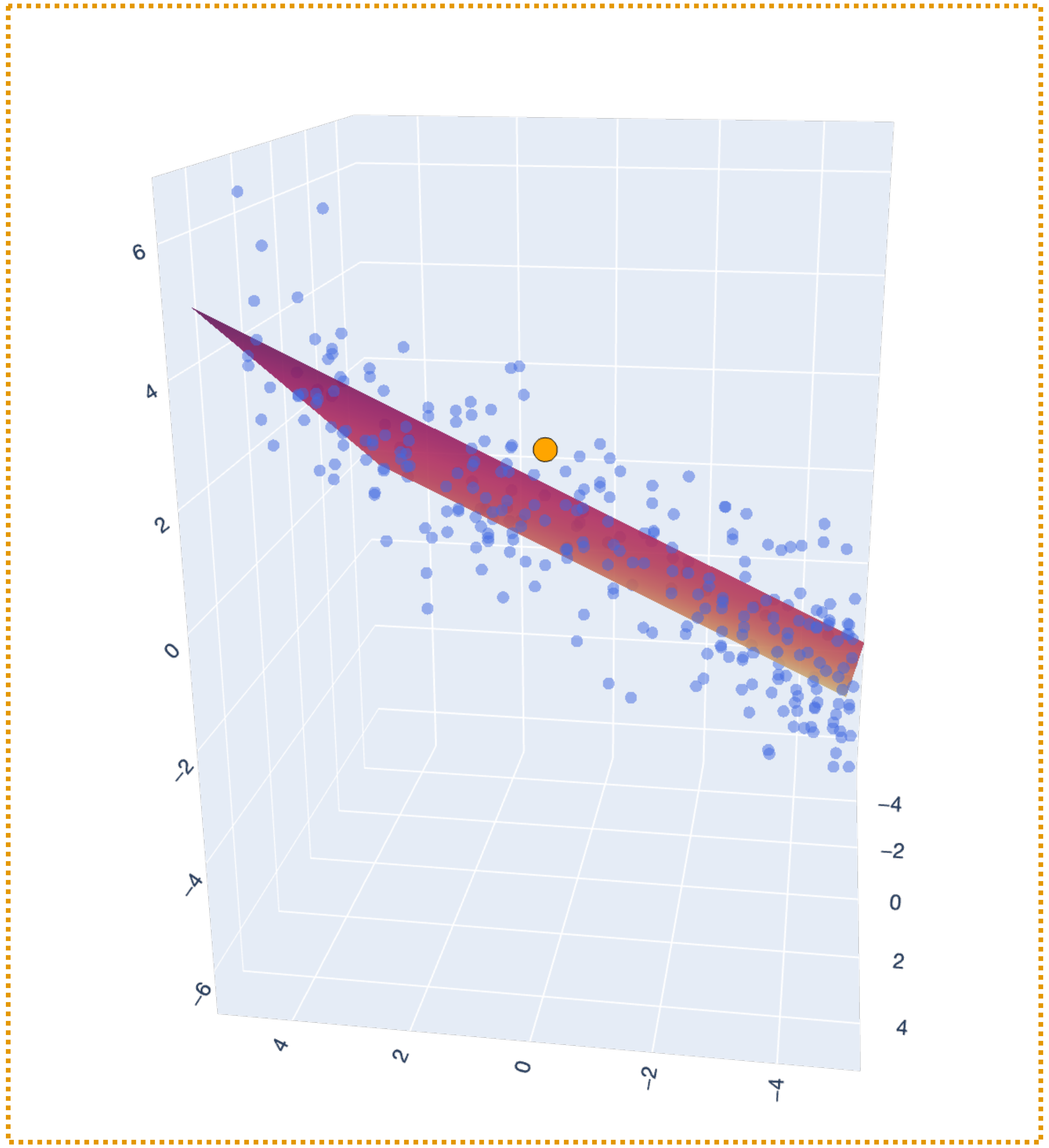
Stochastic gradient descent (SGD). Gradient descent needs to take a gradient over all n training examples, which may be large; SGD *estimates* the gradient to speed up the process.

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

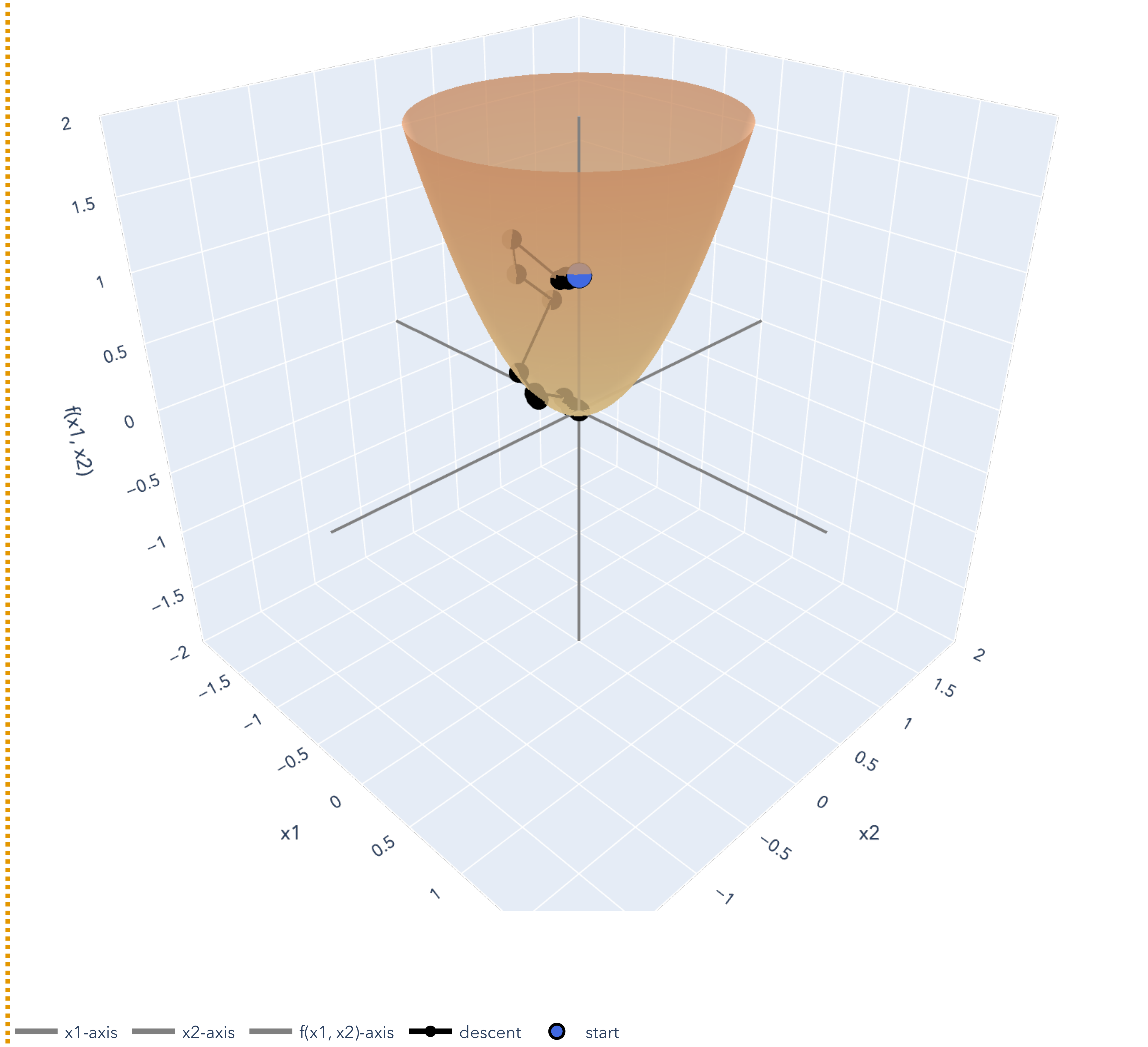
Big Picture: Least Squares

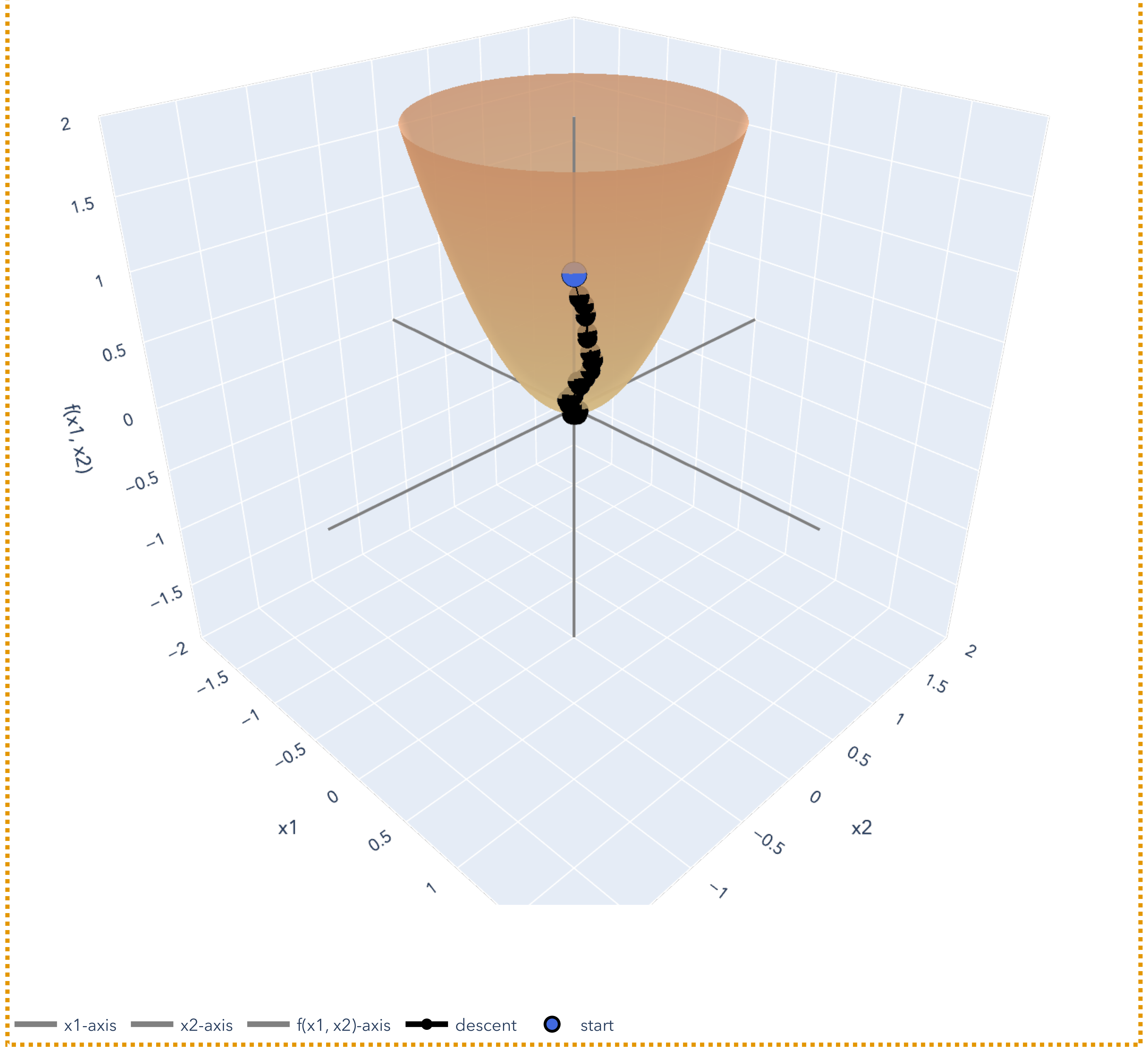




Lesson Overview

Big Picture: Gradient Descent





Law of Large Numbers

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 of probability." We see some data and we try to make inferences about the process that generated the data.

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

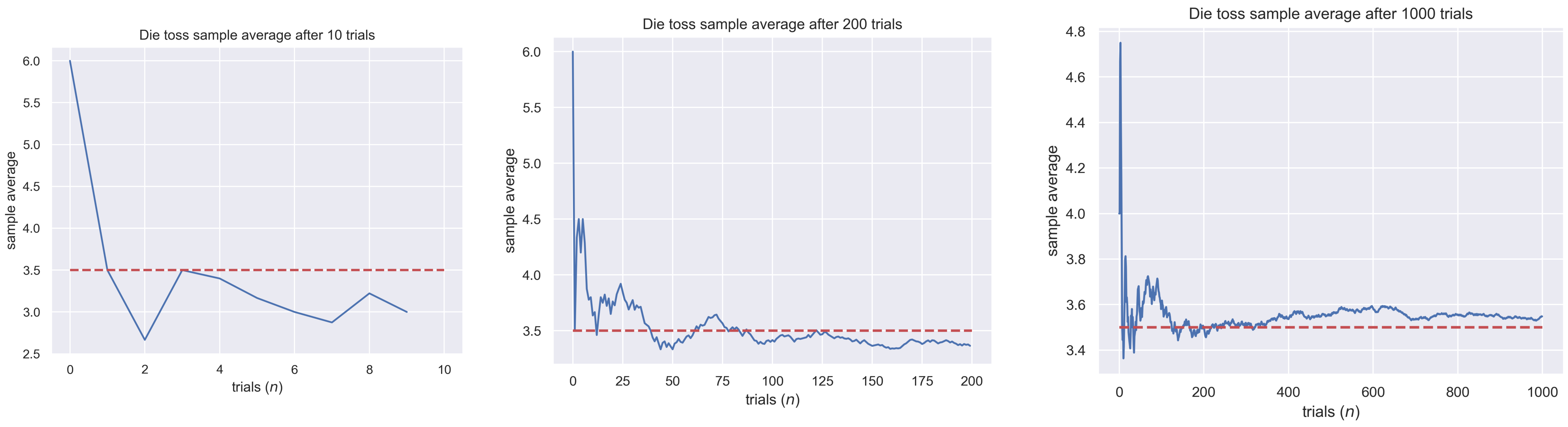
Underlying fact: collecting more and more data gives us sharper conclusions!

Law of Large Numbers

Intuition

Averages of a *large* number of random samples converge to their mean.

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



Independence

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

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

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

and all the X_i have the same distribution.

Very common assumption in ML!

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 $\mu := \mathbb{E}[X_i]$. Their *sample average* is

e.g. X_i is result of die toss i from the same die

If i.i.d. then all have same mean.

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

Then, for any $\epsilon > 0$, the sample average converges to the true mean:

Probability is over the joint distribution of all $X_1, \ldots, X_n$

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

This "kicks in" when n gets very large.

This type of convergence is also called convergence in probability.

Markov's Inequality

Intuition

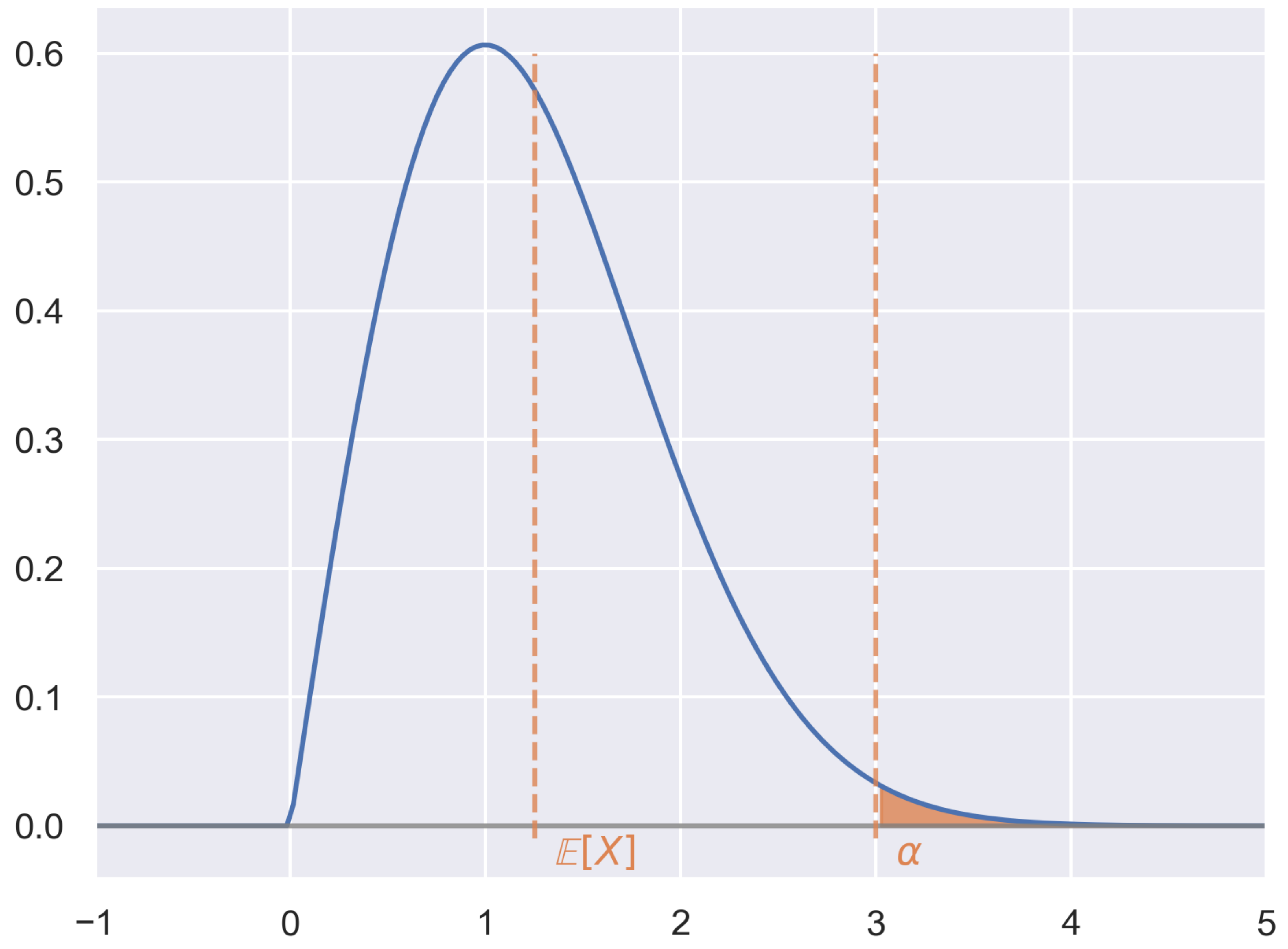
Suppose we have a village where the average salary is \$2 (say). We ask:

What fraction of villagers makes \$10 or more?

Without knowing anything else, Markov's Inequality says:

$$\mathbb{P}(X \geq 10) \leq 2/10 = 0.2.$$

No more than 20% can have more than \$10. Otherwise, we *must* have a higher average!

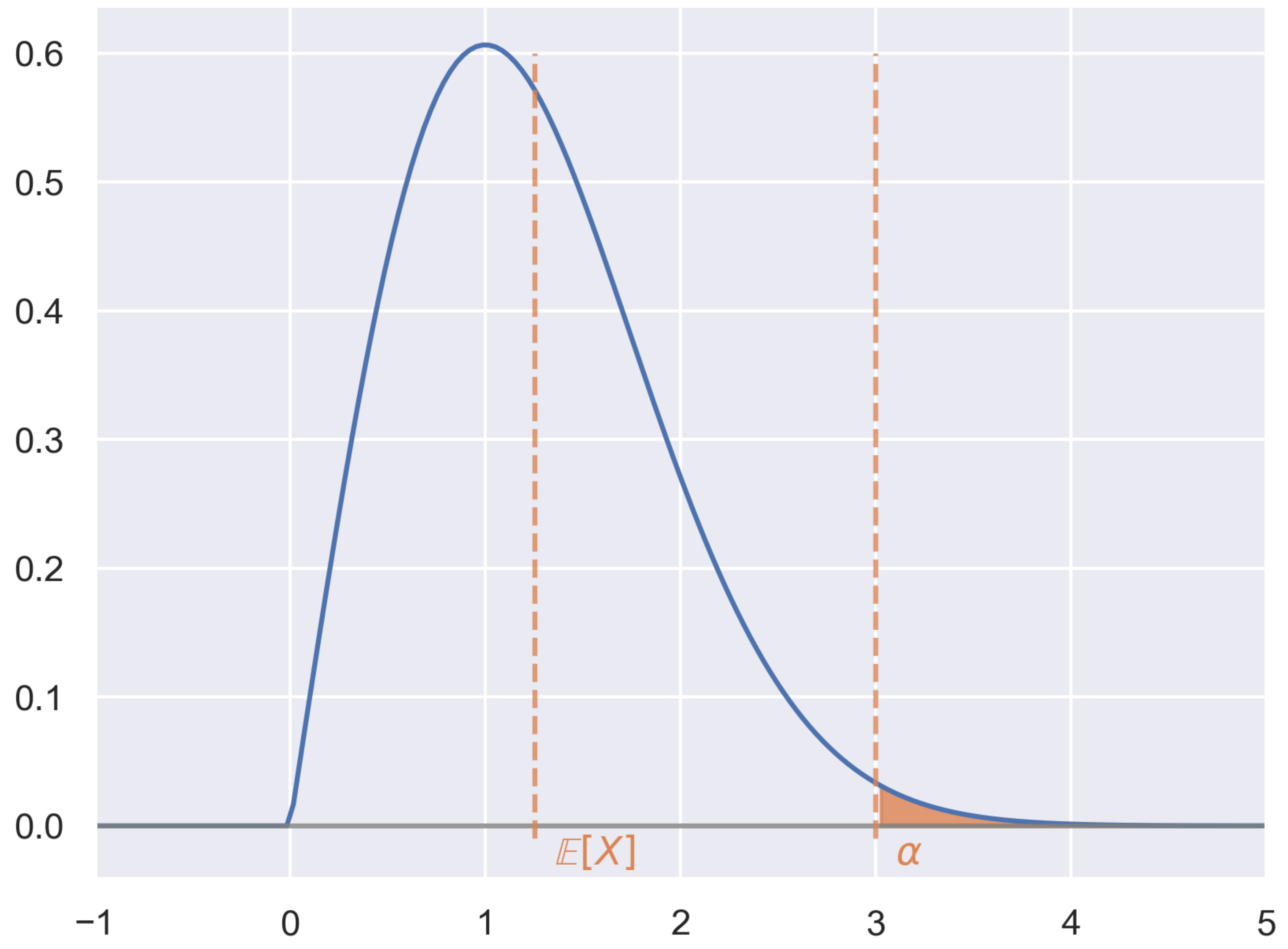


Markov's Inequality

Statement

Theorem (Markov's Inequality). If X is any nonnegative RV with expectation $\mathbb{E}[X]$, then for any $\alpha > 0$,

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



Markov's Inequality

Proof

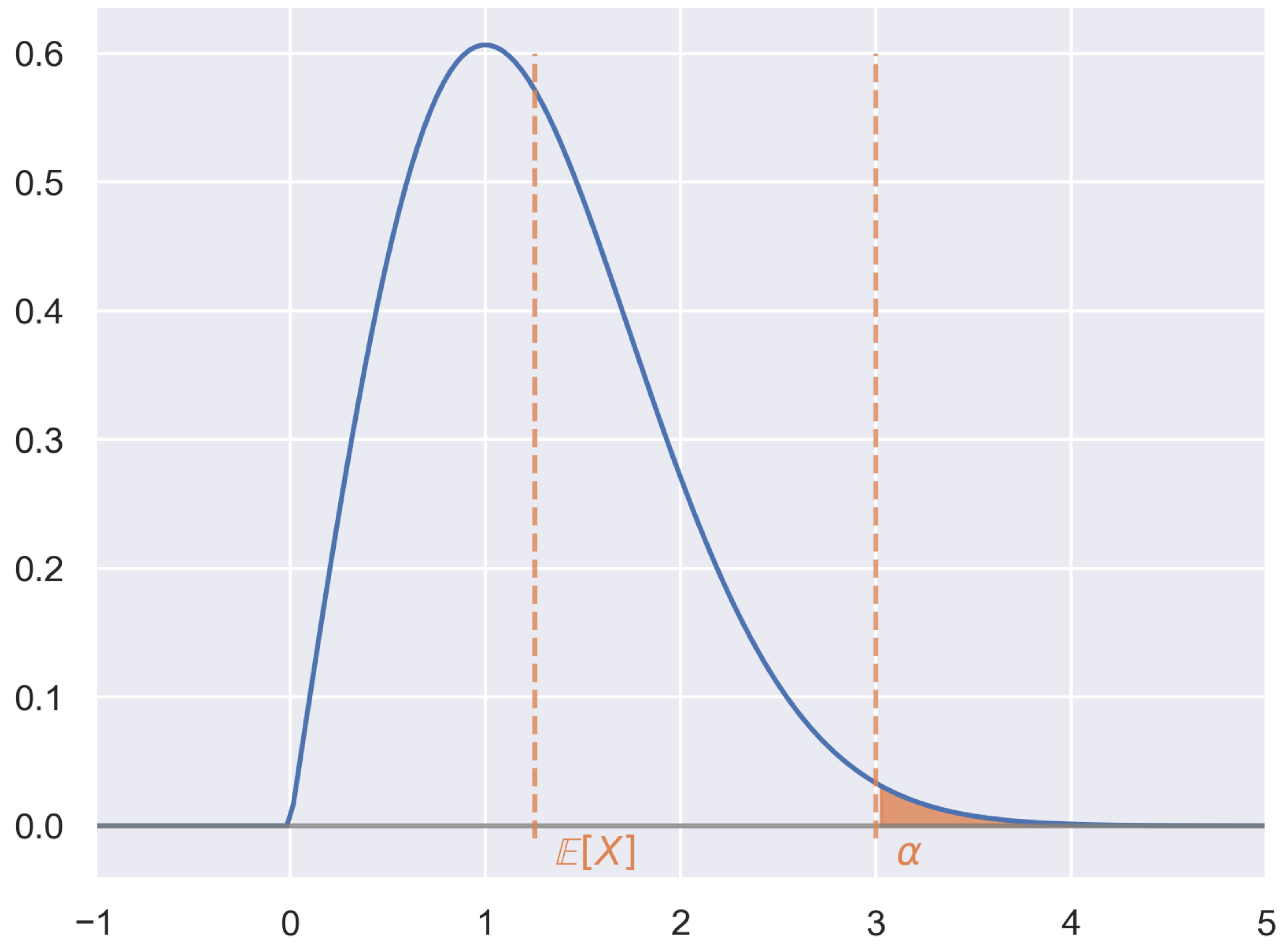
Theorem (Markov's Inequality). If X is any nonnegative RV with expectation $\mathbb{E}[X]$, then for any $\alpha > 0$,

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

Proof. Let $\mathbf{1}\{X \geq \alpha\}$ be the *indicator RV* of the event "$X \geq \alpha$." Then:

$$X \geq \alpha \mathbf{1}\{X \geq \alpha\} \text{ is always true.}$$

Take expectation of both sides, divide by α.

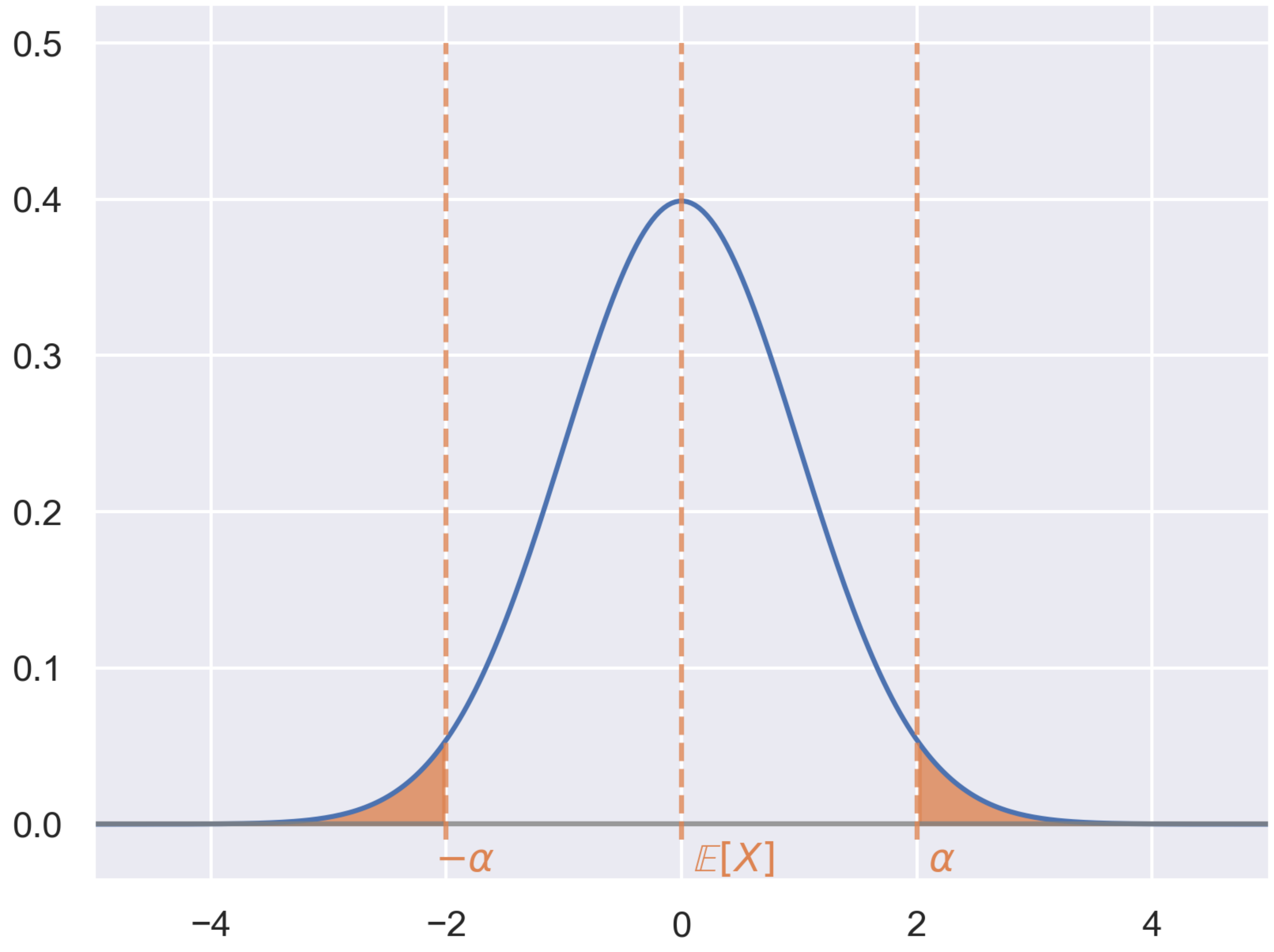


Chebyshev's Inequality

Statement

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

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



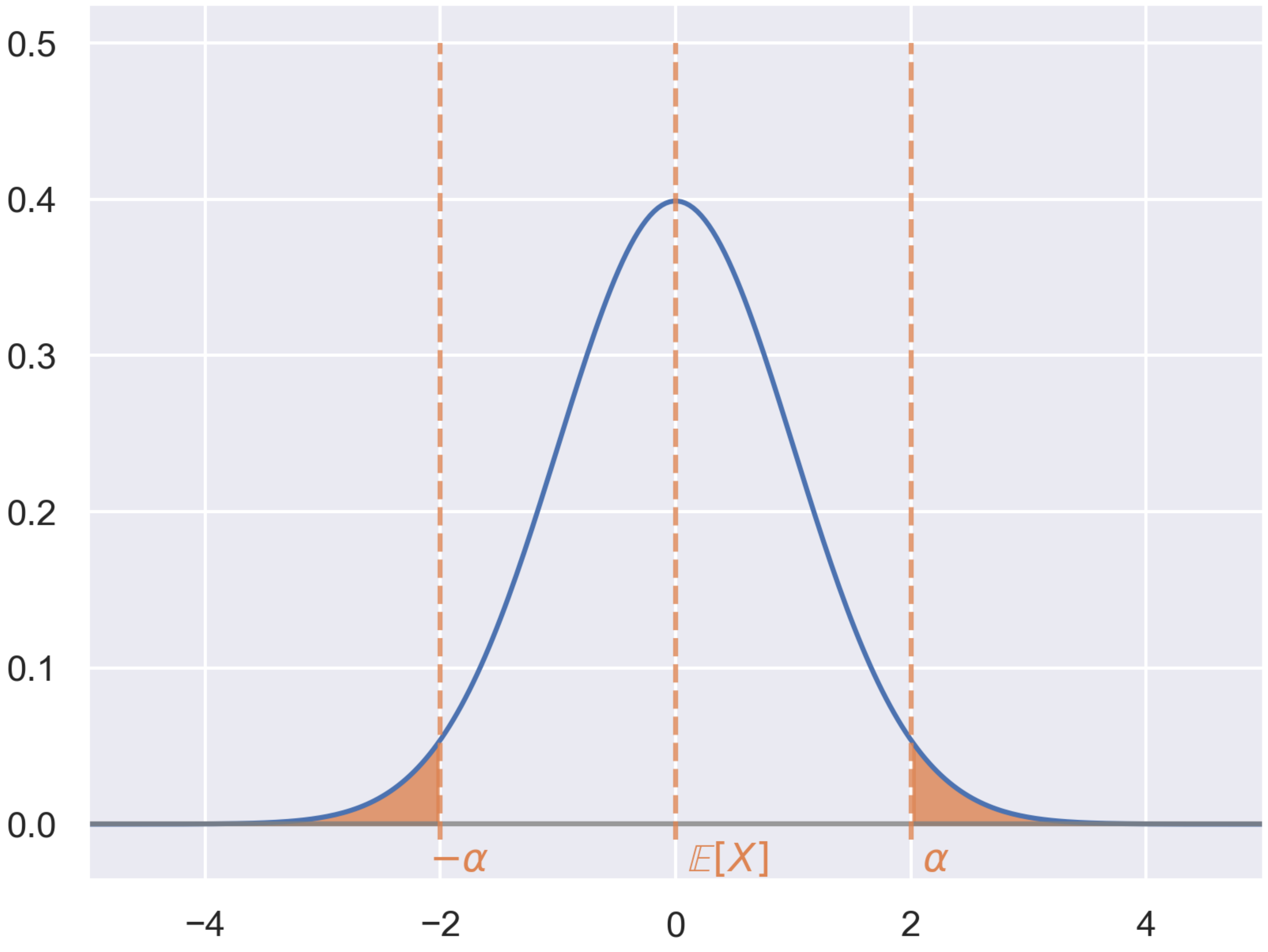
Chebyshev's Inequality

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

LLN: Then, for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - \mu \ < \epsilon\right) = 1$.

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

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

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

Sample Average

Definition

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

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

LLN justifies our "frequentist" view of probability!

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*:

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

Law of Large Numbers

Example: Mean Estimator for Coins

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:

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

Law of Large Numbers

Example: Mean Estimator for Coins

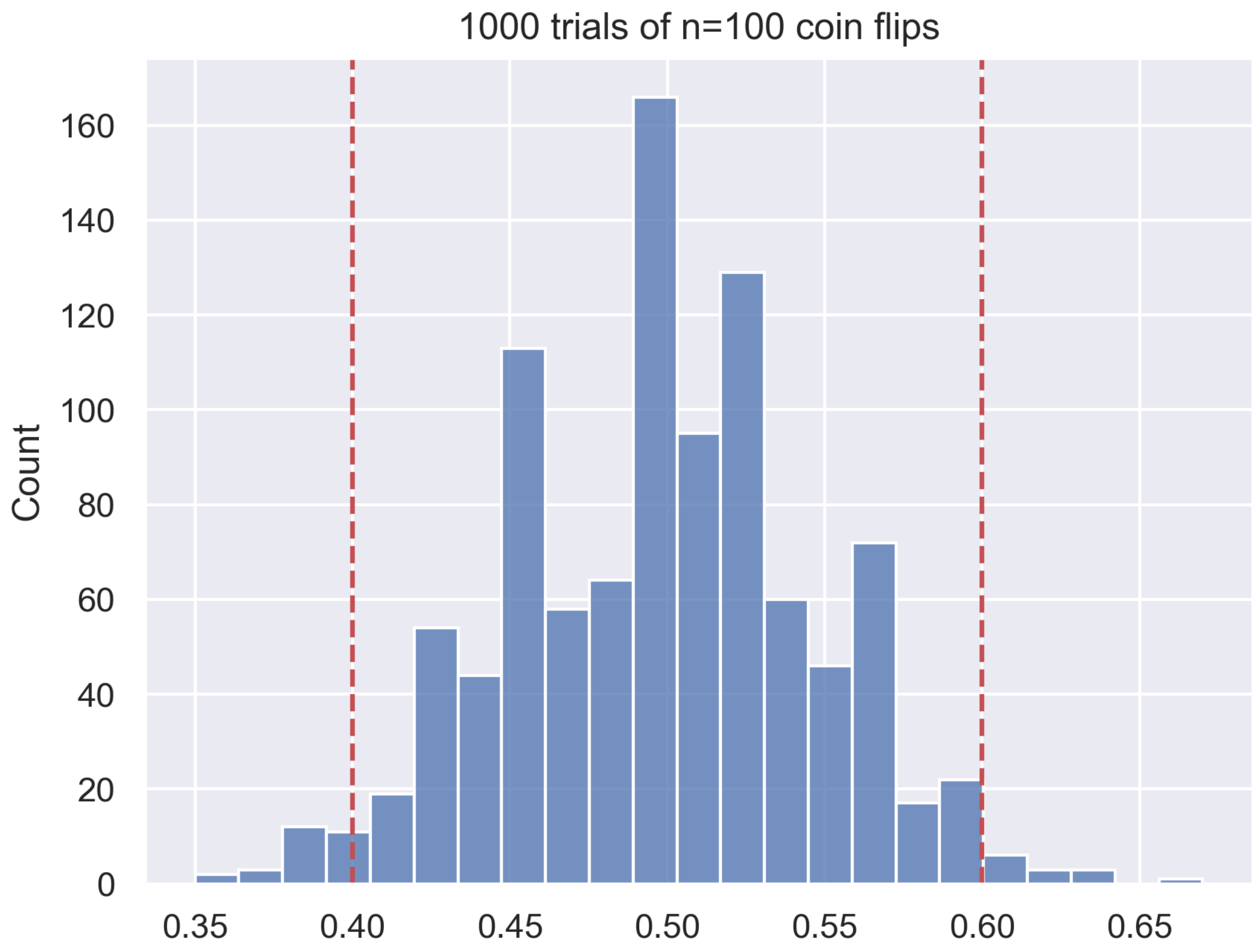
Law of large numbers states that for *any* $\epsilon > 0$, *no matter how small*:

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

Chebyshev's Inequality says:

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

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



Empirical Covariance Matrix

In machine learning

Suppose we draw n examples $\mathbf{x}_1, \ldots \mathbf{x}_n \sim \mathbb{P}_{\mathbf{x}}$ a distribution over $\mathbb{R}^d \ldots$

$\mathbf{x}_i = (x_1, x_2, \ldots, x_d)$ a random vector of d random variables.

Arrange them into a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $\mathbf{x}_i^\top$ are the rows.

Then, if each $\mathbf{x}_i$ is centered (i.e. $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$), the empirical covariance matrix is:

$$\hat{\boldsymbol{\Sigma}}_n := \frac{1}{n}\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}.$$

A property of the a specific observed dataset, $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Empirical Covariance Matrix

Law of Large Numbers

Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ is an observed data matrix where $\mathbf{x}_i \in \mathbb{R}^d$ are the rows, drawn i.i.d. from $\mathbb{P}_{\mathbf{x}}$.

By the law of large numbers,

$$\hat{\mathbf{\Sigma}}_n := \frac{1}{n}\mathbf{X}^\top \mathbf{X} \to \mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Var}(\mathbf{x}), \text{ as } n \to \infty.$$

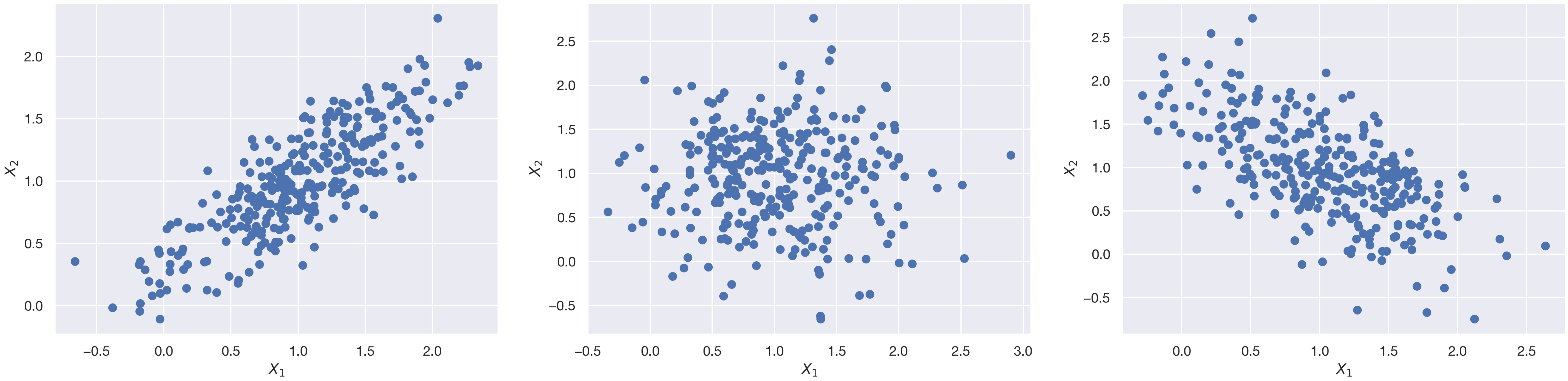
$$\text{Useful fact: } \hat{\mathbf{\Sigma}}_n^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} \sim \frac{1}{n}\mathbf{\Sigma}^{-1}.$$

The empirical covariance matrix is a approaches the true covariance matrix with more data!

Empirical Covariance Matrix

Law of Large Numbers

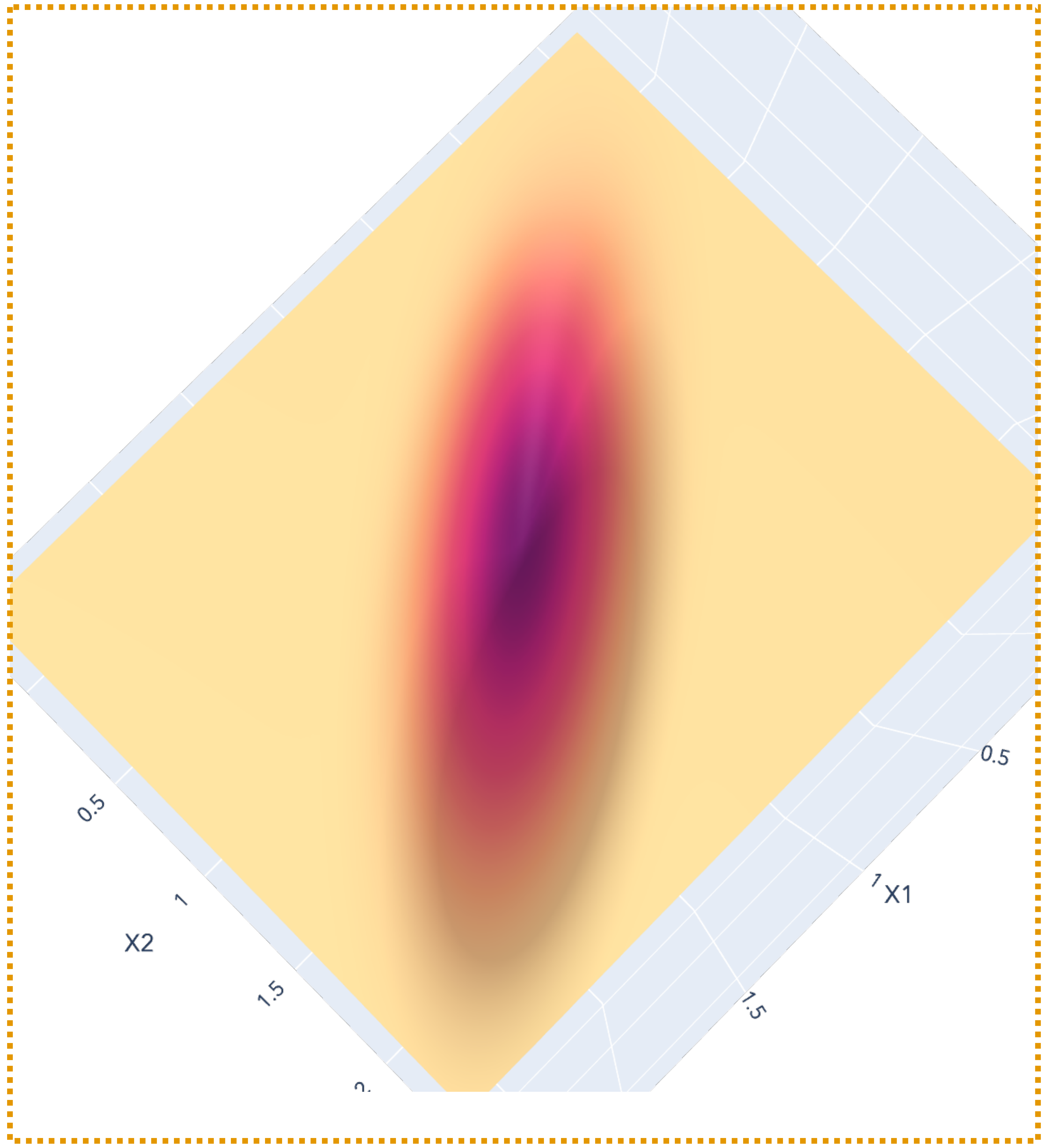
$$\hat{\boldsymbol{\Sigma}}_n := \frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Var}(\mathbf{x}), \text{ as } n \to \infty.$$

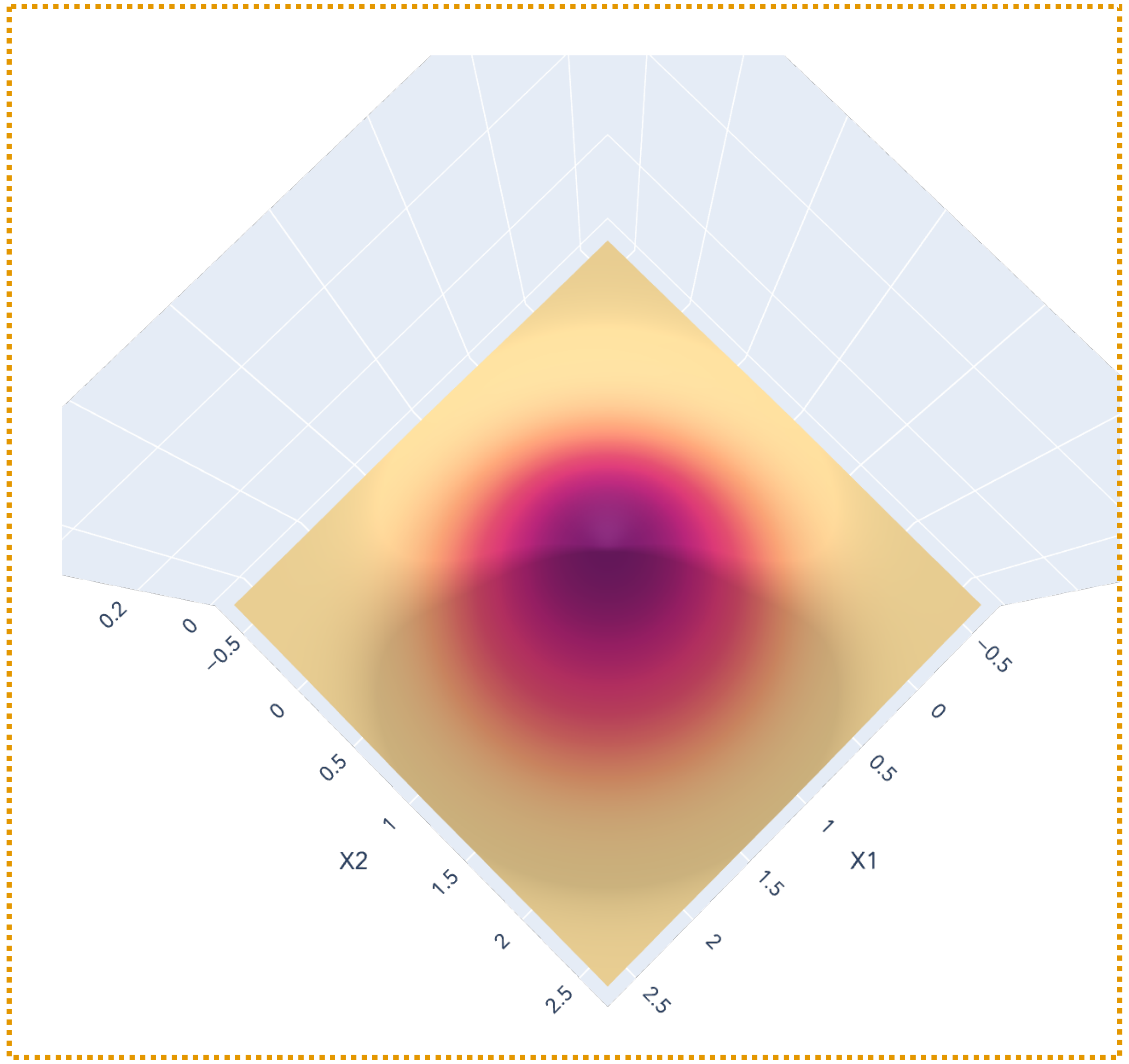


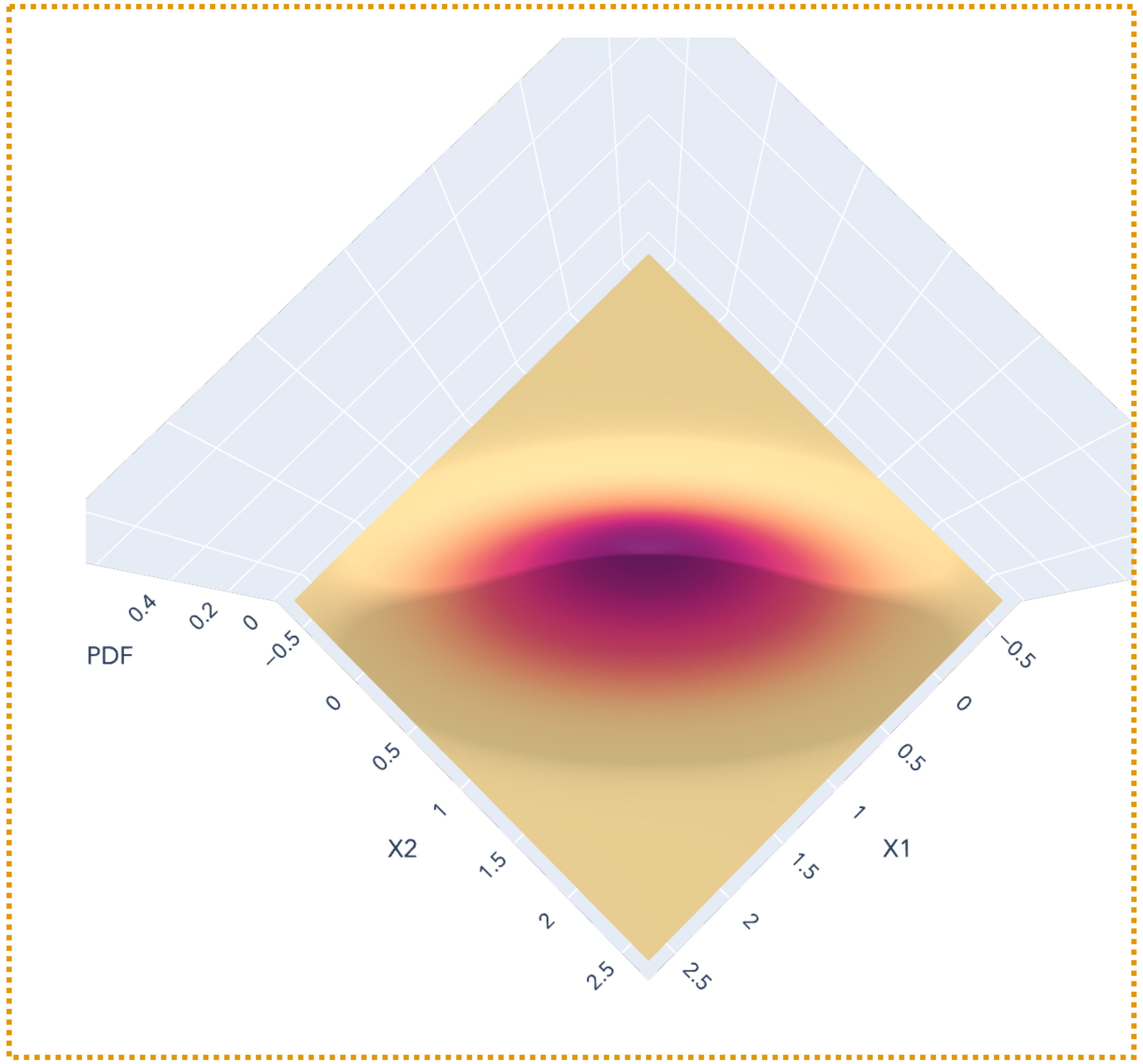
Empirical Covariance Matrix

Law of Large Numbers

$$\hat{\mathbf{\Sigma}}_n := \frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathrm{Var}(\mathbf{x}), \text{ as } n \to \infty.$$

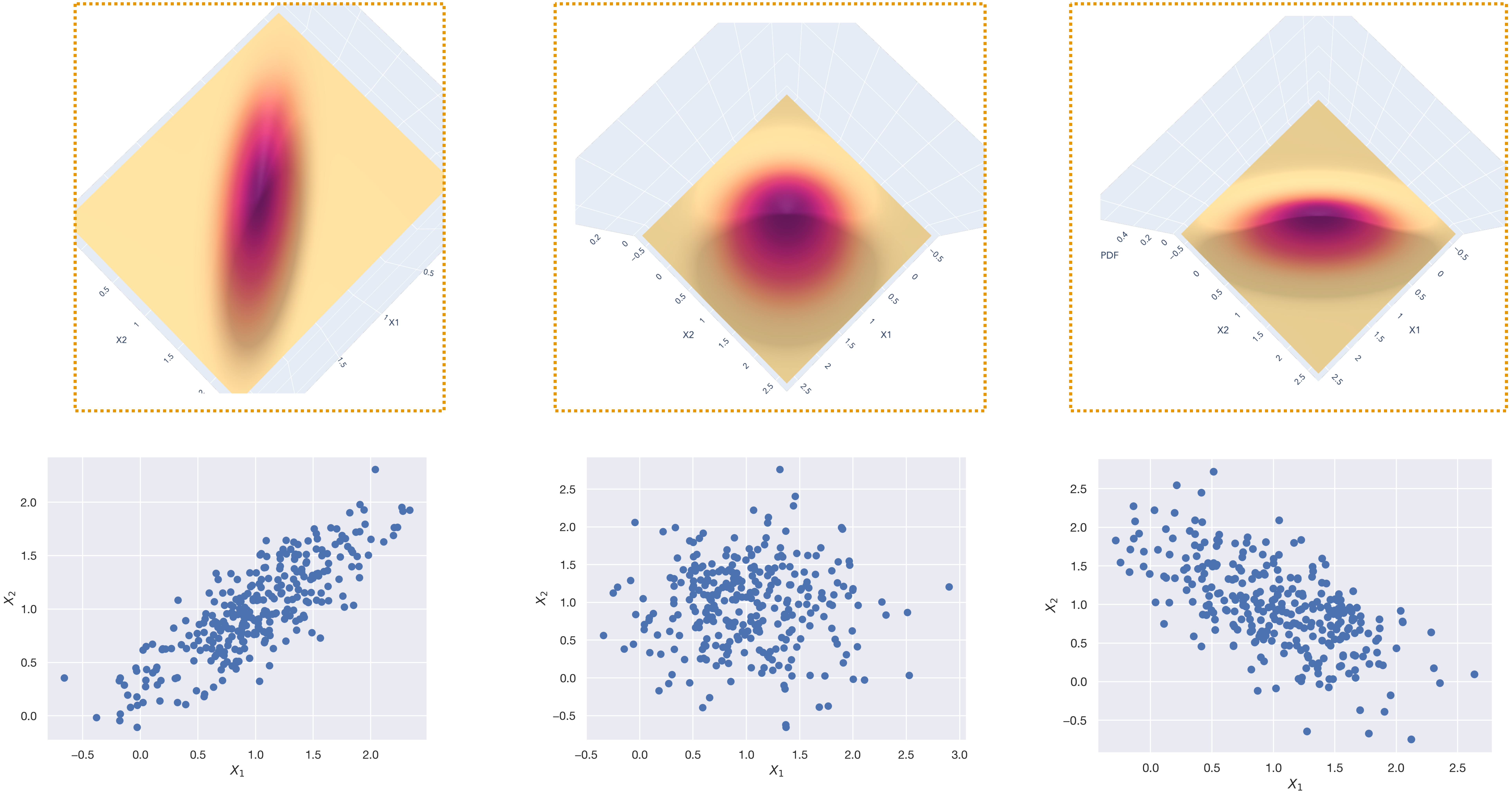






Empirical Covariance Matrix

Law of Large Numbers



Statistical Estimation

Intuition

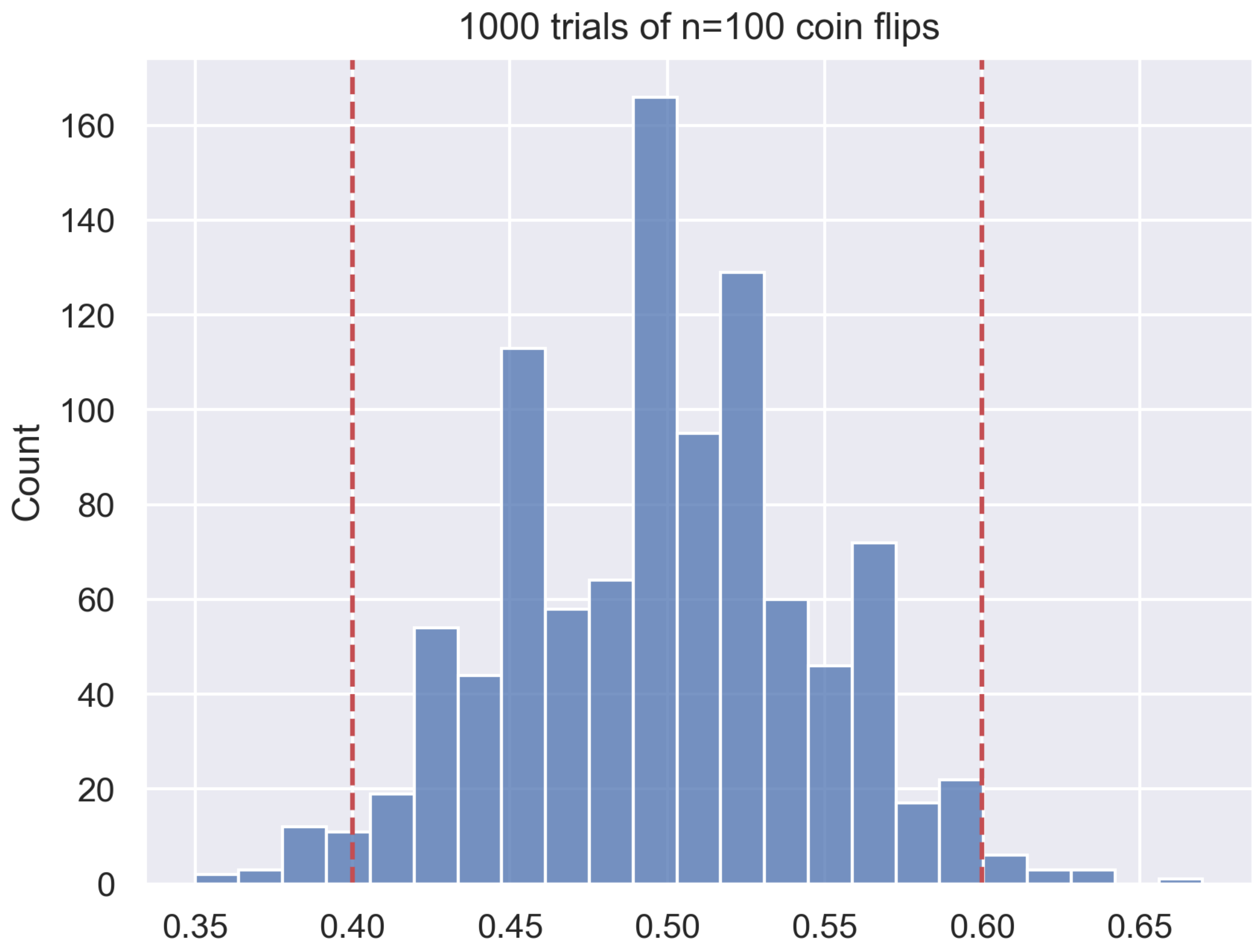
Make some assumptions about data that we're to collect. (i.i.d. assumption).

Collect as much data as we can about the phenomenon. ($n = 100$ *coin flips*).

Use the data to derive characteristics (statistics) about how data were generated (the *true* mean $\mathbb{E}[X_i] = 0.5$)

via some estimator.

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



Generalization

Intuition

Statistics/statistical inference involves drawing conclusions about data we've already seen.

Generalization is a big concern in ML – we want to describe *unseen* data well.

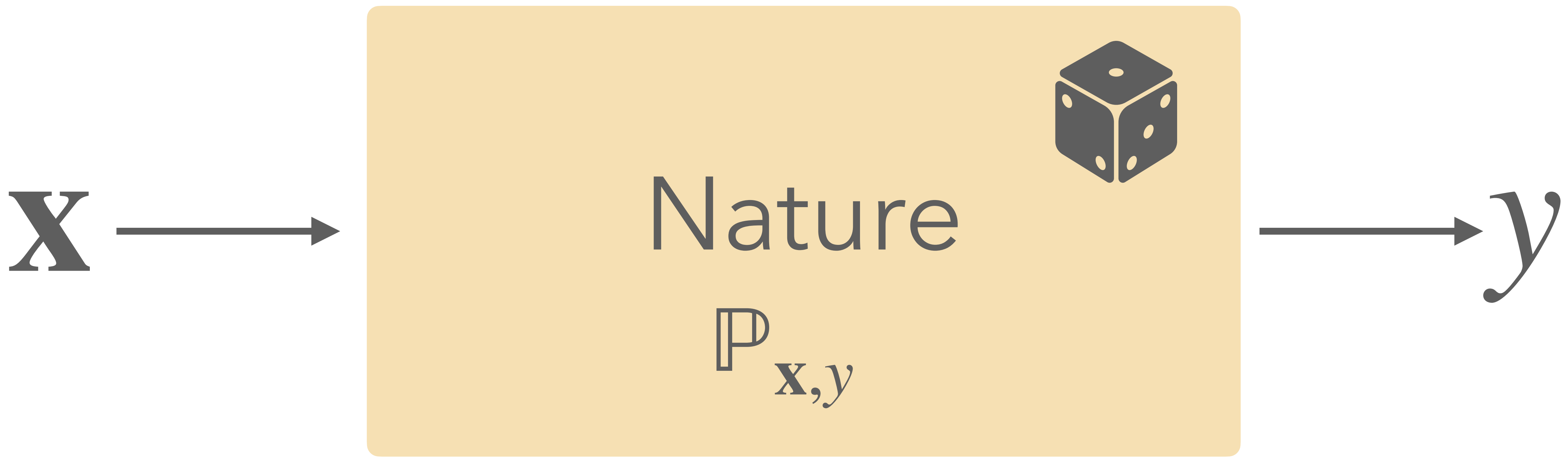
If the future data comes from the same distribution as our past data, then we can hope to generalize by describing our past data well!

Random error model

Our main assumption on $\mathbb{P}_{\mathbf{x},y}$

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

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



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.

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 $\mathbb{P}_X$ with parameter θ.

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 RVs, so they are *themselves* RVs!

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 i.i.d. 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

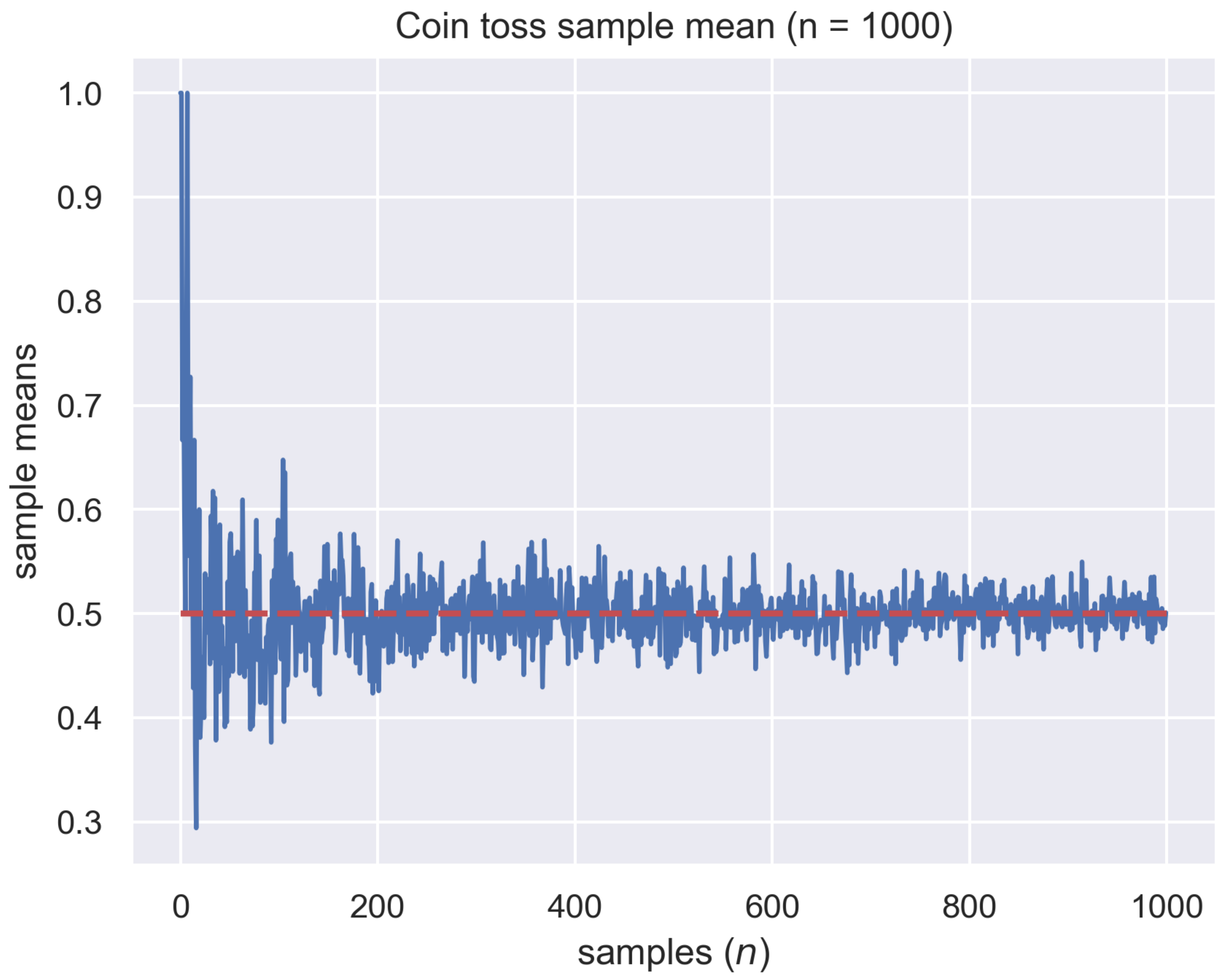
Example: Estimating coin flip

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 i.i.d. 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

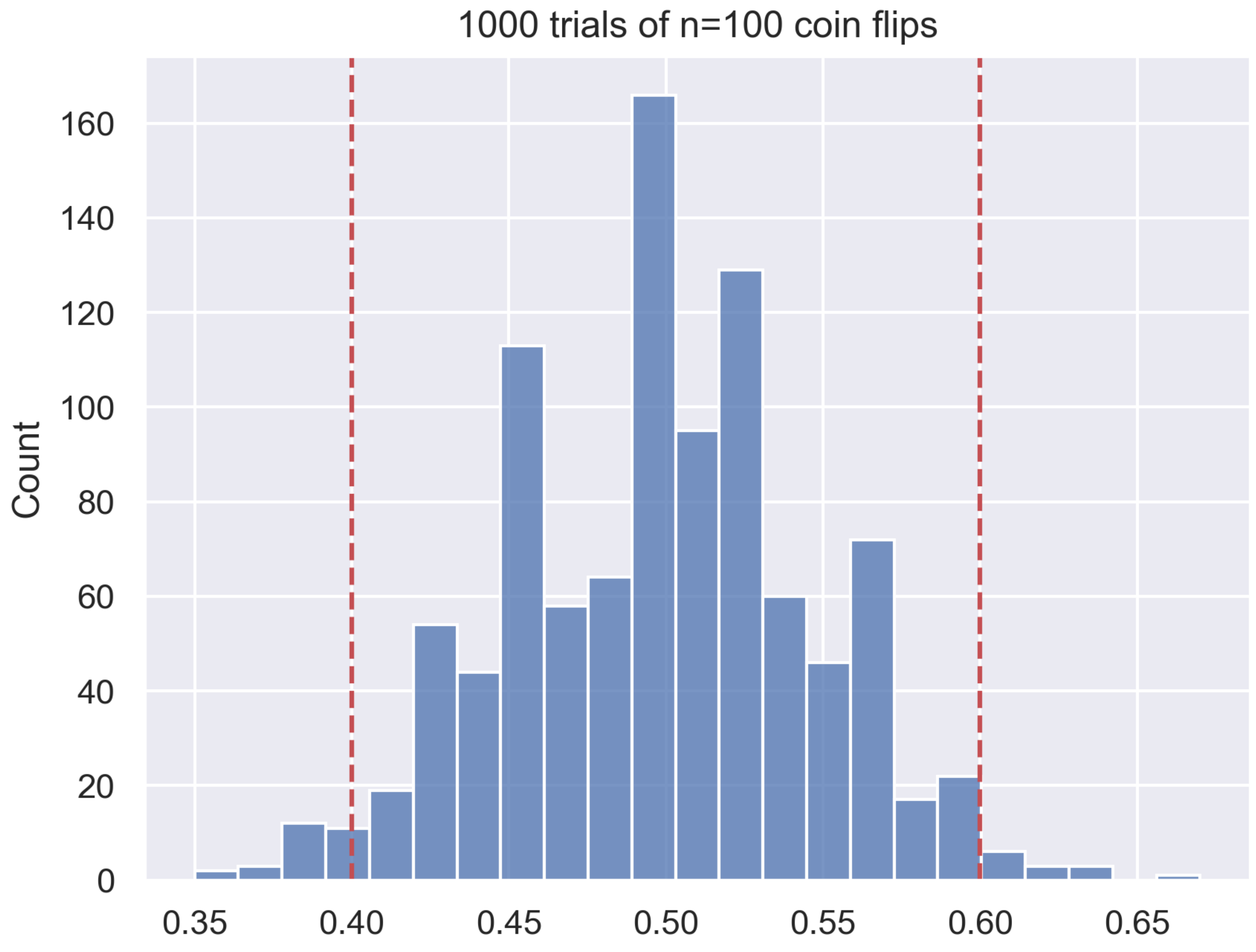
Example: Estimating coin flip

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 i.i.d. 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

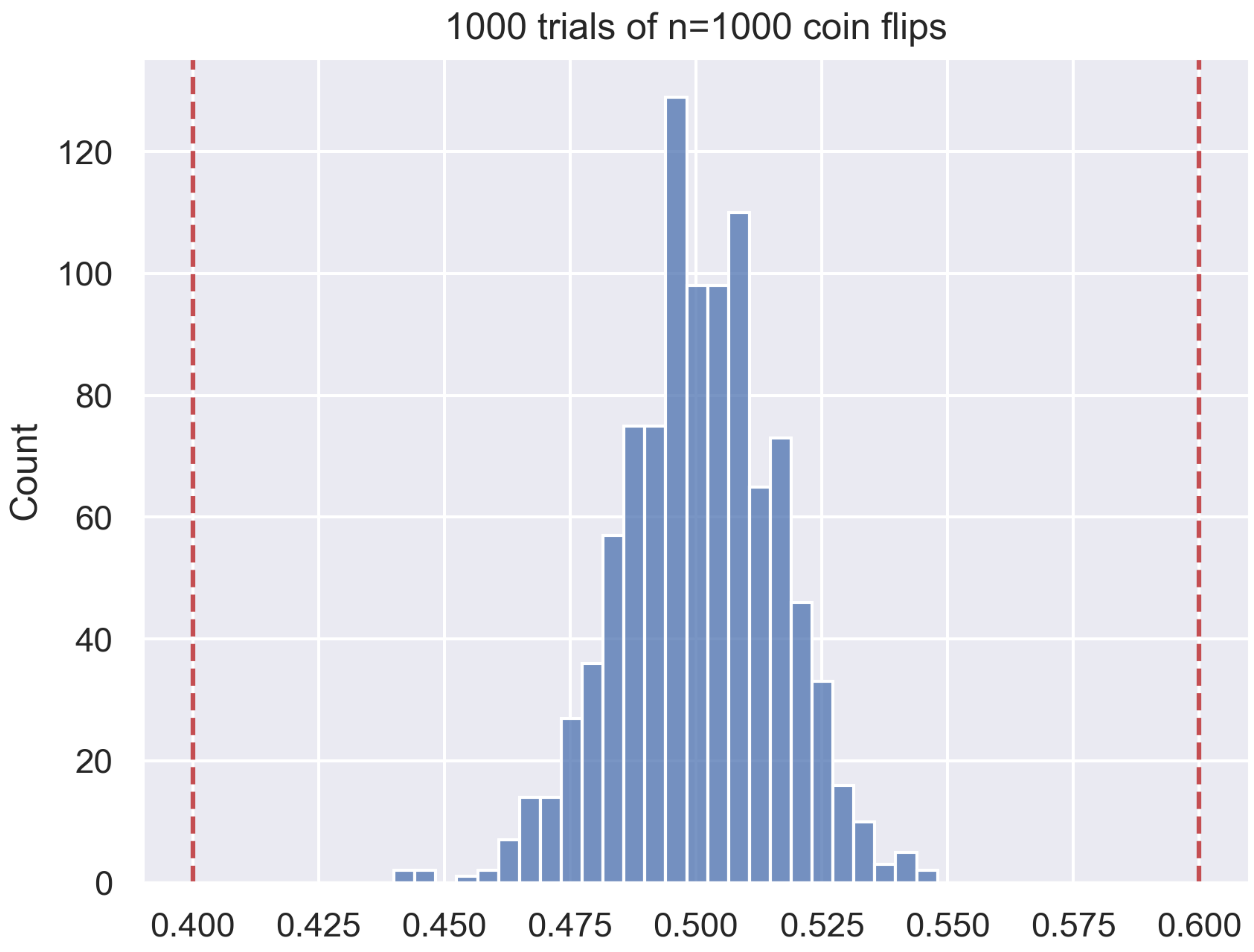
Example: Estimating coin flip

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 i.i.d. 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: 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 i.i.d. RVs $X_1, \ldots, X_n$.

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

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

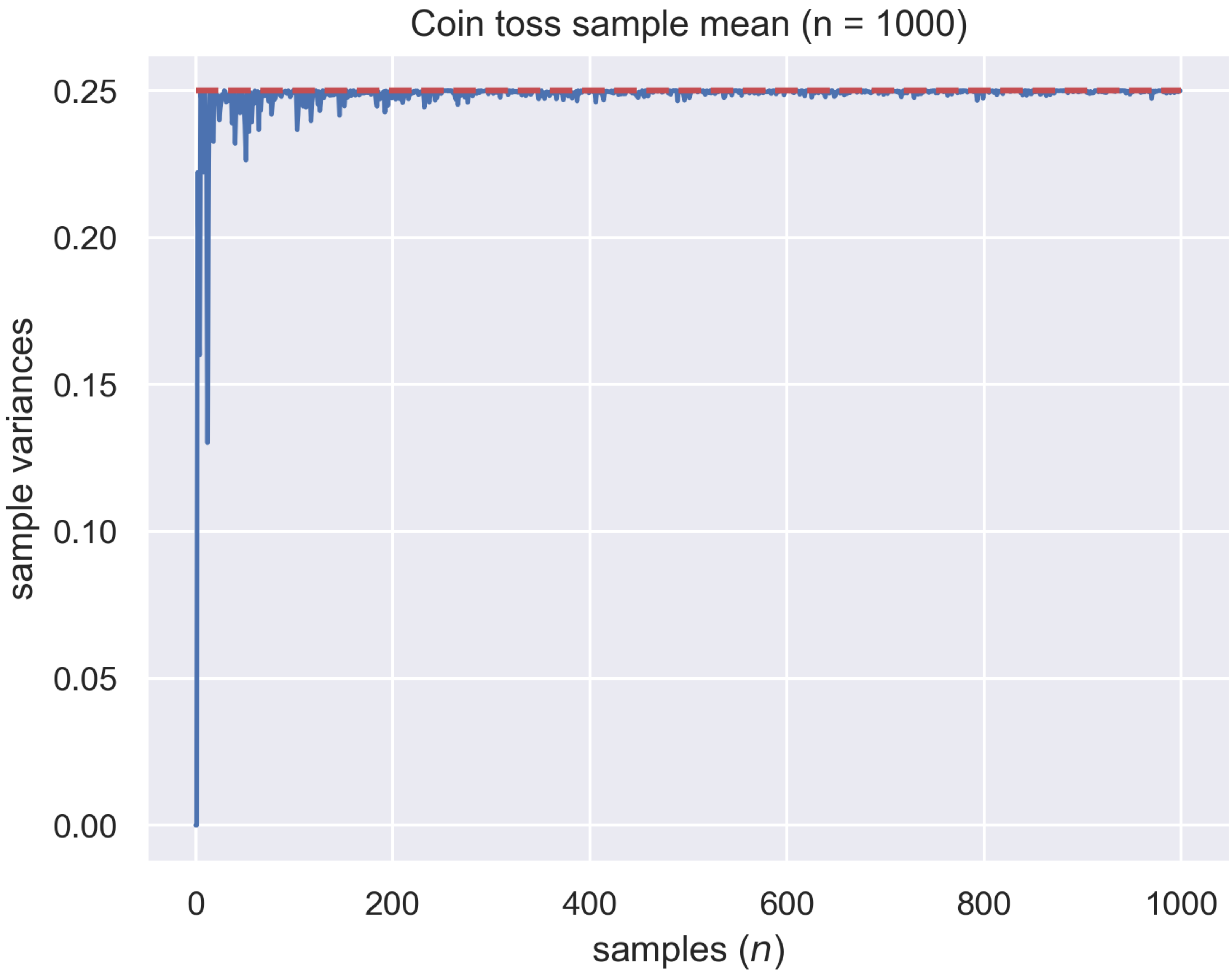
Example: Variance Estimation

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 i.i.d. RVs $X_1, \ldots, X_n$.

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

Estimator: $\hat{\theta}_n = s_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ (*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 $X_1, \ldots, X_n$.

$$\text{Estimand: } \theta = \mu.$$

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

Statistical Estimator

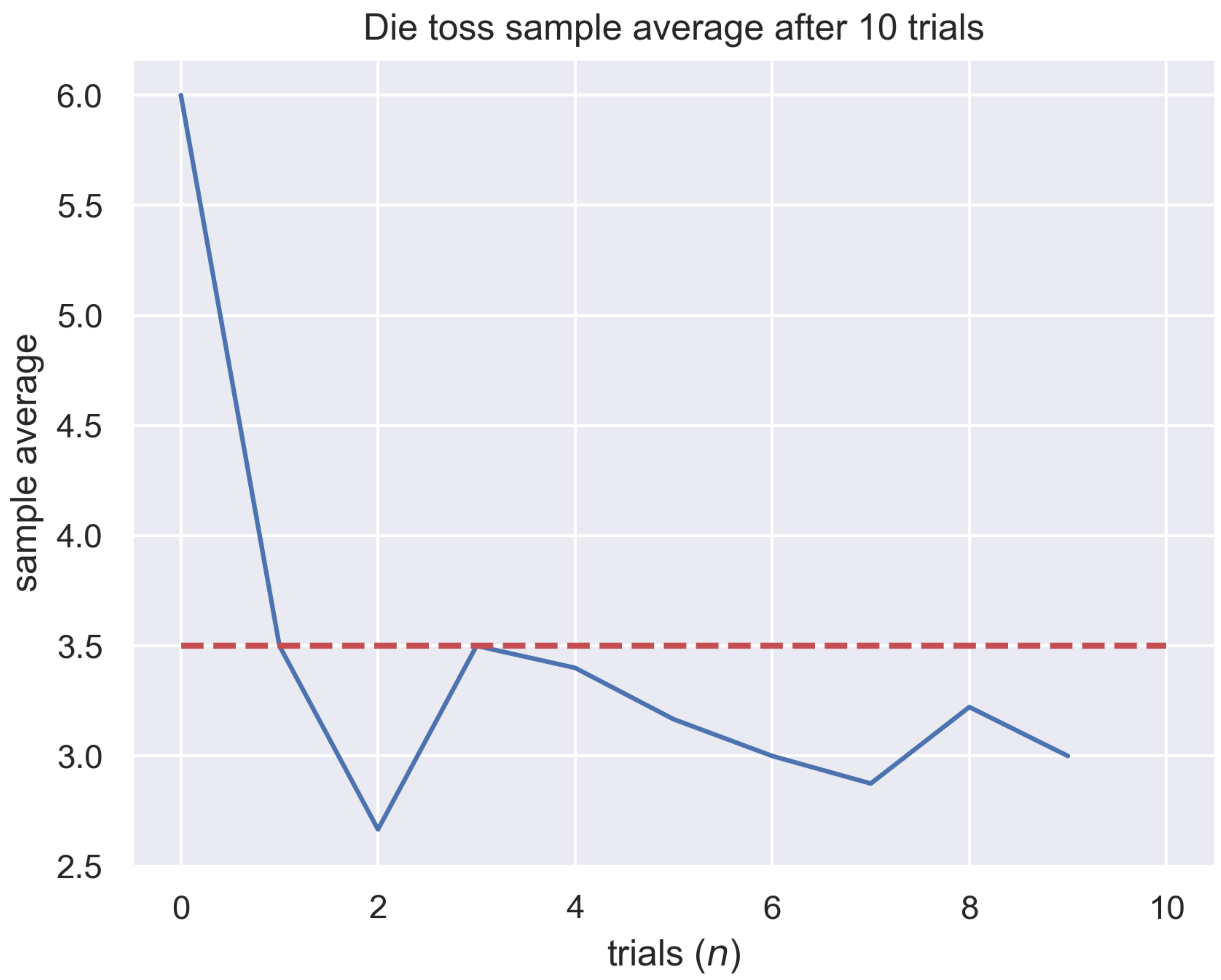
Example: Mean Estimator for Dice

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

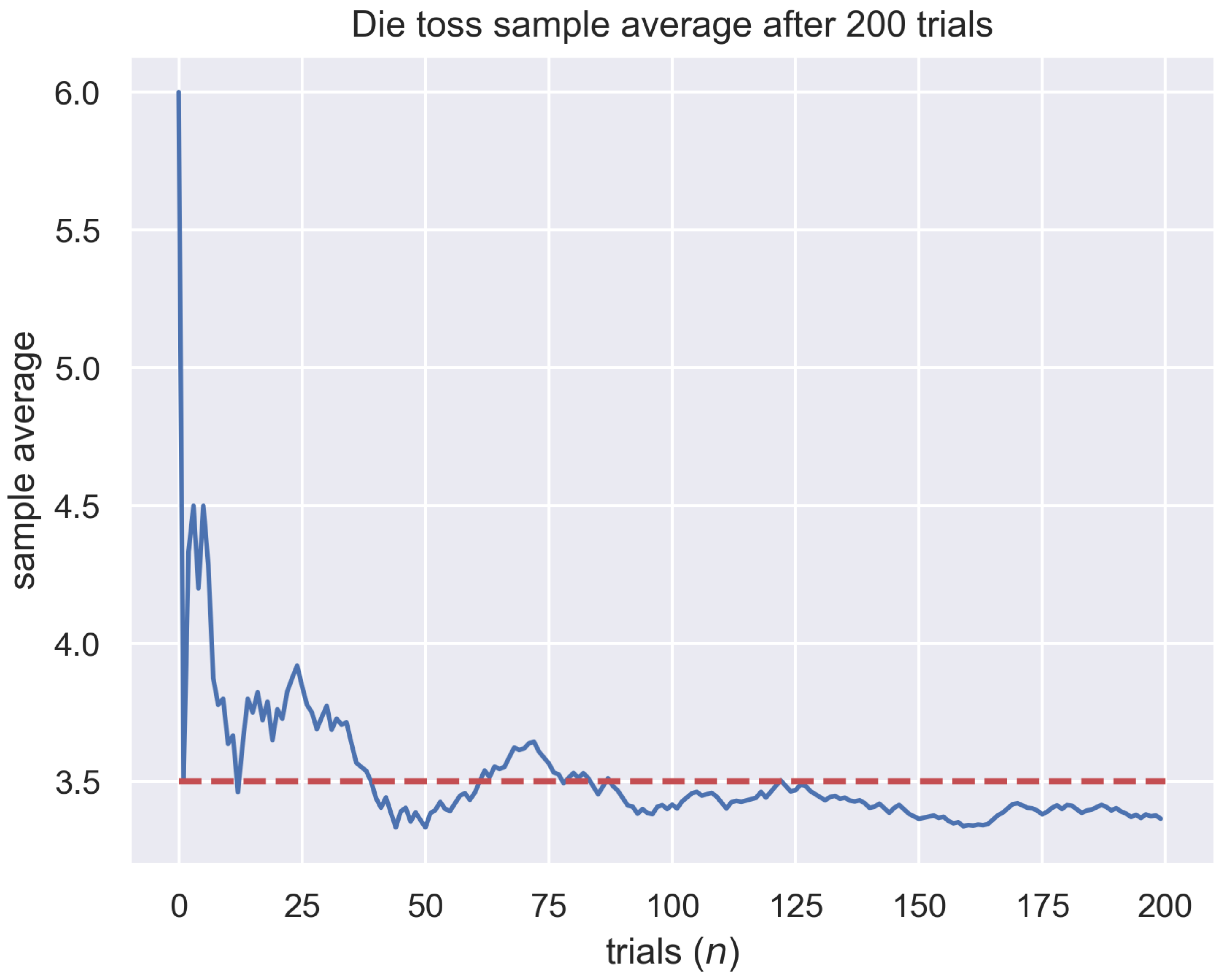
Example: Mean Estimator for Dice

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: $\theta = \mu$.

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



Statistical Estimator

Example: Mean Estimator for Dice

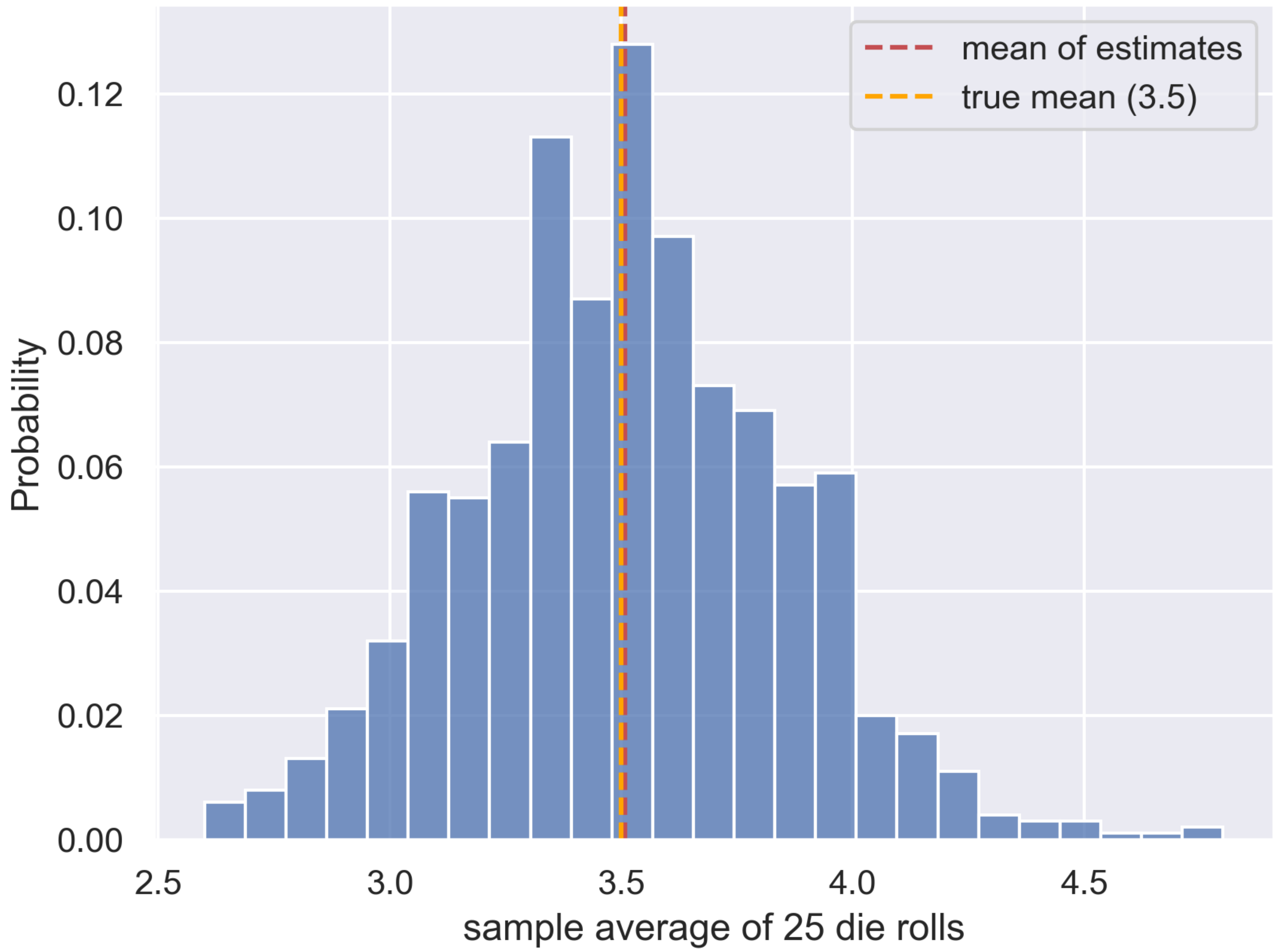
Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

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

Estimand: $\theta = \mu$.

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

Estimator is itself a random variable!



Statistical Estimator

Example: Mean Estimator for Dice

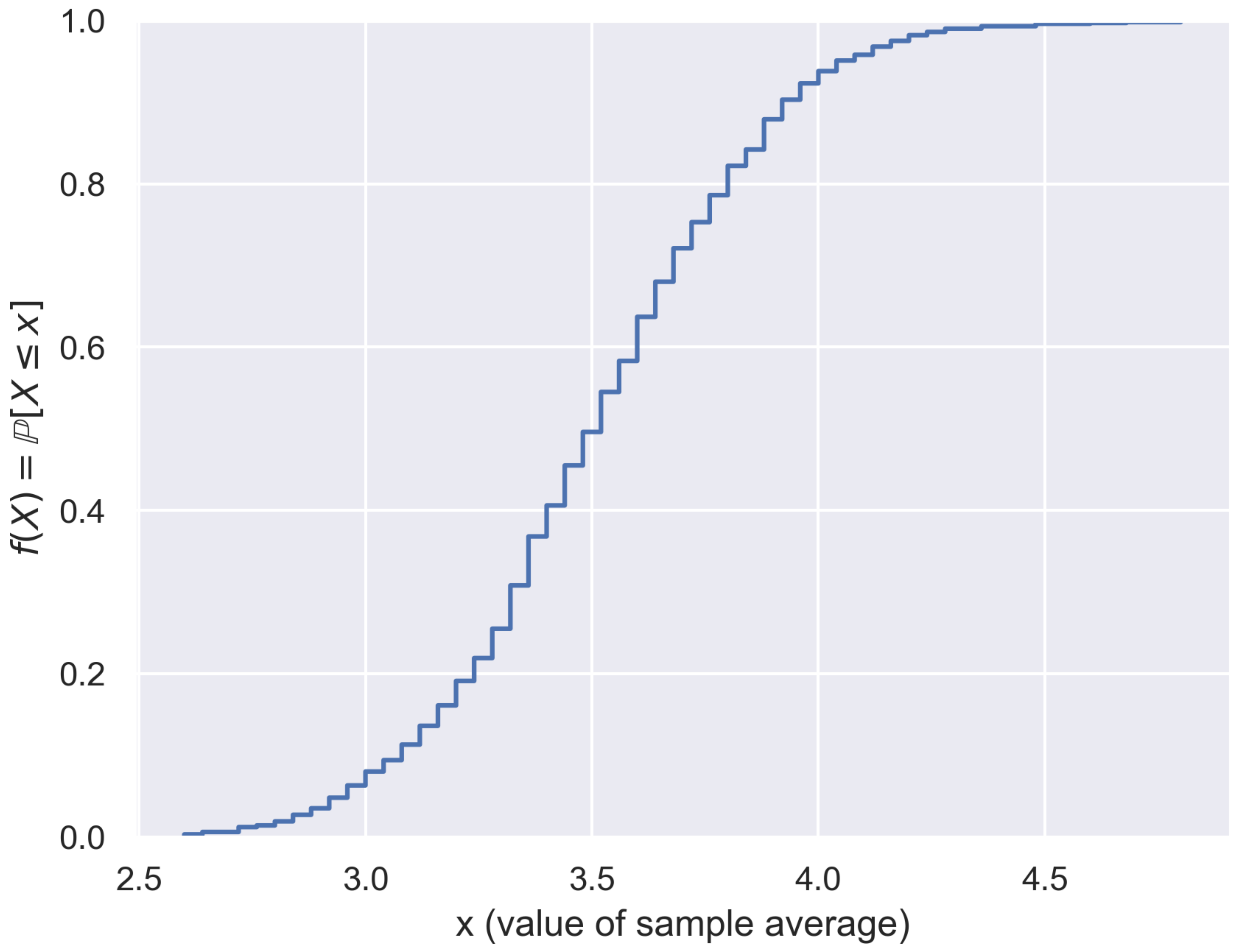
Example. Let X_i be a random variable denoting the face after tossing a six-sided fair die. Clearly, $\mu := \mathbb{E}[X_i] = 3.5$.

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

Estimand: $\theta = \mu$.

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

Estimator is itself a random variable!



Statistical Estimator

Example: OLS 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 $\mathbb{P}_{\mathbf{x},y}$ that follows 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: $\hat{\theta}_n = \hat{\mathbf{w}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

By LLN: $(\mathbf{X}^\top \mathbf{X})^{-1} \sim \frac{1}{n} \boldsymbol{\Sigma}^{-1}$, the true covariance.

Statistical Estimator

Example: OLS Estimator

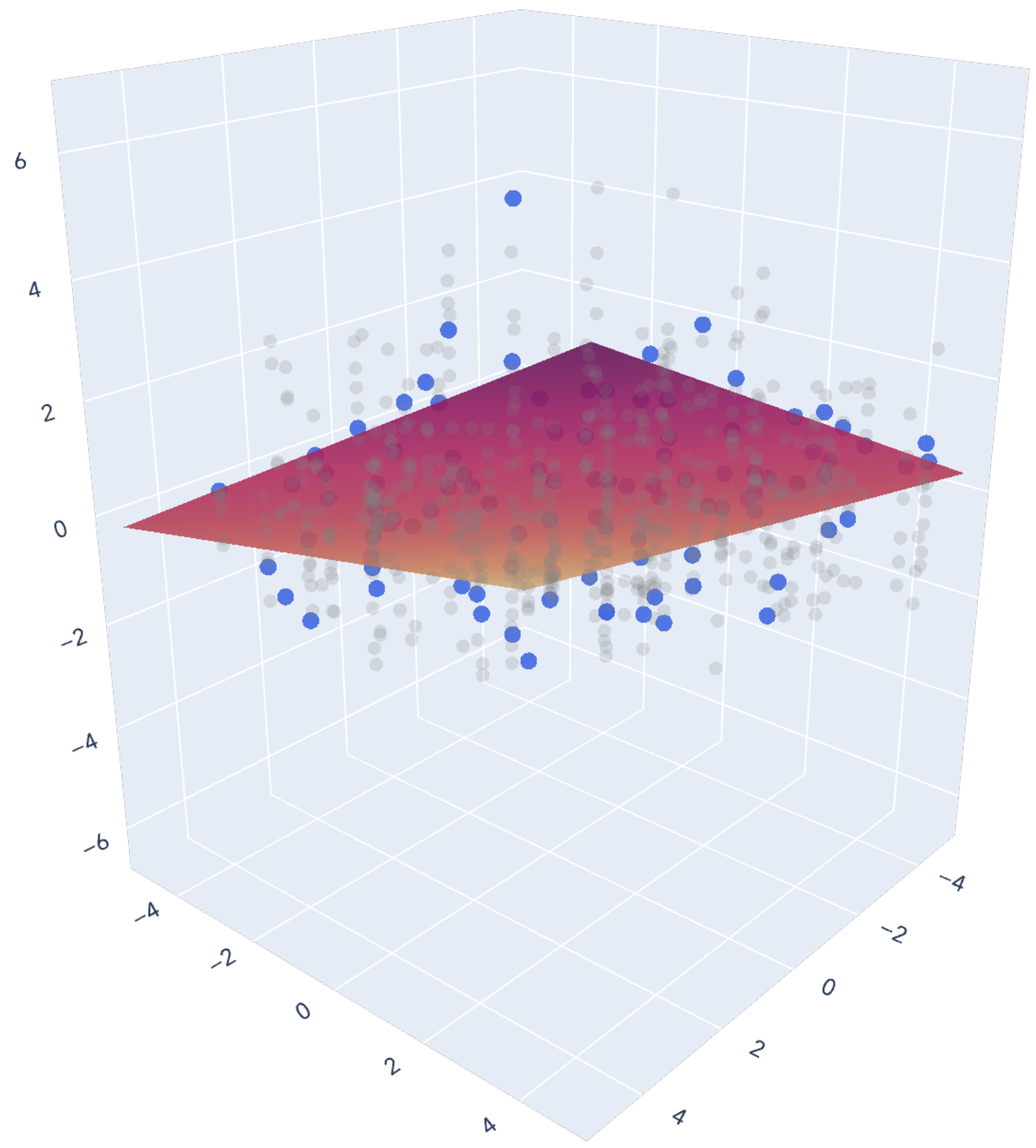
Example. Let $(\mathbf{x}_1, y_1)\ldots,(\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be i.i.d. samples from the joint distribution $\mathbb{P}_{\mathbf{x},y}$ that follows the error model:

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

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: $\hat{\theta}_n = \hat{\mathbf{w}}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$



Statistical Estimator

Example: Ridge Regression 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 $\mathbb{P}_{\mathbf{x},y}$ that follows 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 $\gamma > 0$ is the *regularization parameter*.

Statistical Estimators

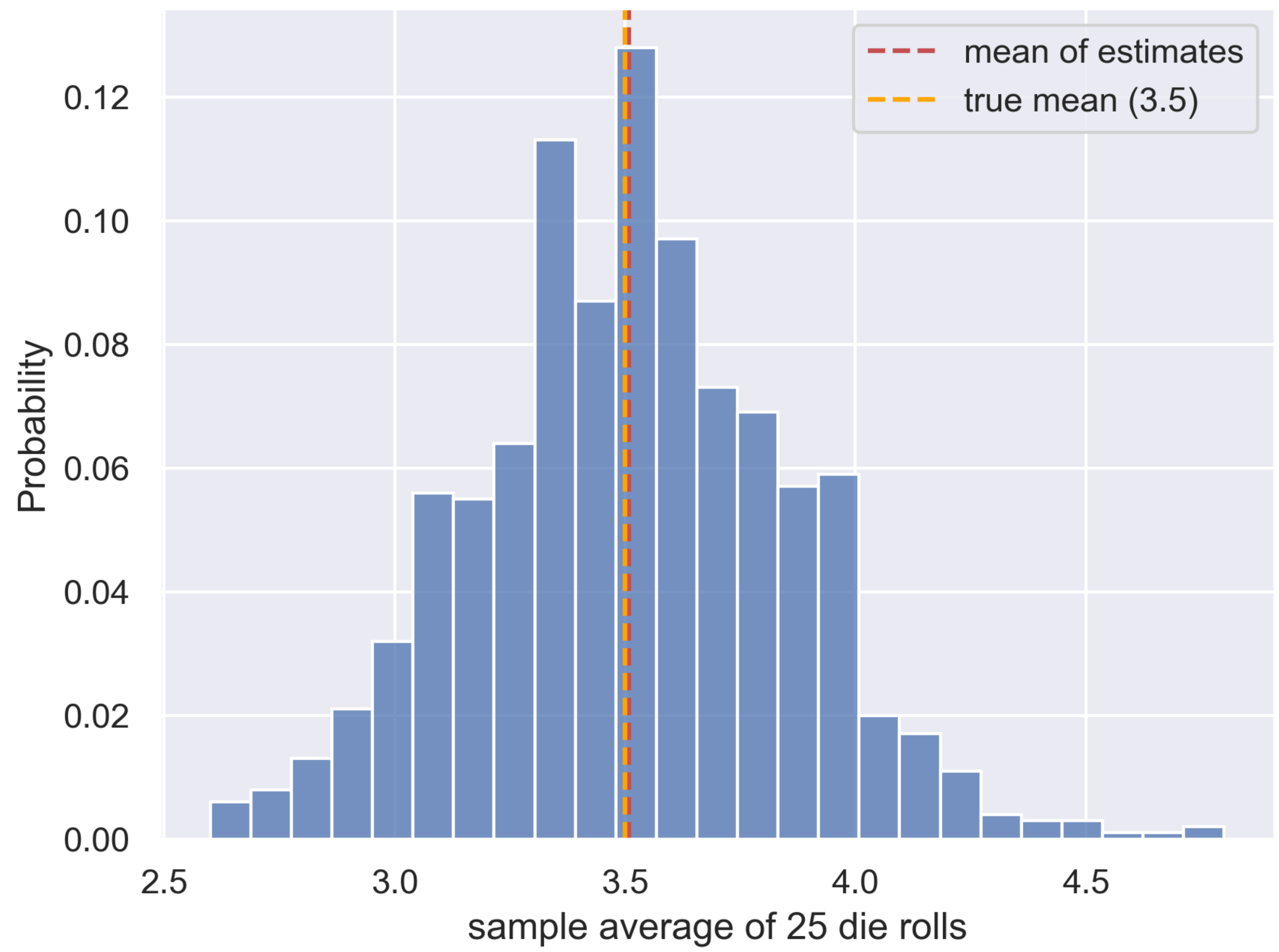
Variance and bias

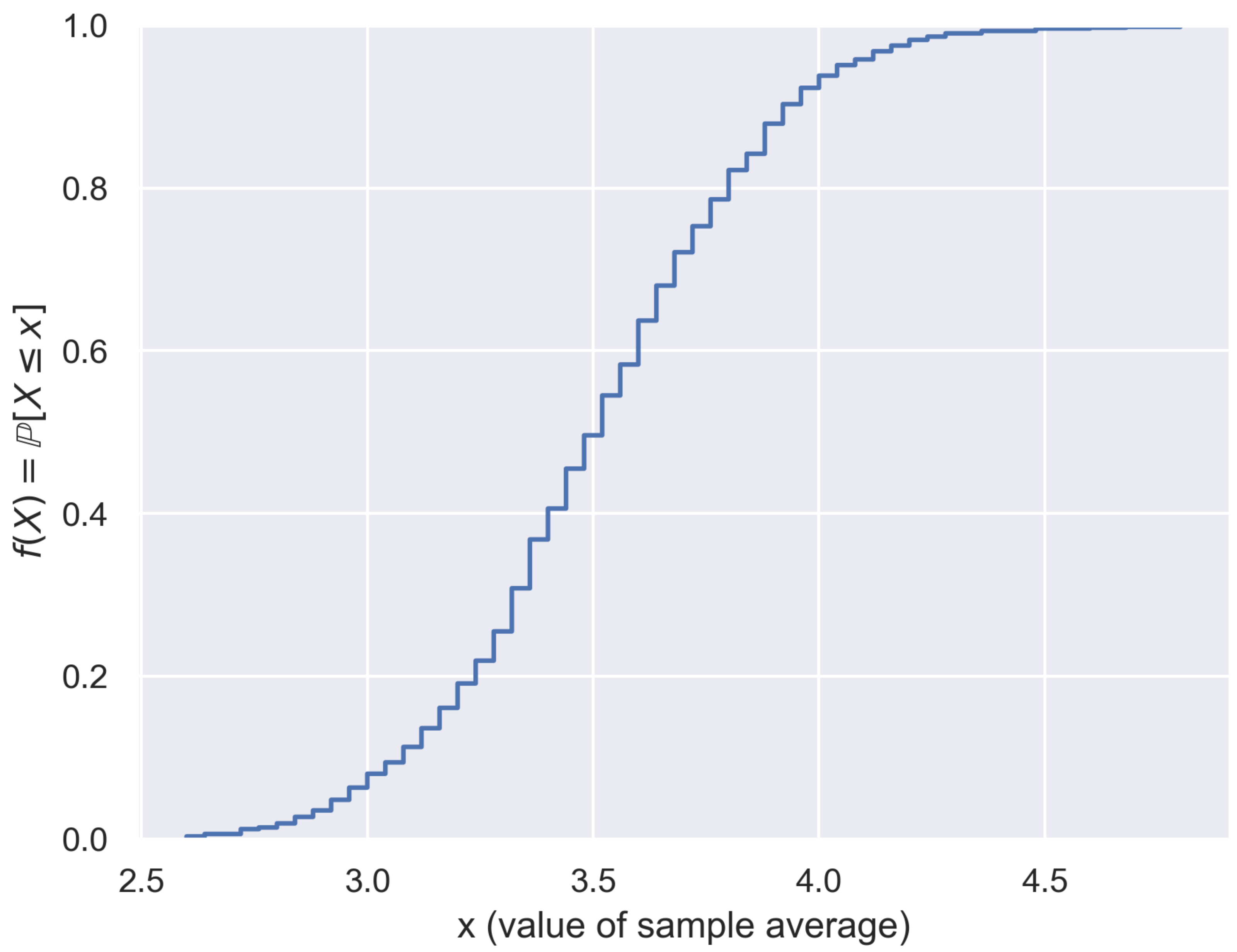
Statistical Estimator

Random Variables

Remember that statistical estimators are random variables!

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

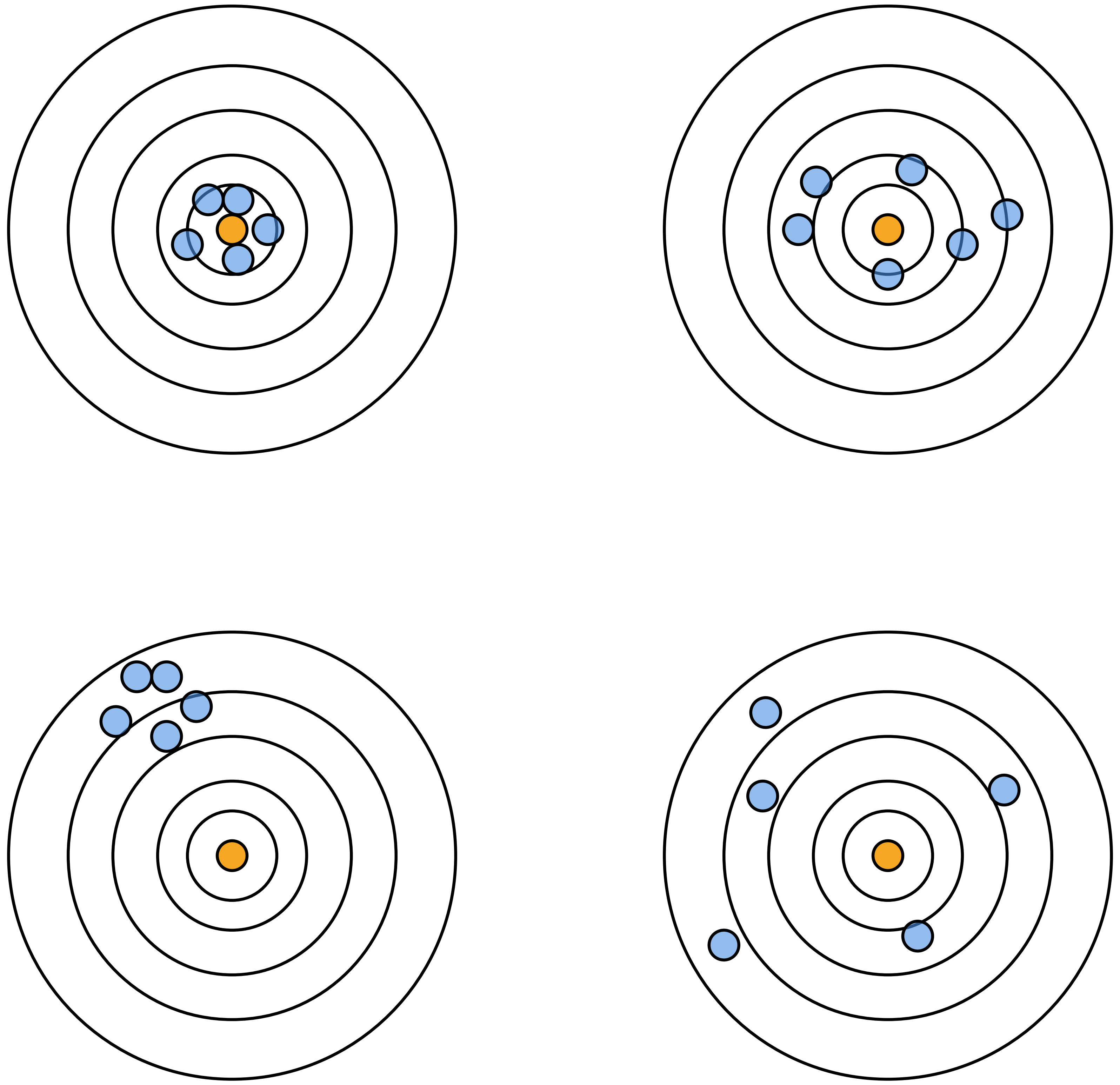




Bias of Estimators

Intuition

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

Bias of Estimators

Examples of Estimators

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, $\theta = \mu$.

What's the bias of the estimator $\hat{\theta}_n = 1$?

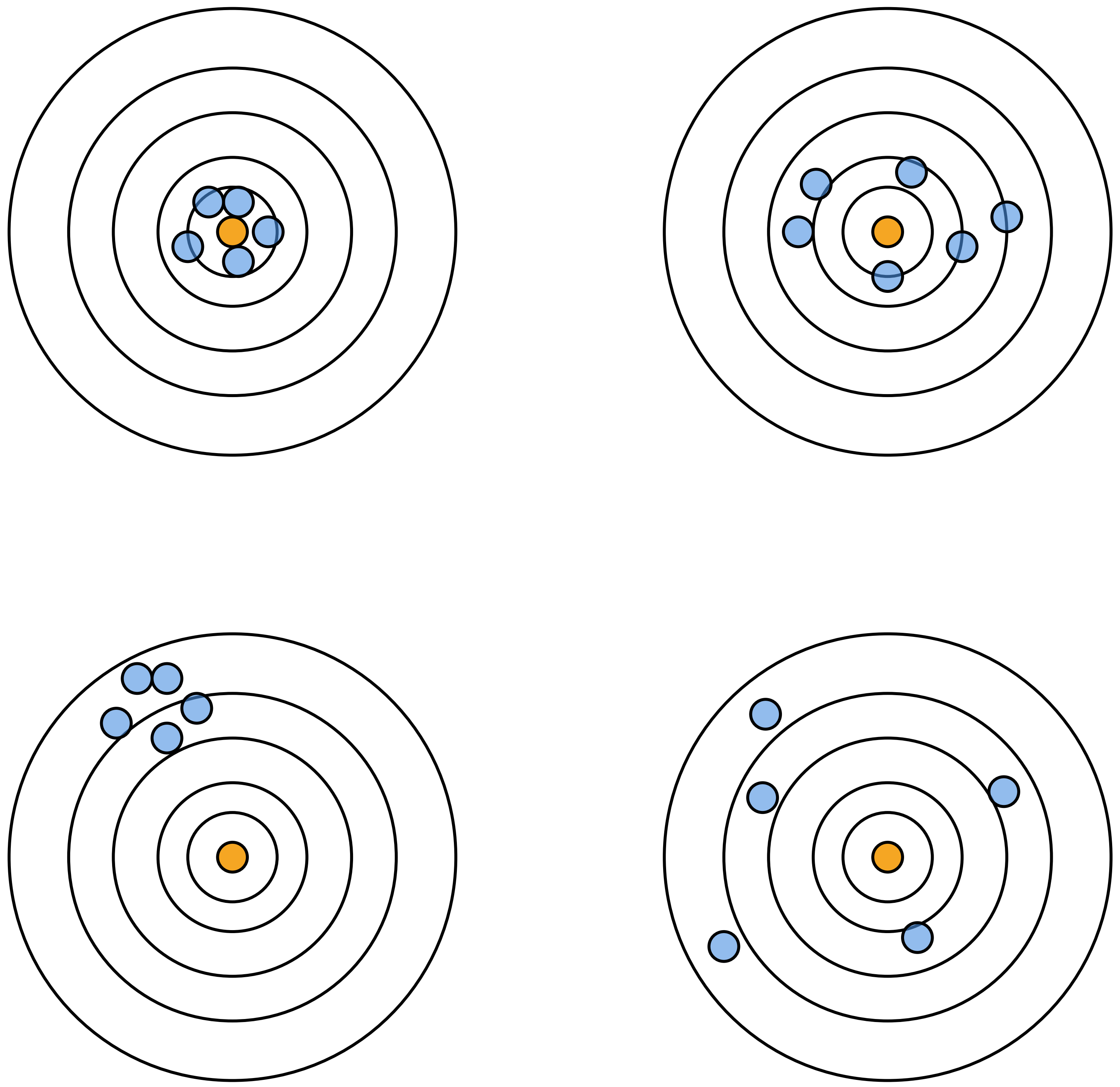
What's the bias of the estimator $\hat{\theta}_n = X_n$?

What's the bias of the estimator $\hat{\theta}_n = \dfrac{1}{n} \sum_{i=1}^{n} X_i$?

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:

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

Notice: The variance of an estimator *does not* concern its estimand (unlike bias).

Variance of Estimators

Examples of Estimators

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, $\theta = \mu$.

What's the variance of the estimator $\hat{\theta}_n = 1$?

What's the variance of the estimator $\hat{\theta}_n = X_n$?

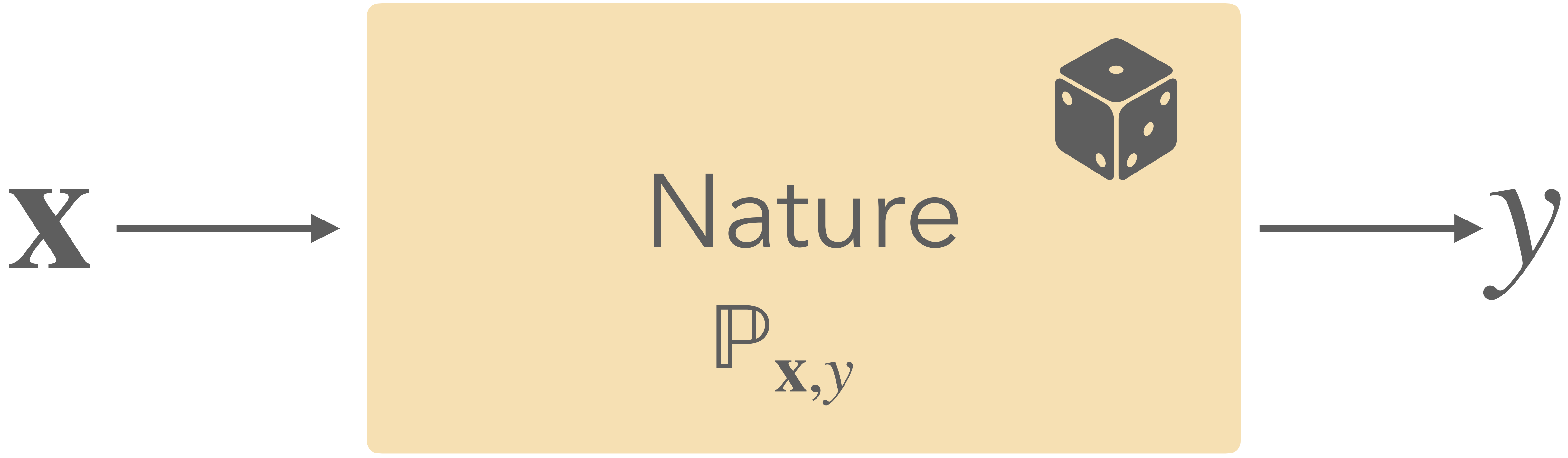
What's the variance of the estimator $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$?

Random error model

Our main assumption on $\mathbb{P}_{\mathbf{x},y}$

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

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



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

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

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

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

Bias and Variance of OLS

Corollaries from Theorem

Under the error model $y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$ the OLS estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ has the following statistical properties *conditional* on $\mathbf{X}$:

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

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

By law of total probability/tower rule, this implies that

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

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

These are a vector and a matrix, respectively.

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

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon, \text{ in the usual random error model.}$$

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\mathrm{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

By LLN: $(\mathbf{X}^\top \mathbf{X})^{-1} \sim \frac{1}{n}\mathbf{\Sigma}^{-1}$, the true covariance.

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

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon, \text{ in the usual random error model.}$$

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\mathrm{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

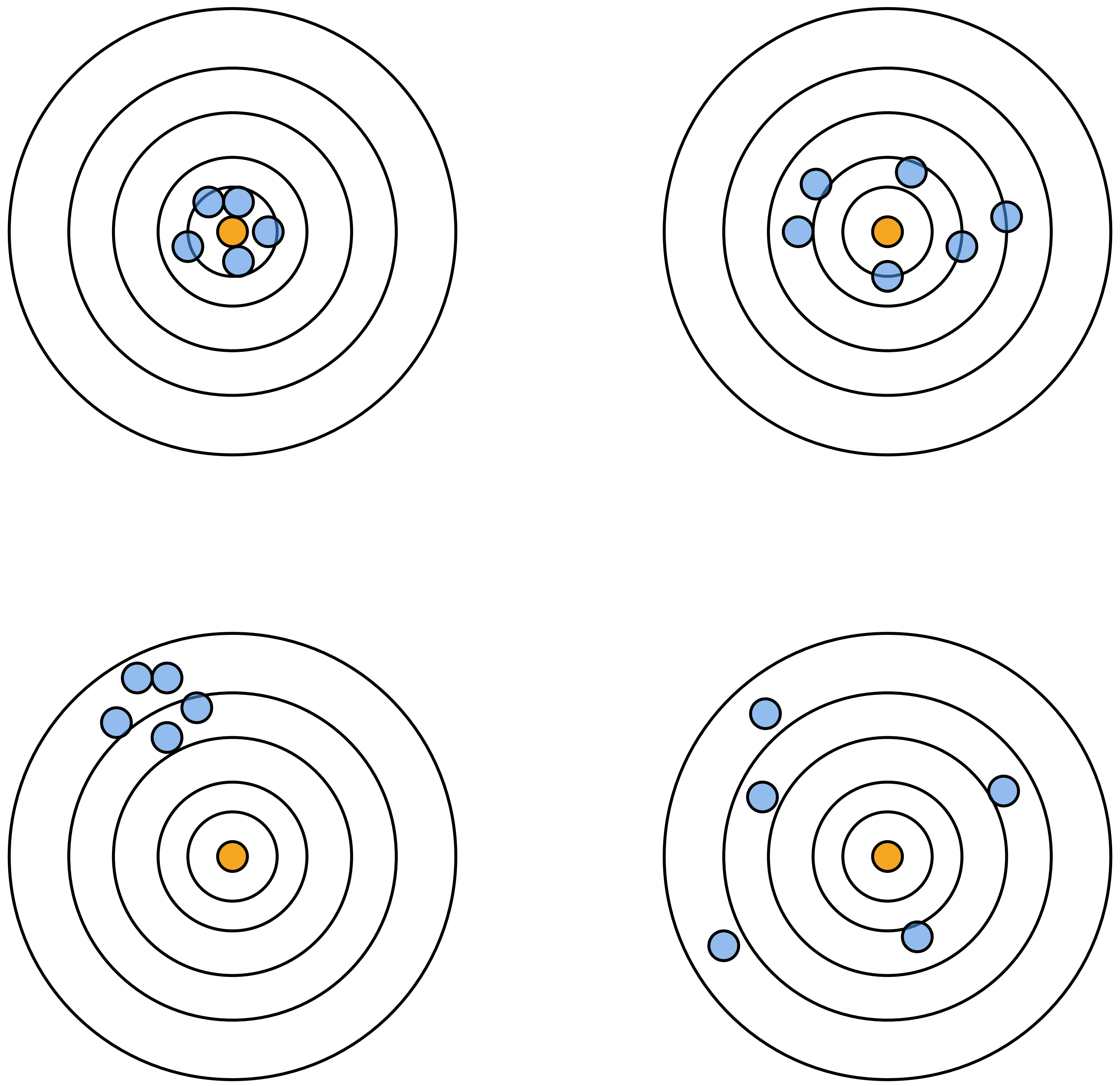
Bias vs. Variance of Estimators

Summary

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

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

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



Mean Squared Error

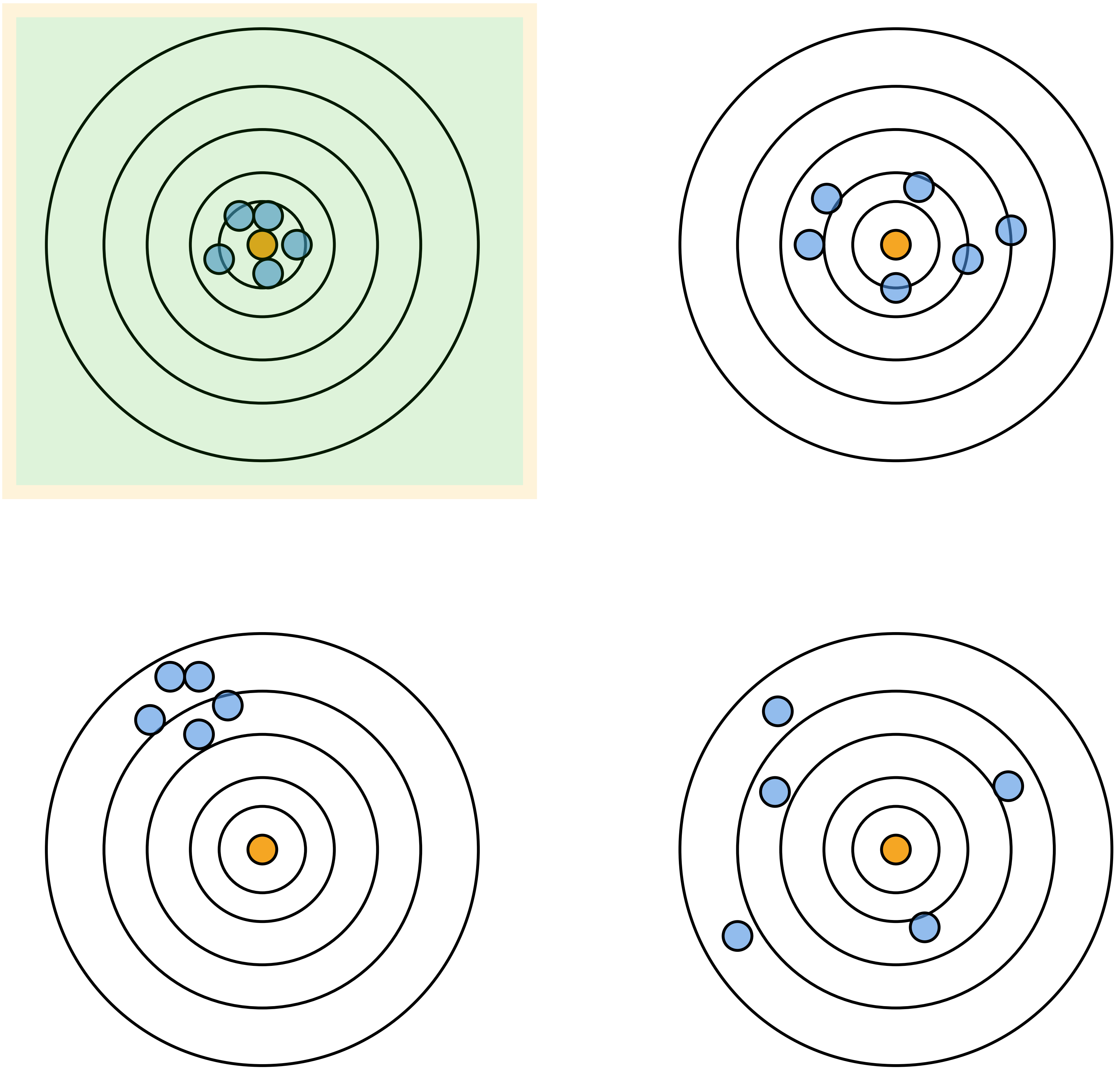
Bias-Variance Tradeoff

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 a scalar estimator $\hat{\theta}_n$ of a scalar 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 a scalar estimator of some scalar estimand θ. The bias-variance decomposition of the mean squared error of $\hat{\theta}_n$ is:

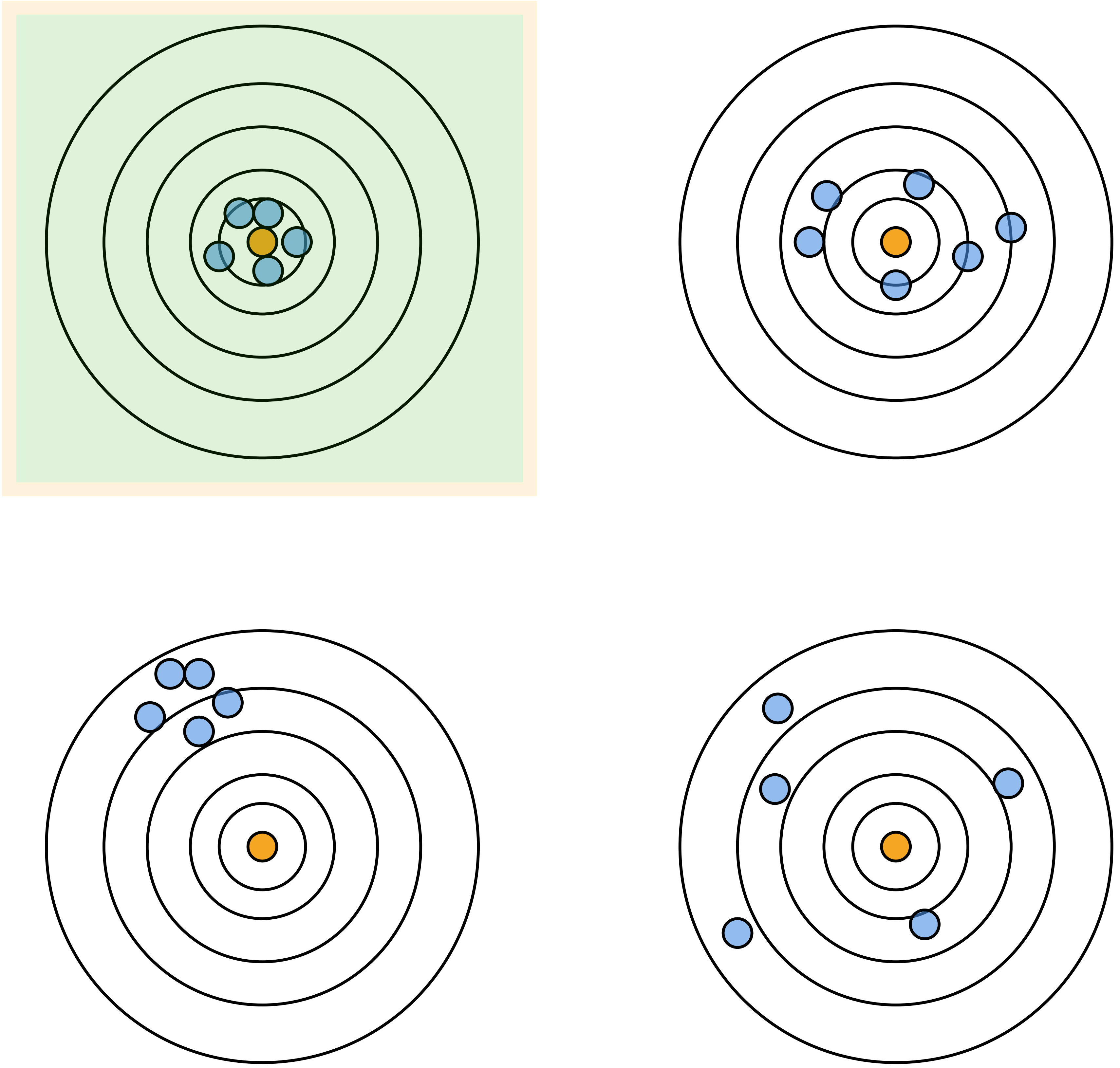
$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n).$$

Bias-Variance Decomposition

Theorem Statement

Theorem (Bias-Variance Decomposition of MSE). Let $\hat{\theta}_n$ be a scalar estimator of some scalar 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 (Scalar Version)

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:

$$\begin{aligned}
\mathbb{E}[(\hat{\theta}_n - \theta)^2] &= \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2] \quad \text{Add and subtract what you need to calculate variance.} \\
&= \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)^2] + 2(\overline{\theta}_n - \theta)\mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)] + \mathbb{E}[(\overline{\theta}_n - \theta)^2] \\
&= (\overline{\theta}_n - \theta)^2 + \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)^2] \quad \text{Notice what is random and non-random.} \\
&= (\mathbb{E}[\hat{\theta}_n] - \theta)^2 + \mathbb{E}[(\hat{\theta}_n - \overline{\theta}_n)^2] = \mathrm{Bias}(\hat{\theta}_n)^2 + \mathrm{Var}(\hat{\theta}_n)
\end{aligned}$$

Bias-Variance Decomposition

Theorem Statement (General)

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

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

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

Sum of diagonal entries of covariance matrix!

Trace

Definition

For any square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, the trace of $\mathbf{A}$, denoted $\mathbf{tr}(\mathbf{A})$, is the sum of its diagonal:

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

For any scalar, $a = a^\top = \operatorname{tr}(a)$.

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

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

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 $\overline{X}_n := \dfrac{1}{n}\sum_{i=1}^{n} X_i$?

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

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

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

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

$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$, in the usual random error model.

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\mathrm{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

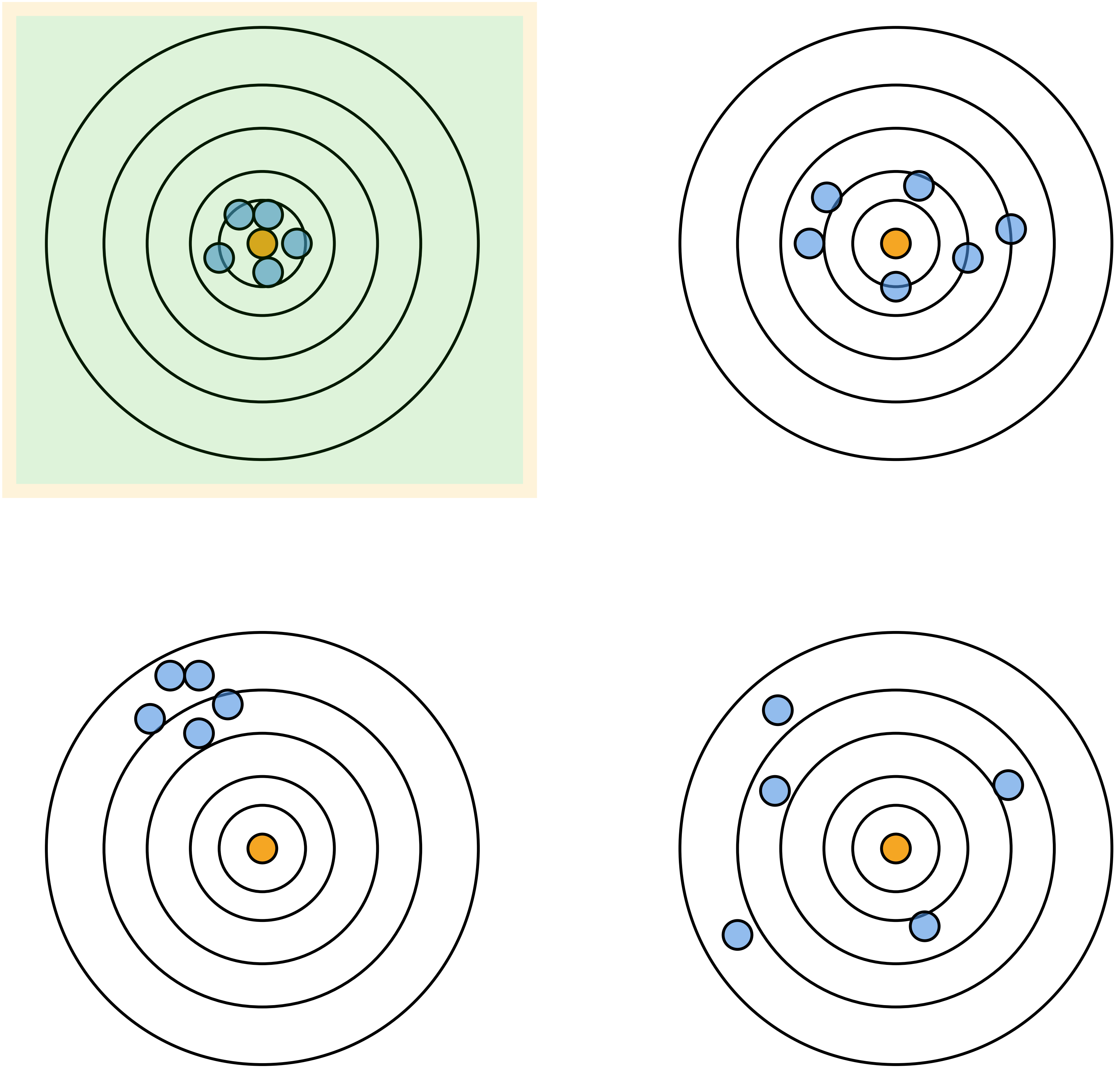
Parameter MSE: $\mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \sigma^2 \mathbb{E}[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^{-1})]$

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:

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



Bias vs. Variance

Stochastic Gradient Descent

Gradient Descent

Algorithm

Initialize at a randomly chosen $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

For iteration $t = 1,2,\ldots,T$:

$$\mathbf{w}^{(t)} \leftarrow \mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$$

Return final $\mathbf{w}^{(T)}$, with objective value $f(\mathbf{w}^{(T)})$.

Gradient Descent

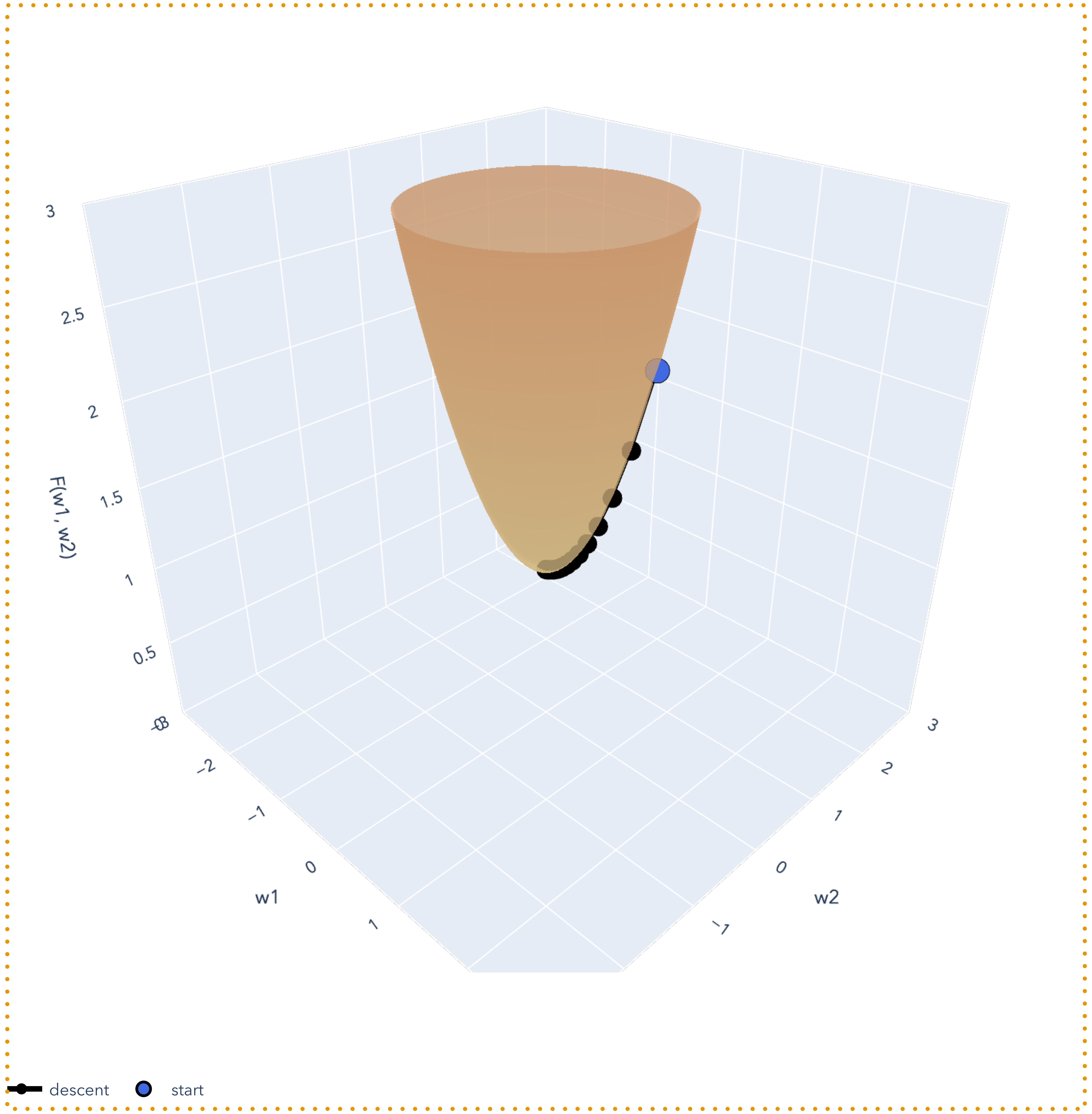
Algorithm for OLS

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

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

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

Computationally expensive,
depends on *entire* dataset.



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?

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

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 batch of k examples $B = \{i_1, \ldots, i_k\}$uniformly from all k-subsets of $1,\ldots,n$:

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

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:

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

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

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

That's exactly what we're estimating!

Stochastic Gradient Descent

Single-sample SGD for OLS

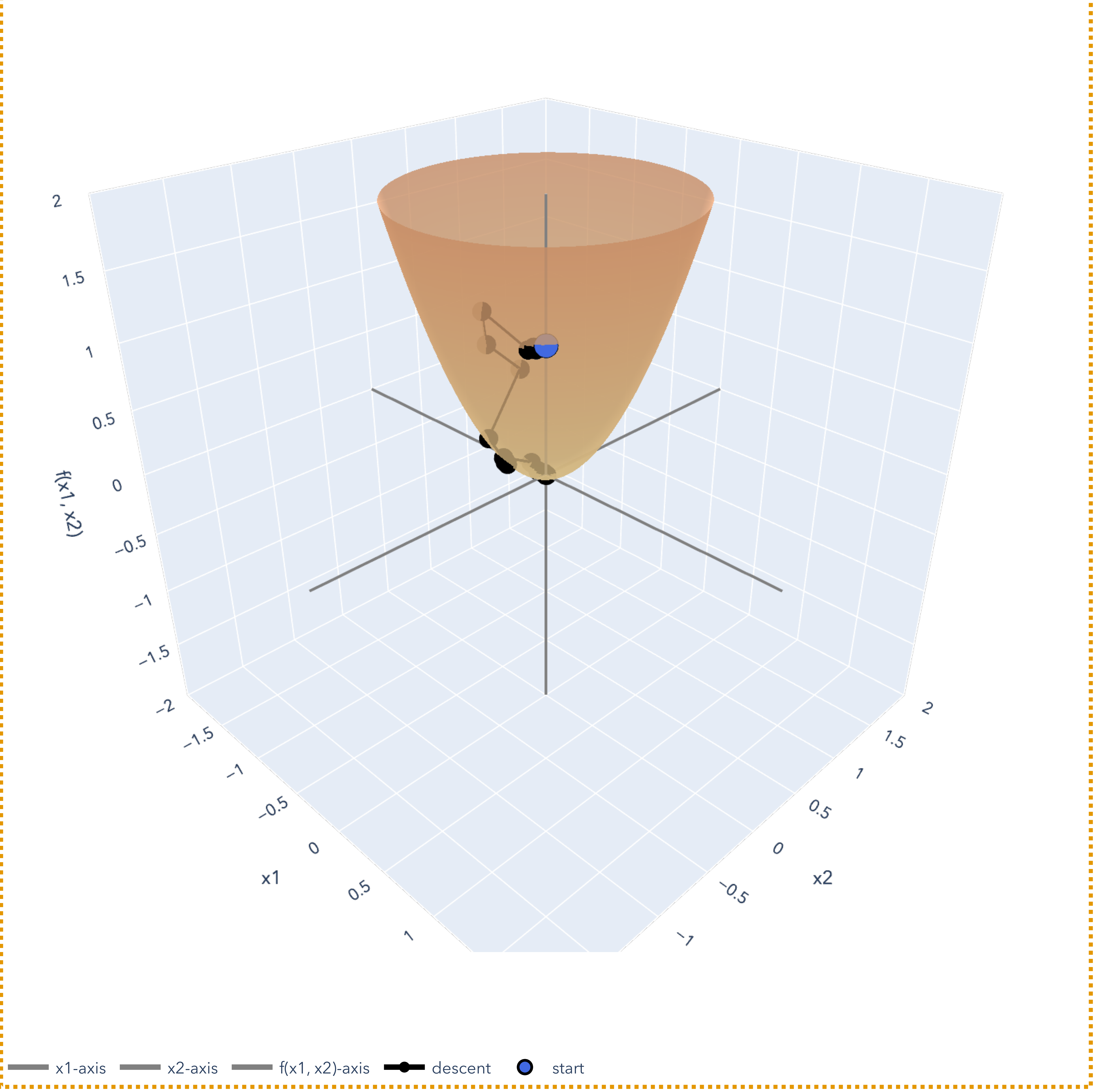
Make an initial guess $\mathbf{w}_0$.

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

Choose $i \sim [n]$ uniformly at random.

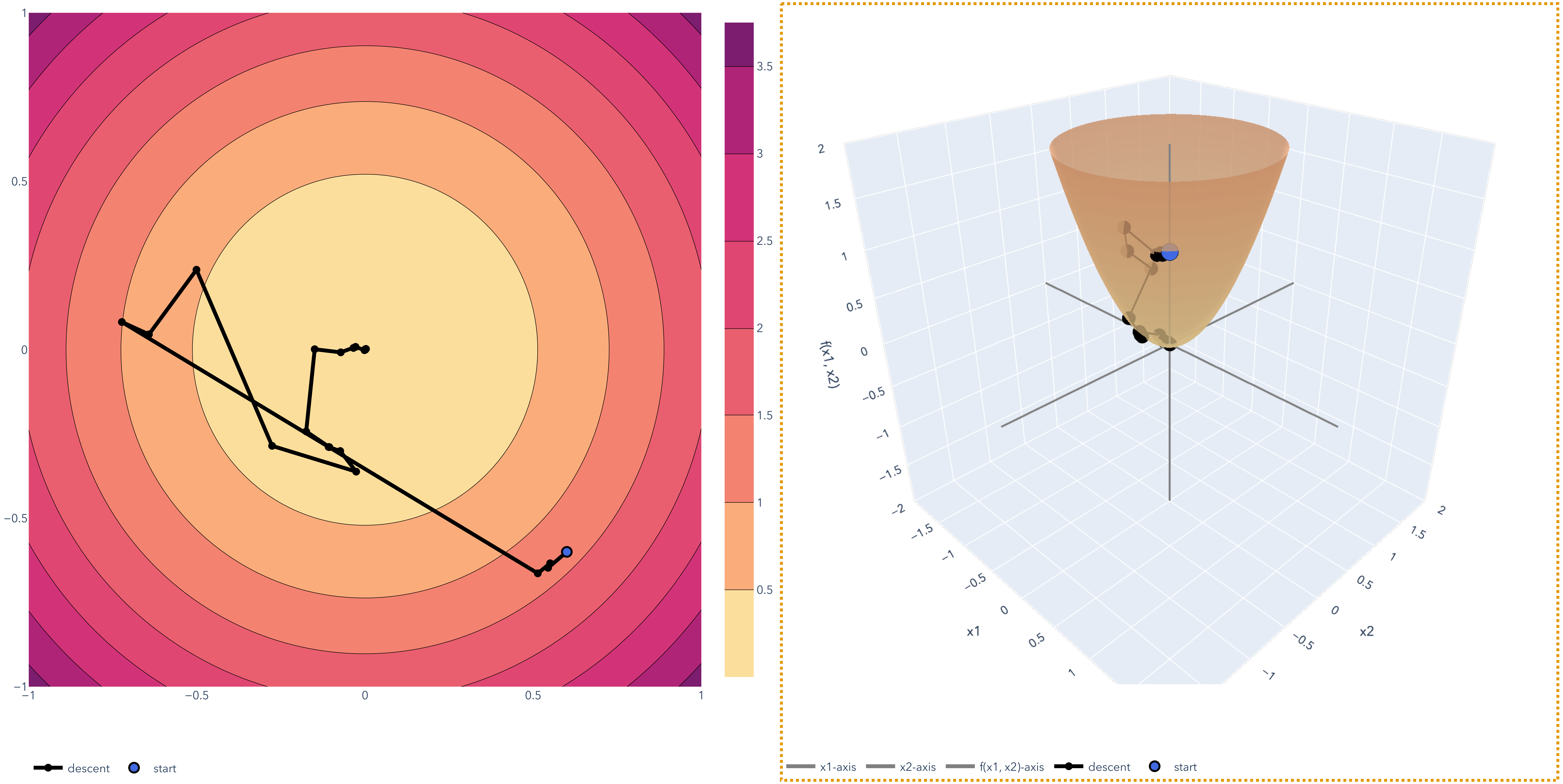
Compute: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \eta \nabla_{\mathbf{w}} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$.

Estimator of the gradient.



Stochastic Gradient Descent

Single-sample SGD for OLS



Stochastic Gradient Descent

Minibatch SGD

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

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

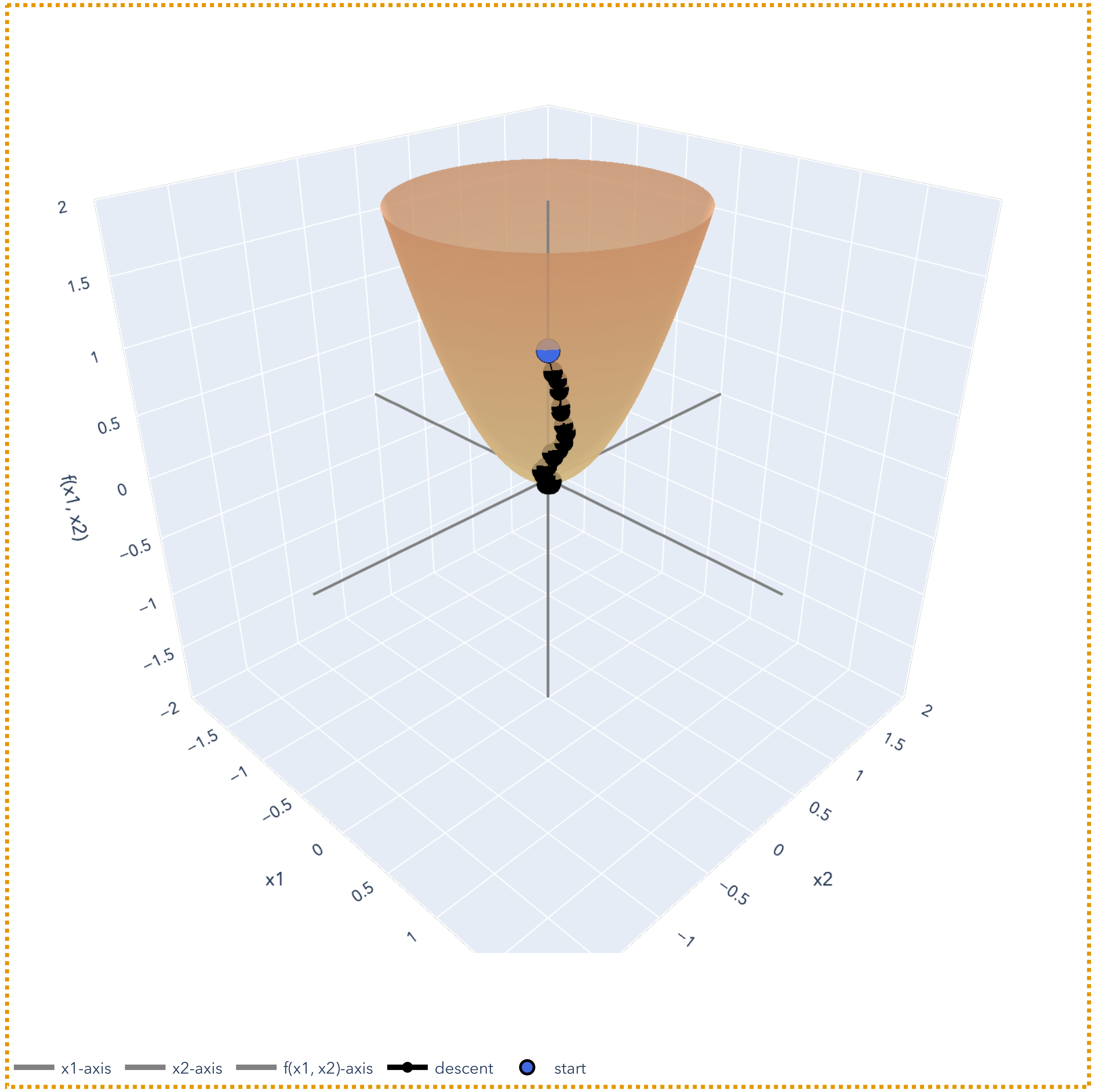
Sample k indices $B = \{i_1, \ldots, i_k\}$ uniformly.

Compute:

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

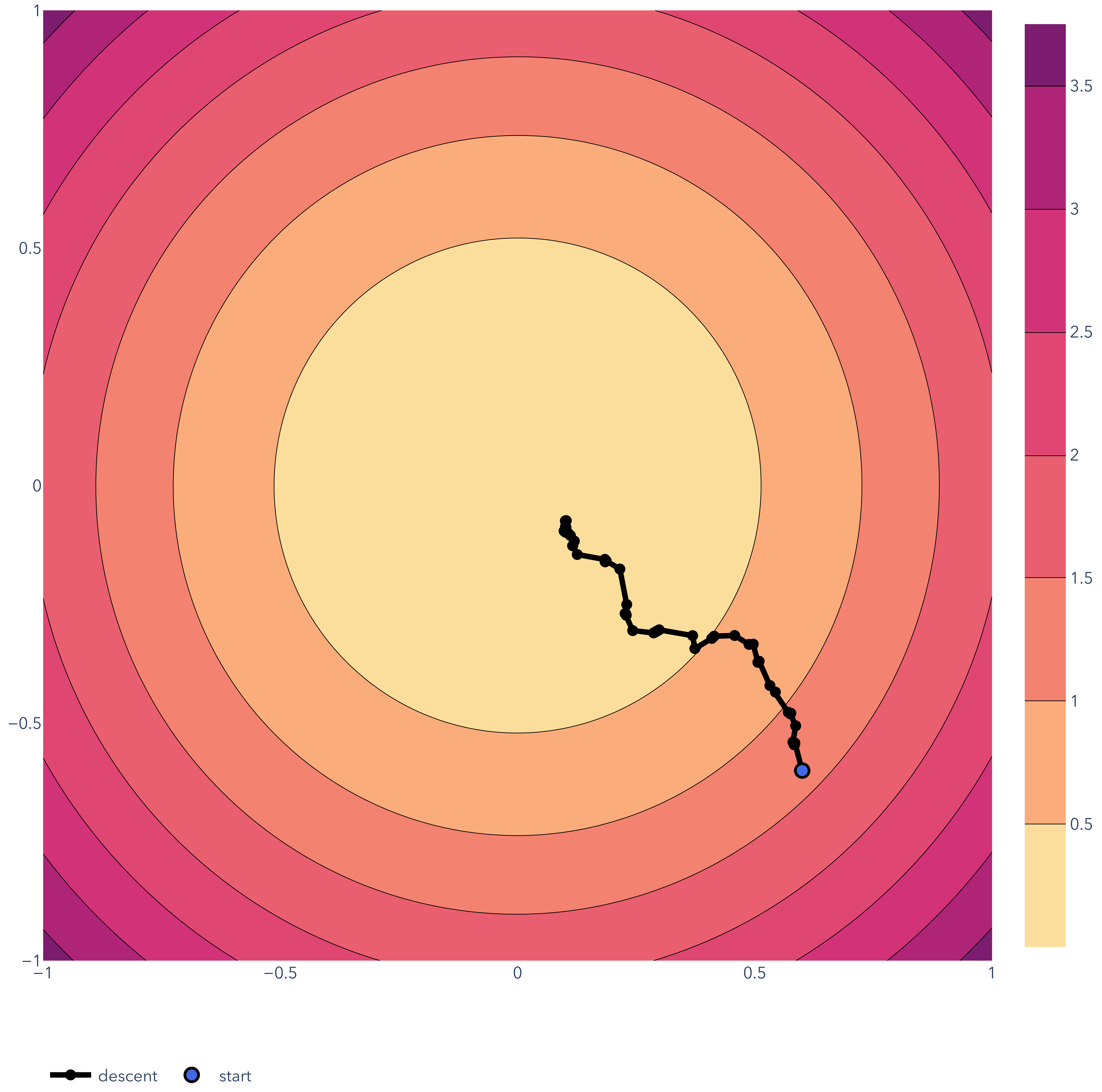
Estimator of the gradient.

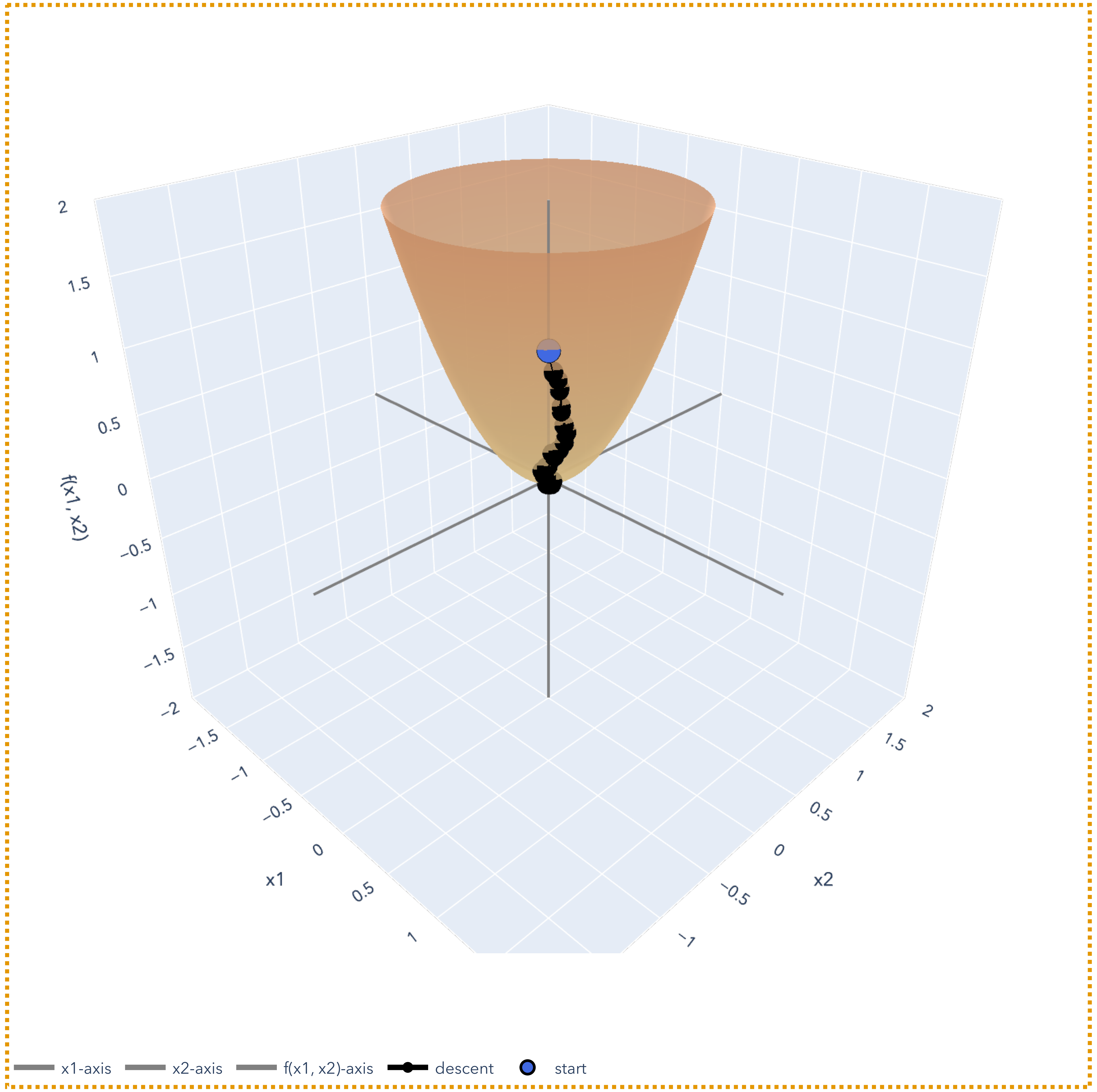
Still unbiased, but improves the **variance**!



Stochastic Gradient Descent

Minibatch SGD





Bias vs. Variance

Ridge Regression

Least Squares

OLS Theorem

Theorem (Ordinary Least Squares). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

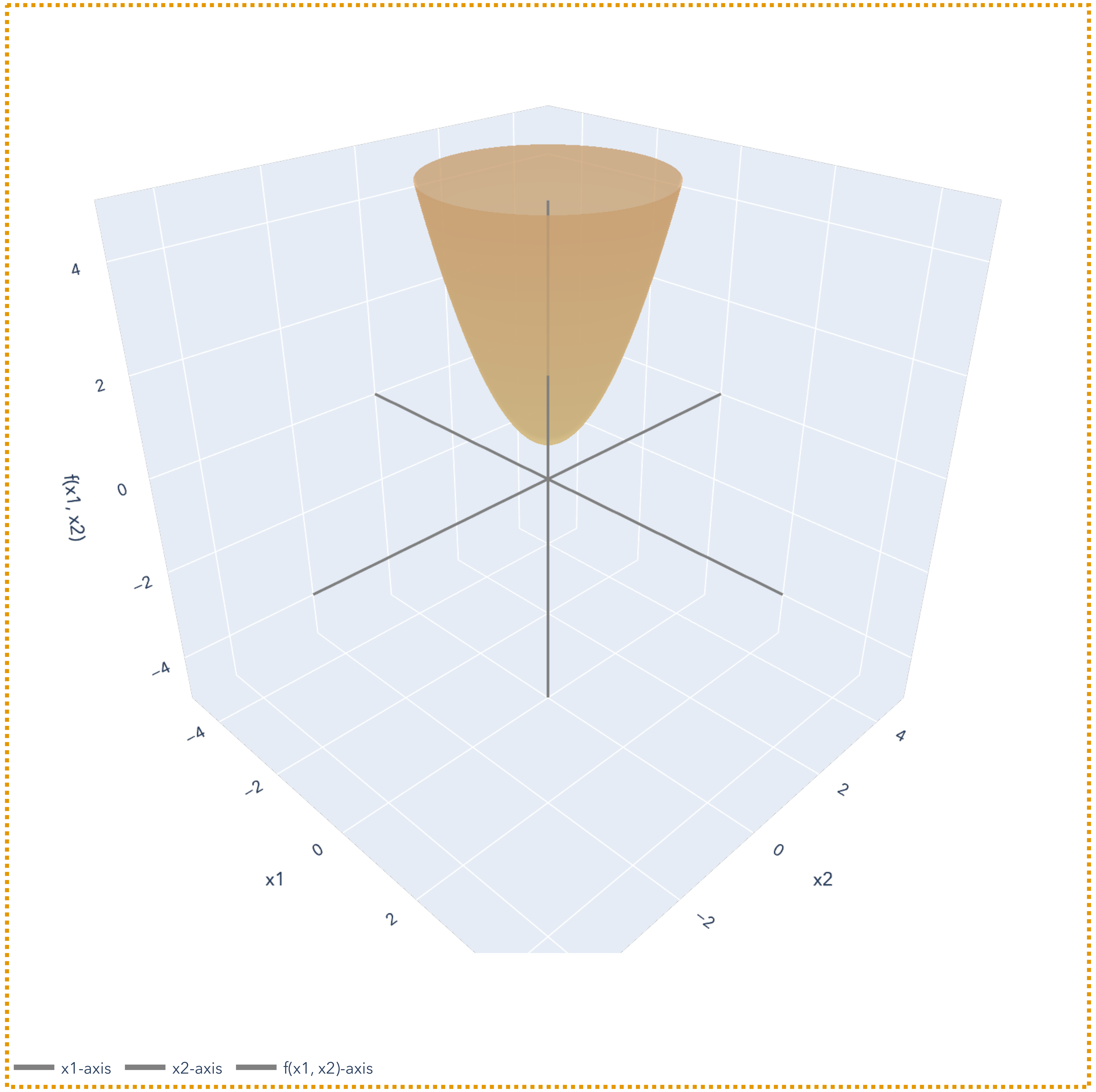
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

If $n \geq d$ and $\mathrm{rank}(\mathbf{X}) = d$, then:

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

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

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



Least Squares

Ridge Regression

Our goal will now be to minimize two objectives:

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

Writing this as an optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

where $\gamma > 0$ is a fixed tuning parameter.

This optimization problem is known as ridge/Tikhonov/ℓ_2-regularized regression.

Least Squares

Ridge Regression

Our goal will now be to minimize two objectives:

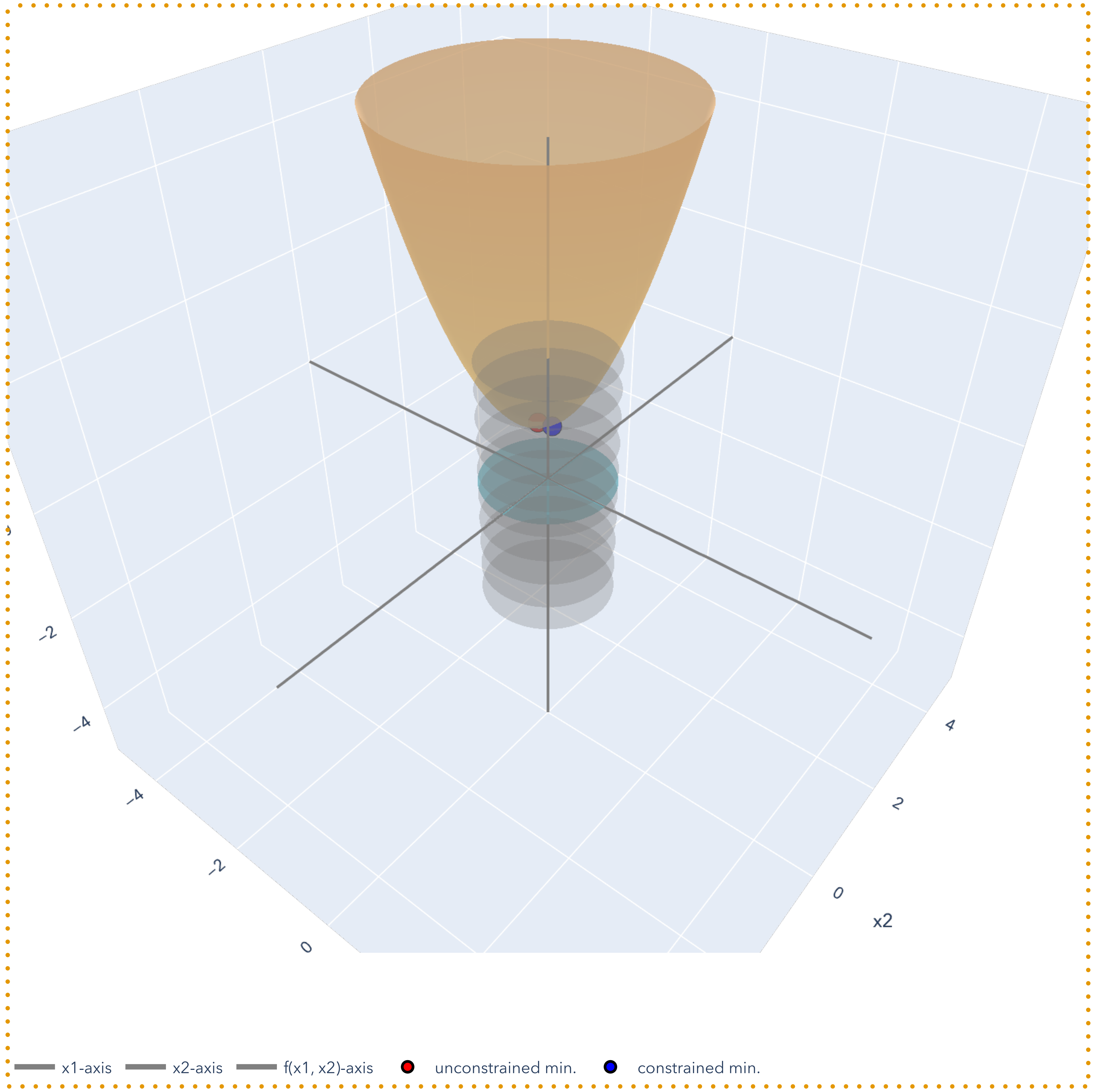
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

Writing this as an optimization problem:

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma \|\mathbf{w}\|^2$$

where $\gamma > 0$ is a fixed tuning parameter.

This optimization problem is known as ridge/Tikhonov/ℓ_2-regularized regression.



Least Squares

Ridge Regression

Our goal will now be to minimize two objectives:

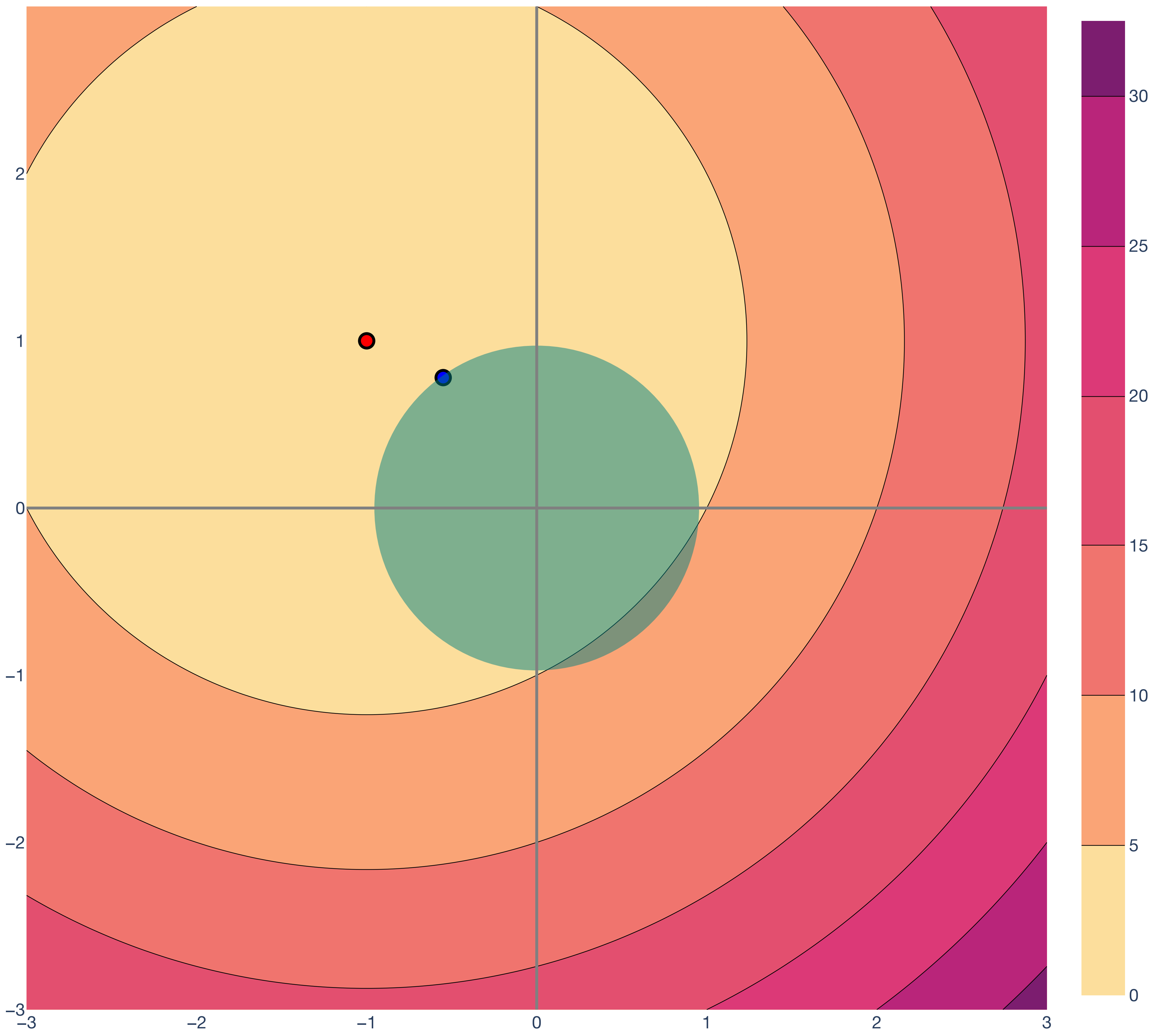
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

Writing this as an optimization problem:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

where $\gamma > 0$ is a fixed tuning parameter.

This optimization problem is known as ridge/Tikhonov/ℓ_2-regularized regression.



For bigger γ, bigger "constraint" ball!

Ridge Regression

Property: PSD to PD matrices

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

How do we solve this using the first and second order conditions?

Property (Perturbing PSD matrices). Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. Then, for any $\gamma > 0$, the matrix $\mathbf{A} + \gamma\mathbf{I}$ is positive definite.

Proof. Let $\mathbf{v} \in \mathbb{R}^d$ be any vector. $\mathbf{v}^\top(\mathbf{A} + \gamma\mathbf{I})\mathbf{v} = \mathbf{v}^\top(\mathbf{A}\mathbf{v} + \gamma\mathbf{v}) = \mathbf{v}^\top\mathbf{A}\mathbf{v} + \gamma\mathbf{v}^\top\mathbf{v}$

$$= \underbrace{\mathbf{v}^\top\mathbf{A}\mathbf{v}}_{\geq 0} + \underbrace{\gamma\|\mathbf{v}\|^2}_{>0 \text{ unless } \mathbf{v}=\mathbf{0}.}$$

Ridge Regression

First-order conditions

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

Take the gradient and set to $\mathbf{0}$:

$$\nabla_{\mathbf{w}}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \nabla_{\mathbf{w}}\|\mathbf{w}\|^2 = 2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{y} + 2\gamma\mathbf{w}$$

$$2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{y} + 2\gamma\mathbf{w} = \mathbf{0} \implies (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})\mathbf{w} = \mathbf{X}^\top\mathbf{y}$$

By property (perturbing PSD matrices), $\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I}$ is PD, so:

$$\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Least Squares

Solving ridge regression

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\gamma\|\mathbf{w}\|^2$$

Candidate minimizer: $\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$.

Gradient: $\nabla_{\mathbf{w}}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\nabla_{\mathbf{w}}\|\mathbf{w}\|^2 = 2\mathbf{X}^\top\mathbf{X}\mathbf{w}-2\mathbf{X}^\top\mathbf{y}+2\gamma\mathbf{w}$

Taking the Hessian,

$\nabla^2 f(\mathbf{w}) = \mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I}$, which is positive definite.

Sufficient condition for optimality applies!

Ridge Regression

Theorem

<u>Theorem (Ridge Regression).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, and $\gamma > 0$. Then,

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

has the form:

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

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

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

Least Squares

Comparison with ridge solution

<u>Theorem (Ridge Regression).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, and $\gamma > 0$. Then, the ridge minimizer:

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

has the form:

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

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

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

<u>Theorem (Ordinary Least Squares).</u> Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ be the least squares minimizer:

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

If $n \geq d$ and $\text{rank}(\mathbf{X}) = d$, then:

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

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

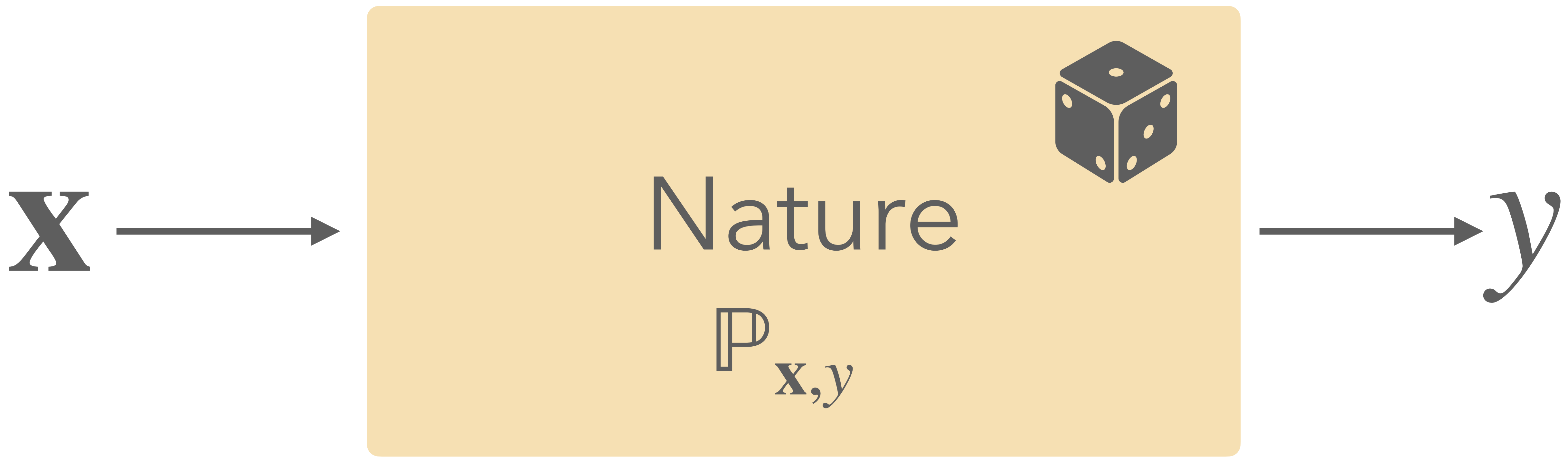
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Random error model

Our main assumption on $\mathbb{P}_{\mathbf{x},y}$

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

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



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

$$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon, \text{ in the usual random error model.}$$

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\text{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

Parameter MSE: $\text{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \sigma^2 \mathbb{E}[\text{tr}((\mathbf{X}^\top \mathbf{X})^{-1})]$

Mean Squared Error (MSE)

Analysis for Least Squares

For $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, the mean squared error is:

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

by the bias-variance decomposition because $\mathrm{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

Mean Squared Error (MSE)

Eigendecomposition analysis

We know that $\mathbf{X}^\top\mathbf{X}$ (the *covariance matrix*) is PSD, so it is diagonalizable:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top \implies (\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{V}^\top\boldsymbol{\Lambda}^{-1}\mathbf{V}\,.$$

The inverse of the diagonal matrix $\boldsymbol{\Lambda}^{-1}$:

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\lambda_d \end{bmatrix}$$, so if λ_i is small, $\mathbb{E}[\mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1})]$ might blow up!

Mean Squared Error (MSE)

Analysis for Ridge Regression

For $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$, the mean squared error is:

$$\mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \mathrm{Bias}(\hat{\mathbf{w}})^2 + \mathrm{tr}(\mathrm{Var}(\hat{\mathbf{w}}))$$

$$\mathrm{Bias}(\hat{\mathbf{w}})^2 = \|\mathbb{E}[\hat{\mathbf{w}}] - \mathbf{w}^*\|^2 = \|((\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} - \mathbf{I})\mathbf{w}^*\|^2$$

$$\mathrm{Var}(\hat{\mathbf{w}}) = \sigma^2 \mathrm{tr} \left[\mathbb{E} \left[(\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{I})^{-1} \right] \right]$$

Error in Ridge Regression

Eigendecomposition perspective

Ridge weights: $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$.

We know that $\mathbf{X}^\top\mathbf{X}$ is positive semidefinite, so it is diagonalizable:

$$\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top + \mathbf{V}(\gamma\mathbf{I})\mathbf{V}^\top \implies (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1} = \mathbf{V}^\top(\boldsymbol{\Lambda} + \gamma\mathbf{I})^{-1}\mathbf{V}.$$

The inverse of the diagonal matrix $(\boldsymbol{\Lambda} + \gamma\mathbf{I})^{-1}$:

$$(\boldsymbol{\Lambda} + \gamma\mathbf{I})^{-1} = \begin{bmatrix} \frac{1}{\lambda_1+\gamma} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_d+\gamma} \end{bmatrix}, \text{ so } \frac{1}{\lambda_i+\gamma} \text{ entries are never bigger than } \frac{1}{\gamma}!$$

Least Squares

Ridge Regression

<u>Theorem (Ridge Regression).</u> Let $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^{n}$, and $\gamma > 0$. Then,

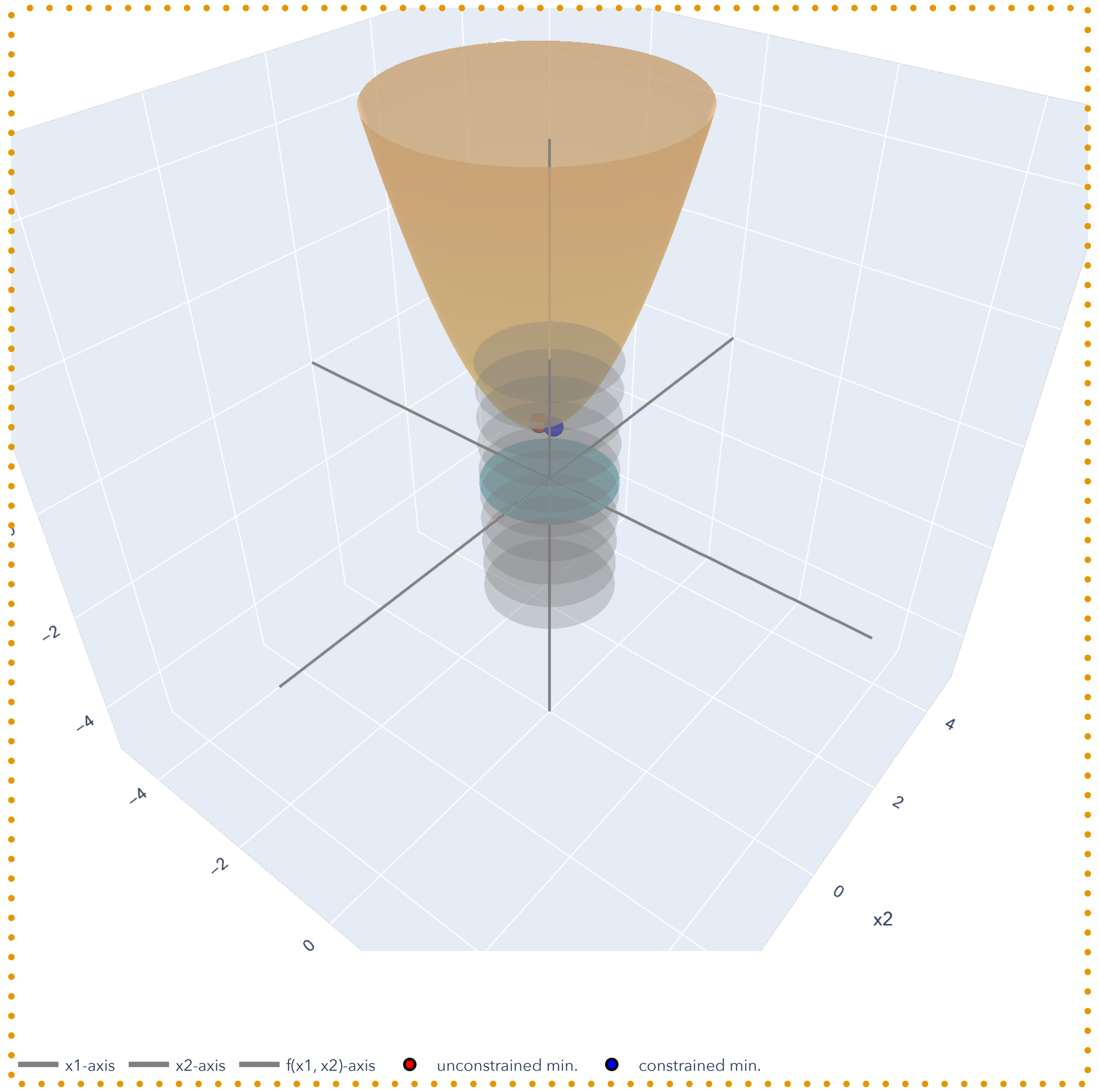
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

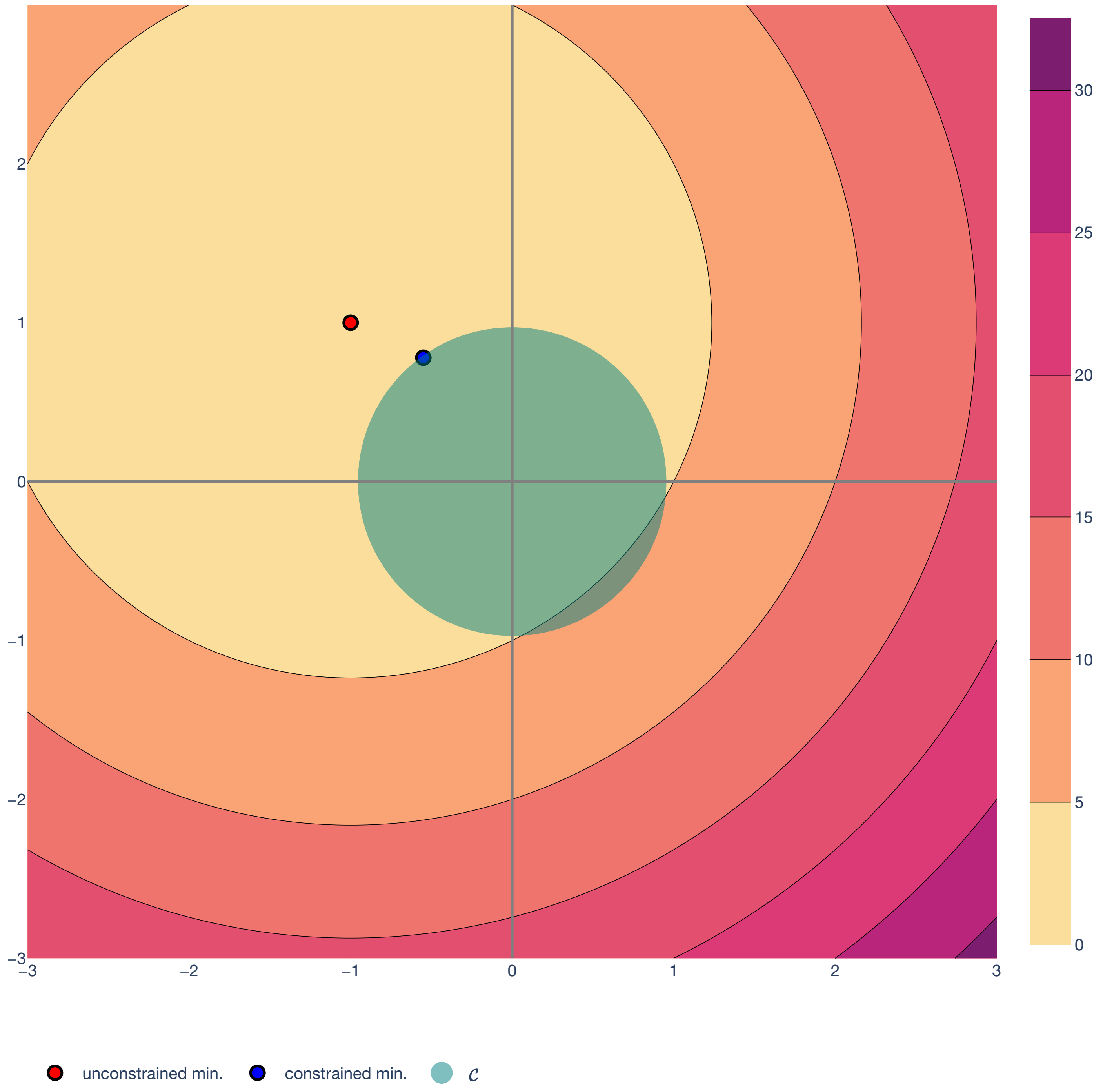
has the form:

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

To get predictions $\hat{\mathbf{y}} \in \mathbb{R}^n$:

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





For lower γ, smaller "constraint" ball: higher bias but lower variance!

Regression

Statistical analysis of risk

Statistics of OLS

Theorem

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

$y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$, in the usual random error model.

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\mathrm{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

Parameter MSE: $\mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \sigma^2 \mathbb{E}[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^{-1})]$

Almost what we want! This is a measure of "distance to $\mathbf{w}^*$" but **not** its accuracy on a new example.

Regression

Setup, with randomness

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

$$R(\hat{f}) := R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x} - y)^2]$$

Note that this is different from the MSE!

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

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

This is what we've been doing!

Regression

Risk vs. MSE

This risk is how well $\hat{\mathbf{w}}$ does *on average on a new example* with respect to squared error:

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

This mean squared error (MSE) is how "far" $\hat{\mathbf{w}}$ is from $\mathbf{w}$ on average:

$$\text{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \sigma^2 \mathbb{E}[\text{tr}((\mathbf{X}^\top \mathbf{X})^{-1})]$$

Conjecture: If $y = \mathbf{x}^\top \mathbf{w} + \epsilon$, then maybe risk is just MSE plus "unavoidable randomness?"

Statistical Analysis of Risk

Theorem Statement

<u>Theorem (Risk of OLS).</u> 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 $\mathbb{P}_{\mathbf{x},y}$ and $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}^\top \mathbf{x}] = \mathrm{Var}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is the true covariance.

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

LLN: $(\mathbf{X}^\top \mathbf{X})^{-1} \approx \frac{1}{n} \mathbf{\Sigma}^{-1}$ as $n \to \infty$.

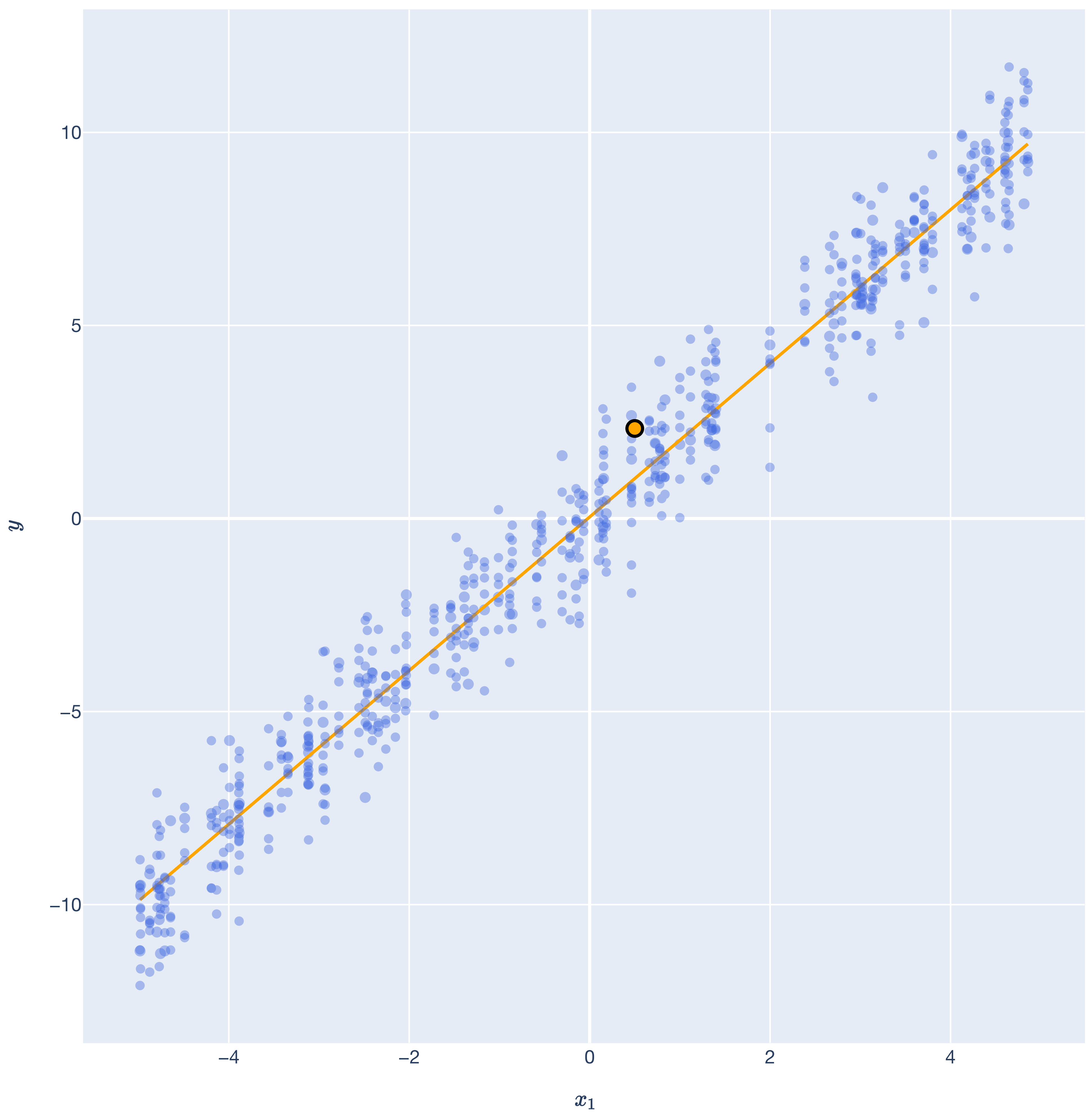
This is "unavoidable" randomness from ϵ!

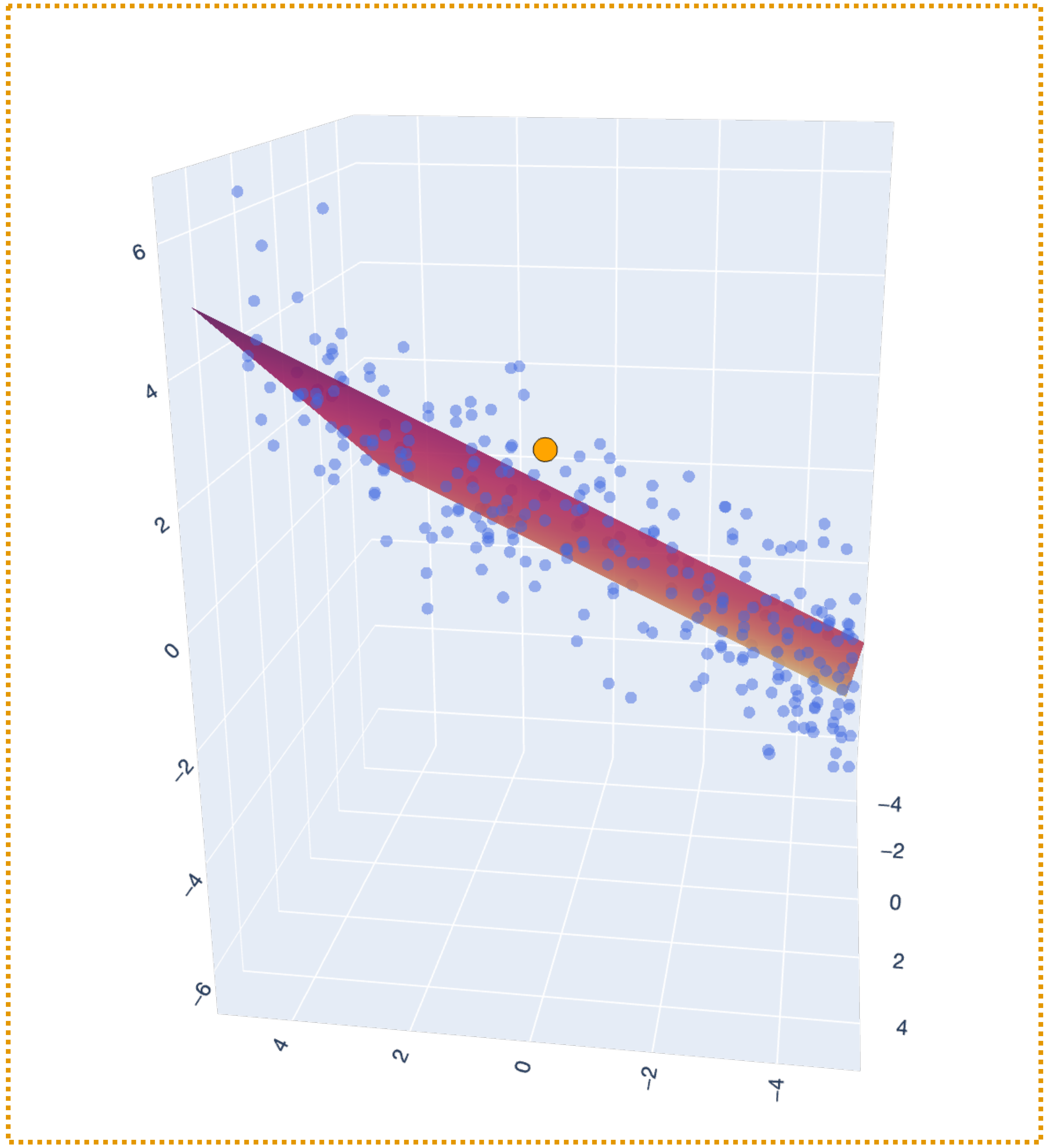
Notice similarity to MSE!

$$R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x} - y)^2] = \sigma^2 + \sigma^2 \mathbb{E}[\mathrm{tr}(\mathbf{\Sigma}(\mathbf{X}^\top \mathbf{X})^{-1})] \approx \sigma^2 + \frac{\sigma^2 d}{n}.$$

Risk of OLS

$d = 1$ and $d = 2$





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 $y = \mathbf{x}^\top \mathbf{w}^* + \epsilon$, in the usual random error model.

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

Expectation: $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}^*$ and $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$, so $\text{Bias}(\hat{\mathbf{w}}) = \mathbf{0}$.

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

Parameter MSE: $\text{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2] = \sigma^2 \mathbb{E}[\text{tr}((\mathbf{X}^\top \mathbf{X})^{-1})]$

Risk (w.r.t. squared error): $R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x} - y)^2] = \sigma^2 + \sigma^2 \mathbb{E}[\text{tr}(\mathbf{\Sigma}(\mathbf{X}^\top \mathbf{X})^{-1})] \approx \sigma^2 + \dfrac{\sigma^2 d}{n}$.

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.

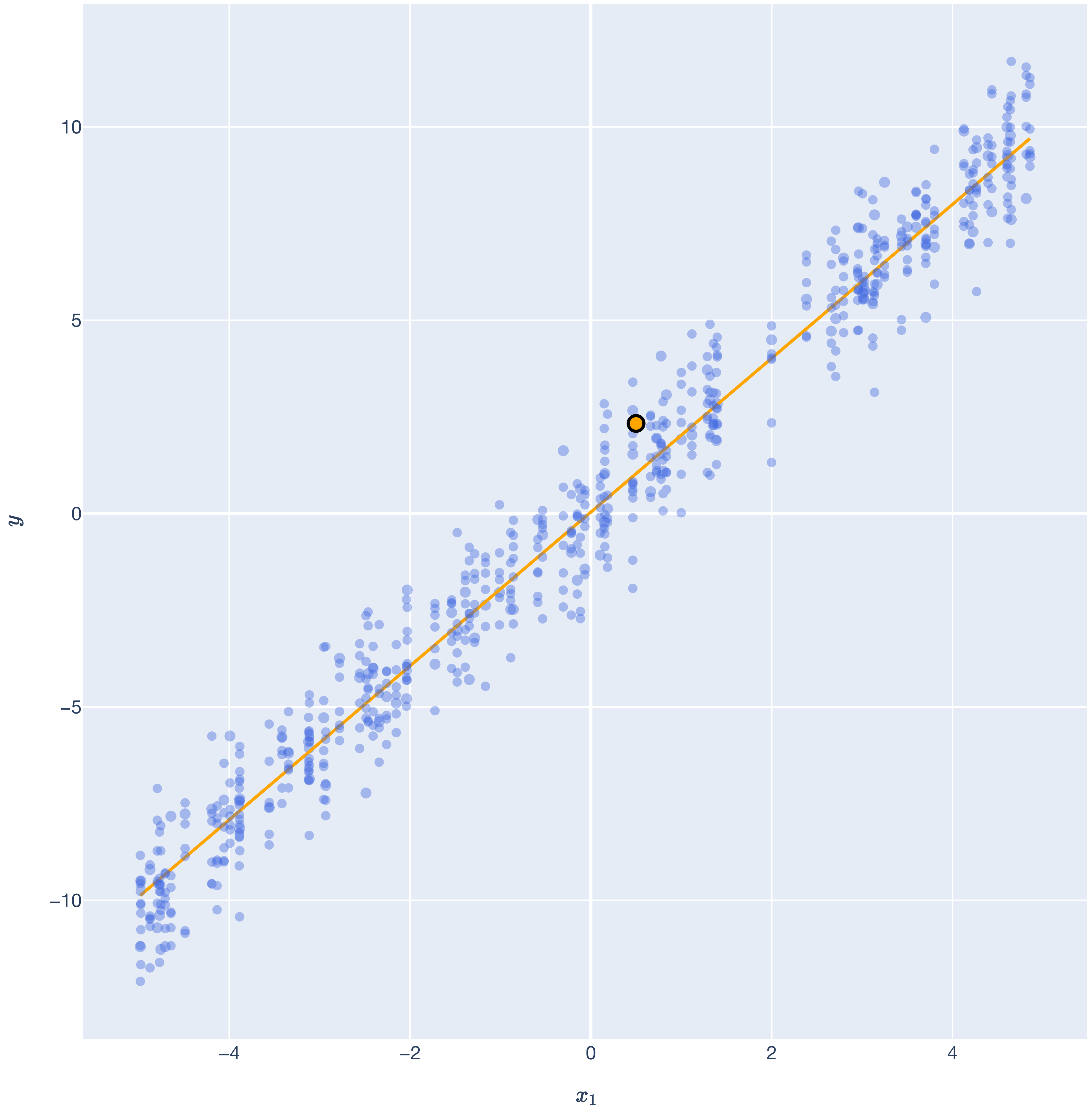
Bias, variance, and MSE. Two important properties of statistical estimators are their *bias* and *variance*, which are measures of how good the estimator is at guessing the target. These form the estimator's MSE.

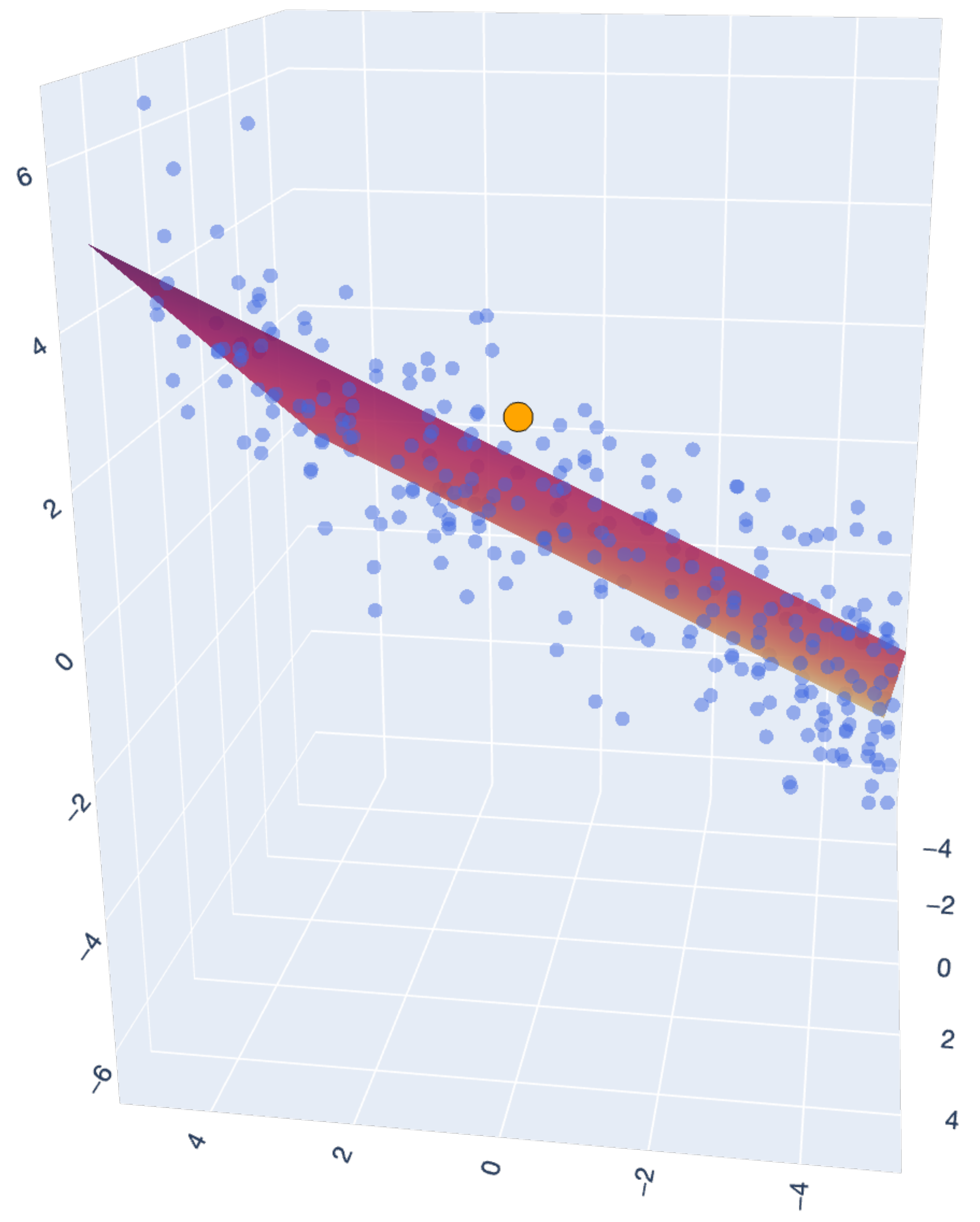
Stochastic gradient descent (SGD). Gradient descent needs to take a gradient over all n training examples, which may be large; SGD *estimates* the gradient to speed up the process.

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