Math for ML

Week 6.2: Multivariate Gaussian Distribution

By: Samuel Deng

Logistics & Announcements

Lesson Overview

OLS under Gaussian Error Model. The distribution of $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ under the Gaussian error model is itself multivariate normal.

Multivariate Gaussian/Normal (MVN) Distribution PDF. We define the multivariate Gaussian/normal distribution and study some simple examples.

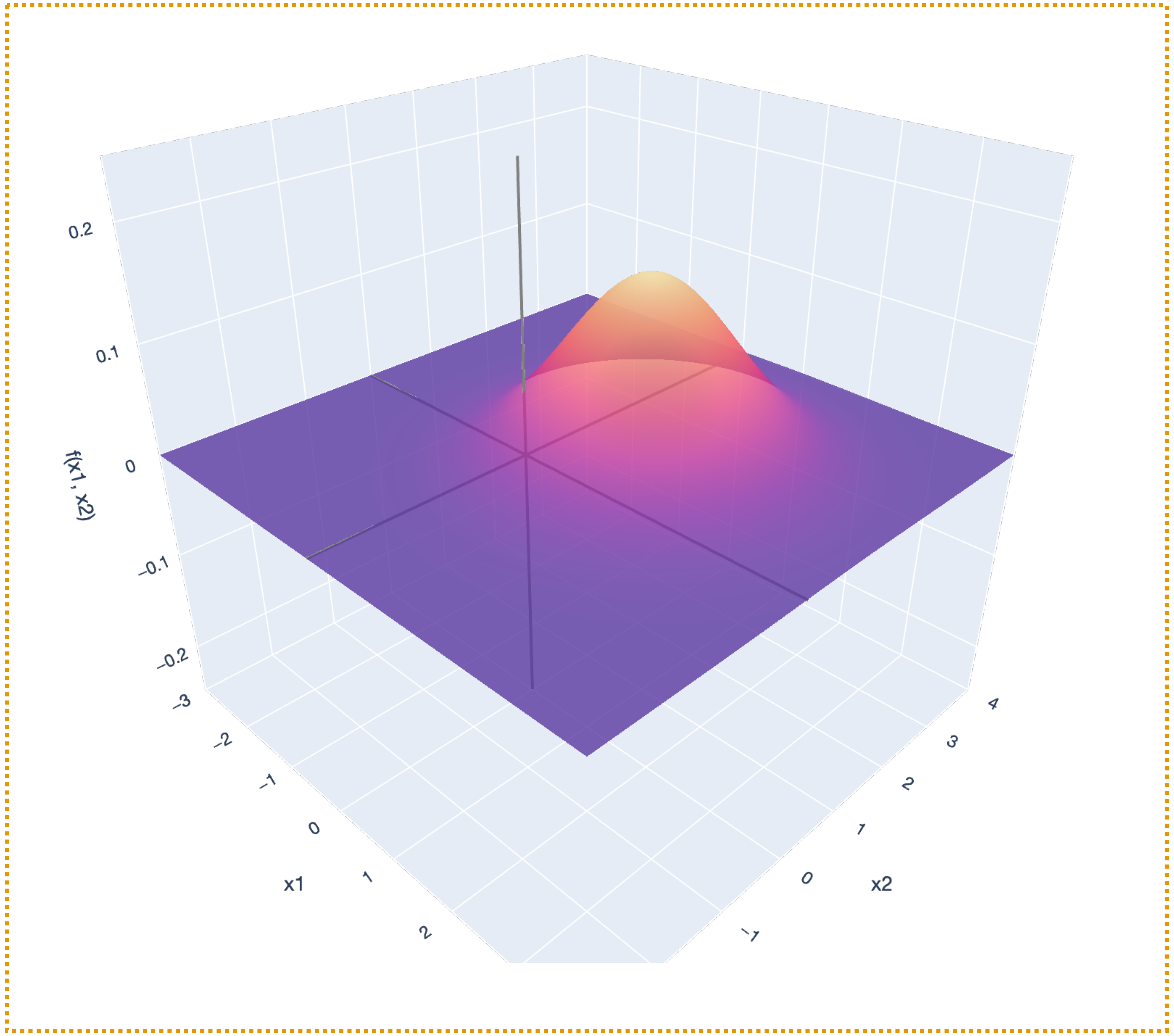
Factorization of the Multivariate Gaussian. We see that a multivariate Gaussian with a diagonal covariance matrix factors into independent Gaussians.

Geometry of the Multivariate Gaussian. We study the geometry of the multivariate Gaussian through its level curves and discover the it is ellipsoidal, with axes determined by the eigenvectors/eigenvalues of the covariance matrix.

Affine Transformations of the Multivariate Gaussian. We establish that any multivariate Gaussian is just an affine transformation away from the standard multivariate Gaussian.

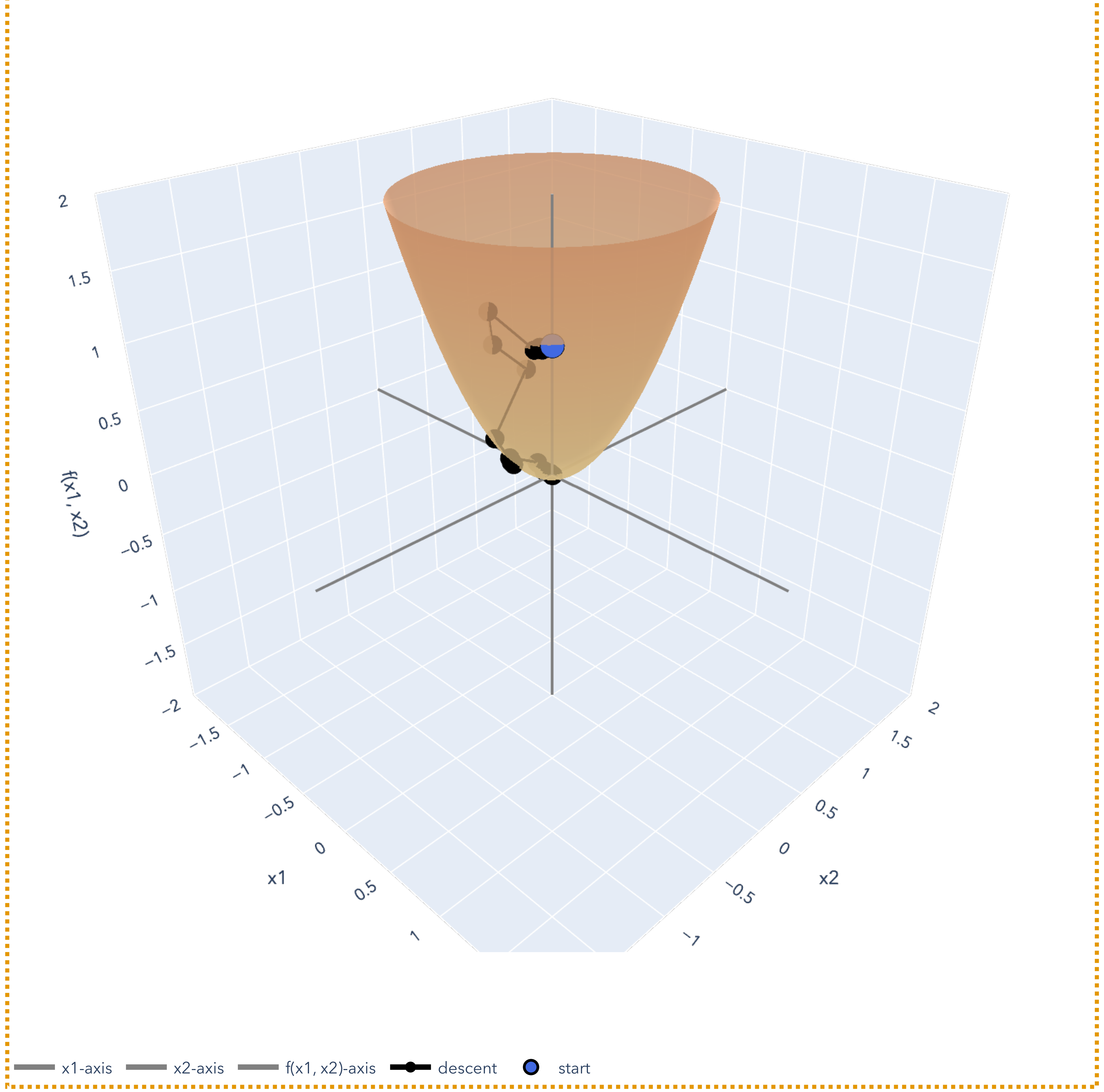
Lesson Overview

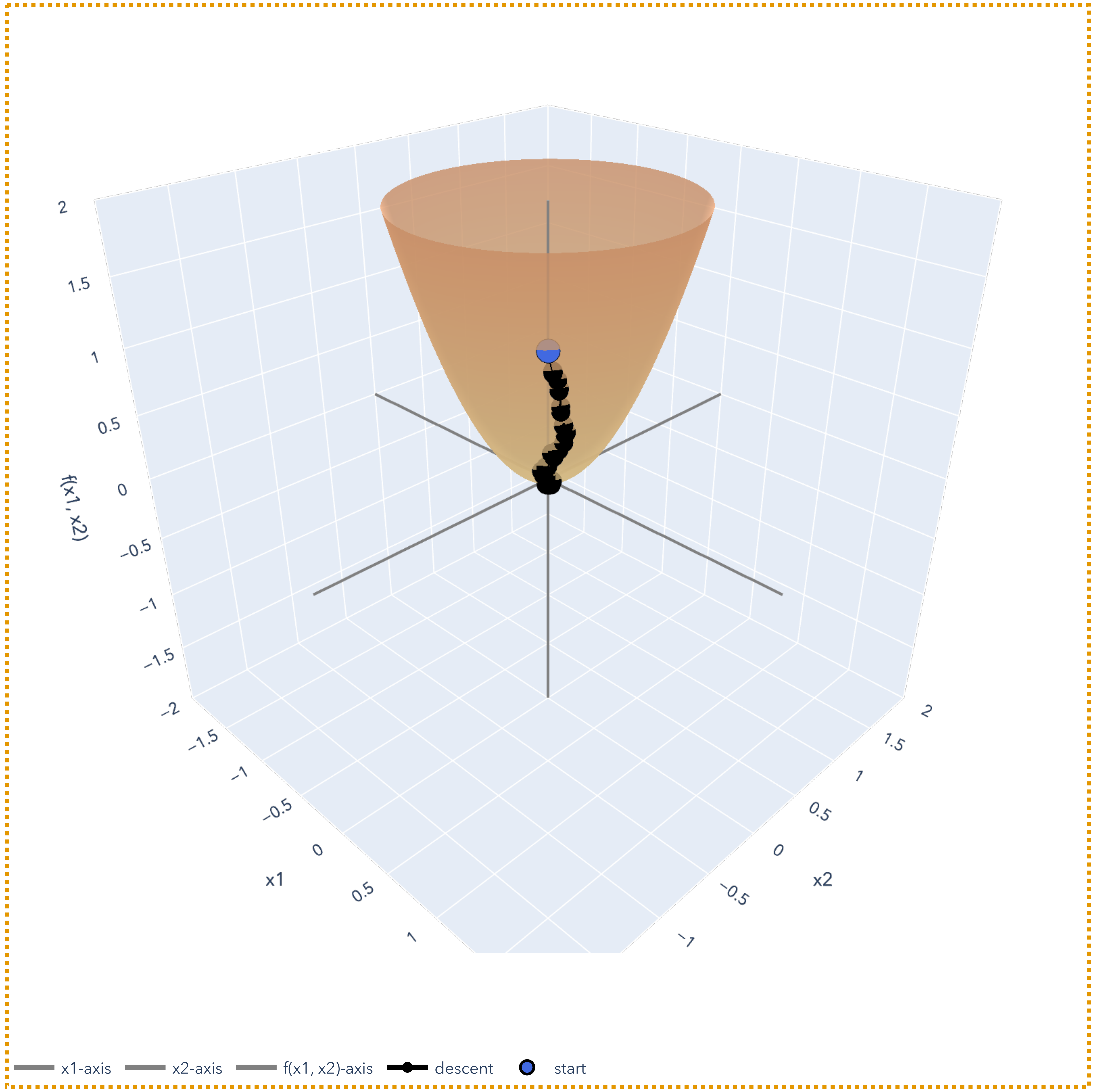
Big Picture: Least Squares



Lesson Overview

Big Picture: Gradient Descent





OLS under Gaussian Errors

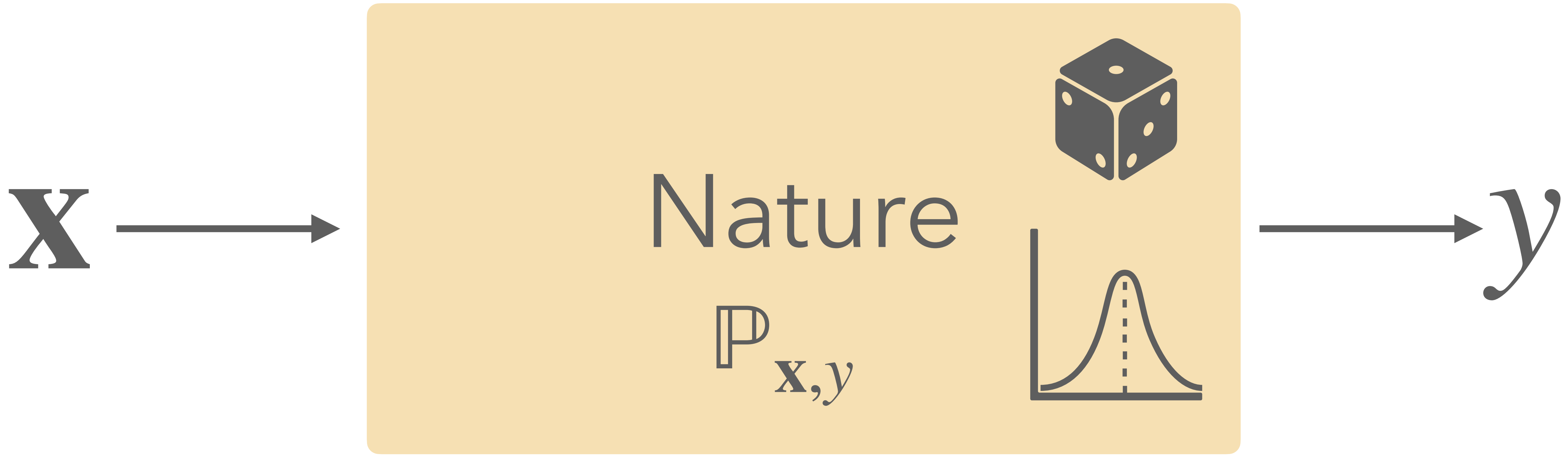
Intuition and Definition

Random error model

Adding Gaussian assumption on ϵ

$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i is independent of $\mathbf{x}_i$.

$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$, where $\epsilon \in \mathbb{R}^n$ is a Gaussian random vector with covariance matrix $\mathrm{Var}(\epsilon) = \sigma^2 \mathbf{I}$.



We can think of ϵ as the randomness from the "unexplained" errors in modeling the relationship of y to $\mathbf{x}$ with a linear model $\mathbf{w}^* \in \mathbb{R}^d$. Possibly very complex!

OLS and MLE

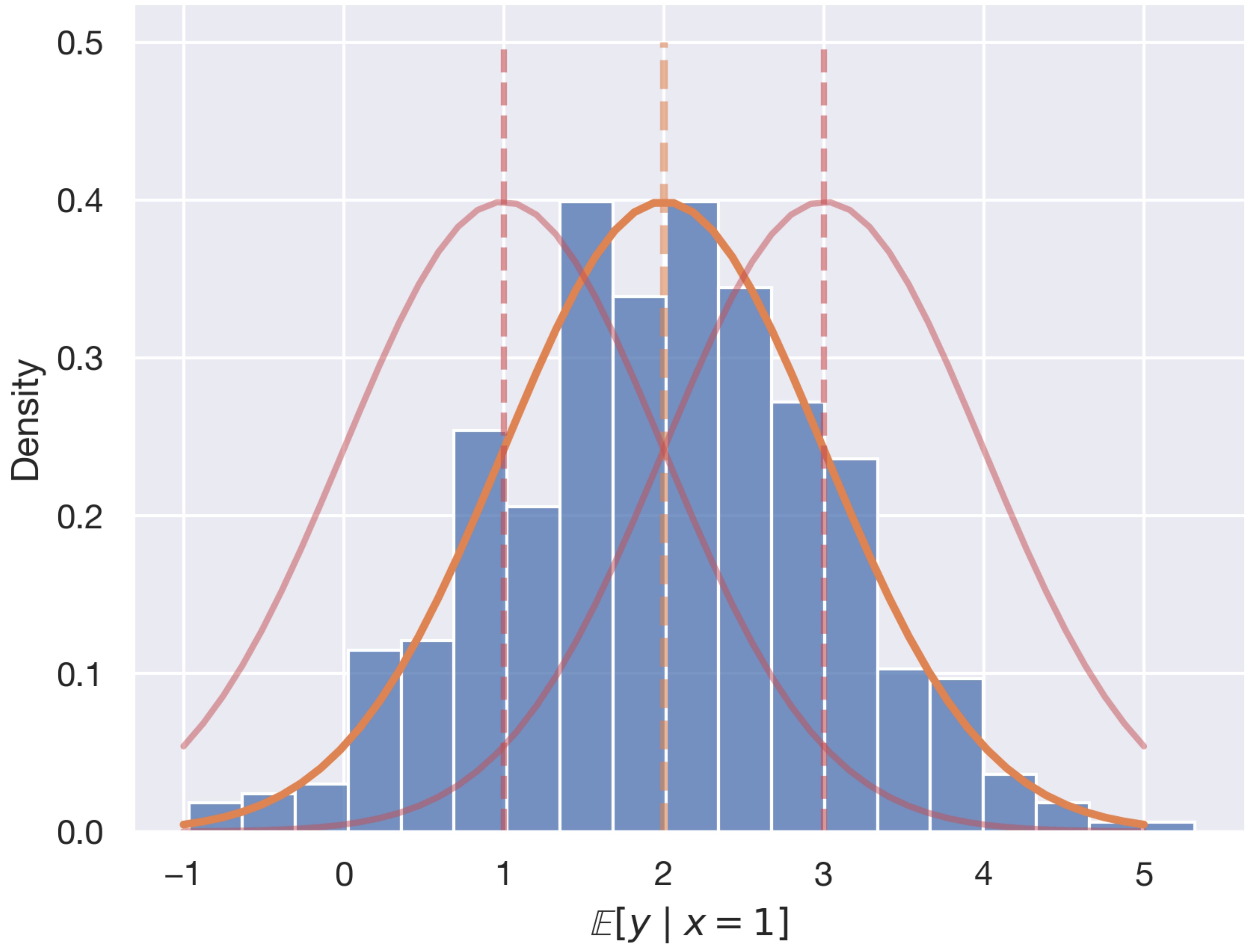
Theorem Statement

Theorem (OLS and MLE). Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d. samples in $\mathbb{R}^d \times \mathbb{R}$ with conditional distribution $\mathbb{P}_{y|\mathbf{x}}$ defined by:

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

where $\epsilon_i \sim N(0, \sigma^2)$ and each ϵ_i is independent. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ contain all the i.i.d. samples. Then, the maximum likelihood estimate (MLE) $\hat{\mathbf{w}}_{MLE}$ of the parameter $\mathbf{w}^*$ is given by the OLS estimator:

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



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. What if we assume Gaussian errors here?

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

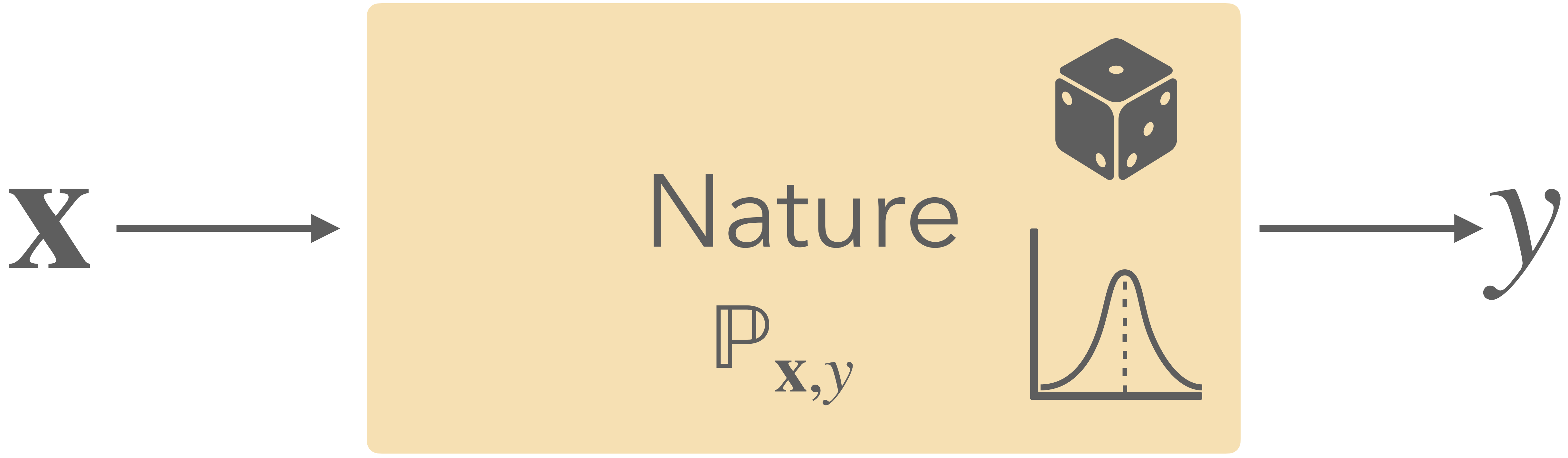
Risk (w.r.t. squared error): $R(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}}^\top \mathbf{x} - y)^2] = \sigma^2 + \sigma^2 \mathbb{E}[\mathrm{tr}(\boldsymbol{\Sigma}(\mathbf{X}^\top \mathbf{X})^{-1})] \approx \sigma^2 + \dfrac{\sigma^2 d}{n}$.

Random error model

Adding Gaussian assumption on ϵ

$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i is independent of $\mathbf{x}_i$.

$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$, where $\epsilon \in \mathbb{R}^n$ is a Gaussian random vector with covariance matrix $\mathrm{Var}(\epsilon) = \sigma^2 \mathbf{I}$.



We can think of ϵ as the randomness from the "unexplained" errors in modeling the relationship of y to $\mathbf{x}$ with a linear model $\mathbf{w}^* \in \mathbb{R}^d$. Possibly very complex!

Statistics of OLS

Under Gaussian Error Model

Question: What is the distribution of $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$?

In matrix-vector form, our Gaussian error model looks like:

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

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, and $\epsilon \in \mathbb{R}^n$ where $\epsilon_i \sim N(0, \sigma^2)$. Condition on $\mathbf{X}$. We can rewrite $\hat{\mathbf{w}}$ as:

$$\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}^* + \epsilon) \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon \\
&= \mathbf{w}^* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon
\end{aligned}$$

Statistics of OLS

Under Gaussian Error Model

Question: What is the distribution of $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$?

Therefore, $\hat{\mathbf{w}}$ *can be expressed as:*

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

With $\mathbf{X}$ fixed, this is a function of the random vector $\epsilon \in \mathbb{R}^n$.

We will show: *If* $\mathbf{x} \in \mathbb{R}^n$ *is a Gaussian random vector, then all affine transformations* $\mathbf{A}\mathbf{x} + \mathbf{b}$ *(where* $\mathbf{A} \in \mathbb{R}^{d \times n}$ *and* $\mathbf{b} \in \mathbb{R}^d$*) of* $\mathbf{x}$ *are also Gaussian random vectors.*

Therefore: $\hat{\mathbf{w}} \sim N(\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}], \mathrm{Var}(\hat{\mathbf{w}} \mid \mathbf{X})).$

Statistics of OLS

Under Gaussian Error Model

Question: What is the distribution of $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$?

So $\hat{\mathbf{w}}$ *is multivariate Gaussian:* $\hat{\mathbf{w}} \sim N(\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}], \mathrm{Var}(\hat{\mathbf{w}} \mid \mathbf{X}))$.

What's $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}]$? Because $\mathbb{E}[\epsilon \mid \mathbf{X}] = \mathbf{0}$ and $\mathbf{w}*$ is fixed, $\mathbb{E}[\hat{\mathbf{w}} \mid \mathbf{X}] = \mathbf{w}*$.

What's $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}]$, the covariance matrix? Already showed: $\mathrm{Var}[\hat{\mathbf{w}} \mid \mathbf{X}] = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2$.

Therefore, $\hat{\mathbf{w}} \sim N(\mathbf{w}*, (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2)$.

Statistics of OLS

Theorem Statement

Theorem (Statistical properties of OLS under Gaussian errors). 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 $\mathbf{x}$, with each $\epsilon \sim N(0,\sigma^2)$.

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 iid 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 a multivariate Gaussian distribution:

$$\hat{\mathbf{w}} \sim N(\mathbf{w}*, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$

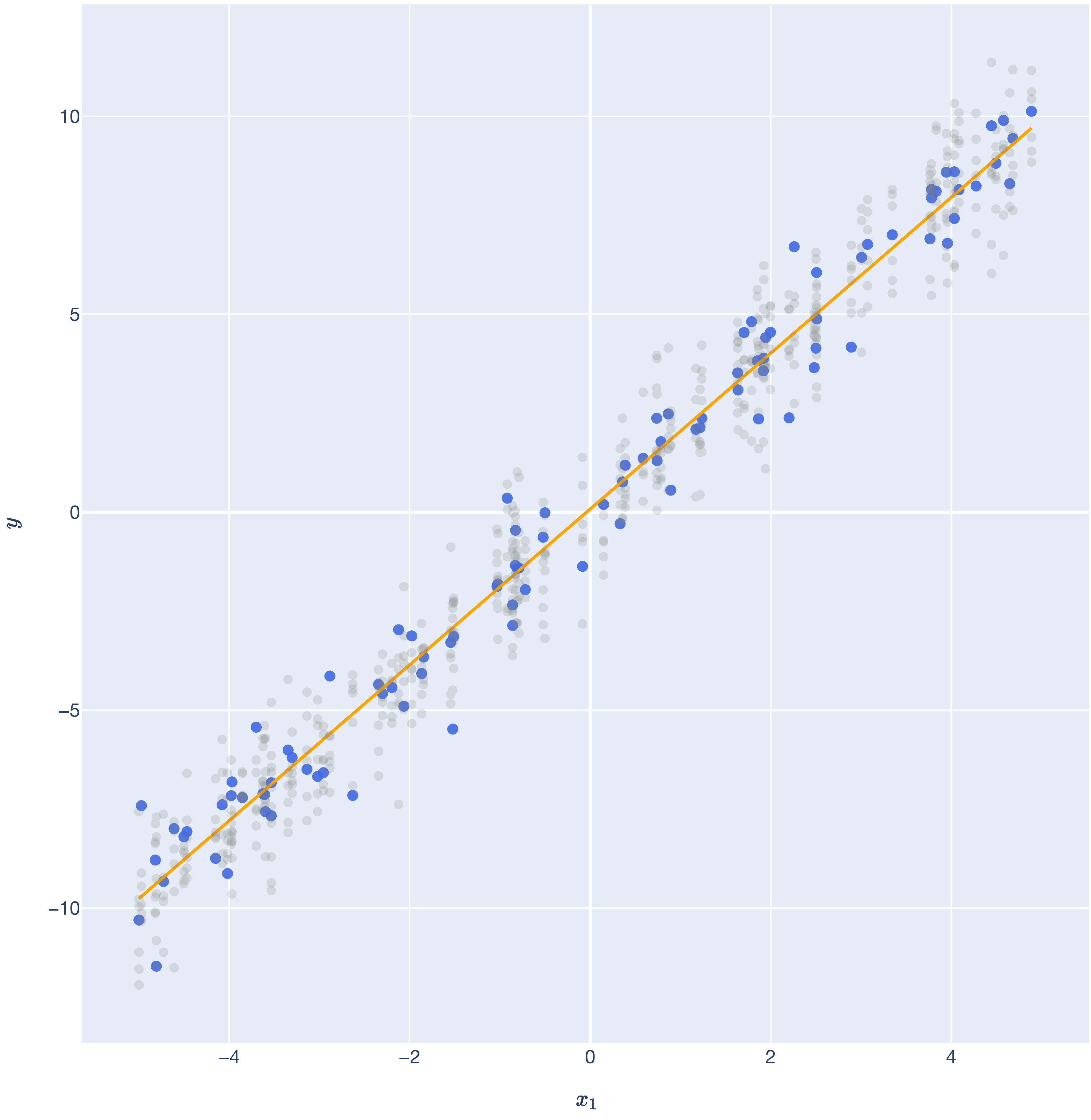
Statistics of OLS

Theorem Statement

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

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

$$\hat{\mathbf{w}} \sim N(\mathbf{w}^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$



Statistics of OLS

Theorem Statement

Theorem (Statistical properties of OLS under Gaussian errors). 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 $\mathbf{x}$, with each $\epsilon \sim N(0,\sigma^2)$.

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 iid 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 a multivariate Gaussian distribution:

$$\hat{\mathbf{w}} \sim N(\mathbf{w}^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

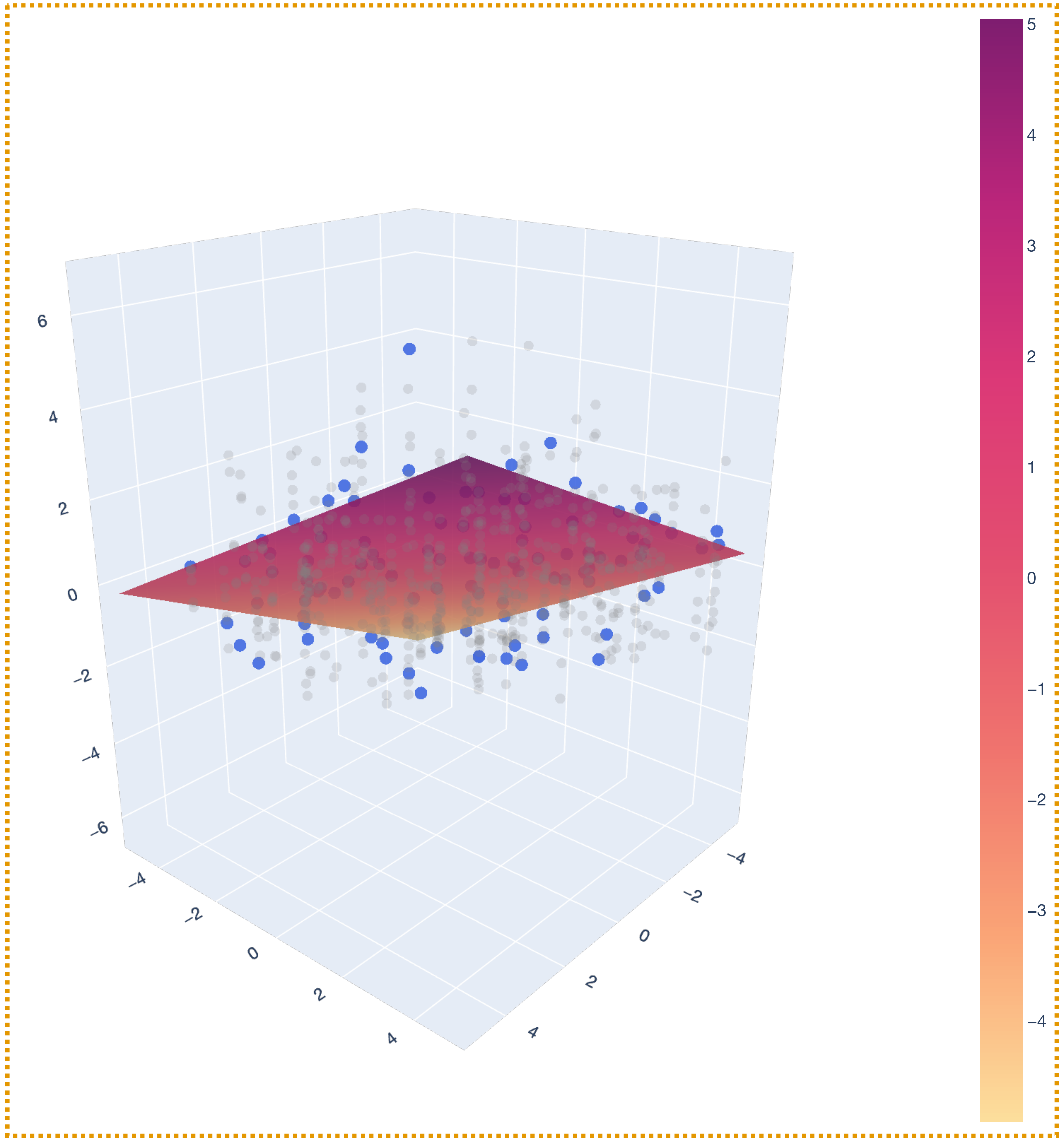
Statistics of OLS

Theorem Statement

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

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

$$\hat{\mathbf{w}} \sim N(\mathbf{w}^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$



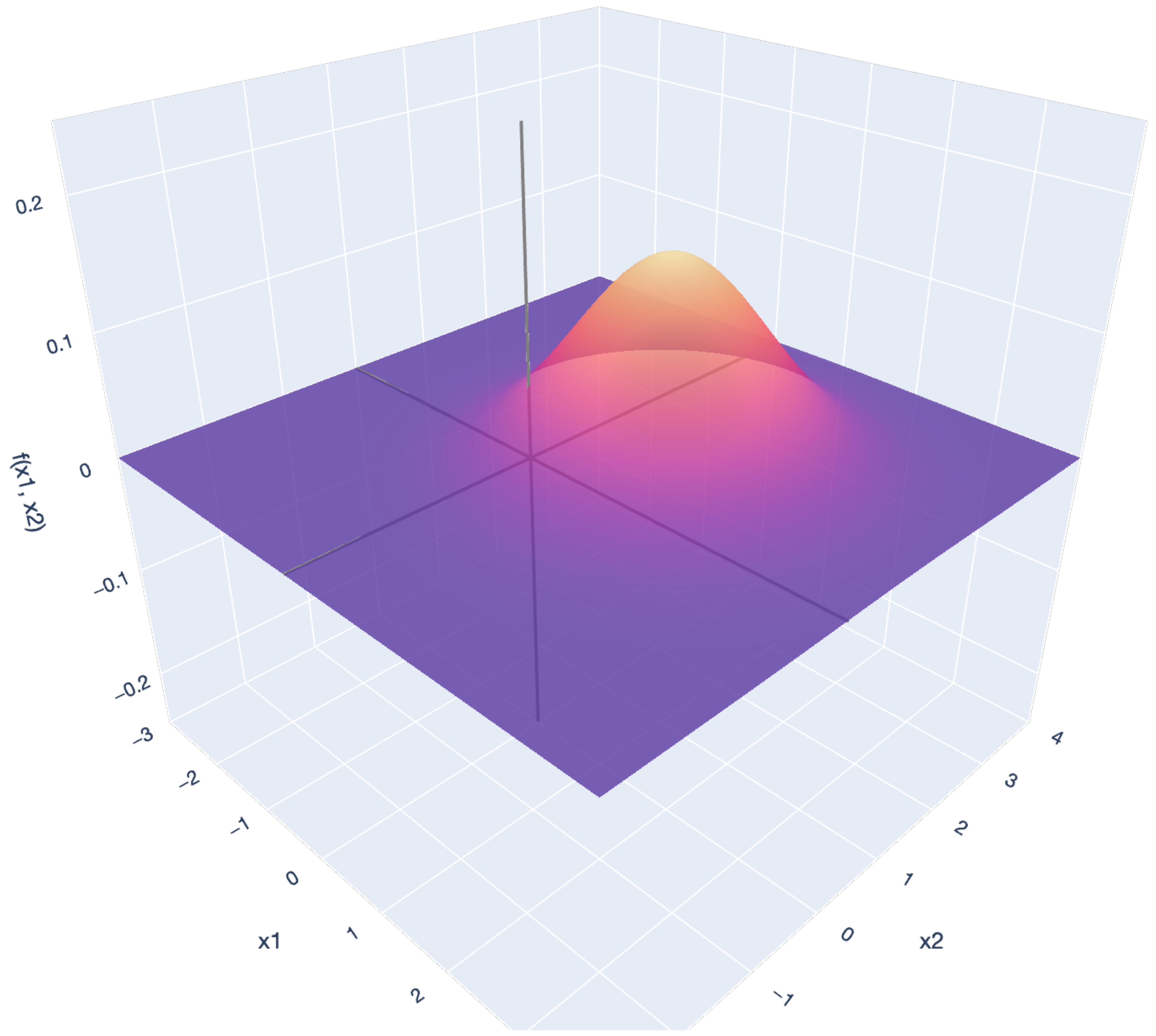
Statistics of OLS

Theorem Statement

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

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

$$\hat{\mathbf{w}} \sim N(\mathbf{w}^*, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$



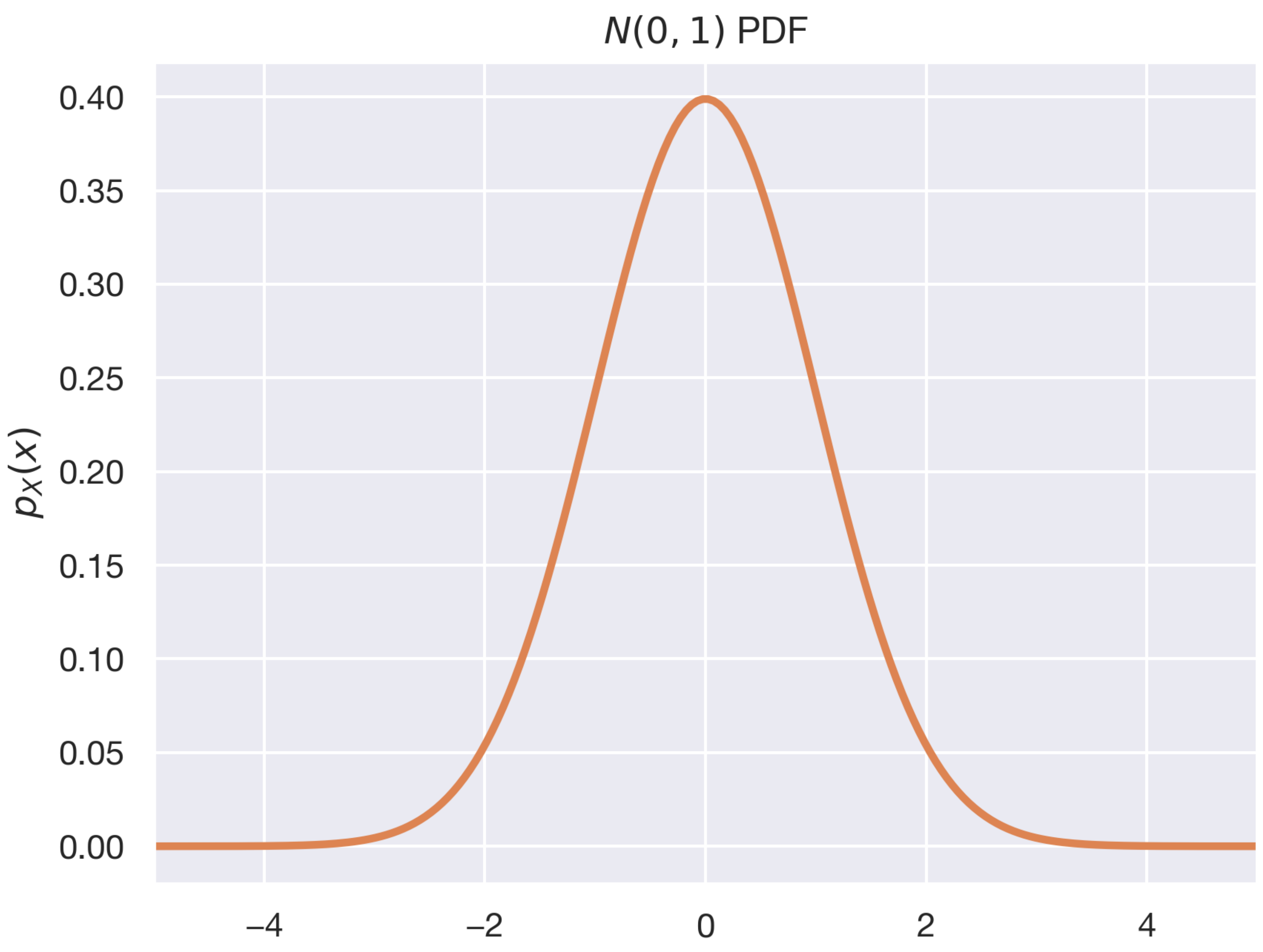
Single-variable Gaussian

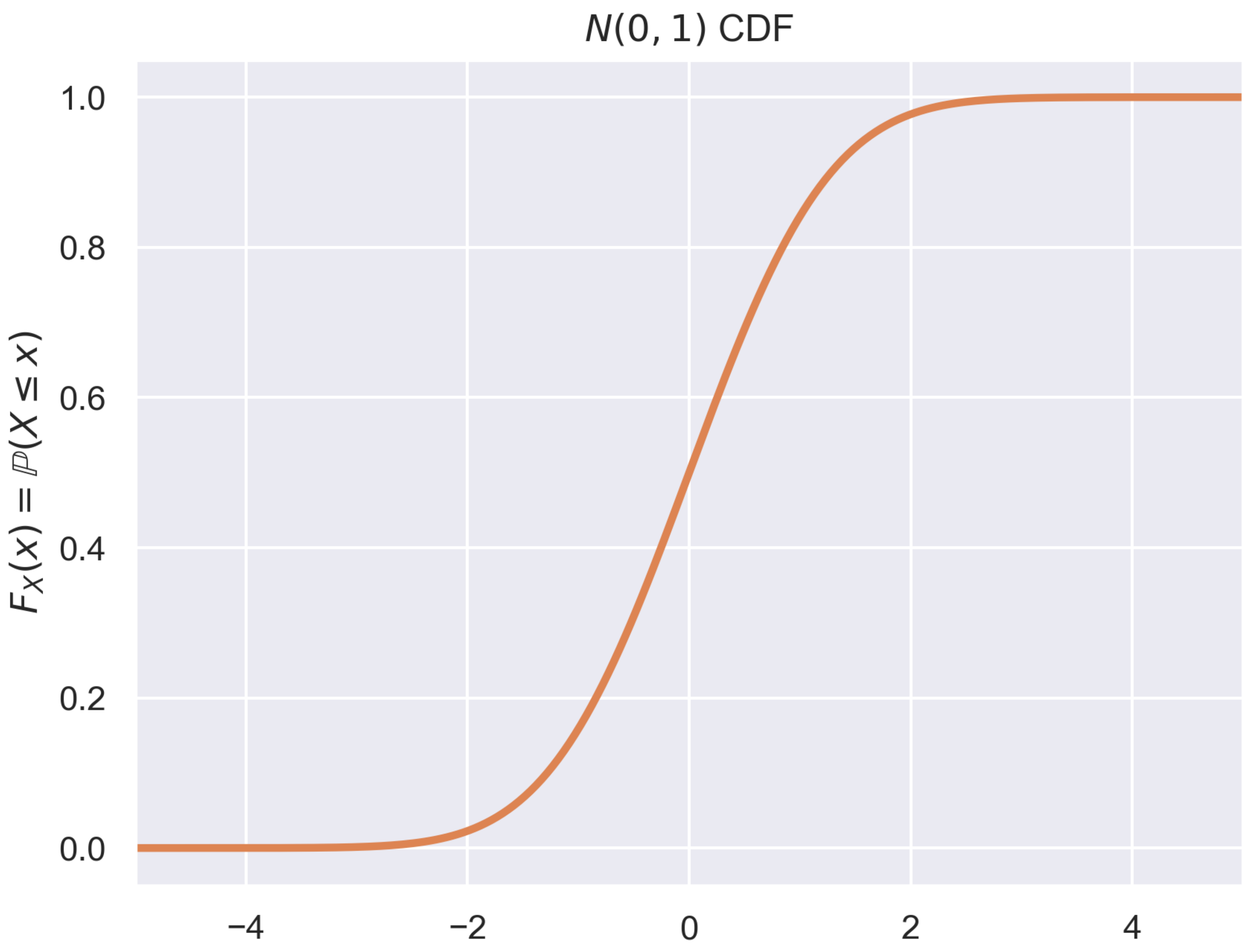
Review and Intuition

The Gaussian Distribution

Intuition and Shape

The Gaussian/Normal distribution with parameters μ and σ has a "bell-shaped" PDF centered at μ and "spread" depending on the parameter σ.





The Gaussian Distribution

Standard Gaussian Definition

A RV Z has a standard Gaussian/Normal distribution denoted $Z \sim N(0,1)$ if it has PDF:

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \text{ for all } z \in \mathbb{R}.$$

This random variable has mean $\mathbb{E}[Z] = 0$ and variance $\text{Var}(Z) = 1$.

Traditionally, standard Gaussians are denoted with Z, PDF $\phi(z)$, and CDF $\Phi(z)$.

The Gaussian Distribution

General Definition

A random variable X has a Gaussian/Normal distribution with parameters μ and σ, denoted $X \sim N(\mu, \sigma^2)$ if it has PDF:

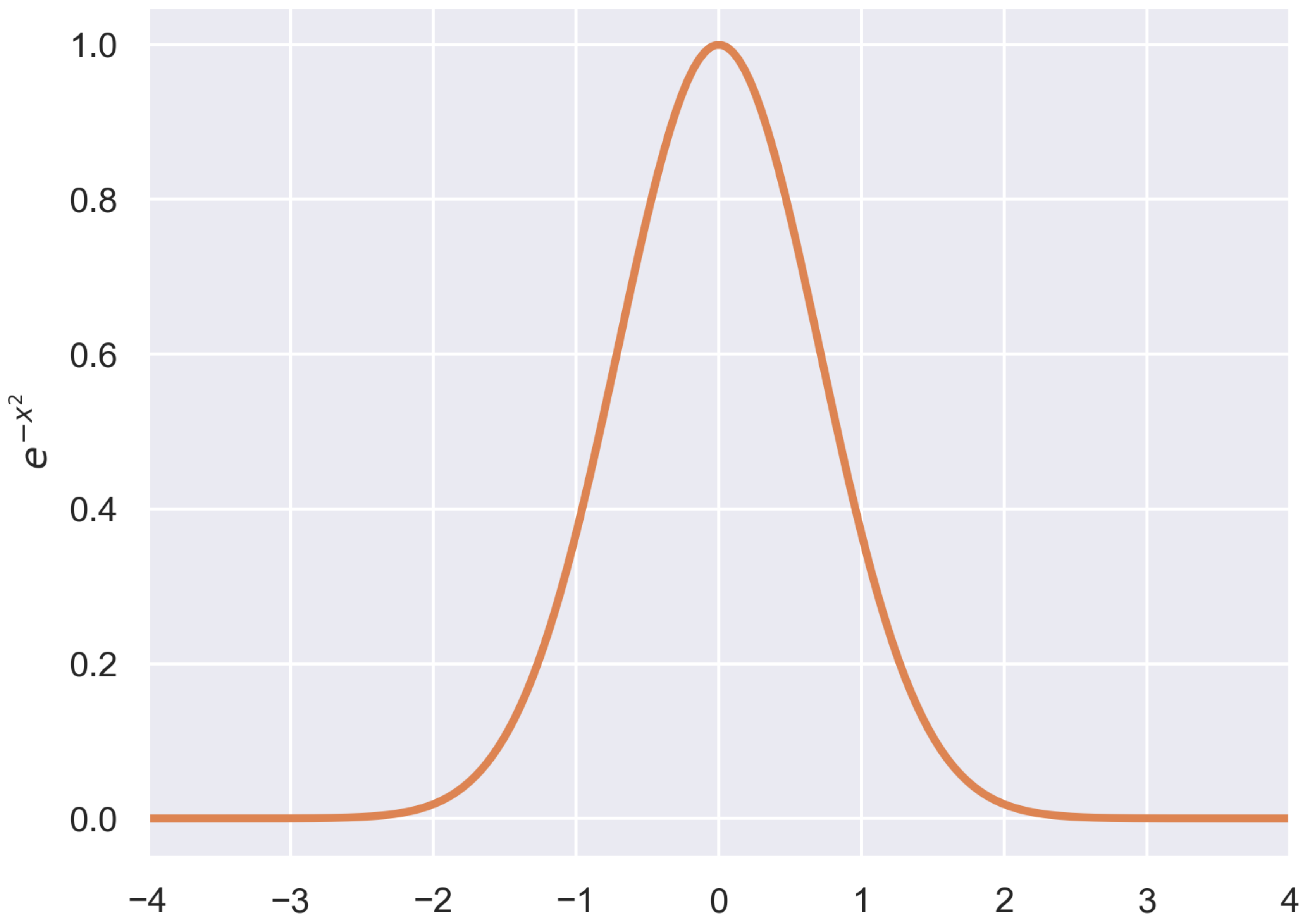
$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \text{ for all } x \in \mathbb{R}.$$

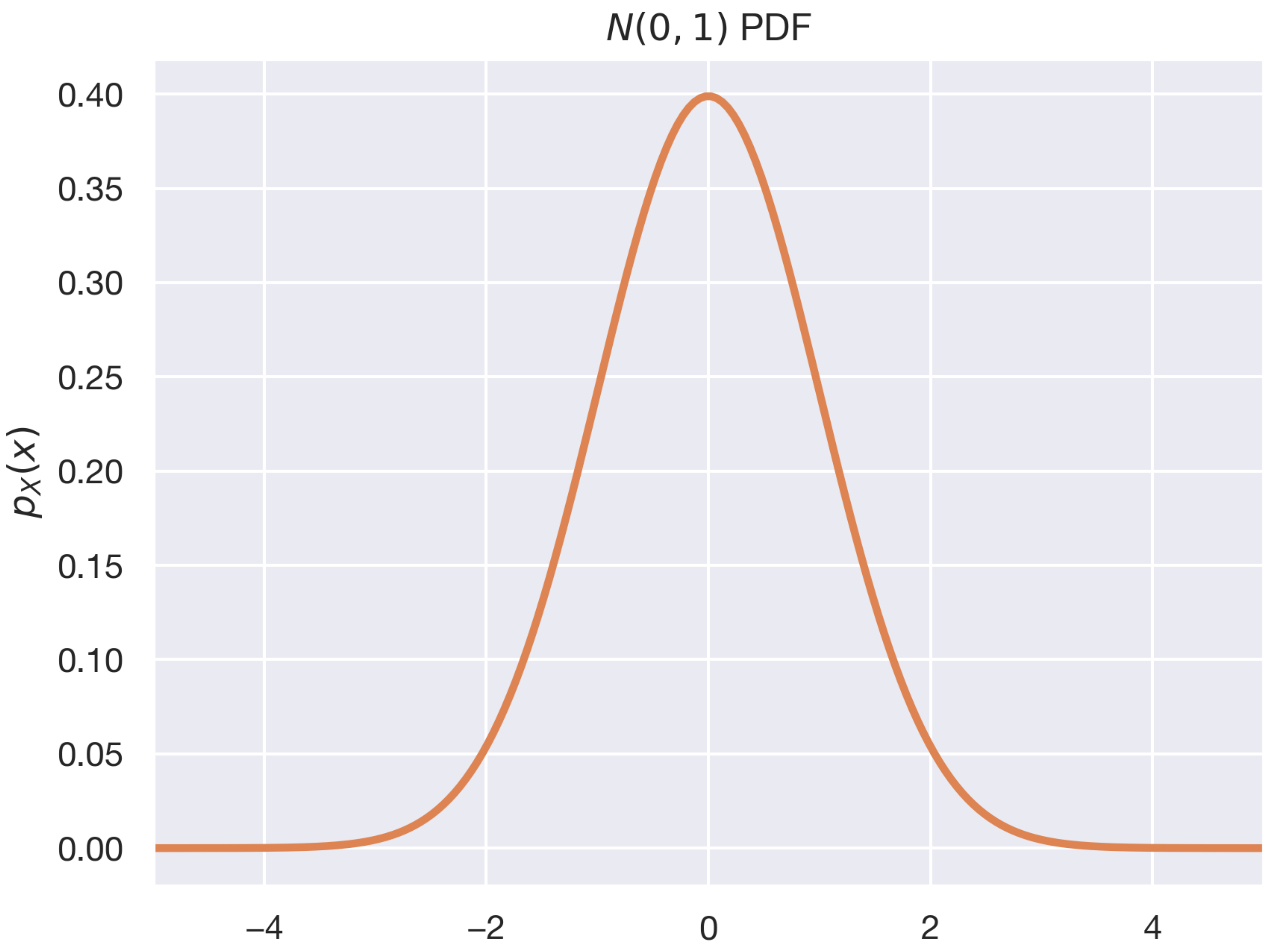
This random variable has mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$.

PDF of the Gaussian

Intuition

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$





PDF of the Gaussian

Intuition

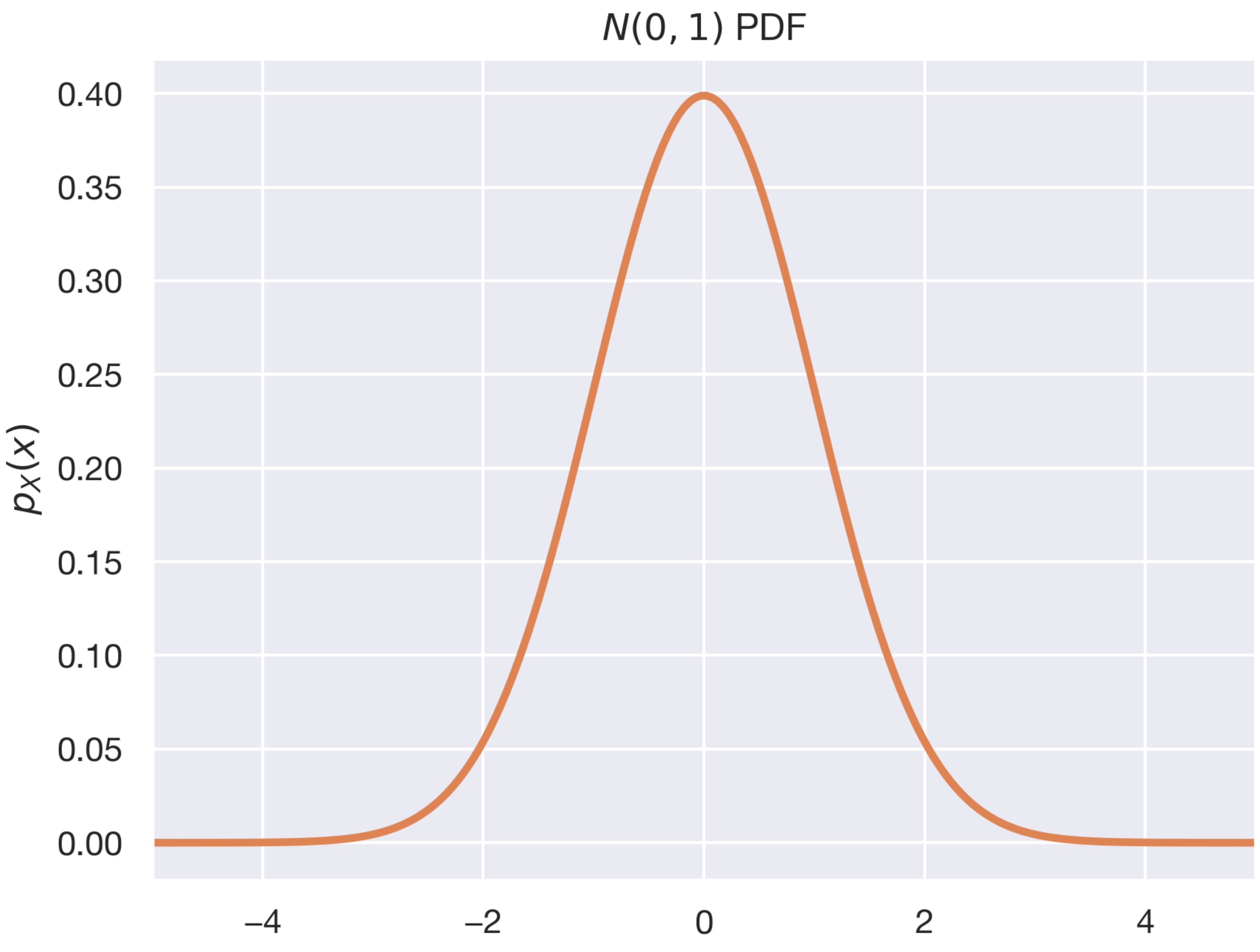
$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

The argument of $\exp(\,\cdot\,)$ is a *quadratic function*:

$$-\frac{1}{2\sigma^2}(x-\mu)^2.$$

The coefficient doesn't depend on x; it's a *normalizing constant*:

$$\frac{1}{\sigma\sqrt{2\pi}}.$$



Multivariate Gaussian

Intuition and Definition

Single-variable to Multivariable

Comparison

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \qquad p(\mathbf{x}) = \frac{1}{\det(\Sigma)^{1/2}(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu) \right\}$$

$-\frac{1}{2\sigma^2}(x-\mu)^2$ is a *quadratic function.*

$\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)$ is a *quadratic form.*

$\frac{1}{\sigma\sqrt{2\pi}}$ is a *normalizing constant.*

$\frac{1}{\det(\Sigma)^{1/2}(2\pi)^{n/2}}$ is a *normalizing constant.*

Single-variable to Multivariable

Comparison

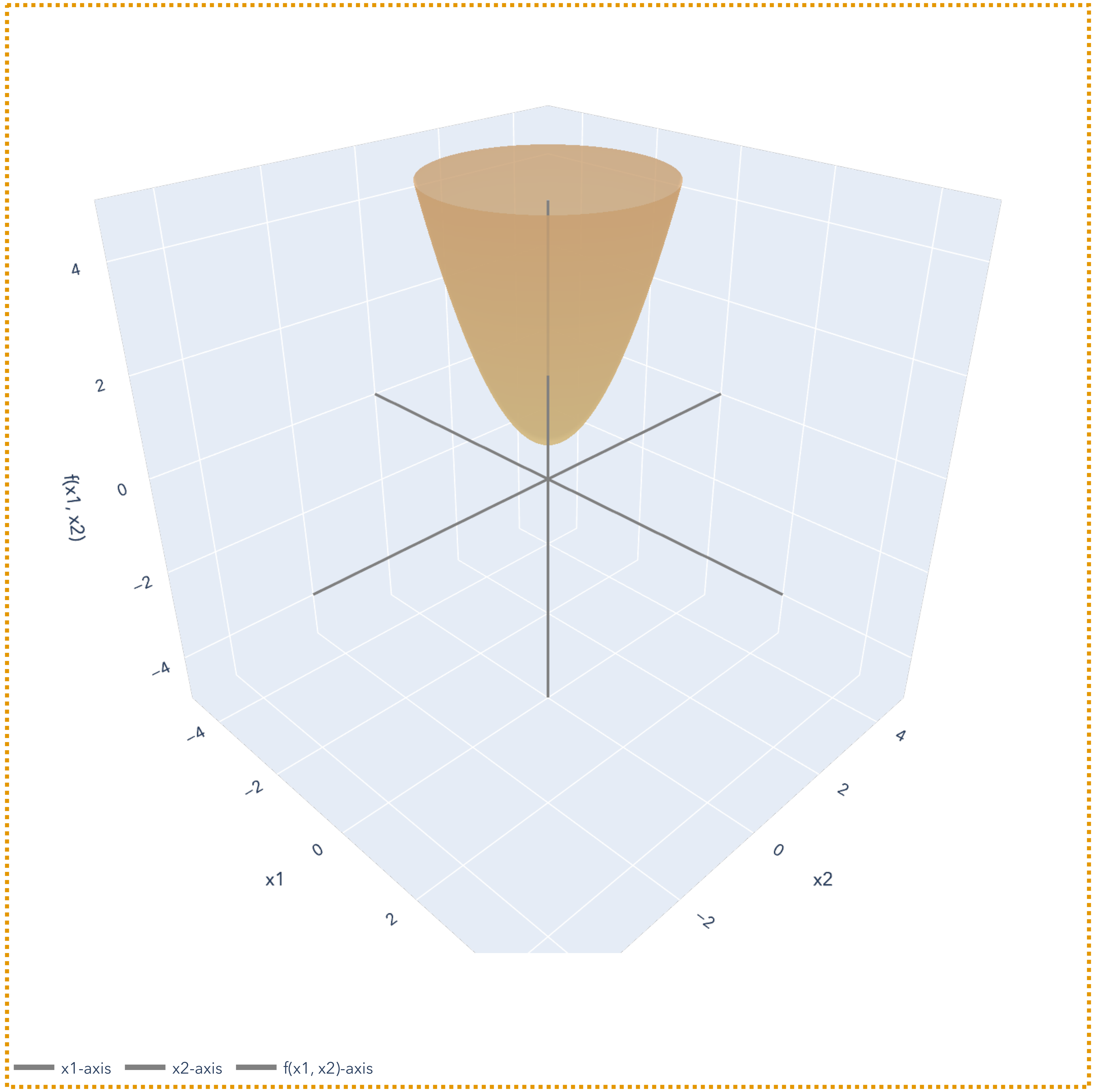
$$p(\mathbf{x}) = \frac{1}{\det(\Sigma)^{1/2}(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

$\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)$ is a *quadratic form.*

Σ is positive definite, Σ^{-1} is also positive definite.

Therefore, $(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu) > 0.$

Therefore, $\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu) < 0.$



Multivariate Gaussian

Definition

A random vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ has the multivariate Gaussian/Normal distribution, denoted $\mathbf{x} \sim N(\mu, \Sigma)$ if it has the density:

$$p(\mathbf{x}) = \frac{1}{\det(\Sigma)^{1/2}(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

where $\det(\Sigma)$ is the *determinant* of $\Sigma \in \mathbb{R}^{d\times d}$, a *positive definite* matrix covariance matrix, and $\mu \in \mathbb{R}^d$ is the mean $\mathbb{E}[\mathbf{x}]$.

Standard Multivariate Gaussian

Definition

A random vector $\mathbf{x} = (z_1, \ldots, z_d) \in \mathbb{R}^d$ has the standard multivariate Gaussian/Normal distribution, denoted $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ if it has the density:

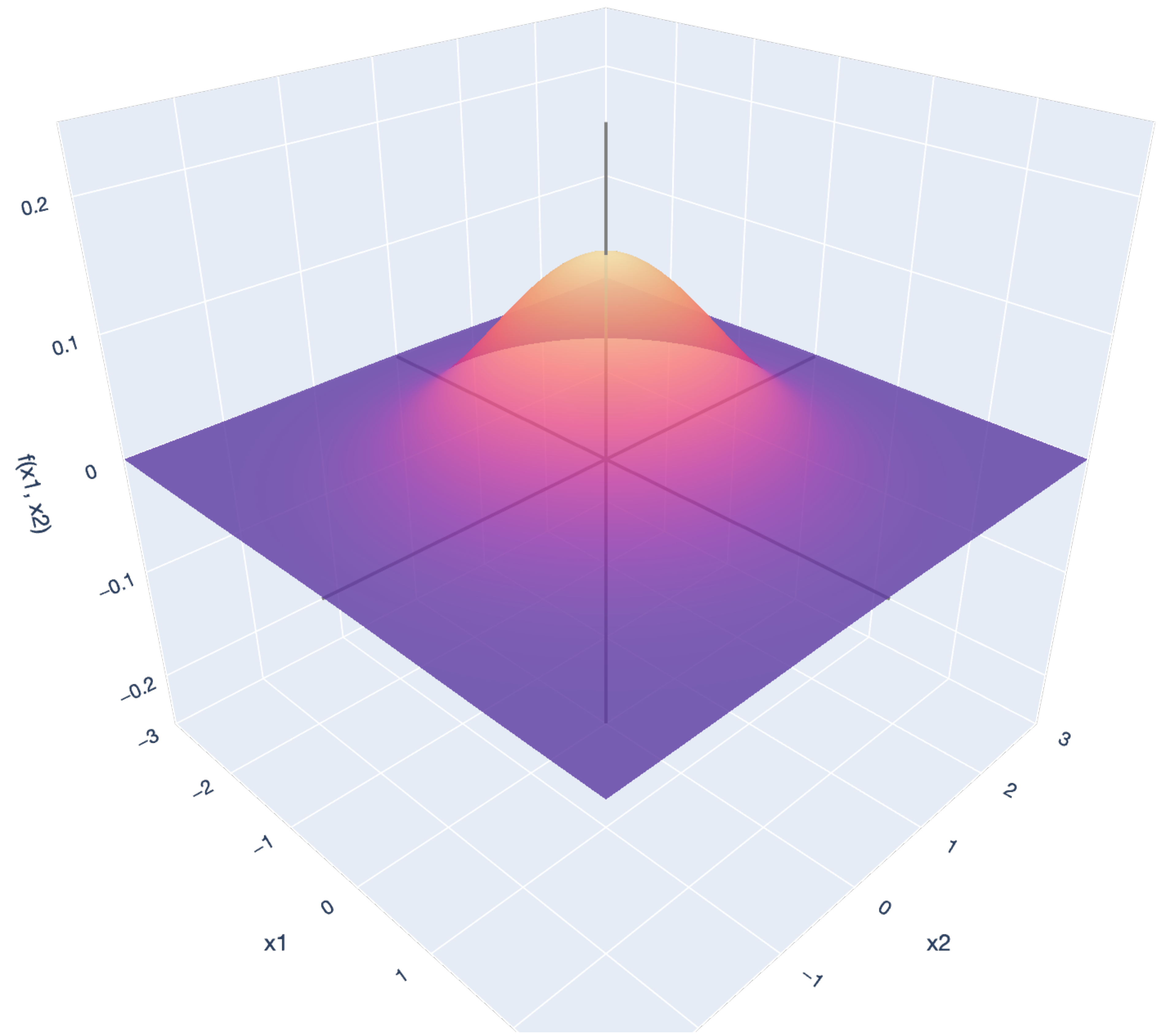
$$p(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right\}.$$

Standard Multivariate Gaussian

Definition

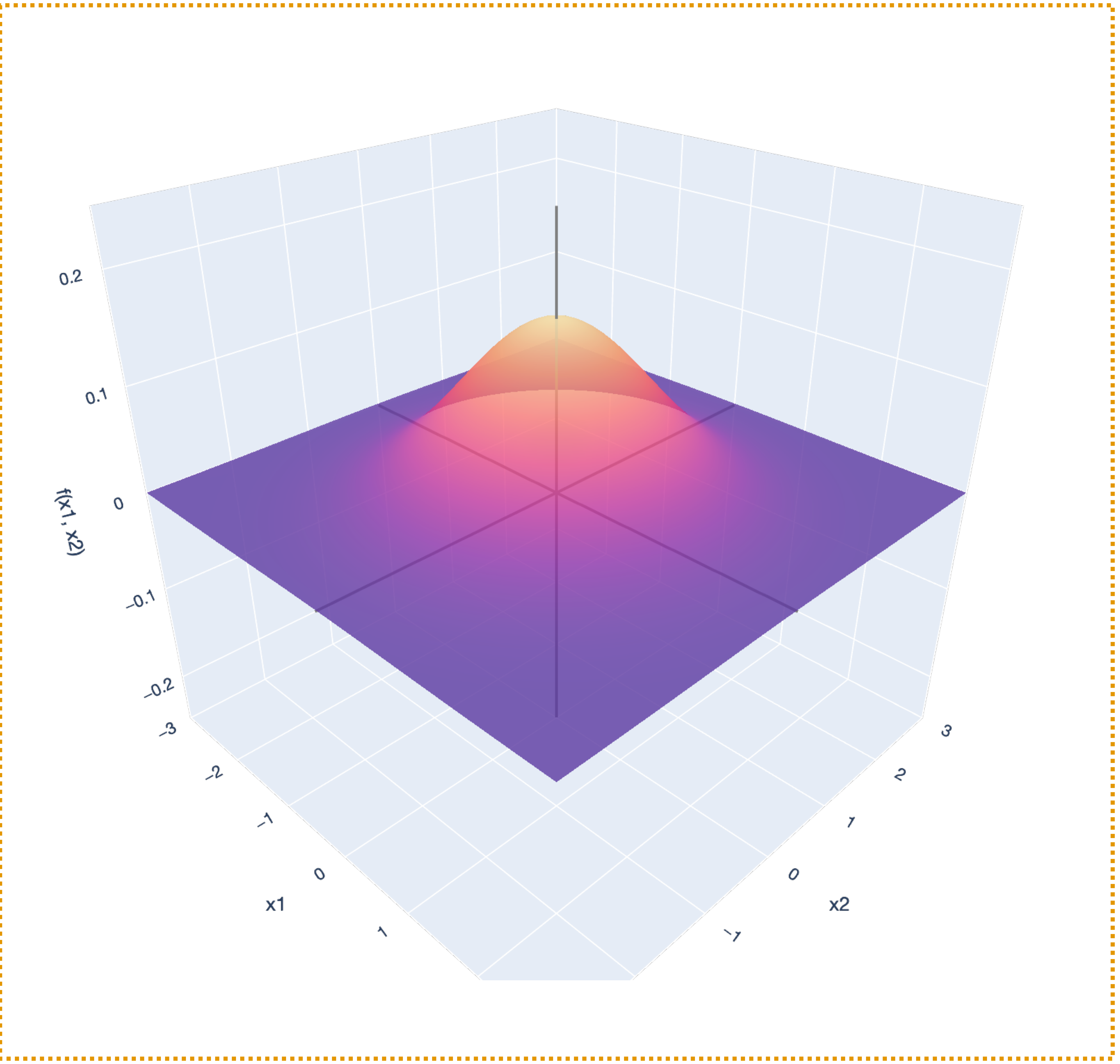
A random vector $\mathbf{x} = (z_1, \ldots, z_d) \in \mathbb{R}^d$ has the standard multivariate Gaussian/Normal distribution, denoted $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ if it has the density:

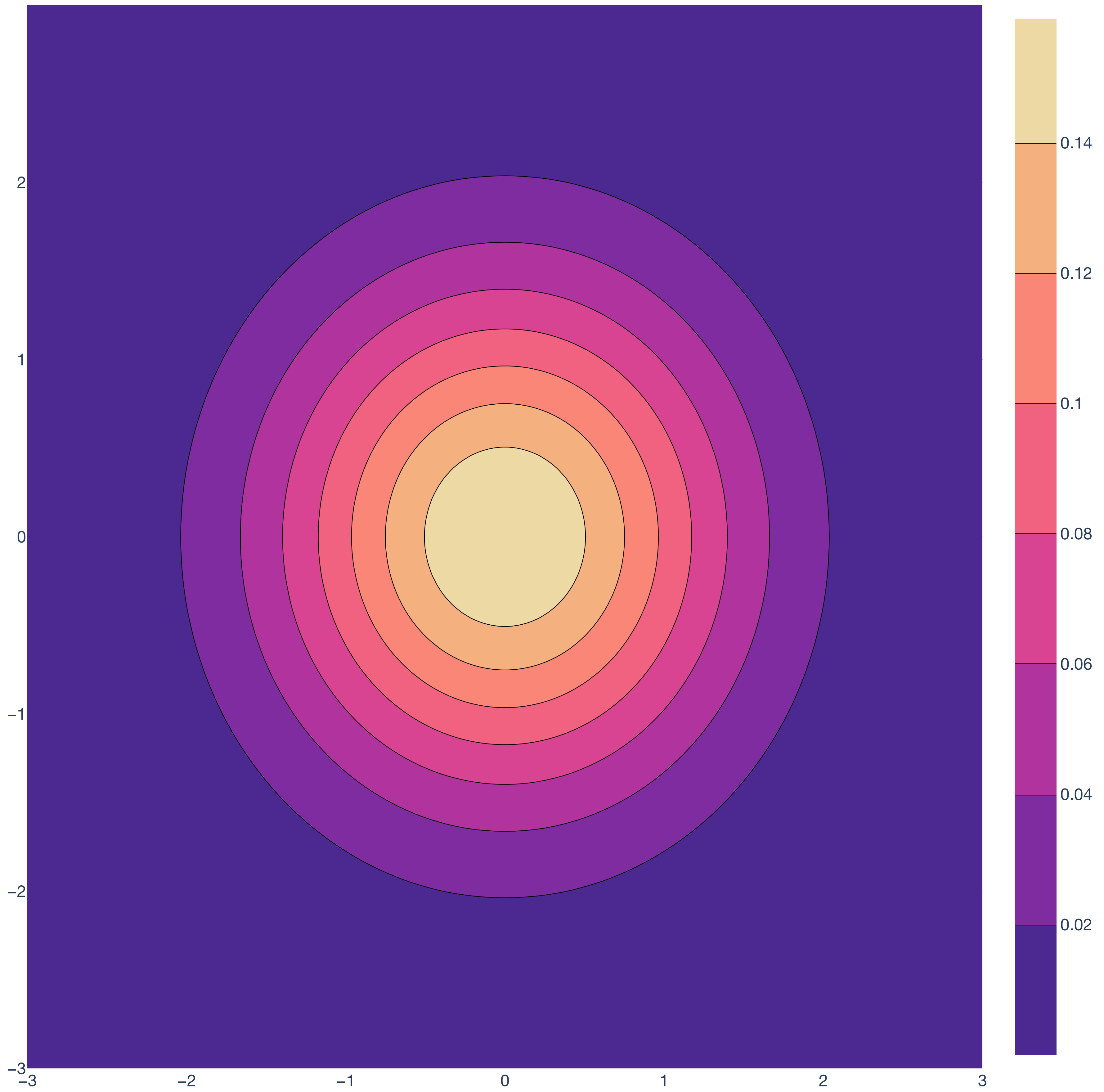
$$p(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}\right\}.$$



Multivariate Gaussian

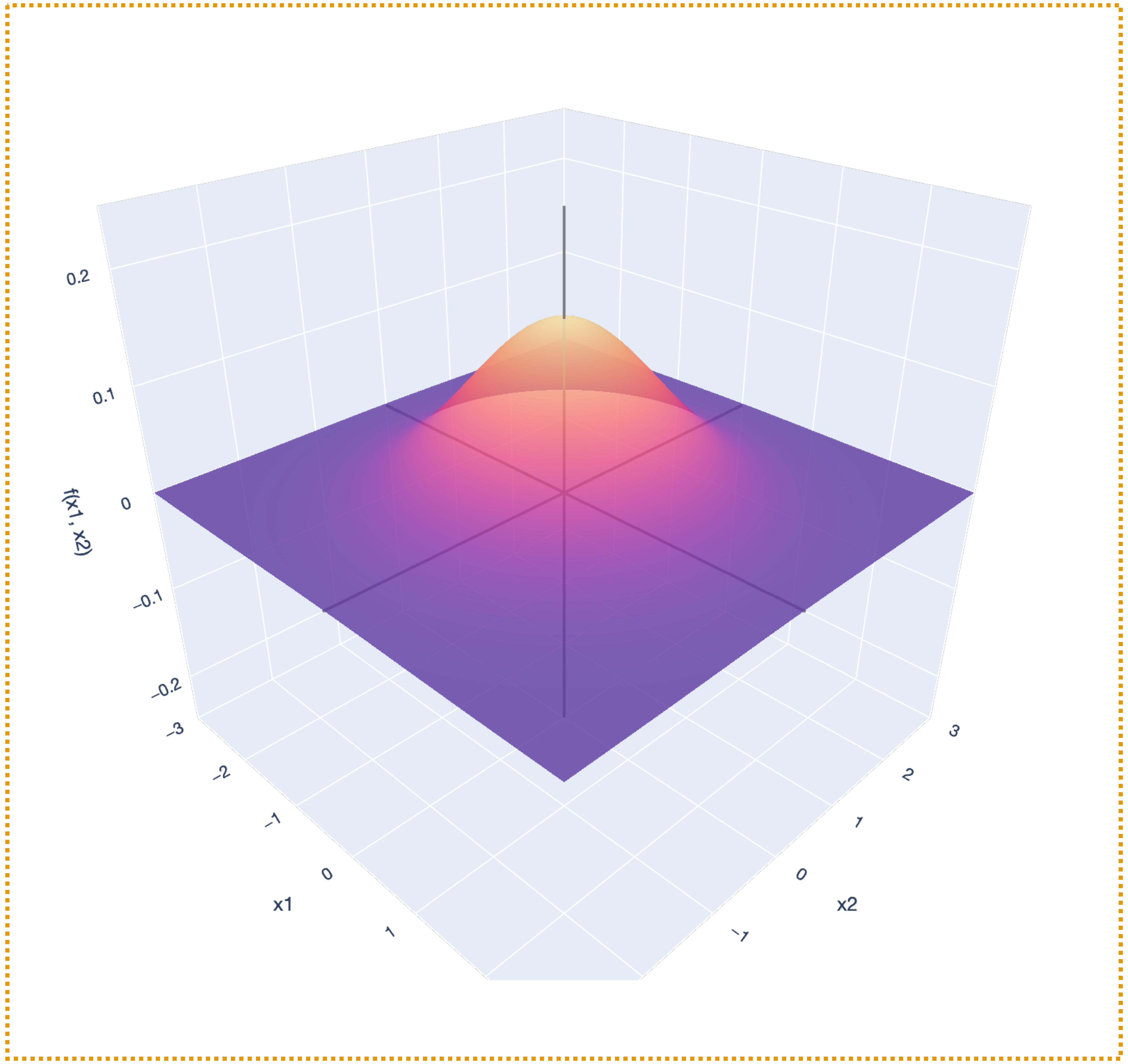
Example: $N(\mathbf{0}, \mathbf{I})$

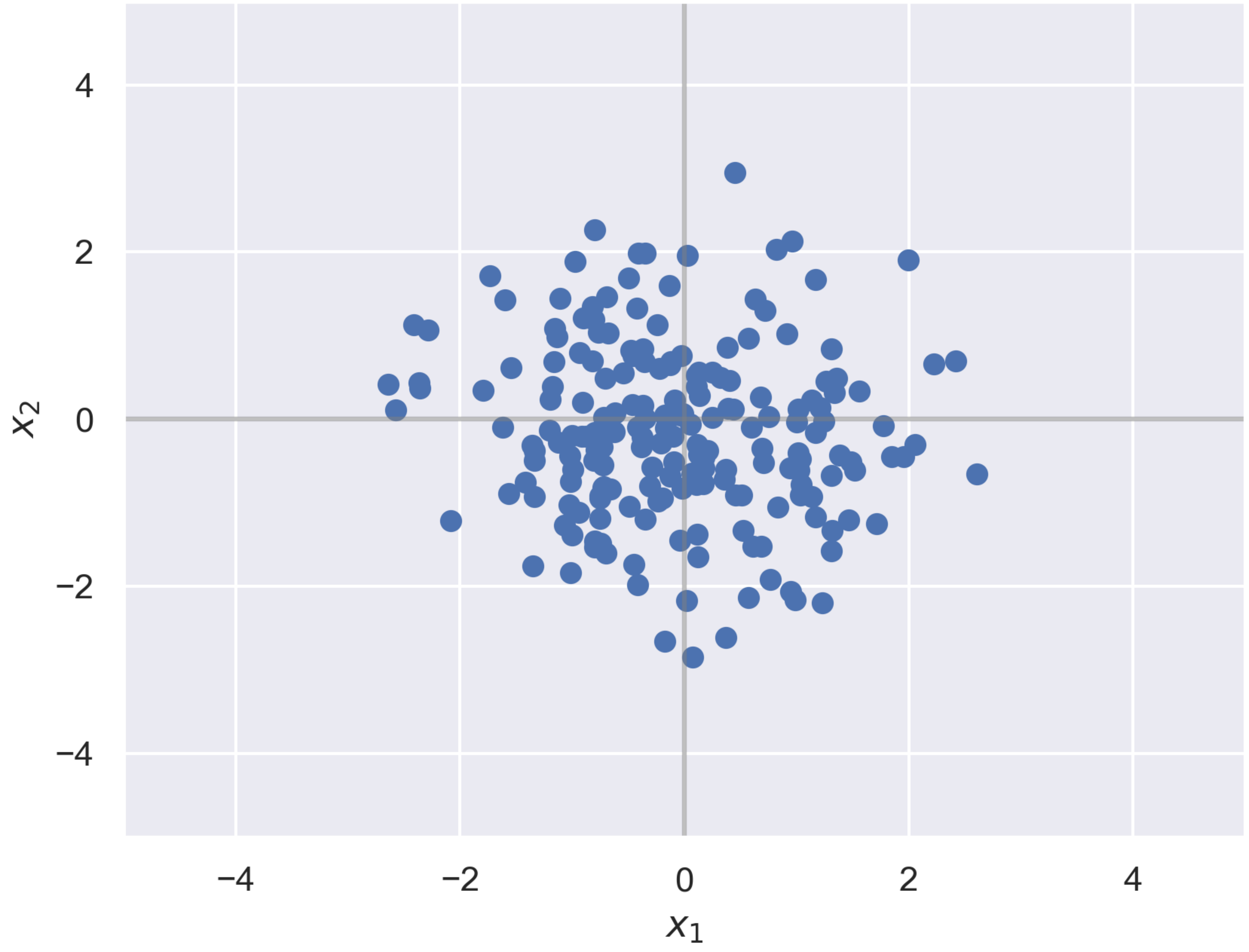




Multivariate Gaussian

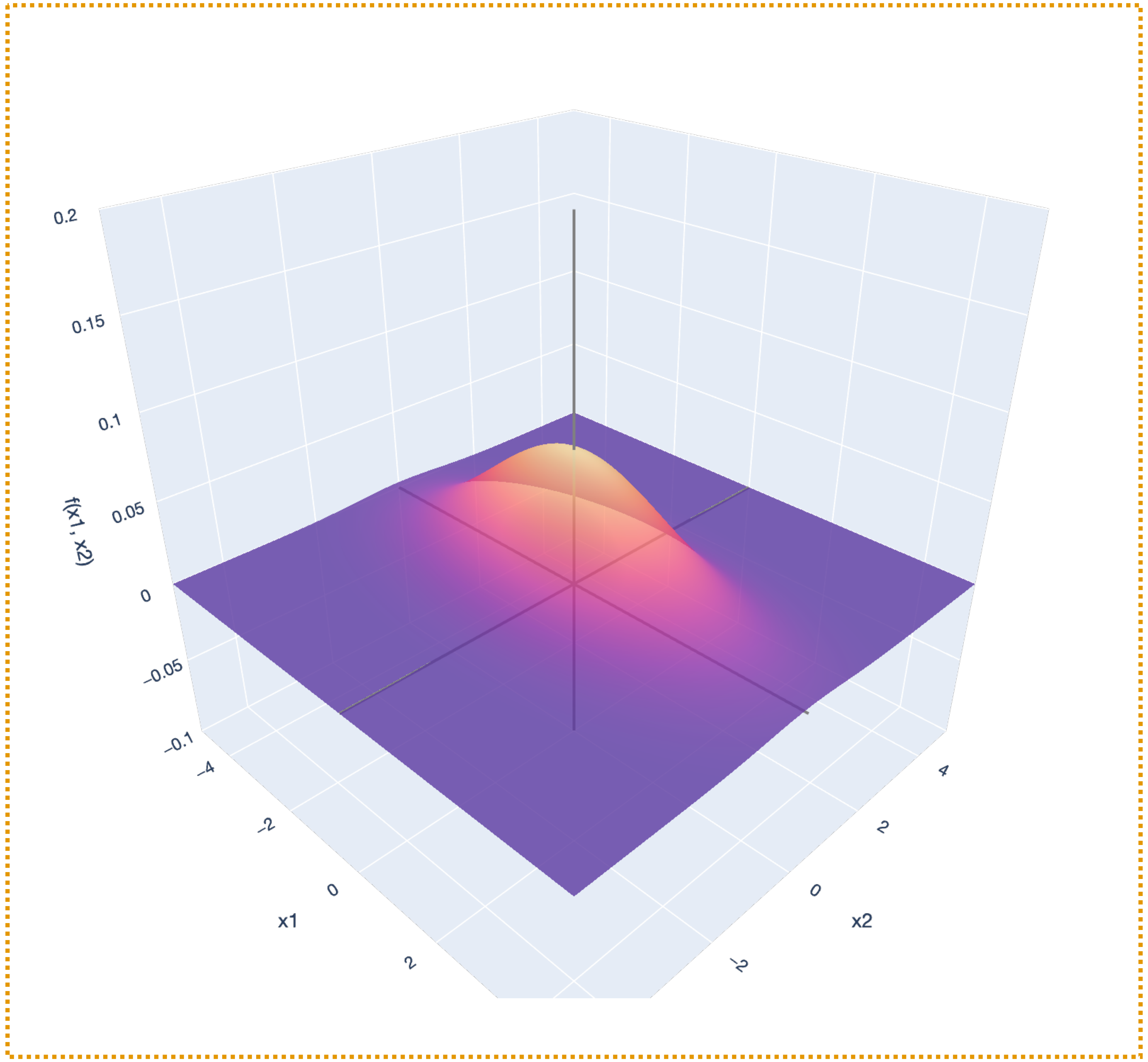
Example: $N(\mathbf{0}, \mathbf{I})$

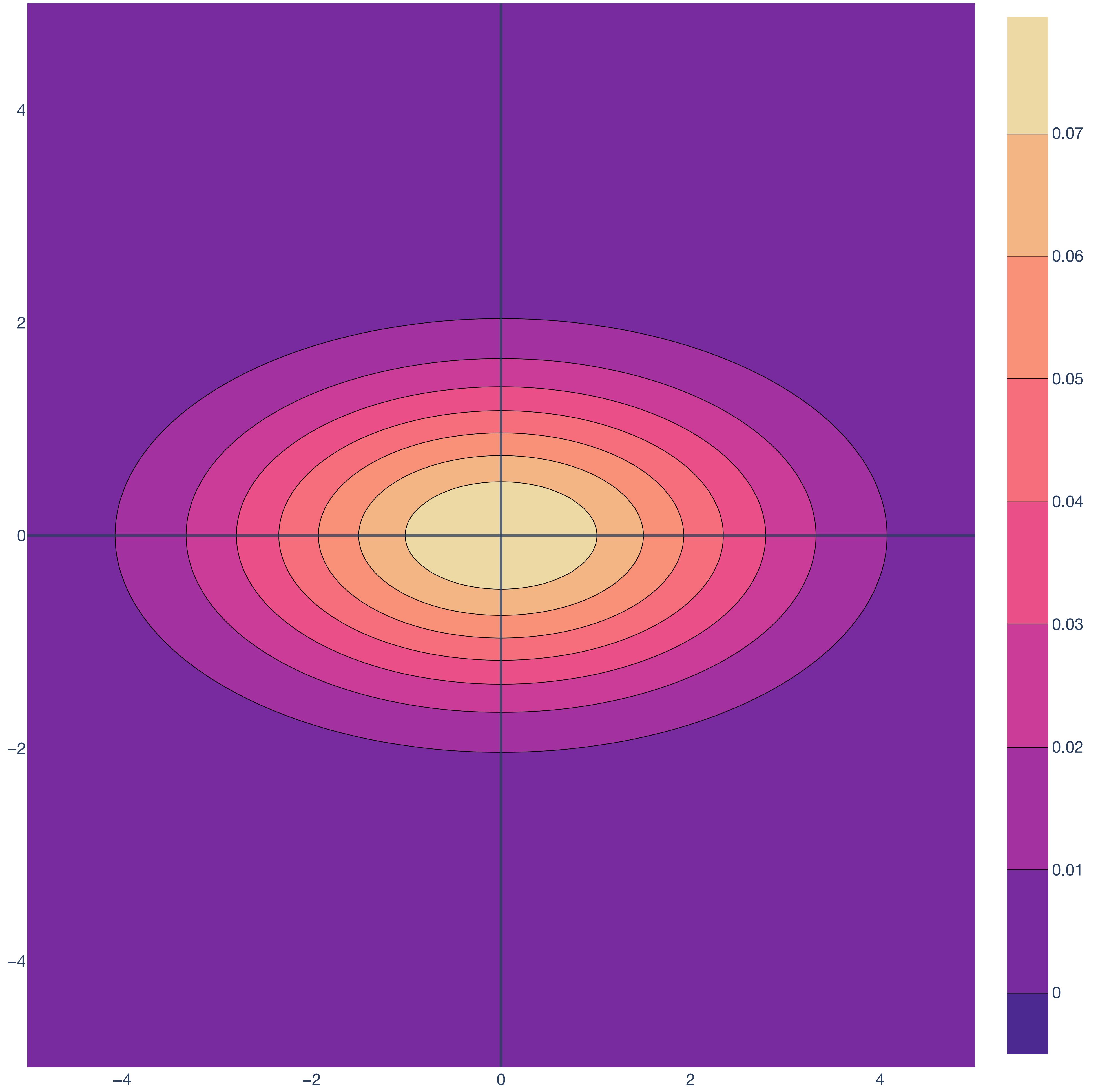




Multivariate Gaussian

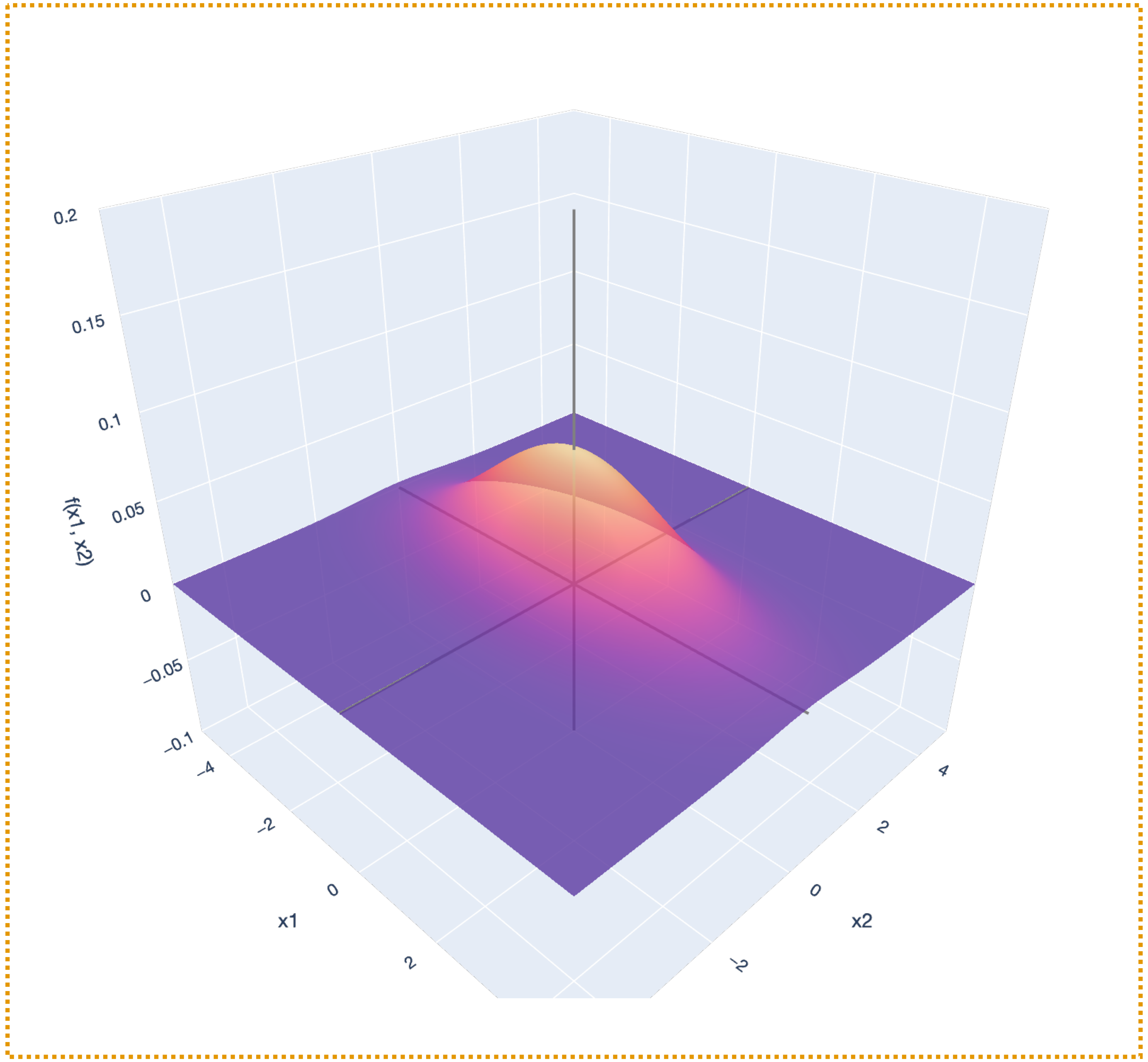
Example: $N(\mathbf{0}, \Sigma)$

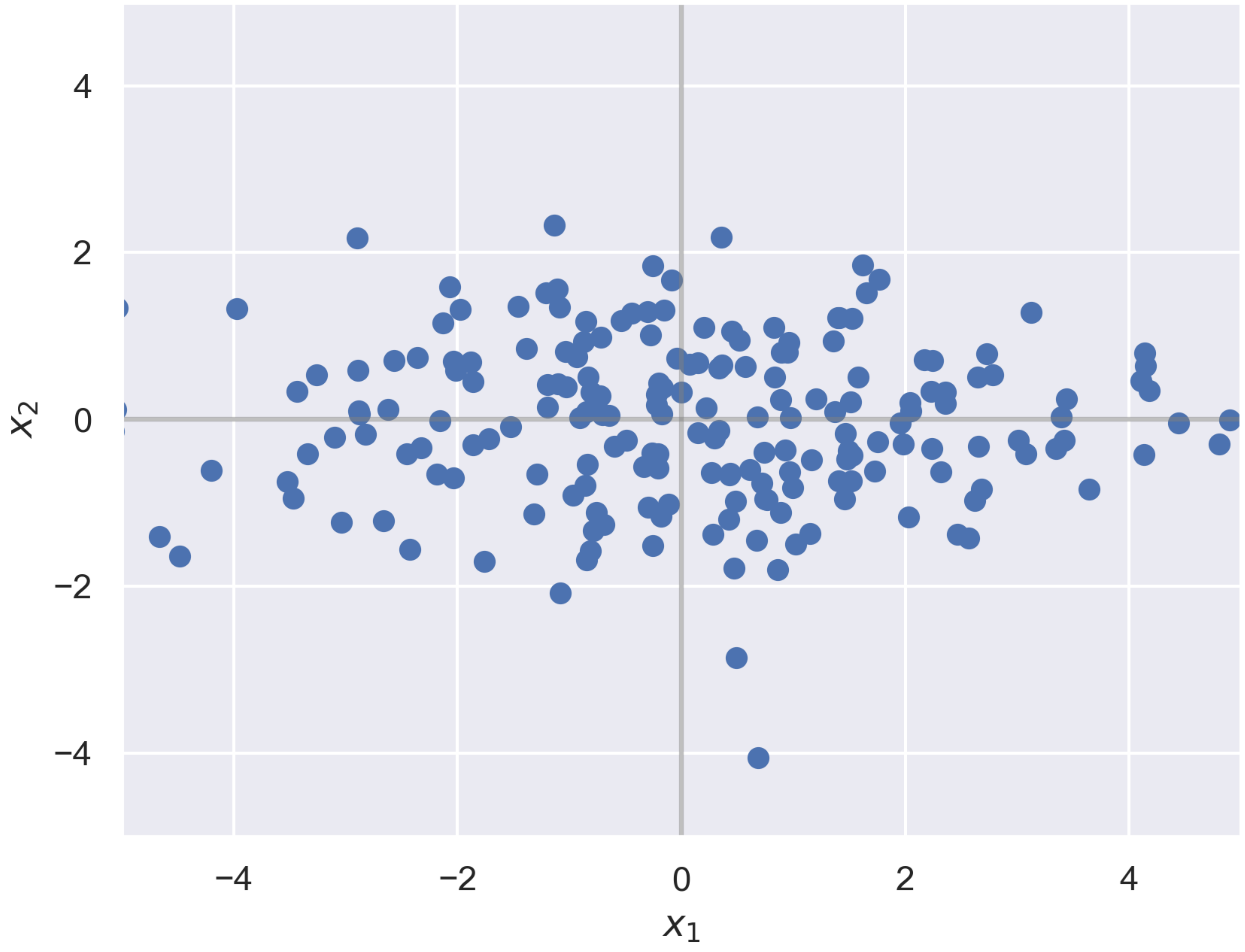




Multivariate Gaussian

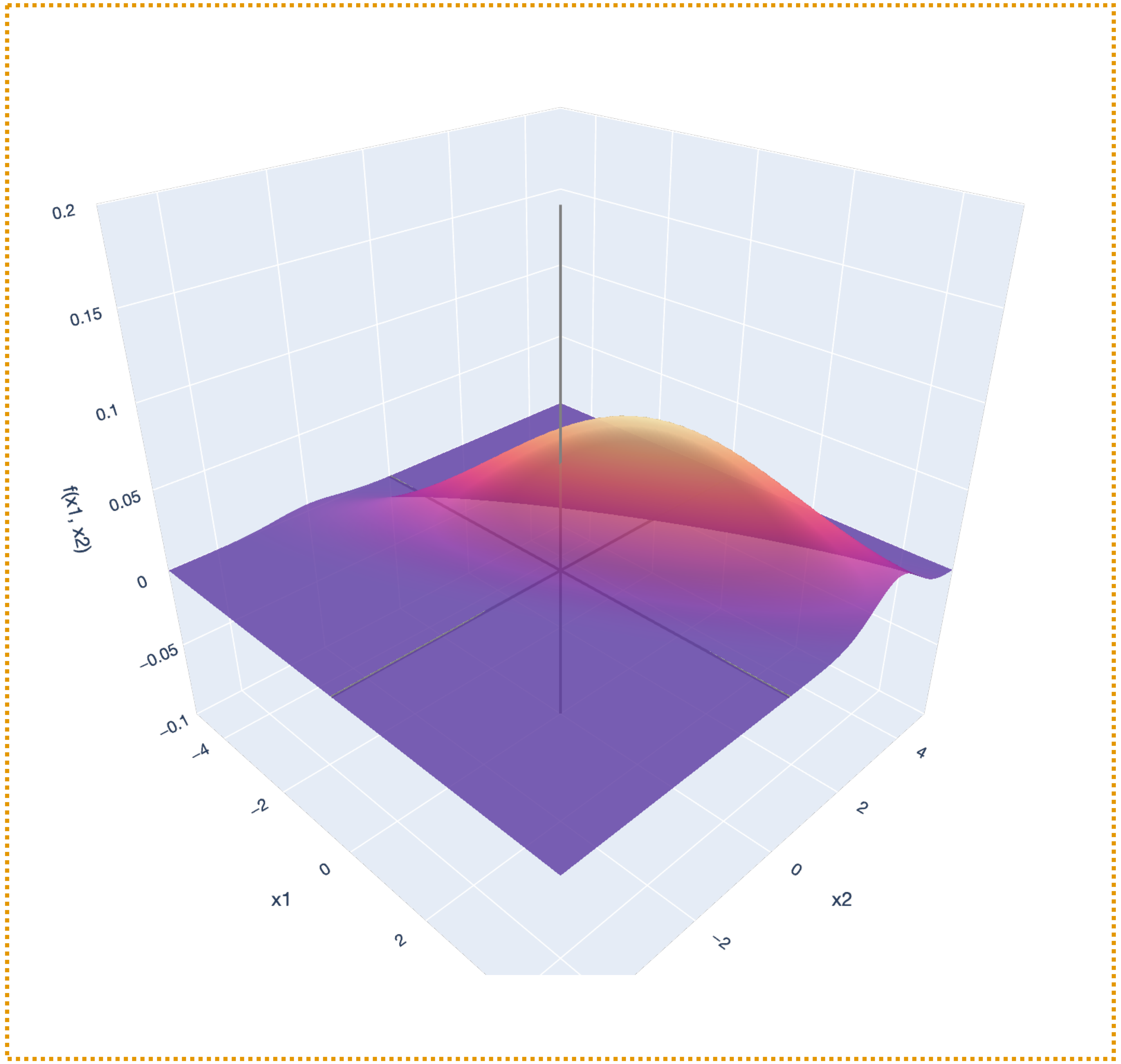
Example: $N(\mathbf{0}, \Sigma)$

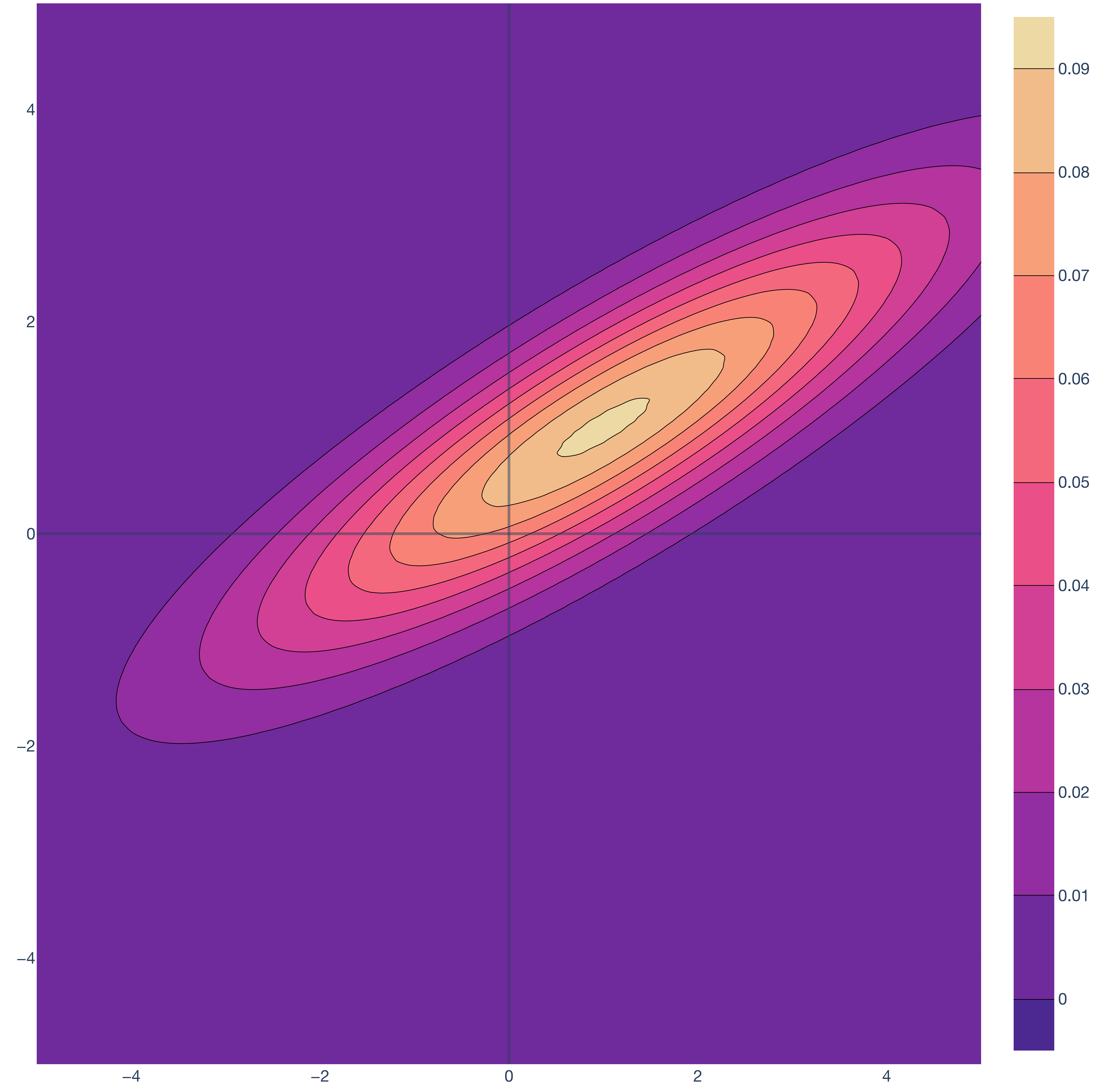




Multivariate Gaussian

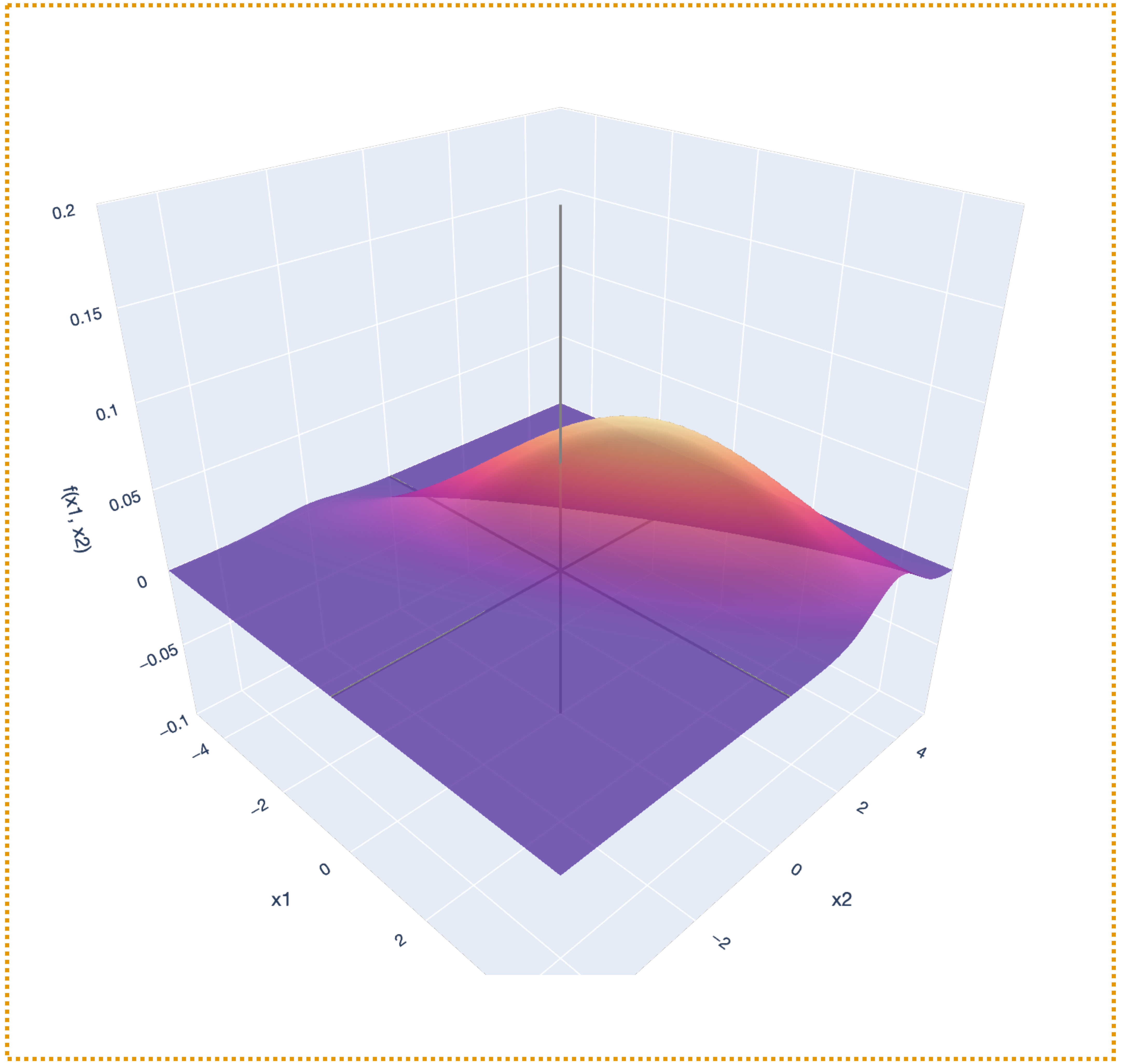
Example: $N(\mu, \Sigma)$

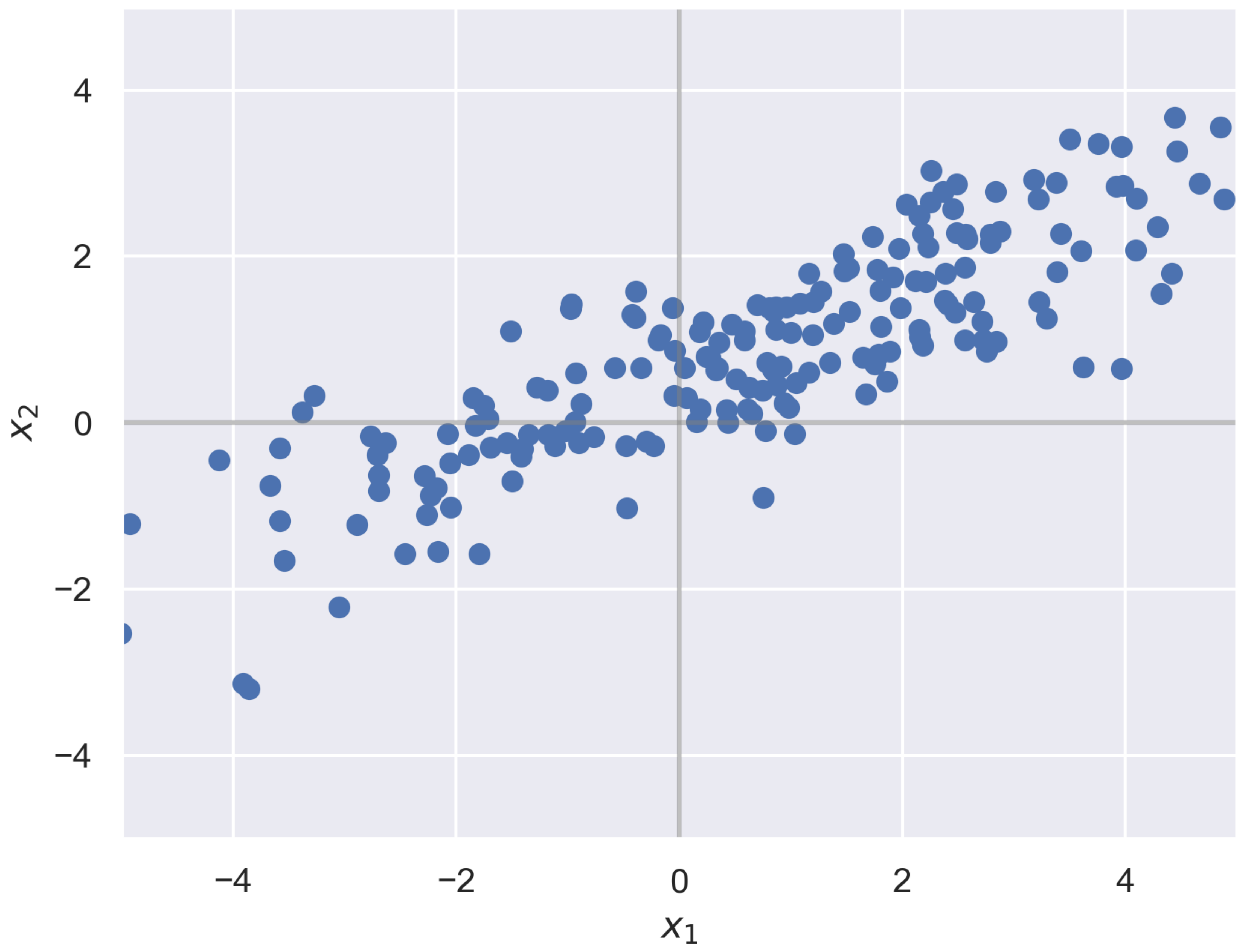




Multivariate Gaussian

Example: $N(\mu, \Sigma)$





Multivariate Gaussian

Diagonal Covariance and Factorization

Covariance of the MVN

Simple $d = 2$ Case

Consider the $d = 2$ case where $\Sigma \in \mathbb{R}^{2\times 2}$ is diagonal:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

What does the MVN density look like?

$$p(\mathbf{x}) = \frac{1}{2\pi \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

Determinant of 2×2 Matrix

Quick Definition

For a matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

the determinant of $\mathbf{A}$ is the scalar quantity:

$$\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}.$$

Determinant of Covariance Matrix

Applied to MVN

For a covariance matrix $\Sigma \in \mathbb{R}^{2\times2}$ written as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix},$$

the determinant of Σ is the scalar quantity:

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2.$$

Covariance of the MVN

Simple $d = 2$ Case

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

What does the MVN density look like?

$$p(\mathbf{x}) = \frac{1}{2\pi \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

$$\implies p(\mathbf{x}) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

Covariance of the MVN

Simple $d = 2$ Case

$$\implies p(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

Multiplying out the quadratic form…

$$\implies p(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

Covariance of the MVN

Simple $d = 2$ Case

$$\implies p(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right)$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right) \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right)$$

But this is just the product of two independent Gaussians!

$p(\mathbf{x}) = p(x_1) \cdot p(x_2)$, where $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$.

Factorization of the MVN

Theorem Statement

Theorem (Factorization of MVN). Let $\mathbf{x} = (x_1, \ldots, x_d) \sim N(\mu, \Sigma)$ be a multivariate Gaussian random vector, where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix and $\mu = (\mu_1, \ldots, \mu_d)$. Then, each coordinate x_i of $\mathbf{x}$ is an *independent* single-variable Gaussian random variable, with:

$$x_i \sim N(\mu_i, \sigma_i^2),$$

and the PDF of $\mathbf{x}$ factorizes into d marginal single-variable Gaussian PDFs:

$$p(\mathbf{x}) = \prod_{i=1}^{d} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\right).$$

Factorization of the MVN

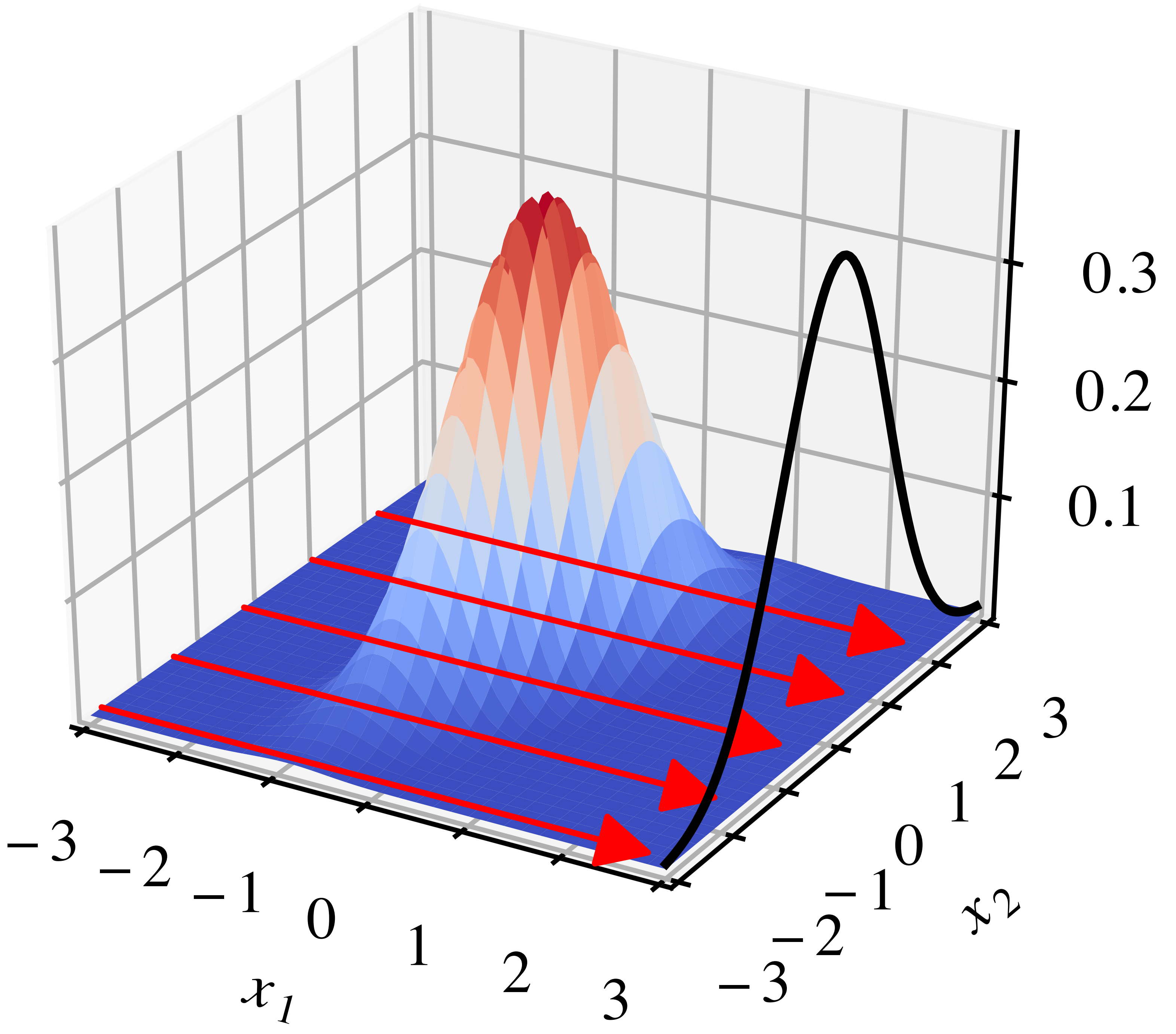
Theorem Statement

Theorem (Factorization of MVN). Let $\mathbf{x} = (x_1, \ldots, x_d) \sim N(\mu, \Sigma)$ be a multivariate Gaussian random vector, where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix and $\mu = (\mu_1, \ldots, \mu_d)$. Then, each coordinate x_i of $\mathbf{x}$ is an *independent* single-variable Gaussian random variable, with:

$$x_i \sim N(\mu_i, \sigma_i^2),$$

and the PDF of $\mathbf{x}$ factorizes into d marginal single-variable Gaussian PDFs:

$$p(\mathbf{x}) = \prod_{i=1}^{d} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2 \right).$$



Multivariate Gaussian

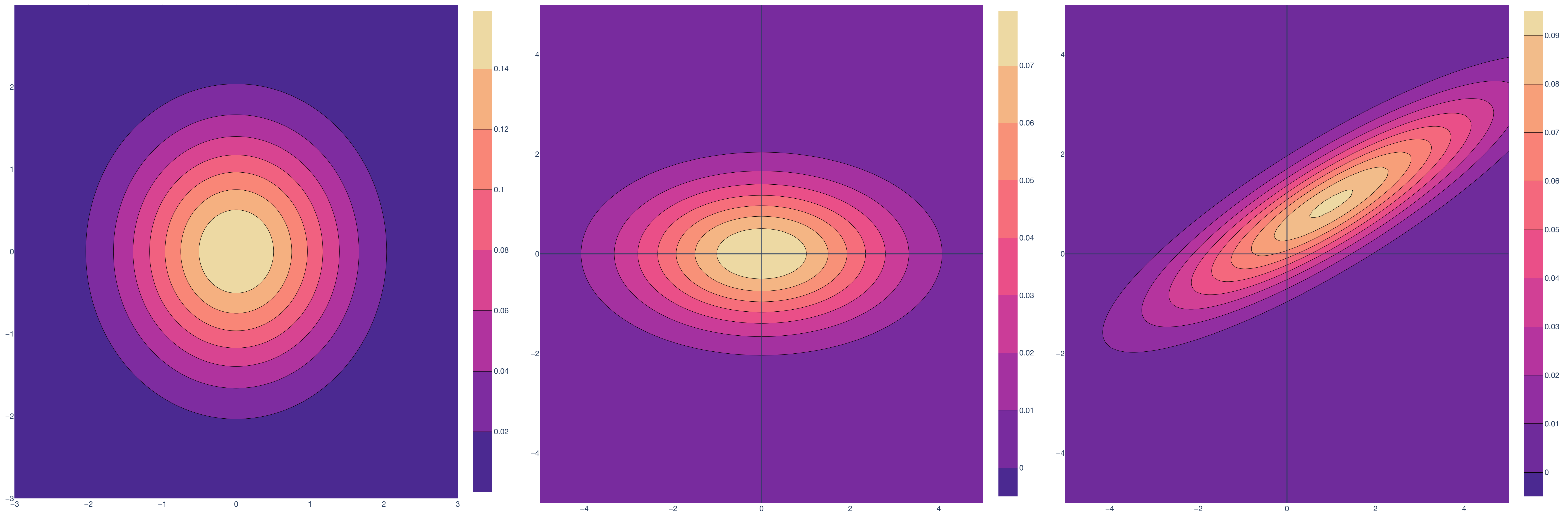
Contours and Geometry

Level Curves

Intuition and Definition

For a function $f : \mathbb{R}^d \to \mathbb{R}$, the level curves or isocontours of f at $c \in \mathbb{R}$ is the set of the form:

$$L_f(c) := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = c\}.$$



Geometry of MVN

Simple $d = 2$ Case

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$p(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

What are the level curves at some c?

Solve for: $p(\mathbf{x}) = c$.

Geometry of MVN

Simple $d = 2$ Case

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$p(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

Using some algebra, we can show that $p(\mathbf{x}) = c$ *when…*

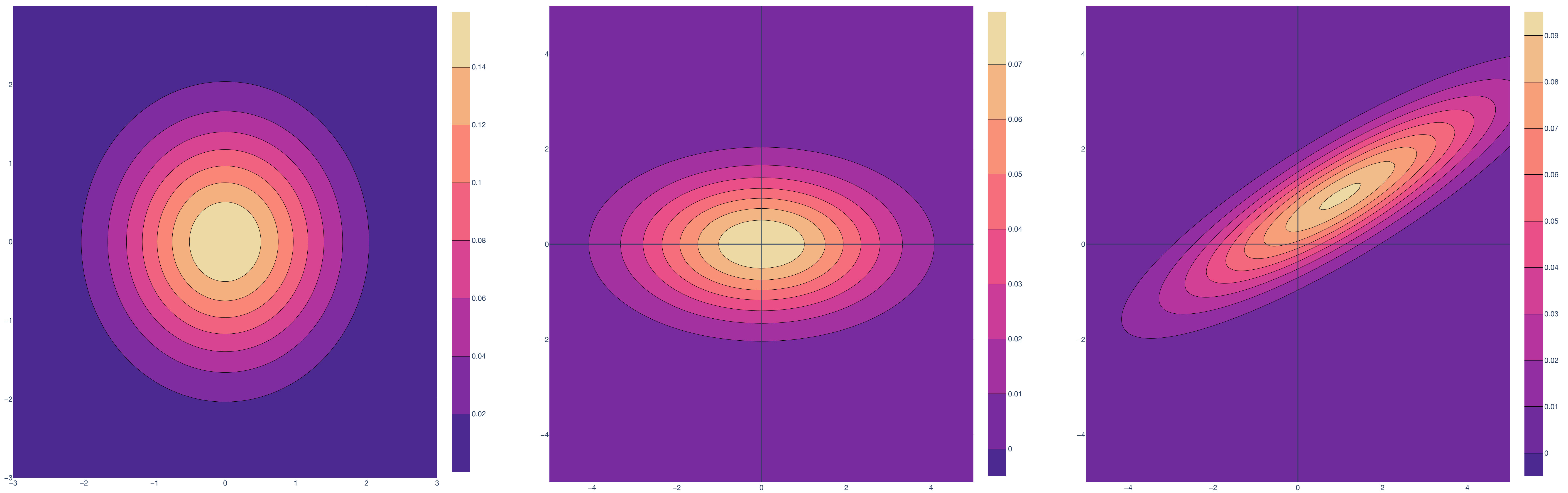
$$1 = \left(\frac{x_1 - \mu_1}{r_1} \right)^2 + \left(\frac{x_2 - \mu_2}{r_2} \right)^2, \text{ where } r_i = \sqrt{2\sigma_i^2 \log\left(\frac{1}{2\pi c \sigma_1 \sigma_2} \right)}.$$

Geometry of MVN

Simple $d = 2$ Case

Therefore, for $c \in \mathbb{R}$, the simple bivariate MVN has ellipse-shaped level curves:

$$1 = \left(\frac{x_1 - \mu_1}{r_1}\right)^2 + \left(\frac{x_2 - \mu_2}{r_2}\right)^2, \text{ where } r_i = \sigma_i \sqrt{2 \log\left(\frac{1}{2\pi c \sigma_1 \sigma_2}\right)}.$$



Geometry of MVN

Simple $d = 2$ Case

Therefore, for $c \in \mathbb{R}$, the simple bivariate MVN has ellipse-shaped level curves:

$$1 = \left(\frac{x_1 - \mu_1}{r_1}\right)^2 + \left(\frac{x_2 - \mu_2}{r_2}\right)^2, \text{ where } r_i = \sigma_i \sqrt{2 \log\left(\frac{1}{2\pi c \sigma_1 \sigma_2}\right)}.$$

For a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$, the eigenvalues are just σ_1 and σ_2 and the standard basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are eigenvectors!

Geometry of MVN

General Case

For positive definite $\mathbf{A}$, the associated quadratic form $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ looks like a *bowl/ellipsoid* with:

Axes in the direction of the eigenvectors of Σ.

Axis lengths proportional to the *inverse* square roots of the eigenvalues of $\mathbf{A}$:

$$r_1 \propto \frac{1}{\sqrt{\lambda_1}}, \ldots, r_d \propto \frac{1}{\sqrt{\lambda_d}}$$

Geometry of MVN

General Case

The quadratic form in the MVN exponent:

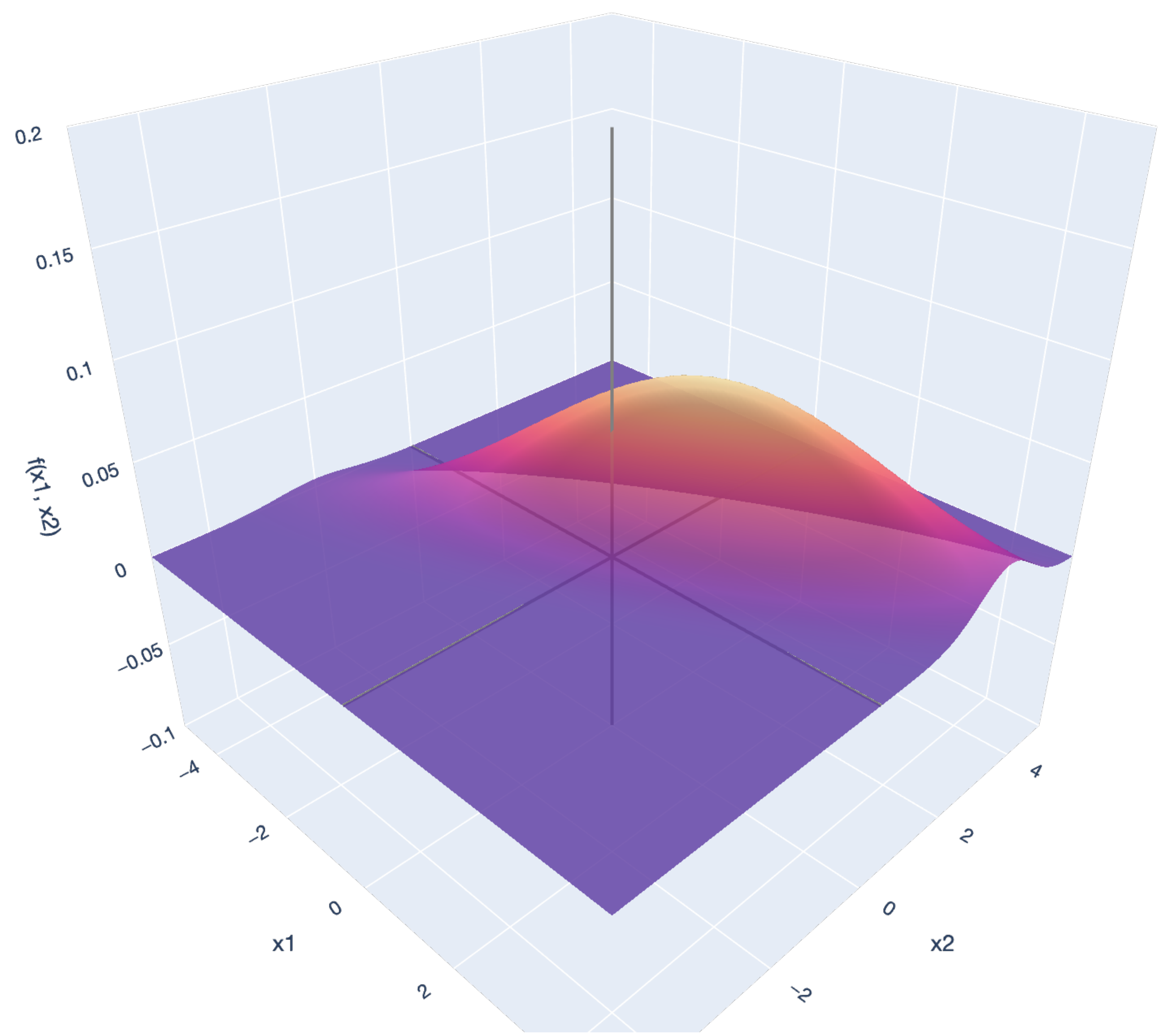
$$-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu).$$

Center of the ellipsoid is at μ.

Axes in the direction of the eigenvectors of Σ^{-1}.

Axis lengths proportional to *inverse* square roots of the eigenvalues of Σ^{-1}, or sq. roots of the eigenvalues of Σ.

$r_1 \propto \sqrt{\lambda_1}, \ldots, r_d \propto \sqrt{\lambda_d}$, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues of Σ.



Geometry of MVN

General Case

The quadratic form in the MVN exponent:

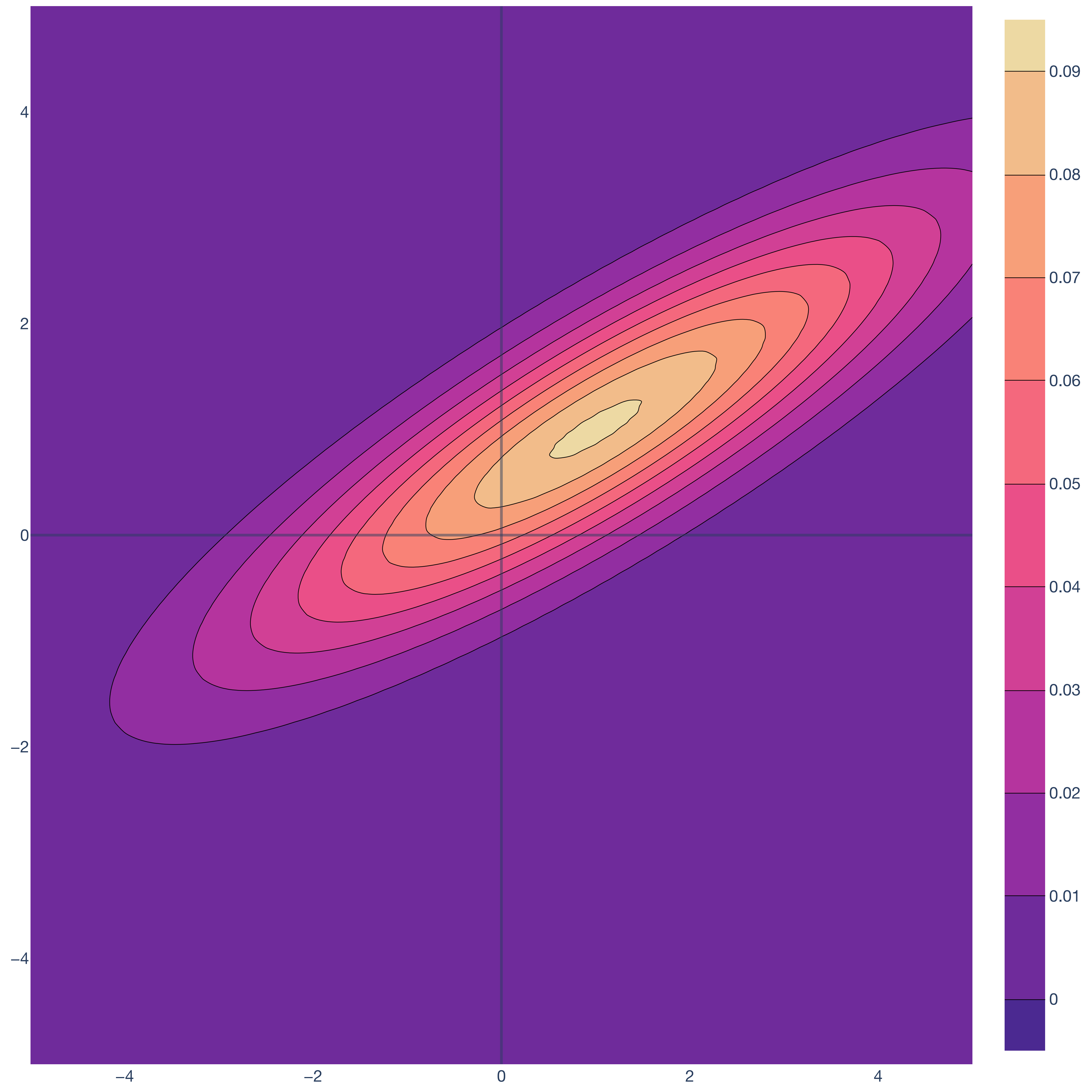
$$-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu).$$

Center of the ellipsoid is at μ.

Axes in the direction of the eigenvectors of Σ^{-1}.

Axis lengths proportional to *inverse* square roots of the eigenvalues of Σ^{-1}, or sq. roots of the eigenvalues of Σ.

$r_1 \propto \sqrt{\lambda_1}, \ldots, r_d \propto \sqrt{\lambda_d}$, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues of Σ.



Geometry of MVN

General Case

The quadratic form in the MVN exponent:

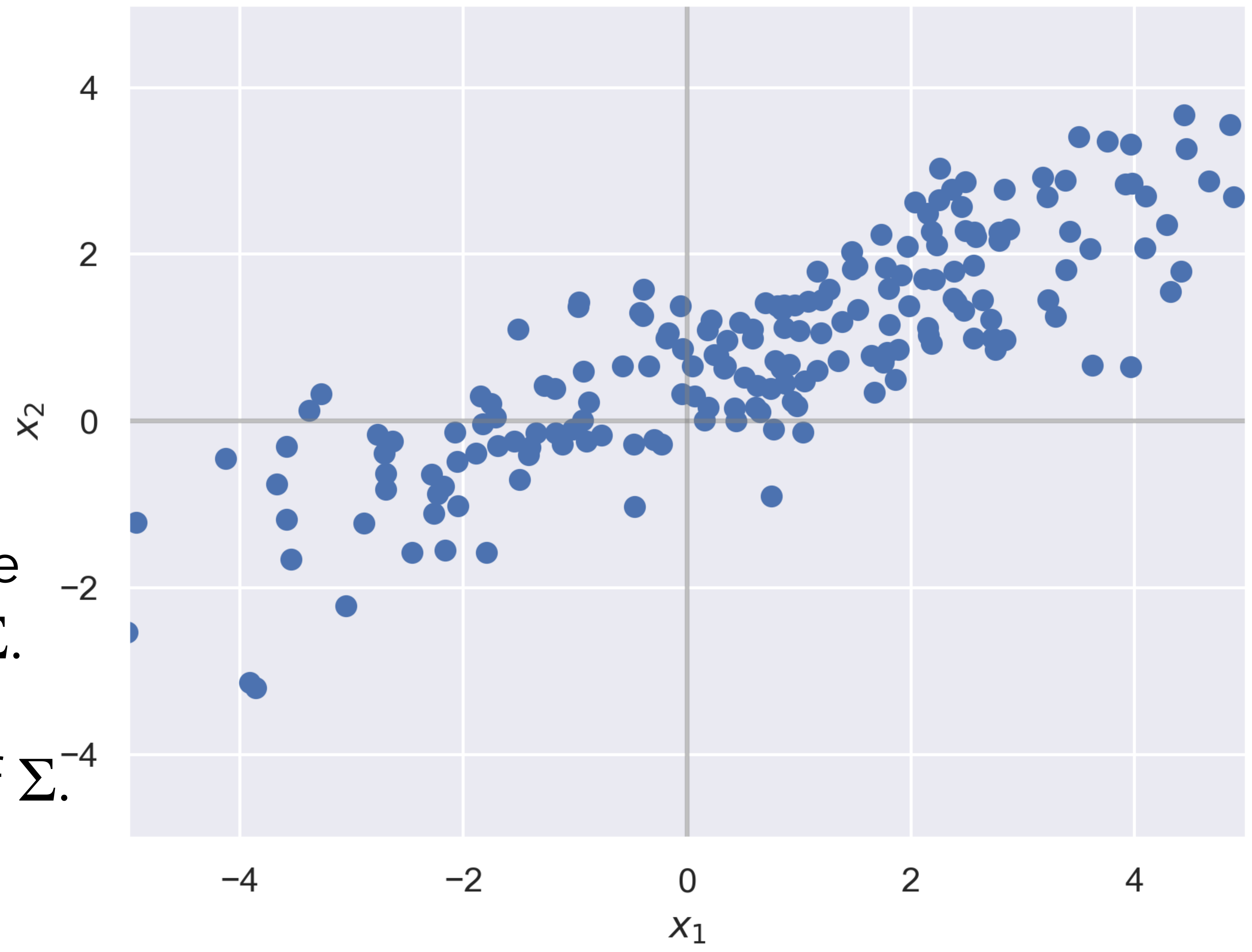
$$-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu).$$

Center of the ellipsoid is at μ.

Axes in the direction of the eigenvectors of Σ^{-1}.

Axis lengths proportional to *inverse* square roots of the eigenvalues of Σ^{-1}, or sq. roots of the eigenvalues of Σ.

$r_1 \propto \sqrt{\lambda_1}, \ldots, r_d \propto \sqrt{\lambda_d}$, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues of Σ.



Multivariate Gaussian

Linear Transformations

Diagonal Covariance Matrices

Why they're nice

If $\mathbf{x} \sim N(\mu, \Sigma)$ is MVN with *diagonal* covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \sigma_d^2 \end{bmatrix},$$

the eigenvectors are $\mathbf{e}_1, \dots, \mathbf{e}_d$ (the principal axes of the ellipsoid),

the eigenvalues are $\sigma_1^2, \dots, \sigma_d^2$ (the squared axes lengths),

the PDF factorizes: $p(\mathbf{x}) = p_{x_i}(s)$ where $p_{x_i}(s)$ is the PDF of $x_i \sim N(\mu_i, \sigma_i^2)$.

Diagonal Covariance Matrices

Why they're nice

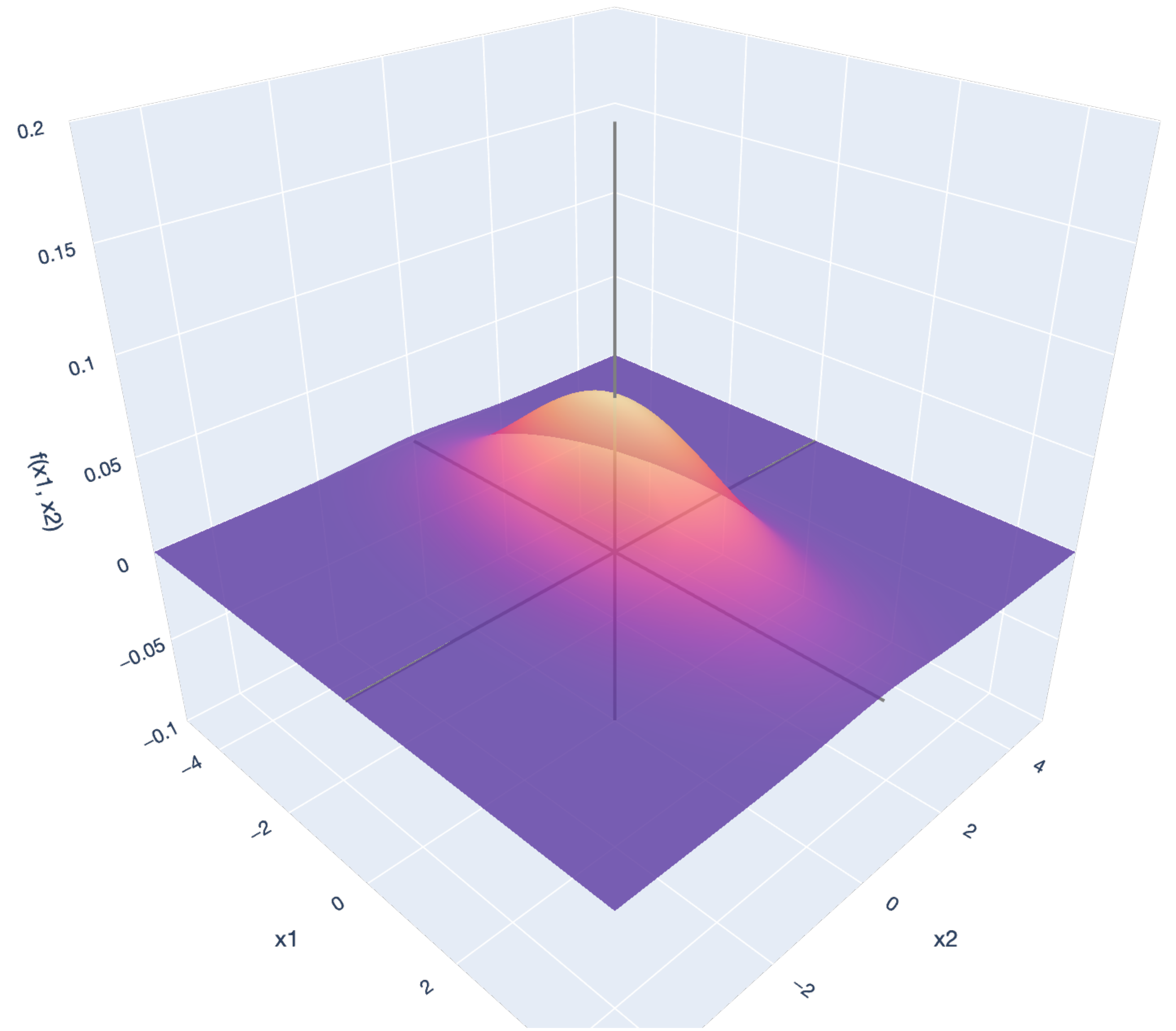
If $\mathbf{x} \sim N(\mu, \Sigma)$ is MVN with *diagonal* covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \sigma_d^2 \end{bmatrix},$$

the eigenvectors are $\mathbf{e}_1, \ldots, \mathbf{e}_d$ (the principal axes of the ellipsoid),

the eigenvalues are $\sigma_1^2, \ldots, \sigma_d^2$ (the squared axes lengths),

the PDF factorizes: $p(\mathbf{x}) = p_{x_i}(s)$ where $p_{x_i}(s)$ is the PDF of $x_i \sim N(\mu_i, \sigma_i^2)$.



Diagonal Covariance Matrices

Why they're nice

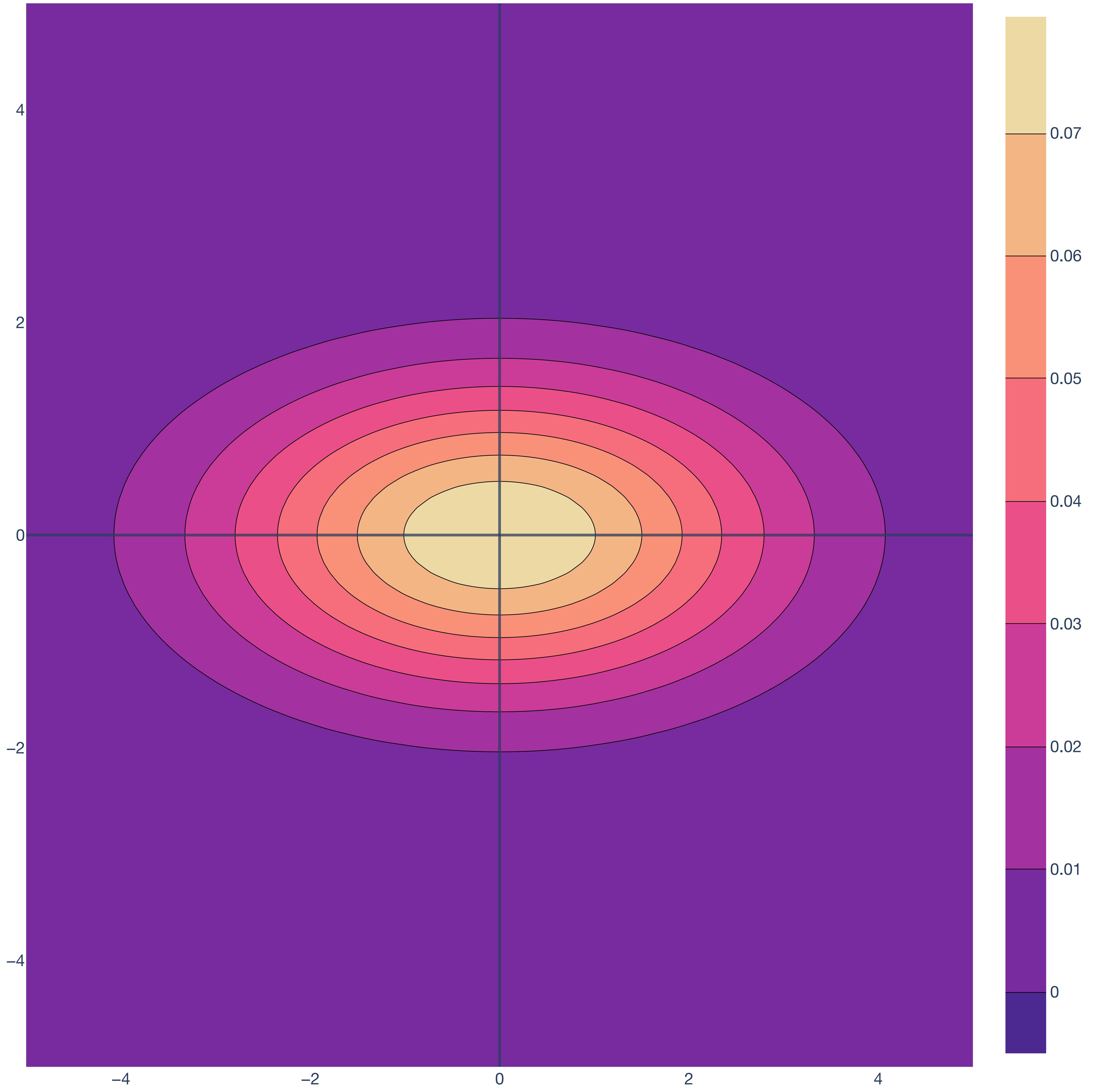
If $\mathbf{x} \sim N(\mu, \Sigma)$ is MVN with *diagonal* covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \sigma_d^2 \end{bmatrix},$$

the eigenvectors are $\mathbf{e}_1, \dots, \mathbf{e}_d$ (the principal axes of the ellipsoid),

the eigenvalues are $\sigma_1^2, \dots, \sigma_d^2$ (the squared axes lengths),

the PDF factorizes: $p(\mathbf{x}) = p_{x_i}(s)$ where $p_{x_i}(s)$ is the PDF of $x_i \sim N(\mu_i, \sigma_i^2)$.



Random Vectors

Variance and Covariance Matrix

The variance of a random vector generalizes to the covariance matrix

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

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

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

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

Nondiagonal MVN Covariance

Connection to Diagonal Covariance MVNs

Theorem (Nondiagonal MVNs). Let $\mathbf{x} \sim N(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then, there exists a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ such that $\mathbf{A}\mathbf{A}^\top = \Sigma$, and if

$$\mathbf{z} = \mathbf{A}^{-1}\left(\mathbf{x} - \mu\right),$$

then $\mathbf{z} \sim N(0, \mathbf{I})$.

Nondiagonal MVN Covariance

Connection to Diagonal Covariance MVNs

Theorem (Nondiagonal MVNs). Let $\mathbf{x} \sim N(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then, there exists a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ such that $\mathbf{A}\mathbf{A}^\top = \Sigma$, and if

$$\mathbf{z} = \mathbf{A}^{-1}\left(\mathbf{x} - \mu\right),$$

then $\mathbf{z} \sim N(0, \mathbf{I})$.

Analogue of single-variable fact: $X \sim N(\mu, \sigma^2)$ *gets "standardized" by taking* $Z = \dfrac{X - \mu}{\sigma}$.

Interpretation: Any multivariate Gaussian random vector $\mathbf{x}$ is the result of applying a linear transformation and translation (*affine transformation*): $\mathbf{X} = \mathbf{A}\mathbf{z}$ to a collection of d independent standard normal random variables $\mathbf{z} = (z_1, \ldots, z_d)$.

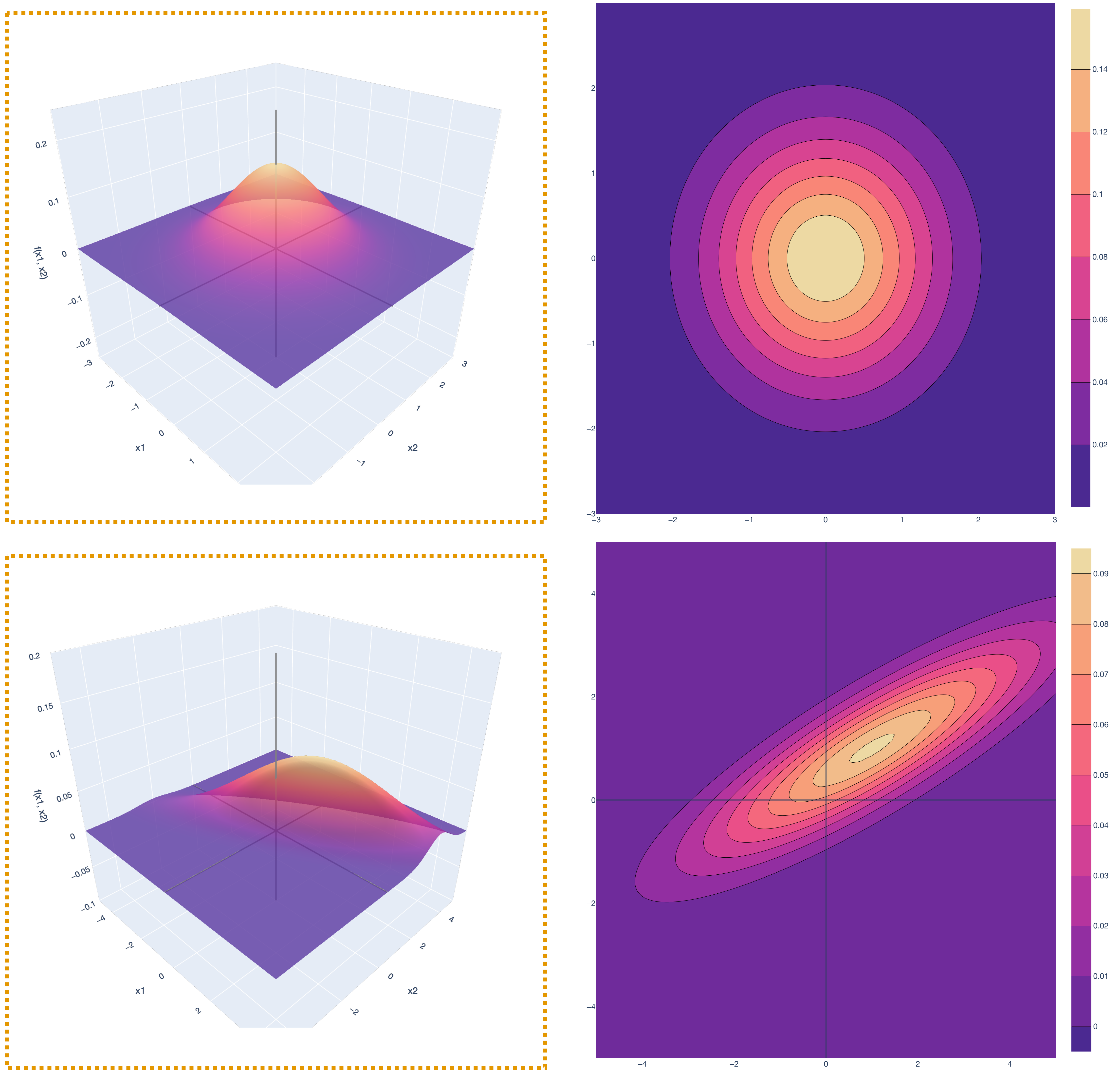
Nondiagonal MVN Covariance

Connection to Diagonal Covariance MVNs

Theorem (Nondiagonal MVNs). Let $\mathbf{x} \sim N(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then, there exists a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ such that $\mathbf{A}\mathbf{A}^\top = \Sigma$, and if

$$\mathbf{z} = \mathbf{A}^{-1}\left(\mathbf{x} - \mu\right),$$

then $\mathbf{z} \sim N(0, \mathbf{I})$.



Multivariate Gaussian

Other Basic Properties

Other Properties of MVN

Linear Combinations

Theorem (Linear Combinations of MVNs). Let $\mathbf{x} \sim N(\mu, \Sigma)$ be an MVN random vector.

Let $\mathbf{b} \in \mathbb{R}^d$. $\mathbf{x} \sim N(\mu, \Sigma)$ if and only if any linear combination $\mathbf{b}^\top \mathbf{x}$ has a single-variable Gaussian distribution, $\mathbf{b}^\top \mathbf{x} \sim N(\mathbf{b}^\top \mu, \mathbf{b}^\top \Sigma \mathbf{b})$.

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. The affine transformation is distributed as MVN: $\mathbf{A}\mathbf{x} + \mathbf{b} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^\top)$.

Other Properties of MVN

Linear Combinations

Theorem (Independence). Let $\mathbf{x} \sim N(\mu, \Sigma)$ be an MVN random vector, written $\mathbf{x} = (x_1, \ldots, x_d)$. Then, x_i and x_j are independent if and only if $\Sigma_{ij} = 0$.

Also, if x_i and x_j are all pairwise independent for $i \neq j$, the set of random variables $x_1, \ldots, x_d$ are completely independent.

Other Properties of MVN

Marginal and Conditional Distributions

Let $\mathbf{x} \sim N(\mu, \Sigma)$ be multivariate normal, *partitioned* into parts:

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2), \text{ where } \mathbf{x}_1 \in \mathbb{R}^k \text{ and } \mathbf{x}_2 \in \mathbb{R}^{d-k}.$$

Also partition μ into

$$\mu = (\mu_1, \mu_2), \text{ where } \mu_1 \in \mathbb{R}^k \text{ and } \mu_2 \in \mathbb{R}^{d-k},$$

and $\Sigma \in \mathbb{R}^{d \times d}$ into

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{ where } \Sigma_{11} \in \mathbb{R}^{k \times k}, \Sigma_{21} \in \mathbb{R}^{(d-k) \times k}, \text{ etc.}$$

Other Properties of MVN

Marginal Distributions

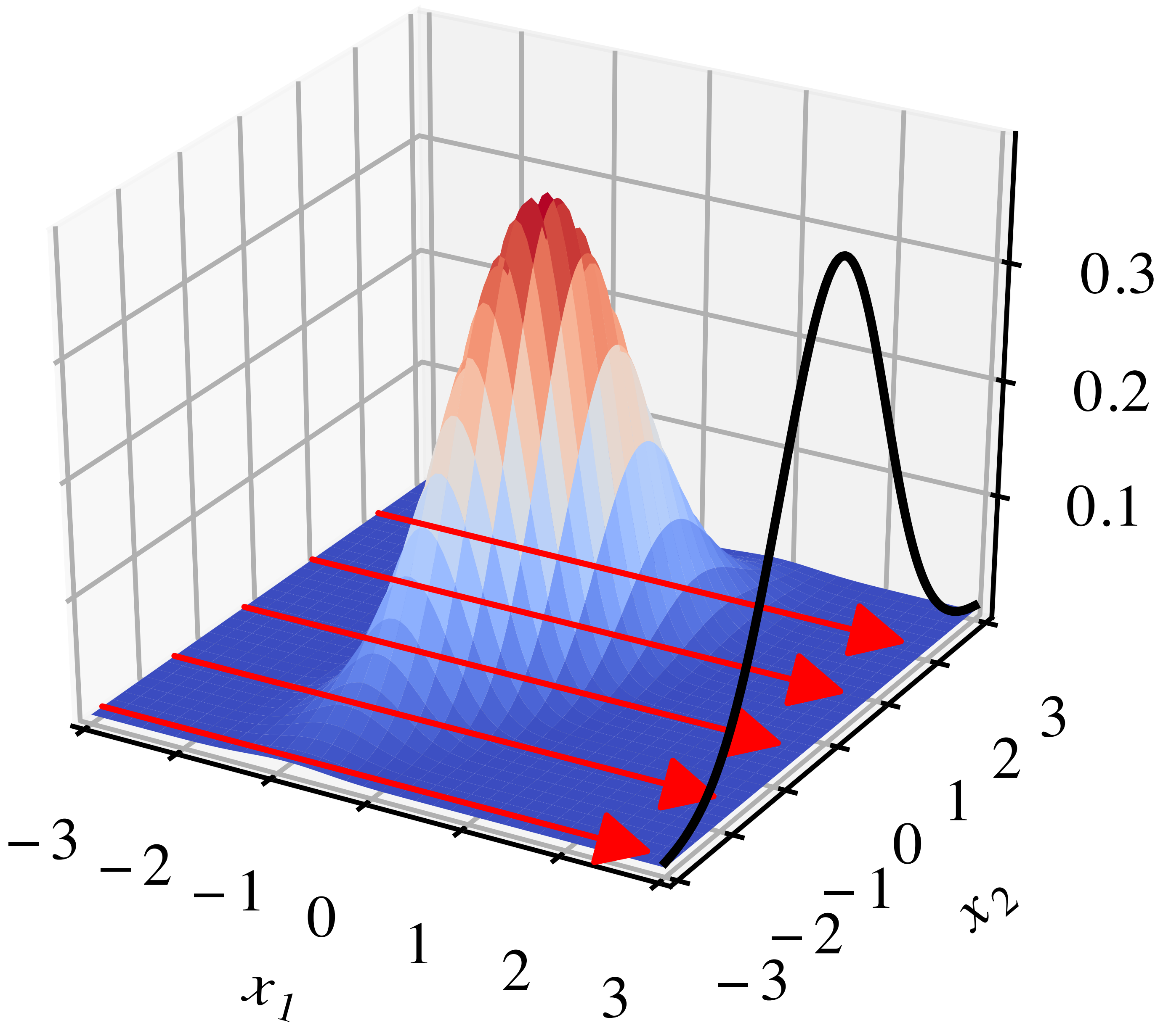
Theorem (Marginal Distributions). Let $\mathbf{x} \sim N(\mu, \Sigma)$ be an MVN random vector, partitioned:

$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1 \in \mathbb{R}^k$ and $\mathbf{x}_2 \in \mathbb{R}^{d-k}$.

$\mu = (\mu_1, \mu_2)$, where $\mu_1 \in \mathbb{R}^k$ and $\mu_2 \in \mathbb{R}^{d-k}$,

$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, where $\Sigma_{11} \in \mathbb{R}^{k \times k}$, $\Sigma_{21} \in \mathbb{R}^{(d-k) \times k}$, etc.

Then, $\mathbf{x}_1 \sim N(\mu_1, \Sigma_{11})$ and $\mathbf{x}_2 \sim N(\mu_2, \Sigma_{22})$ are multivariate Gaussians.



Other Properties of MVN

Conditional Distributions

Theorem (Conditional Distributions). Let $\mathbf{x} \sim N(\mu, \Sigma)$ be an MVN random vector, partitioned:

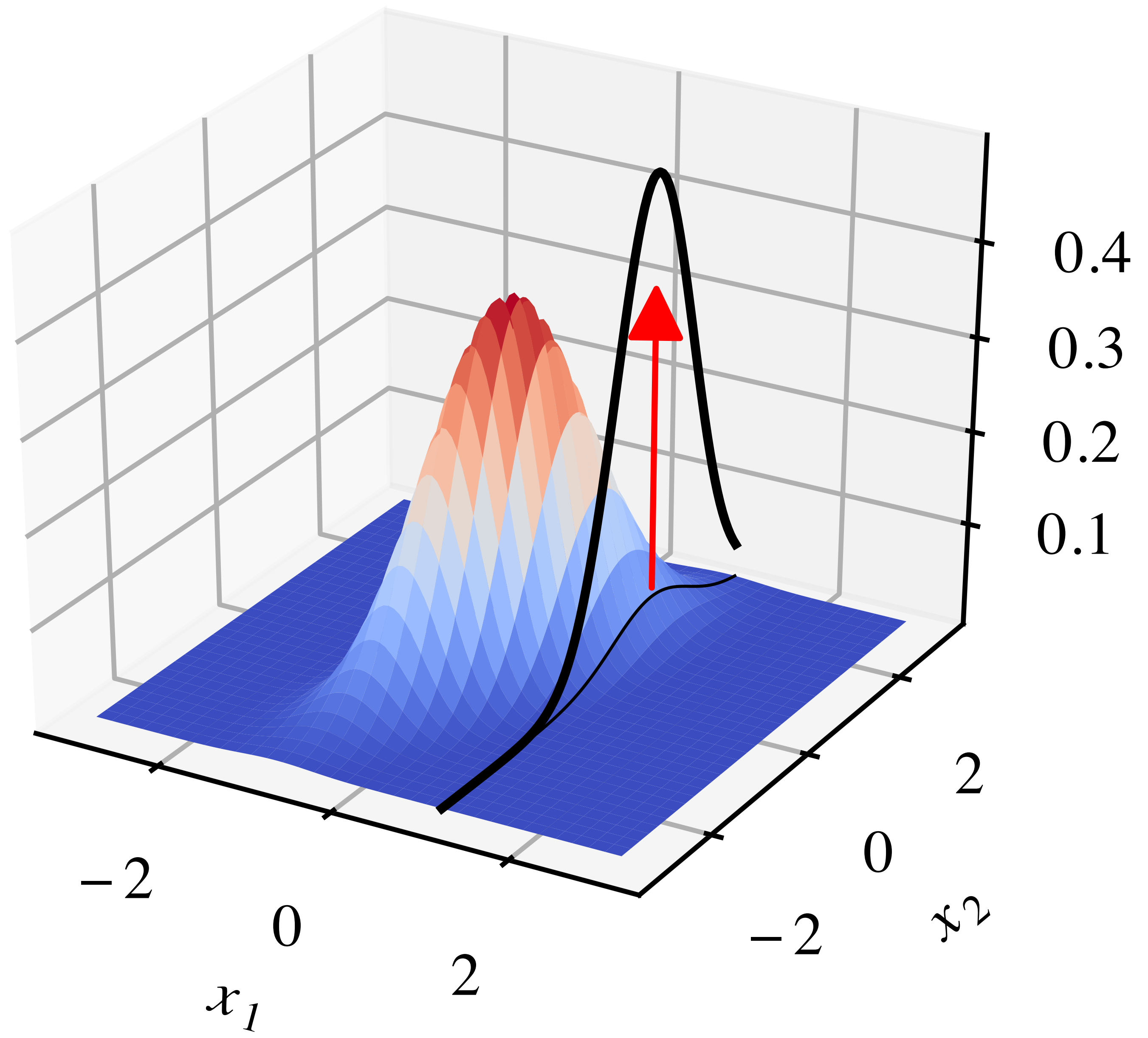
$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1 \in \mathbb{R}^k$ and $\mathbf{x}_2 \in \mathbb{R}^{d-k}$.

$\mu = (\mu_1, \mu_2)$, where $\mu_1 \in \mathbb{R}^k$ and $\mu_2 \in \mathbb{R}^{d-k}$,

$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, where $\Sigma_{11} \in \mathbb{R}^{k \times k}$, $\Sigma_{21} \in \mathbb{R}^{(d-k)\times k}$, etc.

Then, the conditional distribution of $\mathbf{x}_1 \mid \mathbf{x}_2$ is multivariate Gaussian with:

$$\mathbf{x}_1 \mid \mathbf{x}_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \, .$$



Recap

Lesson Overview

OLS under Gaussian Error Model. The distribution of $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ under the Gaussian error model is itself multivariate normal.

Multivariate Gaussian/Normal (MVN) Distribution PDF. We define the multivariate Gaussian/normal distribution and study some simple examples.

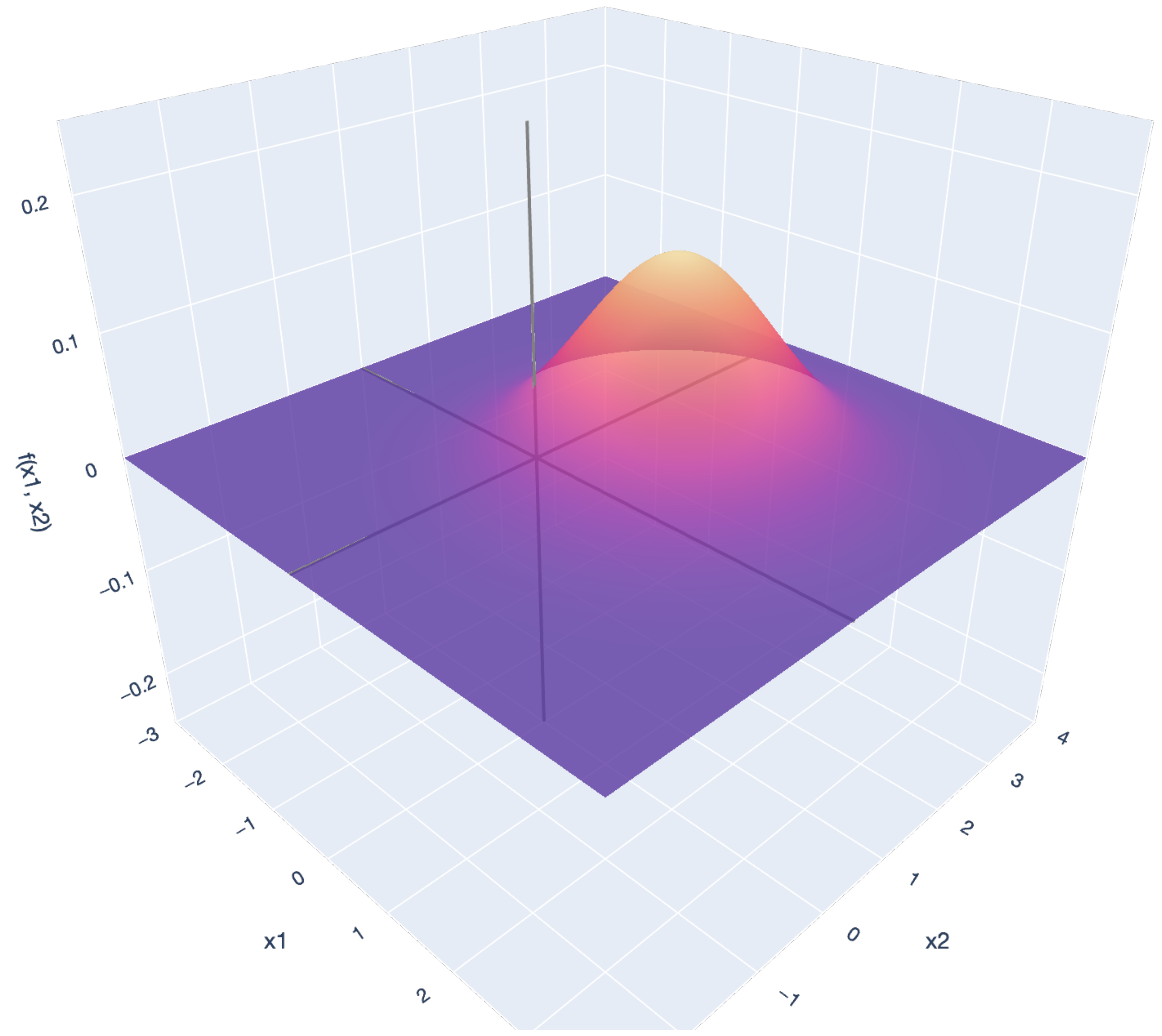
Factorization of the Multivariate Gaussian. We see that a multivariate Gaussian with a diagonal covariance matrix factors into independent Gaussians.

Geometry of the Multivariate Gaussian. We study the geometry of the multivariate Gaussian through its level curves and discover the it is ellipsoidal, with axes determined by the eigenvectors/eigenvalues of the covariance matrix.

Affine Transformations of the Multivariate Gaussian. We establish that any multivariate Gaussian is just an affine transformation away from the standard multivariate Gaussian.

Lesson Overview

Big Picture: Least Squares



Lesson Overview

Big Picture: Gradient Descent