Math for Machine Learning

Week 3.2: Taylor Series, Linearization, and Gradient Descent

By: Samuel Deng

Logistics & Announcements

Lesson Overview

Linearization for approximation. We explore using the ***linearization*** of a function to approximate it. This is also called a "first-order approximation."

Taylor series. We define the ***Taylor series*** of a function, which is an "infinite polynomial" that approximates a function at a point.

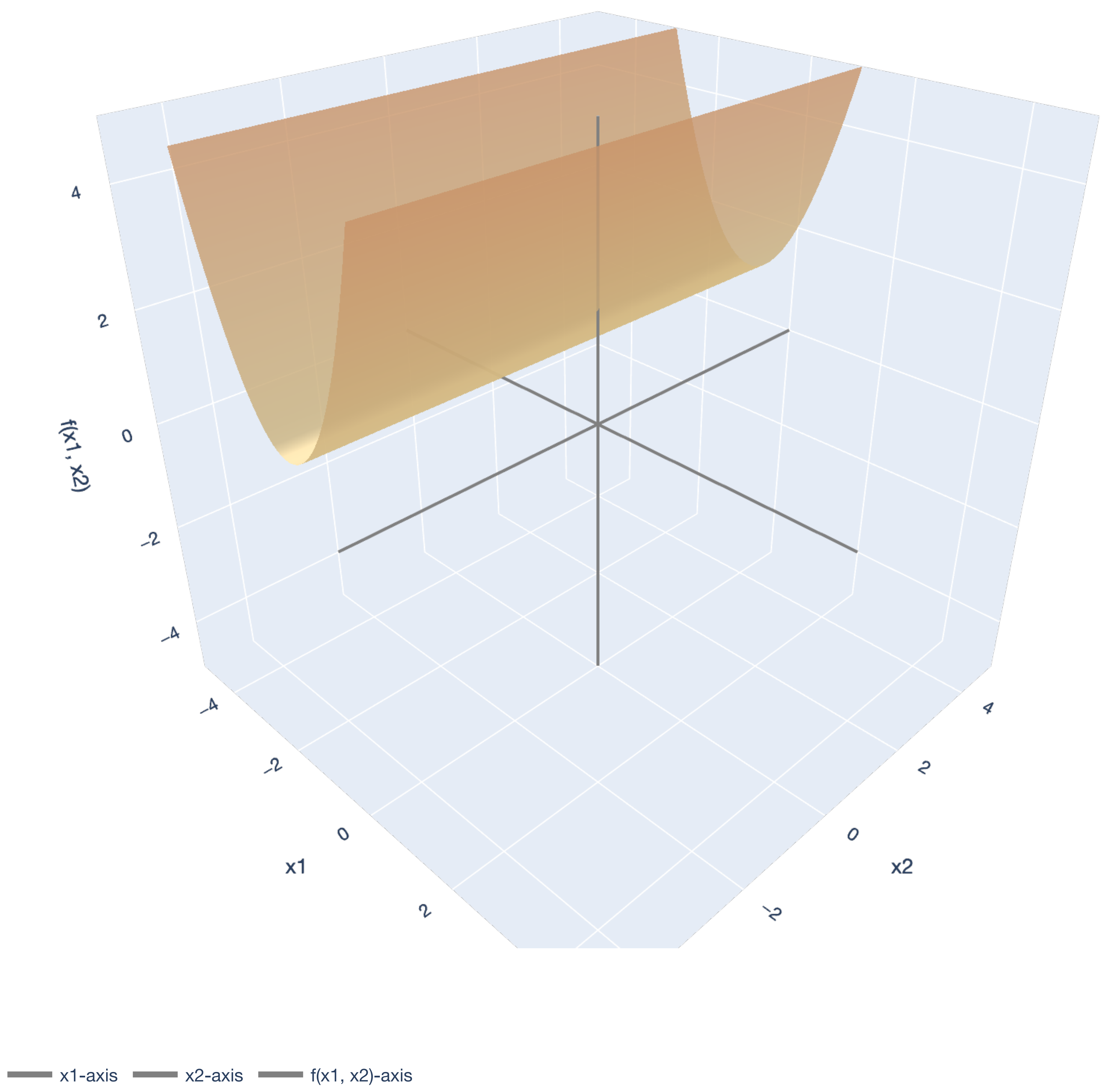
First-order and second-order Taylor approximation. The Taylor polynomial allows us to approximate a funciton by "chopping it off" at a certain degree.

Taylor's Theorem. To quantify how bad our approximations are, we can use ***Taylor's Theorem.*** We present two forms of Taylor's Theorem (Peano and Lagrange).

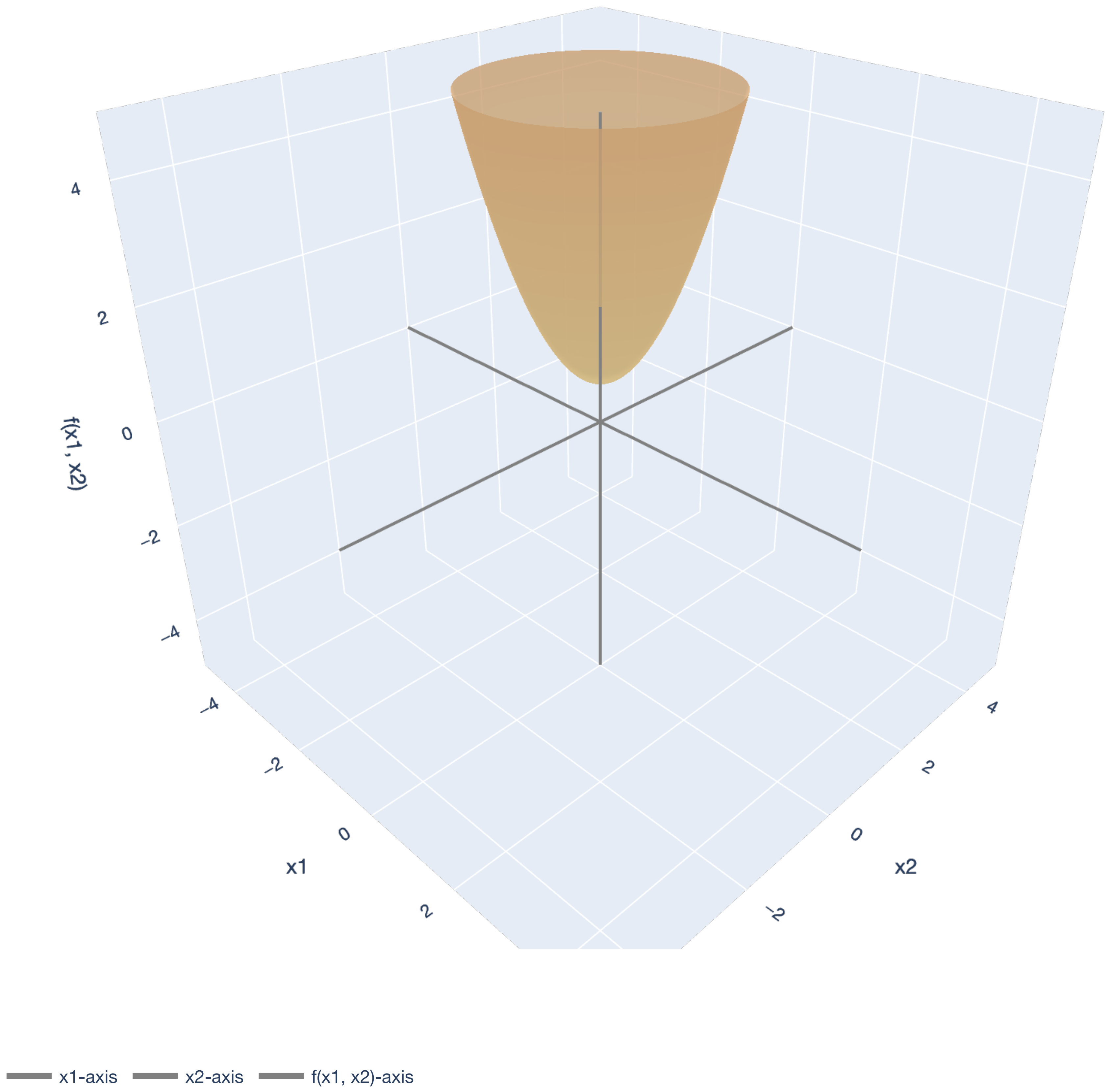
Gradient descent. We write down the full algorithm for ***gradient descent***, the second "story" of our course. Using Taylor's Theorem, we can prove that, for ***β-smooth functions***, GD makes the function value smaller from iteration to iteration, as long as we set the "step size" small enough.

Lesson Overview

Big Picture: Least Squares



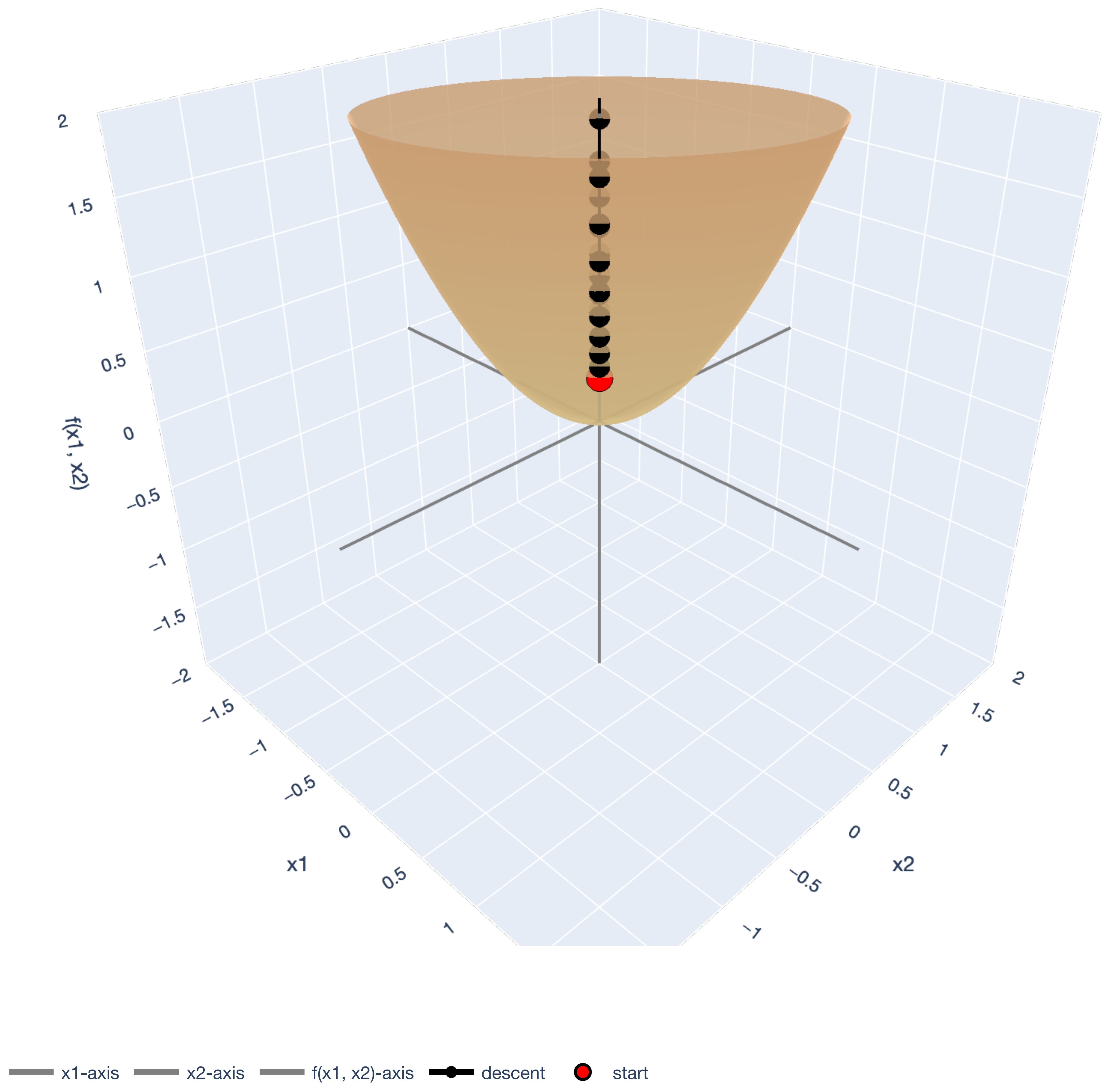
$$\lambda_1, \ldots, \lambda_d \geq 0$$

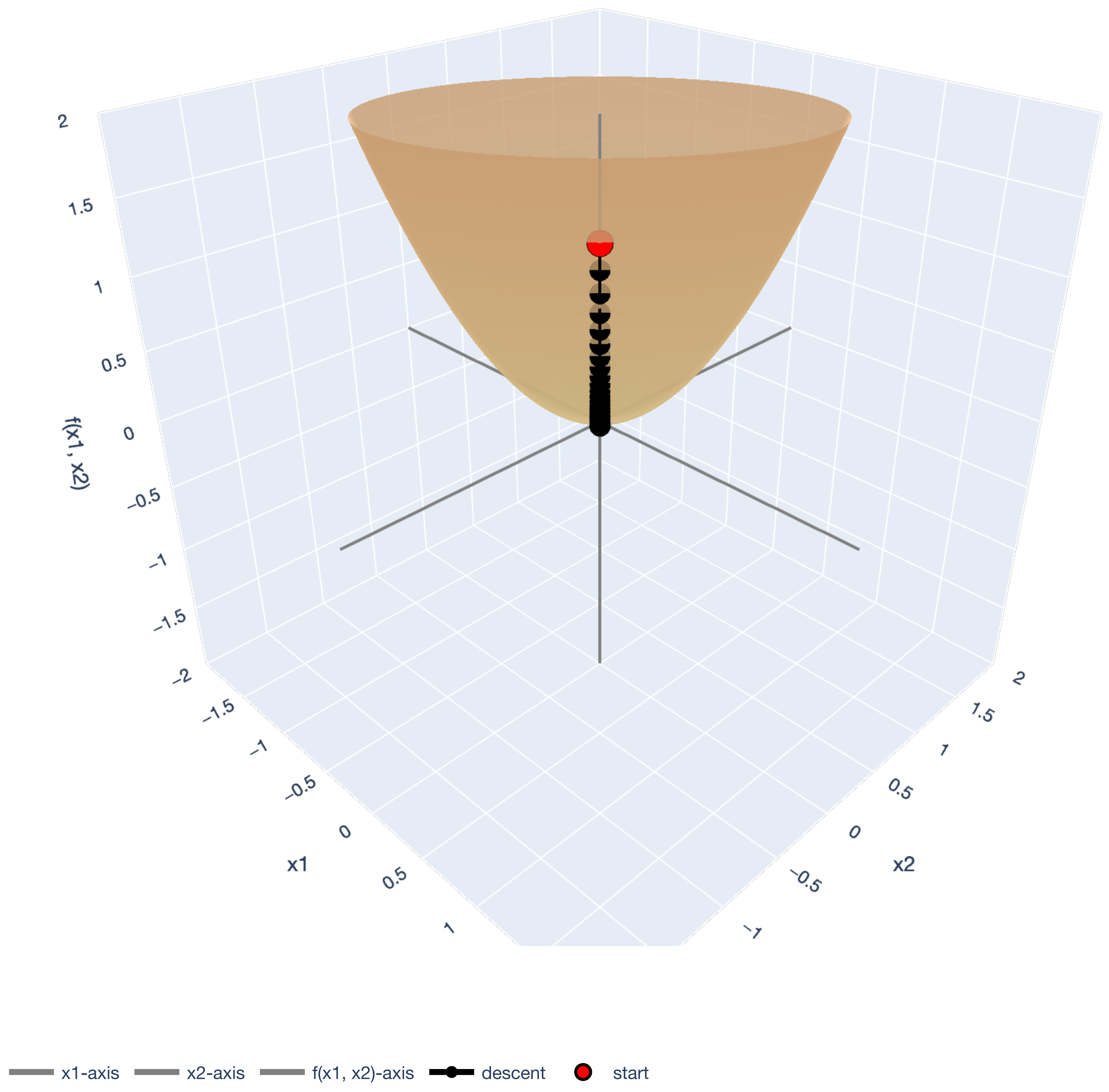


$$\lambda_1, \ldots, \lambda_d > 0$$

Lesson Overview

Big Picture: Gradient Descent





Linearization

Derivatives to find linear approximations

Motivation

Optimization in calculus

In much of machine learning, we design algorithms for well-defined *optimization problems.*

In an optimization problem, we want to minimize an ***objective function*** $f : \mathbb{R}^d \to \mathbb{R}$ with respect to a set of constraints $\mathscr{C} \subseteq \mathbb{R}^d$:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & x \in \mathscr{C} \end{aligned}$$

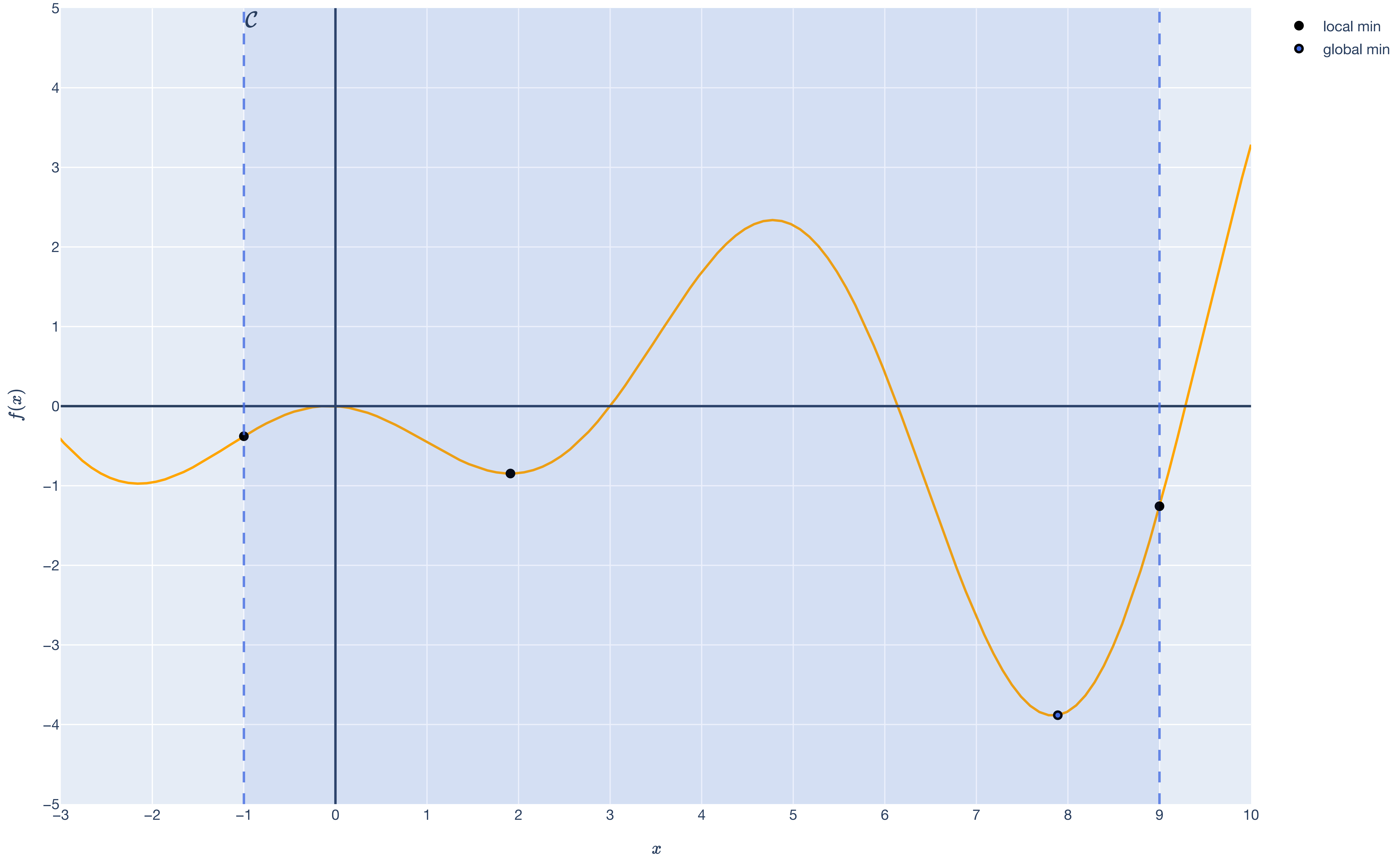
Motivation

Optimization in single-variable calculus

Ultimate goal: Find the *global minimum* of functions.

Intermediary goal: Find the *local minima.*

Derivatives give us the direction of steepest descent!



Multivariable Differentiation

Total Derivative

In this lecture, we'll focus on scalar-valued multivariable functions $f : \mathbb{R}^d \to \mathbb{R}$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function and let $\mathbf{x}_0 \in \mathbb{R}^d$ be a point. If there exists a gradient vector $\nabla f(\mathbf{x}_0) \in \mathbb{R}^d$ such that

$$\lim_{\vec{\delta} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0) - \nabla f(\mathbf{x}_0)^\top \vec{\delta}}{||\vec{\delta}||} = 0,$$

then f is ***differentiable*** *at* $\mathbf{x}_0$ and has the ***(total) derivative*** $\nabla f(\mathbf{x}_0)$.

Think of $\vec{\delta}$ as a "change in $\mathbf{x}$": for a base point $\mathbf{x}_0$ and a "destination point" $\mathbf{x}'$, think of $\vec{\delta} = \mathbf{x}' - \mathbf{x}_0$.

Multivariable Differentiation

Partial Derivative

Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $\mathbf{e}_i$ be the ith standard basis vector in $\mathbb{R}^d$. The *ith partial derivative* of f at $\mathbf{x}_0$ is

$$\frac{\partial f}{\partial x_i}(\mathbf{x}_0) := \lim_{\delta \to 0} \frac{f(\mathbf{x}_0 + \delta \mathbf{e}_i) - f(\mathbf{x}_0)}{\delta}$$

This is the derivative of f when keeping all but one variable constant.

Multivariable Differentiation

Partial Derivative

Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $\mathbf{e}_i$ be the ith standard basis vector in $\mathbb{R}^d$. The *ith partial derivative* of f at $\mathbf{x}_0$ is

$$\frac{\partial f}{\partial x_i}(\mathbf{x}_0) := \lim_{\delta \to 0} \frac{f(\mathbf{x}_0 + \delta \mathbf{e}_i) - f(\mathbf{x}_0)}{\delta}$$

This is the derivative of f when keeping all but one variable constant.

If f is *differentiable* at $\mathbf{x}$, then:

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d}\right]^\top \in \mathbb{R}^d$$

Linearity and Differentiation

Replacing nonlinear functions with linear function

The derivative is a linear transformation that maps changes in inputs to changes in outputs. We like linear transformations!

$$T : \text{change in } \mathbf{x} \to \text{change in } \mathrm{f}(\mathbf{x})$$

$$\nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) \approx f(\mathbf{x}) - f(\mathbf{x}_0)$$

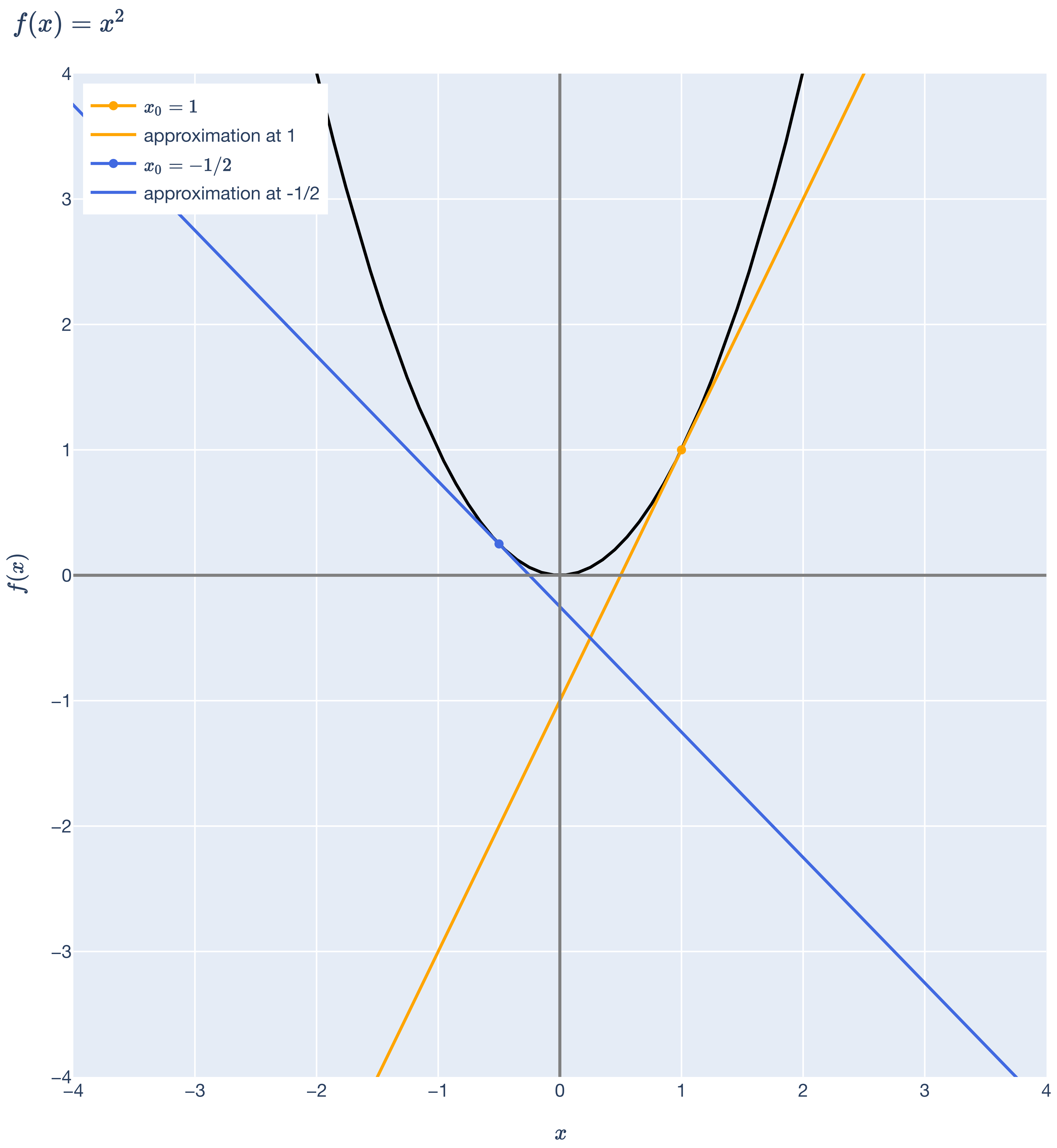
A goal of differential calculus, for us, is to replace nonlinear functions with linear approximations!

Linearization

The behavior of a differentiable function close to a point $\mathbf{x}$ can be approximated with the linear transformation given by its derivative.

For $\mathbf{x}$ close to $\mathbf{x}_0$,

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0).$$



Linearization

Derivative definition, one more time

$$\lim_{\vec{\delta} \to \mathbf{0}} \frac{f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0) - \nabla f(\mathbf{x}_0)^\top \vec{\delta}}{||\vec{\delta}||} = 0$$

The $\vec{\delta}$ vector is the "change in $\mathbf{x}$." Think of it as $\mathbf{x}' - \mathbf{x}_0$ for some "destination" $\mathbf{x}'$.

The term $f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0)$ is the "change in f."

The term $\nabla f(\mathbf{x}_0)^\top \vec{\delta}$ is the "linear approximation of the change in f."

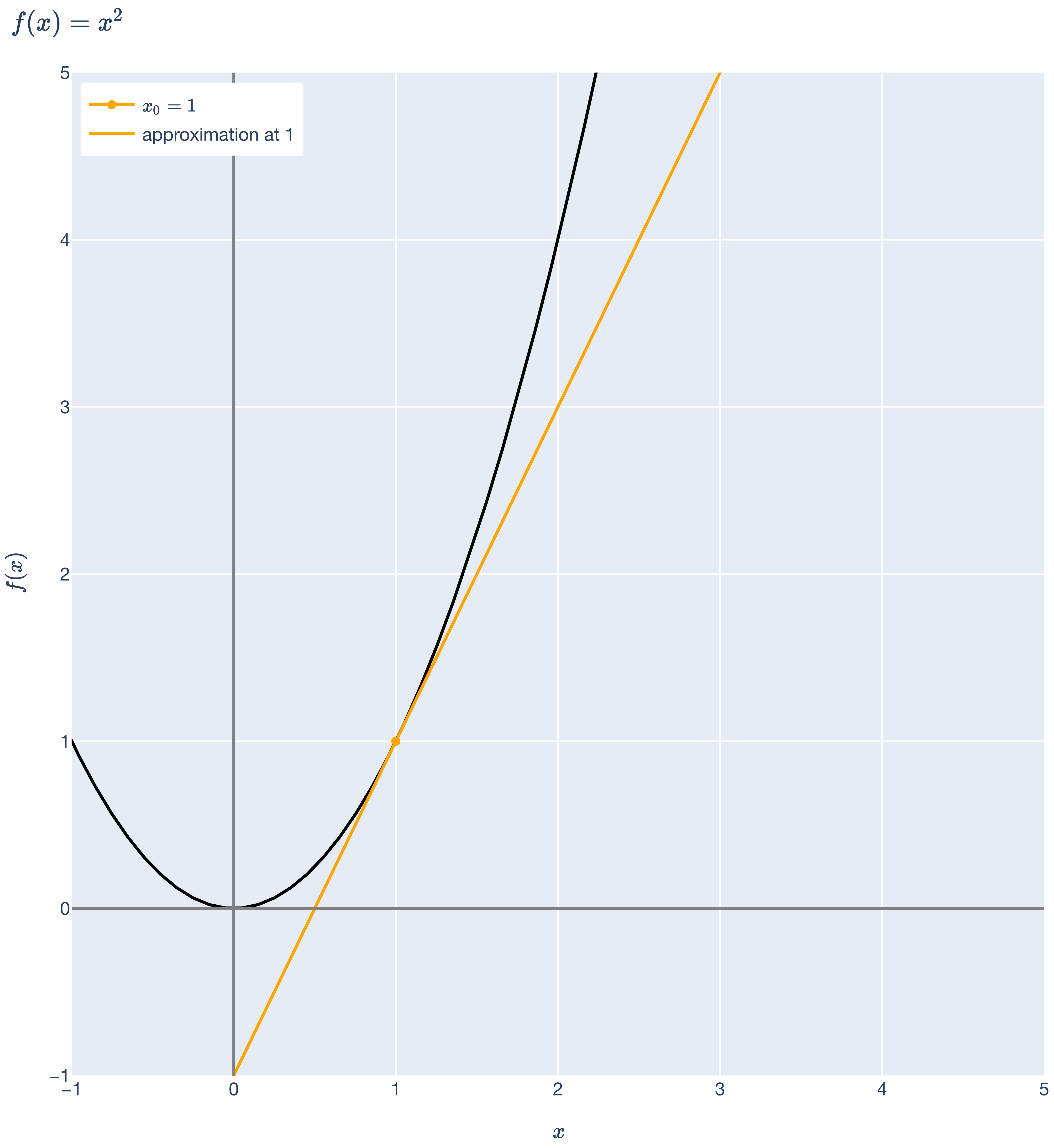
As $\vec{\delta}$ gets smaller (i.e. $\vec{\delta} \to \mathbf{0}$), there is smaller and smaller difference between the "change in f" and the "linear approximation of the change."

Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

What is the linearization?



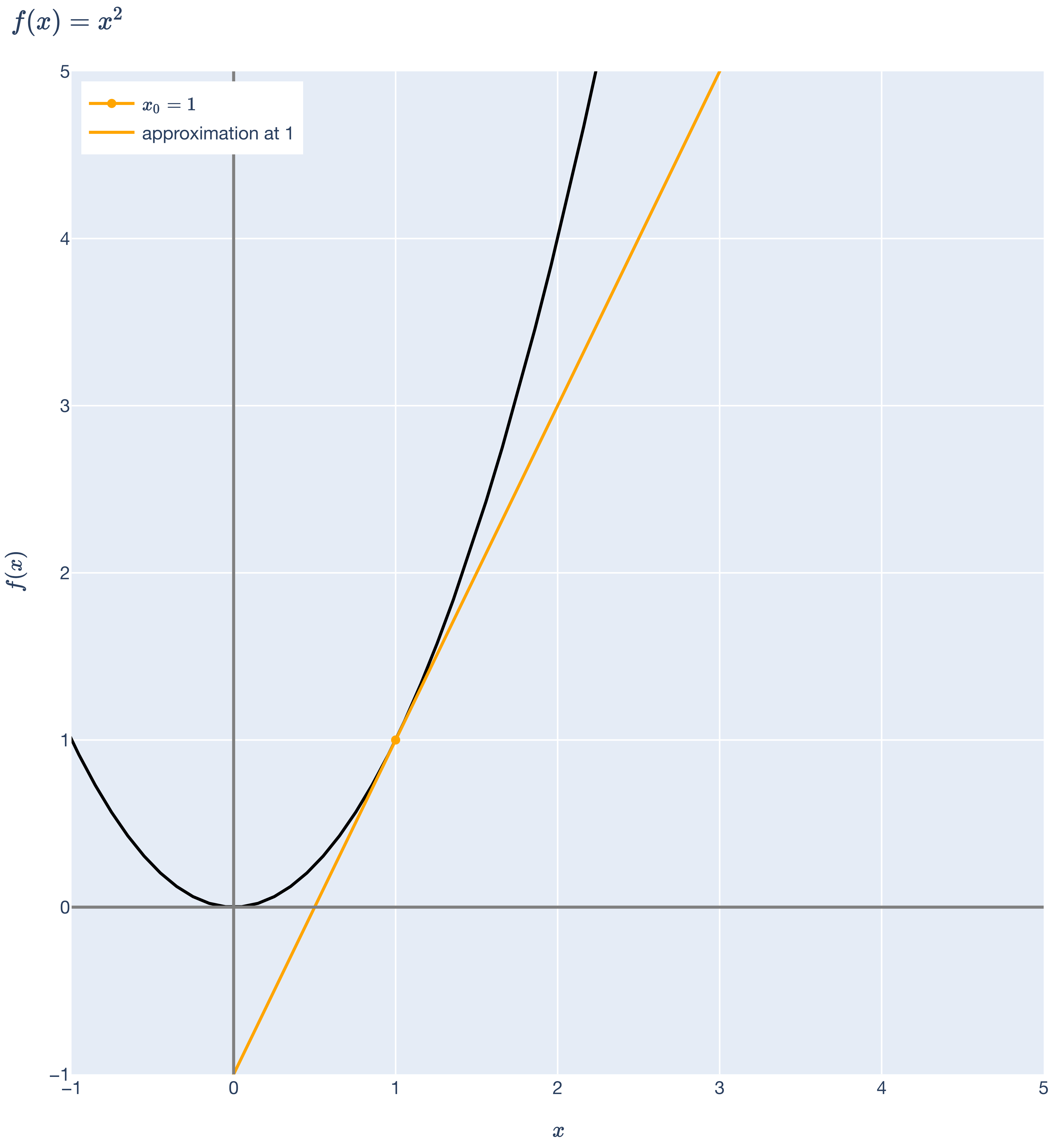
Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

What is the linearization?

$$f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0)$$



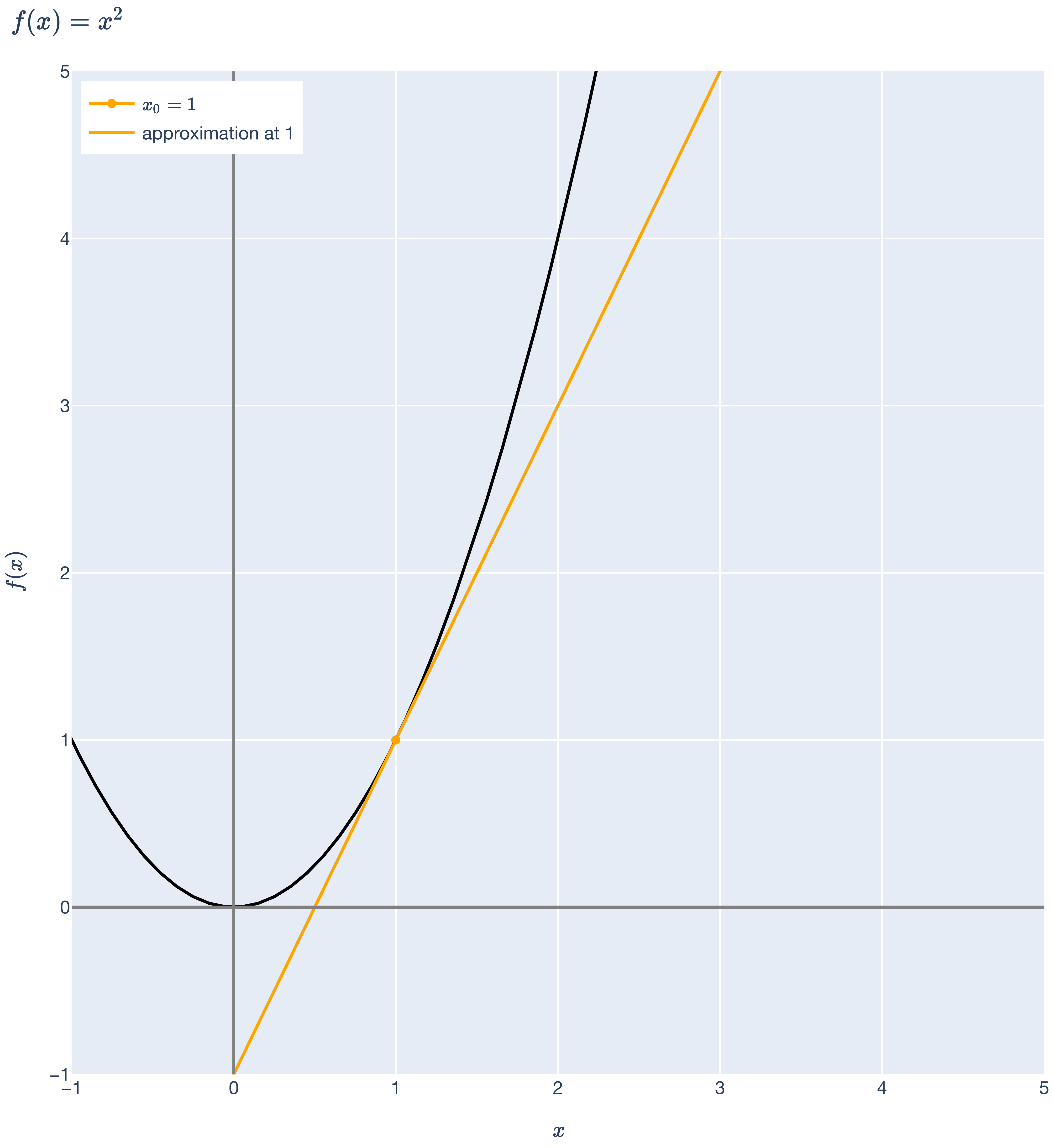
Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

What is the linearization?

$$f(x) \approx 1 + 2(x - 1)$$



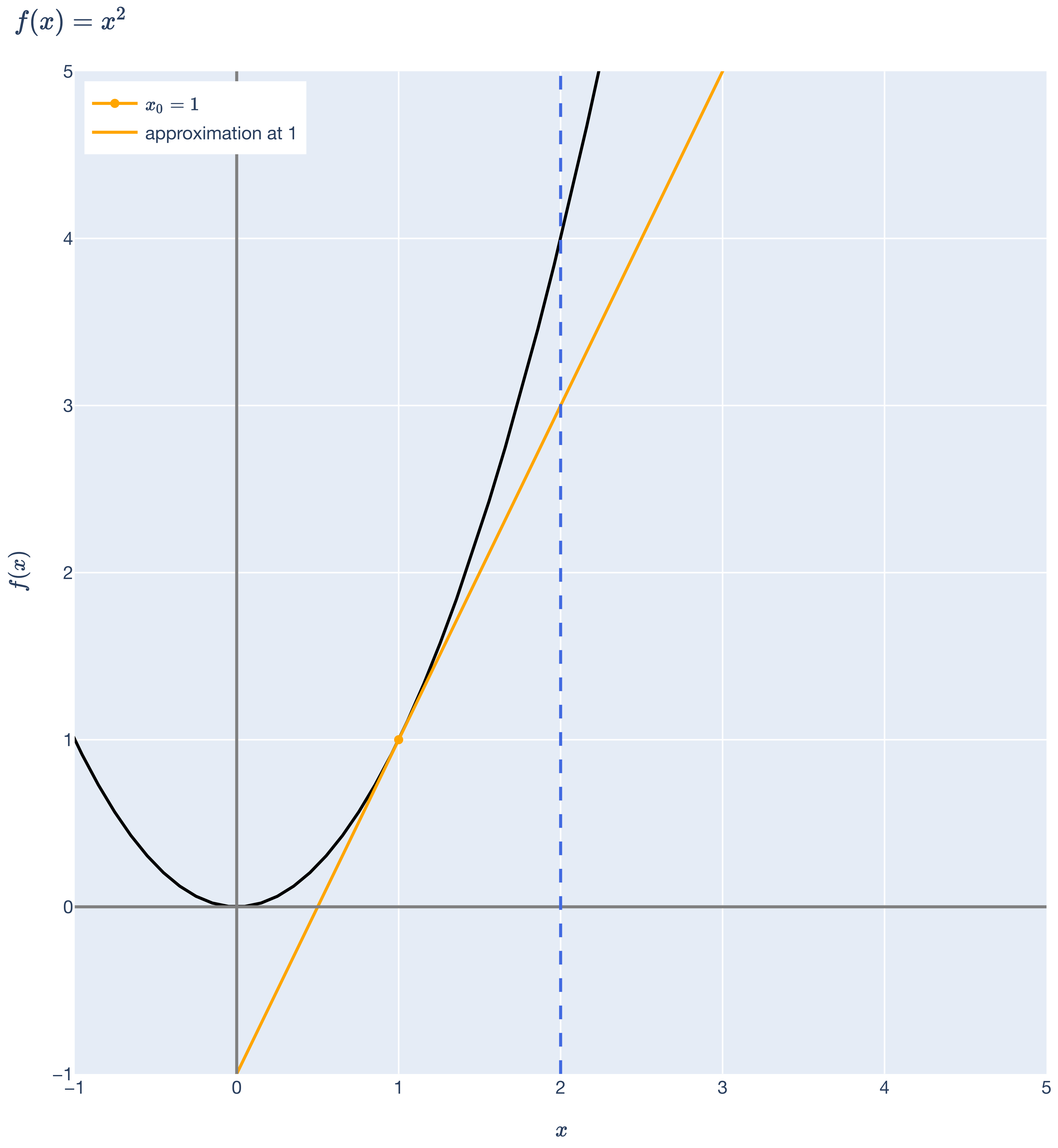
Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

Linearization: $f(x) \approx 1 + 2(x - 1)$

How good is the approximation at $x = 2$?



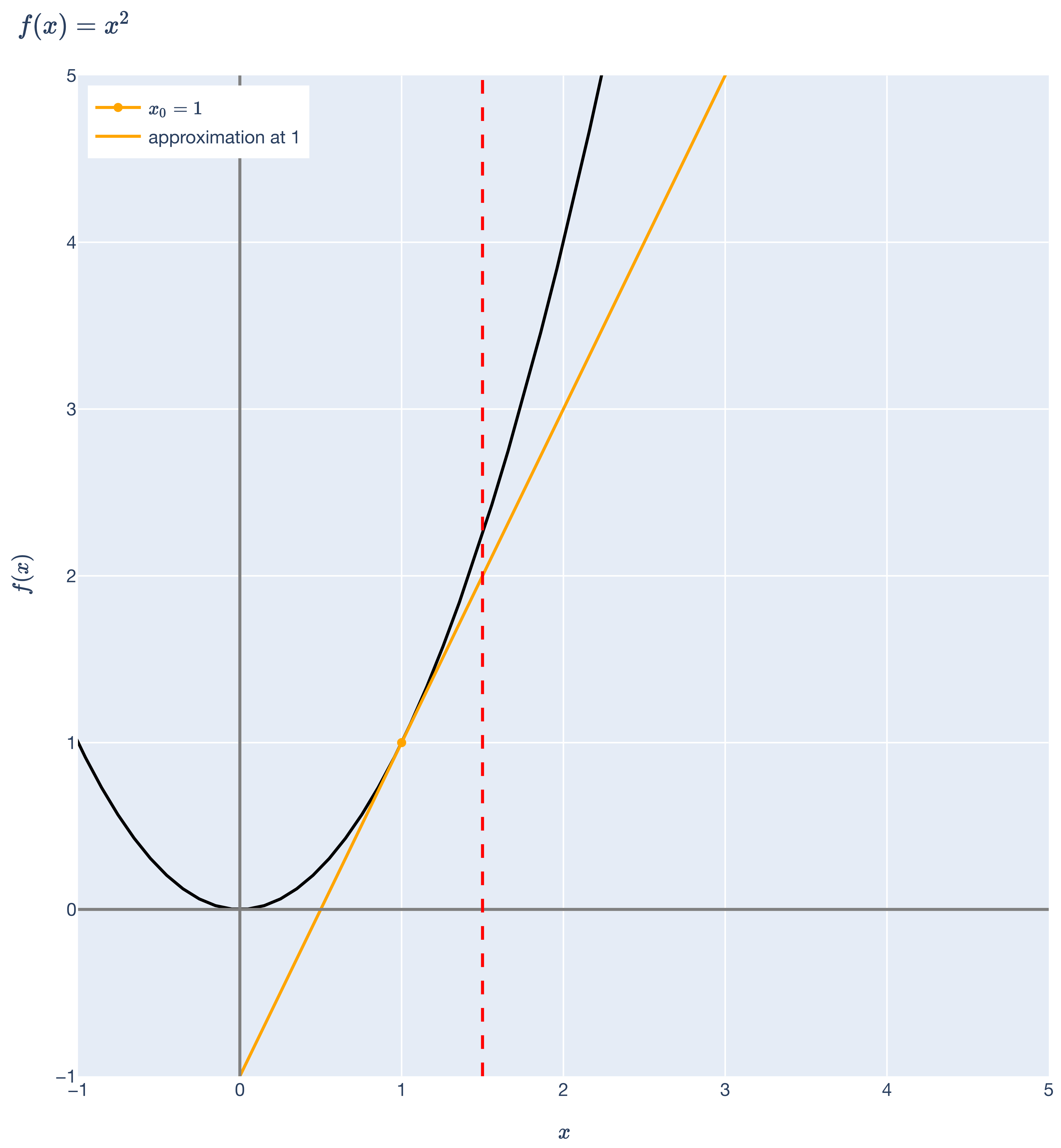
Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

Linearization: $f(x) \approx 1 + 2(x - 1)$

How good is the approximation at $x = 1.5$?



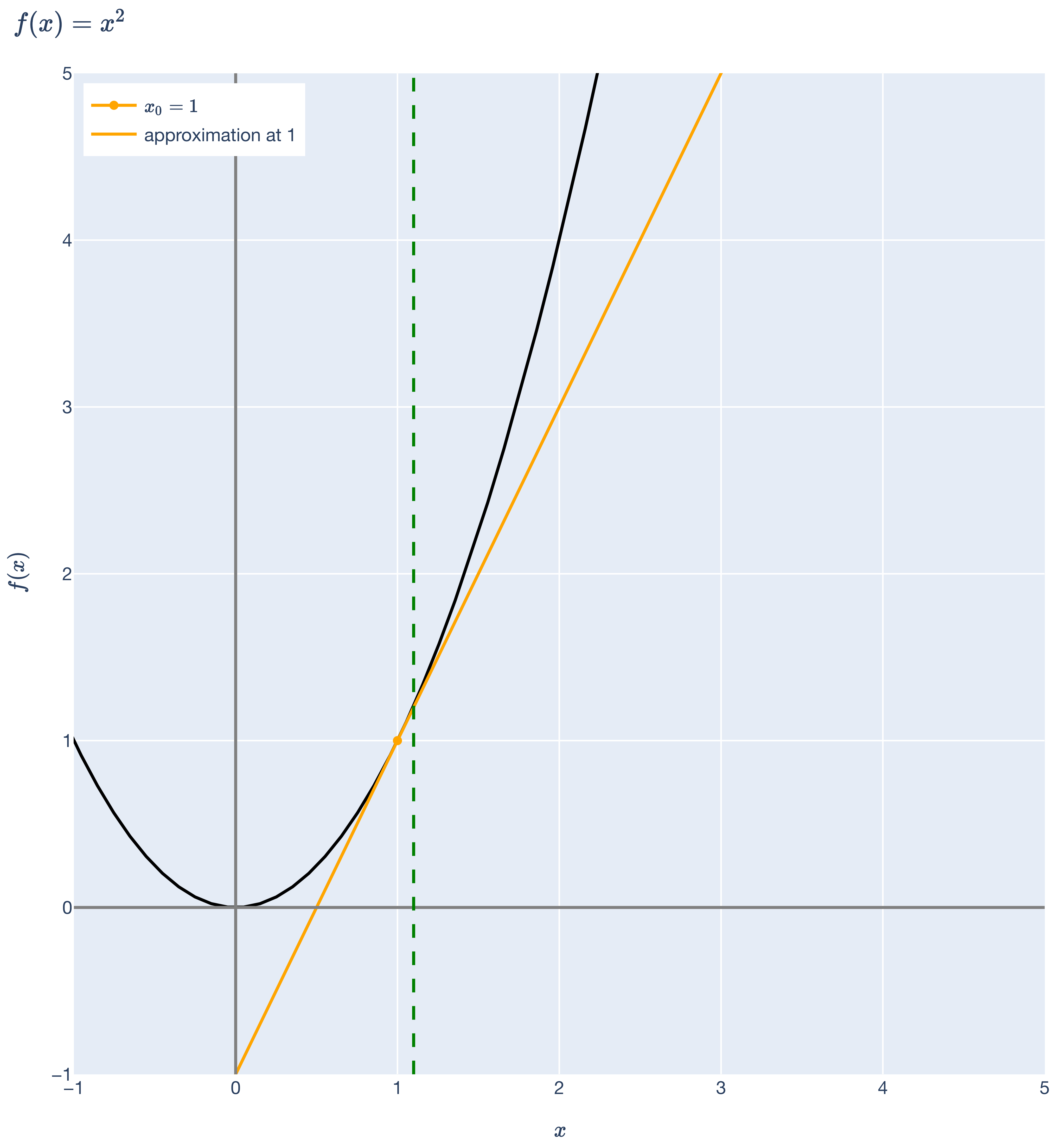
Linearization

$f : \mathbb{R} \to \mathbb{R}$ example

$$f(x) = x^2 \text{ with } x_0 = 1$$

Linearization: $f(x) \approx 1 + 2(x - 1)$

How good is the approximation at $x = 1.1$?

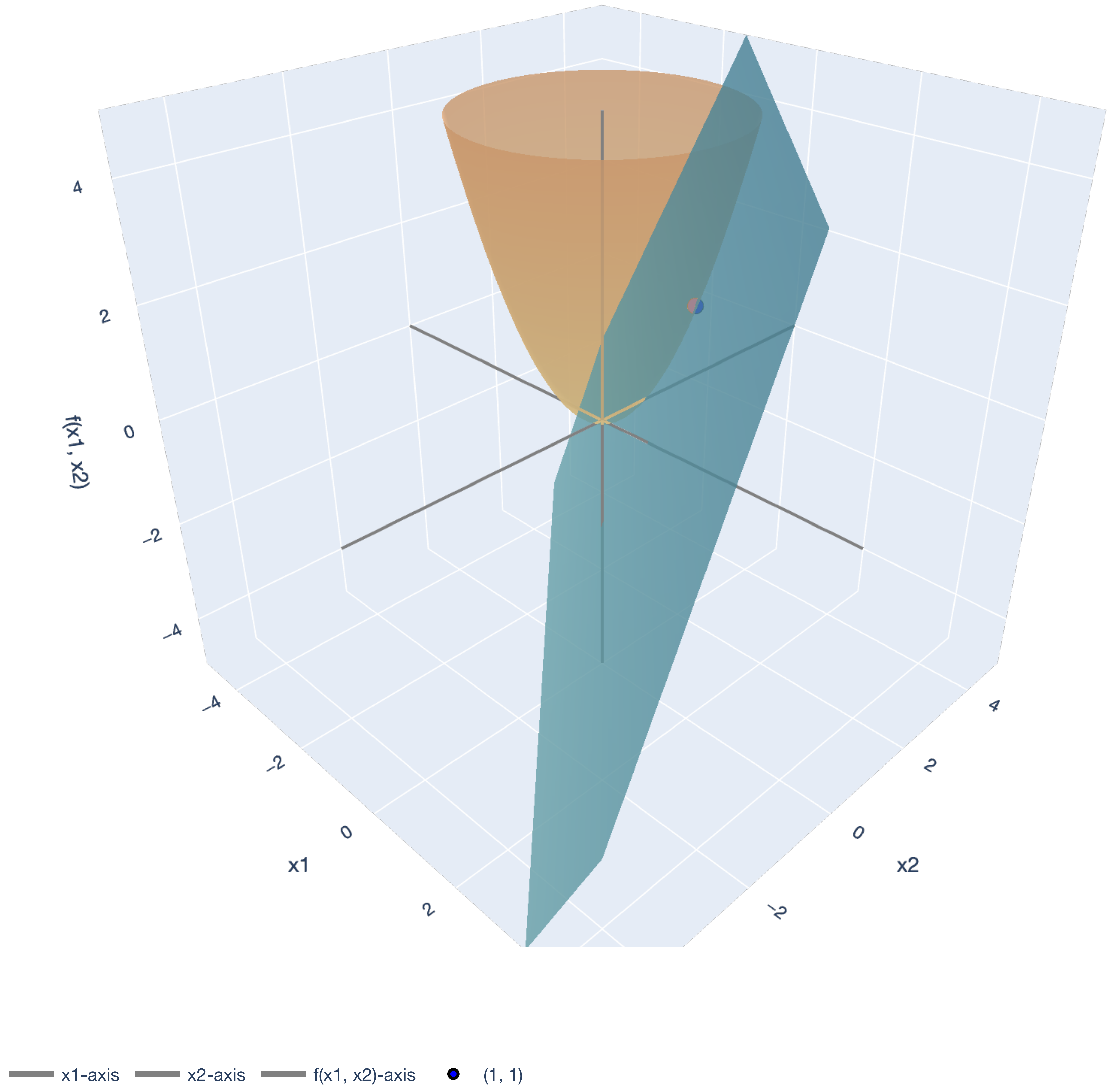


Linearization

$f : \mathbb{R}^2 \to \mathbb{R}$ **example**

$$f(x_1, x_2) = x_1^2 + x_2^2 \text{ with } \mathbf{x}_0 = (1,1)$$

What is the linearization?

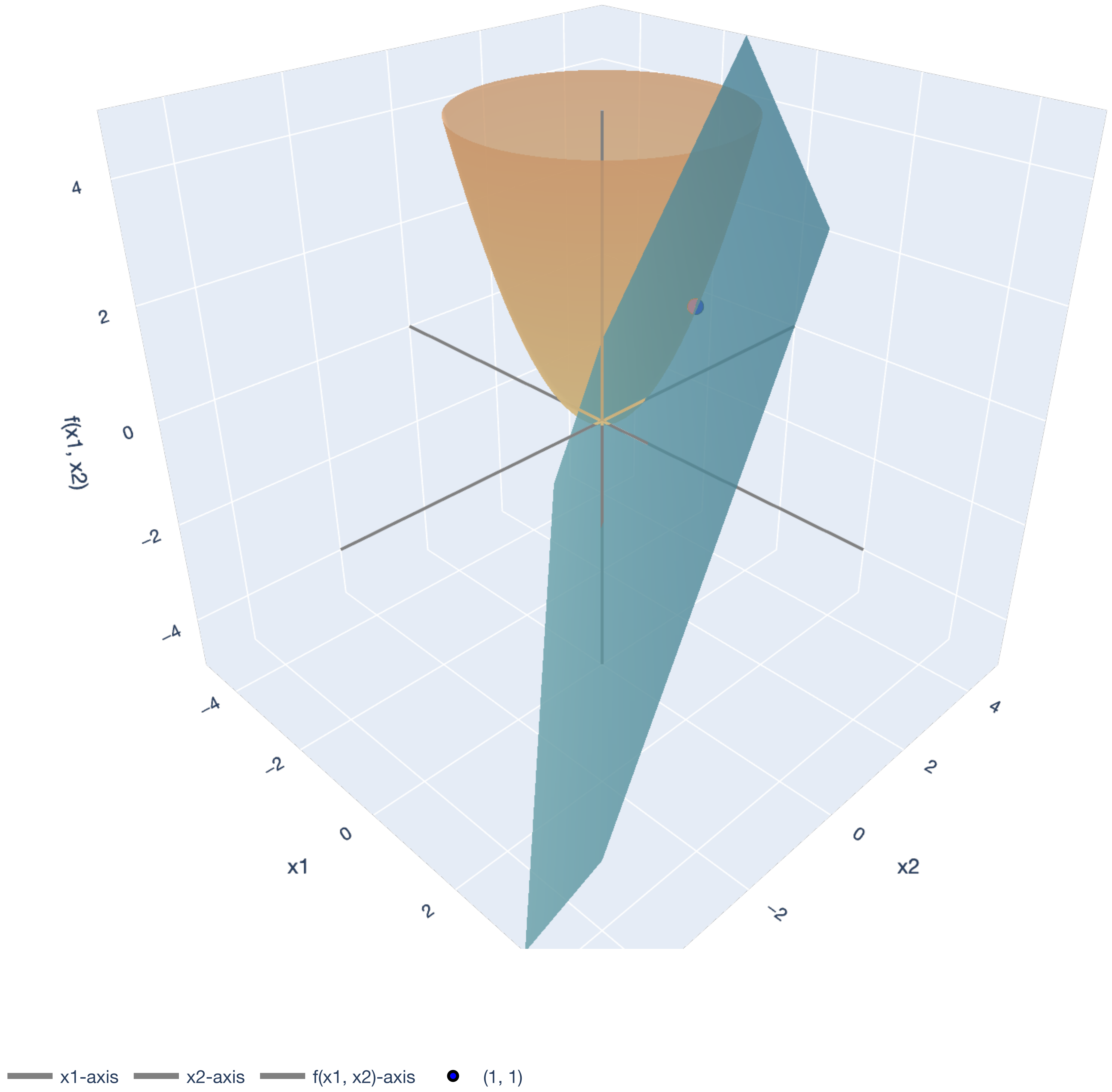


Linearization

$f : \mathbb{R}^2 \to \mathbb{R}$ example

$$f(x_1, x_2) = x_1^2 + x_2^2 \text{ with } \mathbf{x}_0 = (1,1)$$

Linearization: $f(x_1, x_2) \approx 2x_1 + 2x_2 - 2$



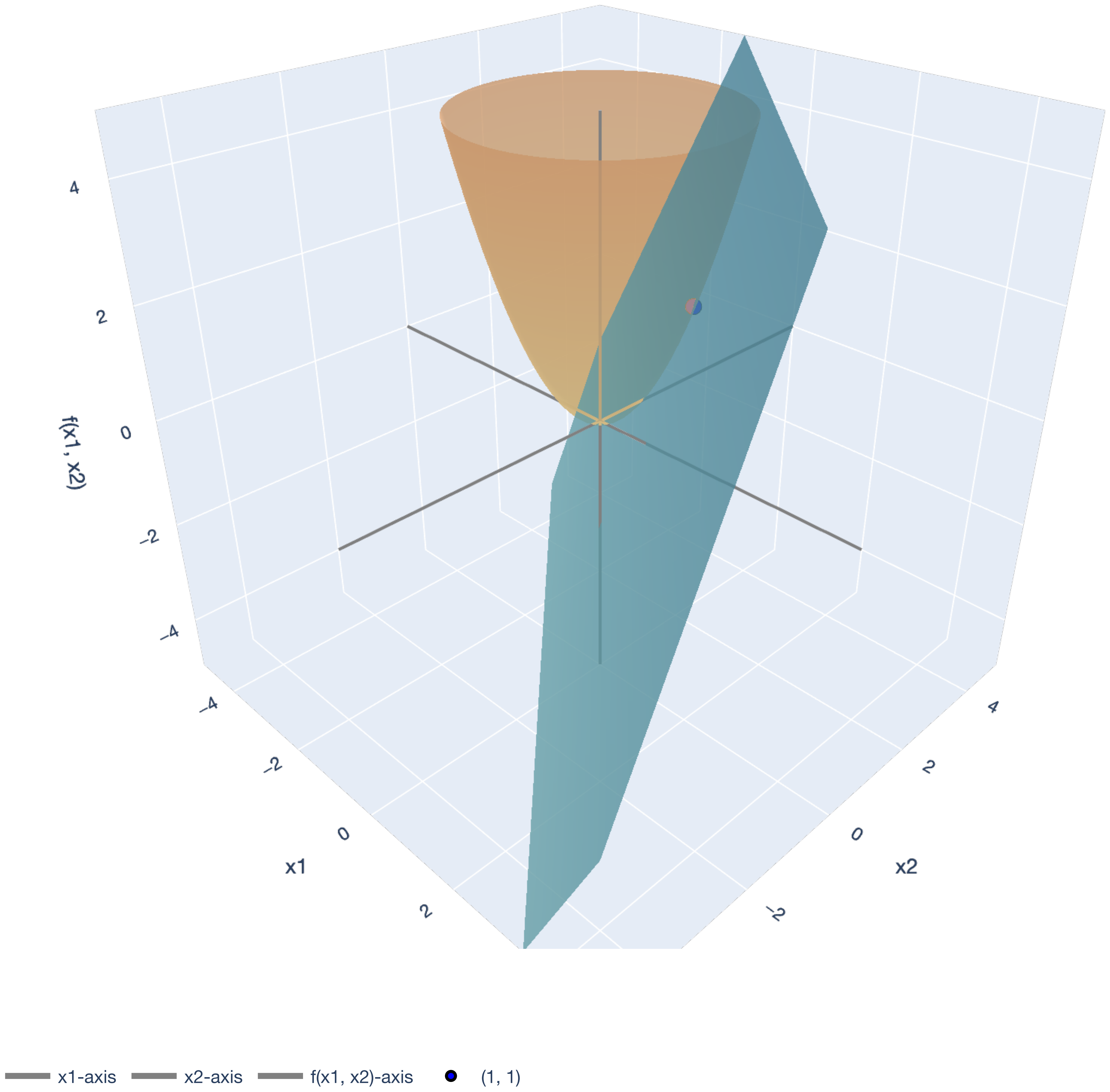
Linearization

$f : \mathbb{R}^2 \to \mathbb{R}$ example

$$f(x_1, x_2) = x_1^2 + x_2^2 \text{ with } \mathbf{x}_0 = (1,1)$$

Linearization: $f(x_1, x_2) \approx 2x_1 + 2x_2 - 2$

How good is the approximation at $\mathbf{x} = (0,1)$?



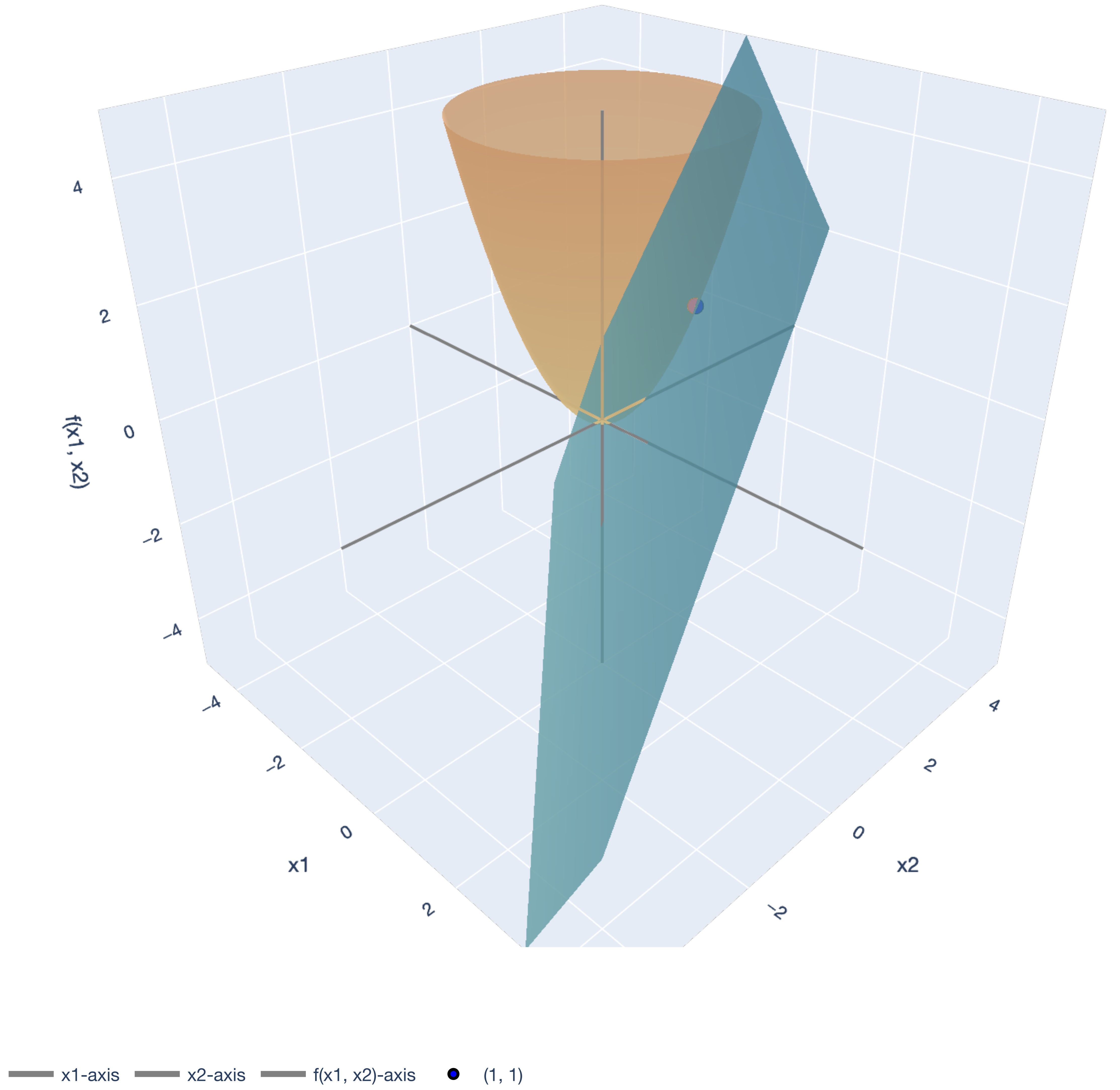
Linearization

$f: \mathbb{R}^2 \to \mathbb{R}$ example

$$f(x_1, x_2) = x_1^2 + x_2^2 \text{ with } \mathbf{x}_0 = (1,1)$$

Linearization: $f(x_1, x_2) \approx 2x_1 + 2x_2 - 2$

How good is the approximation at $\mathbf{x} = (1,0)$?



Taylor Series

In one variable

$\mathscr{C}^p$ functions and "smoothness"

Review of smooth functions

Smooth functions are functions that have (several) continuous derivatives.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is ***continuously differentiable*** if all of the partial derivatives of f exist and are continuous. We call such functions $\mathscr{C}^1$ *functions,* and the collection of all such functions are the class $\mathscr{C}^1$.

The class $\mathscr{C}^\infty$ are the ***infinitely differentiable*** functions — these have derivatives of *any* order.

$\mathscr{C}^p$ functions and "smoothness"

Review of smooth functions

Smooth functions are functions that have (several) continuous derivatives.

A function $f : \mathbb{R}^d \to \mathbb{R}$ is ***continuously differentiable*** if all of the partial derivatives of f exist and are continuous. We call such functions $\mathscr{C}^1$ *functions,* and the collection of all such functions are the class $\mathscr{C}^1$.

The class $\mathscr{C}^\infty$ are the ***infinitely differentiable*** functions — these have derivatives of *any* order.

"Smooth" varies from problem to problem. It usually denotes a function being "sufficiently differentiable."

$\mathscr{C}^p$ functions and "smoothness"

Review of smooth functions

Example. $f(x) = e^x$.

$\mathscr{C}^p$ functions and "smoothness"

Review of smooth functions

Example. $f(x) = \sin x$.

$\mathscr{C}^p$ functions and "smoothness"

Review of smooth functions

Example. $f(x_1, x_2) = x_1^2 + x_2^2$. Polynomials, in general.

Polynomials

Single-variable definition

A single-variable ***polynomial function*** of degree m is a function $f : \mathbb{R} \to \mathbb{R}$ that can be written in the form:

$$a_m x^m + a_{m-1} x^{m-1} + \ldots + a_2 x^2 + a_1 x + a_0,$$

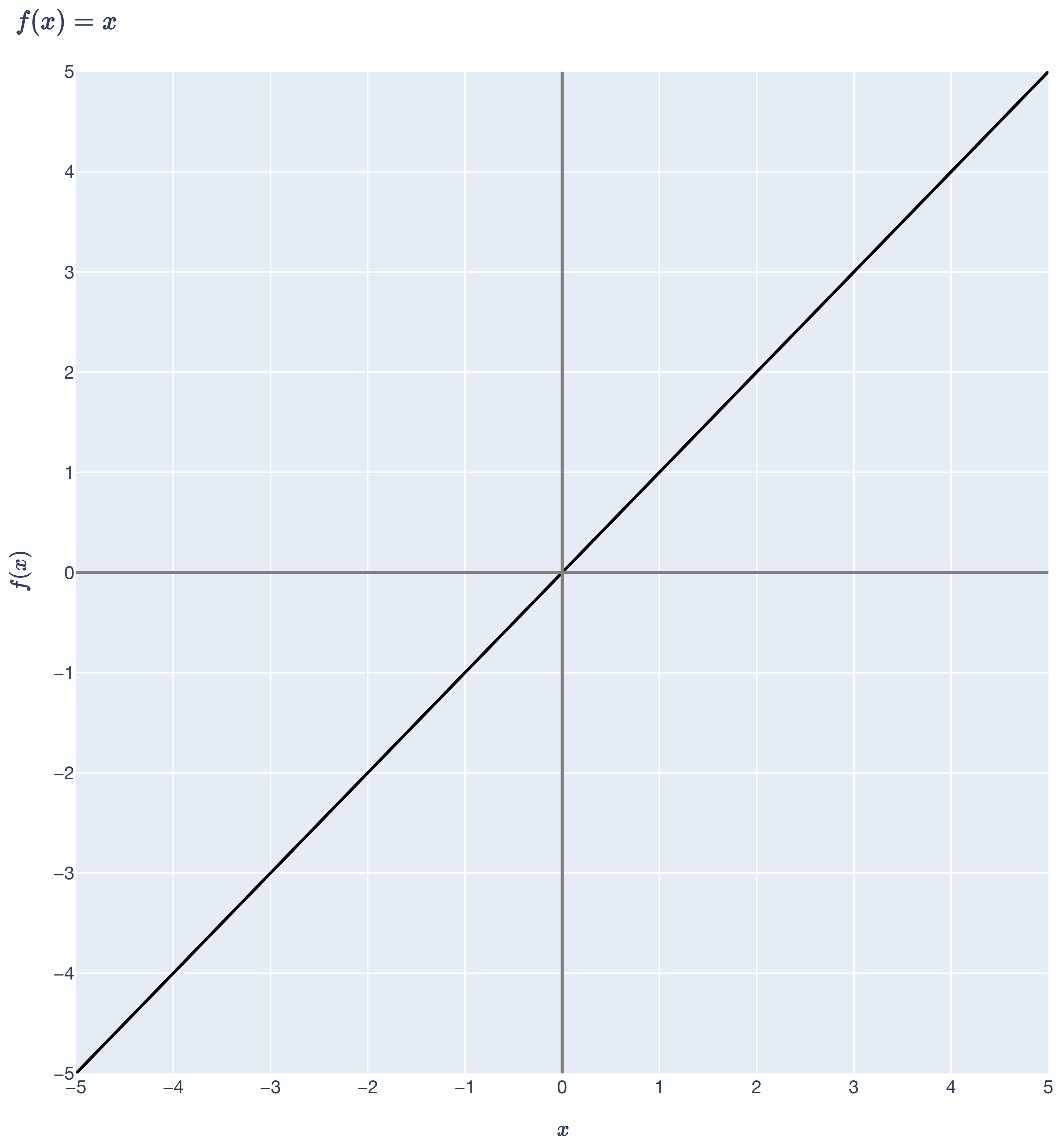
where $a_m, \ldots, a_0 \in \mathbb{R}$ are the *coefficients* of the polynomial.

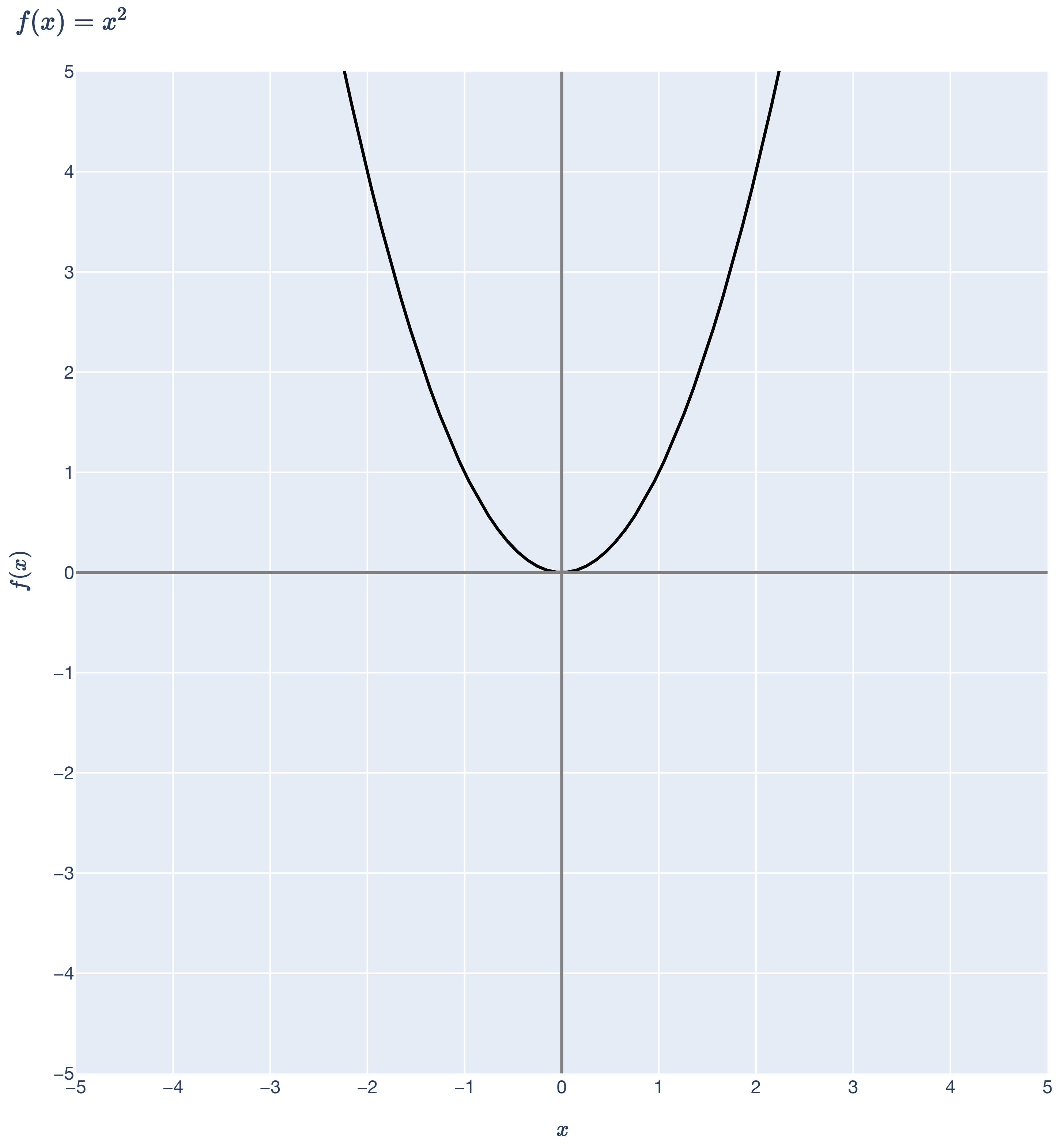
Example: $f(x) = 4x^3 + 2x - 1.$

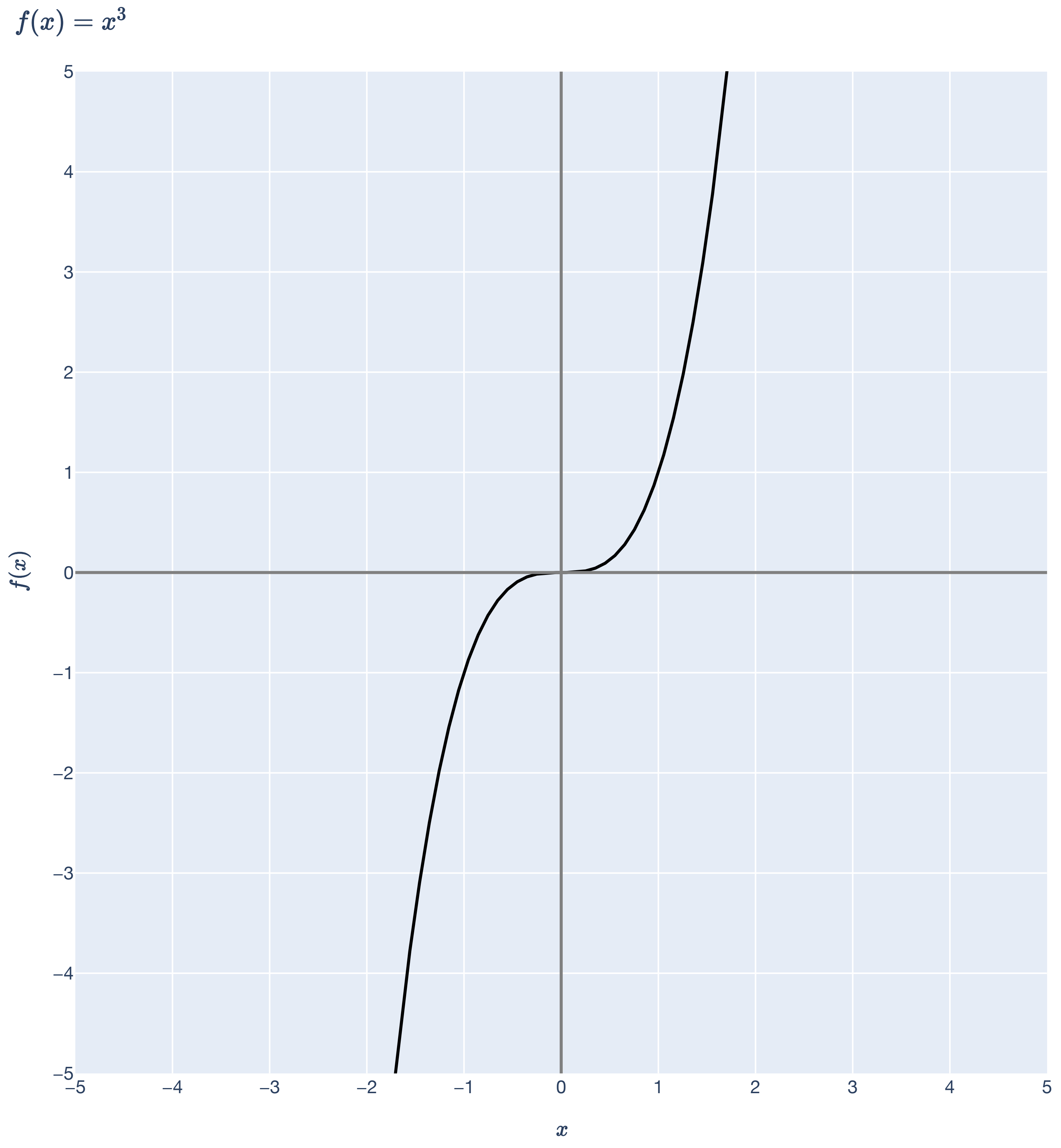
Polynomials

Single-variable definition

$$f(x) = x \qquad f(x) = x^2 \qquad f(x) = x^3$$







Polynomials

Multivariable definition

A ***monomial function*** is a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$x_1^{k_1} \dots x_d^{k_d} \text{ with integer exponents } k_1, \dots, k_d \geq 0.$$

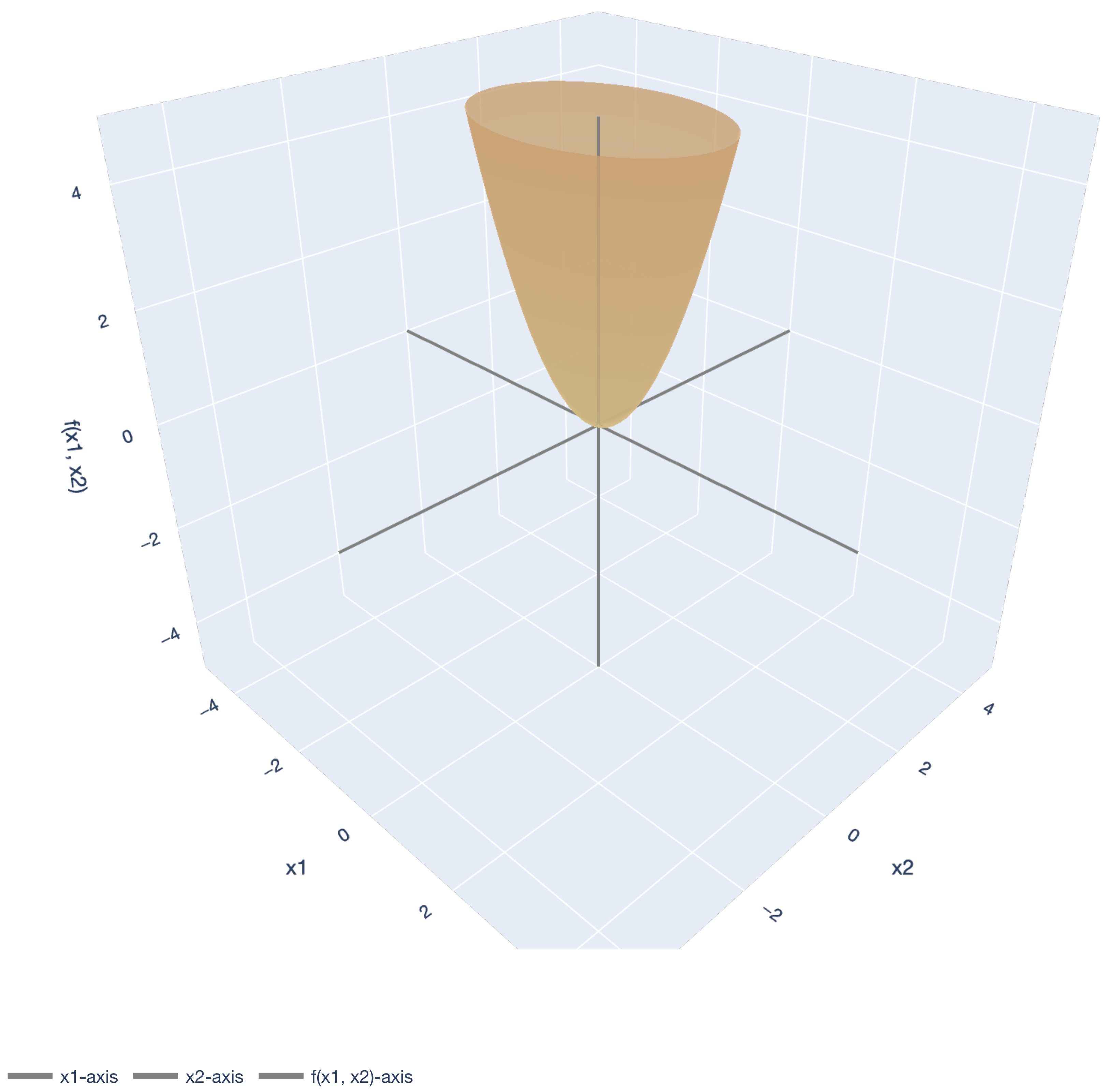
A ***polynomial function*** is a function $f : \mathbb{R}^d \to \mathbb{R}$ is a finite sum of monomials with real coefficients.

Example: $f(x_1, x_2, x_3) := x_1^2 x_2 + 3x_1 x_3$.

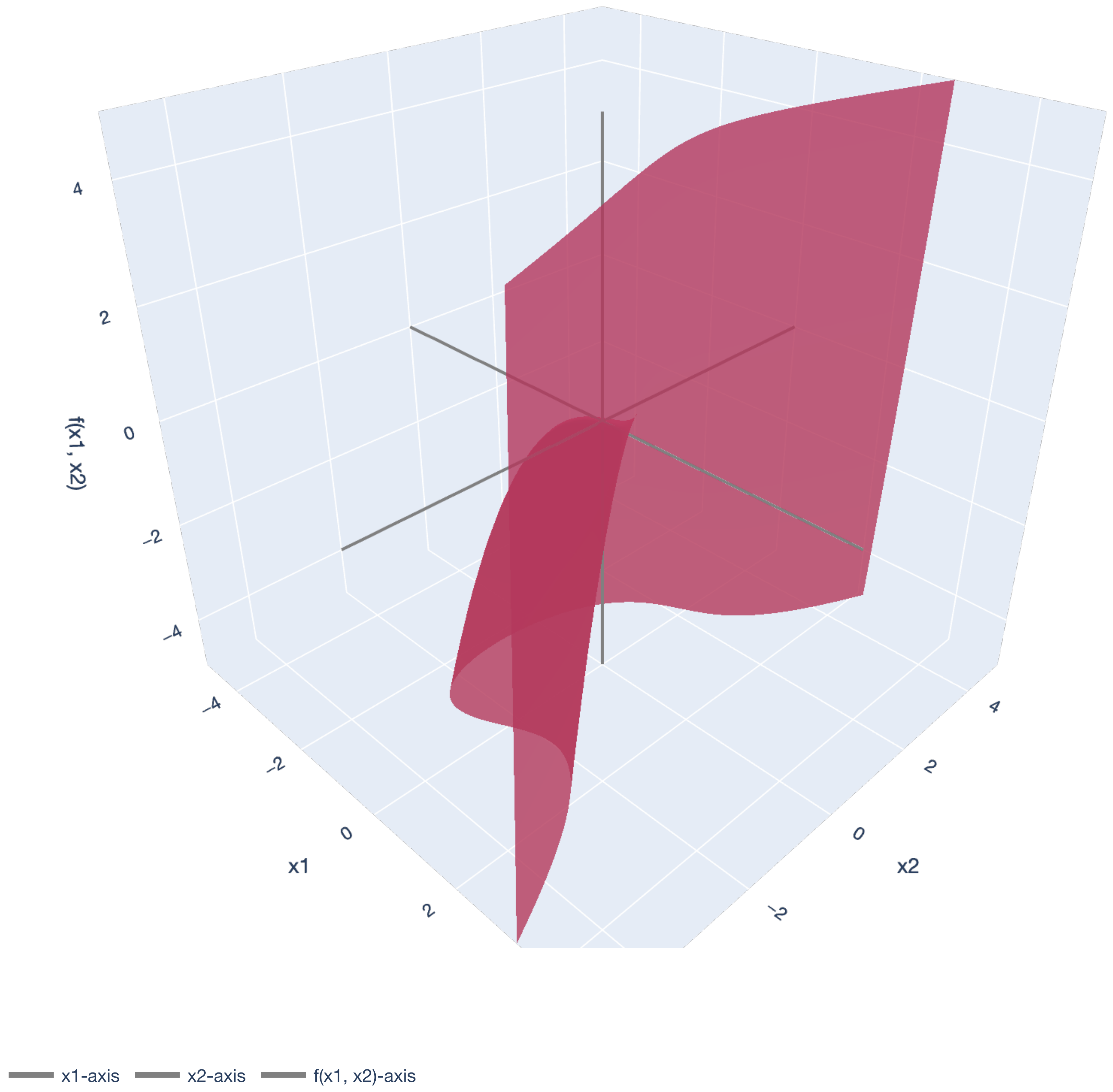
Polynomials

Multi-variable definition

$$f(x_1, x_2) = x_1^2 + 2x_2^2$$



$$f(x_1, x_2) = x_1^3 + x_1 x_2 - x_2^2$$



Taylor Series
Intuition

We like *polynomials* — they're easy to perform calculus on and analyze.

$$f(x) = x^5 + 3x^3 - 2x^2 + 3x - 1$$

A ***Taylor series*** at some point x_0 is the representation of "smooth" functions as an "infinite polynomial," expanded around x_0.

Canonical example (at $x_0 = 0$):

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

Taylor Series

Intuition

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

"Cutting off" the Taylor series at some order p of derivatives gives us the ***pth-order Taylor approximation.***

The first-order Taylor approximation is just the *linearization*!

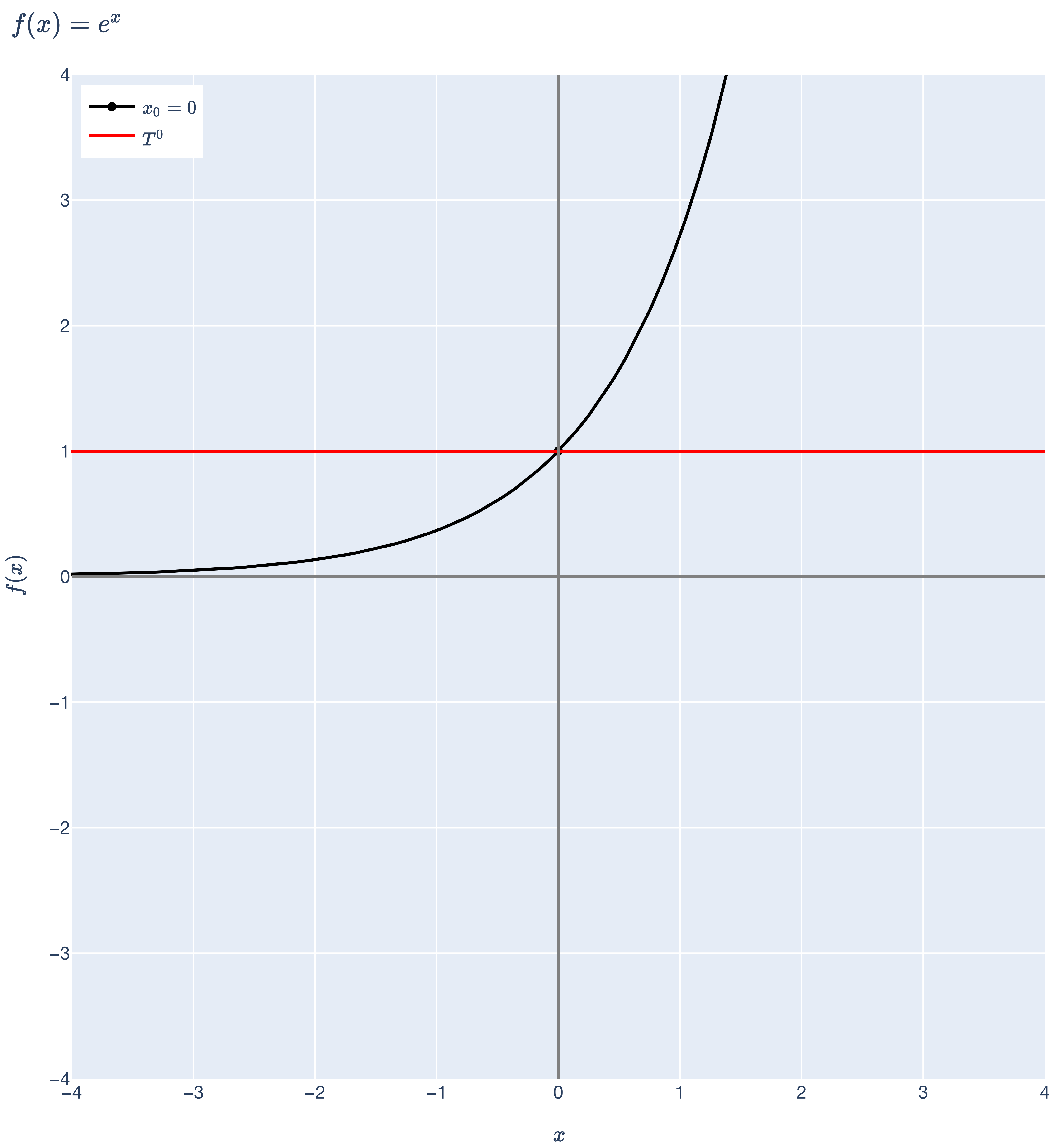
The second-order Taylor approximation is just a quadratic function!

Taylor Series

Example: $f(x) = e^x$

Taylor series at $x_0 = 0$:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots$$

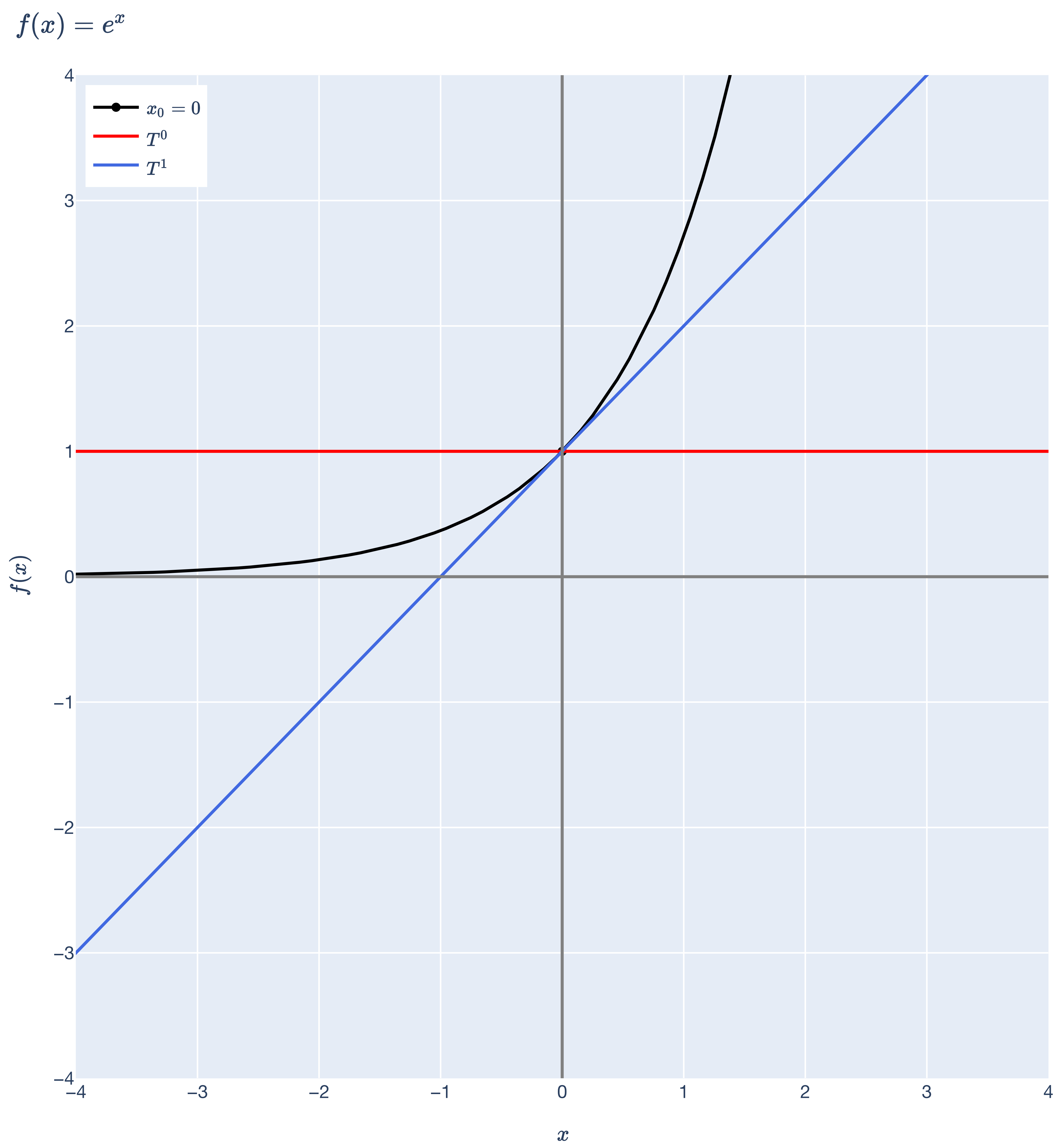


Taylor Series

Example: $f(x) = e^x$

Taylor series at $x_0 = 0$:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots$$

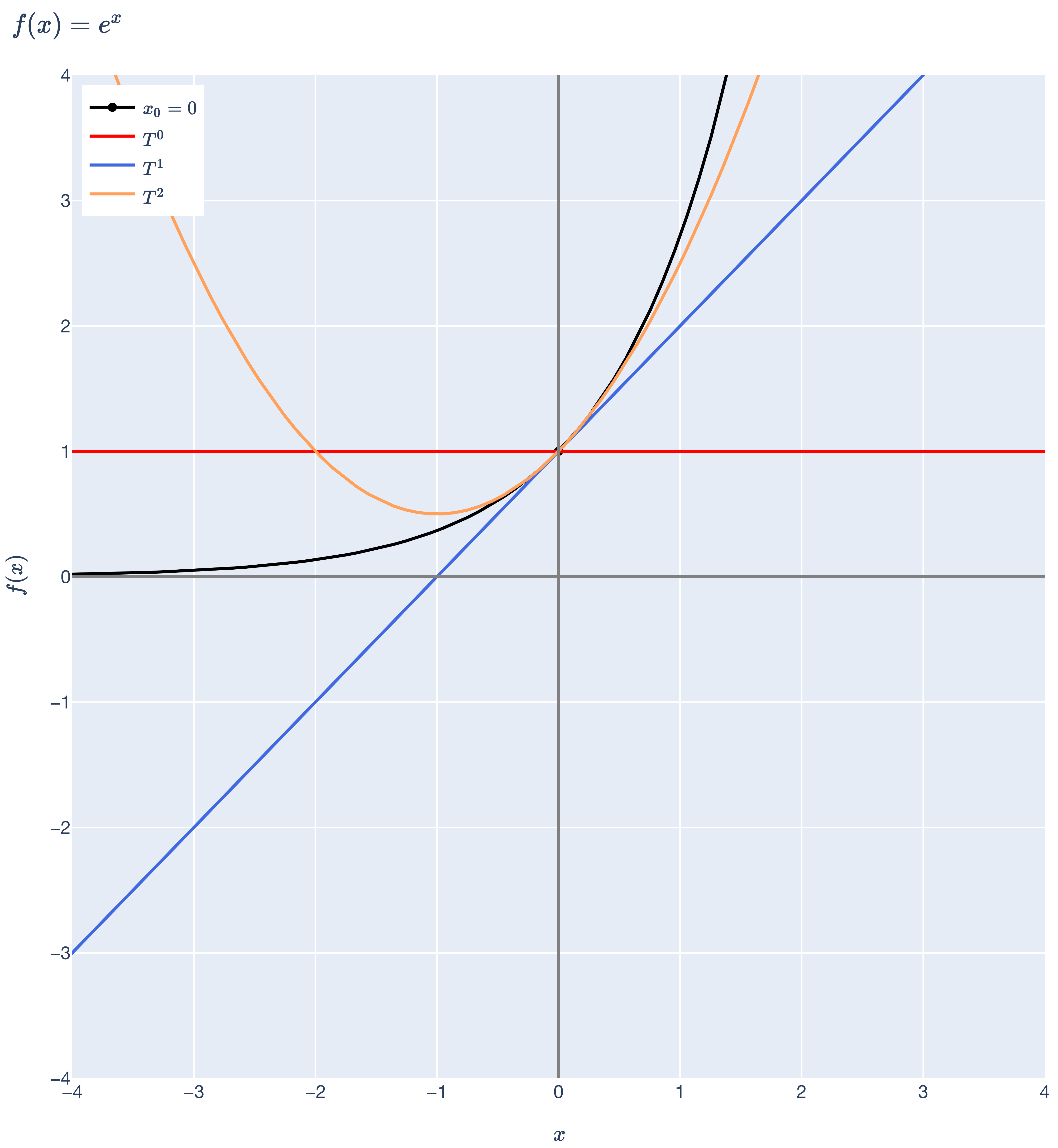


Taylor Series

Example: $f(x) = e^x$

Taylor series at $x_0 = 0$:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

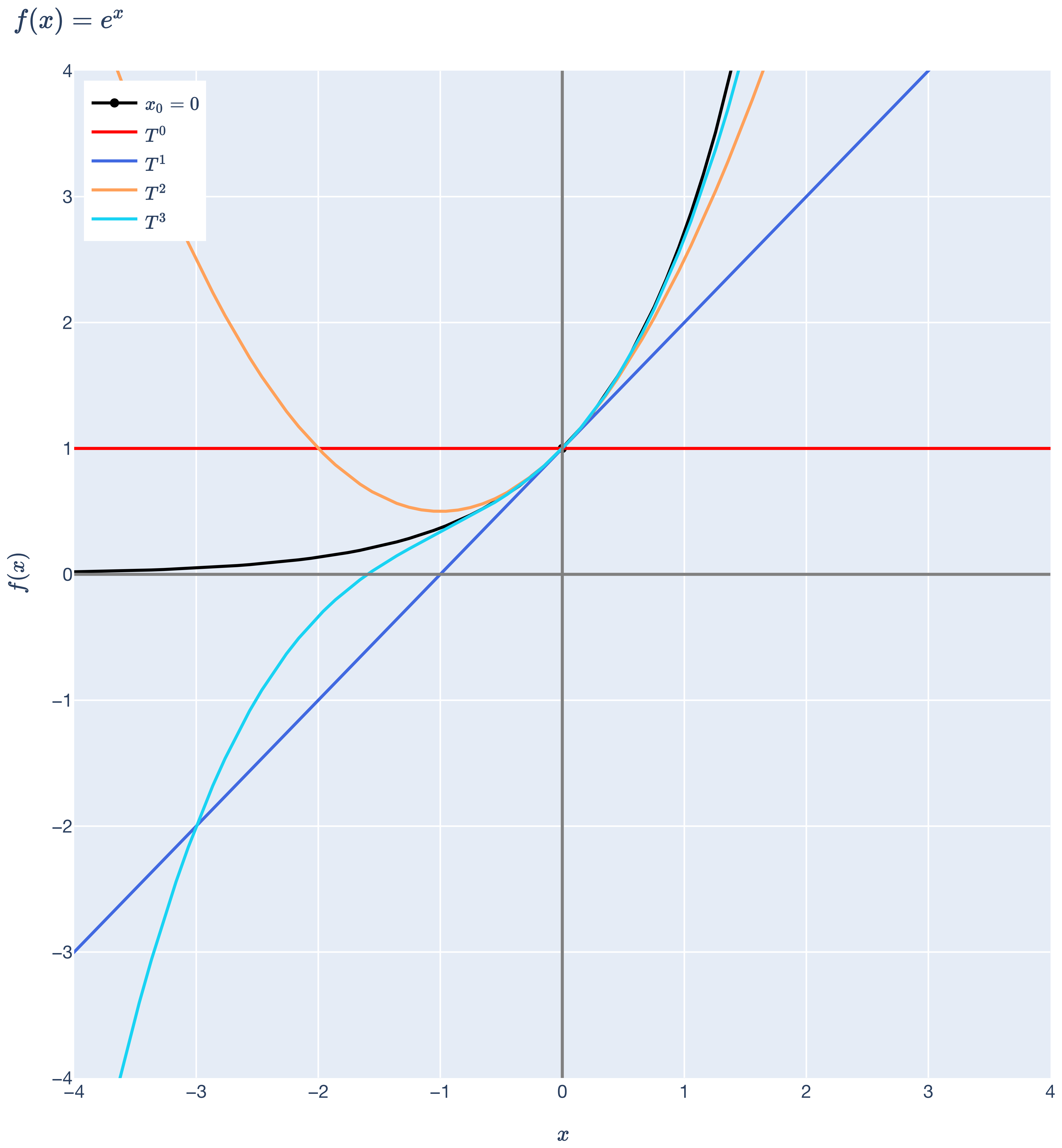


Taylor Series

Example: $f(x) = e^x$

Taylor series at $x_0 = 0$:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

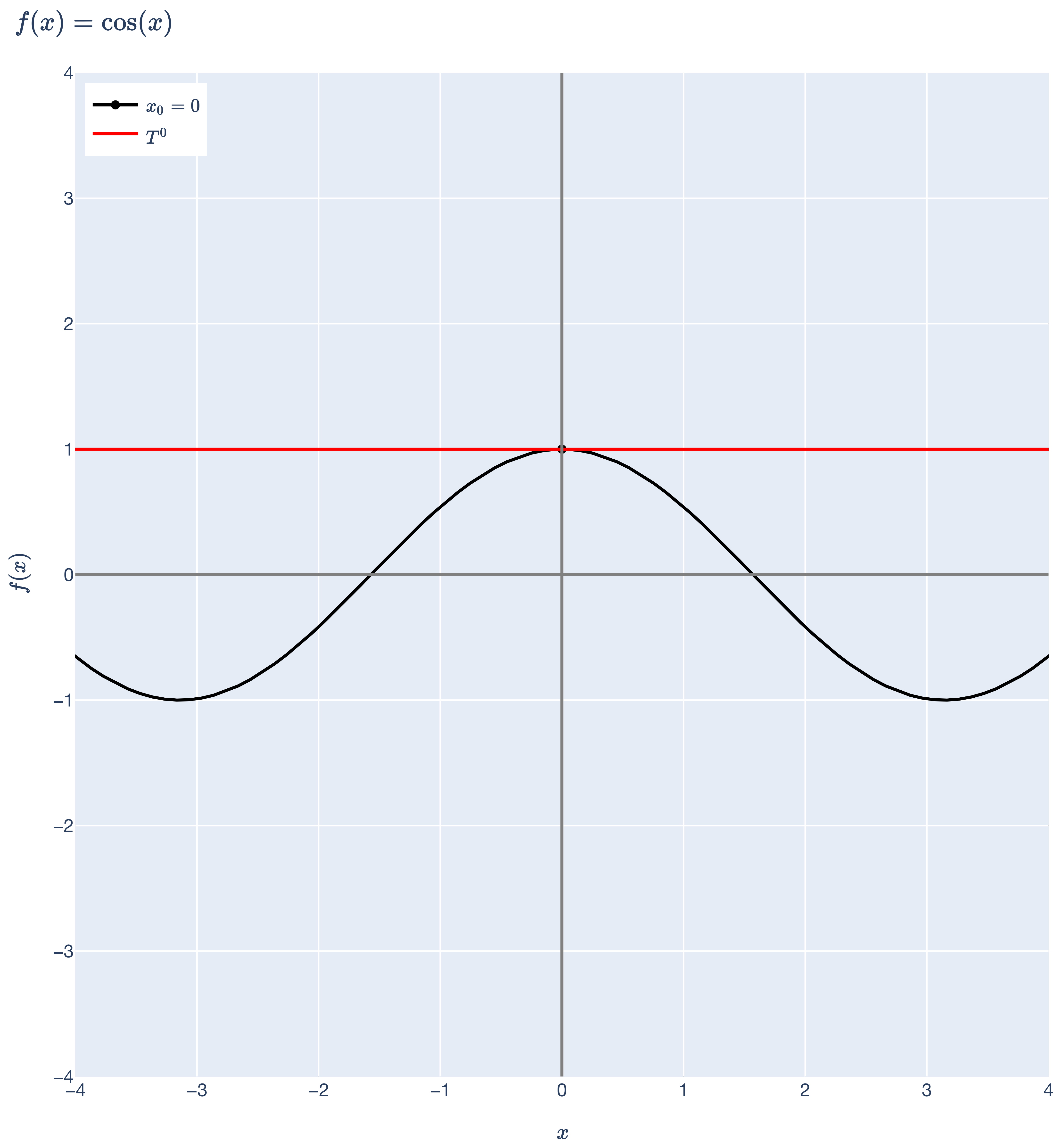


Taylor Series

Example: $f(x) = \cos x$

Taylor series at $x_0 = 0$:

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots$$

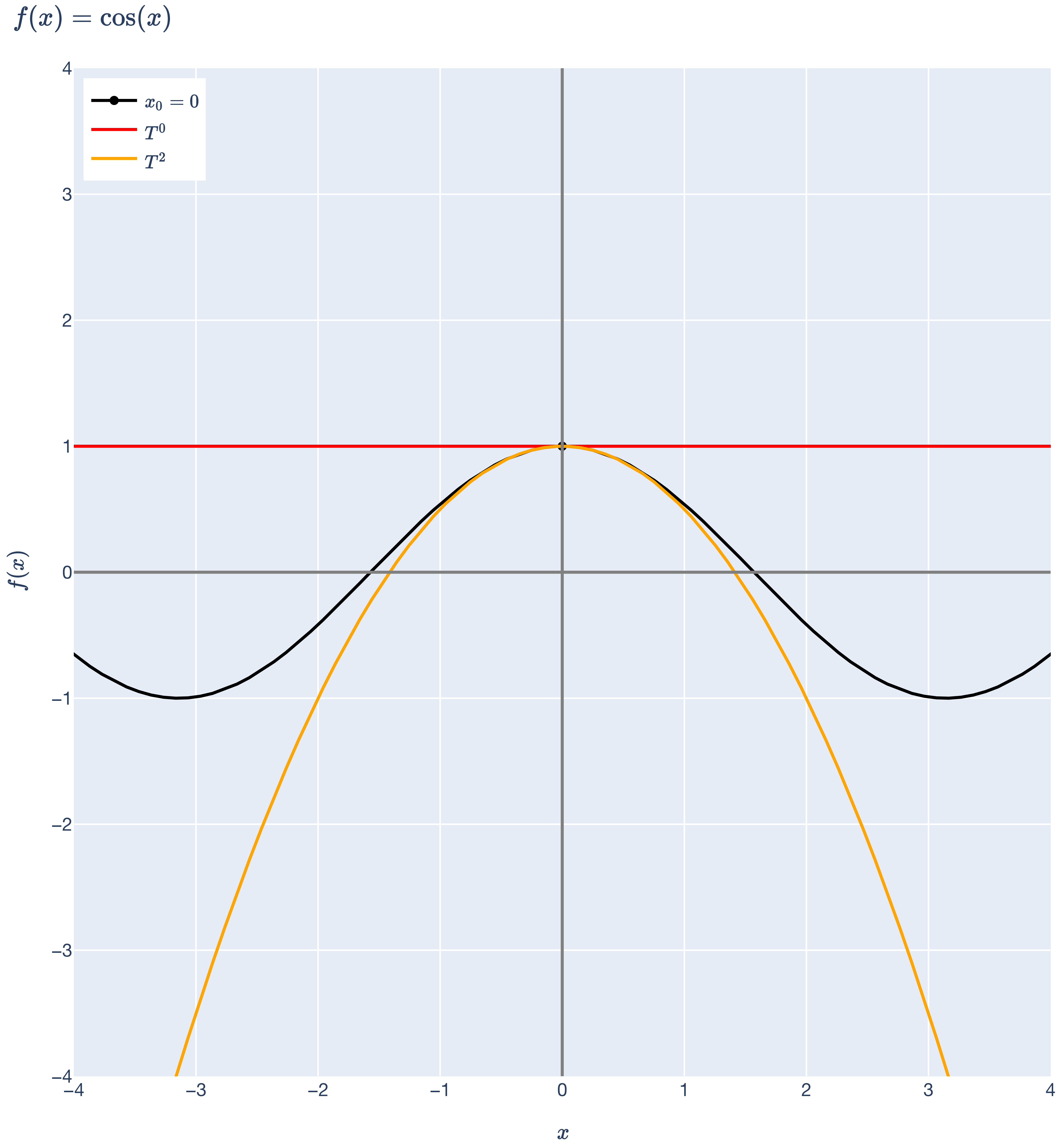


Taylor Series

Example: $f(x) = \cos x$

Taylor series at $x_0 = 0$:

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \ldots$$

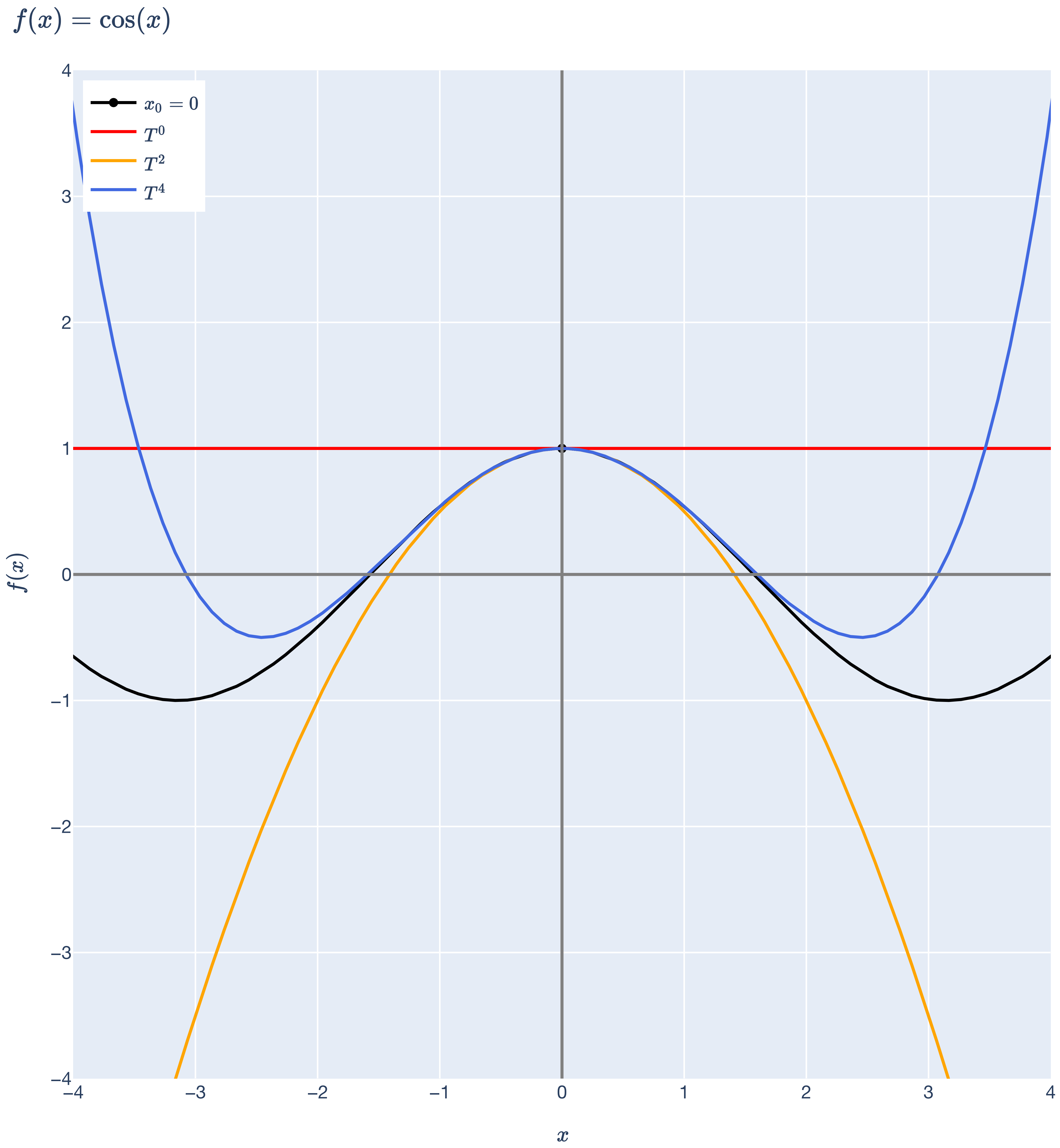


Taylor Series

Example: $f(x) = \cos x$

Taylor series at $x_0 = 0$:

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \ldots$$



Taylor Series

Single-variable definition

For simplicity, let's first consider $f : \mathbb{R} \to \mathbb{R}$.

For a smooth function $f \in \mathscr{C}^\infty$ (f has derivatives of all orders), the ***Taylor series of f at x_0*** is defined as:

$$T_{x_0}(x) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k .$$

The ***Taylor polynomial of degree n*** of f at x_0 is defined as:

$$T^n_{x_0}(x) := \sum_{k=0}^{n} \frac{f^{(k)(x_0)}}{k!}(x - x_0)^k .$$

Note: It only make sense to talk about a Taylor series/polynomial *at a point*!

Taylor Series

When is the Taylor series the function?

A function that is equal to its Taylor series at x_0 in some neighborhood around x_0 is called ***analytic***. *We won't get into the finer points of Taylor series and analytic functions in this course.*

For all intents and purposes,

$$f(x) \approx T^n_{x_0}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k = \underbrace{f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2}_{\text{usually already pretty good!}} + \dots$$

for all x that are sufficiently close to x_0 and sufficiently large n (we'll usually study $n \leq 2$).

Taylor Series

When is the Taylor series the function?

A function that is equal to its Taylor series at x_0 in some neighborhood around x_0 is called ***analytic***.

For all intents and purposes,

$$f(x) \approx T^n_{x_0}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k = \underbrace{f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2}_{\text{usually already pretty good!}} + \dots$$

for all x that are sufficiently close to x_0 and sufficiently large n (we'll usually study $n \leq 2$).

Takeaway. For many common functions, a second-order Taylor polynomial is a good approximation of the function close to the point we do the expansion about.

Taylor Series
Example

All polynomials are in $\mathscr{C}^\infty$ and have *exact* Taylor series representations.

Consider the Taylor series of $f(x) = 2x^3 + x^2 - x + 1$.

Taylor Series

Example

Many of the “nice” functions of calculus are infinitely differentiable.

Consider the Taylor series of $f(x) = \sin x + \cos x$.

Taylor Series

Example

Many of the “nice” functions of calculus are infinitely differentiable.

Consider the Taylor series of $f(x) = e^x$.

Taylor Series

In multiple variables

Taylor Series
Multivariable definition

There's a reason we started with $f : \mathbb{R} \to \mathbb{R}$...

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function with derivatives of all orders (i.e., in $\mathscr{C}^\infty$). The ***Taylor series of f at $\mathbf{x}_0 = (x_{01}, \ldots, x_{0n}) \in \mathbb{R}^n$*** is given by:

$$T(x_1, \ldots, x_n) := \sum_{k_1=0}^{\infty} \ldots \sum_{k_n=0}^{\infty} \frac{(x_1 - x_{01})^{k_1} \ldots (x_n - x_{0n})^{k_n}}{k_1! \ldots k_n!} \left(\frac{\partial^{k_1+\ldots+k_n} f}{\partial x_1^{k_1} \ldots \partial x_n^{k_n}} \right)(x_{01}, \ldots, x_{0n})\,.$$

Thankfully — we won't ever need to use this — at most, we'll use the *second-order Taylor approximation* of a function in multiple variables.

Hessian

The multivariable second derivative

The ***Hessian*** for $f : \mathbb{R}^2 \to \mathbb{R}$ at some point $\mathbf{x}_0$ is the 2×2 matrix of all second-order partial derivatives:

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

The Hessian for general $f : \mathbb{R}^n \to \mathbb{R}$ is given by the $n \times n$ matrix of all second-order partial derivatives, constructed similarly.

For twice-continuously differentiable $f \in \mathscr{C}^2$, the Hessian is symmetric.

Taylor Series

Just the second-order terms

For $f : \mathbb{R}^n \to \mathbb{R}$, the second-order terms of the Taylor series of f at $\mathbf{x}_0$ are:

$$T^2_{\mathbf{x}_0}(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

Taylor Series

Just the second-order terms

For $f : \mathbb{R}^n \to \mathbb{R}$, the second-order terms of the Taylor series of f at $\mathbf{x}_0$ are:

$$T^2_{\mathbf{x}_0}(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

The part $\nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ is a *linear function(al)*!

Taylor Series

Just the second-order terms

For $f : \mathbb{R}^n \to \mathbb{R}$, the second-order terms of the Taylor series of f at $\mathbf{x}_0$ are:

$$T^2_{\mathbf{x}_0}(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

The part $\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ is a *quadratic form*!

First-order Taylor Approximation

Just linearization

For a function $f : \mathbb{R} \to \mathbb{R}$, the *Taylor series at* x_0 is

$$T_{x_0}(x) = \underbrace{f(x_0) + \frac{f'(x_0)}{1!}(x - x_0)}_{\text{first-order terms}} + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots$$

For $f : \mathbb{R}^n \to \mathbb{R}$, the *Taylor series at* $\mathbf{x}_0$ is

$$T_{\mathbf{x}_0}(\mathbf{x}) = \underbrace{f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)}_{\text{first-order terms}} + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \dots$$

Linearization of f at $\mathbf{x}_0$. This is just taking the first-order terms of the Taylor series!

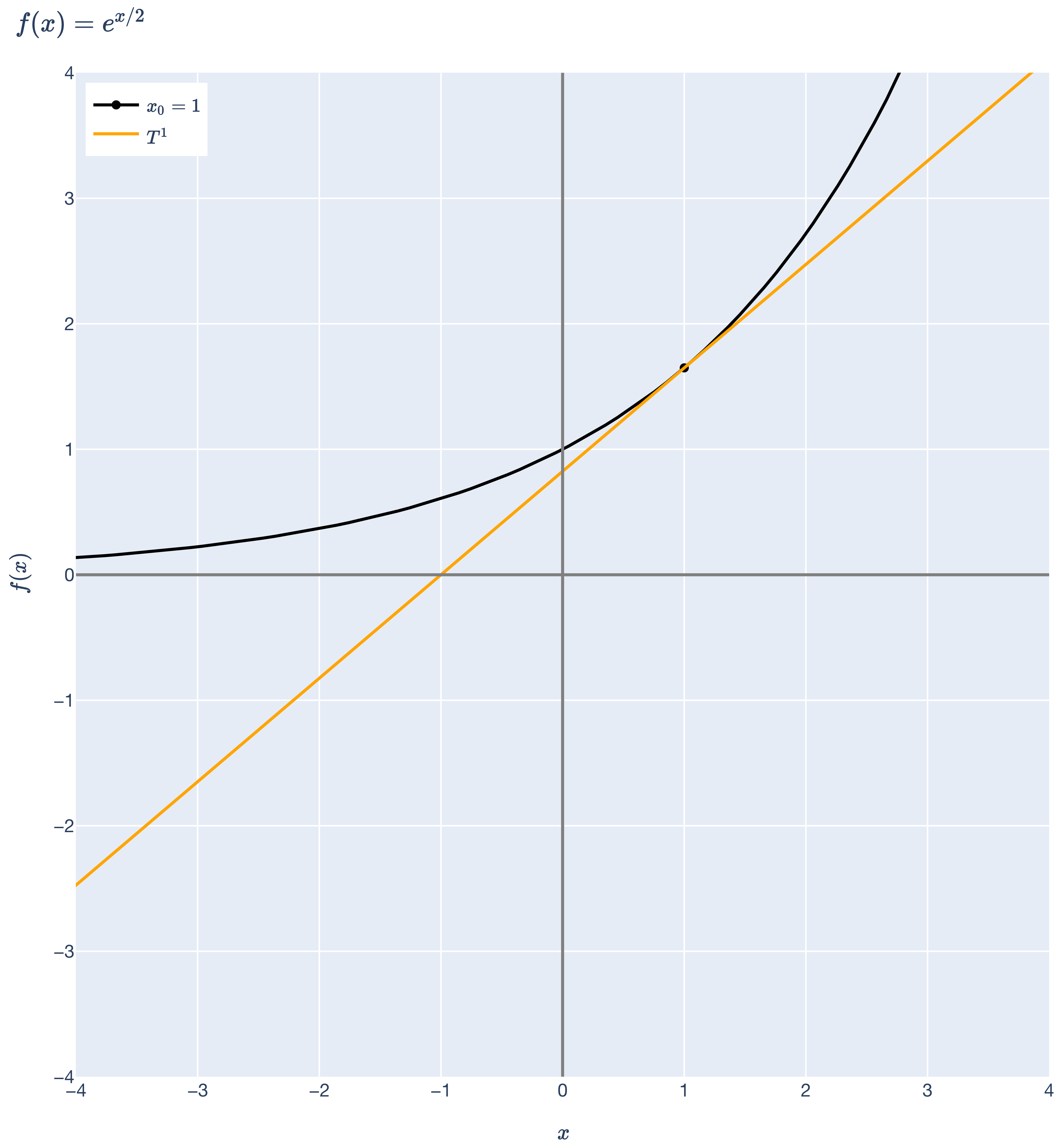
First-order Taylor Approximation

Single-variable example

$$f(x) = e^{x/2}$$

First-order Taylor expansion at $x_0 = 1$:

$$T^1(x) = e^{1/2} + \frac{e^{1/2}(x-1)}{2}$$



Second-order Taylor Approximation

Approximation by a quadratic

For $f : \mathbb{R} \to \mathbb{R}$,

$$T(x) = \underbrace{x_0 + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2}_{\text{second-order terms}} + \frac{f'''(x_0)^3}{3!}(x - x_0)^3 + \ldots$$

For $f : \mathbb{R}^n \to \mathbb{R}$,

$$T_{\mathbf{x}_0}(\mathbf{x}) = \underbrace{f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)}_{\text{second-order terms}} + \ldots$$

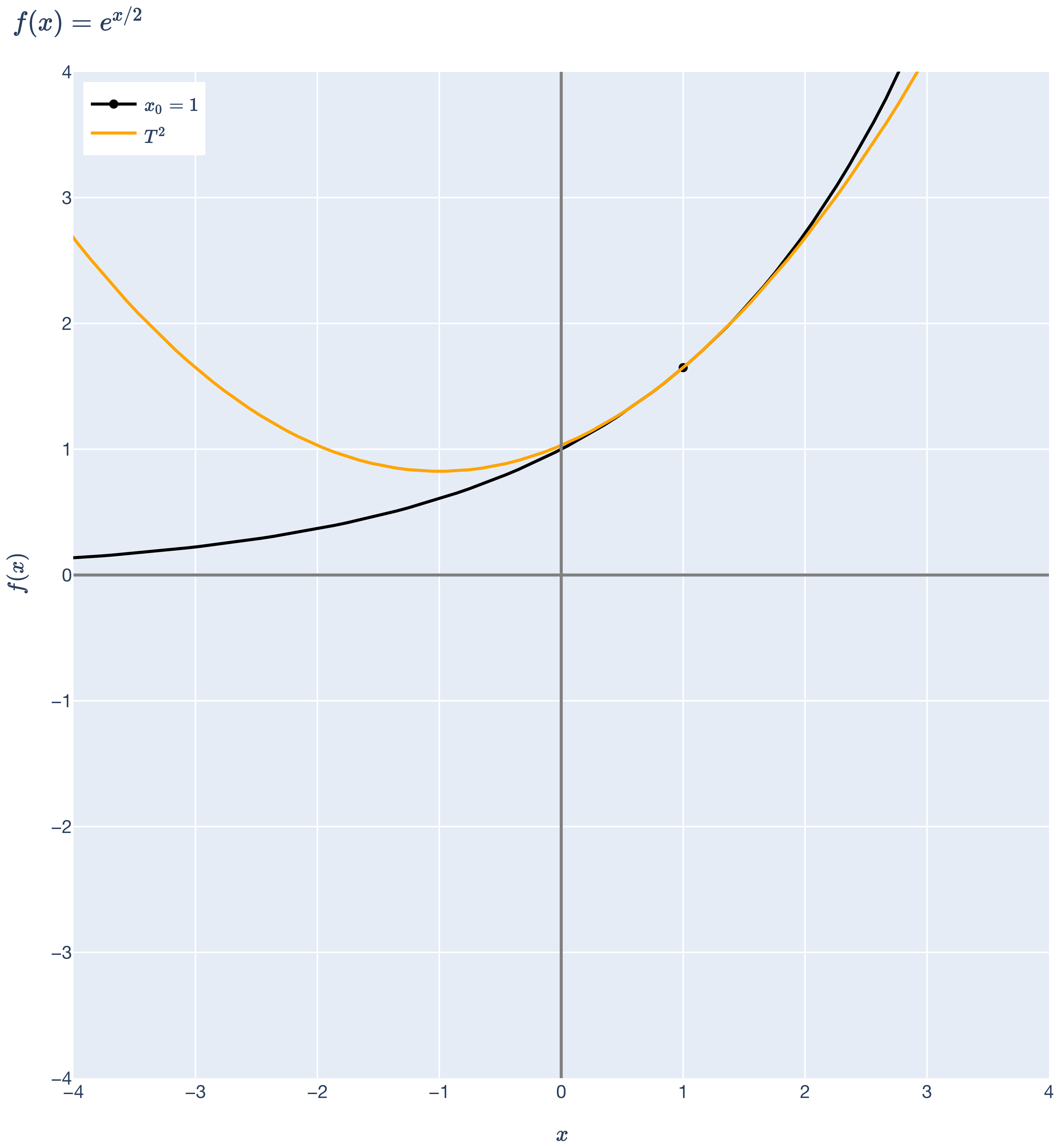
Second-order Taylor Approximation

Single-variable example

$$f(x) = e^{x/2}$$

Second-order Taylor expansion at $x_0 = 1$:

$$T^2(x) = e^{1/2} + \frac{e^{1/2}(x-1)}{2} + \frac{e^{1/2}(x-1)^2}{8}$$



Taylor Approximations
Summary

The *first-order Taylor approximation (linearization)* of a function at $\mathbf{x}_0$ is:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)\,.$$

The *second-order Taylor approximation* of a function at $\mathbf{x}_0$ is:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\,.$$

A natural question to ask is: *how good are these approximations?*

Taylor’s Theorem

Quantifying the approximation

Taylor's Theorem
Intuition

How much do we lose by approximating f with a Taylor approximation? We'll think of this in terms of the "remainder" — how much more Taylor series is left after "chopping it off" at order n.

First-order approximation:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$$

The remainder is:

$$f(\mathbf{x}) - (f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0))$$

Taylor's Theorem

Intuition

How much do we lose by approximating f with a Taylor approximation? We'll think of this in terms of the "remainder" — how much more Taylor series is left after "chopping it off" at order n.

Second-order approximation:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\,.$$

The remainder is:

$$f(\mathbf{x}) - \left(f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \right).$$

Remainder of Taylor Polynomial

Definition

The ***remainder*** of a function and its Taylor polynomial at $\mathbf{x}_0$ is the function:

$$R^n(\mathbf{x}) := f(\mathbf{x}) - T^n_{\mathbf{x}_0}(\mathbf{x})$$

What behavior would we like? Ideally, $R^n(\mathbf{x}) \to 0$ as $\mathbf{x} \to \mathbf{x}_0$ (the approximation gets better as we approach $\mathbf{x}_0$).

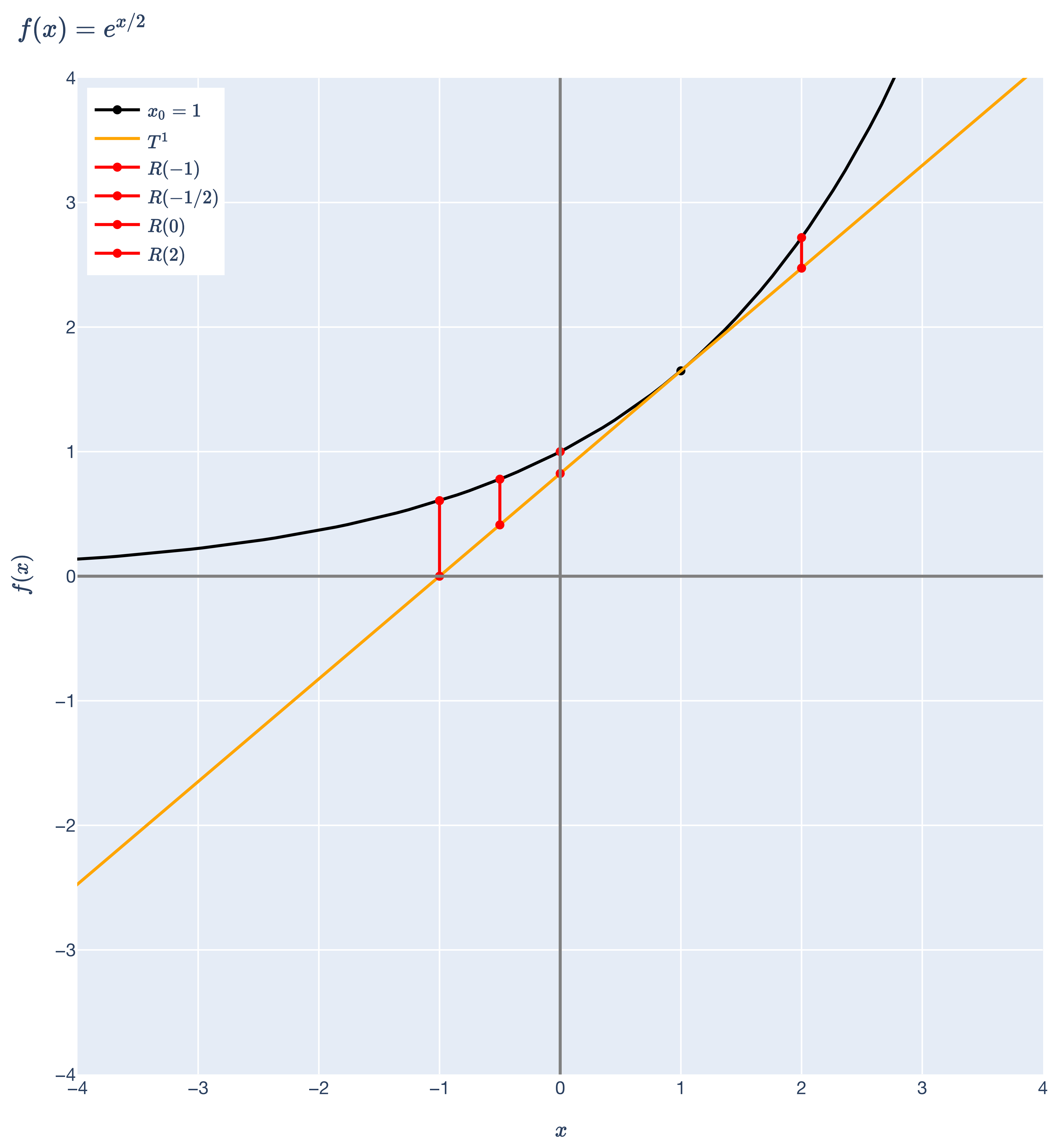
Remainder of Taylor Polynomial

Definition

The ***remainder*** of a function and its Taylor polynomial at $\mathbf{x}_0$ is the function:

$$R^n(\mathbf{x}) := f(\mathbf{x}) - T^n_{\mathbf{x}_0}(\mathbf{x})$$

What behavior would we like? Ideally, $R^n(\mathbf{x}) \to 0$ as $\mathbf{x} \to \mathbf{x}_0$ (the approximation gets better as we approach $\mathbf{x}_0$).



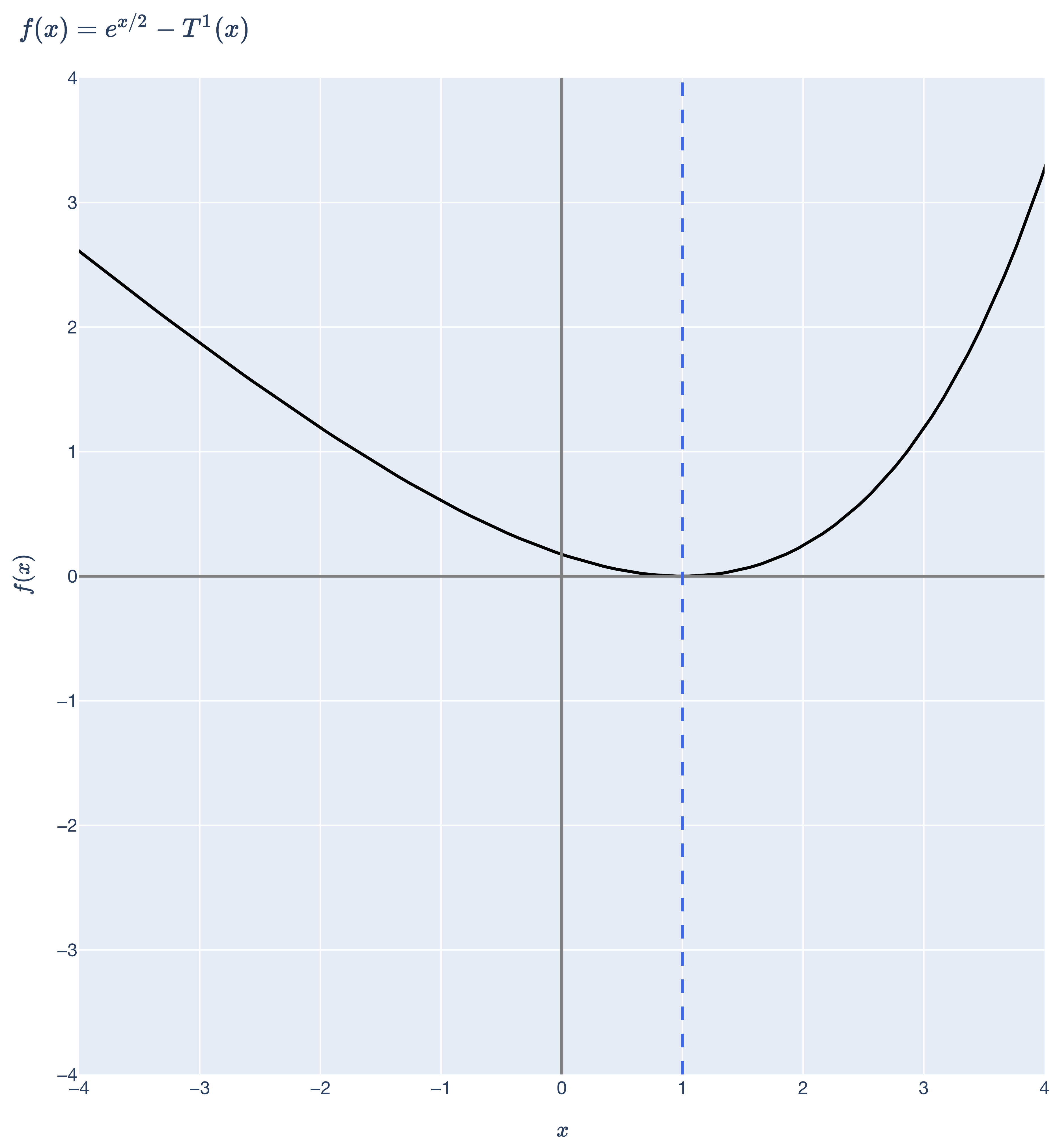
Remainder of Taylor Polynomial

Definition

The ***remainder*** of a function and its Taylor polynomial at $\mathbf{x}_0$ is the function:

$$R^n(\mathbf{x}) := f(\mathbf{x}) - T^n_{\mathbf{x}_0}(\mathbf{x})$$

What behavior would we like? Ideally, $R^n(\mathbf{x}) \to 0$ as $\mathbf{x} \to \mathbf{x}_0$ (the approximation gets better as we approach $\mathbf{x}_0$).



Taylor's Theorem

Idea: Taylor's Theorem (Peano's Form)

Say we want the value of f at $\mathbf{x}$ and we have a Taylor approximation at $\mathbf{x}_0$.

Then, the *direction* to go from $\mathbf{x}$ to $\mathbf{x}_0$ is $\mathbf{d} = \mathbf{x} - \mathbf{x}_0$.

By taking a constant $\alpha > 0$, we can make the direction $\alpha\mathbf{d}$ as small as we want:

$$\|\alpha\mathbf{d}\| = \alpha\|\mathbf{d}\|.$$

Taylor's Theorem

Idea: Taylor's Theorem (Peano's Form)

By taking a constant $\alpha > 0$, we can make the direction $\alpha\mathbf{d}$ as small as we want:

$$||\alpha\mathbf{d}|| = \alpha||\mathbf{d}||.$$

Peano's Form of Taylor's Theorem says that for any direction $\mathbf{d}$, as $\alpha \to 0$,

$$T^n(\mathbf{x}_0 + \alpha\mathbf{d}) \to f(\mathbf{x}) = f(\mathbf{x}_0 + \alpha\mathbf{d}),$$

i.e. the approximation when we "chop off" the Taylor series at n approaches the function's actual value.

Little O Asymptotics

Definition

For two functions, $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$, with g nonnegative, f is ***asympotically smaller than g*** or ***"little-oh" of g***, denoted

$$f(x) = o(g(x))$$

if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0.$$

Taylor's Theorem

Remainder Theorem 1: Peano's Form Taylor's Theorem

Theorem (Taylor's Theorem: Peano's Form). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a k-times differentiable function at $\mathbf{x}_0$. Then, for every direction $\mathbf{d} \in \mathbb{R}^d$:

$$f(\mathbf{x}_0 + \mathbf{d}) = T^k_{\mathbf{x}_0}(\mathbf{x}_0 + \mathbf{d}) + o(\|\mathbf{d}\|^k), \text{ as } \mathbf{d} \to \mathbf{0},$$

where $o(\|\mathbf{d}\|^k)$ as $\mathbf{d} \to \mathbf{0}$ means that if $R^k(\mathbf{x}_0 + \mathbf{d}) := f(\mathbf{x}_0 + \mathbf{d}) - T^k_{\mathbf{x}_0}(\mathbf{x}_0 + \mathbf{d})$,

$$\lim_{\mathbf{d} \to 0} \frac{R^k(\mathbf{x}_0 + \mathbf{d})}{\|\mathbf{d}\|^k} = 0.$$

We'll usually only go up to $k = 2$ (quadratic approximation), so we'll only need…

Taylor's Theorem

Remainder Theorem 1: Peano's Form Taylor's Theorem

Theorem (2nd Order Taylor's Theorem: Peano's Form). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function at $\mathbf{x}_0$. Then, for every direction $\mathbf{d} \in \mathbb{R}^d$:

$$f(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{d} + o(\|\mathbf{d}\|^2)\,.$$

The remainder is

$$R^2(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0 + \mathbf{d}) - \left(f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{d} \right),$$

and the claim is that $R^2(\mathbf{x}_0 + \mathbf{d}) = o(\|\mathbf{d}\|^2)$, meaning that $\lim_{\mathbf{d} \to \mathbf{0}} R^2(\mathbf{x}_0 + \mathbf{d})/\|\mathbf{d}\|^2 = 0$.

Taylor’s Theorem

Remainder Theorem 2: Lagrange’s Form Taylor’s Theorem

Theorem (Taylor’s Theorem: Lagrange Form). Let $f : \mathbb{R} \to \mathbb{R}$ be a $\mathscr{C}^{k+1}$ function on the closed interval between x_0 and x. Then, there exists some number $z \in \mathbb{R}$ between x_0 and x such that

$$f(x) = T^n(x) + \frac{f^{(n+1)}(z)}{(n+1)\,!}(x - x_0)^{n+1}\,.$$

So, in terms of the remainder:

$$R^n(x) = \frac{f^{(n+1)}(z)}{(n+1)\,!}(x - x_0)^{n+1}\,.$$

Taylor's Theorem

Remainder Theorem 2: Lagrange's Form Taylor's Theorem

Theorem (1st Order Taylor's Theorem - Lagrange Form). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$ function. For $\mathbf{x}_0, \mathbf{d} \in \mathbb{R}^n$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{x}} = \mathbf{x}_0 + \lambda\mathbf{d}$ on the line segment between $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{d}$

$$f(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}$$

Or, in terms of the remainder:

$$R^1(\mathbf{x}_0 + \mathbf{d}) = \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}.$$

Gradient Descent

Intuition and Algorithm

Motivation

Optimization in calculus

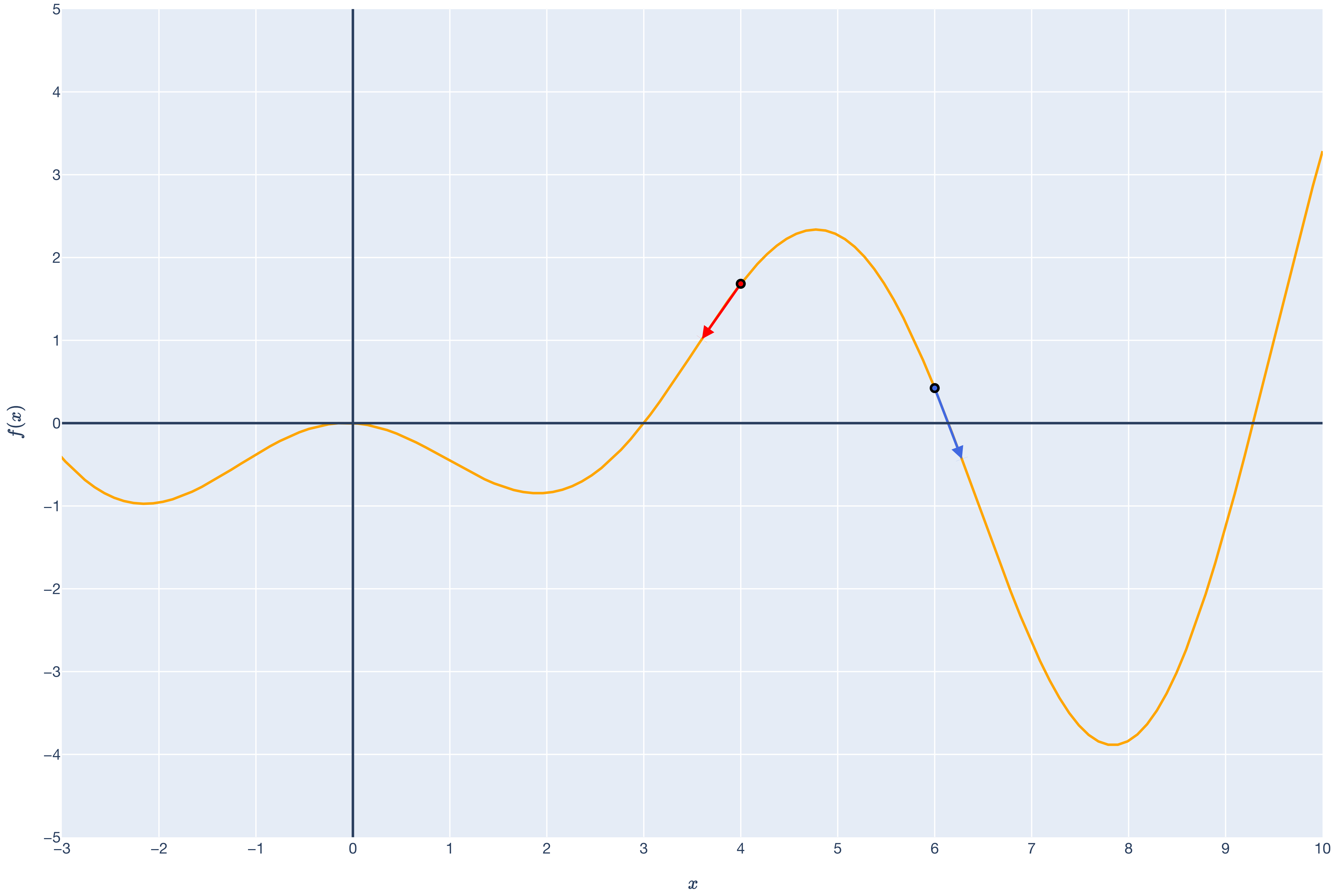
We want to minimize an ***objective function*** $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x})$$

Gradient Descent

Idea

How do you get to the minimum?



Gradient Descent

Gradient as direction of steepest ascent

Theorem (Gradient and direction of steepest ascent). Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable at $\mathbf{x}_0 \in \mathbb{R}^d$. If $\mathbf{d} \in \mathbb{R}^d$ is a *unit* vector making angle θ with the gradient $\nabla f(\mathbf{x}_0)$, then:

$$\nabla f(\mathbf{x}_0)^\top \mathbf{d} = \|\nabla f(\mathbf{x}_0)\| \cos\theta.$$

Therefore, the directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{d}$ is maximized in the direction $\nabla f(\mathbf{x}_0)$!

Gradient is the direction of *steepest ascent* at the rate $\|\nabla f(\mathbf{x}_0)\|$!

Gradient Descent

The direction of steepest descent

Going in the direction $-\nabla f(\mathbf{x}_0)$ gives the direction of *steepest descent.*

Here's a candidate algorithm:

1. Initialize at a point $\mathbf{x}_0$.
2. Obtain $\mathbf{x}_1$ by moving in the direction $-\nabla f(\mathbf{x}_0)$.
3. Obtain $\mathbf{x}_2$ by moving in the direction $-\nabla f(\mathbf{x}_1)$.
4. Repeat until convergence to a minimum…

Gradient Descent
Algorithm

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

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

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

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

Gradient Descent

Taylor’s Theorem for Convergence Theorem

Taylor Approximation

1st Order Taylor Approximation

Recall the first-order Taylor approximation:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)\,.$$

As long as $\mathbf{x}$ is close enough to $\mathbf{x}_0$, this is a good approximation.

At time $t \geq 0$, we are at the point $\mathbf{x}_t \in \mathbb{R}^d$. We want to move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{x}_t + \mathbf{d}) < f(\mathbf{x}_t)$. Our choice? $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

Taylor Approximation

1st Order Taylor Approximation

Recall the first-order Taylor approximation:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)\,.$$

As long as $\mathbf{x}$ is close enough to $\mathbf{x}_0$, this is a good approximation.

At time $t \geq 0$, we are at the point $\mathbf{x}_t \in \mathbb{R}^d$. We want to move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{x}_t + \mathbf{d}) < f(\mathbf{x}_t)$. Our choice? $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

Why? If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$

Taylor Approximation

1st Order Taylor Approximation

At time $t \geq 0$, we are at the point $\mathbf{x}_t \in \mathbb{R}^d$. We want to move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{x}_t + \mathbf{d}) < f(\mathbf{x}_t)$. Our choice? $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

Why? If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$

This explains the gradient descent step: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$.

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) < f(\mathbf{x}_t) \text{ as long as } \eta \text{ is small.}$$

Taylor Approximation

1st Order Taylor Approximation

At time $t \geq 0$, we are at the point $\mathbf{x}_t \in \mathbb{R}^d$. We want to move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{x}_t + \mathbf{d}) < f(\mathbf{x}_t)$. Our choice? $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

Why? If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$

This explains the gradient descent step: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$.

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) < f(\mathbf{x}_t) \text{ as long as } \eta \text{ is small.}$$

To quantify the $\approx$, we had Taylor's theorem. We will use the *Lagrange form* of Taylor's theorem.

Taylor's Theorem

Remainder Theorem 2: Lagrange Form of Taylor's Theorem

Theorem (1st Order Taylor's Theorem - Lagrange Form). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$ function. For $\mathbf{x}_0, \mathbf{d} \in \mathbb{R}^n$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{x}} = \mathbf{x}_0 + \lambda \mathbf{d}$ on the line segment between $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{d}$

$$f(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}$$

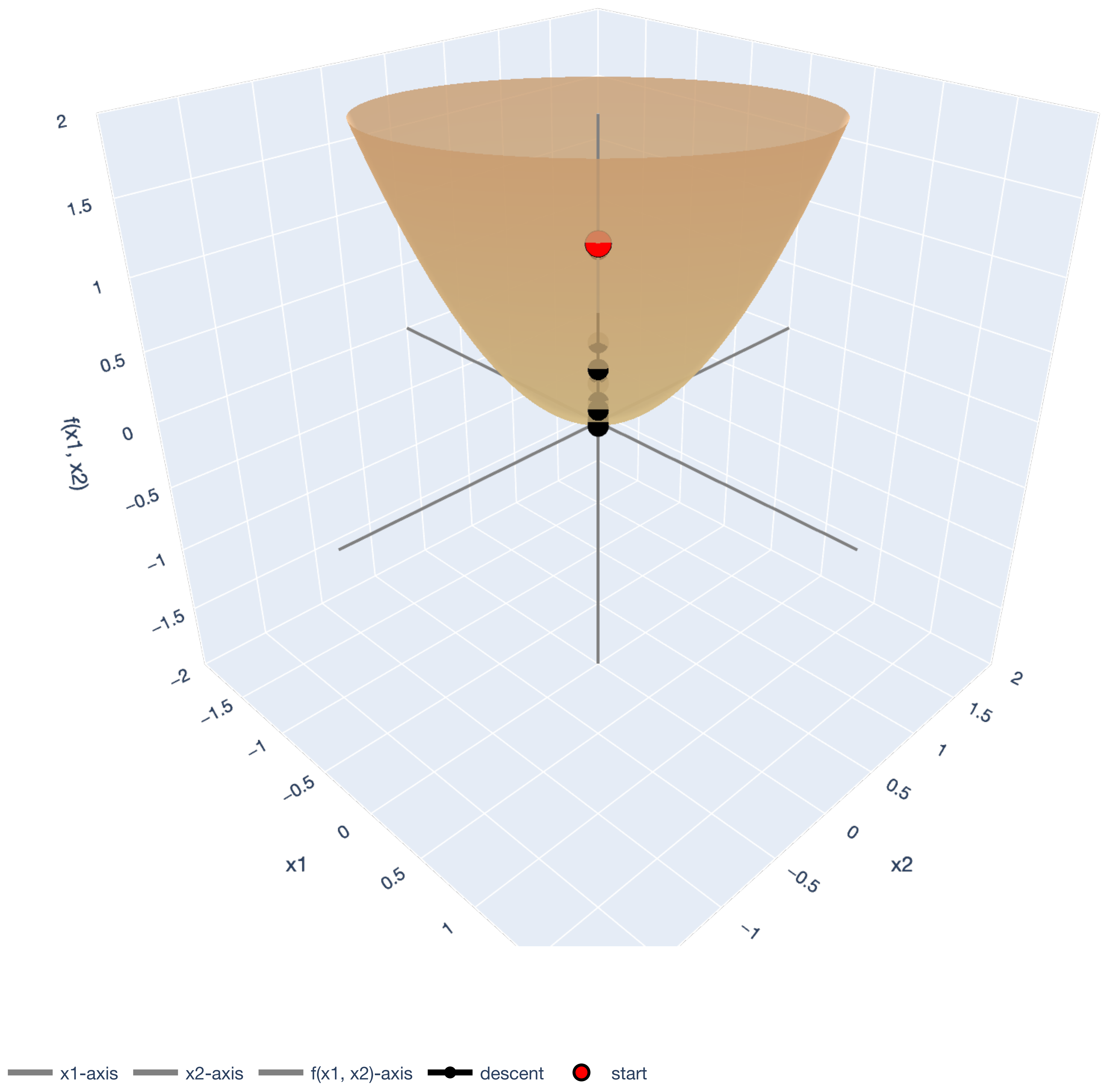
Gradient Descent and η

Example

Move in the direction: $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$



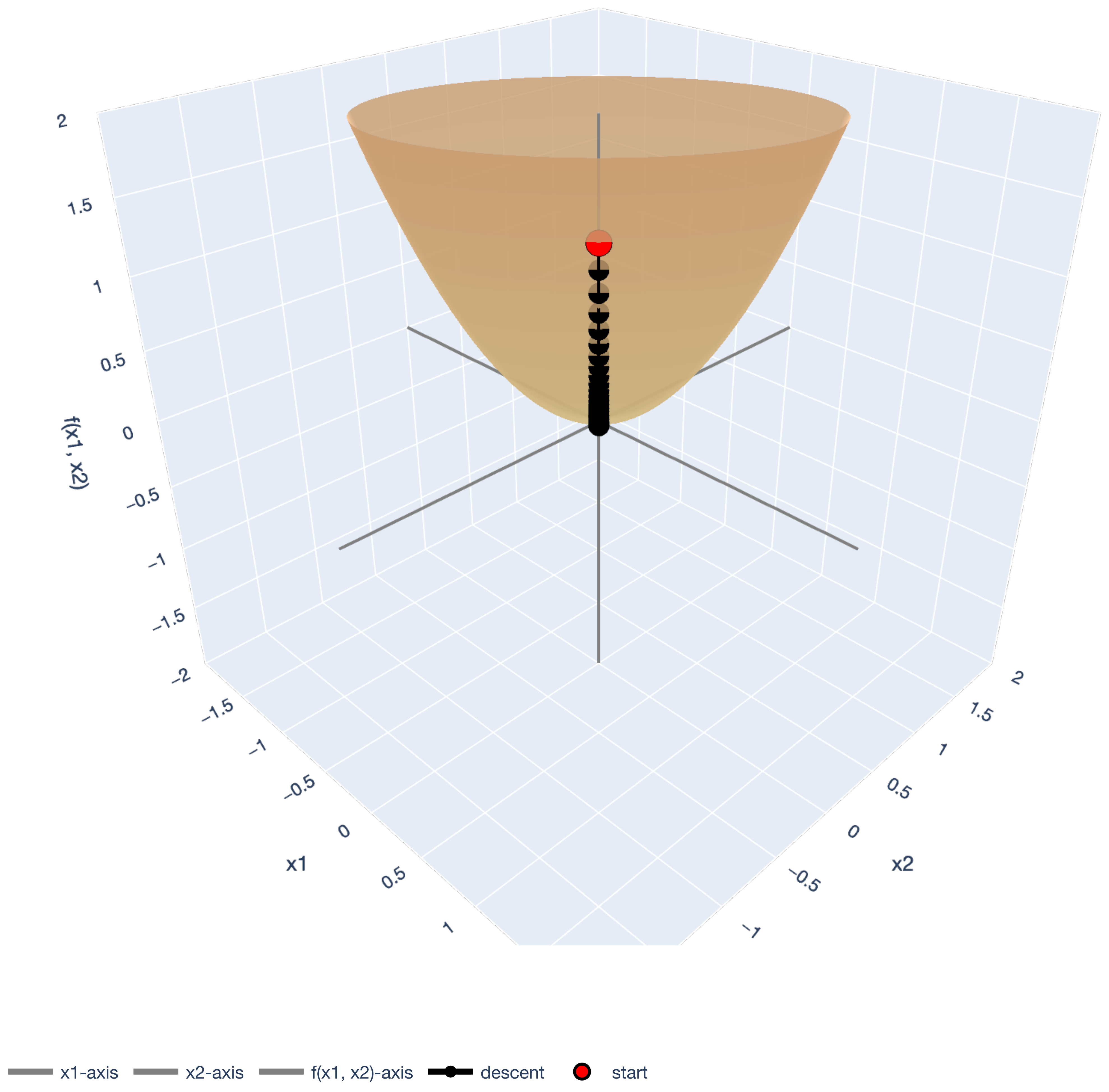
Gradient Descent and η

Example

Move in the direction: $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$



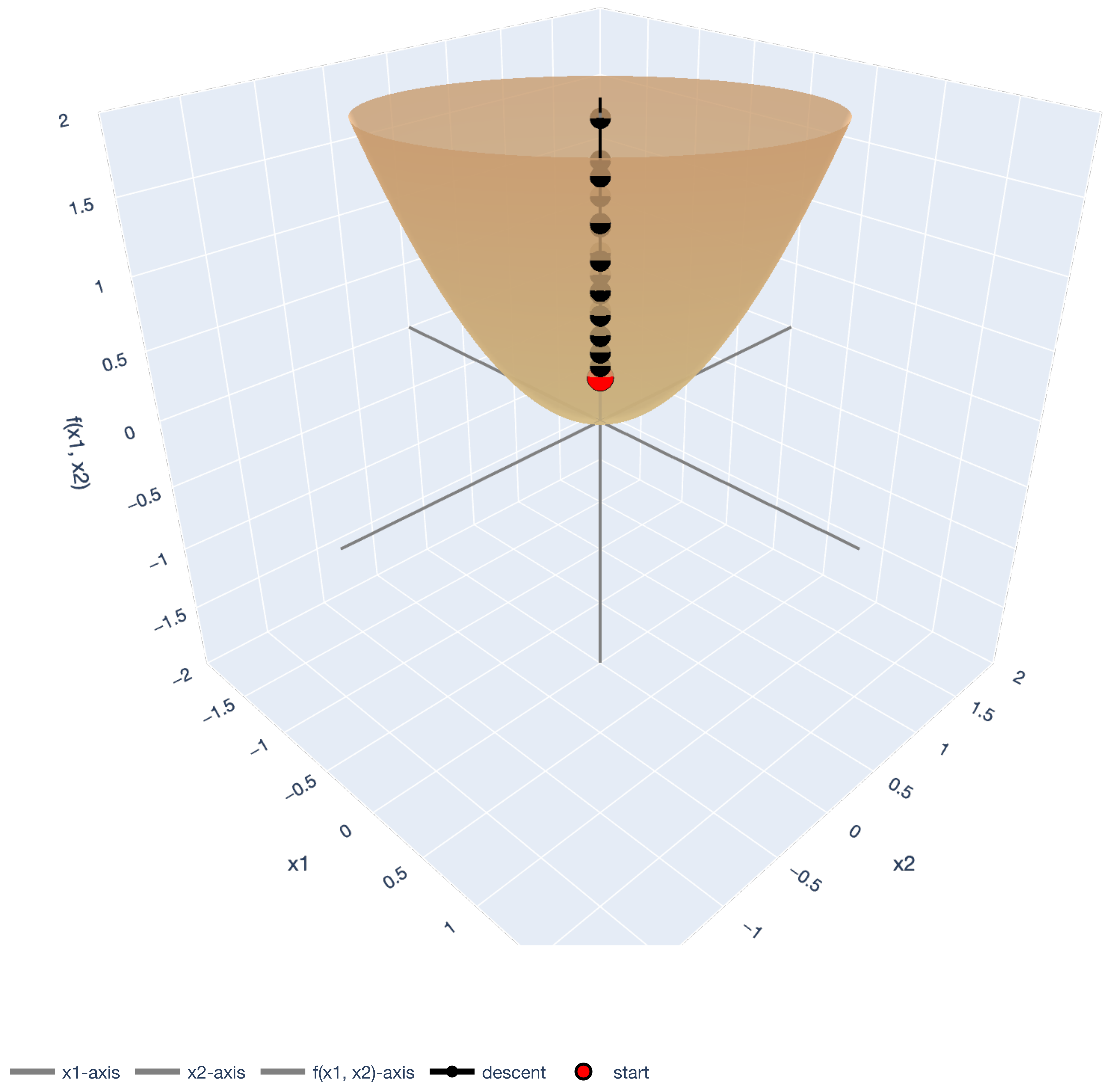
Gradient Descent and η

Example

Move in the direction: $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$



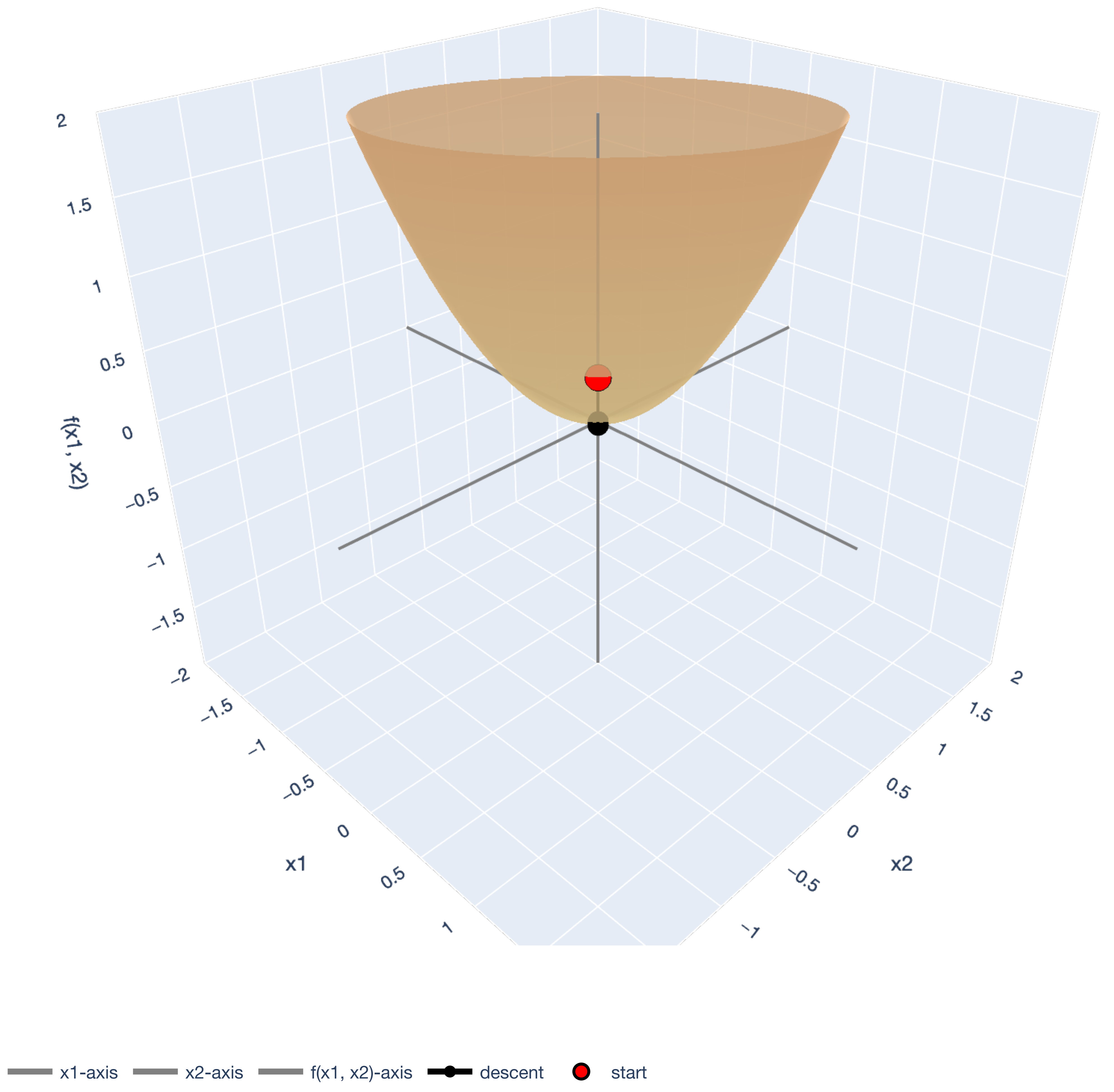
Gradient Descent and η

Example

Move in the direction: $\mathbf{d} = -\eta \nabla f(\mathbf{x}_t)$.

If η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and:

$$f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$



Gradient Descent and η

Applying the first-order Taylor Approximation

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) < f(\mathbf{x}_t) \text{ as long as } \eta \text{ is small.}$$

We would like the assurance that gradient descent is always decreasing our function:

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) \text{ at each step } t.$$

Gradient Descent and η

Applying the first-order Taylor Approximation

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t) - \eta \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) < f(\mathbf{x}_t) \text{ as long as } \eta \text{ is small.}$$

We would like the assurance that gradient descent is always decreasing our function:

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) \text{ at each step } t.$$

Strategy: Use Taylor's Theorem to analyze the first-order approximation! This works if the first derivative doesn't change too much.

Bounding change in gradients

β-smoothness

For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, the largest eigenvalue of $\mathbf{A}$ is $\lambda_{\max}(\mathbf{A})$.

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a ***β-smooth matrix*** if its eigenvalues are at most β:

$$\lambda_{\max}(\mathbf{A}) \leq \beta .$$

Bounding change in gradients

β-smoothness

A twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is a ***β-smooth function*** if the eigenvalues of its Hessian at any point $\mathbf{x} \in \mathbb{R}^d$ are at most β. That is:

$$\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq \beta\,.$$

Bounding change in gradients

β-smoothness

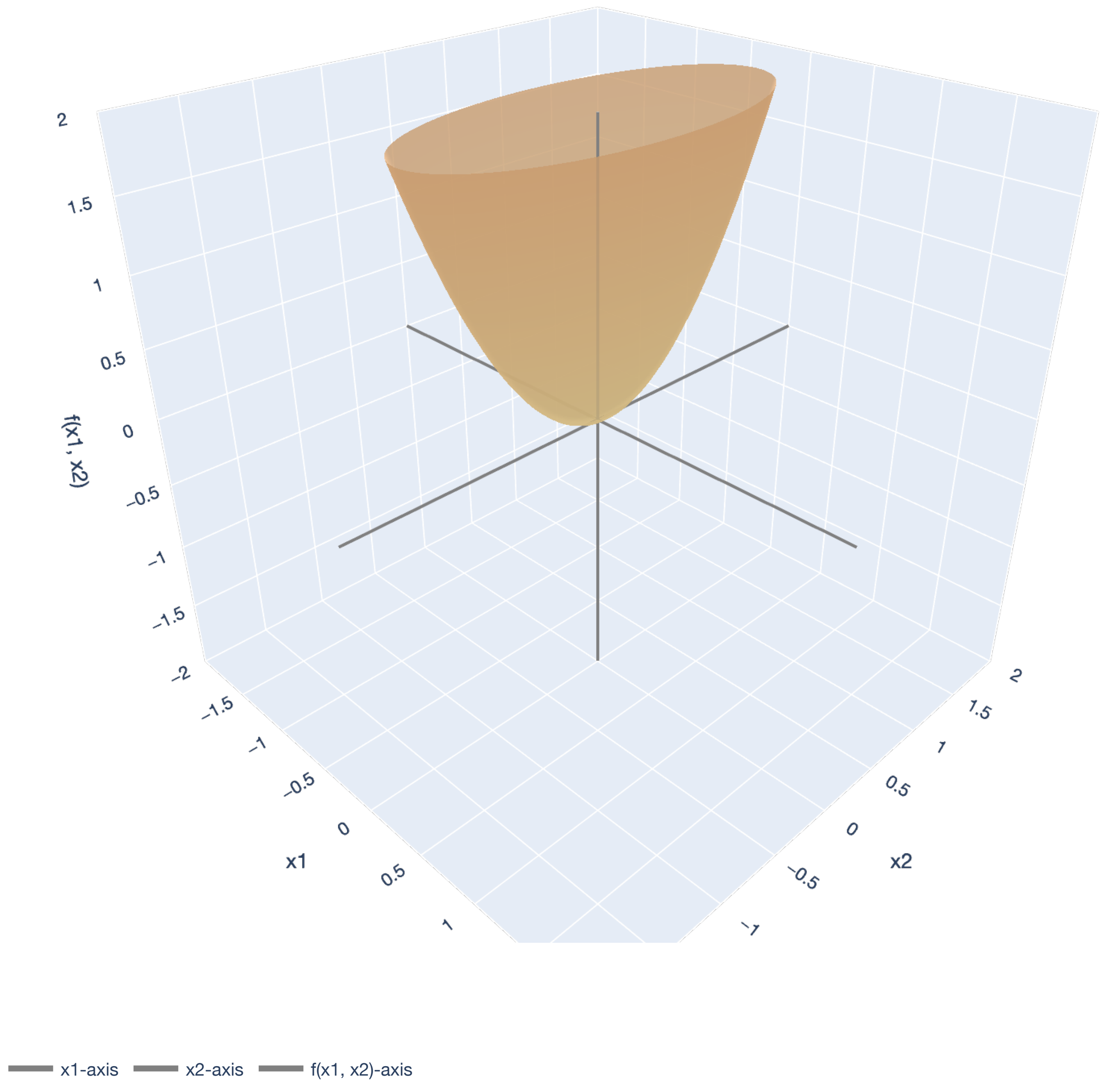
Property (Smoothness bounds quadratic forms). If $\mathbf{A} \in \mathbb{R}^{d \times d}$ is β-smooth, then for any unit vector $\mathbf{v} \in \mathbb{R}^d$,

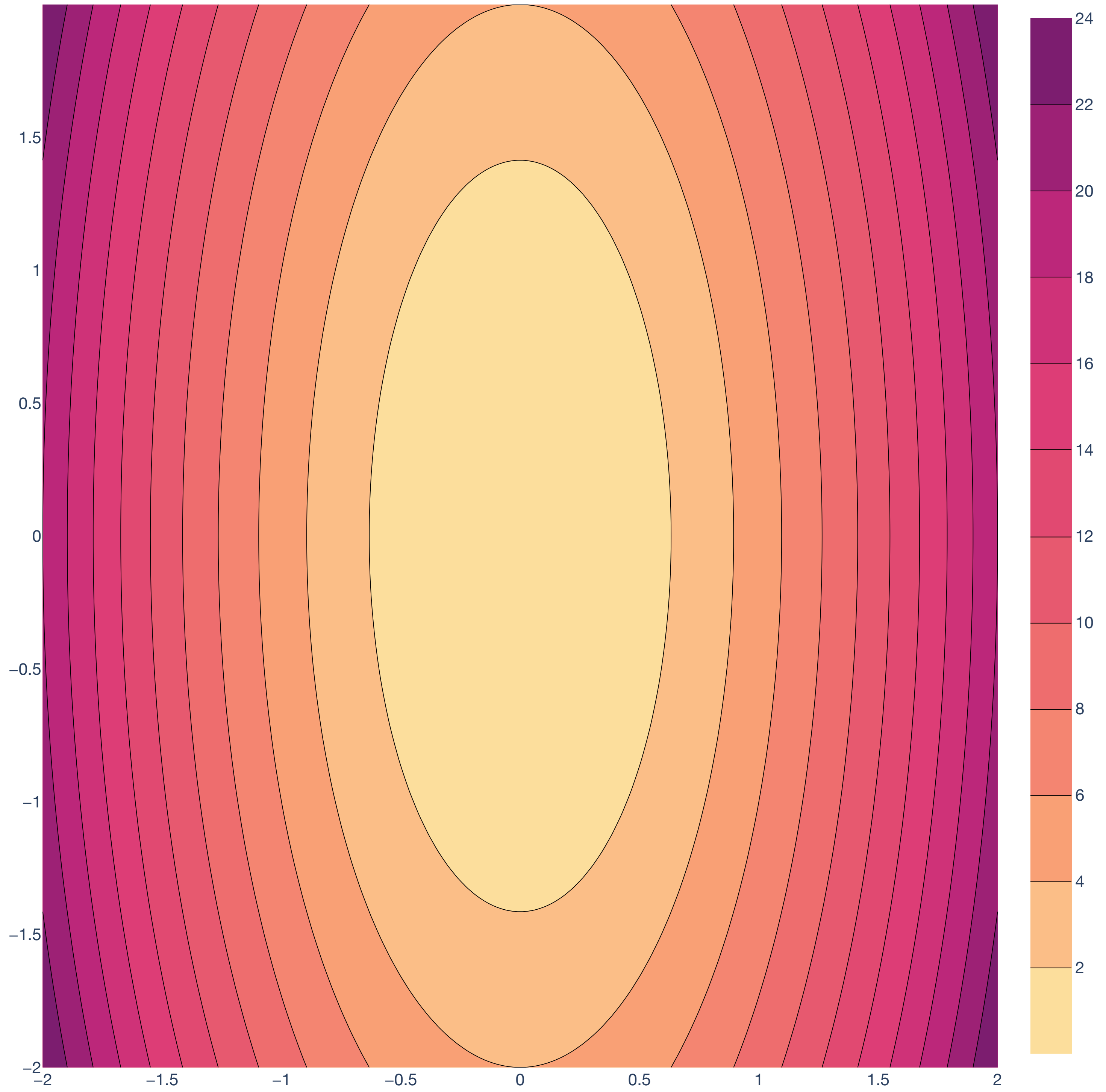
$$\mathbf{v}^\top \mathbf{A} \mathbf{v} \quad \leq \beta.$$

Bounding change in gradients

β-smoothness

$$\mathbf{\Lambda} = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$

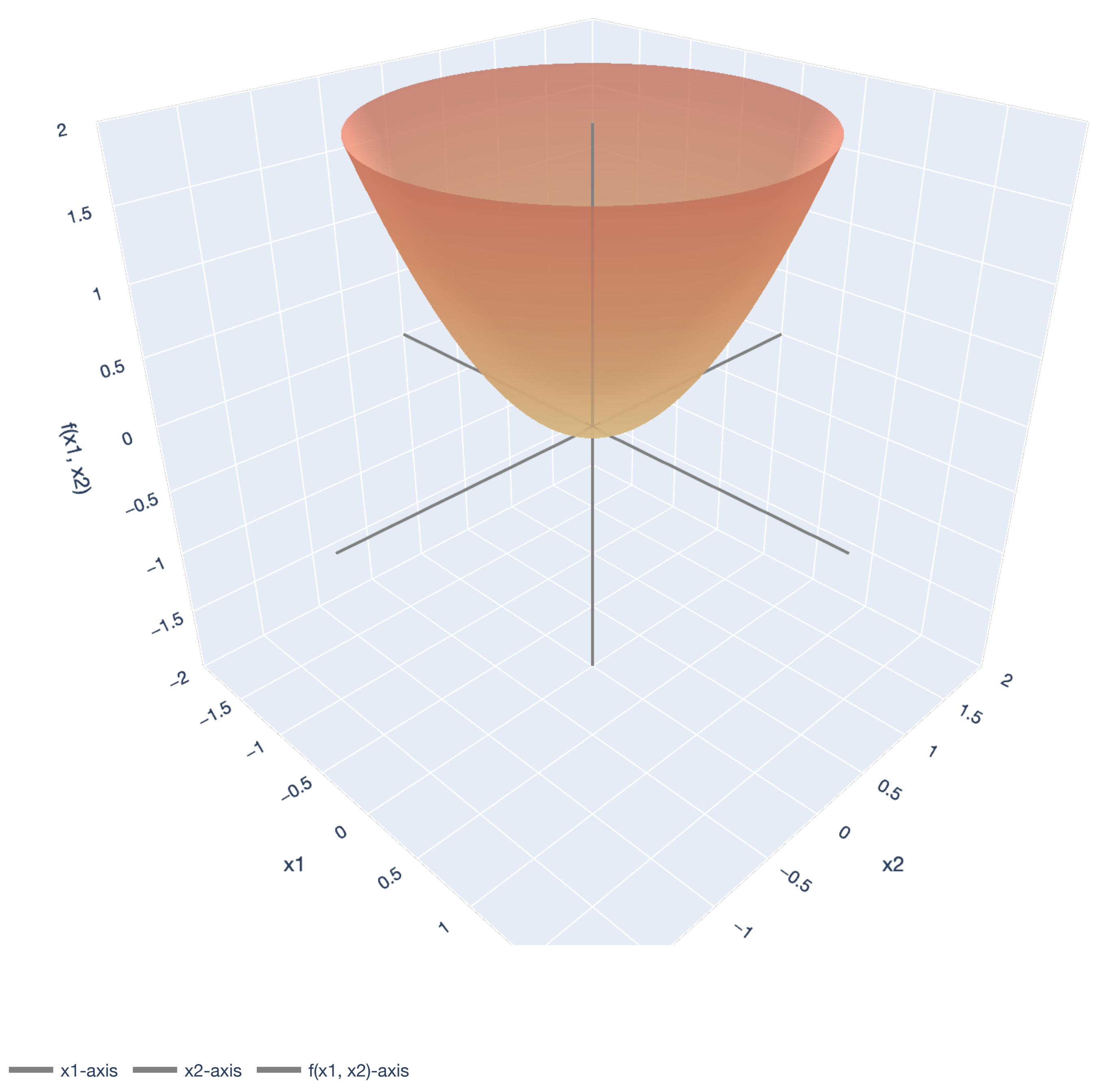


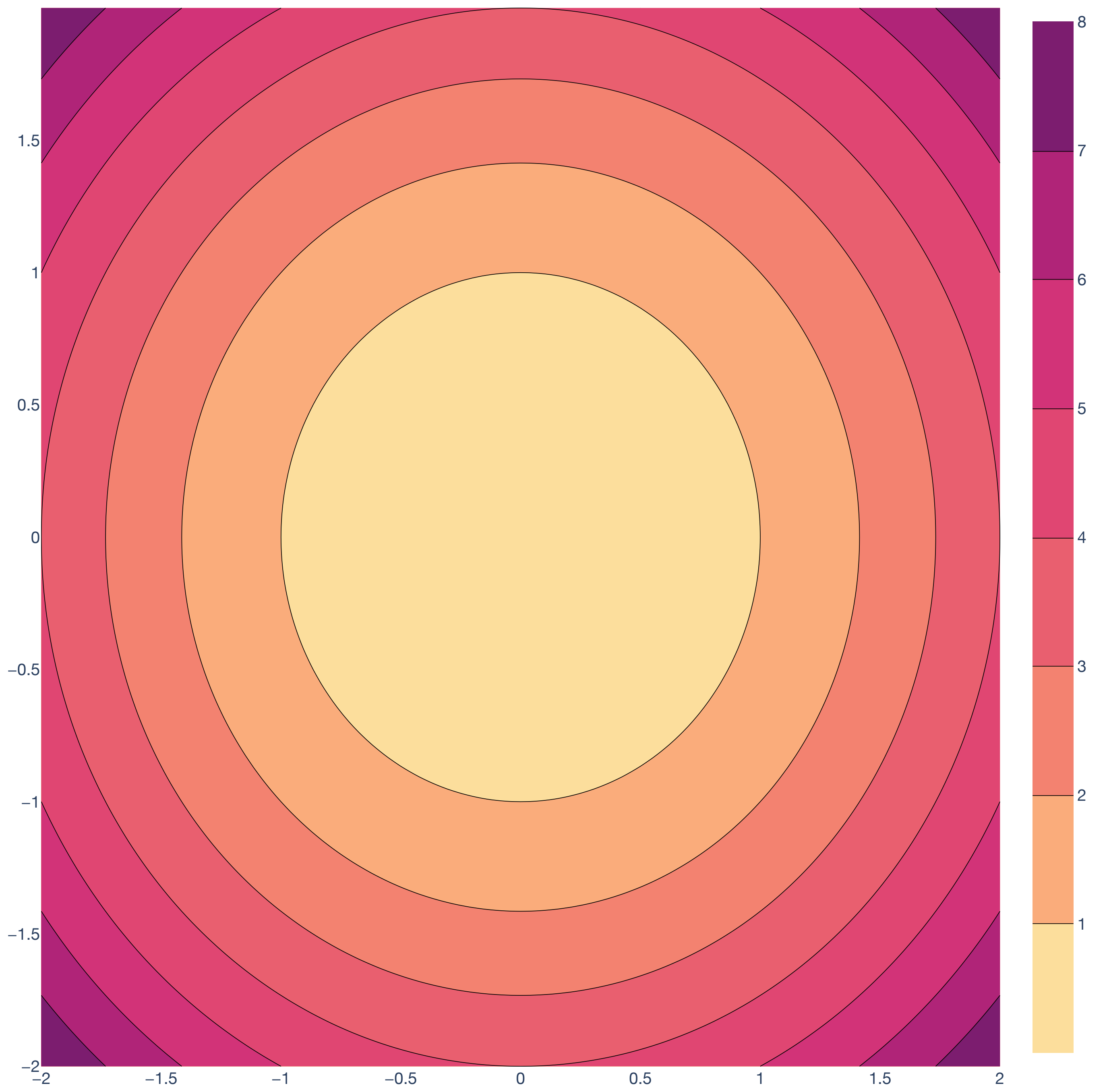


Bounding change in gradients

β-smoothness

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





Gradient Descent

Applying Taylor's Theorem

Theorem (Gradient descent makes the function value smaller). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$, β-smooth function. Then, for any $t = 1,2,3,\ldots$, a gradient descent update

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

with step size $\eta = \dfrac{1}{\beta}$ has the property:

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) - \frac{1}{2\beta} \|\nabla f(\mathbf{x}_{t-1})\|^2.$$

This theorem says that gradient descent always makes our function value smaller, as long as the function's gradients don't change too much!

Gradient Descent

Main tool for proof of GD Theorem

Theorem (1st Order Taylor's Theorem - Lagrange Form). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$ function. For $\mathbf{x}_0, \mathbf{d} \in \mathbb{R}^n$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{x}} = \mathbf{x}_0 + \lambda\mathbf{d}$ on the line segment between $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{d}$

$$f(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}$$

Gradient Descent

Proof of GD Theorem

Want to show: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2.$

Step 1: Use Lagrange's Form of Taylor's Theorem to get an expression for $f(\mathbf{x}_t + \mathbf{d})$.

There exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{x}} = \mathbf{x}_t + \lambda \mathbf{d}$,

$$f(\mathbf{x}_t + \mathbf{d}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}$$

Gradient Descent

Proof of GD Theorem

Want to show: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2.$

Step 2: Use β-smoothness to bound the first-order approximation.

$$f(\mathbf{x}_t + \mathbf{d}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d}$$

Upper bound the quadratic term:

$$\begin{aligned} \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\tilde{\mathbf{x}})\mathbf{d} &= \frac{1}{2}\|\mathbf{d}\|^2(\mathbf{d}/\|\mathbf{d}\|)^\top \nabla^2 f(\tilde{\mathbf{x}})(\mathbf{d}/\|\mathbf{d}\|) \\ &\leq \frac{1}{2}\|\mathbf{d}\|^2\beta \end{aligned}$$

(bound on quadratic forms)

Gradient Descent

Proof of GD Theorem

Want to show: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2.$

Step 3: Optimize the quadratic upper bound to find the direction and magnitude to take a step.

$$f(\mathbf{x}_t + \mathbf{d}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\|\mathbf{d}\|^2\beta$$

We need to choose a direction $\mathbf{d} \in \mathbb{R}^d$ to take a step in. To do this, optimize the RHS:

$$\nabla_{\mathbf{d}}(f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\|\mathbf{d}\|^2\beta) = \nabla f(\mathbf{x}_t) + \beta\mathbf{d}$$

Set the gradient to $\mathbf{0}$ and solve:

$$\nabla f(\mathbf{x}_t) + \beta\mathbf{d} = 0 \implies \mathbf{d} = -\frac{1}{\beta}\nabla f(\mathbf{x}_t)$$

Gradient Descent

Proof of GD Theorem

Want to show: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2$.

Step 4: Plug optimal value of the quadratic upper bound back in to get our result.

Notice that $\mathbf{d} = -\frac{1}{\beta}\nabla f(\mathbf{x}_t)$ is exactly how we get our gradient step:

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) \text{ with } \eta = 1/\beta.$$

Plug this back into the quadratic upper bound: $f(\mathbf{x}_t + \mathbf{d}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\|\mathbf{d}\|^2\beta$

$$f(\mathbf{x}_{t+1}) = f\left(\mathbf{x}_t - \frac{1}{\beta}\nabla f(\mathbf{x}_t)\right) \leq f(\mathbf{x}_t) - \frac{1}{\beta}\nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) + \frac{1}{2\beta}\|\nabla f(\mathbf{x}_t)\|^2$$

$$\leq f(\mathbf{x}_t) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_t)\|^2$$

Gradient Descent

Applying Taylor's Theorem

Theorem (Gradient descent makes the function value smaller). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$, β-smooth function. Then, for any $t = 1,2,3,\ldots$, a gradient descent update

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

with step size $\eta = \dfrac{1}{\beta}$ has the property:

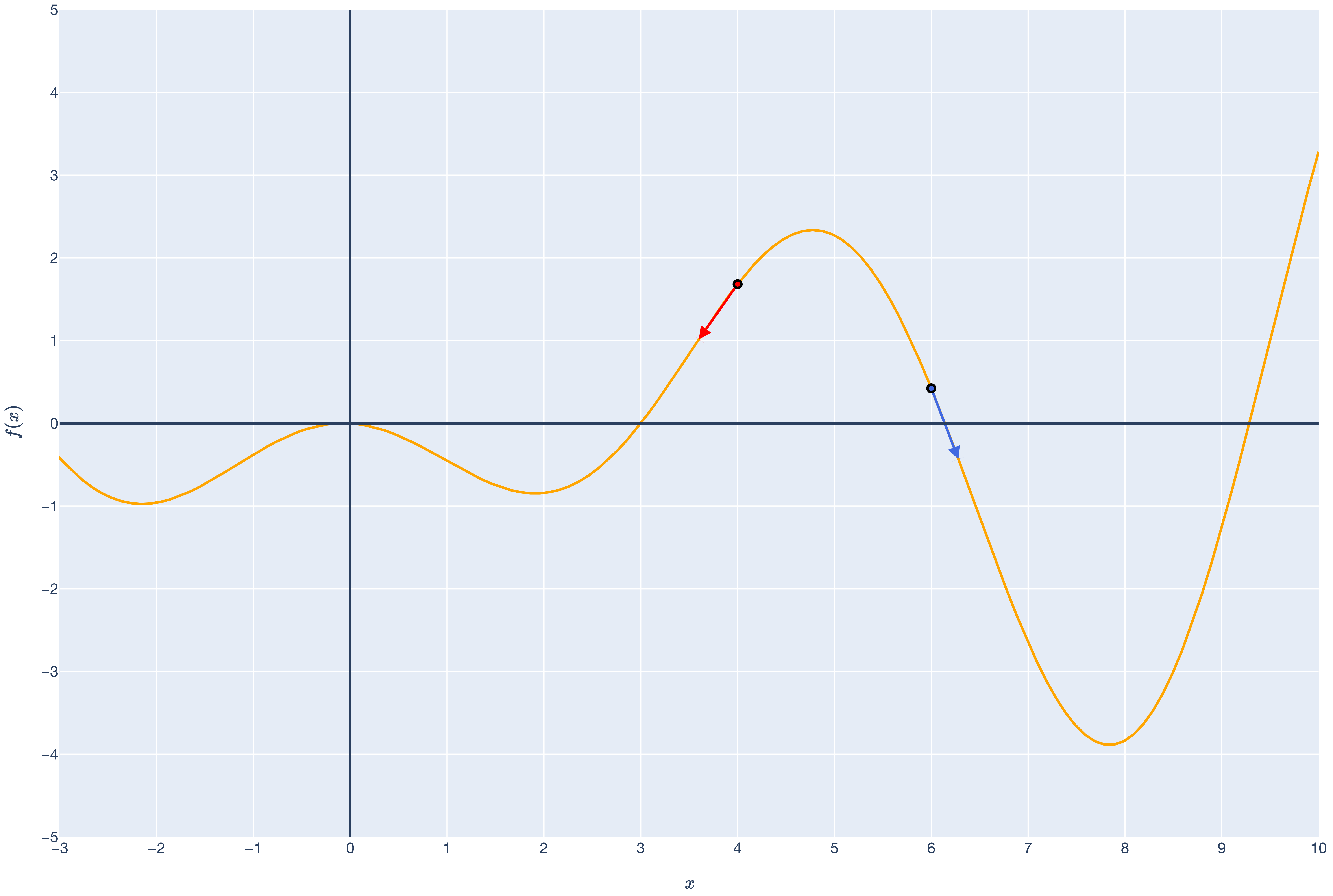
$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) - \frac{1}{2\beta} \|\nabla f(\mathbf{x}_{t-1})\|^2.$$

This theorem says that gradient descent always makes our function value smaller, as long as the function's gradients don't change too much!

Gradient Descent

Preview of convexity

Problem: gradient descent gets us to a *local* minimum, but perhaps not a global minimum.

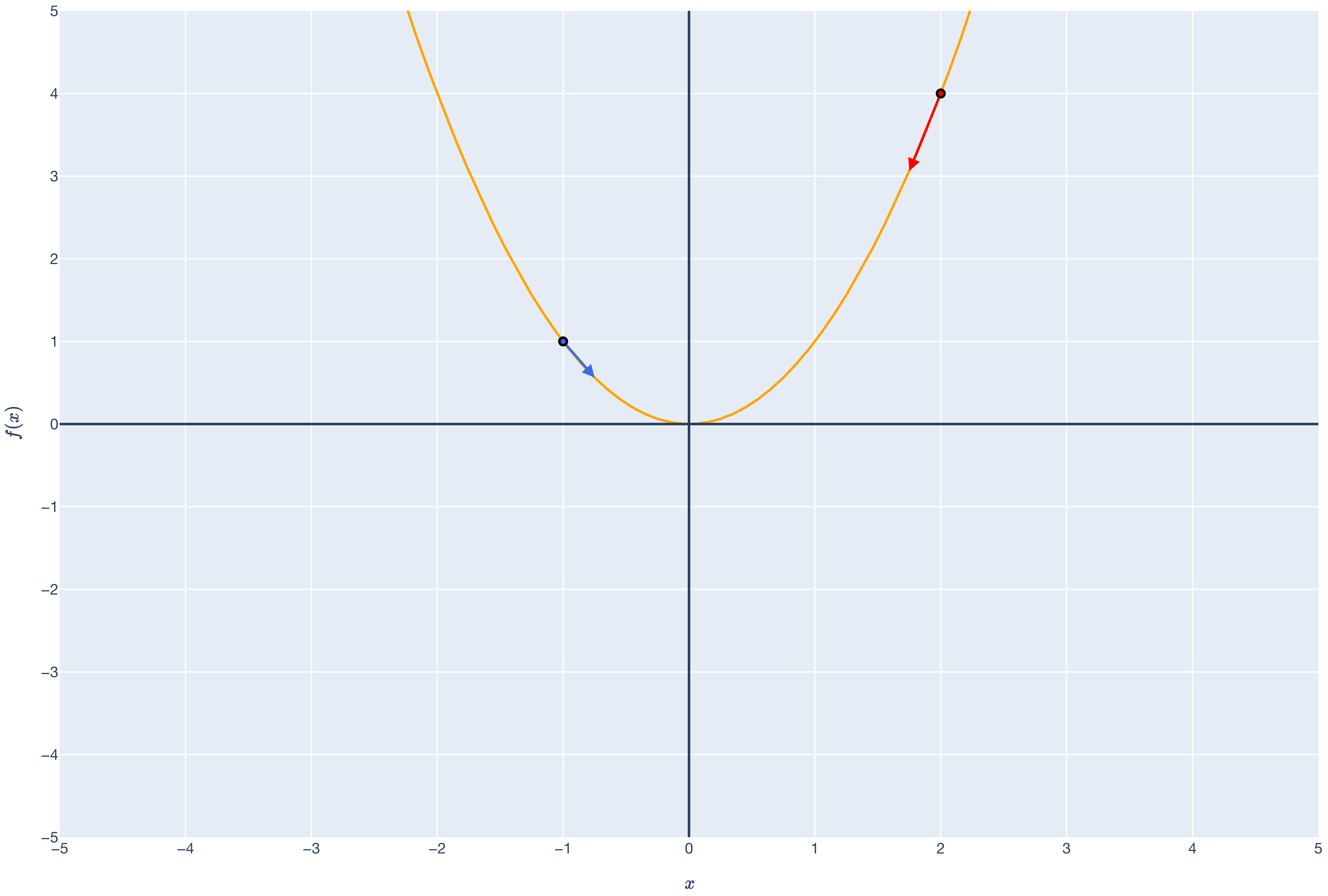


Gradient Descent

Preview of convexity

Solution: *Convex functions* are functions that "look like bowls."

These have nice properties, the main one being: *all local minima are global minima.*



Gradient Descent

Preview of convexity

Theorem (Convergence of GD for smooth, convex functions). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathscr{C}^2$, β-smooth, and convex function. Let $\mathbf{x}^*$ be a minimizer of f, i.e. $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

If we run gradient descent with step size $\eta = \dfrac{1}{\beta}$ and initial point $\mathbf{x}_0 \in \mathbb{R}^n$ for T iterations, we have:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right).$$

Recap

Lesson Overview

Linearization for approximation. We explore using the ***linearization*** of a function to approximate it. This is also called a “first-order approximation.”

Taylor series. We define the ***Taylor series*** of a function, which is an “infinite polynomial” that approximates a function at a point.

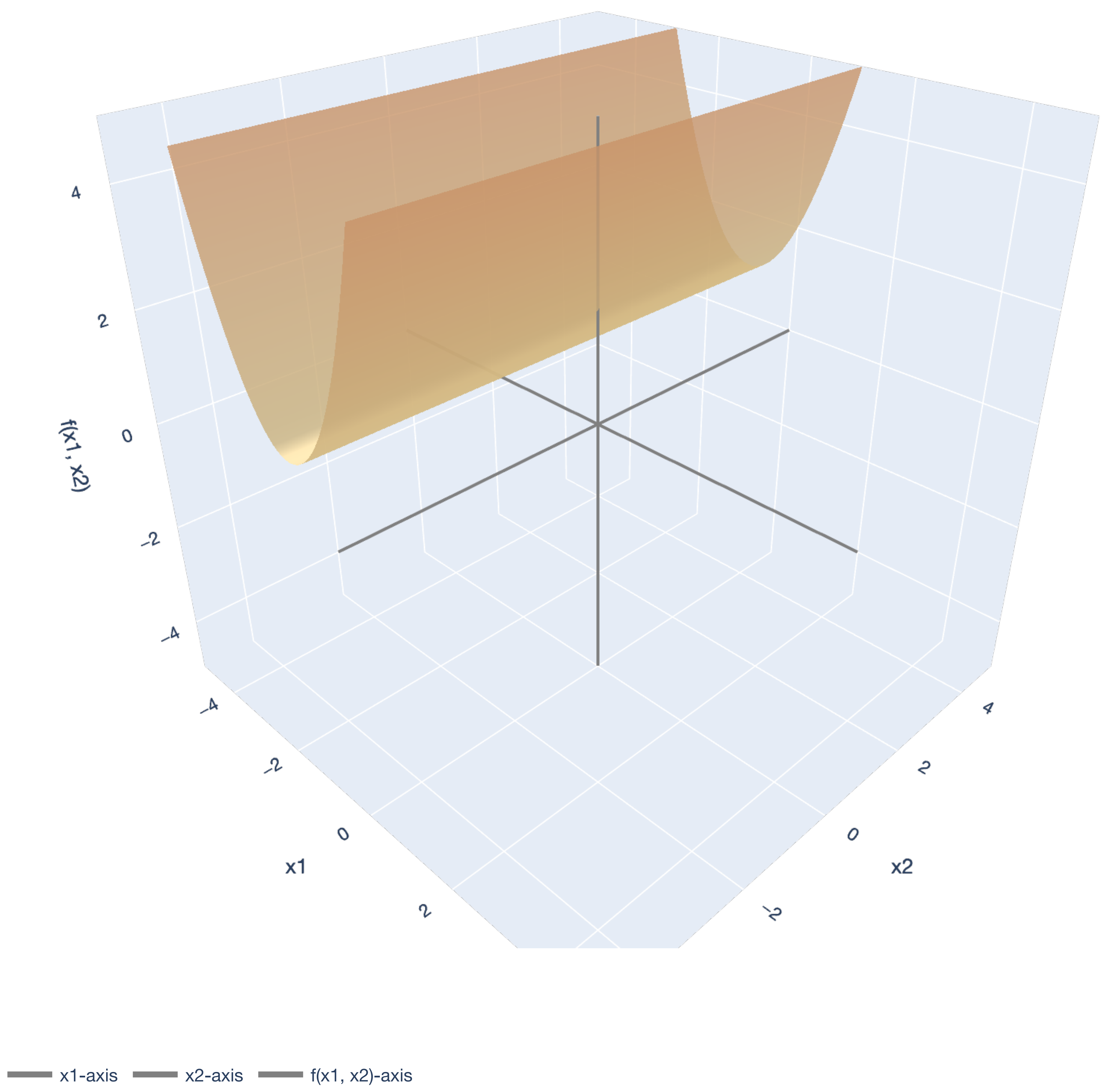
First-order and second-order Taylor approximation. The Taylor polynomial allows us to approximate a funciton by “chopping it off” at a certain degree.

Taylor’s Theorem. To quantify how bad our approximations are, we can use ***Taylor’s Theorem.*** We present two forms of Taylor’s Theorem (Peano and Lagrange).

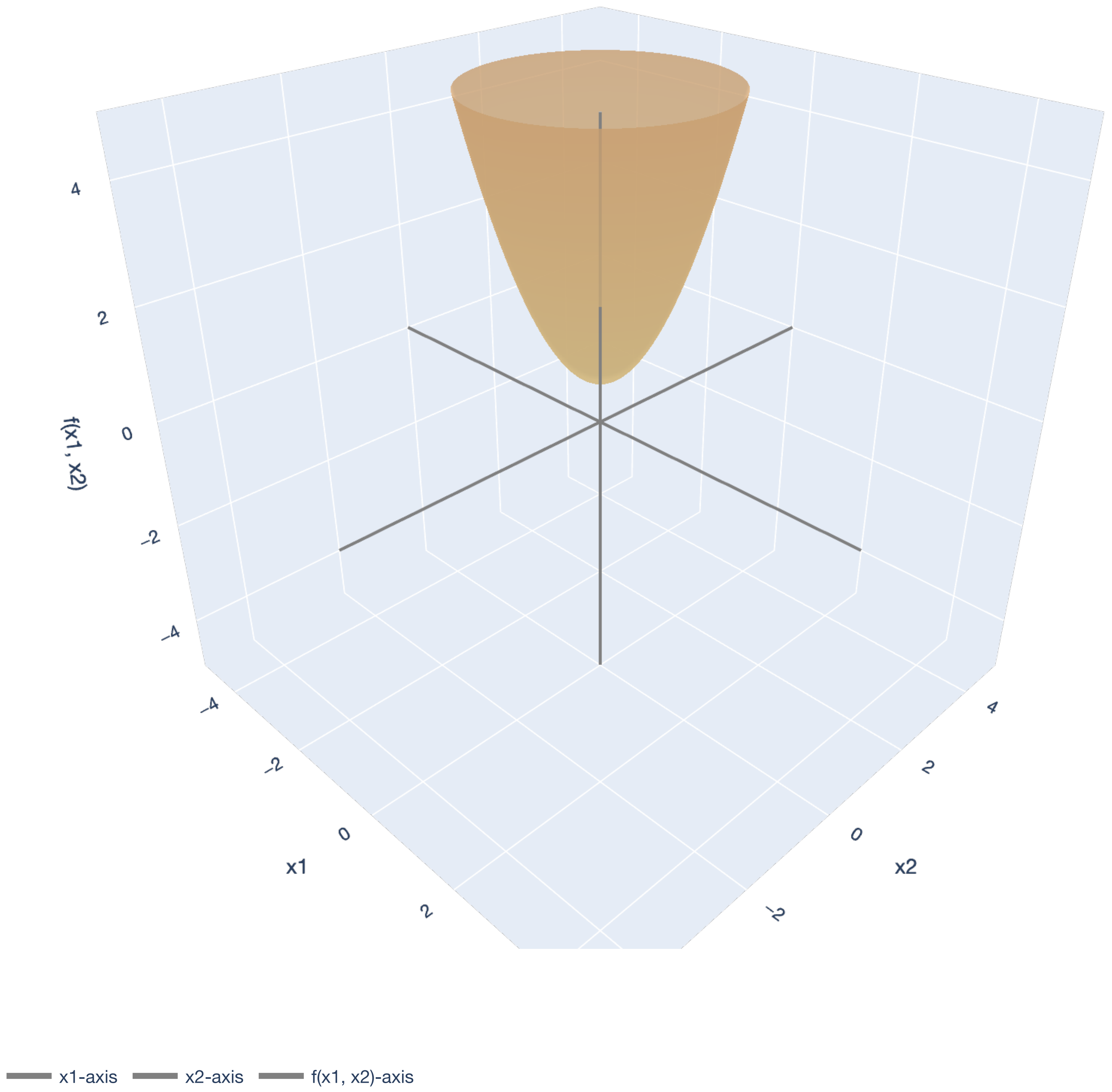
Gradient descent. We write down the full algorithm for ***gradient descent***, the second “story” of our course. Using Taylor’s Theorem, we can prove that, for ***β-smooth functions***, GD makes the function value smaller from iteration to iteration, as long as we set the “step size” small enough.

Lesson Overview

Big Picture: Least Squares



$$\lambda_1, \ldots, \lambda_d \geq 0$$



$$\lambda_1, \ldots, \lambda_d > 0$$

Lesson Overview

Big Picture: Gradient Descent