Math for Machine Learning

Week 3.2: Linearization, Gradient Descent, and Taylor Series

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.”

Gradient descent. We write down the full algorithm for gradient descent, the second “story” of our course. First, we prove the informal descent lemma. Then, we use Taylor series to formalize it.

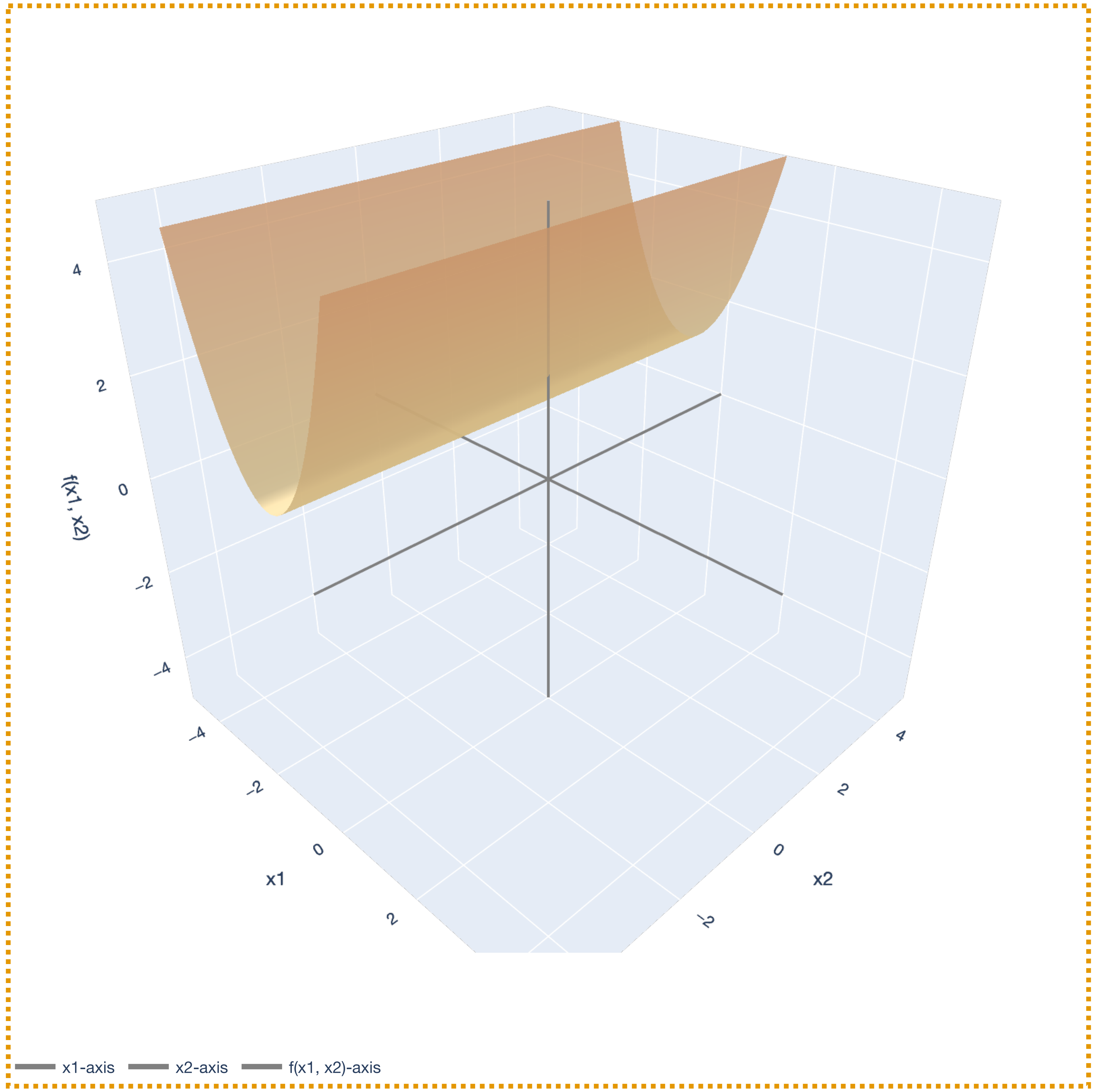
Taylor series. We define the Taylor series of a function, which is an “infinite polynomial” that approximates a function at a point.

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

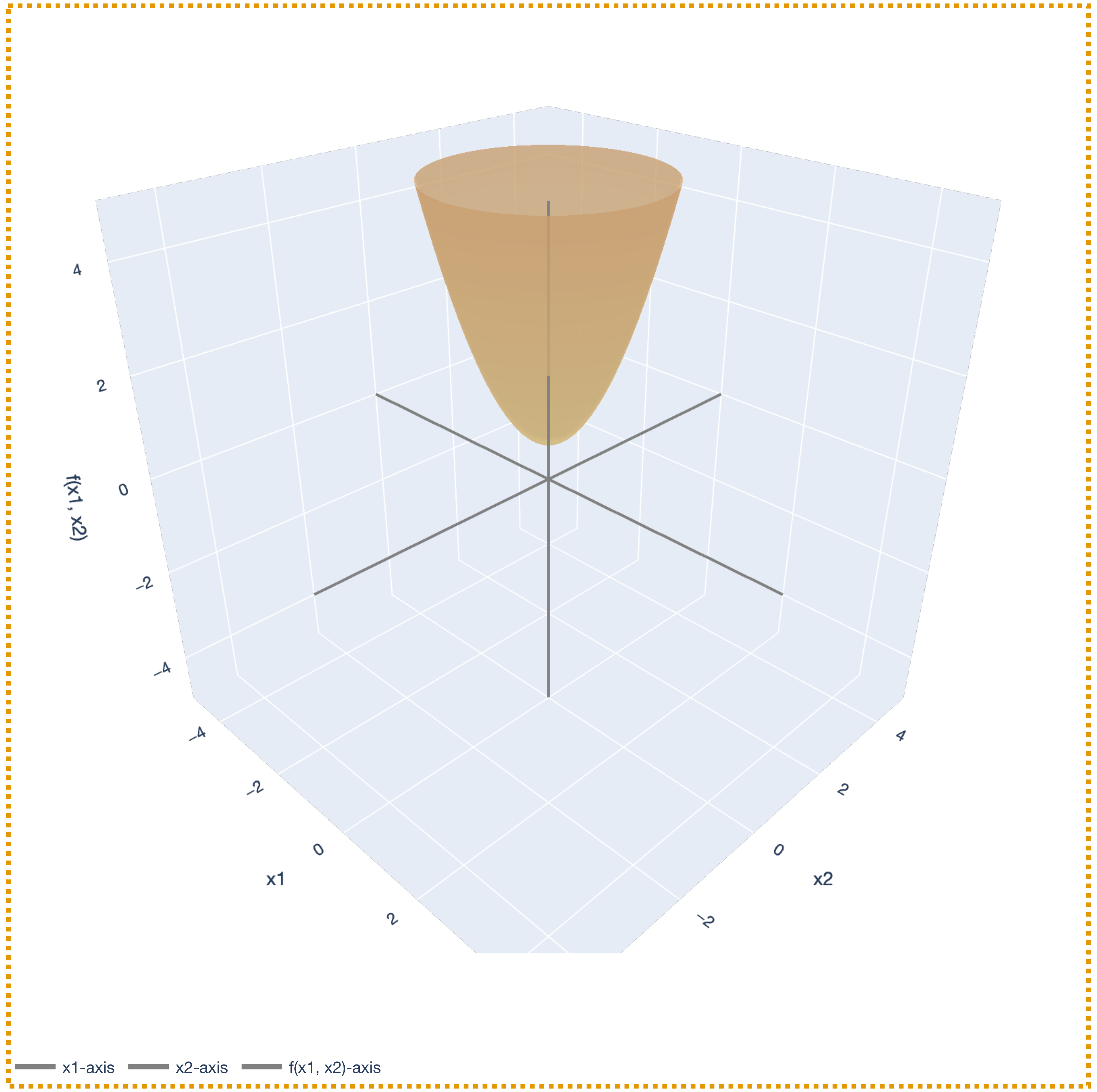
Taylor’s Theorem. To quantify how bad our approximations are, we can use Taylor’s Theorem.

Lesson Overview

Big Picture: Least Squares



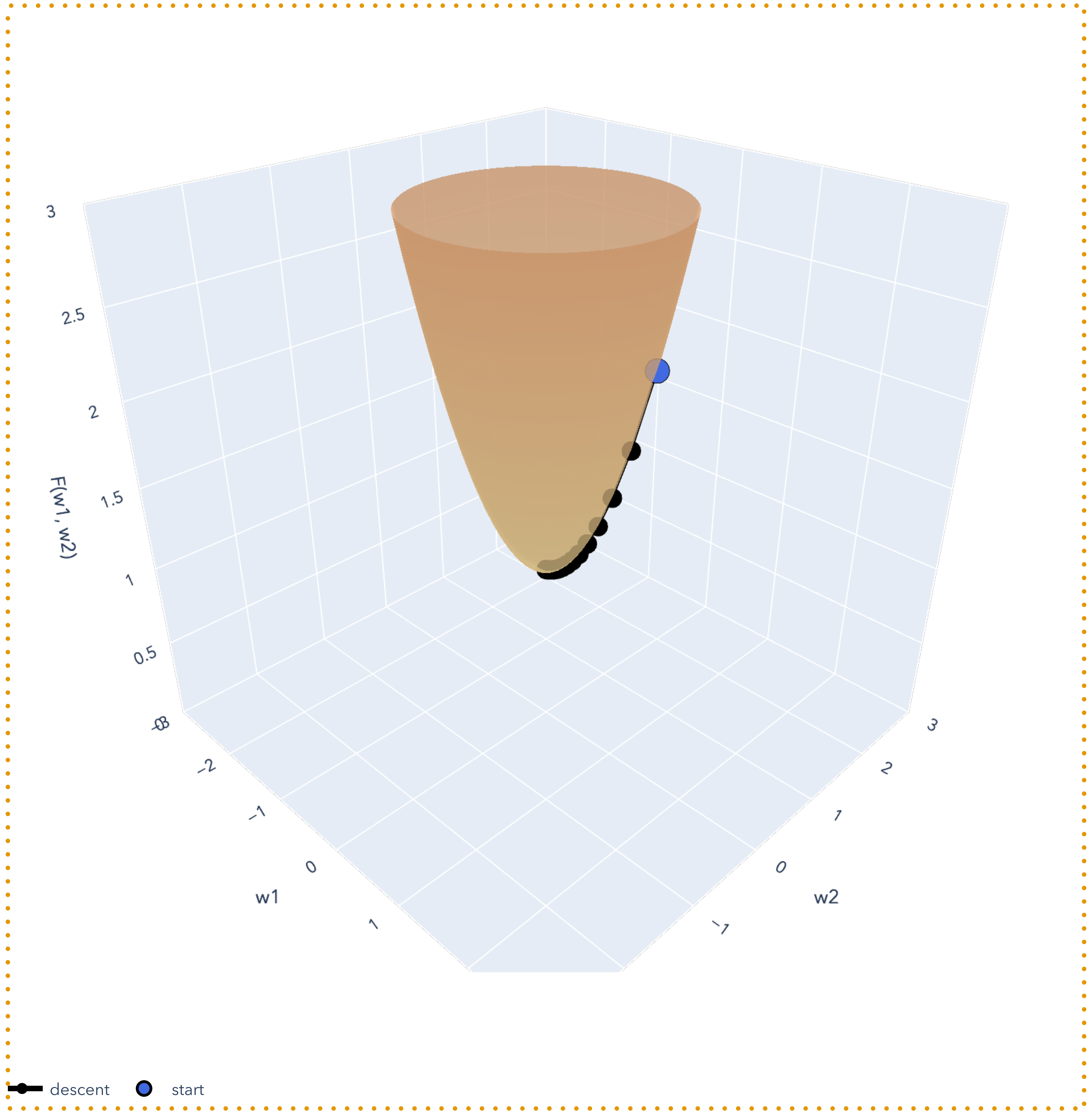
$\lambda_1, \ldots, \lambda_d \geq 0$

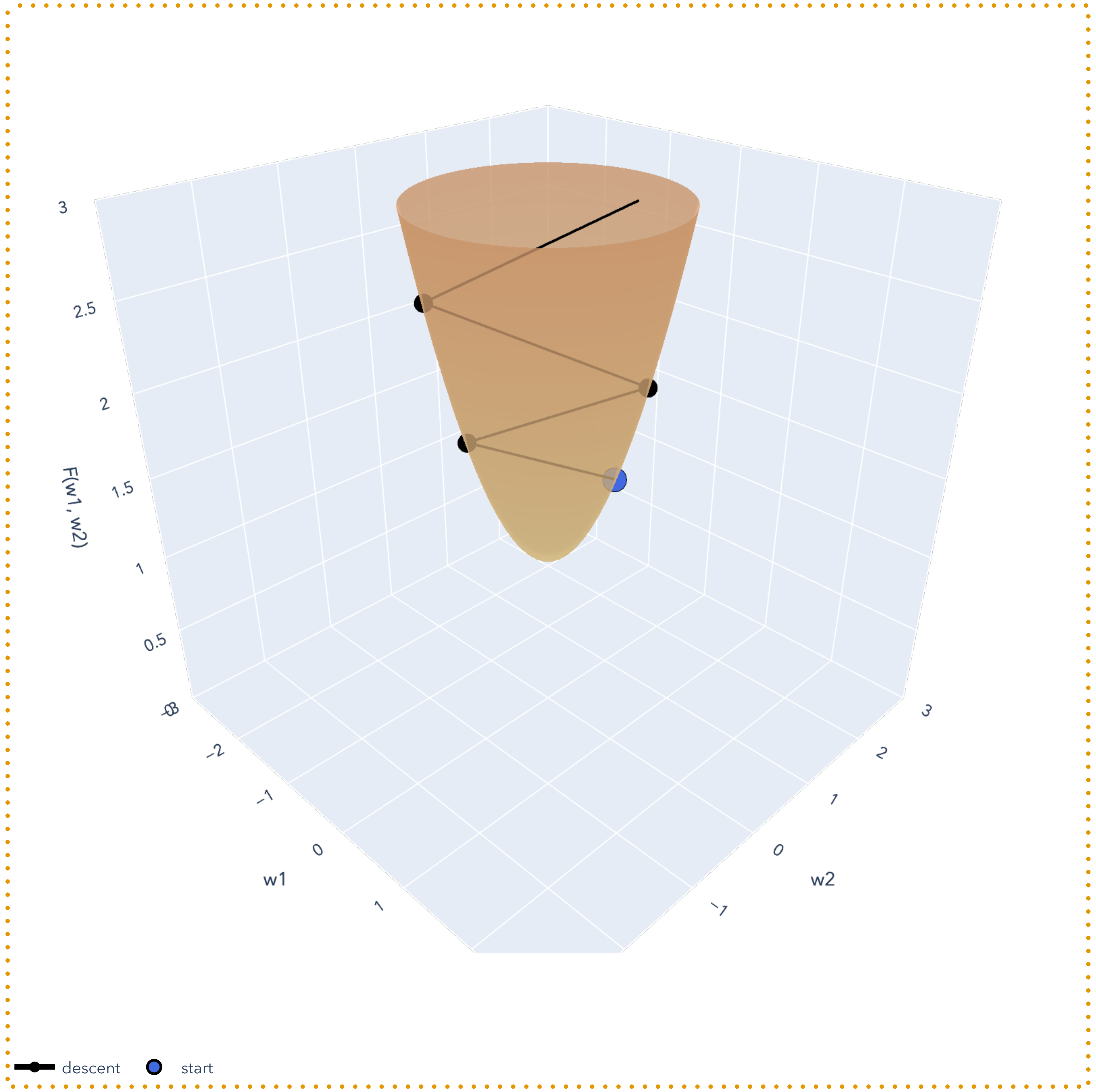


$\lambda_1, \ldots, \lambda_d > 0$

Lesson Overview

Big Picture: Gradient Descent





Linearization

Derivatives to find linear approximations

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

Given an objective function f, find the $\mathbf{w}$ *that makes* $f(\mathbf{w})$ *as small as possible.*

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(3,2,1,\ldots,0) = 48$$

Given an objective function f, find the $\mathbf{w}$ that makes $f(\mathbf{w})$ as small as possible.

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(1,1,1,\ldots,1) = 10.2$$

Given an objective function f, find the $\mathbf{w}$ *that makes* $f(\mathbf{w})$ *as small as possible.*

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(-3, 1, 0, \ldots, 1) = 0.24$$

Given an objective function f, find the $\mathbf{w}$ that makes $f(\mathbf{w})$ as small as possible.

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

Given an objective function f, find the $\mathbf{w}$ that makes $f(\mathbf{w})$ as small as possible.

Assume: $\mathbf{w} \in \mathbb{R}^d$ is unconstrained.

Optimization Problem

Review: Basic Goal

In much of machine learning, we solve well-defined *optimization problems.*

Goal: minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

Given an objective function f, find the $\mathbf{w}$ that makes $f(\mathbf{w})$ as small as possible.

Assume: $\mathbf{w} \in \mathbb{R}^d$ is unconstrained.

Assume: $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

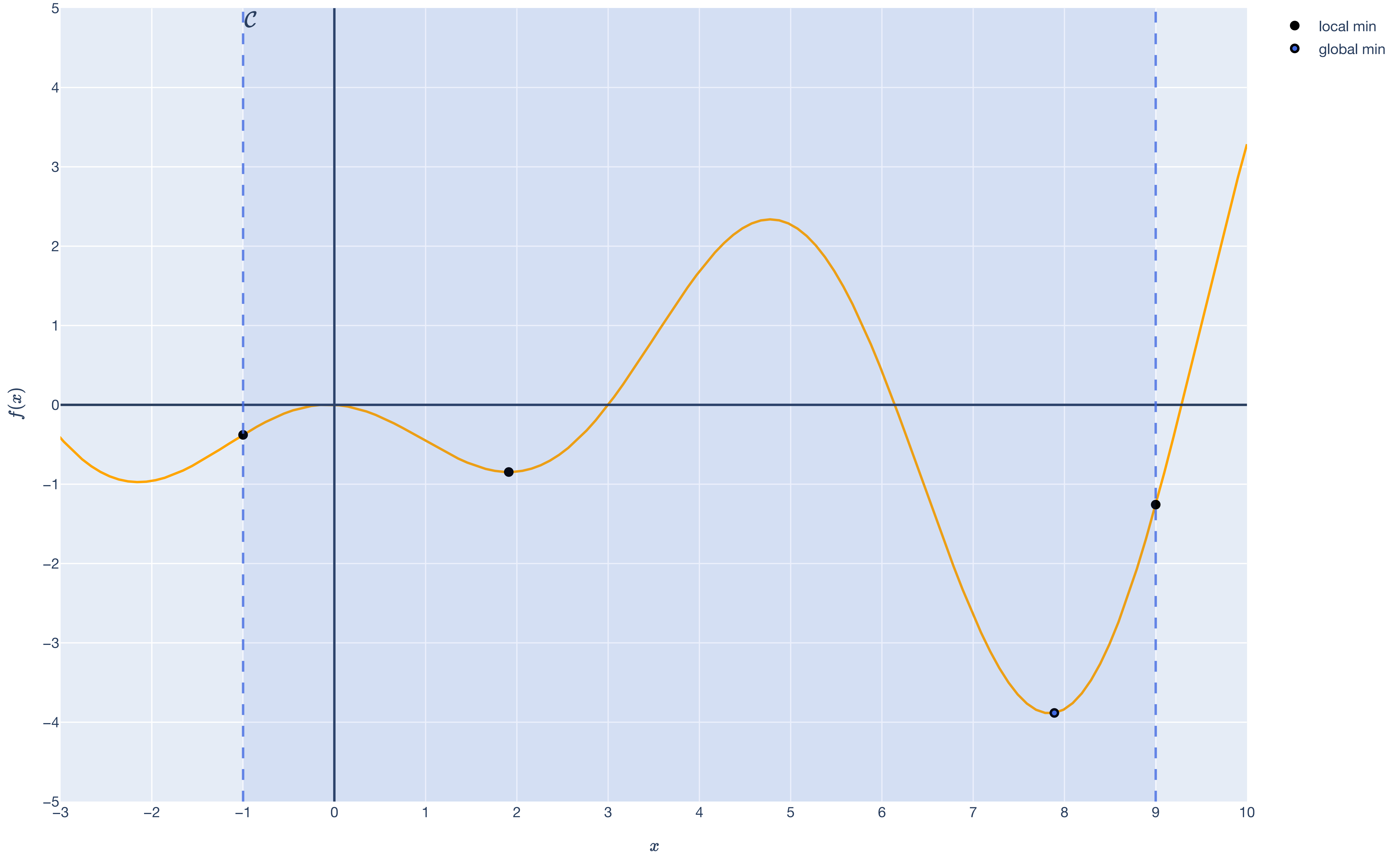
Motivation

Optimization in single-variable calculus

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

Intermediary goal: Find the *local minima*.

Derivatives will give us descent directions!

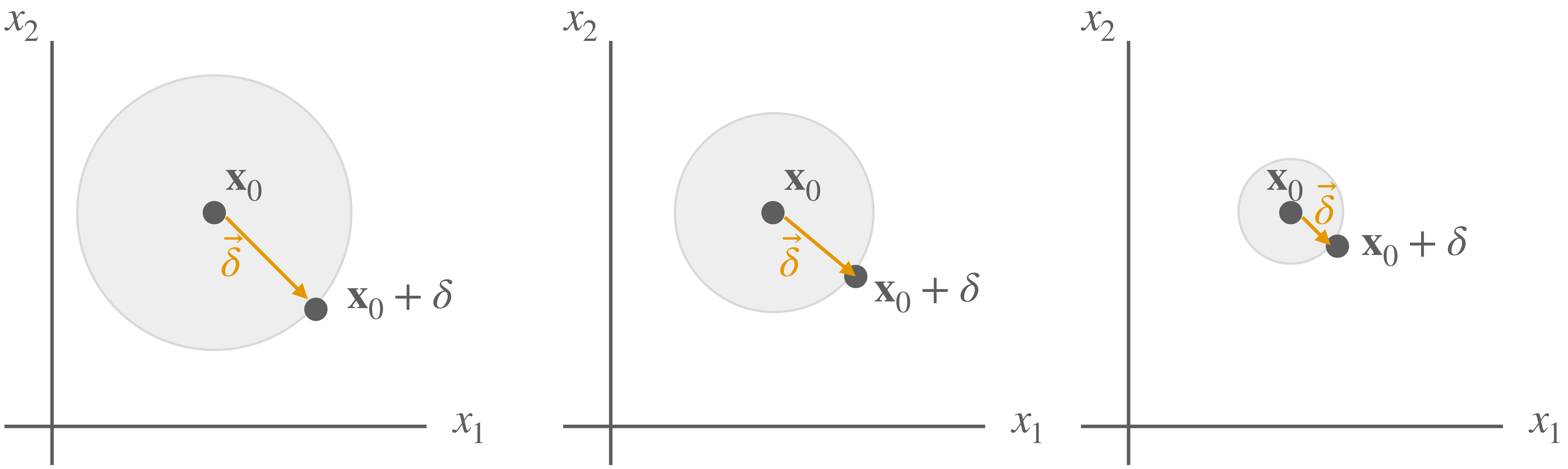


Multivariable Differentiation

Total Derivative for $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{\vec{\delta} \to 0} \frac{1}{||\vec{\delta}||} \left(\left(f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0) \right) - Df_{\mathbf{x}_0}(\vec{\delta}) \right) = 0,$$

Approaching $\mathbf{x}_0$ from any direction $\vec{\delta}$, the change $f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0)$ is approximated by $Df_{\mathbf{x}_0}$.



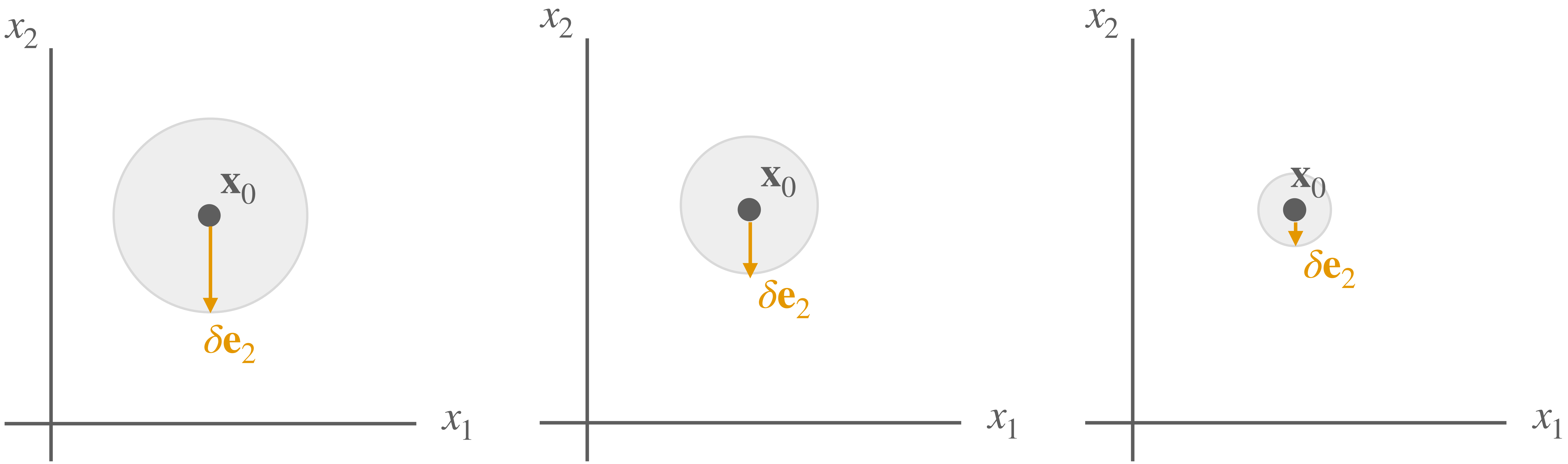
Multivariable Differentiation

Partial Derivative

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

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

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



Multivariable Differentiation

Gradient

Let $f : \mathbb{R}^d \to \mathbb{R}$. The gradient of f at $\mathbf{x}_0$ is the vector $\nabla f(\mathbf{x}_0) \in \mathbb{R}^d$ composed of all the partial derivatives of f at $\mathbf{x}_0$:

$$\nabla f(\mathbf{x}_0) := \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}_0) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}_0) \end{bmatrix}$$

Slogan: Derivatives are linear transformations

Linearity and differentiation

The derivative is a linear transformation that maps changes in $\mathbf{x}$ to changes in f.

For $f: \mathbb{R}^d \to \mathbb{R}$, a scalar-valued function…

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

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

equivalent to:

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

An *affine function* that approximates f.

Differential Calculus

Review: Derivative

If $f : \mathbb{R}^d \to \mathbb{R}$ is *differentiable* at $\mathbf{x}_0 \in \mathbb{R}^d$…

$$\lim_{\vec{\delta} \to 0} \frac{1}{\|\vec{\delta}\|} \left(\left(f(\mathbf{x}_0 + \vec{\delta}) - f(\mathbf{x}_0) \right) - Df_{\mathbf{x}_0}(\vec{\delta}) \right) = 0$$

is equivalent to:

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

Differential Calculus

Review: Derivative

at the point where we're taking derivative…

If $f : \mathbb{R}^d \to \mathbb{R}$ is *differentiable* at $\mathbf{x}_0 \in \mathbb{R}^d$…

linear approximation

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

as $\mathbf{x}$ gets closer to $\mathbf{x}_0$… …the function is closer and closer to its linear approximation!

The linear approximation of f at $\mathbf{x}_0$ is the function:

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

One use of differential calculus: *Analyze nonlinear functions with their linear approximations!*

Differential Calculus

Review: Derivative

at the point where we're taking derivative…

If $f : \mathbb{R}^d \to \mathbb{R}$ is *differentiable* at $\mathbf{x}_0 \in \mathbb{R}^d$…

linear approximation

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

as $\mathbf{x}$ gets closer to $\mathbf{x}_0$… …the function is closer and closer to its linear approximation!

One use of differential calculus: *Analyze nonlinear functions with their linear approximations!*

At any point $\mathbf{x}_0 \in \mathbb{R}^d$, $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

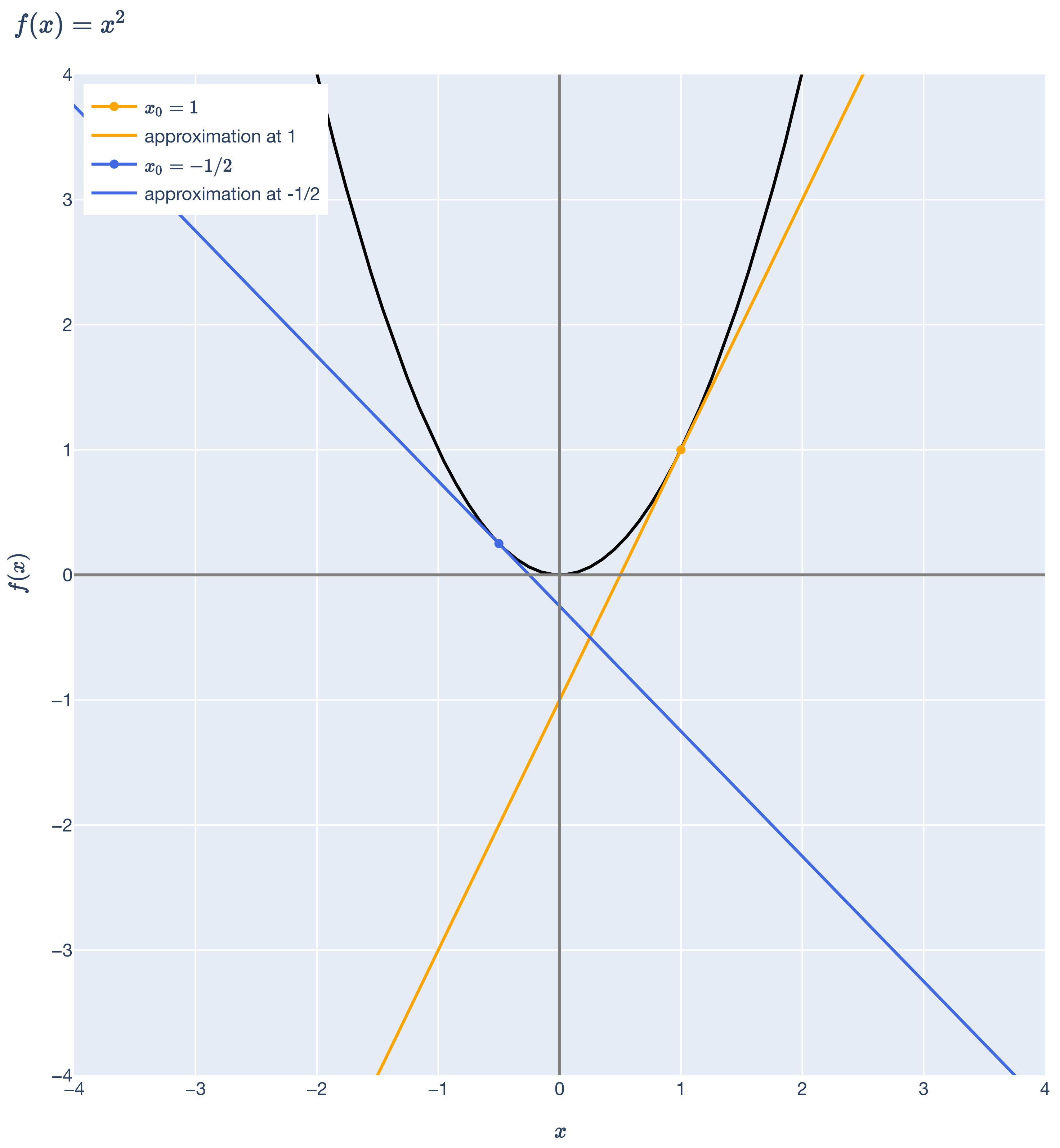
Our main slogan

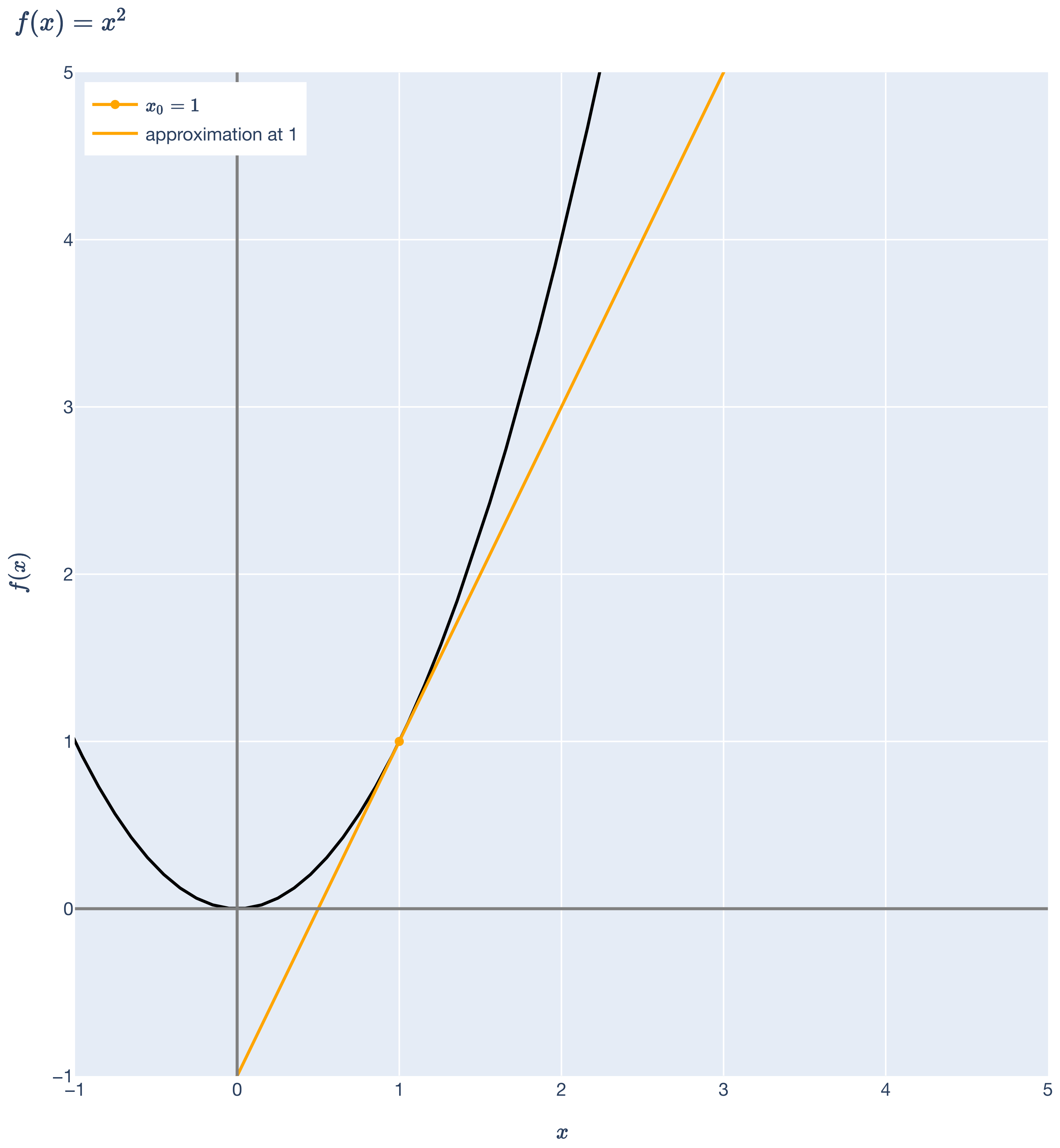
At any point $\mathbf{x}_0 \in \mathbb{R}^d$, $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

Our main slogan

At any point $\mathbf{x}_0 \in \mathbb{R}^d$, $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

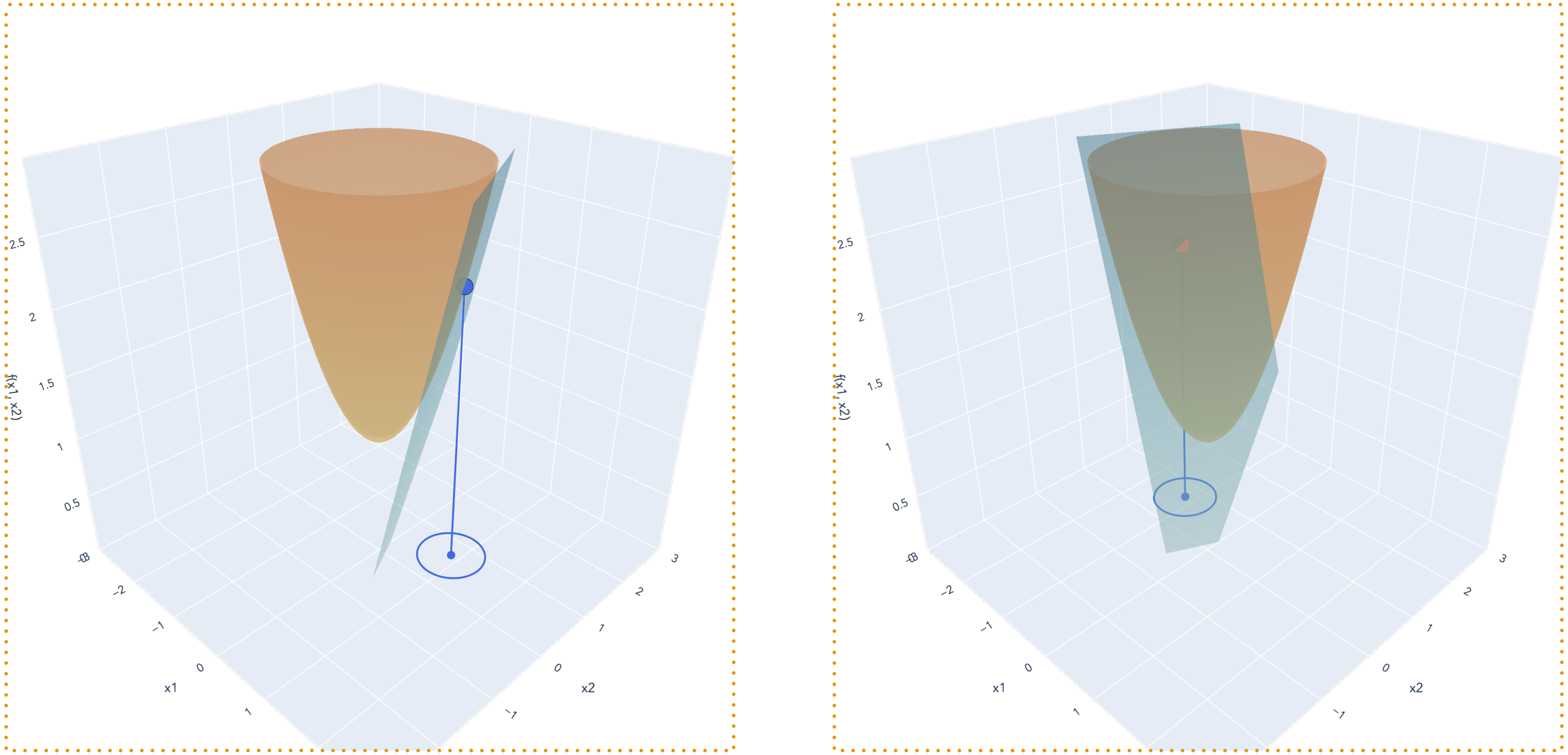




Linear Approximations

Our main slogan

At any point $\mathbf{x}_0 \in \mathbb{R}^d$, $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

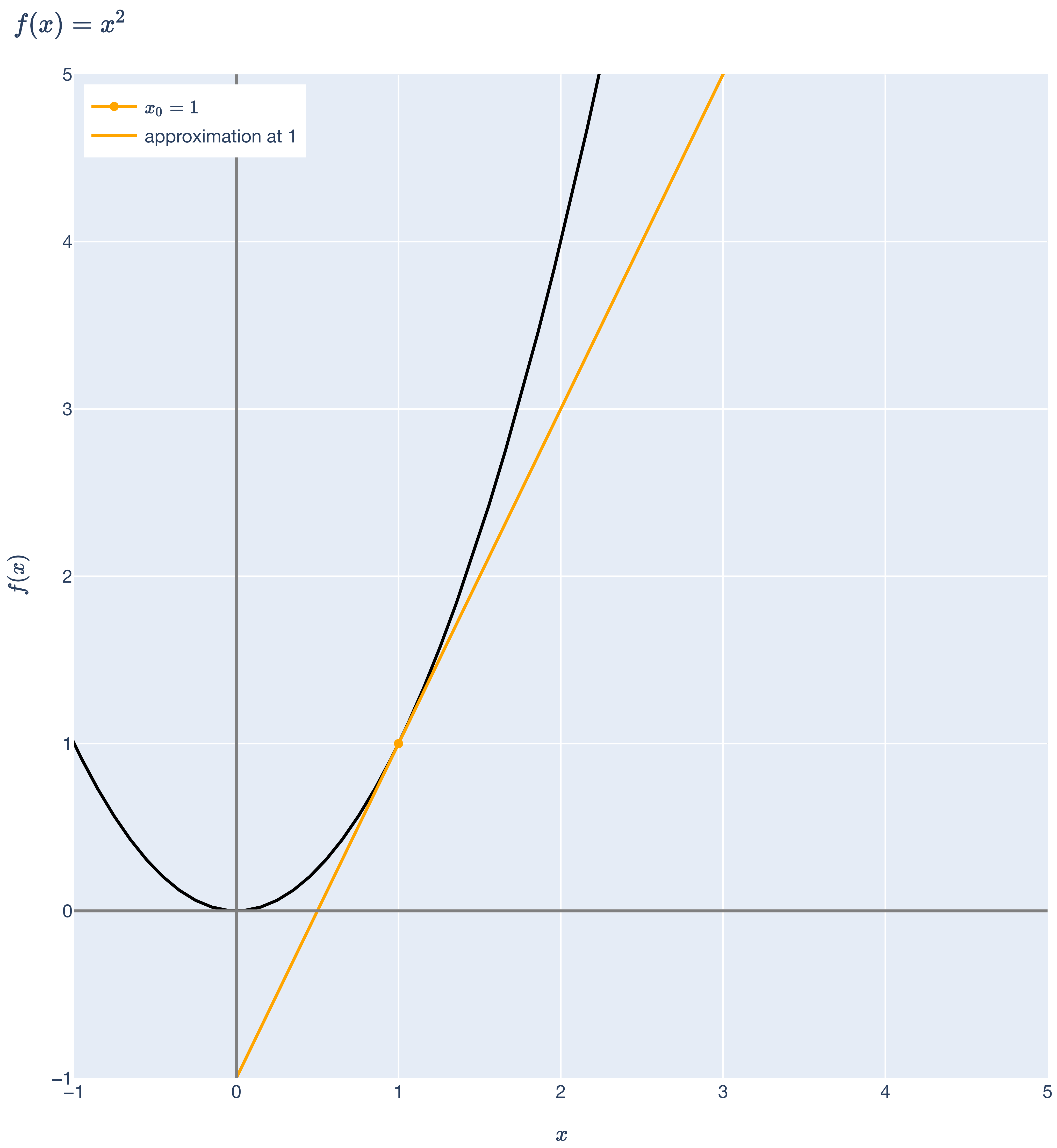


Linear Approximations

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

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

What is the linear approximation?



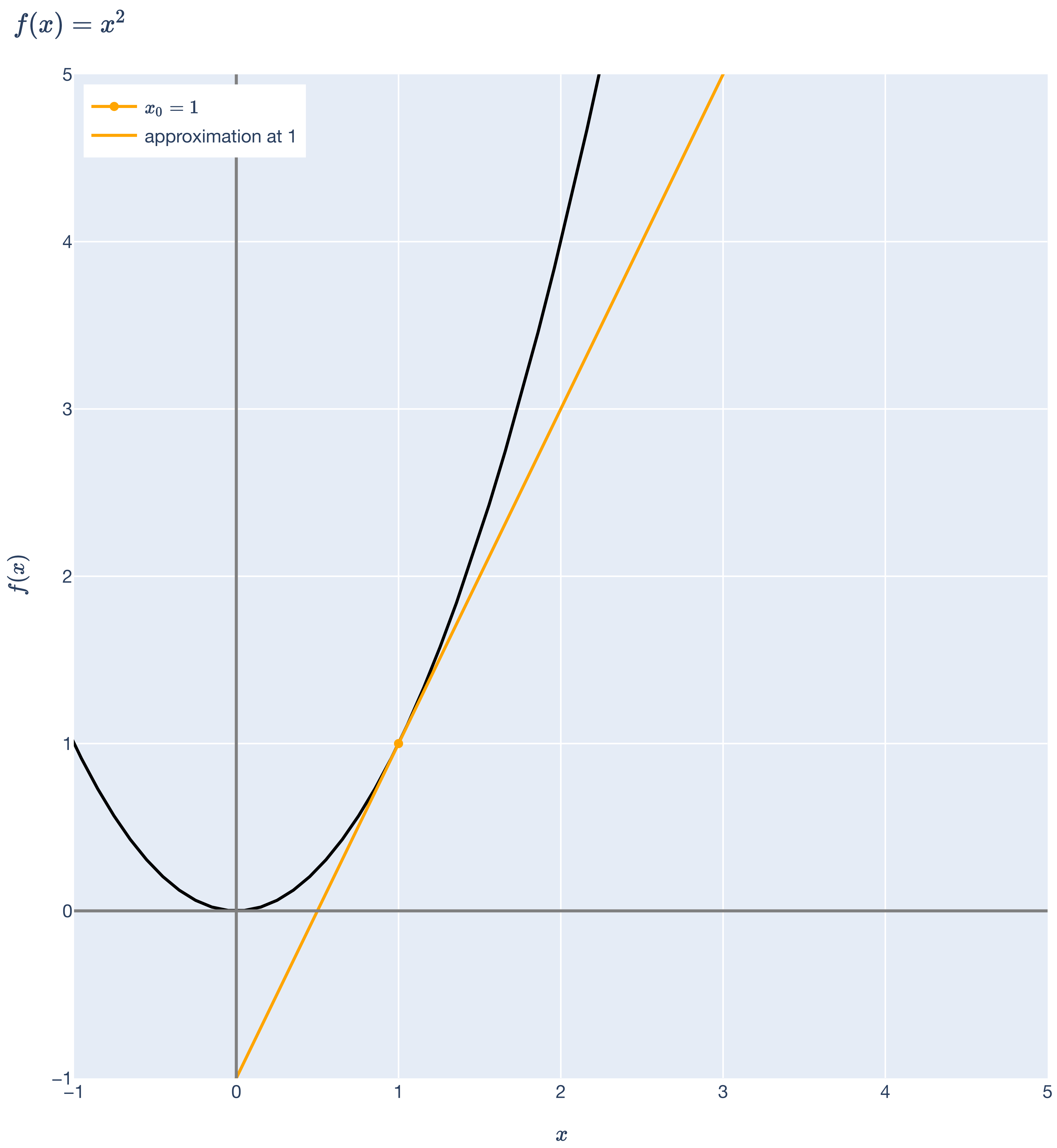
$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

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

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

What is the linear approximation?



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

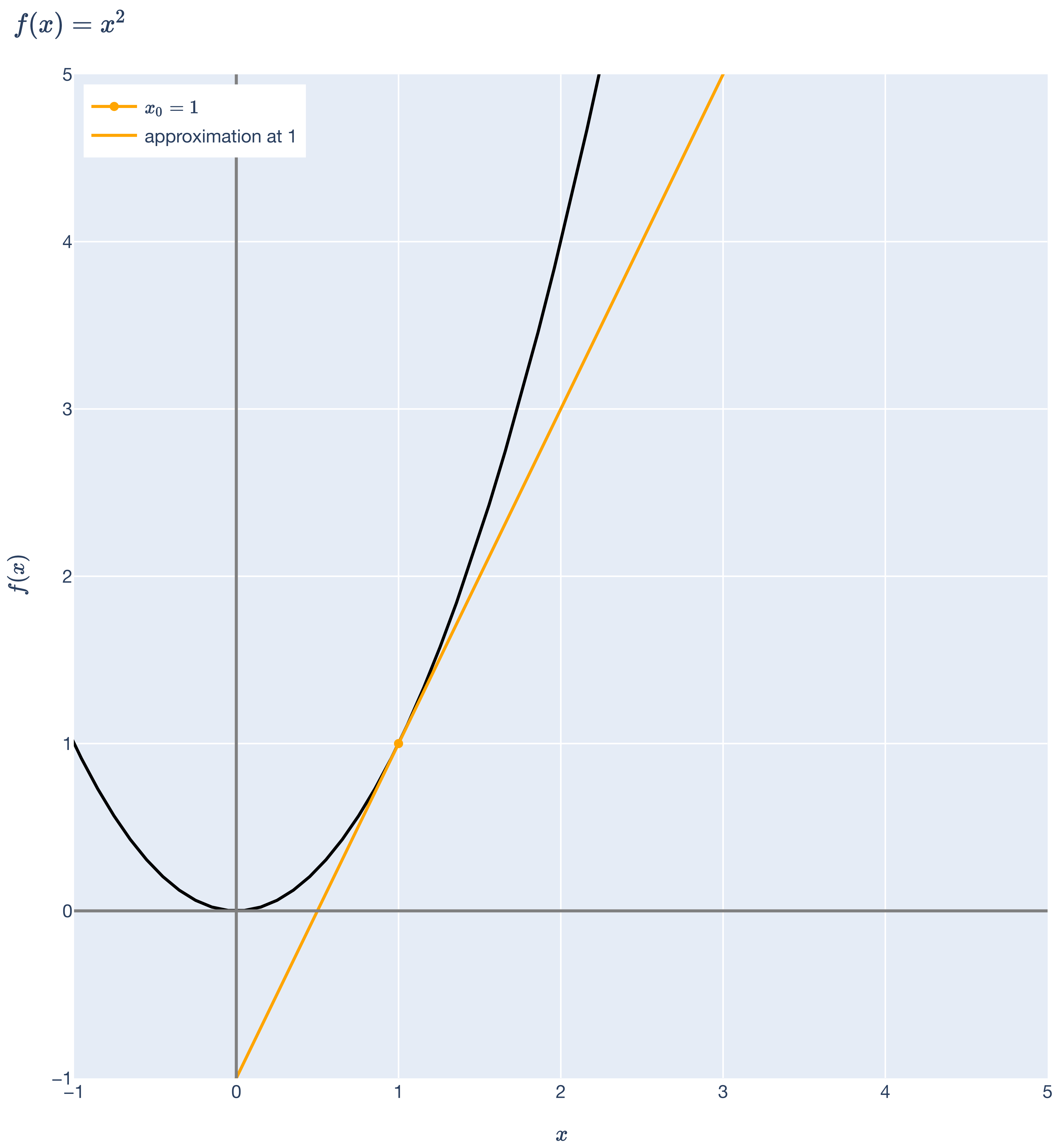
Linear Approximations

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

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

What is the linear approximation?

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



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

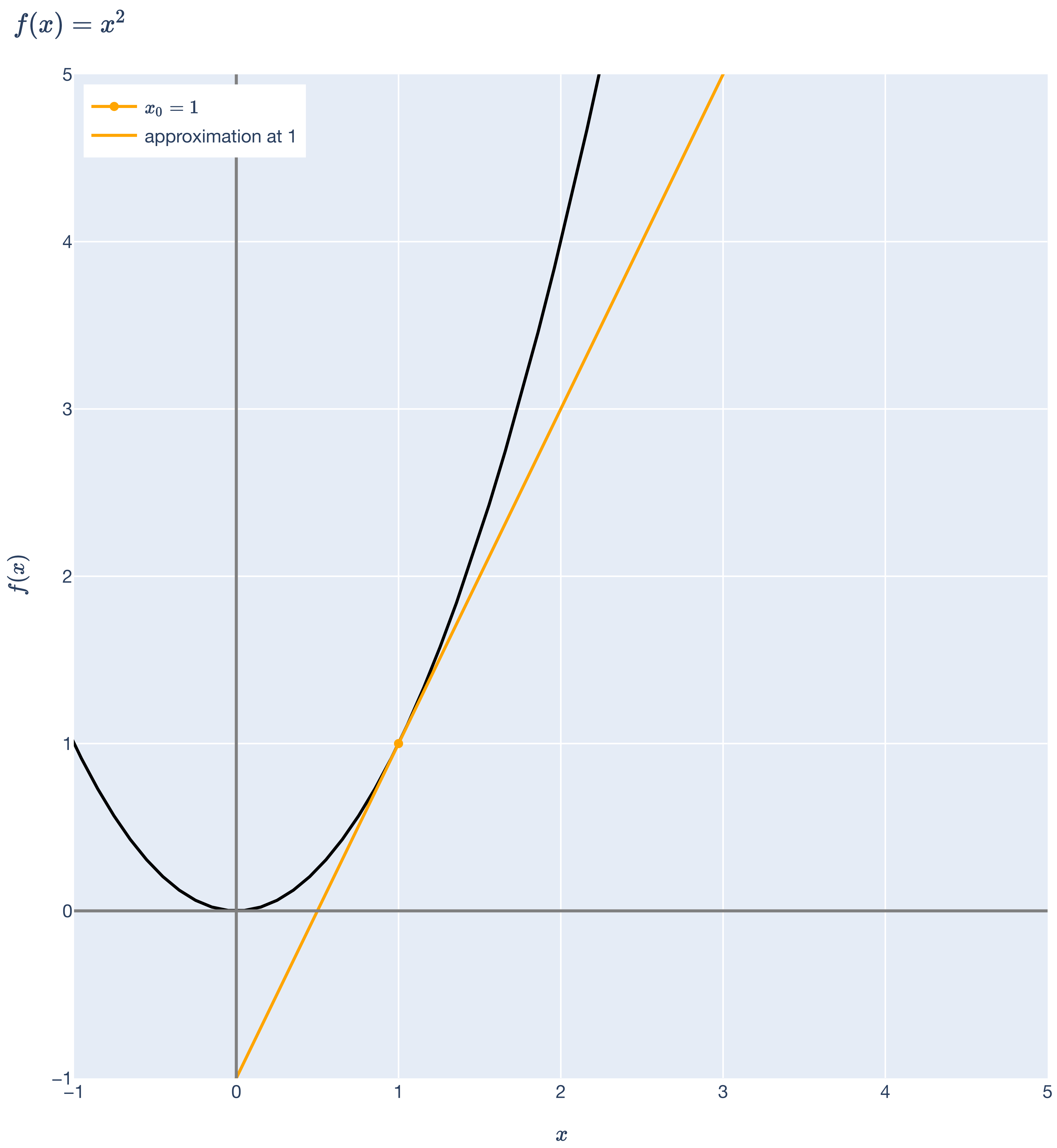
Linear Approximations

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

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

What is the linear approximation?

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



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

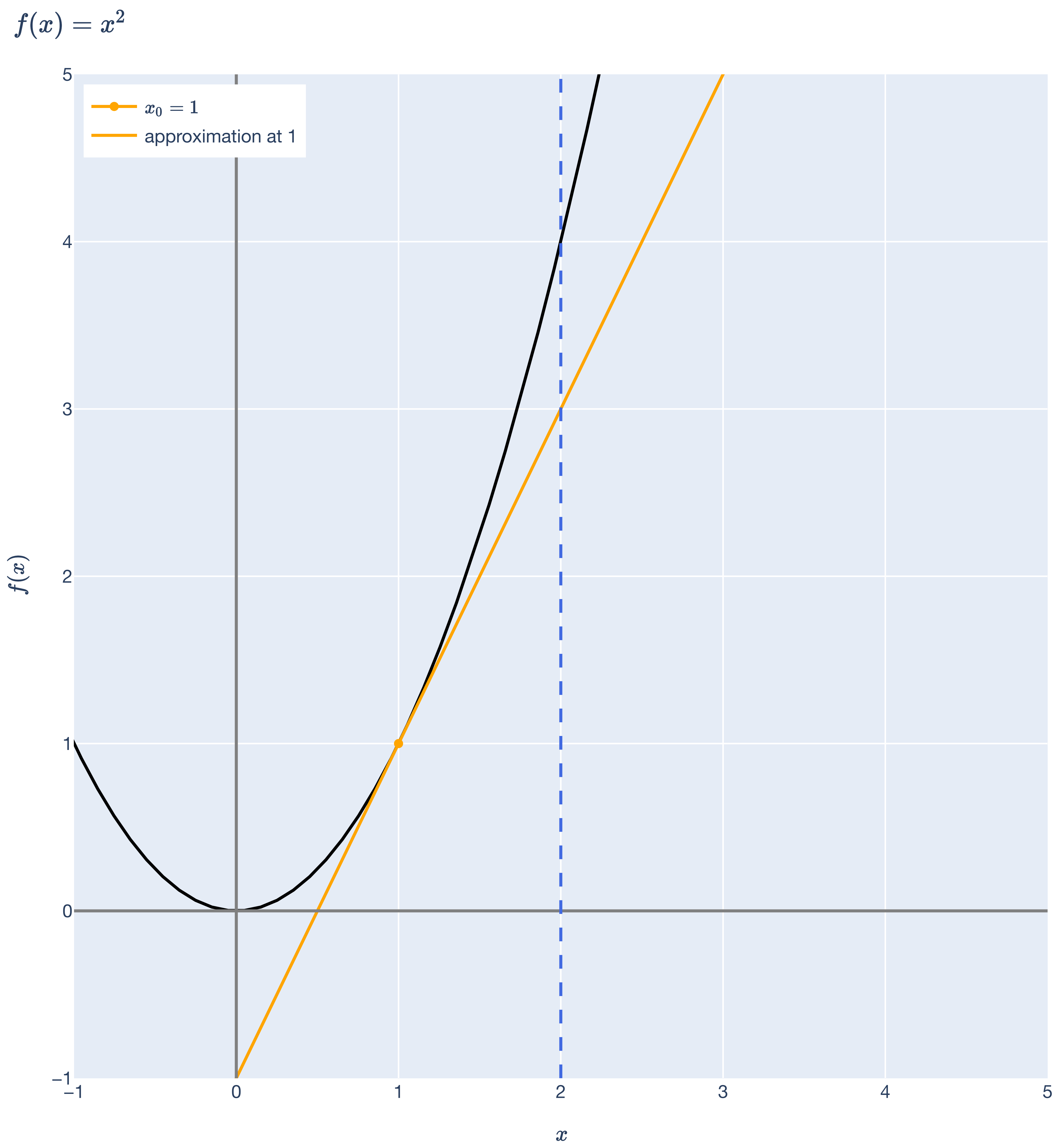
Example: $f : \mathbb{R} \to \mathbb{R}$

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

What is the linear approximation?

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

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



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

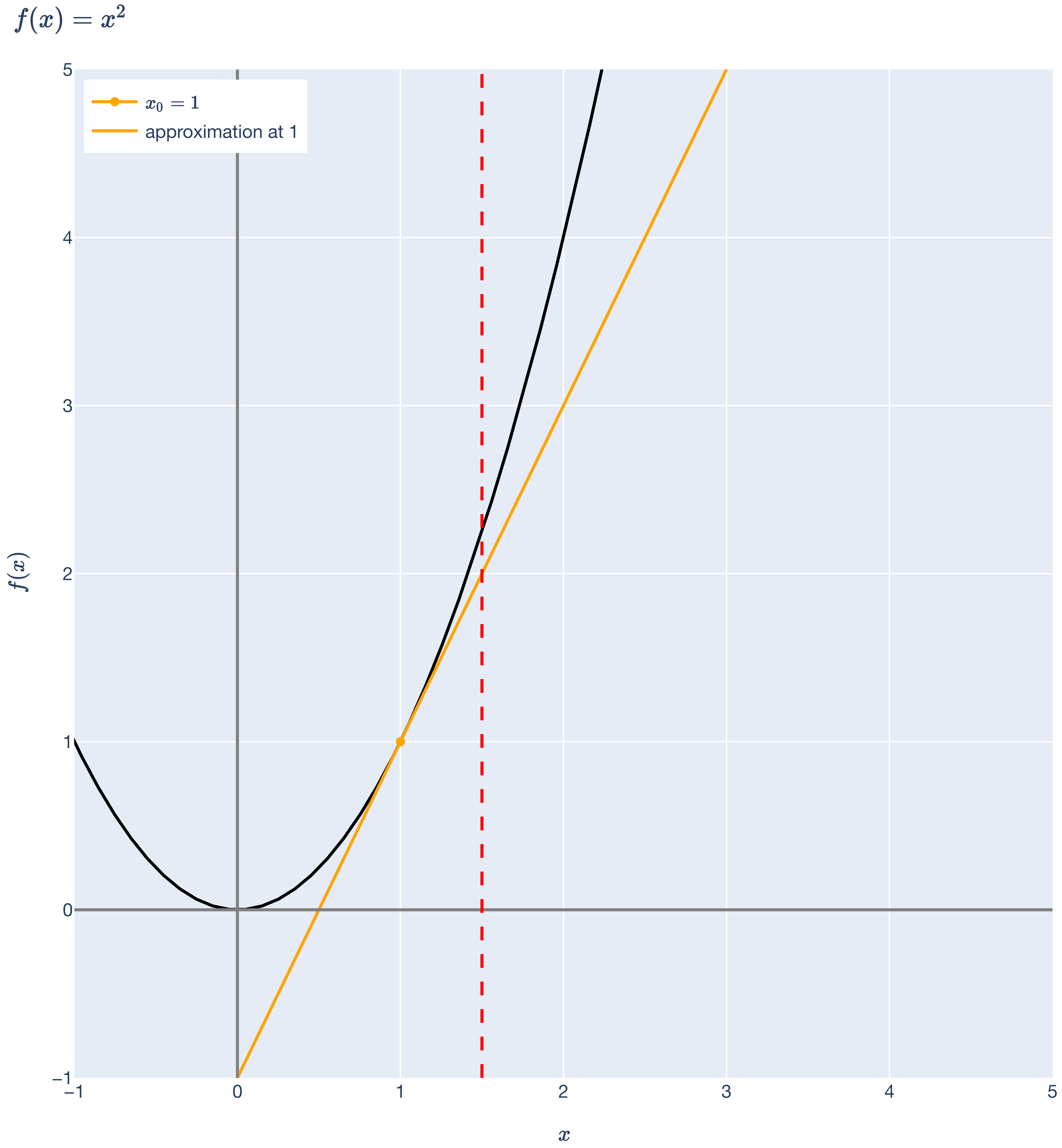
Example: $f : \mathbb{R} \to \mathbb{R}$

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

What is the linear approximation?

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

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



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

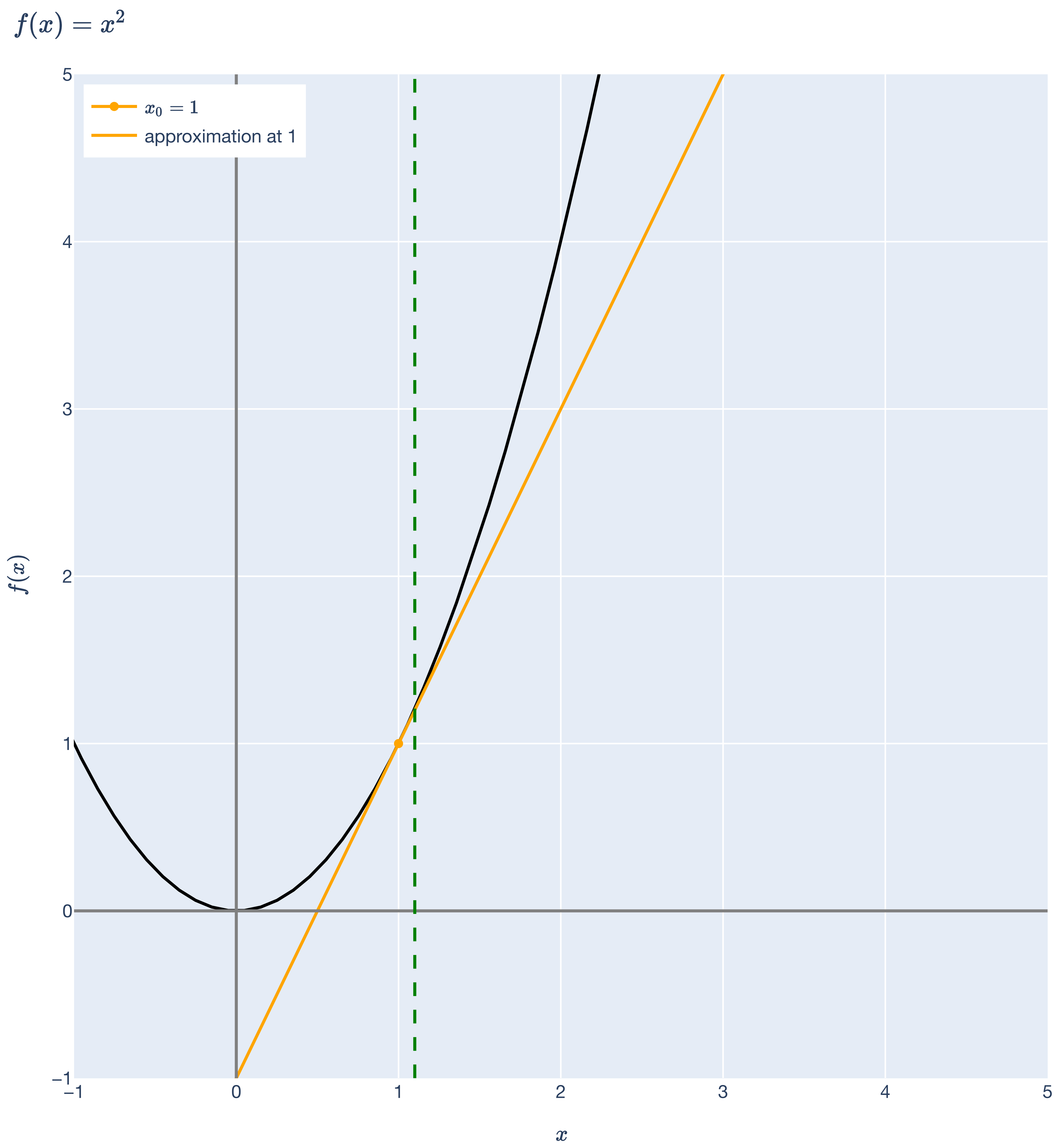
Example: $f : \mathbb{R} \to \mathbb{R}$

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

What is the linear approximation?

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

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



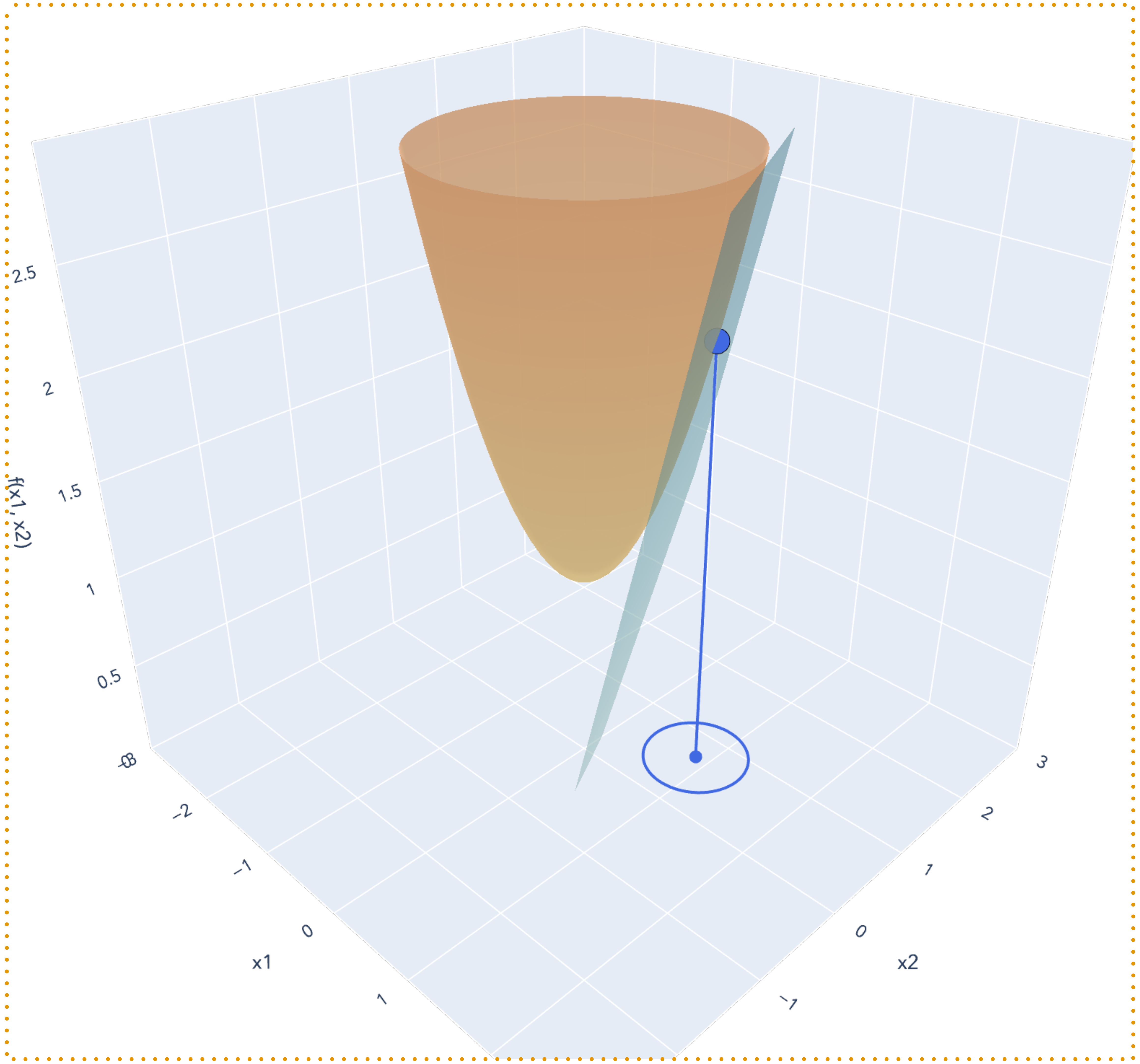
$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

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

$$F(x_1, x_2) = x_1^2 + x_2^2 + 1 \text{ at } \mathbf{x}_0 = (1,\ 0.5)$$

What is the linear approximation?



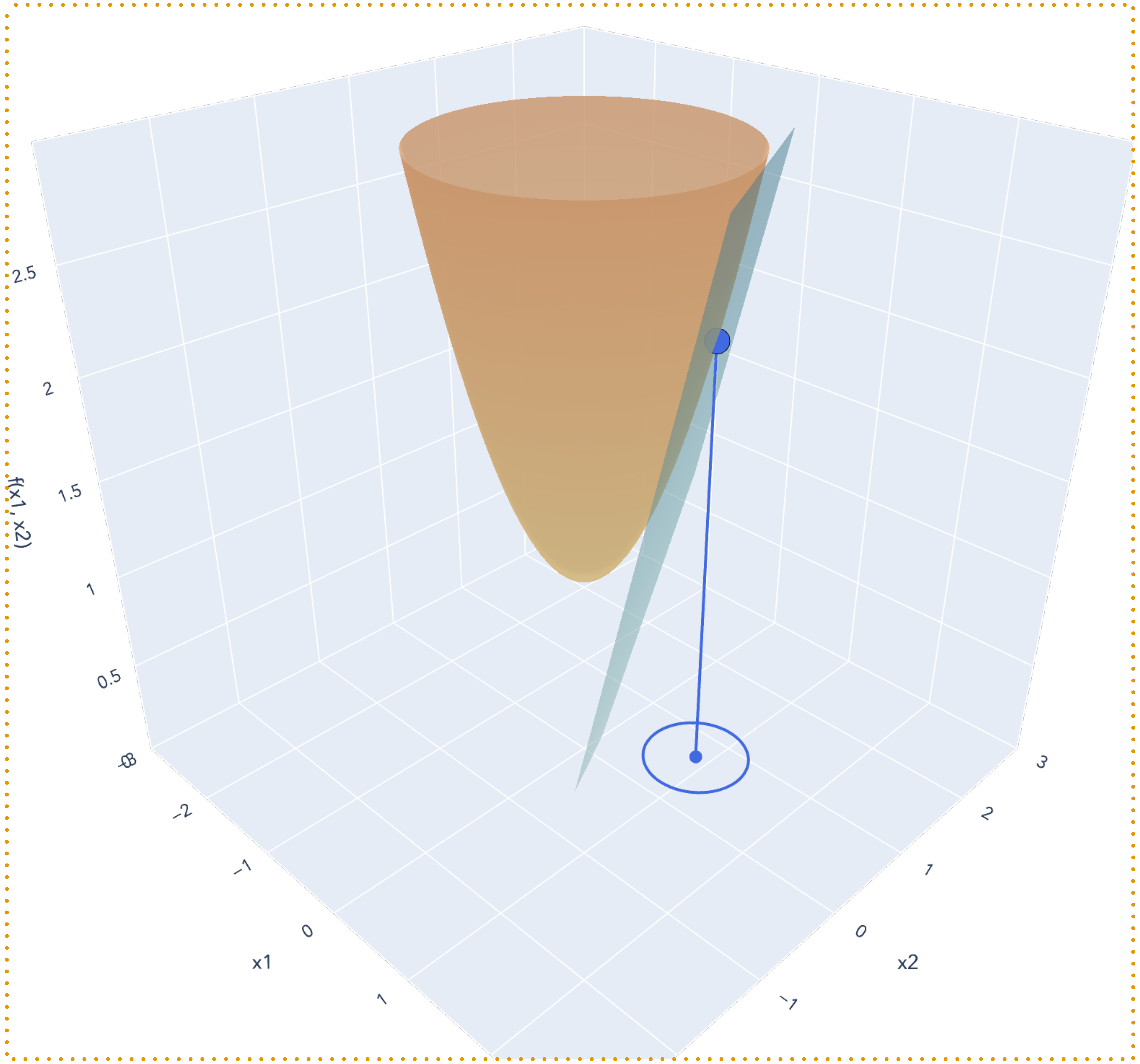
$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

Linear Approximations

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

$$F(x_1, x_2) = x_1^2 + x_2^2 + 1 \text{ at } \mathbf{x}_0 = (1,\ 0.5)$$

What is the linear approximation?



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

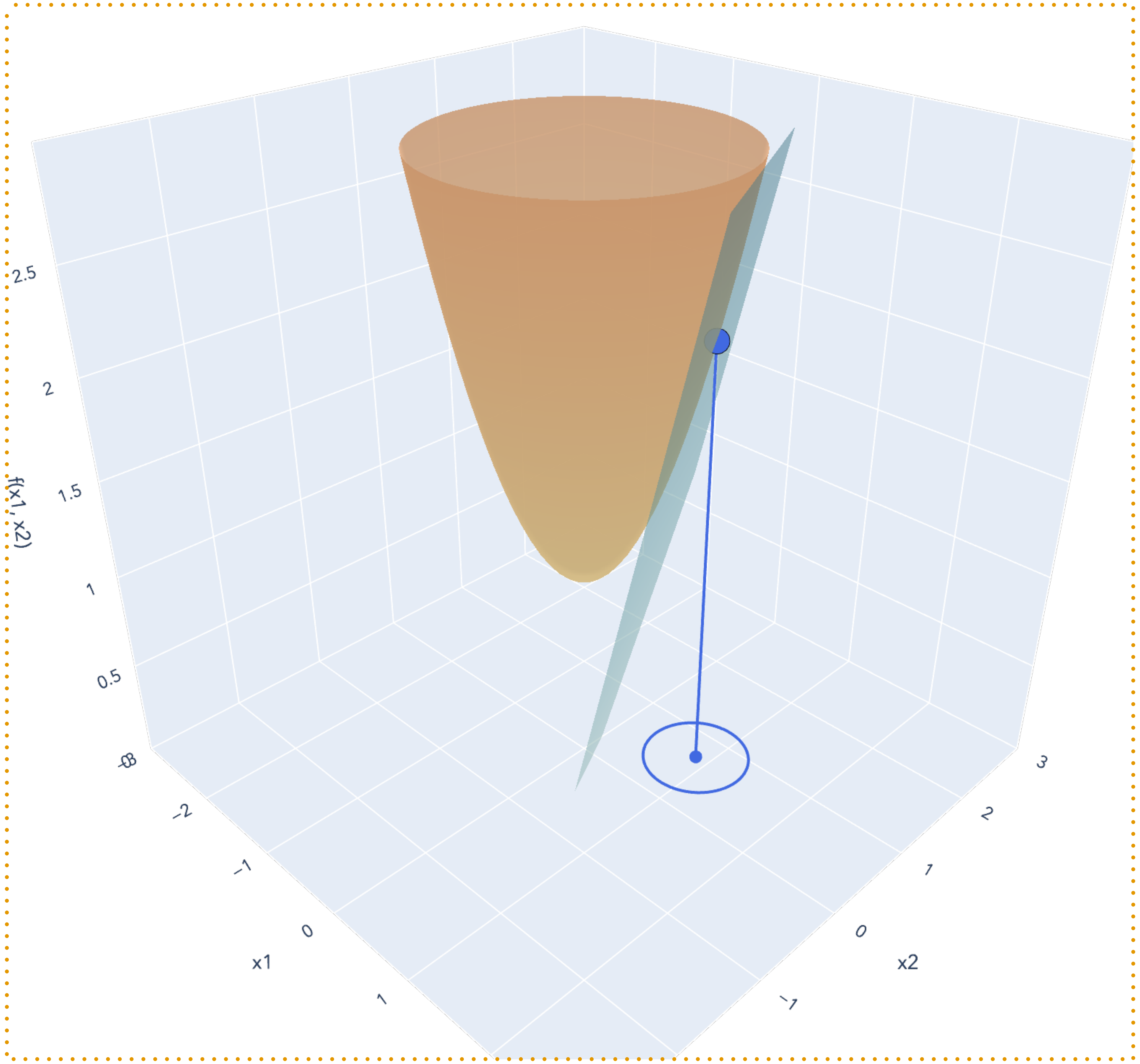
Linear Approximations

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

$$F(x_1, x_2) = x_1^2 + x_2^2 + 1 \text{ at } \mathbf{x}_0 = (1,\ 0.5)$$

What is the linear approximation?

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



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

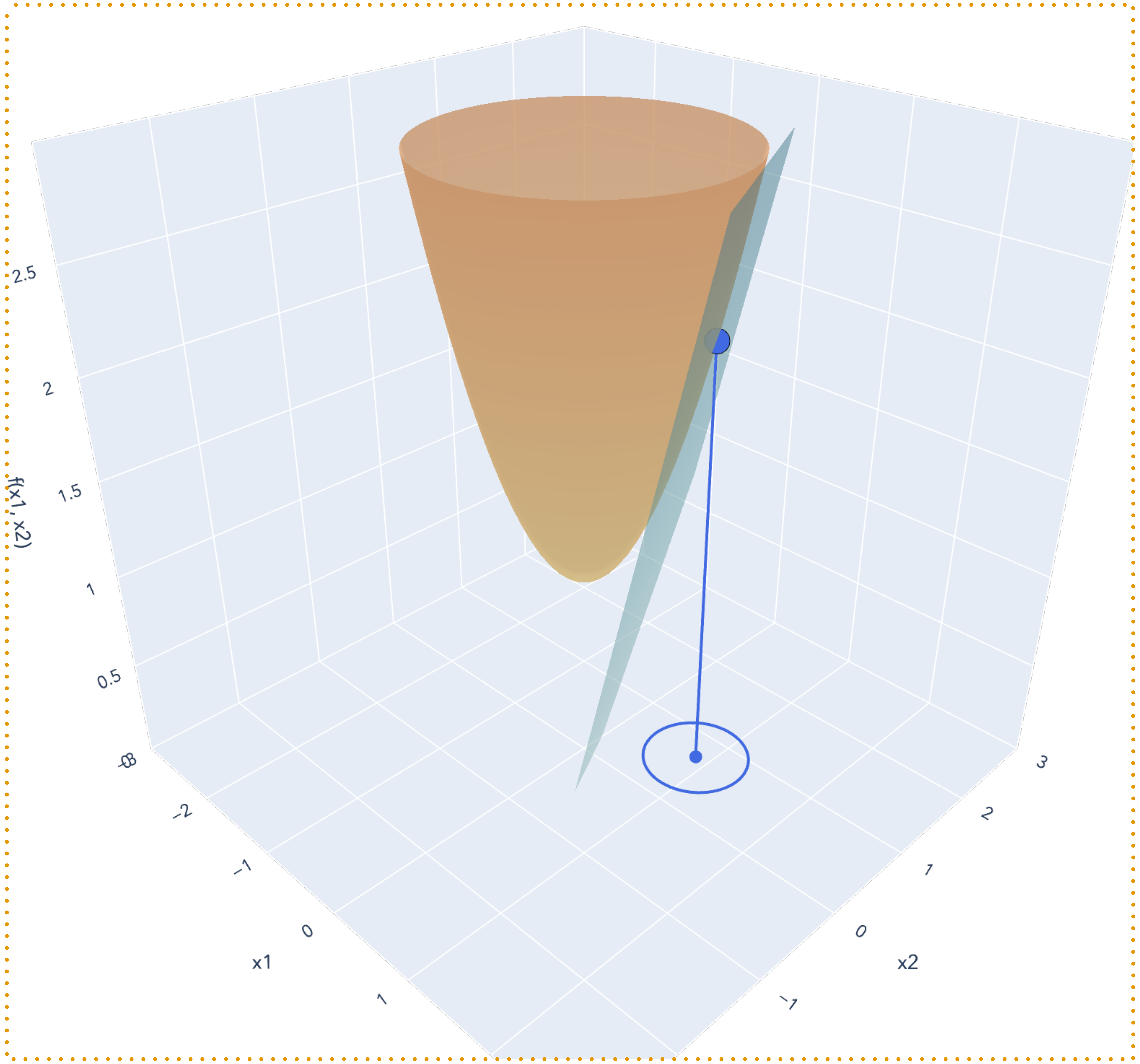
Linear Approximations

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

$$F(x_1, x_2) = x_1^2 + x_2^2 + 1 \text{ at } \mathbf{x}_0 = (1,\ 0.5)$$

What is the linear approximation?

$$F(w_1, w_2) \approx 2x_1 + x_2 - 0.25$$



$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$

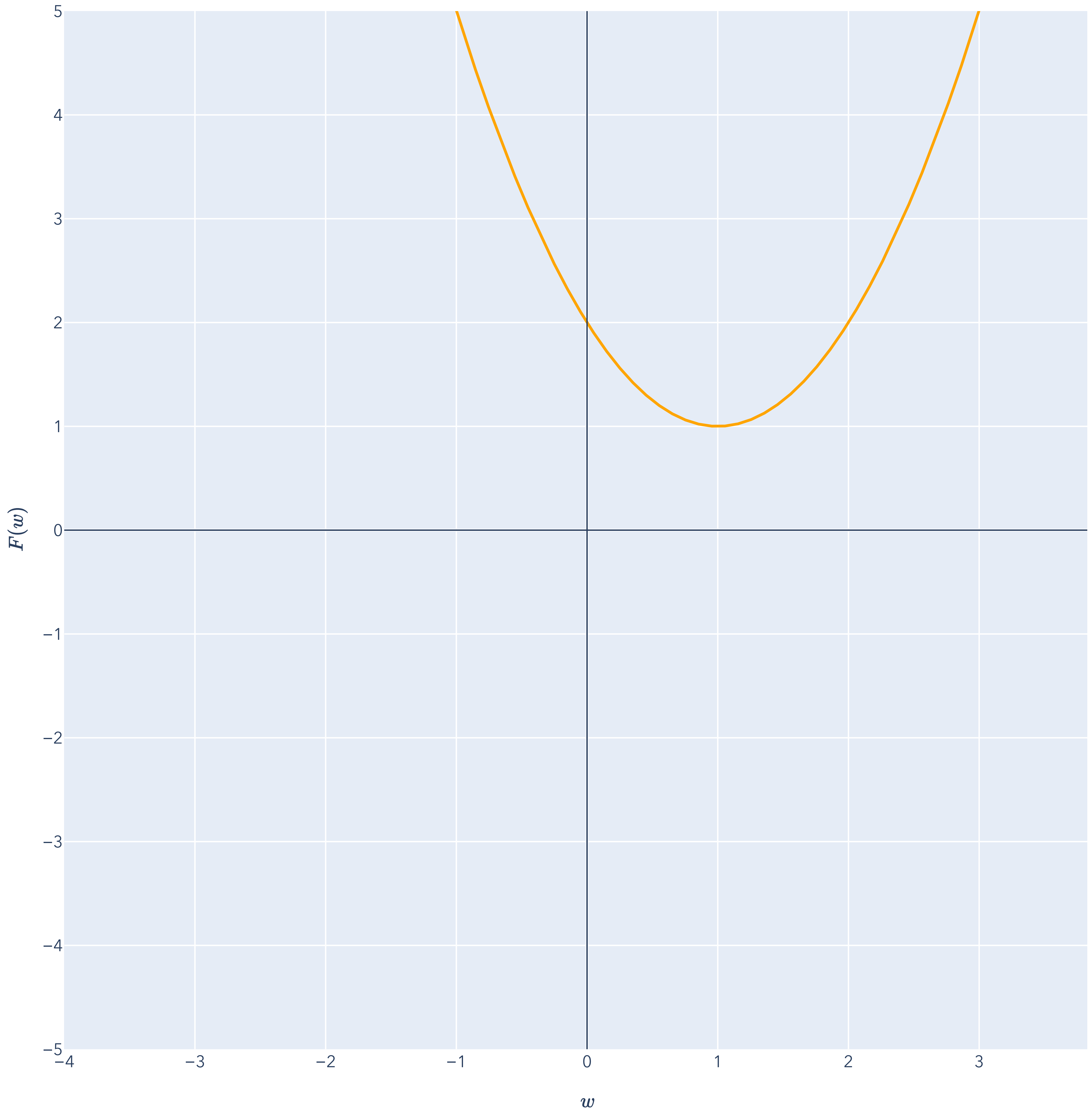
Gradient Descent

Designing a “candidate algorithm”

A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$



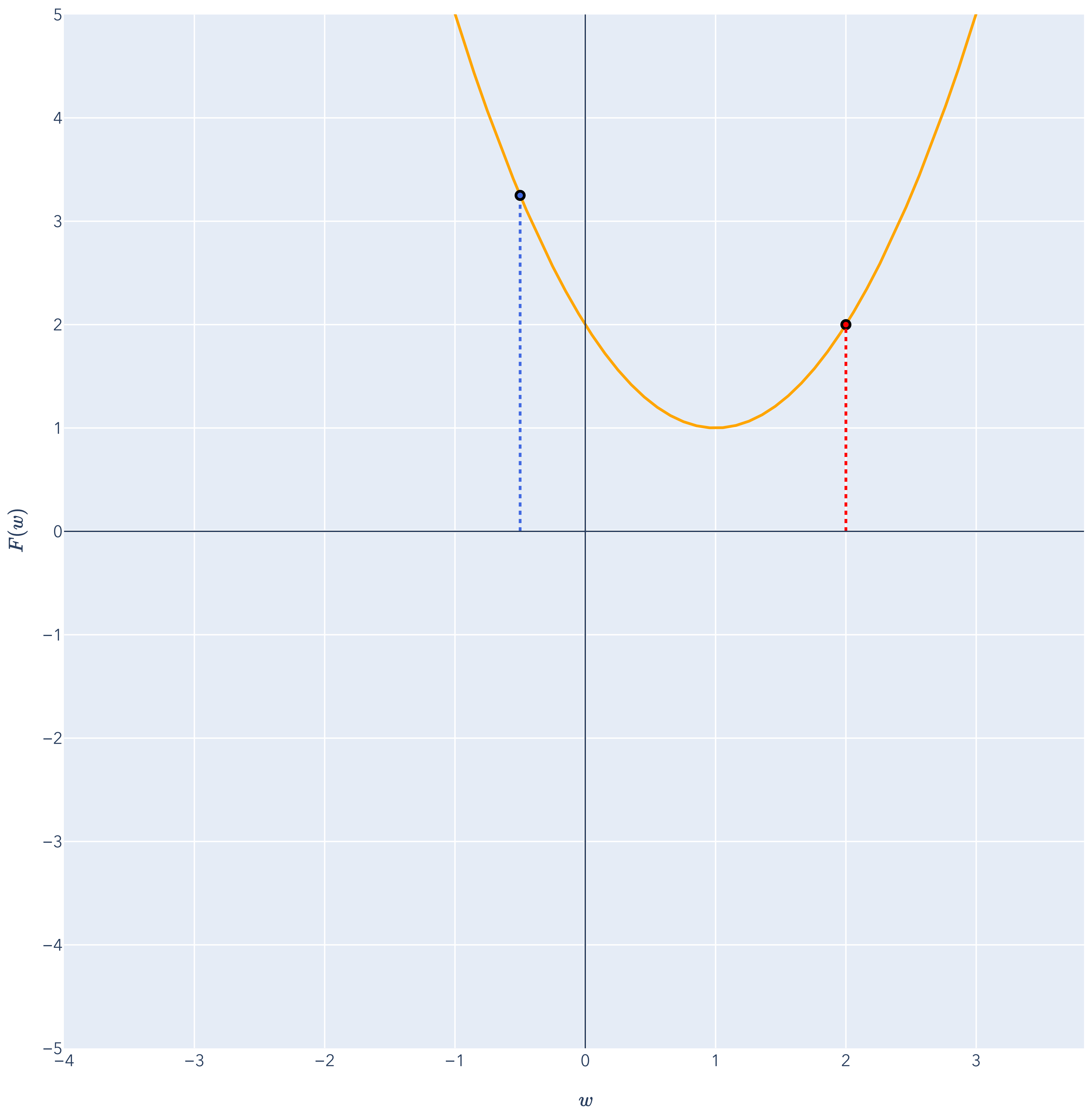
A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.



A candidate algorithm

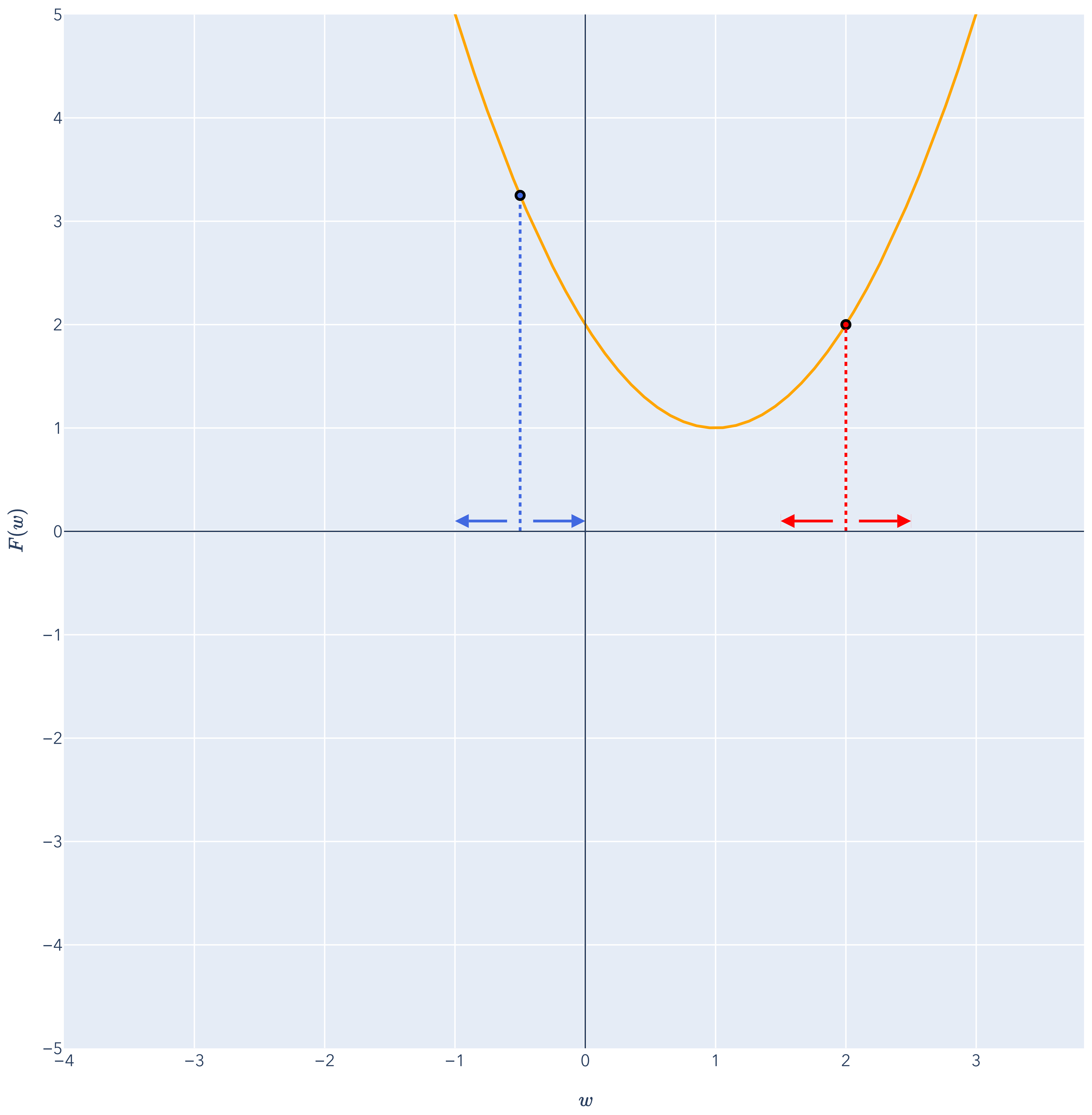
Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease* f?



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

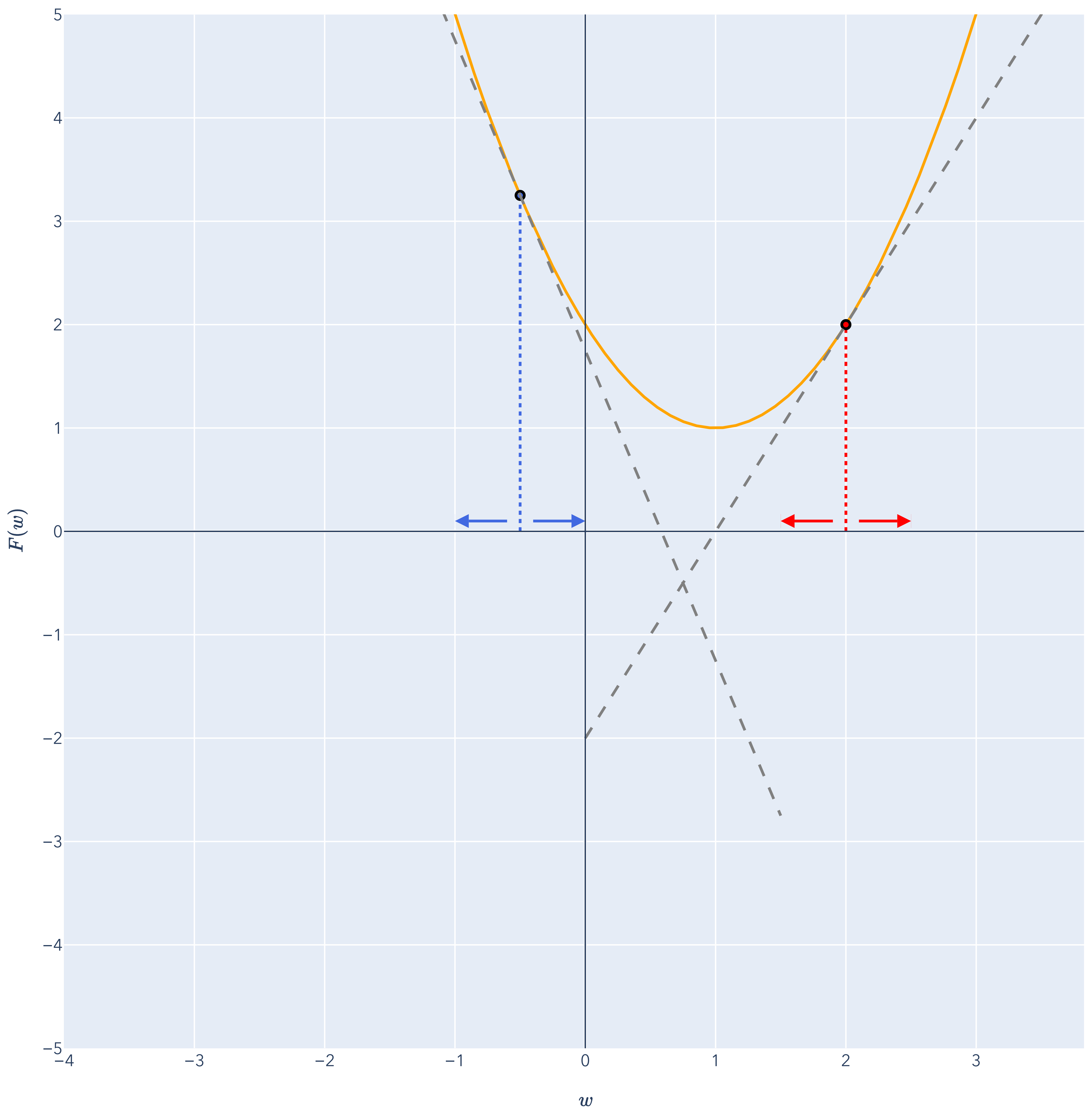
Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease f*?

If slope is negative, go right.

If slope is positive, go left.



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

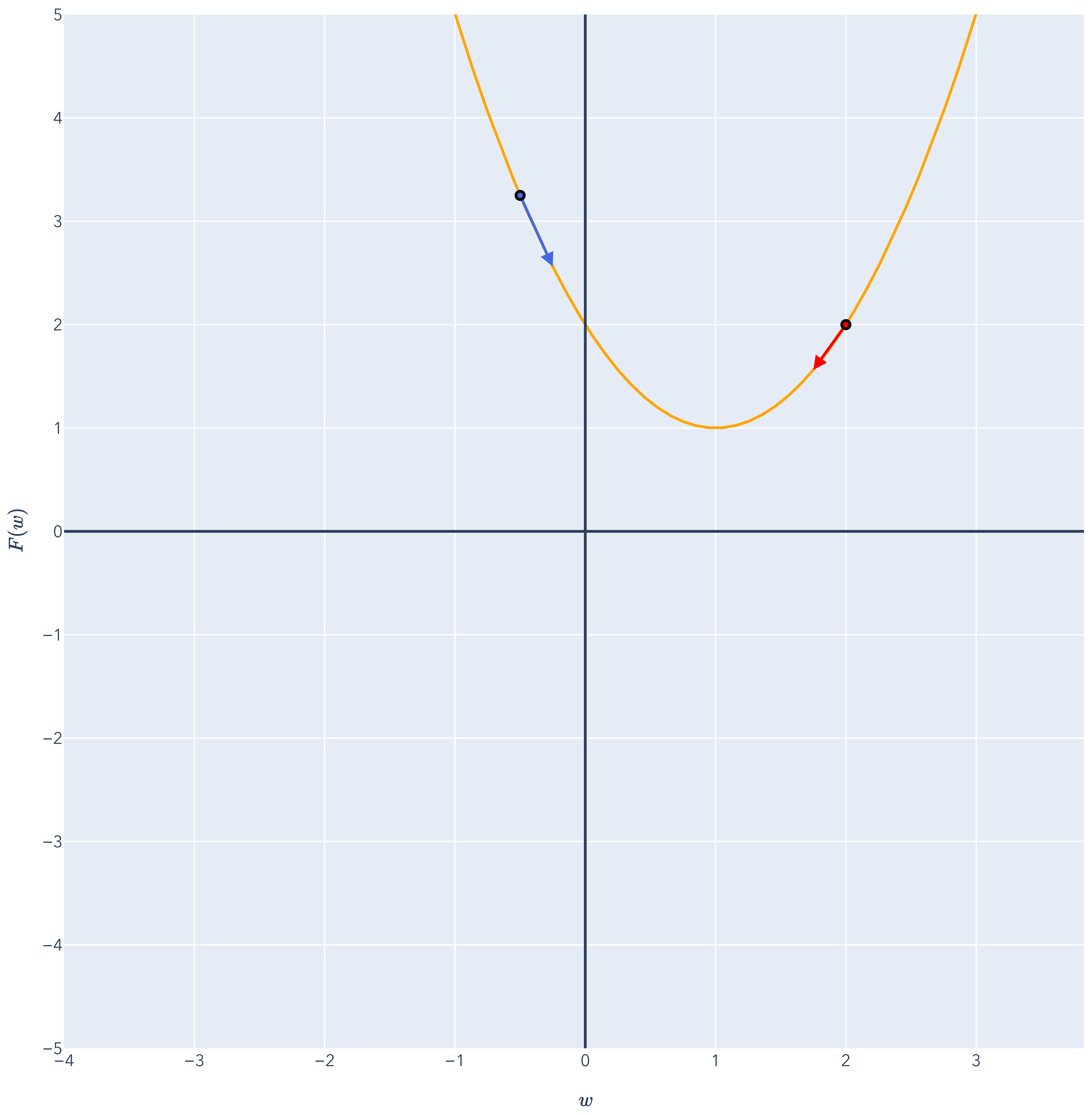
Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease f*?

Follow the derivative (slope at a point)!

Repeat over and over to minimize.



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

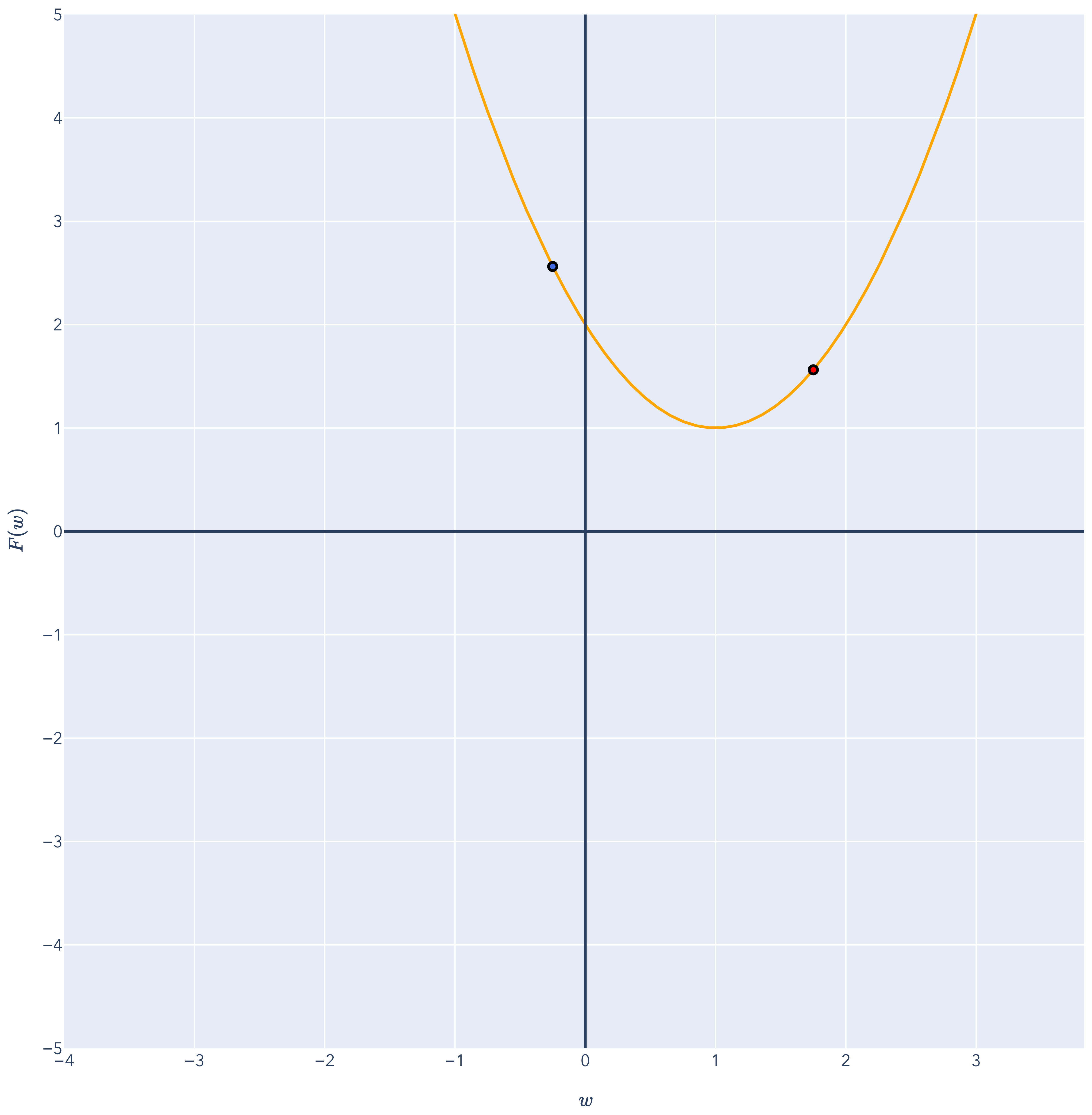
Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease f*?

Follow the derivative (slope at a point)!

Repeat over and over to minimize.



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

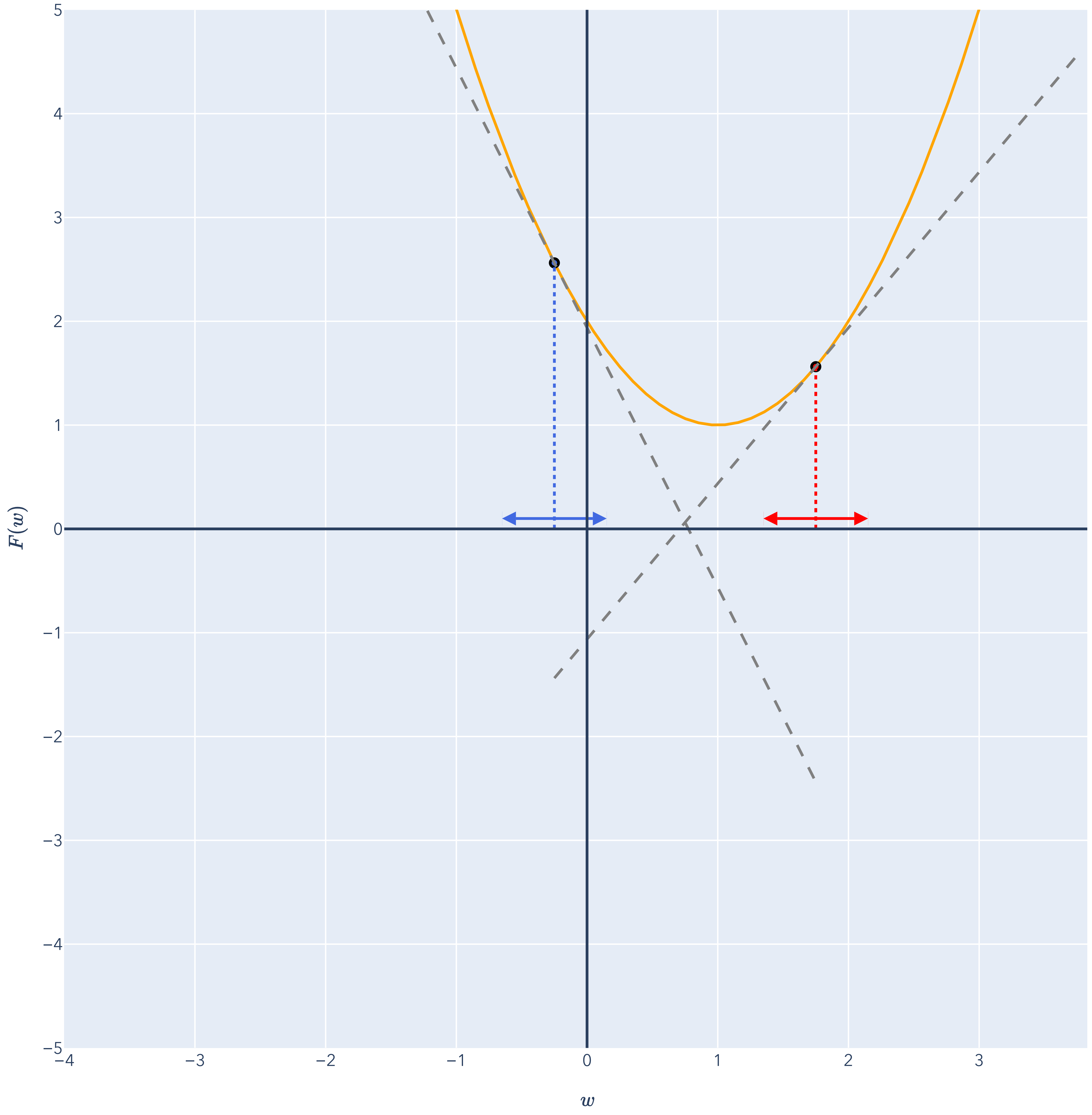
Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease f*?

Follow the derivative (slope at a point)!

Repeat over and over to minimize.



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

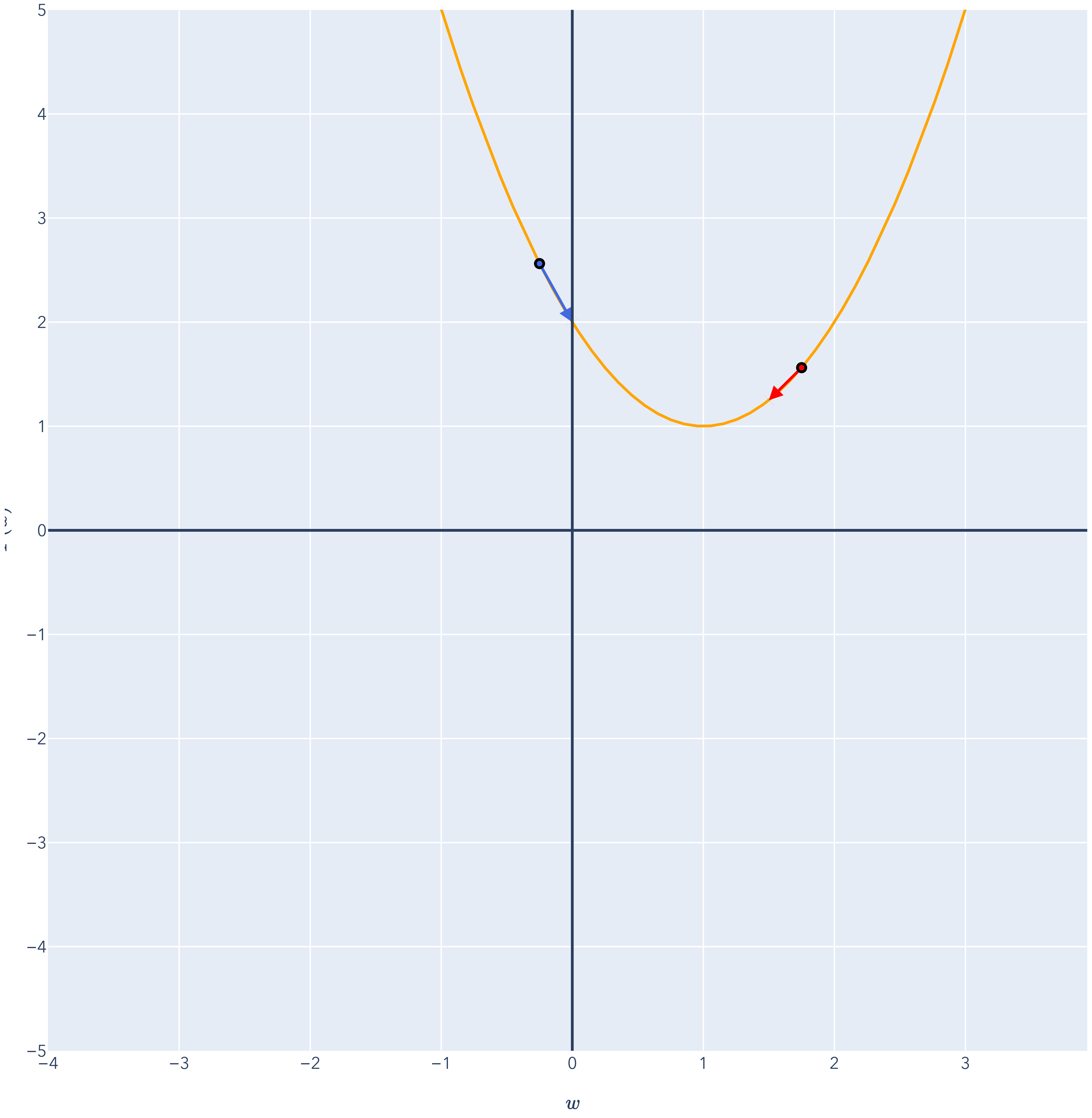
Suppose I drop you off at $w = -0.5$.

Or at $w = 2$.

Which direction to go in to *decrease f*?

Follow the derivative (slope at a point)!

Repeat over and over to minimize.



A candidate algorithm

Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

Suppose I drop you off at $w = -0.5$.

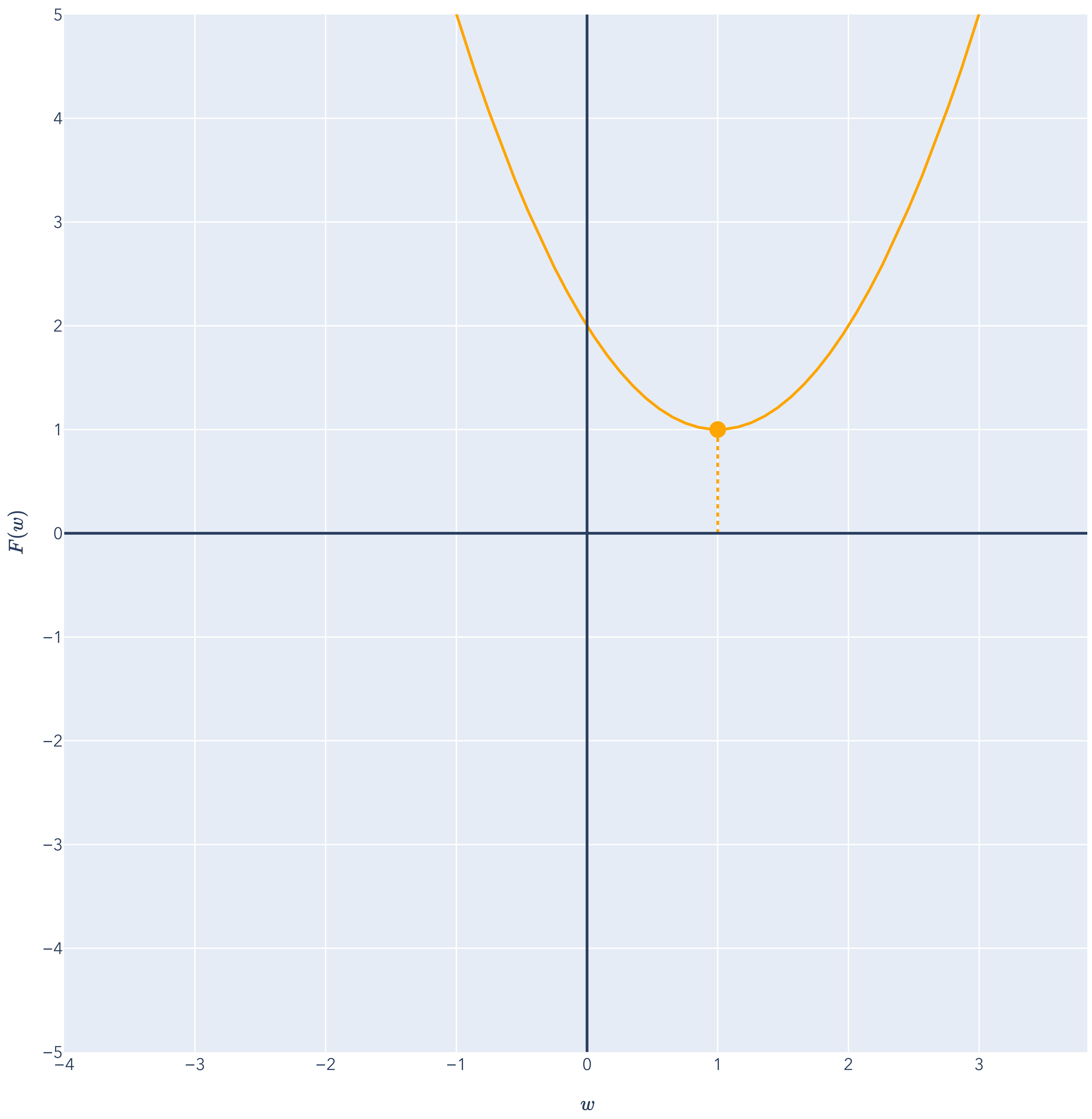
Or at $w = 2$.

Which direction to go in to *decrease* f?

Follow the derivative (slope at a point)!

Repeat over and over to minimize.

Eventually, we might reach a minimum!



A candidate algorithm

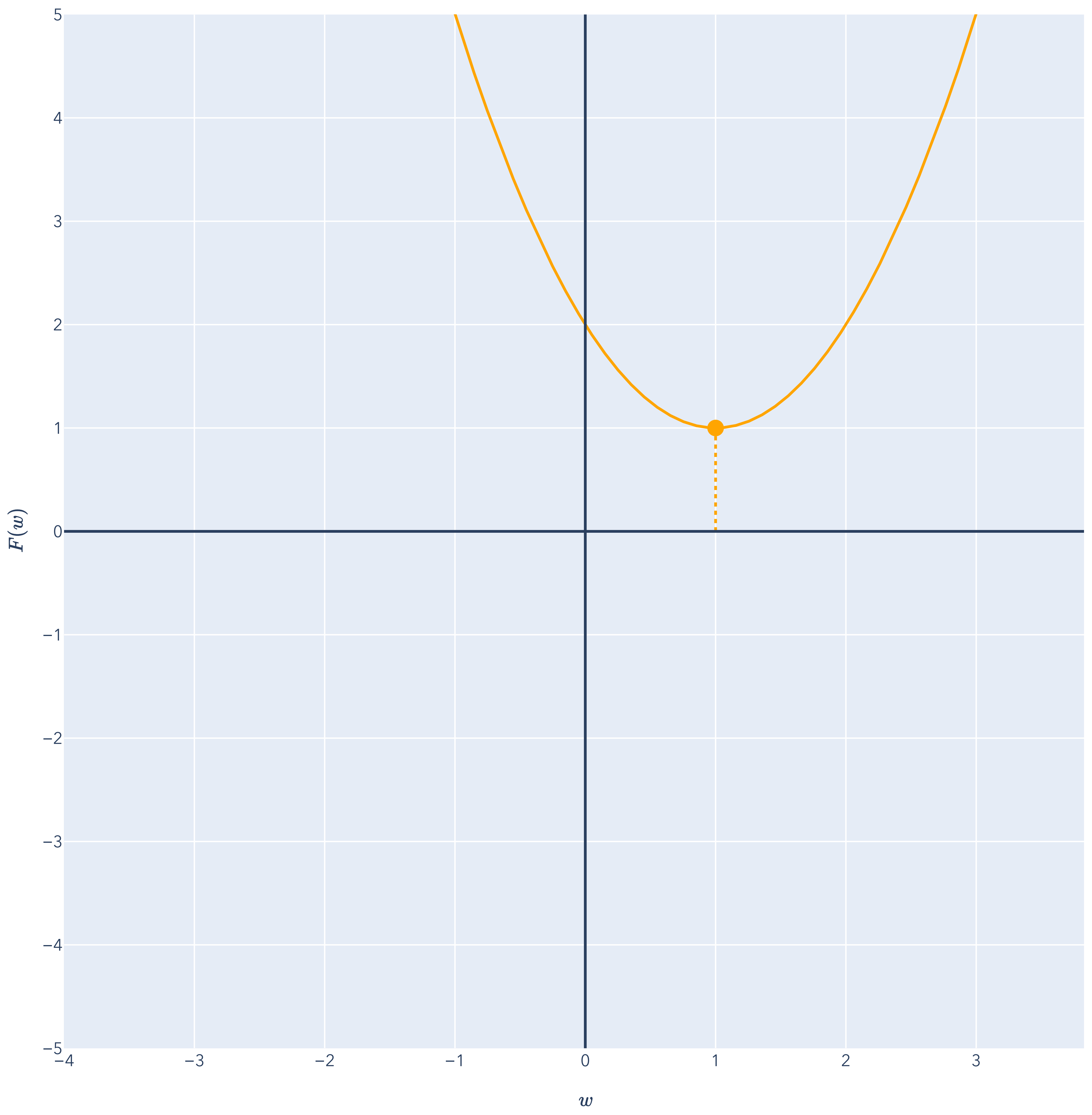
Moving in steepest descent direction

$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

But we can also just minimize in one shot!

$$f'(w) = 0$$

(first order condition)



A candidate algorithm

Moving in steepest descent direction

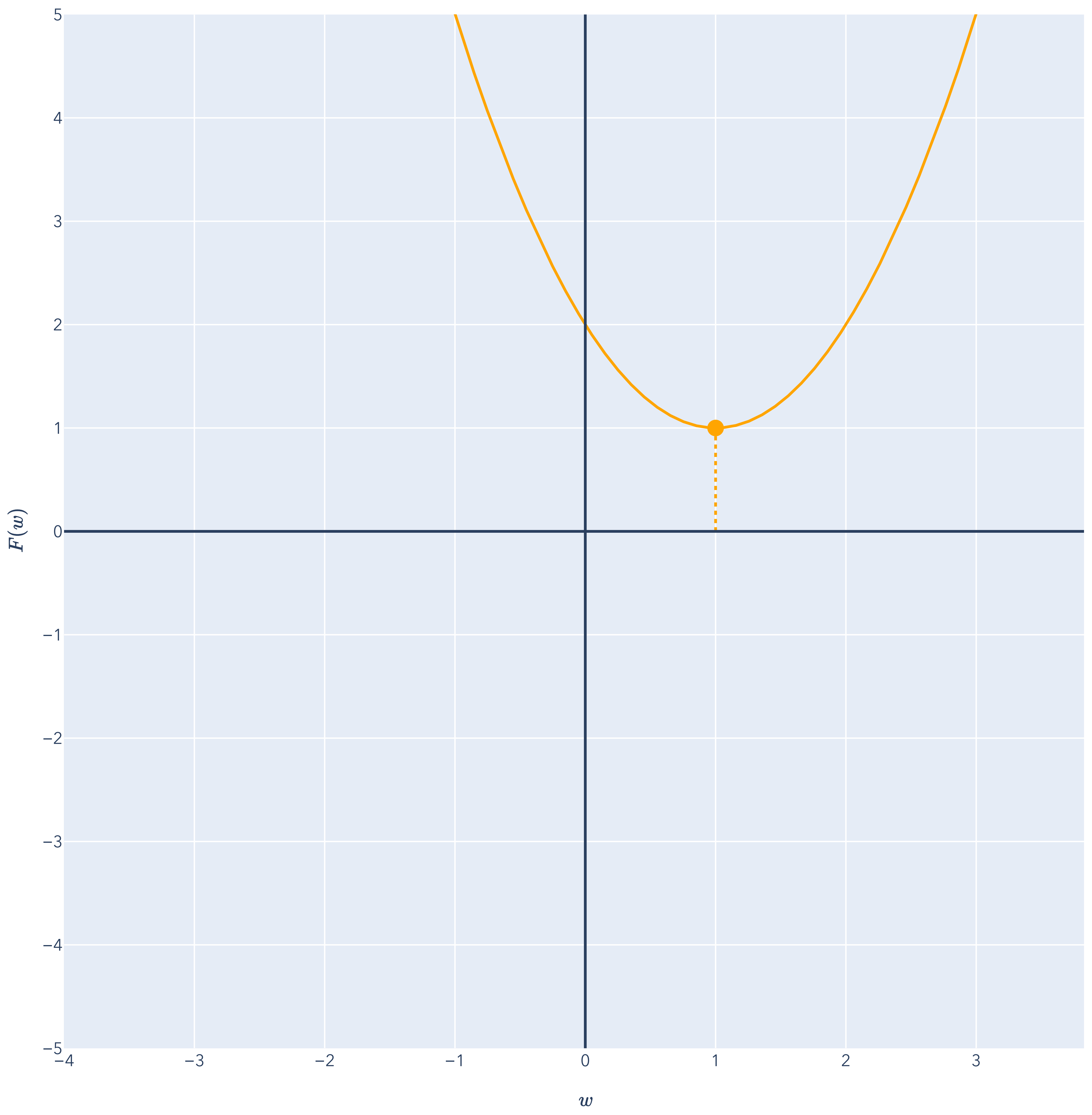
$$\underset{w \in \mathbb{R}}{\text{minimize}} \quad f(w)$$

But we can also just minimize in one shot!

$$f'(w) = 0$$

(first order condition)

Not always possible, so need an *iterative* algorithm.

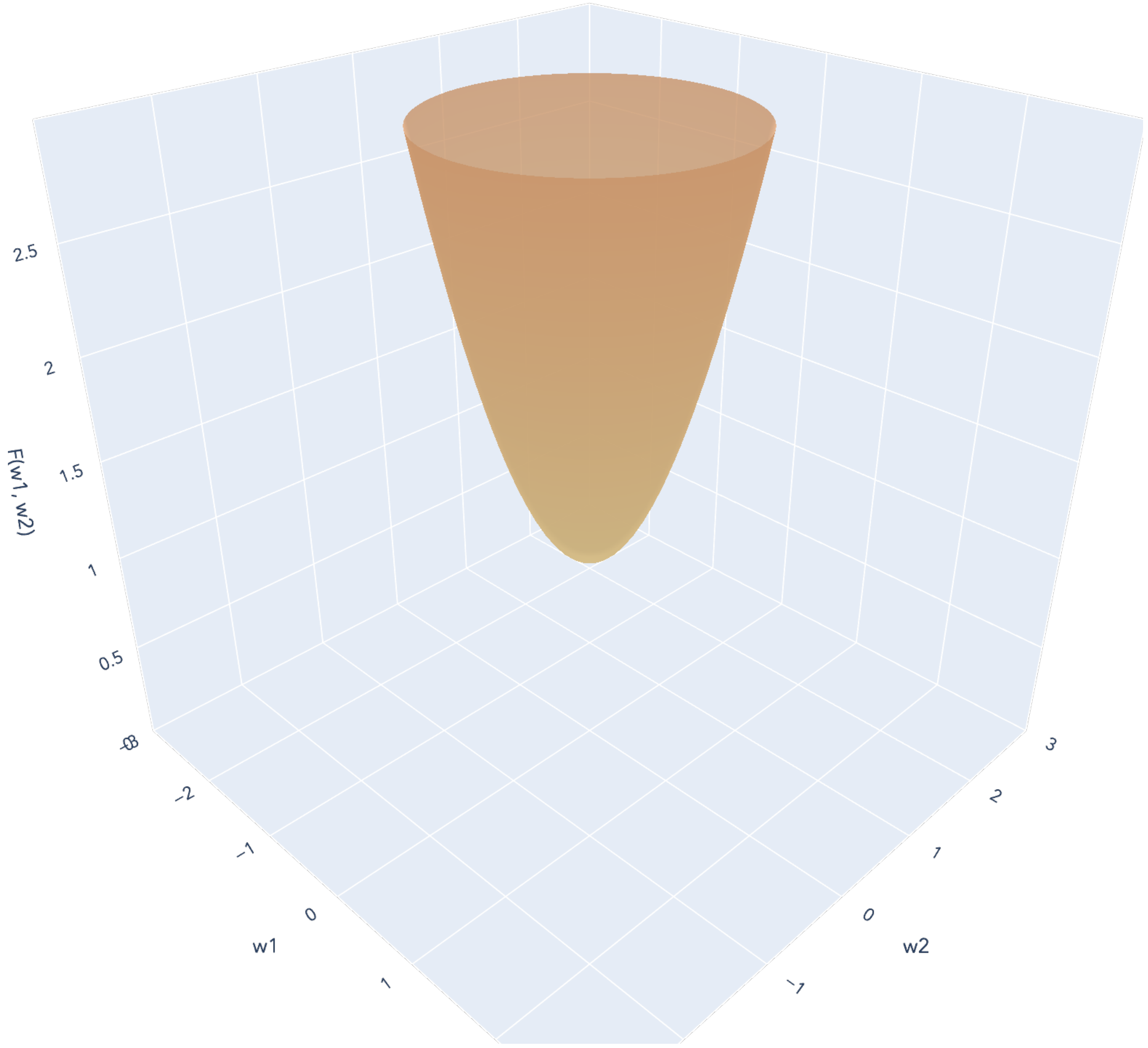


A candidate algorithm

Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$



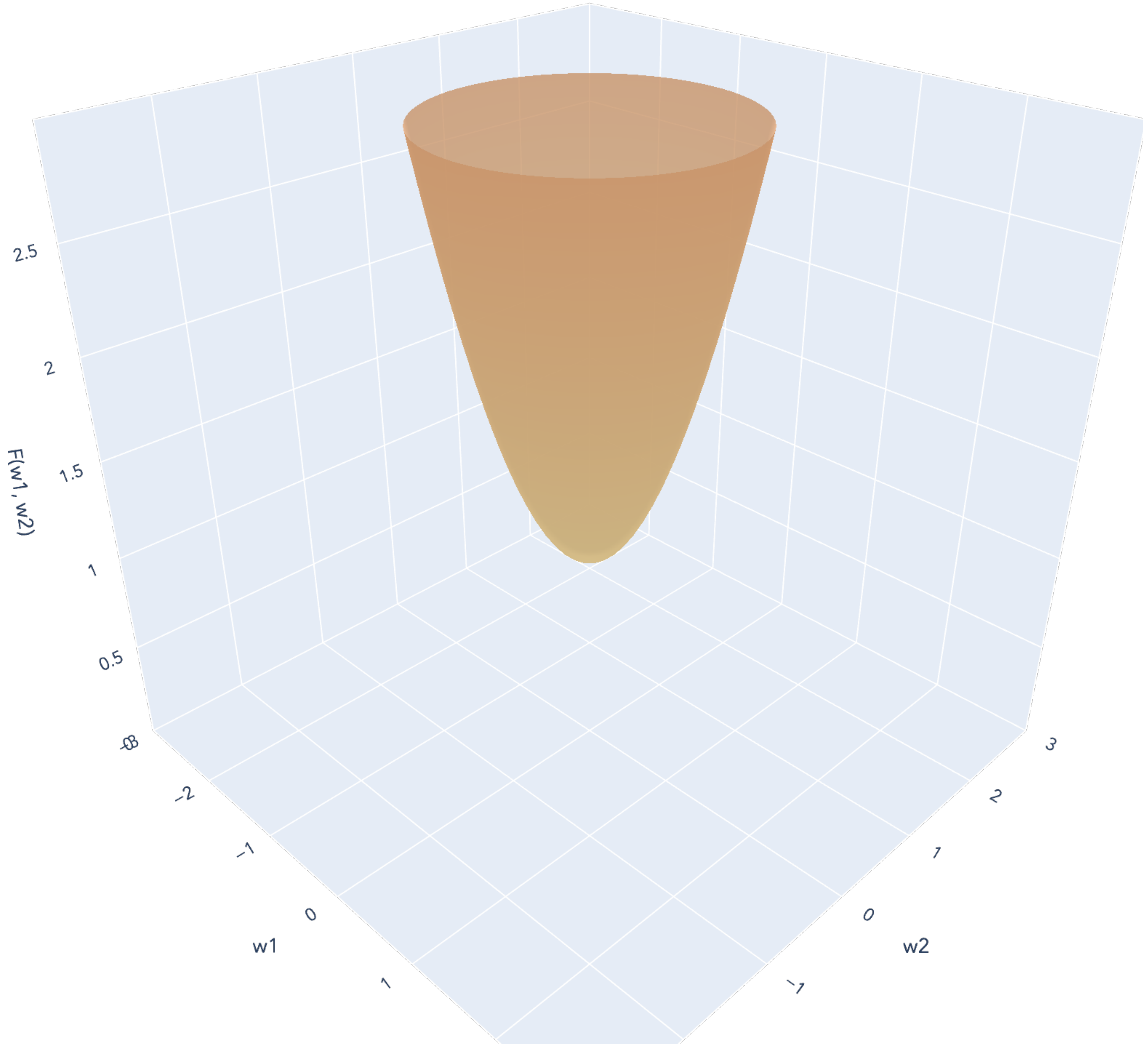
A candidate algorithm

Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

From two directions to infinitely many directions to go in…



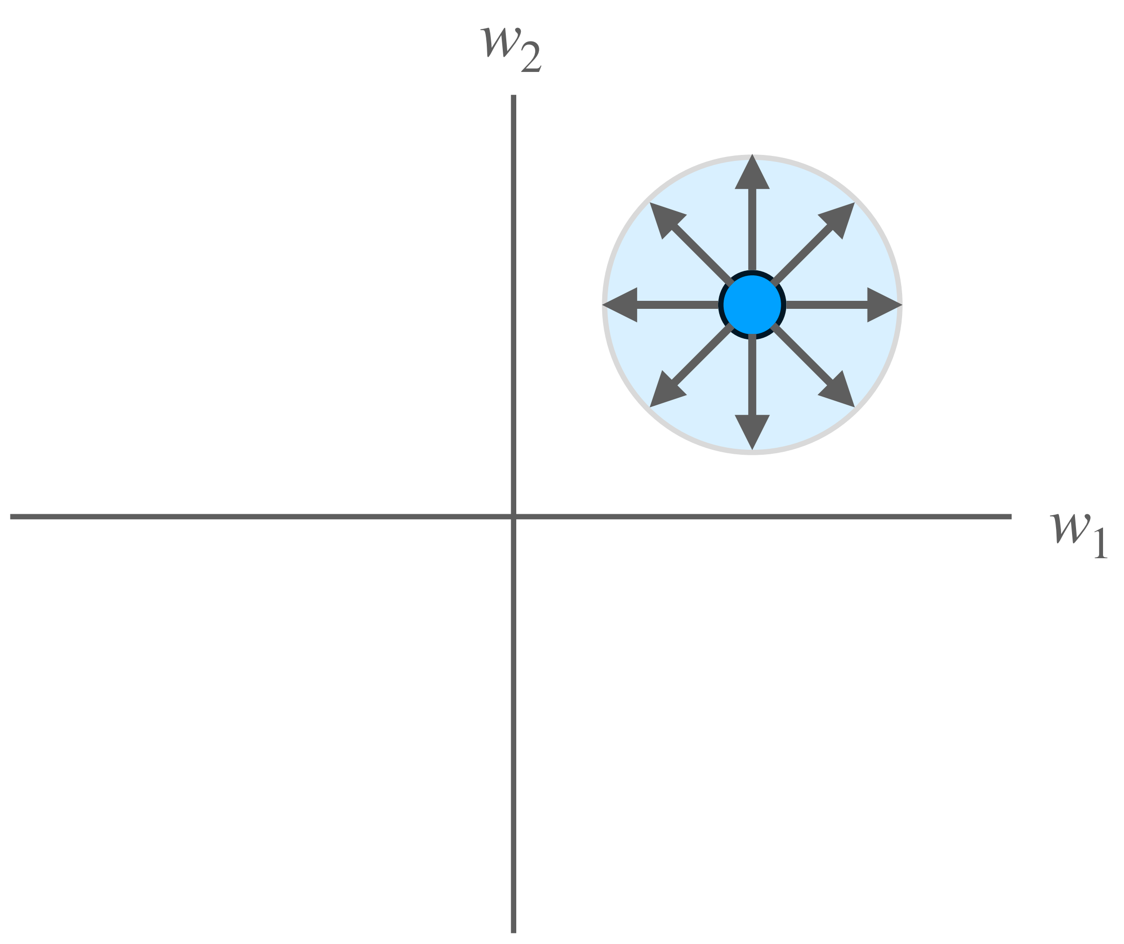
A candidate algorithm

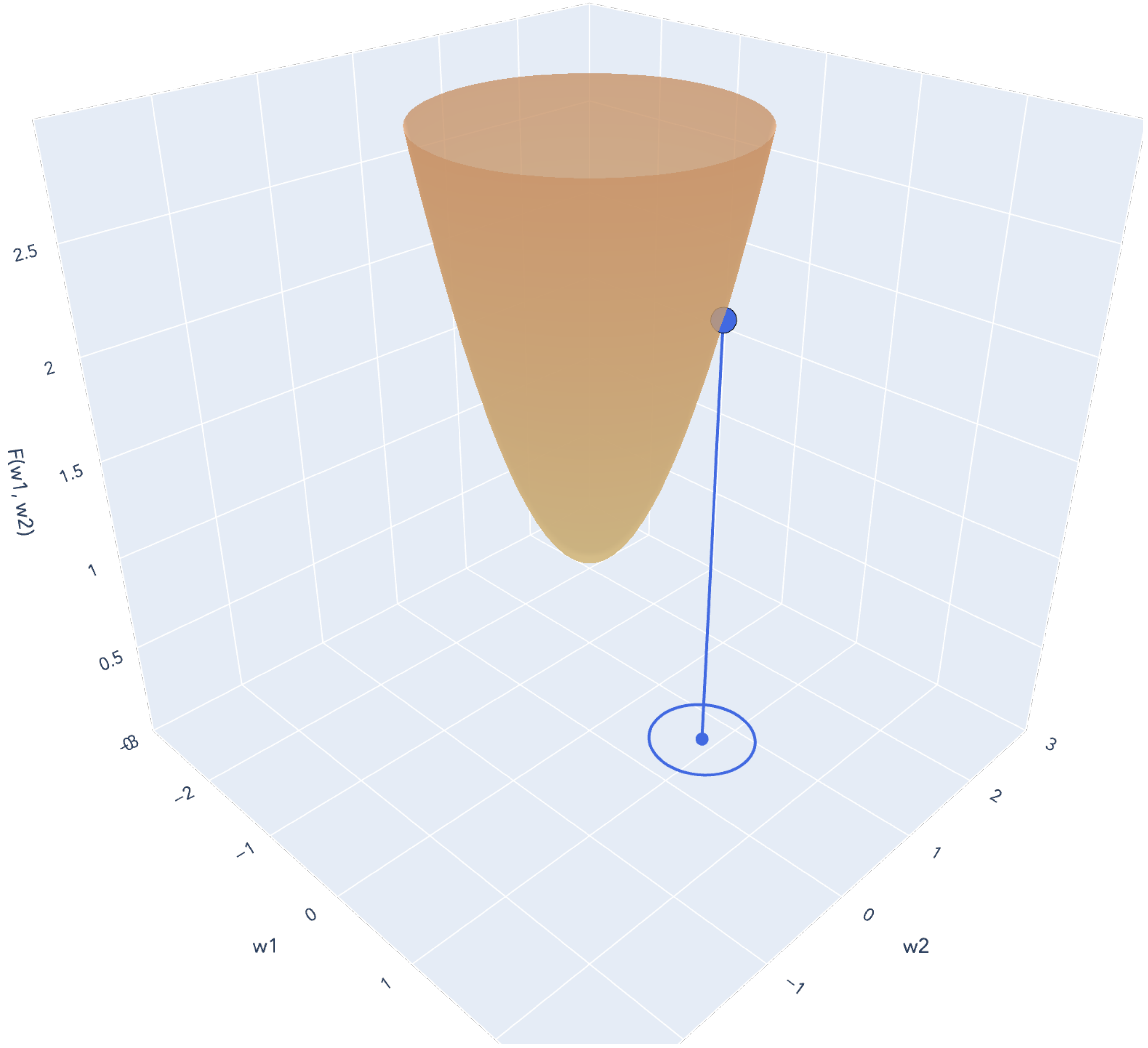
Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

From two directions to infinitely many directions to go in…





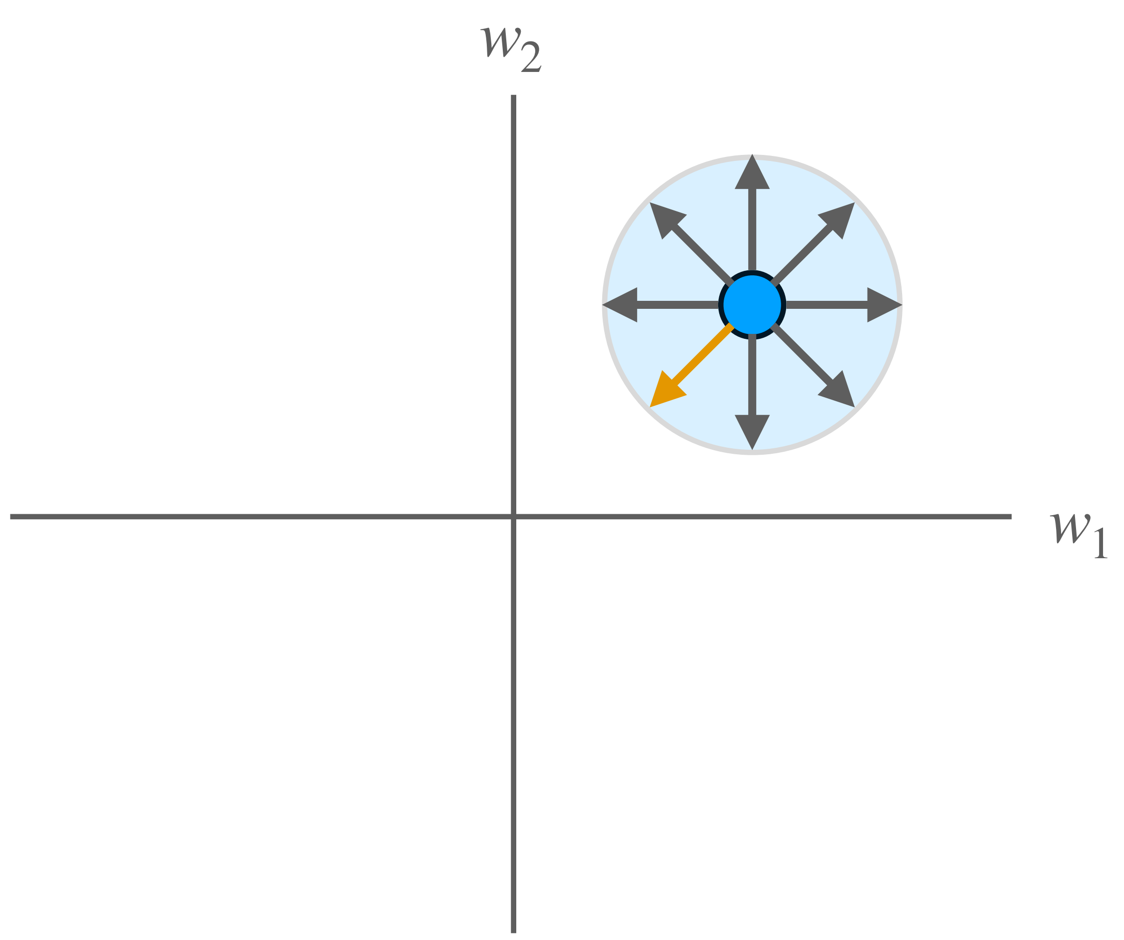
A candidate algorithm

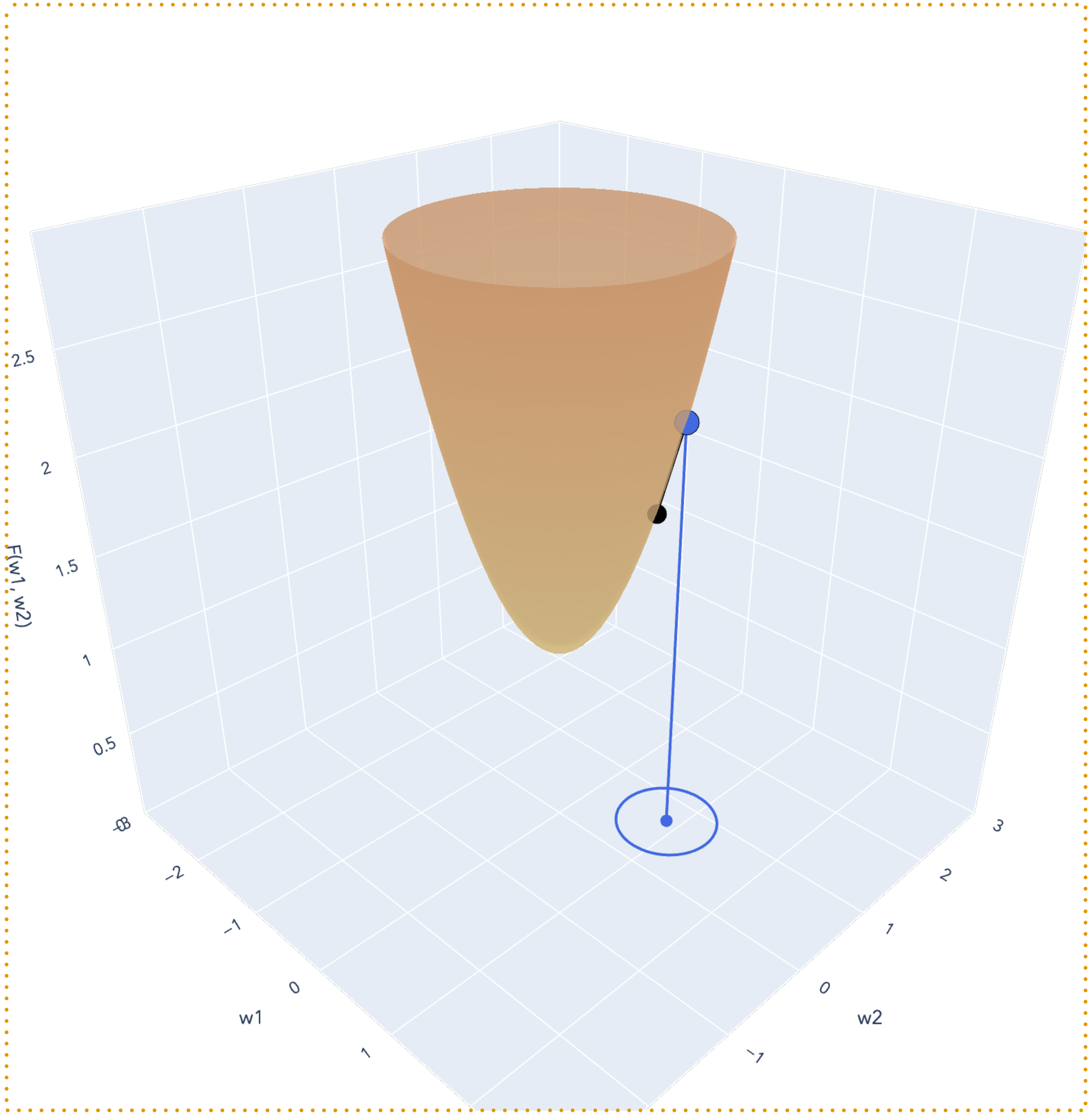
Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

But still can go in the "steepest decrease" direction!





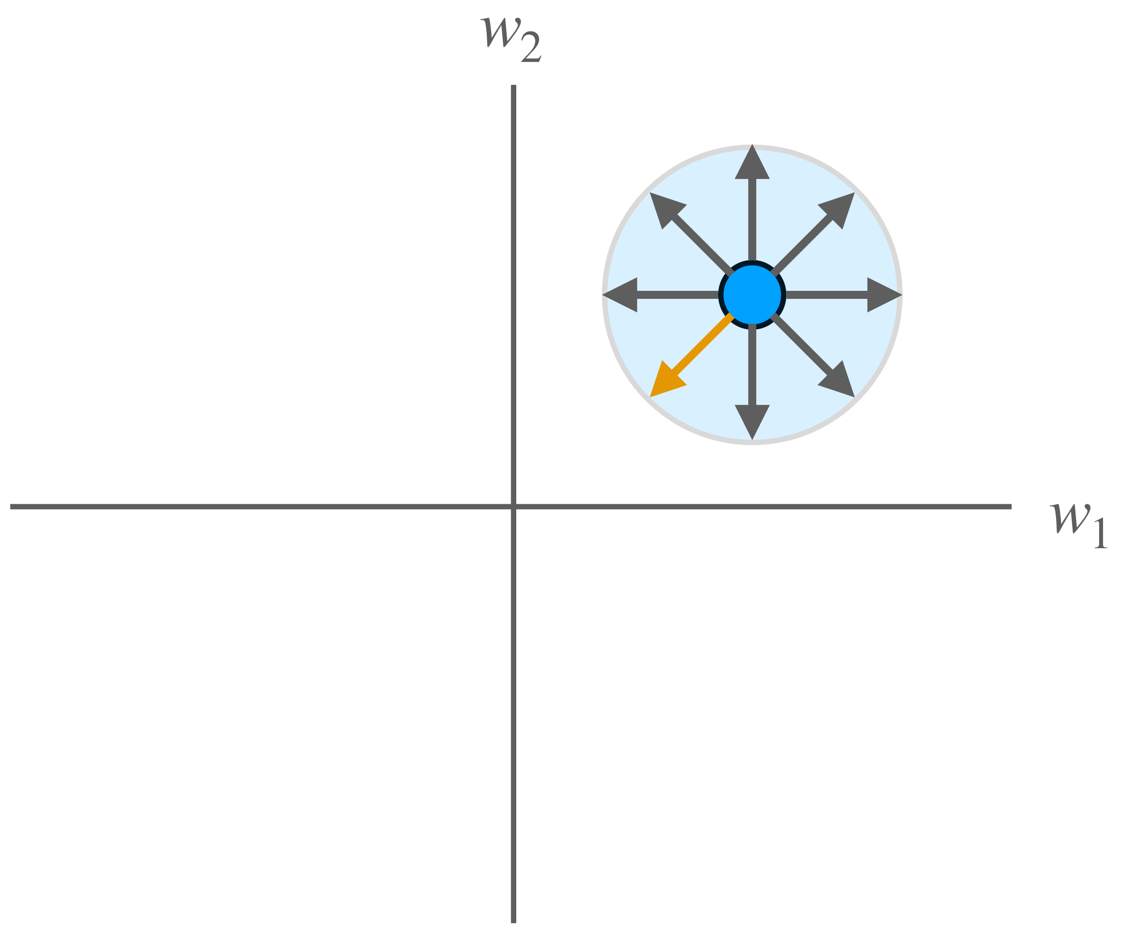
A candidate algorithm

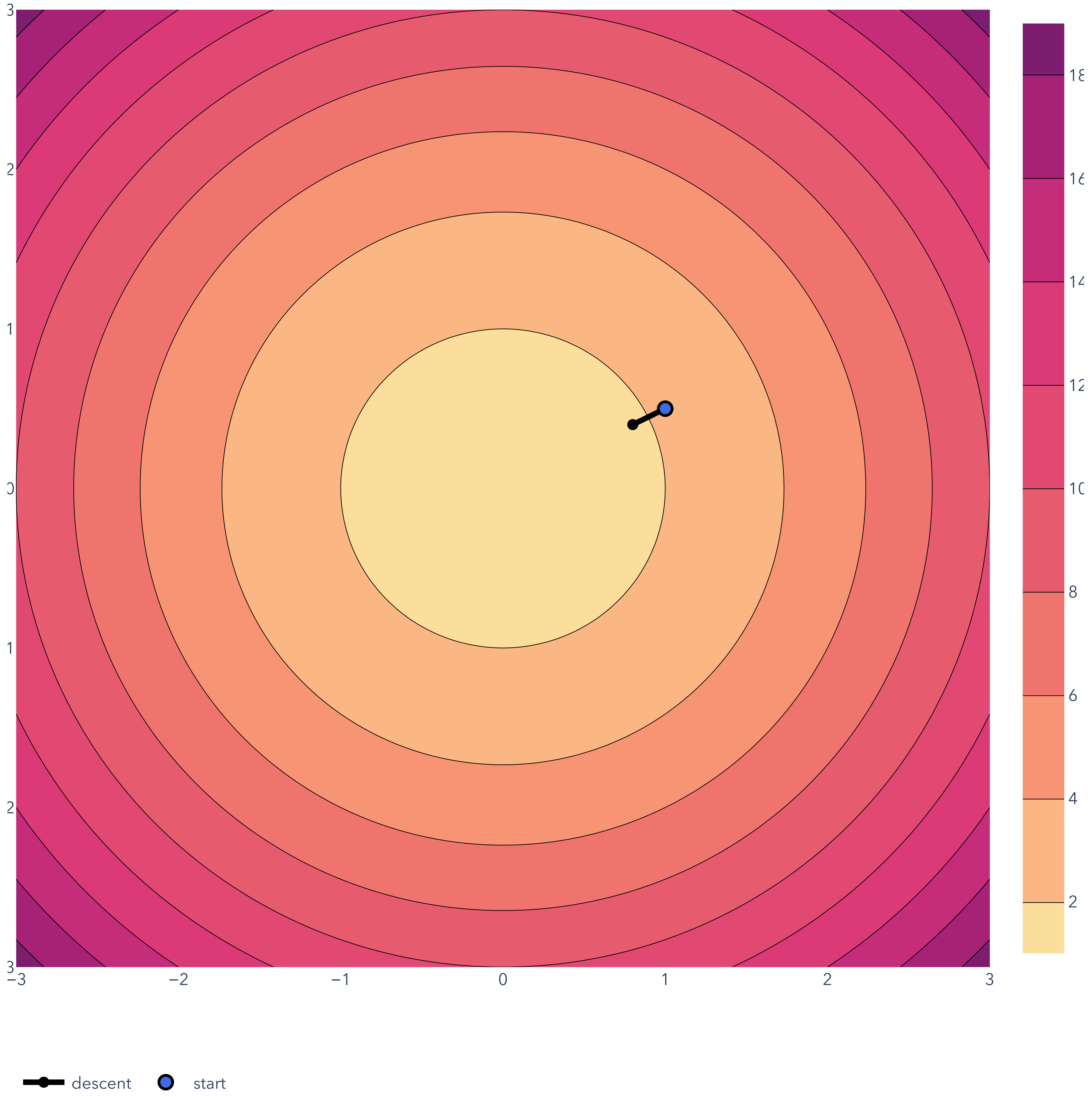
Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

But still can go in the "steepest decrease" direction!





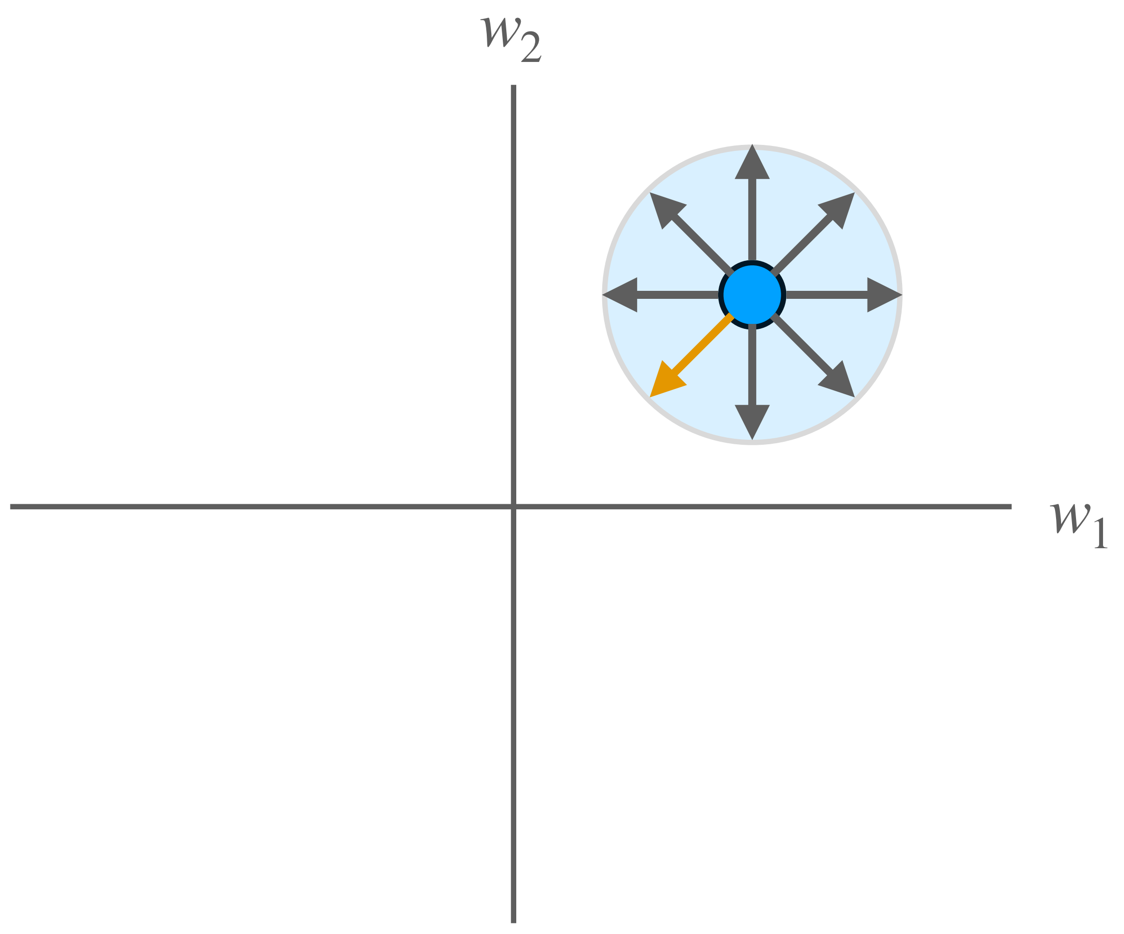
A candidate algorithm

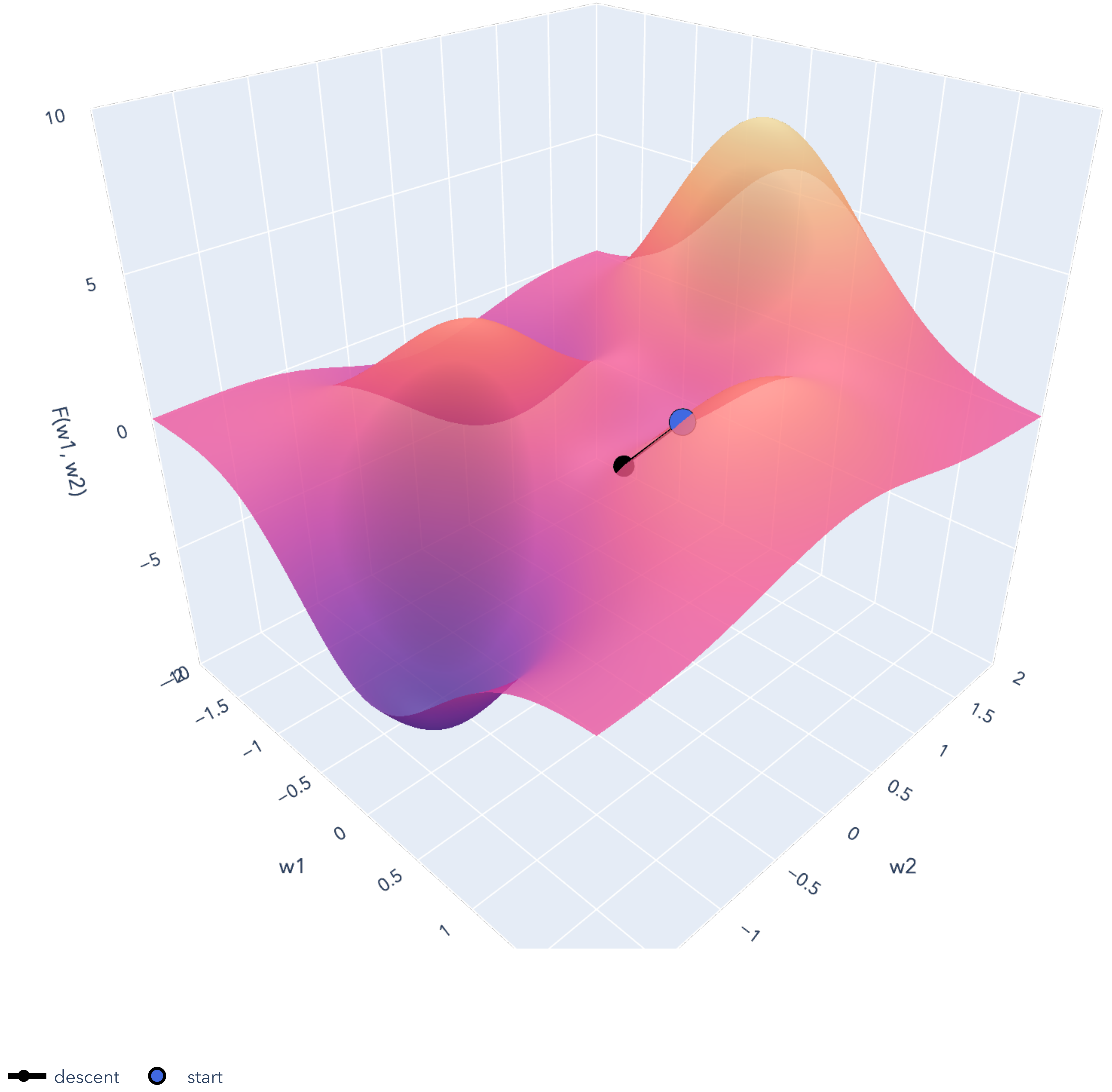
Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

This "myopic" strategy works for arbitrarily complex functions.





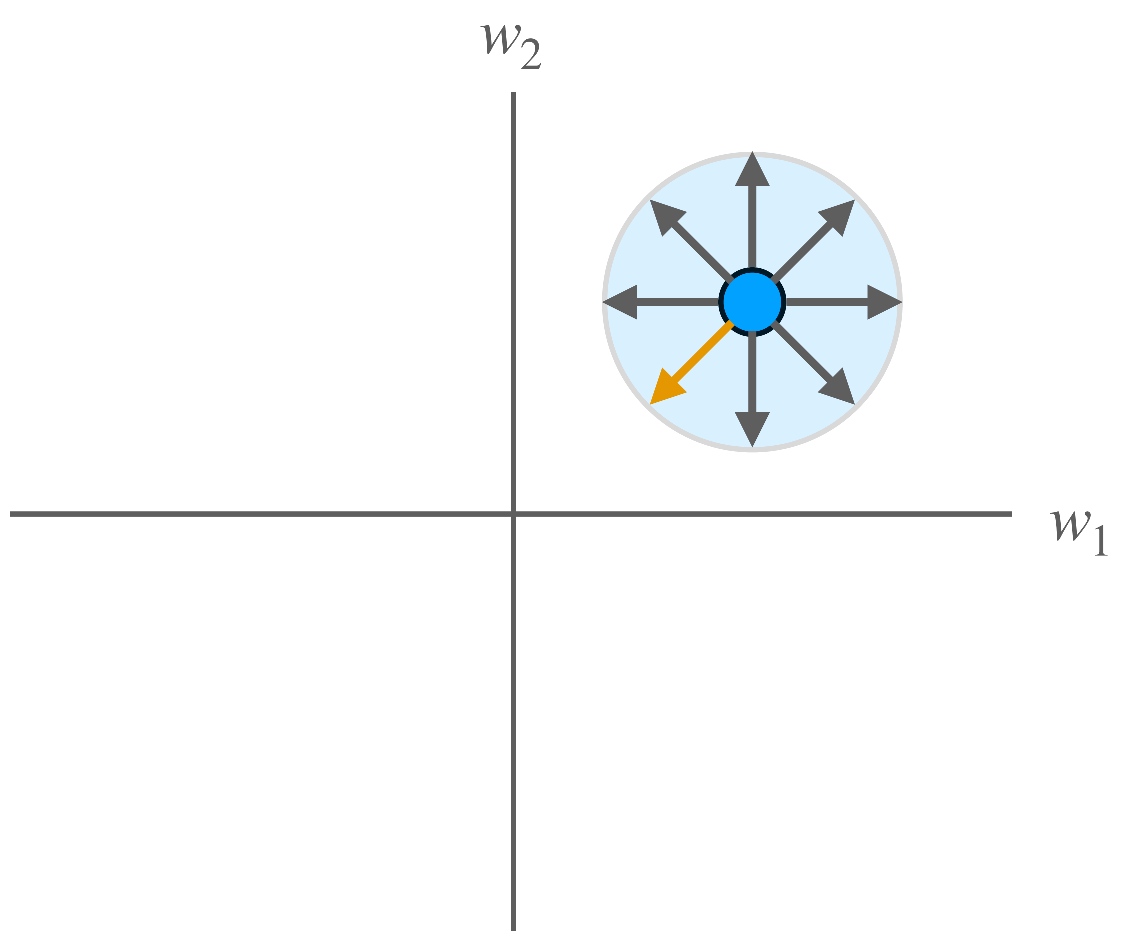
A candidate algorithm

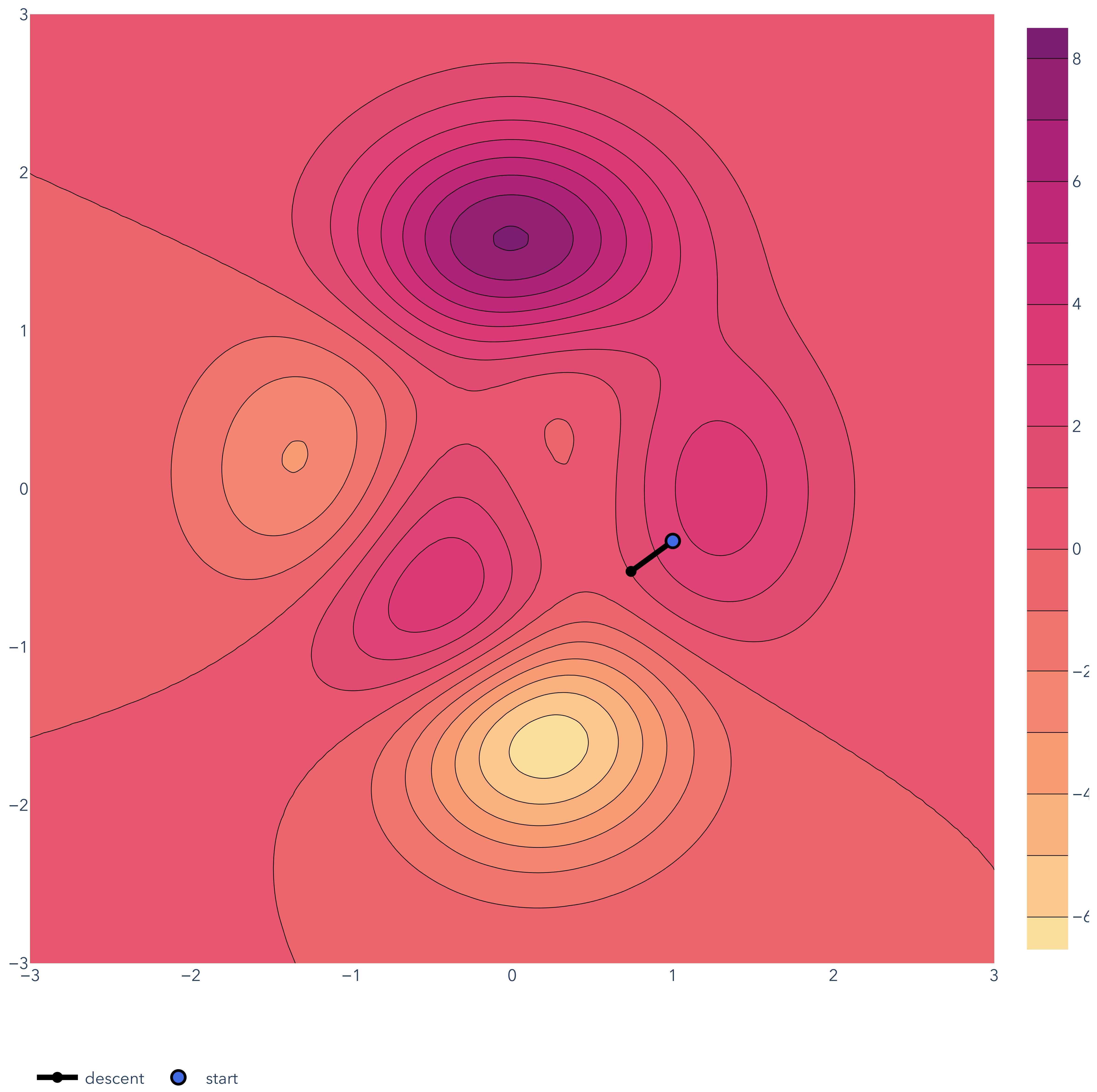
Moving in steepest descent direction

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{w})$$

$$f(w_1, w_2)$$

This "myopic" strategy works for arbitrarily complex functions.





A candidate algorithm

Moving in steepest descent direction

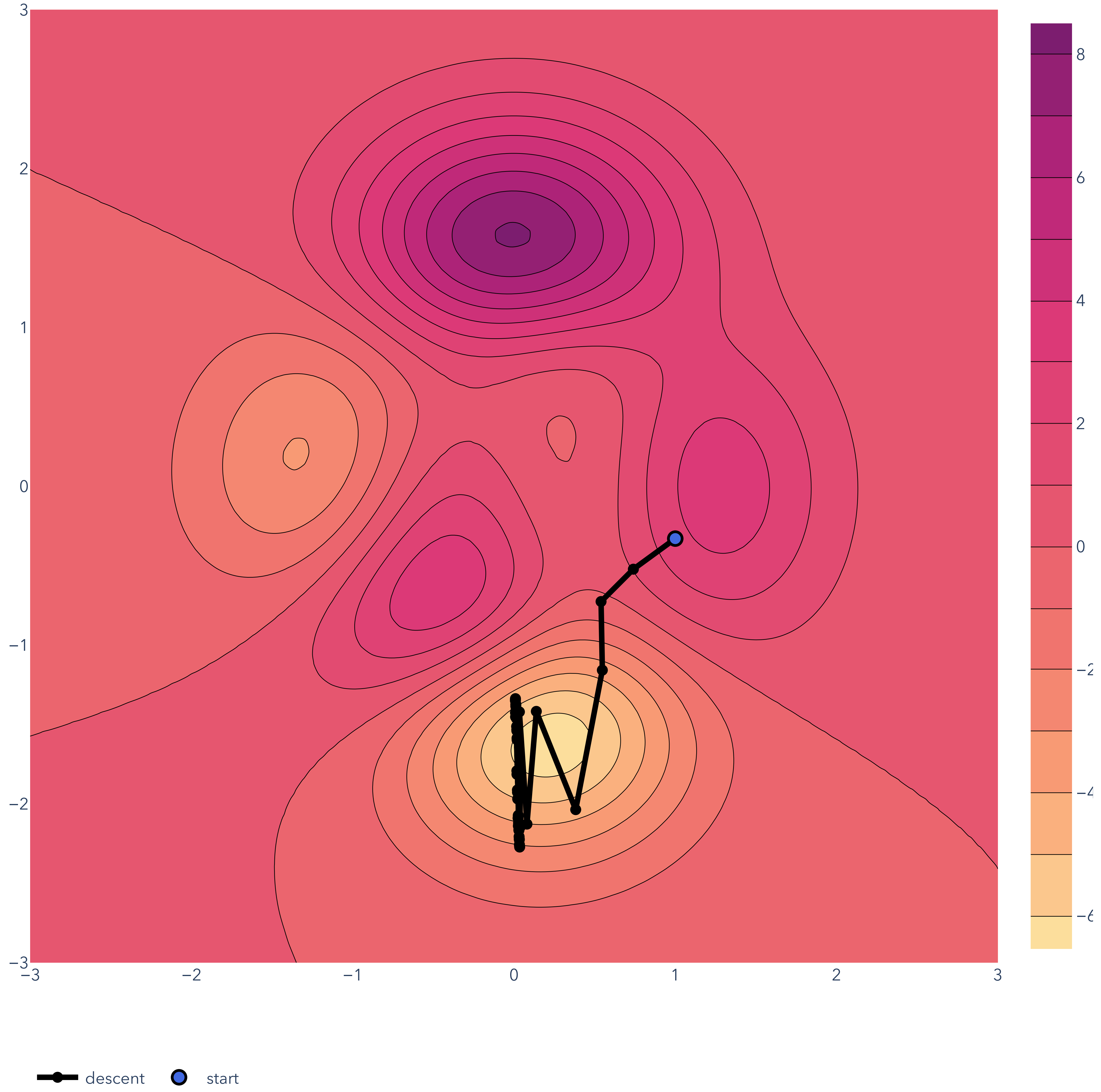
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})\ldots$

Take another step in the direction of steepest decrease for $f(\mathbf{w})\ldots$

$\vdots$

Repeat until satisfied.



A candidate algorithm

Moving in steepest descent direction

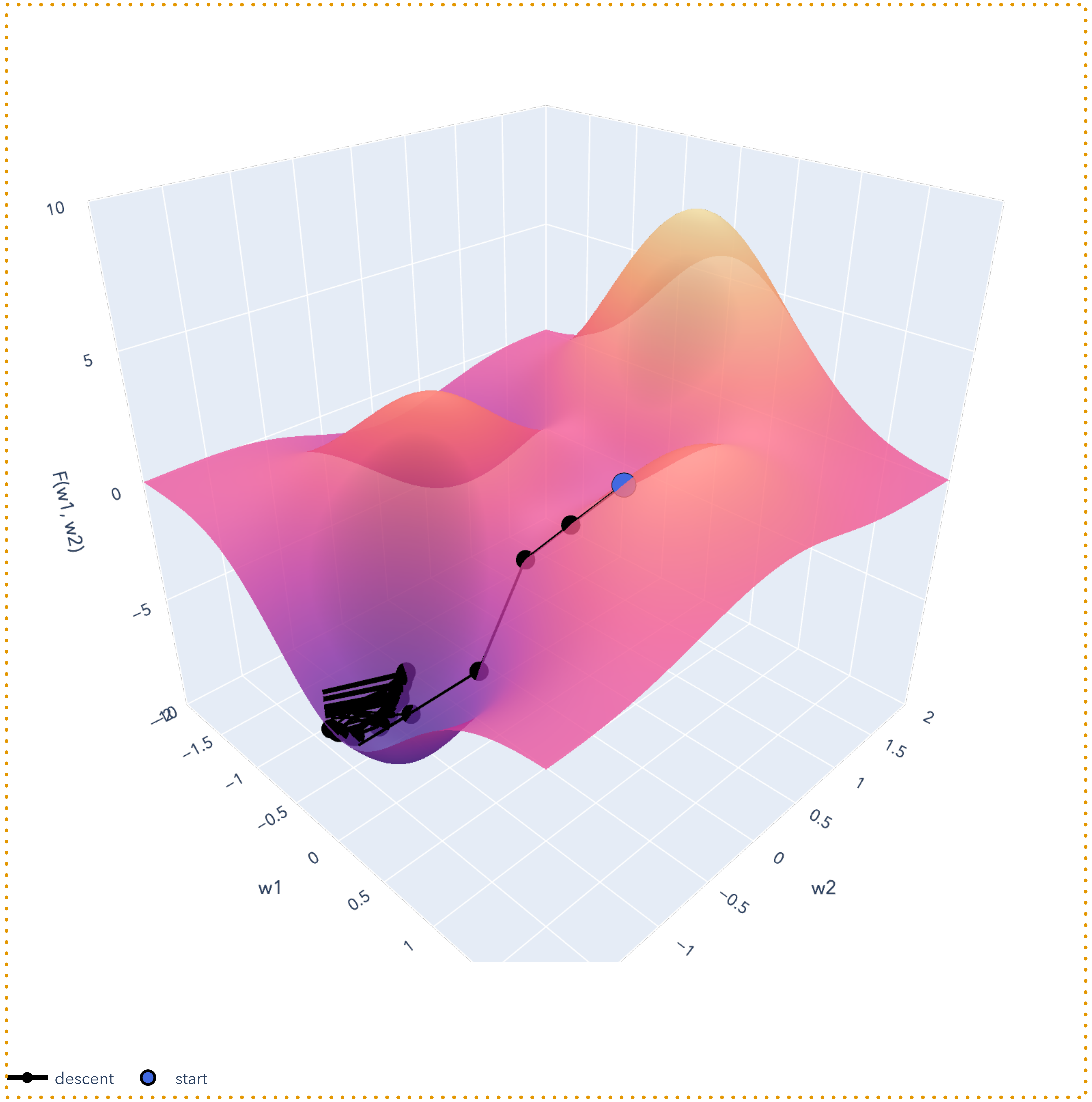
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})$…

Take another step in the direction of steepest decrease for $f(\mathbf{w})$…

⋮

Repeat until satisfied.



A candidate algorithm

Moving in steepest descent direction

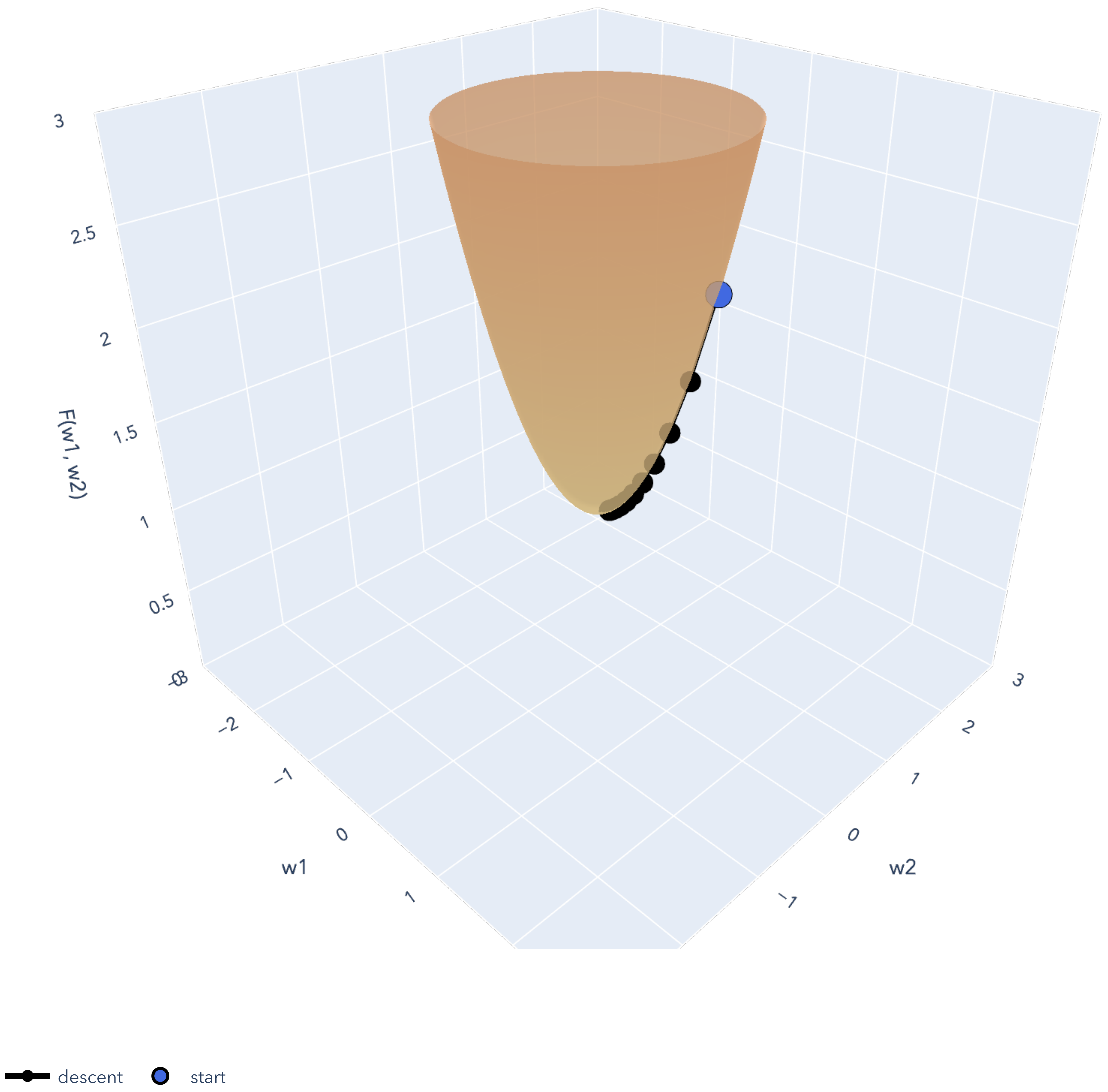
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})$…

Take another step in the direction of steepest decrease for $f(\mathbf{w})$…

⋮

Repeat until satisfied.



A candidate algorithm

Moving in steepest descent direction

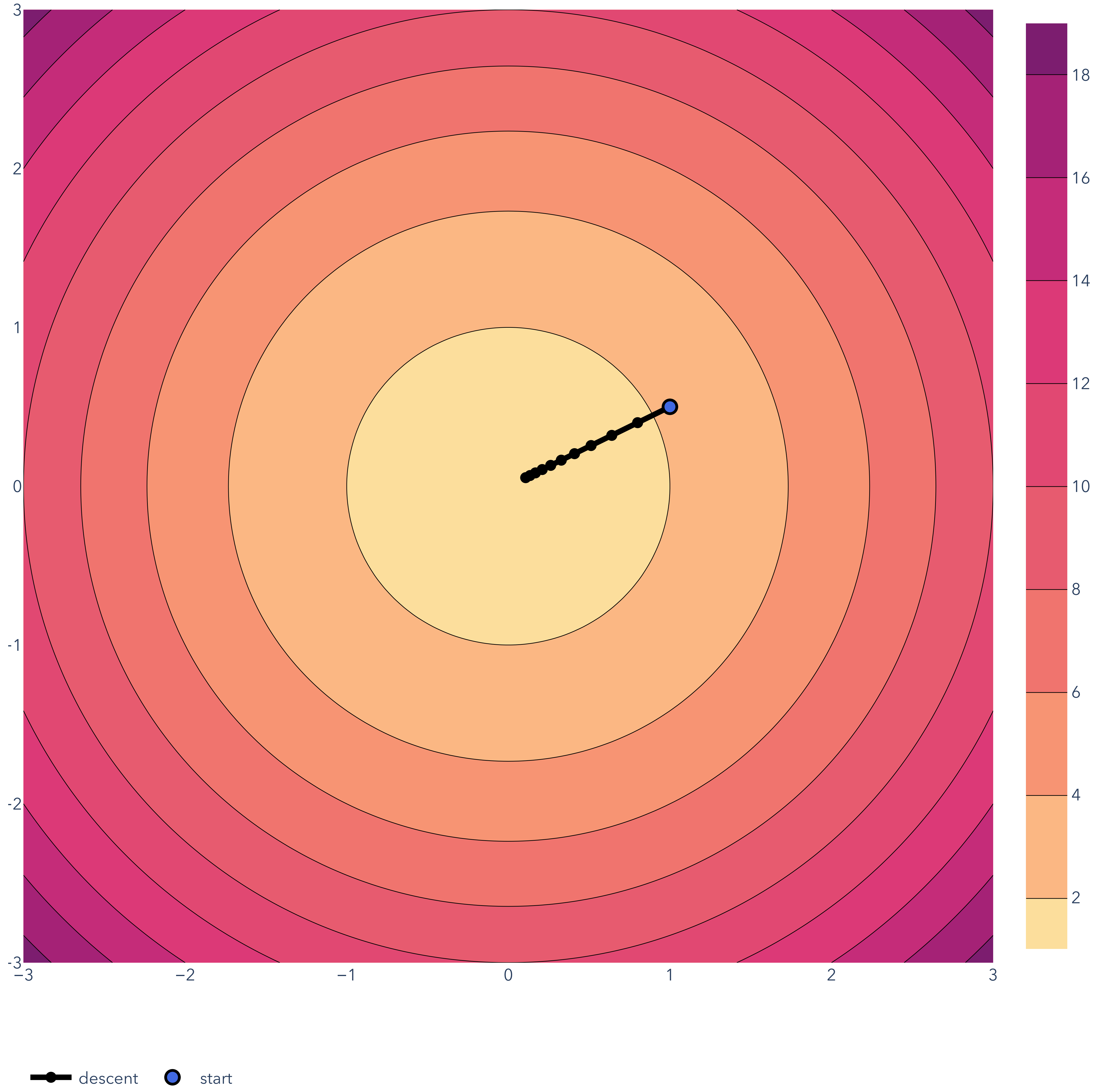
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})$…

Take another step in the direction of steepest decrease for $f(\mathbf{w})$…

⋮

Repeat until satisfied.



A candidate algorithm

Moving in steepest descent direction

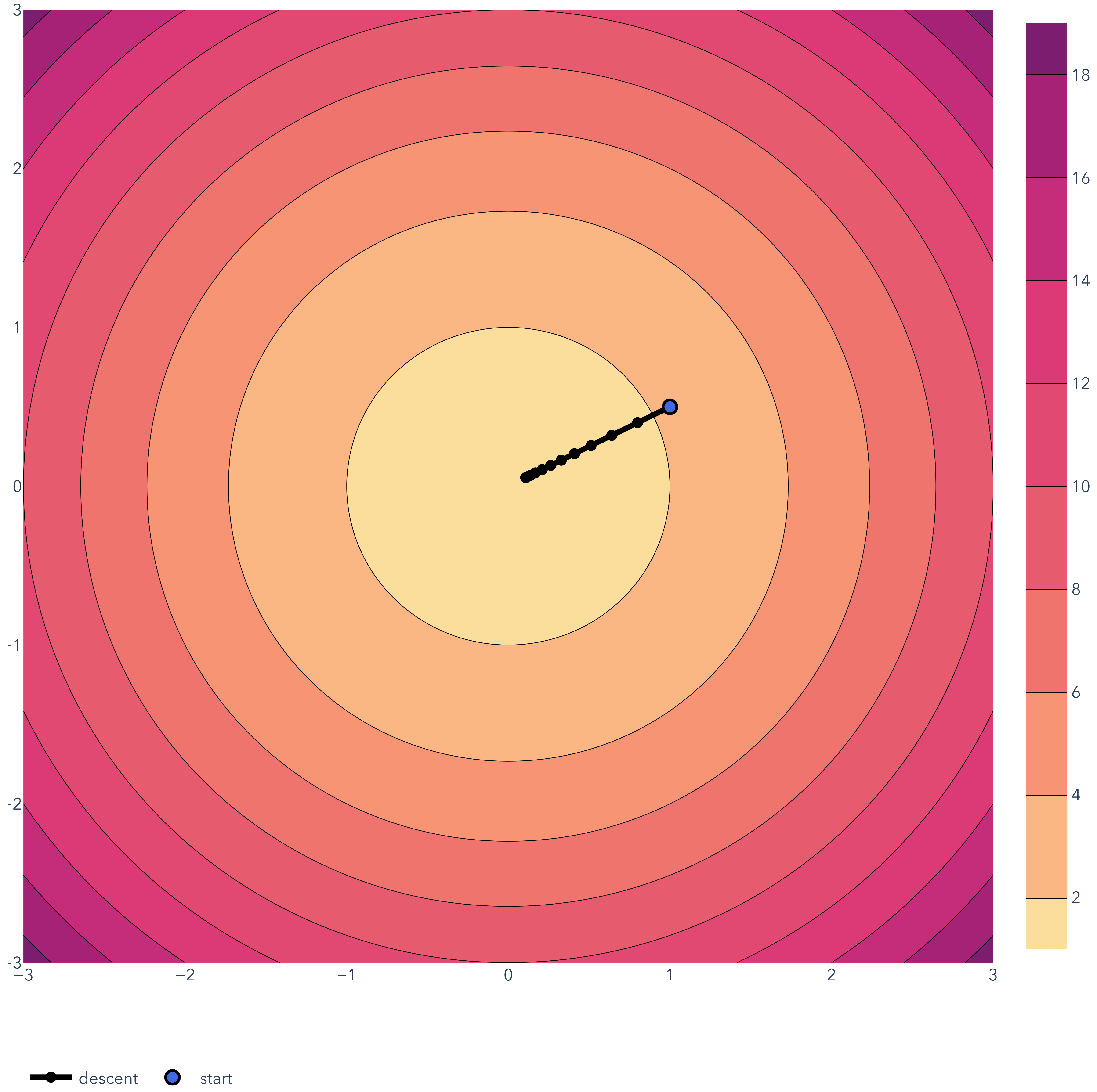
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})$…

Take another step in the direction of steepest decrease for $f(\mathbf{w})$…

⋮

Repeat until satisfied.

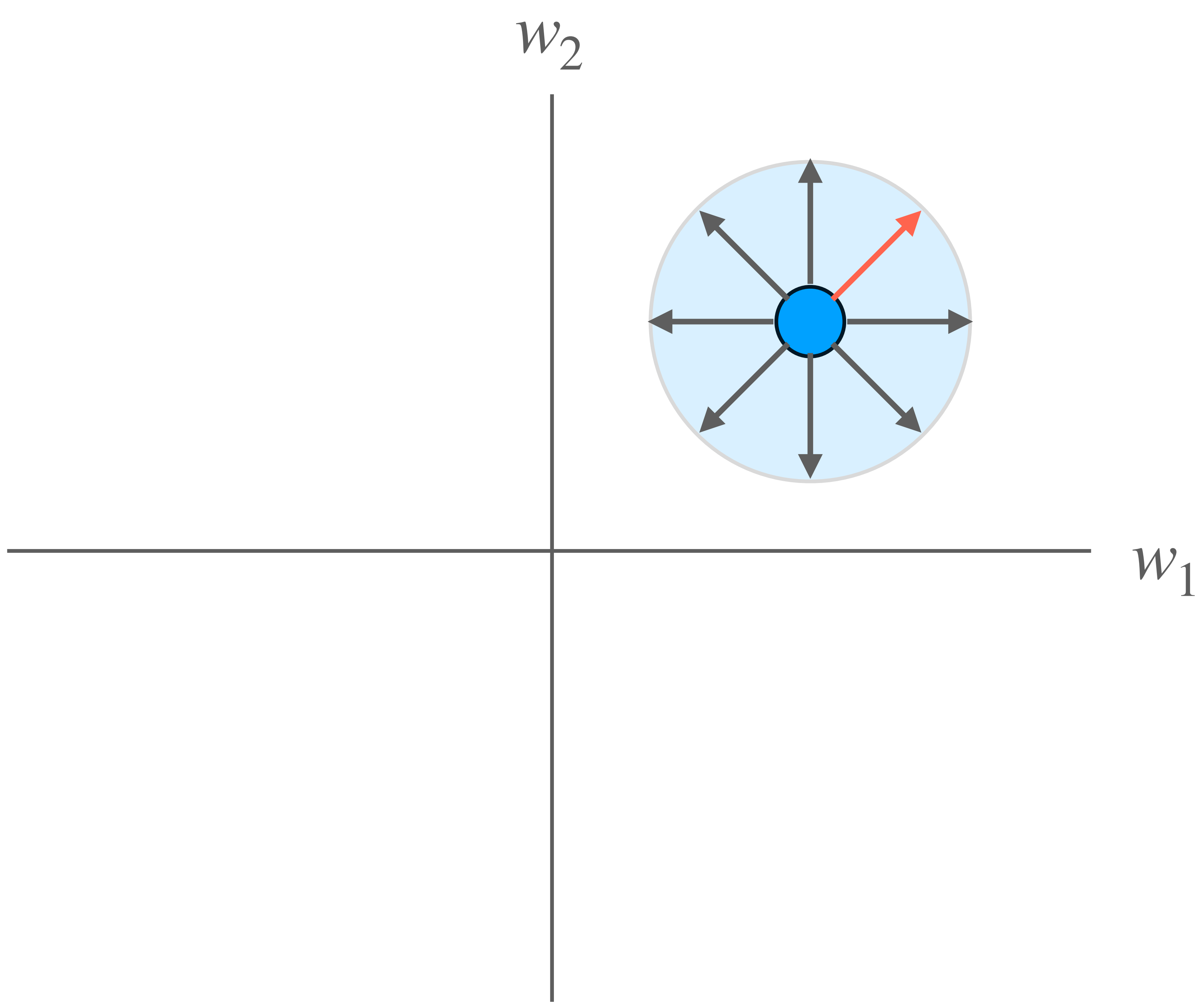


Gradient Descent
Algorithm

Gradient

The direction of steepest ascent

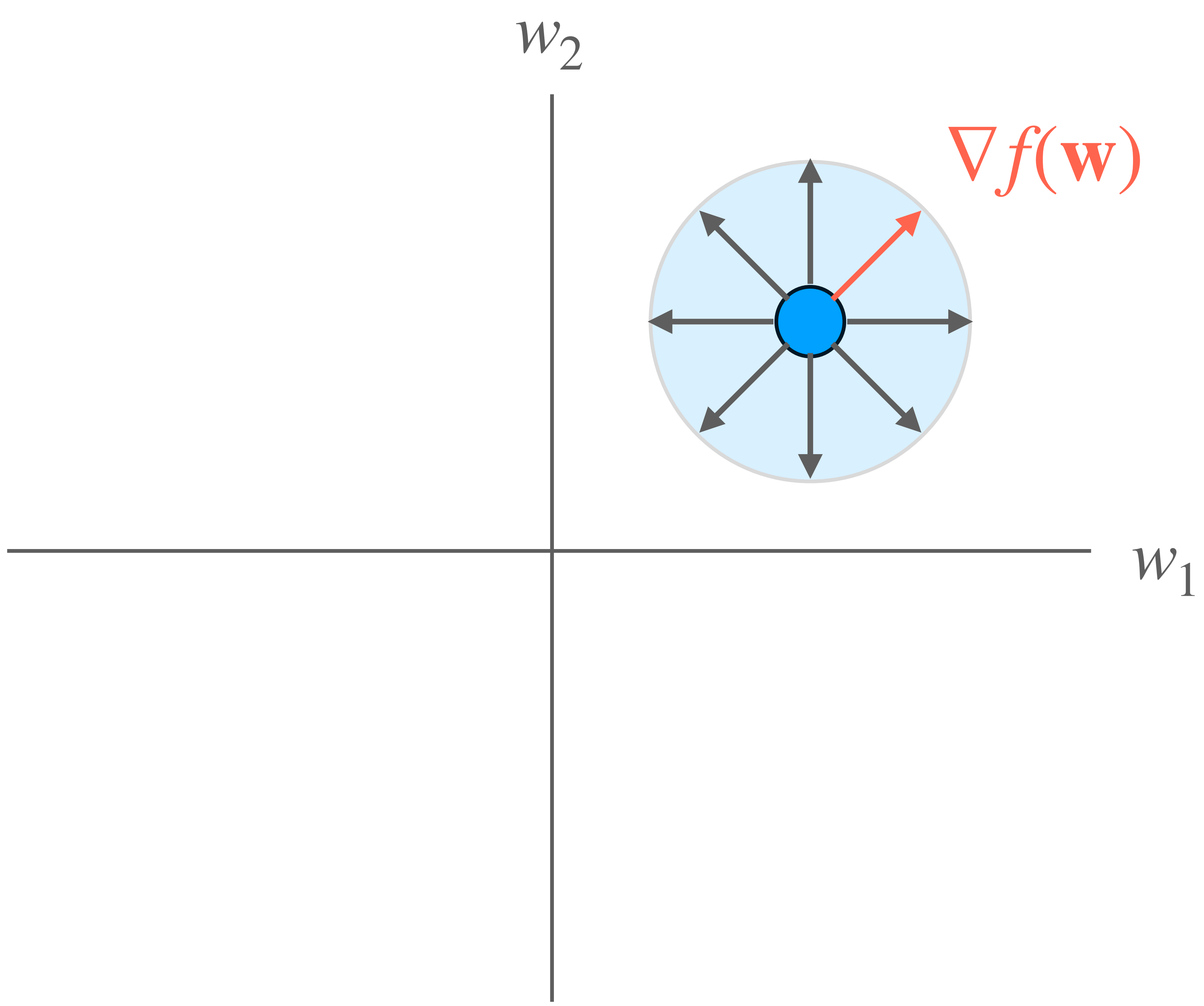
Steepest increase direction?



Gradient

The direction of steepest ascent

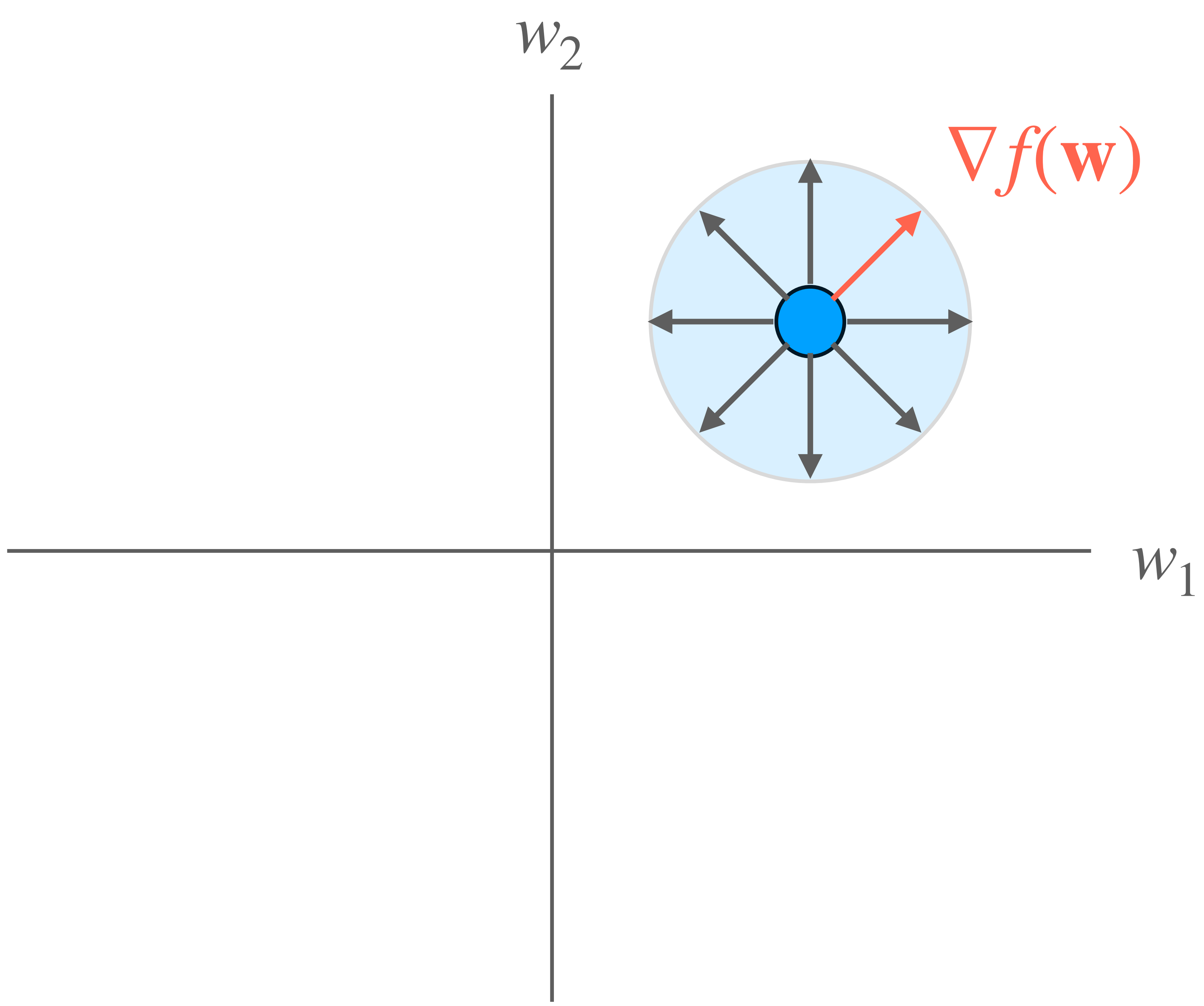
Steepest increase direction?



Gradient

The direction of steepest ascent

Steepest increase direction?

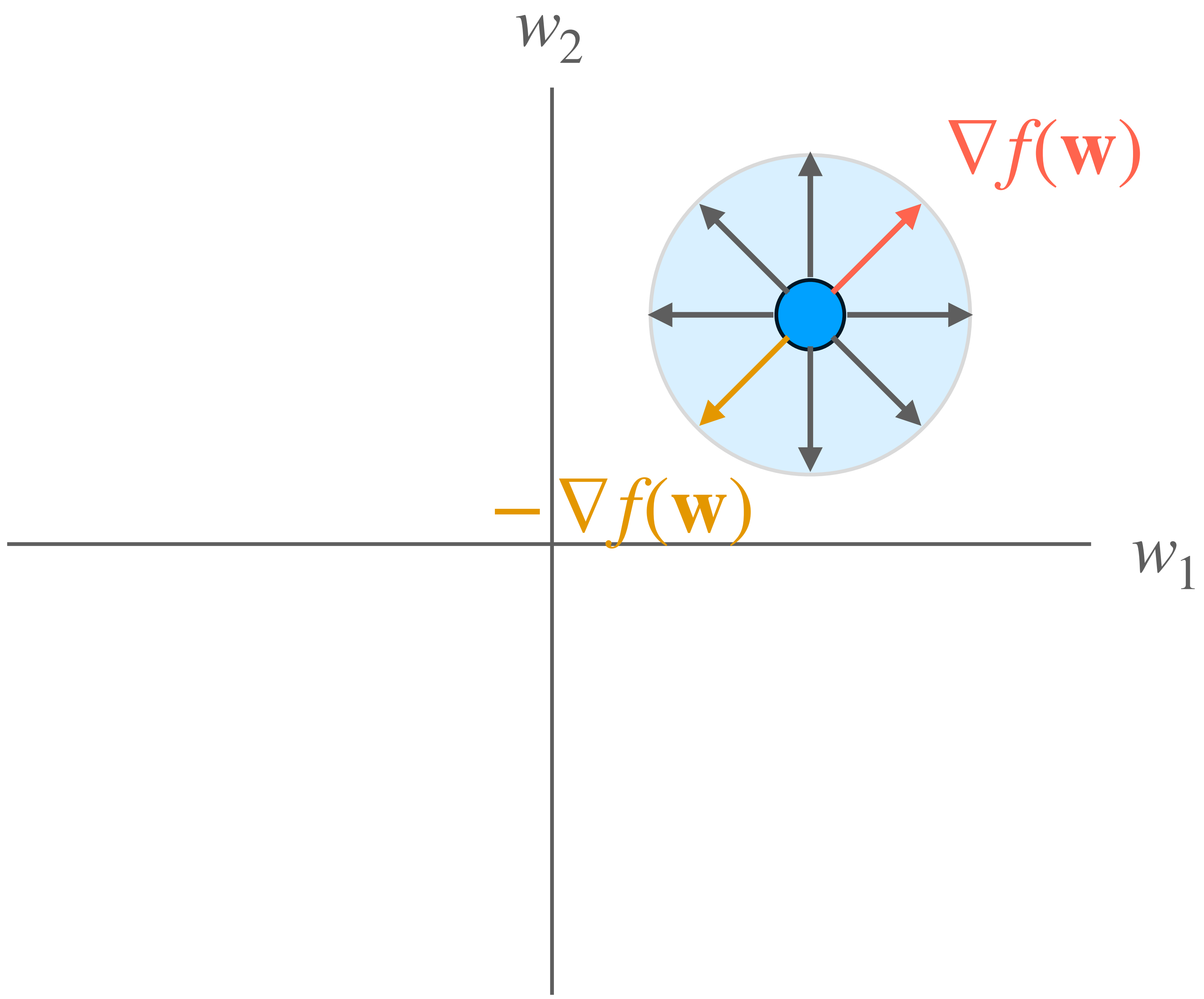


Recall: HW problem on directional derivatives!

Negative Gradient

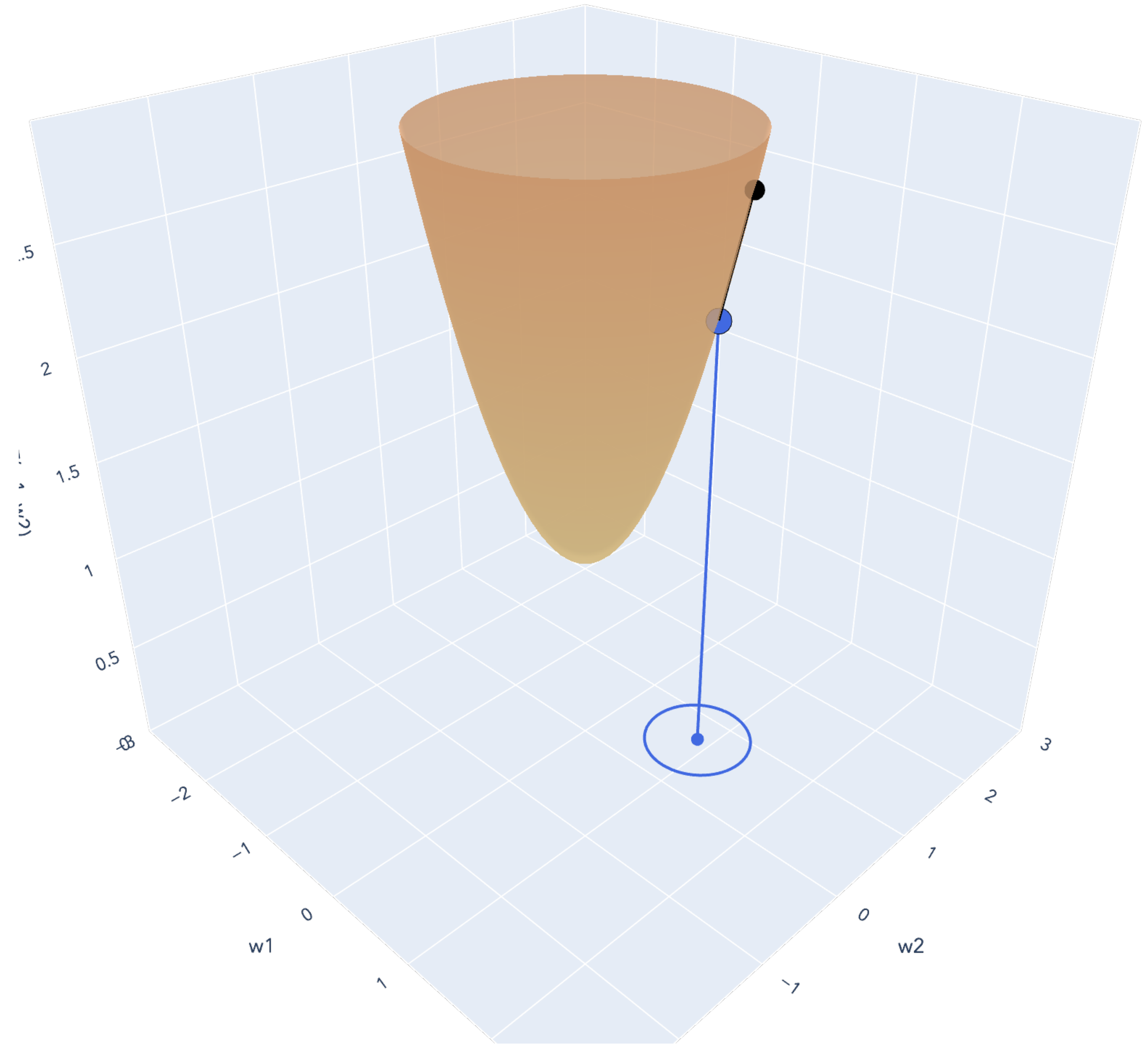
The direction of steepest ascent

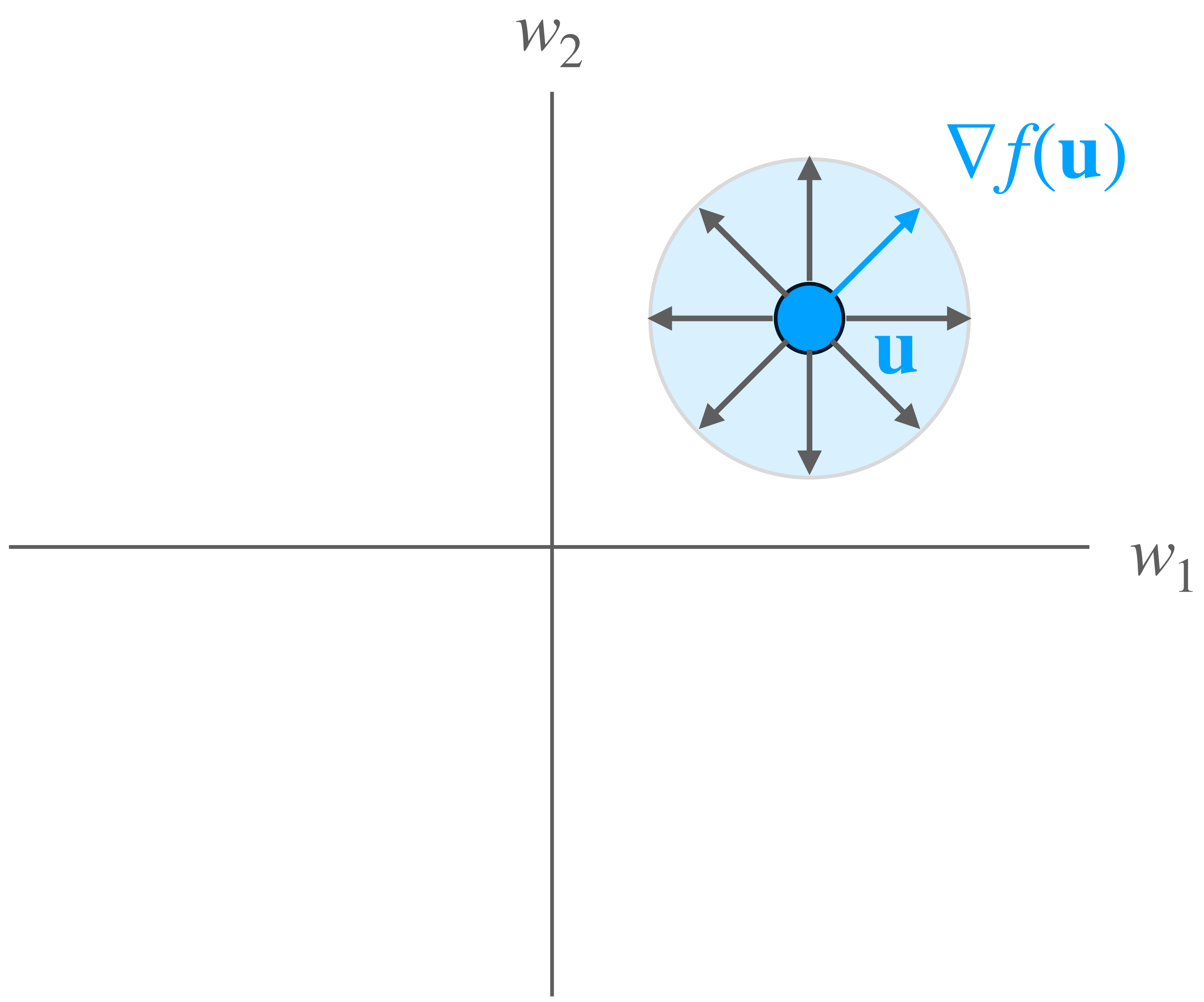
Steepest decrease direction?



Differential Calculus

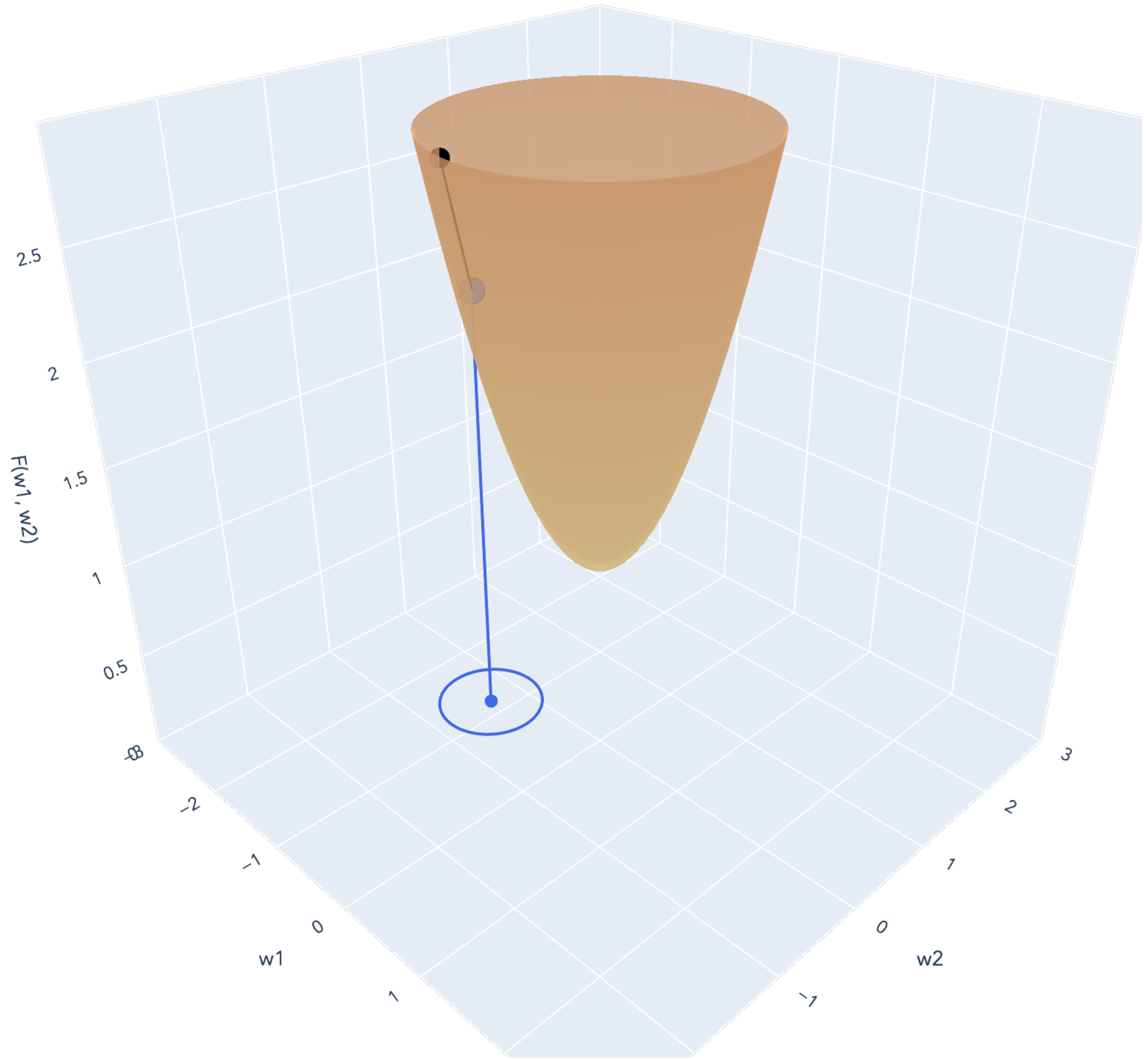
Review: Gradient

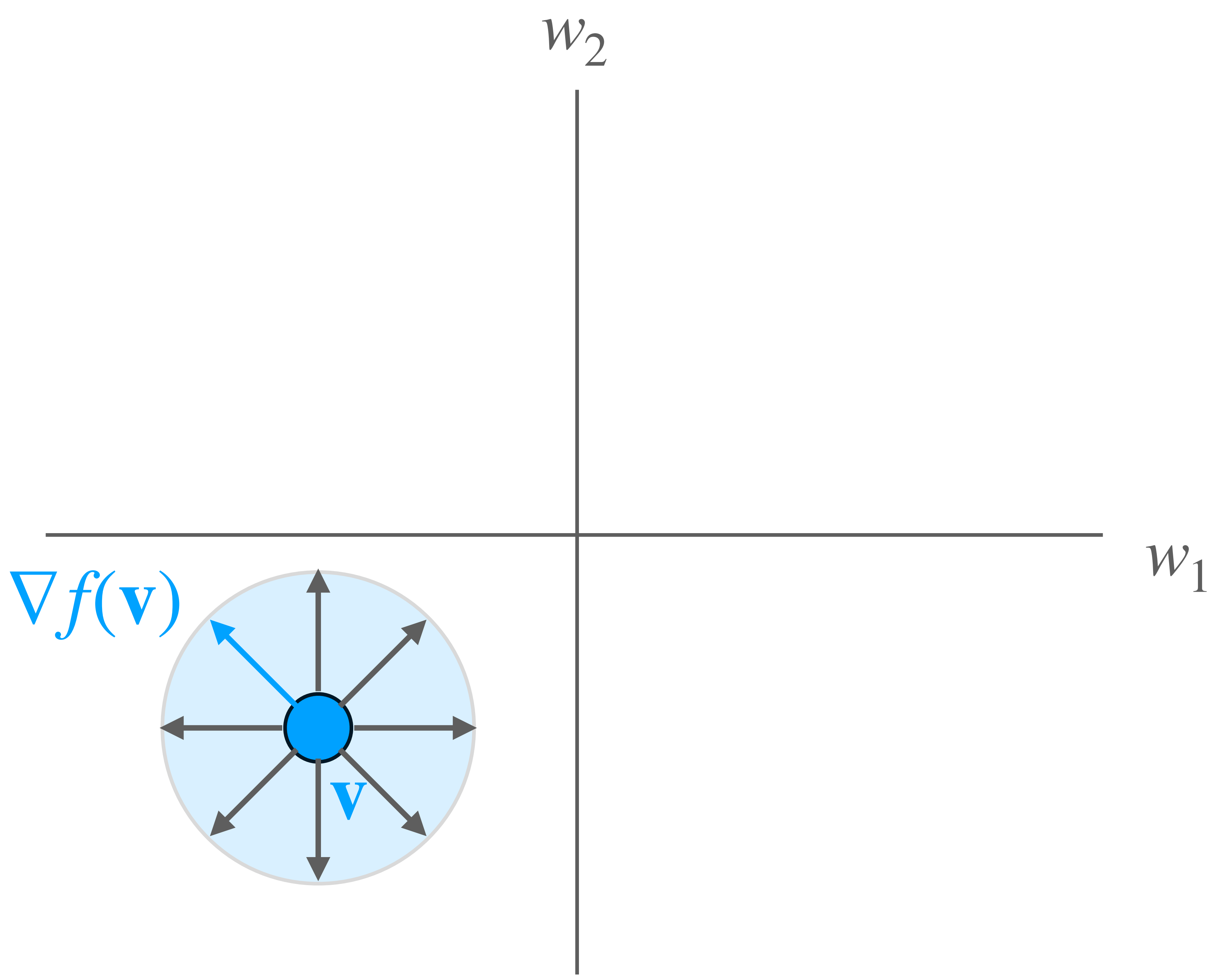




Differential Calculus

Review: Gradient





Gradient Descent

Algorithm

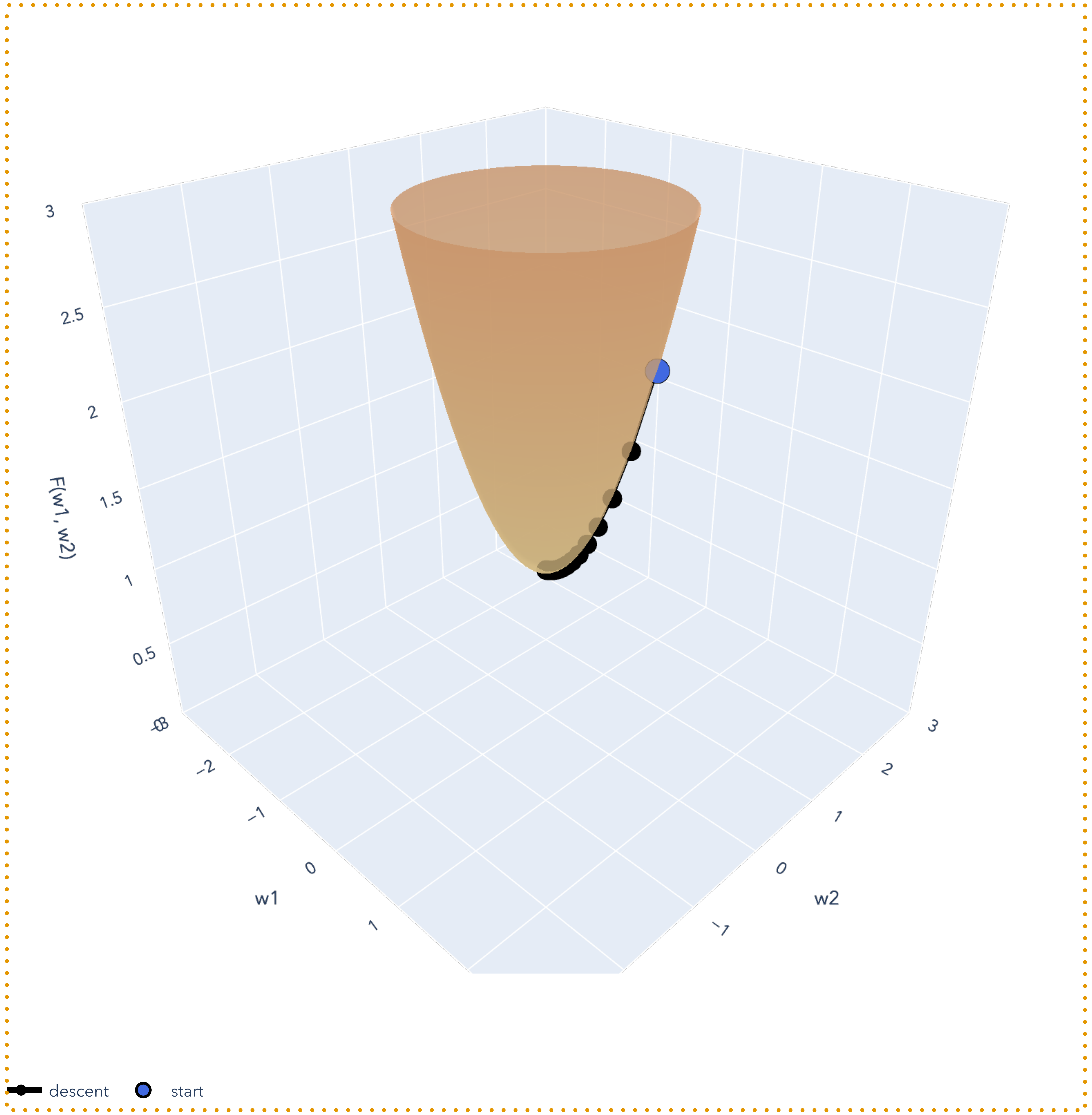
Start at some arbitrary point $\mathbf{w}^{(0)} \in \mathbb{R}^d$.

Step in the direction of steepest decrease for $f(\mathbf{w})$…

Take another step in the direction of steepest decrease for $f(\mathbf{w})$…

⋮

Repeat until satisfied.



Gradient Descent

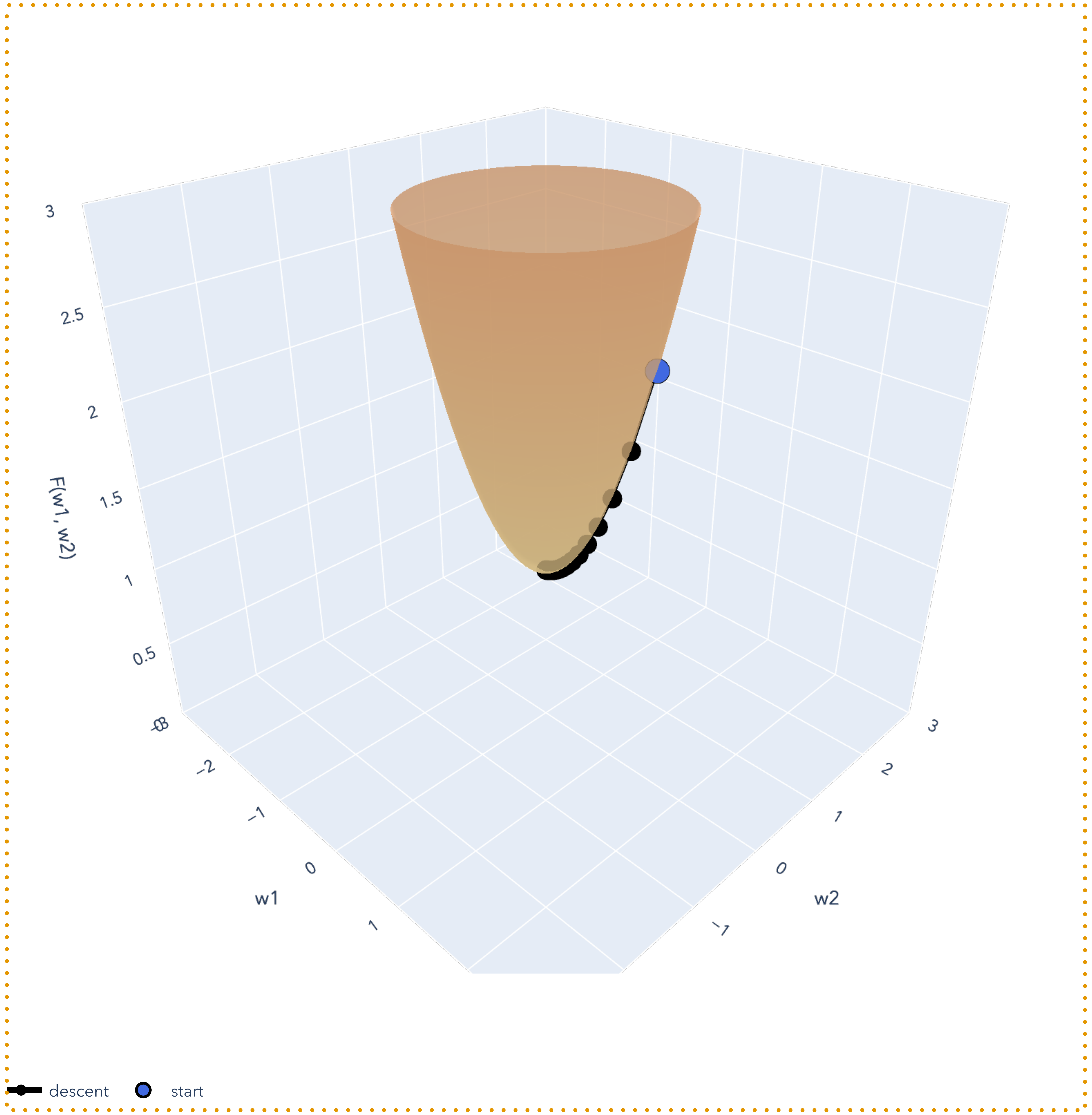
Algorithm

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

For iteration $t = 1,2,\ldots$(until "stopping condition" satisfied):

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

Return final $\mathbf{w}^{(t)}$.



Gradient Descent

Algorithm

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

For iteration $t = 1,2,\ldots$ (until "stopping condition" is satisfied):

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

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

Gradient Descent

Algorithm

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

For iteration $t = 1,2,\ldots$ (until "stopping condition" is satisfied):

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

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

Gradient Descent

Algorithm

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

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

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

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

Gradient Descent

Algorithm

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

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

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

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

learning rate

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

Gradient Descent

Algorithm

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

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

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

learning rate ($\eta > 0$)

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

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

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

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

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

update rule

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

Gradient Descent

Update rule and descent lemma

Gradient Descent

Two questions

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

1. Which direction to step in?

2. How big of a step?

Gradient Descent

Two questions

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

1. Which direction to step in?

Close to $\mathbf{w}^{(t-1)}$, the objective f "looks linear!"

2. How big of a step?

Gradient Descent

Two questions

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

1. Which direction to step in?

Close to $\mathbf{w}^{(t-1)}$, the objective f "looks linear!"

2. How big of a step?

Make η "small enough" for linear approximation to be accurate!

Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

At time t, we are at the point $\mathbf{w}^{(t-1)} \in \mathbb{R}^d$.

Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

At time t, we are at the point $\mathbf{w}^{(t-1)} \in \mathbb{R}^d$.

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

Descent Lemma

Setup and goal

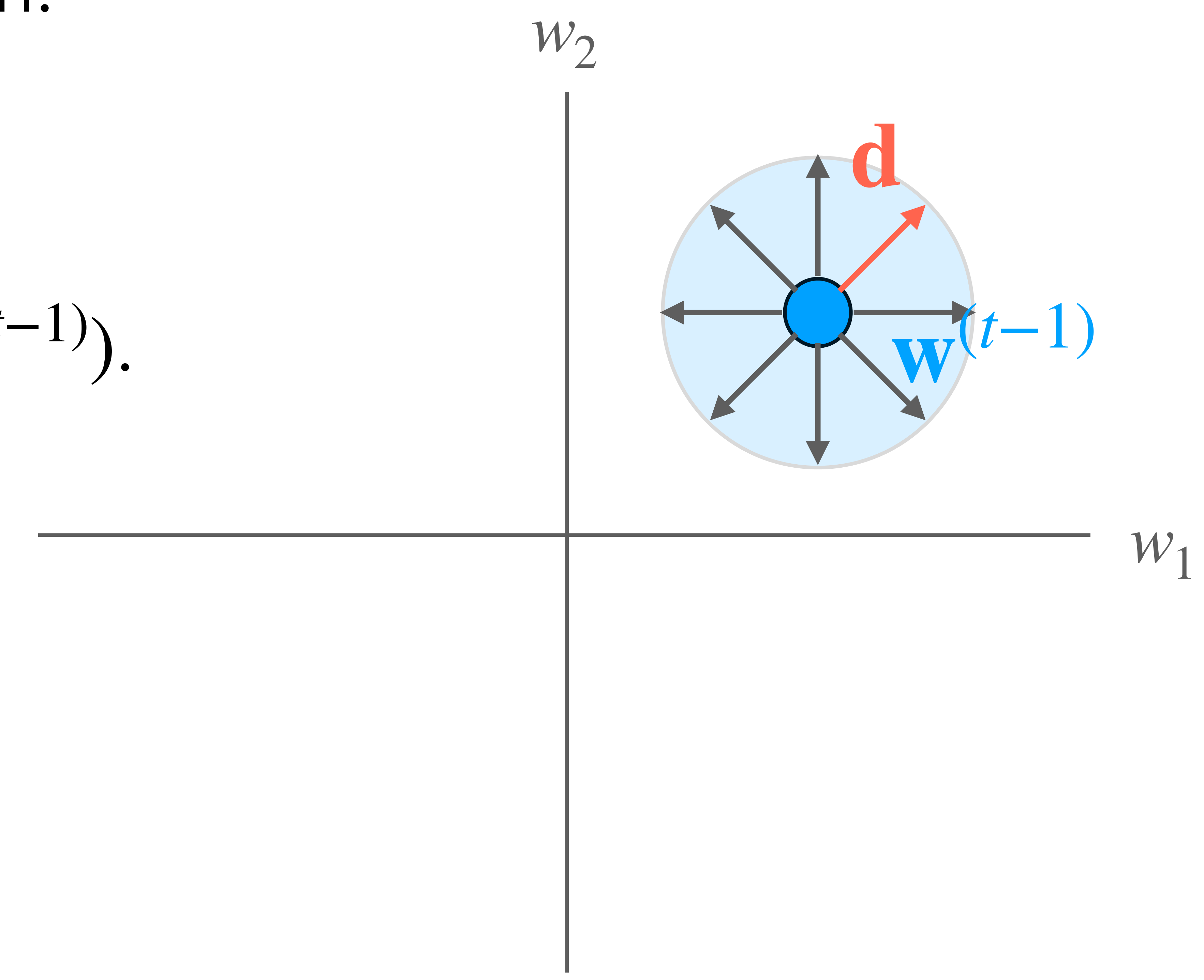
$\mathbf{w}^{(t)} \leftarrow \mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

At time t, we are at the point $\mathbf{w}^{(t-1)} \in \mathbb{R}^d$.

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.



Descent Lemma

Setup and goal

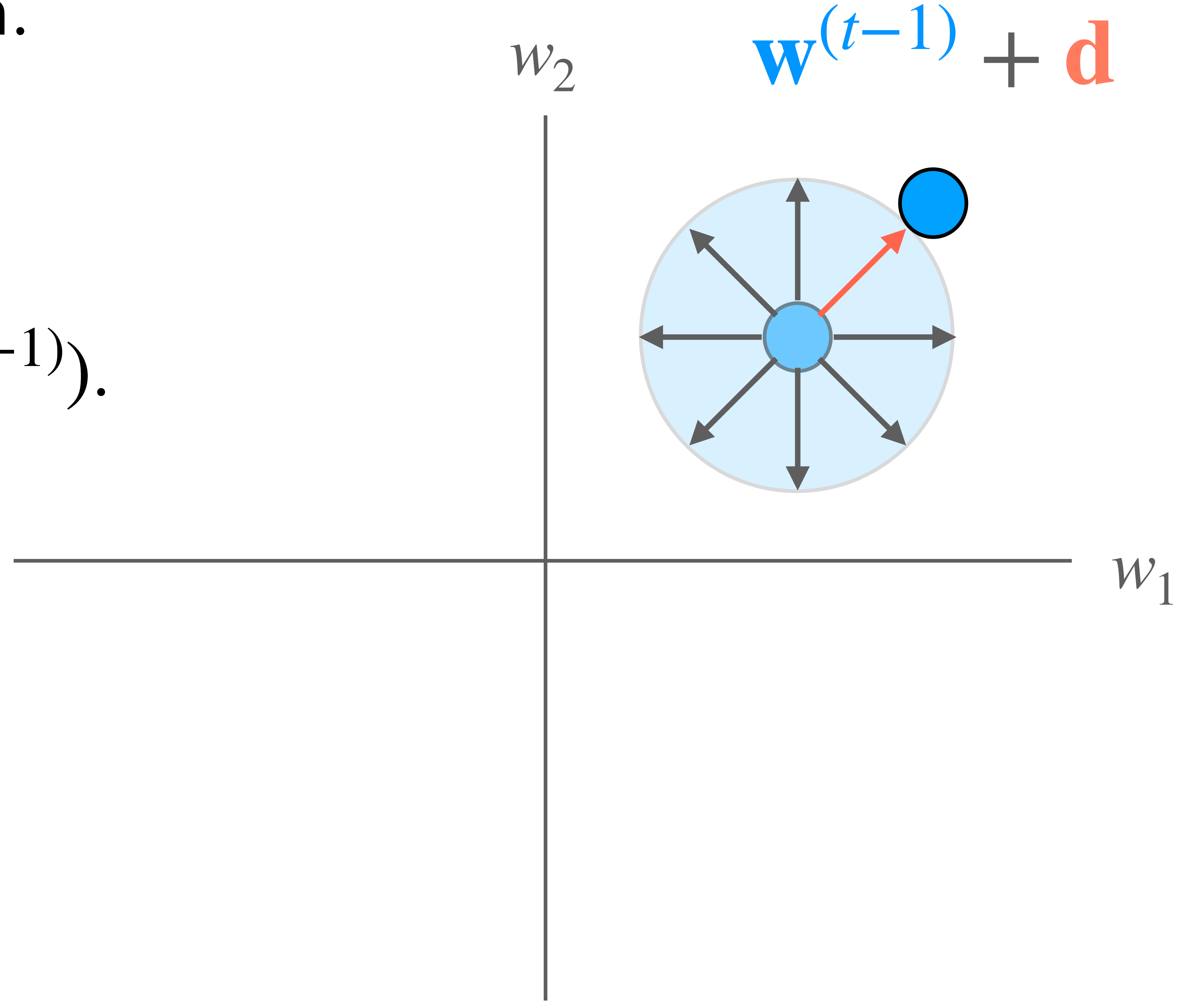
$$\mathbf{w}^{(t)} \leftarrow \mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$$

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

At time t, we are at the point $\mathbf{w}^{(t-1)} \in \mathbb{R}^d$.

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.



Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u})$$

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

At time t, we are at the point $\mathbf{w}^{(t-1)} \in \mathbb{R}^d$.

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

How about: $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$?

Descent Lemma

Setup and goal

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

Descent Lemma

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

Step 1: Take linear approximation

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) + \nabla f(\mathbf{w}^{(t-1)})^\top \left(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) - \mathbf{w}^{(t-1)}\right).$$

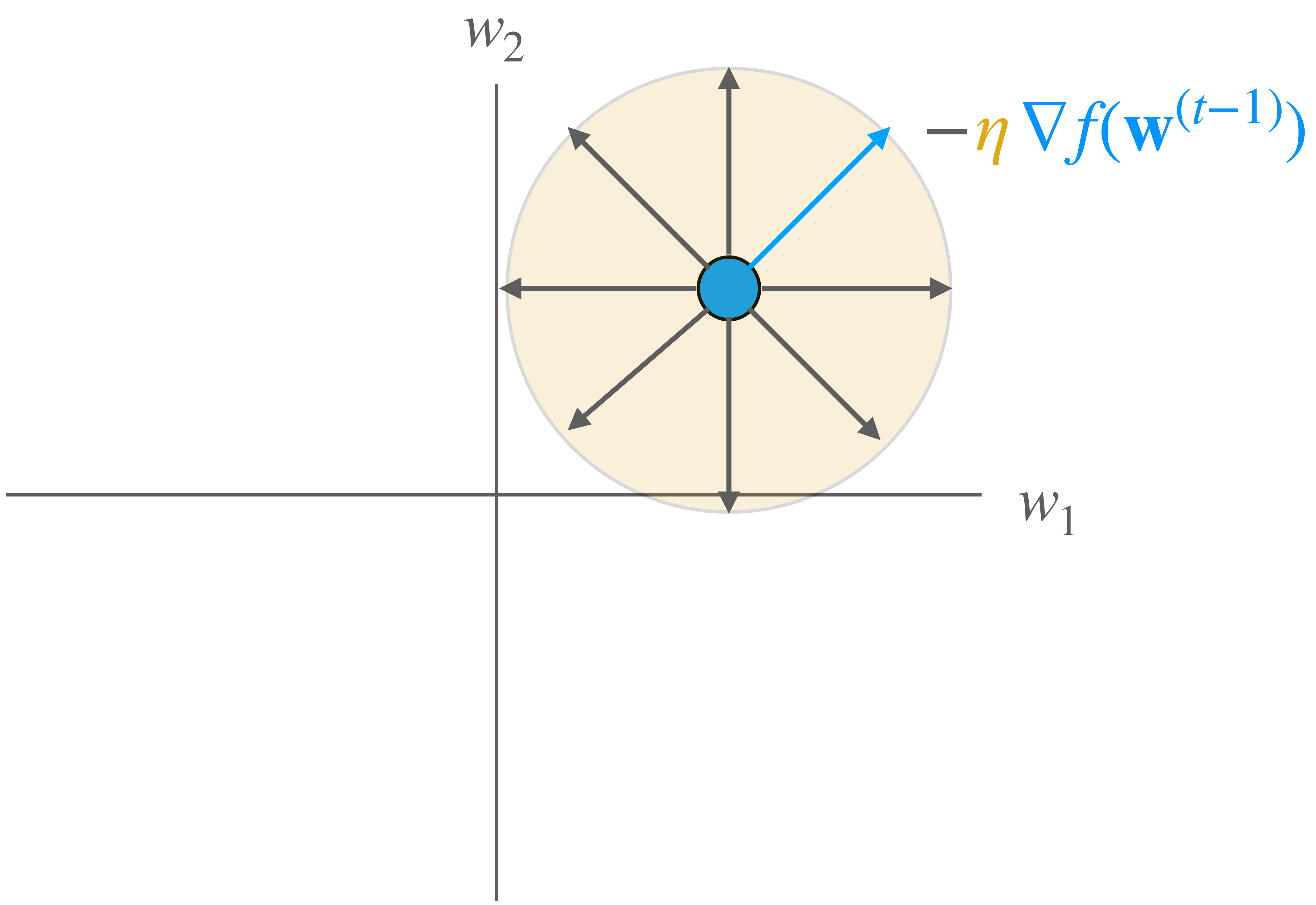
Descent Lemma

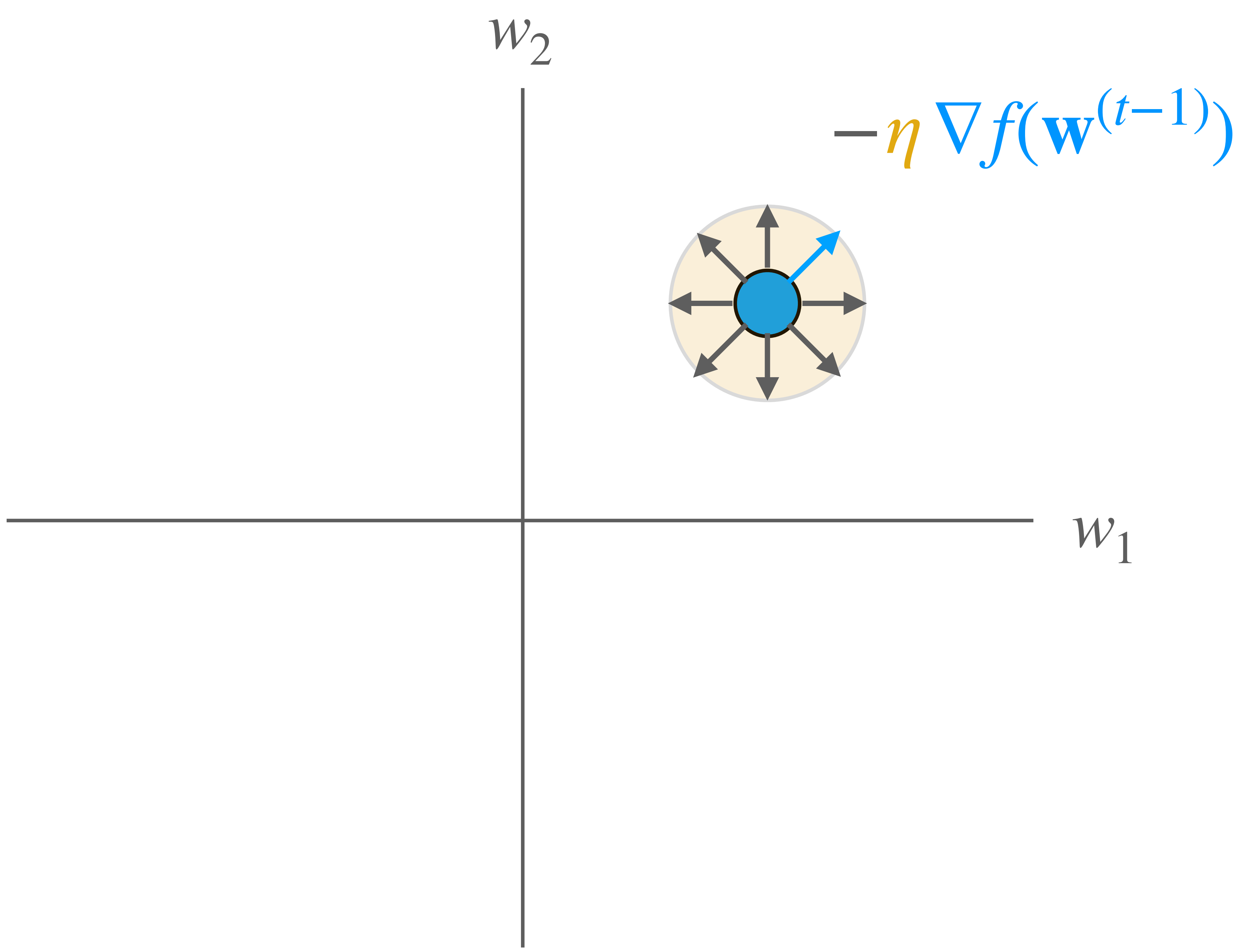
$$\mathbf{w}^{(t)} \leftarrow \mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$$

Step 1: Take linear approximation (make sure η is small)

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) + \nabla f(\mathbf{w}^{(t-1)})^\top \left(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) - \mathbf{w}^{(t-1)}\right).$$





Descent Lemma

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

Step 2: Simplify using linear algebra

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) + \nabla f(\mathbf{w}^{(t-1)})^\top \left(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) - \mathbf{w}^{(t-1)} \right).$$

Descent Lemma

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

Step 2: Simplify using linear algebra

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) + \nabla f(\mathbf{w}^{(t-1)})^\top \left(-\eta \nabla f(\mathbf{w}^{(t-1)})\right).$$

Descent Lemma

Step 3: Non-negativity of squared norm

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

recall: $\eta > 0$

Therefore,

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

Descent Lemma

Step 4: Gradient descent definition

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

Therefore,

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

Descent Lemma

Step 4: Gradient descent definition

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

Therefore,

$$f(\mathbf{w}^{(t)}) \lessapprox f(\mathbf{w}^{(t-1)})!$$

Descent Lemma

Conclusion

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

Therefore,

$$f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^{(t-1)}) \text{ as long as } \eta \text{ is sufficiently small!}$$

Gradient Descent

Two questions

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

1. Which direction to step in?

Close to $\mathbf{w}^{(t-1)}$, the objective f "looks linear" so we can follow the gradient!

2. How big of a step?

Make η "small enough" for linear approximation to be accurate!

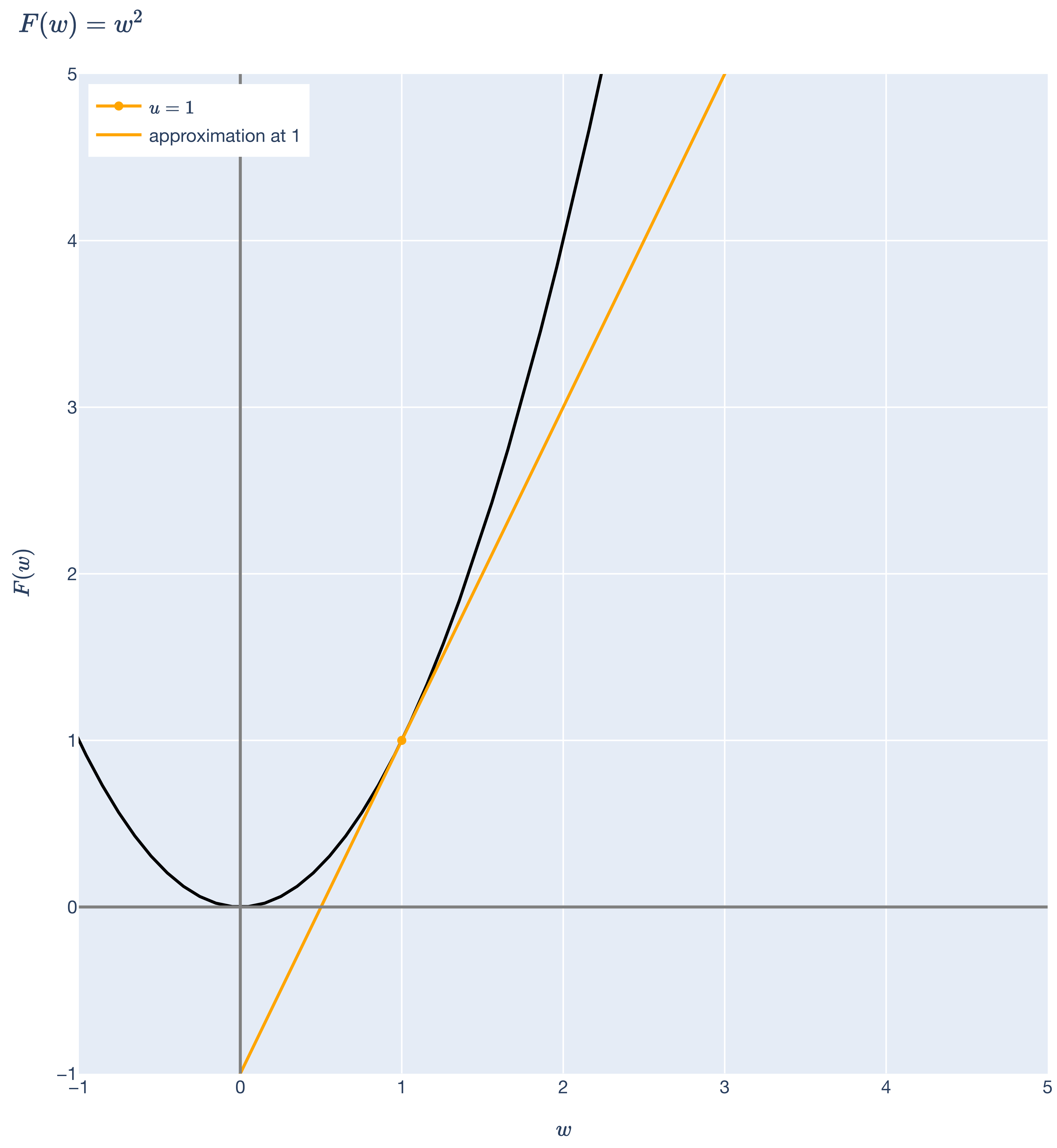
Descent Lemma

Q2: How big of a step?

If η is small enough, then:

$$\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) \text{ is close to } \mathbf{w}^{(t-1)}$$

and our linear approximation is good…



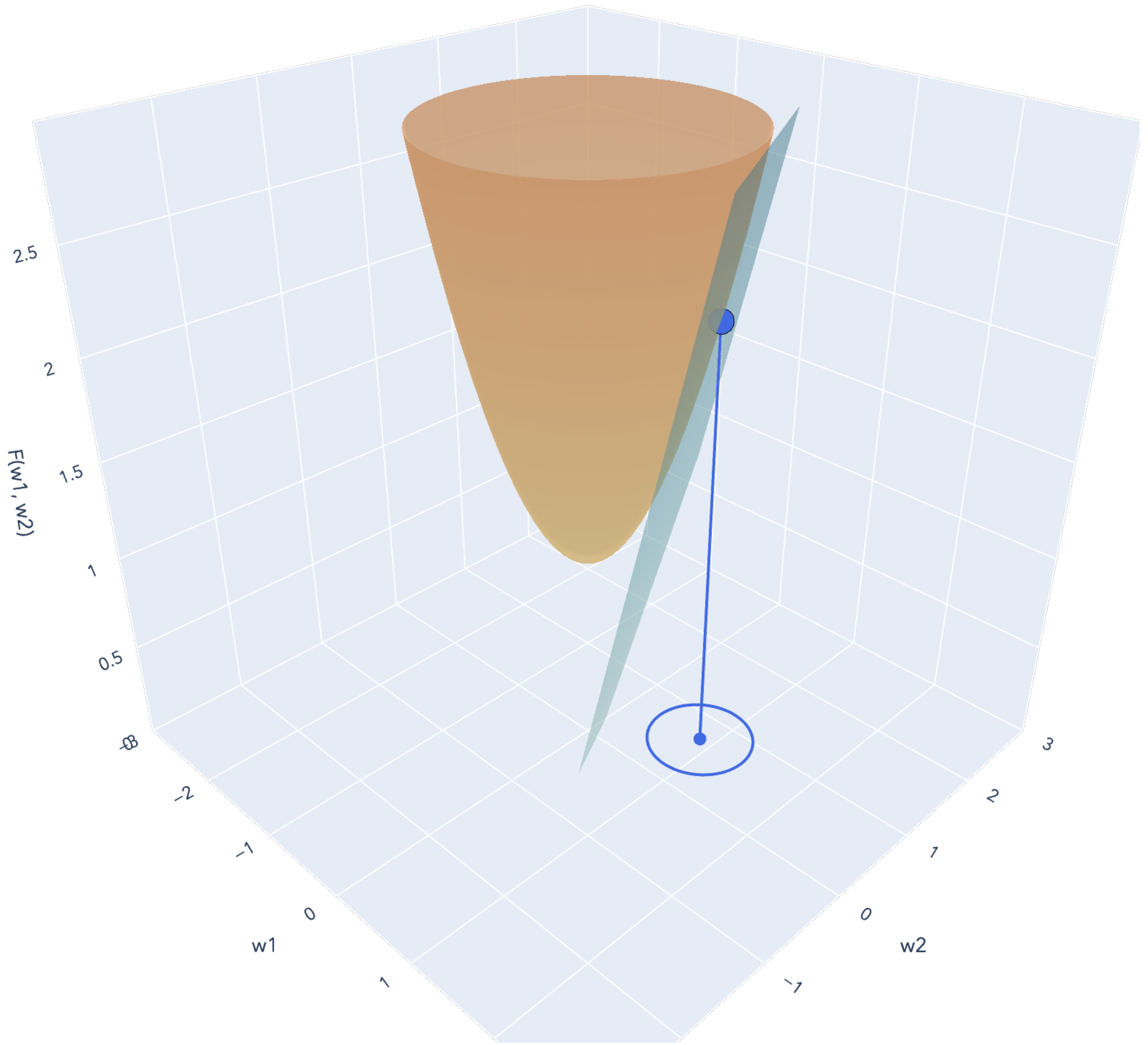
Descent Lemma

Q2: How big of a step?

If η is small enough, then:

$$\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) \text{ is close to } \mathbf{w}^{(t-1)}$$

and our linear approximation is good…



Descent Lemma

Q1: Which direction to step in?

…so we can "replace"

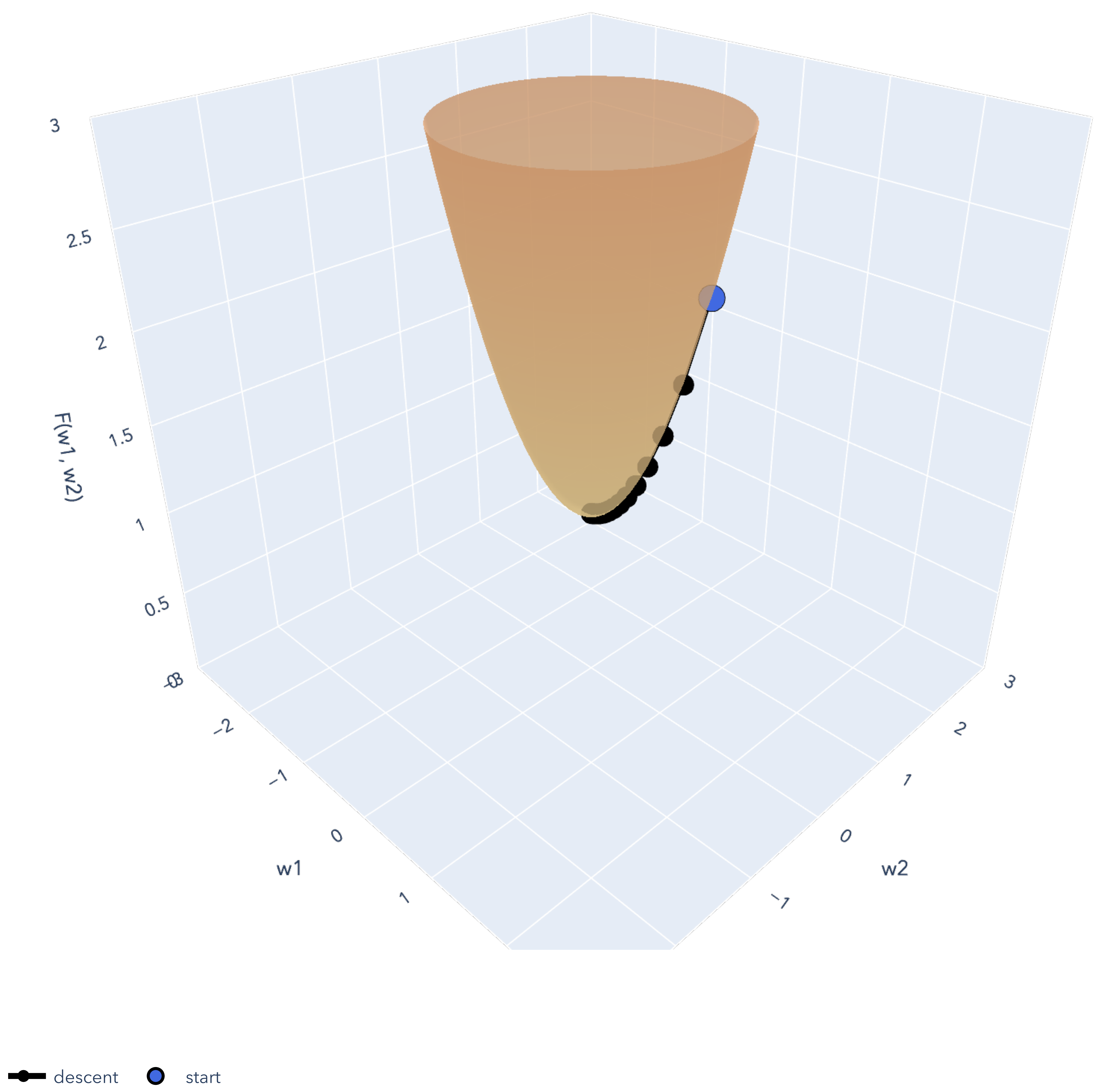
$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}))$$

and instead reason about

$$f(\mathbf{w}^{(t-1)}) + \nabla f(\mathbf{w}^{(t-1)})^{\top} \left(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)}) - \mathbf{w}^{(t-1)} \right)$$

to conclude

$f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^{(t-1)})$ as long as η is small!



Descent Lemma

Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

Descent Lemma

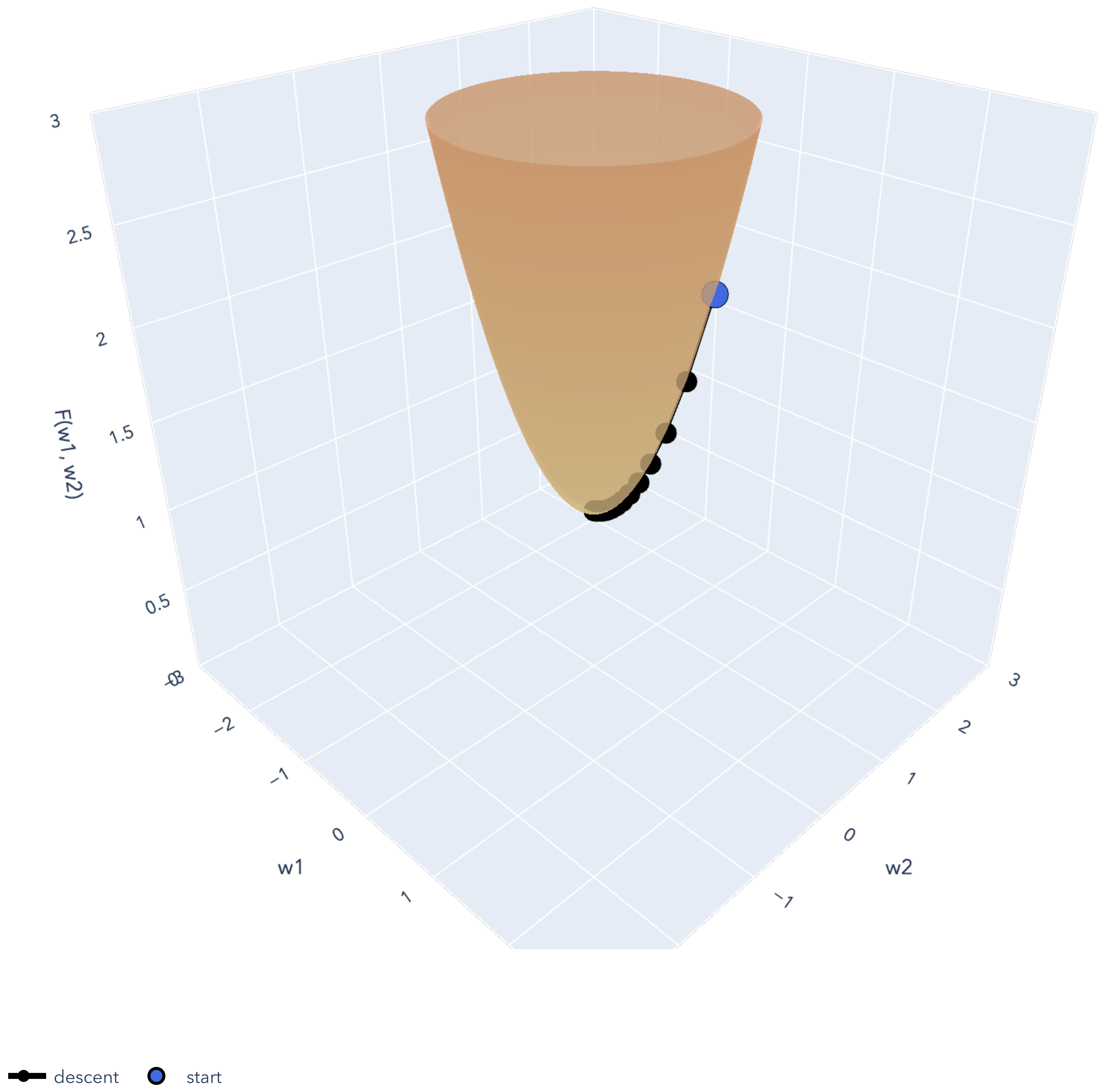
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Descent Lemma

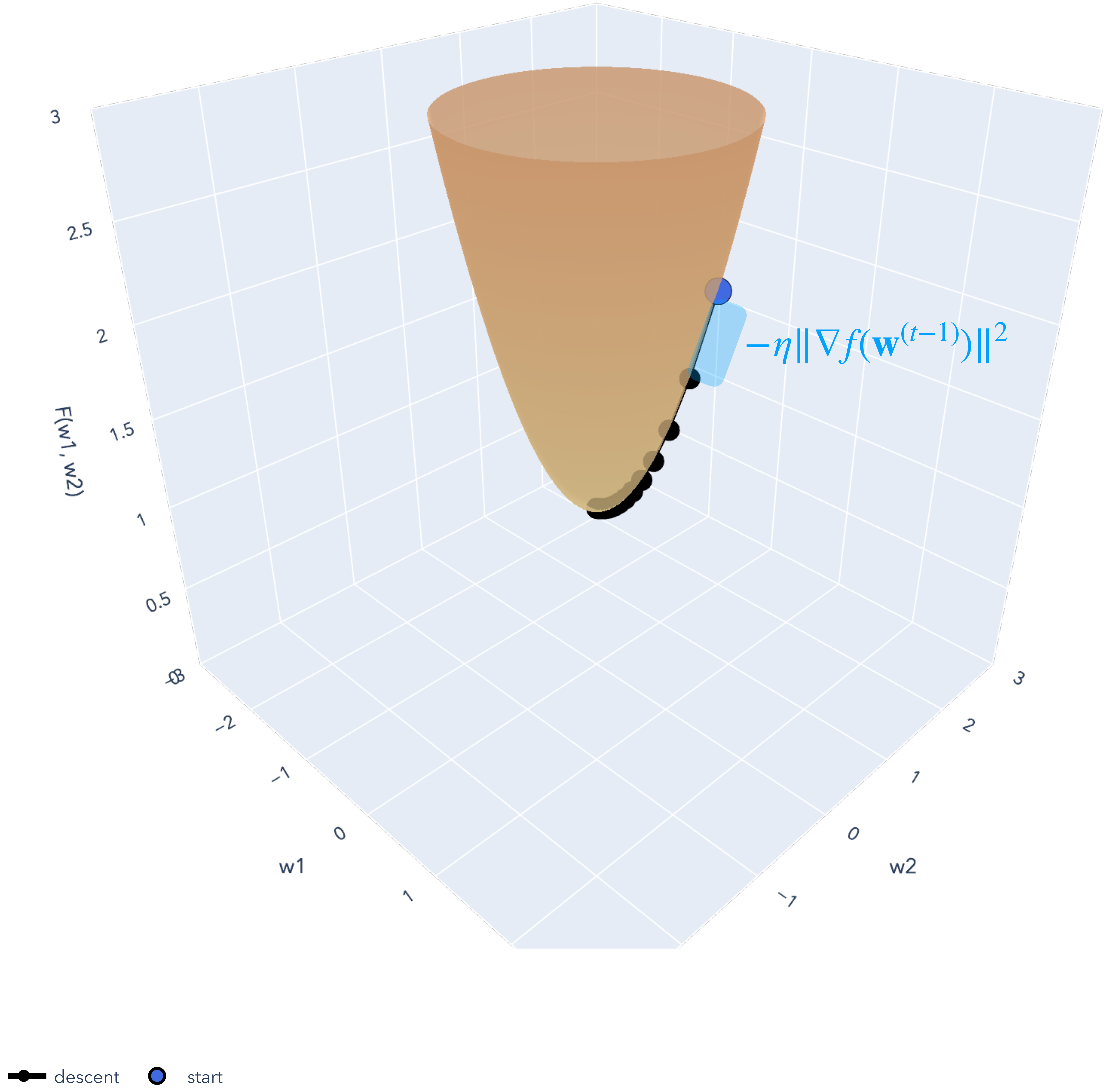
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Descent Lemma

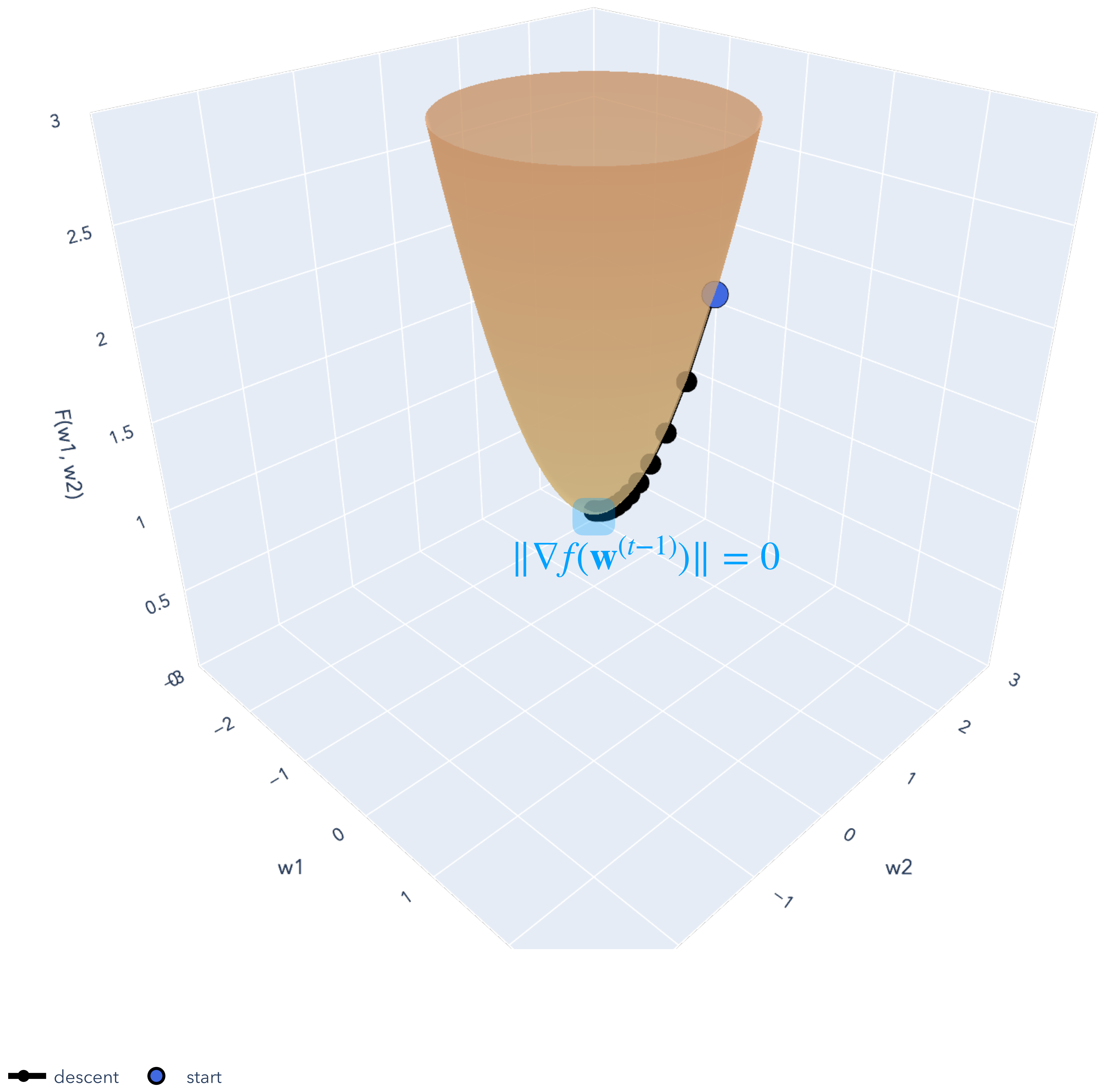
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Descent Lemma

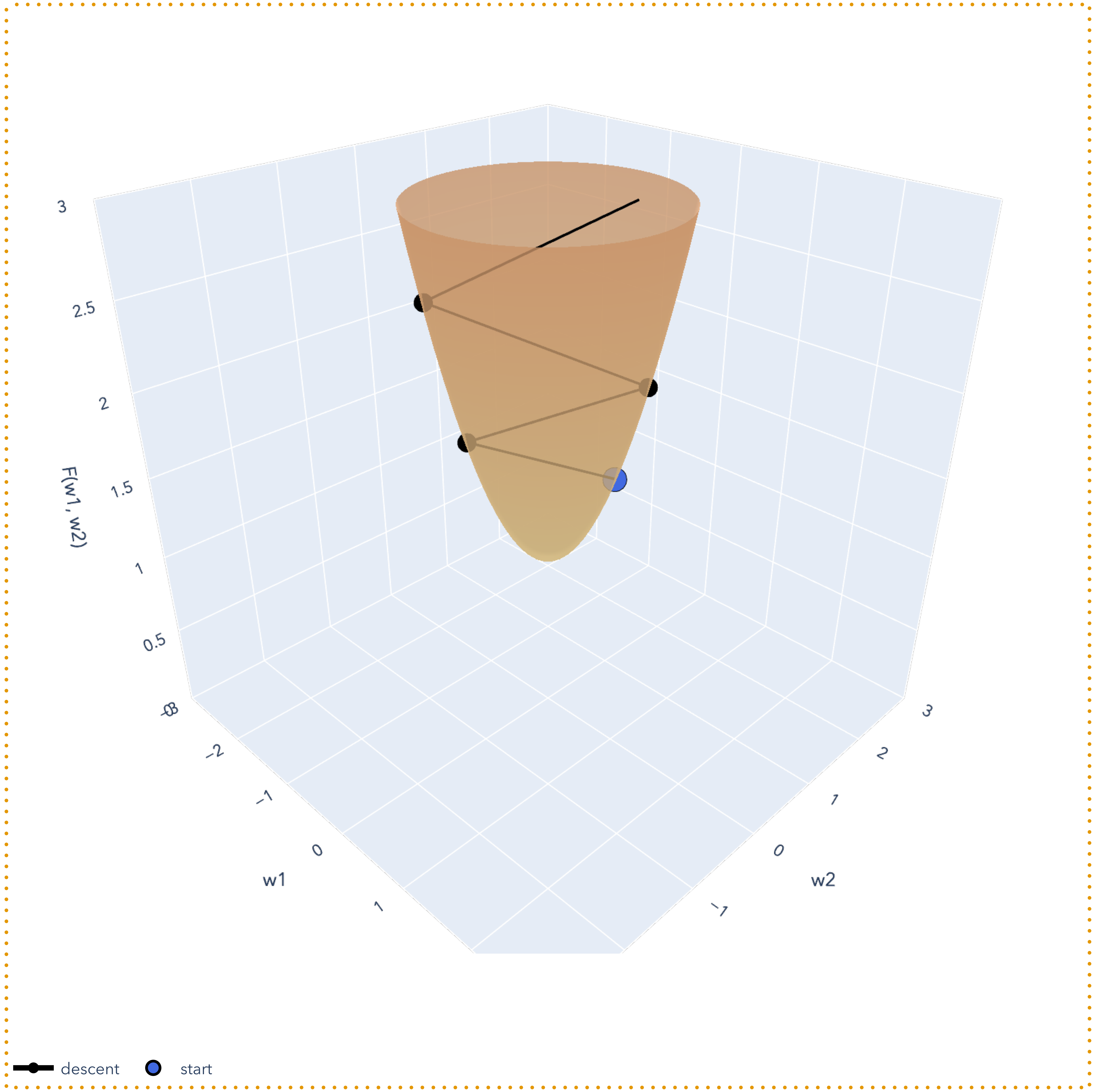
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Gradient Descent Guarantees

Theorem 1: Descent Lemma

Theorem (Descent Lemma). If f is "smooth enough," then there is a choice of $\eta > 0$ such that, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{\eta}{2} \|\nabla f(\mathbf{w})\|^2.$$

"Smooth enough" : f is a β-smooth function.

Taylor's Theorem: makes the $\lessapprox$ rigorous!

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.

"Smooth" varies in context. 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:

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

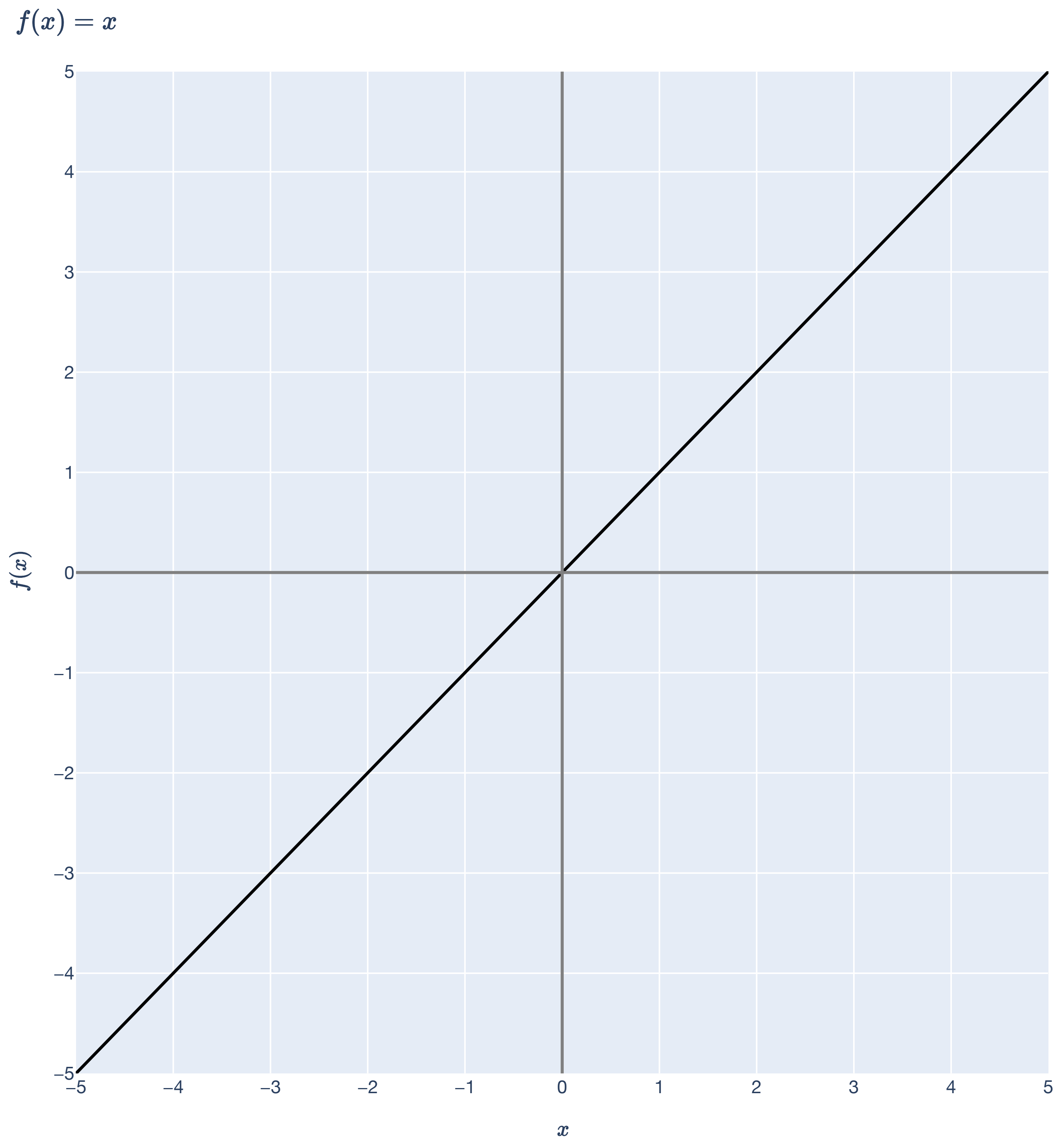
where $a_m, \ldots, a_0 \in \mathbb{R}$ are the *coefficients* of the polynomial.

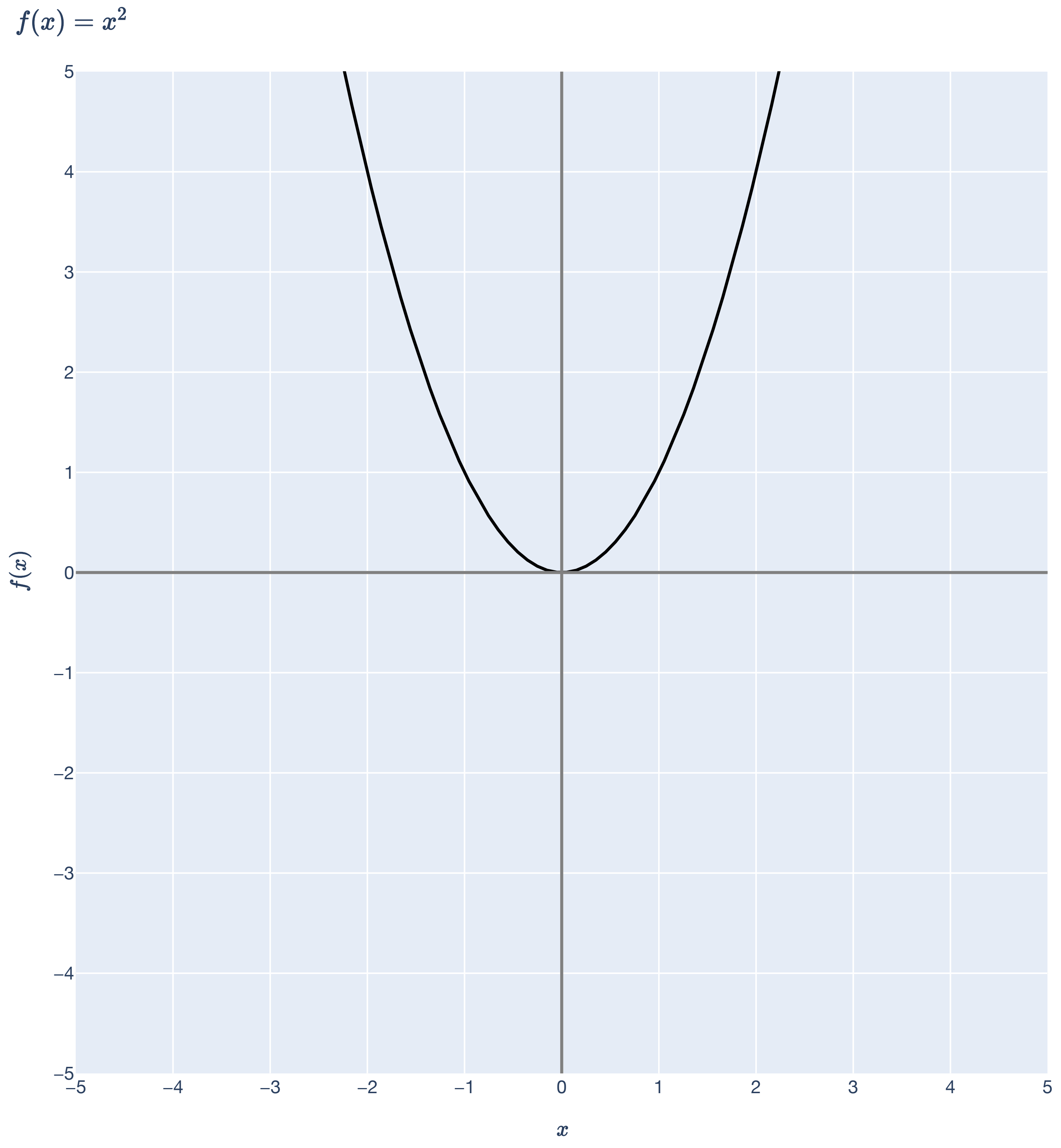
Example: $f(x) = 4x^3 + 2x - 1.$

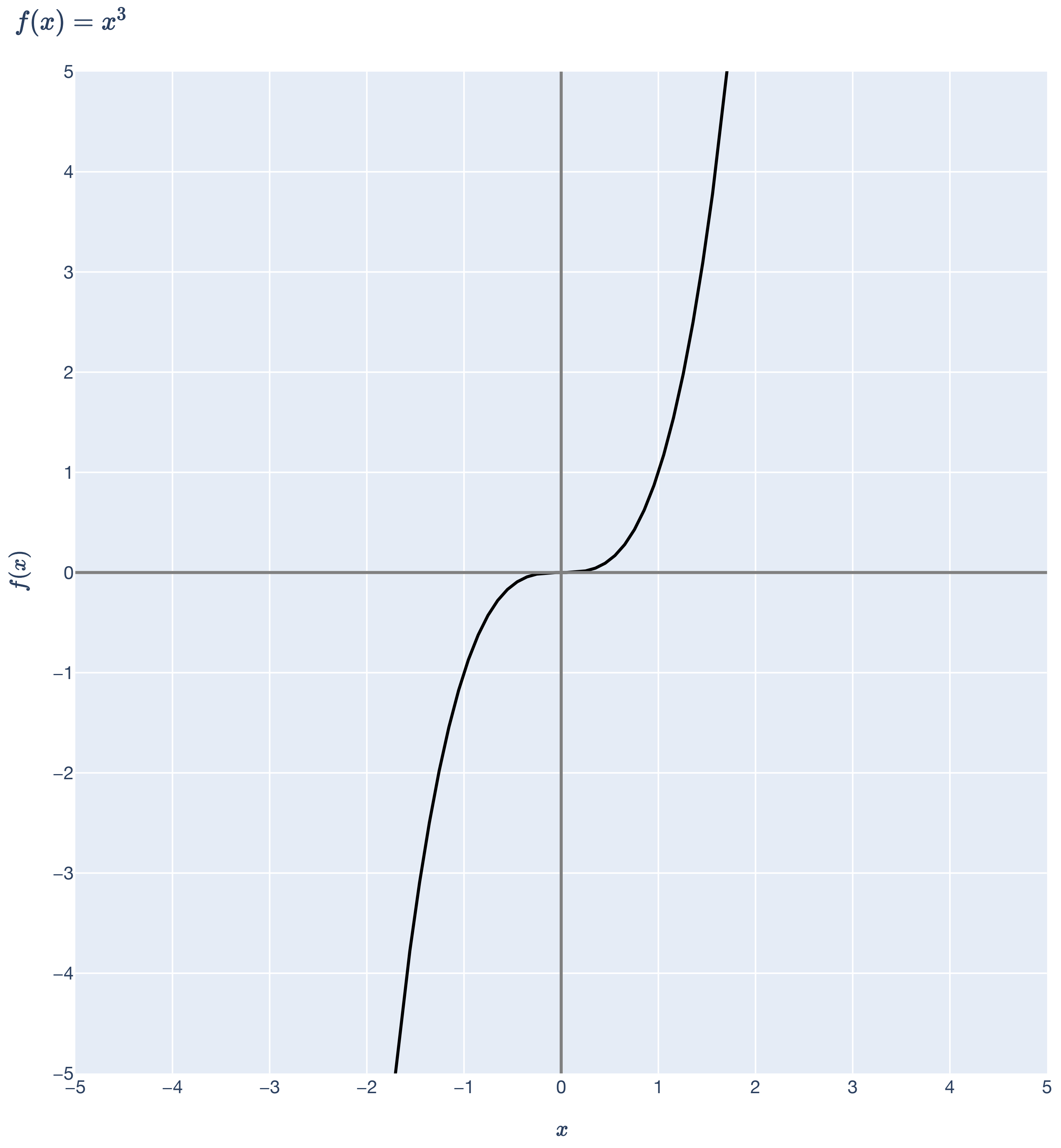
Polynomials

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

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

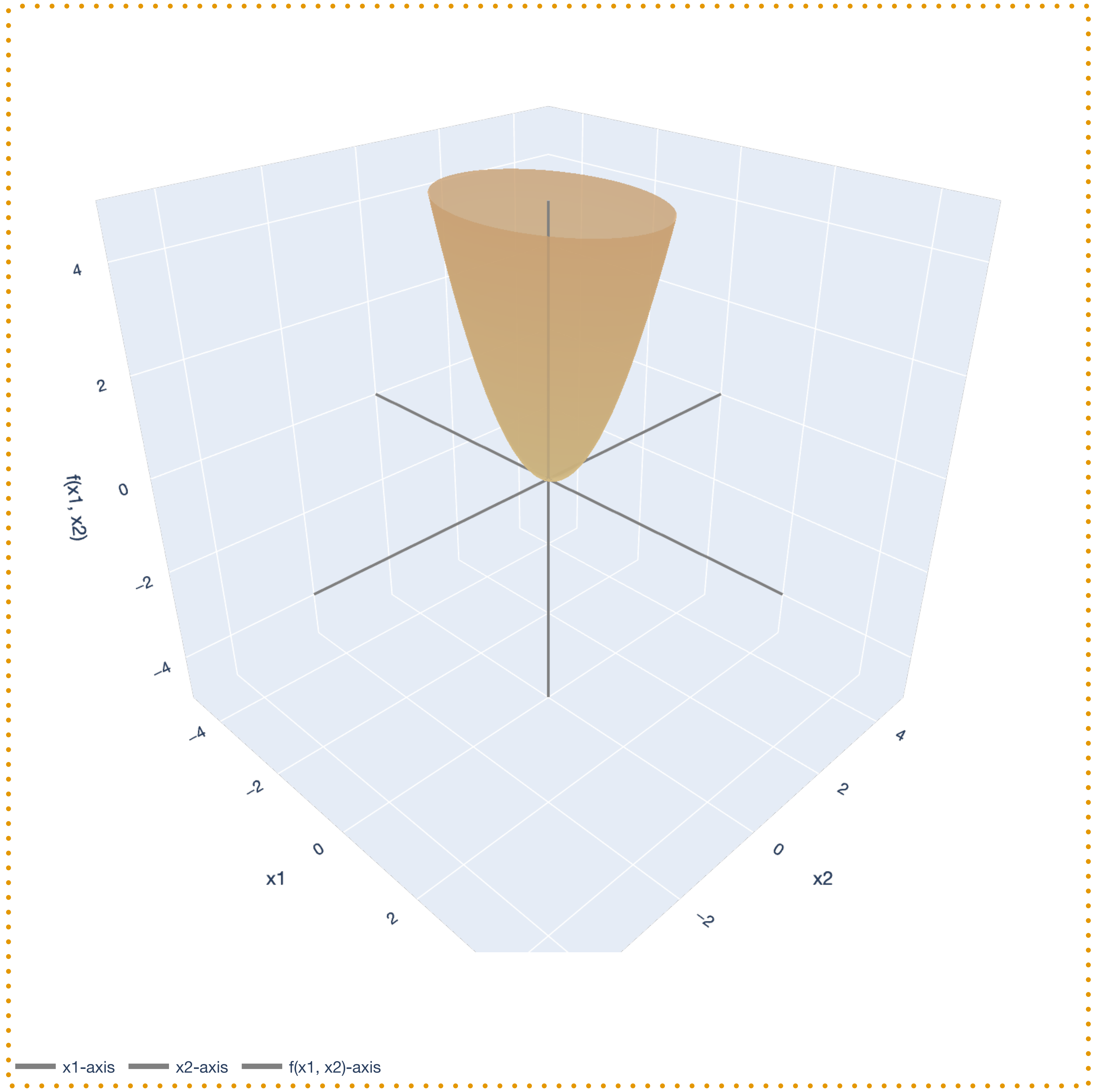
A polynomial function is a function $f : \mathbb{R}^d \to \mathbb{R}$ is a finite sum of monomials with real coefficients.

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

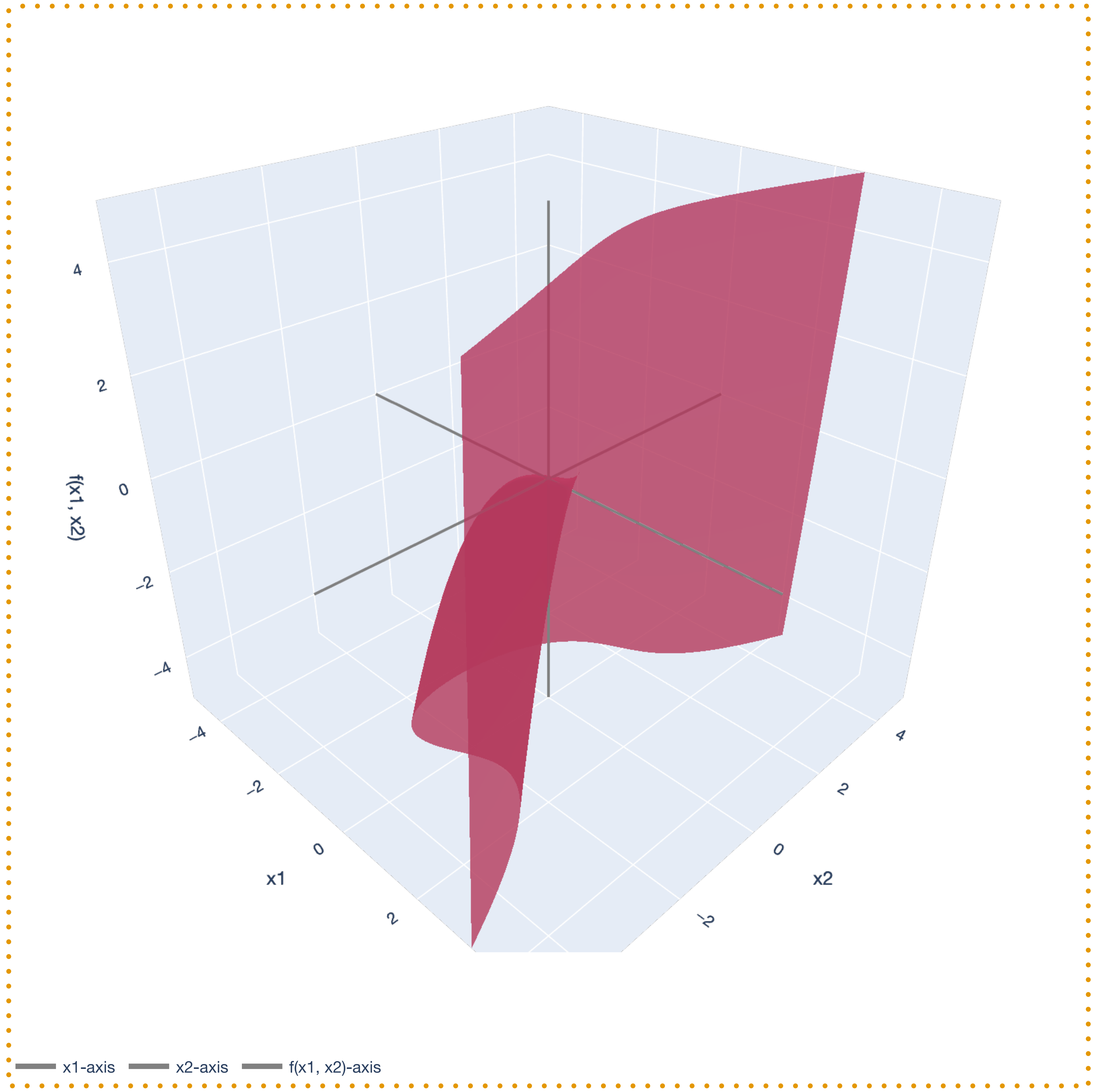
Polynomials

Multivariable definition

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



$$f(x_1, x_2) = x_1^3 + x_1x_2 - x_2^2$$



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*!

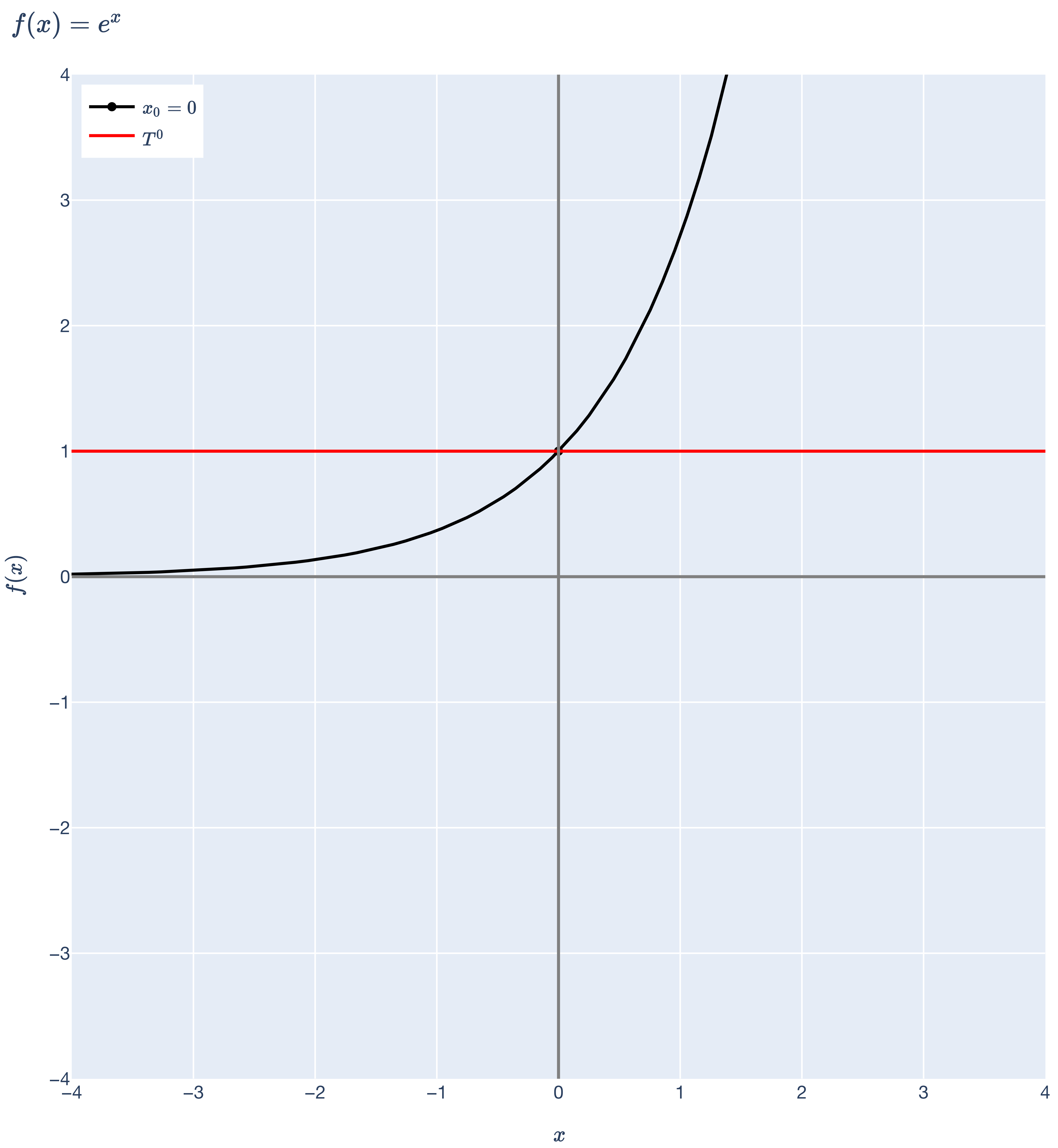
The second-order Taylor approximation is just a quadratic function!

Taylor Series

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

Taylor series at $x_0 = 0$:

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

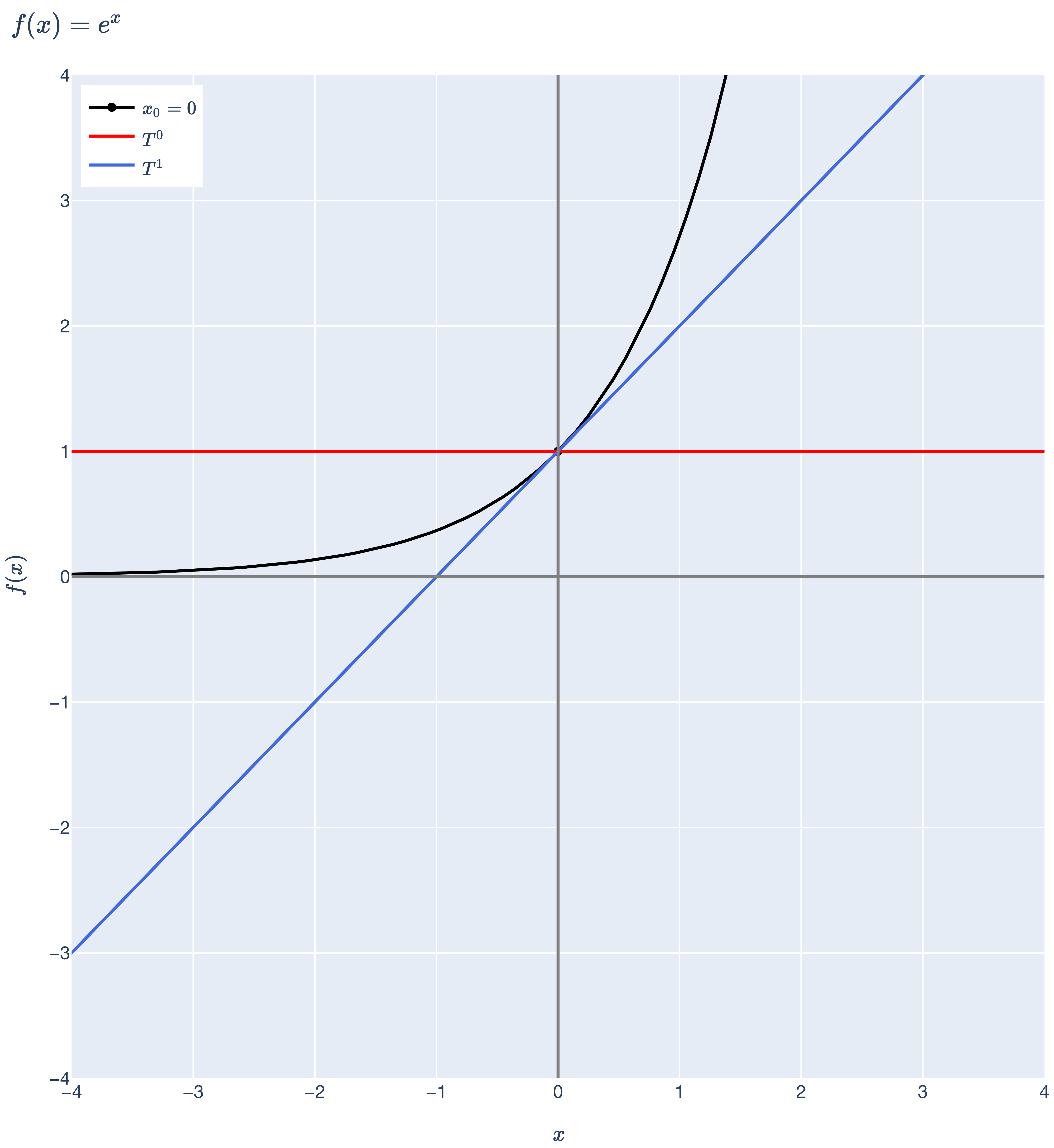


Taylor Series

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

Taylor series at $x_0 = 0$:

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

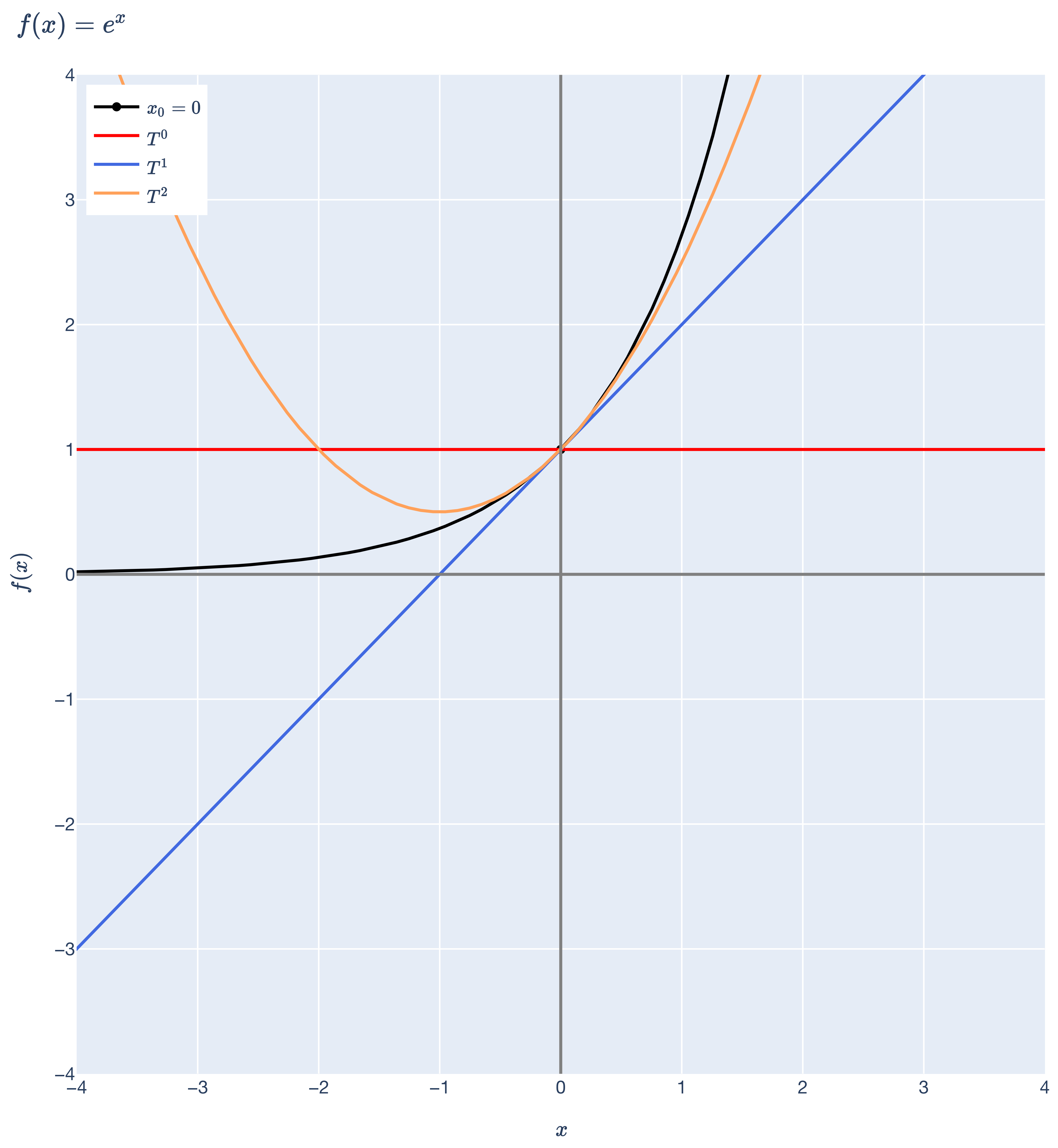


Taylor Series

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

Taylor series at $x_0 = 0$:

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

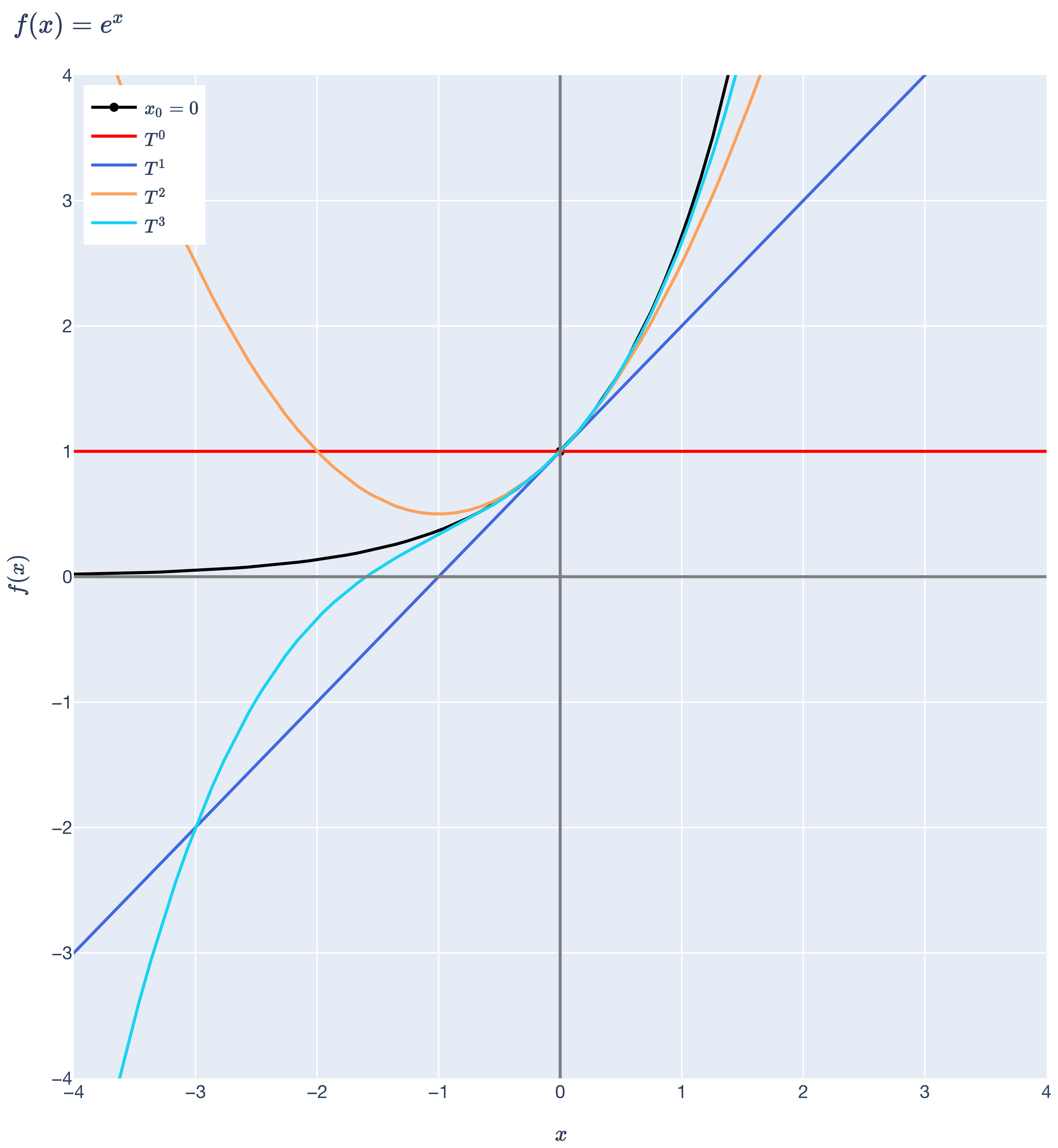


Taylor Series

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

Taylor series at $x_0 = 0$:

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

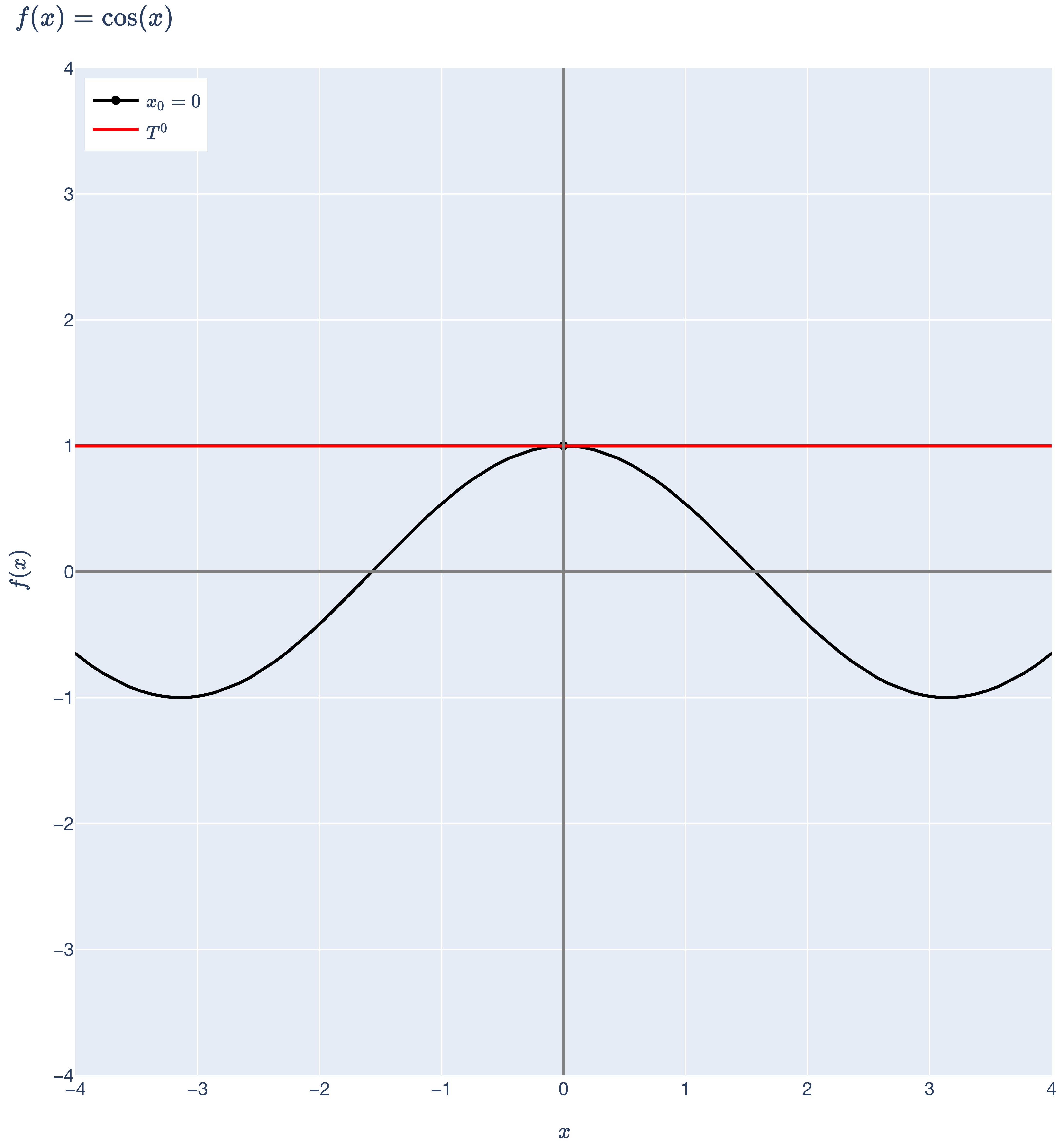


Taylor Series

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

Taylor series at $x_0 = 0$:

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

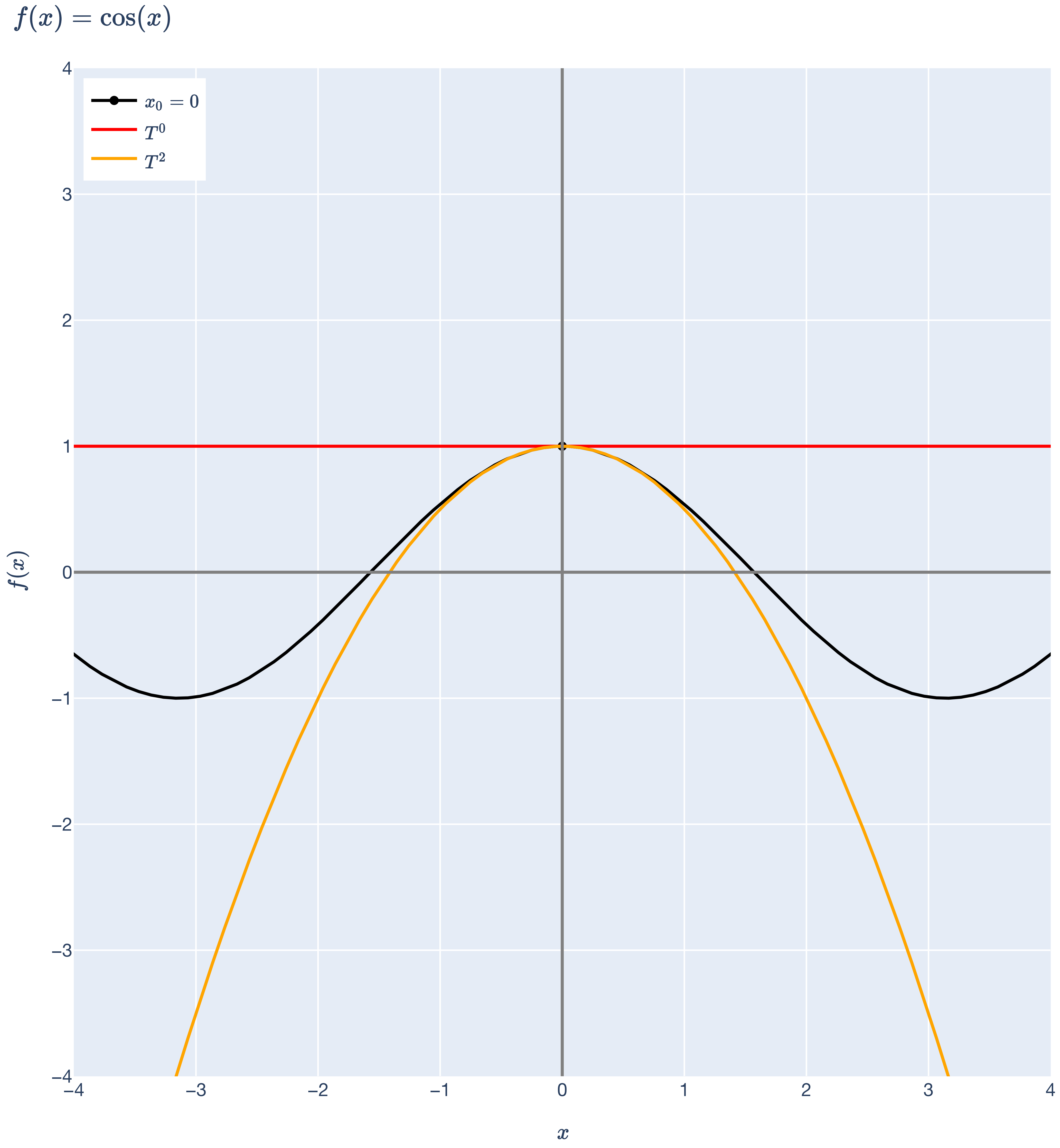


Taylor Series

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

Taylor series at $x_0 = 0$:

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

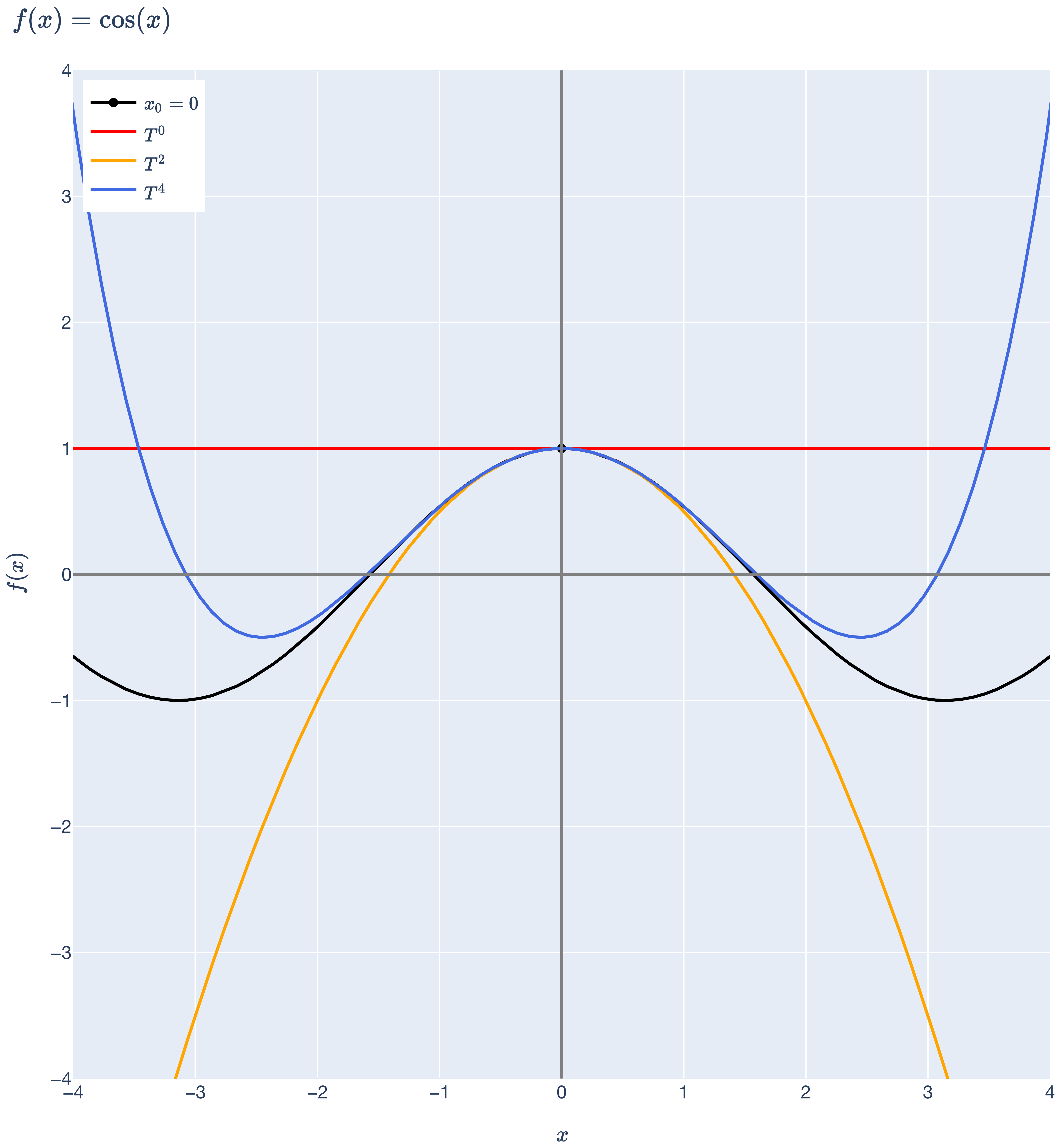


Taylor Series

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

Single-variable definition ($f : \mathbb{R} \to \mathbb{R}$)

For a 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 a 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 $n \leq 2$).

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 ($f : \mathbb{R}^d \to \mathbb{R}$)

Let $f \in \mathscr{C}^\infty$. The Taylor series of f at $\mathbf{x}_0 = (x_{01}, \ldots, x_{0d}) \in \mathbb{R}^d$ is given by:

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

Thankfully, we won't ever need to use this in full generality. 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 $\mathbf{x}_0$ is the 2×2 matrix of all second-order partial derivatives:

$$\nabla^2 f(\mathbf{x}_0) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(\mathbf{x}_0) & \frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x}_0) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x}_0) & \frac{\partial^2}{\partial x_2^2} f(\mathbf{x}_0) \end{bmatrix}$$

The Hessian for general $f : \mathbb{R}^d \to \mathbb{R}$ is given by the $d \times d$ matrix constructed similarly.

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

Taylor Series

Just the second-order terms

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

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

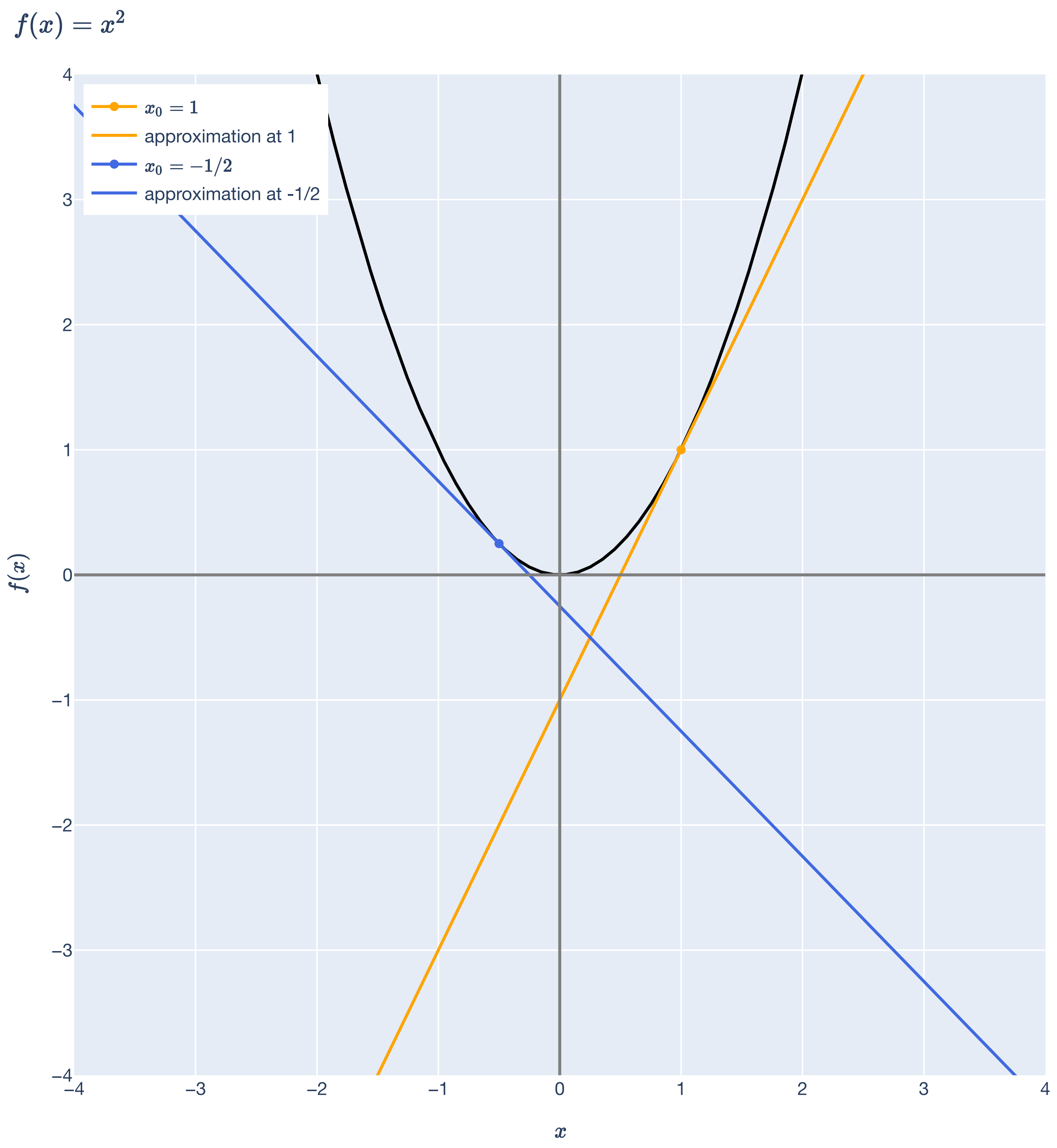
linear function!

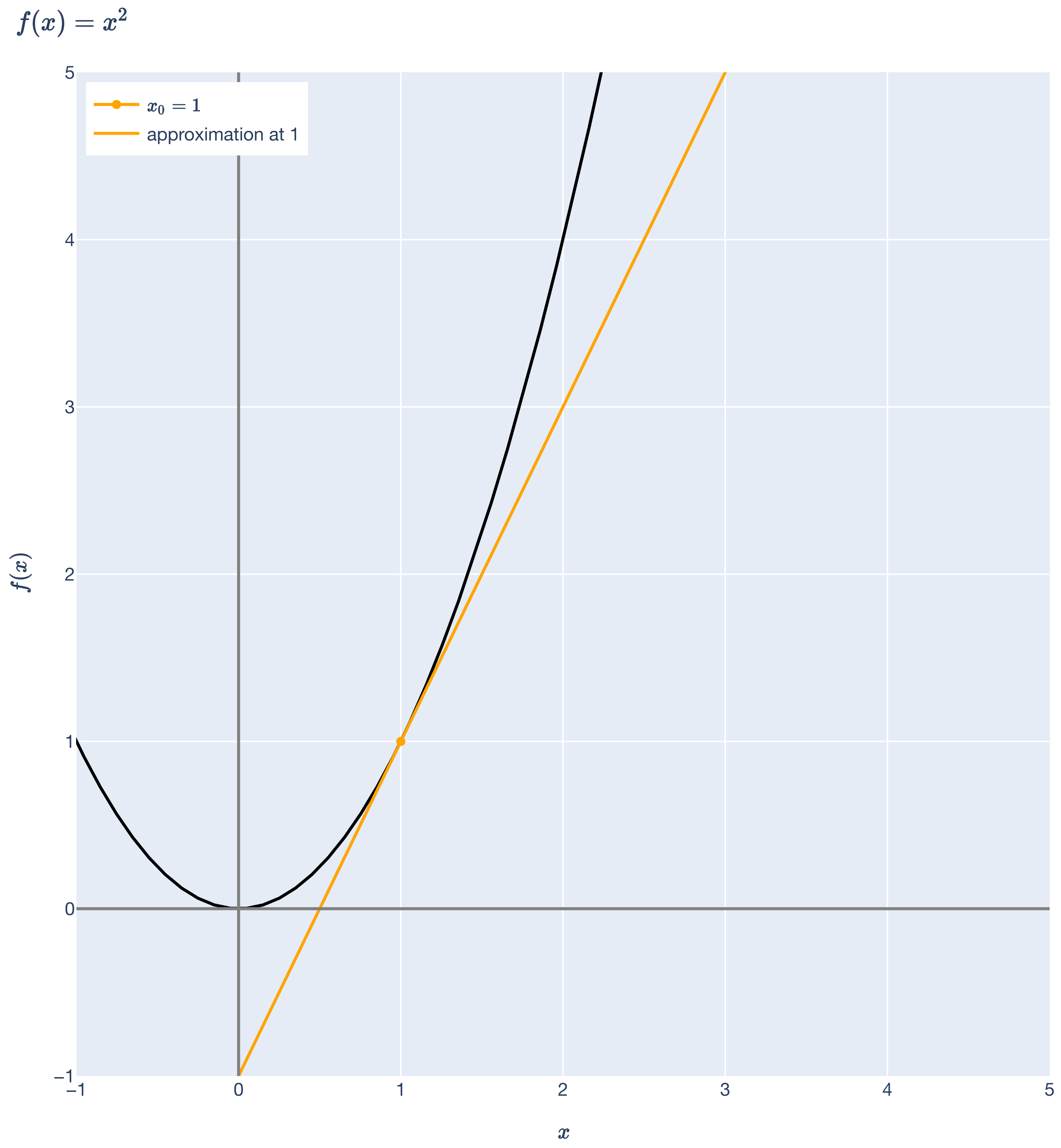
quadratic form!

Linear Approximations

Our main slogan

At any point $\mathbf{x}_0 \in \mathbb{R}^d, f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)$ *for all* $\mathbf{x}$ *close to* $\mathbf{x}_0$





First-order Taylor Approximation

Just linear approximation

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

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

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

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

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

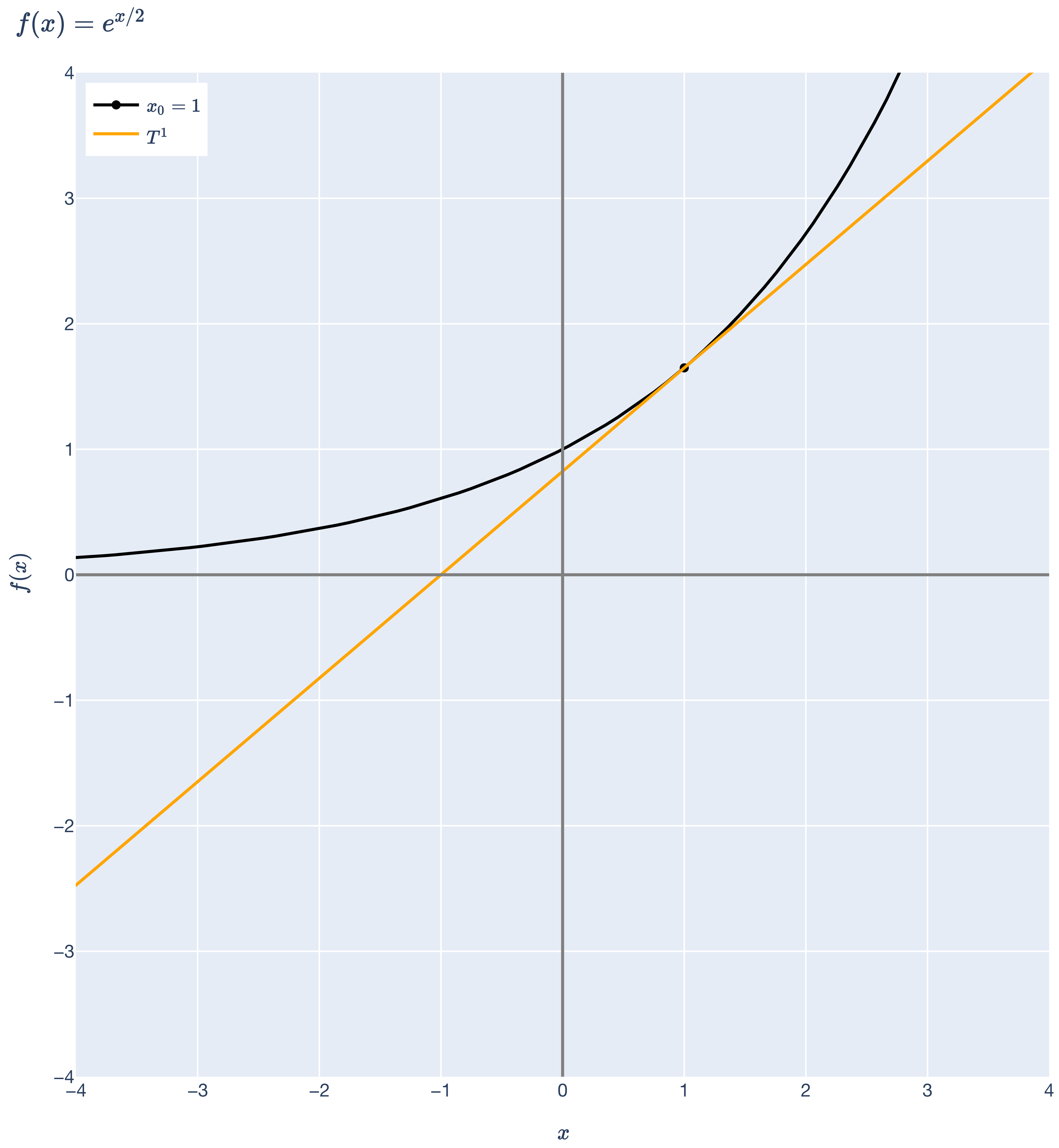
First-order Taylor Approximation

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

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

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

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

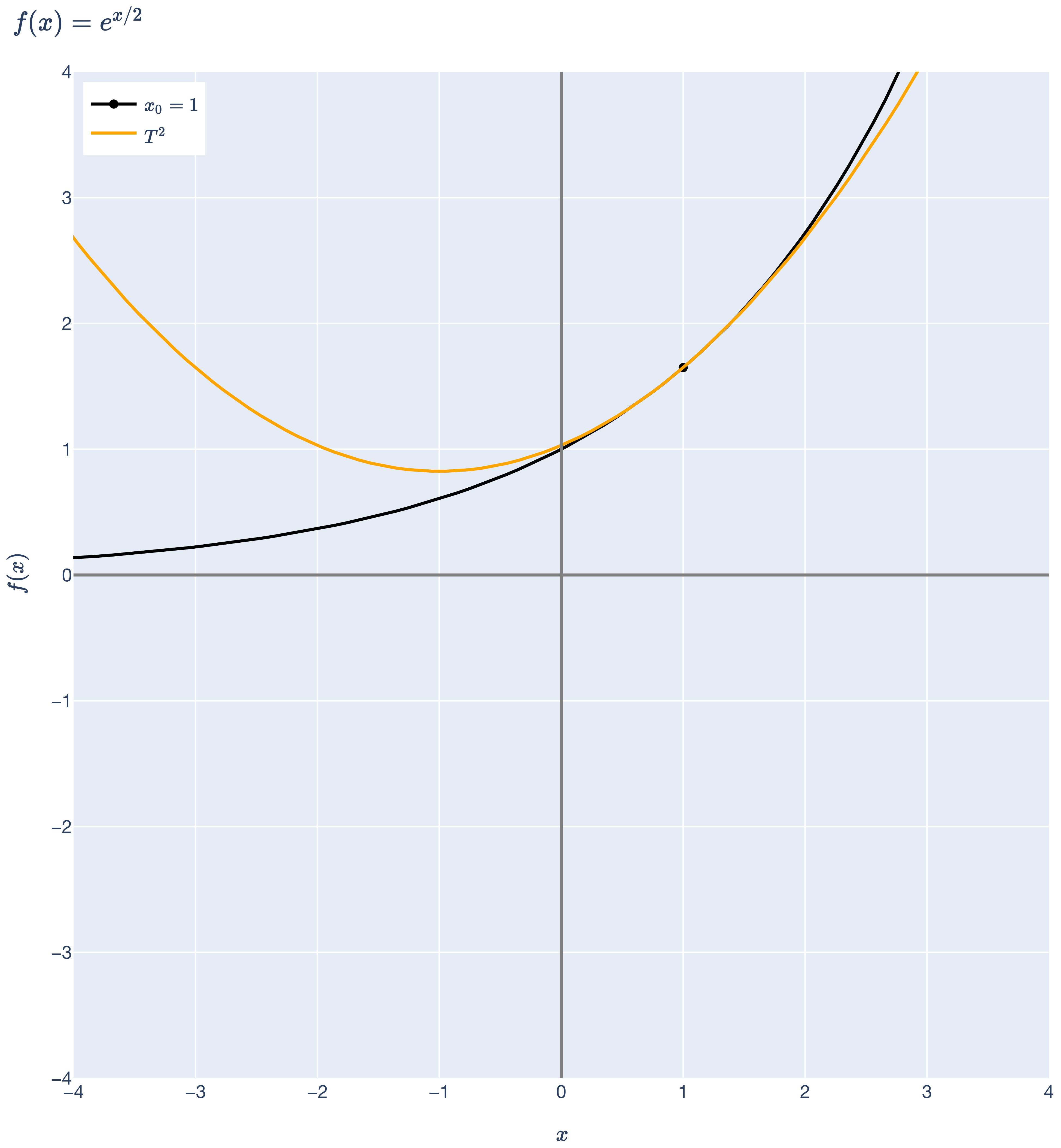
Second-order Taylor Approximation

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 (linear approximation)* of a function at $\mathbf{x}_0$ is:

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

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?

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?

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

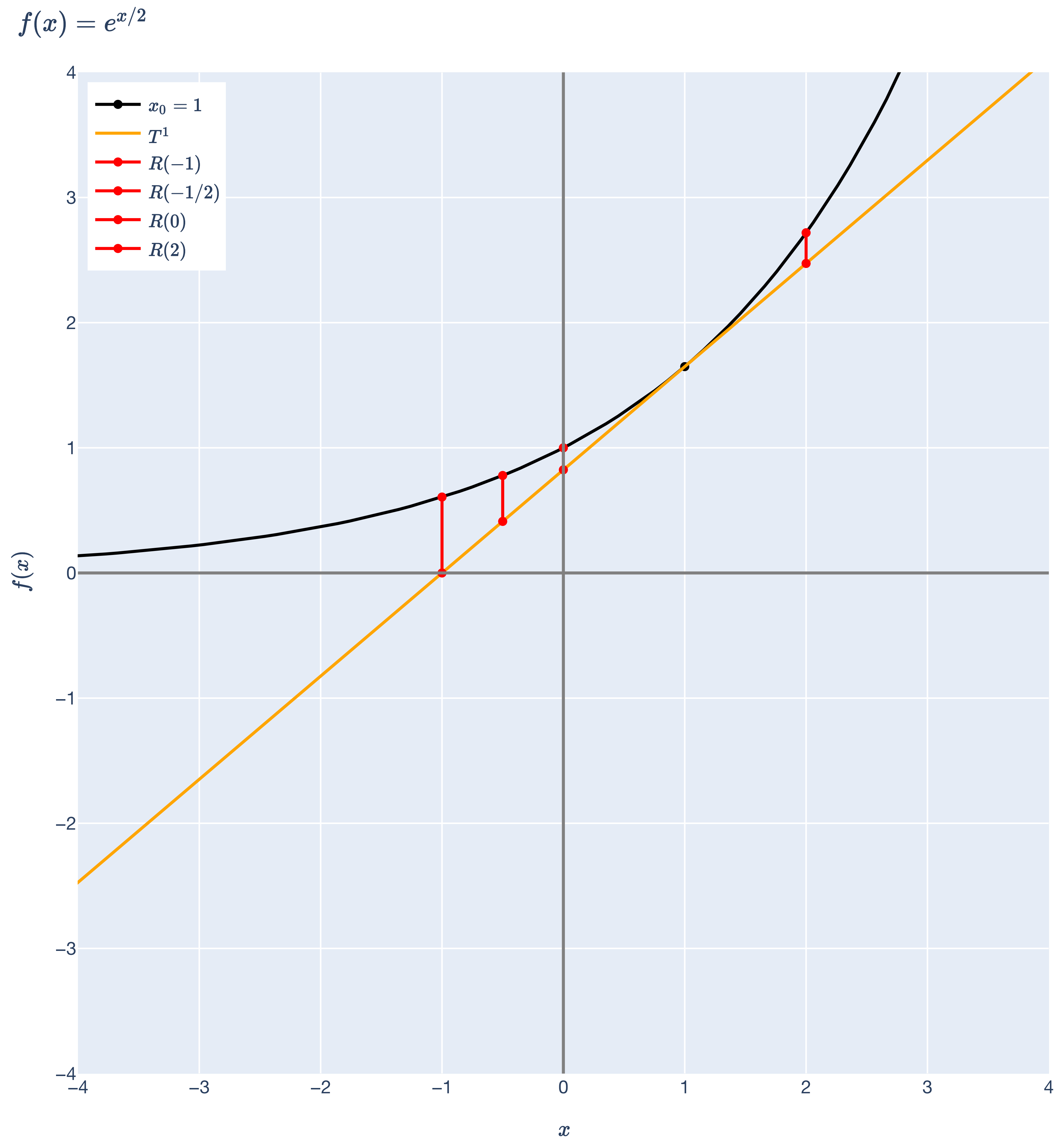
Definition

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

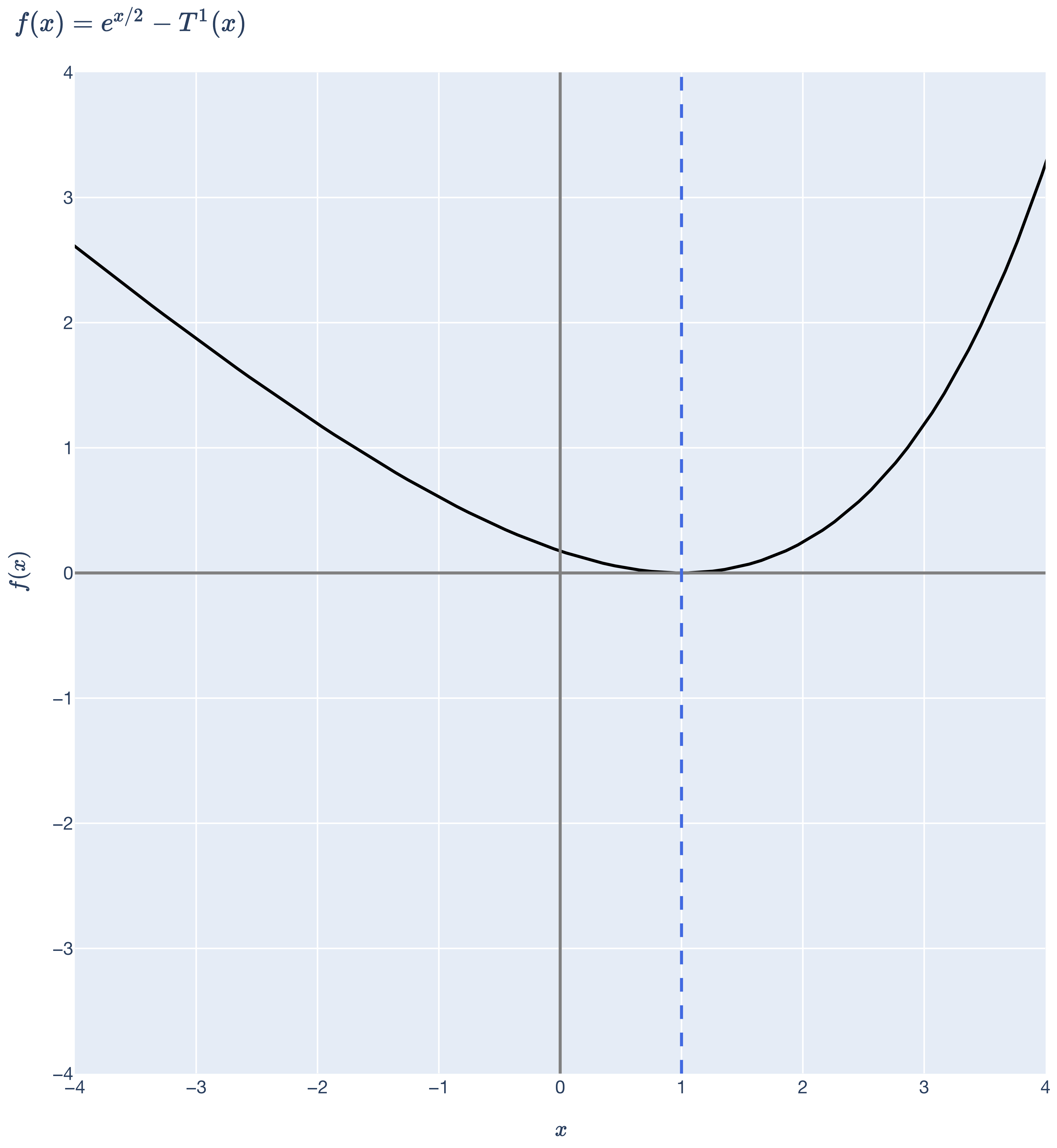
Definition

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

Single variable theorem

Theorem (Taylor’s Theorem, single variable). 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

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

Or, in terms of the remainder:

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

Taylor's Theorem

Multivariable (and first order) theorem

Theorem (Taylor's Theorem, multivariable). 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

Formalizing the descent lemma

Gradient Descent Guarantees

Theorem 1: Descent Lemma

Theorem (Descent Lemma). If f is "smooth enough," then there is a choice of $\eta > 0$ such that, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{\eta}{2} \|\nabla f(\mathbf{w})\|^2.$$

"Smooth enough" : f is a β-smooth function.

Taylor's Theorem: makes the $\lessapprox$ rigorous!

Descent Lemma

Conclusion

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

If η is small enough, then $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$ is close to $\mathbf{w}^{(t-1)}$, and:

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$

Therefore,

$$f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^{(t-1)}) \text{ as long as } \eta \text{ is sufficiently small!}$$

Taylor's Theorem

Multivariable (and first order) theorem

Theorem (Taylor's Theorem, multivariable). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathscr{C}^2$ function. For $\mathbf{x}_0, \mathbf{d} \in \mathbb{R}^d$, 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:

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

Descent Lemma

Applying Taylor's Theorem

Taylor's Theorem

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

For $\mathbf{w}^{(t-1)}$ and $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{w}} = \mathbf{w}^{(t-1)} - \lambda \eta \nabla f(\mathbf{w}^{(t-1)})$ on the line segment between $\mathbf{w}^{(t-1)}$ and $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$,

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) = f(\mathbf{w}^{(t-1)}) - \eta \nabla f(\mathbf{w}^{(t-1)})^\top \nabla f(\mathbf{w}^{(t-1)}) + \frac{1}{2}(-\eta \nabla f(\mathbf{w}^{(t-1)}))^\top \nabla^2 f(\tilde{\mathbf{w}})(-\eta \nabla f(\mathbf{w}^{(t-1)}))$$

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) = f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2}{2} \nabla f(\mathbf{w}^{(t-1)})^\top \nabla^2 f(\tilde{\mathbf{w}}) \nabla f(\mathbf{w}^{(t-1)})$$

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

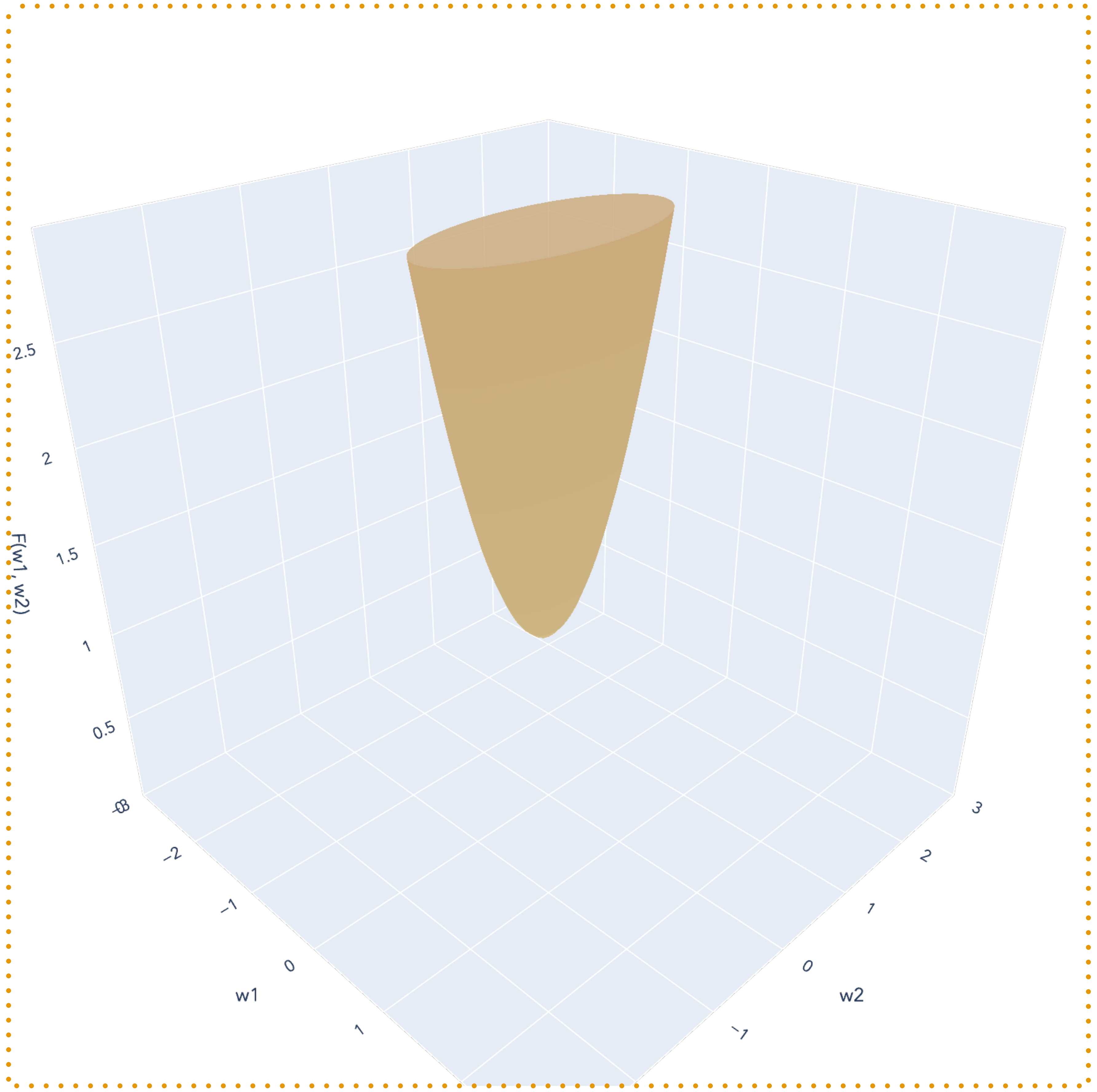
Prop (Smoothness & Quad. Forms). If $\mathbf{A} \in \mathbb{R}^{d \times d}$ is β-smooth, then for any unit vector $\mathbf{v} \in \mathbb{R}^d$,

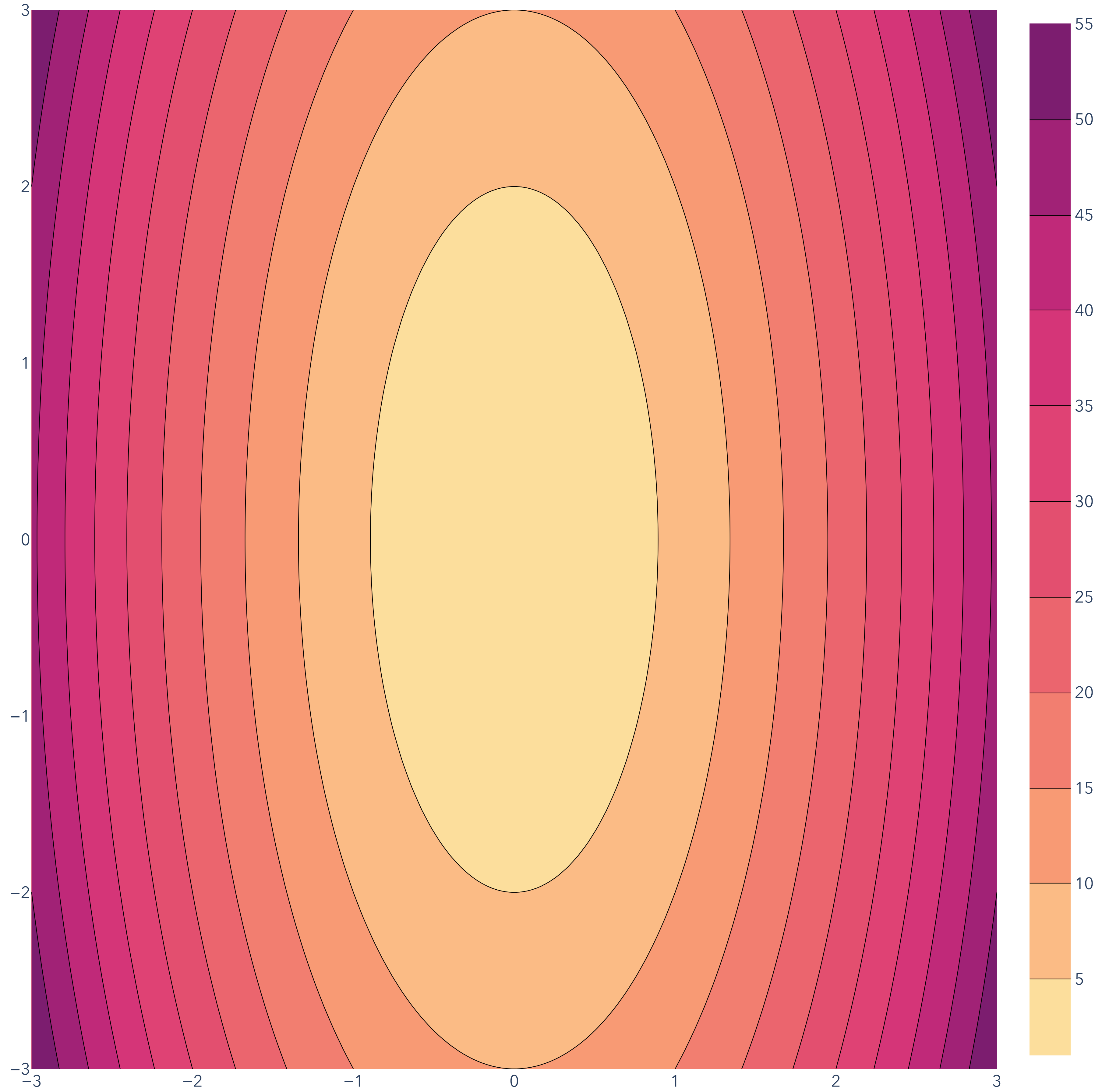
$$\mathbf{v}^\top \mathbf{A} \mathbf{v} \quad \leq \beta.$$

Bounding change in gradients

β-smoothness

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

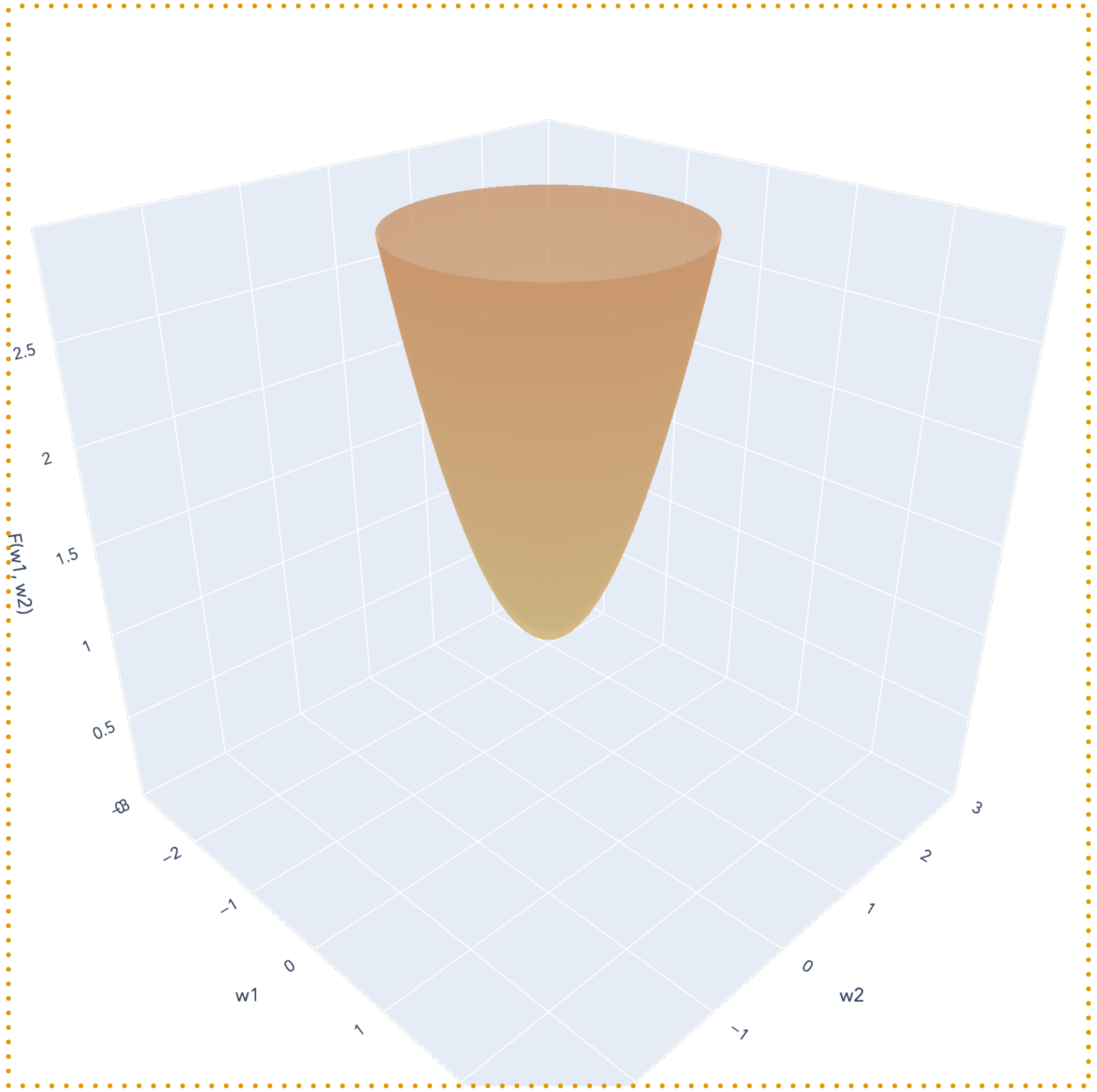


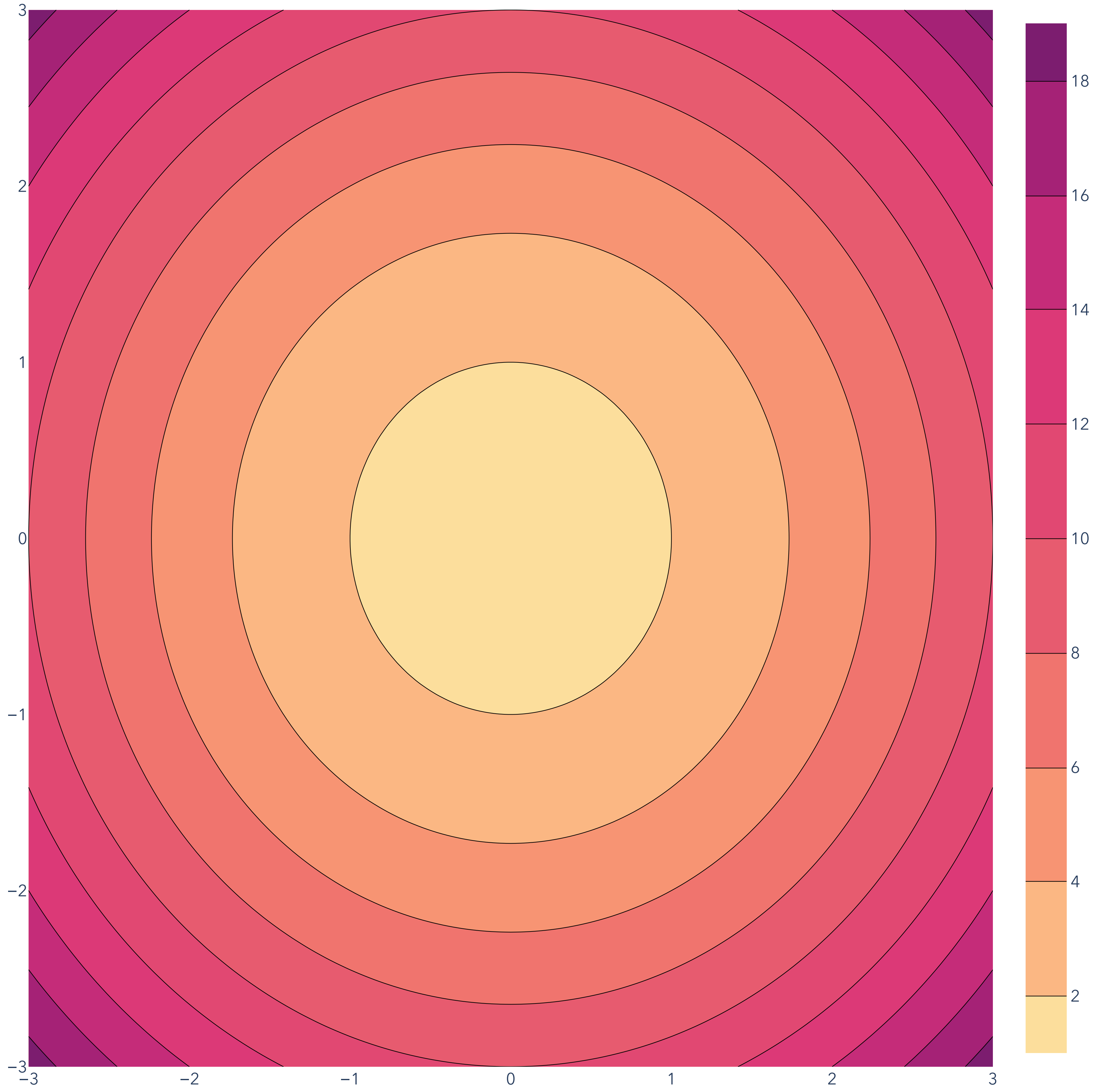


Bounding change in gradients

β-smoothness

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





Descent Lemma

Applying Taylor's Theorem

Taylor's Theorem

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

For $\mathbf{w}^{(t-1)}$ and $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{w}} = \mathbf{w}^{(t-1)} - \lambda \eta \nabla f(\mathbf{w}^{(t-1)})$ on the line segment between $\mathbf{w}^{(t-1)}$ and $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$,

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) = f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2}{2} \nabla f(\mathbf{w}^{(t-1)})^\top \nabla^2 f(\tilde{\mathbf{w}}) \nabla f(\mathbf{w}^{(t-1)})$$

$$= f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2 \|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2} (\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)^\top \nabla^2 f(\tilde{\mathbf{w}}) (\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)$$

Scale to unit vectors to apply smoothness property!

Descent Lemma

Applying Taylor's Theorem

Taylor's Theorem

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

<u>Goal:</u> move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

For $\mathbf{w}^{(t-1)}$ and $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{w}} = \mathbf{w}^{(t-1)} - \lambda \eta \nabla f(\mathbf{w}^{(t-1)})$ on the line segment between $\mathbf{w}^{(t-1)}$ and $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$,

$$f(\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})) = f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2}{2} \nabla f(\mathbf{w}^{(t-1)})^\top \nabla^2 f(\tilde{\mathbf{w}}) \nabla f(\mathbf{w}^{(t-1)})$$

$$= f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2 \|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2} (\nabla f(\mathbf{w}^{(t-1)}) / \|\nabla f\|)^\top \nabla^2 f(\tilde{\mathbf{w}}) (\nabla f(\mathbf{w}^{(t-1)}) / \|\nabla f\|)$$

Apply β smoothness to the quadratic form!

Descent Lemma

Applying Taylor's Theorem

Taylor's Theorem

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

For $\mathbf{w}^{(t-1)}$ and $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{w}} = \mathbf{w}^{(t-1)} - \lambda \eta \nabla f(\mathbf{w}^{(t-1)})$ on the line segment between $\mathbf{w}^{(t-1)}$ and $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$,

Apply β smoothness to the quadratic form!

$$= f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2 \|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2} (\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)^\top \nabla^2 f(\tilde{\mathbf{w}})(\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)$$

$$\leq f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2 \|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2} \beta$$

Descent Lemma

Applying Taylor's Theorem

Taylor's Theorem

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

$$f(\mathbf{w}) \approx f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{w} - \mathbf{u}) \text{ for } \mathbf{w} \text{ close to } \mathbf{u}$$

Goal: move in a direction $\mathbf{d} \in \mathbb{R}^d$ such that $f(\mathbf{w}^{(t-1)} + \mathbf{d}) < f(\mathbf{w}^{(t-1)})$.

For $\mathbf{w}^{(t-1)}$ and $\mathbf{d} = -\eta \nabla f(\mathbf{w}^{(t-1)})$, there exists $\lambda \in (0,1)$ such that for $\tilde{\mathbf{w}} = \mathbf{w}^{(t-1)} - \lambda \eta \nabla f(\mathbf{w}^{(t-1)})$ on the line segment between $\mathbf{w}^{(t-1)}$ and $\mathbf{w}^{(t-1)} - \eta \nabla f(\mathbf{w}^{(t-1)})$,

Apply β smoothness to the quadratic form!

$$= f(\mathbf{w}^{(t-1)}) - \eta\|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2\|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2}(\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)^\top \nabla^2 f(\tilde{\mathbf{w}})(\nabla f(\mathbf{w}^{(t-1)})/\|\nabla f\|)$$

$$\leq f(\mathbf{w}^{(t-1)}) - \eta\|\nabla f(\mathbf{w}^{(t-1)})\|^2 + \frac{\eta^2\|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2}\beta \leq f(\mathbf{w}^{(t-1)}) - \frac{\|\nabla f(\mathbf{w}^{(t-1)})\|^2}{2\beta}$$

Letting $\eta = 1/\beta$, we get the best possible bound.

Gradient Descent Guarantees

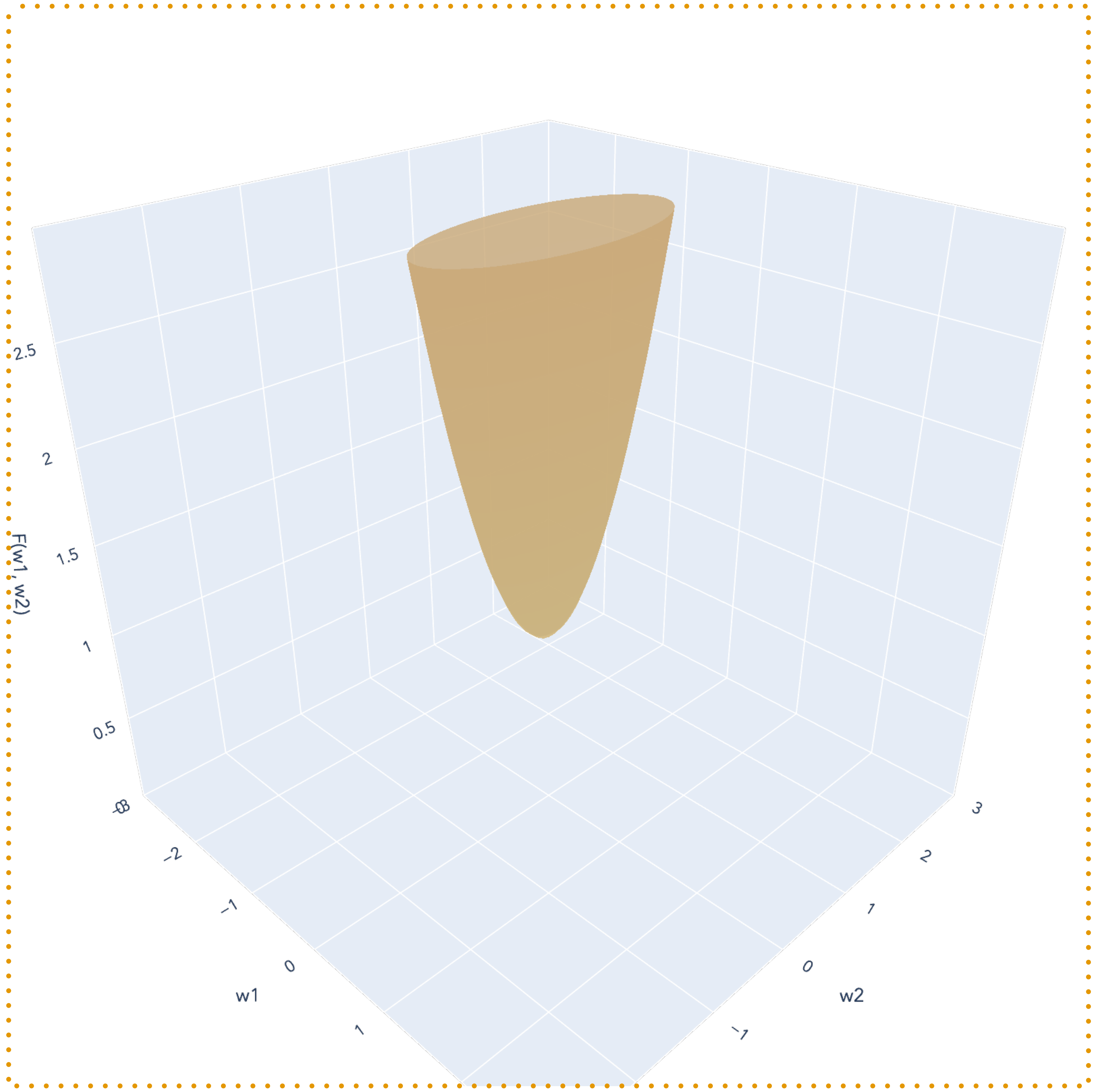
Theorem 1: Descent Lemma

Theorem (Descent Lemma). If f is "smooth enough," then there is a choice of $\eta > 0$ such that, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{\eta}{2} \|\nabla f(\mathbf{w})\|^2.$$

"Smooth enough" : f is a β-smooth function.

Taylor's Theorem: makes the $\lessapprox$ rigorous!



Gradient Descent Guarantees

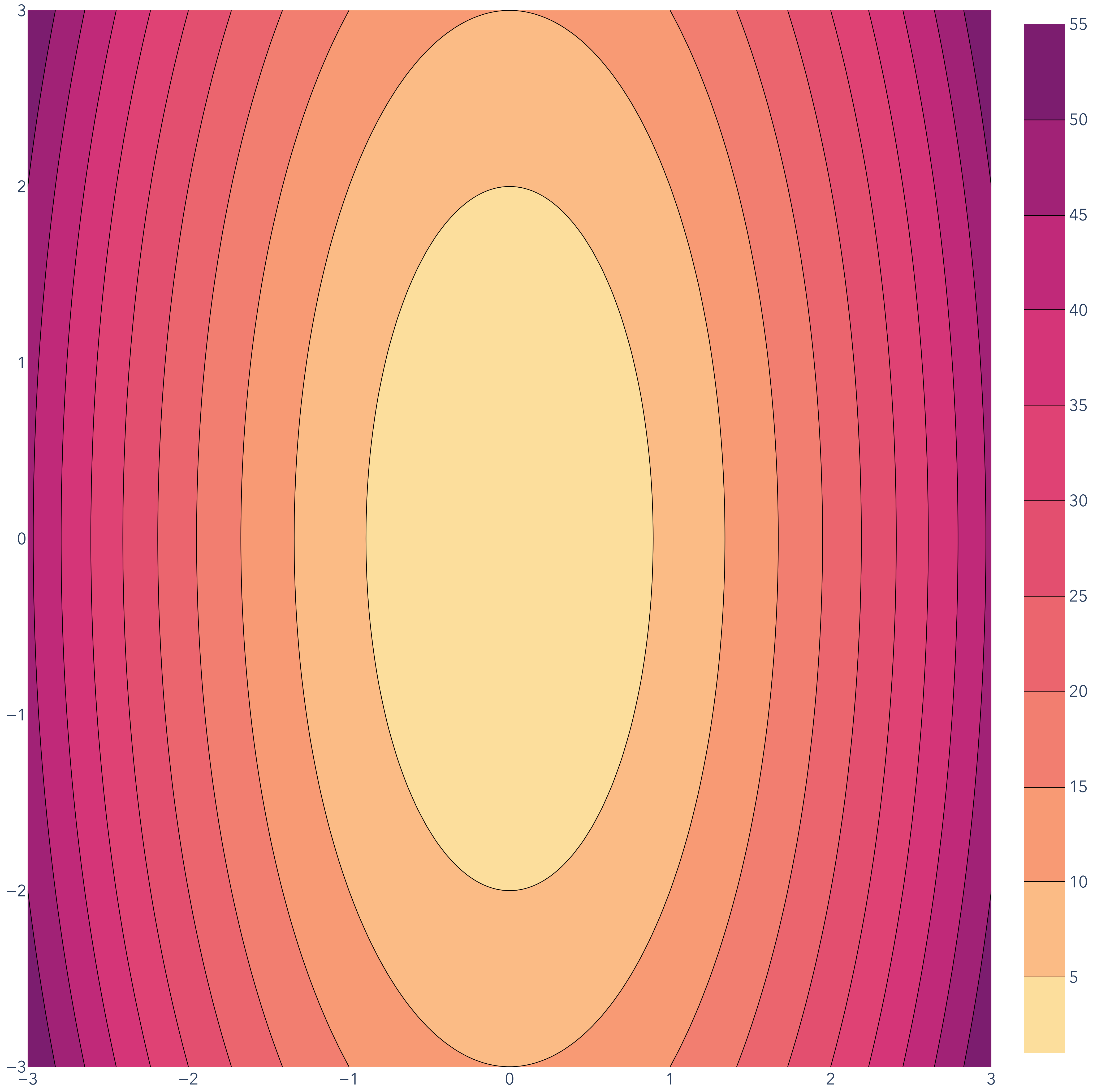
Theorem 1: Descent Lemma

Theorem (Descent Lemma). If f is "smooth enough," then there is a choice of $\eta > 0$ such that, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{\eta}{2}\|\nabla f(\mathbf{w})\|^2.$$

"Smooth enough" : f is a β-smooth function.

Taylor's Theorem: makes the $\lesssim$ rigorous!



Gradient Descent Guarantees

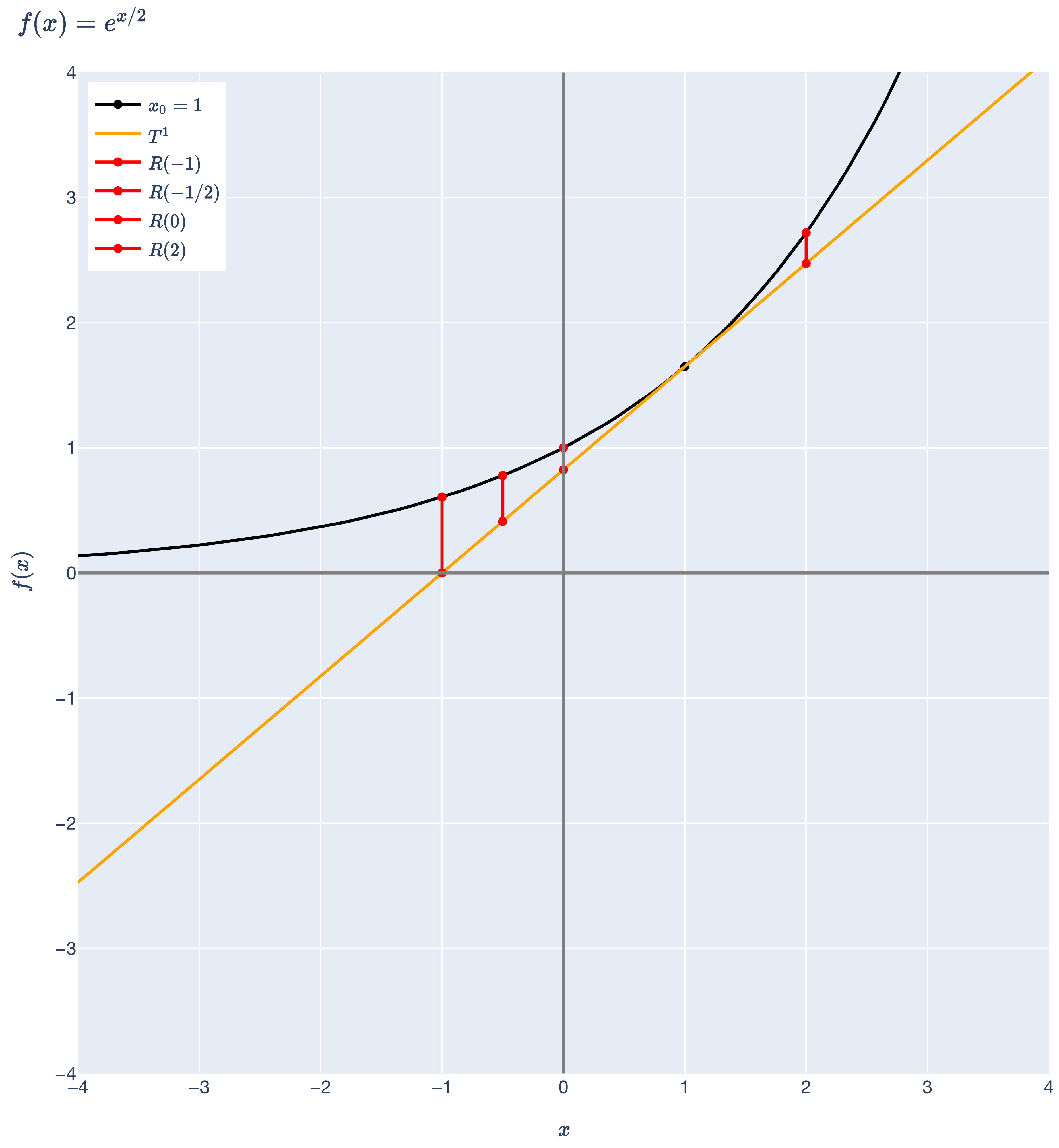
Theorem 1: Descent Lemma

Theorem (Descent Lemma). If f is "smooth enough," then there is a choice of $\eta > 0$ such that, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{\eta}{2} \|\nabla f(\mathbf{w})\|^2.$$

"Smooth enough" : f is a β-smooth function.

Taylor's Theorem: makes the $\lessapprox$ rigorous!

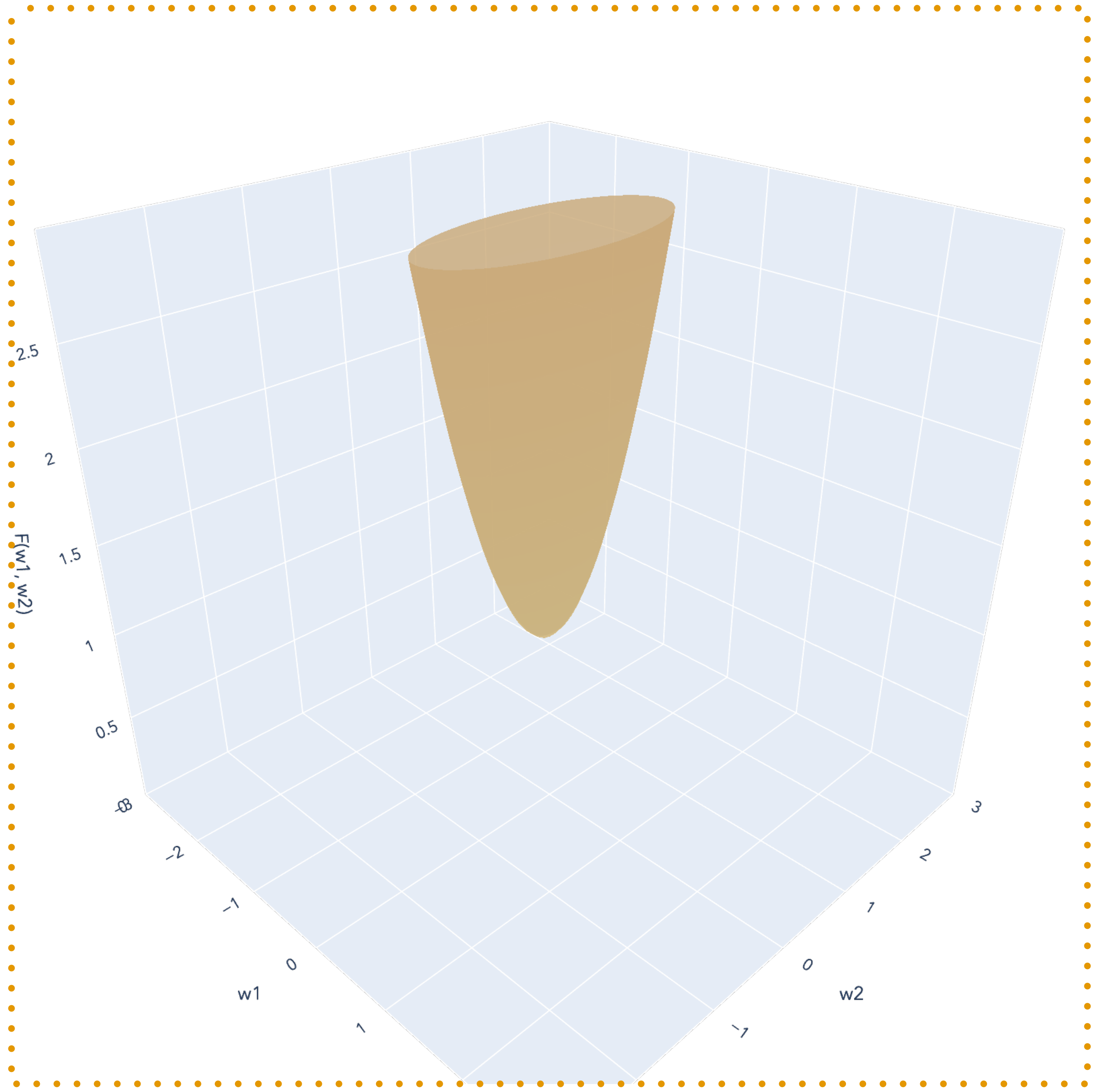


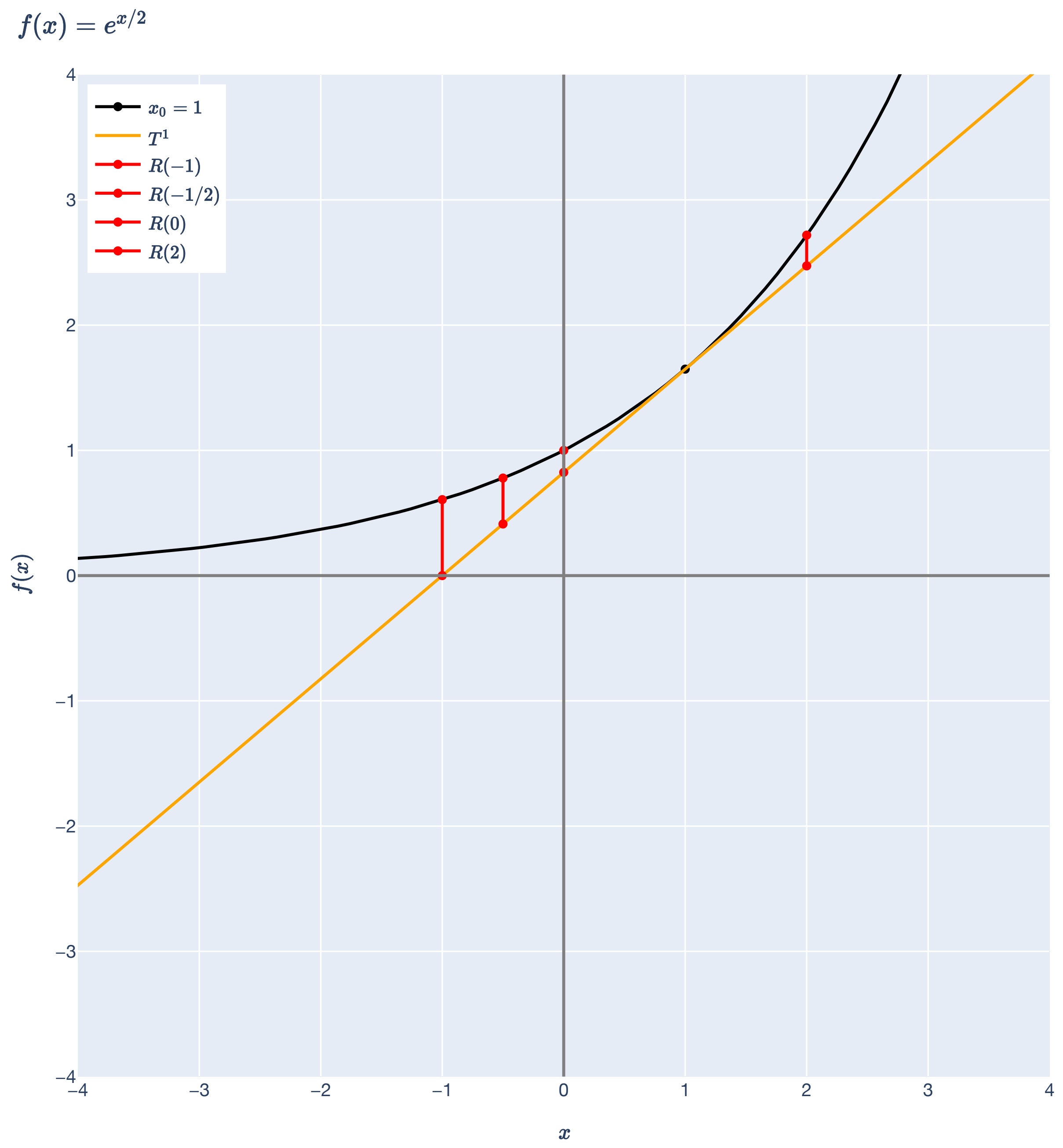
Gradient Descent

Theorem 1: Descent Lemma (Formal)

Theorem (Descent Lemma). If $f \in \mathscr{C}^2$ and is β-smooth, then with $\eta = 1/\beta$, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{1}{2\beta} \|\nabla f(\mathbf{w})\|^2.$$





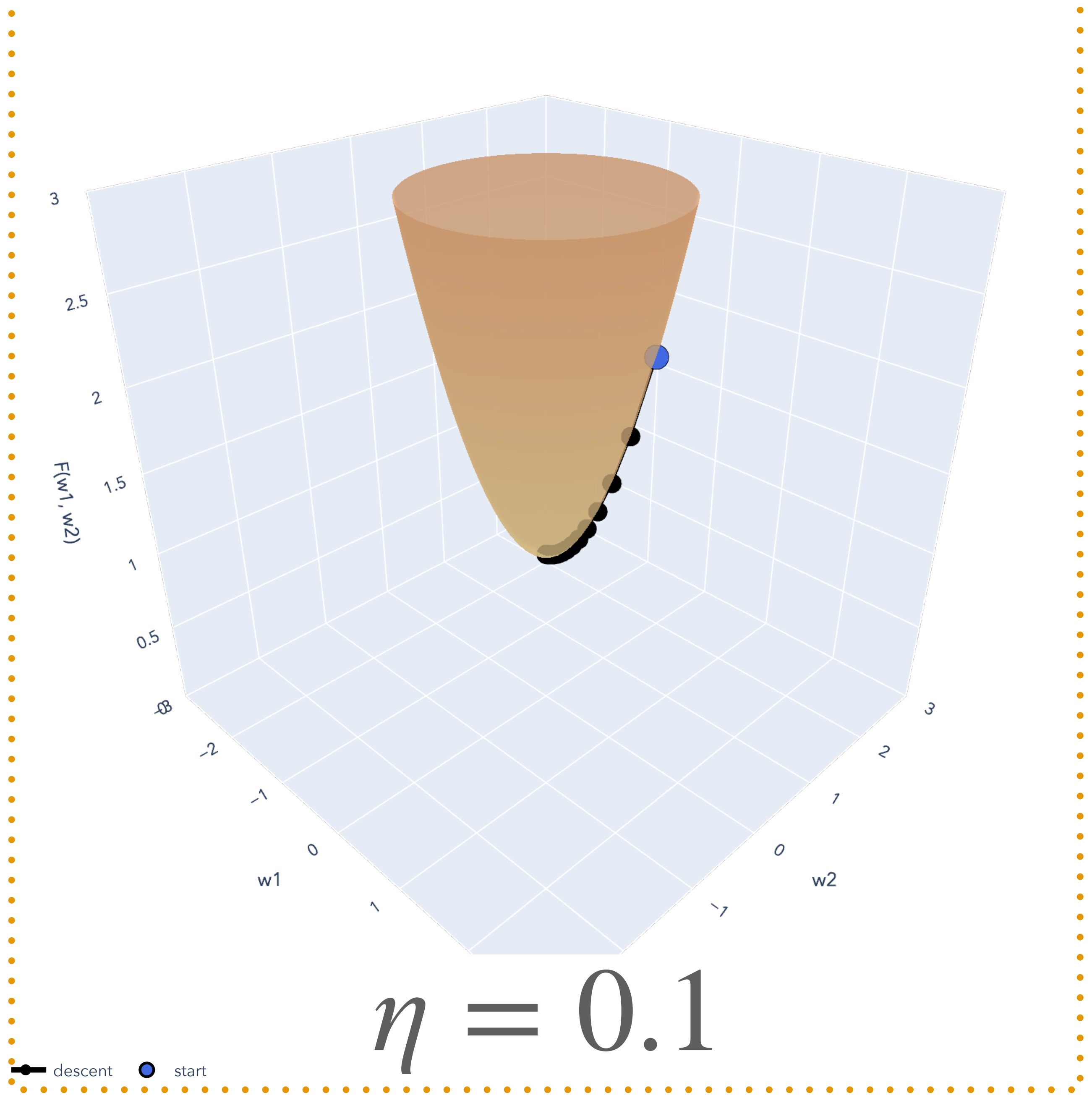
Gradient Descent

Theorem 1: Descent Lemma (Formal)

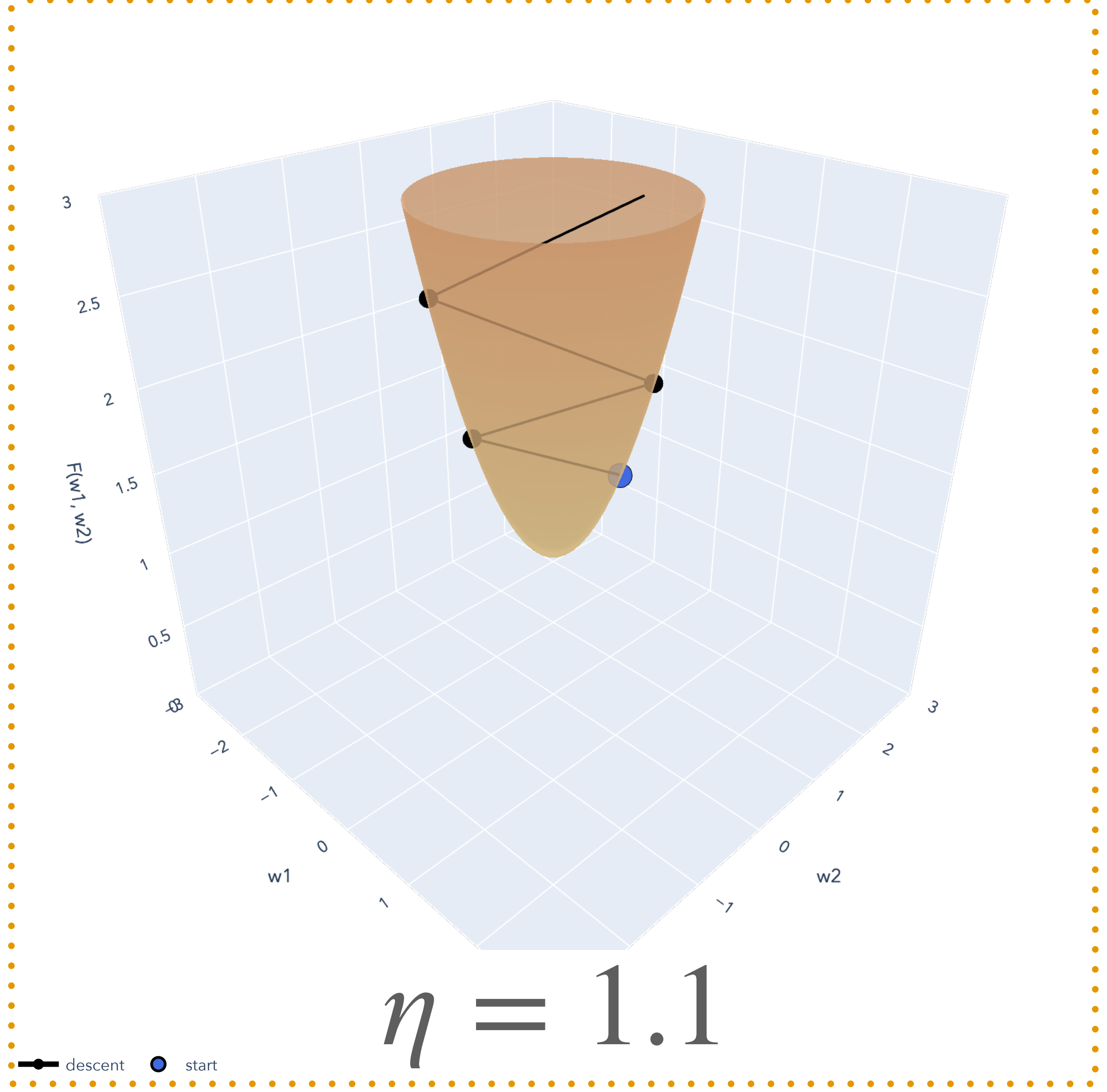
Theorem (Descent Lemma). If $f \in \mathscr{C}^2$ and is β-smooth, then with $\eta = 1/\beta$, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2.$$

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



$\eta = 0.1$



$\eta = 1.1$

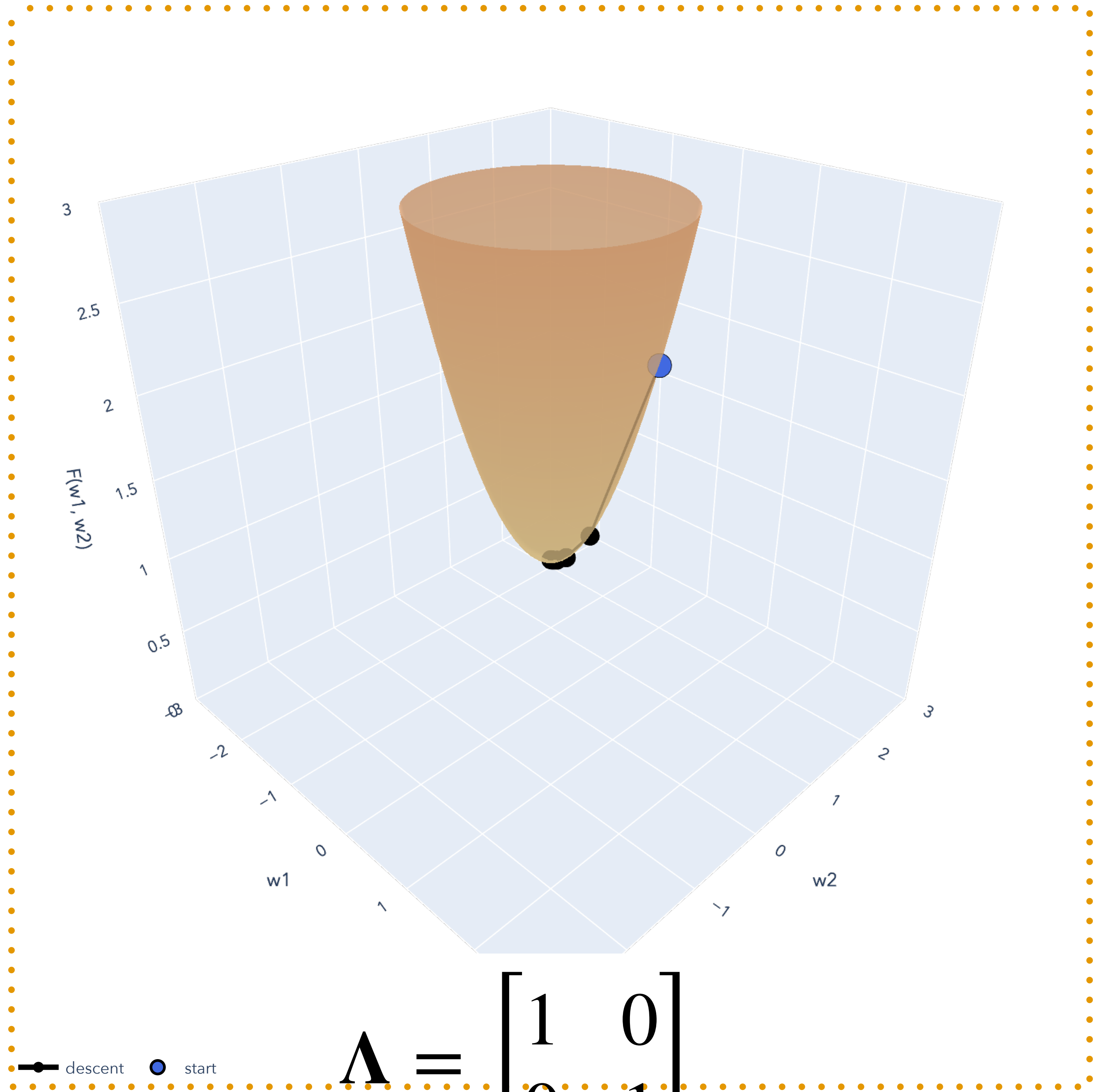
Gradient Descent

Theorem 1: Descent Lemma (Formal)

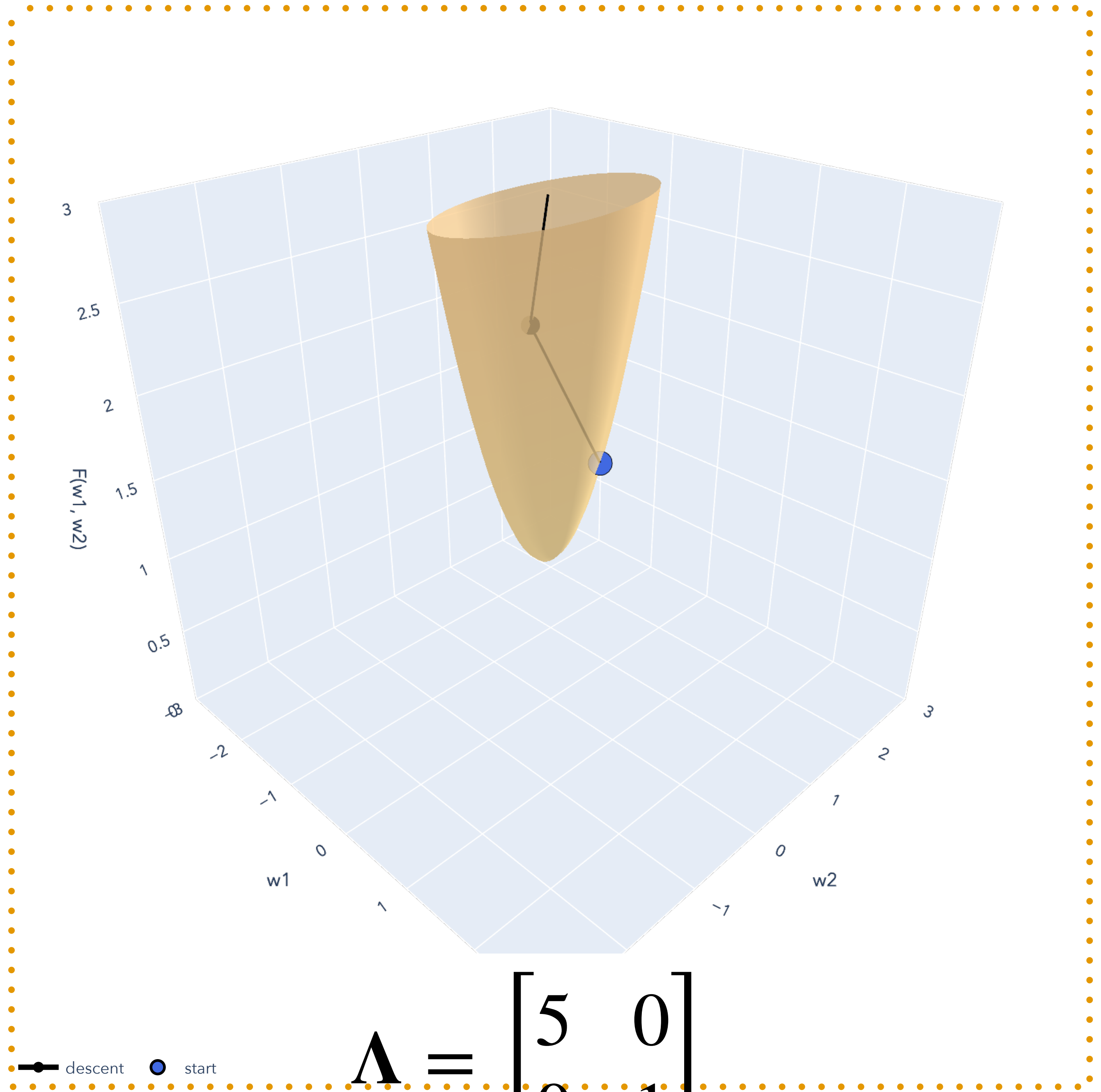
Theorem (Descent Lemma). If $f \in \mathscr{C}^2$ and is β-smooth, then with $\eta = 1/\beta$, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2.$$

$\eta = 0.3$



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



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

Gradient Descent

Theorem 1: Descent Lemma (Formal)

Theorem (Descent Lemma). If $f \in \mathscr{C}^2$ and is β-smooth, then with $\eta = 1/\beta$, for any $\mathbf{w} \in \mathbb{R}^d$,

$$f(\mathbf{w} - \eta \nabla f(\mathbf{w})) \leq f(\mathbf{w}) - \frac{1}{2\beta} \|\nabla f(\mathbf{w})\|^2.$$

Gradient Descent

Preview of convexity

Descent Lemma

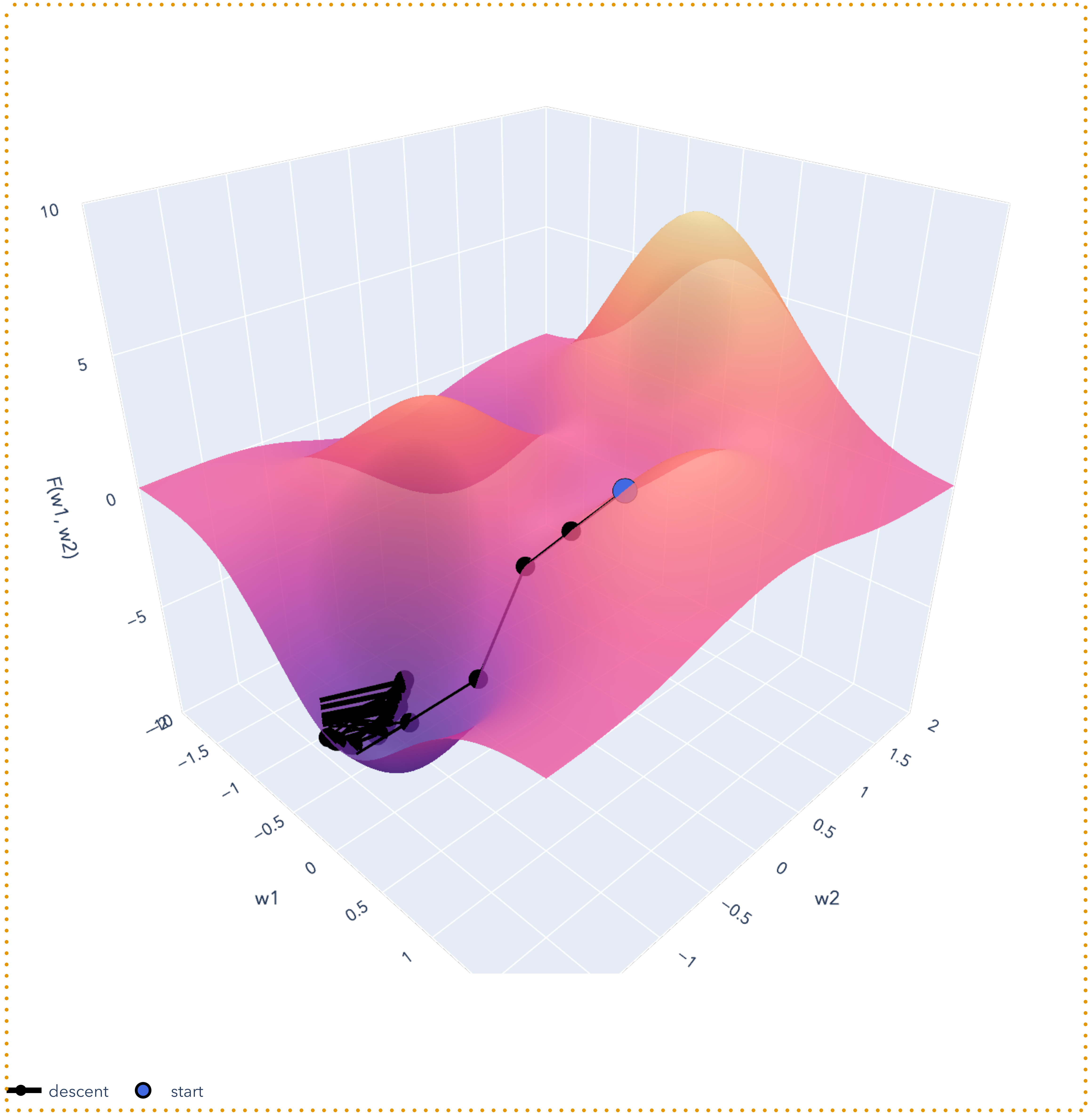
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Descent Lemma

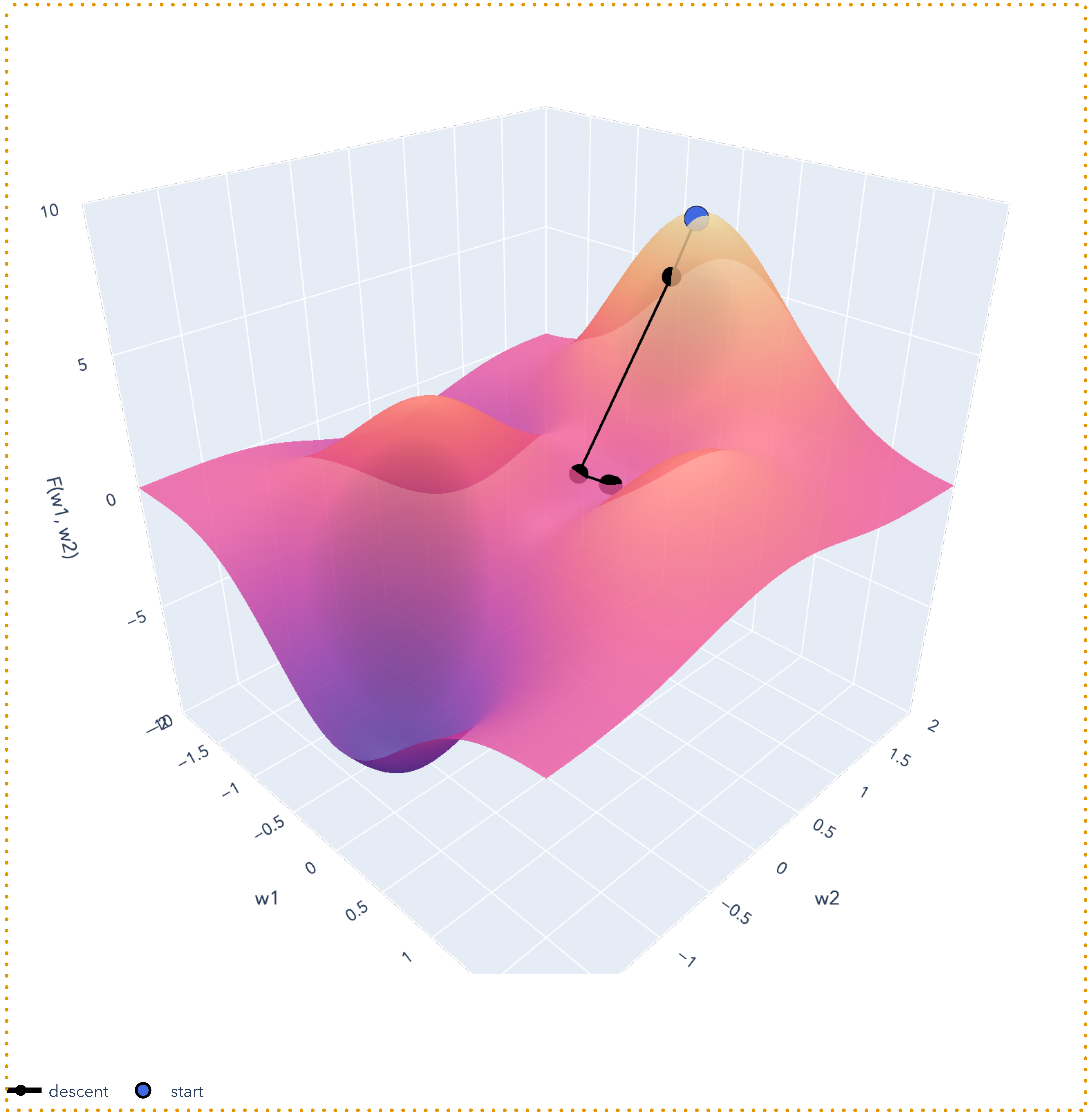
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Descent Lemma

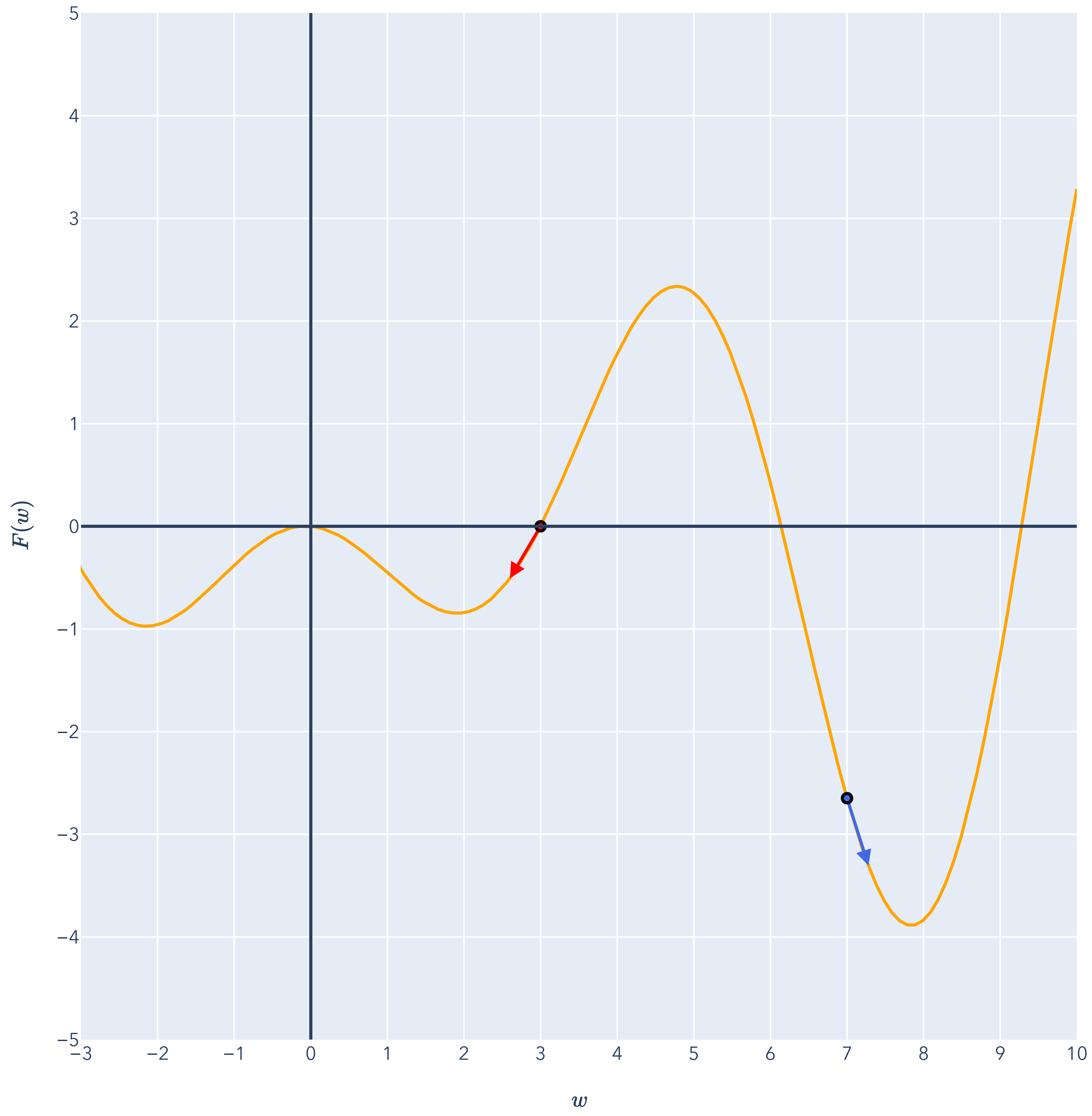
Guarantee (Informal)

If η is small enough, then the gradient descent update rule

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

has the property:

$$f(\mathbf{w}^{(t)}) \approx f(\mathbf{w}^{(t-1)}) - \eta \|\nabla f(\mathbf{w}^{(t-1)})\|^2.$$



Gradient Descent Guarantees

Theorem 2: GD on Convex Functions

Theorem (Gradient descent on convex functions). If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t \to \infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Gradient Descent Guarantees

Theorem 2: GD on Convex Functions

Theorem (Gradient descent on convex functions). If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t \to \infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Gradient Descent Guarantees

Theorem 2: GD on Convex Functions

Theorem (Gradient descent on convex functions). If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t\to\infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}).$$

we'll *eventually* reach a global minimum!

Gradient Descent Guarantees

Theorem 2: GD on Convex Functions

Theorem (Gradient descent on convex functions). If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t \to \infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Convex: the "bowl-shaped" functions!

Gradient Descent Guarantees

Theorem 2: GD on Convex Functions

Theorem (Gradient descent on convex functions). If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t \to \infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

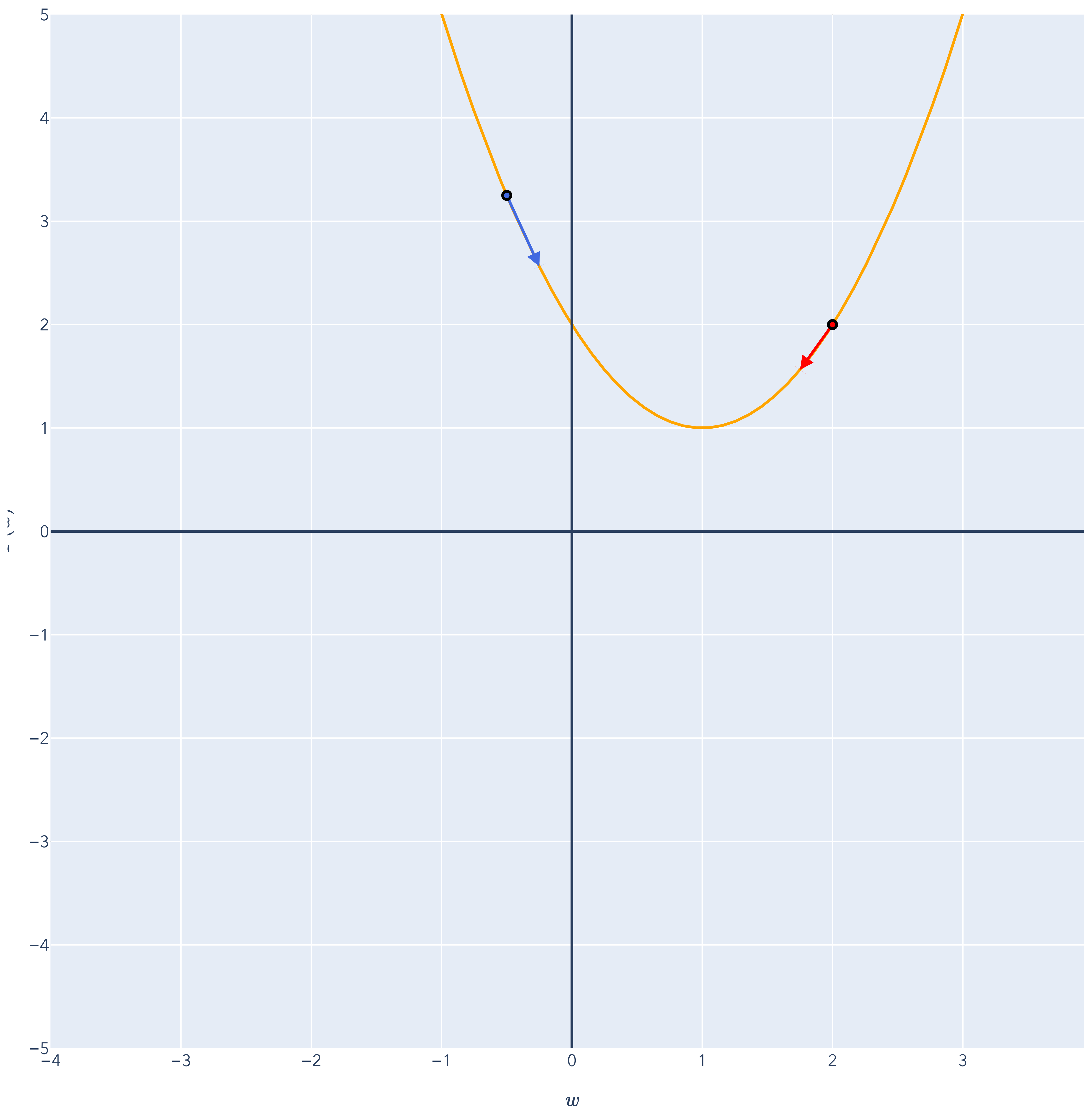
Convex: the "bowl-shaped" functions!

Convex Functions

A preview

If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t\to\infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}).$$

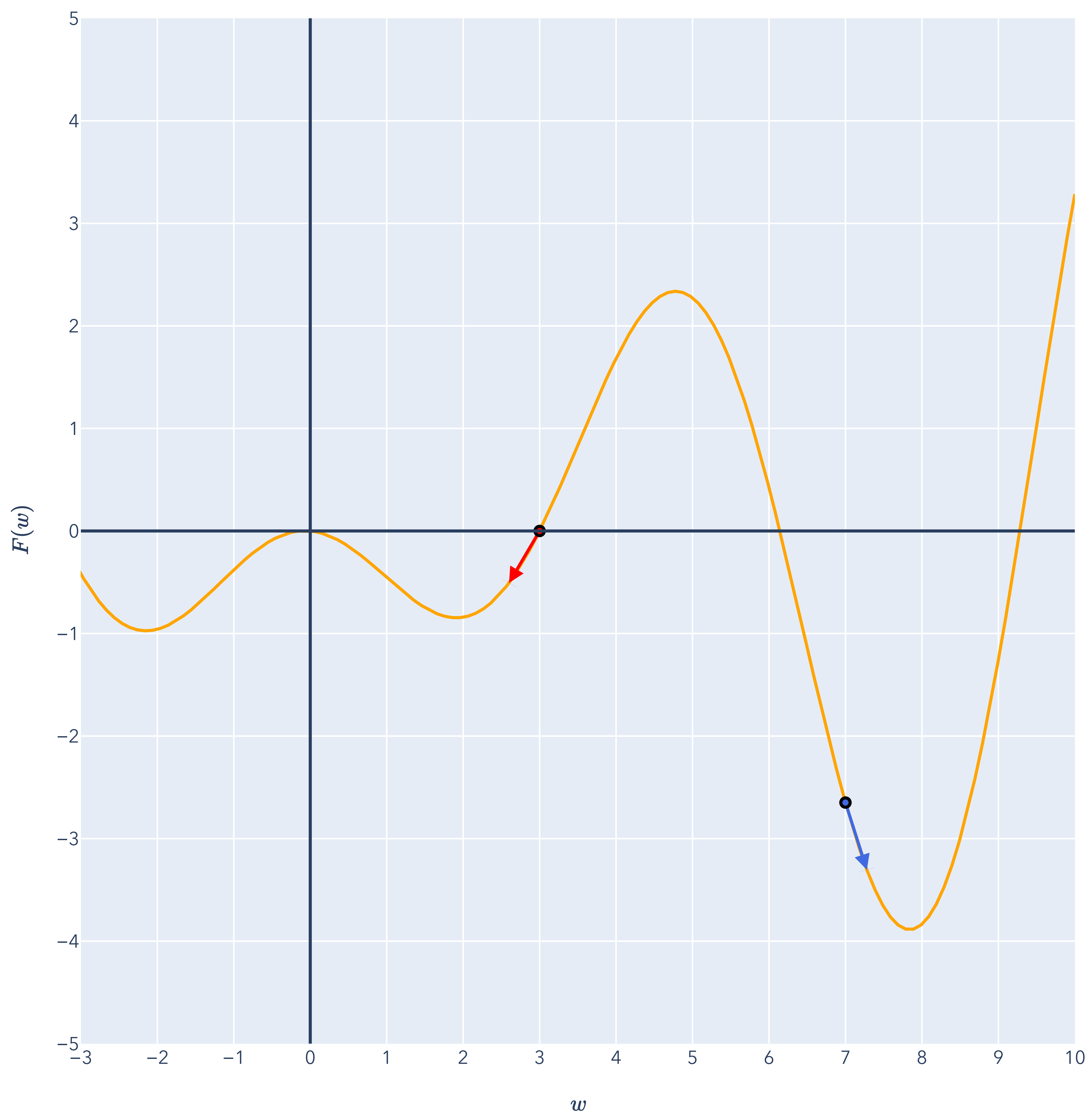


Convex Functions

A preview

If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t \to \infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

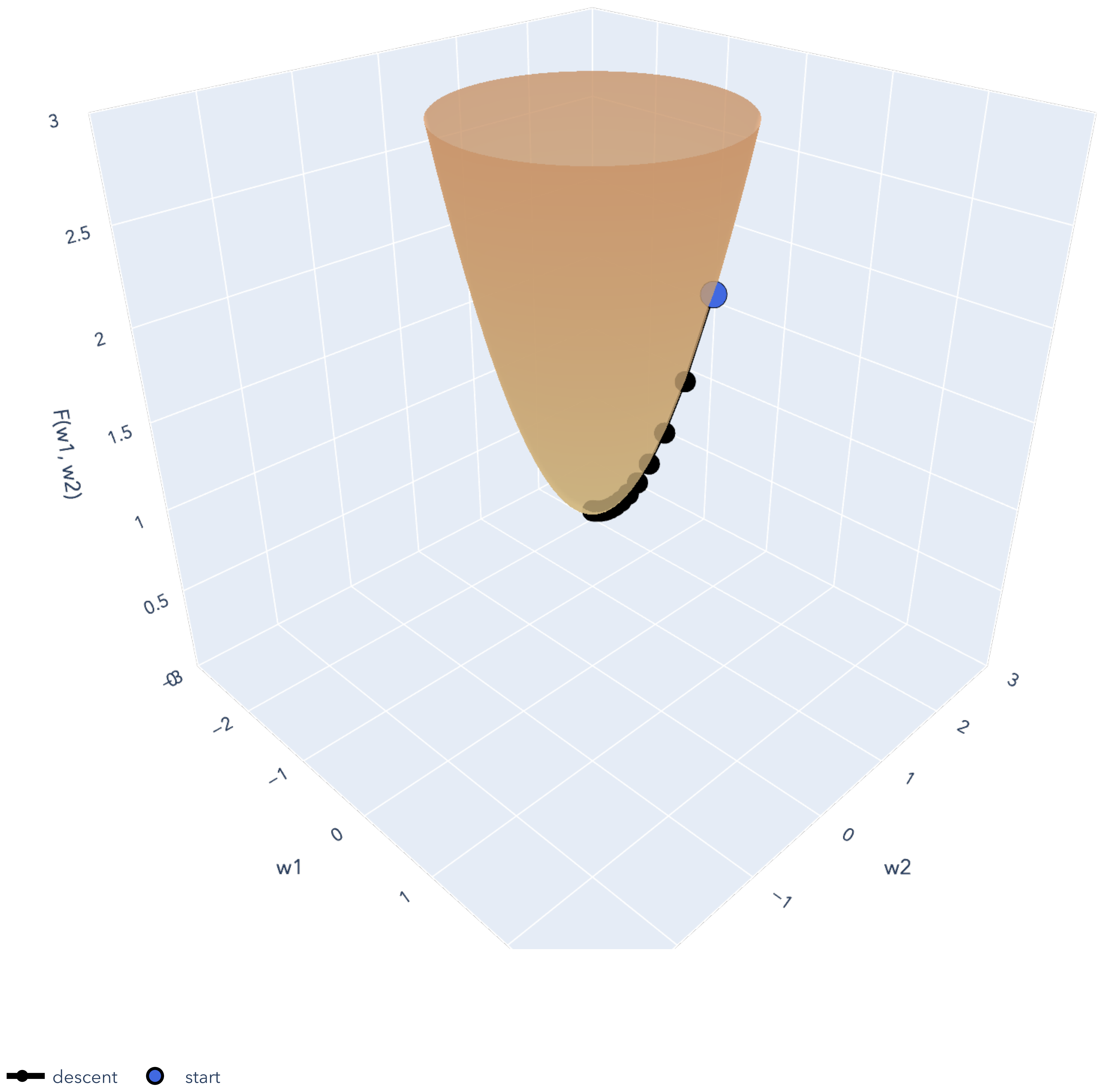


Convex Functions

A preview

If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

$$\lim_{t\to\infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}).$$

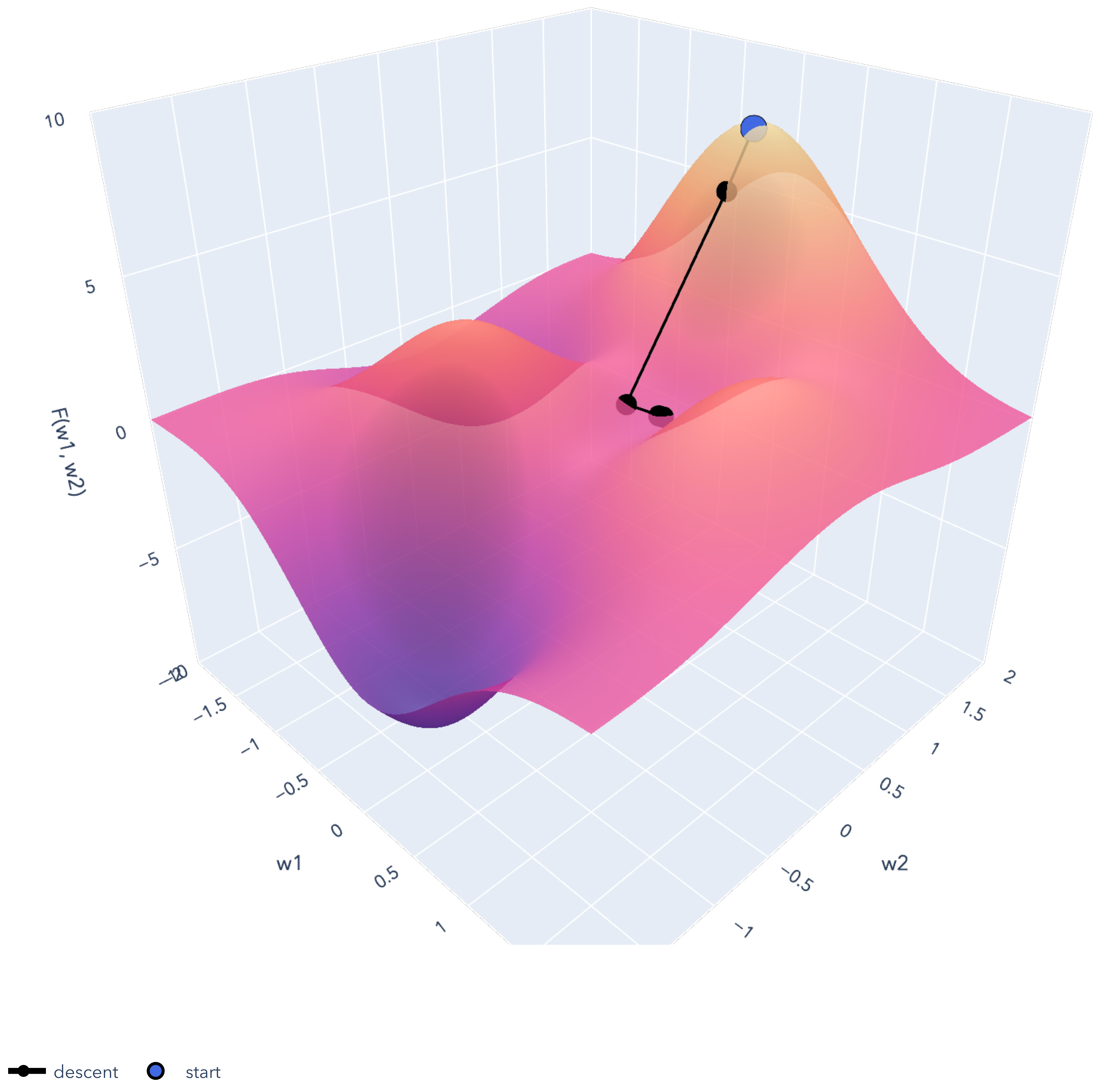


Convex Functions

A preview

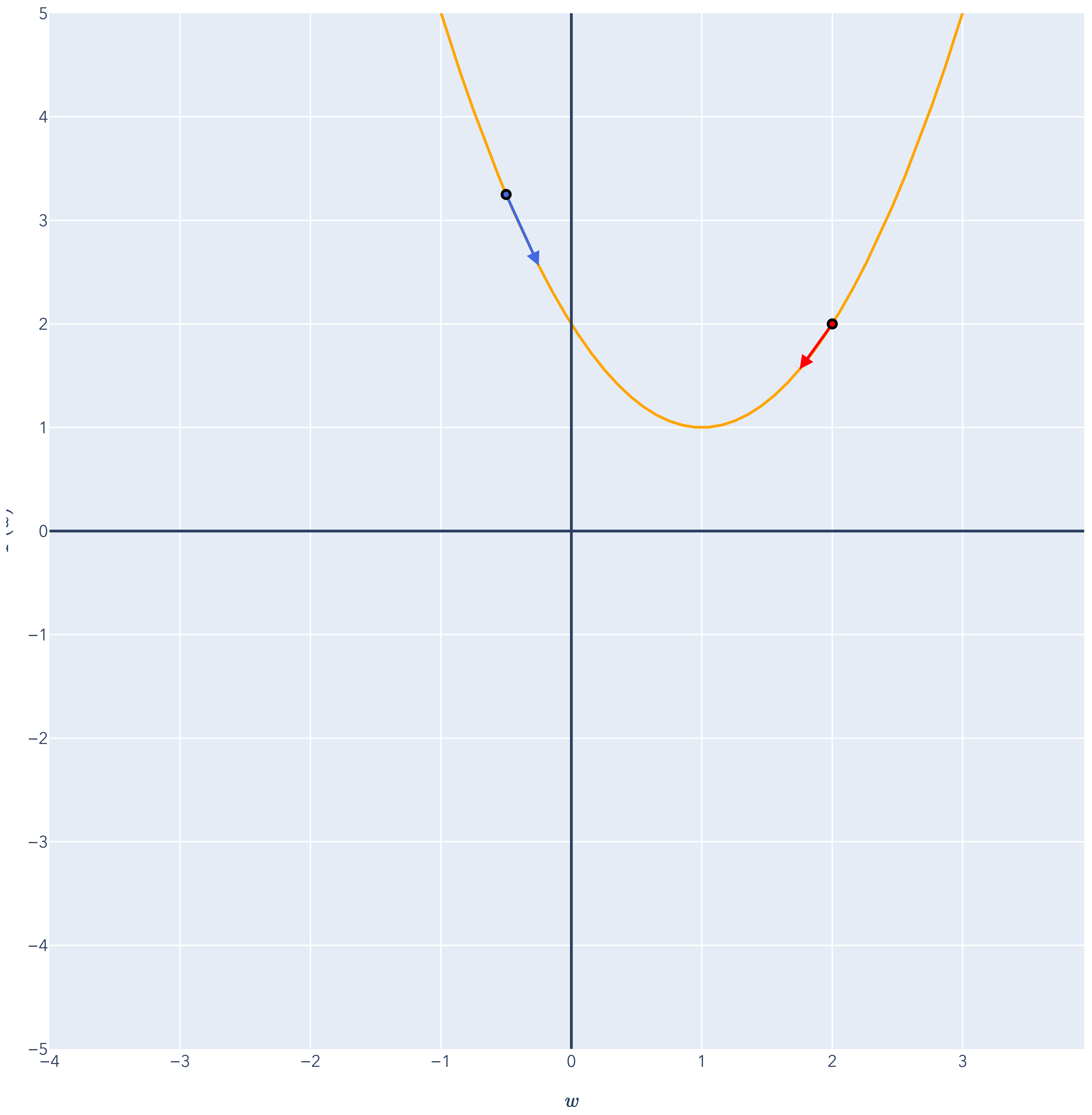
If f is convex and "smooth enough," then there is a choice of $\eta > 0$ such that for any initial $\mathbf{w}^{(0)} \in \mathbb{R}^d$, the iterates of gradient descent $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ satisfy

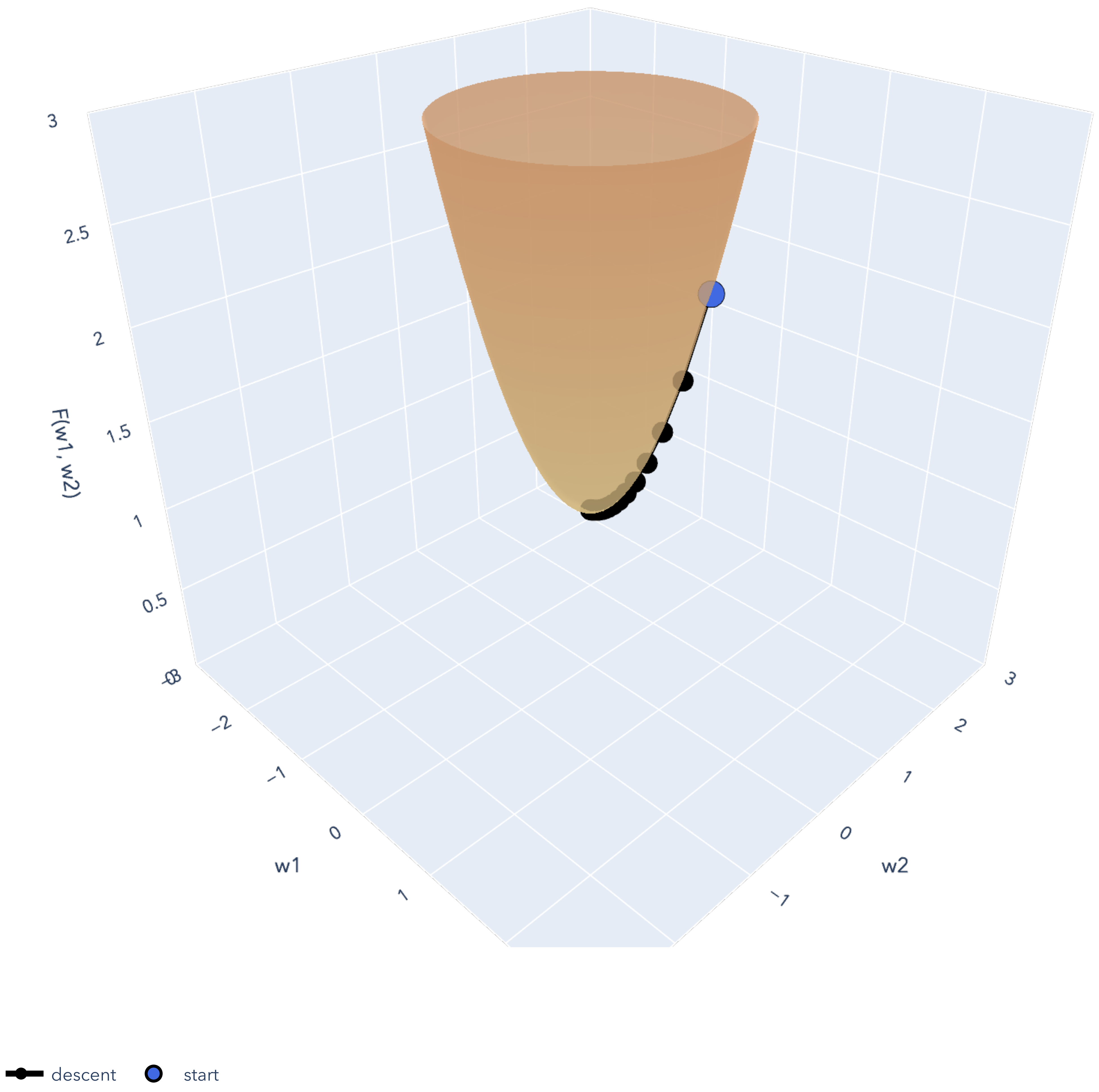
$$\lim_{t\to\infty} f(\mathbf{w}^{(t)}) = \min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}).$$



Convex Functions

A preview





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.”

Gradient descent. We write down the full algorithm for gradient descent, the second “story” of our course. First, we prove the informal descent lemma. Then, we use Taylor series to formalize it.

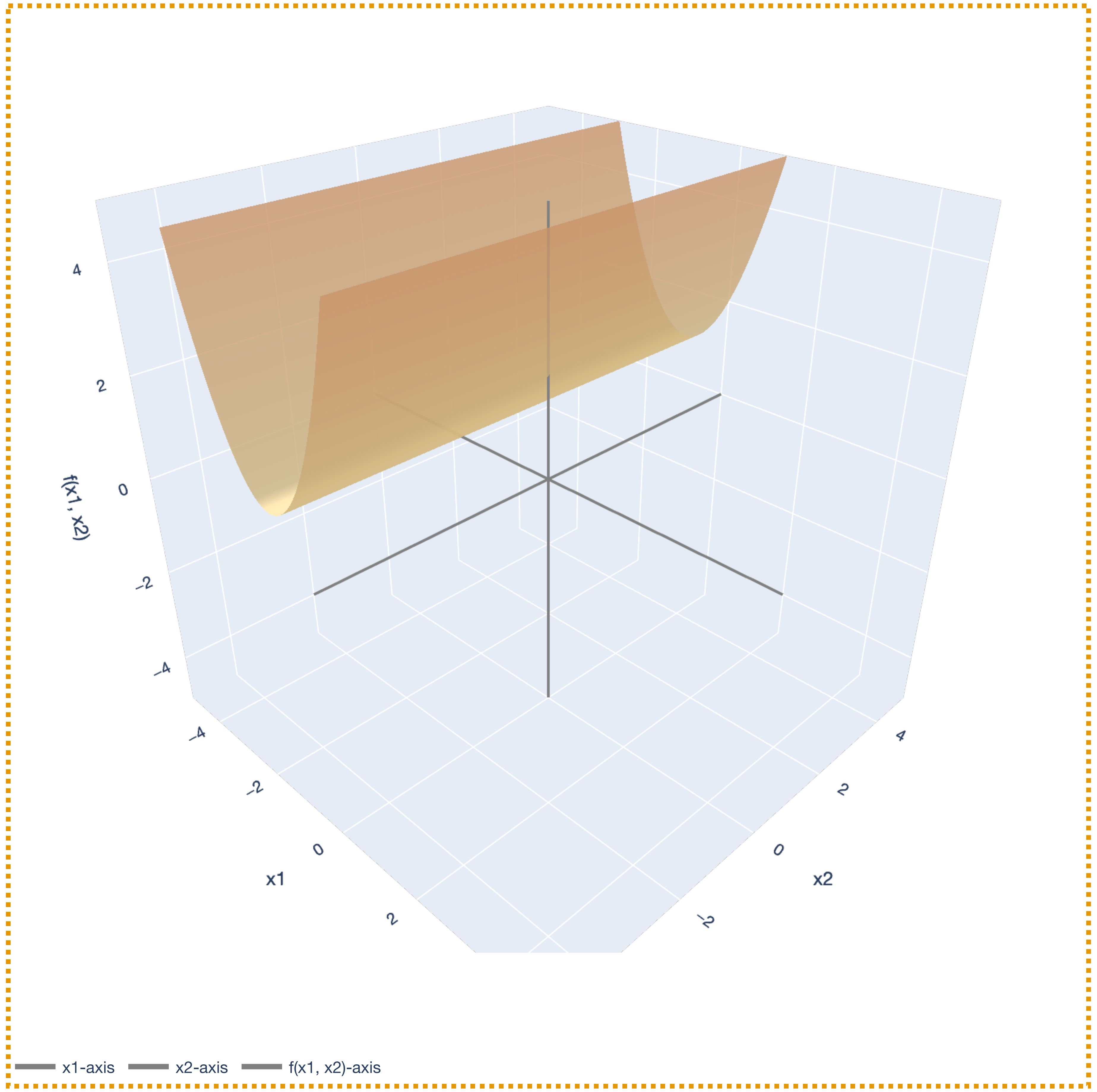
Taylor series. We define the Taylor series of a function, which is an “infinite polynomial” that approximates a function at a point.

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

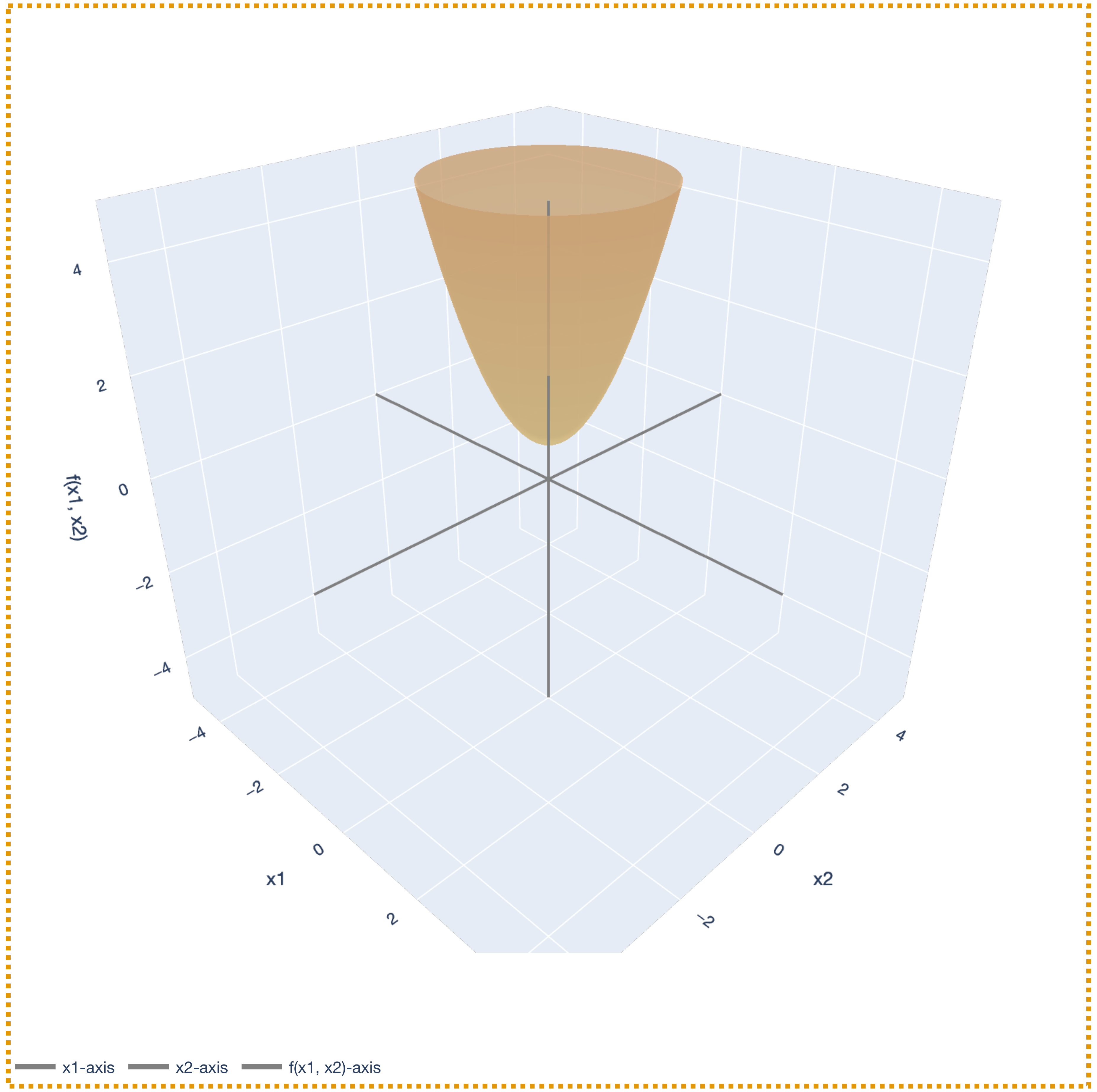
Taylor’s Theorem. To quantify how bad our approximations are, we can use Taylor’s Theorem.

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