Math for Machine Learning

Week 4.1: Optimization and the Lagrangian Method

By: Samuel Deng

Logistics & Announcements

- PS3 RELEASED. (DUE NEXT MONDAY).
- PS2 DUE TMRW (MONDAY JUL. 22 11:59 PM).
- ☆ MID-COURSE SURVEY (ON ED). → optional but highly recommended!
- WEEK 4 LECTURES ONLINE!

⇒ ASK ME QUESTIONS!! (on Ed)

at OH THIS WEEK
3PM - 5PM
Mon. Wed.

EXTRA OH (TBD).

Lesson Overview

Optimization. Minimize an objective function $\boxed{f : \mathbb{R}^d \to \mathbb{R}}$ with the possible requirement that the minimizer $\mathbf{x}^*$ belongs to a constraint set $\boxed{\mathscr{C} \subseteq \mathbb{R}^d}$.

Lagrangian. For optimization problems with $\mathscr{C}$ defined by equalities/inequalities, the Lagrangian is a function $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ that "unconstrains" the problem.

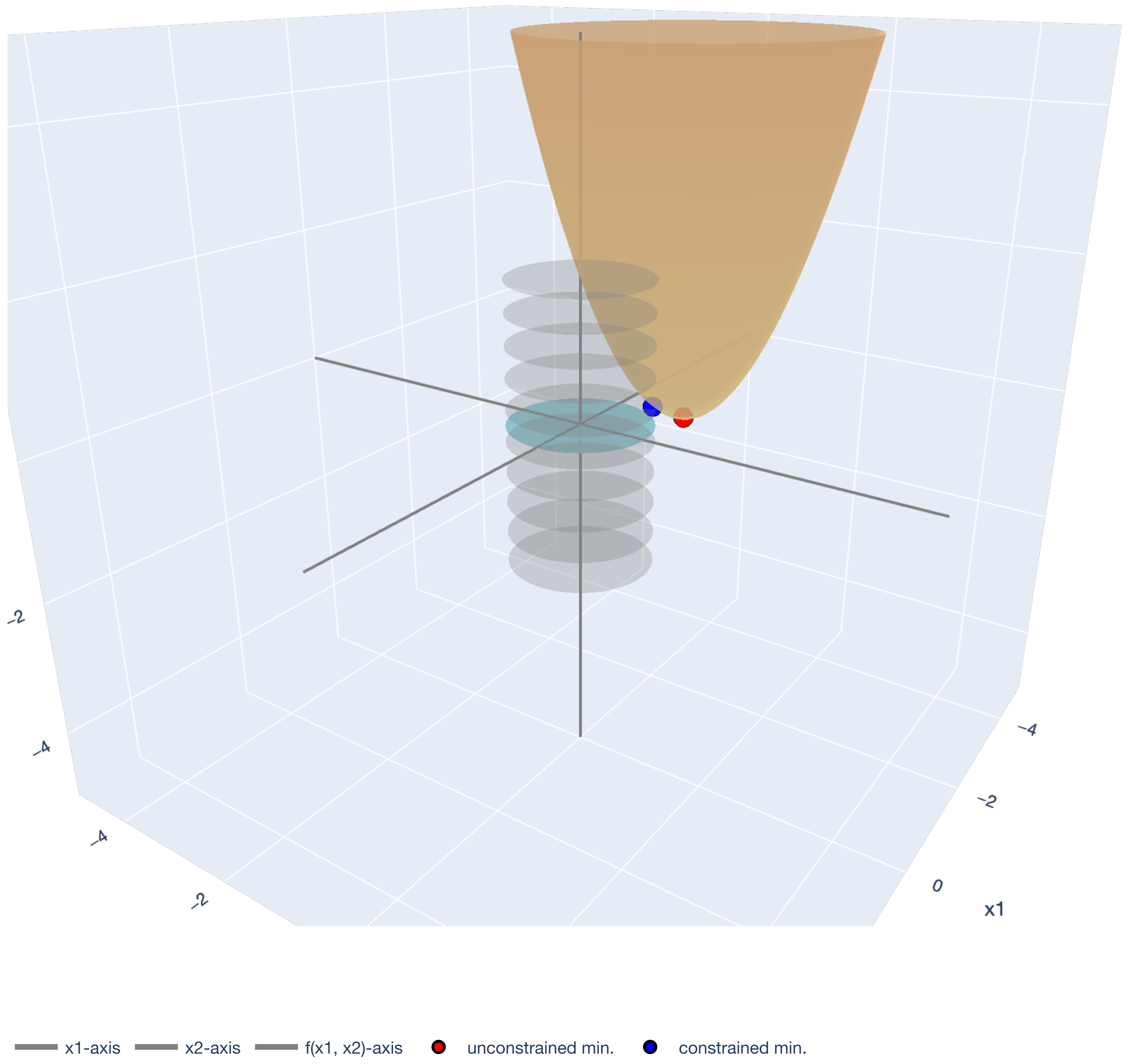
Unconstrained local optima. With no constraints, the standard tools of calculus give conditions for a point $\mathbf{x}^*$ to be optimal, at least to all points close to it. First order condition + second order condition. $\left\{ \begin{array}{l} f'(x)=0 \\ f''(x)>0. \end{array} \right.$

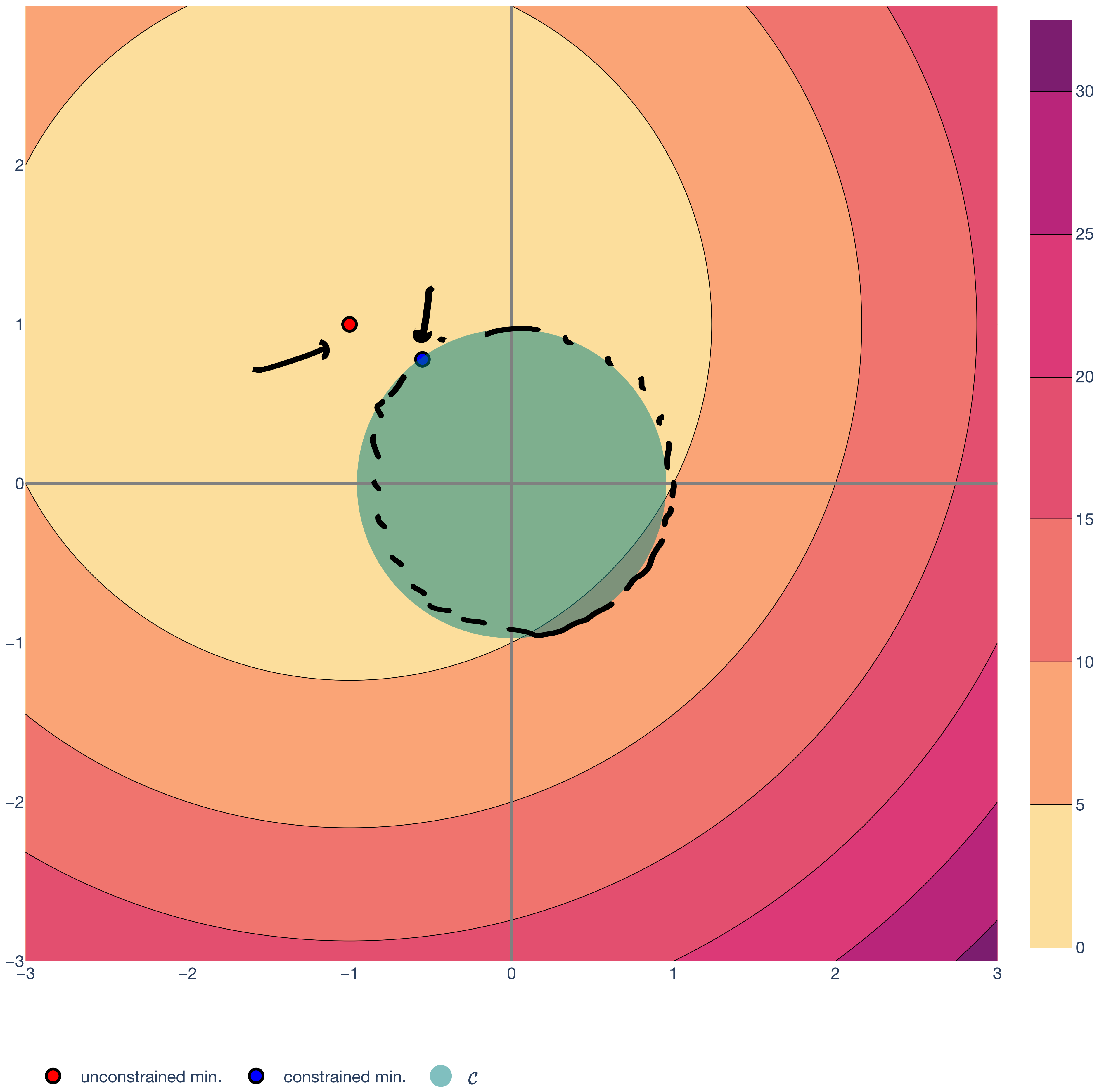
Constrained local optima (Lagrangian and KKT). When $\mathscr{C}$ is represented by inequalities and equalities, we can use the method of *Lagrange multipliers* and the *KKT Theorem* to "unconstrain" the problem.

Ridge regression and minimum norm solutions. By constraining the norm of $\mathbf{w}^* \in \mathbb{R}^d$ of least squares (i.e. $\|\mathbf{w}^*\|$), we obtain more "stable" solutions.

Lesson Overview

Big Picture: Least Squares

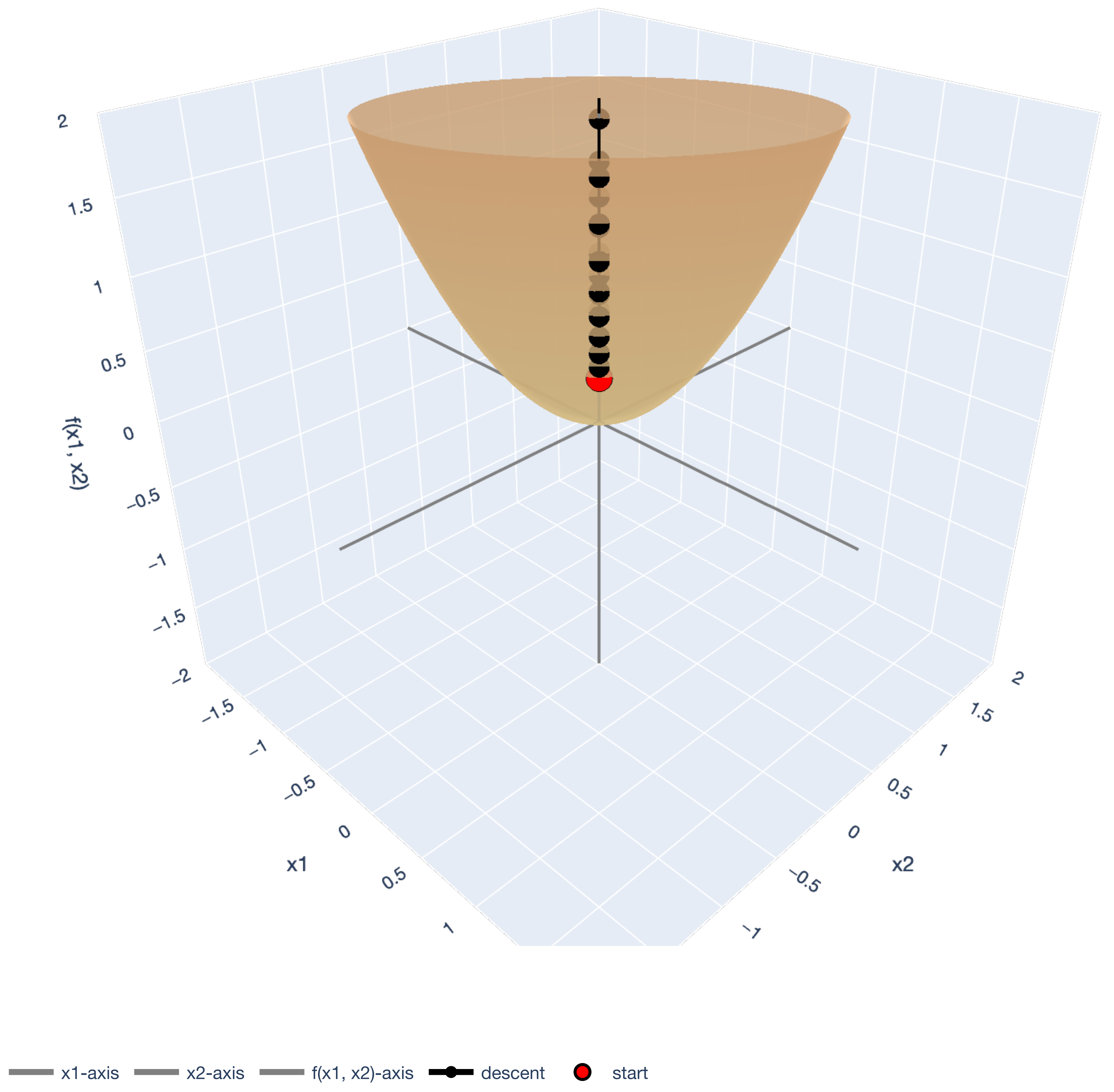


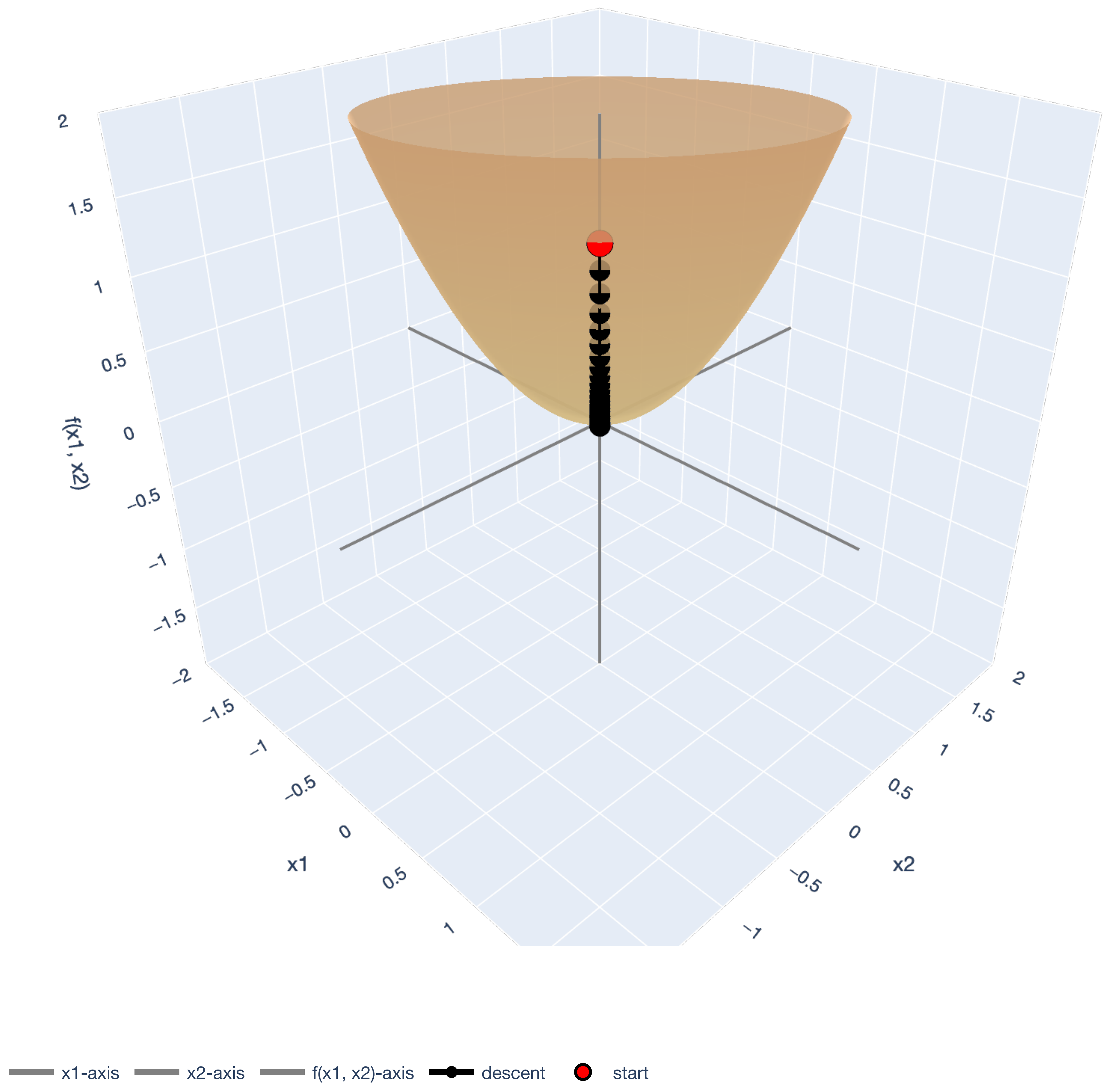


Lesson Overview

Big Picture: Gradient Descent

$\eta > 0$ sufficiently small.





Optimization Problems

Definition and examples

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 $\mathscr{C} \subseteq \mathbb{R}^d$:

$$\begin{aligned} \underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \boxed{\mathbf{x} \in \mathscr{C}} \end{aligned}$$

$\mathscr{C} = \mathbb{R}^d$.

Motivation

Components of an optimization problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \quad & f(\mathbf{x}) \longleftarrow \text{objective.} \\ \text{subject to} \quad & \mathbf{x} \in \mathscr{C} \longleftarrow \text{constraint.} \end{aligned}$$

$f : \mathbb{R}^d \to \mathbb{R}$ is the ***objective function.***

$\mathscr{C} \subseteq \mathbb{R}^{\cancel{x}\,d}$ is the ***constraint/feasible set.*** feasible : $x \in \mathscr{C}$.

Motivation

Components of an optimization problem

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

$f : \mathbb{R}^d \to \mathbb{R}$ is the ***objective function.***

$\mathscr{C} \subseteq \mathbb{R}^n$ is the ***constraint/feasible set***.

GOAL

$\mathbf{x}^*$ is an ***optimal solution (global minimum)*** if "minimizer"

$$\underline{\mathbf{x}^* \in \mathscr{C}} \quad \text{and} \quad \underline{f(\mathbf{x}^*)} \leq \underline{f(\mathbf{x})}, \quad \text{for all } \underline{\mathbf{x} \in \mathscr{C}}.$$

The ***optimal value*** is $f(\mathbf{x}^*)$. Our goal is to find $\mathbf{x}^*$ and $f(\mathbf{x}^*)$.

minimum

(after plugging in x^*).

Motivation

Components of an optimization problem

maximize $-f(x)$ $\iff$

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

$f : \mathbb{R}^d \to \mathbb{R}$ is the ***objective function.***

$\mathscr{C} \subseteq \mathbb{R}^n$ is the ***constraint/feasible set***.

$\mathbf{x}^*$ is an ***optimal solution (global minimum)*** if

$$\mathbf{x}^* \in \mathscr{C} \quad \text{and} \quad f(\mathbf{x}^*) \leq f(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathscr{C}.$$

The ***optimal value*** is $f(\mathbf{x}^*)$. Our goal is to find $\mathbf{x}^*$ and $f(\mathbf{x}^*)$.

Note: to maximize $f(\mathbf{x})$, just minimize $-f(\mathbf{x})$. So we'll only focus on *minimization* problems.

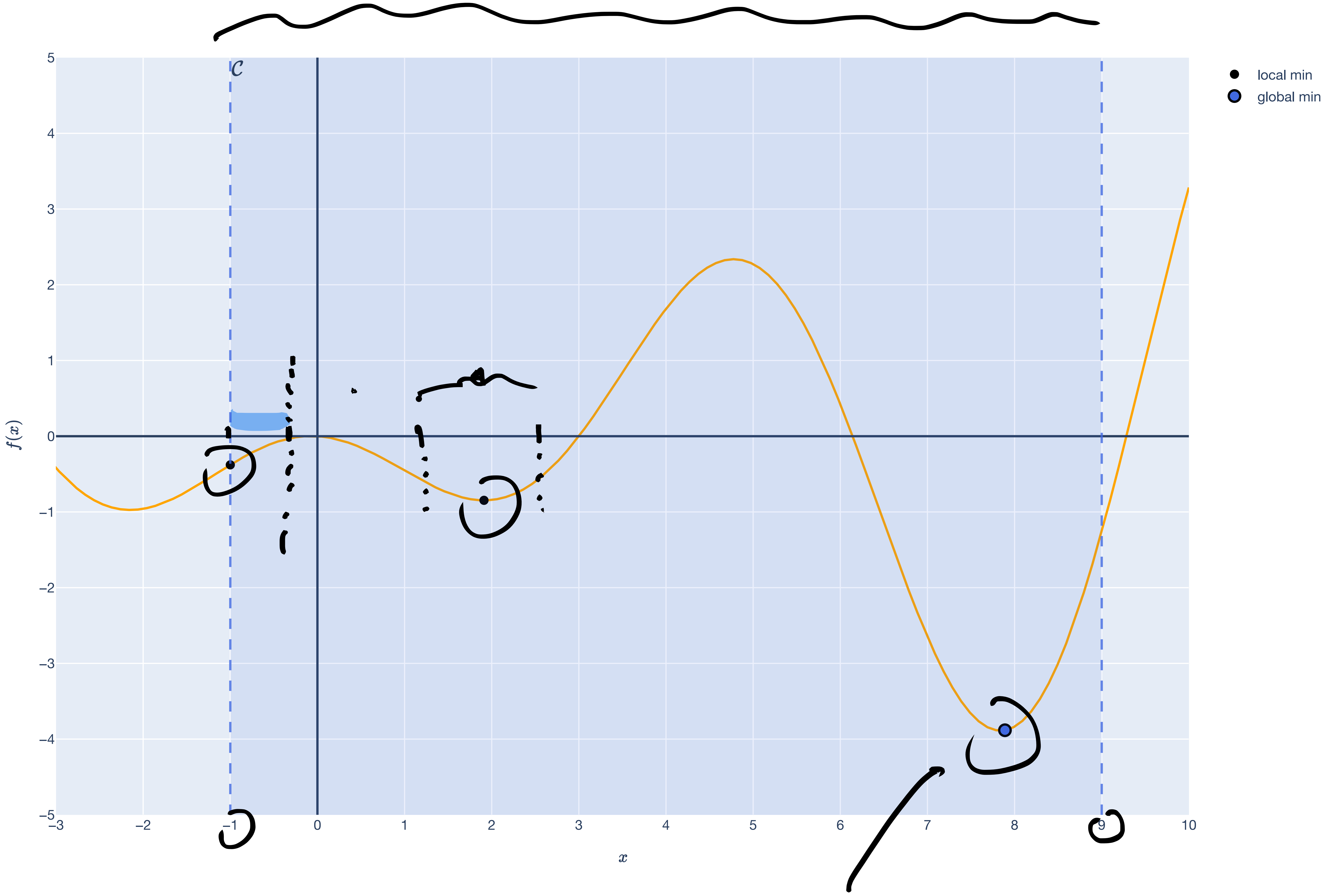
Motivation

Optimization in single-variable calculus

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

Intermediary goal: Find the *local minima.*

⇒ Minimum for points in a neighborhood of x^*.



Motivation

Example: Linear Programming

- OPERATIONS RESEARCH.
- ECONOMICS
- COMPUTER SCIENCE

$\vdots$

"cost"

Let $\mathbf{c} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^n$ be fixed.

$c^T x = \sum_{i=1}^{d} c_i x_i$

Let $\mathbf{x} \in \mathbb{R}^d$ be the ***decision/free variables***.

$x = (x_1, \ldots, x_d)$

Each variable has cost

$$\begin{aligned} \underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimize}} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \end{aligned}$$

n constraints

$$\underbrace{\begin{bmatrix} - a_1 - \\ \vdots \\ - a_n - \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}}_{b}$$

$\preceq$ is *element-wise* inequality: $\mathbf{a}_i^\top \mathbf{x} \le b_i$ for all $i \in [n]$.

- n constraints.
- d variables

Motivation

Example: Linear Programming ($d = 3$, $n = 7$)

We're cooking some NYC classics again. Suppose we have:

100 bacon, 120 egg, 150 cheese, and 300 (sandwich) rolls.

$d=3$

There are three recipes we know:

Bacon egg and cheese (BEC) requires 1 bacon, 1 egg, 1 cheese, and 1 roll.

Cost (including labor): $3

Egg and cheese (EC) requires 0 bacon, 2 egg, 1 cheese, and 1 roll.

Cost (including labor): $2

Bacon egg omelette (BEO) requires 1 bacon, 3 egg, $1/2$ cheese, and 0 roll.

Cost (including labor): $1

Motivation

Example: Linear Programming ($d = 3$, $n = 7$)

We're cooking some NYC classics again. Suppose we have:

100 bacon, 120 egg, 150 cheese, and 300 (sandwich) rolls.

$n = 4$?.

There are three recipes we know:

1. Bacon egg and cheese (BEC) requires 1 bacon, 1 egg, 1 cheese, and 1 roll.

Cost (including labor): \$3

2. Egg and cheese (EC) requires 0 bacon, 2 egg, 1 cheese, and 1 roll.

Cost (including labor): \$2

3. Bacon egg omelette (BEO) requires 1 bacon, 3 egg, $1/2$ cheese, and 0 roll.

Cost (including labor): \$1

$\Longrightarrow$

Decision variables?

$d = 3$

$\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$

$x_1 \geq 0$

$x_2 \geq 0$

$x_3 \geq 0$.

x_1 = number of BEC,

x_2 = number of EC,

x_3 = number of BEO

Constraints?

Bacon: $\mathbf{a}_1 = (1,0,1)$, $b_1 = 100$

Egg: $\mathbf{a}_2 = (1,2,3)$, $b_2 = 120$

Cheese: $\mathbf{a}_3 = (1,1,1/2)$, $b_3 = 150$

Roll: $\mathbf{a}_4 = (1,1,0)$, $b_4 = 300$

Objective?

$c = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$

$\boxed{\mathbf{c}^\top \mathbf{x} = 3x_1 + 2x_2 + x_3}$

Motivation

Example: Linear Programming ($d = 3$, $n = 7$)

Decision variables?

$$\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$$

x_1 = number of BEC,

x_2 = number of EC,

x_3 = number of BEO

Constraints?

Bacon: $\mathbf{a}_1 = (1,0,1)$, $b_1 = 100$

Egg: $\mathbf{a}_2 = (1,2,3)$, $b_2 = 120$

Cheese: $\mathbf{a}_3 = (1,1,1/2)$, $b_3 = 150$

Roll: $\mathbf{a}_4 = (1,1,0)$, $b_4 = 300$

Objective?

$$\mathbf{c}^\top \mathbf{x} = 3x_1 + 2x_2 + x_3$$

$$\Longrightarrow$$

Linear program:

$$\begin{aligned}
\text{minimize} \quad & \boxed{3x_1 + 2x_2 + x_3} \\
\text{subject to} \quad & x_1 + x_3 \leq 100 \\
& x_1 + 2x_2 + 3x_3 \leq 120 \\
& x_1 + x_2 + 0.5x_3 \leq 150 \\
& x_1 + x_2 \leq 300 \\
& x_1 \geq 0 \\
& x_2 \geq 0 \\
& x_3 \geq 0
\end{aligned}$$

TOTAL COST.

Bacon

Egg

⋮

Nonnegative.

Motivation

Example: Linear Programming ($d = 3$, $n = 7$)

$$\begin{bmatrix} -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \leq 0 \implies -x_1 \leq 0 \iff \boxed{0 \leq x_1}$$

Linear program:

$$\begin{aligned}
\text{minimize} \quad & 3x_1 + 2x_2 + x_3 \\
\text{subject to} \quad & x_1 + x_3 \leq 100 \\
& x_1 + 2x_2 + 3x_3 \leq 120 \\
& x_1 + x_2 + 0.5x_3 \leq 150 \\
& x_1 + x_2 \leq 300 \\
& x_1 \geq 0 \\
& x_2 \geq 0 \\
& x_3 \geq 0
\end{aligned}$$

$\Longrightarrow$

LP in matrix form:

$c = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$

$$\begin{aligned}
\text{minimize} \quad & 3x_1 + 2x_2 + x_3 \\
\text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}
\end{aligned}$$

constraint

$$\begin{array}{l} \text{Bacon} \\ \text{Eggs} \\ \text{cheese} \\ \\ \text{Roll} \\ \\ \text{nonnegativity} \\ \\ \end{array} \mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & \frac{1}{2} \\ 1 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 100 \\ 120 \\ 150 \\ 300 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Regression

Setup

FIXED

Observed: Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^{\cancel{d}}$. n.

FIXED
$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

Unknown: *Weight vector* $\mathbf{w} \in \mathbb{R}^d$ with weights $w_1, \ldots, w_d$.

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

Regression

Setup

<u>Goal:</u> For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

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

Least Squares

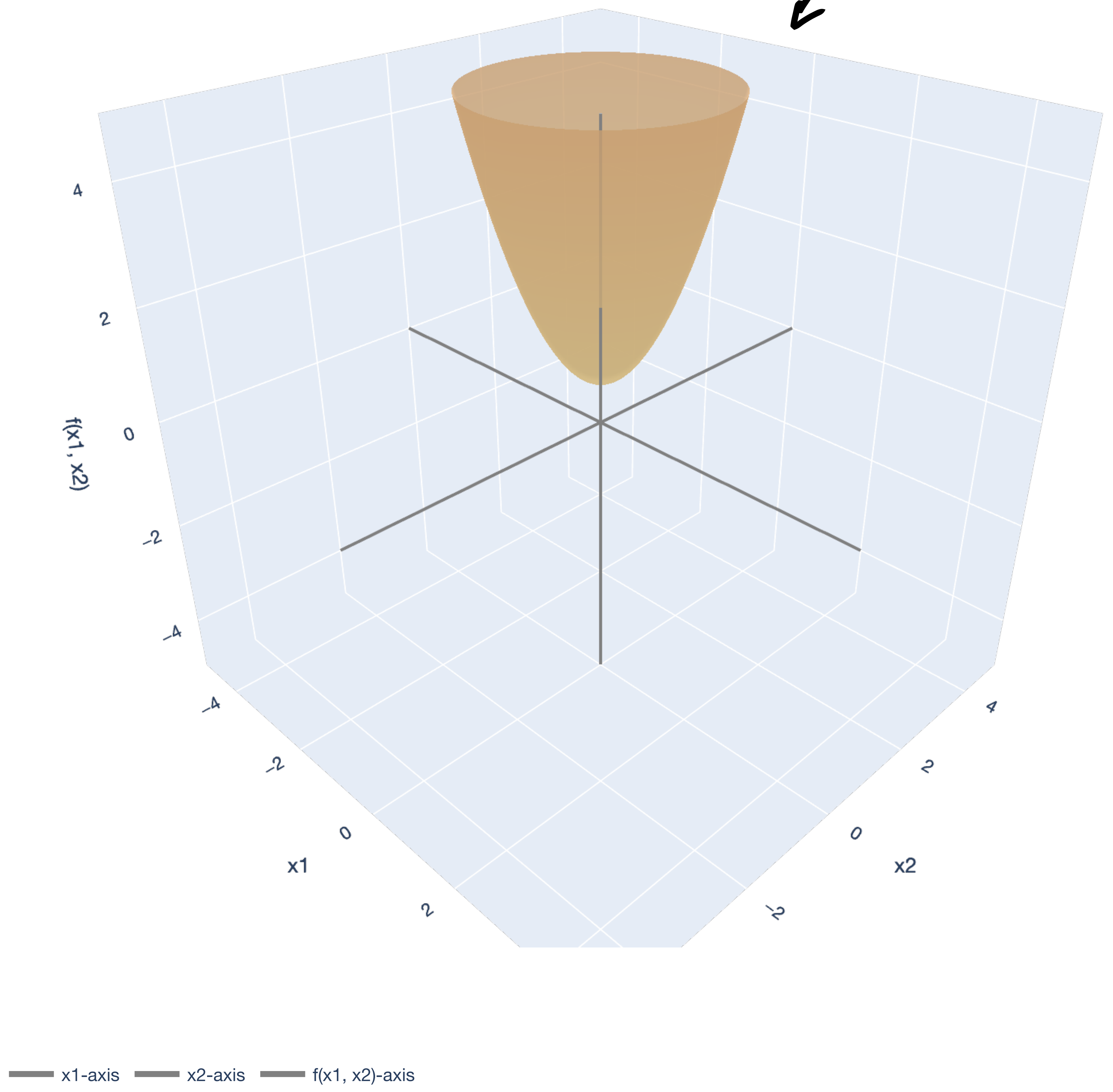
Optimization Problem

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^{n}$ be fixed. Let $\mathbf{w} \in \mathbb{R}^{d}$ be the decision variables.

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

$$\text{subject to} \quad \underbrace{\mathbf{w} \in \mathbb{R}^d}_{\text{UNCONSTRAINED}}$$

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



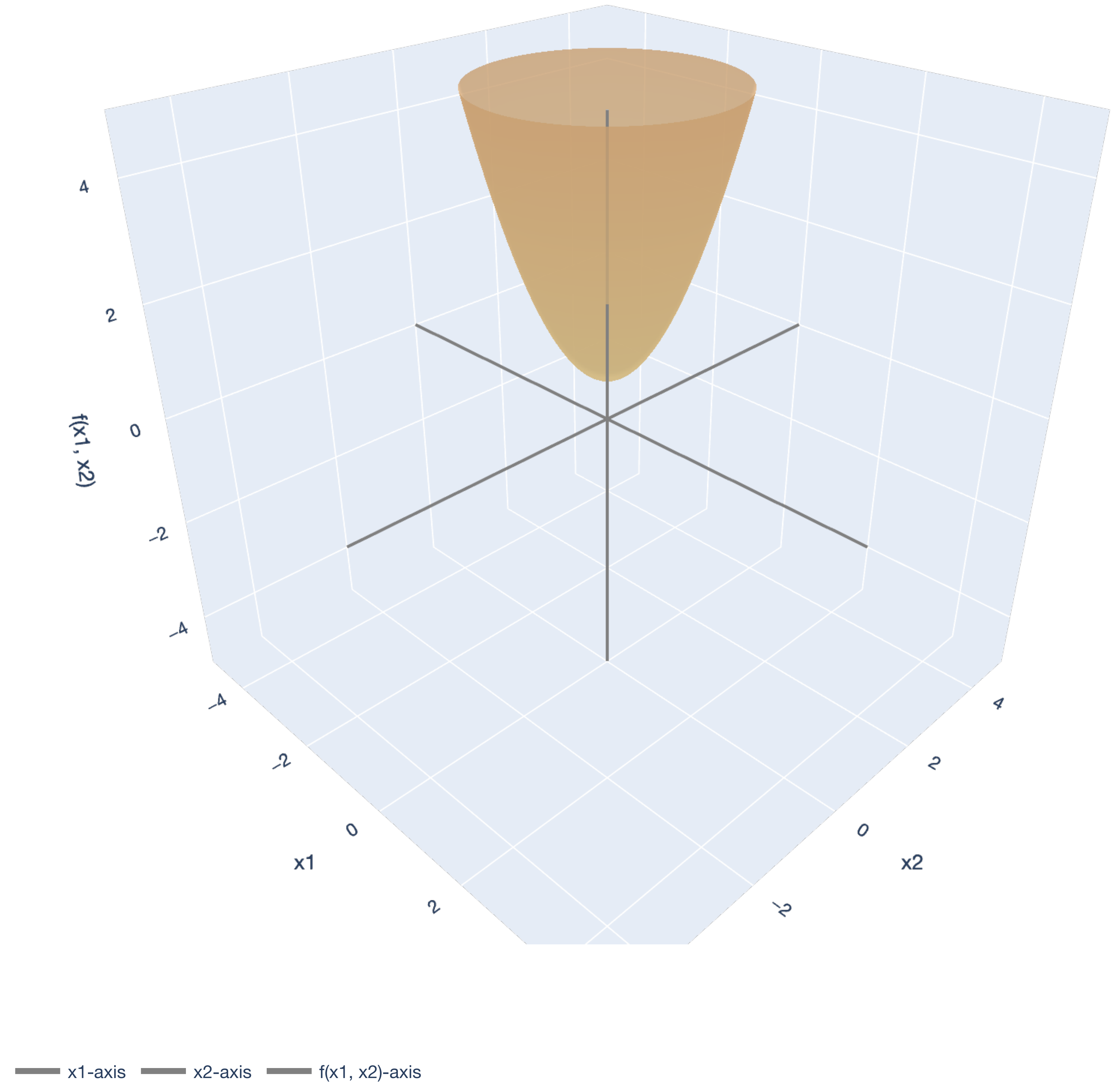
Least Squares

Optimization Problem

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$ be fixed. Let $\mathbf{w} \in \mathbb{R}^d$ be the decision variables.

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \\ \text{subject to} \quad & \mathbf{w} \in \mathbb{R}^d \end{aligned}$$

How to find the minimizer?



Least Squares

OLS Theorem

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

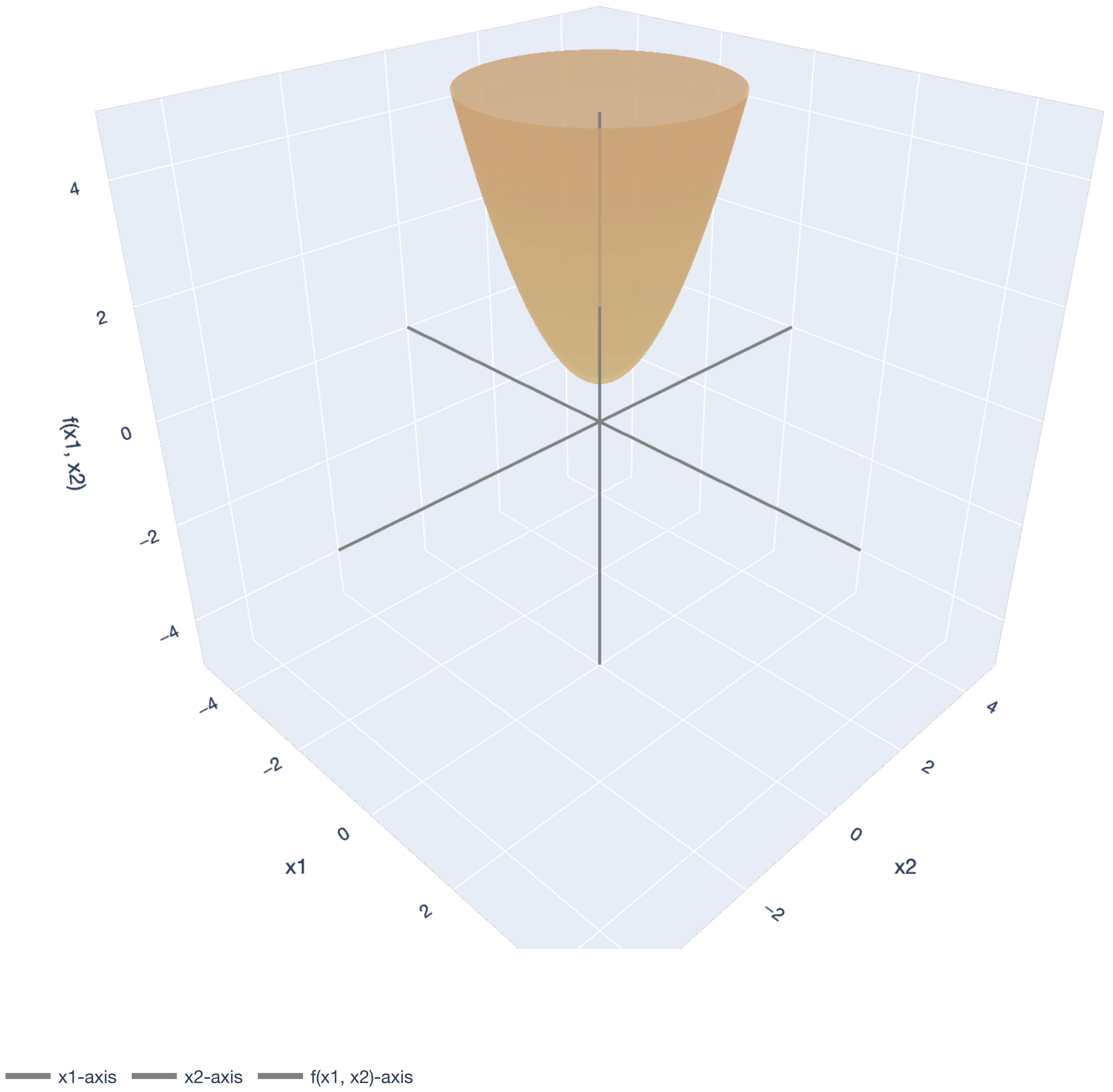
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

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

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

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

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



Least Squares

OLS Theorem

Single-var:

$f'(x) = 0$

$f''(x) > 0$?.

Proof (OLS).

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \iff$$
$$f(\mathbf{w}) = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}$$

"First derivative test." Take the gradient.

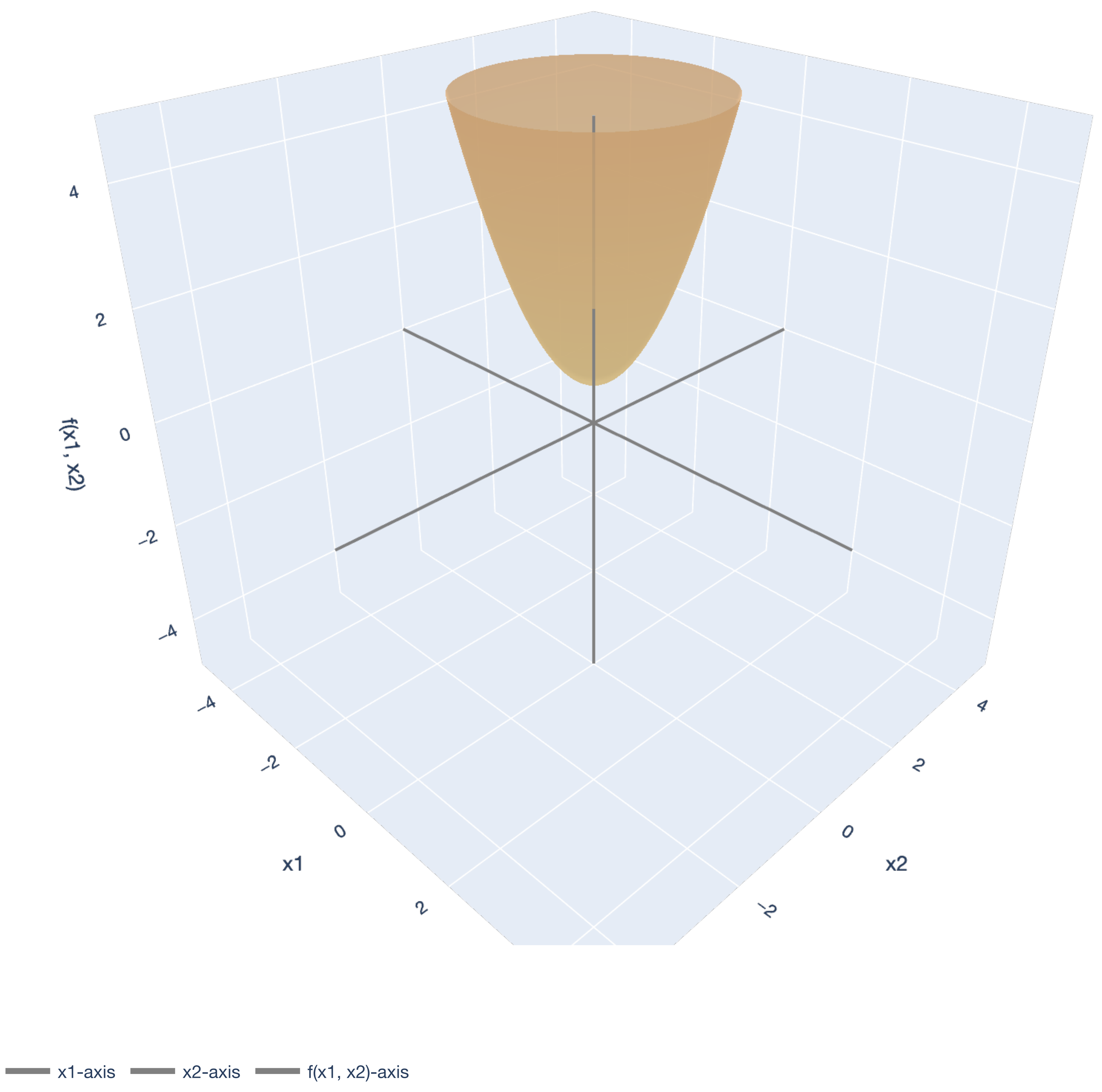
$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}.$$

Set it equal to $\mathbf{0}$.

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

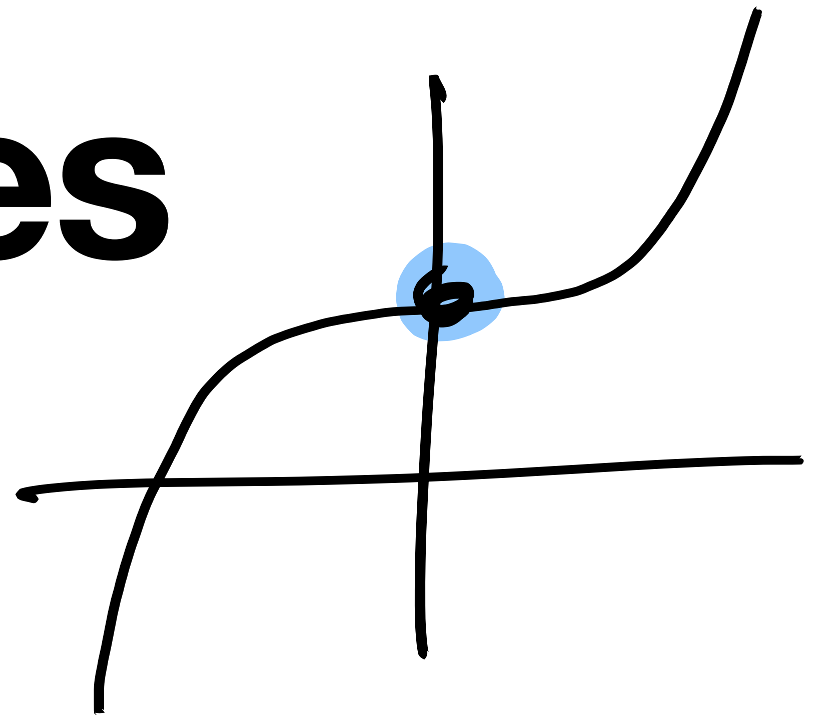
$\mathrm{rank}(\mathbf{X}) = d \implies \mathrm{rank}(\mathbf{X}^\top \mathbf{X}) = d \implies \mathbf{X}^\top \mathbf{X}$ is invertible:

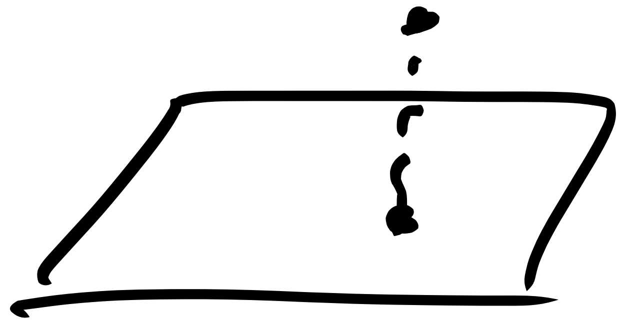
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$



Least Squares

OLS Theorem





Proof (OLS).

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \iff f(\mathbf{w}) = \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}$$

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}.$$

Set it equal to $\mathbf{0}$.

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

$\operatorname{rank}(\mathbf{X}) = d \implies \operatorname{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

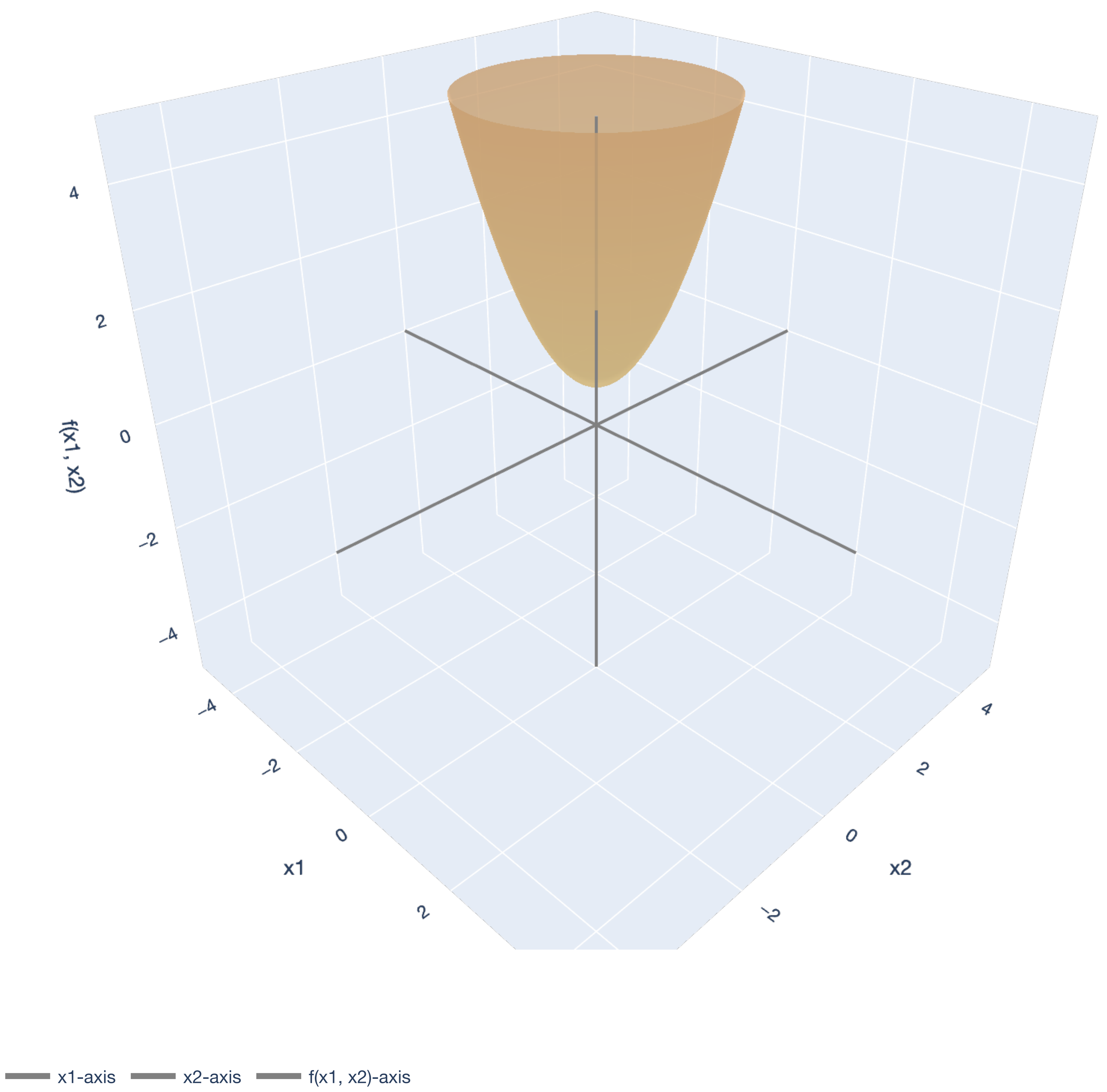
$$\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.} \leftarrow \text{candidate}$$

"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\boxed{\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}.}$$

$$\operatorname{rank}(\mathbf{X}) = d \implies \operatorname{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \lambda_1, \ldots, \lambda_d > 0$$

$$\implies \mathbf{X}^\top\mathbf{X} \text{ is positive definite!} \Longleftarrow f''(x) > 0.$$



Least Squares

OLS Theorem

<u>Proof (OLS).</u>

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \iff f(\mathbf{w}) = \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}$$

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}.$$

Set it equal to $\mathbf{0}$.

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

$\mathrm{rank}(\mathbf{X}) = d \Longrightarrow \mathrm{rank}(\mathbf{X}^\top\mathbf{X}) = d \Longrightarrow \mathbf{X}^\top\mathbf{X}$ is invertible:

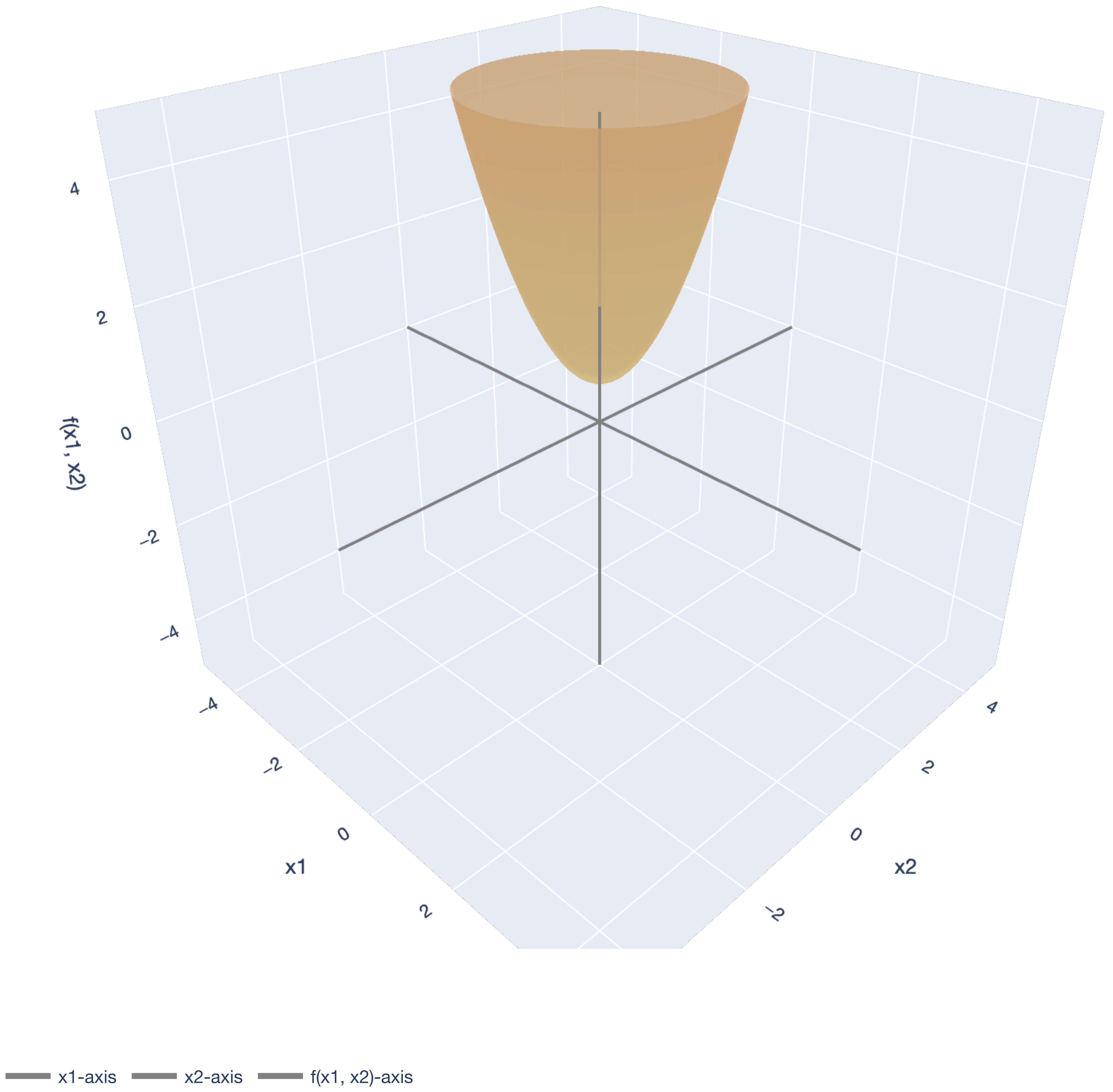
$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}.$$

$$\mathrm{rank}(\mathbf{X}) = d \implies \mathrm{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \lambda_1, \ldots, \lambda_d > 0$$

$$\implies \mathbf{X}^\top\mathbf{X} \text{ is positive definite!}$$



Local and global minima

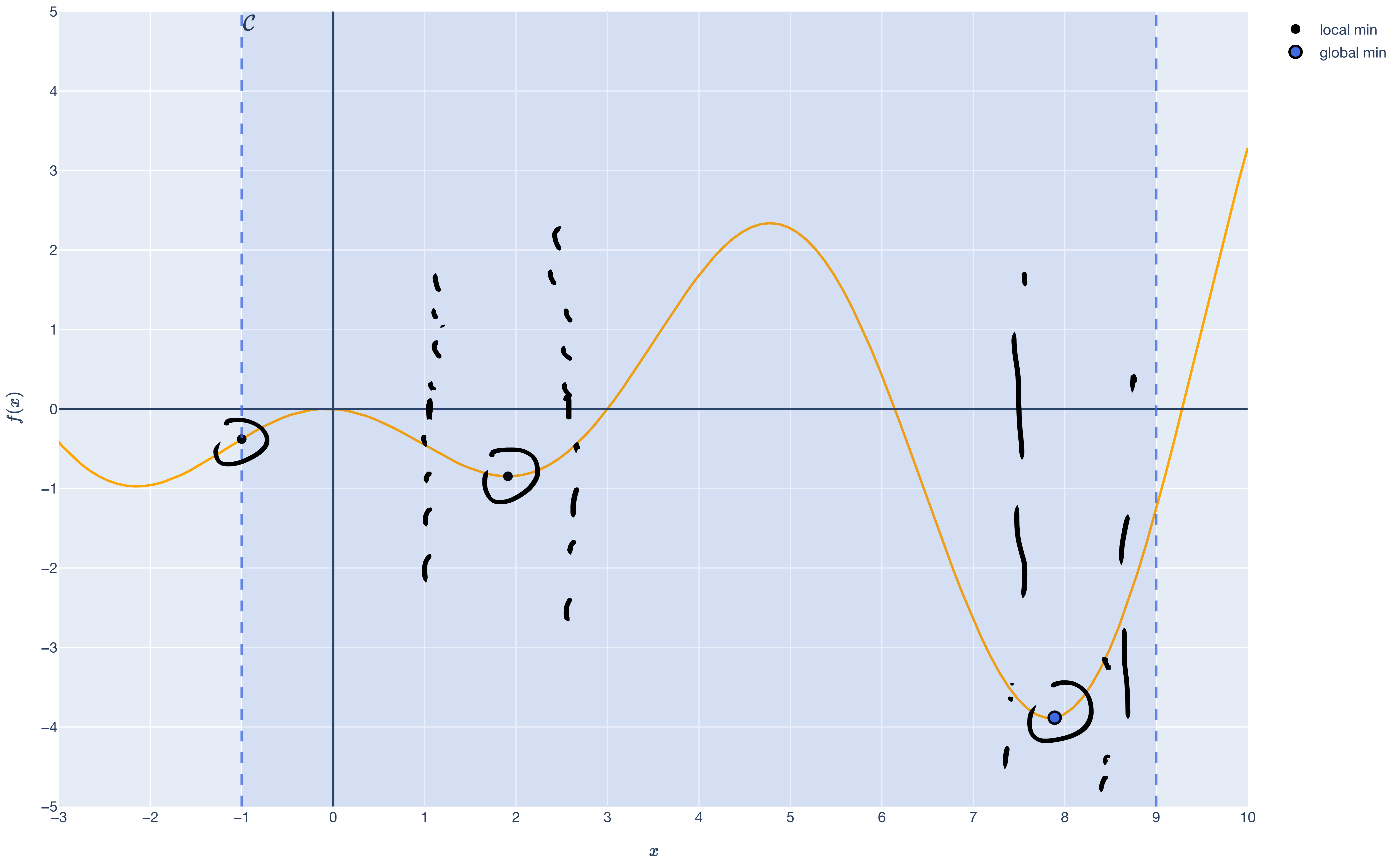
Definition of “locality” and different minima

Motivation

Optimization in single-variable calculus

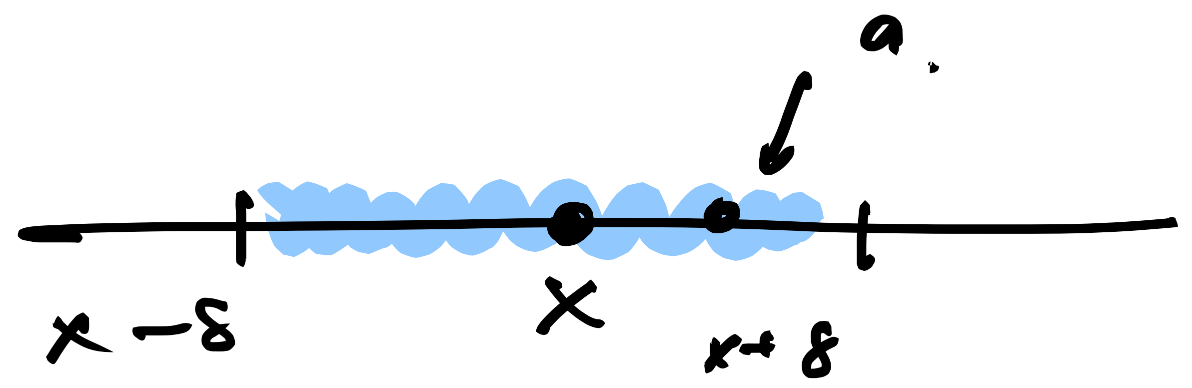
Ultimate goal: Find the *global minimum* of functions.

Intermediary goal: Find the *local minima.*



"Local" to a Point

Definition of an open ball/neighborhood



radius

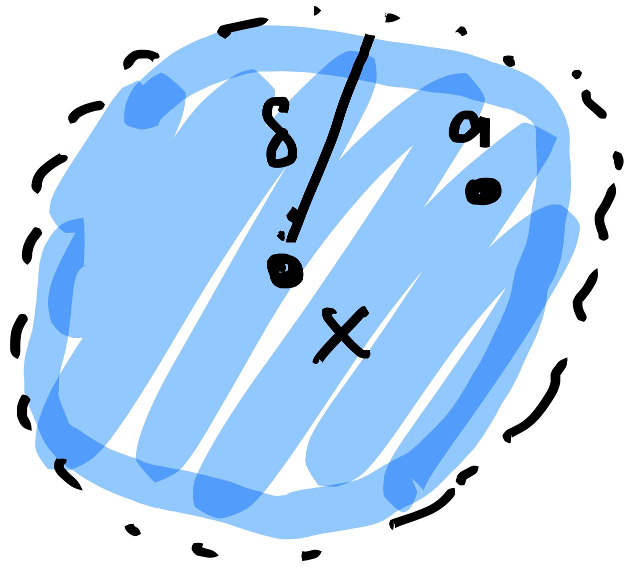
Let $\mathbf{x} \in \mathbb{R}^d$ be a point. For some real value $\delta > 0$, the ***open ball*** or ***neighborhood of radius*** δ around $\mathbf{x}$ is the set of all points: ← INSIDE OF INTERVAL/CIRCLE/SPHERE.

$$B_\delta(\mathbf{x}) := \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\| < \delta\}\,.$$

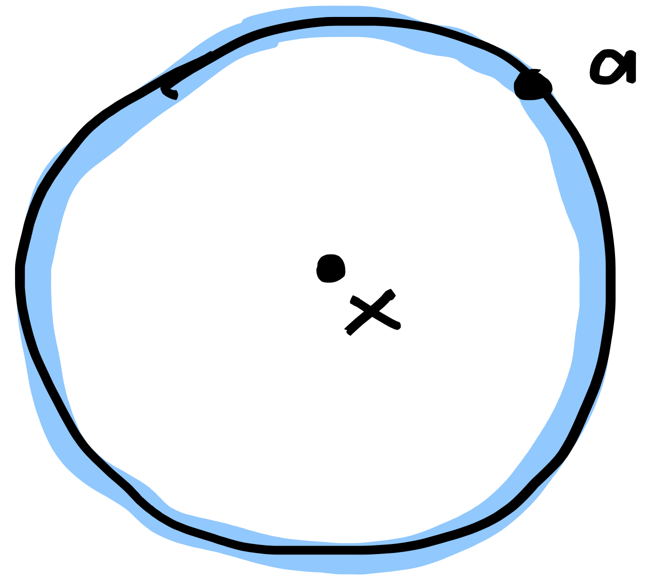
$\| x - a \| < \delta$

$\Rightarrow \sqrt{(x_1 - a_1)^2 + \dots + (x_d - a_d)^2} < \delta$

$\Rightarrow (x_1 - a_1)^2 + \dots + (x_d - a_d)^2 < \delta^2$



$\| x - a \| = \delta$



"Local" to a Point

Definition of an open ball/neighborhood

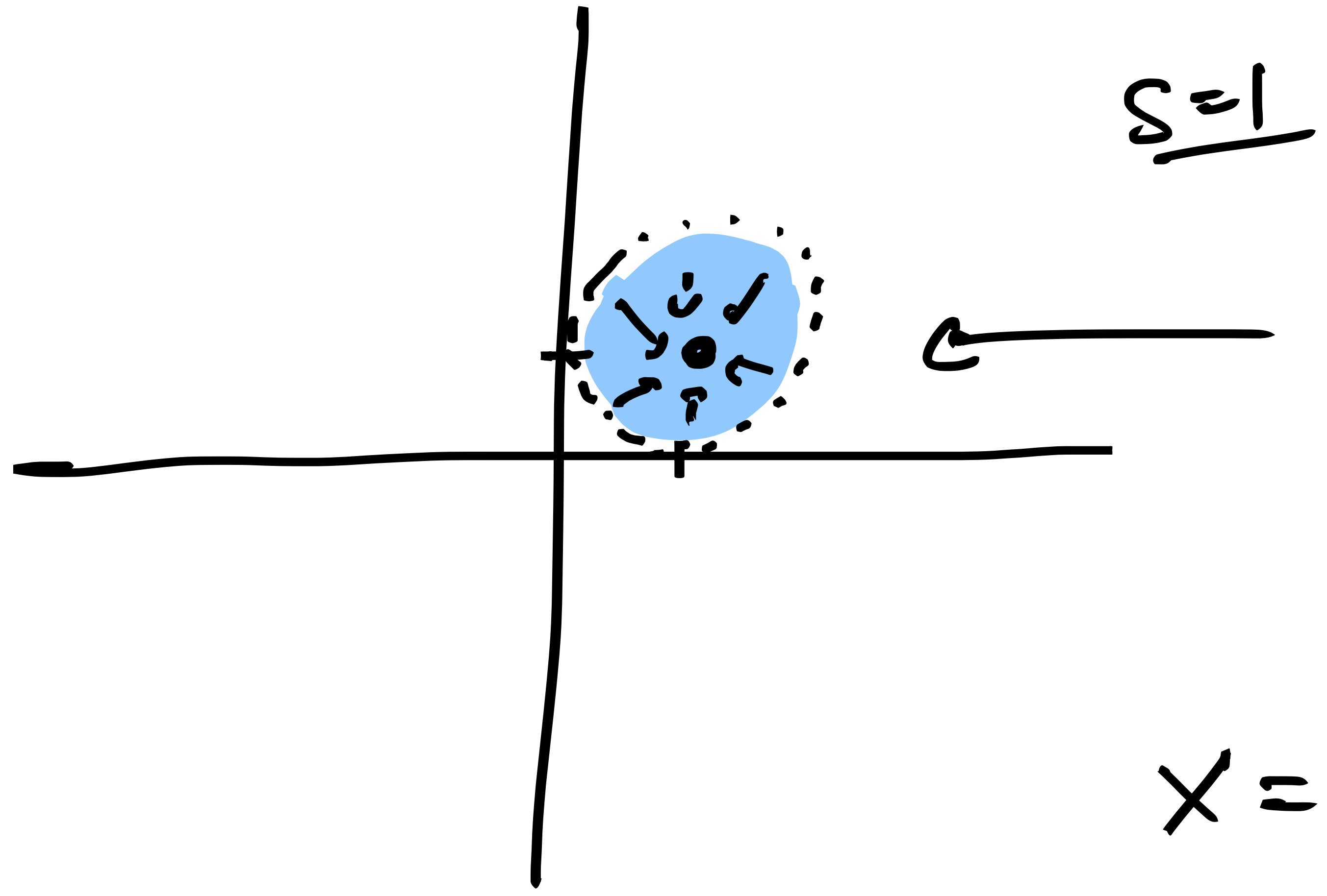
Example. Consider $\mathbf{x} = (1,1) \in \mathbb{R}^2$. What is the open ball of radius $\delta = 1$ around $\mathbf{x}$?

$B_\delta(x) = \{ a \in \mathbb{R}^2 : \|x - a\| < \delta \}$.

$\delta = 1$: $B_1(x) = \{ a \in \mathbb{R}^2 : \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2} < 1 \}$

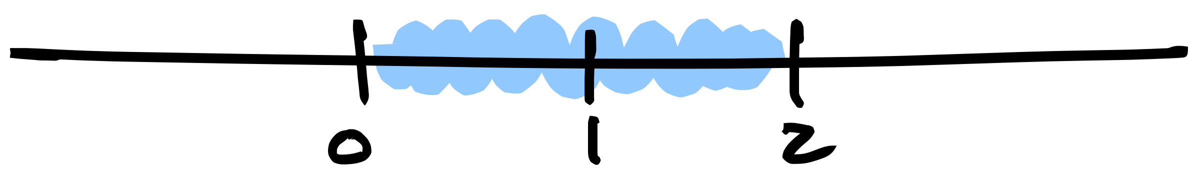
$= \{ a \in \mathbb{R}^2 : (x_1 - a_1)^2 + (x_2 - a_2)^2 < 1 \}$

$= \boxed{\{ a \in \mathbb{R}^2 : (a_1 - 1)^2 + (a_2 - 1)^2 < 1 \}}$.



$x = 1 \in \mathbb{R}$, then neighborhood of radius $\delta = 1$:

$\boxed{(0, 2)}$



“Local” to a Point

Definition of an open ball/neighborhood

Example. Consider $\mathbf{x} = (1,1) \in \mathbb{R}^2$. What is the open ball of radius $\delta = 1$ around $\mathbf{x}$?

An open ball lets us approach $\mathbf{x}$ from all directions.

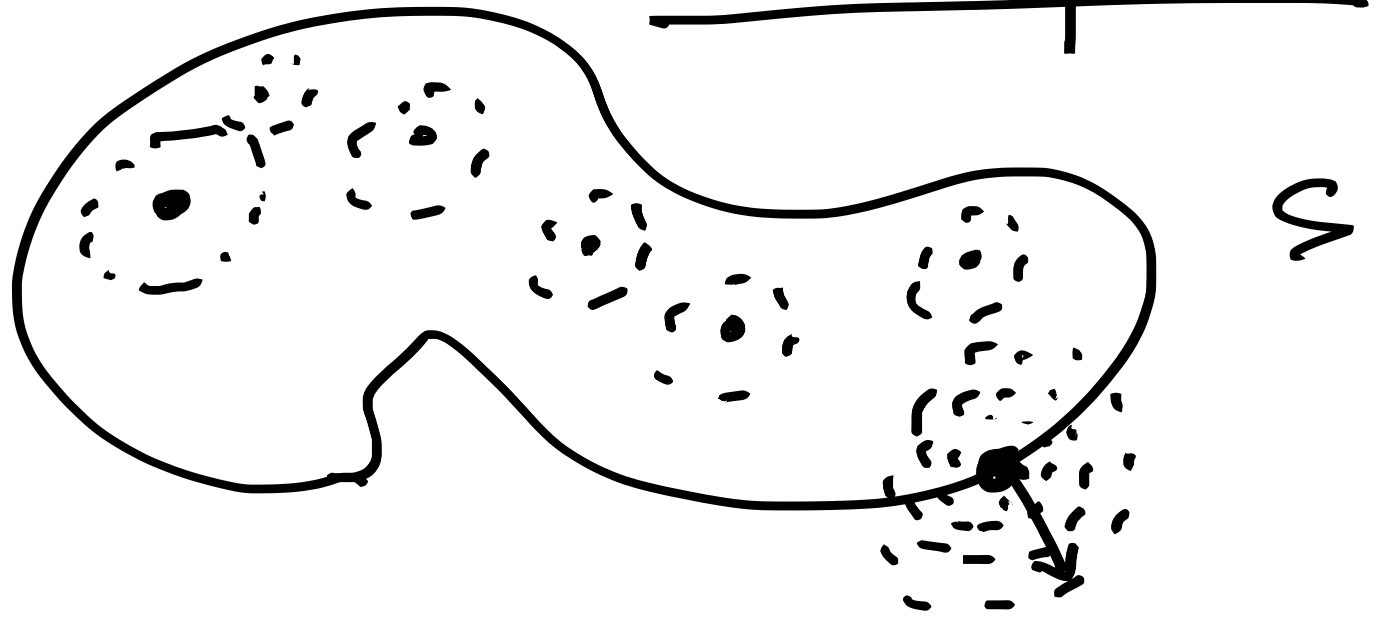
"Local" to a Point

Definition of the interior of a set

$$B_\delta(\mathbf{x}) := \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\| < \delta\}$$

Let $S \subseteq \mathbb{R}^d$ be a set. A point $\mathbf{x} \in S$ is an ***interior point*** if there exists a neighborhood $B_\delta(\mathbf{x})$ around $\mathbf{x}$ such that $B_\delta(\mathbf{x}) \subset S$ (where $\subset$ is *proper subset*).

↳ we can draw an open Ball (doesn't include the border) such that all of the ball is in S.



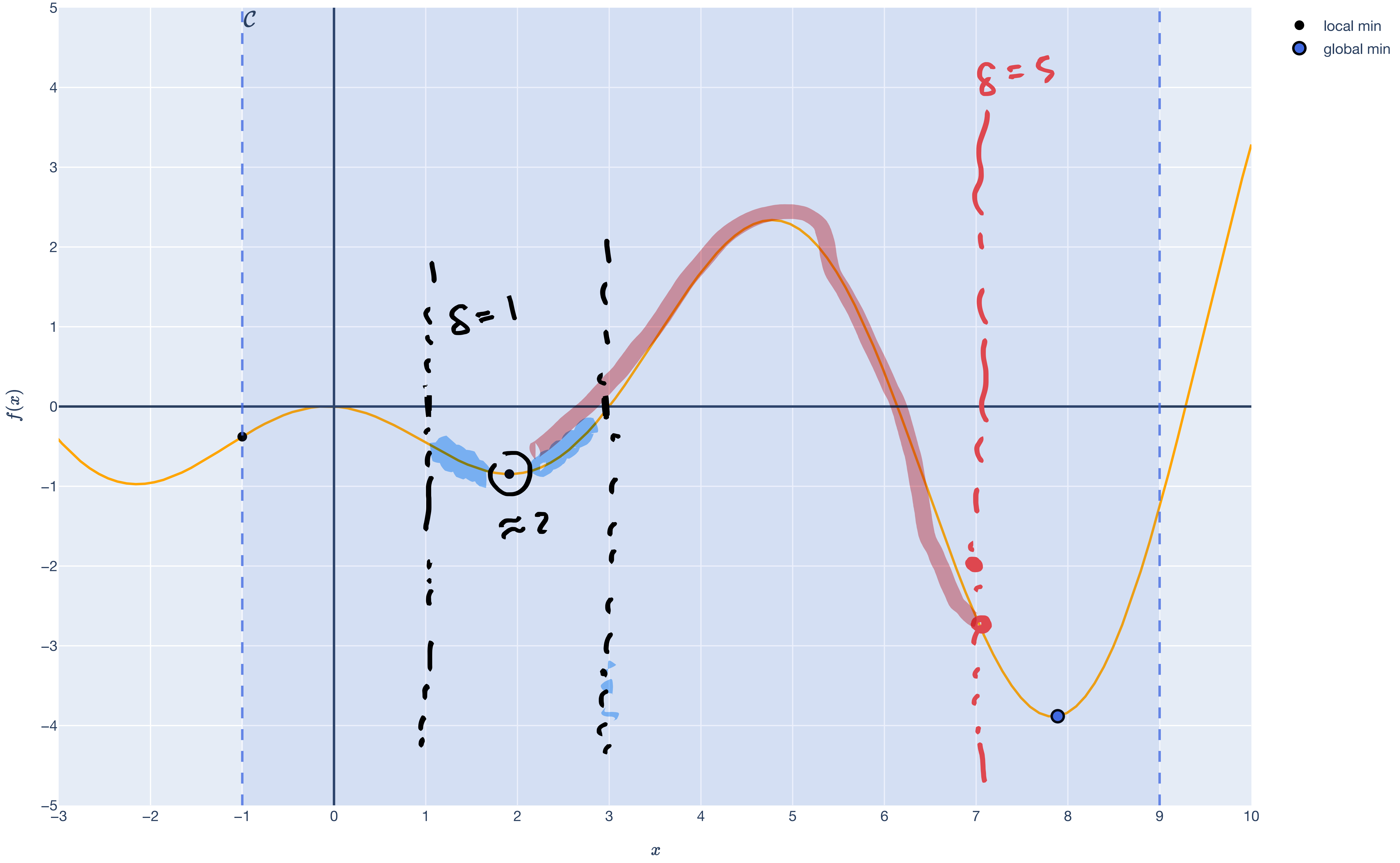
The ***interior of the set*** $\text{int}(S)$ is the set of all interior points of S, i.e.

$$\text{int}(S) := \{\underbrace{\mathbf{x} \in S : B_\delta(\mathbf{x}) \subset}\, S\}\,.$$

"Not on the Boundary."

Types of Minima

Local and global minima



Types of Minima

Local and global minima

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

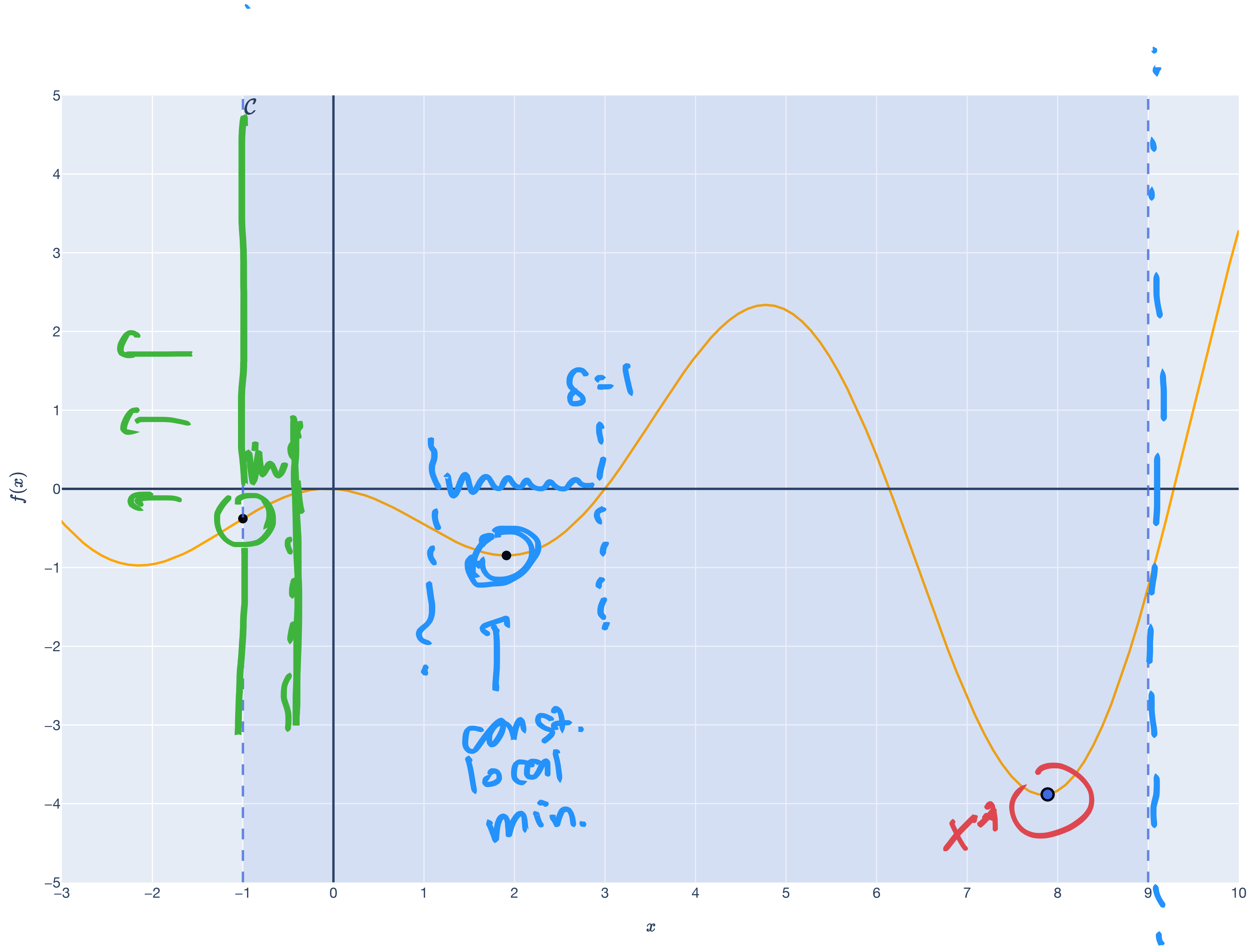
A point $\hat{\mathbf{x}} \in \mathscr{C}$ is a ***local minimum*** if there exists a neighborhood $B_\delta(\hat{\mathbf{x}})$ around $\hat{\mathbf{x}}$ such that

$$\boxed{f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathscr{C} \cap B_\delta(\hat{\mathbf{x}}).}$$

We will also call this a ***constrained local minimum***.

A point $\mathbf{x}^* \in \mathscr{C}$ is a ***global minimum*** if

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathscr{C}.$$



Types of Minima

Local and global minima

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

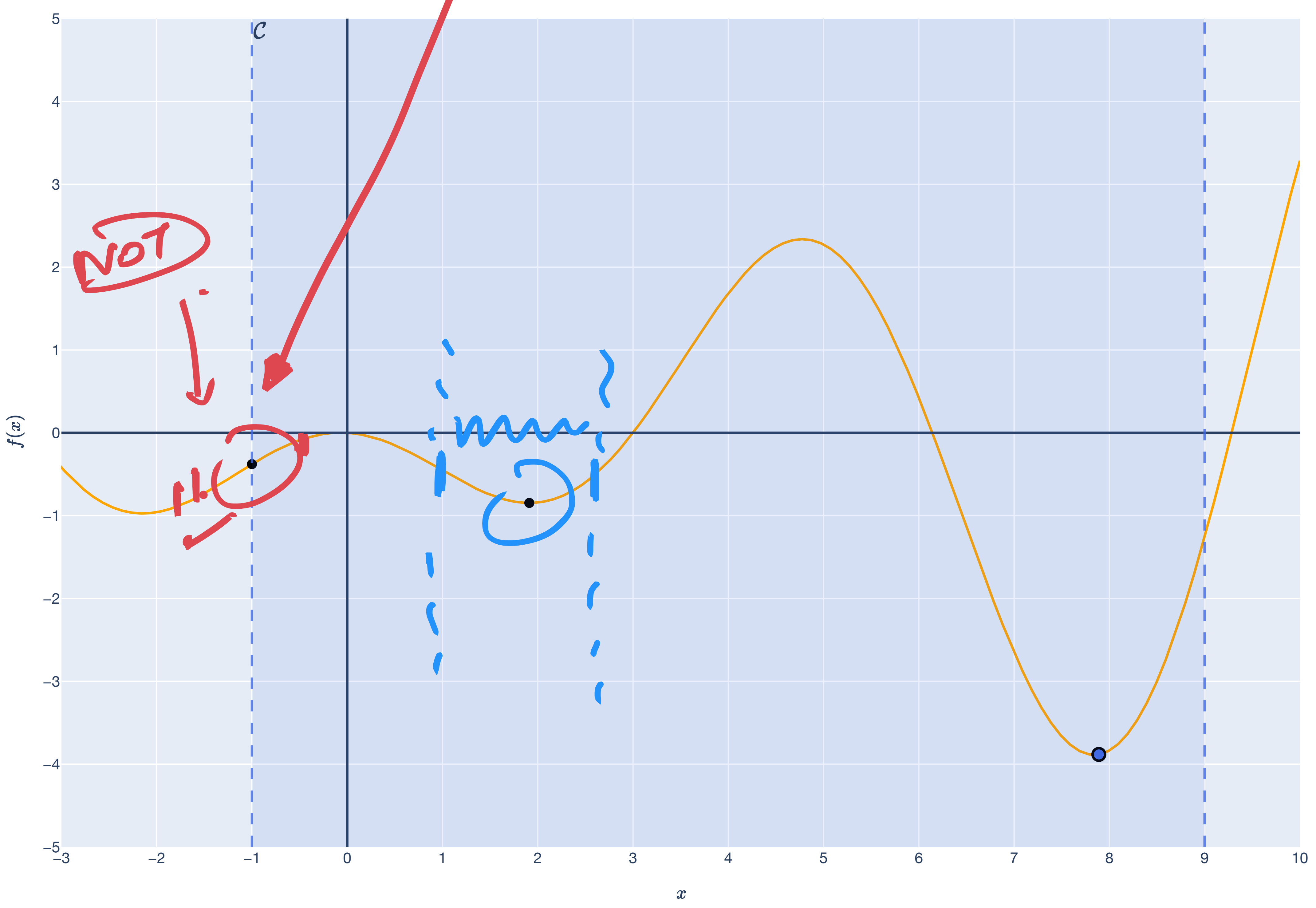
A point $\hat{\mathbf{x}} \in \mathscr{C}$ is an ***unconstrained local minimum*** if there exists a neighborhood $B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}$ around $\hat{\mathbf{x}}$ such that

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in B_\delta(\hat{\mathbf{x}}).$$

PROPERTY

CONSTRAINED local min.

NOT an = unconst. local min.



Types of Minima

Local and global minima

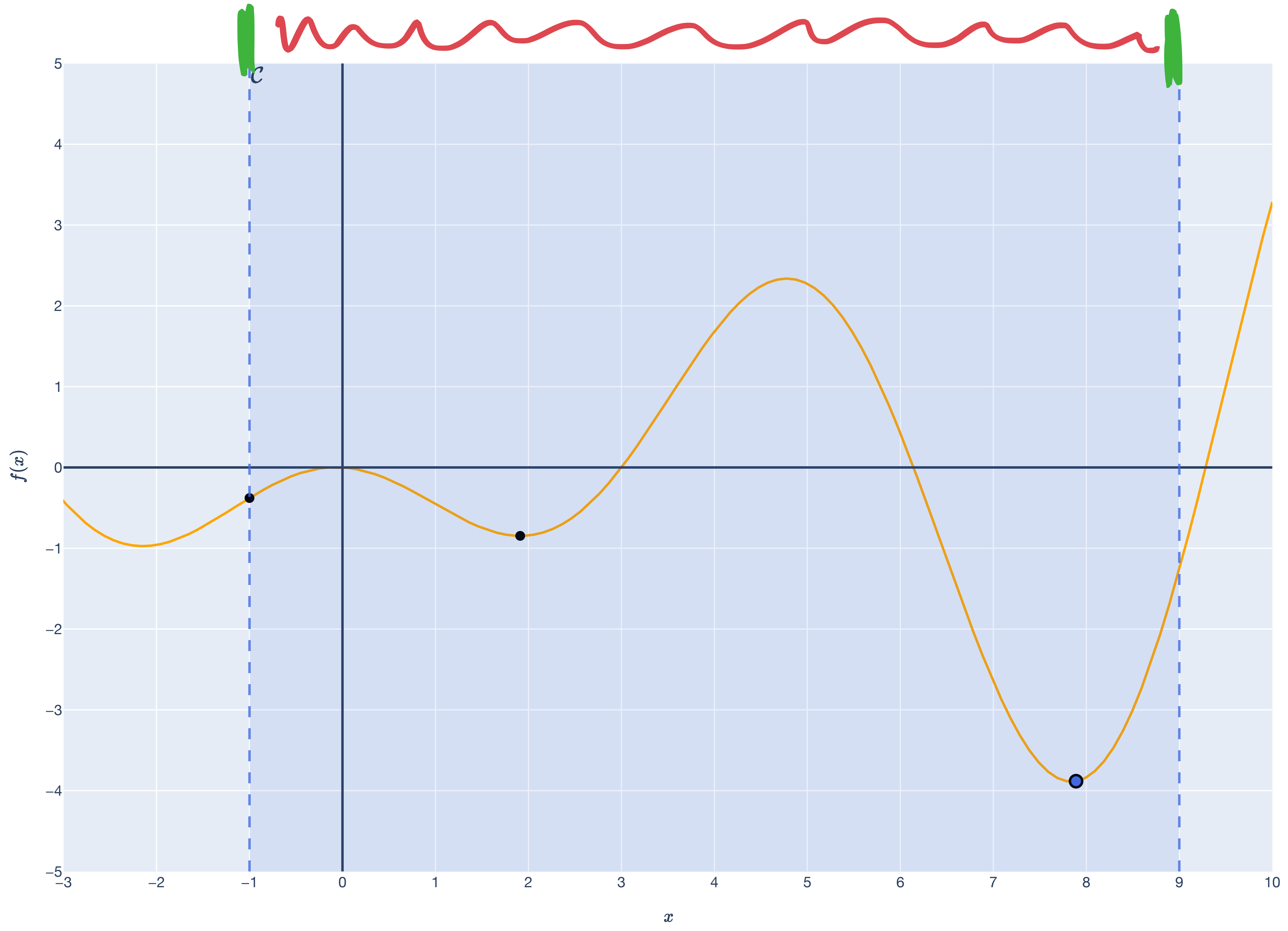
$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathscr{C} \end{array}$$

A point $\hat{\mathbf{x}} \in \mathscr{C}$ is an ***unconstrained local minimum*** if there exists a neighborhood $B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}$ around $\hat{\mathbf{x}}$ such that

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in B_\delta(\hat{\mathbf{x}}).$$

Unconstrained local minima are in the interior $\mathrm{int}(\mathscr{C})$ of the constraint set.

On the other hand, *constrained local minima* can be on the "edge" of the constraint set.



Types of Minima

Which type of minima are each of these points?

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

① ***constrained local:*** ← nearest.

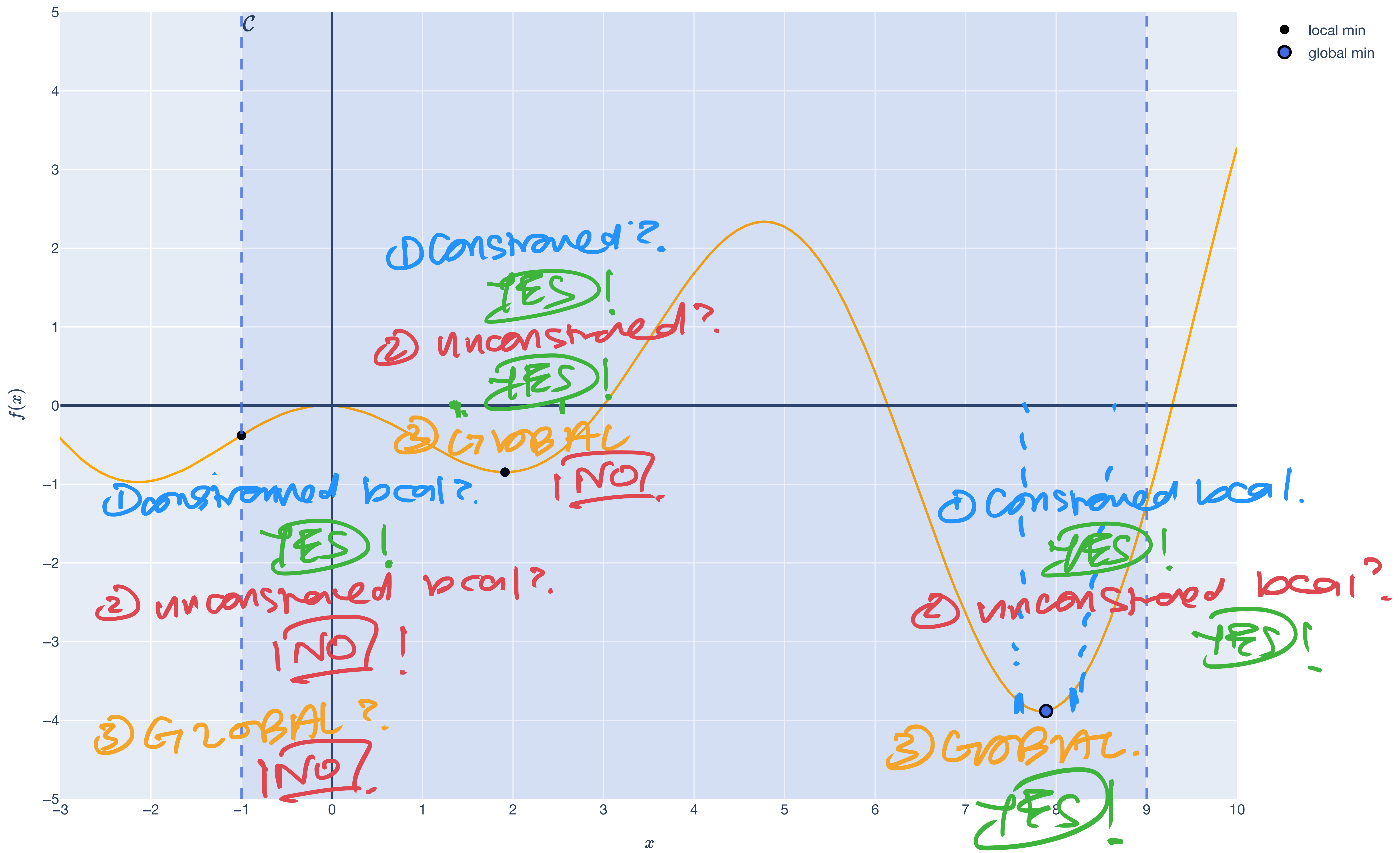
$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathscr{C} \cap B_\delta(\hat{\mathbf{x}})$$

② ***unconstrained local:***

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in B_\delta(\hat{\mathbf{x}}) \text{ and } B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}.$$

global: ← strongest.

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathscr{C}.$$



Types of Minima

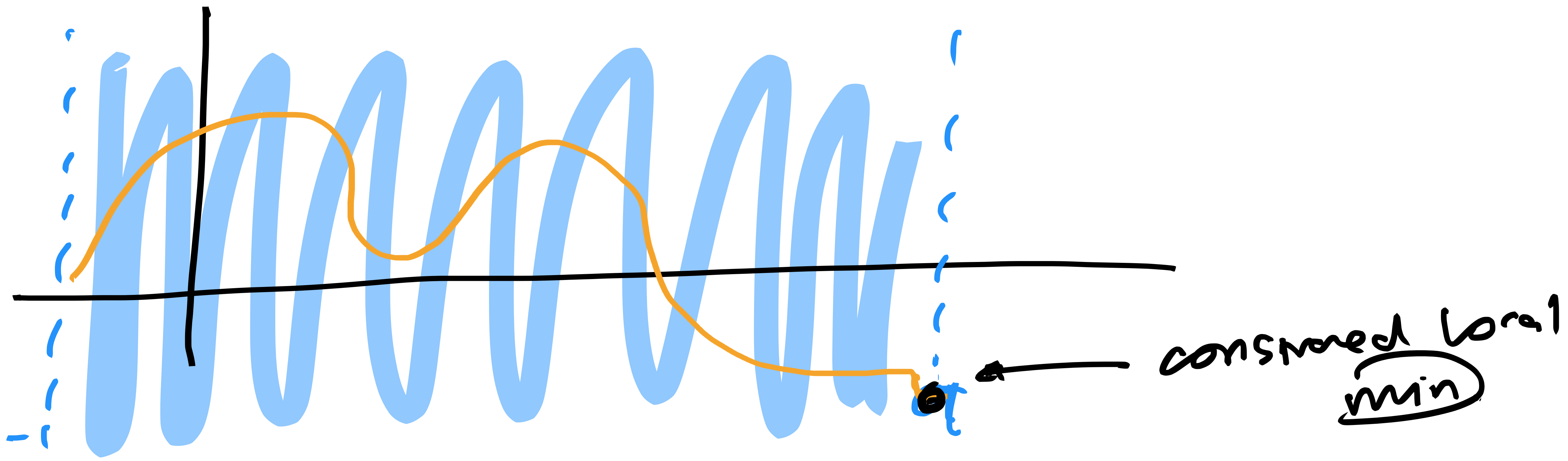
Big picture

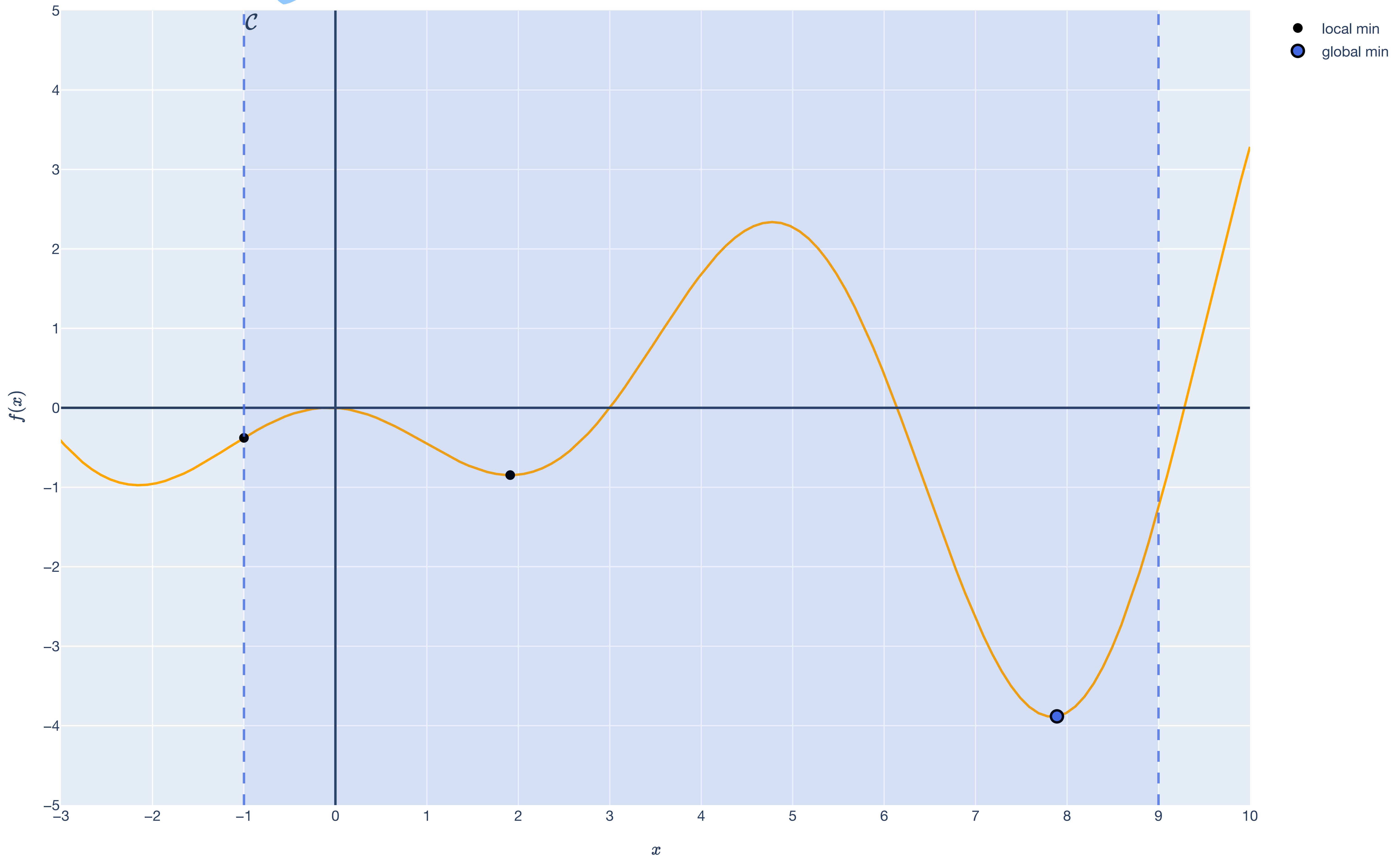
At the end of the day, we want to find ***global minima***.

Global minima could be either ***unconstrained local minima*** or ***constrained local minima***.

Without $\mathscr{C}$, global minima are just one of the *unconstrained local minima.*

With $\mathscr{C}$, global minima may lie on the boundary of the constraint set.





Types of Minima

Big picture

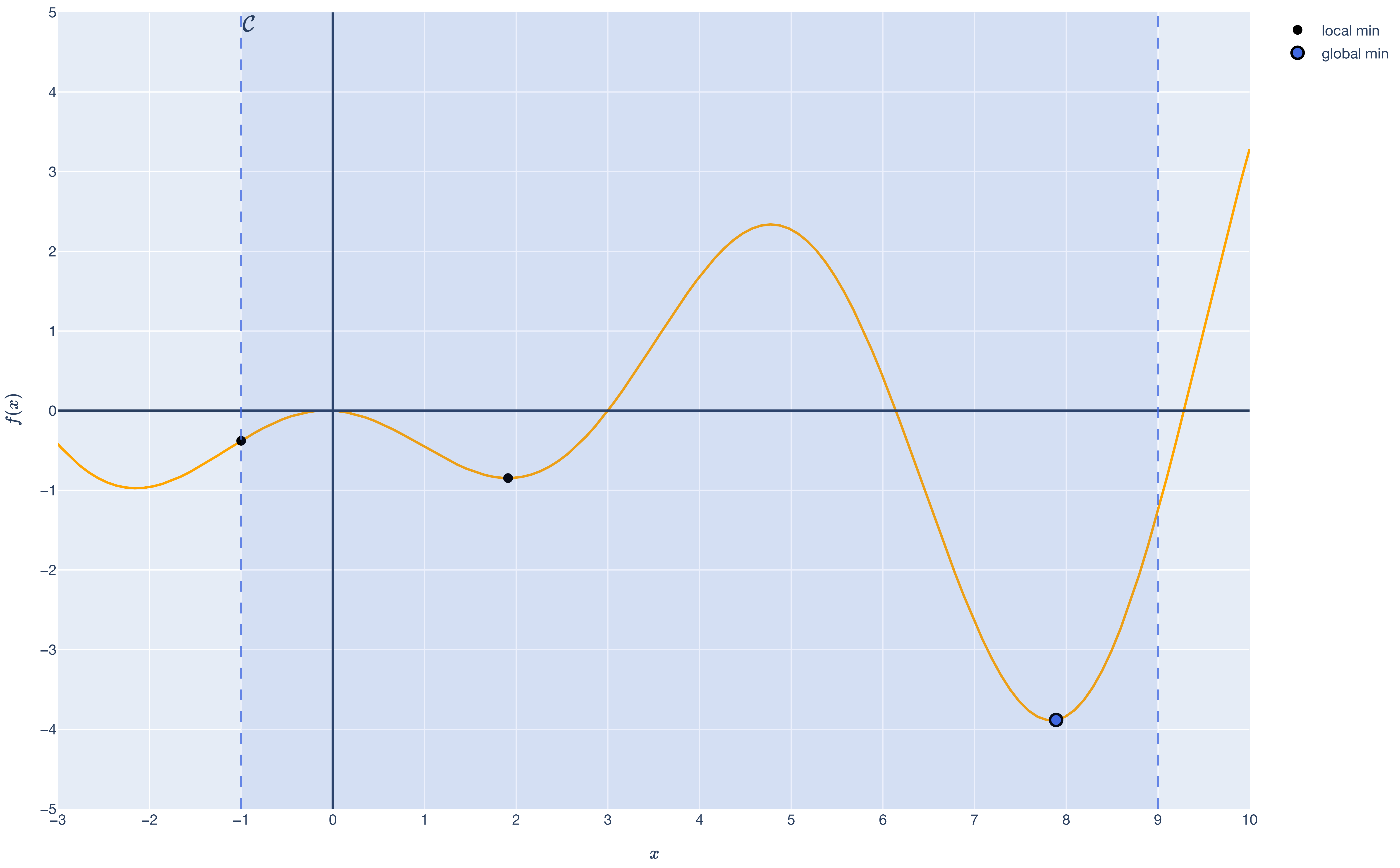
At the end of the day, we want to find ***global minima***.

Global minima could be either ***unconstrained local minima*** or ***constrained local minima***.

> Without $\mathscr{C}$, global minima are just one of the *unconstrained local minima.*

> With $\mathscr{C}$, global minima may lie on the boundary of the constraint set.

Strategy: Find all unconstrained and constrained local minima, then *test* for global minima.



Finding local minima

Big Picture

1) NECESSARY : LOCAL MIN.

2) SUFFICIENT : LOCAL MIN.

Necessary and sufficient conditions

Review

OLS.

$\text{rank}(X) = d \implies (X^T X)^{-1} X^T Y.$

$$P \implies Q$$

Q is ***necessary*** for P. P is ***sufficient*** for Q.

sufficiency: If you assume this, you get your property.

necessity: Your property cannot hold unless you assume this.

Example:

95

A *sufficient* (but not necessary) condition to get an A in this class is to get 100 on every assignment.

A *necessary* (but not sufficient) condition to get an A in this class is to turn in every assignment.

$\Rightarrow$ 33%.

Unconstrained Minima

How do we find unconstrained minima?

multi-variable

A point $\hat{\mathbf{x}} \in \mathscr{C}$ is an ***unconstrained local minimum*** if there exists a neighborhood $B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}$ around $\hat{\mathbf{x}}$ such that

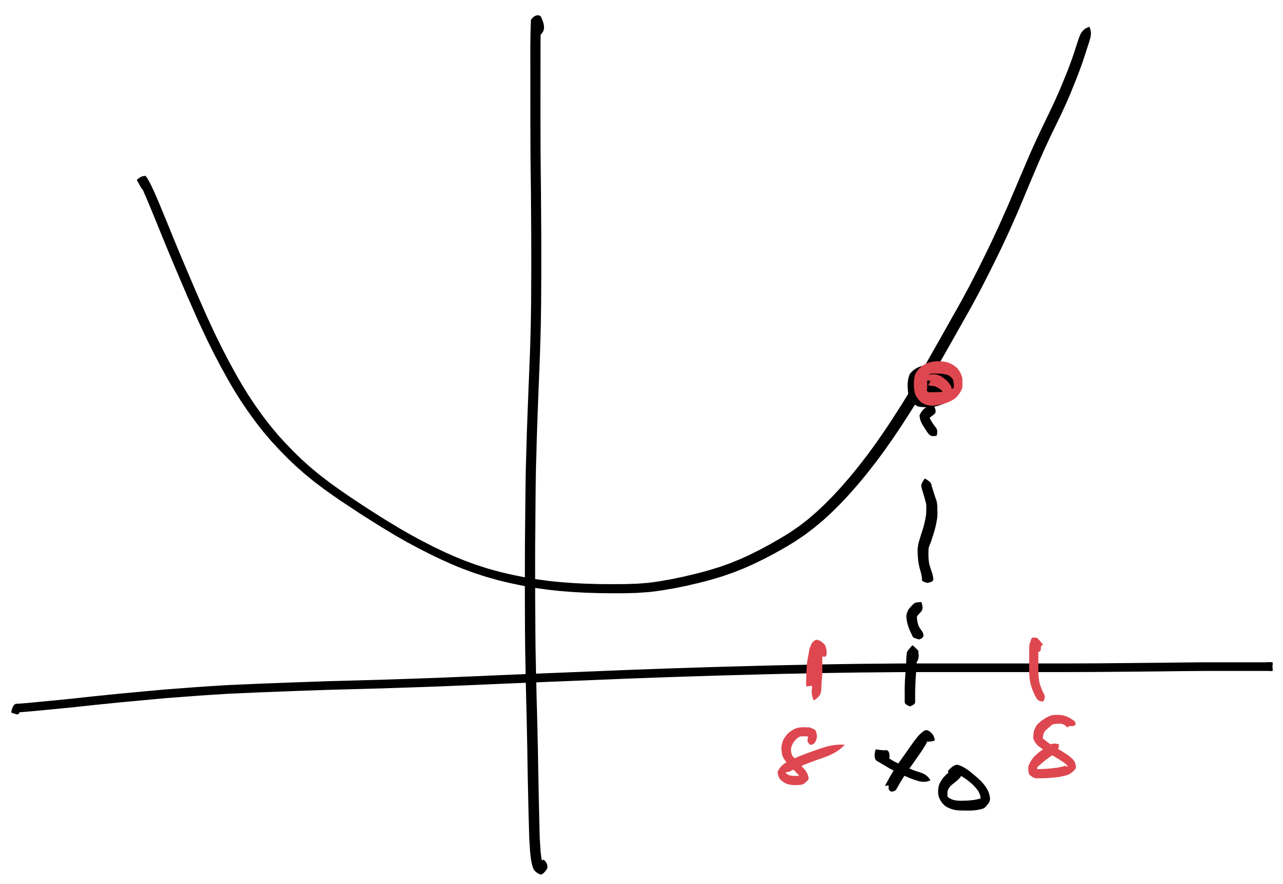
$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in B_\delta(\hat{\mathbf{x}}).$$

From single-variable calculus:

LOCAL MIN. $\Longrightarrow$ $\boxed{f'(x) = 0 \text{ and } f''(x) \geq 0.}$ NECESSARY CONDITIONS

Unconstrained Minima

Intuition from Taylor series



Let $\delta \in \mathbb{R}$ be a scalar increment.

At $x_0 \in \mathbb{R}$, *the second-order Taylor approximation tells us all we need to know:*

MAIN IDEA :

$$f(x_0 + \delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2}f''(x_0)\delta^2.$$

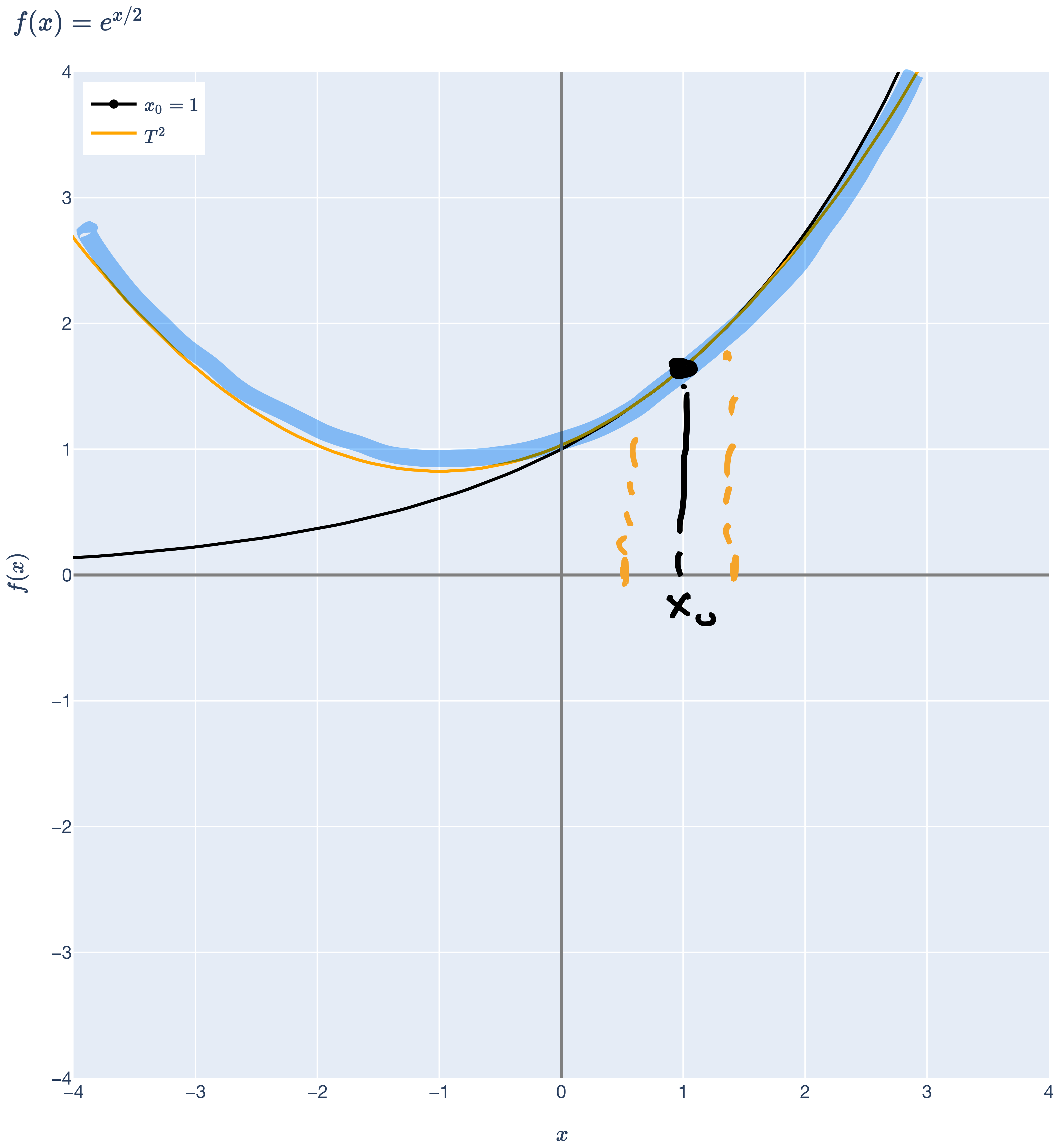
Second-order Taylor Approximation

Single-variable example

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

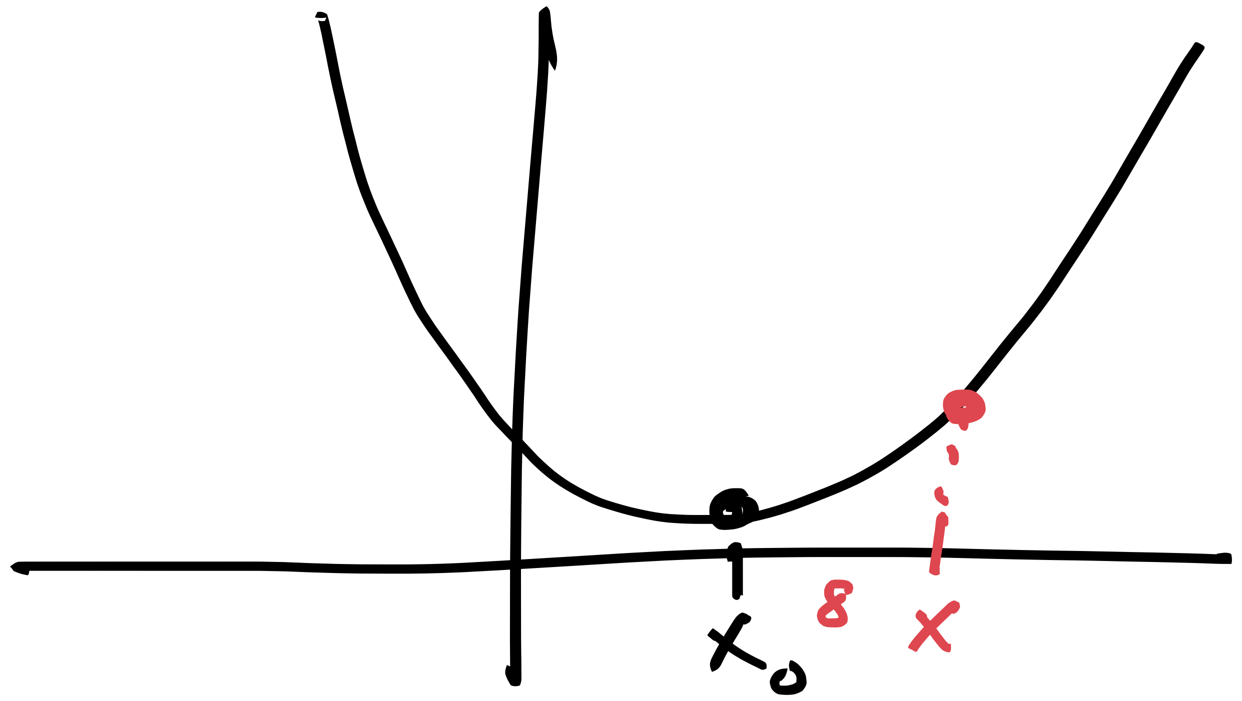
Second-order Taylor expansion at $x_0 = 1$:

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



Unconstrained Minima

Intuition from Taylor series



Let $\delta \in \mathbb{R}$ be a scalar increment.

At $x_0 \in \mathbb{R}$, *the second-order Taylor approximation tells us all we need to know:*

$> 0 \quad < 0 \quad < 0 \quad > 0$

$$f(x_0 + \delta) \approx f(x_0) + \underbrace{f'(x_0)\delta}_{0} + \underbrace{\frac{1}{2} f''(x_0)\delta^2}_{\geq 0,\ > 0}.$$

$f(x_0 + \delta) \leq f(x)$

Pretend that this function approximation is exact. Then…

$f(x_0 + \delta) \approx f(x_0) + \frac{1}{2} f''(x_0)\delta^2$

> 0

What are the *necessary* conditions for x to be a minimum?

$= x_0 + \delta = x.$

$f(x_0 + \delta) < f(x_0) + \frac{1}{2} f''(x_0)\delta^2$

What are the *sufficient* conditions for x to be a minimum?

Unconstrained Minima

Intuition from Taylor series

Let $\delta \in \mathbb{R}$ be a scalar increment.

At $x_0 \in \mathbb{R}$, *the second-order Taylor approximation tells us all we need to know:*

$$f(x_0 + \delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2}f''(x_0)\delta^2.$$

Pretend that this function approximation is exact. Then…

What are the *necessary* conditions for x to be a minimum? $f'(x) = 0, f''(x) \geq 0.$

What are the *sufficient* conditions for x to be a minimum? $f'(x) = 0, f''(x) > 0.$

Unconstrained Minima

Sufficient conditions met

$$\boxed{f(x_0+\delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2}f''(x_0)\delta^2}$$

Necessary conditions: $f'(x_0) = 0, f''(x_0) \geq 0.$

Sufficient conditions: $f'(x_0) = 0, f''(x_0) > 0.$

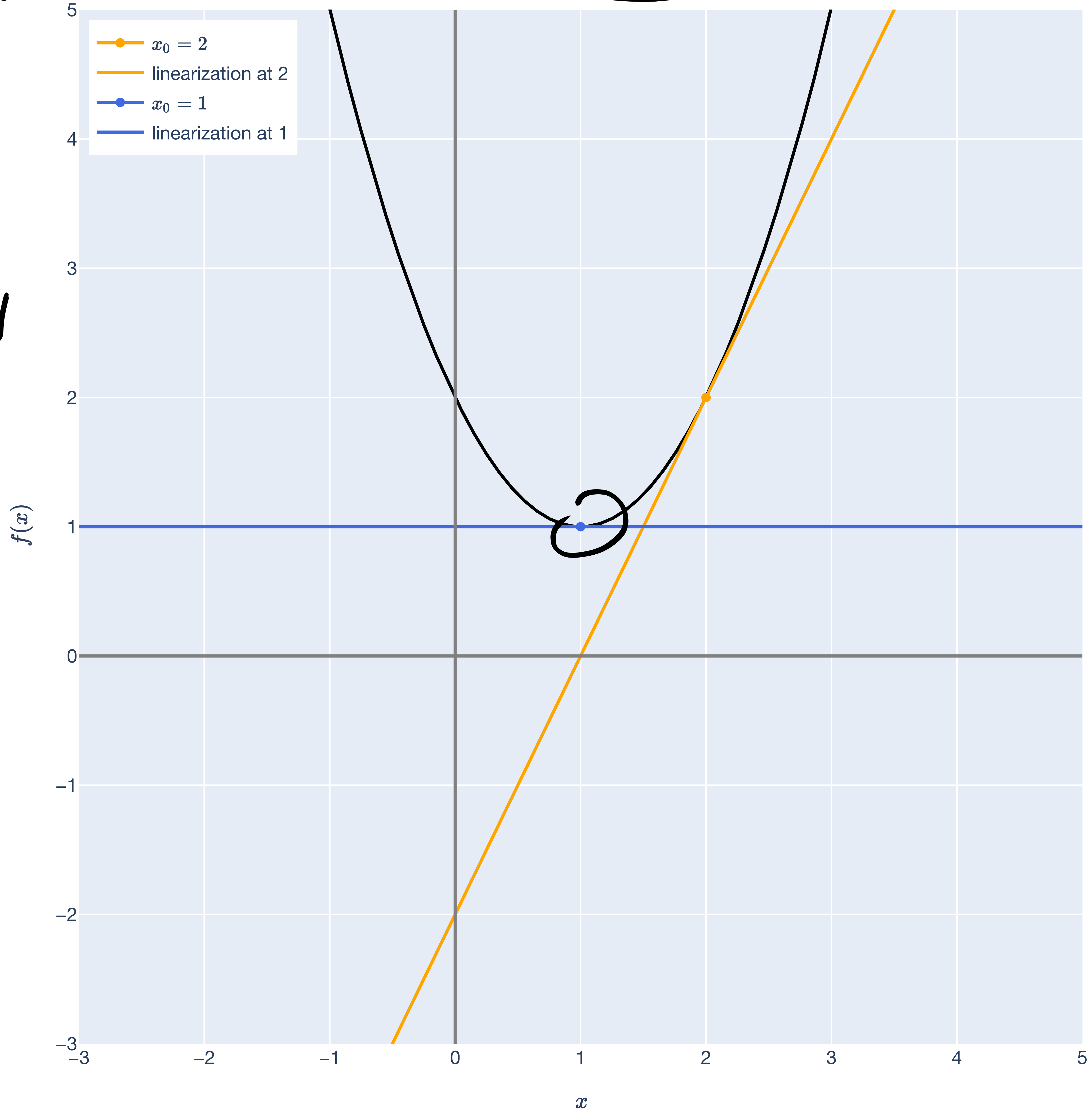
Candidate: $x^* = 1$

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

$f'(x) = 2(x-1) \Rightarrow \boxed{f'(1) = 0.}$

$f''(x) = 2. > 0$

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



Local Min $\Rightarrow$ $f'(x_0) = 0$
$f''(x_0) \geq 0$.

Unconstrained Minima

Necessary, not sufficient

$f'(x_0) = 0$, $f''(x_0) = 0$ $\cancel{\Rightarrow}$ Local Min.

$f'(x_0) = 0$, $f''(x_0) > 0$ ✓ $\Rightarrow$ Local Min.

$$f(x_0 + \delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2} f''(x_0)\delta^2$$

Necessary conditions: $f'(x_0) = 0, f''(x_0) \geq 0.$

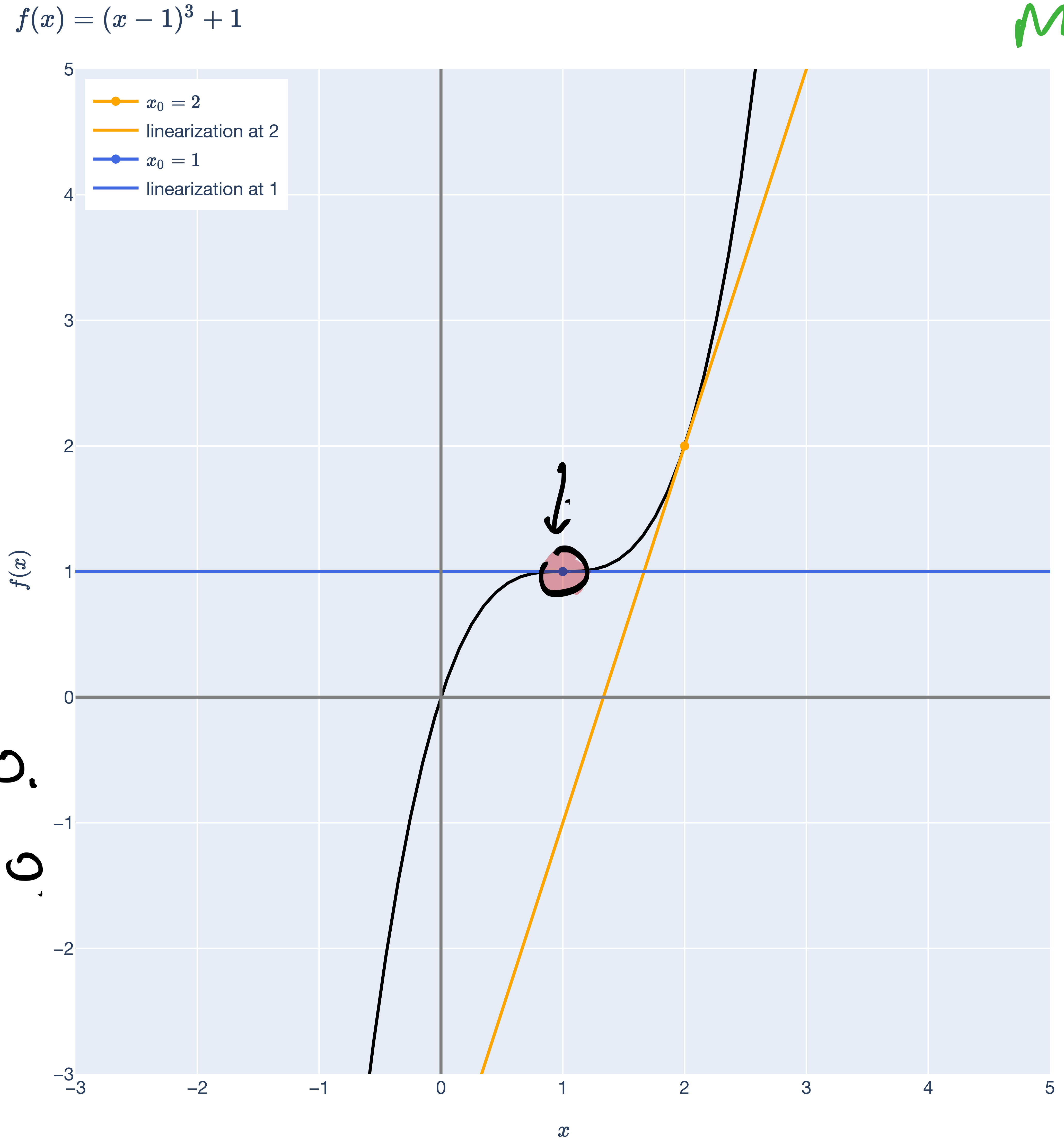
Sufficient conditions: $f'(x_0) = 0, f''(x_0) > 0.$

$x_0 = 1$

$f'(x) = 3(x-1)^2 \Rightarrow f'(x) = 3(1-1)^2 = 0.$

$f''(x) = 6(x-1) \Rightarrow f''(x) = 6(1-1) = 0.$

$\boxed{f''(1) = 0}$



Remainder of Taylor Polynomial

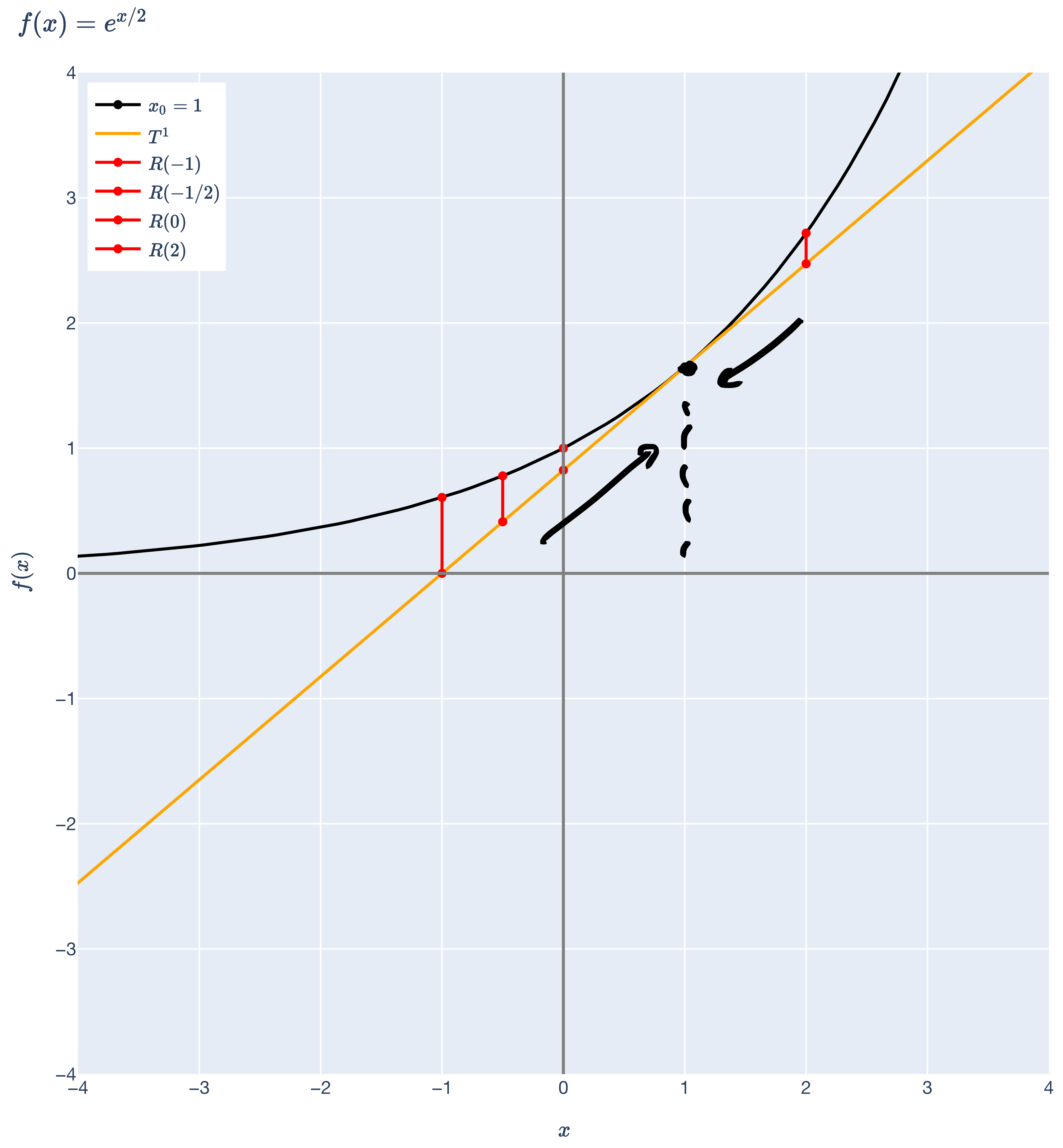
Definition

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

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

Error from chopping off..

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

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

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

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 1: Peano's Form Taylor's Theorem

What does $R^2(\mathbf{x}_0 + \mathbf{d}) = o(\|\mathbf{d}\|^2)$ mean?

For every $C > 0$, there exists a neighborhood $B_\delta(\mathbf{0})$ such that

$$\frac{1}{2} \quad \frac{1}{8} \quad \frac{1}{16}. \qquad R^2(\mathbf{x}_0 + \mathbf{d}) \le C\|\mathbf{d}\|^2, \quad \forall \mathbf{d} \in B_\delta(\mathbf{0}) .$$

δ small enough.

We can make the remainder term as *small as we like* as long as $\|\mathbf{d}\|$ is sufficiently small ($\|\mathbf{d}\| < \delta$ does the trick).

$$\boxed{\frac{R^2(x_0+d)}{\|d\|^2} \le \frac{1}{2}.}$$

Taylor's Theorem

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

What does $R^2(\mathbf{x}_0 + \mathbf{d}) = o(||\mathbf{d}||^2)$ mean?

Let $\mathbf{d} \in \mathbb{R}^d$ be a unit vector with $||\mathbf{d}|| = 1$ and $\alpha > 0$ be a scalar, so:

only direction

$$o(||\alpha\mathbf{d}||^2) = o(\alpha^2).$$

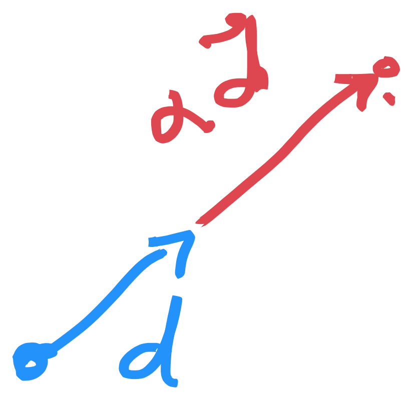
$||d|| = 1.$

$||\alpha d||^2 = \alpha^2 ||d||^2 = \alpha^2.$

Then, $R^2(\mathbf{x}_0 + \alpha\mathbf{d}) = o(\alpha^2)$ means:

$$\boxed{\lim_{\alpha \to 0} \frac{R^2(\mathbf{x}_0 + \alpha\mathbf{d})}{\alpha^2} = 0}$$

(the remainder goes to 0 *faster* than a quadratic).



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 $\mathbf{x}_0$. Let $\mathbf{d} \in \mathbb{R}^d$ be any direction. For every $C > 0$, there exists a neighborhood $B_\delta(\mathbf{0})$ such that

& However small we want.

$$\left| \underbrace{f(\mathbf{x}_0 + \mathbf{d}) - \left(f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}_0) \mathbf{d} \right)}_{R^2.} \right| \leq C \|\mathbf{d}\|^2$$

for all $\mathbf{d} \in B_\delta(\mathbf{0})$.

Unconstrained local minima
Necessary conditions

Least Squares

OLS Theorem

$w^T X^T X w - 2w^T(X^T y)$

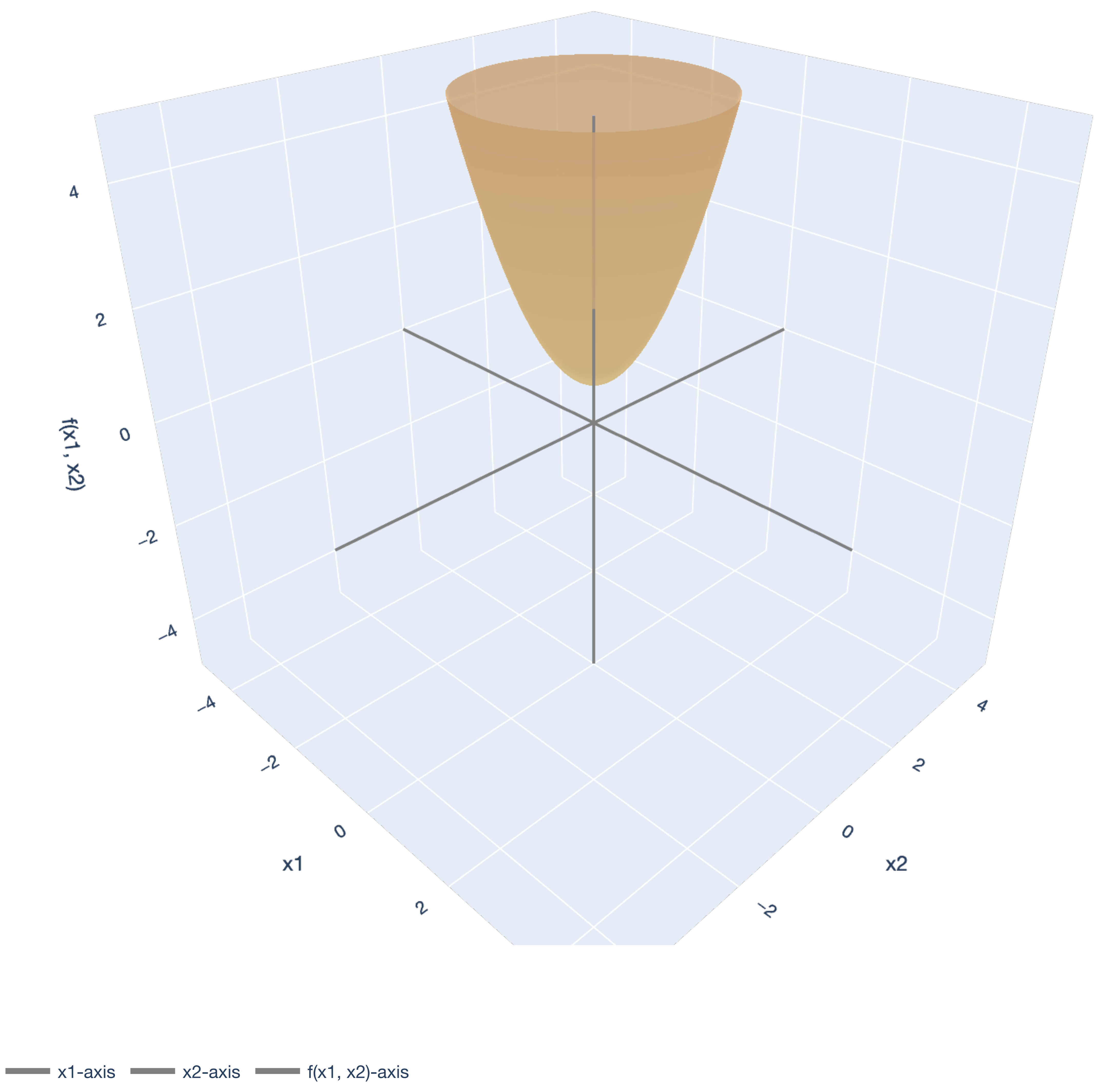
$\downarrow \qquad + y^T y$

Proof (OLS).

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}.$$

Set it equal to $\mathbf{0}$.



Least Squares

OLS Theorem

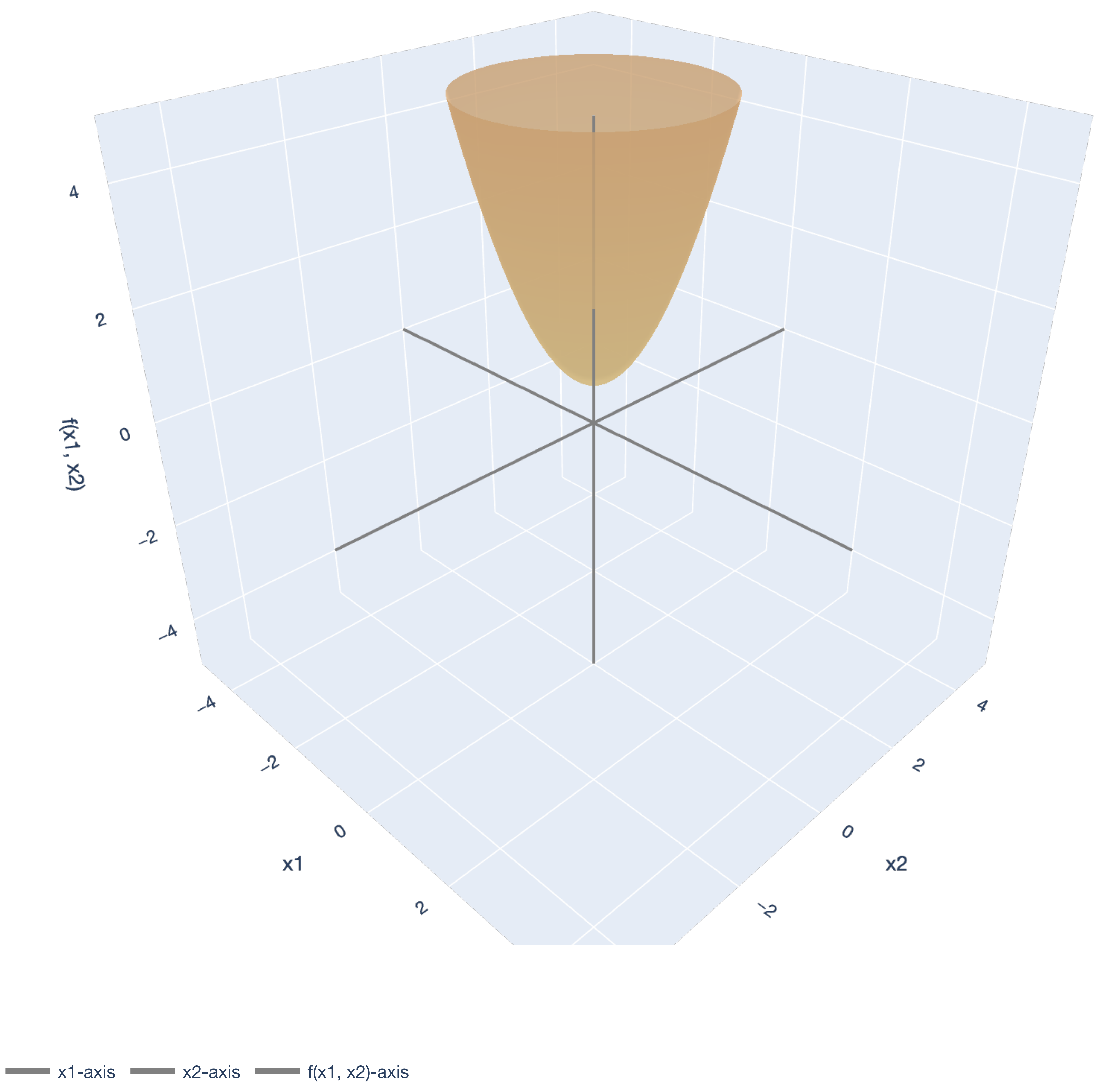
Proof (OLS).

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}.$$

Set it equal to $\mathbf{0}$.

Why is this the right thing to do?



Taylor's Theorem

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

For all intents and purposes,

For $f(x_0)$ to be a min.?

MAIN IDEAS

$$\underbrace{f(x_0+\delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2}f''(x_0)\delta^2}$$ when δ is small enough.

is analogous to:

$$\underbrace{f(\mathbf{x}_0+\mathbf{d}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{d}}$$ when $||\mathbf{d}||$ is small enough.

Unconstrained Minima

Necessary conditions

Hessian (symmetric)

$$f(x_0+\delta) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2} f''(x_0)\delta^2$$

when δ is small enough.

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

when $\|\mathbf{d}\|$ is small enough.

- x_0 to be a minimum.

$\Rightarrow f(x_0) \le f(x_0+\delta)$ for $\delta \in \mathbb{R}$.

$x_0 \quad x_0+\delta$

$f(x_0) + f'(x_0)\delta + \underbrace{\frac{1}{2} f''(x_0)\delta^2}_{\ge 0}$

$\cancel{f(x_0)} \le \cancel{f(x_0)} + f'(x_0)\delta + \frac{1}{2} f''(x_0)\delta^2$

Necessary conditions: $0 \le f'(x_0)\delta + \frac{1}{2} f''(x_0)\delta^2$ Necessary conditions:

" ≥ 0 "

$\boxed{x^T A x \ge 0}$

$>0 \quad >0$

$<0 \quad <0$

$=0$

$=0$

<0

<0

$f'(x_0) = 0$, $f''(x_0) \ge 0$.

$\nabla f(\mathbf{x}_0) = \mathbf{0}$, $\nabla^2 f(\mathbf{x}_0)$ is PSD.

$=$ $\quad \sim \quad \ge 0$

Total Derivative

Review of definition

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

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

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

$\nabla f(x) = 0$ $\quad$ $\nabla^2 f(x)$ is PSD $\quad$ PSD

$\boxed{0 \preceq A}$

Unconstrained Minima

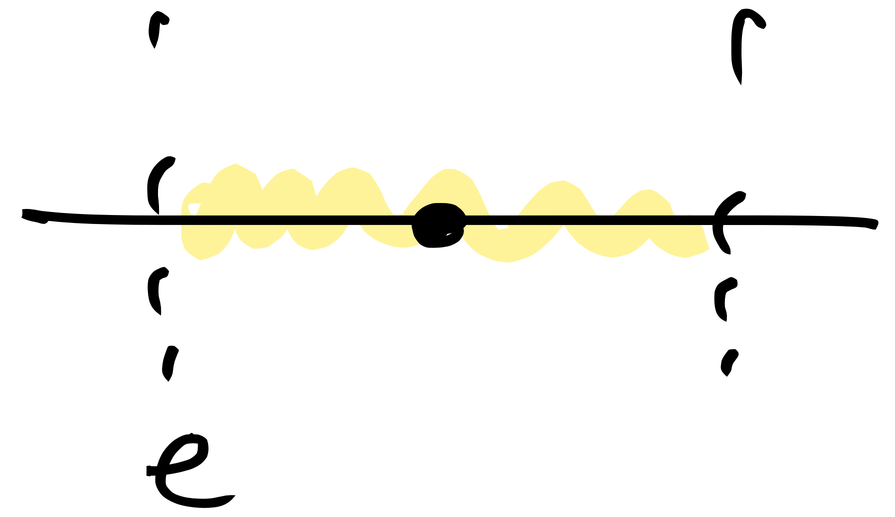
Necessary conditions

$\Leftarrow$ $f'(x) = 0$ $\quad$ $f''(x) \geq 0$.

Theorem (Necessary Conditions for Unconstrained Local Minimum). Consider the optimization problem

Doesn't have anything e.

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



Suppose $\mathbf{x}^* \in \text{int}(\mathscr{C})$ is an *unconstrained local minimum.* Then,

First-order condition. If f is differentiable at $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$. $\Leftrightarrow f'(x)$

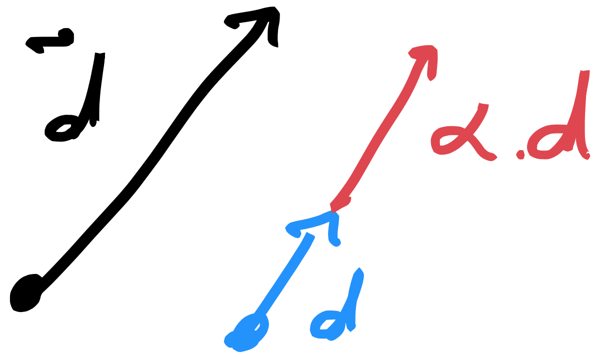
Second-order condition. If f is twice-differentiable at $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is positive semidefinite, i.e. $\mathbf{v}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^d$.

$f''(x) \geq 0$.

Proof of necessary conditions

First order condition

local min $\Longrightarrow \nabla f(x^*) = 0$.



First-order condition. If f is differentiable at $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Step 1: Use definition of the gradient for $\alpha\mathbf{d}$.

Choose an arbitrary direction $\alpha\mathbf{d} \in \mathbb{R}^d$, where $\|\mathbf{d}\| = 1$ is a unit vector and $\alpha > 0$ is a scalar.

f is differentiable, so…

For any direction.

$$\boxed{\lim_{\alpha \to 0} \frac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) - \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d}}{\alpha\|\cancel{\mathbf{d}}\|} = 0}$$

$\|d\| = 1$.

which is the same as stating:

$$\boxed{\lim_{\alpha \to 0} \frac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*)}{\alpha} = \nabla f(\mathbf{x}^*)^\top \mathbf{d}.}$$

Proof of necessary conditions

First order condition

LOCAL MIN. $\Rightarrow \nabla f(x^*) = 0$.

First-order condition. If f is differentiable at $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Step 2: Use local optimality on difference $f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*)$.

From Step 1,

≥ 0

$$\lim_{\alpha \to 0} \frac{f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*)}{\alpha} = \nabla f(\mathbf{x}^*)^\top \mathbf{d}.$$

∅

$\mathbf{x}^*$ is an *unconstrained local minimum*, so there exists a neighborhood $B_\delta(\mathbf{x}^*)$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in B_\delta(\mathbf{x}^*)$. So if $\alpha < \delta$ (sufficiently small),

$$\boxed{f(\mathbf{x}^* + \alpha \mathbf{d}) \geq f(\mathbf{x}^*)} \Longrightarrow \boxed{\nabla f(\mathbf{x}^*)^\top \mathbf{d} \geq 0.}$$

Proof of necessary conditions

First order condition

$[1\ 0\ 0\ 0] \begin{bmatrix} \nabla f(x^*)_1 \\ \vdots \\ \nabla f(x^*)_d \end{bmatrix} = \nabla f(x^*)_1$

First-order condition. If f is differentiable at $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

$e_1^\top \nabla f(x^*) =$

Step 3: Conclude by recalling that $\mathbf{d} \in \mathbb{R}^d$ was an arbitrary direction.

From Step 2, if $\alpha < \delta$ (sufficiently small), $\nabla f(\mathbf{x}^*)^\top \mathbf{d} \geq 0$.

$\nabla f(x^*)_1 \geq 0$

But $\mathbf{d} \in \mathbb{R}^d$ was an arbitrary direction with $\|\mathbf{d}\| = 1$.

TRICK!

$$\mathbf{d} = \mathbf{e}_1 \implies \nabla f(\mathbf{x}^*)_1 \geq 0 \text{ and } \mathbf{d} = -\mathbf{e}_1 \implies \nabla f(\mathbf{x}^*)_1 \leq 0$$

$\Rightarrow \nabla f(x^*)_1 = 0$

$$\mathbf{d} = \mathbf{e}_2 \implies \nabla f(\mathbf{x}^*)_2 \geq 0 \text{ and } \mathbf{d} = -\mathbf{e}_2 \implies \nabla f(\mathbf{x}^*)_2 \leq 0$$

$$\vdots$$

$$\mathbf{d} = \mathbf{e}_d \implies \nabla f(\mathbf{x}^*)_d \geq 0 \text{ and } \mathbf{d} = -\mathbf{e}_d \implies \nabla f(\mathbf{x}^*)_d \leq 0$$

Therefore, $\nabla f(\mathbf{x}^*) = \mathbf{0}$. ✓

Proof of necessary conditions

Second order condition

Second-order condition. If f is twice-differentiable at $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is PSD.

Step 1: Use second-order Taylor's theorem with $\alpha\mathbf{d} \in \mathbb{R}^d$ with $\|\mathbf{d}\| = 1$.

Choose an arbitrary direction $\alpha\mathbf{d} \in \mathbb{R}^d$, where $\|\mathbf{d}\| = 1$ is a unit vector and $\alpha > 0$ is a scalar. By Taylor's Theorem (Peano's form):

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) = \nabla f(\mathbf{x}^*)^\top(\alpha\mathbf{d}) + \frac{1}{2}(\alpha\mathbf{d})^\top \nabla^2 f(\mathbf{x}^*)(\alpha\mathbf{d}) + o(\|\alpha\mathbf{d}\|^2)$$

$$= \alpha\nabla f(\mathbf{x}^*)^\top\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2)$$

$\|d\| = 1$

$o(\|\alpha d\|^2)$

$= o(\alpha^2 \|d\|^2) = o(\alpha^2)$

Proof of necessary conditions

Second order condition

$\Rightarrow \nabla f(x^*) = 0.$

LOCAL MIN $\Rightarrow \nabla^2 f(x^*)$ is PSD

Second-order condition. If f is twice-differentiable at $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is PSD.

Step 2: Use first-order condition on difference $f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*)$.

From Step 1,

$$f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*) = \overset{=0}{\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d}} + \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} + o(\alpha^2)$$

$\mathbf{x}^*$ is an *unconstrained local minimum,* so by first-order condition (just proved):

$$f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*) = \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} + o(\alpha^2)$$

Proof of necessary conditions

Second order condition

Second-order condition. If f is twice-differentiable at $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is PSD.

Step 3: Take $\alpha \to 0$ to get rid of the little-oh terms.

From Step 3,

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) = \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2).$$

$$\lim_{\alpha \to 0} \frac{o(\alpha^2)}{\alpha^2} \to 0$$

Recall that if $g = o(h)$, then $\displaystyle\lim_{\alpha\to 0}\frac{g(\alpha)}{h(\alpha)} = 0$.

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) - \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} = o(\alpha^2) \Longrightarrow \lim_{\alpha\to 0}\frac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*)}{\alpha^2} - \frac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} = 0.$$

By local optimality of $\mathbf{x}^*$,

$$\Downarrow.$$

$$\boxed{f(x^* + \alpha d) \geq f(x^*)}$$

$0 \leq \dfrac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*)}{\alpha^2}$, so $0 \leq \dfrac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d}$. By definition, $\nabla^2 f(\mathbf{x}^*)$ is PSD.

For any d.

Least Squares

OLS Theorem

$f''(x) \geq 0$. $f''(x) > 0 \implies$ local min.

local min $\implies$ $\boxed{\nabla f(x^*) = 0. \quad \nabla^2 f(x^*) \text{ is PSD}}$

Proof (OLS).

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \iff f(\mathbf{w}) = \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}$$

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}.$$

Set it equal to $\mathbf{0}$.

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

$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

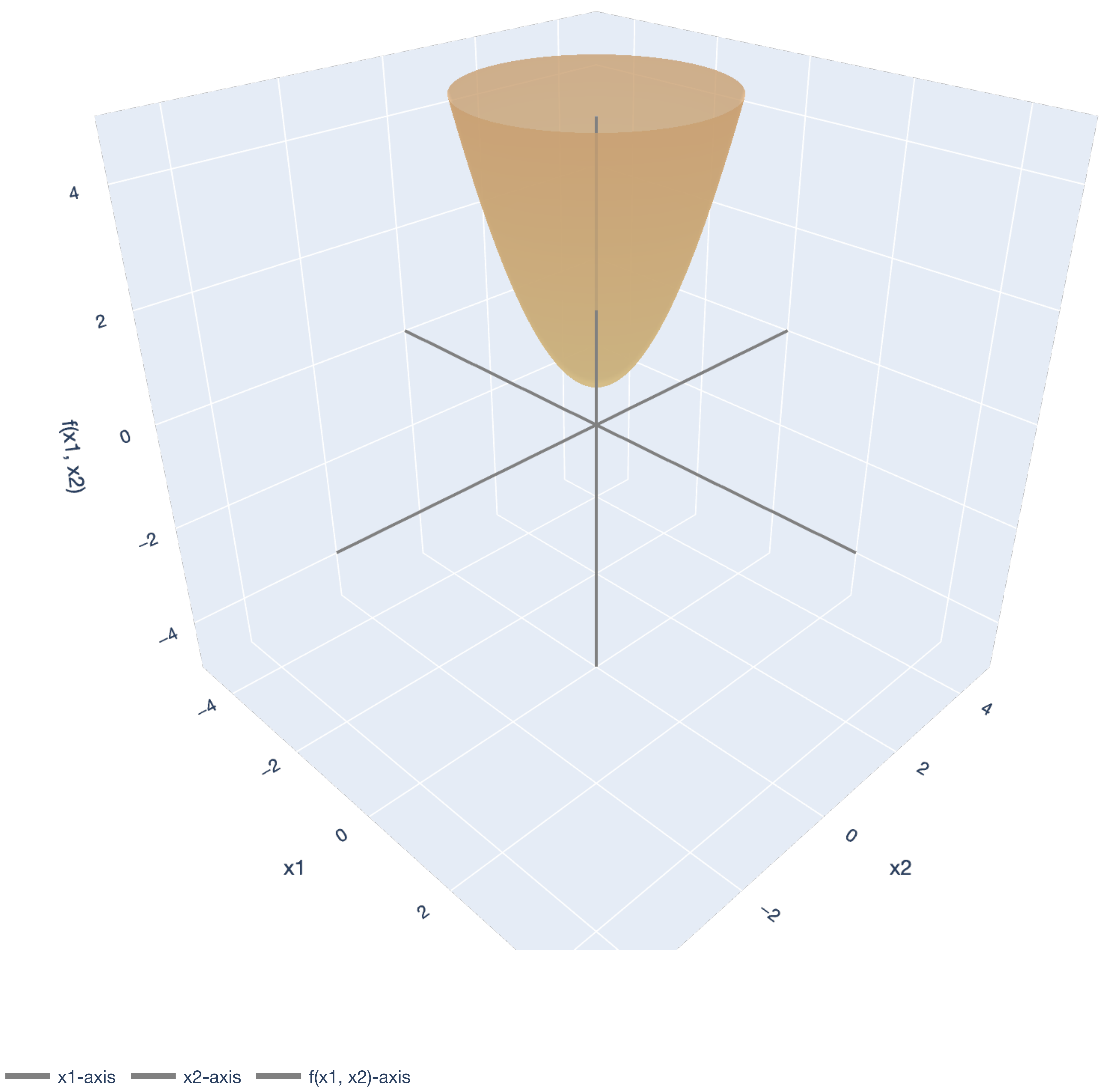
$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}.$$

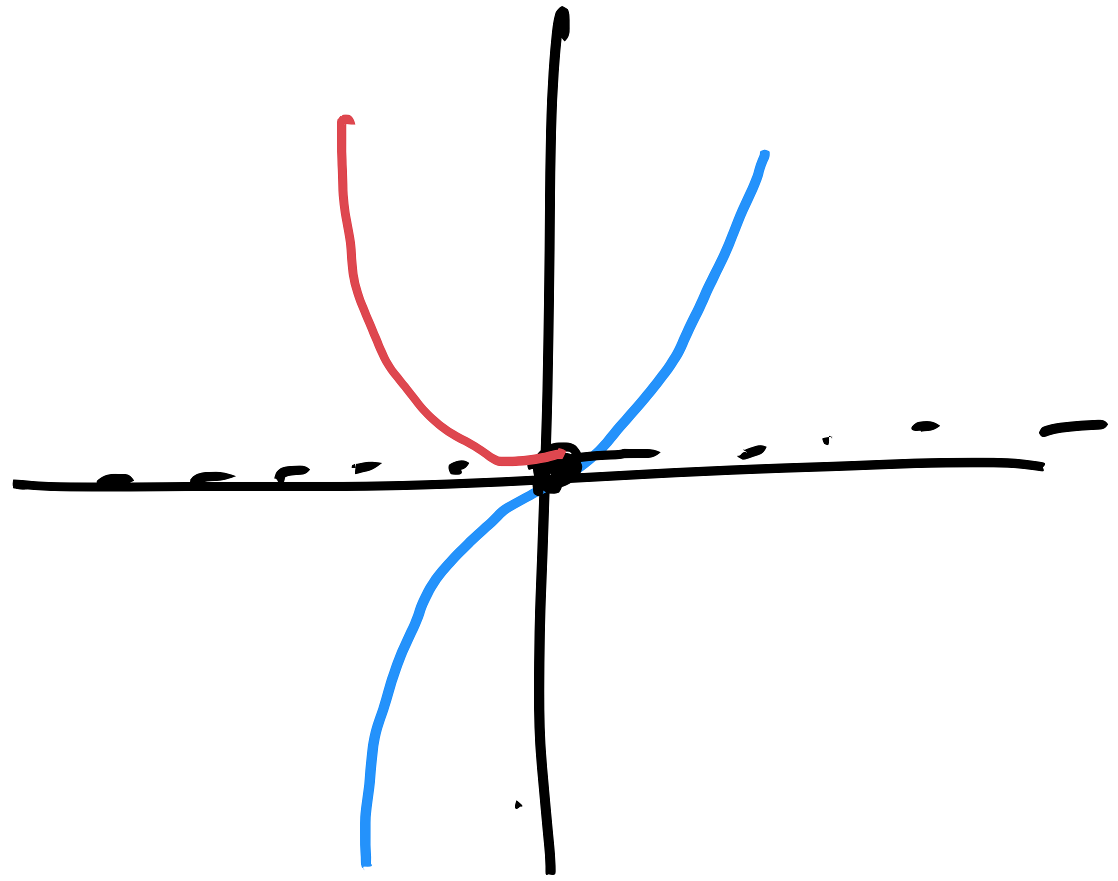
$$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \underline{\lambda_1, \ldots, \lambda_d > 0}$$

$$\implies \mathbf{X}^\top\mathbf{X} \text{ is positive definite!}$$ def.



necessary: x^* LOCAL MIN. $\Longrightarrow$ $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ is PSD.

SUFF. COND. $f'(x) = 0$, $f''(x) > 0$ $\Longrightarrow$ LOCAL MIN.



Unconstrained local minima

Sufficient conditions

Least Squares

OLS Theorem

$\nabla f(w) = 0.$

$\hookrightarrow \hat{w} = (X^T X)^{-1} X^T y$

Proof (OLS).

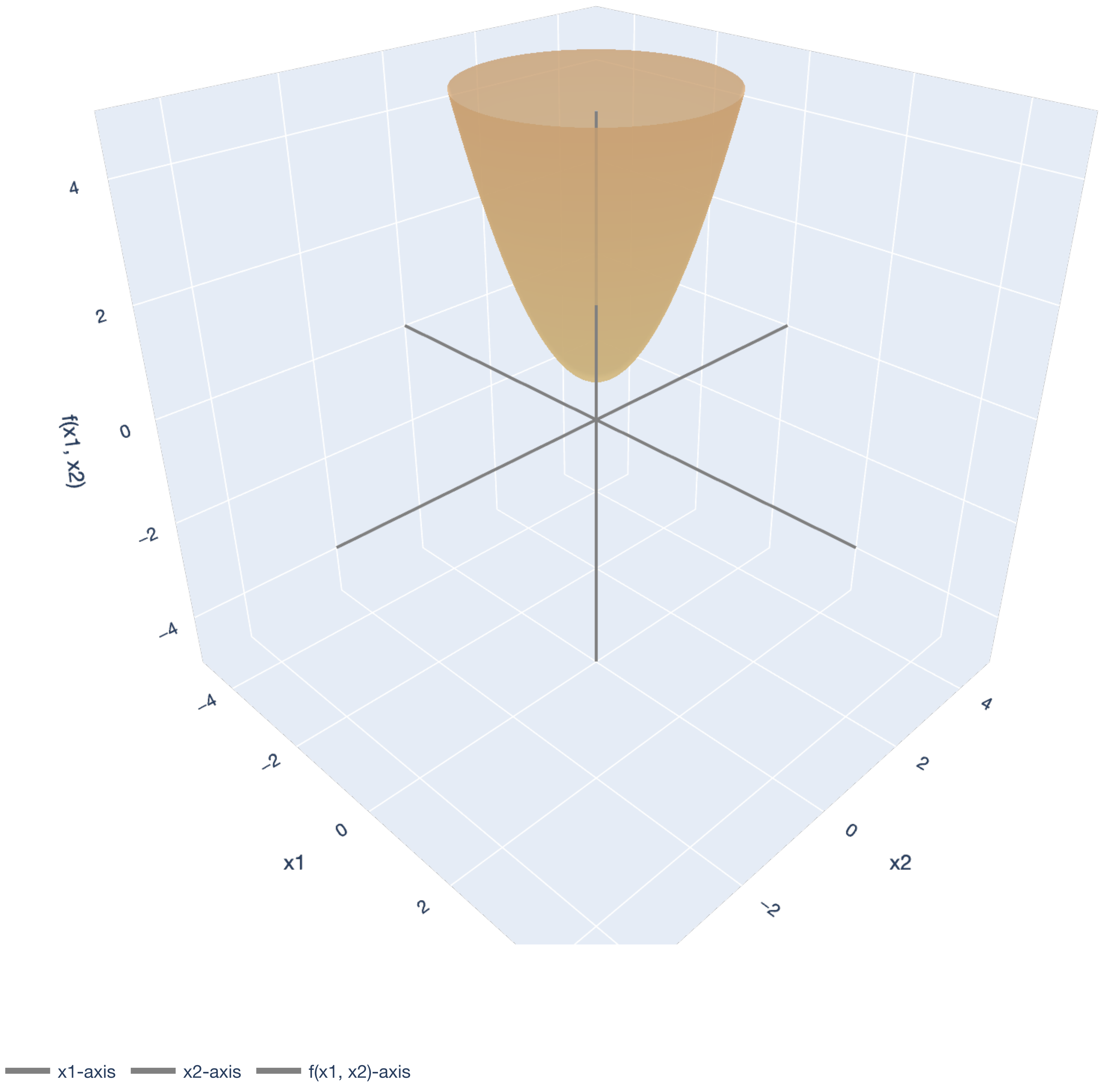
$f(\hat{w}) \leq f(w) \ \forall w \in \mathbb{R}^d.$

"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\boxed{\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}.}$$

$$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \lambda_1, \ldots, \lambda_d > 0$$

$\implies$ $\boxed{\mathbf{X}^\top\mathbf{X} \text{ is positive definite!}}$



Least Squares
OLS Theorem

Proof (OLS).

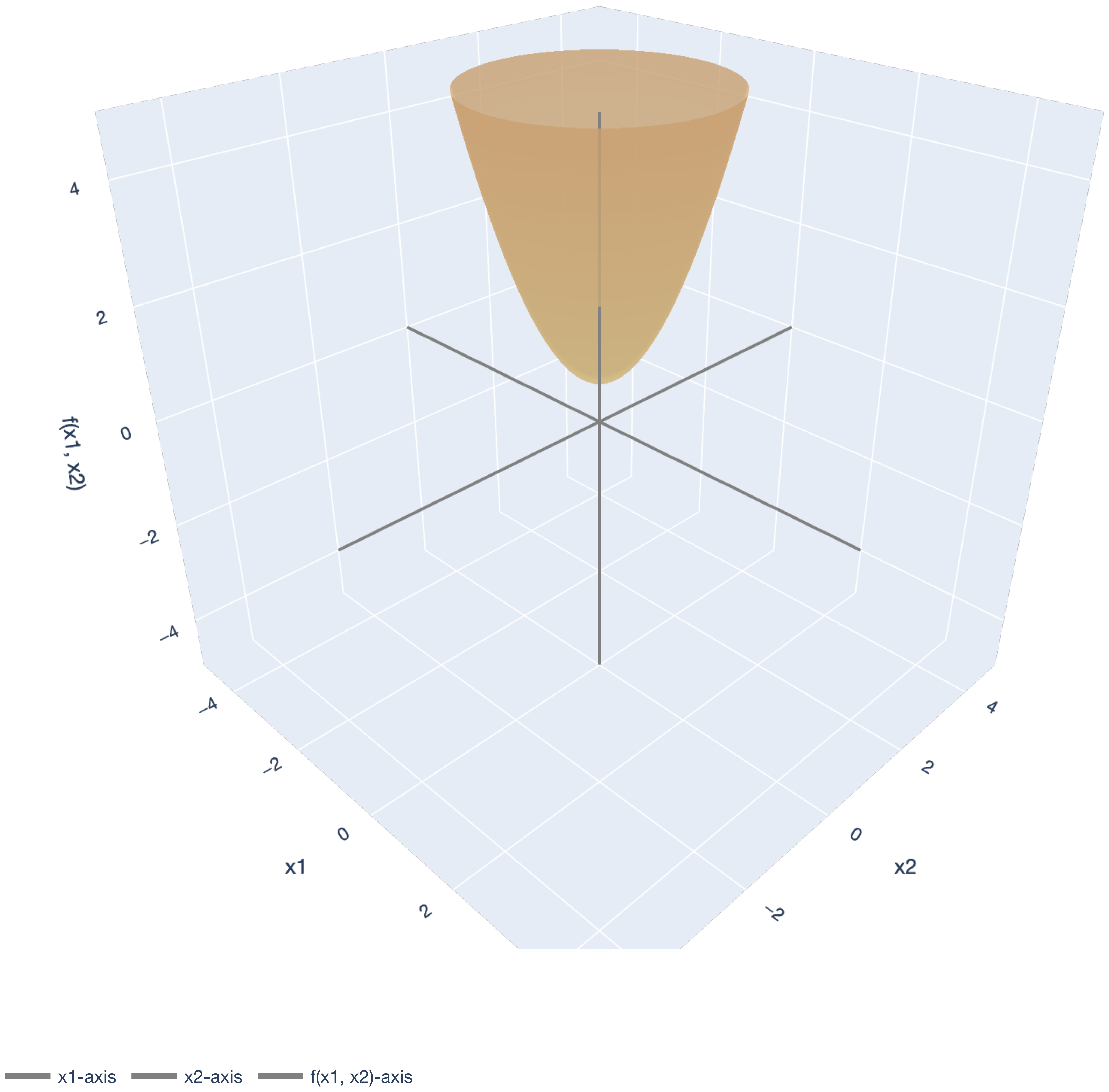
"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top \mathbf{X}.$$

$$\mathrm{rank}(\mathbf{X}) = d \implies \mathrm{rank}(\mathbf{X}^\top \mathbf{X}) = d \implies \lambda_1, \ldots, \lambda_d > 0$$

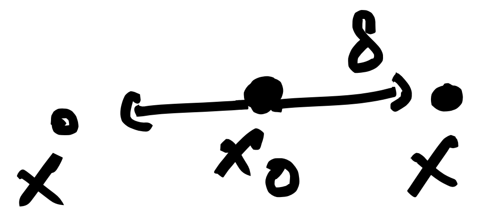
$\implies$ $\mathbf{X}^\top \mathbf{X}$ is positive definite!

Why is this the right thing to do?



Unconstrained Minima

Sufficient conditions



$$f(\overset{x}{\overset{\shortparallel}{x_0 + \delta}}) \approx f(x_0) + f'(x_0)\delta + \frac{1}{2} f''(x_0)\delta^2$$

when δ is small enough.

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

when $\|\mathbf{d}\|$ is small enough.

$f(x_0) \leq f(x) \quad \forall x \in \mathbb{R}^d.$

$f(x) = f(x_0+\delta) \approx f(x_0) + \cancel{f'(x_0)}\delta + \boxed{\frac{1}{2} f''(x_0)\delta^2}$

$f(x) = f(x_0+\delta) \approx f(x_0) +$ POSITIVE TERM.

$f(x_0) < f(x_0) +$ POSITIVE TERM.

$\Longrightarrow$ SAME INTUITION.

Sufficient conditions:

$$f'(x_0) = 0, f''(x_0) > 0.$$

Sufficient conditions:

$$\nabla f(\mathbf{x}_0) = \mathbf{0},\ \nabla^2 f(\mathbf{x}_0) \text{ is PD.}$$

Unconstrained Minima

Sufficient conditions

Theorem (Sufficient Conditions for Unconstrained Local Minimum).
Consider the optimization problem

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

twice diff. continuous
↓

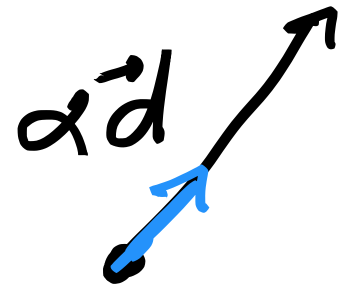
Let $\mathbf{x}^* \in \text{int}(\mathscr{C})$. If $f \in \mathscr{C}^2$ within a neighborhood $N_\delta(\mathbf{x}^*)$ of $\mathbf{x}^*$ and

$$\boxed{\nabla f(\mathbf{x}^*) = \mathbf{0}} \quad \text{and} \quad \boxed{\nabla^2 f(\mathbf{x}^*) \text{ is positive definite,}}$$

then $\mathbf{x}^*$ is a *strict* unconstrained local minimum.

$f(x^*) < f(x) \quad \forall x \in \mathscr{C}.$

Proof of sufficient conditions



Second order condition

Second-order condition. If $\nabla^2 f(\mathbf{x}^*)$ is PD, then $\mathbf{x}^*$ is an unconstrained local minimum.

Step 1: Use second-order Taylor's theorem with $\alpha\mathbf{d} \in \mathbb{R}^d$ with $||\mathbf{d}|| = 1$.

Choose an arbitrary direction $\alpha\mathbf{d} \in \mathbb{R}^d$, where $||\mathbf{d}|| = 1$ is a unit vector and $\alpha > 0$ is a scalar. By Taylor's Theorem (Peano's form):

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) = \nabla f(\mathbf{x}^*)^\top(\alpha\mathbf{d}) + \frac{1}{2}(\alpha\mathbf{d})^\top \nabla^2 f(\mathbf{x}^*)(\alpha\mathbf{d}) + o(||\alpha\mathbf{d}||^2)$$

$$= \alpha\nabla f(\mathbf{x}^*)^\top\mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2)$$

Proof of sufficient conditions

Second order condition

$v_d^T A v_d \le v^T A v \le v_1^T A v.$

Second-order condition. If $\nabla^2 f(\mathbf{x}^*)$ is PD, then $\mathbf{x}^*$ is an unconstrained local minimum.

Step 2: $\nabla^2 f(\mathbf{x}^*)$ is positive definite, so its eigenvalues are all positive. $\lambda_1, \dots, \lambda_d > 0.$

From Step 1, for any $\mathbf{d} \in \mathbb{R}^d$ with $\|\mathbf{d}\| = 1$ and $\alpha > 0$,

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) = \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2).$$

Let the eigenvalues of $\nabla^2 f(\mathbf{x}^*)$ be $\lambda_1 \ge \ldots \ge \lambda_d > 0$, and consider the smallest eigenvalue, $\lambda_d > 0$ with unit eigenvector $\mathbf{v}_d$ with $\|\mathbf{v}_d\| = 1$.

$Av = \lambda v.$

$$\Longrightarrow \underbrace{\frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d}}_{\text{for any } d.} \ge \underbrace{\frac{\alpha^2}{2}\mathbf{v}_d^\top \nabla f(\mathbf{x}^*)\mathbf{v}_d}_{v_d^T v_d = 1.} = \frac{\lambda_d \alpha^2}{2}.$$

Proof of sufficient conditions

$\nabla f(x^*) = 0.$

Second order condition

Second-order condition. If $\nabla^2 f(\mathbf{x}^*)$ is PD, then $\mathbf{x}^*$ is an unconstrained local minimum.

Step 3: We chose $\mathbf{d}$ arbitrarily, so the first-order term can be non-negative.

FOC

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) = \underbrace{\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d}}_{\text{FOC.}} + \underbrace{\frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d}}_{\geq \frac{\lambda_d \alpha^2}{2}} + o(\alpha^2)$$

*

Because $\mathbf{d}$ is an arbitrary direction (could be negative or positive), $\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} \geq 0$, and

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) \geq \frac{\lambda_d \alpha^2}{2} + o\left(\alpha^2\right) = \left(\frac{\lambda_d}{2} + \frac{o(\alpha^2)}{\alpha^2}\right)\alpha^2.$$

Proof of sufficient conditions

Second order condition

Second-order condition. If $\nabla^2 f(\mathbf{x}^*)$ is PD, then $\mathbf{x}^*$ is an unconstrained local minimum.

Step 4: If α is small enough, then $o(\alpha^2)/\alpha^2$ can be as small as we like.

From Step 3,

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) \geq \left(\frac{\lambda_d}{2} + \frac{o(\alpha^2)}{\alpha^2} \right) \alpha^2$$

$\leq \lambda_d/4$

For any $C > 0$, we can choose α small enough so $\left| \dfrac{o(\alpha^2)}{\alpha^2} \right| \leq C$.

Let's make $\left| \dfrac{o(\alpha^2)}{\alpha^2} \right|$ smaller than $\boxed{C = \dfrac{\lambda_d}{4}}$. Then, for any $\alpha > 0$ sufficiently small,

$$f(\mathbf{x}^* + \alpha\mathbf{d}) \geq f(\mathbf{x}^*) + \underbrace{\frac{\lambda_d}{4}\alpha^2}_{>0} > f(\mathbf{x}^*).$$

For any $C > 0$, we can find α small enough s.t.

$o(\alpha^2) \leq \alpha^2 C$.

$\dfrac{o(\alpha^2)}{\alpha^2} \leq C$.

Least Squares
OLS Theorem

Proof (OLS).

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \iff f(\mathbf{w}) = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}$$

"First derivative test." Take the gradient.

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top \mathbf{X})\mathbf{w} - 2\mathbf{X}^\top \mathbf{y}.$$

Set it equal to $\mathbf{0}$.

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

$\mathrm{rank}(\mathbf{X}) = d \Longrightarrow \mathrm{rank}(\mathbf{X}^\top \mathbf{X}) = d \Longrightarrow \mathbf{X}^\top \mathbf{X}$ is invertible:

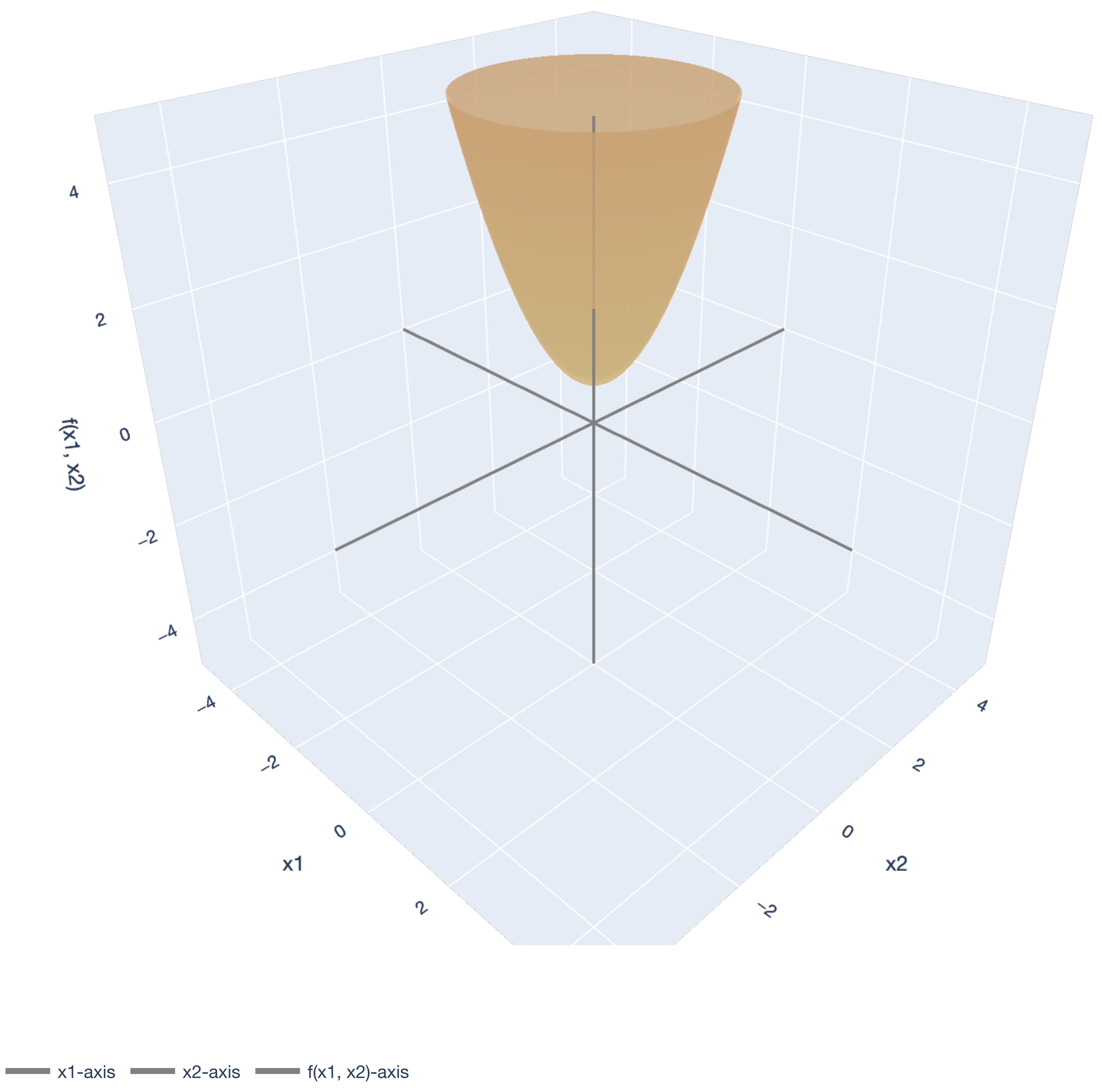
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

"Second derivative test." Take the *Hessian* of $f(\mathbf{w})$.

$$\nabla^2_{\mathbf{w}} f(\mathbf{w}) = 2\mathbf{X}^\top \mathbf{X}.$$

$$\mathrm{rank}(\mathbf{X}) = d \implies \mathrm{rank}(\mathbf{X}^\top \mathbf{X}) = d \implies \lambda_1, \ldots, \lambda_d > 0$$

$\implies \mathbf{X}^\top \mathbf{X}$ is positive definite!

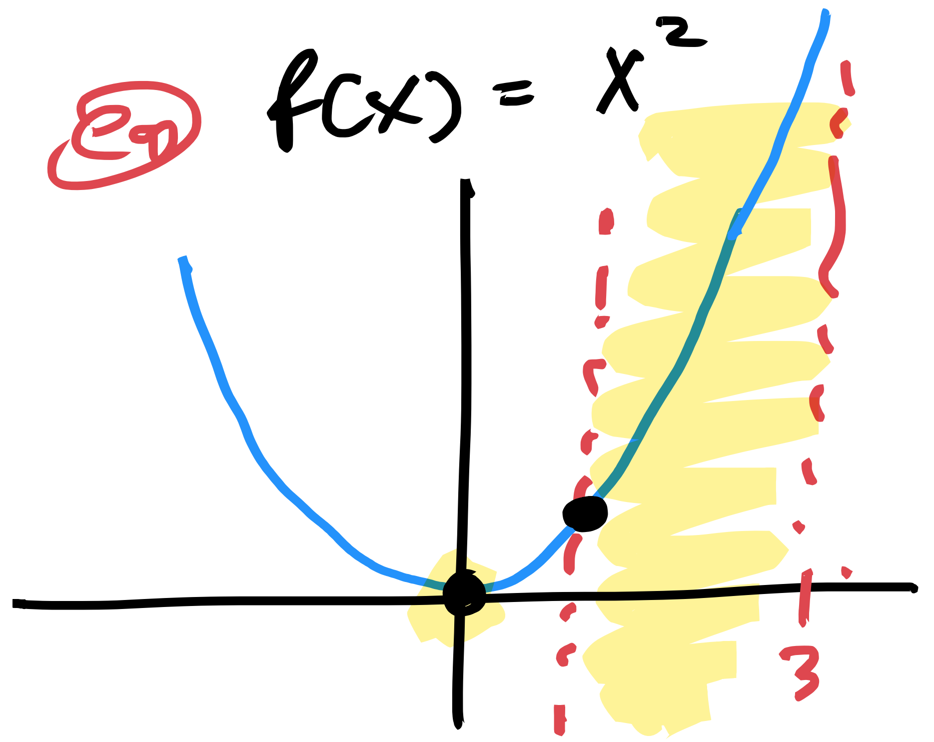


Finding global minima

Introducing constraint sets

Types of Minima

Big picture



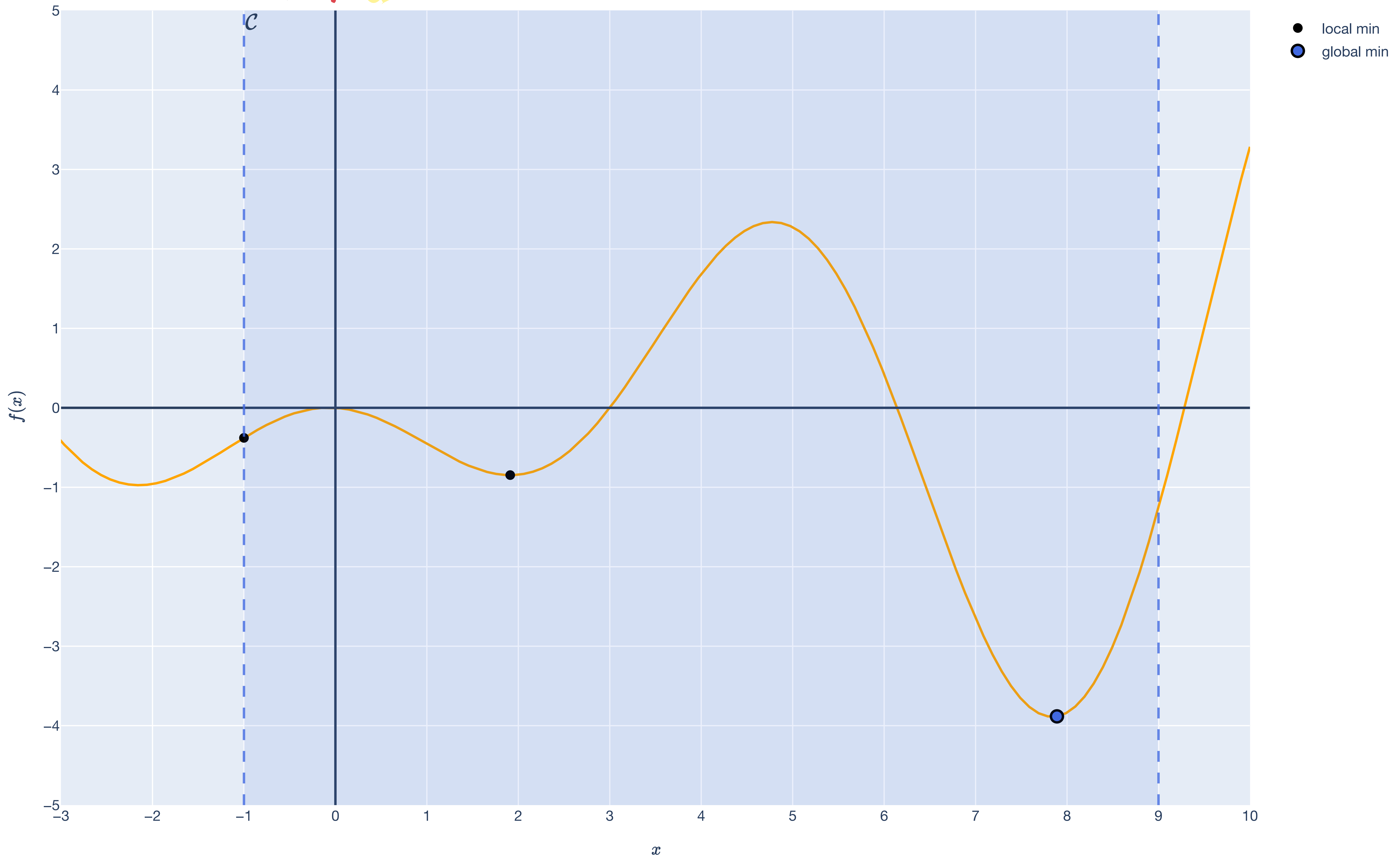
At the end of the day, we want to find ***global minima***.

Global minima could be either ***unconstrained local minima*** or ***constrained local minima***.

> Without $\mathscr{C}$, global minima are just one of the *unconstrained local minima.* (then. 77 slides?)

> With $\mathscr{C}$, global minima may lie on the boundary of the constraint set.

Strategy: Find all unconstrained and constrained local minima, then *test* for global minima.



Unconstrained Minima

Necessary conditions

Theorem (Necessary Conditions for Unconstrained Local Minimum). Consider the optimization problem

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

Suppose $\mathbf{x}^* \in \operatorname{int}(\mathscr{C})$ is an *unconstrained local minimum*. Then,

> *First-order condition.* If f is differentiable at $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
>
> *Second-order condition.* If f is twice-differentiable at $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is positive semidefinite, i.e. $\mathbf{v}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^d$.

Note: *These necessary conditions only apply to* $\mathbf{x}^* \in \operatorname{int}(\mathscr{C})$*!*

Finding global minima

RECIPE

Using necessary conditions with constraints

Necessary conditions for unconstrained local minima:

"CANDIDATES"

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \succeq 0.$$

$\succeq 0$ (PSD)

How do we find the *global* minimum from this?

$x \in \mathscr{C}$.

1. Find the set of possible *unconstrained local minima* from the first-order condition $M := \{\mathbf{x}^* \in \operatorname{int}(\mathscr{C}) : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.

2. Find the set of "boundary" points $B := \mathscr{C} \setminus \operatorname{int}(\mathscr{C}) = \{\mathbf{x} \in \mathscr{C} : \mathbf{x} \notin \operatorname{int}(\mathscr{C})\}$.

3. The global minimum must be in the set $M \cup B$, so evaluate f on all $\mathbf{x} \in M \cup B$ and see which one is smallest.

Finding global minima

Using necessary conditions with constraints

Necessary conditions for unconstrained local minima:

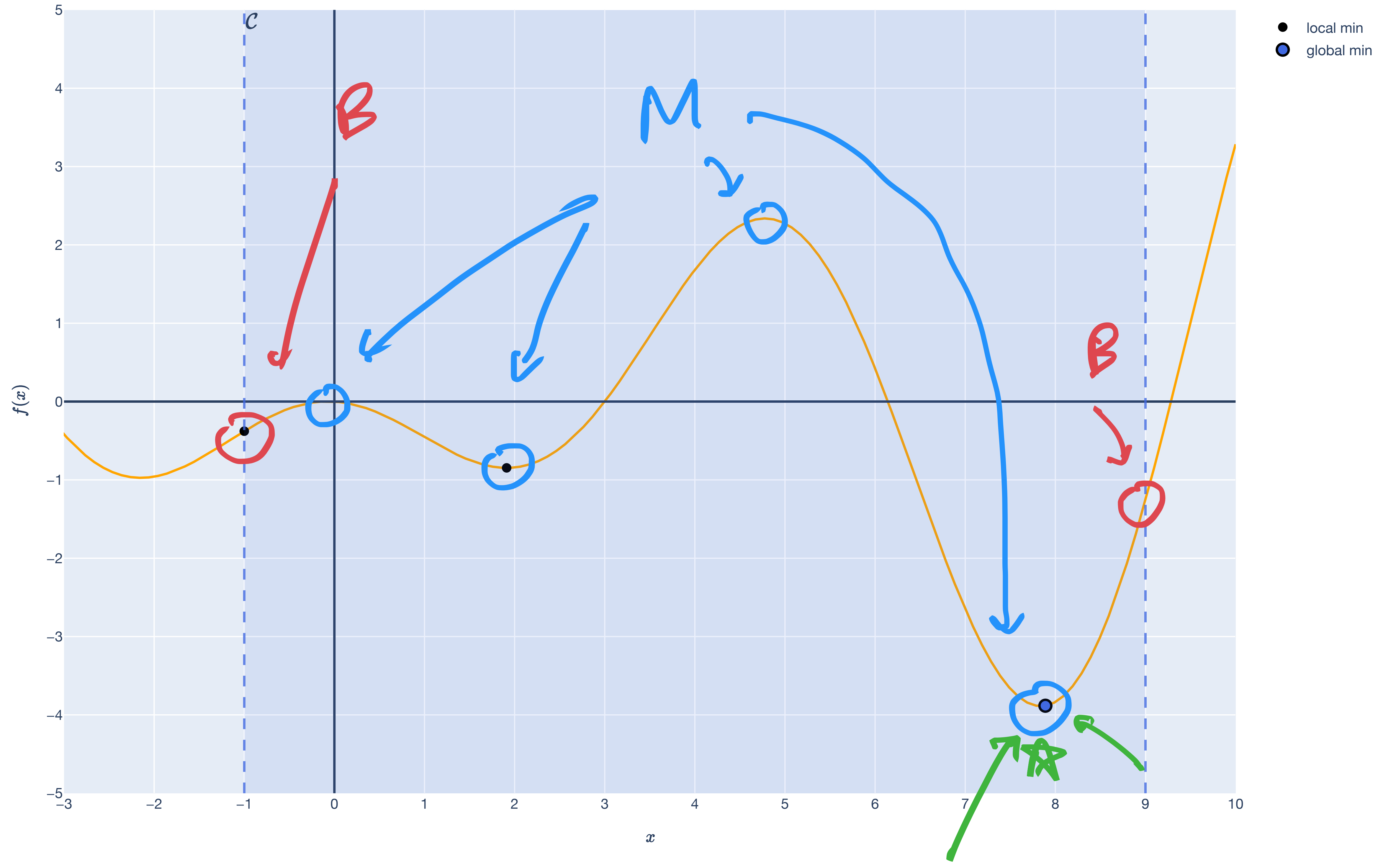
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \geq 0.$$

How do we find the *global* minimum from this?

1. Find the set of possible *unconstrained local minima* from the first-order condition $M := \{\mathbf{x}^* \in \text{int}(\mathscr{C}) : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.

2. Find the set of "boundary" points $B := \mathscr{C} \setminus \text{int}(\mathscr{C}) = \{\mathbf{x} \in \mathscr{C} : \mathbf{x} \notin \text{int}(\mathscr{C})\}$.

3. The global minimum must be in the set $M \cup B$, so evaluate f on all $\mathbf{x} \in M \cup B$ and see which one is smallest.



Finding global minima

Using necessary conditions without constraints

Necessary conditions for unconstrained local minima: $[\mathscr{C} = \mathbb{R}^d]$

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \geq 0.$$

How do we find the *global* minimum from this when $\mathscr{C} = \mathbb{R}^d$?

1. Find the set of possible *unconstrained local minima* from the first-order condition $M := \{\mathbf{x}^* \in \operatorname{int}(\mathscr{C}) : \nabla f(\mathbf{x}^*) = \mathbf{0}\} = \{\mathbf{x}^* \in \mathbb{R}^d : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.
2. There are no boundary points!
3. The global minimum must be in the set M, so evaluate f on all $\mathbf{x} \in M$ and see which one is smallest.

OR USE SUFF. COND. (HESSIAN IS PD?)

Finding global minima

Using necessary conditions without constraints

Necessary conditions for unconstrained local minima:

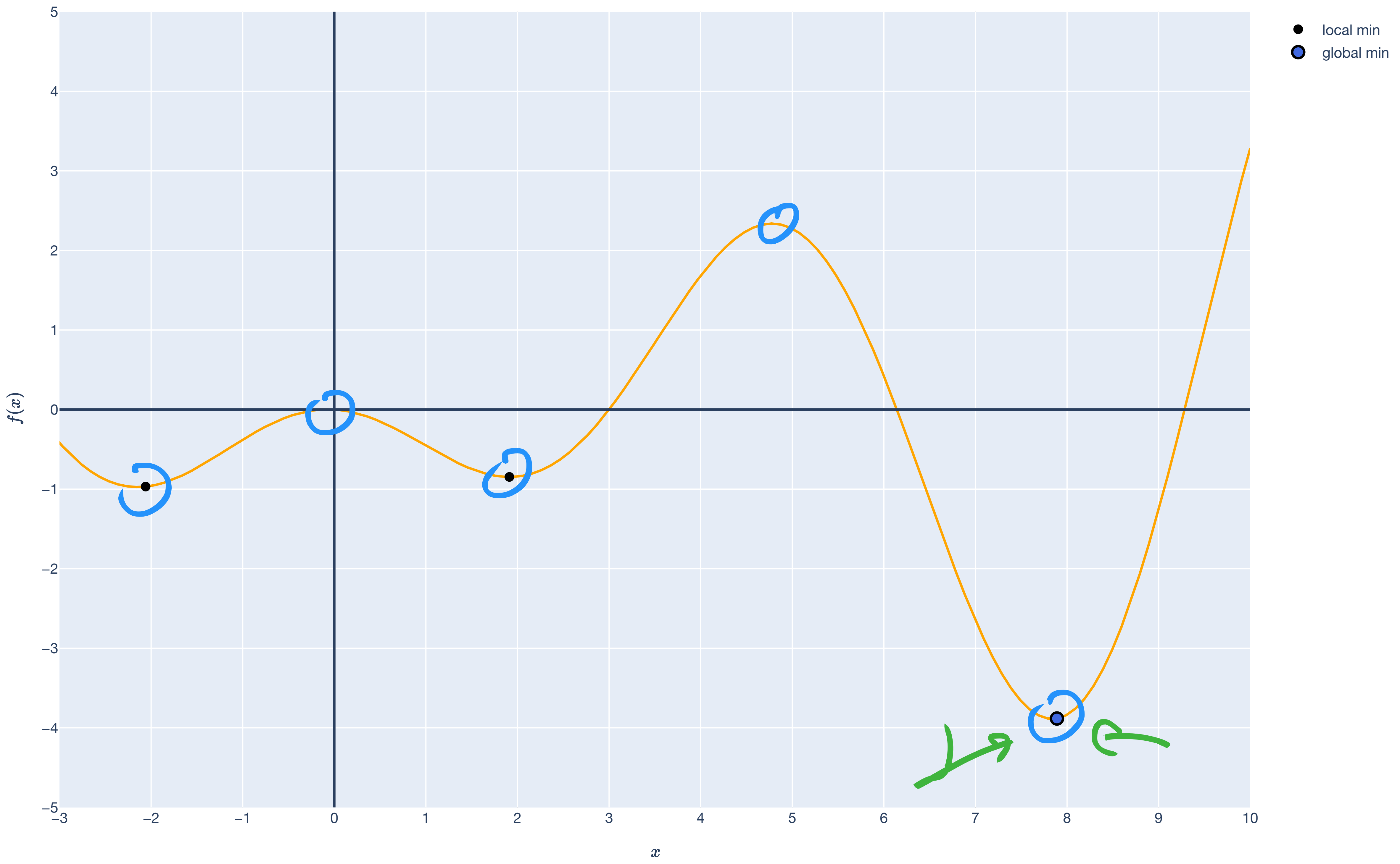
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \geq 0.$$

How do we find the *global* minimum from this when $\mathscr{C} = \mathbb{R}^d$?

1. Find the set of possible *unconstrained local minima* from the first-order condition $M := \{\mathbf{x}^* \in \mathrm{int}(\mathscr{C}) : \nabla f(\mathbf{x}^*) = \mathbf{0}\} = \{\mathbf{x}^* \in \mathbb{R}^d : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.

2. There are no boundary points!

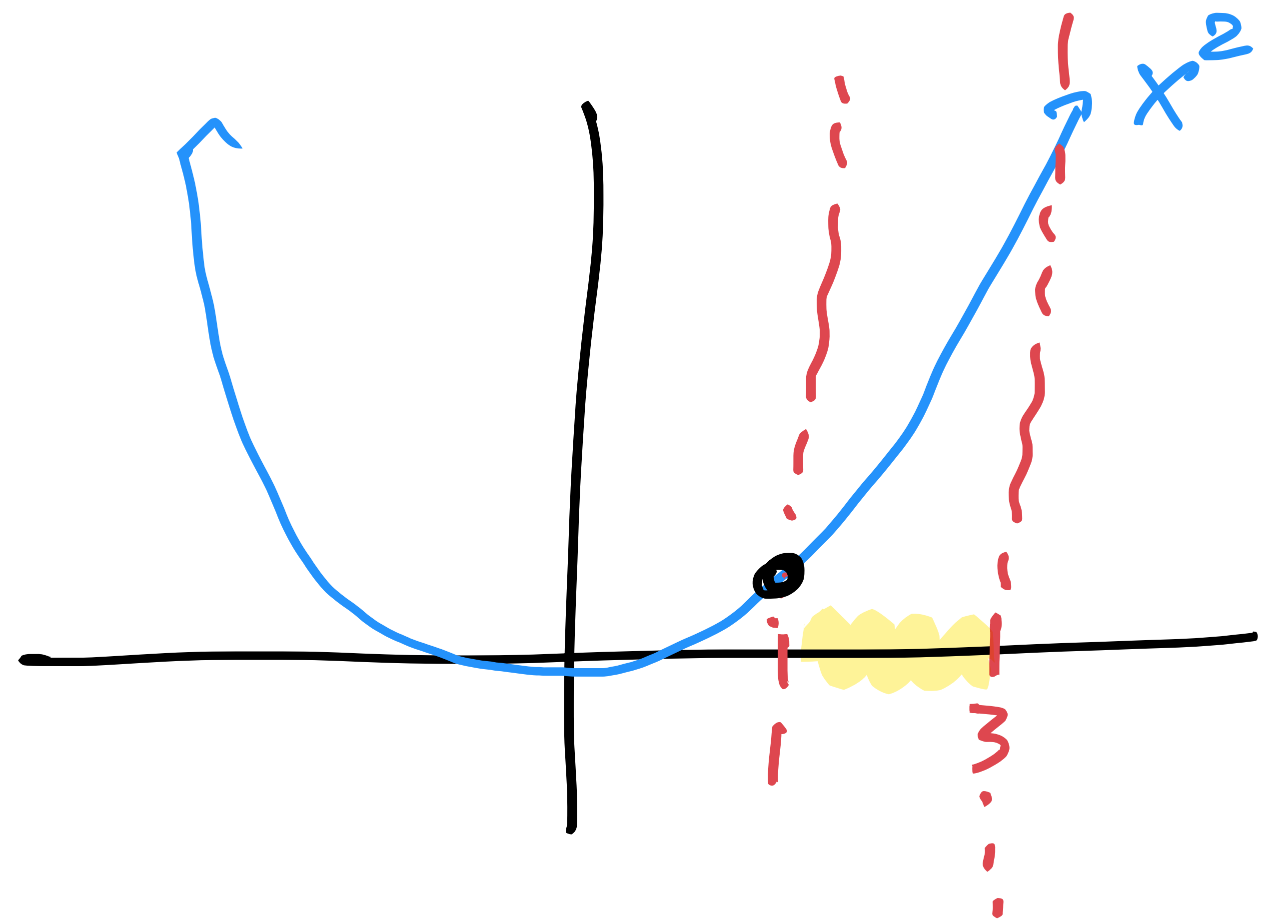
3. The global minimum must be in the set M, so evaluate f on all $\mathbf{x} \in M$ and see which one is smallest.



Unconstrained Minima

Example

Consider the one-dimensional optimization problem



$$\begin{array}{ll} \text{minimize} & x^2 \\ \text{subject to} & x \in [1,3] \end{array}$$

$\mathscr{C}$ $\quad [1,3) \cup [4,5] \cup \ldots$

① Find $\{x \in \mathscr{C} : f'(x) = 0\} := M$

$f'(x) = 2x \Rightarrow f'(x) = 0 = 2x \Rightarrow x = 0$

$\Rightarrow \boxed{M = \emptyset.}$

② $B = \{x \in \mathscr{C} : x \notin \mathrm{int}(\mathscr{C})\} = \{x \in [1,3] : x \notin (1,3)\} = \{1, 3\}$

③ $M \cup B = \{1,3\}$ $\quad \boxed{f(1) = 1^2 = 1.}$ $\quad f(3) = 3^2 = 9.$

In general, this works for any one-dimensional problem where $f : \mathbb{R} \to \mathbb{R}$ is continuous on $\mathscr{C} = [a,b]$ and differentiable on $\mathrm{int}(\mathscr{C}) := (a,b)$.

$$\boxed{x^* = 1}$$

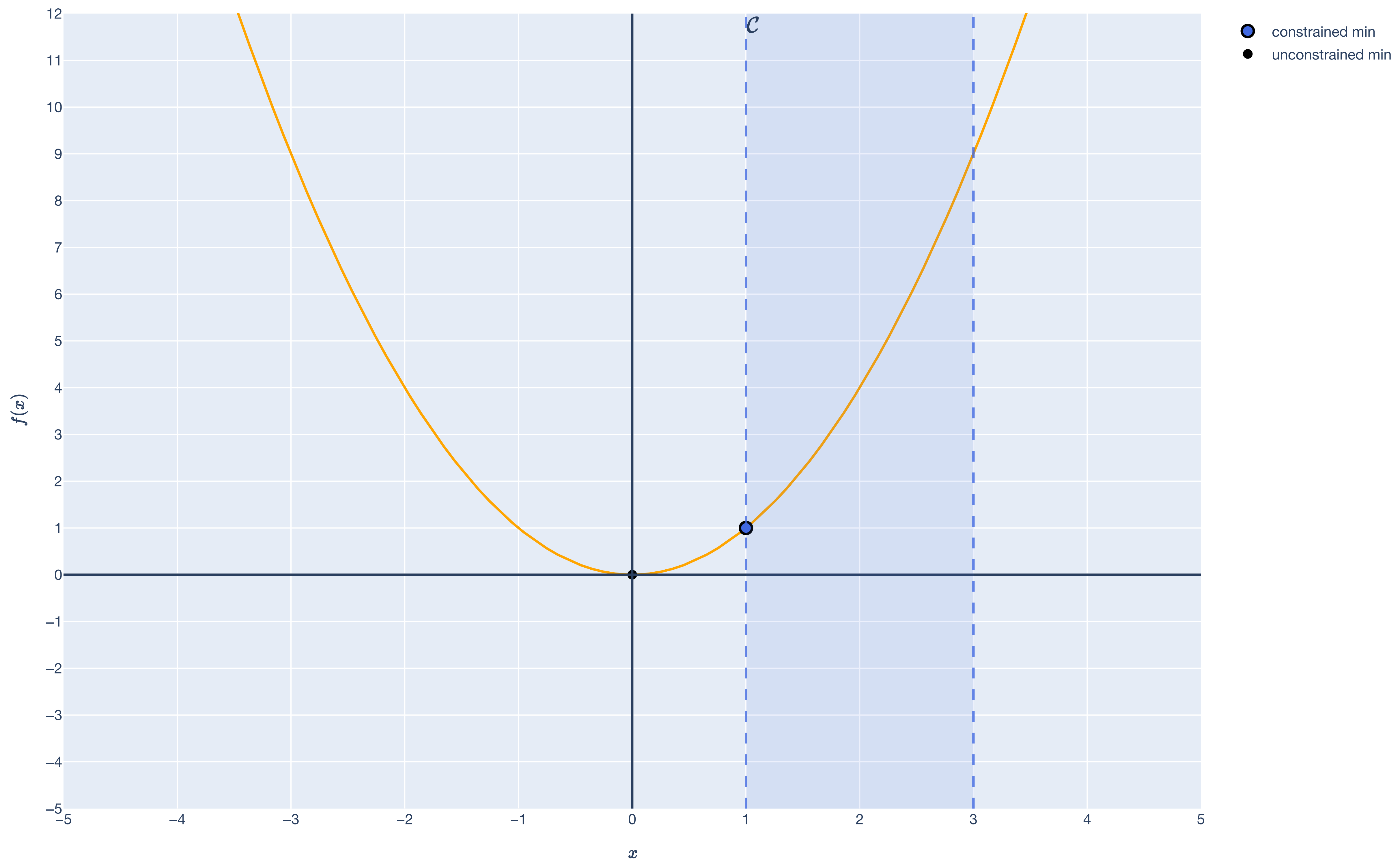
$$\boxed{f(x^*) = 1}$$

Unconstrained Minima

Example

Consider the one-dimensional optimization problem

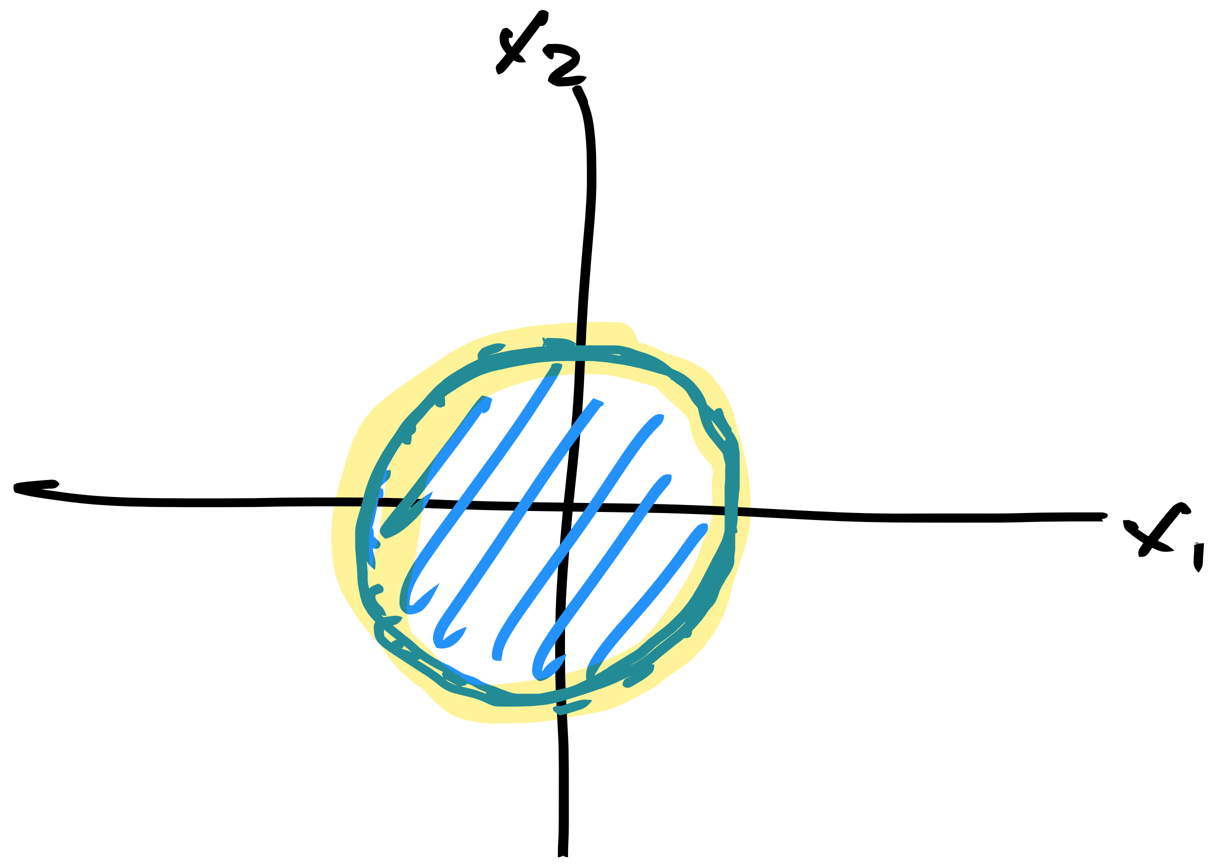
$$\begin{array}{ll} \text{minimize} & x^2 \\ \text{subject to} & x \in [1,3] \end{array}$$



In general, this works for any one-dimensional problem where $f : \mathbb{R} \to \mathbb{R}$ is continuous on $\mathscr{C} = [a, b]$ and differentiable on $\mathrm{int}(\mathscr{C}) := (a, b)$.

Unconstrained Minima

Example: Why haven't we solved optimization?



Consider the two-dimensional optimization problem

$$\begin{array}{ll} \text{minimize} & f(x_1, x_2) \\ \text{subject to} & x_1^2 + x_2^2 \leq 1 \end{array}$$

We might have to evaluate f on the infinite number of points on the boundary of the circle, $\mathscr{C} \setminus \text{int}(\mathscr{C}) := \{\mathbf{x} \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$!

This isn't feasible, so the question is:

How do we deal with the possible constrained local minima induced by $\mathscr{C}$?

Unconstrained Minima

Example: Why haven't we solved optimization?

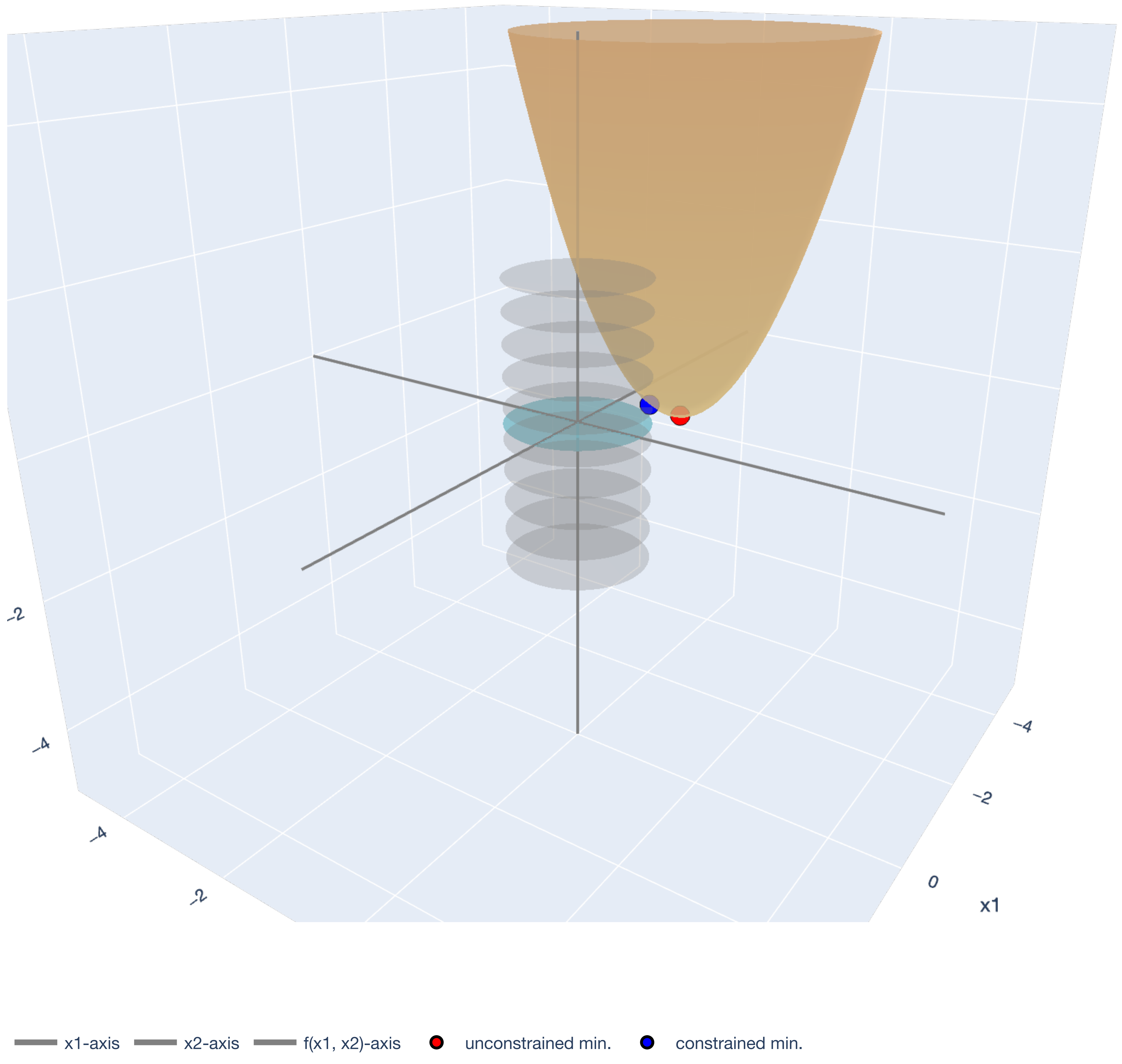
Consider the two-dimensional optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x_1, x_2) \\ \text{subject to} \quad & x_1^2 + x_2^2 \leq 1 \end{aligned}$$

We might have to evaluate f on the infinite number of points on the boundary of the circle, $\mathscr{C} \backslash \text{int}(\mathscr{C}) := \{\mathbf{x} \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$!

This isn't feasible, so the question is:

How do we deal with the possible constrained local minima induced by $\mathscr{C}$?

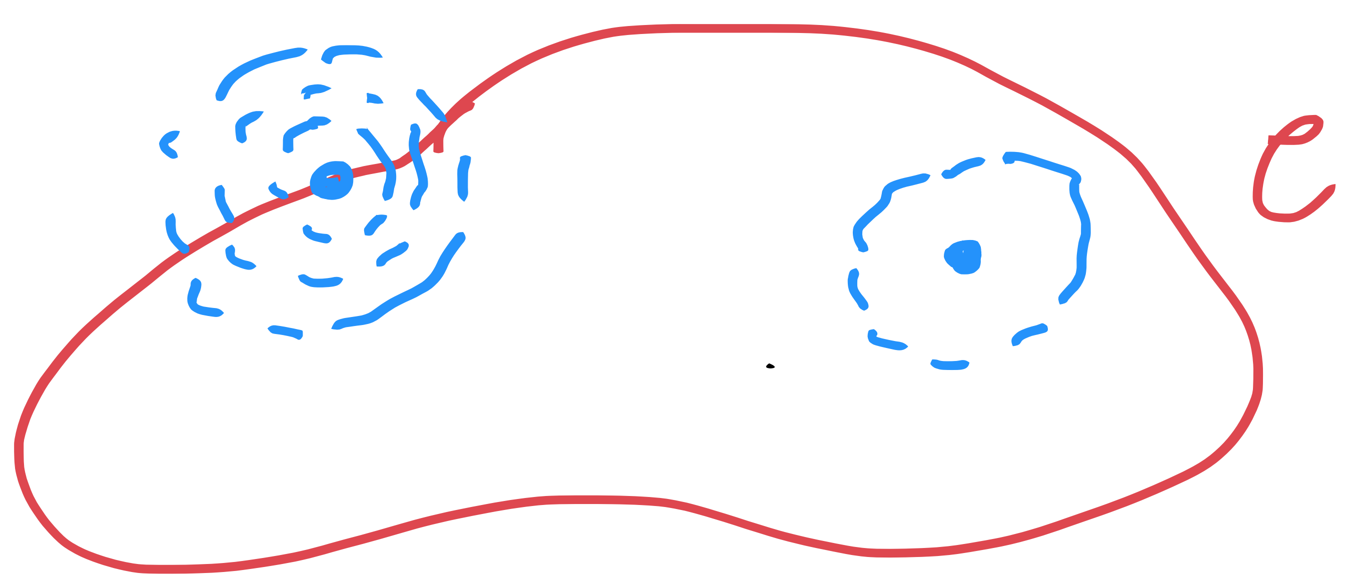


Constrained Minima

Equality Constraints and the Lagrangian

Constrained Minima

What can go wrong?



Recall the definitions of *(unconstrained) local minima* and *constrained local minima.*

A point $\hat{\mathbf{x}} \in \mathscr{C}$ is an ***unconstrained local minimum*** if there exists a neighborhood $B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}$ around $\hat{\mathbf{x}}$ such that

COULD NOT BE ON BOUNDARY

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in B_\delta(\hat{\mathbf{x}}).$$

A point $\hat{\mathbf{x}} \in \mathscr{C}$ is a ***local minimum*** if there exists a neighborhood $B_\delta(\mathbf{x})$ around $\hat{\mathbf{x}}$ such that

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathscr{C} \cap B_\delta(\hat{\mathbf{x}}).$$

COULD BE ON THE BOUNDARY.

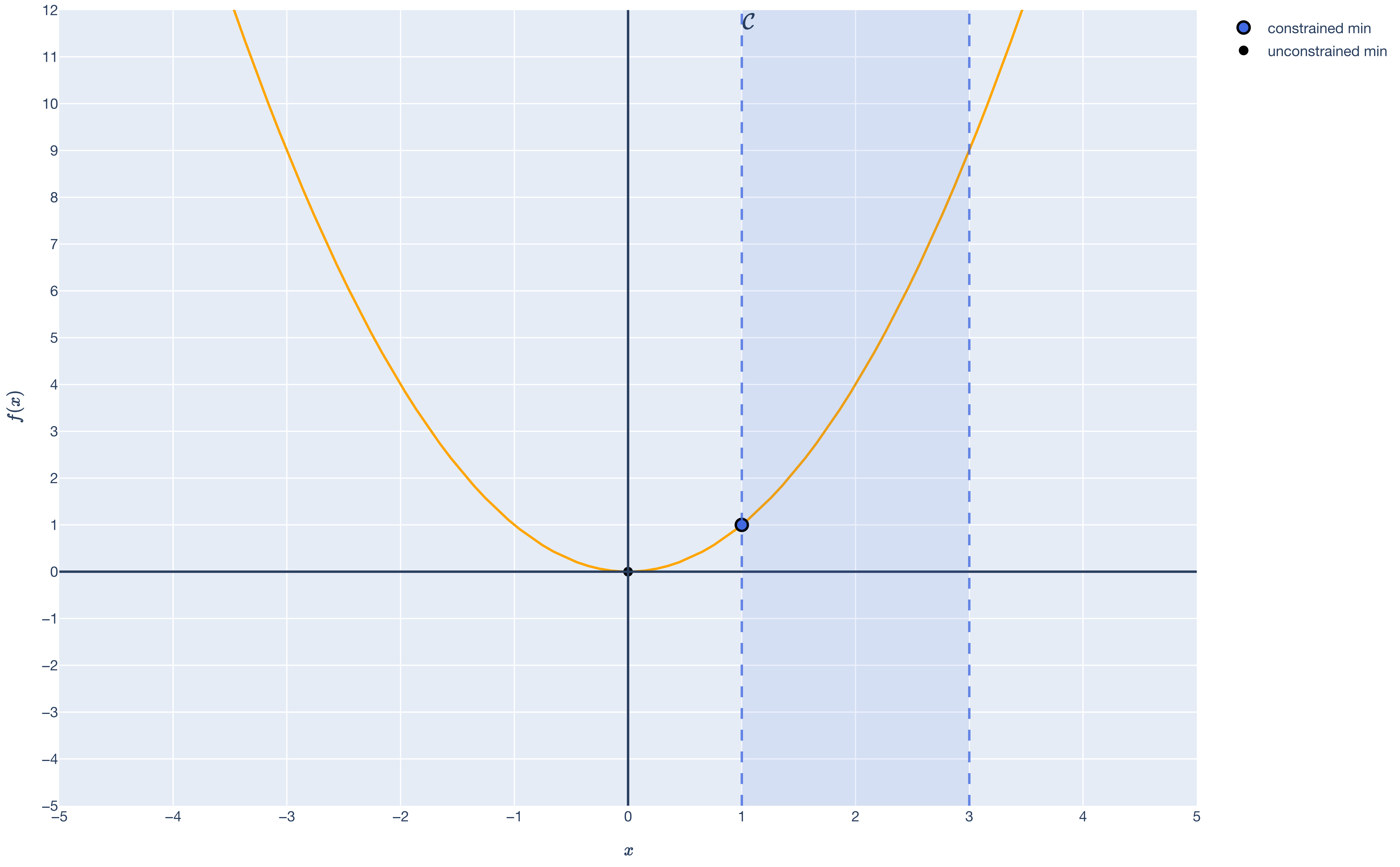
We also call this a ***constrained local minimum***.

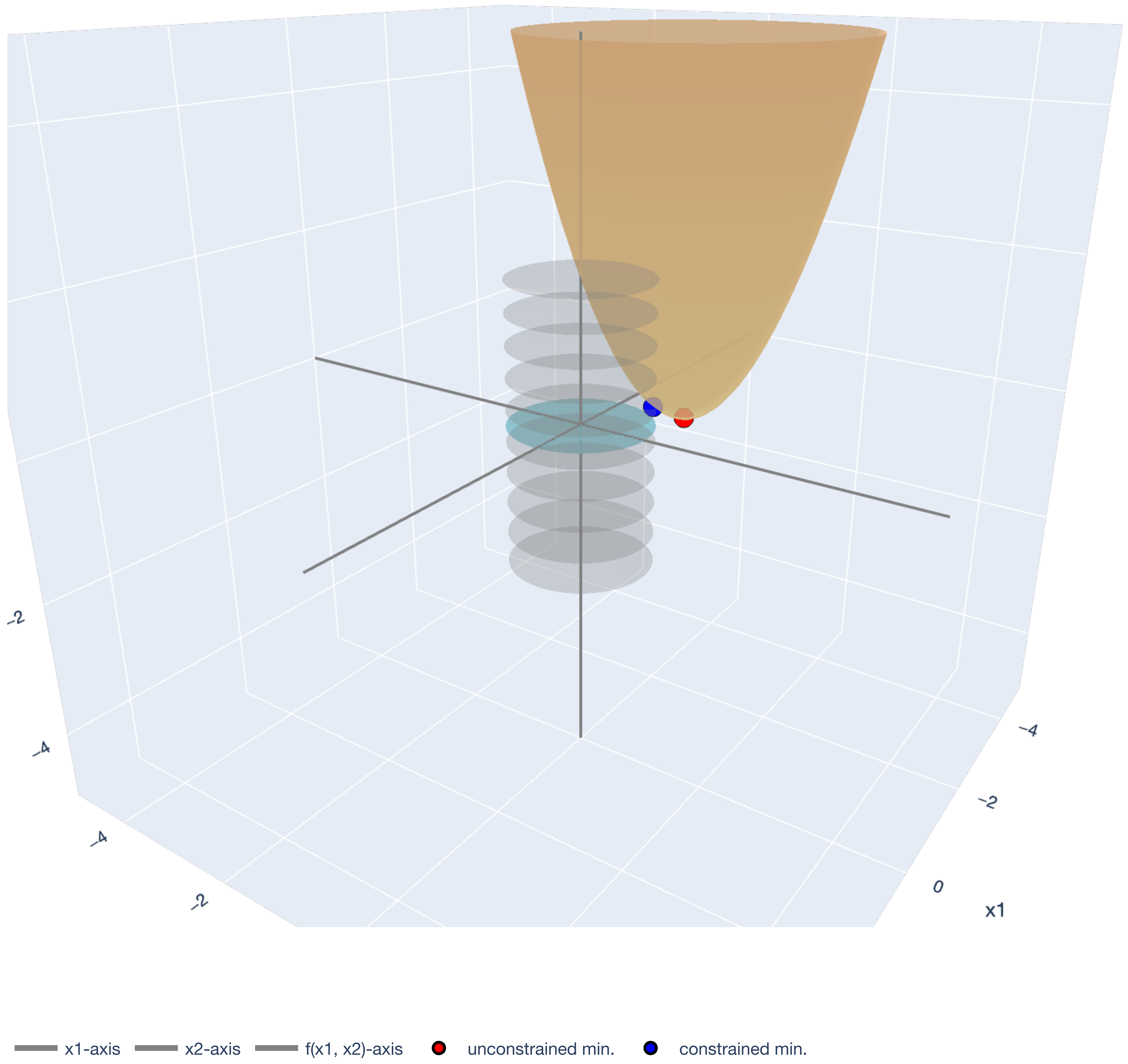
Constrained Local Minima

Minimum values on the "edge of the constraint set"

$\mathbb{R}^2$

$\mathbb{R}^1$.





Constrained Minima

Equality constrained optimization

An ***equality constrained minimization problem*** is an optimization problem defined by an objective function $f : \mathbb{R}^d \to \mathbb{R}$, decision variables $\mathbf{x} \in \mathbb{R}^d$, and constraints $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$ from a $\mathscr{C}^1$ vector-valued function $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^m$, written as follows:

$$\boxed{\begin{array}{ll}\text{minimize} & f(\mathbf{x})\\ \text{subject to} & \left.\begin{array}{l} h_1(\mathbf{x}) = 0\\ \quad\vdots\\ h_m(\mathbf{x}) = 0\end{array}\right\} e \end{array}}$$

m diff. scalar valued functions.

$h_1(x) = 5$

$\Rightarrow h_1(x) - 5 = 0.$

where $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_m(\mathbf{x}))$.

Constrained Minima

Equality constrained optimization

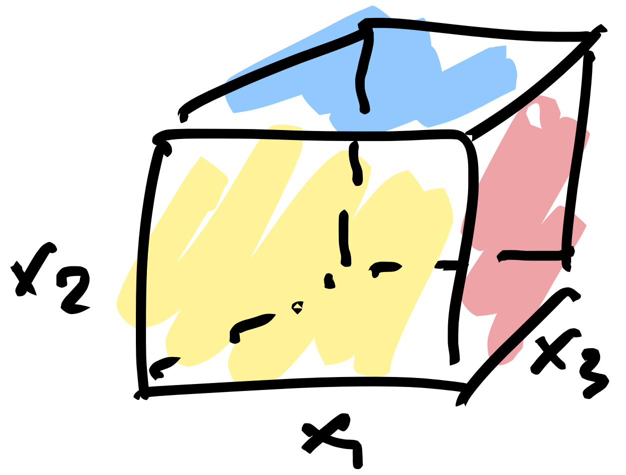
$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_1(\mathbf{x}) = 0 \\ & \vdots \\ & h_m(\mathbf{x}) = 0 \end{aligned}$$

The $= 0$ constraint is WLOG:

If $h_j(\mathbf{x}) = c$ then we can always consider $h_j'(\mathbf{x}) = h_j(\mathbf{x}) - c = 0$ instead.

Constrained Minima: Equality Constraints

Example: Maximum Volume Box



$x_1x_2x_3$

Consider the following optimization problem

$$\begin{array}{ll} \text{minimize} & x_1x_2x_3 \leftarrow \text{volume (length} \times \text{width} \times \text{height)} \\ \text{subject to} & \boxed{x_1x_2 + x_2x_3 + x_1x_3 - c/2 = 0} \end{array}$$

Here, $\mathbf{x} \in \mathbb{R}^3$, the objective is $f(\mathbf{x}) = x_1x_2x_3$, and $h : \mathbb{R}^3 \to \mathbb{R}$ is just scalar-valued (one constraint) with $h(\mathbf{x}) = x_1x_2 + x_2x_3 + x_1x_3 - c/2$.

$$x_1x_2 + x_2x_3 + x_1x_3 = c/2 = 10.$$

Constrained Minima: Equality Constraints

Idea

We will convert the *constrained* optimization problem into an *unconstrained* optimization problem and then use our tools for unconstrained optimization problems:

PSD. $\succeq$.

$$\boxed{\nabla f(\mathbf{x}) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) \geq 0.}$$

The unconstrained optimization problem will have m more variables (for each constraint h_j for $j \in [m]$), represented by a vector $\lambda \in \mathbb{R}^m$ (the ***Lagrange multipliers***).

Constrained Minima: Equality Constraints

Definition of the Lagrangian

For an optimization problem with equality constraints

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_1(\mathbf{x}) = 0 \\ & \vdots \\ & h_{\cancel{r}}(\mathbf{x}) = 0 \end{aligned}$$

$d+m \rightarrow \mathbb{R}$ $\quad m$

the ***Lagrangian function*** $L : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is the function

$\min_{x, \lambda}$

$$L(\mathbf{x}, \lambda) := f(\mathbf{x}) + \boxed{\sum_{i=1}^{m} \lambda_i h_i(\mathbf{x})} = f(\mathbf{x}) + \lambda^{\top} \mathbf{h}(\mathbf{x}).$$

Notice that the function $L(\mathbf{x}, \lambda)$ is an *unconstrained* function.

Constrained Minima: Equality Constraints

Regularity Conditions

For an optimization problem with equality constraints,

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0 \end{array}$$

a point $\mathbf{x} \in \mathbb{R}^{\overset{d}{\cancel{n}}}$ is a ***regular point*** if it is feasible and the gradients $\nabla h_1(\mathbf{x}), \ldots, \nabla h_m(\mathbf{x})$ are linearly independent.

This will be the (usually) easily checkable condition we need for a minimum in the Lagrangian. Another condition is that $h_1, \ldots, h_m$ are linear functions.

Constrained Minima: Equality Constraints

Lagrange Multiplier Theorem

NECESSARY

$\nabla L(x, \lambda) = 0$

$\nabla^2 L(x, \lambda)$ is PSD

Theorem (Lagrange Multiplier Theorem). Let $\mathbf{x}^* \in \mathbb{R}^d$ be a local minimum that is a regular point. Then, there exists a unique vector $\lambda \in \mathbb{R}^m$ called a ***Lagrange multiplier*** such that

$$\boxed{\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0}$$

If, in addition, f and $h_1, \ldots, h_m$ are twice continuously differentiable,

$$\mathbf{d}^\top \left(\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*) \right) \mathbf{d} \geq 0$$

for all $\mathbf{d} \in \mathbb{R}^n$ such that $\nabla \mathbf{h}(\mathbf{x}^*)^\top \mathbf{d} = 0$, where $\nabla \mathbf{h}(\mathbf{x}^*) \in \mathbb{R}^{d \times m}$ is the Jacobian of $\mathbf{h}$ at $\mathbf{x}^*$.

Constrained Minima: Equality Constraints

How to remember the Lagrange multiplier theorem

The Lagrangian function is:

$$L(\mathbf{x}, \lambda) = \nabla f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}) = 0$$

Remember the necessary conditions for local minima:

$$\nabla f(\mathbf{x}) = \mathbf{0} \text{ and } \nabla^2 f(\mathbf{x}) \geq 0.$$

Applying the first-order necessary conditions for the Lagrangian, a local minimum $(\mathbf{x}^*, \lambda^*)$ must satisfy

$$\boxed{\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = \mathbf{0}} \text{ and } \boxed{\nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = \mathbf{0}.}$$

Notice that $\nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = \mathbf{0}$ is the same as requiring feasibility: $\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$.

Constrained Minima: Equality Constraints

Lagrange Multiplier Theorem: Sufficient Conditions

Theorem (Lagrange Multiplier Theorem - Sufficient Conditions). Let f and $\mathbf{h}$ be $\mathscr{C}^2$ functions, such that $\mathbf{x}^* \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}^m$ satisfy

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = 0 \text{ and } \nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = 0$$

$$\mathbf{d}^\top \nabla^2_{\mathbf{x},\mathbf{x}} L(\mathbf{x}^*, \lambda^*)\mathbf{d} > 0, \quad \forall \mathbf{d} \text{ such that } \nabla \mathbf{h}(\mathbf{x}^*)^\top \mathbf{d} = \mathbf{0}.$$

PD

Then, $\mathbf{x}^*$ is a local minimum.

Constrained Minima: Equality Constraints

How do we use the Lagrangian?

RECIPE

Assuming that *a global minimum exists* and f and $\mathbf{h}$ are $\mathscr{C}^1$, let the Lagrangian be:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x})\,.$$

To find a global minimum…

1. Find the set $(\mathbf{x}^*, \lambda^*)$ satisfying the necessary conditions: $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = 0$ and $\nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = 0$. This is just our usual first-order condition applied to $L(\,\cdot\,,\,\cdot\,)$!
2. Find the set of all non-regular points. ← NECESSARY COND. DON'T APPLY TO! (Finding $B = \{x \in \mathcal{C} : x \notin \operatorname{int} \mathcal{C}\}$)
3. The global minima must be among the points in (1) or (2).

① ∪ ②.

Constrained Minima: Equality Constraints

Example: Maximum Volume Box

Consider the following optimization problem

$$\boxed{8\left(\sqrt{\frac{c}{24}}\right)^3}$$

minimize $-x_1x_2x_3$ $\quad c>0.$

$$\begin{array}{ll} \overset{\text{maximize}}{\cancel{\text{minimize}}} & x_1x_2x_3 \\ \text{subject to} & x_1x_2 + x_2x_3 + x_1x_3 - c/2 = 0 \end{array}$$

$x_1 = x_2 = x_3$

$$\boxed{x^* = \boxed{2\sqrt{\frac{c}{24}}}}$$

① LAGRANGIAN . $h_1(x) = x_1x_2 + x_2x_3 + x_1x_3 - c/2.$

$$\Rightarrow \boxed{L(x,\lambda) = x_1x_2x_3 + \underline{\lambda}(x_1x_2 + x_2x_3 + x_1x_3 - c/2)}$$

$\nabla h_1(x) = \nabla h_1\left(2\sqrt{\frac{c}{24}},\ 2\sqrt{\frac{c}{24}},\ 2\sqrt{\frac{c}{24}}\right)$

$$\nabla_x = \begin{bmatrix} x_2x_3 + \lambda(x_2+x_3) \\ x_1x_3 + \lambda(x_1+x_3) \\ x_1x_2 + \lambda(x_2+x_1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\nabla_\lambda = x_1x_2 + x_2x_3 + x_1x_3 - c/2 = 0$

$x_1 =$

$x_2 = -2\cdot\sqrt{\frac{c}{24}}$

$x_3 =$

$= \boxed{2\sqrt{\frac{c}{24}}}$

$x_1 \longrightarrow$ ① $x_2x_3 + \lambda(x_2+x_3) = 0$

$x_2 \longrightarrow$ ② $x_1x_3 + \lambda(x_1+x_3) = 0$

$\longrightarrow$ ③ $x_1x_2 + \lambda(x_2+x_1) = 0$

$\longrightarrow$ ④ $x_1x_2 + x_2x_3 + x_1x_3 - c/2 = 0$

$\Rightarrow$

$$\lambda = -\frac{x_2x_3}{x_2+x_3} = -\frac{x_1x_3}{x_1+x_3} = -\frac{x_1x_2}{x_2+x_1}$$

$\lambda(x_1x_2 + x_1x_3) = \lambda(x_1x_2 + x_2x_3)$

$\Rightarrow \lambda x_1x_3 = \lambda x_2x_3$

$\boxed{x_1 = x_2}$ $\Rightarrow \boxed{x_1 = x_2 = x_3}$

$x_1 = x_2 = x_3 = -2\lambda.$

$$\lambda = -\sqrt{c/24}$$

Constrained Minima

Inequality Constraints and the KKT Theorem

Constrained Minima

Inequality constrained optimization

An ***inequality constrained minimization problem*** with objective $f : \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0 \\ & g_1(\mathbf{x}) \leq 0, \ldots, g_r(\mathbf{x}) \leq 0 \end{aligned}$$

where $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$ are $\mathscr{C}^1$ and $g_1(\mathbf{x}), \ldots, g_r(\mathbf{x})$ are $\mathscr{C}^1$.

Constrained Minima

Inequality constrained optimization

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0 \\ & g_1(\mathbf{x}) \leq 0, \ldots, g_r(\mathbf{x}) \leq 0 \end{aligned}$$

INEQUALITY

$\Downarrow$

EQUALITY

$\Downarrow$

UNCONSTRAINED

Main idea: Reduce to *equality constrained optimization.*

The only difference is that each *inequality constraint* can either be ***active*** or not.

BINDING.

A constraint $j \in [r]$ is ***active*** if $g_j(\mathbf{x}) = 0$.

Constrained Minima: Inequality Constraints

Definition of active constraints

For feasible $\mathbf{x} \in \mathbb{R}^d$ the set of ***active inequality constraints*** is

$$\boxed{\mathscr{A}(\mathbf{x}) := \{j : g_j(\mathbf{x}) = 0\} \subseteq [r] \,.}$$

This means we get a new definition for a *regular point*…

A point $\mathbf{x} \in \mathbb{R}^d$ is a ***regular point*** if it is feasible and the gradients

$$\underbrace{\{\nabla h_1(\mathbf{x}), \ldots, \nabla h_m(\mathbf{x})\}}\cup\underbrace{\{\nabla g_j(\mathbf{x}) : j \in \mathscr{A}(\mathbf{x})\}}$$

are linearly independent.

Constrained Minima: Inequality Constraints

Lagrangian in Inequality Constrained Optimization

For an optimization problem with equality *and* inequality constraints

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0 \\ & g_1(\mathbf{x}) \le 0, \ldots, g_r(\mathbf{x}) \le 0 \end{aligned}$$

the ***Lagrangian function*** $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is the function

$$L(\mathbf{x}, \lambda, \mu) := f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \mu_j g_j(\mathbf{x}) = \underbrace{f(\mathbf{x}) + \lambda^\top \mathbf{h}(\mathbf{x})} + \underbrace{\mu^\top \mathbf{g}(\mathbf{x})}.$$

Notice that the function $L(\mathbf{x}, \lambda, \mu)$ is an *unconstrained* function.

Constrained Minima: Inequality Constraints

Karush-Kuhn-Tucker (KKT) Theorem

(NECESSARY CONDITIONS)

Theorem (KKT Theorem). Let $\mathbf{x}^* \in \mathbb{R}^d$ be a local minimum that is a regular point. Then, there exists unique vectors $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^r$ called ***Lagrange multipliers*** such that

$\nabla h_1, \ldots, \nabla h_m$ and active $\nabla g_1, \ldots, \nabla g_r$.

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0,$$

where $\mu_j^* \geq 0$ for all $j \in [r]$ and $\mu_j^* = 0$ for all non-active constraints $j \notin \mathscr{A}(\mathbf{x}^*)$ (***complementary slackness***).

$g_j(x^*) < 0 \iff \mu_j^* = 0.$

$g_j(x^*) = 0 \iff \mu_j \in \mathbb{R}.$

If, in addition, $f(\cdot)$ and $h(\cdot)$ are twice continuously differentiable, $g(\cdot)$

$$\mathbf{d}^\top \left(\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*) \right) \mathbf{d} \geq 0$$

$\sum_{i=1}^{r} \mu_i \nabla^2 g_i(x^*)$

for all $\mathbf{d} \in \mathbb{R}^d$ such that $\nabla \mathbf{h}(\mathbf{x}^*)^\top \mathbf{d} = 0$, where $\nabla \mathbf{h}(\mathbf{x}^*) \in \mathbb{R}^{d \times m}$ is the Jacobian of $\mathbf{h}$ at $\mathbf{x}^*$.

should include g.

Constrained Minima: Inequality Constraints

Karush-Kuhn-Tucker (KKT) Theorem

For the Lagrangian,

$$L(\mathbf{x}, \lambda, \mu) := f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \mu_j g_j(\mathbf{x}),$$

we can write the previous necessary conditions at the local optimum $(\mathbf{x}^*, \lambda^*, \mu^*)$ as:

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*, \mu^*) = 0, \ \mathbf{h}(\mathbf{x}^*) = 0, \ \mathbf{g}(\mathbf{x}^*) \leq 0$$

where we *also* require the *complementary slackness conditions*:

$$\mu^* \geq 0 \text{ and } \mu_j^* g_j(\mathbf{x}^*) = 0, \ \forall j \in [r].$$

Constrained Minima: Inequality Constraints

Karush-Kuhn-Tucker (KKT) Theorem: Sufficient Conditions

Theorem (KKT Theorem - Sufficient Conditions). Let f, $\mathbf{h}$, and $\mathbf{g}$ be $\mathscr{C}^2$ functions, such that $\mathbf{x}^* \in \mathbb{R}^d$, $\lambda \in \mathbb{R}^m$, $\mu^* \in \mathbb{R}^r$ satisfy

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*, \mu^*) = 0,\ \mathbf{h}(\mathbf{x}^*) = 0,\ \mathbf{g}(\mathbf{x}^*) \le 0$$

comp. slackness. $\longrightarrow$ $\mu^* \geq 0$ **and** $\mu_j^* g_j(\mathbf{x}^*) = 0,\ \forall j \in [r]$

$$\mathbf{d}^\top \nabla^2_{\mathbf{x},\mathbf{x}} L(\mathbf{x}^*, \lambda^*, \mu^*)\mathbf{d} \overset{\text{PD}}{>} 0,$$

$\Downarrow$ LOCAL MIN.

for all $\mathbf{d}$ such that $\nabla \mathbf{h}(\mathbf{x}^*)^\top \mathbf{d} = \mathbf{0}$ and $\nabla g_j(\mathbf{x}^*)^\top \mathbf{d} = 0,\ \forall j \in \mathscr{A}(\mathbf{x}^*)$

Then, $\mathbf{x}^*$ is a local minimum.

Constrained Minima: Inequality Constraints

How do we use the Lagrangian?

Assuming that *a global minimum exists* and f, $\mathbf{h}$, and $\mathbf{g}$ are $\mathscr{C}^1$, let the Lagrangian be:

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \mu_j g_j(\mathbf{x})$$

To find a global minimum…

1. Find the set $(\mathbf{x}^*, \lambda^*, \mu^*)$ satisfying the necessary conditions:
   $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*, \mu^*) = 0,\ \mathbf{h}(\mathbf{x}^*) = 0,\ \mathbf{g}(\mathbf{x}^*) \leq 0$ (*first-order conditions*)
   $\mu^* \geq 0$ and $\mu_j^* g_j(\mathbf{x}^*) = 0,\ \forall j \in [r]$ (*complementary slackness*)
2. Find the set of all non-regular points.
3. The global minima must be among the points in (1) or (2).

Constrained Minima: Inequality Constraints

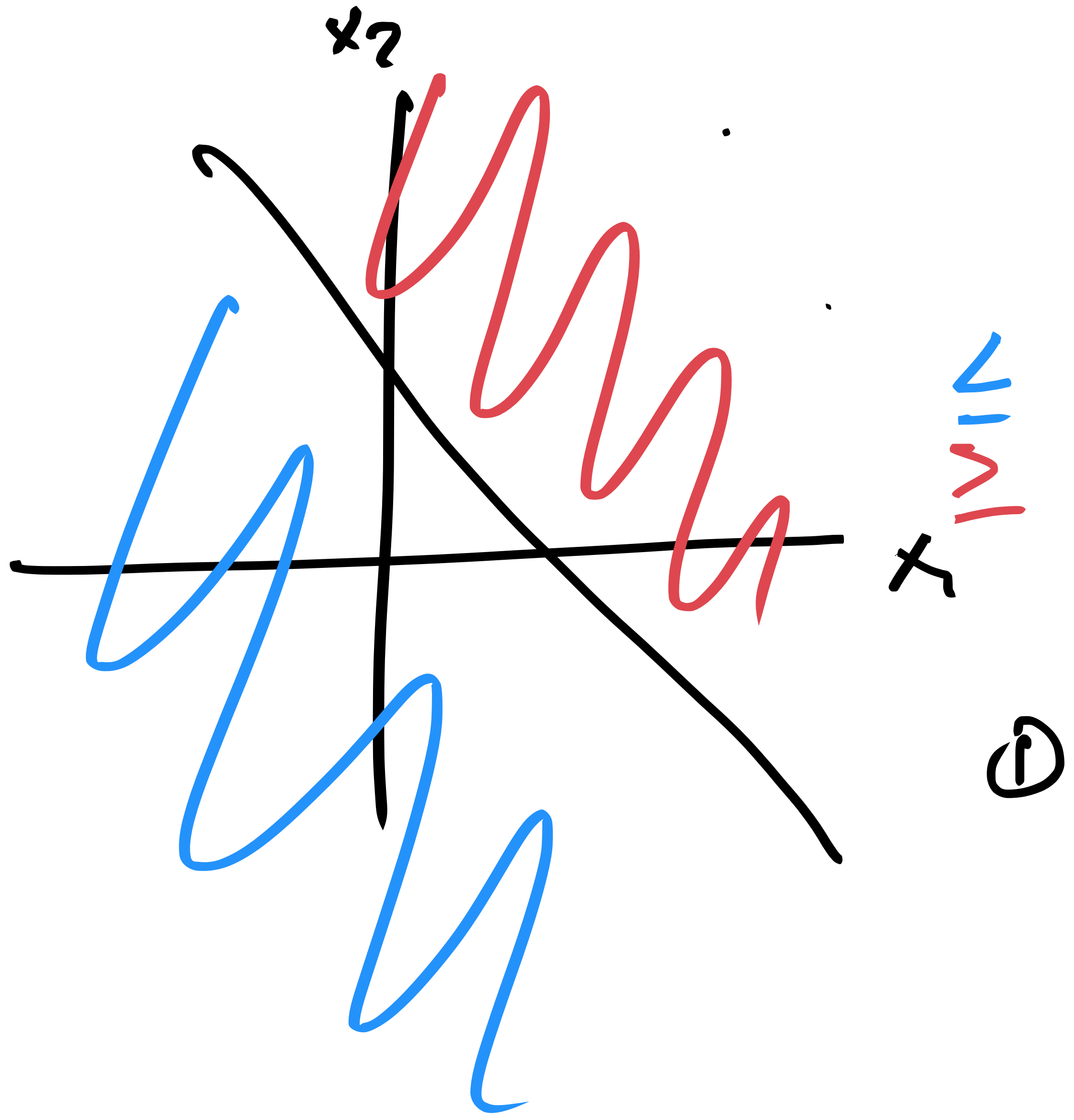
Example: Smallest point in a halfspace

Consider the following optimization problem over $\mathbf{x} \in \mathbb{R}^3$:

$$\boxed{x^* = (-1,-1,-1)}$$

$$x_1 = x_2 = x_3 = -1$$

$$\boxed{f(x^*) = 3/2}$$



$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 \\ \text{subject to} \quad & x_1 + x_2 + x_3 \le -3 \end{aligned}$$

$\frac{1}{2}\|x\|^2$

$$[1 \ 1 \ 1]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + 3 \le 0.$$

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

$w^T x \le b$

① LAGRANGIAN. $g_1(x) = x_1 + x_2 + x_3 + 3 \le 0.$

$$L(x, M) = \frac{1}{2}\|x\|^2 + M(x_1 + x_2 + x_3 + 3) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + M(x_1 + x_2 + x_3 + 3)$$

$$\nabla_x L = \begin{bmatrix} x_1 + M \\ x_2 + M \\ x_3 + M \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \nabla_M L = x_1 + x_2 + x_3 + 3 = 0.$$

$$\Rightarrow \begin{aligned} x_1 + M &= 0 \\ x_2 + M &= 0 \\ x_3 + M &= 0 \\ x_1 + x_2 + x_3 &= -3 \end{aligned} \qquad \begin{aligned} & M = -x_1 = -x_2 = -x_3 \\ & x_1 = x_2 = x_3 \\ & \Rightarrow x_1 = x_2 = x_3 = \boxed{-1} \\ & \boxed{M = 1} \end{aligned}$$

Complementary Slackness: $x_1 = x_2 = x_3$

$-1 - 1 - 1 = -3 \longrightarrow$ ✓

Least Squares Regression

Regularization and Ridge Regression

Regression

Setup

Observed: Matrix of *training samples* $\mathbf{X} \in \mathbb{R}^{n \times d}$ and vector of *training labels* $\mathbf{y} \in \mathbb{R}^d$.

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \dots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}.$$

Unknown: *Weight vector* $\mathbf{w} \in \mathbb{R}^d$ with weights $w_1, \dots, w_d$.

Goal: For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \dots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

Regression
Setup

<u>Goal:</u> For each $i \in [n]$, we predict: $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i = w_1 x_{i1} + \ldots + w_d x_{id} \in \mathbb{R}$.

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\mathbf{X}\hat{\mathbf{w}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

To find $\hat{\mathbf{w}}$, we follow the *principle of least squares.*

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

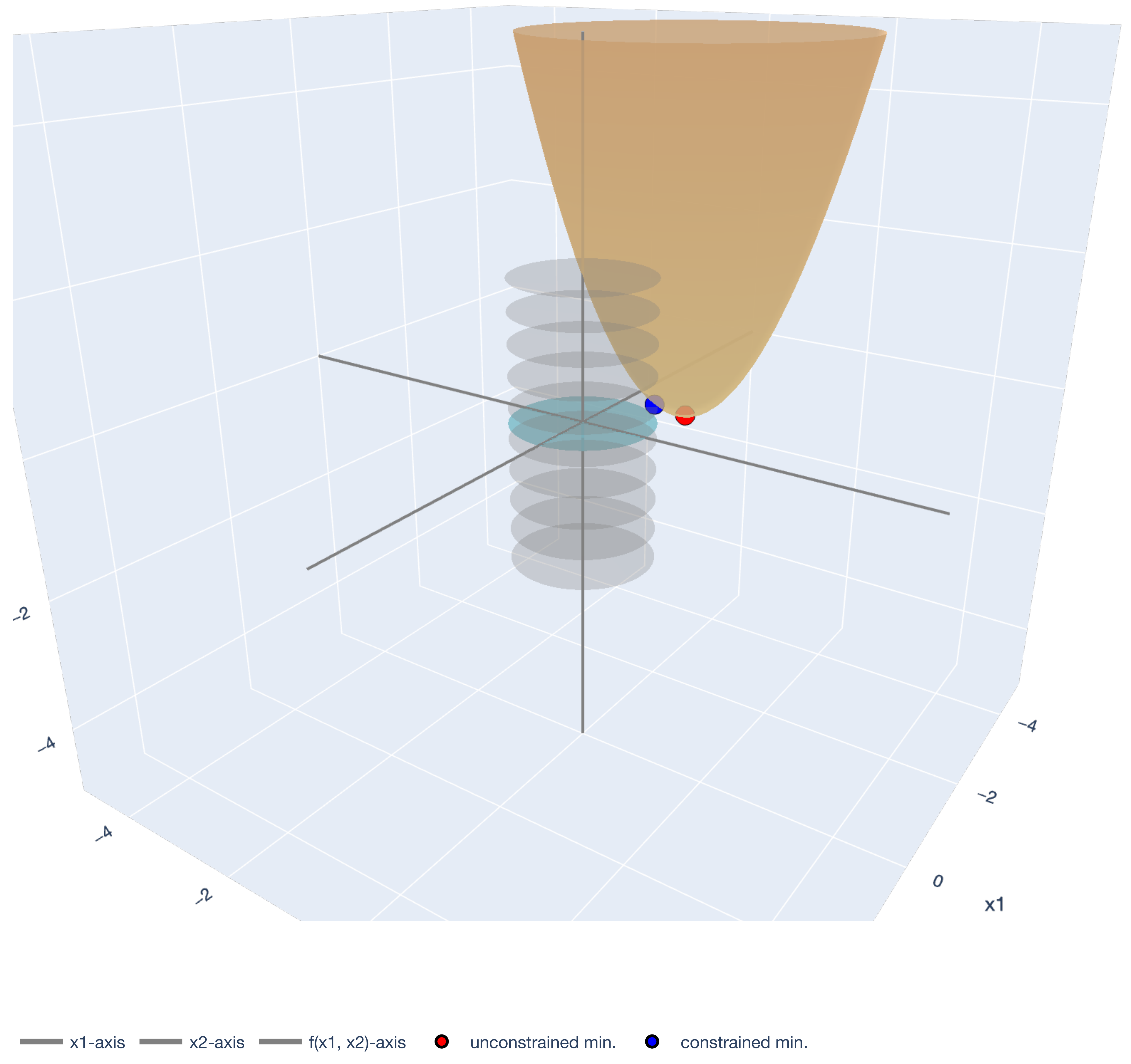
Regression

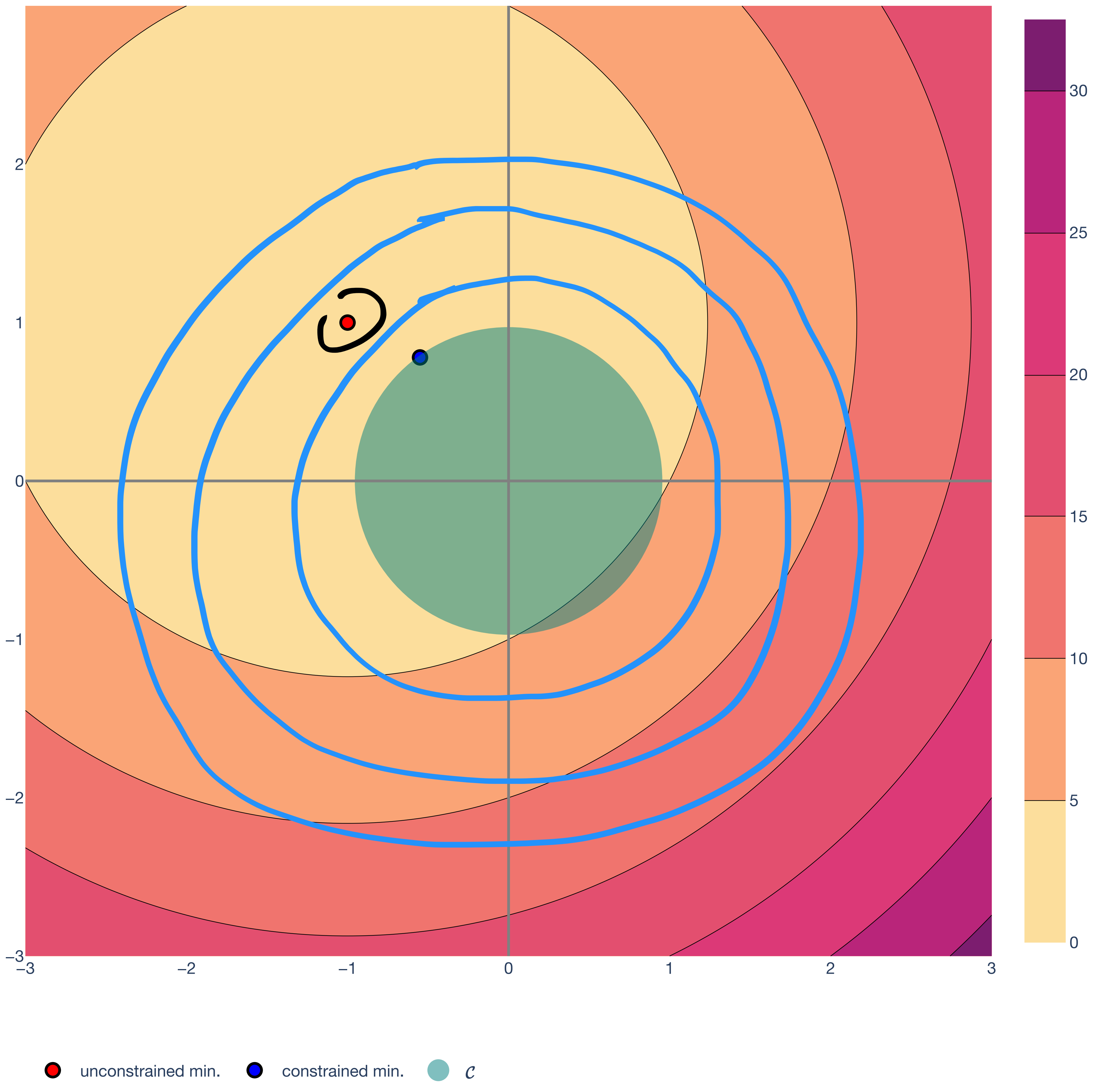
“Regularization” and keeping $\|\mathbf{w}\|$ small

One reasonable

Lesson Overview

Big Picture: Least Squares





Least Squares

Least norm exact solution

OLS $n \geq d$

$d \geq n$

EXACT SOLUTIONS (PSEUDOINVERSE)

For $\mathbf{X} \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}(\mathbf{X}) = n$,

$w = (w_1, \ldots, w_d)$

1000 -900

1   -0.9

$$\boxed{\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad \|\mathbf{w}\|}$$
$$\text{subject to} \quad \mathbf{X}\mathbf{w} = \mathbf{y}$$

SMALL NORM SOLUTIONS = MORE STABLE!

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\operatorname{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

We already know how to solve this — use the pseudoinverse!

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n\times d}$ with $\operatorname{rank}(\mathbf{X}) = n$, → we have exact solution.

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

$X\hat{w} = y.$

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$, let $d \geq n$, and let $\operatorname{rank}(\mathbf{X}) = n$. Then, $\hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{y}$ is the exact solution $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$ with smallest Euclidean norm:

$$\|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\mathrm{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\|^2 \\ \text{subject to} \quad & \mathbf{Xw} = \mathbf{y} \end{aligned}$$

$$Xw = y \rightarrow \left.\begin{array}{l} x_1^T w = y_1 \\ x_2^T w = y_2 \\ \vdots \\ x_n^T w = y_n \end{array}\right\} \begin{array}{c} n \\ \text{consr} \end{array}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \underline{\|\mathbf{w}\|^2} + \underline{\lambda^\top(\mathbf{Xw} - \mathbf{y})}$$

Least Squares

Least norm exact solution

x^2

$\nabla_w w^T w = 2w.$

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\mathrm{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\|^2 \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \|\mathbf{w}\|^2 + \lambda^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

First-order conditions: $\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 2\mathbf{w} + \mathbf{X}^\top \lambda$ and $\nabla_\lambda L(\mathbf{w}, \lambda) = \mathbf{X}\mathbf{w} - \mathbf{y}$.

Setting equal to zero: $2\mathbf{w} + \mathbf{X}^\top \lambda = \mathbf{0}$ and $\mathbf{X}\mathbf{w} - \mathbf{y} = \mathbf{0}$

Least Squares

Least norm exact solution

$$\begin{array}{ll} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} & \|\mathbf{w}\| \\ \text{subject to} & \mathbf{X}\mathbf{w} = \mathbf{y} \end{array}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \|\mathbf{w}\| + \lambda^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

First-order conditions: $\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 2\mathbf{w} + \mathbf{X}^\top \lambda$ and $\nabla_{\lambda} L(\mathbf{w}, \lambda) = \mathbf{X}\mathbf{w} - \mathbf{y}$.

Setting equal to zero: $2\mathbf{w} + \mathbf{X}^\top \lambda = \mathbf{0}$ and $\mathbf{X}\mathbf{w} - \mathbf{y} = \mathbf{0}$

$$\implies \boxed{\mathbf{w} = -\frac{1}{2}\mathbf{X}^\top \lambda} \text{ and } \boxed{\mathbf{X}\mathbf{w} = \mathbf{y}}$$

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\text{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \|\mathbf{w}\| + \lambda^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

First-order conditions: $\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 2\mathbf{w} + \mathbf{X}^\top \lambda$ and $\nabla_\lambda L(\mathbf{w}, \lambda) = \mathbf{X}\mathbf{w} - \mathbf{y}$.

Setting equal to zero: $2\mathbf{w} + \mathbf{X}^\top\lambda = \mathbf{0}$ and $\mathbf{X}\mathbf{w} - \mathbf{y} = \mathbf{0} \implies \mathbf{w} = -\frac{1}{2}\mathbf{X}^\top\lambda$ and $\mathbf{X}\mathbf{w} = \mathbf{y}$

Solve for λ*:* $\mathbf{X}\mathbf{w} = -\frac{1}{2}\mathbf{X}\mathbf{X}^\top\lambda \implies -\frac{1}{2}(\mathbf{X}\mathbf{X}^\top)\lambda = \mathbf{y} \implies \lambda = -2(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}.$

$\text{rank}(X) = n$
$XX^T \in \mathbb{R}^{n \times n}$

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\text{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \|\mathbf{w}\| + \lambda^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

First-order conditions: $\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 2\mathbf{w} + \mathbf{X}^\top \lambda$ and $\nabla_\lambda L(\mathbf{w}, \lambda) = \mathbf{X}\mathbf{w} - \mathbf{y}$.

Setting equal to zero: $2\mathbf{w} + \mathbf{X}^\top\lambda = \mathbf{0}$ and $\mathbf{X}\mathbf{w} - \mathbf{y} = \mathbf{0} \implies \mathbf{w} = -\frac{1}{2}\mathbf{X}^\top\lambda$ and $\mathbf{X}\mathbf{w} = \mathbf{y}$

$d \geq n$
For $\text{rank}(X) = n$ $\}$ Pseudo inverse w/ SVD $\boxed{\Sigma^+}$

Solve for λ*:* $\mathbf{X}\mathbf{w} = -\frac{1}{2}\mathbf{X}\mathbf{X}^\top\lambda \implies -\frac{1}{2}(\mathbf{X}\mathbf{X}^\top)\lambda = \mathbf{y} \implies \lambda = -2(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$.

Plug λ *back in to solve for* $\mathbf{w}$: $\mathbf{w} = \underline{-\frac{1}{2}\mathbf{X}^\top\lambda} = \underline{-\frac{1}{2}\mathbf{X}^\top\left(-2(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}\right)} \implies \boxed{\mathbf{w} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = \mathbf{X}^+\mathbf{y}.}$ *The pseudoinverse!*

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\operatorname{rank}(\mathbf{X}) = n$,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

Alternate proof (through Lagrangian): For Lagrange multipliers $\lambda \in \mathbb{R}^n$,

$$L(\mathbf{w}, \lambda) = \|\mathbf{w}\| + \lambda^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

First-order conditions: $\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 2\mathbf{w} + \mathbf{X}^\top \lambda$ and $\nabla_\lambda L(\mathbf{w}, \lambda) = \mathbf{X}\mathbf{w} - \mathbf{y}$.

Setting equal to zero: $2\mathbf{w} + \mathbf{X}^\top \lambda = \mathbf{0}$ and $\mathbf{X}\mathbf{w} - \mathbf{y} = \mathbf{0} \implies \mathbf{w} = -\dfrac{1}{2}\mathbf{X}^\top \lambda$ and $\mathbf{X}\mathbf{w} = \mathbf{y}$

Solve for λ: $\mathbf{X}\mathbf{w} = -\dfrac{1}{2}\mathbf{X}\mathbf{X}^\top \lambda \implies -\dfrac{1}{2}(\mathbf{X}\mathbf{X}^\top)\lambda = \mathbf{y} \implies \lambda = -2(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$.

Plug λ back in to solve for $\mathbf{w}$: $\mathbf{w} = -\dfrac{1}{2}\mathbf{X}^\top \lambda = -\dfrac{1}{2}\mathbf{X}^\top \left(-2(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}\right) \implies \mathbf{w} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = \mathbf{X}^+\mathbf{y}$. *The pseudoinverse!*

Least Squares

Least norm exact solution

For $\mathbf{X} \in \mathbb{R}^{n\times d}$ with $\mathrm{rank}(\mathbf{X}) = n$,

LAGRANGIAN

$$\begin{aligned} \underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & \mathbf{X}\mathbf{w} = \mathbf{y} \end{aligned}$$

$$\boxed{\begin{array}{l}\textbf{Theorem (Minimum norm least squares solution).}\ \text{Let } \mathbf{X} \in \mathbb{R}^{n\times d}\text{, let } d \geq n\text{, and let } \mathrm{rank}(\mathbf{X}) = n\text{. Then, } \hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{+}\mathbf{U}^{\top}\mathbf{y} \\ \text{is the exact solution } \mathbf{X}\hat{\mathbf{w}} = \mathbf{y} \text{ with smallest Euclidean norm:} \\ \qquad \|\mathbf{w}\|_2^2 \geq \|\hat{\mathbf{w}}\|_2^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d. \end{array}}$$

How about for the approximate solution to $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$*?*

Least Squares

Ridge Regression

Our goal will now be to minimize two objectives:

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

Writing this as an optimization problem:

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

$\gamma \to \infty \implies$ All we care about is $\|w\|$.

$\gamma = 0 \implies$ Back to OLS.

where $\gamma > 0$ is a fixed tuning parameter. This optimization problem is known as ***ridge/Tikhonov/ℓ_2-regularized regression.***

Least Squares

Ridge Regression

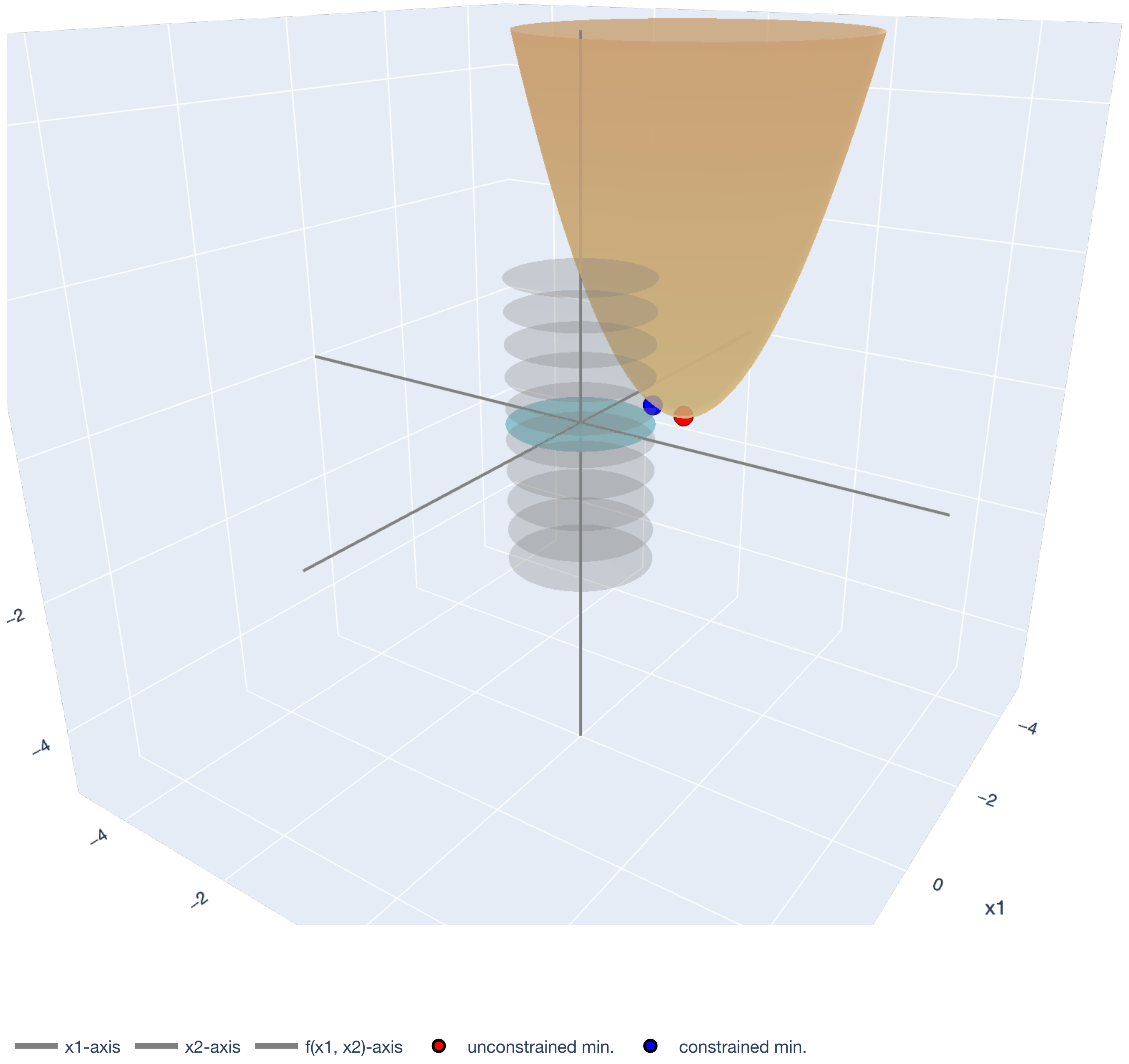
Our goal will now be to minimize two objectives:

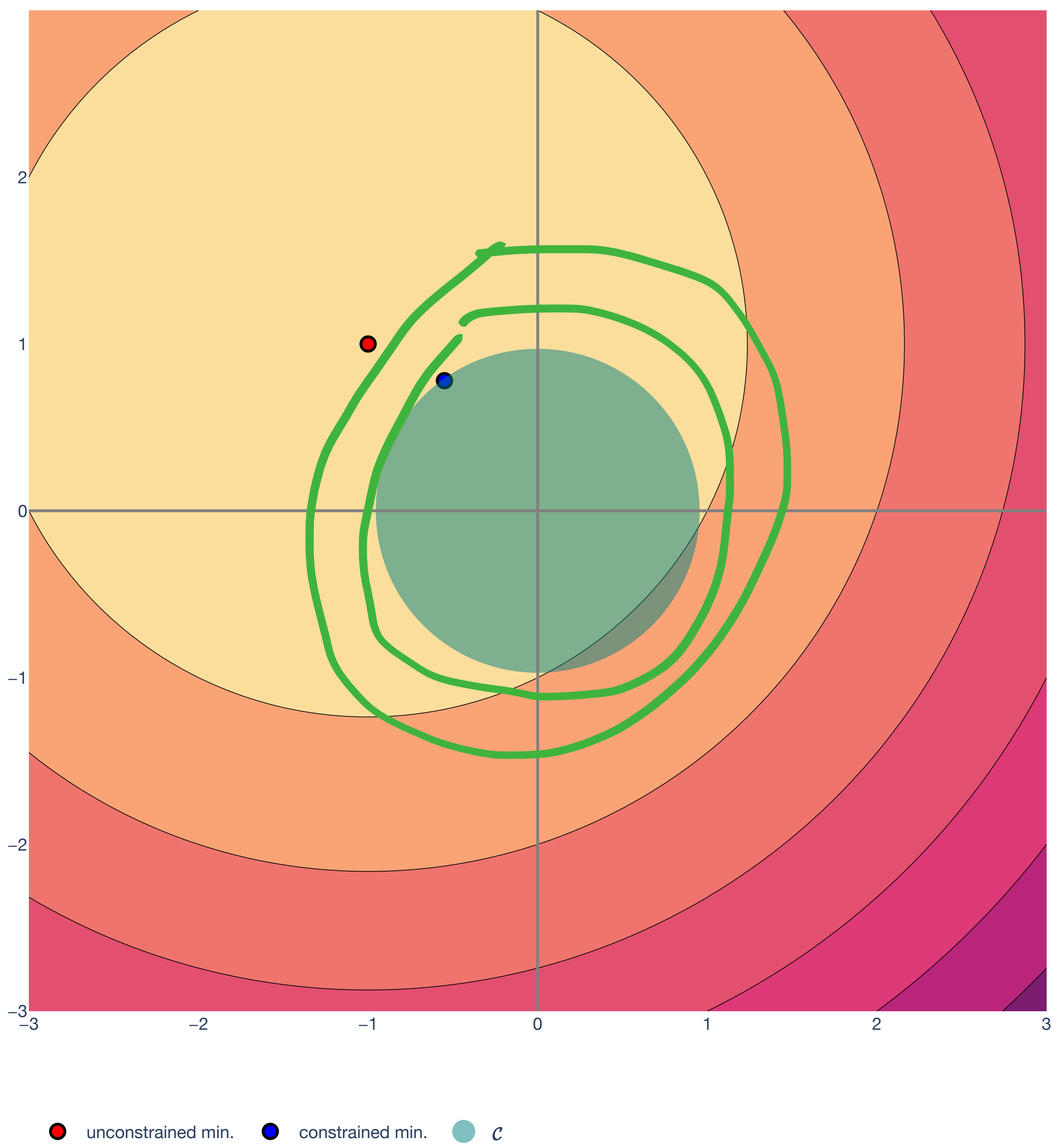
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

Writing this as an optimization problem:

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

where $\gamma > 0$ is a fixed tuning $\gamma(\|w\|^2 - 2)$ parameter. This optimization problem is known as ***ridge/Tikhonov/ℓ_2-regularized regression.***





Least Squares

Ridge Regression

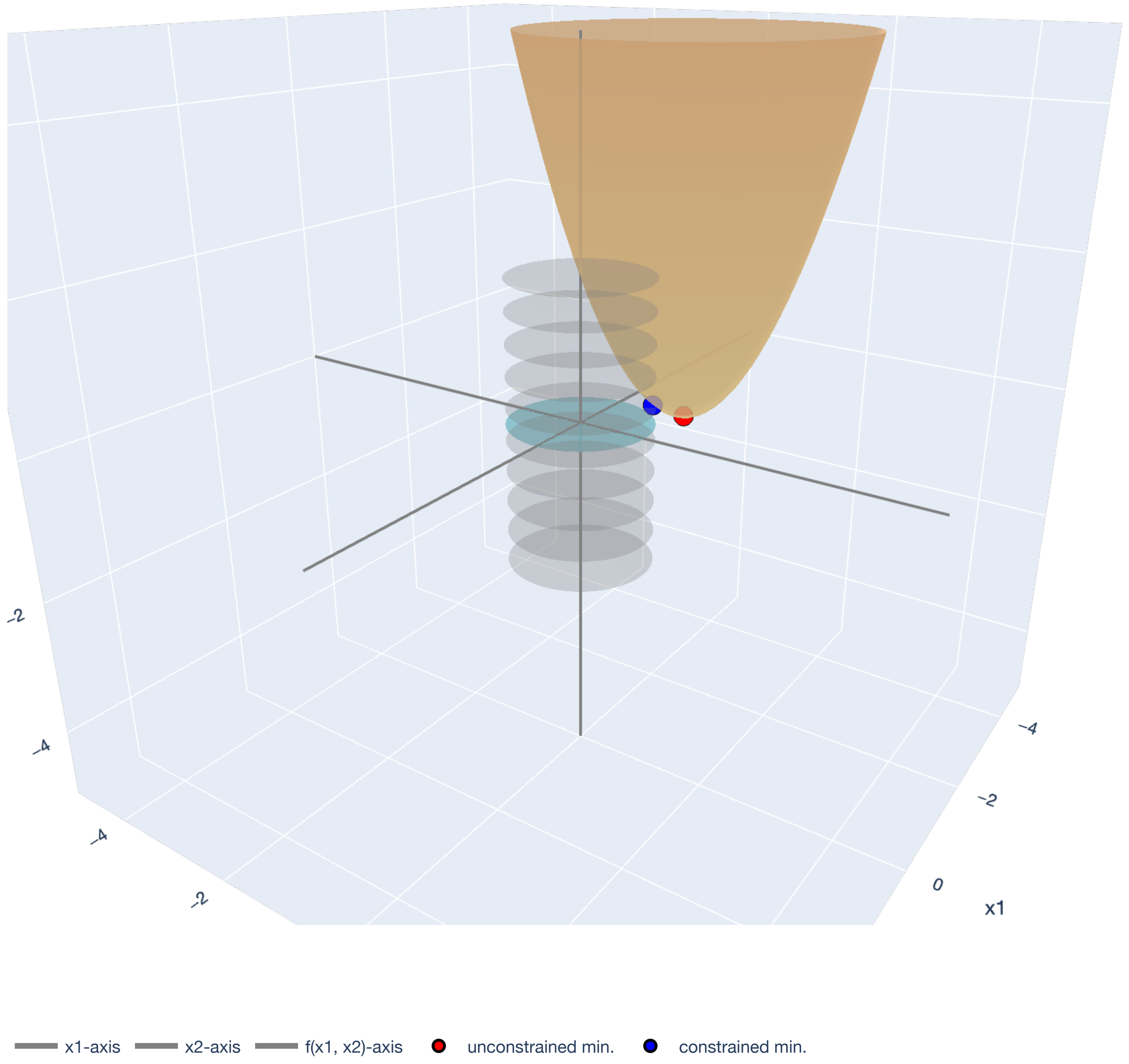
Our goal will now be to minimize two objectives:

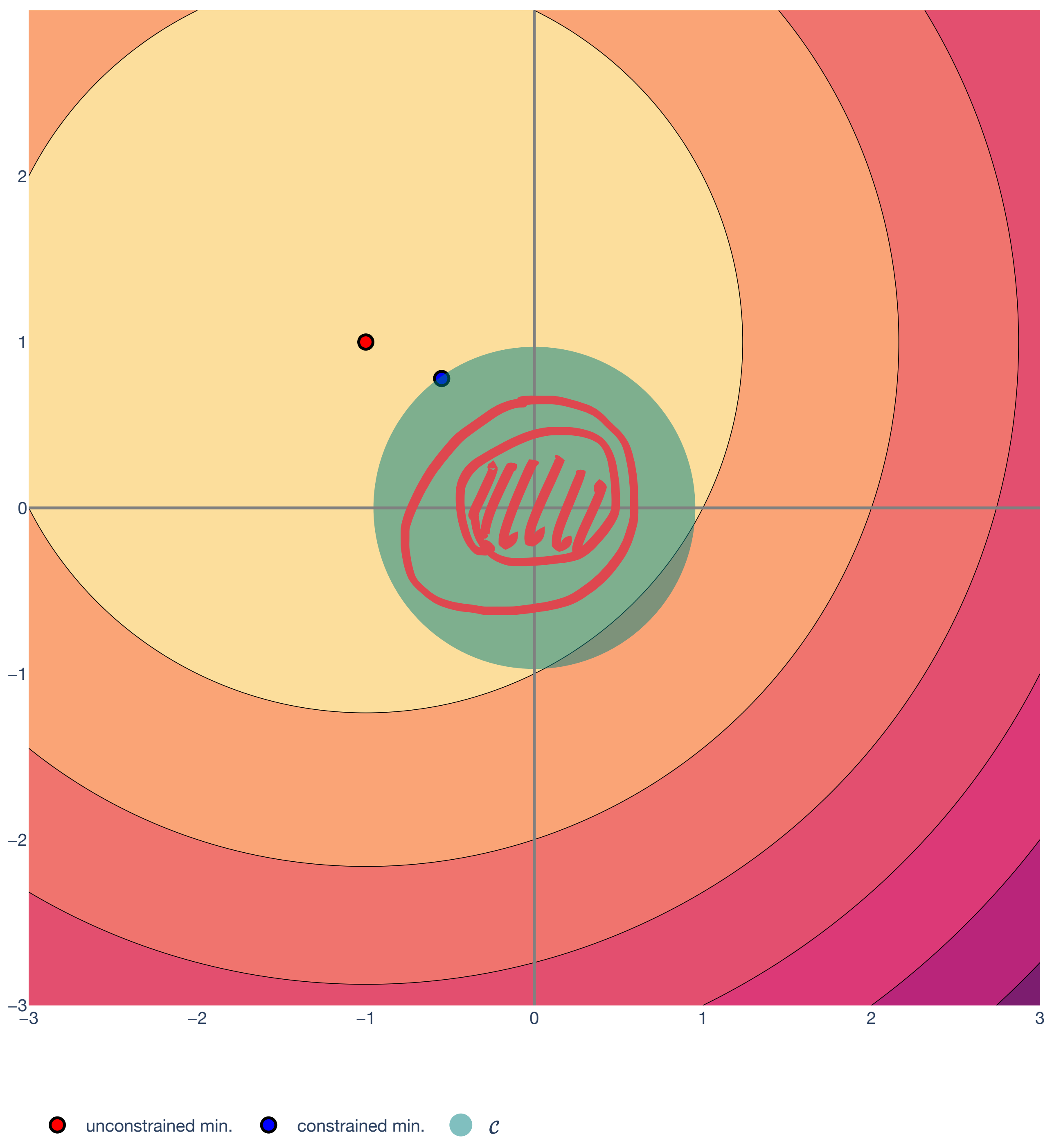
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

Writing this as an optimization problem:

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

where $\gamma > 0$ is a fixed tuning parameter. This optimization problem is known as ***ridge/Tikhonov/ℓ_2-regularized regression.***





For bigger γ, ~~bigger~~ smaller “constraint” ball!

Least Squares

Solving ridge regression

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

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

Least Squares

Solving ridge regression

$$\gamma I = \begin{bmatrix} \gamma & & & 0 \\ & \gamma & & \\ & & \ddots & \\ 0 & & & \gamma \end{bmatrix}$$

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

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

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

Least Squares

Solving ridge regression

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

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

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

Proof. Let $\mathbf{v} \in \mathbb{R}^d$ be any vector.

linearity

$$\mathbf{v}^\top(\mathbf{A}+\gamma\mathbf{I})\mathbf{v} = \mathbf{v}^\top(\mathbf{A}\mathbf{v}+\gamma\mathbf{v}) = \mathbf{v}^\top\mathbf{A}\mathbf{v} + \gamma\mathbf{v}^\top\mathbf{v}$$

$v^T v = \|v\|^2$.

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

PSD

Least Squares

Solving ridge regression

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

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

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

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

Least Squares

Solving ridge regression

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

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

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

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

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

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

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

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

Least Squares

Solving ridge regression

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

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

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

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

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

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

Taking the Hessian,

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

Sufficient condition for optimality applies!

Least Squares

Solving ridge regression

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

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

has the form:

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

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

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

Least Squares

Solving ridge regression

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

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

has the form:

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

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

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

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

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

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

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

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

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

Error in (OLS) Regression

Error using least squares model

Choose a weight vector that "fits the training data": $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $y_i \approx \hat{y}_i$ for $i \in [n]$, or:

$$\boxed{\mathbf{X}\hat{\mathbf{w}}} = \hat{\mathbf{y}} \approx \mathbf{y}\,.$$

But $\hat{\mathbf{y}}$ might not be a perfect fit to $\mathbf{y}$!

Model this using a *true weight vector* $\mathbf{w}^* \in \mathbb{R}^d$ and an *error term* $\underbrace{\epsilon = (\epsilon_1, \ldots, \epsilon_n)}$ $\in \mathbb{R}^n$.

$$y_i = \mathbf{x}_i^\top \mathbf{w}^* + \epsilon_i \text{ for all } i \in [n]$$

$$\underbrace{\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon} \leftarrow \varepsilon \in \mathbb{R}^n$$

Error in (OLS) Regression

Error using least squares model

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the OLS weights $\boxed{\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}?}$

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

Error.

Error in (OLS) Regression

Error using least squares model

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the OLS weights $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$?

$$
\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \epsilon) \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon \\
&= \mathbf{w}^* + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon \longrightarrow 0
\end{aligned}
$$

When $\epsilon = 0$ ($\mathbf{y}$ is linearly related to $\mathbf{X}$), this is perfect: $\hat{\mathbf{w}} = \mathbf{w}^*$!

$\hat{w} = w^*$

Error in (OLS) Regression

Error using least squares model

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the OLS weights $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$?

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

When $\epsilon \neq 0$, we have an error of $\hat{\mathbf{w}} - \mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon$.

Error in (OLS) Regression

Eigendecomposition perspective

$\hat{w} = w^* + (X^T X)^{-1} X^T \varepsilon.$

Weight vector's error: $\hat{\mathbf{w}} - \mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon.$

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

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

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

$\lambda_i \to 0$

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\lambda_d \end{bmatrix}$$, so if λ_i is small, the entries of $\hat{\mathbf{w}}$ blow up!

(For some directions of the error)

Error in Regression

Error using ridge regression

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the ***ridge weights*** $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$?

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

Error in Regression

Error using ridge regression

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the ***ridge weights*** $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$?

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

When $\epsilon = 0$ ($\mathbf{y}$ is linearly related to $\mathbf{X}$), this is no longer perfect:

$$\hat{\mathbf{w}} = \underbrace{(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}}\mathbf{w}^*, \text{ but...}$$

(no longer exactly w^*).

Error in Regression

Error using ridge regression

True labels: $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$.

What happens when we use the ***ridge weights*** $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$?

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

When $\epsilon \neq 0$, we have more stable errors!

Error in Ridge Regression

Eigendecomposition perspective

$$\lambda + \gamma I$$

$$\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix} + \begin{bmatrix} \gamma & & 0 \\ & \ddots & \\ 0 & & \gamma \end{bmatrix} = \begin{bmatrix} \lambda_1+\gamma & & 0 \\ & \ddots & \\ 0 & & \lambda_d+\gamma \end{bmatrix}$$

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

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

$VV^T = I$

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

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

$\gamma \to \infty$

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

$$\boxed{\frac{1}{\lambda_i+\gamma} \leq \frac{1}{\gamma}}$$

$\lambda_i \to 0$

★ DO NOT MAGNIFY ERRORS!

Least Squares

Ridge Regression

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

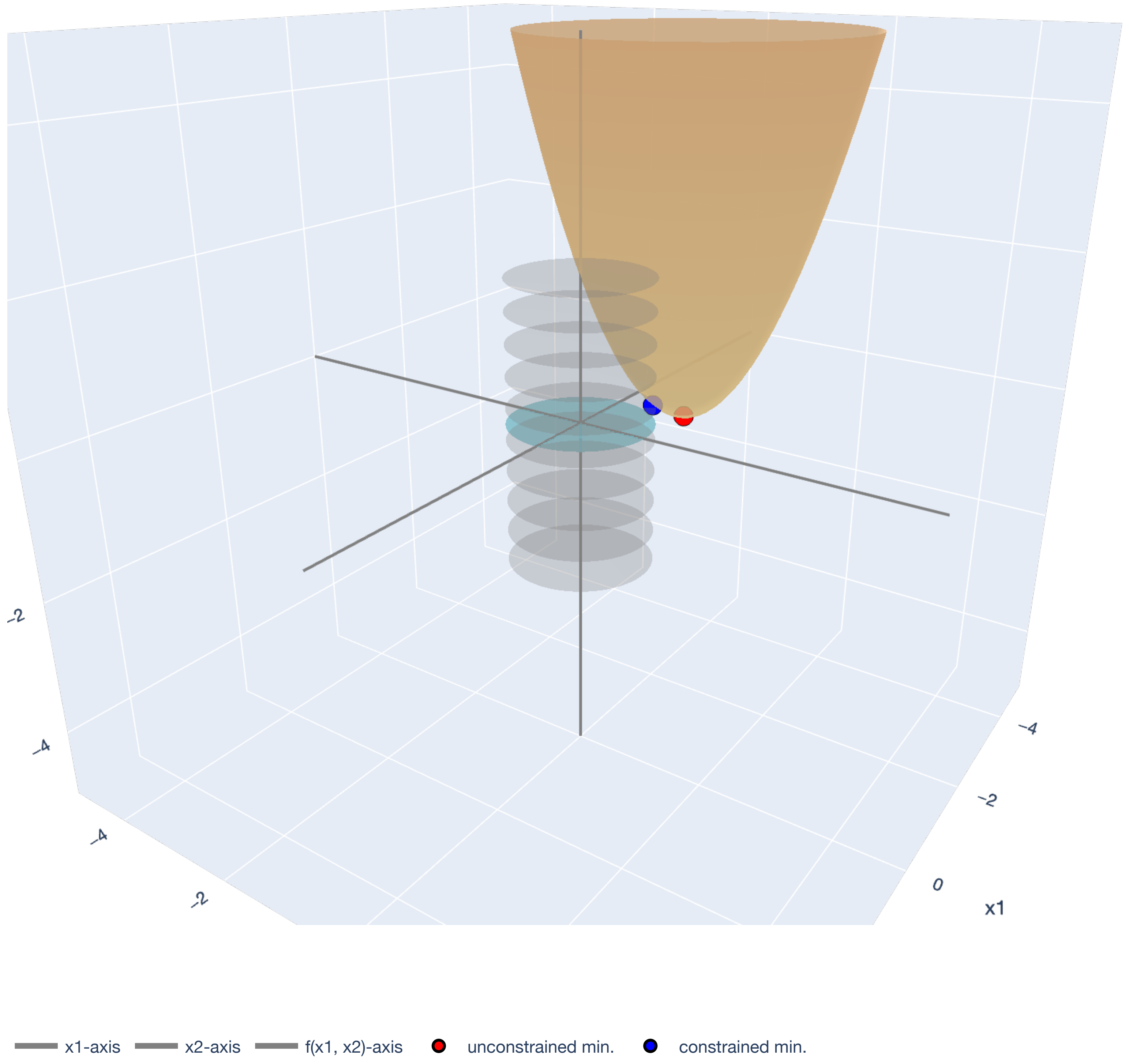
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2$$

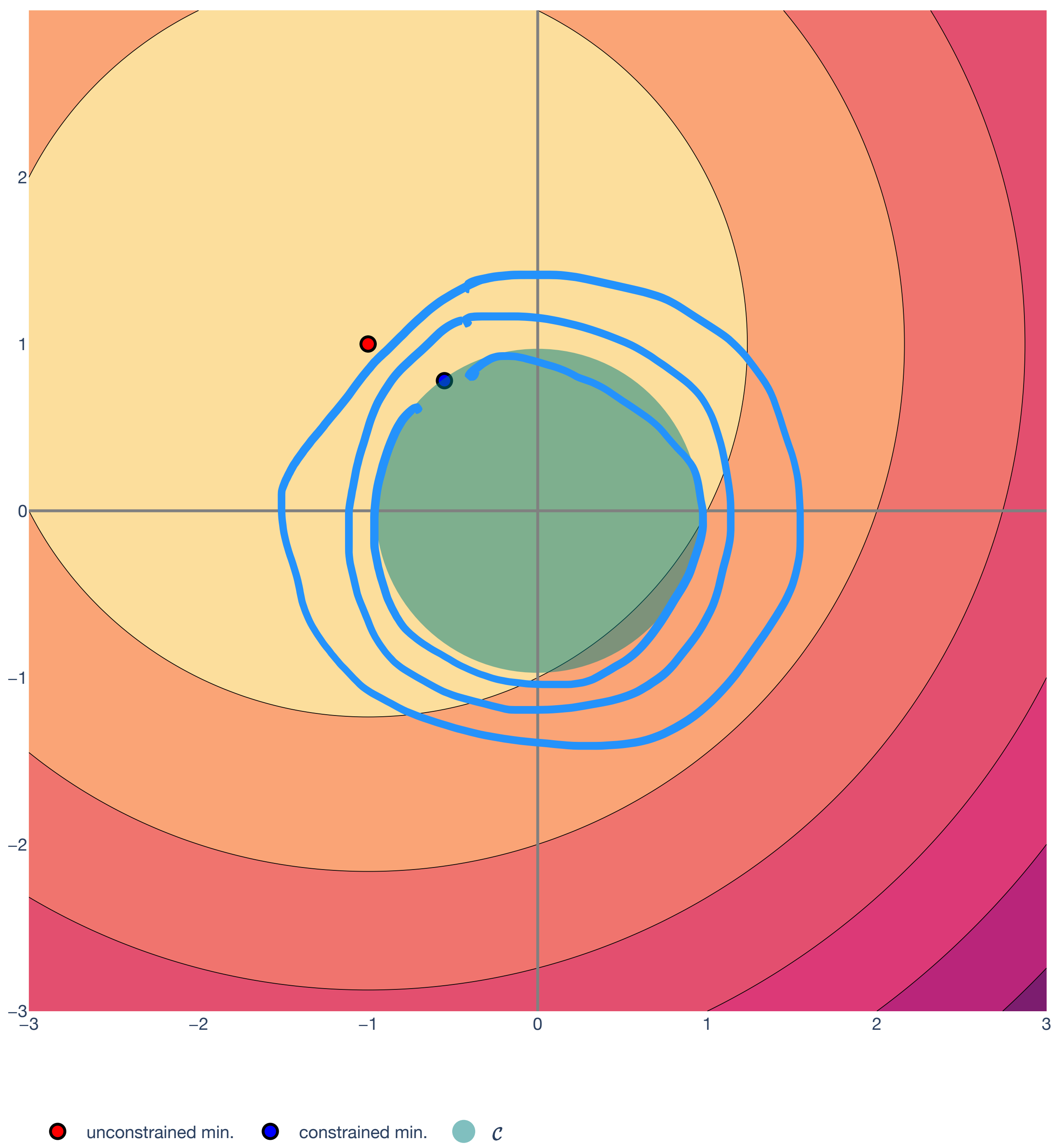
has the form:

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

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

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





For bigger γ, ~~bigger~~ smaller "constraint" ball!

Recap

Lesson Overview

EQUALITY CONSTRAINED ← INEQUALITY CONSTRAINED

LAGRANGIAN.

UNCONSTRAINED

Optimization. Minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$ with the possible requirement that the minimizer $\mathbf{x}^*$ belongs to a constraint set $\mathscr{C} \subseteq \mathbb{R}^d$.

Lagrangian. For optimization problems with $\mathscr{C}$ defined by equalities/inequalities, the Lagrangian is a function $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ that "unconstrains" the problem.

Unconstrained local optima. With no constraints, the standard tools of calculus give conditions for a point $\mathbf{x}^*$ to be optimal, at least to all points close to it.

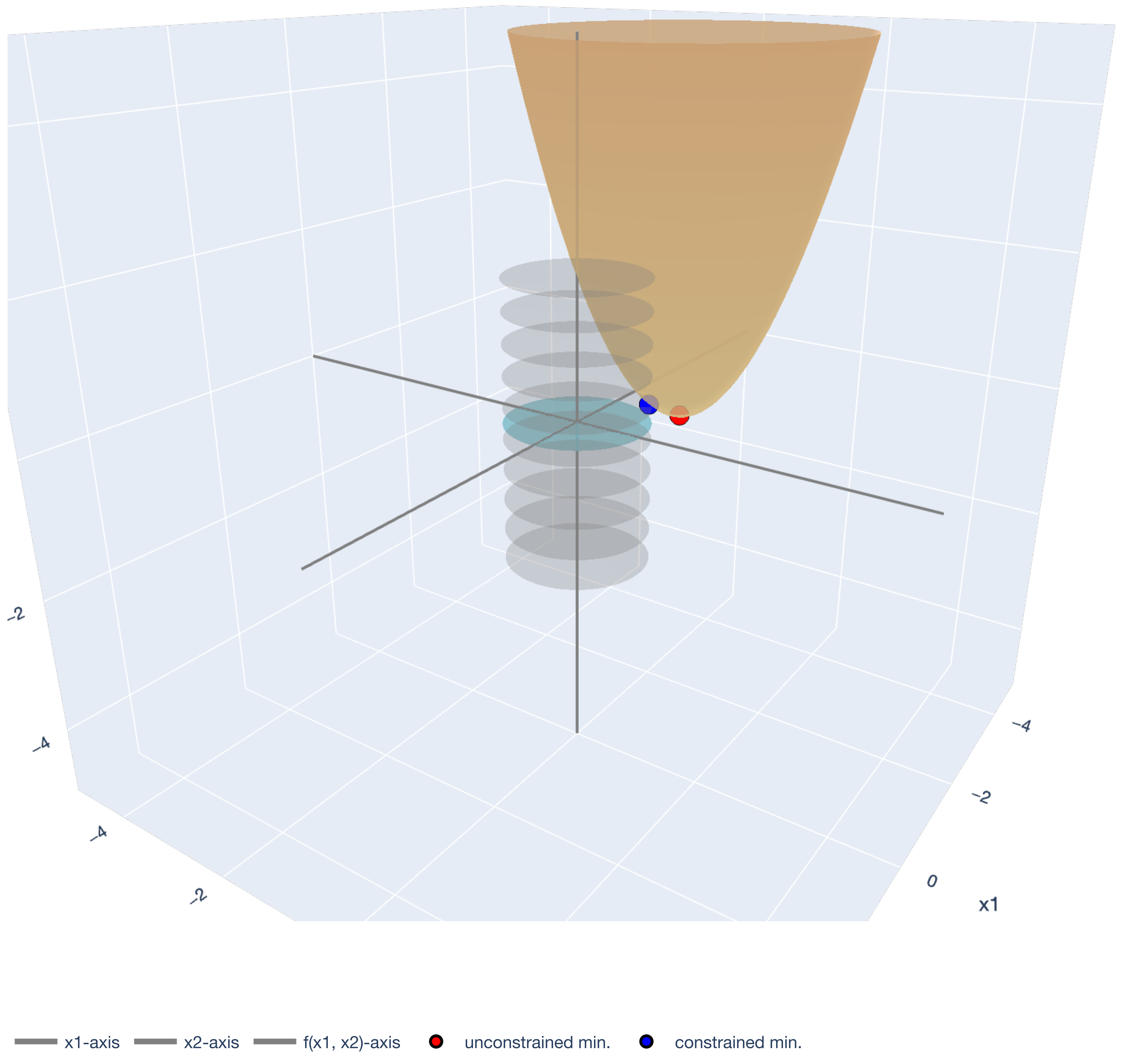
Constrained local optima (Lagrangian and KKT). When $\mathscr{C}$ is represented by inequalities and equalities, we can use the method of *Lagrange multipliers* and the *KKT Theorem* to "unconstrain" the problem.

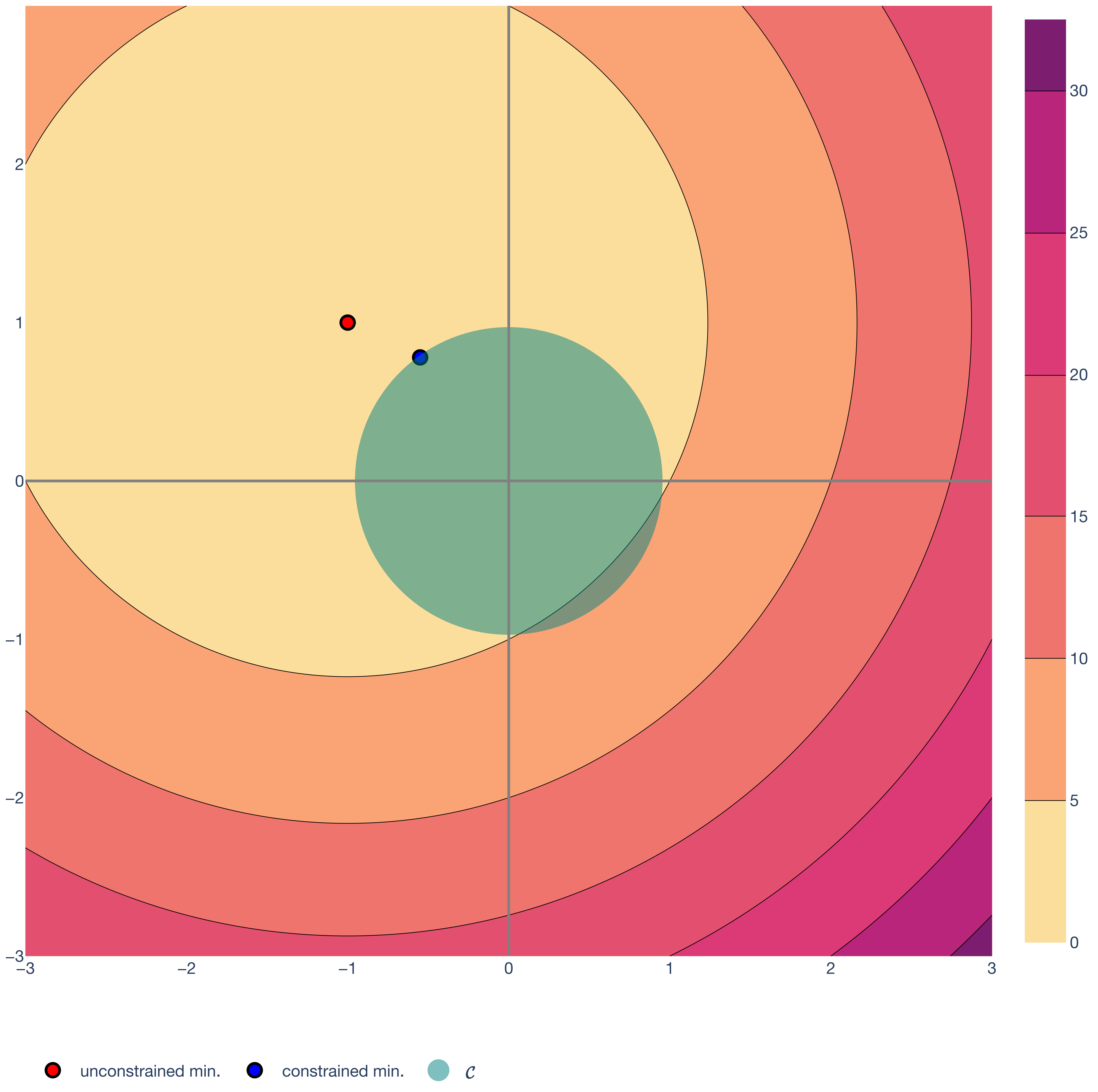
Ridge regression and minimum norm solutions. By constraining the norm of $\mathbf{w}^* \in \mathbb{R}^d$ of least squares (i.e. $||\mathbf{w}^*||$), we obtain more "stable" solutions.

Lesson Overview

Big Picture: Least Squares

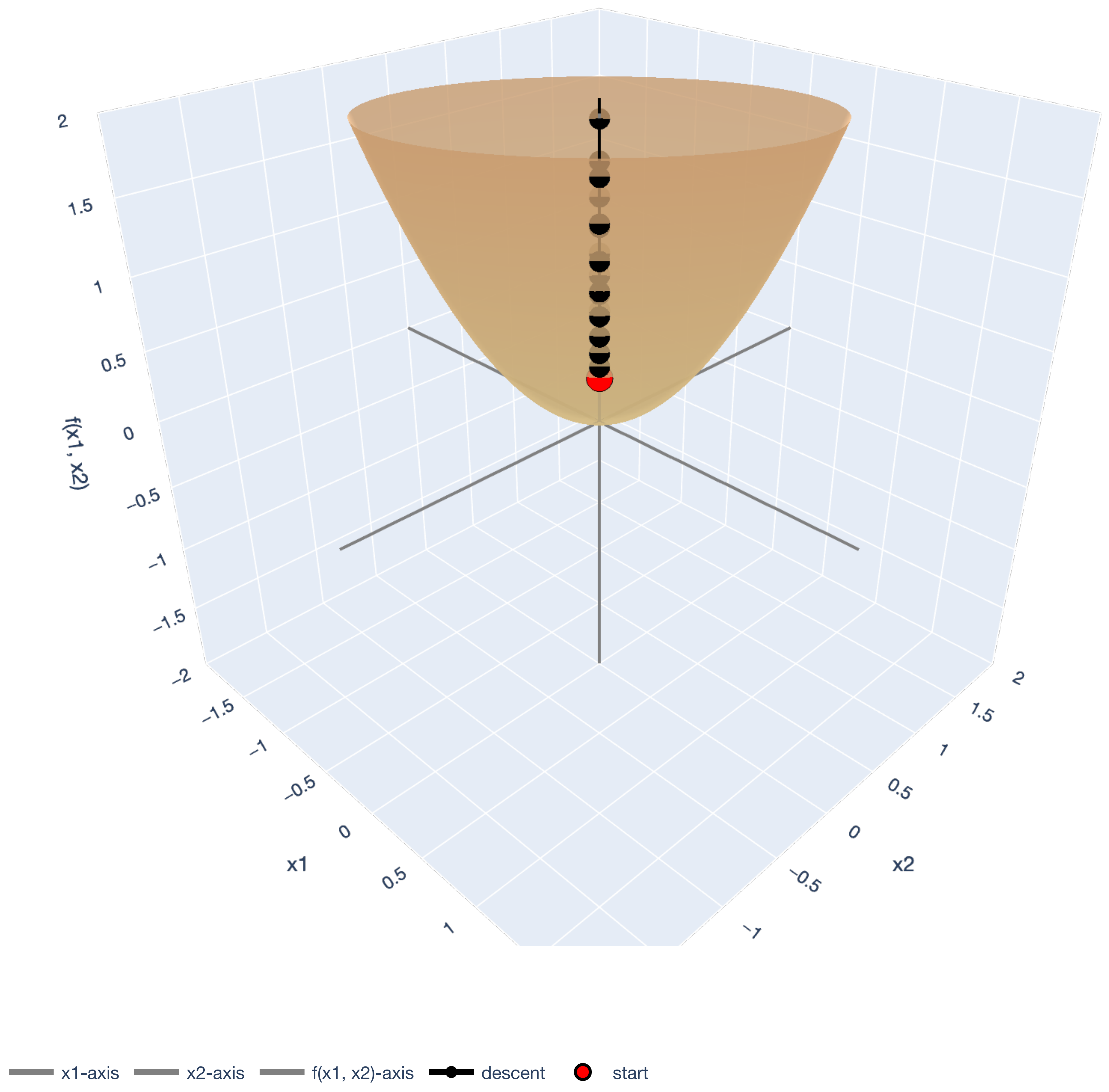
REGULARIZATION → BIAS-VARIANCE TRADEOFF

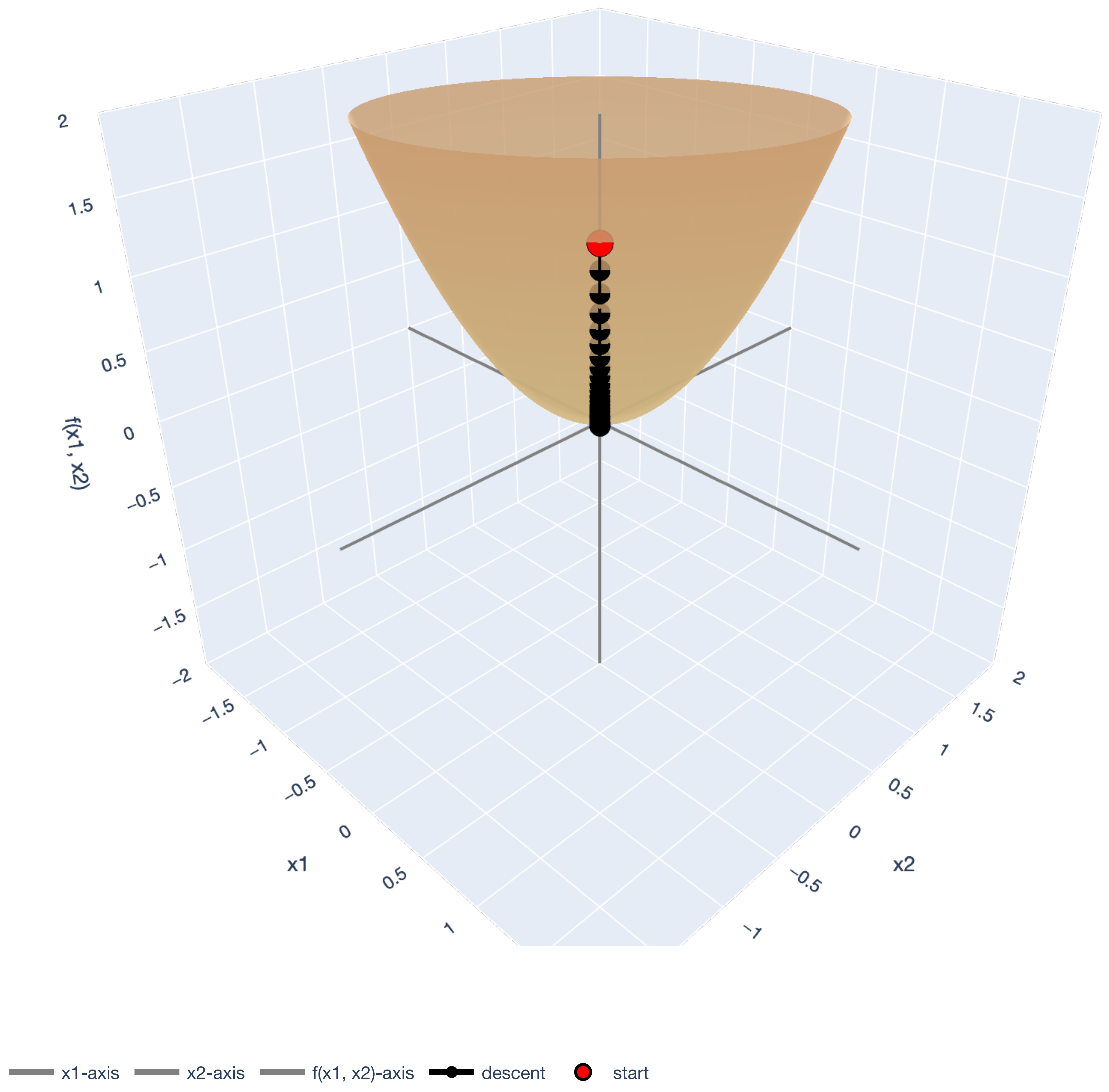




Lesson Overview

Big Picture: Gradient Descent





References

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

Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach. John H. Hubbard and Barbara Burke Hubbard.

“Lecture 1: Introduction.” Santiago Balserio and Ciamac Moallemi. Lecture notes from B9118 Foundations of Optimization, Fall 2023.

“Lecture 2: Local Theory of Optimization.” Santiago Balserio and Ciamac Moallemi. Lecture notes from B9118 Foundations of Optimization, Fall 2023.