# Math for Machine Learning

## Week 4.1: Optimization and the Lagrangian Method

By: Samuel Deng

# Logistics & Announcements

# Lesson Overview

**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.

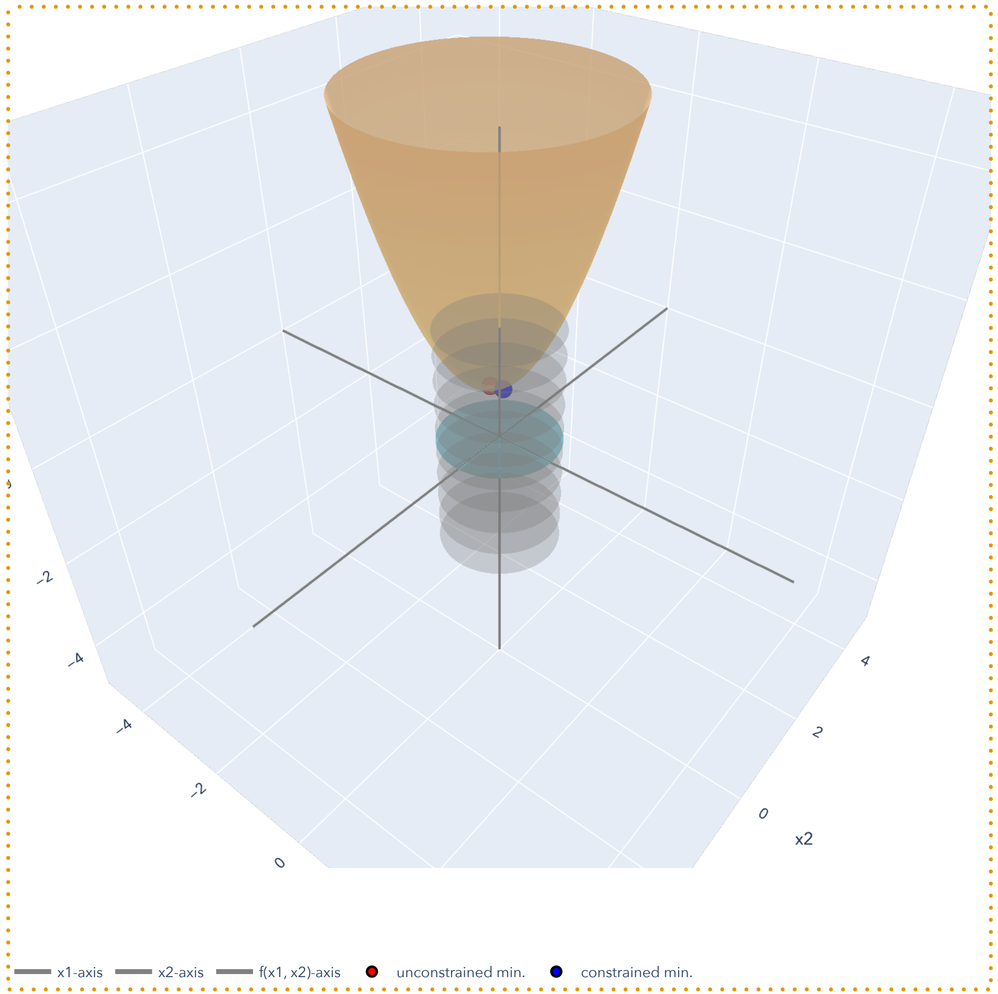
**Unconstrained local optima.** With no constraints, the standard tools of calculus give conditions for a point $\mathbf{x}^*$ to be optimal, at least to all points close to it.

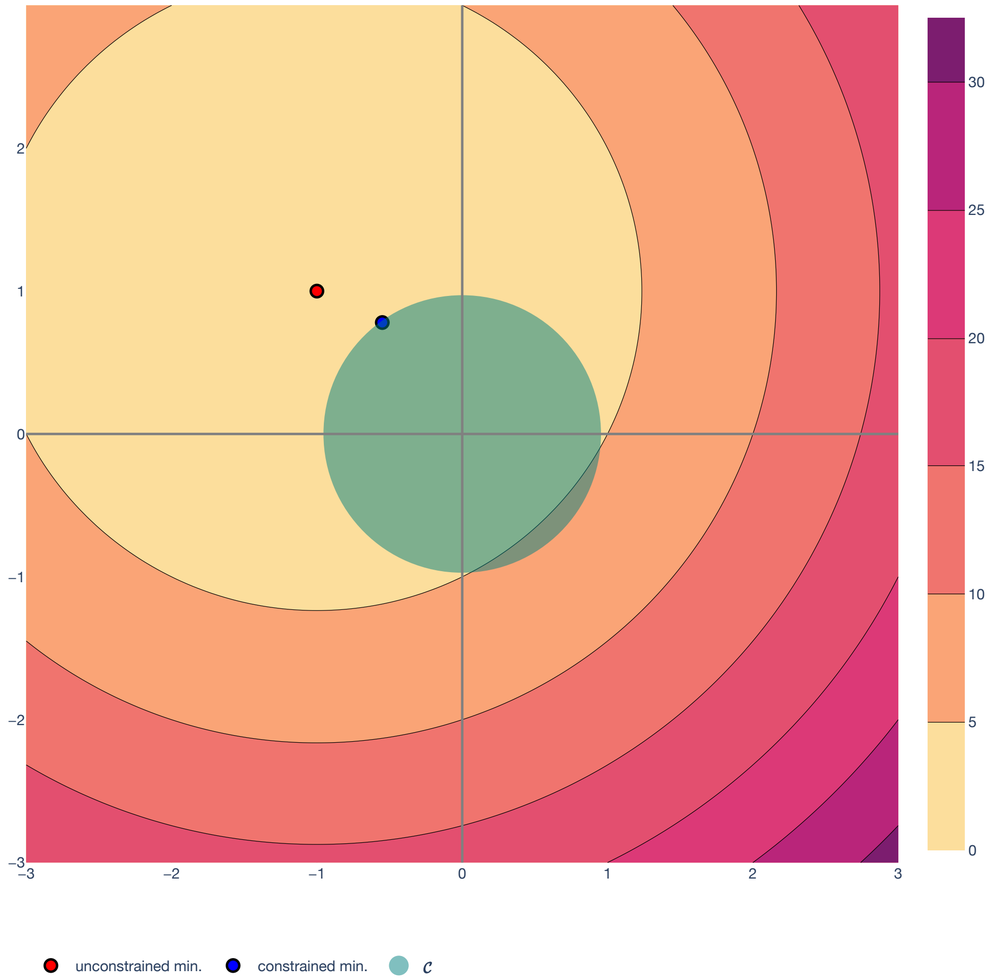
**Constrained local optima (Lagrangian and KKT).** When $\mathscr{C}$ is represented by inequalities and equalities, we can use the method of Lagrange multipliers and the KKT Theorem to "unconstrain" the problem.

**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

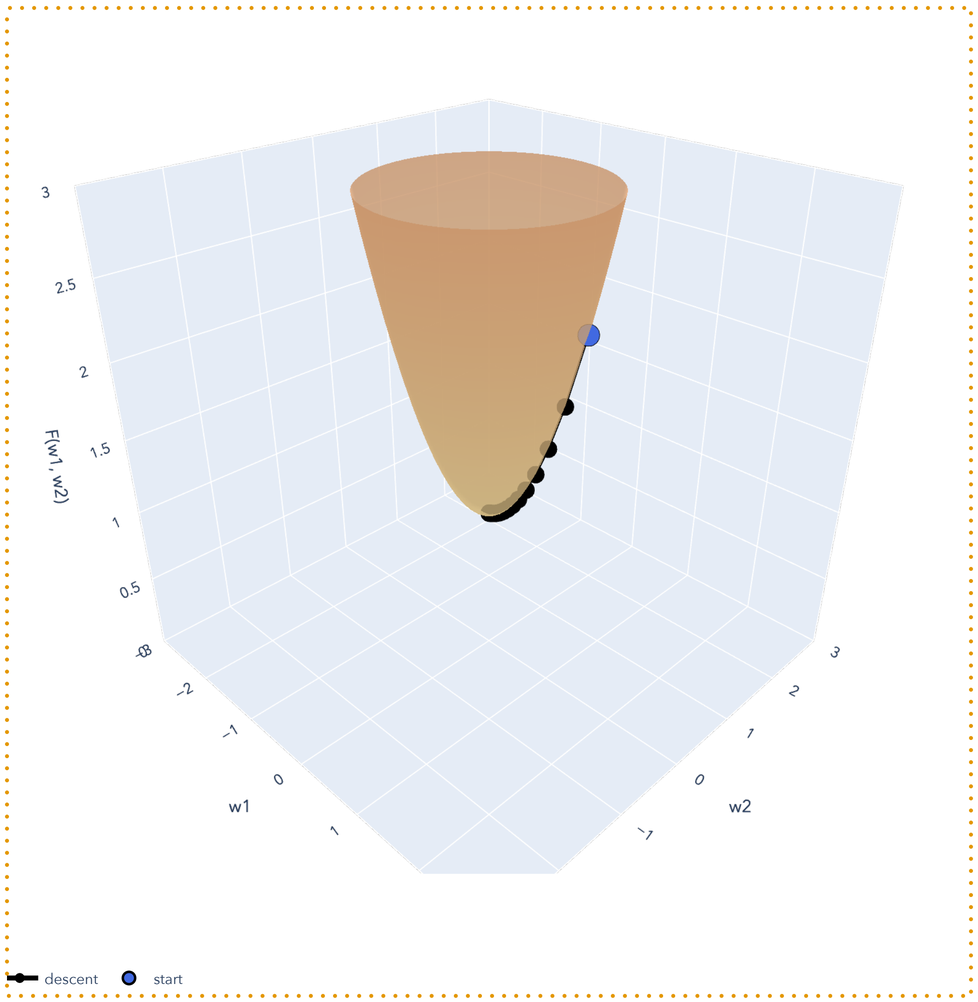
## Big Picture: Least Squares

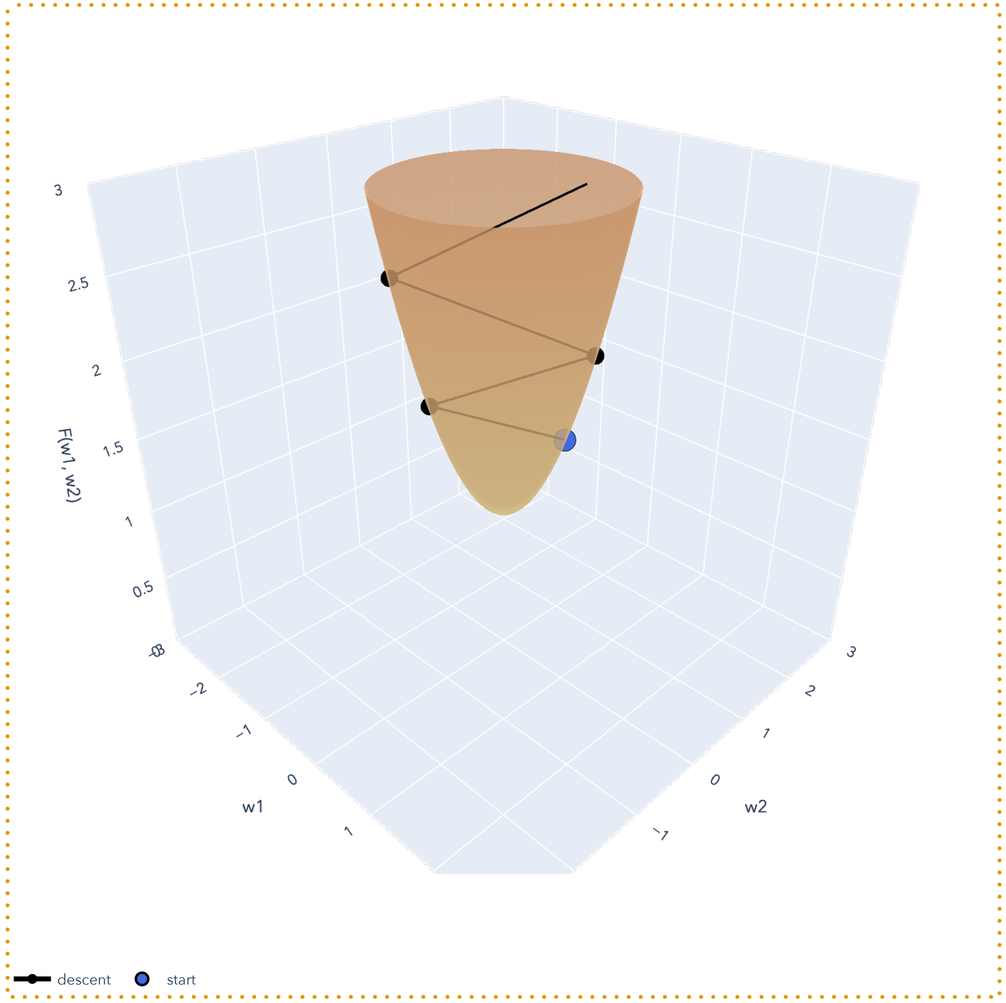


# Lesson Overview

## Big Picture: Gradient Descent



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

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

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

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

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

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.

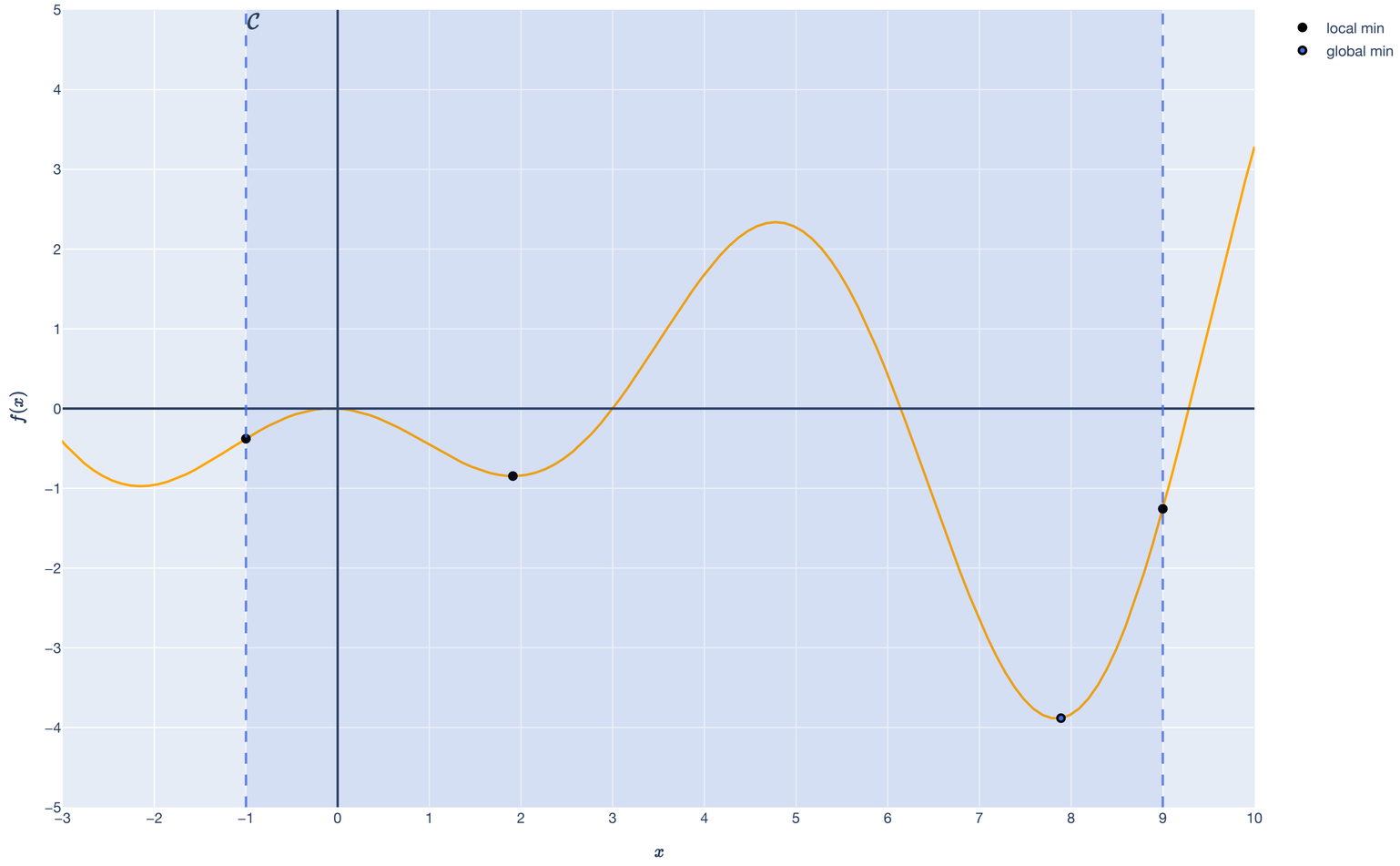
# Motivation

## Optimization in single-variable calculus

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

**Intermediary goal:** Find the *local minima*.

*Now we will focus on constraints!*



# Motivation

## Example: Linear Programming

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

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

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

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

# Motivation

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

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

$$100 \text{ bacon}, 120 \text{ egg}, 150 \text{ cheese, and } 300 \text{ (sandwich) rolls.}$$

**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:

$$100 \text{ bacon}, 120 \text{ egg}, 150 \text{ cheese, and } 300 \text{ (sandwich) rolls.}$$

**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

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?

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

# 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?

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

Linear program:

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

# Motivation

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

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

LP in matrix form:

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

$$
\mathbf{A} = \begin{bmatrix}
1 & 0 & 1 \\
1 & 2 & 3 \\
1 & 1 & \frac{1}{2} \\
1 & 1 & 0 \\
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & -1
\end{bmatrix}
\qquad
\mathbf{b} = \begin{bmatrix}
100 \\
120 \\
150 \\
300 \\
0 \\
0 \\
0
\end{bmatrix}
$$

# Regression

## Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

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 (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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:

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

# 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:

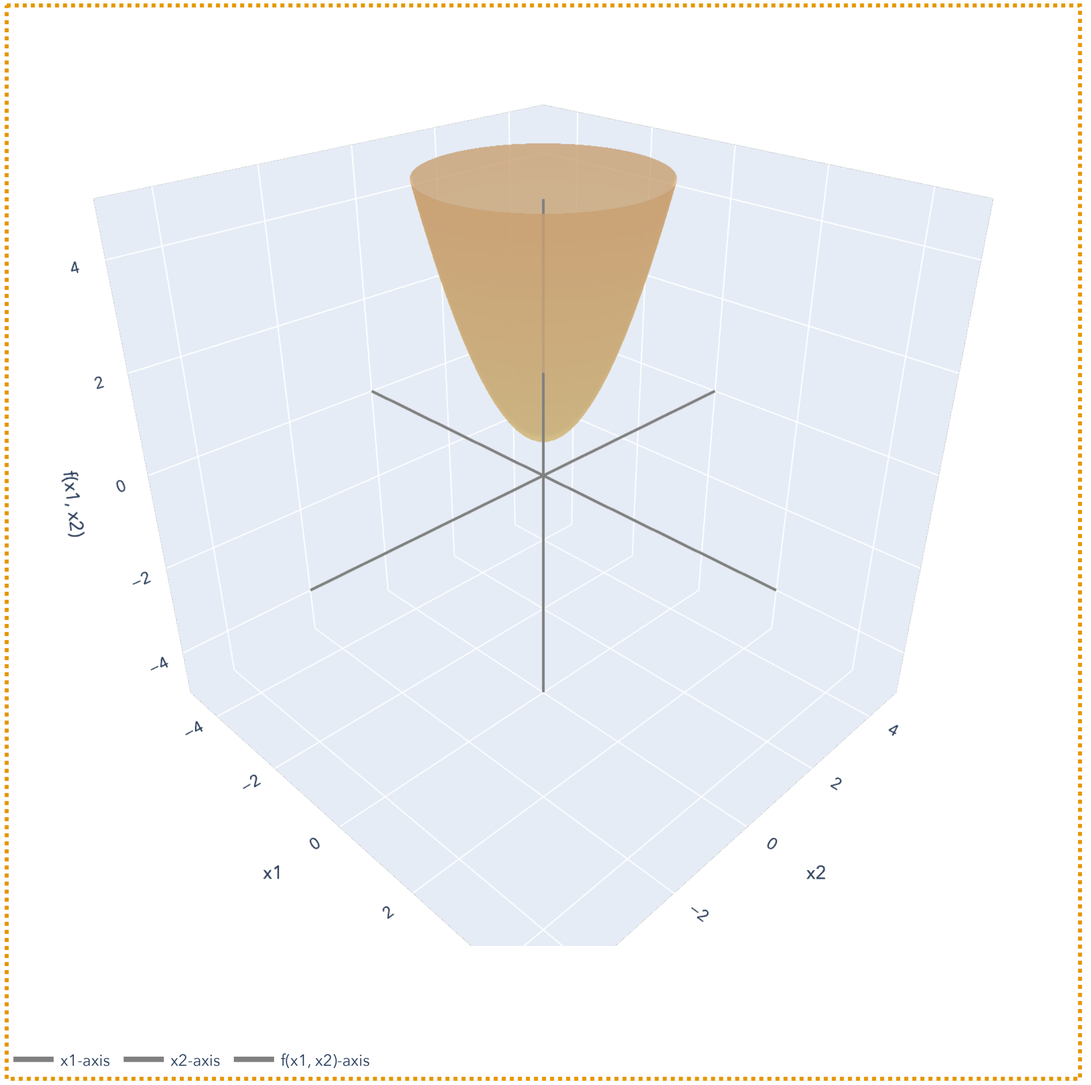
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

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

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

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

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



# Least Squares

## OLS Theorem

Proof (Calculus proof of 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." $\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}$.

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

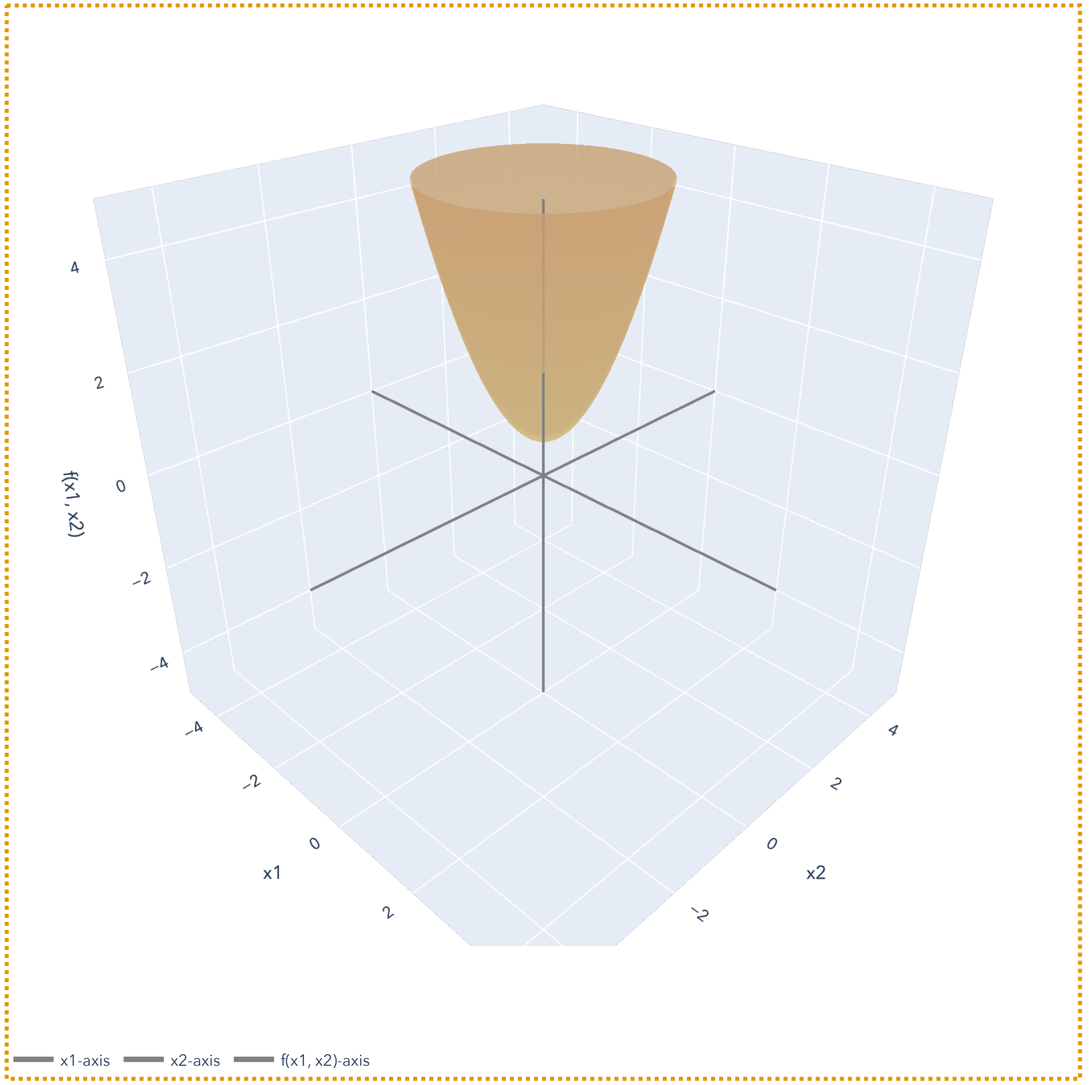
$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

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

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

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

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



# Local and global minima

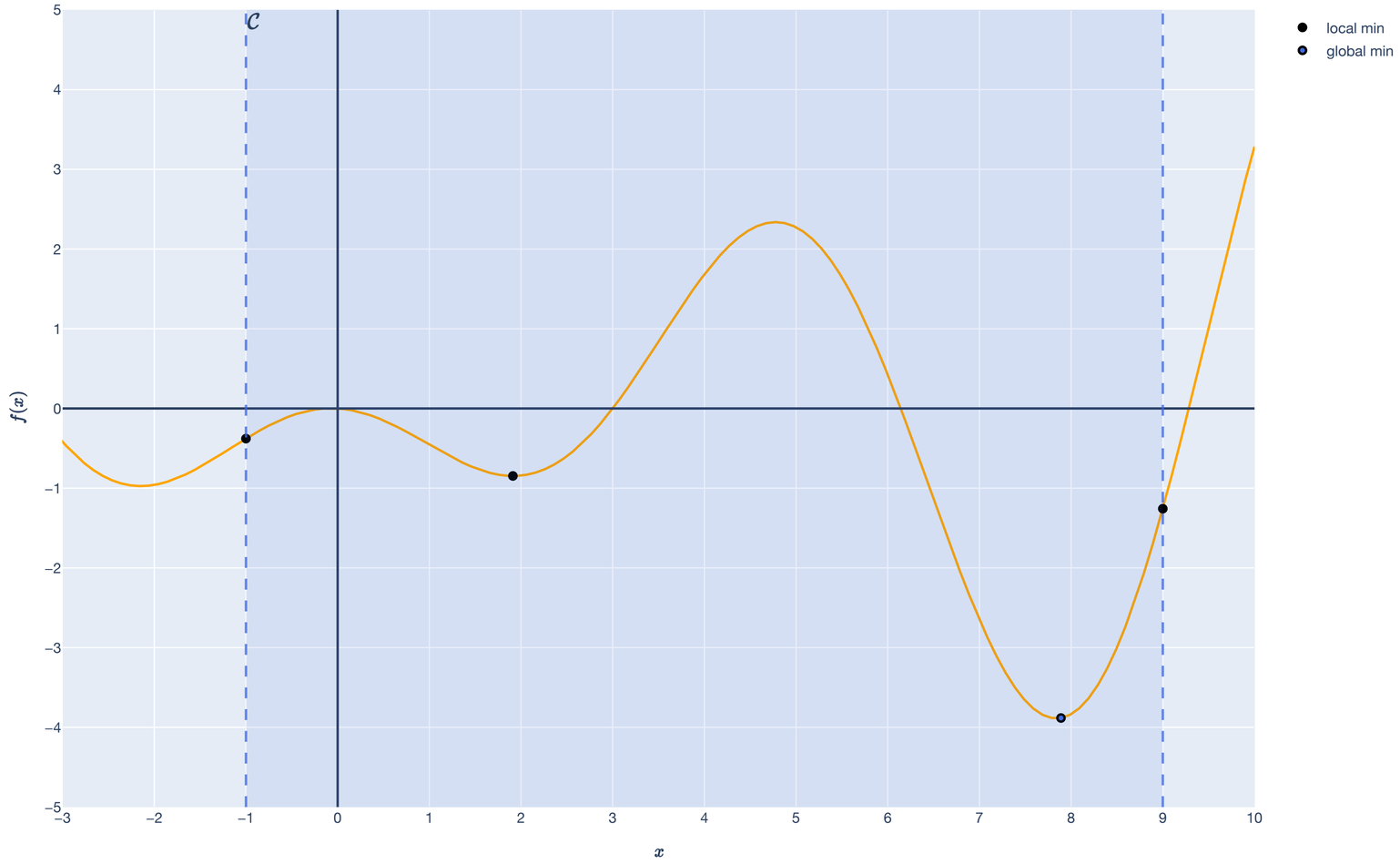
## Definition of “locality” and different minima

# 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

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

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

# “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 $\mathbf{x}$?

# “Local” to a Point

## Definition of the interior of a set

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

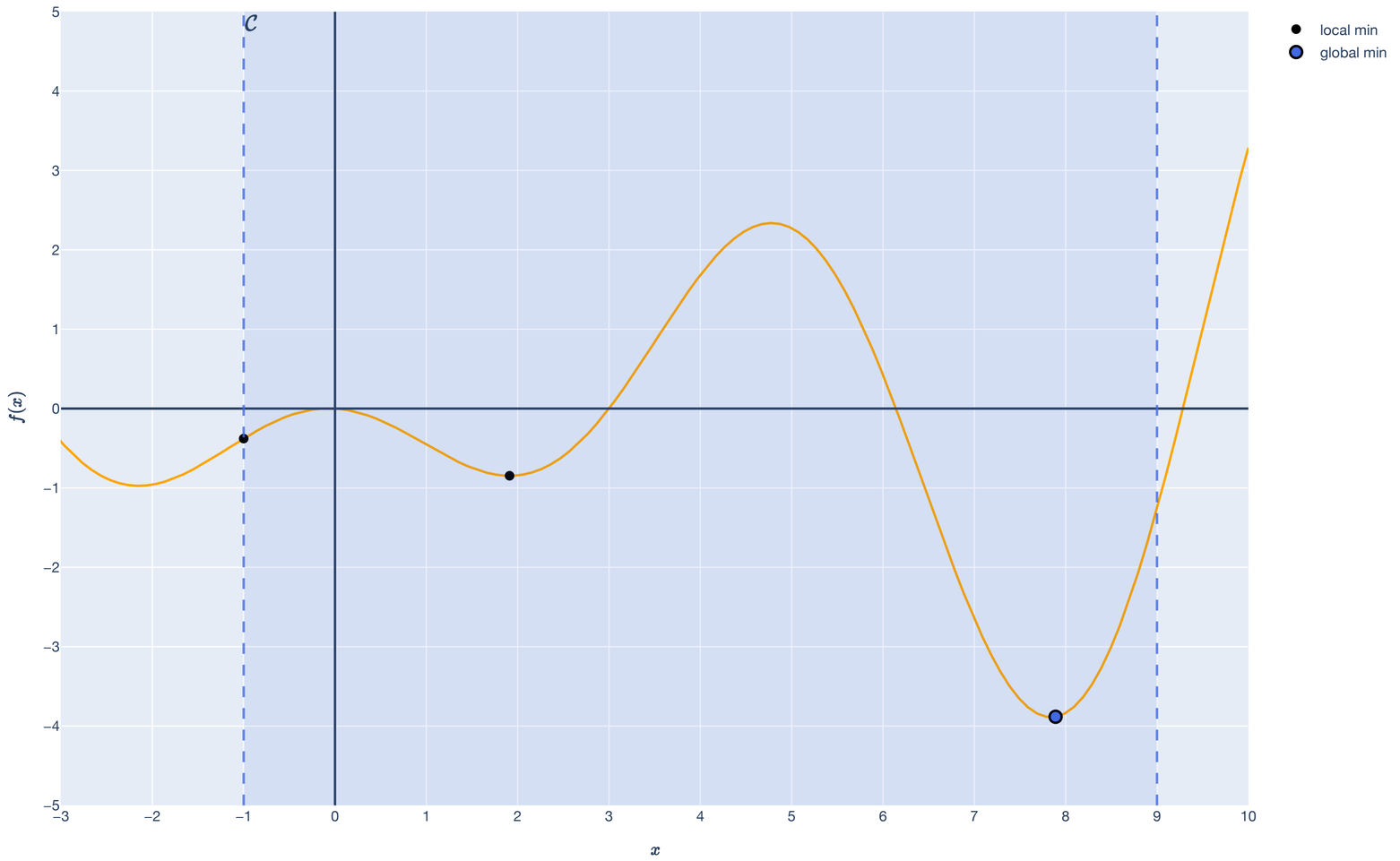
Let $S \subseteq \mathbb{R}^d$ be a set. A point $\mathbf{x} \in S$ is an interior point if there exists a neighborhood $B_\delta(\mathbf{x})$ around $\mathbf{x}$ such that $B_\delta(\mathbf{x}) \subset S$ (where $\subset$ is *proper subset*).

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

$$\mathrm{int}(S) := \{\mathbf{x} \in S : N_\delta(\mathbf{x}) \subset S\}\,.$$

# Types of Minima

## Local and global minima



# Types of Minima

## Local and global minima

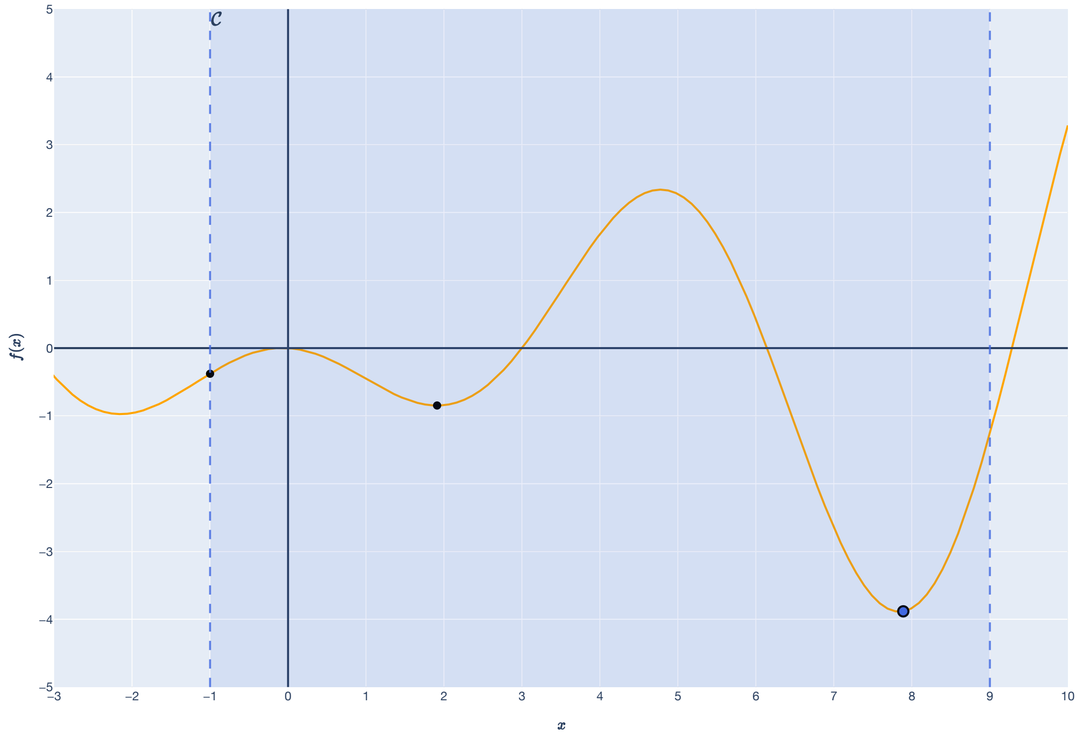
$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathscr{C} \end{aligned}$$

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

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

$\mathbf{x}^* \in \mathscr{C}$ 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

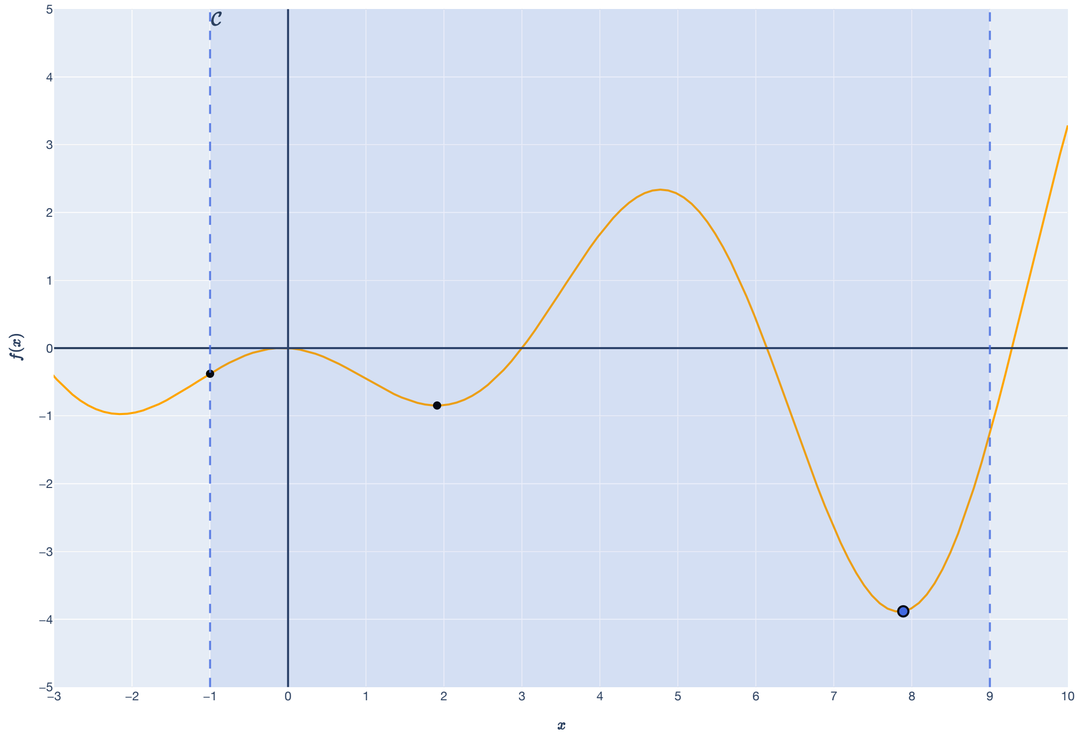
$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathscr{C} \end{aligned}$$

$\hat{\mathbf{x}} \in \mathscr{C}$ is an unconstrained local minimum if there is 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 $\mathbf{int}(\mathscr{C})$.

*Constrained local minima* can be on the "edge" of the constraint set.



# Types of Minima

## Which type of minima are each of these points?

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

constrained local:

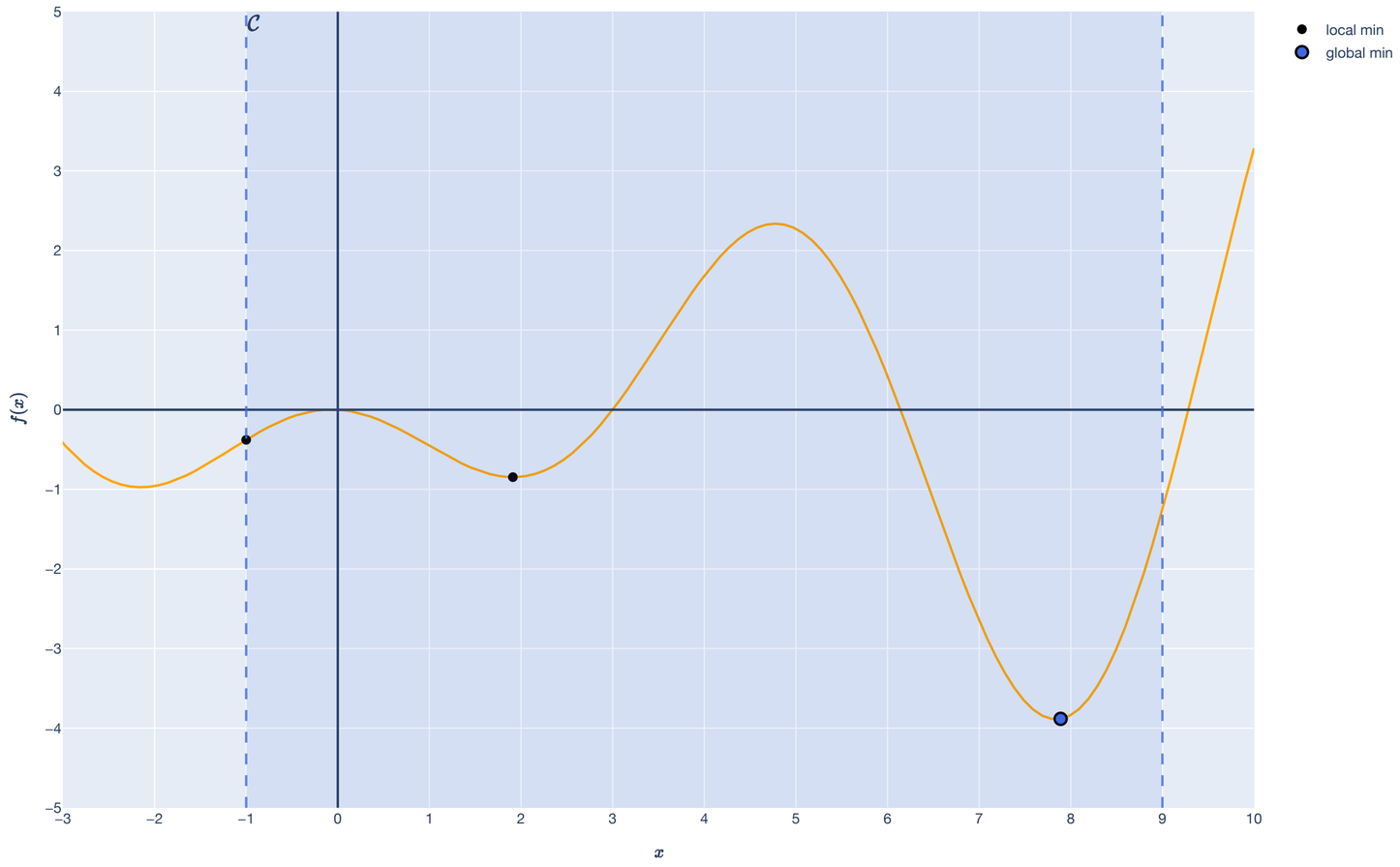
$f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{C} \cap B_\delta(\hat{\mathbf{x}})$

unconstrained local:

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

global:

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



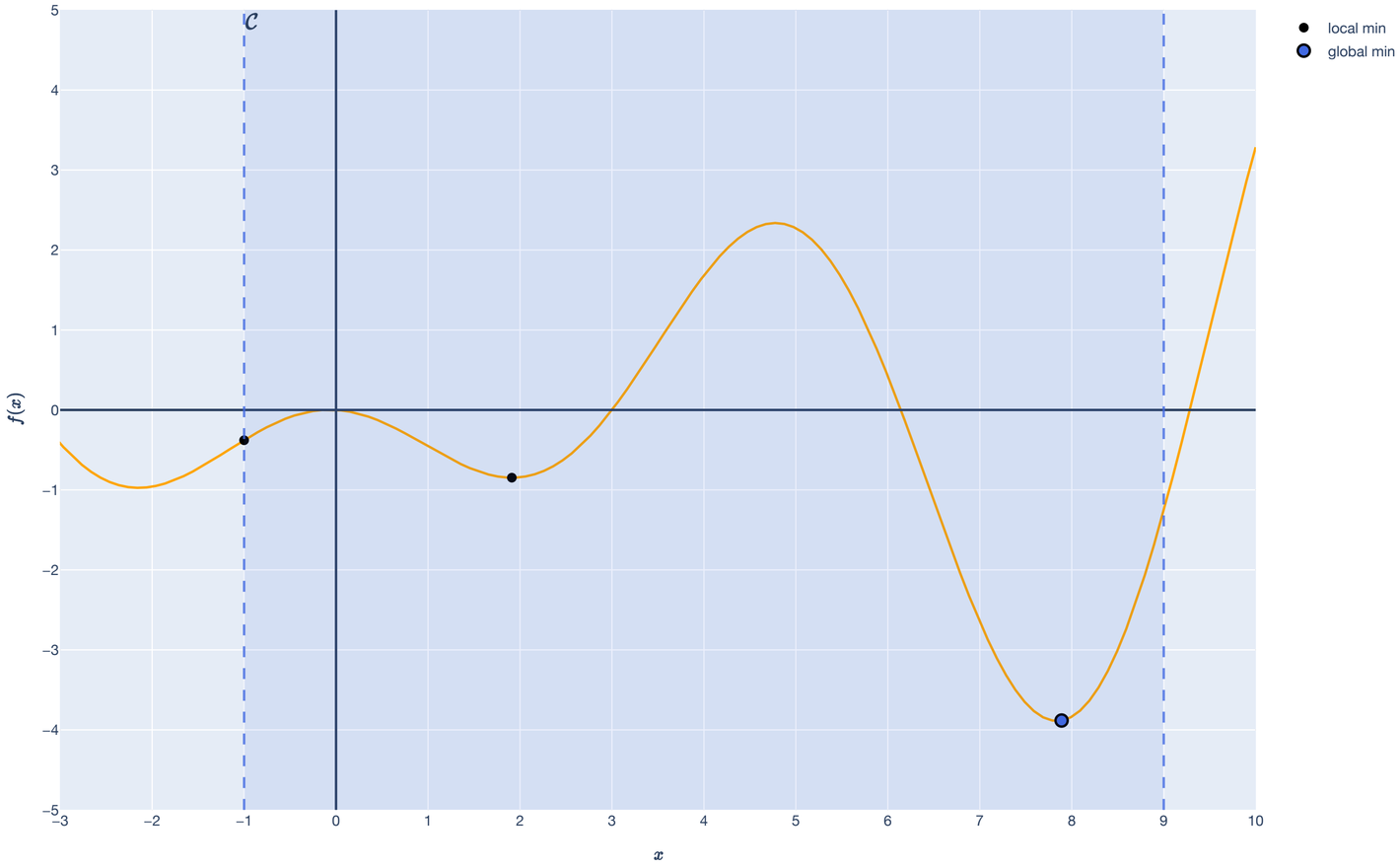
# Types of Minima

## Big picture

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 an *unconstrained local minima*.
>
> With $\mathscr{C}$, global minima may lie on the boundary of the constraint set.
>
> *Find local minima, then test!*



# Finding local minima

## Big Picture

# Necessary and sufficient conditions

## Review

$$P \implies Q$$

$Q$ is necessary for $P$. $P$ is sufficient for $Q$.

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

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

**necessity:** Your property cannot hold unless you assume this.

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

# Unconstrained Minima

## How do we find unconstrained minima?

$\hat{\mathbf{x}} \in \mathscr{C}$ is an unconstrained local minimum if there is a neighborhood $B_\delta(\hat{\mathbf{x}}) \subset \mathscr{C}$ around $\hat{\mathbf{x}}$ s.t.

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

From single-variable calculus, this is true if:

$$f'(x) = 0 \text{ and } f''(x) \geq 0.$$

# 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:*

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

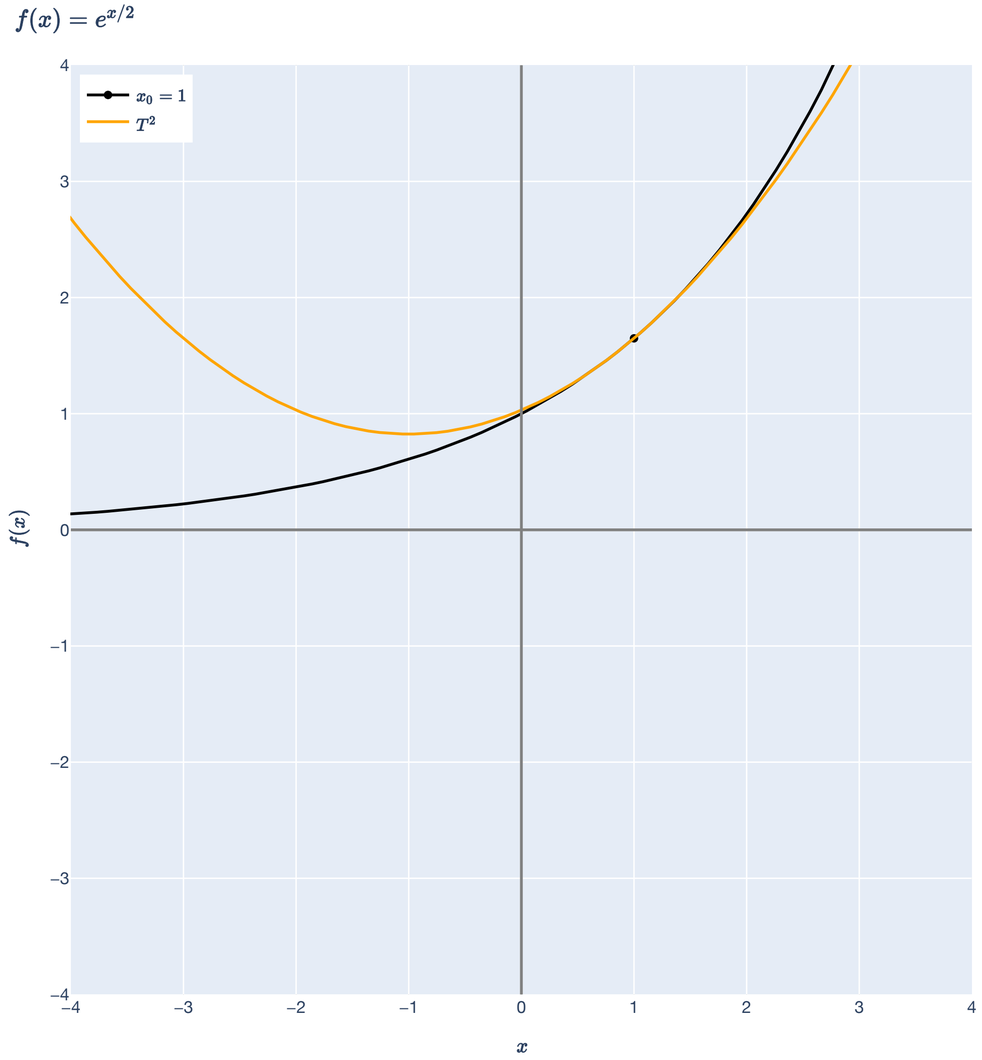
# Second-order Taylor Approximation

## Single-variable example

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

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

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



# 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:*

$$f(x_0 + \delta) \approx f(x_0) + \underset{f'(x) = 0}{f'(x_0)\delta} + \frac{1}{2}\underset{f''(x) \geq 0}{f''(x_0)}\delta^2.$$

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

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

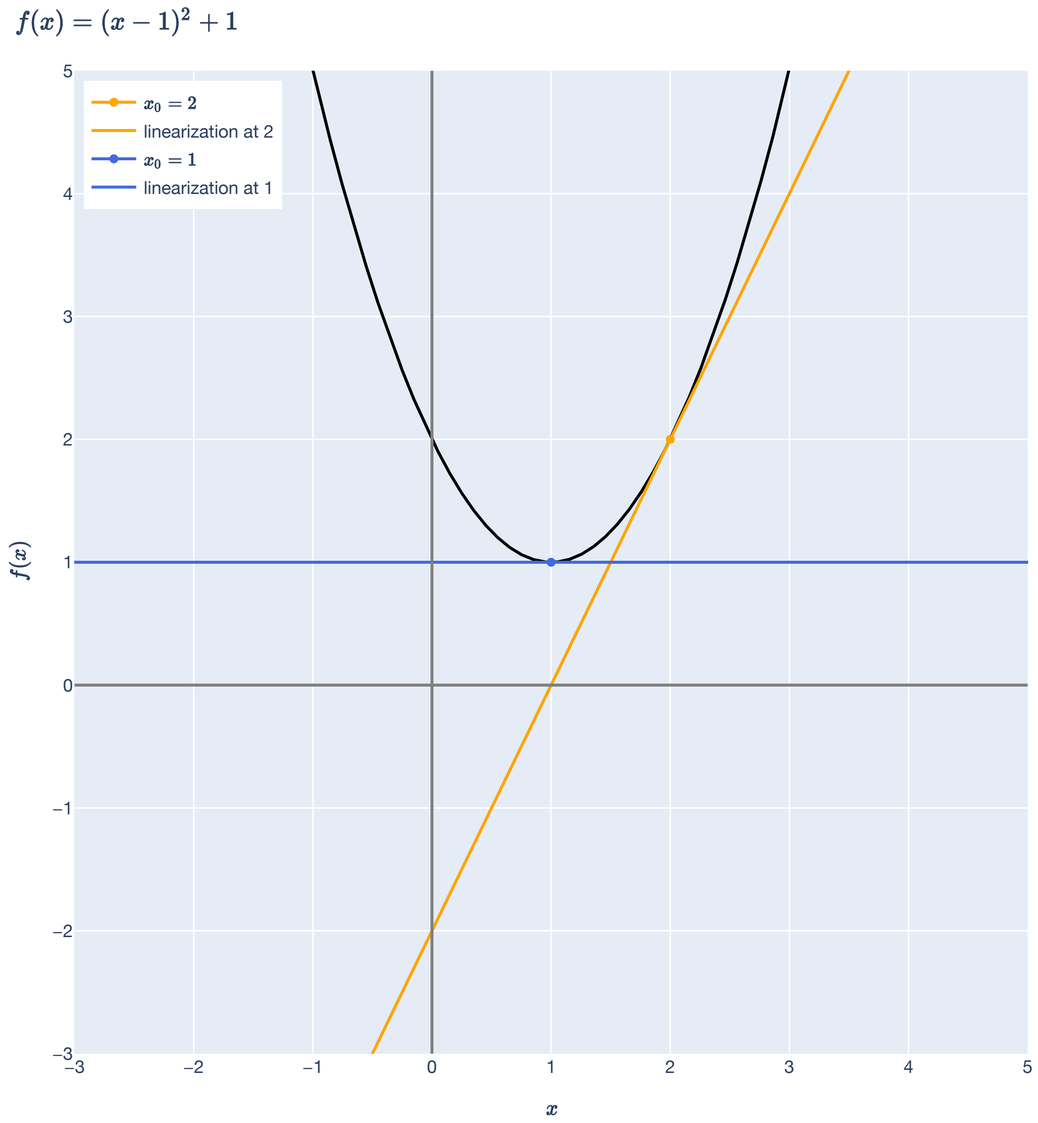
# Unconstrained Minima

## Sufficient conditions met

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

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

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



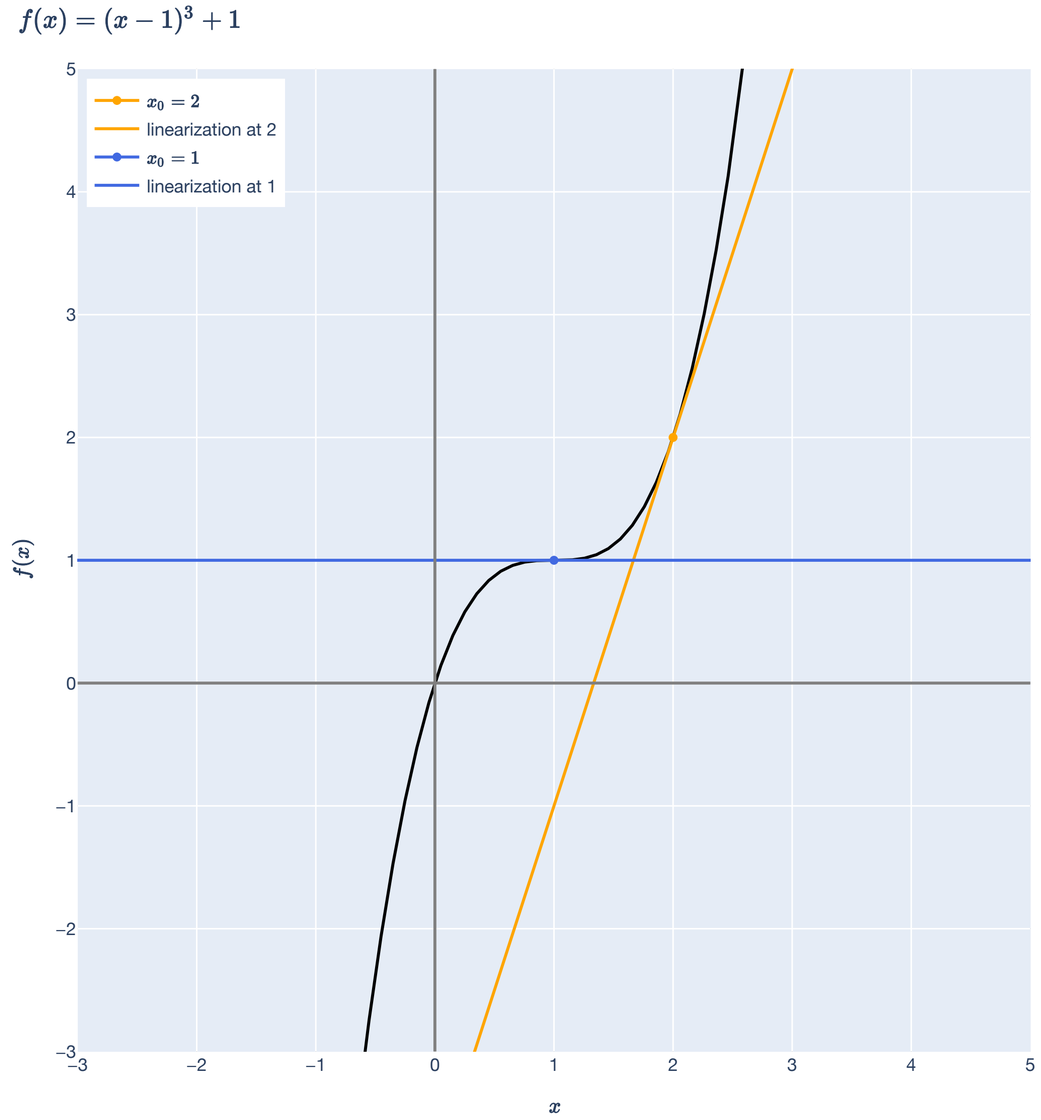
# Unconstrained Minima

## Necessary, not sufficient

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

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

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



# Taylor's Theorem

## Intuition

How much do we lose by approximating $f$ with a Taylor approximation?

**Remainder**: how much more Taylor series is left after "chopping it off" at order $n$.

**First-order approximation:**

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

The remainder is:

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

# Taylor's Theorem

## Intuition

How much do we lose by approximating $f$ with a Taylor approximation?

**Remainder**: how much more Taylor series is left after "chopping it off" at order $n$.

**Second-order approximation:**

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

The remainder is:

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

# Remainder of Taylor Polynomial

## Definition

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

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

What behavior would we like?

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

# Remainder of Taylor Polynomial

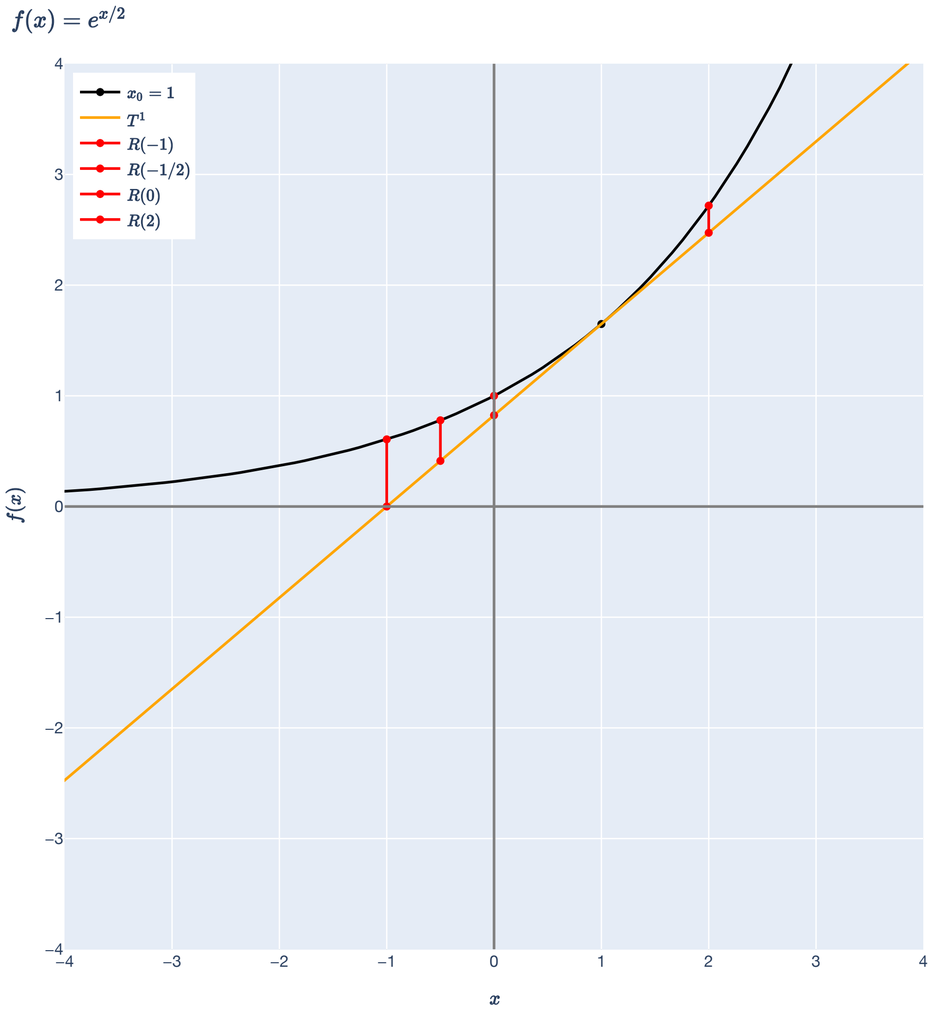
## Definition

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

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

What behavior would we like?

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



# Taylor's Theorem

## Peano's Form

**Theorem (2nd Order Taylor's Theorem: Peano's Form).** Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable at $\mathbf{x}_0$ and let $\mathbf{d} \in \mathbb{R}^d$. For every $\epsilon > 0$, there exists a neighborhood $B_\delta(\mathbf{0})$ such that

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

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

*However small you want the remainder ($\epsilon$), as long as you are $\delta$-close to $\mathbf{x}_0$, the remainder can get $\epsilon\|\mathbf{d}\|^2$ small.*

# Unconstrained local minima

## Necessary conditions

# Least Squares

## OLS Theorem

Proof (Calculus proof of 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." $\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}.$

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

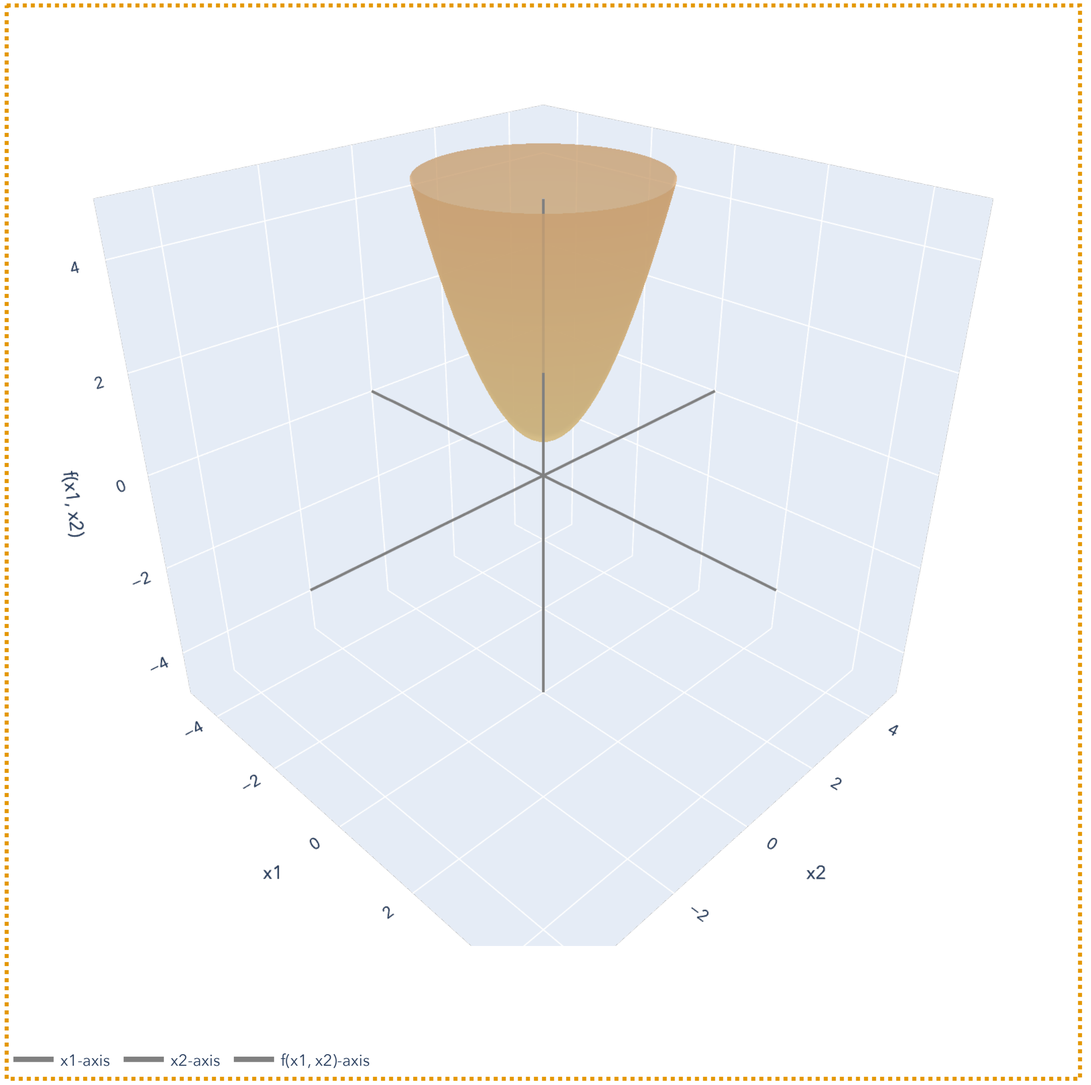
$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

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

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

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

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



# Necessary Conditions

## Comparison to single variable

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

when $\delta$ is small enough.

Necessary conditions:

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

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

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

Necessary conditions:

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

# Differential Calculus

## Review: Derivative

at the point where we're taking derivative…

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

linear approximation

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

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

*Throughout this section,* $\mathbf{d} = \mathbf{x} - \mathbf{x}_0$.

# Unconstrained Minima

## Necessary conditions

Theorem (Necessary Conditions for Unconstrained Local Minimum).

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

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

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

# Proof of first order necessary condition

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

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

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…

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

which is the same as stating:

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

# Proof of first order necessary condition

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

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

From Step 1,

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

$\mathbf{x}^*$ 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),

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

# Proof of first order necessary condition

## Step 3: $\mathbf{d} \in \mathbb{R}^n$ was an arbitrary direction.

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

From Step 2, if $\alpha < \delta$ (sufficiently small), $\nabla f(\mathbf{x}^*)^\top \mathbf{d} \geq 0.$ But $\mathbf{d} \in \mathbb{R}^d$ was an arbitrary direction with $\|\mathbf{d}\| = 1$.

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

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

$$\vdots$$

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

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

# Unconstrained Minima

## Necessary conditions

Theorem (Necessary Conditions for Unconstrained Local Minimum).

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

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

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

# Proof of second order necessary condition

## Step 1: Use second-order Taylor approximation

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

Choose an arbitrary direction $\alpha\mathbf{d} \in \mathbb{R}^d$ where $\alpha > 0$ is a scalar. By Taylor's Theorem (Peano's form) there exists $\delta > 0$ such that for all $\mathbf{d} \in B_\delta(\mathbf{0})$:

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - \left( f(\mathbf{x}^*) + \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} \right) \leq \alpha \|\mathbf{d}\|^2.$$

# Proof of second order necessary condition

## Step 2: Use first-order condition so $\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} = 0$

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

$$f(\mathbf{x}^* + \alpha \mathbf{d}) - \left( f(\mathbf{x}^*) + \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} \right) \leq \alpha \|\mathbf{d}\|^2$$

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

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

# Proof of second order necessary condition

## Step 3: Divide by $\|\mathbf{d}\|^2$ and use local optimality: $f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*) \geq 0$

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

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

Divide by $\|\mathbf{d}\|^2$ everywhere and take the limit as $\alpha \to 0$:

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

By local optimality of $\mathbf{x}^*$ and arbitrary $\mathbf{d} \in \mathbb{R}^d$:

$0 \leq \frac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*)}{\|\mathbf{d}\|^2}$, so $0 \leq \frac{1}{2}(\alpha\mathbf{d}/\|\mathbf{d}\|)^\top \nabla^2 f(\mathbf{x}^*)(\alpha\mathbf{d}/\|\mathbf{d}\|) \implies \nabla^2 f(\mathbf{x}^*)$ is PSD (definition of PSD).

# Unconstrained Minima

## Necessary conditions

Theorem (Necessary Conditions for Unconstrained Local Minimum).

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

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

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

# Unconstrained local minima

## Sufficient conditions

# Least Squares

## OLS Theorem

Proof (Calculus proof of 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." $\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}$.

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

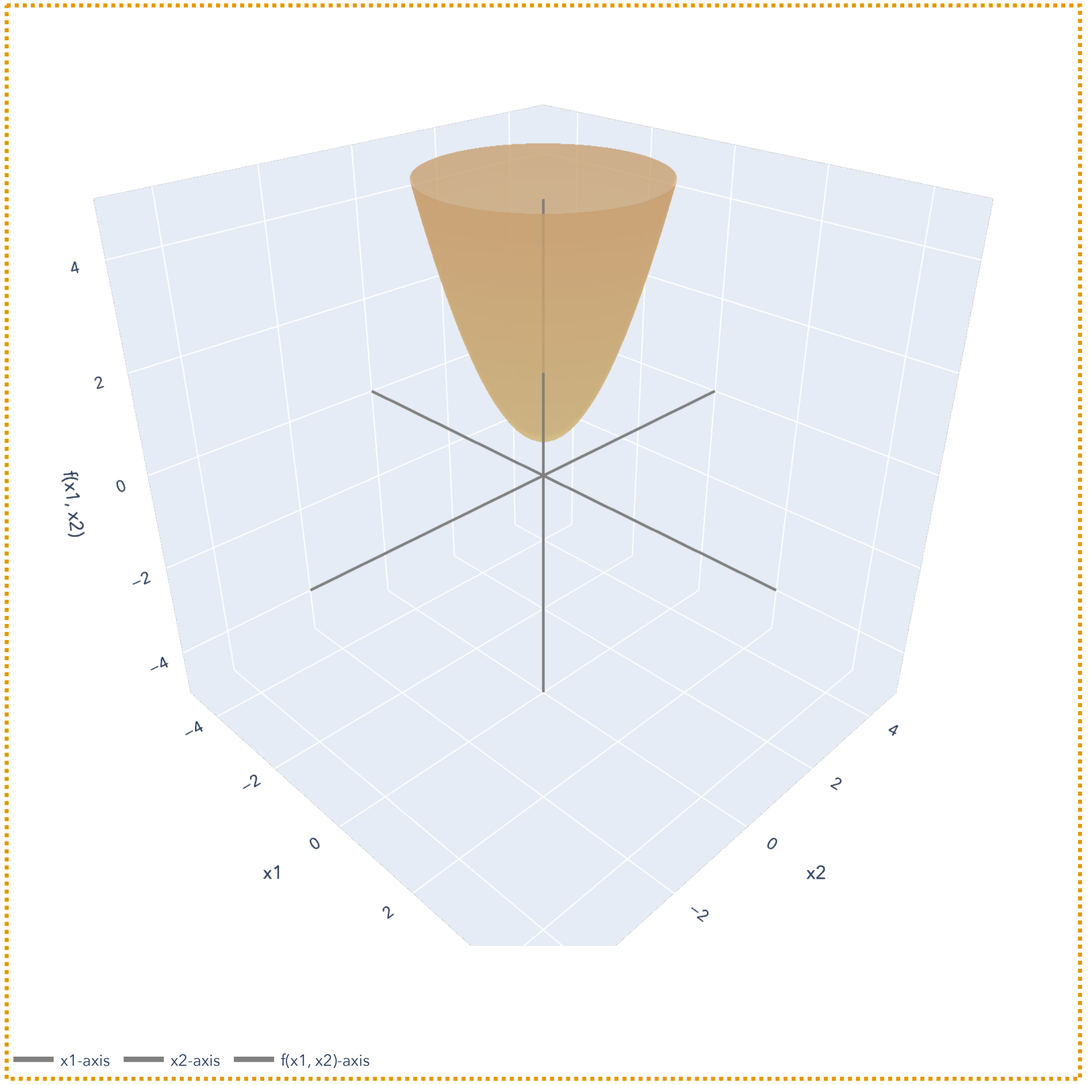
$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

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

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

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

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



# Sufficient Conditions

## Comparison to single variable

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

when $\delta$ is small enough.

Necessary conditions:

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

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

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

Necessary 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).

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

Let $\mathbf{x}^* \in \text{int}(\mathscr{C})$. If $f \in \mathscr{C}^2$ and

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

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

# Proof of second order sufficient condition

## Step 1: Use second-order Taylor approximation

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

Choose an arbitrary direction $\alpha\mathbf{d} \in \mathbb{R}^d$ where $\alpha > 0$ is a scalar. By Taylor's Theorem (Peano's form) there exists $\delta > 0$ such that for all $\mathbf{d} \in B_\delta(\mathbf{0})$:

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - \left( f(\mathbf{x}^*) + \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} \right) \geq -\alpha \|\mathbf{d}\|^2.$$

*Note: Used the negative direction of the statement (which is absolute value).*

# Proof of second order sufficient condition

## Step 2: Eigenvalues of PD matrix are positive

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

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

$$f(\mathbf{x}^* + \alpha\mathbf{d}) - \left( f(\mathbf{x}^*) + \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} \right) \geq -\alpha \|\mathbf{d}\|^2.$$

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

$$\implies \frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d} \geq \frac{\alpha^2}{2} \mathbf{v}_d^\top \nabla f(\mathbf{x}^*) \mathbf{v}_d = \frac{\lambda_d \alpha^2}{2}.$$

# Proof of second order sufficient condition

## Step 3: $\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} = 0$ from first-order condition

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

Cancel out the first-order term $\alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} = 0$ and plugin the eigenvalue lower bound

$$f(\mathbf{x}^* + \alpha \mathbf{d}) - f(\mathbf{x}^*) \geq \alpha \nabla f(\mathbf{x}^*)^\top \mathbf{d} + \underbrace{\frac{\alpha^2}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{d}}_{\geq \frac{\lambda_d \alpha^2}{2}} - \alpha \|\mathbf{d}\|^2$$

so this simplifies to…

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

# Proof of second order sufficient condition

## Step 4: Divide by $\|\mathbf{d}\|^2$ and consider small enough $\mathbf{d} \to \mathbf{0}$

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

Take our inequality

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

and divide by $\|\mathbf{d}\|^2$ to get:

$\frac{f(\mathbf{x}^* + \alpha\mathbf{d}) - f(\mathbf{x}^*)}{\|\mathbf{d}\|^2} \geq \left(\frac{\lambda_d}{2\|\mathbf{d}\|^2} - \frac{1}{\alpha}\right)\alpha^2$, and sufficiently small $\mathbf{d} \to \mathbf{0}$ makes the RHS positive.

# Least Squares

## OLS Theorem

Proof (Calculus proof of 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." $\nabla_{\mathbf{w}} f(\mathbf{w}) = 2(\mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}$.

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

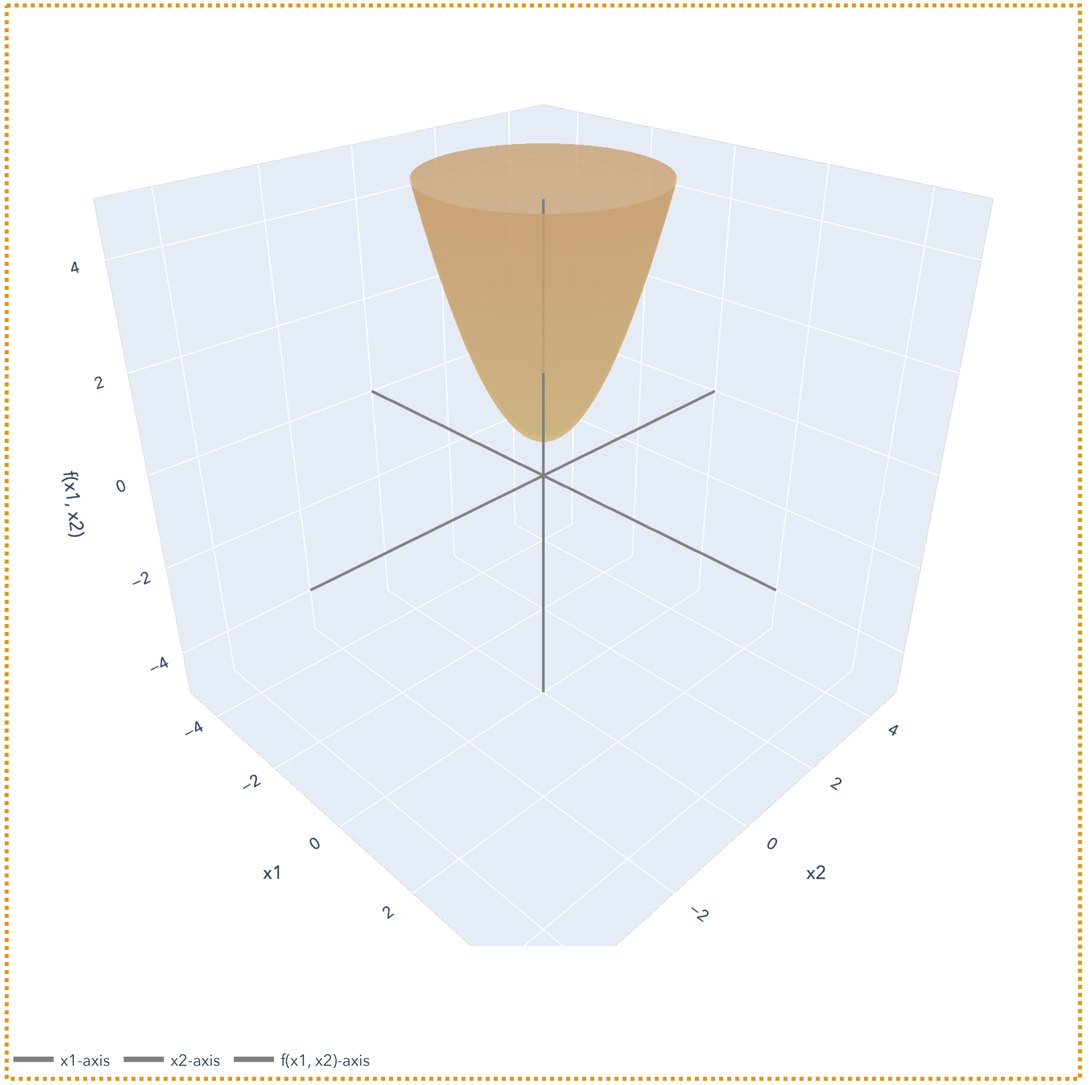
$\text{rank}(\mathbf{X}) = d \implies \text{rank}(\mathbf{X}^\top\mathbf{X}) = d \implies \mathbf{X}^\top\mathbf{X}$ is invertible:

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

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

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

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



# Finding global minima

## Introducing constraint sets

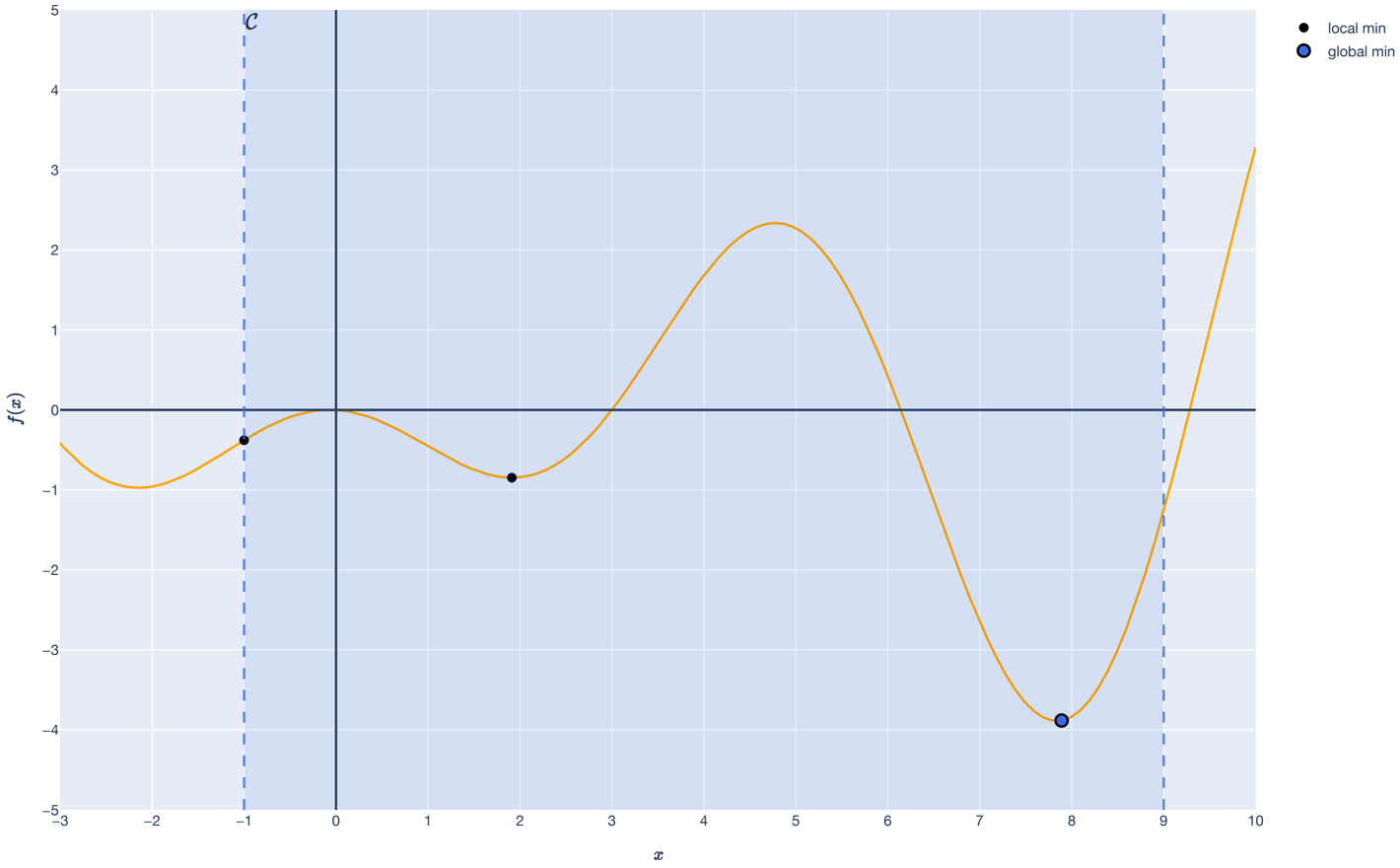
# Types of Minima

## Big picture

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 an *unconstrained local minima*.
>
> With $\mathscr{C}$, global minima may lie on the boundary of the constraint set.
>
> *Find local minima, then test!*



# Unconstrained Minima

## Necessary conditions

Theorem (Necessary Conditions for Unconstrained Local Minimum).

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

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

*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 $\mathbf{v} \in \mathbb{R}^d$.

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

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

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

# Finding global minima

## Using necessary conditions with constraints

Necessary conditions for unconstrained local minima:

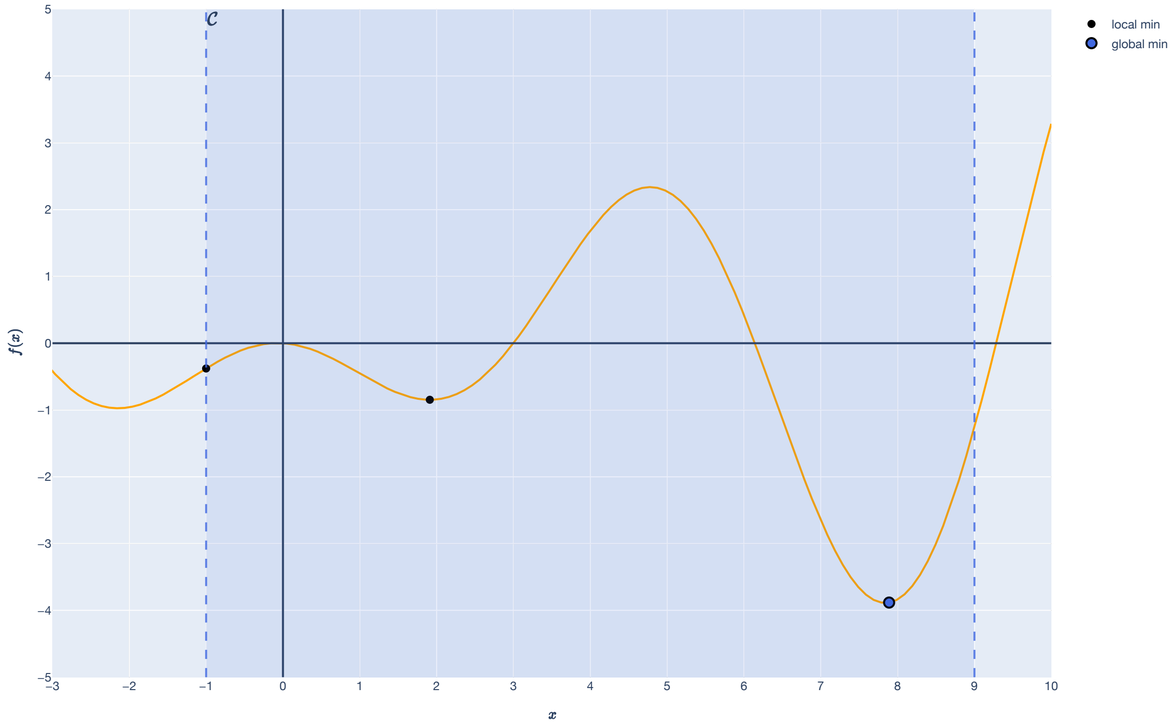
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \geq 0.$$

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

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

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

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



# 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 *unconstrained local minima* from first-order condition $M := \{\mathbf{x}^* \in \mathbb{R}^d : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.
2. There are no boundary points! ($B := \mathscr{C} \backslash \text{int}(\mathscr{C}) = \{\mathbf{x} \in \mathscr{C} : \mathbf{x} \notin \text{int}(\mathscr{C})\} = \varnothing$)
3. The global minimum must be in the set $M$, so evaluate $f$ on all $\mathbf{x} \in M$.

# Finding global minima

## Using necessary conditions without constraints

Necessary conditions for unconstrained local minima:

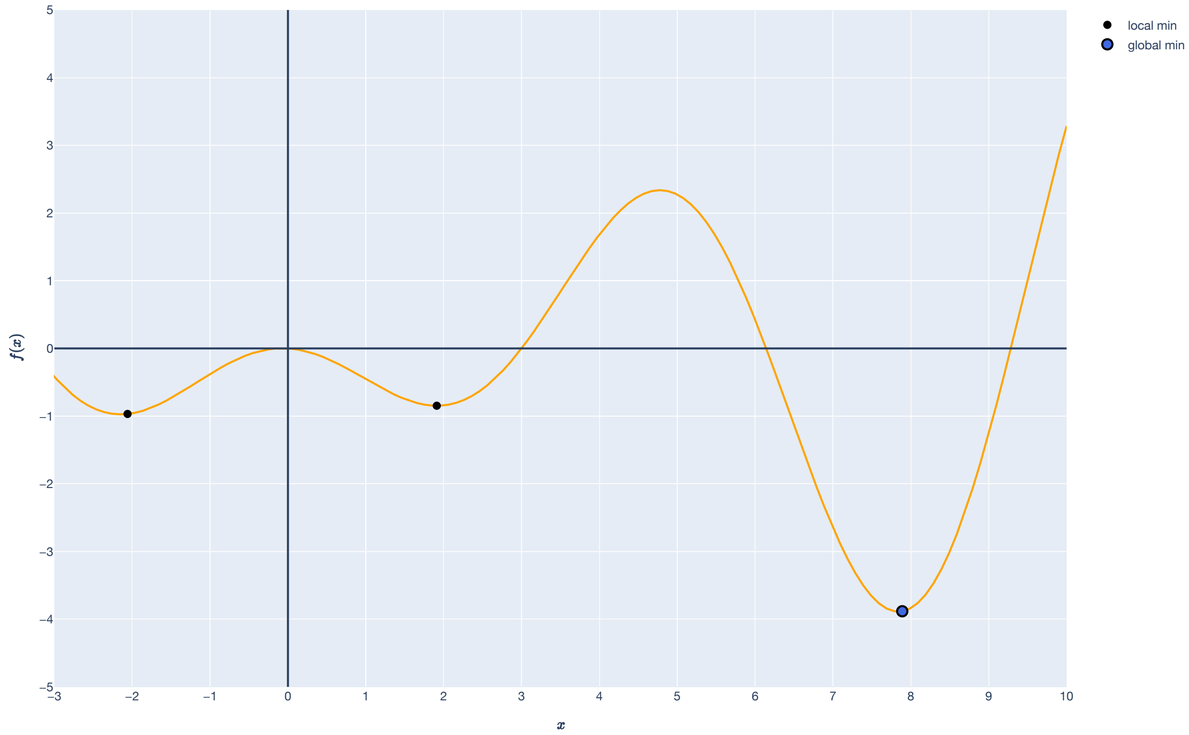
$$\nabla f(\mathbf{x}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}^*) \geq 0.$$

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

1. Find *unconstrained local minima* from first-order condition $M := \{\mathbf{x}^* \in \mathbb{R}^d : \nabla f(\mathbf{x}^*) = \mathbf{0}\}$.

2. There are no boundary points!
   $(B := \mathscr{C} \backslash \text{int}(\mathscr{C}) = \{\mathbf{x} \in \mathscr{C} : \mathbf{x} \notin \text{int}(\mathscr{C})\} = \varnothing)$

3. The global minimum must be in the set $M$, so evaluate $f$ on all $\mathbf{x} \in M$.



# Unconstrained Minima

## Example

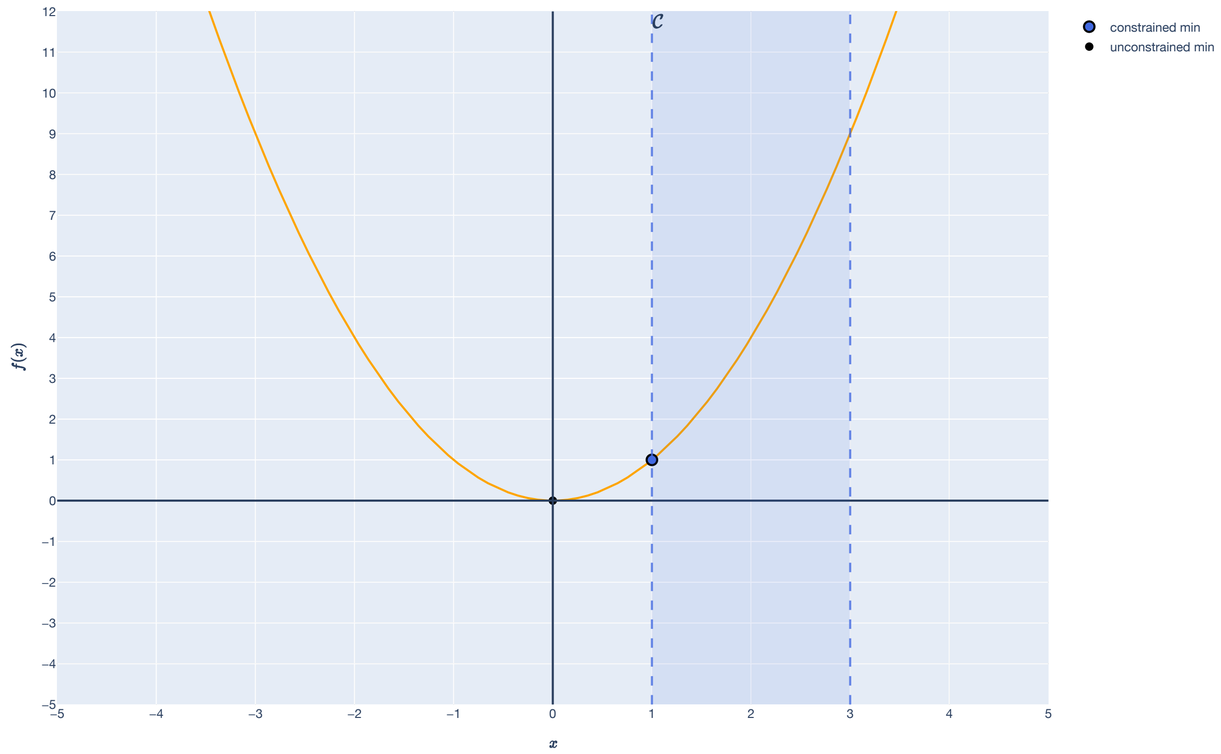
$$\begin{array}{ll} \text{minimize} & x^2 \\ \text{subject to} & x \in [1,3] \end{array}$$

When $f : \mathbb{R} \to \mathbb{R}$ is one-dimensional on $\mathscr{C} = [a, b]$ and differentiable on $\mathbf{int}(\mathscr{C}) := (a, b)$.

# Unconstrained Minima

## Example

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




When $f : \mathbb{R} \to \mathbb{R}$ is one-dimensional on $\mathscr{C} = [a, b]$ and differentiable on $\text{int}(\mathscr{C}) := (a, b)$.

# Unconstrained Minima

## Example: Why haven't we solved optimization?

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

Need to evaluate $f$ on the infinite number of points on the boundary of the circle,

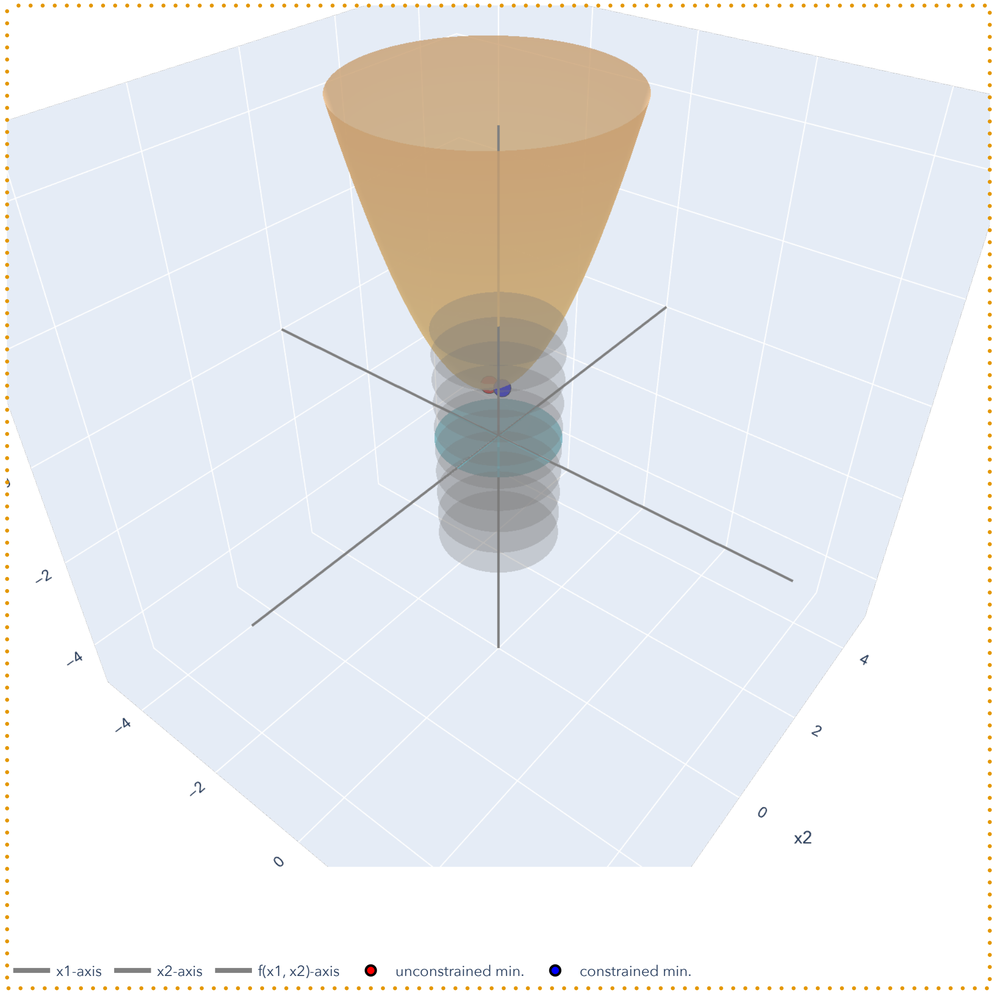
$$\mathscr{C} \backslash \text{int}(\mathscr{C}) := \{\mathbf{x} \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}!$$

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

# Unconstrained Minima

## Example: Why haven't we solved optimization?

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



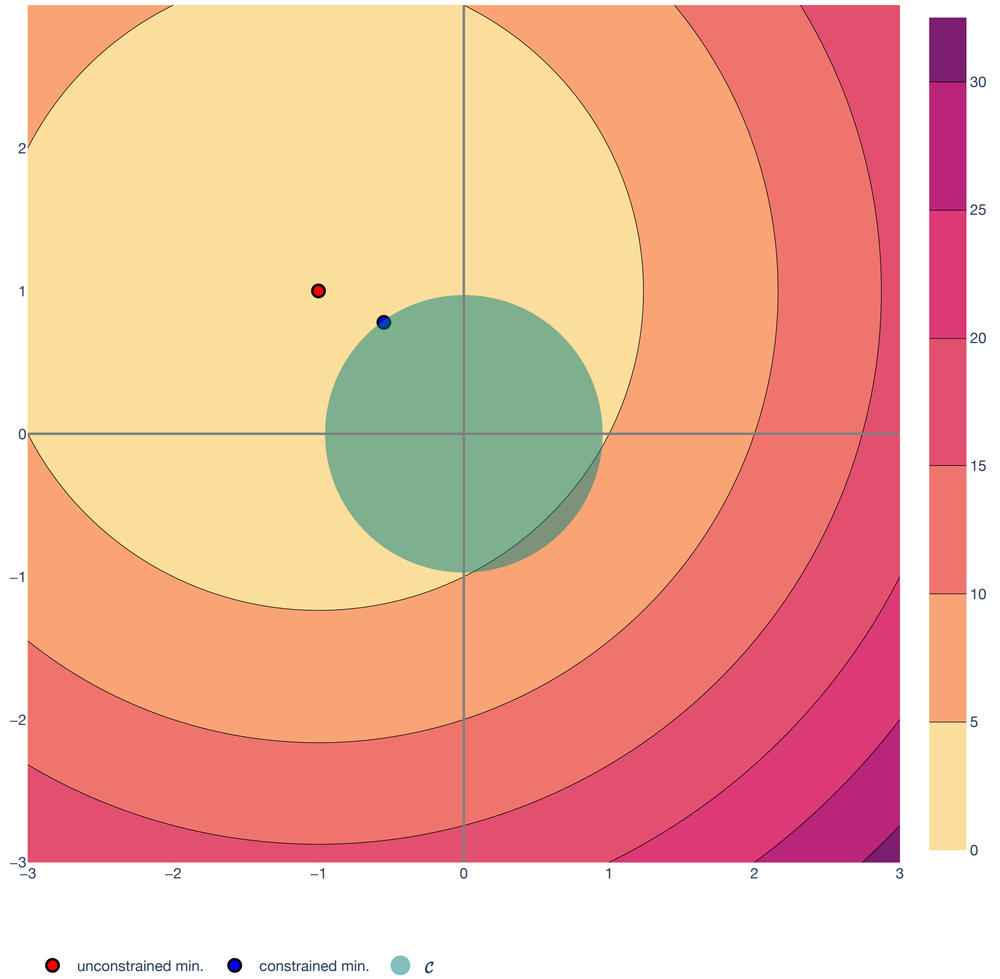

Need to evaluate $f$ on the infinite number of points on the boundary of the circle, $\mathscr{C} \backslash \text{int}(\mathscr{C}) := \{\mathbf{x} \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$!

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

# Unconstrained Minima

## Example: Why haven't we solved optimization?

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




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

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

# Constrained Minima

## Equality Constraints and the Lagrangian

# Types of Minima

## Which type of minima are each of these points?

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

constrained local:

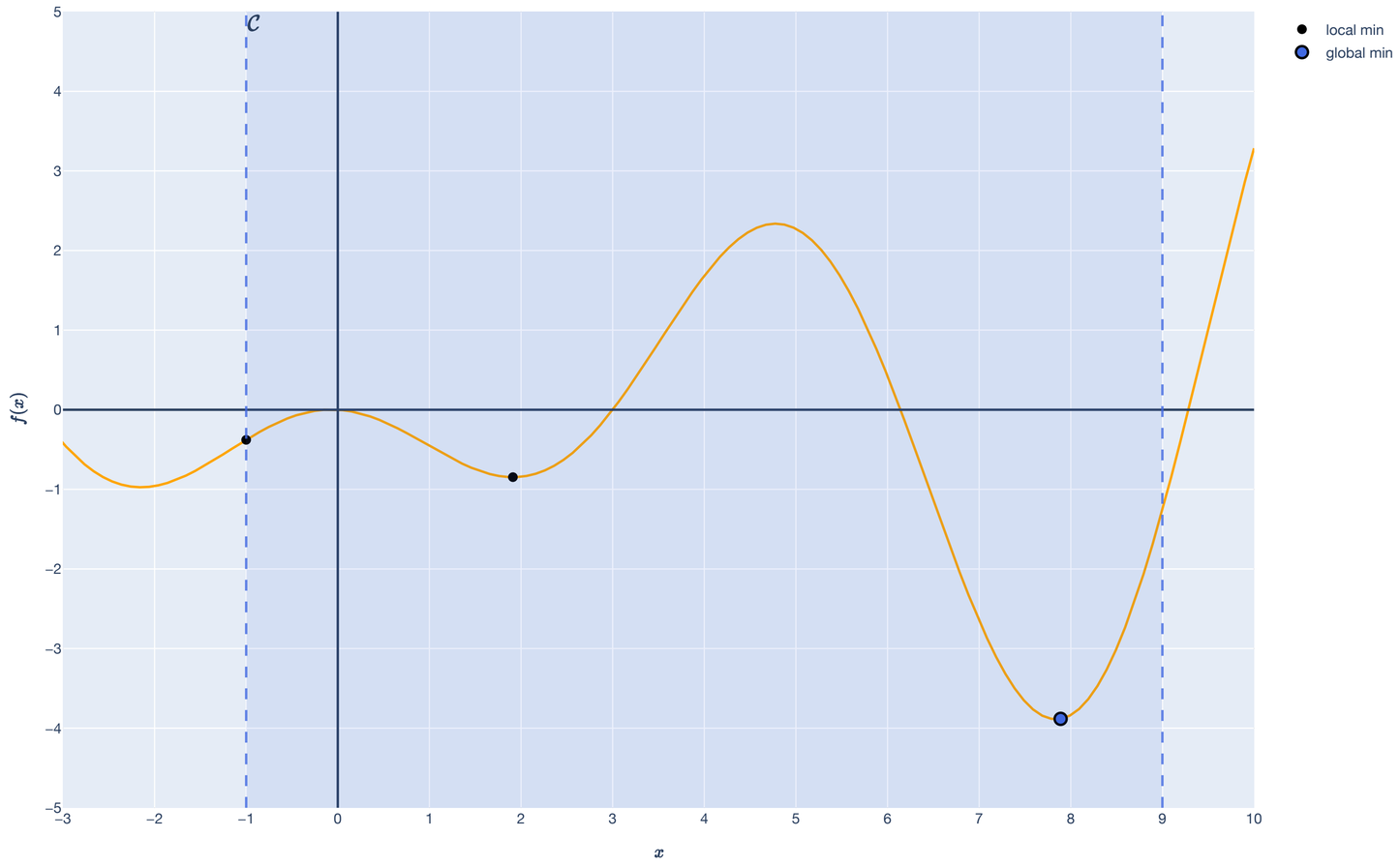
$f(\hat{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{C} \cap B_\delta(\hat{\mathbf{x}})$

unconstrained local:

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

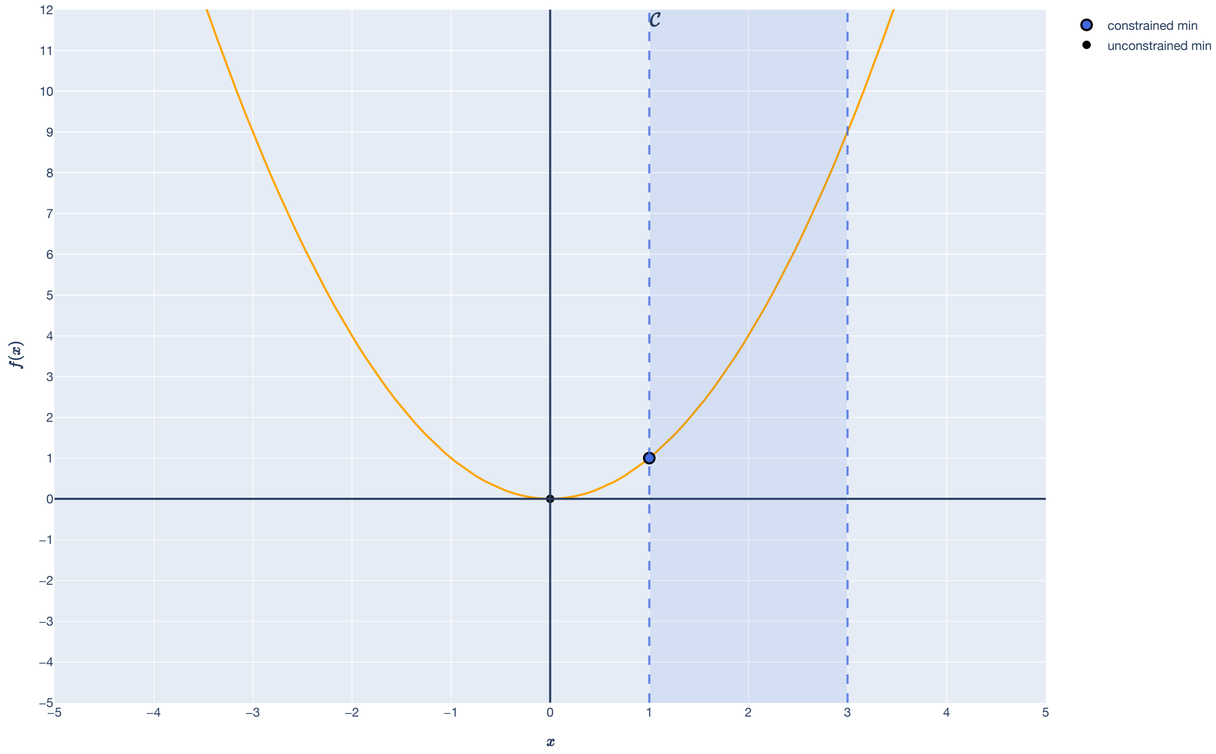
global:

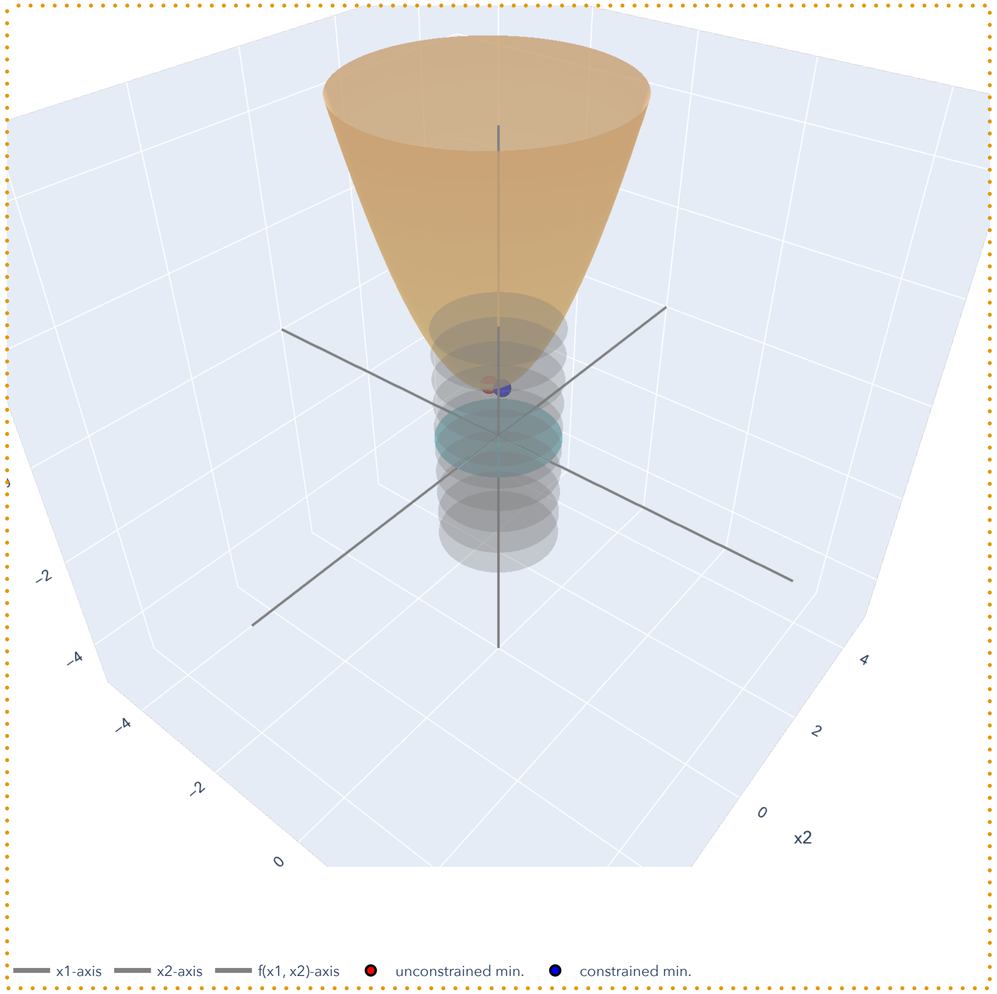
$f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{C}$.



# Constrained Local Minima

## Minimum values on the “edge of the constraint set”



# Constrained Minima

## Equality constrained optimization

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

objective function

equality constraints

Objective function $f : \mathbb{R}^d \to \mathbb{R}$ like before.

$h_1, \ldots, h_m$ are $\mathscr{C}^1$ functions $h_i : \mathbb{R}^d \to \mathbb{R}$ that form $\mathscr{C}$, the constraint set.

# 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 without loss of generality:

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

# Constrained Minima: Equality Constraints

## Example: Maximum Volume Box

$$\begin{aligned} \text{minimize} \quad & x_1 x_2 x_3 \\ \text{subject to} \quad & x_1 x_2 + x_2 x_3 + x_1 x_3 - c/2 = 0 \end{aligned}$$

Objective function: $f(\mathbf{x}) = x_1 x_2 x_3$

Single equality constraint: $h : \mathbb{R}^3 \to \mathbb{R}$, defined as $h(\mathbf{x}) = x_1 x_2 + x_2 x_3 + x_1 x_3 - c/2$.

# Constrained Minima: Equality Constraints

## Idea

Convert *constrained* optimization problem into an *unconstrained* optimization problem.

Then deal with unconstrained problem as we did before:

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

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

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

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

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

# Constrained Minima: Equality Constraints

## Regularity Conditions

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

A point $\mathbf{x} \in \mathbb{R}^d$ is a regular point if:

1. $\mathbf{x}$ is feasible, i.e. $h_1(\mathbf{x}) = 0, \ldots, h_m(\mathbf{x}) = 0$.

2. The gradients $\nabla h_1(\mathbf{x}), \ldots, \nabla h_m(\mathbf{x})$ are linearly independent.

*Constraints are "non-redundant." This is a property of how we write down our problem.*

# Constrained Minima: Equality Constraints

## Lagrange Multiplier Theorem: Necessary Conditions

**Theorem (Lagrange Multiplier Theorem - Necessary).** 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

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

# Constrained Minima: Equality Constraints

## Lagrange Multiplier Theorem: Necessary Conditions

**Theorem (Lagrange Multiplier Theorem - Necessary).** 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

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

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 $\mathbf{d} \in \mathbb{R}^d$ such that $\nabla h_j(\mathbf{x}^*)^\top \mathbf{d} = 0$ for all $j \in [m]$.

# Constrained Minima: Equality Constraints

## How to remember the Lagrange multiplier theorem

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

Remember the necessary conditions for unconstrained local minima:

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

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

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

Notice that $\nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = \mathbf{0}$ is the same as requiring feasibility: $h_j(\mathbf{x}^*) = 0$ for all $j \in [m]$.

# 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

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

$$\mathbf{d}^\top \nabla^2_{\mathbf{x},\mathbf{x}} L(\mathbf{x}^*, \lambda^*)\mathbf{d} > 0, \text{ for all } \mathbf{d} \in \mathbb{R}^d \text{ such that } \nabla h_j(\mathbf{x}^*)^\top \mathbf{d} = 0 \text{ for all } j \in [m].$$

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

# Constrained Minima: Equality Constraints

## How do we use the Lagrangian?

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

Assuming a global minimum exists, to find it…

1. Find the set $(\mathbf{x}^*, \lambda^*)$ of regular points satisfying the first-order necessary conditions:

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

2. Find the set of all non-regular points.

3. The global minima must be among the points in (1) or (2).

# Constrained Minima: Equality Constraints

## Example: Maximum Volume Box

$$\begin{aligned} \text{minimize} \quad & x_1 x_2 x_3 \\ \text{subject to} \quad & x_1 x_2 + x_2 x_3 + x_1 x_3 - c/2 = 0 \end{aligned}$$

# Constrained Minima

## Inequality Constraints and the KKT Theorem

# Constrained Minima

## Inequality constrained optimization

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

Objective function $f : \mathbb{R}^d \to \mathbb{R}$ like before.

$h_1, \ldots, h_m$ are $\mathscr{C}^1$ functions $h_i : \mathbb{R}^d \to \mathbb{R}$ that form $\mathscr{C}$, the constraint set.

$g_1, \ldots, g_r$ are $\mathscr{C}^1$ functions $g_i : \mathbb{R}^d \to \mathbb{R}$ that form $\mathscr{C}$, the constraint set.

# Constrained Minima

## Inequality constrained optimization

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

**To solve:** Reduce to *equality constrained optimization*.

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

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

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

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

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

are linearly independent.

# Constrained Minima: Inequality Constraints

## Lagrangian in Inequality Constrained Optimization

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

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

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

# Constrained Minima: Inequality Constraints

## Karush-Kuhn-Tucker (KKT) Theorem

**Theorem (KKT Theorem - Necessary Conditions).** 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

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

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).

# Constrained Minima: Inequality Constraints

## Karush-Kuhn-Tucker (KKT) Theorem

**Theorem (KKT Theorem - Necessary Conditions).** 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

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

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).

If, in addition, $f$ and the $h_i$ are all 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 $\mathbf{d} \in \mathbb{R}^d$ such that $\nabla h_j(\mathbf{x}^*)^\top \mathbf{d} = 0$ for all $j \in [m]$.

# Constrained Minima: Inequality Constraints

## Karush-Kuhn-Tucker (KKT) Theorem

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

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

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

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

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

for all $\mathbf{d}$ such that $\nabla h_i(\mathbf{x}^*)^\top \mathbf{d} = 0$ for all $i \in [m]$ 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?

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

Assuming a global minimum exists, to find a global minimum…

1. Find the set $(\mathbf{x}^*, \lambda^*, \mu^*)$ satisfying the necessary conditions:

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

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

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

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

# Least Squares Regression

## Regularization and Ridge Regression

# Regression

## Setup (Example View)

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

$$\mathbf{X} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d.$$

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 (Feature View)

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

$$\mathbf{X} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & & \downarrow \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \text{ where } \mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{R}^n.$$

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

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:

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

# 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:

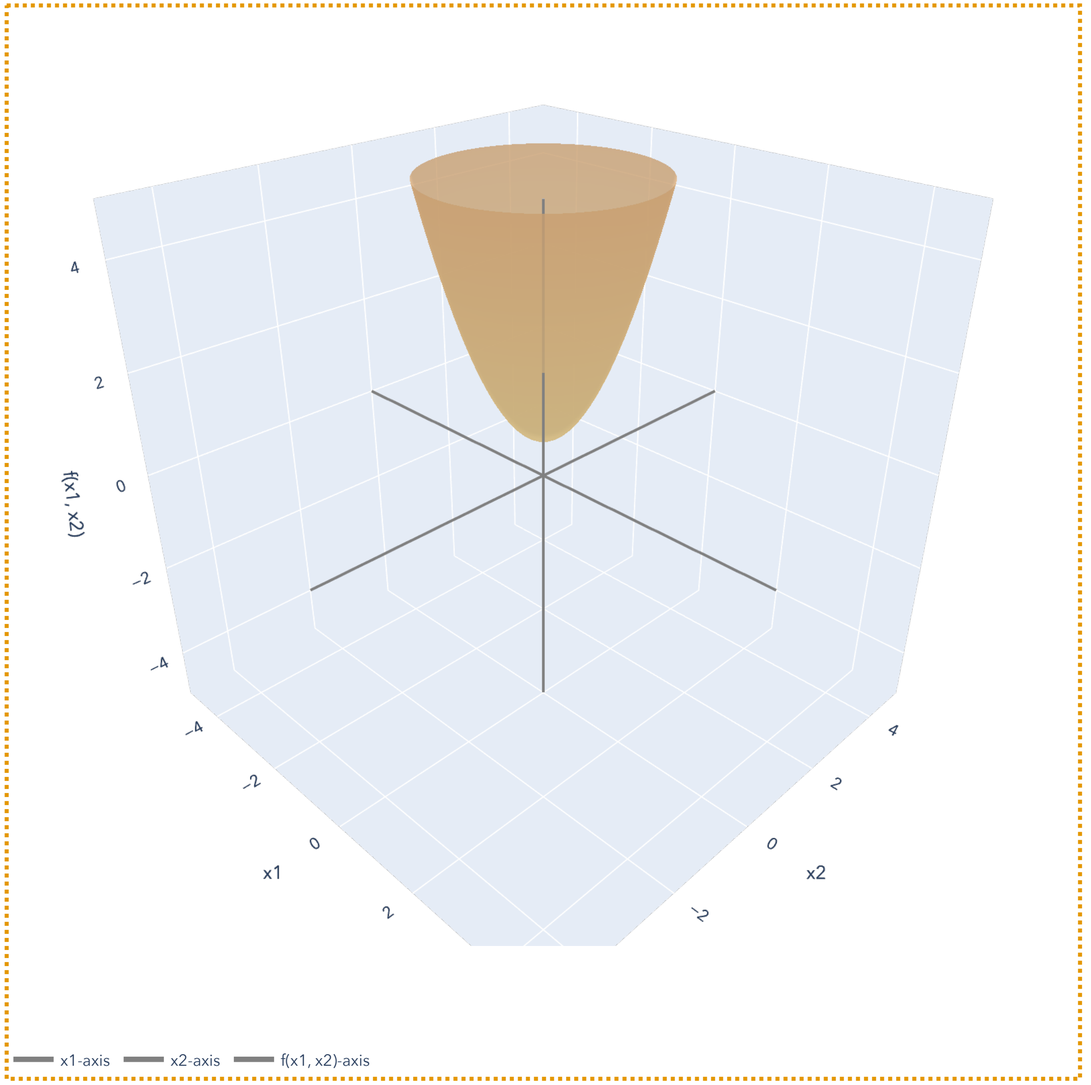
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

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

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

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

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



# Least Squares

## Least norm exact solution

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

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

*We already know how to solve this – use the pseudoinverse!*

# Least Squares

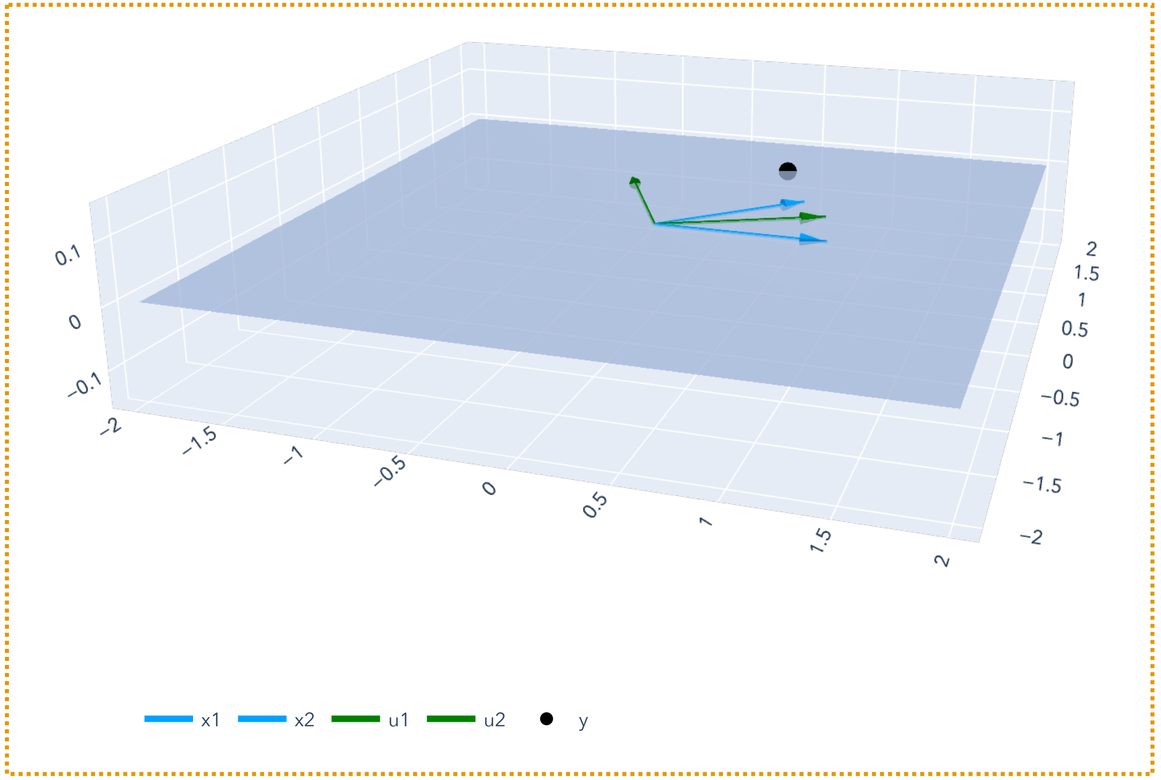
## Least norm exact solution

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

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

<u>Theorem (Minimum norm least squares solution).</u>
Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, and let $\mathbf{rank}(\mathbf{X}) = n$. Then, $\hat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\mathsf{T}}\mathbf{y}$ 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

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

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

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

# Least Squares

## Least norm exact solution

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

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

Theorem (Minimum norm least squares solution). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, let $d \geq n$, and let $\text{rank}(\mathbf{X}) = n$. 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

## 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/$\ell_2$-regularized regression.

# Least Squares

## Ridge Regression

Our goal will now be to minimize two objectives:

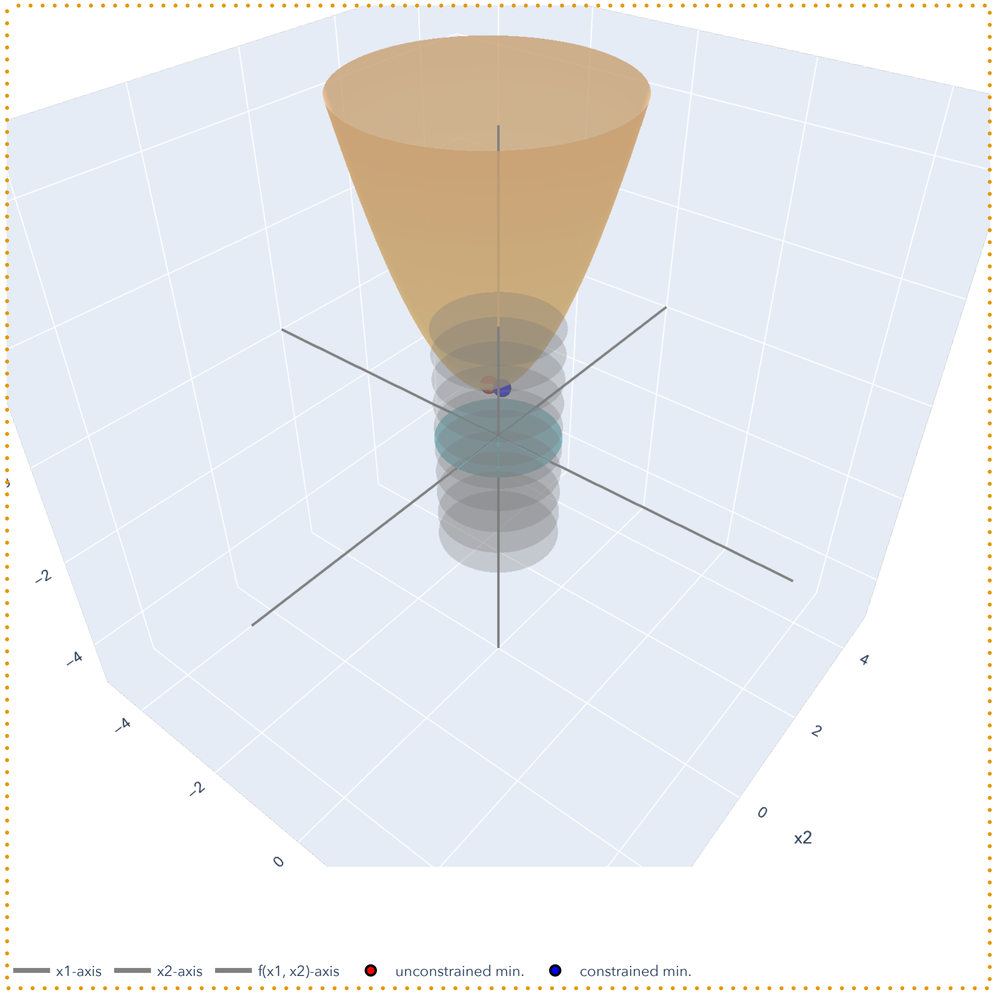
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

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/$\ell_2$-regularized regression.



# Least Squares

## Ridge Regression

Our goal will now be to minimize two objectives:

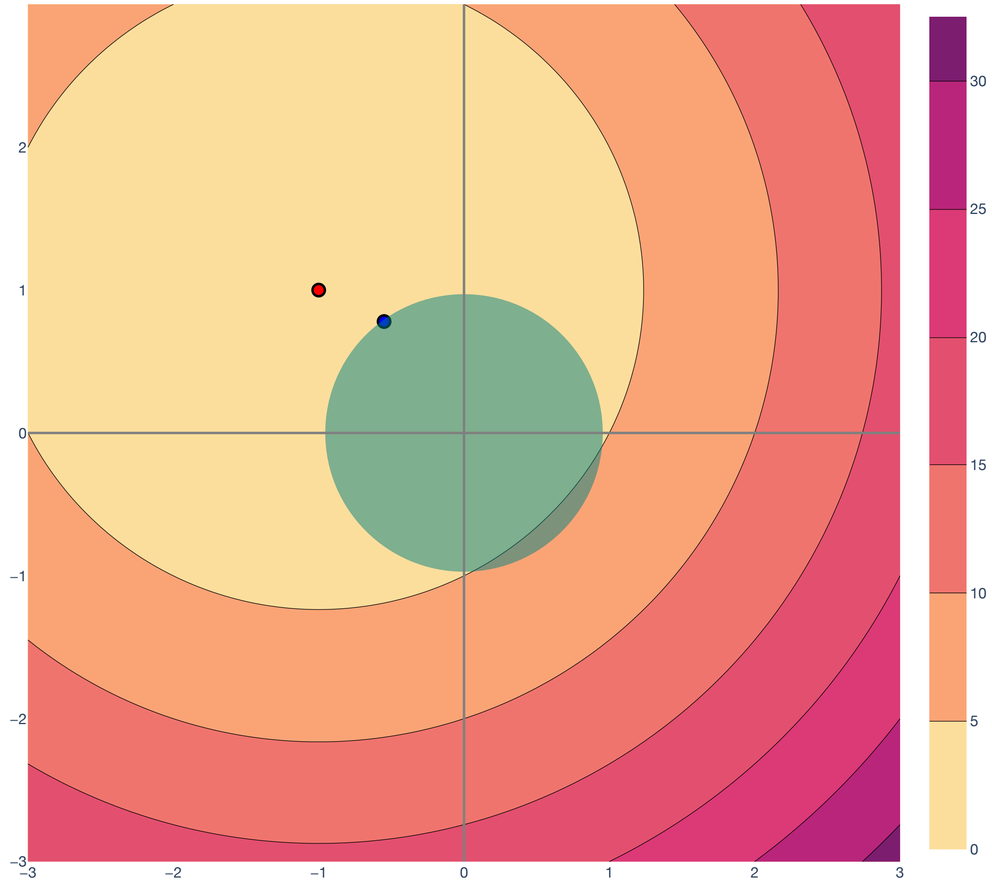
$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ and } \|\mathbf{w}\|^2.$$

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/$\ell_2$-regularized regression.




*For bigger $\gamma$, bigger "constraint" ball!*

# Ridge Regression

## Property: PSD to PD matrices

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

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

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

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

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

# Ridge Regression

## First-order conditions

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

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

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

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

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

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

# Least Squares

## Solving ridge regression

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

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

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

Taking the Hessian,

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

*Sufficient condition for optimality applies!*

# Ridge Regression

## Theorem

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

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

# Least Squares

## Comparison with ridge solution

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

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

has the form:

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

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

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

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

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

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

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

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

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

# 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:

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

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* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$.

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

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

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

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

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

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

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

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

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

When $\epsilon \neq 0$, we are off by $\hat{\mathbf{w}} - \mathbf{w}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon$.

# Error in (OLS) Regression

## Eigendecomposition perspective

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:

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

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

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

# Error in Regression

## Error using ridge regression

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

What happens when we use the ridge regression 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:

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

# Error in Regression

## Error using ridge regression

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

What happens when we use the ridge regression 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

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

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

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

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

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

# Least Squares

## Ridge Regression

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

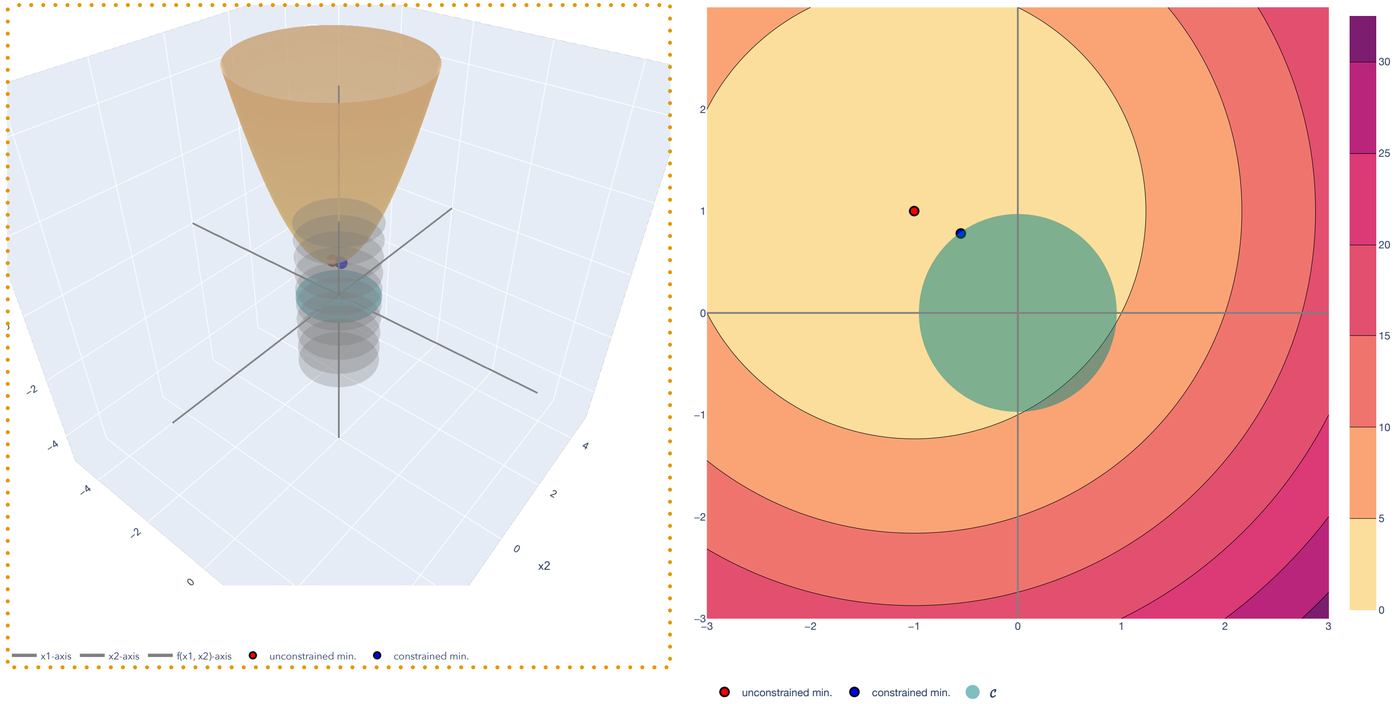
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma \|\mathbf{w}\|^2$$

has the form:

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

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

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




*For bigger $\gamma$, bigger "constraint" ball!*

# Recap

# Lesson Overview

**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.

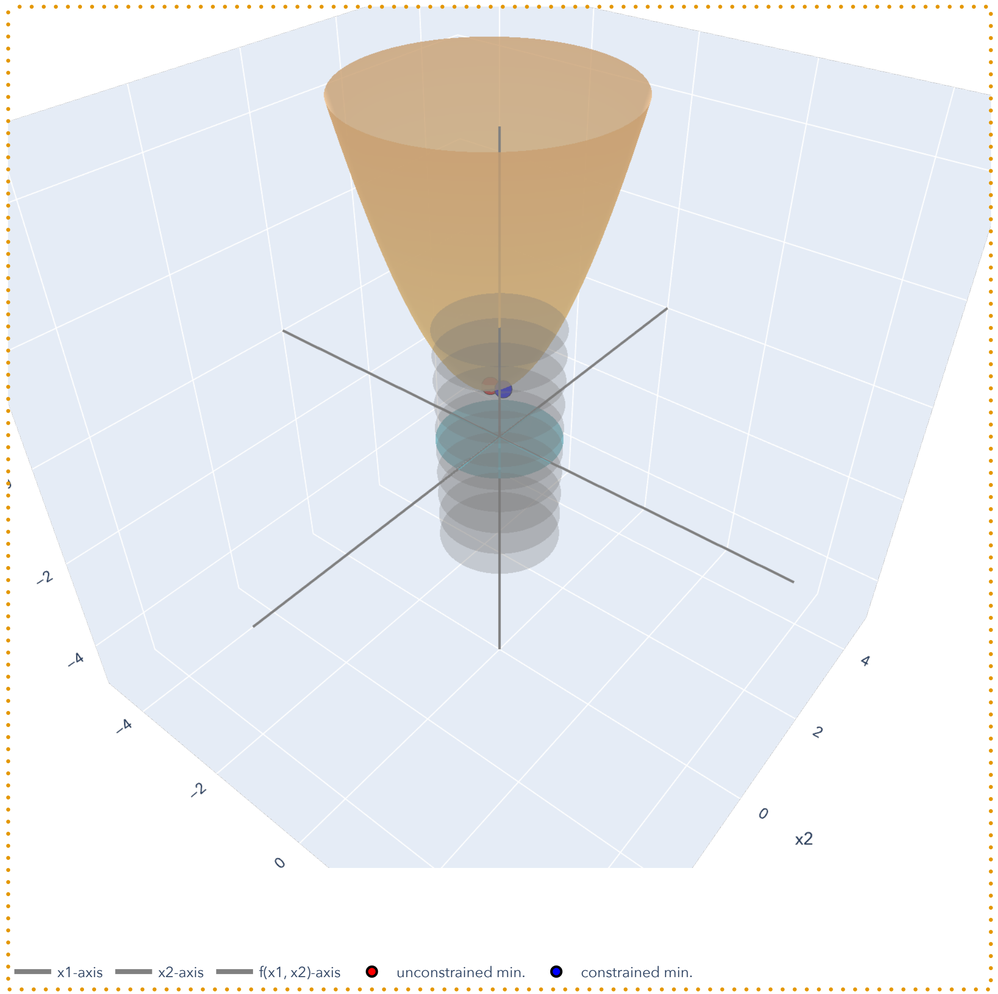
**Unconstrained local optima.** With no constraints, the standard tools of calculus give conditions for a point $\mathbf{x}^*$ to be optimal, at least to all points close to it.

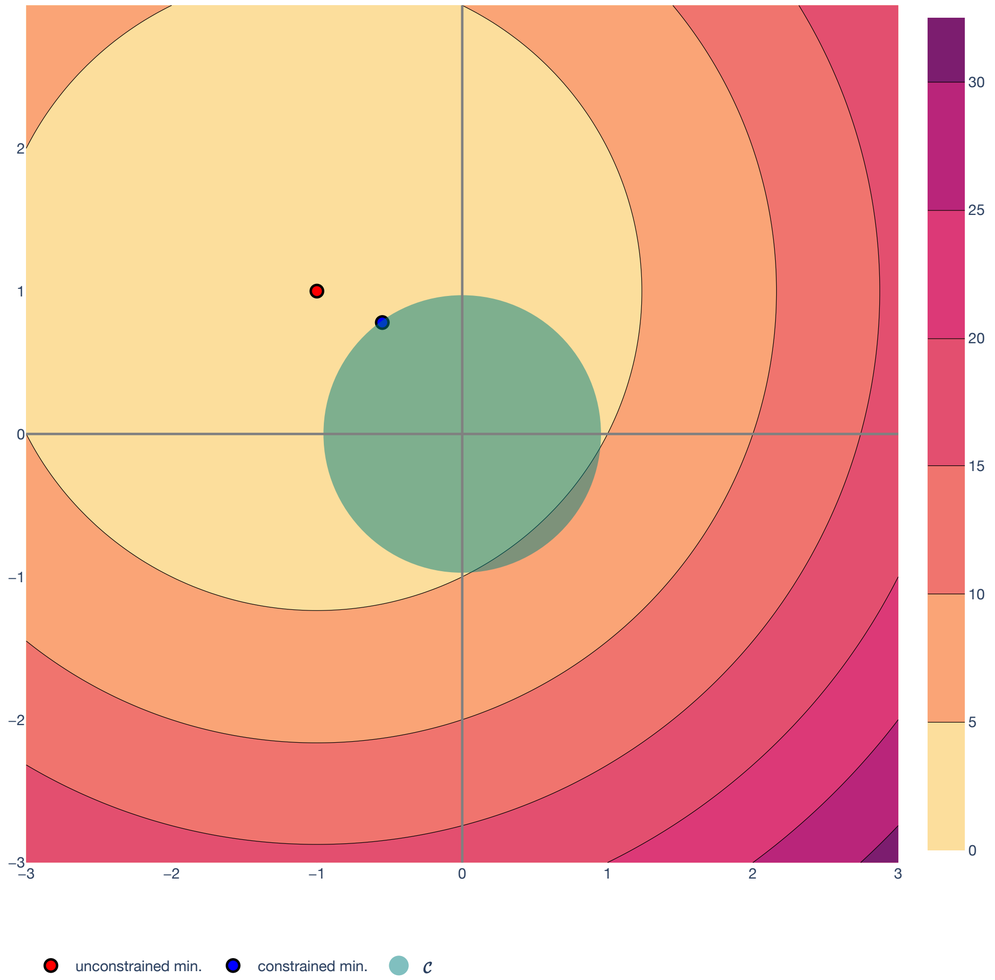
**Constrained local optima (Lagrangian and KKT).** When $\mathscr{C}$ is represented by inequalities and equalities, we can use the method of Lagrange multipliers and the KKT Theorem to "unconstrain" the problem.

**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



# Lesson Overview

## Big Picture: Gradient Descent